Department of Computer Science
Series of Publications A
Report A-2009-6

A Probabilistic Approach to the Primary Visual Cortex

Urs Köster

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium D101, Physicum, on October 5th, 2009, at 12 o'clock noon.

University of Helsinki
Finland

With support from the


Alfried Krupp von Bohlen und Halbach-Stiftung

Contact information

Postal address:
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: postmaster@cs.helsinki.fi (Internet)

URL: http://www.cs.helsinki.fi/

Telephone: +358 9 1911

Telefax: +358 9 191 51120

ISSN 1238-8645
ISBN 978-952-10-5714-4 (paperback)
ISBN 978-952-10-5715-1 (PDF)
Computing Reviews (1998) Classification: I.5.1, I.2.10
Helsinki 2009
Helsinki University Print

A Probabilistic Approach to the Primary Visual Cortex

Urs Köster

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
urs.koster@cs.helsinki.fi
http://cs.helsinki.fi/u/koster

PhD Thesis, Series of Publications A, Report A-2009-6
Helsinki, September 2009, 168 pages
ISSN 1238-8645
ISBN 978-952-10-5714-4 (paperback)
ISBN 978-952-10-5715-1 (PDF)

Abstract

What can the statistical structure of natural images teach us about the human brain? Even though the visual cortex is one of the most studied parts of the brain, surprisingly little is known about how exactly images are processed to leave us with a coherent percept of the world around us, so we can recognize a friend or drive on a crowded street without any effort.

By constructing probabilistic models of natural images, the goal of this thesis is to understand the structure of the stimulus that is the *raison d'être* for the visual system. Following the hypothesis that the optimal processing has to be matched to the structure of that stimulus, we attempt to derive computational principles, features that the visual system should compute, and properties that cells in the visual system should have.

Starting from machine learning techniques such as principal component analysis and independent component analysis we construct a variety of statistical models to discover structure in natural images that can be linked to receptive field properties of neurons in primary visual cortex such as simple and complex cells. We show that by representing images with phase invariant, complex cell-like units, a better statistical description of the visual environment is obtained than with linear simple cell units, and that complex cell pooling can be learned by estimating both layers of a two-layer model of natural images.

We investigate how a simplified model of the processing in the retina, where adaptation and contrast normalization take place, is connected to the natural stimulus statistics. Analyzing the effect that retinal gain control has on later cortical processing, we propose a novel method to perform gain control in a data-driven way. Finally we show how models like those presented here can be extended to capture whole visual scenes rather than just small image patches. By using a Markov random field approach we can model images of arbitrary size, while still being able to estimate the model parameters from the data.

Computing Reviews (1998) Categories and Subject Descriptors:
I.5.1 Models: Statistical
I.2.10 Vision and Scene Understanding: Representations, Data Structures and transforms

General Terms:
Vision, Computational Neuroscience, Unsupervised Machine Learning

Additional Key Words and Phrases:
Natural Image Statistics, Score Matching, Independent Component Analysis

Acknowledgements

First and foremost I wish to thank Aapo Hyvärinen, who was always there for discussions, and gave me just the right amount of supervision for my PhD. Helping me out with ideas when I asked for it, but just as happy to leave me to work on a problem on my own, he taught me the perseverance and state of mind required for academic research.

I would like to acknowledge the *Alfried Krupp von Bohlen und Halbach-Stiftung* and especially Prof. Dr. mult. h.c. Berthold Beitz, who provided me with funding throughout my PhD. I am deeply grateful for the way the *Stiftung* decided to support me even though none of their programs included funding for a PhD abroad and in computer science. Additionaly I thank the HeCSE graduate school and the Academy of Finland for funding.

Especially heartfelt thanks go to my friends and colleagues Jussi Lindgren and Michael Gutmann, without whom many of the ideas this work is based on would not have reached maturity. Helpful discussions with Patrik Hoyer, especially during the early stages of my PhD, were instrumental in introducing me to the world of independent component analysis and natural image statistics. A special contribution to my thesis was made by Malte Spindler who designed the cover artwork. My friend David C.J. Senne deserves thanks for comments on the manuscript and much disport.

In particular, I wish to thank my family: my brother Malte, my father Ulrich and especially my mother Barbara, who did everything in her power to pave the way for an academic career for me. Last, but most certainly not least, I thank the crew at home for being around and keeping me in touch with the world outside academia.

Contents

1

Introduction

Can ye make a model of it?
If ye can, ye understands it,
and if ye canna, ye dinna!
- Lord Kelvin -

1.1 The Challenge of Vision

From our personal experience, vision seems like an automatic process which does not require any conscious effort. In cluttered environments with many competing stimuli, objects can easily be distinguished from backgrounds and identified reliably, even if we have never seen the object at this particular angle, under these particular lighting conditions, or in this particular context before. All in all, vision seems like child's play.

Decades of research into human and machine vision tell a different story.

While vision seems so effortless to us, it is one of the hardest problems that the human brain has to solve. The visual cortex is organized into a highly interconnected hierarchy of dozens of separate areas [23], analyzing visual scenes and combining the information from the stimulus with prior knowledge so a coherent percept of the visual world emerges.

Even though the visual apparatus is the most-studied part of the brain, having drawn the attention of investigators as early as Descartes [19] (see Fig. 1.1), we are far from understanding the neural basis of human vision. After countless studies using methods such as psychophysics, electrophysiology and fMRI (functional magnetic resonance imaging), we have just started scratching the surface and are only beginning to understand what mechanisms the human visual system is employing to pick out an object

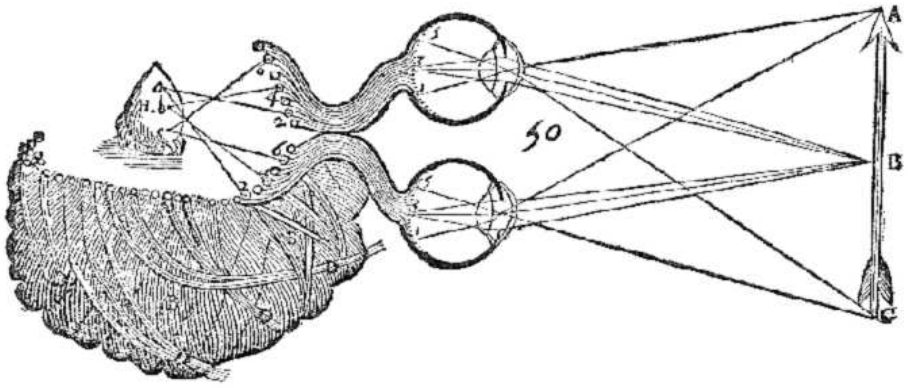


Figure 1.1: In his work *Traité de l'homme* (1664) Descartes gives one of the earliest accounts of visual perception. He postulated the *pineal gland* to be the interface between body and soul, and believed that visual information was relayed to this gland so we can consciously perceive it.

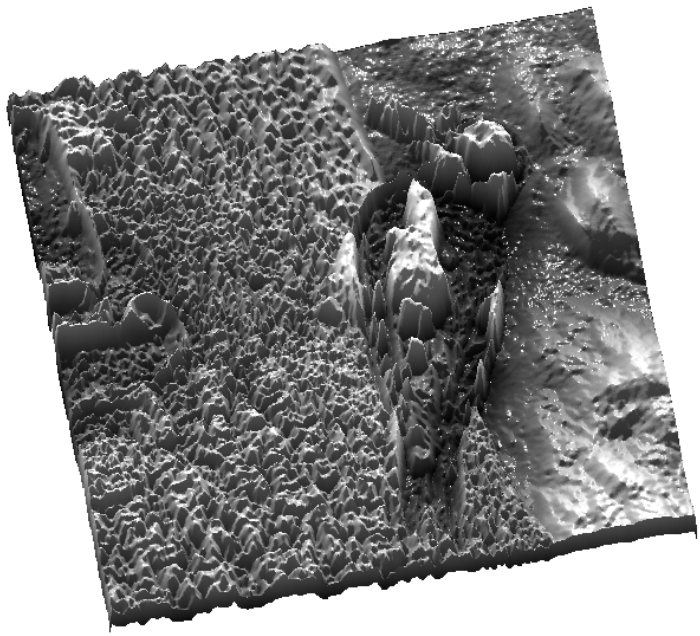


Figure 1.2: A natural image presented in a slightly different way than usually: shades of gray are mapped to elevation in a 3D surface plot. While the information is almost the same as in the original image, it is nearly impossible to tell what the content of the image is. For the curious, the same image is displayed in its ordinary form in Fig. 1.3.

from a cluttered environment or to recognize a familiar face [29].

To get an intuitive feeling for how hard the seemingly trivial process of vision is, consider Fig. 1.2: it shows an image represented in such a way that, while most of the raw information is preserved, many of the cues we take for granted have been distorted or disappeared altogether. This makes it virtually impossible to tell what the image contains. Another way to get a feeling for the sheer complexity of visual perception is to look at the metabolic resources that humans devote to vision. About one quarter of the cortical surface in the brain is dedicated to visual processing [29]. While the brain makes up only 2% of the mass of the human body, it consumes 20% of the energy [14], so an enormous fraction of our total energy intake is consumed just for visual processing.

Understanding the workings of the visual system is not only of interest

Figure 1.3: The image from the previous page in its ordinary form. It depicts a great spotted woodpecker.

to neuroscientists, but also to a variety of fields in engineering and computer science. It is notoriously difficult to design computer vision systems that perform well under real world conditions [108]. Many systems for object recognition [78] have a set of build-in invariances and perform well under the conditions they are designed for, but fail when faced with the great complexity of natural scenes. Inspiration from how the human brain is solving the problem seems to be needed.

In a similar way, image processing is intertwined with biological vision in several ways: reconstruction of missing regions in an image such as *filling-in* or *inpainting* [11] is a problem faced also by the visual system, e.g. when parts of an object are occluded. Denoising based on image priors becomes necessary in low light conditions when the visual signal is limited by photon shot noise, and superresolution [125] is conceivably important in the periphery of the retina where sampling is very sparse. A different kind of example is lossy image compression, where detailed knowledge about visual processing might be used to discard information that the visual system does

not pay attention to.

Obviously, these engineering problems and neuroscientific questions are connected by the properties of the stimulus.

Based on the properties of the visual signal, it is possible to infer much of the required processing of the visual system, without ever having to specify goals such as object detection or classification.

In the 1980's David Marr [82] proposed a theory of visual processing that is highly regarded for its contribution to computer vision. He identified the main goal of the visual system to be the reconstruction of a 3D world from a 2D stimulus, an ill-posed problem that requires prior information about the signal. A key idea in his work is that the algorithms and representations required for vision are distinct from the implementation in the brain, and can be analyzed as a purely computational problem. Similarly, the psychologist James Gibson [30] studied perception under the premise that the properties of the environment dictate many of the properties of the visual system. Another proponent of this *ecological approach* to vision and perception in general was Horace Barlow [7]. In his seminal paper he concluded that in encoding sensory messages, the nervous system should remove redundancy from the stimulus to arrive at an efficient code. This of course requires knowledge about the environment and the statistical structure of sensory signals.

From this early work, combined with advanced statistical techniques like independent component analysis (ICA) [16, 116] a whole field has emerged trying to use the statistical structure of ecologically valid stimuli to infer the optimal processing and understand - or even predict - what kind of processing the visual system is performing. This is the line of work we are following in this thesis.

1.2 Scope of this Work

Over the last two decades, the study of natural image statistics has grown into an important research field. Key properties of the early visual system have been explained as being optimal in a statistical sense. Visual processing seems to be matched to the statistical structure of the environment to better be able to infer the state of the environment from incomplete or noisy stimuli. Some of the receptive field properties of cells in the retina and primary visual cortex can be reproduced by optimizing statistical criteria such as reducing redundancy and maximizing independence between

cells.

Possibly the greatest limitation of the previous work has been that mostly linear models were considered, and only a single linear transformation was estimated from the data. It is clear that to obtain the kind of invariant representations that are required for vision in a natural environment, and that have been found in visual neurons, highly nonlinear transformations of the stimulus are required. Only in recent years it has become possible to build *hierarchical, multi-layer* models which capture more of the structure of the signal by forming nonlinear, invariant representations.

In this work we present advances on several multi-layer models, some of which have only been made possible through new statistical methods developed during recent years. We consider models of *complex cells*, and show that pooling of linear filters provides a better statistical description of the stimulus than a simple linear model. We continue to show a method for learning the optimal pooling from the data, rather than using a fixed pooling. In addition, we consider the effect of incorporating non-linear *gain control* in our models, to obtain a better statistical description of the stimulus. Finally, we consider the problem of extending models for small, localized *patches* of natural images to models for larger natural scenes. We show how this can be done using only local interactions, which makes it computationally tractable to work with high-dimensional stimuli.

The structure of the first part of this thesis is as follows: we give a short introduction to the human visual system in Chapter 2, starting with the processing at the retina and describing some of the visual areas of the cortex. In the first part of that chapter, we focus on *what* kind of features the visual system is computing, i.e. what kind of receptive fields visual neurons have, and look at some of the representations that are formed at various stages of the visual hierarchy. In the second part of Chapter 2 we consider some of the classical models for early visual processing and investigate *how* the visual system can compute certain features. Some of the mechanism we consider have been proposed to be implemented in the neural hardware, whereas others are on a very abstract level and we make no attempt to hypothesize possible neural implementations.

Investigating *what* kind of features the visual system is computing, and *how* this is achieved is begging the question as to *why* it is necessary or at least advantageous to perform these computations. This question is addressed in Chapter 3, where we use the statistical structure of natural images to derive the optimal features with which natural images should be processed. In this chapter much of the earlier work that this thesis is building upon is introduced, and the mathematical and computational

framework in which this thesis is rooted is described in detail.

In Chapter 4 we introduce the publications in the second part of this thesis. We give an introduction to independent subspace analysis and describe some of the results in *Publications 1 and 2*, as well as motivating *Publications 3 and 4*. We discuss the importance of gain control for simple and complex cell in the context of *Publication 5* and finish with a short introduction to Markov random fields in the context of *Publication 6*.

In the final chapter we discuss how the various contributions in this thesis and previous work relate, and where this leaves us in terms of understanding the visual processing in the brain.

1.3 Problem Statement and Research Questions

To investigate what the goal of processing in the primary visual cortex is, we are going to exploit the connection between this processing and the statistical structure of natural images. This naturally breaks down into a number of subproblems. The statistics of natural images are not yet well understood, so the first step is a better characterization of this structure. The second step then is to link the statistical properties to the constraints and goals of the visual system. We will focus on the first part of the problem, building models of image patches that capture as much as possible of their structure. In particular we focus on multi-layer models with more than one layer of features estimated from the data. Thus our primary research question can be cast as:

RQ1: What are suitable statistical models for patches of natural images?

A model is only as good as the estimation methods that are available to fit its parameters. In the past, many promising approaches have found their demise because the estimation was prohibitively expensive in terms of computational resources or could not be scaled up to high-dimensional data. An equally important question to the first is then:

RQ2: How can multi-layer models of natural images be estimated?

After considering the rather general aspects of models and estimation methods, we turn our attention to connecting these models to the properties of the visual system. In particular, we analyze the statistical utility of orientation-selective, phase-invariant complex cell responses. This question can be phrased as:

RQ3: Can we show that complex cell-like units provide a better statistical description of images than linear filters?

Finally we consider how these models relate to another ubiquitous aspect of visual processing, which is gain control. The statistical structure of the visual stimulus is non-stationary, and we analyze how the optimal processing is affected by this. In particular we try to answer the question whether gain control on the pixel level has an effect on the optimal processing in later stages such as simple and complex cells. The general question we are trying to answer is thus:

RQ4: Is gain control in the visual system matched to the optimal processing of the stimulus, and how does gain control affect the later processing?

These four questions will guide us through the rest of this thesis. After exploring to which extend previous work can answer these questions and what aspects have not been addressed, we will present our contribution to these points, and analyze the results in an attempt to obtain a better understanding of the processing in the visual system.

1.4 Overview of the Publications

Publication 1: Aapo Hyvärinen and Urs Köster, "FastISA: a fast fixed-point algorithm for Independent Subspace Analysis" *ESANN2006: 14th European Symposium on Artificial Neural Networks, 371-376, 2006*

In Publication 1 we describe a new algorithm for *Independent Subspace Analysis*, FastISA, which is a generalization of the FastICA algorithm for *Independent Component Analysis*. The FastISA algorithm is simple to use and converges quickly, so it is particularly useful for researches and engineers who require a turn-key algorithm that doe not require fine-tuning.

The algorithm was conceived and originally implemented by A.H., the Author contributed the convergence proof, performed simulations and wrote the article.

Publication 2: Aapo Hyvärinen and Urs Köster, "Complex Cell Pooling and the Statistics of Natural Images" *Network: Computation in Neural Systems, 18:81-100, 2007.*

In Publication 2 we compare the likelihood of ISA models with different subspace sizes. This is made possible by formulating the likelihood of the ISA model including the subspace size as a parameter, and optimizing this parameter. In addition, we generalize from L_2-spherical subspaces to L_p-spherical, and attempt to find the optimal norm. Furthermore we investigate the effect that contrast gain control has on the optimal subspace

size. We conclude that ISA is a better model for natural images, in the sense of a statistical criterion, than ICA. The optimal subspace size strongly depends on the patch size and on preprocessing, but is always larger than one, the ICA case.

The idea for this work and the derivation of the probability density function we used were A.H.'s, the Author contributed the implementation of the algorithm, the methods for gain control, performed all experiments and wrote the article.

Publication 3: Urs Köster and Aapo Hyvärinen, "A two-layer ICA-like model estimated by Score Matching" *Proc. Int. Conf. on Artificial Neural Networks (ICANN2007), 798-807, 2007*

Publication 3 provides a generalization of the previous work on ISA to a full two-layer network. Using the theory of *score matching*, we consider a two-layer model that contains ISA and topographic ICA as special cases. We show that estimating of both layers from natural image patches leads to a pooling in the second layer like in ISA, where a few units with similar tuning properties are squared and summed together.

Using the score matching framework developed by A.H., the Author derived the model and implemented the model for gradient estimation. The author performed all experiments and wrote the article.

Publication 4: Urs Köster and Aapo Hyvärinen, "A Two-Layer Model of Natural Stimuli Estimated with Score Matching" *Submitted Manuscript*

Publication 4 generalizes Publication 3 in several ways. We show that by learning both layers in the hierarchical model simultaneously rather than one after the other, the tuning of the units changes and becomes more complex cell-like. In previous work it had been suggested that sequential estimation of the layers does not lead to a change in receptive fields [62, 91]. Furthermore we apply the model to natural sounds, which gives similar results to those for natural images.

Based on A.H.'s score matching framework, the Author derived and implemented the model, designed and performed the experiments and wrote the article.

Publication 5: Urs Köster, Jussi T. Lindgren and Aapo Hyvärinen, "Estimating Markov Random Field Potentials for Natural Images" *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA2009), 515-522, 2009*

Publication 5 describes another generalization of ICA made possible by

score matching. We consider a Markov random field (MRF) over an image, which allows us to lift the constraint of working on small image patches and generalize ICA to whole images of arbitrary size. The model needs to be trained on patches approximately twice the size of the linear filters to capture spatial dependencies extending beyond the size of the filter, and can be applied to images of arbitrary size. This approach combines the benefits of MRF models which previously used very small filters such as 3×3 pixels, and ICA which was constrained to very small images up to about 32×32 pixels.

The MRF model was conceived by the Author together with J.T.L., implementation and writing the article are the Author's work. The idea to estimate an ICA-like model for whole images in this way was originally proposed by A.H.

Publication 6: Urs Köster, Jussi T. Lindgren, Michael Gutmann and Aapo Hyvärinen, "Learning Natural Image Structure with a Horizontal Product Model" *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA2009), 507-514, 2009*

Publication 6 shows an alternative two-layer model to the hierarchical models considered previously. We consider a generative model that independently samples from two linear models representing two different aspects of the data. The outputs are then combined in a nonlinear way to generate data vectors, i.e. natural image patches. By structuring the model in a horizontal, rather than hierarchical way, we can model complex dependency structures that are more naturally represented at the pixel level, such as lighting influences, rather than having to take into account their influence on the filter outputs.

The idea was developed by the Author together with J. T. L., with small contributions from A.H. and M.G., the model, implementation and writing the article are the Author's work.

2

Vision

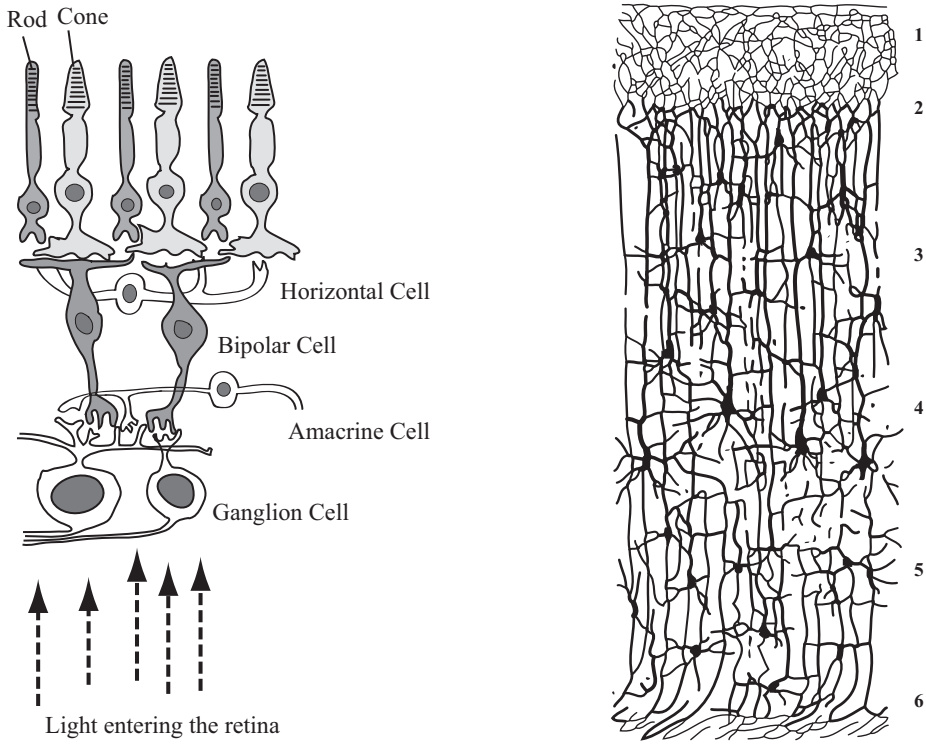
Das Auge hat sein Dasein dem Licht zu danken.
Aus gleichgültigen tierischen Hilfsorganen
ruft sich das Licht ein Organ hervor,
das seinesgleichen werde, und so bildet
sich das Auge am Lichte fürs Licht,
damit das innere Licht
dem äusseren entgegentrete.
- J. W. von Goethe -

2.1 Biological Vision

Of all human senses, vision is arguably the most important: for our ancestors and closest relatives, e.g. primates such as chimpanzees, vision is of prime importance for gathering food, spotting predators and finding mates. Therefore it is not surprising that the primate visual system is highly evolved and makes up a significant fraction of the cortex. But even very primitive organisms have surprisingly complex visual systems. The barnacle, which hardly has a nervous system at all, does not have eyes but a primitive form of vision and can respond to visual stimuli - shadows of predators passing by - by quickly retracting into its shell [33]. Scallops, still very primitive organisms, already posses image forming eyes (some 60 of them), which can detect motion, allowing them to flee from predators. The photosensitive pigment that allows the detection of light, rhodopsin, is ubiquitous across the animal kingdom, which suggest that vision has evolved very early. At the same time, differences in the architecture and protein makeup in the eyes of different animals suggest that eyes have independently emerged many times throughout evolution [33].

2.1.1 The Retina

When light enters the eye, it is focused by the cornea and the lens to form an image on the retina, which contains a number of different cells shown in Fig. 2.1. The retina has two types of light-sensitive cells, rod and cone photoreceptors. If light strikes one of the photoreceptor cells, this will inhibit the release of glutamate, a neurotransmitter. The photoreceptors are connected to two types of bipolar cells (so called because they have two extensions, the axon and the dendrite), the ON and OFF bipolar cells. Bipolar cells are sensitive to *contrast*, so rather than directly encoding the light intensity signaled by the photoreceptors, they compare the intensity in the center of their *receptive field* (RF) to the intensity in the surroundings of this central spot. ON bipolar cells react to a bright center with a comparatively dark surround, whereas OFF cells fire in response to relative darkness in the center [29]. These receptive fields are illustrated in Fig. 2.3 a). There are more than 10 different kinds of bipolar cells specialized for processing of color, temporal information and other properties of the stimulus. Horizontal cells provide lateral connections between the photoreceptors and play an important role in shaping the center-surround receptive fields of the bipolar cells by inhibiting the signals of photoreceptors depending on the activity of neighboring photoreceptors. Amacrine cells, of which there is a great diversity of more than 30 types, perform a



(a) Retina (b) Sketch of Cortex

Figure 2.1: a) Sketch of a piece of retina. Shown are the light sensitive rod and cone cells at the back of the retina, and the main feed-forward pathway of bipolar cells and ganglion cells. Horizontal cells provide lateral connections between photoreceptors and amacrines between bipolar cells. b) Sketch of cortex based on a drawing by Santiago Ramon y Cajal *Textura del Sistema Nervioso del Hombre y de los Vertebrados, 1904*. The cortex consists of six layers, marked 1-6. The cell bodies visible are pyramidal and granular cells. Inputs from the thalamus (LGN) go into layer 4 and layer 6 sends feedback connections back to thalamus.

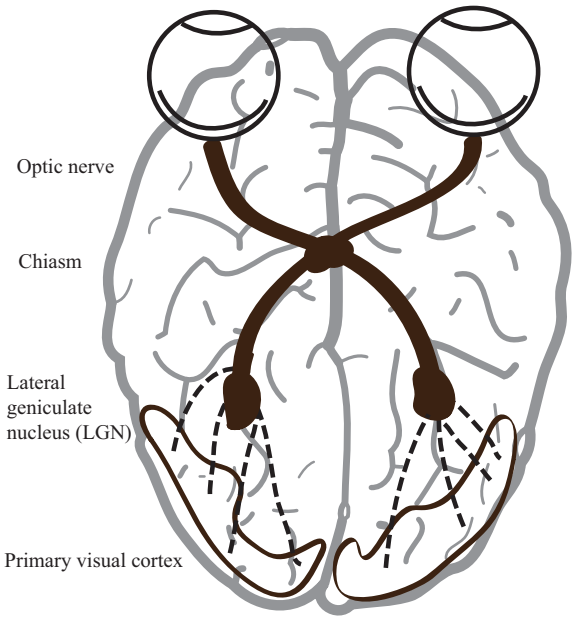


Figure 2.2: Sketch of a horizontal section of the human brain. Highlighted are the eye, optic nerve, LGN and the primary visual cortex.

similar function. They provide lateral connections between the outputs of bipolar cells, i.e. the inputs of retinal ganglion cells (RGC). Their function is not well understood, but is believed to be related to gain control and redundancy reduction [83, 87]. Finally, ganglion cells relay the information from the bipolar cells and amacrine cells to the brain. The axons of these cells form the optic nerve and project to the thalamus, hypothalamus and midbrain [83, 120]. However, of the nearly 20 kinds of ganglion cells, less than 15 actually send axons to the brain, so it is important to keep in mind that our description is strongly simplified, and much of the retinal processing is not well understood at this time.

Since there are about 100 million photo receptors in the retina of each eye, but only about 1 million ganglion cells, information cannot be relayed from the photoreceptors to the brain without further processing [29]. On average, the information from 100 receptors needs to be send down one axon in the optic nerve. The visual system must preserve as much of the information from the photoreceptors as possible, so the signal needs to be encoded in such a way that little information is lost in this compression. To a large extent, this redundancy reduction is implemented by the ON and

OFF center receptive fields, which are sensitive to local contrast and perform *spatial decorrelation*. Additionally, the dynamic range of the signal is compressed by divisive *gain control* [31]. This processing step is important not only in the retina, but throughout all the later processing stages. In the retina, gain control is mediated mainly by amacrine cells. This way the high dynamic range stimulus is compressed to fit the limited bandwidth of neurons. Gain control needs to be dynamic and adaptive on several temporal and spatial scales, so it makes up a large fraction of the processing done on the retinal level.

2.1.2 The Lateral Geniculate Nucleus

About 90% of the axons in the optic nerve project to the *lateral geniculate nucleus* (LGN) which is a structure in the thalamus in the midbrain. As sketched in Fig. 2.2, the LGN relays information to the primary visual cortex, as well as receiving feedback connections from the cortex. The neurons in LGN have center-surround receptive fields similar to those of bipolar cells in the retina.

There are two main types of cells in LGN, with are arranged in two parvocellular and four magnocellular layers. Out of the six layers, three each receive inputs from the *ipsi-* and *contralateral* eye. The LGN in each hemisphere of the brain "sees" only the contralateral half of the visual field, which is organized in a retinotopic way, i.e. the layers in LGN preserve the topographic structure from the retina. The parvo- and magnocellular layers operate on different timescales, with the former operating on a slow timescale but processing details like color information from cone photoreceptors. The latter operates much quicker, but does not process as much detail [29]. Finally, konicellular cells between the layers provide a third pathway which is not well understood at this time.

Not much is known about the functional role of the LGN in visual processing, but there is evidence of processing linked to temporal decorrelation [20] and attentional modulation [85].

2.1.3 The Cortex

Projections from the LGN, called the *optic radiations*, finally carry the visual signals to layer 4 of the primary visual cortex (area V1), the largest and best studied of the visual areas in the brain. Fig 2.1 b) shows the structure of the cortex and its organization into layers. The vast size of the primary visual cortex, as much as 15% of the total cortical area [115], suggests that it is the site of some very complex processing. The number of

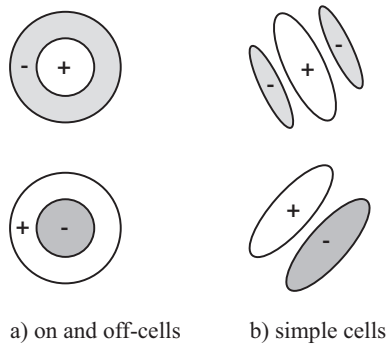


Figure 2.3: a) Receptive fields of retinal ganglion and LGN cells. An ON-center and an OFF-center cell is shown. Shaded in white are facilitatory regions that respond to increased brightness, in gray inhibitory regions that respond to a decrease in brightness.
b) Simple cells in primary visual cortex. Two cells with different orientation selectivity are shown. The preferred stimulus of the top cell is a bright bar on a dark background, the bottom cell prefers a dark to bright edge.

cells in the visual cortex is orders of magnitude larger than in LGN, so most of the synapses in V1 are recurrent connections or feedback connections from higher areas. Similar to LGN, the organization of the cells in V1 takes the form of a retinotopic map, where the visual space is mapped from the retina to the surface of the cortex.

V1 is responsible for processing much of the local structure in the visual input and has cells tuned to location, orientation, color, motion, disparity and various other properties of the input. For the sake of simplicity, we will focus mainly on the spatial receptive field properties, especially orientation selectivity, and ignore most other tuning properties. For a discussion of tuning for binocular disparity and color, as well as spatiotemporal receptive fields which are tuned to motion with a particular speed and direction, the reader is referred to the literature, e.g. [43].

2.1.4 Simple and Complex Cells

In their seminal study in the 1950s, David Hubel and Thorsten Wiesel [44, 45] systematically analyzed the receptive field properties in cat primary visual cortex, work for which they were awarded the Nobel Prize in 1981. In their experiments, they presented stimuli in the form of dark and bright bars to the animals and recorded the activity of cells in V1. They discovered

that many of the cells had a preferred stimulus orientation. In contrast to the center-surround units in the retina and LGN, they fired strongly in response to bars oriented at a particular angle. The cells could be divided into two main classes, which they termed *simple cells* and *complex cells*. The difference between the two classes is that simple cells only react to stimuli of a particular polarity, e.g. to a bright to dark edge, but not the reversed dark to bright edge. Complex cells on the other hand fire irrespective of stimulus polarity. In later studies, the exact shape of the receptive fields was mapped using the technique *reverse correlation* [98]. By presenting a white noise stimulus, and averaging over all the stimuli that preceded a spike by a certain interval (e.g. 100ms), the linear "prototype" stimulus could be obtained that maximally stimulates the cell. Using this technique, the *Gabor-like* shape of simple cell receptive fields, illustrated in Fig 2.3. b) was found. A Gabor function is the product of a sinusoidal grating with a Gaussian envelope.

While a large fraction of the cells in V1 is relatively well described as one of these types of cells, it is important to note that this is a very basic description and ignores much of the subtlety in the neural responses. The most glaring omission is the ubiquitous gain control that we already mentioned in the context of retinal cells. The responses of individual cells are modulated by the level of activity of neighboring cells at different temporal and spatial scales. Other, more complicated nonlinear properties are under active research, such as the effects of contextual modulation [46]. By presenting specific additional stimuli outside of the *classical receptive field* as it was defined by Hubel and Wiesel, the response can be strongly modulated, even though the cell would not fire in response to the extra stimulus alone. Another important property of V1 that is not well understood at this time is *attentional modulation*. Attention is a non-local phenomenon that is notoriously hard to study with electrophysiology, which is how most of the research discussed above has been carried out. More recent studies using *functional Magnetic Resonance Imaging* (fMRI) are beginning to shed more light on this [38].

2.1.5 Higher Visual Areas

Beyond V1 there is a large number of cortical areas involved in higher order visual processing, but for the most part very little is known about the processing that takes place in these areas. We will therefore discuss only a small subset of these areas, where experimental evidence exists that elucidates some of the function. V2, which is almost as large and immediately next to V1 shares most of the receptive field properties, and also forms

a retinotopic map. It has slightly larger receptive fields and responds to some more abstract properties of the visual stimulus such as distinguishing between figure and ground by coding for border ownership [96]. It has been suggested that the visual processing splits into two streams after this initial processing, with the *dorsal stream* performing processing related to the position of objects, and the *ventral stream* being responsible for object representation and recognition[114]. This is controversial however, and more recent studies show that this clear distinction cannot be made [32].

An important area in the dorsal stream that has been subject to much study is V5 or *mediotemporal cortex* (MT), which plays an important role in motion perception [84]. Similarly in the ventral stream, the *inferotemporal cortex* (IT) has received much attention. It contains cells that are highly invariant to location and orientation of an object, so it has been suggested that IT plays an important role in object recognition [112].

2.1.6 Hierarchical Processing in the Cortex

At the end of the cortical hierarchy, which consists of as many as 40 different areas which have been identified, are very specific areas such as the *fusiform face area* which have highly tuned properties such as responding specifically to human faces [105]. Little is known about what computations are performed in the brain to obtain these receptive fields, which are both extremely specific (neurons have been identified that are selective to a particular person) and at the same time highly invariant to distractors like lighting conditions and viewing angle. There is strong evidence that these invariances are gradually build up over a hierarchy of many layers. Conceivably, each of these layers performs only a relatively simple transform on its inputs (such as building the orientation-selective V1 responses by pooling circular-symmetric LGN inputs), and the complexity of the whole system emerges as all these simple transformations add together. Hierarchical processing is a powerful approach that we will further investigate in the rest of this thesis.

2.2 Modeling of Vision

We have now seen some of the properties of neurons in different parts of the visual system. But even if every single neuron in the visual system was characterized under every possible stimulus condition, this would leave us far from understanding the visual system. Being able to look up the correct response for a particular stimulus does not mean that we understand how this response is generated. Furthermore, since the number of possible

stimuli is infinite for all practical purposes, such an approach would not only be highly unsatisfactory, but also impossible.

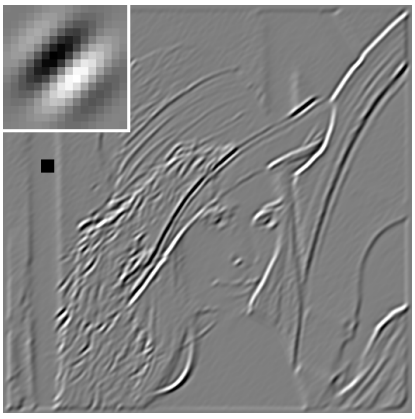
Rather, we would like to understand what kind of features the visual system is extracting, and how it is processing its inputs to arrive at an invariant high-level representation. This requires us to identify the processing steps and put them into the language of mathematics. Models of the visual system can be made at different levels of abstraction, ranging from a detailed physical model of individual synapses, over models of networks of spiking neurons to high-level models that use firing-rate codes or do away with the neuron as a unit of computation altogether. We will here be concerned with the latter kind of models, which are focusing on the underlying computations without paying attention to how these computations may be implemented in the hardware, or rather "wetware", of the brain. This does not only have the advantage of conceptual simplicity, but is also a necessity to make the estimation of the model parameters possible. In a biologically realistic model, the number of parameters would be so great, that given today's computational resources, estimating all the parameters would be utterly impossible, necessitating the selection of model parameters by hand.

2.2.1 Spatial Receptive Fields

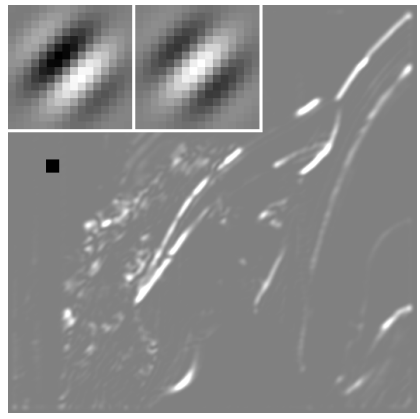
Retinal and LGN cells are only a few synapses away from the photoreceptors, so it is comparatively easy to model their function. In fact, within limits, these cells can be modeled as having a *linear* response, i.e. the firing rate can be computed as a linear function of the stimulus. A linear function that maps a small region of an image to a scalar response is also known as a *linear filter*. Rather than having to specify the filter at many different locations, taking the convolution of the filter with the image immediately gives the response at all locations, and the matrix of filter responses can be displayed as another image. Performing this operation with a set of linear filters is a common first processing stage in many computer vision applications. A similar operation can be thought to be taking place in the brain, where we can imagine the convolution being replaced by a dense tiling of cells with overlapping receptive fields. Fig. 2.4 a) and b) shows an image and the response to a center-surround filter, which is modeled after a bipolar cell in the retina. The filter is designed as a *difference of Gaussians*, giving it a circular symmetric shape. A functional interpretation of this filter would be contrast detection, removing local gray-value information and giving a non-zero response only to sudden changes in gray-value. This is not only useful in object recognition, were we are interested in detecting

a) Original image

b) Center-surround filter

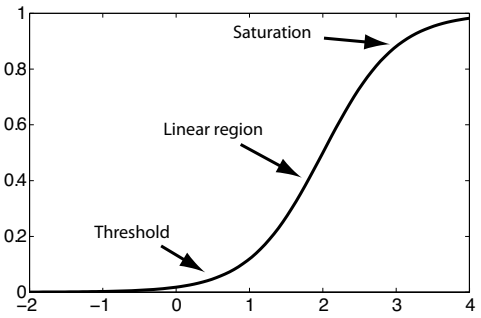


c) Oriented Simple Cell filter

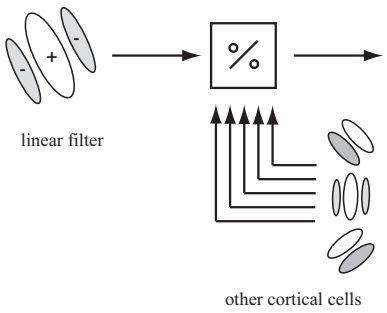


d) Phase-invariant Complex Cell filter

Figure 2.4: a) A natural image, and the response of different filters: b) A center-surround filter, which has a response like a bipolar cell, performs contrast coding. If the image is uniform within the receptive field, the response of the filter is zero. c) A Gabor filter, modeled after a simple cell, detects edges with a particular orientation. d) The response of a phase-invariant complex cell changes more slowly, and the polarity of an edge does not affect the response.
The insert in the upper left shows the filter the image was convolved with, the smaller black insert shows the actual size of the filter.



(a) Linear-Nonlinear model



(b) Normalization model

Figure 2.5: a) In the *linear-nonlinear model*, the scalar outputs are passed through a nonlinear function like the sigmoid shown here. It performs half-wave rectification on the inputs, and saturates at very high input levels. b) The *normalization model* for gain control in cortical neurons. The output of the linear filter is modulated by dividing with a weighted sum of the activities in the neighborhood of the unit.

the contours of an object, but is also related to spatial decorrelation and efficient coding [2].

The simple cells in primary visual cortex can also have a reasonably linear response to visual stimuli, so they can be modeled in the same fashion. The spatial properties of the receptive fields, which are localized both in space and frequency, can be modeled as *Gabor functions* or *filters* [94]. A Gabor filter consists of a Gaussian envelope which is multiplied with a sinusoidal grating. In Fig. 2.4 c) the response of such as filter is shown, as well as the filter itself in the upper left corner. Due to the edge-like shape of the filter, it responds at locations where the structure in the image is oriented the same way as the filter. If an edge in the image is in phase with the filter, a strong positive response is obtained, as can be seen e.g. at the upper edge of the mirror in the upper right hand corner of the image. Likewise, if the stimulus is out of phase with the filter, such as the lower edge of the mirror, the response is negative.

2.2.2 Gain Control and Divisive Normalization

Now we begin to see the limitations of the linear model: it produces negative as well as positive responses, whereas firing rates of neurons can only be positive. While an inhibitory stimulus can in fact depress the activity of a neuron below the spontaneous baseline firing rate, it is usually assumed that information is carried by an increase in firing rate, which is a non-

negative signal. Another limitation of the linear model is that it predicts an arbitrarily large firing rate in response to an arbitrarily large stimulus. In fact the visual system has to deal with an enormous range of signal intensities in the environment, which needs to be encoded with the limited dynamic range of neurons. To a certain extent, both of these problems can be alleviated in a simple way by applying a scalar nonlinearity to the outputs of the linear transformation [47, 12, 36, 103]. This is called the *linear-nonlinear* (LN) model. A suitable nonlinearity is sketched in Fig. 2.5 a). It is nearly zero for all negative inputs, so it performs rectification, and it levels off at very high input values. Between the two extremes is a region where the response is linear.

Another ubiquitous nonlinear effect that is not captured by this model, but can be found throughout the visual system is *gain control*. It is possible to model this using *divisive normalization*, which normalizes the activity of a unit by the average activity of the units around it [37, 104]. This model is illustrated in Fig. 2.5 b). Intuitively, if there is a very high contrast stimulus, many units will be active, driving down the sensitivity of individual units. Conversely, in low contrast conditions, the normalization term will be small and the sensitivity of the units will be boosted. This response can be written as

$$r_{\text{out}} = \frac{r_{\text{in}}}{\sqrt{\sum_i r_i^2}} \tag{2.1}$$

where the output rate r_{out} is computed by dividing by the rectified activity of the i neighboring cells. There are some nonlinear effects other than gain control that can be modeled in this way. For example, the response to a weak Gabor stimulus can be increased by surrounding it with flankers of the same orientation [93], so the nonlinear lateral interactions need not always be suppressive.

2.2.3 Models for Complex Cells

Even with the "trick" of using a nonlinearity after the linear filtering, the models we have considered so far are constrained to situations where the system behaves linearly over a certain range. But as we have seen in the previous chapter, even in primary visual cortex there exist cells that have a highly nonlinear response. Complex cells share the orientation selectivity of simple cells, but are completely invariant to the spatial phase of a stimulus. This kind of response can be modeled by taking the sum of squared simple cell outputs, which has come to be referred to as the *energy model* of complex cells [1, 111] and is illustrated in Fig. 2.6. While there is some evidence from physiology that this model may not reflect the actual processing

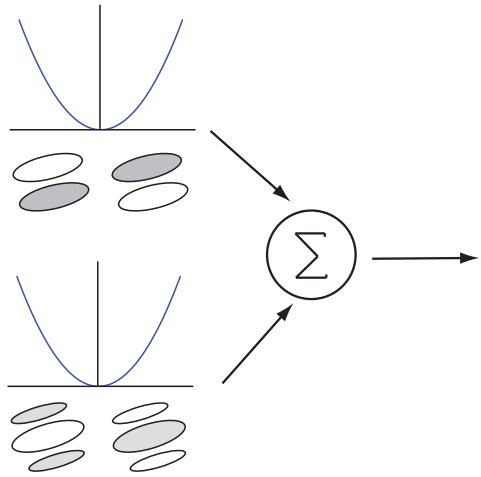


Figure 2.6: The *energy model* for complex cells. The outputs of two simple cells in quadrature are rectified by squaring, where negative responses of the simple cells can be taken as the response from additional cells with opposite polarity receptive fields. The complex cell output is obtained by summing up the squared responses.

in the visual cortex, it provides a very good description of the response of complex cells. In the energy model, the output of the cells is given by

$$r_{\text{out}} = \sqrt{(\mathbf{w}^{+\mathrm{T}}\mathbf{x})^2 + (\mathbf{w}^{-\mathrm{T}}\mathbf{x})^2} \tag{2.2}$$

where $^{\mathrm{T}}$ denotes transpose, $\mathbf{w}^{+}$ and $\mathbf{w}^{-}$ are two Gabor filters that are 90 degrees out of phase, and $\mathbf{x}$ is the visual stimulus. This mechanism is illustrated in In Fig. 2.6, and the result of this processing is shown in Fig. 2.4 d). It can be seen that the response does not depend on the polarity of the edge, and is slightly more "fuzzy" than that of the simple cell.

Notably, this model is not without criticism, since there exists a continuum of cells ranging from prototypical simple to complex cells rather than two disjoint classes. It has been suggested that the observed bimodality in the distribution of outputs is an artifact of cortical amplification, and does not reflect the properties of the underlying population [86, 102].

2.2.4 Theories for Higher Level Processing

Since our focus here is on early vision, we will only briefly mention two models for higher level processing here. Taking an interesting direction from the

simple and complex cells models we have considered so far, the *Neocognitron* by Kunihiko Fukushima consists of a hierarchy with alternating layers of simple and complex cells [27]. While this is a very speculative theory of the architecture of the visual cortex, the model has shown some impressive results in computer vision applications, e.g. in handwritten digit recognition [28]. By using layers with increasing receptive field size, the invariance properties of the complex cell units build up more and more invariance towards shifts in scale, orientation and position. This demonstrates that even relatively simple principles such as those described in this chapter can lead to powerful computations if they are performed in a hierarchical fashion. These ideas have been refined in various ways and successfully used in a variety of object recognition tasks in complex environments [106, 97].

A related approach to object recognition is the use of *convolutional neural networks*, which build up invariant representations through a hierarchy of feature maps, where the feature maps of the previous layers are convolved with a kernel. Again this method is only loosely related to the processing in biological visual systems, so it is hard to say how much, if anything, can be learned from models like this. They are certainly useful in their own right, though, and have been used successfully for handwritten digit recognition [71], object recognition [72] and navigation of autonomous vehicles in natural environments [34].

3

Linking Vision to Natural Image Statistics

Love looks not with the eyes, but with the mind
William Shakespeare

3.1 Natural Image Statistics

In this chapter we will discuss how the processing in the visual system is related to the structure in natural images, and how this structure can be exploited to build visual systems. We follow the assumption that knowledge about the regularities in natural images can help us to determine what the *optimal* way of processing in a visual system is. By matching the processing to the statistical structure of the stimulus, we can optimize the system to make inferences about the stimulus in the presence of noise or with otherwise incomplete information.

This is by no means a novel idea and dates back to the end of the 19^{th} century with ideas from Ernst Mach [81] and Hermann von Helmholtz [119], who proposed that vision was the process of *unconscious inference*, complementing the incomplete information from the eyes with assumptions based on prior experience, to make conclusions about the environment. After the introduction of information theory by Claude Shannon in the 1950's [107], the importance of redundancy reduction in neural coding was proposed as another reason why sensory systems should be adapted to the statistics of their environment. The implications of efficient coding on neural processing were investigated in the context of neural coding by Horace Barlow [7] and in relation to perceptual psychology by Fred Attneave [4].

Thus the systematic study of the statistical structure of natural images started more than 50 years ago, but only with the proliferation of powerful and inexpensive computers in the 1980's the implications for the visual system could be explored in more detail [70, 101, 3]. Initially, efficient coding provided one of the driving forces for understanding the processing, but even when it became clear that most computations are easier to perform in highly overcomplete and redundant representations [6], the study of the visual system in relation to its environment has produced a multitude of fascinating results. In the rest of this chapter we will provide an account of the most important results in the study of natural image statistics, and how neural processing is adapted to the statistical properties of ecologically valid stimuli. For completeness it should be mentioned that processing based on the statistical structure is useful not only for biological vision, but equally for machine vision and image processing applications. Although we will not consider it in more detail in this work, models based on natural image statistics have been successfully used for denoising [109, 95] and in other machine vision applications.

In order to formalize these ideas, let us start by defining what a *natural image* means in the context of this work. We consider photographic images that have been digitized in some form so we have a matrix containing

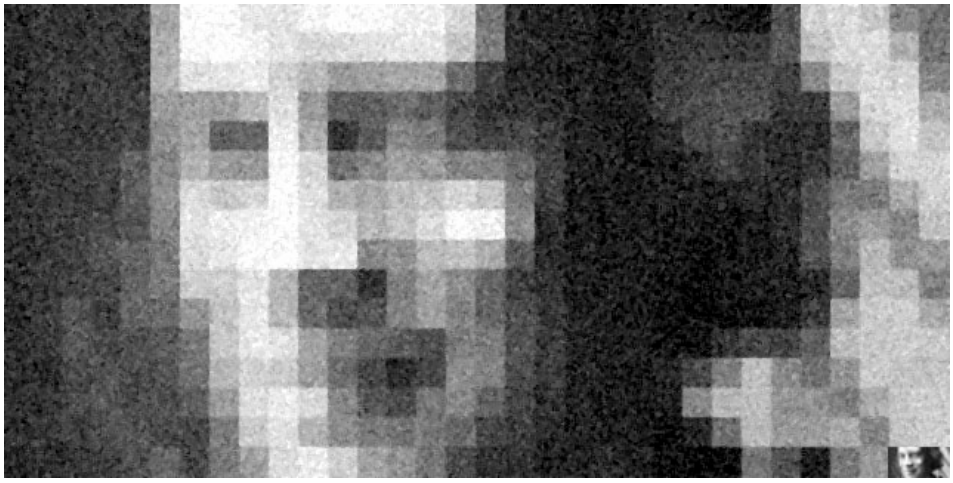
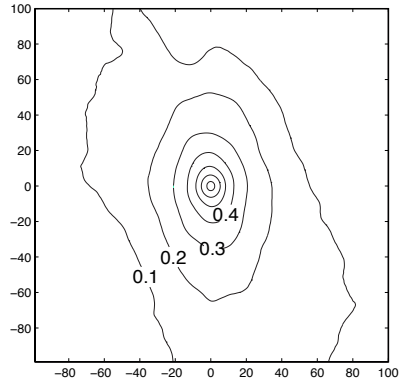


Figure 3.1: Example of a 16×32 pixel image. By squinting or otherwise blurring the image, it becomes possible to recognize that it depicts a human face, and those familiar with him may recognize Aapo Hyvärinen. Note that the two pixels at the bottom right contain the whole image displayed at ordinary scale.

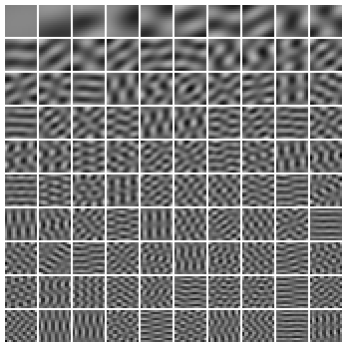
luminance values as a function of spatial location $I(x, y)$. An immediate problem is that typical images are extremely high-dimensional. If we consider the space of 256×256 pixel images quantized to 256 gray levels, there is a space of $2^{8\times256\times256} \approx 10^{150,000}$ possible images. Each of these images would be represented by a $256 \times 256 = 65,536$-dimensional vector, and even if enough images could be obtained to give a fair sample of typical natural images, the task of storing them alone would pose a serious memory problem for a typical workstation computer.

Therefore we need to restrict ourselves to small *images patches*, typically around 12×12 to 32×32 pixels. This reduces the computational load sufficiently for a statistical analysis, but still retains enough information for human observers to extract useful features, as illustrated in Fig. 3.1. As a further simplification we consider only gray scale images. Writing these matrices of gray-values as a long vector, we obtain the data vector $\mathbf{x}$, which we consider to be a realization of a random process. To infer the properties of the *probability density function* $p(\mathbf{x})$ that these data vectors are samples of, we need to consider large samples of image patches, which we will write as the columns of the matrix X.

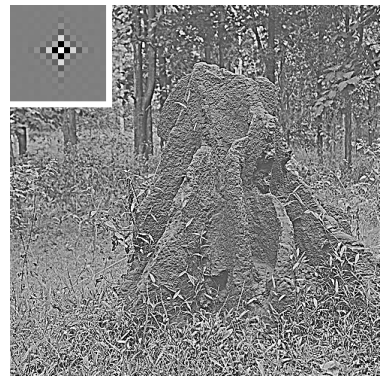
(a) A natural image



(b) Correlational Structure



(c) Principal Components of image patches



(d) Whitened image and whitening filter

Figure 3.2: Gaussian structure in natural images: a) A typical natural image. b) The correlations between pairs of pixels at a range of distances. c) Sampling 16×16 pixel patches from the image and performing an eigenvalue decomposition on the covariance matrix gives the principal components of the image patches. Only the first 100 eigenvectors are shown. d) Using the whitening filter (insert) obtained by PCA and convolving it with the image, the pixels can be approximately decorrelated.

3.2 Gaussian Structure and Whitening

As any statistician would agree, the first analysis to attempt on some data with unknown structure would be to fit a Gaussian model. A Gaussian distribution can be described solely in terms of its mean and covariance matrix, so this amounts to analyzing the covariance structure (the mean is not very informative, so it is usually removed in preprocessing). Since neighboring pixels often have very similar values, it does not come as a surprise that natural images contain strong correlations, which we will now look at in some detail.

To do this, let us consider a typical photographic image of a natural scene such as the one shown in Fig. 3.2 a). The simplest way to quantify the redundancy in this image is to compute pairwise correlations between pixels, as shown in b): for a large sample of randomly chosen pixels in the image we compute the correlation coefficient with surrounding pixels up to 100 pixels distance in the x and y-directions. It can be seen that there is a high correlation even at relatively large distances. The correlation is not uniform, since the image itself is not isotropic. The strong correlations introduce considerable redundancy in the image. This is intuitively clear; given the gray-value of one pixel we would be able to do a good job guessing what the neighboring pixels would be.

This short exposition has shown that the pixels in natural images are highly correlated, so we may try to model the correlations and ultimately remove them. This is straightforward by performing *principal component analysis* (PCA) on a sample of image patches. The principal component vectors can then be used to transform the image pixels to a set of uncorrelated variables. The principal components are displayed in 3.2 c) and take an appearance similar to a *discrete cosine transform* basis. The components are ordered by their contributed variance, so it can be seen that the lowest spatial frequencies carry most of the signal energy. Since the eigenvectors and corresponding eigenvalues exactly describe the covariance structure of the image patches, we can use this knowledge to *decorrelate* or *whiten* the image patches. In mathematical terms this means that we are looking for a transform V that we can apply to image patches $\mathbf{x}$ so that the transformed patches $\mathbf{z} = V\mathbf{x}$ have uncorrelated variables. The covariance matrix of the transformed patches should therefore be identity, i.e. $\mathrm{cov}(\mathbf{z}) = I$, where I denotes the identity matrix. Considering centered data (without loss of generality) we have

$$\mathrm{cov}(V\mathbf{x}) = \mathrm{E}\{V\mathbf{x}\mathbf{x}^{\mathrm{T}}V^{\mathrm{T}}\} = V\mathrm{E}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}V^{\mathrm{T}} = V\mathrm{cov}(\mathbf{x})V^{\mathrm{T}} = I \qquad (3.1)$$

so we are looking for a matrix V that fulfills $V\mathrm{cov}(\mathbf{x})V^{\mathrm{T}} = I$. Here we make

use of the eigenvalue decomposition on the covariance and write $\mathrm{cov}(\mathbf{x}) = U\Lambda U^{\mathrm{T}}$, where U is the matrix of eigenvectors and Λ the diagonal matrix of eigenvalues, so we have $VU\Lambda U^{\mathrm{T}}V^{\mathrm{T}} = I$, which can be satisfied by setting

$$V = \Lambda^{-\frac{1}{2}}U^{\mathrm{T}} \tag{3.2}$$

as can be seen by substituting this expression back in. In geometrical terms, this whitening operation amounts to projecting the image patches on the principal components and then rescaling them with the variance along the direction of that component. It is important to note that the highest frequencies, which are strongly boosted by this processing, have a low signal-to-noise ratio and contain little information relevant to the image. Furthermore, due to the rectangular sampling grid, retaining these frequencies may give rise to filters with aliasing artifacts such as checkerboard patters. It is therefore common practice to reduce the dimensionality of the data at the same time as whitening to attenuate or remove these high frequency components. This can simply be done by projecting only on the first few principal components, discarding as much as 50% of the components which carry very little variance [55].

Since an orthogonal rotation Q does not change the now spherical covariance structure, any $V = Q\Lambda^{-\frac{1}{2}}U^{\mathrm{T}}$ is also a whitening matrix. By choosing $Q = U$, we can perform *zero phase whitening*, which means that after rescaling the variables we rotate back to the original coordinates. This is called "zero phase" because the Fourier phase of the signal is not changed, and the whitened image is closest to the original image in terms of squared distance to the original pixels. Unlike the principal components in 3.2 c), the whitening matrix obtained in this way contains identical copies of a center-surround filter at each pixel location. One such whitening filter is shown in the insert in 3.2 d). Since the effect of multiplying with this whitening matrix is a convolution with the single whitening filter, we can illustrate the effect of whitening on a whole image by "abusing" one of the vectors of the whitening matrix as a whitening filter, and convolving it with the image as is shown in 3.2 d).

The alert reader will have noticed that the whitening filter which we optimized to remove correlations between image pixels is similar in shape to the receptive fields of ON and OFF-bipolar cells in the retina we have seen in Fig. 2.3 a) and that the effect of whitening is very much like that of the center-surround filter in Fig. 2.4 b). This similarity gives strong support to the hypothesis that the coding employed by the visual neurons is utilizing spatial decorrelation to reduce redundancy in the input signal. However, this interpretation is not without criticism, and other mechanisms have

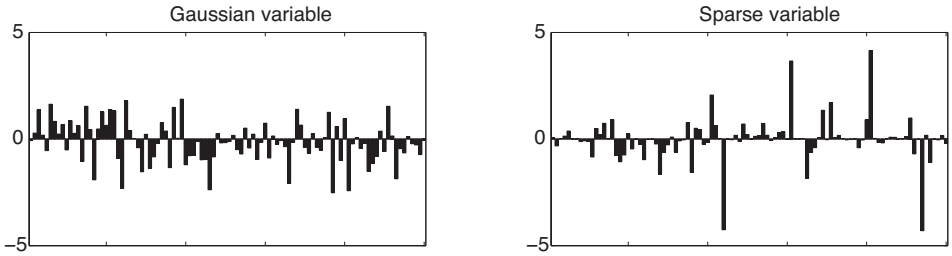


Figure 3.3: Comparison of 100 samples of a Gaussian and a sparse random variable, both with unit variance. The sparsely distributed variable occasionally takes on very large vaules, but stays close to zero most of the time.

been proposed by which the center-surround receptive fields of of bipolar cells can be explained. One alternative hypothesis is that the receptive fields are optimized to satisfy wiring length constraints [118].

By whitening we have transformed the image data to a set of variables that are uncorrelated and of unit variance, which means that we have removed all the second-order structure. Even though this makes the image look rather strange with greatly exaggerated edges, it is still possible to discern most of the content of the image. In fact, from the standpoint of a human observer, not much has been lost from the image at all, and all of the features that are relevant for the perception of objects are still there. Clearly, there is still a lot of rich statistical structure that we can attempt to model. This means that we now need to turn to the *non-Gaussian* structure of the image, which requires more advanced statistical methods. In the rest of this chapter we will look at the non-Gaussian structure in more detail, and analyze how it relates to processing in the visual cortex.

3.3 Sparse Coding and Simple Cells

While the Gaussian distribution, perhaps due to its simplicity or by reference to the central limit theorem [17], is often seen as the most natural probability distribution, it turns out that most ecological signals deviate from a Gaussian in a specific way. These signals have *supergaussian* distributions with heavy tails and a strong peak at zero. A random variable that follows a supergaussian distribution, such as in the right hand panel of Fig. 3.3, is only rarely activated, and close to zero most of the time. Therefore this class of distributions is termed *sparse*. We have already seen a natural signal that follows this kind of distribution: a whitened image like

that in 3.2 has many pixels that are nearly zero, but occasionally pixels have very high or low values. Sparseness is an important concept in neural coding and has been extensively studied [7, 24, 25, 26]. In comparison to *dense distributed codes*, where many units are active simultaneously to represent a pattern, a sparse code can represent any input pattern with just a few active units. In addition to their robustness properties in the presence of noise, sparse codes are advantageous if there is an energy cost associated with a unit being active [122]. This is especially true in the brain, where signals are transmitted by spikes. When a neuron fires a spike, its membrane potential becomes reversed, and restoring the membrane to the resting potential has a substantial metabolic cost. In fact the cost of a single spike is so high, that the fraction of neurons that can be substantially active concurrently is limited to an estimated 1% [74].

Due to the statistical properties of the stimulus, the response of a whitening filter, or retinal bipolar cell, is already quite sparse, without any particular optimization. In fact, by limiting the analysis to the covariance, we have deliberately excluded any measure of sparseness from our previous analysis. But motivated by the useful properties of sparse codes, we can explicitly maximize the sparseness of the representation, following the work of Bruno Olshausen and David Field [88, 89]. Their *sparse coding* algorithm models image patches $\mathbf{x}$ as a linear superposition of basis functions $\mathbf{a}_i$, weighted by coefficient s_i that follow a sparse distribution. Thus we have $\mathbf{x} = A\mathbf{s} + n$ where we use matrix notation for convenience, so A contains the vectors $\mathbf{a}_i$ and n is a small, additive Gaussian noise term. We are trying to find a combination of basis functions and coefficient that gives a good reconstruction $\hat{\mathbf{x}} = A\mathbf{s}$ while at the same time maximizing a measure of spareness of the activation coefficients s_i. We can formalize this as an optimization problem where we trade off reconstruction error for sparseness as

$$\min_{\mathbf{a}_i} \mathrm{E}\{||\mathbf{x} - A\mathbf{s}||^2 + \lambda \sum_i |s_i|\}. \tag{3.3}$$

Here the expectation E{} is taken over a large number of image patches. The constant λ determines the trade-off between sparseness and reconstruction error and therefore sets the noise level. We have used the Euclidean norm for the reconstruction error and use the L_1-norm as a measure of sparseness. This corresponds to a probabilistic model where we are maximizing the posterior of a Gaussian likelihood with a Laplacian sparseness prior. The exact estimation of this model would require integrating over the coefficients, which is intractable. Therefore it is estimated using a *maximum a posteriori* (MAP) approximation, leading to the following optimization:

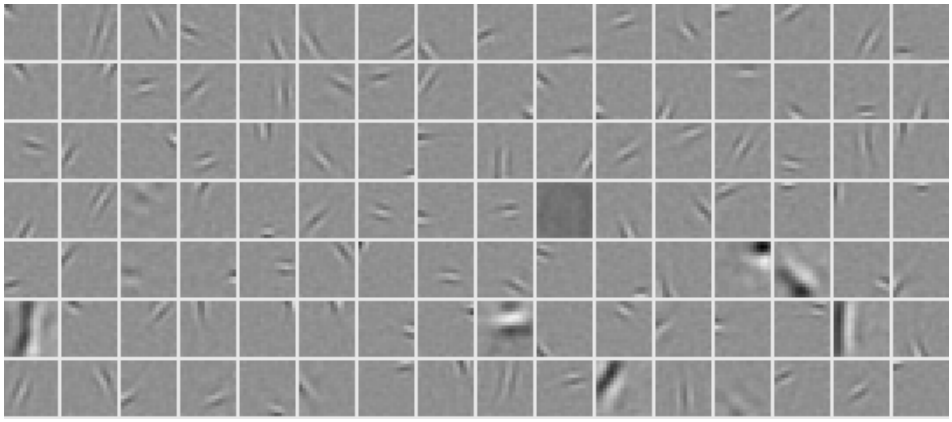


Figure 3.4: Subset of a basis for natural images obtained by sparse coding. Image patches of size 16×16 pixels were pre-processed by approximate whitening, rolling off the highest frequencies. The sparse coding algorithm was then used to estimate a two times overcomplete basis set. Note that basis functions obtained by sparse coding are localized, oriented "edge-detectors", very much like the simple cells of primary visual cortex.

starting from an initial set of units $\mathbf{a}_i$ we compute the coefficients s_i that give the lowest combined reconstruction and sparseness penalty. Keeping these s_i fixed, we then compute the set of basis functions $\mathbf{a}_i$ that improve the reconstruction the most. Alternating between these two steps, we can find the dictionary of basis functions $\mathbf{a}_i$ that can describe the set of natural images in a maximally sparse way.

In Fig. 3.4 we show a subset of the linear filters estimated by applying the sparse coding algorithm to a collection of 10.000 image patches of size 16×16 pixels, randomly sampled from natural images such as that shown in Fig. 3.2 (a). The image patches were pre-processed by performing whitening with a center-surround filter similar to the one we derived in the previous section, but rolling off the highest frequencies to avoid aliasing artifacts.

Rather than a complete basis set, with as many basis functions as pixels, we estimated an *overcomplete* set with twice as many basis vectors. Having more basis functions has the advantage that the basis functions can be more specialized and therefore become active less frequently. This makes for a sparser code, and also provides some robustness, so if individual units become "damaged" or their activations become switched off, the underlying visual stimulus is still represented fairly accurately.

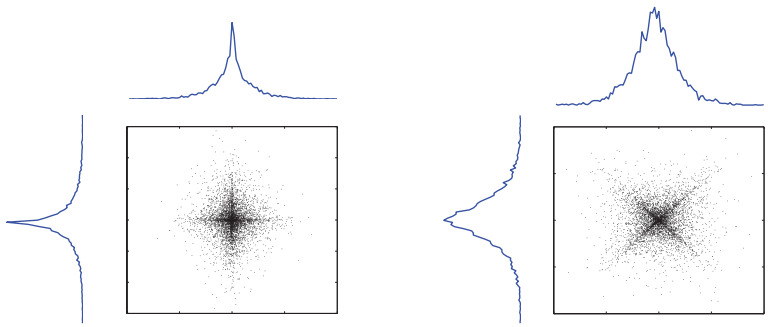
The individual basis functions that provide a dictionary to represent the possible natural image patches, have some very familiar structure: they are

localized within the image patch, are selective for a particular direction, and also cover the different scales of spatial frequencies. In fact, they look very similar to the *Gabor functions* we introduced in section 2.2.1 as a model for the spacial receptive fields of simple cells in primary visual cortex. The key properties of these receptive fields are all reflected in the basis functions learned by sparse coding. This provides some evidence that the first processing steps in the primary visual cortex, which give rise to simple cell receptive fields, are constrained by efficient coding principles and may be optimized to satisfy wiring and metabolic constraints by maximizing the sparseness of the representation. On the other hand, many of the nonlinear properties of simple cells cannot be explained in this simple framework based on a highly simplified stimulus, so caution is required in interpreting these encouraging results.

3.4 Independent Component Analysis

Sparseness is somewhat related to the concept of statistical independence, which states that two variables are independent if and only if $p(x, y) = p(x)p(y)$ i.e. the joint probability of the variables can be factorized and is equal to the product of the marginal densities. The relation may not be immediately obvious, so let us consider a simple example. In Fig. 3.5 we show a scatter plot of the joint density of two random variables. In the left hand plot the two variables are independent, whereas in the right hand plot they have been rotated (mixed), so the probability density function (pdf) can no longer be written as the product of the marginals. In accordance with the *Central Limit Theorem*, this has an interesting implication for the marginals: under certain regularity conditions, mixtures of non-Gaussian variables are always more Gaussian than the original variables. Therefore, if the original variables have a sparse, or supergaussian distribution, maximizing independence is equivalent to maximizing sparseness. In both cases we are looking for directions, or basis functions, that maximize nongaussianity.

Independent Component Analysis (ICA) is a method that attempts to recover these directions by maximizing some measure of nongaussianity. After ICA was first described [15, 60], it took only a few years until it was applied to natural images [9, 116, 55]. While ICA is very similar to sparse coding in some respects, there are a few important differences. In ICA we usually consider a complete model, which has as many basis functions as pixels, so the matrix of basis functions is invertible. Furthermore we restrict the analysis to the noiseless case, i.e. there is no gaussian reconstruction



a) sources aligned with the axes

b) mixtures of the supergaussian sources

Figure 3.5: Illustration of the connection between sparseness and independence. On the left hand side, two independent variables are shown with the histograms of the marginal distributions. On the right hand side, the two variables have been mixed together, so they now have dependencies. In this simple case, the dependency can directly be read off the scatterplot: if one variable takes on a high value, the other has a high chance of also having a strong negative or positive activation. The key point here is that the marginal distributions have changed and have become less sparse, or more Gaussian.

error and the first term in Eq. 3.3 vanishes [75].

Before we discuss the application of ICA to natural images, let us quickly review the basic ICA model and one possible way to estimate it. We will focus on the likelihood-based approach for estimating the ICA model, since it forms the basis for much of the work described later in this thesis. To define ICA as a probabilistic model, we write the data as a mixture of sources $\mathbf{x} = A\mathbf{s}$ and define the distribution of the data in terms of the densities of the independent sources

$$p_x(\mathbf{x}) = |\det W| p_s(W\mathbf{x}) = |\det W| \prod_i p_i(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}) \tag{3.4}$$

where we assume that the mixing is invertible and $W = A^{-1}$ is the inverse of the mixing matrix, so the $\mathbf{w}_i^{\mathrm{T}}$ are the rows of the inverse mixing, or *filter matrix* W. We denote the pdf of the mixtures by p_x and that of the sources by p_s. The p_i denote the marginal distributions of the individual sources. For the first equality, we have used a well-known result for the density of a linear transform, and for the second equality the independence of the the

sources. Given T samples of the data vector, denoted by $\mathbf{x}(t)$, we can now write the log-likelihood of the $\mathbf{w}_i$ as

$$\log p_x(\mathbf{x}|W) = \sum_{t=1}^{T} \sum_i \log p_i(\mathbf{w}_i^{\mathrm{T}} \mathbf{x}(t)) + T \log |\det W|. \tag{3.5}$$

In principle, estimating the model would require estimating not only the matrix W but also the densities of the sources, p_i. This is a nonparametric estimation problem, and estimating the true densities, which tend to be strongly peaked at zero, can lead to problems with gradient optimization methods. Thus it is preferable to use smooth proxy distributions for the estimation of the filters. It turns out that if we know that all of the sources are supergaussian, we can plug in *any* supergaussian pdf for the sources and still get a consistent estimate of the filters $\mathbf{w}_i$ [56]. In cases where we are interested in the densities of the sources in addition to the filters, or in nonlinear models where the above does not hold, we can use a simple family of densities, such as the generalized normal distribution, to infer the shape of the marginals.

A supergaussian pdf for the estimation that naturally comes to mind for its simplicity is the Laplacian distribution, which we have already seen as the sparseness prior in the sparse coding model. Normalized for zero mean and unit variance, it is given by

$$\log p_i(s_i) = -\sqrt{2}|s_i| - \frac{1}{2} \log 2. \tag{3.6}$$

However, the derivative of this density has a discontinuity at zero, so it is convenient to replace it by a smooth version, the *logistic distribution* given by

$$\log p_i(s_i) = -2 \log \cosh \left(\frac{\pi}{\sqrt{12}} s_i \right) - \log 4. \tag{3.7}$$

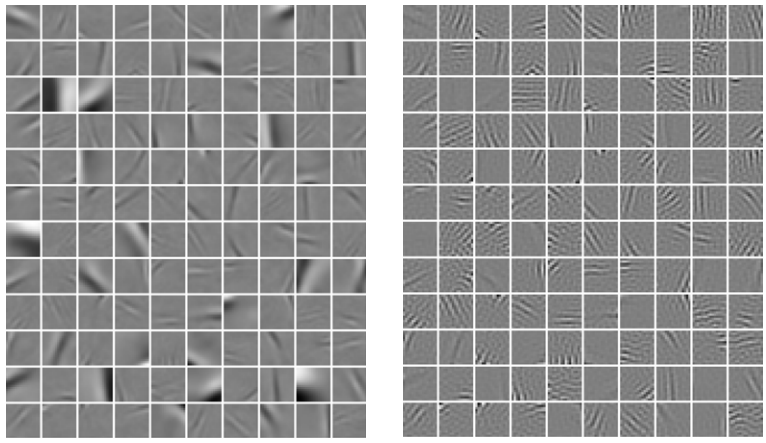
For convenience the various normalization factors are usually omitted, and the density that is used is just $\log p_i(s_i) = -2 \log \cosh s_i$.

The ICA model is estimated by taking the gradient of the log-likelihood w.r.t. the filters $\mathbf{w}$. Substituting the derivative of the marginal distributions, $\frac{\partial}{\partial u} \log \cosh(u) = \tanh(u)$ we obtain the gradient as

$$\frac{\partial}{\partial W} \log p_x(\mathbf{x}|W) = -\sum_{t=1}^{T} \tanh(W\mathbf{x}(t))\mathbf{x}(t)^{\mathrm{T}} + TW^{-T} \tag{3.8}$$

where we have used the identity $\frac{\partial}{\partial W} \log |\det W| = W^{-T}$. Now the maximum likelihood estimation proceeds by taking gradient steps like

$$W \leftarrow W + \mu \frac{\partial}{\partial W} \log p_x(\mathbf{x}|W) \tag{3.9}$$



a) ICA basis functions b) ICA filters

Figure 3.6: Basis functions and filters estimated using the FastICA algorithm on 16×16 pixel natural image patches. The dimensionality of the data has been reduced to 120 dimensions by PCA. For white data the filters are equal to the basis functions, i.e. $A = W^{\mathrm{T}}$, but projecting back to the original, non-white space, the whitening matrix is absorbed in the filters, so the high frequencies are emphasized, whereas the basis functions have the same power spectrum as the image patches. Note that the appearance of the sparse coding basis is in between ICA filters and basis functions since there the whitening is not part of the model.

where the step size is given by μ, a small constant. This has come to be known as the Bell-Sejnowski algorithm [8]. There are various ways to make the estimation of this model computationally more efficient, for example using a modified gradient update rule [13] or with a fixed-point algorithm like FastICA [48]. We will not go into the details here, since most of the work of this thesis is based on the gradient algorithm in the simple form presented above.

Applied to natural image data, ICA produces basis functions (i.e. the columns of the matrix A) that are very similar to the sparse coding basis functions or simple cells of primarily visual cortex. The similarity should not be surprising however, because ICA is a special case of sparse coding [89]. The advantage of ICA here is that estimating the model is much easier and faster, since the sources or independent components can be computed in closed form and do not have to be estimated by gradient descent. The filters that give the independent components can be plotted in the same

way as basis functions as shown in Fig. 3.6.

While it is not possible to draw a clear distinction between independence and sparseness in the linear models we have considered so far, this is not the case in general. In the neural processing hierarchy, there is no reason why higher levels of abstraction should have increasingly sparse encodings, and in fact there is no evidence that this is the case [5]. On the other hand, it is conceivable that a continued maximization of independence to encode for different objects, persons, etc. may be useful even in very high-level representation, so it is important not to confuse sparseness with independence [80, 110].

3.5 Score Matching

In this chapter, we will describe a novel estimation principle that provides an alternative way to learn the filters in ICA and related models. Score matching has been proposed in 2005 and provides a mechanism for learning in energy based models, where the pdf can be computed only up to a multiplicative normalization constant [49, 50, 51]. It has been used in the two-layer model in *Publications 3 and 4* [66, 67] and in the Markov Random Field, *Publication 5* [69] of this thesis.

We will start by describing how the score matching optimization proceeds, and then give some intuition for the method by providing a simple example. Consider a distribution defined up to proportionality by the exponential of a non-negative energy function

$$p(\mathbf{x}|\theta) \propto \exp(-E(\mathbf{x},\theta)) \tag{3.10}$$

so the normalized distribution is given by

$$p(\mathbf{x}|\theta) = \frac{\exp(-E(\mathbf{x},\theta))}{\int \exp(-E(\mathbf{x},\theta))d\mathbf{x}} = \frac{1}{Z}\exp(-E(\mathbf{x},\theta)) \tag{3.11}$$

where the integral is taken over all space. The score function, which we here take to be the derivative w.r.t. the elements of $\mathbf{x}$ is given by

$$\psi_i(\mathbf{x},\theta) = -\frac{\partial}{\partial x_i}E(\mathbf{x},\theta). \tag{3.12}$$

We can now define a new objective function that measures the squared distance of the score functions of the model, denoted by $\psi(\mathbf{x},\theta)$ and of the data $\psi_{\mathbf{x}}(\mathbf{x})$ as

$$J = \frac{1}{2}\int p_{\mathbf{x}}(\mathbf{x})||\psi(\mathbf{x},\theta) - \psi_{\mathbf{x}}(\mathbf{x})||^2 d\mathbf{x} \tag{3.13}$$

and we seek the parameter vector θ that minimizes this distance. This stills seems like a difficult problem, since there is no easy way of estimating the data score function $\psi_{\mathbf{x}}(\mathbf{x})$. However, it can be shown whence expansion of the terms and partial integration that minimizing the score matching objective function reduces to

$$J(\mathbf{x}, \theta) = \frac{1}{T} \sum_{i,t} \frac{1}{2} \psi_i(\mathbf{x}(t), \theta)^2 + \frac{\partial}{\partial x_i} \psi_i(\mathbf{x}(t), \theta) + C \tag{3.14}$$

where we have additionally replaced the expectation by a sample average over T observations, and the constant C does not depend on the parameters. Now supposing that the data follows the model, i.e. there exists a θ^* such that $p_{\mathbf{x}}(\mathbf{x}) = p(\mathbf{x}, \theta^*)$, then under some weak regularity conditions minimizing J gives a consistent estimate of the parameter vector. To show this, consider the case $J(\theta) = 0$ for some θ. Now the non-negativity of the energy implies that $p_{\mathbf{x}}(\mathbf{x}) > 0$ for all $\mathbf{x}$ from which it follows that the score functions $\psi(\mathbf{x}, \theta)$ and $\psi_{\mathbf{x}}(\mathbf{x})$ are equal. This implies that the probabilities are related as $p_{\mathbf{x}}(\mathbf{x}) = cp(\mathbf{x}, \theta)$ but the constant c is necessarily unity since both pdfs have to integrate to zero. From this it follows that $\theta = \theta^*$, showing that the global minimum of the score matching objective corresponds to the true solution.

3.5.1 A Simple Example

Consider the simple problem of fitting the mean and variance of a univariate gaussian to observed data samples $x(t)$. We have the log-likelihood

$$\log p(x(t)|\mu, \sigma^2) = -\frac{1}{2\sigma^2}(x(t) - \mu)^2 - \log Z \tag{3.15}$$

where the partition function Z is treated as unknown. The score function of the model is

$$\psi = \frac{\partial}{\partial x} \log p = -\frac{1}{\sigma^2}(x(t) - \mu) \tag{3.16}$$

and the derivative of the score function

$$\psi' = \frac{\partial^2}{\partial x^2} \log p = -\frac{1}{\sigma^2} \tag{3.17}$$

so the sample version of the score matching objective for T observations is

$$J = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \left(\sigma^{-2}(x(t) - \mu)\right)^2 - \sigma^{-2} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2} \sigma^{-4}(x(t) - \mu)^2 - \sigma^{-2}. \tag{3.18}$$

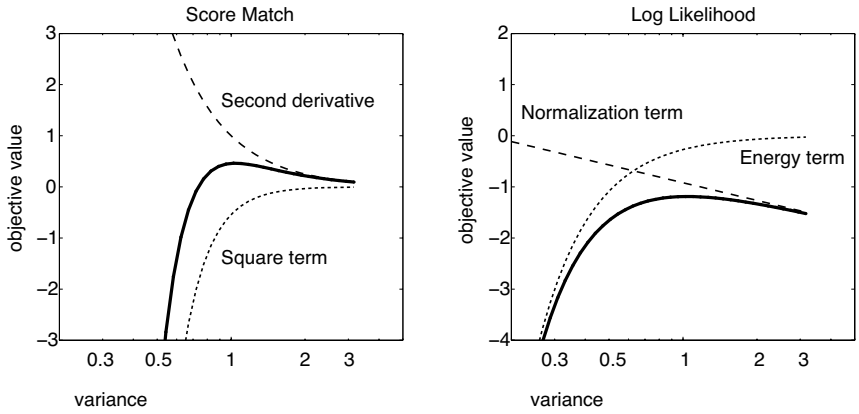


Figure 3.7: Illustration of score matching, applied to infer the variance of a univariate normal distribution. Both the score matching objective and the log-likelihood (solid curves) have the optimum at the correct position, but the functions differ significantly in shape. Comparing the two terms of the score matching objective with the energy and normalization term of the log-likelihood reveals the similarity between the second derivative term in the score matching objective and the normalization term: both penalize unspecific models with unnecessarily high variance.

To obtain the score matching estimate of the mean μ we take the derivative and set it to zero to obtain

$$\frac{\partial J}{\partial \mu} = \frac{1}{T} \sum_{t=1}^{T} \sigma^{-4} (\mathbf{x}(t) - \mu) = 0 \tag{3.19}$$

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}(t) \tag{3.20}$$

which is the sample mean, in agreement with the maximum likelihood estimate. Similarly, to fit the variance σ^2 we take the derivative

$$\frac{\partial J}{\partial \sigma} = \frac{1}{T} \sum_{t=1}^{T} \left(-2\sigma^{-5} (\mathbf{x}(t) - \mu)^2 \right) + 2\sigma^{-3} = 0 \tag{3.21}$$

$$\hat{\sigma} = \sqrt{\sum_{t=1}^{T} (\mathbf{x}(t) - \mu)^2} \tag{3.22}$$

which is the sample variance, again in accordance with maximum likelihood. In general, there is no closed form solution for the optimal parameters, so

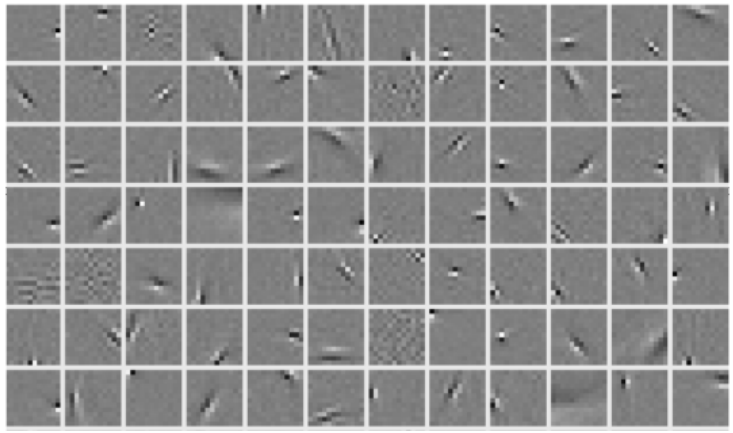


Figure 3.8: Overcomplete ICA model estimated with score matching. Only a subset of the 2304 filters from the 16 times overcomplete model are shown. Note that in contrast to the sparse coding model, we estimate an overcomplete set of filters rather than basis functions. This leads to some very different properties of the model.

gradient methods have to be used for optimization.

The relation between score matching and maximum likelihood are further illustrated in Fig. 3.7. Comparing the terms in the score matching objective with the log-likelihood allows an intuitive interpretation of score matching. The square term in the score matching objective acts similar to the energy (i.e. the non-normalized log-likelihood) and tries to make the model general enough to cover all of the data points. The second derivative has a similar effect to a normalization term, penalizing the unspecific model and forcing it to focus probability mass to where the observed data lies.

3.5.2 Overcomplete ICA Example

Score matching can easily be applied to estimate ICA models, and because the normalization constant is not required to be known, it is an obvious choice for overcomplete ICA. In this case the model is

$$-\log p(\mathbf{x}) \propto E(\mathbf{x}) = \sum_{i=1}^{M} g(\mathbf{w}_i^{\mathrm{T}}\mathbf{x}) \tag{3.23}$$

where the number of filters M is larger than the dimensionality of the data. Models of this kind, also known as *Products of Experts* (PoE) have

been studied extensively [39], where they were estimated with *Contrastive Divergence* (CD) [40], a method that shares some similarities with score matching and is based on Monte Carlo methods.

Estimating this model for natural image patches again leads to familiar Gabor-like filters, which is shown in Fig. 3.8 for a 16 times overcomplete model. Here we show the filters $\mathbf{w}$ rather than basis functions, because in the overcomplete model the filter matrix cannot be inverted, and thus no basis functions are defined.

The complete ICA model which we considered in the previous section can be seen as a bridge between overcomplete sparse coding models on one hand, and overcomplete PoE or ICA models on the other hand, which have some very important differences. In an energy-based model, the internal representation can be computed by a fast, simple feed-forward computation, whereas in the generative sparse coding model, the optimal pattern of activities $\mathbf{s}$ is the solution to an optimization problem. This implicitly nonlinear mapping between the data $\mathbf{x}$ and the components $\mathbf{s}$ has some attractive properties for neuroscience: since the basis functions inhibit each other, behavior such as end-stopping and nonclassical surround effects has been observed in overcomplete sparse coding models [73]. On the other hand, a recent study [22] has shown some advantages in energy based models over generative models in denoising applications, and it is not clear at the time how the two approaches are related and which provides a better model for natural images, and therefore ultimately for visual processing.

4

Novel Models in this Work

Wer kann was Dummes, wer was Kluges denken,
Das nicht die Vorwelt schon gedacht?
- J. W. von Goethe -

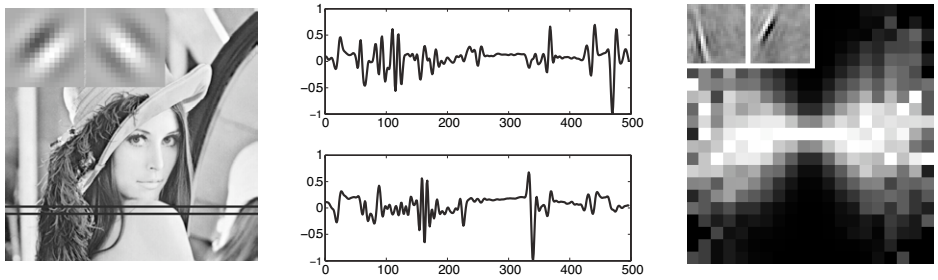
4.1 Limitations of Linear Models

Natural images are not generated as a linear superposition of sources, so the "most independent" components still have strong dependencies between them. In fact the distribution of filter outputs is approximately spherically symmetric for nearby filters, which is in stark contrast to the factorial distribution we assume in the linear model [79]. Thus, while the components are guaranteed to be uncorrelated due to whitening, there are strong variance dependencies, i.e. correlations between squares of components, which are also termed energy dependencies. If two components have energy dependencies, they will tend to be active together, but one may be highly positive and the other negative, leading to no net correlation.

Since the energy dependencies form such an important part of the statistical structure of natural images, it is worth looking in some more detail how this dependency structure is generated. Natural images are quite non-stationary, so different parts of an image can have very different statistics. One reason for this is the way images are generated from occlusions of different objects, another cause are non-uniform lighting conditions. This is illustrated in Fig. 4.1, where the response of two filters to a natural image is analyzed. Although the two filters are orthogonal in orientation, they tend to be active simultaneously in high contrast, heavily textured regions of the image. Similarly, in uniform regions of the image, both filters have low activity. This problem can be alleviated to a certain extent by performing *contrast gain control* on the images before the linear filtering. Similar to the neural processing we have seen in Sec. 2.2.2, this can be done by computing the local variance in the neighborhood of an image pixel and dividing the pixel value by it. By this operation, the image is made more uniform, so there will be less difference in the variance of filter responses to different regions of the image. As we have shown in *Publication 2* [58], this leads to a reduction in energy dependencies.

4.2 Independent Subspace Analysis

In the previous chapters, we have attempted to explain the properties of center-surround cells in the retina and LGN which perform spatial decorrelation, and we have given a statistical justification to the orientation-tuning of simple cells which maximize sparseness or independence. Let us now see if the properties of complex cells, which are invariant to the polarity of the stimulus, can be explained using similar statistical properties of natural images. Clearly, the linear models we have considered so far are not power-



a) Test image b) Filter responese along a line c) Conditional histogram of filter activies

Figure 4.1: a) Natural image filtered along a line with the two Gabor filters shown in the upper left corner.
b) Responses of the two filters: even though the responses are uncorrelated, it can be seen that both filters tend to be active in the same parts of the image.
c) Conditional histogram of two ICA filters, following [104]. The horizontal axis represents the activity of the first filter which is shown in the upper left corner. Each vertical line represents the histogram of activities of the second filter, conditional on the first filter having the activity specified by the horizontal position. Each of the columns is normalized by dividing by the value of the largest bin. It can be see that if the first filter is inactive, so is the second filter. However, if the first filter has a strong positive or negative activity, the second one is also highly active. Note that the two responses are still uncorrelated.

ful enough to capture the highly nonlinear receptive fields of complex cells. Changing the sign of the stimulus does not change the output of a complex cell, whereas in the linear models we have considered so far, the sign of the output would also be flipped. But we have already seen in Sec. 2.2.3 that a simple element-wise squaring nonlinearity combined with pooling of units can reproduce the receptive field properties of complex cells. The remaining question then is, if it is possible to learn this kind of processing from the data. Using the idea of feature subspaces [64], this is the goal of *independent subspace analysis* (ISA) [53]. Here, the components are projected onto a number of small subspaces, and independence is optimized only between, but not within individual subspaces. The distribution of the components inside one subspace is assumed to be a function of the L_2-norm of that subspace only, which is computed by summing the squares of components in the subspace. This is exactly the processing that is required to obtain complex cell responses, so if the linear filters learned with ISA take

a similar form of the quadrature pairs we saw in Sec. 2.2.3, the receptive fields would indeed be those of complex cells. The ISA model is estimated by taking the norms of projections onto subspaces

$$u_j = \sum_{i \in S_j} s_i^2 = \sum_{i \in S_j} (\mathbf{w}_i^{\mathrm{T}} \mathbf{x})^2 \tag{4.1}$$

where S_j indicates the set of components in the j^{th} subspace. The distribution of these features is analogous to the likelihood based ICA model

$$p(\mathbf{x}|W) = |\det W| \prod_j \exp\left(-\sqrt{u_j}\right), \tag{4.2}$$

where the index j runs over all subspaces. The model can be interpreted as a hierarchical two-layer neural network, where a (learned) first layer of weights $\mathbf{w}_i$ is followed by a static nonlinearity and a second layer of linear weights, which is fixed to perform a pooling on groups of inputs. The model can be estimated by ascending the gradient of the log-likelihood, but similar to basic ICA, a computationally more efficient estimation is possible using the FastISA algorithm [57] from *Publication 1*. By performing the ISA estimation with groups of more than two filters per subspace, slightly more position invariance can be gained in addition to the phase invariance, while mostly retaining the selectivity to spatial frequency and orientation. In Fig. 4.2 we show an ISA basis with a subspace size of four. We plot the nonlinear receptive fields in the same way as we have previously visualized complex cell receptive fields, by showing just the linear filters, the outputs of which are squared and pooled to obtain the invariant response. The first four basis functions (from left to right) belong to the first subspace, the next four bases to the second subspace, and so on.

4.2.1 Gain Control for ISA

The squaring nonlinearity which is at the core of ISA can be justified in a statistical sense as a way to model *energy dependencies* between the "independent components" of natural images, which have been identified [126] as an important problem for the independence assumption inherent in the linear models of Chapter 3. With ISA, we identify the pairs or groups of linear filters that have the strongest dependencies and model them with a spherically symmetric distribution, where they have a high probability of being activated together. In this framework, it is possible to interpret complex cells as being optimized to model these energy dependencies.

It turns out however that this view is overly simplistic if we compare the likelihood of ISA models with different subspace sizes [58], as we did in

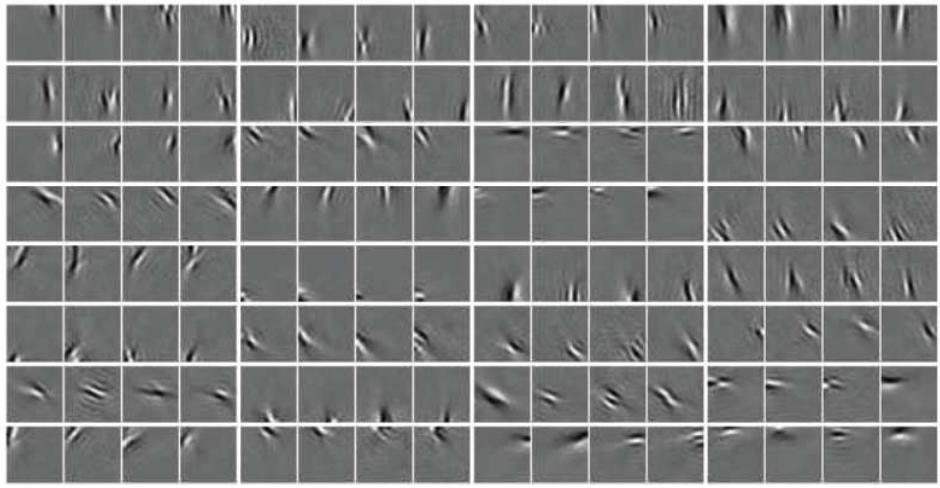


Figure 4.2: Basis estimated with ISA using four components per subspace. The linear filters in one subspace share location and direction selectivity, but differ in the local spatial phase. Thus, the pooled outputs show the typical invariance properties of complex cells.

Publication 2. Without any gain control in the preprocessing, the energy correlations between all pairs of components are so strong that surprisingly large subspaces are found to be optimal, in some cases pooling all the components into a single subspace, which amounts to fitting a spherically symmetric distribution. This suggests that the dependencies are so strong that orientation selective filters do not provide any advantage in encoding the stimulus. This implies that a spherically symmetric distribution gives the best fit to the data, a surprising effect that has been studied in more detail in [110]. Only by performing gain control, the global dependencies are reduced sufficiently for small subspaces to be optimal. This can be done in a very simple way by dividing each of the whitened image patches by its variance, or in a slightly more physiologically plausible way by computing the variance in small gaussian neighborhoods of each pixels and dividing by that variance[1]. Later in this chapter we will see that rather than using this ad-hoc preprocessing, it is also possible to estimate optimized filters for gain control from the data.

[1]This kind of processing was proposed by Bruno Olshausen

4.2.2 Alternatives to ISA

ISA has been criticized for the use of a fixed, rather than learned, pooling [35] and it has been argued that it is not possible to learn complex cell responses from static images in a principled way [65]. As a possible alternative, methods have been proposed that use short sequences of *natural movies* to learn complex cell properties. Körding et al. did this using movie sequences from a head-mounted camera from a cat [21] to learn pairs of linear filters which were subsequently summed and squared, and optimized for temporal stability of the outputs. This lead to similar results as those obtained with ISA, but while it addressed the point of using more naturalistic stimuli, it still used a hard-coded energy pooling.

Another approach, *slow feature analysis* (SFA) [123, 10], has been used to learn phase invariant receptive fields without using a fixed pooling, also by optimizing outputs to change as little as possible over time. Intuitively, this *slowness* criterion favors complex cell-like responses because they have a slight position invariance, so by translating the input image, which is the most common transformation in natural movies, the cell smoothly changes its activity. SFA creates a nonlinear mapping by projecting the normalized inputs into a high-dimensional features space, similar to the kernel spaces used e.g. in support vector machines [117]. In this feature space, a temporal derivative is computed and a set of linear filters is estimated that optimizes the desired slowness property. This is achieved by performing PCA in the temporal derivative feature space and selecting the directions with the smallest eigenvalues. While this method makes less assumptions on the form of the model (the nonlinear mapping is chosen in a very general way as the monomials of degree one and two, including quadratic terms and terms such as $x_1 x_2$, see also [77]), it was only demonstrated on an artificially generated data set, so it is not clear at the time how the model would perform with natural movie data.

Another possibility is to abandon the concept of learned linear filters altogether and to model only the radial component of the density of natural image patches. This has been the main focus in the work of Matthias Bethge [110] and Siwei Lyu [79]. Both authors have shown that for small, whitened image patches, the distribution is much closer to spherical than to factorial, and that a closer fit to the true distribution can be obtained by modeling the radial component of the pdf then to optimize a set of linear filters. The drawback of this approach is that very little structure can be encoded in what is essentially a single parametric or non-parametric fit to a filter output histogram. While the fit of the model to the data, as measured e.g. by the Kullback-Leibler divergence [17], provides an obvious

way of judging model quality, ultimately we are interested in inferring as much as possible about the structure of the data, for which learning linear transforms provides a much more general framework than working with the radial component of the density only.

4.2.3 ISA and Complex Cells

With ISA we can reproduce the receptive field properties of complex cells which can be seen by comparing the basis functions in Fig. 4.2 with the energy model in the second chapter. This suggests that the phase-invariant responses distinguishing complex from simple cells can be understood in terms of statistical optimality. We have shown that computing complex cell responses allows us to obtain a better match of the model distribution to the statistics of natural images than the simple cell model does, explaining why it is advantageous to pool over simple cells for a phase-invariant response. Together with Topographic ICA [54], which is an alternative way to model energy dependencies within groups of linear filter responses, ISA provides a nonlinear extension to ICA which captures more abstract properties of the stimulus, but still follows the objective of maximizing independence between the outputs of different units.

A crucial point about this hierarchical, nonlinear processing is that it gives rise to invariances. The phase invariance we see in complex cells allows for the reliable detection of stimuli with a particular orientation without being sensitive to small shifts in the position of the stimulus. While this is still a long way from full translation and scale invariance, as many object detection tasks require, it shows that even a relatively simple model of natural image statistics can lead to important coding principles beyond simple linear filtering.

Another important point to note about the basis functions in Fig. 4.2 is that the individual linear basis functions are not quite the same as the ICA basis functions we have seen previously. The addition of the second layer leads to subtle changes in the individual units, which are not as Gabor-like as in the simple ICA case, but are adapted to the processing in the next higher layer in the hierarchy. This illustrates an important point; the features at any one layer are not only tuned to the input signal, but also adapted to later processing steps. This should be kept in mind when estimating multi-layer models, where it may be tempting to fix the lowest level e.g. to an ICA basis, but this may seriously impair the performance and validity of the model.

4.3 Multi-Layer Models

In order to further investigate the hypothesis that complex cell responses can be derived from static natural images by efficient coding algorithms, several hierarchical models were developed [61, 62, 91], including *Publications 3 and 4* [66, 67] of this thesis. By learning two layers of linear filters, with a scalar nonlinearity in between, the arbitrary choice of pooling e.g. two or four units into a subspace is replaced by a principled estimation of the correct pooling from the statistical structure of the data. This is important since virtually all pairs of linear filters in an ICA model exhibit energy dependencies, so the correct pooling to account for these dependencies is not at all obvious.

4.3.1 Generative and Energy-Based Models

The existing two layer models can be grouped into two classes, *generative models* and *energy-based models*. As an example of a generative model we have seen sparse coding, where the model specifies how the data is *generated* (in this example as a linear superposition of sources, $\mathbf{x} = A\mathbf{s}$), so it is easy to draw samples from the model, but it is hard to assign a probability to observed data (in the sparse coding example, this required a gradient optimization). On the contrary, in energy based models, it is easy to assign an energy (or log-probability) to an observed data vector, but it is hard to generate, or draw samples from the model. Energy-based models are so called because they work with non-normalized log-probabilities rather than with probabilities, as the normalization factor often cannot be expressed in closed form.

4.3.2 Hierarchical Model with Score Matching Estimation

Let us consider an extension to ISA where the fixed pooling has been replaced by a linear transform estimated from the data. The model, which is described in more detail in *Publications 3 and 4* is of the form

$$E(\mathbf{x}) = \sum_i f\left(\mathbf{v}_i^{\mathrm{T}} g(W\mathbf{x})\right) \tag{4.3}$$

where g and f are fixed scalar nonlinearities, W is the first layer weight matrix and V, containing row vectors $\mathbf{v}^{\mathrm{T}}$, is the non-negative second layer. The first nonlinearity g is a rectifying nonlinearity, following the energy model of complex cells, and the second nonlinearity f serves to produce a supergaussian distribution of the outputs. The pdf of the model is defined as the exponential of the negative energy, normalized to integrate to

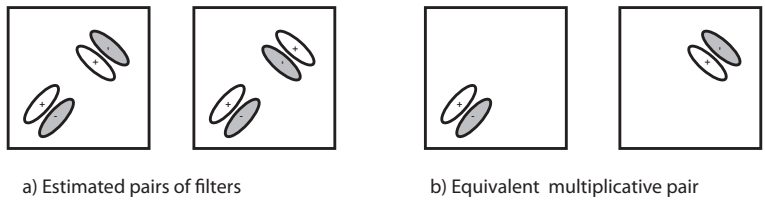


Figure 4.3: Estimating the two layer score matching model with real-valued weights V leads to the emergence of pairs of filters like that in a). Subtracting the squares of these filters is equivalent to multiplying the two filters shown in b). The emergence of these multiplicative pairs possibly indicates that the nonlinearity of the model needs to be matched to the sparseness of the data.

unity. In contrast to the fixed pooling of ISA, the matrix V makes the normalization of the model in closed form impossible. Previously, learning in models like this required model-specific approximations or computationally expensive sampling methods such as Markov chain Monte Carlo (MCMC). With the score matching framework that we introduced in Sec. 3.5 however, estimation is straightforward and does not require approximations or sampling.

Estimating the model for natural image data leads to familiar Gabor-like features in the first layer (not shown), and a pooling of these linear filters in the second layer. This is illustrated in Fig. 4.4 b) for a random selection of second layer units, where all the Gabors from the first layer that contribute strongly to the particular unit are represented as ellipses. It can be seen that each of the second layer units pools over a small number of linear filters which mostly share the same position and orientation, but have different spatial phase (not indicated in the plot). Thus the outputs are very similar to the phase-invariant features in ISA, but without the need for any assumptions on the pooling other than non-negativity. Again the model was estimated on image data pre-processed with contrast gain control, and it is likely that pooling would be much less specific without this pre-processing.

The non-negativity constraint is required for two technical reasons: firstly, the overall pdf obtained by combining the two nonlinearities needs to be supergaussian. To offset the squaring-like effect of the rectifying first nonlinearity, the second nonlinearity needs to be shaped like a square-root, which has a discontinuity at zero. It would seem that this first issue could be alleviated e.g. by compounding two log-cosh nonlinearities, but this leads

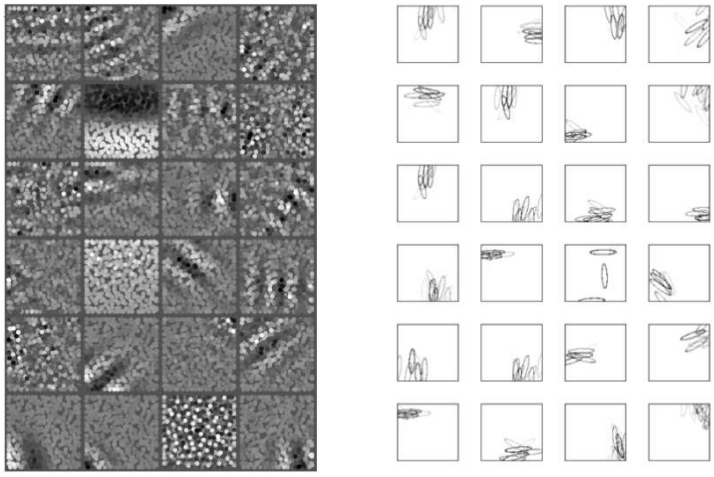
to the second problem: the model then learns to pair two Gabors each into two linear receptive field, as shown in Fig. 4.3 a), and pools these pairs with one negative and one positive weight, effectively subtracting one off the other. By using the identity $a^2-b^2 = (a-b)(a+b)$, we can see that this corresponds to multiplying the responses of two linear filters that contain only one of the Gabors in each receptive field. While it would be tempting to interpret this as some kind of non-linear end-stopping behavior like corner detection, it seems more likely that it points out a flaw in the choice of the nonlinearities. Multiplying pairs of outputs allows a better match of the model pdf to the high sparseness of the data distribution, which is not well captured by the very smooth log-cosh function. It would be a very interesting direction for future work to investigate what nonlinearities would be best suited to the data and would still allow the model to be estimated in the score matching framework. Extremely peaked distributions are generally problematic here because the gradients of the objective include terms up to the third derivative, making the the estimation very cumbersome if the functions are strongly peaked. Once this problem is solved, it would be feasible to extend the model with a third or more layers.

4.3.3 Hierarchical Product of Experts

Similar to the work presented in the previous section, the hierarchical *product of experts* by Simon Osindero is an energy-based model, and due to the intractable normalization factor, straightforward maximum likelihood estimation is not possible. Instead, the authors resort to using *contrastive divergence* (CD) [40], a Markov chain Monte Carlo method that works by comparing the data distribution to the model distribution after taking only a single Monte Carlo step. The model is defined as a product of modified Student-t distributions [41] with the pdf

$$p(\mathbf{x}) \propto \prod_i \frac{1}{1 + \mathbf{v}_i^{\mathrm{T}}(W\mathbf{x})^2}. \tag{4.4}$$

If the second layer weight matrix V, which consists of vectors $\mathbf{v}_i$, is identity, the model reduces to a classical ICA / product of experts model. The weights V are constrained to be non-negative, and estimated at the same time as the first layer W. The authors report that when trained on natural image patches, the rows of V pool over first layer outputs to produce receptive fields resembling those of complex cells.



a) Pooling patterns from Karklin & Lewicki Model

b) Pooling patterns from Score Matching model

Figure 4.4: Comparison of the pooling patterns obtained by the hierarchical Bayesian model by Karklin and Lewicki (reproduced from [62]), and the two layer model estimated by score matching. For the Bayesian model, individual first order units are represented by dots localized at the center of the linear filter and shaded according to activation strength. For the score matching model, units are represented by ellipses with the major axis indicating the orientation of the underlying linear filter. The latter model shows highly specific pooling of co-localized, iso-oriented filters, whereas the pooling in the Bayesian model is much broader, often covering the whole image patch.

4.3.4 Hierarchical Bayesian Model

The *hierarchical Bayesian model* [62] that was developed by Yan Karklin and Mike Lewicki is a generative model and can be viewed as an extension of topographic ICA (TICA). To understand how this model relates to our own hierarchical model, let us quickly review TICA. In contrast to the classic generative ICA model, the components are not generated independently, but in groups with a common variance variable [52]. This leads to a positive correlation of squares between components within groups, but in contrast to ISA the groups are overlapping and therefore define a topography on the

filters. The components can be written as

$$s_i = \phi(\mathbf{b}_i^{\mathrm{T}}\mathbf{u})z_i \tag{4.5}$$

where $\mathbf{u}$ are higher order components giving rise to the variance dependencies and the matrix B, consisting of row vectors $\mathbf{b}_i^{\mathrm{T}}$, generates the topography. The z_i are independent, supergaussian variables, and the data is created by a linear mixing of the s_i. Estimating this model necessitates the use of an approximation, which amounts to estimating an energy-based model similar to ISA.

If we consider the two-layer score matching model as an extension of energy-based ISA, we can view the hierarchical Bayesian model as an extension of the generative TICA model, where the second layer weights are estimated in addition to the first layer. Like in TICA, higher order variables u are drawn from their distribution and mixed with a mixing matrix B. The higher order mixtures then provide the variances for the independent components s in the model. In a first instantiation [62] the authors used a fixed ICA basis in the first layer and only estimated the second layer with a *maximum a posteriori* (MAP) approximation. In a later publication [63] the authors used a full, simultaneous estimation of both the first layer features W and the higher order features B and reported that this leads to a significant change in the first layer features, which is in agreement with our own results. This should not come as a surprise though, since we have already seen in ISA how a particular (fixed) second layer can influence the exact shape of first layer linear filters. In all the experiments the authors used a MAP approximation of the latent variables u to optimize the linear filters. Because of this generative approach, which includes the estimation of latent variables, rather than the feedforward computation of the energy-based models, quite different results are obtained. The model does not give rise to the classical complex cell pooling, but a variety of higher order features pooling over a large fraction of the inputs distributed over various positions and orientations. This can be seen in Fig. 4.4, where the pooling patterns are compared with those from the score matching model.

4.4 Horizontal Model for Gain Control

Even though we have highlighted the importance of gain control in visual processing throughout this thesis, many of the models we discussed so far included only very rudimentary gain control, if any. Linear models such as ICA have traditionally been employed without any gain control, and for the comparison of subspace sizes with ISA we used an ad-hoc method

for divisive normalization as described previously. In accordance with the underlying principle of this work, it would be preferable to estimate the processing from the data. In *Publication 5* [68] we attempted to estimate an ICA model on natural image patches, and at the same time estimate the optimal gain control. This was done using a generative model in some ways similar to the hierarchical model by Karklin and Lewicki. The image patches are generated by a combination of two linear transforms of independent sources; one corresponding to the classical ICA model, and a second one that encodes the local contrast of the image patch. Thus, the generative model is of the form

$$\mathbf{x} = A\mathbf{s} \odot B\mathbf{t} \tag{4.6}$$

where $\mathbf{s}$ and $\mathbf{t}$ are the independent sources of image structure and contrast respectively, and A and B are linear transforms to be estimated from the data. The element-wise product, denoted by $\odot$, is used to modulate each pixel with a scalar gain factor. Due to the multiplication of the two types of sources, the model is highly nonlinear, even though there are no scalar nonlinearities like in the previous models we saw. In comparison to those models, where the goal was to model the residual dependencies in the "independent components" by a nonlinear pooling, the goal here is to perform a nonlinear rescaling on the data before estimating the ICA model. This can be seen by rewriting Eq. 4.6 as

$$\mathbf{x} = \mathrm{diag}(B\mathbf{t})A\mathbf{s} \tag{4.7}$$

$$\mathbf{s} = W\mathrm{diag}(B\mathbf{t})^{-1}\mathbf{x} \tag{4.8}$$

where we have used the ICA convention of writing $W = A^{-1}$ and replaced the element-wise product of the vectors by multiplication with a diagonal matrix containing the elements of the vector. In this type of model we constrain the gain components $\mathbf{t}$ and the rows of B to be non-negative, since the gain of a pixel should not influence the sign. Typically we restrict the dimensionality of $\mathbf{t}$ to be quite small compared to the dimensionality of the data $\mathbf{x}$, since most of the information about the image patch should be encoded in A, and we are trying to find only a few basis vectors that can describe most of the variance patterns in the data.

Estimating this model on natural image patches leads to the two results illustrated in Fig. 4.5: the basis functions estimated for the contrast part of the model take the shape of localized, Gaussian "blobs" that tile the image patch. This is not entirely unexpected, since the variance changes slowly from one image region to the next. In fact, the normalization estimated in

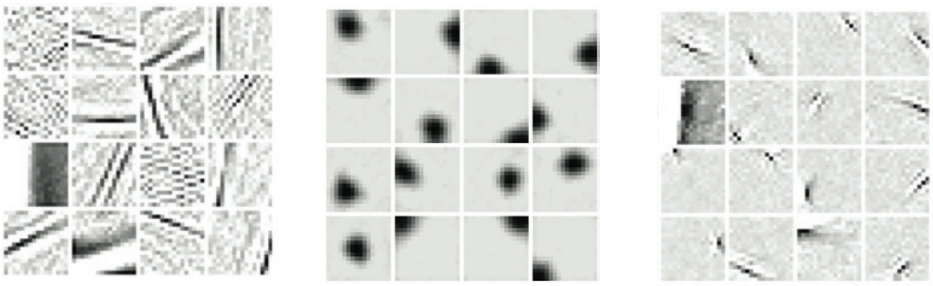


Figure 4.5: The product model for gain control on the pixel level. a) A subset of filter vectors from W, b) the 16 basis function from B and c) ordinary ICA filters for comparison. The filters in the product model are less localized than for the classical ICA model. The extra network layer converges to approximately spherical "blobs" tiling the image patch.

this way is very similar to the ad-hoc gain control we described previously, where we divided by the variance in Gaussian neighborhoods.

The second result from the estimation of the model is that the optimal ICA filters in this framework are very different from those in classical ICA. While the basis functions are still Gabor-like, with orientation and frequency tuning, they lose much of the original location selectivity and cover a significant fraction of the image patch. It could in fact be said that the model learns to separate the Gabor into the sinusoidal part and the Gaussian envelope. While it is hard to interpret this result in the context of physiology and biological visual processing, it shows that gain control has a very significant effect on the optimal processing in the next stage of the hierarchy and should not be ignored.

4.5 Markov Random Fields

In all of the models we discussed previously, we focused on very small image patches and made no attempts to generalize to larger images or whole natural scenes. This was necessary because high dimensional data would make computations excruciatingly slow. It can also be justified by the fact that most cells in the early visual system have very localized receptive fields. Let us now consider a model that attempts to overcome these limitations and which is the subject of *Publication 6* [69].

While there are long-range correlations in natural images, as we have

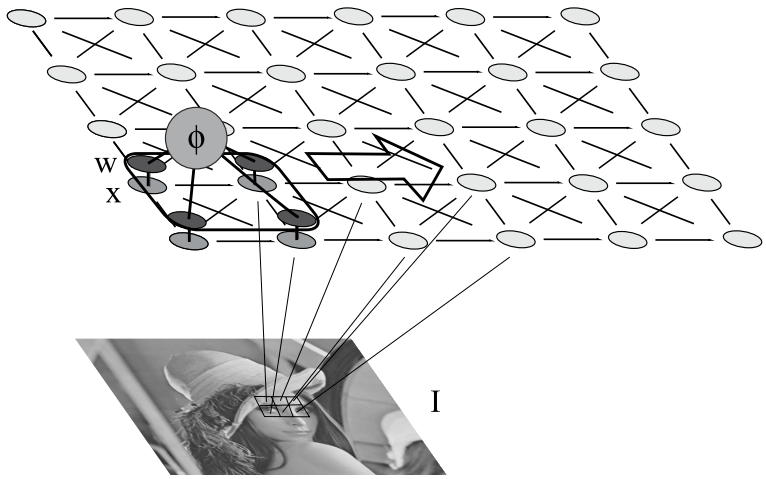


Figure 4.6: Illustration of a Markov random field. The clique size is 2×2, one clique $\mathbf{x}$ is highlighted in the lower left corner. The energy terms of the field are computed by applying the potential function ϕ to the outputs of linear filters $\mathbf{w}$ as indicated for one clique. By convolving the potential functions, or filters, with the image I and summing all terms, the total energy is obtained. The unnormalized probability of the image is then given by the exponential of the negative energy.

seen in Fig. 3.2, and two far-away pixels may have high level dependencies as in e.g. belonging to the same object, it is reasonable to assume that most low-level structure can be modeled in terms of local interactions. This can be formalized as the Markov property: given the values of a clique of neighboring pixels, the one pixel we are considering is conditionally independent of the rest of the image. From this starting point we can build *Markov random fields* (MRF) [76], graphs with dense local connectivity, but no long range connections. The *maximal cliques* have associated *potential functions*, that assign an energy to the data under that clique. These potential functions are repeated for each maximal clique, tiling the image in a convolutional way. This is illustrated in Fig. 4.6.

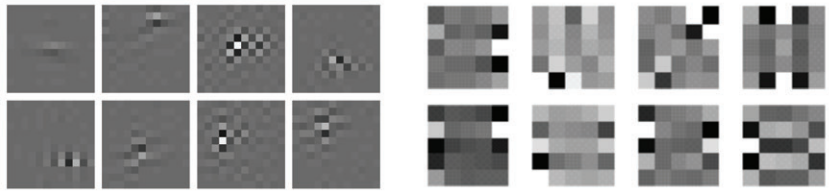
MRFs have traditionally been used with very small potential functions, which have been selected by hand (e.g. [127]) rather than learned. These models have been used for applications such as novel view synthesis [124] and texture modeling [127]. Typically the filters that define the potentials are of only 3×3 pixel size, and are modeled after spatial derivative filters. Only recently Roth and Black have shown that MRF filters can

be estimated from natural image data [100] by generalizing the product of experts framework to the *fields of experts* (FoE). However, with the estimation using contrastive divergence, learning is very slow and the approach is still limited to small potentials of 5×5 pixel size.

By estimating a similar model with score matching, we have shown how MRF potentials can easily be estimated for potentials of 12×12 pixels using "images" of 36×36 pixels size. While the model is virtually identical to the FoE, the estimated filters are quite different: in our MRF, we find filters similar to the Gabor functions obtained by classical ICA, whereas the filters for the FoE are discontinuous, as depicted in Fig. 4.7 b). It is not clear at this time what causes these differences, and in particular why the FoE filters break up into discontinuous regions.

The high frequencies of the FoE filters are easily explained since the model operates on non-witened data. In an energy-based model, the filters preferably take directions in data space that result in low, rather than strong responses, which correspond to the highest frequencies of natural images [121]. This partially explains the good denoising performance of the high frequency FoE filters, which model the high frequencies with the lowest signal-to-noise ratio particularly well. Indeed it has been shown in [59] that the FoE tends to over-smooth, indicating that it is strongly penalizing the high spatial frequency components. However, both our MRF and the FoE model differ strongly even when whitening is accounted for, so there is no clear reason for the filters to be vastly different from the results we obtained. This raises the possibility that the CD algorithm used by the authors did not converge correctly, which would also agree with the observation that the FoE algorithm converges to qualitatively different local minima depending on details of the estimation, sometimes converging to filters which perform worse than random filters in the benchmarks used by the authors [99]. Still we cannot exclude the possibility that the differences are due to treatment of image borders. In particular, we did not proof rigorously that our approach of computing the score matching objective only w.r.t. the central image pixels indeed corresponds to working with infinitely large images, and an empirical verification would require training on significantly larger images than what is practical.

The similarity between the MRF model and ICA should not be surprising though, because the MRF can be considered as a special case of a highly overcomplete ICA model. This works by imposing two constraints on the ICA model, which is estimated for the "images" (of e.g. 36×36 pixel size) rather than for the cliques (which are e.g. 12×12 pixels). The ICA filters are constrained to cover only a region of 12×12 of the image and overcom-



a) 12 x 12 filters estimated with score matching

b) 5 x 5 filters estimated with the field of experts approach

Figure 4.7: A random selection of filters learned with our MRF compared with filters from the fields of experts model (reproduced from [121]). With score matching the model can be estimated for larger maximal cliques, in this example of 12×12 pixels. For the comparison we have absorbed the whitening into the filters. Since no dimensionality reduction was performed, they are dominated by the highest spatial frequencies. Still they are well described as Gabor functions, whereas the PoE model estimation leads to discontinuous filters very different from the Gabors of ICA models.

pleteness is achieved by placing identical copies of the 12×12 region that contains the filter in all possible positions within the 36×36 image. This overcomplete ICA model is identical to the MRF. Due to the fact that the 12×12 filters are implicitly applied to larger images, it is not surprising that they are on average slightly larger than ICA filters estimated on 12×12 image patches. In addition, the extremely high implicit overcompleteness gives an intuitive justification to the fact that the filters, which are shown in Fig. 4.7 a), seem more diverse in appearance than ordinary ICA filters.

In comparison to the previous models discussed here, the assumptions that define the MRF give the model two major advantages. Firstly, the MRF is not limited to small patches but can be applied to images of arbitrary size. While this seems important mainly for image processing applications, it is more than just a technical advancement: by making the explicit model assumption that interactions should be of limited range, the estimation of a model for large images is greatly simplified because there is no longer any need to train on images significantly larger than the patch size. This model constraint is justified from the observation that even when estimated on large image patches, ICA always produces localized basis functions that span only a fraction of the whole image patch.

The second advantage of the MRF is the explicit translation invariance, which is more of technical rather than neuroscientific interest. In an ICA

model, the translation invariance that is inherent in natural images has to be reflected by a spatial tiling of identical filters. This is expensive since it requires the estimation of many more filters in an overcomplete model than the estimation of a model with build-in translation invariance does. The high overcompleteness that is implicit in any MRF model compared to ICA thus allows a much more detailed statistical description of the stimulus, while requiring the estimation of fewer parameters.

While the MRF most certainly does not provide us with a better description of neural processing *per se*, these two advantages make it a significantly more powerful model of natural images, and may therefore lead to new insights about visual processing. We can apply the model to real-world tasks such as filling-in of large missing image regions, which are out of the realm of patch based methods, since the required large patch sizes would lead to an explosion of dimensionality and make learning impractical. With the MRF, we can compare the performance of the model with that of the human visual apparatus and judge how much of the structure of natural images has actually been captured in an immediately useful way.

5

Conclusion

Essentially,
all models are wrong,
but some are useful.
G. E. P. Box

5.1 Discussion

In the first chapter of this thesis, we posed a number of research questions as a guide through this work. Chapters 2 and 3 served mainly to put these questions into perspective by describing the problem at hand in more detail and discussing previous attempts at solving these problems. In Chapter 4 the contribution of our work was presented and the relation to other approaches was established. Here we will revisit the questions and try to answer them using the insights and results we have gained from the models discussed in the previous chapter, and in more detail in the publications in the second part of this thesis.

RQ1: What are suitable statistical models for patches of natural images?

From the beginning, we have focused on hierarchical models, which is clearly not the only and quite possibly not the best choice to capture the structure of natural images. For example, a perceptron with only a single hidden layer can represent any function with arbitrary accuracy given enough units [42]. As we have seen in Chapter 2 though, the brain is very successful using hierarchies of many areas to perform vision, so by constraining our search to methods that fit this framework, we can reduce our search space to something more manageable. Though relatively little is known about processing in biological visual systems, we can attempt to approach a viable solution to the problem, by comparing with - and ultimately trying to predict - the processing of those biological systems. With the added benefits of the conceptual simplicity and computational tractability, hierarchical, energy based models are a very strong candidate for modeling natural images in such a way that both advances in vision as an engineering problem and as a neuroscientific problem can be made.

With the hierarchical model in *Publications 3 and 4* we have proposed a framework that can potentially be extended to more than two layers. There are no fundamental obstacles to this, except that it becomes very tedious to implement the estimation for three or more layers. The hierarchical model gives a quantitatively better statistical description of natural images than previous models such as ISA and TICA, which it includes as special cases.

RQ2: How can multi-layer models of natural images be estimated?

We have repeatedly used score matching and we have shown that it provides a powerful estimation principle for energy-based models. It allows for consistent parameter estimation with much reduced computational load compared to alternative methods, and it is generally quite easy to derive and optimize the objective function. An alternative to the energy-based

approach is to use generative models, which generally require the estimation of latent variables. We have followed this route in *Publication 6*, where we used a MAP-approximation for the latent variables.

Energy based models have the advantage that the probability of a data vector is given by a simple feed-forward computation, and with score matching there is a straightforward way for model estimation. However, it is difficult to draw samples from the model distribution. Generative models can possibly be considered to be more principled, because they provide a mechanistic description of the process that generates the data. They have their own share of problems we alluded to, mainly the difficult estimation which usually needs to be tuned for the particular model at hand and often requires approximations. In conclusion, both of these classes of models can be used to estimate the statistical structure of natural images, but the jury is still out on which model class better reflects the processing in the brain.

RQ3: Can we show that complex cells provide a better statistical description of images than linear filters?

Some previous complex cell models that attempted to explain the receptive fields as being matched to the statistics of natural images, were weakened by rigid model assumptions. In ISA a fixed pooling nonlinearity was used, as was the case in the related method using movie sequences [65]. By directly comparing the likelihood of the ISA model with classical ICA, we have shown in *Publication 2* that the subspace model has a higher likelihood for image data, so we can conclude that phase-invariant, complex cell-like units are in fact better adapted to the statistics of natural images. We explored this further in *Publications 3 and 4*, where the fixed pooling was replaced by a second layer of arbitrary connectivity, estimated from the data. Again the emergence of complex-cell receptive fields provides evidence that pooling in spherical subspaces gives a good description of the statistical structure of the data.

A clear weakness of the latter model is that it uses a fixed nonlinearity, and also the ISA model was only estimated for the relatively constrained family of generalized Gaussian distributions. Estimating the correct form of the nonlinearity has in general been neglected since it is a nonparametric problem. Furthermore, it is not easy to visualize and interpret the influence of the nonlinearity on the distribution. Another drawback of the two-layer model was the restriction to non-negative connections in the second layer for technical reasons, so the model was still restricted to perform some kind of pooling in the second layer. It is an interesting direction for future research to lift this constraint and test whether complex cell responses are still obtained.

RQ4: Is gain control in the visual system matched to the optimal processing of the stimulus, and how does gain control affect the later processing?

In our attempt to answer the last of our research questions, we have taken a rather different approach from previous work. While models of gain control have received much attention, they have almost exclusively been applied on the level of simple and complex cells. Our question, however, was aimed at the effect of mostly retinal gain control mechanisms and how this affects later processing stages.

We already saw that this type of gain control has an important effect in *Publication 2*. The pooling into small subspaces that is typically associated with complex cells was shown to be optimal only after normalizing the variance of the image patches; without this preprocessing it is advantageous to pool a very large number of linear filters, giving an effectively spherical output distribution.

However, this result alone is a rather weak justification to apply gain control as preprocessing. After all it is the spirit of this work to estimate all processing from the data, rather than fixing it by hand. Our results in *Publication 6* show that this is indeed possible, and leads to the emergence of gain control over small Gaussian neighborhoods. This makes it possible to interpret much of the divisive normalization that occurs in the retina and LGN as processing optimized to the statistical structure of the stimulus. Furthermore, the changes we observed in the linear filters compared to the ICA model serve to emphasize that conclusions about any one layer of the model cannot be made in isolation, but it is important to consider several layers of the hierarchy simultaneously. Interactions between the layers greatly affect the resulting outputs, necessitating the estimation of more than one network layer as we have done here.

5.2 Future Outlook

What we have seen in this thesis about natural image statistics and visual processing is that it is a rocky road from the simple and elegant idea of utilizing stimulus statistics for inferring the optimal processing, to making testable predictions about the visual system. Even the idea of interpreting simple cell responses by analogy to ICA on natural images [116], which has been around for more than a decade, is continuing to be challenged. Our understanding of the processing that occurs in the primary visual cortex is incomplete at best [90, 12], and as little as 20-40% of the variance of individual neural responses can be explained [18]. As new methods provide

a better explanation of the underlying neural processing and the classical ideas of single neuron receptive fields make way to more abstract models of population responses [92], simple models such as linear ICA become increasingly hard-pressed to provide a satisfactory explanation of neural properties.

On the other hand, even where statistical models can provide a satisfactory explanation, we can never rule out the possibility that the receptive fields appear to be optimized for statistical criteria purely by coincidence. It is possible that the tuning properties of simple and complex cells have developed for very different reasons than to provide a sparse, independent code.

Beyond this, the stated goal of the study of natural image statistics, namely to provide testable hypotheses about processing in higher cortical areas which are not yet well understood, is facing more serious problems. While the idea of unsupervised learning is to put as few constraints as possible in the model, it turns out that these "few constraints" still greatly influence what the model can and cannot do. The linear transform model in sparse coding and ICA was chosen *because* simple cells could successfully be modeled with a linear transform. Likewise, the pooling of squared responses in ISA and other complex cell models followed the energy model that was created as a way to describe results from physiology. While learning the correct linear filters within such a model framework is by no means a small achievement, what is really needed is a framework to estimate the correct model architecture [113] rather than a set of linear transformations, given the hand-crafted model structure. Since this is a non-parametric problem, where an effectively infinite number of parameters has to be estimated, progress in this direction has been very slow.

From the previous paragraphs we can conclude that this aspect of computational neuroscience is still in its infancy and holds many interesting challenges. It is therefore important to keep in mind the quote at the beginning of this chapter, and to avoid trying to find too close a link between the models of natural image statistics described here on one hand, and the processing in the brain on the other hand. That said, there certainly is much more to be learned about visual processing from models of the kind described here. The field is changing rapidly with new models and estimation methods constantly being developed, and there is much unchartered ground to be explored. Multilayer models such as the hierarchical model we considered here have only been around for a very short time, and we already saw several ways on how they can be extended by adding more layers or lifting connectivity constraints. Additionally, we have so far only

considered particular, non-overlapping aspects of the statistical structure, so they can be combined to form more powerful representations. Over the last 20 years we have seen a rapid development from simple linear models to approaches of ever-increasing sophistication. The current generation of models is using nonlinearities to model relatively simple invariances on the level of complex cells or for contrast gain control, but continuing this line of work and generalizing it to other, less straightforward nonlinear effects, holds the promise to give testable predictions about biological visual processing.

References

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A 2*, pages 284 – 299, 1985.

[2] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.

[3] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, 1992.

[4] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[5] R. Baddeley, L. F. Abbott, M. C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc R Soc Lond B Biol Sci*, 264(1389):1775–1783, 1997.

[6] H. B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001.

[7] H.B. Barlow. *Possible principles underlying the transformation of sensory messages*. Cambridge, MA: MIT Press, 1961. W. Rosenblith (Ed.) Sensory Communication.

[8] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[9] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[10] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5:579–602, 2005.

[11] M. Bertalmío, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *ACM SIGGRAPH*, pages 417–424, 2000.

[12] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

[13] J.-F. Cardoso and B. Hvam Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, 1996.

[14] D. D. Clark and L. Sokoloff. *Basic Neurochemistry: Molecular, Cellular and Medical Aspects.* Philadelphia: Lippincott., 1999. Siegel GJ, Agranoff BW, Albers RW, Fisher SK, Uhler MD (Ed.).

[15] P. Comon. Analyse en composantes indépendantes et identification aveugle. *Traitement du Signal*, 7(5):435–450, 1990.

[16] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[17] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd edition.* Wiley, 2006.

[18] S. V. David, W. E. Vinje, and J. L. Gallant. Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci*, 24(31):6991–7006, August 2004.

[19] R. Descartes. *Traité de l'homme.* Charles Angot, Paris, 1664. Grench.

[20] D. W. Dong and J. J. Atick. Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus. In *Network*, pages 159–178, 1995.

[21] W. Einhäuser, C. Kayser, P. König, and K.P. Körding. Learning the invariance properties of complex cells from natural stimuli. *Eur J Neurosci*, 15(3):475–86, 2002.

[22] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. In *Inverse Problems 23 (2007*, pages 947–968, 2005.

[23] D. J. Felleman and D. C. van Essen. Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[24] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4:2379–2394, 1987.

[25] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[26] P. Földiák and M. P. Young. *Sparse coding in the primate cortex*. MIT Press, Cambridge, MA, USA, 1998.

[27] K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, 1988.

[28] K. Fukushima. Neocognitron for handwritten digit recognition. *Neurocomputing*, 51:161–180, 2003.

[29] M. S. Gazzaniga, R. B. Ivry, and G.R. Mangun. *Cognitive Neuroscience: The biology of the mind*. W. W. Norton, New York, 2002. Second edition.

[30] J. J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin, Boston, 1979.

[31] T. Gollisch and M. Meister. Modeling convergent on and off pathways in the early visual system. *Biological Cybernetics*, 99(4):263–278, 2008.

[32] M. A. Goodale and A. D. Milner. Separate pathways for perception and action. *Trends in Neuroscience*, 15:20–25, 1992.

[33] R. J. Greenspan. *An Introduction to Nervous Systems*. CSHL Press, 2007.

[34] R. Hadsell, P. Sermanet, M. Scoffier, A. Erkan, K. Kavackuoglu, U. Muller, and Y. LeCun. Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics*, 26(2):120–144, February 2009.

[35] W. Hashimoto. Quadratic forms in natural images. *Network: Computation in Neural Systems*, 14(4):765–88, 2003.

[36] D. J. Heeger. Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9:181–198, 1992.

[37] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197, 1992.

[38] D. J. Heeger and M. Rees. Neural correlates of visual attention and perception. In M. S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 341–348. The MIT Press, 2004.

[39] G. E. Hinton. Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*, volume 1, pages 1–6, 1999.

[40] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[41] G. E. Hinton and T. J. Sejnowski, editors. *Unsupervised Learning*. MIT Press, 1999.

[42] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[43] D. H. Hubel. *Eye, Brain, and Vision (Scientific American Library)*. W H Freeman & Co (Sd), 1988.

[44] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J Physiol.*, 148:574 – 591, 1959.

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

[46] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195:215–243, 1968.

[47] I. W. Hunter and M.J. Korenberg. The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biological Cybernetics*, 55(2-3):135–144, 1986.

[48] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[49] A. Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[50] A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007.

[51] A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51:2499–2512, 2007.

[52] A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation.*, 13(7):1527–1558, 2001.

[53] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[54] A. Hyvärinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41:2413 – 2423, 2001.

[55] A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics*. Springer-Verlag, 2009. In press.

[56] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[57] A. Hyvärinen and U. Köster. FastISA: A fast fixed-point algorithm for independent subspace analysis. In *Advances in Computational Intelligence and Learning (ESANN2006)*, pages 798–807, 2006.

[58] A. Hyvärinen and U. Köster. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18:81–100, 2007.

[59] V. Jain and H. S. Seung. Natural image denoising with convolutional networks. In *Advances in Neural Information Processing Systems 21 (NIPS2008)*, 2008.

[60] C. Jutten and J. Herault. Blind separation of sources part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[61] Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499, 2003.

[62] Y. Karklin and M. S. Lewicki. A hierarchical bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423, 2005.

[63] Y. Karklin and M. S. Lewicki. Is early vision optimized for extracting higher-order dependencies? *Advances in Neural Information Processing Systems*, 18:625–642, 2006.

[64] T. Kohonen. Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75:281–291, 1996.

[65] K. Körding, C. Kayser, W. Einhäuser, and P. König. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91(1):206–12, 2004.

[66] U. Köster and A. Hyvärinen. A two-layer ICA-like model estimated by score matching. In *Artificial Neural Networks - ICANN 2007, Lecture Notes in Computer Science*, pages 798–807. Springer Berlin / Heidelberg, 2007.

[67] U. Köster and A. Hyvärinen. A two-layer model of natural stimuli estimated with score matching. 2009. Submitted manuscript.

[68] U. Köster, A. Hyvärinen, M. Gutman, and J. T. Lindgren. Learning natural image structure with a horizontal product model. In *ICA 2007, Lecture Notes in Computer Science*, pages 507–514. Springer Berlin / Heidelberg, 2009.

[69] U. Köster, A. Hyvärinen, and J. T. Lindgren. Estimating Markov random field potentials for natural images. In *ICA 2007, Lecture Notes in Computer Science*, pages 515–522. Springer Berlin / Heidelberg, 2009.

[70] S. B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung*, 36c:910–912, 1981.

[71] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[72] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of CVPR'04*. IEEE Press, 2004.

[73] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, 2006.

[74] P. Lennie. The cost of cortical computation. *Current biology : CB*, 13(6):493–497, March 2003.

[75] M. S. Lewicki and B. A. Olshausen. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A*, 16:1587–1601, 1999.

[76] S. Z. Li. *Markov Random Field modeling in image analysis, 2nd edition*. Springer, 2001.

[77] J. T. Lindgren and A. Hyvärinen. Emergence of conjunctive visual features by quadratic independent component analysis. *Advances in Neural Information Processing Systems*, 2006.

[78] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[79] S. Lyu and E. P. Simoncelli. Nonlinear extraction of 'independent components' of natural images using radial Gaussianization. *Neural Computation*, 21(6):1485–1519, Jun 2009.

[80] S. Lyu and E. P. Simoncelli. Reducing statistical dependencies in natural signals using radial Gaussianization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Adv. Neural Information Processing Systems 21*, volume 21, pages 1009–1016, Cambridge, MA, May 2009. MIT Press.

[81] E. Mach. *Die Analyse der Empfindungen und das Verhältnis des Physischen zum Psychischen*. Fischer, Jena, 1886.

[82] D. Marr. *Vision - A computational investigation into the human representation and processing of visual information*. Freeman, 1982.

[83] R. H. Masland. The fundamental plan of the retina. *Nat Neurosci*, 4(9):877–886, 2001.

[84] J. Maunsell and D. van Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation. *J Neurophysiol*, 49(5):1127–47, 1983.

[85] K. McAlonan, J. Cavanaugh, and R. H. H. Wurtz. Guarding the gateway to cortex with attention in visual thalamus. *Nature*, 2008.

[86] F. Mechler and D. L. Ringach. On the classification of simple and complex cells. *Vision Research*, 42(8):1017–33, 2002.

[87] M. Meister and M. J. Berry. The neural code of the retina. *Neuron*, 22:435–450, 1999.

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

[89] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[90] B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural Computation*, 17:1665–1699, 2005.

[91] S. Osindero, M. Welling, and G. E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18, 2006.

[92] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature*, 454(7206):995–999, Aug 2008.

[93] U. Polat and D. Sagi. Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33(7):993–999, 1993.

[94] D. A. Pollen and S. F. Ronner. Visual cortical neurons as localized spatial frequency filters. *IEEE Transactions on System, Man and Cybernetics*, 13:907–916, 1983.

[95] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, 2003.

[96] F. T. Qiu and R. von der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155–166, 2005.

[97] M. Riesenhuber and T. Poggio. Computational models of object recognition in cortex: A review. Technical report, Massachusetts Institute of Technology AI lab / Center of Biological And Computational Learning, Department of Brain and Cognitive Sciences, 2000.

[98] D. L. Ringach and R. Shapley. Reverse correlation in neurophysiology. *Cognitive Science*, 28(2):147–166, 2004.

[99] S. Roth. *High-Order Markov Random Fields for Low-Level Vision*. PhD thesis, Brown University, 2007.

[100] S. Roth and M. Black. Fields of experts: A framework for learning image priors. *CVPR, vol. 2*, pages 860–867., 2005.

[101] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physics Review Letters*, 73(6):814–817, 1994.

[102] F. S.Chance, S. B. Nelson, and L .F. Abbott. Complex cells as cortically amplified simple cells. *Nature Neuroscience*, 2(3):277–282, 1999.

[103] O. Schwartz, J. W. Pillow, N. C. Rust, and E. P. Simoncelli. Spike-triggered neural characterization. *J. Vis.*, 6(4):484–507, 7 2006.

[104] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.

[105] J. Sergent, S. Ohta, and B. MacDonald. Functional neuroanatomy of face and object processing. a positron emission tomography study. *Brain*, 115(1):15–36, 1992.

[106] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, volume 2, pages 994–1000, 2005.

[107] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[108] L. G. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, 2001.

[109] E. P. Simoncelli and E. Adelson. Noise removal via Bayesian wavelet coding. *Intl Conf. on Image Processing.*, pages 379–382, 1996.

[110] F. Sinz and M. Bethge. The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Neural Information Processing Systems 2008*, Cambridge, MA, USA, 2009. MIT Press.

[111] H. Spitzer and S. Hochstein. A complex-cell receptive-field model. *J. Neurophysiol. 53*, pages 1266 – 1286, 1985.

[112] K. Tanaka. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, 19:109–139, 1996.

[113] J. Tenenbaum. Learning, and learning to learn, with hierarchical bayesian models. In *Frontiers in Systems Neuroscience. Conference Abstract: Computational and systems neuroscience.*, 2009.

[114] L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

[115] D. C. van Essen. Organization of visual areas in macaque and human cerebral cortex. In L. M. Chalupa and J. S. Werner, editors, *The Visual Neurosciences, volume 2*, pages 507–521. The MIT Press, 2004.

[116] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, 265:359–366, 1998.

[117] V. N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, Heidelberg, DE, 1995.

[118] BT Vincent, RJ Baddeley, T Troscianko, and ID Gilchrist. Is the early visual system optimised to be energy efficient? *Network*, 16(2-3):175–190, 2005.

[119] H. von Helmholtz and A. König. *Handbuch der physiologischen Optik.* L. Voss, Leipzig, 1896.

[120] H. Wassle. Parallel processing in the mammalian retina. *Nat Rev Neurosci*, 5(10):747–757, 2004.

[121] Y. Weiss and W. T. Freeman. What makes a good model of natural images? In *Proc. CVPR 2007, Minneapolis*, 2007.

[122] B. Willmore and D. J. Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12:255–270, 2001.

[123] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, April 2002.

[124] O. Woodford, I. Reid, P.H.S. Torr, and A.W. Fitzgibbon. Fields of experts for image-based rendering. *Proceedings British Machine Vision Conference*, 2006.

[125] Z. Zalevsky and D. Mendlovic. *Optical Superresolution.* Springer, 2003.

[126] C. Zetzsche and G. Krieger. Nonlinear neurons and higher-order statistics: new approaches to human vision and electronic image processing. In B. Rogowitz and T.V. Pappas, editors, *Human Vision and Electronic Imaging IV*, pages 2–23. 1999.

[127] S. C. Zhu, Y. N. Wu, and D. Mumford. FRAME: Filters, random field and maximum entropy – towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):1–20, 1998.

TIETOJENKÄSITTELYTIETEEN LAITOS
PL 68 (Gustaf Hällströmin katu 2 b)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Gustaf Hällströmin katu 2 b)
FIN-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports may be ordered from: Kumpula Science Library, P.O. Box 64, FIN-00014 University of Helsinki, FINLAND.

A-2001-1 J. Rousu: Efficient range partitioning in classification learning. 68+74 pp. (Ph.D. Thesis)

A-2001-2 M. Salmenkivi: Computational methods for intensity models. 145 pp. (Ph.D. Thesis)

A-2001-3 K. Fredriksson: Rotation invariant template matching. 138 pp. (Ph.D. Thesis)

A-2002-1 A.-P. Tuovinen: Object-oriented engineering of visual languages. 185 pp. (Ph.D. Thesis)

A-2002-2 V. Ollikainen: Simulation techniques for disease gene localization in isolated populations. 149+5 pp. (Ph.D. Thesis)

A-2002-3 J. Vilo: Discovery from biosequences. 149 pp. (Ph.D. Thesis)

A-2003-1 J. Lindström: Optimistic concurrency control methods for real-time database systems. 111 pp. (Ph.D. Thesis)

A-2003-2 H. Helin: Supporting nomadic agent-based applications in the FIPA agent architecture. 200+17 pp. (Ph.D. Thesis)

A-2003-3 S. Campadello: Middleware infrastructure for distributed mobile applications. 164 pp. (Ph.D. Thesis)

A-2003-4 J. Taina: Design and analysis of a distributed database architecture for IN/GSM data. 130 pp. (Ph.D. Thesis)

A-2003-5 J. Kurhila: Considering individual differences in computer-supported special and elementary education. 135 pp. (Ph.D. Thesis)

A-2003-6 V. Mäkinen: Parameterized approximate string matching and local-similarity-based point-pattern matching. 144 pp. (Ph.D. Thesis)

A-2003-7 M. Luukkainen: A process algebraic reduction strategy for automata theoretic verification of untimed and timed concurrent systems. 141 pp. (Ph.D. Thesis)

A-2003-8 J. Manner: Provision of quality of service in IP-based mobile access networks. 191 pp. (Ph.D. Thesis)

A-2004-1 M. Koivisto: Sum-product algorithms for the analysis of genetic risks. 155 pp. (Ph.D. Thesis)

A-2004-2 A. Gurtov: Efficient data transport in wireless overlay networks. 141 pp. (Ph.D. Thesis)

A-2004-3 K. Vasko: Computational methods and models for paleoecology. 176 pp. (Ph.D. Thesis)

A-2004-4 P. Sevon: Algorithms for Association-Based Gene Mapping. 101 pp. (Ph.D. Thesis)

A-2004-5 J. Viljamaa: Applying Formal Concept Analysis to Extract Framework Reuse Interface Specifications from Source Code. 206 pp. (Ph.D. Thesis)

A-2004-6 J. Ravantti: Computational Methods for Reconstructing Macromolecular Complexes from Cryo-Electron Microscopy Images. 100 pp. (Ph.D. Thesis)

A-2004-7 M. Kääriäinen: Learning Small Trees and Graphs that Generalize. 45+49 pp. (Ph.D. Thesis)

A-2004-8 T. Kivioja: Computational Tools for a Novel Transcriptional Profiling Method. 98 pp. (Ph.D. Thesis)

A-2004-9 H. Tamm: On Minimality and Size Reduction of One-Tape and Multitape Finite Automata. 80 pp. (Ph.D. Thesis)

A-2005-1 T. Mielikäinen: Summarization Techniques for Pattern Collections in Data Mining. 201 pp. (Ph.D. Thesis)

A-2005-2 A. Doucet: Advanced Document Description, a Sequential Approach. 161 pp. (Ph.D. Thesis)

A-2006-1 A. Viljamaa: Specifying Reuse Interfaces for Task-Oriented Framework Specialization. 285 pp. (Ph.D. Thesis)

A-2006-2 S. Tarkoma: Efficient Content-based Routing, Mobility-aware Topologies, and Temporal Subspace Matching. 198 pp. (Ph.D. Thesis)

A-2006-3 M. Lehtonen: Indexing Heterogeneous XML for Full-Text Search. 185+3 pp. (Ph.D. Thesis)

A-2006-4 A. Rantanen: Algorithms for ^{13}C Metabolic Flux Analysis. 92+73 pp. (Ph.D. Thesis)

A-2006-5 E. Terzi: Problems and Algorithms for Sequence Segmentations. 141 pp. (Ph.D. Thesis)

A-2007-1 P. Sarolahti: TCP Performance in Heterogeneous Wireless Networks. (Ph.D. Thesis)

A-2007-2 M. Raento: Exploring privacy for ubiquitous computing: Tools, methods and experiments. (Ph.D. Thesis)

A-2007-3 L. Aunimo: Methods for Answer Extraction in Textual Question Answering. 127+18 pp. (Ph.D. Thesis)

A-2007-4 T. Roos: Statistical and Information-Theoretic Methods for Data Analysis. 82+75 pp. (Ph.D. Thesis)

A-2007-5 S. Leggio: A Decentralized Session Management Framework for Heterogeneous Ad-Hoc and Fixed Networks. 230 pp. (Ph.D. Thesis)

A-2007-6 O. Riva: Middleware for Mobile Sensing Applications in Urban Environments. 195 pp. (Ph.D. Thesis)

A-2007-7 K. Palin: Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements. 130 pp. (Ph.D. Thesis)

A-2008-1 I. Autio: Modeling Efficient Classification as a Process of Confidence Assessment and Delegation. 212 pp. (Ph.D. Thesis)

A-2008-2 J. Kangasharju: XML Messaging for Mobile Devices. 24+255 pp. (Ph.D. Thesis).

A-2008-3 N. Haiminen: Mining Sequential Data – in Search of Segmental Structures. 60+78 pp. (Ph.D. Thesis)

A-2008-4 J. Korhonen: IP Mobility in Wireless Operator Networks. (Ph.D. Thesis)

A-2008-5 J.T. Lindgren: Learning nonlinear visual processing from natural images. 100+64 pp. (Ph.D. Thesis)

A-2009-1 K. Hätönen: Data mining for telecommunications network log analysis. 153 pp. (Ph.D. Thesis)

A-2009-2 T. Silander: The Most Probable Bayesian Network and Beyond. (Ph.D. Thesis)

A-2009-3 K. Laasonen: Mining Cell Transition Data. 148 pp. (Ph.D. Thesis)

A-2009-4 P. Miettinen: Matrix Decomposition Methods for Data Mining: Computational Complexity and Algorithms. 164+6 pp. (Ph.D. Thesis)

A-2009-5 J. Suomela: Optimisation Problems in Wireless Sensor Networks: Local Algorithms and Local Graphs. 106+96 pp. (Ph.D. Thesis)