Ma 117

ACTA POLYTECHNICA SCANDINAVICA

MATHEMATICS AND COMPUTING SERIES No. 117

Probabilistic Models of Early Vision

Patrik O. Hoyer

Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 9800
FIN-02015 HUT, Finland

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 15th of November, 2002, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science

ESPOO 2002

Hoyer, P. O., **Probabilistic Models of Early Vision**. Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 117, Espoo 2002, 65 pp. Published by the Finnish Academies of Technology. ISBN 951-666-613-2. ISSN 1456-9418.

Keywords: Natural images, independent component analysis, latent variable models, unsupervised learning, neural networks, early vision, visual cortex.

ABSTRACT

How do our brains transform patterns of light striking the retina into useful knowledge about objects and events of the external world? Thanks to intense research into the mechanisms of vision, much is now known about this process. However, we do not yet have anything close to a complete picture, and many questions remain unanswered. In addition to its clinical relevance and purely academic significance, research on vision is important because a thorough understanding of biological vision would probably help solve many major problems in computer vision.

A major framework for investigating the computational basis of vision is what might be called the probabilistic view of vision. This approach emphasizes the general importance of uncertainty and probabilities in perception and, in particular, suggests that perception is tightly linked to the statistical structure of the natural environment. This thesis investigates this link by building statistical models of natural images, and relating these to what is known of the information processing performed by the early stages of the primate visual system.

Recently, it was suggested that the response properties of simple cells in the primary visual cortex could be interpreted as the result of the cells performing an independent component analysis of the natural visual sensory input. This thesis provides some further support for that proposal, and, more importantly, extends the theory to also account for complex cell properties and the columnar organization of the primary visual cortex. Finally, the application of these methods to predicting neural response properties further along the visual pathway is considered.

Although the models considered account for only a relatively small part of known facts concerning early visual information processing, it is nonetheless a rather impressive amount considering the simplicity of the models. This is encouraging, and suggests that many of the intricacies of visual information processing might be understood using fairly simple probabilistic models of natural sensory input.

Preface

This thesis is the result of work carried out at the Neural Networks Research Centre of the Laboratory of Computer and Information Science at Helsinki University of Technology. The main funding was provided by the Helsinki Graduate School of Computer Science and Engineering. I am also grateful to the Jenny and Antti Wihuri Foundation and the Finnish Foundation for the Promotion of Technology for additional financial support.

First and foremost, I am deeply indebted to my mentor, collaborator, and friend, Dr. Aapo Hyvärinen. Not only has he given me daily guidance and encouragement, as well as sceptical criticism when I have needed it, he has in fact taught me by example how to carry out scientific research. Thank you, Aapo.

I would also like to thank my supervisor, Academy Professor Erkki Oja, who has constantly encouraged me ever since I began my work in the lab, back in 1997. Together with Professor Olli Simula and the founder of the research centre, Academician Teuvo Kohonen, they have managed to create a research centre which is not only a successful and productive laboratory, but also a truly enjoyable place to work!

Furthermore, I want to thank the members of our research group, Mr. Jarmo Hurri, Mr. Mika Inki, and Mr. Jaakko Väyrynen, for insightful comments and support. I am also grateful to Dr. Harri Valpola, who provided useful observations and assisted me on several occasions, and to Professor Kai Kaila, for supporting my neurobiology studies. I would also like to express my sincere gratitude to Dr. Mikko Lehtokangas and Dr. Pentti Laurinen for reviewing my thesis and providing constructive criticism.

During the course of my studies, I have had the pleasure to interact with numerous distinguished scientists who have provided valuable feedback. I could not possibly list them all here, but would like to mention at least Professor Bruno Olshausen and Professor Eero Simoncelli, whose perceptive comments have significantly affected my work.

For the relaxed atmosphere, stimulating discussions, and for conceding to have lunch as early as eleven o'clock (albeit grudgingly), I wish to thank past and present members of the lab with whom I have had the joy and pleasure of spending my weekdays! For our interesting Monday seminars on Meritullinkatu, I thank all participants, Dr. Laurinen in particular. And for helping me remember that life is actually so much more than just work, I am deeply grateful to my non-scientist friends, with whom I have had so many memorable moments, I love you guys!

Finally, I want to express my sincere gratitude to my sister and my parents. Not only have they encouraged and supported me throughout my studies, but they also originally helped open my eyes to the wonderful world of science, something I will never forget!

Helsinki, November 2002

Patrik Hoyer

Contents

Notation

Generally, bold uppercase letters (e.g. $\mathbf{W}$, $\mathbf{A}$) denote matrices, bold lowercase letters (e.g. $\mathbf{x}$, $\mathbf{s}$, $\mathbf{a}_j$) denote (column) vectors, whereas scalars are displayed in italics (e.g. a_{ij}, s_j).

a_{ij}	Element $\{ij\}$ of matrix $\mathbf{A}$ (element i of vector $\mathbf{a}_j$), see below
$\mathbf{a}_j$	j:th basis vector in sparse coding/ICA model
$\mathbf{A}$	ICA *mixing matrix*, containing the vectors $\mathbf{a}_1 \dots \mathbf{a}_n$ as columns
$\mathbf{A}^{-1}$	Matrix inverse of $\mathbf{A}$
$f_j(\mathbf{x})$	Scalar function of its vector input $\mathbf{x}$
$\mathbf{f}(\mathbf{x})$	Vector with components $f_j(\mathbf{x})$
$I(x, y)$	Image intensity (luminance) as a function of spatial position (x, y)
m	Number of elements in $\mathbf{x}$
n	Number of elements in $\mathbf{s}$
$p(\mathbf{x})$	Probability density of sensory input vector $\mathbf{x}$
$p(\mathbf{x}, \mathbf{s})$	Joint probability density of $\mathbf{x}$ and $\mathbf{s}$
$p(\mathbf{s}\|\mathbf{x})$	Probability density of $\mathbf{s}$ given that we have observed a specific $\mathbf{x}$
r_j	Predicted firing rate of neuron j
σ	Noise level in the generative model
s_j	Element j of the vector $\mathbf{s}$, see below
$\mathbf{s}$	Model response to the input. Specifically: - Responses of model neurons to input image (chapter 5) - Latent variable vector (chapter 6) - Vector of latent variables, whose optimal estimates (upon observing an input $\mathbf{x}$) model neural responses to that input (chapters 7-10)
u_i	Element i of vector $\mathbf{u}$, see below
$\mathbf{u}$	Vector of latent variables, whose estimates model complex cell responses
$w(x, y)$	Receptive field weights as a function of spatial position (x, y)
$\mathbf{w}_j$	Weight vector for neuron j
$\mathbf{W}$	Weight matrix, having $\mathbf{w}_1^T \dots \mathbf{w}_n^T$ as its rows
(x, y)	Image spatial coordinates, not to be confused with x_i or $\mathbf{x}$, explained below
x_i	Element i of the vector $\mathbf{x}$, see below
$\mathbf{x}$	Observed input data vector. Specifically: - Image (or image patch) represented as a vector (chapters 4–5 and 7–8) - Observed data vector in latent variable models (chapter 6) - Vector of model LGN cell responses to images (chapter 9) - Vector of model complex-cell responses to images (chapter 10)

Publications of the thesis

List of publications

1. P. O. Hoyer and A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Computation in Neural Systems*, vol. 11, no. 3, pp. 191–210, 2000.

2. A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

3. 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*, vol. 41, no. 18, pp. 2413–2423, 2001.

4. P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing 2002, Martigny, Switzerland)*, pp. 557–565, 2002.

5. P. O. Hoyer, "Modeling receptive fields with non-negative sparse coding," in *Computational Neuroscience: Trends in Research 2003*, Elsevier, Amsterdam, 2003. In press.

6. P. O. Hoyer and A. Hyvärinen, "A multi-layer sparse coding network learns contour coding from natural images," *Vision Research*, vol. 42, no. 12, pp. 1593–1605, 2002.

Contents of the publications and contributions of the author

In **Publication 1**, independent component analysis (ICA) features were calculated from chromatic and binocular natural image data. It was shown that also in these cases the learned features resemble the receptive fields of simple cells in the primary visual cortex (area V1), yielding further support for the proposal that simple cells in V1 are adapted to provide a maximally independent linear representation of the natural visual sensory input. The current author suggested and performed the experiments, and wrote the paper. Dr. Hyvärinen provided invaluable insight during the whole process and also co-edited the manuscript.

In **Publication 2**, the ICA model was extended so that independent and sparse subspaces were sought rather than scalar components. This model was applied to natural image input,

and it was shown that the norms of projections onto the subspaces resembled V1 complex cell responses. The present author contributed all the experiments, and participated in the editing of the paper.

In **Publication 3**, the model developed in Publication 2 was further extended to model not only V1 simple and complex cell receptive fields, but also V1 topography. This was done by considering a *local pooling* of simple cells into complex cells, and then maximizing the sparseness of the complex cells. All in all, this model provides a very parsimonious model of V1 organization that seems to capture some aspects of ordering that most other models fail to account for. The current author contributed all the experiments and participated in the editing of the paper. The idea of extending the complex cell model of Publication 2 into a topographic map was jointly developed by the authors.

In **Publication 4**, the basic ideas behind ICA/sparse coding and Non-negative Matrix Factorization were combined by introducing non-negativity constraints in the sparse coding model. An algorithm was developed to find the optimal hidden components, given the basis. It was further shown how also the basis could be learned from data.

In **Publication 5**, arguments were presented in favor of non-negativity constraints when modeling visual receptive fields. The technique presented in Publication 4 was then used to learn receptive field profiles from natural images prefiltered to model actual V1 input.

In **Publication 6**, the principle of non-negative sparse coding was applied to the responses of model complex cells. The goal was to predict neural response properties further along the visual pathway. A decomposition was found where contours are represented by units selectively tuned to contour length. Neurons like this (so-called end-stopped neurons) have indeed been observed, but the representation at this level is not yet well understood. Hopefully, this (and future) work can help unravel the mysteries. The present author suggested and performed the experiments, and wrote the paper. Dr. Hyvärinen co-edited the manuscript and provided numerous acute and perceptive comments and abundant constructive criticism.

Purpose and intended audience

Before we begin, it is important to make clear the intended audience and the purpose of this thesis. The thesis consists of two separate parts, which serve quite different audiences.

The six published research articles that form the core of the thesis were written for fellow researchers in this exciting field of study, with the hope of contributing to and advancing the field. They assume a fair degree of familiarity with previous work on natural image statistics and vision, and might be a difficult read for those with little previous knowledge of such research.

The 'introductory part', however, serves very different purposes. First, as part of a doctoral thesis, it is directed to my department, to the pre-examiners of the thesis, and to the appointed opponent of its public defence, with the purpose of fulfilling the requirements for the doctoral degree. Second, it is written for me: it serves to organize and clarify the thoughts and views that have been developing in my mind during the last few years; many of those thoughts have only indirectly found expression in the published articles. Finally, it is written for those with little or no previous exposure to either visual neuroscience or natural image statistics (or both!), who wish to find out what this research really is all about.

It might very well be impossible to completely realize all those purposes simultaneously. Certainly, I do not pretend to have succeeded. Rather it is simply my hope that there be a little bit for everybody. For example, to accommodate readers completely new to the field, chapters 2–4 provide a brief but hopefully quite accessible introduction to all this. If this material feels extremely basic to you, feel free to skip it. On the other hand, I am well aware of the fact that the exposition in chapters 5–7 probably is too brief to adequately explain the concepts involved to someone with no previous familiarity with them. Hopefully, these chapters will nevertheless give even such readers a general idea of these topical issues in the research on vision.

Chapter 1

Introduction

This thesis is about *vision*. Although most people have a strong intuitive idea of what vision is from their own personal experience, it will be useful to define our use of the word more precisely. For our purposes, vision is to be understood as *the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect* [113]. Note that this definition does not specify what kind of a system is implementing this process. In fact, we are interested in both biological and computer vision.

At first glance, it might seem that biological visual systems would have little in common with computers. For one thing, the implementing 'hardware' is very different: the biochemical structures of brains are quite unlike the physical components of a computer of this day. The two also differ radically on a functional level: It is well known that the operating principles of current computers are quite unlike those that seem to operate in the biological brain. So why not keep biological and electronic systems separate?

Indeed, much research on vision *is* separate. Biological vision scientists have not turned into computer science theorists, and a great deal of computer vision research bears little resemblance to anything observed in the brain. However, a growing amount of research is being conducted that is neither purely biology nor purely computer science, but rather something in-between. This is the direct result of the modern approach that sees vision as a computational task separate from the particular medium implementing it [94].

Although the computational approach to vision is already well established, this thesis concerns a related approach that is not perhaps quite as widely accepted in the field of vision science. This is the probabilistic view of vision, which treats vision as an estimation problem [12, 73, 74, 75, 122, 124, 166]: Because vision is severely under-constrained, with many possible world interpretations for any given image, it is in theory impossible to know which interpretation is the correct one. The best we can do is to estimate their probabilities; this is very much in the spirit of the *unconscious inference* suggested by Helmholtz [159] over a century ago. Although the probabilistic framework has a long history, it is currently experiencing a surge of interest from vision theorists. Perhaps the main reason for this is that the theoretical tools and the computational power needed to investigate it have only very recently become available.

The probabilistic framework not only gives the general goal of the computational problem of vision, but in fact points to the general importance of uncertainty and probabilities in perception. In particular, it suggests that the statistical structure of the natural environment is of crucial importance to perception [6, 7, 8, 12, 73, 138, 153].

This thesis attempts to model the information processing of the early stages of the primate visual system using statistical models of the visual input in a natural environment. The goal of this research is to gain a better understanding of how vision is accomplished in the brain. Such an understanding has both clinical importance and academic significance, and in addition would quite probably lead to significant progress in the field of computer vision. This thesis is organized as follows: chapters 2–6 give a thorough introduction to this research, while chapters 7–10 describe the particular models considered and the main results of this thesis. In chapter 11 we end with some conclusions and consider the future of this research field.

Chapter 2

The computational task of vision

2.1 The starting point of vision

In the introduction we defined vision as the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect. The first thing we will need to consider is in what form this information initially is available.

The light emitted and reflected by objects has to be collected and then measured before any information can be extracted from it. Both biological and artificial systems typically perform the first step by focusing light to form a two-dimensional *image* by perspective projection. Although there of course are countless differences between the eye and any camera, the image formation process is essentially the same. When the image has been formed the intensity of the light is measured. In the human eye this is performed by the photoreceptors, whereas artificial systems employ a variety of technologies. However, all systems share the fundamental idea of converting the optical image into some kind of signal that represents the intensity of the light at each point in the image.

Although in general the projected images have both temporal and chromatic dimensions, we will be mostly concerned with static, monochrome (gray-scale) images. Such an image can be defined as a scalar function over two dimensions, $I(x, y)$, giving the intensity (luminance) value at every location (x, y) in the image. Although in the general case both quantities (the position (x, y) and the intensity $I(x, y)$) take continuous values, we will focus on the case where the image has been sampled at discrete points in space. This means that in our discussion x and y take only integer values, and the image can be fully described by an array containing the intensity values at each sample point.[1] Note that although the spatial sampling performed by biological and artificial systems often is not rectangular or even regular, this does not change the fact that on a qualitative level the sampling is quite similar.

It is from this kind of image data that vision extracts information. Information about the physical environment is contained in such images, but only *implicitly*. The visual system must somehow transform this implicit information into an explicit form. This is not a simple problem, as the demonstration of the next section attempts to illustrate.

[1] When images are stored on computers, the entries in the arrays also have to be discretized; this is, however, of less importance in the discussion that follows, and we will assume that this has been done at a high enough resolution so that this step can be ignored.

2.2 The magic of your visual system

Vision is an exceptionally difficult computational task. Although this is clear to vision scientists, it might come as a surprise to others. The reason for this is that we are equipped with a truly amazing visual system that performs the task effortlessly and quite reliably in our daily environment. We are simply not aware of the whole computational process going on in our brains, rather we experience only the result of that computation.

To illustrate the difficulties in vision, figure 2.1 displays an image in its numerical format (as described in the previous section), where light intensities have been measured and are shown as a function of spatial location. In other words, if you were to colour each square with the shade of gray corresponding to the contained number you would see the image in the form we are used to, and it would be easily interpretable. Without looking at the solution just yet, take a minute and try to decipher what the image portrays. You will probably find this extremely difficult.

Now, have a look at the solution (figure 2.3; please note that figure 2.2 was skipped here). It is immediately clear what the image represents! Our visual system performs the task of recognizing the image completely effortlessly. Even though the image at the level of our photoreceptors is represented essentially in the format of figure 2.1, our visual system somehow manages to make sense of all this data and figure out the real-world object that caused the image.

In the discussion thus far, we have made a number of drastic simplifications. Among other things, the human retina contains photoreceptors with varying sensitivity to the different wavelengths of light, and we typically view the world through two eyes, not one. Finally, perhaps the most important difference is that we normally perceive dynamic images rather than static ones. Nonetheless, these differences do not change the fact that the optical information is, at the level of photoreceptors, represented in a format analogous to that we showed in figure 2.1, and that the task of the visual system is to understand all this data.

2.3 The difficulty of vision

Most people would agree that this task initially seems amazingly hard. But after a moment of thought it might seem reasonable to think that perhaps the problem is not so difficult after all? Image intensity edges can be detected by finding oriented segments where small numbers border with large numbers. The detection of such features can be computationally formalized and straightforwardly implemented [94]. Perhaps such oriented segments can be grouped together and subsequently object form be analyzed? Indeed, such computations can be done, and they form the basis of many computer vision algorithms [141]. However, although current computer vision systems work fairly well on synthetic images or on images from highly restricted environments, they still perform quite poorly on images from an unrestricted, natural environment. In fact, perhaps one of the main findings of computer vision research to date has been that the analysis of real-world images is extremely difficult [141]! Even such a basic task as identifying the contours of an object is complicated because often there is no clear image contour along some part of its physical contour, as illustrated in figure 2.2.

In light of the difficulties computer vision research has run into, the computational accomplishment of our own visual system seems all the more amazing. We perceive our environment quite accurately almost all the time, and only relatively rarely make perceptual

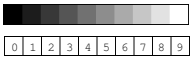

Figure 2.1: An image displayed in numerical format. The shade of gray of each square has been replaced by the corresponding numerical intensity value. What does this mystery image depict?

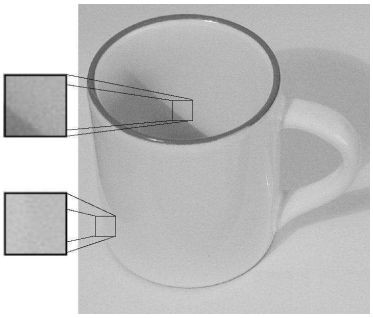

Figure 2.2: This image of a cup demonstrates that physical contours and image contours are often very different. The physical edge of the cup near the lower-left corner of the image yields practically no image contour (as shown by the magnification). On the other hand, the shadow casts a clear image contour where there in fact is no physical edge.

mistakes. Quite clearly, biology has solved the task of everyday vision in a way that is completely superior to any present-day machine vision system.

This being the case, it is natural that computer vision scientists have tried to draw inspiration from biology. Many systems contain image preprocessing and feature extraction steps that mimic the processing that is known to occur in the early parts of the biological visual system. However, beyond the very early stages, little is actually known about the representations used in the brain. Thus, there is actually not much to guide computer vision research at the present.

On the other hand, it is quite clear that good computational theories of vision would be useful in guiding research on biological vision, by allowing hypothesis-driven experiments. So it seems that there is a dilemma: computational theory is needed to guide experimental research, and the results of experiments are needed to guide theoretical investigations. The solution, as we see it, is to seek synergy by multidisciplinary research into the computational basis of vision.

Figure 2.3: The image of figure 2.1. It is immediately clear that the image shows a male face. Many observers will probably even recognize the specific individual (note that it might help to view the image from relatively far away).

Chapter 3

A primate visual system primer

To understand how the computational models proposed in this thesis relate to biology, one must be at least partly familiar with our present knowledge of the biological visual system. This section is intended to give *the absolutely minimal* facts to someone who lacks this familiarity, so that he or she can follow the discussion that follows. Nothing except what is absolutely necessary for the purposes of this thesis is presented, so it is by no means intended to be anything near a comprehensive account of what is currently known. Those with even the slightest background in visual neuroscience are probably better off skipping this section.

3.1 *Which* biological visual system?

You may have noticed (and possibly been annoyed by the fact) that we have this far often talked of *the* biological visual system without actually specifying exactly what is meant. *Which* biological visual system? We have in some cases hinted that we mean the human visual system, and it is indeed the one we are truly interested in. A great deal is known about our visual system from psychophysical experiments (for review, see e.g. [113, 126]), patients with focal brain damage [27, 77, 165], and recently from advanced brain imaging methods [46, 162].

Nevertheless, much of what we know about biological vision actually concern the visual systems of other species. Electrophysiological experiments (combined with anatomical data) on other primates have been particularly revealing (see e.g. [17, 44, 54, 105, 146, 152, 157, 169]). Fortunately, current evidence seems to suggest that the results of these experiments also shed light on human vision. This evidence comes from anatomical similarities as well as parallels between human psychophysics (and brain imaging) and animal physiology, see e.g. [15, 91, 162]. Because it seems highly probable that the basic elements of biological vision considered in this thesis are comparable among human and non-human primates, we do not differentiate between the two. Rather, we talk simply of the primate visual system.[1]

[1] Many of the earliest electrophysiological results concerning the mammalian visual cortex were obtained from cats (e.g. [52]). However, most of the findings relevant for this thesis have subsequently been shown to be valid for primates as well.

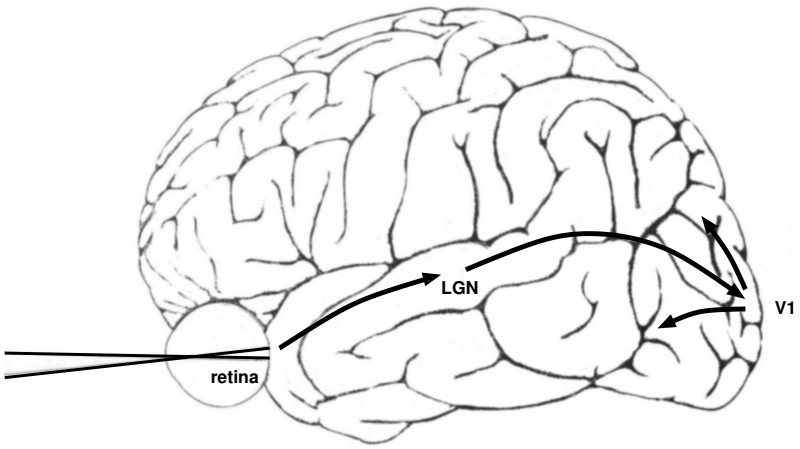

Figure 3.1: The main visual pathway in primates. See main text for details. (Adapted from [50].)

3.2 The main visual pathway

Figure 3.1 illustrates the earliest stages of the main visual pathway. Light is detected by the photoreceptors in the retinas, and the ultimate output of the retinas is sent by the retinal ganglion cells through the optic nerve. These axons eventually synapse in the lateral geniculate nucleus (LGN) of the thalamus. LGN cells subsequently send their axons to the primary visual cortex (area V1) at the very back of the brain. This is the first point where visual signals are processed by the cerebral cortex. From there, the information is sent to the surrounding extrastriate cortex in several pathways for further processing.

It must, however, be stressed that this account is a drastic simplification of biological facts. First, within this main visual pathway, there are actually multiple parallel information channels that carry different aspects of visual information to the cortex. Second, there is a massive feedback projection from the primary visual cortex to the thalamus. Unfortunately, the computational role of this feedback pathway is still a mystery. (For a textbook account of these and other well-known facts about the visual pathway, see, e.g. [68].)

It is interesting to note that a sizable part of the brain is in fact devoted to vision. In macaque monkeys, for example, it has been estimated that approximately half of the neocortex is concerned with this task [152]. It seems likely that the fraction is somewhat smaller for humans, but this does not change the fact that an enormous amount of machinery is dedicated to vision. This further reinforces our sense of the complexity of the problem of vision.

3.3 Neural receptive fields

The main information processing workload of the brain is carried by neurons [68]. The majority of neurons communicate by action potentials (also called *spikes*), stereotyped electrical impulses traveling down the axons of neurons. Although in principle information could be carried by very complex patterns of spikes [147], in practice most research to date

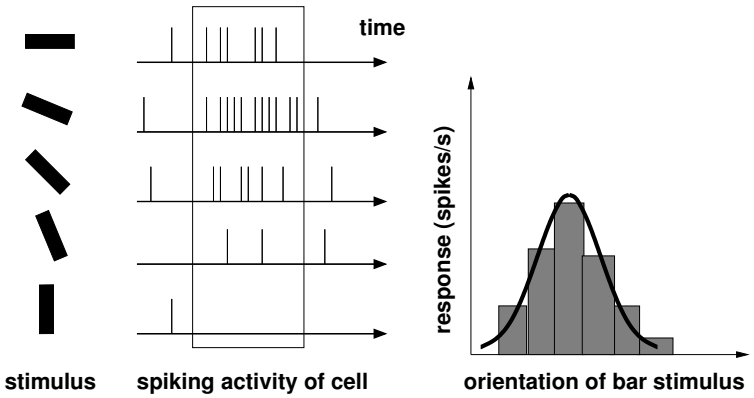

Figure 3.2: A caricature of a typical experiment. A dark bar on a white background is flashed onto the screen, and action potentials are recorded from a neuron. Varying the orientation of the bar yields varying responses. Counting the number of spikes elicited within a fixed time window following the stimulus, and plotting these counts as a function of bar orientation, one can construct a mathematical model of the response of the neuron.

has focused on the neurons' *firing rates*, the number of spikes fired by the neurons within some suitably defined time window. These firing rates are thought to reflect the general level of activity of the cells [2, 52].

Thus, much of visual neuroscience has been concerned with measuring the firing rates of cells as a function of some properties of a visual stimulus. For example, an experiment might run as follows: An image is suddenly projected onto a (previously blank) screen that an animal is watching, and the number of spikes fired by some recorded cell in the next second are counted. By systematically changing some properties of the stimulus and monitoring the elicited response, one can make a quantitative model of the response of the neuron. Such a model mathematically describes the response (firing rate) r_j of a neuron as a function of the stimulus $I(x, y)$, as in

$$r_j = f_j(I(x, y)). \tag{3.1}$$

This is illustrated in figure 3.2.

In the early visual system, the response of a typical neuron depends only on the intensity pattern of a very small part of the visual field. This area, where light increments or decrements can elicit above-baseline firing rates, is called the *classical receptive field* of the neuron. More generally, the concept also refers to the particular light pattern that yields the maximum response [52, 79].

So, what light patterns actually elicit the strongest responses? This of course varies from neuron to neuron. The retinal ganglion cells as well as cells in the LGN typically have circular center-surround receptive field structure [51, 79]: Some neurons are excited by light in a small circular area of the visual field, but inhibited by light in a surrounding annulus. Other cells show the opposite effect, responding maximally to light that fills the surround but not the center. This is depicted in figure 3.3a. Cells in V1 have more interesting receptive fields. The so-called *simple cells* typically have adjacent elongated (instead of concentric circular) regions of excitation and inhibition [54]. This means that these cells

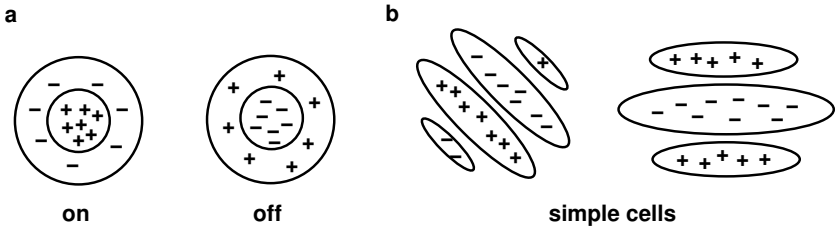

Figure 3.3: Typical classical receptive fields of neurons early in the visual pathway. Plus signs denote regions of the visual field where light causes excitation, minuses regions where light inhibits responses. (a) Retinal ganglion and LGN neurons typically exhibit center-surround receptive fields organization, in one of two arrangements. (b) The majority of simple cells in V1, on the other hand, have oriented receptive fields.

respond maximally to *oriented* image structure. This is illustrated in figure 3.3b.

All of the above classical receptive fields can be modeled by a linear model: the response of a neuron can be reasonably predicted by a weighted sum of the image intensities, as in

$$r_j = \sum_{(x,y)} w_j(x,y) I(x,y), \tag{3.2}$$

where $w_j(x,y)$ contains the pattern of excitation and inhibition for light for the neuron j in question. Typically, center-surround receptive fields are modeled as the difference of two circular Gaussian kernels with different widths (difference-of-gaussians model [130]), whereas oriented receptive fields are most often fitted by Gabor functions (products of sinusoidal gratings and Gaussian envelopes) [28, 32, 65, 93, 129]; however, see also [142]. It goes without saying that these linear models are drastic simplifications of the actual neural firing dynamics of these cells, but they are nevertheless useful starting points on which more detailed models can be built.

Although these linear models are useful in modeling many cells, there are also neurons in V1 called *complex cells* for which these models are completely inadequate [54]. These cells do not show any clear spatial zones of excitation or inhibition. Complex cells respond, just like simple cells, selectively to bars and edges at a particular location and of a particular orientation; they are, however, relatively invariant to the spatial phase of the stimulus. An example of this is that reversing the contrast polarity of the stimulus does not markedly alter the response of a typical complex cell. The responses of complex cells have often been modeled by the classical 'energy model'[2] [1, 98, 121], in which

$$r_j = \left(\sum_{(x,y)} w_{j_1}(x,y) I(x,y) \right)^2 + \left(\sum_{(x,y)} w_{j_2}(x,y) I(x,y) \right)^2, \tag{3.3}$$

where $w_{j_1}(x,y)$ and $w_{j_2}(x,y)$ are quadrature-phase Gabor functions, see figure 3.4.

The idea that V1 complex cells pool the responses of simple cells (as opposed to constructing their response properties directly from the LGN afferents) is here attractive; such an anatomical arrangement was originally suggested by Hubel and Wiesel [54]. There is

[2]The term 'energy' simply denotes the squaring operation.

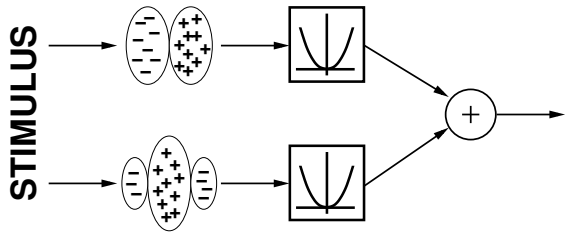

Figure 3.4: The classic energy model for complex cells. The response of a complex cell is modeled by linearly filtering with quadrature-phase Gabor filters (Gabor functions whose sinusoidal components have a 90 degrees phase difference), taking squares, and summing. Note that this is purely a mathematical description of the response and *should not* be directly interpreted as a hierarchical model summing simple cell responses.

evidence both for and against this proposition, however, so it is not yet clear if this is indeed the case (see [3, 95]).

3.4 Topographic organization

It is interesting to consider how the receptive fields of neighboring cells are related. In the retina, the receptive fields of retinal ganglion cells are necessarily linked to the physical position of the cells. This is due to the fact that the visual field is mapped in an orderly fashion to the retina. Thus, neighboring retinal ganglion cells respond to neighboring areas of the visual field. However, there is nothing to guarantee the existence of a similar organization further up the visual pathway.

But the fact of the matter is that, just like in the retina, neighboring neurons in the LGN and in V1 tend to have receptive fields covering neighboring areas of the visual field. Yet this is only one of several types of organization. In V1, the preferred orientation of receptive fields also tends to shift gradually along the surface of the cortex [52]. In fact, neurons are often approximately organized according to several functional parameters simultaneously [52]. This kind of topographic organization also exists in higher visual areas, such as inferotemporal cortex [146].

Topographical representations are not restricted to areas devoted to vision, but are in fact present in various forms throughout the brain (for review, see [100]). Examples include the tonotopic map (frequency-based organization) in the primary auditory cortex and the complete body map for the sense of touch. In fact, one might be pressed to find a brain area that would not exhibit any sort of topography. Even for areas that might seem to fit the bill, it may well be that the underlying organization has just not yet been understood.

Chapter 4

Structure in natural images

4.1 The image state space

Recall how in chapter 2 we described an image representation in which each image is represented as a numerical array containing the intensity values of its picture elements, or *pixels*. To make the following discussion concrete, say that we are dealing with images of a fixed size of 256-by-256 pixels. This gives a total of $65536 = 256^2$ pixels in an image. Each image can then be considered as a point, call it $\mathbf{x}$, in a 65536-dimensional state space, each axis of which specifies the intensity value of one pixel [37]. Conversely, each point in the state space specifies one particular image. This state space concept is illustrated in figure 4.1.

Next, consider taking an enormous set of images, and plotting each as the corresponding point in our state space. (Of course, plotting a 65536-dimensional space is not very easy to do on a two-dimensional page, so we will have to be content with making a thought experiment.) An important question is: how would the points be distributed in this space? In other words, what is the probability density $p(\mathbf{x})$ of our images like? The answer, of course, depends on the set of images chosen. Astronomical images have very different properties from holiday snapshots, for example, and the two sets would yield very different clouds of points in our state space.

4.2 Defining natural images

In this thesis we will be specifically concerned with a particular set of images called *natural images* or images of *natural scenes*. Some images from our data set are shown in figure 4.2. This set is supposed to resemble the natural input of the visual system we are investigating. So what is meant by 'natural input'? This is actually not a trivial question at all. The underlying assumption in this line of research (as explained in chapters 5–6) is that biological visual systems are, through a complex combination of the effects of evolution and development, adapted to process the kind of sensory input that they receive. Natural images is thus some set that we believe has similar statistical structure to that which the visual system is adapted to.

This poses an obvious problem, at least in the case of human vision. The human visual system has evolved in an environment that is in many ways different from the one most of us experience daily today. It is probably quite safe to say that images of skyscrapers, cars,

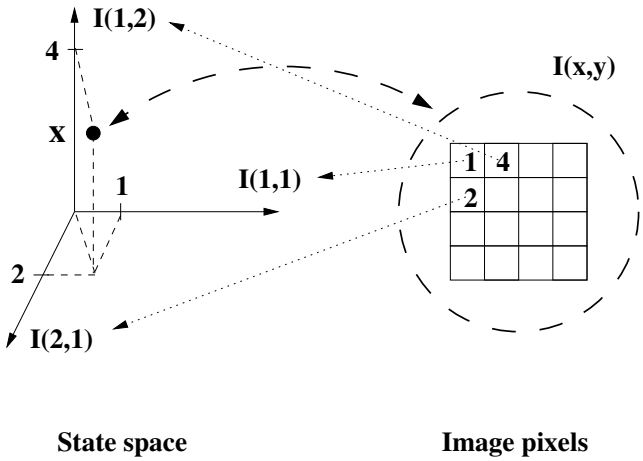

Figure 4.1: The state space representation of images. Images are mapped to points in the space in a one-to-one fashion. Each axis of the state space corresponds to the brightness value of one specific pixel in the image.

Figure 4.2: Three representative examples from our set of natural images.

and other modern entities have not affected our genetic makeup to any significant degree. On the other hand, few people today experience nature as omnipresent as it was tens of thousands of years ago. Thus, the input on the time-scale of evolution has been somewhat different from that on the time-scale of the individual. Should we then choose images of nature or images from a modern, urban environment to model the 'natural input' of our visual system? Most work to date (for review, see [138]) has focused on the former, and this has also been our choice. However, it should by no means be taken for granted that this is the only, 'correct' choice.

Returning to our original question, how would *natural images* be distributed in the image state space? The important thing to note is that they would not be anything like uniformly distributed in this space. It is easy for us to draw images from a uniform distribution, and they do not look anything like our natural images! Figure 4.3 shows three images randomly drawn from a uniform distribution over the image space. As there is no question that we can easily distinguish these images from natural images (figure 4.2) it follows that these are drawn from separate, very different, distributions. In fact, the distribution of natural images is highly non-uniform. This is the same as saying that natural images contain a lot of *redundancy*, an information theoretic term that we turn to now.

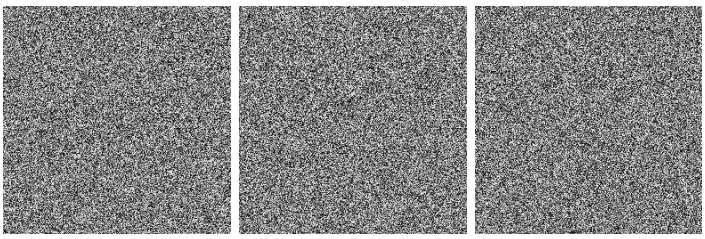

Figure 4.3: Three images drawn randomly from a uniform distribution in the image state space. Each pixel is drawn independently from a uniform distribution from black to white.

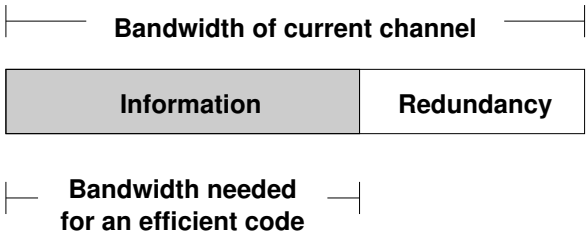

Figure 4.4: Redundancy in a signal. Some of the bandwidth carrying a typical signal is normally 'wasted' because of redundancy (structure) in the signal. If the signal is optimally compressed before transmission, stripping it of all redundancy, it can be transmitted using much less resources. (Note that redundancy is often useful in a noisy channel where it can be used to detect or repair errors, but we only consider noise-free channels here.)

4.3 Redundancy of natural images

The development by Claude Shannon [132] of the theory of information (for a textbook account, see [25]) is one of the true milestones of science. Shannon considered the transmission of a message across a communication channel and developed a mathematical theory that quantified the variables involved: the amount of information transmitted and the capacity of a channel to carry information. Because of its generality the theory has found, and continues to find, an endless number of applications in a variety of disciplines.

One of the key ideas in information theory is that the amount of information carried by a signal is often less than the maximum amount that could be transmitted by the communication channel (also known as the 'bandwidth'). This is because some of the capacity is essentially consumed by structure in the signal. The more rigid the structure, the less room there is to provide the receiver with information. Thus, the contents of any signal can essentially be divided into information and redundancy. This is depicted in figure 4.4.

To make this more concrete, consider the binary image of figure 4.5. The image contains a total of $32 \times 22 = 704$ pixels. Thus, the standard representation (where the color of each pixel is indicated by a '1' or a '0') for this image requires 704 bits. But it is not difficult to imagine that one could compress it into a much smaller number of bits. For example, one could invent a representation that assumes a white background on which black squares (with given positions and sizes) are printed. In such a representation, our image could be coded by simply specifying the top-left corners of the squares ((5,5) and (19,11)) and their

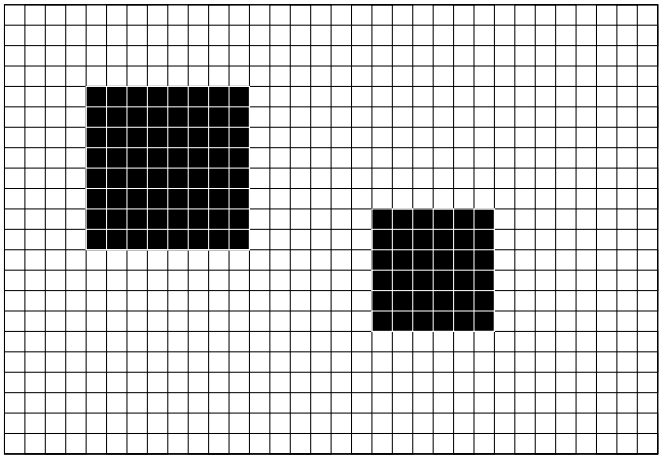

Figure 4.5: A binary image containing a lot of structure. Images like this can be coded efficiently; see main text for discussion.

sizes (8 and 6). This could certainly be coded in less than 704 bits.[1]

The important thing to understand is that this kind of representation is good for certain kinds of images (those with a small number of black squares) but not others (that do not have this structure and thus require a huge amount of squares to be completely represented). Hence, if we are dealing mostly with images of the former kind, and we are using the standard binary coding format, then our representation is highly redundant. By compressing it using our black-squares-on-white representation we achieve an efficient representation. Although natural images are much more variable than this hypothetical class of images, it is nonetheless true that they also show structure and can be compressed.

Attneave [7] was the first to explicitly point out the redundancy in images. The above argument is essentially the same as originally given by Attneave, although he considered a 'guessing game' in which subjects guessed the colour of pixels in the image. The fact that subjects perform much better than chance proves that the image is predictable, and information theory ensures that predictability is essentially the same thing as redundancy [7].

As will be discussed in the next two chapters, making use of this redundancy of images is essential for vision. But the same statistical structure is in fact also crucial for many other tasks involving images. Engineers who seek to find compact digital image formats for storing or transmitting images also need to understand this structure [42, 85, 135]. Image synthesis and noise reduction are other tasks that optimally would make use of this structure [42, 56, 85, 135, 136, 137, 168]. Thus, the analysis of the statistical properties of images has widespread applications indeed, although perhaps understanding vision is the most profound.

[1] The specification of each square requires three numbers which each could be coded in 5 bits, giving a total of 30 bits for two squares. Additionally, a few bits might be needed to indicate how many squares are coded, assuming that we do not know *a priori* that there are exactly two squares.

Chapter 5

Redundancy reduction and efficient coding

5.1 Redundancy reduction

Following its conception, it did not take long before psychologists and biologists understood that information theory was directly relevant to the tasks of biological systems. Indeed, the sensory input is a signal that carries *information* about the outside world. This information is *communicated* by sensory neurons by means of action potentials.

In his original article [7] describing the redundancy inherent in images, Attneave suggested that the visual system recodes the inputs to *reduce redundancy*, providing an 'economical description' of the sensory signals. He likened the task of the visual system to that of an engineer who seeks to represent pictures with the smallest possible number of bits. It is easy to see the intuitive appeal of this idea. Consider again the image of figure 4.5. Recoding images of this kind using our black-squares-on-white representation, we reduce redundancy and obtain an efficient representation. However, at the same time we have *discovered the structure* in the signal: we now have the concept of 'squares' which did not exist in the original representation. More generally: to reduce redundancy one must first identify it. Thus, redundancy reduction *requires* discovering structure.

Although he was arguably the first to spell it out explicitly, Attneave was certainly not the only one to have this idea. Around the same time, Barlow [8] provided similar arguments from a more biological/physiological viewpoint. Barlow has also pointed out [11, 12] that the idea, in the form of 'economy of thought', is clearly expressed already in the writings of Mach [92] and Pearson [116]. Nevertheless, with the writings of Attneave and Barlow, the *redundancy reduction* (or *efficient coding*) *hypothesis* was born.

5.2 Testing the efficient coding hypothesis

Is the efficient coding principle actually implemented in the visual system somehow? How could we test the hypothesis? A recent excellent review of research in this area can be found in [138]. Here we will only briefly describe the part of this work which is most relevant to this thesis.

There are two main approaches to testing the efficient coding hypothesis. One possibil-

ity is to record from neurons at various levels of the visual system and try to directly estimate the information versus redundancy in their patterns of action potentials [16, 128, 143]. One of the key results of this line of inquiry is that many neurons seem to be highly efficient at communicating information (see e.g. [16, 127]), giving some support to the hypothesis.

The other approach, which will be the main focus of this chapter, is to consider the statistics of the typical sensory input and then 'derive' a model for efficient coding of this input. Then, the model is compared to known physiology of the early visual system. If the match is good, this supports the idea that the brain is implementing efficient coding.

To be able to proceed, we must somehow limit the models we will consider. We will assume that our stimuli consist of static images, and that our model neurons respond to these images with scalar outputs denoting their firing rates. This is the model that was introduced in chapter 3, see equation (3.1). Note that this completely ignores all temporal aspects of both stimuli and responses, and also abstracts the individual discrete action potentials into a continuous firing rate. Finally, there is no variability in this model, so the same response is always elicited by a given stimulus. Despite these strong simplifications, this model has been quite useful for understanding neural responses and will also serve our discussion well.

Denoting the sensory inputs (pixel intensity values of our images) x_i and the responses of neuron j by s_j, the response model is $s_j = f_j(x_1, \ldots, x_m)$. This can be written in vector form as

$$\mathbf{s} = \mathbf{f}(\mathbf{x}). \tag{5.1}$$

Given the statistics of the sensory input, $p(\mathbf{x})$, we will seek a response model $\mathbf{f}$ such that the coding scheme adheres to the efficient coding principle.

The basic efficient coding hypothesis states that sensory neurons should be adapted to transmit the maximum amount of information about the natural environment, given limited resources. This has also been called *infomax*, information maximization [13, 89, 103]. The limitations can be, for example, a fixed number of neurons and some fixed mean or maximum firing rates of these neurons. Alternatively, the hypothesis assumes that neurons minimize the resources needed to transmit a fixed amount of information.

Using information theory, it is straightforward to show (e.g. [13, 103]) that in our response model, under certain assumptions,[1] the mutual information between the sensory input and the neural response is maximized when the entropy of the response is at a maximum. For a single neuron, the response distribution that maximizes the entropy depends on the specific constraints applied. As an example, if the range of the response is bounded, the response distribution with maximal entropy is the uniform distribution [80]. However, the more interesting result is that in the case of many neurons, maximal response entropy requires that the responses of the neurons are statistically independent [13, 103].

Assuming that our input $\mathbf{x}$ consists of natural image data, can we find a transformation $\mathbf{f}$ that would give responses s_j which are mutually independent? In the general case, this is a rather complicated problem because a mapping that produces completely independent responses can be quite complex. Thus, researchers have mainly focused on the special case where the mapping is constrained to be linear, as in

$$\mathbf{s} = \mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{x}. \tag{5.2}$$

[1] In particular, it is assumed that the nonlinear transfer functions f_j are bounded and invertible, and that there is infinitesimal additive output noise.

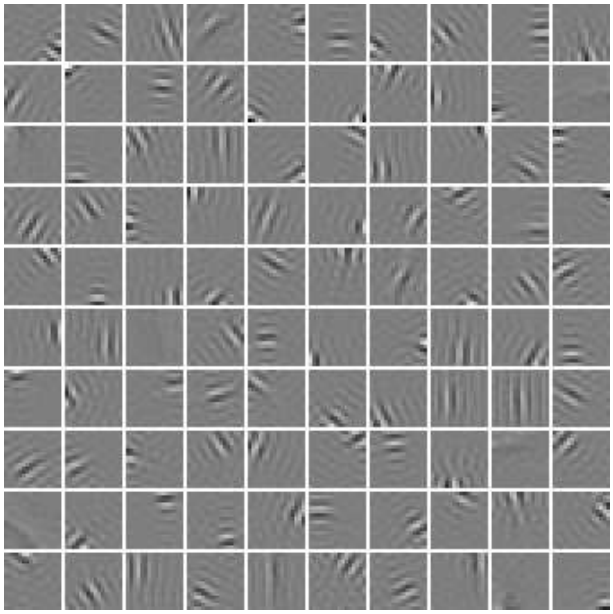

Figure 5.1: Linear filters derived by ICA from natural image patches. Each patch displays one filter $\mathbf{w}_j$ (i.e. $w_j(x, y)$); the filters having been arbitrarily arranged into a 10×10 array. These are the filters that minimize the statistical dependencies between the responses when applied to natural images. Bright pixels denote excitatory, dark pixels inhibitory, responses to light. Compare with figure 3.3b.

In other words, each response s_j is given by a linear sum of the image intensities,

$$s_j = \mathbf{w}_j^T \mathbf{x} = \sum_{(x,y)} w_j(x, y) I(x, y), \tag{5.3}$$

where $\mathbf{w}_j^T$ is the j:th row of the matrix $\mathbf{W}$. Above, we also emphasized that the vectors $\mathbf{w}_j$ and $\mathbf{x}$ exactly correspond to the receptive field weighting profile (see chapter 3) and the input image (see chapter 4), respectively. The question then is: which mapping $\mathbf{W}$ (set of filters $w_j(x, y)$) produces responses that are *as independent as possible* in the natural environment? Several algorithms have recently been developed that attempt to perform this task, called Independent Component Analysis (ICA) [4, 13, 19, 22, 23, 55, 59, 67, 71, 83, 108]; for a textbook account of this technique the reader is referred to [58]. Applied to data consisting of small patches of natural images, such algorithms result in linear filters (see figure 5.1) that resemble the receptive fields of simple cells in the primary visual cortex [14, 155].

The similarity of simple cell receptive fields to the the optimal linear filters suggests that the neurons are adapted to provide approximately independent responses to natural sensory input. But even the optimal filters do not give *completely* independent responses, due to the linearity constraint on the mapping (see Publication 2 and [139, 167]). This is in fact fortunate, since one can then hope to account for further properties of the visual system using the dependencies that remain.

One might thus consider specific forms of nonlinearities that might increase indepen-

dence. For example, Simoncelli has, together with his colleagues, shown that a nonlinearity known as 'contrast gain control' [17, 40, 48] acting on the linear outputs significantly reduces dependencies between units [131, 139]. This gives further evidence that simple cells provide a maximally independent code of the visual input, given some constraints.

How far can this approach be taken? If the mapping $\mathbf{f}$ was not constrained in any way, one could eventually achieve complete independence. But what would this accomplish? Would this constitute some kind of 'ultimate solution' to vision? We would argue that it would not, and that simply seeking a transform that produces independent components cannot be the whole story. In fact, without any constraints on the mapping there are an infinite number of solutions $\mathbf{f}$ that give independent components [60]. Although indeterminancy also exist in the linear case, the indeterminancy in the nonlinear case is much worse since the solutions are not trivially related [60].

Suppose for just a second that we do not mind the indeterminancy described above. (After all, this only means that there are a lot of solutions so one might think that it implies that it would ease the task of finding one!) The important question is: what have we actually accomplished? We have effectively managed to solve the *density estimation problem*, and we are in a position to specify $p(\mathbf{x})$ for any $\mathbf{x}$. But how does this solve vision? In the following chapter we will provide an alternative way of thinking about the connection between image statistics and vision that might provide a more useful framework.

Chapter 6

The latent variable model approach

6.1 Analysis-by-synthesis

The traditional computational approach to vision [94] focuses on how, from the image data $\mathbf{x}$, one can compute quantities of interest, which we will call $\mathbf{s}$. These quantities might be, for instance, scalar variables such as the distances to objects, or binary parameters such as signifying if an object belongs to some given categories.[1] In other words, the emphasis is on a function $\mathbf{f}$ that transforms images into world or object information, as in $\mathbf{s} = \mathbf{f}(\mathbf{x})$. This operation might be called image *analysis*.

Several researchers (see e.g. [43, 74, 101]) have pointed out that the opposite operation, image *synthesis*, more often than not is much simpler. That is, the mapping $\mathbf{g}$ that generates the image given the state of the world ($\mathbf{x} = \mathbf{g}(\mathbf{s})$), is considerably easier to work with than the mapping $\mathbf{f}$. But how does knowing $\mathbf{g}$ help us, one may ask. The answer is that one may then search for the parameters $\hat{\mathbf{s}}$ that produce an image $\hat{\mathbf{x}} = \mathbf{g}(\hat{\mathbf{s}})$ which, as well as possible, matches the observed image $\mathbf{x}$. Under reasonable assumptions, this might lead to a good approximation of the correct parameters $\mathbf{s}$. This approach to vision is known as *analysis-by-synthesis*.

To make all this concrete, consider again the image of the cup in figure 2.2. The traditional approach of vision would propose that an early stage extracts local edge information in the image, after which some sort of grouping of these edge pieces would be done. Finally, the evoked edge pattern would be compared with patterns in memory, and recognized as a cup. Meanwhile, analysis of other scene variables, such as lighting direction or scene depth, would proceed in parallel. The analysis-by-synthesis framework, on the other hand, would suggest that our visual system has an unconscious internal model for image generation. Estimates of object identity, lighting direction, and scene depth are all adjusted until a satisfactory match between the observed image and the internally generated image is achieved.

[1] One point about notation: The motivation for re-using $\mathbf{s}$ here comes from the way it is used in the next chapters and thus related to the neural firing rates that it designated in chapter 5.

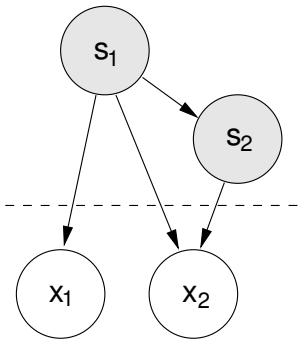

Figure 6.1: A simple latent variable model. The nodes define the observed (x_1 and x_2) as well as the hidden (s_1 and s_2) variables, and the arrows indicate the dependencies among these variables. To complete the specification of the model, one needs to define (1) the allowed states of the variables, (2) the prior probabilities of all nodes lacking parents (i.e. $p(s_1)$), and (3) the conditional probabilities of all other nodes (i.e. $p(s_2|s_1)$, $p(x_1|s_1)$, and $p(x_2|s_1, s_2)$). The dashed line and the shading simply illustrates the division into observed and hidden nodes.

6.2 Latent variable models

In the probabilistic approach to vision, analysis-by-synthesis is naturally implemented in the framework of *latent variable models*. Latent variable models attempt to explain observed data by some underlying hidden causes or factors that we have only indirect information about. The models we shall consider here consist of a set of observed random variables which we will call x_i, a set of latent random variables s_j, and a set of prior and conditional probabilities that define how the states of the variables are dependent on each other. These models are also known as *Bayesian networks* or *belief networks* [63, 66, 115]. A simple example of such a model is given in figure 6.1.

Such a latent variable model defines a joint probability density over all variables, $p(\mathbf{x}, \mathbf{s})$. However, only the x_i are ever observed. Why then introduce hidden variables? The reason is that quite complex patterns in the observed data (dependencies between the x_i) can often be understood as resulting from relatively simple underlying 'causes' that cannot, however, be directly observed. Radford Neal has suggested (on his WWW page) an analogy from medicine: Doctors observe only the various symptoms of patients, and have invented 'diseases' (hidden variables) to explain the various combinations of symptoms.[2] Mathematically, latent variable models approximate the density $p(\mathbf{x})$ of the data as the observed part of the joint density $p(\mathbf{x}, \mathbf{s})$. This is obtained by integrating the joint density over all hidden states, as in

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{s})\, d\mathbf{s}. \tag{6.1}$$

There are two fundamental operations in such networks, *inference* and *learning*. Inference refers to estimating the hidden variables $\mathbf{s}$ given an observed data vector $\mathbf{x}$. In most

[2] Although one might argue that one can actually observe the diseases directly by the use of various tests, one can equally well view these tests are just part of our observations, and the 'diseases' as our explanations for the combination of observed symptoms and test results.

models it is impossible (even in theory) to know the precise values of $\mathbf{s}$, so one must be content with a probability density $p(\mathbf{s}|\mathbf{x})$ [29]. By Bayes rule, this is given as

$$p(\mathbf{s}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{x})}. \tag{6.2}$$

To obtain a point estimate of the hidden variables, many models simply opt to find the particular $\mathbf{s}$ which maximize this density,

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} p(\mathbf{s}|\mathbf{x}). \tag{6.3}$$

The connection to analysis-by-synthesis should now be clear: the latent variable model specifies how data vectors are synthesised (generated) from the 'causes' $\mathbf{s}$ by $p(\mathbf{x}|\mathbf{s})$. The network infers these causes by selecting them so that the fit to the observed data is maximized. However, note the additional factor $p(\mathbf{s})$ providing a bias for causes assumed to occur more frequently than others. This fits common sense: if two world configurations are equally adequate explanations of a given image, the more frequent configuration is the more probable cause.

The second fundamental operation in latent variable models is the estimation, or learning, of the model from observed data. Typically, the dependency *structure* in the model is fixed, but the specific dependency strengths may vary. The dependencies are parametrized, so that a given parameter vector, call it θ, yields a specific joint density $p_\theta(\mathbf{x}, \mathbf{s})$. Now, how can we adapt the model's parameters so that it best fits the observed data? The typical choice is to attempt to fit $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{s})\, d\mathbf{s}$ to the observed density $p(\mathbf{x})$ [66].

6.3 Internal models

The remainder of this thesis will focus on different latent variable models for images. These can be viewed as some kinds of *internal models* [10, 26, 122] of the visual input. In this framework the goal is, utilizing the dependencies inherent in the sensory environment, to learn models that reflect the structure of the world [30, 49, 86, 111, 122, 123, 151].

It must be emphasised that although the models employed can have many free parameters to be learned, some overall structure or form of the dependencies must be fixed. Thus we must use our intuition and experience when specifying the models to be considered. For example, the model structure might be chosen to mimic the way real objects interact to produce images. As we seek to model brain-like computation, another possibility is to try to match the model structure to plausible neural circuitry patterns. In this thesis, the latter has been the dominant heuristic.

Finally, it must be underlined that the models discussed in this thesis are quite low-level models of images. They do not contain variables that would specify the identity of objects or the direction of lighting in a scene. Rather, the models concern low-level features of images, such as bars and edges, and at best elongated contours. This means that no practical object recognition is done by these models. However, the learned models can directly be compared with representations in the early biological visual system, of which much is known. This allows us to investigate if there is anything to the claim that the brain builds an internal model of the world, utilizing the redundancy in the sensory input.

Chapter 7

Independent component analysis based on a latent variable model

7.1 Sparse coding

In a seminal paper published in 1996, Olshausen and Field [110] described how a simple neural network performing *sparse coding* [9, 37, 39, 148] learned features that were qualitatively very similar to the receptive fields of V1 simple cells. This was significant because it was the first study to show how all the basic spatial properties of simple cell classical receptive fields (localization in space and in spatial frequency, and orientation tuning) could emerge in an unsupervised manner from natural images. Olshausen and Field showed that regardless of the mechanisms by which simple cell receptive fields develop, they can be interpreted as providing a maximally sparse code of the natural sensory input.

The basic idea of sparse coding is relatively simple. Each input image patch $\mathbf{x}$ is represented by a linear combination of some set of basis patches $\mathbf{a}_j$, as in

$$\mathbf{x} \approx s_1 \cdot \mathbf{a}_1 + s_2 \cdot \mathbf{a}_2 + \cdots + s_n \cdot \mathbf{a}_n. \tag{7.1}$$

This is illustrated in figure 7.1. Varying the coefficients s_j allows different input vectors $\mathbf{x}$ to be accurately represented. In sparse coding, the goal is to select the set of basis patches so that typical input vectors $\mathbf{x}$ can (by proper choices of the s_j) be represented *accurately* and *sparsely*. The first objective essentially means that the right-hand side of equation 7.1 should closely equal the left-hand side. The second objective favors representations where only a few coefficients s_j are significantly active (non-zero) for any given input $\mathbf{x}$. All basis

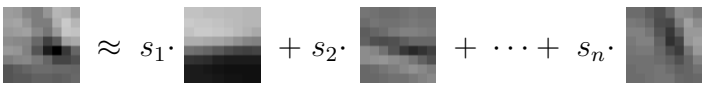

Figure 7.1: The linear image synthesis model. Each input patch $\mathbf{x}$ (corresponding to $I(x, y)$) is represented as a linear combination of basis (feature) patches $\mathbf{a}_j$ (i.e. $a_j(x, y)$). In sparse coding, one attempts to learn basis patches such that in this representation the coefficients s_j are as sparse as possible, meaning that for most image patches only a few of them are significantly non-zero.

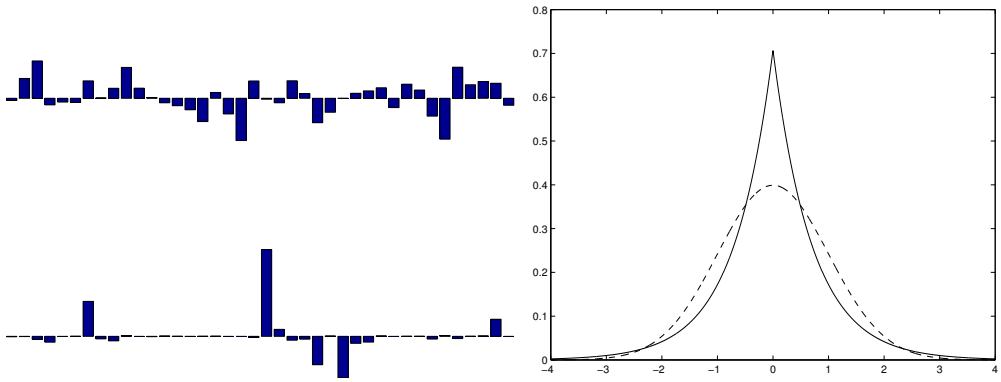

Figure 7.2: Illustration of sparse distributions. Left: samples from a Gaussian random variable (top) versus samples from a sparse random variable (bottom). Both have the same variance (sum of squares) but in the sparse variable the energy is concentrated into only a few samples while the rest are effectively zero. Right: The probability density of a Gaussian variable (dashed line) versus that of a sparser variable (solid line). Both are normalized to have the same variance, but the density of the sparser variable exhibits a higher peak at zero and heavier (enhanced) tails. In other words, the peak at zero of the sparser density is exactly compensated by thicker tails to give the same overall scale (variance) as that of the Gaussian density.

patches $\mathbf{a}_j$ are needed, however, because the set of active coefficients changes from input to input.

If all units s_j are assumed to be equally active in the long run (i.e. when considering the whole set of inputs $\mathbf{x}$), this kind of 'population' sparseness is essentially the same as 'lifetime' sparsess, where any given unit is approximately zero for most inputs and significantly active only rarely [164]. Such a sparse response distribution is also called supergaussian [59] (or leptokurtotic [72]), and is illustrated in figure 7.2.

There are two main tasks in sparse coding. The first one is, given a specific input vector $\mathbf{x}$, to find the optimal values of the coefficients s_j. This is done by defining an objective function that specifies how the goals of accuracy and sparseness are measured and how these measures are combined. On the presentation of an input $\mathbf{x}$, the s_j are chosen which optimize this objective. Note that for typical objective functions the optimal values of the coefficients are a *nonlinear* function of the input, and iterative methods are required to find a local optimum [21, 111].

The second task in sparse coding is to find a set of basis patches $\mathbf{a}_j$ which allows typical input vectors $\mathbf{x}$ to be accurately and sparsely represented. Just as the linear transformation matrix $\mathbf{W}$ giving maximally independent components (see chapter 5) depends on the probability density $p(\mathbf{x})$ of the data, so too does the optimal set of basis patches $\mathbf{a}_j$ in sparse coding. Thus it makes sense to look for the optimal set for natural image input $\mathbf{x}$. This is just what Olshausen and Field [110, 111] did. The result was a set quite similar to that shown in figure 7.3. The resemblance of the features to V1 simple cell receptive fields were then taken as evidence for sparse coding in the primary visual cortex.

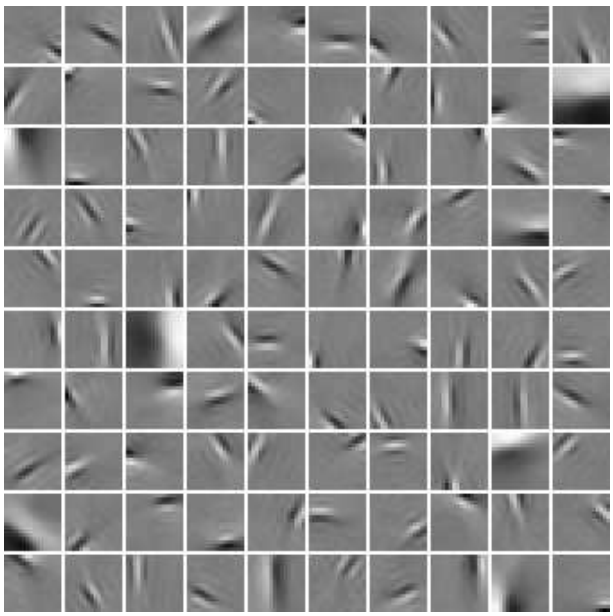

Figure 7.3: A set of basis patches $\mathbf{a}_j$ ($a_j(x, y)$) learned from natural image patches. Each basis patch contains the contribution of one s_j to the input $\mathbf{x}$ ($I(x, y)$), as in figure 7.1. (Note that there is no particular order among the basis patches and that the arrangement into a 10×10 array is completely arbitrary. Note also the resemblance to figure 5.1; see main text for details.)

7.2 The ICA model

It soon became clear [47, 109, 111] how learning in the sparse coding network (developed independently by Olshausen and Field [110] and Harpur and Prager [47]) could be interpreted as the approximate estimation of a latent variable model. This model is graphically shown in figure 7.4. Both the hidden variables s_j and the observed data x_i are continuous random variables. The hidden variables are independent with supergaussian prior densities $p(s_j)$, whereas the observed variables are drawn from Gaussian distributions with constant variance, but whose means are specified by linear combinations of the s_j. A typical choice for the hidden unit prior density is the laplacian (double-sided exponential). Mathematically, we may write

$$p(s_j) = \exp(-\sqrt{2}|s_j|)/\sqrt{2} \tag{7.2}$$

$$p(x_i|\mathbf{s}) = \frac{1}{N_\sigma} \exp\left[\frac{-(x_i - \sum_j s_j a_{ij})^2}{2\sigma^2}\right], \tag{7.3}$$

where σ is the 'noise level' in the model, and N_σ is the normalizing constant of the Gaussian density.

As discussed in chapter 6, two fundamental operations in such a latent variable model are inference and learning. Inference consists of estimating the hidden variables s_j that generated a given data vector $\mathbf{x}$, while learning is the estimation of the model parameters

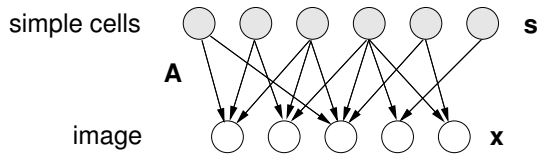

Figure 7.4: The generative ICA model. All nodes are continuous random variables. The hidden variables s_j are generated independently and exhibit sparseness (see figure 7.2). The observed variables x_i are drawn from Gaussian distributions with means given by linear combinations of the latent variables s_j. The weights of the linear combinations are given by the entries a_{ij} of a matrix $\mathbf{A}$.

a_{ij} from a large set of data vectors. It turns out that the objective for finding the most likely hidden unit activations s_j in the above model is exactly the same as the one previously proposed for selecting the optimal sparse coding coefficients s_j [47, 109, 111]. In addition, the algorithm for learning the sparse coding basis patches can be interpreted as performing approximate estimation of the parameters of the latent variable model [109, 111].

Furthermore, it was shown [18, 109, 111] how the ICA algorithm of Bell and Sejnowski [13] could be derived from maximum-likelihood estimation of the above model, for the special case of infinitesimal noise ($\sigma \to 0$) and an equal number of hidden and observed variables. This special case has received a lot of attention, primarily because of its simplicity. The key property is that in this case the posterior density $p(\mathbf{s}|\mathbf{x})$ collapses onto a *single point* which is a *linear* function of the data, $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}$, allowing simple and efficient algorithms for estimating the model.

Hence, both the original sparse coding network and ICA algorithms can be interpreted as estimating the latent variable model given in figure 7.4. In fact, the understanding that has emerged is that redundancy reduction by ICA, sparse coding, and estimation of the generative model are in some sense all just different theoretical approaches leading to essentially the same method.

7.3 Modeling simple cell responses

In figure 7.3, we showed a set of basis patches $\mathbf{a}_j$ learned from data consisting of natural image patches. These basis patches were learned using the FastICA algorithm [55, 59] but, as discussed in the previous section, can also be considered 'sparse coding' basis patches. Each patch contains the model weights $\{a_{ij}\}_{i=1\ldots m}$ that determine the contribution of one hidden unit s_j to the data. Note the similarity to figure 5.1. In fact, the ICA *separating matrix* $\mathbf{W}$ is just the pseudo-inverse of the ICA *mixing matrix* $\mathbf{A}$. For image data, the separating filters $w_j(x, y)$ shown in figure 5.1) are essentially just high-pass filtered versions of the basis patches $a_j(x, y)$ (see discussion in Publication 3).

While the features learned by ICA correspond to simple cell receptive fields, it is the latent variables s_j that represent the firing rates of the neurons. The special case of no noise and $n = m$ makes this particularly clear. As discussed in the previous section, in that case there is no uncertainty in the values of the s_j after observing $\mathbf{x}$, and they can be obtained as $s_j = \mathbf{w}_j^T\mathbf{x} = \sum_{(x,y)} w_j(x, y)I(x, y)$. This is precisely the linear response model of firing rates, introduced in chapter 3. In the presence of noise, or when the number of units n is greater than the dimensionality of the input m, one cannot know precisely

the values s_j that generated a particular input $\mathbf{x}$. But one can still find an estimate, by taking the values that maximize the posterior $p(\mathbf{s}|\mathbf{x})$. The components of this *maximum a posteriori* estimate shall serve to represent the firing rates in the general case [111].

Following the establishment of a *qualitative* correspondence between simple cell responses and the ICA representation, van Hateren and van der Schaaf [155] showed that even *quantitatively* the match was fairly good. They calculated various measures of the properties of the features learned by ICA from natural images, and compared these with previously published data on simple cell receptive fields, finding a good match for most of the parameters.

The question then arose if the ICA model could account for further simple cell properties. In addition to the purely spatial response characteristics described in chapter 3, much is also known about how simple cells respond to spatiotemporal, chromatic, and binocular stimuli. Briefly, spatiotemporal receptive fields can be classified into space-time separable and inseparable groups, with mainly inseparable ones exhibiting directional selectivity [32, 33]. As for responses to chromatic stimuli, many researchers [24, 90, 149, 150] have found that simple cells can be grouped into three main groups that prefer achromatic, red/green, and blue/yellow stimuli, respectively. However, the cortical coding of colour is still not adequately understood, and many investigators [34, 64, 84] have reported results that, at least at first sight, seem incompatible with this picture. More generally, it is believed that cells responding preferentially to chromatic stimuli are not as tuned to orientation as achromatic ones. Finally, the binocular characteristics of receptive fields are also relatively well understood: Cells display varying degrees of binocularity and exhibit interocularly matched preferred orientations and spatial frequencies [52, 140], but have both interocular phase and position differences [5].

To investigate whether the spatiotemporal structure of cortical receptive fields could be accounted for by ICA, van Hateren and Ruderman [154] applied ICA to natural image sequences. The features of the ICA decomposition strongly resembled simple cell spatiotemporal receptive fields, providing further support for the proposed model.

In 1999, we set out to investigate the properties of ICA features learned from colour and stereo images, and to compare them with known chromatic and binocular aspects of simple cell receptive fields. (Several other researchers have recently performed similar investigations, applying ICA to feature extraction from colour images, see [145, 160].) Figure 7.5 depicts the features learned from colour images. As discussed in Publication 1, our results fit many aspects of what is known of chromatic receptive fields. Indeed, the ICA features can be neatly grouped into three distinct chromatic groups, with the achromatic units essentially identical to those learned from monochrome images.

The ICA features learned from stereo image data are shown in figure 7.6. Also in this case, the match between the properties of the ICA features and those of simple cell receptive fields is impressive, as discussed in Publication 1. The features exhibit a range of degrees of binocularity, and further show interocularly matched frequency and orientation, but have prominent interocular phase differences.

In summary, then, it seems that ICA applied to natural images yields features which are in a multitude of ways similar to V1 simple cell receptive fields. This is particularly striking because of the conceptual simplicity of the model; it only assumes sparse, independent hidden variables and linear mixing. No prior information whatsoever on the weights a_{ij} is incorporated. Thus ICA applied to natural images provides a highly parsimonious model of the earliest cortical processing of visual input in the brain. Recently, even physiological evidence for sparse and independent representations has been presented [156].

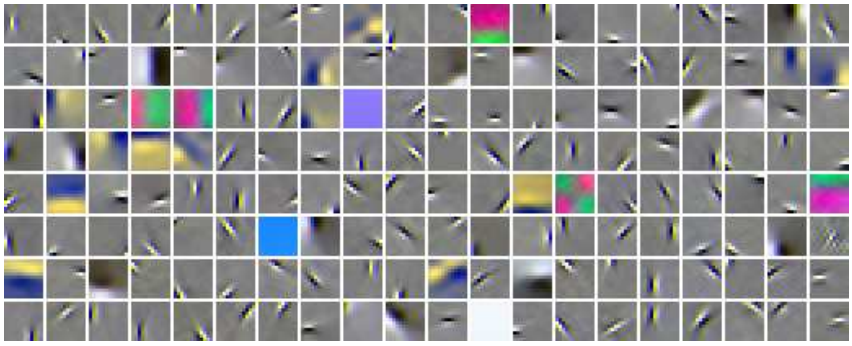

Figure 7.5: ICA basis of colour images. Again, each patch represents the contribution of one latent variable in the generative model. Note how the representation is split into separate red/green, blue/yellow, and bright/dark features, with the last group essentially identical to the features learned from gray-scale images.

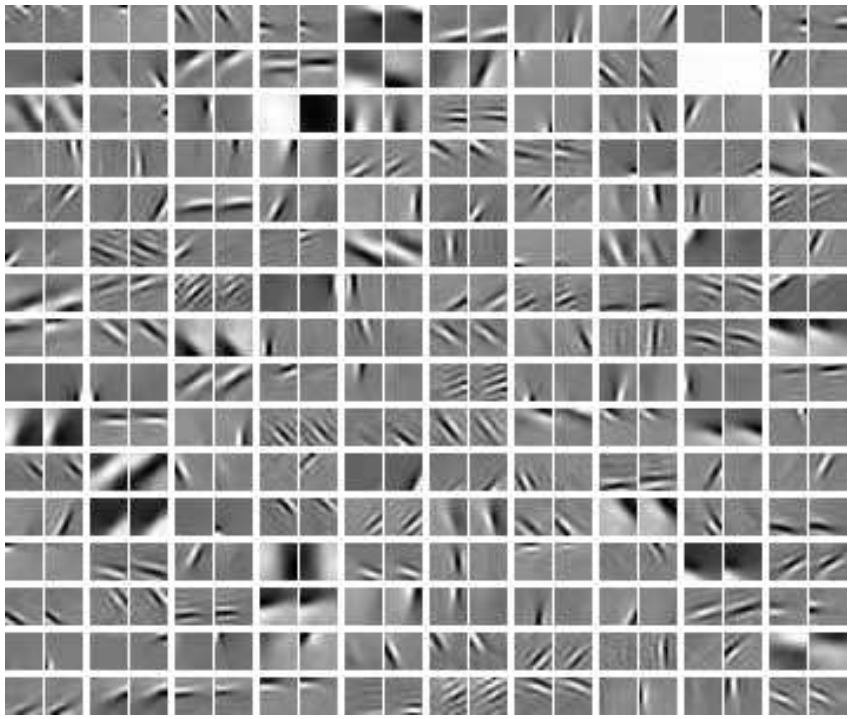

Figure 7.6: ICA basis of stereo images. Here, each neighboring *pair of patches* represents the contribution of one hidden variable s_j to the observed data, with the left and right patches denoting the contributions to the left and right eye data (respectively). Note how the degree of binocularity varies widely, and how binocular features have interocularly matched spatial frequency and orientation preferences.

Chapter 8

Modeling complex cells and topography

As discussed in the previous chapter, the basic ICA model has been used to account for many aspects of simple cell receptive fields. In light of this success, one might seek to model further aspects of V1 by a similar approach. The most conspicuous features of the primary visual cortex, aside from the responses of simple cells, are (a) the response properties of complex cells, and (b) the topography (columnar organization). This chapter will focus on the modeling of these two V1 characteristics.

8.1 Modeling complex cell responses

Could the basic ICA model also account for the response properties of complex cells? As discussed in chapter 3, complex cells respond to bars and edges, just like simple cells, but are not selective as to the *local phase* of the stimulus (for example, a black bar on a white background typically elicits a similar response to that of a white bar on a black background). It is not difficult to see that the ICA model is fundamentally incompatible with this kind of responses. In the basic ICA model, reversing the contrast polarity (that is, flipping the sign of the input vector $\mathbf{x}$) reverses the sign of the representation $\mathbf{s}$, so the representation is certainly not invariant to contrast reversal. Complex cells, on the other hand, typically are relatively invariant to contrast reversal.

Accordingly, the model must be extended. We have considered a model which retains important features of the basic ICA model yet modifies it in a significant way that allows the modeling of complex cell responses. The observed data $\mathbf{x}$ is given as a (noisy) linear combination of sparse hidden variables s_j, just as in the ICA model. However, the hidden variables are no longer completely independent, but can be divided into groups within which dependencies exist. (Similar relaxations of the independence assumption in ICA were proposed in [20, 88].) Within these groups the dependencies take the form of *correlations of energies*: if one component s_j in the group is significantly non-zero, other components in that group are also likely to be active. This means that components in a group tend to be simultaneously active more often than would be the case for independent components. Note, however, that there is no *linear* correlation, because the signs of the individual components are not predictable. This kind of dependency structure can be built into the

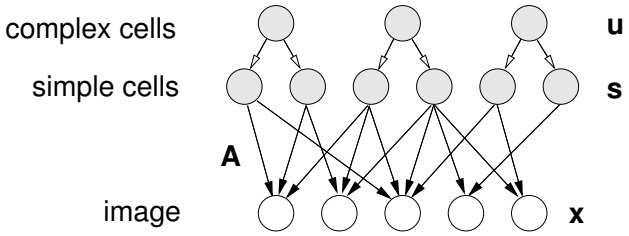

Figure 8.1: The complex cell model. Each higher-order hidden variable u_i is drawn independently and determines the *conditional variance* of the s_j in its group (they all have zero mean, however). The observations $\mathbf{x}$ are then determined as previously.

model by having higher-order hidden variables u_i that determine the conditional variance of the conditionally Gaussian components s_j, as shown in figure 8.1.

This 'independent subspace analysis' (ISA) was introduced in Publication 2, although it must be noted that we did not construct the hierarchical generative model of figure 8.1 until later [57]. The algorithm proposed in Publication 2 estimates the parameters a_{ij} of the model in the special case of infinitesimal noise and equal dimensionality of $\mathbf{s}$ and $\mathbf{x}$ ($n = m$).

Estimating the model from natural image patches, we obtained the features shown in figure 8.2. Each window gives one feature (column $\mathbf{a}_j$ from $\mathbf{A}$), and each set of four features belongs to the same group, as described above. The first thing to note is that the qualitative appearance of the features has not changed: most can still be described as Gabor filters. This means that the individual components s_j in the model still respond quite like simple cells do. But the truly interesting thing is the way the components are grouped: features belonging to the same group tend to have similar spatial frequency and orientation. This means that estimates of the u_i mirror the operation of complex cells.

To see this, remember that the model specified that the conditional variance of the s_j in a group was specified by the u_i connected to that group. To estimate this conditional variance one should sum the squares of the values s_j in the group. In other words, the optimal estimate of a higher-order latent variable u_i is a function of the sum of the squares of the responses of the s_j in its group. This should remind you of the 'energy model' for complex cell responses discussed in chapter 3, in which the responses of simple cells are squared and summed. Specifically, in that model simple cells with receptive fields having similar orientation and spatial frequency are pooled. This is just the kind of pooling that our network learned in a completely unsupervised manner! In the next section, where we describe an elaboration of the model, we show simulations that indicate that the u_i do operate like complex cells, at least qualitatively.

The fact that the model associates filters of similar orientation and frequency, but differing in phase, implies that the outputs of such linear filters (when applied to natural images) are dependent. We are certainly not the only nor the first to discuss such dependencies. As early as 1990, Wegmann and Zetzsche [163] described the energy correlations between such filters. Recently, Simoncelli and his colleagues [131, 139] have also described similar dependencies and have proposed to reduce them by a divisive normalization operation (see chapter 5).

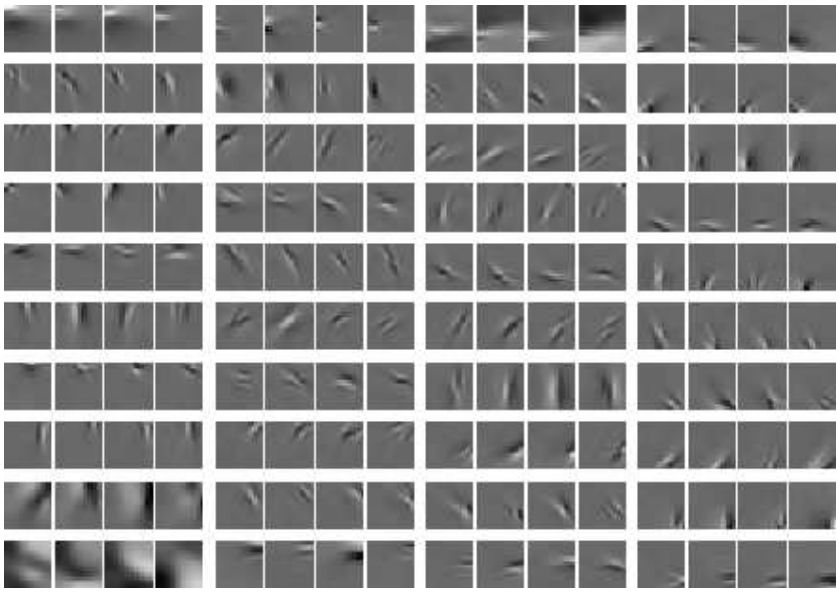

Figure 8.2: Learned weights in the complex cell model. Each patch, as before, shows the contribution of one hidden variable s_j to the observed data. Each set of neighboring four patches belong to the same complex cell group. Note how the features in a group tend to have similar orientation and spatial frequency.

8.2 Modeling topography

As briefly described in chapter 3, one of the most interesting features of the primary visual cortex is its topographic organization. Neurons which are physically close to each other have preferred stimuli which are similar in terms of spatial position, orientation, and spatial frequency [31, 52, 62, 134].[1] Could this columnar organization be modeled by statistical models?

In Publication 3 (see also [57, 61]), we showed how V1-like topography could be learned by a very simple modification of the complex cell model. This modified model, called 'Topographic Independent Component Analysis' (TICA), is illustrated in figure 8.3. The basic idea is that instead of having discrete groups of model simple cells, the groups are in a sense overlapping. Each simple cell is associated with *several* complex cells, instead of just one. In the generative model, the conditional variance of a simple cell is given by the sum of the activities of independent complex cells.

Note that now an explicit organization is imposed on the components s_j, because they are all tied together. Although for simplicity figure 8.3 shows a one-dimensional model, we shall mostly work with a two-dimensional arrangement of components (because we are attempting to model the layout of the cortex, which is essentially a thin two-dimensional sheet [100]). In such an arrangement, complex cells are associated with a small neighborhood (e.g. 3-by-3) of simple cells arranged on a two-dimensional grid.

[1] Neurons are also organized according to various non-spatial receptive field properties, such as ocular dominance, but for simplicity these will not be considered here.

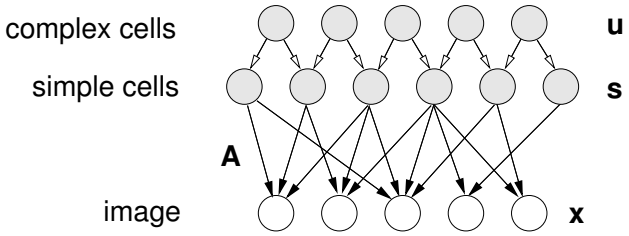

Figure 8.3: The topographic generative model. This model works exactly like the complex cell model (figure 8.1), with the exception that the conditional variances of the s_j are determined by the *sum* of the activity of *near-by* complex cells u_i.

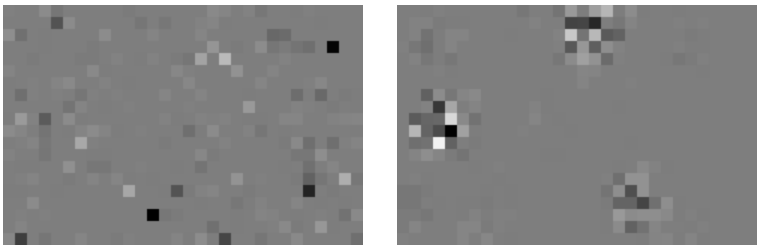

Figure 8.4: Sparse versus clustered sparse activity. Each pixel denotes the activity of one latent variable s_j, with gray representing zero whereas white and black represent strong positive and negative values (respectively). Left: typical activity pattern in ICA model. Most units are effectively zero, and there are only few strong responses. Right: typical activity pattern in topographic model. The responses are similar, but in addition tend to be clustered, so that neighboring units often are active simultaneously.

Just like in the complex cell model of figure 8.1, the topographic model preserves the important property of sparseness of the components s_j but rejects the notion of complete independence between them. And just as in the complex cell model, the components in the topographic model show correlations of energies. The novelty of the topographic model is that the components s_j are ordered on a grid, and those which are *close to each other* on the grid show this dependence. This essentially means that for typical input, activity in the grid tends to be clustered. This point is illustrated in figure 8.4.

In Publication 3 and [57, 61], the model was estimated from natural image patches.[2] The features obtained once again resemble simple cell receptive fields. But now the components are organized. Figure 8.5 shows a typical learned basis. Note that the features with low spatial frequency are clustered in the map. Another notable characteristic of the map is that neighboring units tend to exhibit similar orientation preferences.

To further analyze the structure of the map, one can fit Gabor functions (see chapter 3) to the individual features, and then measure how the various Gabor parameters of neighboring units correlate. This type of an analysis is shown in figure 8.6, using the estimated

[2] Again, the actual algorithm used corresponds to the no-noise and complete basis model special case. In addition, an approximation to the exact likelihood was used in order to get a simple and efficient algorithm, see [57, 61].

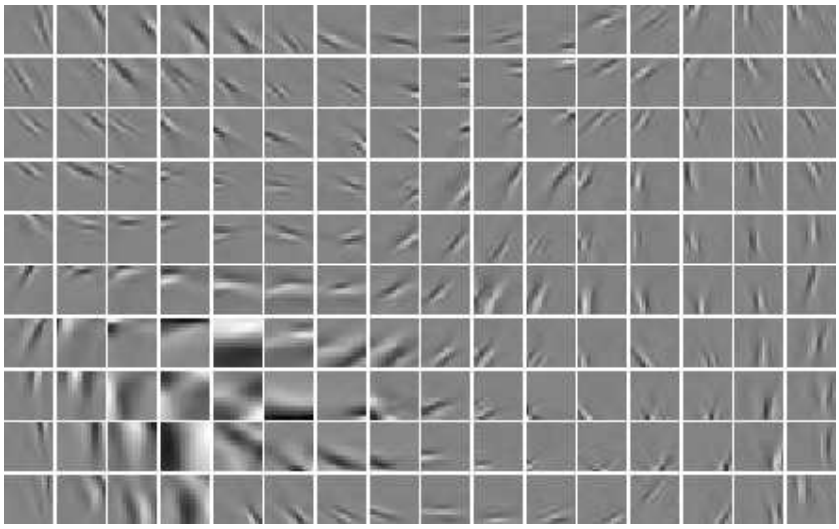

Figure 8.5: Features learned by the topographic model. Again, each patch represents one $\mathbf{a}_j$ ($a_j(x, y)$), the contribution of one hidden unit s_j to the observed data. Note that the learned features still look like simple cell receptive fields, but now they are further arranged so that neighboring units tend to exhibit similar orientation and spatial frequency preferences.

basis shown in figure 4 of Publication 3. How well does the organization produced by our model match that found in V1? Although many aspects of V1 topography have been studied for decades, the recent study by DeAngelis et al. [31] is the first comprehensive account of neighboring neuron receptive field parameter correlations in the primary visual cortex.[3] One of their primary finding was that, surprisingly, there is virtually no correlation with respect to receptive field phase. Their study is in excellent agreement with our results, which also gave strong correlations between the orientation and spatial frequency parameters, but no correlation of the phase parameter.

Exactly like in the complex cell model, the higher-order activities u_i in the topographic model respond like actual complex cells because they pool the responses of filters with similar orientation and frequency but differing in phase. To show that this arrangement works in the model, we calculated tuning curves for phase, location, orientation, and frequency for all model simple cells and complex cells. These tuning curves are summarized in figure 8.7. The tuning curves strongly resemble those measured from real neurons: simple cells are selective for all parameters, whereas complex cells are insensitive to stimulus phase and exhibit a decreased sensitivity to location.

The proposed topographic model for natural images has some interesting connections to other contemporary work on image statistics. Especially the recent work on the dependencies between nodes in a wavelet tree [161] incorporates many similar ideas. The main differences are that we employ a two-dimensional grid structure (as opposed to a wavelet tree) and that we estimate the linear filters from the data instead of fixing them a priori. Another related model is that proposed in [167].

[3] Note that this study concerned the primary visual cortex of cats, not primates.

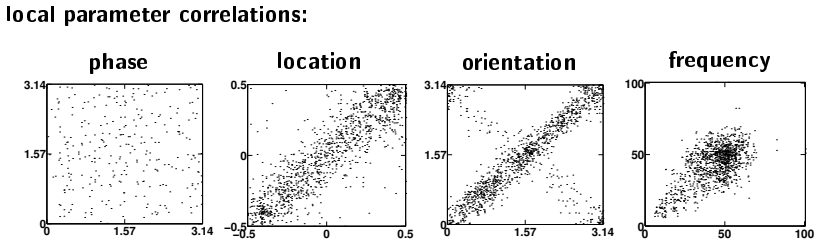

Figure 8.6: Neighboring unit parameter correlations. A Gabor function was fit to each basis vector $a_j(x, y)$, describing it in terms of location within the window, orientation, spatial frequency, and local phase. Each point in each of the four scatter-plots gives the corresponding parameter values of one pair of neighboring units. Note the strong correlations of location, orientation, and frequency. However, there is no organization of local phase. (Cf. figure 11 of [31]).

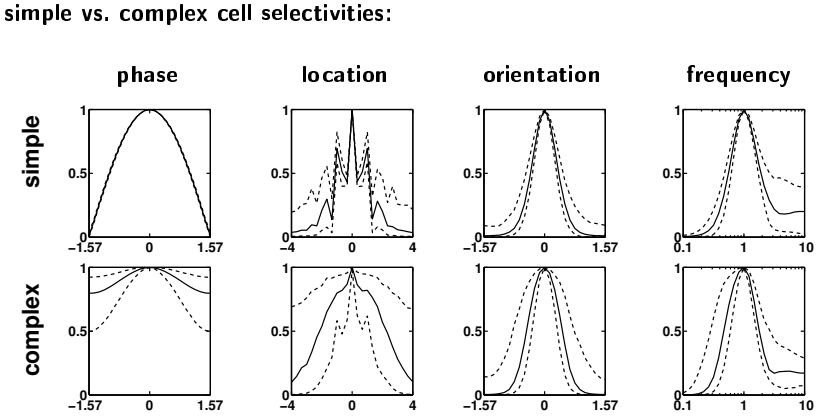

Figure 8.7: Tuning curves for model simple and complex cells. The top row shows tuning curves of the model simple cells (s_j) for the different stimulus parameters; the bottom row gives the corresponding curves for the model complex cells (u_i). In all plots, the solid curves give the median responses; the dashed curves indicate 90 and 10% quantiles.

Chapter 9

Non-negativity constraints

9.1 Why non-negativity?

Although the basic ICA model described in chapter 7 (figure 7.4) has had significant success in modeling the receptive fields of simple cells, there are at least two obvious ways in which it is unrealistic as a model of simple cell behavior. Perhaps the most evident discrepancy is the fact that in the model each unit s_j can, in addition to being effectively silent (close to zero), be either positively or negatively active. This basically means that every feature contributes to representing stimuli of opposing polarity; for example, the same unit that codes for a dark bar on a bright background also codes for a bright bar on a dark background when the sign of the unit is reversed. This is in contrast to the behavior of simple cells in V1: these neurons tend to have quite low background firing rates and, as firing rates cannot go negative, thus can only represent one half of the output distribution of a signed unit s_j.

Another major difference between the model and V1 is that the input data in the model is double-sided (signed), whereas V1 receives the visual data from the lateral geniculate nucleus (LGN) in the form of separated ON- and OFF-channels. Of course, as an abstract model of visual coding, the input data should indeed be (signed) image contrast. But if we on the other hand are interested in how V1 recodes its input signals, we must consider separate ON- and OFF-channel input.

Thus, if we would like to transform the basic ICA model from a relatively abstract model of image representation in V1 to a concrete model of simple cell recoding of inputs coming from the LGN, the model must be changed. First, our input data should consist of hypothetical firing rates of ON- and OFF-center LGN cells in response to natural image patches. Second, all coefficients s_j should be restricted to non-negative values. As both the sources $\mathbf{s}$ and the data $\mathbf{x}$ thus have non-negative values only, it is logical to assume the same of the model parameters a_{ij}. This is because if some weights a_{ij} were negative, the generative model would inevitably produce some negative data as well.[1]

Although we so far considered non-negativity constraints with the objective of making the model better fit known neurophysiology, there are also other equally important

[1] Technically, even with both the coefficients and the weights non-negative, a non-zero noise level σ would generate some negative data. This is however not significant for the relatively low noise levels typically considered.

arguments for non-negativity. In particular, it has been argued that non-negativity can be important for learning parts-based representations [81]: In the standard sparse coding model, the data is described as a combination of elementary features involving both additive and subtractive interactions. The fact that features can 'cancel each other out' using subtraction is contrary to the intuitive notion of combining parts to form a whole. Non-negativity ensures that elementary object features combine additively.

9.2 Non-negative sparse coding

Adding non-negativity constraints to the sparse coding/ICA model (described in chapter 7) is straightforward. $\mathbf{A}$ and $\mathbf{s}$ are both constrained to have non-negative elements only, and the same is assumed of the data $\mathbf{x}$. The latent variables s_j are assumed to have exponential distributions, i.e. $p(s_j) = \exp(-s_j)$, and the data $\mathbf{x}$ are generated from the latent variables as before, see equation (7.3). In Publication 4, we developed a simple algorithm for inferring the *maximum a posteriori* estimate of the s_j given the input $\mathbf{x}$ and the generative weight matrix $\mathbf{A}$, under these non-negativity constraints. We also showed how to learn the generative weights from the observed data.

We are certainly not the first to consider non-negativity constraints in linear models. As early as 1994, Paatero and Tapper [112] described *positive matrix factorization*, which attempts to reconstruct the non-negative input matrix as a product of two non-negative matrices with lower dimensionality. This was subsequently put into a neurobiological context by Lee and Seung [81] (calling it *non-negative matrix factorization*), who also developed efficient algorithms for solving the problem [81, 82]. In the context of ICA, non-negativity constraints (on $\mathbf{A}$, $\mathbf{s}$, or both) have recently been considered by several authors, see e.g. [97, 106, 114, 117, 118].

The main contributions of Publication 4 were the application of non-negativity constraints in the sparse coding framework [110, 111], and the extension of the algorithm proposed in [82] to this case.

9.3 Learning receptive fields

In Publication 5, the algorithm proposed in Publication 4 was applied to learning a non-negative sparse representation of simulated ON/OFF-channel image data. Natural images[2] were filtered to yield inputs x_i that mimicked the responses of ON- and OFF-center cells to the images. That is, each x_i was obtained by filtering the original image patches with a center-surround linear receptive field (see chapter 3) and then performing half-wave rectification:

$$x_i = \text{rect}\left(\sum_{(x,y)} w_i(x,y) I(x,y)\right), \tag{9.1}$$

where $w_i(x,y)$ is a manually constructed center-surround filter and

$$\text{rect}(z) = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$

[2]The images used in these experiments were kindly provided by Bruno Olshausen.

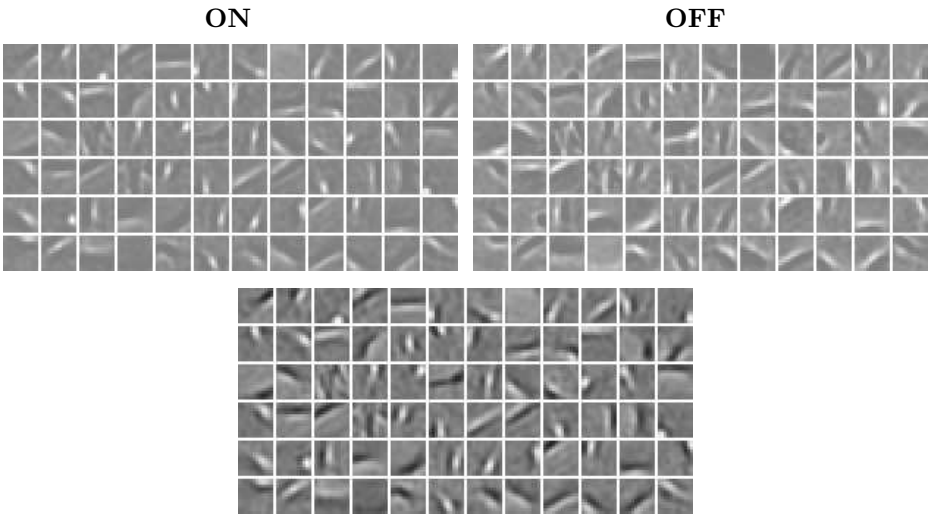

Figure 9.1: Learned features in a non-negative representation of ON/OFF-channel image data. Top left: Generative weights for the ON-channel. Each patch represents the part of one basis vector $\mathbf{a}_j$ corresponding to the ON-channel input. Gray pixels denote zero weights, brighter pixels represent positive weights. Top right: Corresponding weights for the OFF-channel input. Bottom: Weights for ON minus weights for OFF.

The filters $w_i(x, y)$ were varied so that half of them were of ON-type and the other half of OFF-type, and they covered all spatial locations in the 12×12 pixel image patch, for a total of $m = 2 \times 12 \times 12 = 288$ filters. The weight matrix $\mathbf{A}$ was then learned using the algorithm described in Publication 4.

The resulting basis patterns (columns of $\mathbf{A}$) are shown in figure 9.1. Note how the basis vectors $\mathbf{a}_j$ represent Gabor-like image input by elongated ON- and OFF-subregions. This is most clearly seen in the bottom panel, where the weights for the OFF-channel have been subtracted from those for the ON-channel. These learned features are at least qualitatively similar to the ones found with the standard sparse coding model applied to symmetric image data, see figure 7.3.[3]

These results show that it is straightforward to adapt the basic ICA model to yield a representation that more closely corresponds to the neurophysiology of V1. Although this is certainly a meaningful endeavour on its own, our main goal for exploring non-negative representations was that they might be especially useful when considering representations further along in the visual processing pathway. In the next chapter we report on some early steps in that direction.

[3]Note that here, the weights of the matrix $\mathbf{A}$ make up the mapping from simple cell responses to LGN responses, so the patches of figure 9.1 are not image patches but rather LGN activity patterns. However, as the LGN responses are essentially (rectified) band-pass filters, these activity patterns can nevertheless be reasonably compared with the image patches of figure 7.3.

Chapter 10

Contour coding

10.1 A simplified hierarchical network

The preceding chapters described how relatively simple latent variable models are able to account for many aspects of the organization of the primary visual cortex. With these successes in mind, it is tempting next to try to account for response properties of neurons higher in the processing hierarchy. Here, one runs into several problems. The responses of neurons later in the processing chain are increasingly complex nonlinear functions of the input (see, e.g., [44, 146]). This means that simple linear models cannot be directly used any more. Another problem is that not much is actually known about the response properties of such neurons. The two points are actually related: Because of the complicated nonlinear responses of the neurons it is difficult to get a coherent picture of what they are doing. Even when such a picture is available, *how* they are doing it is often not known.

Nevertheless, one may attempt to extend the models described earlier to learn more complex image structure. Perhaps the most straightforward approach is to extend the complex cell model discussed in section 8.1 by assuming that the activities of the complex cells are not independent, but instead are given by the non-negative ICA model discussed in the previous chapter. This would amount to adding a linear layer on top of the model of figure 8.1. In Publication 6 we studied a simplified version of that model, where the lower layers were neglected and the responses of the model complex cells were a straightforward function of the image input. This is depicted in figure 10.1, where the lower layers are grayed out to emphasize that these layers do not play any active role in this simplified model.

To begin with, we had to select a specific form for the responses of our model complex cell. The natural choice seemed to be the classic energy model, introduced in chapter 3. Not only is this model perhaps the most widely used complex cell response model, but it also fits naturally in the framework of latent variable models, as shown in chapter 8. For each image patch, we calculated the responses of such hypothetical complex cells at a variety of spatial positions and preferred orientations. These responses formed the components of a vector $\mathbf{x}$, the 'complex cell activity pattern' elicited by a given natural image patch. A few such activity patterns are shown in figure 10.2.

Having sampled a large number of activation patterns $\mathbf{x}$ (such as those shown in figure 10.2b), we estimated the parameters a_{ij} of a linear ICA model (see chapter 7) with non-negativity constraints (as described in chapter 9). Each complex cell activation pat-

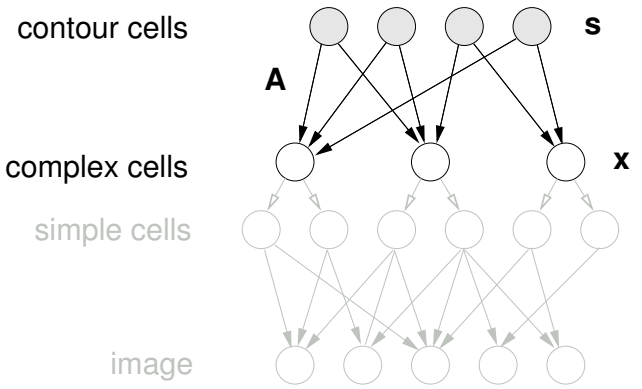

Figure 10.1: The simplified hierarchical model. Model complex cell responses are calculated in a manually specified feedforward manner, and these responses are modeled by the non-negative ICA model discussed in the previous chapter. To emphasise that the lower layers play no active part in the model they have been grayed out.

tern is represented by one data vector $\mathbf{x}$, with each element x_i representing the firing rate of one complex cell. Each s_j represents the response of one higher-order neuron, whose 'receptive field' is closely related to the corresponding $\mathbf{a}_j$. Again, the goal is to find basis patterns $\mathbf{a}_j$ such that typical input patterns $\mathbf{x}$ can be described accurately using only a few significantly active higher-order neurons, see figure 10.3.

10.2 Sparse coding of contours

A representative subset of the estimated basis patterns $\mathbf{a}_j$ is shown in figure 10.4. Note that most basis patterns consist of a variable number of active complex cells arranged collinearly. This makes intuitive sense, as collinearity is a strong feature of the visual world [41, 78, 133]. In addition, visual analysis in terms of smooth contours is supported by evidence from both psychophysics [38, 120] and physiology [69, 70, 119], and is incorporated in many models of contour integration, see e.g. [45, 87, 104]. To our knowledge, ours is the first model to learn this type of a representation from the statistics of natural images.

It is easy to understand why basis patterns consist of collinear complex cell activity patterns: Such patterns are typical in the data set, and can be sparsely coded if a long contour can be represented by only a few higher-level units. The necessity for *different* length basis patterns comes from the fact that long basis patterns simply cannot code short (or curved) contours, and short basis patterns are inefficient at representing long, straight contours.

Although the network is linear from the latent variables s_j to the data $\mathbf{x}$, the inferred (most likely) s_j are a nonlinear function of the data $\mathbf{x}$, due to the noise and the *overcompleteness* of the basis [111], as also discussed in chapter 7. In other words, the contour-coding neurons respond to the complex cell activity patterns in a nonlinear fashion. In particular, there is competition between the neurons [111], so that they respond only when they are better than competing units at representing the stimulus. As a prominent feature of the learned representation is the existence of different-length patterns, this leads to units being selective for contour length, in addition to being tuned to position and orientation.

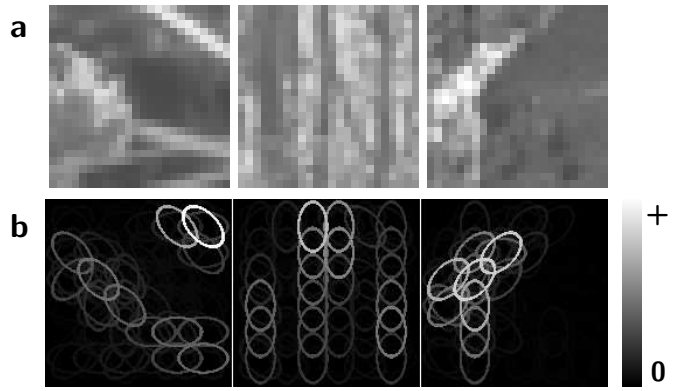

Figure 10.2: Model complex cell responses to natural image patches. (a) Three patches from the set of natural images. (b) Responses of the model complex cells to the patches. The ellipses show the orientation and approximate extent of the receptive fields of individual complex cells. The brightness of the different ellipses indicate the response strengths.

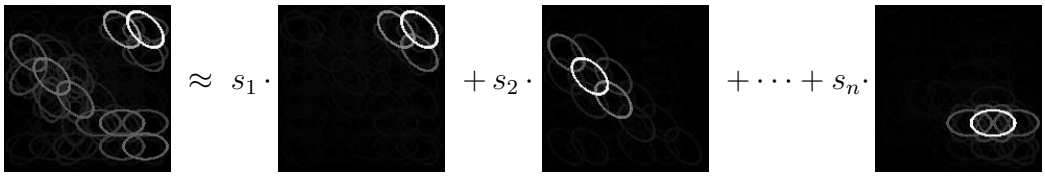

Figure 10.3: Sparse coding of complex cell responses. Each complex cell activity pattern is represented as a linear combination of basis patterns $\mathbf{a}_j$. The goal is to find basis patterns such that the coefficients s_j are as 'sparse' as possible, meaning that for most input patterns only a few of them are needed to represent the pattern accurately.

In other words, units representing long contours do not respond to short ones, whereas units coding short contours exhibit *end-stopping* [53, 54].

To illustrate the nonlinear transform from complex cell activities $\mathbf{x}$ to higher-order activities s_j we can make a linear approximation. Optimal approximating linear filters are shown in figure 10.5b, for the units whose basis patterns are depicted in figure 10.5a. Note that units representing short contour segments tend to have inhibitory regions at one (or both) of the ends of their 'receptive fields', illustrating the end-stopping effect. On the other hand, units which code longer contours have inhibitory weights from complex cells which are positioned on the contour but are of the wrong orientation. This enhances the selectivity of these units so that they don't respond to contours that only partly overlap the receptive field.

The nonlinear effects can also be seen by directly showing length-tuning curves (figure 10.5c). Each plot shows how the response of the corresponding higher-order unit s_j varies with the length of the stimulus, when all other stimulus parameters are held at their optimal values. The length of the stimulus (relative to the length of the sampling window) is given on the horizontal axis (note the logarithmic scale) and the corresponding response is plotted on the vertical axis. Notice how the response of the end-stopped units starts to decrease when the stimulus length increases past its optimal value, eventually falling to

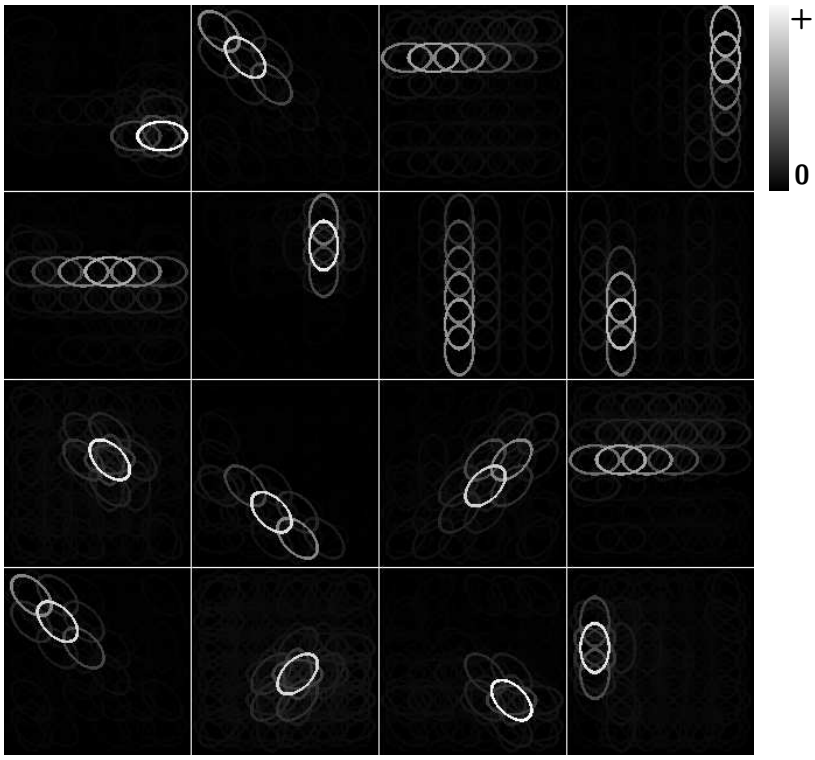

Figure 10.4: A representative set of basis functions from the learned basis. The majority of units code the simultaneous activation of collinear complex cells, indicating a smooth contour in the image.

zero. On the other hand, the response of the unit coding long contours does not decline by any significant degree. These results thus show that our model higher-level units have extra-classical properties that make them clearly distinct from standard complex cells.

It should be noted that those higher-order units which represent long contours bear many similarities to 'collator' (or 'collector') units, proposed in the psychophysical literature [99, 102]. Such units are thought to integrate the responses of smaller, collinear filters, to give a more robust estimate of global orientation than could be achieved with elongated linear mechanisms.

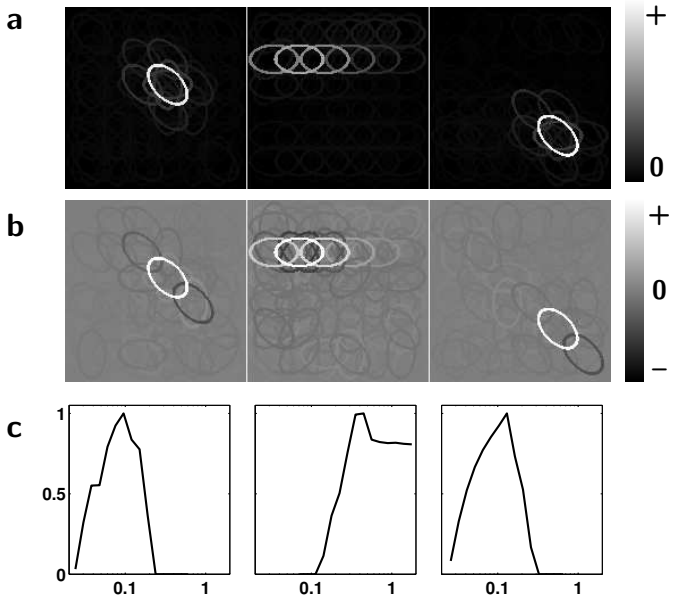

Figure 10.5: (a) Three basis patterns $\mathbf{a}_j$ from the estimated basis. (b) Optimal approximating linear filters for the units in (a). These are the filters that minimize the mean squared error between the linear response (followed by half-rectification) and the optimal activations. White ellipses denote excitatory connections from complex cells, black ellipses inhibitory weights. (c) Length-tuning curves for the units in (a). The horizontal axis gives the length of the stimulus (logarithmic scale, relative to the size of the sampling window) and the vertical axis denotes response strength (normalized to a maximum of one).

Chapter 11

Conclusion

11.1 Main points of this thesis

This thesis has built on the ideas introduced almost half a century ago by Attneave [7] and Barlow [8], and since then further developed by several researchers. Attneave and Barlow emphasized that the natural sensory input of the biological visual system is highly redundant, and that the statistical structure of this input provides knowledge about the structure of the world. This is now widely appreciated. There is also a general agreement that visual systems must make use of the redundancy to be able to optimally extract the information provided.

Despite the consensus on these general issues, there is little agreement on the specifics. How should structure be identified, and how is the information extracted? To what degree is the structure of the environment 'hard-coded' by evolution and to what degree learned during the lifetime of the individual? Furthermore, the problem is confounded by the fact that quite different theoretical frameworks for identifying and utilizing redundancy often lead to exactly the same optimization criteria. The prime example of this is independent component analysis (ICA), which can be motivated as efficient or sparse coding, or as the estimation of a latent variable model.

In this thesis we have advocated the latent variable approach to understanding the connection between natural image statistics and early vision. Specifically, we have extended the ICA model, allowing us to account for complex cell properties and V1 topography in addition to simple cell receptive fields. Furthermore, we have discussed how non-negativity constraints in the ICA model allow a closer correspondence to V1 processing. Finally, we have made a preliminary investigation concerning the application of probabilistic models to the modeling of further stages of early visual processing.

As to the question of nature versus nurture, this thesis is completely neutral. Regardless of the mechanism by which receptive fields in the primary visual cortex are formed, the simulations provided in this thesis demonstrate that they might be interpreted as parameters of a hidden variable model of natural sensory data. In other words, algorithms estimating these models are not meant as explicit models of neural development.

11.2 Future prospects

This thesis has 'explained' many basic receptive field properties of neurons in the primary visual cortex. Of course, numerous explanations have been given for V1 receptive fields (see e.g. [10, 28, 94, 121]) and cortical topography (e.g. [35, 76, 96, 107, 158], for reviews see [36, 144]) in the past. How will we ever be able to say which explanation is the right one? In science, models are compared by (a) their explanatory power, and (b) their simplicity. Given two models that equally well describe some given phenomena, the simpler one is usually taken to be the better one. This principle is known as Occam's razor. On the other hand, given two models which are equally simple, the one that explains more phenomena and/or more accurately, is superior. (These intuitively pleasing rules can actually be motivated using Bayesian theory, see e.g. [125].)

Our explanation for V1 structure must thus be pitted against other explanations in terms of explanatory power and simplicity. Obviously, latent variable models such as ours are much more complicated than simply saying 'neurons in V1 perform a local spatial frequency analysis of the input', so our model must also account for more experimental facts to stand a chance of survival. Fortunately, there are many facts requiring explanation: the exact structure of linear receptive fields (including spatiotemporal, binocular, and chromatic structure), complex cell response properties, the precise topographical structure of V1, etc. Due to the broad success of latent variable models such as those proposed in this thesis, we believe that they are competitive as theories of cortical function.

The ultimate test of such probabilistic models, however, will be how well they predict facts as yet unknown about the visual system. So far, much of theoretical neuroscience has been lagging behind experimental work, 'explaining' earlier observations. But as theory develops one would hope that it could also give useful predictions that could subsequently be verified in biological experiments. In parts of physics, for example, theory has been steering experiments, instead of the other way around. It is my belief that this will soon, at least partly, happen in neuroscience as well. Investigating the visual processing 'hierarchy' using multi-level latent variable models, such as those proposed in this thesis, might take us one step in that direction.

Bibliography

[1] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A*, vol. 2, no. 2, pp. 284–299, 1985.

[2] E. D. Adrian, *The basis of sensation: The action of the sense organs*. W. W. Norton, 1926.

[3] J.-M. Alonso and L. M. Martinez, "Functional connectivity between simple cells and complex cells in cat striate cortex," *Nature Neuroscience*, vol. 1, no. 5, pp. 395–403, 1998.

[4] S.-I. Amari and A. Cichocki, "Adaptive blind signal processing – neural network approaches," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2026–2048, 1998.

[5] A. Anzai, I. Ohzawa, and R. D. Freeman, "Neural mechanisms for encoding binocular disparity: Receptive field position vs. phase," *Journal of Neurophysiology*, vol. 82, no. 2, pp. 874–890, 1999.

[6] J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network: Computation in neural systems*, vol. 3, pp. 213–251, 1992.

[7] F. Attneave, "Some informational aspects of visual perception," *Psychological Review*, vol. 61, pp. 183–193, 1954.

[8] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication* (W. A. Rosenblith, ed.), pp. 217–234, MIT Press, 1961.

[9] H. B. Barlow, "Single units and sensation: A neuron doctrine for perceptual psychology?," *Perception*, vol. 1, pp. 371–394, 1972.

[10] H. B. Barlow, "What is the computational goal of the neocortex?," in *Large-scale neuronal theories of the brain* (C. Koch and J. L. Davis, eds.), pp. 1–22, MIT Press, 1994.

[11] H. B. Barlow, "The exploitation of regularities in the environment by the brain," *Behavioral and Brain Sciences*, vol. 24, no. 3, 2001.

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

[13] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[14] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327–3338, 1997.

[15] R. Blake and N. K. Logothetis, "Visual competition," *Nature Reviews Neuroscience*, vol. 3, pp. 13–21, 2002.

[16] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, no. 11, pp. 947–957, 1999.

[17] M. Carandini, D. J. Heeger, and J. A. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *Journal of Neuroscience*, vol. 17, pp. 8621–8644, 1997.

[18] J.-F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Signal Processing Letters*, vol. 4, pp. 112–114, 1997.

[19] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[20] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, (Seattle, WA), 1998.

[21] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[22] A. Cichocki and R. Unbehauen, *Neural Networks for Signal Processing and Optimization*. Wiley, 1994.

[23] P. Comon, "Independent component analysis – a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[24] B. R. Conway, "Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (v-1)," *Journal of Neuroscience*, vol. 21, no. 8, pp. 2768–2783, 2001.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[26] K. J. W. Craik, *The Nature of Explanation*. Cambridge University Press, 1943.

[27] H. Damasio and A. R. Damasio, *Lesion analysis in neuropsychology*. Oxford University Press, 1989.

[28] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A*, vol. 2, pp. 1160–1169, 1985.

[29] P. Dayan, "Recognition in hierarchical models," in *Foundations of Computational Mathematics* (F. Cucker and M. Shub, eds.), Springer, 1997.

[30] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz machine," *Neural Computation*, vol. 7, pp. 889–904, 1995.

[31] G. C. DeAngelis, G. M. Ghose, I. Ohzawa, and R. D. Freeman, "Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons," *Journal of Neuroscience*, vol. 19, no. 10, pp. 4046–4064, 1999.

[32] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development," *Journal of Neurophysiology*, vol. 69, pp. 1091–1117, 1993.

[33] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation," *Journal of Neurophysiology*, vol. 69, pp. 1118–1135, 1993.

[34] B. M. Dow and R. G. Vautin, "Horizontal segregation of color information in the middle layers of foveal striate cortex," *Journal of Neurophysiology*, vol. 57, pp. 712–739, 1987.

[35] R. Durbin and G. Mitchison, "A dimension reduction framework for understanding cortical maps," *Nature*, vol. 343, pp. 644–647, 1990.

[36] E. Erwin, K. Obermayer, and K. Schulten, "Models of orientation and ocular dominance columns in the visual cortex: A critical comparison," *Neural Computation*, vol. 7, pp. 425–468, 1995.

[37] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, pp. 559–601, 1994.

[38] D. J. Field, A. Hayes, and R. F. Hess, "Contour integration by the human visual system: Evidence for a local 'association field'," *Vision Research*, vol. 33, no. 2, pp. 173–193, 1993.

[39] P. Földiák and M. P. Young, "Sparse coding in the primate cortex," in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), pp. 895–898, The MIT Press, 1995.

[40] W. S. Geisler and D. G. Albrecht, "Cortical neurons: isolation of contrast gain control," *Vision Research*, vol. 32, no. 8, pp. 1409–1410, 1992.

[41] W. S. Geisler, J. S. Perry, B. J. Super, and D. P. Gallogly, "Edge co-occurence in natural images predicts contour grouping performance," *Vision Research*, vol. 41, pp. 711–724, 2001.

[42] R. Gonzalez and R. Woods, *Digital Image Processing*. Addison-Wesley, 1992.

[43] U. Grenader, *Lectures in Pattern Theory I, II, and III: Pattern analysis, pattern synthesis and regular structures*. Springer-Verlag, 1976–1981.

[44] C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender, "Visual properties of neurons in inferotemporal cortex of the macaque," *Journal of Neurophysiology*, vol. 35, pp. 96–111, 1972.

[45] S. Grossberg and E. Mingolla, "Neural dynamics of perceptual grouping: textures, boundaries and emergent segmentations," *Perception and psychophysics*, vol. 38, no. 2, pp. 141–171, 1985.

[46] M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of Modern Physics*, vol. 65, no. 2, pp. 413–497, 1993.

[47] G. F. Harpur and R. W. Prager, "Development of low entropy coding in a recurrent network," *Network: Computation in Neural Systems*, vol. 7, pp. 277–284, 1996.

[48] D. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, pp. 181–198, 1992.

[49] G. E. Hinton and Z. Ghahramani, "Generative models for discovering sparse distributed representations," *Phil. Trans. R. Soc. Lond. B*, vol. 352, pp. 1177–1190, 1997.

[50] D. H. Hubel, *Eye, brain, and vision*. Scientific American Library, 1988.

[51] D. H. Hubel and T. N. Wiesel, "Integrative action in the cat's lateral geniculate body," *Journal of Physiology*, vol. 155, pp. 385–398, 1961.

[52] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106–154, 1962.

[53] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat," *Journal of Neurophysiology*, vol. 28, pp. 229–289, 1965.

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

[55] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[56] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Computation*, vol. 11, no. 7, pp. 1739–1768, 1999.

[57] A. Hyvärinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, 2001.

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

[59] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[60] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, pp. 429–439, 1999.

[61] M. Inki, "Topographic independent component analysis: Theory and applications," Master's thesis, Helsinki University of Technology, Lab of Computer and Information Science, 2000.

[62] N. P. Issa, C. Trepel, and M. P. Stryker, "Spatial frequency maps in cat visual cortex," *Journal of Neuroscience*, vol. 20, no. 22, pp. 8504–8514, 2000.

[63] F. V. Jensen, *Bayesian Networks and Decision Diagrams*. Springer, 2001.

[64] E. N. Johnson, M. J. Hawken, and R. Shapley, "The spatial transformation of color in the primary visual cortex of the macaque monkey," *Nature Neuroscience*, vol. 4, no. 4, pp. 409–416, 2000.

[65] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, pp. 1187–1211, 1987.

[66] M. I. Jordan, ed., *Learning in Graphical Models*. MIT Press, 1999.

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

[68] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of neural science*. McGraw-Hill, 4th ed., 2000.

[69] M. K. Kapadia, M. Ito, C. D. Gilbert, and G. Westheimer, "Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys," *Neuron*, vol. 15, no. 4, pp. 843–856, 1995.

[70] M. K. Kapadia, G. Westheimer, and C. D. Gilbert, "Spatial distribution of contextual interactions in primary visual cortex and in visual perception," *Journal of Neurophysiology*, vol. 84, pp. 2048–2062, 2000.

[71] J. Karhunen, E. Oja, L. Wang, R. Vigário, and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 8, no. 3, pp. 486–504, 1997.

[72] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*. Charles Griffin & Company, 1958.

[73] D. Kersten, "High-level vision as statistical inference," in *The new cognitive neurosciences* (M. S. Gazzaniga, ed.), pp. 353–363, MIT Press, 2000.

[74] D. Kersten and P. Schrater, "Pattern inference theory: A probabilistic approach to vision," in *Perception and the Physical World* (R. Mausfeld and D. Heyer, eds.), Wiley & Sons, 2002.

[75] D. C. Knill and W. Richards, eds., *Perception as Bayesian Inference*. Cambridge University Press, 1996.

[76] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 56–69, 1982.

[77] B. Kolb and I. Q. Whishaw, *Fundamentals of human neuropsychology*. W. H. Freeman, 4th ed., 1995.

[78] N. Krüger, "Collinearity and parallelism are statistically significant second order relations of complex cell responses," *Neural Processing Letters*, vol. 8, pp. 117–129, 1998.

[79] S. W. Kuffler, "Discharge patterns and functional organization of mammalian retina," *Journal of Neurophysiology*, vol. 16, pp. 37–68, 1953.

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

[81] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[82] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing 13 (Proc. NIPS*2000)*, MIT Press, 2001.

[83] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources," *Neural Computation*, vol. 11, no. 2, pp. 417–441, 1999.

[84] P. Lennie, J. Krauskopf, and G. Sclar, "Chromatic mechanisms in striate cortex of macaque," *Journal of Neuroscience*, vol. 10, no. 2, pp. 649–669, 1990.

[85] M. Lewicki and B. Olshausen, "Probabilistic framework for the adaption and comparison of image codes," *J. Opt. Soc. Am. A: Optics, Image Science, and Vision*, vol. 16, no. 7, pp. 1587–1601, 1999.

[86] M. Lewicki and T. J. Sejnowski, "Bayesian unsupervised learning of higher order structure," in *Advances in Neural Information Processing 9 (Proc. NIPS*96)*, pp. 529–535, MIT Press, 1997.

[87] Z. Li, "Pre-attentive segmentation in the primary visual cortex," *Spatial Vision*, vol. 13, no. 1, pp. 25–50, 1999.

[88] J. K. Lin, "Factorizing multivariate function classes," in *Advances in Neural Information Processing 10 (Proc. NIPS*97)*, MIT Press, 1998.

[89] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, pp. 105–117, 1988.

[90] M. S. Livingstone and D. H. Hubel, "Anatomy and physiology of a color system in the primate visual cortex," *Journal of Neuroscience*, vol. 4, pp. 309–356, 1984.

[91] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annual Review of Neuroscience*, vol. 19, pp. 577–621, 1996.

[92] E. Mach, *The analysis of sensations, and the relation of the physical to the phychical* (Translation of the 1st, revised from the 5th, German edition by S. Waterlow). Open Court (also reprinted 1959 by Dover), 1886.

[93] S. Marcelja, "Mathematical description of the responses of simple cortical cells," *Journal of the Optical Society of America*, vol. 70, no. 11, pp. 1297–1300, 1980.

[94] D. Marr, *Vision*. W. H. Freeman and Company, 1982.

[95] B. W. Mel, D. L. Ruderman, and K. A. Archie, "Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations," *Journal of Neuroscience*, vol. 18, pp. 4325–4334, 1998.

[96] K. D. Miller, "Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns," in *Models of Neural Networks III* (E. Domany, J. L. van Hemmen, and K. Schulten, eds.), pp. 55–78, Springer-Verlag, New York, 1995.

[97] J. Miskin and D. J. C. MacKay, "Ensemble learning for blind image separation and deconvolution," in *Advances in Independent Component Analysis* (M. Girolami, ed.), Springer-Verlag, 2000.

[98] M. C. Morrone and D. C. Burr, "Feature detection in human vision: a phase-dependent energy model," *Proc. Royal Soc. London Ser. B*, vol. 235, no. 1280, pp. 221–245, 1988.

[99] B. Moulden, "Collator units: Second-stage orientational filters," in *Higher-order processing in the visual system. Ciba Foundation Symposium 184*, pp. 170–192, John Wiley, 1994.

[100] V. B. Mountcastle, "The columnar organization of the neocortex," *Brain*, vol. 120, pp. 701–722, 1997.

[101] D. Mumford, "Neuronal architectures for pattern-theoretic problems," in *Large-scale neuronal theories of the brain* (C. Koch and J. Davis, eds.), MIT Press, 1994.

[102] A. J. Mussap and D. M. Levi, "Spatial properties of filters underlying vernier acuity revealed by masking: Evidence for collator mechanisms," *Vision Research*, vol. 36, no. 16, pp. 2459–2473, 1996.

[103] J.-P. Nadal and N. Parga, "Non-linear neurons in the low noise limit: a factorial code maximizes information transfer," *Network*, vol. 5, pp. 565–581, 1994.

[104] H. Neumann and W. Sepp, "Recurrent V1-V2 interaction in early visual boundary processing," *Biological Cybernetics*, vol. 81, pp. 425–444, 1999.

[105] W. T. Newsome, M. N. Shadlen, E. Zohary, K. H. Britten, and J. A. Movshon, "Visual motion: Linking neuronal activity to psychophysical performance," in *The Cognitive Neurosciences* (M. S. Gazzaniga, ed.), pp. 401–414, MIT Press, 1995.

[106] D. Nuzillard and J.-M. Nuzillard, "Blind source separation applied to non-orthogonal signals," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, (Aussois, France), pp. 25–30, 1999.

[107] K. Obermayer, H. Ritter, and K. Schulten, "A principle for the formation of the spatial structure of cortical feature maps," *Proceedings of the National Academy of Science, USA*, vol. 87, pp. 8345–8349, 1990.

[108] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomputing*, vol. 17, no. 1, pp. 25–46, 1997.

[109] B. A. Olshausen, "Learning linear, sparse, factorial codes," Tech. Rep. AIM-1580, Artificial Intelligence Laboratory, MIT, 1996.

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

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

[112] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[113] S. E. Palmer, *Vision Science – Photons to Phenomenology*. MIT Press, 1999.

[114] L. Parra, C. Spence, P. Sajda, A. Ziehe, and K.-R. Müller, "Unmixing hyperspectral data," in *Advances in Neural Information Processing 12 (Proc. NIPS*99)*, pp. 942–948, MIT Press, 2000.

[115] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[116] K. Pearson, *The grammar of science*. Scott, 1892.

[117] M. Plumbley, "Adaptive lateral inhibition for non-negative ICA," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, (San Diego, CA, USA), 2001.

[118] M. Plumbley, "Conditions for non-negative independent component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, 2002.

[119] U. Polat, K. Mizobe, M. W. Pettet, T. Kasamatsu, and A. M. Norcia, "Collinear stimuli regulate visual responses depending on cell's contrast threshold," *Nature*, vol. 391, pp. 580–584, 1998.

[120] U. Polat and D. Sagi, "Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments," *Vision Research*, vol. 33, pp. 993–999, 1993.

[121] D. A. Pollen and S. F. Ronner, "Visual cortical neurons as localized spatial frequency filters," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 13, pp. 907–916, 1983.

[122] R. P. N. Rao, "An optimal estimation approach to visual perception and learning," *Vision Research*, vol. 39, no. 11, pp. 1963–1989, 1999.

[123] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[124] R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki, eds., *Probabilistic Models of the Brain*. MIT Press, 2002.

[125] C. E. Rasmussen and Z. Ghahramani, "Occam's razor," in *Advances in Neural Information Processing 13 (Proc. NIPS*2000)*, MIT Press, 2001.

[126] D. Regan, *Human perception of objects*. Sinauer Associates, 2000.

[127] P. Reinagel and R. C. Reid, "Temporal coding of visual information in the thalamus," *Journal of Neuroscience*, vol. 20, no. 14, pp. 5392–5400, 2000.

[128] F. Rieke, D. Warland, R. de R. van Steveninck, and W. Bialek, *Spikes: Exploring the neural code*. MIT Press, 1997.

[129] D. Ringach, "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex," *Journal of Neurophysiology*, vol. 88, pp. 455–463, 2002.

[130] R. W. Rodieck, "Quantitative analysis of cat retinal ganglion cell response to visual stimuli," *Vision Research*, vol. 5, pp. 583–601, 1965.

[131] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, 2001.

[132] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[133] M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco, "On a common circle: Natural scenes and gestalt rules," *Proceedings of the National Academy of Science, USA*, vol. 98, pp. 1935–1940, 2001.

[134] M. S. Silverman, D. H. Grosof, R. L. D. Valois, and S. D. Elfar, "Spatial-frequency organization in primate striate cortex," *Proceedings of the National Academy of Science, USA*, vol. 86, no. 2, pp. 711–715, 1989.

[135] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *31st Asilomar Conf on Signals, Systems and Computers*, (Pacific Grove, CA), pp. 673–678, IEEE Computer Society, November 1997.

[136] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models* (P. Müller and B. Vidakovic, eds.), pp. 291–308, Springer-Verlag, 1999.

[137] E. P. Simoncelli and E. H. Adelson, "Noise removal via bayesian wavelet coring," in *Proc. Third IEEE Int. Conf. on Image Processing*, (Lausanne, Switzerland), pp. 379–382, 1996.

[138] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1215, 2001.

[139] E. P. Simoncelli and O. Schwartz, "Modeling surround suppression in V1 neurons with a statistically-derived normalization model.," in *Advances in Neural Information Processing Systems 11*, pp. 153–159, MIT Press, 1999.

[140] B. C. Skottun and R. D. Freeman, "Stimulus specificity of binocular cells in the cat's visual cortex: Ocular dominance and the matching of left and right eyes," *Experimental Brain Research*, vol. 56, no. 2, pp. 206–216, 1984.

[141] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision.* Brooks/Cole publishing company, 1998.

[142] D. G. Stork and H. R. Wilson, "Do Gabor functions provide appropriate descriptions of visual cortical receptive fields?," *Journal of the Optical Society of America A*, vol. 7, no. 8, pp. 1362–1373, 1990.

[143] S. P. Strong, R. Koberle, R. de R. van Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Physical Review Letters*, vol. 80, no. 1, pp. 197–200, 1998.

[144] N. V. Swindale, "The development of topography in the visual cortex: a review of models," *Network: Computation in Neural Systems*, vol. 7, pp. 161–247, 1996.

[145] D. R. Tailor, L. H. Finkel, and G. Buchsbaum, "Color-opponent receptive fields derived from independent component analysis of natural images," *Vision Research*, vol. 40, pp. 2671–2676, 2000.

[146] K. Tanaka, "Mechanisms of visual object recognition: monkey and human studies," *Current Opinion in Neurobiology*, vol. 7, pp. 523–529, 1997.

[147] F. E. Theunissen and J. P. Miller, "Temporal encoding in nervous systems: A rigorous definition," *Journal of Computational Neuroscience*, vol. 2, pp. 149–162, 1995.

[148] S. Thorpe, "Localized versus distributed representations," in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), pp. 549–552, The MIT Press, 1995.

[149] D. Y. Ts'o and C. D. Gilbert, "The organization of chromatic and spatial interactions in the primate striate cortex," *Journal of Neuroscience*, vol. 8, no. 5, pp. 1712–1727, 1988.

[150] D. Y. Ts'o and A. W. Roe, "Functional compartments in visual cortex: Segregation and interaction," in *The Cognitive Neurosciences* (M. S. Gazzaniga, ed.), pp. 325–337, MIT Press, 1995.

[151] H. Valpola, *Bayesian ensemble learning for nonlinear factor analysis.* PhD thesis, Helsinki University of Technology, 2000.

[152] D. C. van Essen, "Concurrent processing in the primate visual cortex," in *The Cognitive Neurosciences* (M. S. Gazzaniga, ed.), pp. 383–400, MIT Press, 1995.

[153] J. H. van Hateren, "A theory of maximising sensory information," *Biological Cybernetics*, vol. 68, pp. 23–29, 1992.

[154] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex," *Proc. Royal Society ser. B*, vol. 265, pp. 2315–2320, 1998.

[155] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. Royal Society ser. B*, vol. 265, pp. 359–366, 1998.

[156] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[157] R. von der Heydt, "Form analysis in visual cortex," in *The Cognitive Neurosciences* (M. S. Gazzaniga, ed.), pp. 365–382, MIT Press, 1995.

[158] C. von der Malsburg, "Self-organization of orientation-sensitive cells in the striate cortex," *Kybernetik*, vol. 14, pp. 85–100, 1973.

[159] H. von Helmholtz, *Physiological Optics. Volume III. The theory of the perceptions of vision.* (Translated from 3rd German edition, 1910.) Ch. 26. Optical Society of America, 1925.

[160] T. Wachtler, T.-W. Lee, and T. J. Sejnowski, "Chromatic structure of natural scenes," *Journal of the Optical Society of America A*, vol. 18, no. 1, pp. 65–77, 2001.

[161] M. J. Wainwright and E. P. Simoncelli, "Scale mixture of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems 12*, pp. 855–861, MIT Press, 2000.

[162] B. A. Wandell, "Computational neuroimaging of human visual cortex," *Annual Review of Neuroscience*, vol. 22, pp. 145–173, 1999.

[163] B. Wegmann and C. Zetzsche, "Visual-system-based polar quantization of local amplitude and local phase of orientation filter outputs," *Proc. SPIE*, vol. 1249, pp. 306–317, 1990.

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

[165] S. Zeki, *A vision of the brain*. Blackwell Science, 1993.

[166] R. S. Zemel, P. Dayan, and A. Pouget, "Probabilistic interpretation of population codes," *Neural Computation*, vol. 10, no. 2, pp. 403–430, 1998.

[167] C. Zetzsche and G. Krieger, "Nonlinear neurons and higher-order statistics: new approaches to human vision and electronic image processing," *Proc. SPIE*, vol. 3644, pp. 2–33, 1999.

[168] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, no. 8, pp. 1627–1660, 1997.

[169] K. Zipser, V. A. F. Lamme, and P. H. Schiller, "Contextual modulation in primary visual cortex," *Journal of Neuroscience*, vol. 16, no. 22, pp. 7376–7389, 1996.