Department of Computer Science
Series of Publications A
Report A-2008-5

Learning Nonlinear Visual Processing from Natural Images

Jussi T. Lindgren

Academic Dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Hall 5, University Main Building, on Nov. 28th, 2008, at 12 o'clock noon.

University of Helsinki
Finland

ISSN 1238-8645
ISBN 978-952-10-5028-2 (paperback)
ISBN 978-952-10-5029-9 (PDF)
http://ethesis.helsinki.fi/

Computing Reviews (1998) Classification:
G.3,I.2.6,I.2.10,I.4.7,I.4.8,I.5.1,I.5.4

Helsinki University Print
Helsinki, November 2008 (100+52 pages)

Learning Nonlinear Visual Processing from Natural Images

Jussi T. Lindgren
Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
jtlindgr@iki.fi
http://www.iki.fi/jtlindgr/

Abstract

The paradigm of computational vision hypothesizes that any visual function – such as the recognition of your grandparent – can be replicated by computational processing of the visual input. What are these computations that the brain performs? What should or could they be? Working on the latter question, this dissertation takes the statistical approach, where the suitable computations are attempted to be learned from the natural visual data itself. In particular, we empirically study the computational processing that emerges from the statistical properties of the visual world and the constraints and objectives specified for the learning process.

This thesis consists of an introduction and 7 peer-reviewed publications, where the purpose of the introduction is to illustrate the area of study to a reader who is not familiar with computational vision research. In the scope of the introduction, we will briefly overview the primary challenges to visual processing, as well as recall some of the current opinions on visual processing in the early visual systems of animals. Next, we describe the methodology we have used in our research, and discuss the presented results. We have included some additional remarks, speculations and conclusions to this discussion that were not featured in the original publications.

We present the following results in the publications of this thesis. First, we empirically demonstrate that luminance and contrast are strongly dependent in natural images, contradicting previous theories suggesting that luminance and contrast were processed separately in natural systems due to their independence in the visual data. Second, we show that simple cell -like receptive fields of the primary visual cortex can be learned in the nonlinear contrast domain by maximization of independence. Further, we provide first-time reports of the emergence of conjunctive (corner-detecting) and subtractive (opponent orientation) processing due to nonlinear projection pursuit with simple objective functions related to sparseness and response energy optimization. Then, we show that attempting to extract indepen-

dent components of nonlinear histogram statistics of a biologically plausible representation leads to projection directions that appear to differentiate between visual contexts. Such processing might be applicable for priming, i.e. the selection and tuning of later visual processing. We continue by showing that a different kind of thresholded low-frequency priming can be learned and used to make object detection faster with little loss in accuracy. Finally, we show that in a computational object detection setting, nonlinearly gain-controlled visual features of medium complexity can be acquired sequentially as images are encountered and discarded. We present two online algorithms to perform this feature selection, and propose the idea that for artificial systems, some processing mechanisms could be selectable from the environment without optimizing the mechanisms themselves.

In summary, this thesis explores learning visual processing on several levels. The learning can be understood as interplay of input data, model structures, learning objectives, and estimation algorithms. The presented work adds to the growing body of evidence showing that statistical methods can be used to acquire intuitively meaningful visual processing mechanisms. The work also presents some predictions and ideas regarding biological visual processing.

Computing Reviews (1998) Categories and Subject Descriptors:
G.3 Probability and Statistics: multivariate statistics
I.2.6 Learning: concept learning, connectionism and neural nets, parameter learning
I.2.10 Vision and Scene Understanding: representations, data structures, and transforms
I.4.7 Feature Measurement: feature representation
I.4.8 Scene Analysis: object recognition
I.5.1 Models: statistical
I.5.4 Applications: computer vision

General Terms:
vision research, machine learning, statistical modelling

Additional Key Words and Phrases:
natural image statistics, statistical dependencies, independent component analysis, object recognition, feature extraction, feature selection, data transformations

Acknowledgements

I am thankful to my supervisors Aapo Hyvärinen and Esko Ukkonen for their guidance, and to HeCSE, FDK, ALGODAN, HIIT, and PASCAL for funding. In addition I appreciate the concerns that Tapio Elomaa and Jyrki Kivinen showed towards my well-being during my studies, and I am grateful to Ilkka Autio, Christian and Krista Grothoff, Jarmo Hurri, Urs Köster, and Juho Rousu for scientific collaboration. I am also indebted to Patrik Hoyer for reading and commenting on an earlier draft of this thesis, and to Heikki Kälviäinen and Jorma Laaksonen for reviewing a later version. As my research has also relied on more mundane matters, the valuable assistance of the department's technical and administrative support is to be noted – in particular, I fondly remember the help from Pekka Niklander and Päivi Karimäki-Suvanto. For contributions to hallway discussions and banter, I would like to congratulate Lauri Eronen, Michael Gutmann, Mika Inki, Matti Kääriäinen, Tuomo Malinen, Taneli Mielikäinen, Tommi Mononen, Jukka Perkiö, Ari Rantanen, and Pasi Rastas. Yet perhaps the most notable moments of academic delight during my studies were obtained from excursions to the works of R. Brooks, G. Chaitin, S. Lehar, and S. Wolfram. Here, reaching the end of my list of these most esteemed personages, a necessarily glib greeting to Raatis and Naamis is in order; the reader is recommended to imagine an appropriately flippant one. Finally, I thank my family.

Publications of the thesis

This thesis consists of 7 peer-reviewed publications and an introduction reviewing the area of study. In the thesis, the included publications are referred to as Publications 1-7 with the numbering in the publishing order,

1. J. T. Lindgren and A. Hyvärinen. Learning High-level Independent Components of Images through a Spectral Representation. *Proc. 17th International Conference on Pattern Recognition (ICPR)*, volume 2, pp. 72-75, 2004.

2. I. Autio and J. T. Lindgren. Attention-driven Parts-based Object Detection. *Proc. 16th European Conference on Artificial Intelligence (ECAI)*, pp. 917-921, 2004.

3. I. Autio and J. T. Lindgren. Online learning of discriminative patterns from unlimited sequences of candidates. *Proc. 18th International Conference on Pattern Recognition (ICPR)*, volume 2, pp. 437-440, 2006.

4. J. T. Lindgren and A. Hyvärinen. Emergence of conjunctive visual features by quadratic independent component analysis. *Advances in Neural Information Processing Systems 19: Proc. of the 2006 conference (NIPS)*, pp. 897–904, 2007.

5. J. T. Lindgren, J. Hurri and A. Hyvärinen. The statistical properties of local log-contrast in natural images. *Proc. 15th Scandinavian Conference on Image Analysis (SCIA)*, pp. 354-363, 2007.

6. J. T. Lindgren and A. Hyvärinen. On the learning of nonlinear visual features from natural images by optimizing response energies. *Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 1027–1034, 2008.

7. J. T. Lindgren, J. Hurri and A. Hyvärinen. Spatial dependencies between local luminance and contrast in natural images. *Journal of Vision*, 8(12):6, 1-13, 2008.

These publications mainly present empirical, explorative work on learning from natural images. Most of the applied methodologies – such as Independent Component Analysis – are well-established methods that are not radically extended here.

The role of the author of this thesis (in the following, "the author") in the numbered publications is described below. In the publications, all authors participated in discussing the subject and the used methodologies, and in editing the paper.

1. Aapo Hyvärinen proposed to study the spectral representation. The author designed and performed the experiments and wrote the paper.

2. Ilkka Autio designed the proposed method and performed the formal analysis. The author studied the applicability of low-frequency priming in the the context of the method. Ilkka Autio and the author jointly performed experiments and wrote the paper.

3. Ilkka Autio designed the proposed Bayesian selection method and performed the formal analysis. The author designed the proposed heuristic selection method. Ilkka Autio and the author jointly performed experiments and wrote the paper.

4. The author devised the study, performed the experiments and wrote the paper. Aapo Hyvärinen suggested the way to represent products of filter responses as filter response energy differences.

5. The author devised the study and performed the experiments. Jarmo Hurri wrote the paper.

6. The author experimented with different objective functions, designed and performed the experiments and analysis, and wrote the paper.

7. The author devised the study, performed the experiments, and wrote the paper. Jarmo Hurri suggested the experimental design of studying luminance and contrast dependencies by the triplet method.

The main results[1] of the numbered publications are discussed in Chapter 5. Here we summarize the results for convenience:

1. We demonstrate that a statistical model learned with Independent Component Analysis on top of a nonlinear filter response histogram representation is able to segregate the gists of natural scenes to some extent.

2. We present a statistically learned system for object recognition where the computationally more expensive discriminative processing is chosen based on initial, faster mechanisms. We study the low-frequency priming hypothesis in the context of the system.

3. We propose two online feature selection algorithms, one based on Bayesian analysis and the other on heuristics. We evaluate the algorithms on selecting nonlinear visual features for object recognition.

4. We show that Independent Component Analysis, when applied to quadratically basis-expanded natural image data, can learn nonlinear visual processing that functionally resembles angle and corner detection.

5. We study the statistical structure of nonlinear local contrast in natural images by applying Fourier techniques and Independent Component Analysis. We show that in terms of the used statistical methods, the local contrast retains strong similarities to the raw images.

6. We show that statistical minimization or maximization of paired filter response energies over natural image data can lead to emergence of nonlinear filters that exhibit conjunctive (angle and corner detecting) and subtractive (orientation opponency) behaviour, respectively.

7. We study and describe the statistical relationships between local luminance and contrast. These two image properties appear approximately pairwise independent in natural images. We show that this independence does not extend to spatial analysis and hence that independence can not be used as an argument to support the segregation of luminance and contrast processing in a spatial sense.

[1]The usual c-word is avoided here; its proper context can be seen e.g. in Locke (1933).

Glossary

The following technical terms and symbols are common in the introductory part of this thesis. The publications may have slightly different notations.

ICA	Independent Component Analysis
mechanism	an operation that does some processing on information
model	an object with tunable parameters, can also be a density
modelling	selecting a model/mechanism structure, possibly optimizing its parameters by data and constraints
nonlinear	any computation on $\mathbf{x}$ not representable as $\sum_i w_i x_i$
PCA	Principal Component Analysis
SVM	Support Vector Machine
V1	the primary visual cortex
V2, V4	other cortical visual areas

$\|\| \cdot \|\|_2$	Euclidean norm
$\mathbf{A}$	feature matrix, columns are features or receptive field models
det	determinant of a matrix
$g(\cdot)$	some nonlinear scalar function, on vectors applied pointwise
$P(\cdot)$	probability of an event
$p_{\mathbf{x}}(\cdot)$	density function with relation to the distribution of $\mathbf{x}$
s	an output value of a computation, a "response"
$\mathbf{W}$	weight matrix, a filter bank, rows are filters $\mathbf{w}^T$
$\mathbf{v}, \mathbf{w}$	weight vectors, linear filters
$\mathbf{w}^T\mathbf{x}$	dot product, same as $\sum_i w_i x_i$, filtration, "mechanism" example
$\mathbf{x}$	data vector, information, as input
$\mathbf{x}_i$	the i:th row or column vector of a matrix (depending on context)
x_i	the i:th attribute/dimension/variable of vector $\mathbf{x}$
$\mathbf{z}$	data vector from a whitened source (i.e. has identity covariance)

Preface

The research area of this thesis is inherently multidisciplinary and the amount of relevant literature is staggering: the fields under consideration include vision research, computer vision, machine learning, neuroscience and psychology. When applicable and available, I have tried to cite recent review work to provide understandable yet authoritative entry points to the discussed topics. I have also attempted to re-use references in different contexts. In many cases, scores of recommendable reports exist concerning some specific issue. I apologize for the committed sins of omission.

Aside the acknowledgements and this preface, I will use the plural "we" to denote the author, the author and the audience, or the joint authors with respect to the publications, depending on the context. I may also use the plural as a passive – I request the reader to be tolerant towards this.

Contents

1 On studying vision **1**
- 1.1 Vision as computational processing . . . 3
- 1.2 Thesis organization . . . 5

2 Visual processing **9**
- 2.1 Challenges of seeing . . . 10
  - 2.1.1 Ill-posedness . . . 12
  - 2.1.2 Visual variation and natural transformations . . . 12
  - 2.1.3 Semantic concerns . . . 14
  - 2.1.4 Ecological aspects . . . 14
- 2.2 Biological vision . . . 15
  - 2.2.1 Neural processing in the visual system . . . 15
  - 2.2.2 Visual modules and pathways . . . 21
  - 2.2.3 Formation and plasticity of visual function . . . 26

3 Ecology-driven modelling of vision **29**
- 3.1 Historical background . . . 30
- 3.2 Statistics and function . . . 32
  - 3.2.1 Natural image statistics . . . 32
  - 3.2.2 Statistical models of visual input . . . 35
  - 3.2.3 Statistical models of visual function . . . 39
- 3.3 Are there independent mechanisms in perception? . . . 43

4 Statistical modelling, methods, and visual data **47**
- 4.1 Modelling with different objectives . . . 47
  - 4.1.1 Independence objective . . . 48
  - 4.1.2 Response energy objective . . . 50
  - 4.1.3 Object recognition objective . . . 51
  - 4.1.4 Feature selection objective . . . 54
- 4.2 Intricacies in statistical learning . . . 56
  - 4.2.1 Local optima . . . 56

4.2.2 Overfitting and model selection 57
4.2.3 Further issues 58
4.3 Natural image data and its preprocessing 59

5 Learning visual processing **65**
5.1 Low-level statistical dependencies in images 65
5.1.1 Structure of local contrast 66
5.1.2 Relations between local luminance and contrast . . . 67
5.2 Quadratic processing 68
5.2.1 Quadratic processing by ICA 69
5.2.2 Quadratic processing by energy optimization 71
5.3 Simple priming mechanisms 72
5.3.1 Low-frequency priming 73
5.3.2 Gists of visual scenes 76
5.4 Online feature selection 77

6 Conclusion **81**

References **85**

Reprints of the original publications **101**

Chapter 1

On studying vision

Perceiving visual scenes seems relatively effortless to us. Our brains interpret our visual environments with seemingly little delay, turning the received barrage of photons into perceptions of our surroundings. This rapid, unconscious interpretation is what allows us to *see* the world conveniently as shapes, objects, surfaces, patterns, colours and so on.

The introspective feeling that seeing is easy is misplaced. As an invitation, the reader is referred to Figure 1.1 that could well be taken as an artist's illustration of at least three different vision-related issues that we will discuss in the remainder of this chapter. First, the illustration can be used as a teaching example demonstrating how difficult it is to model the processes of seeing. Second, the illustration could represent the brain activity as it becomes our perception of the visual world. Third, the illustration could portray the happy cross-disciplinarity of vision research. We will shortly explain these three points, and hope that this thesis will further convince the reader how fitting the suggested allegory is.

First, consider how Figure 1.1 reflects the problems of seeing. If time is spent thinking on what might be interesting in the scene, we might formulate these interests as questions posed to the visual apparatus. These questions could be such as "what is shown in the image?", "where is that place?", "what objects are present?" and yet more specific ones like "is that Harry in the middle left?", "is there anyone drowning in the image?", or "if you see something like that, should you run?". Or, we could consider tasks involving other images, such as "examine some additional images and find those that resemble this one". It should be apparent that it is difficult to mathematically specify how the given image in Figure 1.1 should be used to address such questions, as the challenges the task poses may range from simple image processing to meaningful incorporation of human cultural semantics. Subsequently, should there be a model to acceptably answer

these (and other equally arbitrary) questions for images, the system would practically pass a Turing test (Turing, 1950) customized for images: an acceptable performance for the system would be that a human interrogator would not be able to tell if the answers are returned by another human or an artificial model. Hence we suppose that cognition and high-level visual tasks are not ultimately separable (see e.g. Chalmers et al. (1992); A. J. Bell (1999); Thelen et al. (2001), for a contrary view see Pylyshyn (1999) and the heated peer commentary). Instead we accept that human-like seeing may be a complicated and convoluted effort, with the required machinery not necessarily simpler than the human brain.

Our second point was that Figure 1.1 can be used as an allegory of system level neural visual processing in the brain. Making a convincing case of this is not entirely possible without a proper overview of the biological brain. At the moment we have to content ourselves by pointing out that the cortical processing in animals is performed by diverse sets of parallel elements and areas of computation that influence each other (Gilbert & Sigman, 2007). These entities may perform at different latencies, having an order of processing that may rather be cyclic instead of stagewise or serial (Bullier, 2004). Further, these elements and areas may use different kinds of signaling strategies (Krahe & Gabbiani, 2004) and codes (deCharms & Zador, 2000) to relay the results of their operation. Areas that are considered segregated on the cortex may be devoted to separate aspects of the visual input (Zeki, 1978; Livingstone & Hubel, 1988), but also several visual aspects can be considered by a single cortical area (Ts'o & Roe, 1995). Possibly different visual properties can be represented by the same computational element at different points in time after the stimulus onset (Roelfsema et al., 2007). To make things even more interesting, to some extent this visual machinery can change its general operation over time (Kaas et al., 1990; Kohn, 2007). The whole process of seeing, then, somewhat resembles the performance of a well-tuned orchestra – or the parallel baying from a well-behaving zoo (although provocative, the flavour of this idea is not new, see e.g. Minsky (1986)). Together, the processing elements make up a system of complicated interactions, analogous to one in Figure 1.1.

Finally, as our third point we suggested that Figure 1.1 could illustrate the research community that studies vision. Given that vision may be studied on many partially overlapping levels of abstraction, including molecular, biochemical, neural, computational, psychological and cultural levels (see e.g. the scope of Palmer (1999)), it is not surprising that the work towards understanding vision is very multidisciplinary, with contributions

coming from diverse research areas including neurophysiology, vision research, brain research, neurophysiology, cognitive science, psychology, biochemistry, physics, mathematics, statistics, computer science, artificial intelligence, machine learning, and even economics. Subsequently these fields have brought their own research traditions and tend to have characteristic scopes for the questions they are addressing, possibly incomprehensible to a researcher from a different background[1]. Often there are contradicting results regarding vision even from inside a single discipline, but the different areas also fruitfully interact with each other, and the situation can be summed up as not being completely unlike the one shown in Figure 1.1.

1.1 Vision as computational processing

In this thesis, vision is studied on the abstraction level of information transformations. Central to this idea is *a model system* that receives natural visual input, and performs transformations on the input to produce useful behaviour (some kind of desired functionality). These transformations that we call *visual processing* are assumed to be partly fixed (roughly corresponding to optimization that has been done by evolution) and partly learned from exposure to visual data (corresponding to plasticity during the lifetime of an organism). In this setting, the exposure to data and some specified goals are used to adapt parameters of the processing mechanisms, i.e. the mechanisms may belong to some fixed function classes, but the function parameters are learned from experience to attain the goals. A major part of this thesis concerns learning different functions from visual data. We are also interested in statistical properties of such data, as learning, statistics and statistically meaningful behaviour are closely connected.

In the scope of this work, visual processing is modelled to operate on the abstraction level of algebra on real numbers, vectors, and functions of such. Combined and tuned appropriately, these models represent and process abstract information, typically in arbitrary units. Aspects of lower levels, for example how real neurons are formed from molecules or how they actually transmit information or produce an electric discharge, are taken to be below the used level of abstraction (but these lower-level issues may still be important for higher-level function, see e.g. A. J. Bell (1999)). Likewise, it should be emphasized that in this thesis we do not propose

[1] I recently participated in a workshop on "neuroinformatics", and saw a poster that used the abstraction level of dynamical systems and biochemistry in neuronal modelling. The presentation was quite beyond me. Likewise, had I asked how the proposed model helped in some high-level functionality, I might have seen a blank stare matching my own.

new models of neurons, nor are we proposing biologically plausible learning algorithms. Neither are we presenting new, improved mechanisms for computer vision. On the contrary, it could be claimed that in such regards, the mechanisms studied in this thesis do not incorporate all the complexity of current systems level biology, nor do they meet the finesse of the state of the art computer vision systems. This is mainly because of traditions in philosophy of science suggesting we should not complicate issues needlessly (an idea often known as *Occam's razor*, after William of Ockham, c. 1288 - c. 1348, later elaborated by several others, e.g. Mach (1882)), but it is also a matter of practical tractability. Hence, as we study computation and phenomena, we pick the most simple and feasible computation we can think of, given that it produces or verifies some of the phenomena we are interested in. Subsequently, with some control on the complexity, we can more easily reason about the limitations of the approach, think of possible extensions and consider resemblances to natural processing.

The main underlying hypothesis in our setting is that vision is amenable to computational simulation (as in e.g. Churchland and Sejnowski (1992)). Should this *computational hypothesis* be true, it would mean that mathematical descriptions can mechanically explain and replicate the transformations from the environmental visual data to the eventual animal perceptions and behaviours. To chart the validity of the computational hypothesis, we can in principle search among the multitude of mathematical descriptions (models) by requiring that the mathematical description, when simulated, can produce appropriate behaviours on the given visual inputs. In this thesis, we perform this search for suitable models by adapting the model system parameters to natural visual data and some behavioural objectives.

The benefit of using behavioural objectives and a large amount of natural image data to guide selection of mathematical models is that it allows us to study and estimate model mechanisms of visual function without having to resort to neurophysiological experimentation. Although the results can be compared to neurophysiological data, the models can also be evaluated with relation to the quality of behaviour they exhibit. Then, this approach to studying vision can be taken to combine natural image statistics research (Simoncelli & Olshausen, 2001) with the more goal-oriented methodology from computer vision and machine learning research. In this thesis we call this combination the *ecology-driven approach*, and we will elaborate on the connotations of the name in Chapter 3.

Should computational modelling of high-level vision succeed, the scope of applications would be enormous. Already in the eighties, methods from the machine vision community were in production use in tasks such as ma-

chine inspection of factory products (Robinson & Miller, 1989). At the time, this success was made possible by the tightly controlled factory conditions. Later, statistically fitted neural network models could be deployed in e.g. cheque recognition (LeCun et al., 1998), a problem that is more challenging due to the letters on the cheques having been written by humans. Currently, computer vision methods are advanced enough to be deployed in even more demanding settings, such as in autonomous planetary exploration vehicles (Matthies et al., 2007). As the research progresses, old applications such as autonomous robotics, face recognition and biometrics are expected to become more successful, and yet new applications may become feasible. One example of a future application is a personal assistant for browsing, filtering, searching, and recommending visual media; this problem can be taken as particularly demanding as semantics and feelings affect our judgments regarding visual content. Also, as the understanding of the visual processing mechanisms in humans and primates grows, neural interfaces transmitting sensory information directly in and out of the brain may greatly improve, allowing revolutions in e.g. entertainment, prostheses development, and quite possibly in society in general.

In the scope of this thesis, it should be prudently admitted that the results we present here do not trivially alleviate such future applications as described above. Here, our results are related to learning simple and abstracted mechanisms of visual processing from natural visual data. In particular, we add to the surmounting evidence that meaningful visual features and processing can be learned from the natural visual data, and we explore how including certain nonlinearities to the processing affects the emerging mechanisms. Our analyses of the models and the input data enlargens our understanding of the complex statistical structure of the visual input, and thus may not only help in the efforts to realize visual processing in machines, but also in understanding biological vision.

1.2 Thesis organization

The rest of this thesis is organized as follows. In Chapter 2 we outline our view on visual processing and describe some of the problems that are currently understood to be associated with it. Next, we review some of what is known of the operation of natural visual systems and how they process visual information. In Chapter 3 we describe the modelling approach used in this thesis, along with its historical connections. In Chapter 4, we give a review of the statistical estimation methods and learning objectives we have used, along with an account of some of the challenges related to the

application of such methods. We also discuss the properties of the used visual datasets in the same chapter. Then, in Chapter 5 we overview the technical content of this thesis with additional discussion and hindsight that was not part of the original publications. Finally, Chapter 6 concludes with a speculative outlook at possible future directions and developments. The main technical content of this thesis is appended to the end as reprints of the original publications.

We recommend the following reading order: readers familiar with vision research and machine learning should skip to the publications at the end of this thesis and then return to read Chapter 5. For other audiences, Chapter 2 and Chapter 3 provide introductory material. The publications at the end could be glanced next, and should the technical learning methods require some additional explanations, Chapter 4 provides a starting point. Although Chapter 5 reviews the publications of the thesis, the provided discussions may be more understandable after studying the publications. The last chapter, Chapter 6, concludes in a nontechnical manner.

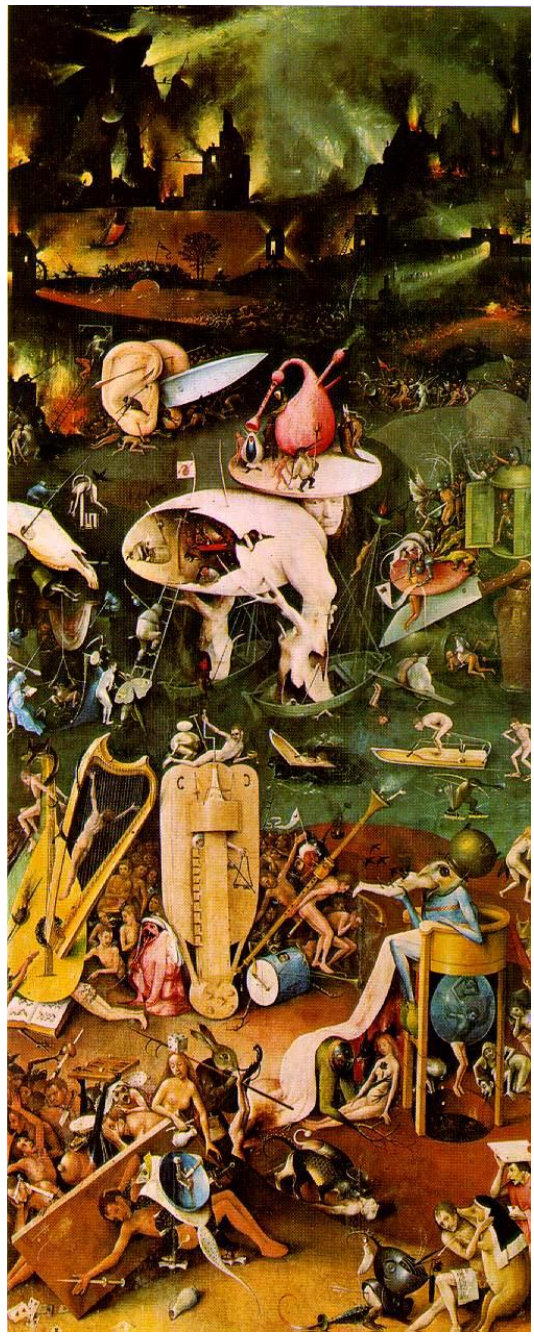

Figure 1.1: Hell, the right panel of Garden of Earthly Delights, by Hieronymus Bosch, ca. 1504. Currently in Museo del Prado, Madrid.

Chapter 2

Visual processing

"Sans [...] le canard de Vaucanson vous n'auriez rien qui fit ressouvenir de la gloire de la France." – Voltaire

We start our account of modelling-based vision research by recalling the underlying fundamental hypothesis. This hypothesis is that mathematical models can be constructed that are functionally similar to the biological visual systems, albeit in simulation. To put it another way, it is expected that if the mathematical mechanisms are designed appropriately, they can replicate behaviour at some required level of analysis. For example, a model of a real neuron could be expected to predict the responses of the real neuron when both are subjected to the same stimulations. A model of a network of such neurons could be required to reproduce the dynamic behaviours that such networks have in biological systems. Further, a yet higher-level model might be formulated to perform a task like object recognition.

It is important to note that as all such models are essentially evaluated on a (digital) computer, it follows that the mathematics involved are necessarily *mechanistic*. Should such simulations be able to replicate arbitrary visual function to any required level of precision, this would mean that vision in itself is *computable* in the sense of Turing computability[1].

Here we accept these underlying premises for now, and consider *visual processing* as a process where photons are caught from the environment to form measurements that are further transformed by the visual system to support ecologically useful behaviour. It is of some interest how to

[1] In general, "computable" should not be confused with "computational". Although in this thesis we do work in the paradigm of "computational science", i.e. use computing and large datasets as tools for scientific discovery, here this setting has also the consequence that if the visual functions under study can be eventually simulated by computation, we have shown them to be "computable".

characterize this process. Should the characterization be laid out in terms of chemistry, or perhaps physics? In Chapter 1 we mentioned that vision research is a multidisciplinary effort. This is true, but when the goal of the research is to provide a mathematical model that takes some form of visual input and makes some computations on it to generate an output, on a philosophical level we converge to a single discipline of *information processing*. This is because mathematics work by manipulating abstract objects such as values or concepts, that is, *data*. Mathematical models of vision can never directly manipulate some mysterious quantas of nature, but only their abstracted representations in the form of some input data. The data is no less data, whether it contains measurements describing photons, concentrations, voltages, or time-series of electric pulses. Similarly, the only thing mathematical models ever output is information. It follows that computational models as presented by vision research (Marr, 1982; Palmer, 1999) and theoretical neuroscience (Churchland & Sejnowski, 1992; Dayan & Abbott, 2001; Eliasmith, 2007) are efforts in designing mechanisms for information processing and transformation.

But is this view appropriate? If vision (and more generally, cognition) would be amenable to mechanistic modelling in the sense of classic mechanics and such mathematical descriptions as can be simulated on computers, then very little separation would be left between animals and machines. Interestingly, for those who would prefer bio-mysticism over the mechanistically definable, at least a few other possibilities remain. One is that some functionality would be amenable to mathematical description, but the description being necessarily such that it can not be evaluated on a Turing machine in a reasonable time (see Copeland (2000)), for example due to hypothetical involvement of quantum phenomena (Penrose (1994, 1997)). As these issues do not seem to greatly concern the mainstream neuroscience (see e.g. Litt et al. (2006)), we feel justified to leave these issues to future philosophers and move on to overview the challenges involved in processing of visual information.

2.1 Challenges of seeing

The process of seeing classically starts from the stage where the properties of the environment are measured. In this, visual systems and cameras are in the same line of business: both use photons from the environment as their input. The eyes and the camera, both in their own way, measure the densities and wavelengths of the photonic bombardment from the environment. Thus in essence the early retinal image in the eye, a photographic

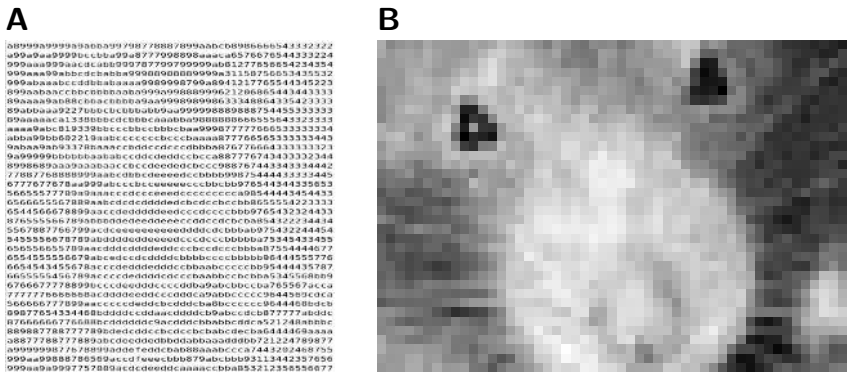

Figure 2.1: **A)** A greyscale picture of 16 shades represented as hexadecimals. **B)** The same picture in a more ordinary representation as shades of grey. Although the information is roughly the same in both images, the character-composed image seems difficult to interpret for the human visual system. On the other hand, the character image on the left is analogous to the initial numeric representation that computers and digital cameras use for greyscale images.

image, and an image on a computer can be taken to be similar as they count the amount of light at different positions across a spatial map, as well as incorporating information about the wavelengths of the light (colours). These low-level measurements are then collected continuously over time by a (video) camera or the retina to produce a stream of visual information for further analysis (for details regarding this sampling of visual information, see e.g. Sonka et al. (2007)).

In the nineteenth century, the replication of the visual scene (as in *camera obscura*, a simple photographic device involving a painter) was thought to be all that there is to seeing. As we now know, a photograph of a scene understands very little of it. Figure 2.1 shows that simply replicating the scene contents does not equal perception in human vision either: although Figure 2.1A can be seen well, its symbolic representation does not allow the later human visual processing to make sense of it. On the contrary, Figure 2.1B allows useful perceptions, while it has roughly the same information as Figure 2.1A. Now, given that images represented appropriately can lead to useful percepts, what are the processes that transform the grey-level image into a perception, and what kind of challenges do they face? In Figure 2.2 we list some of the grand challenges related to visual processing, and we will discuss them briefly in the following.

- Core issue
  - Ill-posedness of the inversion problem
- Visual variation and natural transformations
  - Arbitrariness in location, orientation, and distance of "things"
  - Intra-class diversity of visual properties of "things"
  - Variability due to illumination, shadows, occlusions, and colour
- Semantic concerns
  - Visual interpretation may require "understanding"
- Ecological aspects
  - Requirement for quick and prioritized processing
  - Requirement for plasticity and learning

Figure 2.2: Grand challenges that visual systems face in the natural environment as perceived by the author.

2.1.1 Ill-posedness

A central challenge of vision is that both retinal and camera images are two-dimensional projections of the three-dimensional external reality (for a description of the optics involved, see e.g. Palmer (1999); Sonka et al. (2007)). The external reality cannot uniquely be reconstructed from only two such projections – many different states of reality can map to the same image, or to two stereo images. One simple example is to think of one object occluding something from our sight. Although we can make some more or less conscious inferences regarding how the world should look behind the occluding surface, in practice any number of different things could lurk there. This problem does not have a unique solution; the best any system can do is to make educated guesses about the unseen parts of the world, based on its prior experiences and inbuilt biases. Collecting and consolidating such experience into a model system clearly is a problem in itself.

2.1.2 Visual variation and natural transformations

Another problem in seeing is that perceptually similar images may not be similar in terms of the input representation and such metrics as are typically

Figure 2.3: According to Euclidean distance applied in the greyscale pixel space, the flanking images at the two sides are closer to the uniform grey image in the middle than to each other.

considered in elementary mathematics. In terms of linear algebra, an image – such as the one on the retina – is a point in a multidimensional space. A digital greyscale photograph of 1024×1024 variables (in the case of images, the variables are called *pixels*) is a point in a space of roughly a million dimensions. Now imagine an object of interest to be first positioned on the left in the image, and then on the right in the image. Although the objects of interest are the same, if we consider metrics such as the Euclidean distance, these two images are worlds apart. This is illustrated in Figure 2.3: if the Euclidean distance is used to measure the closeness of the images, the left and the right images are closer to the blank grey image in between than to each other. Not only changes in position, but also other *natural transformations* such as changes in rotation and distance of an object of interest are enough to make the traditional metrics in the input space return distance estimates that feel incorrect to human intuition. Similar effects can be attained from the classic metrics by changing lighting conditions or adding shadows.

The issue is that depending on the viewing conditions, the same object of interest may really appear very different on the level of the spatial light intensity configurations that the system receives as input. It is known that these differences may pose difficulties for artificial systems (e.g. Pinto et al. (2008)), whereas the human visual processing can often discount the confounding factors and identify the object in question. A lower-level example can be given from the context of colour processing: an object of a certain colour is necessarily represented differently to the retina under different lighting conditions, yet the human visual system is often able to infer the correct object colour; this phenomenon is called *colour constancy* (Land & McCann, 1971). However, human vision is not totally *invariant* to all natural transformations, but only to some extent (Kingdom et al., 2007; Kravitz et al., 2008), and one challenge for modelling human-like vision is to achieve similar invariances in a model system.

2.1.3 Semantic concerns

Suppose for a moment that we had an artificial visual system that would take an image and always represent the objects of interest in some standardized, object-centered coordinate system where the object representation could be easily matched against stored memories of objects, without having to worry about issues such as position and lighting. Yet this would not ultimately solve the problem of e.g. object recognition. The textbook example is the recognition of chairs. Imagine you had memorized a set of prototypes of chairs. These might already look wildly different, but nevertheless, by themselves this collection would not explicitly capture or highlight the "semantic" rule that a chair is something that can be sat upon (Gibson, 1979). Hence, we can not easily disentangle all visual function from cognitive, semantic issues. This was already recognized by Koffka (1935), though it is commonly – and perhaps conveniently – forgotten by many modellers working in the modern computer vision and learning-based vision paradigms. Here, although such concerns have not been forgotten, we have to admit that semantic issues are also outside the scope of the models examined in this thesis.

2.1.4 Ecological aspects

In nature, visual systems do not exist in conditions where leisure rumination could always be performed on the scene before acting on it. Instead, evolutionary pressure prompts the approaches to be fast: threats need to be recognized quickly to be able to react appropriately. In such cases, there may not be time for a serial processor to run sluggish comparisons between thousands of stored prototypes to see if the currently seen visual element is dangerous. It seems also reasonable that vision does not need to be equally fast for everything, and nor it is. Instead, natural visual processing seems to prioritize important aspects such as threats (Fox et al., 2000) and sexual saliency (Anokhin et al., 2006). Hence speed and biases of processing can be seen as constraints to natural systems and subsequently reasonable requirements for model systems as well.

Another issue related to ecological aspects is adaptability. In some regards, the visual processing mechanisms are encoded in the genetic code (DNA, deoxyribonucleic acid), and in others, the mechanisms are learned for each animal anew. Although the interactions and divisions between nature and nurture are not yet completely understood, it is clear that adaptive approaches may have evolutionary edge in being able to learn from experience and incorporate new information: the natural environment and its

events are not static or deterministic from the viewpoint of the animal. In humans, vision seems especially plastic. For example, object identities are learned during the lifetime, and this learning may require only a single glance at the object. Whatever the model systems are, eventually they also should be amenable to quick learning and adaptive visual behaviour to match their natural counterparts.

In this thesis, learning different types of visual processing from the natural visual experience is a central subject that will recur in the sections and chapters to come.

2.2 Biological vision

We have seen that visual processing can be an interwoven affair of different, complex issues. At the moment there is no unified, accepted theory that could describe vision and allow vision to be simulated in general. However, most animals implement some kind of vision, and at the level relevant to the species in question, these natural systems can handle the grand challenges we listed in Figure 2.2. Although it remains an open question how these systems precisely work, it is clear that investigating biological vision is one way to shed more light on the required processes[2], just as studies in artificial vision can help to understand biological processing and the problems it faces.

In this section we give a brief overview on the current opinion regarding the early visual processing in biological systems. Although the research we cite is based on studying a variety of species – such as cats, monkeys, and humans – on the level of our account, these differences can be taken as unimportant, as the mammalian mechanisms of vision tend to be made up from qualitatively similar components. Here we consider these natural visual mechanisms largely from the viewpoint of data transformations, i.e. how they transform and route visual information in the early visual processing. We also review some propositions from the literature regarding the functional significance of such transformations.

2.2.1 Neural processing in the visual system

After the influential works of Ramón y Cajal, c. 1852 - c. 1934, the classic building blocks of computation in the brain have been thought to be the cells called *neurons*. According to the *neuron doctrine* that Ramón y

[2]In Chapter 3 we will describe a complementary approach where natural environment is studied to provide suggestive answers to questions about vision.

Cajal proposed (see e.g. Bullock et al. (2005)), neurons perform the bulk of the signal processing in the brain regardless of the area of the brain in question. A single neuron is thought to perform only a relatively simple computation, whereas higher-level functionality is considered to arise from the joint interaction of interconnected neurons of diverse types. This practically amounts to saying that every mental activity is in correspondence with some neuronal computation, an idea often attributed to McCulloch and Pitts (1943).

For convenience, we show a drawing of a neuron in Figure 2.4A, where the typical neuronal parts are clearly visible. Figure 2.4B shows a corresponding artificial neuron model, as to be described on page 18. In Figure 2.4A, the blob in the middle is called *soma*, and *dendrites* are the spindly fibers that neurons "receive" their inputs with. The neuron relays the results of its processing through the *axon*, which is the protruding spike extending upwards from the soma in Figure 2.4A. The axons allow neurons to communicate with other neurons (but possibly also with themselves) through connections called *synapses*. The actual information is transmitted via the axon by the neuron *firing* a time-series of electric, binary discharges called *spike trains* that get converted into chemicals at the synapses. For details, see e.g. Churchland and Sejnowski (1992); Dayan and Abbott (2001).

Neural coding and receptive fields. One way to attempt to understand the computation that a neuron carries out is to provide the studied neuron some input (possibly indirectly) and see how its spike trains are affected. But how to measure this change in the firing? One longstanding possibility is that the number of spikes as averaged over some time window is how neuron outputs represent information (Adrian & Zotterman, 1926), suggesting the relevant measurements to be *firing rates* or firing frequencies. The corresponding representation that a neuron creates for its inputs is in this case called a *rate coding* scheme (for a review see e.g. Dayan and Abbott (2001)). Subsequently, to see how the firing is affected by stimulus change, we could look at the changes in the firing rates. Yet this is by no means the only possibility of how the spike trains could code for information; for example, in the more recent idea of *temporal coding* it is thought that the amount of time passed between spikes may also be an informative quantity. A spike train, and its rate- and temporal codes are shown in Figure 2.5.

As stated, visual neurons can be studied in the rate coding paradigm by displaying stimuli to the retina and measuring changes on the neural

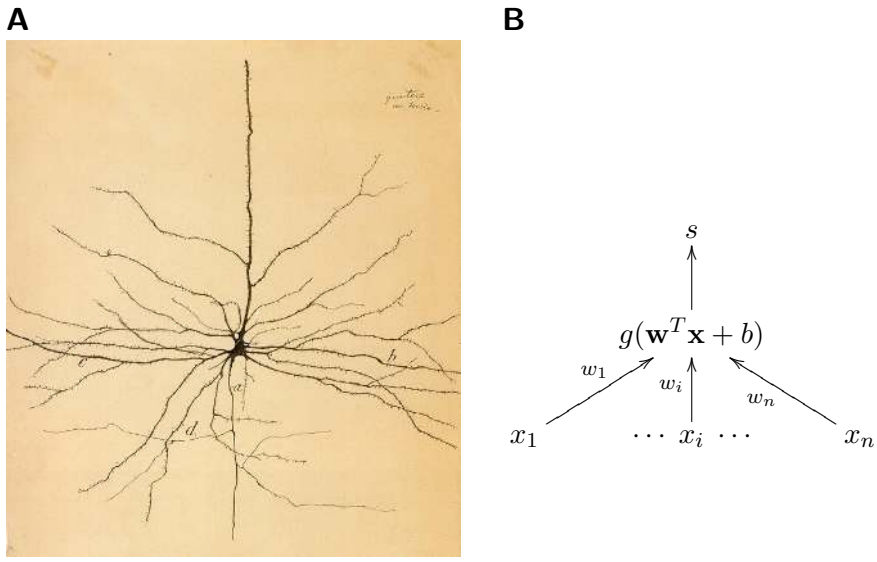

Figure 2.4: **A)** A neuron as drawn by Ramón y Cajal, ca. 1899. The extensions around the blob (soma) are dendrites, and the long upwards-poking extension is the axon. **B)** A schematic of a simple artificial neuron model reading inputs x_i and returning output value $s = g(\mathbf{w}^T\mathbf{x} + b)$.

firing rate. Although close-to-zero firing rate does not entail that a neuron was not participating in the encoding of the currently shown stimulus (Churchland & Sejnowski, 1992), examining the rate-coding responses of single neurons has led to some practical characterizations of their input/output relationships. In such studies it was found that visual neurons might respond only to modulation at some part of the visual field, and in a literal sense this spatial region was then labelled the neuron's *receptive field*. Early studies (e.g. Hubel and Wiesel (1959)) proposed that modulation of light intersecting the receptive field is what alters the firing rate of the neuron, whereas modulation outside the receptive field has no effect (but see also Bair (2005)). In more recent literature, the receptive field is taken to denote the shape of the favourite input stimulus for the neuron, i.e. the stimulus that coaxes the highest firing rate from the neuron. To illustrate the kind of stimuli that simple visual neurons might prefer in the rate coding setting, some receptive field models are shown in Figure 2.6. In the figure, black and white code for inhibitory and excitatory effects of a dot of light at that spatial location, respectively.

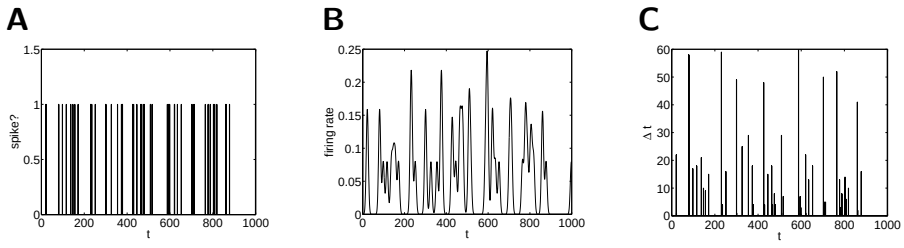

Figure 2.5: Interpreting a neuron's output. **A)** An artificial spike train from a thresholded Poisson process. **B)** Counting the firing rate (frequency) in a localized time window estimates a rate code. Here a Gaussian weighting window was used to linearly filter the spike train in A. **C)** In temporal coding, the time elapsed between subsequent spikes carries relevant information. In this plot, a mark at time t denotes the number of time units (here discrete) that passed between spike at t and the previous spike. It is assumed that neurons have non-negligible recharging times, and thus a zero-height marking at time t can be used to denote that no spike occurred at that point.

Neuron as a function. Discussion in terms of receptive fields provides a high-level abstraction of how light may affect the output rates of the simplest visual neurons, but how exactly do the neurons compute their responses? Instead of getting lost in the elaborate swamp of the current opinion, here we illustrate one possible process by showing a simple, classic model of neural computation of an *integrate-and-fire* neuron (McCulloch & Pitts, 1943). This model, also known as *perceptron* after the learning algorithm of Rosenblatt (1958), computes its response rate s to input $\mathbf{x}$ as

$$s = g(\mathbf{w}^T\mathbf{x} + b), \tag{2.1}$$

where the magnitude of each coefficient in $\mathbf{w}$ represents the *synaptic strength* of the corresponding neural connection, and the sign of w_i encodes whether the connection i is excitatory or inhibitory (we do not explicitly consider interneurons here). Unlike in real neurons, depending on the nonlinearity $g()$, the response rate s may be negative. In that case the response may be interpreted as a difference to some base level of firing, or the model may be taken to model two neurons in one. If $g()$ is half-wave rectification, then the output is always non-negative, and the bias term b has the interpretation of representing the firing threshold. The inputs received in $\mathbf{x}$ by the neuron may be output rates of other neurons.

It can be seen that in the context of the model of eq. (2.1), the vectors $\mathbf{w}$ are practically linear *filters*, and the whole model implements simple

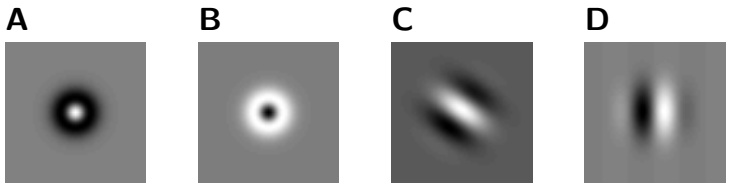

Figure 2.6: The receptive field of a neuron is the spatial area of the visual field where modulations can cause the neuron to fire. Commonly the term also denotes the visual shape the neuron responds most actively to. In these images, black corresponds to inhibitory effects and white to excitatory effects that spots of light have on the firing rate when they are presented at the corresponding spatial locations. Spots of light introduced at the base level (grey) locations have no effect on the firing rate in these models. The units are arbitrary, and these simple receptive field models do not incorporate possible spatiotemporal aspects. **A-B)** Two centre-surround receptive field models, an ON-centre, OFF-surround receptive field, as well as an OFF-centre and ON-surround one. These models would respond strongly to white and black spots, providing that they do not extend to the surround. **C-D)** Oriented receptive field models. The receptive field in C responds strongly to a diagonal white bar if its orientation matches the main axis of elongation of the receptive field. The field in D would prefer a vertical step edge.

nonlinear filtering. In general, all the model neurons having the general form of eq. (2.1) are called perceptrons. One such perceptron was shown schematically in Figure 2.4B. As visual images can also be represented as vectors $\mathbf{x}$ by reshaping them (i.e. $n \times n$ pixel matrix becomes a vector of n^2 dimensions) and the same can be done to spatial filters, any of the receptive fields of Figure 2.6 could be plugged into eq. (2.1) as $\mathbf{w}$ to get a simple model of neural computation that can be simulated numerically. Given a monotonously increasing $g()$, this model would then predict a steady-state rate-coding response s to any stimuli $\mathbf{x}$, with a property that among all stimuli of fixed norm, the receptive field $\mathbf{w}$ itself would be the stimulus $\mathbf{x}^*$ to give the highest response s^*.

Neural networks and model plausibility. If perceptrons are layered into networks, universal function approximators are attained (Hornik et al., 1989). This has the consequence that in principle any computation that can be carried out by a function can be approximated by a layered network of perceptrons (or functionally equivalent real neurons), and subse-

quently more complex computations could be achieved by assembling such simple components as parts of more complex networks. This approach is often called *connectionism*. However, even if models similar to perceptron are used in computational studies (e.g. Serre, Oliva, and Poggio (2007)), it should be kept in mind that reducing neurons to perceptrons is a gross simplification. One reason is that the only aspects that vary in perceptrons are the parameters $\mathbf{w}$ and b and the used nonlinearity $g()$. In contrast, real neurons can vary in several more dimensions: some classification schemes suggest that mammalian retinas alone have approximately 55 different types of neurons (Masland, 2001). In addition, the influence that neurons can exert on one another can be much more complicated and nonlinear than what is possible with eq. (2.1). The cortical connections also include recurrencies.

Just as the perceptron model of a neuron can be said to be convenient but an exaggerated simplification, the rate-coding idea that such models typically implement has also been under recent debate, and not only from the direction of temporal coding that we mentioned earlier: recent findings suggest that some neurons fire in different manners such as in bursts (Krahe & Gabbiani, 2004), and instead each neuron encoding information independently, behaviour such as synchronous firings in neural populations have been observed (Gray, 1999; Jermakowicz & Casagrande, 2007). Yet more theoretical proposals exist claiming that real neurons might not signal the stimuli presented, but the amount of difference of what is seen to what is expected to be seen (Rao & Ballard, 1999; T. S. Lee & Mumford, 2003). Basically the rate coding idea (and especially that of steady-state models) is convenient for mathematical modelling as it often allows cheap computer simulation and tractable parameter learning. Subsequently, equating neurons with some simple fixed functions such as the one of eq. (2.1) remains tempting. This simplification would be more acceptable, if, for a given input, a real neuron would always return the same firing sequence or rate, just as a function does. However, simple high-level phenomena suffices to illustrate that visual processing and neuronal operation do not work as static functions do: looking at a *bi-stable image* – such as the Necker cube shown on page 44 in Figure 3.4B – demonstrates that it is commonly difficult for a human viewer to hold a fixed interpretation of such images for long. It follows that the substrate of perception is not stable or static in the manner that response of eq. (2.1) would be given any representation of the Necker cube as $\mathbf{x}$.

2.2.2 Visual modules and pathways

The classic view of visual processing has been that of a conveyor belt where information is processed and modified by stages of neurons, where each stage does some particular kind of processing before passing the information to the next stage (e.g. Marr (1982)). Although this *feedforward view* seems to be appropriate in some situations (see e.g. Serre, Oliva, and Poggio (2007)), a growing body of recent research embraces the contrary view that visual processing is not a stagewise pipeline with a beginning and an end, but that it may resemble a cyclical process (e.g. A. J. Bell (1999); Rao and Ballard (1999); T. S. Lee and Mumford (2003); Grossberg (2003); Bullier (2004); Olshausen and Field (2005); Gilbert and Sigman (2007)). Still, convincing evidence exists that brain is not a confounding concoction of homogeneous porridge, but that it can be meaningfully subdivided in different ways, for example into visual areas (Essen, 2004). Commonly at least the gross anatomical units such as the retina, the optical tract and the thalamus are agreed to exist as anatomical entities in mammals. These three parts make up a major pathway of visual information from the eyes, and they are shown for clarity in Figure 2.7. Receiving input from the retina, lateral geniculate nucleus (LGN) in the thalamus further feeds into V1 (primary visual cortex), the first cortical visual area at the back of the head. But not even this initial stream is a purely feedforward information queue from the eyes to V1: according to some measurements, only 5-10% of the total inputs to the thalamus are directly from the retina, whereas a larger amount comes as feedback inputs from the cortex (Sherman & Guillery, 2002).

However, although the brain can be subdivided into components such as the retina, LGN, V1, and further areas, this picture is a compromise, as these areas are not necessarily devoted to a single function, nor do they operate independently. Regarding the first issue, evidence is starting to accumulate that in V1, the same neurons may perform different kinds of computations, where the nature of the current computation may depend on how much time has passed since the stimulus onset (Roelfsema et al., 2007). On the level of cortical areas, there may also be interactions between very different neural systems, as for example it is known than sight can affect hearing (Sams et al., 1991), suggesting that not even visual and auditory "subsystems" are independent.

The idea where the signal first enters the retina, and then travels forward via the waypoints of thalamus, V1, and further, is sometimes called the *classic visual hierarchy* (for details on the taxonomy see e.g. Felleman and Essen (1991); Essen (2004)). Although it can be argued that this

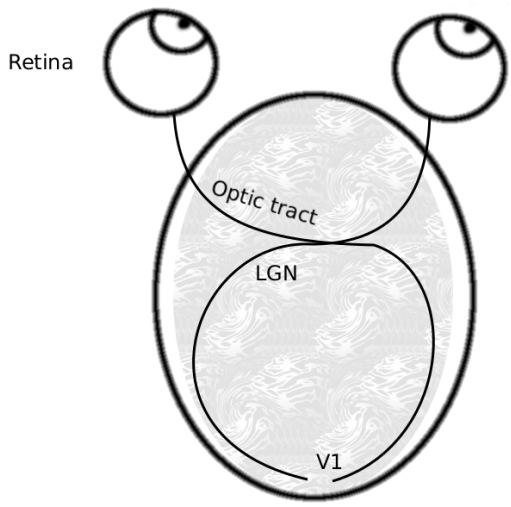

Figure 2.7: The primary pathway of visual information from retina to V1 goes through the LGN in thalamus, as illustrated by a computer science student. The two fiber bundles from the eyes cross at the optic chiasm. Not to scale.

hierarchy might not be a hierarchy functionally, it can still be said that the further we go from the retina in this scheme, the less detailed is our understanding regarding the precise nature of the computations that are performed. We will now cursorily overview the early elements in the classic visual processing view and describe what is known of their computational purposes, as well as what can be reasoned about the visual systems from the properties of these elements.

Retina. The first and perhaps the most researched mechanism in visual processing is the retina (for reviews, see e.g. Hood (1998); Meister and Berry (1999); Masland (2001)). The mammalian retina contains approximately 55 different types of neurons, though not all of them are necessarily required for perception. For example, the retinal melanopsin positive (spindly) ganglion cells are considered to be related to the maintenance of circadian rhythms. For perception, arguably the most important cells are the rods and cones, utilized for night and day vision, respectively. These cells are responsible for measuring the amounts and properties of the incoming photons.

The retinal characteristics can be used to illustrate that the perceived world is an inferred construction and not an honest replication of the external reality. For example, the cones are densely packed in the fovea,

explaining the higher resolution in the center of the visual field. The resolution near the edges of the visual field is much worse, though we are often not consciously aware of this. A similar phenomenon happens with the well-known retinal blind spot and with retinal lesions and scotomas: the missing contents are apparently predicted by the visual system in a process called *filling-in* (Ramachandran & Gregory, 1991). Also, the fact that the rods and cones are actually shadowed by blood vessels (e.g. Adams and Horton (2002)) does not reach conscious perception. Further, although we have two types of cells to sample the photons, this does not result us in perceiving two different modes of vision, nor are there separate rod or cone pathways leaving the retina. These examples from retinal physiology combined with psychophysical measurements suffice to illustrate that the perception is a construction whose mechanisms may not become apparent by simple introspection.

But what computational purpose does the retina serve? One accepted function for the retina is *sampling*, the estimation of amounts, wavelengths and positions of photons that reach the eye. Retina appears to be a very sophisticated device for this purpose, as it is both matched and adaptive to the statistics of the environment (Tadmor & Tolhurst, 2000; Mante et al., 2005), possibly attempting to transmit the visual data efficiently using the limited signalling capacities (Laughlin, 1981). Also, the retina mainly does not transmit light levels per se, but centre surround differences by the operation of retinal ganglion cells (see Figure 2.6A,B for abstractions of receptive fields of two such cells). In the case of the ON-centre cell, the neuron fires strongly if a white light hits the center, as long as the white light does not extend to the surround. Such cells are often modelled by a difference of two Gaussian receptive fields, where a difference of responses to two Gaussians is computed as the cell response, see (Meister & Berry, 1999). In image processing terms, the ganglion cell performs centre-surround, or bandpass filtering (Sonka et al., 2007). It has been suggested that one function of such filtering in the retina is to whiten the signal (Atick and Redlich (1992), also D. J. Graham et al. (2006)), meaning that all spatial frequencies will have approximately the same power in the output. This addresses a problem with the power spectrum of natural scenes, which are dominated by low frequencies, their power decreasing approximately following a power law (for a review, see e.g. Billock (2000), and Section 3.2.1 of this thesis). Whitening also has the consequence of making the covariance structure of the data an identity matrix, i.e. the responses of the centre-surround neurons may become approximately decorrelated over the data in general. We will briefly return to models of whitening in Section 4.3.

After the retinal processing, according to some new results and classification schemes, as many as eight pathways may leave from the retina to the LGN (Casagrande & Xu, 2004). In the classic taxonomy, the best known of such pathways are the parvocellular pathway, which codes for static stimuli (form and colour), and the magnocellular pathway, which is concerned with temporal aspects, i.e. what moves in the environment. These classic pathways are reviewed e.g. in DeYoe and Essen (1988); Livingstone and Hubel (1988), and they end up in different layers in the thalamus.

Thalamus. After the retina, the next distinctive area to receive the visual signal is the lateral geniculate nucleus (LGN) in the thalamus (Sherman & Guillery, 2002). The role of this processing stage is not well understood, possibly due to most of its inputs coming from cortical sources, not from the retina. The cortical inputs to the thalamus are often thought to be related to attentional modulation, that is, the responses of the LGN ganglion cells are affected by later-stage attention. If this attentional component is removed, the LGN ganglion cells appear to behave similarly to their retinal ganglion cell counterparts, i.e. their receptive fields have similar center surround organization. Perhaps due to this similarity, computational models of visual operation that do not include attentional effects do not model effects of LGN, as if LGN did not exist or was a simple relay station[3]. As with the V1 area later, LGN is layered, and different layers e.g. read afferents from different retinas (Sherman & Guillery, 2002). The LGN cells are known to fire in burst mode while the animal is watching natural scenes (Wang et al., 2007) and they have been suggested to code signals with more emphasis on temporal patterns than later V1 neurons do (Kumbhani et al., 2007). These findings take us further from being able to take LGN as a simple relay station, yet the functional significance of these new results is not yet well understood.

V1. The primary visual cortex (area V1) is the first cortical area to receive visual input, and it has been extensively studied since the initial work of Hubel and Wiesel (1959), followed e.g. by Movshon et al. (1978b, 1978a); Ringach (2002), and others. For a brief review of the classic results, see Carandini (2006), and for a critical outlook, see Olshausen and Field (2005).

In V1, some of the receptive fields for the first time take clearly orien-

[3]For example, models such as in Olshausen and Field (1997); A. J. Bell and Sejnowski (1997); Hateren and Schaaf (1998); Hyvärinen and Hoyer (2000); Hyvärinen, Hoyer, and Inki (2001) do not have an LGN component. Perhaps due to this, these models are often called receptive field models, not models of the primary visual pathway.

tation selective shapes, somewhat resembling Gabor filters (e.g. Daugman (1985), see Figure 2.6C,D for illustration). The cells having such receptive fields are typically called *simple cells* while being operationally likened to linear filters that are localized, oriented, and bandpass, and whose response is later rectified (Heeger, 1992a), and possibly influenced by later lateral feedback (e.g. Heeger (1992b); Schwartz and Simoncelli (2001)). To put it another way, discounting the feedback effects, the classic view of a simple cell computation can be implemented using the perceptron model, as defined in eq. (2.1) and shown in Figure 2.4B. A simple cell model is obtained simply by setting the weights of the perceptron to encode the neuron's receptive field, giving the perceptron an appropriate firing threshold, and choosing the used nonlinearity to be halfwave rectification.

The hypotheses for the purposes of the Gabor-like filters found in V1 range from feature or edge detection theories (Hubel & Wiesel, 1959) to local Fourier analysis of the scene content (De Valois & De Valois, 1980). On the other hand, it has also been shown that receptive fields of V1 simple cells are particularly effective for simultaneous measurements of frequency, orientation and spatial position (Daugman, 1985). The simple cell receptive fields also allow for sparse coding of the natural visual environment (Field, 1994; Olshausen & Field, 1997). Another class of cells found in V1, the *complex cells* are often modelled as combinations of rectified simple cell responses, and traditionally taken to compute phase-independent responses to edges and bars (Pollen & Ronner, 1983).

However, as so many other things in visual neuroscience, the concepts of simple and complex cells are under some debate. For example, some authors suggest that this dichotomy to two cell types may be only weakly supported (Chance et al., 1999; Mechler & Ringach, 2002), and yet other theories claim that instead of being a simple pass-through stage for initial visual data analysis, V1 could be a general high-resolution buffer or a scratchboard (T. S. Lee & Mumford, 2003). If time is taken into account, recent results have shown that cells in V1 actively participate in different computations at different times after stimulus onset (Roelfsema et al., 2007). These findings seem to call for readjustment of the classic feedforward view, where V1 cells first perform some simple functionality, and then pass the data on for some more complex processing in subsequent stages. On the contrary, V1 seems to be re-utilized in this further processing as well.

Later areas. The areas following V1 are typically considered to handle motion processing, object identification, colour processing, and so on (Palmer, 1999), with the property that the further we get from the retina,

the neuronal receptive fields get larger and larger spatially, whereas their tuning becomes tighter. For example, in area IT (inferotemporal cortex) that is commonly considered to be involved in object recognition, a receptive field might be more invariant to position, size and orientation of the shown stimulus, but be tightly tuned to more abstract properties such as shape (Tanaka, 1996) Although relatively little is known of the actual computations carried out by these later areas, there already exists a large body of models for high-level tasks such as object recognition. These high-level models are typically designed by engineers to solve the actual problem of object recognition, and sometimes they disregard all biological constraints and plausibility. Yet even such attempts can give valuable insight into vision in general. We will discuss such models in Section 3.2.3.

2.2.3 Formation and plasticity of visual function

So far we have described the challenges of seeing and the early visual mechanisms that partake in addressing these challenges for biological vision. We will now briefly consider the reasons why the visual mechanisms are as they are. Looking from afar, the visual mechanisms can be taken to originate from the interaction between evolution and the properties of the environment. Evolution, on a timescale spanning generations of animals, exerts pressure not only towards optimizing the function of animals as entities, but also towards optimizing the functional quality of their parts (Alexander, 1996). This kind of optimization results in genetically determined differences between species, such as whether the animal will grow compound eyes typical to insects or the more familiar retinal eyes we know from vertebrates. In the scope of this thesis, we take the genetic determination that forces a certain solution mechanism to correspond to a technical problem of model class selection. To give an example, in model selection we might have to decide whether we pick our model from the function class of all the possible perceptrons, or some other class of functions. We will discuss the model selection issue further in Section 4.2.2; here we simply note that it is interesting that the evolutionary convergence towards a certain kind of solution may lead to a dead end (just as a model class choice may do). For example, the compound eye is a solution to sampling has a poor resolution and a design that seems difficult to improve further by evolution-like mechanisms (Nilsson, 1989).

As well as the genetic code getting optimized by evolution, biological systems have evolved to change their own function during the lifetime of the system. These changes that may span timescales ranging from seconds to years have also to do with how the biological mechanisms are; consider, for

example, that the optimization of function – such as learning to recognize a new object – necessarily alters some structures in the visual system.

Here we divide the changes that occur in visual processing during the lifetime into two separate categories that we call *adaptation* and *learning*. By adaptation, we mean such rapid, non-permanent changes that are caused by alterations in the environmental conditions. These rapid adaptation mechanisms could include for example the retinal light level adaptation (Cleland & Freeman, 1988; Hood, 1998; Mante et al., 2005) that adjusts the retinal processing for the current lighting conditions. A classic model of how such adaptation could work is the Naka-Rushton equation for light adaptation (Naka & Rushton, 1966). Given a vectorized grey-scale image $\mathbf{x}$ of length n, the pointwise equation gives

$$\hat{x_i} = \frac{x_i^c}{x_i^c + (1/n \sum_{j=1}^{n} x_j)^c}, \tag{2.2}$$

where c is a fixed parameter. Thus, the adaptation in the model depends on the average lightness level of the scene.

To the same category with light adaptation we include the various gain control phenomena appearing at various stages of visual processing, for example in V1 (Heeger, 1992b; Schwartz & Simoncelli, 2001; Finn et al., 2007; Duong & Freeman, 2007). These kind of adaptation mechanisms appear related to such useful activities as maximization of neuronal information transfer (Laughlin, 1981) as well as minimization of dependencies in the neural codes (Schwartz & Simoncelli, 2001). In other words, the adaptation mechanisms perform real-time optimization of visual processing. However, in the current work our focus is on learning, and the reader interested in adaptation is referred to Kohn (2007).

By *learning* we denote the processes that cause slower and more lasting effects to visual processing. Here we consider such phenomena as the reorganization of cortical maps after lesions (Kaas et al., 1990) and the remembering of previously encountered objects as results of learning, although in actuality these results may be due to different physiological origins.

The type of learning we are considering has been traditionally taken to originate from changes to synaptic strengths between neurons, and to the neuronal firing thresholds (Dayan & Abbott, 2001). Hebb (1964) proposed that these synaptic weights were changed by co-activity, i.e. the neurons that fired at the same time were to enhance the synaptic strength between them. Later, this learning paradigm became known as Hebbian learning, and it is an example of an *algorithm* in the sense that the description specifies how learning is done, but not its objective. However, it can be shown that Hebbian learning corresponds to the objective of maximizing neuronal

response variances (e.g. Dayan and Abbott (2001)), just as Principal Components Analysis (PCA) does (see Hyvärinen, Karhunen, and Oja (2001), also briefly described in Section 4.3).

In model systems such as the perceptrons of eq. (2.1), the learning simulating synaptic strength change would modify the weights $\mathbf{w}$ and the firing threshold b of the models. But does this modification have any connection to biological learning as it is currently known? At the time of writing, the intuitive Hebbian learning and simple synaptic strength modification has been supplanted by a host of low-level learning phenomena, including short- and long-term potentiation (STP/LTP), and the corresponding depressions (STD/LTD), possibly mediated through mechanisms such as spike-time dependent plasticity (STDP, i.e. Dan and Poo (2006)). But in computational studies, learning works solely through modifying values of scalars and vectors with arithmetic operations. In order to incorporate simulations of biochemical mechanisms of potentiation to change scalars and vectors, a convincing account should be presented of biochemical learning being somehow superior in attaining some particular functionality. Meanwhile, models and their estimation are plausibly two different issues, and models that exhibit some functional properties have those properties regardless of how their parameters were obtained. In this thesis we will try to keep the models as simple as possible, and we will further discuss the benefits of this choice in Section 4.2.

We conclude this section by summarizing that the biological visual processing and its impressive capabilities seem to be due to optimization done at different, possibly overlapping timescales. These timescales include the grand evolutionary scale, but also the scales of days and even of seconds. None of these adjustments to form and function of visual processing happen in a void, but in a continuous interaction with the rich, natural environment. In the next chapter, we will describe how these realizations about the formation of natural visual systems can be utilized in computational modelling of visual processing.

Chapter 3

Ecology-driven modelling of vision

To attain models of visual processing, we make the following conceptual segregations in this thesis: we assume some level of analysis (a mechanism), its environment (data) and a purpose for the processing (learning objectives and constraints). In addition, we select some optimization (learning) algorithm to adapt the mechanism. Now, given the data, the optimization algorithm is used to modify the mechanism parameters to try to meet the learning objective and the constraints as well as possible. As an example, consider making the following choices. First, assume a mechanism (model class) of a function $s = g(\mathbf{w}^T\mathbf{x} + b)$ of eq. (2.1), and further choices of $g(y) = y$ and $b = 0$ (making the mechanism a linear filter model). Next, select a set of vectorized natural image patches to represent the environment (multiple $\mathbf{x}$), and a goal that the response s should have as high variance on the data as possible. The preferred model is then a single linear projection that explains as much of the data variation as possible. Finally, to keep the model $\mathbf{w}$ bounded, enforce an additional constraint $||\mathbf{w}||_2 = 1$. The task of the learning algorithm is then to find suitable parameters for $\mathbf{w}$. For natural image data, the optimal $\mathbf{w}$ under these conditions would be a non-oriented low-pass filter, as the linear computation estimating the mean has the highest variance out of all the linear mappings on natural images (see e.g. the first PCA component in Hancock et al. (1992), or in A. J. Bell and Sejnowski (1997)). The attained processing with $g(\mathbf{w}^T\mathbf{x}+b)$ that computes the mean of $\mathbf{x}$ is thus an *emergent* property arising from the interplay of the function class, the constraints, the given environmental data and the learning algorithm used.

The above optimization setting is a concise example of how learning visual models can be carried out in the absence of physiological measurement data: we do not have such data, nor are we attempting to fit neuronal models to it. Instead, we work in the paradigm of information processing

and assume that the physiological neurons and networks are processors of sensory information. Further, we assume that their task is to process and modify information in ways that are optimized for either subsequent processing or to fulfil some more immediate goal (Marr, 1982; Glimcher, 2003). This setting as a whole is what we call the *ecology-driven approach* to visual modelling.

A benefit of the ecology-driven approach is that it is possible to evaluate the learned models by measuring how well they work in the light of the specified objectives. And despite the fact that no physiological data was used, we can compare the learned models to machinery found in physiology: suppose first that a learned mechanism is particularly good at some task and at odds with physiology. This is an interesting result helping us to understand vision better, as it is by no means the case that some decent solution to visual processing was unique, or that the optimization done by evolution should have found the best solution to some particular problem. On the other hand, if the learning results in a model that has resemblance to physiology in some ways, this gives an interesting hypothesis that the physiological mechanism is also doing well with relation to the objective we had specified for the model. This kind of emergence allows predicting a functional explanation for the corresponding biological processing.

3.1 Historical background

The setting that we call the ecology-driven approach is a mixture of several very common ideas that can be traced back at least to Darwin, c. 1809 - c. 1882. The main underlying principle is to consider the biological processing mechanisms as functional entities that have been optimized by evolution. Exaggerating and simplifying, the mechanisms have survived natural selection under the "law of the jungle". Hence, Darwin can have been said to have brought optimization to the ecology-driven approach.

But what is this Darwinian jungle like? As we already described in Section 2.1, the signals the retina encounters are not made of such simple percepts as introspection may lead us to believe. This was recognized relatively early by e.g. James (1899), who emphasized the complexity of the visual world. But albeit complex, clearly the natural conditions are not arbitrary. From these grounds, it is not a long leap to suppose that the visual machineries are particularly good in addressing just the kind of complexity that exists in the nature. Subsequently, the school of Gestalt psychology encouraged to study the environment in order to understand vision (as in e.g. Gibson (1979)). These historical developments deliver

the idea that natural environments (data) are the relevant and challenging working conditions for the visual mechanisms.

Getting more specific from the vague goal of evolutionary fitness, it was Marr who popularized the idea that individual visual mechanisms may have more specific functional purposes (Marr, 1982). In his work, Marr admonished the then-typical straightforward approaches to visual modelling. The problem was that with e.g. neurophysiological data, it is perfectly possible to make a neuron model – of, say, a V1 simple cell – without any account of the purpose the neuron serves: simply create a model that responds similarly to a real neuron to the presented stimuli. Forgetting for a moment the stochastic and interactive nature of neurons, at the worst such a model could be a large table of input-output mappings. Contrary to this kind of modelling that explains nothing, Marr argued that the purposes of the processing should be understood (Marr, 1982). Thus, by promoting purpose, Marr resounded Mach's economical principle (Mach, 1882) and paved way for e.g. the study of neuroeconomics (Glimcher, 2003) and the realization that the whole idea of neurons encoding and representing information is only meaningful if there is a decoder to interpret and make use of the relayed information (deCharms & Zador, 2000; Barlow, 2001).

We have now collected the main ideas of the ecology-driven approach: optimization, ecologically valid environmental data, and learning objectives (or purpose) for the model. Only the actors are missing from this scene we have portrayed. These actors, the models, and the methods to fit their parameters and evaluate them, are due to the recent advances in probabilistic modelling, computational learning theory and machine learning (Duda et al., 2000; Hastie et al., 2001; Bishop, 2006). These fields examine the conditions required to make learning from data feasible. They are also the source of the learning methodologies used in this thesis.

The reason why we do not call the ecology-driven approach "machine learning" is mainly connotational. Although machine learning methodology is utilized in this thesis, our emphasis differs from that of typical machine learning, as we are not particularly interested in the technical properties of learning algorithms nor in deriving new ones, unless the existing ones are deemed insufficient in some crucial way. Instead, in our setting the learning algorithm is more or less a tool that is used to estimate mechanisms for information processing, and it is these estimated mechanisms and their properties that are of interest. Specifically, we are interested in the nature of the emergent computations, should they help us either in solving some particular visual processing task or in understanding visual processing to some larger degree. This is contrary to typical attitudes in machine learn-

ing, where the properties of the data are often not carefully analysed and where the learned models may be considered black boxes. For example, in the context of machine learning and probabilistic models, Bishop claims that "model parameters such as $\mathbf{w}$ are of little direct interest" (Bishop (2006), p. 364). Clearly such stances do not reflect the ones taken in this thesis[1].

3.2 Statistics and function

The ecology-driven approach implicitly assumes that the environment is experienced as *samples* and that the world is uncertain from the viewpoint of the visual system, as we do not know for sure what kind of sample is attained ten minutes from now, and nor can we exactly know the current state of the world based on the current sensory sample alone. A prime example of the latter is the inverse problem mentioned in Section 2.1: due to occlusions, different states of the world can account for the seen image. The best that can be done in this situation is to behave statistically well, i.e. make the choices and inferences that seem appropriate based on our previous experiences, possibly coupling along previously received environmental feedback. Hence, behaviour is partly based on sensory input statistics, i.e. the behaviour is *probabilistic*[2], and thus appropriate behaviour requires measurements of the input statistics (or probabilities) to be able to make predictions about the environment, see e.g. Barlow (2001). These measurements can then be used to carry out such subsequent probabilistic inferences as discussed e.g. in Helmholtz (1867); T. S. Lee and Mumford (2003); Glimcher (2003); Körding and Wolpert (2004).

3.2.1 Natural image statistics

It is relatively easy to estimate how frequently a coin lands as heads, or the correlation between two different coins being tossed (usually zero), and then base inferences and decisions on empirical frequency estimates. But what are the frequencies of different constituents in the natural visual environment? Further, what are these constituents, and how do they relate to each other? For example, can the presence of one constituent be predicted from some others? How do we detect the presence of the constituents from

[1] Note that even in the wildest Bayesian procedures of integrating out parameters, some concrete entities must remain in order to carry out the model function, and it is these entities we are interested in, here encoded in part by the parameters such as $\mathbf{w}$.

[2] The word 'probabilistic' as in 'a probabilistic model' here means that the behaviour is grounded on statistical experience, and not that the behaviour is in some way random.

the retinal inputs in the first place, in order to base any decisions on their presence or absence?

Generally, we can study what kind of elements and statistical regularities exist in the natural environment, and then examine if and how the natural visual systems extract these elements or reflect their statistical regularities. In this thesis, the way to study the environmental statistics is through the statistical characteristics of natural, photographic images. This is equivalent to using photographic images as an approximation of the signal that is received by the retina and then processed further. A central property of natural, photographic images is that they force the modeller to face the complexity that is present in the natural environment. This contrasts to the artificial stimuli such as sine gratings used in research areas such as psychophysics, where artificial data remains a valid research tool (Rust & Movshon, 2005). However, artificial data suffers from ecological invalidity, as it is quite clear that the visual systems did not evolve to survive in a world consisting of sine gratings, but rather in one that is, as William James famously put it, a "great blooming, buzzing confusion" (James, 1899). Working in a significantly more tidy artificial world can help in understanding vision, but it can also lull the modeller into a false sense of tidiness and security, as well as lure to presumptions that may not be ultimately tenable (for example presuming that segmentation can be completed before recognition, see Section 3.3).

There is now ample evidence that the statistical structure of natural images and the processing mechanisms of the early visual system are interrelated in various ways (e.g. Laughlin (1981); Atick and Redlich (1992); Ruderman (1994); A. J. Bell and Sejnowski (1997); Olshausen and Field (1997); Hateren and Schaaf (1998); Sigman et al. (2001); Elder and Goldberg (2002); Mante et al. (2005); Kingdom et al. (2007), for a review see e.g. Simoncelli and Olshausen (2001)). Also research areas that have traditionally used artificial stimuli are now considering experiments based on more natural stimuli (Felsen & Dan, 2005), but see also Rust and Movshon (2005); Pinto et al. (2008). In addition, a lot of applied work in computer vision, content-based image retrieval and low-level image modelling can be taken to analyse or utilize the statistical regularities present in the data.

The tight connection between the environmental characteristics and the appropriate processing is understandable, as it can be argued that incorporating some kind of exploitation of statistical properties in the data can be useful for a wide variety of tasks ranging from low- to high-level ones. For example, the statistical regularities can be utilized in very general frameworks, such as efficient coding and data compression (e.g. Barlow

(1969); Cover and Thomas (2006)), probabilistic modelling and inference (Helmholtz, 1867; Bishop, 2006), economical behaviour (Glimcher, 2003), and so on.

Thus, the study of the statistical characteristics of visual input can be motivated from several fundamental premises. As the data dimensionality of natural images is high, this line of analysis became possible only after the computing infrastructures reached acceptable computational power-per-cost ratios. Subsequently, a substantial amount of research has been carried out to study the statistical properties of natural images (e.g. Ruderman (1994, 1997); Thomson (1999); Zetzsche and Krieger (1999); Brady and Field (2000); Simoncelli and Olshausen (2001); A. B. Lee et al. (2001); Torralba and Oliva (2003); Johnson and Baker (2004); Simoncelli (2005)). At the simplest, these studies measure well-known statistical characteristics of large sets of natural images (Ruderman, 1997; Thomson, 1999; Zetzsche & Krieger, 1999). One such characteristic could be the average Fourier power spectrum, corresponding to the autocorrelation function of the data and its second order dependency structure (i.e. covariances). This analysis reveals that natural images commonly follow a power law structure $1/f^{\alpha}$, where f is the spatial frequency, and α a coefficient typically close to 2 (for a review, see Billock (2000)). The lowest frequencies then have the highest power, corresponding to strongest pixel correlations being between nearby pixels. This characterizes one type of redundancy in the images. But as roughly similar power spectra and other simple statistics can be attained from simple artificial data as well (Ruderman, 1997; A. B. Lee et al., 2001), this illustrates that such simple statistics do not yet capture the rich statistical structure of natural images (as demonstrated in Figure 3.1).

More involved research on natural image statistics considers statistical regularities between responses of filters (Johnson & Baker, 2004; Simoncelli, 2005). Here, linear filters such as those resembling V1 simple cell receptive fields are applied on the images (see Figure 2.6C,D), and the statistical dependencies in the filter responses are analysed. With appropriate gain-control applied to such filter responses, it has been shown that pairwise dependencies are minimized, giving an interesting hypothesis regarding the function of gain-control in V1 (Schwartz & Simoncelli, 2001).

In this thesis we often consider a setting where instead of estimating predefined statistical descriptors of the data by some fixed computation, we attempt to learn the computation that fulfils some predefined objectives, and this learning is based on the natural image data and constraints. The models learned in this manner then incorporate the statistical aspects of the data in various ways, ranging e.g. from capturing higher-order statistical

regularities of the data to performing object recognition on natural images. In the following, we will briefly overview some statistical models of this kind. They have been classified either as models of visual input or as models of visual function, and we will start by describing the visual input models.

3.2.2 Statistical models of visual input

A *model of visual input* is one that given an image $\mathbf{x}$, returns the density of the estimated distribution at that point, i.e. $p_{\mathbf{x}}(\mathbf{x})$. As the only training data required for estimating densities $p_{\mathbf{x}}(\cdot)$ is a set of vectors $\mathbf{x}$, this learning method is called *unsupervised* (Becker & Zemel, 2003) as there is no feedback either from the environment or from a teacher.

One possible learning constraint in unsupervised learning comes from the assumption regarding the functional family of the density of $\mathbf{x}$. For example, this family could be assumed to be Gaussian. In that case, if we wished to estimate the parameters of the model so that our observed data were maximally likely under the model, the model learning would amount to estimating the multivariate mean and covariance from a set of training images (Bishop, 2006). In the case of the Gaussian distribution, these two sets of parameters suffice to specify the distribution exactly.

However, it would be surprising if the rich variability of the visual world could be captured in the few parameters of the Gaussian density. This would entail that more complex dependencies than covariances would not exist in the data, and that the Fourier spectrum would be sufficient to capture the statistical structure of the visual environment: similarly to the Gaussian model, the Fourier spectrum also misses regularities beyond the second order. However, the visual reality is not as unstructured. This is clear from Figure 3.1 where we show a sampled image from a Gaussian model whose parameters capture the spectral decay common to natural images (Billock (2000); for the sampling recipe, see Simoncelli (2005)). Clearly the image resembles natural scenery very little. This lack in the Gaussian model can be somewhat addressed by assuming more structured distributions, and depending on the distribution in question and the success in learning its parameters, the resulting statistical models can be applied to purposes such as de-noising and image compression (e.g. Portilla et al. (2003); Simoncelli (2005)).

Although the mentioned Gaussian model can be used to generate new images by sampling (and in theory this holds for any probability density), the model is holistic in the sense that it does not explicitly describe how the images are formed. Another subclass of input models, often called

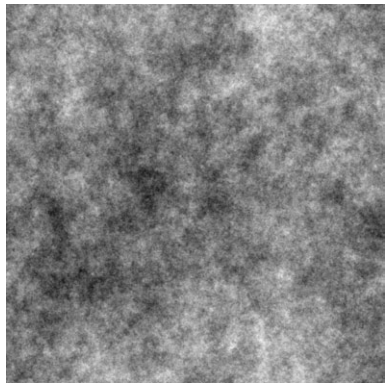

Figure 3.1: An image drawn from a Gaussian model with an approximately $1/f^{\alpha}$ decay in the Fourier power spectrum, using parameter $\alpha = 2.1$.

generative models, are ones that make the assumptions regarding the image formation process explicit. Perhaps the best known generative model is the linear superposition model (e.g. A. J. Bell and Sejnowski (1997); Hateren and Schaaf (1998)),

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{n} s_i \mathbf{a}_i, \tag{3.1}$$

where the s_i are stochastic scalar coefficients (*hidden variables*) of the model that are used to weight n fixed image templates $\mathbf{a}_i$ in a linear sum. The benefit of eq. (3.1) is that it makes the assumed process of image generation explicit: the model *represents* the image $\mathbf{x}$ using the templates in $\mathbf{A}$, as visualized in Figure 3.2. A further benefit is that considering an invertible $\mathbf{A}$ in the linear model of eq. (3.1), finding the coefficients in $\mathbf{s}$ is particularly simple for a given $\mathbf{x}$. The coefficients can be attained with

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x}, \tag{3.2}$$

i.e. by linear filtering. The coefficients $\mathbf{s}$ are then the responses to simple linear filtrations, and explains why we denoted them similarly to the responses of neural models in the previous chapter.

It should be noted that the decomposition of eq. (3.1) is by no means unique for some given dataset. On the contrary, without further constraints, any square and invertible $\mathbf{A}$ is sufficient, and the particular choice of $\mathbf{A}$ will affect the statistical characteristics of the stochastic variables $\mathbf{s}$. Also, instead of being square, $\mathbf{A}$ can be under- or overcomplete, and in some cases eq. (3.1) may hold only approximately after optimization of $\mathbf{A}$

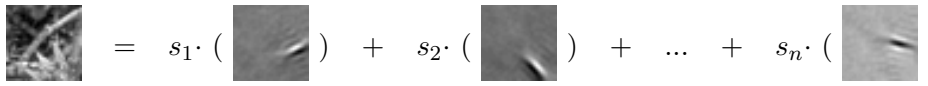

Figure 3.2: In a linear superposition model, images or image patches are represented as linear sums of image templates. Each template ($\mathbf{a}_i$ in the text) is weighted by some image-specific scalar coefficient s_i to generate the image. The coefficients in $\mathbf{s}$ provide a representation for the image in a system defined by the templates.

and $\mathbf{s}$ for a given set of vectors $\mathbf{x}$. Also, if $\mathbf{A}$ was invertible, we could equivalently optimize $\mathbf{W}$, trying to attain some particular statistical properties for $\mathbf{s}$. Then the approach would in essence amount to projection pursuit (Friedman, 1987). The invertibility of $\mathbf{A}$ also allows us to relate the density of $\mathbf{s}$ to the density of $\mathbf{x}$ as

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}}(As) = |\det \mathbf{A}| p_{\mathbf{s}}(\mathbf{s}), \tag{3.3}$$

from the well-known result about the density of invertible transforms (see e.g. Hyvärinen, Karhunen, and Oja (2001), p.35). In effect, if $\mathbf{s}$ has a more tractable distribution than $\mathbf{x}$, then knowing $\mathbf{A}$ allows us to get a more simple model for $p_{\mathbf{x}}(\cdot)$ by eq. (3.3). In such a case, $\mathbf{A}$ could be thought of as a mixing matrix that needs to be found in order to recover the more simple hidden variables $\mathbf{s}$.

In practice, the estimation of $\mathbf{A}$ can be further constrained by specifying some statistical objectives for the hidden variables $\mathbf{s}$. The objectives studied in the publications of this thesis will be reviewed in Section 4.1, but for now we appetize the reader with some examples from the previous research: in particular we mention that we can optimize the model for the sparseness of the hidden variables (Olshausen & Field, 1996), for the independence of the coefficients (A. J. Bell & Sejnowski, 1997; Hateren & Schaaf, 1998), or for the robustness of the representation (Doi et al., 2007). Of these, sparseness can have utility in conserving metabolic energy in biological systems (Levy & Baxter, 1996), but also in attaining better signal-to-noise ratios and in assisting in feature detection (Field, 1994). In addition, sparseness is intimately connected to independence (Olshausen & Field, 1997), and this latter property is beneficial for making probabilistic inference more tractable (as will be explained in Section 4.1.1). The robustness objective, on the other hand, tries to attain visual processing that is not fragile against internal or external noise. Yet further unsupervised objectives can be envisioned if the inputs $\mathbf{x}$ are spatiotemporal data (such as video segments). In that case, objectives such as temporal coherence

(Becker, 1992; Hurri & Hyvärinen, 2003) and slowness (Berkes & Wiskott, 2005) become applicable. In such approaches, the idea is to learn representational mechanisms whose responses change coherently or slowly over time.

An interesting result is that in the context of the linear superposition models and their simple extensions, practically all the objectives mentioned above lead to representing images in terms of features or filters that are localized, oriented, and bandpass. The learned filters resemble both V1 simple cells and the Gabor functions that were used to model receptive fields in Figure 2.6C,D on page 19. How the learned receptive fields actually look like can be seen from most of the papers mentioned above, or from Publication 5 of this thesis. These various examples of "emergence" of Gabor-like filtering suggests that the Gabor-like image decomposition may be beneficial for several slightly different low-level purposes.

One problem with the linear superposition models is that if $\mathbf{A}$ is invertible, then given $\mathbf{s}$ and $\mathbf{A}$, the signal $\mathbf{x}$ can be reconstructed perfectly. This shows that filtering with $\mathbf{W}$ is not discarding any information, and that the representations are "autistic" in the sense that the models will use the coefficients $\mathbf{s}$ to represent aspects of natural scenes that are not required for any of the usually imaginable high level tasks, i.e. the model mechanisms are happy to faithfully create authentic reconstructions of meaningless clutter or stochastic stimuli such as textures and noise. Although we can see this kind of complexity (i.e. it is relayed to the conscious level), it is questionable if such modelling approaches can be extended for higher-level nonlinear modelling, as it arguably becomes more important to concentrate on the relevant aspects of the stimuli and abstract away behaviourally irrelevant detail. To put it another way, for high-level vision it may become more important to consider what the data representations or encodings are used for (deCharms & Zador, 2000; Barlow, 2001; Eliasmith, 2007). For example, it seems unlikely that a brain could contain a different neuron for every imaginable combinatorial configuration of some elementary parts such as edges. Instead, the stored representations (or memories) are more likely to be made only of ecologically important visual configurations, and doing this requires ability to ignore irrelevant information and combinations.

Another issue with the linear superposition models stems from the fact that the generative process they specify is not compatible with the natural image formation process of the physical reality. Instead of summing transparent templates, the real formation process of natural images is more closely analogous to one where opaque surfaces occlude each other from view (Ruderman, 1997; A. B. Lee et al., 2001). Unfortunately, although

image generation from such occlusion models is easy, it seems an open research problem how to invert the computation, i.e. to answer the question how a given image should be represented using a set of templates and some occlusion operator. Also, using such models to give probabilities for images seems difficult. This is likely the main reason for the prevalence of the linear superposition models that are differentiable and hence mathematically convenient.

3.2.3 Statistical models of visual function

From a high-level viewpoint, the objectives of the previous section were *indirect* as the resulting models were not optimized to solve any particular high-level task for some given single $\mathbf{x}$, and it has been proposed that such simple objectives may be insufficient to explain visual mechanisms (Baddeley, 1996). However, the ecology-driven approach is by no means limited to modelling visual data using low-level, unsupervised objectives. Instead, we can directly model some visual function. A *model of visual function* is one that given an image, performs some task on the image, such as *edge detection* or *object recognition*. Both of these practically amount to modelling decisions given an image or some region of it. For modelling decision making, explicit density modelling of the data distribution may not be needed.

The direct modelling of visual function can be said to have been pioneered by works such as Roberts (1965) on edge detection, although the early computer vision models were not in general based on data-driven statistical estimation. Later, the visual processing theories and functional emphasis of Marr (1982) inspired a host of research that progressed by dividing the grand problem of perception into functional subproblems and then solving these subproblems separately. Common problems included e.g. the already mentioned object recognition, but also scene segmentation (Jain & Farrokhnia, 1991; Shi & Malik, 2000; Sharon et al., 2006), colour constancy (Land & McCann, 1971), representations by more specific and invariant features (Lowe, 2004), and so on. Typical to these approaches is the underlying hope that eventually these independent solutions could be designed to work reliably and fast, and then combined into a fully operational general visual system.

In addition to the early computer vision approaches that were more based on engineering than on learning, the statistically learned models of visual function have lately become more prominent. Statistical approaches are very common especially in object detection or recognition (e.g. Turk and Pentland (1991); Riesenhuber and Poggio (2000); Schneiderman and

Kanade (2002); Ullman et al. (2002); Viola and Jones (2003); Amit et al. (2004); Agarwal et al. (2004); LeCun et al. (2004); Fei-Fei et al. (2006); Serre, Wolf, et al. (2007); Leibe et al. (2008)). Typically these methods are based on teaching with labelled examples, a setting that is often called *supervised learning* (Duda et al., 2000). In supervised learning, instead of having only a set of vectors $\mathbf{x}$ as training examples, a set of pairs $(\mathbf{x}, y)$ are provided, where each example $\mathbf{x}$ is supplied with a truth-value y from a teacher, for example the correct action to take (but y could also be a vector). The discrepancy between the action selected by the model and the action suggested by the environment can easily be converted into a feedback signal penalizing or rewarding the actions taken by the model and subsequently to update the model parameters.

For object recognition, the truth-value y could be identities of the objects in the image $\mathbf{x}$, and the task of the learning would be to tune the model parameters so that the model can predict well the objects present in new images that were not used in the training. As the model parameters in these cases are more or less learned from just the images of the objects in question and the corresponding labels, this approach is prone to the earlier criticism in Section 2.1.3, as it is implausible that the methods would learn e.g. the semantic meaning of chairs from chair images alone. However, the models resulting from supervised learning provide an interesting first approximation to objects recognition, as well as one that is very successful in favourable conditions (Serre, Oliva, & Poggio, 2007).

An interesting by-product from modelling of object recognition is the commentary the successful models allow on earlier theories of visual data representation. In particular, the object recognition successes suggest that 3D reconstruction of surfaces (Marr, 1982) or decomposition of scenes into 3D primitives such as geons (Biederman, 1987) may not be universally required as preprocessing steps for all high-level visual functions. On the contrary, quite good recognition rates (in non-pathological conditions) can be attained with methods that do not perform any such explicit reconstructions (e.g. LeCun et al. (2004); Serre, Wolf, et al. (2007)). Another point worth mentioning is that these models also demonstrate that neural mechanisms such as the centre-surround in the retina and the Gabor-filter like machinery in the V1 (as described in Section 2.2) may not be strictly necessary for object recognition either, as a variety of computational systems (Turk & Pentland, 1991; Ullman et al., 2002) are able to recognize objects decently without using Gabor-like filters.

The fact that it is possible to carry out visual function in different ways is parallelled in how such functional mechanisms can be implemented. Here

we wish to emphasize the point that explicit modelling of visual input is not required for modelling of visual function. On the contrary, although models of visual function can be based on models of visual input $p_{\mathbf{x}}(\mathbf{x})$ and application of Bayesian methods (Bishop, 2006), this is not strictly required, and recognition may be based on direct functional modelling (as in LeCun et al. (2004); Serre, Oliva, and Poggio (2007)). One reason why direct modelling of function may be preferable is that it may be unnecessarily complicated to solve the problem through the estimation of the density of the data. We will illustrate this through a slightly technical example in Figure 3.3. There, Figure 3.3A shows a density of two multinomial variables x_1 and x_2, where each variable can have discrete values from the range $[1, 1024]$. Suppose that due to some external rule, pairs with $w_1 x_1 + x_2 < b$ are acceptable (objects) and the rest are unacceptable (non-objects). To put it another way, in supervised learning the instances $\mathbf{x}$ have been *decided by someone* to correspond to outcomes y, possibly based on a *feeling*. Then, assuming the decision rule does not change over time, the decision can be modelled with a mapping $f(\mathbf{x}) \mapsto y$. In the case of our example, $f()$ would evaluate the truth value of the inequality that we specified before. Suppose further that *a'priori* it is unknown how $f()$ is computed. This is a reasonable assumption in tasks such as object recognition, where we can easily tell which objects are present in the image, but where we cannot as easily specify how we reach this decision given the pixels. However, supervised learning can attempt to estimate an approximation for the decision function based on set of pairs $(\mathbf{x}, y)$. A conceptually simple way to proceed would be through modelling $p(\mathbf{x}, y)$, and then selecting the $\hat{y}$ that has maximal density for given $\mathbf{x}$. But already with the example case and density $p(\mathbf{x})$, this would require the estimation of frequencies of roughly a million possible pairs to get just $p(\mathbf{x})$, and this would arguably be less complex than estimating the whole $p(\mathbf{x}, y)$. The former is illustrated in Figure 3.3B, where we used a massive amount of 10^7 examples sampled from the distribution to estimate the density $p(\mathbf{x})$. Subsequently it can be understood that the amount of labelled samples $(\mathbf{x}, y)$ needed to get a good joint density estimate could be enormous. A slightly simpler alternative would proceed through the modelling of $p(\mathbf{x}|y)$. In that case, we would need to separately estimate the frequencies of pairs on the different sides of the red dividing line in Figure 3.3A, but not gaining a major reduction in the number of parameters to estimate. However, if we decided to approximate the decision surface with a hyperplane, the parameter complexity would be brought down from the order of a million parameters to just two parameters in the case of our

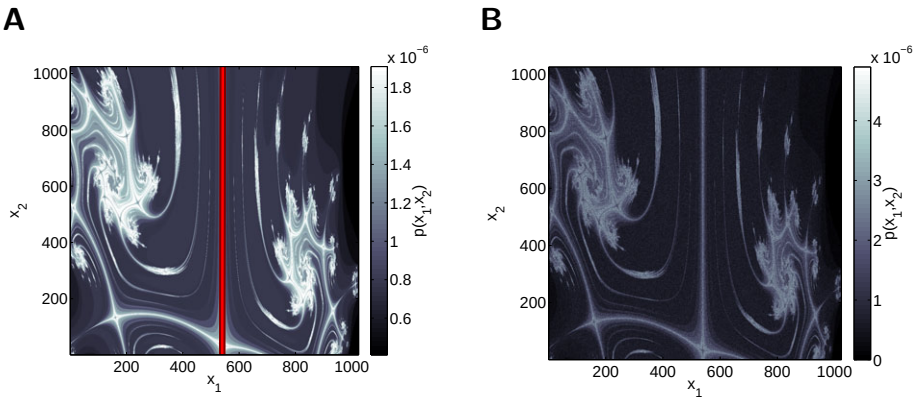

Figure 3.3: **A)** Visualization of an arbitrary fractal density $p(\mathbf{x})$ of two multinomial variables x_1 and x_2. The shading shows the frequency of each of the 1024^2 possible ordered pairs (x_1, x_2). Each such unique pair corresponds to a parameter in the multinomial density. The red dividing line is an arbitrary decision surface of only two parameters. **B)** A maximum likelihood estimate of the density in A, after seeing 10^7 examples sampled from the true distribution. With 10^6 samples, the shape is only barely recognizable (not shown).

example[3]. To summarize, modelling of function can require significantly fewer parameters than density estimation, and thus function may be easier to learn. We will return to these issues in Section 4.2, where we discuss the connections between model complexity and learning in the light of statistical theory.

In this thesis, we study direct models of visual function in object recognition in Publication 2 and Publication 3. Yet, an important question concerns the philosophical underpinnings of such efforts in general: to what extent is it tenable to model different visual functions and operations isolated from each other? We will devote the next section to discussing this question.

[3]Note that Figure 3.3 does have regularity, suggesting that some simple function might exist to compute $p(\mathbf{x})$. However, recovering the unknown function may be challenging from just a few samples of data $\mathbf{x} \sim p_{\mathbf{x}}(\cdot)$.

3.3 Are there independent mechanisms in perception?

The approaches we have discussed either try to formulate a probabilistic model for the input data, or present a data-driven model for some visual function. From the viewpoint of a visual system, both approaches make functional independence assumptions (not exactly the same as statistical independence). By functional independence we mean the hypothesis that there are some functions that can be carried out optimally without any interaction between the functional mechanisms.

In the case of the unsupervised models of input data, there is a fundamental independence assumption between the levels of analysis: the objectives mentioned in Section 3.2.2 are low-level ones, and they do not explicitly tune the representations towards being optimized for any such *behavioural* function that animals can be observed to perform on the macro scale, including foraging, mating, avoiding obstacles and so on. Bridging this gap between objective levels would seem to require a feedback signal from the environment, turning the approach essentially into supervised learning. Yet other, less fundamental functional independence assumptions can be made on the level of the models themselves. To give an example, in eq. (3.2), each s_i is computed independently by a dot product, and these computations in no way influence one another (whereas in natural neural systems, neurons may affect computations of nearby neurons through lateral connections (Dayan & Abbott, 2001)).

Also supervised learning can make suspicious independence assumptions, although this setting is often used to learn behaviour closer to the macro scale. In the context of supervised models, common assumptions include that objects to be recognized are independent of their spatial environments, that images can be segmented before interpreted, or that object parts such as edges can be detected before the object itself. In the two last cases, this entails the existence of early visual processing mechanisms that feed later mechanisms, but that are independent of them; the functional dependency structure is one-directional. But are these assumptions tenable?

For some tasks, assuming functional independence seems well founded. For example, motion and shape could be relatively close to statistical independence in the visual environment, and have so different characteristics that they could require different kinds of processing machineries. This is reflected in the existence of separate processing streams for "what" and "where" in the biological systems (Ungerleider & Mishkin, 1982). Unfor-

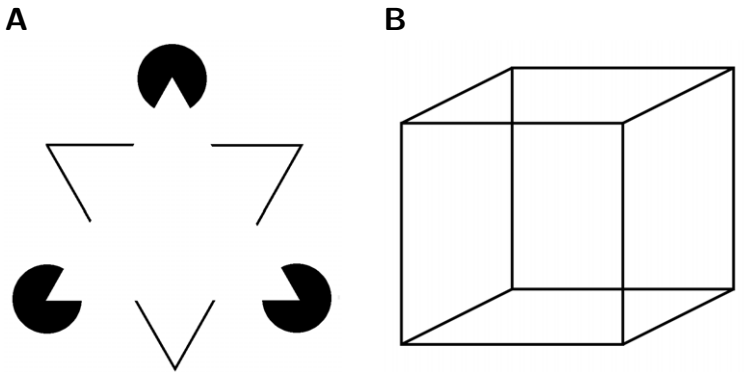

Figure 3.4: **A)** A Kanizsa triangle that causes percepts of illusionary contours. **B)** A Necker cube as an example of bi-stable perception.

tunately, many problems that are attempted to be solved separately in the literature do not seem to be segregated in human vision. One example of such a problem is local edge detection. In a traditional edge detecting setting, a local image area (corresponding to a receptive field) is analysed to decide whether it contains an edge or not (e.g. Roberts (1965)), and this decision is final. For example, a perceptron could attempt to perform such edge detection. But what if the edge is very faint, or does not exist, but according to some global analysis, actually should be there? The result is that an incorrect judgment may be passed on to further processing. Such behaviour is not agreeing with the human visual system that typically can perceive missing edges as *illusionary contours*, as edges that are not there but yet there is a feeling as if they were (for a review see Eagleman (2001)). One such example is the famous Kanizsa triangle, shown in Figure 3.4A. As neurons react to illusionary edges already at the level of V1 (T. S. Lee & Nguyen, 2001), this reveals a mode of non-local analysis, where the responses of the local operators are influenced by responses of other units.

The purely local edge detection described above fails to mimic human perception as it assumes independence between spatial analysers, and does not model the neural interactions of natural visual processing. Another example where isolated local mechanisms fail is segmentation. For example, the Kanizsa triangle in Figure 3.4A suffices to illustrate the difficulty of separating figure (the triangle) from background, as there is no difference either in texture between the triangle and the background (both are uniform white), nor is there a separating, closed contour. Natural images featuring camouflaged animals are another example where the animal can be correctly

segmented from the background only after it has been recognized, possibly by some of its distinct parts or other cues (see the famous Dalmatian image in e.g. Palmer (1999), p.267). This latter example illustrates that human perception is not "completed" by some static segmentation algorithm, but by one that is influenced by experience acquired during the lifetime, an idea already proposed by James (1899). The conclusion is that segmentation cannot be a preprocessing step that is completed before recognition, and nor can it be a visual subproblem that is studied separately from recognition.

It thus seems that in natural visual processing, many visual subproblems are not considered in isolation. On the contrary, perception seems more like a process of *unconscious inference* (Helmholtz, 1867). In such a setting, the hidden variables regarding the unknown world states are inferred by combining measurement data and prior beliefs, hopefully ending in some maximally likely interpretation fulfilling ecological constraints. Again, the Kanizsa triangle in Figure 3.4A is an example of this: given the biases of the human visual system, most people interpret the figure to represent a white triangle occluding three black discs, instead of taking the discs as three black "pac-men". In this case the inference is called unconscious as no rationalization or logical thinking seems to be involved in making the interpretation. Further, these inferences are dynamic and do not necessarily converge to a stable interpretation, i.e. there may not be a "final" output of visual processing. A classic example of this is the Necker cube, shown in Figure 3.4B. For most people, the perception of the Necker cube is bistable, as it seems difficult to decide which side of the cube is the front side and which is the back, the interpretation oscillating over time.

This kind of dynamic inference in the biological systems may be supported by feedback, and it is well-known that most of the connections in natural visual systems are feedback, not feedforward connections (for a review see e.g. Gilbert and Sigman (2007)). These connections carry information either from one neuron to another in the same area (lateral connections), or relay information back from higher visual areas to earlier ones. Although some reports claim that some functionality such as rapid object detection can be carried out without feedback processing (e.g. Serre, Wolf, et al. (2007)), this claim appears mainly to hold for conditions where there are no obfuscating factors present. If the situation is more complex involving e.g. problems due to lighting, occlusion, camouflage or simply variance in spatial location of the objects to be recognized, object detection may start to involve aspects of more involved inference, including reasoning, visual search and scene interpretation. In such situations it is plausible that looped or feedback processing becomes more important: recent results

suggest that feedback has a major role in such processing phenomena as top-down control (e.g. Bar et al. (2006); Saalmann et al. (2007), for a brief review see Miller and D'Esposito (2005)), figure-ground segregation (Hupe et al., 1998) and contour integration at the level of V1 (W. Li et al., 2008), but also featuring in contextual modulation when viewing natural images (Felsen et al., 2005) as well as in visual illusions (T. S. Lee & Nguyen, 2001). Finally, feedback processing has been proposed to be required for conscious perception (for reviews, see e.g. Tong (2003); Fahrenfort et al. (2008)),

What these research reports amount to is a strong account against the traditional approaches where more or less arbitrary problems are studied in isolation, either isolation on the level of analysis or isolated from other problems. Based on such recent results as mentioned above, it has started to look more questionable whether such independent areas or clear functional segregations exist in the cortex as had been previously proposed (Livingstone & Hubel, 1988). Some theorists already argue that high-level aspects of neural processing in general can neither be segregated from low-level processing (A. J. Bell, 1999) nor from each other: these latter claims include e.g. the nonseparability of perception and cognition (Chalmers et al., 1992), of the connectedness of perceiving, remembering and acting (Thelen et al., 2001). Supportive arguments have even been put forth regarding entanglement of cognitive and affective aspects of biological processing (Pessoa, 2008).

To model and chart the dependencies and interactions between tasks and modules seems one of the great challenges for future modelling. This is demanding due to the difficulty of learning parameters for dynamic systems where multiple components interact over time. Although some interplay between high- and low-level aspects is modelled in Publication 2 of this thesis, here we do not actually model dynamic feedback phenomena in the sense of unconscious inference (Helmholtz, 1867) or Bayesian message passing (see Bishop (2006)). Some studies about dynamic interactions have already been published (Z. Li, 1998; Maass et al., 2002; T. S. Lee & Mumford, 2003; Grossberg, 2003; Deco & Lee, 2004; Deco & Rolls, 2004) that may be more compatible with theories of cognition that attempt to avoid the computational approach (e.g. Gelder (1995)). Likewise, some recent studies have taken critical attitudes towards reductionistic, isolated investigation of low-level function (e.g. Olshausen and Field (2005)). In the future we are likely to have a better understanding as to what kind of visual functionality benefits from feedback processing, and what can be handled with the traditional isolated models.

Chapter 4

Statistical modelling, methods, and visual data

We now turn from the more general issues to the practical details of learning the model parameters in the ecology-driven approach. In particular, we describe the methods we applied in the research for the included publications. We also review some of the issues that should be kept in mind when using such methods. In addition, these issues also allow us to provide modern support for minimalistic philosophies such as the Occam's razor mentioned in the first chapter.

4.1 Modelling with different objectives

As discussed in Sections 3.2.2 and 3.2.3, model parameters can be estimated with relation to different principles and learning objectives. These learning objectives in connection with the data, the constraints and the estimation algorithm designate how the eventual learned model turns out. These factors also have a say in the functional properties of the model when it is used in visual processing.

In the publications of this thesis, we have studied four different objectives for estimating model parameters: independence maximization, response energy optimization, object detection accuracy, and feature selection. Here, the first two objectives lead to statistical models of visual input (as in Section 3.2.2), and the latter two to statistical models of visual function (as in Section 3.2.3). These objectives will be briefly reviewed in the following.

4.1.1 Independence objective

One of the main tools we use to learn models in the publications of this thesis is Independent Component Analysis (ICA), see e.g. Jutten and Herault (1991); Comon (1994); A. Bell and Sejnowski (1995) and Hyvärinen, Karhunen, and Oja (2001) for a textbook that this brief overview is mostly based on. The linear ICA is an example of a linear superposition model $\mathbf{x} = \mathbf{A}\mathbf{s}$, as was described in eq. (3.1) and shown in Figure 3.2 on page 37. As the approach tries to model the density of the data, it is a model of visual input (Section 3.2.2). In particular, the task of the learning is to estimate $\mathbf{A}$ in a manner that the coefficients $\mathbf{s}$ are as statistically independent from each other as possible. Often ICA is proposed in the context of an idea that the signals $\mathbf{x}$ received from the environment would be linear mixtures of several different independent components. As an example, a sound heard might be composed of sounds from several independent sources. The purpose of ICA then is to attempt to find transformations that separate the mixed signal back into its original signal templates $\mathbf{a}_i$ and their weights s_i.

Here we do not assume that a linear transformation can decompose natural images to truly independent components, not only because the real generative process is nonlinear (A. B. Lee et al., 2001), but also due to the contrary having been empirically shown (Schwartz & Simoncelli, 2001; Hyvärinen, Hoyer, & Inki, 2001). Instead we motivate the independence maximization from the viewpoint that even approximative independence may have utility. In particular, we are interested in the consequences that independence maximization transformation may have for further probabilistic modelling using the coefficients $\mathbf{s}$. Now, considering a non-singular $\mathbf{A}$ in the linear superposition model, then in the first place $p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{A}| p_{\mathbf{s}}(\mathbf{s})$, e.g. Hyvärinen, Karhunen, and Oja (2001). Further, if the dimensions of $\mathbf{s}$ are independent, the joint distribution of the variables is factorizable (Papoulis, 1991), allowing eq. (3.3) to simplify into the following decomposition of the density of $\mathbf{x}$,

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{A}| p_{s_1}(s_1) p_{s_2}(s_2) \dots p_{s_n}(s_n). \tag{4.1}$$

Sometimes the processing that attains eq. (4.1) is called a *factorial coding* (Field, 1994). The benefit of such a decomposition is that the potentially intractable density $p_{\mathbf{x}}$ can be represented in terms of n one-dimensional marginal distributions p_{s_i}. These marginal distributions can be significantly simpler to estimate and handle: for the sake of argument, consider k binomial random variables. Then, the number of parameters in the joint distribution may grow as $O(2^k)$ if the variables are dependent. But, should

the variables be independent, this complexity is reduced to $O(k)$ parameters. Another benefit of factorial densities is that it is very simple to sample new data from them, as it suffices to sample separately from each marginal $p_{s_i}(s_i)$ to get a random vector $\mathbf{s} \sim p_{\mathbf{s}}(\cdot)$. In the case of linear superposition model, eq. (3.1) can then be used to construct the actual image.

In practice, eq. (4.1) may hold only approximately for $\mathbf{A}$ estimated by ICA, and the subsequent inferences utilizing such factorization may be biased accordingly. Also, the linear superposition model may not closely correspond to the way that the data in question was formed. Nevertheless, ICA presents an interesting and tractable starting point that can suggest meaningful decomposition of natural images, especially in terms of finding image constituent types with weak dependencies. One intuitively pleasing example of this is the segregation of shape and colour in ICA modelling of colour images. In that case, some ICA features in $\mathbf{A}$ turn out to code for colours, and others for shapes (Hoyer & Hyvärinen, 2000), although the division is only approximative (possibly due to the fact that for a linear model, there can not be a spatial colour change template without it including a shape).

It should be emphasized that the transformation performed by ICA may not be helpful for arbitrary further processing. For example, if the linear transform that ICA performs in eq. (3.2) is invertible, all the information in $\mathbf{x}$ is retained, and learning algorithms may be able to incorporate the inverse of the ICA transform by $\mathbf{W}$ into their learned models, should their objective functions prefer that. Subsequently, the learning may be oblivious to the ICA transform having been performed in the first place. For further similar reasoning, see (Vicente et al., 2007).

The only ICA learning algorithm applied in the publications of this thesis is the FastICA algorithm (Hyvärinen, 1999). This choice is mainly due to the fast convergence of the method and the finding that the qualitative differences between the models learned by the different ICA methods are often small (Hyvärinen, Karhunen, & Oja, 2001). Although we do not go into the algorithmic details of the efficiency behind FastICA here (the reader is referred to Hyvärinen, Karhunen, and Oja (2001)), a few words are in order regarding the principles that the estimation relies on.

At its core, FastICA assumes that input data $\mathbf{x}$ is whitened and here we denote such data as $\mathbf{z}$. The data being white means that it has a mean $E[\mathbf{z}] = \mathbf{0}$ and covariance $E[\mathbf{z}\mathbf{z}^T] = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Starting from $\mathbf{x}$, these properties can be attained by removing the empiric means of each variable of $\mathbf{x}$, followed by a whitening transform, as described in Section 4.3. The whitening transformation has the effect that the dependencies

as described by covariances are eradicated, and hence the optimization that FastICA performs is necessarily based on higher-order statistics beyond covariances. In effect, FastICA tries to search for a matrix $\mathbf{W}$ that among the different possible decorrelating transforms has the smallest higher-order statistical dependencies between the dimensions of $\mathbf{s}$. In FastICA, this is approached through the realization that non-Gaussianity and independence are related, as due to the central limit theorem, sums of independent random variables tend towards Gaussianity under suitable conditions. Then, it makes sense to attempt to search for projection directions that are not composed of sums of several independent variables and deviate from Gaussianity. A practical way towards this is to maximize such higher-order statistics that are fixed for any Gaussian data. For example, the fourth moment or kurtosis of the filter responses could be maximized (Hyvärinen, Karhunen, & Oja, 2001). As non-Gaussian properties such as high kurtosis are also the distinguishing property of sparse random variables, FastICA and sparse coding approaches are closely connected (Olshausen & Field, 1997).

One caveat of linear ICA estimation is that it may not converge if the data contains more than one Gaussian direction (Hyvärinen, Karhunen, & Oja, 2001). The reason for this is that the Gaussian subspace does not have any structure beyond that of covariances, and hence all projection directions in $\mathbf{W}$ with relation to it appear equally good. But due to the real generative process of natural images, image data do not seem to have such Gaussian subspaces. Another potential source of problems is that to the ICA estimation in practice requires preprocessing and dimensionality reduction, neither of which are exactly part of the probabilistic formulation of the invertible linear ICA. We will discuss preprocessing further in Section 4.3.

4.1.2 Response energy objective

Although the perceptual environment has properties that may be close to independent in general (for example, location, shape, colour, and motion) it is nontrivial to recover these properties from the input images. In particular, linear filters do not appear to be flexible enough to recover independent properties from natural images without residual dependencies remaining in the filter outputs. One well-known class of residual dependencies is that of *energy dependencies*: the filter responses can be correlated if the response signs are discounted. In other words, the covariance between $|s_i|$ and $|s_j|$ for $i \neq j$ tends to be non-zero when the two responses originate from two different linear filtrations. These energy dependencies do not seem to be

due to the used filter parameter estimation method, but rather because of the nature of the data and the linearity of the filtering. These dependencies have been studied in previous work e.g. by including a nonlinear normalization of the filter outputs to the model (Schwartz & Simoncelli, 2001), leading to reduction in these kinds of dependencies, but also in learning complex cell models (Hyvärinen & Hoyer, 2000) and in forming topographic organizations for a bank of linear filters (Hyvärinen, Hoyer, & Inki, 2001).

In Publication 6, we study the learning of parameters for a subclass of quadratic filters by optimizing filter response energies. In particular, we assume that model responses are computed nonlinearly as $s = (\mathbf{w}^T\mathbf{x})(\mathbf{v}^T\mathbf{x})$, and then optimize for sparsity of s (for example $E_{\mathbf{x}}[|s|]$). In Publication 6, we show that when s originates from such a product, optimizing sparseness entails optimizing the energy correlations of the paired filters. As this is also shown to provide implicit sparseness objectives for the underlying linear filters, the method can be also seen as performing ICA, atleast before the multiplicative computation of the paired responses takes place. This relation to ICA is due to the intimate connections between ICA and optimization of sparseness (Olshausen & Field, 1997). Thus, as the model practically includes an underlying ICA model for the linear filters, it can be seen to belong to the category of models of visual input. In Section 5.2.2 we will describe how both maximization and minimization of response energies (and thus energy correlations) in this setting can lead to interesting results and emergent features from natural images.

4.1.3 Object recognition objective

Publication 2 and Publication 3 of this thesis study supervised models of object recognition, a learning setting that leads to estimation of models of visual function, as was described in Section 3.2.3. As a reminder, in the supervised setting we work with training datasets of pairs $(\mathbf{x}, y)$ where $\mathbf{x}$ is a vectorized image (or some vector derived from the image), and y is the preferred response value supplied by a teacher. We also assume here that a single image either contains an object out of $k+1$ known classes C, where the last class $k+1$ is used to denote background or "no object" images. Then, $y \in \{1, ..., k+1\}$.

In Publication 3, the main emphasis is on online selection of visual features that are used for object recognition, but given such features the decision mechanism we use is particularly simple. In Publication 3, we chose to use a Naive Bayes classifier (e.g. Hastie et al. (2001)), and here we give a more elementary treatment of it. The main simplifying assumption

in the Naive Bayes classifier is to assume the input variables in $\mathbf{x}$ to be conditionally independent given the class (a simple example of such a case is shown in Figure 4.1 for convenience). Now consider

$$P(C = j|\mathbf{x}) = \frac{p_{\mathbf{x}|C}(\mathbf{x}|C = j)P(C = j)}{p_{\mathbf{x}}(\mathbf{x})}, \tag{4.2}$$

where the right hand side follows from the Bayes formula. The Naive Bayes classifier now predicts the class $\hat{y}$ that has the highest probability given $\mathbf{x}$,

$$\hat{y} = \underset{j \in 1,\ldots,k+1}{\operatorname{argmax}} P(C = j|\mathbf{x}) \tag{4.3}$$

$$= \underset{j \in 1,\ldots,k+1}{\operatorname{argmax}} \frac{p_{\mathbf{x}|C}(\mathbf{x}|C = j)P(C = j)}{p_{\mathbf{x}}(\mathbf{x})}. \tag{4.4}$$

As $p_{\mathbf{x}}(\mathbf{x})$ is equal for each class, it can be ignored, and to learn the classifier it is sufficient to estimate the probabilities $P(\mathbf{x}|C = j)$ and $P(C = j)$ from the training data. The first quantity, given the assumption of conditional independence, simplifies to a product of one-dimensional marginal densities,

$$p_{\mathbf{x}|C}(\mathbf{x}|C = j) = \prod_{i=1}^{n} p_{x_i|C}(x_i|C = j). \tag{4.5}$$

Now, if the variables x_i are binary (as they are in Publication 3, where we operate on a binary feature space derived from natural images), to learn the classifier it suffices to compute empiric estimates of marginal Bernoulli distribution parameters from the training data, i.e. estimate the probability of one for each feature x_i. In addition, we need to measure the class frequencies $P(C = j)$ from the labels y. In this case, the Naive Bayes classifier can be implemented very simply as a vector counting the number of occurrences of each object class, and measuring one additional vector per class, where each dimension i counts the number of occurrences of $x_i = 1$ inside that class. These counters can be easily updated one training example $(\mathbf{x}, y)$ at a time. Despite this favourable property and the simplicity due to the conditional independence assumption, Naive Bayes classifiers are generally noted to perform well (e.g. Hastie et al. (2001), p. 185).

For Publication 2, we utilized a more involved learning mechanism called a Support Vector Machine (SVM), e.g. Vapnik (1998); Cristianini and Shawe-Taylor (2000). This method proceeds directly from optimization principles concerning classification accuracy and does not estimate conditional distributions as were required for the Naive Bayes classifier.

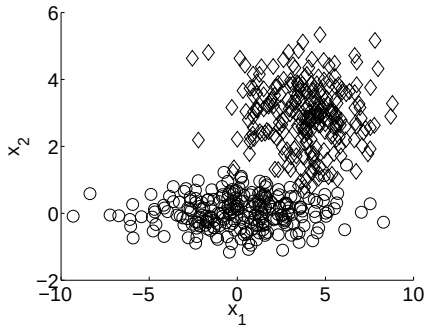

Figure 4.1: A dataset of two-dimensional examples $\mathbf{x}$ from two different classes, marked with circles and diamonds. The instances of each class originate from two-dimensional Gaussian distributions with in-equal variances and means. Although the variables x_1 and x_2 are correlated over the data in general, inside each class they are independent of each other, i.e. the variables are conditionally independent given the class.

In principle this could mean less parameters to fit. The SVM approach typically incorporates several different ideas, including regularization, margin maximization and possibility to learn nonlinear classifiers through the use of kernels. Of these, regularization is used to constrain the complexity of the classifier (see Section 4.2.3) to avoid overfitting. The margin maximization idea follows from the observation that in high-dimensional spaces, typically many different decision surfaces allow to classify training data correctly. In such a case, it seems natural to choose the decision surface that is furthest away from the examples of the classes (possibly in some weighted sense, should the classes overlap). Finally, the kernel trick is equal to fitting a linear classifier after a basis expansion (Hastie et al., 2001), but with application of kernels, this expansion does not have to be explicitly performed.

The only SVM variant we consider in Publication 2 is the 1-norm *soft margin* linear SVM (note that linear classifiers are not literally linear, as the decision is a nonlinear operation – if the decision is based on a computation of $\mathbf{w}^T\mathbf{x}$, the classifier is commonly called linear). The soft-margin SVM learning mechanism for binary classification problems ($y \in \{-1, 1\}$) and l

training examples is defined as an optimization problem as follows,

$$\texttt{minimize} \quad \mathbf{w}^T\mathbf{w} + \mathcal{C}\sum_{i=1}^{l}\xi_i, \tag{4.6}$$

$$\texttt{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i, \tag{4.7}$$

$$\xi_i \geq 0, \forall i, \tag{4.8}$$

where $\mathbf{w}, b, \boldsymbol{\xi}_\mathbf{i}$ are the parameters to optimize, and $\mathbf{w}^T\mathbf{w}$ can be taken as a regularizer to suppress solutions $\mathbf{w}$ with high norms. The prediction of the learned classifier is computed as $\text{sign}(\mathbf{w}^T\mathbf{x} + b)$, and hence the prediction is correct for example $\mathbf{x}_i$ if and only if the left hand side of the inequality in the constraint (4.7) is positive. But in that constraint, we also prefer the margin of each prediction to be 1 or more. If that is the case, no extra benefit is gained in the definition by pushing correctly classified examples even further from the hyperplane defined by $\mathbf{w}$. But if the margin is smaller than preferred, or negative, the constraint (4.7) becomes active, and we incur a penalty in the objective (4.6).

For solving the above optimization problem, as well as for details, proofs, and bounds about the SVM approach, the reader is referred to Vapnik (1998); Cristianini and Shawe-Taylor (2000) and the references therein.

For us, the main reason for using SVMs was that the approach is typically competitive in a wide variety of learning tasks, and it is not relying on any independence assumptions between the variables. Also, as the optimization task is typically convex (a quadratic program in the case of 1-norm soft-margin SVMs), the optimization can be guaranteed to converge to the globally best solution (Cristianini & Shawe-Taylor, 2000).

A downside of the SVM framework is that the estimation of the common SVM models is done using batch learning (i.e. all examples are required to be available at the same time for learning the model). At the time that we worked on Publication 3, established online mechanisms for SVM estimation did not exist, and that was the main reason we resorted to the simpler Naive Bayes approach in that work.

4.1.4 Feature selection objective

With the previous objectives, we assumed that there is a function class with parameters to tune, for example a set of linear filters. In such a case the optimization modifies these parameters according to the used objective. As a consequence, the learned model parameters typically end up representing features of the data.

An alternative way to end up with useful features is to select the features from the data instead of optimizing the feature parameters. This approach, called *feature selection* (Dash & Liu, 1997; Guyon & Elisseeff, 2003) is typical in such machine learning problems where a multitude of measurements is available for each training instance. In such a setting, we wish to select a small set of the features that suffice for the classification or regression problem at hand. In the case of images, the original features are pixels, and selecting some set of pixels may not be very useful for subsequent tasks. Instead, we can select some more complex visual features, for example small image templates, and then match them by correlation to new images (e.g. Ullman et al. (2002)). The resulting binary feature will then work as an indicator of the template presence.

In general, learning features and learning classification can be seen to belong to a single continuum: at one end, some high-level feature may closely reflect the class label (and then very simple classifier may be sufficient), whereas in the other end, the features can be very low-level ones, such as pixels (and a very complex classifier may be needed). Hence, if the set of features is selected for a model of some particular function, the features do not need to faithfully describe the statistical distribution of the whole visual input. Instead, the selected features are better seen as simple computations that may allow further processing to be simple, i.e. they are used to simplify a later model of function. The features do not need to represent anything of the input that the function does not need. It follows that features selected for function can also be taxonomized as parts of modelling visual function rather than the input in general.

In Publication 2 and Publication 3, we selected image template features with object recognition in mind. In Publication 2, we studied an information-theoretic objective to maximize the additional information that each new feature brings about the class (for a textbook on information theory, see Cover and Thomas (2006)). The actual equations and the proposed algorithm are described in Publication 2 and not reproduced here.

Measuring the additional information in the traditional way requires the whole training set of data to be present at the time of learning. In Publication 3, we proposed a novel, probabilistically grounded approach to select visual features online in a process that sees one training example at a time and then discards it. This method is technically described in Publication 3, and the reader is referred there.

We will discuss feature selection further when we review the corresponding included publications in Sections 5.3.1 and 5.4.

4.2 Intricacies in statistical learning

A serious challenge in the ecology-driven approach is that statistical modelling is a science in itself, and textbooks on the subject are ripe with describing and characterizing the issues involved (Vapnik, 1998; Cristianini & Shawe-Taylor, 2000; Duda et al., 2000; Hastie et al., 2001; Bishop, 2006). In the following we will discuss some of these well-known issues, and to make the presentation more intuitive, we describe how this learning theory supports the classic tradition in science that descriptions and models should be as simple as possible, but not simpler.

4.2.1 Local optima

If a model is optimized with relation to some objective, the resulting objective function may have several variable configurations that are the best possible in their own local neighbourhoods in the parameter space, but not necessarily globally optimal. This is intuitively explained through the concept of optimization landscape: assuming that $h(f)$ is the objective function value for a given model f, then as we try out different parameters of f (such as the weights in eq. (2.1)), a landscape of objective values of h is drawn. For example, Figure 3.3 could also represent an optimization landscape for two parameters x_1 and x_2 in some optimization problem. The goal of optimization is then to find the highest peak or the deepest pit in the landscape, depending on if we are maximizing or minimizing. Generally, starting from an arbitrary point in an optimization landscape, the local neighbourhood might not give any indication what the globally best parameter configuration is, and simply following the gradient of the landscape can lead to a locally optimal set of the parameters instead of getting to the globally optimal solution.

A significant exception to the general difficulty of global optimization are the so-called *convex* landscapes that only have single optima, allowing for efficient search. Unfortunately, aside convex functions, no function classes seem to exist that would in a certain sense be "larger" than the class of convex functions and still guaranteed "tractable" in the optimization sense (Kreinovich & Kearfott, 2005). Also, it is well-known that there is no optimization algorithm that would perform well for all optimization problems (Schaffer, 1993; Wolpert & Macready, 1997). In practice these latter results are overly pessimistic, as natural problems tend to have structure, and a local optimum might be good enough.

The problem of local optima holds for *any* nonconvex optimization that tries to fit models to measurements, regardless of the manner the fitting is

performed. Hence any mathematical theory (e.g. in physics) that is based on fitting parameters to measurements is prone to being only a locally best model, and this seems more likely the more complex the related optimization landscape is. It should be noted that even if the objective function is convex, the landscape can still be nonconvex if the model is nonlinear. As a result, optimization theory would suggest using as simple models as possible in order to increase the chances of finding good model parameters.

4.2.2 Overfitting and model selection

Suppose for a moment that we could always find the optimal parameters for our model given the data. Then in a naïve sense, the more complicated model we can fit to the data, the better, as a more complex function will always have a better potential to fit the data than a simpler function. But in the worst case this will result in memorizing the training set. A simple example of this is curve fitting: allowing a complex curve to be fitted to noisy data, a learning algorithm may make the curve pass through all the points, and not learn the "true" curve. An example more in line with the context of this thesis is to consider the linear superposition model and an objective of sparseness. Now, if we allow the matrix $\mathbf{A}$ to have any number of features, then each training example $\mathbf{x}_i$ can be taken as a feature $\mathbf{a}_i$, and subsequently $\mathbf{x}_i$ can be represented simply by setting the coefficient s_i to 1 and the rest to zero for all $s_j, j \neq i$. But, given a new $\mathbf{x}$, this model may however fail to find a sparse representation for the new data point. This behaviour where the model performs as wanted on the training data but fails on new data is called *overfitting*, with its opposite called the *ability to generalize*. Overfitting is made more likely by an expressive model class and an under-constrained objective. As a result, the model can not provide a concise explanation of the data, and at the extreme, the overfitted model may fail to explain anything about the data.

Overfitting can be controlled either implicitly or explicitly. Implicit controls include such choices as selecting a simple model class, and reducing the dimensionality of the input data. Such implicit mechanisms are used in the ICA experiments of this thesis. The other option is to explicitly control overfitting during the learning by *regularization* that penalizes complex function shapes. Regularization is especially typical in the context of Support Vector Machines. For example, in the soft-margin objective (4.6), the term $\mathbf{w}^T\mathbf{w}$ works as a norm-penalizing regularizer (Cristianini & Shawe-Taylor, 2000), and this kind of regularization is used in Publication 2 of this thesis. Yet another choice is to perform *model selection* in the more traditional sense of using some measure that weights the models confor-

mance to the data against the models complexity. The Akaike Information Criterion (AIC) is one example of such a measure, see Akaike (1974). In the publications of this thesis we do not use such traditional model selection mechanisms as AIC. Instead, we tend to evaluate the quality and generalization capability of the learned models by a practical, empirical technique called *cross-validation*, which involves partitioning the used dataset and then training and testing the model on separate data partitions (e.g. Kohavi (1995)).

Hence, in addition to optimization considerations, the philosophical stance of preferring simple models can also be supported from the viewpoint that they are more likely to generalize to new data than complex models that are more easily overfitted.

4.2.3 Further issues

Not only it is difficult to select a suitably constrained model class, but also it may be cumbersome to perform the processing as suggested by the model. For example, in the case of probabilistic models, getting the functional forms properly normalized (so that they integrate over the data space to one) typically requires the computation of a tedious integral called the *partition function*. This has consequences for maximum likelihood parameter estimation, as the approach requires the evaluation of the partition function. For more elaborate densities, analytical evaluation of the partition function is often considered intractable. Techniques such as Markov-Chain Monte Carlo (J. S. Liu, 2001), Contrastive Divergence (Hinton, 2002) and Score Matching (Hyvärinen, 2005) can be attempted to alleviate this problem.

Another issue is that the data often does not follow the assumptions of the learning method. One concrete example of this is the natural image data not following the linear superposition model of ICA and eq. (3.1). Another requirement that we have so far ignored is that strictly speaking, derivations of many estimation methods (such as the maximum likelihood estimation) assume the training examples to be independent and identically distributed (the so-called *i.i.d.* assumption). Especially in vision where the visual input is naturally seen in a continuous sequence, it is not ultimately valid to assume that one moment of visual data would be independent from the previous moment of data. Further, the data distribution may not be identical from one moment to the next, for example because living beings can alter their environments. Nevertheless, the learning methods often seem to perform well regardless of the data violating the i.i.d. assumption. A partial explanation for this is that the batch methods are typically invariant

to the order that the training data is presented in.

A final problem we mention that may have consequences in many modelling attempts is that the learning algorithms and the statistical estimators they rely on may not be *robust*. A simple example is the computation of the empiric mean of k samples y_i, computed with $1/k \sum_{i=1}^{k} y_i$. Here a single erraneous y being far away from the bulk of the data can pull the weighted sum away from the "true" population mean. In a similar fashion, more complex model parameterizations and parameter vectors may end up seriously affected by just a few noisy examples or other outliers in the training data. These issues may be addressed by using more robust estimators (such as the median) or removing the rogue samples from the training data (Huber, 1981; Hampel et al., 1986). In the scope of the publications of this thesis, image preprocessing and the choices of optimization objectives (tanh() nonlinearity in FastICA and soft-margin maximization in SVMs) can be taken to be partly due to enhance robustness.

4.3 Natural image data and its preprocessing

All the publications of this thesis have to do with natural images in one way or another; in some publications we examine generic low-level models, decompositions and statistics of natural images that are estimated in an unsupervised manner, and yet in others we have explored the problem of learning object recognition from such images. Hence, in all the cases, the data is a set of natural images, meaning photographic, greyscale images as captured by a camera. In some studies, natural images are taken to denote photographs of rural scenes with no man-made objects (e.g. in Frazor and Geisler (2006)). In this thesis we do not rule out images with man-made objects; their naturality can be understood by realizing that man-made objects are natural in the environments of the civilized animal. Hence we define that natural images are those that portray *some ecologically meaningful environment*.

For clarity, we show one natural image in Figure 4.2, taken from a set provided by Hateren and Schaaf (1998), a dataset that we use in the majority of the publication of this thesis. Typical to these images is the relatively high resolution, and this exactly corresponds to high data dimensionality in the sense of machine learning should the raw images be considered as training examples. The input image shown in Figure 4.2 with its resolution of 1020×1532 pixels would make a data vector of over a million variables, each with a precision of 12 bits (in the case of the mentioned dataset). This resolution seems still smaller than the discerning ability of the human

Figure 4.2: A typical image of 1020×1532 pixels from van Hateren's and van der Schaaf's natural image dataset (image no. 123). Such images are used as data in the majority of the publications of this thesis. The displayed image was logarithmically transformed to enhance visibility.

retina (by a naïve comparison to the number of retinal sampling elements (Williams & Moody, 2004)), but large enough to make such natural images as reasonable first approximative models of the signal that is received by the retina.

In studies that estimate statistical models based on natural images, the visual data is often not used as-is to learn a model, but only after being subjected to some "preprocessing", often off-the-sleeve transformations that are performed on the images, but that may not have been learned from the data and that may lack analytical motivation. They may also be outside the probabilistic modelling framework, and be performed before such modelling – hence preprocessing can sometimes be taken as an indication of less rigorous aspects being present in the used modelling approach.

In the case of van Hateren and van der Schaaf's images, the preprocessing transformations typically applied include a compressive transform similar to one performed by the retina, as well as block averaging, and finally whitening (Hateren & Schaaf, 1998). We will now briefly describe these three steps in the context of this dataset.

A compressive transform such as a logarithm (Hateren & Schaaf, 1998) or Naka-Rushton equation (Naka and Rushton (1966), also in eq. (2.2)) is used to address the highly skewed intensity distribution of these images.

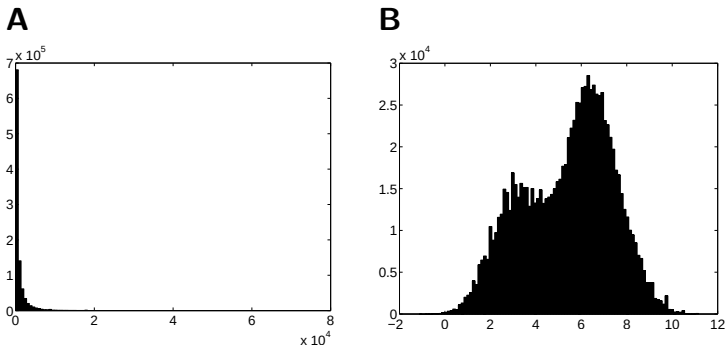

Figure 4.3: Intensity distribution of pixels sampled from the van Hateren and van der Schaaf dataset. **A)** A histogram of intensities (in cd/m^2) of the raw pixels. The thin right tail is very long but has a negligible density: 90% of the data is dimmer than 4×10^3, a value that is roughly 3% of the maximum intensity seen in the data. **B)** The histogram after a logarithmic transform of the pixels x, $\log(x + \epsilon)$. The new dynamic range is better utilized (see text).

This transformation makes the distribution more balanced, as shown in Figure 4.3. In terms of image processing, natural images with highly skewed and concentrated histograms can be considered problematic, as they use the dynamic range inefficiently (Bovik, 2000), and controlling for this can help information transfer (Laughlin, 1981) in neural coding. In unpublished experiments, we examined the changes in the dynamic ranges for the used dataset by doing the following: first, we measured the spread of the studied image dataset by estimating the difference between the 75% and the 25% intensity quartiles in each image, and averaged the obtained estimates for 500 randomly selected images from the dataset. We call this the spread of the distribution. We then compared this average spread to the average spread computed from artificial data with an uniform distribution having the same intensity range, and obtained that the spread of van Hateren's images is approximately only 3% of the spread of the uniformly distributed data (i.e. the one that has maximal equalization of its histogram). After taking a logarithm of the natural images, this relative spread increased to 16%, showing that the new dynamic range is better utilized and that the distribution tails have been brought to the same scale with the bulk of the data. A more principled way to compare the dynamic ranges would have been through entropy (Cover & Thomas, 2006), but the more simple experiment used here seems sufficient to illustrate the point.

Although the effect of highly skewed distributions for learning is not completely understood, such distributions can cause problems for learning algorithms. This is due to the rare occurrences of very high intensities (corresponding to the right tail of the distribution) may appear as outliers in the data, and have very strong effects for e.g. non-robust gradient computations. The compressive transform may mitigate such effects and cause empirical statistics to behave better (e.g. Ruderman (1994)).

The block averaging, on the other hand, can be done to address noise and calibration deficiencies in the acquired image data (Hateren & Schaaf, 1998). It also has another kind of smoothing effect on the data: for the linear intensity version of the used dataset (".iml" filename extensions), we measured that approximately only 38% of the discrete intensity values between zero and the dataset maximum value are actually used (unpublished experiments). This is likely due to the image capturing process, and not a sign of such quantification being a natural phenomenon. A block averaging can smooth away this effect by taking averages of the nearby pixel values. For example, in 2×2 block averaging, each square block of 4 pixels gets replaced by its mean value. This effectually halves the image resolution, but the resulting images are still large enough to be intractable as raw data for many learning methods.

The last preprocessing step often applied is called *whitening* (or sphering), a transformation that makes the data have an identity covariance, i.e. the transformed variables are decorrelated. Whitening is required for the application of certain ICA methods such as FastICA (Hyvärinen, 1999), and the whitening stage seems reasonable if the objective is to find independent directions. This is because whitening removes the second-order statistical dependencies of the data and allows the further measures of dependency to be oblivious to the data covariance structure. The fact that covariance can be taken as unity can also be useful for technical derivations of the learning methods (Hyvärinen, Karhunen, & Oja, 2001).

Whitening can be performed in several ways, for example through filtering (Olshausen & Field, 1996) that mimics retinal processing (Atick & Redlich, 1992; D. J. Graham et al., 2006), or via Principal Component Analysis (PCA, e.g. Hyvärinen, Karhunen, and Oja (2001)). In the majority of the publications of this thesis, whitening is performed by the latter method. Given that PCA is an eigenvalue decomposition of the data covariance matrix[1], it can be shown that projecting the data to the PCA axes (eigenvectors) with each axis rescaled by its inverted eigenvalue leads to the

[1]PCA finds a basis of orthonormal directions where every $i + 1$:th direction has the maximal variance on the condition that it is orthogonal to the previous i directions.

projected data having an identity covariance (e.g. Hyvärinen, Karhunen, and Oja (2001)). As the covariance matrices of natural data can be close to singular, dimensions corresponding to very low eigenvalues may need to be dropped in order to invert the eigenvalues for whitening. This leads to dimensionality reduction that may also have the salient effect of noise reduction. In terms of the square error of data reconstruction, PCA is th optimal linear method, supporting its application in dimensionality reduction. Although there does not seem to be a single universal way to choose the appropriate number of retained dimensions in PCA, in the case of natural images a reasonable value can be picked heuristically, i.e. by reconstructing the data and verifying visually if the reconstructions retain an acceptable amount of visual structure and detail. In natural images, the discarded PCA components correspond to high frequencies in the Fourier analysis sense (e.g. Hancock et al. (1992); A. J. Bell and Sejnowski (1997)). This is in agreement with the $1/f^{\alpha}$ power spectrum of natural images: the high frequencies with the least power are discarded. Subsequently, the whitening that rescales the PCA axes by the eigenvalue inverses boosts the high frequencies and dampens the low frequencies, effectually corresponding to making the power spectrum approximately uniform over the data.

Together, the described preprocessing steps can be taken to correspond to a very simple model of retinal operation. However, it should be noted that these preprocessing steps are by no means mandatory for all "natural" image data, as the images may already be preprocessed, either explicitly by image processing techniques or implicitly in the cameras that may capriciously perform diverse operations on the data without the user being knowledgeable of the algorithms used. Eventually the data we receive and think of as natural may have already gone through a lot of processing. For example, for the car detection dataset (Agarwal et al., 2004) that we used in some of the publications of this thesis, none of the preprocessing steps mentioned above are called for: the images do not have skewed intensity distributions, the images are already scaled to a small size of 100×40 pixels (scaling is effectively similar as block averaging), and we did not examine the effect of whitening with that data.

Hence static, fixed preprocessing steps are not needed for all images, and in settings like computer vision or content-based image retrieval, it is not clear that nonadaptive, fixed preprocessing should be used. In the natural operating conditions of deployed systems, the images may come from a wide variety of cameras, have different resolutions and portray a variety of visual environments. Subsequently, the necessary processing steps must take into account the nature of each input image individually.

Chapter 5

Learning visual processing

After overviewing the area of study and the used methodologies, this chapter finally reviews the main technical content of this thesis. The actual content is included at the end of this thesis as reprints of the original publications. The original publications have been categorized into four topical categories, and instead of proceeding in the chronological order of our research, we present the topics with respect to their perceived position in the classic visual hierarchy (see Section 2.2.2). We start from the publications studying low-level issues, and move towards higher-level aspects. The publications studying higher-level phenomena should be taken as more speculative, as they study the question if something is possible in general, with less regard to the biological plausibility of the used mechanisms.

In Section 5.1, we describe our results related to learning low-level visual processing from dependencies in images. Next, Section 5.2 overviews our attempts towards learning more complex visual features with unsupervised learning. This is followed by Section 5.3 describing learning mechanisms for priming. Finally, we conclude with Section 5.4 about online feature selection for object recognition.

5.1 Low-level statistical dependencies in images

As explained in Section 3.2.1, studies of dependencies in images are not only motivated by the need to design tractable probabilistic models for natural image data, but also to understand what is appropriate processing for such data. Of specific interest is to examine what properties are independent in natural images, and how such properties could be extracted by the visual system. This is due to the fact that in both computational modelling as well as in biological systems, if independent constituents of images can be

extracted by suitable computations, these parts can be modelled separately with no need for communication between the involved parties (see also Section 3.3).

Previous work on learning image dependencies with ICA-like methods suggests that lines and bars are the independent components of natural images (Olshausen & Field, 1996; A. J. Bell & Sejnowski, 1997; Hateren & Schaaf, 1998), although residual dependencies are known to remain (Hyvärinen & Hoyer, 2000; Schwartz & Simoncelli, 2001; Hyvärinen, Hoyer, & Inki, 2001; Inki, 2004). ICA-like techniques have also been used to study colour and stereo images (Hoyer & Hyvärinen, 2000) as well as in spatiotemporal (video) data (Hateren & Ruderman, 1998). Here we will examine the representation and processing that ICA learns from local contrast images. We also study how the structure of local contrast and local luminance relate statistically.

5.1.1 Structure of local contrast

Local contrast has been of significant interest to vision research for several reasons. First, the used artificial test stimuli in psychophysics typically vary in two primary dimensions that are called *luminance* and *contrast* (Peli, 1990; Moulden et al., 1990; Bex & Makous, 2002; Badcock et al., 2005; Sukumar & Waugh, 2007; Allard & Faubert, 2007). Second, the retinal ganglion cells encode and transmit the visual data as contrasts according to the textbook view (Meister & Berry, 1999; Masland, 2001). Third, recent research suggests that splitting images to luminance and contrast might allow access to two independent properties of the data (Mante et al., 2005; Frazor & Geisler, 2006). Thus, it makes sense to examine how the statistical structure of the images is changed when they are converted to local contrast images.

In **Publication 5**, we add to the previous statistical studies of contrast (Peli, 1990; Brady & Field, 2000) by examining the statistical redundancies in a certain biologically plausible contrast representation. We show that the statistical, spatial redundancy structure of these contrast images is not very different from that of the original intensity images. To compute the nonlinear measure of local contrast, we first perform the usual whitening transform on the images (see Section 4.3). In particular, we specify the applied transform to be centre-surround filtering, roughly similar to the one of the retinal ganglion cells. After the whitening, we rectify the image to arrive at a contrast image. These new contrast images are the ones we subsequently study, first by applying some traditional measurements, and then by ICA.

Based on the results described in Publication 5, we suggest that the contrast transformation does not alter the spatial structure of natural images in any revolutionary manner. Instead, the contrast images are quite similar to the original images in terms of the applied statistical techniques: the Fourier power spectrum of the local contrast images resembles that of the original images, having a power-law like correlation structure on average. This suggests that the decorrelation achieved by the centre-surround whitening transform can be abolished by pointwise rectification. Likewise, the familiar, localized, oriented and bandpass processing that resembles Gabor-filtering can be learned by ICA from these contrast images as well. However, it should be noted that with the contrast transformation included, the linear filters learned by ICA are nonlinear filters with relation to the original image data. This learned nonlinear processing may be able to detect such texture variations that are not salient to linear filtering, just as second-order filter-rectify-filter processing models can do (Johnson & Baker, 2004).

5.1.2 Relations between local luminance and contrast

In **Publication 7** we extend the previous results of Mante et al. (2005); Frazor and Geisler (2006). We show that the weak dependencies between local luminance and contrast as observed by single variable analysis are no longer the norm in a spatial setting, where multiple measurements of local luminance and contrast are available. Instead, the local contrast measurements become highly predictable from the local luminance measurements, and optimization for independence leads to processing that integrates luminance and contrast instead of segregating them. Thus, contrary to previous proposals of independent processing of luminance and contrast (e.g. Sukumar and Waugh (2007); Allard and Faubert (2007)), our results suggest that these two properties are highly redundant, and at least based on independence arguments, there is no immediate reason to suppose separate pathways or other segregations for spatial processing of these qualities.

The main difference between our approach and that of Mante et al. (2005) is that the latter considers local luminance and contrast for gain control of retinal ganglion cells and not as actual information to be encoded and transmitted. For purposes such as retinal gain control, taking the local luminance and contrast as independent may remain meaningful, as one gain controlled neuron may have to tune its gain according to the data under its own receptive field, and in this local context, the variables of local luminance and contrast are typically only weakly dependent (Frazor & Geisler, 2006). Our results add to this understanding by showing that later

spatial processing may not trivially benefit from similar local luminance and contrast split, which would be the case if the split allowed the system to access two independent channels of spatial information.

Together, Publication 5 and Publication 7 lead to question the functional significance of having an explicit spatial contrast representation in the first place. For example, numerous models of object detection (e.g. Riesenhuber and Poggio (2000); Schneiderman and Kanade (2002); Viola and Jones (2003); Amit et al. (2004); LeCun et al. (2004); Agarwal et al. (2004); Fei-Fei et al. (2006); Leibe et al. (2008)) do not resort to computing an explicit contrast image in the sense of our publications, and neither they follow the psychophysical idea of a strict dichotomy between luminance and contrast.

Instead of being directly useful for spatial analysis, it remains a possibility that contrast may have more use in gain control and dependency reduction than being something worth explicit representation (see the discussion on 'variance fields' in Simoncelli (2005)). However, it is known that signed contrasts, in terms of wavelet coding (Daubechies, 1992), can be used to attain coding-efficient image representations with only a few variables representing luminance (or base levels) and the most representing signed contrasts. This kind of coding, and Gabor and V1-like transforms in general, can assist in tasks such as object recognition (Jain et al., 1997; Serre, Wolf, et al., 2007) and segmentation (Jain & Farrokhnia, 1991; Sharon et al., 2006). Nevertheless, the practical utility of unsigned contrast as studied in Publication 5 and Publication 7 remains to be explored.

5.2 Quadratic processing

As we saw in the previous section, ICA on natural images usually leads to emergence of features that resemble lines and bars. Some simple nonlinear image preprocessing such as the contrast transformations we examined may be insufficient to attain more interesting results with ICA. These more interesting results could include representations of more complex visual structures, and learning the corresponding features would give a hopeful outlook towards acquiring higher-level visual processing machinery with the used methodology: some higher-level processing may have the representation of more structured visual shapes as a prerequisite. For example, it has been shown that more structured features than simple lines and bars are useful in object recognition (Ullman et al., 2002) and also that as we move further on the cortex from V1 to areas such as V2 and V4, we start to encounter neurons that are tuned to more structured features than simple edges and

bars (Hegdé & Essen, 2000; Heider et al., 2000; Ito & Komatsu, 2004; Anzai et al., 2007). In particular, these medium-complexity neurons seem to prefer visual stimuli that resemble feature combinations.

In this section we study the question how similar medium-complexity features could be learned in an unsupervised manner. Of some interest is to examine whether the approaches of ICA and sparse coding could be extended in some way to account for the formation of more structured features and the corresponding filtering. As it is difficult to make linear filter models more specific for detection of feature combinations while remaining inactive for their constituents (Zetzsche & Krieger, 1999), nonlinear processing and modelling seems called for.

5.2.1 Quadratic processing by ICA

In Publication 4 and Publication 6 we examined whether extending the usual linear models to quadratic ones would allow to learn visual features tuned to more complex image properties than edges and bars, and found positive evidence for this. In **Publication 4**, we performed a simple quadratic basis expansion (Hastie et al., 2001) on the input data, after which linear ICA learning was applied. This led to the emergence of quadratic filters that responded to combinations of Gabor-like, features, but not to the features alone. Thus, on a qualitative level this behaviour corresponded to one observed in V2, where some neurons appear to have preference for feature conjunctions (Hegdé & Essen, 2000; Heider et al., 2000; Ito & Komatsu, 2004; Anzai et al., 2007).

It is interesting that our model learned preference for conjunctions, as this result differs from those previously obtained with ICA on similarly transformed data (Bartsch & Obermayer, 2003; Hashimoto, 2003; Theis & Nakamura, 2004) and also from those obtained with other two-layer models (Hoyer & Hyvärinen, 2001; Berkes & Wiskott, 2005; Köster & Hyvärinen, 2007). In all the mentioned studies, typically simple and complex cell behaviour of V1 were learned. However, reading Theis and Nakamura (2004) carefully shows that feature combinations were also learned with their quadratic model, but the authors did not focus on the conjunctive processing in the analysis of their results.

It should be noted that there are some open problems related to applying ICA on transformed data in the way we did. First, it is not well understood why one should look for independent components in a quadratic space, or why the quadratic space should be represented in particular. Although meaningful features were learned by the model, these questions still lack satisfactory theoretical answers. Second, in further, unpublished ex-

periments performed after Publication 4, we found out that the method we proposed is also able to learn conjunctive structure on noise data. In particular, consider an ICA model learned on Laplacian white noise data of n dimensions taken through the homogeneous quadratic basis expansion, with no dimensionality reduction. In this case, we found out that the learned models will have n directions weighting the n squared original dimensions while exhibiting no paired structure, whereas the remaining $n(n-1)/2$ components become essentially conjunctive. This emergence happens even though it is unlikely that the Laplacian noise data had some hidden conjunctive features. Instead, the learned structures can be understood through the learned model $\mathbf{W}$, which is in this case an identity matrix, up to the permutation of the columns and the signs of the columns. Decomposing the related quadratic forms for each projection direction in such a matrix results in a split to non-conjunctive and conjunctive components as we described. However, the learned $\mathbf{W}$ does not have such a form for arbitrary data: for Gaussian white noise, the components do not end up as such that could be characterized as exhibiting paired structure. And, as is to be expected, the underlying learned filters do not resemble Gabors with either Laplacian or Gaussian data.

Finally, a remark is in order regarding the enterprise of attempting to learn more complex visual structures from the data using the principle of independence maximization. Consider a corner, made of two edges. Now, if independence is maximized, detectors for the corner and detectors for the edges can not exist together, as the activation of the corner detector would predict activation of the edge detectors. Hence with maximization of independence, random variables reacting both to objects and their parts can not coexist if there is a constraint that responses must be perfectly independent. To conclude, the independence maximization approach would at least need to allow dependencies between features of different levels of complexity, perhaps corresponding to allowing dependencies between different layers of a multilayer model. If perfect independence is required between all elements of such systems, compound features are unlikely to be formed[1].

[1]Note that the result of Publication 4 does not contradict this theoretical claim: the model is forced to estimate a predefined number of features from the data, and it is forced to do the best it can, even if there are not as many independent directions in the data. Further, we did not optimize for the independence of the squared subfilters that make up each quadratic form.

5.2.2 Quadratic processing by energy optimization

We followed the study of Publication 4 by another one in **Publication 6**, where we attempted to replicate the results of Publication 4 in a more simple setting. To start with, instead of working in a quadratically transformed space, we assumed the response of each component to be a product of the responses of two linear filters, i.e. $s = (\mathbf{w}^T\mathbf{z})(\mathbf{v}^T\mathbf{z})$. We assumed that all the vectors $\mathbf{v}$ and $\mathbf{w}$ form an orthogonal basis, and that $\mathbf{z}$ is whitened data. As an objective function, instead of maximizing for independence, we optimized energy of s, i.e. $E_{\mathbf{x}}[|s|]$, as described in Section 4.1.2. When minimizing energy, paired filters were learned again, similar to those we attained in Publication 4. The main difference was that now there were no pairs of features that shared orientation. When maximizing the energy instead, the method learned opponent orientation filters, i.e. filters that respond highly positively to one orientation, and highly negatively to another. The presence of such opponent orientation processes in natural visual systems appears to remain an open issue, with proponents (Motoyoshi & Kingdom, 2003) and opponents (N. Graham & Wolfson, 2004).

Due to the better scalability of the setting of Publication 6, we were able to learn the filters in higher resolution, and show that the feature pairings are not arbitrary, but instead seem to reflect some statistics of the input data. Interestingly, also in this case the paired filters turned out to resemble Gabor filters, a phenomenon we discuss to some length in the paper. But as the used objective is somewhat different from that of ICA in the quadratic space, this does not allow us to make claims of the pairings learned with quadratic ICA to be non-arbitrary. Instead, the work of Publication 6 strengthens the hypothesis that meaningful feature combinations can be learned from natural image data with appropriate models and unsupervised objective functions. To what extent the learned processing resembles the conjunctive behaviour in e.g. V2 neurons remains to be seen.

Although we have shown that conjunctive, nonlinear processing can be learned from the visual data in an unsupervised manner, this does not logically force the conclusion that the objectives we use are the reason for the formation of similar features in biological systems. Instead, as research from object detection has indicated, features of medium complexity are more discriminative in object detection than simple Gabor features (Ullman et al., 2002). Likewise, many recent methods either learn combined features (LeCun et al., 2004) or use them after manual design (Serre, Wolf, et al., 2007) for state of the art object detection results. Hence higher-level objectives might call for the presence of conjunctive processing as found in

V2 neurophysiology.

At the time of writing this, we have not yet shown that the emergent paired features or filters have any practical use. For inherently 2D problems such as corner detection, it is known that tighter tuning can be attained with quadratic filtering than linear filtering (Krieger & Zetzsche, 1996; Zetzsche & Krieger, 1999). Clearly our conjunctive filters have very tight tunings for angles and corners, but nevertheless some kind of quantitative analysis should be performed as future work. One option is to validate these unsupervised features in object recognition to verify if better classification accuracies could be attained. Yet quadratic features could also be learned for object recognition to start with, possibly using established methods such as the backpropagation algorithm (Rumelhart et al., 1986). In a partially fixed setting where the first layer were defined to be Gabor filters, this question has already been studied (Weber & Casasent, 2001).

5.3 Simple priming mechanisms

Typically in feedforward network models (Riesenhuber & Poggio, 2000; LeCun et al., 2004; Serre, Wolf, et al., 2007), and in the models of the previous sections, all features are computed regardless of what the input or task is. For example, in the case of the quadratic features we learned in Section 5.2, both constituents in the pairs $(\mathbf{w}^T\mathbf{z})(\mathbf{v}^T\mathbf{z})$ are evaluated regardless of $\mathbf{z}$. If the purpose of the computation was to detect a corner made of two constituents, it could be possible to stop or postpone the evaluation of the second constituent in the case that the first one returned a value close to zero. Contrary to this kind of conditional processing, in the traditional feedforward view the neural processing evaluates more and more complicated features, finally culminating in a stage that evaluates a full set of object models, picking the best matching object model (or background) as the detected one. But does natural vision operate in this manner? Is it the best way to proceed, and is it an efficient way to recognize objects?

Some studies argue that natural visual systems might not work in this way, and suggest an alternative view where computations would be guided by quick preliminary analyses or prior expectations (Bar, 2003). For example, the later visual areas in natural visual processing can shape their selectivity based on interactions with the earlier areas (Jehee et al., 2007), the later areas thus being *primed* by information from the earlier processing. Also, it is known that in natural processing, the information does not exactly proceed in a first-come-first-served manner through a series of stages as in a waterfall, but instead there are overlaps due to latencies of

different areas (Bullier, 2004). In principle this would allow one area to be already primed before more detailed information reaches it through some slower pathway.

Several studies of object detection have examined ideas where the later computations are somehow dictated by the earlier ones, usually in an attempt to reduce the number of performed computations in different ways (Mirmehdi et al., 1999; Viola & Jones, 2001; Blanchard & Geman, 2005). One of the most famous examples of this approach for object detection is the cascade model by Viola and Jones (2003), reminiscent of decision trees, where the test instances $\mathbf{x}$ are routed in a tree-like model towards decisive nodes using simple attribute tests (Quinlan, 1993). In the cascade model, instead of learning a balanced tree of attribute tests, a sequence of very simple rejective classifiers are learned, and the input $\mathbf{x}$ is either rejected by the classifier ("there is no way that an object of interest is present in $\mathbf{x}$") or passed on to the next classifier for further analysis. By proper tuning of the rejective classifiers, high recognition rates can be achieved and yet most examples $\mathbf{x}$ can be rejected after only a few simple tests.

5.3.1 Low-frequency priming

The hypothesis of low-frequency priming (Bar, 2003) is a specific priming hypothesis that suggests that a low-frequency version of the scene could be processed first in natural vision, and the results of this computation could be used to prime the subsequent later processing. In natural systems, this may be alleviated by different streams of information that process at different speeds (Bullier, 2004), and in computational models of object recognition it is parallelled by coarse-to-fine processing (Amit et al., 2004), where the processing starts with coarse or low frequency representations and moves towards more detailed and specific processing.

In **Publication 2** we examined the applicability of low-frequency priming in a parts-based object recognition paradigm (Ullman et al., 2002; Schneiderman & Kanade, 2002; Agarwal et al., 2004; Leibe et al., 2008). In this approach, each object is represented as a set of object parts, or small image fragments. A simple realization of the parts-based setting could be one where the image representation for classification is a binary vector, where each dimension x_i corresponds to a fragment i. If the fragment is found in the image, the value of x_i is set to 1, otherwise to 0. A classifier is then learned from pairs of such feature vectors and their labellings $(\mathbf{x}, y)$ in a supervised manner, for example by an SVM method (see Section 4.1.3).

To represent the image in the parts-based model, in typical approaches all the parts known to the system are tested to see if they are present in the

current image. Although the tests can be restricted to be performed near the likely spatial locations of such features, the parts-based approach in this simple form is reminiscent of the bag-of-words approach known in text document classification (Manning & Schütze, 1999). For text documents, counting the number of word occurrences can be done reasonably fast, but with images, the number of possible object parts in terms of small fragments can be large, and it may be computationally demanding to test for presences of all the parts known by the system in all image locations. For this reason we examined in Publication 2 whether some quick computations would allow us to decide which parts to test.

We designed the method of the Publication 2 as follows. First, a linear classifier to do object vs. background classification was learned separately for each object class to be recognized. These initial classifiers were intended to work as a quick test to choose between three different options: 1) decide that the image contains no object known by the system, 2) decide that we remain uncertain, or, 3) decide a certain object has been recognized. In cases 1 and 3, the evaluation stops and a decision of the object class is made. In the case 2, the actual decision is delegated to be performed by one or more instances of object-specific parts-based classifiers. As a result of the pruning performed by these initial classifiers, often not all parts known by the system have to be evaluated. The specific decision rules between the three options and the subsequent routing of the instances are described in Publication 2. For more recent work and detail on this kind of delegation approach, see Autio (2008).

To be computationally as fast as possible, the initial linear classifiers were designed to work in the greyscale pixel space and trained by an SVM soft margin method. As a result, the weights $\mathbf{w}$ learned for each object class by the SVM are practically holistic, global templates, as shown in Figure 5.1B.

We examined whether the low-frequency priming would be applicable in the scope of these initial linear mechanisms, i.e. does the linear SVM classifier succeed in the initial pruning task, relying on low-frequency information. We evaluated this by processing the training images with varying degrees of Gaussian blur, as shown in Figure 5.1A, the blur increasing towards the right. What is interesting in these blurred cases is that the learned receptive fields look even less like the target object class, but more like noise, as shown in Figure 5.1B. But despite this, they classify the training instances correctly and succeed in working as initial pruning mechanisms when evaluated by cross-validation.

A few problems arise from our methodological choice of using linear clas-

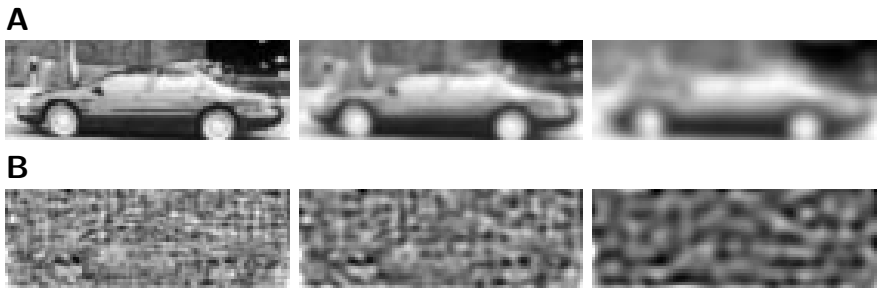

Figure 5.1: **A)** An example of a training image for car detection with varying amounts of Gaussian blur, with blur variance $\sigma^2 \in \{0.5, 1, 2\}$. **B)** "Receptive fields" learned by a linear support vector machine for detection of cars, when the training images had been blurred as in A. An interesting property of these receptive fields is that they all are sufficient to classify the whole training set correctly, yet they do not closely resemble car templates.

sifiers on the raw greyscale images. First, the cortical object-specific neurons are considered highly specific and nonlinear (Tanaka, 1996), whereas in a linear classifier the nonlinearity is a simple thresholding operation. This already makes it implausible that our linear classifiers would be operationally similar to the usually encountered neurons in the inferotemporal cortex. The receptive fields in Figure 5.1B may explain why the evolution is unlikely to have resorted to as a simple model as we used. Consider one of the receptive fields $\mathbf{w}$ that is essentially used by testing the inequality $\mathbf{w}^T\mathbf{x} > b$ for object presence, where b is a threshold. It is well-known that the highest activation that $\mathbf{w}^T\mathbf{x}$ can have with fixed-norm $\mathbf{w}$ and $\mathbf{x}$ is attained by the choice $\mathbf{x} = \mathbf{w}$. But this means that the model based on $\mathbf{w}$ predicts $\mathbf{w}$ to look as much the object as possible, and $-\mathbf{w}$ the opposite. It can clearly be seen from Figure 5.1B that neither case is true. Even if the receptive field $\mathbf{w}$ had been learned to look like a car, clearly $-\mathbf{w}$ would still look like one, only with its colours inverted. Despite this, the linear classifier would claim it to be background. This creates some serious problems that do not seem to be addressable with linear classifiers. First, the model of Publication 2 will accept examples as objects that look like the receptive fields of Figure 5.1B, but that may not pass as the target objects to human vision. On the positive side, this illustrates the limitations of the typical machine learning practice: the training data of objects and backgrounds that we used to train the model of Figure 5.1B does not include such "background" images that would have prevented the learner from estimating a

function that behaves badly by predicting an object to be present in such senseless images as the receptive fields we showed. And even if the dataset had contained such images, the fact that both $\mathbf{x}$ and $-\mathbf{x}$ would represent cars can not be consolidated within a linear classifier framework. The linear separation of the training dataset of Publication 2 testifies to the fact that without understanding the properties of the data and the problem at hand, such invalid models as estimated by machine learning methods may not be caught by blindly cross-validating on some dataset alone, but require both the understanding of the data, the problem, and the learned models, just as we suggested in Section 3.1.

Besides the initial linear model we used, the class of rejective cascades in general can be criticized for the fact that they can make unalterable decisions, and should the initial decisions be wrong, this can not be repaired by later processing, as there is none. It is possible to speculate that yet different approaches are used by natural systems that could resemble evidence-guided inference loops that stop only after the most demanding of analyses have been carried out (possibly in Bayesian models of object detection – consider for example the hypothesis testing in Fei-Fei et al. (2006)).

5.3.2 Gists of visual scenes

Although the low-frequency priming study of Publication 2 was motivated to investigate a hypothesis from biology (Bar, 2003), in principle we could try to base the priming on *any* computations with informative outputs, and not necessarily just on low-frequency data. In **Publication 1**, we examined if a so-called *spectral representation* (X. Liu & Cheng, 2003) could be used as a source of quickly computable information that roughly delivers the gist of the current scene. This gist could possibly be used to rapidly infer the current environmental context, for example if we are indoors or outdoors, or in a forest or on a plain, as statistics of such visual environments are arguably different (Torralba & Oliva, 2003).

The spectral representation we studied is made up of activity histograms of a V1-like bank of filters. To get a better intuition of the representation, imagine a Gabor-filter in some particular orientation and frequency, and then convolve this filter over the input image, and finally compute a histogram of the convolution response. Such marginal histograms from different filters concatenated into a vector form the spectral representation. In Publication 1 we studied the information content of such a nonlinear representation by ICA, with the underlying filters also learned by ICA and not specified as Gabor filters. We demonstrated tentative evidence towards

the ICA components of this nonlinear representation being able to capture something of the gists of the scenes. The extent that the gist was captured by the components was comparatively evaluated against a few other methods by naïve human subjects, and the results of the ICA-based method were found to be preferable to e.g. PCA-based results. However, we never succeeded in qualitatively describing what higher-level or holistic aspects of the scene the learned ICA components represented.

In hindsight, we think that the information content of the spectral representation as we used it may be relatively poor. The problem is that the linear ICA on natural images practically learns the same filters in different spatial locations. But when two slightly differently positioned filter masks are convolved over larger images, and ignoring negligible border effects, the marginal histograms are almost identical. Hence, unless the representation is augmented suitably, for example by multiscale analysis (as in X. Liu and Cheng (2003)), its capability to relate the global structure of the image may be weak. However, at least for the ICA analysis of the representation, we could not find clear benefits in resorting to the multiscale version of the spectral representation. It is also prudent to admit that in Publication 1, the images were not standardized or gain-controlled to bring them to some uniform scale, and hence we speculate that some of the learned features may be capturing some rather simple concepts from the images, such as the amount of global illumination or the amount of contrast oriented to some specific direction.

5.4 Online feature selection

The work of Publication 2 and that of Ullman et al. (2002); Schneiderman and Kanade (2002); Leibe et al. (2008), and others, demonstrate that to some degree of success, object recognition can be based on small image fragments (local templates) given that they are informative regarding the class of the object. In contrast to the traditional neural network models (e.g. LeCun et al. (2004)), in the usual parts-based approaches these local templates are not adjusted or tuned in any way. Instead, they are simply used as they were found from the data, and the main learning process concerns their selection. A design choice in Ullman et al. (2002) and in Publication 2 is that these features are selected with a batch mechanism: a large set of object and non-object images are expected to be available, and the selection is done using information theoretic criterion regarding the discriminability of the features. Although other feature selection methods could be easily applied for parts-based modelling (see Dash and Liu (1997);

Guyon and Elisseeff (2003) for reviews on feature selection), these selection methods largely work in the batch setting, requiring all training data at once.

For the research of **Publication 3**, we started from the simple observation that for humans, learning to recognize objects seems to be very different to the framework of supervised learning: in natural conditions, an object is perceived, someone perhaps tells us its category (but not necessarily), and possibly we are later able to recognize the same object and remember its name (if we were told it). We seem to be able to do this even without going through a whole training set of objects and non-objects. We hence asked if the features used for object recognition could be collected as they are encountered, image by image, in an incremental or *online* manner.

Publication 3 answers this question affirmatively, and we propose a few simple methods to perform this kind of online feature acquisition. An additional benefit of the proposed approach and the algorithms is that they are by no means limited to vision, but could in principle be used and extended for any selection problem. In particular, they could be used to select from any sets of functions or mechanisms as they are encountered in the environment. We suggest that this is a very natural approach. As an allegory, consider that we have a pile of textbooks on elementary mathematics, and we wish to solve the problem of long division efficiently. Do we *develop* an algorithm from scratch, or would it be more effective to scan the book pile, and *acquire* one? In Publication 3 the philosophy is that we might as well acquire the algorithm, and we evaluated the idea in the context of acquiring visual processing mechanisms from the environment. In terms of our example, note that not only we proposed collecting the patches ("the numbers to divide with"), we also proposed that the computation itself ("how to perform long division") could be acquired during the lifetime of the system; possibly from a teacher or another system.

With a bit of reflection, it seemed that a setting as general as the one we proposed in Publication 3 could not be novel. It soon turned out that although such online selection approaches have not lately been in the limelight of vision or machine learning research, similar settings have definitely been studied. In 1992, Poggio et al. presented a model for perceptual learning (Poggio et al., 1992), where new nodes were added to a neural network in the course of learning. Similar dynamically growing networks have also been studied in the scope Adaptive Resonance Theory (ART) networks (Carpenter & Grossberg, 2003). The selection of cases to the database in Case-Based Reasoning (CBR) approach to image interpretation (Perner, 2001) is semantically similar to our approach. More lately, the online fea-

ture selection problem has been studied as streamwise selection (Zhou et al., 2006). Further, this problem setting is even better known in economics as *portfolio selection* (Elton et al., 2002), and regards the maintenance of a set of good stocks (features). In statistics, similar problems are investigated under the subfield of sequential analysis (Lai, 2001). As adapting the methods from statistics did not seem straightforward in our setting (and we were not aware of e.g. Poggio et al. (1992); Perner (2001); Carpenter and Grossberg (2003) at the time of Publication 3), it remains to be examined how the algorithms we proposed in Publication 3 compare to selection algorithms proposed in these other research traditions. We leave this as future work.

Finally, the methods we proposed in Publication 3 are intimately connected to the assumption that the usefulness of the acquired features can be evaluated, in our case against a training signal. Although we demonstrated in Section 5.2 that more complex, inherently two-dimensional features can be learned from visual data with unsupervised techniques, it would be interesting to understand better the acquisition of such features in an unsupervised setting so that they would end up corresponding to meaningful object parts and not just angles and corners. We note that replicating such acquisition may require different datasets than the commonly used one from Hateren and Schaaf (1998), as that set can be argued not to contain frequently recurring object shapes that would retain their shape across instances (i.e. the typical objects in that set – such as trees and bushes – look very different in a low-level sense from one instance to another). There is some work towards the direction of learning medium-complexity features in unsupervised manner, see (Edelman et al., 2002; Ranzato et al., 2007), and in a slightly more supervised setting (Shams & von der Malsburg, 2002). Carrying on from such approaches may turn out worthwhile.

Chapter 6

Conclusion

In the scope of this thesis we worked in the setting of ecology-driven approach to modelling of vision. We presented new results by learning visual processing from natural images using various objectives, while examining the roles that some simple nonlinearities could play in the learning process. Often the resulting models exhibited intuitively meaningful visual processing, and the knowledge gained from these experiments may help us understand the nature of the visual data and the problems of vision to a greater extent.

It can be said that directly, our learned models did not turn out to solve any of the grand challenges of vision. A'posteriori, this is somewhat expected, as it is becoming increasingly clear that the ecology-driven approach is still in its infancy, and a significant amount of work remains to be done. In particular, there is a large gap between the natural reality and ecology as reduced to natural images and learning objectives. In the bulk of today's machine learning, the model and the learning algorithm are passive viewers of the data (but see also Sutton and Barto (1998)), as neither the learning algorithm nor the resulting processing mechanism can move around in the world and sample its own training examples. In contrast, animals can utilize their motor systems to look at their surroundings with much more freedom. The limitation to passive models is becoming increasingly untenable, as knowledge regarding the interconnectedness of different aspects of perception, embodiment and action in natural systems grows (e.g. A. J. Bell (1999); Thelen et al. (2001); Wilson (2002)). It is understandable that in the current learning settings with either too specific or too nonspecific objectives, the statistical methodology may lack all incentive to emerge such interesting processing properties as figure/ground segregation, depth perception, or understanding that objects may occlude one another. Even less will the statistical approach spontaneously develop

capability for passing such visual Turing tests as described in Chapter 1.

Our opinion is that it may be time to move on from the constraining settings of supervised and unsupervised learning and incorporate even more ecology into the ecology-driven approach. For us, this means acknowledging the fact that pathological exceptions aside, each real-life visual apparatus is a part of an embodied system that interacts with its environment and has objectives molded by evolutionary pressure. That carrots should be appreciated and whips avoided are not two absolute truths dictated in some holy book of objectives, but indirect consequences following from the necessities of carbon-based life. In other words, it is the entity and the environment that together define the salient and the nonsalient, and what should be transformed, represented, or ignored.

Accepting these facts does not necessarily lead to anything overly complicated or intractable: we could start by modelling very simple visuomotor learning systems such as that in the bee or the fruit fly, similarly as both evolution and Brooks (1999) have found the simple mechanisms first. Essentially in the limits of their capabilities, the embodied model systems could explore the world and subsequently learn about its structure. This would not be unlike how infants learn about the natural constraints and properties of the world by manipulating objects and interacting with their environment (Rochat, 1989). It should be emphasized that the proposed approach is not at all whimsical, as tentative embodied visual learning efforts are already underway in robotics (e.g. Krichmar and Edelman (2005); Kassahun et al. (2007)). Further, it should be pointed out that making the ecological approach more ecological does not require physical robotics, as more and more realistic virtual environments are being developed, some of them published as open source, and some likely to become standards. In such modern environments, experiments could be carried out in simulation with smaller risks of oversimplification than the *blocks world* experiments in the seventies had (as described e.g. in Marr (1982); Palmer (1999)).

Regardless of how the current modelling frameworks are extended, statistical models, learning and inference appear to remain central methodological tools, as the uncertain and partially observable nature of the external world will remain. However, the idea of systems that interact with the environment and each other seems to call for new, dynamic learning paradigms and algorithms. The idea also forces us to consider how aspects like motor control and location awareness are incorporated into the framework of vision and learning. Nevertheless, facing this complexity seems worthwhile, as embracing these more general settings of learning could eventually result in models of visual behaviour that can make distinctions based

on such old Gestalt concepts as *affordance*. Perhaps one day some model will recognize the odd-looking object for a chair because it *affords* to be sat upon. When some model eventually makes that realization, we can say that the models of recognition are no longer "mere template matchers" but ones that realize Helmholtz' classic proposal of perception as (un)conscious inference (Helmholtz, 1867).

References

Adams, D. L., & Horton, J. C. (2002). Shadows cast by retinal blood vessels mapped in primary visual cortex. *Science*, *298*(5593), 572–576.

Adrian, E. D., & Zotterman, Y. (1926). The impulses produced by sensory nerve endings: Part II: The response of a single end organ. *J Physiol (Lond.)*, *61*, 151–171.

Agarwal, S., Awan, A., & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(11), 1475–1490.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alexander, R. M. (1996). *Optima for animals, revised ed.* Princeton University Press.

Allard, R., & Faubert, J. (2007). Double dissociation between first- and second-order processing. *Vision Research*, *47*(9), 1129–1141.

Amit, Y., Geman, D., & Fan, X. (2004). A coarse-to-fine strategy for multi-class shape detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(12), 1606–1621.

Anokhin, A. P., Golosheykin, S., Sirevaag, E., Kristjansson, S., Rohrbaugh, J. W., & Heath, A. C. (2006). Rapid discrimination of visual scene content in the human brain. *Brain Research*, *1093*, 167–177.

Anzai, A., Peng, X., & Essen, D. C. van. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, *10*(10), 1313–1321.

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

Autio, I. (2008). *Modeling efficient classification as a process of confidence assessment and delegation* (Tech. Rep. No. A-2008-1). Department of Computer Science, University of Helsinki.

Badcock, D. R., Clifford, C. W. G., & Khuu, S. K. (2005). Interactions between luminance and contrast signals in global form detection. *Vision Research*, *45*, 881–889.

Baddeley, R. (1996). Searching for filters with 'interesting' output distributions: an uninteresting direction to explore? *Network: Computation in Neural Systems*, *7*(2), 409–421.

Bair, W. (2005). Visual receptive field organization. *Current Opinion in Neurobiology*, *15*, 459–464.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*, 600-609.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *PNAS*, *103*(8), 449–454.

Barlow, H. (1969). The coding of sensory messages. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behaviour* (pp. 331–360). Cambridge: Cambridge University Press.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*, 241–253.

Bartsch, H., & Obermayer, K. (2003). Second-order statistics of natural images. *Neurocomputing*, *52-54*, 467–472.

Becker, S. (1992). Learning to categorize objects using temporal coherence. In *Advances in neural information processing systems 5* (pp. 361 – 368). Morgan Kauffman.

Becker, S., & Zemel, R. (2003). Unsupervised learning with global objective functions. In M. Arbib (Ed.), *The handbook of brain theory and neural networks, second edition.* Cambridge, MA: The MIT Press.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129-1159.

Bell, A. J. (1999). Levels and loops: the future of artificial intelligence and neuroscience. *Phil. Trans. R. Soc. Lond. B*, *354*, 2013–2020.

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

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, *5*, 579–602.

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *Journal of the Optical Society of America A*, *19*(6), 1096–1106.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–117.

Billock, V. A. (2000). Neural acclimation to 1/f spatial frequency spectra in natural images transduced by the human visual system. *Physica*

D: Nonlinear Phenomena, *137*(3-4), 379-391.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blanchard, G., & Geman, D. (2005). Hierarchical testing designs for pattern recognition. *Annals of Statistics*, *33*(3), 1155–1202.

Bovik, A. (2000). Basic gray-level image processing. In A. Bovik (Ed.), *Handbook of image and video processing* (pp. 21–36). London: Academic Press.

Brady, N., & Field, D. J. (2000). Local contrast in natural images: normalization and coding efficiency. *Perception*, *29*(9), 1041–1055.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new AI*. The MIT Press.

Bullier, J. (2004). Communications between cortical areas of the visual system. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences, volume 1* (pp. 522–540). The MIT Press.

Bullock, T. H., Bennett, M. V. L., Johnston, D., Josephson, R., Marder, E., & Fields, R. D. (2005). The neuron doctrine, redux. *Science*, *310*, 791–793.

Carandini, M. (2006). What simple and complex cells compute. *J Physiol*, *577*(2), 463–466.

Carpenter, G. A., & Grossberg, S. (2003). Adaptive resonance theory. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks, second edition* (pp. 87–90). Cambridge, MA: MIT Press.

Casagrande, V. A., & Xu, X. (2004). Parallel visual pathways: A comparative perspective. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences, volume 1* (pp. 494–505). The MIT Press.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, *4*, 185-211.

Chance, F. S., Nelson, S. B., & Abbott, L. F. (1999). Complex cells as cortically amplified simple cells. *Nature Neuroscience*, *2*, 277–282.

Churchland, P. S., & Sejnowski, T. (1992). *The computational brain*. MIT Press.

Cleland, B. G., & Freeman, A. W. (1988). Visual adaptation is highly localized in the cat's retina. *Journal of Physiology*, *404*, 591-511.

Comon, P. (1994). Independent component analysis – a new concept? *Signal Processing*, *36*, 287–314.

Copeland, J. (2000). Narrow versus wide mechanism: Including a reexamination of Turing's views on the mind-machine issue. *The Journal of Philosophy*, *97*(1), 5–32.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory, 2nd edition*. Wiley.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press.

Dan, Y., & Poo, M. (2006). Spike timing-dependent plasticity: From synapse to perception. *Physiol Rev*, *86*, 1033–1048.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(3), 131–156.

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: Society for Industrial and Applied Math.

Daugman, J. G. (1985). Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. of the Optical Society of America A*, *2*, 1160–1179.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, Massachusetts: The MIT Press.

De Valois, R. L., & De Valois, K. K. (1980). Spatial vision. *Annual Review of Psychology*, *31*, 309–341.

deCharms, R. C., & Zador, A. (2000). Neural representation and cortical code. *Annu. Rev. Neurosci.*, *23*, 613–647.

Deco, G., & Lee, T. S. (2004). The role of early visual cortex in visual integration: a neural model of recurrent interaction. *European Journal of Neuroscience*, *20*, 1089–1100.

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object detection. *Vision Research*, *44*, 621–642.

DeYoe, E. A., & Essen, D. C. van. (1988). Concurrent processing streams in monkey visual cortex. *Trends in Neuroscience*, *11*, 219–226.

Doi, E., Balkan, D. C., & Lewicki, M. S. (2007). Robust coding over noisy overcomplete channels. *IEEE Transactions on Image Processing*, *16*, 442–452.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification, second edition*. Wiley.

Duong, T., & Freeman, R. D. (2007). Spatial frequency-specific contrast adaptation originates in the primary visual cortex. *Journal of Neurophysiology*, *98*, 187–195.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*, 920–926.

Edelman, S., Intrator, N., & Jacobson, J. S. (2002). Unsupervised learning of visual structure. In H. H. Bülthoff, T. Poggio, S. W. Lee, &

C. Wallraven (Eds.), *Lecture notes in computer science* (Vol. 2025, pp. 629–643). Springer.

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, *2*(4), 324-353.

Eliasmith, C. (2007). Computational neuroscience. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science. handbook of philosophy of science (vol. 4).* Elsevier.

Elton, E., Gruber, M., Brown, S., & Goetzmann, W. (2002). *Modern portfolio theory and investment analysis, 6th edition.* Wiley.

Essen, D. C. van. (2004). Organization of visual areas in macaque and human cerebral cortex. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences, volume 2* (pp. 507–521). The MIT Press.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2008). The spatiotemporal profile of cortical processing leading up to visual perception. *Journal of Vision*, *8*(1), 1-12.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *28*(4), 594 – 611.

Felleman, D. J., & Essen, D. C. van. (1991). Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12).

Felsen, G., Touryan, J., Han, F., & Dan, Y. (2005). Cortical sensitivity to visual features in natural scenes. *PLoS Biology*, *3*(10), e342.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Finn, I. M., Priebe, N. J., & Ferster, D. (2007). The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, *54*, 137–152.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition and Emotion*, 61–92.

Frazor, R. A., & Geisler, W. S. (2006). Local luminance and contrast in natural images. *Vision Research*, *46*(10), 1585–1598.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, *82*(397), 249–266.

Gelder, T. van. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston:

Houghton Mifflin.

Gilbert, C. D., & Sigman, M. (2007). Brain states: Top-down influences in sensory processing. *Neuron*, *54*, 677–696.

Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. The MIT Press.

Graham, D. J., Chandler, D. M., & Field, D. J. (2006). Can the theory of 'whitening' explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Research*, *46*, 2901–2913.

Graham, N., & Wolfson, S. S. (2004). Is there opponent-orientation coding in the second-order channels of pattern vision? *Vision Research*, *44*(27), 3145–3175.

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, *24*, 31–47.

Grossberg, S. (2003). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*(1), 47–76.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hampel, F. R., Ronchetti, E. M., Rousseuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley.

Hancock, P. J., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network*, *3*, 61–70.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems*, *14*(4), 765–788.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.

Hateren, J. H. van, & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, *265*, 2315-2320.

Hateren, J. H. van, & Schaaf, A. van der. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, *265*, 359-366.

Hebb, D. O. (1964). *Organization of behavior*. New York: Wiley.

Heeger, D. J. (1992a). Half-squaring in responses of cat striate cells. *Visual Neuroscience*, *9*, 181–198.

Heeger, D. J. (1992b). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197.

Hegdé, J., & Essen, D. C. van. (2000). Selectivity for complex shapes in primate visual area V2. *The Journal of Neuroscience*, *20*(5), RC61-

66.

Heider, B., Meskenaite, V., & Peterhans, E. (2000). Anatomy and physiology of a neural mechanism defining depth order and contrast polarity at illusionary contours. *European Journal of Neuroscience*, *12*(11).

Helmholtz, H. von. (1867). *Treatise on physiological optics (from 3rd german edition, trans.) (3rd ed., vol iii)*. New York: Dover Publications.

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

Hood, D. C. (1998). Lower-level visual processing and models of light adaptation. *Annual Reviews of Psychology*, *49*, 503–535.

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

Hoyer, P. O., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, *11*(3), 191–210.

Hoyer, P. O., & Hyvärinen, A. (2001). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, *42*(12), 1593–1605.

Hubel, D., & Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. (Lond.)*, *148*, 574-579.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, *394*, 784–787.

Hurri, J., & Hyvärinen, A. (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, *15*(3), 663–691.

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

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

Hyvärinen, A., Hoyer, P., & Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, *13*(7), 1527-1558.

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

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* Wiley.

Inki, M. (2004). *Extensions of independent component analysis for natural image data* (Tech. Rep. No. D9). Helsinki University of Technology.

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neu.sci.*, *24*(13), 3313–3324.

Jain, A., & Farrokhnia, F. (1991). Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, *24*(12), 1167–1186.

Jain, A., Ratha, N., & Lakshmanan, S. (1997). Object detection using Gabor filters. *Pattern Recognition*, *30*, 295-309.

James, W. (1899). *The principles of psychology: Vol 1.* New York, NY: Henry Holt and Company.

Jehee, J. F., Roelfsema, P. R., Deco, G., Murre, J. M., & Lamme, V. A. (2007). Interactions between higher and lower visual areas improve shape selectivity of higher level neurons – Explaining crowding phenomena. *Brain Research*, *1157*, 167–176.

Jermakowicz, W. J., & Casagrande, V. A. (2007). Neural networks a century after Cajal. *Brain Research Reviews*, *55*, 264–284.

Johnson, A. P., & Baker, C. L. (2004). First- and second-order information in natural images: a filter-based approach to image statistics. *Journal of the Optical Society of America A*, *21*(6), 913–925.

Jutten, C., & Herault, J. (1991). Blind separation of signals, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, *27*, 1–10.

Kaas, J. H., Krubitzer, L. H., Chino, Y. M., Langston, A. L., Polley, E. H., & Blair, N. (1990). Reorganization of retinotopic cortical maps in adult mammals after lesions of the retina. *Science*, *248*(4952), 229–231.

Kassahun, M., Edginton, M., Gea, J. de, & Kirchner, F. (2007). Exploiting sensorimotor coordination for learning to recognize objects. In *International joint conference on artificial intelligence (ijcai-07)* (pp. 883–888).

Kingdom, F. A. A., Field, D. J., & Olmos, A. (2007). Does spatial invariance result from insensitivity to change? *Journal of Vision*, *7*(14), 1–13.

Koffka, K. (1935). *Principles of gestalt psychology.* New York: Harcourt, Brace.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th interna-*

tional joint conference on artificial intelligence (p. 1137-1145).

Kohn, A. (2007). Visual adaptation: Physiology, mechanisms, and functional benefits. *J. Neurophysiol*, *97*, 3155–3164.

Krahe, R., & Gabbiani, F. (2004). Burst firing in sensory systems. *Nature Reviews Neuroscience*, *5*, 13–23.

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences*, *12*(3), 114–122.

Kreinovich, V., & Kearfott, R. B. (2005). Beyond convex? Global optimization is feasible only for convex objective functions: A theorem. *Journal of Global Optimization*, *33*, 617–624.

Krichmar, J. L., & Edelman, G. M. (2005). Brain-based devices for the study of nervous systems and the development of intelligent machines. *Artificial Life*, *11*, 63–77.

Krieger, G., & Zetzsche, C. (1996). Nonlinear image operators for the evaluation of local intrisic dimensionality. *IEEE Transactions on Image Processing*, *5*(6), 1026–1042.

Kumbhani, R. D., Nolt, M. J., & Palmer, L. A. (2007). Precision, reliability, and information-theoretic analysis of visual thalamocortical neurons. *J. Neurophysiol*, *98*, 2647–2663.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Köster, U., & Hyvärinen, A. (2007). A two-layer ICA-like model estimated by score matching. In *Proc. int. conf. on artificial neural networks (ICANN2007)* (pp. 798–807).

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, *11*, 303–408.

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61*, 1-11.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. Teil C*, *36*, 910-912.

LeCun, Y., Bottou, L., Bengyo, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Huang, F. J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of cvpr'04.*

Lee, A. B., Mumford, D., & Huang, J. (2001). Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, *41*(1/2), 35-59.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America*, *20*(7), 1434-1448.

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *PNAS*, *98*(4), 1907–1911.

Leibe, B., Leonardis, A., & Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, *77*(1–3), 259–289.

Levy, W. B., & Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation*, *8*, 531–543.

Li, W., Piech, V., & Gilbert, C. D. (2008). Learning to link visual contours. *Neuron*, *57*, 442–451.

Li, Z. (1998). A neural model of contour integration in the primary visual cortex. *Neural Computation*, *10*, 903–940.

Litt, A., Eliasmith, C., Kroon, F. W., Weinstein, S., & Thagard, P. (2006). Is the brain a quantum computer? *Cognitive Science*, *30*(3), 593–603.

Liu, J. S. (2001). *Monte carlo strategies in scientific computing*. Springer.

Liu, X., & Cheng, L. (2003). Independent spectral representations of images for recognition. *Journal of the Optical Society of America*, *20*(7), 1271-1282.

Livingstone, M. S., & Hubel, D. H. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, *240*, 740–749.

Locke, L. L. (1933). The contributions of Leibniz to the art of mechanical calculation. *Scripta Mathematica*, *1*, 315–321.

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

Maass, W., Legenstein, R. A., & Markram, H. (2002). A new approach towards vision suggested by biologically realistic neural microcircuit models. In *Bmcv '02: Proceedings of the second international workshop on biologically motivated computer vision* (pp. 282–293). London, UK: Springer-Verlag.

Mach, E. (1882). The economical nature of physical inquiry. In *Reprinted in E. Mach, popular scientific lectures (5th ed.) (T. J. McCormack, trans.)* (pp. 186–213). La Salle, IL: Open Court.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., & Carandini, M. (2005). Independence of luminance and contrast in natural scenes and in the

early visual system. *Nature Neuroscience*, *8*(12), 1690–1697.

Marr, D. (1982). *Vision*. Freeman.

Masland, R. H. (2001). The fundamental plan of the retina. *Nature Neuroscience*, *4*(9), 877–886.

Matthies, L., Maimone, M., Johnson, A., Cheng, Y., Willson, R., Villalpando, C., et al. (2007). Computer vision on Mars. *International Journal of Computer Vision*, *75*(1), 67–92.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Mechler, F., & Ringach, D. L. (2002). On the classification of simple and complex cells. *Vision Research*, *42*, 1017–1033.

Meister, M., & Berry, M. J. (1999). The neural code of the retina. *Neuron*, *22*, 435–450.

Miller, B. T., & D'Esposito, M. (2005). Searching for "the top" in top-down control. *Neuron*, *48*, 535–538.

Minsky, M. (1986). *The society of mind*. New York: Simon and Schuster.

Mirmehdi, M., Palmer, P. L., Kittler, J., & Dabis, H. (1999). Feedback control strategies for object detection. *IEEE Transactions on Image Processing*, *8*(8), 1084–1101.

Motoyoshi, I., & Kingdom, F. A. (2003). Orientation opponency in human vision revealed by energy-frequency analysis. *Vision Research*, *43*(9), 2197-2205.

Moulden, B., Kingdom, F., & Gatley, L. F. (1990). The standard deviation of luminance as a metric for contrast in random-dot images. *Perception*, *19*, 79-101.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978a). Nonlinear spatial summation in the receptive fields of complex cells in the cat striate cortex. *J Physiol*, *283*, 78–100.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978b). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J Physiol*, *283*, 53–77.

Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from luminosity units in the retina of fish (cyprinidae). *Journal of Physiology*, *185*, 587–599.

Nilsson, D. E. (1989). Optics and evolution of the compound eye. In D. G. Stavenga & R. C. Hardie (Eds.), *Facets of vision* (pp. 30–73). Springer.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*,

381, 607-609.

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

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, *17*, 1665–1699.

Palmer, S. E. (1999). *Vision science.* The MIT Press.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes.* McGraw-Hill.

Peli, E. (1990). Contrast in complex images. *Journal of the Optical Society of America A*, *7*(10), 2032–2040.

Penrose, R. (1994). *Shadows of the mind.* Oxford University Press.

Penrose, R. (1997). Physics of the mind. In *The large, the small and the human mind* (pp. 93–143). Cambridge University Press.

Perner, P. (2001). Why case-based reasoning is attractive for image interpretation. In D. Aha & I. Watson (Eds.), *Case-based reasoning research and developments* (pp. 27–44).

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*, 148–158.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), 151–156.

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, *256*, 1018–1021.

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

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

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*, 341–423.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* Morgan Kaufmann.

Ramachandran, V. S., & Gregory, R. L. (1991). Perceptual filling in of artificially induced scotomas in human vision. *Nature*, *350*, 699–702.

Ranzato, M., Huang, F., Boureau, Y., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. computer vision and pattern recognition confer-*

ence, cvpr'07) (pp. 1–8). IEEE Press.

Rao, R. J., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, *2*(1).

Riesenhuber, M., & Poggio, T. (2000). *Computational models of object recognition in cortex: A review* (Tech. Rep.). Massachusetts Institute of Technology AI lab / Center of Biological And Computational Learning, Department of Brain and Cognitive Sciences.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*, 455–463.

Roberts, L. G. (1965). Machine perception of three-dimensional solids. In J. T. Tippett, D. A. Berkowitz, & L. C. C. et al. (Eds.), *Optical and electro-optical information processing.* MIT Press.

Robinson, S. L., & Miller, R. K. (1989). *Automated inspection and quality assurance.* Marcel Dekker, Inc.

Rochat, P. (1989). Object manipulation and exploration in 2- to 5-month-old infants. *Developmental Psychology*, *25*(6), 871–884.

Roelfsema, P. R., Tolboom, M., & Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron*, *56*, 785–792.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*(4), 517–548.

Ruderman, D. L. (1997). Origins of scaling in natural images. *Vision Research*, *37*, 3385–3398.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. vol. i: Foundations* (pp. 318–362). Bradford Books/MIT Press, Cambridge, MA.

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650.

Saalmann, Y. B., Pigarev, I. N., & Vidyasagar, T. R. (2007). Neural mechanisms of visual attention: How top-down feedback highlights relevant locations. *Science*, *316*(5831), 1612–1615.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S., et al. (1991). Seeing speech: visual information from lip move-

ments modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*(1), 141–145.

Schaffer, C. (1993). A conservation law for generalization performance. In *Proceedings of the 11th international conference on machine learning* (pp. 259–265).

Schneiderman, H., & Kanade, T. (2002). Object detection using the statistics of parts. *International Journal of Computer Vision*, *56*(3), 151–177.

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

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *PNAS*, *104*(15), 6424–6429.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426.

Shams, L., & von der Malsburg, C. (2002). Acquisition of visual shape primitives. *Vision Research*, *42*, 2105–2122.

Sharon, E., Galun, M., Sharon, D., Basri, R., & Brandt, A. (2006). Hierarchy and adaptation in segmenting visual scenes. *Nature*, *442*(17), 810–813.

Sherman, S. M., & Guillery, R. W. (2002). The role of the thalamus in the flow of information to the cortex. *Phil. Trans. R. Soc. Lond. B*, *357*, 1695–1708.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and Gestalt rules. *PNAS*, *98*(4), 1935–1940.

Simoncelli, E. P. (2005). Statistical modeling of photographic images. In A. Bovik (Ed.), *Handbook of image and video processing, 2nd edition* (pp. 431–441). Academic Press.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural presentation. *Annu. Rev. Neurosci.*, *24*, 1193–1216.

Sonka, M., Hlavac, V., & Boyle, R. (2007). *Image processing, analysis, and machine vision, 3rd edition.* Thomson-Engineering.

Sukumar, S., & Waugh, S. (2007). Separate first- and second-order processing is supported by spatial summation estimates at the fovea and eccentrically. *Vision Research*, *47*, 581–596.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An intro-*

duction. MIT Press.

Tadmor, Y., & Tolhurst, D. J. (2000). Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. *Vision Research*, *40*(22), 3145–3157.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, *19*, 109–139.

Theis, F., & Nakamura, W. (2004). Quadratic independent component analysis. *IEICE Trans. Fundamentals*, *E87-A*(9), 2355-2363.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant persevertive reaching. *Behavioral and Brain Sciences*(24), 1–86.

Thomson, M. G. A. (1999). Higher-order structure in natural scenes. *Journal of the Optical Society of America A*, *16*(7), 1549–1553.

Tong, F. (2003). Primary visual cortex and visual awareness. *Nature Reviews Neuroscience*, *4*, 219–229.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391-412.

Ts'o, D. Y., & Roe, A. W. (1995). Functional compartments in visual cortex: segregation and interactions. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (p. 325-337). Cambridge, MA: M.I.T. Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind*(236), 433-460.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682-687.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

Vicente, M. A., Hoyer, P. O., & Hyvärinen, A. (2007). Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(5), 896–900.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. of the 2001 IEEE computer society conference on computer vision and pattern recognition* (pp. 511–518).

Viola, P., & Jones, M. (2003). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Wang, X., Wei, Y., Vaingankar, V., Wang, Q., Koepsell, K., Sommer, F. T., et al. (2007). Feedforward excitation and inhibition evoke dual modes of firing in the cat's visual thalamus during naturalistic viewing. *Neuron*, *55*, 465–478.

Weber, D. M., & Casasent, D. P. (2001). Quadratic Gabor filters for object detection. *IEEE Transactions on Image Processing*, *10*(2), 218–230.

Williams, R. W., & Moody, S. A. (2004). Developmental and genetic control of cell number in the retina. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences, volume 1* (pp. 63–76). The MIT Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, *9*(4), 625–636.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Zeki, S. M. (1978). Functional specialization in the visual cortex of the rhesus monkey. *Nature*(274), 423–428.

Zetzsche, C., & Krieger, G. (1999). Nonlinear neurons and higher-order statistics: new approaches to human vision and electronic image processing. In B. Rogowitz & T. Pappas (Eds.), *Human vision and electronic imaging iv* (pp. 2–23).

Zhou, J., Foster, D. P., Stine, R. A., & Ungar, L. H. (2006). Streamwise feature selection. *Journal of Machine Learning Research*, *7*, 1861–1885.

Publication 1

J. T. Lindgren and A. Hyvärinen:

Learning High-level Independent Components of Images through a Spectral Representation

Proc. 17th International Conference on Pattern Recognition (ICPR), volume 2, pp. 72-75, 2004.

Publication 2

I. Autio and J. T. Lindgren:

Attention-driven Parts-based Object Detection

Proc. 16th European Conference on Artificial Intelligence (ECAI), pp. 917-921, 2004.

Publication 3

I. Autio and J. T. Lindgren:

Online learning of discriminative patterns from unlimited sequences of candidates

Proc. 18th International Conference on Pattern Recognition (ICPR), volume 2, pp. 437-440, 2006.

Publication 4

J. T. Lindgren and A. Hyvärinen:

Emergence of conjunctive visual features by quadratic independent component analysis

In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19: Proc. of the 2006 conference (NIPS)*, pp. 897–904, 2007.

Publication 5

J. T. Lindgren, J. Hurri and A. Hyvärinen:

The statistical properties of local log-contrast in natural images

Proc. 15th Scandinavian Conference on Image Analysis (SCIA), pp. 354-363, 2007.

Publication 6

J. T. Lindgren and A. Hyvärinen:

On the learning of nonlinear visual features from natural images by optimizing response energies

Proc. International Joint Conference on Neural Networks (IJCNN), pp. 1027–1034, 2008.

Publication 7

J. T. Lindgren, J. Hurri and A. Hyvärinen:

Spatial dependencies between local luminance and contrast in natural images

Journal of Vision, 8(12):6, 1-13, 2008.

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

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

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

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

A-2000-2 B. Heikkinen: Generalization of document structures and document assembly. 179 pp. (Ph.D. Thesis).

A-2000-3 P. Kähkipuro: Performance modeling framework for CORBA based distributed systems. 151+15 pp. (Ph.D. Thesis).

A-2000-4 K. Lemström: String matching techniques for music retrieval. 56+56 pp. (Ph.D.Thesis).

A-2000-5 T. Karvi: Partially defined Lotos specifications and their refinement relations. 157 pp. (Ph.D.Thesis).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A-2008-5 J.T. Lindgren: Learning Nonlinear Visual Processing from Natural Images. 100+52 pp. (Ph.D. Thesis).