Doctoral Thesis

Bottom-up and top-down processing of natural scenes

Author(s):
Einhäuser Treyer, Wolfgang

Publication Date:
2004

Permanent Link:
https://doi.org/10.3929/ethz-a-004751928

Rights / License:
In Copyright - Non-Commercial Use Permitted

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.
Bottom-up and Top-down Processing
of Natural Scenes

A dissertation submitted to the
Swiss Federal Institute of Technology Zurich
for the degree of
Doctor of Natural Sciences

presented by
Wolfgang Einhäuser Treyer
Dipl.-Phys. ETH (ETH Zurich)
born July, 6th, 1978
citizen of Germany

accepted on the recommendation of
Prof. Dr. Rodney J. Douglas, examiner
Prof. Dr. Peter König, co-examiner
Prof. Dr. Wulfram Gerstner, co-examiner

2004
Curriculum Vitae

Personal
Name: Wolfgang Einhäuser Treyer
Day of birth: July 6th, 1978
Place of birth: Mannheim (Germany)
Citizenship: German
Marital status: married, no children

Education
2001 - present: PhD student at Institute of Neuroinformatics of ETH Zürich (Switzerland)
1997 - 2001: studying physics for a total of 8 terms (4 years)
   2000 - 2001: at ETH Zürich, (Switzerland)
      October 2001: Graduation as "Dipl-Phys ETH"
1997 - 2000: at Universität Heidelberg (Germany)
   April 1999: "Vordiplom"
1996 - 1997: National Service
1988 - 1996: Peter-Petersen-Gymnasium in Mannheim (Germany)
1984 - 1988: Elementary School in Mannheim (Germany)
Languages
   English, Latin, (ancient) Hebrew
Professional qualifications
   Qualification to carry out animal experiments. (FELASA Category B)

Industry research internships
July/August 1996 & March/April 1998
   Daimler-Benz/Daimler-Chrysler AG ["Direktion: Forschung und Technologie 3 - Assis-
tenzsysteme" - Project: "Technisches Sehen Basis" (Fundamentals of Machine Vision)].

Teaching experience
Winterterms 2001/02, 2002/03 & 2003/04: TA for the class "Biophysics of Neural
Computation - Introduction into Neuroinformatics" of R.J. Douglas & K.A.C.Martin

Grants & Fellowships
PhD (2001-present): Honda RI Europe
Undergraduate support (1997-2001): Studienstiftung des deutschen Volkes (German
National Merit Foundation)
Publications

2004

2003

2002

2001

2000

Submitted manuscripts
W. Einhäuser, K.A.C. Martin, and P. König Are switches in perception of the Necker cube related to eye-position? submitted
W. Einhäuser and P. König Coincidences in feature space: the notion of an object is subject to plasticity. submitted


Disclaimers

All experiments involving human and/or animal subjects conformed with the respective National and Institutional Guidelines. Experiments involving human subjects conformed with the Declaration of Helsinki. Written informed consent was obtained from all human subjects.

Parts of this thesis (text and figures) were published or submitted for publication in various scientific journals as listed below. To achieve a consistent style throughout this thesis both text and figures might slightly differ from submitted/published versions. Copyright of text and figures has been or will be transferred to the respective publishers. The author hereby guarantees that he significantly contributed to all material used in this thesis. This does not apply to figures 1.1, 2.6 and 2.9 as well as to part of the data in figures 3.6, 4.1 and 4.12, whose origin and copyright is referenced in the respective captions. Text and figures of publications of the author prior to the start of his PhD (Weber et al., 2000; Kayser et al., 2001; Einhäuser et al., 2002a; Einhäuser, 2001) are not used in this thesis, but cited as references where appropriate.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sections/Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Betsch et al., 2004)</td>
<td>1.1; 2.1.1; 3</td>
</tr>
<tr>
<td>(Einhäuser et al., 2002b)</td>
<td>4.2</td>
</tr>
<tr>
<td>(Einhäuser &amp; König, 2003)</td>
<td>1.5; 2.1.2; 2.2.1; 6</td>
</tr>
<tr>
<td>(Einhäuser et al., 2003)</td>
<td>1.2; 2.6.2; 4.2</td>
</tr>
<tr>
<td>(Einhäuser, Kruse et al., in preparation)</td>
<td>2.1.2; 2.2.2; 2.2.3; 6</td>
</tr>
<tr>
<td>(Einhäuser, Martín et al., submitted)</td>
<td>1.6; 7</td>
</tr>
<tr>
<td>(Einhäuser &amp; König, submitted)</td>
<td>1.7; 8</td>
</tr>
<tr>
<td>(Einhäuser, Hipp et al., submitted)</td>
<td>2.1.3; 4.4</td>
</tr>
<tr>
<td>(Franzius et al., submitted; Franzius, 2004)</td>
<td>4.3</td>
</tr>
<tr>
<td>(Frey et al., in preparation)</td>
<td>2.4; 6.2.4; 6.3.3</td>
</tr>
<tr>
<td>(Kayser, Einhäuser et al., 2003)</td>
<td>1.1; 3</td>
</tr>
<tr>
<td>(Körding et al., 2004)</td>
<td>4.1</td>
</tr>
<tr>
<td>(Moeller et al., in preparation)</td>
<td>1.4; 2.2.4; 5</td>
</tr>
</tbody>
</table>
Zusammenfassung


Summary

Real world behaviour requires the understanding of complex stimuli. What are the properties of such natural stimuli as compared to artificial ones? Can we derive properties of the brain from statistical properties of the natural environment by applying simple rules? Which stimulus features control visual attention? What on the other hand is the role of stimulus independent signals? Can attention as such generate the percept of an object entity from multiple components? The present thesis addresses these issues.

To obtain natural stimuli, we recorded videos in part by a camera mounted to the head of a freely behaving cat. The spatio-temporal statistics of the thus acquired videos well approximates the natural statistic of the visual system’s input signal. Using these stimuli we first unsupervised train simulated neurons. We show, that by applying a simple rule, namely a variant of the “temporal coherence” principle, to natural stimuli leads to the emergence of neurons, whose response properties correspond to those of complex cells in primary visual cortex. These cells are characterised by their specificity to orientation paired with invariance to position. Likewise other complementary stimulus properties can be segregated based on their different temporal coherence, as we exemplify using colour and orientation. A further application of the same principle leads to texture-specific cells, whose classification behavior is well compatible with data measured in human observers. Finally, the application of the same principle to a standard computer vision database of natural objects generates object representations that are significantly less sensitive to viewpoint than the original representation. In summary we show that a single principle may explain several properties of the visual system - from position invariance in primary visual cortex up to invariant object representations.

Sequential shifts of visual attention represent a further central factor in understanding complex natural scenes. The so-called “overt” attention, i.e. attention associated with the shift of gaze, can be assessed directly by measuring the eye-position of the observer. Using this technique we asses the stimulus-driven (“bottom-up”) contribution to attention. We first confirm earlier studies, that showed a correlation between the guidance of voluntary eye-movements and luminance-contrast in natural scenes. Furthermore we show that the same correlation also holds for non-human primates. By selectively modifying natural stimuli we can on the other hand show, that this relation - for natural scenes - is correlative but not causal. Our data suggest that object content of natural scenes overwrites a residual causal dependence on contrast. The overwritten causal effect, however, is stronger in monkeys than in humans. We test the contribution of other features to fixation behaviour. For the example feature of colour we find, that interaction with other features only occurs if colour appears naturally. All these results strongly point to a decisive role of higher cognitive (“top-down”) processes in guiding attention. In order to investigate this role independently of changes in the stimulus, we use an ambiguous figure, i.e. a stimulus, whose percept changes although the stimulus itself remains unchanged. We investigate the relation of eye-movements and
the percept of the so-called “Necker-cube”. We find a mutual coupling between overt attention and the percept of the cube: While the change in eye-position follows a change in perception, the new eye-position in turn inhibits this new percept. In summary there is a strong mutual interaction between attention and perception. In order to investigate whether attention by itself suffices to bind various features, which have only simultaneous attention in common, into a unified object entity, we use chromatic gratings. A particular combination of colour and orientation is defined as target “object”. We find, that humans - in the case of such artificial stimuli - do not automatically bind multiple features into an object percept. This implies that in the case of natural stimuli binding of features into objects results from particular statistical properties of natural stimuli. Summarizing we show in the present thesis, that the perception of natural stimuli can to some extent be explained by their statistical properties. Although attentional processes are primarily guided by higher cognitive interpretation and in turn act back on it, this interpretation can also be understood as a consequence of learnt statistical relations. A deeper understanding of active perception thus inevitably requires the understanding of the statistical probities of natural stimuli.
# Contents

1 Introduction          1  
1.1 Statistical regularities of natural stimuli ................................... 1  
1.2 Invariance and specificity ............................................................ 3  
1.2.1 Invariance and specificity in the mammalian visual system ............ 3  
1.2.2 Invariant recognition & machine vision ........................................ 4  
1.3 Coding principles in the visual system ........................................... 4  
1.3.1 Sparseness .................................................................................. 5  
1.3.2 Temporal coherence ................................................................. 5  
1.3.3 Unifying temporal coherence and sparseness ................................ 7  
1.3.4 Physiological implementation ..................................................... 7  
1.4 Oculomotor behaviour in cats and primates ....................................... 8  
1.5 Saliency maps .................................................................................. 8  
1.6 Perceptual rivalry ............................................................................. 10  
1.7 Tracking features vs. tracking objects .............................................. 11  

2 General Methods 13  
2.1 Acquisition of real-world stimuli .................................................... 13  
2.1.1 CatCam-videos ............................................................................ 13  
2.1.2 Static images ............................................................................... 15  
2.1.3 Computer vision test images ....................................................... 15  
2.2 Measuring eye-movements ............................................................... 16  
2.2.1 Dr. Bouis oculometer .................................................................. 17  
2.2.2 Eyelink system ............................................................................ 18  
2.2.3 Scleral search coil in monkeys ................................................... 19  
2.2.4 Eye-head coordination in cats ..................................................... 19  
2.3 Presentation setups .......................................................................... 21  
2.4 Colour spaces .................................................................................. 21  
2.4.1 Screen calibration for DKL space ................................................. 23  
2.5 Human subjects ............................................................................... 25  
2.6 Modeling temporal coherence .......................................................... 25  
2.6.1 Neuronal models ......................................................................... 25  
2.6.2 The stability objective function .................................................. 26  

3 Statistics of Natural Scenes 27  
3.1 Specific methods ............................................................................... 27  
3.1.1 Stimuli ....................................................................................... 27  
3.1.2 Spectral analysis .......................................................................... 27  
3.1.3 Wavelet analysis .......................................................................... 28
3.1.4 Correlation analysis .................................................. 29
3.2 Results .............................................................................. 29
  3.2.1 Spatial distribution of luminance and contrast ................. 29
  3.2.2 Contours at the cardinal axes ........................................... 30
  3.2.3 Temporal correlations ................................................... 34
  3.2.4 Time constants of orientation and phase ......................... 39
  3.2.5 Correlations of colour .................................................. 39
  3.2.6 Interaction between different orientations ...................... 40
3.3 Discussion .......................................................................... 40
  3.3.1 Contrast at the fixation spot ............................................ 40
  3.3.2 The oblique effect ......................................................... 41
  3.3.3 Stimulus statistics and neuronal properties ..................... 42
  3.3.4 Anisotropy of lateral interactions .................................... 42

4 Temporal Coherence ................................................................. 43
  4.1 Complex cells ................................................................. 43
    4.1.1 Specific methods ......................................................... 43
    4.1.2 Results .................................................................... 46
    4.1.3 Discussion ............................................................... 50
  4.2 Functional segregation of colour and orientation ................. 51
    4.2.1 Specific methods ......................................................... 51
    4.2.2 Results .................................................................... 52
    4.2.3 Discussion ............................................................... 59
  4.3 Higher order properties: Texture cells ............................... 61
    4.3.1 Specific methods ......................................................... 61
    4.3.2 Data analysis ............................................................. 62
    4.3.3 Results .................................................................... 63
    4.3.4 Discussion ............................................................... 67
  4.4 Invariant representations of objects .................................... 67
    4.4.1 Specific methods ......................................................... 67
    4.4.2 Results .................................................................... 69
    4.4.3 Discussion ............................................................... 75
  4.5 Discussion temporal coherence ........................................... 76

5 Eye-Head coordination in the cat ............................................. 78
  5.1 Results .............................................................................. 78
  5.2 Discussion ........................................................................ 81

6 Bottom-up Attention .............................................................. 83
  6.1 Specific methods ............................................................. 83
    6.1.1 Local modifications of luminance-contrast ..................... 83
    6.1.2 Modification detection paradigm .................................... 85
    6.1.3 Global modification of colour ........................................ 86
    6.1.4 Analysis of eye-traces: 'actual' vs. 'control' fixations ........ 87
  6.2 Results .............................................................................. 87
    6.2.1 Unmodified stimuli ....................................................... 87
    6.2.2 Modified stimuli ......................................................... 89
    6.2.3 Detection paradigm ...................................................... 92
    6.2.4 Interaction of luminance and colour .............................. 93
  6.3 Discussion ......................................................................... 94
6.3.1 Luminance contrast & Saliency .......................... 94
6.3.2 A modified saliency map model ......................... 95
6.3.3 Interaction of different features ......................... 96
6.3.4 Inter-species differences .............................. 96

7 Perceptual Rivalry .......................... 97
7.1 Specific methods ........................................ 97
7.1.1 Experimental setup .................................. 97
7.1.2 Experimental protocol .............................. 97
7.1.3 Data analysis ....................................... 99
7.2 Results ............................................. 100
7.3 Discussion ........................................... 104

8 Automatic Feature binding? .......................... 107
8.1 Specific methods .................................... 107
8.1.1 Subjects ........................................ 107
8.1.2 Stimuli .......................................... 108
8.1.3 Experimental design .............................. 109
8.2 Results ............................................. 110
8.2.1 Discrimination thresholds .......................... 110
8.2.2 Instruction session ................................ 111
8.2.3 Single feature tasks .............................. 111
8.2.4 Learning ........................................ 112
8.2.5 Dual feature task ................................ 112
8.2.6 Learning ........................................ 114
8.3 Discussion ........................................... 115

9 Concluding Remarks & Outlook .......................... 118
9.1 Summary of results .................................. 118
9.2 Natural stimuli and Attention .......................... 118
9.3 Learning from natural stimuli ......................... 119
9.4 On the use of natural stimuli ......................... 119

Bibliography ........................................... 121

Acknowledgements ..................................... 136
Chapter 1

Introduction

Any system that behaves in a real-world environment is continuously confronted with rich sensory stimuli. In comparison to the complexity of such natural stimuli the input- as well as the output-capacity of the behaving system are usually extremely limited. Two strategies are typically employed to deal with this challenge: First, the system may selectively allocate its resources to distinct parts of the input, i.e. shift its attention to process only a small subset of the sensory stimuli at one time (James, 1890). Second, the system may exploit statistical regularities of real-world stimuli and use this a-priori knowledge to process the stimuli more efficiently (Barlow, 1961; Barlow, 1997). While both strategies are employed by humans and other higher animals, they come at a price: Selectively directing attention can lead to ignore large part of the stimulus despite its potential behvioural relevance, as seen quite drastically in examples of inattentional (Neisser & Becklen, 1975; Simons, 2000 for review) or change (Rensink et al., 1997) blindness. Relying on a-priori knowledge fails if the stimulus does not obey the learnt or acquired natural regularities, most evident in case of visual illusions. Nevertheless, in most real-world situations humans and animals by far outperform contemporary artificial systems. In order to develop artificial systems, which are based on the same principles as these animated systems and that shall behave likewise in real-world scenarios, it is therefore vital to achieve a sound understanding of the statistical properties of real world stimuli, of the mechanisms directing attention and of the interaction between these two strategies. In this thesis we restrict ourselves to the visual domain; we investigate statistical regularities of natural visual stimuli and relate them to properties of the mammalian visual system, we address which features of natural scenes guide visual attention and how visual attention in combination with a priori knowledge of the real world influences the percep of well-controlled visual stimuli.

1.1 Statistical regularities of natural stimuli

The first and probably most decisive step to understand the processing of natural stimuli is to understand the statistical properties of the natural stimuli themselves. One of the most elementary questions in natural scene statistics is whether or not they can be assigned a characteristic scale. Analysing the principal components of 15 natural scene (Hancock et al., 1992) showed that those are independent from scale. An important consequence of a general scale invariance would be that the power-spectrum does not follow a distribution which has a scale attached to it (like an exponential), but has to obey a power law. Although the power-spectrum can vary dramatically between natural scenes, it on average roughly
falls off with $\frac{1}{r}$ which is consistent with the image being scale independent (Field, 1987; Ruderman & Bialek, 1994; van der Schaaf & van Hateren, 1996). Hence it seems likely that no general characteristic scale can be attributed to natural scenes. Another relevant question regards the distribution of orientation in natural scenes: It turns out that the first principle components of natural scenes are oriented horizontally and vertically and only the 7th principle component is the first to show a non-trivial (the first component which represents the image mean is of course isotropic) oblique contribution (Hancock et al., 1992). This finding is supported by analysis of power spectra: more power is found at cardinal than at oblique orientations (van der Schaaf & van Hateren, 1996). While these studies used scenes of naturalistic environment (plants, forest, etc.), the same effect is found to be even more prominent in outdoor scenes that contain human artifacts (street scenes, tennis courts, etc.) and indoor scenes (Coppola, Purves, et al.,1998). Consequently the predominance of cardinal orientations seems to be a general feature of natural scenes. The properties described so far are, however, in no way specific for natural scenes: appropriately orientation filtered pink noise would e.g. share the same properties without looking “natural” to human observers. In the context of perceptual grouping, Gestalt psychologists developed the rule of “good continuation” (Wertheimer, 1938): a line tends to be perceived continuing in the established direction. This rule is reflected in natural scene statistics: there a strong long-range correlations of orientations that can extend over the whole scene (Sigman et al.,2001). In addition the same authors find the rule of circularity: line segments are preferentially oriented tangentially to a circle of arbitrary size. This, however, does not necessarily imply that natural scenes are composed of a large number of circular structures, but can be understood as the result of the presence of many smooth closed structures (Chow et al., 2002). This interpretation does, however, not affect the proposed link between the Gestalt rule of good continuation to natural scene statistics. All of aforementioned studies exclusively addressed the spatial properties of static images and did not deal with the temporal domain. However, the temporal domain is vital to human perception as most clearly demonstrated in patients suffering from akinetopsia, i.e. who lack the ability to perceive movement (Zeki, 1991; Rizzo et al., 1995). Using recordings of television broadcasts, (Dong & Attick, 1995a) showed that natural scenes in general are not separable in space and time. They derive from general assumption on objects and velocity distributions, that instead the power-spectrum of natural scenes obeys a $f^n F(\omega)$ law, where F is a non-linear function, f denotes spatial and $\omega$ temporal frequency. However, for high and low spatial frequencies, natural scenes are spatiotemporally separable and follow a $\frac{1}{\omega^n}$ behaviour. How can the described statistical properties of natural scenes be linked to properties of the visual system? Besides the idea of redundancy reduction employed in sparse coding, which we will address in more detail below (section 1.3.1), several other relations between statistics of natural scenes and properties of the visual systems are known, of which we shall give some examples:

- The so-called oblique effect, the fact, that human observers better discriminate gratings at cardinal orientations than at oblique ones (Boltz et al., 1997; Krebs et al., 2000) is linked to the predominance of the former over the latter in natural scenes (Coppola, White et al., 1998; Keil & Cristobal, 2000). More recently (Essock et al., 2003) could show that the oblique effect reverses when stimuli with naturalistic frequency content are used. Again this result can be linked to natural scenes, in that it facilitates the discrimination of orientations that are less frequent and thus more salient (Hansen et al., 2003).

- On the spatiotemporal side, the psychophysically measured human visual sensitivity can directly be derived from the spatio-temporal power-spectrum of natural scenes
(Dong & Attick, 1995b) as can response properties of neurons in the lateral geniculate nucleus (Dong & Attick, 1995c).

- Visual illusions are often thought to be a by-product of using a-priori knowledge to interpret visual scenes. For example the well-known illusion that apparent distance depends on stimulus orientation may result from such a priori knowledge on natural scene statistics: Since the distance of an object in the real world is related to its orientation in the retinal projection, taking this a priori knowledge into account facilitates judging real size from retinal size, but leads to wrong interpretations when non-natural stimuli are encountered (Howe & Purves, 2002).

- Another typical example, in which perception is likely to rely on a-priori knowledge is colour constancy, the fact that the colour of a surface can be perceived identically under a wide range of lightning conditions\(^1\). Recent studies indeed suggest that human observers use their prior knowledge on natural scene statistics to judge the colour of a surface (Golz & MacLeod, 2002; Long & Purves, 2003).

All these examples illustrate that the human visual system exploits a-priori knowledge on natural scene statistics. Our detailed analysis of the spatio-temporal statistics of natural videos in chapter 3 will continue on this idea, and link our results to recent anatomical, physiological and psychophysical data.

1.2 Invariance and specificity

1.2.1 Invariance and specificity in the mammalian visual system

When proceeding upstream through the hierarchy of the mammalian visual system (Felleman & van Essen, 1991) neuronal responses on the one hand get more and more specific: While neurons in primary visual cortex (V1) exhibit specificity to simple stimulus features like orientation (Hubel & Wiesel, 1962), neurons in inferior temporal cortex (IT) can be specific to whole objects as complex as human faces (Perret et al., 1982; Rolls, 1992). In parallel to this increase in specificity, however, there is also an increase in invariance to stimulus features that are irrelevant for the neuron’s coding task: The earliest example of invariance in the visual system is already found at the level of photo-receptors in the retina: By adaptation their response gets largely insensitive to global variations in illumination. One of the most well-known examples for invariant responses is found in primary visual cortex: while one class of neurons - so called “simple” cells - are specific to both orientation and phase the other large neuron class - “complex” cells - are invariant to phase while preserving the specificity to orientation. This sudden emergence of invariance is even more remarkable given that the textbook model for complex cells (Hubel & Wiesel, 1962) assumes a selective connectivity pattern from simple cells to complex cells to form the particular response properties of the latter\(^2\). Higher up in the visual system’s hierarchy, face-selective neurons in IT are largely invariant to complex transformations, such as translations, rotations, changes in view-point or even non-rigid transformations like changes in facial expression (Rolls, 1992; Hietanen et al., 1992; Ito et al., 1995). Experimental evidence furthermore suggests that neurons with comparable response properties are organized into distinct functional pathways (Livingstone & Hubel, 1984; DeYoe & van Essen, 1985; Zeki & Shipp, 1988; Livingstone & Hubel, 1988). The notion of a distinct neuronal pathway also implies that a group of

---

1The phenomenon of colour constancy is discussed in more detail in section 2.4
2see (Mechler & Ringach, 2002) for a recent discussion on simple vs. complex classification
neurons will be selective to one stimulus dimension, but invariant to other dimensions. It is generally agreed that within a given sensory processing level, neurons can be segregated according to their selectivity to a particular stimulus dimension. However, this segregation may not be strictly maintained across multiple processing levels in the brain (Merigan & Maunsell, 1993). At higher processing stages, selectivity to a given stimulus property may be increased following exposure to that stimulus (Logothetis et al., 1995; Sigala & Logothetis, 2002). This suggests that selectivity - and consequently also invariance, may be established through experience, and depend on properties of the sensory input.

1.2.2 Invariant recognition & machine vision

When it comes to invariant object recognition, humans and other higher animals by far outperform contemporary machine vision systems. Since storing every recognizable object under every possible transformation is combinatorially impossible, invariant representations of objects are desirable for any vision system. Consequently a lot of effort in computer vision research is put in generating such invariant representations. If the transformations to which invariance shall be achieved are known, the recognition system may be constructed to achieve this particular invariance: One typical approach is to include the invariance step into image preprocessing, e.g. to achieve invariance to translation by discarding phase information in Fourier space or to perform a log-polar transform (Bradsie & Groesberg, 1995) for invariance to rotations in the image plane. Invariance can also be built-in in the processing as such, a typical example in the context of neural networks being weight-sharing architectures like the Neocognitron (Fukushima, 1980) for translation invariance. Since invariance is a built-in property neither of these approaches can generalize to invariance to different transformations. Another approach overcoming this problem is to train a supervised learning mechanism with large datasets that include examples which have undergone the desired transformation. Although this approach can provide invariances as complex as viewpoint invariance (Weber et al., 2000), it suffers from the drawback, that the number of needed training examples grows exponentially with the number of desired transformations. Consequently, this approach cannot be employed in more general tasks either. All these approaches, that achieve invariance as a built-in property (of the system or the training set), thus lack the ability to generalize to invariance to various transformations. While this might not be too much of a disadvantage if a system is tailored to perform a very specific task, it is more so for a system that has to perform more general recognition tasks in complex natural scenarios. For a system that needs to express a variety of different behaviours in a real-world environment such a restricted recognition system therefore does not seem to well-suit. How can we then overcome the need for hand-crafted invariances or large training sets? In chapter 4 we will follow a different approach: We investigate to what extent a general principle - like temporal coherence, can generate invariant representations, just by exploiting the statistical structure of the (small) input set.

1.3 Coding principles in the visual system

For evolutionary reasons, it seems likely that sensory systems are well adapted to the stimuli they naturally process. Indeed, the human visual system seems to deal more readily with complex natural stimuli than with comparably simple artificial ones, at least in rapid classification tasks (Thorpe et al., 1996; Li et al., 2002). These results imply that the visual system has learnt to exploit some of the observed statistical regularities of natural
CHAPTER 1. INTRODUCTION

scenes\(^3\). This raises the question whether one can find simple rules, that can guide such an adaptive process, and link the statistical structure of natural scenes to properties of the mammalian visual system. Two general coding principles have been investigated extensively in the past two decades: sparseness and temporal coherence. Although the present thesis mainly restricts on the latter, we will introduce both briefly here.

1.3.1 Sparseness

It is often assumed that neuronal codes should represent their typical stimuli as efficiently - be it in metabolic or information theoretical terms, as possible, i.e. to remove redundant information from the input (Barlow, 1961; Barlow, 2001). It was suggested that this requirement can be fulfilled by neurons or neuronal populations that encode their input sparsely (Field, 1987). A random variable herein defined to be sparsely distributed, or sparse, if it takes very high and very low absolute values more frequently than a Gaussian random variable. Thus a variable is sparse, if its probability density function (pdf) has a large peak at 0 and a heavy tail (compared to a Gaussian of same variance). In the context of neuronal activities, two types of sparseness have to be carefully distinguished, the sparseness of a single neuron’s activity over time (“lifetime-sparseness”) and the sparseness across a population at a given point in time (“population sparseness”), as pointed out in (Willmore & Tolhurst, 2001). Several experimental and modeling studies have recently addressed the sparseness of neuronal codes in the visual system (Fyfe & Baddeley, 1995; Rolls & Tovee, 1995; Olshausen & Field, 1996; Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998; Vinje & Gallant, 2000; Weliky et al., 2003):

Using natural scenes as stimuli, both lifetime sparseness and population sparseness were investigated in various areas of the visual cortex. Under natural scene stimulation single unit activity of V1 neurons in the awake rhesus monkey exhibit high life-time and population sparseness, which can be increased by stimulation outside the classical receptive field (Vinje & Gallant, 2000). More recently, multiunit activity of layer 2/3 cells in ferret V1, recorded by electrodes from the cortical surface, were found to also exhibit high population sparseness and life-time sparseness (Weliky et al., 2003). Further up in the visual hierarchy it is known that neurons in primate IT show a considerably high population sparseness (Rolls & Tovee, 1995). These results suggest that sparseness is a general coding principle used throughout the visual cortex.

Given the importance of sparseness in representing natural scenes, can in turn the properties of the mammalian system be deprived from the principle of sparseness? In 1996 Olshausen & Field showed that model neurons that optimize population sparseness and are at the same time constrained to faithfully encode the natural input image, resemble simple cell receptive fields (Olshausen & Field, 1996). This result is qualitatively independent of the exact form of the sparse distribution\(^4\). Using the method of independent component analysis (ICA), which is closely related to sparse-coding (Hyvärinen et al., 2001), (Bell & Sejnowski, 1997) obtained simple cells as the independent components of natural scenes. The ICA approach was subsequently also applied to videos and found to not only represent the spatial but also the temporal structure of simple cell receptive fields (van Hateren & van der Schaaf, 1998). The group of Hyvärinen then applied the principle of ICA also successfully to derive various properties from natural scenes: complex cell receptive fields (Hyvärinen & Hoyer, 2000), simple cell receptive fields (Hyvärinen & Hoyer, 2001), colour and disparity selective cells

\(^3\)\textit{When we use the term “learning” for such an adaptive process in this thesis, this is not meant to imply that the adaption in the animal’s visual system takes place on an ontogenetic timescale, but also stands loosely for some evolutionary adaptive process.}

\(^4\)Olshausen & Field tested \(-e^{-x^2}, \log(1 + x^2)\) and \(|x|\)
(Hoyer & Hyvärinen, 2000) and contour selective cells (Hoyer & Hyvärinen, 2002). In summary, applying the principle of sparseness to natural scenes can explain neuronal response properties at various levels of cortical processing.

1.3.2 Temporal coherence

The idea underlying all temporal coherence approaches is the trivia that different features vary on different timescales. Furthermore, there is the intuition that more “interesting” or “relevant” features vary on slower time-scales than “irrelevant” features, often illustrated by the extremes of the identity of the object (which is slow) and the luminance fluctuations at the input (which are fast). It is tempting to assume that if we can get rid of all the fast transformations of a scene we will finally be left only with the object that we then can easily classify. Although a lot of simulation and theoretical work has been put in the temporal coherence idea in the last two decades since Földiak’s original paper (Földiak, 1991), even the proof of concept that temporal coherence indeed can facilitate such invariant object recognition for real world objects had not been delivered. Instead most studies focused on either simple transformations or on properties of the early visual system. This is further complicated by the fact that the idea of temporal coherence comes in different flavors: Starting from Földiak’s trace rule implementation and network architectures based on it, several implementations have been put forward: so-called slow feature analysis, our stability implementation as well as physiologically inspired models.

Trace rule

The original implementation of temporal coherence (Földiak, 1991) introduces the so-called trace rule: the modification of the synaptic weight is not proportional to pre- and postsynaptic activity as in classical Hebbian learning, but the postsynaptic activity is replaced by a running average over previous postsynaptic activity (the “trace”). Training a grid of cells by repetitively sweeping an orientated bar over it, Földiak by this rule reproduced orientation specificity and translation invariance of complex cells. To resemble the structure of the ventral stream of the visual cortex (Wallis & Rolls, 1997) implemented the trace rule in a hierarchical architecture (“VISNET”). Cells trained with translated, rotated and scaled images of faces obtained responses, invariant to the respective transformation for that limited data-set. Restricting themselves to translation invariance, (Rolls & Milward, 2000) extended their scheme (“VISNET2”) using different variants of the trace rule, a larger stimulus set and different information theoretic measurements. Consequently the trace rule still is a promising implementation of the temporal coherence principle, especially when large hierarchical structures are concerned.

SFA and stability

So-called “slow feature analysis” (SFA) puts the idea underlying all temporal coherence implementations, to extract the slow features out of a given (fast) input signal, directly into a mathematical framework (Wiskott & Sejnowski, 2002): In brief, they state the optimization problem of SFA as follows: Given an input vector $\mathbf{x}(t)$ one searches for the functional $y(t) = g(x(t))$ for which the squared temporal derivative averaged over time $\left(\frac{d^2 y(t)}{dt^2}\right)^2$ is minimal. Without further constraints $g=const$ would be a solution. To avoid this trivial solution the $y_j$ are constrained to zero mean and unit variance. So far there is no interaction between different $y_j$ and $y_i$ for $j \neq i$ and optimization would thus lead to identical
\( y_i = y_j \forall (i,j) \). Therefore the \( y_i \) additionally constrained to have vanishing covariance, i.e. including the unit variance constraint: \( <y_i y_j> = \delta_{ij} \). Wiskott and colleagues have shown that SFA yields invariance to a huge variety of transformations using artificial (Wiskott & Sejnowski, 2002) as well as natural (Berkes & Wiskott, 2003) stimuli. In the latter study the authors can also relate the properties of cells obtained with SFA to a variety of properties of complex cells in primary visual cortex. In a recent study (Wiskott, 2003) could analytically analyse the free solutions of SFA, i.e. the optimal \( y_i \) with no input signal (but boundary conditions) given. This showed the great value of the SFA implementation as framework to treat temporal coherence in a mathematically rigorous manner. The stability implementation of the temporal coherence principle used in this thesis (section 2.6.2), that we first proposed in (Kayser et al., 2001), is closely related to the SFA formulation. We also try and minimize the squared temporal derivative and we use the variance and covariance to avoid trivial solutions. However, we do not “hard-wire” the constraints, but implement them softly, i.e. on the same level as the minimization of the temporal derivative, by integrating them into a single objective function, which is then optimized. In chapter 4 we will show how this implementation does not only describe properties of the early visual system, but also can be used to facilitate invariant object classification.

1.3.3 Unifying temporal coherence and sparseness

Although both sparseness and temporal coherence are principles that can be used to learn properties of either V1 cell type, only few attempts have been made to unify both in a single framework. An obvious approach to combine sparseness and temporal coherence, is using the former for a different definition of the latter: Instead of optimizing the squared derivative as in SFA and our “stability” implementation, (Hashimoto, 2003) instead maximize the sparseness of the derivative. Although the authors do not provide quantitative measures of complex cell properties, their cells seem to more closely resemble complex cells than those obtained - in their simulations, by optimizing the squared derivative. Including sparseness in the definition of temporal coherence thus may lead to a quantitative improvement of the receptive field properties also in other implementations.

The group of Hyvärinen tried to put sparseness and temporal coherence in a unified framework: As a first step (Hurri & Hyvärinen, 2003) optimized their so-called “temporal response strength correlation”, which for certain non-linearities approaches a sparseness measure, when the temporal delay goes to 0. By this implementation of temporal coherence the authors obtain simple cell type receptive fields. In (Hyvärinen et al., 2003) the same authors continue and assume signals which are sparse and temporally correlated (in their terms “spare temporal activity bubble”). By additionally introducing a topography to the cells, they obtain “spatio-temporal activity bubbles”, whose linear mixture then represent an image-sequence. Using natural images (Hyvärinen et al., 2003) obtain simple cell like receptive fields, that are topographically ordered according to their orientation, spatial frequency and location preferences. In addition they use the same formulation to learn spatio-temporal receptive fields. Although the difference in results to previous topographic models, which did not involve temporal coherence (Hyvärinen & Hoyer, 2001), are only of quantitative nature, the “bubbles” formulation offers a promising possibility to combine sparseness and temporal coherence in one framework.

1.3.4 Physiological implementation

In a study not described in this thesis (Einhäuser, 2001; Einhäuser et al., 2002a) we have also put forward a more physiologically inspired implementation of the temporal coherence
principle. We implemented a two layer network, which is trained using natural videos as input. The first layer employs a Hebbian learning scheme with strong lateral inhibition. This can be interpreted as a sparseness constraint and leads to the emergence of simple cell type receptive fields. Second layer neurons implement the idea of spike-timing dependent plasticity: connections to a first layer cells get strengthened when it was highly active a time-step before the top-layer cell is active and weakened otherwise. This mechanism can be interpreted as physiologically inspired implementation of the temporal coherence principle and leads to the emergence of complex cell like receptive fields. Besides implementing sparseness and temporal coherence in the same framework, the model has the further advantage that it is local in time (±1 timestep) and thus can be used for online and real-time learning. Several concerns may, however, arise: First, although the idea is based on spike-timing-dependent plasticity, the model is formulated in terms of firing rates. Second, no attempt has been made to realistically model the temporal dynamics of receptive fields. Finally - at least in the case of complex cells - learning from natural stimuli on an ontogenetic timescale seems unlikely, since primary visual cortex neurons already exhibit adult like properties in the visually unexperienced newborn (Hubel & Wiesel, 1963; Wiesel & Hubel, 1974). Consequently the use of the physiological model in its current form beyond providing a proof of concept is limited. Since online learning is not yet an issue, we thus - in this thesis, restrict ourselves to the more abstract stability implementation of the temporal coherence principle.

1.4 Oculomotor behaviour in cats and primates

Our analysis of attentional processes in this thesis will mainly restrict to overt attention, i.e. attention associated with the shift of gaze. Hence we here will briefly summarise some facts on oculomotor behaviour of the species used in this thesis. Both in cats and primates one may distinguish three types of gaze shifting eye-movements: saccades, pursuit and vergence movements (Goldberg, 2000). Saccades are high velocity eye movements that orient the eye to a target of interest. Saccades are well investigated in a variety of paradigms and species (Evinger & Fuchs, 1978; Evinger et al., 1981; Guitton et al., 1990; Araujo et al., 2001). Pursuit movements keep the image of a, typically small, moving target on the central part of the retina (Missal et al., 1995). Vergence movements adjust the convergence of the two eyes and will not be considered in this thesis. The other two types of gaze shifting eye movements, saccades and pursuit movements, induce a shift of a large part of the visual scene relative to the retina and thus create global optical flow. A second class of eye movements serves the stabilization of the image on the retina: The optokinetic reflex and the vestibulo-ocular reflex. The optokinetic reflex is triggered by a global unidirectional motion of the visual environment and serves to minimize the retinal slip. As a result the eye movement is in the same direction as the optical flow. The VOR results from the detection of angular velocity of the head by the semicircular canals. It elicits eye movements that compensate for the head movement. Thus, both types of gaze stabilizing eye movements stabilize the image on the retina and try to minimize optical flow.

When comparing the feline and the primate eye there are two striking differences that will be relevant for oculomotor behaviour: While the primate retina achieves its highest resolution in only a small region (the so-called fovea) spanning about 0.5° of visual angle, the resolution in the feline retina is about an order of magnitude lower but about constant over 5° in the central retina (the so-called area centralis). Hence the need to actively shift the gaze to optimize resolution is larger in primates than in cats. Secondly the oculomotor range of cats is restricted to about ±25° (Guitton et al., 1984), which compares to about ±70° in primates. Consequently single eye-saccades are not sufficient in cats to cover the full visual
space, which spans about \( \pm 70^\circ \). Likely as a consequence of these two anatomical differences primates perform saccades frequently (about 3-5 per second) and largely independently of head-movements, while - at least under laboratory conditions - cats do not perform any saccades in the absence of head-movements (Guitton et al., 1984). Given this species-difference, we - in addition to our measurement in human and non-human primates (chapters 6 and 7), will also briefly address eye-head coordination in a cat that freely behaves in its natural environment (chapter 5).

1.5 Saliency maps

How humans direct their overt attention in complex natural scenes is influenced by the task, the past experience of the observer and the stimulus properties (Yarbus, 1967). Most theories on the selection of the focus of attention, which address the latter - "bottom-up"-driven part, rely on the concept of a "saliency map" (Koch & Ullman, 1985; Itti & Koch, 2000): The input image is analyzed locally with respect to various stimulus properties such as luminance, orientated contours, and colour. Retinotopic gradients in the resulting maps (contrasts) are summed up, and the location to be attended is selected by a subsequent winner take all process (figure 1.1). In short, regions with high contrasts (e.g. luminance-contrast, colour-contrast, orientation-contrast) attract attention.

As physiological substrates for encoding saliency - besides the idea of distributed emergence from locally competitive interactions (Corchs & Deco, 2002) - various individual brain regions have been proposed: Recent studies, which found increased luminance-contrast at the center of gaze when viewing natural images (Reinagel & Zador, 1999; Krieger et al., 2000), suggest that luminance-contrast is a major contributor to the saliency map for visual atten-
tion (Parkhurst et al., 2002). As neuronal responses in early visual areas are highly sensitive to luminance-contrast, these findings are consistent with the encoding of saliency in these early areas (Lee et al., 2002; Li, 2002). On the other hand, studies using lesions of brain areas as well as electrophysiological studies that used specific artificial stimuli, found saliency to be represented in different, non-exclusively visual, brain regions. The pulvinar (Posner & Petersen, 1990; Robinson & Petersen, 1992), the superior colliculus (Posner & Petersen, 1990; Kustov & Robinson, 1996; Horwitz & Newsome, 1999; McPeek & Keller 2002), the frontal eye field (Thompson et al., 1997) as well as the lateral intraparietal area (Gottlieb et al., 1998) have been associated with the encoding of saliency maps. This evidence suggests that the computation of a saliency map is not strictly performed in early visual areas. Furthermore, as luminance-contrast sensitivity decreases in the course of the hierarchy of the visual system (Avidan et al., 2002), these findings are in conflict with the above observation of a correlation between overt attention and luminance-contrast. Therefore, the question has to be reconsidered, whether luminance-contrast indeed causally contributes to the selection of fixation points, or just happens to be correlated to attention-attracting higher-order properties of natural scenes. This distinction is decisive for the question on the underlying neural substrate. In chapter 6 of this thesis we address this issue by using natural visual stimuli with locally modified luminance-contrast while keeping the other stimulus parameters constant. We investigate whether these modifications influence the direction of gaze in humans and macaques. In a separate paradigm it is additionally tested whether the used modifications fall within a perceptually relevant range for humans. Since the “classical” saliency map model (Koch & Ulman, 1985; Itti & Koch, 2000) furthermore assumes independence and linear interaction between features we - in a third paradigm, test the effect of colour on the use of luminance-contrast in guiding human overt attention.

1.6 Perceptual rivalry

Although the experiments on the stimulus-driven (bottom-up) influences on overt attention try to minimize the influence of task and higher cognitive processes, they - in principle, can never rule out these top-down effects completely. To systematically dissect top-down from bottom-up influences on overt attention, we therefore follow a different paradigm. We exclude any bottom-up effect by using stimuli, which itself do not change, while their perceptual (i.e. top-down) interpretation does. Such rivalrous stimuli are widely used as a tool to assess the neural correlates of subjective perception, independently from physical changes to the stimulus. Several models have been put forward as to what mechanism these perceptual alterations arise from. These suggestions include the neural fatigue hypothesis (Köhler, 1940; Hochberg, 1950; Long & Toppino, 1981) and variants (Gomez et al., 1996), changes in neuronal synchronization (Fries et al., 1997), planned interventions from non-visual cortical areas (Leopold & Logothetis, 1999) as well as inter-hemispheric switching (Miller et al., 2000). In most of the used paradigms (animal or human) subjects are instructed to maintain their fixation throughout the task (Leopold & Logothetis, 1996; Miller et al., 2000), or oculomotor responses are explicitly used to assess the perceptual state (Fries et al., 1997). Both eliminate possible feedback by changes in the retinal input or feedback from the oculomotor system itself. On the other hand, there seems to be a consensus that eye-movements relate to the switch of percepts, at least for viewing bistable figures. The causal direction of this relation, i.e. whether eye-movements cause the perceptual switch or vice versa, however, has been subject to debate for nearly a century, without reaching a final decision. (Sission, 1935; Glen, 1940; Pheiffer et al., 1956; Magnusen, 1970; Flamm & Bergum, 1977; Ellis &
Stark, 1978; Kawabata et al., 1978; Peterson & Hochberg, 1983). In spite of the importance of this question for models of multistable visual phenomena, modern studies on the relation of gaze position and perception are rare. Most recent studies that relate ambiguous figures to the oculomotor system rather focus on other parameters, like fixation duration (Ellis & Stark, 1978), induced smooth eye movements (Scotto, 1991) as well as vergence and pupil diameter (Enright, 1987). Along this line, it has recently been found that voluntary blinking (as well as blanking of the stimulus) stabilizes the percept of a bistable figure (Leopold et al., 2002).

In chapter 7 we address this issue for the perhaps best known ambiguous figure, Necker’s eponymous cube, which actually started its life as a rhomboid. Necker, then professor of Mineralogy in Geneva, noticed the switch while studying engraved plates of rhomboid-shaped crystals. In a letter to Sir David Brewster in 1832 (Necker, 1832), Necker described the switch, which, in his opinion, resulted from ‘an involuntary change in the adjustment of the eye for obtaining distinct vision.’ Wheatstone soon showed that Necker’s interpretation was wrong on geometric grounds and claimed instead that the switch depended only on ‘our mental contemplation’ (Wheatstone, 1838). Wheatstone’s interpretation was of course strongly influenced by his systematic studies of the related phenomenon of binocular rivalry using his recently devised stereoscope. Hering, contrariwise, claimed that the switching rate could be influenced by eye-movements and differential retinal adaptation (Hering, 1879).

A century later the debate remains very much alive: Do eye-movements at all relate to perceptual switching or not (Flamm & Bergum, 1977; Ellis & Stark, 1978)? And if so, do eye-movements precede the perceptual switch, as some have suggested (Glen, 1940; Kawabata et al., 1978) or are they an involuntary consequence of the perceptual switch, as others argued (Zimmer, 1913; Pfeiffer et al., 1956)? Unfortunately, most attempts to decide between these options have either had difficulties in measuring accurately the temporal relation of eye-movements and perceptual switches, or had specifically instructed their subjects to make (or to suppress) eye-movements. In chapter 7 of this thesis we measure eye-position while subjects were presented Necker cubes under free-viewing conditions to establish the temporal relation between perceptual switching and eye-movements. We found that perceptual switching consistently preceded changes in eye-position. The change in eye-position evoked by the new percept may itself provide negative feedback on the new percept.

1.7 Tracking features vs. tracking objects

Humans are capable of accurately tracking moving targets through physical space, even in the presence of visually identical distractors (Pylyshyn & Storm, 1988). When two of the objects coincide - i.e. their spatio-temporal trajectories overlap - observers need to employ additional constraints to disambiguate the situation. In the absence of contextual cues, there seems to be a concept of inertia: If one object moves at constant speed through the point of coincidence, it is typically perceived as passing across the other object, no matter whether the second object approaches at the same speed from the opposing direction (Metzger, 1934) or is at rest (Scholl & Nakayama, 2002); if, however, the first object stops before fully overlapping with the second object, the second object is typically perceived as being launched by the first one (Michotte, 1963). The judgment of either object’s identity after the coincidence event, however, depends strongly on the visual (Scholl & Nakayama, 2002) and non-visual (Sekuler et al., 1997; Watanabe & Shimojo, 2001) context. (Sekuler et al., 1997) report that an auditory stimulus at the time of coincidence biases observers to interpret the two objects bouncing from each other rather than steaming past each other. Using this effect in infants Scheier et al. (Scheier et al., 2004) provide evidence that the
concept of object knowledge relevant for this task undergoes a change at about 6 months of age. Consequently, the notion of object identity after coincidence seems to be determined by physical concepts, but is also subject to plasticity. In a recent study (Blaser et al., 2000) show that tracking of objects is not restricted to spatio-temporal trajectories in physical space: In their study 'objects' are Gabor wavelets, that vary smoothly in orientation, colour and spatial-frequency. This variation is described by a smooth trajectory in the 3 dimensional "feature-space" spanned by orientation, colour and spatial-frequency. (Blaser et al., 2000) find that this "feature-temporal" trajectory is sufficient to track a target Gabor, even if a distractor is spatially superimposed and follows a different feature-temporal trajectory. Their result suggests that humans can perform tracking not only on the basis of spatio-temporal trajectories, but can also use abstract feature-temporal trajectories. This striking similarity between tracking in physical space and tracking in feature space raises the question, to what extent the aforementioned principles that disambiguate coincidences in physical space can also be transferred to feature space? When compositions of features are tracked simultaneously through feature space, the question arises whether tracking along different feature dimensions is done independently, or interactions occur. Interactions would imply that subjects bind the target features into the percept of a single object. Indeed, (Blaser et al., 2000) suggest that their stimuli are not tracked on the basis of their individual features, but as a single entity. This implies that not the features themselves, but their composition are the basic units in tracking tasks. Consequently compositions of features that are distinct from other features just by being the target, are bound to an object, solely on the basis of this fact. The (Blaser et al., 2000) paradigm, however, uses direct feedback, that might effectively train subjects to bind compositions of features into an object entity for tracking. This raises the further question, whether the tracking of different features as common object is as a consequence of training or does also naturally occur in observers, who do not receive feedback? Is the sole fact, that some features are defined as targets and others not, sufficient to induce binding of features into objects?

In chapter 8 we use a variation of the (Blaser et al., 2000) stimuli to investigate how - in the absence of feedback, naïve adult observers disambiguate feature-temporal coincidences in one and two dimensional feature spaces. Our paradigm sheds new light on the issue to what extent the notion of what defines an "object" is subject to plasticity in the adult.
Chapter 2

General Methods

This chapter provides the basic methods used throughout this thesis. Methods that are specific to a single experimental paradigm, will be provided directly in the corresponding chapter. An overview of the different types of natural stimuli used in this thesis (section 2.1) is followed by a detailed description of our eye-tracking systems used in cats, monkeys and humans (section 2.2). The presentation setup used are described in section 2.3. The overview over different colour spaces includes a detailed description on monitor calibration 2.4. After a brief statement on the choice of human subjects (section 2.5), we conclude the methods with the methods used in the theoretical part of this thesis, including a detailed description of our “stability” implementation of the temporal coherence principle (section 2.6). Unless otherwise stated, simulations and data-analysis were performed using the MatLab (Mathworks, Natick, MA, USA) programming environment.

2.1 Acquisition of real-world stimuli

While “real-world” stimuli should be as naturalistic as possible, the priority of requirements posed on natural stimuli may differ depending on the particular experiment. Whereas the use and investigation of temporal statistics obviously requires the natural timecourse to be preserved, experiments involving human subjects on the other hand require high spatial resolution. While acquisition of natural scenes should be as unbiased as possible, object recognition requires the presence of many well-defined objects. Given limited recording and processing resources, these requirements cannot be fulfilled by one single stimulus set. Consequently, three different types of real-world stimuli are used in this thesis, which are described in detail below: videos acquired by a camera mounted to a freely-behaving cat (section 2.1.1), high resolution still images of natural outdoor scenes (section 2.1.2) and object images taken from a standard database (section 2.1.3). All self-recorded video and image data are available from the author upon request.

2.1.1 CatCam-videos

Four cats are used for collecting natural videos (figure 2.1). They wear chronic skull implants, containing microelectrodes, which are used in another study. This implant is made of dental cement and carries 2 nuts, to which we temporarily attach a CCD-camera (Conrad Electronics, Hirschau, Germany). The camera is mounted on a circuit board measuring 38 mm by 38 mm yielding a total weight of 34 g. This camera has a global gain control, which
partly mimics adaptation processes of the visual system. The time constant of this adaptation process (2.5s) does not quantitatively match the characteristics of luminance adaptation of the cat visual system precisely. However, it is at least an order of magnitude longer than the time constants of processes investigated in this thesis. Hence, it thus does not confound any of the presented analysis. The vertical distance of the camera from the eyes is about 4 cm. The camera is aligned to the visual axis of the animal, when it is looking straight ahead. We control this by examining whether objects, the cat is fixating on from various distances, are represented in the central image area. When the cat is very close to an object (less than about 20 cm) and potentially touching it, the images taken by the camera do not match the visual input of the animal anymore. Therefore, data originating from such instances are discarded. The camera contains a 10 mm lens featuring a visual angle of 71 degrees in the horizontal direction and 53 degrees in the vertical direction. Due to the short focal length of the lens the whole visual field is in focus. The camera connects to a backpack carried by the scientist. It contains the recording equipment: an analog PAL-VCR (Lucky Goldstar, Seoul, Korea), a 12V-battery for power supply and a miniature television for online control during outside walks.

We collected movies in colour of an average length of 20 minutes, spread over the 4 seasons, pertaining to various weather situations and locations (forest, university campus, grassland, residential areas, figure 2.2). Sections of the videos containing recording artifacts (e.g. glare which saturates whole columns of the CCD array) were excluded from further analysis. Of our database we thereby chose sequences of about 3000-5000 frames (53884 frames in total), recorded at a sampling rate of 25 frames per second. These are digitized at a resolution of 320*240 pixels using a miroVIDEO DC30 graphics card (Pinnacle Systems, Mountain View,
California) and Adobe Premiere software (San Jose, CA). The data are stored in 24-bit (RGB) TIFF-format. In experiments using grey-scale images the conversion is done using the default mapping of matlab’s rgb2gray.m function. The pixel values are mapped linearly on the interval from 0 to 1 with identical scaling for all frames.

### 2.1.2 Static images

For all experiments of this thesis that involve eye-tracking and natural stimuli, we use stimuli that are based on natural photographs of local outdoor environment. They depict scenes of open area or in the forest and no man-made objects are visible on these images. Subjectively, they resemble stills from the aforementioned natural videos, that were taken by a camera mounted on the head of a cat (2.1.1), but are taken with a high quality digital camera (3.3Mega pixel colour mosaic CCD, Nikon Coolpix 995, Tokyo, Japan) for increased spatial resolution. For display purposes they are downscaled by a factor of 2 using bicubic interpolation to a final resolution of 1024x768. For eye-tracking studies not involving colour they are converted to 8-bit grayscale using the default mapping of matlab’s function rgb2gray.m. Twelve examples out of the database of 128 images are shown in figure 2.3.

### 2.1.3 Computer vision test images

Computer vision test-images used in this thesis were derived from the Columbia University Object Image Library (COIL-100,(Nayer et al., 1996)). This database provides photographs of 100 objects from 72 different viewpoints each (5 degree steps). Objects are located in front of a black background at a original resolution of 128x128 pixels and 3 channel (RGB) colour representation. In this thesis we use stimuli in two ways: in a plain condition, where one object ('target') was present on the retina, and in a 'cluttered' condition, where another object ('distractor') was placed on the retina in the background of the target. To allow for partial overlap in the cluttered condition from all 7200 stimuli the background was removed before processing. This was done by pixel-wise applying a threshold to the luminance (weighted
Figure 2.4:  a) Plain stimuli. Several training views of each object are shown repeatedly at different positions; the square retina has 64x64 pixel resolution in 3 colour channels. b) In the cluttered condition a random object is added as distractor to the background of the target object at a random location.

sum of all colour channels) and such that rather parts of the object boundary were cut than leaving black fringes around the object. In both conditions objects were placed at a random position on a 256x256 pixel wide retina, which afterwards is down-sampled to 64x64 using bi-cubic interpolation (figure 2.4a).

For training only certain views of the objects were used, the number of training views per object being a divider of the total 72 views. From one time-step to next the viewpoint of each object was changed by \( \frac{\text{number of training views}}{360} \), while the position was randomly selected independently across time-steps (figure 2.4a). All views of each object were presented 100 times. In the cluttered condition the distractor was selected and placed independently for each step (figure 2.4b). Distractor objects were always taken from the whole database even if the network was trained only on a subset.

For testing, all 72 views of each trained object were presented to the converged network 100 times at random position. For testing in the cluttered condition a distractor was added for each test presentation. The object used as this testing distractor was chosen randomly from a subset of objects disjoint from the trained objects. That means, if objects #1 to #10 were used as targets for training, objects #11-#100 were used as distractors; for cells trained with object #1-#50 as target testing distractors were objects #51 to #100. To ensure a testing distractor set of at least equal size as the trained target set, the maximum number of trained objects was restricted to 50.

### 2.2 Measuring eye-movements

A variety of methods exists to accurately measure eye-position in humans and non-human animals. Data presented in this thesis uses four different eye-tracking methods depending on species and task. For humans we use as well a stationary system (section 2.2.1) as a head mounted system (section 2.2.2), both of which are non-invasive. For monkeys we use an implanted eye-coil (section 2.2.3) and for cats we use a custom-designed non-invasive optical system (section 2.2.4).
\subsection{Dr. Bouis oculometer}

When the primate eye is illuminated by infrared light, several reflections resulting from different boundary layers inside the eye occur. The most dominant reflection, however, is the so-called first Purkinje image, the reflection occurring at the front of the cornea. Being agnostic about the exact source of the reflection, the oculometer (Dr Bouis, Karlsruhe, Germany) described in this section utilizes the centre of gravity of the eye’s reflection when illuminated with diffuse low-intensity infrared light. Empirically, the dependence of centre of gravity on eye-position is linear, which is why the center of gravity provides a good measure of eye-position (Bach et al., 1983).

The eye is diffusely illuminated by an array of infrared diodes via a mirror that is transparent to visible light. Via the same mirror the eye’s reflection is reflected back to a 20x20mm wide photodiode, which senses the center of gravity of the reflection. The resulting photocurrents are transformed electronically into two output voltages, one encoding horizontal eye-position, the other vertical eye-position. The voltages recorded by a CIO-DAS16/CTR/JR card (Computer boards, Mansfield, MA, USA) installed into an i80486DX40 PC (Image Microsystems, McLean, VA, USA) at a sampling rate of 1kHz. To ensure precise timing between stimulus, subjects’ responses and eye-position, the same board is used to record signals corresponding to such events.

The measured oculometer output voltages are - given optimal adjustment - approximately linear to horizontal and vertical eye-position respectively. However, since the subject’s position might be shifted with respect to the eye-tracker or to the screen, this linearity does not hold perfectly. In addition some cross-talk between horizontal and vertical position may occur, i.e. the transformation from screen coordinates to eye-tracking output-voltage for horizontal and vertical position might mutually depend on each other. All these effects are accounted for by assuming a bi-linear transformation between screen-coordinates \((x',y')\) and voltage output \((x,y)\):

\[
\begin{align*}
x' &= a_1 x + a_2 y + a_3 xy + a_4 \\
y' &= a_5 x + a_6 y + a_7 xy + a_8
\end{align*}
\] (2.1)

To obtain the parameters \(a_i\) subjects are required to fixate defined points presented on the screen before each experimental block. In those calibration trials \((x',y')\) and \((x,y)\) are both known and the \(a_i\) are chosen that the estimated screen coordinates \((x',y')\) given by equation 2.1 optimally fit (in least squares sense) the known real screen coordinates \((x',y')\). The thus obtained \(a_i\) are used for the subsequent experimental block.
Minimizing subjects’ movements

A major problem in stationary systems is to fix the subjects relative to the setup. While the combination of a forehead-rest and a chin-rest successfully suppresses the head from tilting and rotating around its vertical axis, it turns out to be insufficient to suppress horizontal rotations. We overcome this problem by additionally using a bite-bar made of thermoplastic material (Kerr). While this procedure stabilizes the subjects’ head reasonably well, it also increases the subjects’ discomfort and thus limits continuous and overall experimental duration. The requirement of strict head restraint is a general problem associated with the use of stationary eye-tracking systems in humans. As a consequence head-mountable systems - as described in the following section - are of advantage for most applications.

2.2.2 Eyelink system

Given the problems associated with the use of stationary systems in human subjects, we for some experiments also use a commercial head-mounted system, eyelink I (SR Research, Toronto, Canada). The system consists of there high-speed (500Hz) cameras, two of which record one eye each, while the third serves to compute compensations for head-movements, all mounted on a head-band (figure 2.6). Eye-position is computed from the camera’s image using the center of mass of the pupil in combination with the first Purkinje image. The third camera, pointing parallel to the subject’s line of sight, records four infrared markers located at the corners of the screen. Movements of the markers relative to camera are used to compute head-position. Gaze direction is computed from eye-position and corrected for (slight) head movements using software provided by the manufacturer. To achieve comparable results, the same calibration procedure as above is applied in addition and verified with the calibration of the “eyelink” system itself. The terms corresponding to the cross-talk between channels ($a_3$ and $a_6$ in equation 2.1) turn out to be negligible. Comparing the results of calibration trials performed before and after experimental blocks the absolute gaze direction
error can be estimated to be below 1° including drift. In conclusion, if high precision and long experimental blocks are required, but timing in the sub-millisecond range is not an issue, the “eyelink” system seems optimally suited for eye-tracking in human subjects.

2.2.3 Scleral search coil in monkeys

Eye-tracking in monkeys was performed using scleral search coils, following the method described in (Judge et al., 1980). The two monkeys used for experiments in this thesis had their search-coils implanted at the University of Bochum in the context of different studies. Calibration of the coils’ output signal relative to screen coordinates was done using the same bi-linear method as described above. During calibration trials monkeys were rewarded manually for correct fixation. Unlike for experiments with humans, the mapping between coil-output and screen coordinates was very stable, such that all calibration trials could be used for calibrating all experimental trials of the same day. From calibration trials the precision of the eye-position signal is estimated to about 0.5° in a head free monkey and far below this value in a head fixed monkey. Although this high precision in eye-tracking is desirable and justifies the invasive procedure, one should keep in mind that most errors result from head movements, such that - with head fixed monkeys - similar precision can also be reached with non-invasive methods.

2.2.4 Eye-head coordination in cats

To simultaneously record eye and head movements in the freely behaving cat, we constructed a light-weight (89.0 g) setup consisting of two identical CCD-cameras (DFM 5303, The Imaging Source, Germany). As for the recording of ‘catcam’ videos (section 2.1.1), this setup was reversible attached to the mounts of the cat’s cranial implant and recordings were made while the cat explored an outdoor environment. One of the cameras (‘front camera’) pointed parallel to the cat’s head and captures a large part of the cat’s visual field (80 degrees viewing angle). The second camera (‘eye camera’) captured the cat’s eye via a small magnifying mirror (figure 2.7).

Recording and processing for both cameras was as described above, with the exception that the spatial resolution was 360(H) x 270(V) instead of 320x240. The videos obtained from both cameras were temporally aligned by synchronous automatic inversion of the cameras’ colour look-up tables every 250 frames. This preserves all recorded data and additionally allows controlling for potential frame-loss during digitization.

In order to compute relative eye position from the eye camera’s signal, the pupil was modeled as the intersection of two circles of equal radius centered on the same horizontal line. The radius, the position of one circle center and its distance to the other center were varied for fitting the pupil. In each frame these free parameters were chosen such that the circles’ intersection optimally matched the black (< 70% of mean luminance) area of the pupil. When the quality of the match between pupil model and black area was below a threshold the respective frame was treated as missing data. Otherwise the eye position was defined as the midpoint of the circles’ centers. (Figure 2.8)

In order to compute head movements from the front camera’s signal, the image flow between two subsequent frames was computed. 71x71 pixel wide image patches were extracted at 35 (5x7) equally spaced grid points from each frame. These patches were correlated to the subsequent frame. For each grid point, the position of the maximum correlation defines the end point of the local flow vector. A local flow vector is discarded as invalid if its length exceeded 100 pixels, as these were most often false matches. When there were less than 20 valid local flow vectors or the standard deviation of valid local flow vectors exceeded 50
Figure 2.7: Lightweight camera system to be mounted on the freely behaving cat. Top camera measures head movements in space, bottom camera eye-position via magnification mirror.

Figure 2.8: Example frame of eye-camera, with detected eye-position overlayed.
pixels, the frame was treated as missing data. Otherwise global flow was defined as the median of all valid local flow vectors.

The clear advantage of the proposed combined eye- and head-tracking system is its robustness that makes it suitable for outdoor use. In addition, its small weight in combination with the little obstruction of the cat’s visual field, guarantees little to no interference with the cat’s natural behavior. These advantages come at the cost of comparably low temporal and spatial resolution. Furthermore, since cats - unlike primates - cannot readily be instructed to fixate certain spots, calibration of absolute eye-position is extremely cumbersome (for details how absolute eye-position calibration can be achieved in cats see (Körding et al., 2001)). Nevertheless, for the specific purpose to compare eye and head-*movements* in the freely behaving cat, the system turns out to be well suited.

### 2.3 Presentation setups

For psychophysical experiments we generated stimuli on a Macintosh G4/800 computer (Apple Inc, Cupertino, CA, USA) using MatLab including its Psychophysics toolbox extension (Brainard, 1997; Pelli, 1997). Stimuli were presented centrally on a 19” inch computer screen (Hitachi CM772E, Tokyo, Japan), which was located in 57 cm distance from the subject. The monitor’s frame rate was set to be 120 Hz. This frequency is far above the maximum human flicker fusion frequency of about 75 Hz in human vision at high eccentricities. For simultaneous presentation of overlapping stimuli we presented each stimulus every other frame only. The resulting effective rate of 60 Hz is still more than double the human flicker fusion frequency in the central visual field (about 24 Hz). Consequently the stimulus presentation can be regarded as flicker free for human observers. To keep the subject in a defined position during psychophysical experiments that did not involve eye-tracking, chin-rest and forehead-rest but no bite-bar were used.

Deviating from the setup just described, part of the data presented in chapter 6 were obtained using a different presentation setup: Stimuli were projected from the back on a transparent screen with a black circular aperture (diameter: 110 cm) using a NEC LT 157 (NEC Solutions, Inc, Itasca, IL) projector (resolution: 1024x768 pixel at 60 Hz). Two different projection sizes were used: In the "large screen" condition subjects were located at 50 cm from the screen and the image spanned 133 cm x 100 cm, which corresponds to 106° x 90° of visual angle. In the "small screen" condition subjects were located 70 cm from the screen and the image spanned 47 cm x 56 cm, corresponding to 37° x 28°. The small screen condition approximates the visual angle of a 19" monitor located in 57 cm distance that was used in previous experiments. Since in the large screen condition, part of the field of view was obstructed by the setup, for analysis we only used the inner 600x400 pixels, which were clearly visible in all conditions. As a further advantage of this restriction, potential effects of the image boundaries cannot confound the analysis. The gamma factor of the projector was corrected for, yielding a linear relation between pixel-values and luminance. As the total power of projection was unchanged between screen sizes, peak luminance ("white") on the screen was $0.24 \text{cd/m}^2$ for the large screen and $1.85 \text{cd/m}^2$ for the small screen condition respectively.

### 2.4 Colour spaces

While physically a colour can be described completely by its spectral composition, this description is inadequate for most physiological and psychophysical purposes. In photopic
vision the normal primate visual system uses there different types of receptors with different spectral selectivities (dubbed L, M and S cones for their sensitivity peaking at long, medium or short wavelengths respectively). The physical colour of a surface, i.e. the convolution of the spectrum of incoming light with the surface's reflectance, on the other hand is continuous in wavelength. For each location, the continuum describing the colour physically has to be mapped on 3 numbers, the activations of the 3 different cone types. This reduction induces ambiguities: Different physical colours may result in exactly the same physiological signal, while the same reflectance in turn can induce different signals (Wandell, 1995). Nevertheless - under natural conditions, human observers are very well capable to perceive the same reflectance as the same colour nearly independent of illumination. To achieve this constancy in colour perception, the visual system has to take additional constraints into account, such as the surrounding surfaces or a priori knowledge on natural scene statistics (Golz & MacLeod, 2002). The discrepancy between the variability in the spectral signal arriving at the retina and the "colour constancy" in perception emphasizes the need to define an appropriate colour space, when investigating the feature of colour.

Colour spaces for human colour perception may be defined on physiological or perceptual grounds. The most widely used perceptual colour spaces are those defined by the CIE (Commission Internationale de l'Eclairage - International Commission on Illumination). The starting point of their definition is based on the observation that most monochromatic colours can be perceptually matched by additive mixture of three primary colours (Young, 1802; von Helmholtz, 1852; Maxwell, 1857). Each colour is then defined by the relative contributions of those 3 primaries for an average observer (colour matching functions). Following this approach, the CIE defined the so-called CIE1931 standard. Its primaries (X,Y,Z) are defined such, that the ideal observer can match every visible colour by positive combinations of those primaries. Using the usual convention we will denote the relative contributions of the primaries X,Y and Z at each wavelength λ as x(λ), y(λ) and z(λ) respectively, the so called "colour-matching functions" (with x(λ) + y(λ) + z(λ) = 1∀λ). There are several problems attached to this CIE1931 definition. Most importantly the colour matching functions depend on the size of the stimulus. The CIE1931 is defined only for a 2° stim-

\[^{1}\text{http://www.cie.co.at/cie/}\]
This problem lead to the definition of the CIE1964 standard, that is based on a 10° stimulus. The second major issue of the CIE1931 and CIE1964 standard is that they are not perceptually uniform. This problem lead to the so-called CIE1976 standard (for reflective surfaces also called CIELAB), which is obtained by a non-linear transformation of the CIE1931 (or CIE1964) space. The resulting coordinates are representing the luminance (L) perpendicularly to a plane spanned by a “red-green” (a*) and a “yellow-blue” axes (b*). Since the CIELAB space is designed to be perceptually uniform it is well suited for simple psychophysical experiments. However the principle problem, the dependence of the defining observation conditions, namely the size of the stimulus, remains. For the experiments presented in this thesis, we thus chose a colour space that is defined on a physiological instead of a perceptual basis: the so-called DKL (for Derrington-Krauskopf-Lennie) - space (Derrington et al., 1984): Extending the work of (MacLeod & Boynton, 1979) and based on measurements of cone excitations (Smith & Pankon, 1975), DKL space is spanned by three perpendicular axis; a “luminance” axis and the two “cardinal” colour axes “constant-blue” and “constant-red&green” (figure 2.9). Along the constant-blue axis the excitation of the S cone is kept constant, while L and M vary such that their sum remains constant (L-M axis). Along the constant-red&green axis L and M cone activities are constant while S cone activities vary. The third axis defines the luminance, along which activities of all cones vary proportionally to each other. The origin of the system is chosen to be achromatic and of average (referred to the display device) luminance. In the polar coordinate system spanned by the three axes, the azimuth (φ) defines the hue, the absolute value corresponds to the saturation. 0° azimuth is chosen on the constant-blue axis in the direction of increasing L cone activation and thus corresponds to red, 90° is chosen at S cone activations larger than at the origin and thus corresponds to blue (figure 2.9).

2.4.1 Screen calibration for DKL space

In order to represent the DKL space on the a CRT-computer screen we need to map the values that corresponds to its guns’ activities to the axis of the DKL-space. As above x(λ), y(λ) and z(λ) denote the CIE1931 colour matching functions; in analogy we define r(λ), g(λ) and b(λ) as the colour matching function of the monitor’s phosphors. We denote the CIE1931 coordinates of the phosphors as x_r, y_r, x_g, y_g and x_b, y_b respectively. The z-coordinate is given by x_i + y_i + z_i = 1. Conversion from rgb to DKL space is then done in two steps; first we transform the r,g,b representation in the CIE1931 (x,y,z) representation. Then we transform the x,y,z coordinates into DKL space.

- First we need to define the origin of the colour space. This means we must find the relative phosphor activities R_E,G_E,B_E (with R_E + G_E + B_E = 1) that yield equal energy white. Where to find even energy white in the x,y,z space obviously depends on the chosen primaries; the CIE1931 standard is defined such that the equal energy white corresponds to all primaries being activated equally, i.e. x_E = y_E = z_E = 0.33. The coordinates in x,y,z space are given by the coordinates of the phosphors weighted by their relative activation:

\[ x_E = R_Ex_r + G_Ex_g + B_Ex_b \]  \[ y_E = R_Ey_r + G_Ey_g + B_Ey_b \]

The intensities on the screen, and thus the r,g,b values, normally relate non-linearly to the pixel values provided in the software r, g, b by some power function \( r = r^r, r = r^g, r = r^b \). We assume r,g,b values to be already gamma corrected, i.e. to be linear in luminance.

The procedure described here mainly follows a detailed instruction kindly provided by Sophie Würger
Solving this equation for $R_E$ and $G_E$ yields the relative phosphor activations at equal energy white $R_E$ and $G_E$; by using $R_E + G_E + B_E = 1$ we also obtain $B_E$.

- Next we need to use the transformation to map CIE coordinates to relative cone excitations (denoted here as $l, m$ and $s$). We use the transformation provided in (MacLeod & Boynton, 1979) that is based on the data of (Smith & Porkony, 1975). Denoting the transformation matrix as $T$, we obtain

$$
\begin{pmatrix}
  l(\lambda) \\
  m(\lambda) \\
  s(\lambda)
\end{pmatrix}
= T
\begin{pmatrix}
  x(\lambda) \\
  y(\lambda) \\
  z(\lambda)
\end{pmatrix}
= \begin{pmatrix}
  0.155 & 0.543 & -0.033 \\
  -0.155 & 0.457 & 0.033 \\
  0 & 0 & 0.016
\end{pmatrix}
\begin{pmatrix}
  x(\lambda) \\
  y(\lambda) \\
  z(\lambda)
\end{pmatrix}
$$

(2.4)

In particular we can compute the phosphor coordinates in the space spanned by the cone excitations:

$$
\begin{pmatrix}
  L_r \\
  M_r \\
  S_r
\end{pmatrix}
= T
\begin{pmatrix}
  x_r \\
  y_r \\
  z_r
\end{pmatrix}
= T
\begin{pmatrix}
  L_g \\
  M_g \\
  S_g
\end{pmatrix}
= T
\begin{pmatrix}
  L_b \\
  M_b \\
  S_b
\end{pmatrix}
= T
\begin{pmatrix}
  x_b \\
  y_b \\
  z_b
\end{pmatrix}
$$

(2.5)

and similar the equal energy white in cone excitation space:

$$
\begin{pmatrix}
  L_E \\
  M_E \\
  S_E
\end{pmatrix}
= T
\begin{pmatrix}
  x_E \\
  y_E \\
  z_E
\end{pmatrix}
$$

(2.6)

- We now proceed to define the cardinal axes constant-red&green (tc - for tritanopic confusion) and constant-blue (cb) of the DKL space. They will be scaled such that their end-points ($r_\| = \pm 1, c_\| = \pm 1$), the so called cardinal lights (denoted $R, G, B, Y$) fall within the gamut of the monitor. In addition the cardinal lights shall be symmetric around the origin, the equal energy white. First we compute the constant blue axis: The s cone activation is constant along this axis, hence $S_E = S_R = S_G$. Symmetry implies: $L_E - L_G = L_R - L_E$. To use the gamut of the monitor efficiently\(^4\), we choose $(G)$ to be on the axis connecting the green and blue phosphor. This implies $B_G = 1 - G_G$ and thus we obtain:

$$
L_G = G_G L_G + (1 - G_G)L_b
$$

(2.7)

$$
M_G = G_G M_G + (1 - G_G)M_b
$$

(2.8)

Since we know $s_{(G)} = s_E$ (cb axis) we can compute

$$
G_G = \frac{S_G - S_b}{S_g - S_b}
$$

(2.9)

and $B_G = 1 - G_G$. Analogously we obtain the phosphor coordinates for the cardinal colours along the tc axis, $R_G$ and $G_Y$.

- Using the phosphor activities for $R$ and $Y$ we can now compute express any colour in DKL space using the phosphor intensities $(i_R, i_G, i_E)$ by adding the luminance axis

\(^4\)This assumes that the the cb axis is restricted most by $G$ than by $R$, which is true for our monitor.
“lum”:

\[
\begin{align*}
    i_R &= lum + cb + \frac{R_E - R_g}{R_E} tc \\
    i_G &= lum + \frac{G_E - G_g}{R_E} cb + \frac{G_E - G_v}{R_E} tc \\
    i_B &= lum + \frac{R_E - R_g}{R_E} cb + tc \\
\end{align*}
\]

(2.10)

where \( B_Y = 0 \) and \( R_g = 0 \) has already been inserted. The intensities are in arbitrary units, their scaling is defined by the values the phosphors can provide. We choose to map the possible phosphor intensities to [-1,1], then equal energy light at average luminance (grey) is at \((\text{lum},t,c,b) = (0,0,0)\), white (maximal luminance) at \((1,0,0)\), black at \((-1,0,0)\), cardinal red at \((0,1,0)\) and so on. Note, that the gamut of the monitor does not fill the DKL unit sphere completely.

The CIE coordinates of the Hitachi CM772E monitor’s phosphors are provided by the manufacturer as \( x_r = 0.625, y_r = 0.34, x_g = 0.285, y_g = 0.605, x_b = 0.15, y_b = 0.065 \). This yields the following transformation matrix from DKL in phosphor coordinates:

\[
\begin{pmatrix}
    i_R \\
    i_G \\
    i_B
\end{pmatrix} = \begin{pmatrix}
    1 & 0.120 & 1 \\
    -0.461 & -0.166 & 1 \\
    0.027 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
    cb \\
    tc \\
    lum
\end{pmatrix}
\]

(2.11)

### 2.5 Human subjects

All experiments described in this thesis involving human subjects were conducted in healthy volunteers, who did receive no payment or similar reward for participation. All subjects used for eye-tracking experiments with the Dr Bous oculometer had normal vision, subjects used in other experiments had normal or corrected to normal vision. Subjects who participated in experiments involving chromatic stimuli had normal colour vision as assessed by Ishihira colour plates. All subjects gave written informed consent to participate in the respective experiments. Experiments conformed with the National (German/Swiss) and Institutional guidelines for the use of human subjects and with the Declaration of Helsinki.

### 2.6 Modeling temporal coherence

#### 2.6.1 Neuronal models

The outcome of a simulation study on a general principle such as temporal coherence is determined by three factors, the objective function used (i.e. the principle to be implemented), the stimuli presented and the network architecture, especially the neuronal model. While we restrict ourselves to a single objective function, “stability” implementation of the temporal coherence principle (section 2.6.2), the neuronal models as well as the input will vary from simulation to simulation. Nevertheless can the neuronal model be described in a general form that applies to all simulations. We use simulated neurons, whose activity \( A(t) \) is a (potentially non-linear) function \( \phi \) of the sum of activities of \( S \) subunits. Each subunit’s
activity is given as a non-linear function of the product of the subunit’s synaptic weights $W$ and the input vector $\mathbf{I}(t)$:

$$A_j = \phi \left( \sum_{x=1}^{S} \left( \sum_{i=1}^{D} W_{ij} I_i \right)^n \right)$$

(2.12)

where $D$ denotes the dimensionality of the input. Depending on the simulation, the input vector $\mathbf{I}(t)$ is either formed directly by the pixel-values of an image patch, the first principle components thereof or - in hierarchical systems - given by the activity of the previous neuronal layer. The exponent $n$ and the function $\phi$ define the non-linearity, while $S$ defines the number of subunits. In this thesis we use different types of neuronal models: linear threshold neurons ($\phi(x) = \max(x, 0)$, $n=1, S=1$), two sub-unit energy model neurons ($\phi(x) = x, n = 2, S = 2$) and neurons with multiple subunits and higher order non-linearities (usually $\phi(x) = x, n = 4, S = 8$).

### 2.6.2 The stability objective function

In this thesis we use the ‘stability’ formulation of the temporal coherence principle (Kayser et al., 2001): The squared temporal derivative of the neurons’ activity $A$ is minimized, while the trivial solution of constant zero activity is avoided by normalization by the temporal variance.

$$\Psi^{(\text{stable})} = \sum_i \psi_i^{(\text{stable})} = \sum_i \frac{\left( \frac{d}{dt} A_i \right)^2}{\text{var}_t(A_i)}$$

(2.13)

where $\langle \cdot \rangle_t$ denotes temporal averaging. The derivative is implemented as a finite difference $\frac{A_i(t) - A_i(t - \Delta t)}{\Delta t}$, where $\Delta t$ is the step size between subsequent stimuli, as defined below. $\psi_i^{(\text{stable})}$ measures the stability of a single neuron and will be referred to as “individual stability”.

The stability objective does not include interaction between neurons in the network. Consequently, maximizing $\Psi^{(\text{stable})}$ alone would lead to a population of identical neurons. An additional de-correlation term forces neurons to acquire dissimilar receptive fields:

$$\Psi^{\text{decorr}} = -\frac{\sum_i \sum_{j \neq i} \left( \sigma_{ij}^2(t) \right)}{(N-1)^2}$$

(2.14)

$$\sigma_{ij} = \frac{\left( A_i(t) - \langle A_i(t) \rangle \right) \left( A_j(t) - \langle A_j(t) \rangle \right)}{\sqrt{\text{var}_t(A_i) \cdot \text{var}_t(A_j)}}$$

denotes the coefficient of correlation and $N$ the number of neurons. The total objective is then defined as

$$\Psi = \Psi^{(\text{stable})} + \beta \Psi^{\text{decorr}}$$

(2.15)

Unless otherwise stated $\beta$ is set to 1.

Since the neuronal activity $A$ is a function of the neurons’ receptive fields $W$, $\Psi$ is maximized by adaptive gradient ascent with respect to $W$. The optimization of the stability objective function is based on a custom-designed matlab toolbox, whose first version we described in (Kayser et al., 2001) and which has been modified by several authors since then. Detailed description and source code can be obtained at http://www.koerding.com/software.html.
Chapter 3

The Spatio-temporal Statistics of Natural Scenes

Investigating the processing of Natural Stimuli requires a sound knowledge of their statistical properties. In this chapter we will analyse the spatio-temporal structure of the cat-cam movies (section 2.1.1) and link them to known physiological effects.

3.1 Specific methods

3.1.1 Stimuli

Stimuli used in this chapter are based on the videos recorded by a camera mounted to a freely behaving cat as described in section 2.1.1. Apart from section 3.2.5, we ignore chromaticity and convert images to greyscale using the default mapping provided by matlab’s rgb2gray.m function. We use only the central 200x200 pixels, where the imaging properties of the camera are homogeneous, for further analysis. Throughout this chapter the term ’image’ will refers to this central region. For the calculation images are low-pass filtered by convolution with a Gaussian at a cutoff frequency of the transfer function of the CCD camera (0.53\frac{\text{mm}}{\text{deg}}). This is lower than the spatial acuity at the fixation spot, but of similar order of magnitude at moderate eccentricities (Pasternak & Horn, 1991). Contrast is then defined as the standard deviation of intensity within a 20x20 pixel (4.6 x 4.6 degrees) patch divided by the mean of intensity in the image.

3.1.2 Spectral analysis

For the spatial spectral analysis of each frame we use standard methods (Papoulis, 1991). The mean intensity of each image is subtracted. To reduce boundary effects and to avoid any orientation bias, images are pointwise multiplied with a 2-D isotropic Gaussian-window of standard deviation 70 pixels. To facilitate the subsequent Fourier-transform, the resulting image is zeropadded to size 256x256, denoted as I. J denotes the set of all pixels of I; brackets ⟨,⟩ denote averaging. The Fourier transform is conducted by the 2D-Fast Fourier Transform function of MatLab (fft2.m) and the 2D-power-spectrum computed as

\[ P_k = \left| \sum_{r \in J} I_r e^{-\frac{2\pi i k r}{256}} \right|^2 \]  

(3.1)
\( k \) denotes the 2D spatial frequency. To explore scaling behavior, we fit the power spectrum along the horizontal \( (k = \begin{pmatrix} k \\ 0 \end{pmatrix}) \) and vertical \( (k = \begin{pmatrix} 0 \\ k \end{pmatrix}) \) axis with the function \( f(k) = \left( \frac{1}{k} \right)^{2-\eta} \) in the range of spatial frequencies up to a quarter of the spatial Nyquist frequency.

### 3.1.3 Wavelet analysis

For wavelet analysis we use complex-valued 2D Gabor-wavelets, which are corrected for zero mean with a smooth decline towards 0. Formulated for the complete pixelset \( J \), the wavelet corresponding to frequency \( \omega \) and center at \( r_0 \) has consequently the following functional form:

\[
G^{(\omega,\phi, r_0)}_r = e^{-\frac{(r-r_0)^2}{\sigma^2}} e^{-i(r-r_0) + \frac{\pi}{2}} \sum_{r \in J} e^{-\frac{(r-r_0)^2}{\sigma^2}} e^{-i(r-r_0) + \frac{\pi}{2}}
\]

where \( \omega = \begin{pmatrix} \sin \phi \\ \cos \phi \end{pmatrix} \) is the spatial frequency vector. The representation of the dependence of the vector \( \omega \) on \( \phi \) (\( \omega = \omega(\phi) \)) is omitted for readability throughout the paper. The angle \( \phi \) runs over 6 different angles, \( \omega \) over 4 different spatial scales: \( \phi \in \Phi = \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ\} \) and \( \omega \in \{\omega_n = \sqrt{2}^{(n-5)} \text{pix}^{-1} | 0 \leq n \leq 3\} \). The width of the Gaussian envelope is chosen dependent on \( \omega \) to yield constant bandwidth: \( \sigma = 2.5\omega^{-1} \). For computational efficiency a complete convolution of the Gabor with the image is avoided by multiplying \( G^{(\omega,\phi, r_0)} \) with \( I \) only at specific points, which form an evenly spaced grid \( \Gamma \) and are separated by 16 pixels along each dimension (for \( \omega_2 \) we chose a slightly denser grid with 14 pixel distance):

\[
\Gamma = \left\{ r | r = \frac{16n}{16m} | n, m \in \{1, \ldots, 11\} \right\}
\]

or

\[
\Gamma = \left\{ r | r = \frac{14n}{14m} | n, m \in \{1, \ldots, 11\} \right\}
\]

respectively. We denote the amplitude of the product as \( L^{(\omega,\phi, r_0)} \):

\[
L^{(\omega,\phi, r_0)} = \left| \sum_{r \in J} G^{(\omega,\phi, r_0)}_r I_r \right|
\]

referred to as "local amplitude" and the phase

\[
\Lambda^{(\omega,\phi, r_0)} = \text{phase} \left( \sum_{r \in J} G^{(\omega,\phi, r_0)}_r I_r \right)
\]

referred to as "local phase".

The mean wavelet amplitudes are computed by averaging over time and/or space as described above. To examine the significance level of differences of mean wavelet amplitudes a paired sign test is used. To obtain the incidences of the dominant orientation we calculate for each grid point \( r_0 \in \Gamma \) and time the \( \phi \in \Phi \) for which the local amplitude \( L^{(\omega,\phi, r_0)} \) reaches its maximum:

\[
\phi^{(\omega, r_0)}_D = \arg \max_{\phi \in \Phi} L^{(\omega,\phi, r_0)}
\]
3.1.4 Correlation analysis

The spatio-temporal auto-correlations \( X_L(\omega, \phi, r_0) (\delta, \tau) \) and \( X_A(\omega, \phi, r_0) (\delta, \tau) \) of local amplitude and local phase respectively are calculated according to standard algorithms using MatLab’s \textit{xcorr.m} function. \( \delta, \tau \) denote the spatial and temporal lag; phases are unwrapped using MatLab’s \textit{unwrap.m} and amplitudes subtractively normalized to zero mean before computing the correlation.

3.2 Results

3.2.1 Spatial distribution of luminance and contrast

To get a first impression on the average image structure, we analyse the mean luminance distribution over all images. We find a vertical intensity gradient from bottom to top, ranging from about 40\% to 70\% of the luminance range. In horizontal direction no such clear gradient is found (figure 3.1). On average the top part of the image is thus found to be brighter, which is likely to be a consequence of the fact, that part of the sky is visible in most images. A recent study reports a 20\% higher contrast at fixation points selected by humans scanning still images (Reinagel & Zador, 1999). While we will address this issue for human and non-human primates in detail in chapter 6 and will also adress eye-movements in freely behaving
cats (chapter 5), we here exclusively use the video material of the front camera to achieve a first qualitative estimate of whether or not contrast also affects fixation behaviour also in cats. Thereby we can make use of a natural stimuli database, which is much larger than the database that includes simultaneous acquisition of eye-movements. Under the assumption that the head-direction in the vast majority of frames corresponds to the gaze direction of the animal (Guitton et al., 1984), the contrast at the center of the image provides a good estimate of the contrast the animal is fixating. We then compare this central contrast to the contrast at higher eccentricities (figure 3.2). Even using the assumption that subsequent frames are independent, only three movies indeed show significantly ($p < 0.05$, t-test) higher contrast values in the central part of the image. In one example (figure 3.2, left) we find a 24% higher contrast in the central image area. A frame taken from this movie illustrates this contrast distribution (Figure 2.2, left): a bright sky and a dark ground form the low-contrast image regions, whereas most high-contrast objects are located in the middle image regions. Other movies, however, show a different contrast distribution. In one movie taken during a walk through the woods for example, the lower image-region has approximately 28% higher contrast values compared to the upper half (Figure 3.2, middle). A frame of this movie (Figure 2.2, right) depicts a low-contrast sky, whereas $\frac{2}{3}$ of the lower image are filled with high-contrast features like branches, leaves and brushwood. The average contrast over the complete set of movies is shown in Figure 3.2 (right) and the spatial distribution of its coefficient of variation ($CV = \frac{std}{min}$) is illustrated in Figure 3.3. Indeed the average contrast is not homogeneously distributed over the whole region. The top margin of the image shows a band of low-contrast, whereas the $\frac{2}{3}$ forming the lower half of the image shows about 16% higher contrast values. The central image area, however, does not reveal a significantly higher contrast profile. The contrast for the central 8.8 degrees deviates from the mean by only 3.8%. Concluding, in contrast to the study of human vision, here, using cats exploring a natural environment, we could not observe a bias of fixating high contrast regions.

3.2.2 Contours at the cardinal axes

How is the distribution of oriented contours encountered by a freely behaving animal? The analysis of the stimulus sequences reveals that the local amplitude of horizontal features for the standard spatial frequency of $\omega_2$ ($\{L^{\omega_2, \phi=90^\circ, r_0}\}_{r_0 \in G}$, see section 3.1.3) is significantly higher than the average of all orientations (6.79%, $p < 0.001$, paired sign test). For the other spatial frequencies we obtain values of 15.24%, 11.07% and 4.70% ($\omega_0, \omega_2, \omega_3$).

This result can be illustrated by complementing the wavelet analysis with a standard frequency analysis (Figure 3.4). The shape of the power spectrum is anisotropic and the power is highest at horizontal orientations. When plotted on a log-log-scale (figure 3.5) the falloff

\[
CV \text{ of contrast}
\]

\[
\begin{array}{ccc}
& 0.42 & \\
0.20 & &
\end{array}
\]

Figure 3.3: Standard error of the mean contrast distribution shown in Figure 3.2.
Figure 3.4: The power spectrum averaged over all movies. Maximal power is found in the center at low spatial frequencies.

Figure 3.5: The two main axes of figure 3.4 are plotted on a log-log-axis. The gray shading indicates the region unaffected by the low pass filter, which is used for the fit. Both data sets fall approximately on a straight line and are well fitted to the $\frac{1}{k^{2-\eta}}$ model (see methods) with an $\eta$ of -0.10 for the horizontal and 0.16 for the vertical orientation. This behavior is described by the "Power Law" (Field, 1987; Burton & Moorhead, 1987; Tolhurst et al., 1992; van der Schaaf & van Hateren, 1996) that is compatible with scale invariance. This scaling invariance seems to be a general property of natural scenes and has been implicated to be reflected in the properties of the mammalian visual system (Hughes, 1977).
Figure 3.6: The mean wavelet amplitude is demonstrated (red) for the 6 different orientations (0° denotes the vertical orientation) averaged over the complete database. The abscissa shows the orientation of the image region ranging from 0 to 150°. The incidences of the dominant orientation (green) are illustrated for the 6 different orientations. The mean fMRI response amplitude in V1 (blue) plotted as a function of orientation (data replotted from Furmanski and Engel, 2000, Fig. 1c).

Figure 3.7: The mean wavelet amplitude in all movies, averaged over columns is plotted colour coded as a function of height in the image (vertical axis) and of the orientation of the wavelet (horizontal axis).
in power for the higher spatial frequencies along the 2 main axes is clearly visible. More energy pertains to the horizontal orientation (green) than to the vertical orientation (red) by a factor of 1.17. Also note that the fit is not perfect at low frequencies, indicating a limited range of scale invariance of natural images.

To ensure that the dominance of horizontal contours observed in the wavelet analysis is not a result of a setup artifact, we performed the following control: we mounted the camera to a pivotable tripod and took images using the camera in normal orientation and rotated by 90 degrees within the image plane. The data acquisition and processing was done exactly identical to the original setup, including recording the camera signal to the VCR and digitizing offline from tape. For the "normal" image we found (total over all used omega) in 22.6% of the contours to be dominantly horizontal, for the rotated image 23.5% to be dominantly vertical (i.e. horizontal in the real world). These numbers closely match the result for the whole image set and thus exclude that the result could be a consequence of an anisotropy of the camera setup.

To control that the visibility of the horizon does not produce the dominance of horizontal orientations we perform 3 controls. Firstly we calculate the incidences of the dominant orientation \( \phi^{(o,\omega)} \) (figure 3.6, green). In contrast to above we are not calculating the mean power at each orientation, but determine at each location the dominant orientation and compare the statistics of this distribution: In 24.1% of all local regions (i.e. all grid points \( r_0 \in \Gamma \) at spatial frequency \( \omega_2 \) ) analyzed the amplitude of the horizontal wavelet is largest. This is 58.6% higher than the mean of incidence of the other orientations being dominant. Thus, large image areas are affected by the horizontal feature predominance and as the horizon contributes only to a limited area within each image it cannot fully explain this effect. Secondly, we perform the analysis on a subset of movies that do not contain the horizon and obtain comparable results (data not shown). Thirdly, we investigate the spatial distribution of the mean wavelet amplitudes within the images. We are interested in the dependence of prevalent orientations along the vertical and horizontal axis. Averaging the wavelet amplitudes over image rows and orientations results in one value per column. These values have a standard deviation that is small compared to their mean (0.10 \( \ll \) 3.21). Thus, the wavelet amplitudes show little variation in the horizontal direction. Instead, we observe a strong variation of wavelet amplitudes along the vertical axis of the image (figure 3.7).

The average amplitude of all orientations of the upper image half is 7.48% smaller than in the lower part of the image. This is a natural consequence of the spatial distribution of contrast reported above. Furthermore, the average amplitude for the horizontal orientation in the upper third of the image is even 8.64% smaller than in the lower two thirds of the image. Thus the lower image regions contribute most to the predominance of the horizontal orientation, the upper regions of the image contribute most vertical orientations. Upright trees against a bright sky (figure 2.2, right), for example, contribute to this effect. This third controls shows that the horizon may contribute to some extent to the preponderance of horizontal orientations, but it does not explain the full effect. Taken together, these three controls show that the predominance of horizontal contours is a general property of the stimuli the animal experiences in its natural habitat.

This phenomenon of a predominance of horizontal and vertical orientations has been described as the oblique effect. It is not only exemplified in statistics on natural images but has also been described in anatomical and physiological studies on the visual processing. In figure 3.7 (blue line) we demonstrate this analogy in particular for data on fMRI (Furmanski & Engel, 2000). The study shows a significantly greater behavioral sensitivity with larger responses in primary visual cortex to stimuli of cardinal orientation (horizontal and vertical) as opposed to oblique features. It also demonstrates that the signal induced by the
Figure 3.8: The correlation of orientation amplitude over time at the same pixel. Squares: 0°, stars: 90°, dashed: all other orientations (spaced 22.5°)

Figure 3.9: Correlations of different combinations of orientations and different spatial arrangements of the two points in the same frame. The orientations are (from top to bottom and left to right): 0°, 135°, 90°, 45°

two cardinal orientations is comparable. The former aspect matches our data on natural images. With respect to the latter a difference to the results of this study is noticeable.

3.2.3 Temporal correlations

First we investigate temporal correlations at the same point in space. Figure 3.8 demonstrates that if an orientation is present at one point in time then the amplitude of this orientation in the next frames at the same point is also likely to be high. Temporal correlations are strongest for the cardinal orientations, i.e. horizontal and vertical. For the other orientations correlations decay faster but are still significant over several hundreds of milliseconds (decay time constants for 0°: > 1 s, 45°: 490 ms, 90°: 900 ms, 135°: 360 ms). Thus the presence of an oriented segment gives a strong prediction for the orientation at the same point later in time.

The Gestalt law of "good continuation" has been proposed as the underlying principle of feature binding. This inspires us to investigate the relation of orientations in neighboring image locations (Krüger, 1998; Krüger & Wörgötter, 2002). We analyze the spatial aspect
of the spatio-temporal autocorrelation of wavelet-amounts for different image locations. Hence we analyse the two dimensional spatial distribution of correlations as well as correlations of different orientations. Figure 3.9 shows the correlations between segments of 4 different orientations (0°, 45°, 90°, 135°) situated at different relative locations in the same frame. Iso-orientation correlations (panels on the diagonal) are stronger than cross-orientation correlations. We compare this result with recent anatomical studies on the connectivity in primary visual cortex (Bosking et al., 1997). These authors report a 1.53 larger range (median of maximum distance at which boutons were found) of tangential connections along the axis of orientation preference as compared to tangential connections in the orthogonal direction. This number matches fairly well to our results. In this respect we were interested if we could observe different correlation values along the collinearly displaced image regions when comparing different orientations. Indeed, the contour lines of the iso-orientation correlations are elongated along the direction of the particular orientation. This shows that collinear structures are more prevalent than parallel shifted contours. Also parallel contours occur more likely than T-junctions since the iso-orientation correlations are at all points stronger than the correlations of this orientation with the orthogonal. An example of how the spatial correlations decay independently of the spatial direction is shown in figure 3.10 for the horizontal (90°) orientation. Correlations decay fastest during the first 2 degrees of spatial distance but extend well up to 8 degrees.

Our dataset allows analyzing how these spatial correlations evolve over time. Figure 3.11 shows the same data as in figure 3.9 but for segments 400 ms apart in time. The spatial arrangement of correlations is the same as for zero time lag but the amplitudes decayed by a factor higher than 2. For the cardinal orientations again collinear interactions are prevalent. This is in agreement with figure 3.8 which shows that these orientations are very stable over time. Since the oblique orientations are less well correlated over time we would expect that collinearity will here be less prominent for larger time lags. Indeed the contour lines of the correlations for the oblique orientations are more circular symmetric. To quantify these changes over time we measure the aspect ratio \( \frac{\text{length}}{\text{width}} \) of the contour lines for the different time lags. Collinearity means a high aspect ratio and a loss of collinearity therefore is accompanied with a decrease in aspect ratio. Using this measure, figure 3.12 shows that collinearity is preserved over long temporal lags and is strongest for the cardinal orientations.

To quantify the change in amplitude of the spatial correlations in a different way, we define areas of strong interactions by thresholding correlations. We chose a threshold of 0.4
Figure 3.11: Correlations of different combinations of orientations and different spatial arrangements of the two points in frames separated by 400ms. Orientations as in figure 3.9

Figure 3.12: Relative change of the aspect ratio of the correlation contours in figure 3.10 as a function of time. Shown is the aspect ratio at each point in time divided by the aspect ratio at t=0. Squares: 0°; stars: 90°; solid: 45°, dashed: 135°
Figure 3.13: Areas of strong correlations. We defined spatio temporal separations with a correlation over 0.4 as strong. The figure shows these areas for the correlation diagram of figure 3.10.

Figure 3.14: Size of the areas of strong correlations relative to the total patch size over time. Squares: 0°, stars: 90°, solid: 45°, dashed: 135°.
Figure 3.15: (A) For an example orientation (45°) we show the mean (left) over 12 video sequences together with the standard deviation (right). (B) The mean plus/minus the standard deviation.

Figure 3.16: Cross orientation correlations over space for filters of a higher spatial frequency.

to ensure that even for zero time-lag only iso-orientation correlations exceed this threshold (figure 3.13). As expected the decay times are slowest for the cardinal orientations but independent of the orientation there exist points with strong correlations for at least 280 ms (Figure 3.14).

We performed controls to see how these results depend on the amount of data used. The above data were averaged over our whole data-base. Since one feature of our video sequences is their variety in terms of landscapes etc. we look at the differences between different sequences. In figure 3.15 we show the correlations for one oblique orientation (135°). The mean and standard deviation over 12 video sequences is shown in Figure 3.15. The error is rather small compared to the correlation values. More importantly, the correlation surface plus minus the error (figure 3.15B) shows the same spatial structure as the mean. Also, the distinct pattern of correlations is visible in averages over shorter sequences (data not shown). Thus the distinct patterns of spatial correlations are not introduced by averaging over a large data set.

As a further control, we use filters of a different spatial scale and frequency to measure the orientation content. The filters used for figure 3.16 are twice as large as the ones used for the other experiments. The results are basically the same as with the lower frequency filters.
Again collinearity is most prevalent. Therefore our results generalize over a wide range of filter parameters.

### 3.2.4 Time constants of orientation and phase

In order to compare the correlation time constants of different features, we average the spatio-temporal autocorrelation function of wavelet amplitude and phase at zero spatial lag over orientations \( \langle X_L(\omega, \phi, r_0) (\delta = 0, \tau) \rangle_{\delta \in \phi} \) and \( \langle X_L(\omega, \phi, r_0) (\delta = 0, \tau) \rangle_{\delta \in \phi} \) respectively. Figure 3.17 shows the decay of the correlation functions of amplitudes and phases over time. As expected from the results above, the autocorrelation function of the local amplitude drops to 10% after 1.4s. In comparison the auto-correlation function of the phase reaches this level already after 160 ms. This demonstrates that the orientation of local structures changes much slower than their position. In chapter 4 this difference of correlation time constants will enable the temporal coherence objective to segregate orientation from phase: responses will get invariant to the fast feature (phase) and specific to the slow one (orientation), leading to the emergence of complex cell type receptive field. Consequently, the improved understanding of temporal image properties will foster our understanding of coding principles in the visual system.

### 3.2.5 Correlations of colour

Besides local phase and orientation, colour is the arguably most basic feature of a natural scene. Since the camera used for recording the catcam videos is far from having a well-defined luminance independent spectral sensitivity, only a coarse definition of colour is available here. As for the modeling experiments we therefore transform the rgb-representation of the video to a representation based on hue, saturation and brightness using the default mapping of matlab’s rgb2hsv.m. We then investigate use the hue in analogy to the azimuth of the DKL-space representation, that we use for psychophysical experiments (section 2.4)

\[1\] The measured correlation time constant for hue is 14.3s, about an order of magnitude larger

\[1\] Although for our psychophysical experiments the true input colour is also unknown, we in that case can precisely calibrate our output device to faithfully represent colour as defined in DKL space.
different orientations. Coherence (green): The average coupling strength is illustrated between pairs of cortical recording sites as a function of the relative orientation preference of the neurons. The coupling strength is quantified as a spectral coherence (data replotted from (Frien & Eckhorn, 2000)). Bouton distribution (red): The number of labeled terminals found in the sector surrounding an injection site is plotted as a function of the preferred site (data replotted from (Bosking et al., 1997)).

than the 1.4s of orientation. In chapter 4 we will also use this result to segregate colour and orientation on the basis of their different temporal coherence.

3.2.6 Interaction between different orientations

The relation between the statistics of natural visual stimuli, the anatomical structure of tangential connections in cortex and interaction of orientation tuned neurons is illustrated in Figure 3.18. Hereby we juxtapose the results of recent anatomical and physiological studies to our data. The correlation of the wavelet amplitude at neighboring image regions declines with increasing difference in orientation tuning. Interestingly, a similar phenomenon can be observed in anatomical results on the bouton distribution, which is reflected in the number and in the distance of the labeled terminals (Figure 3.18, red line, data from (Bosking et al., 1997)). Hereby the relationship between the axis of elongation of the terminal distribution and the preferred stimulus orientation of the injection site has been quantified. Labeled boutons were consistently found to extend greater distances from the injection site along the axis of visual field that corresponds to their preferred stimulus orientation (Bosking et al., 1997; Schmidt et al., 1997). A similar dependency on the stimulus orientation could be exemplified in a physiological study measuring the coherence (Figure 3.18, green line, data from (Frien & Eckhorn, 2000)) among recording pairs in striate cortex of awake monkeys. Similar results have been obtained in anesthetized cats (Engel et al., 1991): the more the orientations were similar, the higher was the observed coherence. Thus we can observe that both anatomical results and physiological data match the distribution and correlation of local image features in our natural videos.

3.3 Discussion

In this chapter we have cats explore their environment to collect natural visual stimuli. The statistical analysis of this large database allows studying properties of natural time varying stimuli. Results from previous studies on still images, which report "scaling behavior" (Field, 1987; Hancock et al., 1992; Tolhurst et al., 1992; Ruderman & Bialek, 1994) and the Power-Law (Ruderman & Bialek, 1994; van der Schaaf & van Hateren, 1996) are also demonstrated for our data-set. Analyzing the spatial distribution of contrast, we do not find an increased contrast in the presumed direction of gaze. The precise position of contours changes more rapidly than their orientation. Finally, image features of the cardinal orientation or spatially correlated with collinear contours are more prevalent than parallel shifted contours.
3.3.1 Contrast at the fixation spot

Previous studies on human eye movements (Reinagel & Zador, 1999) demonstrated a higher contrast at the fixation points compared to randomly chosen parts of static pictures. Therefore, one might expect a higher contrast at the positions where the cats are looking at. However, the variance of contrast over movies taken in different environments outweighs the difference between the central and peripheral regions. Whereas three movies reveal a higher contrast in the central image area, movies recorded in other surroundings show a drastically different contrast distribution. Thus, this aspect of the image statistics strongly depends on the explored environment. However, comparing this result with the results of (Reinagel & Zador, 1999) several differences in subjects and experimental setups are evident: In the present study we assume the central part of the image to represent the point the cat is actually fixating. We control this by examining whether objects, the cat is fixating on from various distances, are represented in the central image area. Furthermore, data we will present below (chapter 5) will indicate that freely behaving cats make mostly eye movements of small amplitude. Small eye movements, however, can potentially smooth out a contrast peak at the center of the image and we would not detect this maximum in the statistics any more. Furthermore, allowing head movements, makes correcting "head saccades" very effective (Guitton et al., 1984). Thus, it seems reasonable to assume that the animals primarily gaze into the direction of the center of the image.

The most important difference, however, is the specific anatomy of the feline as compared to the primate eye. The small fovea endows the primate visual system with a very high spatial acuity. In comparison, the region of the best visual acuity of the cat, the area centralis, is much larger than the fovea and its spatial acuity is an order of magnitude lower (Blake, 1988). The need to focus on high spatial detail might thus be much reduced in cats as compared to primates. Defining a cat’s gaze direction in perceptual terms to the same angular precision as that of primates therefore seems questionable in itself. Consequently the coarse center vs. periphery analysis performed here, might capture the essence of feline eye-movement behaviour equally well as a highly precisely laboratory setup, while coming at the advantage of large datasets acquired in a freely behaving animal.

3.3.2 The oblique effect

The data analysis showed a predominance of cardinal over oblique contours, with a further preference of horizontal over vertical orientations. The oblique effect is well known in the psychophysical and physiological literature. However, the number of previous publications on this issue, using natural image material is surprisingly small (Switkes et al., 1978; Hancock et al., 1992; van der Schaaf & van Hateren, 1996; Coppola, Purves, et al., 1998; Kell & Cristobal, 2000). Those studies explore the oblique effect though, mainly examine human selected natural visual scenes. However, the interest in the ‘oblique effect’ provoked many anatomical and physiological studies, extensively illuminating the orientation selectivity in different organisms and experimental conditions. The major consensus is that both cardinal oriented contours are more abundant than oblique contours. Psychophysical studies could demonstrate a better visual performance regarding visual acuity and contrast sensitivity when using cardinal oriented stimuli (Boltz et al., 1997; Krebs et al., 2000). Single-unit recordings in the primary visual cortex demonstrated higher response properties for stimuli at the cardinal axes, using cats (Pettigrew et al., 1968; Fregnac & Imbert, 1978; Orban & Kennedy, 1981; Leventhal & Schall, 1983) or monkeys (Mansfield & Ronner, 1978; Blakemore et al., 1981). Indeed, a recent study based on a large database of recorded cells confirms the larger fraction of neurons responding to the cardinal axes, and furthermore provides
evidence that horizontal contours activate more neurons than do vertical contours (Li et al., 2003). This is in line with the results reported in the present study. Stimuli at the cardinal orientations produced a larger cortical representation in optical imaging studies in ferrets (Chapman & Bonhoeffer, 1998; Coppola, White et al., 1998) and cats (Dragoi et al., 2001). Similarly, they induced a higher performance (faster latency, larger amplitude) when evoked potentials were measured (Arakawa et al., 2000). The cardinal effect is not purely 'cortical' but could also be demonstrated for cortical sites as the retina and the LGN, whereby retinal ganglion cells and their dendrites are preferentially arranged along the vertical and horizontal meridians (Wässele et al., 1975). A further differentiation has been described by applying stimuli with different spatial frequencies. It was shown that neurons in the Macaque visual cortex that are tuned to high spatial frequency preferred the cardinal orientations (de Valois et al., 1982), whereas another study applying middle to low spatial frequencies even described a preference for oblique orientations (Nelson et al., 1984).

The data presented here show that when our cats explore a natural environment, horizontal contours dominate at all spatial frequencies investigated. Thus, results on the increased representation of cardinal orientations with an emphasis on horizontal contours in primary visual cortex (Coppola, Purves, et al., 1998; Li et al., 2003) might be a natural consequence of the statistics of visual stimuli. This argument would favor the assumption that the natural visual input has an instructive role towards the orientation preference of cortical neurons (Sengpiel et al., 1999). In hand with this idea go single-unit-studies examining the effect of offering a single visual orientation during early visual experience and demonstrating a dominance of neurons responsive to that orientation (Hirsch & Spinelli, 1970). Apparently the selection of the visual input results in an overrepresentation of the experienced image features. Thus we would suggest that the statistics of our natural visual images reflect the structure and processing of the visual system.

### 3.3.3 Stimulus statistics and neuronal properties

When we analyse the temporal properties of natural scenes, we find that local positions (on the scale of receptive field sizes) of oriented contours change much faster than their orientation. Retrospectively, this property of natural stimuli can be understood as consequence that even pure rotations, within a small aperture far away from the center of rotation, locally generate rather a translation than a rotation. In chapter 4 we will relate this general property of natural scenes to the invariant response properties of complex cells in primary visual cortex (Hubel & Wiesel, 1962). In connection with the results presented there and with other modeling studies that also relate the properties of natural scenes to neuronal responses (Law & Cooper, 1994; Dong & Attick, 1995b; Olshausen & Field, 1996; Dan et al., 1996; Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998; Hyvärinen & Hoyer, 2000), a better knowledge of natural scene statistics is likely to greatly contribute to a better understanding of neuronal coding.

### 3.3.4 Anisotropy of lateral interactions

Being interested in what extent neighboring image-regions refer to each other, we found that image regions lying on the same axis as the examined orientation are more strongly correlated. Thus, collinear features are more prevalent than other combinations. This is in agreement with previous results on the statistics of human selected still images (Krüger, 1998). Furthermore, it relates closely to the connectivity of primary visual cortex. Recent anatomical studies on the topography of lateral connections and orientation maps (Gilbert & Wiesel, 1989; Kisvarday et al., 1997; Schmidt et al., 1997) show that intra-areal projec-
tions in cat visual cortex preferentially connect regions of similar orientation preference and regions with a collinear alignment of preferred orientations. Similarly were the results in psychophysical studies reporting a lower detection thresholds of aligned Gabor patches (Gilbert & Wiesel, 1990; Hess & Dakin, 1997; Polat & Norcia 1998; Keeble & Hess, 1999) with larger response amplitudes and higher contrast sensitivity. Likewise, physiological experiments demonstrate a dependency of neuronal activity and of synchronization (Engel et al., 1991; Gray et al., 1989; Löwel S & Singer, 1992; König et al., 1993; Frien & Eckhorn, 2000) on the alignment of visual stimuli. Furthermore, we observe that the anatomical results on the bouton distribution (Bosking et al., 1997) and the physiological data on coherence (Frien & Eckhorn, 2000) match the distribution and correlation of local image features in our natural time series. Like it has been suggested for the oblique effect we would suggest that the cortical anatomy and processing reflects the natural visual input experienced during the formation of 'cortical wiring'. When observing that co-axial image regions on the cardinal axes are stronger correlated over time than those on the oblique axes and keeping in mind its anatomical and physiological correspondent/analogy, we would suggest that this anisotropy could be related to the stability on the processing of cardinal orientations during visual perception as proposed in earlier studies (Dragoi et al., 2001).
Chapter 4

Temporal Coherence: From V1 to object recognition

Despite the specialization of areas and pathways within the visual system, recent theoretical studies raise the hope, that many of its functional properties may be derived from few general coding principles. In this chapter we will analyse our implementation of the temporal coherence principle - the so-called “stability” objective. We will first show that simulated neurons trained to optimize the stability objective resemble several physiological properties of real complex cells (section 4.1). We then show how the functional segregation of colour and orientation can be derived from the same principle applied on natural colour movies (section 4.2). Using a two layer network, in which we repeatedly apply the temporal coherence principle on natural movies, we obtain cells, whose specificities match the classification of naïve human observers (section 4.3). Applying the same scheme to a standard computer vision database, we show that “stability” leads to a representation of real-world objects that facilitates view-point invariant object recognition (section 4.4). In summary, this chapter shows how the same implementation of a simple principle - temporal coherence - might explain invariant response properties throughout the visual hierarchy.

4.1 Complex cells

4.1.1 Specific methods

Stimuli

For the experiment described in this section, we use the grayscale converted version of the catcam videos described in section 2.1.1. Ideally we would like to take a single long sequence from the central region of the video. Such a sequence, however, would need to be prohibitively long to uniformly sample the stimulus material. This is why we instead take pairs of patches measuring 30 x 30 pixels from randomly selected but matching locations within two subsequent frames in the movie. Temporal coherence is evaluated between the patches of the same pair, approximating the optimal sampling process. The patches are first multiplied pointwise with a Gaussian kernel centered over the patch the SD (width) of which was 10 pixels. This procedure has a limited effect on the amount of information available in the input stream but avoids edge effects and the anisotropy inherent in square patches. Repeating the simulations below without this windowing leads to qualitatively similar results.
CHAPTER 4. TEMPORAL COHERENCE

(data not shown). The receptive field obtained in such simulations are localized, do not cover the full patch, and are approximately round too. The resulting patches are decomposed into their principal components. The first component, representing the mean patch brightness, is removed. Components 2 to 100 carry > 95% of the variance and define the input vector I to the simulated neurons (equation 2.12). As the activity of each subunit linearly depends on the input, the preprocessing of the input by a principal component analysis, which is also linear transformation, has no influence on the optimization process. Discarding the higher-order components, however, is an non-linear process and thus could potentially have an effect. As these components carry only a small part of the total variance, however, we do not expect an influence of on the results obtained. Indeed, this assumption is supported by the results of one of our preliminary studies (Kayser et al., 2001).

Neuronal model

Complex cells, in contrast to simple cells, display several strong nonlinear properties (Chance et al., 1999; Movshon et al., 1978; Ohzawa et al., 1997; Spitzer & Hochstein, 1988). Hence, it is not possible to describe them adequately by linear models, and we have to consider non-linear model neurons. Identical to the choice in a number of other studies e.g. (Hyvärinen & Hoyer, 2000) we chose the two subunit energy model (Adelson & Bergen, 1985) by setting the parameters in equation 2.12 to \( \phi(x) = x, n = 2, S = 2 \). For the experiment described in the current section we use \( N=100 \) neurons.

These two subunit energy model neurons can, given appropriate weights, exhibit a large variety of response properties. Most of these properties are never observed for real neurons. The simulated neurons can, however, also act like a complex cell if both subunits have Gabor-wavelet-like receptive fields with identical orientation and spatial frequency, and the two wavelets have a relative phase-shifting of 90°. If such a neuron is excited by a visual stimulus in form of a bar that is moved over its receptive field, each subunit has an activity that depends on the bar’s position. As the bar is shifted, the subunits alternate in having large squared activity. Thus the neuron’s activity, the sum of the squared subunits activities, changes only little as the bar is moved within the receptive field. It should be noted, however, that - given the large number of parameters (twice the length of the weight vector, \( S*D \) in equation 2.12) involved in determining the response properties of these model neurons, such complex cell like properties are only one among many other conceivable outcomes.

Data analysis

In analogy to physiological experiments, we characterize the response properties of the model neurons by several indices. The orientation tuning width is calculated as the range of orientations for which the response to a bar of optimal position is above \( \frac{1}{\sqrt{2}} \) of the maximal activity. The best orientation is defined as the stimulus orientation that leads to maximal responses. The selectivity for spatial frequency is defined via the range of spatial frequencies to which the response exceeds \( \frac{1}{\sqrt{2}} \) of the maximal level (Schiller et al., 1976b). The difference between the lower and upper bound of this range is then multiplied by 100. We measure the responses of neurons to drifting sinusoidal gratings of optimal orientation and spatial frequency. The neurons AC/DC ratio is the maximum minus the minimum divided by the mean of the resulting activity.

The models that are used for the modeling of complex cells, such as the two subunit energy model used here, always respond to moving gratings with twice the temporal frequency of the moving grating as they respond equally well to bright and dark edges. This implies that the simulated neurons have a vanishing first harmonic (F1) while the second harmonic
(F2) does not vanish. Real complex cells, however, show such frequency doubling only to a limited degree, and both components are small (Heeger, 1992; Spitzer & Hochstein, 1988). How should the AC/DC ratios of such simulated neurons be compared with the relative modulation of real neurons? Either we could compare the AC/DC ratio to the F2/F0 ratio of real neurons, assuming that the frequency doubling is just an artifact of the simulation method. Alternatively we could compare the AC/DC ratio of the simulated neurons to the F1 of the real neurons; this is the preferable method to distinguish complex cells from simple cells. In the scenario followed in this paper, the simulated neurons should have small AC/DC ratio compared with the relative modulation of real neurons. The envelope of the receptive field is defined as: \( E_i(x, y) = W_{i,1}(x, y)^2 + W_{i,2}(x, y)^2 \). The length \( L_i \) and width \( V_i \) (defined via the SDs) of the receptive field is calculated (using the abbreviation \([.]_+ = \max(., 0)\)):

\[
L_i = \sqrt{\sum_{x,y} \left( x \sin(\phi) + y \cos(\phi) \right)^2 \left[ E_i - 0.5 \text{std} (E_i) \right]_+}
\]

\[
V_i = \sqrt{\sum_{x,y} \left( x \cos(\phi) + y \sin(\phi) \right)^2 \left[ E_i - 0.5 \text{std} (E_i) \right]_+}
\]

(4.1)

where \( x \) and \( y \) are the positions relative to the center of gravity of the receptive field. The aspect ratio is defined as \( \frac{L}{V} \). The subtraction and rectification prevents points with low values, lying far from the receptive field, from strongly influencing the aspect ratio. This is comparable to removing values below the noise level in physiological experiments. Histograms are compared using a one-sided Kolmogorov-Smirnov (KS) test yielding the probability of both histograms being drawn from the same distribution.

**Parametric studies**

In parametric studies we characterize the dependence of \( \Psi(\theta, \alpha, \kappa) \) on the receptive field properties. To elucidate why sparse coding alone is not expected to result in complex cell type responses, we also measure the dependence of a specific definition of sparseness on the receptive field properties

\[
\Psi_{\{\text{Kurtosis}\}} = \sum_i \left( \frac{A_i}{\langle A_i \rangle} \right)^4
\]

(4.2)

We repeat this simulation with the objective function derived from the Cauchy prior and the SD obtaining essentially the same results. We use the same two-subunit model as in the optimization procedures in the preceding text albeit with simplified receptive fields. Because the optimization methods result in Gabor type receptive fields and neuronal receptive fields are well approximated by these, we choose the subunits to be Gabor wavelets of fixed orientation and spatial frequency. The phase and aspect ratio of each subunit, however, remain free parameters

\[
G(a, s, s_x, s_y; x, y) = \sin \left( 180 \frac{a}{\alpha} + s \right) e^{-\frac{x^2}{(a_x)^2} - \frac{y^2}{(a_y)^2}}
\]

(4.3)

where \( a \), which is fixed to a value of 5 pixels, is the size of the Gabor, \( s \) is the relative shift between the subunits, \( s_x \) and \( s_y \) are the relative length and width, and \( x \) and \( y \) the relative positions of the pixels. For figure 4.1A, B and C, we choose identical shapes: \( W_1 = G(5, 0, 1, 1), W_2 = G(5, s, 1, 1) \) and vary the shift, \( s \), between the subunits. For figure
4.1C, we choose a fixed shift of 90°: \( W_1 = G(5,0,\lambda,w), W_2 = G(5,90,\lambda,w) \) and vary length, \( \lambda \), and width, \( w \), between 0.5 and 4 in steps of 0.1. Aspect ratios are binned in steps of 0.2 between 0.2 and 5.

### 4.1.2 Results

We simulate neurons and adapt them to display optimally stable activity over time. The resulting response properties are characterized by the receptive fields of their two subunits (figure 4.2A). Most of the subunits exhibit a receptive field that is well described by a Gabor wavelet. They thus have receptive fields that are localized in the visual space and that are selective to orientation and spatial frequency. Most neurons exhibit a phase shift between the Gabor wavelets representing the receptive fields of each of its subunits that is close to a quarter cycle (90°). This suggests that the response properties of the simulated neurons exhibit some translation invariance, a key property of complex cells. The neurons are furthermore tuned to orientation and spatial frequency (Figure 4.2, panels B and C) (see also (Webster & De Valois, 1985)).

In the following, we quantitatively compare the simulated neurons’ responses to bars and gratings to those of real neurons. First we investigate the orientation specificity. In response to a bar of optimal width, the population of optimized neurons displays a narrow orientation tuning (38° width, Fig. 2A). This specificity is somewhat tighter than the tuning width of real complex cells (56°, \( p < 0.001 \), KS test) (Schiller et al., 1976a). The simulated neurons also exhibit a tight tuning (index of 51.9) to spatial frequency comparable to the tuning index of cortical neurons (average index of 46.9, (Schiller et al., 1976b)), although the small difference is significant (\( p < 0.01 \), KS-test).

Next we compare real and simulated neurons on the basis of their response to moving gratings. In primary visual cortex, a bimodal distribution of relative modulation strengths is observed (Skottun et al., 1991) (figure 4.3C). Complex cells are defined as having a relative modulation \( > 1.0 \), whereas simple cells are defined by larger values of the modulation ratio. In our simulations, a wide bimodal distribution of AC/DC values is also observed. The AC/DC ratios of the optimally adapted complex cells have a mean (0.41) that is not significantly larger than the experimentally observed relative modulations (0.40, \( p > 0.3 \), KS test). Last we compare the aspect ratios of the receptive fields defined as the ratio of its width relative to its length. Real complex cells have an aspect ratio of 1.02° ± 0.2 (Olzawa & Freeman., 1997) (mean ± SD; figure, 4.3D). The optimally adapted neurons have an aspect ratio of 1.00 ± 0.3, closely matching the experimental values (\( p > 0.3 \), t-test).

AC/DC ratio and aspect ratio define the invariant processing performed by complex cells. Thus the simulated neurons with optimally stable activity result in good fits to the measured properties of complex cells in the primary visual cortex. It has been proposed that combining sparse coding with appropriate boundary conditions also leads to complex cells (Hyvärinen & Hoyer, 2000). We repeat that simulation using our stimulus database. This simulation yields neurons with an orientation selectivity of 37° and a spatial frequency selectivity of 40.5, both well in the range of the physiological values (56°, 46.9, respectively) and comparable to optimizing a stability objective (38°, 51.9, respectively). For the AC/DC ratio, this simulation, however, results in a value of 0.65 that is far larger than the physiological value (0.40) and the result of optimizing a stability objective (0.41; \( p < 0.001 \) KS-test). Thus combining a sparseness objective with additional boundary conditions does not result in sufficiently translation invariant neurons. Furthermore, the aspect ratio of 1.73 is far larger than the one observed for real complex cells (1.02, \( p < 0.001 \) t-test). Similar results and equally significant deviations are found if we exchange stable in our simulations by the objective function derived from a Cauchy prior as used in (Olshausen & Field, 1996) or
Figure 4.1: The influence of the parameters. A: neurons consisting of 8 half-squaring \( f(x) = x^2 \) for \( x > 0 \), \( f(x) = 0 \) otherwise subunits are modeled. The histogram of AC/DC ratios is shown for cat complex cells (Skottun et al., 1991) and for the optimized neurons with 8 half-squaring subunits each. B: back in the simple 2-subunit energy model, the objective functions \( \Psi^{(stable)} \) (thick lines) and \( \Psi^{(kurtosis)} \) (thin lines) are plotted as a function of the relative phase between subunits with Gabor shaped receptive fields. C: objective functions are plotted as a function of the aspect ratio. Error bars denote the SE.
Figure 4.2: Qualitative properties of the simulated neurons. A: pairs of receptive fields are shown of neurons with optimally stable activity. B: the responses of 2 representative neurons to bars of changing orientations and widths (displayed at the optimal positions) are shown. C: responses of the same neurons to gratings of optimal phase and orientation but varying spatial frequency are shown. D: the optimal orientation and optimal spatial frequency are plotted for all the simulated neurons.
Figure 4.3: Density distribution of properties of complex cells in primary visual cortex and of neurons with optimally stable activity. A: the orientation tuning widths are shown for cortical complex cells in monkey cortex (Schiller et al., 1976a) and for the simulated neurons. B: the selectivities to spatial frequency are shown for cortical complex cells in monkey cortex (Schiller et al., 1976b) and for the simulated neurons. C: the relative modulation strengths are shown for a collection of 1,061 cat complex cells replotted from (Skottun et al., 1991) along with the AC/DC ratio of the simulated neurons. D: the distribution of aspect ratios is sketched for cat cortical neurons (Ohzawa & Freeman, 1997). This is compared with the aspect ratios of the simulated neurons.

the Kurtosis. This suggests that only the objective of stability adequately explains the properties of complex cells.

The head-mounted camera does not register changes in gaze associated with movements of the eyes. However, the results of chapter 5 will indicate that under the conditions the stimuli were recorded eye movements contribute little to stabilizing the retinal image. To control for possible residual stabilizing effects of eye movements, we perform two experiments: 1) we simulate eye movements that randomly stabilize 50% of the patches. And 2) we randomly shuffle 10% of the patches. The resulting receptive field properties are essentially unchanged in both cases. In particular in both cases, they are translation invariant and have AC/DC ratios close to the relative modulation of physiological data (p > 0.3 for both controls, KS test). Therefore we do not expect major changes of the reported results if eye-movements of the cats under free viewing conditions were taken into account.

To investigate if the results generalize to a more general nonlinear model or if the results are due to the way, we constructed our model neurons we perform an additional simulation (figure 4.1A). Simulated neurons consisting of eight half-squaring subunits are modeled. The neural properties resulting from optimizing stable are similar to those found for the two-subunit energy model described in the preceding text. Importantly, the AC/DC ratio distribution is not significantly larger than the relative modulations of real complex cells (p > 0.3, KS test). Thus the results do not critically depend on the constraints on the model neurons’ nonlinear properties defined by the two-subunit energy model. The type of the nonlinearity is set in our simulations. For the neurons to exhibit complex cell properties, however, the subunits need to obtain identical orientation and spatial frequency as well as the right phase shift. This simulation thus shows that these properties can be obtained from natural scenes even for varied neuron models.
To better understand the preceding results, we proceed to characterize some important nonlinear statistical properties of videos natural scenes. To do so, we measure the objective values of simulated neurons in response to the videos of natural scenes. We choose the subunits of the same model as in the preceding text to be Gabor wavelets of fixed orientation and spatial frequency, leaving the aspect ratio and the relative phase as free parameters. With this more restricted set of subunit receptive fields, we can systematically analyze the influence of the receptive field properties on various objective functions. Varying the relative phase of the subunits reveals that stable is maximal if the simulated neuron is translation invariant and the wavelets have a relative phase of 90° (figure 4.1B). Neurons then represent localized oriented energy detectors and are translation invariant as are real complex cells. We furthermore analyze the influence of the aspect ratio on the objective functions (figure 4.1C). stable reaches its highest value for spherical receptive fields with an aspect ratio of 1 similar to the value of real complex cells (Ohzawa & Freeman, 1997). For comparison with other studies, we also plot sparseness as a function of phase and aspect ratio, which peaks at values that are far from those found in physiology. It thus seems that stability is a good candidate for an adaptation criterion that links complex cells with the statistics of natural scenes.

4.1.3 Discussion

We have show that adaptation to a stability objective leads to simulated neurons sharing important spatial properties of complex cells in the primary visual cortex. Sparseness can be derived from several ideas such as minimizing energy consumption, optimal channel coding, or searching for a meaningful representation of data. Stability can also be derived from various ideas: high level variables such as object identities are stable, stable variables can be transmitted through channels with lower bandwidth, and learning is easier in a system where variables change slowly.

Recently (Hurri & Hyvärinen, 2003) have proposed that optimizing stability of linear neurons in response to natural stimuli leads to receptive fields like those of simple cells. The stability of linear neurons, however, is always considerably lower than the stability of the nonlinear complex cells in our study. The authors furthermore use a slightly different objective that biases the neurons to be both stable and sparse. These results might still indicate that both simple and complex cell responses could be understood in a coherent framework derived from the idea of stability.

In our simulations, each neuron only saw the input stimulus windowed by a Gaussian. Parts of the properties of the neurons, in particular the aspect ratio could thus be affected by this preprocessing. Some of the simulated neurons, however, do have receptive fields that are smaller than the size of the Gaussian. There is a tendency for neurons to obtain localized receptive fields. It would be interesting for future studies to analyze if the distribution of receptive field sizes can be obtained exclusively from optimizing stability. Such studies would, however, need very large numbers of simulated neurons as they would need to jointly encode the retinal space in addition to the orientation and spatial frequency space.

Do neurons found in primary visual cortex exhibit sparse or stable or maybe both types of response properties? Both objectives seem useful for processing in the nervous system. The question of which objective links the properties of natural scenes to the properties of complex cells is experimentally accessible. On one hand, for these analyses, recordings from neurons in response to natural scenes would need to be compared with response to artificial stimuli such as bars or gratings. With respect to sparseness some experiments started to address this issue (Baddeley et al., 1997; Vinje & Gallant, 2000). If a large set of natural visual patterns is presented in sequence, most of these are not effectively stimulating the
recorded neuron. A small subset of stimuli, however, can activate the neuron strongly and elicit very high firing rates. Similar experiments could address how stable neural responses are.

The good match between physiological and modeling data raises the question whether there is also a possible physiological implementation of the proposed objective function. In (Einhäuser et al., 2002a; Einhäuser, 2001) we provide a physiologically inspired implementation of the temporal coherence principle: A possible mechanism for the de-correlation part of the objective \( \Psi^{(\text{decorr})} \) is provided by inhibitory lateral connections. Physiologically, this process does not necessarily require synaptic changes at the tangential connections, but can also exploit modifications to afferent synapses that are in turn influenced by lateral interactions (Körting & König, 2000). This mechanism utilizes action potentials propagating retrogradely through the dendritic tree (Stuart & Sakmann, 1994). Building on the same physiological mechanism, a temporally asymmetric learning rule (Markram et al., 1997; Larkum et al., 1999) could plausibly subserve the function of the stability objective \( \Psi^{(\text{stable})} \) (Einhäuser et al., 2002a). These mechanisms could provide a physiological substrate for the proposed objective function.

The fact that complex cells of adult animals are well described as an adaptation to a stability objective raises the question whether this adaptation occurs on ontogenetic time scales. If there is an ontogenetic component to the development of complex cells, it allows the following experimental test of the stability hypothesis. Changing the environment during an animal’s critical period (e.g., by strobe rearing) would impair the development of complex cell type receptive fields. In particular, there should be a range of strobe rates in which complex cells are severely affected, whereas simple cells are not. Based on the data of chapter 3, this rate is expected to be of the order of 10Hz.

If simple cells optimize a sparseness criterion and complex cells optimize a stability criterion, it is tempting to speculate, whether such a division of labor is repeated in higher areas. Indeed, in a widely used architecture for invariant object recognition, the Neocognitron (Fukushima, 1980), a hierarchical network with an alternation of simple and complex type cells is used. Hence it is interesting to build larger systems consisting of several layers, each optimizing an adequate objective. In later sections of this chapter (4.3 and 4.4) we will address this issue using a fixed complex cell input layer, whose output feeds a second layer, that will also be trained to optimize the stability objective.

### 4.2 Functional segregation of colour and orientation

After having demonstrated how complex cell properties arise from optimizing the stability objective on natural videos, we here extend on the same principle. As discussed above, the emergence of complex cells can be interpreted as a consequence of local position (phase) having a shorter correlation time constant than orientation (section 3.2.4). The same temporal relation is true for colour relative to orientation in natural videos (section 3.2.5). As a decisive step towards a more general application of the temporal coherence objective, we in this section therefore show, that orientation and colour can be segregated by temporal coherence, just like position and orientation were for complex cells. This result indicates that temporal coherence can segregate any two features that vary on different time scales and thereby emphasizes the generalizability of the temporal coherence principle.
4.2.1 Specific methods

Stimuli

For this experiment we use a total of 4900 consecutive frames from the catcam movies described in section 2.1.1. Stimulus extraction and preprocessing is done as in the previous section (4.1) with the following exceptions: frames are not converted to grayscale, but all 3 colour channels are used (R,G,B); Gaussian smoothing is applied to each colour channel individually. Since the original input dimensionality increases by a factor of 3 (3 colour channels), we here use principal components 2 to 200 unless otherwise stated.

Neuronal models

For the modeling experiment described in this section we use a neuron that consists of a single subunit ($S=1,n=1$ in equation 2.12). The neurons’ non-linearity is either of the linear threshold type ($\phi(x) = \max(x,0)$, equation. 2.12) or full-wave-rectifying ($\phi(x) = \max|\text{ }x\text{ }|$). In each case the output of a neuron cannot be less than zero, reflecting the fact that real neurons cannot spike at negative rates. Unless otherwise stated we use $N=200$ neurons in this experiment.

Analysis of chromatic receptive fields

We analyse the properties of the receptive fields in the network after convergence. First, the neuron’s receptive field representation in input space is obtained by inverting the PCA on the weight matrix, W. Then, each receptive field is scaled individually such that the values of each pixel fall between 0 and 1. The scaling is the same for each colour channel and thus does not bias either the chromatic or the spatial properties of the optimized receptive field. To analyse colour and spatial content independently, the receptive fields are transformed from a colour channel representation (RGB format) into a representation separating hue, saturation and brightness channels (HSV format), achieved using a standard function in MATLAB rgb2hsv.m

We characterise the chromatic selectivity of a receptive field by calculating the mean saturation across pixels. Neurons with a mean saturation greater than 0.2 are classed as chromatic neurons, while the remaining neurons are classed as achromatic. Spatial properties of the receptive field are assessed by measuring the anisotropy of the receptive field using standard methods (Jähne, 1997). Briefly, the tensor of inertia is computed on the values from the brightness channel. Anisotropy is defined as the ratio of the difference between the tensor’s long and short principal axis, divided by their sum. This measure is 0 for an isotropic (non-oriented) receptive field and approaches 1 for a perfectly oriented receptive field. Neurons with an anisotropy value less than 0.2 were considered to be non-oriented.

4.2.2 Results

Under all conditions analysed here, the optimisation process converges rapidly and reaches steady state after less than 60 iterations (Figure 4.4). The optimised receptive fields are analysed in input space (by inverting the PCA on the weight-matrix, see Methods). As a starting point, we chose the following parameters values: 200 neurons, using PCA components 2-200, $\beta = 5$ and full-wave rectifying neurons. Using this as a baseline, approximately 80% of the neurons exhibit colour selectivity. Figure 4.5 shows a complete set of receptive fields, sorted by the individual stability value $\psi_i^{(stab)}$
Figure 4.4: Development of objective function $\Psi$ in the course of the optimisation process for a simulation of full-wave-rectifying neurons with $\beta = 5$.

Figure 4.5: Receptive fields produced following 60 iterations of the simulation shown in figure 4.5. The receptive fields are sorted according their individual stability value, with low stability values at the top of the figure and high stability values at the bottom.
Figure 4.6: Coefficient of correlation between all the activities over one stimulus presentation of the neurons of figure 4.5.

Figure 4.7: Dependence of neuron chromaticity on individual objective value $\psi_i^{(stabil)}$. Chromaticity was assessed by calculating mean saturation for each receptive field, averaged across pixels.

The stability objective ($\Psi_i^{(stabil)}$) does not include interactions between neurons in the network. The de-correlation term ($\Psi_i^{(decorr)}$) force neurons to acquire different receptive fields. As stability and de-correlation are competing mechanisms being added into a single objective function, neurons with sub-optimal stability emerge as well as neurons with suboptimal de-correlation. As one expects, activities of neurons with high stability values $\psi_i^{(stabl)}$, tend to be more correlated than those with low stability values (Figure 4.6).

Visual inspection of the neurons in figure 4.5 already suggests that chromatic neurons tend to have higher values of $\psi_i^{(stabl)}$ than achromatic neurons. Quantitative analysis reveals that chromatic neurons indeed have higher stability (mean $\psi_i^{(stabl)}$, -0.0041) than achromatic neurons (mean $\psi_i^{(stabl)}$, -0.0053). There is a pronounced relation between a neuron’s chromaticity and its individual stability (correlation coefficient: r=0.70, Figure 4.7). This demonstrates that chromatic neurons have optimally stable responses to natural stimuli, while the achromatic stimuli are a consequence of sub-optimal stability due the de-correlation objective.

We also investigate the spatial properties of the neurons in the optimised network. Orientation selectivity was estimated from the degree of anisotropy in the receptive field (see Methods). In the simulation using the baseline parameter values described above, 66% of
Figure 4.8: Chromaticity versus anisotropy for $\beta = 5$ and full-wave rectifying neurons.

Figure 4.9: Chromaticity versus anisotropy for $\beta = 5$, linear-threshold model.
the neurons are non-oriented. In the simulations using the linear-threshold model, 81% are non-oriented. These results suggest that there is a relationship between spatial and chromatic properties of our model neurons. We find a strong correlation between chromaticity (mean saturation) and isotropy (defined as 1 minus anisotropy), both for the linear-threshold model (correlation coefficient: 0.72, Figure 4.8) and full-wave rectifying model (correlation coefficient: 0.79, Figure 4.9). Thus, colour selective neurons tend to be non-oriented, while achromatic neurons tend to be tuned for orientation.

Our results show that achromatic neurons emerge as a consequence of adding a de-correlation term ($\Psi^{(\text{decor})}$) to our objective function. The de-correlation objective permits interactions between neurons in the network, and results in receptive fields that are sub-optimal with respect to the stability objective alone. We test whether increases in the relative contribution of the de-correlation objective (using the term $\beta$ see section 2.6.2) would further increase the fraction of achromatic neurons in the network. The limit case of $\beta \to \infty$ is simulated by omitting stability from the objective function altogether. For the simulations using the linear-threshold model, the population average of mean saturation (chromaticity) is reduced from 0.50 where $\beta = 0$ (no contribution from $\Psi^{(\text{decor})}$) to 0.19 where $\beta \to \infty$ (no contribution from $\Psi^{(\text{stab})}$). The results from the simulations using the full-wave rectifying neurons are qualitatively similar, with a reduction in mean saturation from 0.5 ($\beta = 0$) to 0.25 ($\beta \to \infty$). For both neuron types a 50% drop in mean saturation is achieved for values of $\beta = 20$ (Figure 4.10).

The reduction in mean saturation is associated with a reduction in the proportion of chromatic neurons. In the simulations using the linear-threshold model, increasing $\beta$ from 1 to 20 reduces the proportion of chromatic neurons from 100% to 61%. In the full-wave rectifying model simulation, the same increment in $\beta$ reduces the proportion of chromatic neurons from 93% to 63%. These results show that the proportion of chromatic versus achromatic neurons as well as the average chromaticity across the network is determined by the relative contribution of the de-correlation versus the stability objective. Therefore a single parameter, $\beta$, is sufficient to determine the proportion of chromatic versus achromatic neurons in the network.

Next, we investigate the robustness of our model to reductions in the dimensionality of the training input. We determine the change in mean saturation when as few as 25
Figure 4.11: Dependence of mean neuron chromaticity on PCA dimension for different values of $\beta$ and the two neuron models.
principle components were used, roughly $\frac{1}{3}$ of the principle components used in the main simulations (see Methods). The change in mean saturation across both linear and full-wave rectifying models, and for values of $\beta$ between 5 and 10, was not larger than 53% (Figure 4.11). This minimal dependence on the number of input dimensions contrasts with a study by Hoyer & Hyvärinen (Hoyer & Hyvärinen, 2000) using independent component analysis, which reported a strong correlation between input dimensionality and the proportion of colour selective neurons. Increasing input dimension by a factor of 2.5 yielded an increase in the number of chromatic neurons of approximately 290% (estimated from Figure 10 in (Hoyer & Hyvärinen, 2000)). Doubling the number of independent components (from 100 to 200 dimensions) yielded a 77% increase in chromatic neurons, while in our simulations, which use PCA, the same increase in dimensionality produces an increase of just 11% (Figure 4.12). This indicates that in simulations using ICA, the emergence of colour selective neurons depends on dimensionality of the input, while in simulations using the stability objective, colour selectivity is determined by the relative strength of the de-correlation objective.

We sought to verify that the results of our simulations are indeed a consequence of the temporal structure in the natural stimuli used as training input. Additional simulations were performed in which we destroy this temporal structure by randomly shuffling the frames comprising a stimulus pair. We compare the value of the objective function between initialisation and steady state conditions. In the simulations of full-wave rectified neurons, $\beta = 10$, steady state values are 8% higher than at initialisation (from -2.23 to -2.04). In the simulations using the linear-threshold model, an increase of 4% is found (from -2.05 to -1.96). These values are modest compared to those found in the simulations with natural temporal ordering in the training input (from -1.96 to -1.32, or 33% in the full-wave rectified model simulations and from -1.72 to -1.25, or 27% in the linear-threshold model simulations).

The impact of shuffling the stimulus pairs was even more dramatic when the objective function was defined by $\beta$ stable alone ($\beta = 0$). In this case, the increase in the value of the objective function when using the natural temporal stimuli is more than 10 times larger (31% increase for the full-wave rectified model, 43% increase for the linear-threshold model) than for the simulations using the temporally shuffled stimuli (3% increase for the full-wave
rectified model, 4% increase for the linear-threshold model).

We also determined the effect of shuffling the stimulus pairs on the emergence of colour selectivity in the network. Mean pixel saturation was 0.19 in the network trained on the shuffled stimuli, half the value that is obtained when the natural temporal order is intact (0.38). This resulted in the proportion of chromatic neurons in the network falling from 65% in the naturally ordered condition to just 19% in the shuffled condition. Furthermore, most neurons in the simulation using the linear-threshold model are close to the chromatic/achromatic threshold of 0.2 (96% fall between 0.1 and 0.3 mean pixel saturation), whereas for naturally ordered condition, mean saturation shows a wider distribution (43% between 0.1 and 0.3). While the change in mean saturation is less pronounced for the full-wave rectifying model (0.24 compared to 0.29), a similar narrowing in chromaticity is observed (63% compared to 30% of neurons fall between 0.1 and 0.3 mean saturation, Figure 4.13). This indicates that in the shuffled condition, most neurons cannot be clearly identified as either chromatic or achromatic.

Taken together, these findings indicate that natural temporal structure is critical to the attainment of an optimal solution for the stability objective function at the level of the network, and for the emergence of distinct chromatic and achromatic neuron populations.

4.2.3 Discussion

In this section we addressed how neurons selective for different stimulus dimensions can emerge from an initially homogeneous population. We have shown that optimising the stability objective alone yields non-oriented chromatic neurons. Forcing neurons, to acquire dissimilar receptive fields (and thus sub-optimal stability) leads to the emergence of a second subpopulation of oriented achromatic neurons. The stability of each neuron serves as system inherent measure to separate the two groups of neurons.

Furthermore, our simulations show that the relative size of each subpopulation is determined by a single parameter. Thereby, we have shown that the proposed objective function successfully segregates neurons into distinct populations that are selective to one property of the stimulus while being relatively non-selective to another property. By adopting complementary selectivities, a small number of neuronal populations can encode a complete set of features in the stimulus independently of each other, and thus achieve a parallel mode of processing.

We considered two distinct neuronal models in our simulations, the linear-threshold model and the full-wave rectifying model. The results are similar regardless of which neuronal model is used, indicating that our objective function can succeed irrespective of the type of model. This result highlights the potential utility of our approach. Our objective function may provide insights into the mechanisms underlying learning and development, not just at early stages of the visual system, but at higher levels of the visual hierarchy as well. The generalisability of our approach also holds promise for its application in the development of artificial vision systems.

Simulations of the development of colour selective responses using natural stimuli as input has been addressed in several recent studies. These adopt a version of the sparseness principle, using independent component analysis (ICA), and use standard colour images (Hoyer & Hyvärinen, 2000; Tallor et al., 2000) or hyperspectral images as training input (Wachtler et al., 2001). All studies find colour selective neurons, similar to those described here. However, neither study quantifies the relation of the neurons' spatial receptive fields to their chromatic properties. Here we find a strong correlation between chromatic and spatial properties. Another remarkable difference between the stability objective and the sparseness objective, as modelled using ICA, is the dependence of the latter approach on the
CHAPTER 4. TEMPORAL COHERENCE

dimensionality of the training input. This indicates that when using the ICA approach, the segregation of chromatic and achromatic subpopulations will depend on properties of the dimensionality of the external input. In the present approach this segregation is regulated by an internal parameter specifying the strength of the interactions between neurons within the network.

Our as well as earlier experiments on the extraction of complex cells by the temporal coherence principle were motivated by the idea that, if different transformations of the visual input naturally happen on different timescales, selectivity and invariance towards these transformations can be extracted using solely this fact (Fögliak, 1991). For example, the invariance to position and selectivity to orientation results from local orientations being correlated over longer time scales than positions (see section 3.2.4). As the temporal structure of the stimulus is therefore the decisive property exploited by this type of objective function, it is crucial that we demonstrate that these results are confirmed when using stimuli that preserve temporal structure in the natural input. Our unique stimuli, derived from recordings made from a camera mounted on a freely behaving cat, provide image sequences that preserve the natural temporal structure. Note however, that we found it technically infeasible to use more realistic colour representations, as the current sampling rates of devices capable of recording hyperspectral images are several orders of magnitude below the correlation time constants observed in our natural videos (chapter 3). Thus, we were not able to capture this aspect of the natural input. Note that the results of the ICA studies using hyperspectral images (Wachtler et al., 2001) are not qualitatively different to those in which the standard RGB-representation was used (Tailor et al., 2000; Hoyer & Hyvärinen, 2000). Thus, we can be confident that the RGB format is adequate for our purposes. Furthermore, the present 'Catcam' videos were not recorded at sufficient temporal resolution to model the temporal properties of receptive fields. Modeling chromatic spatio-temporal receptive fields with the stability objective thus remains an interesting issue for future research.

There is physiological evidence that colour-sensitive neurons in primate V1 tend to be non-oriented, whereas achromatic neurons tend to exhibit the precise orientation tuning seen to be characteristic of V1 simple and complex neurons (Gouras, 1974; Lennie et al., 1990). This is further supported by psychophysical experiments that show that humans can resolve higher spatial frequencies for isochromatic patterns than for isoluminant patterns (Webster et al., 1990; Sekiguchi et al., 1993). In addition to the finding of the previous section (4.1), this again offers a link between properties of real neurons and our simulated cells. As for complex cells, the segregation of colour and orientation on the basis of their different temporal statistics can also be implemented in a physiologically inspired framework (Einhäuser, 2001). Therefore the biological plausibility of our approach also in this case may go beyond the similarity between our simulated receptive fields and those found in visual cortex.

We speculate that important properties of the visual system, including the establishment of distinct functional pathways, may develop using learning rules resembling the stability objective.

4.3 Higher order properties: Texture cells

So far we have shown, that the stability objective can segregate different simple features on basis of their different temporal coherence. As a first step to extend this approach towards representing more complex structures, we implement the stability objective into a hierarchical network that learns “texture” representations. Although there seems to be some intuitive notion of “texture”, no consensus exists on a positive definition. Consequently - unlike selectivity to simple features, texture selectivity cannot readily be quantified. One
common working definition for texture is through spatial repetivity of a small scale structure on a larger scale. According to this definition texture already implies some translation invariance. We in this section therefore characterize texture cells mainly by their invariance properties. In addition we characterize the cells according to an intuitive notion of texture by comparing their responses to human observers that are asked to classify stimuli according to their texture. This psychophysically defined mid-level property of texture thereby also bridges the gap from the physiological properties described in the previous sections to system level behaviour we are aiming for by using temporal coherence.

4.3.1 Specific methods

Stimuli

For the modeling experiments presented in this section we use the complete cat-cam (section 2.1.1) video database, converted to greyscale as above (section 4.1). To obtain a larger variety of stimuli, we here extract two (instead of 20) patches per frame, which span 46 by 46 pixels, yielding a stimulus set of 100,000 pairs of patches for training the network.

Network

We simulate a two-layer feed-forward network (see Figure 4.14). The first layer consists of a grid of complex cells (CC). In this section we do not learn the CC receptive fields, but model their activity as the localized Fourier component:

$$A^{(CC)}_{\phi, \sigma}(x, y, t) = \left| \sum_x \sum_y J(x, y, t) e^{i \phi \sin \frac{y}{\sigma_y} \cos \frac{x}{\sigma_x}} e^{-\frac{x^2+y^2}{\sigma^2}} \right|$$  \hspace{1cm} (4.4)

where $J(x, y, t)$ is the image patch and $(x, y)$ the pixel position relative to the centre of the patch. We use the orientations $\phi \in \{45^\circ, 90^\circ, 135^\circ, 180^\circ\}$ 3 grids of CCs exist. For the smallest sigma of 1.4 pixels the grid is comprises 13 by 13 CCs. For the medium sigma of
CHAPTER 4. TEMPORAL COHERENCE

2.8 pixels the grid consists of 7 by 7 medium CCs, and for the large sigma of 5.6 pixels the grid consists of 4 by 4 CCs. The CCs are distributed regularly over the patch, leaving free a 4 pixel margin where no CCs have their center points. We thus obtain a pyramid of filters. This process yields a total of 936 orientation-specific, localized simulated complex cells in three different spatial scales. We perform a PCA then performed on the activity-traces of the 936 CC and 100 components, accounting for 77% of the variance are used as input (I in equation 2.12) to a second layer of texture cells (TC). The TC’s activity $A^{(TC)}$ is computed according to equation 2.12 using 10 neurons (N=10), 9 subunits (S=9), the 4-norm for subunit activities (n=4) and a linear threshold on the neurons’ activity ($\phi(x) = max(x, 0)$).

4.3.2 Data analysis

Unlike the specificity to a simple feature like position or orientation, the texture specificity of a cell cannot be readily defined. Hence we here define some indices to characterize the cells resulting from the simulation. In order to measure a TC’s activation independent of the stimulus’ contrast, we define its relative activation as the activity of a neuron divided by the sum of the activities of all other neurons:

$$A^{(TC, rel)}_i(t) = \frac{A_i^{(TC)}(t)}{\sum_j A_j^{(TC)}(t)}$$

(4.5)

We will refer to the stimulus, that elicits the strongest relative activation as “relatively best stimulus” for a given TC.

To characterize the response of the neurons with respect to translation and rotation we perform the following analysis. We determine the 1000 relatively best patches for each TC. An enlarged patch of 92 by 92 pixels that contains the optimum patch in its center is then copied from the respective input video frame. To analyse translation invariance a window of 46 by 46 pixels size is moved pixel by pixel diagonally over this big patch from the top left to the bottom right corner. For each window position the TC activity is calculated. The activity of each TC is normalized by the activity of the stimulus from the center of the total 92x92 patch. The average value and standard deviation of all activities and TCs is then used to characterize the TC’s translation invariance. Rotation invariance is computed analogously, by rotating the window in $5^\circ$ using the `imrotate.m` function of the MATLAB image processing toolbox.

Human classification

In order to test to which degree the classification by the TCs matches a subjective human texture classification, five naive subjects (3 female; 2 male, age: 25 to 31) were tested using the following paradigm: In each trial, one of the 2nd to 21st best stimuli for each TC was randomly chosen and displayed at the center of the screen. Each of the stimuli was used exactly once for each subject, yielding 200 trials per subject. Additionally, on the top of the screen the best stimulus for each cell was displayed. To avoid any spatial bias, the order of these stimuli on the screen was changed from trial to trial. The subject had to indicate by button press, which of the top-row stimuli “matches the texture of the central stimulus most closely.” The central stimulus disappeared upon the button-press and the subject had to confirm its choice. In rare cases ($\leq 2.5\%$ of the trials for each subject), when the subject indicated that it mistyped the button, it was allowed to correct its choice at this point. These corrected trials were excluded from analysis. Otherwise, after pressing the confirmation key, the next trial was started after a 200 ms blank. The presentation setup
was as described in section 2.3. Stimuli were presented at original resolution (46x46 pixel) corresponding to about 2° of visual angle.

4.3.3 Results

Single TC

We simulate a two layer network, whose second layer is trained to optimize the stability objective. We first quantify the response of individual cells, dubbed texture cells (TC), in this layer. We find that each neuron responds strongly only to a small subset of training stimuli, while most stimuli elicit only a weak response (histogram to the left of figure 4.15). In other words, the distribution of a single neuron’s activities is sparse. This high selectivity of a neuron raises the question on which features the neuron is selective to. We analyse the stimuli out of the training set that elicit the highest and lowest relative activities for the example neuron. While the best stimuli show a pattern that matches the intuitive definition of a texture, the worst do not share an obvious property (figure 4.15). This is a first indication that the TCs are indeed selective for texture.

TC population

We proceed to analyse the properties of the set of all our 10 TCs. Visual inspection of the best stimuli for each cell again indicates texture specificity (figure 4.16a). Furthermore, each TC seems to code for different type of texture than the other neurons. Since our aforementioned working definition of texture implies spatial continuity, we also analyse this property on nine representative frames of the training set. Indeed we find that the regions, for which each neuron codes, are spatially connected (figure 4.16b).

Invariance properties

Since translation invariance is a characteristic property of textures, we quantify this invariance. Translation by maximally 23 pixels (or half the receptive field size) to either side causes a mean drop from the maximal activity at 0 offset of less than 25%, the neuron is thus highly translation invariant (figure 4.17a). In comparision, when the input patch is rotated however the activity changes dramatically; neurons are highly selective to the orientation of a texture (figure 4.17a).

Human classification

So far we made use of our intuitive notion of texture. To quantify this intuition, we have 5 naive human subjects classify the best 21 textures of each TC. We find that subjective classification of stimuli matches the TCs classification in 56% of the cases (figure 4.18). This is far above the chance level of 10% for random guessing. Since subjects are instructed to classify according to texture it is unlikely that they rely on lower-level features. Furthermore, if people would base their decision on pixelwise Euclidean distance, the TCs would only match human performance in 14% of the cases. This strongly suggests that the selectivities of the TCs are indeed close to a intuitive human notion of texture.

4.3.4 Discussion

In the present experiment we integrated the stability objective into a hierarchical network. Since the input layer consists of complex cells, which can also be learnt by using the sta-
Figure 4.15: Characterization of a single TC The histogram of the activity of one TC is depicted on the left hand side of this figure. The value of the Activity is shown along the vertical axis while the number of times this activity is measured is shown along the horizontal axis in a logarithmic scales. The right part of this figure shows rows of input stimuli. Each row exhibits the ten best stimuli which activate the TC just below the denoted percentage of the maximum activity of this TC. The gray bars indicate the range of activities caused by the patches.
Figure 4.16: Characterization of the TC population A: the ten relatively best stimuli for each TC are shown (rows correspond to cells, best stimuli are sorted from left to right by activity). B: 9 frames from the Catcam videos are composed into a single grayscale image. The colour code denotes the TC that is most active in this region of the image and is identical to the colours used in panel A.
Figure 4.17: Response to translation and rotation. Solid lines indicate mean activity values, dashed lines indicate the standard deviation. Left: Translation. X-axis indicates shift from optimal position, y-axis indicates mean activity. Right: Rotation. Original patch is at 0°.

Figure 4.18: Texture classification by humans compared to classification by model for individual subjects and sum over subjects (lower right panel). X-axis: texture classification by model (stimulus belongs to 21 best stimuli for that cell), Y-axis: classification by human observer; box size: number of correspondences. Full box equals 100% match human vs. cell (i.e., 20 incidences for individual subjects, 100 incidences for sum over subjects). Number of valid trials and number of matches indicated on top of each plot; chance level at 10%.
bility objective, we show that repetitive application of the same principle may facilitate the representation of relevant structures in the input. This homogeneity is a desirable property, since it allows scaling up the system to more hierarchical levels. It is nevertheless tempting to speculate about the potential of a more heterogeneous architecture, e.g., by using alternating different principles when proceeding through the hierarchy, akin to the alternating simple and complex layers in the Neocognitron (Fukushima, 1980). One obvious candidate principle for temporal coherence to alternate with, is sparseness: A recent study showed that hierarchical application of sparseness on natural scenes can extract contours (Hoyer & Hyvärinen, 2002). As local contours are a property complementary to texture, it is likely that a combination of sparseness and temporal coherence is better suited to represent relevant features of natural scenes than either objective alone.

4.4 Invariant representations of objects

So far we have shown that cells obtained by optimizing the stability objective match known physiological and psychophysical properties of the mammalian/human visual system. We demonstrated that these cells are specific to one input feature (orientation, colour, texture) while invariant to others (position, orientation). We explained these results based on the different timescales these features vary on in natural scenes. However the invariances demonstrated up to now were rather trivial (mainly translation and rotation in the image plane) compared to the non-isomorphic transformations the human visual system can also readily cope with (changes in viewpoint, size, deformations of non-rigid objects). In this section we use the stability objective to representations of real-world objects, that are less sensitive to changes in viewpoint than the input and thus facilitate invariant classification.

4.4.1 Specific methods

Stimuli

In this section only stimuli that are based on the Columbia Object Image Library are used and preprocessed as described in section 2.1.3.
CHAPTER 4. TEMPORAL COHERENCE

Network

Like in the previous experiment in input layer of our network we use neurons that resemble complex cells (CC), which are fixed. To obtain a relatively coarse representation of the input, we use 36 CCs, which are again modelled as complex Gabor filters V.

\[ V_{k,\phi}(x, y) = e^{-2k((x-x_0)^2+(y-y_0)^2)}e^{-2\pi ik((x-x_0)\cos(\phi) -(y-y_0)\sin(\phi))}; \]

\[ 1 \leq x, y \leq 32; x_0 = y_0 = 16.5 \]  

(4.6)

where k denotes the spatial frequency and \( \phi \) the orientation of the Gabor; x and y the pixel coordinate in a 32x32 pixel wide patch and \( x_0, y_0 \) the patch centre. To compute the CCs’ activities on each input image \( J(x,y,t) \) the colour channels \( J_C \) of J are first convolved separately with the Gabors:

\[ J_{c,k,\phi}(x, y, t) = \sum_{x'} \sum_{y'} V_{k,\phi}(x' - x, y' - y)J_C(x', y', t) \]  

(4.7)

The according CC activity is then given by the absolute value of pooled over all locations:

\[ A_i^{(CC)}(t) = \sum_{x} \sum_{y} |J_i(x, y, t)| \]  

(4.8)

where the index i summarizes the subscripts \( (c, k, \phi) \). Each CC’s activity is normalized to zero mean and unit standard deviation. We use 4 orientations \( (0^\circ, 45^\circ, 90^\circ, 135^\circ) \), 3 spatial frequencies \( \left( \frac{1}{8 \text{pixel}}, \frac{1}{16 \text{pixel}}, \frac{1}{32 \text{pixel}} \right) \) and 3 colour channels (R,G,B), resulting in \( 4^4 \cdot 3 = 36 \) different CCs (figure 4.19). Note that - unlike in the previous experiment - CC activities are already pooled over locations. In physiological terms the CCs used here rather correspond to a functional column of complex cells than to a single complex cell. The CCs project on a second layer of N neurons, which will be referred to as object cells (OCs) and whose activities are computed according to equation 2.12 with input vector \( I = A^{(CC)} \). Note that, unlike in the previous experiment - no principal component analysis is performed, but the full vector is used. Unless otherwise stated the number of OCs is set to N=36 and the number of their subunits to S=8. The subunit’s non-linearity in equation 2.12 is set to n=4 and no other non-linearity is applied (\( \phi(x) = x \)).

Quantifying viewpoint invariance

In order to quantify viewpoint-invariance, we probed CCs as well as converged OCs with all views of the objects used for training (Note that for the training usually only a subset of views was used). Each view of each object is presented 100 times at different locations and - for simulations including distractors - with changing random distractors. The cell activity was averaged over these 100 presentations and this average will be referred to as the response of the cell to a certain view of a certain object. The response of each cell over this complete set of objects and views was then normalized to zero mean and unit standard deviation. To quantify a cell’s dependence on the viewpoint we take the standard-deviation of the response over all views of each object. 1 minus this standard-deviation defines a measure of viewpoint invariance for each cell and object. Note that viewpoint invariance can be calculated for each object separately to distinguish effects that result from object inherent rotational symmetries from those that are an acquired property of a cell.
Figure 4.20: Single cell activity during presentation of 10 objects rotating over all 72 viewpoints during testing. Top row: complex cell activity, bottom row: object cell activity, middle: presented object (For illustration depicted centrally in full database resolution and from 0 degree viewpoint only). Data of each viewpoint are averaged over 100 test presentations at different locations. In the used example object cells were trained with 12 views (30° steps) of each of the 10 shown objects.

Figure 4.21: a) Invariance index histogram for all 36 CCs for object 1. Vertical line indicates mean. (b) Invariance index histogram for all 36 OCs for object 1. Vertical line indicates mean. (c) Invariance index mean over all CCs (x-axis) versus mean over all OCs (y-axis). Each datapoint corresponds to one object. The example of panels a and b is marked by a circle.

Classification performance

In order to analyse how well OC and CC outputs can be used for object classification, we performed a clustering on the cells’ activities during testing. We repeatedly applied the k-means algorithm (implementation from MATLAB’s statistics toolbox) with 10 different random initial settings. The number of clusters was chosen to match the number of objects. To assess the number of correctly classified objects, each cluster had to be assigned an object, which was done by the procedure described in the appendix. The fraction of correctly classified test presentations was computed independently for each repetition of the k-means algorithm and the mean across repetitions is referred to as unsupervised classification performance.
Figure 4.22: Invariance index for each cell (averaged over objects) plotted versus individual stability of each cell. Solid lines: Regression lines for best linear fit. (a) CCs (r = 0.28) (b) OCs (r = 0.81)

4.4.2 Results

Single cell properties

We train object cells (OCs) to optimise the stability objective on a subset of viewpoints of several objects of the COIL-100 database. After convergence we probe both OCs and their inputs (complex cells, CCs) on all trained objects from all viewpoints. The response trace of one example cell of each cell type is depicted in figure 4.20: For objects, whose appearance changes considerably with changing viewpoint, the response of the CC dramatically depends on the viewpoint (Figure 4.20, top row). The response of the OC, on the other hand, does not exhibit a strong modulation while the same object is presented, but responds rather independently of the viewpoint (Figure 4.20, bottom row). Consequently the OC can be characterised view-point invariant cell.

We now investigate to what extent the discussed example is typical for the whole population of cells. Unless otherwise stated, all data presented in the next paragraphs refer to a simulation, in which OCs were trained on 12 views (30° steps) of 50 objects in the absence of a distractor. We first measure viewpoint invariance - as defined in the methods, for an object that by itself exhibits only little rotational symmetry, object #1 (a flat box). We find that, while most CCs show only little viewpoint invariance (mean for object #1: 0.57, figure 4.21a), the response of most OCs have invariance values close to the theoretical maximum 1 (mean for object #1: 0.78, figure 4.21b), i.e. are nearly viewpoint invariant. Indeed for 36 out of the 50 objects, mean viewpoint invariance is larger for the OCs than for the CCs (figure 4.21c), a highly significant effect (p < 0.005, sign-test). This result indicates that the stability objective indeed leads to the emergence of viewpoint invariance. To further investigate how the stability objective relates to viewpoint invariance, we directly compare the individual stability values of each cell to its mean viewpoint invariance (averaged over objects). As expected - since OCs are trained to optimise stability, the individual stability values of CCs (mean: -0.046) are significantly smaller (p < 10^{-4}, t-test) than those of OCs (mean: -0.038). The same is true for the mean viewpoint invariance (CC: 0.671, OC: 0.783, p < 10^{-11}, t-test). Nevertheless, there is a highly significant correlation between viewpoint invariance and individual stability for both CCs (r=0.71, p < 10^{-5} figure 4.22a), and OCs (r=0.53, p < 10^{-3}, figure 4.22b). This shows that stability is indeed a good correlate of viewpoint invariance, and the optimisation of stability is the key to the emergence of viewpoint invariance in the OCs.
CHAPTER 4. TEMPORAL COHERENCE

Figure 4.23: Dependence of invariance on the number of subunits of each OC. Mean and standard error over objects and OCs.

Dependence on number of subunits

All simulations reported so far used 8 subunits (S=8). Do the observed invariance properties critically depend on this choice? We perform additional simulations varying the number of subunits from 1 to 8 (figure 4.23). We find that invariance is significantly dependent on the number of subunits \( p < 10^{-15}, \text{ANOVA} \). However, invariance increases monotonically only up to 6 subunits and then saturates. This shows that increasing the number of subunits further does not affect invariance anymore. Consequently 8 subunits is a reasonable choice.

The emergence of viewpoint invariance

In order to achieve a sound understanding as to how the described OC invariance emerges from appropriately combining their CC afferents, we more closely analyse these projections. figure 4.24 illustrates how OC activity derives from CC activity. Most CCs exhibit a strong modulation with viewpoint (figure 4.24, left). However, only few of the CCs have strong projections to a given OC subunit (figure 4.24, right). As the viewpoint dependent modulation is strong for inhibitory as well as for excitatory projections to a subunit, the peaks get flattened out (figure 4.24, third row from the bottom). After passing through the even non-linearity the negative troughs become peaks (figure 4.24, second row from the bottom). Consequently the sum over subunits only shows a weak and biphasic modulation with viewpoint (figure 4.24, bottom), and thus increased viewpoint invariance. The chosen OC exemplifies three major factors that shape the OC response: (i) Only few CCs have strong connections to a given subunit, i.e. the connectivity is sparse. (ii) Subunit activities get more invariant by excitatory and inhibitory projections from CCs, that antagonize each other. (iii) Subunits that receive mainly inhibitory and mainly excitatory projections antagonize each other and flatten the response further; while the absolute integral, and thus the object specific response, is persevered by the even non-linearity. To quantify these observations for the whole population of cells we first analyse connectivity weights over all cells and subunits, and find that this distribution is indeed sparse, i.e. heavy tailed (figure 4.25). We measure subunit invariance analogously to CC and OC invariance. We find that the mean invariance of the subunits (0.72) is in between those of CCs (0.67) and OCs (0.78), and significantly different from either \( p < 0.005 \) for both pair-wise t-tests. To test how many subunits are actually needed to achieve invariance, we compute each OC’s activity only including its n most invariant subunits. We find that invariance increases gradually with an increasing number of subunits (figure 4.26). Adding an additional subunit leads to a significant increase compared up to leaving that subunit out until 5 subunits are included (pair-wise t-tests: 1 \( \rightarrow \) 2: \( p < 10^{-11} \); 2 \( \rightarrow \) 3: \( p < 10^{-4} \); 3 \( \rightarrow \) 4: \( p < 0.01 \); 4 \( \rightarrow \) 5: \( p < 0.05 \); ANOVA over all n: \( p < 10^{-15} \)). This shows, that the mutually antagonizing effect of two or more subunits is a factor in the emergence of invariance. Concluding, the three factors
Figure 4.24: Example how OC activity emerges. Example OC is the same as in figure 4.20, response to object 1 is depicted. *Left column:* Response of all CCs. Spacing between base-lines corresponds to 3 standard deviations (over all objects) in activity. Colours are used to easily assign each activity trace to its baseline. *Right column:* Weight matrix W of projections of CCs (rows) to an OC's subunits (columns). Weights are pseudo-colour-coded as indicated by the colour-bar beneath W. *Third row from bottom:* activity traces of the subunits before 4th power non-linearity (A(CC)*W,s). *Second row from bottom:* subunit activity (non-normalized as used for OC activity computation). *Bottom row:* OC activity, normalized to mean and std over whole set of objects.
leading to the emergence of OC invariance described in the example of panel 6a (sparse connectivity, antagonizing CCs→subunit projections, antagonizing subunit→OC projections), are typical for the emergence of invariance over all OCs.

**Training and testing in the presence of a distractor**

We have shown so far that optimising the stability objective leads to viewpoint invariant cells when trained on plain stimuli. Do these results persist if a distractor is present as well during training and testing? We train and OCs with the same stimuli and parameters as in the baseline situation (50 objects, 12 views), but add to each presentation a randomly chosen second stimulus as distractor. For testing, the distractors are taken only out of the 50 objects, that are not trained. In the simulation with distractor viewpoint invariance is larger for OCs as compared to CCs for 39 out of 50 objects ($p < 10^{-4}$, sign-test; figure 4.27). This means that the difference between OCs and CCs in viewpoint invariance is not only conserved for training and testing with distractors but gets even more pronounced. This shows that the stability objective is also well suited to suppress distractors solely on the basis of input statistics, a process that would typically be attributed to top-down attention mechanisms.

**Unsupervised clustering**

The stability objective increases viewpoint invariance for individual cells, but does this also facilitate invariant object classification on the population level? In order to test this issue, we perform an unsupervised clustering on both the CC and OC activities. As before OCs are trained on a subset of 12 views per object (30 degree steps). After convergence k-means clustering is performed on the activities in response to all 72 views of all trained objects. In the absence of a distractor, classification performance reaches 79% (10 objects, figure 4.28 left) and 75% (50 objects, figure 4.28 right) for OCs, which is larger than the 65% and 62% reached for CCs respectively. These values are obtained as mean over 10 different random initial conditions for the clustering algorithm and the observed difference between
Figure 4.27: Analogous to figure 4.21c for cluttered condition: Mean OC invariance for individual objects plotted versus mean CC invariance. OCs are trained in the presence of a distractor, both cell types are tested in presence of distractor. All other parameters unchanged as compared to figure 4.21c.

Figure 4.28: Classification performance for kmeans clustering on CCs and OCs using 10 and 50 objects in presence and absence of distractor. Horizontal lines indicate chance level (10% and 2% respectively); error-bars denote standard deviation over 10 repetitions of k-means clustering.
CCs and OCs is highly significant ($p < 0.005$ and $p < 10^{-7}$ respectively, t-test). When training and testing is performed in the presence of a distractor the difference becomes even more remarkable: both for 10 and 50 objects, classification performance increases over 50% (from 38% to 60% and from 14% to 22% respectively) and are both highly significant ($p < 10^{-9}$, t-test). In the simulations discussed so far, a rather large number of views (12) was presented for training the OCs. Do the results persist if this number is reduced? We tested classification performance for OCs trained with 50 objects and 3 different views per object (120° steps). The results do not differ remarkably, still OCs classification performance reaches 71% (50 objects) in the absence of a distractor and 20% in the presence of a distractor. In both conditions the OC performance is still significantly higher than CC performance ($p < 10^{-5}$, t-test). Concluding, even for a small number of training views and in the presence of a distractor, unsupervised classification performance is significantly increased for cells trained with the stability objective (OCs) as compared to their afferents (CCs). Consequently, the stability objective is a valuable pre-processing step for viewpoint-invariant object classification, especially if the object to be classified cannot readily be segmented from the background and can be trained from few viewpoints only.

4.4.3 Discussion

We show that unsupervised training with a general optimisation principle facilitates the classification of real-world stimuli. The facilitation still holds, if stimuli undergo a complex, non-isomorphic, transformation - in our case the change of viewpoint - and only few examples are seen during training. The degree of facilitation is furthermore maintained in the presence of a distractor during training and classification. This result rests on the fact that cells trained to optimise a temporal coherence objective represent the input more invariant to viewpoint than their afferent cells. This result is especially remarkable as it is achieved in a homogeneous architecture: Already the input neurons, modelled as complex cells, can be understood as the temporally coherent features of a natural stimulus sequence (section 4.1). Nevertheless, the degree of invariance still increases for the second layer trained here. Hence, the arguably most difficult step in object classification, namely achieving a representation of the input, which is invariant under the desired transformations, does not require a large set of labeled examples. Instead, training on few unlabeled examples according to a general principle like temporal coherence, generates such a representation. A simple classification algorithm operating on the resulting representation then suffices for successful classification. Classifying real-world objects invariant to complex transformations has also been addressed in other physiologically inspired network models. Mel's carefully designed "SEEMORE" architecture (Mel, 1997) allows very reliable recognition of the trained objects from novel view-points and is robust to various other distortions. However, SEEMORE needs a huge variety of pre-defined feature channels, while we can restricted ourselves to few very simple input 'feature' channels (colour, orientation and spatial frequency preferences of the CCs), which in addition are only coarsely sampled. Although engineered solutions like SEEMORE might reach higher classification rates for specific problems, the appropriate combination with general learning principles like temporal coherence therefore reduces design and computation costs.

Feature-based recognition systems often rely on the precise segmentation of the object from its background. A recent study shows that a network trained by the principle of sparse-coding achieves some degree of viewpoint invariance for classifying real-world objects in cluttered scenes (Wersing & Köner, 2003). Here we show, that the principle of temporal coherence generates viewpoint invariant representations. We do not use a fully cluttered background but just a single distractor. It is hence likely that most information extracted by
our system is still based on the outline of each object, especially since the retinal resolution is low and using fine internal object structure thus difficult. These considerations could be addressed by a complementary segregation based on independent cues, like motion, which are presently not used. Concluding the present study provides an unsupervised learning scheme that is insensitive to the background during both classification and training.

When learning takes places in the presence of distractor, learning of the specific target-distractor combination needs to be avoided. Instead the isolated target has to be learnt. This can be achieved by a variety of segregation mechanisms. Human and non-human primates can achieve segregation by selectively directing attention to the target (Desimone & Duncan, 1995). When the target to be trained is presented with a sufficient number of different distractors, this statistical property can be exploited. Our network model uses the higher temporal coherence of the target relative to the distractor. Hence statistical regularities inherent in the input enable our system to unsupervised learn the relevant representation.

4.5 Discussion temporal coherence

A large and growing body of recent studies applies the temporal coherence principle to achieve invariant representations and to explain some properties of the visual system (Földiak, 1991; Stone, 1996; Wallis & Rolls, 1997; Rolls & Milward, 2000; Kayser et al., 2001; Ehnhäuser et al., 2002a; Stringer & Rolls, 2002; Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2003; Hurri & Hyvärinen, 2003; Wiskott, 2003). These studies as well as those on similar coding schemes (see (Ohlhauser, 2002) for review) are often taken as evidence, that the functional organization of large parts of the visual cortex can be derived from such general principles. In this chapter we first replicated well-known properties of the early visual system and provided quantitative comparisons, where the physiological data is well known. We proceeded to show that temporal coherence can be implemented in a hierarchical architecture and learnt cells that resemble texture cells. As much less is known about the physiology of cells that might code such mid-level properties, we instead compared our results to psychophysical experiments, showing again a good match. The important criterion during evolution and in development of technical applications, however, is whether a general coding principle boosts a system’s performance. Invariant object classification of real-world objects under complex, i.e. non-isomorphic, transformations is thus a critical test case for the above conjecture. Only few of the aforementioned studies, however, fully address this issue: Wallis and Rolls (Wallis & Rolls, 1997) show that cells of their trace-rule based 'VISNET' architecture can acquire translation and viewpoint invariance for face stimuli. Thereby they delivered the proof of concept, that temporal coherence based learning rules in principle can learn viewpoint invariant representations. The analysis of their results is mainly based on a so-called "discrimination factor", which does not exclude that information on the stimulus is lost across their hierarchical levels. While in case of translation invariance the authors control for this possibility by adding a simple supervised classifier to the output of their network, they do not report a similar control for viewpoint invariance. Furthermore, the number of stimuli to be discriminated in their viewpoint invariance experiment is very small (3). Although more recent studies on a similar architecture ("VISNET2") use more stimuli, the authors either restrict themselves to translation invariance (Rolls & Milward, 2000) or do not use real world stimuli (Stringer & Rolls, 2002). Consequently the present study is the first to apply the temporal coherence principle to viewpoint invariant classification of a large number of real-world objects. In addition - unlike (Wallis & Rolls, 1997) - we demonstrate the suppression of distractors and generalization to unseen viewpoints. Hence the present
experiments are the first to show the functional usefulness of a general coding principle like temporal coherence on the system level.
Chapter 5

Eye-Head coordination in a freely behaving cat

Animals use eye movements to shift their gaze relative to the environment as well as to stabilize the retinal image. Under natural conditions the corresponding systems need to interact continuously. Using the eye-tracking setup described in section 2.2.4, we simultaneously record eye and head movements in the freely behaving cat in its natural environment. We find that cats perform only few isolated eye movements and that combined head and eye movements dominate. In about two thirds of combined eye and head movements both point in opposing directions, indicating a slightly dominating compensatory role for eye movements. However, velocities of eye and head cover different frequency ranges and are only weakly correlated. Further analysis shows that the compensatory role of eye movements is usually shadowed by an accompanying explorative function.

5.1 Results

Eye- and head movements were recorded in the freely behaving cat in its natural environment. For the following analysis about 10 minutes of continuous data are used. Head-movements were estimated from the image motion of the front camera. After discarding potentially invalid data points (see Methods) we obtained a robust estimate of head in space motion in 94% of our data. Figure 5.1 (top) shows an example of a continuous trace of horizontal and vertical velocity thus recorded.

The relative position of the right eye was obtained from movies recorded by the eye camera. Slightly more data (16%) had to be discarded as invalid than in the case of head-movements (6%). Most invalid data here resulted from blinking or from grass occasionally obstructing the camera’s view of the eye. Figure 5.1 (bottom) shows an example of a continuous trace of horizontal and vertical eye position temporally aligned to the head-movement data of Figure 5.1 (top). Please note, that - as a consequence of the mirror’s magnification - the setup obstructs less of the cat’s visual field than it might appear.

To validate the automated method for determining pupil position, we compared the positions determined by the algorithm to those of two naive observers in a randomly chosen 10s interval. Correlation in horizontal eye position between the automated and manual method was 0.90, compared to 0.98 between the two subjects. Hence, the presented video-based analysis is a valid method to investigate eye-head coordination.

We sought to assess the relation between horizontal head and eye movements. For the first
Figure 5.1: Five seconds of raw eye-position and raw head-velocity. From top to bottom: horizontal head velocity, vertical head velocity, horizontal eye-position, vertical eye-position. Red dots indicate missing data.

Figure 5.2: Eye head coupling: TOP: The temporal correlation is shown in the top diagram. Absolute values of the horizontal component of the eye velocity are correlated with the absolute values of the horizontal component of the head velocity (red curve). The vertical black dotted line is plotted at the correlation of 0.5. The horizontal black dotted line is plotted at delta $t = 0$. MIDDLE: The coherence between eye and head movements is shown for the frequencies 0 - 12.5 Hz. The red line shows the coherence for the absolute values of the horizontal eye velocity component with the horizontal head velocity component. The green line shows the coherence for the absolute values of the horizontal simulated eye velocity component with the horizontal head velocity component. BOTTOM: The power spectrum gives information about the individual frequency composition of each data set. The power spectrum is calculated for the horizontal component of the eye velocity (red curve), the simulated eye velocity (green curve) and the head velocity (blue curve). All data are normalized.
Figure 5.3: Compensation: A: Pseudo-colour-coded 2D-histogram of head (x-axis) and simulated eye velocities (y-axis) at the same point in time. B: Pseudo-colour-coded 2D-histogram of head (x-axis) and eye velocities (y-axis) at the same point in time with the related colour scale. C: Pseudo-colour-coded plot of the normalized cumulative distribution function from flow field velocity (y-axis) minus 0-200 % of simulated eye movement (x-axis). Subtracting percentages of simulated eye velocity from the flow field velocity results in different velocity distributions which can be seen as retinal slip for the idealized data. D: Pseudo-colour-coded plot of the normalized cumulative distribution function made from flow field velocity (y-axis) minus 0-200 % of eye movement (x-axis). Subtracting percentages of eye velocity from the flow field velocity results in different velocity distributions which can be seen as retinal slip.

In order to detect potentially more complicated regularities we plot the probability of co-occurrence of the velocity of head and eye movements. For the simulated eye movements, by construction, we observe a strict linear relationship (Figure 5.3). Only during the sac-
CHAPTER 5. EYE-HEAD COORDINATION IN THE CAT

5.2 Discussion

Mobile head and eye tracking in the cat gives interesting insights into head-eye coordination under natural conditions. We find a surprisingly weak temporal correlation of head and eye movements. Furthermore, in our data eye movements contribute little if at all to the stabilization of the retinal image.

Studying the relationship between eye-head coordination is mainly done in head-restrained subjects under artificial conditions. However, some studies address an approach towards more natural conditions, giving insight into eye-head coordination (human: (Pelz et al., 2001; Herst et al., 2001; Guittton et al., 1990; Fuller et al., 1983); monkey: (Wellenius & Cullen, 2000; Freedman & Sparks, 1997)). These studies reveal the task-dependence of eye-head coordination but address mostly the question if eye or head precedes a coordinated movement. But still, both studies assess the natural environment and natural circumstances
as crucial influences on behavior e.g. because of the possibility of motor planning. Two food-preparing studies in humans reveal as well the importance of eye movements in motor planning in an everyday task (Land & Hayhoe, 2001). Using these new ideas could also lead via a better understanding of oculomotor behavior as well as eye-head coordination to a better understanding of visual perception.

Our results give critical insight into the coordination of head and eye movements made under natural, unrestrained conditions. In our former study we examined the different eye movement systems showing natural stimuli (Moeller et al., 2004). We presented natural movies and natural pictures taken by the cat to head-restrained humans and cats in a laboratory setup. We found that for cats the natural pictures, with a superimposed linear global motion, are a more effective stimulus than gratings for eliciting stabilizing eye movements. In addition we superimposed the natural movie with a global horizontal movement but we couldn’t elicit stabilizing eye movements. This was interpreted as the inability of the cat to compensate for the missing vestibular input. In the light of the present results, we have to conclude that under natural conditions the role of stabilizing eye movements might be overrated.

We find here that eye-head coordination is highly complex, involving compensatory eye movements (i.e. those made in the opposite direction to head movements) as well as exploratory movements that are poorly correlated with head movements. The latter probably include scanning eye movements superimposed upon compensatory eye movements. (Fuller et al., 1983) supports this idea that only the beginning and the end of head movement is compensated by the eye via the vestibular-oculomotor interaction. This is then reflected not only in the weak correlation but also in the different frequency domains covered by eye and head.

While our results suggest that eye-head coordination exhibits a high degree of variety and complexity, one clear emerging feature is that the two movement systems work in a complementary fashion. Eyes are responsible for rapid and dynamic movements and the head moves more slowly and provides bigger displacements. This would to some extent account for the poor correlation and coherence between head and eye movements because of their different frequency domains. It may be that a more sophisticated analysis of eye-head coordination during a predicted, trained behavior sequence would reveal a greater degree of correlation or coherence. But this is a different paradigm leading away of the interest in natural, unrestrained behavior. Only when we examine visually guided movements under the natural and unconstrained conditions described here will we begin to refine our definitions of head and eye movements and eye-head coordination to reflect the movements that are made under natural conditions. These considerations further emphasise the necessity to use more natural paradigms in order to understand natural behavior.
Chapter 6

Bottom-up Overt Attention: Correlations and Causality

The “classical” saliency map model (Koch & Ullman, 1985; Itti & Koch, 2000) states that human overt attention in natural scenes is guided by contrasts of low-level image features, such as luminance, colour, orientation or motion. The spatial maps of these feature-contrasts are then added linearly to yield the “saliency map”, whose maximum defines the next location to be fixated. Although this model quite successfully predicts fixation behaviour on the system level, little is known on the actual causal contributions of each individual features. In this chapter we address the causal contribution of luminance-contrast to the saliency map in humans and monkeys. We in addition briefly touch the issue, how different features, colour and luminance in our case, interact in the guidance of human overt attention.

6.1 Specific methods

Eye-tracking data for luminance contrast experiments were recorded at the University of Bochum with the “Eyelink” system (human, see section 2.2.2) or the seleral search coil (monkey, section 2.2.3). Eye-tracking data on colour images were recored using the Dr. Bouis oculometer system (section 2.2.1). For the former experiment the backprojection setup was used, while for the latter as well as for experiments not involving eye-tracking, the G4/CM772E setup was used (see section 2.3 for details).

6.1.1 Local modifications of luminance-contrast

All stimuli used in this study are based on the static images described in section 2.1.2. As a measure of contrast we follow the definition of (Reinagel & Zador, 1999): contrast equals the standard deviation of the luminance within a square image region divided by the mean intensity of the image. This definition canonically extends the two-point contrast and is numerically robust. Unless otherwise stated the length of the square region was chosen to be 80 pixels. For some aspects of later analysis, images were low-pass filtered. This was done by applying a 2-dimensional Fourier transform to the image, multiplying it with a Gaussian kernel centered at the DC-component and transforming the result back to image space. The width (HWHM) of the kernel will be referred to as cut-off-frequency $f_{cut}$. As the filtering suppresses the contribution of high-frequency components the total contrast
is reduced. Note, that low-pass filtering was performed for analysis only, and in all cases stimuli were presented at full resolution.

In order to modify contrast within a stimulus without introducing intensity boundaries, we applied the following procedure: Five points \((x_i, y_i)\) were randomly chosen in the image, excluding points closer than 160 pixels to the image boundary. Centered at each of these points a two-dimensional Gaussian \(G_i(x, y) = e^{-\frac{(x-x_i)^2+(y-y_i)^2}{\lambda^2}}\) with \(\lambda = 80 \text{pixels}\) was placed. If any of the points were closer to each other than 160 pixels (i.e. 2\(\lambda\)) different random points were selected. Taking the maximum over \(G_i\) resulted in the mask \(G(x, y) = \max_{i\in\{1...5}\} G_i(x, y)\). At each image point the original pixel intensity \(I_0(x, y)\) was then modified to

\[
I(x, y) = I_0(x, y) + \alpha G(x, y) (I_0(x, y) - \langle I_0 \rangle)
\]

where \(\langle \cdot \rangle\) denotes the mean over the image and \(\alpha\) the peak contrast modification level. If \(I(x, y)\) exceeded the 8-bit range of possible pixel values, the result was cropped to the maximum respectively minimum possible value. By this procedure local contrast around the points \((x_i, y_i)\) was increased or decreased, while minimally affecting overall intensity and avoiding the introduction of artificial "object" boundaries. For the experiments described below contrast modification levels from -0.6 to +1.0 were used (figure 6.1).

For stimulus presentation the 972 stimuli (108 basis images times 9 modification levels) were split into 18 presentation blocks. The blocks were balanced such, that each basis image occurs once per block. One half of the basis images was used for even blocks, the other half for odd blocks only. Each modification level was used exactly 6 times per block. Within each block the order of stimuli was random. Each stimulus was presented for 6s, stimuli were interleaved by at least 0.5s of blank time.
Figure 6.2: Stimulus example of the modification detection paradigm: the whole image is subjected to the same type of modification (i.e., 5 locations are contrast modified) as in the eye-tracking experiment; one half is then replaced by the unmodified image; modified and unmodified part are separated by a overlaid bar. In the example there are three modifications (+100%) visible in the left part of the image (above the bifurcation of the black tree, at the uppermost corner of the rock, at the birch tree at about one third of the image below the top).

Subjects

Two male monkeys (*macaca mulatta*) were used for this experiment. Both performed all 18 blocks in both large and small screen condition. To exclude possible memory effects as much as possible, between large screen and small screen sessions more than five months elapsed. In between monkeys were performing tasks unrelated to the present experiment; especially no naturalistic photographs were presented to them during this period. Monkeys received a drop of water or apple-juice after each stimulus presentation, irrespective of the eye-movements they performed.

A total of 13 human subjects (undergraduate students at university of Bochum, 6 male, 7 female) participated in the experiment. Six subjects performed 6 blocks each in the large screen condition, six different subjects performed 6 blocks in the small screen condition. In both conditions, two subjects performed blocks 1 to 6, two subjects blocks 7 to 12 and two subjects blocks 13 to 18. This yields equivalent amounts of data for humans as for monkeys. One additional subject (SH) performed all 18 blocks for large screen condition; from her data blocks 3 to 6 are missing due to setup failure.

6.1.2 Modification detection paradigm

In order to measure the degree to which contrast modification is detectable by subjects the following two-alternative forced choice protocol was performed on the 8 of the basis images: Modified stimuli were generated as in the eye-tracking paradigm, i.e., contrast modifications of identical peak modification level were applied in 5 randomly chosen locations across the image. Then one randomly chosen half of the image was replaced by its unmodified version. To avoid potential boundary cues between modified locations in one part and their unmodified counterpart, the midline of the complete stimulus was covered by a 60 pixel wide gray bar (Figure 6.2). Additionally, 50% of the resulting stimuli were mirror-reversed on the vertical midline. Subjects were instructed to indicate by a button press, which part of the stimulus was contrast-modified. The stimulus was shown until the subject
made its decision and then followed by a 0.5s blank. At no time templates of unmodified images were presented. Therefore, subjects had to rely on general "knowledge" how natural images look like to detect the contrast manipulations. The extreme modifications of -60% and +100% are described as foggy areas and very high contrast structures. Subjective visibility of the moderate contrast modifications was low, and subjects reported that they were often guessing which side contained the modification. However, no feedback was given to the subjects on the correctness of their decision. Five instances of each of the 16 image versions (8 images + 8 reversed) and each of the 9 contrast modifications were presented in random order, yielding a total of 720 stimuli. For each stimulus presentation, correctness of response as well as reaction time were recorded. To account for inter-subject differences in total reaction time, reaction times were normalized to unit standard deviation within each subject. Four human subjects (three male, one female, age between 23 and 41) participated in this experiment.

6.1.3 Global modification of colour

For the experiment on the interaction of colour and luminance, 12 out of the 108 static images are used. Images are presented in four different conditions (figure 6.3): the original colour image, a grayscale (just luminance) version and two versions which are globally modified in colour. For the latter the azimuth in DKL-space of each pixel is changed by 90°, while luminance and elevation are kept constant. If by this procedure pixels of an image fall outside the gamut of the presentation screen, the saturation (vectorial sum of luminance and elevation) is reduced for that pixel in all conditions. Thereby we ensure that all features besides the azimuth, such as luminance, luminance contrast, but also colour-contrast is identical across conditions. In each experimental session, each condition of each image was shown once for 6s, yielding a total of 48 different stimuli per session. The order of stimuli within each session was random. Eight subjects participated in this experiment, each performing 10 sessions. Eye-tracking was conducted using the Dr Bouis oculometer and the G4/GM772E presentation setup.
6.1.4 Analysis of eye-traces: 'actual' vs. 'control' fixations

When measuring the effect of any image property (contrast, contrast-modification, etc.) on fixation we apply the following method: For each image we compute the median value of the property at the fixations of the subject in that image. We denote this result as 'actual', since it is obtained on the fixations that actually belong to the that particular image. For comparison we take the median of the same property in the same image at fixations obtained on all other images. This median provides an unbiased estimate of the value the property would take in the actual condition, if the property had no influence on fixation. Hence we denote this condition as 'control' condition throughout this chapter.

6.2 Results

We measure the eye-position of human and monkey observers while they are viewing greyscale pictures of natural scenes (figure 2.3). During the course of the experiment 108 images are shown unmodified as well as after being locally modified in luminance-contrast (figure 6.1). While the unmodified images are used to measure correlations between fixation and luminance-contrast, using the modified images we can probe the causal component of luminance-contrast in attracting overt attention.

6.2.1 Unmodified stimuli

Each experimental block of 54 stimuli includes 6 images that were not modified. To investigate, whether there is any correlation between fixation probability and luminance-contrast, we first analyse data from those stimuli. For each of the images we compute the contrast distribution on the fixation spots in that image (green dots in figure 6.4a). This is then compared to the contrast in the same image along all fixations recorded in the same subject on all unmodified images (red dots in figure 6.4a). This 'control distribution' resembles an unbiased estimate of the contrast along the fixation points if fixation were entirely independent from contrast. For the example image of figure 6.4a, the control distribution is depicted as red histogram in figure 6.4b. Its median is lower (0.21, red line) than the median of the actual fixations (green data-points, median: 0.24, green line). If increased contrast had an attractive effect on fixation, this relation should hold for the majority of images. Analysing all unmodified images for the same subject (monkey N) and the same screen size (large screen), shows that actual contrast indeed exceeds control contrast in more images (65) than vice versa (43 images, figure 6.4c). Pooling the data within species and within screen size (figure 6.4d), shows that this effect is highly significant for large screen presentation in both monkey (p < 0.01, sign-test) and human (p < 0.001, sign-test). However, for the small screen no significant effect is found for neither human (p > 0.09) nor monkey (p > 0.48). Since in the large screen condition the spatial frequencies are lower, we check whether the observed difference between presentation conditions is a consequence of a dependence on spatial frequency. When applying a low-pass filter on the images before analysis, we find highly significant effects of contrast also for the small screen condition (figure 6.4e). This correlative effect of contrast is preserved for a broad range cut-off frequencies of the low-pass filter in both screen size conditions. It drops to chance level only for very low cut-off frequencies (< 10 pixels^{-1}), which contribute only little to local image structure. Concluding, the results on unmodified images show a clear correlation between fixation and luminance-contrast across a broad spatial frequency range for both screen sizes and species.
Figure 6.4: (a) Green circles: fixations of one subject (monkey N in large screen condition) on the unmodified image shown in the background ('actual' fixations). Red dots: Fixations of the same monkey on all other unmodified images shown in the course of the experiment ('control' fixations). Only the 600x400 pixels wide central region used for analysis is shown. (b) Green dots: contrast at actual fixations of panel a. Red histogram: contrast distribution at control fixations for the image of panel a. Vertical lines indicate the medians of these actual and control contrast distributions. (c) Contrast-Medians of all images for monkey N, control condition plotted versus actual condition. Example of panel a & b marked in blue and indicated by green and red line. Points below diagonal indicate that contrast is higher at fixation points than expected if contrast did not affect fixation. (d) Number of trials, for which contrast median at fixation points is higher (green bars) or lower (red bars) than predicted by control condition. Data are pooled over trials with same species and same screen size. Arrows from panel c illustrate where its data enter (monkey, large screen). Significance values from sign-test. Note, that the total number of trials analysed is slightly lower than the number of trials performed, since in some trials no actual fixation was within the analysed image region. (e) Dependence of contrast influence on spatial frequency. Vertical axis: Percentage of trials, for which contrast median at fixation points is higher than predicted by control condition (chance at 50%), horizontal axis: cut-off frequency of low-pass filter applied for contrast calculation of the images. Note, that the eye-tracking data is the same along the horizontal axis, low-pass has just been applied for analysis. Data are pooled over trials with same species and same screen size (monkey, small screen: cyan diamonds; monkey, large screen: blue dots; human, small screen: magenta stars; human, large screen: red circles). Symbols are in grey where individual sign-tests do not show significant difference from chance (50%), i.e. p > 0.05. Data of panel (d) plotted to the right (corresponds to infinite cut-off frequency)
6.2.2 Modified stimuli

Having confirmed the correlation between fixation and luminance-contrast in humans and monkeys, we continue to address to what extent this effect is causal in both species. This is done using images with locally increased or decreased contrast at a total of 8 different peak modification levels (see methods for details). A causal effect of contrast would imply that positive modifications attract attention, while negative repel attention. Figure 6.5a shows the fixations of a single subject (monkey N) in the analyzed region of an image that is contrast reduced at -40% peak modification level. The modification levels at these 'actual' fixations are then compared to the modification levels of the same stimulus at control fixations. Control fixations are given by all fixations of all experimental trials of the same subject (within the analyzed region). Thus they are independent of the current image and provide an unbiased estimate of the distribution of fixations, that would occur, if modifications had no influence on fixation. For the example of figure 6.5a this procedure is illustrated in figure 6.5b. Green points resemble the modification at the fixation points as shown in figure 6.5a. The red histogram resembles the distribution of modification levels of the same stimulus at all control fixations. The medians of both distributions are depicted by vertical lines. If the modifications had no influence on fixation the medians should (on average over trials) not differ. If modifications repel fixation, the actual median (green) is closer to 0 than the control median. If modifications attract fixation the actual median is farther from 0 than the control median, as it is true for the example of figure 6.5a/b. In the example the negative modifications have an attractive effect on fixation, thus contradicting a causal role of luminance-contrast. In order to quantify this effect, we define the saliency of a modification as the difference between actual and control median divided by the peak modification level. In the example this value equals \( \frac{70.144(-0.095)}{-0.4}=0.12 \). By construction a positive saliency implies at all (positive and negative) peak modification levels that modifications are attractive for fixation. If luminance-contrast was a causal factor to shifting overt attention, one should observe negative saliency values for negative peak modification levels and positive otherwise. Pooled over images and subjects at constant peak modification levels within species and screen size, however, one obtains the contrary result: mean saliency is positive at all measured peak modification levels for both species and both screen sizes (data-points in figure 6.5c). In humans, this confirms the previous finding that luminance-contrast has no causal role in attracting overt attention. Additionally we show here that the same result also holds for monkeys.

For natural scenes there does not seem to be a causal effect of luminance-contrast on saliency. However, for simple stimuli like a dot on an otherwise uniform background, that will attract overt attention, there is no other cue but contrast available. This leads to the assumption that in natural scenes causal effects of contrast might be shadowed by the object content of the image. If the modifications as such can have such object-like saliency, this effect should be independent of the sign and the strength of the modification. Consequently such an 'object-ness' of modifications shall kick in at a certain threshold and then add a constant offset to the saliency of the modification. Consequently we model the total effect of the modifications on saliency as combination of two effects: First we assume a residual linear dependence of saliency on contrast (figure 6.5d, left). The slope of this linear relation, \( "a" \), is a measure of the causal effect of luminance-contrast. Second we model the object-ness of the modifications by a V shaped influence on saliency (figure 6.5d, right). The asymptotic value of this function, \( "b" \), models the maximum influence of this object-ness on fixation. The width of the function, \( "c" \), provides a measure for the threshold modification level at which modifications reach object-like quality. The total saliency of a modification is then modelled as sum of both functions. We fit this model to the measured mean saliency (figure
**Figure 6.5:** (a) Green dots: actual fixations of monkey N on the stimulus shown as background (large screen). White lines indicate modifications analogous to figure 6.1. Only analysed central region is shown. Peak modification level in the example is $\alpha = 40\%$.

(b) Green dots: contrast modification level along actual fixations for example of panel a. Red histogram: distribution of modification for stimulus of panel a at control fixations (All fixations during all other trials of same subject and screen size). Vertical lines indicate medians of distributions. (c) Data points: measured saliency (Difference between median modifications at actual fixations and control fixations divided by peak modification level) averaged over stimuli and subjects. Error-bars denote standard error across subjects. Solid lines: saliency model best fitting (in $\chi^2$-sense) the data. (d) Illustration of the saliency model: the causal effect of contrast (linear in $a$) is added by an $V$-shaped effect of the modifications reaching object like saliency as such. (e) Values of fit parameters for different species and screen sizes (colours as in panel c). Error-bars denote confidence intervals obtained from the covariance matrix of the fit (see methods for details).
6.5c, solid lines). The c2-values obtained from the fits (given the known standard error across subjects) range from 0.26 (monkey, large screen) to 5.25 (human, small screen). This corresponds to Q-values between 26.3% and 99.2% which indicates a good prediction of the model. This result implies that the measured saliency is very well compatible with a linear attractive effect of contrast that this fully overruled by object-ness over the measured range. The model-parameters furthermore provide quantitative hints to differences between species (figure 6.5e): The slope of the causal contrast influence (fit-parameter a) is larger for monkeys as compared to humans for both small screen (0.14±0.03 compared to 0.09±0.01) and large screen (0.11±0.03 to 0.07±0.01) presentation condition. This suggests that the residual causal effect of contrast is stronger in monkeys that in humans. The effect of object-ness (parameter b) is slightly larger in monkey than in human (small screen: 0.13±0.02 vs. 0.09±0.01; large screen: 0.07±0.02 vs. 0.05±0.01). On the other hand the threshold peak modification level (fit parameter c) at which modifications as such become relevant is lower for monkeys than for humans (small screen: 0.21±0.04 vs. 0.28±0.05; large screen: 0.12±0.09 vs. 0.30±0.06) Both suggests that monkeys might be more susceptible to use the modifications as such for directing overt attention instead of real-world objects than humans. In summary, the analysis of the fit-parameters point to a difference between monkeys and humans: Although in either species any causal effect of contrast over a broad range is shadowed by higher level cues, the residual effect seems to be stronger in monkeys than in humans. In both species this shadowing is likely to arise not only from the object like quality (object-ness) of the modifications, but also from objects inherent in the natural scene. In order to quantify the difference between humans and monkeys further and independently of any model assumptions, we directly compare the influence of the modifications between species on a stimulus by stimulus basis. (Note, that exactly the same stimulus-set was used for humans and monkeys). Using the data from the 100% peak-modification trials in large screen presentation condition, we find that the mean modification level along the actual fixations of monkeys was larger than that for humans in 69 out of 108 stimuli (figure 6.6a), a fraction which is highly significant (p < 0.01, sign-test). The fact that the modification level at monkey fixations is larger than at human fixations for more images than vice versa is valid for both screen sizes and at all peak modification levels, with the sole exception of -20% peak modification level in the large screen condition. This means for 15 out of 16 (2
screen sizes times 8 non-zero modification levels) modification conditions, monkeys are more likely to be attracted by increased contrast. Consequently this results in highly significant difference between monkeys and humans when pooling the above analysis over all non-zero modification levels ($p < 10^{-3}$, sign-test). For both positive and negative modifications monkeys thus tend to fixate at higher contrast modification levels than humans. This is independent whether this higher modification level results from a positive modification at the fixated location or from negative modifications at a different, non-fixated location. Concluding, although the causal effect of contrast is shadowed in both species by higher-order effects, monkeys are more causally driven by contrast than humans. Comparing the contrast at fixation-points in unmodified images, however, does not reveal any significant difference (figure 6.6b; $p > 0.63$ and $p = 1$, sign-test for large and small screen respectively). Taken together, these results imply that humans substitute the causal use of contrast by higher order cues, that - in unmodified natural scenes - are well correlated to the original contrast cues.

### 6.2.3 Detection paradigm

The fact, that moderate contrast modifications do not attract attention, but strong modifications of either sign do, raises the question whether the moderate modifications are in a relevant range, i.e. available to the subject on the system level. This issue is addressed in a two-alternative forced choice detection paradigm in which the side containing contrast modifications has to be detected (see methods). These measurements reveal that already a ±20% modification - as all considered non-zero modification levels - can be detected above chance level. Furthermore, it shows that -40% modification - which significantly attracts the gaze - has similar detection probability (78%±6% correct) as +80% modification - which is the maximum positive modification that does not significantly attract the gaze (81%±9% correct, Figure 6.7a). The analysis of reaction times reveals a similar pattern as the detection probability. The normalized reaction times are slightly smaller for ±20% modification than at 0% and about identical (2.9±1.1 vs. 3.1±1.2) at -40% and +80% (Figure 6.7b). As non-gaze-attracting positive modifications show similar detection and reaction-time patterns as gaze-attracting negative modifications, the increased fixation probability at decreased contrast is unlikely to be based solely on a better detectability of the negative modification. This finding does not exclude, however, that the better detectibility partly contributes to the increased fixation probability at highly decreased contrasts, as this unnatural modification
might consciously attract the attention of the subjects. As the main finding, the modification detection experiment shows that the result, that gaze is not attracted by increased contrast, cannot be explained as consequence of a lack of detectability, i.e., availability on the system level.

### 6.2.4 Interaction of luminance and colour

In a separate experiment we investigate the interaction of colour and luminance-contrast in natural and globally modified natural scenes. We compare the luminance-contrast along fixations in naturally coloured scenes to grayscale scenes, which are identical in luminance. For 42 trials fixated luminance-contrast is higher in the grayscale images than in the coloured images, while the opposite is true in only 21 trials (figure 6.8). A sign-test reveals that this difference is significant ($p < 0.05$). Luminance-contrast is thus more relevant in guiding overt attention, when colour is not present. This indicates that natural colour-content can overwrite even the correlative effect of luminance-contrast. Is this a consequence of the low-level feature of colour itself or of higher-order image properties associated with colour? If the former was the case, one would expect the effect to be identical for naturally coloured scenes and for colour rotated scenes. One finds, however, no such significant effect between grayscale images and colour rotated images. On the contrary, there is a significant effect between colour rotated and naturally coloured scenes (table 6.1). Hence, colour is found to only overrule the effect of luminance-contrast if it occurs naturally. This suggests that not colour itself, but higher order image properties associated to colour affects human overt attention.
<table>
<thead>
<tr>
<th></th>
<th>Natural</th>
<th>Gray</th>
<th>DKL cw.</th>
<th>DKL ccw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>X</td>
<td>33%*</td>
<td>29%*</td>
<td>40%*</td>
</tr>
<tr>
<td>Gray</td>
<td>67%*</td>
<td>X</td>
<td>53%</td>
<td>56%</td>
</tr>
<tr>
<td>DKL cw.</td>
<td>71%*</td>
<td>47%</td>
<td>X</td>
<td>56%</td>
</tr>
<tr>
<td>DKL ccw</td>
<td>60%*</td>
<td>44%</td>
<td>44%</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 6.1: Percentage of trials in which luminance contrast is larger in the condition indicated in the row than in the column. * denotes significance at p < 0.05, sign-test

6.3 Discussion

6.3.1 Luminance contrast & Saliency

Here we show that - despite the observed correlation between contrast and fixation probability - high contrast by itself does not causally attract overt attention in natural scenes. As revealed by the detection paradigm, this is true, although the contrast information is available on the system level. In natural scenes luminance-contrast thus does not measurably contribute to the saliency map for human overt visual attention.

The idea of saliency maps originated in the search for mechanisms that can quickly select possibly interesting locations in complex scenes (Koch & Ullman, 1985; Desimone & Duncan, 1995). Purely bottom-up driven saliency maps dramatically reduce the demand of computational resources and have been proven to perform efficiently even in very complex natural scenes (Itti & Koch, 2000). The relevance of the saliency map notion thus often derives from their human-comparable performance even on such difficult stimuli. On the other hand, for the question on the neuronal substrate of saliency maps, the comparison has to be extended from the system level towards the individual processing stages.

Previous studies address the correlation between contrast and human fixation probability in natural scenes (Reinagel & Zador, 1999; Krieger et al., 2000). Over a certain spatial-frequency and temporal range our results confirm their finding that luminance-contrast in (unmodified) natural photographs is higher at attended locations than along random eye-traces. The fact that both studies find this effect without low-pass filtering already on their intrinsic - however arbitrary - spatial scale may partly be attributed to technical differences: the lower resolution of their images, their choice of images and to the patch-size contrast was computed in. Neither of the cited studies, however, has performed an explicit analysis of the influence of spatial scale. The reduced contribution of high frequencies to saccade direction found in the present study seems reasonable, as saccade-preparation has to be based on information from higher retinal eccentricities and can thus rely on only reduced spatial-resolution. (Parkhurst et al., 2002) explicitly model the contributions of different low-level stimulus properties (luminance, colour and orientation) to a saliency map. The predictions of their model excellently match their results on human subjects. For natural scenes they find that - among the proposed stimulus properties, luminance-contrast is most salient. Like the studies above, it addresses only the correlation of the low-level features to attention, but does not investigate, to which degree they causally determine attention. Thus, it remains open whether the observed predominance of luminance-contrast actually derives from the low-level property itself or from higher level properties that happen to be correlated to both - contrast and attention. This distinction between correlation and causality, however, is decisive to draw conclusions on the underlying neuronal substrate and is therefore dealt with in the present study.
In order to circumvent the problem of confounding higher order properties of complex stimuli with properties of the selection process, most studies on attention rely on more simplified stimuli. Typically stimuli consist of several identical objects, of which one is changed in at least one property (Treisman & Gelade, 1980). This change implies a locally increased feature-contrast and leads the object to 'pop-out' from the background (Nothdurft, 1993). This is also true in the case of increased luminance-contrast. Since it is not obvious whether results obtained on artificial pop-out stimuli can directly be transferred to more natural conditions, it does not, however, contradict the present findings. Indeed, for the particular question on the relative contributions of different stimulus properties to a saliency map, the type of image profoundly influences the results (Parkhurst et al., 2002). In the present study natural stimuli are used and modifications are applied such that they are globally within the range of natural variations. Therefore we consider our experiments on natural stimuli well suited to address the contribution of a particular stimulus property under realistic conditions.

As the image type is crucial, care also needs to be taken that the stimuli closely resemble a "natural" environment. Our static photographs, that resemble stills from the catcam videos, well fulfill this requirement. In particular - unlike for example artistic photographs, street scenes or pictures of urban environment - the present stimuli rarely imply an obvious segregation into clearly defined verbally describable objects. This, along with the fact that the general pattern of contrast (in)dependence does neither change in the course of the experiment nor during viewing an individual stimulus, we are confident that the selected stimuli do not induce an implicit task to the subject. As by the instruction to "study the images carefully" no explicit task is provided either, the possible task dependent (top-down) influences are reduced as much as possible. Nevertheless, we find that the most direct bottom-up explanation of a saliency map largely based on luminance-contrast is not consistent with our data. Therefore, even if neither an explicit nor an obvious implicit task is involved, top-down processes seem to be dominant for the direction of human visual attention.

Various brain regions have been found to encode stimulus saliency (Posner & Petersen, 1990; Robinson & Petersen, 1992; Kustov & Robinson, 1996; Thompson et al., 1997; Gottlieb et al., 1998; Horwitz & Newsome, 1999; McPeek & Keller 2002). All of the suggested areas, higher cortical regions, like the frontal eye field or the lateral intraparietal area, as well as sub-cortical structures, like pulvinar or the superior colliculus, are strongly interconnected to visual areas. Recent studies measuring (Lee et al., 2002) and modeling (Li, 2002) the contribution of early visual areas to saliency map mechanisms also build on pop-out stimuli. Nevertheless, despite the amount of data where saliency is represented, the origin of this representation under natural conditions remains unresolved. Exploiting the different sensitivities to luminance contrast of different cortical areas, however, our results can shed some light on the latter issue: Neuronal responses in the early visual system are strongly sensitive to luminance-contrast as tuning curves to most low-level features scale with luminance-contrast (Tolhurst et al., 1981; Carandini & Heeger, 1994). While proceeding through the visual hierarchy, on the other hand, responses become more and more invariant to contrast (Avidan et al., 2002). Our finding that saliency does not causally depend on contrast thus yields the conclusion that -for natural scenes - the saliency map for overt visual attention does not originate in early cortical visual areas.

6.3.2 A modified saliency map model

In reply to our study preliminary to the data presented here (Einhäuser & König, 2003), a recent modeling study (Parkhurst & Niebur, 2004) has extended the "classical" saliency model (Koch & Ullman, 1985; Itti & Koch, 2000) to account for our observations. The
authors argue that a bottom-up saliency map can well account for our results, if one assumes a second order (or texture) contrast in addition to the usual (first order) luminance contrast. Using the assumption that texture contrast has an about 10 times stronger effect than luminance-contrast, their model fits both our eye-tracking data and our detection paradigm data extremely well. Consequently this model supports well the idea that luminance-contrast has little to no causal role in attracting overt attention. In addition it provides a hint, what the nature of the higher order properties that attract attention could be. One should note, however, that a causal role of texture contrast has yet to be shown, e.g. by using appropriately modified images.

In spite of using our data to build their model, (Parkhurst & Niebur, 2004) also question the general methodology employed here. Most of their concerns were related to small number of images (8) used in our study (Einhäuser & König, 2003), an issue that has been overcome with the experiments presented in this thesis. Given that each subject sees each basis image only 3 times (humans) or 18 times (monkeys) among a total of up to 1944 trials, also makes memory effects unlikely to play a major role. The most important concern, however, questions the use of local modifications as such. (Parkhurst & Niebur, 2004) argue that not only luminance-contrast is locally modified, but also e.g. local spatial frequency content. In the analysis presented here, we summarized all these potential influences in our 'objectness' measure. A closer analysis, how this objectness can be dissected into various higher order features will remain an interesting issue for further research.

6.3.3 Interaction of different features

The classical saliency map model (Koch & Ullman, 1985; Itti & Koch, 2000) treats its different low-level features independently and assumes a linear interaction between them. The data on colour images however, suggests, that there is crosstalk between the features: the effect of one feature can be influenced by the presence of another feature. The fact that this effect is only present when the feature occurs naturally points again to the importance of higher level scene interpretation as compared to the low-level feature as such. Further investigation on feature interactions will thus be a key step towards understanding the nature of the higher order properties that guide overt attention.

6.3.4 Inter-species differences

For the role of luminance-contrast in overt attention we showed a difference between humans and non-human primates: the causal effects of contrast are stronger in monkeys than in humans, while the correlative effects are similar. This indicates that in humans the stimulus-driven (bottom-up) influences on overt attention are overruled by prior knowledge on natural stimulus statistics. This prior knowledge then acts on overt attention through top-down signals and overrules the currently present bottom-up information. This substitution of bottom-up by top-down signals plays a larger role in humans than in monkeys. Since rhesus monkeys are a typical model for attention, our results underline the care that has to be taken when applying data from the non-human primate to human visual processing.
Chapter 7

Top-down Overt Attention: Perceptual Rivalry

In the previous chapter we saw that even in conditions which do not involve specific tasks, a simple bottom up saliency map cannot explain fixation behaviour of human and monkey subjects. Instead top-down scene interpretation seems to play a major role. Since there is apparently no possibility to rule out top-down effects for prolonged viewing of a natural scene, we in this chapter use a complementary approach. We measure the interaction of top-down scene interpretation and overt attention in the absence of bottom-up signals, by using a rivalrous figure, i.e. a stimulus which induces different percepts while itself remaining unchanged.

7.1 Specific methods

7.1.1 Experimental setup

Stimuli were generated and presented using the G4/CM72E setup as described in section 2.3. Three different illumination levels were used, which will be referred to as 'white' (79.8 cd/m²), 'gray' (0.71 cd/m²) and 'black' (< 10⁻³ cd/m²) throughout the paper. Eyeposition was recorded using the Dr Bouis oculometer (section 2.2.1).

7.1.2 Experimental protocol

The experiment consisted of five distinct blocks: First a text-based trial; second a block of three trials using differently sized biased (non-ambiguous) Necker-Cubes; third the main block of three trials using differently sized unbiased (ambiguous) Necker-Cubes. After the block with unbiased cubes, a second block of three biased cube trials was performed and finally a second text-based reaction time trial concluded the experiment (Figure 7.1). Each trial was preceded by a Calibration trial.

Text-based trials

The words "LEFT" and "RIGHT" were displayed alternating (20 presentations of each stimulus per trial) at the center of the screen at a font height of 12 pixels in white on black background. Subjects were instructed to press the according button of a two-button mouse.
The time for each stimulus presentation was independent of the subject’s response and drawn from a normal distribution with mean 3s and standard-deviation 1s and truncated not to fall below 0.5s.

Unbiased cube trials
For the unbiased cube trials, Necker-Cubes of three different sizes (size of faces: 100, 200 and 300 pixels) were used. Edges were white and 2 pixels wide and presented on a black background. The angle between the "diagonal" edges to the horizontal was 45 degrees and diagonals were half the length of the horizontal and vertical edges. The cubes were placed such that the top right corner of the lower left face and the lower left corner of the upper right face were at equal distance from the screen center. The percept assuming the lower left face in front will be referred to percept I, the percept of the upper right face in front will be referred to percept II throughout this paper.

Each cube-size was used for exactly one trial, the sequence of the sizes was random. To avoid effects possible arising from the anticipation of the end of a trial, the duration of each trial was drawn from a normal distribution of mean 120s and standard deviation 20s. Subjects were instructed to indicate their percept by pressing the left mouse button when switching to percept I, the right button for percept II. To avoid biasing subjects to a specific strategy, no explicit reference to faces, edges or corners were made, but the instruction instead was done with pictures of the biased cube. Subjects (2 out of 9) who consistently used the buttons the other way (in both biased and unbiased cube trials), were included in the analysis, whereas one subject (1/9), whose button usage was not consistent throughout the biased cube trials, was excluded from analysis.

Biased cube trials
For the biased cube trials, the same cube sizes and locations of the cubes were used as for the unbiased cubes. The biased cubes differed from the unbiased cubes only in that the edges invisible in the suggested percept were displayed in gray instead of white. In each trial, both "percepts" were alternating displayed (10 presentations of each percept per trial). Presentation duration of each cube was computed analogously to the reaction time trials.
The order of cube sizes was random: each size was used once in trials 1-3 (before unbiased trials) and once in trials 4-6 (after unbiased trials); arrangements that would present biased and unbiased cube of the same size in trials immediately following each other were excluded. Subjects were instructed to indicate the percept shown by pressing the appropriate mouse button.

**Blinking correction**

Eye-blinks result in a stereotypical signal, which is reliably detectable in the eye-position recordings. Each blinking results in a 200ms period of unreliable data, which is treated as missing data for analysis.

### 7.1.3 Data analysis

Biased cube trials were used to determine the reaction time of each subject from the change in percept to the button press. The rare cases in which subjects failed to press the correct button before the next presentation of the other "percept" of the biased cube, were excluded from that computation. For the unbiased trials the reaction time thus obtained was subtracted from each time-point of button press. The resulting time-points will be referred to as times of perceptual switch for unbiased cube trials throughout the paper. For biased cube trials and text-based trials the analogous time-points are the times of the onsets of either of the alternating stimuli. Therefore these onset times are to be understood as times of perceptual switch for all analyses that are described for unbiased trials only but also performed as control for biased cube trials and text-based trials.

In order to compare the distribution of percept-durations of the unbiased cubes across subjects, the times between subsequent perceptual switches are divided by the trial mean for that percept. This yields the normalized perception duration for each percept, i.e. each inter-switch interval. The distribution of this normalized perception durations over all subjects (which by construction has mean 1), was fitted by gamma distributions and normal distributions of mean 1. The similarity of distributions was assessed by a two-sided Kolomogorov-Smirnov (KS)-test.

If subjects pressed the same button twice in an unbiased cube trial without pressing the other button in between, the period from the button press before this "double press" to the button press after this double press was excluded from the analysis. This correction had to be applied very rarely and was to ensure that only data was used, for which subjects reported their percept unambiguously.

To analyze eye-position data relative to the perceptual switches and independently from the absolute duration of a single percept, the following procedure was applied for each trial: All eye-traces between subsequent perceptual switches were resampled by nearest-neighbor interpolation to equal length (chosen to be equal to the maximum length occurring for that trial). Then the eye-traces that belong to the same percept were aligned and averaged. In this normalized timeframe the time, in which percept I is perceived is chosen as interval (0,0.5), while percept II is perceived in (0.5,1). The perceptual switch from percept I to percept II thus occurs at time 0.5, while the switch from II to I is from time point 1 to time-point 0 in this periodic timeframe. We will refer to this representation as the mean time-normalized eye-trace of each trial.

In order to get a one-dimensional representation of the eye-position data, a principal component analysis (PCA) is performed on each of the mean time-normalized eye-traces, and the trace is projected on the (long) principle axis. However, the PCA leaves the direction (sign) of the principle axis open. As we set out to analyze the relative timing of eye-movements
CHAPTER 7. PERCEPTUAL RIVALRY

and percep, this sign cannot be inferred from the unbiased trial itself. Instead the mean
time-normalized eye-trace of both biased cube trials, which have the same cube size as the
considered unbiased trial, is used: this biased cube eye-trace is projected on the principal
axis of the unbiased trace; the sign then is defined such that the mean of the projected trace
is larger for (biased) percept I than for (biased) percept II. Therefore, a larger value of the
projected trace implies that the eye-position corresponds to percept I, a smaller value corre-
"ponds to percept II. For each biased trace, the sign was defined using the considered trace
itself. For the text-based trials, the biased mean time-normalized eye-traces corresponding
to the minimum cube size (100 pixel) were used.

To assess the relation between eye-movements and perceptual switches, one can make use
of the fact, that - by construction - the mean time-normalized eye-traces are periodic with
period 1, and the perceptual switches occur at the fixed time-points 0 (=1) and 0.5. There-
fore we applied a Fourier-expansion to these traces, after additionally normalizing the traces
to mean 0 and standard?deviation 1. If there is a strong relation between eye-position and
perception the F1-component of the transform can be expected to be most prominent;
the F1-amplitude thus is a measure of this relation. More important, the phase of the
F1 component provides a precise measure of the relative timing between eye-position and
perception.

7.2 Results

Eight subjects (4 female, 4 male, age from 25 to 47, mean: 29, seven right-handed, one (TF)
left-handed) were used for the experiment. One additional subject was excluded from analy-
"sis, as it could not report its percept consistently as inferred from the biased-cube trials.

The reaction time from percept to button-press is crucial to infer the time of perceptual switch
from the time of button press. Therefore, within the same subject the reaction times
of the three different biased cube sizes are compared among each other and to the text-based
trials. For none of the subjects we find a significant dependence on the cube size for the
left mouse button (ANOVA: $p > 0.07$ for each subject, figure 7.2a), and only one subject
(RS) exhibits a significant difference for the right button (ANOVA for each of the other
subjects: $p > 0.09$, figure 7.2a). The latter can most likely be attributed to a subjectively
reported change in finger position between trials. On the other hand, half of the subjects
show significant ($p < 0.05$, t-test) differences between the two mouse buttons in biased cube
trials (figure 7.2a). Differences between biased cube trials and text-based-trials are found
only for the right button in subjects PB and RS ($p > 0.24$, t-test for each of the other
subjects), and none for the left button ($p > 0.11$, t-test for all subjects). Building on these
results, the reaction time used in the following for calculating the time of perceptual switch
in each subject, is averaged over cube-sizes but separated for the two different percepts
(mouse buttons).

For the unbiased cube trials, we first analyze the durations each percept is maintained. We
find a wide range of the average time a subject keeps one percept, reaching from 1.9s (AS)
to 6.4s (CH). All but one subject (RS) exhibit a slight preference for percept I, which is
significant for 4 individual subjects ($p < 0.05$, t-test) in AS, CH, PB and TF. Although
significant differences ($p < 0.05$, ANOVA) between different unbiased cube trials can be
found in four subjects (both percepts: AS, RS; percept II only: PRK, TF), there is no consistent systematic effect of cube-sizes (figure 7.2b) or trial number. For all subjects the
mean percept duration is at least 11 times larger than the standard-deviation of the subject’s
reaction time, such that the variation of the reaction time cannot confound the analysis in
the time-normalized frame. The distribution of normalized perception durations (normal-
Figure 7.2: a) Mean reaction times for biased trials. For each subject from left to right: reaction time for text-based trial "left", reaction times for biased cube percept I increasing Cube Size (100, 200, 300 pixel), reaction-time for text-based trial "right", reaction times for biased cube percept II increasing cube Size. Errorbars denote standard deviation. b) Mean switch time at unbiased cubes for all trials. For each subject from left to right: mean duration percept I for increasing cube sizes, mean duration percept II for increasing cube sizes. c) Normalized switch-time distribution. Durations of each percept are normalized by the mean percept duration of the according trial and subject. These distributions are added over all trials of all subjects. Solid line shows best fitting gamma function.

ized to mean 1 for each trial, see methods) across all subjects does not significantly differ from a gamma distribution of mean 1 and gamma = 6.86 (p > 0.46, KS-test, figure 7.2c), whereas the best fitting normal distribution highly significantly deviates from the observed (p < 0.0001, KS-test). Hereby we confirm earlier findings (Borsellino et al. 1972) on the distribution of normalized perceptual durations to be gamma distributed.

The measured eye-positions at given time-points shortly after the perceptual switches in most unbiased cube trials show a clear separation depending on the polarity of the switch (i.e., whether from I to II or vice versa). In the example eye-trace shown in figure 7.3a, this separation is evident 300ms after the switch (7.3b). The variation in switch times, however, imposes a difficulty on quantitative analysis in an absolute time frame: Since the relative timing to the surrounding perceptual switches is different at this fixed absolute time-point, the eye-position can differently be influenced by the perception maintained or by the perception prepared. The heavy tail of the gamma-distribution of the switch times enhances this difficulty. To circumvent this problem we normalize each period between two perceptual switches to equal length and use this time-normalized eye-trace for further analysis. The resulting mean time-normalized eye-traces are shown in figure 7.5. Besides its 2-dimensional nature this representation in absolute space coordinates has an additional drawback for inter-subject comparison: As we infer from subjects' reports as well as from the mean normalized eye-traces of biased cube trials (not shown), subjects use different cues (cube edges or corners vs. faces) when looking at a cube; and the percepts belonging to the same eye-position by that can be just inverted. To overcome these problems and achieve a representation, which is 1-dimensional and independent from the subjective cues, we project the mean time-normalized eye-traces on their principal axis (figure 7.5, blue lines). For most unbiased trials (19/24) the main PCA axis explains more than 75% of the variance (figure...
Figure 7.3: a) Excerpt of an eye-trace of a single subject (RS). Green indicates that subject perceives percept I, red percept II. Timing of perceptual switches is corrected for reaction-times, which are estimated from biased cube trials. b) Eye-positions relative to the stimulus, 300ms after the times of perceptual switches. Data of the inset is taken over the whole trial, from which the example eye-trace is taken (Cubsize: 300 pixel). Latin numbers indicate corresponding points between inset and time-course plot.

Figure 7.4: Projections of normalized eye-traces on principal axis for all subjects and all unbiased cube trials. Rows correspond to cube sizes (from top to bottom: 100, 200, 300 pixels), columns to subjects. From biased cube trials the sign of the principal axis is chosen such that high values of eye-position correspond to (biased) percept I low values to (biased) percept II. Times of perceptual switches are at 0 (--1) and 0.5 in this normalized time frame.

Figure 7.5: Mean time-normalized eye-traces of 3 representative unbiased trials of 3 different subjects (KC, 200 pixel; CH, 300 pixel; RS, 300 pixel). Green denotes percept I, red denotes percept II. Blue line indicates the orientation of the first principal axis and its direction (from center to outside) as inferred from the biased cube trials.

Figure 7.6: Direction of principal axis and variance explained by this axis for all subjects and unbiased cube trials.
CHAPTER 7. PERCEPTUAL RIVALRY

Figure 7.7: a) F1 values of all figure 7.4 plots plotted in complex plane. Symbols denote subjects; gray arrow average F1 value. b) Control: F1 values for biased cube trials. c) Control: F1 values for text trials

7.6, left). This is also true for the biased trials (31/48) and for taking biased cube trials of the same cube size and same subject together (17/24), whereas in only half (8/16) of the text-based trials such a clear defined axis is found. This indicates that for most subjects and cube sizes there is a strongly preferred axis along which eye-movements takes place, which is not the case for the text-based trials. The orientation of the principal axis for most unbiased trials, matches the main diagonal of the cube (lower left to upper right): of all unbiased trials 11/24 are within 10 degrees of this main diagonal, 15/24 within 20 degrees and 20/24 within 30 degrees around this main diagonal (figure 7.6, right). A similar, however slightly smaller, predominance of the main diagonal is also present for the biased cube trials (14/48 within 10 degrees, 25/48 within 20 degrees and 31/48 within 30 degrees), ensuring that most subjects indeed use the same eye-movement strategy for looking at the biased cubes. However, no such predominance is found for the text-based trials (0/12 within 10 degrees, 3/12 within 20 degrees, 5/12 within 30 degrees of main axis, close to chance-levels predicted for homogenously distributed preferred axes that would be 1.3/12, 2.6/12 and 4/12 respectively). This shows that most subjects look along the main diagonal for viewing the cubes and that this is not a consequence of a general preference for this orientation of eye-movements.

To analyze the relative timing of perceptual switches and eye-position, the mean time-normalized eye-traces of unbiased cube trials are projected onto their principal axis. The direction (sign) of this axis is inferred from biased cube trials of same subject and same cube size, such that high values correspond to percept I and low values to percept II. This representation reveals that most subjects’ eye-position exhibit a strong modulation with fixed phase relative to their perceptual switches. While the amplitude of the eye-positions for most subjects depends on the cube size, the relative phase to the perceptual switches is consistent within each subject across cube-sizes (figure 7.4). In order to quantify the temporal relation of eye-position and perceptual switches further, we apply a Fourier transform to this representation (after normalization to zero mean and unit standard-deviation) and analyze the resulting F1 component. A large F1-amplitude corresponds to a strong modulation between of unit wavelength (in the used normalized time-frame) and thus indicated a strong relation between perceptual switches and eye-position, while a low F1-amplitude indicates an only weak or absent relation. The phase of the F1 component corresponds to the relative timing of perception and eye-position.

Most unbiased cube trials show a large F1 amplitude (for 18/24 trials |F1| > 0.5, mean:
(\langle F1 \rangle = 0.58), indicating that there is a strong relation between eye-position and perceptual switches. More important, the F1 components of 19 out of 24 unbiased cube trials have a negative imaginary part, indicating that in these about 80% of the cases, eye-movements are a consequence of perception rather than the cause. Additionally 13 out of 24 trials fall within 45 degrees to 90 degrees phase shift and the mean F1 component (also taking the amplitudes into account) also falls into that range (mean F1 angle: 114 degrees; \langle F1 \rangle = 0.31, figure 7.7a). This argues that the model most consistent with the data is an arrangement, in which the perception causes the eyes to move towards a position specific to this perception, and this eye-position in turn inhibits the percept.

As control we perform the same analysis for the biased cube trials and the text-based trials. As one expects (since the "percept" is externally triggered), the F1 component of the vast majority (37/48) of trials has a negative imaginary part, consistent with perception causing eye-movements. Most of those trials (36/48) furthermore show a phase-shift between -90 and 0 degrees (angle of \langle F1 \rangle is -40 degrees, figure 7.7b), consistent with the perception strongly driven by an external timer, as it is indeed the case. In the text-based trials the relative phase is random as quantified by the low amplitude of the mean F1 component (\langle |F1| \rangle = 0.06, compared to 0.31 for unbiased and 0.39 for biased cubes, figure 7.7c). This control on the text-based trials excludes the possibility that the observed relation between eye-position and perceptual switch could arise from eye-hand coupling when pressing the mouse-button to indicate the percept.

### 7.3 Discussion

We measured the temporal relation of eye position to perceptual switches in the appearance of a Necker Cube. Unprompted, subjects' perceptual switches were followed by consistent changes in eye-position, regardless of the particular features of the Necker Cube an individual subject chooses to foveate. This reflex-like shift of eye position is remarkable given that the target itself has not shifted, only the perception of its 3D-geometry has. This result throws light on a longstanding debate, which was begun by some of the most eminent scientists of the 19th century. They divided into two camps on the issue of whether the perceptual switch reported by Necker was 'mental' or whether it was driven by a reflexive 'involuntary' motor act. Necker (Necker, 1832), (Brewster, 1847) and (Hering, 1879) were nativists, who interpreted sensations as a true account of the material world. For them, the 'involuntary' eye movements produced a changed sensation that caused the perceptual switch. On the other side Wheatstone (Wheatstone, 1838), von Helmholtz (von Helmholtz, 1867) and Wundt (Wundt, 1897) were empiricists, for whom sensations provided only signs that had to be interpreted, often unconsciously, through our prior experience of the world. Hence Wheatstone’s opinion that the switches were the results of 'mental contemplation'.

One of the most detailed studies on the relation of eye-movements to the percept of the Necker cube, Glen found a clear correlation between eye-movements and reversal. When subjects were instructed to make switches to happen "as rapidly as possible", Glen's data suggest that eye-movements slightly precede the perceptual switch (Glen, 1940). These results directly challenged Zimmer's (Zimmer, 1913) earlier observation, that eye-movements do occur subsequent to perceptual switching. However Glen thought his analysis indicated that the eye-movements occurring after the perceptual switch, 'had become associated with whatever type of objective response is adopted for reporting perceived reversions. The random phase found for the text control in our study, however, excludes such an association between response and eye movement as one possible interpretation. In his 'spontaneous observation' task, which corresponds most closely to our experimental condition, Glen could
only state the "close proximity" of eye-movements and time of perceptual switching, without a clear statement of causal direction.

Using a rather crude technique to mark perceptual switches, Pfeiffer et al. (Pfeiffer et al., 1956) again supported Zimmer's result and argued in favour of perceptual changes leading to eye-movements and not vice versa. Their results had in turn been challenged on technical grounds by Flamm & Bergum (Flamm & Bergum, 1977), who found no significance difference in the number of eye-movements in a 1 second interval around the perceptual switch compared to other intervals. However, they instructed their subjects "not to be concerned if no switch occurred", which might bias subjects to fail to report a switch. Such failures to report a switch would confound their "control" intervals assumed to be free of perceptual switches. Furthermore, since Flamm & Bergum (Flamm & Bergum, 1977) did not incorporate the sizes of eye-movements, observing no difference in the number of eye-movements to perceptual switches does not argue against any relation of eye-position to percept.

The issue of eye-position rather than saccade frequency or direction was also addressed by Ellis & Stark (Ellis & Stark, 1978). Looking at a single subject at the instance of the perceptual switch they found a clear clustering of eye-positions depending on the polarity of the switch, similar to the one observed in the present study. However, unlike the present study (Ellis & Stark, 1978) did not try and explicitly address the question on the relative timing of eye-position change to perceptual switching. Ellis & Stark reported prolonged fixation times around the instant of perceptual reversal. Their result has been contested by Ito et al. (Ito et al., 2003), who could not replicate it and interpreted it as an artifact of a sampling bias known as the 'bus paradox'. Here long fixations would be more likely to contain an independent random event (the perceptual switch) than shorter fixations.

Kawabata et al. (Kawabata et al., 1978) used a clever technique to argue indirectly that eye-position causes the percept and not vice versa. They made subjects fixate different corners of the Necker cube. Hereby they found that subjects significantly favour the interpretation that the fixated corner belonged to the frontal part of the figures. However, corners are clearly a no more 'natural' cue for the Necker cube than edges or faces. In fact, in the present study we avoided biasing the subject as to what cue to use. Using the biased cube trials as reference (which are consistent with the subjective reports), we found that some subjects use the inner corners instead of the cube's faces or outer corners. This would yield seemingly inverted eye-movement patterns, which would lead to incorrect interpretation, if one had relied only on the unbiased cube data combined with the wrong a-priori assumption, that all subjects use the faces or outer corners as cues. Only by using the biased cube trials as reference this problem is resolved. Furthermore, it is not clear that the instruction to fixate does not interfere with the subjects' report, especially as Glen reported a dramatic dependence of switching rates on fixation instructions (Glen, 1940). Hence, providing explicit fixation instructions, especially in combination with instructions to facilitate or inhibit switching percepts (Peterson & Hochberg, 1983), is unlikely to produce results which are valid for free, spontaneous viewing.

As an interesting corollary, Leopold et al. showed that perceptual switching depends on the continuous presence of the percept (Leopold et al., 2002): if the percept is interrupted, e.g. by periodically removing the stimulus for some seconds, one particular view of the Necker cube can be stabilized: This implies that a memory of the percept persists, suggesting that the persistent activity of 'working memory' neurons in parietal or prefrontal cortex may be involved. Since the same regions are associated with planning and control of saccadic eye-movements, the close link between eye-movements (involuntary or not) to switches in the percept of the Necker cube might point to a common neuronal substrate. Our interpretation
is consistent with the findings of Leopold et al., i.e. interruption of the stimulus might suppress the suggested negative feedback mechanism of eye-position on and thus stabilize the current percept.
In summary, most evidence in favour of eye-movements causing perceptual switches had been indirect, using switching rates and effects of different instructions as main line of argument. On the other hand, most studies arguing in favour of the opposite causal relation, had been bedevilled by technical or conceptual problems. Our results may reconcile both views: In free viewing conditions we find a tight link between changes in perception and involuntary changes in eye-position. Perceptual switches precede a consistent change in eye-position. One simple interpretation of the phase relation of percept and change in eye-position is that the eye-position provides a feedback signal, that suppresses the current percept. By this negative feedback mechanism voluntary eye-movements or deliberate suppression of eye-movements can influence perceptual switching rates, although our results make it clear that eye-movements are the consequence rather than the cause of perceptual switching.
Chapter 8

Binding Features into Objects - An automated process?

We demonstrated in chapter 4 that different features of a natural scene can be segregated using a temporal coherence objective. We then went on to show that the bottom-up contribution of such simple low-level features like contrast or colour to overt attention only holds when they appear in their natural context (chapter 6). The Necker cube experiment in chapter 7 on the other hand demonstrated a strong bi-directional coupling between attention and scene interpretation. This raises the question whether attention alone is sufficient to induce a complex perceptual effect, such as binding multiple features into an object entity. In this chapter we investigate this question using an attentional tracking paradigm similar to (Blaser et al., 2000). We demonstrate, that - for artificial stimuli and adult observers - binding of features into objects is not an automatic process that can be induced by attention alone, but requires feedback.

8.1 Specific methods

8.1.1 Subjects

Ten subjects (3 female, 7 male, aged 25-41) participated in the experiment described in this chapter. One of the subjects is an author (PK), all other subjects were naive to the purpose of the experiment.

Figure 8.1: Possible coincidence types in a one dimensional feature space; from left to right: touch (T), none (0) and collision (C). Each colour corresponds to a different object; the feature angle is a periodic variable (360° for colour and 180° for orientation)
Figure 8.2: Illustration of the stimuli corresponding to different feature space trajectories. Each panel shows six time-points during stimulus presentation as indicated in lower left panel. Coincidence occurs at $t=3s$. Rows: different coincidence conditions for colour, columns for orientation. Upper 3x3 panels: dual feature task, bottom and rightmost panels single feature trials; bottom right: phase control. Note that the additional velocity for illustration was set to 0 in all stimuli. For real stimuli, this additional constant velocity prevents subjects from basing their decision solely on the final stimulus.

8.1.2 Stimuli

Feature-Trajectories

For the main experiment target and distractor were defined by two cyclic features, orientation and colour. This implies a two-dimensional feature space spanned by these features. As both features change over time, target and distractor are characterized by their trajectories in feature-space. In analogy to physical space, we refer to the 1st and 2nd temporal derivative of the features’ trajectories as their ‘velocity’ and ‘acceleration’ respectively. When target and distractor are present, their trajectories may intersect at certain points in time. We define three different basic types of coincidence for each feature: (1) In the "T"-type (for "touch"), trajectories do intersect at their extrema (i.e. at zero velocity) and then return into the direction of feature space they originated from. The acceleration of both trajectories is constant (2) In the "0"-type (no coincidence), trajectories do not intersect. The velocity of each trajectory is constant (3) In the "C"-type (for "crossing") features intersect. Both trajectories have constant velocity Figure 8.1 illustrates these 3 types of coincidence for a one dimensional feature space. In a two dimensional feature space both features can show all three types of coincidence. Consequently in this two dimensional space, there are 9 (3x3) different combinations of intersections, referred to as T/T, T/0, and so on.

In our experiments we identified the first feature with the orientation of a Gabor, the second feature with its colour. Colour here was defined as the azimuth in Derrington-Lennie-Krauskopf (DKL) colour space. Since both features are cyclic, the full cycle gives a natural scale for comparisons of distance, velocity and acceleration.
Dual feature task

In the 'dual feature task' subjects were first presented a single coloured Gabor. They were instructed that this Gabor "defines the target" which they "are supposed to track" and will "be overlaid by another Gabor after button-press". Upon button press the target was spatially superimposed by the distractor Gabor and both target and distractor started to move through feature space as described above. Target and distractor remained spatially superimposed throughout the trial (Figure 8.2, 3x3 panels on the upper left). To avoid potentially remaining spatial cues the gratings drifted in phase with constant velocity. After 6s both target and distractor stopped and the subjects were asked to indicate which of the Gabors they had tracked. Identification at the end of each trial was done by small numbers extending along the gratings’ direction, which also shared the final colour of their corresponding Gabor.

The initial position of the target in feature space was random. To ensure that target and distractor can be well discriminated at the start of the trial, the initial feature space position of the distractor was at maximal distance from the target. Feature space velocities were chosen such, that this was also true at the end of each trial. To avoid the possibility to track the target solely on the basis of initial and final position in feature space, a constant velocity was independently added to each feature, not affecting the above constraints. For clarity of illustration this aspect has been omitted in figures 8.1 and 8.2.

While two Gabors were presented superimposed, each Gabor was presented in every other frame only, yielding an effective temporal resolution of 60 Hz.

Single feature tasks

Single feature tasks were identical to the dual feature task apart from one feature being removed from the stimulus. In the orientation-only task, achromatic Gabors were used instead of coloured Gabors. (figure 8.2, bottom) In the colour-only task circular gratings were used instead of oriented gratings (figure 8.2, right). To control for the possibility to track the target on the basis of phase information achromatic circular gratings were used (figure 8.2, bottom right).

8.1.3 Experimental design

Discrimination thresholds

To test whether colour and orientation discrimination were comparable, in the first session each subject's discrimination threshold for each feature was estimated. This was done by a 3-alternative forced choice (3AFC) oddity task: For orientation three achromatic Gabors were shown for 83 ms each, interleaved by 0.5 s blanks. Subjects had to decide which of the Gabors deviated from the other two in orientation. An analogous procedure was used for colour. Acoustic feedback was provided on the correctness of the decision. Discrimination threshold was defined as 66 % correct level (chance at 33%). The distance between the 'odd' stimulus to the other two stimuli was adapted by a staircase procedure. For each feature 8 pedestal orientations/colours were tested (0°, 22.5°, …, 180° for orientation; 0°, 45°, …, 360° for colour).

Instruction session

To ensure, that there are no problems related to the principle understanding of the tracking task or to indication of the tracked target, a short (36 trials) instruction session was per-
formed as second session. To avoid any possible influence on how subjects deal with coincidences in feature space, for this session only stimuli with no coincidence in either feature dimension (i.e. 0/0 type stimuli) were used. This was the only tracking session in which subjects received feedback. Single feature sessions Stimuli in the single feature tasks were as described above. Half of the subjects performed a colour-only and an orientation-only task (108 trials each) before the dual feature sessions (i.e. as 3rd and 4th session) and after dual feature sessions and phase control session (as 9th and 10th session). The other half of the subjects performed these single feature tasks only after the dual feature sessions (i.e. as 6th and 7th session).

### Dual feature sessions
Stimuli were as described above, the nine different intersection types randomly intermixed. Each subject performed three dual-feature sessions of 108 trials each.

### Phase control session
To control whether subjects could use the phase for tracking, the phase control session (36 trials) was performed as 8th session in each subject. The schedule for all sessions is given in table 8.1.

### 8.2 Results

#### 8.2.1 Discrimination thresholds
As both colour and orientation are cyclic features, discrimination thresholds can be directly compared. The mean discrimination threshold over all subjects is 11.2° for orientation (Figure 8.3a). We observe best discrimination at the cardinal axis (0°, 90°) and on average 92% higher threshold at oblique (45°, 135°) orientations. The mean discrimination threshold for colour is 26.7° (Figure 8.3b). We observe best discrimination at greenish-yellow (270°) and redish-purple (45°) hues. The poorest discrimination threshold is 91% above the best threshold. Referred to the full cycle of either feature this results in 62% of the full range (11.2°) for orientation and 7.4% of the full range (26.7°) for colour. 8/10 subjects have slightly lower discrimination thresholds for orientation than for colour, although the

<table>
<thead>
<tr>
<th>Session</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>discrimination thresholds</td>
<td>discrimination thresholds</td>
</tr>
<tr>
<td>2</td>
<td>instruction session (0/0)</td>
<td>instruction session (0/0)</td>
</tr>
<tr>
<td>3</td>
<td>dual feature task</td>
<td>single feature task</td>
</tr>
<tr>
<td>4</td>
<td>dual feature task</td>
<td>single feature task</td>
</tr>
<tr>
<td>5</td>
<td>dual feature task</td>
<td>dual feature task</td>
</tr>
<tr>
<td>6</td>
<td>single feature task</td>
<td>dual feature task</td>
</tr>
<tr>
<td>7</td>
<td>single feature task</td>
<td>dual feature task</td>
</tr>
<tr>
<td>8</td>
<td>phase control task</td>
<td>phase control task</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>single feature task</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>single feature task</td>
</tr>
</tbody>
</table>

Table 8.1: Experimental design
overall difference is just above significance (p=0.051, t-test). Thus, in the context of the present task, we do not observe relevant differences in discrimination thresholds of these two features.

### 8.2.2 Instruction session

To ensure subjects understand the instructions and can accurately perform the tracking task as such, we run one initial session on 0/0 stimuli (Figure 8.2, 2nd row, 2nd column) with acoustic feedback. Since these stimuli do not coincide in feature space, the feedback in the instruction session cannot be used to disambiguate coincidences in later sessions. After 36 trials the subjects reached an average performance of 88% with a small standard deviation (Figure 8.4). Thus, all subjects quickly learnt the principle of the task.

### 8.2.3 Single feature tasks

First we analyze the performance in single feature tasks. The baseline is formed by the phase-only control stimulus (Figure 8.2, bottom right corner). Here the subjects reach on average
55% performance. This average is not significantly different from chance level (p > 0.48, t-test). Thus, there is no residual information contained in the phase of the stimuli that could confound the results of other experiments. In the case of orientation-only (Figure 8.2, bottom row), there is a highly significant dependence of performance on the type of intersection (p < 10^{-5} ANOVA, Figure 8.5a). "C" and "O" type trials do not significantly differ from each other (p > 0.33, t-test), but the dependence arises from the difference between 'T'-type and the other two types (p < 10^{-3} in each case). This remarkable difference is a consequence of a wide scatter of performance across subjects, some subjects even systematically interpreting the T-type intersection analogously to the C-type intersection. This is surprising, as in the T-type trials at the moment of coincidence, both target and distractor have an instantaneous velocity of zero and a negative acceleration (relative to their previous movement direction). Thus a physical interpretation would be a touch. Nevertheless, some subjects consistently track the 'wrong' grating after the coincidence occurred. These results imply that orientation inertia is used to faithfully track orientation in case of crossings, and by some subjects even in cases when it is physically incorrect. In the case of colour-only (Figure 8.2, right column), there is no difference in performance between any of the three intersection types (p > 0.29, ANOVA, figure 8.5b). The mean performance over all "colour only" trials is 77% and significantly greater than chance (p < 10^{-3}, t-test). This implies that subjects use colour 'inertia' for tracking colour, and in contrast to orientation, no subject shows a consistent misuse of the concept of colour-inertia. For the interpretation of the dual feature results, it is important to know, whether - judging from the single feature results, any of the features is more reliable for tracking than the other. Comparing all single feature trials independently of the intersection type, performance for both features is about the identical (colour: 77%, orientation: 80%, p > 0.56, t-test). Since subjects have no information on the different intersection types, this shows, that objectively none of the features is better or worse for tracking than the other.

8.2.4 Learning

Does the performance change in the course of the experiments, or is the performance dependent on the sequence of single and dual feature tracking tasks? Comparing the single feature results obtained before and after the dual feature task for the 5 subjects of group 2 (table 1) shows no improvement neither for colour (p > 0.55, t-test) nor for orientation (p > 0.42, t-test). This implies that the capability to track a single feature is not affected by learning during the experiment. Comparing the performance in the first dual feature session (figure 8.6 left column) of subjects in group 1 (74.2% correct) to those of group 2 (74.4% correct) we observe no difference (p > 0.95, t-test). The same is true over all dual feature sessions (79.4% to 79.3%, p > 0.99). This indicates that under the conditions of the present experiment, i.e. absence of feedback, the dual feature task has no impact on the performance in the single feature task.

8.2.5 Dual feature task

In the dual feature task we address whether subjects track features individually or they track the target features as a single object entity. In case objects are tracked as a whole, we would expect an interaction of the two feature dimensions. First we analyze the performance for the different combinations of coincidence types averaged across subjects and sessions. This results in a 3x3 (T/T, T/O, ...C/C) performance matrix (figure 8.6a). In case these two features - orientation and colour, were used independently for tracking, this matrix could be represented as outer product of two performance vectors. We compute the performance
Figure 8.5: Single feature experiments. a) Percentage correct for orientation-only trials versus coincidence type (T/0/C). Each data point represents 1 subject, horizontal lines indicate mean over all subjects. For group 2 data is averaged over both orientation-only sessions. (b) Results of colour-only experiment, presentation analogous to panel a.
vectors, which optimally (in a least squares sense) approximate the measured performance matrix (figure 8.6 b,c). Indeed the outer product of these vectors (Figure 8.6d) resembles the original data well, indicating a lack of interaction between the two tracked features. To quantify this lack of interaction further, we perform a balanced 2 way ANOVA on the tracking performance for the 3x3 different types of coincidence. Pooling each subject’s results from all 3 dual feature sessions we find that the results individual factors are consistent with the single feature task results: While there is a highly significant dependence on orientation (p-value $\ll 10^{-10}$, below numerical accuracy, $F=66.1$), there is no dependence on colour (p=0.41). However, we do not observe any interaction between the two features (p = 0.59). Concluding, we find no evidence, that the two target features are tracked as a single object entity.

8.2.6 Learning

So far we analyzed all dual feature sessions together. To investigate whether there is a change during the course of the experiment we additionally perform the analysis individually for each session (figure 8.7). We neither find a significant interaction of colour and orientation, nor a significant influence of colour in any session. If anything, any remaining interaction decreases over sessions (p = 0.20, 1st session; p= 0.57, 2nd session; p = 0.95, 3rd session). There is also a trend that any residual dependence on colour decreases over sessions (p=0.17, p=0.62, p=0.71 for 1st, 2nd and 3rd session respectively). This may indicate, that there is a small residual effect of colour very early in the experiment, that completely vanishes during its course. To check whether this residual effect could have an impact on tracking,
we analyze the development of overall tracking performance. In the first dual feature session overall performance is below orientation-only performance in 8 out of 10 subjects (p < 0.05, Wilcoxon signed rank test). Performance increases significantly from 1st to 2nd dual feature session (increase for 8 subjects; decrease for 1 subject; p < 0.05, Wilcoxon signed rank test) and is then comparable to orientation only performance (p > 0.92; Wilcoxon signed rank test). This cannot be explained by a general training effect on the task, as such an effect would also affect performance in the single feature tasks, which is not the case (see above). This result indicates that in early stages of the dual feature experiment colour might take away attentional resources from tracking orientation.

8.3 Discussion

The experiments of this chapter show that humans can faithfully track perform tracking tasks not only in physical space (Pylyshyn & Storm, 1988), but also in abstract feature spaces. Beyond this confirmation of earlier work (Blaser et al., 2000), we additionally show that this correspondence between physical and feature space extends to the disambiguation of coincidences. Subjects interpret coincidences in feature space on the basis of the immediate object history and of coincidence parameters using strategies known from physical space (Metzger, 1934; Michotte, 1963). As seen for the "T-type" coincidence in the orientation domain, the disambiguation strategy is usually preserved within the same subject, although it may differ across subjects. This indicates, that concepts like velocity, acceleration and inertia are also imminently present for abstract feature spaces.

Tracking in physical space requires directing spatial attention towards the target. Both widely used models for spatial attention, the feature integration theory with its spotlight metaphor (Treisman & Gelade, 1980) and the biased competition model (Duncan & Humphreys, 1989; Desimone & Duncan, 1995; Lee et al., 1999; Deco & Lee, 2002), utilize the retinotopic organization of early visual areas (Daniel & Whitteridge, 1961)). Neuromorphic technical solutions of tracking also rely on this principle (Horiuchi et al., 1997; Indiveri, 1999). Retinotopy thus is likely to be the key for attentional tracking in physical space. Besides retinotopy there is also a featro-topic cortical organization: The primary visual cortex is organized in orientation domains with nearby cells tending to have similar orientation preference (Hubel & Wiesel, 1962; Hubel & Wiesel, 1998; Hubel et al., 1977; Bonhoeffer & Grinvald, 1991). Furthermore cells of similar orientation preference are more likely to be connected than those coding for different orientations (Malach et al., 1993; Boaking et al., 1997). In a recent study on macaque V2, Xiao et al. (Xiao et al., 2003) found a similar featro-topically organization also for colour. Furthermore, electrophysiological (Treue & Martínez Trujillo, 1999; McAdams & Maunsell, 2000) as well as human imaging (Saenz et al., 2002) studies confirm that feature-based attention acts similar to spatial attention. This evidence makes it seem likely that tracking in feature space is based on similar mechanisms as those suggested for tracking in physical space.

In our dual feature experiments we do not find any interaction between the features that form the target. This is in contrast to the results of (Blaser et al., 2000), who find that subjects bind features into a single object entity. While the stimuli are similar in both studies, there is one decisive difference: (Blaser et al., 2000) provide positive feedback for correct tracking, which is even more supported by a 'point score'. In contrast, in the present study no feedback is provided. Note, that the short instruction session with feedback only uses 0/0 stimuli and thus cannot serve as training for the main experiments. The conjunction of both studies implies that binding of features to an object is dependent on feedback. Consequently, the notion of which combination of features form an object entity is subject to
### Table 8.7: Results of dual feature experiments of all subjects (rows) separated by sessions (columns). For each subject and session, the 3x3 coincidence combinations (T/T, T/0, ..., C/C) are displayed as indicated for the 1st session of JM: columns correspond to different coincidence types for orientation, rows for colour. Percentage correct trials denoted by white number and are colour-coded. Top half of subjects (JM to PK) form group 1, bottom half (GM to JHZ) group 2.
plasticity. This is in line with results that human infants can bind different features into objects only at about 3 months of age (Cohen, 1973; Mundy, 1985). This view has been challenged by (Slater et al., 1991), who find that already newborns perceive objects as feature compounds. However, (Taga et al., 2002) argue that this early type of feature binding differs from adult-like binding. Most likely newborns do not differentiate the features and thus circumvent the binding problem and adult-like attention-based binding arises only at 2 to 3 months of age (Taga et al., 2002). Furthermore, there is evidence that the notion of an object required for disambiguation of physical space coincidences undergoes a change at the age of 6 months (Scheier et al., 2004). Concluding, binding of features into an object is not an inherent property of the visual system but subject to plasticity, and this plasticity requires direct reinforcement in adult subjects.
Chapter 9

Concluding Remarks & Outlook

9.1 Summary of results

In this thesis we addressed the processing of natural scenes, both by theoretical modeling and psychophysical experiments. We first analysed the statistical regularities of natural scenes and linked them to known experimental data. We then showed that a single learning rule, temporal coherence, can extract relevant properties from these stimuli. The same rule explains processing at several levels of the visual hierarchy in a purely bottom-up (stimulus-driven) fashion. Consequently we asked the question, whether overt attention can also be accounted for by bottom-up mechanisms. First we showed that even in the cat the role of eye-movements extends much beyond the reflexive, gaze-stabilizing function that previous laboratory studies had suggested. We then tested individual contributions of features suggested in the saliency-map model for primate attention. This revealed little to no causal effect of low-level stimulus features, indicating that primate attention cannot be explained by a simple bottom-up model. To directly test the top-down effects we then used the paradigm of perceptual rivalry, which excludes bottom-up influences. We found a strong bilateral coupling between human scene perception and overt attention. In a final experiment we showed that attention alone on the other hand is not sufficient to define a perceptual object. Taken together these results suggest that the visual system exploits natural scene statistics to form representations that are then employed for bottom-up perception and the top-down direction of attention.

We have already discussed our results and their relation to the existing literature at the end of the corresponding chapters. Therefore we will restrict ourselves here to some additional links among our results and focus on an outlook on potential lines of further research that build on the presented findings.

9.2 Natural stimuli and Attention

One of the main findings of this thesis is that simple bottom-up saliency-maps cannot account for all aspects of directing overt attention in natural scenes. Despite being a negative result, it can guide the way to further experimental paradigms: The proposal to use second order luminance-contrast (texture contrast) to account for part of our results (Parkhurst & Niebur, 2004) provides a first step into this direction. Likewise second-order contrasts in other feature channels (colour and orientation) may be considered. Along the same line and building on results on interactions between features, mixed second order contrasts between
feature channels are likely to be also of relevance and to improve bottom-up saliency models. Our results on natural scenes as well as on Necker cubes, however, indicate that - even in the absence of a well-defined task, top-down influences play a major role in directing overt attention. To assess the bottom-up component alone, we need paradigms that eliminate or at least limit the role of top-down signals. Furthermore, given the tool of eye-tracking, we restricted ourself to overt attention, while saliency map models also have been suggested to model covert attentional shifts (Itti & Koch, 2000). One possibility to address covert attention and at the same time limit top-down influences are concurrent dual rapid classification tasks, that show dramatic differences between natural scenes and artificial stimuli (Li et al., 2002). Building on the results presented here, the combination of both eye-tracking and rapid classification during free viewing of appropriately modified natural scenes offers a promising way to address bottom-up driven attention - both overt and covert, in a single paradigm.

9.3 Learning from natural stimuli

In spite of the importance of top-down interactions for attention, we derived several properties of the visual system in a purely bottom-up model. As we used one single rule, our model can readily be scaled into a large hierarchical model, that simultaneously incorporates various stages of visual processing. This suggests that a large part of early visual perception can be performed in a bottom-up fashion, solely on the basis of stimulus statistics. This even includes a process, that is typically expected to involve top-down interactions, namely the suppression of distractors. Combining the results on the top-down role in attention and the bottom-up models for early visual processing suggests, that experience gained through learning from natural stimulus statistics may be the basis for top-down signals in later learning phases. Including top-down signals into our theoretical learning models will thus be a decisive step for further research.

While our results show that in principle natural visual systems could make use of the statistical properties of natural scenes, it obviously does not proof that they must indeed do so. Furthermore, in most cases it is not clear to what extent the adaptation to the natural environment occurs on ontogenetic or on phylogenetic timescales. One approach to this question would be rearing animals in modified environment akin to the seminal visual deprivation experiments of Hubel & Wiesel in the 60s and 70s for the case of primary visual cortex (see (Hubel & Wiesel, 1998) for overview). However, to appropriately design such experiments a sound understanding of statistical properties of natural scenes is inevitable. We are confident, that the data presented in this thesis provides a good starting point for such research.

9.4 On the use of natural stimuli

When investigating sensory processing, one always faces a trade-off between using well-controlled artificial stimuli as compared to stimuli that closely reflect those the system naturally encounters: While the latter are harder to control and consequently make interpretation of data more difficult, it is not clear to what extent results obtained on artificial stimuli generalize to the real-world. Although most of our knowledge on the visual system to date results from experiments with artificial stimuli, we demonstrated the limitations of too simplified generalizations to natural conditions:
• While luminance-contrast definitely is a factor to overt attention, when no other cues are present, it gets completely overruled in natural scenes.

• While eye-movements were thought to contribute little to exploratory behaviour in the cat under laboratory conditions (Guitton et al., 1984), our results suggest that this does not hold under natural conditions.

• While infants in a natural environment learn to bind features into objects apparently without any teacher (Scheier et al., 2004), artificial features - like our chromatic gratings - require feedback.

Another example referred to in this thesis, in which results obtained on simple stimuli cannot be readily transferred to the natural world is the oblique effect: originally described for gratings the effect reverses for natural scenes and patterns that share their frequency spectrum (Essock et al., 2003). Only the knowledge of natural scene statistics resolves this seeming discrepancy (Hansen et al., 2003). A similar example is found in V1 physiology: while neuronal responses to natural scenes cannot be utterly explained from the responses to gratings (Kayser, Salazar et al., 2003; Smyth et al., 2003) most of multiunit activity and local field potential structure can be reproduced by using artificial stimuli that share the spatio-temporal spectrum of natural scenes (Kayser, Salazar et al., 2003). Both examples stress the importance to know, which features of natural scenes are relevant for a certain phenomenon. Only then one can design experiments that are both - well-controlled and generalizable to natural conditions. In most parts of this thesis we used stimuli and experimental conditions that resembled the natural environment as closely as possible. The two experiments, in which we used artificial stimuli, either had the purpose to directly stress their limitations (tracking experiments) or investigated the interpretation of a plain figure as natural object (rivalry experiments) under natural viewing conditions. In conclusion, all data presented in this thesis are likely to provide a relevant contribution to the understanding of natural behaviour.
Bibliography


BIBLIOGRAPHY


Einhäuser W, Kruse W, Hoffmann K, König P. Contrast and overt attention: Correlation and causality in monkey and human. manuscript in preparation

Einhäuser W, Martin KAC, König P. Are switches in perception of the Necker cube related to eye-position? submitted manuscript

Einhäuser W, König P. Coincidences in feature space: the notion of an object is subject to plasticity. submitted manuscript

Einhäuser W, Hipp J, Eggert J, Körner E, König P Learning viewpoint invariant object representations using a temporal coherence principle. submitted manuscript
Einhäuser W (2001) Selfsupervised Online Learning of Invariances from Natural Stimuli (Diplomarbeit - ETH Zürich 2001)


Franzius M, Einhäuser W, König P, Körding KP. Learning a hierarchical model of cortical function from natural stimuli. submitted manuscript


Frey H, König P, Einhäuser W The influence of coulor on a human saliency map. manuscript in preparation


Hochberg JE (1950) Figure-Ground Reversal as a function of visual satiation. J. Exp. Psychol. 40:682-686.


James, W (1890) Principles of Psychology, New York, Holt.


BIBLIOGRAPHY


Maxwell JC (1857) Experiments on colour, as perceived by the eye, with remarks on colour-blindness. Transactions of the Royal Society of Edinburgh, 21:275-299.


Moeller GU, Einhäuser W, Conradt J, König P. Interaction of eye- and head movements in the freely havaving cat. manuscript in preparation


Necker LA (1832) Observations on some remarkable Optical Phaenomena seen in Switzerland; and on an Optical Phaenomenon which occurs on viewing a figure of a crystal or geometrical solid. The London & Edinburgh Philosophical Magazine and Journal of Science, 1:329-337.


Pettigrew JD, Nikara T, Bishop PO (1968) Responses to moving slits by single units in cat striate cortex. Exp. Brain Res. 6, 373-390.


Rensink RA, O'Regan JK, Clark JJ (1997) To see or not to see: the need for attention to perceive changes in scenes. Psychol. Sci. 8:368-373.


Acknowledgements

Financial support

We are grateful to Honda RI Europe GmbH for their generous financial support of this work. In addition we thank the following organizations for providing travel grants: the European Neural Network Society (ENNS), the European Neuroscience Society (FENS) and the Center of Neuroscience Zurich (ZNZ).

Personal

I would like to thank my supervisor Peter König, without whom this work would not have been possible, as well as Rodney Douglas and Kevan Martin for their invaluable support and for the creative and inspiring atmosphere at the Institute of Neuroinformatics. Furthermore I would like to thank all Co-Authors that directly contributed to the data presented this thesis: Christoph Kayser, Edgar Körner (Honda), Gudrun Moeller, Matthias Franzius, Hans-Peter Frey, Jörg Hipp, Jörg Conradt, Julian Eggert (Honda), Kevan Martin, Klaus-Peter Hoffmann (Uni Bochum) Konrad Körding and Wolfgang Kruse (Uni Bochum). Special Thanks also to Dave Lawerence for solving all the computer problems of his “best customer” and to Daniel Kiper for help in setting up the psychophysical experiments. Finally I would like to thank my wife Valerie for her love and support.