What is the Goal of Complex Cell Coding in the Primary Visual Cortex?

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Jörn-Philipp Lies
aus Heidelberg, Deutschland

April 2013

Tag der mündlichen Prüfung:	24. Juli 2013
Dekan der Math.-Nat. Fakultät	Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. Matthias Bethge
2. Berichterstatter:	Prof. Dr. Felix A. Wichmann
Prüfungskomission:	Prof. Dr. Matthias Bethge Dr. Laura Busse Prof. Dr. Martin Giese Prof. Dr. Felix A. Wichmann

I hereby declare that I have produced the work entitled: "What is the Goal of Complex Cell Coding in the Primary Visual Cotex?", submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, ____________________ ____________________

Date Signature

Danksagung

Zuerst danke an Matthias, dessen Ratschläge und Unterstützung mich immer wieder auf die richtige Spur gebracht haben, und der mich motiviert hat wann immer ich dachte komplett den Faden verloren zu haben. Mit einem anderen Doktorvater wäre ich nie so weit gekommen. Und danke an alle anderen im Bethgelab für die familiäre und inspirierende Atmosphäre. Ganz besonders an Alex, Fabi, Jakob, Ralf und Seb, die mir mit viel Geduld immer wieder weitergeholfen haben.

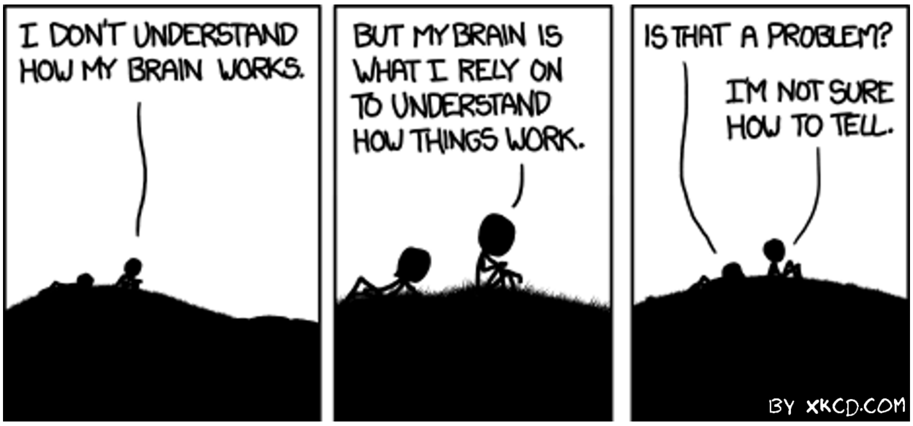
Dann danke an all meine Freunde, die alle auf ihre Weise die letzten Jahre unvergesslich gemacht haben. Danke an Anne und Philip, für die konstante Quelle an Unterhaltung Tag aus Tag ein und all die gemeinsam bestrittenen Abenteuer. Danke an Alex, Christian, Frank, Holly, Julia, Lucas, Regine und Sonja für all die schöne Zeit außerhalb des Labs. Und danke an Martin für die vielen Gespräche über Gott und die Welt, ob bei Kaffee oder Bier. Ganz speziell auch danke an Björn, Catherine, Janina und Nina - ohne euch wäre Tübingen nicht das gleiche gewesen und ohne eure Ablenkung wäre ich vermutlich längst wahnsinnig geworden. Und danke an Mazze und Georg für die letzten (fast) 12 Jahre.

Danke natürlich auch an meine Eltern, Großeltern und Geschwister, für die bedingungslose Unterstützung all die Jahre. Nichts gibt einem so viel Selbstvertrauen wie das Wissen, dass immer jemand hinter einem steht.

Abschließend und hervorzuhebend danke an Leona. Für alles.

It's the kid inside of us that keeps us all from going crazy.

Dr. John Dorian


I DON'T UNDERSTAND HOW MY BRAIN WORKS.
BUT MY BRAIN IS WHAT I RELY ON TO UNDERSTAND HOW THINGS WORK.
IS THAT A PROBLEM?
I'M NOT SURE HOW TO TELL.
BY XKCD.COM

Abstract

Complex cells in the primary visual cortex are the first cells to exhibit geometrical invariance, namely they are insensitive to the phase of a stimulus. It has been suggested that complex cells learn this property from the statistics of their input. Two differing unsupervised learning paradigms have mainly been used: slowness and redundancy reduction. This thesis provides a quantitative comparison of slowness objective and redundancy reduction objective with respect to their ability to account for complex cell properties. Both objectives have been proposed as principle underlying the formation of complex cell features, however, we show that—contrary to widespread belief—the two objectives lead to quite different predictions for the receptive field properties. For this, we compare both objectives on a population and a single cell level. The redundancy reduction objective is represented by *independent subspace analysis* (ISA) and the slowness objective by *slow subspace analysis* (SSA). We show that SSA is favorable over the better known slow feature analysis (SFA) algorithm, as SFA is unable to reproduce key properties of complex cell receptive fields and SSA uses the same energy model structure as ISA. We find that slowness leads to global receptive fields in both single cell and population approaches. The receptive field size is only limited by the patch size and SSA can be seen as a generalization of the Fourier transform. Redundancy reduction, in contrast, leads to clearly localized receptive fields but with spatial frequency and aspect ratio higher than those found in physiological studies. We also find that in a combined optimization of slowness and redundancy reduction the filters obtained resemble those found with redundancy reduction alone even though the individual optima are quite different. In summary, both slowness and redundancy reduction cannot account for all complex cell properties evaluated here, but would require additional constraints such as wiring length to lead to physiologically plausible receptive fields. However, studying the quite opposing demands of both objectives can lead to a better understanding of the computational strategy employed in the visual system.

Contents

1 Introduction

Less than 50 years ago, the complexity of visual processing in the brain was largely underestimated. In 1966, artificial intelligence pioneer Marvin Minsky proposed the project of designing a computer vision system as a summer project to one of his undergraduate students. Today, in 2013, the vision system is still in development. We made tremendous advantages and countless researchers dedicated their career to extend the knowledge about the visual system. Experts in neurophysiology, psychophysics, signal processing, or machine learning, to name a few, worked together and recorded cell responses, mapped out cell connectivity, and developed computational models to explain the found properties. From my point of view particularly interesting is the early visual system, which can be reasonably well modeled as signal processing network. More specifically, complex cells in the primary visual cortex exhibit invariance to stimulus position, and invariance is an interesting and well-studied subject in dynamical systems. There are different models which explain invariance and other complex cell properties; an important question to ask is *why* do the models find these properties. And this is the crucial question of this thesis. To provide an introduction to complex cell models, I divided the first chapter of my thesis in three major sections. First, I give a short introduction into the physiology and cell models of the early visual system on a coarse level. In the next section, I introduce the concept of invariance from a theoretical point of view. Finally, I combine the physiological models and the invariance theory in presenting how invariance can be learned in a neuronal framework.

1.1 Physiology of the early visual system

The ability to see fascinates humans for centuries. Basic understanding of optics goes back to Greek philosophers like Euclid or Aristotle almost 3000 years ago. The first fundamental publication on optics was the seven volume treatise *Book of Optics* written by the Arabic scientist Alhazen around 1000 AD (Verma, 1969). Alhazen put forward the hypothesis that visual perception takes place in the brain rather than in the eye. More recently, the computational role of the brain was also emphasized when Herrmann Von Helmholtz

(1867) found that given the rather poor optical properties of the eye, the brain must do some kind of inference to evoke the high quality perception that humans have. He called this idea *unconscious inference*. At the end of the 19th century, scientists found that the visual signal from the retina travels through the lateral geniculate nucleus (LGN) to the primary visual cortex in the occipital lobe (Henschen, 1893; Flechsig, 1896). An illustration of the visual pathway is shown in Figure 1.1. After the visual information reaches the primary visual cortex, it passes though the extrastriate visual cortical areas (V2, V3, V4, and V5) before it is split into two processing streams: the dorsal stream, which represents the spatial position of objects, and the ventral stream, which represents the identity of an object (Mishkin and Ungerleider, 1982; Ettlinger, 1990; Goodale and Milner, 1992).

A milestone in vision was the Nobel price awarded findings of Hubel and Wiesel in the primary visual cortex of cats (Hubel and Wiesel, 1962, 1963) and monkeys (Hubel and Wiesel, 1968). Until then it was known that cells in the retina and LGN respond to bright spots on dark background or vice versa (Kuffler, 1953), but response properties of cortical cells were unknown. By probing cells in the striate cortex while flashing oriented bars in the visual field of the animal, they found that the V1 cells respond to oriented bars instead of single spots. Further they found that the cells can be separated into two different classes according to their response properties: *simple cells* and *complex cells*. A cell is classified as simple cell if its receptive field can be subdivided into distinct excitatory and inhibitory regions with linear summation (Hubel and Wiesel, 1962). All cells which failed to comply with any of the criteria were called *complex cells*. In the following years, the primary visual cortex has been the most studied visual area in mammalian brain (for reviews see e.g. Callaway (1998); Bruce and Green (2003); Carandini et al. (2005)).

The basic response properties of simple cells can be described by a linear receptive field with a point-wise output nonlinearity (Movshon et al., 1978b; Andrews and Pollen, 1979). At more detail, the responses vary largely with stimulus contrast, which can be explained by contrast gain control mechanisms that incorporate the responses of neighboring cells (Heeger, 1992a,b; Carandini and Heeger, 1994; Carandini et al., 1997). However, these output nonlinearities do not alter the linear feature detection carried out by the simple cell (Smyth et al., 2003).

The response properties of complex cells are generally nonlinear and therefore more difficult to identify. A quantitative study by Movshon et al. (1978a) showed that complex cells can be described as the integration of the output of several simple cell subunits. Spike triggered covariance (STC) (de Ruyter van Steveninck and Bialek, 1988; Touryan et al., 2002; Rust et al., 2005) analysis allowed a flexible estimation of the linear subunits of a complex cell. In STC, the stimuli which elicit a spike are collected and the covariance matrix of

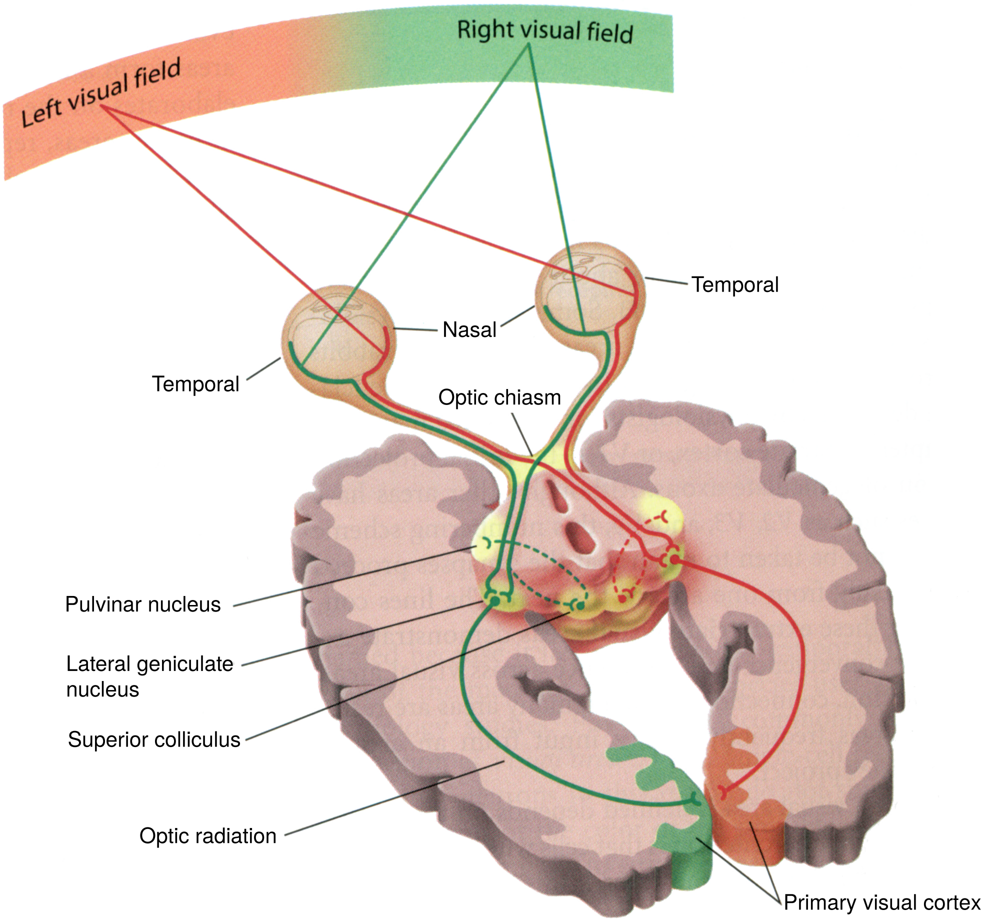


Figure 1.1. Primary visual pathway. The visual information from the visual field passes through the eye and stimulates the receptors on the retina. The retinal cells transmit the signal through the optic nerve to the lateral geniculate nucleus and then to the primary visual cortex in the occipital lobe in the back of the brain.
From COGNITIVE NEUROSCIENCE: THE BIOLOGY OF THE MIND, THIRD EDITION by Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun. Copyright ©2009, 2002, 1998 by Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun. Used by permission of W.W. Norton & Company, Inc. (Gazzaniga et al., 2009)

these stimuli is computed. The eigenvectors of this covariance matrix with eigenvalue significantly larger or smaller than the eigenvalues of the covariance matrix over all stimuli reveal the linear subunits of the complex cell. STC analysis has shown that complex cells consist of very few subunits with clear orientation selectivity, localization and bandpass filtering (Touryan et al., 2005; Rust et al., 2005; Chen et al., 2007). Besides the extensive characterization of simple and complex cell response properties, the proposed classification of V1 cells into simple and complex is still debated. Instead of two distinct classes there is rather a continuum ranging from cells with more simple cell-like responses to cells with more complex cell-like responses (Dean and Tolhurst, 1983; Chance et al., 1999; Priebe et al., 2004). Mechler and Ringach (2002) provide an overview of various quantitative studies with evidence for and against a segregation into two distinct classes. More recently Fournier et al. (2011) found that the ratio of complex cells to simple cells also depends on the kind of stimulus used for classification rather than being a property of the cell itself.

1.1.1 Cell Models

For the description of simple cells it is common to use a parsimonious model consisting of two stages (Daugman, 1980; Carandini et al., 1997, 1998): a linear filter followed by a nonlinear output stage (Figure 1.2 A). Marcelja (1980) found that the static filter properties of simple cells can be described by a two-dimensional Gabor function (Gabor, 1946a,b,c). A Gabor function provides orientation tuning and spatial frequency tuning which, given the right set of parameters, lies within the range of physiological data (Kulikowski et al., 1982; Field and Tolhurst, 1986; Kulikowski and Bishop, 1981b; Jones and Palmer, 1987a). The nonlinear stage consists of a static half-wave rectification or squaring, allowing for positive outputs only. An important extension of the simple linear-nonlinear model is the addition of contrast gain control mechanisms. The most common contrast gain control model is divisive normalization (Albrecht and Hamilton, 1982; Bonds, 1989; Heeger, 1992a; Geisler and Albrecht, 1992; Carandini et al., 1997).

Adelson and Bergen (1985) proposed a parsimonious model for complex cells: the *energy model*. It consists of two Gabor filters with identical orientation, spatial frequency, and Gaussian envelope, but a 90° phase offset. In contrast to the simple cell model, the output nonlinearities are not half-wave rectifications but quadratic functions. The contrast gain control mechanism for complex cells is commonly modeled just like for simple cells (Ohzawa et al., 1982, 1985).

1.1.2 Visual Adaptation

Following David Marr, the visual system can be analyzed on three different levels: (1) the computational goal or objective to achieve, (2) the algorithm and the representation of its in- and output, and (3) the (biophysical) substrate within which it is implemented (Marr and Poggio, 1976). All these levels are largely independent (Marr, 1982), just like a software designer should not have to care about the details of the underlying hardware. However, most models incorporate more than one level (Serre and Poggio, 2011). Marr

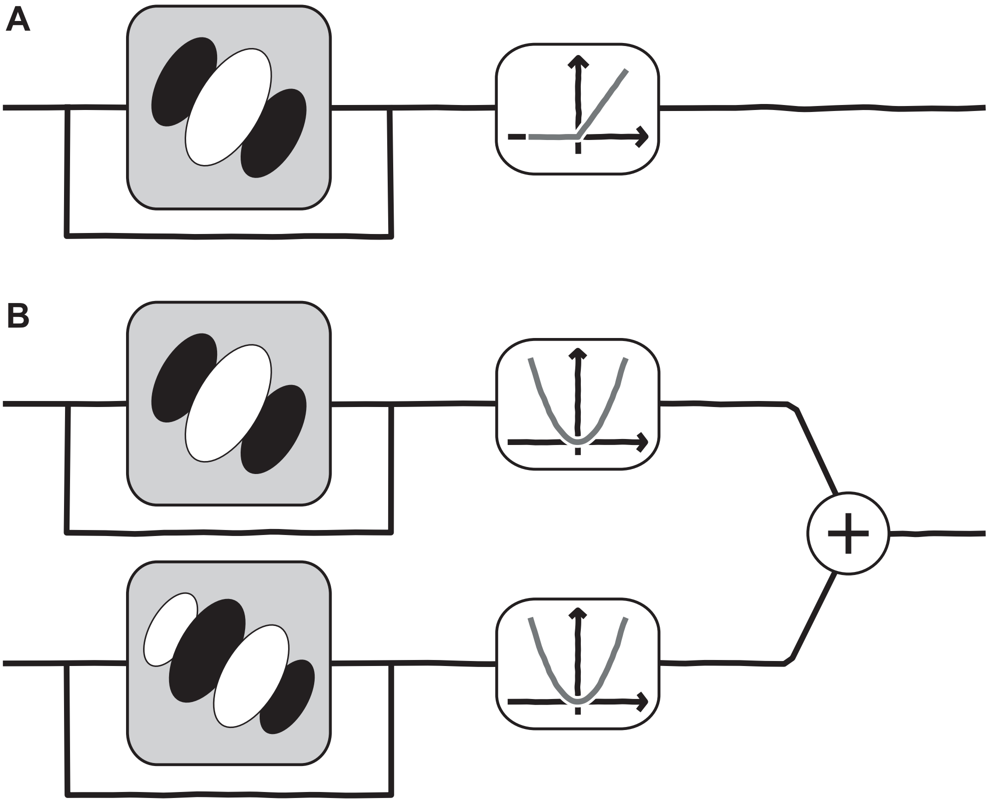


Figure 1.2. Computational model for simple and complex cells. The simple cell model (A) consists of one Gabor filter succeeded by a rectification nonlinearity. The filter output can be fed back for contrast gain control. The complex cell model (B) consists of two Gabor filters which only differ by a phase difference of 90°. The filter outputs are squared and summed up to produce the complex cell response. Adapted from Carandini et al. (2005).

particularly emphasizes that there is a strong relationship between algorithm and representation. Whether a representation is useful or not depends on the algorithm. For example, Arabic numbers are well suited for addition or multiplication while Roman numbers are not.

For the early visual system, the representation of the input is likely to reflect the scene statistics of the natural environment surrounding us. It is reasonable to assume that neural systems, especially sensory systems, are highly adapted to the statistics of the input signal (Attneave, 1954; Barlow, 1961; Field, 1987; Simoncelli and Olshausen, 2001). The dynamics of this adaptation may take place on different timescales: evolutionary over several generations, developmental during the forming of an individual, and behavioral at any point during the lifespan of an individual (Simoncelli and Olshausen, 2001). An example for evolutionary adaption to the input statistics would be the distribution of blue- and green-sensitive cones in mice (Nikonov et al., 2006; Baden et al., 2013), where blue-sensitive photoreceptors are mainly in the lower part and green-sensitive photoreceptor cells are mainly in the upper part of the retina, as the mouse field of view usually consists of plants on the bottom and sky on top. Contrast adaptation is an example of fast adaptation process that correlates with behavioral time scales. In contrast adaptation, the response to a low contrast stimulus is attenuated if the same stimulus with high contrast has been presented in the preceeding history (Kohn, 2007). Finally, the characteristic features of the primary visual cortex, like orientation tuning or bandpass filtering, are learned or at least refined during development (Hirsch and Spinelli, 1970, 1971; Blakemore and Cooper, 1970; Löwel and Singer, 1992; Wong, 1999; Albert et al., 2008). Kittens raised in an environment purely defined by vertical black and white stripes are virtually blind to horizontal bars, or vice versa. Even months after the kittens have been living in a natural environment, the selective blindness was still present and receptive fields selective to the non-exposed orientation were not found in their visual cortices.

Barlow (1981) stated that an important factor in understanding how the visual system works is understanding what limitations are imposed on it. These limitations can be within the system or external. Examples for limitations within the system are limited bandwidth of nerve fibers or different temporal response times of different cell types. External limitations are those depending on the habitat of animal, such as the large average viewing distance of a bird of prey high up in the sky or the refraction on the water surface for a fish hunting for above-surface food. Field (1987) and Burton and Moorhead (1987) were the first who took a detailed look at the statistics of natural images to gain a better understanding why neurons favor Gabor-shaped receptive fields. Based on the statement of Gibson (1950) that one must understand the nature of the environment to understand the nature of visual processing, Field investigated what the optimal code for natural images might be. He found that, in contrast to what was commonly believed, natural image statistics are not random but exhibit common properties independent of the scene. Natural images have a $^1/_{f^2}$ power spectrum, i.e. the power within one octave is constant over all frequencies (Kretzmer, 1952; Deriugin, 1956). Gabor filters with a bandwidth of 0.5 - 1.5 octaves showed to be optimal to encode natural images with the least amount of filters. This is very well within the range of physiological data (Ringach, 2002).

In order to adapt to the statistics of the input without additional guidance, the visual system has to have some kind of *unsupervised learning* mechanism implemented (Barlow, 1997). This closes the loop to Marr's tri-level hypothesis, as the unsupervised learning mechanism can be seen as the algorithmic level. The substrate level concerns the anatomy and biophysics of cells and will not be covered in this thesis. Here, the main question concerns the computational goal of complex cell representations which will be covered in detail later.

1.2 Invariance

Invariant feature representations are believed to be a key building block for high level vision tasks such as object detection and recognition. The appearance of objects changes dramatically with lighting, pose, and orientation. This requires a representation which is invariant to the infinite number of possible lighting and viewing conditions. In the visual system, invariance is believed to be constructed hierarchically (Riesenhuber and Poggio, 1999). Starting from the image on the retina, every stage creates a more invariant representation of the input (Figure 1.3). While retina and LGN, omitted in Figure 1.3, create brightness and contrast invariance, the complex cell layer is the first layer with geometrical invariance. To be precise, the complex cells provide invariance to the phase and thus, to some extent, the spatial position of the stimulus. This can be seen as an invariance under local transformations built from shifts. The *Neocognitron* was an early example of such a network model proposed by Fukushima (1980).

Evidence for a view-invariant object representation in the brain was found by Biederman and Cooper (1991) and several findings were recently reviewed by Biederman et al. (2009). However, there is still a big controversy if view-invariant (Biederman et al., 2009) or view-based (Tarr and Bülthoff, 1998) representations are used by the visual system. The interested reader is referred to extensive reviews on that topic (Logothetis and Sheinberg, 1996; DiCarlo et al., 2012).

1.2.1 Steerable filters

Transformation invariance is an extensively studied research area mathematics and machine learning. A well known example for a transformation invariant representation is the Fourier power spectrum of a signal. The Fourier power spectrum is invariant under global shifts with periodic boundary conditions, as these change only the phase of the Fourier components but not their (squared) amplitude. Essentially this invariance results from the well-known identity

$$(\sin kx)^2 + (\cos kx)^2 = 1 \qquad \forall x. \tag{1.1}$$

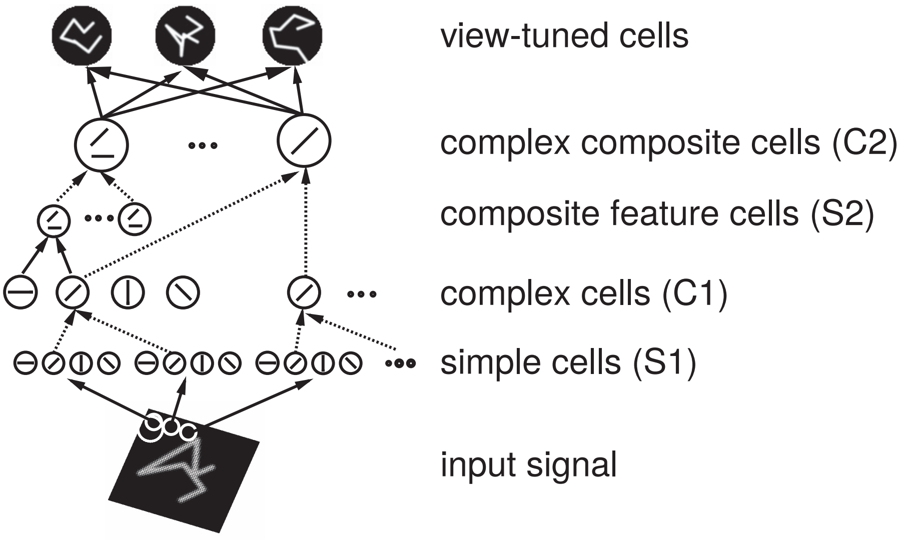


Figure 1.3. Model of hierarchical structure of the visual cortex. With each layer the representation of the feature becomes more invariant. While the first layer (S1, simple cells) does not have any geometrical invariance, the second layer (C1, complex cells) already shows spatial invariance with further increasing complexity of the object. The model can be extended further up to provide view-invariant object representations.
Adapted by permission from Macmillan Publishers Ltd: Nature Neuroscience. Riesenhuber and Poggio (1999) Figure 2, copyright 1999

Since all signals, for which a Fourier transform exists, can be represented as a combination of (possibly infinitely many) sine waves, the Fourier power spectrum provides an invariant representation under periodic shifts.

The field of steerable filter theory follows an analytical approach to invariance. First pioneered by Knutsson and Granlund (1983) and then refined and popularized by Freeman and Adelson (1991) and Granlund and Knutsson (1995), steerable filter theory provides a method to synthesize filters with certain properties from a set of basis filters. This allows an efficient computation of filter responses with arbitrary precision compared to a predefined filter bank (Perona, 1991, 1995). The simplest example of a steerable filter would be a sine wave pair, where two filters with identical spatial frequency but a phase offset of 90° are combined like in Eq 1.1 above to achieve a phase invariant response. Filters with identical spatial frequency and orientation which only differ in phase by 90° are called *quadrature pair*. The main motivation for steerable filters was orientation steerability (Freeman and Adelson, 1991), such that a small set of basis filters can synthesize any arbitrary orientation. The classical example is the two-dimensional isotrope Gaussian function $G(x,y) = e^{-(x^2+y^2)}$ and its directional derivatives $G_x^{(1)} = \frac{\partial}{\partial x}G(x,y)$ and $G_y^{(1)} = \frac{\partial}{\partial y}G(x,y)$ along the x- and y-axis, respectively. The two directional derivatives are the basis functions which can be used to synthesize the directional derivative along any direction θ via

$$G_\theta^{(1)} = \cos(\theta)\, G_x^{(1)} + \sin(\theta)\, G_y^{(1)} \tag{1.2}$$

with $\cos\theta$ and $\sin\theta$ as interpolation functions. In general, all functions which are a combination of a polynomial and a radially symmetric windowing function are steerable (Freeman and Adelson, 1991). Further, the steerable filters can be used in a multi-scale representation known as the steerable pyramid, which is widely used in image processing (Simoncelli and Freeman, 1995).

A couple of studies investigated how steerable filters can be learned in an unsupervised way (Rao and Ruderman, 1999; Miao and Rao, 2007; Bethge et al., 2007; Wang et al., 2009; Sohl-Dickstein et al., 2010).

1.2.2 Lie groups

Continuous transformation groups or Lie groups, named after Norwegian mathematician Sophos Lie who introduced them at the end of the 19th century, are a set of transformations which fulfill the algebraic group properties[1] and are differentiable. For every Lie group there exists a Lie algebra which is the tangent space at the identity element of the Lie group with the exponential map mapping the Lie algebra into the Lie group. The full theory and capabilities of Lie groups, Lie algebras and their wide variety of applications goes far beyond the scope of this thesis. In the following, I will give an intuition how Lie groups can be used for invariance learning. A complete and comprehensible introduction can be found in (Gilmore, 2008).

The interesting feature of a Lie group is that the complete group, i.e. all transformations it contains, can be generated from infinitesimal generators. A simple examples is the special orthogonal group $SO(2)$, the Lie group of all rotations around the origin in $\mathbb{R}^2$:

$$\underbrace{\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}}_{\text{rotation of angle } \varphi} = \exp\left(\varphi \underbrace{\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}}_{\text{infinitesimal generator}}\right) \tag{1.3}$$

[1] The characteristic properties of an algebraic group $(G, \cdot)$ are

associativity $f \cdot (g \cdot h) = (f \cdot g) \cdot h$

identity element $f \cdot e = f = e \cdot f$

inverse $f \cdot f^{-1} = e = f^{-1} \cdot f$

The scalar φ determines the angle of rotation, thus allowing to create any arbitrary rotation with the generator and the matrix exponential[2] only. This principle generalizes to all Lie groups in arbitrary dimensions, which makes them interesting with respect to transformation learning.

Lie groups have been used in various fields of computational vision (Hoffman, 1966; Dodwell, 1983; Nordberg et al., 1994; Van Gool et al., 1995) but mainly with predefined generators. Rao and Ruderman (1999) were the first who specified only a general Lie group structure but learned the specific operators from the data. They were able to learn Lie group operators in an unsupervised learning framework for infinitesimal 2D rotations and 1D translations which were artificially introduced to the data. To overcome the restriction on infinitesimal transformation, Miao and Rao (2007) extended the framework and learned a Taylor approximation of the generator, which was able to learn transformations from natural movie sequences as well, but was computationally expensive. Instead of learning the generator directly, Sohl-Dickstein et al. (2010) learned the eigenvectors and eigenvalues of the Lie group operator separately, reducing the matrix exponential to a simple exponential.

1.3 Invariance Learning

The visual system extracts invariant features from its input without an external supervisory signal. This requires some kind of *unsupervised learning* mechanism. In this thesis, I focus on *normative models* for unsupervised learning.

The main focus of normative models is answering *why* neurons respond as they do, thus providing valuable insight into the emergence of features. This is usually achieved by finding the optimal filters for a given objective function. Different objective functions make different predictions for the filter shapes. One can obtain hints about the underlying computational goal by verifying which of the filter shapes is more consistent with neural data. A further advantage is the interpretability and applicability of the results (Serre and Poggio, 2011). As the models emerged from signal processing and machine learning methods with simple structures their filter properties can be analyzed easily. For the same reason, although not directly neuroscience related, the gained insight can be transfered to applications like image/video compression. While normative models ignore many details about the biophysics of the neurons (for a list of biophysically plausible mathematical

[2] The matrix exponential is defined as $\exp(M) = \sum_{k=0}^{\infty} \frac{1}{k!} M^k$

operations see (Koch, 1999)), this abstraction may be acceptable as suggested by Marr's independence assumption regarding the three explanatory levels.

The number of possible objectives for normative models is infinitely large. For the purpose of this thesis, I focus on two distinct objective classes which have been proposed in the context of learning in the early visual system; the redundancy reduction objective (Barlow, 1961) aims to model the detailed statistical structure of the signal, and the slowness objective (Hinton, 1989) maximizes temporal stability or minimizes variations in the output over time. Both objective classes will be explained in the following sections.

1.3.1 Redundancy Reduction

The idea to explain neural representations based on a redundancy reduction principle was first proposed by (Barlow, 1961). Based on the concepts of information theory (Shannon and Weaver, 1949), he suggested that the visual system tries to remove redundant information from its internal representation of the environment. The inputs and outputs of neurons are therefore treated as random variables and the transformation is determined by the goal of minimizing the redundancy in the output. The statistical measure of redundancy between two random variables is the *mutual information*. It quantifies (in bits) how much information both random variables share, is strictly positive and 0 only if both random variables are statistically independent (Cover and Thomas, 1991). The extension to n random variables is called *multi information*. It is important to stress the difference between correlation and dependence or their counterparts uncorrelation and independence. Two random variables can be uncorrelated, i.e. have 0 covariance, but still be dependent. On the other hand, independent random variables are always uncorrelated.

The simplest approach to reduce redundancy is removing the second-order correlations from the input such that the covariance becomes a diagonal matrix. Decorrelation on natural images has successfully reproduced properties of the early visual system. It has been shown that decorrelation of the red, green, and blue color channels of natural images leads to the color opponency found in the retina (Buchsbaum and Gottschalk, 1983; Ruderman et al., 1998). Spatial and spatio-temporal decorrelation of natural images lead to bandpass filters similar to those found in LGN and retina (Atick and Redlich, 1990, 1992; van Hateren, 1992, 1993; Dong and Atick, 1995). However, decorrelation cannot explain the oriented filters found in the primary visual cortex nor the nonlinear features of complex cells or higher visual areas.

A classical approach to maximize statistical independence of the model output is *independent component analysis* (ICA). First proposed by Jutten and Herault (1991) in the context of blind source separation and later popularized by Comon (1994) and Bell and Sejnowski (1995), ICA is searching for a linear transformation of the input such that statistical dependencies of the output components are minimal. Formally, ICA assumes that the signal is a mixture of several independent sources plus noise. The mixture of independent components is defined in the mixture matrix. The ICA matrix is (up to scale and order of the components) identifiable under the assumptions that the output components are statistically independent, have non-Gaussian distributions and the matrix is square (Comon, 1994). When the input to the ICA matrix is white, i.e. decorrelated and with unit variance, the optimization is reduced to the set of orthogonal matrices, which significantly simplifies the optimization. Several methods have been proposed to find the orthogonal matrix, such as minimizing the multi information (Comon, 1994) or maximum likelihood estimation (Pham and Garat, 1997). Hyvärinen and Oja (2000) give a comprehensive tutorial on ICA including various methods to find the independent components with an update published recently (Hyvärinen, 2013).

On a practical level, maximizing the independence of filter responses can be described as maximizing the sparseness of the filter outputs. For a random variable, such as the output of a filter, sparseness is defined as being more likely to take very small absolute values or very large absolute values than a Gaussian random variable with identical variance. Distributions with this property are called *sparse*, *super-Gaussian*, or *leptokurtotic*. An example for a sparse distribution is the Laplacian distribution, which is shown together with a Gaussian distribution in Figure 1.4. The Laplacian (red) has a higher peak at the mean but then drops off faster than the Gaussian (blue). For large values, however, the Gaussian converges faster towards 0 than the Laplacian. One measure of sparseness is *kurtosis*, which is the 4th moment of the distribution divided by the squared variance.

Sparseness in neuron ensembles can be seen in two ways; a single neuron which is only active very rarely shows *lifetime sparseness*, while a population of neurons where only few neurons are active at any point in time exhibits *population sparseness* (Willmore and Tolhurst, 2001). A favorable side eddect of the sparseness objective is that low activity levels yields low energy consumption (Lennie, 2003).

An intuition about a link between sparseness and independence can be shown with the central limit theorem (CLT). The CLT says that if we add up infinitely many independent random variables, we would end up with a Gaussian random variable. More specifically, it holds that the sum of two independent random variables is more Gaussian, i.e. less sparse, than each of the two independent ones. Hence, if the data one generated is a linear mixture of independent components, then the maximally sparse components must be the independent components (Hyvärinen et al., 2009).

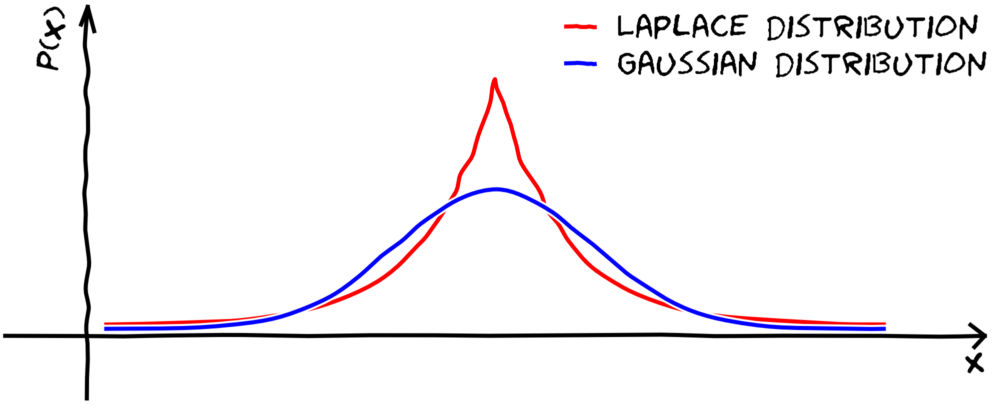


Figure 1.4. Gaussian vs Laplacian distribution. Both distributions have equal mean and variance. The Laplacian distribution (red) has a higher peak at the mean than the Gaussian distribution (blue). The Laplacian first drops off faster but at long distance from the mean the Laplacian converges slower towards 0 than the Gaussian, resulting in a higher probability for large values. Distributions with these properties are called *sparse*, *leptokurtotic*, or *super-Gaussian*.

Olshausen and Field (1996, 1997) used sparseness maximization on natural images in an overcomplete representation and obtained simple cell-like receptive fields. Plain ICA, where it is assumed that the number of independent sources equals the number of dimensions in the data, was first applied to natural images by Bell and Sejnowski (1997), who showed that the obtained filters have simple cell properties as well. Van Hateren and colleagues did a quantitative analysis of the filter properties obtained from still images (van Hateren and van der Schaaf, 1998) and image sequences (van Hateren and Ruderman, 1998) and compared them to simple cell recordings from macaque visual cortex.

As all redundancy reduction models in this thesis have been linear models so far, they cannot account for the nonlinear complex cell features. Hyvärinen and Hoyer (2000) were the first to show that complex cell like features can be learned via redundancy reduction using *independent subspace analysis* (ISA). ISA uses the same objective as ICA but instead of a linear filter the objective is evaluated on the squared radial component of an n-dimensional subspace (Kohonen, 1996). This corresponds to ICA on the output of the energy model (Adelson and Bergen, 1985) with n linear filters.

Hyvärinen et al. (2001) introduced a relaxation of the hard separation into n-dimensional subspaces via a topographical organization of the components. The output space is organized on a 2D grid with each node representing the squared output of one linear ICA filter. This can be seen as a simple cell layer. On top of the simple cell layer is a second layer, the complex cell layer, which pools over a topographically local area of simple cell outputs. With the pooling area fixed, the complex cell layer output is used as input for the unsupervised learning of the ICA filters. The obtained complex cell responses are not only phase invariant, but the topographical organization also resembles the organization in cortex (Hyvärinen et al., 2001).

1.3.2 Slowness

The classical motivation for slowness is the observation that one looks at an object such as a zebra and the animal moves perpendicular to the viewing direction, the local light intensities vary largely between the stripes but the global percept is still a zebra. The position of the animal over time changes a lot slower than the light intensities of single receptors on the retina. Thus the visual system could try to extract the perceptually stable structure within the rapidly varying input stream by searching for the slowest components. Therefore "slowness", "temporal stability", or "temporal smoothness", may serve as an objective function for invariant representation learning.

The idea of using slowness as a learning objective was first stated by Hinton (1989) in the context of learning in neural networks. The first implementation in a neural network was given by Földiák (1991) and is called the *trace rule*. The trace rule was able to obtain shift invariance for simple moving bar stimuli. Slowness has also been used to learn invariant subspaces (Kayser et al., 2001; Körding et al., 2004) using the energy model (Adelson and Bergen, 1985) similar to ISA (Hyvärinen and Hoyer, 2000). Their objective is minimizing the variance of the temporal derivative of the squared subspace response divided by the variance of the subspace response. The learned subspaces were phase and position invariant, similar to V1 complex cells.

The best known implementation of the slowness principle is Slow Feature Analysis (SFA) (Wiskott and Sejnowski, 2002). The algorithm has been applied to image sequences with artificial induced transformations to obtain features which reproduce complex cell properties (Berkes and Wiskott, 2005). The objective of SFA is very similar to the objective defined by Kayser et al. (2001) but instead of the subspace energies SFA uses a nonlinear expansion of the feature space. The maximally slow features in the nonlinear feature space are then defined as generalized eigenvectors as in oriented PCA (Diamantaras and Kung, 1994; Bethge et al., 2007). This allows one to use standard eigenvector solvers to find the SFA components. For complex cell learning the nonlinear expansion is the expansion into the quadratic feature space, the space of all monomials of degree 1 and 2. This roughly squares the input dimensionality and is thus usually combined with significant dimensionality reduction. SFA has also been used in hierarchical networks, where each layer applies SFA to the outputs of the previous SFA layer. With hierarchical approaches, SFA was able to reproduce properties like place and grid cells (Franzius et al., 2007) or invariant object recognition (Franzius et al., 2011). Sprekeler et al. (2007) showed that SFA is in fact identical to the trace rule used by Földiák (1991). For a recent review on SFA and all its applications see (Wiskott et al., 2011).

Hurri and Hyvärinen (2003) provide a slightly different definition of slowness. Their objective is called *temporal coherence* and is defined as the correlation between the model outputs of consecutive time steps, which has to be maximized. The model consists of a simple linear filter with static point-wise nonlinear function. If the temporal distance between the two time steps approaches 0, temporal coherence becomes identical to ICA with kurtosis maximization as objective. The significant difference to the previous slowness approaches is that the temporal coherence principle is a forth-order rather than a second-order objective function and thus leads to simple cell properties.

1.3.3 Combination of slowness and redundancy reduction

A straight forward idea is to combine both, the redundancy reduction and slowness objective. The approach of Einhäuser et al. (2002) is a three-layer neural network where the first layer are the input pixel, the middle layer optimizes for sparseness, and the top layer for slowness. The middle layer showed classical simple cell receptive field properties while the top layer was largely invariant under phase and position of a stimulus.

A direct combination of sparseness and temporal correlation is the *bubbles* framework by Hyvärinen et al. (2003). The framework consists of two layers, the first layer are spatial or spatio-temporal linear filters and a second pooling layer whose nodes pool over a predefined spatio-temporal area. The filter outputs are optimized for three different objectives simultaneously; temporal coherence is the correlation of the squared outputs of the same filter in time, energy correlation is the correlation of the squared output of two different filters at the same point in time, and a sparseness objective on each filter. This leads to an extension of topographical ICA (Hyvärinen et al., 2001) from spatially active circles to spatio-temporal activity bubbles. While the first layer corresponds to simple cells, the second layer, called bubble detector, corresponds to complex cells.

Closely related to the bubbles framework is the study by Berkes et al. (2009). Their model consists of binary *identity* variables which indicate the presence or absence of a certain feature in the scene. Each feature is a mixture of linear filters where the mixture weights are called *attributes*. All 3 parameter sets - filters, identities, and attributes - are learned from natural image sequences using a Bayesian optimization where the prior was defined such that the components are spatially independent and temporally smooth. This approach differs from the bubbles framework in so far as here the identity is binary while its counterpart the bubble detector is a continuous variable, and the attributes are continuous while their counterpart, the pooling area, has fixed weights of 1. Further, the filters of the attributes are neither topographically ordered nor overlapping, but every attribute has its independent set of filters of variable size.

Cadieu and Olshausen (2008, 2009, 2012) used a different approach. They propose a two-layer model where the first layer represents local features and the second layer groups the local features to encode form and motion. As form and motion are beyond V1, the first layer is of interest here. The first layer consists of a complex-valued sparse coding layer where the amplitudes of the complex-valued output are optimized for sparseness and slowness simultaneously. The real and imaginary part of the filter correspond to the two filters of the energy model (Adelson and Bergen, 1985) and resemble complex cell receptive fields.

1.4 Receptive field model

Receptive fields of cells in the primary visual cortex are mapped out using, for example, reverse correlation, spike-triggered average, or spike-triggered covariance. But to work with the obtained receptive field, we have to find a compact mathematical description. Movshon et al. (1978b) found that the V1 simple cell receptive fields compute weighted sums of their input. Shortly after that Marcelja (1980) showed that 1D Gabor functions (Gabor, 1946a,b,c) are well suited to describe the weighting profile of simple cells. Daugman (1980, 1985) extended the model to 2D Gabor functions which have been used in countless studies since. An alternative to the Gabor function is the Gaussian Derivative or Hermite function (Young, 1978; Young et al., 2001; Young and Lesperance, 2001). I will present both models with their individual advantages and disadvantages in the following subsections and motivate why the Gabor model is used in this thesis.

1.4.1 Gabor filter

Gabor filters have been defined by Gabor (1946a,b,c) in the context of radio transmission as intermediate representation between frequency and spatio-temporal representations. They are defined as the product of a sinusoidal carrier wave and a Gaussian envelope,

$$g\left(x|\omega,\sigma,x_0,\phi\right)=\cos\left(2\pi\omega\left(x-\phi\right)\right)e^{-\frac{\left(x-x_0\right)^2}{2\sigma^2}}, \tag{1.4}$$

where ω and ϕ are the spatial frequency and phase, and σ and x_0 are the width and spatial position, respectively. In the Fourier domain, the Gabor filter is the convolution of delta peaks at $\pm\,\omega$ with a Gaussian envelope of width $1/\sigma$, resulting in a frequency spectrum of two Gaussians with envelope size $1/\sigma$ at $\pm\,\omega$. Gabor filters are widely used in image processing, e.g. Lee (1996); Daugman (1988); Jain and Bhattacharjee (1992); Weldon et al. (1996); Hamamoto et al. (1998); Kamarainen et al. (2006) to name a few.

The simplicity of the Gabor filter is one of its main advantages. The parameters ω and σ provide intuitive understanding of the shape and properties of the filter. They further allow any combination of parameters to create an arbitrary tiling of the space-frequency domain. This allows perfect tuning to the requirements of any signal processing task.

However, Gabor filters also have notable downsides. Depending on the spatial frequency and envelope size, Gabor filters can have a significant DC component. This means that the filter contains information about the mean value of the input. This can be addressed by removing the DC component from the filter but this naturally changes the filter properties. Further, Gabor filters of identical spatial frequency and envelope size but with an offset of 90°, i.e. a quadrature pair, are not perfectly orthogonal.

1.4.2 Hermite filter

Hermite filters are the combination of Hermite polynomials (Hermite, 1864) with a Gaussian envelope. Hermite polynomials are defined by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}. \tag{1.5}$$

To ensure orthogonality the Hermite polynomials have to be normalized, which leads to the Hermite filter definition of Martens (1990, 1997):

$$d_n(x) = \frac{(-1)^n}{\sqrt{n!2^n\pi\sigma^2}} H_n\left(\frac{x}{\sigma}\right) e^{\frac{-x^2}{\sigma^2}} \tag{1.6}$$

where n is the degree of the Hermite polynomial and σ is the width of the envelope. Young (1978) showed that 1D Hermite filters match 1D representations of V1 receptive fields and later Young (1986) extended this concept to 2D Hermite filters.

Intuitively, Hermite filters are the n-th order derivative of a Gaussian function

$$g_n(x) = \frac{d^n}{dx^n} e^{-\frac{x^2}{2}} \tag{1.7}$$

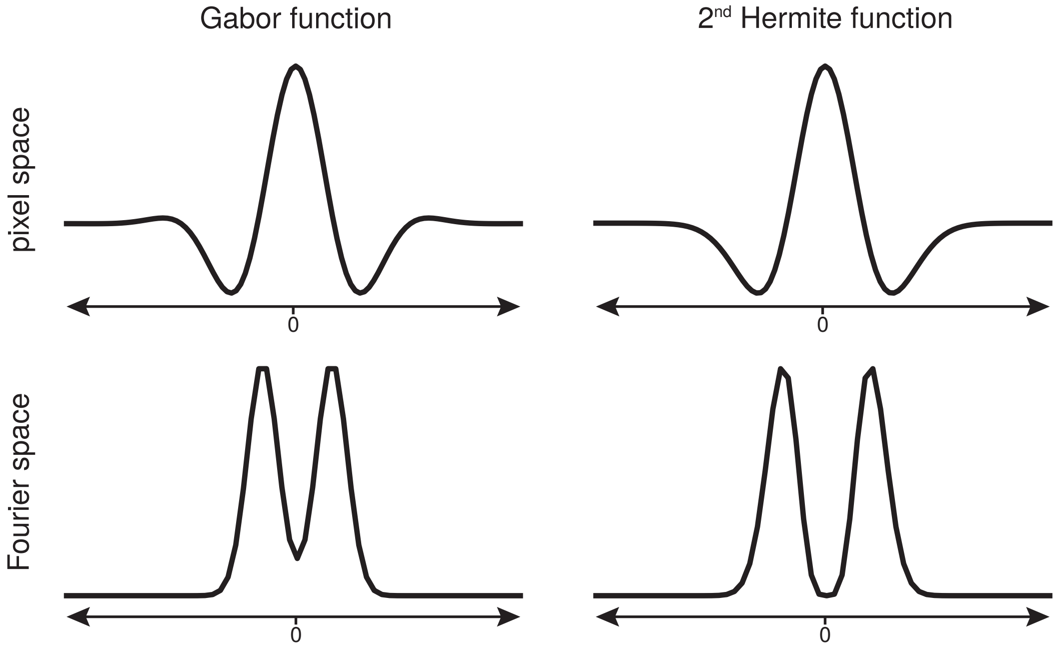


Figure 1.5. Example of Gabor and Hermite function in pixel and Fourier space. The left side shows an even-symmetric Gabor function with three visible subfields - the peak at the center pixel and the two side troughs. Below the Gabor is the amplitude spectrum of its Fourier transform with the DC component at 0 on the abscissa. The baseline of the amplitude spectrum corresponds to 0 energy at that frequency. On the right side is a 2nd degree Hermite function and its amplitude spectrum for comparison. In pixel space, the differences are quite minimal. If we look closely, there is a small bump next to the troughs for the Gabor function. The Hermite function does not show this bump. The difference is more prominently visible in the amplitude spectrum. While the position and the width of both peaks are comparable, the Gabor function has a value significantly larger than 0 for the DC component, i.e. the function conveys information about the average value of the filtered signal. In contrast, the amplitude spectrum of the Hermite function drops to 0 at the DC component thus only conveys information around the peak spatial frequency.

with appropriate scaling. However, Gaussian derivative filters are not identical to Hermite filters but differ by a scale factor.

The advantage of Hermite filters is that the DC component is 0 for all filters with degree larger than 0. This is illustrated in Figure 1.5. On the left side is a Gaussian function in pixel space and its amplitude spectrum and on the right side a 2nd degree Hermite function and its amplitude spectrum. While both pixel space representations look quite similar, the amplitude spectrum reveals that the Gabor function conveys a significant amount of information about the DC component while the amplitude spectrum of the Hermite function drops to 0 at the DC component. Also, Hermite filters of different degrees are always orthogonal due to the normalization. This allows to create a perfectly orthogonal filter basis. One disadvantage of Hermite filters is that the frequency selectivity around the peak spatial frequency ω is slightly antisymmetric, thus the filter does not respond equally to stimuli with frequency $\omega + \epsilon$ and $\omega - \epsilon$. Further, Hermite filters are computationally more complex and less intuitive in the interpretation of their parameters.

1.4.3 Comparison

A comparison of Hermite and Gabor filters for cortical data has been published by Young et al. (2001); Young and Lesperance (2001) and a theoretical comparison between Hermite and Gabor filters has been published by Rivero-Moreno and Bres (2003). The studies found that the performance of both models is comparable and lead to very satisfactory approximations of V1 receptive fields.

In preliminary simulations of our complex cell model with both receptive field models, we also found comparable results with marginally better performance of the Gabor filters. Further, the Hermite models require more complex computations and provide less intuitive insight into the filter properties. For example, the Hermite polynomial requires the computation of the factorial function

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} = n! \sum_{m=0}^{\lfloor n/2 \rfloor} \frac{(-1)^m}{m!(n-2m)!} (2x)^{n-2m} \tag{1.8}$$

which is computationally more complex and less numerically stable than the multiplication of a cosine and an exponential. Also, the concept of envelope size and peak spatial frequency allows easier comparisons than degree and envelope size.

Given that there was no significant difference and the more intuitive and less computationally complex design of the Gabor filter we decided to use Gabor receptive field models in this thesis.

1.5 Goals of this thesis

The main goal of this work is to provide a quantitative comparison of the two unsupervised learning paradigms *redundancy reduction* and *slowness* with respect to complex cell properties. Both objectives have been proposed as principle underlying the formation of complex cell receptive fields in the primary visual cortex. However, it turns out that the two objectives lead to rather different predictions for the filter properties.

To compare both objectives on common ground, I present how the slowness principle can be used in a subspace framework to learn invariant representations from artificial translations and natural movies. So far, the different studies on redundancy reduction and slowness varied largely in data sets, preprocessing, experimental parameters, feature space, evaluation methods, and many more details. The framework I use is identical to the one used by ISA (Hyvärinen and Hoyer, 2000) and physiologically well motivated (Pollen and Ronner, 1983). This allows a fair comparison between the redundancy reduction objective and the slowness objective as both are working with the same model structure on the identical feature space. I show how slowness and redundancy reduction lead to completely different filter sets when optimized under the same conditions. Subsequently I present a weighted combination of slowness and redundancy reduction objective and show how the filter sets change with the weighting factor.

Slowness has mainly been used through the implementation of SFA. However, while SFA has advantages in cases where, for example, a closed-form solution is required, Slow Subspace Analysis (SSA) as defined by Kayser et al. (2001) has several advantages over SFA. I show the differences and commonalities of SFA and SSA and discuss why SSA is favorable for complex cell modeling. Therefore SSA is our algorithm of choice for a comparison with redundancy reduction.

Localization is a key feature of complex cells in the primary visual cortex and a feature found by both redundancy reduction and slowness studies. However, steerable filter theory suggests that global filters are perfect for translations. Therefore, it is worthwhile to investigate if slowness or redundancy reduction can account for localization. Further physiological properties found in monkey (Ringach, 2002) and cat (Jones and Palmer, 1987a) are the scaling of preferred spatial frequency with envelope size (*wavelet scaling)* and the aspect ratio of the envelope. These three features form an interesting set of receptive field properties which have not been quantitatively evaluated for slowness and redundancy reduction at the same time.

To compare both objectives with physiological data, I use an approach similar to Field and Tolhurst (1986) for simple cells. The idea is to model the V1 complex cells by the energy model with two quadrature phase Gabor filters with fixed position, orientation and spatial frequency. The remaining free parameters are the horizontal and vertical Gaussian envelope size. If we now maximize the objective of the respective paradigm we can compare the optimal envelope size parameters with empirically found parameters. The advantage of the single cell approach is avoiding mutual influence, thus providing a clearer insight into the truly optimal filter for both objectives.

2 Slowness and sparseness have diverging effects on complex cell learning

This article is joint work of Jörn-Philipp Lies, Ralf M. Häfner, and Matthias Bethge. It was submitted on 21 April 2013 to PLoS Computational Biology. All simulations and computations as well as the documentation of methods and results including figures were done by JPL. The design of the experiments, the evaluation of the results and the discussion were jointly done by all 3 authors.

The article is contained as submitted with only 3 changes, namely the citation style (using author name and year instead of numbers), the figures are at the position in the text where they are referenced instead of at the end of the article, and the bibliography is at the end of the thesis. All changes are for enhanced readability only and do not alter the content.

2.1 Abstract

A key question in visual neuroscience is how neural representations achieve invariance against appearance changes of objects. In particular, complex cells are often interpreted as a signature of an invariant coding strategy. Following earlier studies which showed that a sparse coding principle may explain the receptive field properties of complex cells, it has been concluded that the same properties may be equally derived from a slowness principle. Here we show that—contrary to widespread belief—slowness and sparsity drive the representations towards substantially different receptive field properties. To do so, we present complete sets of basis functions learned with *slow subspace analysis* (SSA) in case of natural movies as well as translations, rotations, and scalings of natural images. SSA directly parallels independent subspace analysis (ISA) with the only difference that SSA maximizes slowness instead of sparsity. We find a large discrepancy between the filter shapes learned with SSA and ISA. We argue that SSA can be understood as a generalization of the Fourier transform where the power spectrum corresponds to the maximally

slow subspace energies in SSA. Finally, we investigate how much performance can be achieved if one optimizes for both slowness and sparsity simultaneously and how this trade-off effects the filter shapes.

2.2 Author Summary

A key question in visual neuroscience is how neural representations achieve invariance against appearance changes of objects. In particular, invariance of complex cell responses against small translations is commonly interpreted as a signature of an invariant coding strategy possibly originating from an unsupervised learning principle. Various models have been proposed to explain the response properties of complex cells using a sparsity or a slowness criterion and it has been concluded that physiologically plausible receptive field properties can be derived from either criterion. Here, we show that the effect of the two objectives on the resulting receptive field properties is in fact very different. We conclude that slowness alone cannot explain the filter shapes of complex cells and discuss how both slowness and sparsity may be two signatures of a representation that is adapted to a complex generative model of vision for which we still need to gain a much better understanding.

2.3 Introduction

The appearance of objects in an image can change dramatically depending on their pose, distance, and illumination. Learning representations that are invariant against such appearance changes can be viewed as an important preprocessing step which removes distracting variance from a data set in order to improve performance of downstream classifiers or regression estimators (Burges, 2005). Clearly, it is an inherent part of training a classifier to make its response invariant against all within-class variations. Rather than learning these invariances for each object class individually, however, we observe that many transformations such as translation, rotation and scaling apply to any object independent of its specific shape. This suggests that signatures of such transformations exist in the spatio-temporal statistics of natural images which allow one to learn invariant representations in an unsupervised way.

Complex cells in primary visual cortex are commonly seen as building blocks for such invariant image representations (e.g. (Riesenhuber and Poggio, 1999)). While complex cells, like simple cells, respond to edges of particular orientation they are less sensitive to the precise location of the edge (Hubel and Wiesel, 1962). A variety of neural algorithms have been proposed that aim at explaining the response properties of complex cells as components of an invariant representation that is optimized for the spatio-temporal statistics of the visual input (Hyvärinen and Hoyer, 2000; Hyvärinen et al., 2001; Berkes and Wiskott, 2005; Karklin and Lewicki, 2009; Berkes et al., 2009; Kayser et al., 2001; Einhäuser et al., 2002; Kayser et al., 2003; Körding et al., 2004).

The two main objectives used for the optimization of the neural representations are *sparseness* and *slowness*. At first sight, the slowness objective seems to be more directly related to invariance learning than the sparsity objective: While for natural signals it may be impossible to find perfectly invariant representations, one seeks to find features that at least change as slowly as possible under the appearance transformations exhibited in the data (Sutton and Barto, 1981; Klopf, 1982; Földiák, 1991; Mitchison, 1991; Stone and Bray, 1995; Stone, 1996; Wallis and Rolls, 1997; Kayser et al., 2001; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Kayser et al., 2003; Hurri and Hyvärinen, 2003; Körding et al., 2004; Berkes and Wiskott, 2005; Spratling, 2005; Maurer, 2006; Turner and Sahani, 2007; Masquelier et al., 2007; Maurer, 2008).

In contrast to sparse representation learning which is tightly linked to generative modeling, many slow feature learning algorithms follow a discriminative or coarse-graining approach: they do not aim at modeling all variations in the sensory data but rather classify parts of it as noise (or some dimensions as being dominated by noise) and then discard this information. This is most obvious in case of slow feature analysis (Wiskott and Sejnowski, 2002). It can be seen as a special case of oriented principal component analysis which seeks to determine the most informative subspace under the assumption that fast changes are noise (Bethge et al., 2007). While it is very likely that some information is discarded along the visual pathway, throwing away information in modeling studies requires great caution. For example, if one discards all high spatial frequency information in natural images one would easily obtain a representation which changes more slowly in time. Yet, this improvement in slowness is not productive as high spatial frequency information in natural images cannot be equated with noise but often carries critical information.

Thus, for better comparability of different unsupervised learning algorithms it is often useful to exclude the possibility of information reduction. More specifically, if we want to compare the effect of slowness and sparseness on complex cell learning this can be done most easily by comparing complete sets of filters learned with *slow subspace analysis* (SSA) (Kayser et al., 2001) and *independent subspace analysis* (ISA) (Hyvärinen and Hoyer, 2000), respectively. The two algorithms are completely identical with the only difference that SSA maximizes slowness while ISA maximizes sparsity.

While for sparseness it is common to show complete sets of filters this is not so in case of slowness. Based on the analysis of a small subset of filters, it has been argued that SSA may generally yield similar results to ISA (Kayser et al., 2001). In contrast, we here arrive at quite the opposite conclusion: by looking at the complete representation we find a large discrepancy between the filter shapes derived with SSA and those derived with ISA.

Complete representations optimizing slowness have previously been studied only for mixed objective functions that combined slowness with sparseness (Hyvärinen et al., 2003; Cadieu and Olshausen, 2009, 2012; Berkes et al., 2009) but never when optimizing exclusively for slowness alone. Here we systematically investigate how a complete set of filters changes when varying the objective function from a pure slowness objective to a pure sparsity objective by using a weighted mixture of the two and gradually increasing the ratio of their respective weights. From this analysis we will conclude that the receptive field shapes shown in (Hyvärinen et al., 2003; Cadieu and Olshausen, 2009, 2012; Berkes et al., 2009) are mostly determined by the sparsity objective rather than the slowness objective. That is the receptive fields would change relatively little if the slowness objective was dropped but it would change drastically if the sparsity objective was removed. These findings changes our view of the effect of slowness and raises new questions that can guide us to a more profound understanding of unsupervised complex cell learning.

2.4 Results

The main take home message of the paper is the observation that the effect of the slowness objective on complex cell learning is substantially different from that of sparseness. Most likely this has gone unnoticed to date because previous work either did not derive complete representations from slowness or combined the slowness objective with a sparsity constraint which masked the genuine effect of slowness. Therefore, we here put large effort into characterizing the effect of slow subspace learning on the complete set of filter shapes under various conditions. We first study a number of analytically defined transformations such as translations, rotations, and scalings before we turn to natural movies and the comparison between slowness and sparseness.

The general design of SSA is illustrated in Figure 2.1. We apply a set of filters to the input $\mathbf{x}(t)$ and square the filter responses. Two filters form a 2-dimensional subspace (gray box in Figure 2.1) and the sum of squared filter responses of these two filters yield the subspace

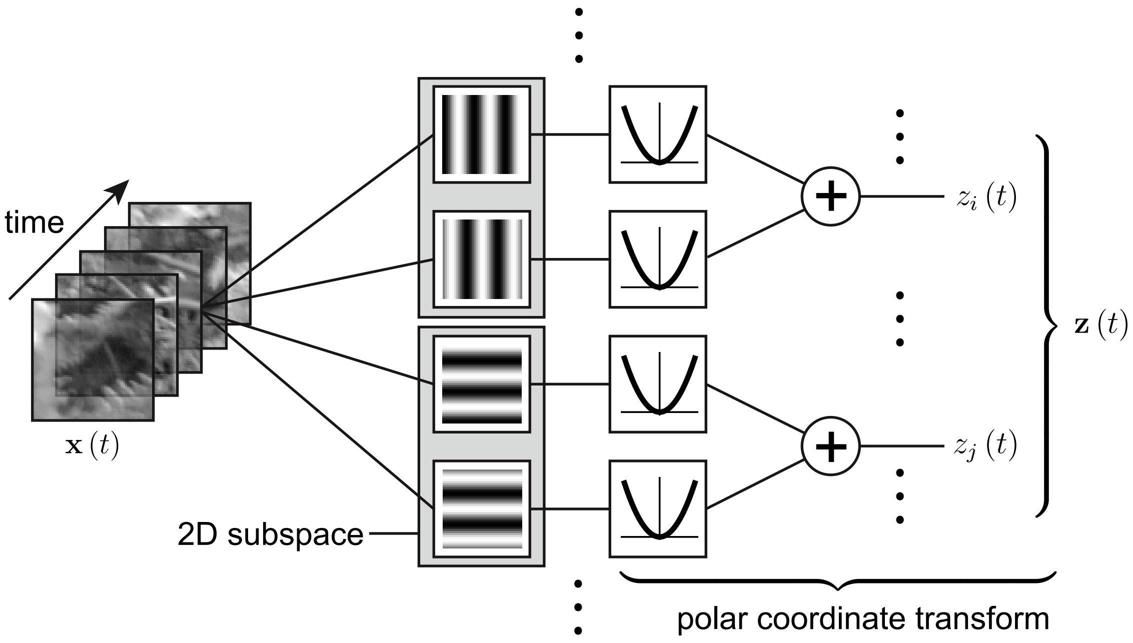


Figure 2.1. Illustration of slow subspace analysis (SSA). This Figure shows the energy model structure of SSA. The input signal, e.g. a movie sequence, is applied to several filters. Two filters form a subspace. The output of the filters are passed through a quadratic nonlinearity before the responses of the filters within one subspace are summed up. The output corresponds to the radial component of the 2D subspace. The $n/2$ responses $z_i(t)$ then form the multidimensional output signal $\mathbf{z}(t)$. If the filters are the discrete Fourier transform basis where each subspace consists of the two filters which only differ in phase, then the output $\mathbf{z}(t)$ is the power spectrum of the input signal $\mathbf{x}(t)$.

energy response. This can be seen as the squared radial component of the projection of the signal into the 2D subspace formed by the two respective filters. For example, if the filters are taken from the Fourier basis and grouped such that the two filters within each subspace have the same spatial frequency and orientation and 90° phase difference, the SSA output $\mathbf{z}(t)$ at a fixed time instant t is the power spectrum of the image $\mathbf{x}(t)$. As input $\mathbf{x}(t)$ we used 11×11 image patches sampled from the van Hateren image database (van Hateren and van der Schaaf, 1998) and from the video database (van Hateren and Ruderman, 1998), vectorized to 121-dimensions, and applied SSA to all remaining 120 AC components after projecting out the DC component.

In the first part of our study, the input sequence consisted of translations. As time-varying process for the translations, we implemented a two-dimensional random walk of an 11×11 window over the full image. The shift amplitudes were drawn from a continuous uniform distribution between 0 and 2 pixels, allowing for subpixel shifts. The filters obtained are shown in Figure 2.2A. Each row contains the filter pairs of 6 subspaces, sorted by descending slowness from left to right and top to bottom. The filters clearly resemble global sine wave functions. The wave functions differ in spatial frequency and orientation between the different subspaces. Within each subspace, orientation and spatial frequency are almost identical, but phases differ significantly. In fact, the phase difference is close to $90°$ ($90.2° \pm 3.8°$), resembling quadrature pairs of sine and cosine functions as it is the case for the two-dimensional Fourier basis. Accordingly, the subspace energy output $\mathbf{z}(t)$ of the resulting SSA representation is very similar to the power spectrum of the image $\mathbf{x}(t)$.

In fact, one can think of SSA as learning a generalized power spectrum based on a slowness criterion. While the power spectrum is known to be invariant against translations with periodic boundary conditions, perfect invariance—or infinite slowness—is not achieved for the translations with open boundary conditions studied here (see Figure 2.2 B). The slowness criterion is best understood as a penalty of fast changes since it decomposes into an average over penalties of fast changes for each individual component (see methods). Therefore, we will always show the inverse slowness v for each component such that the *smaller* the area under the curve the *better* the average slowness.

The decrease in v, i.e. the increase in slowness, is substantial: the average inverse slowness $\langle v \rangle$ decreases approximately by a factor of three. The low frequency subspaces are clearly the slowest subspaces, and slowness decreases with increasing spatial frequency. At the same time, however, the inverse slowness of all learned subspaces is still larger than 0, i.e. even for the slowest components, perfect invariance is not achieved. This is not surprising, as perfect invariance is impossible whenever unpredictable variations exist as it is the case for open boundary conditions.

In Figure 2.2 C, we show that SSA can indeed find perfectly invariant filter starting from a random initial filter set if one imposes periodic boundary conditions. To this end, we created 11×11 pink noise patches with circulant covariance structure, i.e. the pixels on the left border of the image are correlated with pixels on the right border as if they were direct neighbors. As time-varying process, we implemented a random walk with cyclic shifts where the patches were translated randomly with periodic boundary conditions. As in the previous study, the shift amplitudes were drawn from a continuous uniform distribution between 0 and 2 pixels. Since the Fourier basis is the eigenbasis of the cyclic shift operator it should yield infinite slowness for the cyclic boundary conditions. Indeed, the filters learned from these data recover the Fourier basis with arbitrary precision. Perfect invariance is equivalent with the objective function converging to 0. This means that the response of each subspace is identical for all shifts. Figure 2.2D shows the inverse

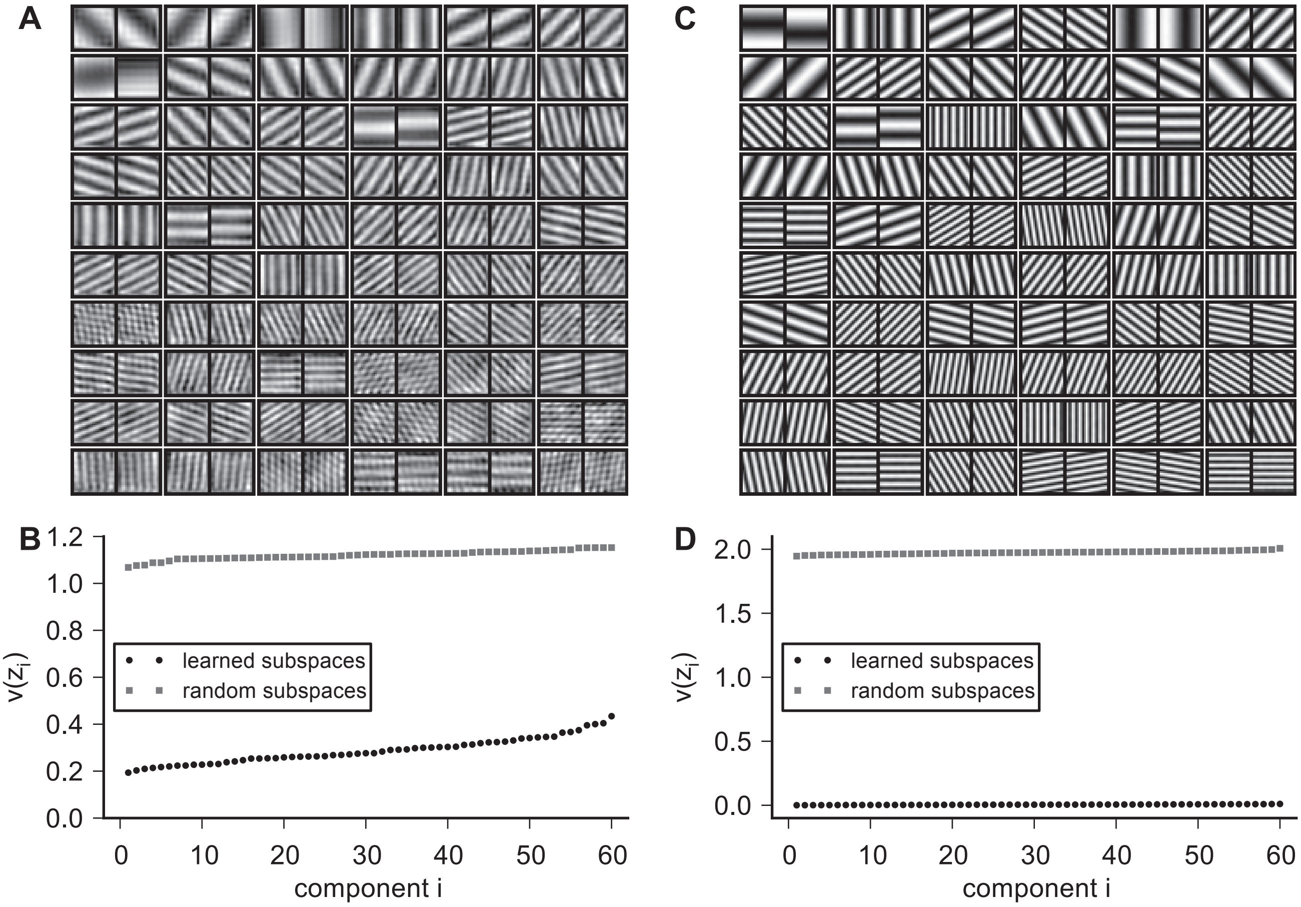


Figure 2.2. SSA on translations with open and cyclic boundary conditions. The complete set of filters learned from translated images with open and cyclic boundary conditions are shown in (A) and (C), respectively. Each row shows the filters of 6 subspaces with 2 dimensions. The subspaces are ordered according to their slowness, with the slowest filter in the upper left corner and decreasing slowness from left to right and top to bottom. The *inverse* slowness v for the individual subspaces after learning (black dots) and for the initial random filters (gray squares) is shown in (B) and (D), respectively. For open boundary conditions (B), the inverse slowness does not converge to 0, hence perfect invariance is not achieved. For cyclic shifts, however, the inverse slowness approaches 0 with arbitrary precision (D), indicating convergence to perfect invariance.

slowness v of the individual components. For all filters, v is very small ($< 10^{-3}$), close to perfect invariance and infinite slowness.

Given that the SSA representation learned for translations is very similar to the Fourier basis and since the Fourier basis achieves perfect invariance for cyclic shifts we proceeded to investigate whether the Fourier basis is optimal even for non-cyclic translations as well.

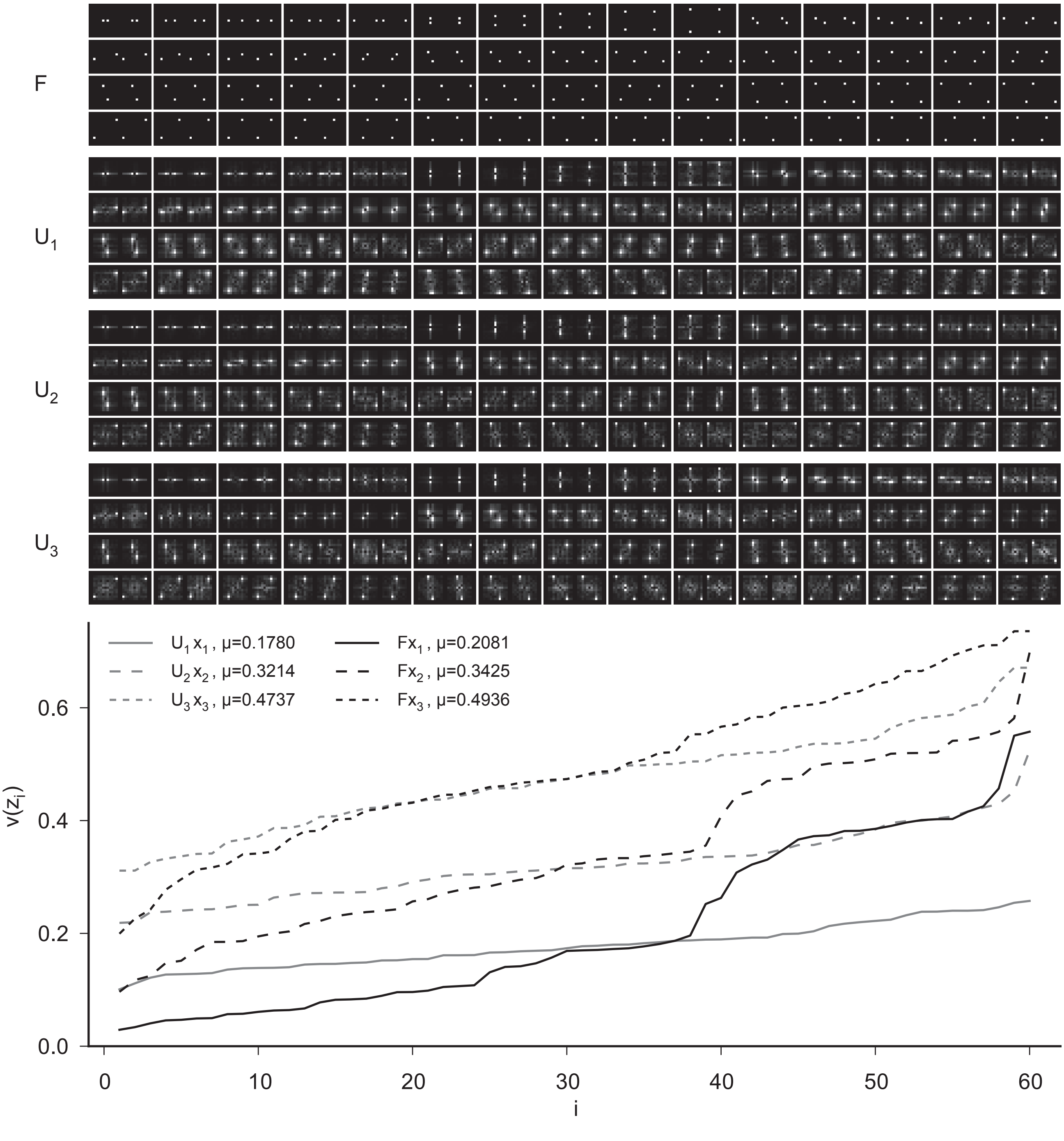
We created three different data sets, with random translations as in the first study, but the maximal shift amplitude of the 2D random walk was 1, 2, and 3 pixels, respectively. As initial condition, we used the Fourier basis (Figure 2.3, 'F') instead of a random matrix. The optimized bases are denoted as U_i where i indicates the maximal shift amplitude. We show the 2D-Fourier amplitude spectrum of the filters rather than the filters in pixel space because it is easier to access the differences between the different bases. The DC component is located at the center of the spectrum.

	Fourier basis		optimized basis	
	training	test	training	test
1 pixel shift	0.17838	0.17725	0.13801	0.15359
2 pixel shift	0.29469	0.29185	0.24680	0.27570
3 pixel shift	0.41521	0.41943	0.36569	0.40423

Table 2.1. Control for overfitting. Objective on training and test set for optimized filters and Fourier basis.

During optimization, the basis slightly departs from the initial condition but remains very localized in the Fourier domain (Figure 2.3, 'U_1'). The low frequency filters become sensitive to higher frequencies while the high frequency filters become also sensitive to lower frequencies as the initial filters blur out towards the border or center, respectively. The objective function is improved for the optimized filters not only on the training but also on the test set (cf. Table 2.1). The slowness of the 60 individual components z_i evaluated on identically created test sets (x_1, x_2, and x_3, respectively) is shown in Figure 2.3. The Fourier filters are slower than the optimized filters for the first 20-30 components, then about equal for 10 components, and significantly faster for the remaining components.

Figure 2.3 *(facing page)*. Deviations from the Fourier basis for translations with open boundary conditions. Here, we started the optimization with the Fourier basis (F) as initial condition. We used 3 different data sets sampled from the van Hateren image database using 2D translations with a shift amplitude of maximally 1, 2, or 3 pixels. The optimized filters U_n, where n is the maximal shift amplitude, do not deviate dramatically from the initial condition. The amplitude spectra of all filters are shown in the *upper panel* with the DC component being at the center. The amplitude spectra of the optimized filters blur out towards the lower frequencies except for the lowest frequencies, which blur out towards the higher frequencies. Only the highest frequencies show additional sensitivity at the lowest spatial frequencies which cannot be explained by spatial localization. The slowness of the individual components is shown in the *lower panel*. The black lines indicate the performance of the Fourier basis applied to test data with shift amplitudes of up to 1 (solid), 2 (long dashes), or 3 (short dashes) pixels. The gray lines show the performance of the optimal filters. SSA sacrifices slowness on the slower filters to gain a comparatively larger amount of slowness on the faster filters. In this way, overall SSA achieves better slowness.


F
U1
U2
U3
U1 x1 , μ=0.1780
U2 x2 , μ=0.3214
U3 x3 , μ=0.4737
Fx1 , μ=0.2081
Fx2 , μ=0.3425
Fx3 , μ=0.4936
v(zi)
0.0
0.2
0.4
0.6
0
10
20
30
40
50
60
i

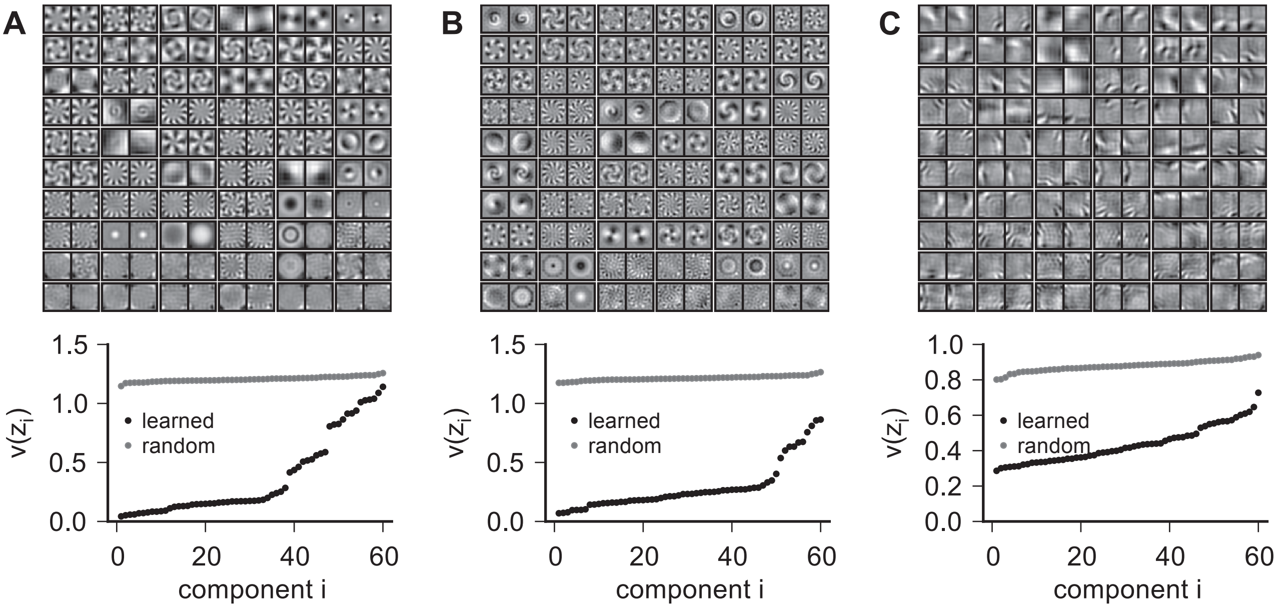


Figure 2.4. SSA filters for local rotation and scaling. Illustration of the filters obtained from patch-centered rotation sequences (A,B) and patch-centered scaling sequences (C) with the slowness of the individual filter subspaces before (*random*) and after the optimization (*learned*). The filters are ordered in ascending inverse slowness v (row-wise) with the slowest feature in the upper left and the fastest feature in the lower right corner. The data in (A) and (C) consist of 11×11 square patches from the van Hateren data set while the data for (B) consist of 121-dimensional round patches which are, for visualization, embedded in a 14x14 square patch. The rotation filters match those found in steerable filter theory (Bethge et al., 2007). The filters of the patch-centered anisotropic scaling exhibit localized edge filters centered towards the patch boundaries.

Apparently, the SSA objective sacrifices a little bit of the slowness of the low frequency components to get a comparatively larger gain in slowness from modifying the high frequency components. The optimization of average inverse slowness in contrast to searching for a single maximally slow component is a characteristic feature of SSA.

Even though we expect changes in natural movies to be dominated by local translations, it is instructive to study other global affine transforms as well. Therefore, we applied SSA to 3 additional data sets: The first data set contains 11×11 patches from the van Hateren image set which were rotated around the center pixel. The second data set consists of 14×14 patches from the van Hateren image set which were also rotated around the center pixel but where we kept only the pixels within a predefined circle. Specifically, we reduced the number of dimensions again to 121 pixels by cutting out the corners which

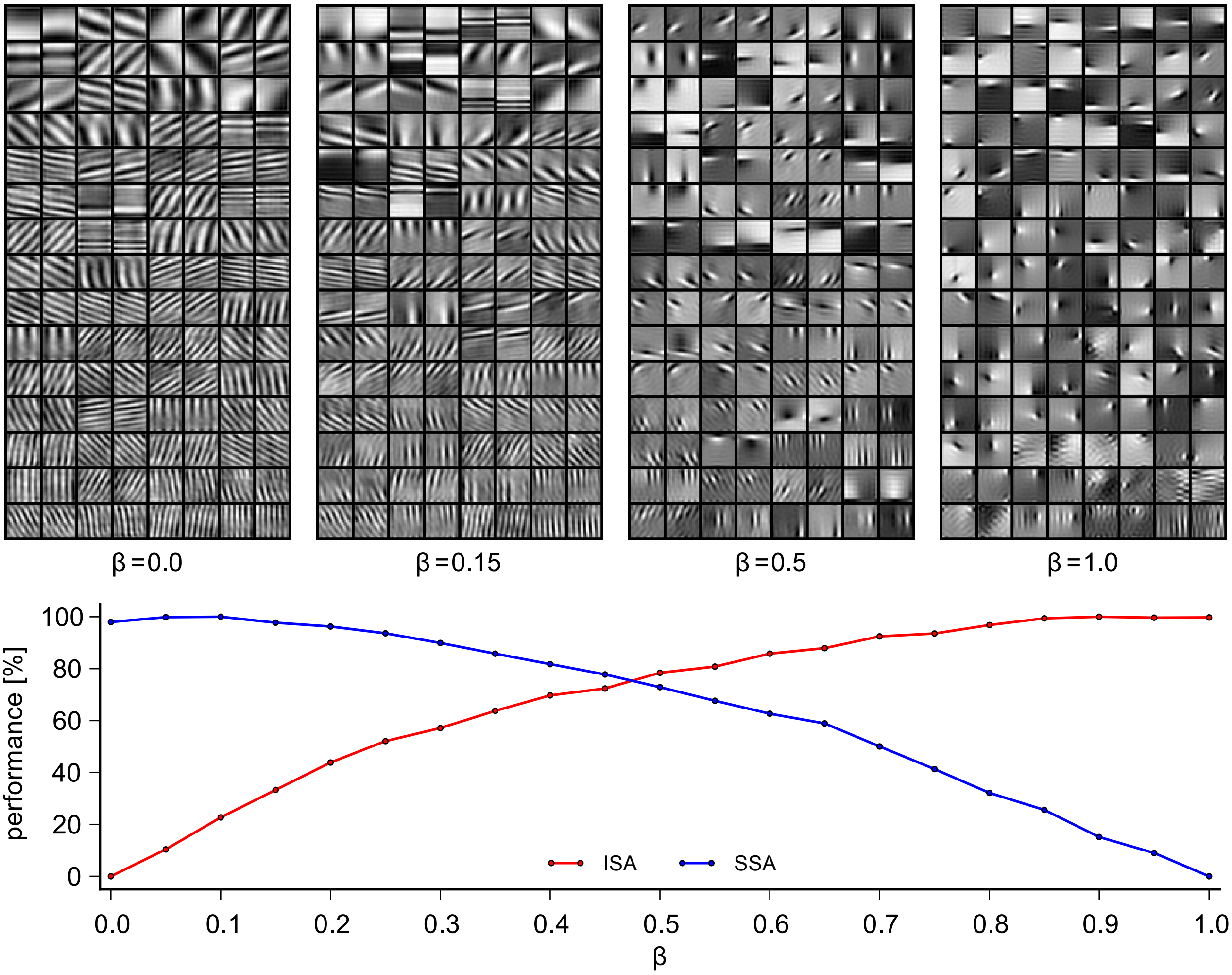
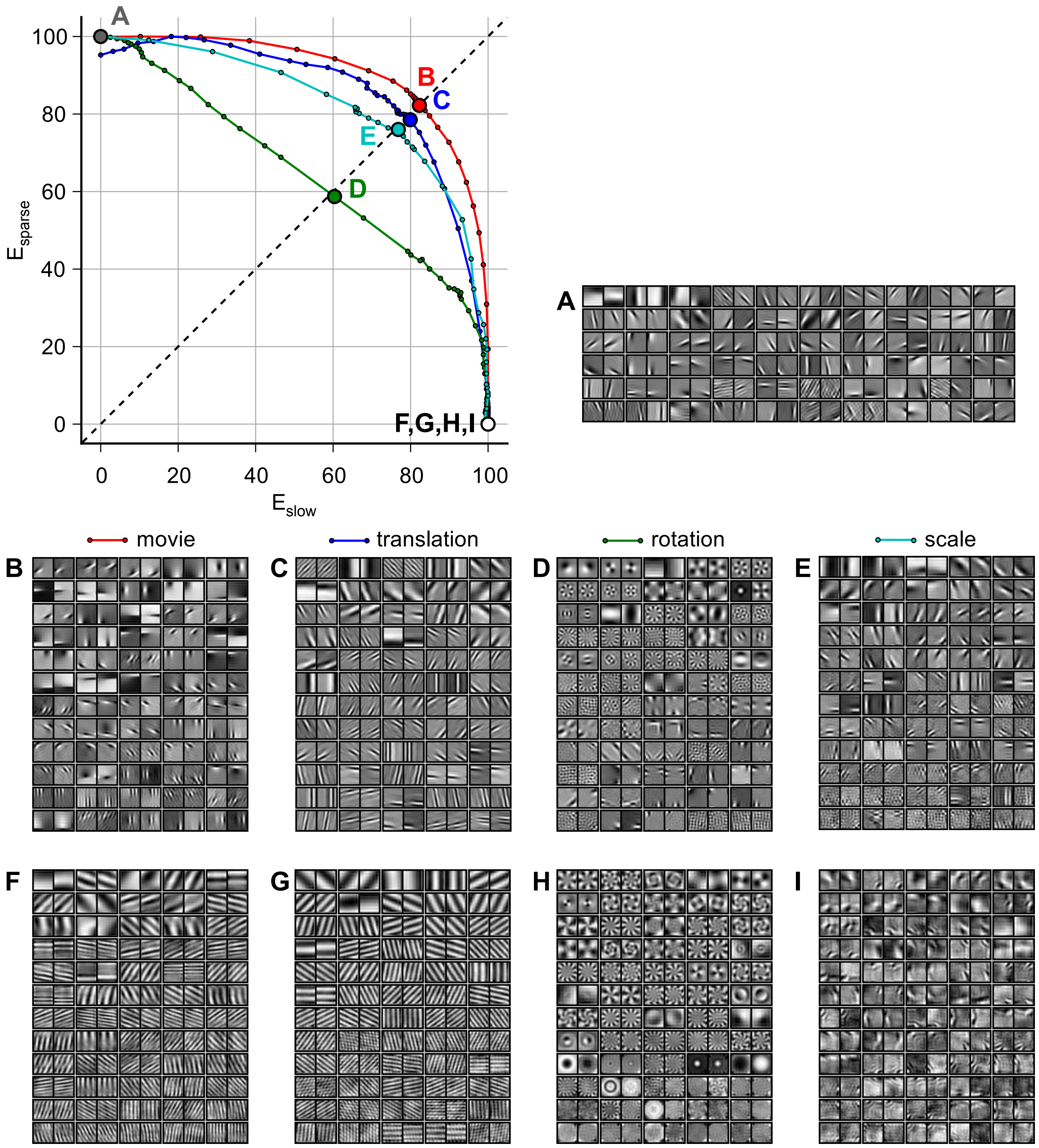


Figure 2.5. Filters of slowness, independence and mixture objective learned on movies. The lower panel shows the performance with respect to both the slowness objective E_{slow} (blue) and the sparsity objective E_{sparse} (red) and the upper panel displays four sets of filters as obtained for different values for the trade-off parameter β: The leftmost case ($\beta = 0$) is equivalent to SSA and the rightmost case ($\beta = 1$) is equivalent to ISA. There is a large difference between the two that can easily be grasped by eye. The example for $\beta = 0.5$ reflects the crossing point in performance (see lower panel) meaning that the representation performs slightly better than 80% of its maximal performance with respect to both objectives simultaneously. The case $\beta = 0.15$ was hand-picked to represent the point where the filters perceptually look similarly close to ISA and SSA.

left an 11×11 circular image patch. The patches in the third data set were sampled with sizes ranging from 9×9 to 13×13 pixels and then rescaled to 11×11 pixels, in order to obtain a patch-centered anisotropic scaling transformation. The preprocessing was identical to the previous studies and the initial filter matrix was a random orthonormal matrix. The filters and the objective of the individual subspaces of the 11×11 rotation data are shown in Figure 2.4A. The filters resemble the rotation filters found with steerable filter theory (Bethge et al., 2007). The slowness of all components is significantly larger than for random filters, but with clearly decreasing slowness for the last subspaces. Notably, the last subspaces have no systematic structure. This can be explained by the fact that when rotating a square patch, the pixels in the 4 corners are not predictable unless for multiples of $90°$ rotations. Therefore the algorithm cannot find meaningful subspaces that would preserve the energy for the pixels in the corners. The filters in Figure 2.4B from the disc shaped patches do not show these artifacts. Here, all filters nicely resemble angular wave functions as expected from steerable filter theory and also exhibit better slowness. Finally, the scaling filters are shown in Figure 2.4C. All filters resemble windowed wave functions that are localized towards the boundaries of the patch. This indicates that a scaling can be seen as a combination of local translations which go inward for downscaling and outward for upscaling. All subspaces defined by the learned filters are significantly slower than the random subspaces.

After characterizing the result of slow subspace learning for analytically defined transformations we now turn to natural movies and the comparison between slowness and sparseness. Specifically, we compare slow subspace analysis (SSA) to independent subspace analysis (ISA) in order to show how the slowness and the sparsity objective have different effects on the receptive field shapes learned. To this end, we combine the two objectives to obtain a weighted mixture of them for which we can gradually tune the trade-off between the slowness and the sparseness objective. In this way, we obtain a 1-parametric

Figure 2.6 *(facing page)*. Trade-off in the performance with respect to slowness and sparsity. When optimizing the filter set for a weighted superposition of the slowness and sparsity objectives the performance with respect to E_{sparse} decreases monotonically with E_{slow} (*upper left*). The steepness of decay indicates the impact of the trade-off. The different colors correspond to different datasets (see legend). While the performance with respect to E_{sparse} for the rotation data falls off quickly (green), the differences between scaling, translation and movie data (cyan, blue, red) are not significant. The concave shapes of the curves indicate a rather gentle trade-off. The dashed diagonal line indicates the break even point where both objectives are reduced by the same factor relative to their optimal performance. The corresponding filters are shown in the adjacent panels: The ISA filters are shown in (A) which are independent of the temporal statistics. The ISSA filters at the break even point are shown in (B) for movies, in (C) for translations, in (D) for rotations, and in (E) for scalings. The last row shows the SSA filters in the same order: (F) for movies, in (G) for translations, in (H) for rotations, and in (I) for scalings.


A
B
C
E
D
F,G,H,I
100
80
60
40
20
0
0
20
40
60
80
100
E_sparse
E_slow
movie
translation
rotation
scale
A
B
C
D
E
F
G
H
I

family of objective functions

$$E_\beta := \beta E_{sparse} + (1-\beta)E_{slow} \tag{2.1}$$

for which the parameter β determines the trade-off between slowness and sparseness. Specifically, we obtain SSA in case of $\beta = 0$ and ISA for $\beta = 1$. As one can see in Figures 2.5 the filters learned with SSA ($\beta = 0$) look very different from those learned with ISA ($\beta = 1$). This finding contradicts earlier claims that the filters learned with SSA are comparable to those learned with ISA. The most obvious difference is that the slowness objective works against the localization of filters that is brought forward by the sparsity objective.

For $0 < \beta < 1$ we will refer to the resulting algorithm as *independent slow subspace analysis* (ISSA). If a representation is optimized for E_β its performance with respect to the slowness objective E_{slow} decreases monotonically with β. At the same time, its performance with respect to E_{sparse} increases with β. Remarkably, it is possible to derive a representation which performs reasonably well with respect to both sparseness and slowness simultaneously. At the crossing point where both objectives, E_{slow} and E_{sparse}, are reduced by the same factor the performance is still larger than 80% for each. Interestingly, for this optimal trade-off the receptive fields look quite similar to those obtained with ISA. This may explain why previous work on unsupervised learning with combinations of sparseness and slowness did not reveal that the two objectives drive the receptive fields towards very different shapes.

The trade-off in performance with respect to slowness and sparsity for natural movies, translation, rotation, and scaling is summarized in Figure 2.6. It shows the ISA filters (A), the ISSA filters at the break even point of slowness and sparsity performance for natural movies (B), translation (C), rotation (D), and scaling (E) and in the same order the SSA filters in (F,G,H,I). The concave shape of the curves (upper left) indicates that the trade-off between the two objectives is rather graceful such that it is possible to achieve a reasonably good performance for both objectives at the same time.

2.5 Discussion

Unsupervised learning algorithms are a widespread approach to study candidate computational principles that may underly the formation of neural representations in sensory systems. Slowness and sparsity both have been suggested as objectives driving the formation of complex cell representations. More specifically, it was widely believed that the filter properties obtained from slow subspace analysis would resemble those obtained with independent subspace analysis. Here, we showed that there is a striking difference between the sets of SSA and ISA filters: While the sparsity objective of ISA facilitates localized filter shapes, maximal slowness can be achieved only with global receptive fields as found by SSA.

The difference between slowness and sparseness in their effect on filter shapes is most salient for the high spatial frequency filters. For low spatial frequency filters the number of cycles is small because clearly it cannot get larger than the product of spatial frequency and envelope size. Since previous studies have inspected only low spatial frequency filters the different effect of sparseness and slowness has gone unnoticed or at least not sufficiently appreciated. A signature of the drive towards global filters generated by slowness can be found in the bandwidth statistics presented in (Berkes and Wiskott, 2005). Global filter shapes correspond to small bandwidth. While the authors mention that the fraction of small bandwidth filters exceeds that found for physiological receptive fields they rather suggested that this may be an artifact of their preprocessing specifically referring to dimensionality reduction based on principal component analysis. However, the opposite is the case: the preprocessing rather leads to *underestimation* of the fraction of small bandwidth filters. Principal component analysis will always select for low spatial frequency components and thus reduce the fraction of small bandwidth filters because it is the high spatial frequency components which have the smallest bandwidth.

Also analytical considerations suggest that slowness is likely to generate global filters with small bandwidth. For small image patches it is reasonable to assume that the spatio-temporal statistics is dominated by translational motion. Thus, it is not surprising that the filter properties of SSA found for natural movies resemble those for translations. In computer vision, there is a large number of studies which derive features that are invariant under specific types of transformations such as translations, scalings and rotations. An analytical approach to invariance is provided by steerable filter theory (Knutsson and Granlund, 1983; Freeman and Adelson, 1991) which allows one to design perfectly invariant filters for any compact Lie group transformation (Hel-Or and Teo, 1998). The best known example is the power spectrum which is perfectly invariant under translations with periodic boundary conditions (Bethge et al., 2007). For the other Lie group transformations studied in this paper, the symmetry was broken due to discretization and boundary effects. In these cases the representations found with SSA can be seen as a generalization of the Fourier transform whose subspace energies are not perfectly invariant anymore but at least maximally stable under the given spatio-temporal statistics.

The receptive fields of complex cells determined from physiological experiments rarely exhibit multiple cycles as predicted by SSA. This indicates that complex cells in the brain are not fully optimized for slowness. It may still be possible though that slowness plays some role in the formation of complex cells. The trade-off analysis with the mixed objective has shown that giving up some sparsity allows one to achieve both relatively large sparsity and slowness at the same time with localized receptive fields.

The deeper principle underlying both sparsity and slowness is the idea of generative modeling (Turner and Sahani, 2007). From a generative modeling perspective, one is most concerned about modeling the precise shape of all variations in the data rather than just optimizing some fixed architecture or feature space to be as invariant or sparse as possible. More specifically, in a generative modeling framework all ingredients of the model are formalized by a density model and thus the likelihood becomes the natural objective function. This holds also true for the studies which combined the slowness objective with a sparsity objective in the past (Hyvärinen et al., 2003; Cadieu and Olshausen, 2009, 2012; Berkes et al., 2009). The generative power of these models, however, still needs to be significantly improved in order to be able to explain object recognition performance of humans and animals. A better understanding of the partially opposing demands of slowness and sparseness on the response properties of visual neurons will help us understand the computational strategy employed by the visual system in reaching that performance.

2.6 Methods

2.6.1 Slow Subspace Analysis

The algorithm of slow subspace analysis (SSA) has previously been described by Kayser et al (Kayser et al., 2001). Just like in independent subspace analysis (Hyvärinen and Hoyer, 2000) also in SSA the N-dimensional input space is separated into $M = \frac{N}{K}$ independent subspaces of dimensionality K and the (squared) norm of each subspaces should vary as slowly as possible. The output function of the i-th subspace is then defined as

$$z_i(t) = g_i(\mathbf{x}(t)) = \sum_{k=0}^{K-1} \left(\mathbf{u}_{iK+k}^{\top}\mathbf{x}(t)\right)^2, \tag{2.2}$$

where K is the dimensionality of the subspace, m the number of the subspace, and $U = [\mathbf{u}_0, \ldots, \mathbf{u}_{N-1}]$ is the orthonormal filter matrix. It is important to notice that, for an input signal $\mathbf{x}(t)$ with zero mean and unit variance, $\mathbf{z}(t)$ has mean K. For $K = 2$, the set of squared subspace norms corresponds to the power spectrum of the Fourier transform if the set of filters are the discrete Fourier transform.

The objective function of SSA has been called "temporal smoothness" objective by Kayser *et al.* (Kayser et al., 2001) and is given by

$$E_{slow}(U) = \frac{1}{M}\sum_{i=0}^{M-1} v\left(z_i\right) = \frac{1}{M}\sum_{i=0}^{M-1}\frac{\text{Var}\left[\dot{z}_i\right]}{\text{Var}\left[z_i\right]} = \frac{1}{M}\sum_{i=0}^{M-1}\frac{\left\langle \dot{z}_i^2\right\rangle_t - \left\langle \dot{z}_i\right\rangle_t^2}{\left\langle z_i^2\right\rangle_t - \left\langle z_i\right\rangle_t^2}. \qquad (2.3)$$

Note, however, that E_{slow} increases with the amount of rapid changes and is minimized subject to $UU^\top = I$. To find the optimal set of filters U under the given constraints we use a variant of the gradient projection method of Rosen (Luenberger, 1969) which was successfully used for simple cell learning before (Hurri and Hyvärinen, 2003).

In order to compute the gradient of the objective function we have to compute the temporal derivative of the output signal $\mathbf{z}(t)$ first, using the difference quotient as approximation:

$$\dot{\mathbf{z}}(t) = \frac{\mathbf{z}(t+\Delta t) - \mathbf{z}(t)}{\Delta t}. \qquad (2.4)$$

As we use discrete time steps, we can set $\Delta t = 1$ which leads to $\dot{\mathbf{z}}(t) = \mathbf{z}(t+1) - \mathbf{z}(t)$. This simplifies the objective function (2.3) as the temporal difference mean $\langle \dot{z}_i\rangle_t^2 = 0$. The objective function can be further simplified by using the fact that $\left\langle \left(\mathbf{u}^\top \mathbf{x}(t)\right)^2 \right\rangle_t = 1$ for $||\mathbf{u}||_2^2 = 1$ and $\mathbf{x}(t)$ having zero mean and unit variance, which leads to $\langle z_i\rangle_t = K$. The complete objective function is then

$$E_{slow}(U) = \frac{1}{M}\sum_{i=0}^{M-1}\frac{\left\langle \left[\sum_{k=0}^{K-1}\left(\mathbf{u}_{iK+k}^\top \mathbf{x}(t+1)\right)^2 - \sum_{k=0}^{K-1}\left(\mathbf{u}_{iK+k}^\top \mathbf{x}(t)\right)^2\right]^2 \right\rangle_t}{\left\langle \left[\sum_{k=0}^{K-1}\left(\mathbf{u}_{iK+k}^\top \mathbf{x}(t)\right)^2\right]^2 \right\rangle_t - K^2} \qquad (2.5)$$

For every iteration, the gradient of the objective function is computed, scaled by the step length α, and subtracted from the current filter set

$$\hat{U}_{i+1} = U_i - \alpha \nabla f\left(U_i\right). \qquad (2.6)$$

The partial gradient with respect to $\mathbf{u}_{iK+k}$ is

$$\frac{\partial E_{slow}(U)}{\partial \mathbf{u}_{iK+k}} = \frac{2\left\langle \left[z_i(t+1) - z_i(t)\right]\left[z_i'(t+1) - z_i'(t)\right]\right\rangle_t \left[\left\langle z_i(t)^2\right\rangle_t - K^2\right] - 2\left[z_i(t+1) - z_i(t)\right]^2 \left\langle z_i(t) z_i'(t)\right\rangle}{M\left[\left\langle z_i(t)^2\right\rangle_t - K^2\right]^2} \qquad (2.7)$$

with

$$z_i'(t) = \frac{\partial z_i(t)}{\partial \mathbf{u}_{iK+k}} = \left\langle \left[\sum_{k=0}^{K-1}\left(\mathbf{u}_{iK+k}^\top \mathbf{x}(t)\right)^2\right] \mathbf{u}_{iK+k}^\top \mathbf{x}(t)\mathbf{x}(t)^\top \right\rangle_t. \qquad (2.8)$$

The matrix containing the resulting filter set is then projected onto the orthogonal group using symmetric orthogonalization (Löwdin, 1950)

$$U_{i+1} = \hat{U}_{i+1} \left(\hat{U}_{i+1}^{\top} \hat{U}_{i+1} \right)^{-0.5}, \tag{2.9}$$

yielding the closest orthonormal matrix with respect to the Frobenius norm (Fan and Hoffman, 1955). Along this gradient a line search is performed where the initial step length α is reduced until the objective function on U_{i+1} is smaller than U_i before the iteration proceeds.

The optimization is initialized with a random orthonormal matrix U_0. As stopping criterion the optimization terminates when the change in the objective function is smaller than the threshold $\epsilon = 1e - 8$. In all our simulations we used a subspace dimension of $K = 2$. A python implementation of the algorithm can be found as part of the natter toolbox `http://bethgelab.org/software/natter/` (Sinz et al., 2013).

2.6.2 Independent Subspace Analysis

Independent subspace analysis (ISA) has originally been proposed by Hyvärinen and Hoyer (Hyvärinen and Hoyer, 2000). The only difference between SSA and ISA is the objective function. Generally speaking, ISA is characterized by a density model for which the density factorizes over a decomposition of linear subspaces. In most cases the subspaces all have the same dimension, and in case of natural images the marginal distributions over the individual subspaces are modeled as sparse spherically symmetric distributions. Like Hyvärinen and Hoyer (Hyvärinen and Hoyer, 2000) we chose the spherical exponential distribution

$$\log p\left(z_i(t)\right) = -\alpha \left[z_i(t)\right]^{0.5} + \beta \tag{2.10}$$

where z_i is the subspace response as defined in Equation 2.2, α is a scaling constant and β the normalization constant. Correspondingly, the objective function reads

$$E_{sparse}(U) = \frac{1}{M} \sum_{i=0}^{M-1} \left\langle \left[z_i(t)\right]^{0.5} \right\rangle_t = \frac{1}{M} \sum_{i=0}^{M-1} \left\langle \left[\sum_{k=0}^{K-1} \left(\mathbf{u}_{iK+k}^{\top} \mathbf{x}(t+1) \right)^2 \right]^{0.5} \right\rangle_t . \tag{2.11}$$

The scaling and normalization constants α and β can be omitted. This leads to the gradient

$$\frac{\partial E_{sparse}(U)}{\partial \mathbf{u}_{iK+k}} = 0.5 \left\langle \left[z_i(t)\right]^{-0.5} z_i'(t) \right\rangle_t \tag{2.12}$$

with $z_i'(t)$ as defined in Equation 2.8. The optimization is identical to SSA where only objective and gradient are replaced. For the numerical implementation of ISA we used a python translation of the code provided by the original authors at `http://research.ics.aalto.fi/ica/imageica/`.

2.6.3 Data Collection

The time-varying input signal $\mathbf{x}(t)$ was derived from the van Hateren image database (van Hateren and van der Schaaf, 1998) for translations, rotations and scalings and the van Hateren movie database (van Hateren and Ruderman, 1998) for movie sequences. The image database contains over 4000 calibrated monochrome images of 1536×1024 pixels, where each pixel corresponds to 0.1 deg of visual angle. We created a temporal sequence by sliding a 11×11 window over the image. Step length and direction for translation, angle for rotation and anisotropic scaling factors were sampled from a uniform random process. If not stated otherwise, the translation was sampled independently for x- and y direction from a uniform distribution on $[-2; 2]$, the rotation angle from a uniform distribution on $[-180; 180)$ and the scaling factors independently for x- and y-direction from a uniform distribution on $[0.8; 1.2]$. The movie database consists of 216 movies of 128×128 pixels with a duration of 192 s and 25 frames per second. The images were taken in Holland and show the landscape consisting mostly of bushes, trees and lakes with the occasional streets and houses. The video clips were recorded from Dutch, German and British television with mostly wildlife scenes but also sports and movies. For each stimulus set we sampled $120,000$ patches.

2.6.4 Preprocessing

The extracted 11×11 image patches were treated as vectors by stacking up the columns of the image patches, resulting in a 121-dimensional input vector $\mathbf{x}(t)$. We projected out the DC component, i.e. removed the mean from the patches, and applied symmetric whitening to the remaining 120 AC components. No low pass filtering or further dimensionality reduction was applied. All computations were done in the 120-dimensional whitened space and the optimized filters then projected back into the original pixel space.

2.7 Acknowledgments

This work was supported by the Max Planck Society and the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601) and the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307). We thank Philipp Berens and Alexander Ecker for helpful discussions and comments on the manuscript.

3 What is the Computational Goal of Complex Cell Coding in V1?

This article is joint work of Jörn-Philipp Lies, Ralf M. Häfner, and Matthias Bethge. It is a draft prepared to be submitted to PLoS Computational Biology. All simulations and computations as well as the documentation of methods and results including figures were done by JPL. The design of the experiments, the evaluation of the results and the discussion were jointly done by all 3 authors.

The article is contained as submitted with only 3 changes, namely the citation style (using author name and year instead of numbers), the figures are at the position in the text where they are referenced instead of at the end of the article, and the bibliography is at the end of the thesis. All changes are for enhanced readability only and do not alter the content.

3.1 Abstract

We seek to identify the computational goal underlying the response properties of complex cells which are ubiquitous in primary visual cortex. They have localized and orientation selective receptive fields (RF) but, in contrast to simple cells, are insensitive to phase. This property can be useful to encode information about the contrast statistics of the visual input. More specifically, the response properties of complex cells have been derived from the redundancy reduction principle using independent subspace analysis (ISA). Slow feature analysis (SFA) provides an alternative model for complex cell learning which seeks to make the neural responses as invariant to fast changes in the input as possible. Here we set out to evaluate the slowness and redundancy reduction objectives with respect to three important empirical findings about of complex cell RFs: 1) locality (i.e. finite, nonzero RF size), 2) the linear relationship between RF size and RF spatial frequency (wavelet scaling), and the aspect ratio of the RF envelope. We first use an approach similar to that employed by Field 1986 for sparse coding. Instead of single Gabor functions we use the energy model of complex cells. We evaluate the objective function of SFA and ISA on the energy model responses to motion sequences of natural images for different RF spatial fre-

quencies, RF envelope sizes and patch sizes. We find that SFA and ISA lead to completely different optima. The objective function of SFA grows without bound for increasing envelope size and is in simulations only limited by a finite patch size. Consequently, SFA learning by itself cannot explain spatially localized RFs but would need other mechanisms such as anatomical wiring constraints to limit the RF size. In contrast, the objective function of ISA yields a clear optimum for RFs of finite, non-zero envelope size, regardless of assumed patch size. However, ISA leads to physiologically unlikely large aspect ratios, such that neither objective alone can explain all compared physiological RF properties.

3.2 Introduction

Neurons in the primary visual cortex are commonly classified into two different types: simple cells and complex cells (Hubel and Wiesel, 1962). While both cell types are sensitive to orientation and spatial frequency, only complex cells are invariant to phase. Accordingly, complex cell representations are often seen as an important building block for higher-order visual tasks, such as object recognition (Riesenhuber and Poggio, 1999).

A variety of neural algorithms have been proposed that aim at explaining the response properties of complex cells as components of an invariant representation that is optimized for the spatio-temporal statistics of the visual input (Hyvärinen and Hoyer, 2000; Hyvärinen et al., 2001; Berkes and Wiskott, 2005; Cadieu and Olshausen, 2009, 2012; Karklin and Lewicki, 2009; Kayser et al., 2001; Körding et al., 2004; Hyvärinen et al., 2003). The objective functions used for the optimization can coarsely be separated into two different approaches. The first approach comes from the observation that the visual information is highly redundant and thus the visual system should remove these redundancies to obtain a robust and efficient code (Attneave, 1954; Barlow, 1961). Redundancy reduction by means of sparseness maximization (Olshausen and Field, 1996) or independent component analysis (ICA) (Bell and Sejnowski, 1997; Comon, 1994; Hyvärinen, 1997) led to Gabor-shaped filters which resemble the receptive fields of V1 simple cells. Independent subspace analysis (ISA) (Hyvärinen and Hoyer, 2000) finds complex cell properties, however, a quantitative comparison to physiological data has never been made. ISA combines the independence objective (Jutten and Herault, 1991; Comon, 1994; Hyvärinen, 1997) with the subspace pooling of the energy model of complex cells (Adelson and Bergen, 1985). The energy model is a parsimonious model for complex cells with invariance to the phase of the stimulus but sensitive to orientation and spatial frequency.

The slowness objective, as defined in slow feature analysis (SFA) (Wiskott, 2003), provides an alternative model for complex cell learning (Berkes and Wiskott, 2005) which seeks to

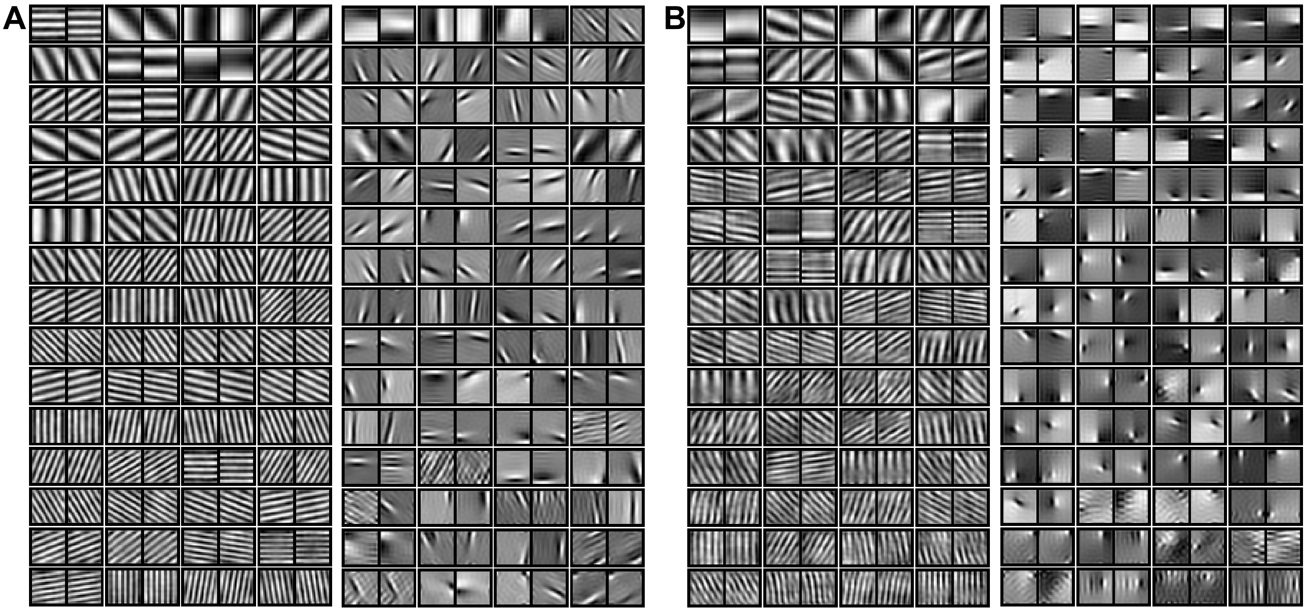


Figure 3.1. Filters learned using slowness (left) and redundancy reduction (right) objective on two different datasets. Filters in (A) are learned on the van Hateren movie dataset while (B) are image patches with 2D translations obtained from the van Hateren image database. For both data sets, the filters learned under the slowness objective (Slow Subspace Analysis, left) resemble the Fourier basis with global filters of all spatial frequencies and orientations. The filters learned under the redundancy reduction objective (Independent Subspace Analysis, right) are localized and resemble receptive fields recorded from cells in the primary visual cortex. Both filters were learned on the identical data set and identical initial random filter matrix, only with different objectives.

make the neural responses as invariant to fast changes in the input as possible. To obtain complex cell features SFA requires the expansion of the input into the quadratic feature space. Computationally, this expansion is costly. Slow Subspace Analysis (SSA) (Kayser et al., 2001; Lies et al., 2013) combines the slowness objective with the same subspace pooling as in ISA (Adelson and Bergen, 1985). Therefore, using SSA facilitates the comparison of both objectives.

The filters obtained with SSA (left) and ISA (right) are shown in Figure 3.1. The filters in (A) are learned on movie sequences sampled from the van Hateren movie dataset (van Hateren and Ruderman, 1998) and the filters in (B) are learned on patches sampled from the van Hateren image database (van Hateren and van der Schaaf, 1998) with 2D translations. In both cases, the filters of SSA and ISA are very different. SSA leads to global, Fourier-like filters while ISA produces localized, Gabor-shaped filters.

Here we set out to further examine the differences in the slowness and redundancy reduction objective with respect to three important empirical properties of complex cell RFs: 1) locality (i.e. finite, non-zero RF size), 2) the relationship between RF size and RF spatial frequency (wavelet scaling), and 3) the RF aspect ratio. We first use an approach similar to that employed by (Field and Tolhurst, 1986) for sparse coding. Instead of single Gabor functions we use the energy model of complex cells (Adelson and Bergen, 1985). We evaluate both objectives on the energy model responses to motion sequences of natural images for different spatial frequencies and envelope sizes with patch sizes ranging from 16×16 to 256×256. Patch sizes of 64×64, 96×96, and 128×128 are assumed to be the more physiologically plausible patch sizes given the receptor density in the retina (van Essen and Anderson, 1995) and the RF sizes in V1 (Gattass et al., 1981; Freeman and Simoncelli, 2011). The receptive field size increases similar to cone spacing with eccentricity, thus V1 receptive fields pool over roughly 5,000 to 15,000 cones.

3.3 Results

The general design of the simulations is shown in Figure 3.2. The model consists of two static linear filters in quadrature phase whose responses are squared and summed, which corresponds to the energy model (Adelson and Bergen, 1985). The filters are even and odd symmetric Gabor filters with vertical orientation selectivity (i.e. $0°$) and centered at the patch center. The four parameters which we varied throughout the experiments are wavelength λ, envelope width σ_x, envelope height σ_y, and image patch size. The two orthogonal filters form a 2-dimensional subspace and the output $z(t)$ can be seen as the squared norm (i.e. the radial component) of the 2D subspace. The input consists of image patches sampled from the van Hateren image database (van Hateren and van der Schaaf, 1998). As temporal transformation we simulated a random walk by randomly shifting the sampling window over the image with shift amplitudes drawn from a continuous uniform distribution on $[-2, 2]$, allowing for subpixel shifts.

We optimized the subspace responses $z(t)$ for two different objectives: slowness and redundancy reduction. The slowness objective is defined as maximizing the signal-to-noise ratio (SNR)

$$SNR\,(z(t)) = \frac{\langle z(t)\rangle^2}{\langle \dot{z}(t)\rangle^2}. \tag{3.1}$$

where the instantaneous variance $\langle z(t)\rangle^2$ of the output $z(t)$ is used to determine the signal energy and the variance $\langle \dot{z}(t)\rangle^2$ of the temporal changes defines the noise energy. With $z(t)$ being the output of the energy model, this optimization is identical to the optimiza-

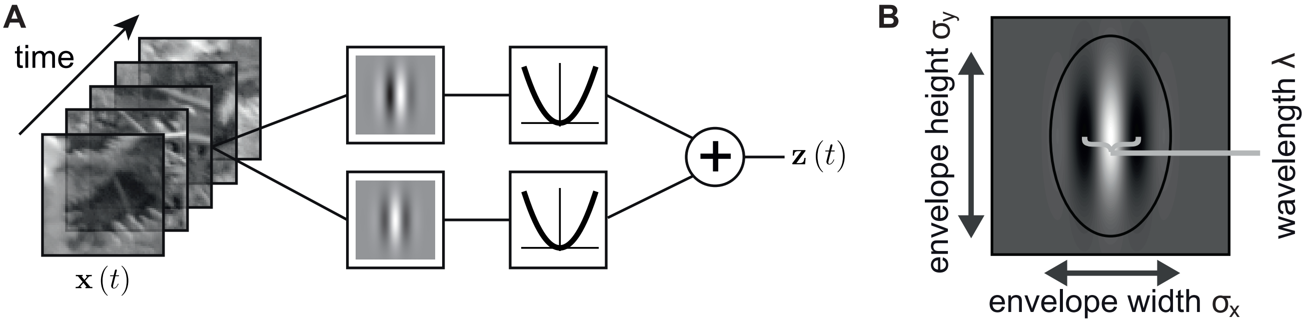


Figure 3.2. Schematic of the simulation. (A) The general design of the simulations. The time-varying input $\mathbf{x}(t)$ is passed through an even and odd symmetric Gabor filter and the filter responses are squared and summed up to form the energy model (or complex cell) response $z(t)$. (B) shows the varied parameters of the Gabor filters: wavelength λ, envelope width σ_x, envelope height σ_y, and the filter size. The position of the Gabor is fixed at the center of the patch and the preferred orientation is $0°$ for all simulations.

tion of slow subspace analysis (SSA) (Kayser et al., 2001; Lies et al., 2013). In the case of independent subspace analysis (ISA) (Hyvärinen and Hoyer, 2000) we fit an elliptically contoured gamma distribution to the response $z(t)$ of the energy model and maximize its negentropy, which is the difference of the entropy of a Gaussian distribution and the entropy of the elliptically contoured gamma distribution with identical variance. Maximizing the negentropy of a distribution is one way to maximize its sparseness. For details see Methods section.

For the first simulation we fixed the patch size to 96×96 and the aspect ratio σ_y/σ_x to 1.5, both lie within the physiologically plausible range (Ohzawa and Freeman, 1997; van Hateren and van der Schaaf, 1998; Ringach, 2002; Gattass et al., 1981; van Essen and Anderson, 1995). We computed the negentropy and the SNR for 12 different wavelengths, from 4 pixels per cycle to 48 pixels per cycle in steps of 4, and 24 envelope widths σ_x of 2 to 48 in steps of 2. The SNR in dB and negentropy in nats are shown in Figure 3.3 A and B, respectively, with the maxima marked with $\times$.

The SNR is maximized by an envelope size which increases slowly linear with wavelength. Very small and very large envelope sizes are significantly worse than envelope widths σ_x between 16 and 32 for the patch size used here. The optimal receptive field shape with wavelength 48 pixels per cycle and envelope size $\sigma_x = 26$ and $\sigma_y = 39$ is shown in Figure 3.3 C. The Gabor filter extends over the complete image patch in both vertical and horizontal direction and expresses 2 and 3 subfields in the odd and even symmetric cases,

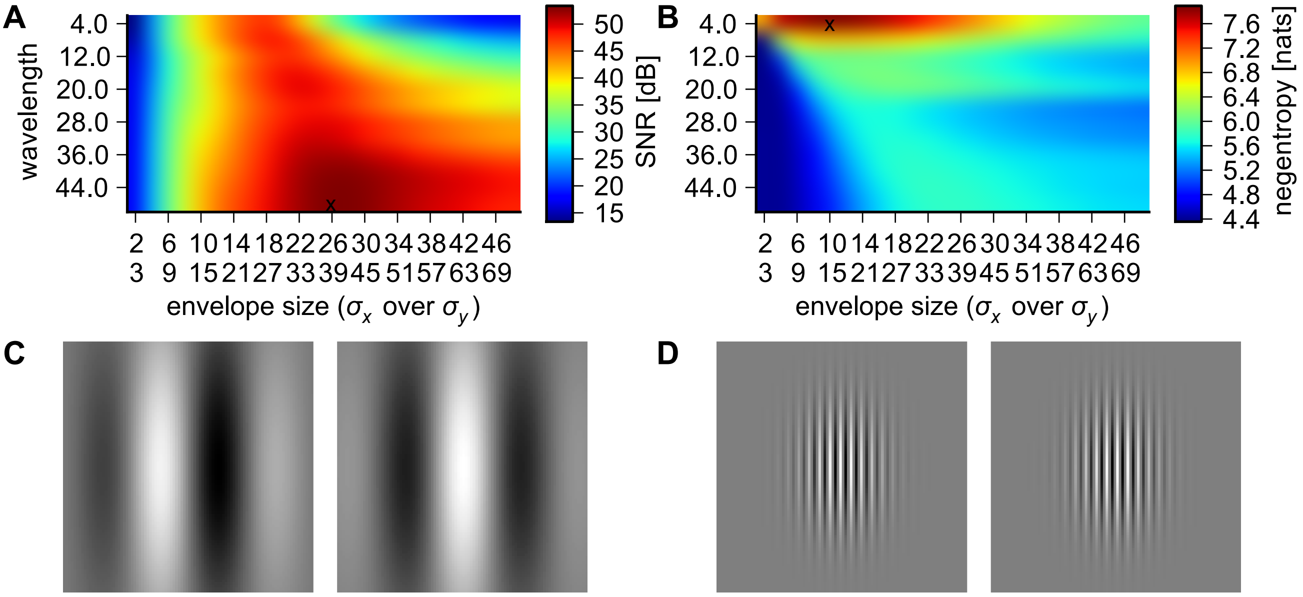


Figure 3.3. Optimal complex cell receptive fields comparison of slowness and redundancy reduction. Slowness (A) and redundancy reduction (B) of patch size 96×96 with envelope aspect ratio $\sigma_y/\sigma_x = 1.5$ computed on 1,000,000 image sequences. Slowness is quantified as SNR in dB, redundancy reduction as negentropy in nats. The energy landscape is quite different for both objectives: the slowness optimum is at high wavelengths and larger envelope size, the redundancy reduction optimum is at low wavelengths and small envelope size. The maxima are marked with $\times$. The optimal RF for slowness (C) resembles complex cell RFs found in primary visual cortex. The optimal RF for redundancy reduction (D), however, does not resemble physiological RFs.

respectively. The receptive fields resemble those found in physiological studies (Ringach, 2002; Jones and Palmer, 1987b; DeAngelis et al., 1993a,b).

The negentropy behaves quite differently. The optimal envelope size increases more rapidly with wavelength for negentropy (0.15λ for SNR, 0.3λ for negentropy) and more strikingly the optimum is at the lower end of the wavelengths and for significantly smaller receptive field size. The best receptive field (within our test set) with wavelength 4 pixels per cycle, envelope width $\sigma_x = 10$ and envelope height $\sigma_y = 15$ is shown in Figure 3.3 D. The filters extend almost to the border in vertical direction but are further away from the border in horizontal direction. The number of visible subfields is 23 and 24 for the odd and even symmetric cases, respectively. Receptive fields with such a large number of visible subfields are incompatible with current empirical findings.

In the second simulation, we investigated how the energy landscape changes if we remove the subspace with the optimal filter dimension from the data (Figure 3.4). For this we projected the data onto the (96×96 minus 2)-dimensional space orthogonal to the optimal filter pair for slowness and redundancy reduction, respectively. The SNR after projecting out the slowest filter is shown in Figure 3.4 A. All filters with similar envelope size and wavelength have reduced SNR, such that the new optimum (marked with ×) has a significantly smaller wavelength of 20 pixel per cycle (Figure 3.4 C). The receptive field still covers the complete patch but with smaller wavelength, resulting in more visible subfields. The filters no longer resemble the classical receptive field shape with low subfield number.

The negentropy of the remaining filters after projecting out the sparsest filter (Figure 3.4 B) does not change much. Therefore the next sparsest filter has the identical wavelength (4 pixels per cycle) but slightly larger envelope (Figure 3.4 D, $\sigma_x = 16$). This increases the number of visible subfields even more such that the filter does not resemble physiological receptive fields.

In a second step, we then projected out the new slowest and sparsest filters and repeated the simulation on the remaining (96×96 minus 4)-dimensional data to obtain the third optimal filter set. For the slowness objective, as before, all filters with similar wavelength and envelope size to the removed filter have lower SNRs (Figure 3.4 E). The new optimum has a wavelength of 8 pixels per cycle and an envelope size of $\sigma_x = 18$ and $\sigma_y = 27$ (Figure 3.4 G). While the envelope size stayed roughly constant for the three slowest filters, the wavelength decreased significantly such that only the slowest filter resembles receptive fields found in V1.

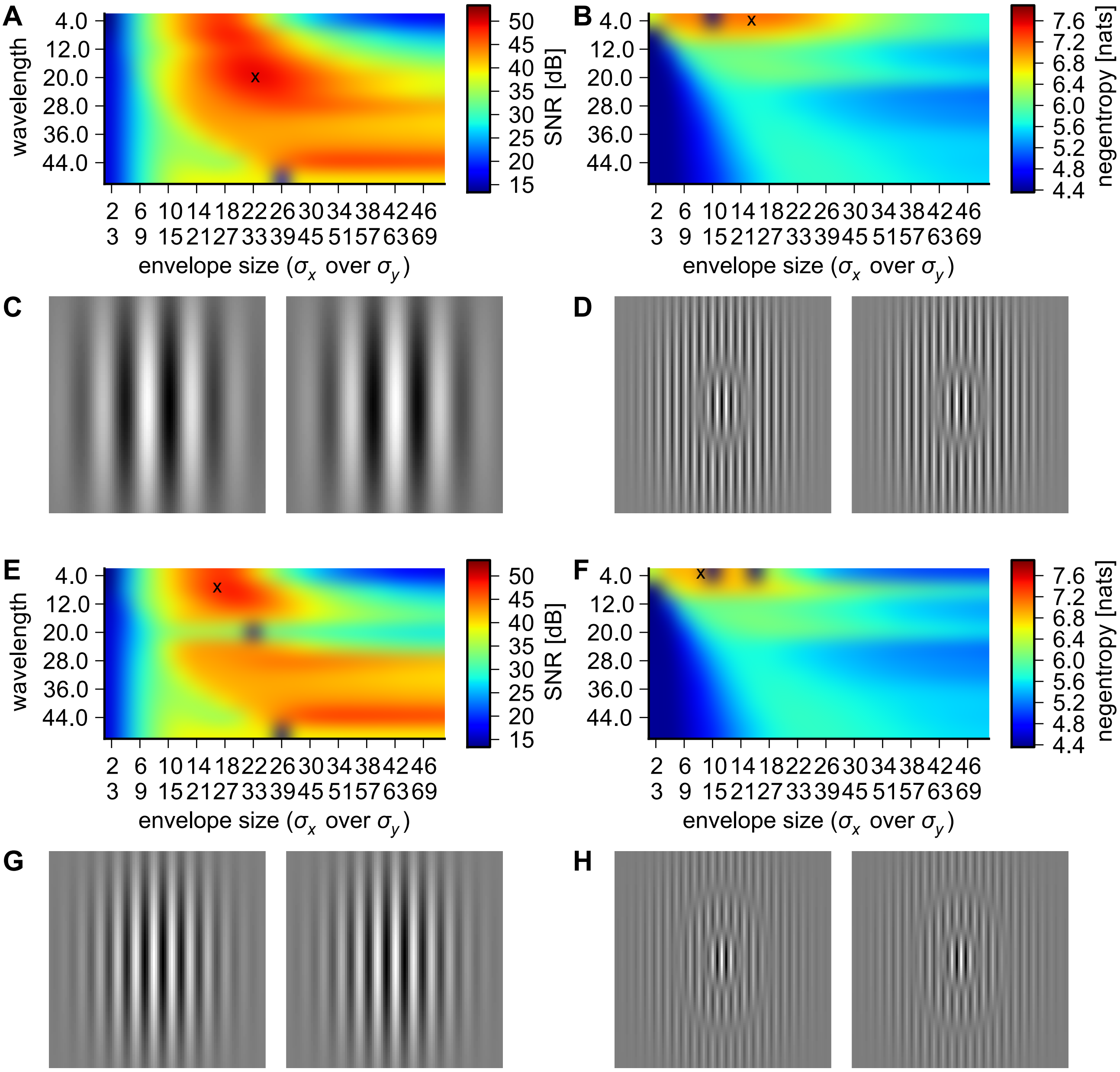
The changes in the negentropy are again rather small (Figure 3.4 F). Just like the previous two filter sets, the third sparsest filter set hast a wavelength of 4 pixels per cycle but a smaller envelope size of $\sigma_x = 8$ and $\sigma_y = 12$ (Figure 3.4 H). However, the envelope size is still too large compared to the wavelength to resemble physiologically plausible receptive fields.

The slowest filters found in the previous simulation (Figure 3.3 B) extended over the complete image patch, just as all filters in the complete SSA filter set (Figure 3.1 A & B, left). We therefore ran simulations for 3 patch sizes (64×64, 128×128, and 256×256), 3 wavelengths λ (8, 12, and 16 pixels per cycle), 6 aspect ratios (1.0, 1.5, 2.0, 3.0, 4.0, 6.0), 16 envelope widths σ_x of 2 to 32 in steps of 2 for 64×64, 32 envelope widths σ_x of 2 to 64 in steps of 2 for 128×128, and 32 envelope widths σ_x of 2 to 128 in steps of 4 for 128×128 to verify if the SNR maximum scales with patch size. The SNR as a function of envelope size is shown in Figure 3.5 A, B, and C for wavelengths 8, 12, and 16 pixels/cycle, respectively.

The optimal σ_x as well as the maximum SNR increase with patch size for the slowness objective while wavelength shows little influence on envelope size and SNR. For 64×64 (blue lines) the optimal envelope size is between 14 and 16 with a SNR of 42 dB, for 128×128 (red lines) the optimal envelope size is between 24 and 28 with a SNR of 53 dB, and for 256×256 (green lines) the optimal envelope size is between 46 and 50 with a SNR of 65 dB. The aspect ratio (different curves of same color) does not have a significant influence on the optimal σ_x as the differences in SNR at the optimum are minimal (SD < 1). The optimal aspect ratio is 1.5 for all but one configuration, where 1.0 and 1.5 have equal SNR. This shows that the slowness objective leads to physiologically plausible aspect ratios but the envelope size is only restricted by the patch size.

Receptive fields in the primary visual cortex have a similar number of subfields independent of their eccentricity while the receptive field size increases substantially with eccentricity (Ringach, 2002). From the equal increases of envelope size and wavelength follows that the bandwidth of the filter is constant. This property is known as *wavelet scaling*. For Gabor filters, the normalized bandwidth can be computed as $B_{df} = \sqrt{2\ln 2}\pi^{-1}n_x^{-1}$ where $n_x = \sigma_x/\lambda$ (Thompson and Tolhurst, 1979; Kulikowski et al., 1982). The ratio $n_x = \sigma_x/\lambda$ is widely used in physiological studies as scale-invariant receptive field property (Kulikowski and Bishop, 1981a,b; Jones and Palmer, 1987a; Ringach, 2002; Thompson and Tolhurst, 1979) which is proportional to the number of subfields for Gabor filters. In monkey and cat n_x is smaller than 0.8 (Ringach, 2002; Jones and Palmer, 1987a). Figure 3.5 showed that the optimal envelope size under the slowness objective is defined by the patch size and does not scale with wavelength. This implies that B_{df} scales with inverse patch size and λ but is independent of aspect ratio. We computed the optimal bandwidth for different wavelengths, patch sizes and aspect ratios to verify this (Figure 3.6 A). Here, constant bandwidth for all wavelengths represents a filter with perfect wavelet scaling properties. The optimal filter under the slowness objective scales in fact perfectly with λ and patch

Figure 3.4 ***(facing page).*** **Slowness and negentropy after removing the best complex cell subspaces.** The setup is identical to Figure 3.3. (A) and (B) are the slowness and redundancy reduction objective, respectively, after projecting out the corresponding optimal complex cell subspace defined by 3.3 C and D. The new optimal complex cell RF for slowness (C) has a lower wavelength (48 to 22 pixels/cycle) but similarly large envelope size (26 to 22) while the new optimal RF for redundancy reduction (D) has the identical wavelength (4 pixels/cycle) but larger envelope size (10 to 16). The feature space was then reduced by the subspaces spanned by C and D, respectively, and the SNR and negentropy were computed again. For the slowness objective (E) the optimum has an even lower wavelength now (22 to 8 cycles/pixel) while the envelope size decreased only slightly (22 to 18). The redundancy reduction objective (F) has the maximum at the same wavelength as the previous two RFs (4 pixels/cycle) but a smaller envelope size (16 to 8). The maximally slow RF (G) and maximally sparse RF (H) both do not resemble RFs found in the primary visual cortex.


A
wavelength
4.0
12.0
20.0
28.0
36.0
44.0
2 6 10 14 18 22 26 30 34 38 42 46
3 9 15 21 27 33 39 45 51 57 63 69
envelope size (σ_x over σ_y)
x
50
45
40
35
30
25
20
15
SNR [dB]
B
4.0
12.0
20.0
28.0
36.0
44.0
2 6 10 14 18 22 26 30 34 38 42 46
3 9 15 21 27 33 39 45 51 57 63 69
envelope size (σ_x over σ_y)
x
7.6
7.2
6.8
6.4
6.0
5.6
5.2
4.8
4.4
negentropy [nats]
C
D
E
wavelength
4.0
12.0
20.0
28.0
36.0
44.0
2 6 10 14 18 22 26 30 34 38 42 46
3 9 15 21 27 33 39 45 51 57 63 69
envelope size (σ_x over σ_y)
x
50
45
40
35
30
25
20
15
SNR [dB]
F
4.0
12.0
20.0
28.0
36.0
44.0
2 6 10 14 18 22 26 30 34 38 42 46
3 9 15 21 27 33 39 45 51 57 63 69
envelope size (σ_x over σ_y)
x
7.6
7.2
6.8
6.4
6.0
5.6
5.2
4.8
4.4
negentropy [nats]
G
H

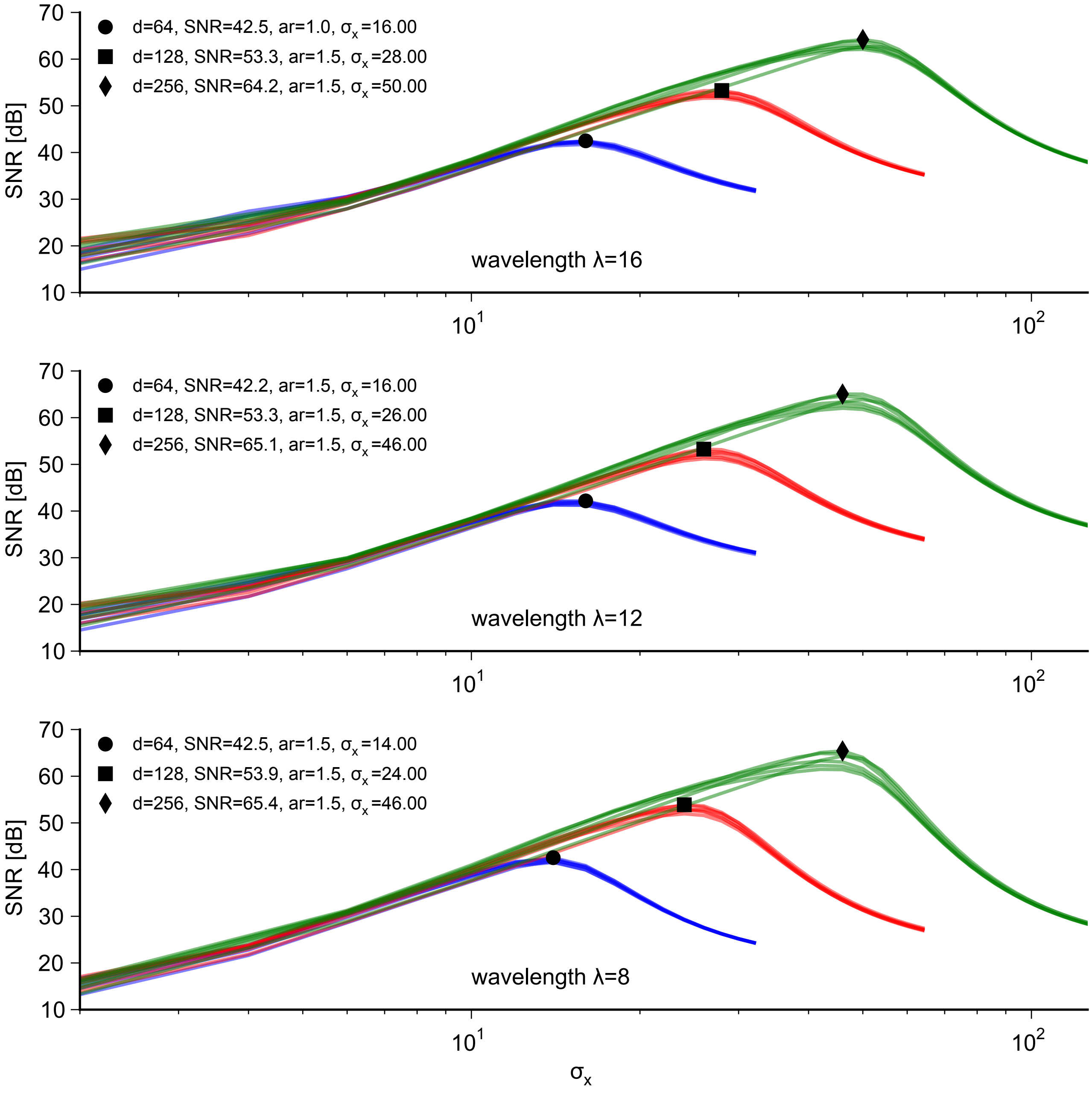
size and is only for large wavelengths (relative to patch size) within the physiologically plausible range of $B_{df} > 0.44$ (gray area) (Ringach, 2002; De Valois et al., 1982). Thus slowness does not exhibit any wavelet scaling property. The aspect ratio has no influence, as the solid (aspect ratio of 1.5) and dashed (aspect ratio of 6) lines coincide. For redundancy reduction (Figure 3.6 B) the aspect ratio has significant influence on the optimal bandwidth. For an aspect ratio of 1.5 (solid line) and small wavelengths, the bandwidth is not within the physiological range and scales with λ. However, for an aspect ratio of 6 the bandwidth does not scale significantly with λ and patch size. Since the bandwidth is approximately constant the redundancy reduction filter with an aspect ratio of 6 exhibits wavelet scaling properties.

To further investigate how well the two objectives can explain physiological data we compared the optimal filters of slowness and redundancy reduction to the receptive field properties from (Ringach, 2002) and Gabor fits to 16×16 ISA filter (Figure 3.7). The parameters are similar to the previous simulations: 8 different patch sizes (16×16, 24×24, 32×32, 64×64, 96×96, 128×128, 196×196, and 256×256), 7 aspect ratios (0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 6.0), 3 wavelengths λ (8, 12, and 16 pixels per cycle), and envelope width σ_x from 2 to half patch size in steps of 2. Figure 3.7 shows ratio $n_x = \sigma_x/\lambda$ at the optimum as a function of the aspect ratio (data of 256×256 with $n_x > 4$ for all configurations not shown for clarity of the figure).

The aspect ratio of the Ringach data (cyan circles) is clustered between 0.5 and 2 and n_x between 0 and 1. The optimal n_x for slowness (gray and green, green for physiologically plausible patch size, shaded area variation over wavelengths) is invariant under the aspect ratio, as to be expected from the previous simulation. However, only for patch sizes of 16×16 and 24×24 the optimal n_x lies within the physiological data. The optimal n_x for larger patch sizes are significantly larger than those found in monkey.

For the redundancy reduction objective (red line, shaded area variation over wavelengths and patch sizes) the aspect ratio has strong influence on the optimal n_x while patch size

Figure 3.5 ***(facing page).*** **Slowness increases with patch size.** The three plots show the slowness as SNR in dB for wavelengths 8 (top), 12 (middle), and 16 (bottom) cycles per patch, patch sizes 64×64 (blue), 128×128 (red), and 256×256 (green), and aspect ratios 1.0, 1.5, 2.0, 3.0, 4.0, and 6.0 as set of curves with identical color. The SNR scales with patch size from 42 dB for 64×64 to 65 dB for 256×256. The variation with wavelength or aspect ratio is around 1 dB. While the variation with aspect ratio is rather small, the optimal aspect ratio is always between 1.0 and 1.5. The optimal envelope width scales significantly with patch size and to a lesser extend with wavelength. The circle, square, and diamond mark the maximum of blue, red, and green set of curves, respectively, with the exact values for SNR, aspect ratio and envelope width given in the legend.


● d=64, SNR=42.5, ar=1.0, σ_x=16.00
■ d=128, SNR=53.3, ar=1.5, σ_x=28.00
◆ d=256, SNR=64.2, ar=1.5, σ_x=50.00
wavelength λ=16
● d=64, SNR=42.2, ar=1.5, σ_x=16.00
■ d=128, SNR=53.3, ar=1.5, σ_x=26.00
◆ d=256, SNR=65.1, ar=1.5, σ_x=46.00
wavelength λ=12
● d=64, SNR=42.5, ar=1.5, σ_x=14.00
■ d=128, SNR=53.9, ar=1.5, σ_x=24.00
◆ d=256, SNR=65.4, ar=1.5, σ_x=46.00
wavelength λ=8
SNR [dB]
10
20
30
40
50
60
70
10^1
10^2
σ_x

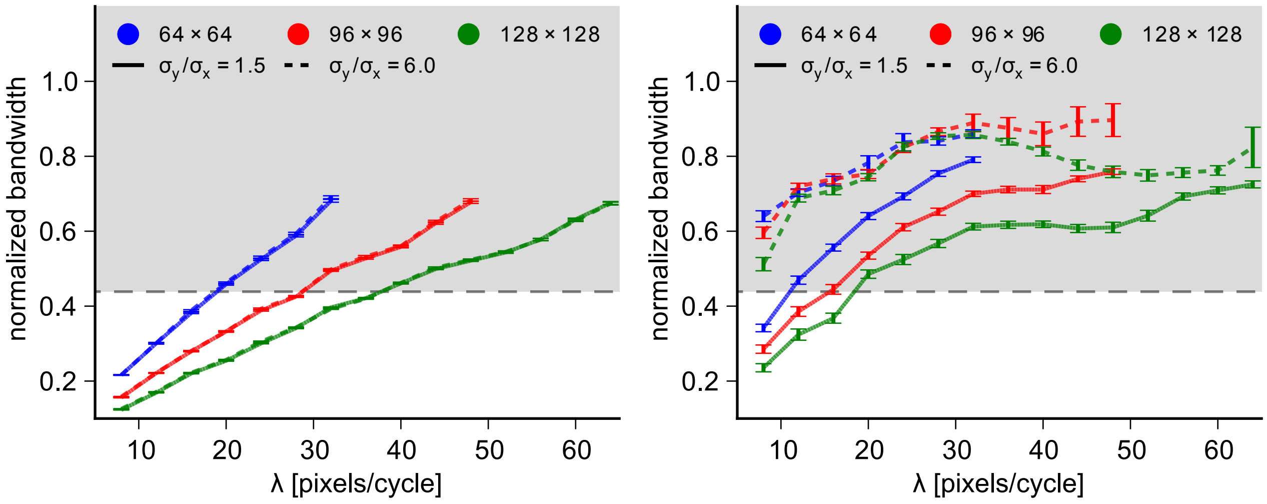


Figure 3.6. Wavelet scaling of slowness and redundancy reduction. Optimal normalized bandwidth B_{df} as function of λ for patch sizes 64×64 (blue), 96×96 (red), and 128×128 (green) and aspect ratios σ_y/σ_x 1.5 (solid) and 6.0 (dashed). The slowness optimal RFs (A) scale with λ and inverse patch size. The aspect ratio has no influence on the optimum, as the dashed and solid lines coincide. Only for large λ (relative to patch size) the bandwidth of the optimal filter is within the physiologically found range (gray area). For redundancy reduction (B) the optimal bandwidth scales with λ and inverse patch size for small aspect ratios and small λ. For large aspect ratios patch size and wavelength have no strong influence, the filter bandwidth is within the physiologically plausible range and the bandwidth of the optimal filters is constant, i.e. the filters exhibit wavelet scaling properties.

and wavelength have no significant influence. The optimal n_x is around 2.0 for an aspect ratio of 0.5 (i.e. double as wide as long) and it approaches 0.5 for aspect ratios larger than 3. This means that ISA leads to filters which have either physiologically plausible envelope sizes or aspect ratios but not both. This is supported by the parameters of the Gabors fitted to 16 × 16 ISA filters trained on van Hateren data (magenta diamonds). The Gabor fits cluster densely around the redundancy reduction optimum with minimal overlap with the physiological data. Fits to a complete SSA filter basis are not provided for the following reason: SSA filters converge to the Fourier basis as shown in (Lies et al., 2013). The best fitting Gabor filter to a Fourier filter has the identical orientation, phase, and spatial frequency with an infinitely large envelope. Thus numerical fitting of SSA filters leads to unreliable results in envelope size and was not included in the evaluation. These results show that neither slowness nor redundancy reduction can explain all tested properties of physiologically found receptive fields.

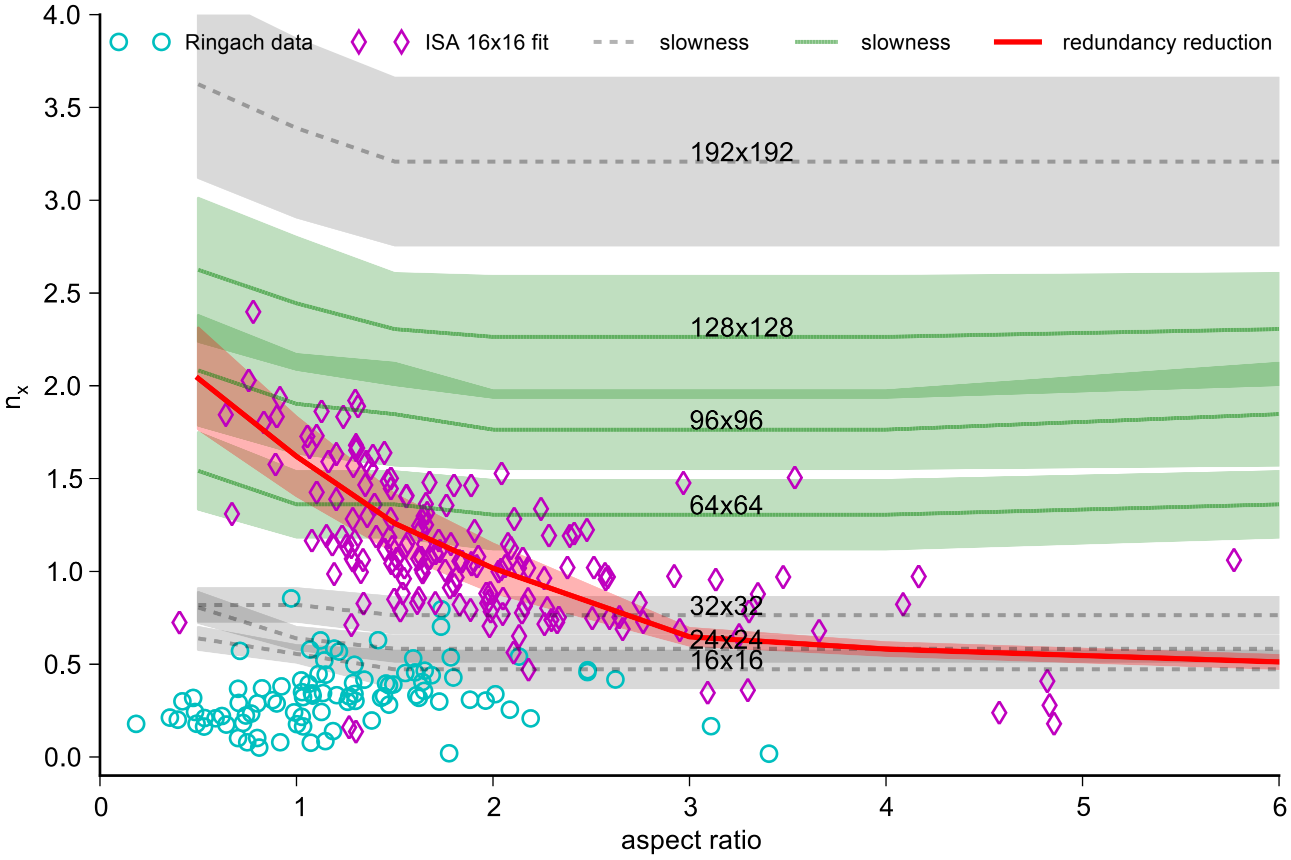


Figure 3.7. Optimal receptive field size and aspect ratio of slowness, redundancy reduction and physiological data. The cyan circles are receptive field data from monkey V1 (Ringach, 2002). The gray and green shaded areas are the ranges of slowness optimal RFs of patch size 16×16 to 192×192. The green areas represent patch sizes with a comparable number of pixels as cones in a V1 receptive field, gray areas have significantly more or less pixels. The red area are the redundancy reduction optimal RFs for all patch sizes. The magenta diamonds are fits to the filters obtained from ISA on 16×16 patches. The optimum under the slowness objective does not vary with aspect ratio. Only for very small patch sizes slowness is within the physiological range. The redundancy reduction objective leads to RFs which either have a physiologically plausible aspect ratio but too many subfields or the correct number of subfields but with a physiologically implausible aspect ratio. The Gabor fits to the ISA filter supports this finding as they cluster around the redundancy reduction curve. A fit to SSA filters is not shown as SSA filters are close to the Fourier basis and a Gabor fit to a Fourier filter has aspect ratio 1.0 but infinitely envelope size.

3.4 Discussion

The central question we wanted to answer here is: given the energy model with Gabor filters, can the slowness and redundancy reduction principles account for complex cell receptive field properties found in physiological studies? Both principles have been used to obtain complex cell properties in the past, however, both approaches used different feature spaces and data sets and were never directly compared. We brought both principles together with identical feature spaces and data sets such that the only difference was the objective function itself.

Here, we evaluated the slowness and redundancy reduction objectives by their ability to explain three important empirical findings of complex cell receptive fields: locality, wavelet scaling, and aspect ratio. We found that neither of the two objectives could explain all three criteria. The redundancy reduction objective led to receptive fields with wavelet scaling properties which were localized orthogonal to the preferred orientation (i.e. along σ_x) but required implausibly large aspect ratios. The slowness-optimized receptive fields had an aspect ratio well within the physiologically plausible range but were only limited in extend by the patch size hence did not show wavelet scaling properties. Moreover, we found that the optima are in principle different. While slowness prefers large, square, full-field waves, redundancy reduction favors localized, elongated, high-frequency filters.

Previous studies on SFA used image sequences with patch sizes of 16×16 patches and low pass filtering of two patches to the first 100 principal components (Berkes and Wiskott, 2005), effectively reducing the patch size to 50 dimensions per patch. In natural images with their characteristic $1/f^2$ power spectrum, low frequencies convey more information than high frequencies, thus PCA retains mostly low frequency components. Our results show that for these parameter, finding physiologically plausible receptive field properties is not surprising. The findings suggest that slowness prefers Gabor filters with large envelopes and of varying spatial frequency with preference to low frequencies. This limits the range of possible ratios n_x thus may lead to physiologically more plausible parameter.

An important note is that patch size and discretization is not a true physiological parameter. It is a necessary artifact of the numerical simulations, however, one can estimate the patch size, or resolution, where a single pixel becomes undetectable. Every V1 complex cell pools over approximately 5,000-15,000 receptors in the retina based on the findings of V1 receptive field size (Gattass et al., 1981; Freeman and Simoncelli, 2011) and receptor density in the retina (van Essen and Anderson, 1995). This sets an upper limit on the spatial frequency theoretically available to V1. However, all physiological studies found receptive fields with rather low spatial frequencies far from the upper limit. This could be an artifact of the mapping procedure where cells with high frequency receptive fields would be more sensitive to eye movements and more likely discarded due to their instability.

The redundancy reduction objective prefers filters with high spatial frequency, large aspect ratio and a limited range of spatial bandwidth. The high frequency selectivity of redundancy reduction implemented in ISA is well known and was already stated by the original authors (Hyvärinen and Hoyer, 2000). Also the clear localization has been present in previous studies, but the tendency to create rather large aspect ratios has not been reported. This behavior is plausible within our findings, as there is a trade-off between aspect ratio and envelope size, as depicted in Figure 3.7, which leads to physiologically plausible aspect ratios but implausibly large envelope sizes.

One question that remains open is what the effect of a combined objective to the receptive field would be. A combination of slowness and redundancy reduction has been used in the bubbles framework (Hyvärinen et al., 2003), topographical ICA (Hyvärinen et al., 2001), the structured model from video (Berkes et al., 2009), the multi-layer invariance model of (Cadieu and Olshausen, 2009, 2012), and a simple linear combination of ISA and SSA (Lies et al., 2013). All models were able to learn filters similar to those obtained with ISA, i.e. localized and elongated. This indicates that the redundancy reduction objective dominates the slowness objective in the combined objective thus leading to filters closer to the pure redundancy reduction optimum. Given the energy landscapes of both objectives, a simple linear combination should not lead to physiologically plausible receptive fields but rather a bimodal energy landscape with the maximum depending on the mixing factor and possibly a very narrow range of intermediate optima when shifting from the slowness optimum to the redundancy reduction optimum. However, both objectives are not directly numerically comparable. Negentropy and SNR are two completely different measures with different dynamical ranges.

We found that neither objective alone can explain the response properties of V1 complex cells. Both objectives can explain a subset of the features but would require additional constraints, such as wiring length constraints, to produce physiologically plausible receptive fields.

3.5 Methods

The methods section is separated in three parts. The first part briefly describes the energy model (Adelson and Bergen, 1985) on an abstract level and presents the framework of our simulations. The second part describes the two different objective functions we apply to the output of the energy model. The last part describes the data sets we used throughout this paper. A python implementation is available on our website at `http://www.bethgelab.org/code/LiesEtAl2013b`.

3.5.1 Energy model

The model underlying all experiments in this paper is the energy model (Adelson and Bergen, 1985) known from steerable filter theory (Gattass et al., 1981; Freeman and Adelson, 1991). The energy model was successfully used in complex cell modeling before (Hyvärinen and Hoyer, 2000; Körding et al., 2004; Lies et al., 2013). The structure of the energy model is depicted in Figure 3.2 A. It consists of two static linear filters $\mathbf{u}_e$ and $\mathbf{u}_o$ whose responses are squared and summed to form the energy model response

$$z(\mathbf{x}(t)) = \left(\mathbf{u}_e^\top \mathbf{x}(t)\right)^2 + \left(\mathbf{u}_o^\top \mathbf{x}(t)\right)^2 . \tag{3.2}$$

The two filters form a quadrature pair, i.e. are orthonormal and have a phase difference of $\pi/2$. They span a 2-dimensional subspace and the energy model response is the squared norm (i.e. radial component) of the response of this subspace. The energy model is a parsimonious model for complex cells in the primary visual cortex, as it provides invariance to stimulus phase and, to some extend, spatial position. As linear filters we use even and odd symmetric Gabor filters centered at the center of the patch. The preferred orientation is always 0°, i.e. vertical. The remaining parameters of the Gabor filters are wavelength λ, envelope height σ_y, envelope width σ_x, and the size of the filter in general. An illustration of the parameters is shown in Figure 3.2 B. To quantify the filter properties we compute the ratio $n_x = \sigma_x/\lambda$ where σ_x is the envelope width orthogonal to the preferred orientation and λ is the preferred wavelength of the Gabor filter. n_x is proportional to the number of antagonistic subregions within the Gaussian envelope, which is widely used as measure in physiological studies (Kulikowski and Bishop, 1981a,b; Jones and Palmer, 1987a; Ringach, 2002; Thompson and Tolhurst, 1979). The normalized bandwidth (Thompson and Tolhurst, 1979; Kulikowski et al., 1982) of the Gabor filter can be derived directly from n_x through

$$B_{df} = \frac{f_u - f_l}{f_0} = \frac{\sqrt{2\ln 2}}{\pi n_x} \tag{3.3}$$

where B_{df} is the normalized bandwidth, f_0 is the frequency with maximum response, and f_u, f_l are the frequencies above and below f_0, respectively, where the response has dropped to 50% compared to the maximum. The bandwidth converges to 0 or n_x increases without bounds, respectively, if the Gaussian envelope of the Gabor filter increases for constant wavelength.

The bandwidth in octaves for a Gabor can be computed from the normalized bandwidth

$$B_{oct} = \log_2 \left[\frac{1 + \frac{B_{df}}{2}}{1 - \frac{B_{df}}{2}} \right] \tag{3.4}$$

under the constraint that $B_{df} < 2$ or $n_x > \sqrt{0.5 \ln 2}\pi$, respectively, as the bandwidth in octaves converges to infinity and is not defined for $B_{df} > 2$. As the data from (Ringach, 2002) do not fulfill the constraint we used the normalized bandwidth.

For each configuration of λ, σ_x, σ_y, and patch size we computed the model response to 1,000,000 image sequences. These responses were then evaluated according to the objectives presented in the following section.

To find the second best filter we project the data onto the n-2-dimensional subspace orthogonal to the two filters $\mathbf{u}_e$ and $\mathbf{u}_o$. The input $\mathbf{x}(t)$ is reduced to $\hat{\mathbf{x}}(t) = W\mathbf{x}(t)$ where W is the $n-2$-dimensional subspace. This procedure is repeated with first and second best filter pair to obtain the third best filters on the remaining n-4-dimensional subspace.

3.5.2 Objectives

Slowness objective

The general idea behind the slowness objective (Hinton, 1989) is the observation that objects, on a coarse scale, vary only slowly over time while the light intensity of a single pixel changes rapidly. The responses of cells should therefore be invariant to the fast pixel variations and represent the slowly moving object. The best-known implementation of the slowness objective is Slow Feature Analysis (SFA) (Wiskott and Sejnowski, 2002). In SFA, the objective is to find for a multi-dimensional time-varying input signal $\mathbf{x}(t) = (x_0(t), \ldots, x_{N-1}(t))^\top,\ t \in [0, T-1]$ a real-valued (nonlinear) output function $\mathbf{z}(t) = \mathbf{g}(\mathbf{x}(t))$, $\mathbf{g} : \mathbb{R}^N \mapsto \mathbb{R}^M$ which minimizes the temporal variance

$$\Delta(z_i) := \left\langle \dot{z}_i^2 \right\rangle_t \tag{3.5}$$

under the constraints that the output signal has zero mean, unit variance, and the output dimensions are uncorrelated

$$\langle z_i \rangle_t = 0 \tag{3.6}$$

$$\left\langle z_i^2 \right\rangle_t = 1 \tag{3.7}$$

$$\langle z_i z_j \rangle_t = 0 \quad \forall i \neq j. \tag{3.8}$$

Here, $\langle\cdot\rangle_t$ is the temporal average and $\dot{z}_i$ the temporal derivative. We dropped the explicit notation of the temporal dependence of z_i to make the equations more readable. Constraint 3.6 is merely for convenience to keep the following constraints simpler. Constraint 3.7 ensures that the trivial solution $\mathbf{g}(\mathbf{x}(t)) = \mathbf{0}$ is not feasible and constraint 3.8 enforces that the output encodes different information in each dimension. Note, however, that $\mathbf{g}(\mathbf{x})$ could be a mapping that entirely ignores $x_2, \ldots, x_n$ by using a set of nonlinear functions $\{g_k(x_1)\}_k$ that all only depend on x_1 and that Constraint 3.7 also precludes the possibility to model perfect invariances as otherwise possible in the case of compact Lie groups. Constraints 3.7 and 3.8 can be combined into

$$\left\langle \mathbf{z}\mathbf{z}^\top \right\rangle_t = \mathbb{1} \tag{3.9}$$

where $\mathbb{1}$ is the identity matrix of matching dimensionality.

From a machine learning point of view, SFA is equivalent to oriented PCA where the static variance (Eq. 3.9) is interpreted as signal energy, the temporal variance (Eq. 3.5) as noise energy, and the goal is to maximize the signal-to-noise ratio (SNR)

$$SNR\,(z_i) = \frac{\mathrm{Var}\,[z_i]}{\mathrm{Var}\,[\dot{z}_i]}. \tag{3.10}$$

If the temporal variance $\mathrm{Var}\,[\dot{z}_i]$ converges to 0, i.e. the signal is perfectly invariant, the SNR converges to infinity. In this paper, the output function $\mathbf{g}(\cdot)$ is the energy model $z(\mathbf{x}(t)) = \left(\mathbf{u}_e^\top \mathbf{x}(t)\right)^2 + \left(\mathbf{u}_o^\top \mathbf{x}(t)\right)^2$ defined in Eq. 3.2 in the previous section. Maximizing the SNR of energy model outputs is known as Slow Subspace Analysis (SSA) (Kayser et al., 2001; Lies et al., 2013). Here, we use only one complex cell, thus $\mathbf{z}(t) = z_0(t)$ is a one-dimensional output signal.

Reduncancy reduction objective

The idea that the visual system might apply a redundancy reduction objective roots in the observation that the visual pathway has only limited capacity and cells therefore want to remove redundant information in their responses (Attneave, 1954; Barlow, 1961; Atick and Redlich, 1990).

One mean of achieving redundancy reduction is to make the model responses as statistically independent as possible, for example using independent component analysis (ICA) (Jutten and Herault, 1991; Comon, 1994; Bell and Sejnowski, 1995). As independence maximization is not defined for a one-dimensional output, we optimize $z(t)$ for sparseness (Olshausen and Field, 1996). For a detailed explanation how independence and sparseness are linked see e.g. (Hyvärinen, 2010). Note that, even though we use a time-varying input $\mathbf{x}(t)$, the redundancy reduction objective used here does not require temporal correlations between following time steps. In fact, any permutation of the time series would lead to the same results as the objective does not contain a temporal component.

A widely used measure of sparseness of a distribution is the negentropy (Comon, 1994; Hyvärinen, 1997). The negentropy of a distribution is defined as the difference of its entropy to the entropy of a Gaussian distribution with the same covariance Σ, i.e. it quantifies the non-Gaussianity of the distribution. To compute the negentropy we need to fit a distribution to the filter outputs of our simulation data. We chose the elliptically contoured Gamma distribution as it matches the fact that the energy model response is the squared radial component of the 2-dimensional linear subspace spanned by its filters. We estimated the shape parameter a and scale parameter b using standard maximum likelihood estimation implemented in SciPy.

For the elliptically contoured gamma distribution with shape parameter a, scale parameter b, and of dimension q, the entropy can be computed analytically (Hosseini and Bethge, 2013):

$$H\left(\mathbf{X}_{ECG}\right) = +\frac{1}{2}\log_2\left(|\Sigma|\right) - \log_2\left(\frac{\Gamma\left(q/2\right)}{\pi^{\frac{q}{2}}\Gamma\left(a\right)b^a}\right) - \left(a - q/2\right)\big(\Psi\left(a\right) + \log_2\left(b\right)\big) + a. \tag{3.11}$$

where Ψ is the digamma function. Given the entropy of a Gaussian distribution $H\left(\mathbf{X}_{Gauss}\right)$, the resulting negentropy objective is

$$\begin{aligned} J\left(\mathbf{X}_{ECG}\right) &= H\left(\mathbf{X}_{Gauss}\right) - H\left(\mathbf{X}_{ECG}\right) \qquad (3.12)\\ &= \log_2\left(\frac{\pi e\sigma^2}{\sqrt{|\Sigma|}}\right) + \log_2\left(\frac{\Gamma\left(q/2\right)}{\pi^{\frac{q}{2}}\Gamma\left(a\right)b^a}\right) - \left(a - q/2\right)\big(\Psi\left(a\right) + \log_2\left(b\right)\big) + a. \qquad (3.13)\end{aligned}$$

This approach corresponds to Independent Subspace Analysis (ISA) (Hyvärinen and Hoyer, 2000) with only one subspace, the elliptically contoured gamma distribution as target distribution, and dimensionality $q = 2$.

3.5.3 Data sets

The time-varying input signal $\mathbf{x}(t)$ was derived from the van Hateren image database (van Hateren and van der Schaaf, 1998). The image database contains over 4000 calibrated monochrome images of 1536×1024 pixels with 12 bit pixel depth. Each pixel corresponds to 0.1 deg of visual angle. We created a temporal sequence by sliding a window of size 16×16, 24×24, 32×32, 64×64, 96×96, 128×128, 196×196, and 256×256, respectively, over the image. Step length and direction for translation were sampled from a continuous uniform distribution on $[-2; 2]$, allowing for subpixel shifts. For each patch size we sampled $1,000,000$ image patch pairs (before and after translation).

The filters in Figure 3.1 (A) were learned on 11×11 patches obtained from the van Hateren movie dataset (van Hateren and Ruderman, 1998) and the filters in (B) were learned on 11×11 patches sampled identically from the van Hateren image database mentioned above with identical translations.

3.6 Acknowledgments

This work was supported by the Max Planck Society and the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601) and the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307).

4 Slow Feature Analysis versus Slow Subspace Analysis

This chapter contains results obtained in the process of developing the previous two articles but which have not been published (yet). The main focus here lies on the differences and commonalities of Slow Feature Analysis (SFA) (Wiskott and Sejnowski, 2002) and Slow Subspace Analysis (SSA) (Kayser et al., 2001; Lies et al., 2013).

In the following, I will first give a short introduction into SFA and subsequently discuss 3 studies on the differences of SFA and SSA and how SSA is better suited for explaining physiological propertied of V1 complex cells.

4.1 Slow Feature Analysis

The general idea is for a given multi-dimensional, time-varying input signal

$$\mathbf{x}(t) = (x_0(t), \ldots, x_{N-1}(t))^\top ,\ t \in [0, T-1] \tag{4.1}$$

to find a real-valued output function $\mathbf{z}(t) = \mathbf{g}(\mathbf{x}(t)), \mathbf{g} : \mathbb{R}^N \mapsto \mathbb{R}^M$ which minimizes the temporal variance

$$\Delta(z_i) := \left\langle \dot{z}_i^2 \right\rangle_t \tag{4.2}$$

under the constraints that the output signal has zero mean, unit variance, and the output dimensions are uncorrelated

$$\langle z_i \rangle_t = 0 \tag{4.3}$$

$$\left\langle z_i^2 \right\rangle_t = 1 \tag{4.4}$$

$$\langle z_i z_j \rangle_t = 0 \quad \forall i \neq j. \tag{4.5}$$

Here, $\langle \cdot \rangle_t$ is the temporal average and $\dot{z}_i$ the temporal derivative. We dropped the explicit notation of the temporal dependence of z_i to make the equations more readable. Constraint 4.3 is merely for convenience to keep the following constraints simpler. Constraint

4.4 ensures that the trivial solution $\mathbf{g}\left(\mathbf{x}(t)\right) = \mathbf{0}$ is not feasible and constraint 4.5 enforces that the output encodes different information in each dimension. Note, however, that $\mathbf{g}(\mathbf{x})$ could be a mapping that entirely ignores $x_2, \ldots, x_n$ by using a set of nonlinear functions $\{g_k(x_1)\}_k$ that all only depend on x_1 and that Constraint 4.4 also precludes the possibility to model perfect invariances as otherwise possible in case of compact Lie groups. Constraints 4.4 and 4.5 can be combined into

$$\left\langle \mathbf{z}\mathbf{z}^\top \right\rangle_t = \mathbb{1} \tag{4.6}$$

where $\mathbb{1}$ is the identity matrix of matching dimensionality.

The common approach (Wiskott and Sejnowski, 2002; Berkes and Wiskott, 2005) is to decompose the transformation $\mathbf{g}(\mathbf{x}(t))$ into a nonlinear stage followed by a linear transformation

$$\mathbf{z}(t) = g\left(\mathbf{x}(t)\right) = U^\top \mathbf{h}(\mathbf{x}(t)). \tag{4.7}$$

with $U \in \mathbb{R}^{Q \times M}$ and $\mathbf{h} : \mathbb{R}^N \mapsto \mathbb{R}^Q$. In the following, we write $\mathbf{h}$ instead of $\mathbf{h}\left(\mathbf{x}\left(t\right)\right)$ to improve readability. If we assume that, without loss of generality, $\langle h_i \rangle_t = 0$ then the constraint (4.3) is fulfilled trivially. Constraints (4.4) and (4.5) can be combined and then take the form of

$$\begin{aligned} \langle z_i z_j \rangle_t &= \left\langle \left(\mathbf{u}_i^\top \mathbf{h}\right)\left(\mathbf{u}_j^\top \mathbf{h}\right) \right\rangle_t = \mathbf{u}_i^\top \left\langle \mathbf{h}\mathbf{h}^\top \right\rangle_t \mathbf{u}_j \\ &= \mathbf{u}_i^\top C_h \mathbf{u}_j \end{aligned} \tag{4.8}$$

where C_h is the covariance of $\mathbf{h}\left(\mathbf{x}\left(t\right)\right)$. Similarly, temporal variance can be rewritten as

$$\begin{aligned} \Delta\left(z_i\right) = \langle \dot{z}_i^2 \rangle_t &= \left\langle \left(\mathbf{u}_i^\top \dot{\mathbf{h}}\right)^2 \right\rangle_t = \mathbf{u}_i^\top \left\langle \dot{\mathbf{h}}\dot{\mathbf{h}}^\top \right\rangle_t \mathbf{u}_i \\ &= \mathbf{u}_i^\top C_{\dot{h}} \mathbf{u}_i. \end{aligned} \tag{4.9}$$

From a machine learning perspective, SFA can be seen as a special case of oriented principal component analysis (PCA) (Diamantaras and Kung, 1994; Bethge et al., 2007) where the static variance (Eq. 4.8) is interpreted as signal energy and the temporal variance (Eq. 4.9) as noise energy and the goal is to maximize the signal-to-noise ratio (SNR) or identically minimize its inverse

$$v\left(z_i\right) = \frac{\mathrm{Var}\left[\dot{z}_i\right]}{\mathrm{Var}\left[z_i\right]} = \frac{\left\langle \dot{z}_i^2 \right\rangle_t}{\left\langle z_i^2 \right\rangle_t} = \frac{\mathbf{u}_i^\top C_{\dot{h}} \mathbf{u}_i}{\mathbf{u}_i^\top C_h \mathbf{u}_i}. \tag{4.10}$$

It is well known that this type of optimization problem can be solved by solving the generalized eigenvalue problem (Golub and Van Loan, 1996):

$$C_h U V = C_{\dot{h}} U, \tag{4.11}$$

where U is the matrix of the generalized eigenvectors and V is the diagonal matrix of the generalized eigenvalues. The smallest eigenvalue corresponds to the smallest inverse SNR, i.e. the largest slowness, and the corresponding eigenvector describes the slowest direction in feature space.

For the learning of complex cells (Berkes and Wiskott, 2005) the algorithm has been applied in the full quadratic feature space, i.e. the expansion of the input into the space of all monomials of degree 2. One potential shortcoming of this expansion is that it leads to a drastic increase in the dimensionality of the data, which presents a computational challenge. Moreover, it weakens the link between the representation and the raw-data, which may cause the representation to focus on irrelevant aspects.

4.2 Spike-triggered covariance analysis of SFA

In this first study we show that SFA fails to reproduce the characteristic findings of spike-triggered covariance (STC) analysis for real neurons. For STC analysis (de Ruyter van Steveninck and Bialek, 1988; Bialek and van Steveninck, 2005), the neuron is presented with a series of stimuli which are then grouped into the set of stimuli which elicited a spike in the neuron and the set of all stimuli. To evaluate which features of the stimuli drove the cell the eigendecomposition of the covariance matrix of the spike set C_{spike} and the complete set C_{all} is computed. The eigenvectors of C_{spike} with eigenvalues significantly larger or smaller than the range of eigenvalues of C_{all} define the subspace of excitatory and inhibitory features, respectively. In contrast to spike-triggered average (STA) where only the average excitatory stimulus is computed, STC allows for more complex features. The range of eigenvalues of C_{all} is called the noise level.

Applied to recordings from real neurons, spike-triggered covariance indicates that only a small fraction of dimensions are necessary to predict the neural responses (Rust et al., 2005; Chen et al., 2007), the quadratic forms learned with slow feature analysis, however, have very broad eigenspectra. This implies that a large fraction of dimensions is necessary to predict the neural responses in the SFA model. In order to compare the theoretically derived SFA features to the STC data, we generated Poisson spike responses with matched firing rates and an identical time window.

We computed SFA filters on the quadratic feature space of the 100 lowest Fourier components of 11×11 image patches sampled from the van Hateren image database (van Hateren and van der Schaaf, 1998). As temporal transformation we applied a 2D translation with shift amplitudes drawn from a 2D uniform continuous distribution on $[-2:2]$ to the data. Subsequently, we applied exactly the same STC analysis in (Chen et al., 2007; Rust et al., 2005) to the model data. We applied a sequence of 50000 Gaussian white noise patterns to the SFA filter. The filter responses were centered at their respective median and split in two firing rate sets, the excitatory from all positive responses (i.e. larger than median) and the inhibitory from the absolute value of all negative responses (i.e. smaller than median). The firing rates were then used to generate Poisson spike counts. Given spike counts and stimuli, we computed the spike triggered covariance (STC) for 100 different noise stimulus sets per SFA filter. To determine which eigenvectors are significant, we computed the STC with shuffled spike counts as control.

We obtained a very large numbers of significant eigenvalues, in clear contradiction to the experimental data shown in Figure 4.1. The first row shows the eigenspectrum of the STC matrix of a complex cell recorded in monkey (A, replotted from (Chen et al., 2007)) and the eigenspectrum of the SFA STC matrix (B). The dashed lines correspond to mean noise level ± 4.4 SD, which corresponds to a confidence interval of $p < 10^{-4}$ for Gaussian distributed eigenvalues. The number of eigenvalues which are significantly larger or smaller than the noise level (big circles) is larger for SFA than for the monkey cell. This becomes most obvious when looking at the histogram of significant excitatory (larger than noise level) and inhibitory (smaller than noise level) eigenvectors. For 130 monkey cells (C, replot from (Chen et al., 2007)) there are only few significant eigenvectors while for 980,000 SFA cells (D), out of the 100 eigenvectors almost all are significant when using the same criterion and a matched noise level.

For SSA, the number of significantly larger eigenvalues is given by design. The number of eigenvalues is equal to the dimensionality of the energy model, i.e. 2 for all simulations shown here. This is because the SSA model only responds to stimuli within the subspace spanned by its simple cell receptive fields. Therefore, the covariance matrix would have non-zero eigenvalues for the basis vectors of the responsive subspace and zero for all non-responding feature dimensions. Note, that the eigenvectors need not be exactly the simple cell receptive fields of the energy model but can be any basis of the feature space. However, for the 2D energy model with quadrature pair Gabor filters this would only lead to an offset in the phase for the two filters.

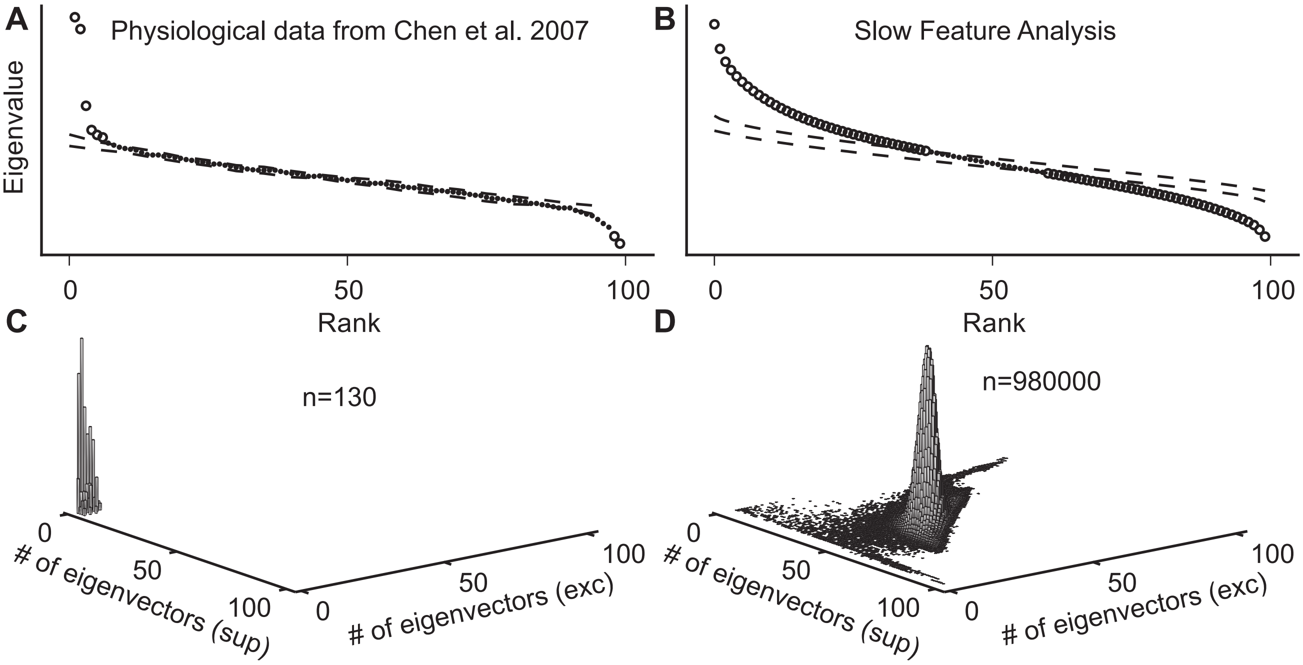


Figure 4.1. Model complex cells derived with SFA fail to reproduce the small numbers of significant eigenvalues found empirically with STC analysis. The spectrum of eigenvalues (eigenspectrum) of the STC matrix of one cell recorded from V1 in an awake monkey (Chen et al., 2007) is shown in (A), the eigenspectrum of the STC of one SFA filter is shown in (B). The dashed lines correspond to the mean noise level ± 4.4 SD, which corresponds to a confidence interval of $p < 10^{-4}$ for Gaussian distributed eigenvalues. One clear difference is the number of significant eigenvectors. While for the V1 cell, only a few eigenvectors are significant, for the SFA model almost all eigenvectors are significant. The histogram of the number of significant excitatory and inhibitory eigenvectors is shown in (C) for the physiological data and in (D) for the SFA model. While the V1 cells have only few significant eigenvectors for all 130 recorded cells, the 980000 cells of the SFA model have on average 80 significant excitatory and inhibitory eigenvectors out of the 100 dimensions. The histogram bins with 0 entries were not plotted for clarity of the figure.

4.3 Average slowness

In the following study, we compare the slowness, i.e. v defined in Equation 4.10, of SSA to SFA. Strictly speaking, Equation 4.10 defines the inverse slowness, this means the smaller v the "slower" are the filter responses. A meaningful comparison can only be carried out with respect to the same feature space. Therefore, we chose as feature space the squared filter responses of the optimized SSA filters $\mathbf{y}(t) = \left(U^\top \mathbf{x}\left(t\right)\right)^2$, where U is the filter matrix learned with SSA, which spans a subspace in the full quadratic feature space. We then applied linear SFA to the feature vectors $\mathbf{y}(t)$ to obtain the SFA filter matrix Q_{SFA}. The

SSA filter matrix Q_{SSA} has the following structure

$$Q_{SSA} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{4.12}$$

which simply sums up the squared filter responses and preserves only half of the dimensions.

Figure 4.2 shows the cumulative sum of the objective

$$v\left(z_i(t)\right) = \frac{\text{Var}\left[\dot{z}_i\right]}{\text{Var}\left[z_i\right]}, \tag{4.13}$$

with $z_i(t) = Q^\top y_i(t)$ and Q being Q_{SFA} for linear SFA (dashed), Q_{SSA} for SSA (solid), and a random filter matrix (dotted). SFA clearly finds the slowest feature, but for more than the first few components, the SFA filters are on average less slow, i.e. have larger v and are therefore "faster", than the SSA filters. These findings are consistent for translation (Figure 4.2A), rotation (Figure 4.2B), scaling (Figure 4.2C), and natural movie sequences (Figure 4.2D). The effect is the strongest on natural movie sequences, with the SFA filters being 56% "faster" on average than the SSA filters and only slightly slower than random filters.

4.4 Optimal quadratic feature

To further evaluate how SFA and SSA differ we compared the 8 slowest filters of SSA and SFA over the complete quadratic feature space. We therefore applied SFA and SSA to the quadratic feature space over the 16 lowest Fourier components of the circulant pink noise, 2D translation, and movie data set used in Chapter 2 (Lies et al., 2013). The filters in Figure 4.3 show the quadratic feature space where the diagonal are the squared components and the off-diagonal entries are the product of component i and j. Since the matrix is symmetric, only the upper triangular part was used for training. The filters are ordered in decreasing slowness from top to bottom.

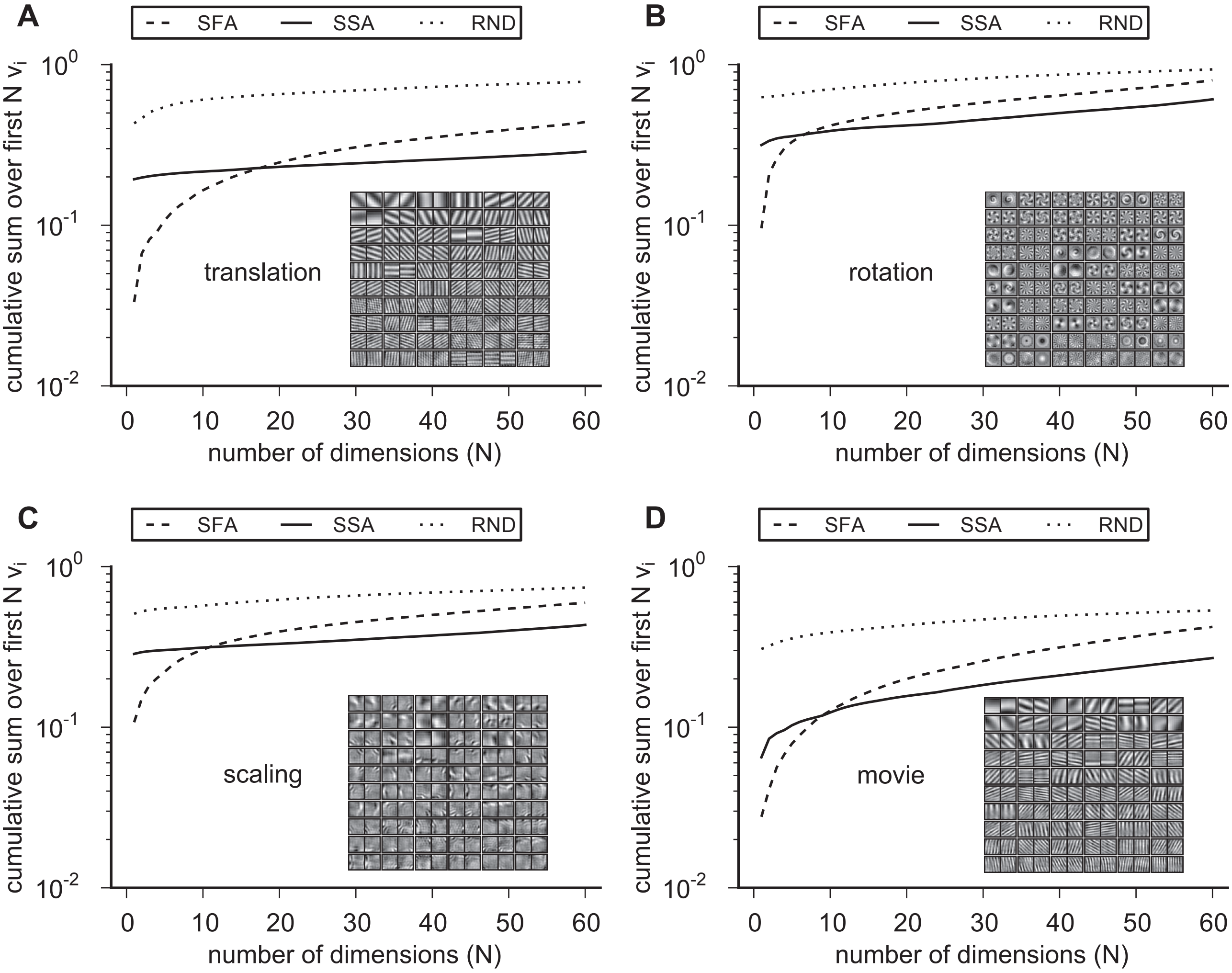


Figure 4.2. Linear Slow Feature Analysis (SFA) learns faster filters than SSA. Linear SFA applied on the squared filter responses of the SSA filters obtained in Figures 2.2 and 2.4, shown as inlays. The panels show the cumulative sum of the objective v of SFA (dashed line), SSA (solid line), and a random filter matrix (dotted line). (A) shows the results of translation, (B) of patch-centered rotation, (C) of anisotropic scaling and (D) of the van Hateren movie sequences. SFA is slower than SSA for the first few components, but SSA has the better overall slowness for all 4 data sets.

Both SSA and SFA recover the quadrature pairs of the Fourier power spectrum for the ideal case of shifts with periodic boundary conditions (A). For all filters only two squared components are active which correspond to the Fourier components with equal spatial frequency and orientation. For translation (B) and movie data (C) SSA still finds the quadrature pairs as slowest components, although not as noise-free as for the circular

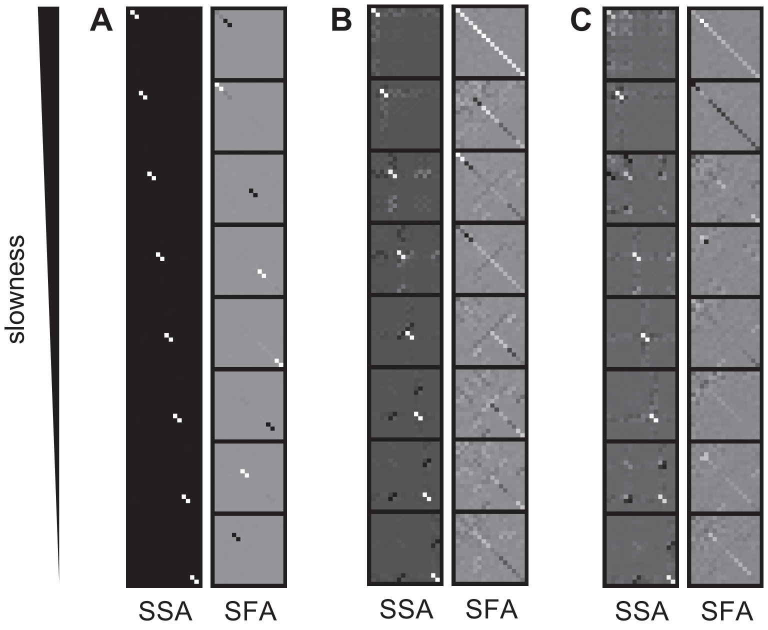


Figure 4.3. Comparison of quadratic features learned with SSA and SFA. Each column visualizes the features learned with SSA and SFA for three different cases: shifts with periodic boundary conditions (A), shifts without periodic boundary conditions (B), and van Hateren movie (C). Within a column each patch shows the weights in the full quadratic feature space spanned by all pairwise products of sixteen Fourier basis coefficients. In the ideal case of shifts with periodic boundary conditions, the learned features of both SSA and SFA perfectly resemble the quadrature pairs of the Fourier power spectrum (A). However, SFA extracts completely different features if the artificially imposed periodic boundary conditions are dropped, whereas the features of SSA are still similar to the quadrature pairs of the Fourier power spectrum (B). The same is true for the case of the van Hateren movie (C).

data. SFA, in contrast, no longer finds the quadrature pairs. The greedy optimization strategy of SFA over the full quadratic feature space determines the total image variance as the slowest feature, since the slowest filter pools over all quadratic components. Also, the following features are not selective for a characteristic subspace but integrate over a large set of dimensions for both data sets.

4.5 Conclusion

The conclusion from these studies is that the simple subspace energy model underlying SSA provides a more suitable framework to learn transformation invariances than the quadratic feature space approach of SFA, because it better resembles the structure found in primary visual cortex. Furthermore, employing the full quadratic feature space leads to a rapid inflation of the number of parameters—growing as $\mathcal{O}(n^4)$—because the input dimensionality increases from n to $(n^2 + n)/2$ dimensions. While still tractable for small dimensions, SFA learning becomes quickly impracticable and noise sensitive for increasing n. SSA, in contrast, makes use of a principled inductive bias borrowed from steerable filter theory to set up the feature space. By working directly on the input (pixel) space, the number of parameters to optimize is in $\mathcal{O}(n^2)$ and all information is preserved by providing a complete filter set over the input space.

While these results indicate that SFA has a lot of drawbacks it still has clear advantages such as providing a closed-form solution and very effective dimensionality reduction. The focus here is clearly on the ability to explain physiologically found properties of complex cells and the ability to extract Lie group operators, for which SSA is an advantageous choice over SFA.

As a final difference between SSA and SFA that I investigated in the course of studying the slowness objective is the average "slowness" of the obtained filters. In the case of slowness optimization over the complete space of squared filter responses, SSA exhibits better overall slowness while SFA finds the slowest features. This means that for finding the one or two slowest components, e.g. for extracting a single source from a noisy signal, SFA is the better choice. However, for building a complete and slow representation of the feature space, SSA finds the on average slower features thus providing a more invariant representation.

5 Discussion

The main conclusions of this work are 1) that slowness and redundancy reduction are contradictory objectives with completely different optima and 2) neither can solely account for the different complex cell properties. I will discuss the individual findings in detail in the following sections.

5.1 Slowness

We found that slowness, as defined by Kayser et al. (2001) or Wiskott and Sejnowski (2002), leads to global, Fourier-like receptive fields in an energy model framework. This result holds for a range of stimuli from periodic shifts to natural movie sequences. By optimizing the envelope size of Gabor filters we found that the optimal size is limited only by the patch size and boundary effects independent of the spatial frequency of the Gabor. This stands in contrast to previous findings who all found complex cell properties. Kayser et al. (2001) used undercomplete SSA, Körding et al. (2004) used overcomplete SSA, Berkes and Wiskott (2005, 2002) used SFA on the quadratic feature space, and Kayser et al. (2003) used SSA but learned the exponent of the energy model. All these previous studies found localized receptive fields with physiologically plausible bandwidths. However, my findings suggest that this is not surprising but rather caused by the choice of simulation parameters. As shown in Figure 3.7 in the previous chapter, for small patch sizes up to 24×24 pixels, the bandwidths of the optimal Gabors are within the physiological range. Kayser et al. (2001) used 10×10 patches without dimensionality reduction, Kayser et al. (2003) used two 30×30 patches windowed by a Gaussian and then reduced to the first 120 principal components (out of 1800) and Körding et al. (2004) used only 99 principal components. Berkes and Wiskott (2005) reduced two 16×16 patches to the first 100 principal components before they expanded the quadratic feature space and ran two 10×10 patches without dimensionality reduction as control. All these data lie well below the minimum patch size where the lack of localization would become imminent. While the two studies by Kayser et al. (2003) and Körding et al. (2004) use relatively large image patches of 30×30, the significant dimensionality reduction using PCA reduces the input signal to the low frequency components thus truncating the high frequency components which could lead to narrow bandwidth filters. A repetition of the aforementioned studies should lead to receptive fields dissimilar to complex cells when patches of at least 64 × 64 become computationally feasible as input data, as suggested by my findings. However, 64 × 64 patches require optimization over $SO(4096)$ with approximately 50 million patches to fully determine the 4096 × 4096 SSA filter matrix. For SFA, the expansion of the 4096-dimensional input space into the quadratic feature space would lead to a feature space in the regime of 10^7 dimensions thus requiring the eigenvalue decomposition of a covariance matrix with more than 10^{14} degrees of freedom. Computationally, such a large feature space is beyond what is reasonable today. The use of such large feature spaces would also pose an enormous challenge for the early visual system. The complete human primary visual cortex has approximately $1.4 \cdot 10^8$ cells (Leuba and Kraftsik, 1994) and each cell receives input from 1,000 to 10,000 cells and its receptive field covers approximately 5,000 to 15,000 photo receptors. Thus if the visual system would use the quadratic feature space expansion it would have to use only a subspace of the complete feature space. However, even if we take the lower bound on the number of receptors a V1 cell pools over, patch sizes of 24 × 24 or below largely underestimate the "physiological" patch size and thus inducing a bias towards low frequency components.

Not only dimensionality reduction but also the attempt to avoid edge effects and anisotropy of square patches significantly shifts the results towards a regime in which complex cell-like results are seen. As an example I repeated the simulation of Figure 2.2 A where SSA was applied to planar shifts but applied the windowing used by Körding et al. (2004), in which patchs were multiplied with a Gaussian to attenuate the pixels further away from the center. Patch and window size were chosen as in (Körding et al., 2004). While the filters without the Gaussian envelope are clearly the Fourier basis (Figure 5.1, left), the filters learned on the preprocessed image patches exhibit localization and resemble V1 receptive fields (Figure 5.1, right). This shows that preprocessing can introduce a bias towards physiological receptive field properties even though the image statistics might not give rise to them otherwise.

Sprekeler and Wiskott (2011) realized recently that SFA does not depend on the (static) statistics of the input images but only on the transformational information. Similar to Bethge et al. (2007) they showed in a theoretical framework that SFA learns Lie generators where their derivation is independent of the spatial statistics of the input signal and depends only on the transformation applied to it. Complex cell properties like side- and end-inhibition, as found by Berkes and Wiskott (2005), can be explained within the framework of Sprekeler and Wiskott (2011) as breaking the translation invariance induced

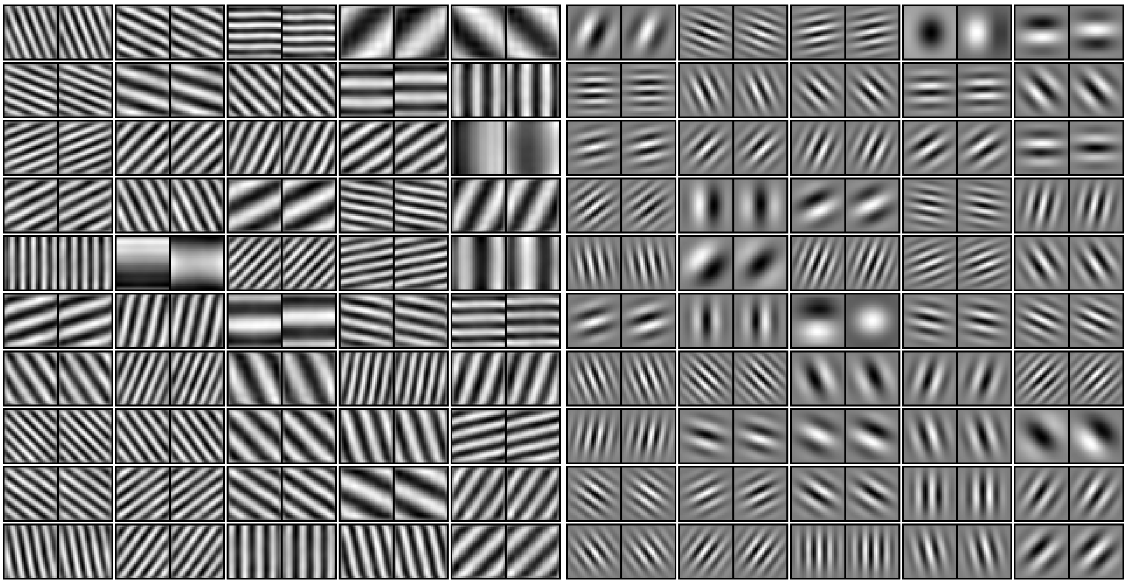


Figure 5.1. SSA on planar shifts with and without preprocessing. The filters are learned using SSA on van Hateren image patches with planar shifts as in Figure 2.2. The only difference between the two sets of filters is that the left filters are learned without special preprocessing while the right filters were filtered by a Gaussian envelope as done by Körding et al. (2004). While the left filters clearly resemble the Fourier basis, the right filters largely resemble V1 receptive fields. Thus the preprocessing induced a bias towards physiological receptive field appearance.

by the finite data. However, the independence of the spatial statistics makes it difficult for the slowness objective to account for degenerations in receptive field properties due to changes in the environment (Hirsch and Spinelli, 1970, 1971; Blakemore and Cooper, 1970; Löwel and Singer, 1992; Wong, 1999; Albert et al., 2008), for example as introduced by raising kittens in an environment consisting purely of vertical or horizontal bars (Blakemore and Cooper, 1970).

5.2 Redundancy Reduction

Redundancy reduction has been successfully used to derive bandpass filtering properties in retina and LGN (Buchsbaum and Gottschalk, 1983; Ruderman et al., 1998; Atick and Redlich, 1990, 1992; van Hateren, 1992, 1993; Dong and Atick, 1995) as well as to derive localization and orientation selectivity in V1 simple cells (Olshausen and Field, 1996, 1997; Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998) and phase invariance in V1 complex cells (Hyvärinen and Hoyer, 2000). Here, we investigated the explanatory power of redundancy reduction for V1 complex cell properties by means of sparseness maximization. We used the ISA framework (Hyvärinen and Hoyer, 2000) and modeled the subspace responses as elliptically contoured Gamma distributions. The objective was the negentropy, i.e. the non-Gaussianity, of the subspace responses which can be computed analytically (Hosseini and Bethge, 2013).

We compared the optimal filters with different physiological properties and found that redundancy reduction can only explain some of them. The optimal bandwidth of the filters showed a clear finite optimum within the range of physiological data (Ringach, 2002; Jones and Palmer, 1987a) but only for unphysiologically large aspect ratios of the Gaussian envelope. Furthermore, the preferred spatial frequency was as high as possible and only limited to the Nyquist frequency defined by the discretization of the image patch. We also found wavelet scaling properties, i.e. the Gaussian envelope of the Gabor filter increased proportionally to its wavelength. This property is also found in cells in the primary visual cortex of monkey (Ringach, 2002) and cat (Jones and Palmer, 1987a).

The redundancy reduction objectives used here have no temporal component, i.e. only depend on the spatial statistics of the image. Thus taking a natural image sequence and shuffling the frames and destroying the temporal correlations would lead to exactly the same results. Destroying (parts of the) temporal correlations with strobe rearing in cats leads to receptive field deformations (Humphrey and Saul, 1998; Crémieux et al., 1987). However, the strobe lighting not only destroys temporal correlations but also alters the spatial lighting in the scene. It is not clear to what extent the receptive field changes are caused by the temporal disruptions or by the lighting changes of the scene. Thus the redundancy reduction objective could be able to explain the receptive field changes.

5.3 Combining slowness and redundancy reduction

A straightforward approach to combining the advantages of slowness and redundancy reduction is a joint optimization. Several models with combined slowness and redundancy reduction objective have been published (Hyvärinen et al., 2003; Berkes et al., 2009; Einhäuser et al., 2002; Masquelier et al., 2007; Cadieu and Olshausen, 2009, 2012), but the effect of slowness and redundancy reduction individually on the resulting filters has never been investigated. Here, I present some of the most recognized studies and how they relate to our findings.

Einhäuser et al. (2002) and Masquelier et al. (2007) presented a multi-layer model with a simple cell layer and a complex cell layer as proposed by Riesenhuber and Poggio (1999). In their 3-layer model, the first layer is the input, the second layer learns simple cell properties and the third layer learns complex cell properties. Einhäuser et al. (2002) derived an interesting prediction from their neural network model; they found that the simple cell layer optimizes sparseness while the complex cell layer optimizes slowness. Based on this observation, they predicted that temporal disruptions of the input at about 3 Hz would destroy the formation of complex cell properties in their network but not simple cell properties. Masquelier et al. (2007) verified this with their model. However, Crémieux et al. (1987) carried out an experiment where they found contradictory results. They raised cats in stroboscopic environments with 2 Hz frequency and light pulse duration of 200 µs. According to their data, the number of complex cells as well as their properties are comparable between strobe-reared cats and normal cats. The alterations found, for example increases in receptive field size or decreases in number of direction selective cells, were comparable between all cell types. This indicates that the formation of complex cells is not disturbed by temporal disruptions. However, the different findings of the physiological and computational studies could be caused by the slightly different experimental conditions, i.e. the difference in strobe frequency and duration.

If simple cells and complex cells use different objectives to obtain their properties, they would have to be two different cell classes. Yet, the segregation into simple cells and complex cells is still a matter of debate (Dean and Tolhurst, 1983; Chance et al., 1999; Mechler and Ringach, 2002; Priebe et al., 2004). Further, Fournier et al. (2011) found that the ratio of complex cells to simple cells depends on the stimulus used to classify them. This would require that the cells have both learning paradigms implemented in parallel to obtain simple and complex cell properties depending on the input statistics.

The bubbles framework (Hyvärinen et al., 2003) and the similar framework of (Berkes et al., 2009) offer joint optimization of slowness and redundancy reduction. Simple and complex cells have the same objective thus making the bubbles framework more likely to be able to explain the shift from complex cell to simple cell response as found by Fournier et al. (2011). The spatial filters largely resemble those from ISA (Hyvärinen and Hoyer, 2000), or its extension to topographical ICA (Hyvärinen et al., 2001), which uses only independence as objective. This leaves the question how much the slowness objective contributes to the final form of the receptive fields.

The temporal coherence objective (Hurri and Hyvärinen, 2003) is not identical to the slowness objective of Wiskott (2003) and Kayser et al. (2001), as temporal coherence maximizes the product of the responses $r(\cdot)$ of consecutive time steps

$$\max_t f_{tc}(t) \propto r(t)r(t-1) \tag{5.1}$$

while SFA and SSA minimize the variance of the difference between consecutive time steps

$$\min_t f_{sfa}(t) \propto (r(t) - r(t-1))^2 . \tag{5.2}$$

Assuming that the variance of $r(t)$ and $r(t-1)$ over t is identical (i.e. $r(\cdot)$ is a stationary process), we see that

$$f_{sfa}(t) \propto (r(t) - r(t-1))^2 = 2r(t)^2 - 2r(t)r(t-1) \propto r(t)^2 - f_{tc}. \tag{5.3}$$

This means that the slowness objective is the difference of response variance (if the response has zero mean) and temporal coherence objective.

Cadieu and Olshausen (2008, 2009, 2012) proposed a two-layer model where the first layer represents local features, i.e. simple and complex cell-like features, and the second layer encodes higher order functionality like form and motion, as found in higher visual areas. The first layer learns complex-valued filters where the filter response amplitudes are optimized for sparseness and slowness simultaneously. Real and imaginary part form a two-dimensional energy model (Adelson and Bergen, 1985). The filters resemble complex cell receptive fields, in agreement with our findings (Lies et al., 2013).

While all models presented here combine slowness and redundancy reduction in quite different ways, the optimal filters are perceptually similar and resemble those found with ISA, i.e. with redundancy reduction alone. Our results show that even if the filters in a combined objective are perceptually closer to the redundancy reduction objective, their performance with respect to both objectives can be larger than 80%. This can explain why the combined models find ISA-like receptive fields. However, as the filter differences are only marginal, the combined models do not explain what advantage the combined objectives provide compared to the simpler redundancy reduction only objective or how much their performance would change if the slowness objective would have been dropped.

5.4 Conclusion

We compared the redundancy reduction and slowness objective on two levels: first on the level of complete filter sets (or a population level) and subsequently on a single cell level. In both cases we found that the objectives have quite opposing optima.

Contrary to previous experimental findings, we found that the slowness objective leads to global receptive fields. The size of the receptive fields is only limited by the patch size and edge effects. Therefore the envelope size did not depend on the wavelength and thus slowness did not exhibit wavelet scaling properties. The optimal aspect ratio of the envelope was between 1 and 1.5 and therefore within the physiologically plausible range. However, this could be caused by the fact that our simulations were run on square patches only, i.e. the aspect ratio of the patches was 1. If patches with, for example, an aspect ratio of 2 would be used it is not clear if the optimal receptive fields would be global with an aspect ratio around 2 or localized in one direction with an aspect ratio between 1 and 1.5.

The slowness optimization over the complete space led to Fourier-like receptive fields for translation data and natural movies and led to circular Fourier-like receptive fields for rotation data. This underlines the pre-dominance of translations in small-scale transformations as previously shown by Wang and Simoncelli (2005). Our findings suggest that slowness would require external constraints, for example wiring length constraints, to limit the extent of the receptive field to a physiologically plausible size. The retina provides this kind of constraint, as the receptive field size is limited and increases with eccentricity.

For the redundancy reduction objective our results are mostly in agreement with previous findings. Redundancy reduction leads to localized, high-frequency filters with large aspect ratios. The optimal aspect ratios are larger than any aspect ratios found in monkey or cat visual cortex. The optimal bandwidth of the filters depends only on the content, not on the size of the filter. It is also independent of the underlying spatial frequency, thus exhibiting wavelet scaling properties. The optimal spatial frequency is as high as possible and only limited by the discretization in the single cell simulation and the requirement of all filters to cover the complete space in the population approach. Receptive fields with very high frequencies have never been recorded in the visual cortex. Therefore redundancy reduction would require physiological constraints which limit the upper frequency available to the primary visual cortex in order to explain low frequency receptive fields. An alternative explanation is that the experimental techniques are not able to map high frequencies and obtain low frequency receptive fields due to undersampling.

A combination of slowness and redundancy reduction has successfully been used in several studies (Hyvärinen et al., 2003; Berkes et al., 2009; Einhäuser et al., 2002; Masquelier et al., 2007; Cadieu and Olshausen, 2009, 2012); however, the optimal filters always resembled more those obtained with redundancy reduction only. We found that the combination of slowness and redundancy reduction leads to better performance than both objectives alone but the redundancy reduction objective dominates the filter shape at the point of equal performance. However, one cannot simply ignore the opposing demands of slowness and redundancy reduction on the complex cell responses. More work is needed to further investigate how both objectives interact and why the combined objective is perceptually closer to redundancy reduction than to slowness.

In summary, both objectives—redundancy reduction and slowness—have its advantages and disadvantages but neither can explain all complex cell properties evaluated in this thesis. Even though they have been seen as equal candidates for complex cell coding, the respective optima are quite different. In models with combined objectives the redundancy reduction seems to dominate the filter shapes. One way to achieve a better understanding of the computational strategy embedded in the visual system could be a better understanding of the quite opposing demands of slowness and sparseness on the response properties of neurons in the visual cortex.

5.5 How to continue

There are several options how to further investigate the role of slowness and redundancy reduction in the emergence of complex cell properties and thus proceed with the line of research presented in this thesis.

The lack of computational power and memory to cope with high-dimensional input data might be resolved by increase in available memory and computational power e.g. of specialized numerical GPUs. However, currently with state-of-the-art systems an experiment of this dimension would require an unreasonably large amount of time for algorithmic design and execution.

Even though the model used here, as most V1 models, assumes a clear segregation into simple and complex cells, the physiological evidence suggests that there is rather a continuum ranging from more linear, simple cell like responses to more non-linear, complex cell like responses (Dean and Tolhurst, 1983; Chance et al., 1999; Priebe et al., 2004; Mechler and Ringach, 2002). The classification of a cell into more simple or more complex can even depend on the experimental stimulus (Fournier et al., 2011). Using a redundancy reduction or slowness measure on recorded cell responses could provide an alternative classification method. For example computing the negentropy of the responses of several cells which are stimulated with natural movies and comparing the negentropy with "classical" simple/complex classification methods. Or does the negentropy classification change with the statistics of the input, as shown for classical methods by Fournier et al. (2011)? And what about the slowness objective? Investigating in that direction would present an interesting experimental project.

Bibliography

Edward H Adelson and James R Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A: Optics and Image Science*, 2(2):284–299, February 1985. ISSN 0740-3232.

Mark V Albert, Adam Schnabel, and David J Field. Innate Visual Learning through Spontaneous Activity Patterns. *PLoS Computational Biology*, 4(8):8, 2008.

Duane G Albrecht and David B Hamilton. Striate cortex of monkey and cat: contrast response function. *Journal of Neurophysiology*, 48(1):217–237, 1982.

Brian W Andrews and Daniel A Pollen. Relationship between spatial frequency selectivity and receptive field profile of simple cells. *The Journal of Physiology*, 287(1): 163–176, 1979.

Joseph J Atick and A Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990.

Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural Computation*, 210(2):196–210, 1992. ISSN 08997667. doi: 10.3406/ahess.1985.283242.

Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61 (3):183–193, 1954. ISSN 0033295X. doi: 10.1037/h0054663.

Tom Baden, Timm Schubert, Le Chang, Tao Wei, Mariana Zaichuk, Bernd Wissinger, and Thomas Euler. Beyond Colour Vision: Dichromacy Provides for Optimal Sampling of Contrast Statistics in Natural Scenes. *in preparation*, 2013.

Horace B Barlow. Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234, 1961. ISSN 15459624.

Horace B Barlow. The Ferrier Lecture, 1980: Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 212(1186):1–34, 1981. ISSN 00804649.

Horace B Barlow. The knowledge used in vision and where it comes from. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 352(1358): 1141–1147, 1997.

Anthony J Bell and Terrence J Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Anthony J Bell and Terrence J Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997. ISSN 00426989. doi: 10.1016/S0042-6989(97)00121-1.

Pietro Berkes and Laurenz Wiskott. Applying Slow Feature Analysis to Image Sequences Yields a Rich Repertoire of Complex Cell Properties. In *Artificial Neural Networks—ICANN 2002*, pages 81–86. Springer Verlag, January 2002.

Pietro Berkes and Laurenz Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):579–602, July 2005. ISSN 1534-7362. doi: 10.1167/5.6.9.

Pietro Berkes, Richard E Turner, and Maneesh Sahani. A Structured Model of Video Reproduces Primary Visual Cortical Organisation. *PLoS Computational Biology*, 5(9):16, 2009.

Matthias Bethge, Sebastian Gerwinn, and Jakob H Macke. Unsupervised learning of a steerable basis for invariant image representations. In *Proceedings of SPIE Human Vision and Electronic Imaging XII (EI105)*, volume 6492, page 12, 2007. doi: 10.1117/12.711119.

W Bialek and RR van Steveninck. Features and dimensions: Motion estimation in fly vision. *arXiv preprint q-bio/0505003*, 2005.

Irving Biederman and Eric E Cooper. Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20(5):585–593, 1991.

Irving Biederman, Eric E Cooper, Zoe Kourtzi, Pawan Sinha, and Johan Wagemans. Biederman and Cooper's 1991 paper. *Perception*, 38(6):809–826, 2009. ISSN 03010066. doi: 10.1068/ldmk-bie.

Colin Blakemore and Grahame F Cooper. Development of the brain depends on the visual environment. *Nature*, 228:477–478, 1970. ISSN 00280836. doi: 10.1038/228477a0.

Alfred B Bonds. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, 2(1):41–55, 1989.

Vicki Bruce and Patrick R Green. *Visual Perception: Physiology, Psychology and Ecology*. Psychology Press, 4 edition, 2003. ISBN 1841692387.

Gershon Buchsbaum and Allan Gottschalk. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society B: Biological Sciences*, 220(1218):89–113, 1983.

Christopher J C Burges. Geometric Methods for Feature Extraction and Dimensional Reduction. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, chapter 4, pages 59–92. Kluwer Academic Publishers, 2005. ISBN 0387244352. doi: 10.1007/0-387-25465-X_4.

Geoffrey J Burton and Ian R Moorhead. Color and spatial structure in natural scenes. *Applied Optics*, 26(1):157, January 1987. ISSN 0003-6935. doi: 10.1364/AO.26.000157.

Charles Cadieu and Bruno A Olshausen. Learning Transformational Invariants from Time-Varying Natural Images, 2008.

Charles Cadieu and Bruno A Olshausen. Learning transformational invariants from natural movies. *Advances in Neural Information Processing Systems*, 21:209–216, 2009.

Charles Cadieu and Bruno A Olshausen. Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, 24(4):827–66, 2012. ISSN 1530888X. doi: 10.1162/NECO_a_00247.

Edward M Callaway. Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, 21(1):47–74, 1998.

Matteo Carandini and David J Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264(5163):1333–1336, 1994.

Matteo Carandini, David J Heeger, and J Anthony Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644, 1997.

Matteo Carandini, David J Heeger, and J Anthony Movshon. Linearity and gain control in V1 simple cells. *Cerebral Cortex. Models of Cortical Circuits*, 13:401–443, 1998.

Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–97, November 2005. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.3726-05.2005.

Frances S Chance, Sacha B Nelson, and Larry F Abbott. Complex cells as cortically amplified simple cells. *Nature Neuroscience*, 2(3):277–282, 1999.

Xiaodong Chen, Feng Han, Mu-Ming Poo, and Yang Dan. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proceedings of the National Academy of Sciences of the United States of America*, 104(48):19120–19125, 2007.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, April 1994. ISSN 0165-1684. doi: 10.1016/0165-1684(94)90029-9.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*, volume 6 of *Wiley Series in Telecommunications*. Wiley, 1991. ISBN 0471062596. doi: 10.1177/0022219410375001.

Jacques Crémieux, Guy A Orban, Jacques Duysens, and Bernard Amblard. Response properties of area 17 neurons in cats reared in stroboscopic illumination. *Journal of neurophysiology*, 57(5):1511–35, May 1987. ISSN 0022-3077.

John G Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

John G Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A: Optics and Image Science*, 2(7):1160–1169, July 1985. doi: 10.1364/JOSAA.2.001160.

John G Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(7):1169–1179, 1988. doi: 10.1109/29.1644.

Rob de Ruyter van Steveninck and William Bialek. Real-Time Performance of a Movement-Sensitive Neuron in the Blowfly Visual System: Coding and Information Transfer in Short Spike Sequences. *Proceedings of the Royal Society B: Biological Sciences*, 234(1277):379–414, September 1988. ISSN 0962-8452. doi: 10.1098/rspb.1988.0055.

R L De Valois, E W Yund, and N Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):531–544, 1982. ISSN 00426989. doi: 10.1016/0042-6989(82)90112-2.

Andrew F Dean and David J Tolhurst. On the distinctness of simple and complex cells in the visual cortex of the cat. *Journal Of Physiology London*, 344(1):305–325, 1983. ISSN 00223751.

G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J Neurophysiol*, 69(4):1091–1117, April 1993a.

G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *J Neurophysiol*, 69(4):1118–1135, April 1993b.

NG Deriugin. The power spectrum and the correlation function of the television signal. *Telecommunications*, 1(7):1–12, 1956.

Konstantinos I Diamantaras and Sun-Yuan Kung. Cross-Correlation Neural Network Models. *Signal Processing, IEEE Transactions on*, 42(1):3218–3223, 1994.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2012. ISSN 10974199. doi: 10.1016/j.neuron.2012.01.010.

Peter C. Dodwell. The Lie transformation group model of visual perception. *Perception & Psychophysics*, 34(1):1–16, January 1983. ISSN 0031-5117. doi: 10.3758/BF03205890.

Dawei W. Dong and Joseph J. Atick. Statistics of natural time-varying images. *Network Computation in Neural Systems*, 6(3):345–358, 1995. ISSN 0954898X. doi: 10.1088/0954-898X/6/3/003.

Wolfgang Einhäuser, Christoph Kayser, Peter König, and Konrad P Körding. Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience*, 15(3):475–486, 2002.

George Ettlinger. "Object vision" and "spatial vision": the neuropsychological evidence for the distinction. *Cortex*, 26(3):319–341, 1990.

Ky Fan and Alan J Hoffman. Some Metric Inequalities in the Space of Matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, 1955. ISSN 00029939.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A: Optics and Image Science*, 4(12):2379–2394, December 1987. ISSN 0740-3232.

David J Field and David J Tolhurst. The structure and symmetry of simple-cell receptive-field profiles in the cat's visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 228(1253):379, September 1986. ISSN 0080-4649.

Paul Emil Flechsig. *Gehirn und Seele*. Veit & Co, 2nd edition, 1896.

Peter Földiák. Learning Invariance from Transformation Sequences. *Neural Computation*, 3(2):194–200, 1991. ISSN 08997667. doi: 10.1162/neco.1991.3.2.194.

Julien Fournier, Cyril Monier, Marc Pananceau, and Yves Frégnac. Adaptation of the simple or complex nature of V1 receptive fields to visual statistics. *Nature Neuroscience*, 14(8):1053–1060, 2011. doi: 10.1038/nn.2861.

Mathias Franzius, Henning Sprekeler, and Laurenz Wiskott. Slowness and Sparseness Lead to Place, Head-Direction, and Spatial-View Cells. *PLoS Computational Biology*, 3 (8):18, 2007.

Mathias Franzius, Niko Wilbert, and Laurenz Wiskott. Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, 23(9):2289–2323, 2011.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–201, September 2011. ISSN 1546-1726. doi: 10.1038/nn.2889.

William T Freeman and Edward H Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

Dennis Gabor. Theory of communication. Part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93 (26):429–441, 1946a. ISSN 17579708. doi: 10.1049/ji-3-2.1946.0074.

Dennis Gabor. Theory of communication. Part 2: The analysis of hearing. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93 (26):442–445, 1946b.

Dennis Gabor. Theory of communication. Part 3: Frequency compression and expansion. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):446–457, 1946c.

Ricardo Gattass, Charles C Gross, and J H Sandell. Visual topography of V2 in the macaque. *Journal of Comparative Neurology*, 201(4):519—-539, 1981.

Michael S. Gazzaniga, Richard B. Ivry, and George Ronald Mangun. *Cognitive Neuroscience: The Biology of the Mind*. 3rd edition, 2009. ISBN 9780393927955.

Wilson S Geisler and Duane G Albrecht. Cortical neurons: isolation of contrast gain control. *Vision Research*, 32(8):1409–1410, 1992.

James J Gibson. *The Perception of the Visual World*. Houghton Mifflin, 1950. ISBN 0837178363. doi: 10.1192/bjp.98.413.717-a.

Robert Gilmore. *Lie Groups, Physics, and Geometry: An Introduction for Physicists, Engineers, and Chemists*, volume 62. Cambridge University Press, 1st edition, 2008. ISBN 978-0521884006.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 3rd edition, October 1996. ISBN 0801854148. doi: 10.1063/1.3060478.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–5, 1992. ISSN 01662236. doi: 10.1016/0166-2236(92)90344-8.

Gösta H Granlund and Hans Knutsson. *Signal Processing for Computer Vision*. Springer US, 1995. ISBN 978-1-4419-5151-9. doi: 10.1007/978-1-4757-2377-9.

Yoshihiko Hamamoto, Shunji Uchimura, Masanori Watanabe, Tetsuya Yasuda, Yoshihiro Mitani, and Shingo Tomita. A Gabor filter-based method for recognizing handwritten numerals. *Pattern Recognition*, 31(4):395–400, 1998. doi: 10.1016/S0031-3203(97)00057-5.

David J Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2):181–197, 1992a.

David J Heeger. Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9(5): 427–443, 1992b. ISSN 09525238. doi: 10.1063/1.2808271.

Yacov Hel-Or and Patrick C Teo. Canonical decomposition of steerable functions. *Journal of Mathematical Imaging and Vision*, 9(1):83–95, 1998.

Salomon E Henschen. On the visual path and centre. *Brain*, 16(1-2):170–180, 1893.

Charles Hermite. *Sur un nouveau développement en série des fonctions*. Gauthier-Villars, 1864.

Geoffrey E Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1-3): 185–234, 1989. ISSN 00043702. doi: 10.1016/0004-3702(89)90049-0.

Helmut V B Hirsch and D N Spinelli. Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science*, 168(3933):869–871, 1970.

Helmut V B Hirsch and D N Spinelli. Modification of the distribution of receptive field orientation in cats by selective visual exposure during development. *Experimental Brain Research*, 12(5):509–527, 1971.

William C. Hoffman. The Lie algebra of visual perception. *Journal of Mathematical Psychology*, 3(1):65–98, February 1966. ISSN 00222496. doi: 10.1016/0022-2496(66)90005-8.

Reshad Hosseini and Matthias Bethge. Elliptically contoured gamma distributions. *in preparation*, 2013.

David H Hubel and Thorsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1): 106–154, March 1962. ISSN 0022-3751.

David H Hubel and Thorsten N Wiesel. Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, 26(6):994–1002, 1963.

David H Hubel and Thorsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, 1968.

Allan L Humphrey and Alan B Saul. Strobe rearing reduces direction selectivity in area 17 by altering spatiotemporal receptive-field structure. *Journal of neurophysiology*, 80(6): 2991–3004, December 1998. ISSN 0022-3077.

Jarmo Hurri and Aapo Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–91, March 2003. ISSN 0899-7667. doi: 10.1162/089976603321192121.

Aapo Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 5, pages 3917–3920. IEEE Comput. Soc. Press, 1997. ISBN 0818679190. doi: 10.1109/ICASSP.1997.604766.

Aapo Hyvärinen. Statistical Models of Natural Images and Cortical Visual Representation. *Topics in Cognitive Science*, 2(2):251–264, April 2010. ISSN 17568757. doi: 10.1111/j.1756-8765.2009.01057.x.

Aapo Hyvärinen. Independent Component Analysis: Recent Advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (184), 2013. doi: 10.1098/rsta.2011.0534.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural*

Computation, 12(7):1705–1720, July 2000. ISSN 0899-7667. doi: 10.1162/089976600300015312.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*, volume 21 of *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley-Interscience, 2001. ISBN 047140540X.

Aapo Hyvärinen, Jarmo Hurri, and Jaakko Väyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, 20(7):1237–1252, 2003.

Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer London, 2009. ISBN 1848824904. doi: 10.1007/978-1-84882-491-1.

Anil K Jain and Sushil Bhattacharjee. Text segmentation using Gabor filters for automatic document processing. *Machine Vision and Applications*, 5(3):169–184, 1992.

Judson P Jones and Larry A Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6): 1233–58, December 1987a. ISSN 0022-3077.

Judson P Jones and Larry A Palmer. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1187–1211, 1987b.

Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. ISSN 01651684. doi: 10.1016/0165-1684(91)90079-X.

Joni-Kristian Kamarainen, Ville Kyrki, and Heikki Kälviäinen. Invariance properties of Gabor filter-based features-overview and applications. *Image Processing, IEEE Transactions on*, 15(5):1088–1099, 2006. doi: 10.1109/TIP.2005.864174.

Yan Karklin and Michael S Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–86, January 2009. ISSN 1476-4687. doi: 10.1038/nature07481.

Christoph Kayser, Wolfgang Einhäuser, Olaf Dümmer, Peter König, and Konrad P Körding. Extracting Slow Subspaces from Natural Videos Leads to Complex Cells. In *Artificial Neural Networks - ICANN 2001*, volume 2130, pages 1075–1080. Austrian Res Inst Artifical Intelligence, 2001. ISBN 3540424865. doi: 10.1007/3-540-44668-0_149.

Christoph Kayser, Konrad P Körding, and Peter König. Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation*, 15(8):1751–9, August 2003. ISSN 0899-7667. doi: 10.1162/08997660360675026.

A. Harry Klopf. *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere Publishing Corporation, Washington DC, 1982. ISBN 089116202X.

Hans Knutsson and Gösta H Granlund. Texture Analysis Using Two-Dimensional Quadrature Filters. In *IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management*, pages 206–213, 1983.

Christof Koch. *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, USA, 1999. ISBN 0-19-510491-9.

Adam Kohn. Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of neurophysiology*, 97(5):3155–64, May 2007. ISSN 0022-3077. doi: 10.1152/jn.00086.2007.

Teuvo Kohonen. Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75(4):281–291, 1996. ISSN 03401200. doi: 10.1007/s004220050295.

Konrad P Körding, Christoph Kayser, Wolfgang Einhäuser, and Peter König. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91(1):206–212, 2004.

Ernest R Kretzmer. Statistics of television signals. *The Bell System Technical Journal*, 31: 751–763, 1952.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, January 1953. ISSN 0022-3077.

Janus J Kulikowski and Peter O Bishop. Fourier analysis and spatial representation in the visual cortex. *Experimentia*, 37:160–163, 1981a.

Janus J Kulikowski and Peter O Bishop. Linear Analysis of the Responses of Simple Cells in the Cat Visual Cortex. *Experimental Brain Research*, 44(4):386–400, 1981b.

Janus J Kulikowski, Stjepan Marcelja, and Peter O Bishop. Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biological Cybernetics*, 43(3):187–198, 1982.

Tai Sing Lee. Image representation using 2D Gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):959–971, 1996. ISSN 01628828. doi:

10.1109/34.541406.

Peter Lennie. The cost of cortical computation. *Current Biology*, 13(6):493–497, 2003.

Geneviève Leuba and Rudolf Kraftsik. Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. *Anatomy and Embryology*, 190(4):351–66, October 1994. ISSN 0340-2061.

Jörn-Philipp Lies, Ralf M Häfner, and Matthias Bethge. Slowness and sparseness have diverging effects on complex cell learning. *under review*, 2013.

Nikos K Logothetis and David L Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621, 1996.

Per-Olov Löwdin. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *The Journal of Chemical Physics*, 18(3):365–375, 1950. ISSN 00219606. doi: 10.1063/1.1747632.

Siegrid Löwel and Wolf Singer. Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, 255(5041):209–212, 1992.

David G Luenberger. *Optimization by vector space methods*. Wiley-Interscience, 1969.

Stjepan Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America A: Optics and Image Science*, 70(11):1297–1300, 1980. ISSN 00303941. doi: 10.1364/JOSA.70.001297.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, volume 8. W.H. Freeman, 1982. ISBN 0716715678. doi: 10.1007/s11097-009-9141-7.

David Marr and Tomaso Poggio. From understanding computation to understanding neural circuitry. *AI Memo*, 357:1–22, 1976.

Jean-Bernard Martens. The Hermite transform - theory. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 38(9):1595–1606, 1990. doi: 10.1109/29.60086.

Jean-Bernard Martens. Local orientation analysis in images by means of the Hermite transform. *Image Processing, IEEE Transactions on*, 6(8):1103–1116, 1997.

Timothee Masquelier, Thomas Serre, and Tomaso Poggio. Learning complex cell invariance from natural videos: A plausibility proof. Technical report, Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, 2007.

Andreas Maurer. Unsupervised slow subspace-learning from stationary processes. In *Proceedings of the 17th international conference on Algorithmic Learning Theory*, volume 4264 of *Lecture Notes in Computer Science*, pages 363–377, Berlin, Heidelberg, October 2006. Springer-Verlag. ISBN 978-3-540-46649-9. doi: 10.1007/11894841.

Andreas Maurer. Unsupervised slow subspace-learning from stationary processes. *Theoretical Computer Science*, 405(3):237–255, October 2008. ISSN 03043975. doi: 10.1016/j.tcs.2008.06.054.

Ferenc Mechler and Dario L Ringach. On the classification of simple and complex cells. *Vision Research*, 42(8):1017–1033, April 2002. ISSN 00426989.

Xu Miao and Rajesh P N Rao. Learning the Lie groups of visual invariance. *Neural computation*, 19(10):2665–2693, 2007. doi: 10.1.1.3.8367.

Mortimer Mishkin and Leslie G Ungerleider. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioural Brain Research*, 6(1):57–77, 1982.

Graeme Mitchison. Removing Time Variation with the Anti-Hebbian Differential Synapse. *Neural Computation*, 3(3):312–320, September 1991. ISSN 0899-7667. doi: 10.1162/neco.1991.3.3.312.

J Anthony Movshon, Ian D Thompson, and David J Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *The Journal of Physiology*, 283(1): 79, October 1978a. ISSN 0022-3751. doi: VL-283.

J Anthony Movshon, Ian D Thompson, and David J Tolhurst. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of Physiology*, 283 (1978):53–77, 1978b.

Sergei S Nikonov, Roman Kholodenko, Janis Lem, and Edward N Pugh. Physiological features of the S- and M-cone photoreceptors of wild-type mice from single-cell recordings. *The Journal of General Physiology*, 127(4):359–374, 2006.

Klas Nordberg, Gösta H Granlund, and Hans Knutsson. Representation and learning of invariance. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 585–589. IEEE Comput. Soc. Press, 1994. ISBN 0-8186-6952-7. doi: 10.1109/ICIP.1994.413638.

I Ohzawa and W Freeman. Spatial pooling of subunits in complex cell receptive fields. *Soc Neurosci Abstr*, 1997.

Izumi Ohzawa, Gary Sclar, and Ralph D Freeman. Contrast gain control in the cat visual cortex. *Nature*, 298(5871):266–268, 1982.

Izumi Ohzawa, Gary Sclar, and Ralph D Freeman. Contrast gain control in the cat's visual system. *Journal of Neurophysiology*, 54(3):651–667, September 1985.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 0028-0836. doi: 10.1038/381607a0.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997. ISSN 00426989. doi: 10.1016/S0042-6989(97)00169-7.

Pietro Perona. Deformable kernels for early vision. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 17, pages 222–227. IEEE Comput. Sco. Press, May 1991. ISBN 0-8186-2148-6.

Pietro Perona. Deformable kernels for early vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(5):448–499, 1995.

Dinh Tuan Pham and Philippe Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *Signal Processing, IEEE Transactions on*, 45(7):1712–1725, 1997. ISSN 1053587X. doi: 10.1109/78.599941.

D.A. Pollen and S.F. Ronner. Visual cortical neurons as localized spatial frequency filters. *IEEE Transactions on Systems, Man, & Cybernetics*, 1983.

Nicholas J Priebe, Ferenc Mechler, Matteo Carandini, and David Ferster. The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience*, 7(10):1113–22, 2004.

Rajesh P N Rao and Daniel L Ruderman. Learning Lie groups for invariant visual perception. *Advances in Neural Information Processing Systems*, 11:810–816, 1999. doi: 10.1.1.50.8859.

Max Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–25, November 1999. ISSN 1097-6256. doi: 10.1038/14819.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, 88(1):455–63, July 2002. ISSN 0022-3077.

Carlos Joel Rivero-Moreno and Stéphane Bres. Conditions of Similarity between Hermite and Gabor Filters as Models of the Human Visual System. *Computer Analysis of Images and Patterns Proceedings*, 2756:762–769, 2003. ISSN 03029743.

Daniel L Ruderman, Thomas W Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America A: Optics and Image Science*, 15(8):2036–2045, 1998. ISSN 10847529. doi: 10.1364/JOSAA.15.002036.

Nicole C Rust, Odelia Schwartz, J Anthony Movshon, and Eero P Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46(6):945–56, June 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.05.021.

Thomas Serre and Tomaso Poggio. Models of visual cortex, 2011. ISSN 19416016.

Claude E Shannon and Warren Weaver. *Mathematical Theory of Communication*, volume 27. University of Illinois Press, 1949. ISBN 0252725484.

Eero P Simoncelli and William T Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Image Processing, International Conference on*, volume 3, pages 444–447. IEEE Comput. Soc. Press, 1995. ISBN 0818673109. doi: 10.1109/ICIP.1995.537667.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. ISSN 0147-006X. doi: 10.1146/annurev.neuro.24.1.1193.

Fabian Sinz, Jörn-Philipp Lies, Sebastian Gerwinn, and Matthias Bethge. NATTER: A Python Natural Image Statistics Toolbox. *submitted*, 2013.

Darragh Smyth, Ben Willmore, Gary E Baker, Ian D Thompson, and David J Tolhurst. The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience*, 23(11):4746–59, 2003. ISSN 15292401.

Jascha Sohl-Dickstein, Jimmy C Wang, and Bruno A Olshausen. An Unsupervised Algorithm For Learning Lie Group Transformations. *Arxiv preprint arXiv*, abs/1001.1: 8, 2010.

Michael W Spratling. Learning viewpoint invariant perceptual representations from cluttered images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5): 753–61, May 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.105.

Henning Sprekeler and Laurenz Wiskott. A theory of slow feature analysis for transformation-based input signals with an application to complex cells. *Neural Computation*, 23(2):303–335, 2011.

Henning Sprekeler, Christian Michaelis, and Laurenz Wiskott. Slowness: An Objective for Spike-Timing–Dependent Plasticity? *PLoS Computational Biology*, 3(6):13, 2007.

James V Stone. Learning Perceptually Salient Visual Parameters Using Spatiotemporal Smoothness Constraints. *Neural Computation*, 8(7):1463–1492, October 1996. ISSN 0899-7667. doi: 10.1162/neco.1996.8.7.1463.

James V Stone and Alistair Bray. A learning rule for extracting spatio-temporal invariances. *Network Computation in Neural Systems*, 6(3):429–436, 1995. ISSN 0954898X.

Richard S Sutton and Andrew G Barto. An adaptive network that constructs and uses an internal model of its world. *Cognition and Brain Theory*, 4(3):217–246, 1981.

Michael J Tarr and Heinrich H Bülthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67(1-2):1–20, 1998.

Ian D Thompson and Tolhurst. Variation in the spatial frequency selectivity of neurones in the cat visual cortex. *The Journal of Physiology*, 295:33P, 1979.

Jon Touryan, Brian Lau, and Yang Dan. Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience*, 22(24):10811–10818, 2002.

Jon Touryan, Gidon Felsen, and Yang Dan. Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, 45(5):781–91, March 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.01.029.

Richard Turner and Maneesh Sahani. A Maximum-Likelihood Interpretation for Slow Feature Analysis. *Neural Computation*, 19(4):1022–38, March 2007. doi: 10.1162/neco.2007.19.4.1022.

David C van Essen and Charles H Anderson. Information Processing Strategies and Pathways in the Primate Visual System . *Knowledge Creation Diffusion Utilization*, 2nd: 45–76, 1995. ISSN 00219673.

Luc J Van Gool, Theo Moons, Eric J Pauwels, and André Oosterlinck. Vision and Lie's approach to invariance. *Image and Vision Computing*, 13(4):259–277, May 1995. ISSN 0262-8856. doi: 10.1016/0262-8856(95)99715-D.

J Hans van Hateren. Real and optimal neural images in early vision. *Nature*, 360(6399): 68–70, 1992.

J Hans van Hateren. Spatiotemporal contrast sensitivity of early vision. *Vision Research*, 33(2):257–67, 1993. ISSN 00426989.

J Hans van Hateren and Daniel L Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1412): 2315–20, December 1998. ISSN 0962-8452. doi: 10.1098/rspb.1998.0577.

J Hans van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):359–366, March 1998. ISSN 0962-8452.

R L Verma. Al-Hazen: father of modern optics. *Al-Arabi*, 8:12–13, August 1969.

Herrmann Von Helmholtz. *Handbuch der physiologischen Optik*, volume II. Voss, 1867.

Guy Wallis and Edmund T Rolls. A model of invariant object recognition in the visual system. *Progress in Neurobiology*, 51(11):167–194, 1997. ISSN 08997667.

Jimmy Wang, Jascha Sohl-Dickstein, and Bruno Olshausen. Unsupervised Learning of Lie Group Operators from Natural Movies. In *Front. Syst. Neurosci. Conference Abstract: Computational and Systems Neuroscience 2009*, volume 21080, 2009. doi: 10.3389/conf.neuro.06.2009.03.344.

Zhihong Wang and Eero P Simoncelli. Translation Insensitive Image Similarity in Complex Wavelet Domain. In *Acoustics, Speech and Signal Processing, IEEE Transactions on*, number March, pages 573–576, 2005. ISBN 0780388747.

Thomas P Weldon, William E Higgins, and Dennis F Dunn. Efficient Gabor filter design for texture segmentation. *Pattern Recognition*, 29(12):2005–2015, 1996.

Ben Willmore and David J Tolhurst. Characterizing the sparseness of neural codes. *Network*, 12(3):255–270, 2001.

Laurenz Wiskott. Estimating Driving Forces of Nonstationary Time Series with Slow Feature Analysis. *Arxiv preprint condmat0312317*, (December):8, 2003.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, April 2002. ISSN 0899-7667. doi: i:10.1162/089976602317318938</p>.

Laurenz Wiskott, Pietro Berkes, Mathias Franzius, Henning Sprekeler, and Niko Wilbert. Slow feature analysis. *Scholarpedia*, 6(4):5282, April 2011. ISSN 1941-6016. doi: 10.4249/scholarpedia.5282.

Rachel O L Wong. Retinal waves and visual system development. *Annual Review of Neuroscience*, 22(1):29–47, 1999.

Richard A Young. Orthogonal basis functions for form vision derived from eigenvector analysis. In *ARVO Abstracts*, page 22, Sarasota, FL, 1978.

Richard A Young. *The Gaussian derivative model for machine vision : visual cortex simulation*. General Motors Research Laboratories, Warren Mich., 1986.

Richard A Young and Ronald M Lesperance. The Gaussian derivative model for spatial-temporal vision: II. Cortical data. *Spatial vision*, 14(3-4):321–89, January 2001. ISSN 0169-1015.

Richard A Young, Ronald M Lesperance, and W Weston Meyer. The Gaussian derivative model for spatial-temporal vision: I. Cortical model. *Spatial vision*, 14(3-4):261–319, January 2001. ISSN 0169-1015.

Acknowledgements

This work was supported by the Max Planck Society and the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601) and the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307).

There will come a time when you believe everything is finished; that will be the beginning.

Louis L'Amour