Visual Receptive Fields

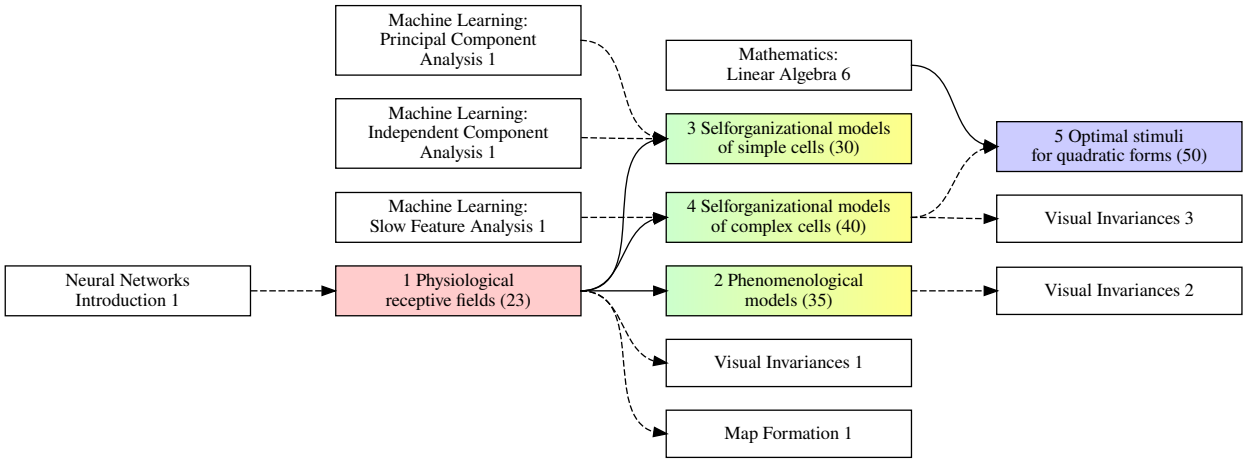
— Lecture Notes —

Laurenz Wiskott
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany, EU

27 March 2021

— Summary —

Receptive fields are a conceptualization of the first step of perceptual information processing in, e.g., the visual, auditory, or somatosensory system. The sonsory input is encoded in a way that makes further processing easier, e.g. by reducing dimensionality or by feature extraction. In the visual system, receptive fields are particularly sensitive to local spots of light, to bars or edges, to direction of motion etc. There are different levels at which one can analyze, model, and understand receptive fields. Abstract mathematical descriptions serve as phenomenological models of what receptive fields do. More detailed biologically motivated models can be used to investigate how receptive fields work. With objective function models one can test hypotheses about the purpose of the receptive fields. And self-organizing models can illustrate how receptive fields develop. Often, models combine more than one of these aspects.



If applicable, core text and formulas are set in dark red, one can repeat the lecture notes quickly by just reading these; ♦ marks important formulas or items worth remembering and learning for an exam; ◊ marks less important formulas or items that I would usually also present in a lecture; + marks sections that I would usually skip in a lecture.
You can also download the teaching material of this topic as zip files and then view them locally on your computer.

— **Lecture 1/2** — → Lecture 1 Exercises, Lecture 1 Solutions

1 Physiological receptive fields (→ slides) in the visual system include center surround receptive fields sensitive to a bright/dark spot on a dark/bright background, simple cells that are sensitive to a bar or grating at a particular location, and complex cells that are also sensitive to bars and gratings but much less to location. Center surround receptive fields offer an explanation for the so-called Hermann grid illusion. Complex cells as compared to simple cells already realize some translation invariance, one prominent ability of the visual system as a whole.

2 Phenomenological models (→ slides) provide a mathematical formalization of the computation of receptive fields. Center surround and simple cells can be modeled by linear filters. For the latter we use a Gabor-wavelet, i.e. a (co)sine wave windowed by a Gaussian envelope function, to achieve the sensitivity to edges and gratings. Quadratically combining a sine and a cosine Gabor-wavelet is a minimal model achieving the invariance of complex cells, much like in the $\sin(\phi)^2 + \cos(\phi)^2 = 1$ rule.
→ Section 2 Exercises, Section 2 Solutions

3 Selforganizational models of simple cells (→ slides) are usually linear and based on various objectives, which yield certain receptive fields as optimal solutions. If these receptive fields are similar to the physiological ones, this provides support for the objective used to be the underlying reason for the shape of the physiological receptive fields. This section shows that sparseness and statistical independence are plausible objectives, while linear compression is not.

— **Lecture 2/2** — → Lecture 2 Exercises, Lecture 2 Solutions

4 Selforganizational model of complex cells (→ slides) is nonlinear (quadratic form), since otherwise the translation invariance could not be achieve. This model is based on the objective that the output should vary slowly over time. It self-organizes from image sequences receptive fields that share many properties with complex cells.

5 Optimal stimuli for quadratic forms can be derived with methods known from linear algebra, in particular the *trust region problem*. Quadratic forms are often used to model complex cells, and being able to derive optimal stimuli helps in comparing the results with physiological experiments.
→ Section 5 Exercises, Section 5 Solutions

Contents

LECTURE 1/2

- Lecture 1 Exercises, Lecture 1 Solutions

1 Physiological receptive fields (→ slides)

Learning material:[1]

☐ 26 min video ~~1 Physiological receptive fields~~

☐ Text below

[1]Generic instruction: Consider the (possibly nested) list of resources like a horizontal tree with an invisible root on the very left, and decide from left to right what you want to select to work through. The invisible root node has to be selected. For any selected parent node all children nodes marked with ■ or ● are mandatory and have to be selected. Children nodes marked with □ or ○ are optional and may be selected in addition to get a better understanding of the material. If a parent node has no mandatory child, then at least one optional child has to be selected. Children marked with + provide additional voluntary material that can be safely ignored, typically going beyond the scope of the section. Children of non-selected parents may be ignored. ■ and □ indicate children that cover (almost) the whole material of the section. Missing content might then be indicated by struck through references to the corresponding learning objectives. Items tend to be ordered by precedence and/or recommended temporal order from top to bottom, assuming that you prefer to first watch a video before reading through lecture notes. If a detailed table of content for videos or lecture notes is given, references to learning objectives might be provided in green, 1:30 should be read as 1 min and 30 seconds, and 1'30 should be read as page 1 at about 30% of the page. Video times may be linked directly to the indicated position in the video, but be aware that the video might be downloaded anew each time you click on a time. Resources without author name are usually authored by Laurenz Wiskott and his team.

Let us first consider some basic properties of physiological receptive fields, mainly in the visual system-

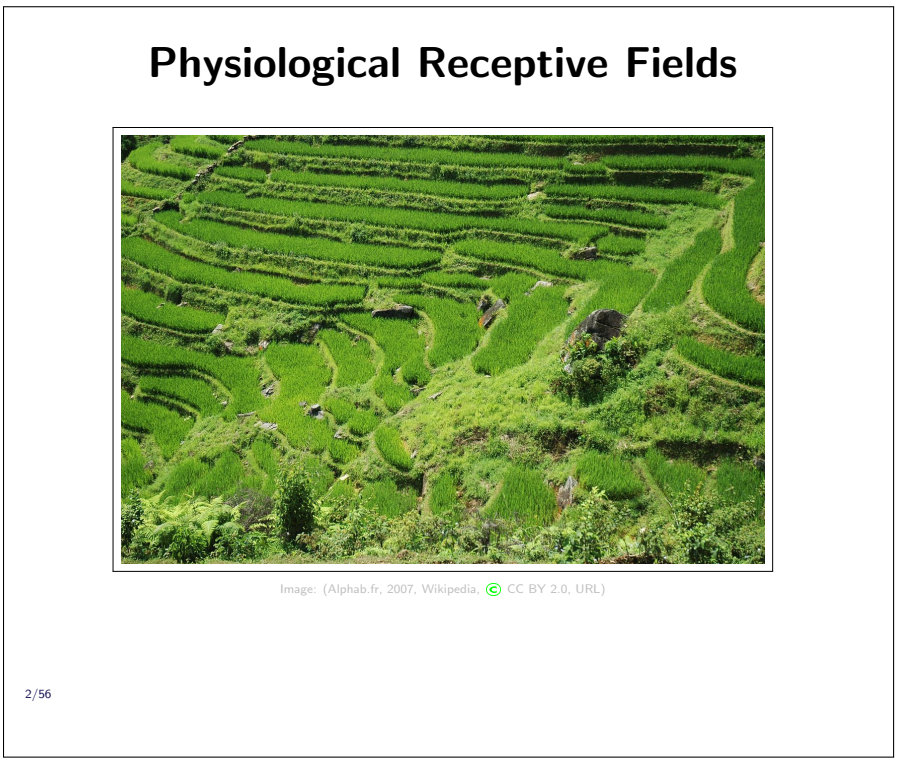
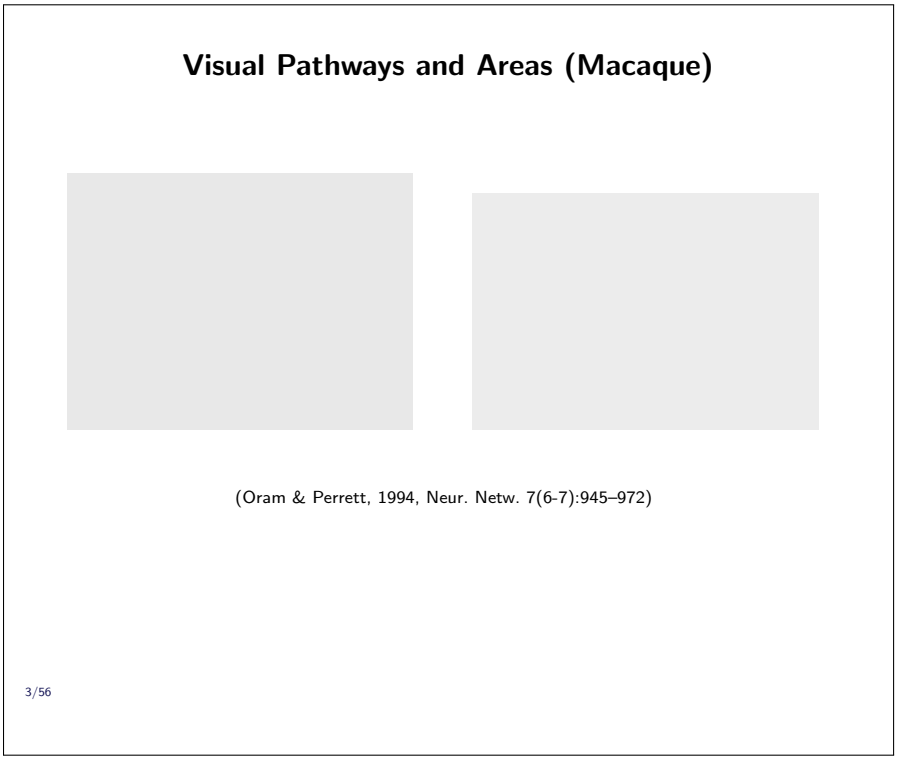


Image: (Alphab.fr, 2007, Wikipedia, © CC BY 2.0, URL)[1.1]



Visual input from the retina gets projected through the lateral geniculate nucleus (LGN, not shown here), which is subcortical, to the primary visual cortex (V1, 'V' and '1' indicating 'visual' and 'primary' respectively). From there it goes through V2 and V4 to the inferior temporal cortex (IT), which can be further subdivided into posterior (PIT), central (CIT), and anterior (AIT) IT. IT is thought to be instrumental for object recognition, while V1–V4 extract more elementary features. This path is referred to as the *what*-path, because it tells us what we see.
Another path goes through V2 and V4 to areas MT/MST, which are particularly responsive to motion. This path has been termed the *where*- or *how*-path, because it is thought to tell us where the objects are or how we can handle them, e.g. grasp them. The paths converge in the posterior (STPp) and anterior (STPa) superior temporal polysensory area. Cells in STPa, for instance, have been found to be sensitive to body motion, such as walking. Figure by Oram and Perrett (1994).

The Visual System

(Wiskott, 2003, CogPrints 3321; after Oram & Perrett, 1994, J. Cogn. Neurosci., 6(2):99–116)

4/56

This schematic drawing (Wiskott, 2003) highlights some organizational principles of the visual system. It is hierarchically structured in areas (listed on the left; in models usually referred to as layers, not to be confused with the layers of cortex), which are coupled by feedforward (gray upward arrows) as well as feedback (gray downward arrows) connections with some shortcut connections that skip an area. Processing in each area takes about 10ms (latencies are shown on the left). Along the hierarchy the receptive field sizes increase (indicated by the triangles in the middle), the feature complexity increases (indicated by some typical stimuli on the right to which a neuron might respond or not), and the invariance, e.g. to shift (or translation), scaling, and rotation, increases.

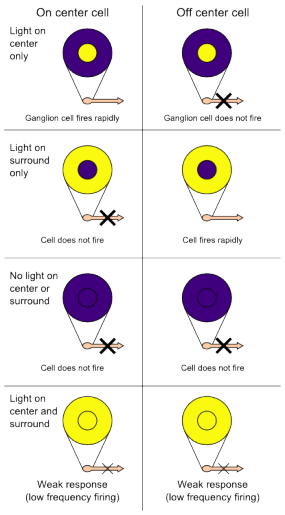
Measuring Receptive Fields

(http://www.medinfo.ufl.edu/... 2001-10-29 (outdated))

5/56

It is possible to make quite detailed measurements of response properties of single cells in awake or anaesthetized animals. To measure visual receptive fields, one typically places an animal in front of a computer monitor, let the animal fixate the center of the screen, presents visual stimuli, and simultaneously records extracellularly from individual neurons. Visually driven neurons usually respond only to stimuli within a particular region, which is referred to as the receptive field. They also only respond to particular shapes or features, such as orientation, color, or motion. One says the cell has a tuning for one or several of these features. One also speaks of such cells as feature detectors.

Center-Surround Cells in Retina and LGN

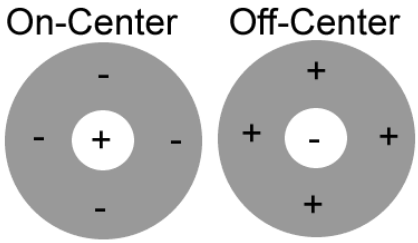


(Delldot/Xoneca, 2005/2008, Wikimedia, © CC0, URL)

6/56

Cells in retina and LGN (lateral geniculate nucleus, which is a relay station between retina and cortex, have center-surround receptive fields. Some of them respond best to a bright spot on a dark background (on-center cell/stimulus), others to a dark spot on a bright background (off-center cell/stimulus). They do not respond well to full field stimuli (dark or bright). Interestingly, an on-center cell gives a response if an off-center stimulus disappears (release-of-inhibition reponse) and the other way around.

Assumed Connectivity



(Paskari, 2007, Wikipedia, © CC0, URL)

7/56

Center-surround receptive fields can be set up easily by a corresponding feedforward connectivity. For an on-center cell, connections coming from the center of the recptive field would be excitatory and those coming from the surround would be inhibitory. For off-center cell it would be the other way around. A canonical way of plotting such a receptive field is to plot the excitatory and inhibitory regions in the visual field (see lower left). Such receptive fields are conceptually linear.

Hermann Grid

(http://www-psych.stanford.edu/ 2001-10-29 (outdated))

8/56

A Hermann grid is a white square grid of appropriate size on a black background. If you look at it you might notice that the white looks darkened somehow at the crosses, but only in the periphery and not at the point of fixation. This is an optical illusion that can be explained with the center-surround receptive fields in the retina or LGN.
(http://www-psych.stanford.edu/ 2001-10-29 (outdated))

Hermann Grid

$8(1 + 4(-1/16)) + 4(4(-1/16)) = 5$

$8(1 + 2(-1/16)) + 4(6(-1/16)) = 5.5$

(http://www-psych.stanford.edu/ 2001-10-29 (outdated))

9/56

To explain the Hermann grid illusion, place a simple center-surround receptive field at a cross and at a line. By adding the product of the image gray value with the receptive field weight one gets a somewhat lower response at a cross (value 5) than a line (value 5 1/2) due to the stronger surround inhibition. This effect depends on the width of the stripes compared to the size of the receptive fields. In the fovea, i.e. around the point of fixation, the receptive fields are very small and the Hermann grid illusion cannot be observed with a coarse grid.
(http://www-psych.stanford.edu/ 2001-10-29 (outdated))

Simple Cells in Primary Visual Cortex (V1)

(Hubel, 1989, Auge und Gehirn: Neurobiologie des Sehens, Fig. 4.10)

10/56

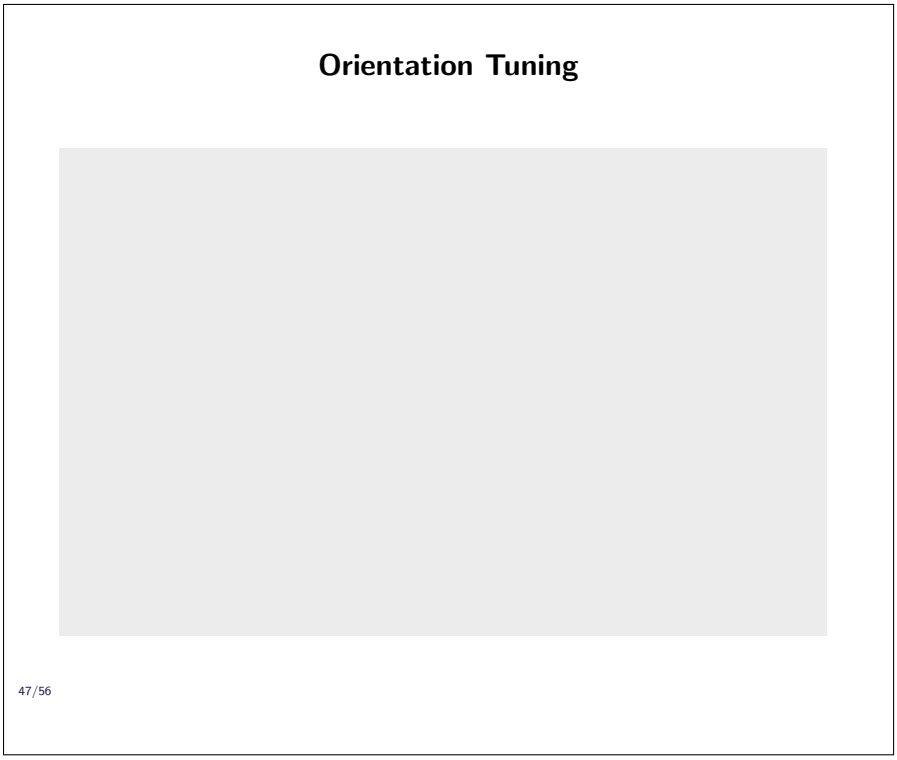
In the first cortical area dedicated to visual processing, referred to as primary visual cortex or, for short, V1, one mainly distinguishes between two types of cells based on their receptive fields: simple cells and complex cells. Both cell types prefer oriented stimuli, such as bars and stripes, but simple cells care about the exact location of the stimuli while complex cells don't. Thus, in some sense complex cells have a higher degree of invariance than simple cells.
(Hubel, 1989, Auge und Gehirn: Neurobiologie des Sehens, Fig. 4.10)

Complex Cells in Primary Visual Cortex (V1)

Figure: (Hubel, 1989, Auge und Gehirn: Neurobiologie des Sehens, Fig. 4.13)

11/56

In the first cortical area dedicated to visual processing, referred to as primary visual cortex or, for short, V1, one mainly distinguishes between two types of cells based on their receptive fields: simple cells and complex cells. Both cell types prefer oriented stimuli, such as bars and stripes, but simple cells care about the exact location of the stimuli while complex cells don't. Thus, in some sense complex cells have a higher degree of invariance than simple cells.
Figure: (Hubel, 1989, Auge und Gehirn: Neurobiologie des Sehens, Fig. 4.13)

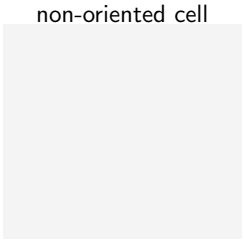


Simple and complex cells usually have preferences for certain orientations. This can be measured by presenting gratings of different orientation to the cell (or rather the animal) and recording the corresponding neural responses. Normally, drifting gratings are used, because the cells respond stronger to moving stimuli.

The responses to different orientations can be conveniently visualized in a polar plot. One simply plots the firing rate in radial direction as a function of orientation in azimuthal direction. The graph shows a standard orientation tuning with one preferred orientation at about 160°, which appears here as two lobes in 180° distance due to the two different drifting directions. Since the two lobes have same size, the cell does not have a preference for a particular drifting direction.

Orientation Tuning

non-oriented cell



(De Valois et al., 1982)

standard orientation tuning

(De Valois et al., 1982)

secondary response-lobes

(De Valois et al., 1982)

direction selectivity

(De Valois et al., 1982)

13/56

Complex cells have a great variety of different orientation tunings. In standard tuning the cell responds well to one orientatoin regardless of the direction of drifting of the grating (upper right panel), giving rise to two large response lobes in the polar plot. Some cells have additional secondary response lobes at a different orientation (lower left panel). This means that theses cells respond well to two different orientations. Other cells don't care about orientation at all (upper left panel), although they are selective for oriented stimuli, such as gratings. Some cells are direction selective (lower right panel), meaning that they respond to a particular orientation only if the grating drifts in the right direction.

End- and Side-Inhibition

14/56

A somewhat peculiar behavior of some complex cells is end- and side-inhibition. Such cells do not respond maximally if the whole receptive field is filled with a grating but only if part of the receptive field is filled. If one slowly moves a grating into such a receptive field from the preferred side, the response first increases but then drops again. End-inhibition means that the grid has to be moved into the receptive field with the ends of the stripes first; side-inhibition means the the grid has to be moved sideways into the receptive field.
One way to interpret this behavior is to assume that the receptive field consists of an excitatory subfield and an inhibitory subfield of same preferred orientation. Only only the excitatory subfield is stimulated with a grating, the response is higher than if both the excitatory and the inhibitory subfield are stimulated.

Cell Tuning in Inferior Temporal Cortex (IT)

(Sato, Uchida et al, 2013, J. Neuroscience, Fig. 5, URL)

15/56

In higher areas, such as inferior temporal cortex, it may be a bit of a stretch to speak of receptive fields, but cells still show a clear preference for certain objects or features. One therefore also speaks of these cells as feature detectors.
In this figure red circles, blue triangles, and green squares represent face, monkey body, and animal body categories, respectively. The ordinate indicates the evoked response of a cell recorded from, the abscissa indicates the rank of an image in terms of the evoked reponse. The top five images in each graph indicate the first rank stimuli, the bottom five images the last rank stimuli.
Figure: (Sato et al., 2013, Fig. 5, URL)[1.2]

Somatosensory Cells

(http://zeus.rutgers.edu/∼ikovacs/SandP/c_fig7.jpg 2001-10-30 (outdated))

16/56

The concept of a receptive field is also used in other modalities. A cell in the somatosensoric areas might, for instance, respond only to tactile stimuli in a certain region of the palm and might even have an orientation preference, like the cell shown here, which prefers horizontal stimuli.

Physiological Receptive Fields - Summary

- Retina and LGN have center-surround receptive fields.
- V1 has two major types of receptive fields, those of simple cells and complex cells.
- Both types of cells respond well to bars or gratings of a particular orientation and frequency.
- Simple cells are sensitive to the exact location of the bar or grating, complex cells are not.
- Higher areas show more complex receptive fields that may even be tuned to specific objects.

17/56

2 Phenomenological models (→ slides)

Learning material:

- ☐
  - ○ 6 min video Introducing Convolutions: Intuition + Convolution Theorem (by Faculty of Khan on YouTube) [00:12-06:35]
  - ● 31 min video 2 Phenomenological models
- ☐ Text below

Models of Visual Receptive Fields

Figure: (http://www.uwec.edu/geography/Ivogeler/w111/Images/srilankatopsheet.jpg 2005-11-16 (outdated))

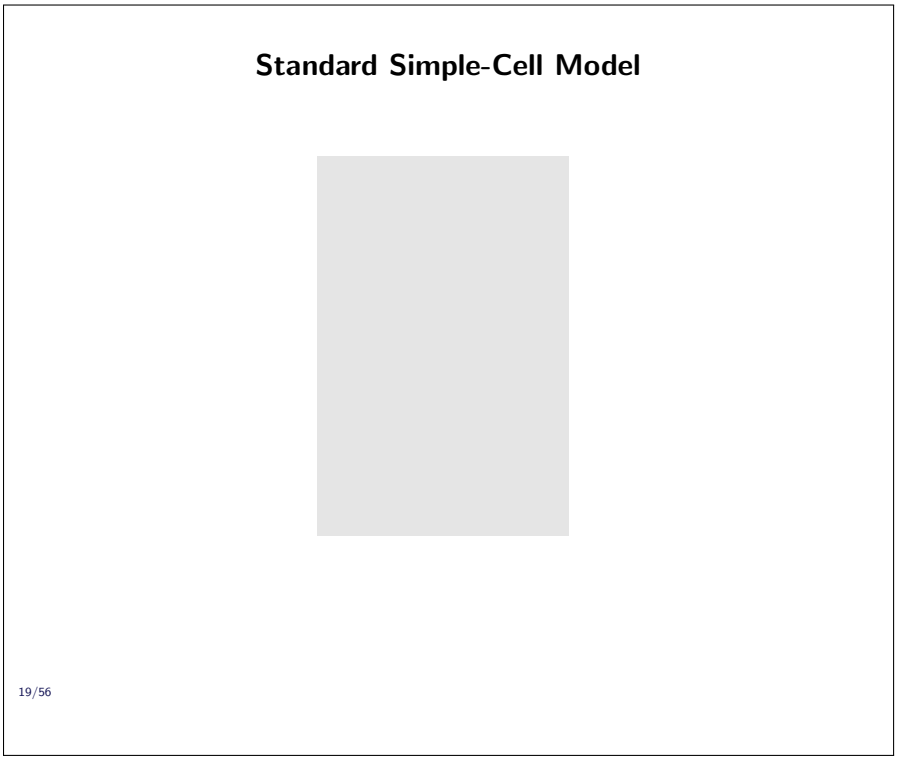
18/56

Figure: (http://www.uwec.edu/geography/Ivogeler/w111/Images/srilankatopsheet.jpg 2005-11-16 (outdated))

There are several levels at which one can model receptive fields (or any other neural subsystem). One can ask:

- What does it do?
- How does it do it?
- Why does it do it?

The what-question tries to get at the phenomenology of the receptive field. It yields descriptive models of the function. We consider this in this section. The how-question gets at the mechanistic realization of the function. We are not going to consider this at all. The why-question finally gets at the role the receptive fields play in the context of the whole brain. It is closely linked to the question of what would be optimal to do in this area. We consider this in Section 3.



The standard model of a simple cell is simply a linear filter having the shape of a wavelet. The response is the inner product $\boldsymbol{w}^T\boldsymbol{x}$ (sum over pointwise products) between the filter (weight vector $\boldsymbol{w}$) and the image (input vector $\boldsymbol{x}$). Such a filter is strongly excited by a bar or grating of the correct frequency (in case of a grating), orientation, and exact position. If the grating is shifted in phase by 180°, or in position by one wavelength orthogonal to the wave fronts, the model unit gives a strong negative response.

Standard Complex-Cell Model

20/56

The standard model for a complex cell is the so-called quadrature filter pair model. The response of two standard simple cell models are squared and added. The filters of the two simple cells form a so called quadrature filter pair, in this case two wavelets that differ only by a slight shift of the stripes by half a stripe width. Their relationship is therefore similar to that of sin and cos, for which $\sin(\phi)^2 + \cos(\phi)^2 = 1$ holds, which implies that the square sum is invariant to a change of ϕ. Similarly, the response of the standard complex cell model is approximately invariant to a shift of stimulus. This invariance is the defining property of an ideal complex cell.

Gabor Wavelets

Gabor wavelets (with DC-correction) are defined as

$$\psi_j(\boldsymbol{x}) := \frac{k_j^2}{\sigma^2} \exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\left(\exp(\mathrm{i}\boldsymbol{k}_j^T\boldsymbol{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right),$$

with wave vectors $\boldsymbol{k}_j$ having different orientations and different frequencies.

Gabor wavelets fulfill the uncertainty relationship exactly.

22/56

Gabor wavelets are often used for image processing and to model simple and complex cells. They are localized in space and frequency, and they actually do that as precise as theoretically possible, i.e. they fulfill Heisenberg's uncertainty relationship exactly (side note for physicists and electrical engineers). A Gabor wavelet is essentially the product of a Gaussian (black solid line) with a (co)sine wave and could therefore be written in its simplest one-dimensional form as $\exp(-x^2)\sin(x)$ (green dashed line) or $\exp(-x^2)\cos(x)$ (blue solid line); together these form a quadrature filter pair.
The equation given on the slide is more complicated and simpler in some aspects for several reasons. This is not essential for the lecture, but for the technically interested reader I explain the differences.

- The $\sin(x)$ and $\cos(x)$ wavelets are combined into one complex wavelet with $\exp(\mathrm{i}x) = \cos(x) + \mathrm{i}\sin(x)$ (second exponential in the equation). This makes in particular the convolution more efficient. Since a convolution is always complex, the second convolution in the imaginary part comes for free.
- The simple x in $\exp(\mathrm{i}x)$ is multiplied by a wave number k_j to allow chosing a spatial frequence different from 1, and index j allows to chose different wave numbers for different cells, yielding $\exp(\mathrm{i}k_j x)$
- In two (or higher) dimensions a wave not only has a frequency but also a direction, thus k_j becomes a two (or higher) dimensional vector and the product $k_j x$ an inner product, yielding $\exp(\mathrm{i}\boldsymbol{k}_j^T\boldsymbol{x})$. (Please note the difference between $\boldsymbol{x}$ representing an image, in which case it might be a 10000-dimensional vector for a 100×100-pixel image, and $\boldsymbol{x}$ representing space, in which case it is just two-dimensional for an image. Here we use the latter version.)
- It is common to add a parameter σ to the Gaussian $\exp(-x^2)$ to control its width, yielding $\exp(-\frac{x^2}{2\sigma^2})$.
- The additional factor k_j^2 in $\exp(-\frac{k_j^2 x^2}{2\sigma^2})$ scales the Gaussian such that all Gabor wavelets look alike, no matter what frequency they have. This is referred to as *self-similarity* of the family of Gabor wavelets with constant σ.
- The term $-\exp(-\frac{\sigma^2}{2})$ at the end pulls the cosine wavelet a bit down in the center to make it really DC-free (DC stands for *direct current* here), i.e. the integral over the whole filter is zero. This is guaranteed for symmetry reasons for the sine filter, but for the cosine filter it must be taken care of explicitely. The filter being DC-free has the advantage that the response of the modeled simple or complex cell does not depend on overall brightness of the image, which is a simple form of visual invariance.
- The prefactor $\frac{k_j^2}{\sigma^2}$ finally scales the Gabor wavelets such that the average magnitude of the responses of the convolution on natural images are more balanced for different k and σ.

Gabor Wavelets

Convolution

$$J_j(\boldsymbol{x}) = a_j(\boldsymbol{x}) \exp(i\phi_j(\boldsymbol{x})) = \int I(\boldsymbol{x}')\psi_j(\boldsymbol{x} - \boldsymbol{x}')\mathrm{d}^2\boldsymbol{x}' \,,$$

with smoothly varying amplitudes

$$a_j = \sqrt{\Re(J_j)^2 + \Im(J_j)^2}$$

(in analogy to the rule $\sqrt{(a\cos(\phi))^2 + (a\sin(\phi))^2} = a$) and quickly varying phase ϕ_j.

23/56

Phase is not globally consistent.

Applying a linear filter at all locations of an image is mathematically a cross-correlation. If one mirrors the filter at its origine and applies that to all locations of an image, that is a convolution. The latter is less intuitive but has some nice mathematical properties, e.g. it is symmetric and it corresponds to a multiplication in Fourier space. Thus, we use convolution and just have to keep in mind that the filter is flipped.

The first equation defines a convolution of an image $I(\boldsymbol{x}')$ with a complex filter (or kernel) $\psi_j(\boldsymbol{x}')$. Since the filter is complex, i.e. it has a real and an imaginary part, the convolution result is complex as well. In our case the real and imaginary part represent the cos- and sin-Gabor wavelet, respectively.

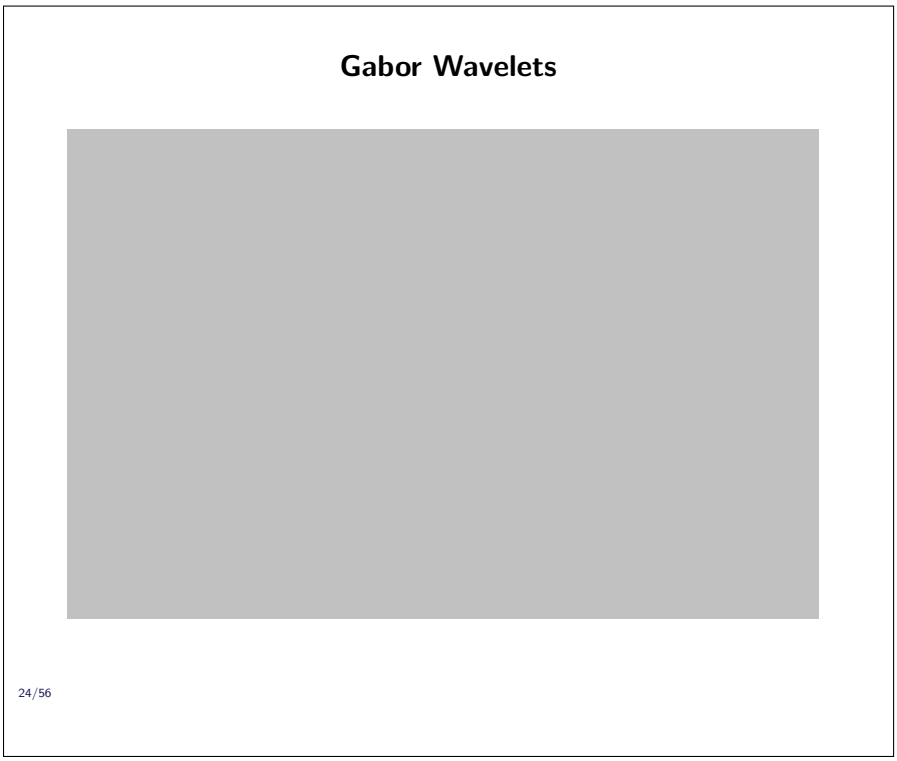
To get an intuitive understanding of the convolution it might help to start with the expression $\int I(\boldsymbol{x}')\psi_j(\boldsymbol{x}')\mathrm{d}^2\boldsymbol{x}'$, which simply means you put the image and the filter on top of each other, do a pointwise multiplication, and then integrate the result, which gives you one scalar value. The operation actually is an inner product if you consider image and filter as vectors. It is large if image and filter are similar (point in the same direction) and close to zero, if they are unrelated (orthogonal to each other). If the filter is a Gabor wavelet, this scalar models a simple-cell response at the origin of the image.

To get the response at another location $\boldsymbol{x}$, one needs to shift the filter, which can be done by subtracting a constant from the argument, yielding $\int I(\boldsymbol{x}')\psi_j(\boldsymbol{x}' - \boldsymbol{x})\mathrm{d}^2\boldsymbol{x}'$, which is a cross-correlation. This then models a simple-cell response at location $\boldsymbol{x}$. If $\boldsymbol{x}$ is considered a variable rather than a constant, the cross-correlation becomes a cross-correlation function in $\boldsymbol{x}$.

Now the final step is to simply mirror the filter at its origin, so that the sign of its argument changes, yielding $\int I(\boldsymbol{x}')\psi_j(\boldsymbol{x} - \boldsymbol{x}')\mathrm{d}^2\boldsymbol{x}'$, which is a convolution.

A complex number can always be written in terms amplitude a and phase ϕ rather than real and imaginary part, yielding $a_j(\boldsymbol{x}) \exp(i\phi_j(\boldsymbol{x}))$.

The plot shows the response of a hypothetical cos- and sin-Gabor wavelet in blue and green, respectively, and the resulting amplitude in black. The phase changes approximately with the same spatial frequency as that of the Gabor wavelet but not exactly, so that phase is not globally consistent with the frequency of the filter.



This figure illustrates the response of a standard simple or complex cell model. White in the first and fourth column indicate zero, black some maximum value. Grey in the second and third column indicate zero, black a negative and white a positive value.
If one applies a simple cell model with a wavelet like weight vector (second column) to all locations of an input image (first column), one gets a response distribution as shown in the third column. The operation is mathematically a convolution. Please realize that the convolution result represents the activity of many identical simple cells at different locations, one for each pixel. There are several things to note here:

- The simple cell response oscillates with the spatial frequency (and orientation) of its wavelet filter. This results from the shift of the filter relative to the image and the wavelet structure of the filter.
- The magnitude of the oscillating response is largest at sharp edges of the correct orientation. It would actually be even larger for gratings of the right frequency and orientation, but these are rare (and not present in this image). Since sharp edges contain all frequencies, they are very good stimuli for all filters.

For the standard complex cell model, one needs a quadrature filter pair, so imagine the filters in the second column complemented by a partner with slightly shifted stripes. Using that filter alone would yield very similar responses like those shown in the third column, just shifted by half a stripe width. If one squares the two responses (the one shown and the one complemented) and adds them, one gets the comlex cell responses shown in the fourth column. They are non-negative and do not oscillate, and they are strong at sharp edges of the correct orientation again.

3 Selforganizational models of simple cells ($\rightarrow$ slides)

While the what-question can be addressed rather directly, because one can measure the responses of the cells and make a model of the responses as a function of the stimulus, the why-question is more difficult and needs to be adressed indirectly. One approach is to make a hypothesis about why the cells have developed their response properties, formulate that as an optimization problem, solve the optimization problem, and then see whether the result bares similarity with the physiological response properties. If it does, it supports the hypothesis, if not, it discredits the hypothesis. For simple cells we consider here the following three hypotheses:

- Simple cells are there to (linearly) compress the visual input.
- Simple cells are there to decompose the visual input into statistically independent components.
- Simple cells are there to yield a sparse representation of the visual input.

3.1 Principal component analysis does not lead to simple cells

Learning material:

☐ 6 min video ~~3.1 Principal Component Analysis does not Lead to Simple Cells~~

☐ Text below

A common way to linearly compress data is *principal component anaylsis (PCA)* (D: Hauptkomponentenanalyse) (see Wiskott, 2016b, for an introduction). The data is considered as points in a vector space, and PCA finds an ordered set of orthogonal directions, called principal components (PC) (D: Hauptkomponenten), such that the variance of the data along the first PC (or projected onto the first PC) is maximal, along the second PC it is maximal under the constraint of being uncorrelated to the first one, along the third PC it is maximal under the constraint of being uncorrelated to the first and second one, ect. For optimal linear compression one keeps the first few principal components and discards the other. How many PCs to keep depends on several factors such as how much compression one needs and how much variance there is along the individual PCs.

Principal Components of Natural Images

(Hancock, Baddeley, & Smith, 1992, Network 3(1):61–70, Fig. 1)

15 natural images of size 256×256 pixels.

20,000 random samples of size 64×64 pixels.

For each pixel the mean gray value over the 20,000 samples was removed.

The samples were windowed with a Gaussian with std. dev. 10 pixels.

Sanger's rule was applied to the samples.

25/56

An early hypothsis was that the purpose of simple cells is to compress images for further processing. Hancock et al. (1992) have tested this by taking 15 natural images of size 256×256 pixels, from these cutting out 20,000 random samples of size 64×64 pixels, removing the mean from each pixel across the 20,000 samples, windowing the samples with a Gaussian, and finally calculating the first principal components with Sanger's rule, which is a neural learning rule for performing PCA. The image patches of size 64×64 are cast into vectors by simply concatenating the rows (or columns) into a vector of length 4096, a transformation that can be easily inverted by rearranging the components of the vector back into a matrix.

Principal Components of Natural Images

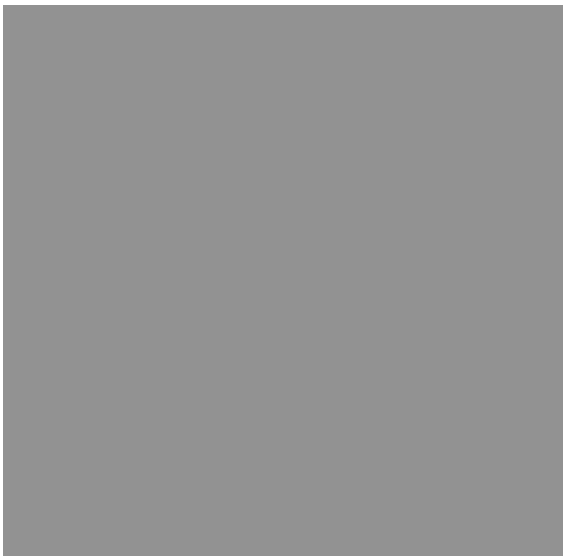
(Hancock, Baddeley, & Smith, 1992, Network 3(1):61–70, Fig. 1)

The first principal components resemble simple-cell receptive fields in the primary visual cortex, the later ones do not.

26/56

The first principal components extracted from natural image patches windowed with a Gaussian somewhat resemble simple cells (Hancock et al., 1992). However, later ones do not, and the Gaussian window plays an important role in making the filters look plausible at all.

Principal Components of Natural Images



(Olshausen & Field, 1996, Nature 381:607–9, Fig. 1)

27/56

Olshausen and Field (1996) have applied principal component analysis (PCA) to natural image patches of size 8×8 and have found filters as shown here.

One can understand this result, if one resorts to Fourier theory and considers the image patches as a linear superposition of sine waves of different frequency, orientation, and phase. Since natural images are known to have a $1/f^2$ power spectrum, i.e. low frequencies f are stronger and thus carry more variance, it is clear that the early principal components (PCs) should focus on low frequencies and the later ones on high frequencies. If one furthermore assumes that the statistics of natural images is translation and rotation invariant (which is at least approximately true), one can see that sine waves of different phase (related by translation) and orientation (related by rotation) but same frequency can be randomly mixed, since they carry identical variance. Taking this together yields the PCs shown here.

3.2 Sparseness leads to simple cells

Learning material:

- ☐ 14 min video 3.2 Sparseness Leads to Simple Cells
- ☐ Text below

Sparseness Principle

(Olshausen & Field, 2004, Curr. Opp. Neurobiol 14:481)

A sparse representation

- can reduce metabolic costs, because fewer units are active,
- can reduce wiring, because fewer units need to be connected,
- can be more robust, because units tend to be more binary,
- can simplify learning and processing, because relevant information is more localized,
- ...

28/56

Olshausen and Field (1996) have argued that the goal of sensory coding is to yield a *sparse* (D: spärliche(?)) representation. A sparse representation is one, where for any given input only few units are strongly active, all others are close to zero. This code might have various advantages for the brain.
The figure (Olshausen and Field, 2004) shows a non-sparse representation at the top and a sparse representation at the bottom.

Sparse Coding

Assumption: Images can be written as a superposition of basis functions,

$$I(\mathbf{x}) = \sum_i a_i \phi_i(\mathbf{x}) \,, \tag{1}$$

with fixed functions $\phi_i(\mathbf{x})$ and variable coefficients a_i.

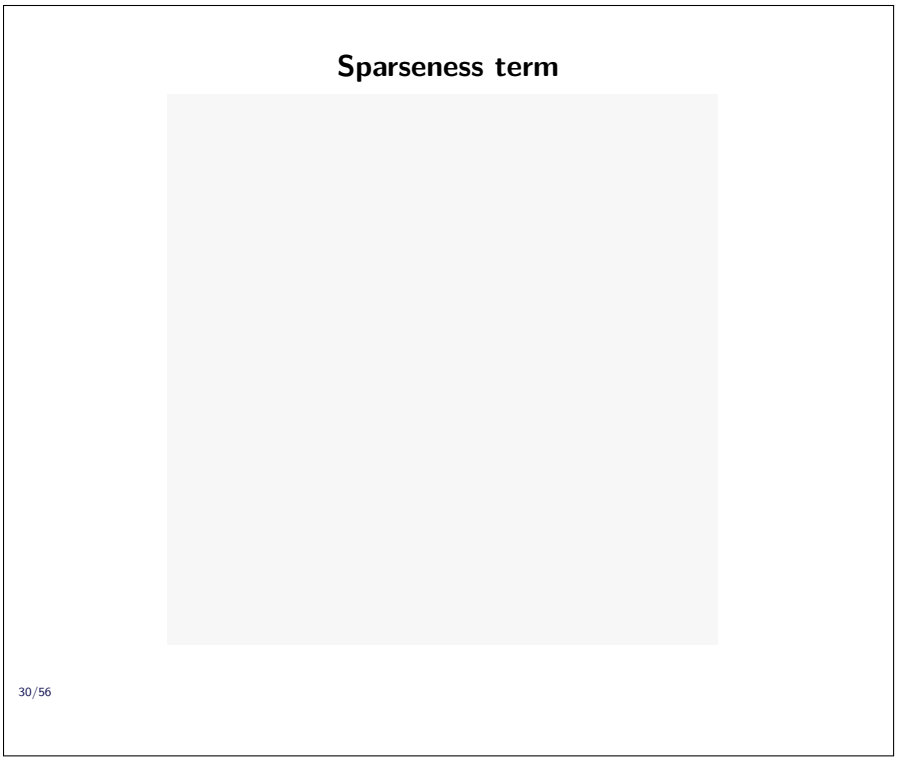
Objective: Choose the (probably normalized) functions such that the reconstruction error is small and the distribution of coefficients sparse, i.e.

$$\text{minimize} \quad E := \underbrace{\int_{\mathbf{x}} (I(\mathbf{x}) - \sum_i a_i \phi_i(\mathbf{x}))^2 \, \mathrm{d}^2\mathbf{x}}_{\text{reconstruction term}} + \lambda \underbrace{\sum_i |a_i|}_{\text{sparseness term}} \,. \tag{2}$$

(Olshausen & Field, 1996, Nature 381:607–9)

29/56

The model by Olshausen and Field (1996) assumes that images $I(\mathbf{x})$ can be represented by a linear superposition of some fixed basis functions $\phi_i(\mathbf{x})$, wich leads to the first term in the cost function E. The basis functions may be overcomplete, i.e. there may be more functions than pixels in the image, and non-orthogonal, which they must be in cast of an overcomplete set.
The weighting coefficients a_i vary from image to image and should be sparsely distributed, i.e. should be near zero most of the time and only occasionally have a large positive or negative value. The second term in the cost function E formalizes the sparseness objective.
An optimization procedure optimizes both, the basis functions across all images as well as the weighting coefficients for each image individually.



Consider the case where we want to represent a vector $\boldsymbol{I}$ (the image) as a linear combination of some basis vectors $\boldsymbol{\phi}_i$ with weighting factors a_i, i.e. $\boldsymbol{I} = \sum_i \boldsymbol{\phi}_i a_i$. If we combine the basis vectors in a matrix $\boldsymbol{\Phi} := (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_N)$ and the weighting factors in a vector $\boldsymbol{a} := (a_1, ..., a_N)^T$, we can write $\boldsymbol{I} = \boldsymbol{\Phi}\boldsymbol{a}$. With any orthogonal (rotation) matrix $\boldsymbol{U}$ we can define a new $\boldsymbol{\Phi}' := \boldsymbol{\Phi}\boldsymbol{U}^T$ and $\boldsymbol{a}' := \boldsymbol{U}\boldsymbol{a}$, so that the image is preserved,

$$\boldsymbol{\Phi}'\boldsymbol{a}' = \boldsymbol{\Phi}\underbrace{\boldsymbol{U}^T\boldsymbol{U}}_{=\mathbf{1}}\boldsymbol{a} = \boldsymbol{I}\,,$$

but the basis vectors as well as the weighting factors change. Thus, we can rotate the representation without compromising the qualitiy of the represented image, which leaves room to optimize sparseness in addition.

The figure illustrates with a dashed circle all the weight vectors with length 3, which can be realized by rotating one weight vector of same length. The solid lines represent the level lines of the sparseness term $|a_1| + |a_2|$. One can see that on the circle the points on the axes, namely $(0, 3), (3, 0), (0, -3), (-3, 0)$, have the smallest value for the sparseness term, which corresponds to the intuition that the coefficients should be either close to zero or large.

In the model (Olshausen and Field, 1996), the solution is not as clean, since the code is optimized for many images simultaneously. Also, some normalization must be imposed on either the basis functions or the weighting factors, because otherwise the latter could be made arbitrarily small while the former grow larger and larger.

Filters Generating a Sparse Code of Natural Images

(Olshausen & Field, 2004, Curr. Opp. Neurobiol 14:481, Fig. 1a)

31/56

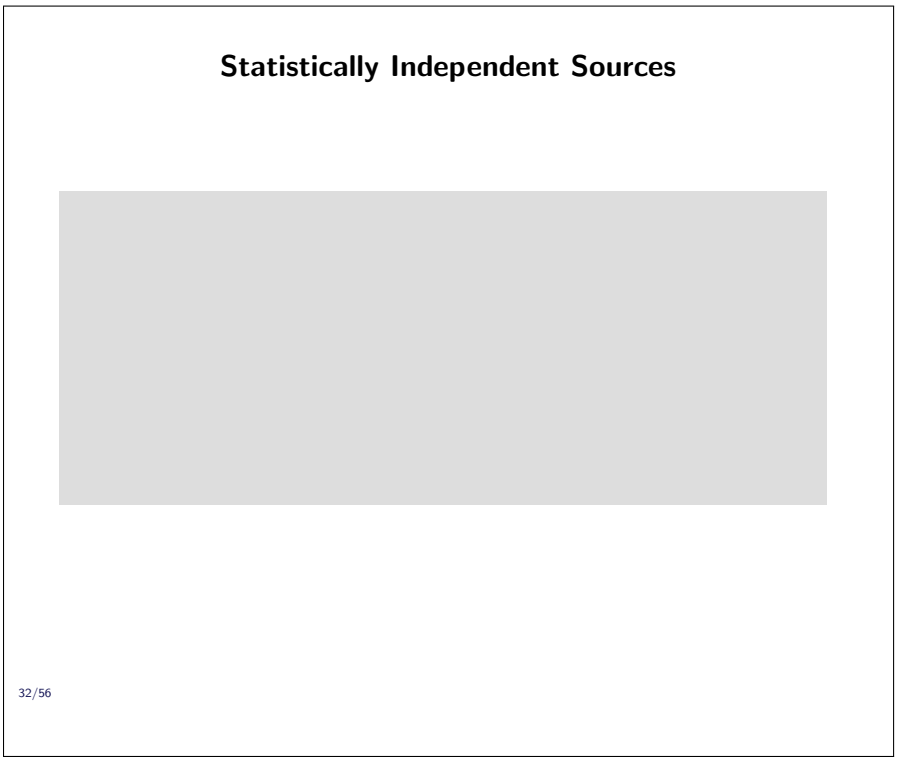
The filters obtained by optimizing the sparseness of the code in the model by Olshausen and Field (1996) resemble simple cell receptive fields fairly well (figure from Olshausen and Field, 2004).

3.3 Statistical independence leads to simple cells

Learning material:

- ☐ 8 min video ~~3.3 Statistical Independence Leads to Simple Cells~~
- ☐ Text below

For a more in depth introduction into independent component analysis see (Wiskott, 2016a).



Assume two stastically independent sources s_1 and s_2 are given, in this case sound sources (left). If one plots samples from the two sources in a common coordinate system such that one component always comes from one source and the other component from the other source, then one gets a two-dimensional data distribution with two statistically independent components. Please notice that the time structure of the signal is now gone and actually irrelevant for what follows.

Intuitively *statistical independence* (D: statistische Unabhängigkeit) means that knowing the value of one component does not tell you anything about the other component of that sample. Visually this roughly means that there may not be any diagonal structures in the plot.

Formaly statistical independence means that the joint probability density function (pdf) equals the product of its marginal pdfs $p(s_1, s_2) = p(s_1)p(s_2)$. This implies that if you cut through the distribution horizontally anywhere, you always get the same 1D curve (namely $p(s_2)$) just scaled differently (by $p(s_1)$), and the same holds for the vertical dimension.

Linear Blind Source Separation

Whitening can be done with PCA.

Rotation can be done based on the objective that the components y_i be statistically independent, here $p(y_1, y_2) = p(y_1)p(y_2)$.

33/56

If one takes samples $\boldsymbol{s}$ from two statistically independent sources and mixes them linearly with an invertible matrix $\boldsymbol{A}$, one gets a mixed signal $\boldsymbol{x}$. If one knew $\boldsymbol{A}$ it would be easy to unmix the data again. One would simply calculate the inverse of $\boldsymbol{A}$ and multiply the data vectors with it. However, even if $\boldsymbol{A}$ is unknown can one unmix the data, up to permutation and scaling, a process called *linear blind source separation*, 'blind' because neither the mixing matrix $\boldsymbol{A}$ nor the sources s_i are known (except that at most one may be Gaussian). The linear algorithm is usually referred to as *Independent Component Analysis (ICA)*.
The first step is whitening, with the argument that statistically independent components must at least be uncorrelated, and that is what whitening gives us. The second step is a rotation, because any skewing or stretching would ruin our whitening again. The rotation angle is dertermined such that some measure of statistical independence or non-Gaussianity is optimized.
It is interesting that making the individual components as non-Gaussian as possible is equivalent to making them as statistically independent as possible. The converse is known from the central limit theorem, if one mixes (adds) random variables, the resulting distribution gets more Gaussian.
The statistically independent components being extracted are all normalized to unit variance and their assignment to the components as well as their sign is arbitrary. This is why I stated above 'up to permutation and scaling'.

Independent Component Analysis on Natural Images

(Bell & Sejnowski, 1997, Vision Research 37:3327–38, Fig. 1)

$\mathbf{y} = \mathbf{Rx} = \mathbf{RAs}$ (with whitened $\mathbf{y}$, i.e. $\langle \mathbf{y}\mathbf{y}^T \rangle = \mathbf{I}$)

34/56

When applying ICA to natural images, the view is that each image itself is a mixture, i.e. a linear superposition, of some statistically independent sources in the real world, and the task of the visual system is to extract these underlying sources from the image (Bell and Sejnowski, 1997).

ICA-Filters for Natural Images

(Bell & Sejnowski, 1997, Vision Research 37:3327–38, Fig. 4)

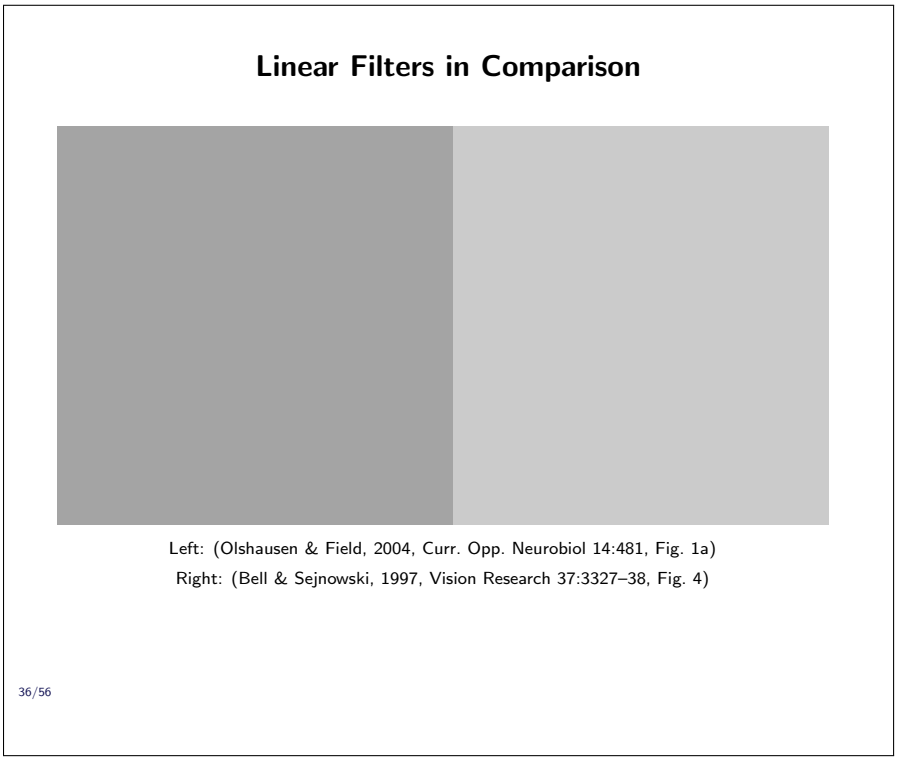
35/56

When one applies ICA to natural images, one gets filters that resemble simple cell receptive fields fairly well (Bell and Sejnowski, 1997).

3.4 Sparseness vs statistical independence

Learning material:

- ☐ 8 min video ~~3.4 Sparseness vs Statistical Independence~~
- ☐ Text below



The filters obtained by the sparseness objective (left) (Olshausen and Field, 2004) and by ICA (right) (Bell and Sejnowski, 1997) look very similar. The reason is that in the linear and complete case and if the underlying sources are sparse the two objectives are equivalent.

Relation Between Sparseness and Independence

37/56

The left figure shows two sparse and statistically independent sources plotted in a common graph with their joint pdf. If one rotates the joint pdf, thereby mixing the components, two things happen: (i) The individual components get less sparse; (ii) The components get more statistically dependent on each other. Thus optimizing sparseness as well as statistical independence both lead to an unmixing of the data; the objectives are equivalent. This would not be true if the sources were non-sparse to begin with.

Linear Models of Visual Receptive Fields - Summary

- Linear filters resulting from principal component analysis on natural images do not resemble simple cell receptive fields.
- Linear filters optimized for sparseness on natural images resemble simple cell receptive fields.
- Linear filters optimized for statistical independence on natural images also resemble simple cell receptive fields.
- Sparseness and statistical independence lead to similar results in linear systems if the underlying sources are sparse.

38/56

LECTURE 2/2

- Lecture 2 Exercises, Lecture 2 Solutions

4 Selforganizational model of complex cells (→ slides)

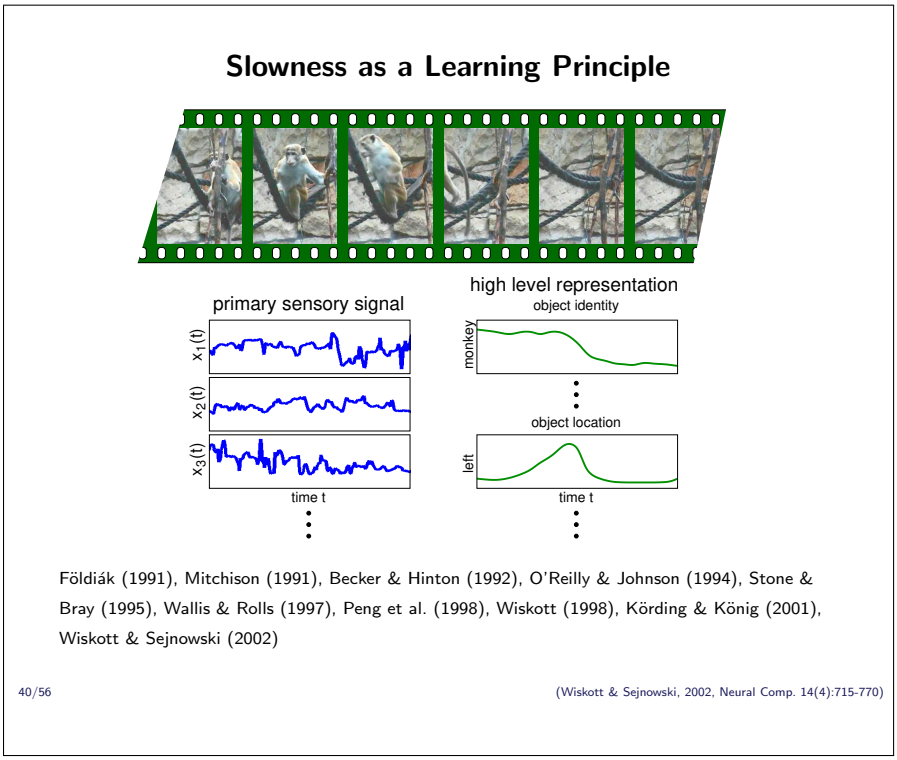
4.1 Slow Feature Analysis (SFA)

Learning material:

☐ 9 min video 4.1 Slow Feature Analysis

☐ Text below

Section title: **Slow Feature Analysis**
Image: (maxmann, 2016, pixabay, © CC0, URL)[4.1]



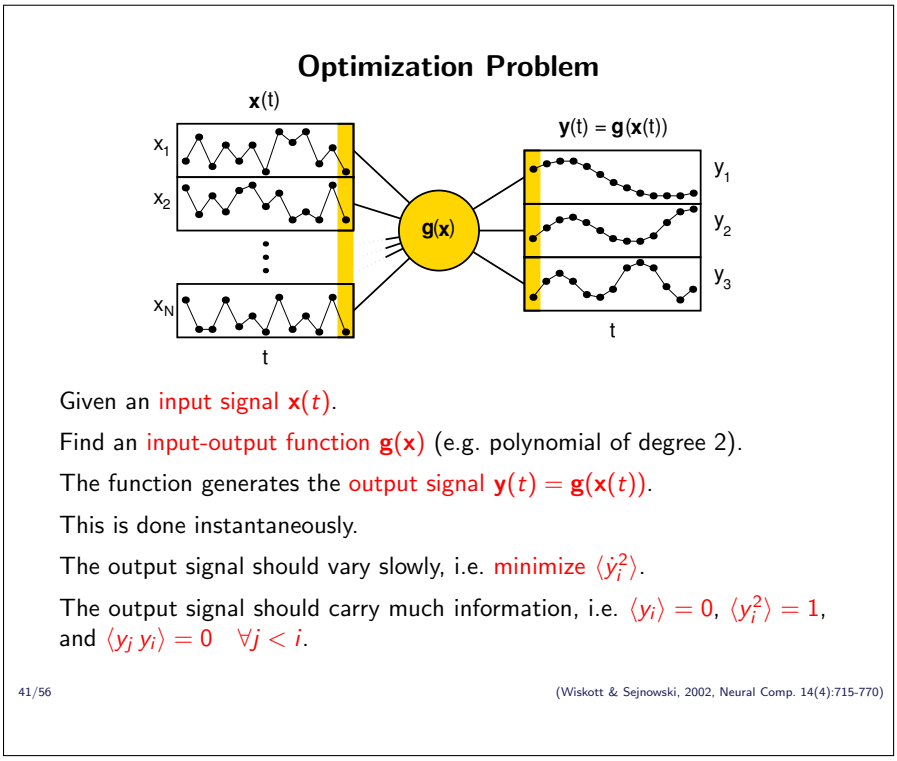
Slowness as a learning principle is based on the observation that different representations of the visual sensory input vary on different time scales. Our visual environment itself is rather stable. It varies on a time scale of seconds.

The primary sensory signal on the hand, e.g. responses of single receptors in our retina or the gray value of a single pixel of a CCD camera, vary on a faster time scale of milliseconds, simply as a consequence of the very small receptive field sizes combined with gaze changes or moving objects. As an example imagine you are looking at a quietly grazing zebra. As your eyes scan the zebra, single receptors rapidly change from black to white and back again because of the stripes of the zebra. But the scenery itself does not change much.

Finally, your internal high-level representation of the environment changes on a similar time scale as the environment itself, namely on a slow time scale. The brain is somehow able to extract the slowly varying high-level representation from the quickly varying primary sensory input. The hypothesis of the slowness learning principle is that the time scale itself provides the cue for this extraction. The idea is that if the system manages to extract slowly varying features from the quickly varying sensory input, then there is a good chance that the features are a good representation of the visual environment.

A number of people have worked along these lines. Slow feature analysis is within this tradition but differs in some significant technical aspects from all previous approaches.

Figure: (Wiskott et al., 2011, Fig. 2, © CC BY 4.0, URL)[4.2]



Slow feature analysis is based on a clearcut optimization problem. The goal is to find input-output functions that extract most slowly varying features from a quickly varying input signal.

It is important that the functions are instantaneous, i.e. one time slice of the output signal is based on just one time slice of the input signal (marked in yellow). Otherwise low-pass filtering would be a valid but not particularly useful method of extracting slow output signals. Instantaneous functions also make the system fast after training, as is important in visual processing, for instance. It is also possible to take a few input time slices into account, e.g. to make the system sensitive to motion or to process scalar input signals with a fast dynamics on a short time scale. However, low-pass filtering should never be the main method by which slowness is achieved.

Without any constraints, the optimal but not very useful output signal would be constant. We thus impose the constraints of unit variance $\langle y_i^2 \rangle = 1$ and, for mathematical convenience, zero mean $\langle y_i \rangle = 0$. To make different output signal components represent different information, we impose the decorrelation constraint $\langle y_j y_i \rangle = 0$. Without this constraint, all output components would typically be the same. Notice that the constraint is asymmetric, later components have to be uncorrelated to earlier ones but not the other way around. This induces an order. The first component is the slowest possible one, the second component is the next slowest one under the constraint of being uncorrelated to the first, the third component is the next slowest one under the constraint of being uncorrelated to the first two, etc.

Figure: (Wiskott et al., 2011, Fig. 1, © CC BY 4.0, URL)[4.3]

4.2 Complex cells with SFA on natural images

Learning material:

- ☐ 15 min video ~~4.2 Complex Cells with SFA on Natural Images~~
- ☐ Text below

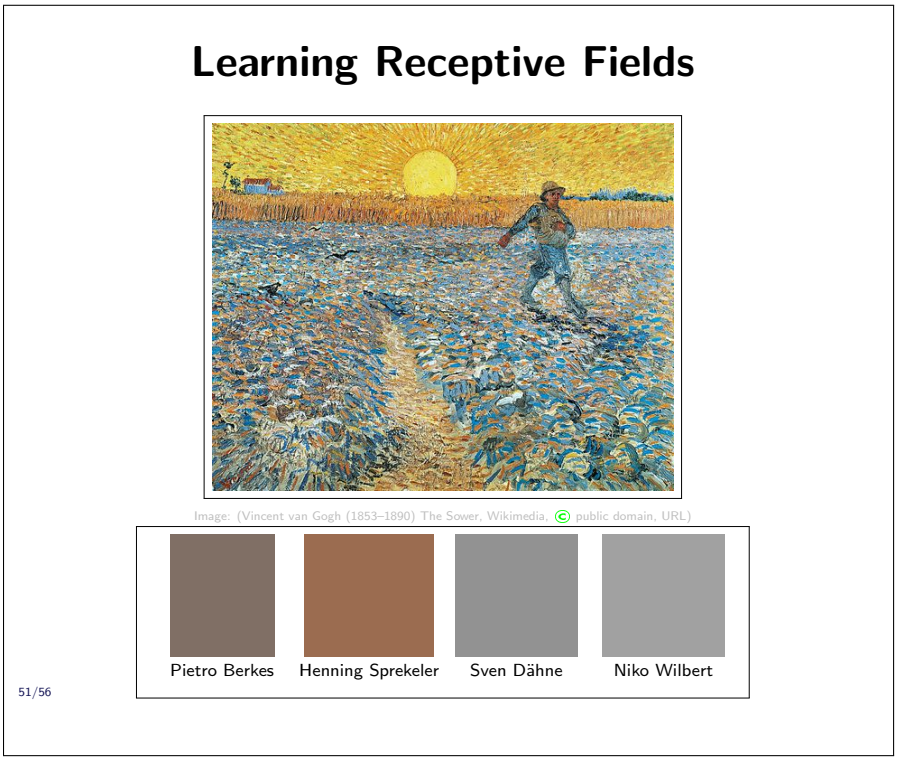
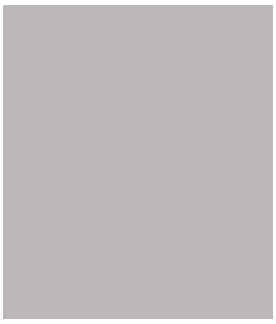


Image: (Vincent van Gogh (1853–1890) The Sower, Wikimedia, © public domain, URL)[4.4]

Training sequences are generated from natural images by moving a selection window across the image by varying translation, rotation, and zoom. Each selected frame is resampled to a size of 16×16 pixels. By concatenating the rows of each frame one obtains the 256-dimensional input vectors for training SFA. To reduce the dimensionality, we perform principal component analysis on the input images and only keep the first 100 components.

Trained with such image sequences, SFA yields a set of functions that extract the most slowly varying features. The functions are ordered by slowness, so that the first one extracts the slowest feature, the second one the next slowest feature, etc. We typically keep the first 100 functions, because the input is 100-dimensional and because later functions yield output signals that vary faster than even the input and the slowness objective is not met anymore.

Units are Polynomials



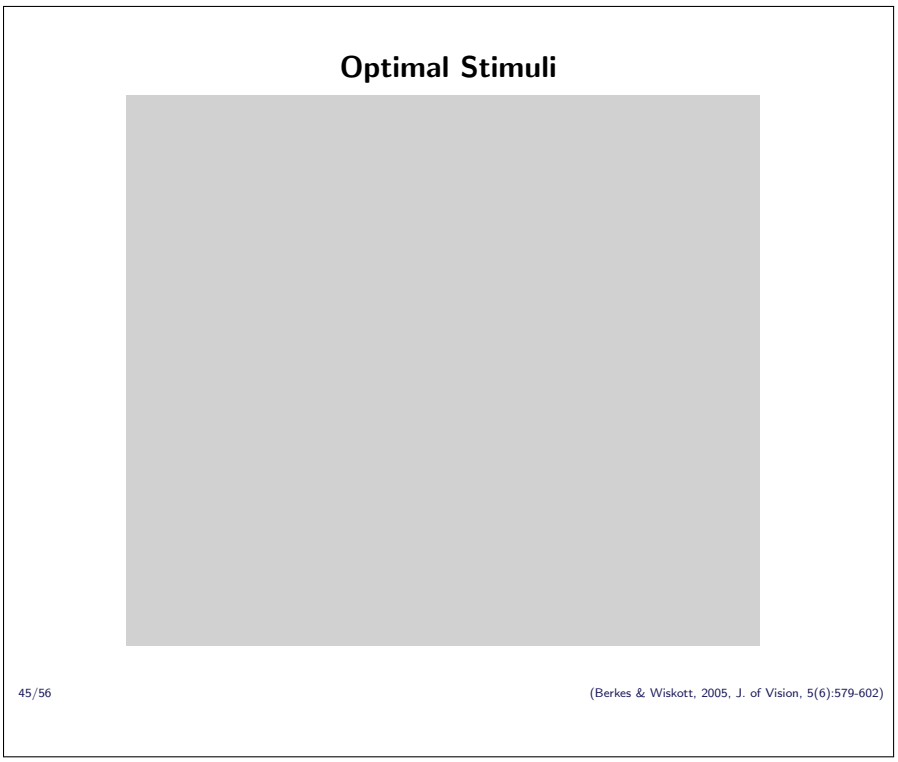
$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \boldsymbol{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c$$
$$= (h_{11}x_1^2 + h_{12}x_1x_2 + ... + h_{NN}x_N^2)/2 + f_1x_1 + ... + f_Nx_N + c$$

Each $g_i(\mathbf{x})$ has $(N(N+1)/2 + N + 1) = 5151$ free parameters!

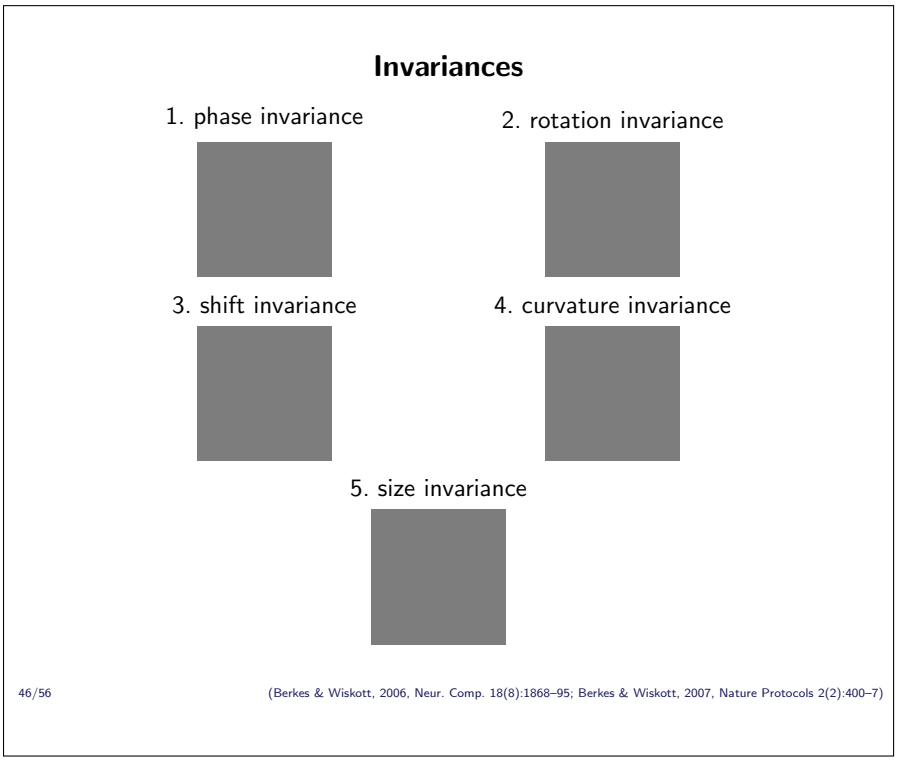
44/56

(Berkes & Wiskott, 2005, J. of Vision, 5(6):579-602)

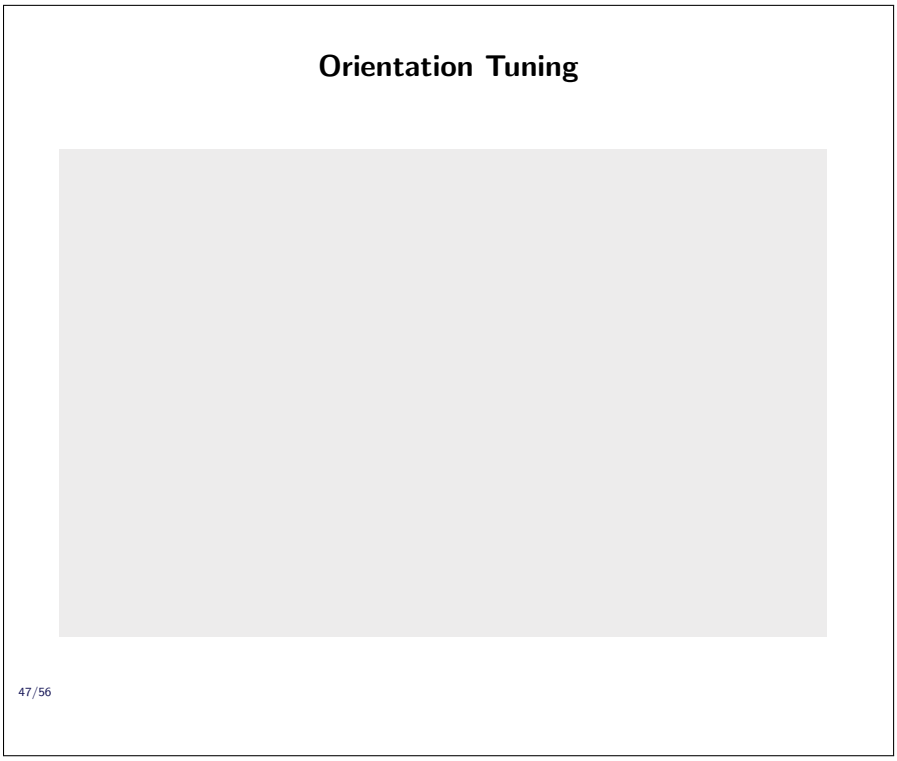
The function space used here is the set of polynomials of degree two in the 256 pixel gray values, or rather their 100 first principal components. This yields 1 constant term, 100 linear terms, 100 quadratic terms, and 4950 mixed products of two different input components, which makes 5151 terms in total. This is a large function space (therefore the dimensionality reduction of the images down to 100, to keep the dimensionality manageable).
Using such a large function space is important as it provides sufficient computational power and reduces built-in prejudices about expected results. Using even larger function spaces might be interesting but was computationally prohibitive for us. One might also argue that physiological cells have computational limitations and that there are experimental results suggesting that polynomials of degree two might be appropriate also for that reason.



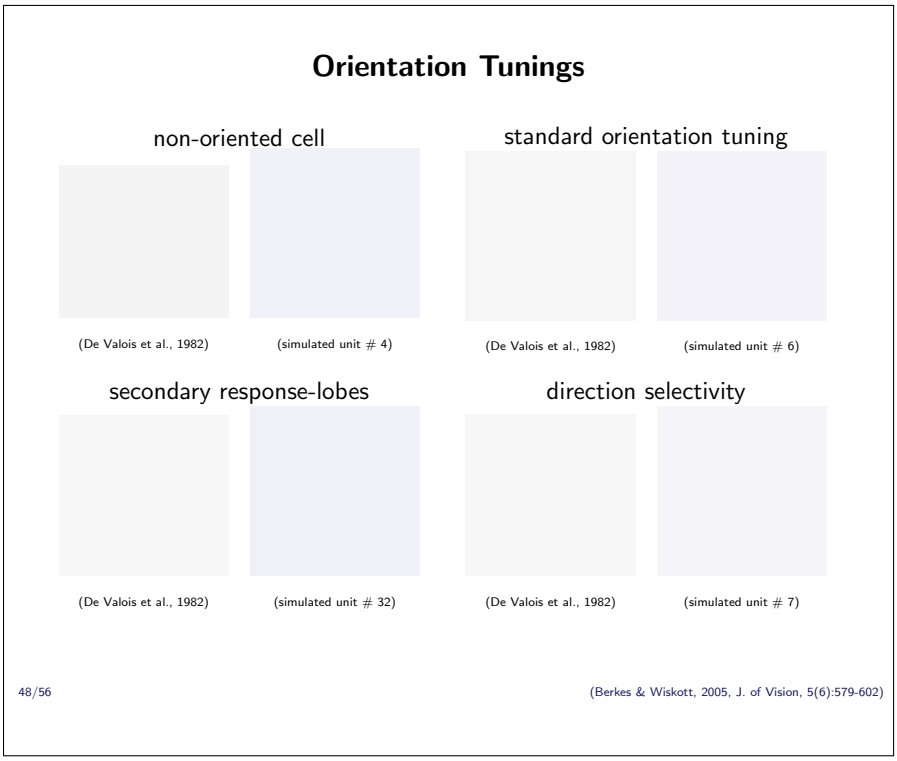
The optimal stimuli of the units obtained with SFA have the shape of Gabor wavelets, which is in good agreement with physiological complex cells.
Notice that the receptive fields are not particularly localized but cover a large fraction of the image patch. This is because responses are typically more slowly varying if the receptive fields are large. On the other hand, the receptive fields do not extend to the borders of the image patch at full strength, because that would cause rapid changes in the output if new image gray values move into the receptive field. Maybe, one can say that a large Gaussian envelop function is optimal for slowness.
Notice also that since the SFA units are nonlinear, the optimal stimuli only give a first hint at the full response properties of the units. If the units were linear, the optimal stimuli would characterize them completely.



The first few invariances of the units can typically be interpreted intuitively. Every unit shows phase invariance, which means that the exact position of the black and white stripes within the wavelet does not matter. This makes the units similar to complex cells rather then simple cells. The unit shown here has in addition rotation invariance, i.e. the wavelet may rotate a bit, shift invariance, i.e. the wavelet may move a little bit, curvature invariance, i.e. the stripes of the wavelet may bend a little bit, and size invariance, i.e. the wavelet may vary in size a little bit without changing the response too much.



Simple and complex cells usually have preferences for certain orientations. This can be measured by presenting gratings of different orientation to the cell (or rather the animal) and recording the corresponding neural responses. Normally, drifting gratings are used, because the cells respond stronger to moving stimuli.
The responses to different orientations can be conveniently visualized in a polar plot. One simply plots the firing rate in radial direction as a function of orientation in azimuthal direction. The graph shows a standard orientation tuning with one preferred orientation at about 160°, which appears here as two lobes in 180° distance due to the two different drifting directions. Since the two lobes have same size, the cell does not have a preference for a particular drifting direction.



A common way of characterizing physiological complex cells is to measure their orientation tuning, which can be conveniently plotted in polar plots like shown here. Complex cells have a great variety of different orientation tunings (black curves). Interestingly the units trained with SFA reproduce quite a few of these (blue curves). Some units are not selective for orientation at all (non-oriented cells). Many have only one preferred orientation (standard orientation tuning). Some units have an additional positive reponse at a second orientation (secondary response lobes). Others are direction selective, i.e. they respond only if the grating moves in one particular direction.

End- and Side-Inhibition

49/56 (Berkes & Wiskott, 2005, J. of Vision, 5(6):579-602)

Some SFA units also show end- or side-inhibition (blue curves). The level to which the response drops when the grating covers the whole receptive field relative to the maximum response varies a lot for both types of cells in any case and is not a discrepancy between the simulation results and physiological measurements.

Histograms

50/56 (Berkes & Wiskott, 2005, J. of Vision, 5(6):579-602)

The good agreement of the simulation results with physiological measurements is not only on a cell-by-cell level but also on a population level. The histograms of orientation bandwidth (upper left panel) and relative orientation between maximum excitation and maximum inhibition (upper right panel) fit quite nicely experimental data. Frequency bandwidth of the simulated cells is biased towards low values (lower left panel). This is a consequence of the relatively small image patches used in combination with the dimensionality reduction, which leaves only a small bandwidth available. We have no good explanation for the fact that simulated units are less direction selective than physiological ones (lower right panel).

4.3 Complex cells with SFA on colored noise images

Learning material:

- ☐ 10 min video 4.3 Complex Cells with SFA on Colored Noise Images
- ☐ Text below

Learning Receptive Fields

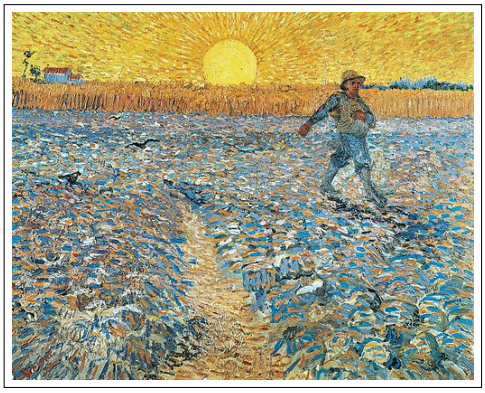


Image: (Vincent van Gogh (1853–1890) The Sower, Wikimedia, © public domain, URL)

Pietro Berkes Henning Sprekeler Sven Dähne Niko Wilbert

51/56

Image: (Vincent van Gogh (1853–1890) The Sower, Wikimedia, © public domain, URL)[4.5]

Colored Noise Image Sequences

It also works with colored noise but not with white noise image sequences.

The results do not depend on higher order statistics in the images.

52/56 (Berkes & Wiskott, 2005, J. of Vision, 5(6):579-602)

Interestingly, the results obtained do not depend on the natural images used. In fact, one gets qualitatively identical results when colored noise images are used. This means that the higher-order statistics of natural images is not essential for the development of complex cell properties based on the slowness principle. Instead, the transformations are essential. Control experiments show that without translation, i.e. with only rotation and zoom around a common center, the optimal stimuli become spirals, funnels, and tunnels, but not wavelets. Thus, translation is essential. Rotation and zoom only limit the size of the receptive field and are less influential.

The fact that colored noise images are sufficient for the development of complex cells with SFA implies that the problem is in principle amenable to an analytical treatment, since all conditions can be formulated analytically.

Theory

If we assume infinitely large receptive fields, variational calculus leads to the eigenvalue equation

$$\mathbf{D} g_j(\mathbf{r}, \mathbf{r}') = \Delta_j g_j(\mathbf{r}, \mathbf{r}') \,,$$

with the differential operator

$$\begin{aligned} \mathbf{D} \;=\; & -\langle v^2 \rangle (\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'})^2 \\ & -\langle \omega^2 \rangle (\mathbf{r} \times \nabla_{\mathbf{r}} + \mathbf{r}' \times \nabla_{\mathbf{r}'})^2 \\ & -\langle \alpha^2 \rangle (\nabla_{\mathbf{r}} \cdot \mathbf{r} + \nabla_{\mathbf{r}'} \cdot \mathbf{r}')^2 \,. \end{aligned}$$

If we assume perfect translation invariance, this eigenvalue problem can be solved and yields input-output functions with plane waves as optimal stimuli.

53/56 (Sprekeler & Wiskott, 2011, Neural Computation, 23(2):303-335)

An analytical treatment (variational calculus) of the self-organization of complex cell receptive fields with SFA under the assumption of infinitely large receptive fields leads to an eigenvalue equation with a differential operator $\mathbf{D}$ containing three terms. They result from translation, rotation, and zoom. If we assume perfect translation invariance, this eigenvalue problem can be solved and yields input-output functions with plane waves as optimal stimuli.

Retinal Waves

(2011-06-14 `http://mcb.berkeley.edu/labs/feller/Movies.htm`; Blankenship et al, 2009, Neuron)

(model by Godfrey & Swindale, 2007; implementation by Sven Dähne, 2009)

(Dähne, Wilbert, & Wiskott, 2014, PLoS Comp Biol, 10(5):e1003564)

Results with Retinal Waves

optimal stimuli

orientation tuning

(Dähne, Wilbert, & Wiskott, 2014, PLoS Comp Biol, 10(5):e1003564)

Learning Receptive Fields - Summary

- Slow feature analysis applied to image sequences with translation, rotation, and zoom yields many receptive-field properties of complex cells in V1.
- Also the histograms of a number of cell properties fit well.
- The results do not seem to depend on higher-order image-statistics.
- The results depend on the presence of translational motion.
- Complex cell receptive fields can therefore emerge also based on retinal waves.
- The model predicts a systematic relationship between reponse time-scale and receptive-field properties.
- The results can be partially understood theoretically.

5 Optimal stimuli for quadratic forms

Learning material:

☐ Text below

This section is based on (Berkes and Wiskott, 2006, sec. 4).

We have seen above that SFA with polynomials of degree two applied to quasi-natural image sequences yields many properties of complex cells. One way to visualize the results was to plot the optimal excitatory (inhibitory) stimuli, which yield the maximal (minimal) output under a fixed norm constraint. In this section we will se how one can find these optimal stimuli.

The problem of finding the optimal excitatory stimulus for quadratic forms under a fixed norm constraint can be mathematically formulated as follows:

$$\begin{array}{ll} \text{maximize} & \blacklozenge \quad g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x} + c \\ \text{under the constraint} & \blacklozenge \quad \mathbf{x}^T\mathbf{x} = r^2\,. \end{array} \tag{5.1}$$

This problem is known as the *Trust Region Subproblem* and has been extensively studied in the context of numerical optimization, where a nonlinear function is minimized by successively approximating it by an inhomogeneous quadratic form, which is in turn minimized in a small neighborhood.

If the linear term is equal to zero (i.e., $\mathbf{f} = \mathbf{0}$), the problem can be solved easily (see the exercises). In the following we consider the more general case where $\mathbf{f} \neq \mathbf{0}$. If the problem were unconstrained, we would simply look for points where the gradient vanishes, i.e. where $\nabla g(\mathbf{x}) = \mathbf{0}$. **To incorporate the constraint $\mathbf{x}^T\mathbf{x} = r^2$ into the objective function we use a Lagrange formulation** to find the necessary conditions for the extremum:

$$\lozenge \qquad \mathbf{x}^T\mathbf{x} = r^2 \tag{5.2}$$

$$\text{and} \quad \lozenge \qquad \nabla[g(\mathbf{x}) - \lambda\mathbf{x}^T\mathbf{x}/2] = \mathbf{0} \tag{5.3}$$

$$\lozenge \quad \Longleftrightarrow \quad \mathbf{H}\mathbf{x} + \mathbf{f} - \lambda\mathbf{x} = \mathbf{0} \tag{5.4}$$

$$\lozenge \quad \Longleftrightarrow \quad \lambda\mathbf{I}\mathbf{x} - \mathbf{H}\mathbf{x} = \mathbf{f} \tag{5.5}$$

$$\lozenge \quad \Longleftrightarrow \quad \mathbf{x} = (\lambda\mathbf{I} - \mathbf{H})^{-1}\,\mathbf{f}\,, \tag{5.6}$$

where we inserted the factor $1/2$ for mathematical convenience.

It can be shown that, if an x that satisfies Equation (5.6) is a solution to (5.1), then $(\lambda\mathbf{I} - \mathbf{H})$ is positive semidefinite, i.e. all eigenvalues are greater or equal to 0 (Fortin, 2000, Theorem 3.1). **This imposes a tight lower bound on the range of possible values for λ.** Note that the matrix $(\lambda\mathbf{I} - \mathbf{H})$ has the same eigenvectors $\mathbf{v}_i$ as $\mathbf{H}$ with eigenvalues $(\lambda - \mu_i)$. For $(\lambda\mathbf{I} - \mathbf{H})$ to be positive semidefinite all eigenvalues must be nonnegative, and thus λ must be greater than the largest eigenvalue μ_1,

$$\lozenge \quad \mu_1 \leq \lambda\,. \tag{5.7}$$

Proving the abovementioned theorem is beyond the scope of this lecture, but one can get an intuitive understanding of it by considering a quadratic form with a Hesse matrix with identical eigenvalues. ...

An upper bound for λ can be found by considering an upper bound for the norm of x. First we note that matrix $(\lambda\mathbf{I} - \mathbf{H})^{-1}$ is symmetric and has the same eigenvectors as $\mathbf{H}$ with eigenvalues $1/(\lambda - \mu_i)$. We also know that $\|\mathbf{A}\mathbf{v}\| \leq \|\mathbf{A}\|\|\mathbf{v}\|$ for every matrix $\mathbf{A}$ and vector $\mathbf{v}$. $\|\mathbf{A}\|$ is here the spectral norm of $\mathbf{A}$,

which for symmetric matrices is simply the largest absolute eigenvalue. With this we find an upper bound for λ:

$$\begin{aligned}
\Diamond \qquad r &= \|\mathbf{x}\| && (5.8)\\
\Diamond \qquad &= \|(\lambda\mathbf{I}-\mathbf{H})^{-1}\mathbf{f}\| && (5.9)\\
\Diamond \qquad &\leq \|(\lambda\mathbf{I}-\mathbf{H})^{-1}\|\,\|\mathbf{f}\| && (5.10)\\
\Diamond \qquad &= \max_i \left\{\left|\frac{1}{\lambda-\mu_i}\right|\right\}\|\mathbf{f}\| && (5.11)\\
\Diamond \qquad &\overset{(5.7)}{=} \frac{1}{\lambda-\mu_1}\|\mathbf{f}\| && (5.12)\\
\Diamond \quad \Longleftrightarrow \quad & \lambda \leq \frac{\|\mathbf{f}\|}{r}+\mu_1\,. && (5.13)
\end{aligned}$$

The optimization problem (5.1) is thus reduced to a search over λ on the interval $\left[\mu_1, \left(\frac{\|\mathbf{f}\|}{r}+\mu_1\right)\right]$ until $\mathbf{x}$ defined by (5.6) fulfills the constraint $\|\mathbf{x}\| = r$ (Eq. 5.2). **Vector $\mathbf{x}$ and norm $\|\mathbf{x}\|$ can be efficiently computed for each λ** using the eigenvalue decomposition of $\mathbf{f}$:

$$\begin{aligned}
\Diamond \quad \mathbf{x} &\overset{(5.6)}{=} (\lambda\mathbf{I}-\mathbf{H})^{-1}\mathbf{f} && (5.14)\\
\Diamond \qquad &= (\lambda\mathbf{I}-\mathbf{H})^{-1}\sum_i \mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f}) && (5.15)\\
\Diamond \qquad &= \sum_i (\lambda\mathbf{I}-\mathbf{H})^{-1}\,\mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f}) && (5.16)\\
\Diamond \qquad &= \sum_i \frac{1}{\lambda-\mu_i}\mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f}) && (5.17)
\end{aligned}$$

and

$$\begin{aligned}
\Diamond| \quad \|\mathbf{x}\|^2 &\overset{(5.17)}{=} \left(\sum_i \frac{1}{\lambda-\mu_i}\mathbf{v}_i\,(\mathbf{v}_i^T\mathbf{f})\right)^T \left(\sum_j \frac{1}{\lambda-\mu_j}\mathbf{v}_j\,(\mathbf{v}_j^T\mathbf{f})\right) && (5.18)\\
\Diamond \qquad &= \sum_{ij} \frac{1}{\lambda-\mu_i}\frac{1}{\lambda-\mu_j}\,(\mathbf{v}_i^T\mathbf{f})\,(\mathbf{v}_j^T\mathbf{f})\,\underbrace{(\mathbf{v}_i^T\mathbf{v}_j)}_{\delta_{ij}} && (5.19)\\
\Diamond \qquad &= \sum_i \left(\frac{1}{\lambda-\mu_i}\right)^2 (\mathbf{v}_i^T\mathbf{f})^2\,, && (5.20)
\end{aligned}$$

where the terms $\mathbf{v}_i^T\mathbf{f}$ and $(\mathbf{v}_i^T\mathbf{f})^2$ are constant for each quadratic form and **can be computed in advance.** The last equation also shows that **the norm of $\mathbf{x}$ is monotonically decreasing in the considered interval,** so that there is exactly one solution and the search can be efficiently performed by a bisection method. $\mathbf{x}^-$ can be found in the same way by maximizing the negative of g.

If the matrix $\mathbf{H}$ is negative definite (i.e., all its eigenvalues are negative) there is a global maximum that may not lie on the sphere, which might be used in substitution for $\mathbf{x}^+$ if it lies in a region of the input space that has a high probability of being reached (the criterion is quite arbitrary, but the region could be chosen to include, for example, 75% of the input data with highest density). The gradient of the function disappears at the global extremum such that it can be found by solving a simple linear equation system:

$$\begin{aligned}
& \nabla g(\mathbf{x}) = \mathbf{H}\mathbf{x}+\mathbf{f} = \mathbf{0} && (5.21)\\
\Longleftrightarrow \qquad & \mathbf{x} = -\mathbf{H}^{-1}\mathbf{f}. && (5.22)
\end{aligned}$$

In the same way a positive definite matrix $\mathbf{H}$ has a negative global minimum, which might be used in substitution for $\mathbf{x}^-$.

Note that although $\mathbf{x}^+$ is the stimulus that elicits the strongest response in the function, it doesn't necessarily mean that it is representative of the class of stimuli that give the most important contribution to its output. This depends on the distribution of the input vectors: If $\mathbf{x}^+$ lies in a low-density region of the input space, it is possible that other kinds of stimuli drive the function more often. In that case they might be considered more relevant than $\mathbf{x}^+$ to characterize the function. Symptomatic for this effect would be if the output of a function when applied to its optimal stimulus would lie far outside the range of normal activity. This means that $\mathbf{x}^+$ can be an atypical, artificial input that pushes the function in an uncommon state. However, the optimal stimuli remain extremely informative in practice.

References

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

Berkes, P. and Wiskott, L. (2006). On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural Computation*, 18(8):1868–1895.

Fortin, C. (2000). A survey of the trust region subproblem within a semidefinite framework. Master's thesis, University of Waterloo.

Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3(1):61–70.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–7.

Oram, M. W. and Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7(6/7):945–972.

Sato, T., Uchida, G., Lescroart, M. D., Kitazono, J., Okada, M., and Tanifuji, M. (2013). Object representation in inferior temporal cortex is organized hierarchically in a mosaic-like structure. *Journal of Neuroscience*, 33(42):16642–16656.

Wiskott, L. (2003). How does our visual system achieve shift and size invariance? Cognitive Sciences EPrint Archive (CogPrints).

Wiskott, L. (2016a). Lecture notes on independent component analysis. Available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/index.html.

Wiskott, L. (2016b). Lecture notes on principal component analysis. Available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/index.html.

Wiskott, L., Berkes, P., Franzius, M., Sprekeler, H., and Wilbert, N. (2011). Slow feature analysis. *Scholarpedia*, 6(4):5282.

Notes

[1.1] Alphab.fr, 2007, Wikipedia, © CC BY 2.0, https://en.wikipedia.org/wiki/File:Rice_fields_near_Sapa,_Vi%C3%AAt_Nam.jpg

[1.2] Sato, Uchida et al, 2013, J. Neuroscience, Fig. 5, http://www.jneurosci.org/content/33/42/16642

[4.1] maxmann, 2016, pixabay, © CC0, https://pixabay.com/en/snail-shell-crawl-mollusk-1330766/

[4.2] Wiskott et al., 2011, Scholarpedia 5(2):1362, Fig. 2, © CC BY 4.0, http://scholarpedia.org/article/Slow_feature_analysis

[4.3] Wiskott et al., 2011, Scholarpedia 5(2):1362, Fig. 1, © CC BY 4.0, http://scholarpedia.org/article/Slow_feature_analysis

[4.4] Vincent van Gogh (1853–1890) The Sower, Wikimedia, © public domain, https://commons.wikimedia.org/wiki/File:The_Sower.jpg

[4.5] Vincent van Gogh (1853–1890) The Sower, Wikimedia, © public domain, https://commons.wikimedia.org/wiki/File:The_Sower.jpg