A. Hyvärinen and P. O. Hoyer

A Two-Layer Sparse Coding Model Learns Simple and Complex Cell Receptive Fields and Topography from Natural Images.

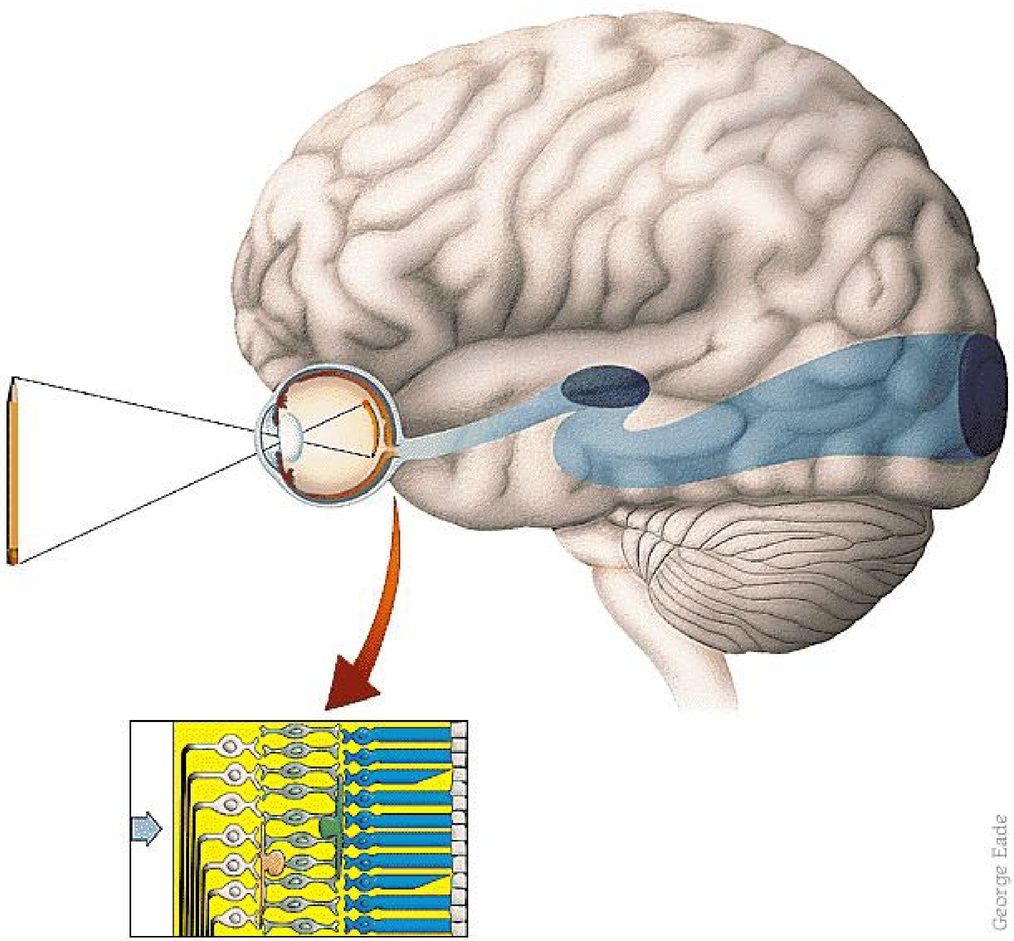
presented by

Hsin-Hao Yu

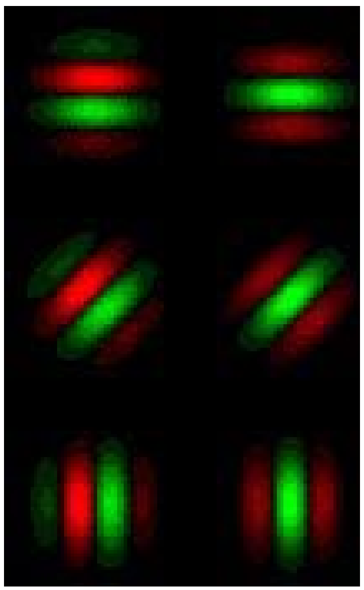
Department of Cognitive Science

November 7, 2001

An overview of the visual pathway



Basic V1 physiology



Simple cells	approximately linear filters
	localized, oriented, band-pass
	phase sensitive
Complex cells	non-linear
	phase insensitive

Question: Why do we have these neurons?

The principle of redundancy reduction

The Principle of redundancy reduction: The world is highly structured. The purpose of early sensory processing is to transform the redundant sensory input to an efficient code. [Barlow 1961]

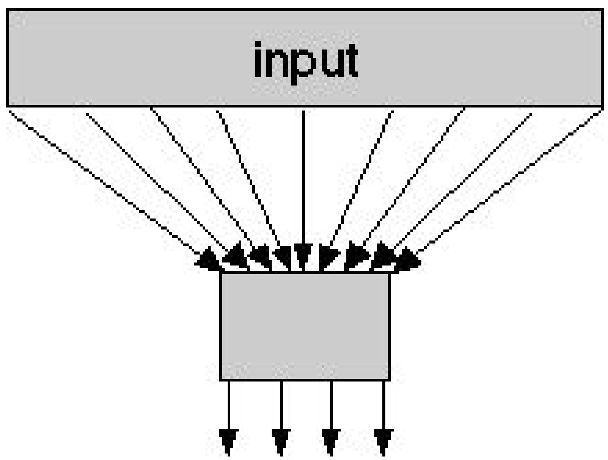
Two approaches have been developed to apply this idea to study the visual cortex:

1. Sparse coding (eg. Olshausen and Field)
2. Independent Component Analysis (eg. Bell and Sejnowski)

Compact coding vs. Sparse coding

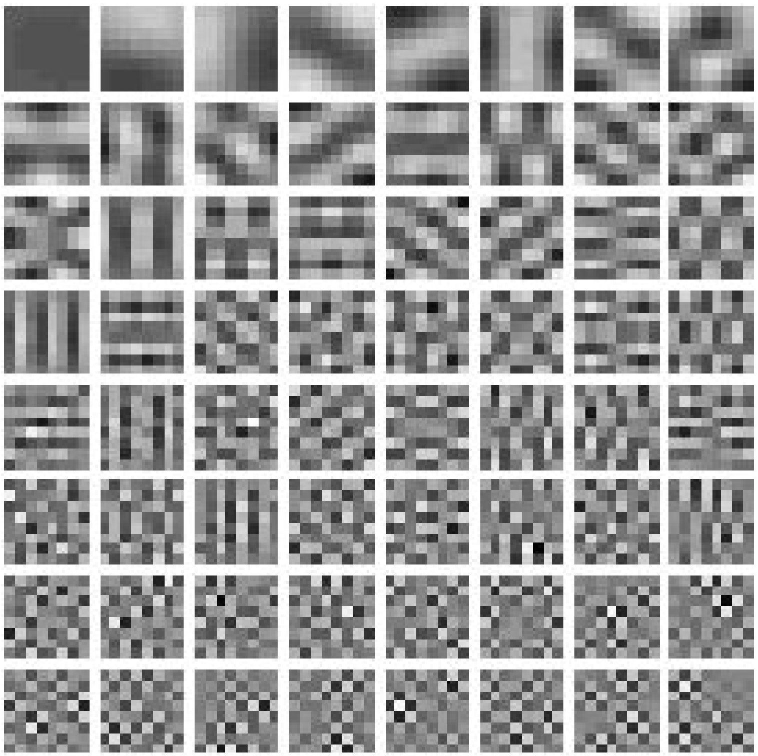
What does a ***efficient code*** **means?**

Strategy 1: *Compact coding* represents data with minimum number of units.



This requirement often produces solutions that's similar to *Principal Component Analysis*, but the principal components do not resemble any receptive field structures found in the visual cortex.

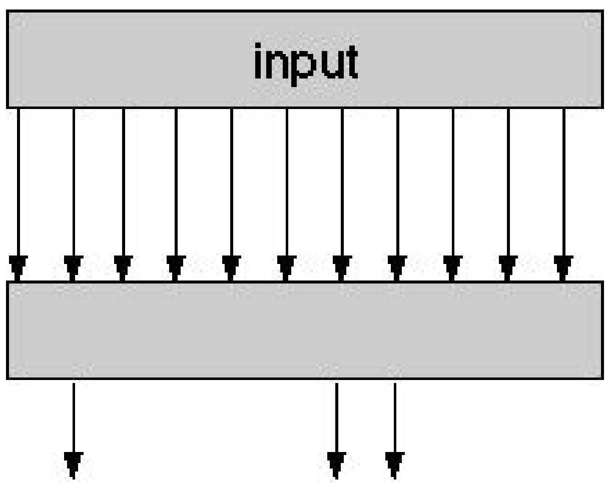
Principal components of natural images



Not localized, and no orientational selectivity.

Compact coding vs. Sparse coding

Strategy 2: *Sparse coding* represents data with minimum number of *active* units, but the dimensionality of the representation is the same as (or even larger than) the dimensionality of the input data.



Learning sparse codes: image model

We use the linear generative mode. That is,

$$I(x, y) = \sum_i a_i \phi_i(x, y)$$

where $I(x, y)$ is a patch of natural image, and $\{a_i\}$ are coefficients to the *basis functions* $\{\phi_i(x, y)\}$.

A neural network interpretation: writing images as column vectors,

$$\begin{bmatrix} \\ I \\ \\ \end{bmatrix} = \begin{bmatrix} \cdots & \phi_1 & \cdots \\ & & \\ & & \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

or $I = \Phi A$. Thus, $A = WI$ where $W = \Phi^{-1}$. A is the output layer of a linear network, and W is the weight matrix (ie. *filters*.)

Learning sparse codes: algorithm

[Olshausen and Field, 1996] For the image model

$$I(x,y) = \sum_i a_i \phi_i(x,y)$$

We require that the distributions of the coefficients, a_i, are "sparse". This can be achieved by minimizing the following cost function:

$$\begin{aligned} E &= -[fidelity] - \lambda[sparseness] \\ fidelity &= -\sum_{x,y} \left[I(x,y) - \sum_i a_i \phi_i(x,y)\right]^2 \\ sparseness &= -\sum_i S(a_i) \\ S(x) &= log(1 + x^2). \end{aligned}$$

Maximum-likelihood and sparse codes

The sparse-coding algorithm can be interpreted as finding ϕ that maximizes the average log-likelihood of the images under a sparse, *independent* prior.

fidelity — negative log-likelihood of the image given ϕ and a, assuming gaussian noise.

$$P(I|a,\phi) = \frac{1}{Z_{\rho_N}} e^{-\frac{|I-a\rho|^2}{2\rho_N^2}}$$

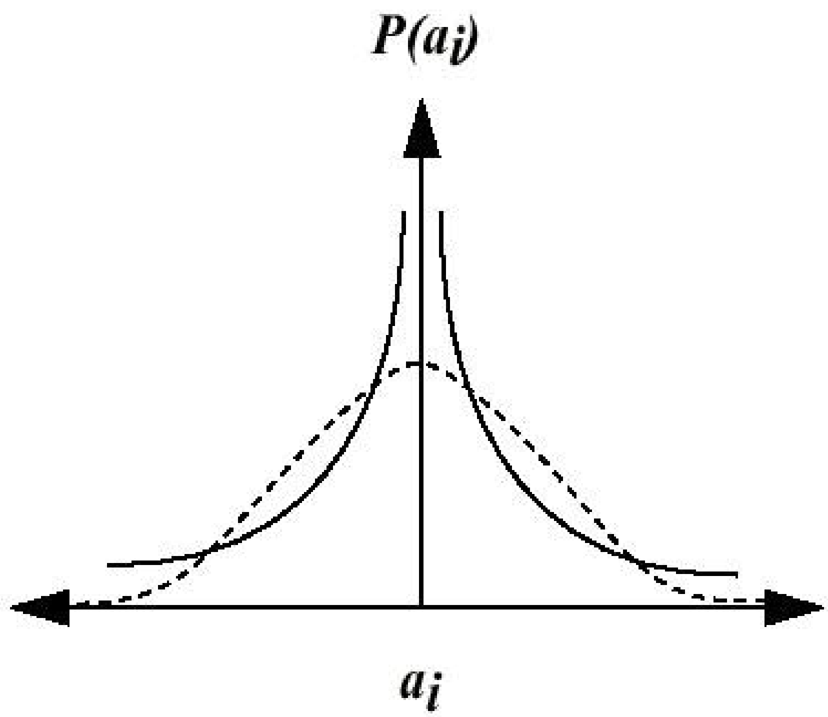
sparseness — sparse, independent prior for a.

$$P(a) = \prod_i e^{-\beta S(a_i)}$$

So $E \propto -log(P(I|a,\phi)P(a))$. It can be shown that minimizing E is equal to maximizing $P(I|\phi)$, given some approximation assumptions.

Supergaussian distributions

$S(a_i) = log(1 + a_i^2)$	$P(a_i) = \frac{1}{1+a_i^2}$	Cauchy distribution
$S(a_i) = \|a_i\|$	$P(a_i) = e^{-\|x\|}$	Laplace distribution



Independent Component Analysis

In the context of natural image analysis:

$$I(x,y) = \sum_i a_i \phi_i(x,y)$$

where the number of a_i equals to the dimensionality of I. We require that $\{a_i\}$, as random variables, are independent to each other. That is, $P(a_i|a_j) = P(a_i)$.

In a more general context, let I be a random vector. The goal of the Independent Component Analysis is to find a matrix W, such that the components of $A = WI$ are non-gaussian, and independent to each other.

The Infomax ICA

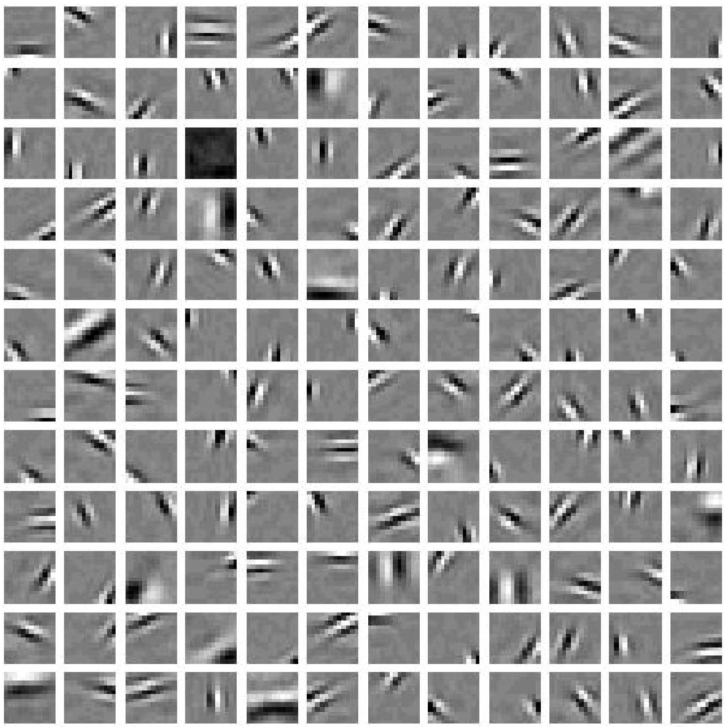
[Bell and Sejnowski 1995] derived a learning rule for ICA by maximizing the entropy of a neural network with logistic (or Laplace) neurons. Similar or equivalent algorithms can be derived from many other frameworks.

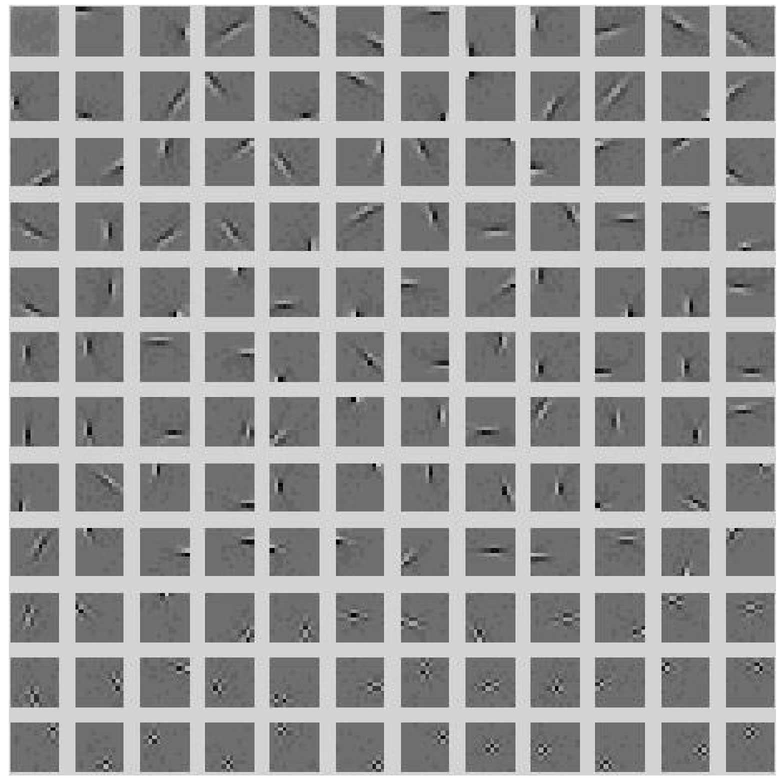
Let $H(X)$ be the entropy of X. The joint entropy of a_1 and a_2 can be written as:

$$H(a_1, a_2) = H(a_1) + H(a_2) - I(a_1, a_2)$$

where $I(a_1, a_2)$ is the mutual information between a_1 and a_2. $\{a_1, a_2\}$ are independent to each other when $I(a_1, a_2) = 0$. We approximate the solution by maximizing $H(a_1, a_2)$.

Independent components of natural images





Olshausen and Field 1996

16x16 basis patches

Bell and Sejnowski 1996

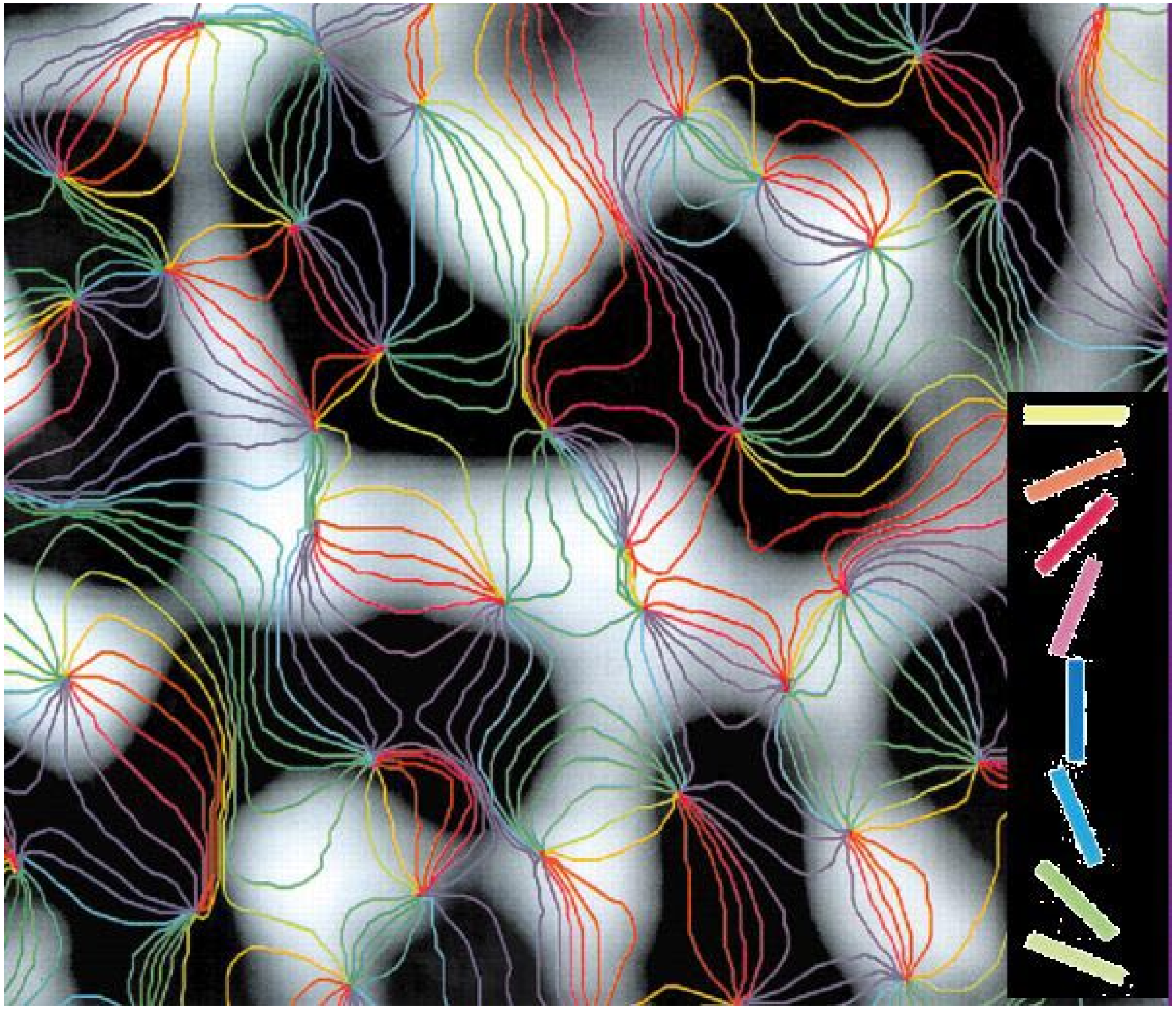
12x12 filters

More ICA applications

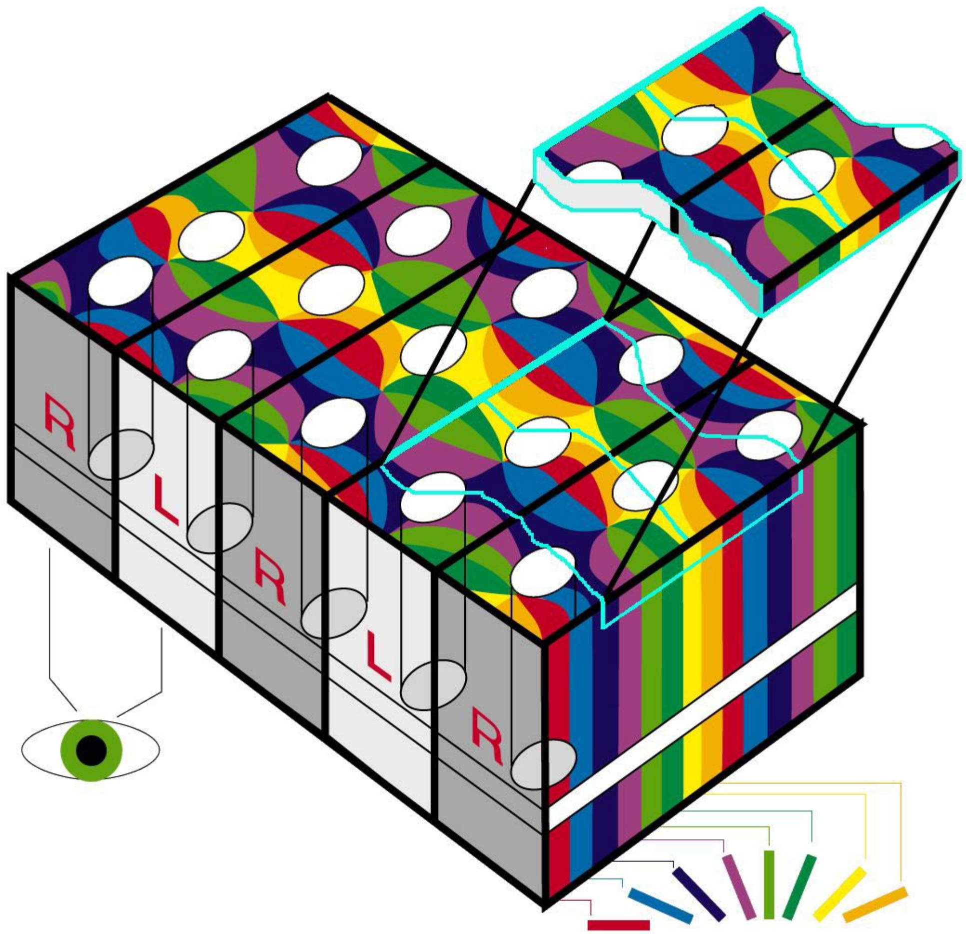
1. Direction selectivity [van Hatern et al., 1998]
2. Flow-field templates [Park and Jabri, 2000]
3. Color [Hoyer, 2000; Tailor, 2000; Lee, 2001]
4. Binocular vision [Hoyer, 2000]
5. Audition [Bell and Sejnowski 1996; Lewicki??]

Complex cells and topography

[Hyvärinen and Hoyer, 2001] uses a hierarchical network and the sparse coding principle to explain the emergence of complex-cell-like receptive fields and topographic structures of simple cells.

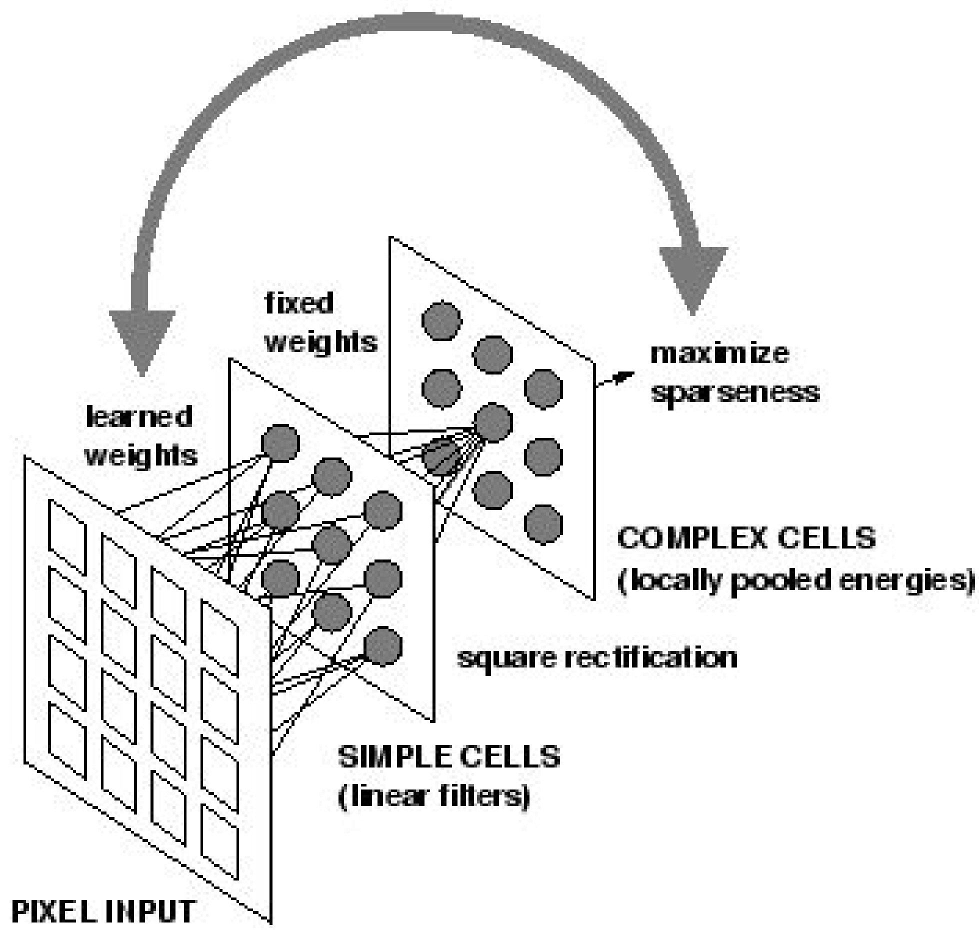


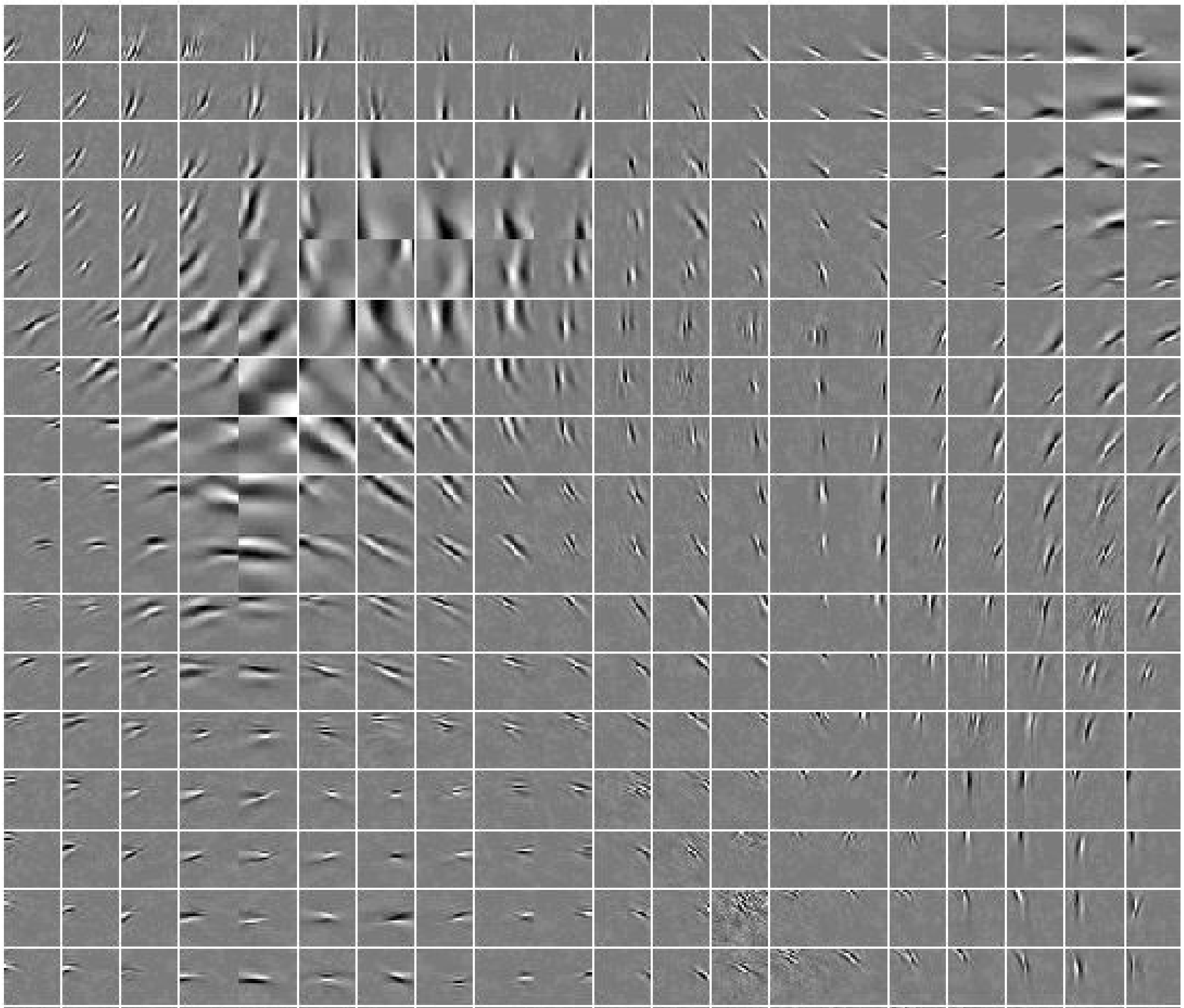
from [Hübener et al. 1997]



The “ice-cube” model of V1 layer 4c

Network architecture





Results: summary

simple cell physiology

orientation/freq selective

phase/position senstive

simple cell topography

orientation continuity, but not phase

orientation singularities, or “pinwheels”

“blob” - grouping of low-freq

complex cells physiology

orientation/freq selective

phase/position insensitive