Perception:
Inference, Priors and Codes

Maneesh Sahani

Gatsby Computational Neuroscience Unit, UCL

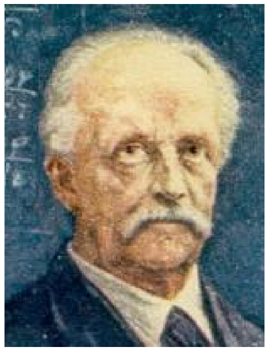
What is Perception *for*

- Control? (After all, the only point of having a brain is to move. . .)
- Forecasting and planning?
- Finding prey, mates, forage . . .

Presumably all of the above, but there is useful intermediate abstraction.

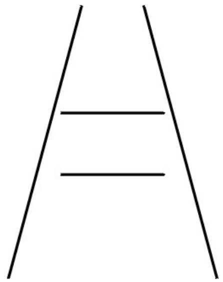
- *work out what's "out there".*

Helmholtz



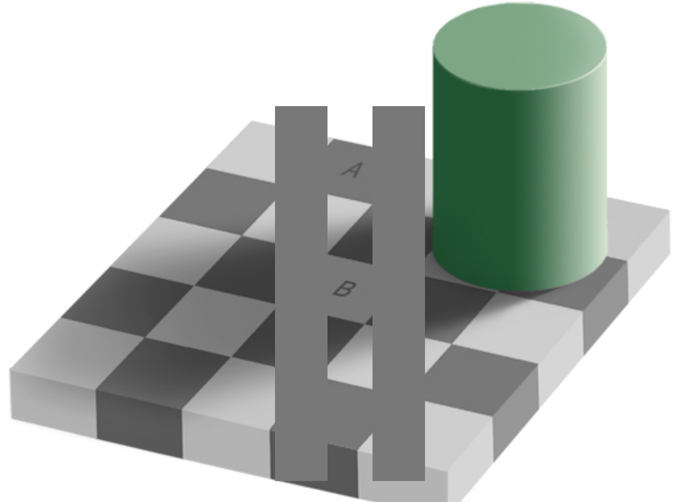
What information, then, can the qualities of such sensations give us about the characteristics of the external causes and influences which produce them? Only this: our sensations are **signs, not images**, *of such characteristics.*

Illusions



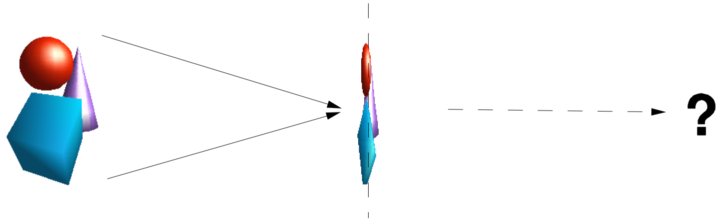
Gregory 1968

Illusions



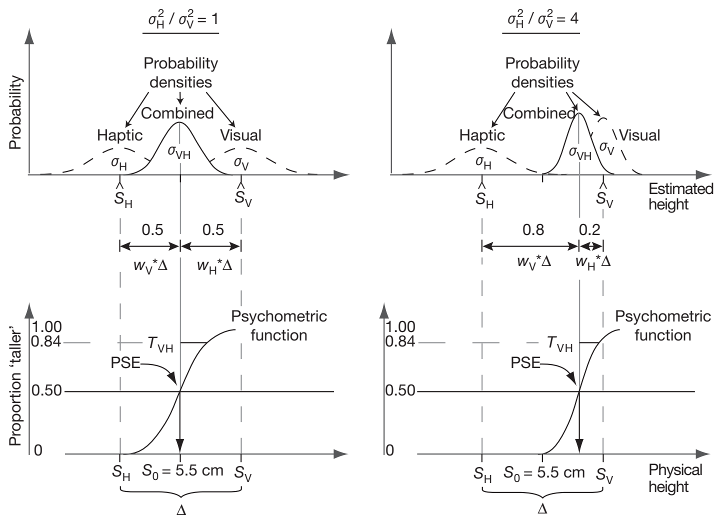
Edward H. Adelson

Perception and Generative Models



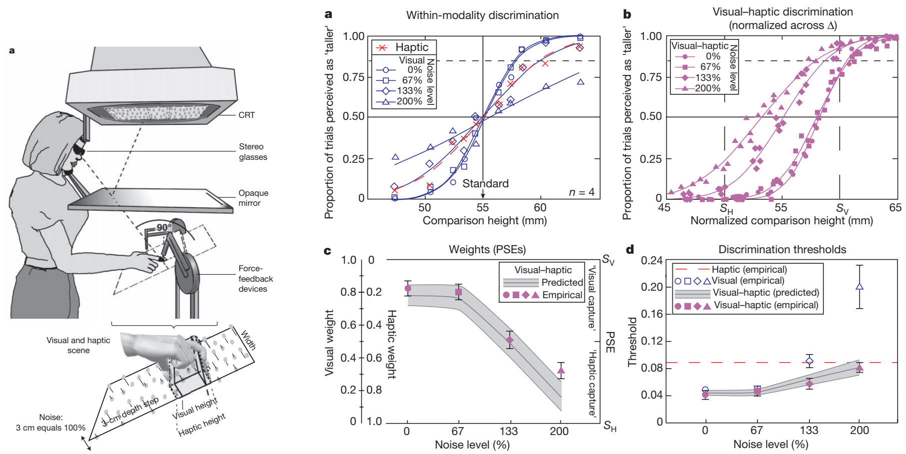
- Sensor activations reflect the state of the world through a (usually non-invertible and noisy) physical transformation.
- The goal of perception is to invert this transformation as best as possible: to infer the state of the world from the sensor signals.
- To do this, we need to know something about the forward (generative) process: both the transformation and the statistics of the world
- . . . and to use every available source of information.

Cue combination



Ernst & Banks 2002

Cue combination

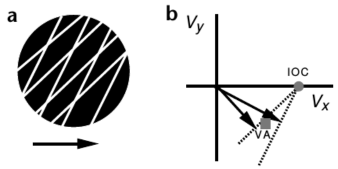


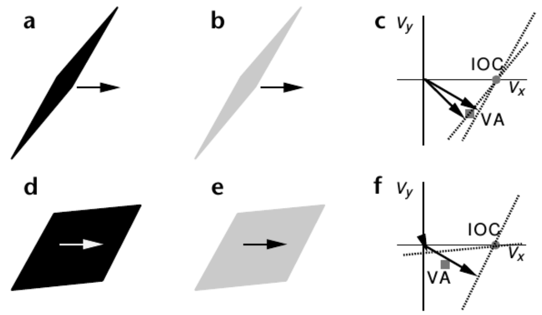
Ernst & Banks 2002

Incorporating priors – long-term priors

https://www.cs.huji.ac.il/~yweiss/Rhombus/rhombus.html

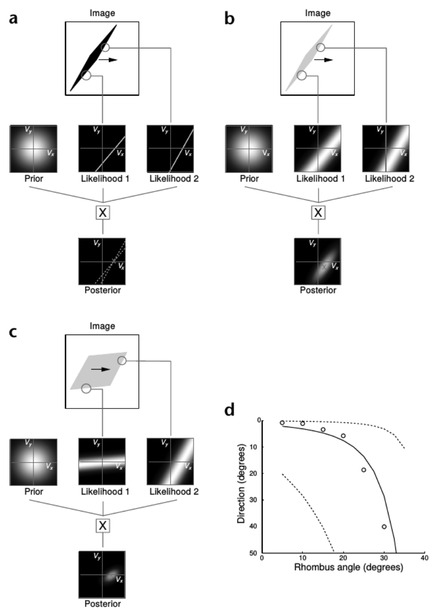
No simple rule





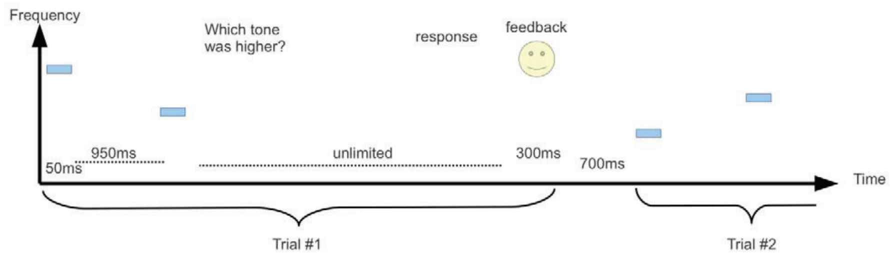
Weiss, Simoncelli, Adelson, 2002

Bayesian inference under a ’slow’ prior



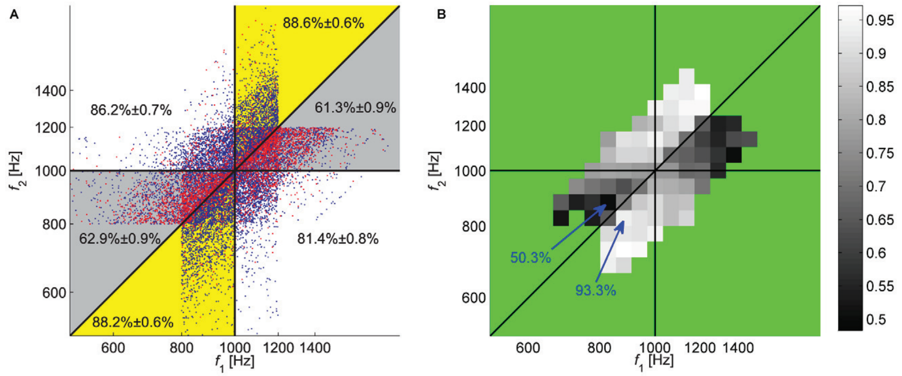
Weiss, Simoncelli, Adelson, 2002

Incorporating priors – short-term adaptation



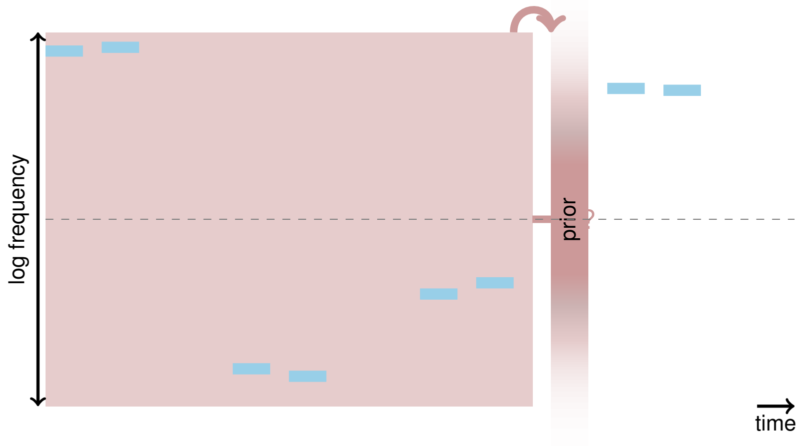
Raviv, Ahissar, Loewenstein, 2012

Frequency discrimination – contraction bias



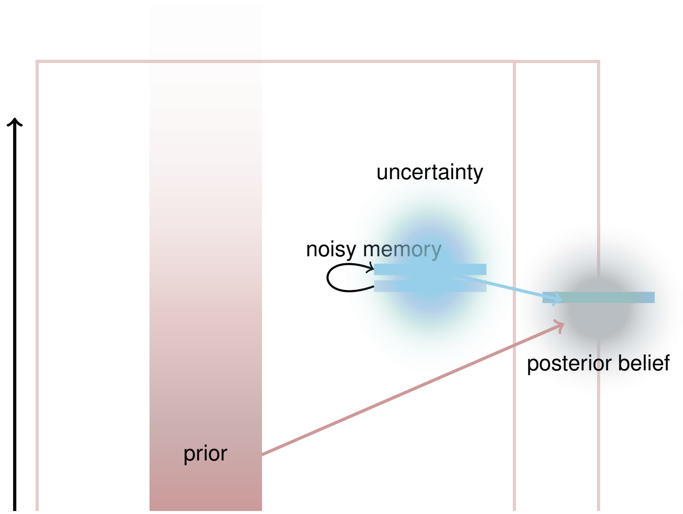
Raviv, Ahissar, Loewenstein, 2012

Prior context



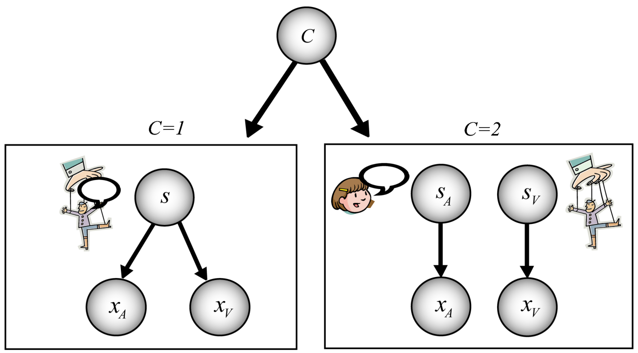
Ashourian, Loewenstein (2011)

Memory



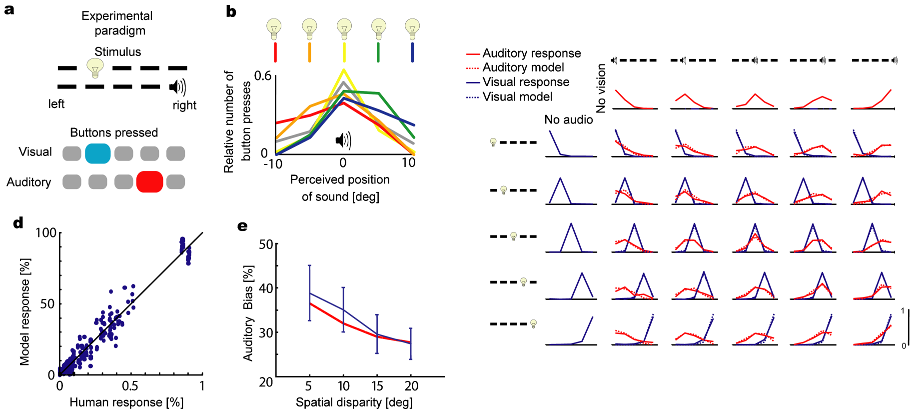
Ashourian, Loewenstein (2011)

Structured inference



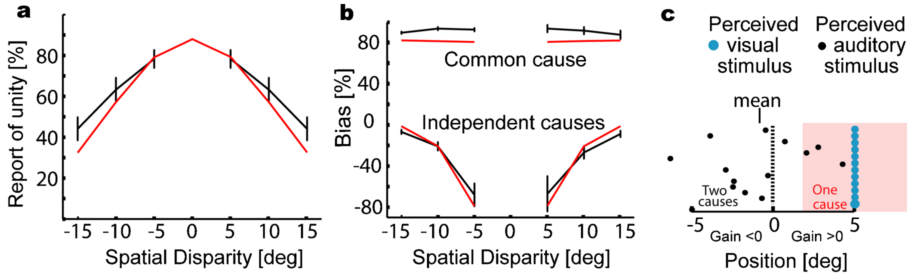
Körding, Beierholm, et al. 2007

Structured inference



Kördig, Beierholm, et al. 2007

Structured inference



Kördig, Beierholm, et al. 2007

Some neural consequences (in theory)

- *Sensory* systems (possibly for low-level control) should feed into *Perceptual* systems.
  - See Goodale & Milner on (visual) ventral and dorsal streams.
- Response properties and receptive fields in the perceptual pathway reflect properties of elements within an inferential system.
  - We should be able to predict those properties by fitting generative models to data.
  - Representations should to represent and manipulate uncertainties, priors and other elements of inference.

Physical vs. Generic Models

- If the physics is known and simple (or if evolution is lucky), it may be possible to invert the exact physical model. This will give the most accurate results.
  - Often difficult, particularly from an evolutionary standpoint.
  - Not flexible (e.g. if the statistics of the world change).
  - May be difficult to invert.
  - Neocortex appears to be generic.
- We consider the case where a generic generative model, with only some elements of physicality, is adapted through learning to describe the generative process in the world.

Inference and Learning

Latent variable model:

$$\mathrm{P}_\theta(\mathbf{y}_i) = \int d\mathbf{x}\, \mathrm{P}_\theta(\mathbf{y}_i \mid \mathbf{x})\, \mathrm{P}_\theta(\mathbf{x})$$

Inference (find $\mathbf{x}_i$ given $\mathbf{y}_i$ and θ):

$$\mathrm{P}_\theta(\mathbf{x}_i \mid \mathbf{y}_i) = \frac{\mathrm{P}_\theta(\mathbf{y}_i \mid \mathbf{x}_i)\, \mathrm{P}_\theta(\mathbf{x}_i)}{\mathrm{P}_\theta(\mathbf{y}_i)}$$

Learning (find θ given $\{\mathbf{y}\}$)

$$\mathrm{P}(\theta \mid \{\mathbf{y}\}) \propto \prod_i \mathrm{P}_\theta(\mathbf{y}_i)\, \mathrm{P}(\theta)$$

usually by ML approximation

$$\theta^* = \operatorname*{argmax}_\theta \prod_i \mathrm{P}_\theta(\mathbf{y}_i)$$

Unsupervised Learning

- Even if the ultimate goal is supervised or reinforcement learning, unsupervised learning can serve as a useful "front end" for finding good representations.
- Generative models provide an extremely successful framework for unsupervised learning.
- Other viewpoints, such as redundancy reduction, can be viewed as special cases of the generative modelling approach.

The Wake-Sleep Algorithm

- **Wake phase**: use recognition model for inference. Train generative weights by online gradient descent.
- **Sleep phase**: use generative model to create pseudo-data ("dreams"). Train recognition weights by online gradient descent.

Learning in Boltzmann Machines

$$\mathrm{P}(\{\mathbf{s}\} \mid W) = \frac{1}{Z} e^{-\sum_i E(\mathbf{s}_i; W)} \qquad Z = \int d\mathbf{r}\; e^{-E(\mathbf{r}; W)}$$

$$\begin{aligned}
\frac{\partial}{\partial W} \log \mathrm{P}(\{\mathbf{s}\} \mid W) &= -\sum_i \frac{\partial E(\mathbf{s}_i; W)}{\partial W} - \frac{N}{Z}\frac{\partial Z}{\partial W} \\
&= -\sum_i \frac{\partial E(\mathbf{s}_i; W)}{\partial W} - \frac{N}{Z}\int d\mathbf{r}\; \partial e^{-E(\mathbf{r}; W)} W \\
&= -\sum_i \frac{\partial E(\mathbf{s}_i; W)}{\partial W} + N \int d\mathbf{r}\; \underbrace{\frac{e^{-E(\mathbf{r}; W)}}{Z}}_{\mathrm{P}(\mathbf{r} \mid W)} \frac{\partial E(\mathbf{r}; W)}{\partial W} \\
&= -N \left\langle \frac{\partial E(\mathbf{s}; W)}{\partial W} \right\rangle_{\mathrm{P}_0(\mathbf{s})} + N \left\langle \frac{\partial E(\mathbf{s}; W)}{\partial W} \right\rangle_{\mathrm{P}_\infty(\mathbf{s} \mid W)}
\end{aligned}$$

$\mathrm{P}_0(\mathbf{s})$ is the data distribution. $\mathrm{P}_\infty(\mathbf{s} \mid W)$ is the usually the distribution of a Gibbs sampler.

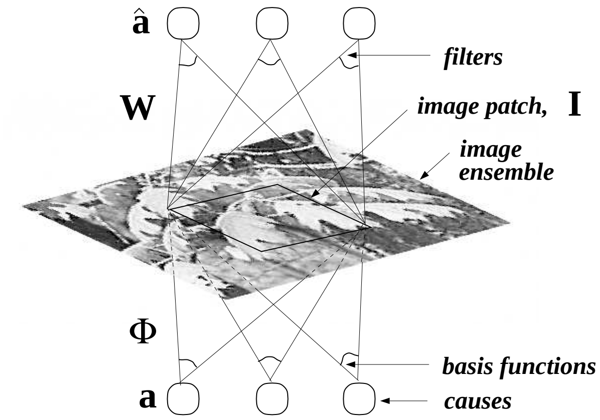
Constrastive Divergence

$$\Delta W \propto -\left\langle \frac{\partial E(\mathbf{s}; W)}{\partial W} \right\rangle_{\mathrm{P}_0(\mathbf{s})} + \left\langle \frac{\partial E(\mathbf{s}; W)}{\partial W} \right\rangle_{\mathrm{P}_n(\mathbf{s}|W)}$$

$\mathrm{P}_n(\mathbf{s} \mid W)$ is the distribution obtained by running a limited number, n, of Gibbs sampler iterations, starting at the observed data.

- Intuitively, try to avoid having the Markov chain leave the data distribution.
- Can be shown that this update is zero if(f) gradient is zero.
- Convergence does not seem to be guaranteed, but many experiments have shown good results.
- Useful in situations where energy can be easily calculated; e.g. product models where $\mathrm{P}(\mathbf{y} \mid \mathbf{x}) \propto \prod_i \mathrm{P}(\mathbf{y} \mid x_i)$ (such as the Boltzmann Machine).

Linear Image Codes

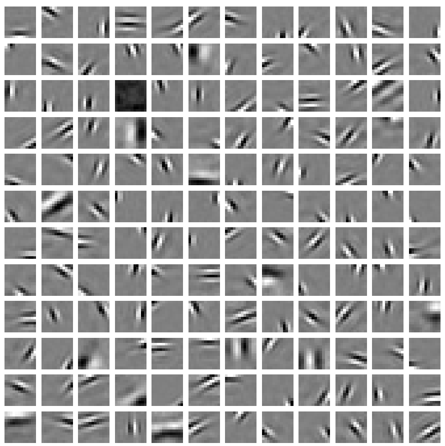


adapted from Bell and Sejnowski (1997)

Sparse Coding

$$E = \min_{\{a_i\}} \underbrace{\sum_{x,y} \left[I(x,y) - \sum_i a_i \phi_i(x,y) \right]^2}_{\log \mathrm{P}(Y \mid X)} + \underbrace{\lambda \sum_i S(a_i)}_{\log \mathrm{P}(X)}$$

$$S(a) = \log(1 + (a/\sigma)^2)$$

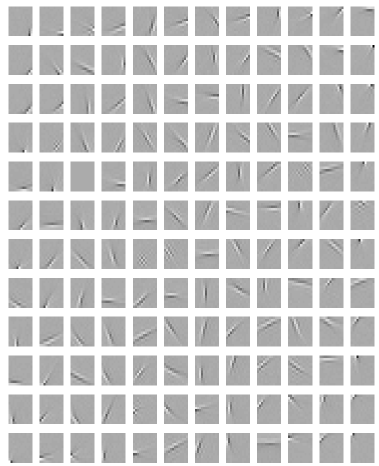


Olshausen & Field (1996)

Infomax

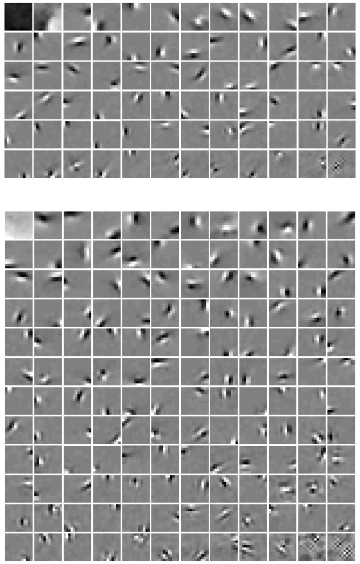
$$E = -H\left[g\left(\sum_{x,y} W_i(x,y) I(x,y) \right) \right]$$

$$g(a) = \frac{1}{1 + e^{-a}}$$



Bell & Sejnowski (1997)

Overcompleteness

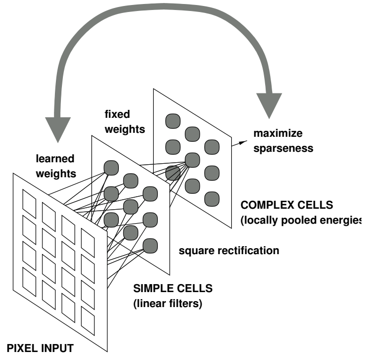


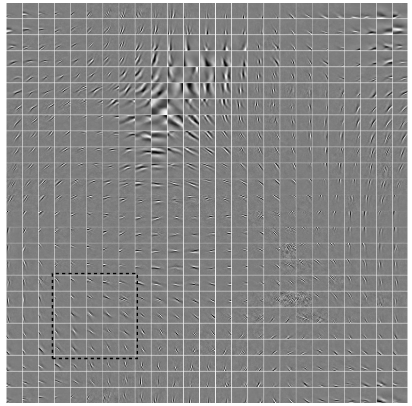
$$E = -\int d\mathbf{a}\, \mathrm{P}_\phi\,(I \mid \mathbf{a})\, \mathrm{P}_S\,(\mathbf{a})$$

(Integral is approximated by saddle-point method.)

Lewicki & Sejnowski (2000); Lewicki & Olshausen (1999)

Topographic ICA - Hyvärinen & Hoyer

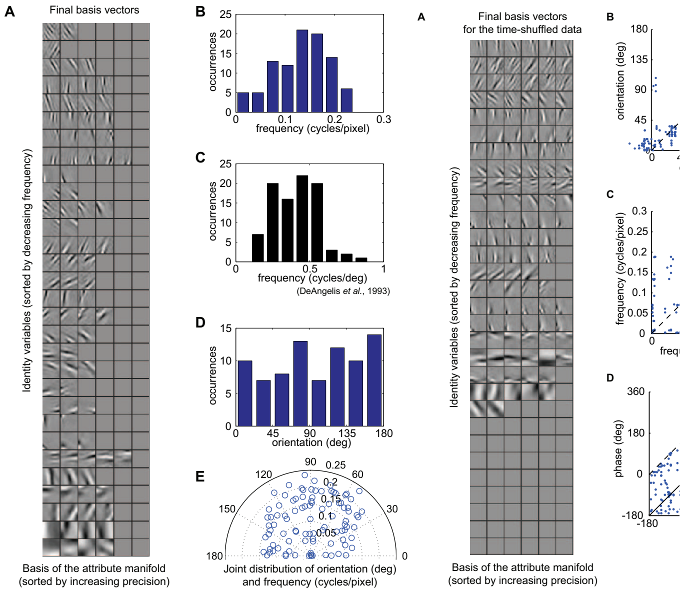




Hyvarinen & Hoyer (2001)

Dynamic constancy

- Dynamic images and latent variables $I(x, y, t) \Rightarrow a_i(t)$.
- Impose prior limiting change in $a_i(t)$.
- With suitably constrained models, results in phase insensitivity (complex cells).



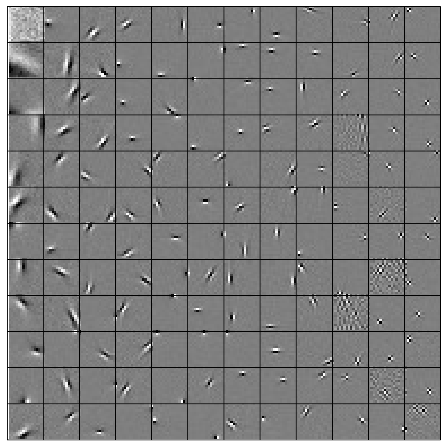
Wiskott & Sejnowski; Körding et al.; Berkes, Turner & Sahani

Recognition models

$$\mathrm{P}\,(I(x,y)) = \frac{e^{-E(\hat{\mathbf{a}})}}{\int d\mathbf{b}\; e^{-E(\mathbf{b})}}$$

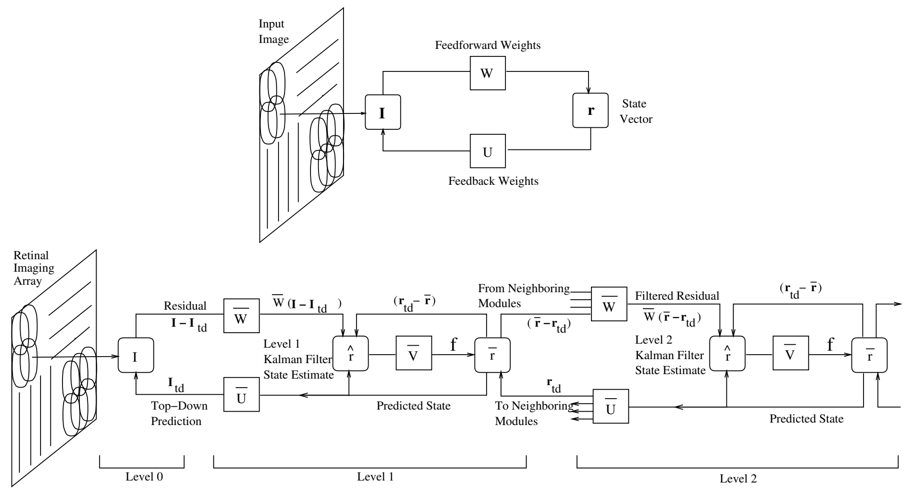
$$E(\hat{\mathbf{a}}) = -\sum_i \log \mathrm{P}_i\,(\hat{a}_i)$$

$$\hat{a}_i = \sum_{x,y} W(x,y) I(x,y)$$



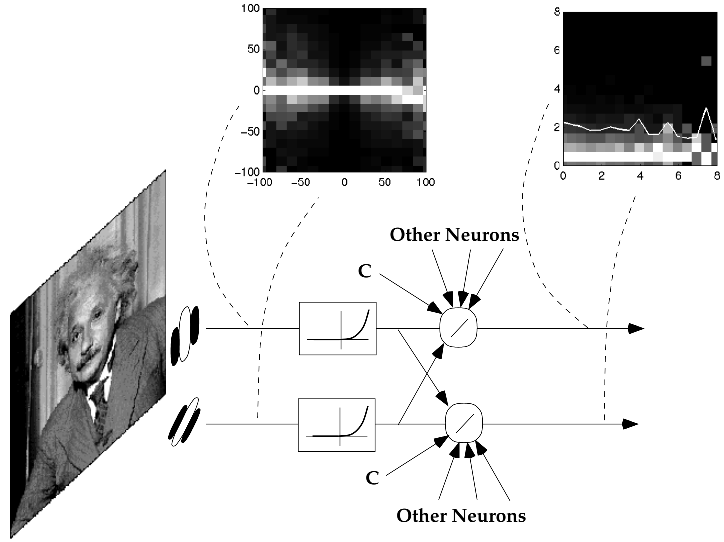
Hinton, Welling, Teh & Osindero (2002)

Feedback cancellation



Rao & Ballard (1997) (cf Friston)

Lateral normalization



Wainwright, Schwartz, & Simoncelli 2001