A three-layer model of natural image statistics

Michael U. Gutmann
Dept of Mathematics and Statistics
University of Helsinki
michael.gutmann@helsinki.fi

The presentation is based on the paper:

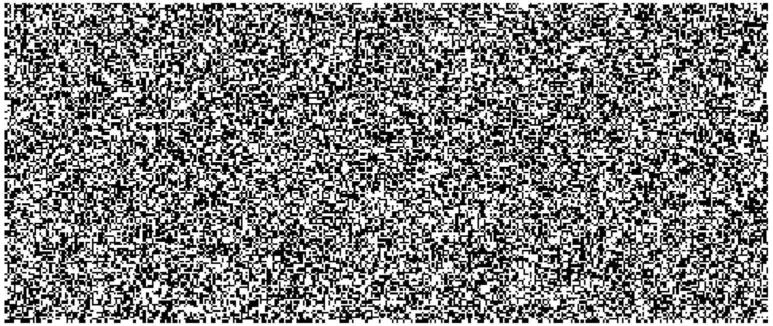
M. Gutmann and A. Hyvärinen,
A three-layer model of natural image statistics,
Journal of Physiology-Paris, 2013, in press.

Introduction

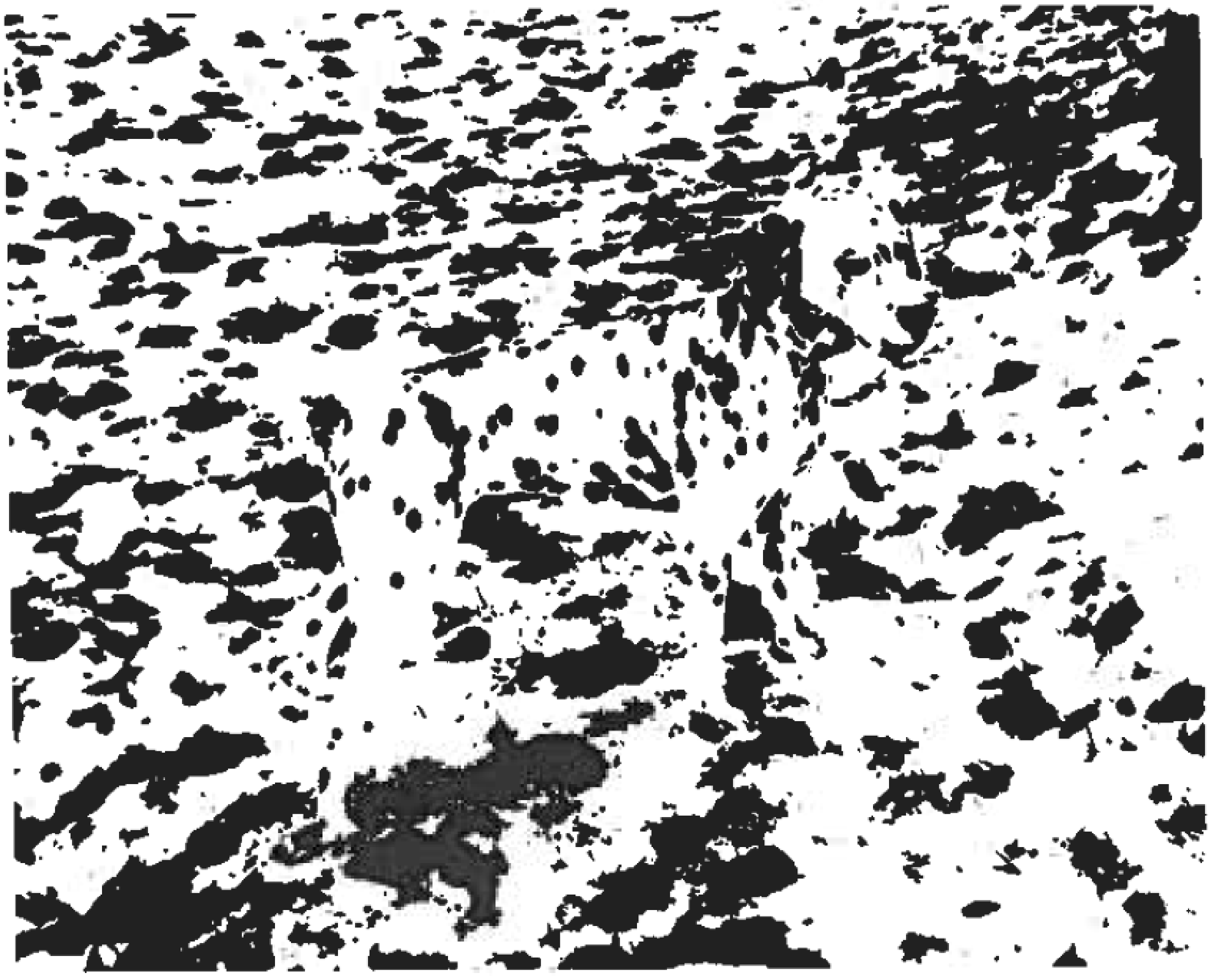
Natural scenes contain regularities

"Apgar 10/10; Feet", by Jacquelyn Berl.

- Dimensions of the image: 150×360. (54000 pixels).
- There are $2^{54000} > 10^{16000}$ different binary 150×360 images.
- Only a very small fraction depicts scenes that we may see in our natural environment.



The regularities are used by the visual system



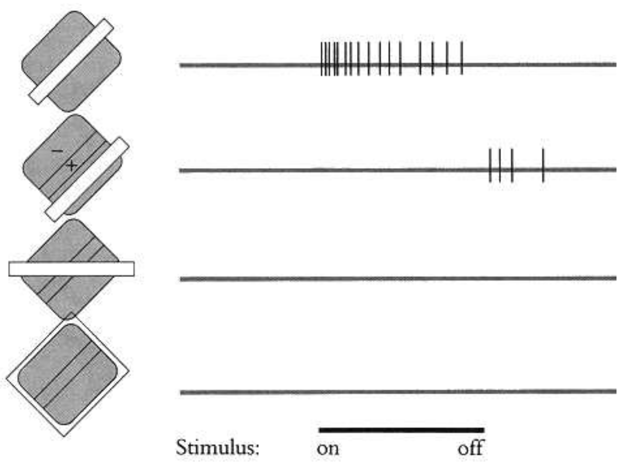
They serve as prior information in perception.

Natural environment and the brain

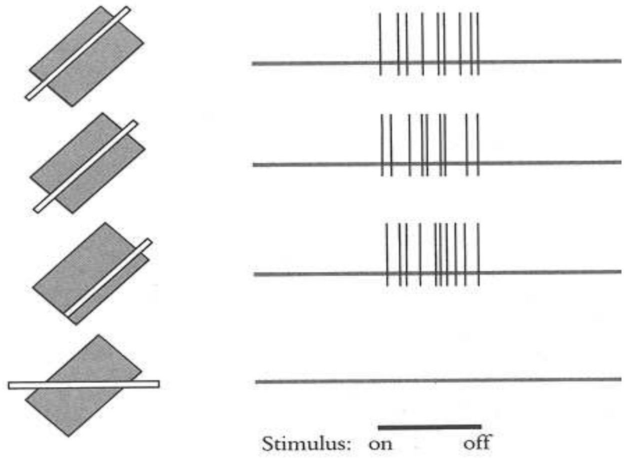
- Natural scenes contain a lot of structure (regularities).
- Basic assumption: The sensory system is adapted to its sensory environment (ecological adaptation).
- Research topic in general: Relate properties of the natural environment to properties of the sensory (visual) system.
- This talk: Its relation to neural selectivity and invariance (tolerance).

Neural selectivity and tolerance

- Some “definitions” of neural selectivity and tolerance:
  - Neurons are selective to certain properties of the stimulus if their response increases strongly when the stimulus properties become present.
  - Neurons are tolerant to them if their response does not change much.
- Example for cells in the primary visual cortex:



Simple cells:
Selective to orientation
and location of the bar



Complex cells:
Tolerant to exact location

Tolerant selectivities

- Combining selectivity with tolerance (tolerant selectivities) is helpful in higher visual tasks.
- Example: To recognize a face, we need to find visual clues that are
  - specific for the person at hand (selectivity), and
  - somewhat invariant to the facial expressions (tolerance).

(Figure from “Facial Expressions – A Visual Reference for Artists” by M. Simon.)

Emergence of higher-level tolerant selectivities (1/3)

- Basic hypothesis: Higher-level tolerant selectivities emerge through a sequence of elementary selectivity and tolerance computations.
- Hypothesis goes back to Kunihiko Fukushima's "neocognitron", which is a multi-layer extension of Hubel& Wiesel's simple-cell, complex-cell cascade.

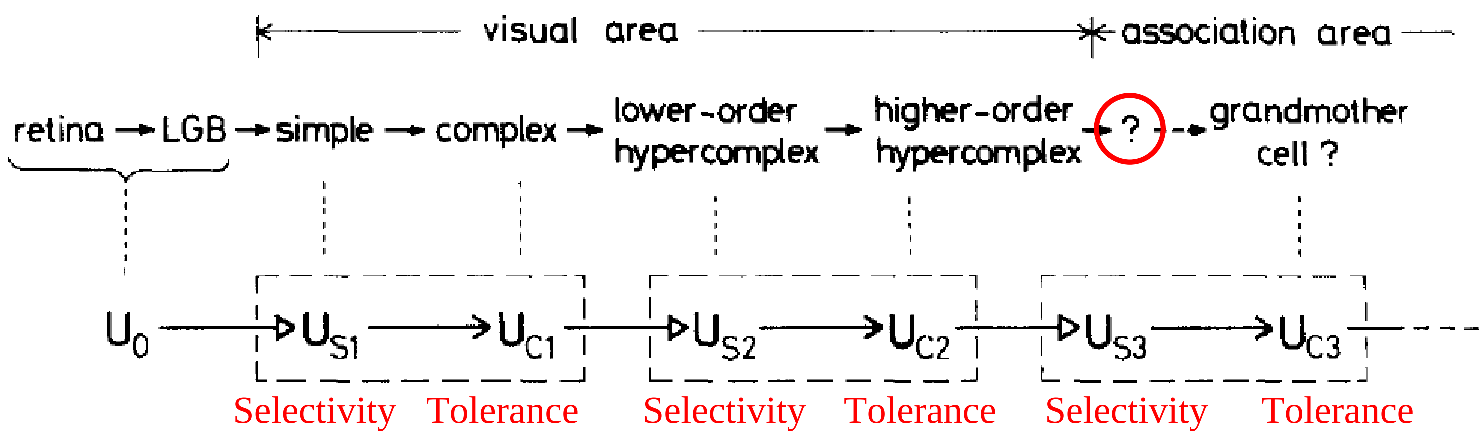


Figure adapted from "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", Biol Cybernetics, 1980.

Emergence of higher-level tolerant selectivities (2/3)

Similar idea was put forward by Riesenhuber and Poggio, Nature 1999, and others.

Representation that is suitable for object recognition tasks

Sequence of elementary computations

Tolerance

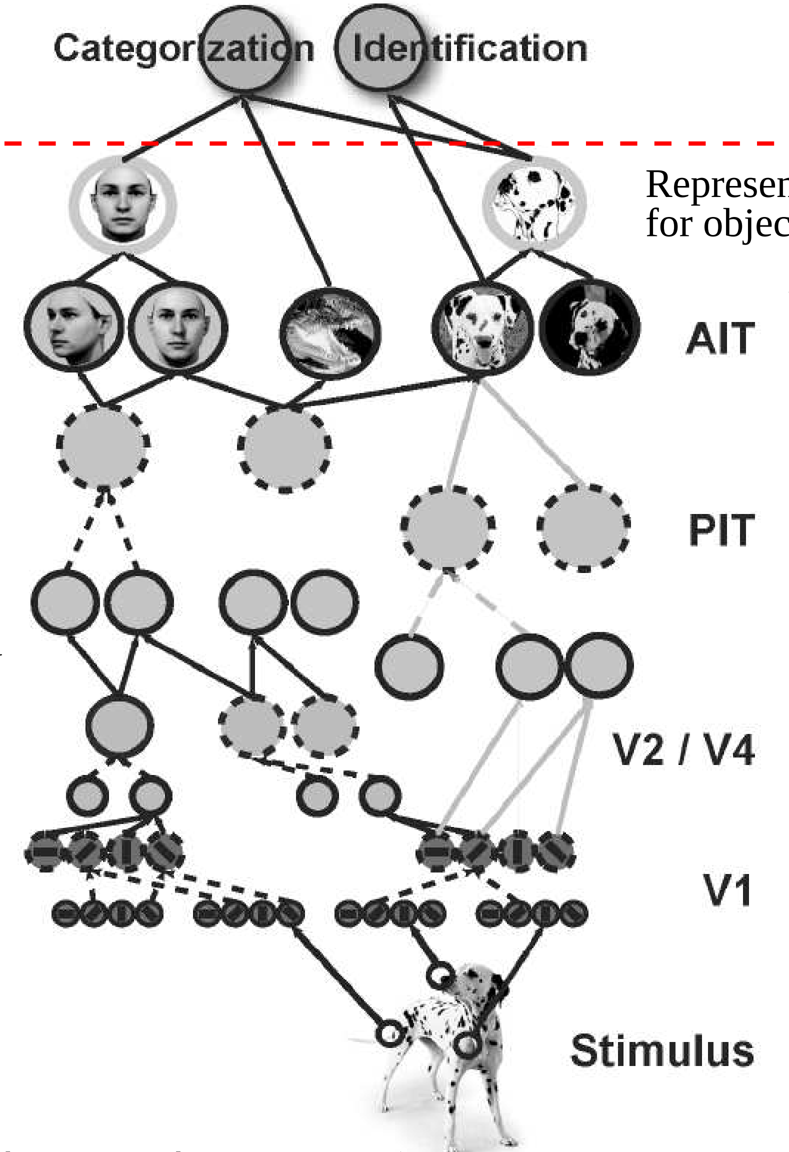
Selectivity

Tolerance

Selectivity

Tolerance

Selectivity



Reformatting of the neural representation of the stimulus

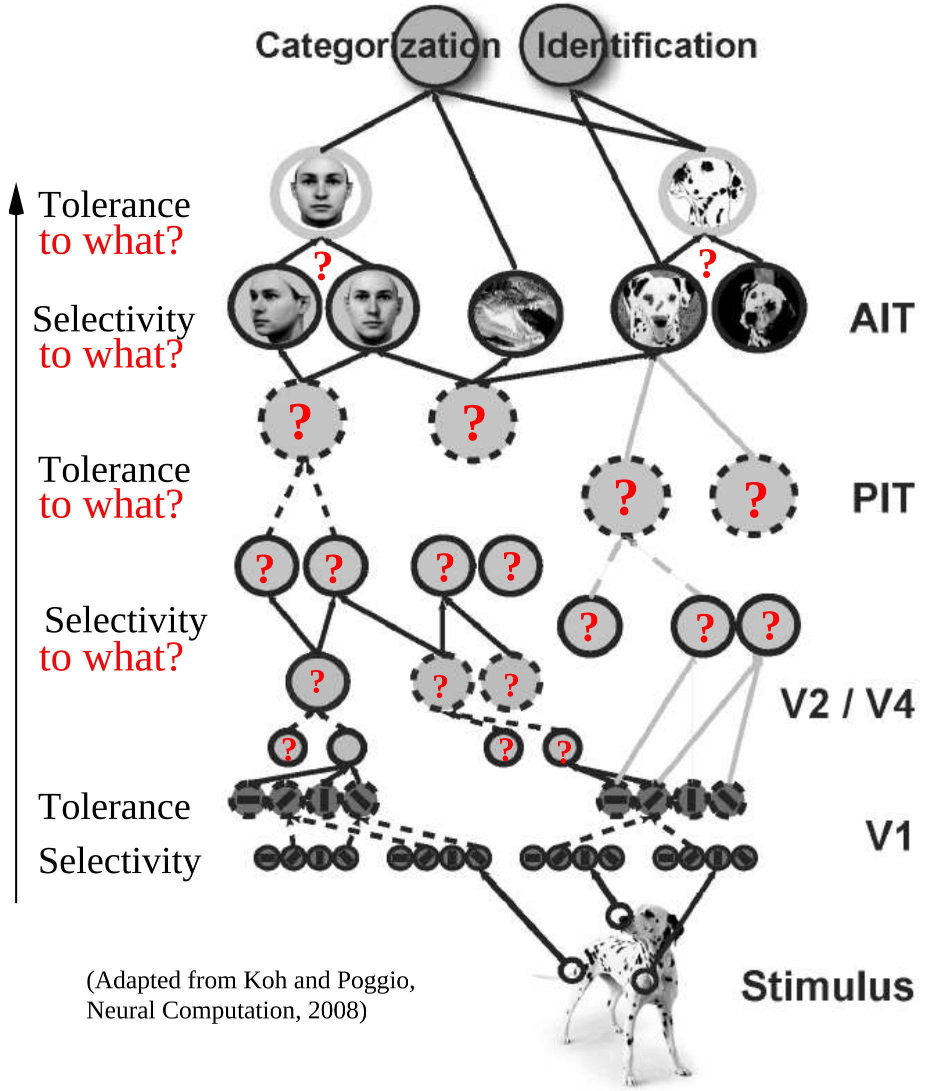
(Adapted from Koh and Poggio, Neural Computation, 2008)

Emergence of higher-level tolerant selectivities (3/3)

- There is (indirect) experimental evidence for an increase in selectivity and tolerance along the ventral pathway

  Rust and DiCarlo, J. Neurosci., 2010

- What remains poorly understood is the nature of the tolerance and selectivity computations along the hierarchy.



Question asked and methodology

- Basic hypothesis:
  Higher level tolerant selectivities emerge through a sequence of elementary selectivity and invariance computations.
- Question asked:
  In a visual system with three processing layers, what should be selected and tolerated at each level of the hierarchy?
- Methodology: Learn the selectivity and invariance computations from natural images.

  Learning = fitting a statistical model to natural image data.

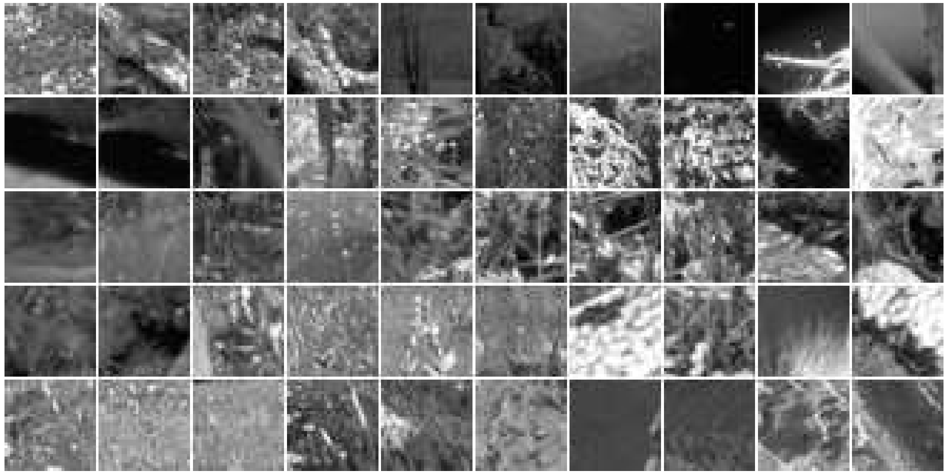
Methods

Data

We learn the computations for two kinds of image data sets:

1. Image patches of size 32 by 32, extracted from larger images (left).
2. "Tiny images" dataset, converted to gray scale: complete scenes downsampled to 32 by 32 images (right)

   (Torralba et al, TPAMI 2008)



The three processing layers (1/2)

- Let $\mathbf{x}$ be a vectorized image after preprocessing (luminance and contrast gain control, low-pass filtering).
- The three processing layers are:

$$y_i^{(1)} = \text{max}\left(\mathbf{w}_i^{(1)} \cdot \mathbf{x}, 0\right), \qquad i = 1 \dots 600$$

$$y_i^{(2)} = \ln\left(\mathbf{w}_i^{(2)} \cdot (\mathbf{y}^{(1)})^2 + 1\right), \qquad i = 1 \dots 100$$

$$\mathbf{z}^{(2)} = \text{gain control}\left(\mathbf{y}^{(2)}\right),$$

$$y_i^{(3)} = \text{max}\left(\mathbf{w}_i^{(3)} \cdot \mathbf{z}^{(2)}, 0\right), \qquad i = 1 \dots 50$$

  Gain control is similar to the preprocessing: centering, normalizing the norm after whitening, possibly dimension reduction
- Free parameters: $\mathbf{w}_i^{(1)}$, $\mathbf{w}_i^{(2)}$, $\mathbf{w}_i^{(3)}$. They govern the computations of the three layers.
- Constraint: the $\mathbf{w}_i^{(2)}$ have nonnegative elements, $w_{ki}^{(2)} \geq 0$.

The three processing layers (2/2)

- First and third layer: $y_i^{(1)} = \max\left(\mathbf{w}_i^{(1)} \cdot \mathbf{x}, 0\right)$

  Linear projection followed by rectification.
  This is a (very) simple model for the steady-state firing rate of neurons.
- Second layer: $y_i^{(2)} = \ln\left(\mathbf{w}_i^{(2)} \cdot (\mathbf{y}^{(1)})^2 + 1\right)$

  Functional form of the energy model for complex cells (Adelson, J Opt Soc Am, A, 1985)
- Linear projections/pooling patterns are not yet specified, but learned from the data.
- The outputs $y_i^{(1)}, y_i^{(2)}, y_i^{(3)}$ are used to define the statistical model (probability density function) of the natural images. (see paper for details)
- Fitting the model allows us to learn the parameters $\mathbf{w}_i^{(1)}$, $\mathbf{w}_i^{(2)}$, $\mathbf{w}_i^{(3)}$.

Results

Computations on the first two layers (in brief)

$$y_i^{(2)} = \ln\left(\sum_k w_{ki}^{(2)}(\mathbf{w}_k^{(1)} \cdot \mathbf{x})^2 + 1\right)$$

- First layer: Selectivity to localized oriented ("Gabor-like") image structure. ("simple cells", similar to prev work)
- The learned computation on the second layer resembles a max operation over selected first-layer outputs.
- Second layer: Selectivity to localized oriented image structure. Tolerance to exact localization. ("complex cells", similar to prev work)

Layer three: example unit for patch data

$$\mathbf{z}^{(2)} = \text{gain control}\left(\mathbf{y}^{(2)}\right) \qquad y_i^{(3)} = \max\left(\mathbf{w}_i^{(3)} \cdot \mathbf{z}^{(2)}, 0\right)$$

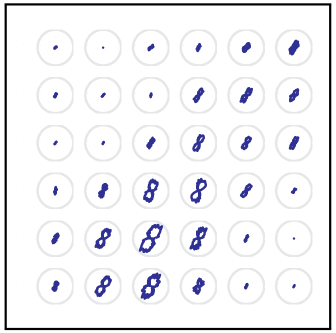
- Black frame: space-orientation receptive field. Visualizes the response to local gratings of different orientations.

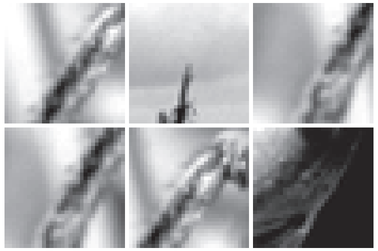
  (Anzai et al, Neurons in monkey visual area V2 encode combinations of orientations, Nat Neurosci, 2007)

- Red frame: “inhibitory” space-orientation receptive field. Shows the location and orientation of local gratings which inhibit the units most.

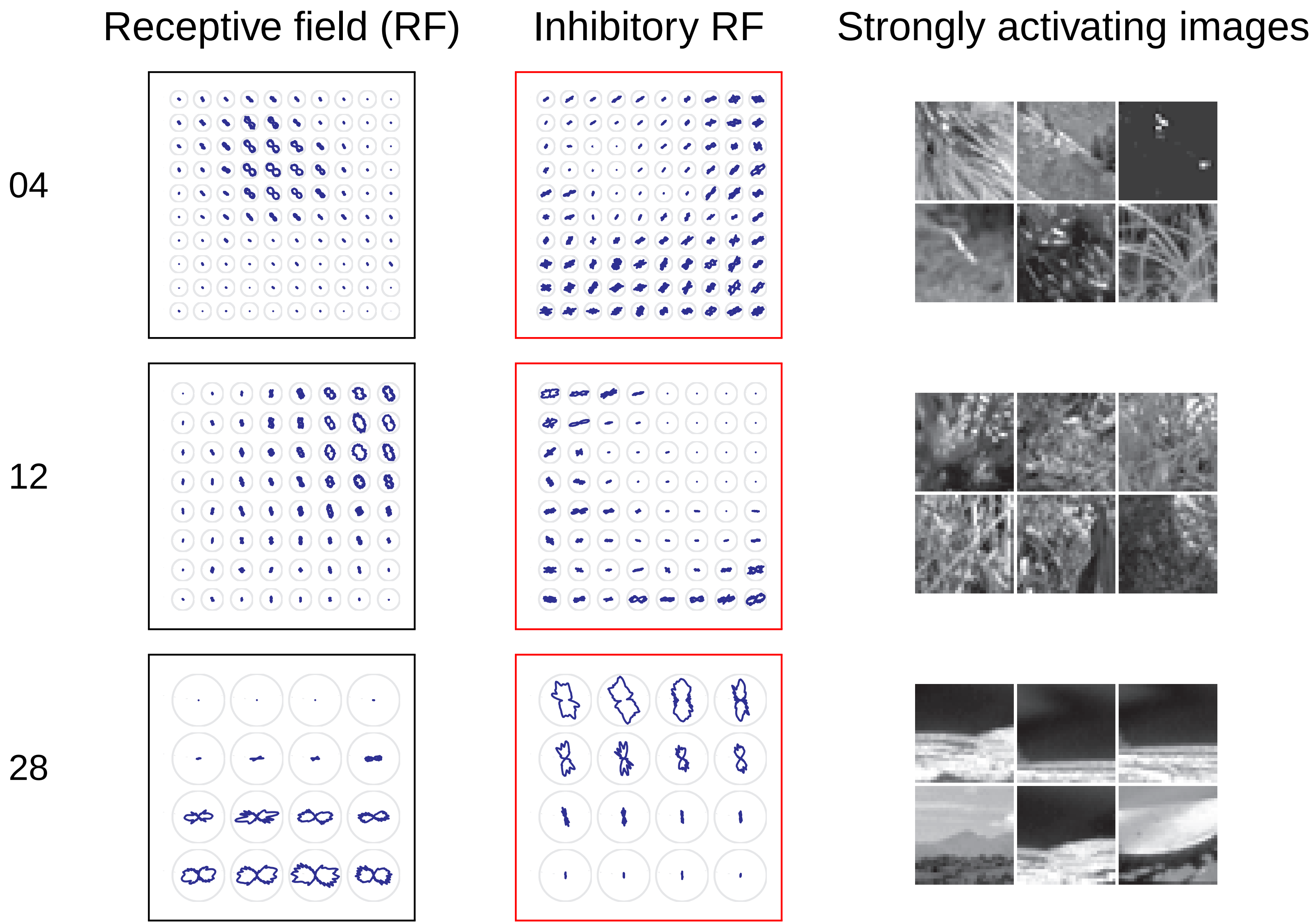
Receptive field (RF) Inhibitory RF Strongly activating images

07

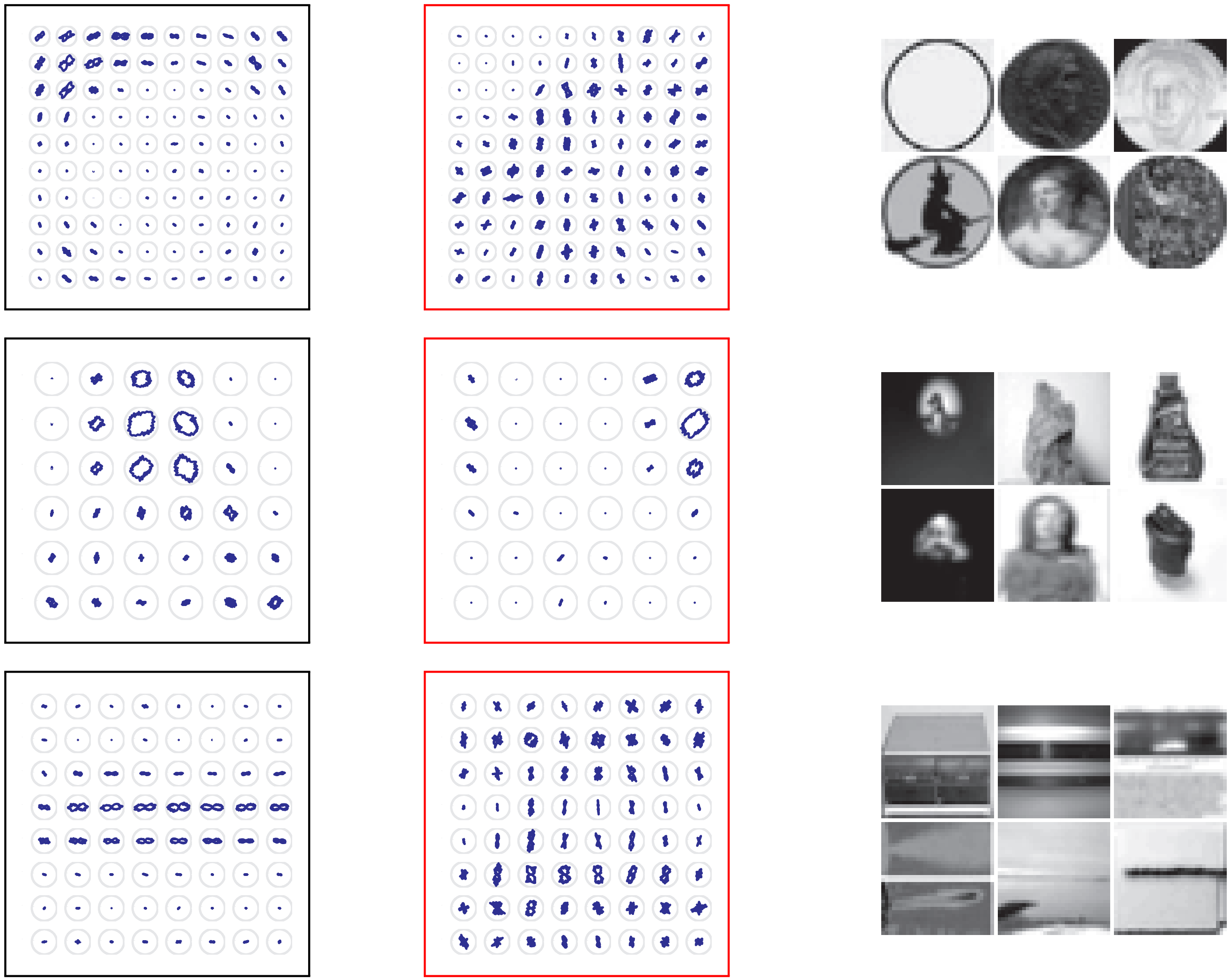




Layer three results: more examples for patch data



Layer three results: examples for tiny image data

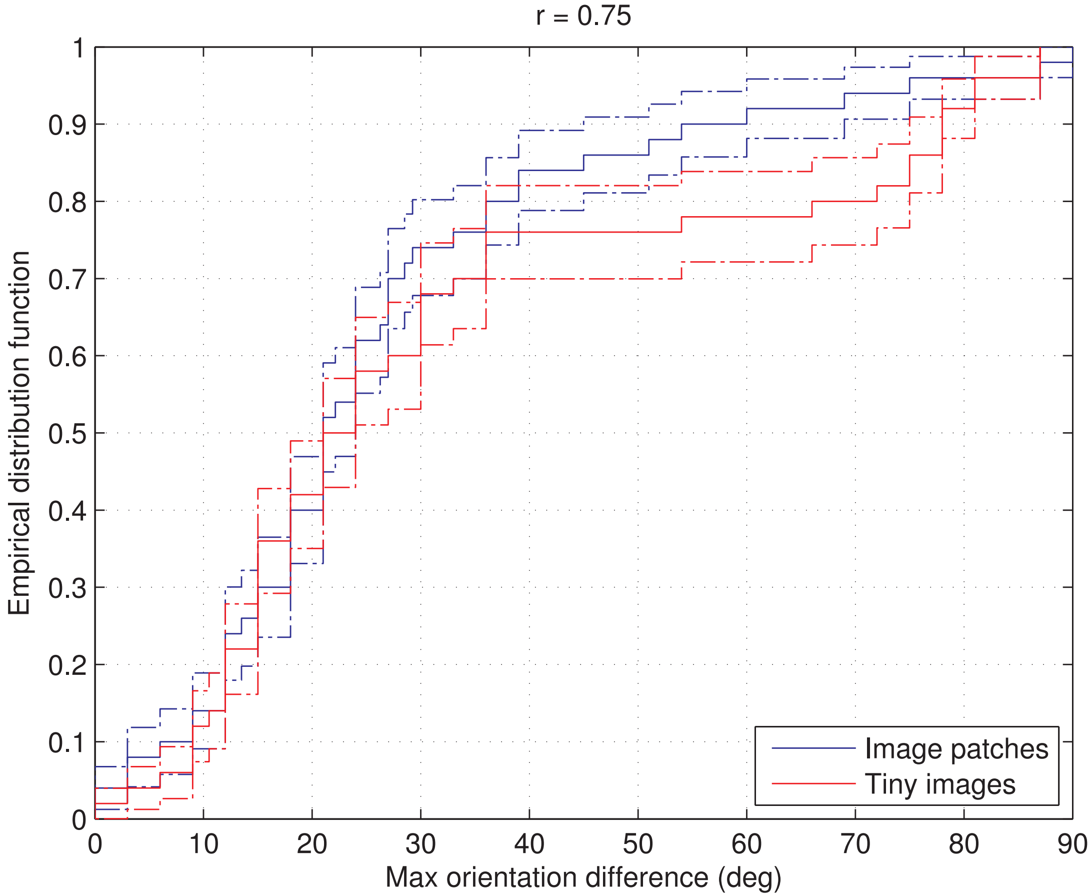


Qualitative observations

- Receptive fields are well structured and often localized.
- Emergence of non-classical receptive fields.
- For tiny images, the receptive fields are more inhomogeneous than for patch data.
- Excitatory and inhibitory gratings form large angles (orientation inhibition).
- Selectivity on the third layer:
  - For patch data: longer contours and texture
  - For tiny images: longer contours, curvatures

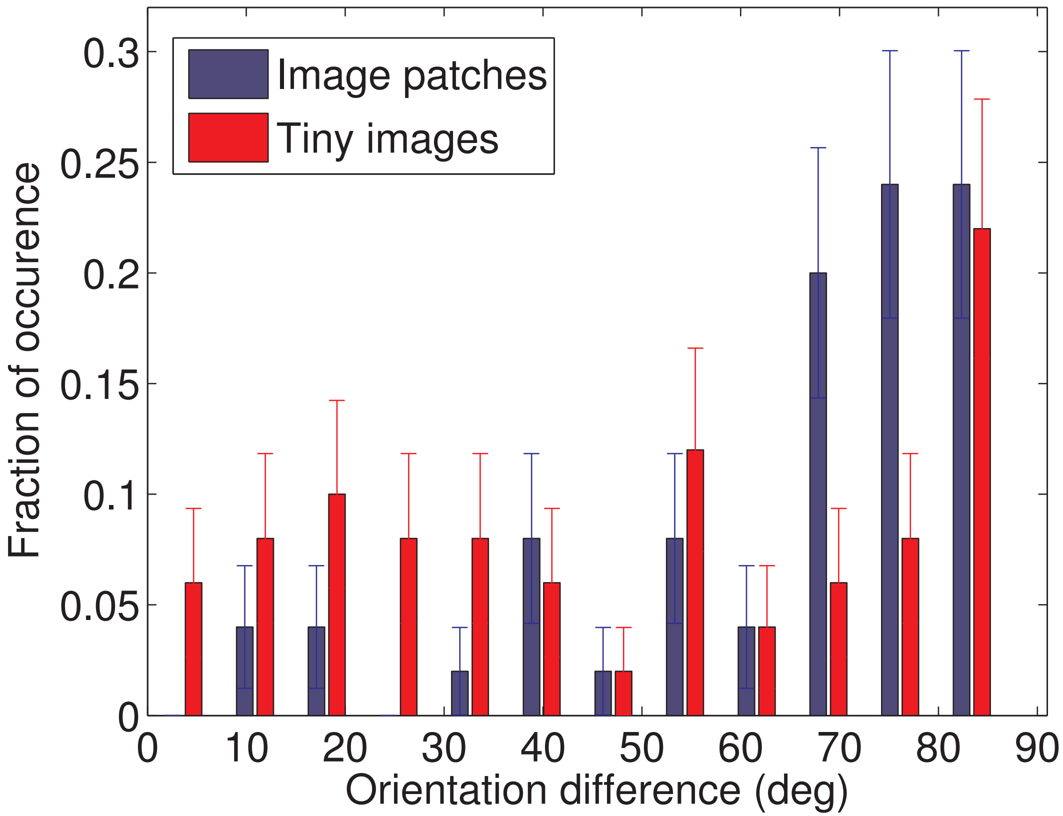
Population analysis of homogeneity

- Maximal difference δ in orientation tuning within a RF on L3:
  $\delta < 30^\circ : 70\%; \quad \delta > 60^\circ : 10\%$ (patches), 20% (tiny images)
- Experimental findings (V2 in Macaque monkeys):
  - Anzai, 2007: $\delta < 30^\circ : 60 - 70\%; \quad \delta > 60^\circ : 30\%$
  - Tao, 2012: $\delta < 30^\circ : 80\%; \quad \delta > 60^\circ : 5\%$



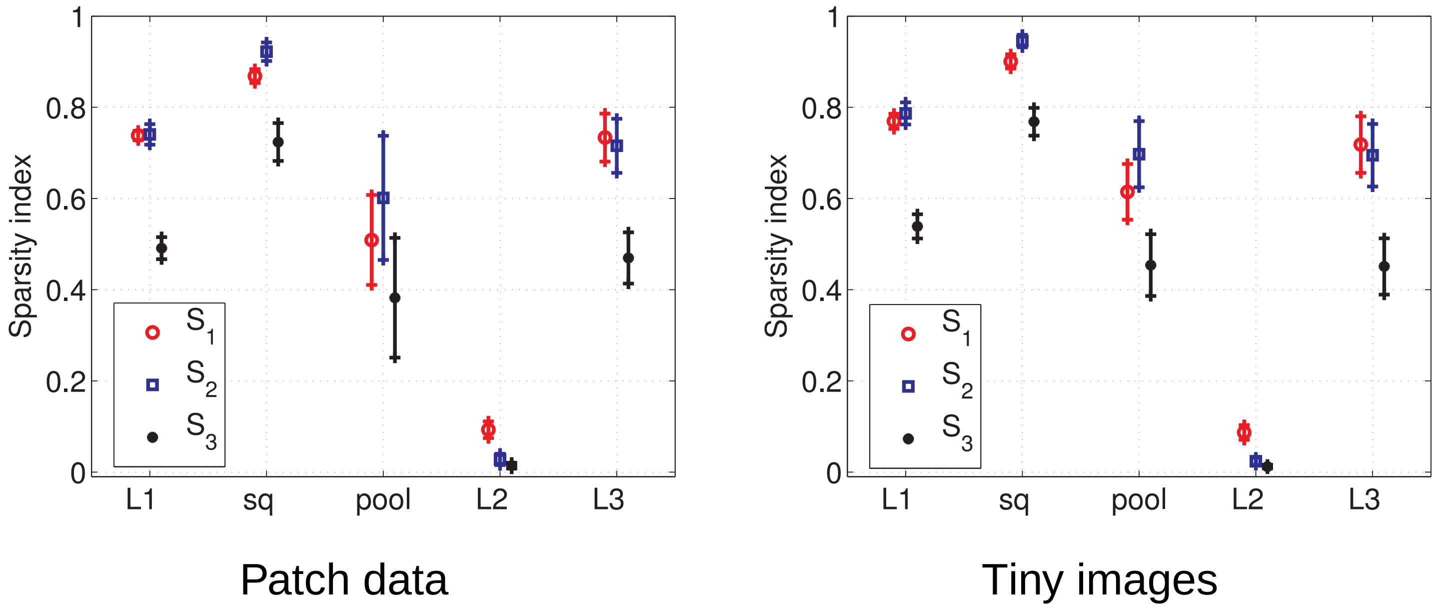
Population analysis of orientation inhibition

- We computed the angle between preferred and least preferred orientation for all third-layer units.
- The mode of the distribution is at $83^\circ \pm 7^\circ$.
- Strongest inhibition occurs for local gratings which are (roughly) orthogonal to the preferred orientation.



Lifetime sparsity across the three layers

- We use three different indices S_1, S_2, S_3 to measure lifetime sparsity (see paper for details).
- Sparsity on layer one ("L1") and three ("L3") are about the same.
- Squaring ("sq") increases sparsity. Pooling ("pool") and taking the logarithm ("L2") reduces it.
- Iterating between selectivity and tolerance computations balances sparsity (no net increase).



Conclusions

What the talk was about

- Basic hypothesis of our work is:
  Higher level tolerant selectivities emerge through a sequence of elementary selectivity and invariance computations.
- We asked:
  In a visual system with three processing layers, what should be selected and tolerated at each level of the hierarchy?
- Our approach was:
  Learn the selectivity and invariance computations from natural images by fitting a statistical model.

What we found

- Computations in the first two layers are in line with previous research. For both patch data and tiny images:
  - First layer: Emergence of selectivity to Gabor-like image structure (“simple cells”)
  - Second layer: Emergence of tolerance to exact orientation or localization of the stimulus (“complex-cells”)
- Computations on the third layer:
  - Patch data: Emergence of selectivity to longer contours and, to some extent, texture.
  - Tiny images: Emergence of selectivity to longer contours and, to some extent, curvature.
  - The receptive fields are mostly homogeneous, in line with experimental results. They are more inhomogeneous for tiny images than for patch data.
  - Emergence of (orientation) inhibition to facilitate the selectivity computations.
- No net increase of sparsity as we go from layer one to layer three.