Object recognition

Hierarchical models of object recognition

Suggested reading:

- Fukushima, K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36:193-202.
- Riesenhuber, M, Poggio, T (1999) Hierarchical models of object recognition in cortex, Nat. Neurosci. 2:1019-1025.
- Serre, T, Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines. PhD Thesis, MIT, 2006.

Four stages of representation (Marr, 1982)

1) pixel-based (light intensity)

2) primal sketch (discontinuities in intensity)

3) 2D sketch (oriented surfaces, relative depth between surfaces)

4) 3D model (shapes, spatial relationships, volumes)

Problem: computationally intractable!

Recognition by Components (Biederman, 1987)

Structural approach to object recognition:

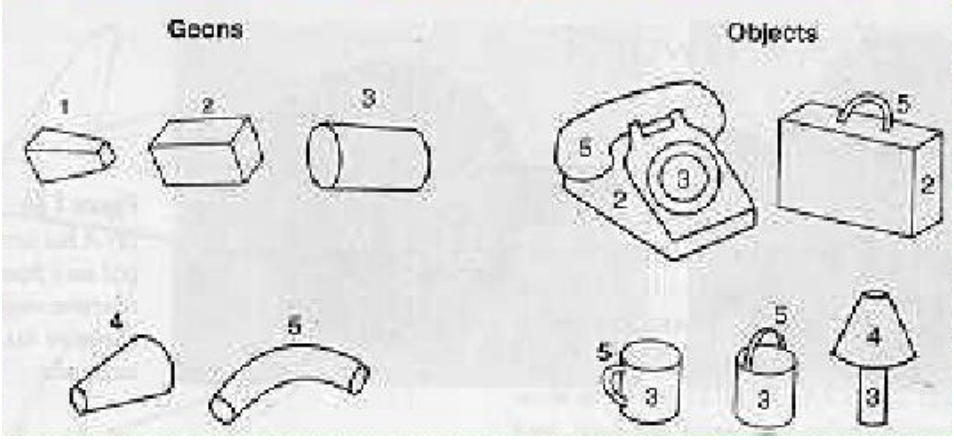
- Complex objects are composed to simpler pieces
- Recognize a novel/unfamiliar object by parsing it in terms of its component pieces, then comparing the assemblage of pieces to those of known objects.

A computational model does not exist !

Recognition by Components

- ***GEONS:*** geometric elements of which all objects are composed (cylinders, cones, etc). On the order of 30 different shapes.

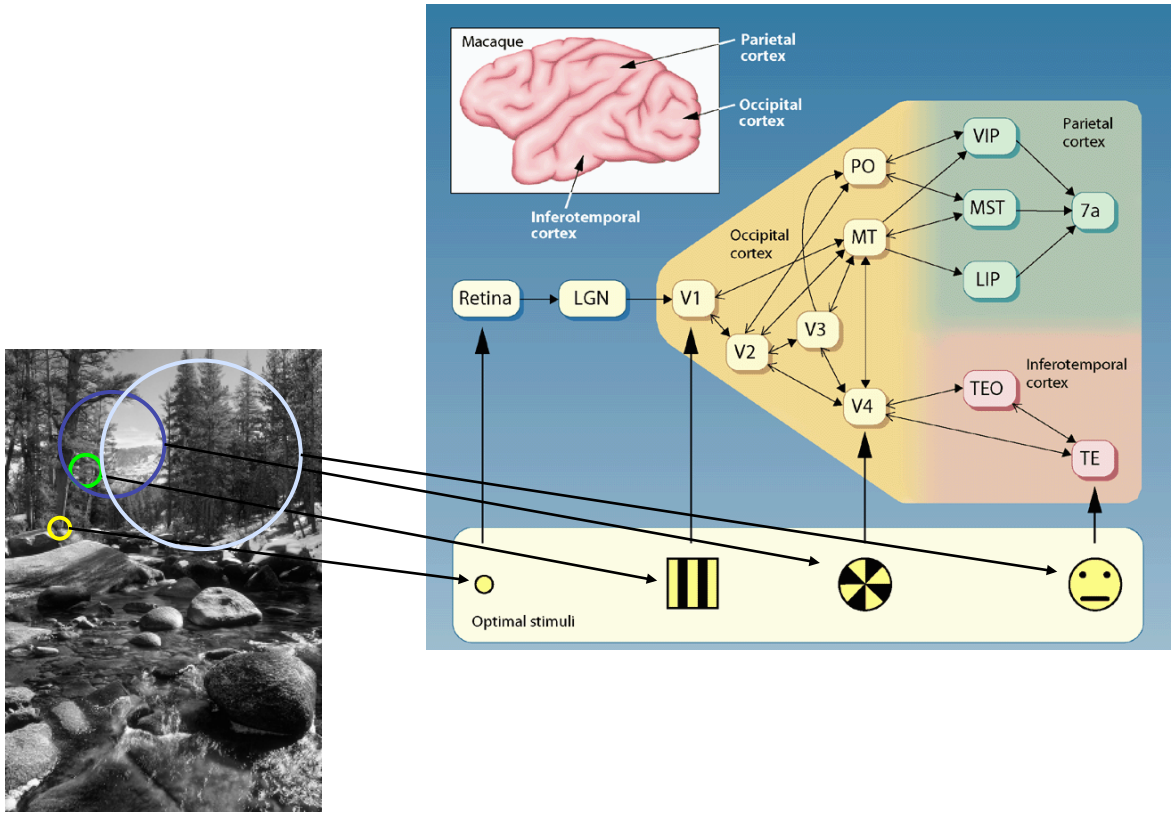
- Skips 2D sketch: Geons are directly recognized from edges, based on their nonaccidental properties (i.e., 3D features that are usually preserved by the projective imaging process).

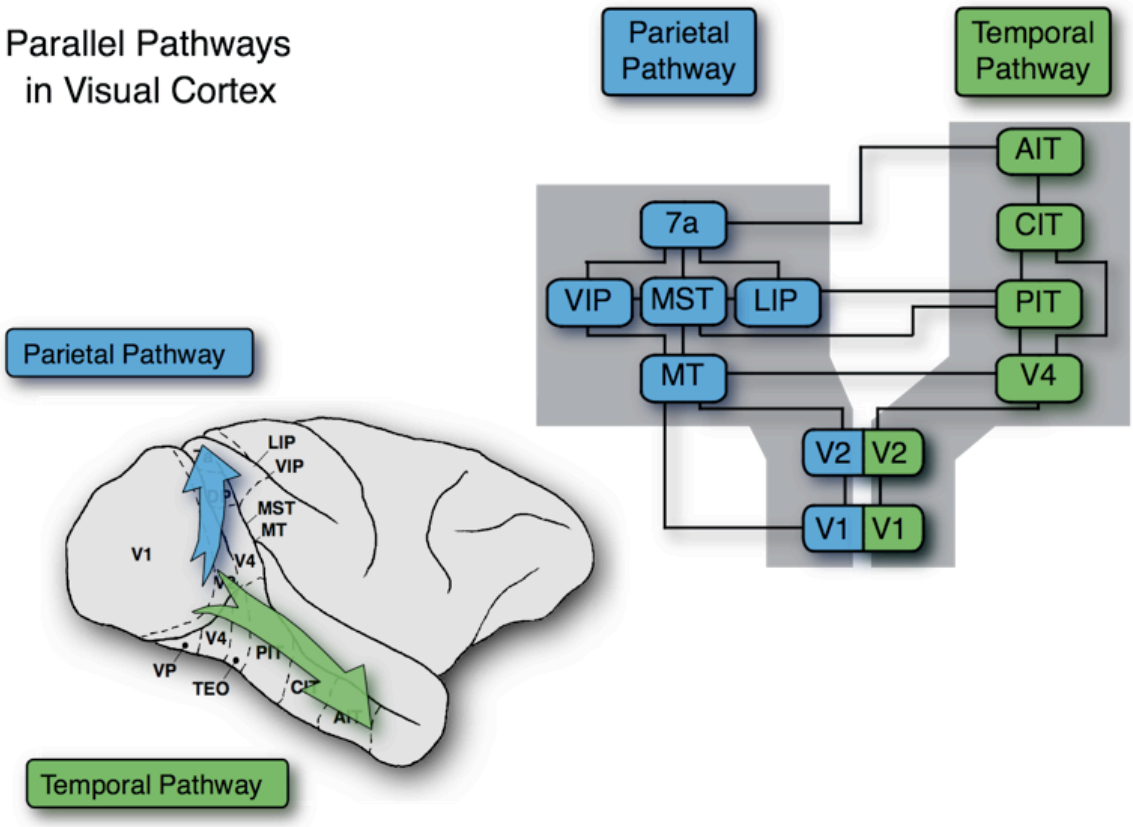


Basic Properties of GEONs

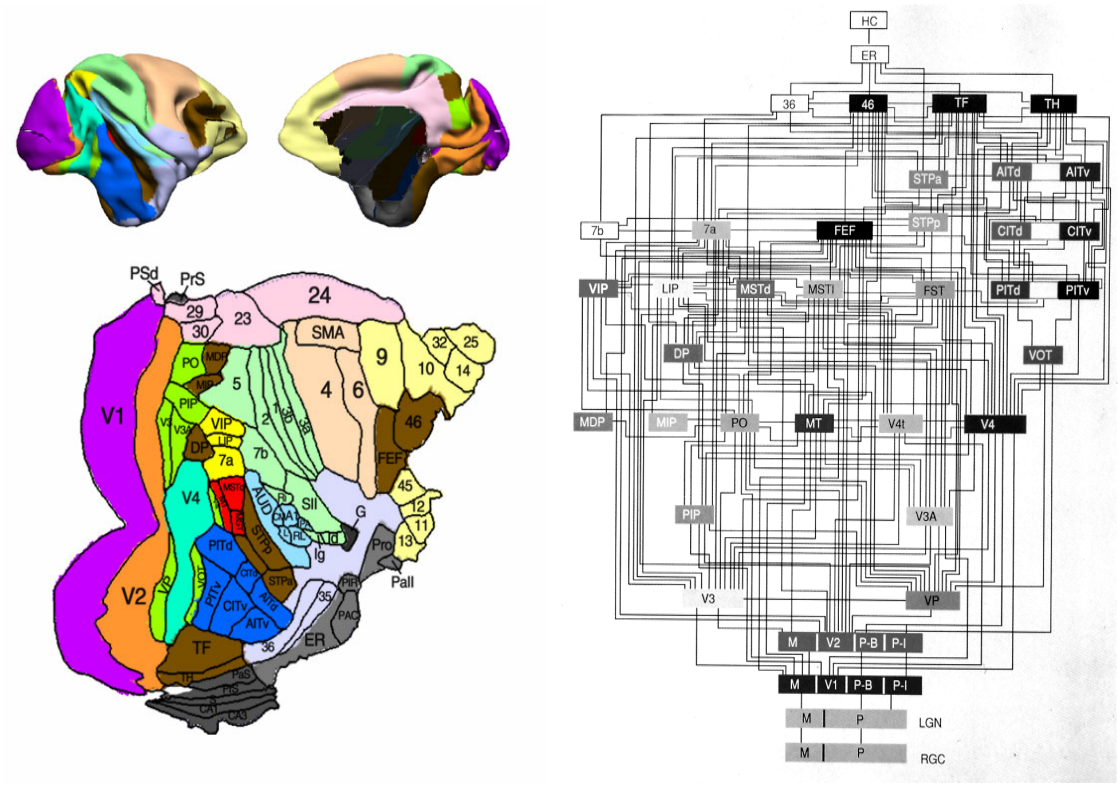
- They are sufficiently different from each other to be easily discriminated
- They are view-invariant (look identical from most viewpoints)
- They are robust to noise (can be identified even with parts of image missing)

Two pathways of visual perception

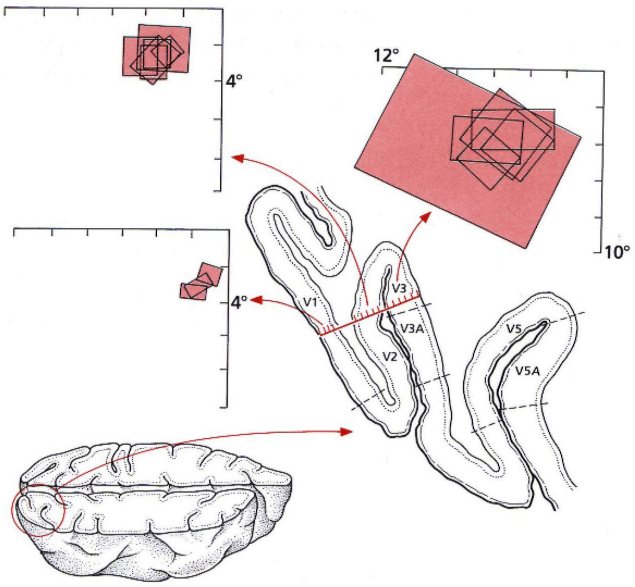


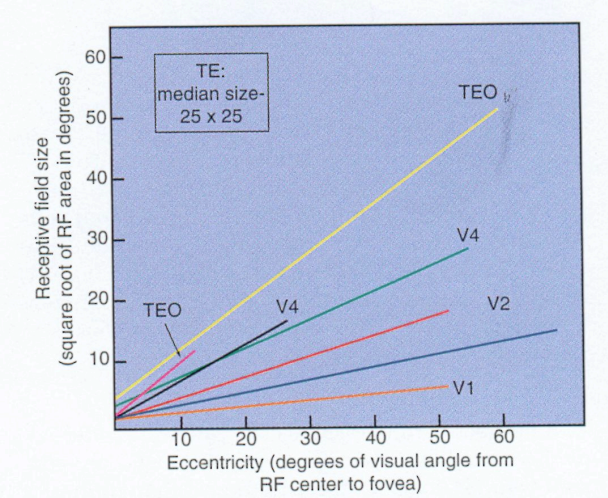


General principle I: Higher areas become smaller

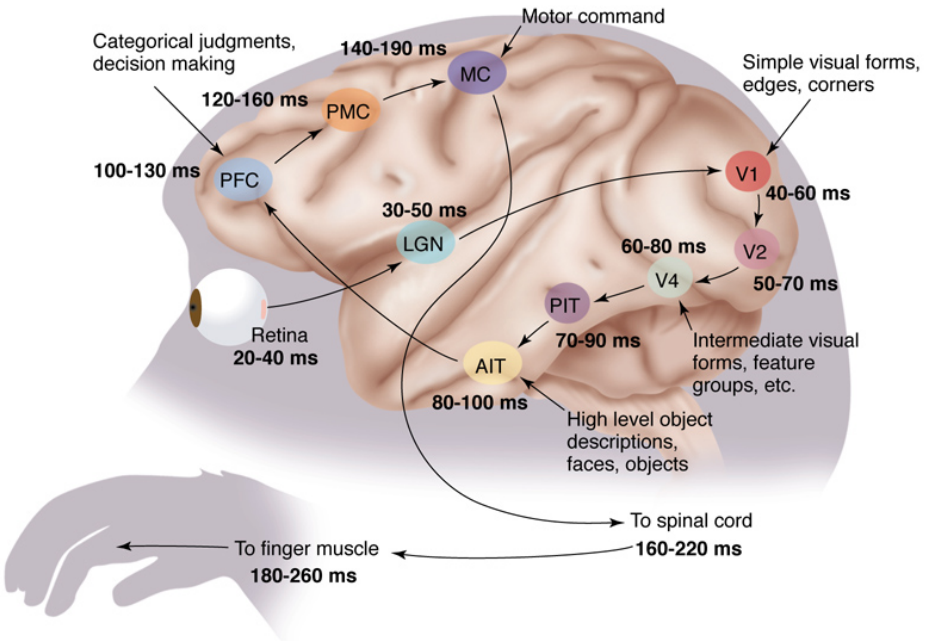


General principle II: Receptive fields become larger



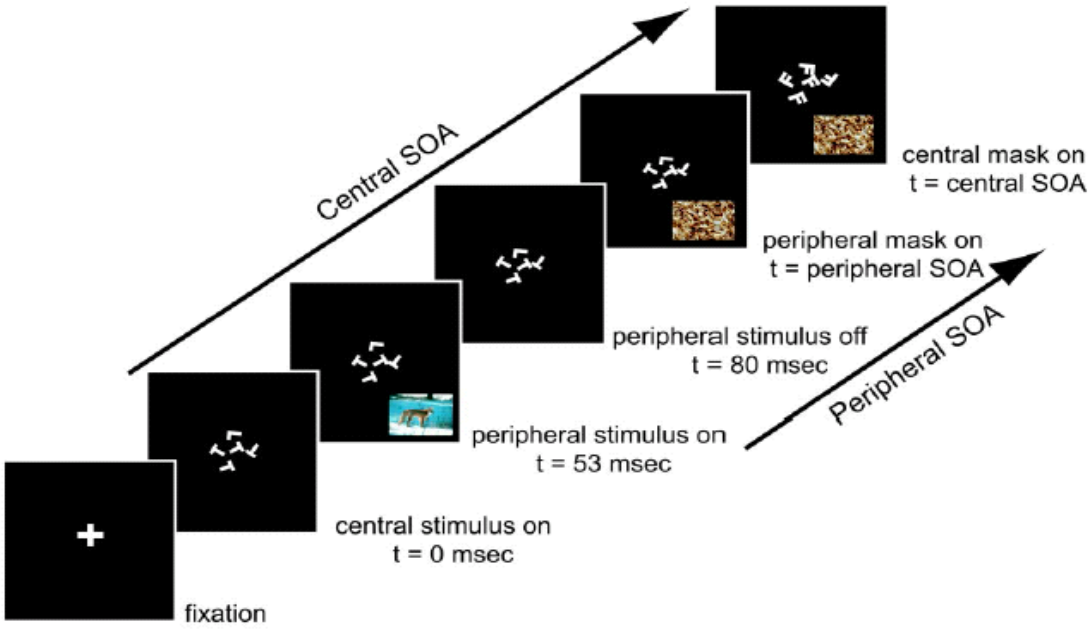


Category pathway

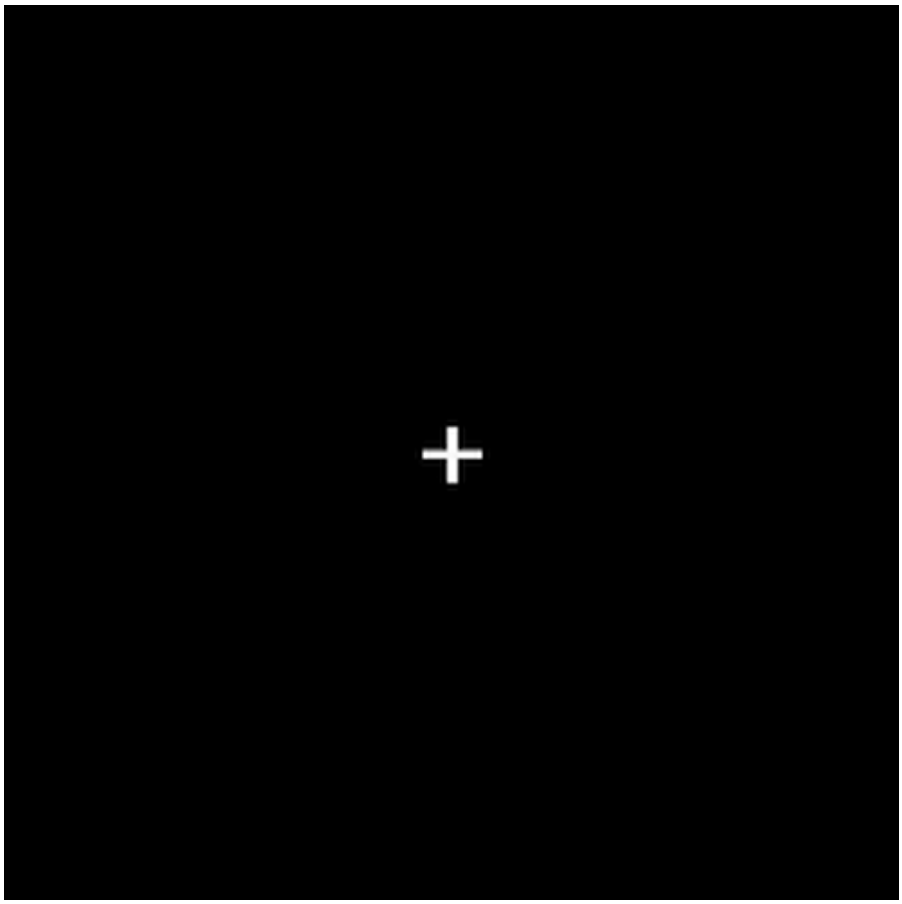


Thorpe, SJ & Fabre-Thorpe, M (2001)
Seeking categories in the brain. Science, 291:260-263.

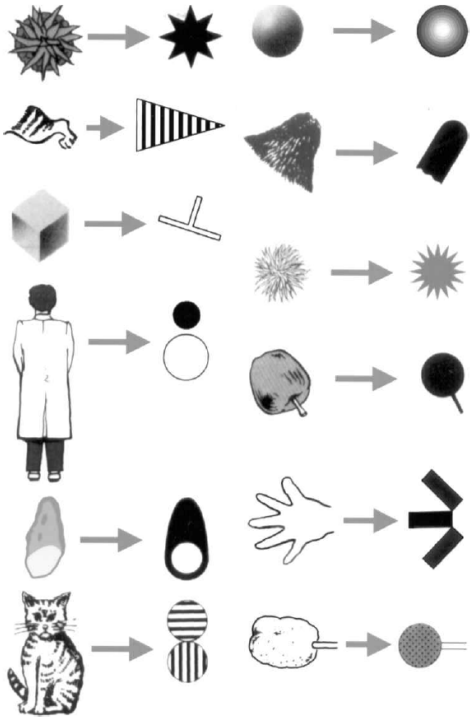
Category detection in dual-task conditions



Li, F.-F., VanRullen, R., Koch, C., Perona, P. (2002) Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci. USA. 99:9596-9601.



Representation of features in IT

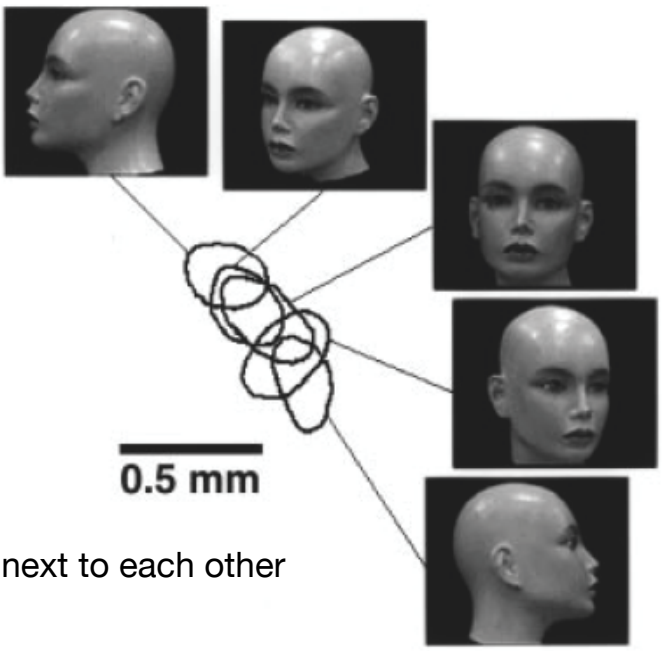


Experimental approach to investigate the encoding of objects in the inferior temporal cortex.

Search a cell that responds to a natural stimulus and simplify the stimulus as much as possible under the constraint that the cell continues to respond to the simplified stimulus.

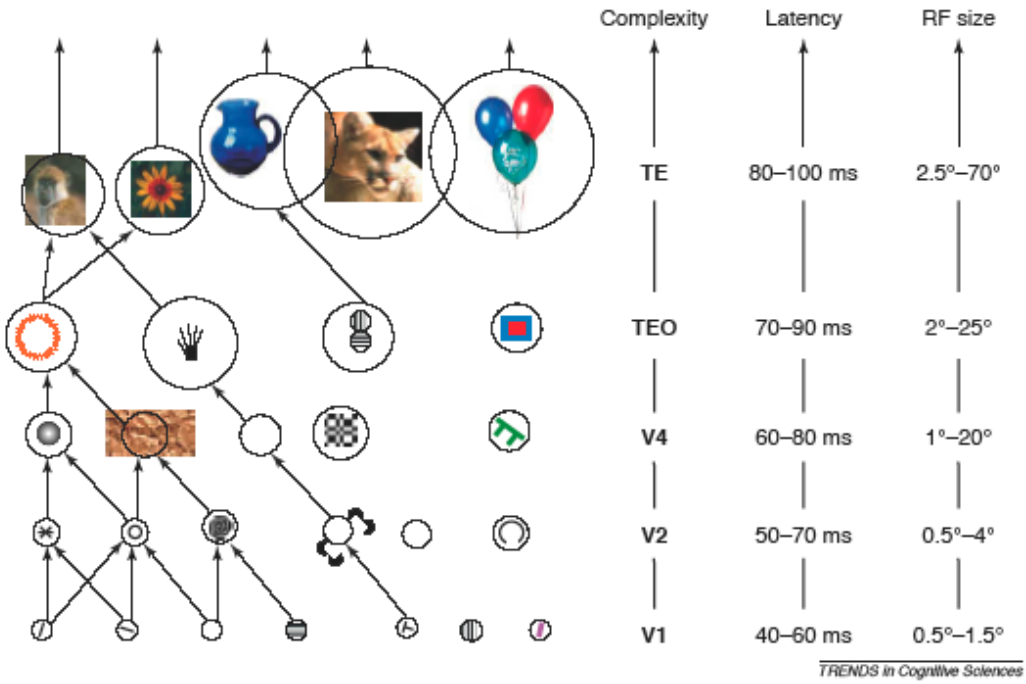
Tanaka,K

View dependent representations



Similar views are represented next to each other in the cortex.

Visual perception in the ventral pathway



Challenges of Object Recognition

- The binding problem

  Binding different features (color, orientation, etc) to yield a unitary percept !

- Location Invariance

  Recognize an object regardless of its location !

- Top-Down Influence

  How much in object recognition is top-down directed ?

- Viewpoint invariance

  Matching 2D views or 3D representation ?

Models of Object Recognition

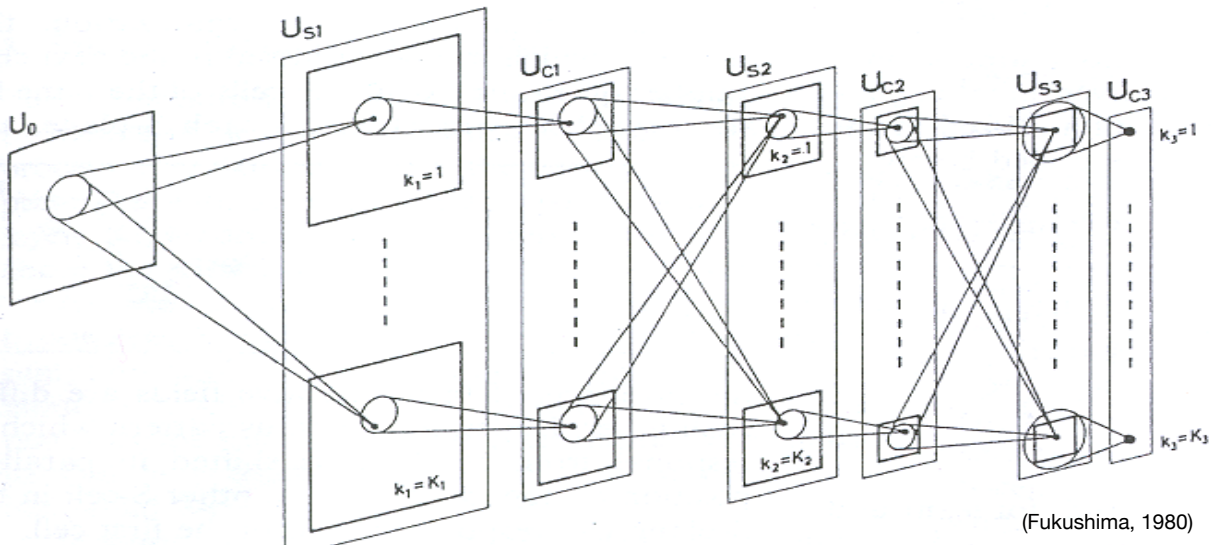
Hierarchical Template Matching:

- Image passed through layers of units with progressively more complex features at progressively less specific locations.
- Hierarchical in that features at one stage are built from features at earlier stages.
- Processing hierarchy yields activation of view-tuned units.
- A collection of view-tuned units is associated with one object.
- View tuned units are built from V4-like units,using sets of weights which differ for each object.

Normalized Template Matching:

- Achieve invariances by a transformation of the visual scene.
- Match the normalized view with template

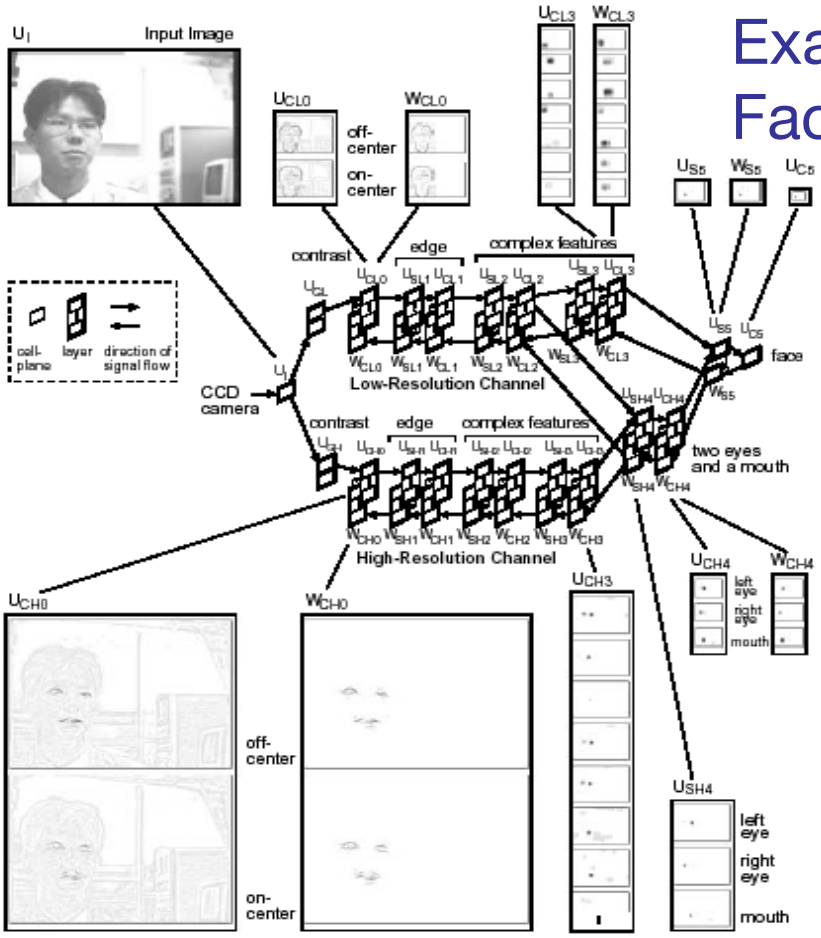
Position invariant recognition in the Neocognitron



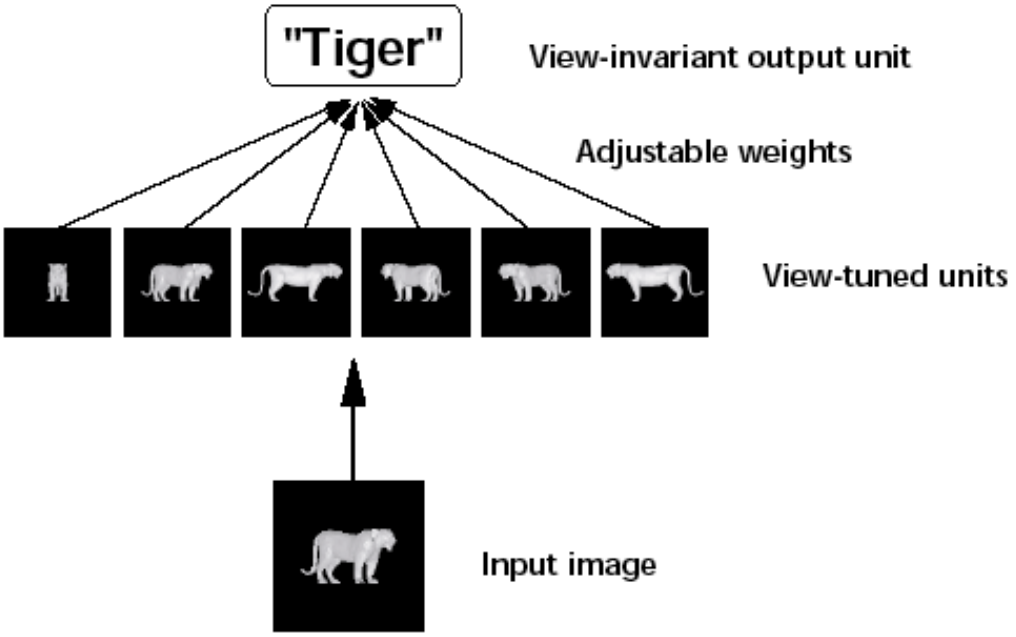
(Fukushima, 1980)

Several processing layers, comprising simple (S) and complex (C) cells.
S-cells in one layer respond to conjunctions of C-cells in previous layer.
C-cells in one layer are excited by small neighborhoods of S-cells.

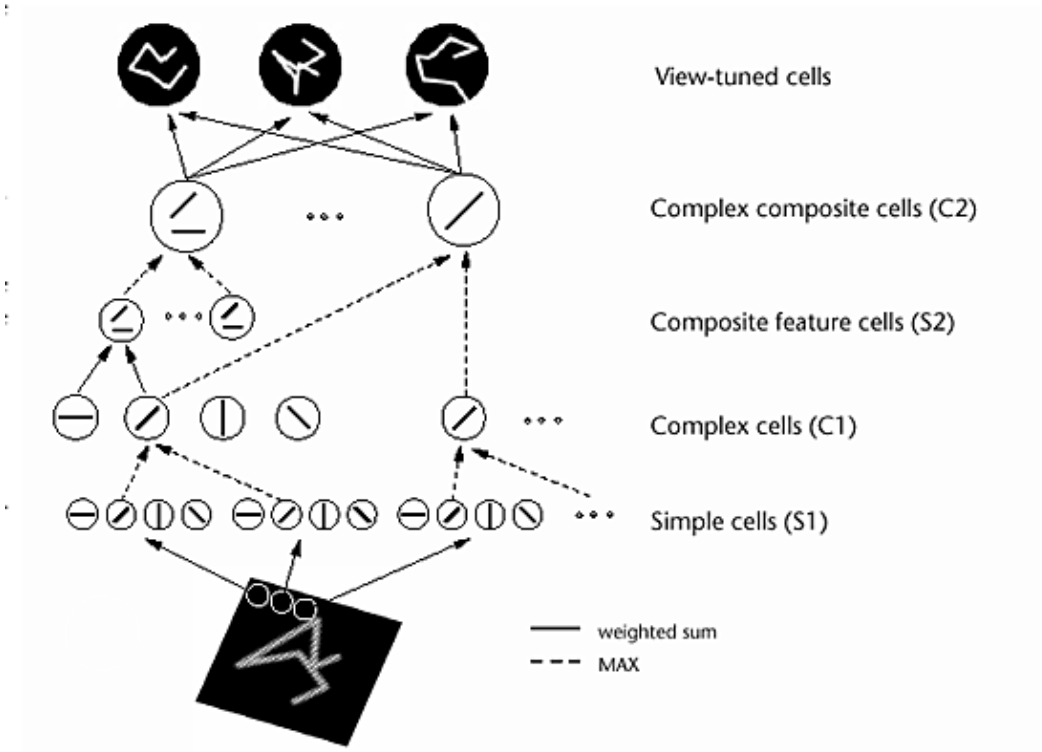
Example: Face Recognition



Viewpoint-dependent recognition

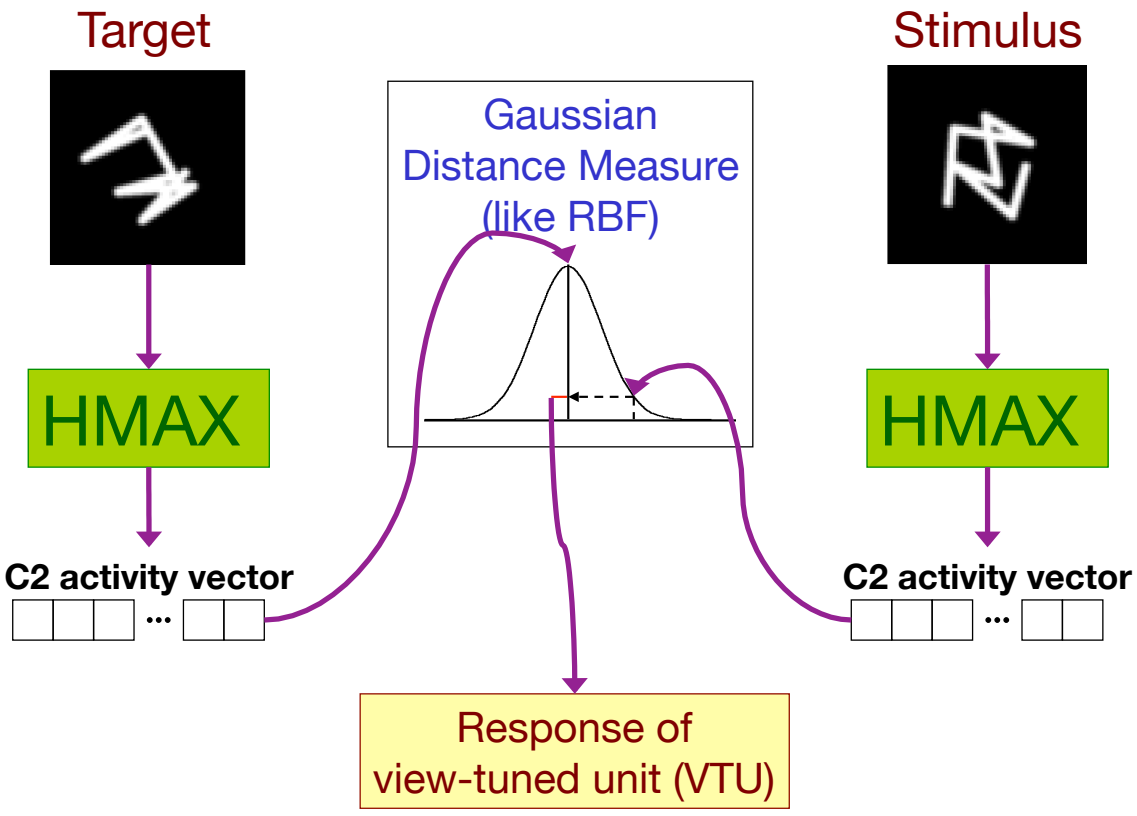


HMAX - basic principles

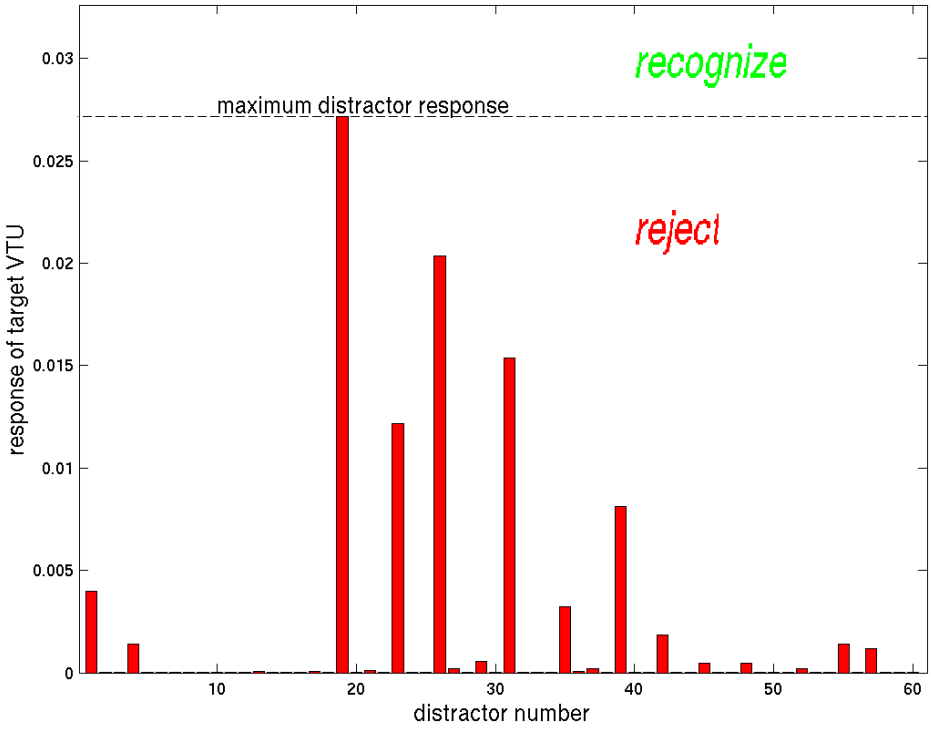


From: Riesenhuber and Poggio, 1999

HMAX - distance measure for recognition

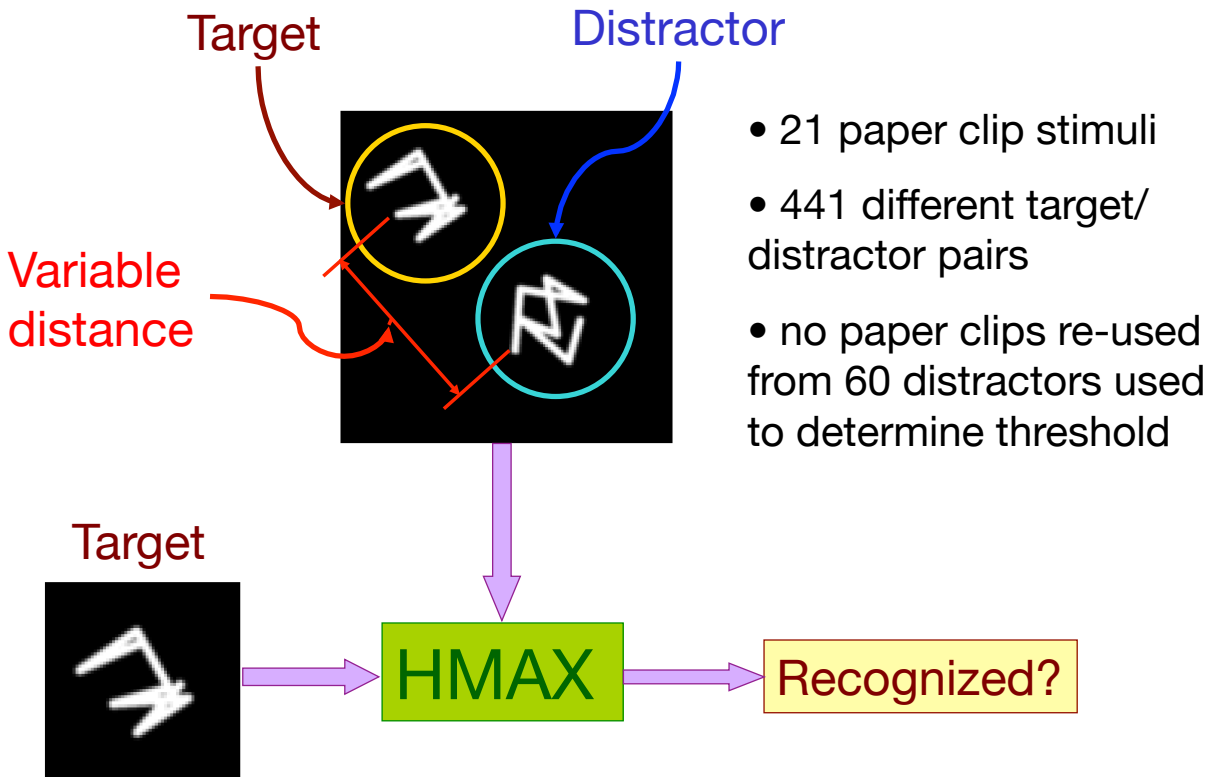


HMAX - distance measure for recognition



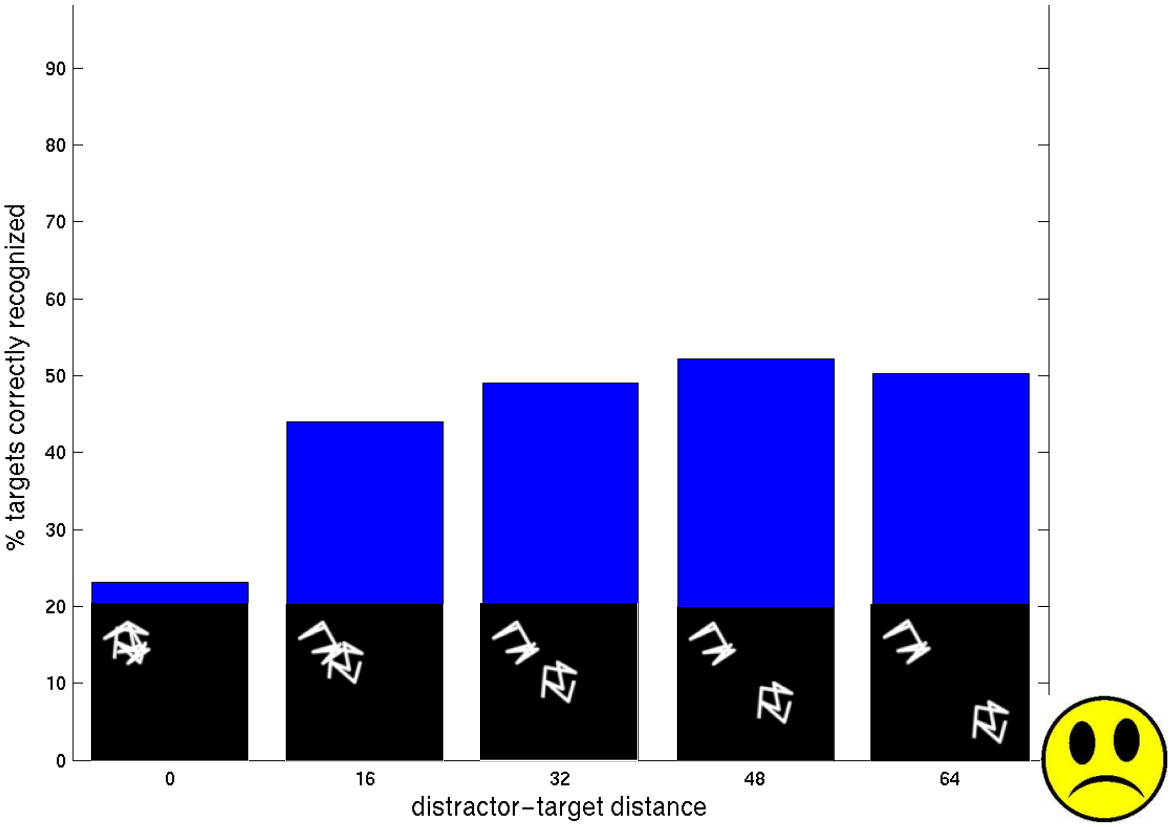
60 randomly chosen distracter paper clips

HMAX - Two paper clips example

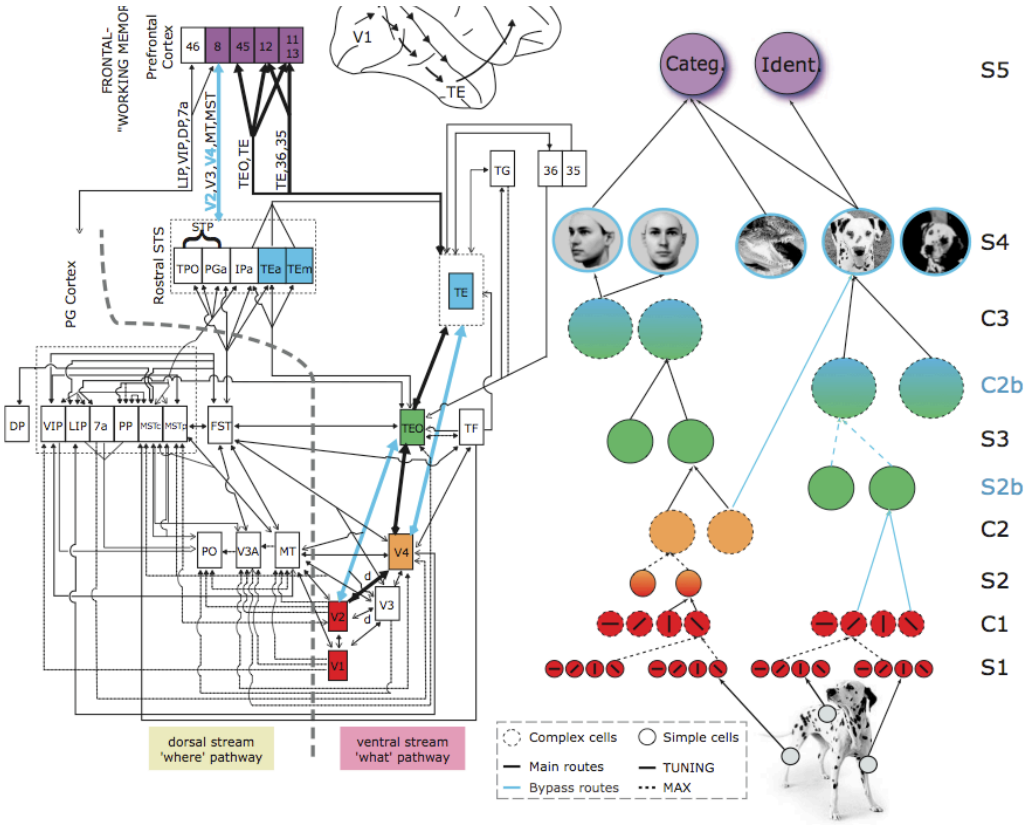


- 21 paper clip stimuli
- 441 different target/ distractor pairs
- no paper clips re-used from 60 distractors used to determine threshold

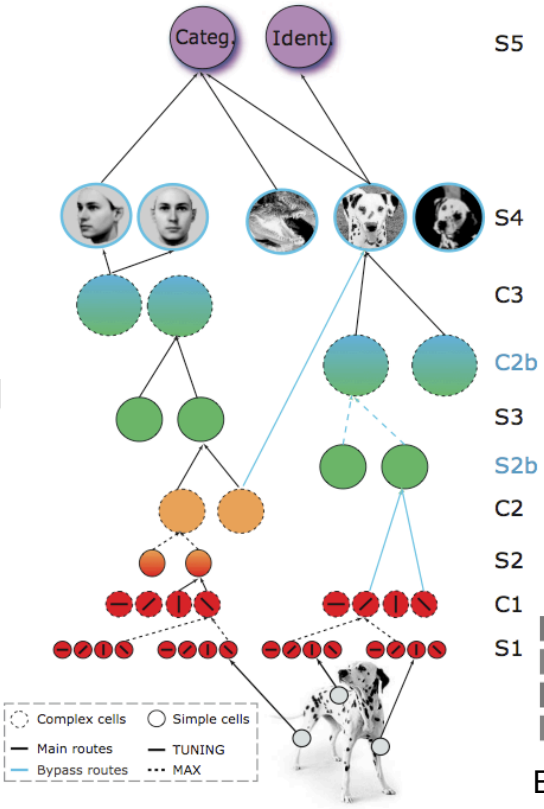
HMAX - Two paper clips example

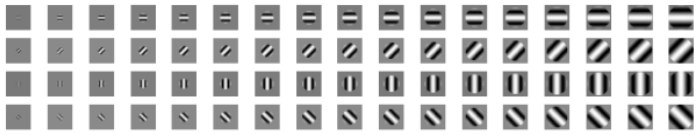


HMAX V2 - model details



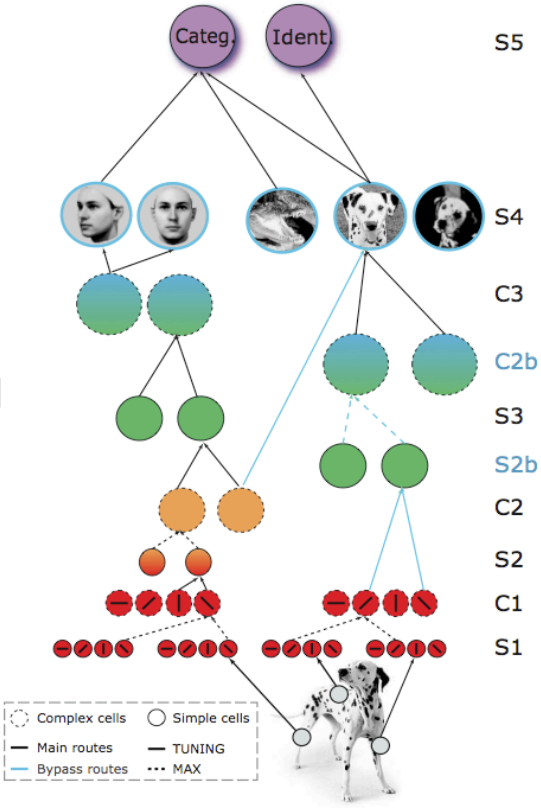
HMAX V2 - model details





Examples of S1 units (4 orientations x 17 sizes)

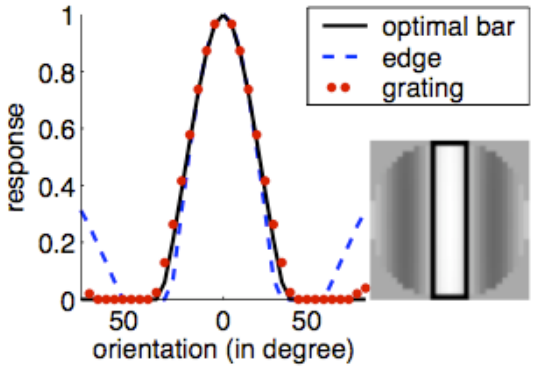
HMAX V2 - model details



$$F(u_1, u_2) = \exp\left(-\frac{(\hat{u_1}^2 + \gamma^2 \hat{u_2}^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}\hat{u_1}\right)$$

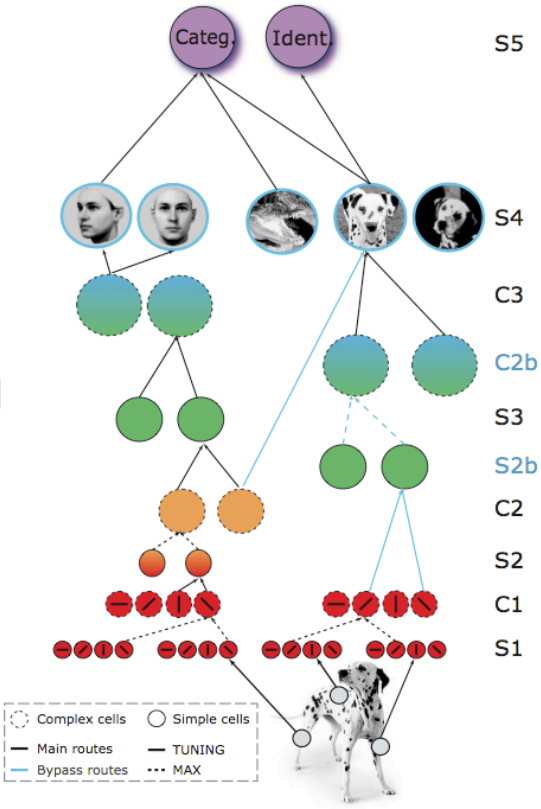
$$\hat{u_1} = u_1 \cos\theta + u_2 \sin\theta \quad \text{and}$$

$$\hat{u_2} = -u_1 \sin\theta + u_2 \cos\theta,$$



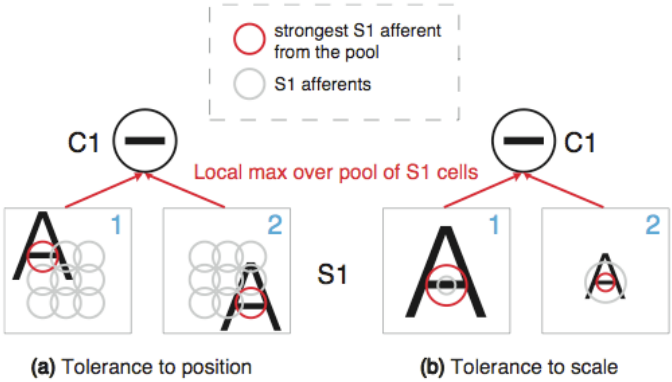
Example S1 unit and its corresponding orientation tuning curve obtained with three classical stimuli, i.e., optimal bar (see black rectangle superimposed on the unit receptive field), optimal grating and edge. A circular mask to the receptive field of the S1 units has been applied.

HMAX V2 - model details



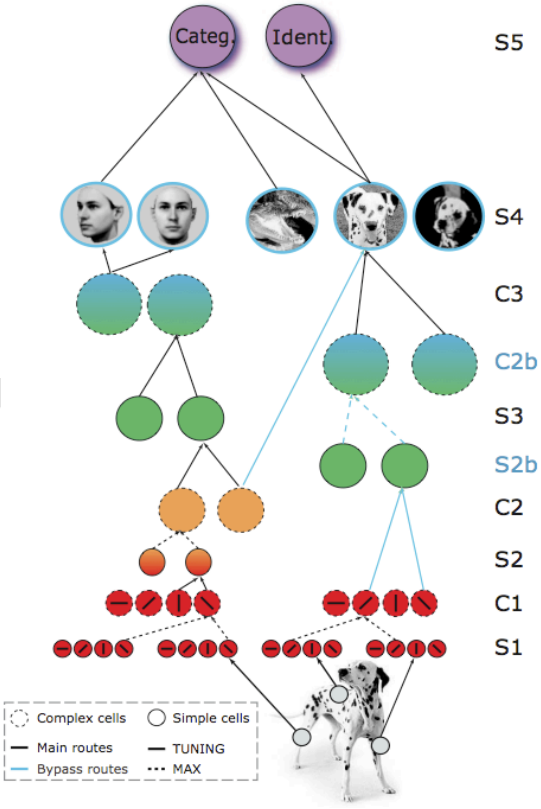
Spatial pooling:

$$y = \max_{j=1\ldots n_{C_k}} x_j.$$



By pooling the activity of all the units in the neighborhood the C1 unit becomes insensitive to the location of the stimulus (a). Similarly for invariance to scale (b).

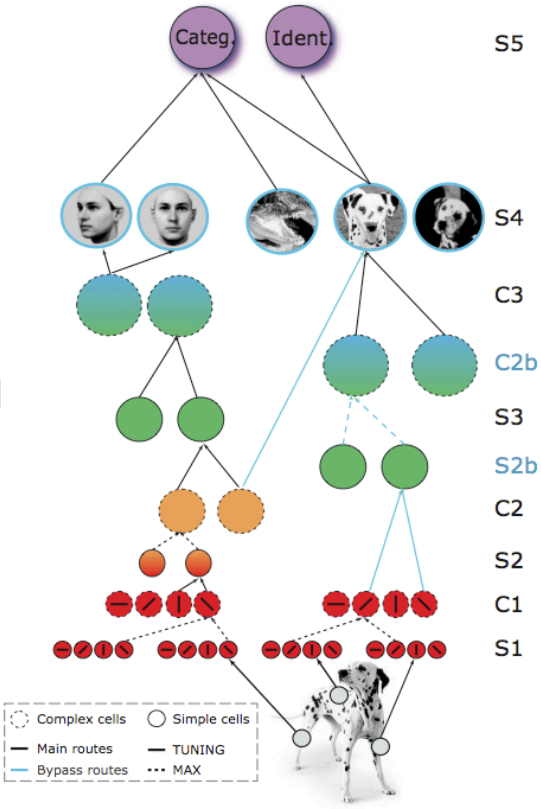
HMAX V2 - model details



S2 and S3 units:

$$y = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_{S_k}}(w_j - x_j)^2\right)$$

HMAX V2 - model details

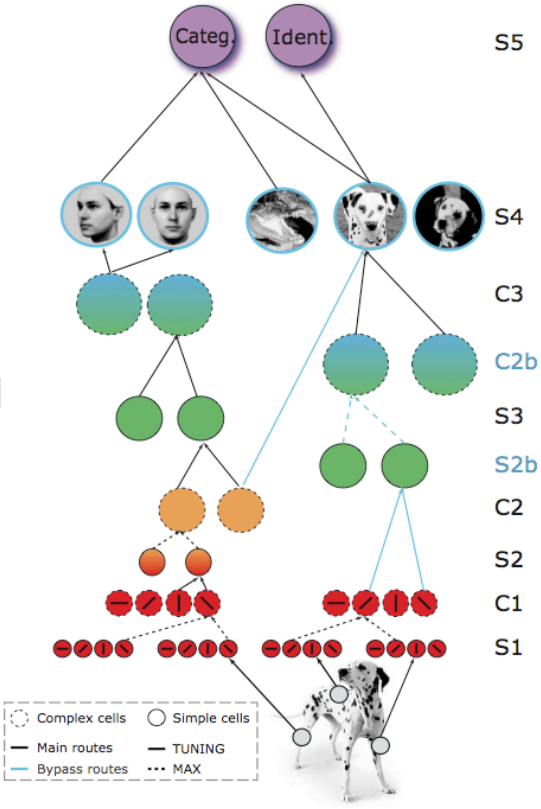


Imprinting S2 and S3 units:

$$y = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_{S_k}}(w_j - x_j)^2\right)$$

At the k[th] image presentation, one macro-column (which corresponds to a particular portion of the visual fild and scale) is selected (at random) and unit w_k from this macro-column is imprinted, i.e., the unit stores in its synaptic weights the current pattern of activity from its afferent inputs in response to the part of the natural image that fell within its receptive field. This is done by setting w_k to be equal to the current pattern of pre-synaptic activity x.
A weight sharing approach is used that leads to identical weights at different locations in the visual field.

HMAX V2 - model details



Classification units:

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x^i}, \mathbf{x})$$

S4 unit:

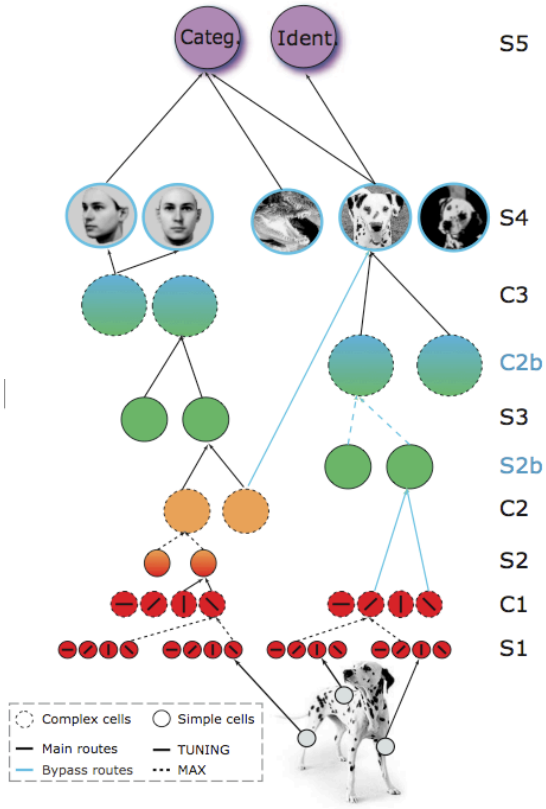
$$K(\mathbf{x^i}, \mathbf{x}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j^i - x_j)^2\right)$$

Supervised learning:

Adjust the synaptic weights c_i so as to minimize the overall classification error on the training set

$$E = \sum_{i=1}^{l} ||f(\mathbf{x^i}) - y^i||^2 + R(f)$$

HMAX V2 - model details



Layers	Number of units
S_1	1.6×10^6
C_1	2.0×10^4
S_2	1.0×10^7
C_2	2.8×10^5
S_3	7.4×10^4
C_3	1.0×10^4
S_4	1.5×10^2
S_{2b}	1.0×10^7
C_{2b}	2.0×10^3
Total	2.3×10^7

Discussion

- The hierarchical template matching model can primarily be used to label the contents of an image.
- It is consistent with our ability to recognize objects very fast.
- This approach cannot deal with the dynamics of visual perception, such as masking and bistable figures.
- Learning of features from images is a topic of research and still not solved.