Visual Object Recognition

Computational Models and Neurophysiological Mechanisms

Neurobiology 130/230. Harvard College/GSAS 78454

Web site: http://tinyurl.com/visionclass

→ Class notes, Class slides, Readings Assignments

Location: Biolabs 2062

Time: Mondays 03:00 – 05:00

Lectures:

Faculty: Gabriel Kreiman and invited guests

TA: Emma Giles

Contact information:

Gabriel Kreiman | Emma Giles

gabriel.kreiman@tch.harvard.edu | emmagiles@g.harvard.edu

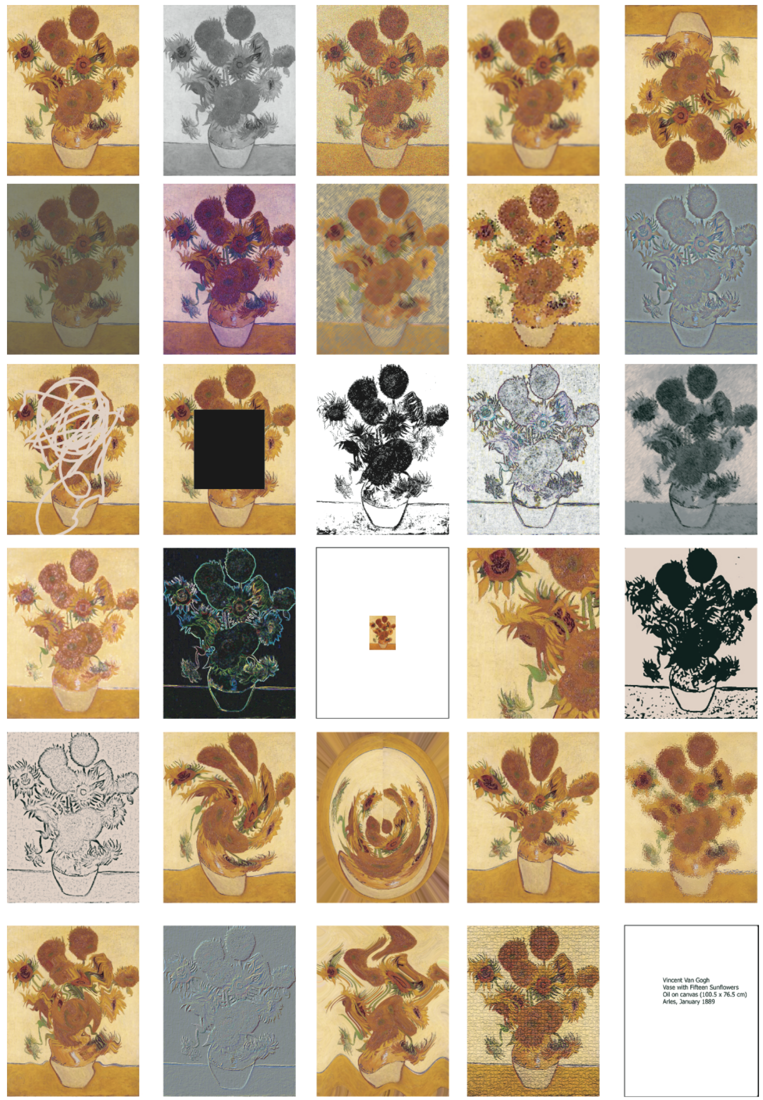
617-919-2530

Office Hours: After Class. Mondays 5pm, or by appointment

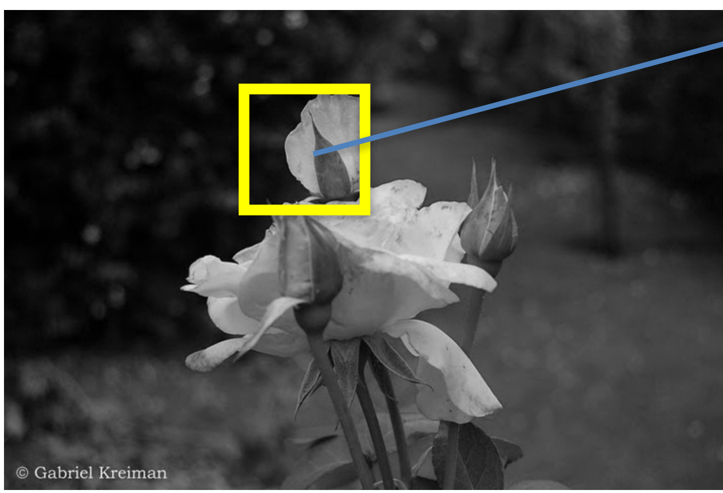
Visual Object Recognition
Computational Models and Neurophysiological Mechanisms
Neurobiology 130/230. Harvard College/GSAS 78454

Class 1 [09/10/2018]. Introduction to pattern recognition [Kreiman]
Class 2 [09/17/2018]. Why is vision difficult? Natural image statistics. The retina. [Kreiman]
Class 3 [09/24/2018]. Lesions and neurological studies [Kreiman].
Class 4 [10/01/2018]. Psychophysics of visual object recognition [Sarit Szpiro]
October 8: University Holiday
Class 5 [10/15/2018]. Primary visual cortex [Hartmann]
Class 6 [10/22/2018]. Adventures into *terra incognita* [Frederico Azevedo]
Class 7 [10/29/2018]. High-level visual cognition [Diego Mendoza-Haliday]
Class 8 [11/05/2018]. Correlation and causality. Electrical stimulation in visual cortex [Kreiman]
Class 9 [11/12/2018]. Visual consciousness [Kreiman]
Class 10 [11/19/2018]. Computational models of neurons and neural networks. [Kreiman]
Class 11 [11/26/2018]. Computer vision. Artificial Intelligence in Visual Cognition [Bill Lotter]
Class 12 [12/03/2018]. The operating system for vision. [Xavier Boix]
FINAL EXAM, PAPER DUE 12/13/2018. No extensions.

An object can cast an infinite number of projections on the retina

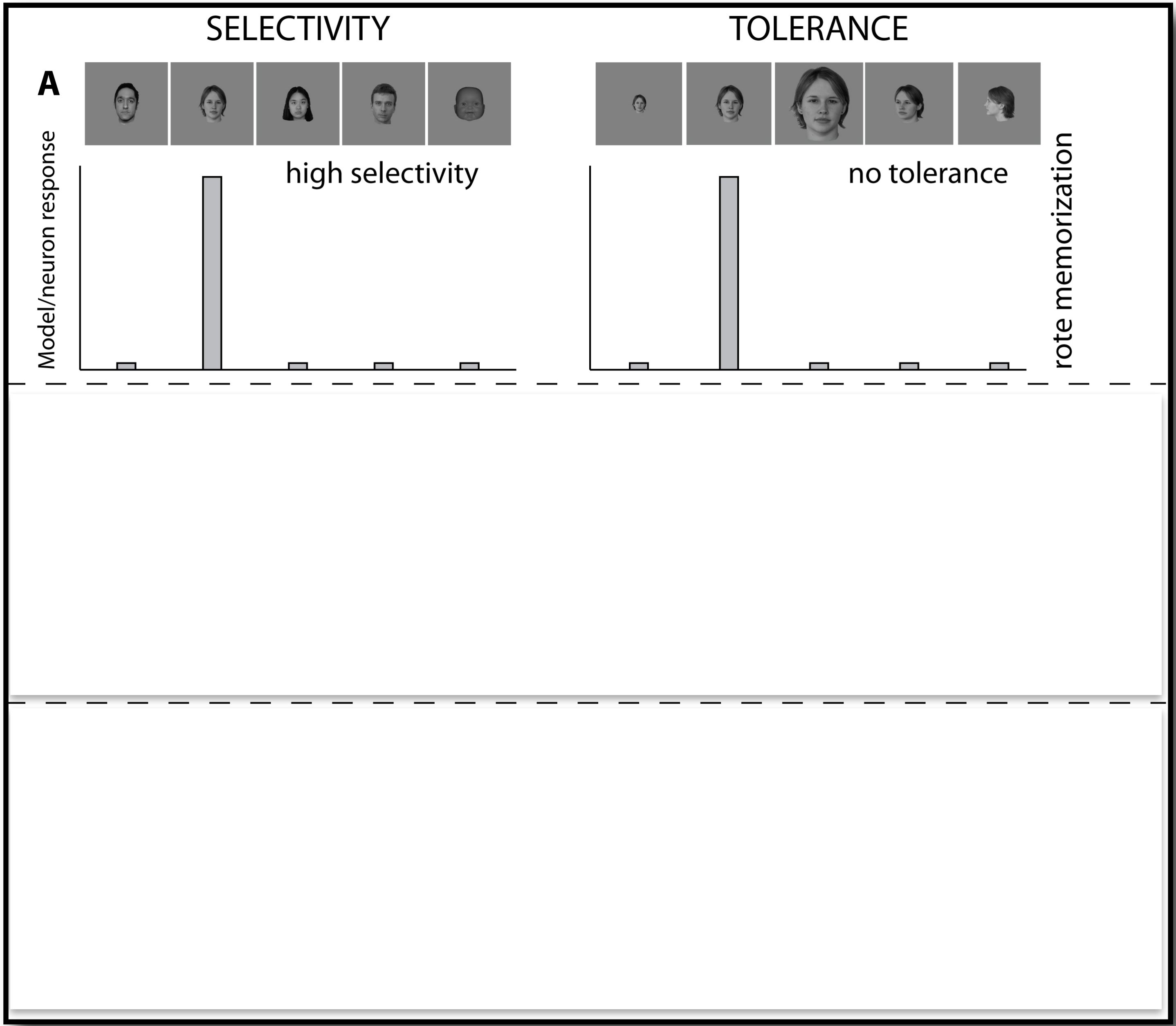


A flower, as seen by a computer

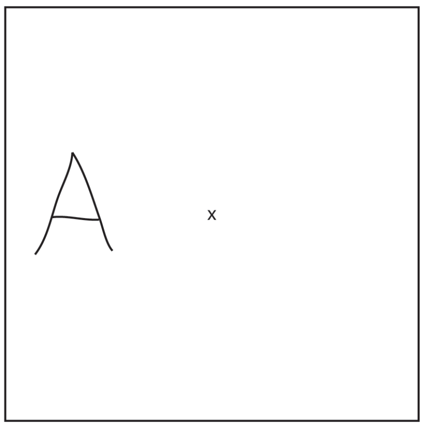


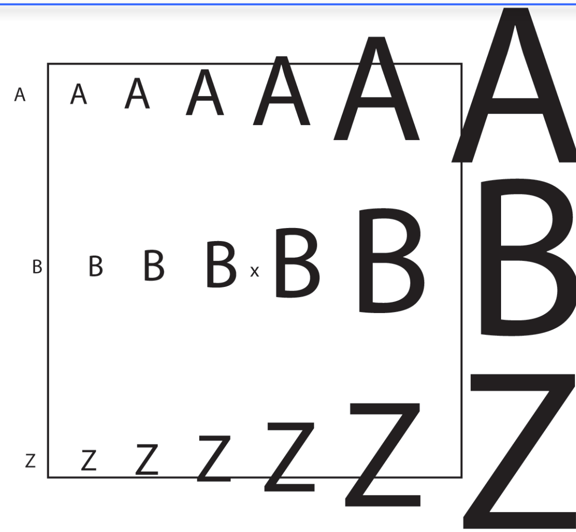
23	16	13	12	13	13	12	12	12	14	16	19	21	22	25	24	20	90	127	101
31	22	13	13	12	12	11	11	13	16	18	18	23	22	21	19	39	83	96	78
34	24	16	14	13	12	21	14	13	17	15	22	15	29	42	82	147	118	63	36
30	20	15	13	14	12	26	34	10	11	79	139	88	91	119	174	172	137	96	78
20	14	12	12	14	14	21	77	35	16	136	148	110	109	127	137	168	157	144	175
13	10	10	12	15	16	14	81	86	52	155	123	91	114	149	120	154	139	138	186
9	9	9	11	14	17	18	54	110	111	143	99	105	104	148	128	103	148	162	172
9	8	9	11	14	18	20	26	97	99	99	91	116	116	141	139	77	88	117	156
9	9	12	12	15	18	15	29	107	99	88	86	121	124	115	123	79	78	98	92
9	10	11	13	15	16	30	97	121	112	98	68	102	125	115	101	100	60	105	109
9	9	11	14	17	13	96	127	145	115	95	60	90	114	118	98	107	72	60	111
9	10	12	13	16	17	117	128	122	114	89	65	94	108	118	116	117	93	59	67
10	10	10	7	9	78	152	127	118	114	77	72	95	109	116	120	128	96	68	50
7	1	10	54	114	166	145	121	125	113	65	70	97	107	110	107	103	93	67	54
33	92	129	151	157	158	146	130	125	104	66	77	100	105	111	108	94	85	62	58
145	144	135	142	151	152	149	137	131	98	69	82	102	111	102	93	89	84	59	54
125	125	140	156	144	150	145	133	128	98	74	87	110	110	106	93	86	80	56	48
147	147	161	143	143	144	138	129	121	94	69	86	107	106	102	91	82	77	50	43
182	181	164	140	143	140	132	128	121	97	71	82	100	109	97	91	93	80	44	40
188	174	143	147	146	144	137	127	119	97	78	83	100	105	104	92	86	81	46	38

Two simple and useless solutions



A brute force approach to object recognition

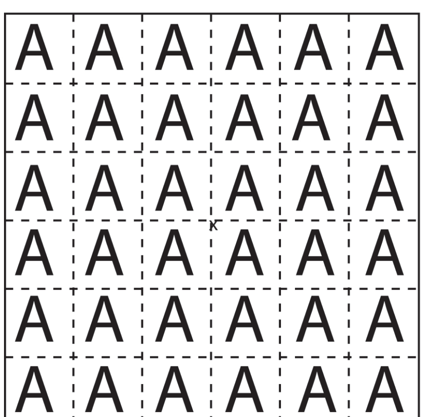


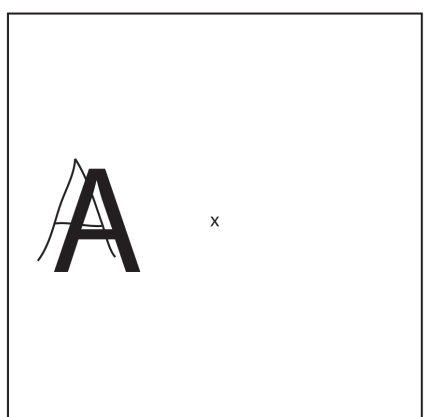


Task: Recognize the handwritten "A"

A "brute force" solution:

- Use templates for each letter
- Use multiple scales per template
- Use multiple positions per template
- Use multiple rotations per template
- Etc.





Problems with this approach:

- Large amount of storage for each object
- No extrapolation, no intelligent learning
- Need to learn about each object under each condition

Recognizing objects by part decomposition

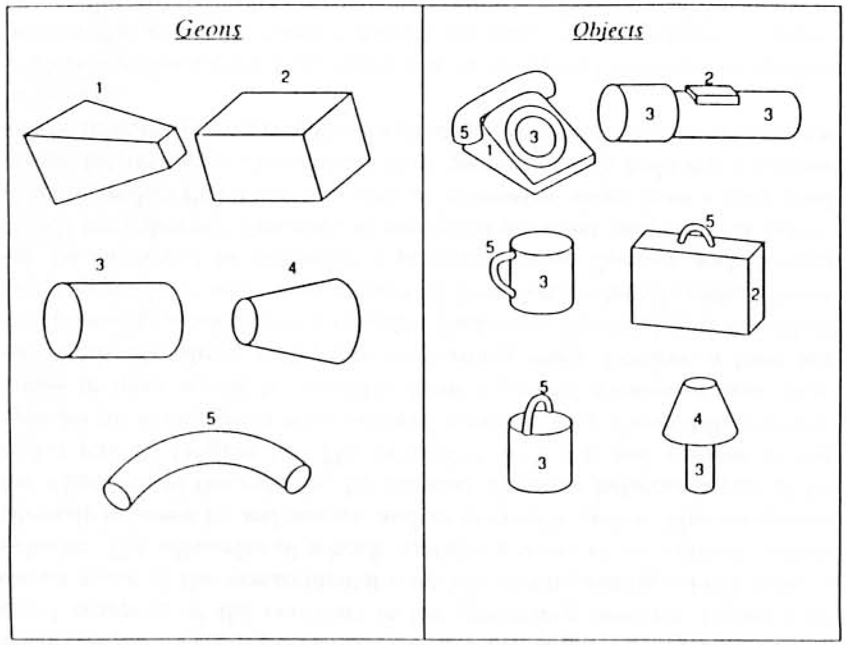


Figure 4.8
(Left) Five geons. (Right) Only two or three geons are required to uniquely specify an object. The relations among the geons matter, as illustrated by the pail and the cup.

Biederman (1987) Psychological Review

A non-exhaustive list of computational models

K. Fukushima, Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 1980. 36: 193-202.*

Y. LeCun, L. Bottou, Y. Bengioand P. Haffner, Gradient-based learning applied to document recognition. Proc of the IEEE, 1998. 86: 2278-2324.

G. Wallis and E.T. Rolls, Invariant face and object recognition in the visual system. Progress in Neurobiology, 1997. 51: 167-94.

B. Mel, SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. Neural Computation, 1997. 9: 777.

B.A. Olshausen, C.H. Anderson and D.C. Van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J Neurosci, 1993. 13: 4700-19.

M. Riesenhuber and T. Poggio, Hierarchical models of object recognition in cortex. Nature Neuroscience, 1999. 2: 1019-1025.

G. Deco and E.T. Rolls, A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res, 2004. 44: 621-42.

P. Foldiak, Learning Invariance from Transformation Sequences. Neural Computation, 1991. 3: 194-200.

Common themes across multiple object recognition models

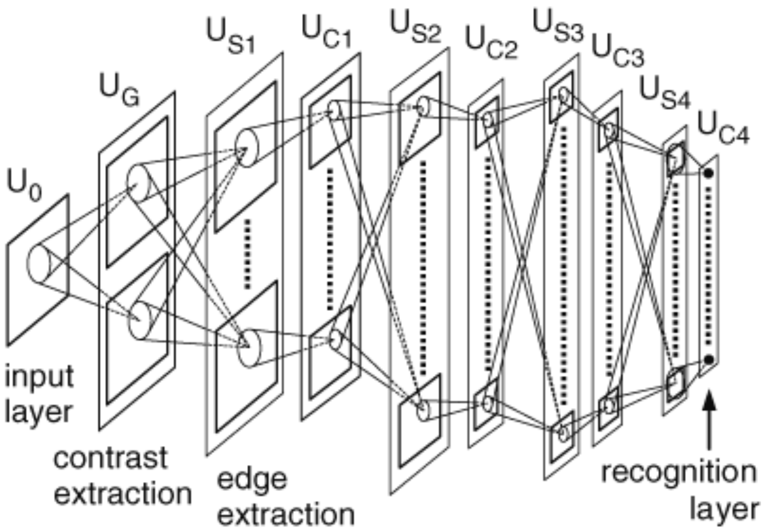
- Hierarchical structure
  "Divide and conquer" strategy

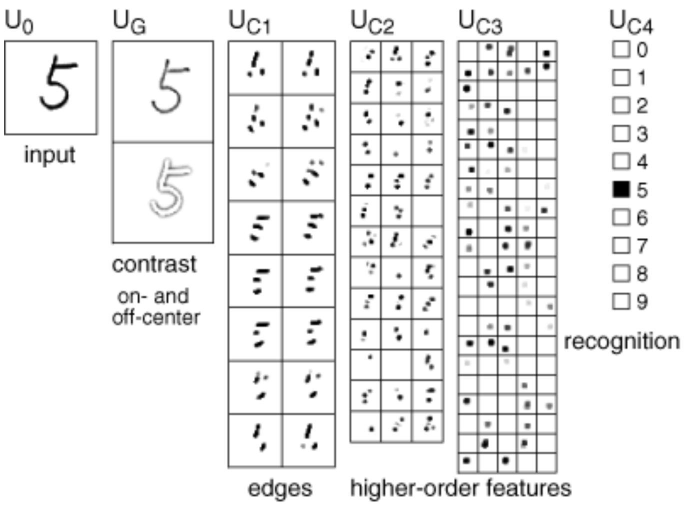
- Increased receptive field size along the hierarchy

- Increased complexity in shape preferences along the hierarchy

- Increased tolerance to (affine) feature transformations along the hierarchy

Neocognitron

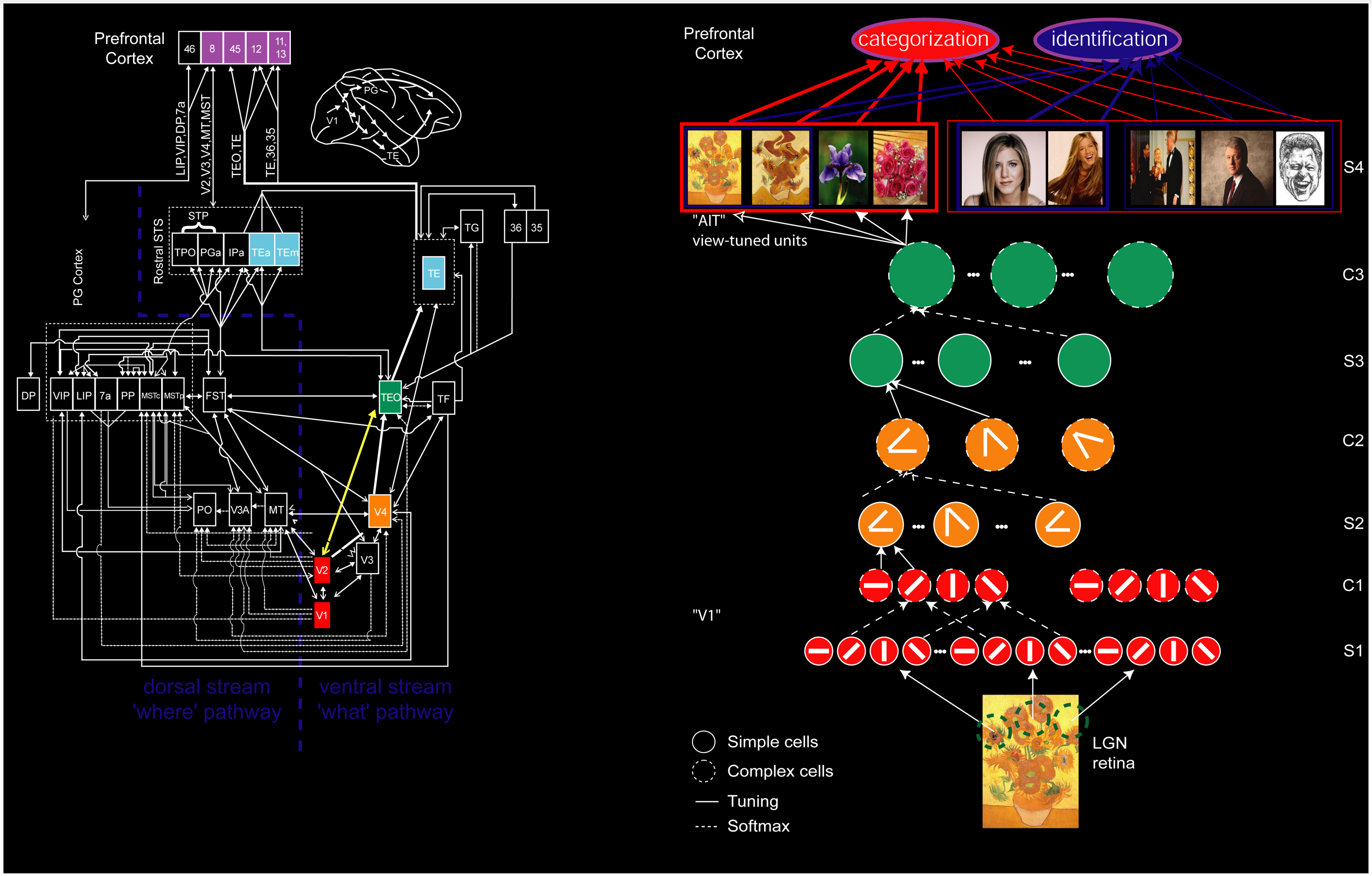




Retinotopically arranged connections between layers
Feature extracting "S" cells
C-cells performing a local "OR" operation
Increasing buildup of position tolerance
Unsupervised learning in S layers

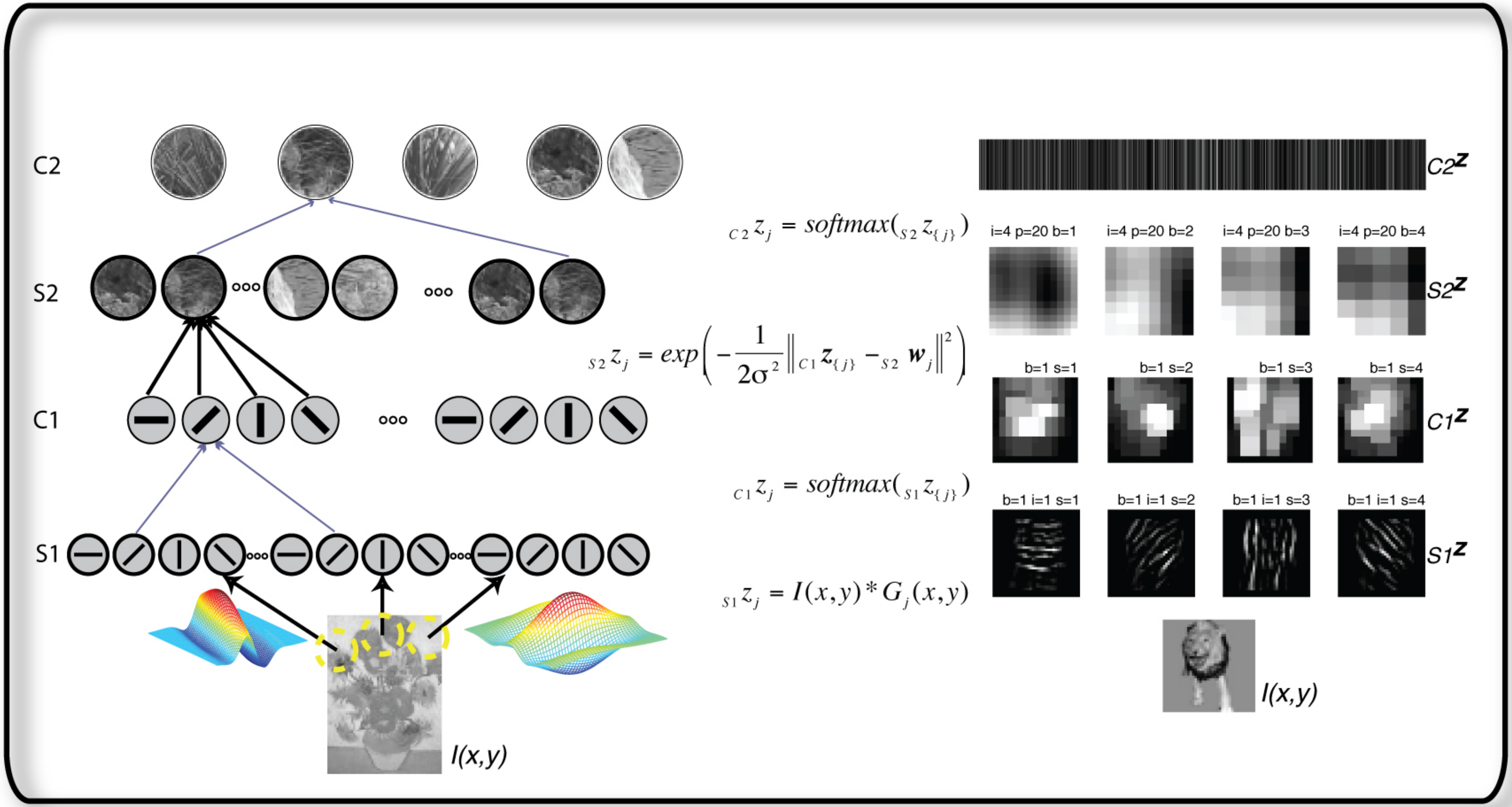
Fukushima K. (1980) Neocognitron: a self organizing neural network model for a mechanism fo pattern recognition unaffected by shift in position. Biological Cybernetics **36**, 193-202

A hierarchical feed-forward model of visual recognition



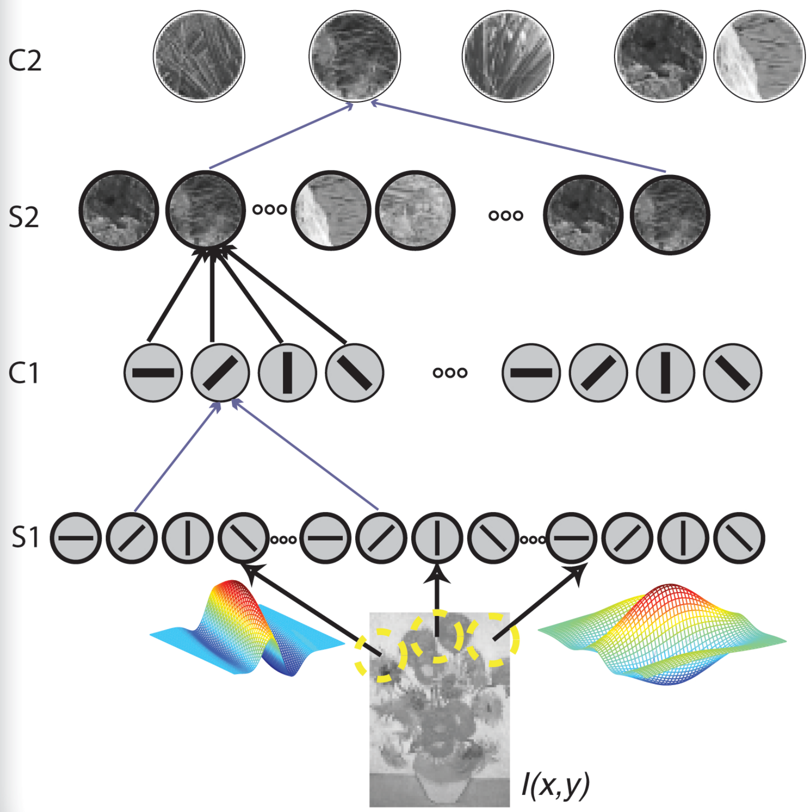
Serre et al. Progress in Brain Research 165C:33-56 (2007)

The CBCL model: A biologically-inspired, bottom-up, hierarchical model of object recognition

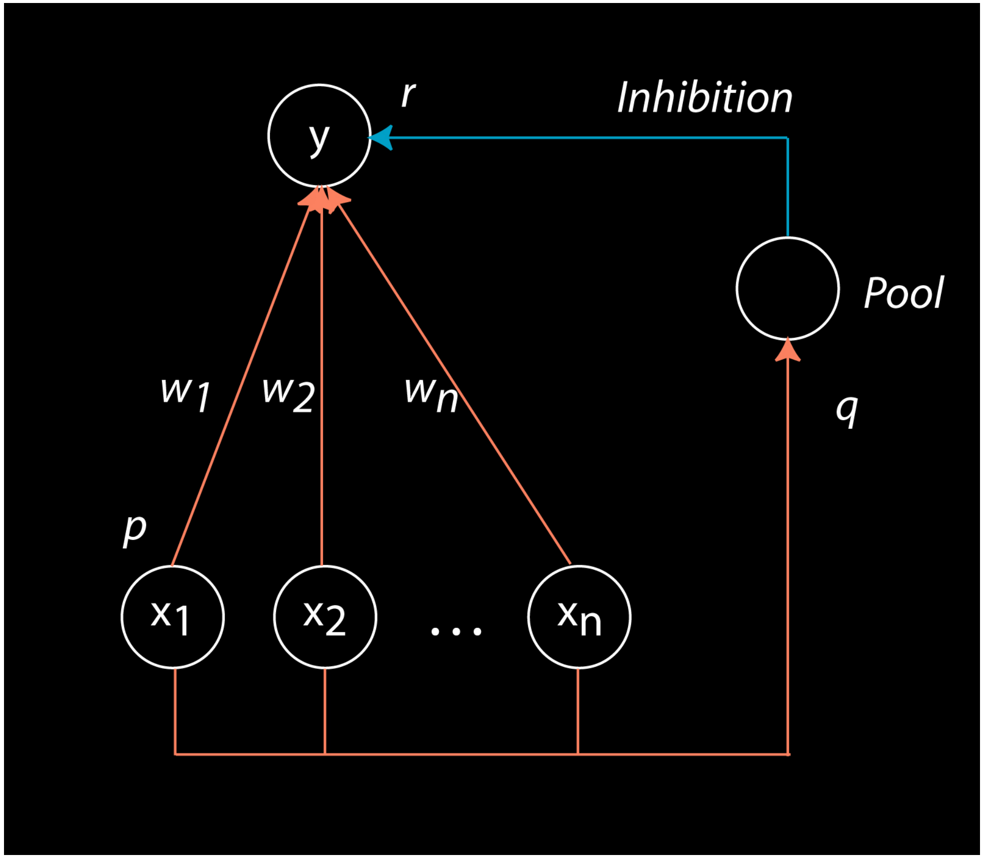


Cadieu, Knoblich, Kouh, Mutch, Riesenhuber, Serre, Poggio

A biologically-inspired, bottom-up, hierarchical model of object recognition



Biophysical implementation of cortical nonlinear operations



$$y = \frac{\sum_{j=1}^{n} w_j x_j^p}{k + \left(\sum_{j=1}^{n} x_j^q\right)^r}$$ Canonical

$$y = \sum_{j=1}^{n} x_j^2$$ Energy model

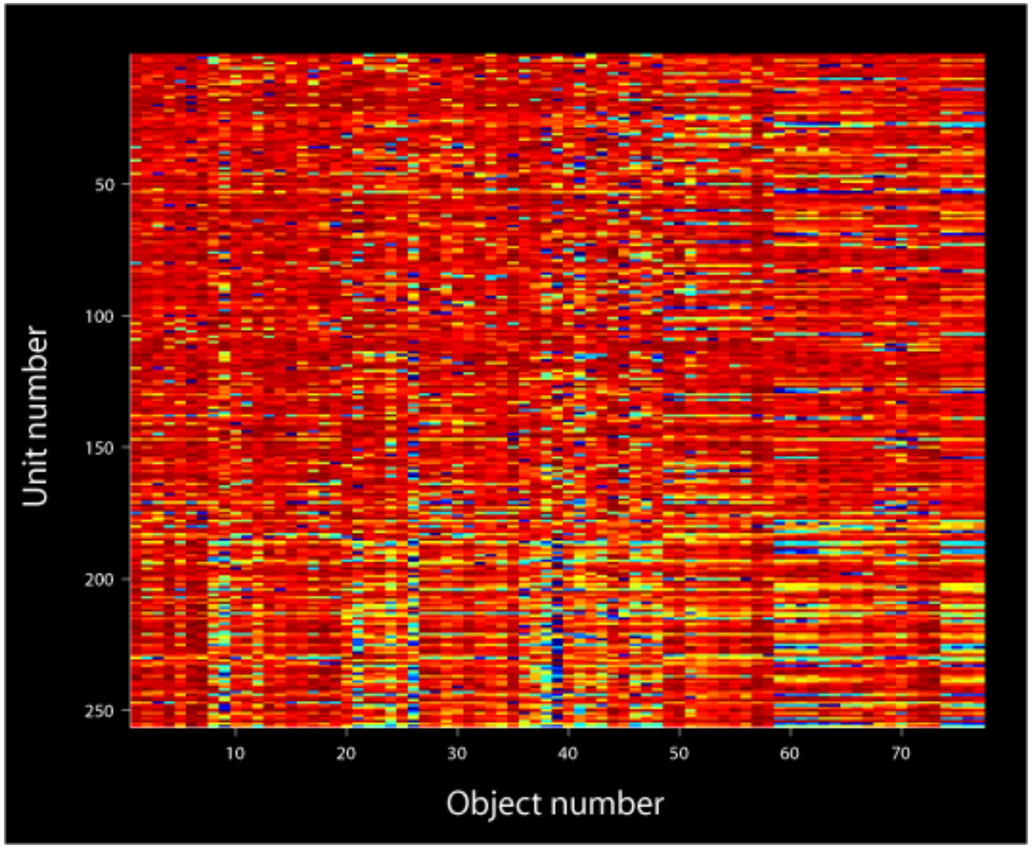
$$y = \frac{\sum_{j=1}^{n} x_j^2}{k + \sum_{j=1}^{n} x_j^2}$$ Sigmoid-like

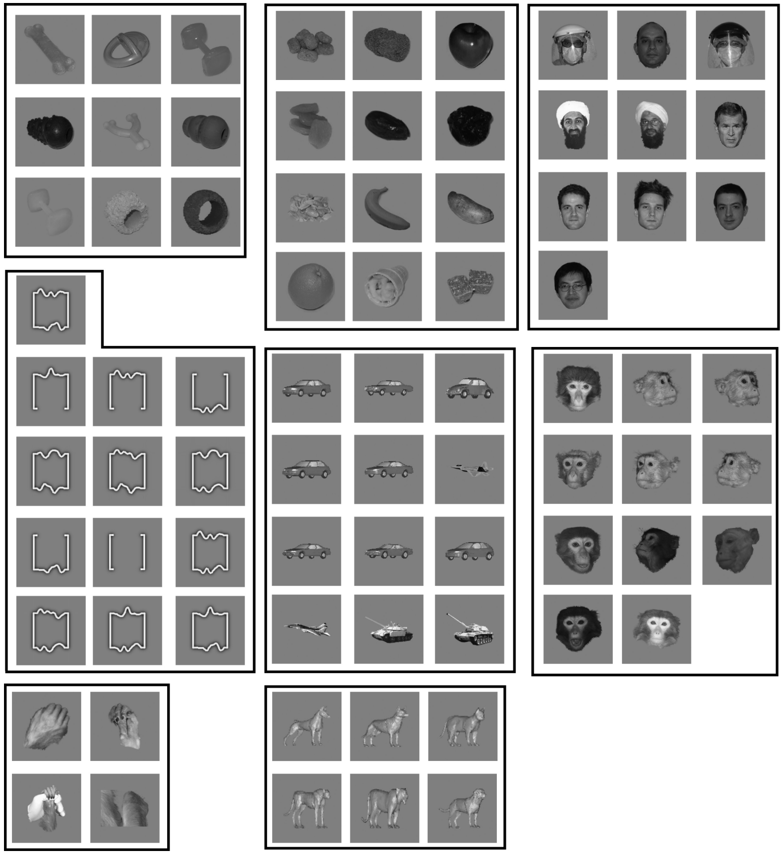
$$y = \frac{\sum_{j=1}^{n} w_j x_j}{k + \sum_{j=1}^{n} x_j^2}$$ Gaussian-like

$$y = \frac{\sum_{j=1}^{n} x_j^3}{k + \sum_{j=1}^{n} x_j^2}$$ Max-like

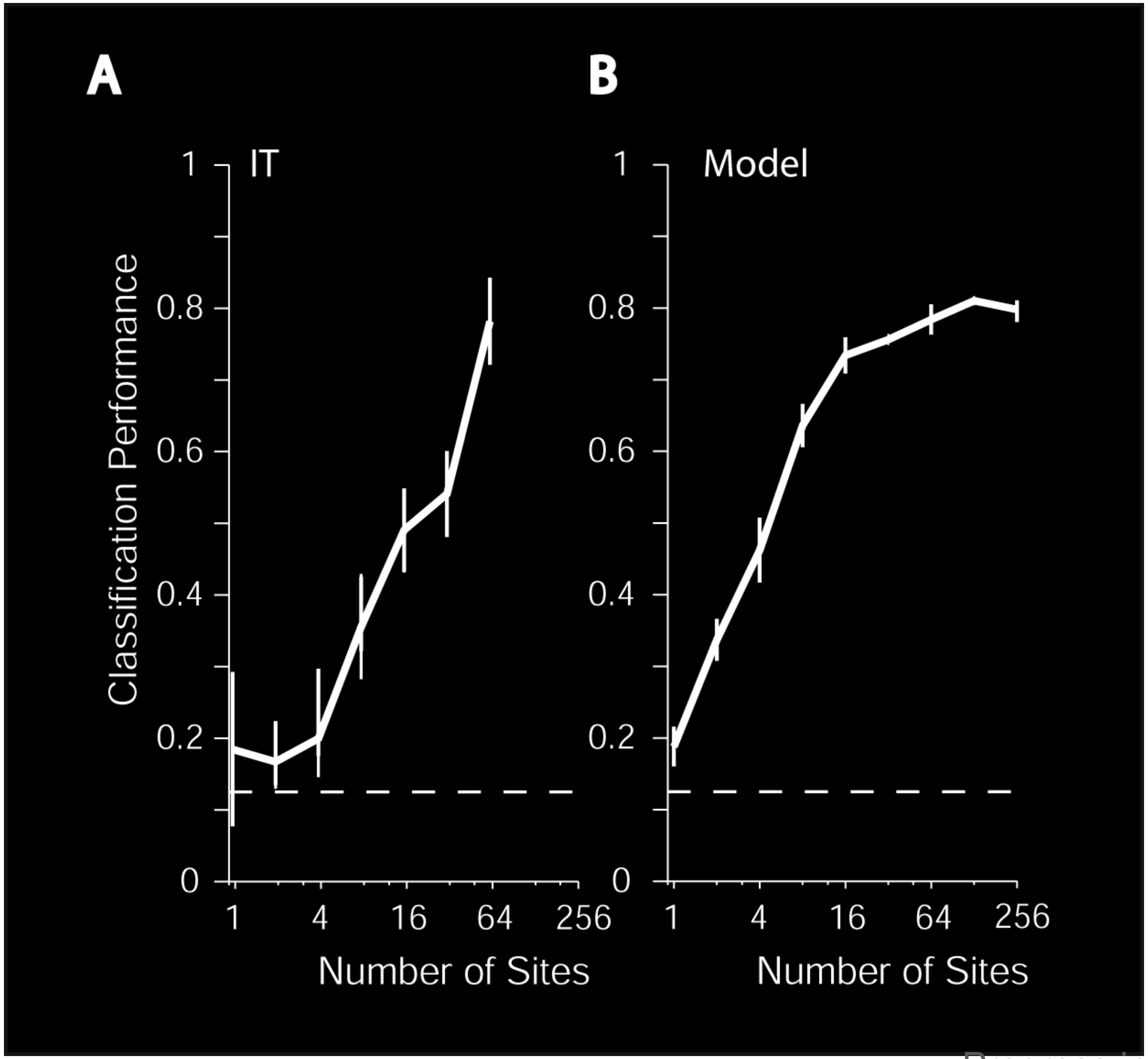
Neural Computation 2008

Example: responses of the top-level units



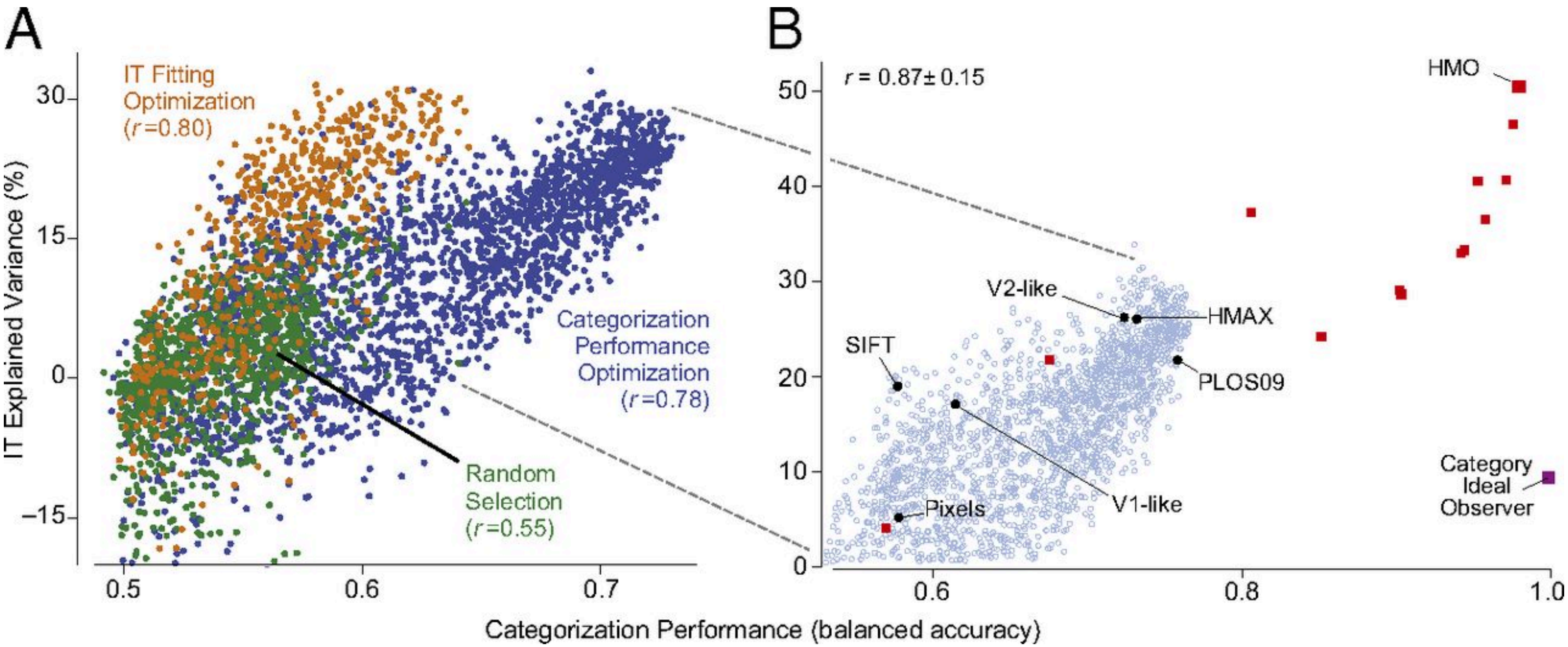


We can decode object information from the model units



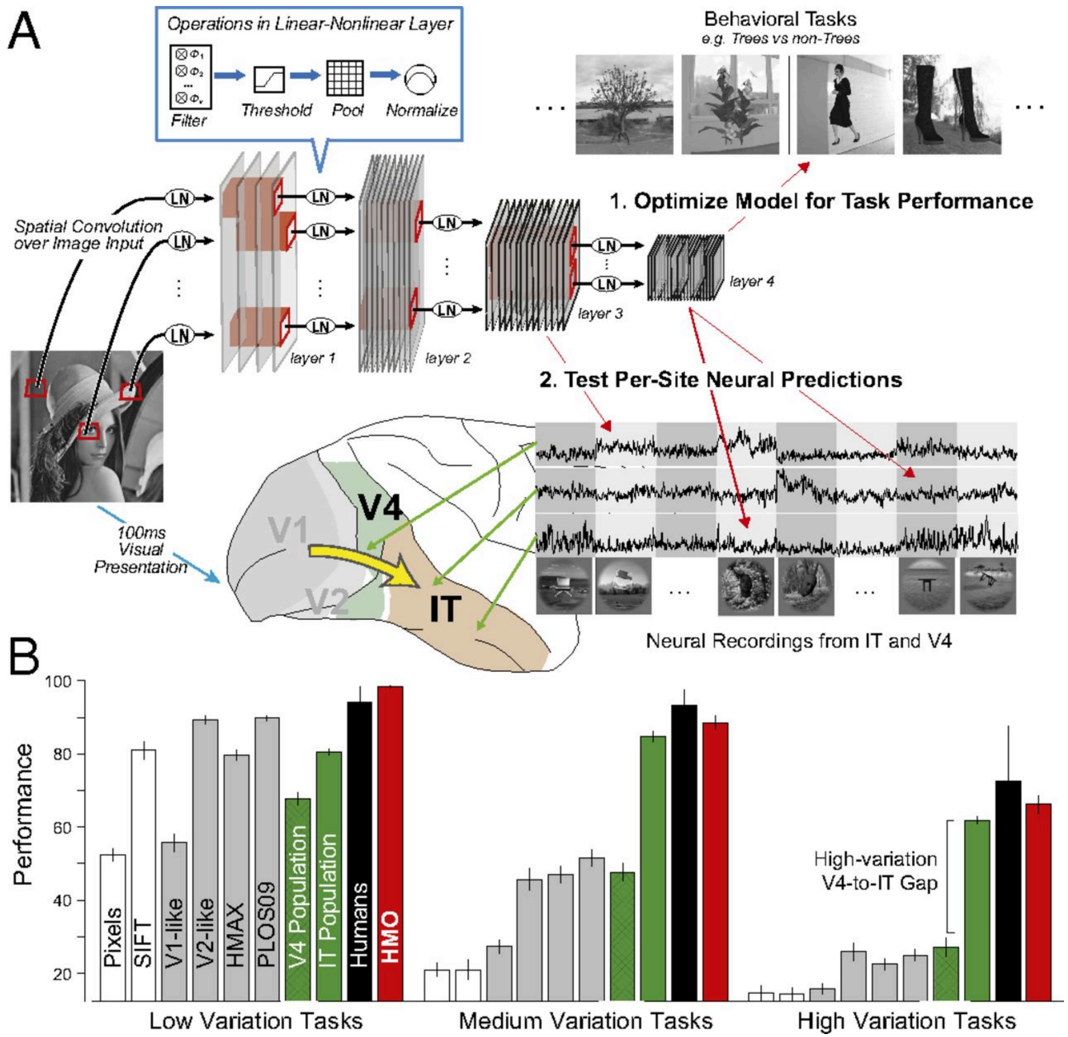
Progress in Brain Research 2007

Correlation between model performance and IT variance explained



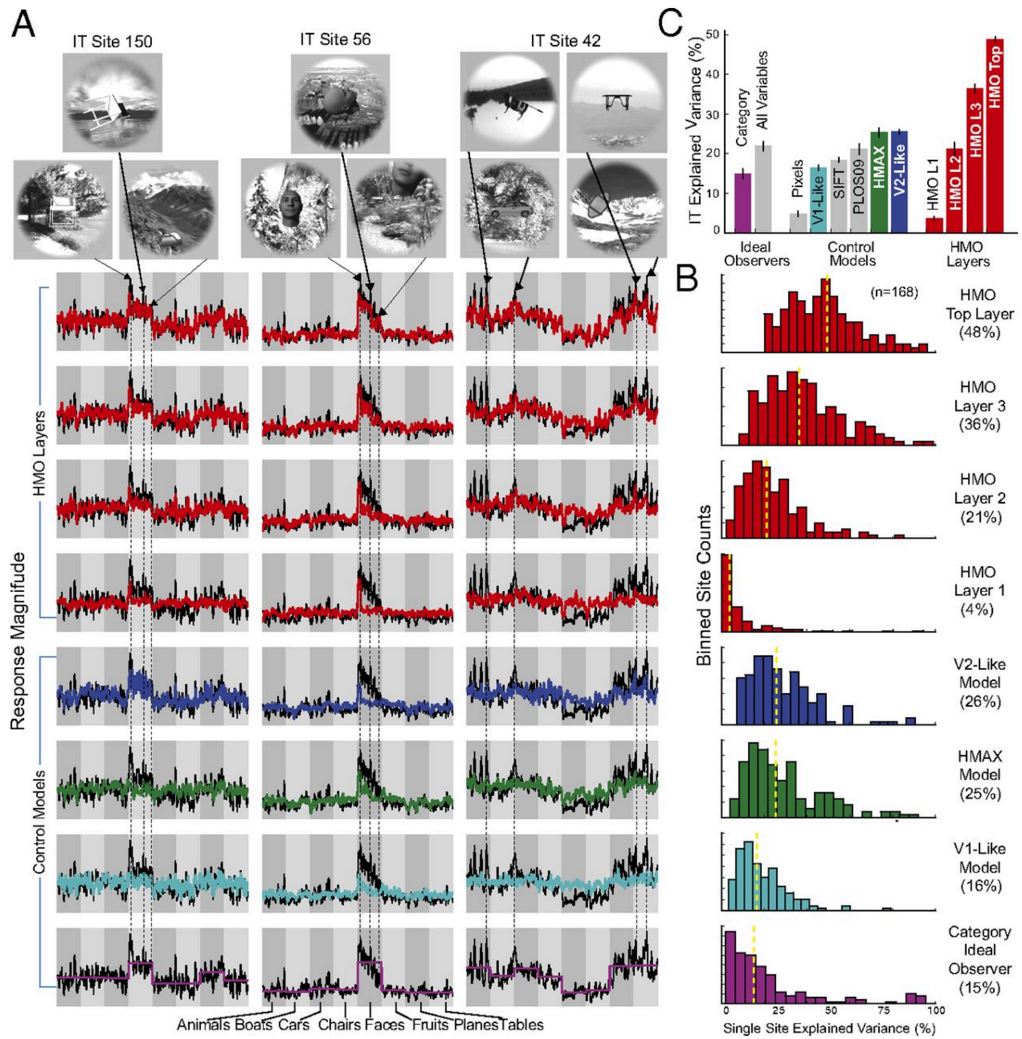
Daniel L. K. Yamins et al. PNAS 2014;111:8619-8624

Neural-like models via performance optimization



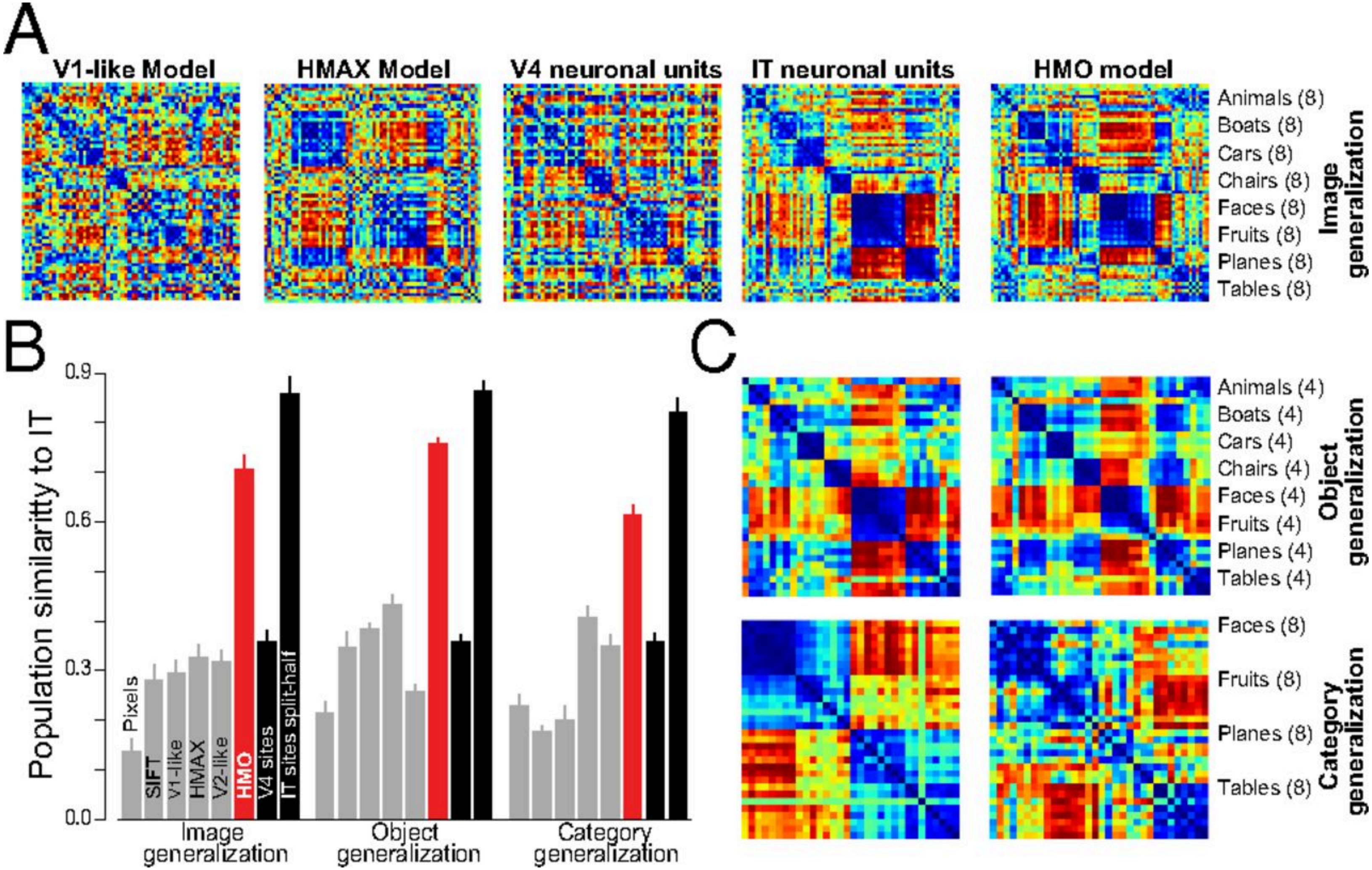
Daniel L. K. Yamins et al. PNAS 2014;111:8619-8624

Example IT neural predictions



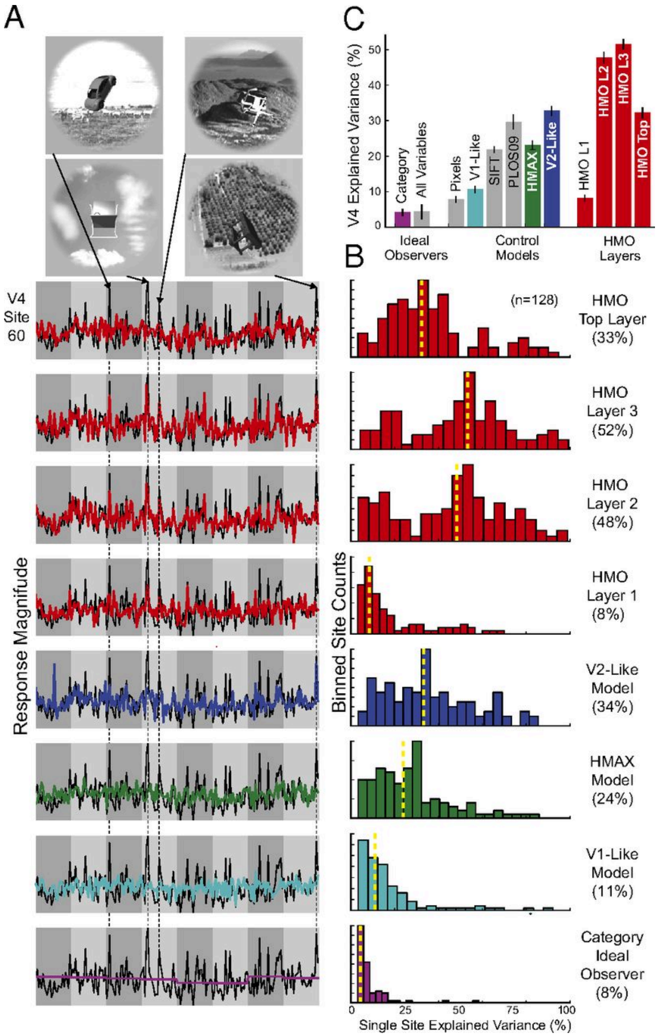
Daniel L. K. Yamins et al. PNAS 2014;111:8619-8624

Population-level similarity



Daniel L. K. Yamins et al. PNAS 2014;111:8619-8624

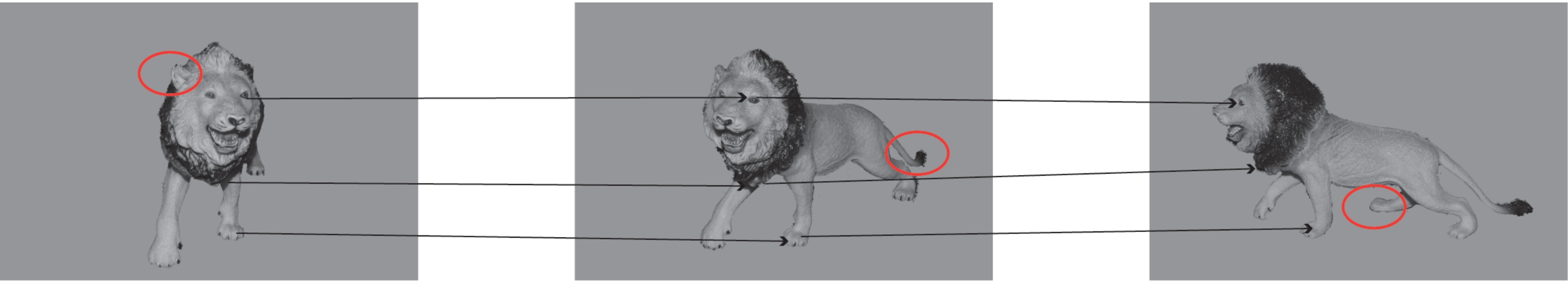
V4 neural predictions



Daniel L. K. Yamins et al. PNAS 2014;111:8619-8624

Object recognition by alignment to prototypes

Prototype



Alignment of 3 points to the prototype (black arrows)
Note: some points may not align (red ellipses)

Ullman (1996) High-level vision

Some ideas about viewpoint invariance: learning from examples

FIG. 1 *a*, Network representation of approximation by GRBFs. In a special simple case, there are as many basis functions (K) as views in the training set (M; in general, $K \leq M$). The centres of the radial functions are then fixed and are identical with the training views. Each basis unit in the 'hidden' layer computes the distance of the new view from its centre and applies to it the radial function. The resulting value $G(\|\mathbf{x}-\mathbf{t}_\alpha\|)$ can be regarded as the 'activity' of the unit. If the function G is gaussian, a basis unit will attain maximum activity when the input exactly matches its centre. The output of the network is the linear superposition of the activities of all the basis units in the network. *b*, An equivalent interpretation of *a* for the case of gaussian radial basis functions. A multidimensional gaussian function can be synthesized as the product of 2-D gaussian receptive fields operating on retinotopic maps of features. The solid circles in the image plane represent the 2-D gaussian functions associated with the first radial basis function, which corresponds to the first view of the object. The dotted circles represent the 2-D receptive fields that synthesize the gaussian radial function associated with another view. The gaussian receptive fields transduce positions of features represented implicitly as activity in a retinotopic array, and their product 'computes' the radial function without the need of calculating norms and exponentials explicitly.

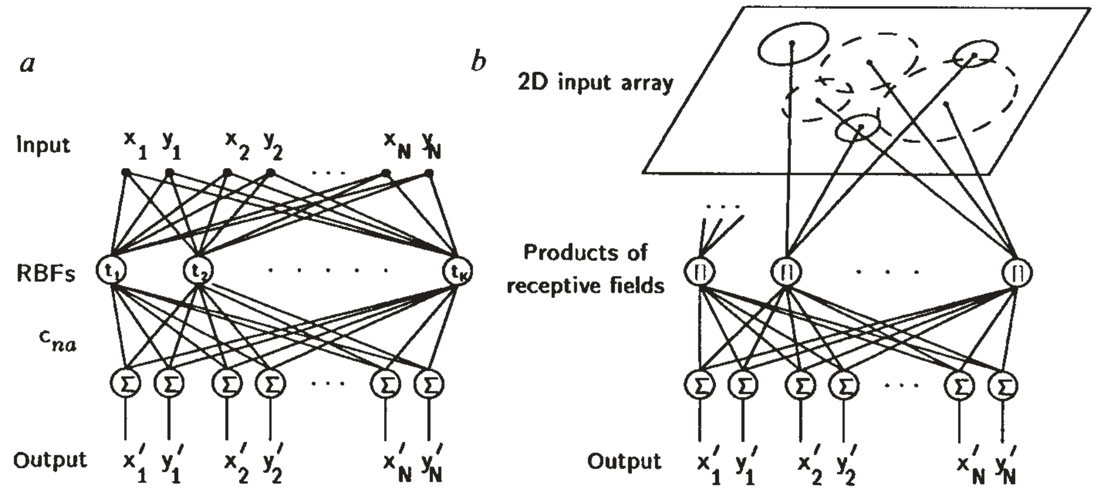
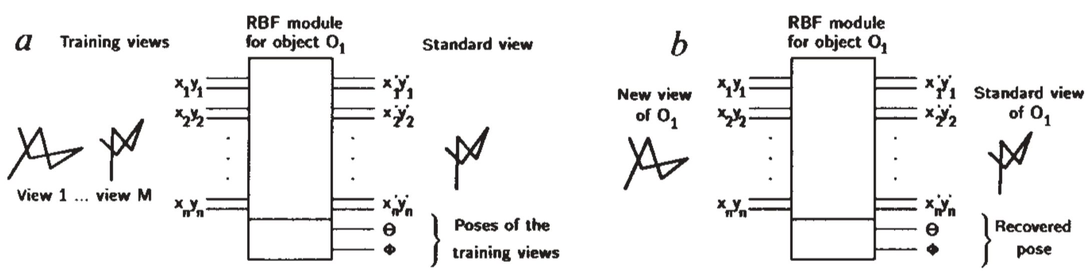
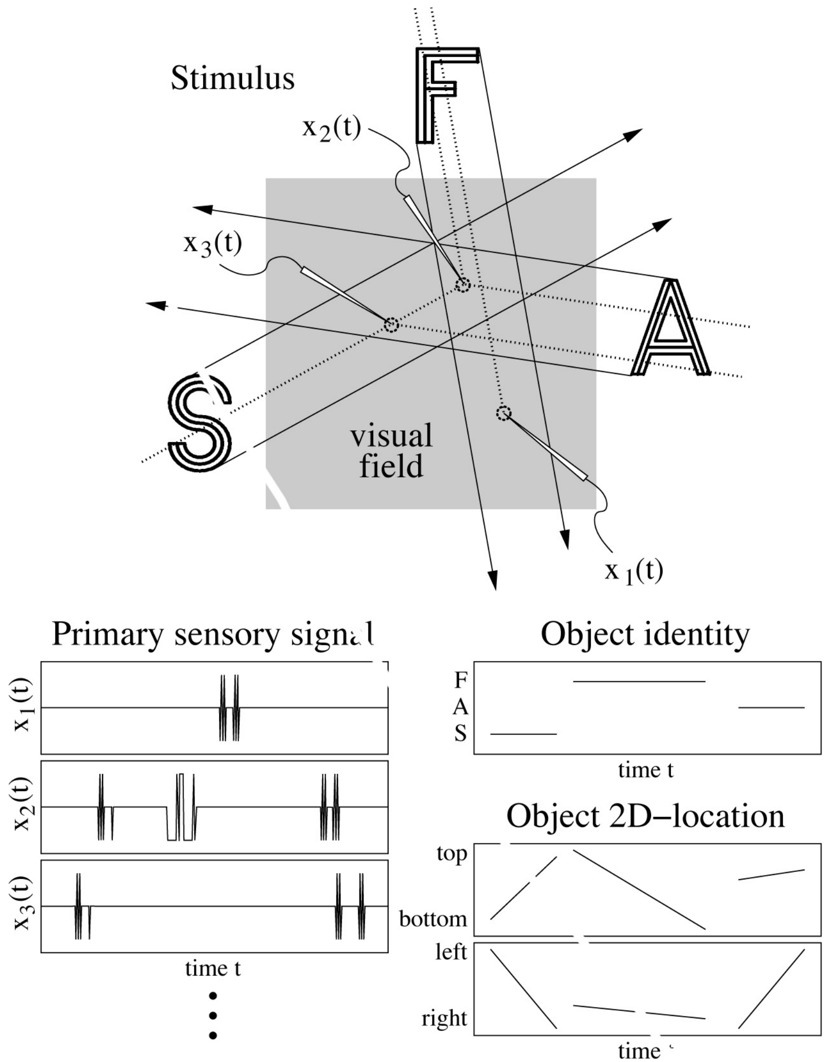


FIG. 2 Application of a general module for multivariate function approximation to the problem of recognizing a 3-D object from any of its perspective views. *a*, Module is trained to produce the vector representing the standard view of the object, given a set of examples of random perspective views of the same object. The module is also capable of recovering the viewpoint coordinates θ, ϕ (the latitude and the longitude of the camera on an imaginary sphere centred at the object) that correspond to the training views. When given a new random view of the same object (*b*), the module recognizes it by producing the standard view. Other objects are rejected by thresholding the euclidean distance between the actual output of the model and the standard view (this step corresponds to the action of a single radial function with a sharp cut-off centred on the standard view).



Poggio and Edelman, Nature 1990

Learning about object transformations by exploiting slowness



Foldiak et al 1991.
Wiskott & Sejnowski 2002

Bottom-up versus Top-down approaches

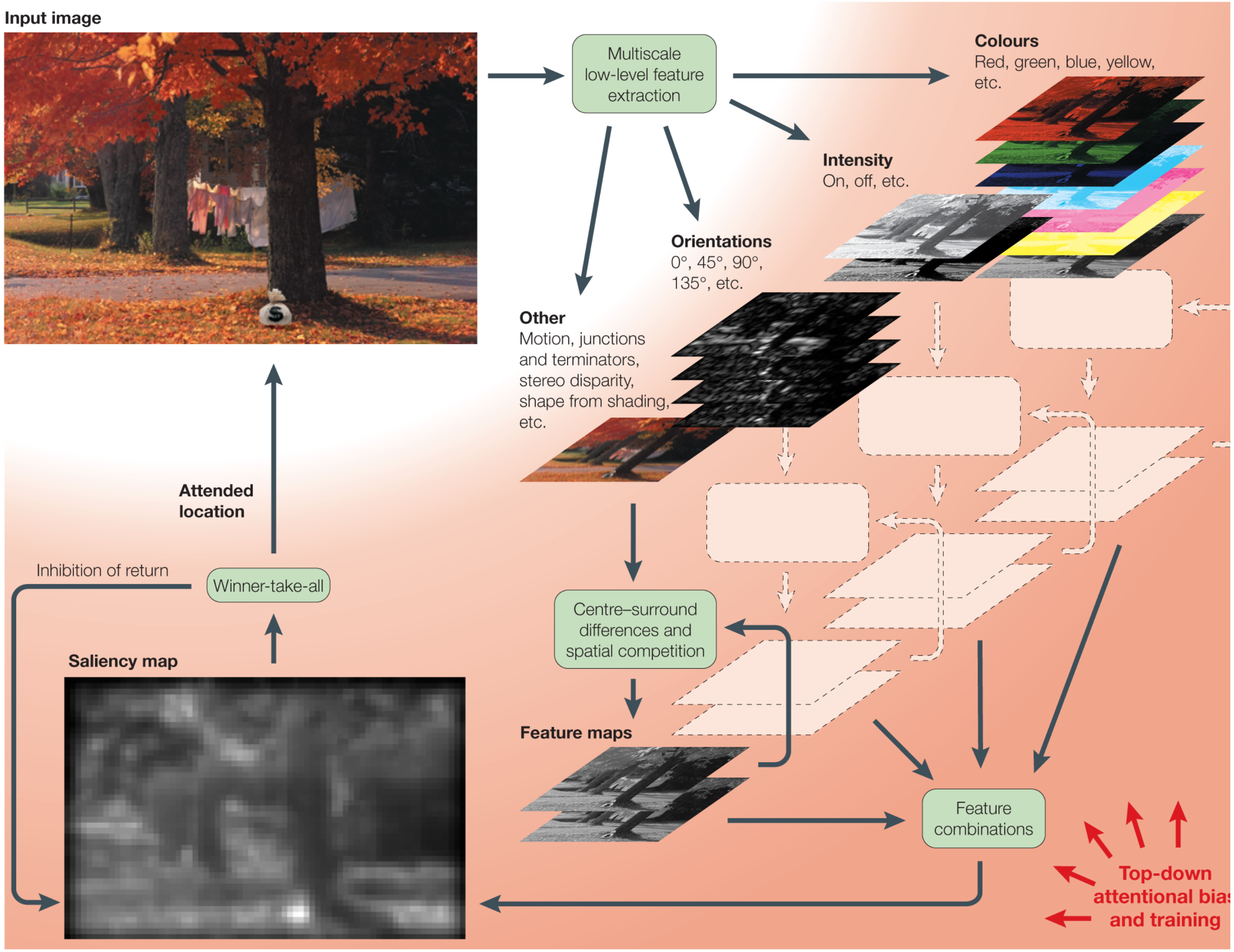
Bottom-up, horizontal and top-down connections intermixed throughout neocortex

The speed of visual recognition places a strong constraint on computational models:
- Scalp EEG: complex categorization by ~150 ms (Thorpe et al 1996)
- ITC responses show latencies of ~100 ms (e.g. Richmond et al 1983)
- Visual recognition in RSVP sequences (e.g. Potter et al 1969)

"Long" versus "short" loops in neuronal circuits underlying recognition
- "Short" loops: Horizontal connections; V1 → V2 → V1
- "Long" loops: ITC → V1 → ITC

There is more to life than vision… Memory, attention, emotions, planning, consciousness, etc. Top-down connections are likely to play key roles in this and other aspects of visual recognition.

Bottom-up saliency models



Itti and Koch 2001

Spatial and feature attention through feedback

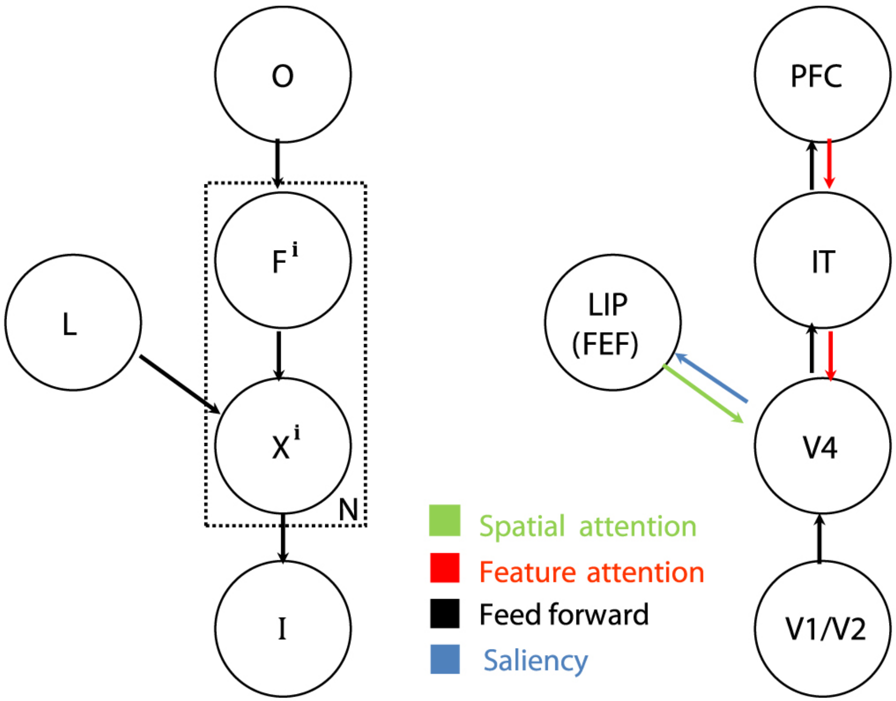
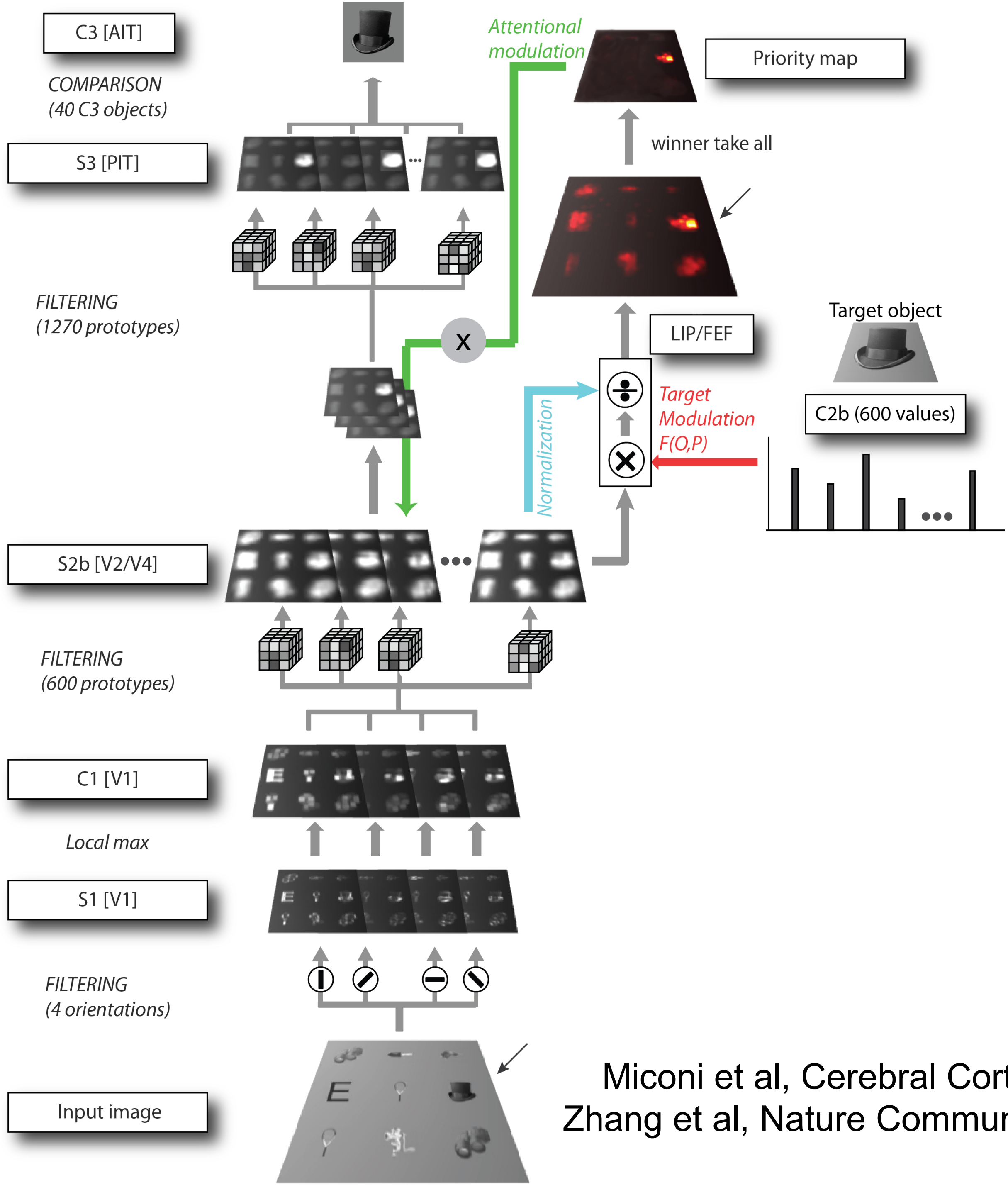


Fig. 2. Left: Proposed Bayesian model. Right: A model illustrating the interaction between the parietal and ventral streams mediated by feedforward and feedback connections. The main additions to the original feedforward model (Serre, Kouh, et al., 2005) (see also Supplementary Online Information) are (i) the cortical feedback within the ventral stream (providing feature-based attention); (ii) the cortical feedback from areas of the parietal cortex onto areas of the ventral stream (providing spatial attention); and (iii) feedforward connections to the parietal cortex that serves as a 'saliency map' encoding the visual relevance of image locations (Koch & Ullman, 1985).

Chikkerur and Poggio 2010

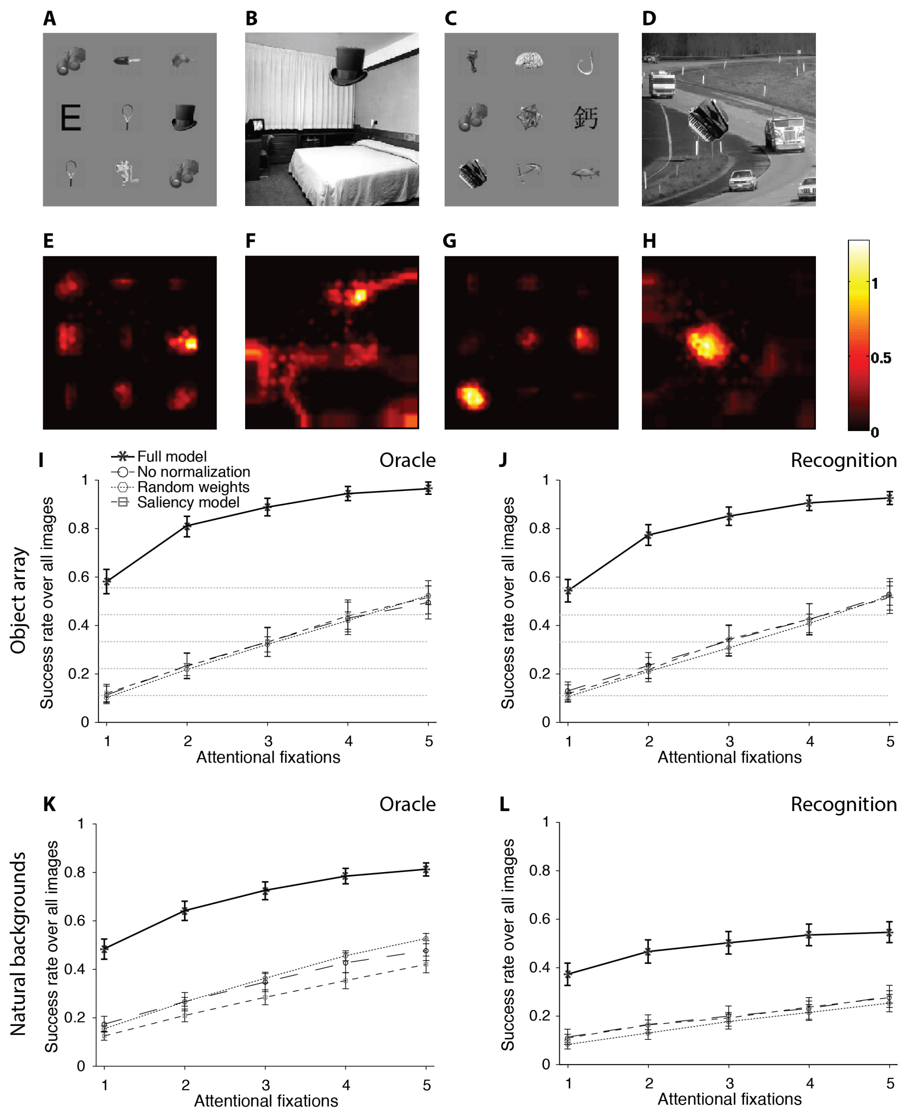
Top-down signals in visual search

Feedback signals in visual search

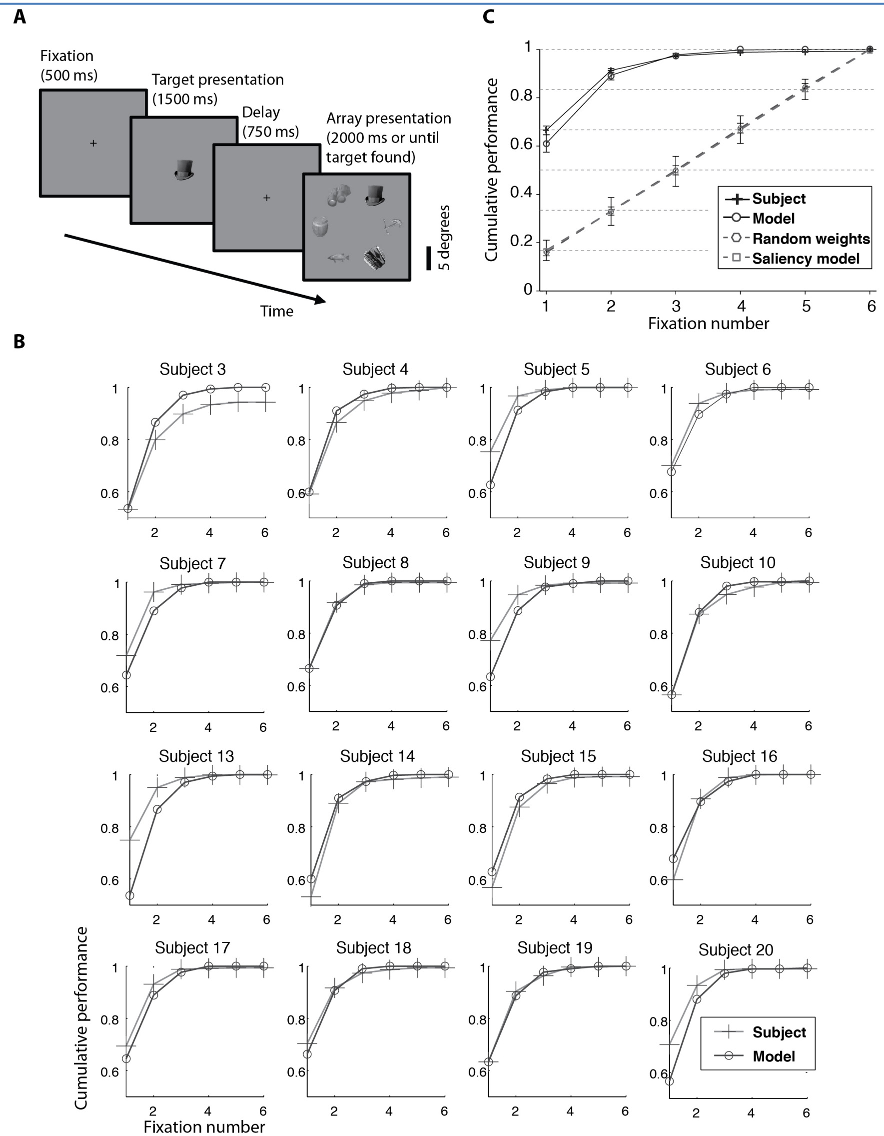


Miconi et al, Cerebral Cortex 2015
Zhang et al, Nature Communications 2018

The model can search for objects in cluttered images



The model's performance is comparable to human performance in the same visual search task



Further reading

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193-202.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. Progress In Brain Research, 165C, 33-56.

Deco, G., & Rolls, E. T. (2004). Computational Neuroscience of Vision. Oxford Oxford University Press.

Ullman, S. (1996). High-Level Vision. Cambridge, MA: The MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2(11), 1019-1025.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc of the IEEE, 86(11), 2278-2324.

Mel, B. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. Neural Computation, 9, 777.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J Neurosci, 13(11), 4700-4719.

Foldiak, P. (1991). Learning Invariance from Transformation Sequences. Neural Computation, 3, 194-200.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput, 14(4), 715-770.

Miconi and Kreiman (2016). There’s Waldo! A normalization model of visual search predicts single-trial human fixations in an object search task. Cerebral Cortex.

Yamins, D. L., et al. (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex." Proc Natl Acad Sci U S A **111**(23): 8619-8624.

Zhang, M., et al. (2018). "Finding any Waldo: zero-shot invariant and efficent visual search." Nat Commun.