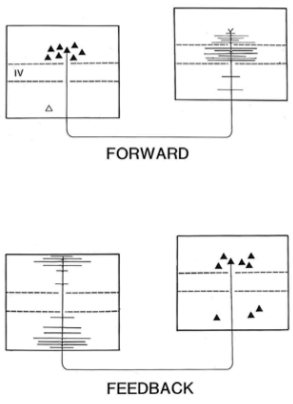
Object recognition and hierarchical computation

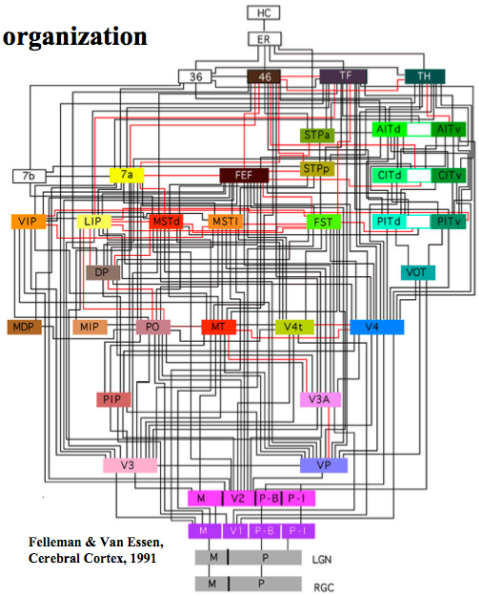
- Challenges in object recognition.
- Fukushima's Neocognitron
- View-based representations of objects
- Poggio's HMAX
- Forward and Feedback in visual hierarchy
- Hierarchical Bayes

15-883 Computational models of neural systems. Visual system lecture 2. Tai Sing Lee.

Distributed hierarchical organization

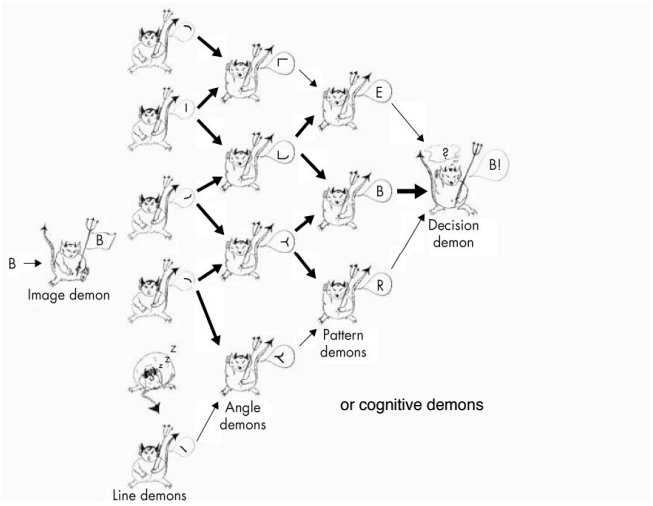


Maunsell & Van Essen, J. Neurophysiol., 1983



Felleman & Van Essen, Cerebral Cortex, 1991

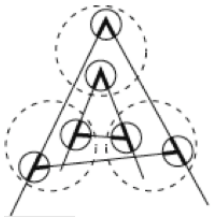
Selfridge's Pandemonium Model (1959) for object recognition



Challenges in object recognition

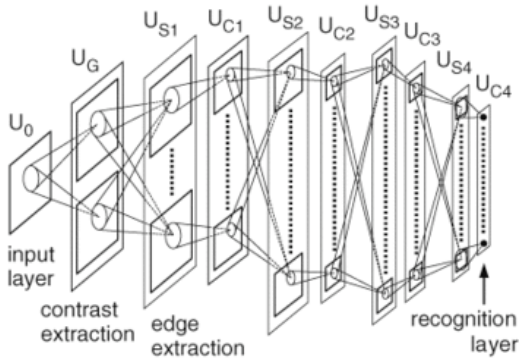
- What are the computational logic and concerns?

Invariance versus *Specificity*.

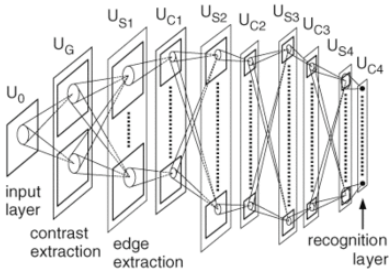


Fukushima's Neocognitron (1980)

A hierarchical multi-layered neural network for visual pattern recognition.

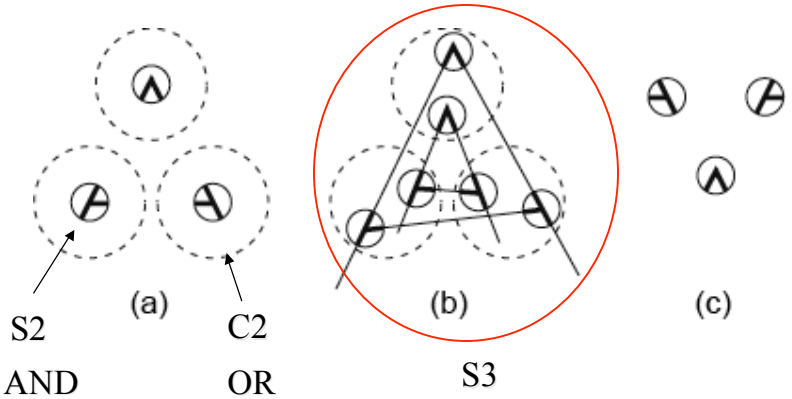


Generalization (invariance) vs discrimination (specificity)



Generalizing simple-complex cells: alternating S and C layers: S - feature detectors (e.g. simple cells) detect conjunction of features (AND, specificity), C - invariance pooling (e.g. complex cells) (OR, MAX, invariance).

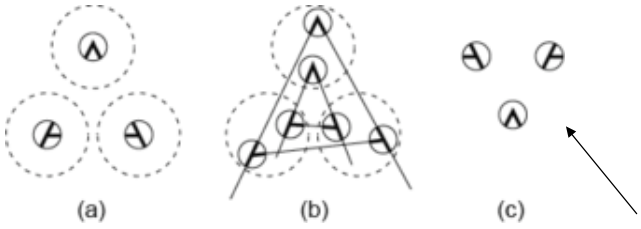
Gradual specificity and invariance



Conjunction of the C2 feature detectors

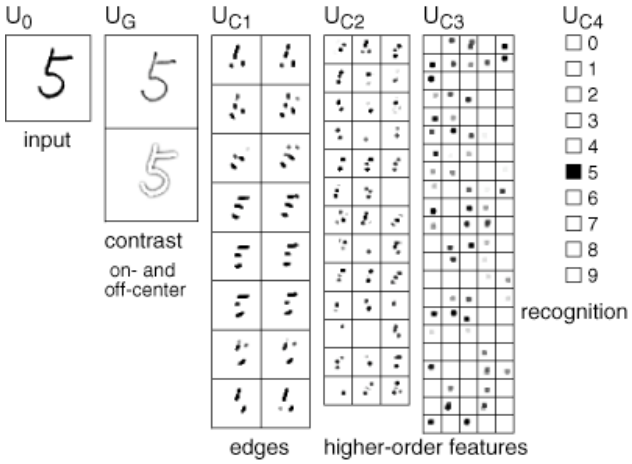
Hierarchical Features

Local features gradually integrated into more global features. With C stage performing `blurring', the next S stage detects more global features even if the components are deformed and shifted.



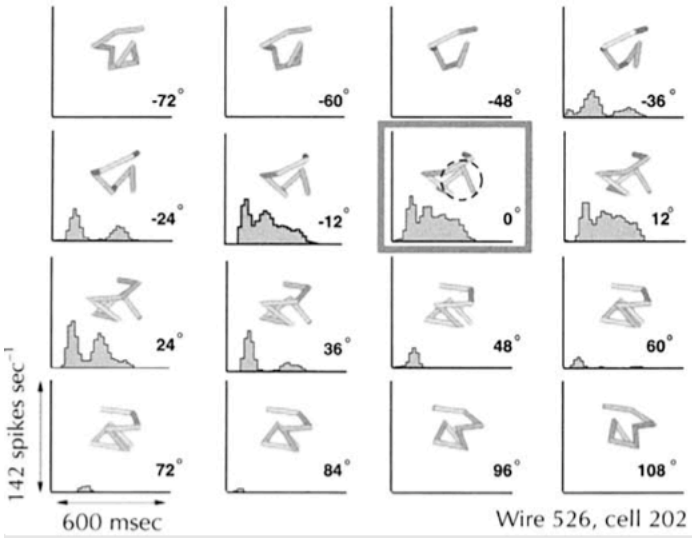
Not tolerated.

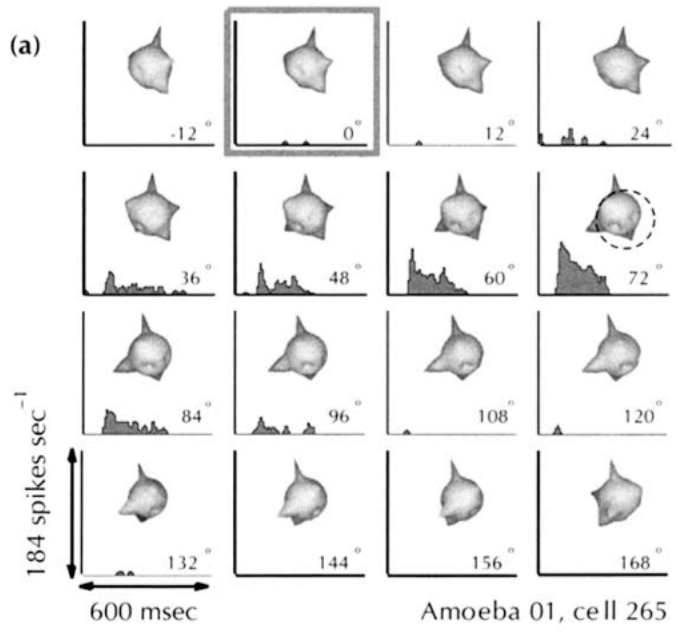
Character recognition



Revolt against 3D model

- Psychophysical experiments (Buelthoff 1992; Gauthier 1997) and neurophysiological experiment (Logothetis 1995) provided strong support for the view-based representation of objects.
- Tested novel object in a particular view, people and monkeys tend to recognize the novel object only within +/- 40 degrees in rotation.
- Neurons also exhibit Gaussian tuning curves with peaks 40-50 degrees apart.


-72°
-60°
-48°
-36°
-24°
-12°
0°
12°
24°
36°
48°
60°
72°
84°
96°
108°
142 spikes sec⁻¹
600 msec
Wire 526, cell 202


(a)
-12°
0°
12°
24°
36°
48°
60°
72°
84°
96°
108°
120°
132°
144°
156°
168°
184 spikes sec⁻¹
600 msec
Amoeba 01, cell 265

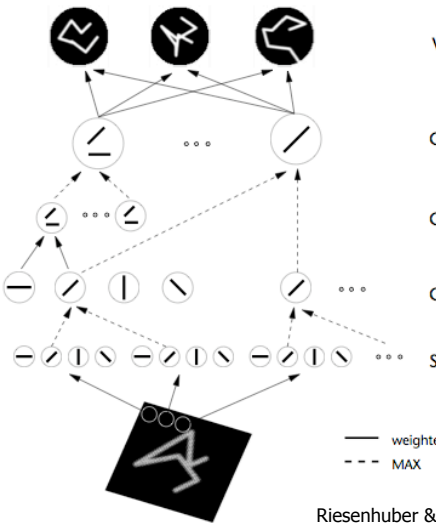
Basic facts

- IT -- neurons coded for object, input to prefrontal cortex.
- IT neurons are relatively scale and position invariance, but view-point and lighting dependence.
- View-point interpolation between different object views to achieve view-point invariance object recognition.
- Learning specific to an individual novel object is not required to be scale and translation invariant.
- Recognition can be very fast. 8/seconds, possibly mediated by a feed-forward model of ventral stream processing.

Basic motivation for the HMAX model

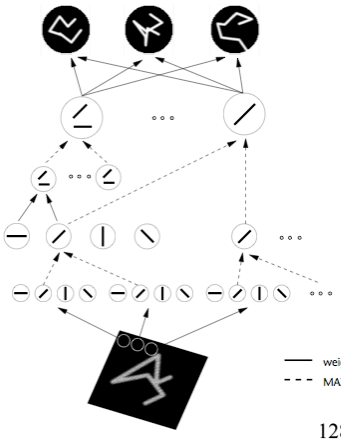
- Generalizing simple cell to complex cell relationship -- invariance to changes in the position of an optimal stimulus (within a range) is obtained in the model by means of a *maximum operation* (max) performed on the simple cell inputs to the complex cells, where the strongest input determines the cell's output. This preserves feature specificity.
- The model alternates layers of units combining simple filters into more complex ones - to increase pattern selectivity with layers based on the max operation - to build invariance to position and scale while preserving pattern selectivity.
- The RBF (Radial Basis Function) -like learning network learns a specific task based on a set of cells tuned to example views.

Hierarchical model of object recognition (HMAX)



Riesenhuber & Poggio, *Nature Neuroscience*, 1999

Hierarchical model and X (HMAX)



128x128 input

Supervised training with one hidden layer

256 units, each max of all scales all positions

4 neighboring (non-overlapping) C1 units form a conjunctive S2 unit - 256 types in each band.

Spatial pooling (4-12 depending on input RF size), 4 scales band.

4 orientations, 2 pixel apart, 7x7 to 29x29 scales, normalized.

http://maxlab.neuro.georgetown.edu/hmax.html#c2

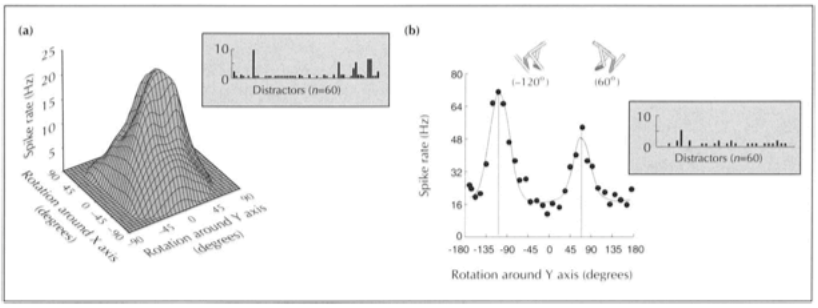


Fig. 4. (a) Response of a view-selective neuron to rotations around the preferred view along four axes. The *z* dimension of the plot is the spike rate, and the *x* and *y* dimensions show the degrees of rotation of the target object around either or both of the X or Y axes, respectively. The volume was generated by testing the cell's response for rotations out to ± 60 ° around the X and Y axes as well as along the two diagonals. The magnitude of response declined by approximately the same extent for rotations away from 0° along all of the axes tested. The activity of the neuron for the 60 distractors is shown in the inset box. Each distractor was a view of a different wire object. **(b)** Response of a neuron selective for pseudo-mirror-symmetric views, 180° apart, of a wire-like object. The filled circles are the mean spike rates for target views around one axis of rotation. The black line is the view-tuning curve obtained by 'distance-weighted least squares' (DWLS) smoothing. The two inset images depict the –120 ° and 60 ° views, around both of which the neuron showed view-selective tuning. The activity of the neuron for the 60 different distractor objects used during testing is shown in the inset box.

Logothetis, Pauls and Poggio Current Biology 1995.

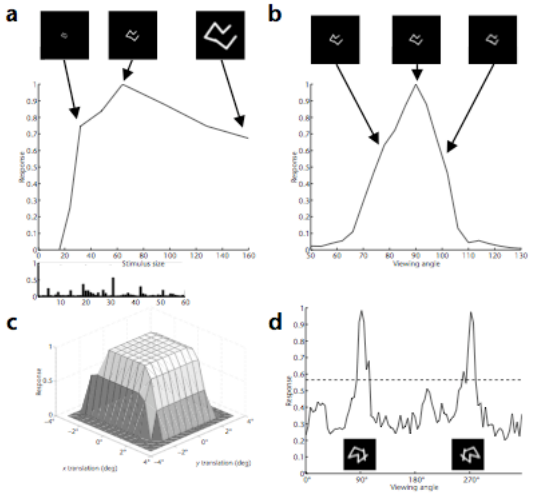
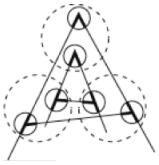


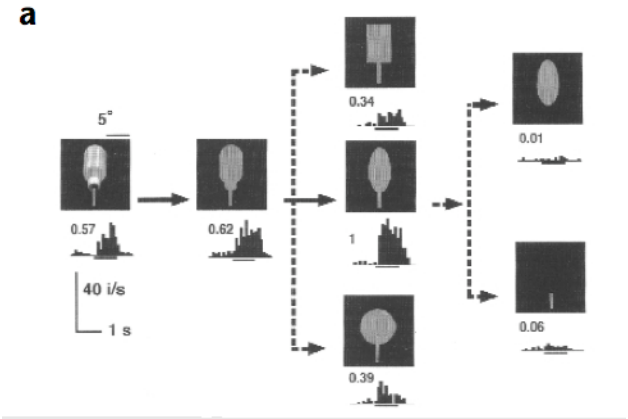
Fig. 4. Responses of a sample model neuron to different transformations of its preferred stimulus. Panels show the same neuron's response to (**a**) varying stimulus sizes (inset shows response to 60 distractor objects, selected randomly from the paperclips used in the physiology experiments[21]) (**b**) rotation in depth and (**c**) translation. Training stimulus size was 64 × 64 pixels, corresponding to 2° of visual angle. (**d**) Another neuron's response to pseudo-mirror views (see text), with the dashed line indicating the neuron's response to the 'best' distractor.

Why does it work at all?

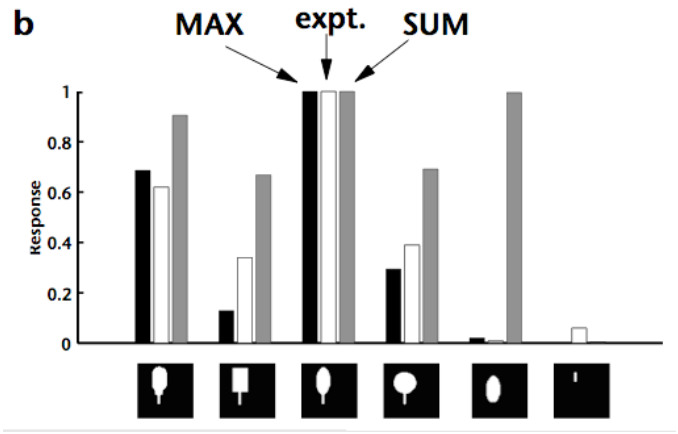
- S2 computes conjunction of oriented features within a neighborhood to create 256 types of higher order features for each scale band and at every position.
- C2 pools over all positions and scale bands, thus is a bag of features.
- This bag of feature approach, also popular in computer vision at the time, apparently is sufficient for discriminating many objects. WHY?
- Because there are a lot of features, some of the features are rather unique and discriminative. It is unlikely two objects tested will share the same set of conjunctive features.



Highly nonlinear integration of component parts



Nonlinear MAX operation approximates physiology better

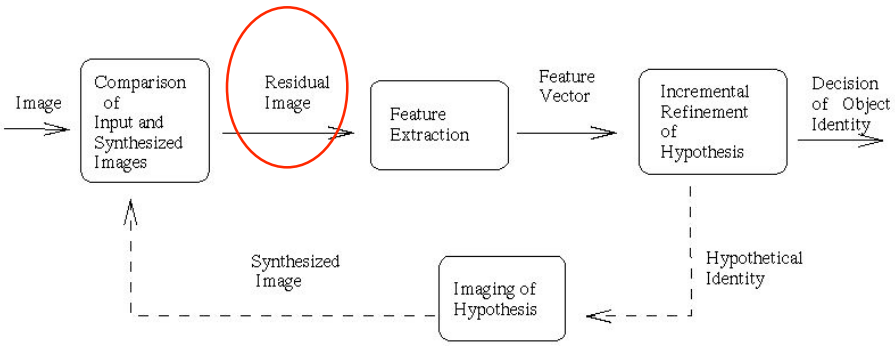


But what does feedback do?

- Attention -- feature and spatial attentional selection.
- Interactive activation -McClelland and Rumelhart, Adaptive resonance - Grossberg, bringing in global structures and context to influence the interpretation of low-level interpretation and pattern completion.
- A modern view: generative model and hierarchical Bayes.

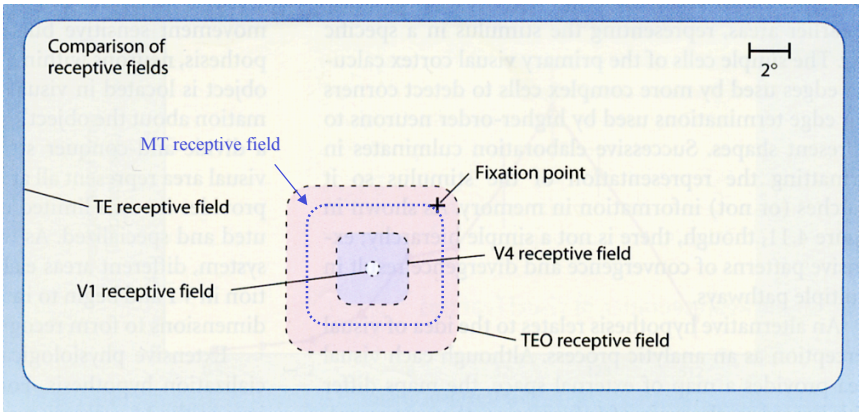
A modern view:
generative model and hierarchical Bayesian inference.

Analysis by Synthesis



Mumford (1992)

Multi-scale analysis across visual areas.



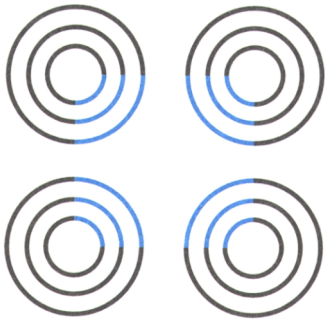
Increase in RF size and complexity of encoded features along the visual hierarchy.

Lee and Mumford's conjecture
High-resolution buffer hypothesis of V1

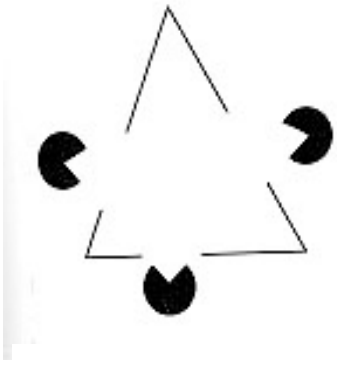
V1 is not simply a filter-bank for extracting features, but is a unique high-resolution buffer that is used by higher visual cortex for performing any visual inference that requires spatial precision and high-resolution feature details.

Lee, Mumford, Romero and Lamme (1998), Lee and Mumford (2003)

The interplay of priors, context and memory in visual inference.

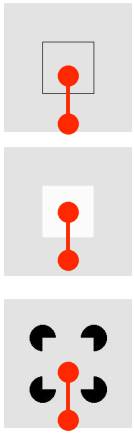


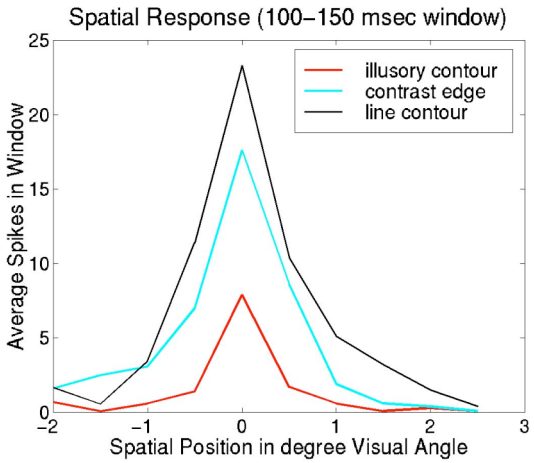
The need to see depth and segregate surfaces is so strong that we can hallucinate surface and contours at locations where there is no visual evidence for them.



Marr (1981)

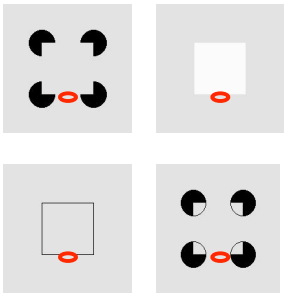
Spatial response to the illusory contour is found at precisely the expected location at V1

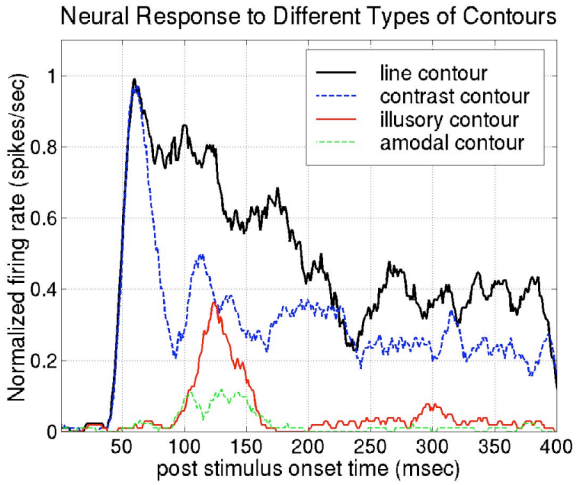




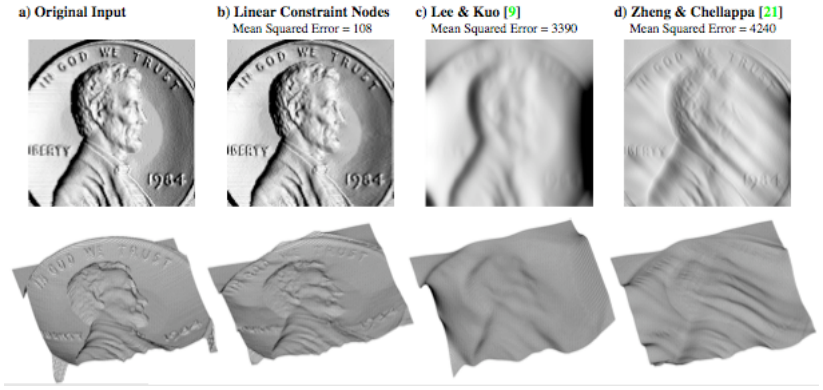
Lee and Nguyen 2001 PNAS.

Temporal Response to the Illusory Contours at 100 msec, 40 msec later than V2





Shape from shading with simple priors with factor graph and BBP and higher order priors



Potetz CVPR (2007).

Texton/token primal sketch

(a) dictionary Δ_B (b) input image (c) reconstructed $K = 300$

Zhu and Mumford (2997)

Dictionary of visual primitives

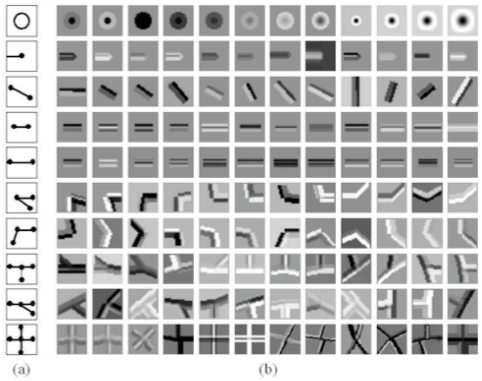
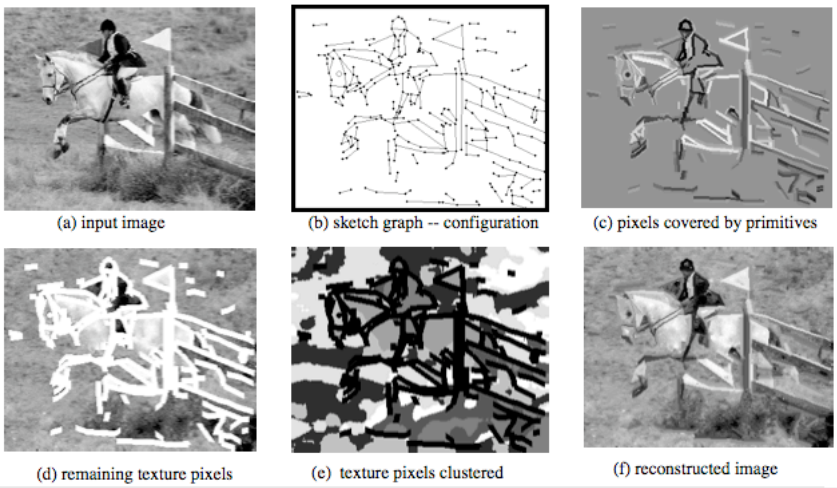


Figure 3: Samples from the visual primitive dictionary, consisting of eight types: blobs, end points, edges, ridges, multi-ridges, corners, junctions and crosses of different degrees. (a) The landmarks on the patches for topological and geometric attributes. (b) The photometric representation of the patches.

One implementation of Primal Sketch

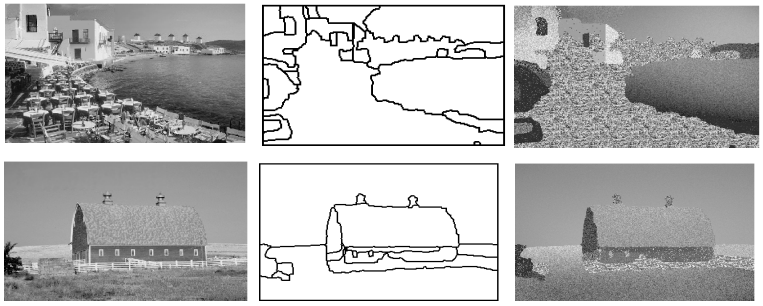


(a) input image (b) sketch graph -- configuration (c) pixels covered by primitives

(d) remaining texture pixels (e) texture pixels clustered (f) reconstructed image

Segmentation as simplifying parsing

Input | Segmentation | Synthesis from model $I \sim p(I \mid W^*)$



Hidden variables describe segments and their texture, allowing both slow and abrupt intensity and texture changes.

Hierarchical Generative models

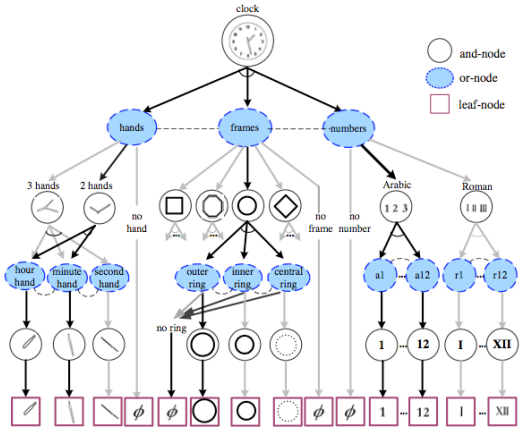


FIGURE 27. An And-Or graph example for the object category – clock. It has two parsing graphs shown in Figure 25, one of which is illustrated in dark arrows. Some leaf nodes are omitted from the graph for clarity. From [73].

Zhu and Mumford (2007) Quest for a stochastic grammar for vision

Parsing graph in response to object in an image

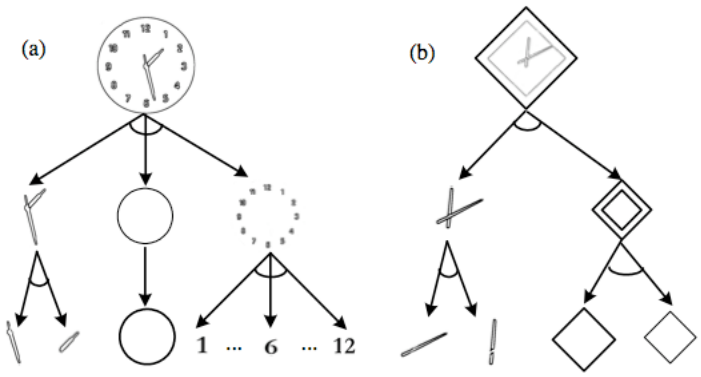


FIGURE 25. Two parsing graph examples for clocks which are generated from the And-Or-graph in Fig. 27. From [73].

Generation of clock images by hallucination

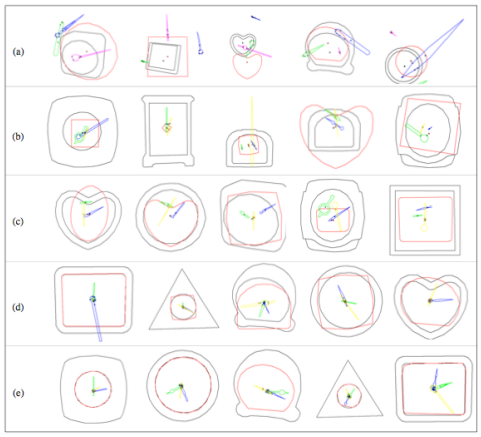


FIGURE 28. Learning the And-Or graph parameters for the clock category. (a) Sampled clock examples (synthesis) based on SCFG (Markov tree) that accounts for the frequency of occurrence. (b-e) Synthesis examples at four incremental statges of the minimax entropy pursuit process. (b) Matching the relation positions between parts, (c) further matching the relative scales, (d) further pursuing the hinge relation, (e) further macthing the containing relation. From [49]

Summary

- Invariance and specificity in object recognition.
- Fukushima's Neocognitron
- View-based representations of objects
- Poggio's HMAX
- Forward and Feedback in visual hierarchy
- Feedback for generating explanation for images
- Hierarchical Bayes and generative models
- High-resolution buffer hypothesis of V1

Readings

- Riesenhuber, M. & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience* 2: 1019-1025.
- Lee, T.S., Mumford, D. (2003) Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America, A.* . 20(7): 1434-1448.