Lecture 6. Object Recognition

Reading Assignments:

None

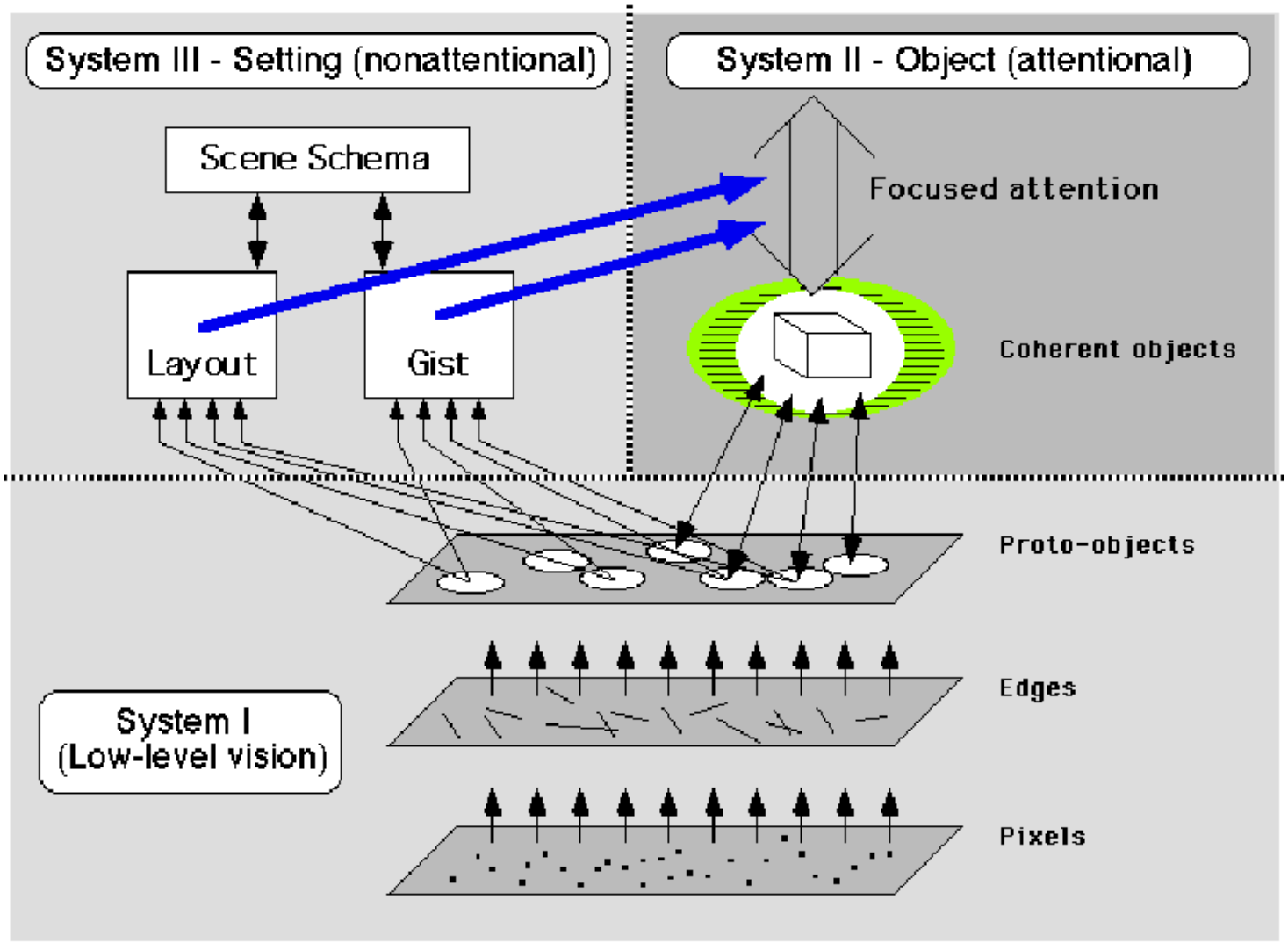
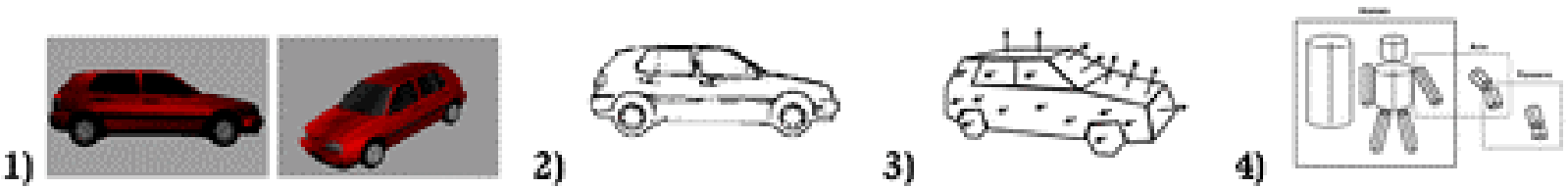


Figure 4

Figure 4. Triadic Architecture. It is suggested that the visual perception of scenes may be carried out via the interaction of three different systems. System I: Early-level processes produce volatile proto-objects rapidly and in parallel across the visual field. System II: Focused attention acts as a hand to "grab" these structures; as long as these structures are held, they form an individuated object with both temporal and spatial coherence. System III: Setting information—obtained via a nonattentional stream—guides the allocation of focused attention to various parts of the scene, and allows priorities to be given to the various possible objects.

Four stages of representation (Marr, 1982)



1) pixel-based (light intensity)

2) primal sketch (discontinuities in intensity)

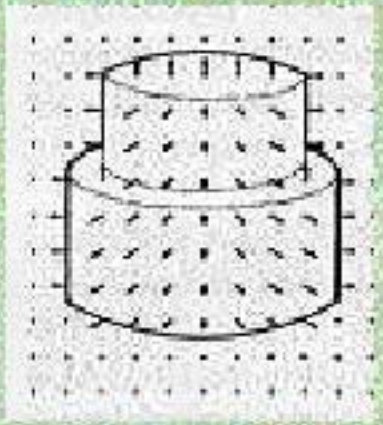
3) 2 ½ D sketch (oriented surfaces, relative depth between surfaces)

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

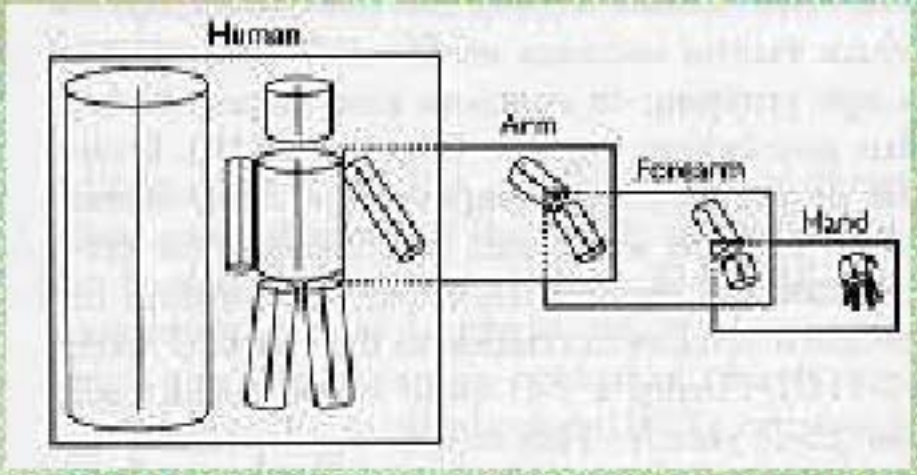
Real world example of Marr

Retinal image

Primal sketch



2 1/2 D sketch



3D model

Challenges of Object Recognition

The binding problem: binding different features (color, orientation, etc) to yield a unitary percept. (see next slide)

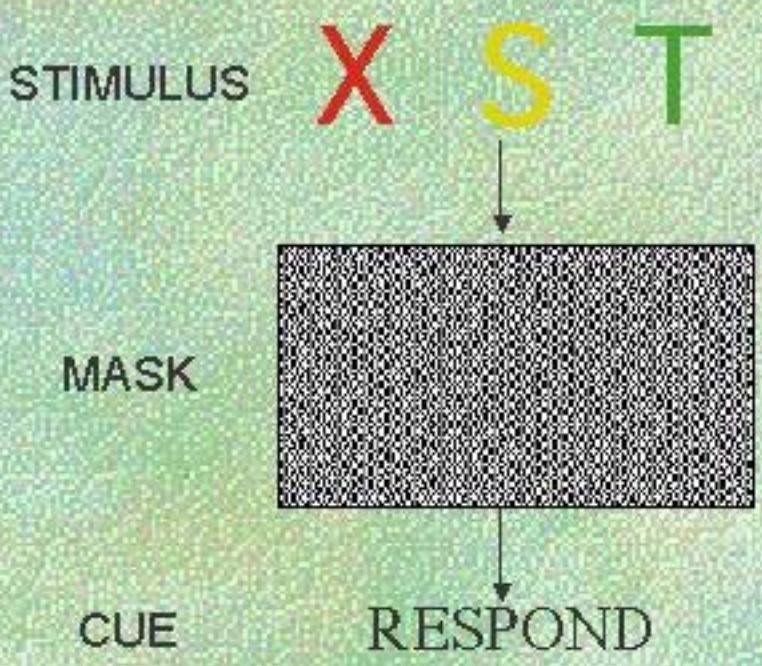
Bottom-up vs. top-down processing: how much is assumed top-down vs. extracted from the image?

Perception vs. recognition vs. categorization: seeing an object vs. seeing is *as* something. Matching views of known objects to memory vs. matching a novel object to object categories in memory.

Viewpoint invariance: a major issue is to recognize objects irrespectively of the viewpoint from which we see them.

illusory conjunctions

- Present colored letters very briefly, then mask
- On report, subjects correctly report the letters and the color but on ~30% or trials, attach wrong color to letter
  - e.g. Green X



Viewpoint Invariance

Major problem for recognition.

Biederman & Gerhardstein, 1994:

We can recognize two views of an unfamiliar object as being the same object.

Thus, viewpoint invariance cannot only rely on matching views to memory.

Models of Object Recognition

See Hummel, 1995, The Handbook of Brain Theory & Neural Networks

Direct Template Matching:

Processing hierarchy yields activation of view-tuned units.

A collection of view-tuned units is associated with one object.

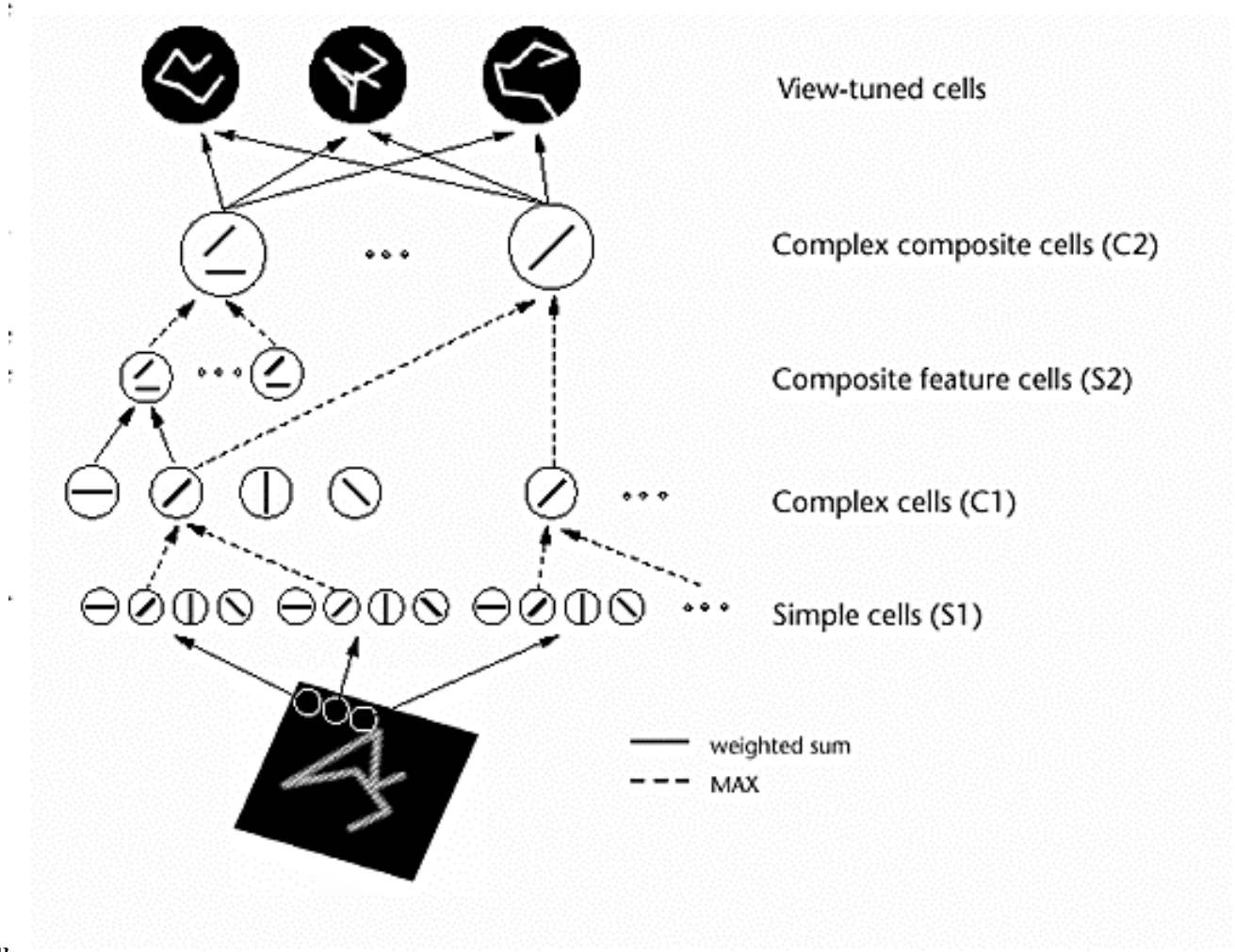
View tuned units are built from V4-like units,

using sets of weights which differ for each object.

e.g., Poggio & Edelman, 1990; Riesenhuber & Poggio, 1999

Computational Model of Object Recognition

(Riesenhuber and Poggio, 1999)



the model neurons are
tuned for size
and 3D orientation
of object

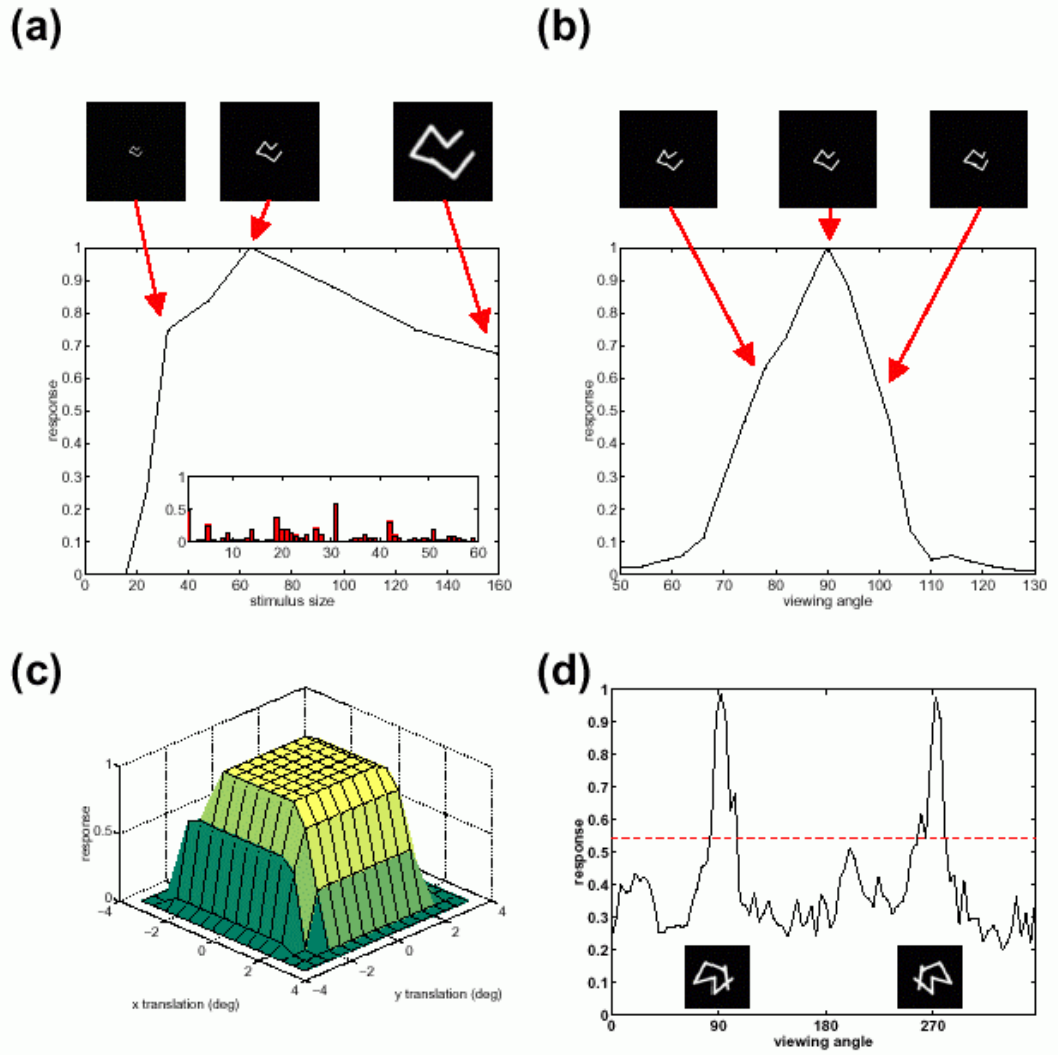


Figure 4: Responses of a sample model neuron to different transformations of its preferred stimulus. The different panels show the same neuron's response to **(a)** varying stimulus sizes (inset shows response to 60 distractor objects, selected randomly from the paperclips used in the physiology experiments[21]),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

Models of Object Recognition

Hierarchical Template Matching:

Image passed through layers of units with progressively more complex features at progressively less specific locations.

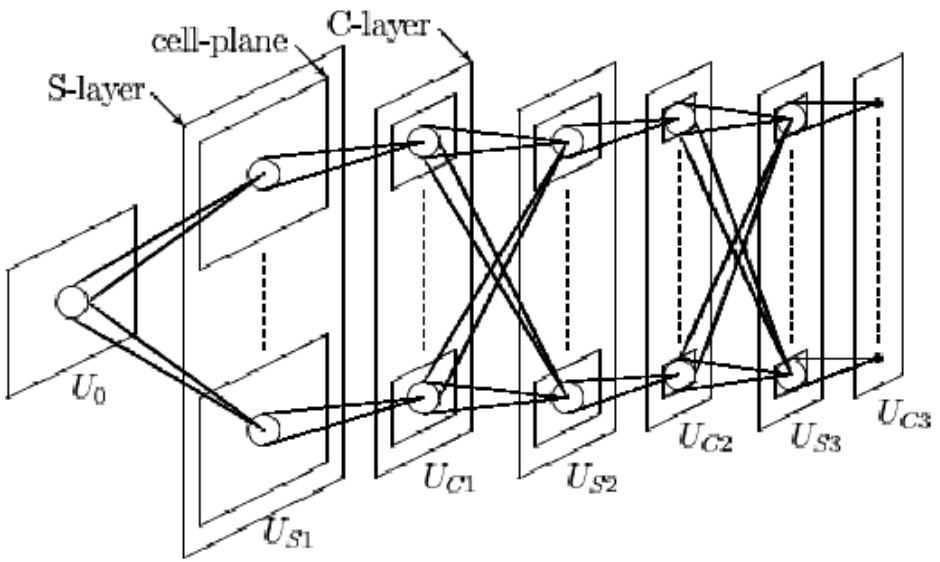
Hierarchical in that features at one stage are built from features at earlier stages.

e.g., Fukushima & Miyake (1982)'s Neocognitron:

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.



Models of Object Recognition

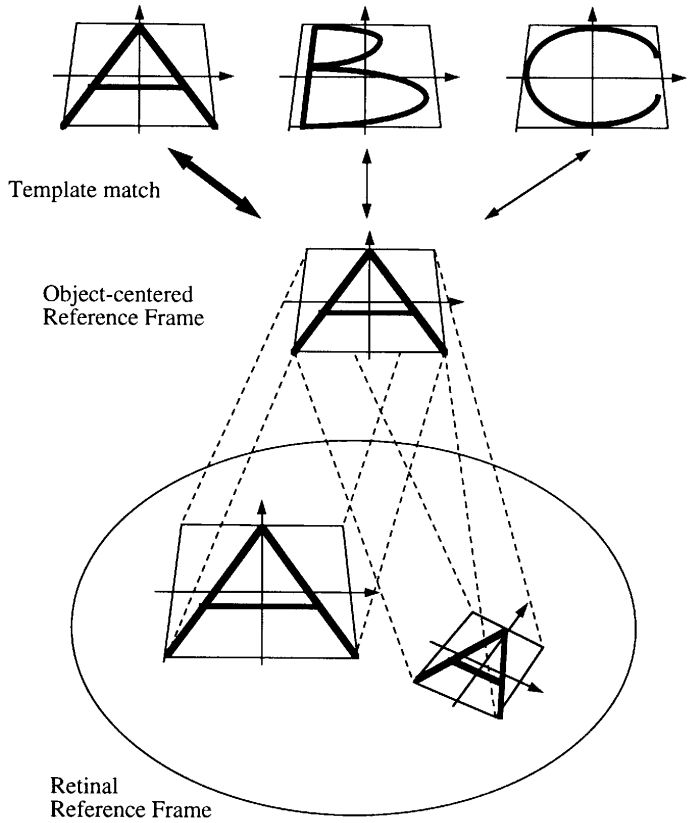
Transform & Match:

First take care of rotation, translation, scale, etc. invariances.

Then recognize based on standardized pixel representation of objects.

e.g., Olshausen et al, 1993,
dynamic routing model

Template match: e.g., with
an associative memory based on
a Hopfield network.



Recognition by Components

Structural approach to object recognition:

Biederman, 1987:

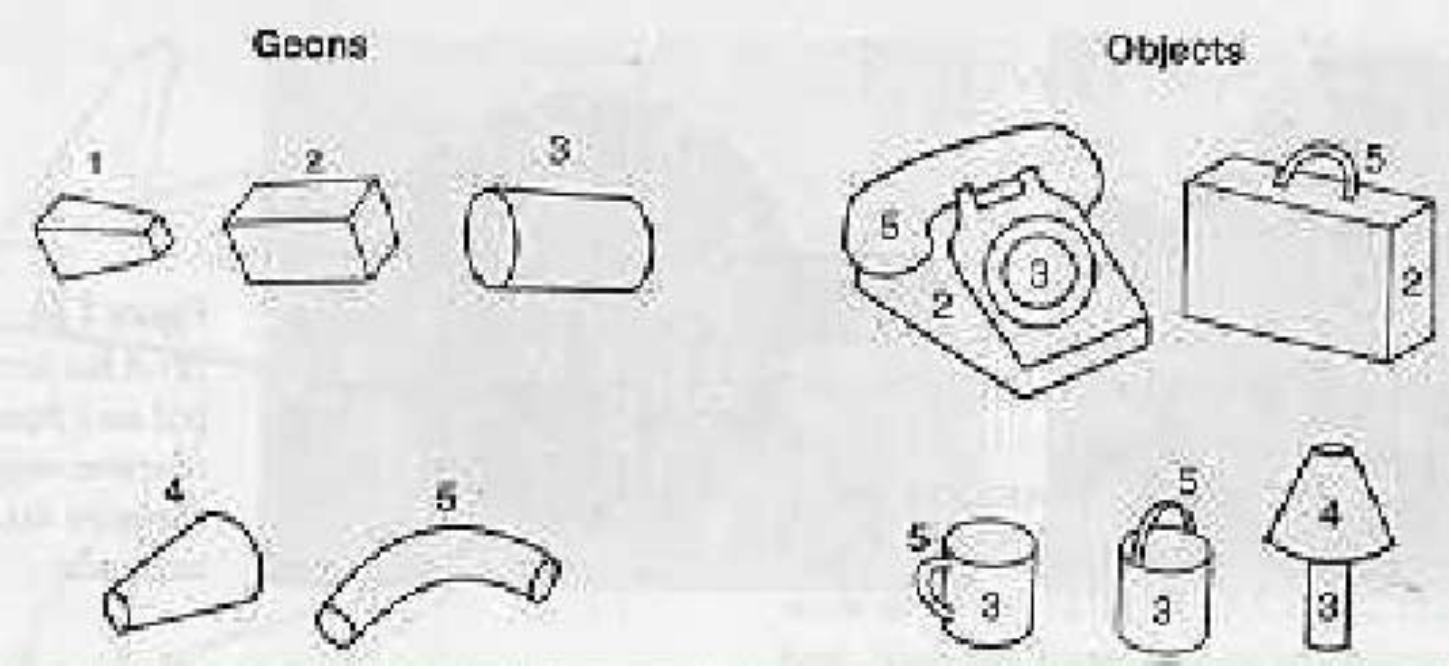
Complex objects are composed so simpler pieces

We can 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.

Recognition by components (Biederman, 1987)

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

Skips 2 ½ D 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).

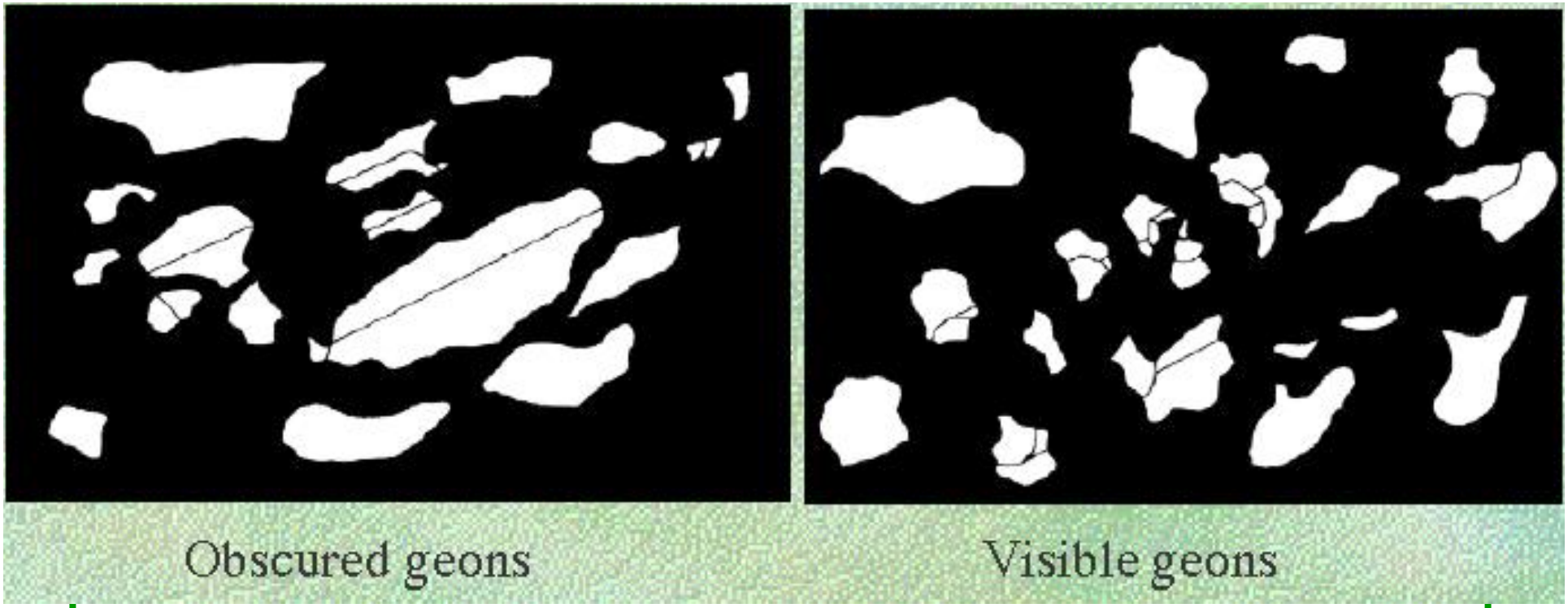


Basic Properties of GEONs

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)



Support for RBC: We can recognize partially occluded objects easily if the occlusions do not obscure the set of geons which constitute the object.

Potential difficulties

A. Structural description not enough, also need metric info

B. Difficult to extract geons from real images

C. Ambiguity in the structural description: most often we have several candidates

D. For some objects, deriving a structural representation can be difficult

Edelman, 1997

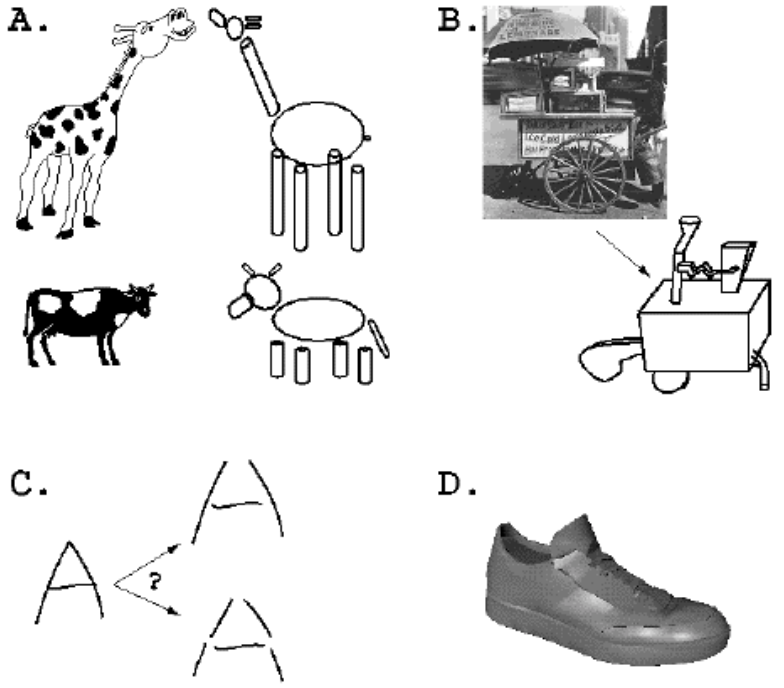
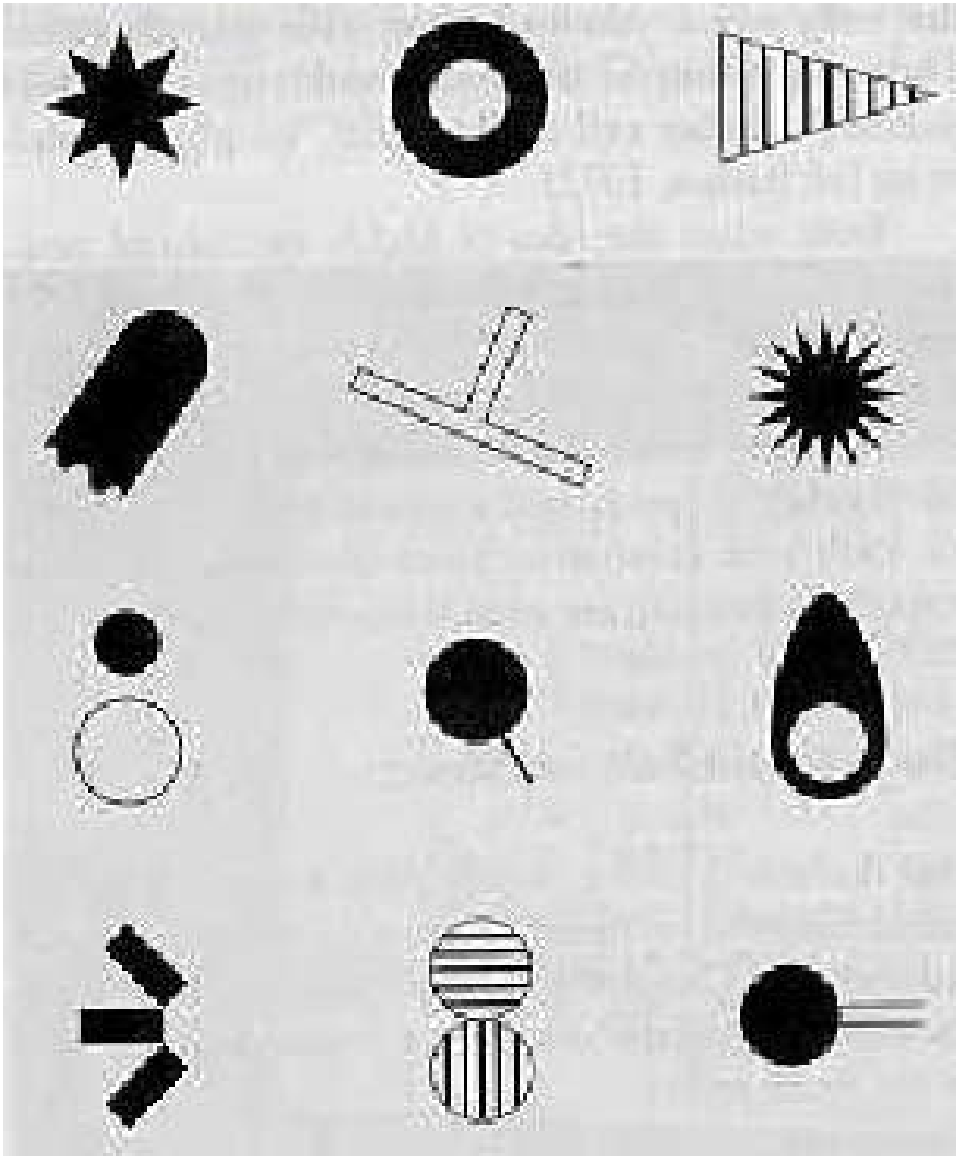


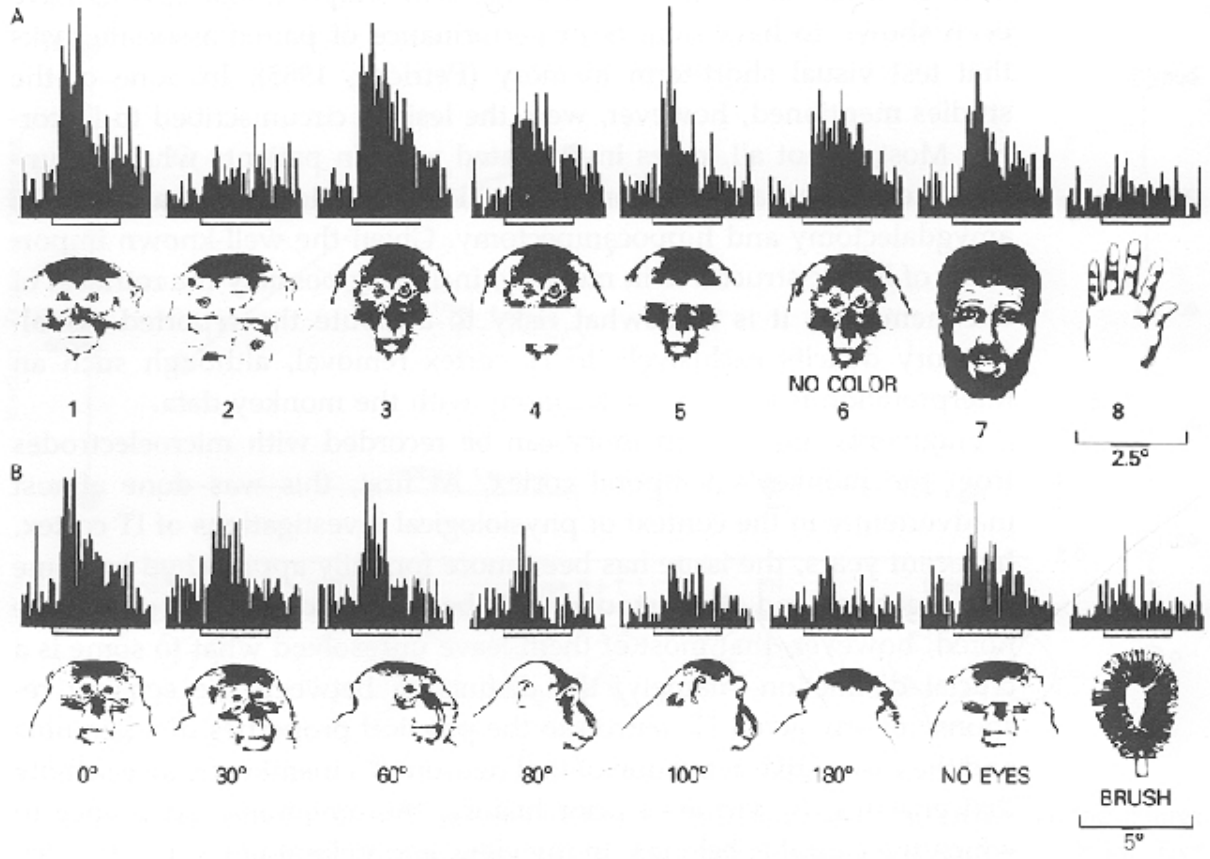
Figure 1: Computational problems with structural representations. A. Structural descriptions must be accompanied by metric information, to represent differences among commonly encountered categories. The inclusion of metric details reduces the ability of structural methods to deal with novel objects. B. A picture of a New York City street-corner hot dog cart, and a stylized object, which, as Biederman [8] suggests, may be described as such following a structural decomposition in the visual system. At present, there is no reliable method for mapping a gray-level image into a collection of (labeled) primitives (lines, corners, etc.) from which RBC's geons are constructed. Thus, although a carefully engineered system such as that described in [22] can form a structural description of the line drawing of a cart-like object, the goal of deriving such a description directly from an image remains elusive. C. Even in simpler tasks (e.g., in character recognition, where the figure is readily separable from the ground), the derivation of a structural description is problematic. The difficulty here stems from the possibility to assign multiple structural descriptions to the same image. D. In some tasks, coming up even with one structural description is problematic; how does one represent a shoe in terms of RBC's geons [7]?

Geon Neurons in IT?

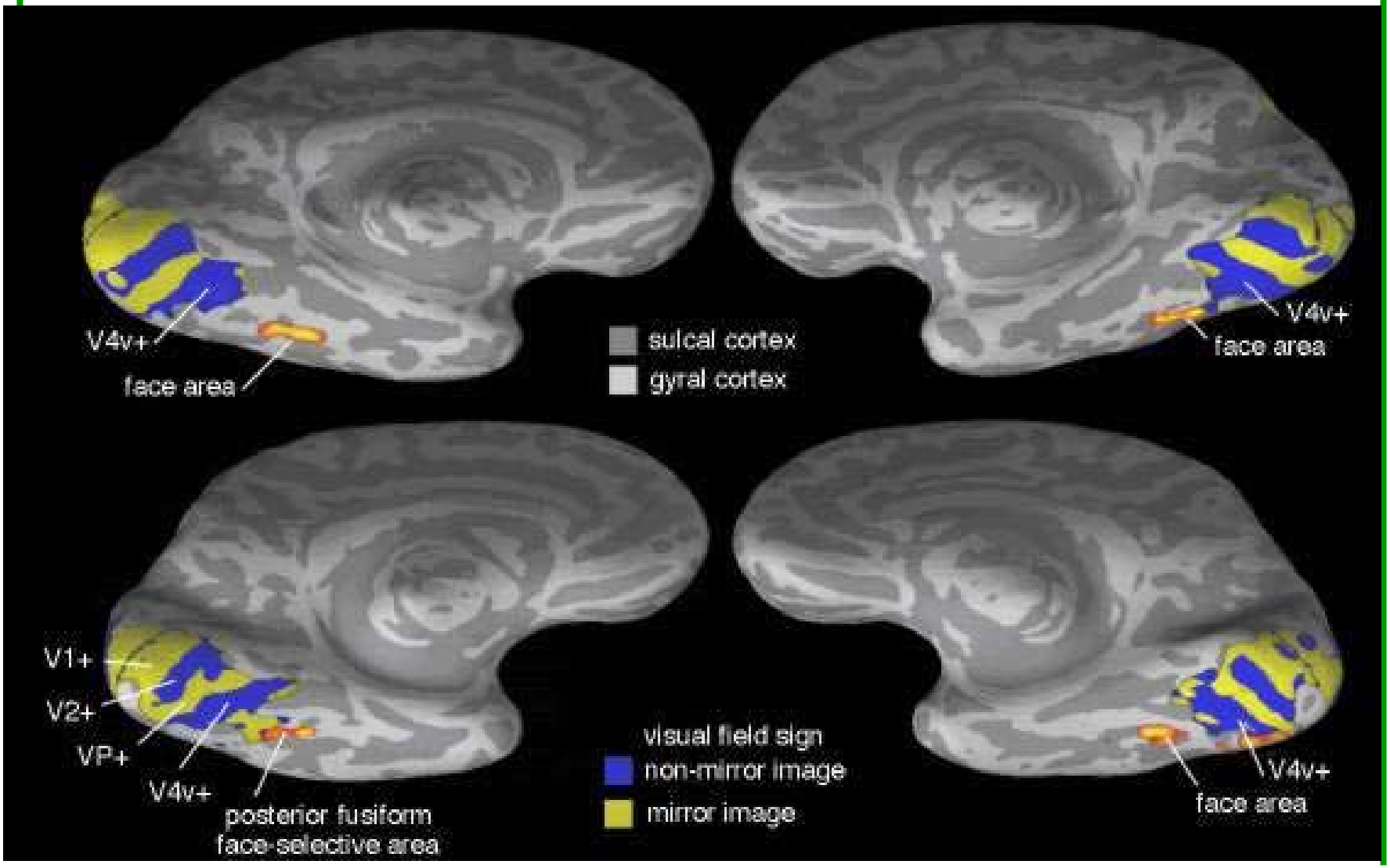
These are preferred stimuli for some IT neurons.



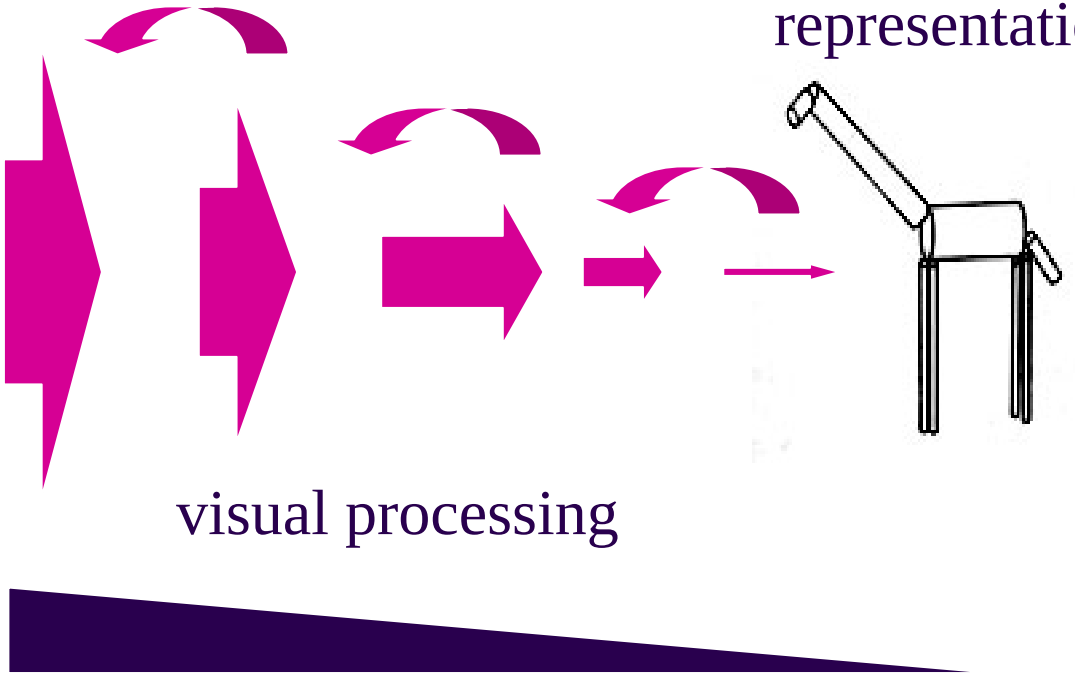
“Face cells”



Fusiform Face Area in Humans



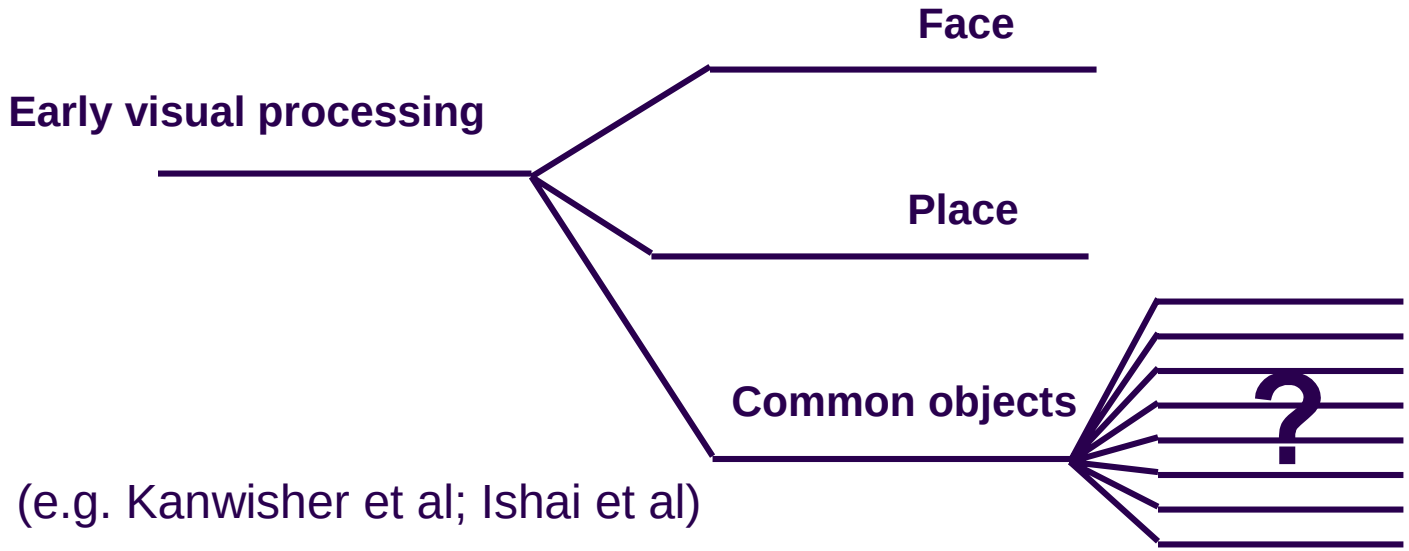
Standard View on Visual Processing



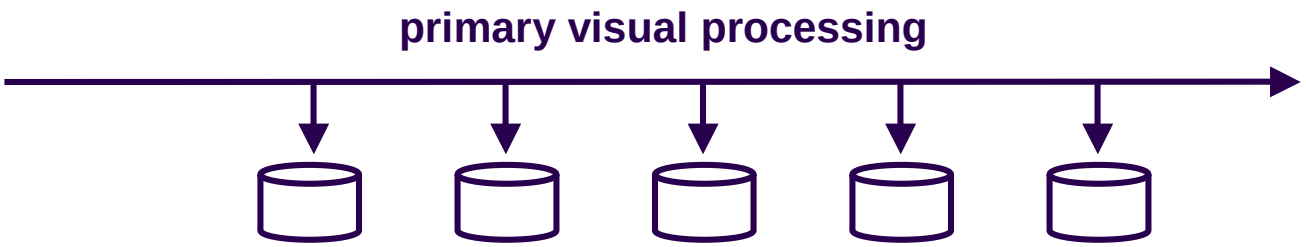
- **Image specific**
- **Supports fine discrimination**
- **Noise tolerant**

- **Image invariant**
- **Supports generalization**
- **Noise sensitive**

Tjan, 1999



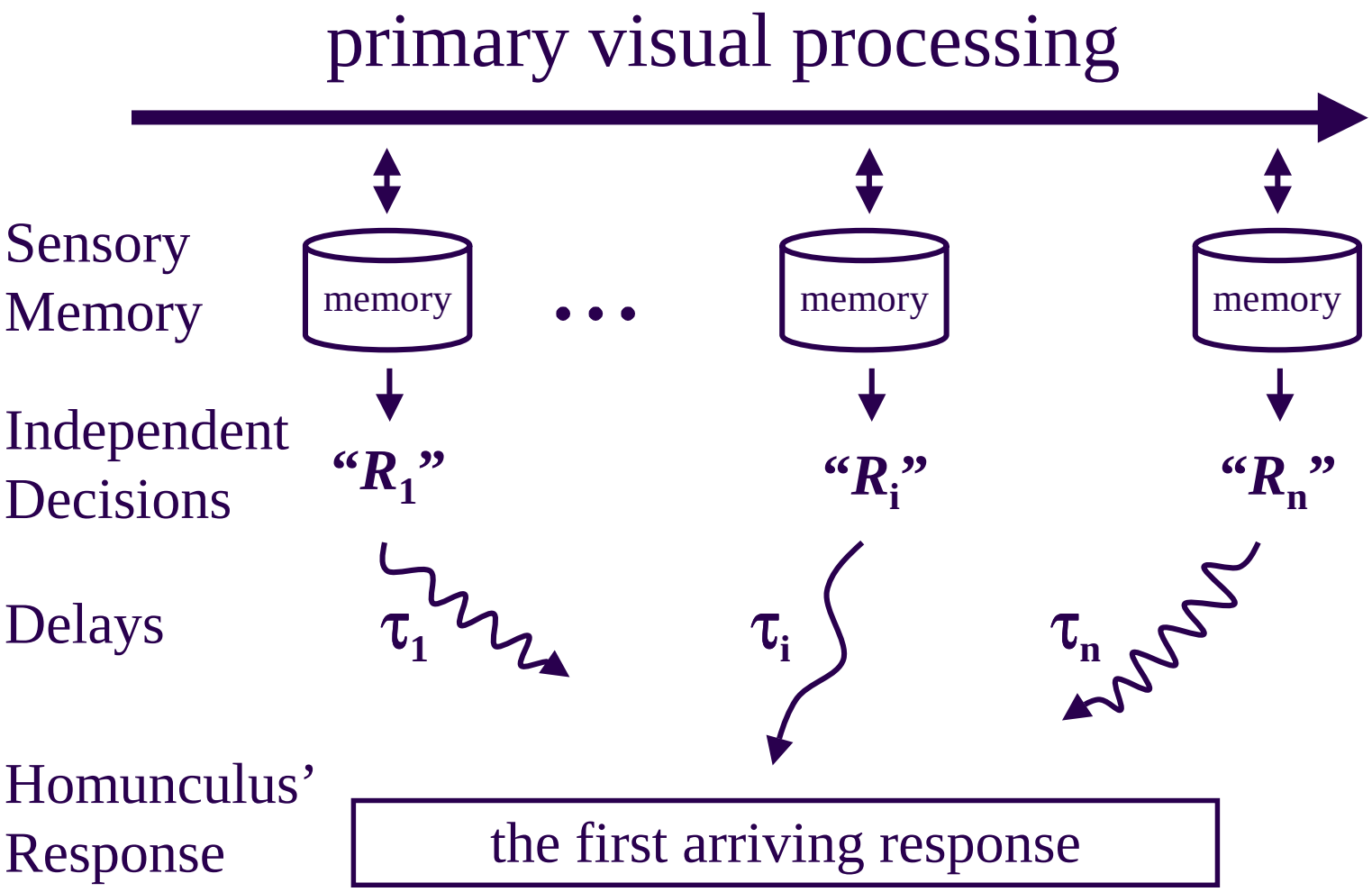
(e.g. Kanwisher et al; Ishai et al)



(Tjan, 1999)

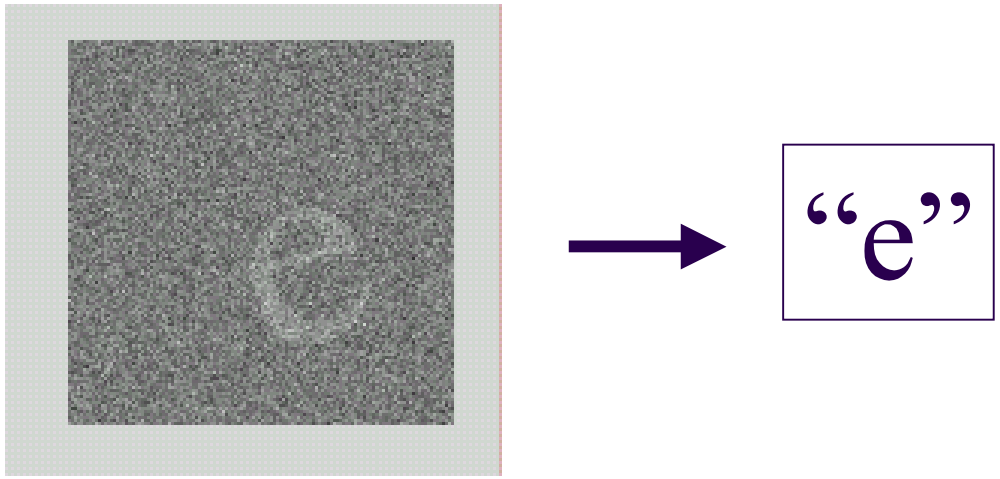
Multiple memory/decision sites

Tjan's "Recognition by Anarchy"



A toy visual system

Task: ***Identify letters from arbitrary positions & orientations***



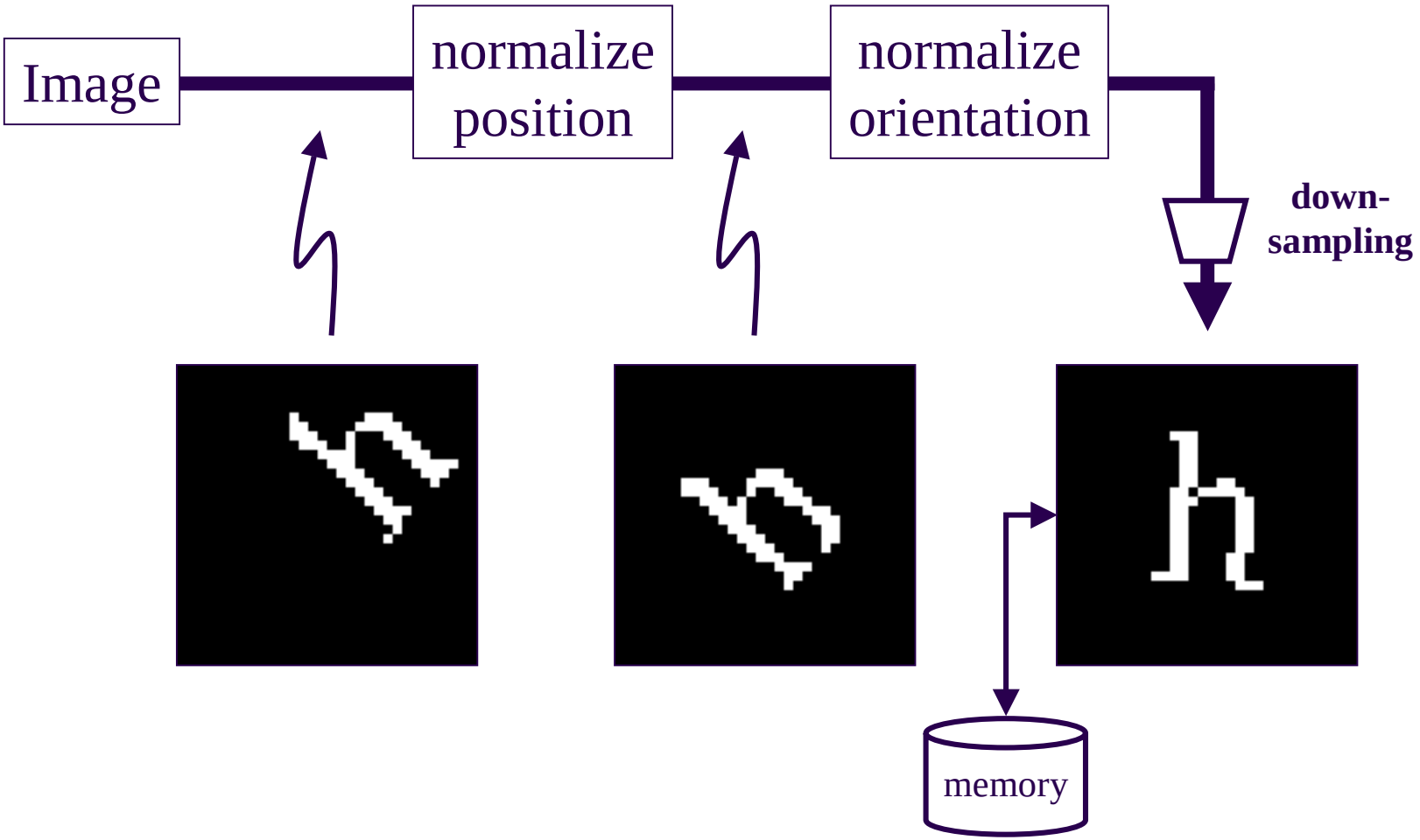


Image
normalize position
normalize orientation
down-sampling
memory

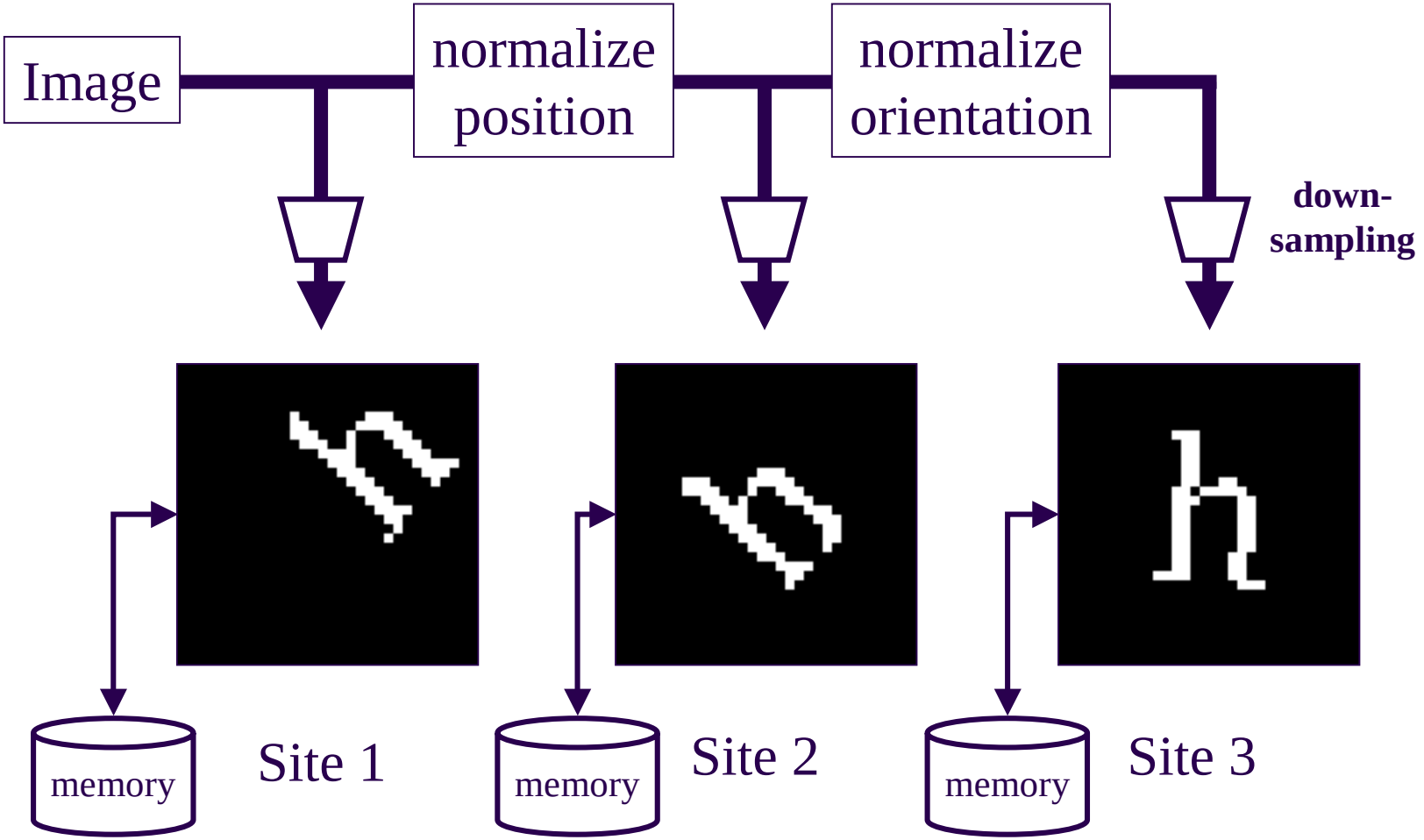


Image
normalize position
normalize orientation
down-sampling
memory
Site 1
memory
Site 2
memory
Site 3

Study stimuli:

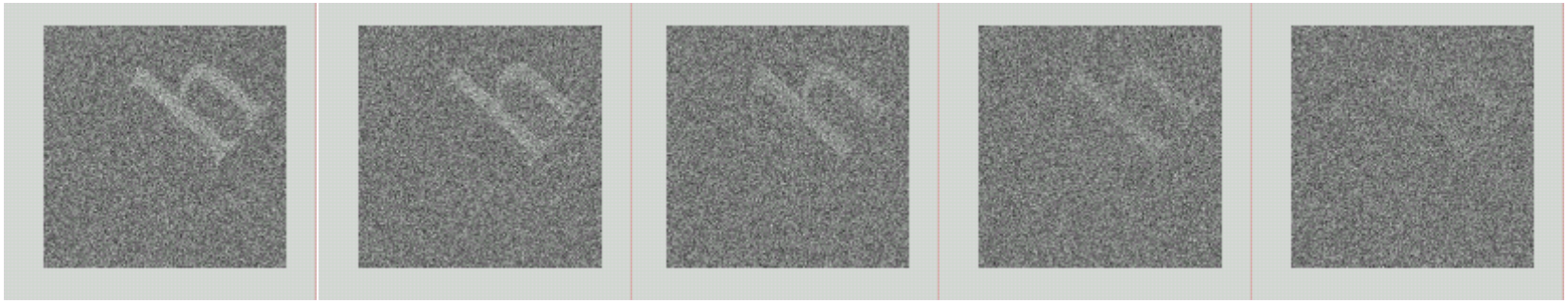
5 orientations ×20 positions at high SNR

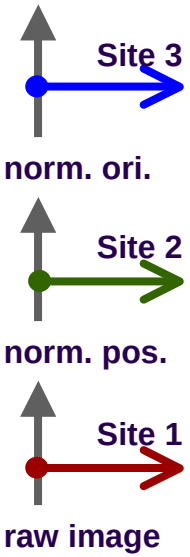
Test stimuli:

1) familiar (studied) views,

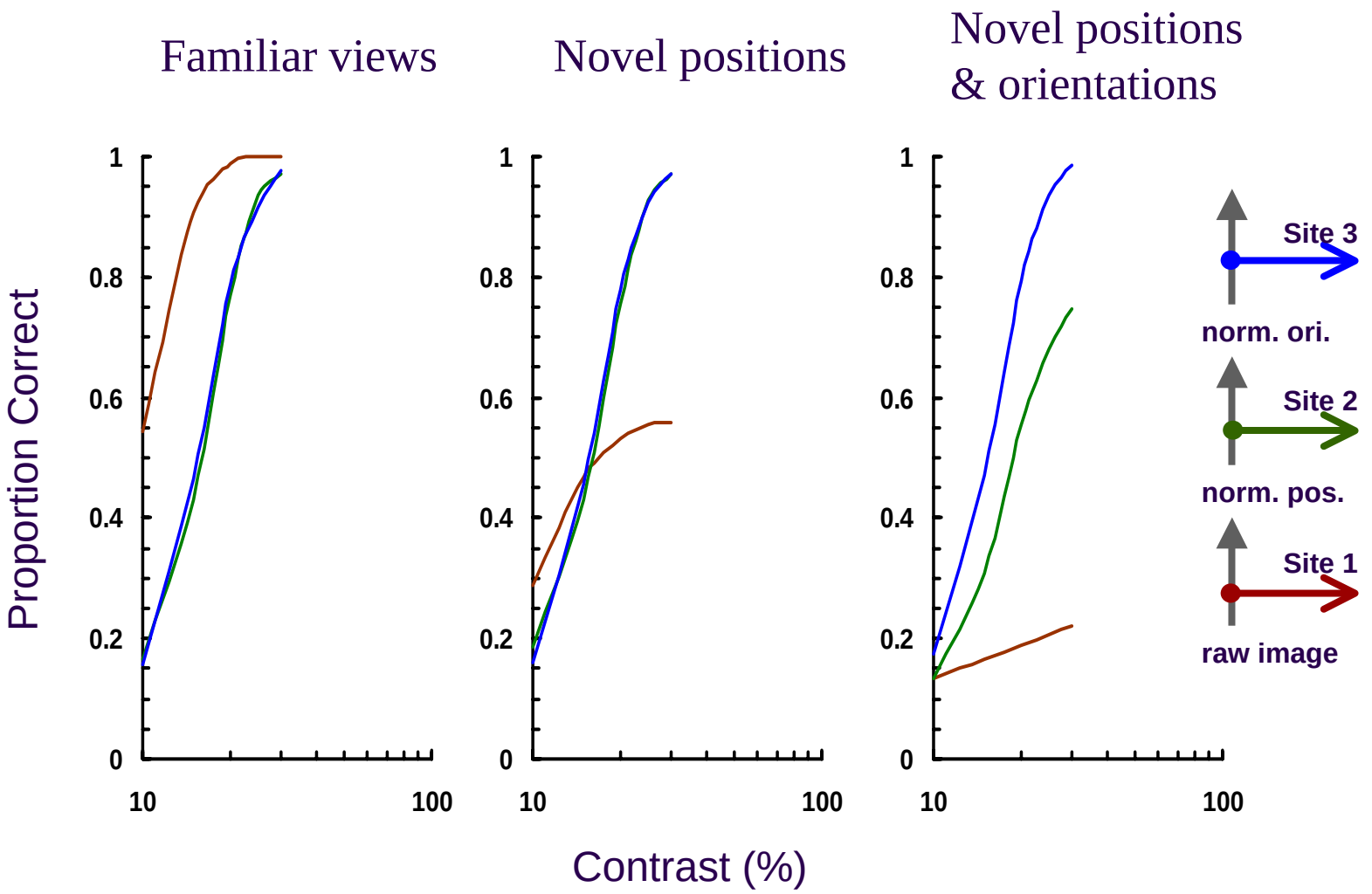
2) new positions,

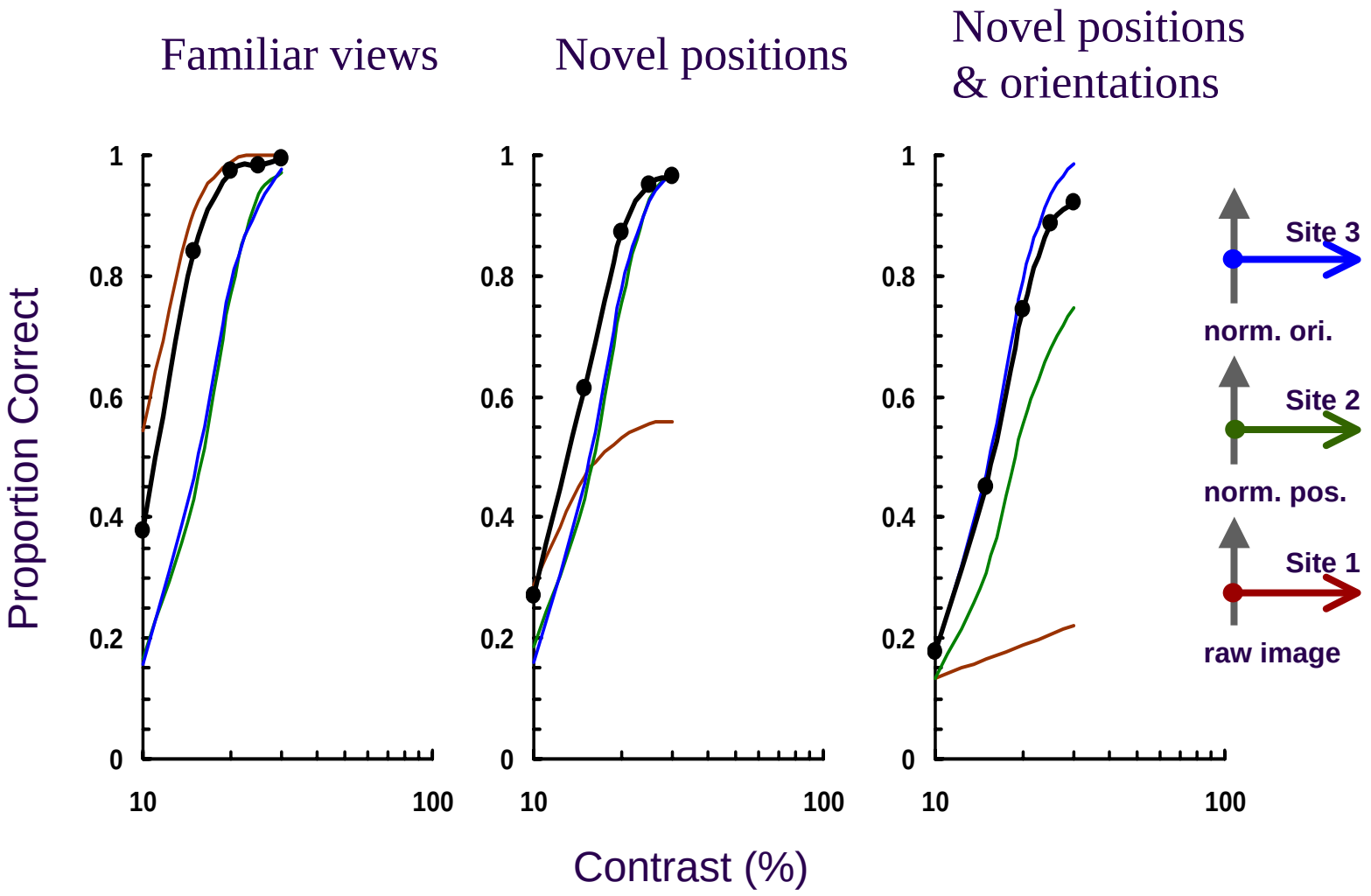
3) new position & orientations





Processing speed for each recognition module depends on recognition difficulty by that module.


Familiar views
Novel positions
Novel positions & orientations
Proportion Correct
1
0.8
0.6
0.4
0.2
0
10
100
Contrast (%)
Site 3
norm. ori.
Site 2
norm. pos.
Site 1
raw image



Black curve: full model in which recognition is based on the fastest of the responses from the three stages.

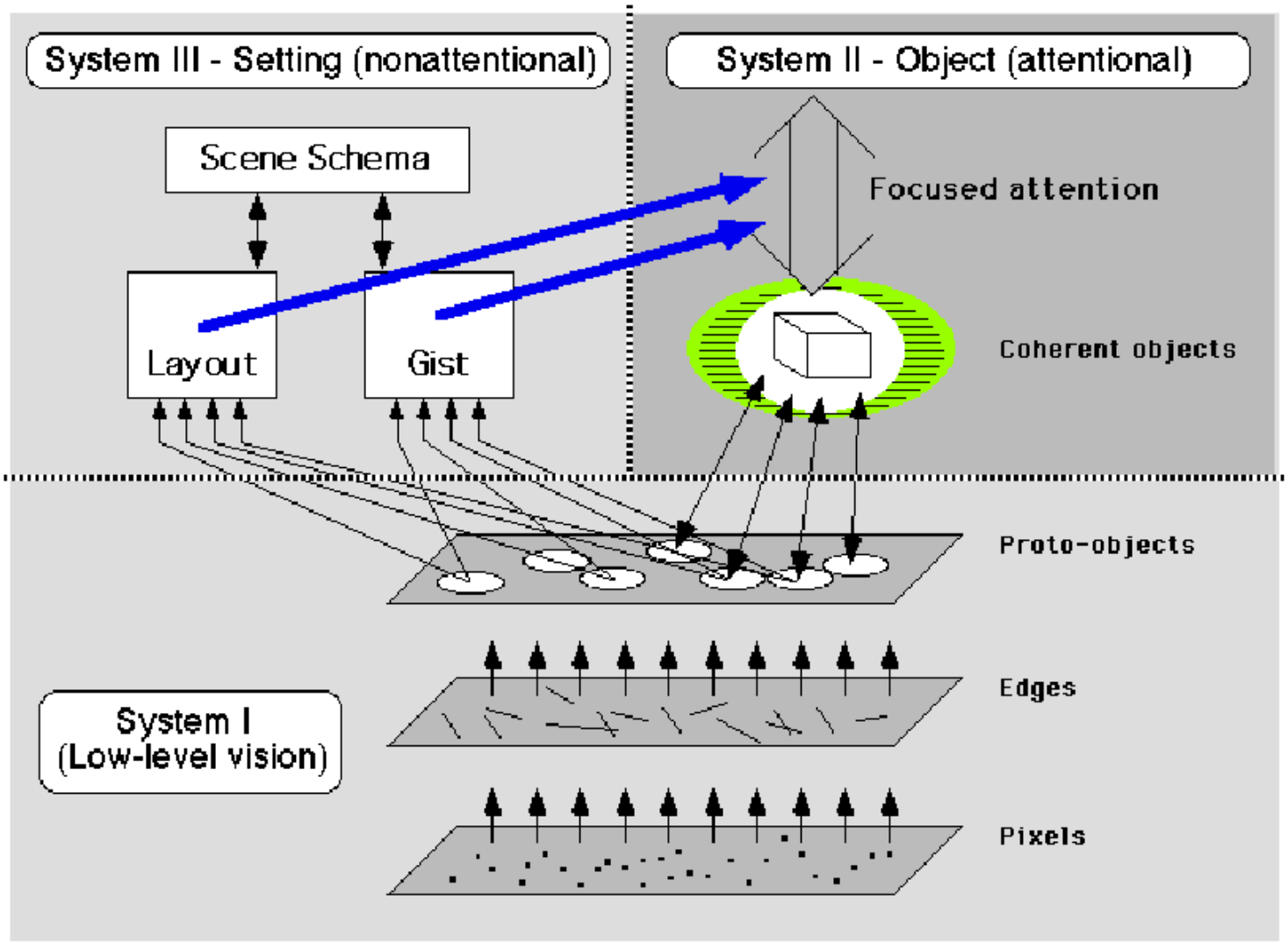


Figure 4

Figure 4. Triadic Architecture. It is suggested that the visual perception of scenes may be carried out via the interaction of three different systems. System I: Early-level processes produce volatile proto-objects rapidly and in parallel across the visual field. System II: Focused attention acts as a hand to "grab" these structures; as long as these structures are held, they form an individuated object with both temporal and spatial coherence. System III: Setting information—obtained via a nonattentional stream—guides the allocation of focused attention to various parts of the scene, and allows priorities to be given to the various possible objects.