
Tourists.mov
File Edit Movie QTV Window Help
00:00:00
TV

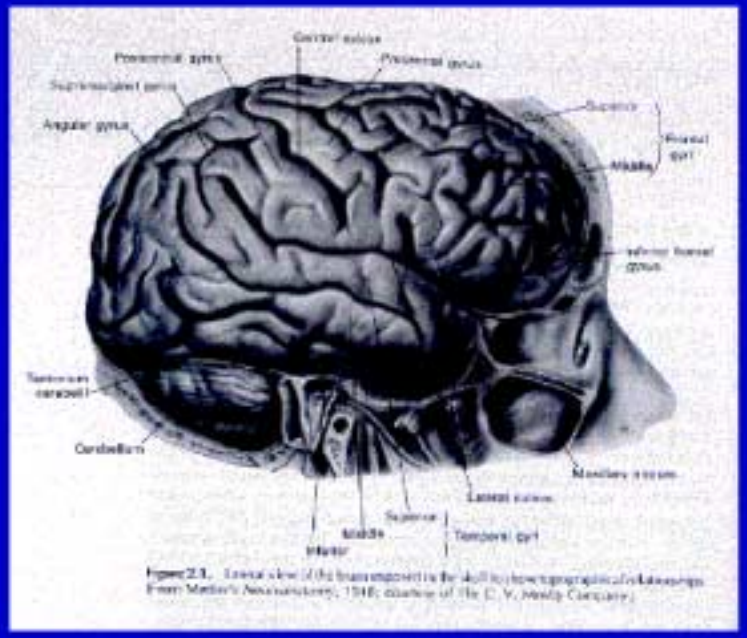

Tourists.mov
File Edit Movie QTV Window Help
00:00:00
TV

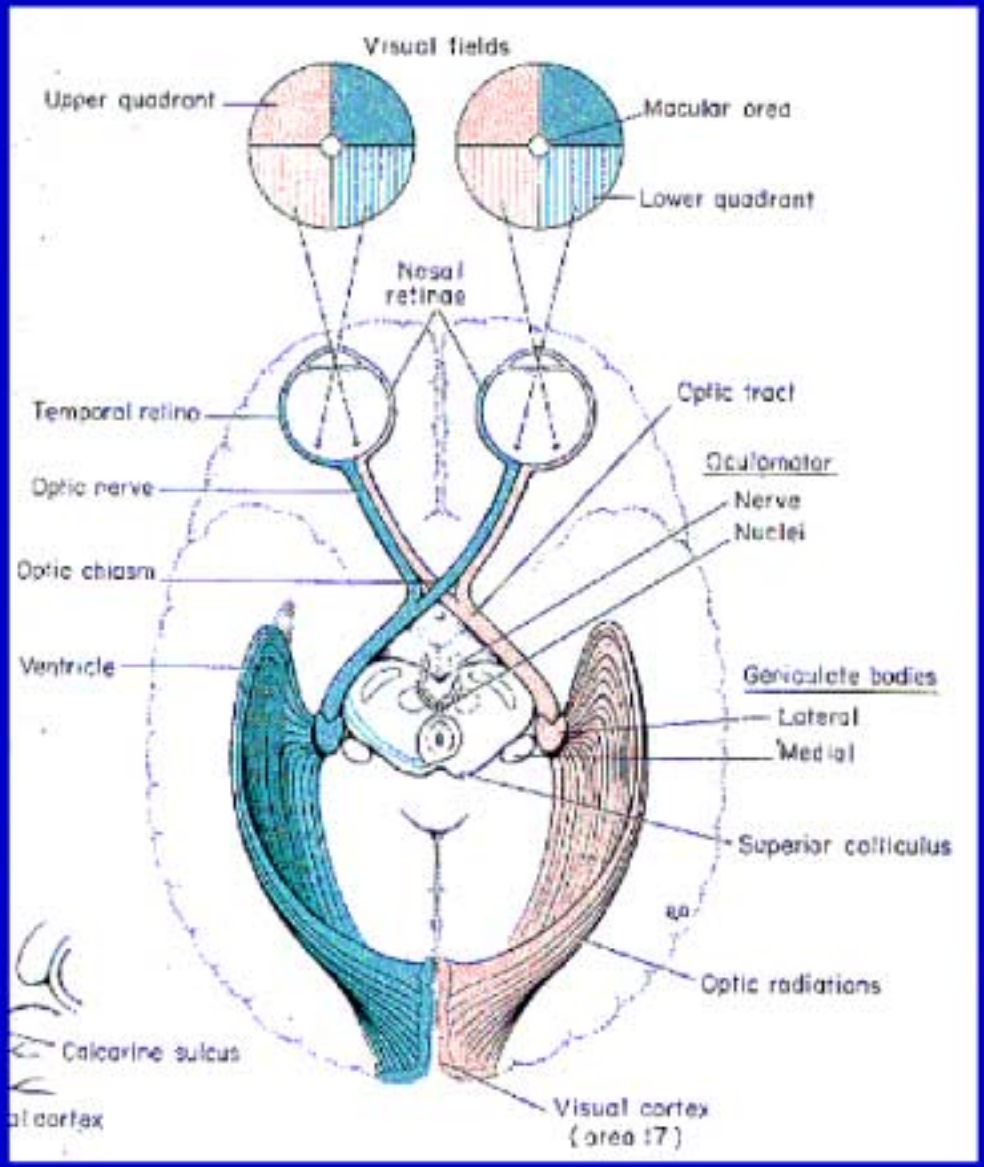

Tourists.mov
File Edit Movie QTV Window Help
00:00:00
TV

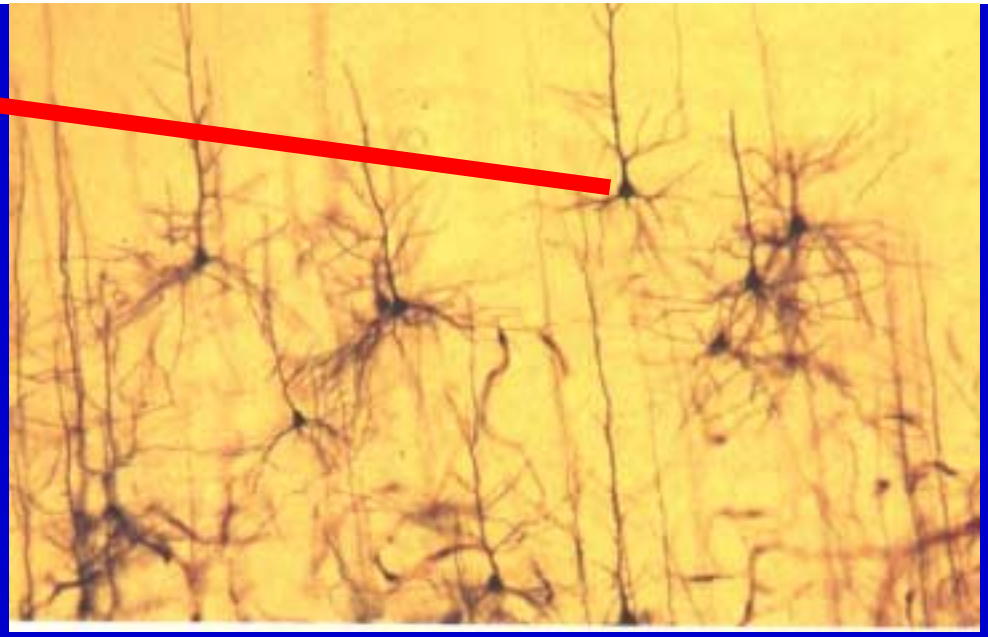
Why model visual attention?

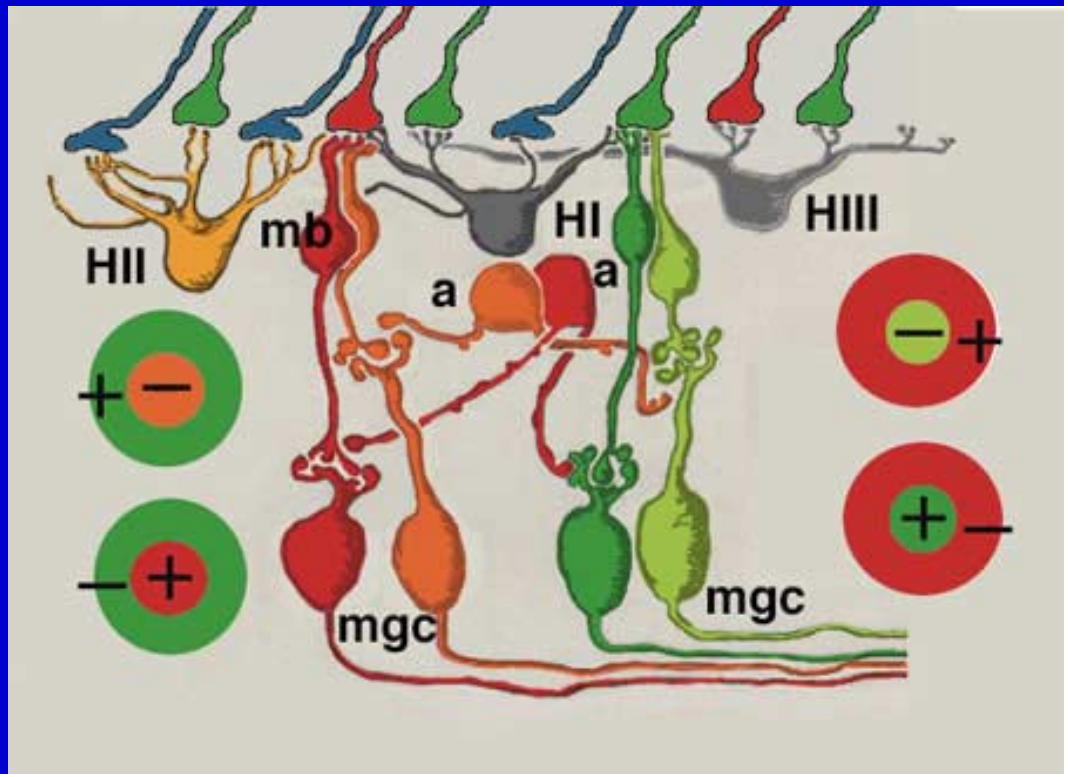
- To computationally understand how the brain works
- To interpret psychophysical experiments
- To guide object recognition systems
- To build robust and adaptive active vision systems

Human Visual System



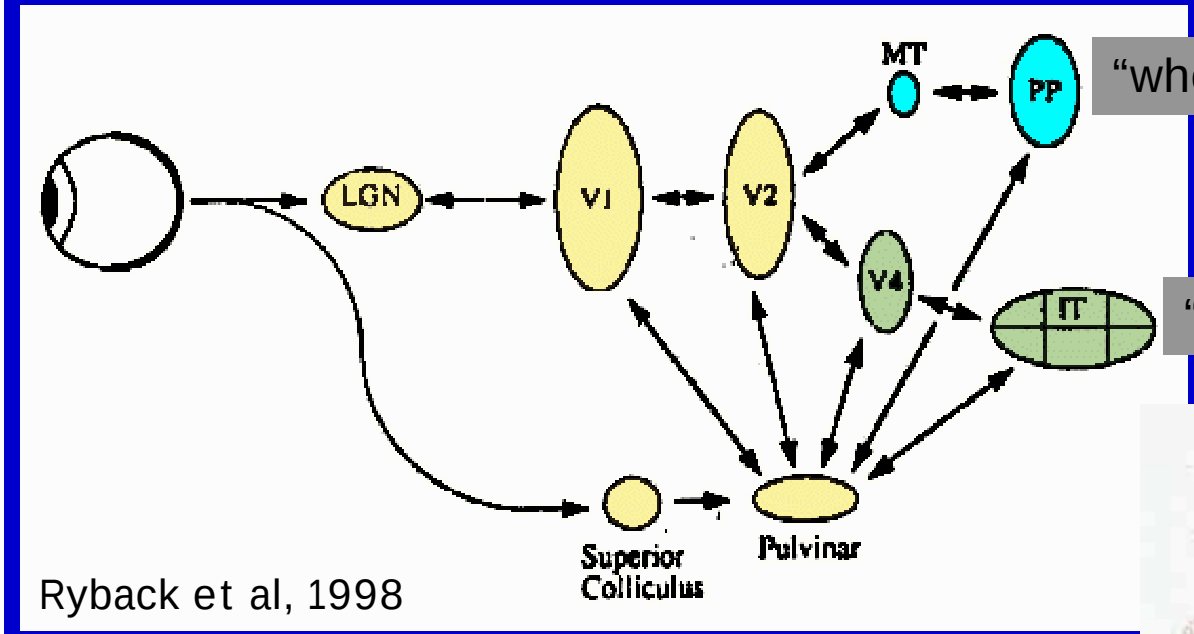




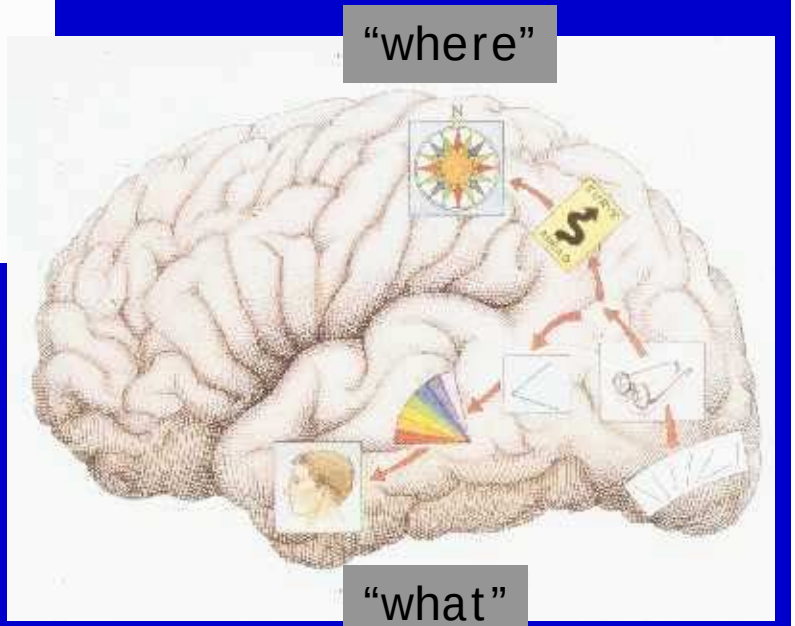

HII
mb
HI
a
a
HIII
mgc
mgc

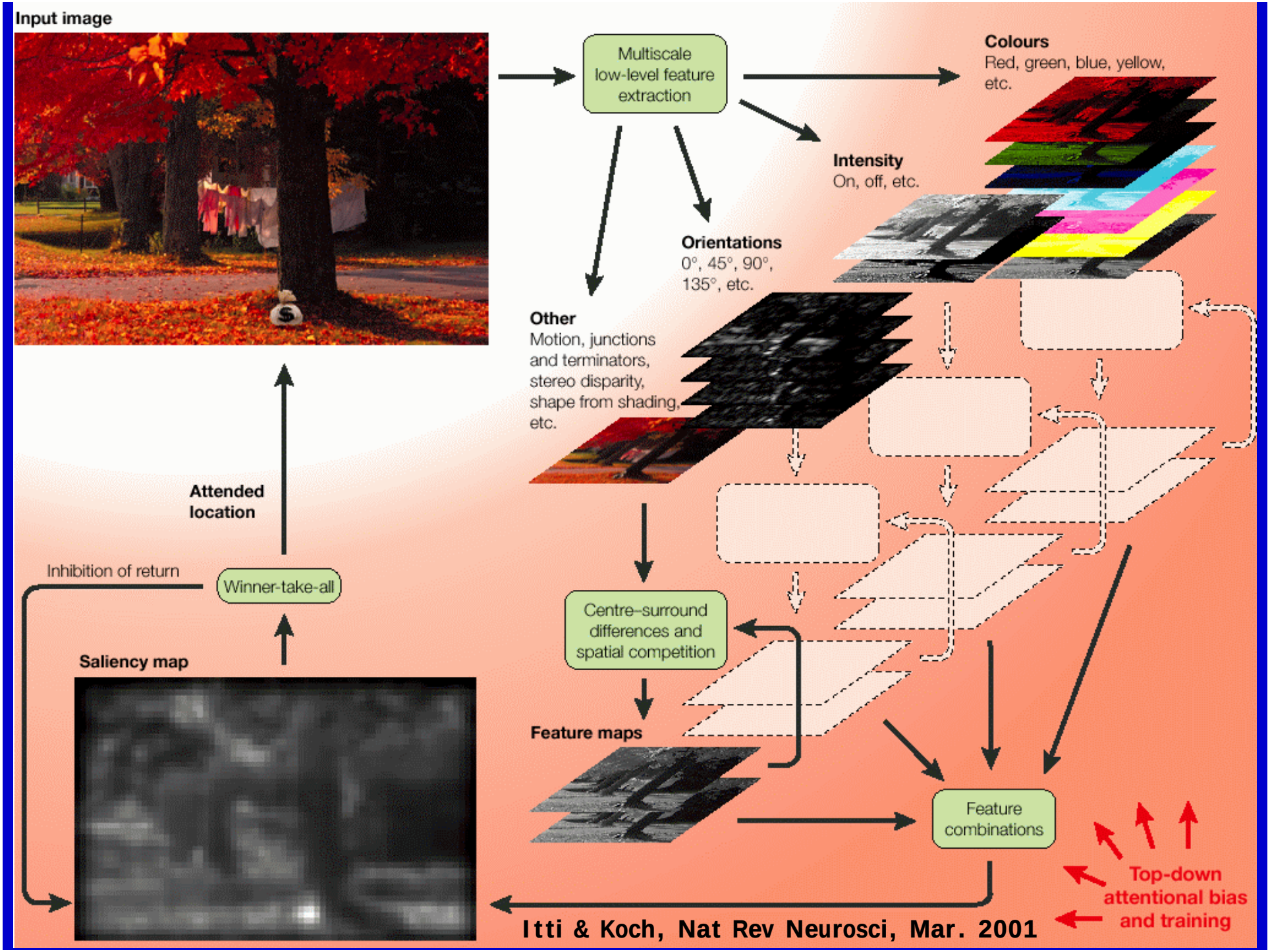
“Where” and “What” Visual Pathways

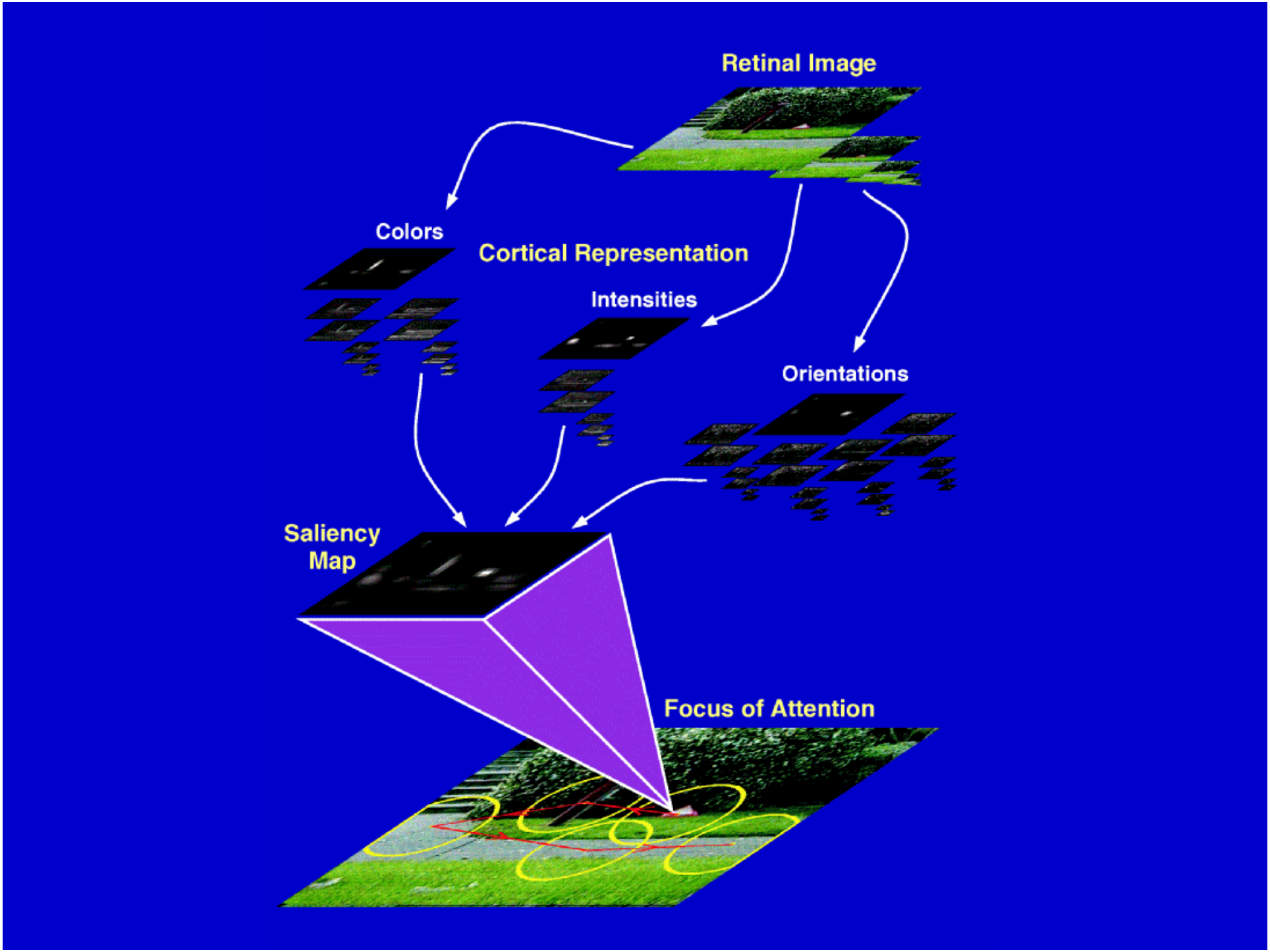
Dorsal stream (to posterior parietal): object localization
Ventral stream (to infero-temporal): object identification

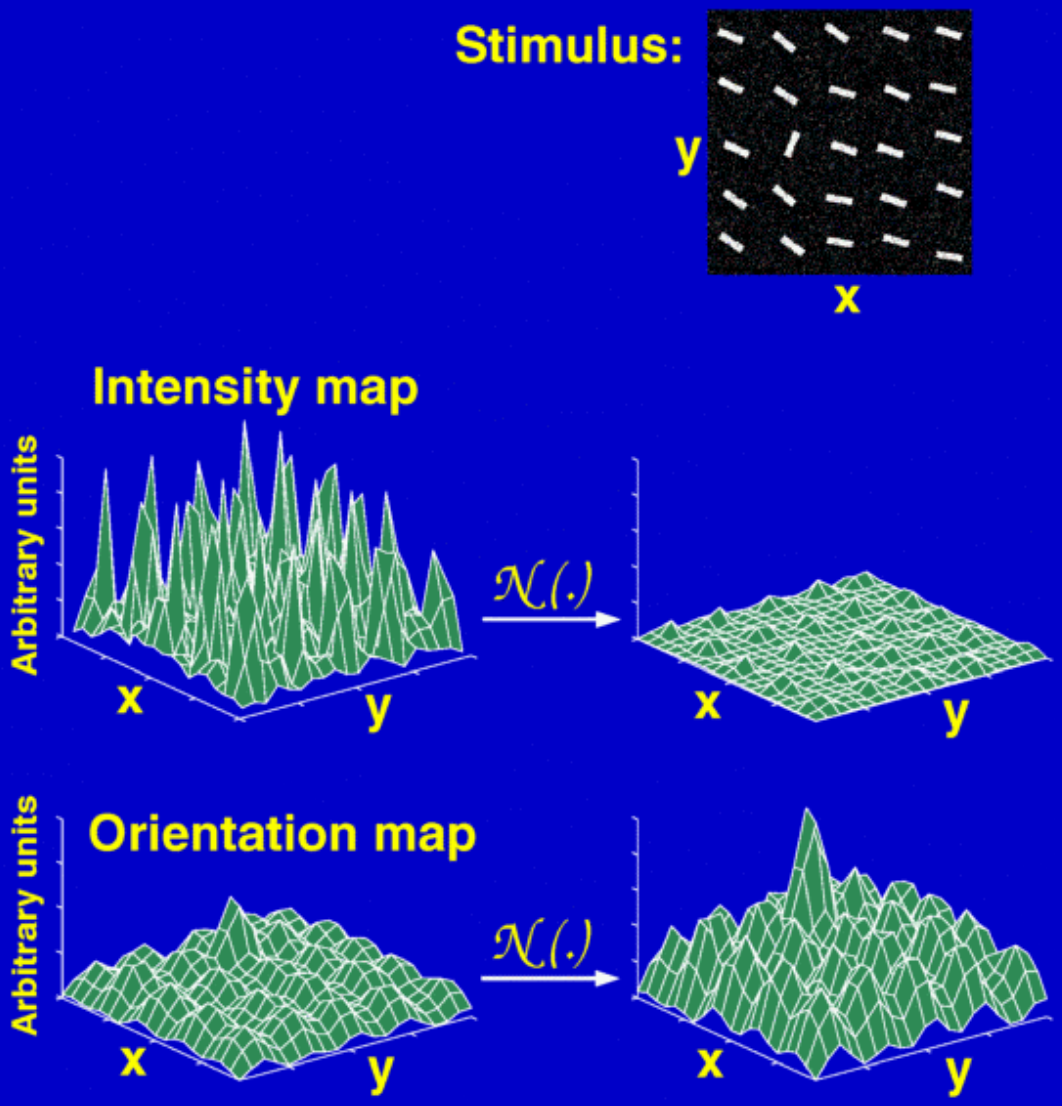


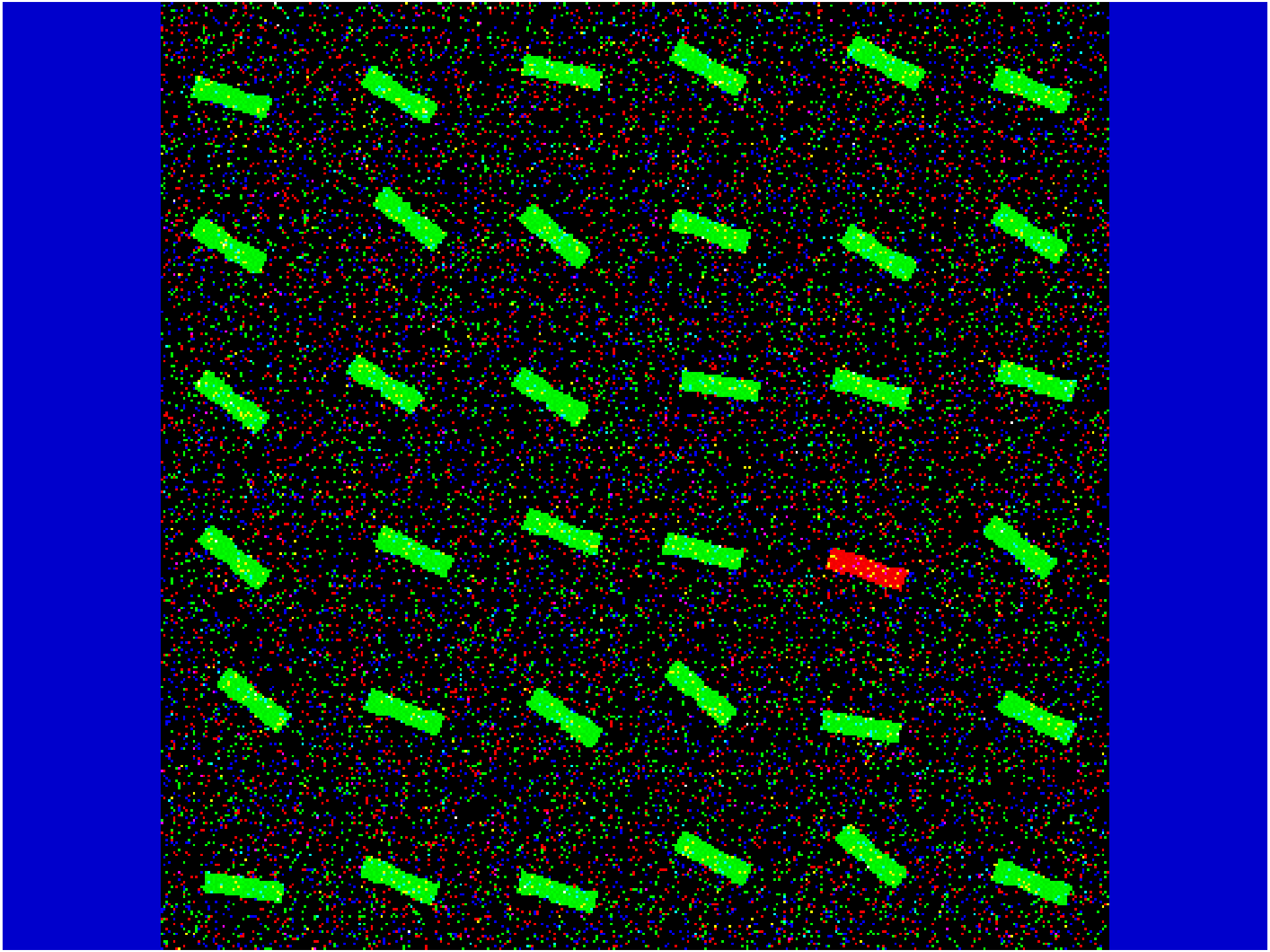
Ryback et al, 1998

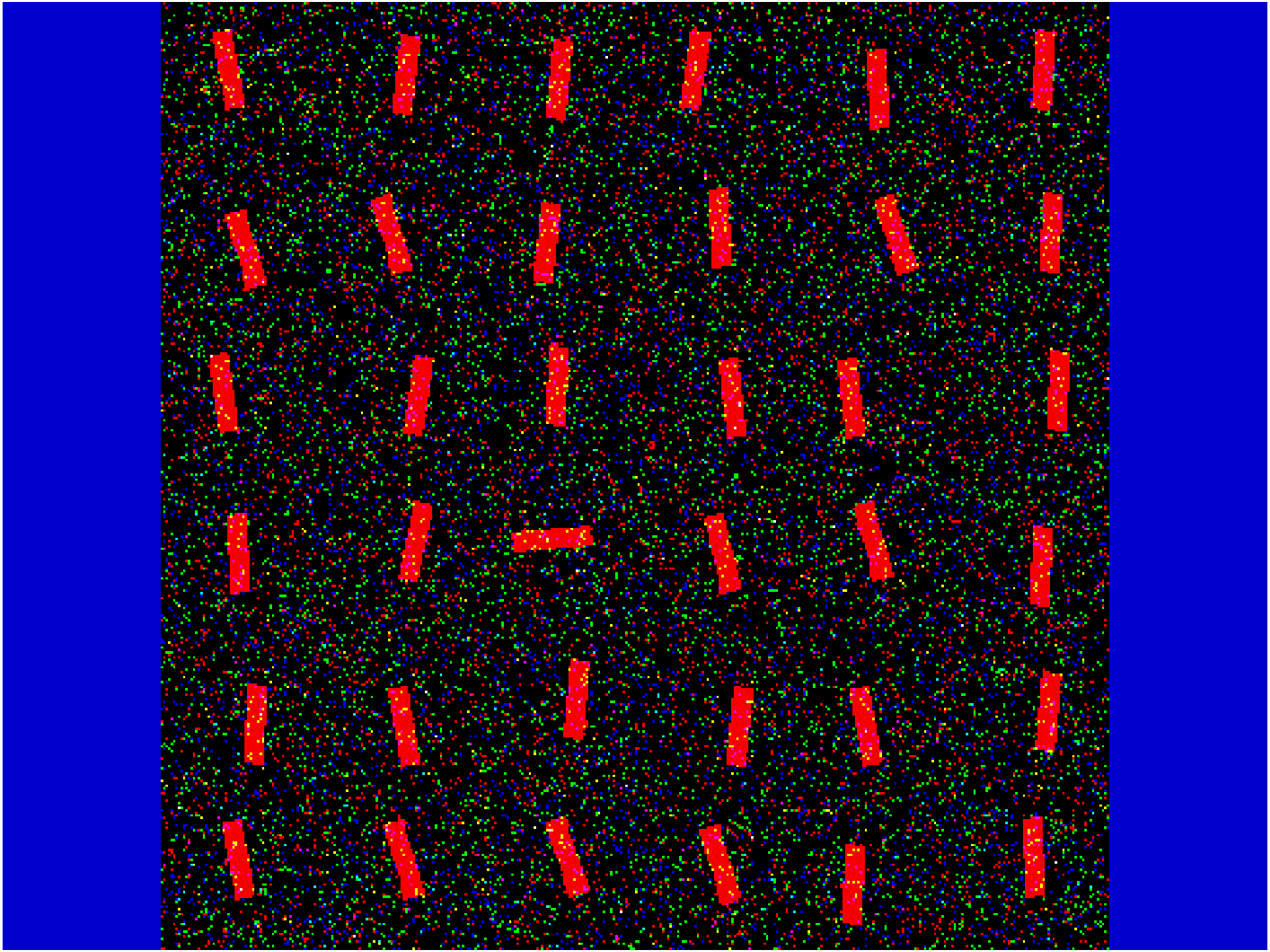


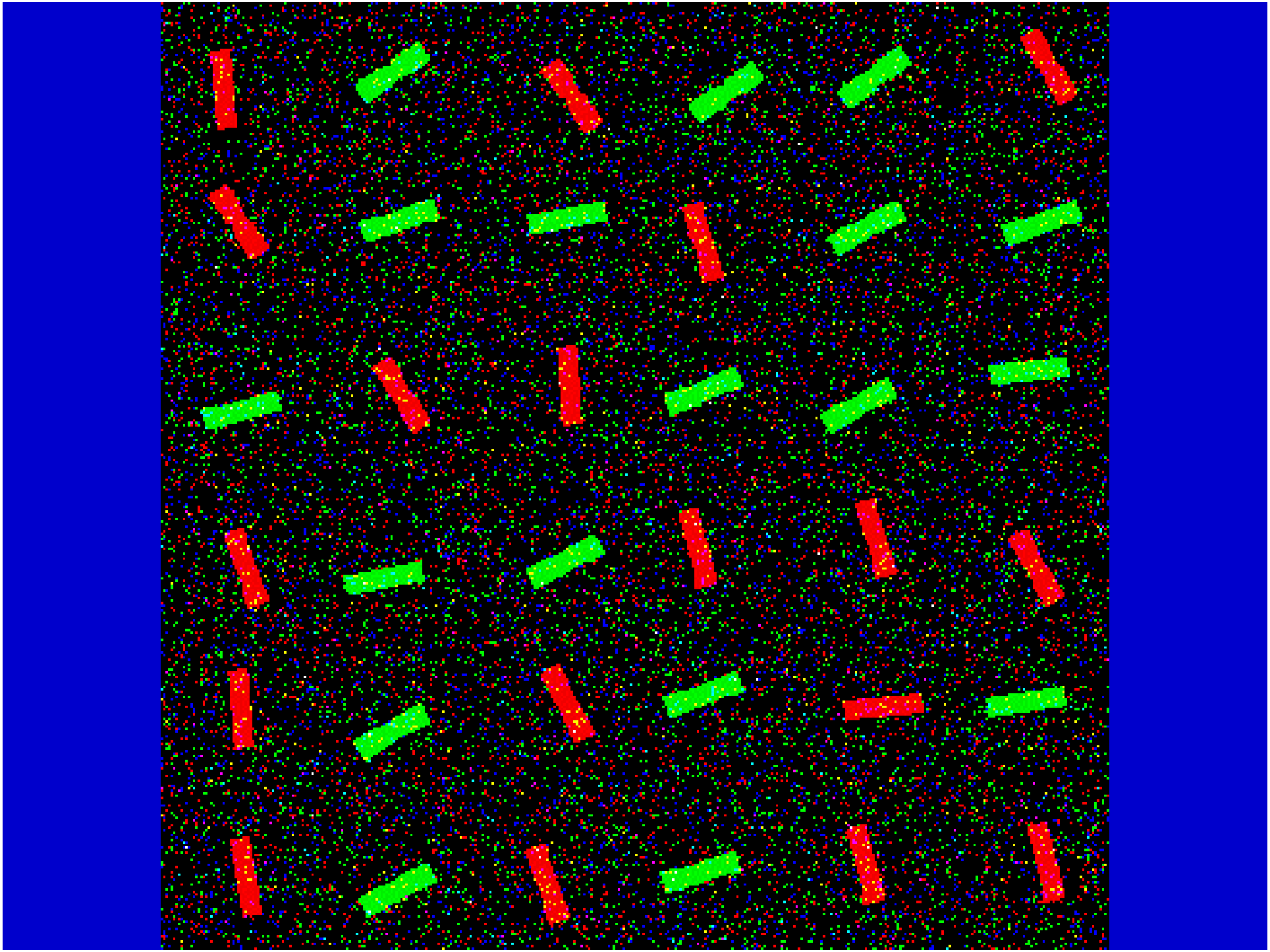

Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training
Itti & Koch, Nat Rev Neurosci, Mar. 2001

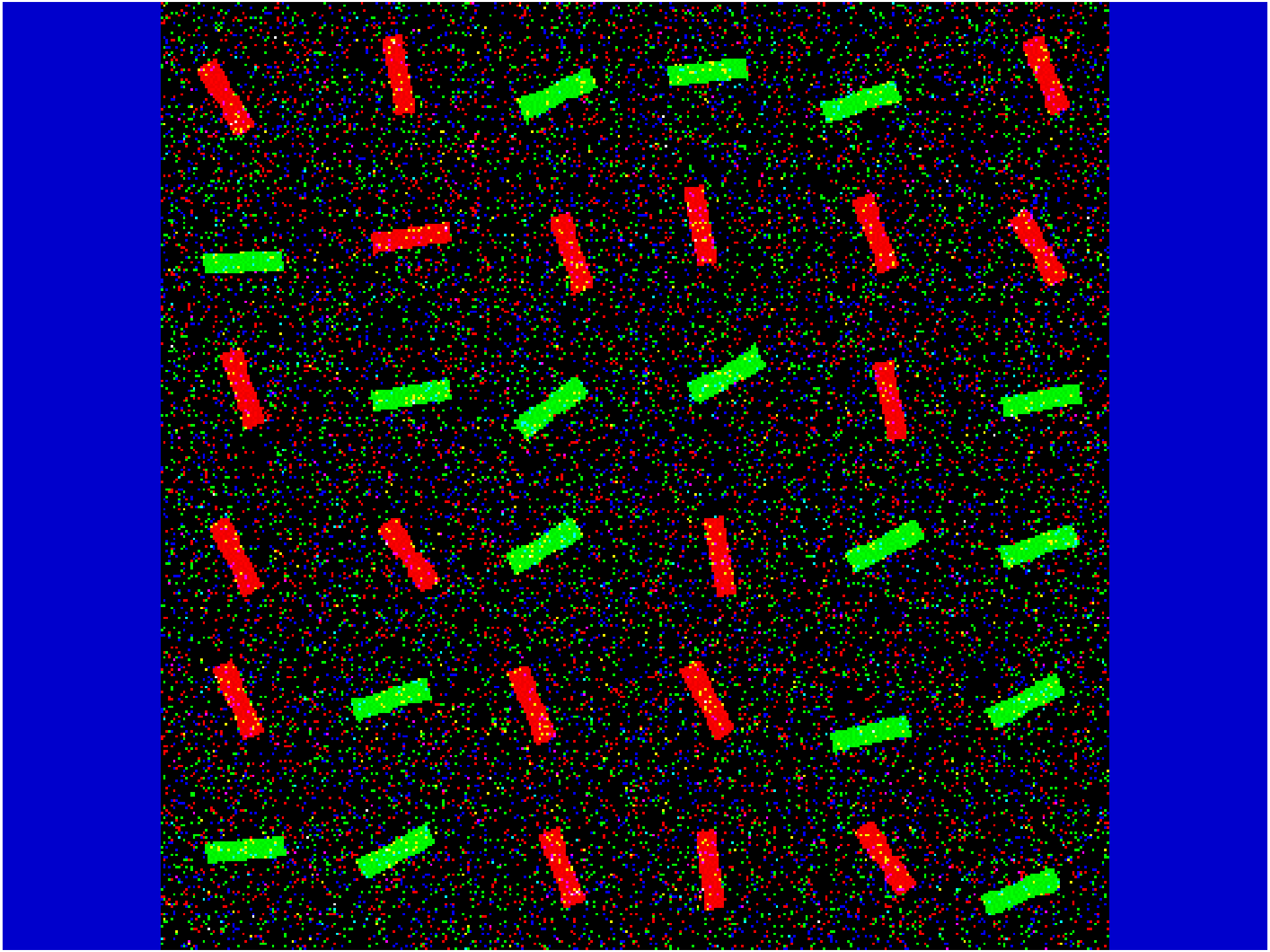

Retinal Image
Colors
Cortical Representation
Intensities
Orientations
Saliency Map
Focus of Attention


Stimulus:
y
x
Intensity map
Arbitrary units
x
y
$\mathcal{N}(.)$
x
y
Orientation map
Arbitrary units
x
y
$\mathcal{N}(.)$
x
y

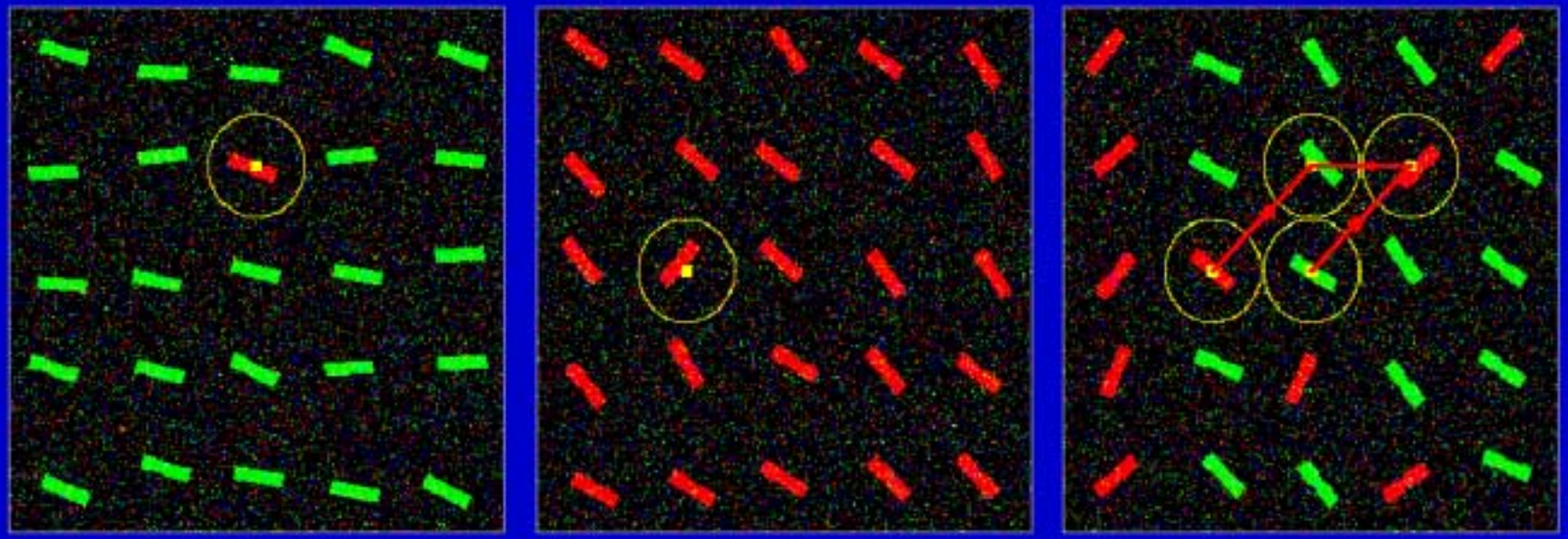




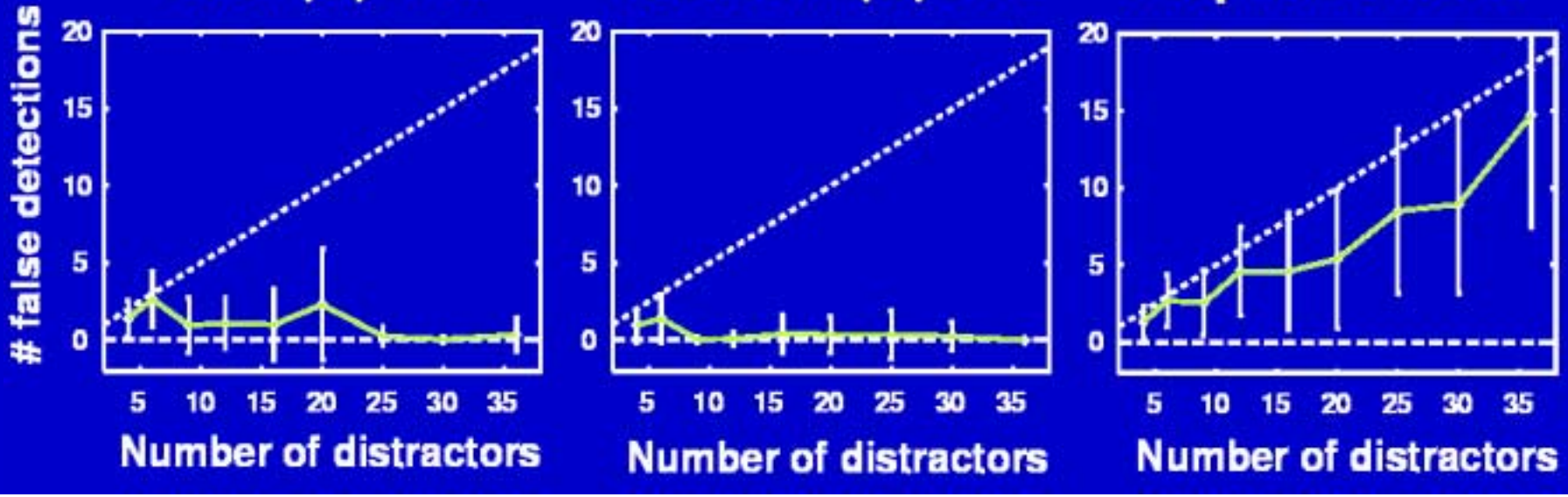




Simulated Psychophysics



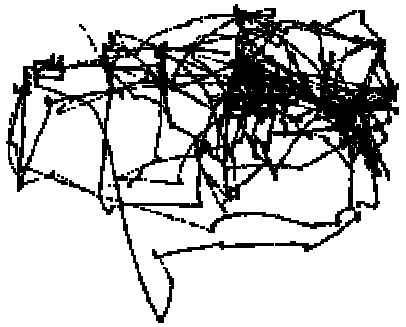
Color pop-out | Orientation pop-out | Conjunctive search

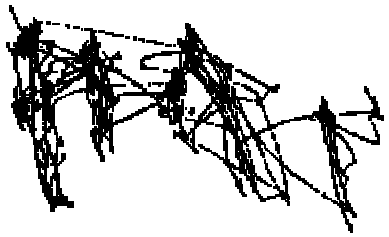


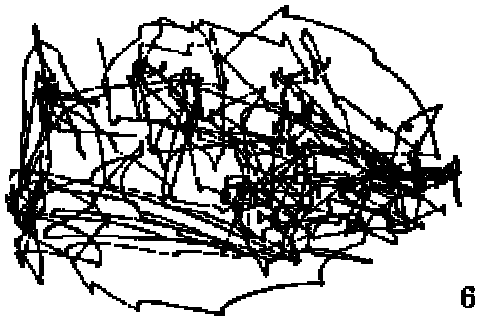
Eye Movements

1) Free examination
2) estimate material circumstances of family
3) give ages of the people
4) surmise what family has been doing before arrival of “unexpected visitor”
5) remember clothes worn by the people
6) remember position of people and objects
7) estimate how long the “unexpected visitor” has been away from family

Yarbus, 1967







Rensink, 1999

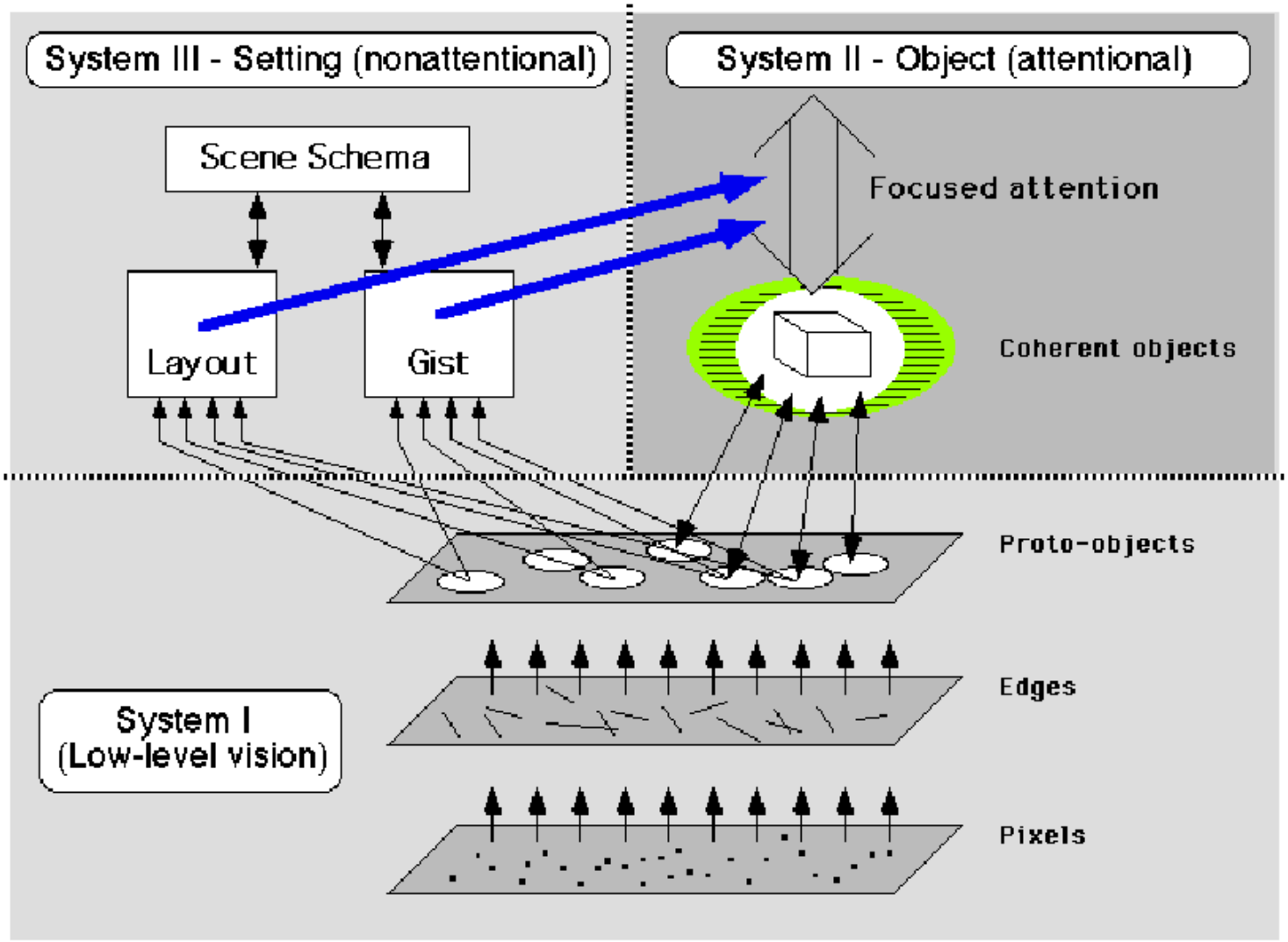
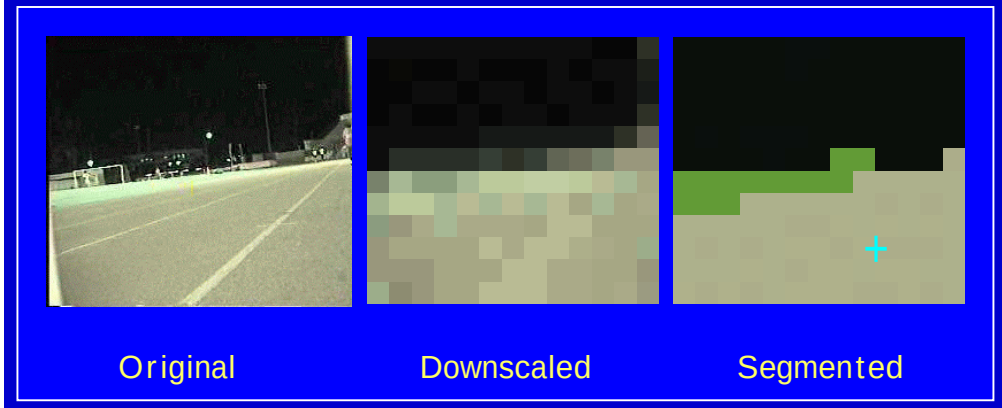


Figure 4

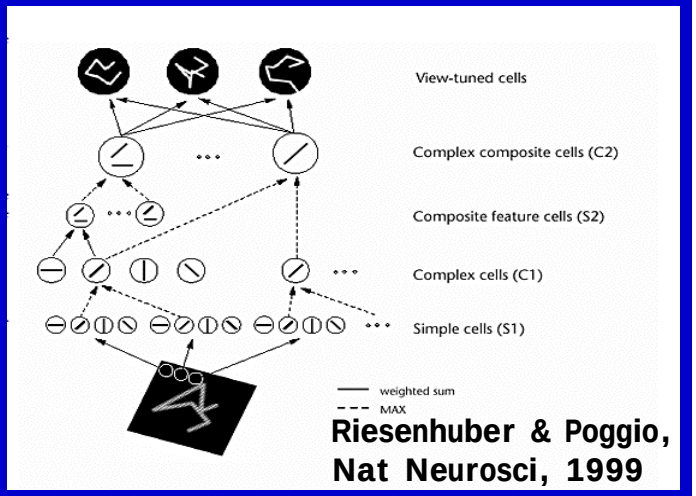
Figure 4. Triadic Architecture. It is suggested that the visual perception of scenes may be carried out via the interaction of three different systems. System I: Early-level processes produce volatile proto-objects rapidly and in parallel across the visual field. System II: Focused attention acts as 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.

The next step…

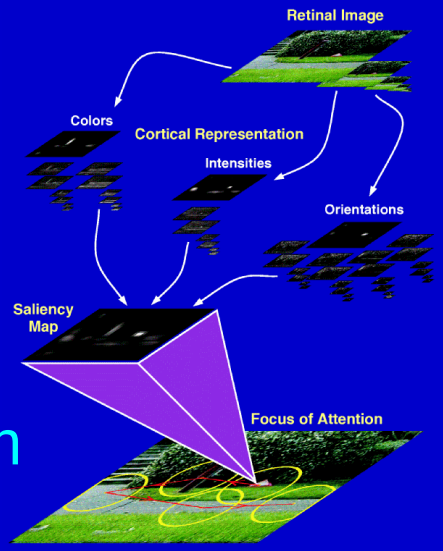
Develop scene understanding/navigation/orienting mechanisms that can exploit the (very noisy) "rich scanpaths" (i.e., with location and sometimes identification) generated by the model.



Scene Layout & Gist



Localized Object Recognition



Attention

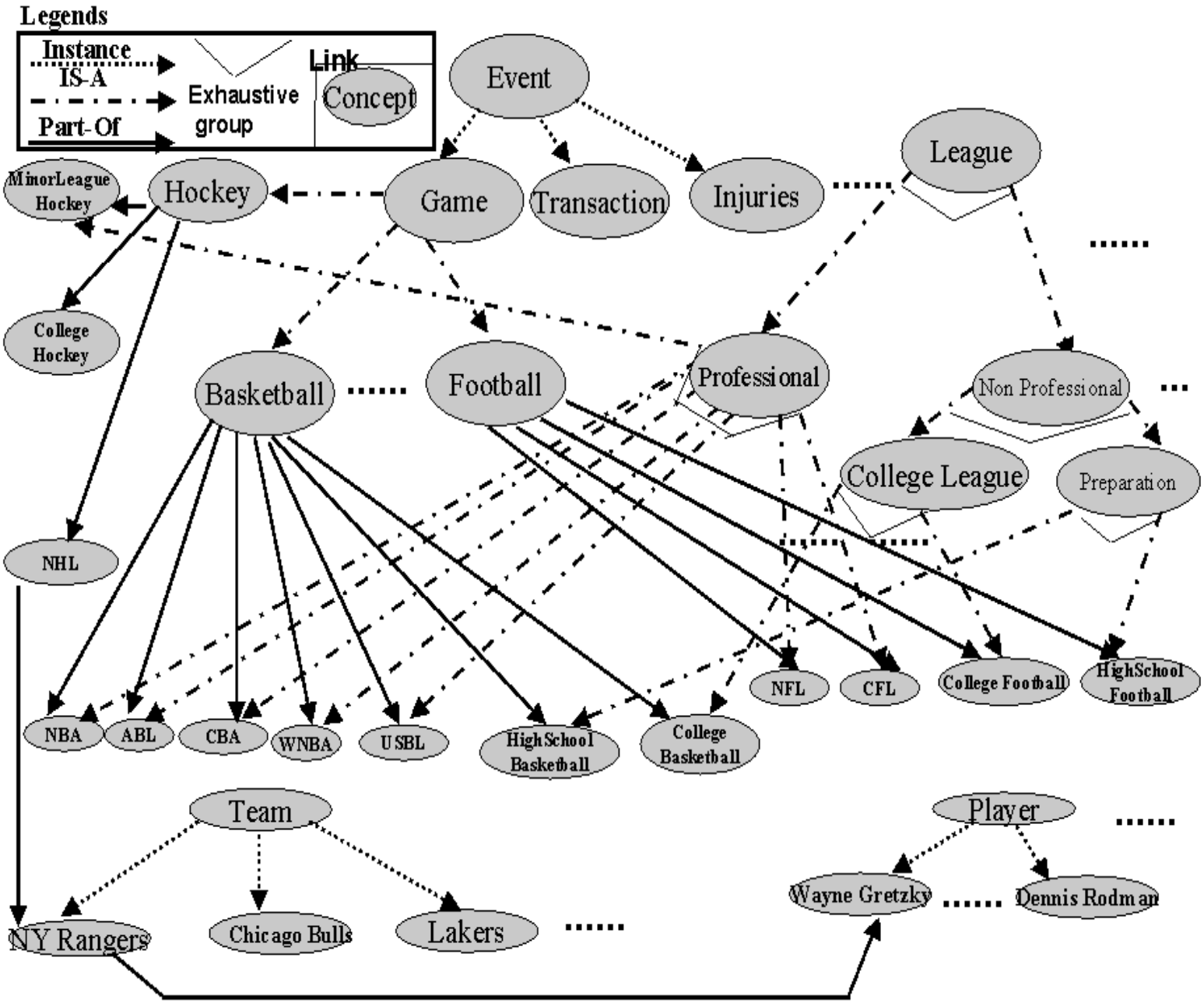
Extract "minimal subscene" (i.e., small number of objects and actions) that is relevant to present behavior.

This requires biasing and pruning the "rich scanpaths" in a task-dependent manner, such as to improve hit rate and to eliminate false positives

Components of the model

- Question/task, e.g., “who is doing what to whom?”
- Lexical parser to extract key concepts from question
- Ontology of world concepts and their inter-relationships, to expand concepts explicitly looked for to related ones
- Attention/recognition/gist+layout visual subsystems to locate candidate relevant objects/actors/actions
- Working memory of concepts relevant to current task
- Spatial map of locations relevant to current task

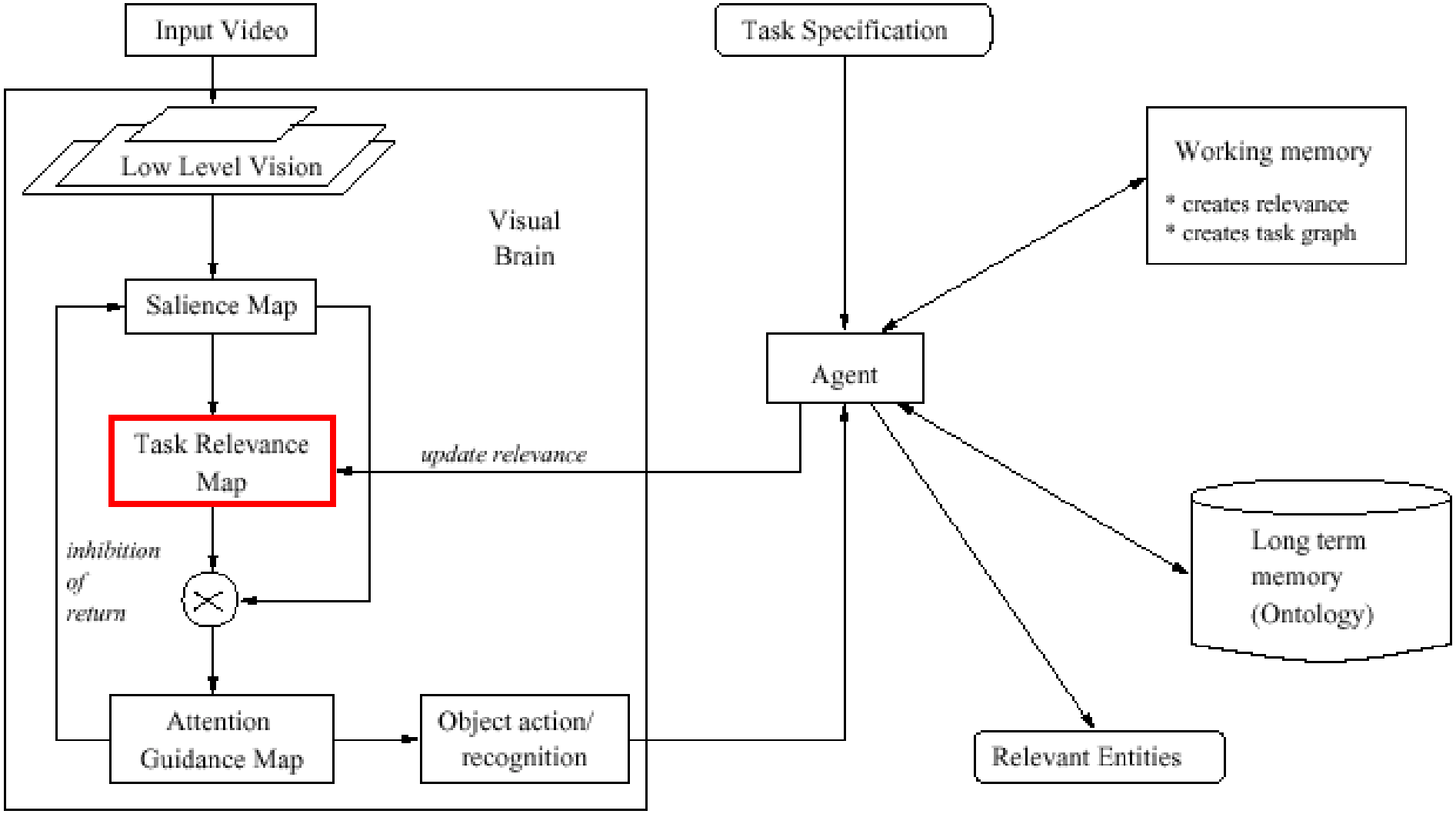
Ontology



Khan & McLeod, 2000

The task-relevance map

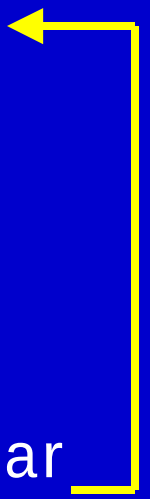
Scalar topographic map, with higher values at more relevant locations



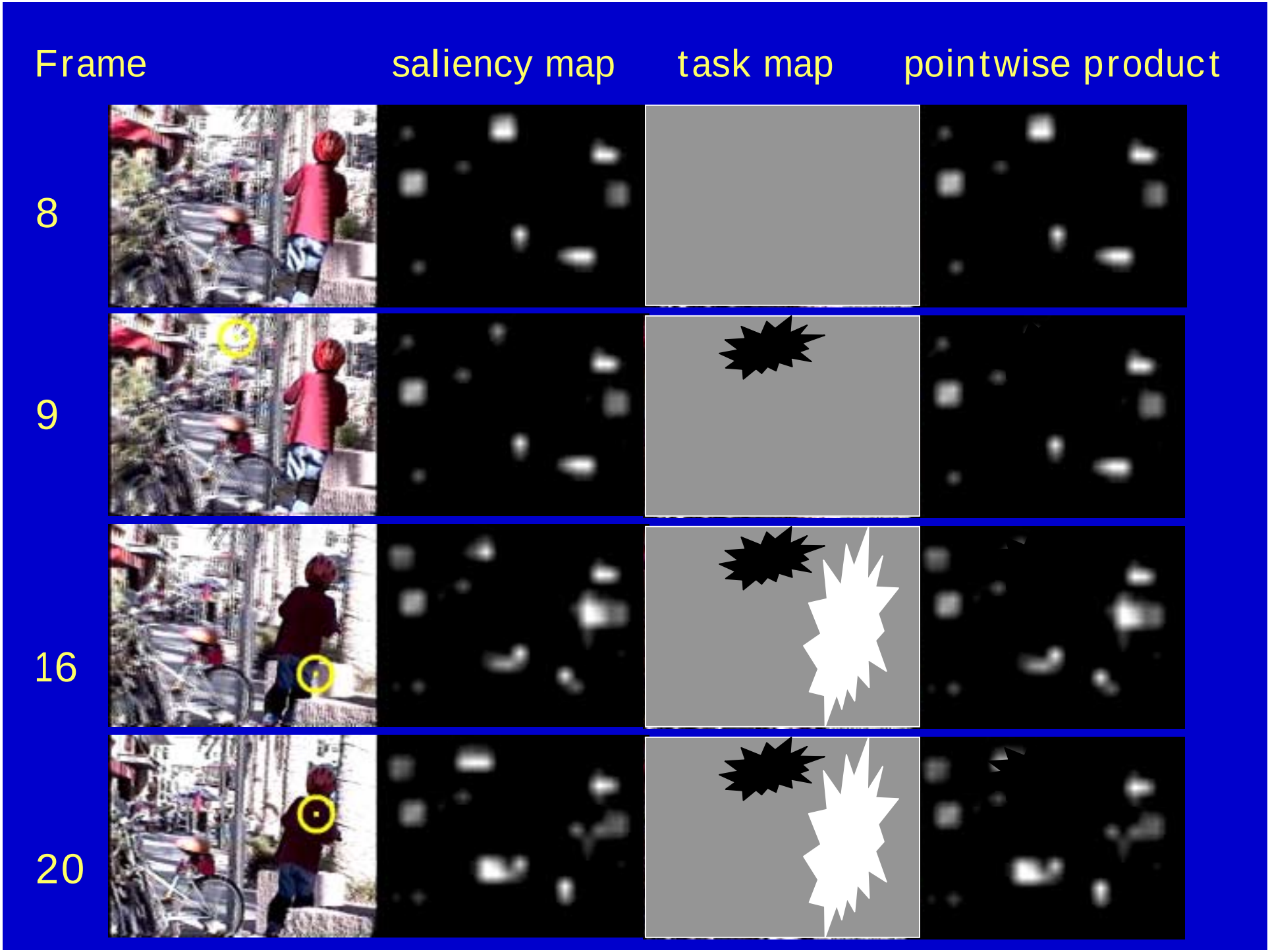
Model operation

- Receive and parse task specification; extract concepts being looked for
- Expand to wider collection of relevant concepts using ontology

- Bias attention towards the visual features of most relevant concept
- Attend to and recognize an object
- If relevant, increase local activity in task map
- Update working memory based on understanding so far



After a while: task map contains only relevant regions, and attention primarily cycles through relevant objects


Frame
saliency map
task map
pointwise product
8
9
16
20

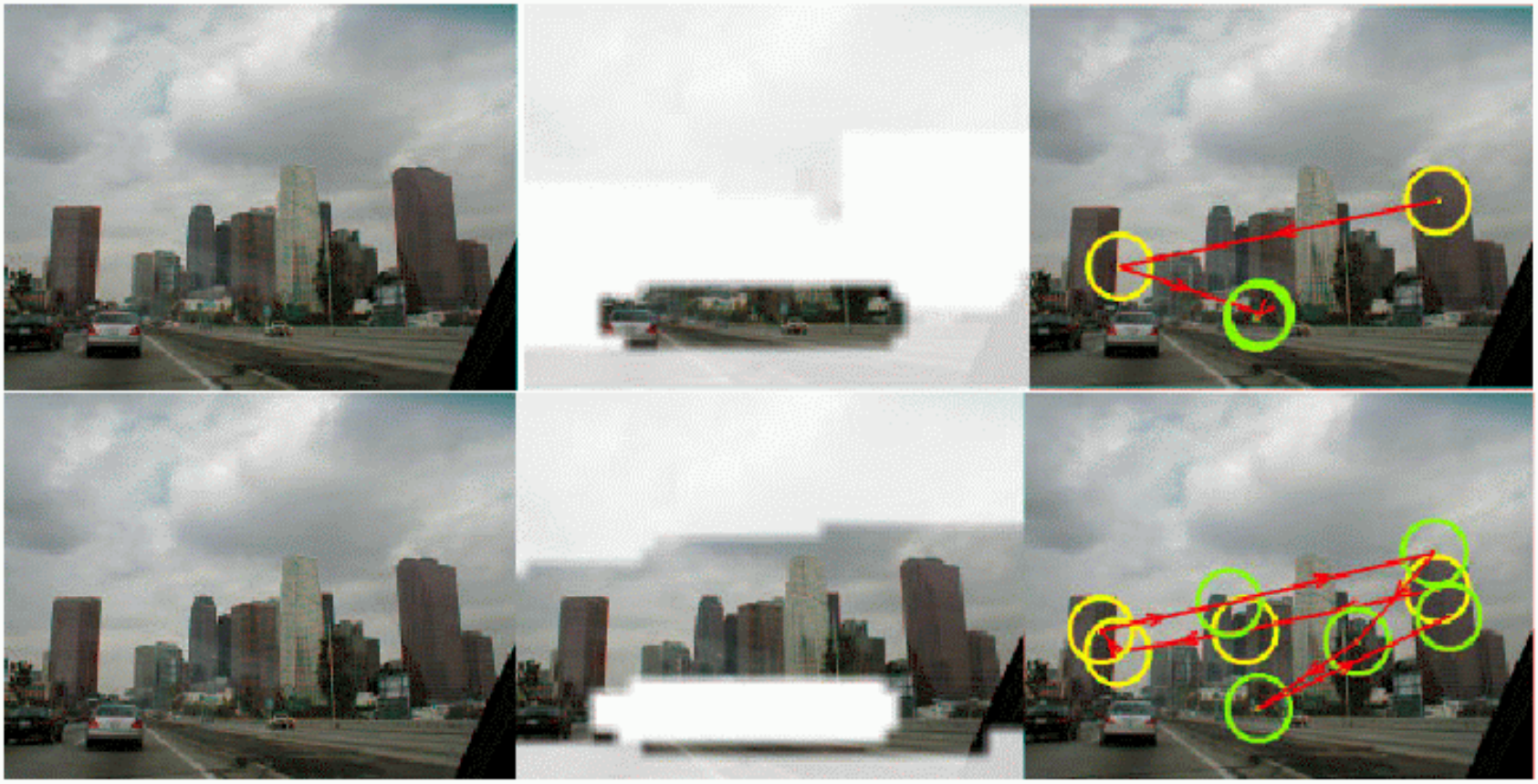
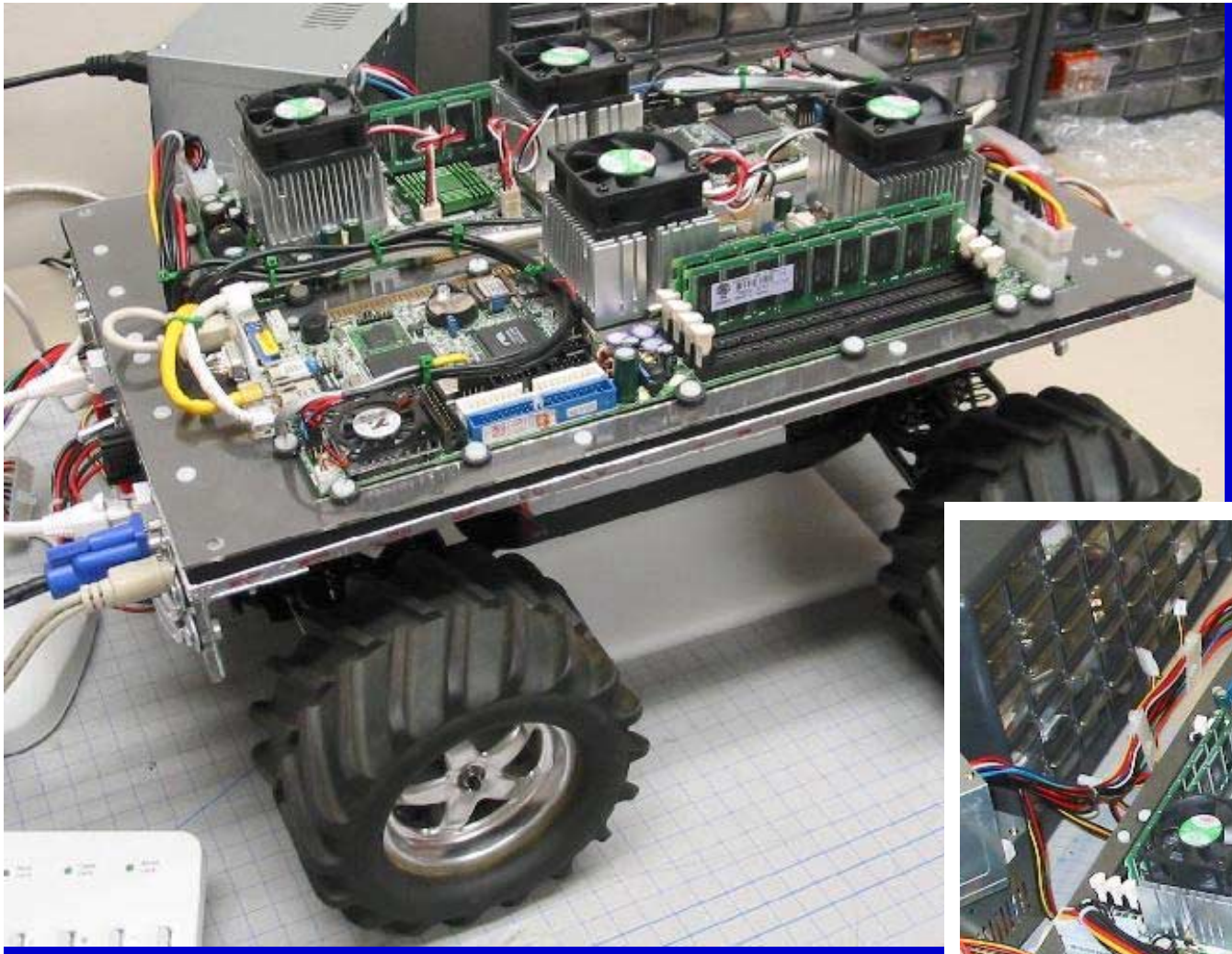


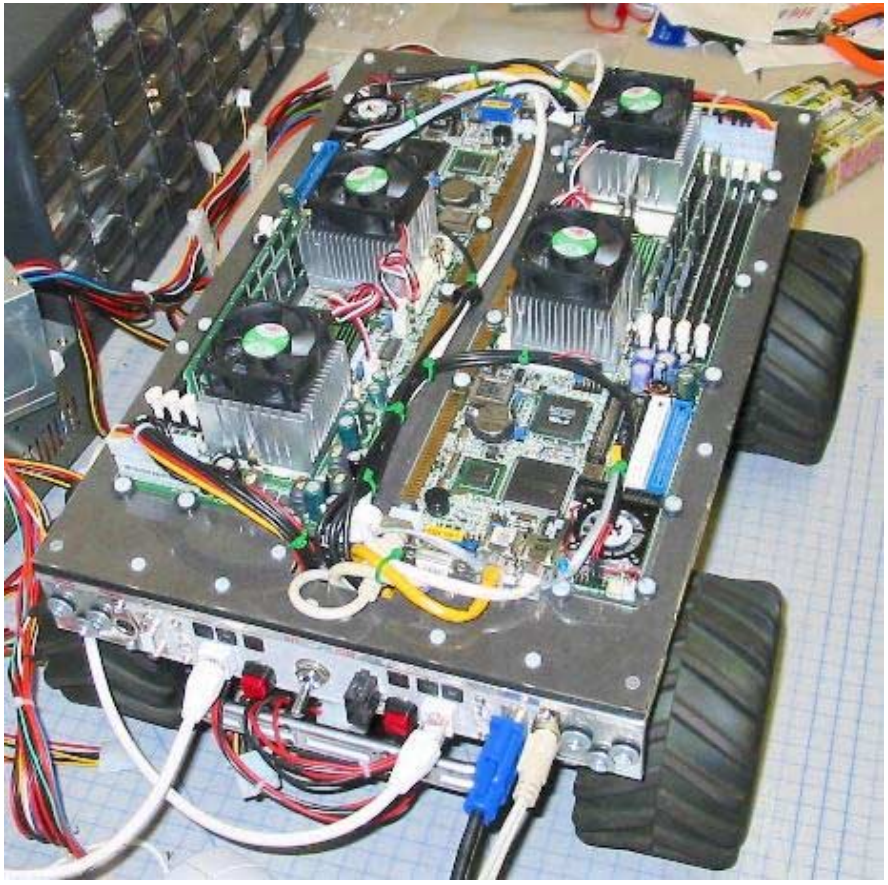
Figure 4: On a city scene, we asked the model to find the cars in the scene. Without any prior knowledge of a city scene, our model picked the relevant portions of the scene in as few as three attentional shifts. Though the initial couple of fixations were primarily salience driven, once the model determined that those fixations i.e. buildings were irrelevant to the task, it refused to attend to them any further despite their salient features. On the same scene, when the model was asked to find the buildings, it attended to all the salient features in the buildings and determined the roads and cars to be irrelevant. In the figure, the first column shows the original scene, followed by the TRM (locations relevant to the task) and finally, the attentional trajectory. The circles represent fixations and each fixation is on a scene segment that is approximately the size of the object.

Figure 5: On a natural cluttered scene, we asked the model to determine the people in the scene and find what they were eating. As expected, the model showed that the relevance of entities in the scene varied with the nature of the task. For the first task, the model looked for people and consequently, it fixated more on human body parts than other irrelevant objects in the scene. While in the second task, the model looked for hand related objects near the human faces and hands to determine what the people were eating. In the figure, the first column is the original image, followed by the TRM after three attentional shifts and the final TRM after ten attentional shifts.



Beowulf + Robot =

"Beobot"



Outlook

- The idea of a unique topographic saliency map yields robust predictions of human bottom-up attention
- The challenge now is to develop algorithms that can extract the currently relevant "minimal subscene" from incoming rich scanpaths
- Such algorithms will, we believe, endow autonomous system with more powerful perceptual senses
- Publications, C++ source code, robot specs: http://iLab.usc.edu


82ms