Lecture 13. Scene Perception

Reading Assignments:

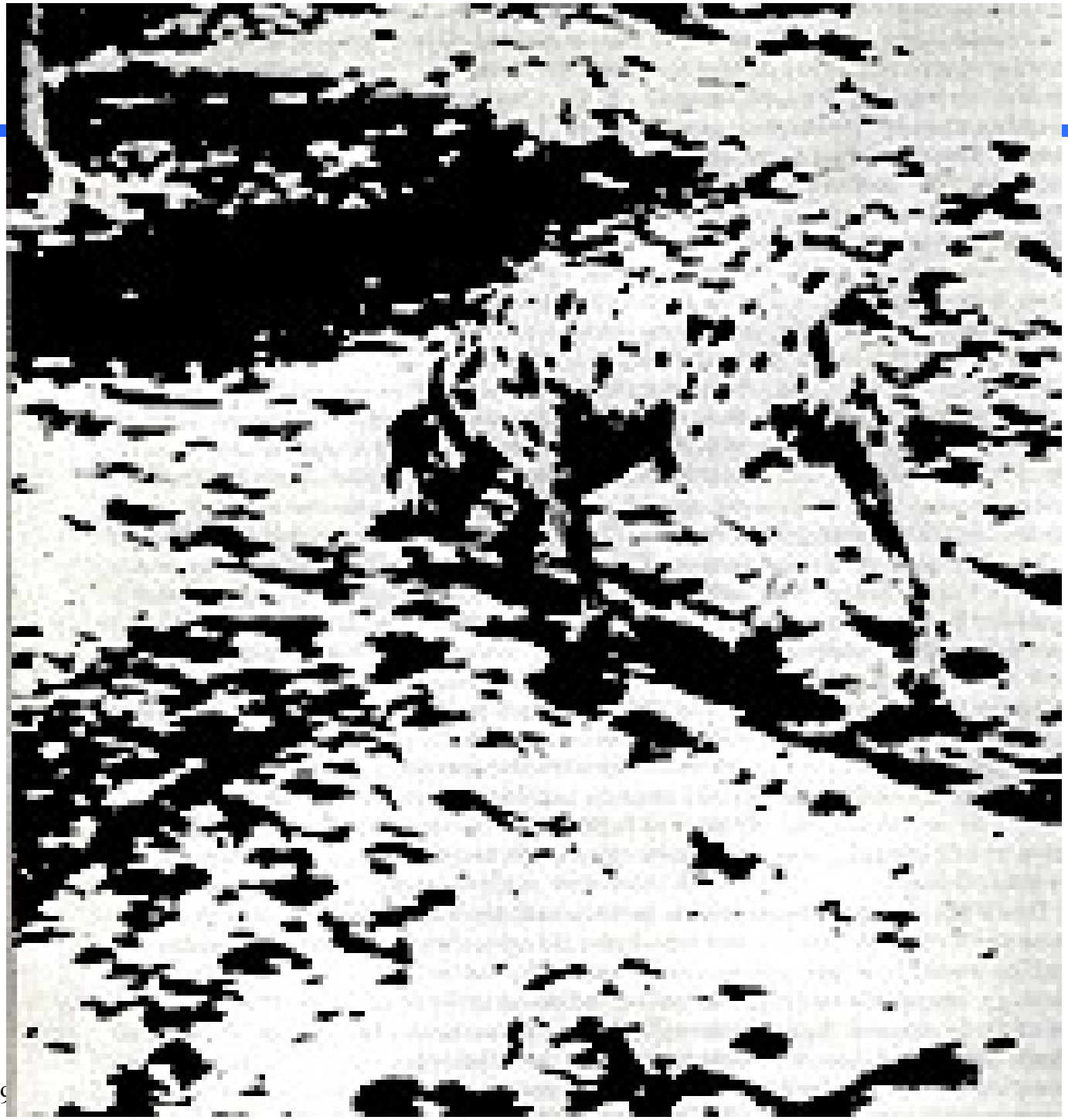
None

How much can we remember?

Incompleteness of memory:

how many windows in the Taj Mahal?

despite conscious experience of picture-perfect, iconic memorization.



But…

We can recognize complex scenes which we have seen before.

So, we do have some form of iconic memory.

In this lecture:

- examine how we can perceive scenes
- what is the representation (that can be memorized)
- what are the mechanisms

Extended Scene Perception

Attention-based analysis: Scan scene with attention, accumulate evidence from detailed local analysis at each attended location.

Main issues:

- what is the internal representation?
- how detailed is memory?
- do we really have a detailed internal representation at all!!?

Gist: Can very quickly (120ms) classify entire scenes or do simple recognition tasks; can only shift attention twice in that much time!

Accumulating Evidence

Combine information across multiple eye fixations.

Build detailed representation of scene in memory.

Eye Movements

1) Free examination

2) estimate material circumstances of family

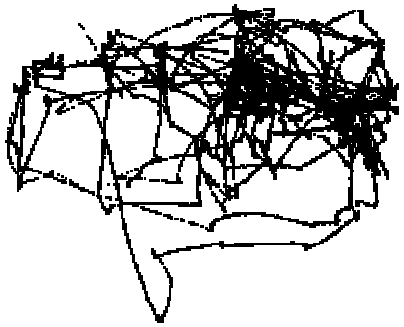
3) give ages of the people

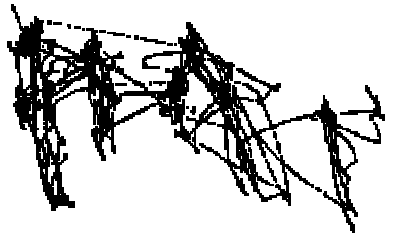
4) surmise what family has been doing before arrival of “unexpected visitor”

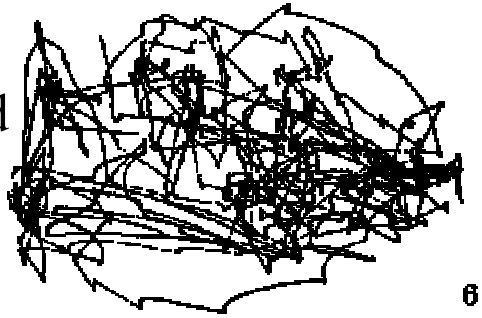
5) remember clothes worn by the people

6) remember position of people and objects

7) estimate how long the “unexpected visitor” has been away from family



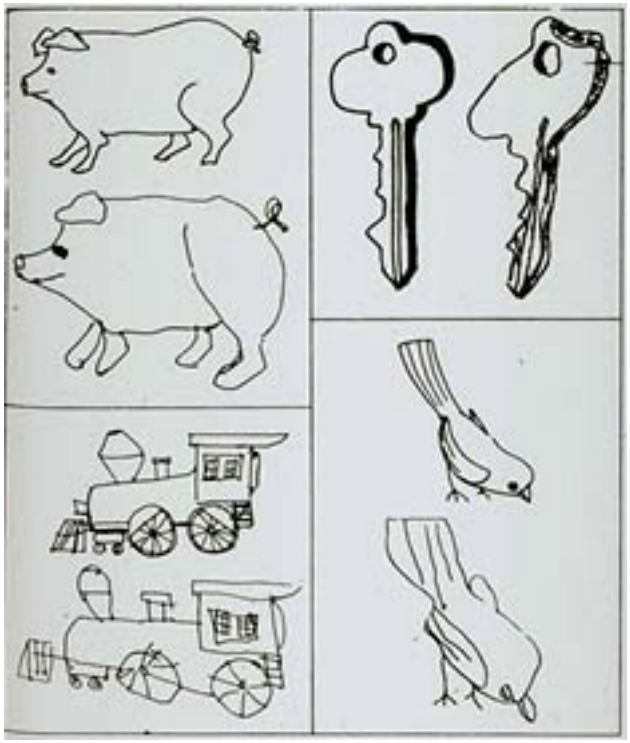




Clinical Studies

Studies with patients with some visual deficits strongly argue that tight interaction between where and what/how visual streams are necessary for scene interpretation.

Visual agnosia: can see objects, copy drawings of them, etc., but cannot recognize or name them!



Dorsal agnosia: cannot recognize objects if more than two are presented simultaneously: problem with localization

Ventral agnosia: cannot identify objects.

These studies suggest…

We bind features of objects into objects (feature binding)

We bind objects in space into some arrangement (space binding)

We perceive the scene.

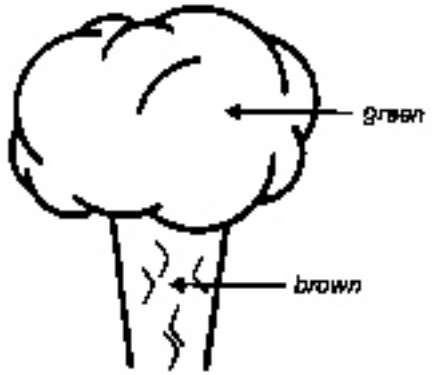
Feature binding = what/how stream
Space binding = where stream

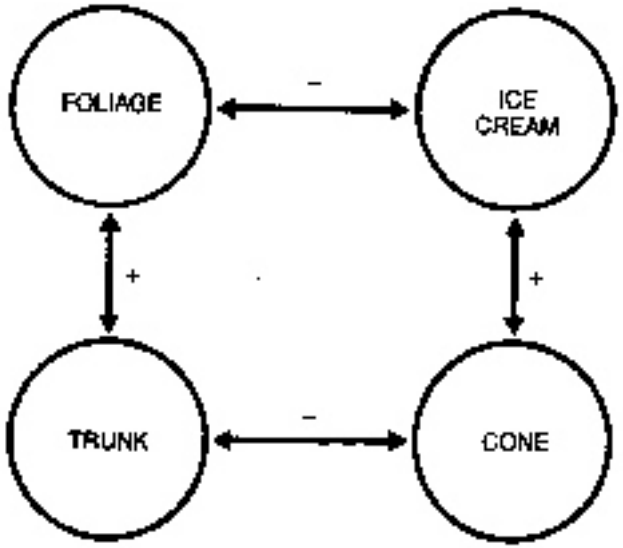
Schema-based Approaches

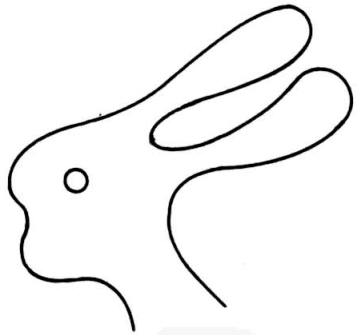
Schema (Arbib, 1989): describes objects in terms of their physical properties and spatial arrangements.

Abstract representation of scenes, objects, actions, and other brain processes. Intermediate level between neural firing and overall behavior.

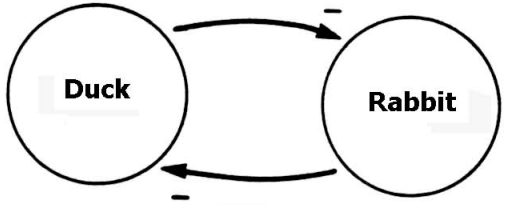
Schemas both cooperate and compete in describing the visual world:



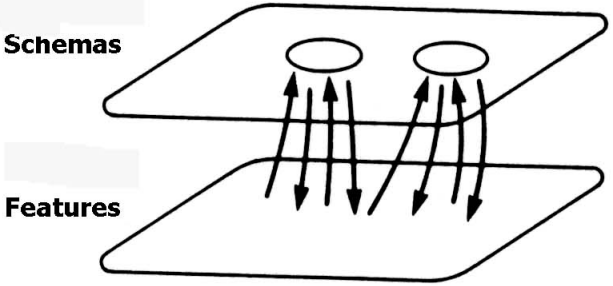




(a)



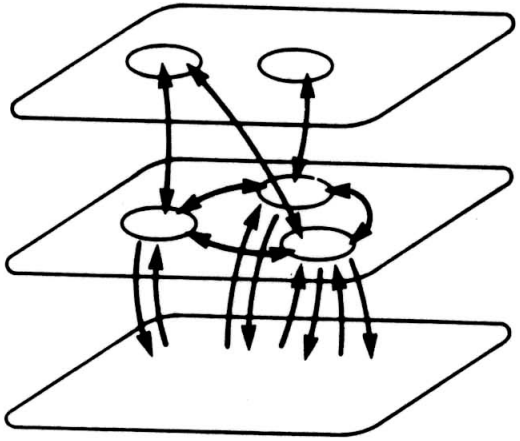
(b)



Lower-level features activate schemas

Schemas cover lower-level features

(c)



Schemas

Episodes: Schema instances

Features

(d)

VISOR

Leow & Miikkulainen, 1994: low-level -> sub-schema activity maps (coarse description of components of objects) -> competition across several candidate schemas -> one schema wins and is the percept.

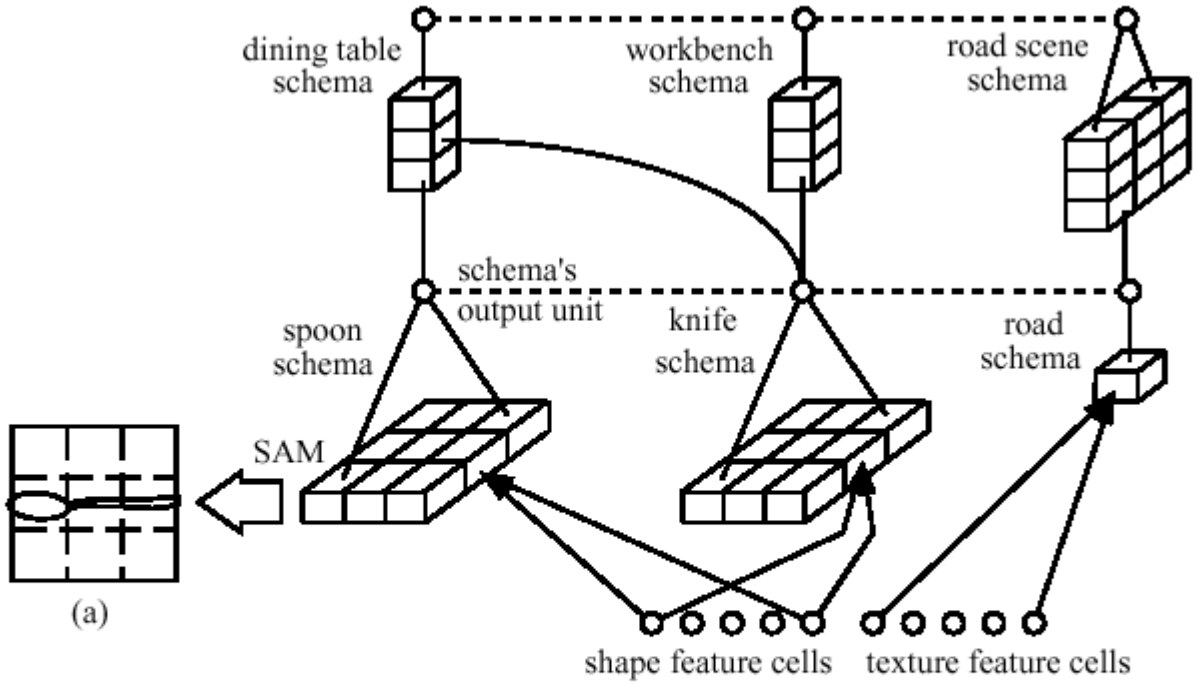
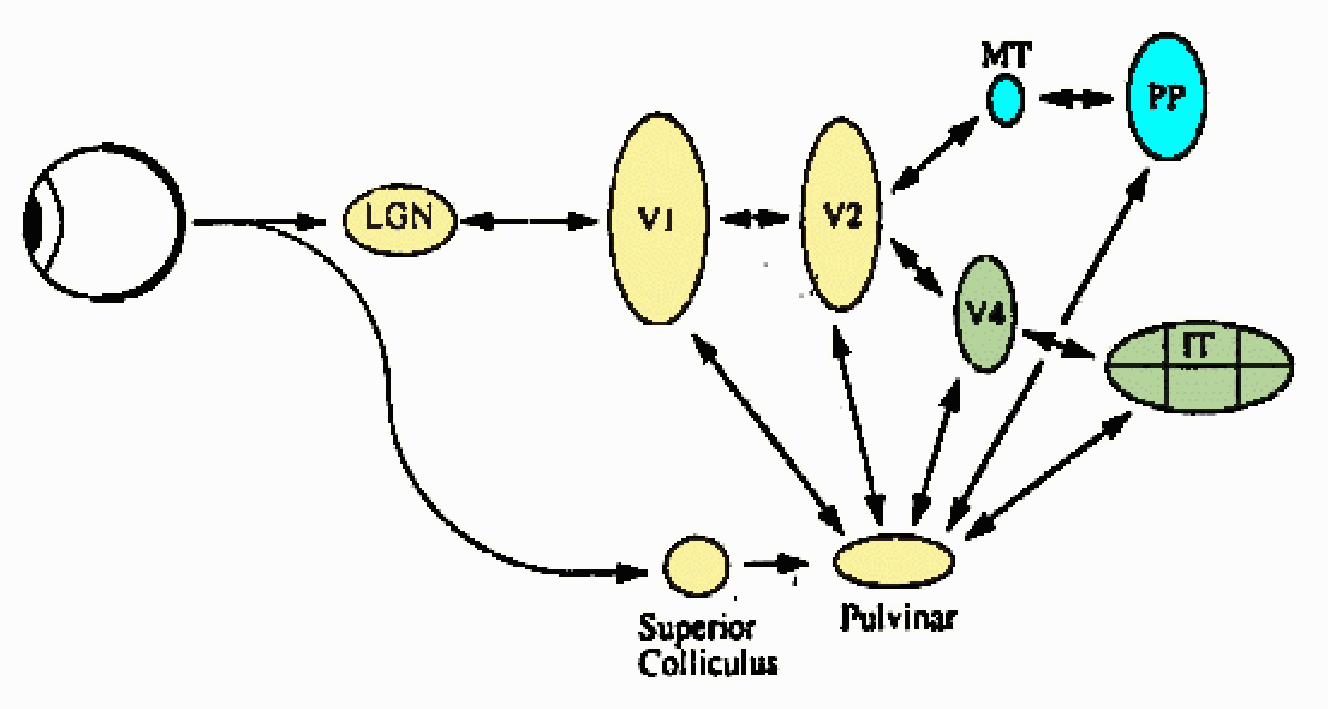


Figure 2: The hierarchy of schema-nets. Schemas' output units are represented as circles, and the units of the Sub-schema Activity Maps (SAMs) as cubes. Arrows represent one-way connections, solid lines represent both the bottom-up and top-down connections (which are different), and dotted lines denote the inhibition among the schemas' output units. (a) The SAM units representing the relative positions of the spoon's components.

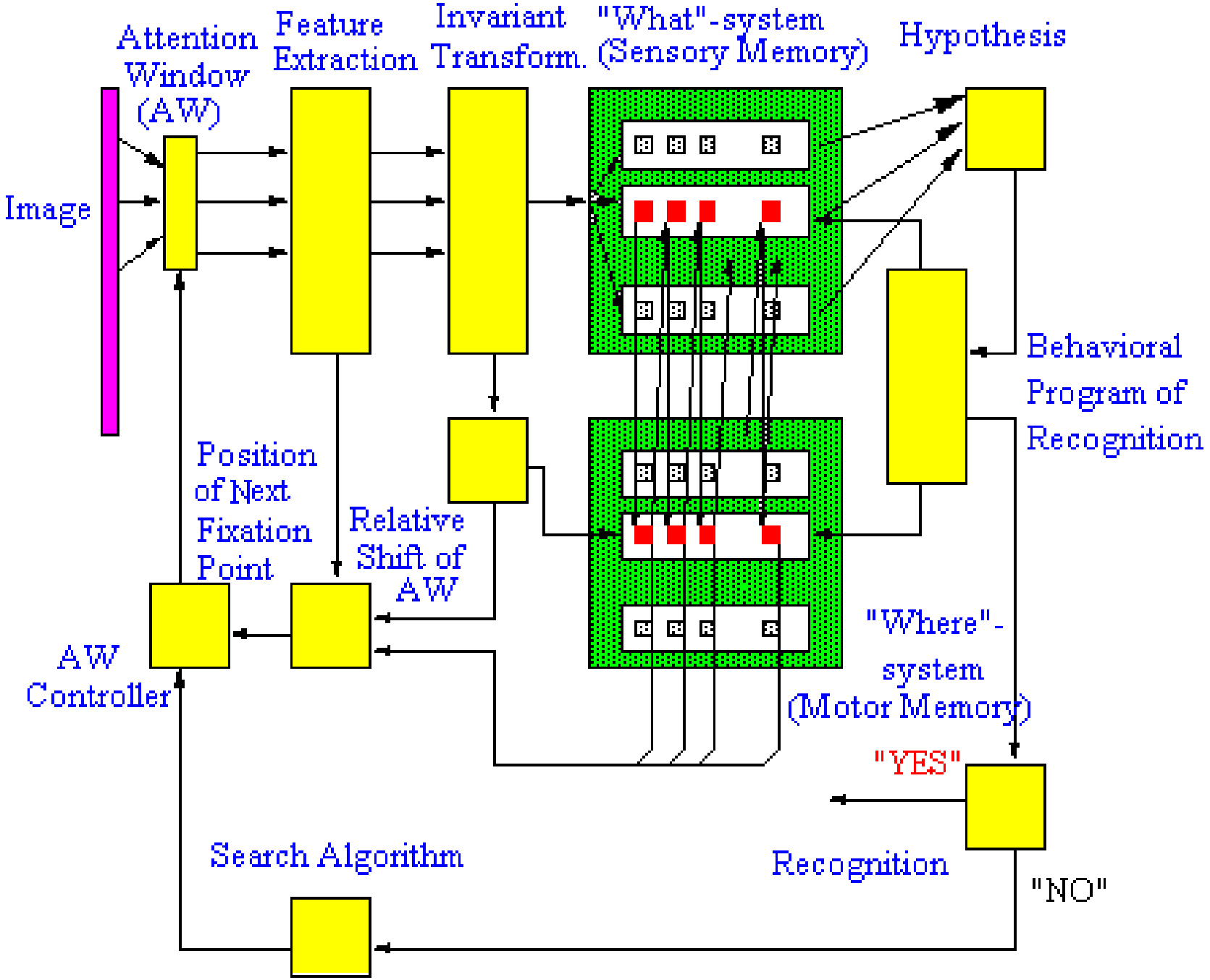
Biologically-Inspired Models

Rybak et al, Vision Research, 1998.



What & Where.

Feature-based frame of reference.


Attention Window (AW)
Feature Extraction
Invariant Transform.
"What"-system (Sensory Memory)
Hypothesis
Image
Behavioral Program of Recognition
Position of Next Fixation Point
Relative Shift of AW
AW Controller
"Where"-system (Motor Memory)
"YES"
Search Algorithm
Recognition
"NO"

Algorithm

- At each fixation, extract central edge orientation, as well as a number of "context" edges;

- Transform those low-level features into more invariant "second order" features, represented in a referential attached to the central edge;

- Learning: manually select fixat store sequence of second-order features found at each fixation into "what" memory; also store vector for next fixation, based on context points and in the second-order referential;

Selection of the Next Point of Fixation

Figure 4. The next fixation point is selected from the set of context points in the current retinal image. The current and next fixation points are marked by *crosses* (*right*). Shift to the next fixation point is shown by the *black arrow* (*right*).

Algorithm

Scanpath

Figure 5.The scanpath of image viewing is shown on background of the initial image (*left*) and on background of the sequence of retinal images along the scanpath (*right*).

As a result, sequence of retinal images is stored in “what” memory, and corresponding sequence of attentional shifts in the “where” memory.

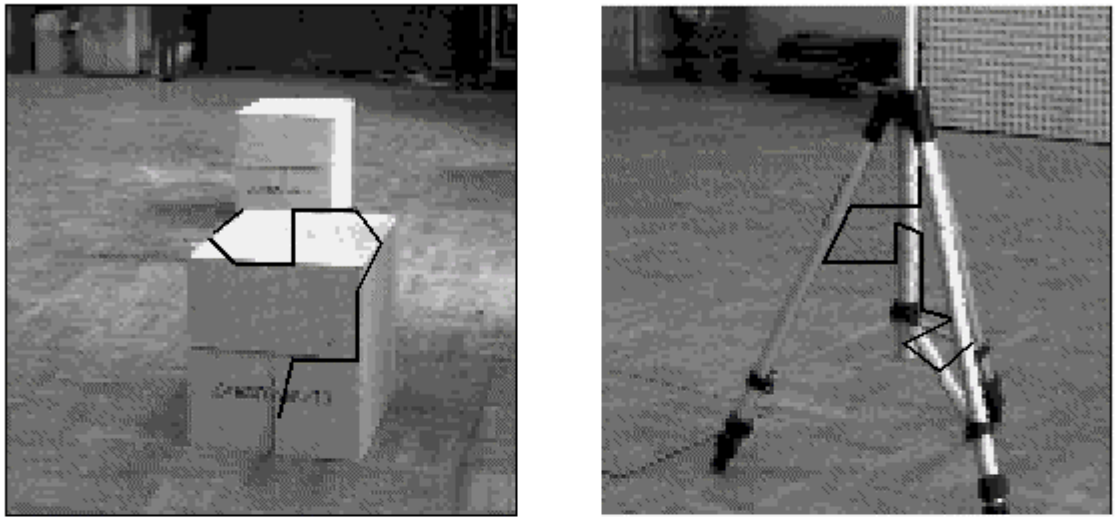
Algorithm

- Search mode: look for an image patch that matches one of the patches stored in the "what" memory;

- Recognition mode: reproduce scanpath stored in memory and determine whether we have a match.

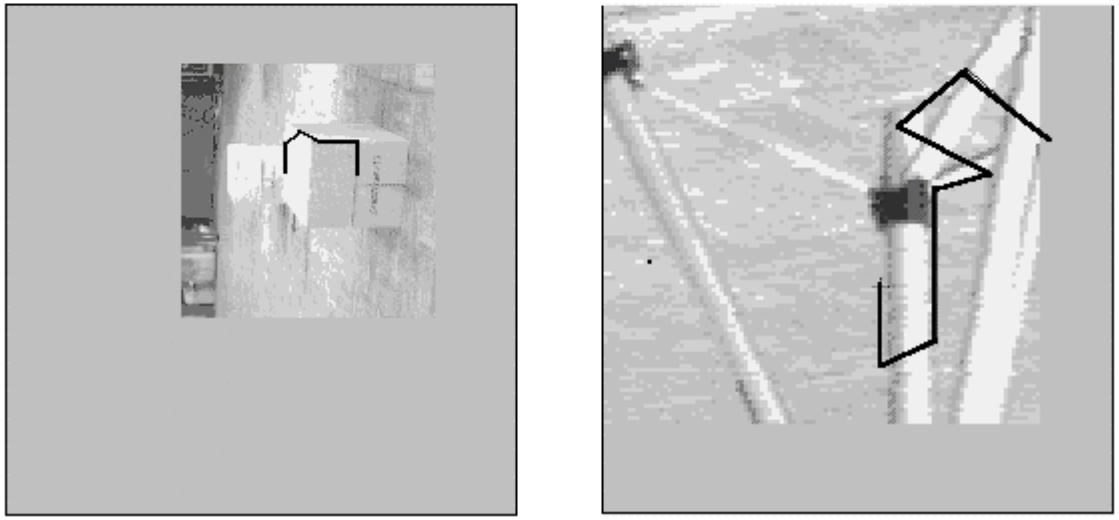
(A) Active viewing and perception of the image:

The scanpaths of viewing are shown black

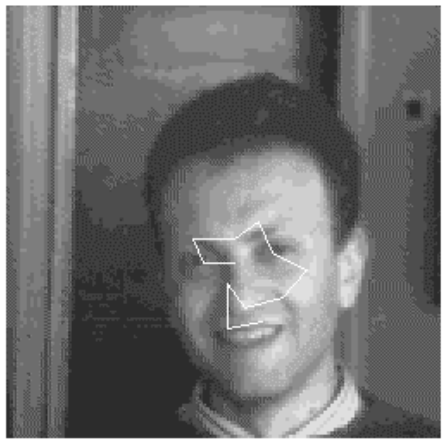


(B) Behavioral (active) process of image recognition:

The scanpaths of recognition are shown black

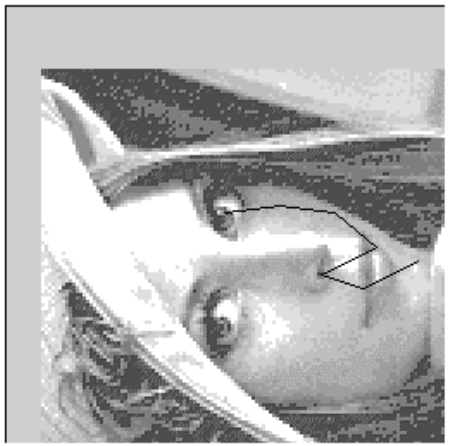


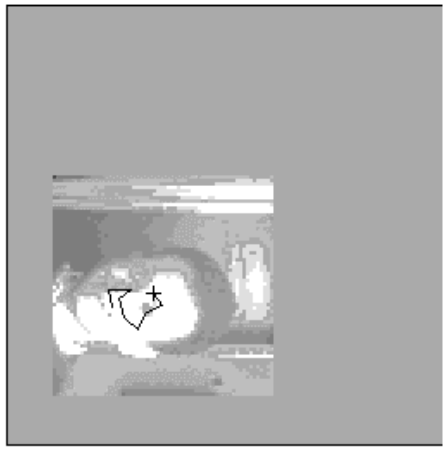
Robust to variations in scale, rotation, illumination, but not 3D pose.



(B) Behavioral (active) process of image recognition:

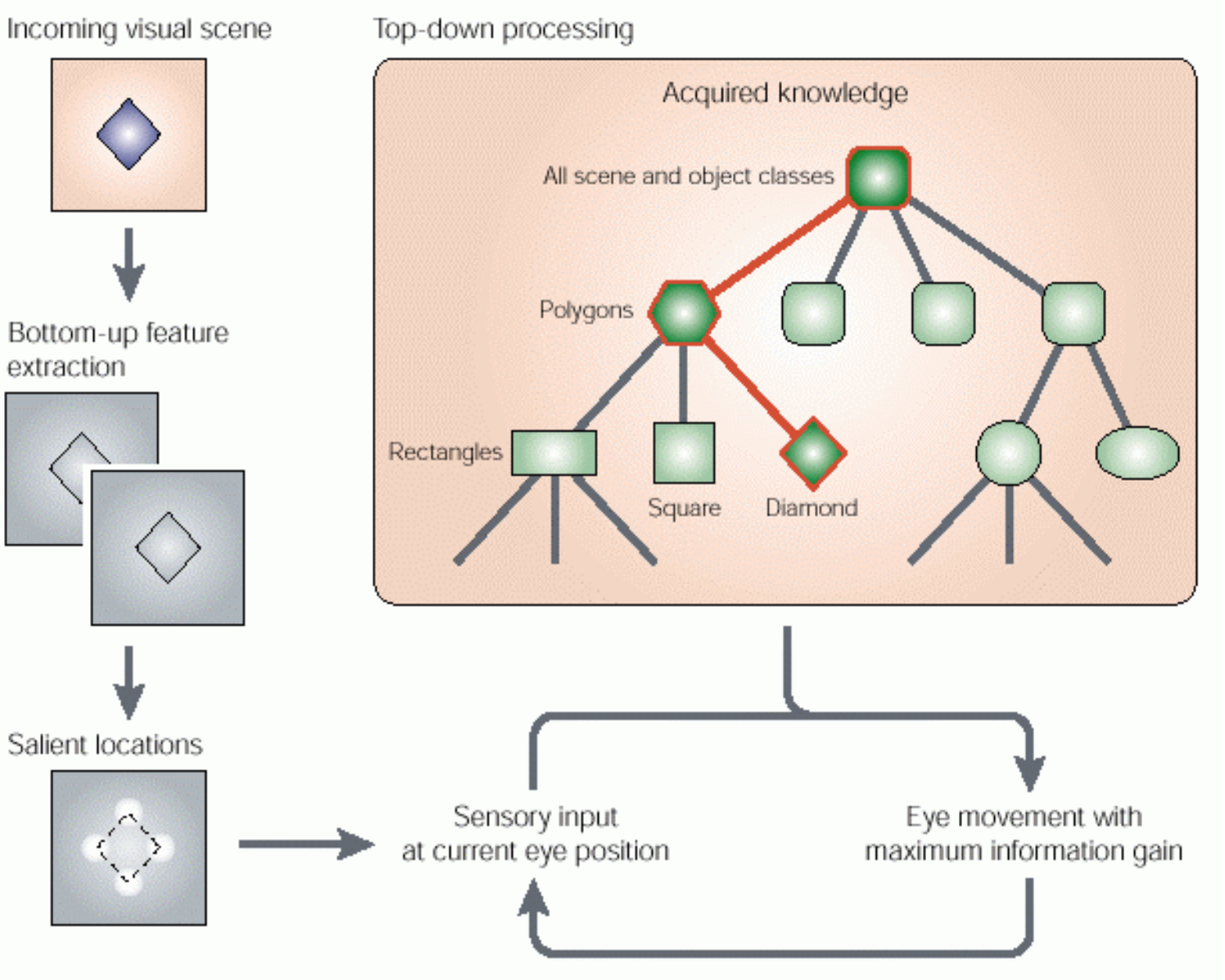
The scanpaths of recognition are shown black





(C) Results of recognition

Schill et al, JEI, 2001



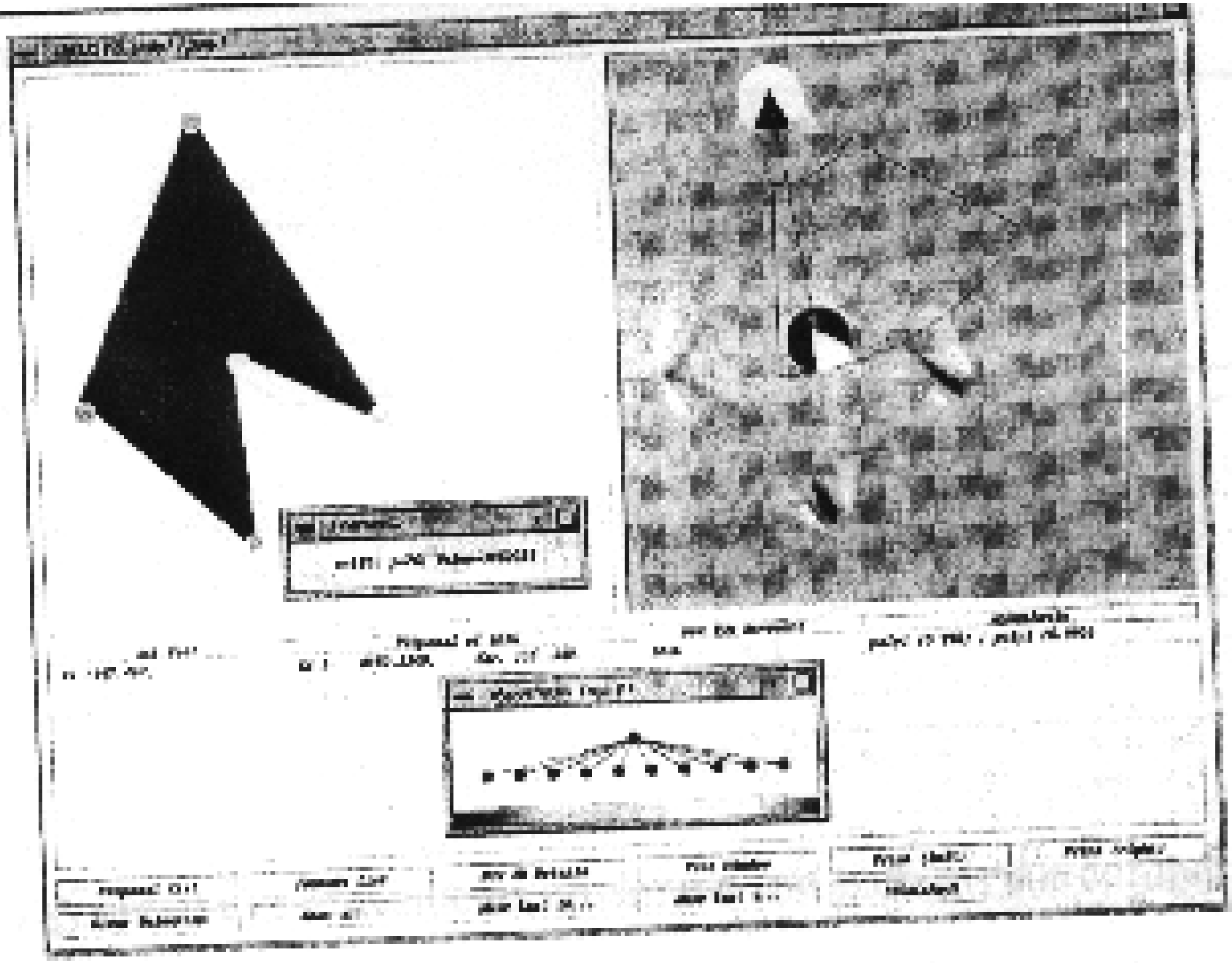


Figure 6: System tests with polygon patterns

Dynamic Scenes

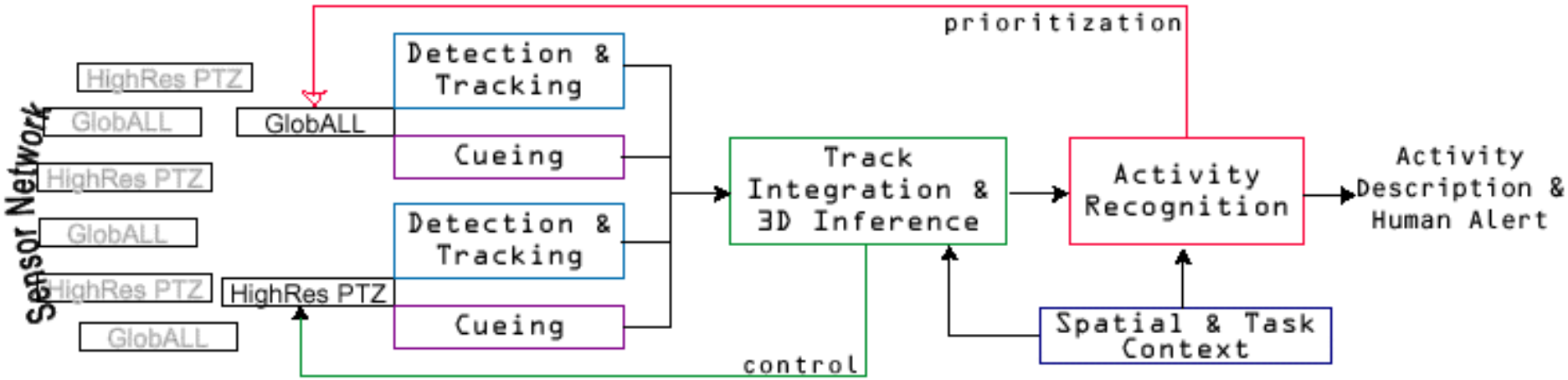
Extension to moving objects and dynamic environment.

Rizzolatti: mirror neurons in monkey area F5 respond when monkey observes an action (e.g., grasping an object) as well as when he executes the same action.

Computer vision models: decompose complex actions using grammars of elementary actions and precise composition rules. Resembles temporal extension of schema-based systems. Is this what the brain does?

Human activity detection

Nevatia/Medioni/Cohen/Itti



Low-level processing

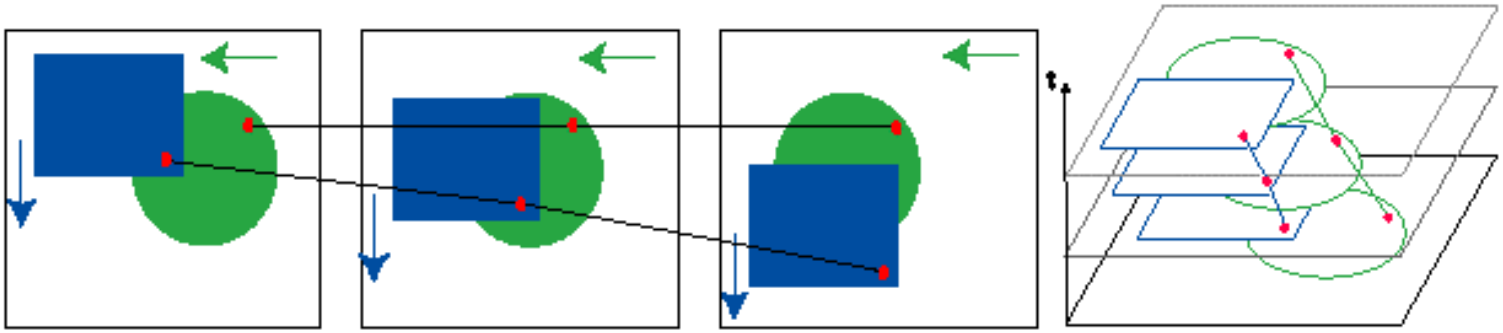
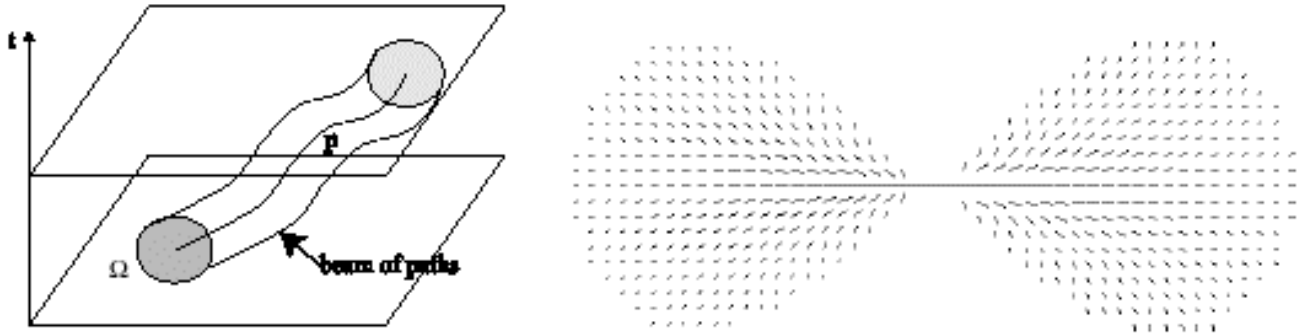


Figure 4: *Example of construction of paths from optical flow field in the* $2D + t$ *space.*



a b

Figure 5: *Integration along a beam of paths of the motion field for robust inference of a pixel trajectory.* **a.** *Illustration of the beam for a circular domain* Ω. **b.** *Illustration of the measure function* $\mu(\omega)$ *along the* x*—axis.*

Spatio-temporal representation

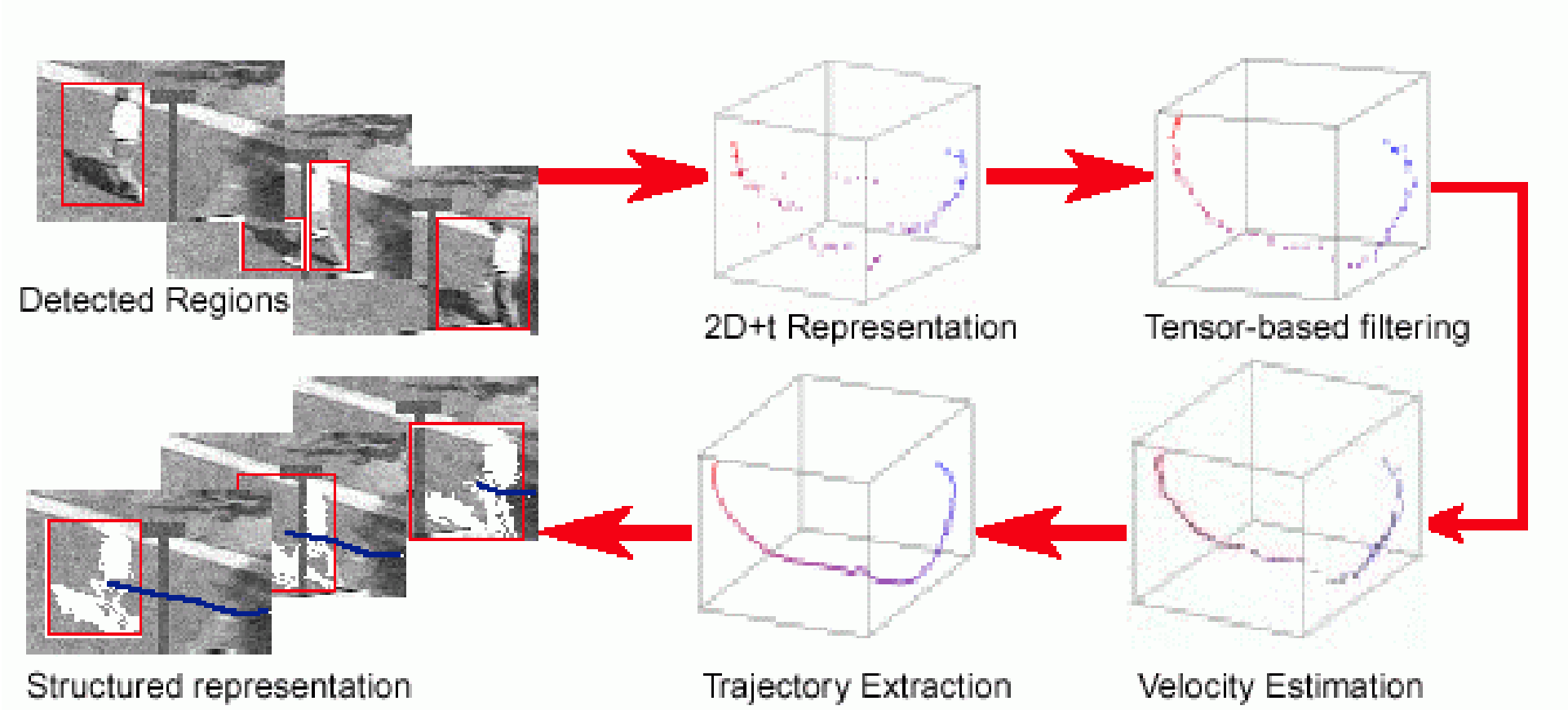


Figure 6: *Inferring the structured representation of a video stream.*

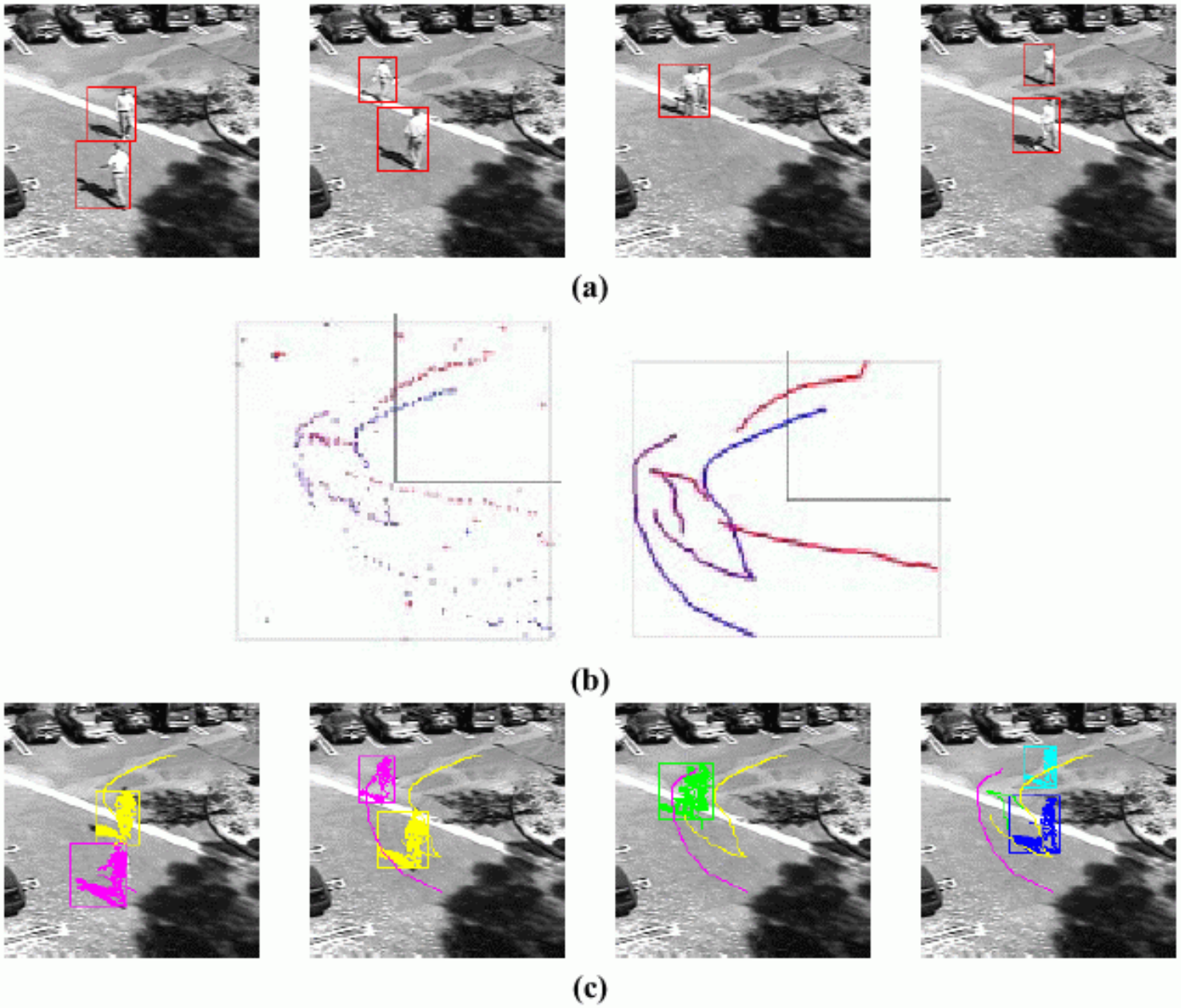


Figure 7: *Structured representation of a video stream of two persons moving in a parking lot.* **(a)** *Detected moving regions,* **(b)** $2D + t$ *representation and inference of trajectories,* **(c)** *Mapping of the structured representation onto the original video frames.*

Modeling Events

Spatial Location			Primary Motion	
at	between	above / below	toward / away	along
inside / outside	among	the front/back of	up / down	around
near / far	on top of	the left / right of	into / out of	through / across
next to	on bottom of		past	after / before

Table 1: English spatial prepositions (simplified from [27])

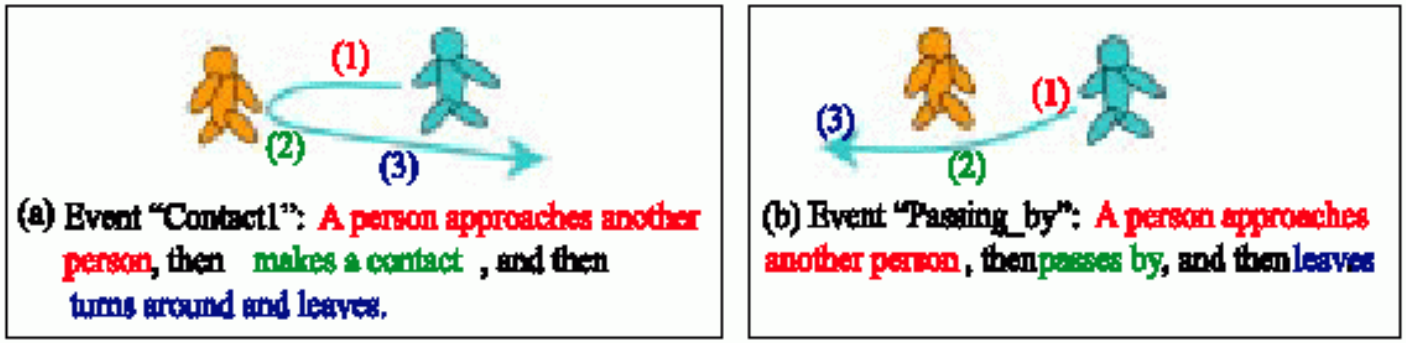
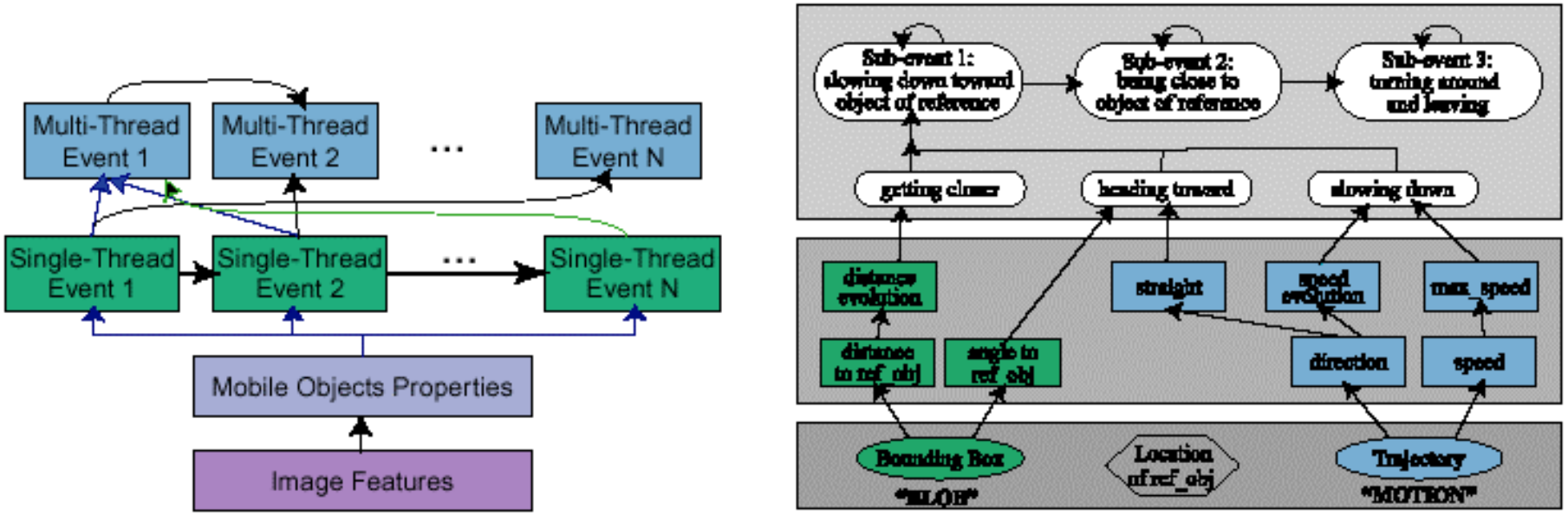


Figure 12: Modeling of two similar complex, single-thread events related to the meeting pattern of two persons. Each event is composed of three simple sub-events.

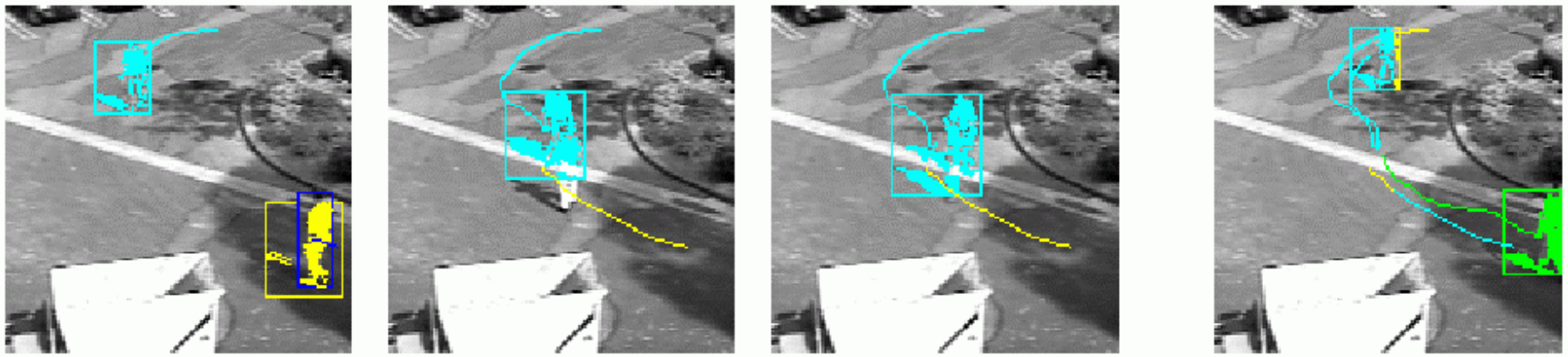
Modeling Events



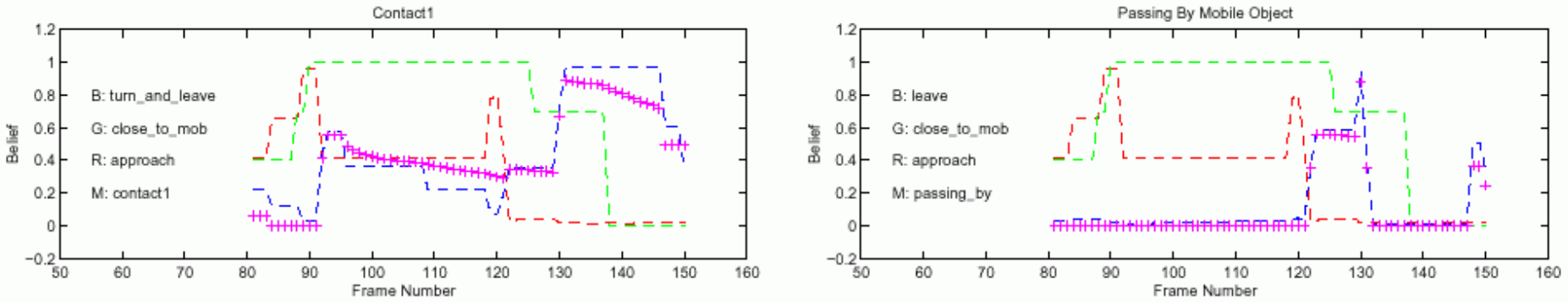
(a) Event Modeling Schema

(b) A representation of complex event "Contact1"

Figure 13: A global view of our proposed scenario modeling; scenarios are defined as a single-thread or a multi-thread event which is described by the associated mobile object properties and image features.



(a) Detection and tracking of moving regions for scenario "CONTACT1".



I) CONTACT1 II) PASSING_BY

(b) Recognition results of two competing activities.

Figure 15: *(a) Input sequence* **A** *shows a complex, single thread event "Contact1". Object 1 (at the top) approaches object 2 (at the bottom), makes contact (both objects have merged as they meet), turns around and leaves. (b) Event "Contact1" is recognized with* $P(MS^*|O) = 0.7$. *Event "Passing By" is recognized with lower probability (almost 0 at the end) since sub-event "leaving without turning around" is not established.*

Several Problems…

with the “progressive visual buffer hypothesis:”

Change blindness:

Attention seems to be required for us to perceive change in images, while these could be easily detected in a visual buffer!

Amount of memory required is huge!

Interpretation of buffer contents by high-level vision is very difficult if buffer contains very detailed representations (Tsotsos, 1990)!

The World as an Outside Memory

Kevin O'Regan, early 90s:

why build a detailed internal representation of the world?

too complex…
not enough memory…

… and **useless**?

The world ***is*** the memory. Attention and the eyes are a look-up tool!

The "Attention Hypothesis"

Rensink, 2000

No "integrative buffer"

Early processing extracts information up to "proto-object" complexity in massively parallel manner

Attention is necessary to bind the different proto-objects into complete objects, as well as to bind object and location

Once attention leaves an object, the binding "dissolves." Not a problem, it can be formed again whenever needed, by shifting attention back to the object.

Only a rather sketchy "virtual representation" is kept in memory, and attention/eye movements are used to gather details as needed

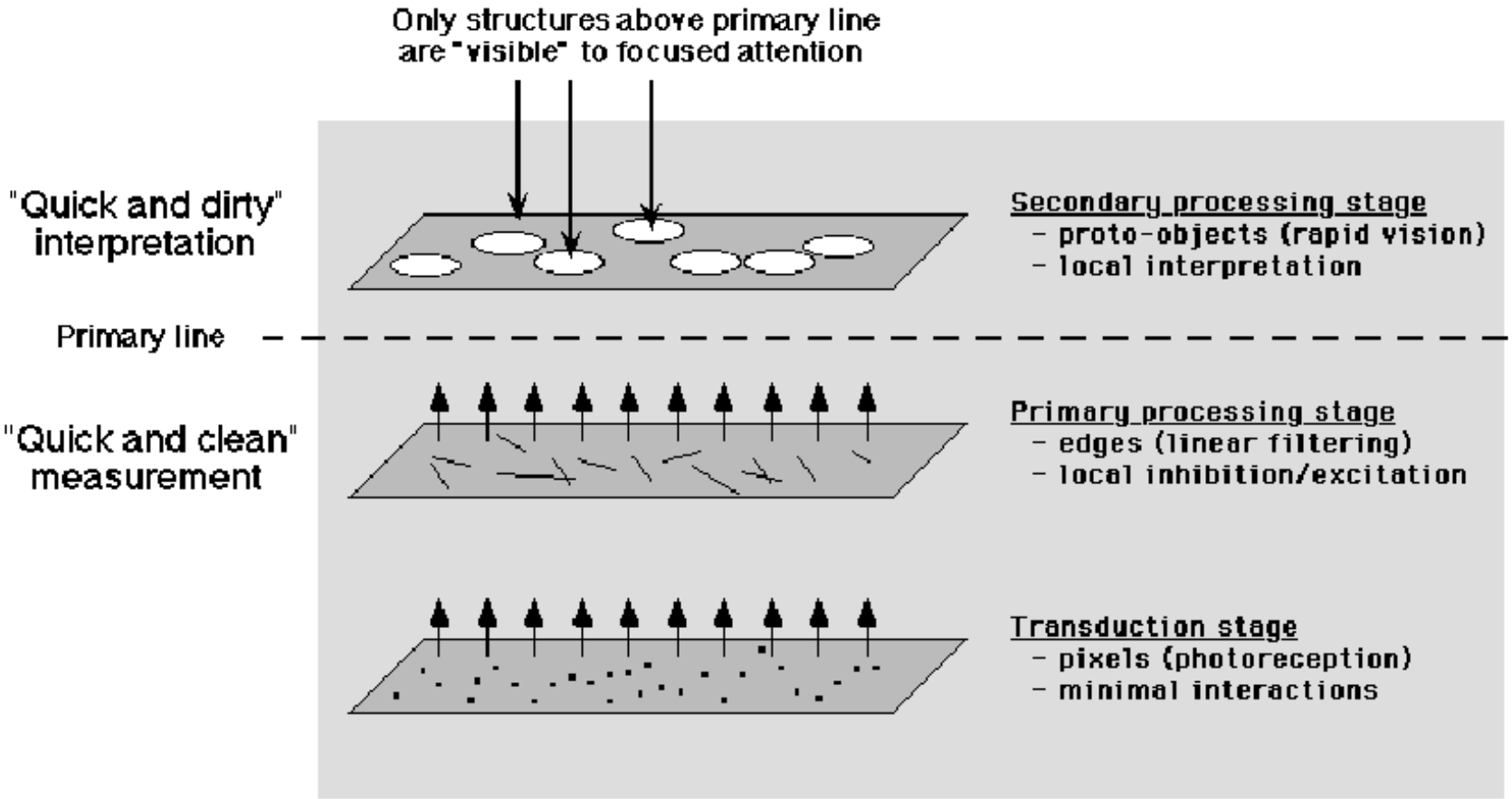


Figure 1

Figure 1. Schematic of Low-Level Vision. Three main stages are distinguished here: (i) the transduction stage, where photoreception occurs, (ii) the primary processing stage, where linear or quasi-linear filters measure image properties, and (iii) the secondary processing stage of rapid non-linear interpretation. Operations at all three stages are carried out in parallel across the visual field. The transduction and primary stages obtain their speed at the expense of complexity; in essence, they perform "quick and clean" measurements. The limits to these kinds of operations are given by the primary line. The secondary stage obtains its speed at the expense of reliability, opting for "quick and dirty" interpretations that may not always be correct. The outputs of this stage are proto-objects that become the operands for attentional processes.

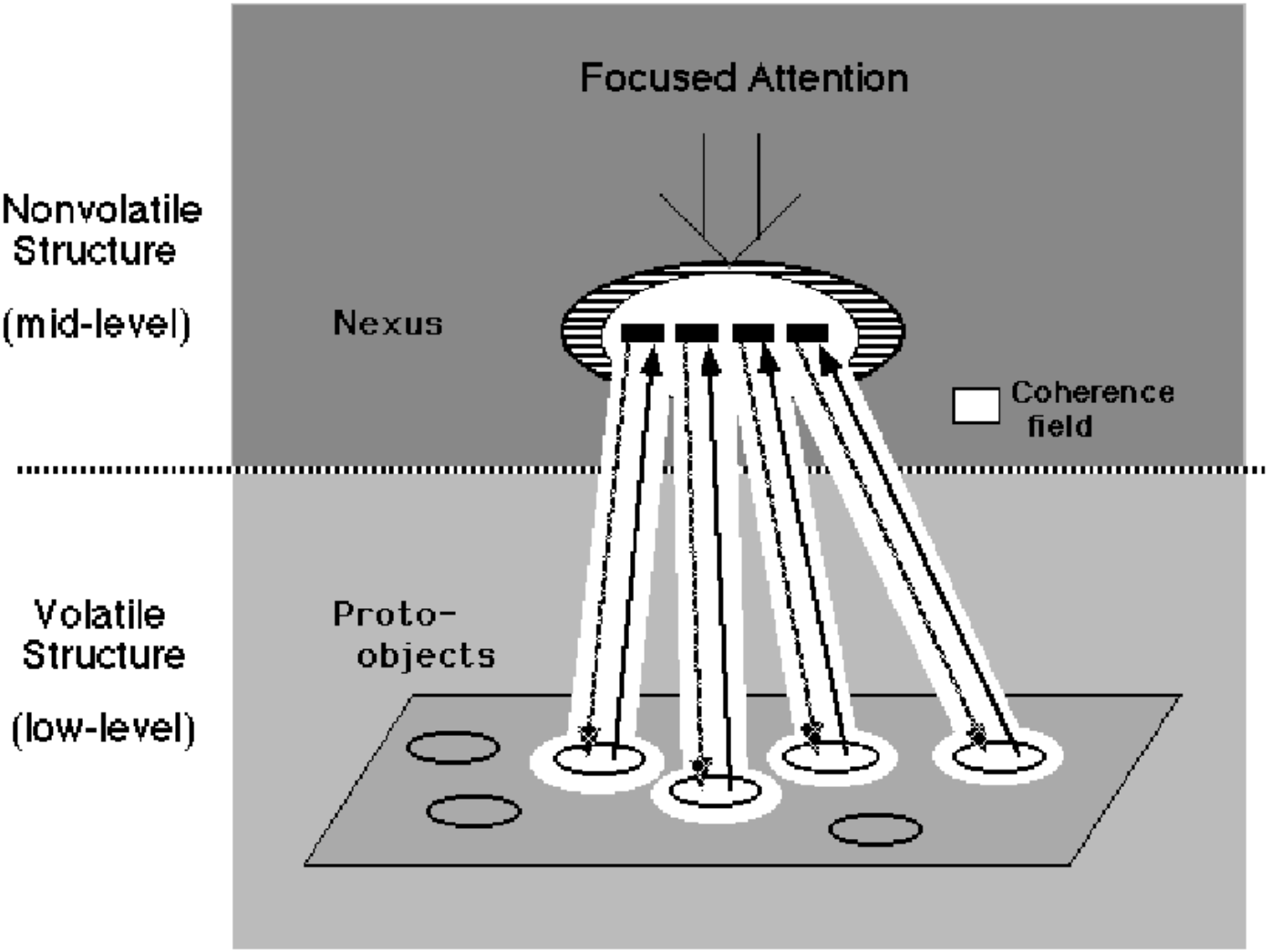
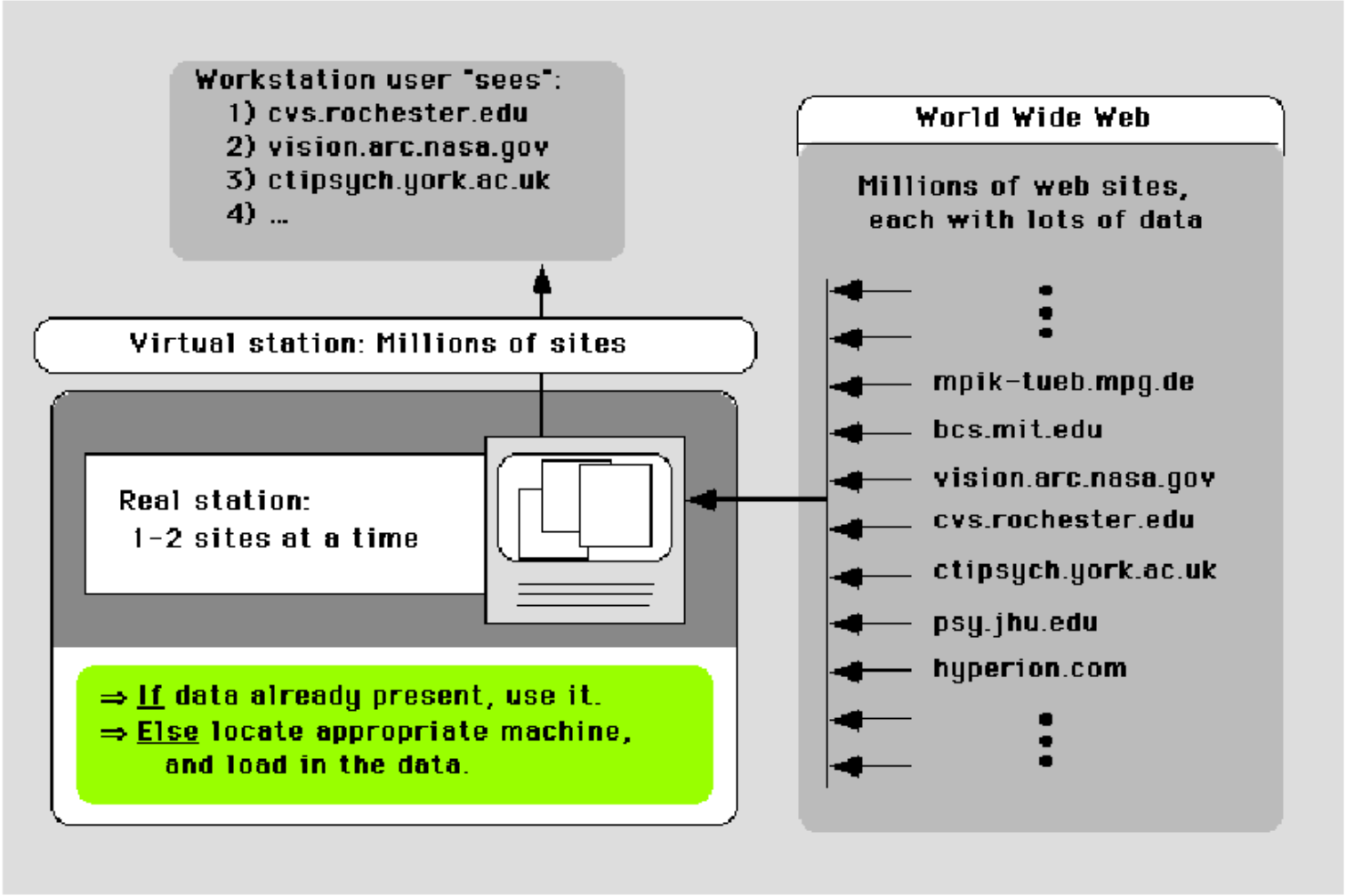


Figure 2

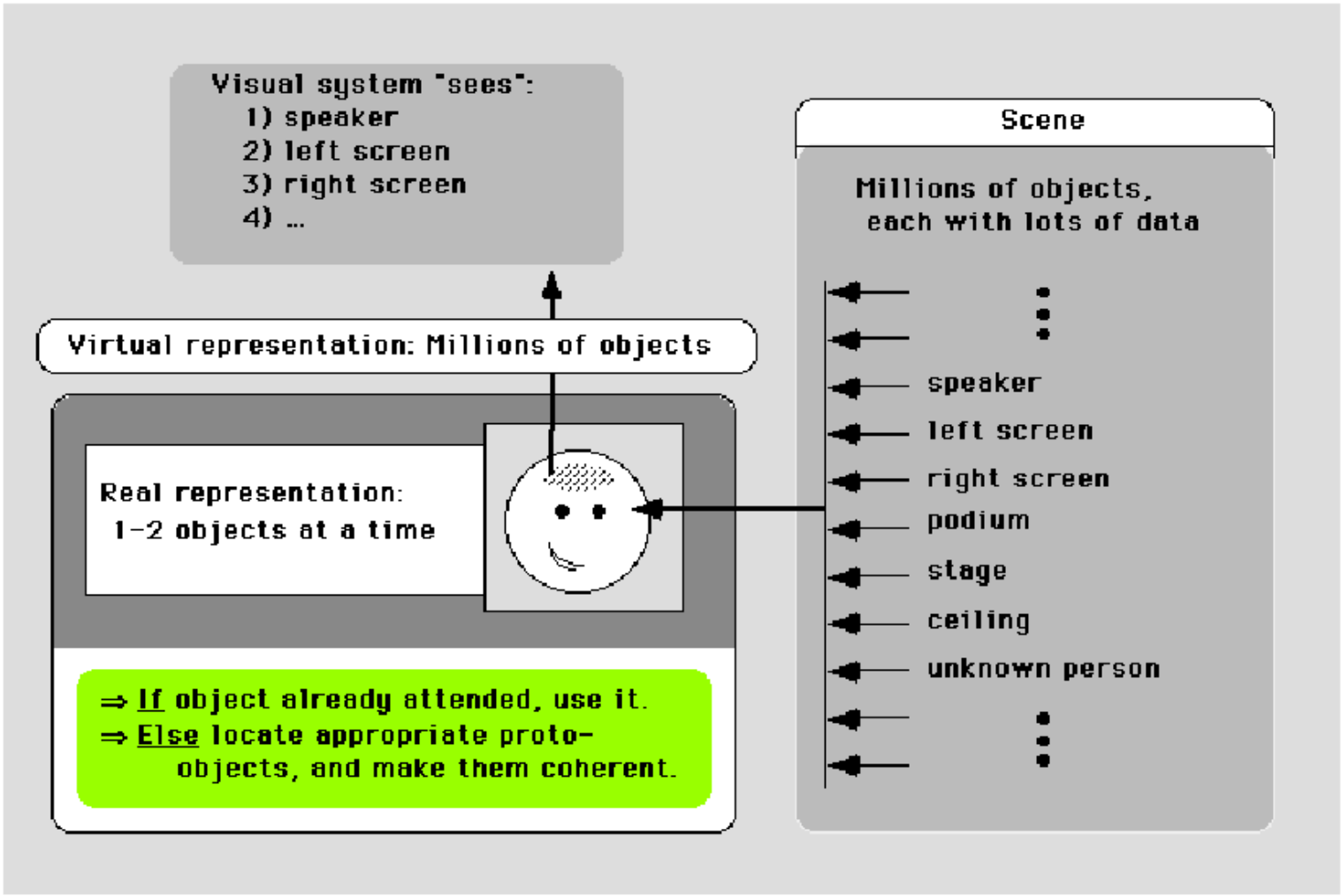
Figure 2. Schematic of a Coherence Field. This structure is composed of three kinds of components: (i) a nexus, corresponding to a single object, (ii) a set of 4-6 proto-objects, corresponding to object parts, and (iii) bidirectional links between the nexus and the proto-objects. Coherence is established when a recurrent flow of information exists between the nexus and its proto-objects, as well as within the nexus itself. Selected information is transmitted up the links to enter into the description of the object. Information from the nexus is also transmitted back down the links to provide stability (and perhaps refinement) to the proto-objects.



(a) Virtual representation: computer network

Figure 3(a)

Figure 3. Virtual Representation. (a) Computer Network. If a limited-capacity workstation can access information from the computer network whenever requested, it will appear to contain all the information from all sites on the network. (b) Human Vision. If a limited-capacity attentional system can access information from the visible scene whenever requested, it will appear to contain all the information from all objects in the visible scene.



(b) Virtual representation: visual system

Figure 3(b)

Figure 3. Virtual Representation. (a) Computer Network. If a limited-capacity workstation can access information from the computer network whenever requested, it will appear to contain all the information from all sites on the network. (b) Human Vision. If a limited-capacity attentional system can access information from the visible scene whenever requested, it will appear to contain all the information from all objects in the visible scene.

Back to accumulated evidence!

Hollingworth et al, 2000 argue against the disintegration of coherent visual representations as soon as attention is withdrawn.

Experiment:

- line drawings of natural scenes
- change one object (target) during a saccadic eye movement away from that object
- instruct subjects to examine scene, and they would later be asked questions about what was in it
- also instruct subjects to monitor for object changes and press a button as soon as a change detected

Hypothesis:

It is known that attention will precede eye movements. So the change is outside the focus of attention. If subjects can notice it, it means that some detailed memory of the object is retained.

Hollingworth et al, 2000

Subjects can see the change (26% correct overall)

Even if they only notice it a long time afterwards, at their next visit of the object

Measures of change detection and eye movement behavior.

	Token change		Control (no change)	
	Consistent	Inconsistent	Consistent	Inconsistent
% correct detection	18.1	35.2	--	--
% false alarms	--	--	4.2	0.0
Gaze duration on target for first entry after change (ms)	753*	744*	419	579

* Misses only.

Figure 1. An example of a stimulus scene. The top panels illustrate a token substitution for a semantically consistent target object (microscope), and the bottom panels illustrate a token substitution for a semantically inconsistent target object (teddy bear). The laboratory scene was paired with bedroom scene in which the teddy bears were consistent and the microscopes inconsistent.

Hollingworth et al

So, these results suggest that

"the online representation of a scene can contain detailed visual information in memory from previously attended objects.

Contrary to the proposal of the attention hypothesis (see Rensink, 2000), the results indicate that visual object representations do not disintegrate upon the withdrawal of attention."

Gist of a Scene

Biederman, 1981:

from very brief exposure to a scene (120ms or less), we can already extract a lot of information about its global structure, its category (indoors, outdoors, etc) and some of its components.

"riding the first spike:" 120ms is the time it takes the first spike to travel from the retina to IT!

Thorpe, van Rullen:

very fast classification (down to 27ms exposure, no mask), e.g., for tasks such as "was there an animal in the scene?"

Gist of a Scene

Oliva & Schyns, Cognitive Psychology, 2000

Investigate effect of color on fast scene perception.

Idea: Rather than looking at the properties of the constituent objects in a given scene, look at the global effect of color on recognition.

Hypothesis:

diagnostic colors (predictive of scene category) will help recognition.

Color & Gist

Figure 1. The middle picture illustrates the application of the operator axis_swap (L*a*b* -> L*b*a*) to the top beach picture: every pixel of the green/red opposition is transposed in blue/yellow and vice versa. Note, for example, that the yellow sand becomes red and that the green sea becomes blue. The bottom picture illustrates the application of the axis inversion operator to the "axis swapped" middle image: each red pixel (e.g., the sand in the middle picture) becomes green and each blue pixel (e.g., the sea in the middle picture) become yellow in the bottom picture. Applications of axis swap and inversion were used to synthesize the abnormally colored scenes of our experiments.

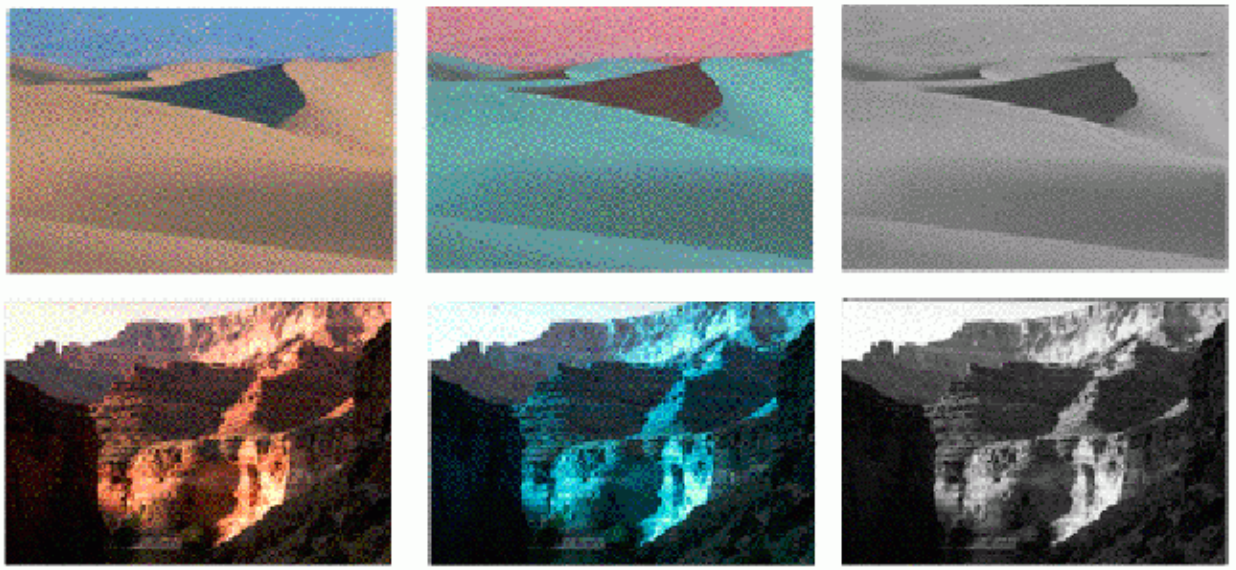


Figure 4. This figure illustrates the three versions of two of the scene pictures used in Experiment 1. The top pictures show normally colored scenes (Norm). The middle pictures show their abnormally (Abn) colored versions and the bottom pictures their luminance-only (Lum) versions.

Color & Gist

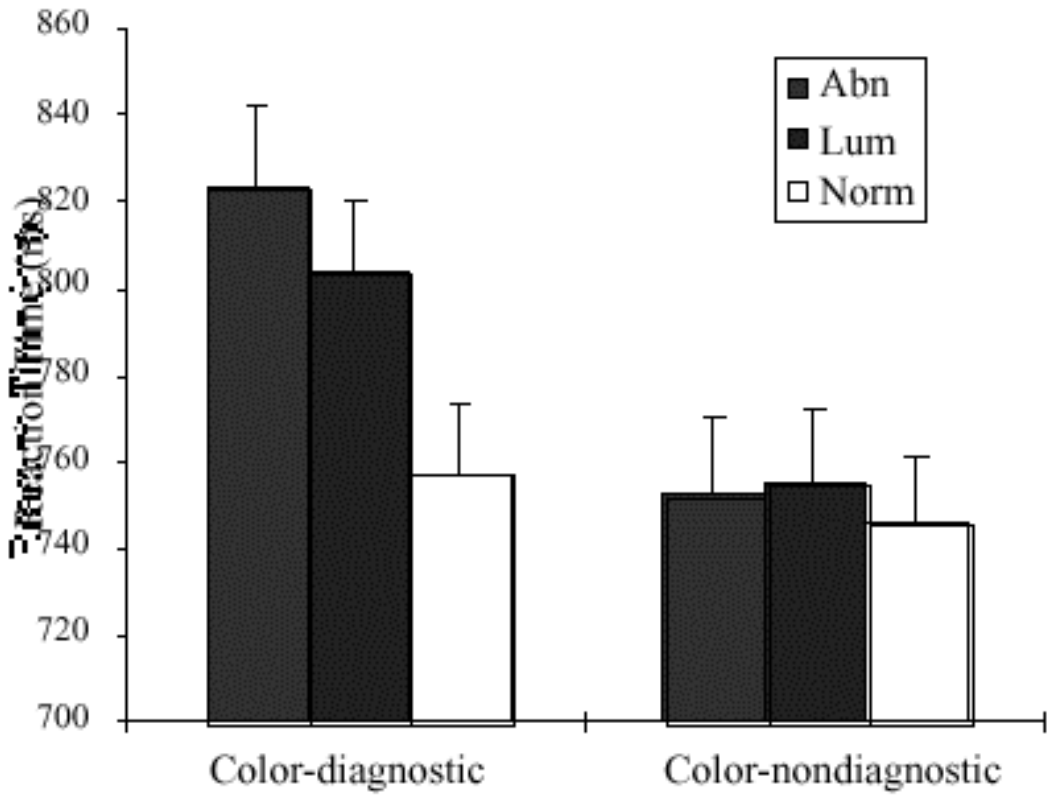


Figure 5. This figure illustrates subjects' naming reaction times in the within-subjects design of Experiment 1. Performance was very similar across luminance-only, normally and abnormally colored versions of the color-nondiagnostic scenes. In contrast, a facilitation of normally colored pictures, and an intereference of abnormally colored pictures were observed for color-diagnostic scenes.

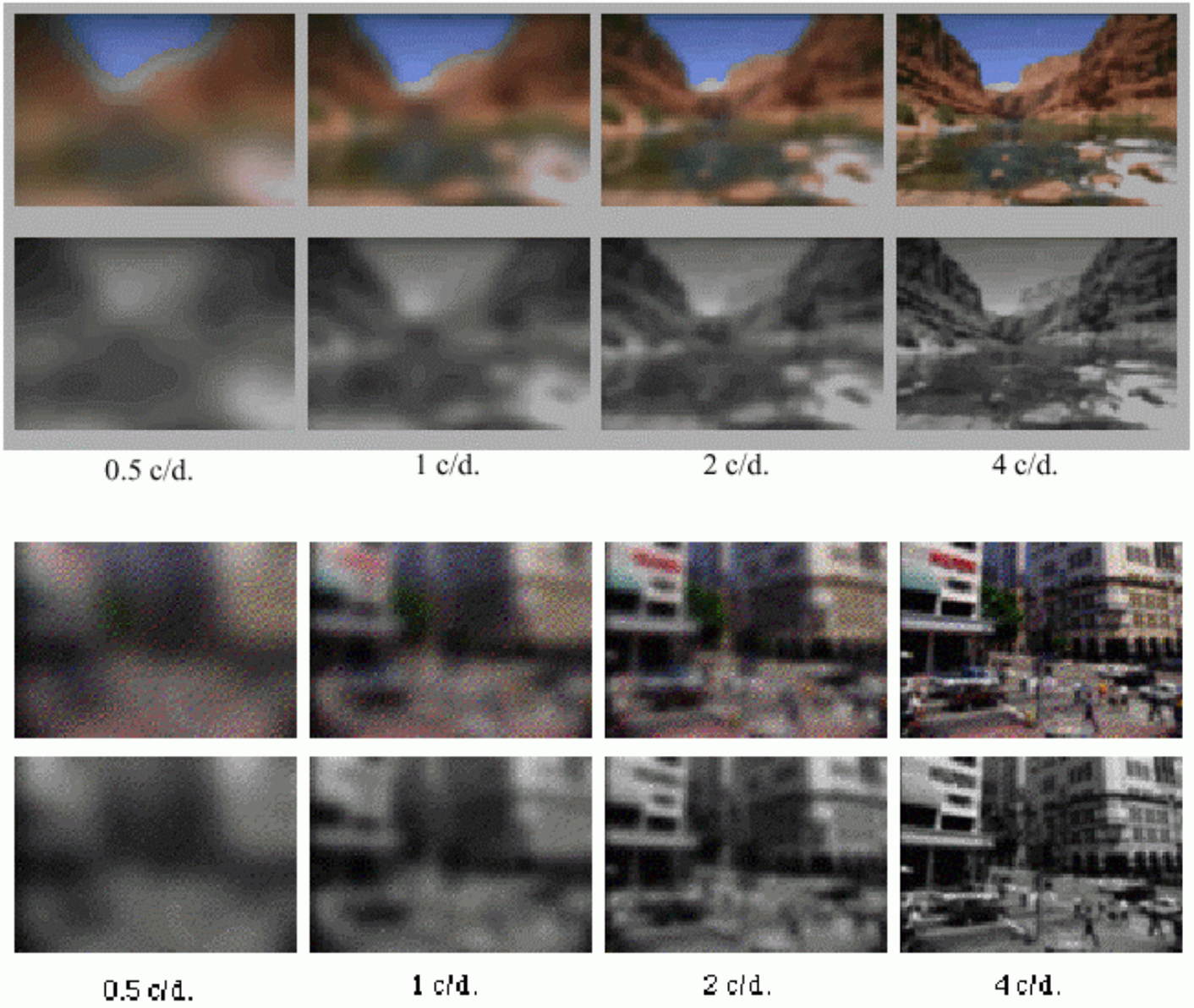


Figure 8. Figure 8 illustrate the different conditions of stimulation used in the free categorization task of Experiment 3 in the colored and luminance-only conditions. The top scene is color-diagnostic whereas the bottom scene is not

Color & Gist

Conclusion from Oliva & Schyns study:

"colored blobs at a coarse spatial scale concur with luminance cues to form the relevant spatial layout that mediates express scene recognition."

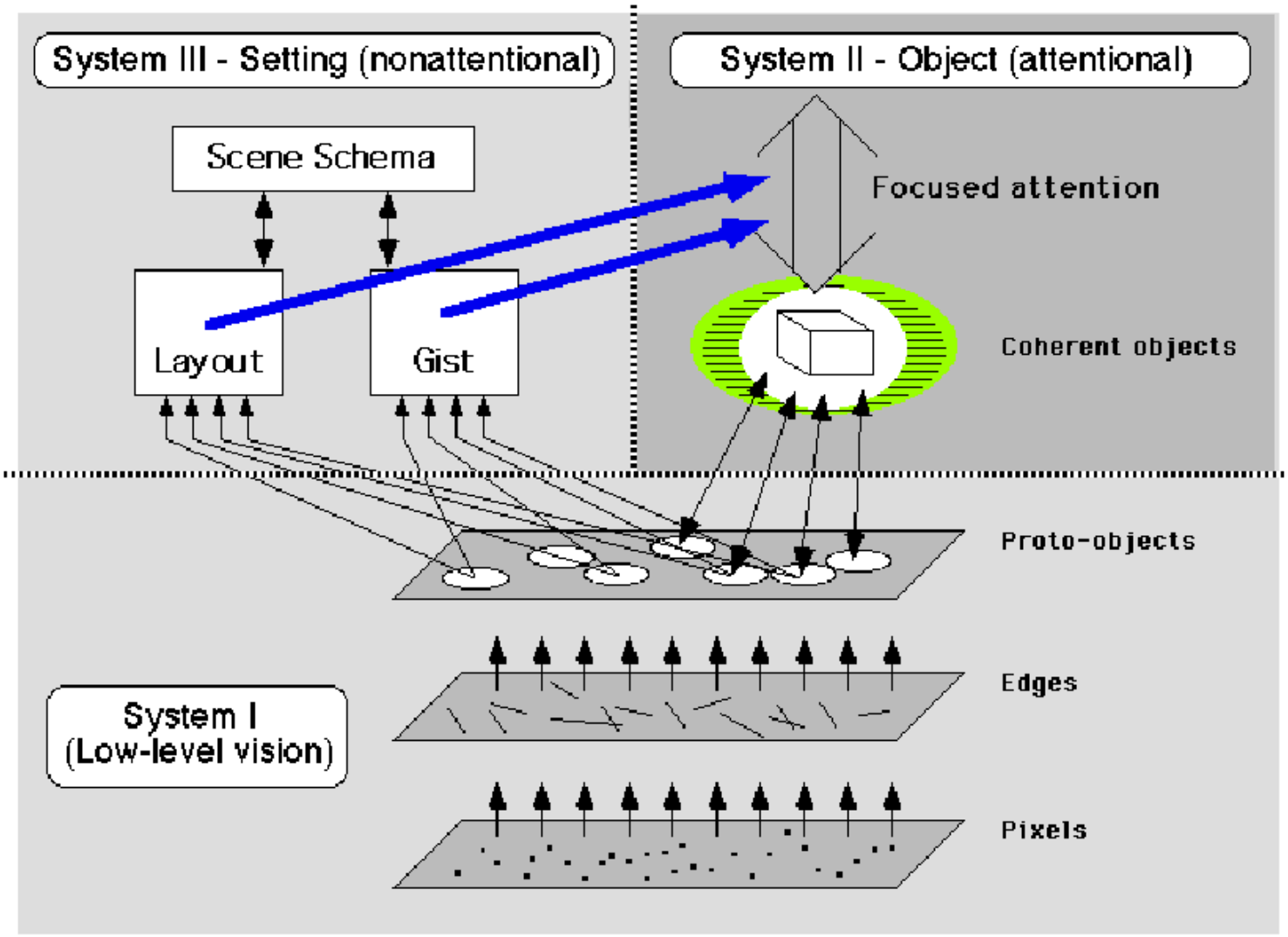


Figure 4

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.

Outlook

It seems unlikely that we perceive scenes by building a progressive buffer and accumulating detailed evidence into it. It would take to much resources and be too complex to use.

Rather, we may only have an illusion of detailed representation, and the availability of our eyes/attention to get the details whenever they are needed. The world as an outside memory.

In addition to attention-based scene analysis, we are able to very rapidly extract the gist of a scene – much faster than we can shift attention around.

This gist may be constructed by fairly simple processes that operate in parallel. It can then be used to prime memory and attention.