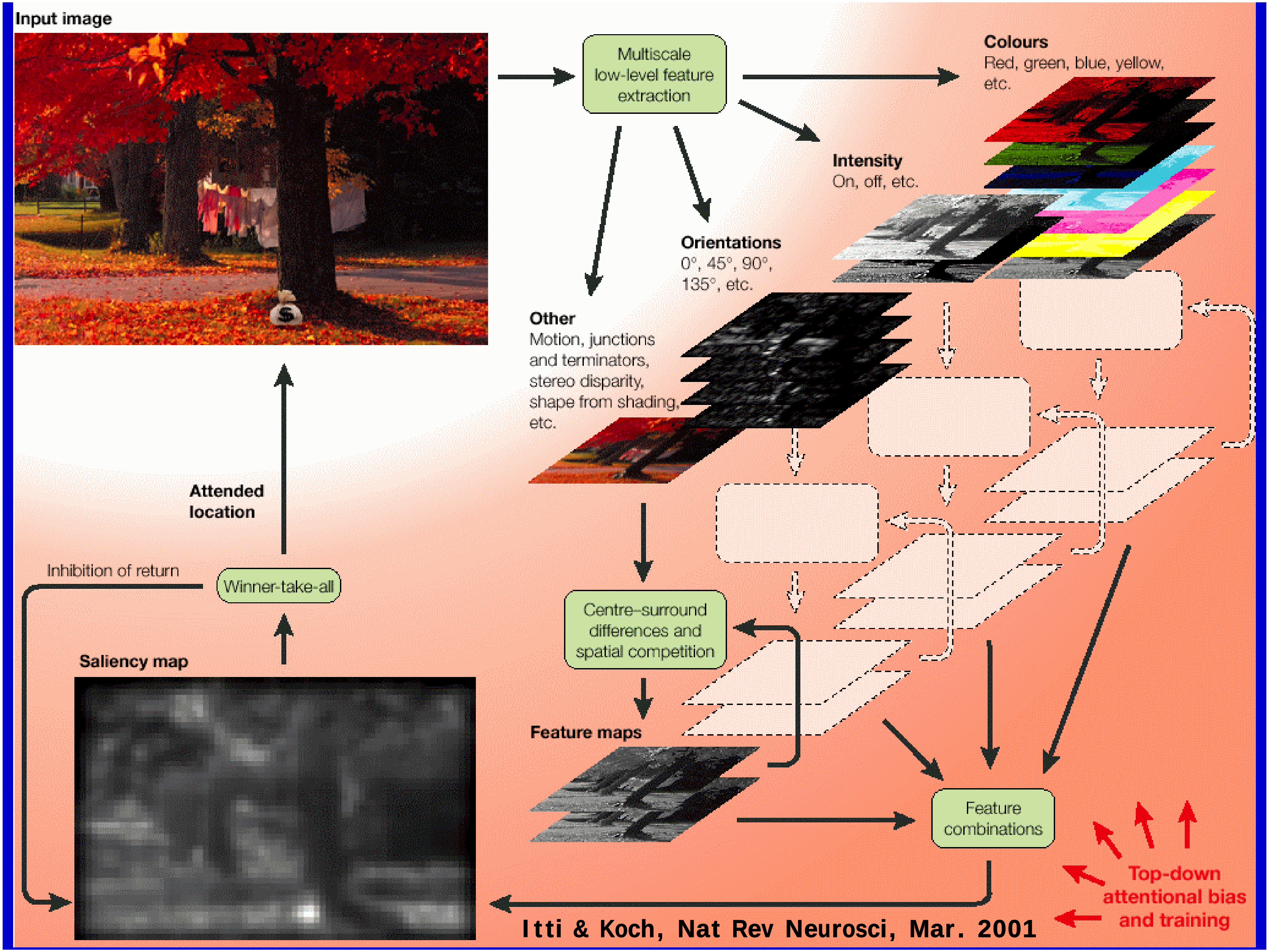
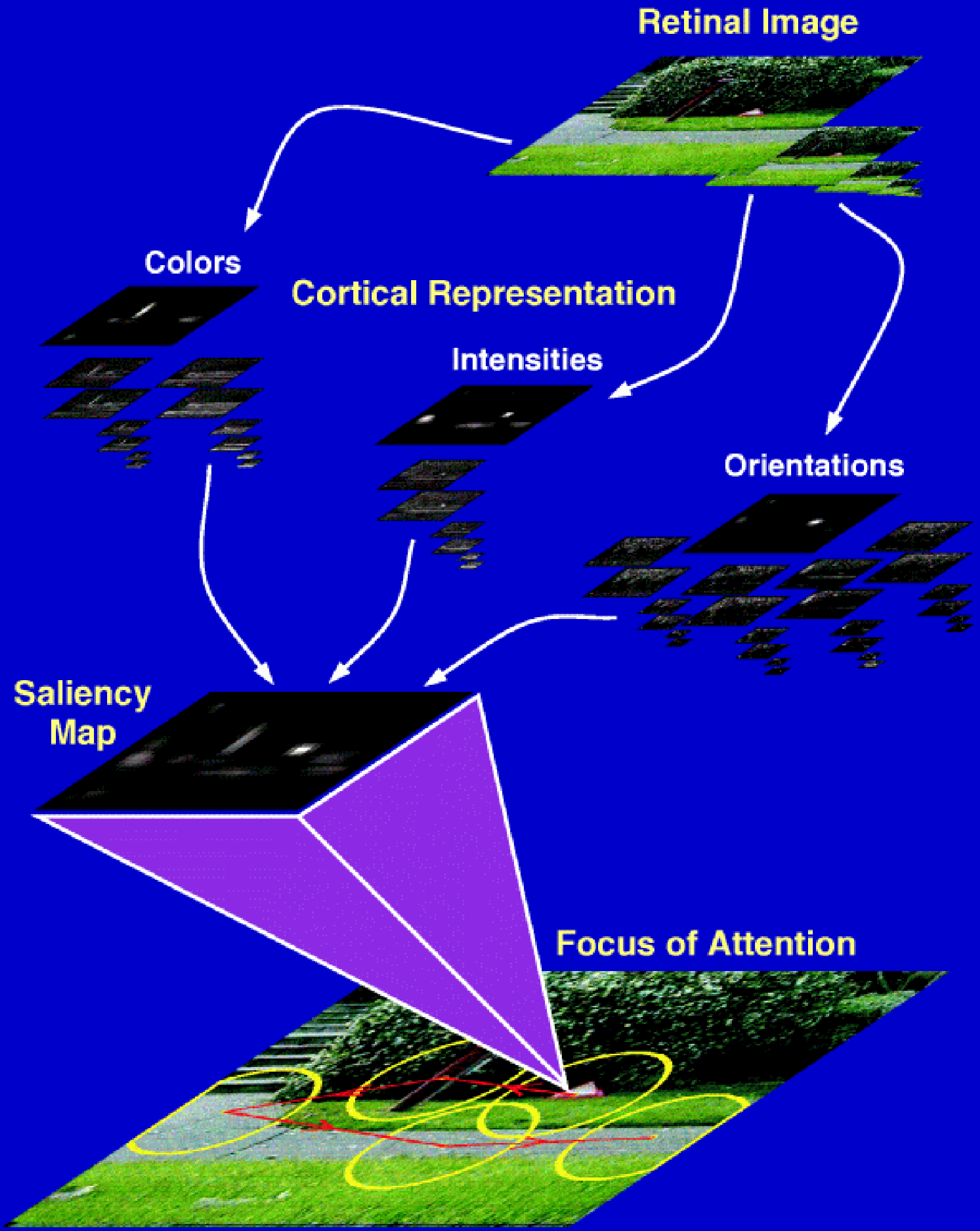
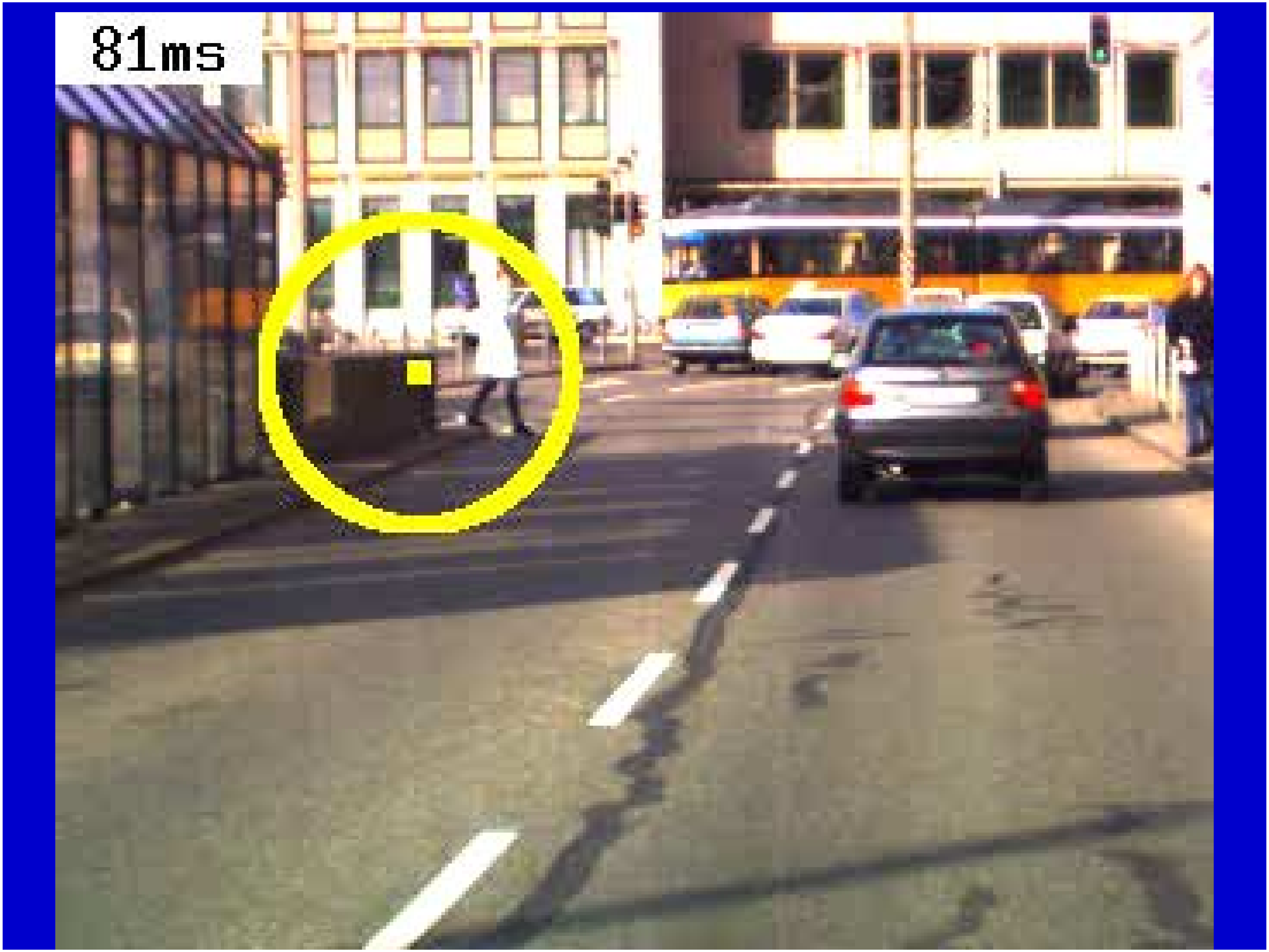

Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Centre–surround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

Itti & Koch, Nat Rev Neurosci, Mar. 2001

3 questions

1) Discuss how the main components of human vision may cooperate to yield scene understanding
2) Discuss why the representation and memorization of scenes is a complex issue
3) Discuss mechanisms by which task and behavioral demands may influence early vision

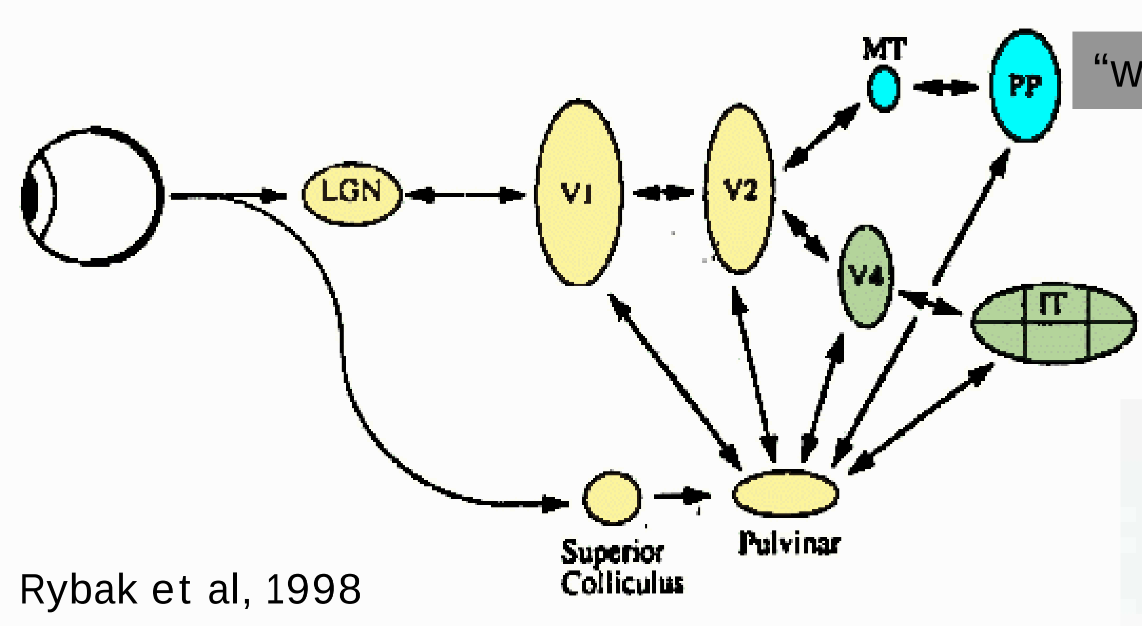

Retinal Image
Colors
Cortical Representation
Intensities
Orientations
Saliency Map
Focus of Attention


81ms

Eye movements
and Character Animation

“Where” and “What” Visual Pathways

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

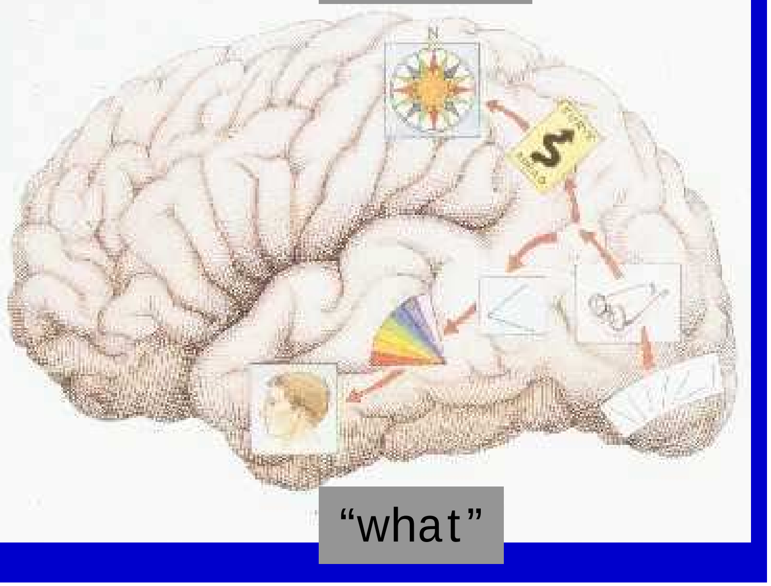


“where”

“what”

Rybak et al, 1998

“where”



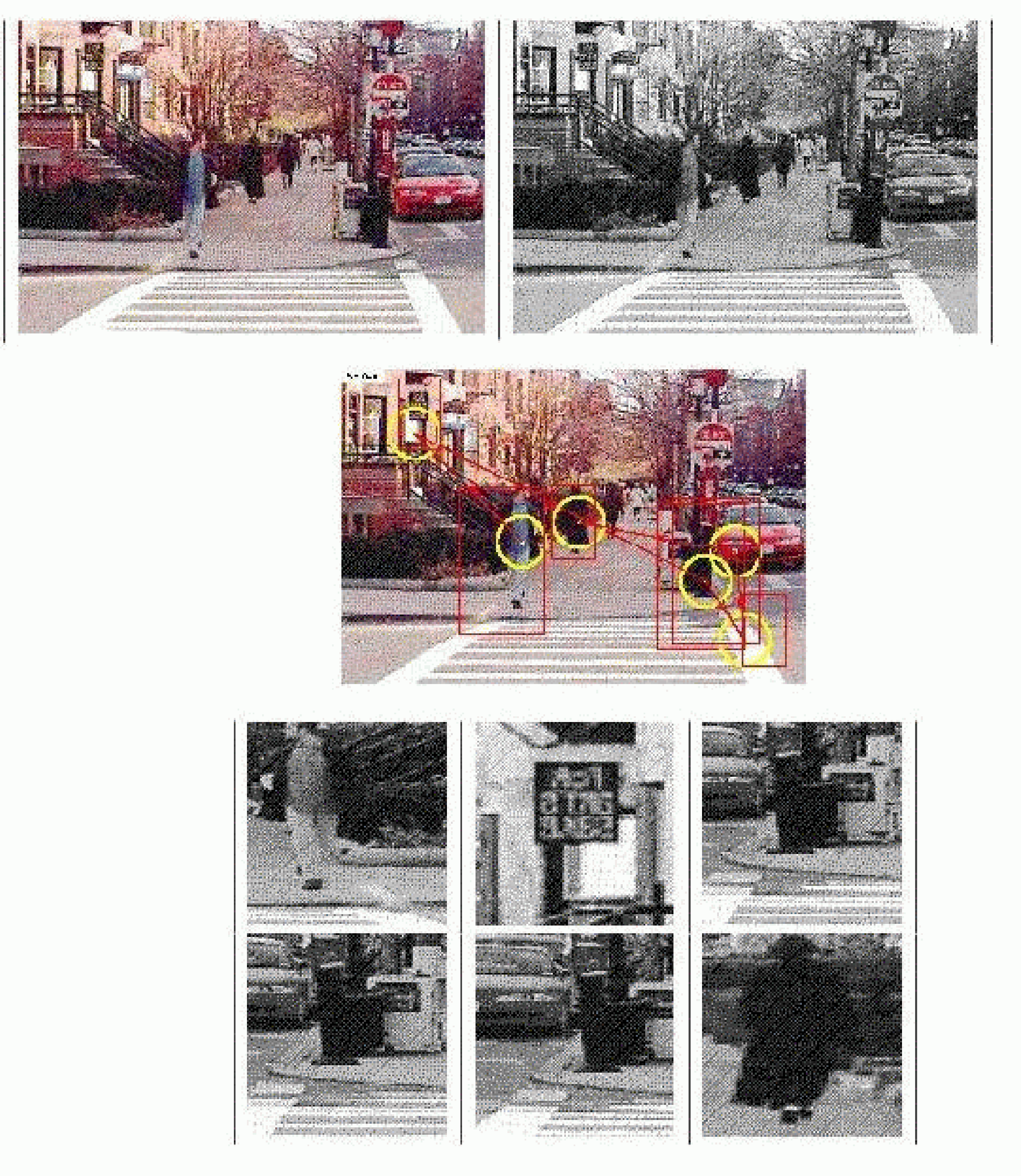
“what”

72 color outdoors images

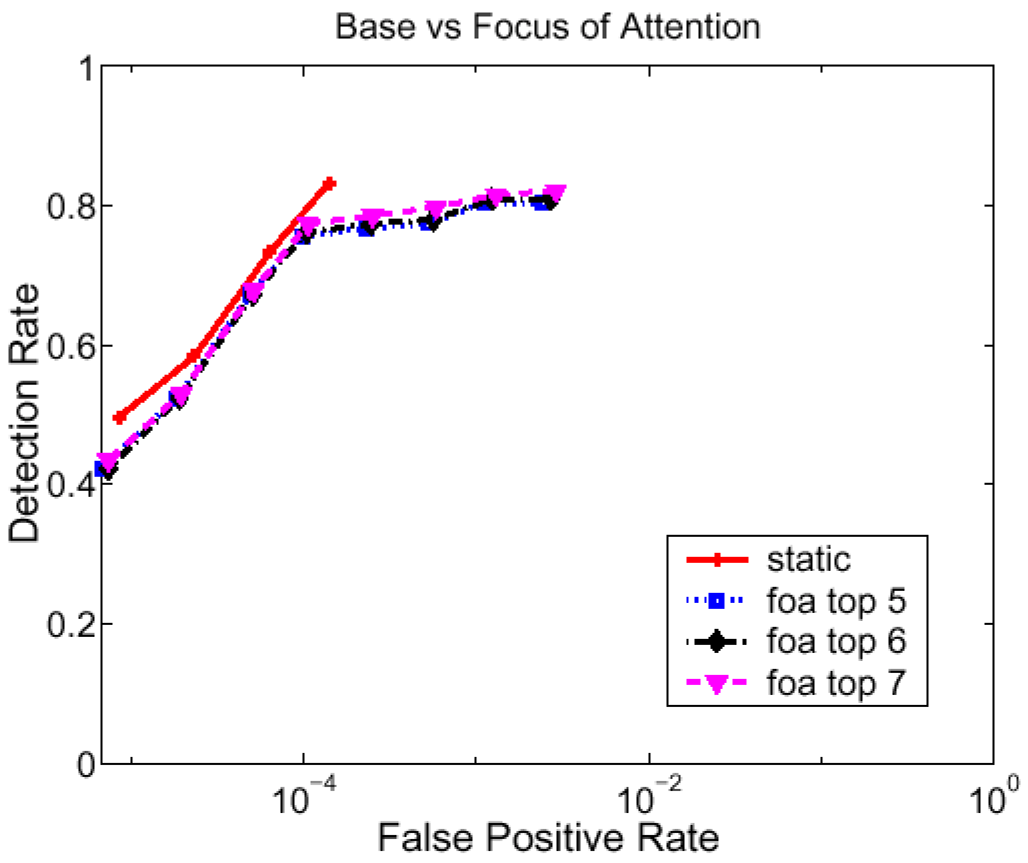
Use attention model to select most salient locations

Crop image around each selected location

Feed cropped sections to recognition model



Combined Where/What Model Performance

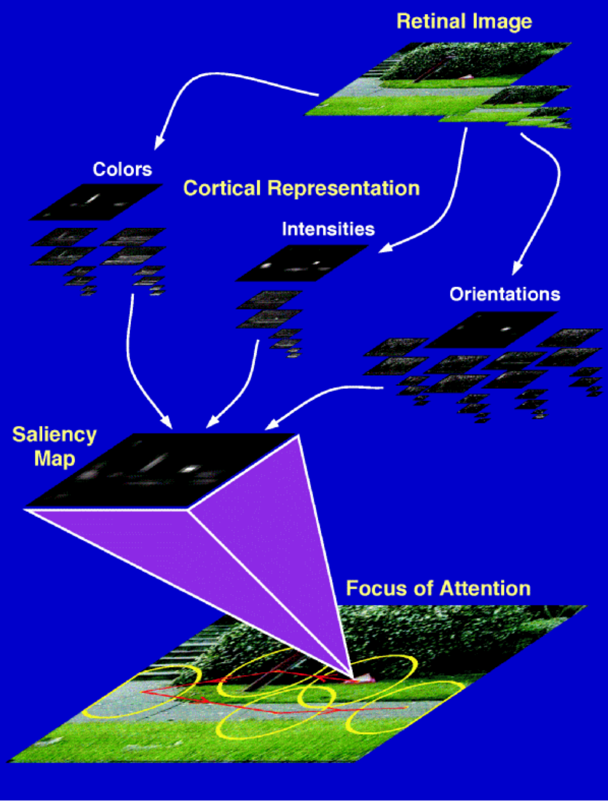


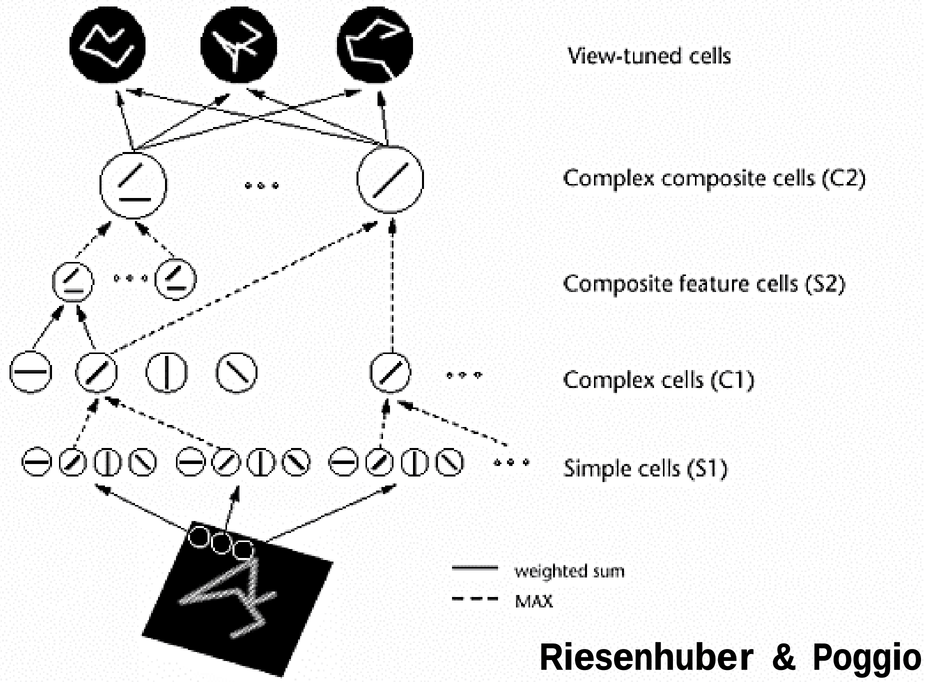
Front-end	Patterns Processed
Base	100.00%
FOA (top 7)	38.97%
FOA (top 6)	35.14%
FOA (top 5)	29.22%

- 3x speed gain
- Overall recognition rate drops <5% (from about 80%)
- No difference between using top 5, 6, or 7 attended locations

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.





Riesenhuber & Poggio, Nat Neurosci, 1999

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

Achieve representation for it that is robust and stable against noise, world motion, and egomotion.

How plausible is that?

3—5 eye movements/sec, that’s 150,000—250,000/day

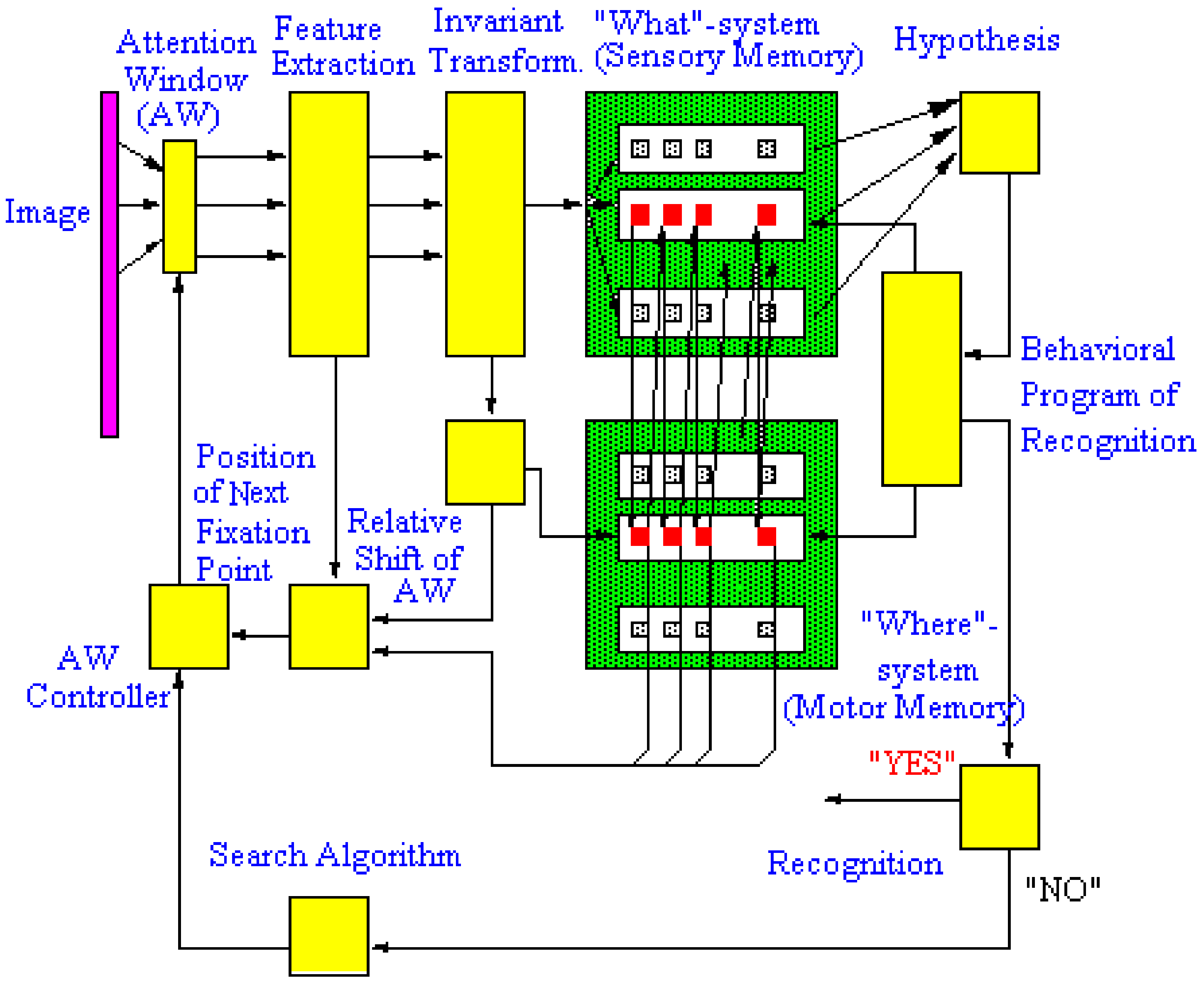
Only central 2deg of retinas (our foveas) carry high-resolution information

Attention helps us bring our foveas onto relevant objects.

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?



Rybak et al, 1998

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 fixation points; 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

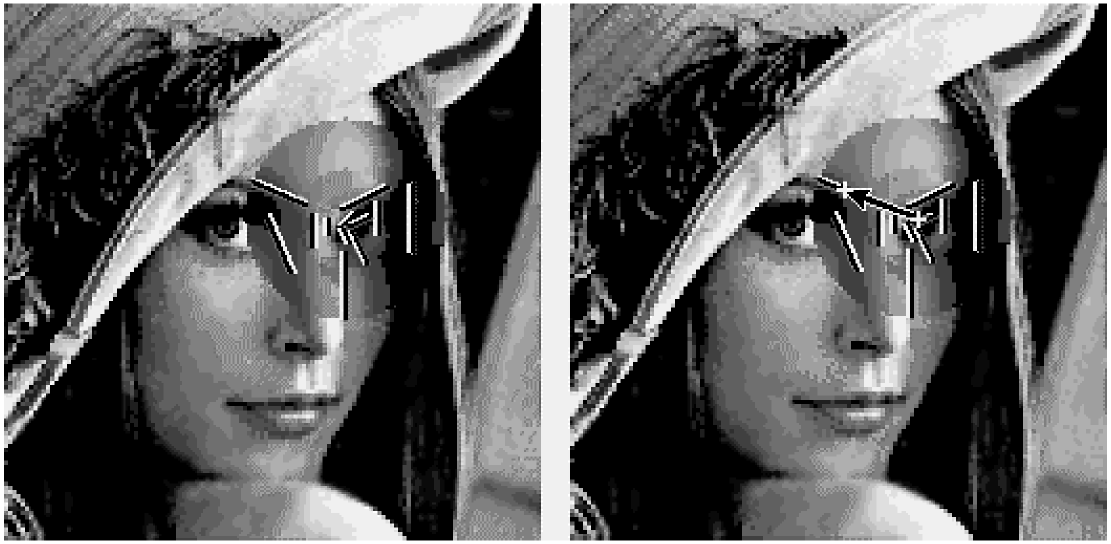


Figure 4. The next fixation point is selected from the set of context points in the current retinal image.
The current and next fixation points are marked by *crosses* (*right*).
Shift to the next fixation point is shown by the *black arrow* (*right*).

Algorithm

Scanpath

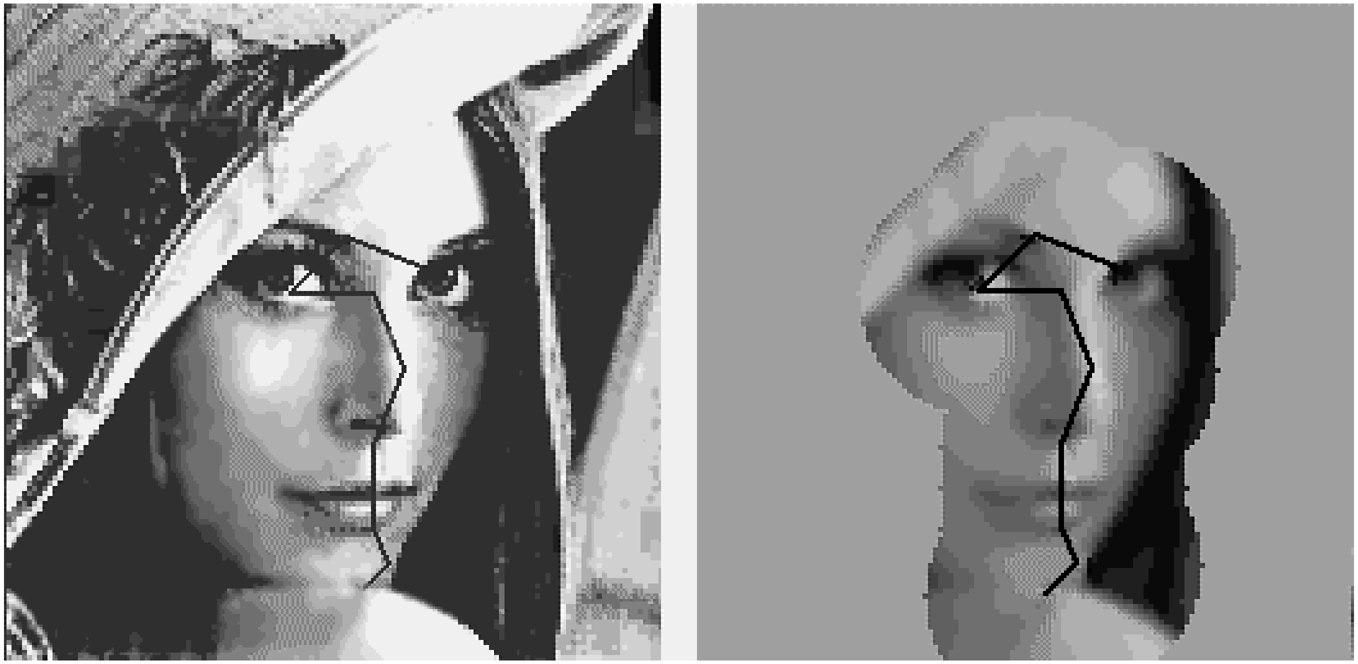
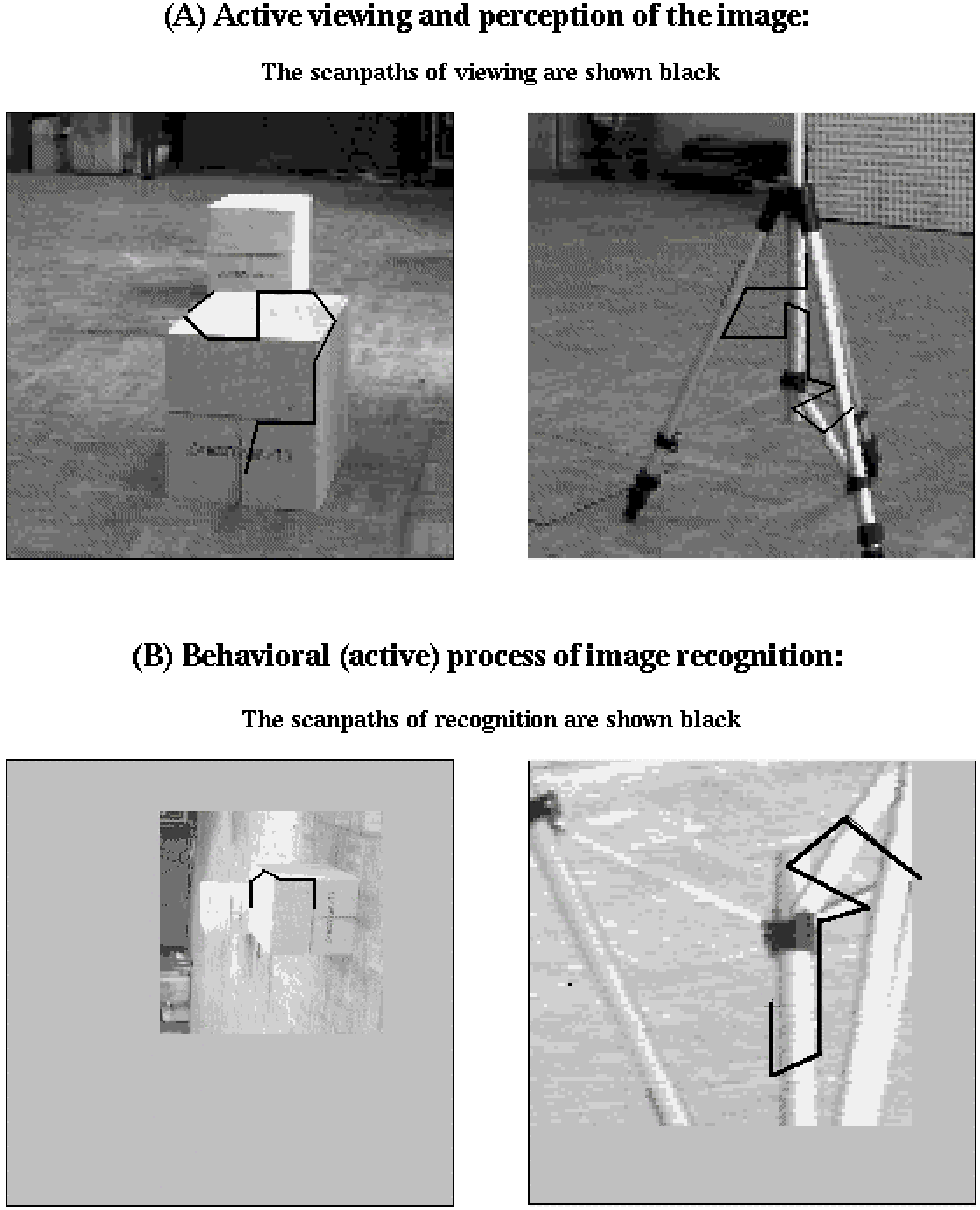


Figure 5. The scanpath of image viewing is shown on background of the initial image (*left*) and on background of the sequence of retinal images along the scanpath (*right*).

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.



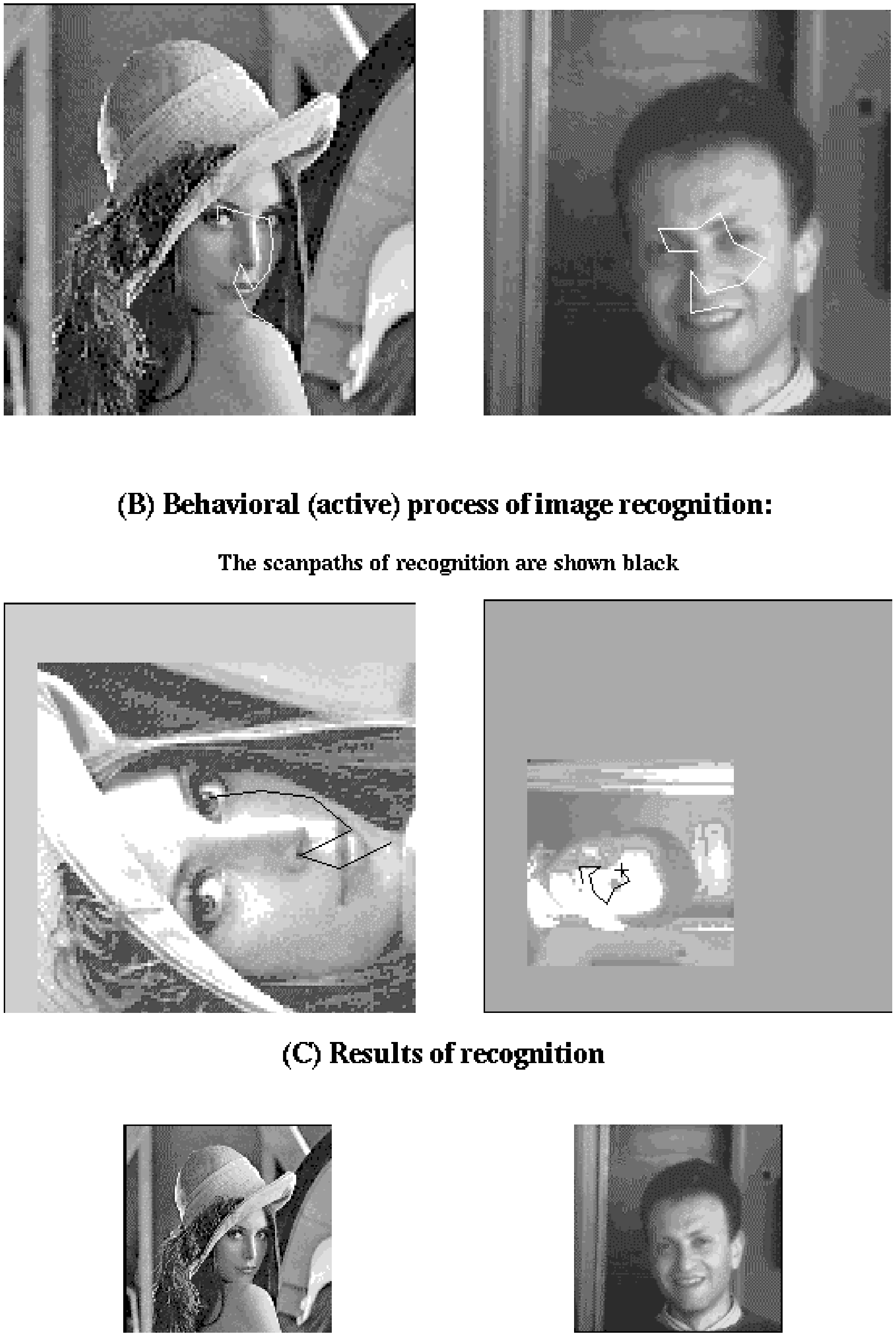
(A) Active viewing and perception of the image:

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.

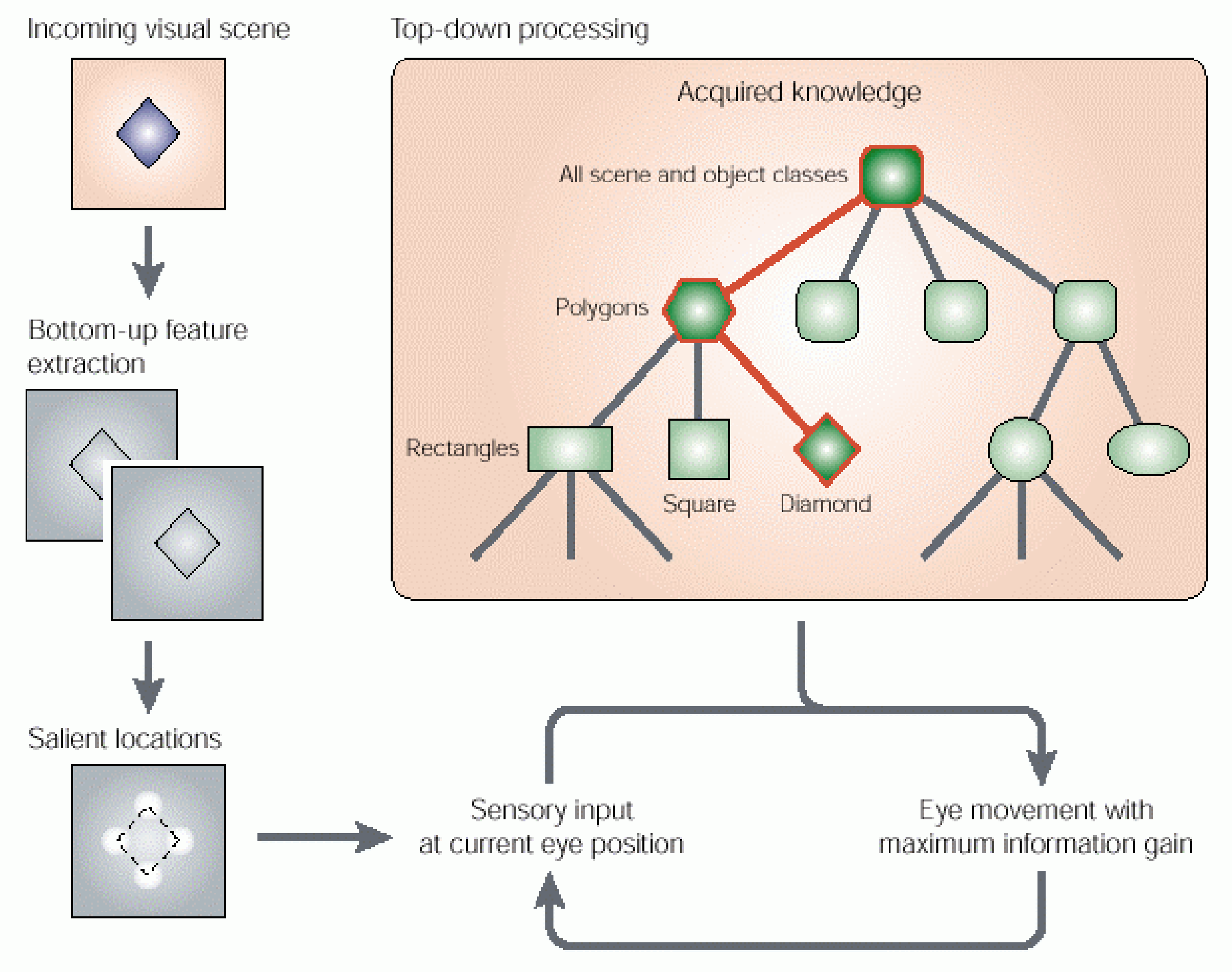


(B) Behavioral (active) process of image recognition:

The scanpaths of recognition are shown black

(C) Results of recognition

Schill et al, JEI, 2001

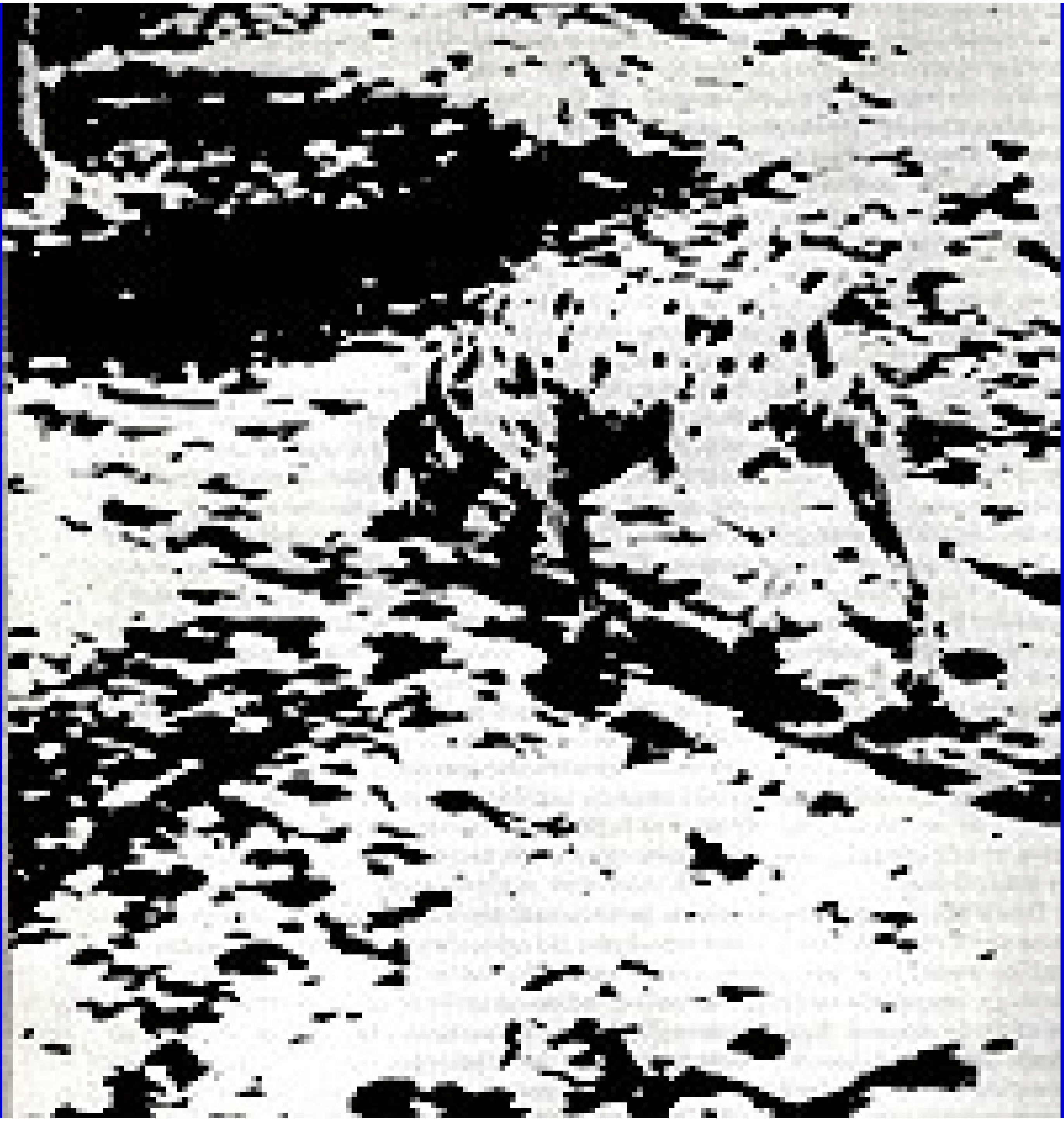

Incoming visual scene
Top-down processing
Acquired knowledge
All scene and object classes
Polygons
Rectangles
Square
Diamond
Bottom-up feature extraction
Salient locations
Sensory input at current eye position
Eye movement with maximum information gain

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.

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!

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?”

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

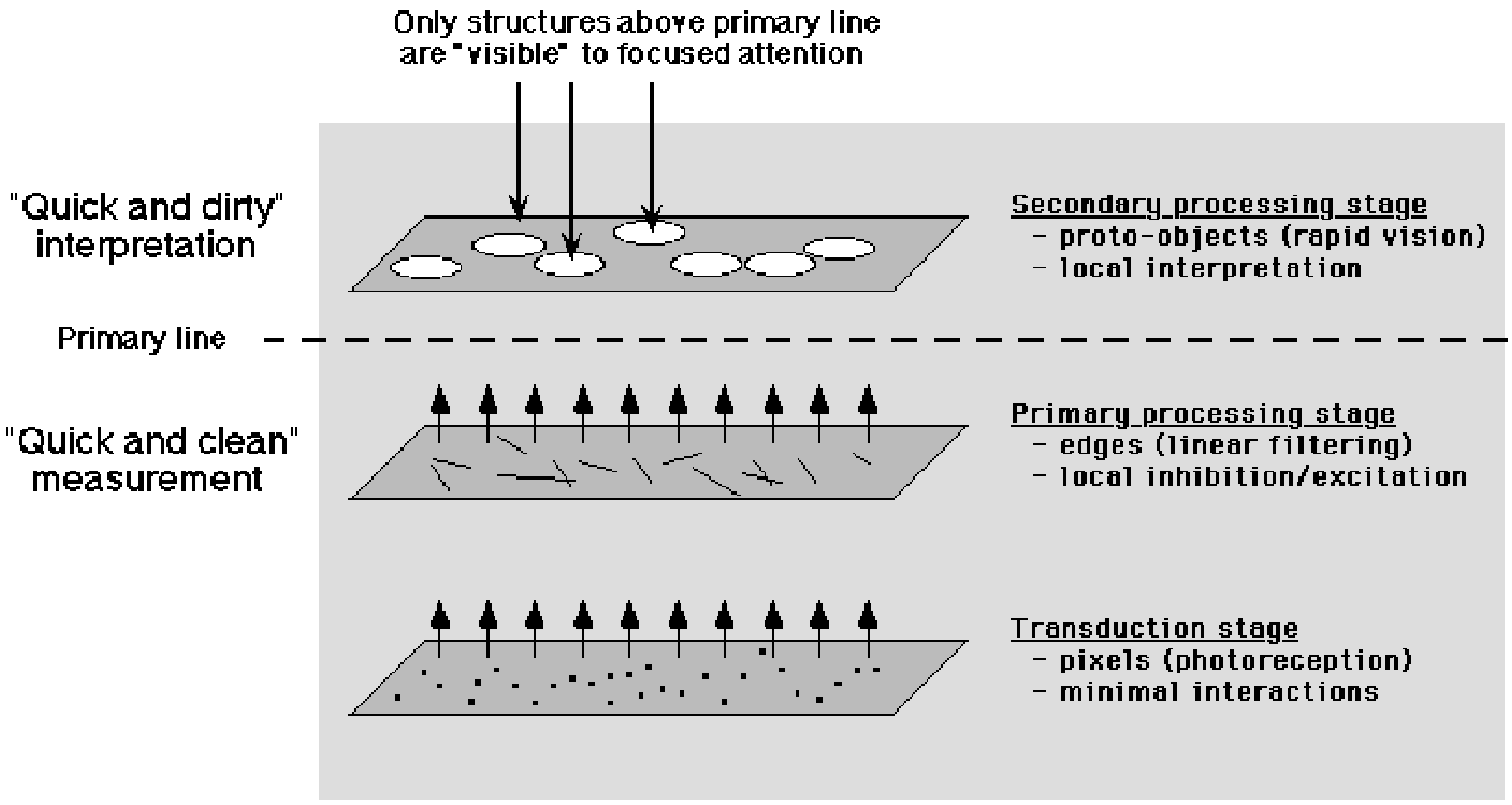

Only structures above primary line
are "visible" to focused attention
"Quick and dirty" interpretation
Secondary processing stage
- proto-objects (rapid vision)
- local interpretation
Primary line
"Quick and clean" measurement
Primary processing stage
- edges (linear filtering)
- local inhibition/excitation
Transduction stage
- pixels (photoreception)
- minimal interactions

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.

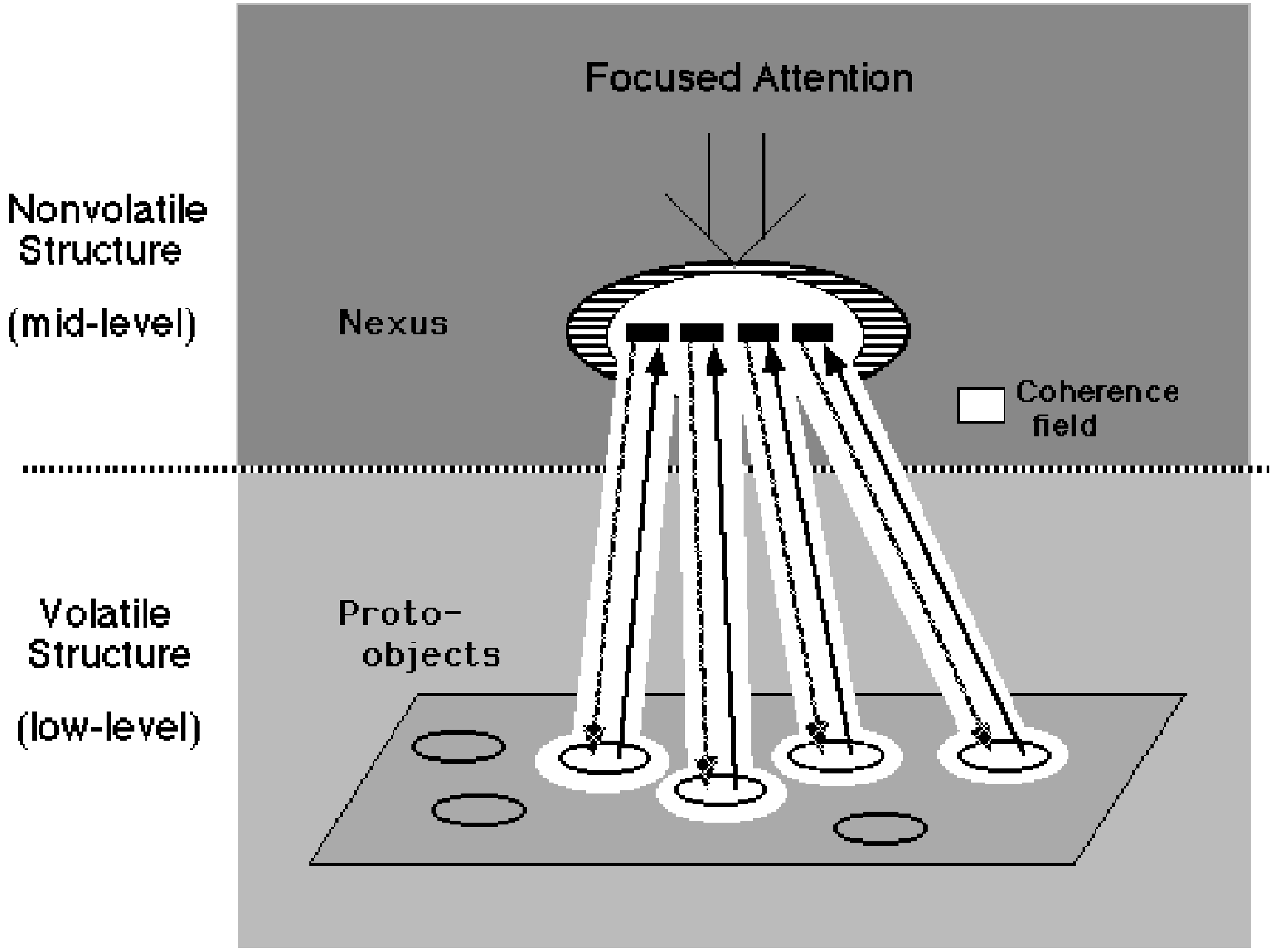
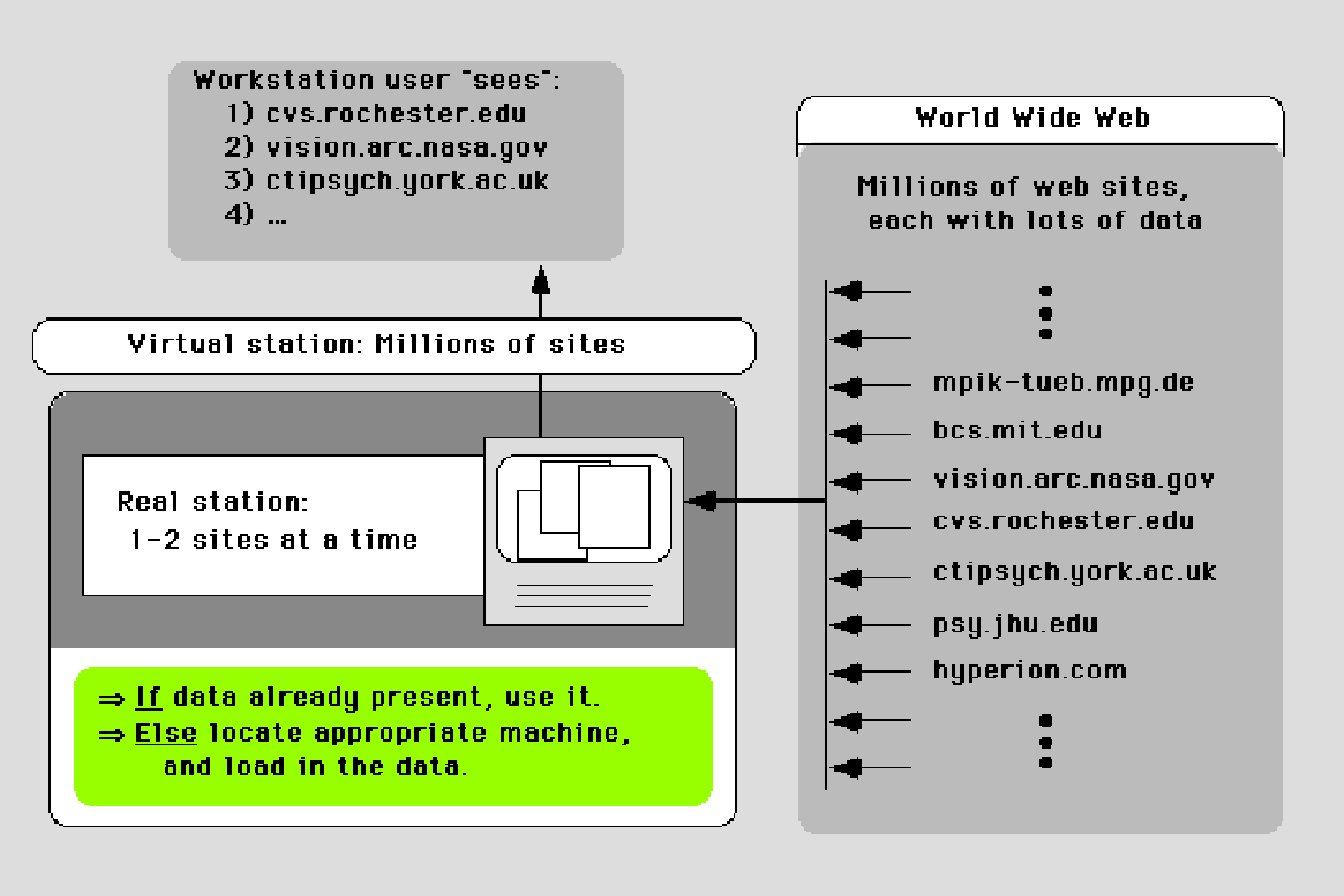


Figure 2

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.



Workstation user "sees":
1) cvs.rochester.edu
2) vision.arc.nasa.gov
3) ctipsych.york.ac.uk
4) ...

World Wide Web

Millions of web sites,
each with lots of data

mpik-tueb.mpg.de
bcs.mit.edu
vision.arc.nasa.gov
cvs.rochester.edu
ctipsych.york.ac.uk
psy.jhu.edu
hyperion.com

Virtual station: Millions of sites

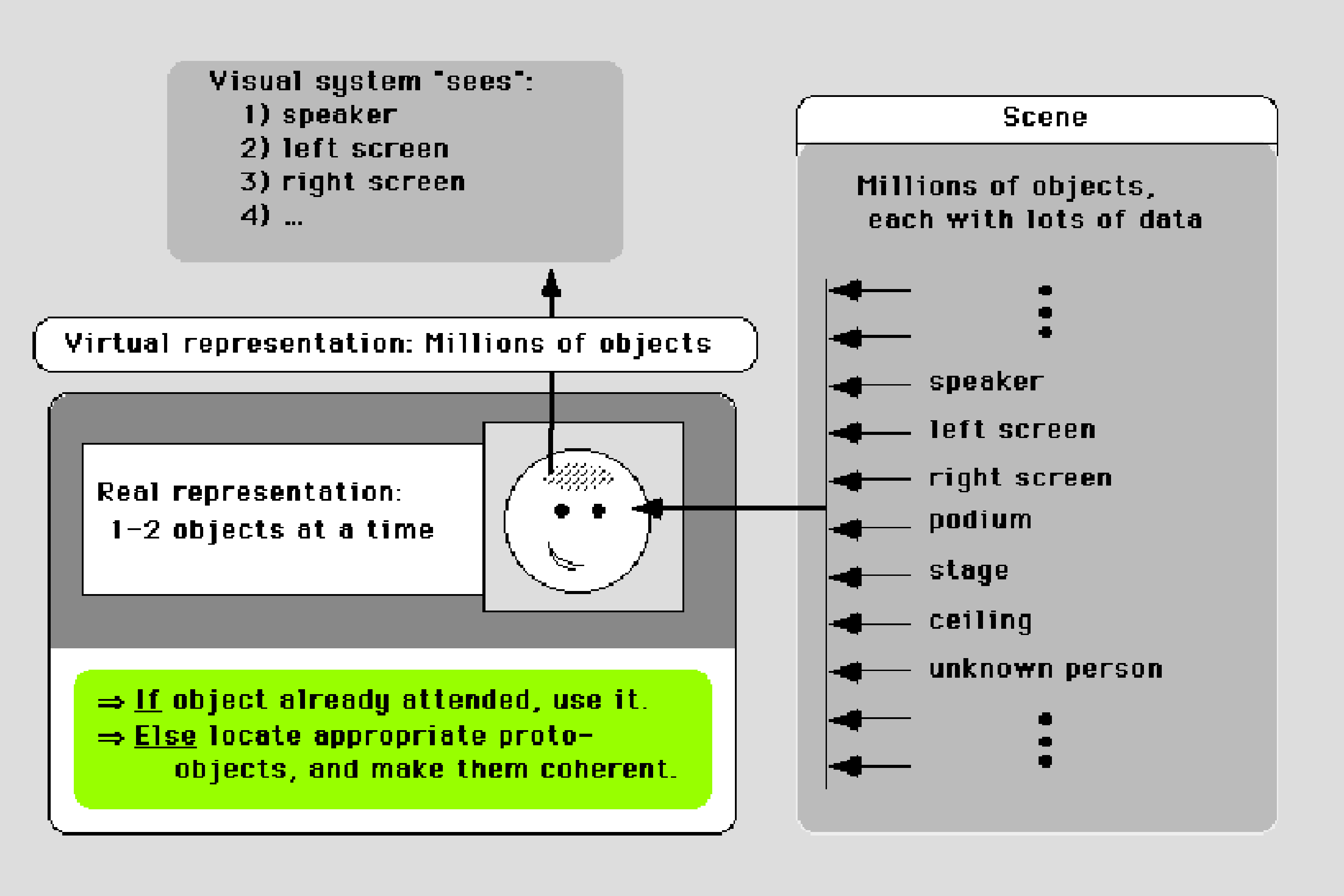
Real station:
1-2 sites at a time

⇒ If data already present, use it.
⇒ Else locate appropriate machine,
and load in the data.

(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.



Visual system "sees":
1) speaker
2) left screen
3) right screen
4) ...

Virtual representation: Millions of objects

Real representation:
1-2 objects at a time

⇒ If object already attended, use it.
⇒ Else locate appropriate proto-objects, and make them coherent.

Scene

Millions of objects, each with lots of data

speaker
left screen
right screen
podium
stage
ceiling
unknown person

(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.

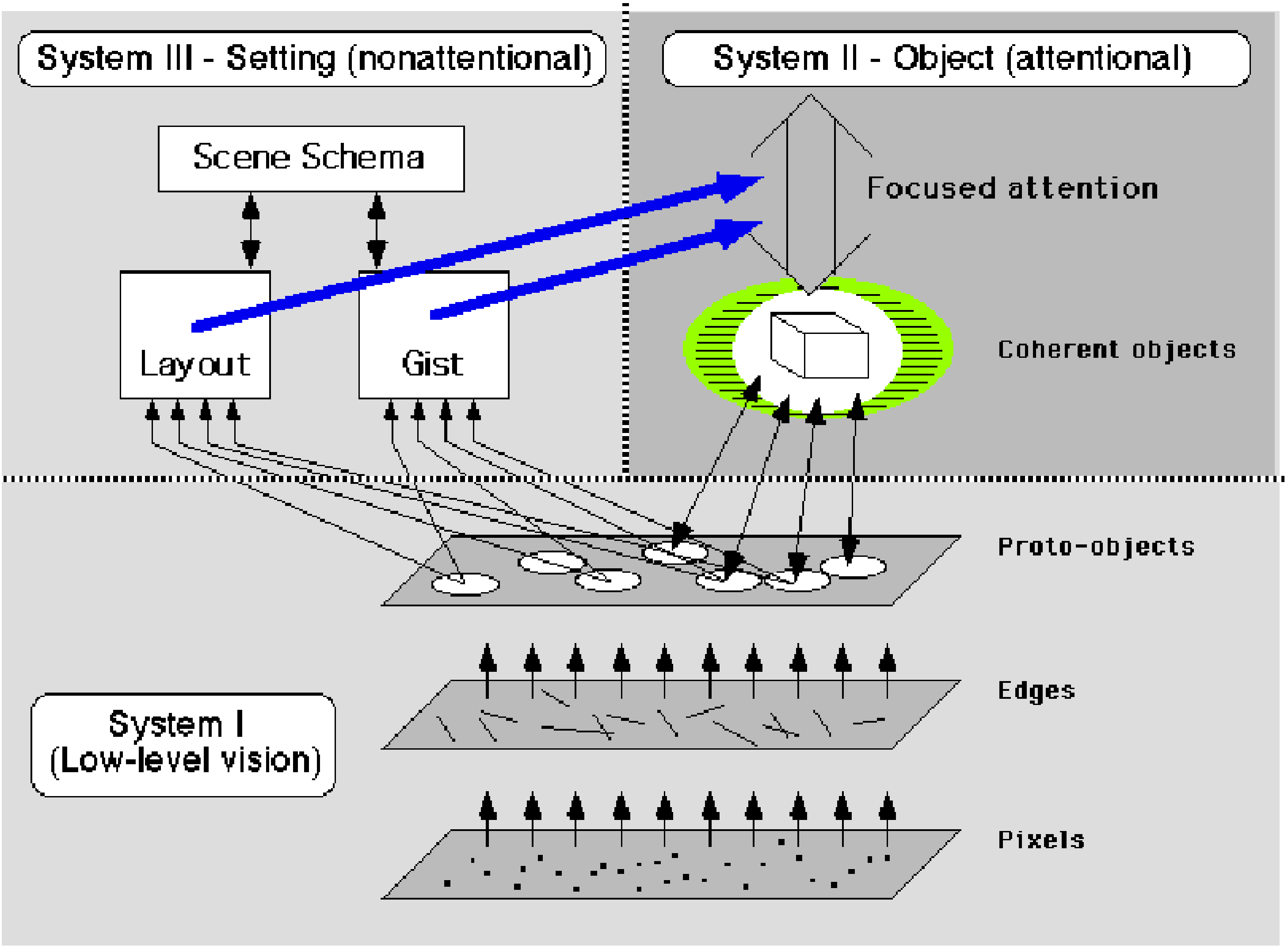


Figure 4

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 on human vision

Unlikely that we perceive scenes by building a progressive buffer and accumulating detailed evidence into it.

- too much resources
- too complex to use.

Rather, we may only have an illusion of detailed representation.

- use eyes/attention to get the details as needed
- the world as an outside memory.

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

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

Eye Movements

1) Free examination

2) estimate material circumstances of family

3) give ages of the people

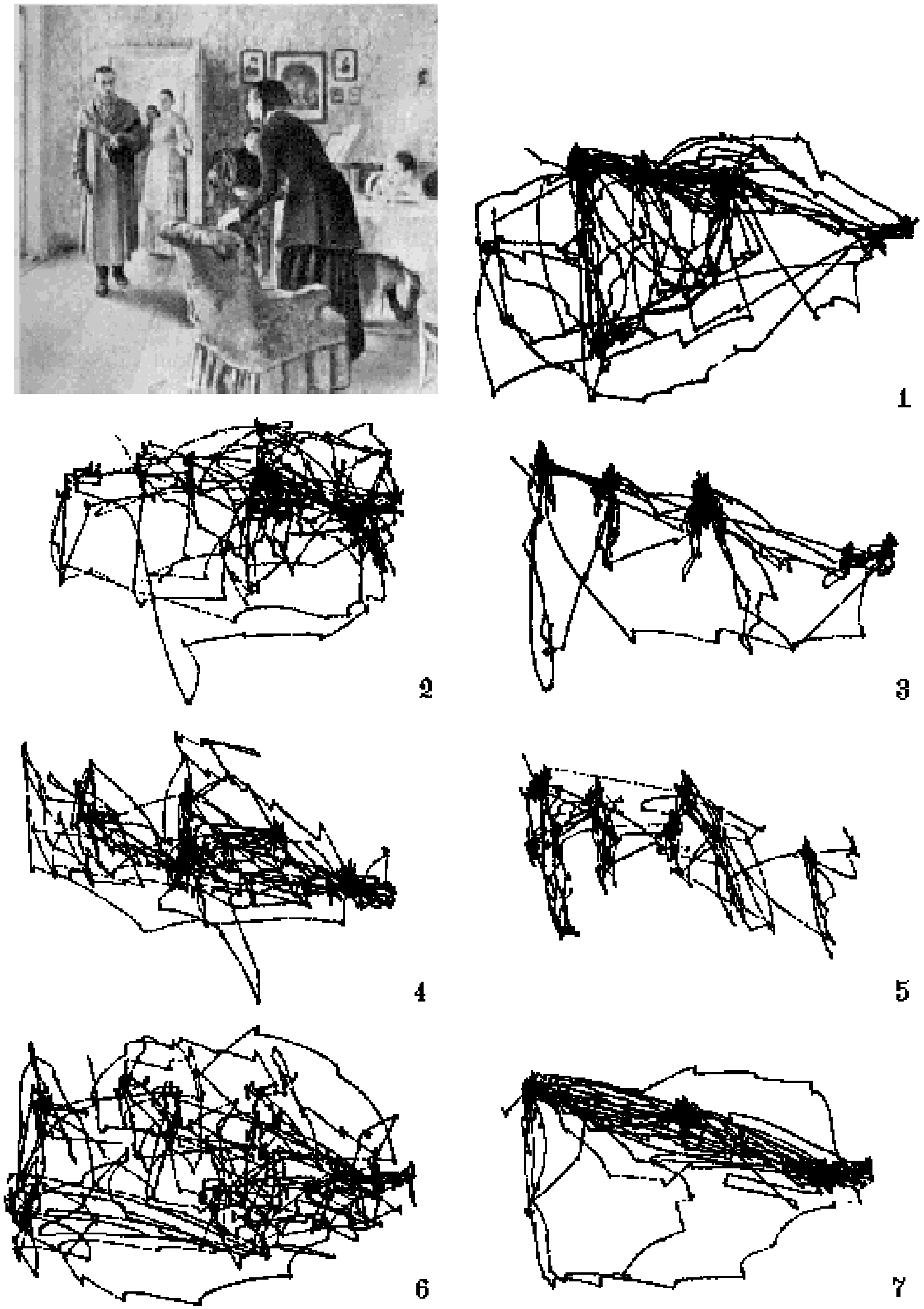
4) surmise what family has been doing before arrival of “unexpected visitor”

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



Goal-directed scene understanding

- **Goal:** develop vision/language-enabled AI system.

Architecture it after the primate brain

- **Test:** ask a question to system about a video clip that it is watching

  e.g., “Who is doing what to whom?”

- **Test:** implement system on mobile robot and give it some instructions

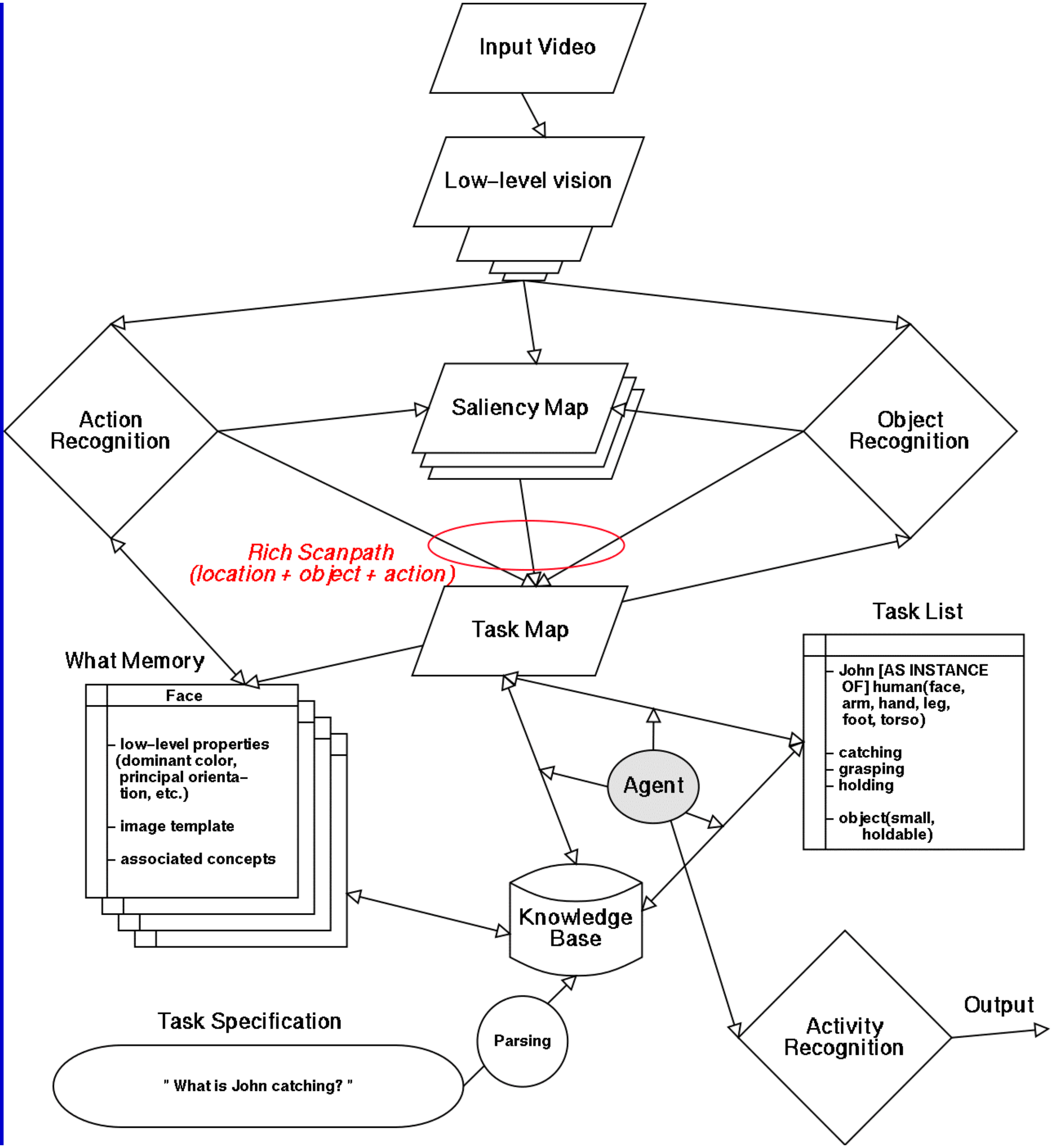
  e.g., “Go to the library”

Example

- **Question:** "who is doing what to whom?"

- **Answer:** "Eric passes, turns around and passes again"

General architecture



Example of operation

- Question: “What is John catching?”
- Video clip: John catching a ball

1) Initially: empty task map and task list

2) Question mapped onto a sentence frame

allows agent to fill some entries in the task list:

- concepts specifically mentioned in the question
- related concepts inferred from KB (ontology)

e.g., task list contains:

“John [AS INSTANCE OF] human(face, arm, hand,
leg, foot, torso)” (all derived from “John”)
“catching, grasping, holding” (derived from “catching”)
“object(small, holdable)” (derived from “what”).

More formally: how do we do it?

- Use ontology to describe categories, objects and relationships:

  Either with unary predicates, e.g., Human(John),

  Or with reified categories, e.g., John $\in$ Humans,

  And with rules that express relationships or properties,

  e.g., $\forall x$ Human(x) $\Rightarrow$ SinglePiece(x) $\wedge$ Mobile(x) $\wedge$ Deformable(x)

- Use ontology to expand concepts to related concepts:

  E.g., parsing question yields "LookFor(catching)"

  Assume a category HandActions and a taxonomy defined by

  catching $\in$ HandActions, grasping $\in$ HandActions, etc.

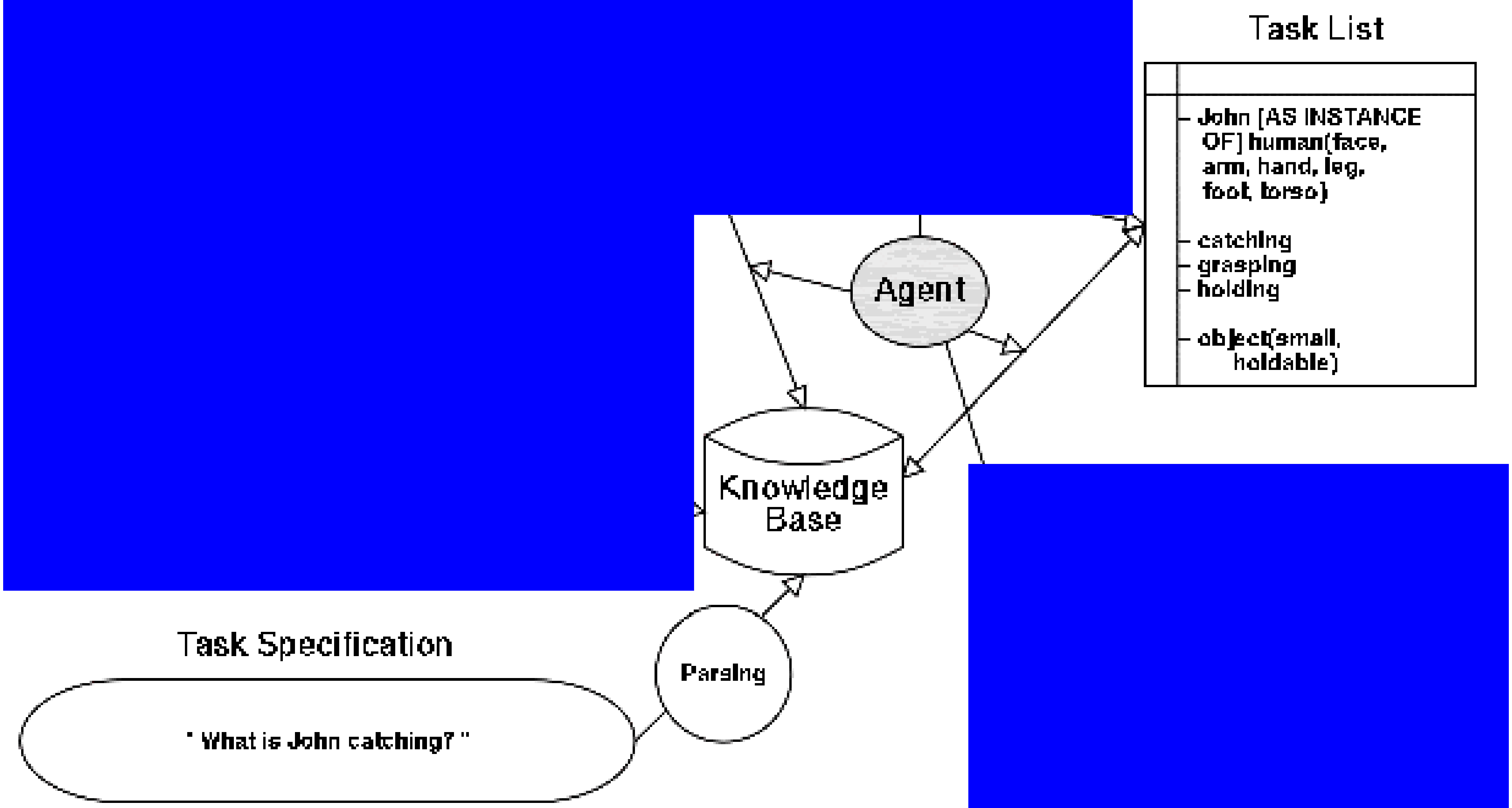
  We can expand "LookFor(catching)" to looking for other actions in the category where catching belongs through a simple expansion rule:

  $\forall a,b,c \quad a \in c \wedge b \in c \wedge \text{LookFor}(a) \Rightarrow \text{LookFor}(b)$

More formally: how do we do it?

- Use composite objects to describe structure and parts:

$$
\begin{aligned}
&\forall h\ \text{Human}(h) \Rightarrow \exists\ f, la, ra, lh, rh, ll, rl, lf, rf, t\\
&\qquad \text{Face}(f) \wedge \text{Arm}(la) \wedge \text{Arm}(ra) \wedge \text{Hand}(lh) \wedge \text{Hand}(rh) \wedge\\
&\qquad\qquad \text{Leg}(ll) \wedge \text{Leg}(rl) \wedge \text{Foot}(lf) \wedge \text{Foot}(rf) \wedge \text{Torso}(t) \wedge\\
&\qquad \text{PartOf}(f, h) \wedge \text{PartOf}(la, h) \wedge \text{PartOf}(ra, h) \wedge \text{PartOf}(lh, h) \wedge\\
&\qquad\qquad \text{PartOf}(rh, h) \wedge \text{PartOf}(ll, h) \wedge \text{PartOf}(rl, h) \wedge \text{PartOf}(lf, h) \wedge\\
&\qquad\qquad \text{PartOf}(rf, h) \wedge \text{PartOf}(t, h) \wedge\\
&\qquad \text{Attached}(f, t) \wedge \text{Attached}(la, b) \wedge \text{Attached}(ra, b) \wedge \text{Attached}(ll, b) \wedge\\
&\qquad\qquad \text{Attached}(rl, t) \wedge \text{Attached}(lh, la) \wedge \text{Attached}(rh, ra) \wedge\\
&\qquad\qquad \text{Attached}(lf, ll) \wedge \text{Attached}(rf, rl) \wedge \text{Attached}(rh, ra) \wedge\\
&\qquad la \neq ra \wedge lh \neq rh \wedge ll \neq rl \wedge lf \neq rf \wedge\\
&\qquad \forall x\ \text{Leg}(x) \wedge \text{PartOf}(x, a) \Rightarrow (x = ll \vee x = rl) \wedge \quad \text{[etc...]}
\end{aligned}
$$



Task List

- John [AS INSTANCE OF] human(face, arm, hand, leg, foot, torso)
- catching
- grasping
- holding
- object(small, holdable)

Agent

Knowledge Base

Parsing

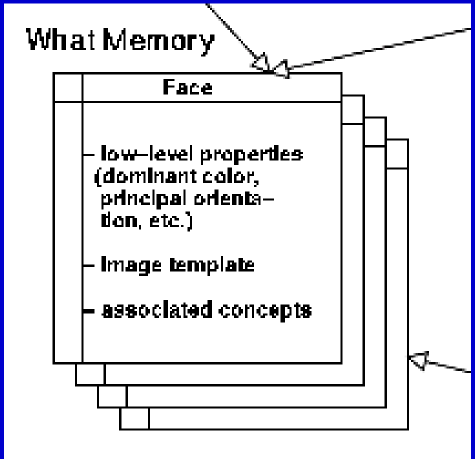
Task Specification

" What is John catching? "

3) Task list creates top-down biasing signals onto vision, by associating concepts in task list to low-level image features in “what memory”

e.g., “human” => look for strong vertically-oriented features

“catching” => look for some type of motion


What Memory
Face
- low-level properties (dominant color, principal orientation, etc.)
- Image template
- associated concepts

In more complex scenarios, not only low-level visual features, but also feature interactions, spatial location, and spatial scale and resolution may thus be biased top-down.

More formally: how do we do it?

- Use measures to quantify low-level visual features and weights:

e.g., describing the color of a face:

$\forall f\ \mathrm{Face}(f) \Rightarrow$

$\mathrm{Red}(f) = \mathrm{Fweight}(0.8) \wedge \mathrm{Green}(f) = \mathrm{Fweight}(0.5) \wedge \mathrm{Blue}(f) = \mathrm{Fweight}(0.5)$

[or use predicates similar to those for intervals to express ranges of feature weights]

e.g., recognizing a face by measuring how well it matches a template:

$\forall f\ \mathrm{RMSdistance}(f, \mathrm{FaceTemplate}) < \mathrm{Score}(0.1) \Rightarrow \mathrm{Face}(f)$

e.g., biasing the visual system to look for face color:

$\forall f\ \mathrm{Face}(f) \wedge \mathrm{LookFor}(f) \Rightarrow \mathrm{RedWeight} = \mathrm{Red}(f) \wedge \mathrm{GreenWeight} = \mathrm{Green}(f) \wedge$

$\mathrm{BlueWeight} = \mathrm{Blue}(f)$

[may eliminate Face(f) if Red(), Green() and Blue() defined for all objects we might look for]

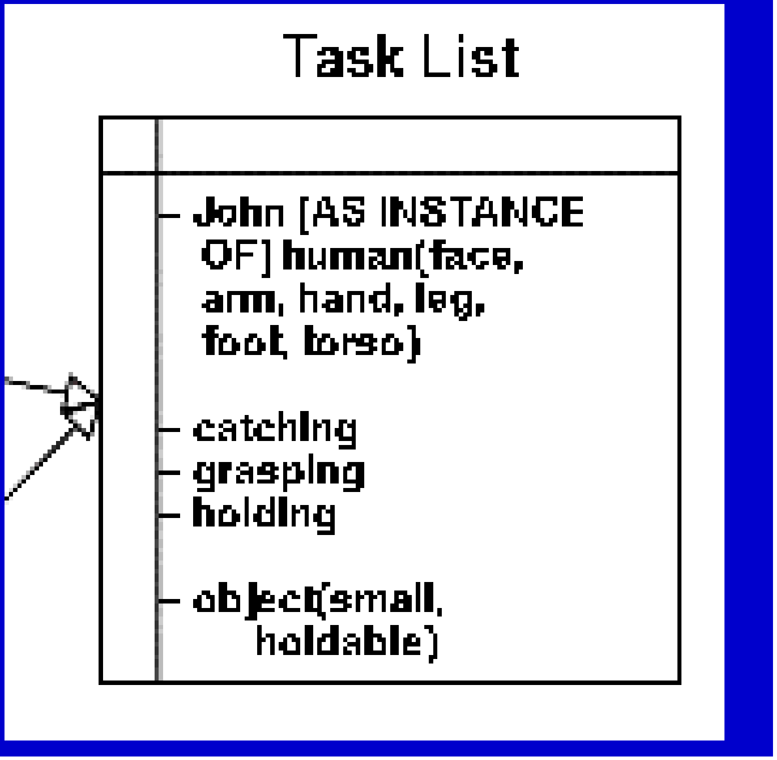
Example of operation

4) Suppose that the visual system first attends to a bright-red chair in the scene.

Going through current task list, agent determines that this object is most probably irrelevant (not really "holdable")

Discard it from further consideration as a component of the minimal subscene.

Task map and task list remain unaltered.



Task List

- John [AS INSTANCE OF] human(face, arm, hand, leg, foot, torso)
- catching
- grasping
- holding
- object(small, holdable)

More formally: how do we do it?

- What is the task list, given our formalism?

  it's a question to the KB: $\text{ASK}(\text{KB}, \exists x \ \text{LookFor}(x))$

- Is the currently attended and recognized object, o, of interest?

  $\text{ASK}(\text{KB}, \text{LookFor}(o))$

- How could we express that if the currently attended & recognized object is being looked for, we should add it to the minimal subscene?

$$\forall x \ \ \text{Attended}(x) \land \text{Recognized}(x) \land \text{LookFor}(x) \land$$
$$x \notin \text{MinimalSubscene} \Rightarrow x \in \text{MinimalSubscene}$$

with:

$$\forall x \ \exists t \ \ \text{RMSdistance}(x, t) < \text{Score}(0.1) \Rightarrow \text{Recognized}(x)$$

and similar for Attended() [Note: should be temporally tagged; see next]

Example of operation

5) Suppose next attended and identified object is John’s rapidly tapping foot.

This would match the “foot” concept in the task list.

Because of relationship between foot and human (in KB), agent can now prime visual system to look for a human that overlap with foot found:

- feature bias derived from what memory for human
- spatial bias for location and scale

Task map marks this spatial region as part of the current minimal subscene.

Example of operation

6) Assume human is next detected and recognized

System should then look for its face
how? from KB we should be able to infer that resolving

“? [AS INSTANCE OF] human”

can be done by looking at the face of the human.

Once John has been localized and identified, entry
“John [AS INSTANCE OF] human(face, arm, hand, leg, foot, torso)”
simplifies into simpler entry
“John [AT] (x, y, scale)”

Thus, further visual biasing will not attempt to further localize John.

More formally: how do we do it?

- How do we introduce the idea of successive attentional shifts and progressive scene understanding to our formalism?

  Using situation calculus!

- Effect axioms (describing change):

  $\forall x,s \;\; Attended(x, s) \wedge Recognized(x, s) \wedge LookFor(x, s) \Rightarrow$
  $\qquad \neg LookFor(x, Result(AddToMinimalSubscene, s))$

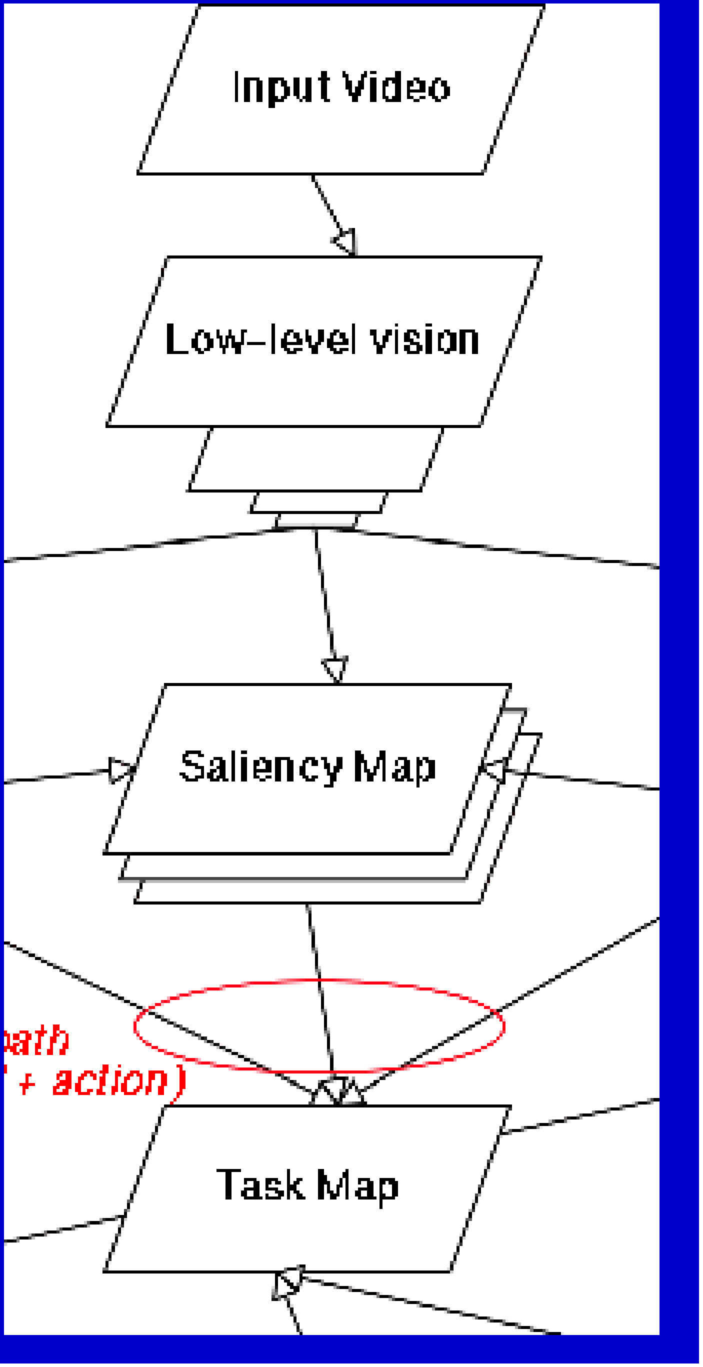
- Successor-state axioms (better than the frame axioms for non-change):

  $\forall x,a,s \quad x \in MinimalSubscene(Result(a, s)) \Leftrightarrow$
  $\qquad (a = AddToMinimalSubscene) \vee$
  $\qquad (x \in MinimalSubscene(s) \wedge a \neq DeleteFromMinimalSubscene)$

7) Suppose system then attends to the bright green emergency exit sign in the room

This object would be immediately discarded because it is too far from the currently activated regions in the task map.

Thus, once non-empty, the task map acts as a filter that makes it more difficult (but not impossible) for new information to reach higher levels of processing, that is, in our model, matching what has been identified to entries in the task list and deciding what to do next.



8) Assume that now the system attends to John’s arm motion

This action will pass through the task map (that contains John)

It will be related to the identified John (as the task map will not only specify spatial weighting but also local identity)

Using the knowledge base, what memory, and current task list the system would prime the expected location of John’s hand as well as some generic object features.

9) If the system attends to the flying ball, it would be incorporated into the minimal subscene in a manner similar to that by which John was (i.e., update task list and task map).

10) Finally: activity recognition.

The various trajectories of the various objects that have been recognized as being relevant, as well as the elementary actions and motions of those objects, will feed into the activity recognition sub-system

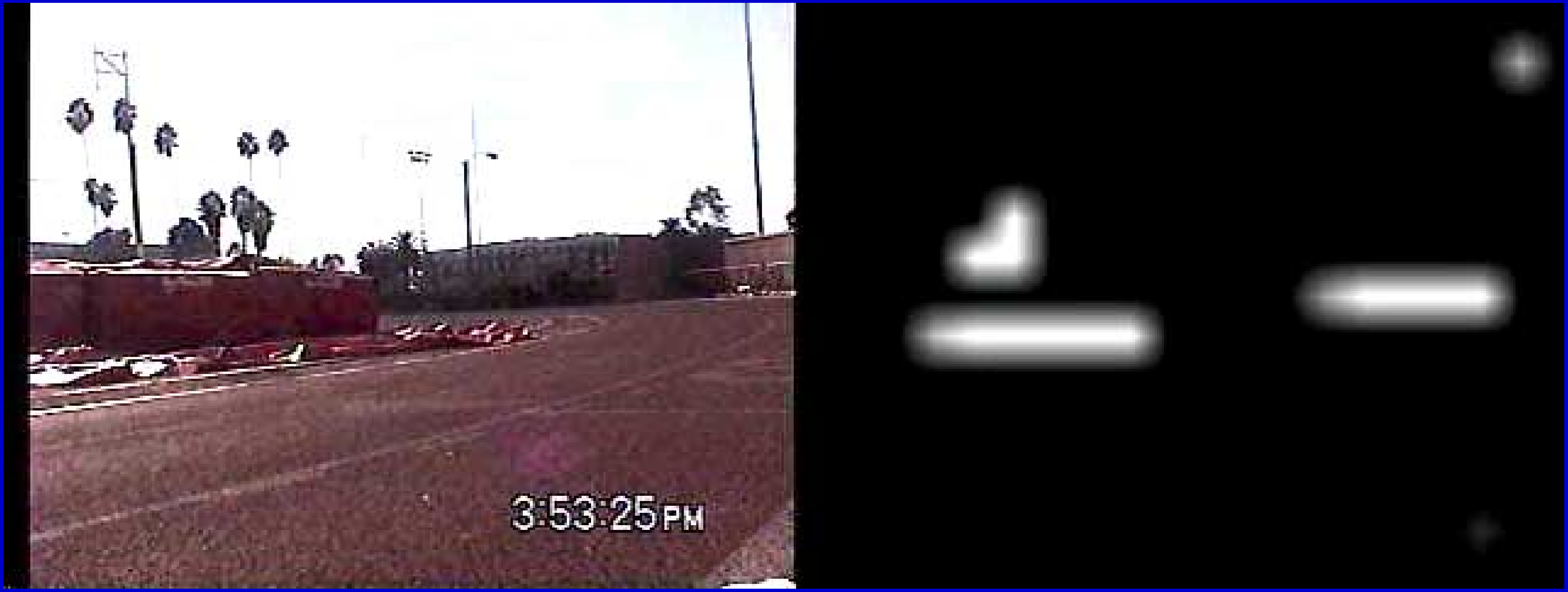
=> will progressively build the higher-level, symbolic understanding of the minimal subscene.

e.g., will put together the trajectories of John's body, hand, and of the ball into recognizing the complex multi-threaded event "human catching flying object."

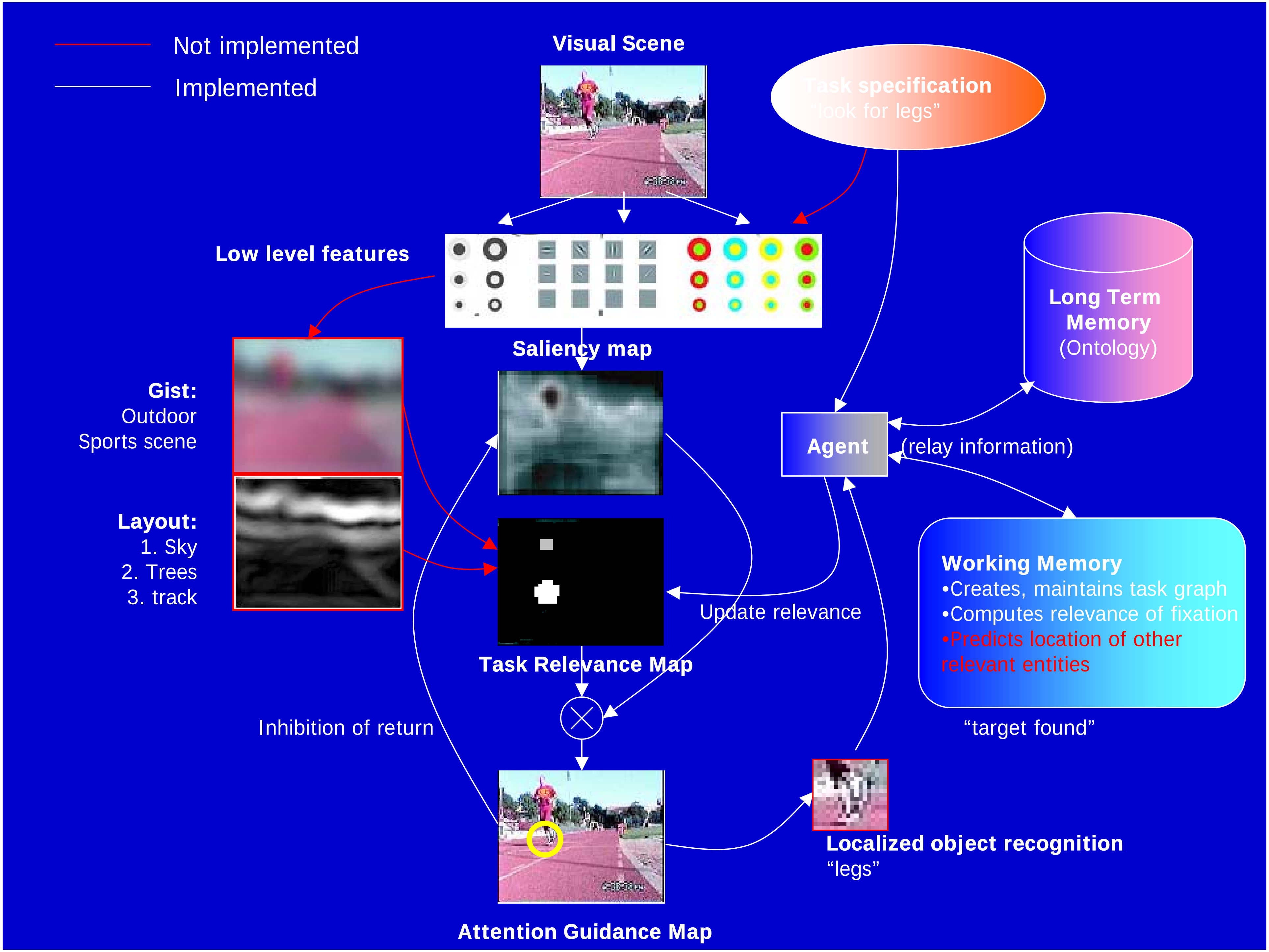
11) Once this level of understanding is reached, the data needed for the system’s answer will be in the form of the task map, task list, and these recognized complex events, and these data will be used to fill in an appropriate sentence frame and apply the answer.

Example

- Question: “who is doing what to whom?”



- Answer: “Eric passes, turns around and passes again”



Not implemented

Implemented

Visual Scene

Task specification
“look for legs”

Low level features

Long Term Memory
(Ontology)

Saliency map

Gist:
Outdoor
Sports scene

Layout:
1. Sky
2. Trees
3. track

Agent

(relay information)

Working Memory

- Creates, maintains task graph
- Computes relevance of fixation
- Predicts location of other relevant entities

Task Relevance Map

Update relevance

Inhibition of return

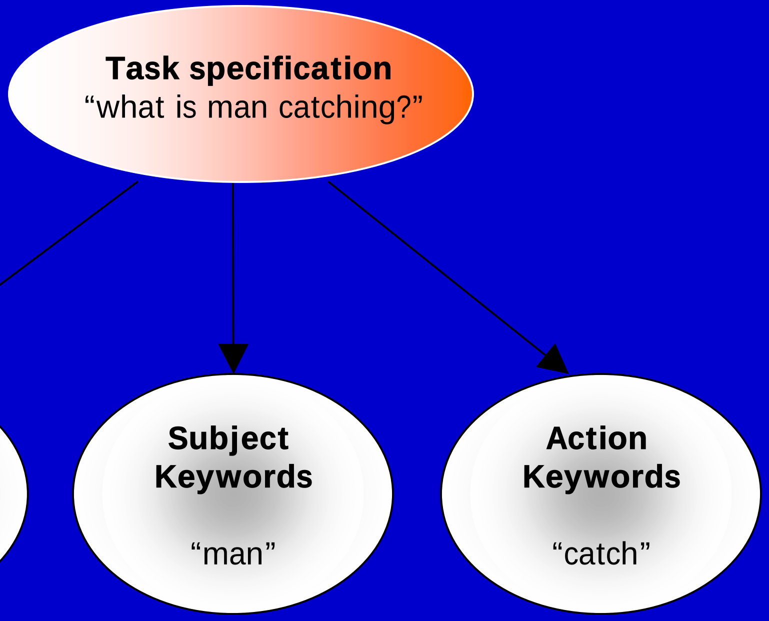
“target found”

Localized object recognition
“legs”

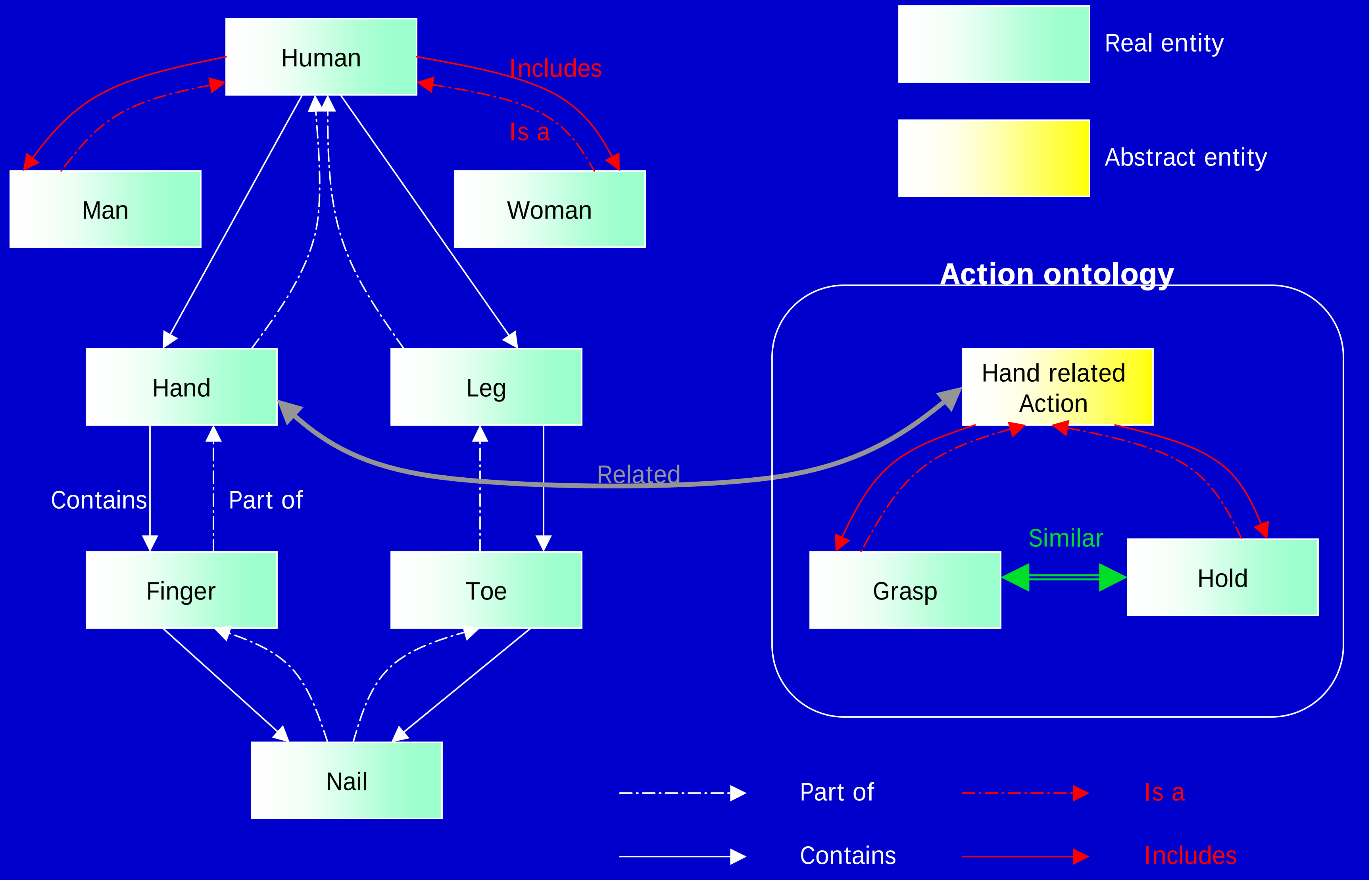
Attention Guidance Map

Task Specification

- Currently, we accept tasks such as "who is doing what to whom?"



Subject ontology



Human

Includes

Is a

Man

Woman

Hand

Leg

Contains

Part of

Finger

Toe

Nail

Related

Action ontology

Hand related Action

Similar

Grasp

Hold

🟩 Real entity

🟨 Abstract entity

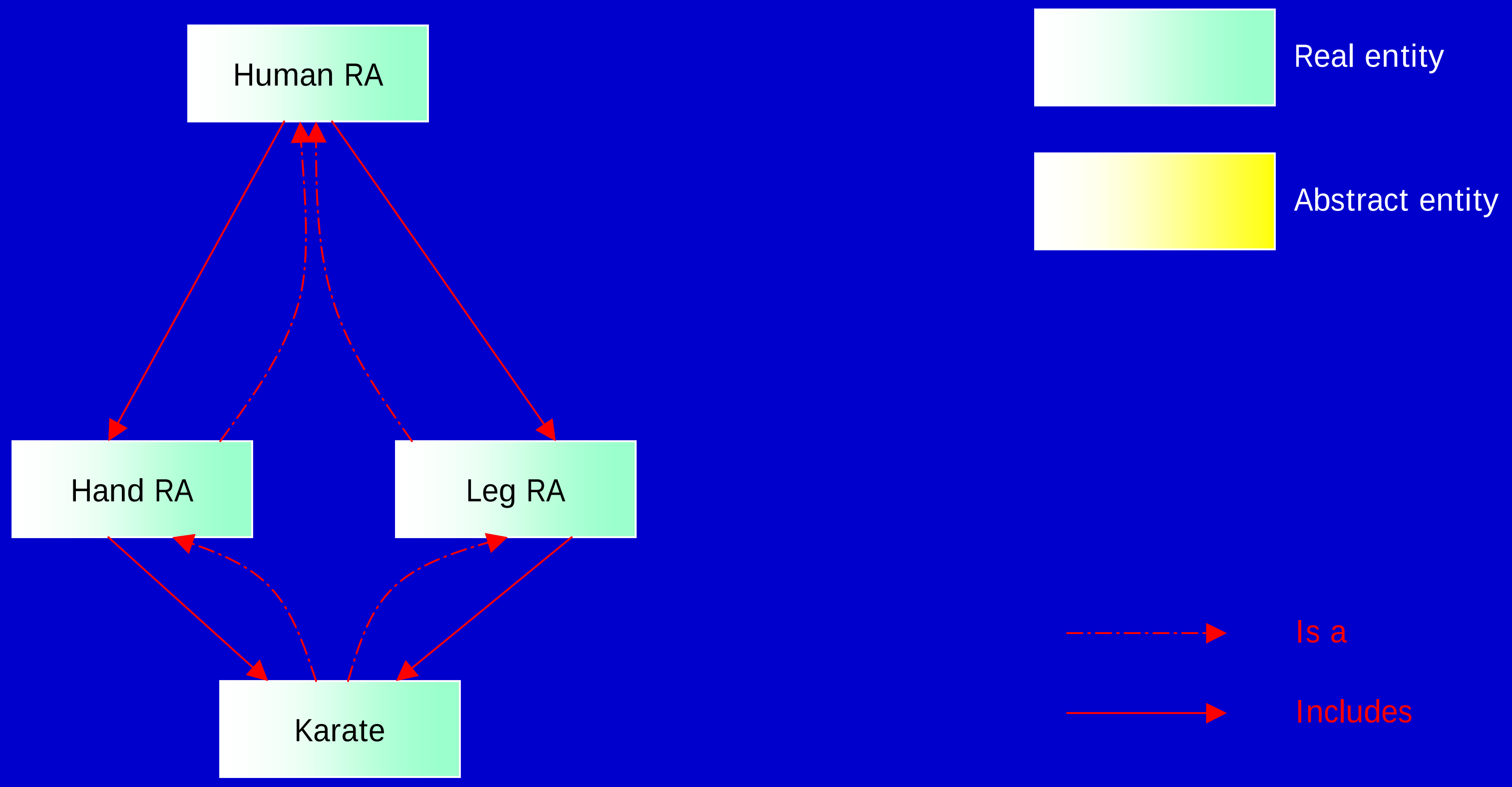
Part of

Is a

Contains

Includes

What to store in the nodes?



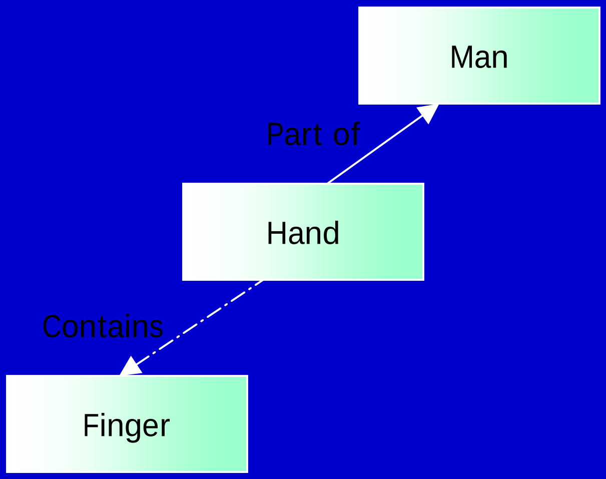
	Leg RA	$\overline{\text{Leg RA}}$
Hand RA	0.65	0.35
$\overline{\text{Hand RA}}$	0	0

Properties

Probability of occurrence

What to store in the edges?

Task: “find hand”



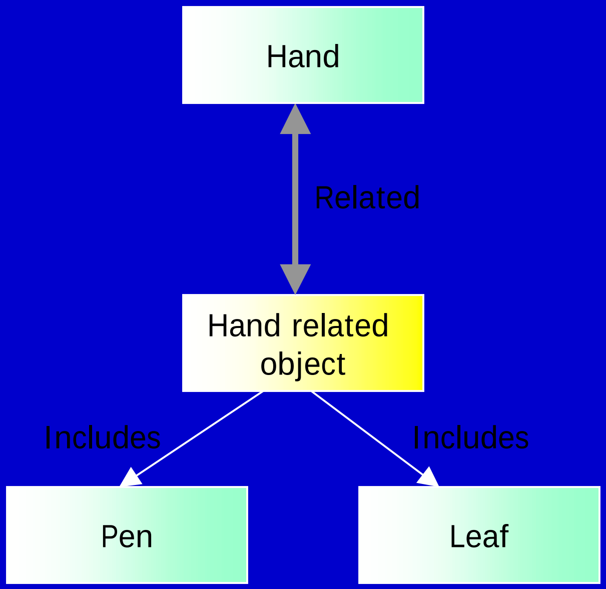
Suppose we find Finger and Man, what is more relevant?

Granularity g(u,v)

g((Hand, Finger)) > g((Hand, Man))

In general, **g(contains) > g(part of)**
g(includes) > g(is a)
g(similar) = g(related)

Edge information



Hand

Related

Hand related object

Includes

Includes

Pen

Leaf

Co-occurrence(u,v)

Probability of joint occurrence of u and v

Task: “find hand”

Suppose we find Pen and Leaf, what is more relevant?

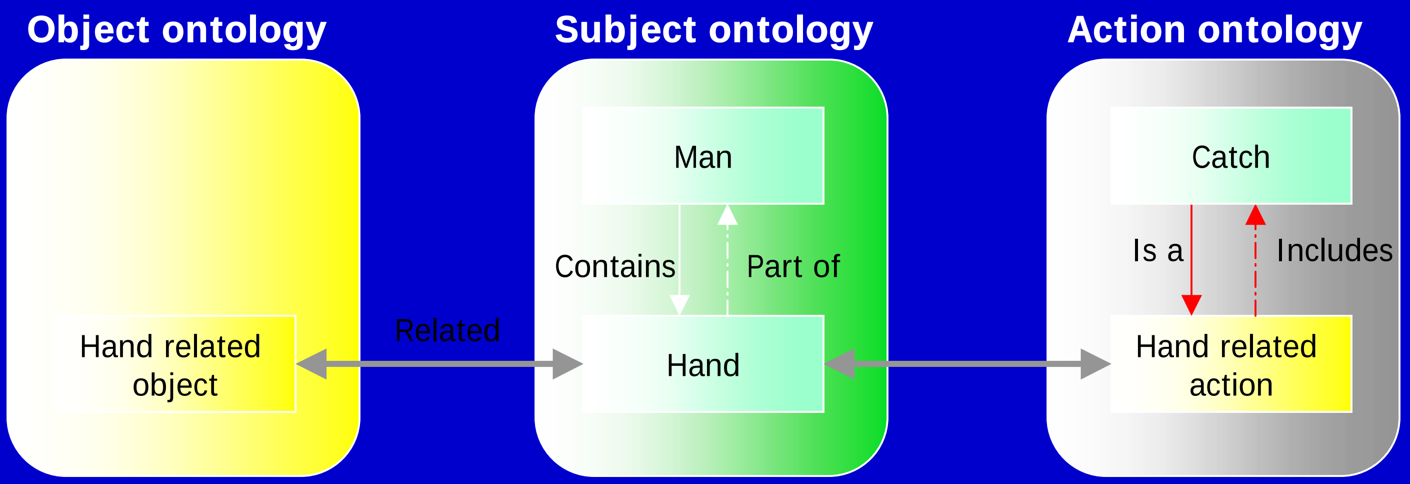
P(Pen is relevant/ Hand is relevant)
Vs
P(Leaf is relevant/ Hand is relevant)

P(Hand occurs/Pen occurs)
Vs
P(Hand occurs/ Leaf occurs)

P(Hand, Pen/ Pen)
Vs
P(Hand, Leaf/ Leaf)

Working Memory and Task Graph

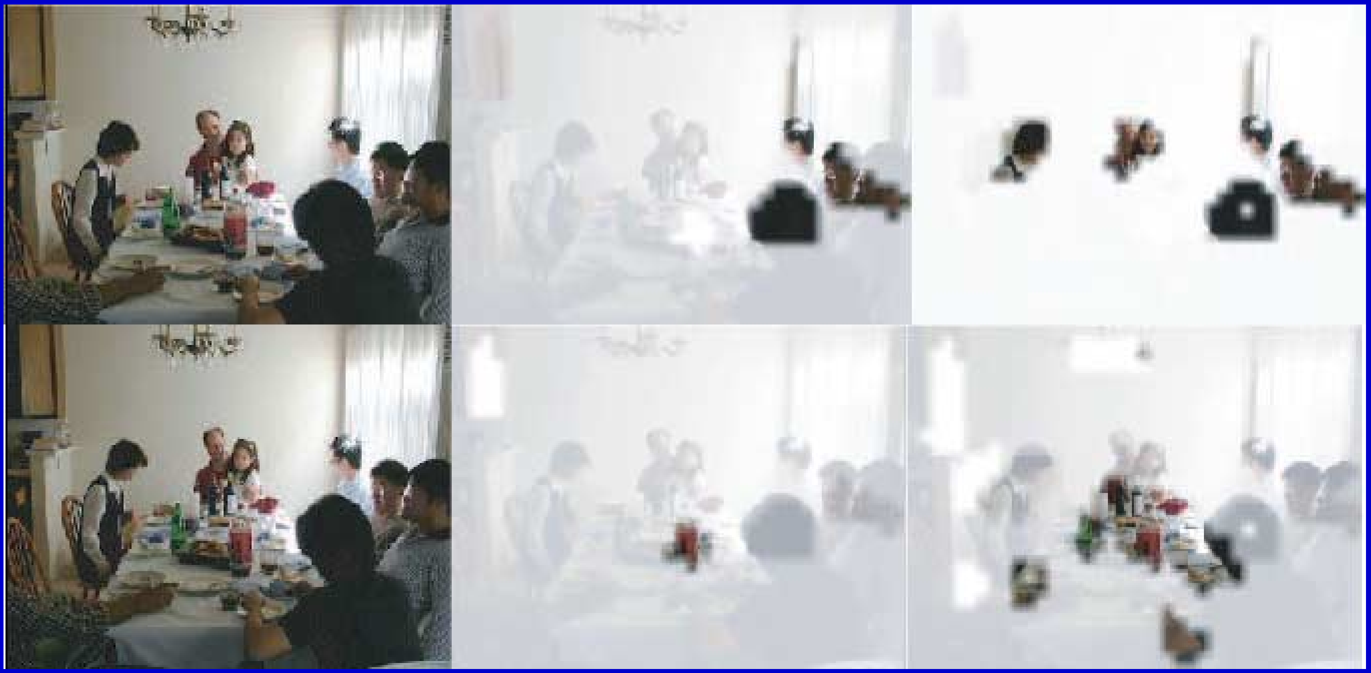
- Working memory creates and maintains the task graph
- Initial task graph is created using the task keywords and is expanded using “is a” and “related” relations.



Task: What is man catching?

Example 1

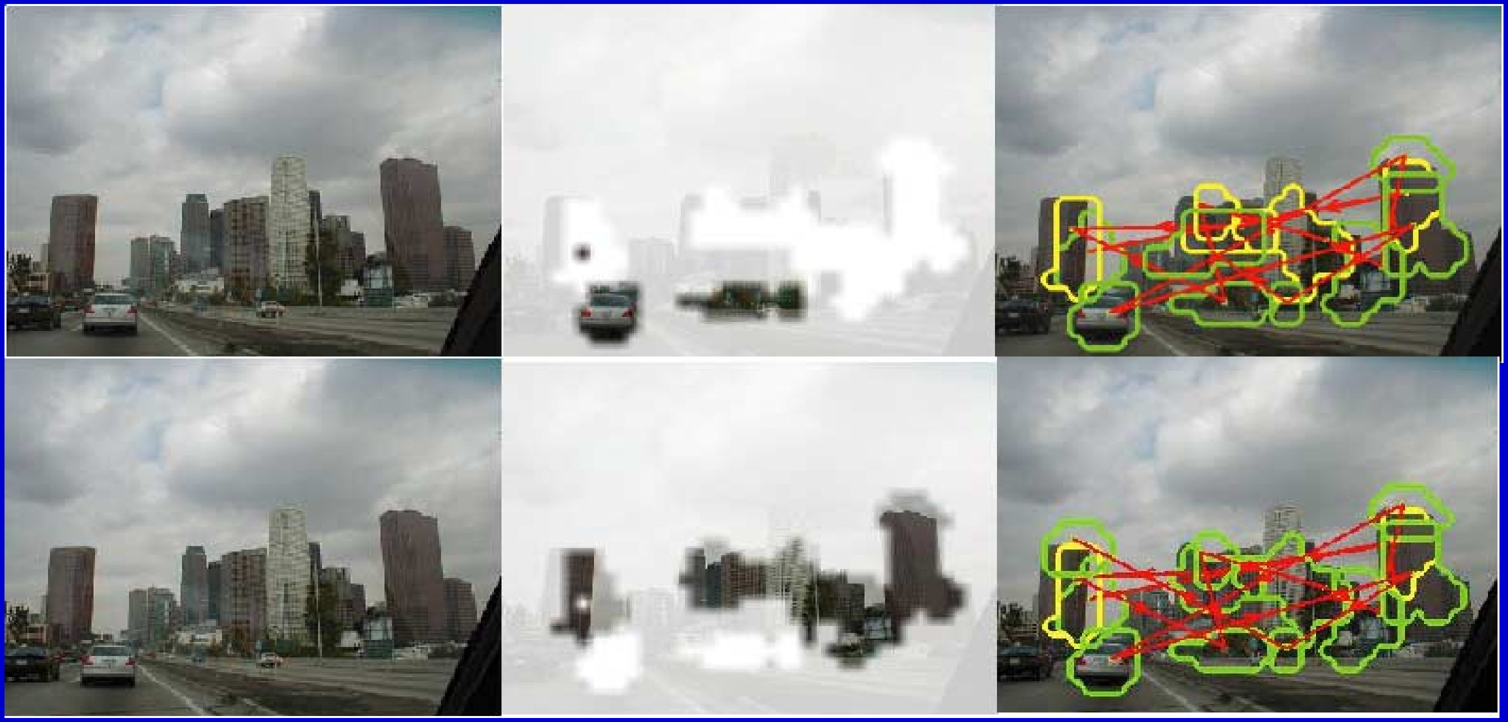
- Task1: find the faces in the scene
- Task2: find what the people are eating



Original scene TRM after 5 fixations TRM after 20 fixations

Example 2

- Task1: find the cars in the scene
- Task2: find the buildings in the scene



Original scene TRM after 20 fixations Attention trajectory

Conclusion

- Our broader goal is to model how internal scene representations are influenced by current behavioral goals.
- As a first step, we estimate task-relevance of attended locations.
- At each instant, our model guides attention based on relevance and salience of entities in the scene.

Outlook

- Neuromorphic vision algorithms provide robust front-end for extended scene analysis
- To be useful, analysis needs to highly depend on task and behavioral priorities
- Thus the challenge is to develop algorithms that can extract the currently relevant "minimal subscene" from incoming rich scanpaths
- Such algorithms will use a collaboration between fast parallel computation of scene gist/layout and slower attentional scanning of scene details