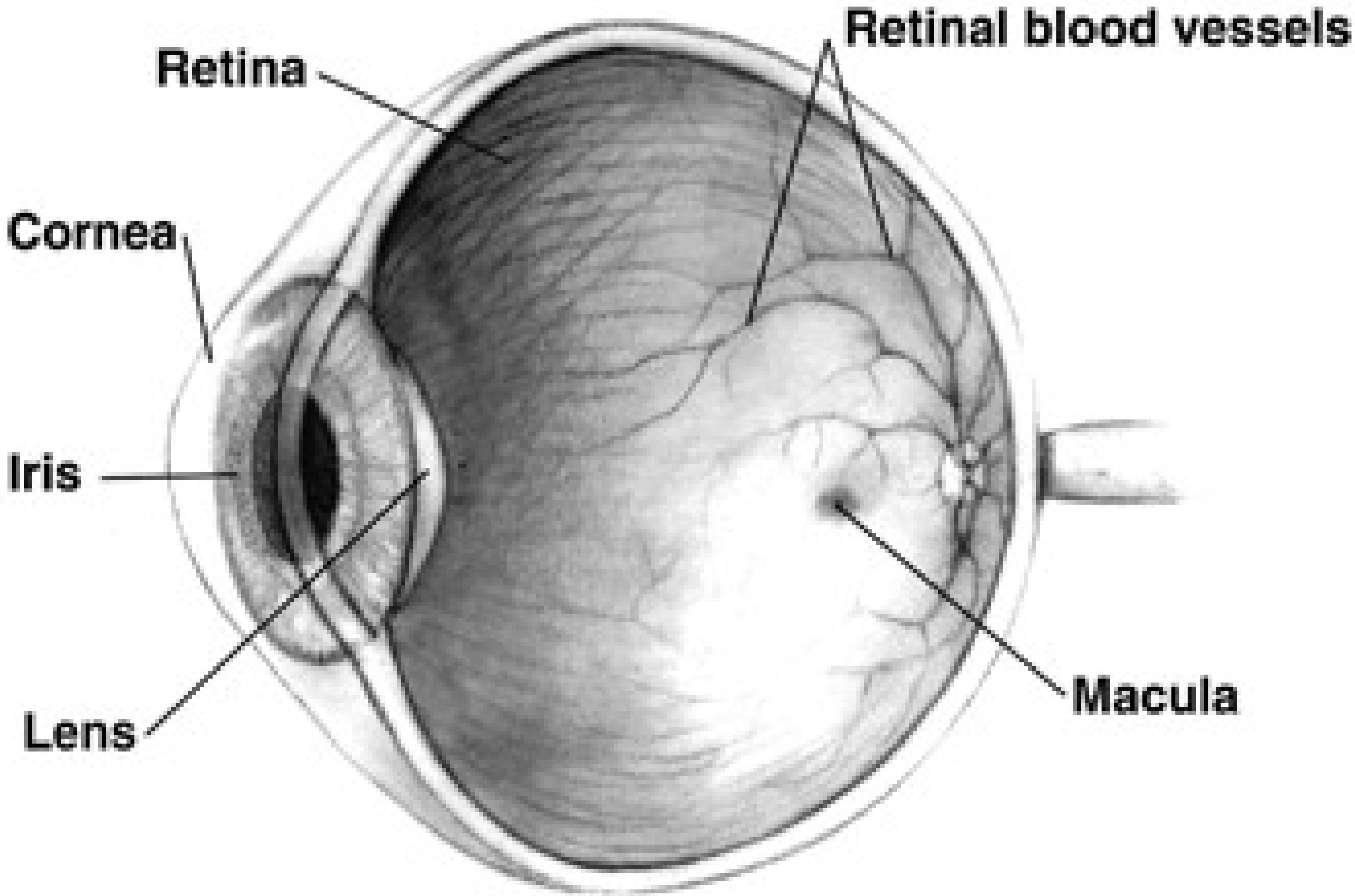
Attention, Saliency, Grouping

Thomas Breuel

EYE MOVEMENTS

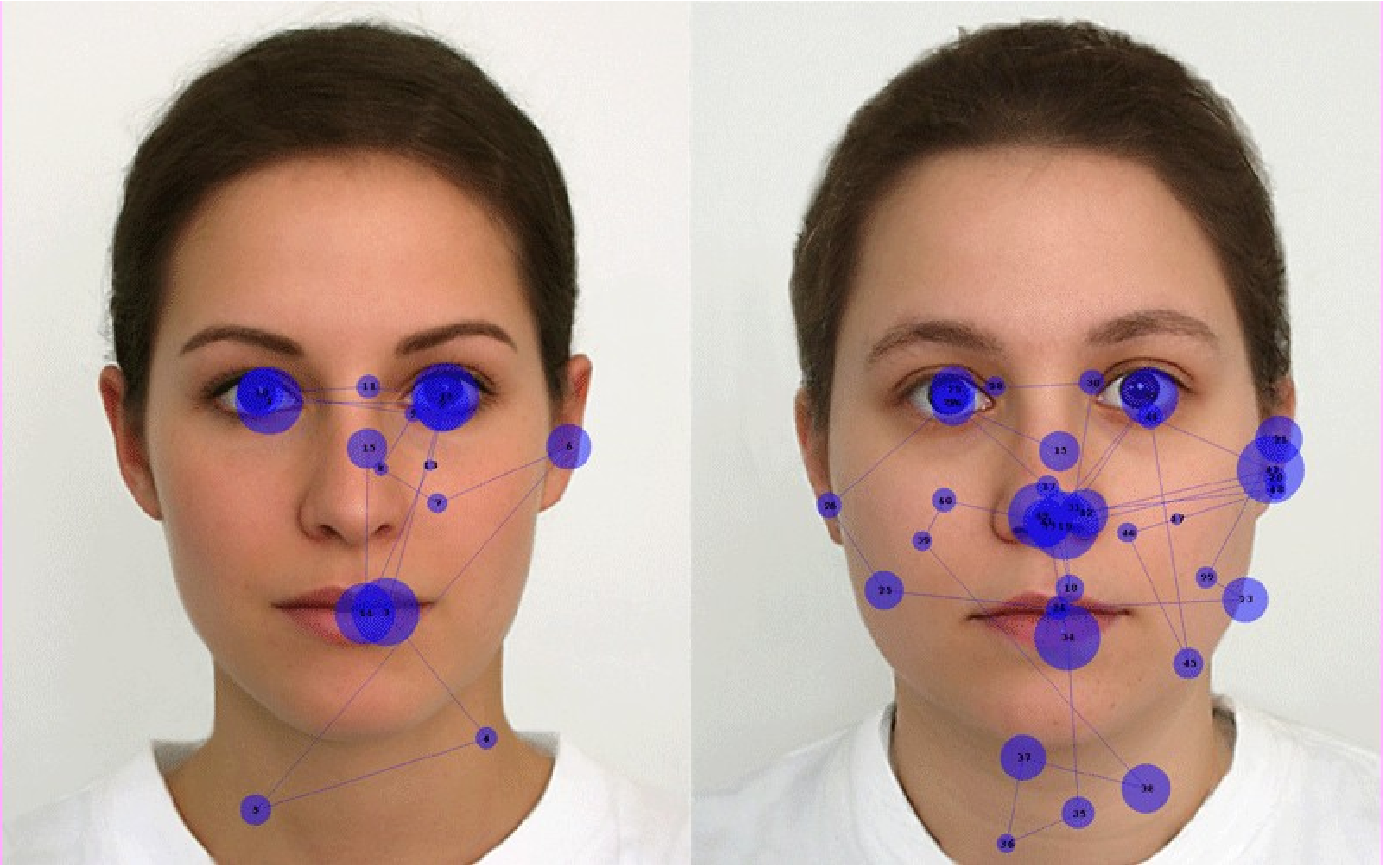
eye



what you think you see

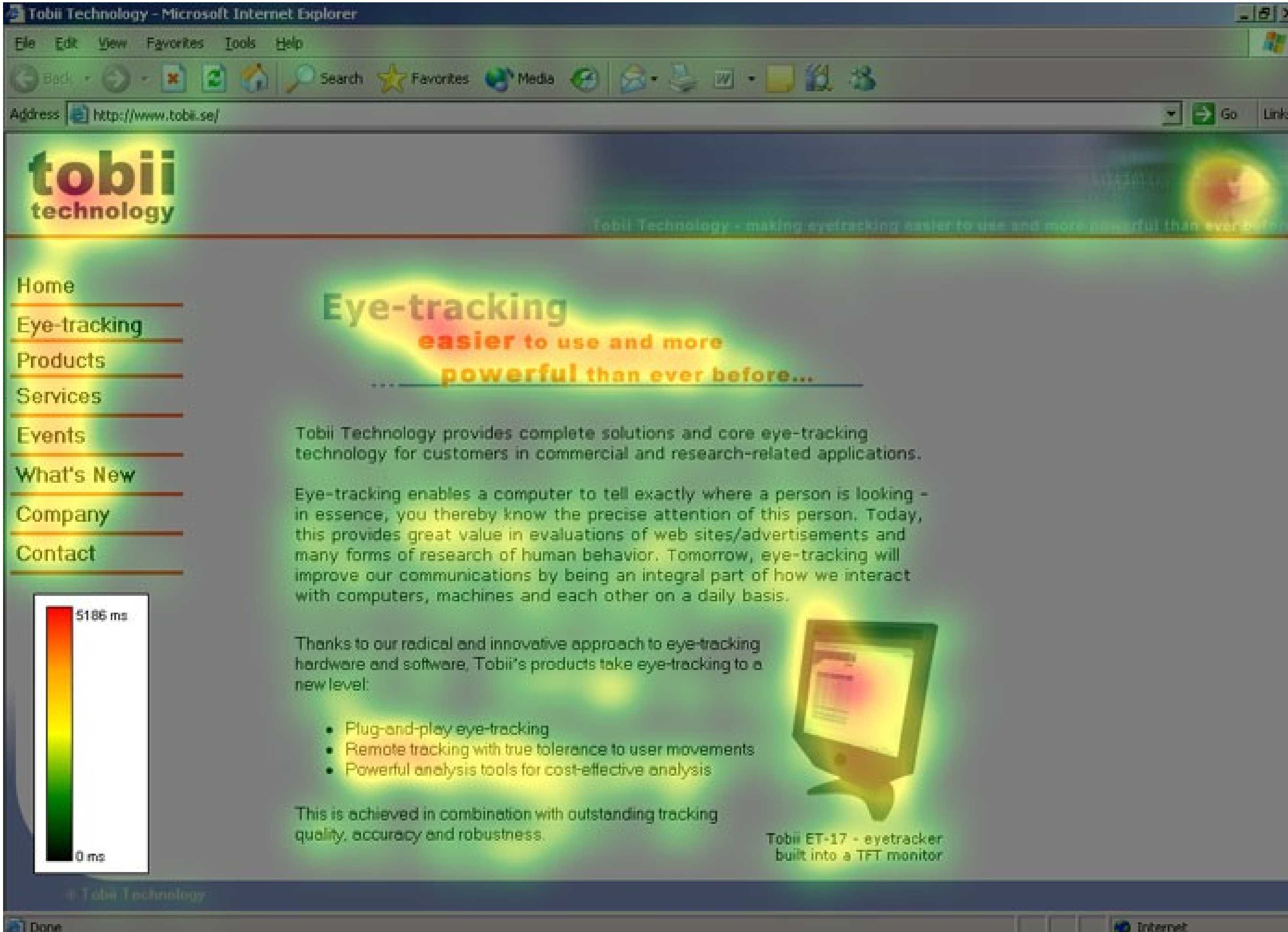
what you actually see

foveation

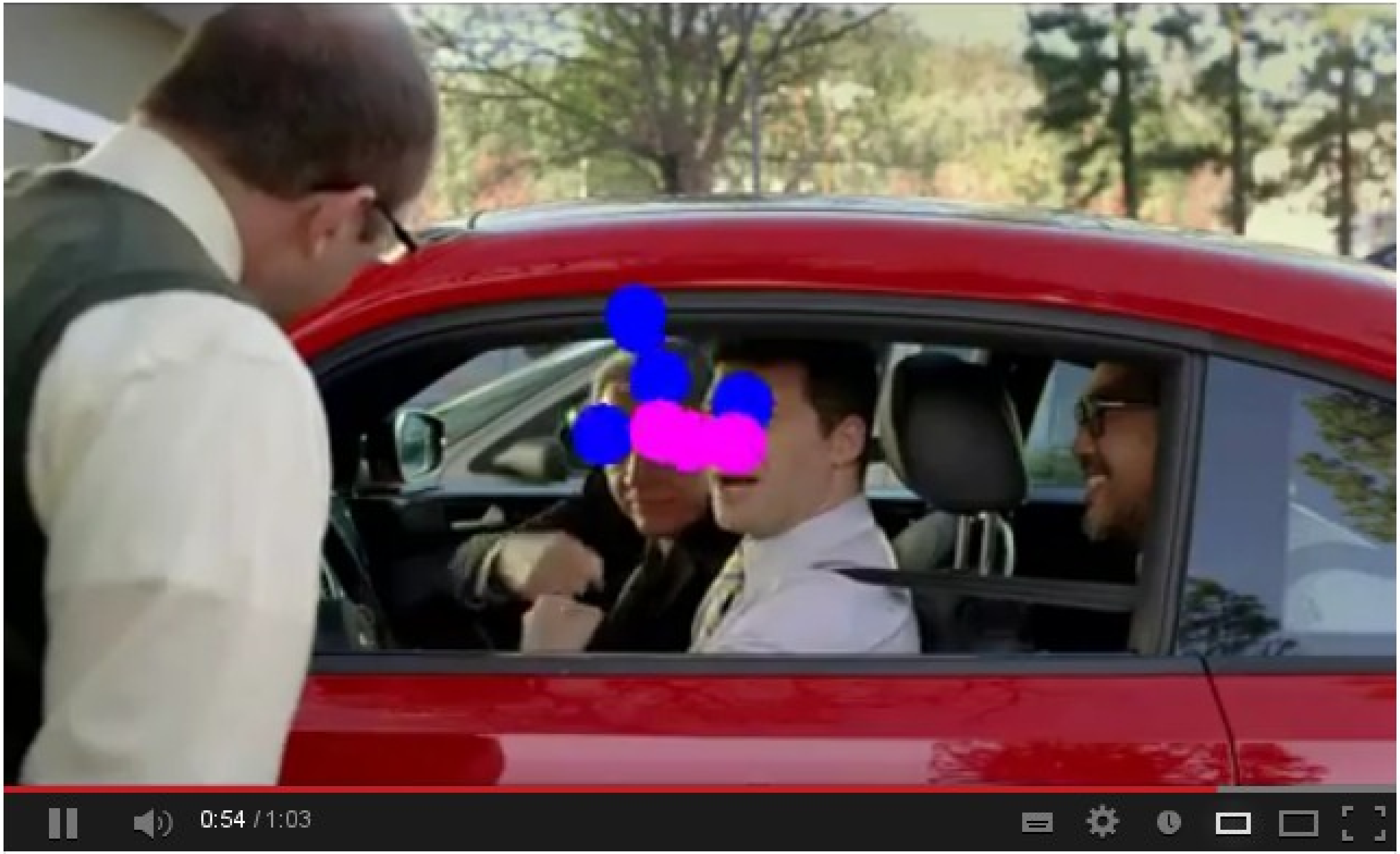


www.tobii.se

eye tracking in user interfaces

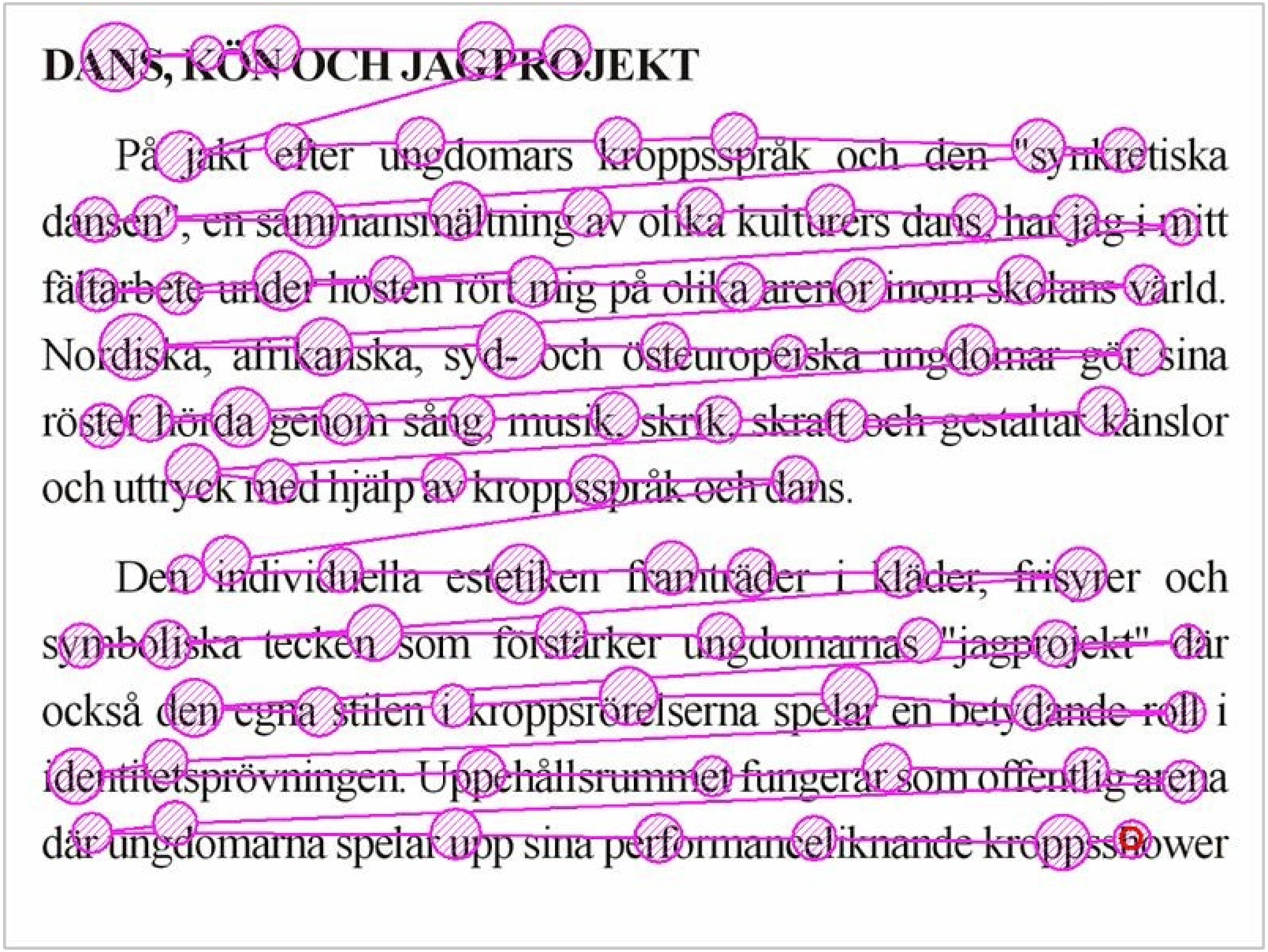


http://www.tobii.se



Eye Tracking - Volkswagen - 2013 Super Bowl Commercial

http://www.youtube.com/watch?v=bITGPkPxbac

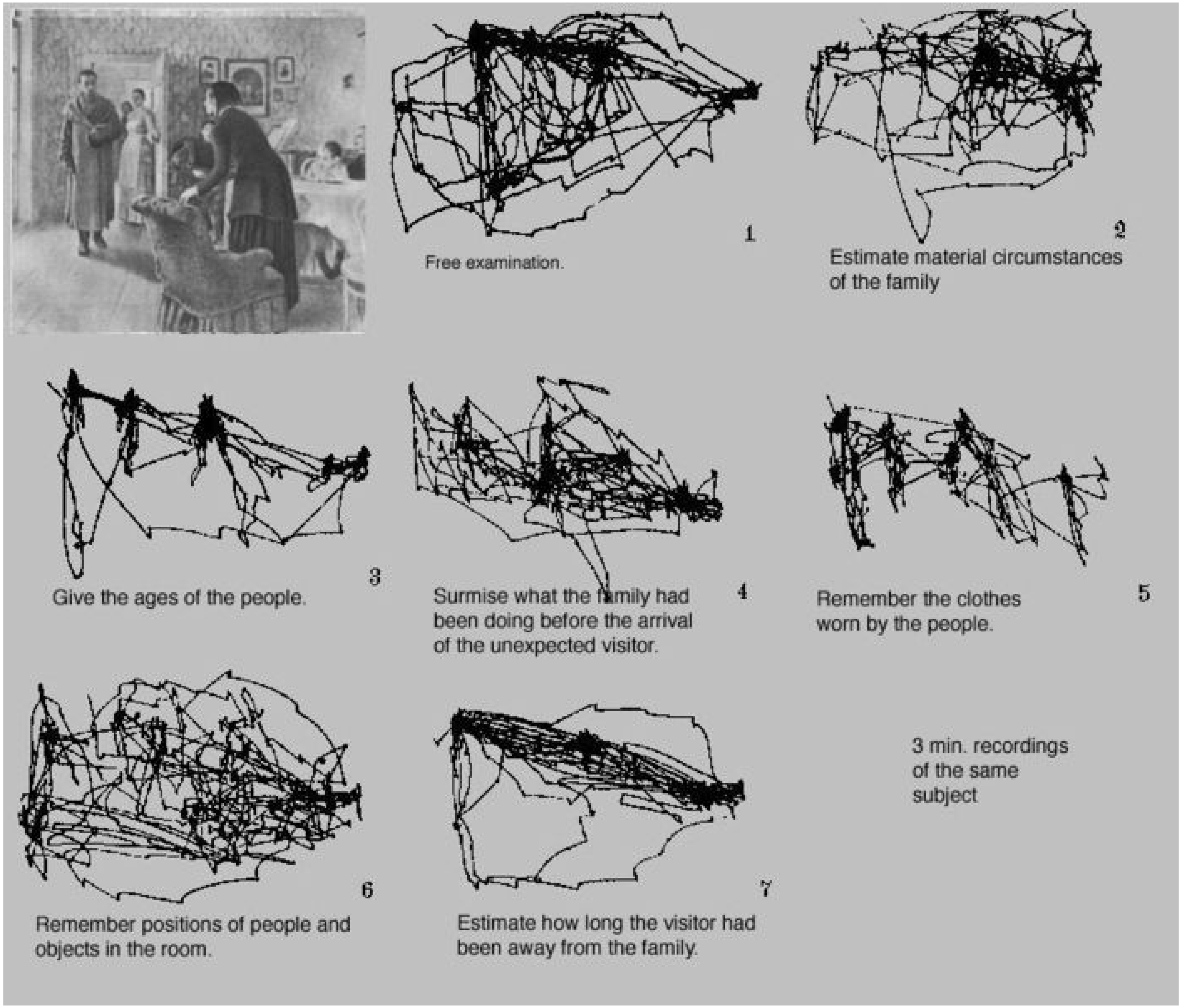


DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten rört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och östeuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestaltar känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträder i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Uppehållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performanceliknande kroppsshower

http://en.wikipedia.org/wiki/File:Reading_Fixations_Saccades.jpg



http://en.wikipedia.org/wiki/File:Yarbus_The_Visitor.jpg

eye movements

saccades

eyes scan the scene to bring different parts into the fovea (sharp, color vision)

perception is masked during saccade

microsaccades

retinal cells only respond to change

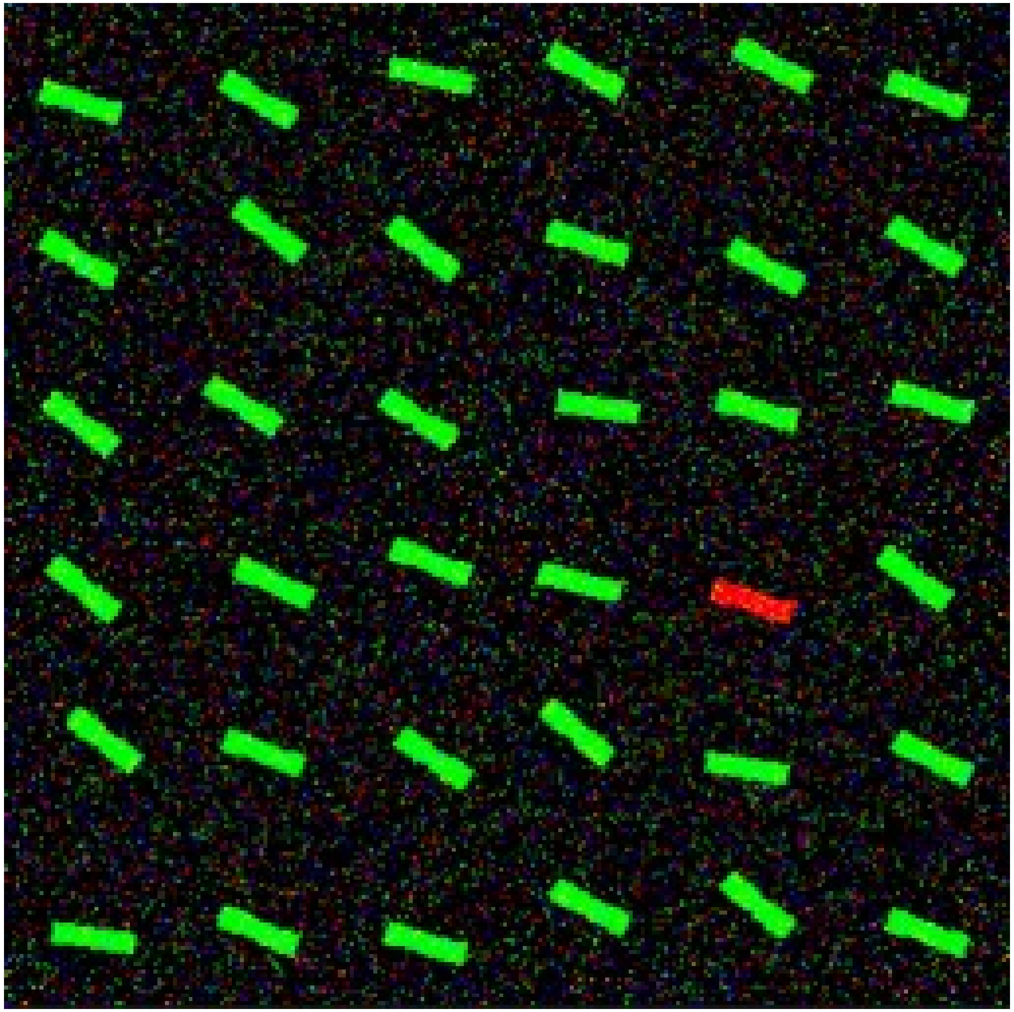
microsaccades are tiny movements

about 60/second

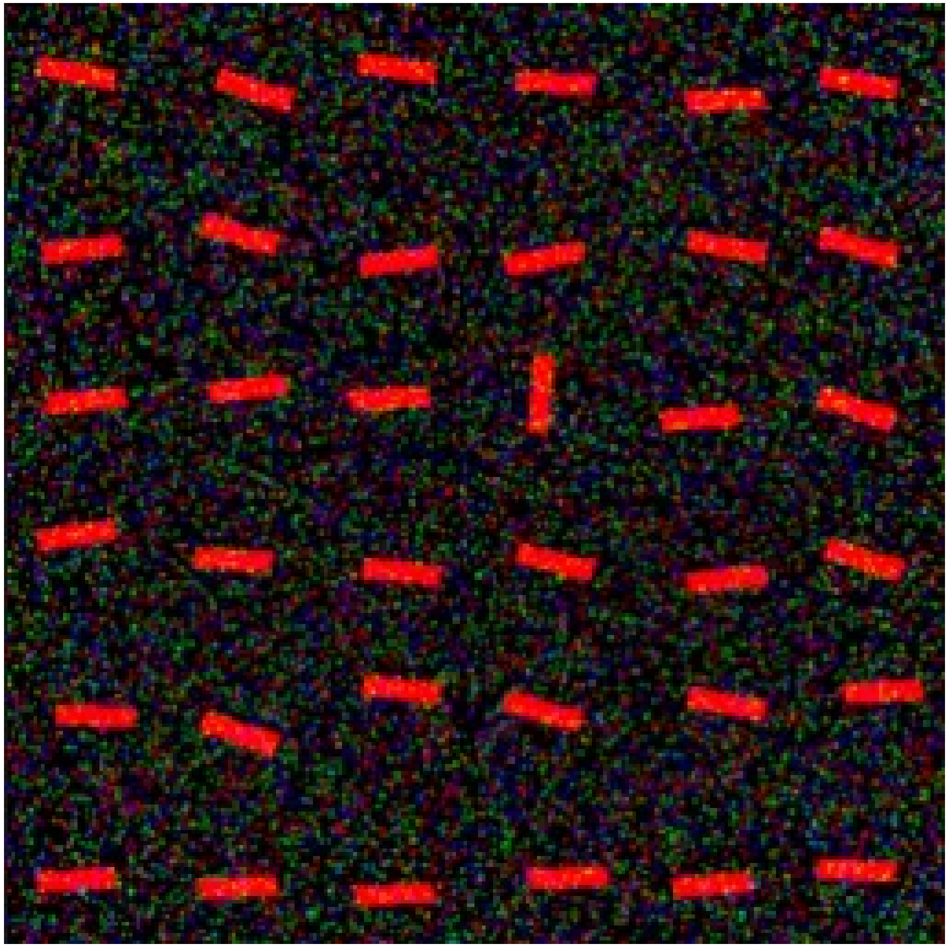
SEARCH AND SALIENCY

serial vs parallel search

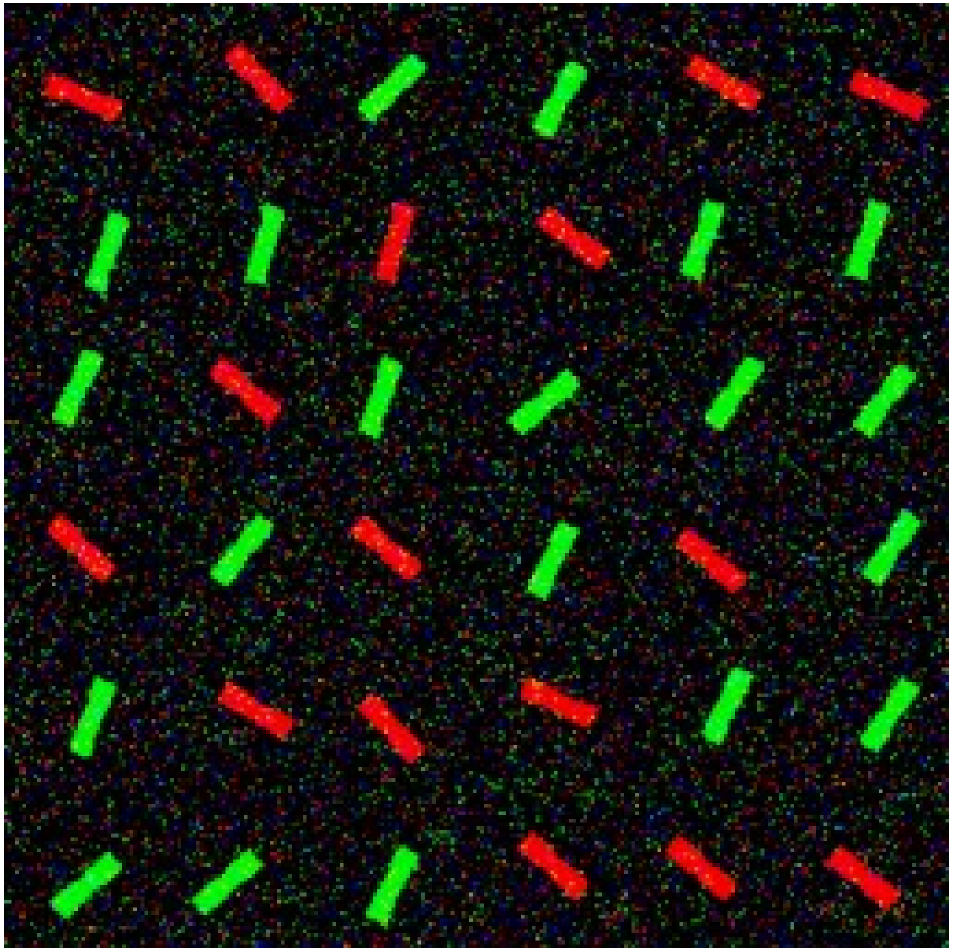
- **sample tasks**
  - “find an object with property X”
  - “find an object that differs from the objects around it”
- **execution**
  - does search time depend on # objects?
  - serial search – linear dependence on # objects
  - parallel search – constant in # objects



http://www.scholarpedia.org/article/Visual_salience

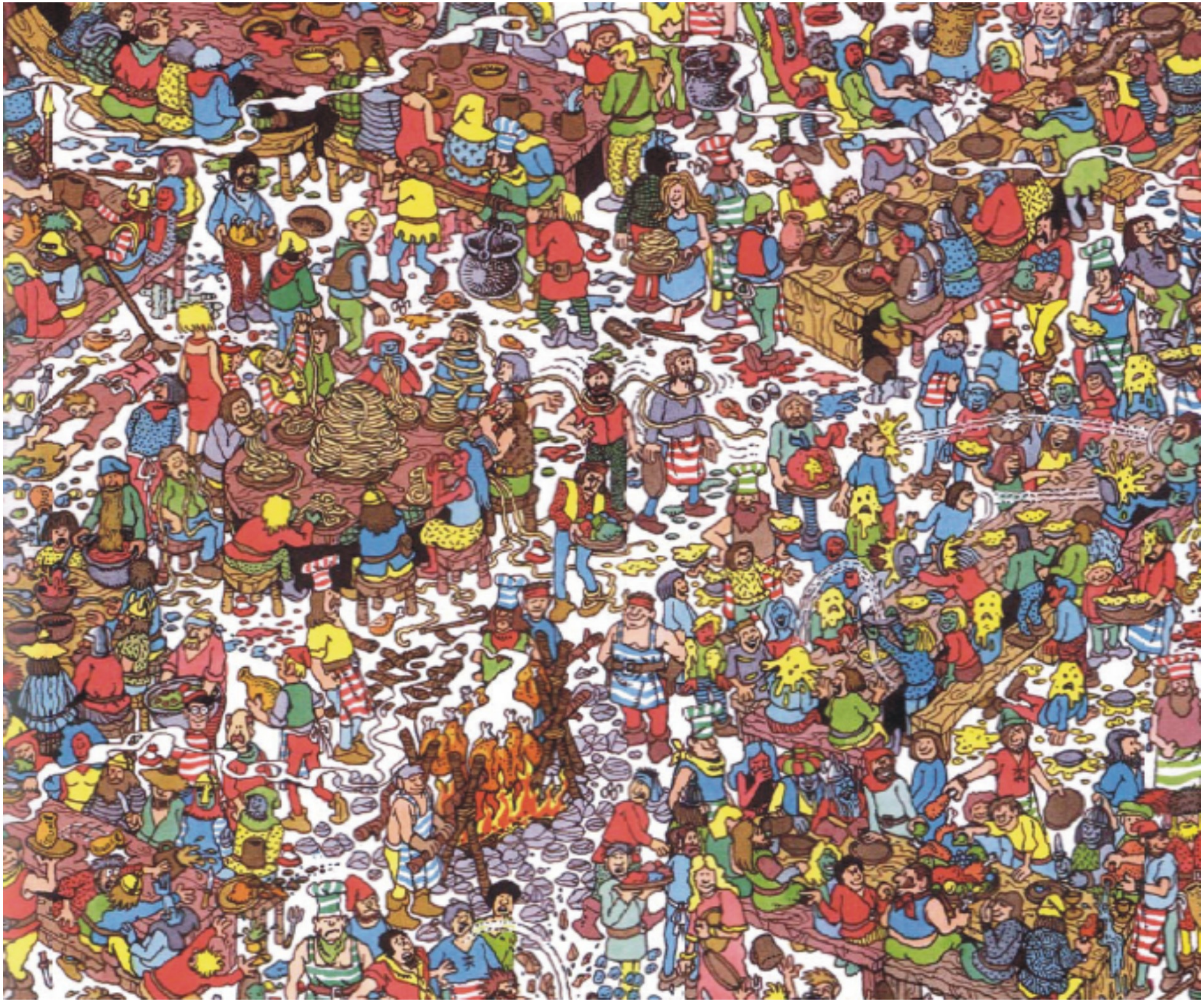


http://www.scholarpedia.org/article/Visual_salience



http://www.scholarpedia.org/article/Visual_salience

http://www.scholarpedia.org/article/Visual_salience



Where is Waldo?

foveation vs attention

different concepts

foveation: sensory

attention: neural / psychological

differences

you can't attend to what you can't see

you may not pay attention even to things you look at

why two mechanisms?

- **attentional / saliency mechanisms**
  - fast, requires no eye movements
  - can solve some tasks directly
  - used to choose targets for further saccades
- **eye movement**
  - finer / more detailed analysis
  - target of eye movement needs to be precomputed

attentional mechanism exist in other sensory modalities

- cocktail party effect
- smell
- touch

cocktail party effect

https://www.youtube.com/watch?v=iO3jTl0WuS4

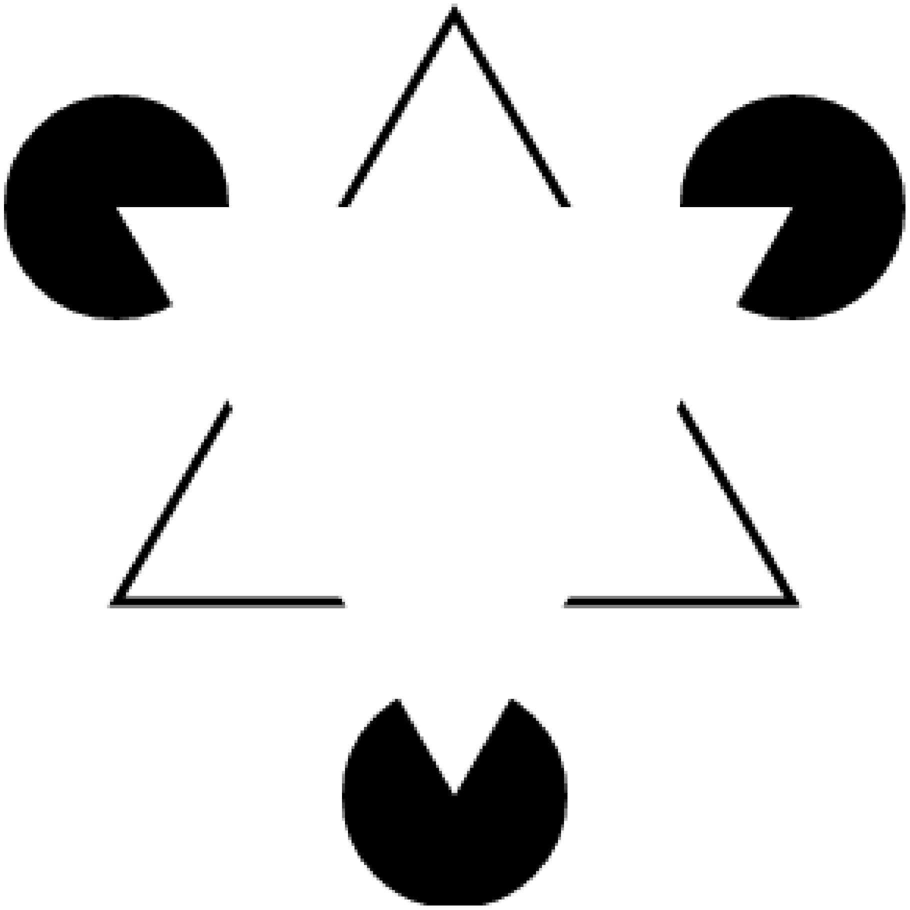
attention deficit disorder in kids

- easily distracted, miss details, forgetful
- trouble maintaining focus on a task
- become bored quickly
- daydreaming
- less good at fast, accurate information processing
- doesn't follow instructions

attention deficit disorder?

- **“attention” to tasks is important for effective learning and normal functioning**
- **neural basis is not well understood, but strong genetic component**
- **may be a deficit in high-level control of attentional mechanisms, or thresholds for saliency**

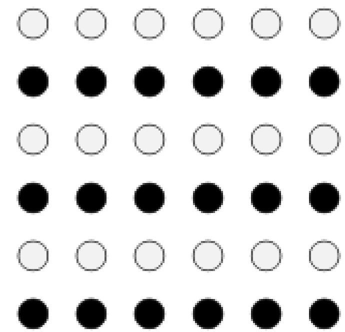
PERCEPTUAL ORGANIZATION



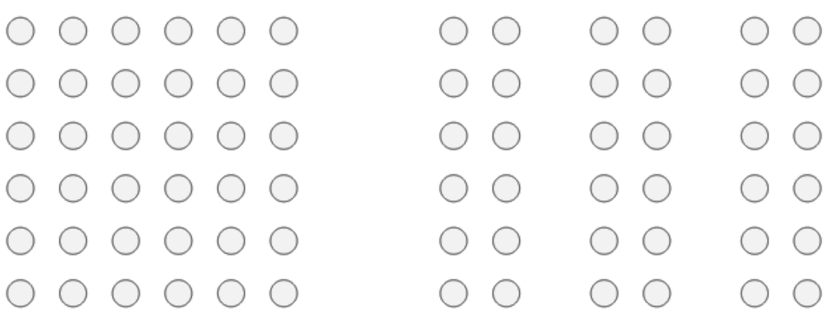
http://upload.wikimedia.org/wikipedia/commons/5/55/Kanizsa_triangle.svg

gestalt laws of grouping

closure

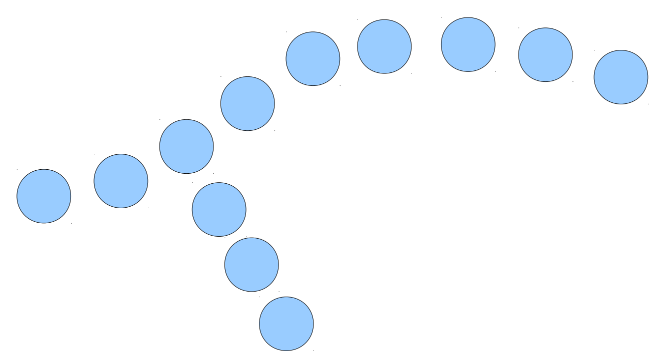


similarity



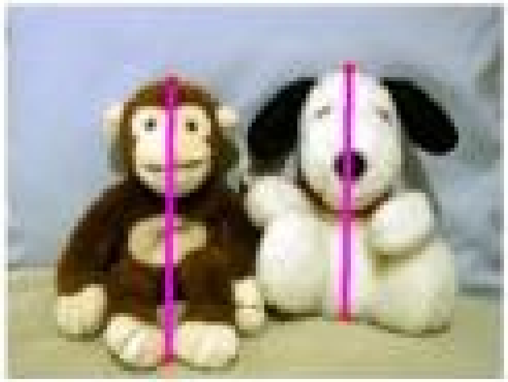
proximity

symmetry

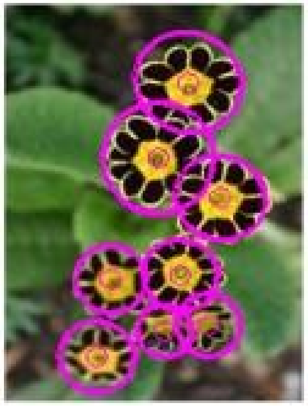


continuity

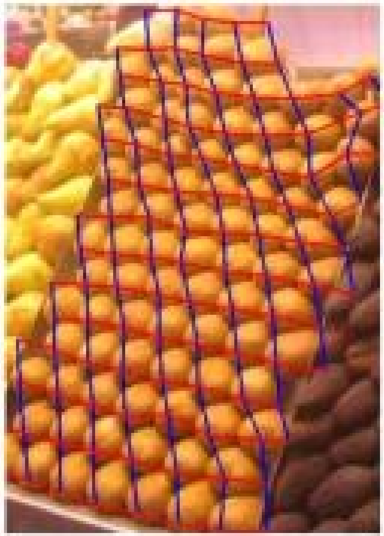
symmetry



reflection



rotation



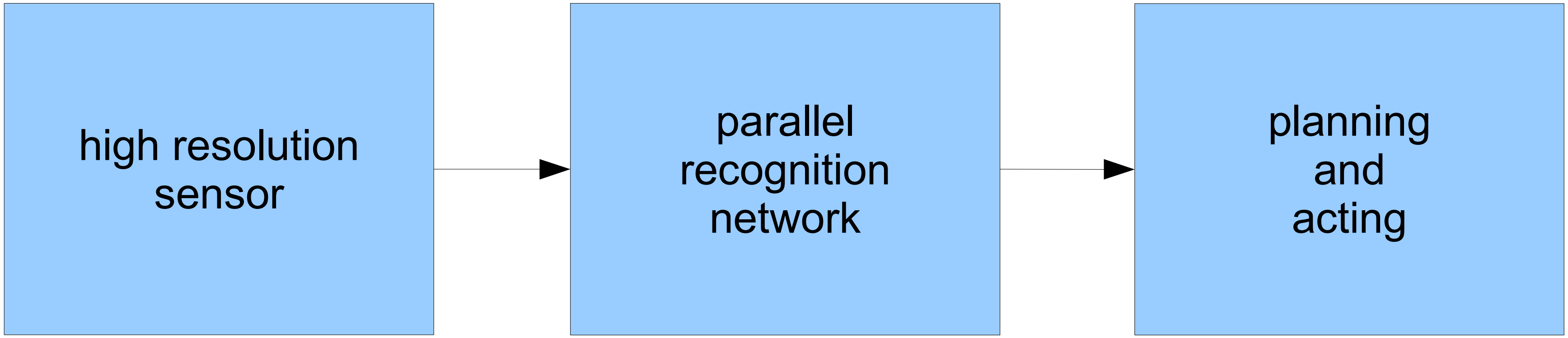
translation

perceptual organization

- **grouping**
  - given a scene element, find others that are probably part of the same object
- **segmentation**
  - divide the scene into regions that are likely part of the same object (oversegmentation: allow segmentation to be parts)
- **saliency**
  - find locations in the image that are likely part of interesting / important objects

WHY?

parallel feed-forward architecture

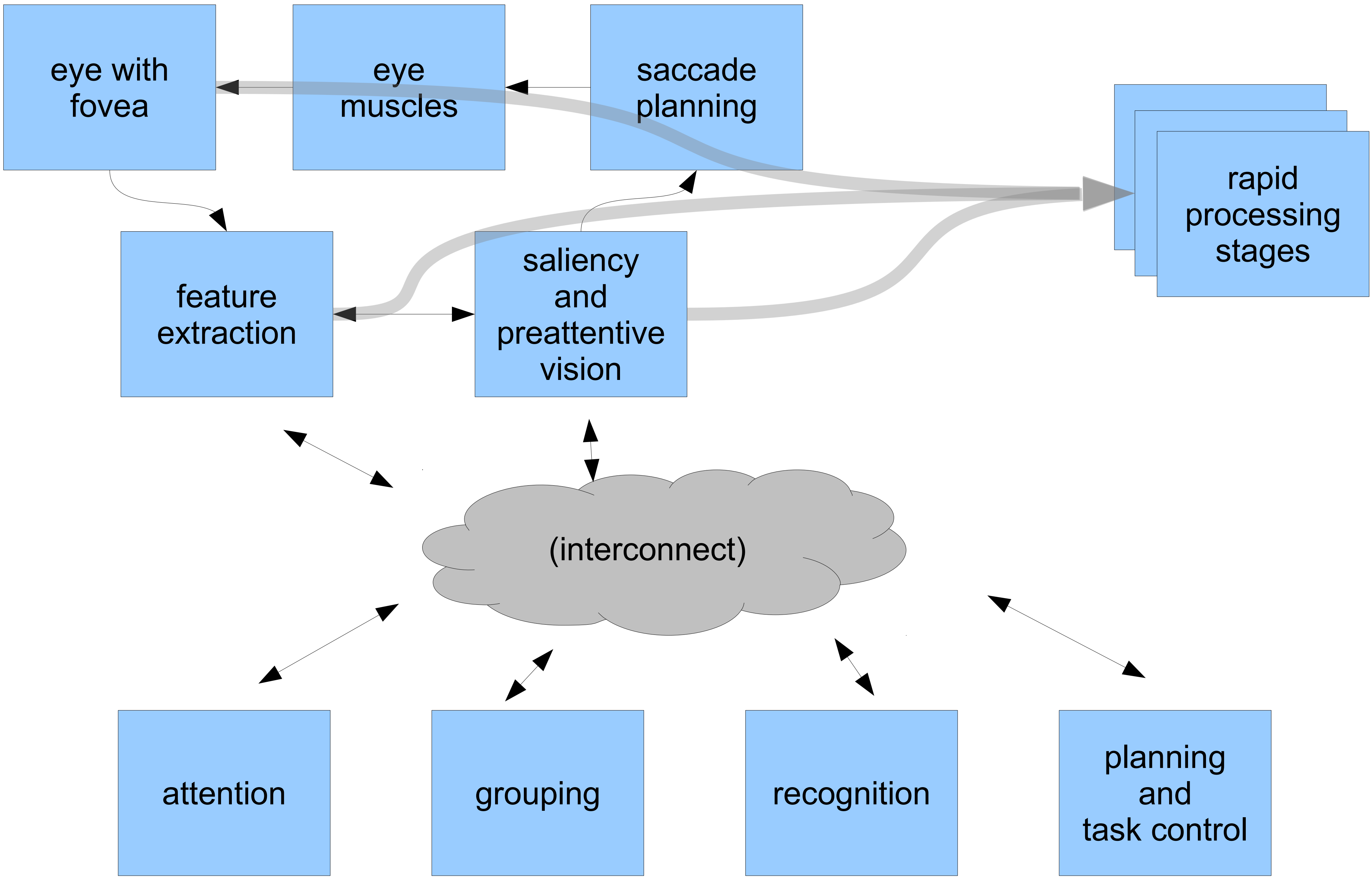


… no eye movement, but eyes bigger than brains

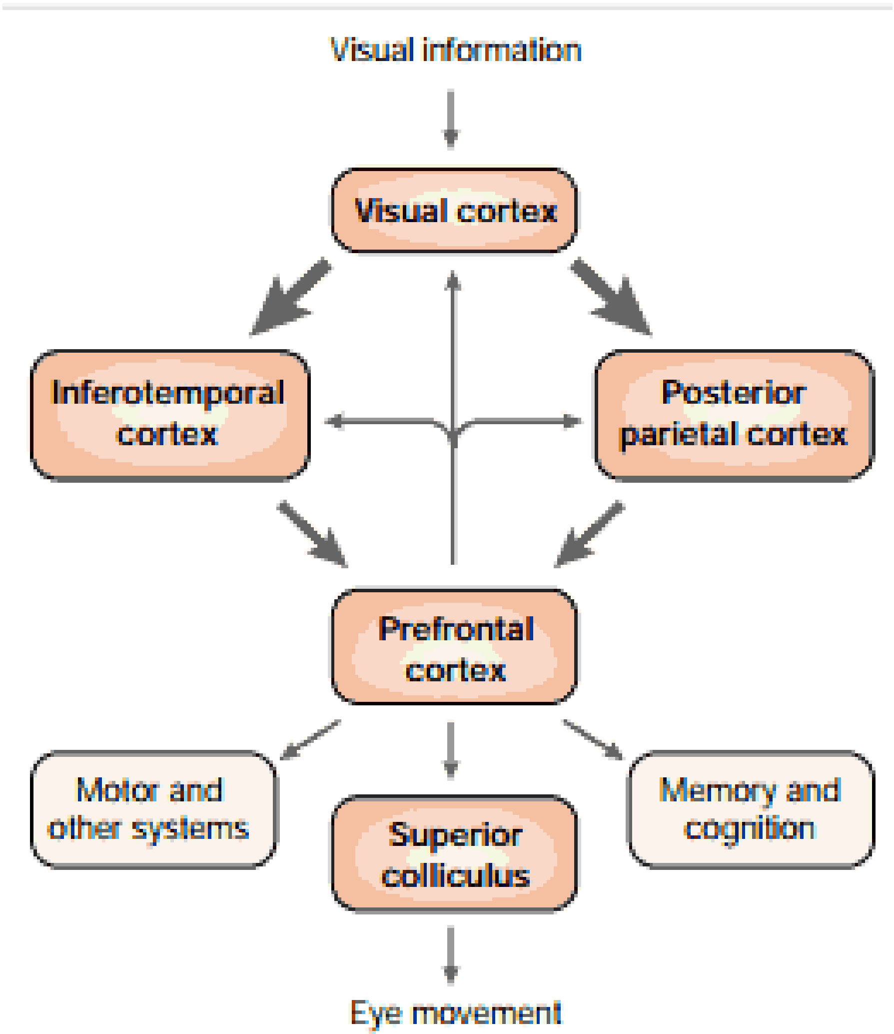
(actually, just uses head movements instead of eye movements)

https://www.youtube.com/watch?v=6Jz0JcQYtqo

components of the visual system

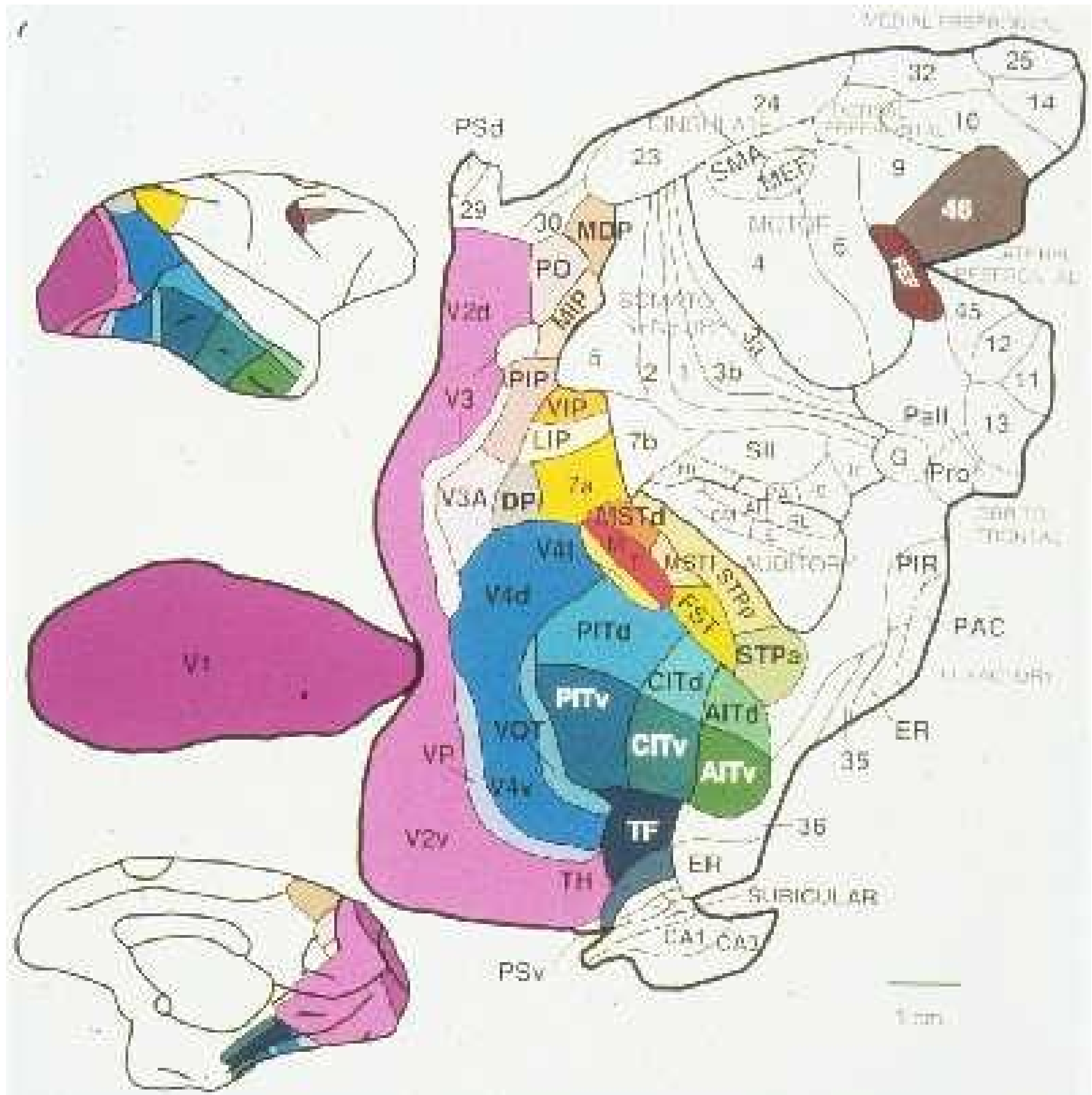


attentional subsystem



Itti and Koch, 2001

why?


PSd
29
30
MDP
PO
MIP
V2d
V3
PIP
VIP
LIP
V3A
DP
7a
7b
MSTd
MSTl
FST
STPp
STPa
V4t
V4d
PITd
PITv
CITd
CITv
AITd
AITv
VOT
V4v
VP
V2v
TF
TH
ER
35
36
SUBICULAR
CA1
CA3
PSv
V1
32
25
24
14
10
23
9
SMA
46
4
6
5
2
1
3b
SII
45
12
11
13
PaII
G
Pro
PIR
PAC
AUDITORY

hardware cost

- **tradeoff – for object recognition**
  - speed of visual processing and recognition
  - amount of neural hardware
  - foveation allows a small amount of hardware to cover a large field of view
- **object recognition is only one of many tasks**
  - sub-cortical vision
  - pre-attentive processing in the cortex

computational reasons

- **without segmentation, error rates would be higher**
- **without grouping, computational complexity would be higher**

COMPUTATIONAL CONSIDERATIONS

correspondence problem

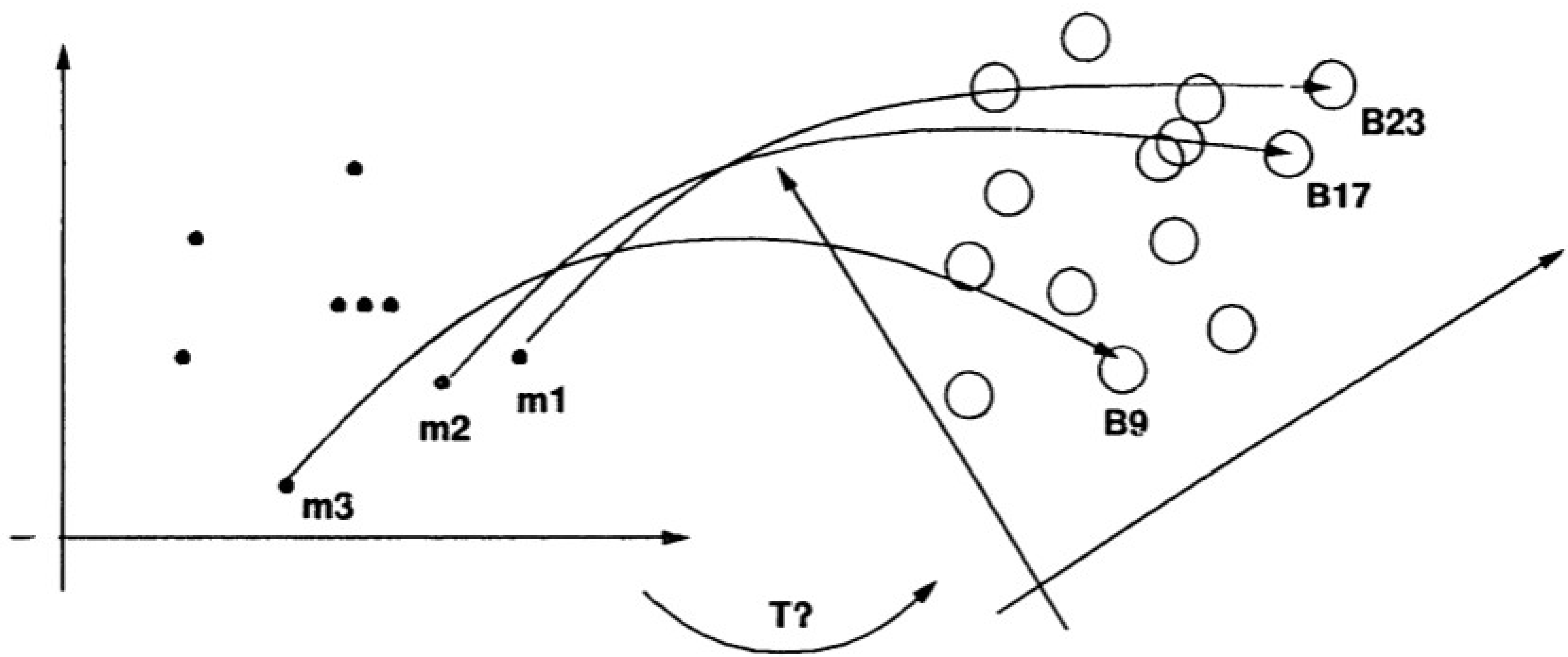
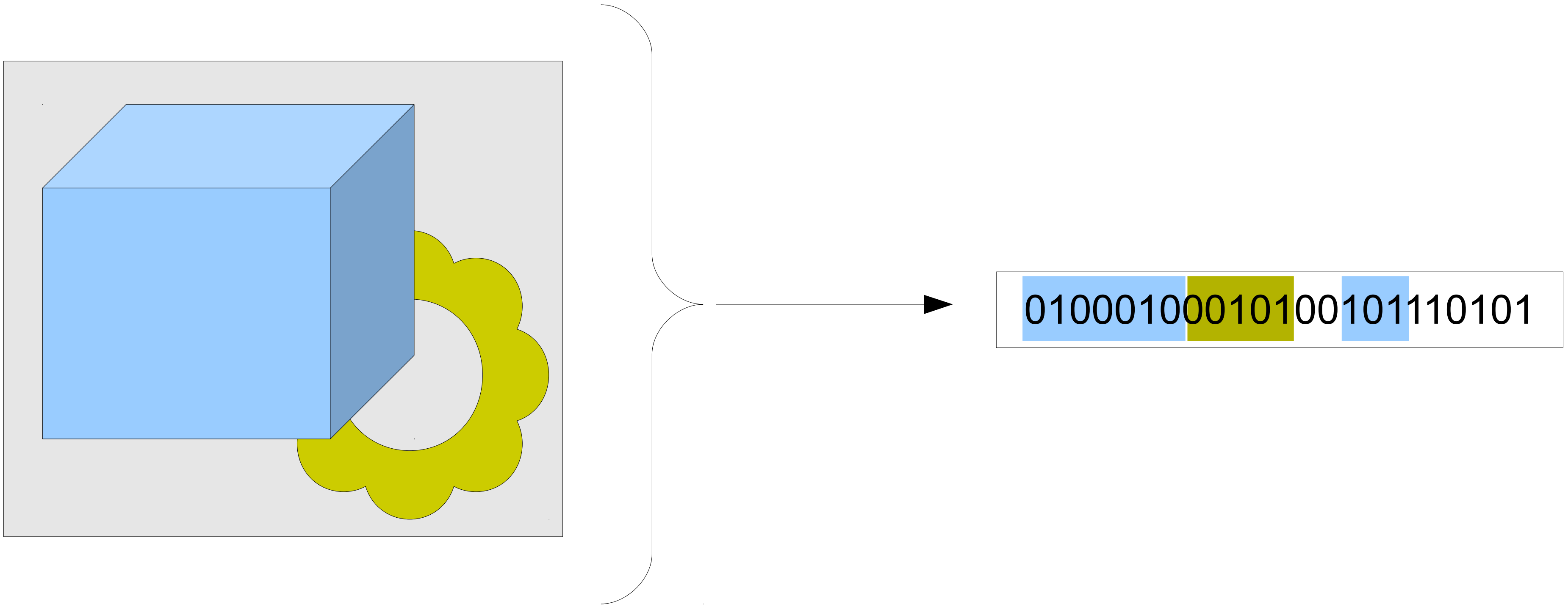


Figure 2-1: A formalization of the recognition problem with bounded error: Find the largest subset of points m_i on the left such that there exists a transformation T (translation, rotation, scale) of the plane that maps the points into the error bounds $B_j = b_j + E_j$ given by image points b_j together with error bounds given as sets E_j on the right.

complexity

- **RANSAC algorithm**
  - model, image: collection of feature points
  - pick 3 model, 3 image points
  - compute transformation
  - evaluate remaining points
- **consider algorithm in presence of clutter**
  - probability that image points are all from object

signal vs noise



Cluttered Scene

Feature Vector

error rate

- **consider simple binary features**
- **two kinds of features**
  - binary features derived from object – give information about the object
  - binary features derived from other objects, background – random, no information about the object
- **consider extreme case**
  - very cluttered, most binary features will be "on" most of the time
  - detection of feature from object doesn't give a lot of additional information

CONCEPTS

common approaches to attention

- **sliding window methods**
  - primarily used with classifiers
  - move rectangular windows across the input image
  - treat the background between outline and window as noise
- **statistical grouping methods**
  - primarily used with feature-based recognition
  - identify feature points / line segments
  - compute statistics of which segments likely belong together
  - deal with multiple possibilities through sampling

approaches to saliency

- **saliency computations**
  - manually constructed or learned "saliency detector"
  - evaluate across the entire image
  - find salient regions within the saliency map
  - apply object recognition method to salient region

STATISTICAL GROUPING

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1177 November 1989

GROUPING FOR RECOGNITION

David W. Jacobs

Abstract: This paper presents a new method of grouping edges in order to recognize objects. This grouping method succeeds on images of both two- and three-dimensional objects. So that the recognition system can consider first the collections of edges most likely to lead to the correct recognition of objects, we order groups of edges based on the likelihood that a single object produced them. The grouping module estimates this likelihood using the distance that separates edges and their relative orientation. This ordering greatly reduces the amount of computation required to locate objects. Surprisingly, in some circumstances grouping can also improve the accuracy of a recognition system. We test the grouping system in two ways. First, we use it in a recognition system that handles libraries of two-dimensional, polygonal objects. Second, we show comparable performance of the grouping system on images of two- and three-dimensional objects. This provides evidence that the grouping system could produce significant improvements in the performance of a three-dimensional recognition system.

object recognition

- framework

- object recognition based on outlines
- outlines may be broken, overlapping
- outlines represented as line segments
- many objects present in images

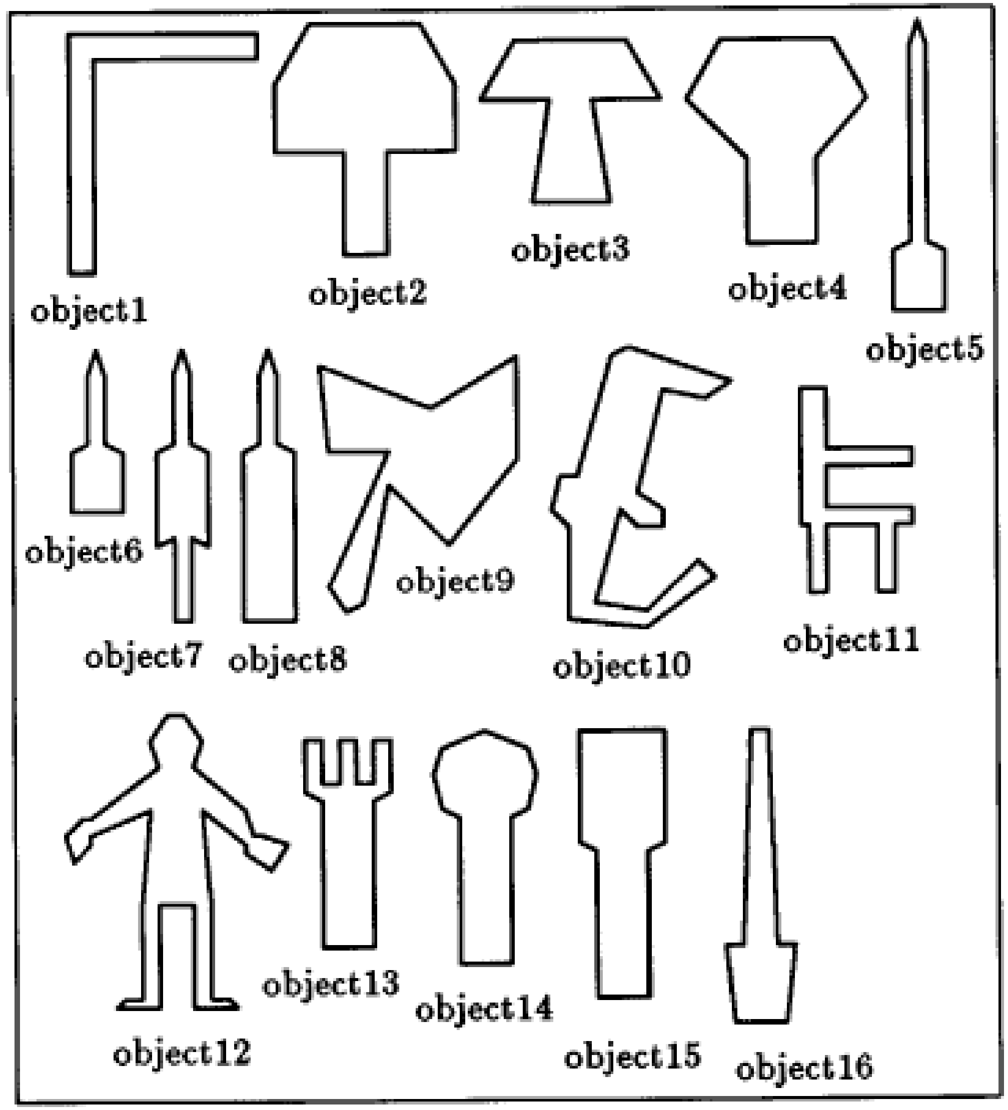


Figure 13: The perimeters of the objects used in tests. The first set of tests used objects 3, 4, 8, 9, 10, and 14. The second and third sets of tests used all sixteen objects.

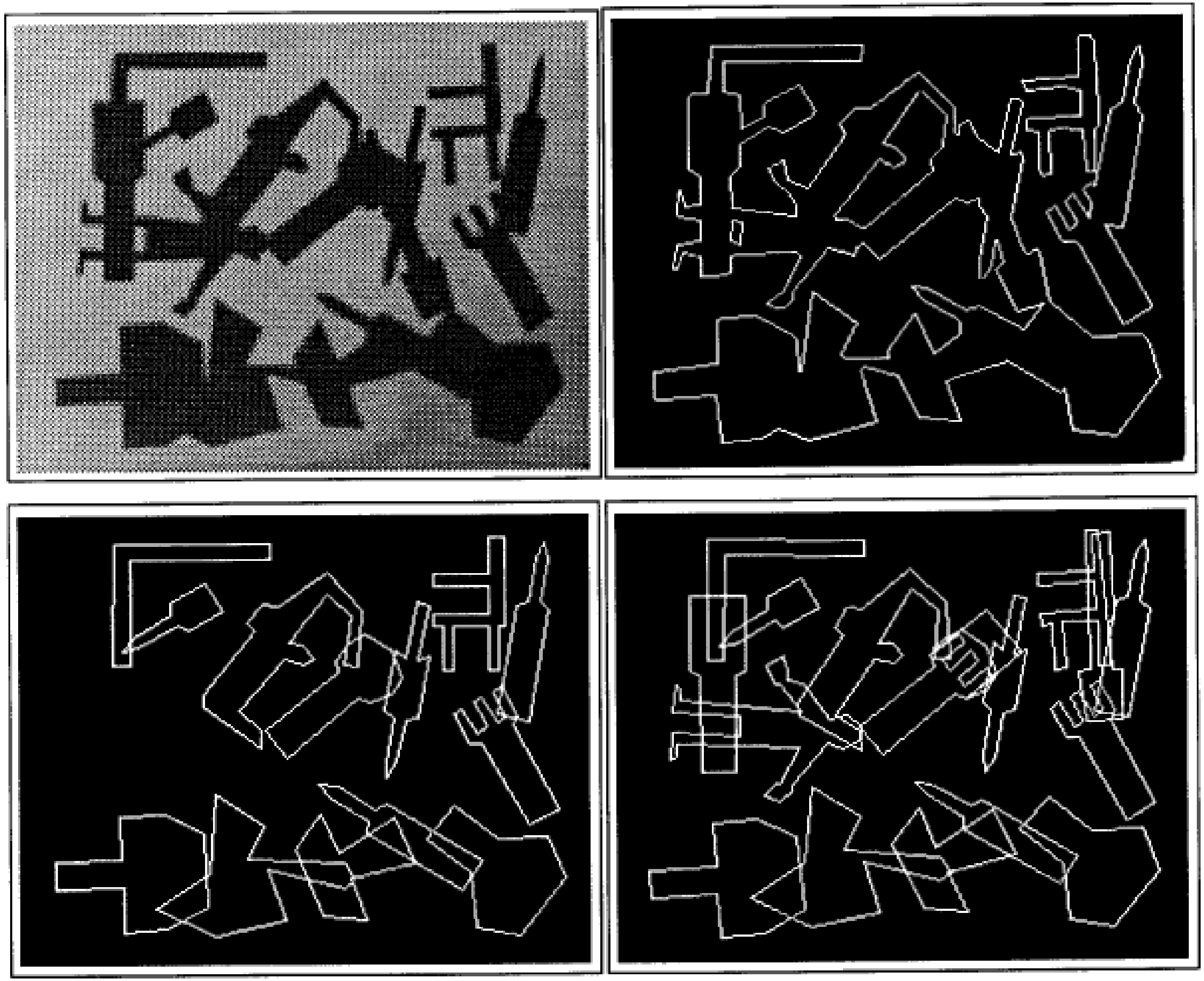


Figure 15: An image from the third set of tests. Again, the picture on the lower left shows the objects GROPER found, and the picture in the lower right shows SEARCHER's finds. Three hypotheses are overlaid where SEARCHER found three models that explained the same edges. 50% of each object was accounted for, with maximum errors of 3 pixels and $sin \frac{\pi}{15}$ radians.

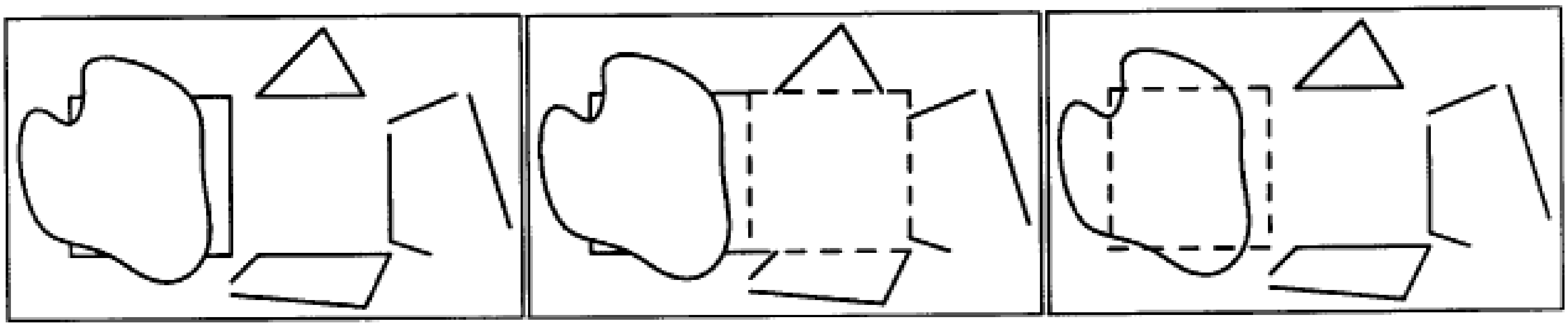


Figure 17: On the left, an hypothetical scene. In the middle, dashed lines show an incorrect hypothesis that accounts for most of a square's perimeter. On the right, an hypothesis that seems more likely to be correct.

The details of edge relationships matter...

approach

- find convex sequences of edge segments
- take pairs of convex groups
- parameterize their relationship
- compute densities for the parameters
- perform a best-first search guided by probability

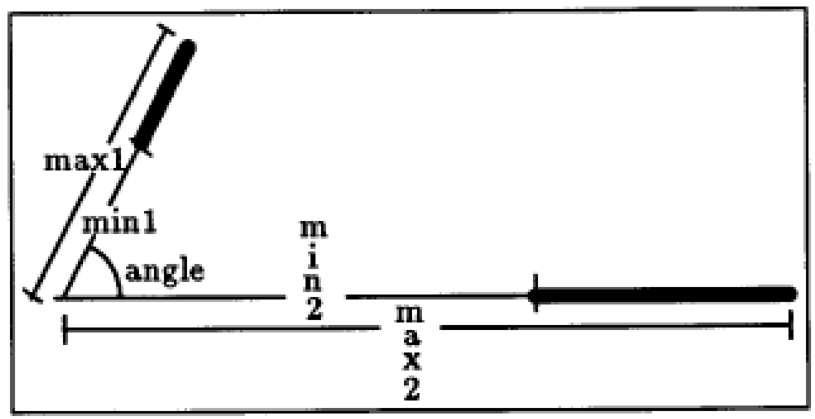
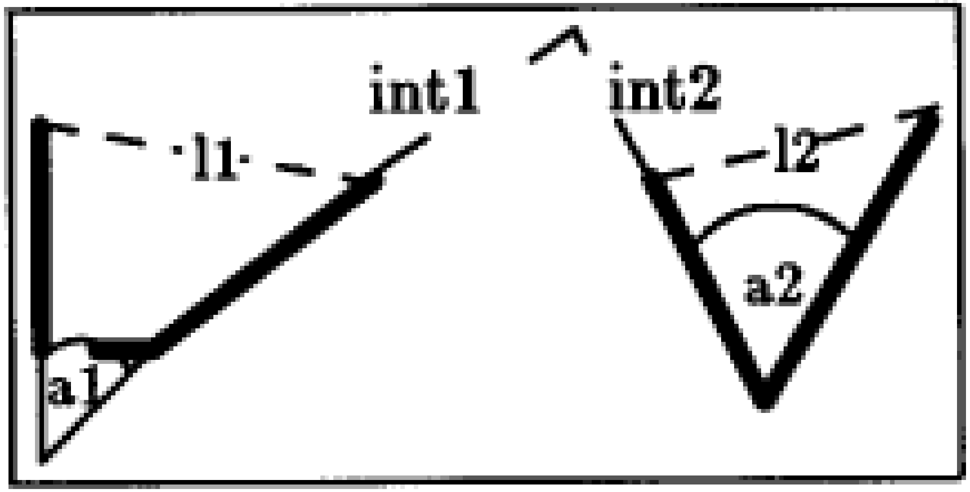


Figure 12: Two edges are in bold. Five parameters describe their relationship.



$$P(type_i|adj, d, l_1, l_2, a_1, a_2, O_1 = O_2)$$

$$= \frac{P(type_i|adj, d, l_1, l_2, a_1, a_2, O_1 \neq O_2)}{P(type_1|adj, d, l_1, l_2, a_1, a_2, O_1 \neq O_2) + P(type_2|adj, d, l_1, l_2, a_1, a_2, O_1 \neq O_2)}$$

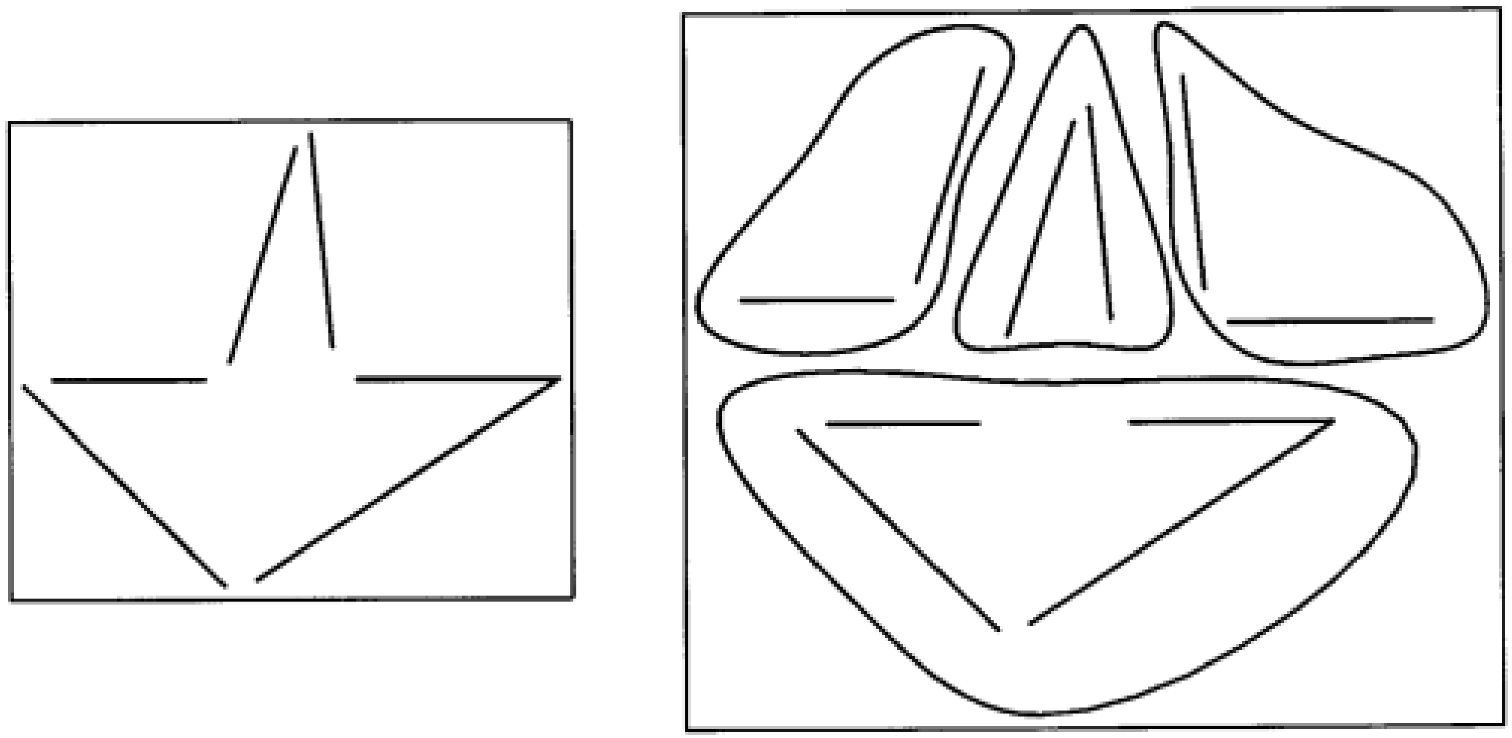


Figure 5: Left: some straight lines. Right: the convex groups they form, circled, and offset slightly from their original position.

non-accidental properties

- **randomly placed line segments always have some kind of relationship (angle, distance)**
- **line segments that come from the same object have a different distribution of angles and distances**

categorize by types of relationships

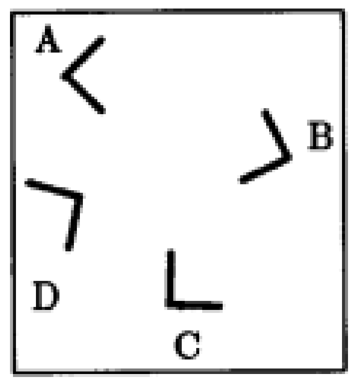


Figure 10: A and B have a $type_1$ relationship, and seem more likely to come from the same object than A and C, which are the same, but have a $type_2$ relationship. These go better together than A and D, which have a $type_3$ orientation.

distributions for types of relationships

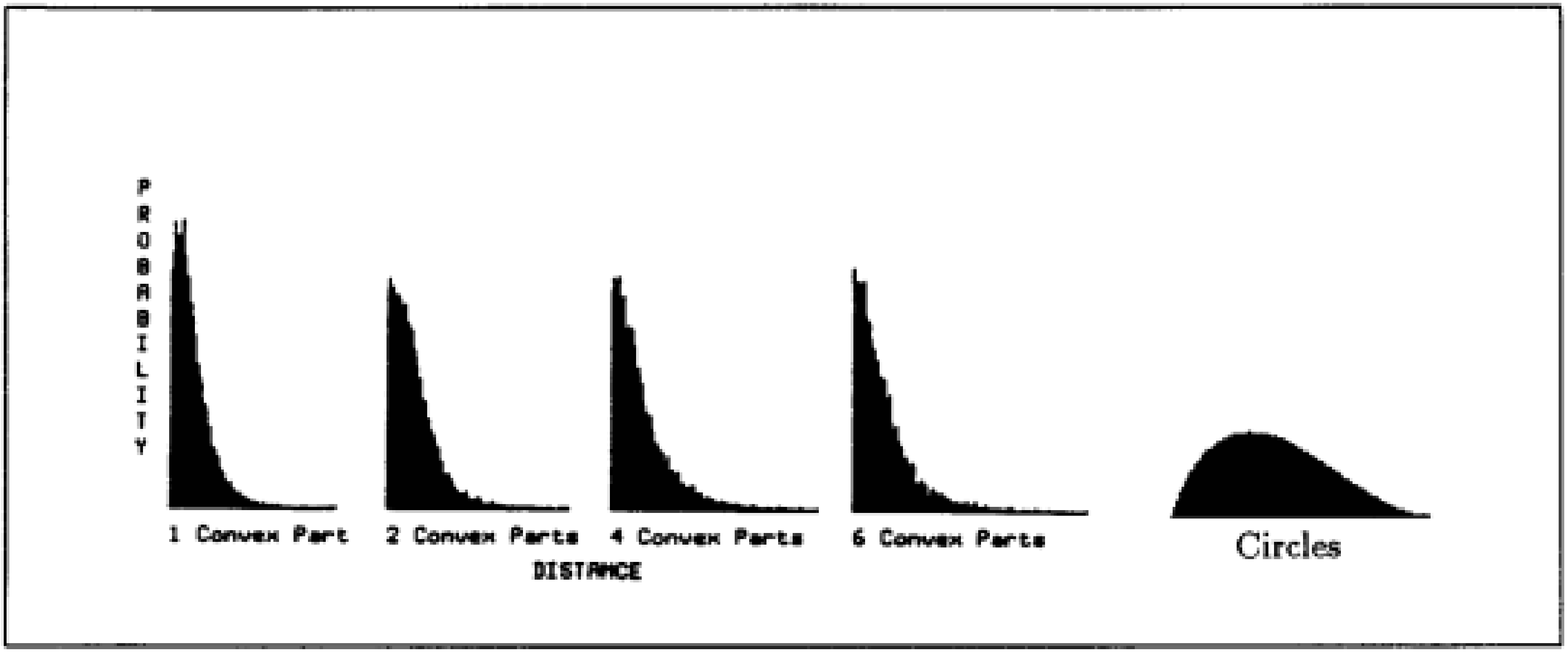


Figure 6: The four graphs on the left show the lengths of occlusions that occur from randomly intersecting random objects. On the right is the distribution resulting from randomly intersecting circles.

GROPER

- **form simple groups**
  - close line segment endpoints
  - collection of line segments is convex
- **form complex groups**
  - take simple groups and join together if it is likely that they came from the same object

recognition with GROPER

- compute line segment approximation to img
- form simple convex groups
- form all pairs of simple groups
- order pairs by P(same object)
- match pairs of groups against model database

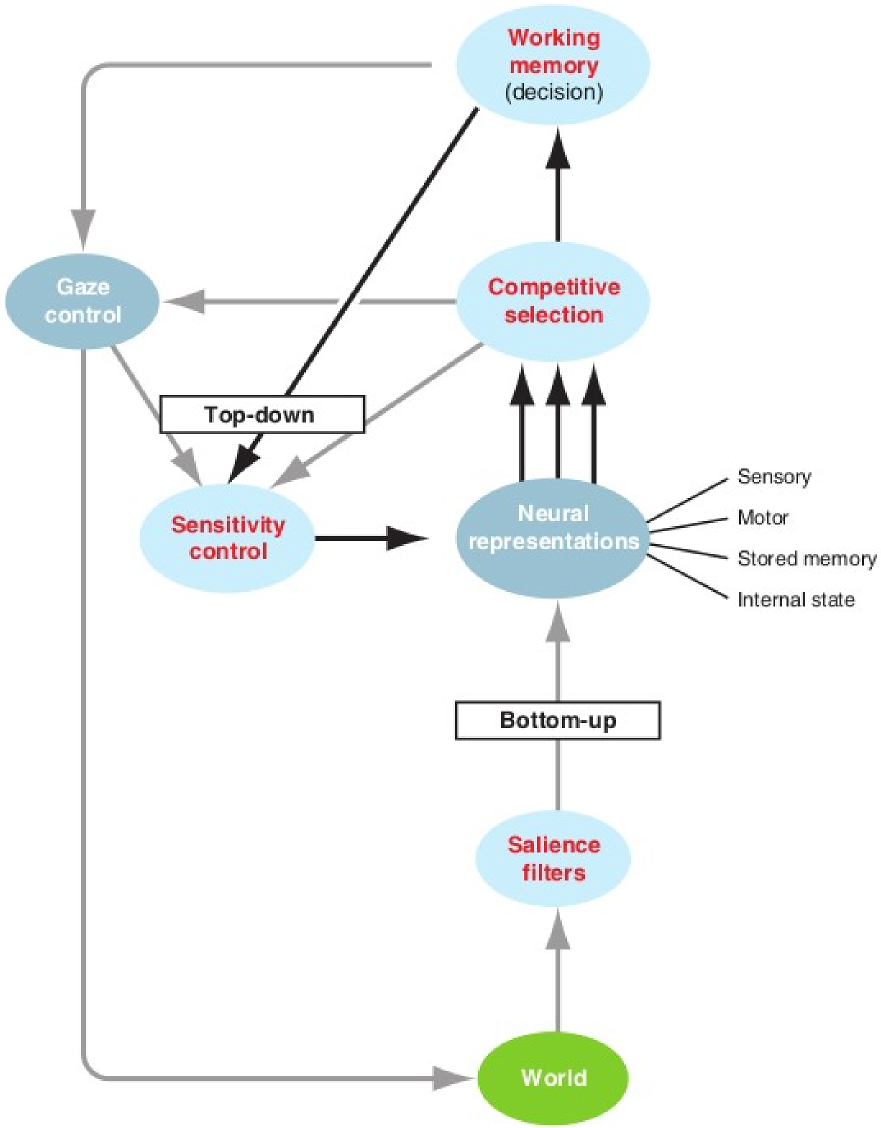
NEURAL MODELS

COMPUTATIONAL MODELLING OF VISUAL ATTENTION

Laurent Itti and Christof Koch‡*

We review recent work on computational models of focal visual attention, with emphasis on the bottom-up, image-based control of attentional deployment. We highlight five important trends that have emerged from the computational literature. First, the perceptual saliency of stimuli critically depends on the surrounding context. Second, a unique 'saliency map' that topographically encodes for stimulus conspicuity over the visual scene has proved to be an efficient and plausible bottom-up control strategy. Third, inhibition-of-return, the process by which the currently attended location is prevented from being attended again, is a crucial element of attentional deployment. Fourth, attention and eye movements tightly interplay, posing computational challenges with respect to the coordinate system used to control attention. And last, scene understanding and object recognition strongly constrain the selection of attended

model of visual attention



components

- **saccades, fixations**
- **modulation of neural responses**
- **control of modulation**
- **integration in working memory**

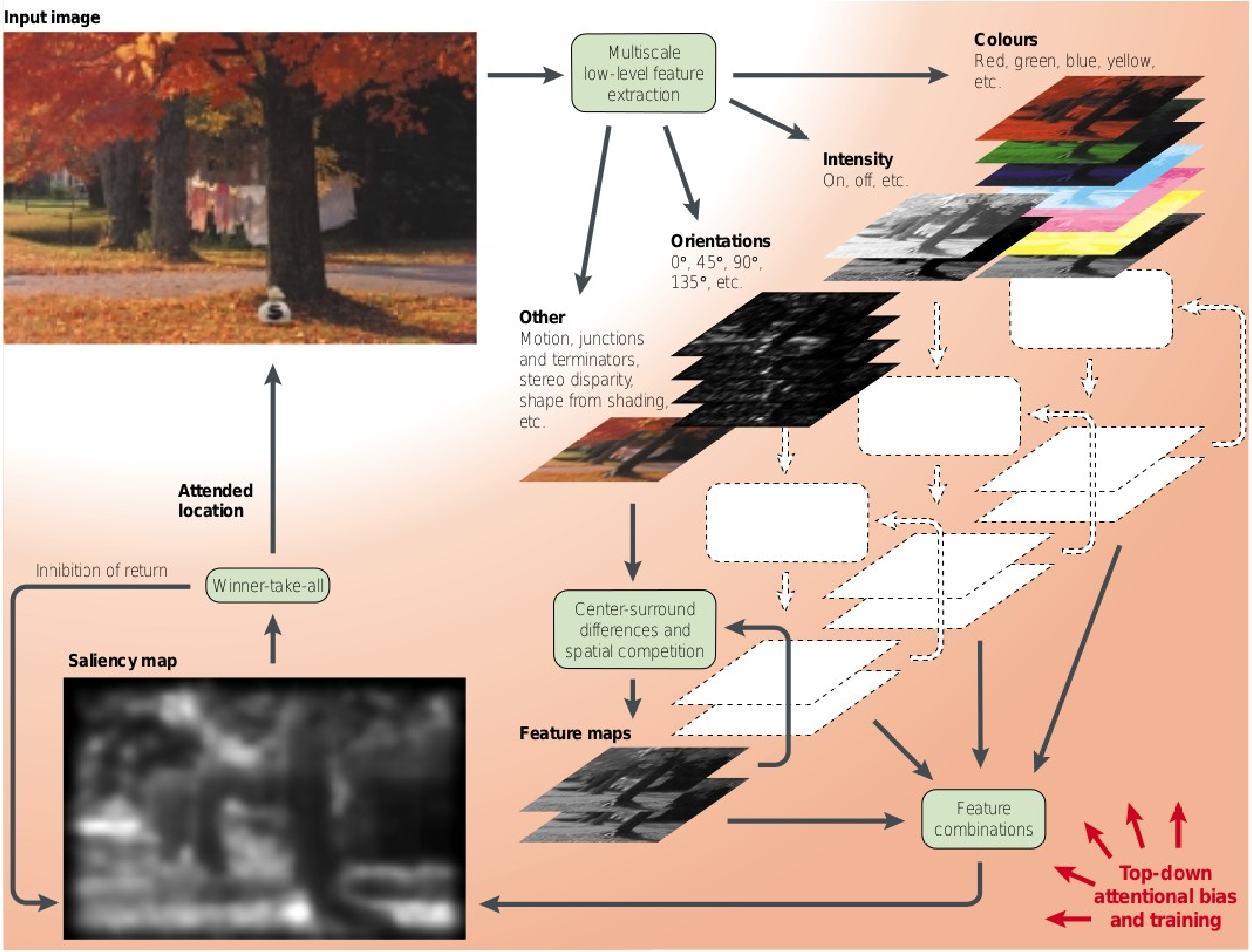
(Knudsen, 2007, Ann. Rev. Neuroscience)

saliency and attention

- **saliency**
  - direct gaze to new locations
  - direct attentional mechanisms without shifting gaze
  - saliency is task-independent
  - computed bottom-up
  - takes 25-50ms per item
- **volitional attention**
  - task dependent
  - deliberate control ("find red, horizontal bars")
- **both mechanisms operate in parallel**

computational model

- **Koch & Ullmann, 1995**
  - saliency map
  - topographic map where intensity represents saliency
- **only models bottom-up component**
  - top-down attention separate concern



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
Center-surround differences and spatial competition
Saliency map
Feature maps
Feature combinations
Top-down attentional bias and training

saliency map

- extract features across visual field
- similar to HMAX model
- find features that are statistically unusual via competition and inhibition
- combine different sources into a single map

attentional selection

- find the highest value in the saliency map
- attend to that location, process visual input
- inhibit the location in the saliency map
- repeat

- "inhibition of return" – widely observed in experiments

attention in recognition

- **MORSEL model**
  - connectionist word recognition model
  - attention selection one word at a time
  - the attended word is then recognized as a whole
- **object recognition**
  - (already discussed)

neural feedback

- **attention modifies low-level perceptual processes**

- **attention...**
  - three-fold increase in orientation sensitivity
  - 20% increase in contrast discrimination

- **explanation**
  - activation of winner-take-all mechanisms within columns (regions)

SUMMARY

attention

- **attention improves both error rates and speed of recognition processes**
- **pre-attentive and top-down mechanisms**
- **components**
  - saliency map – via statistics, symmetry
  - grouping, segmentation – via gestalt principles, statistics
  - task driven – (not well understood yet)
- **when building visual recognition systems**
  - consider implementing saliency, grouping
  - sliding window is simple form of this