RobotCub

Computational models of Visual Attention

Francesco Orabona

LIRA-Lab, DIST
University of Genova – Italy
http://www.liralab.it

Cogsys
Cognitive Systems

From Biology to Robotic

- Biology
  - Biology of visual attention
  - Attention like a filter for the brain
  - Computational models
- Applications
  - To guide a active camera
  - Before of a recognition system

Why Model Visual Attention?

The world is sensed continuously instead of maintaining and updating some complicated internal model:

"World is its own best model".

Moreover, we constantly select a part of the (elaborated) information coming through our senses.

A model of the visual attention is needed:

- To build robust and adaptive active vision systems
- To guide object recognition systems in real world
- To understand how the brain works by building computational models

Just an example…

- Visual search and matching are NP-complete problems
- N x N image, 8 bit and a set of C prototypes of objects
- $8^{N\hat{}2}$ possible images (but not all the possible images are plausible…)
- 2^C combination of prototypes.
- Each image has a possible interpretation
- $8^{N\hat{}2} \cdot 2^C$ possible correspondences
- An image 128 x 128 and 10 prototypes => about $10^{14,800}$ possibilities.
- The problem is the ***dimensionality*** of the images

RobotCub

Active vision is the trick!

- Wide field of view
- High resolution (in the center)
- Translation invariance
- One item at the time

- A mechanism to guide to eyes is needed: the visual attention

Cogsys
Cognitive Systems

Visual attention

- Only a small fraction of the information registered by the visual system at any given time reaches levels of processing that directly influence behavior.
- Visual attention controls access to this privileged level and ensures that the selected information is relevant to behavioral priorities and objectives.
- Thus, visual attention is closely linked to visual awareness and introspection.

RobotCub

The illusion of a perfect vision (1)

Cogsys
Cognitive Systems

RobotCub

The illusion of a perfect vision (2)

Cogsys
Cognitive Systems

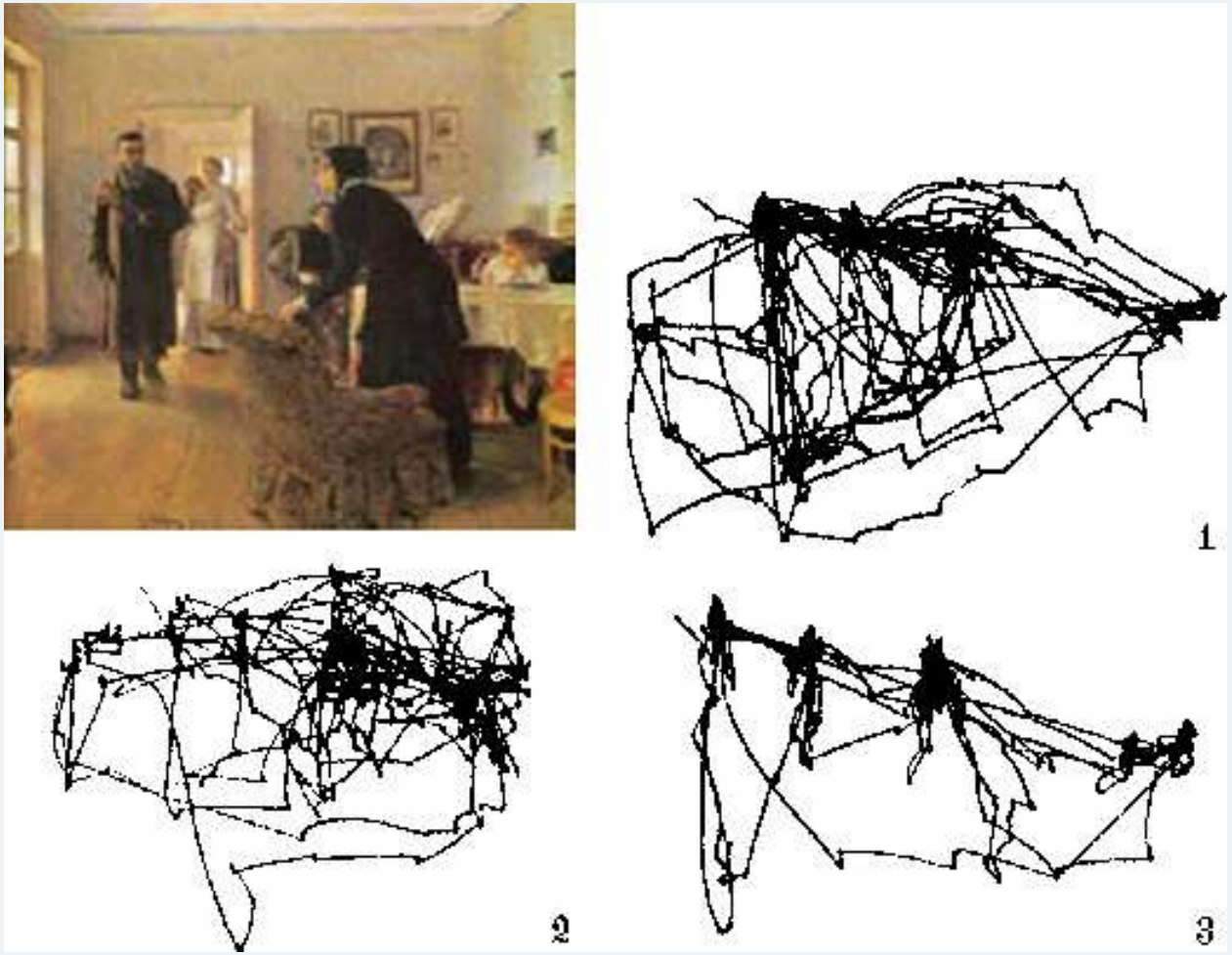
The spotlight paradigm

- *A spotlight* that enhances important information.
- The attention can diverge from the direction of the gaze

Attention & Saccades

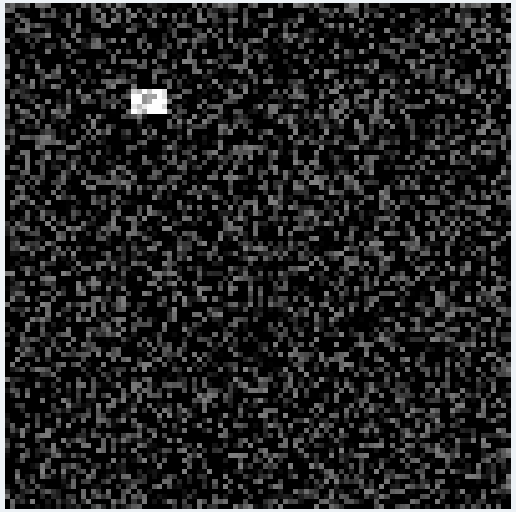
- Covert attention = shifts of the focus of attention in the absence of eye movements.
- Overt attention = shifts of the FOA with the presence of eye movements.
- We move our eyes 3-5 times per second (that is, 150,000 to 250,000 times every day), to align locations of interest with our foveas.
- Overt and covert attention are closely related, as revealed by psychophysical, physiological and imaging studies.
- Experiments suggest that covert attention is deployed to the endpoint of a saccade, in preparation for it.

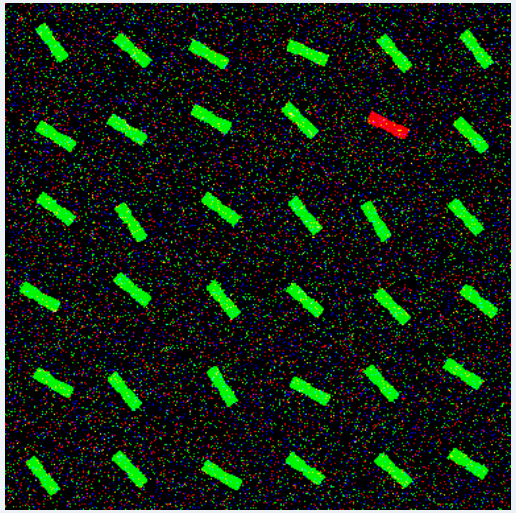
Bottom-Up & Top-Down cue (1)

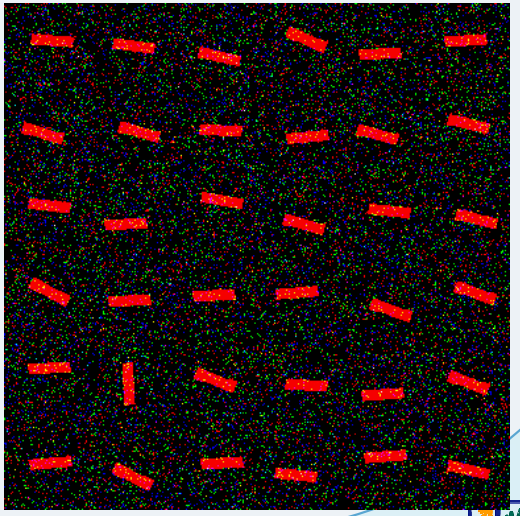


(Yarbus, 1967)

Bottom-Up & Top-Down cue (2)







Bottom-Up & Top-Down cue (3)

Bottom-Up

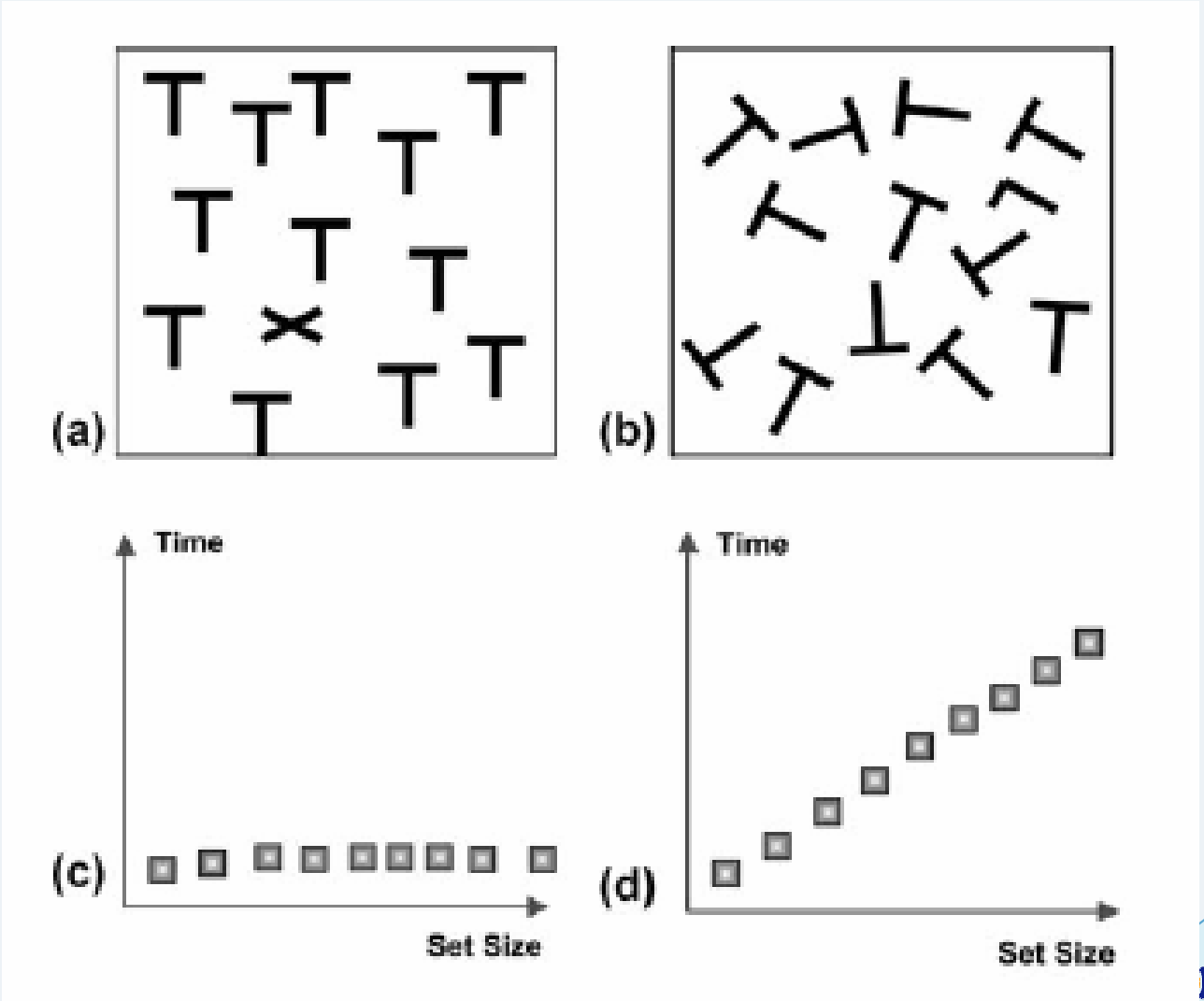
- Stimulus-driven
- For exploration
- Difficult to define (and validate?)

Top-Down

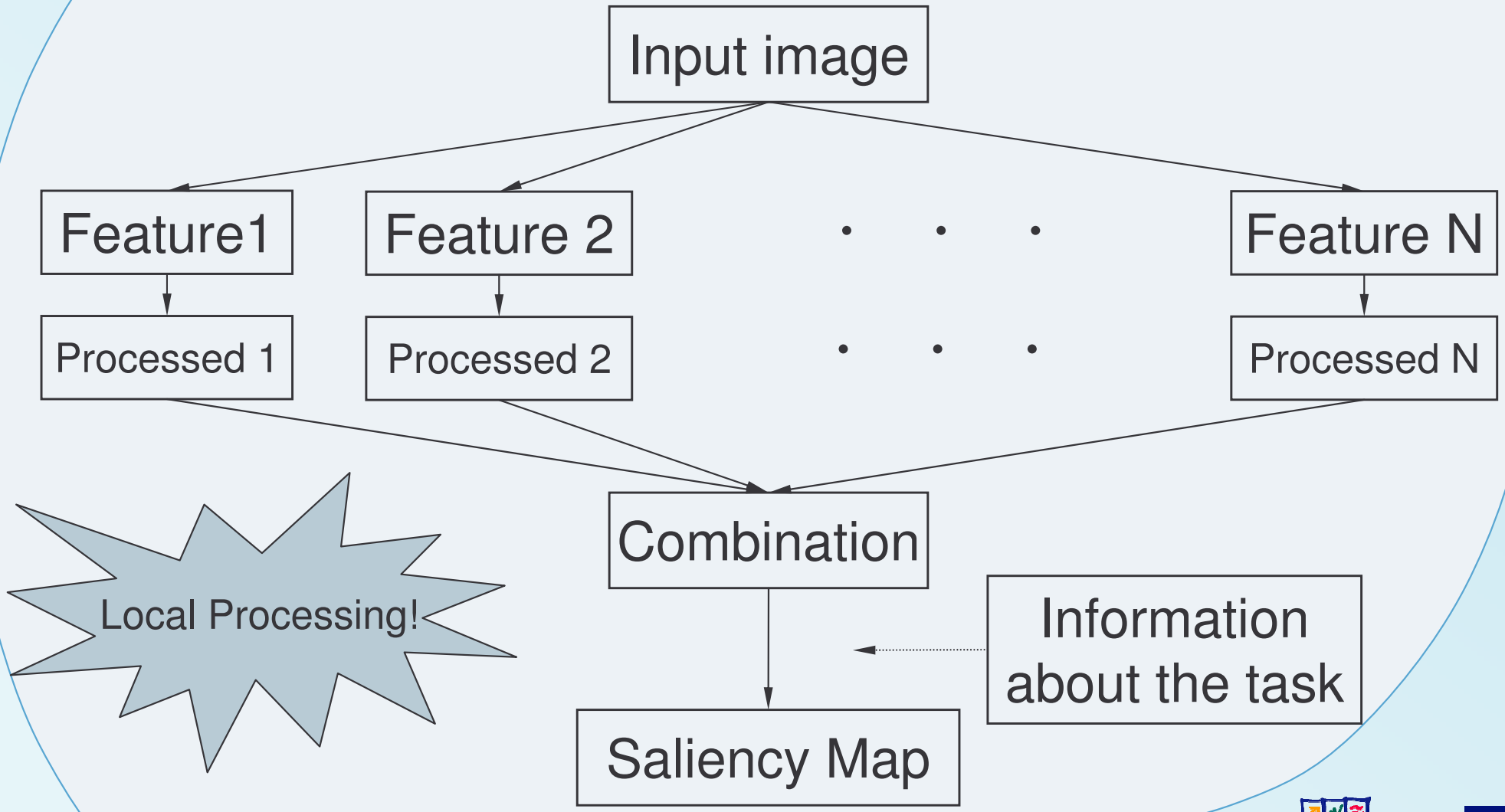
- User-driven
- Need some knowledge
- Test with visual search experiment

Do purely Bottom-Up cues really exist?
Maybe it is an implicit Top-Down modulation, a scan of the scene to fast gather as much information as possible

Parallel and serial visual search

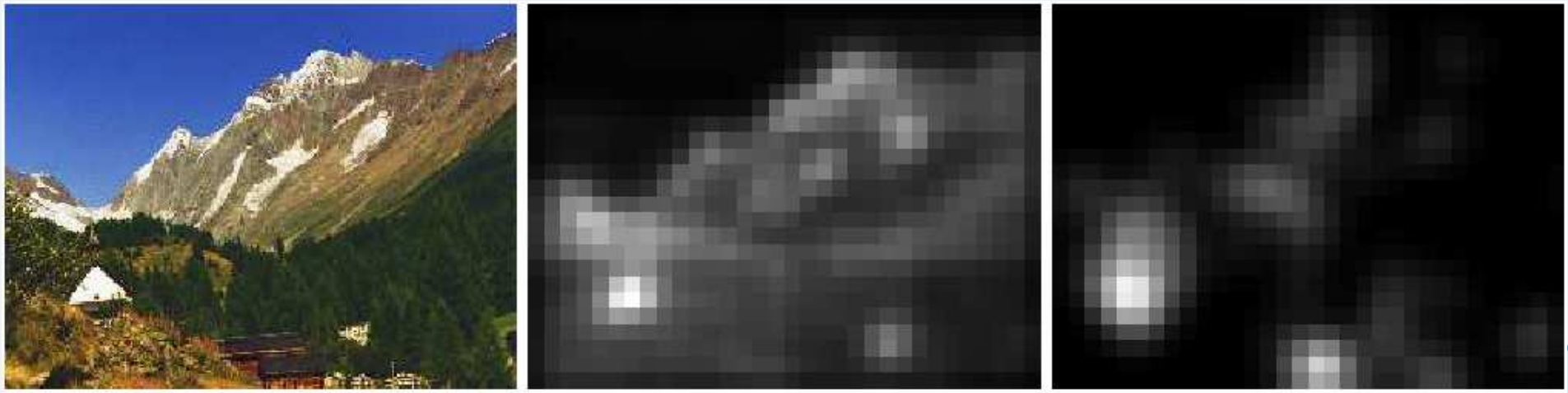


Feature Integration Theory



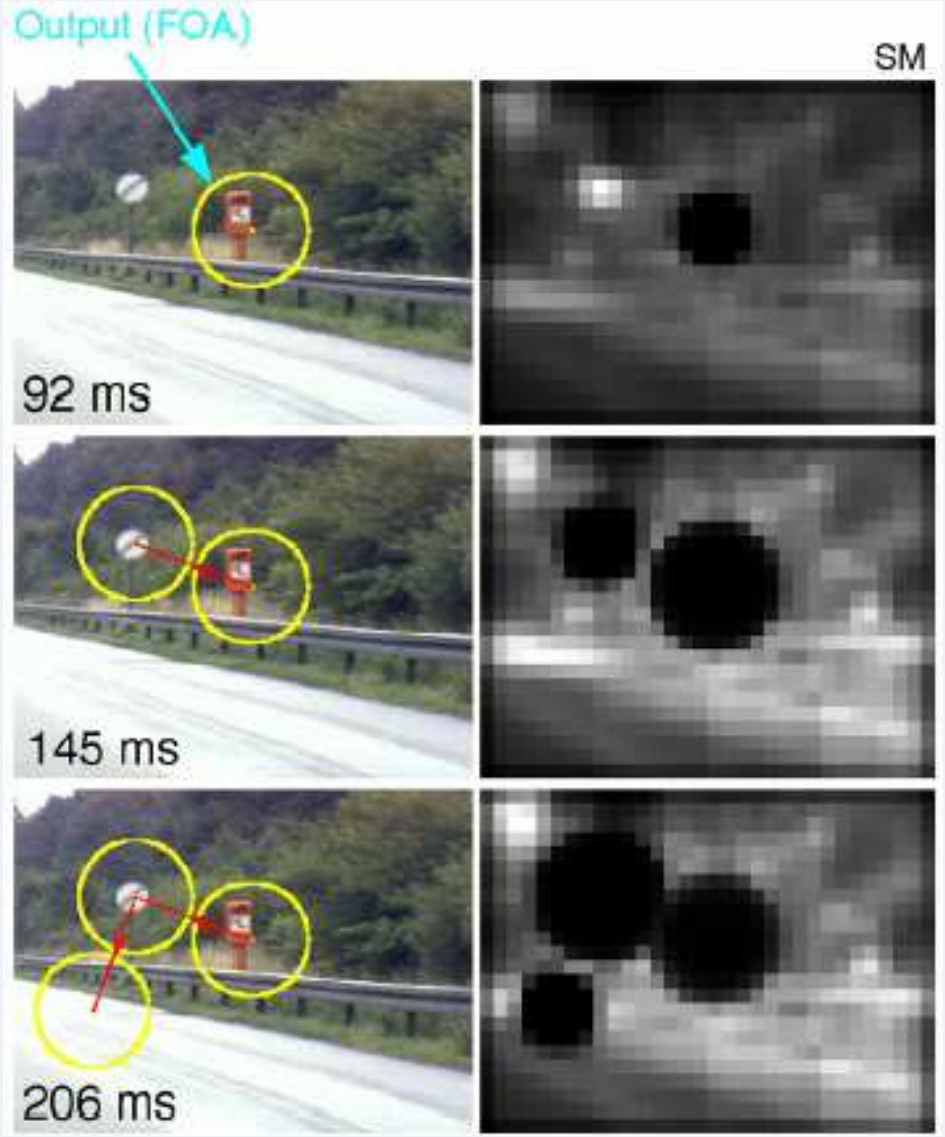
(Treisman, 1980; Wolfe 1994)

Saliency maps

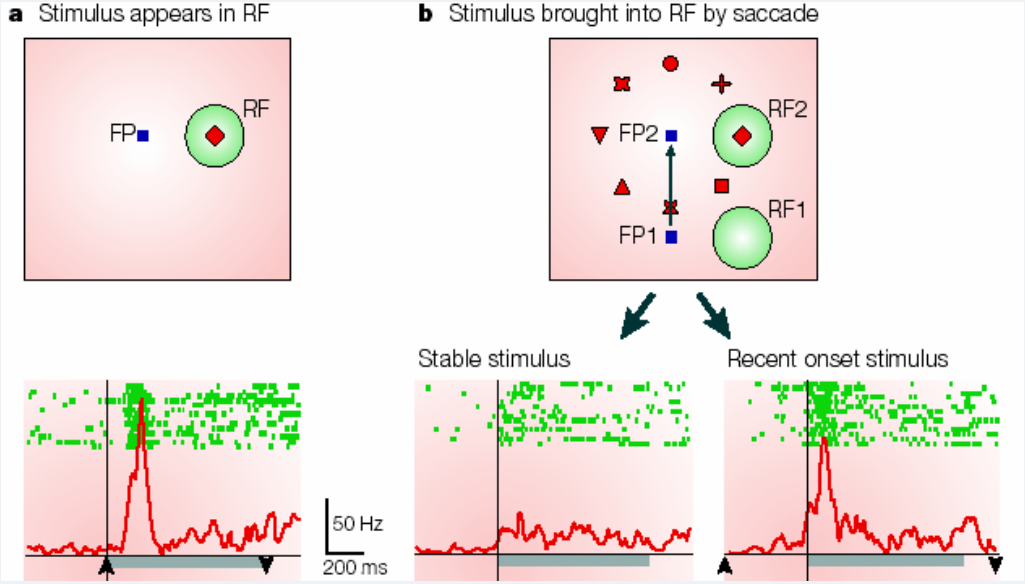


Color image *Computational map* *Human map*

IOR



Saliency maps in the brain

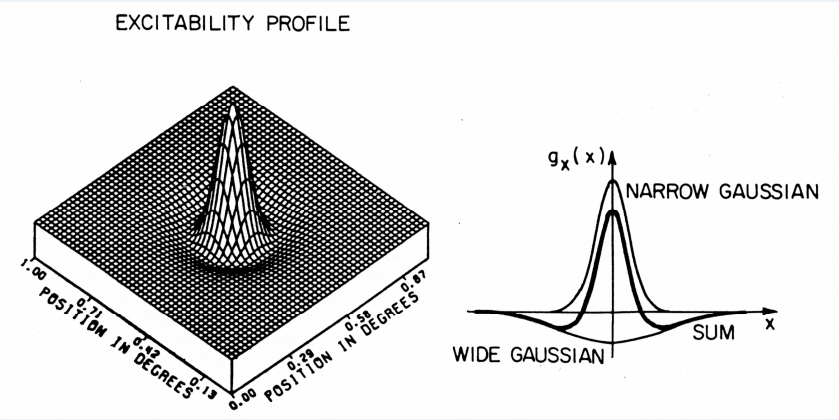


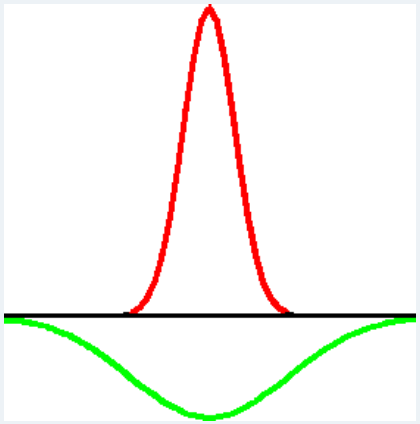
- Gottlieb et al. 1998, recording from the lateral intraparietal sulcus of the awake monkey, found neurons that responded to visual stimuli only when those stimuli were made salient.
- In the control condition, with the eyes stable at the fixation point (FP), a stimulus is presented in the RF of the neuron being recorded from. The response elicited by this stimulus could be simply visual, or indicate the saliency of this stimulus suddenly appearing in the visual field.
- A stimulus entered the RF through a saccade.A vigorous response was observed only when the stimulus had been made salient shortly before the beginning of the trial (by flashing it on and off while it was still outside the RF of the neuron; 'recent onset' condition).

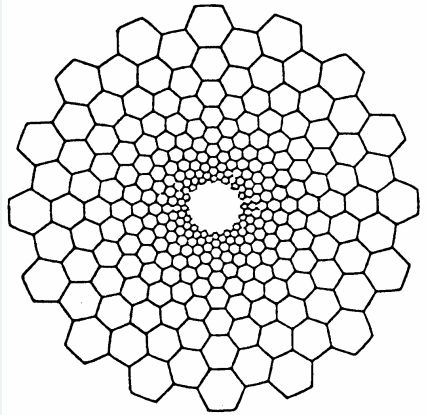
First stage: the Retina

Properties of the receptive fields:

- Linear
- Difference of Gaussians (DoG)
- Chromatic Opponency
- Size depends on the distance from the center (Log-Polar)







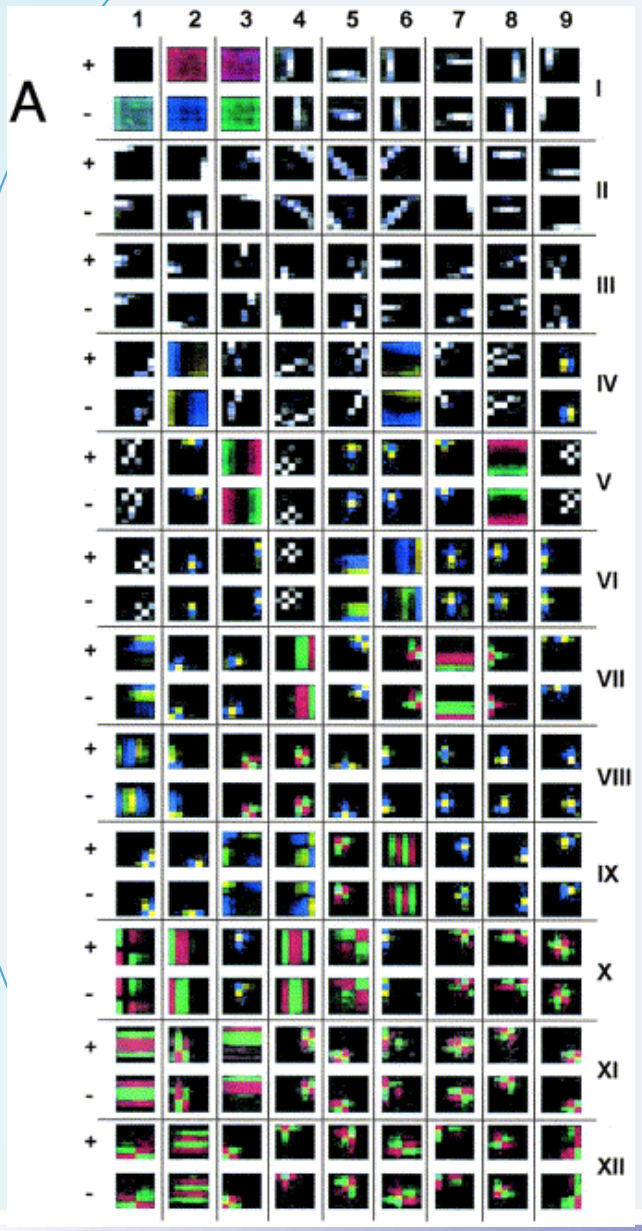
$$R^{+}G^{-}(x,y) = R(x,y) * \frac{\alpha}{2\pi \cdot r^2 k^2 \sigma_c^2} e^{\frac{-(x^2+y^2)}{2k^2 r^2 \sigma_c^2}} - G(x,y) * \frac{\beta}{2\pi \cdot r^2 k^2 \sigma_s^2} e^{\frac{-(x^2+y^2)}{2r^2 k^2 \sigma_s^2}}$$

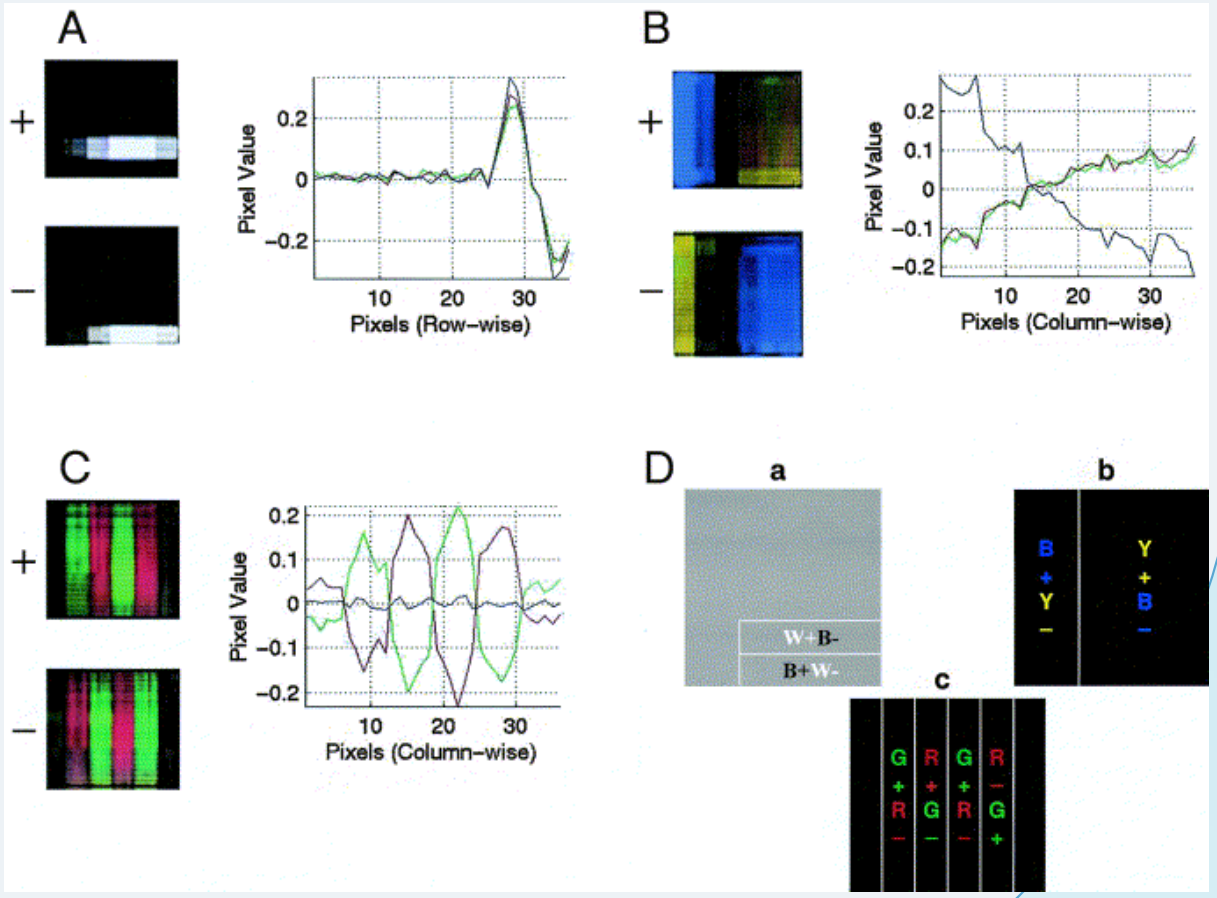
RobotCub

Why difference of gaussians?

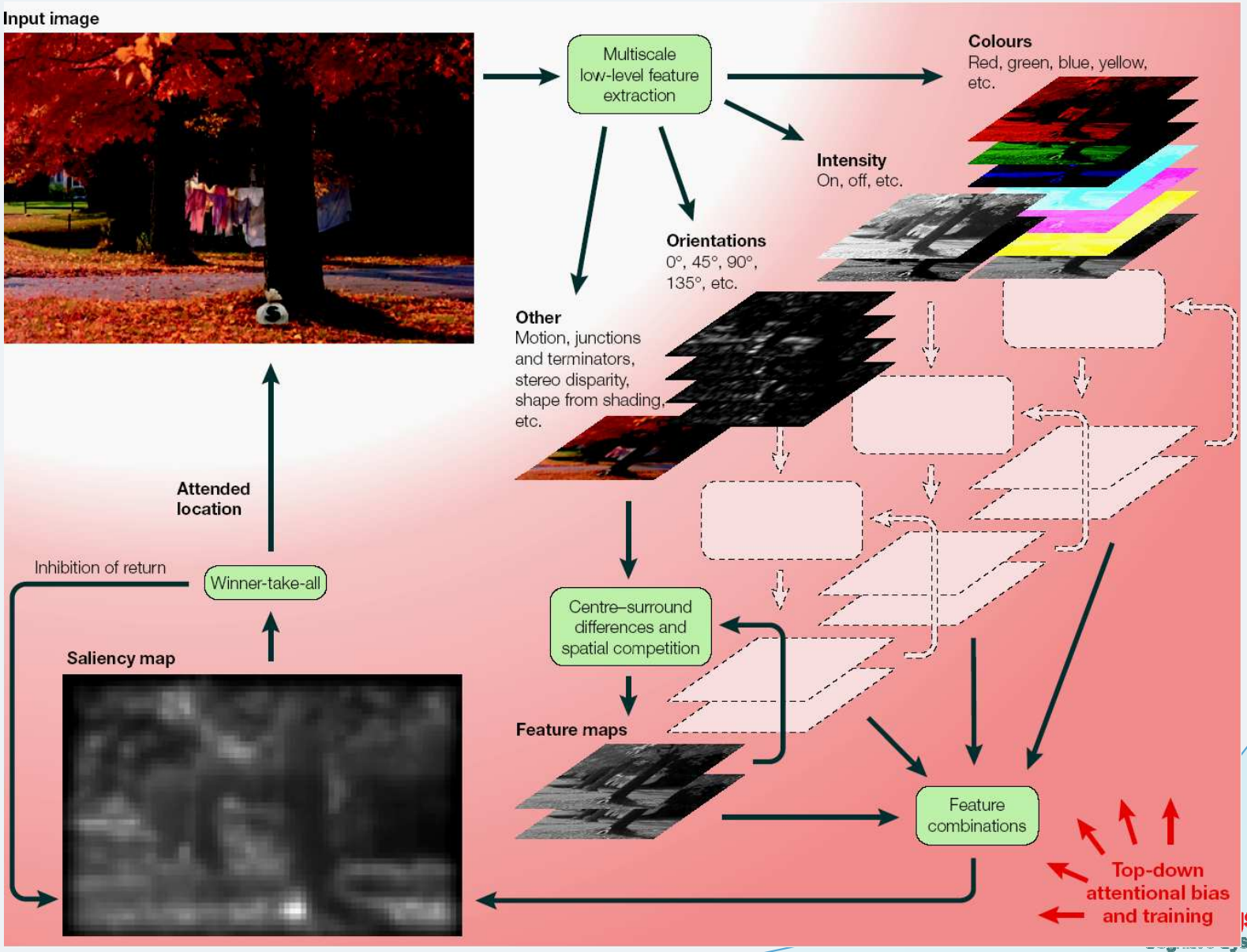
Cogsys
Cognitive Systems

The statistics of the world and the evolution

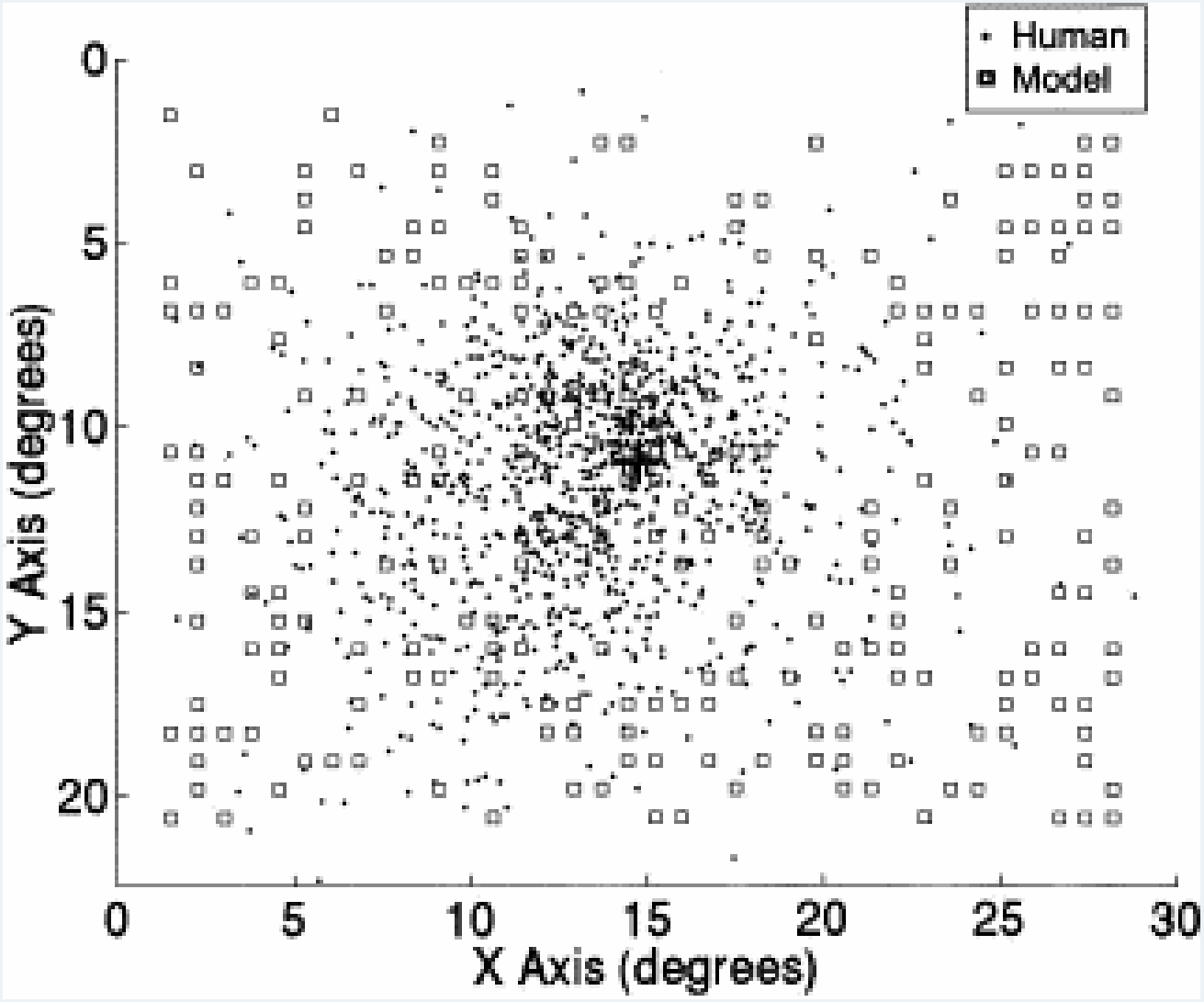




Itti's Model


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

Influence of eccentricity



(D. Parkhurst, K. Law and E. Niebur, 1997)

Log-polar Images (1)

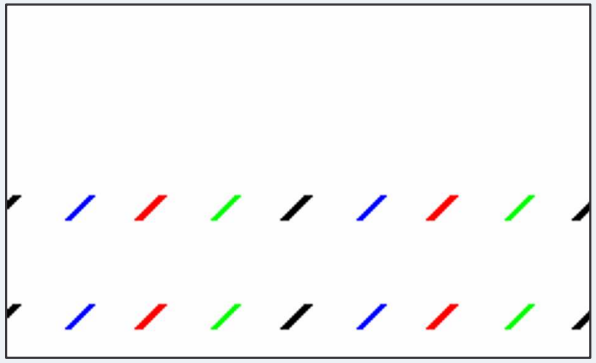
$$\begin{cases} \eta = q \cdot \theta \\ \xi = \log_a \dfrac{\rho}{\rho_0} \end{cases}$$

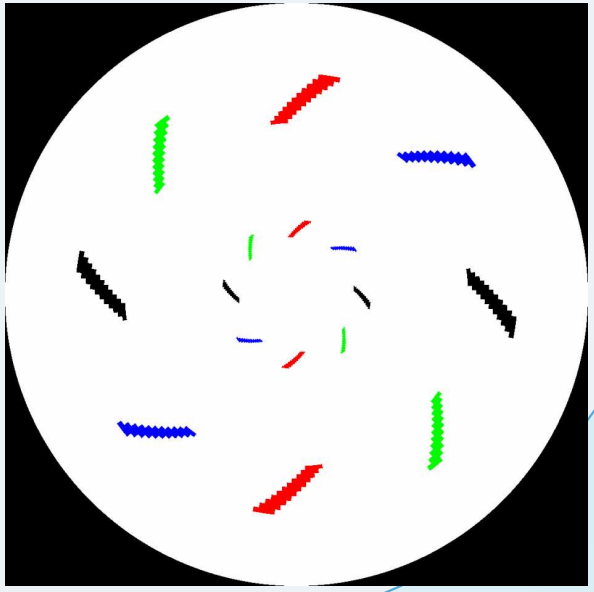
- Mimic photoreceptor distribution and cortical topology
- Conformal transformation, scale and rotation invariant
- Lossy compression of the image
- Influence in visual search task (Enoch 1959, Wolfe 1996)

Log-polar Images (2)

- ✓ Wide field of view
- ✓ Lossy compression (25:1)
- ✓ Space-variant filtering

- × Deformation of the filter
- × Extraction of the orientations



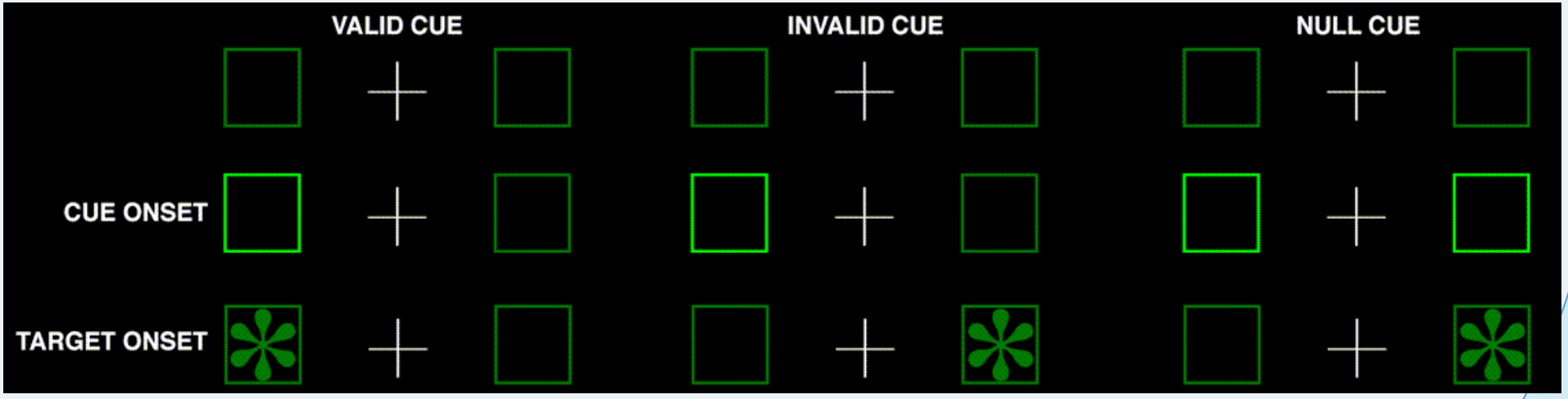


Space-based or Object-based?

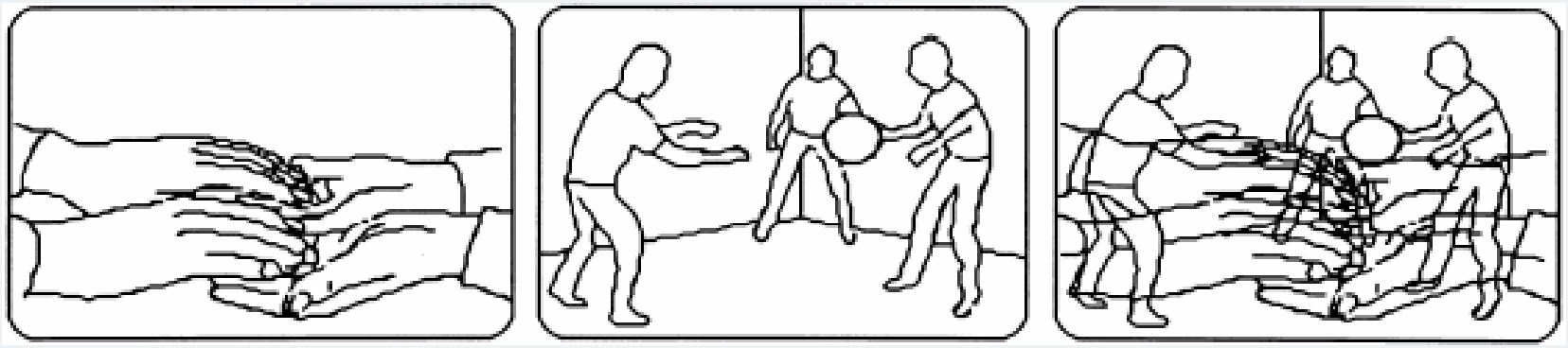
- Space-based models hold that the units of attentional selection are space locations, regions of space
  - Simpler to implement?
  - Doesn't have to define "objects"

- Object-based models hold that the units of attentional selection are "visual objects"
  - Humans tend to fixate objects' centers, not borders
  - Explains "same object advantage", multiple object tracking, etc.

Experiments on spatial attention



Experiments on object-based attention



- The two scenes - the "hand game" and the "ballgame" - are superimposed, and subjects are then induced to attend to only one of them, for example to count the number of times the hands clap each other.
- In this case, subjects fail to perceive incredible sustained events which occur in the other scene, despite the superimposition. (Neisser and Becklen, 1975)

Objects & Proto-Objects

"Proto-objects" or "pre-attentive objects" are a step above the mere localized features, possessing some but not all of the characteristics of objects.

Moreover, through action the robot can go beyond proto-objects, learning and defining its own concept of "objecthood".

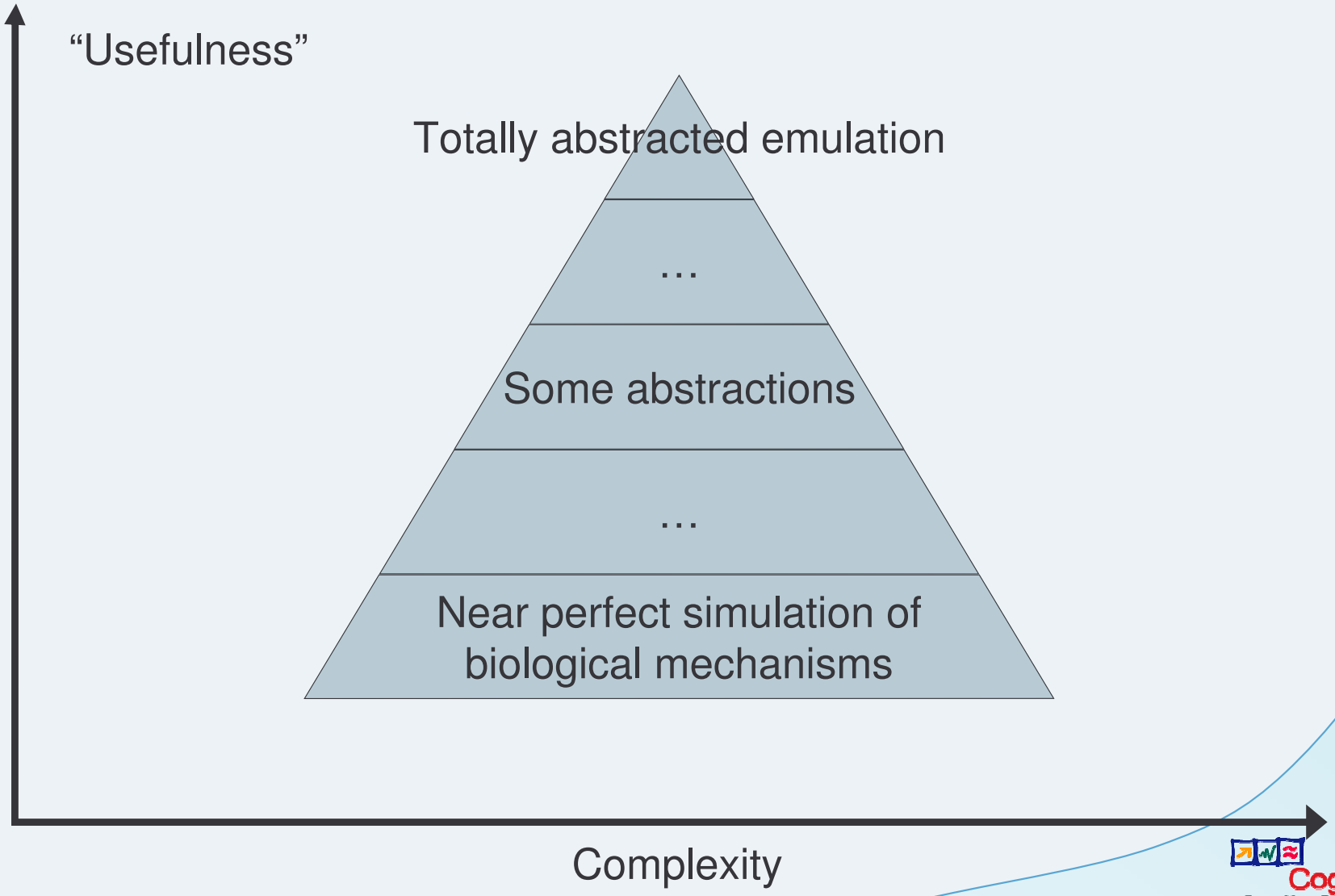
In particular the model of the object is composed of proto-objects and their spatial relations.

Proto-objects in psychology

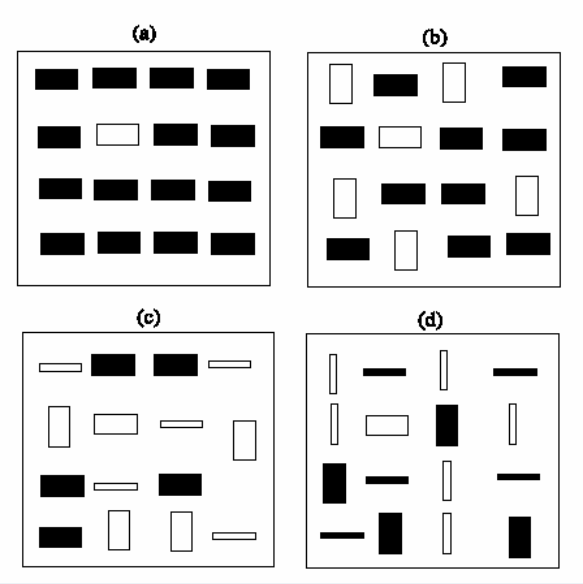
"The concept of a `proto-object' is a general one that has been used by a number of writers [...] in reference to clusters of proximal features that serve as precursors in the detection of real physical objects. What these uses have in common is that they refer to something more than a localized property or `feature' and less than a recognized 3D distal object. Beyond that, the exact nature of a proto-object depends on the theory in question.", Z. W. Pylyshyn, 2001

"It seems clear that a major role of segmentation processes is to distinguish between those parts of the sensory input that are heuristically likely, on the basis of segmentation factors, to correspond to distinct objects (or separate image sources) in the real world, thus constituting `proto-objects'. [...] As the term `mid-level' vision implies, they typically reflect bundles of visual information which are packaged on the basis of properties that go beyond raw image statistics (or primitive `features'), yet which fall short of conceptually recognized entities.", J. Drivera et al., 2001

Computational Models



Different levels of detail



?

RobotCub

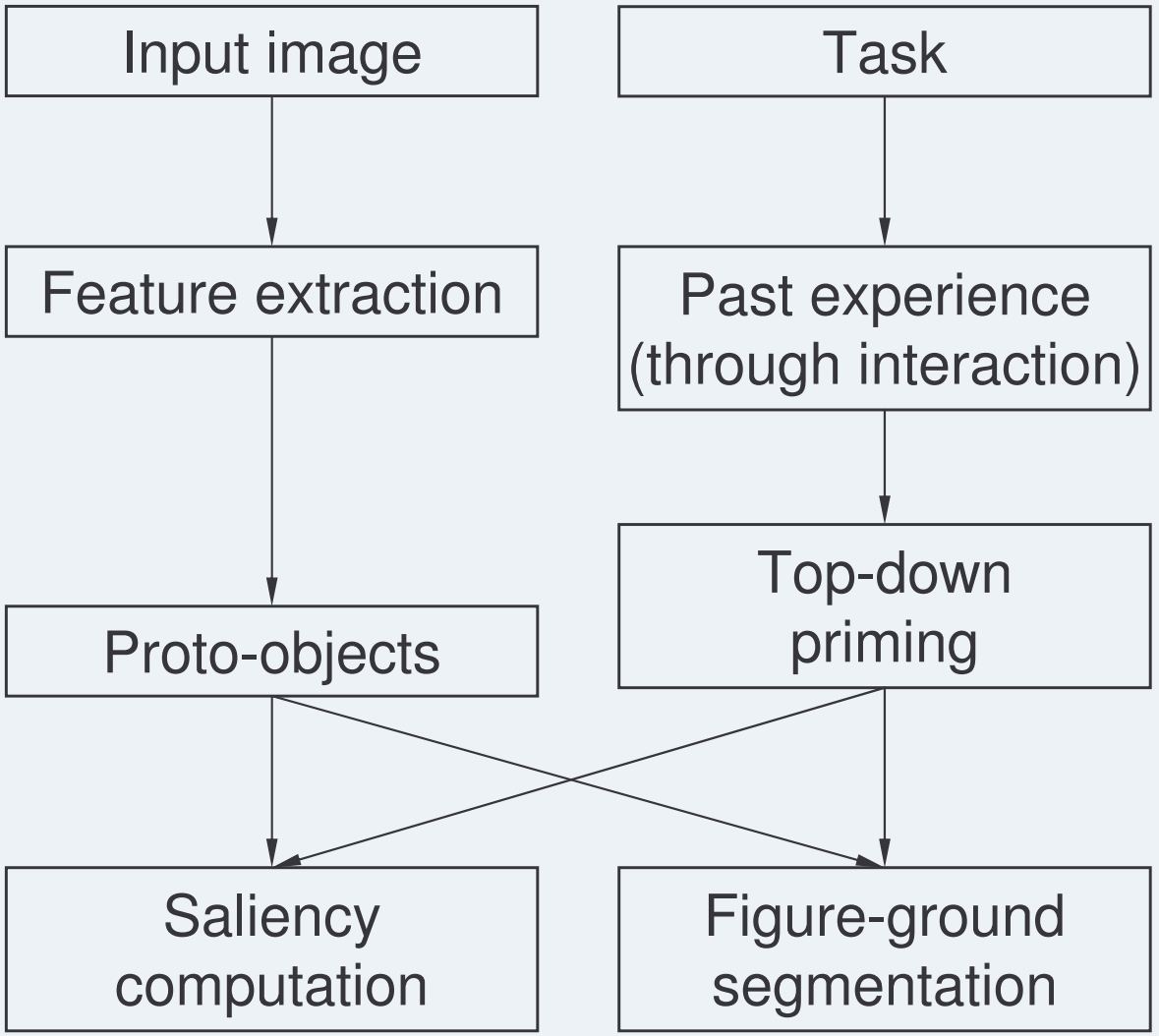
A new model

- Log-polar images
- Object-based model
- Define objects using proto-objects
- Account for Top down priming explicitly

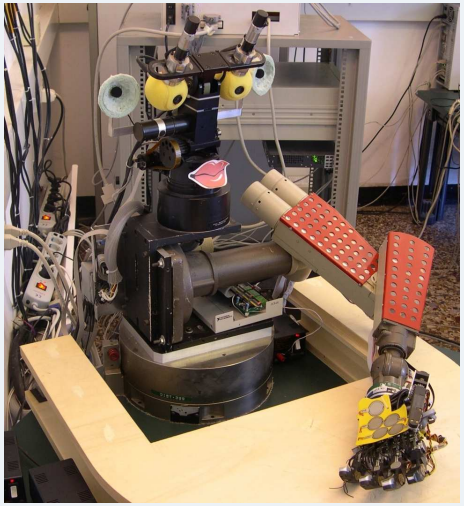
- Medium level of details
- Biological plausibility

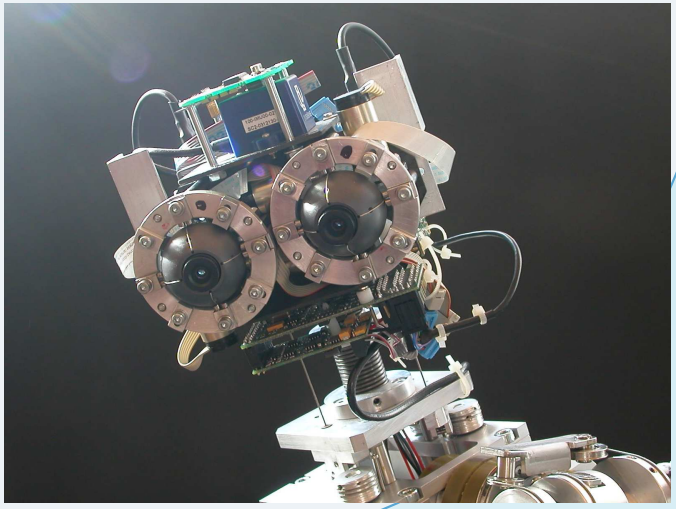
Cogsys
Cognitive Systems

Attention, Segmentation, Action



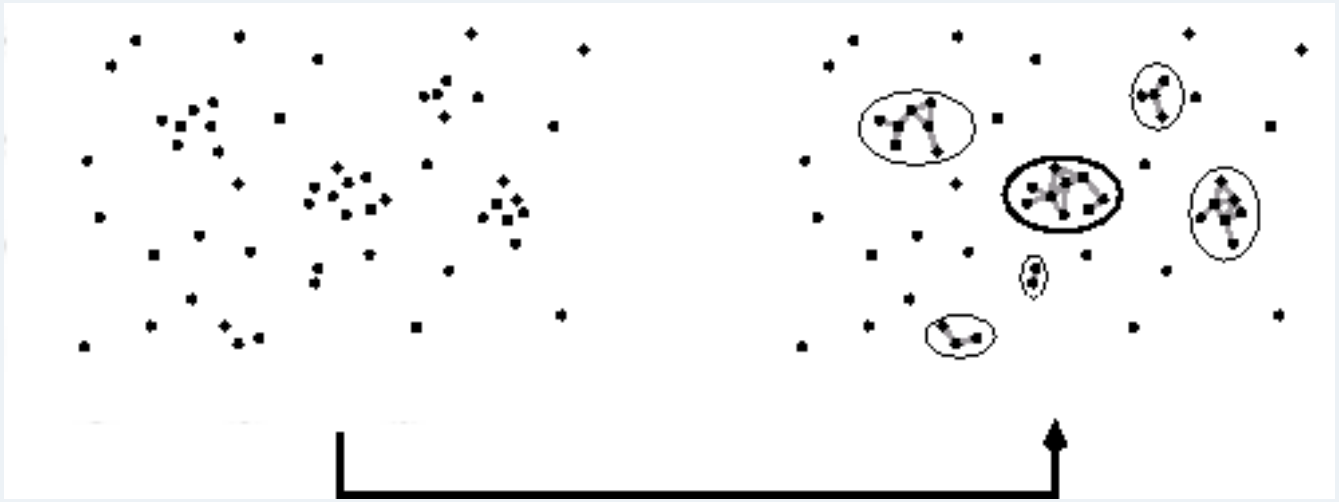
Our Inspiration and our Robots…

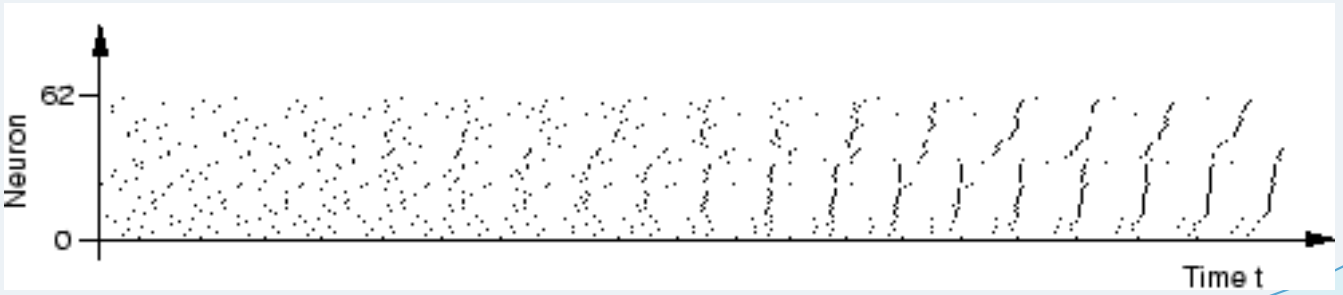




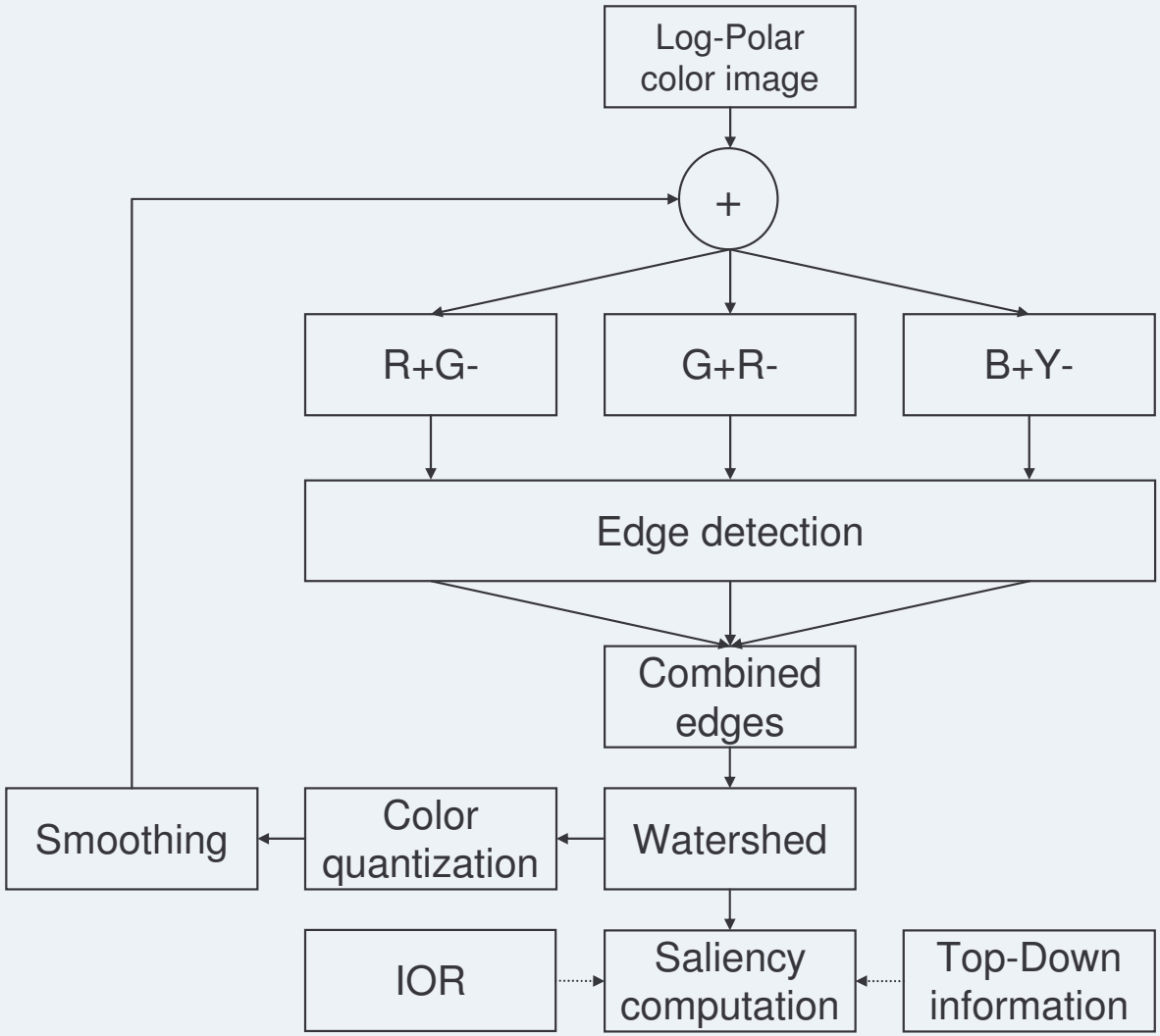
Synchronizations of Visual Cortical Neurons

May serve as the carrier for the observed perceptual grouping phenomenon



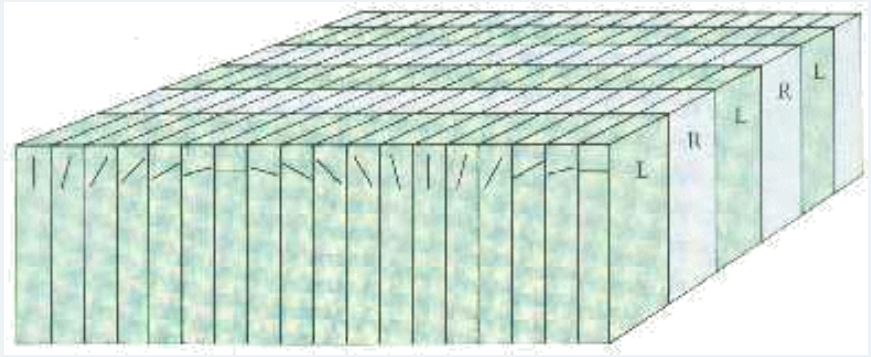


Model Schema



Extraction of Local Orientations

- Inspired to the hypercolumns of V1
- Using a set of generalized Sobel filters on the output of the previous stage



Watershed Transform

- Simulation of the results of the spread of the activation via synchronization
- Currently the proto-objects are defined as uniform closed regions of somewhat constant color (blobs)
- Results doesn't depend on the order in which the points are examined, like in region growing
- Fast (Real time performance)

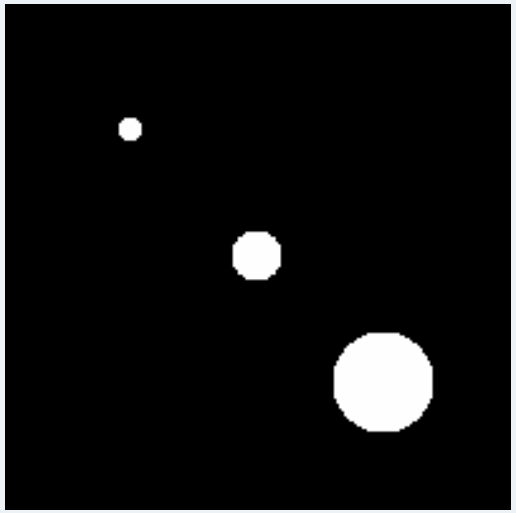
Salience Computation

- Top-down salience is calculated as the distance between the average color of the blob and the target blob color
- Bottom-Up salience as the distance between the average color of the blob and the surrounding region (Euclidean distance in the color opponent space)

$$S = k_{td} \cdot S_{top-down} + k_{bu} \cdot S_{bottom-up}$$

Local Competition

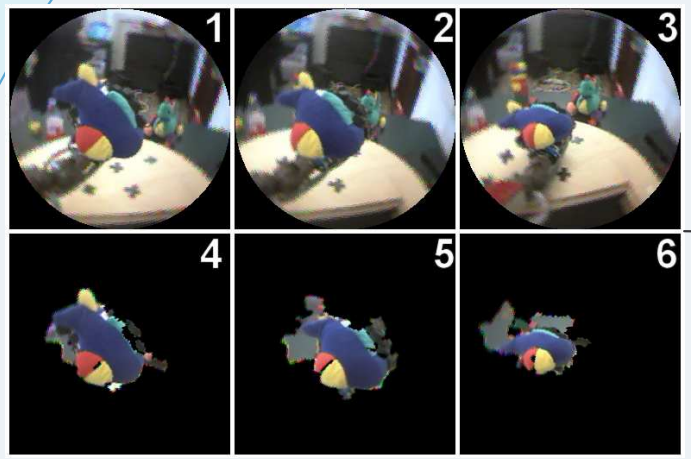
- A DoG filter would enhance the activity only of stimulus of the same size of the positive lobe of the filter
- Multiscale filters -> lack of continuity, combination problem
- Solution: each blobs defines a specific own scale
  - Using difference of rectangular filters (Viola and Jones, 2004) for speed



Learning about objects through interaction (1)

- The actions are used to define a concept of object
- Acquiring a model of the object without using pre-segmented images
- Actions can help to test "hypothesis" about the world
- Even the hand of the robot is a special object

Learning about objects through interaction (2)

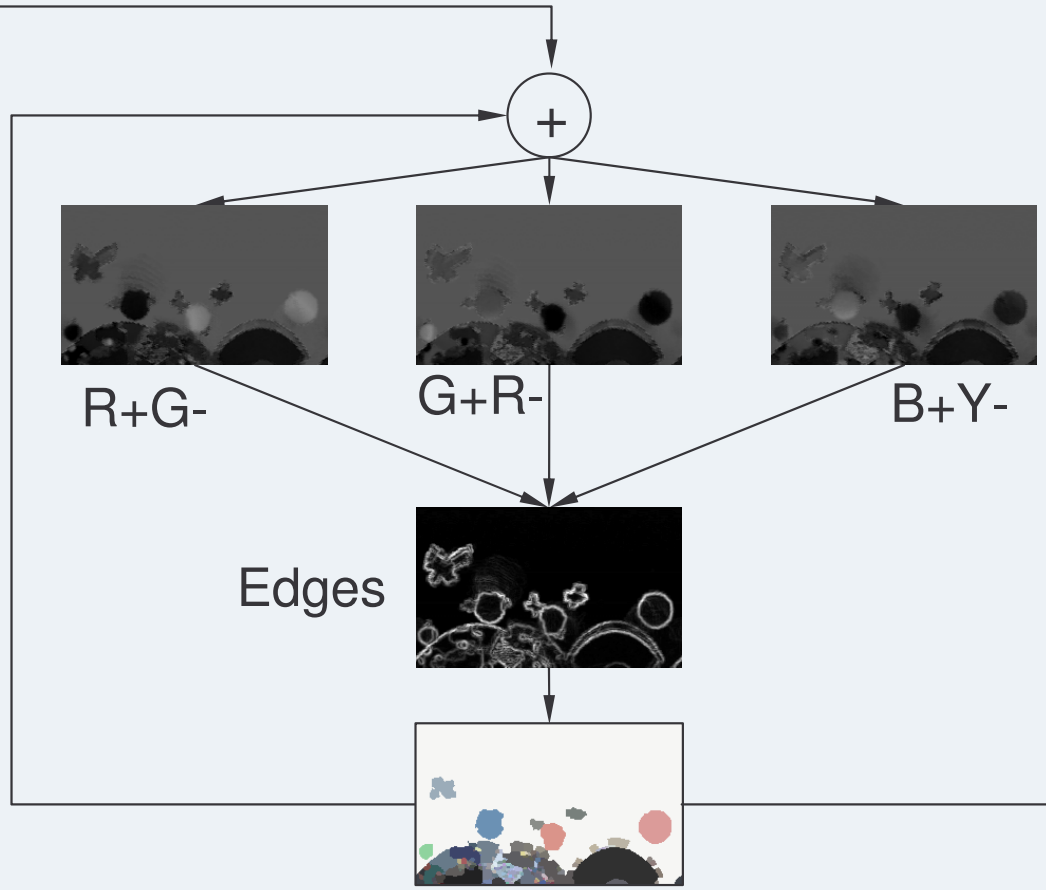


Blob processing → Estimating probabilities

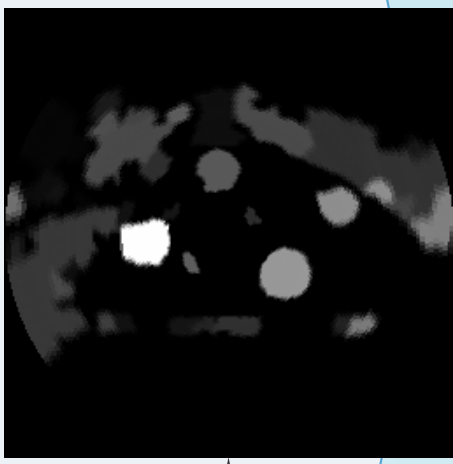
- Watching the hand holding the object
- Hypothesis: central blob $\in$ object
- Estimate during learning:
  $P(blob_i \in object / blob_i \text{ adjacent central blob, object is fixated})$
- Calculate during recognition:
  $P(\text{object is fixated} / blob_i \text{ adjacent central blob}, blob_i \in object)$
- Color of the central blob is used as the top-down information during the visual search of the object

An example

Input image

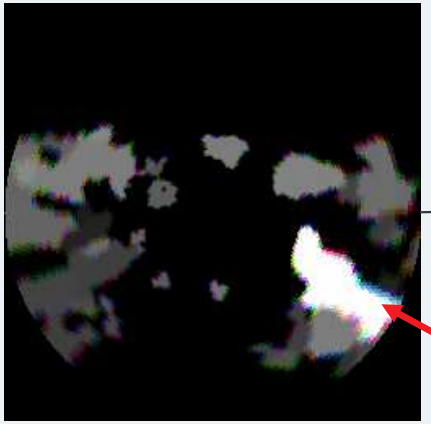


Saliency Map



Some results (1)

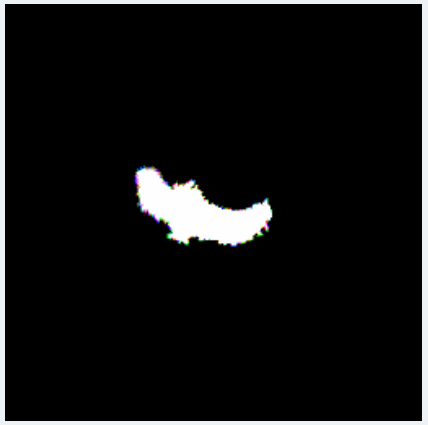
Looking for airplane…



Salient blobs

Moving…

Is it the airplane?

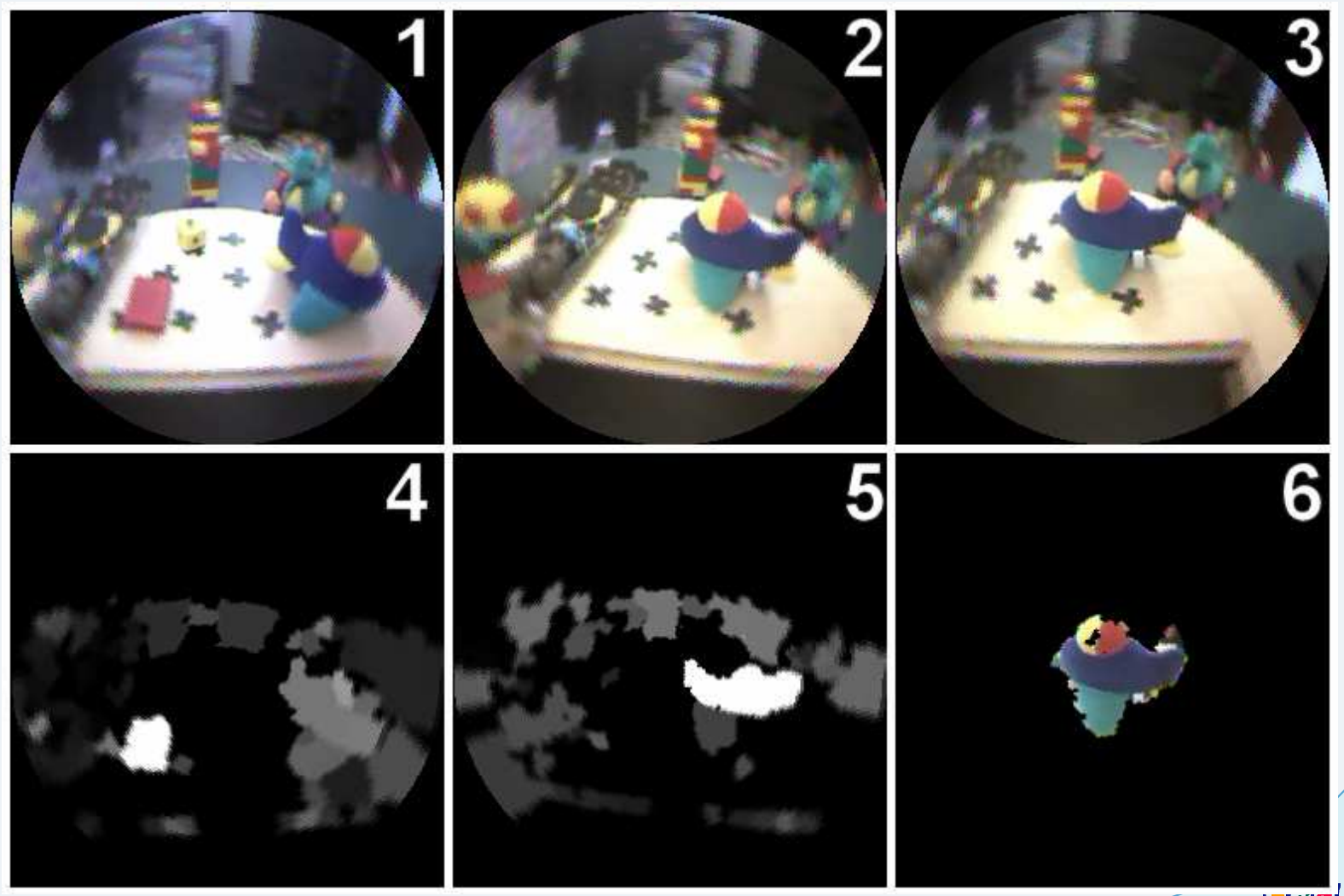


Color segmentation

Full segmentation

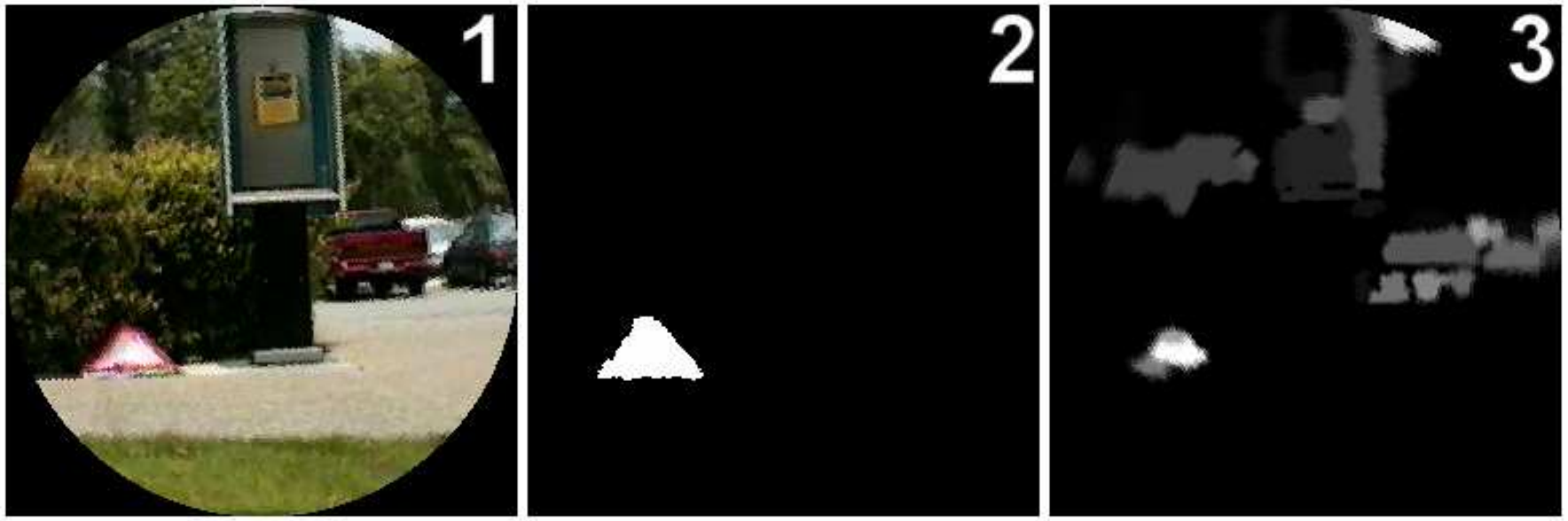
RobotCub

Some results (2)



Cogsys
Cognitive Systems

Some results (3)



The original images and target mask from Itti's database (Itti & Koch, 2001) are transformed in log-polar ones.

In 49% of the images a point inside the emergency triangle was selected as the most salient.

RobotCub

Conclusions

- Wide field of view & high resolution => active vision
- Active vision => Visual attention
- Object-based Visual attention => concept of object
- Object as a graph of proto-objects

- Attention priming though learning, integration with a recognition system, used to guide a robot manipulation task in real-time with some biological plausibility

Cogsys
Cognitive Systems

Future work (thesis?)

- Add other low-level feature maps, following the indications of the findings of the psychologists (e.g. texture)
- To define salience also on size
- Extending geometric relations of the object model

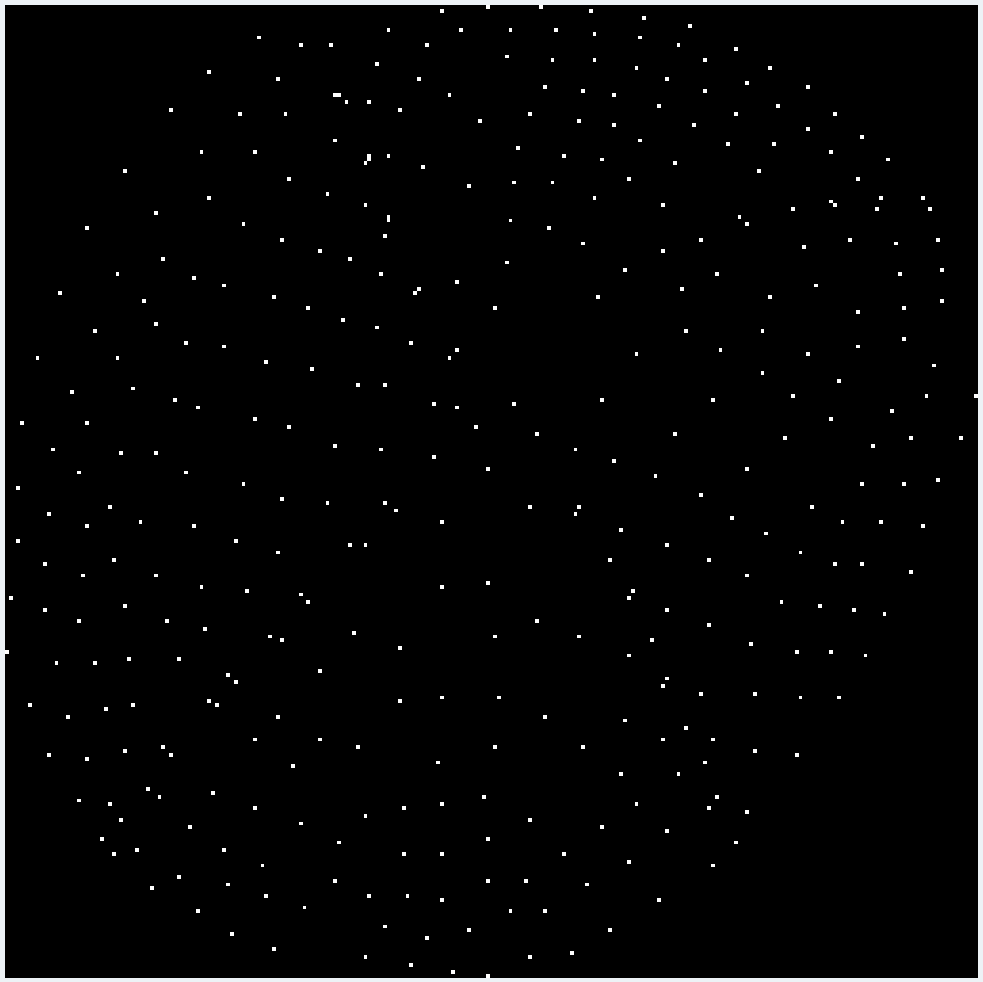
Thesis @ LIRA-Lab (1)

- Elena Curti, Paola Lanza

- Reconstruction of the geometry of a visual sensor, starting from raw data
- Hypothesis: pixel are not independent, the more they are near the more are dependent
- Methods: Mutual information, Renyi's entropies, mean color distance

RobotCub

Thesis @ LIRA-Lab (2)



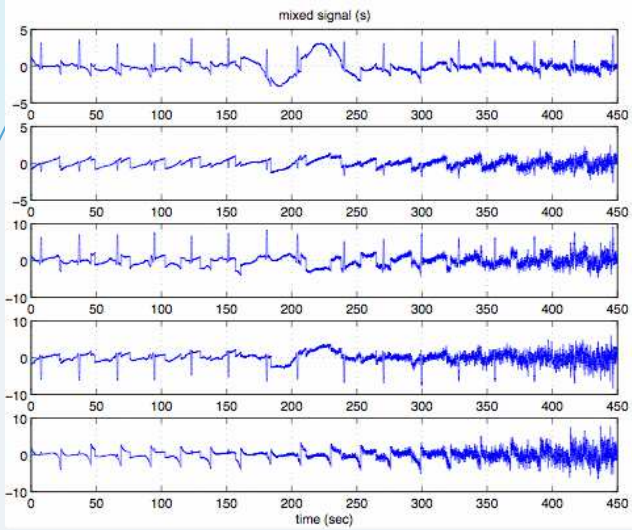
Cogsys
Cognitive Systems

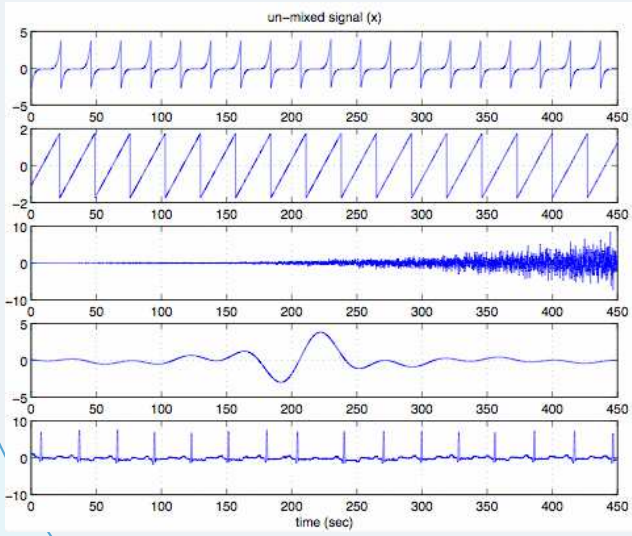
Thesis @ LIRA-Lab (3)

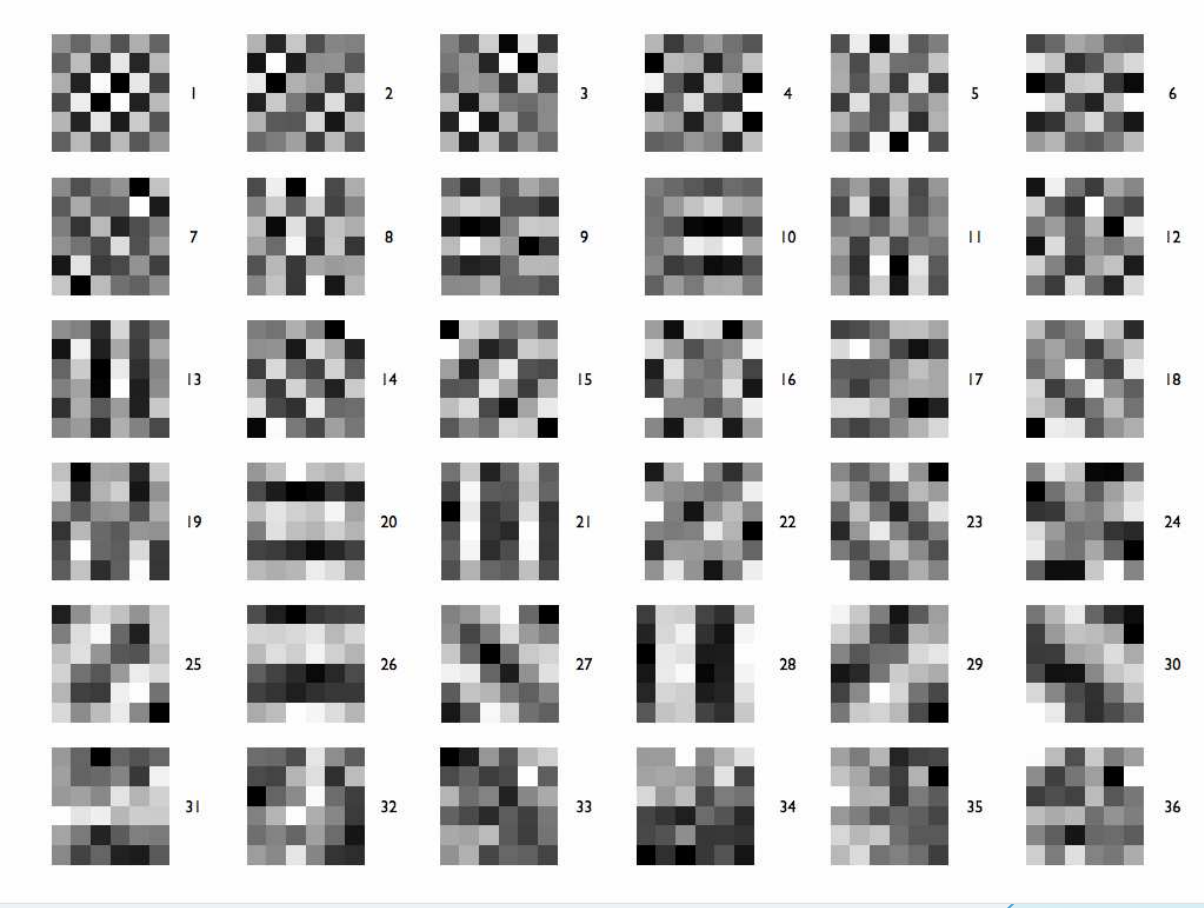
- Michele Tavella, Marianna Semprini

- Independent Component Analysis on images and on sounds
- Hypothesis: images are linear superposition of "patches"; linear mixing of sounds
- Methods: Information theory, image processing

Thesis @ LIRA-Lab (4)







RobotCub

Some links…

- My homepage: http://bremen.liralab.it
- Our papers: http://www.liralab.it/papers.htm

- Itti's homepage: http://ilab.usc.edu

- RobotCub homepage: http://www.robotcub.org

Cogsys
Cognitive Systems

RobotCub

Thanks for your Attention!

Cogsys
Cognitive Systems