TÉCNICO LISBOA

Institute for Systems and Robotics | LISBOA

VVV 2018
Computer Vision

Alex Bernardino

alex@isr.tecnico.ulisboa.pt

LARSyS
Laboratory of Robotics and Engineering Systems

Computer and Robot Vision Lab

VISLAB

Outline

- Introduction
- Basics of Computer Vision
  - Geometry
  - Image Processing
- From Human Vision to Robot Vision.
  - Foveal Vision
  - Visual Attention

Institute for Systems and Robotics | LISBOA

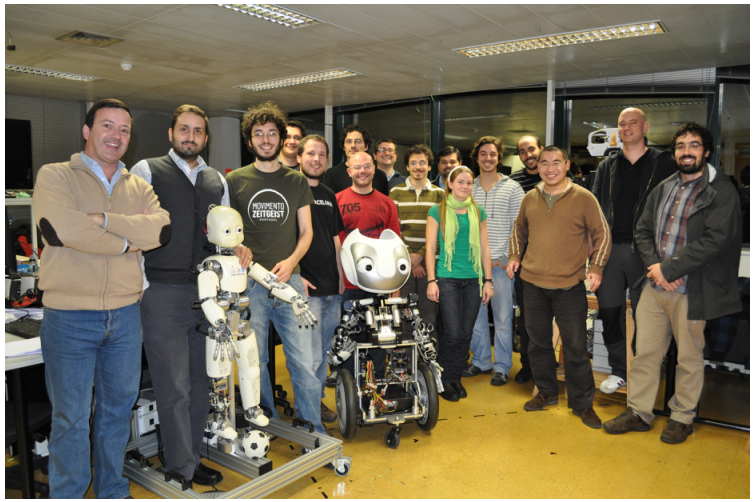
The Vislab Research Team

PEOPLE (October 2017)

Prof. José Santos-Victor
Prof. Alex Bernardino
Prof. José Gaspar
Prof. Jorge Marques

Dr. Mihai Andries
Dr. Mirko Rakovic
Dr. Ricardo Ribeiro
Dr. Moreno Coco
Dr. Nino Cauli
Dr. Plinio Moreno

Eng. Giovanni Saponaro
Eng. Pedro Vicente
Eng. Gonçalo Cruz
Eng. Rui Figueiredo
Eng. Nuno Pessanha-Santos
Eng. Nuno Duarte
Eng. Nuno Monteiro
Eng. João Martins
Eng. Mih Hun Lee
Eng. Atabak Dehban
Eng. João Avelino
Eng. Mehmet Mutlu
Eng. Catarina Barata
Eng. Carlos Cardoso



TÉCNICO LISBOA

Computer and Robot Vision Lab VISLAB

Institute for Systems and Robotics | LISBOA

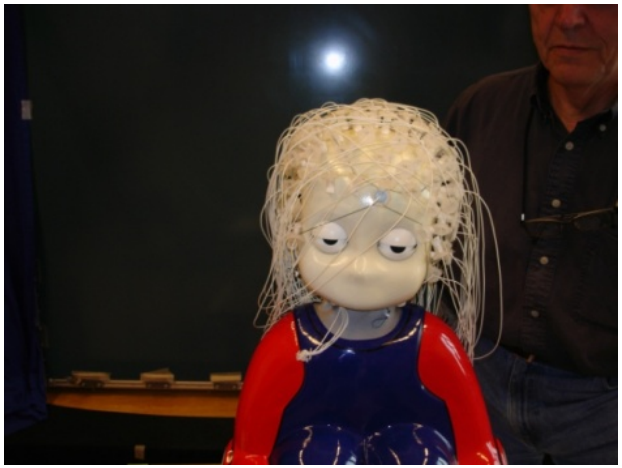
Vislab Mission and Goals

The ultimate goal of our research is twofold: understanding (natural and artificial) vision and building systems and applications that "see".

By understanding how natural systems work, we can devise innovative solutions for artificial ones. By synthesizing artificial systems, we can understand better the mechanisms that drive the natural ones.

TOPICS

Image Analysis and Surveillance
Visual Navigation and Calibration
Bio-inspired Vision and Learning
Cognitive Robots



The iCub in Uppsala, May 2008

Institute for Systems and Robotics | LISBOA

Vislab @ VVV

Basics of Computer Vision

Alex Bernardino, VVV 2018

Why Should We Study Vision

(+) It is a powerfull perceptual modality (the most powerfull?), allowing the acquisition of very rich information of the surrounding environment:

- *Object position and velocites*
- *Relationships among objects*
- *Object identity*
- *Interact with the world in a non-invasive way (without physical contact)*

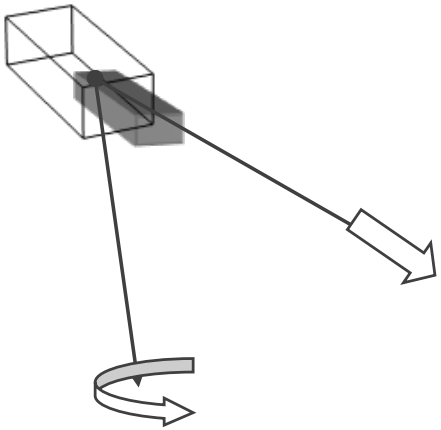
(+/-) Complex perceptual system. Above 50% of the human visual cortex is dedicated to processing visual information.

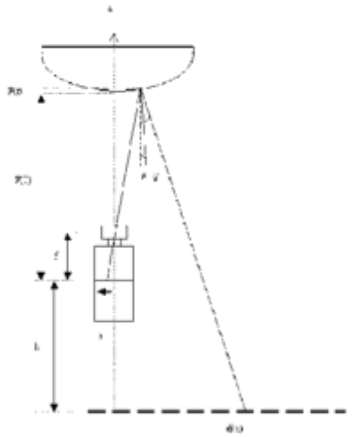
(-) Biological vision systems are still not very well understood.

Geometry vs Photometry

Geometry:

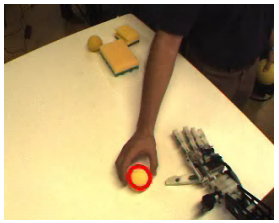
- How do 3D (real world) positions and velocities relate to 2D (images) positions and velocities ?

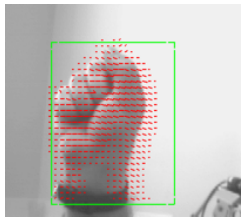


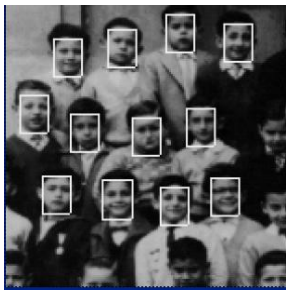


Photometry:

- How do color, brightness and texture information can be used to determine the positions and velocities of objects in the environment.



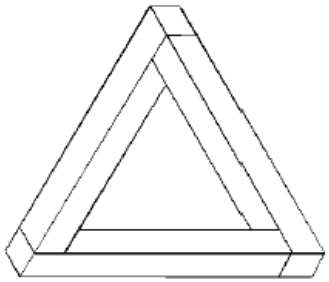




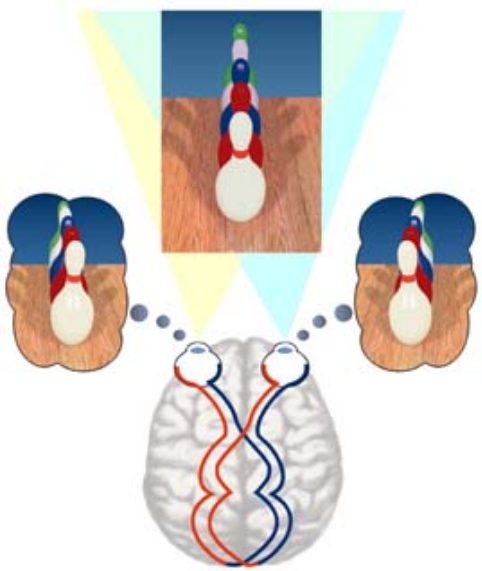
Depth Perception

3D space perception of the surrounding environment is possible due to the conjuntion of several visual cues:

- ***Stereo***
- ***Motion***
- ***Shading***
- ***Texture (gradient)***
- ***Prior Knowledge***
- ***Focus-Accomodation***

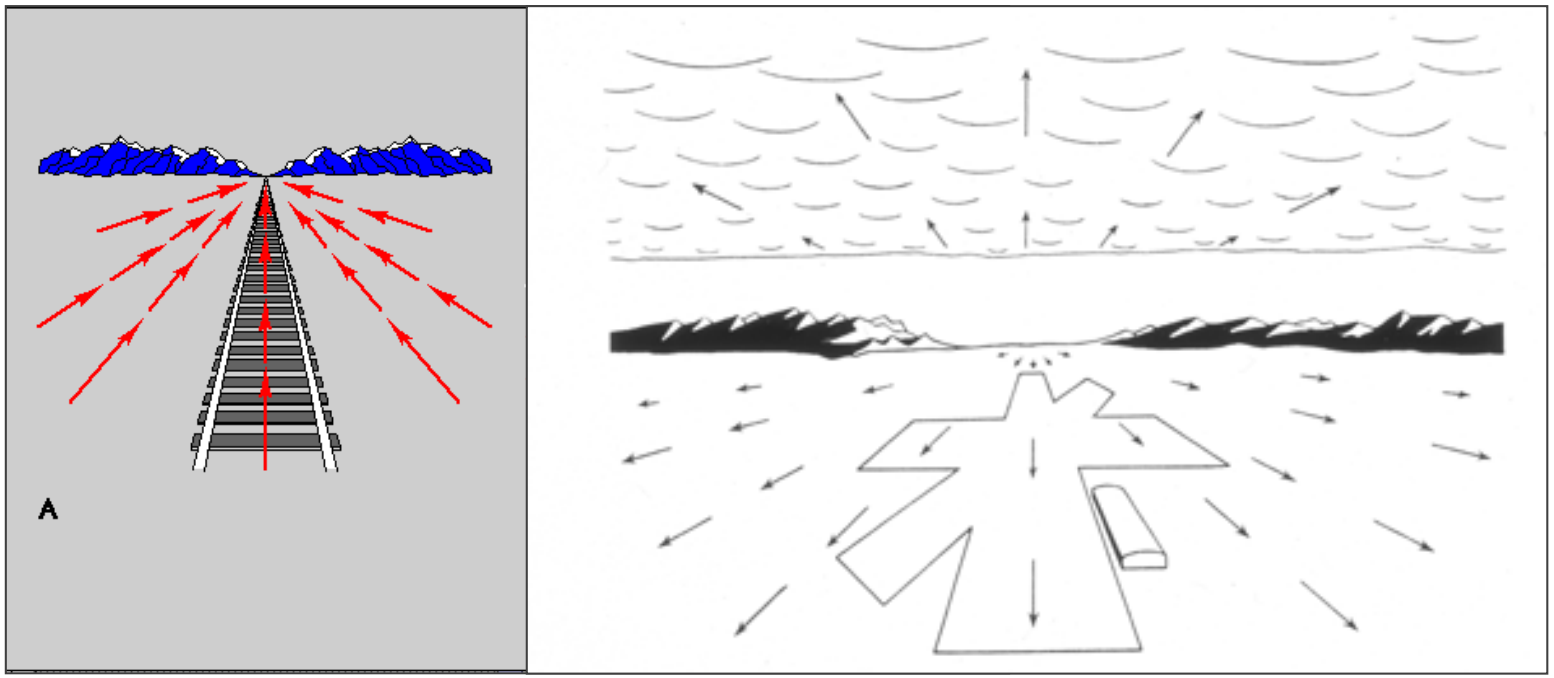


Stereo

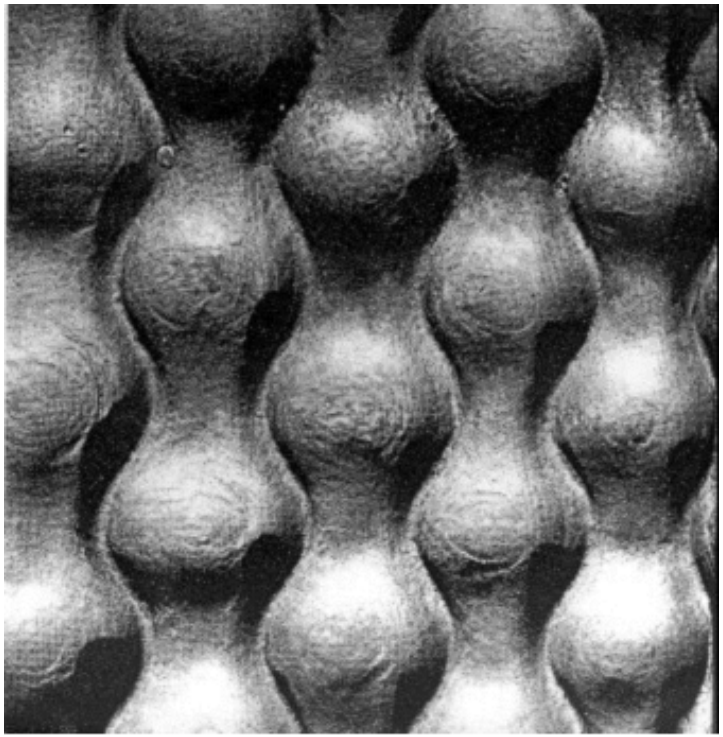


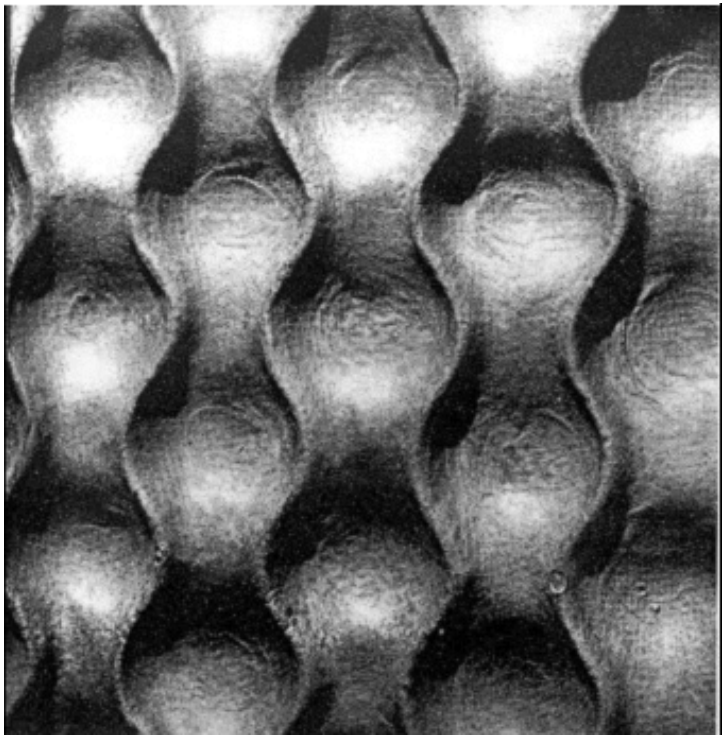
Taj Mahal - India

Motion Parallax

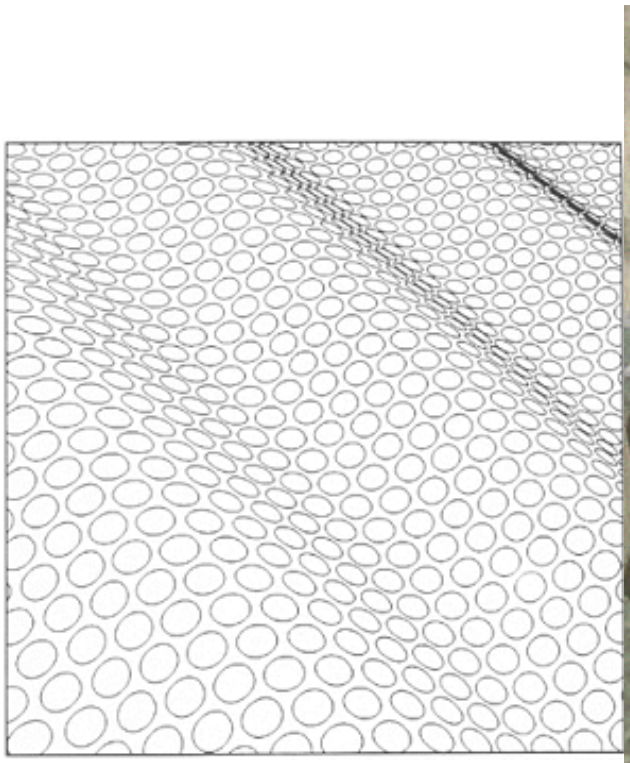


Shading



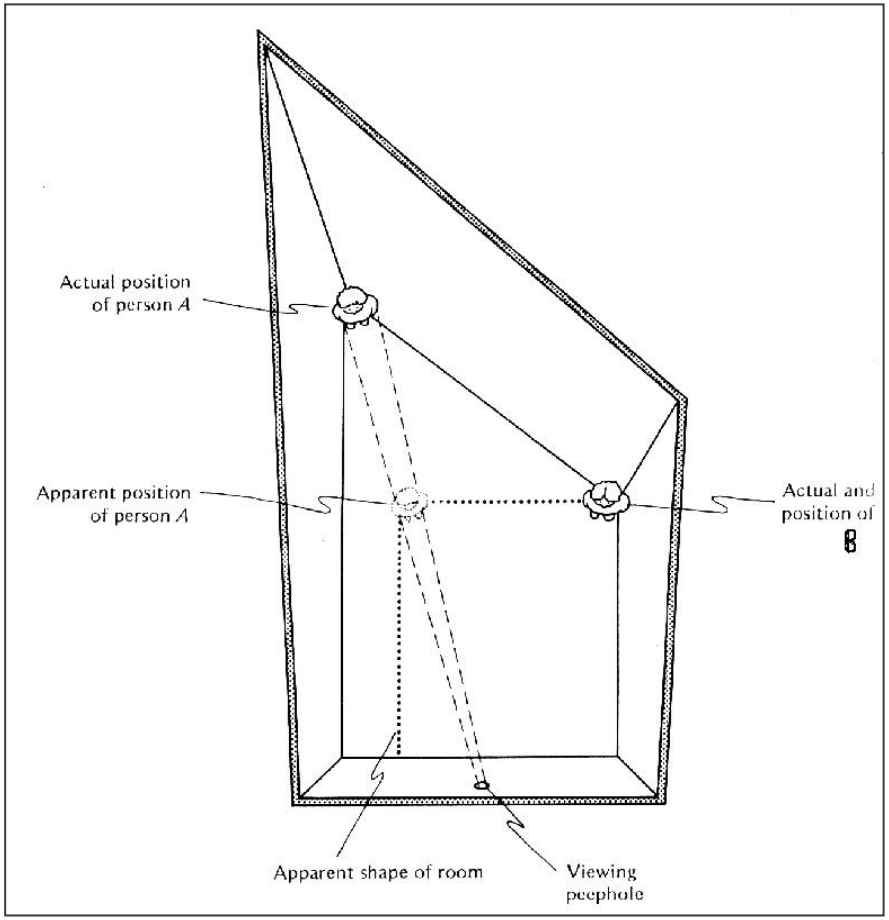


Texture Gradient

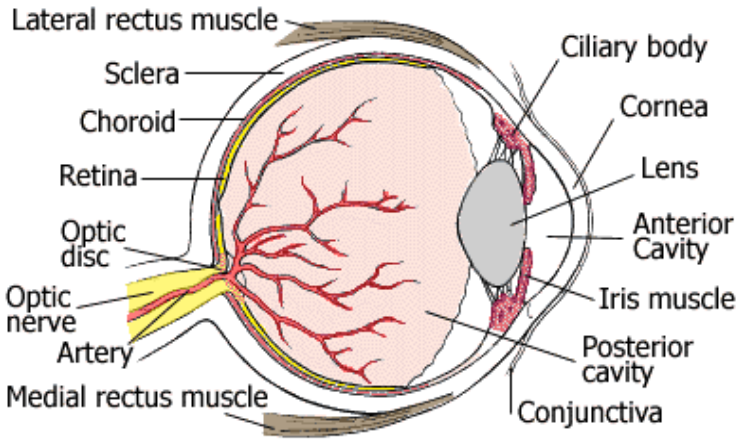


Institute for Systems and Robotics | LISBOA

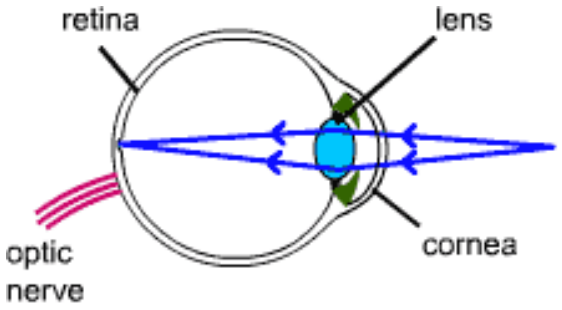
Prior Knowledge



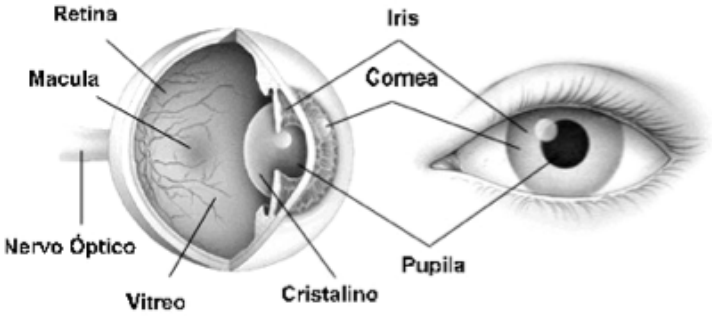
Focus/Accomodation

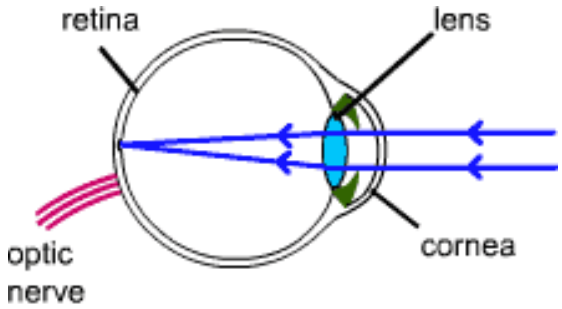


Transverse (horizontal) section of eyeball



eye focussing on near object



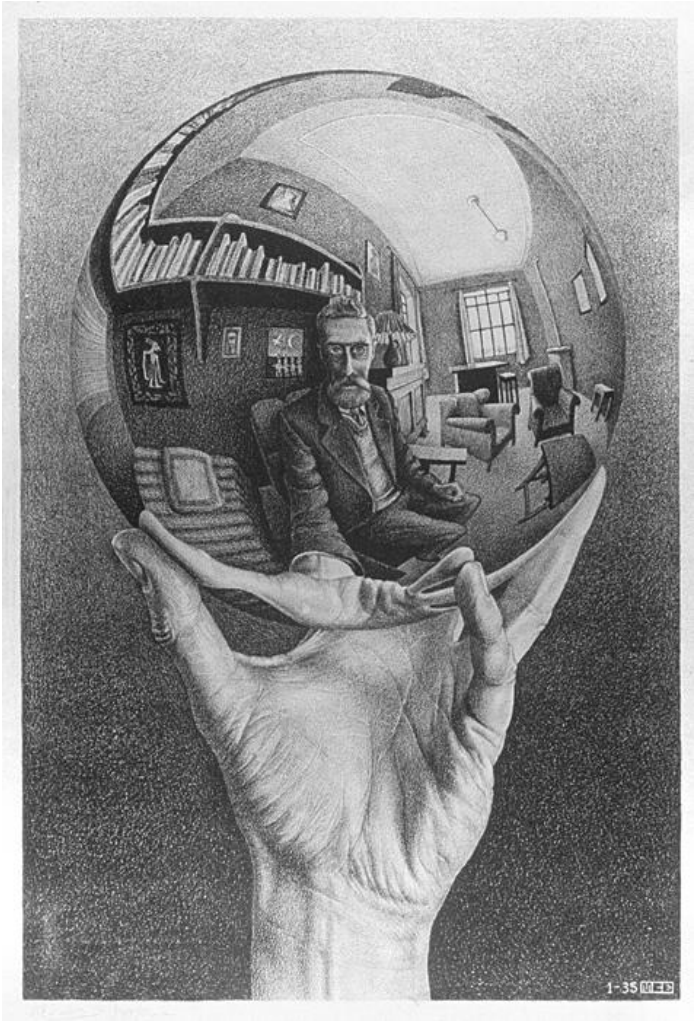


eye focussing on distant object

Image Formation

Image Formation

To understand how 3D world points project in 2D images.



Historical View

Egiptian

Chinese

Renaissance

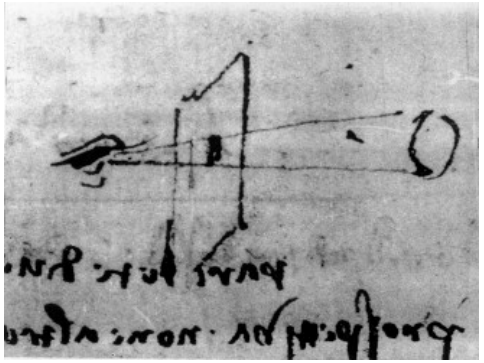
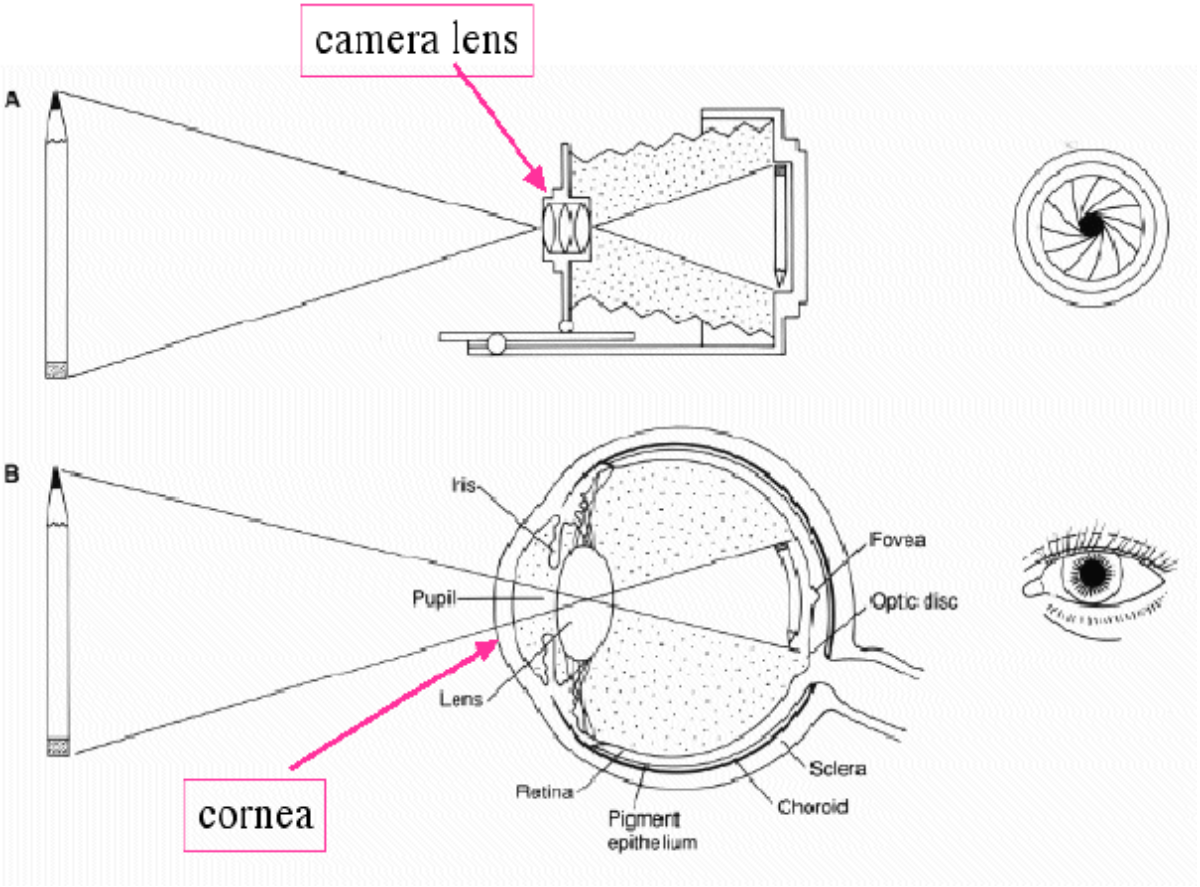
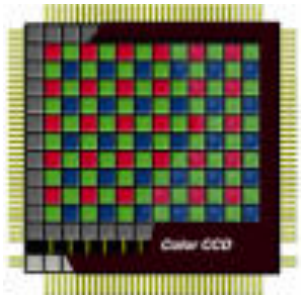
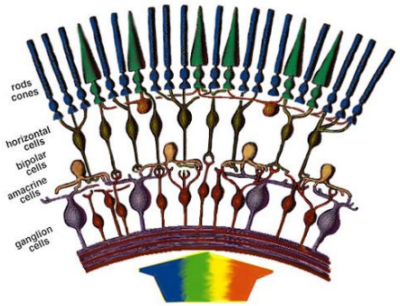


Image Projection







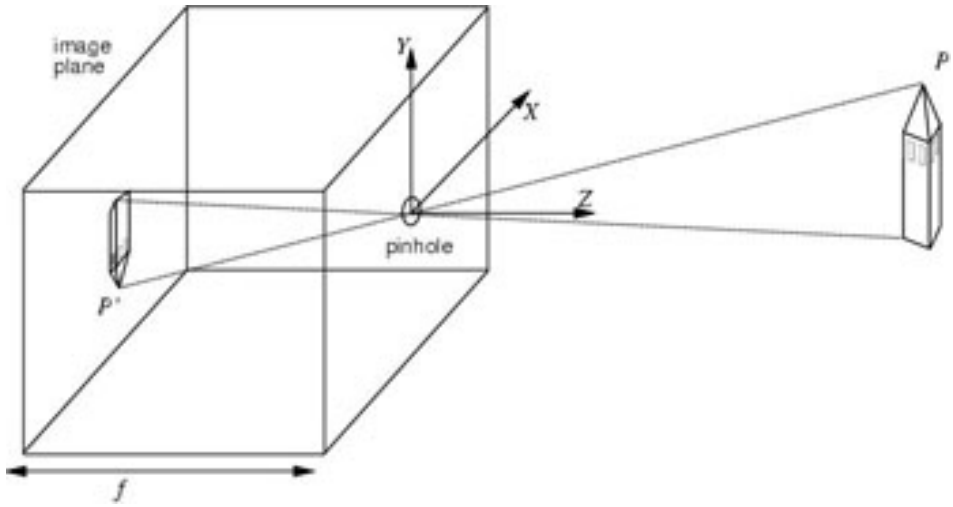
127 10^6 photoreceptors (120 10^6 rods + 7 10^6 cones)

10^6 ganglion cells (receptive fields)

Pinhole Camera Model

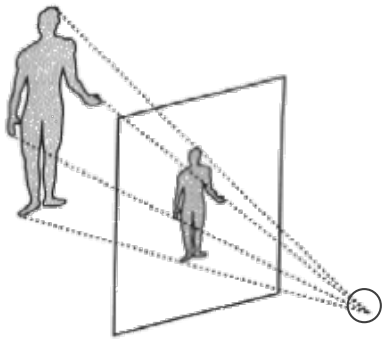
Ideal case:

Pin-Hole Camera



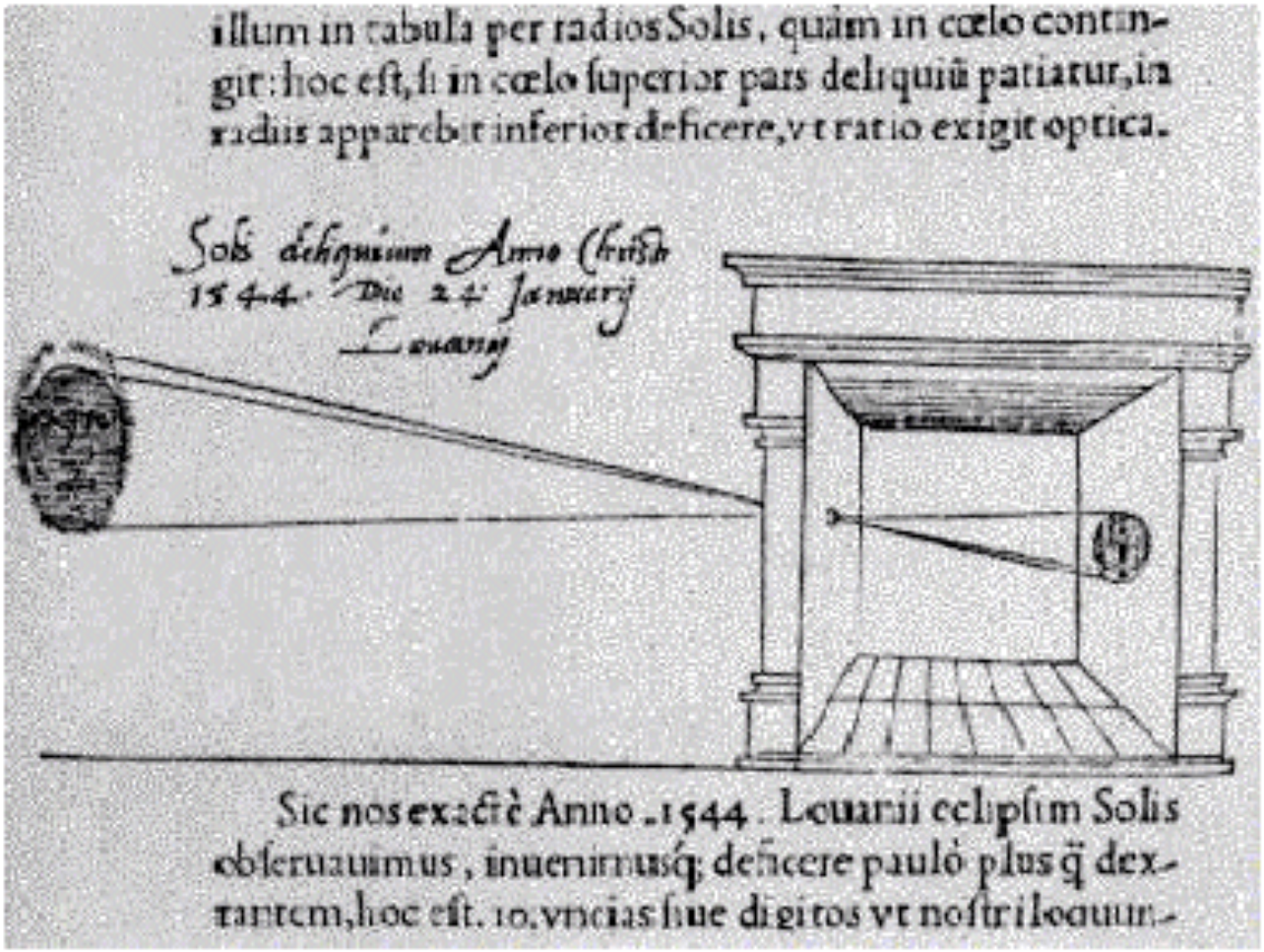
- Only light passing the pin-hole reaches the image plane.
- Each image point corresponds to a unique 3D world point

Alternative Representation:



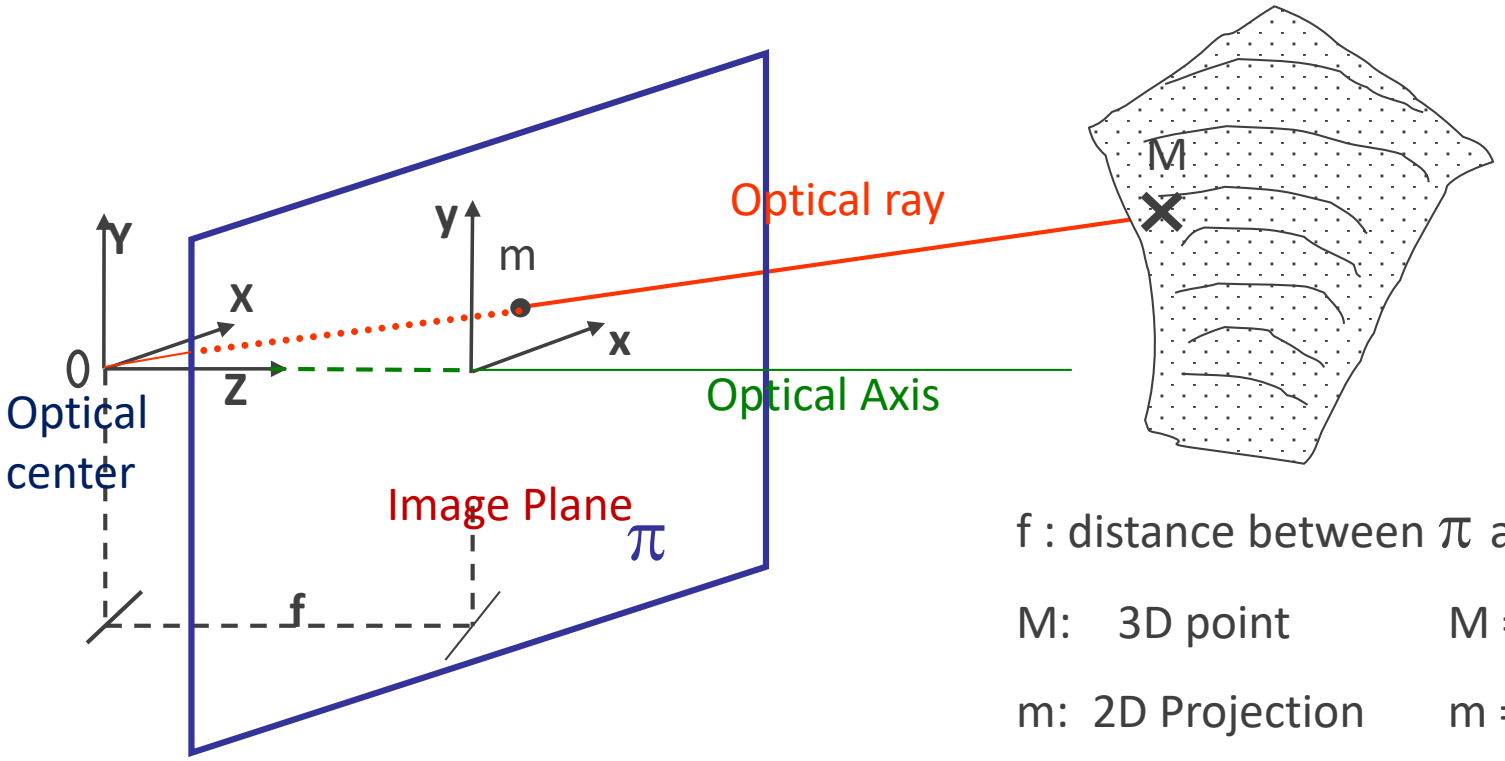
Projection/Optical Centre

Observation of a Solar Eclipse



Gemma Frisius, De Radio Astronomica et Geometrica, 1545

The Perspective Projection



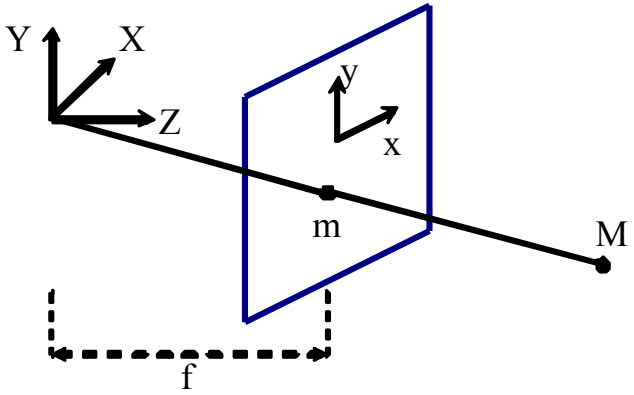
f : distance between π and O.

M: 3D point M = (X, Y, Z)

m: 2D Projection m = (x, y, f)

$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}$$

The Perspective Projection



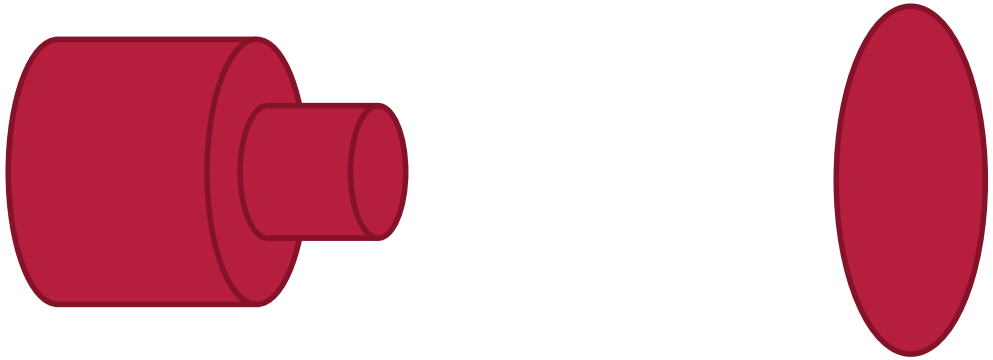
$$x = f\frac{X}{Z}, \quad y = f\frac{Y}{Z}$$

Homogeneous coordinates:

$$\begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Exercise

A disc of radius 50cm is placed in front of a camera, orthogonal the the optical axis, at a distance of 2m. The camera has a focal distance of 4mm. What is the radius of the disc projection in the image plane ?



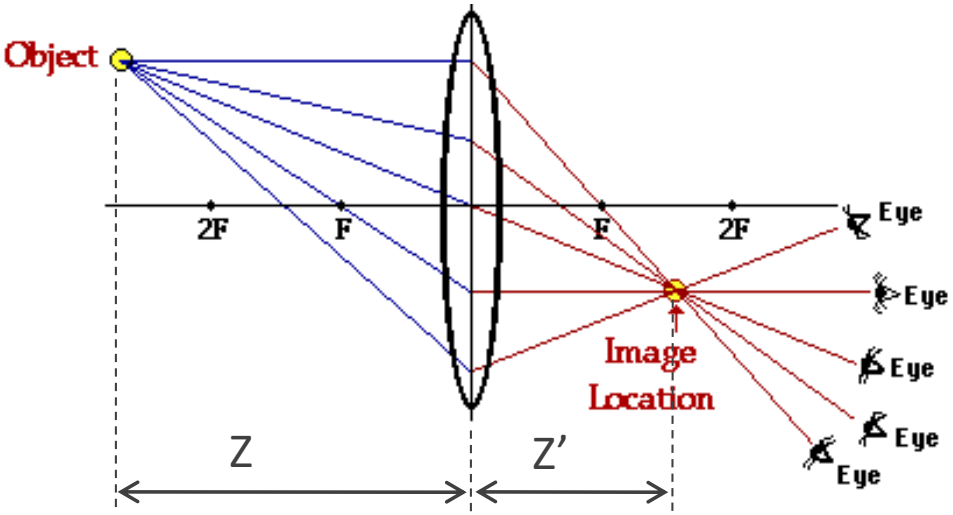
Lenses

Lenses allow:

- The same projection as a pin-hole camera.
- Acquire a sufficient amount of light (as a function of the lens solid angle, seen from the object)

Ideal Lens:

- The optic ray passing on the lens centre is not deflected.
- The remaining rays intersect in a unique point in the image plane, together with the central ray.



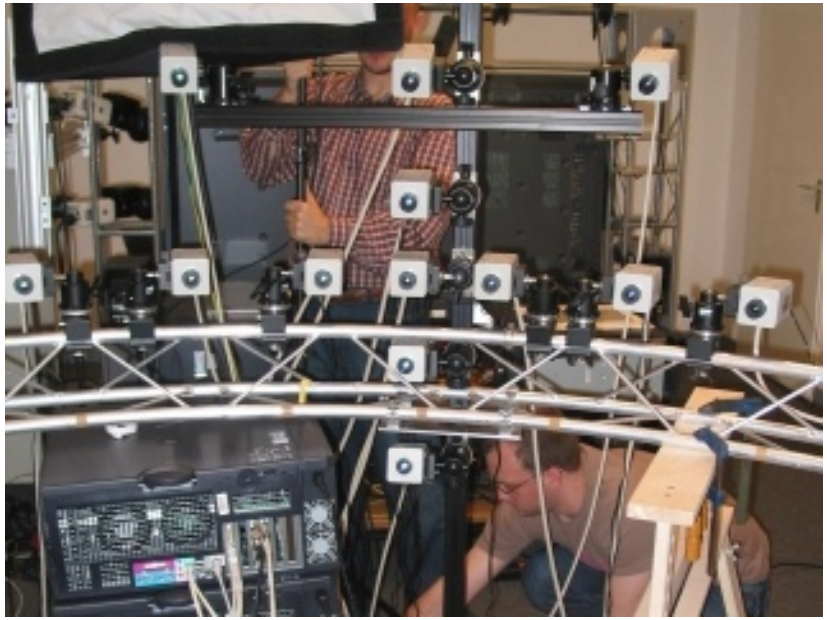
$$\frac{1}{Z'} + \frac{1}{Z} = \frac{1}{f}$$

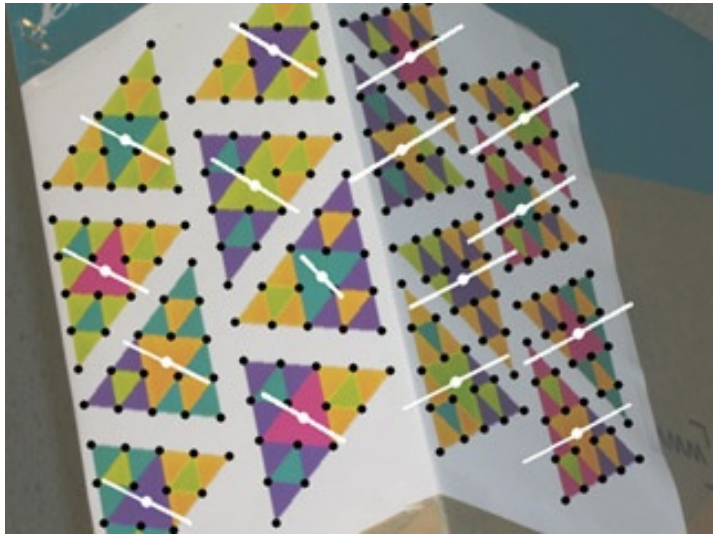
Lens equation

f-focal distance

Good focus only on a single plane!

Camera Calibration





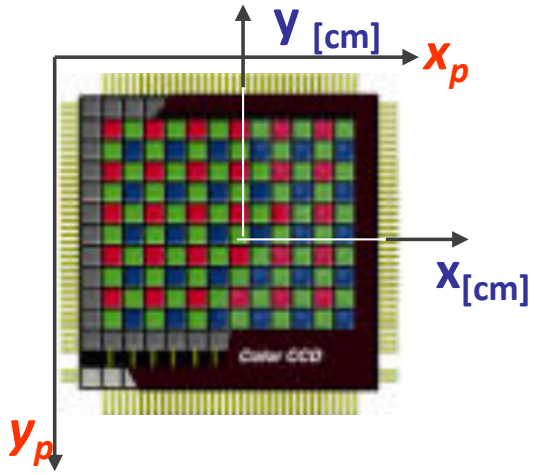
Intrinsic Parameters

In the perspective projection equations, image point coordinates are expressed in metric units and not in pixel indices (integers).

The intrinsic parameters describe:

- Image *pixel's size*.

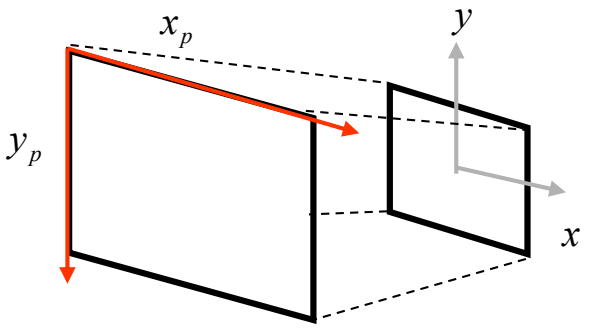
- Position of the *principal point or image center* (intersection of the optical axis with the image plane.



$$\begin{cases} x'_p = k_x x + C_x \\ y'_p = k_y y + C_y \end{cases}, \quad (in\ pixels)$$

In homogeneous coordinates:

$$\begin{bmatrix} \lambda x'_p \\ \lambda y'_p \\ \lambda \end{bmatrix} = \begin{bmatrix} k_x & 0 & C_x \\ 0 & k_y & C_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix}$$



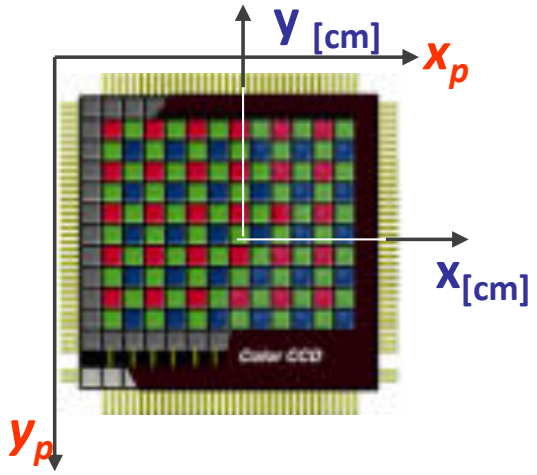
Intrinsic Parameters

In the perspective projection equations, image point coordinates are expressed in metric units and not in pixel indices (integers).

The intrinsic parameters describe:

- Image *pixel's size*.

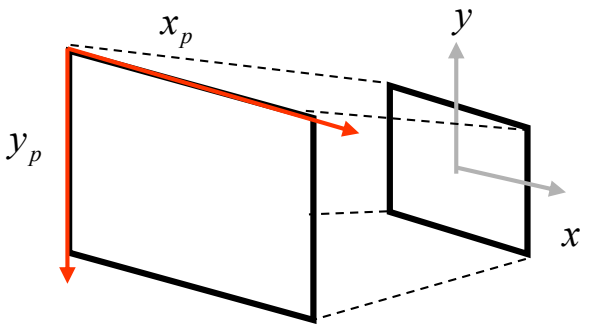
- Position of the *principal point or image center* (intersection of the optical axis with the image plane.



$$\begin{cases} x'_p = k_x x + C_x \\ y'_p = k_y y + C_y \end{cases}, \quad (in\ pixels)$$

In homogeneous coordinates:

$$\begin{bmatrix} \lambda x'_p \\ \lambda y'_p \\ \lambda \end{bmatrix} = \begin{bmatrix} k_x & 0 & C_x \\ 0 & k_y & C_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix}$$



Extrinsic Parameters

Often it is better to express the 3D points' coordinates with respect to a reference frame {W}, external to the camera (e.g in object centered coordinates).

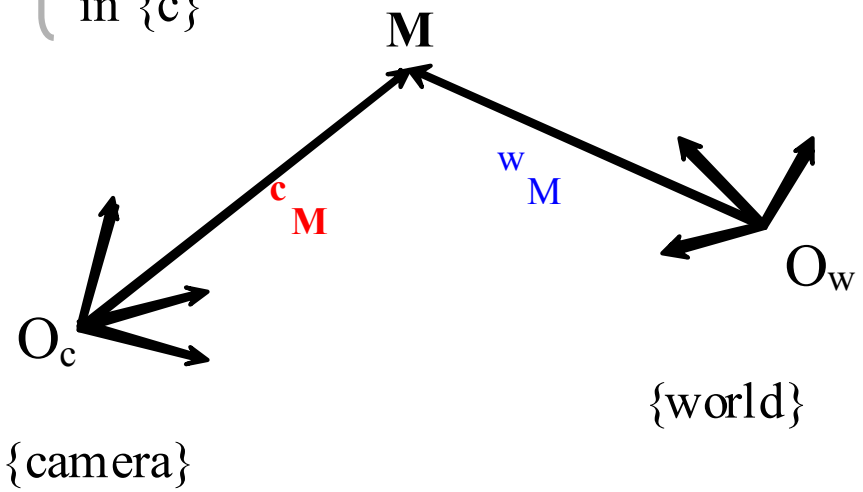
Rigid Motion: Rotation, Translation

$$^{c}M = {}^{c}R_{w}\ {}^{w}M + {}^{c}O_{w}$$

- $^{c}M \rightarrow$ Point M expressed in frame {c}
- $^{c}R_{w} \rightarrow$ Rotation {w, c}
- $^{c}O_{w} \rightarrow$ Origin of frame {w} expressed in {c}

In homogeneous coordinates:

$$^{c}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \left[\begin{array}{c|c} {}^{c}R_{w} & {}^{c}O_{w} \\ \hline (3\times 3) & \\ O_3^T & 1 \end{array}\right] {}^{w}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$



The Projection Matrix

Putting together the previous equation (projection, intrinsic and extrinsic parameters) we obtain the full model.

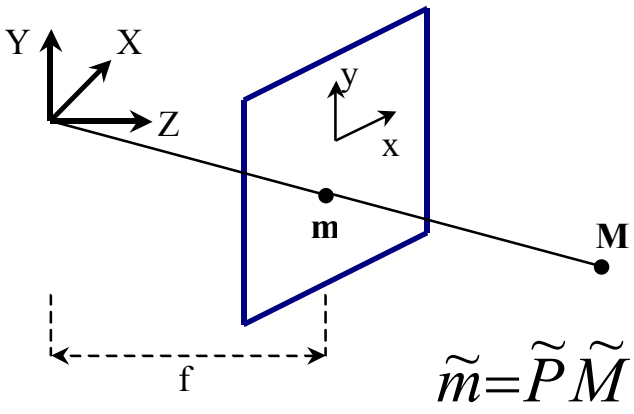
$$\begin{bmatrix} \lambda x_p \\ \lambda y_p \\ \lambda \end{bmatrix} = \begin{bmatrix} fk_x & 0 & C_x \\ 0 & fk_y & C_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \vec{r}_1 & t_x \\ \vec{r}_2 & t_y \\ \vec{r}_3 & t_z \\ \vec{0}_3 & 1 \end{bmatrix}^W \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Intrinsic parameters	Perspective Projection	External Orientation

The Projection Matrix

Multiplying the previous matrices :

$$\begin{bmatrix} \lambda x_p \\ \lambda y_p \\ \lambda \end{bmatrix} = \underbrace{\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix}}_{\tilde{P}} w \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$



$\tilde{P}$ is defined up to a scale factor. We can set : $p_{34} = 1$

Camera Calibration

Matrix $\widetilde{P}$ contains information about the camera's intrinsic and extrinsic parameters, often unknown, as well as the perspective projection. Camera calibration is the process of estimating matrix $\widetilde{P}$

$$\begin{bmatrix} \lambda x_p \\ \lambda y_p \\ \lambda \end{bmatrix} = \underbrace{\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix}}_{\widetilde{P}} \overset{w}{\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}}$$

Expressions can be written as:

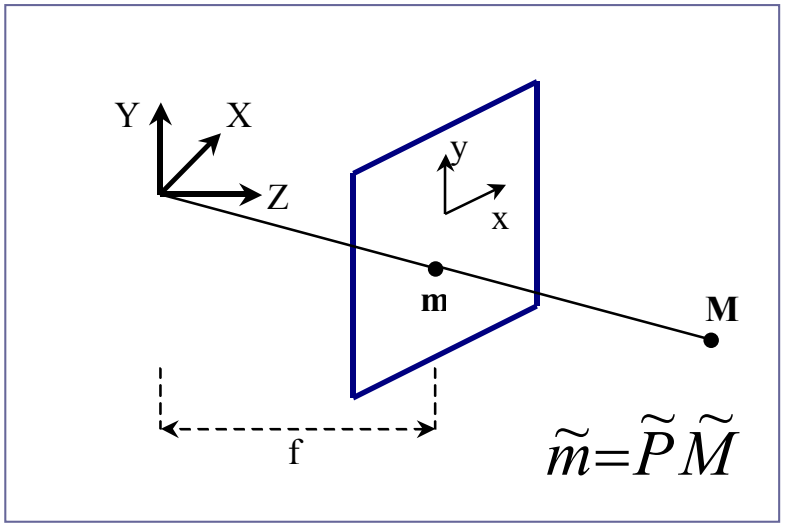
$$x_p = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + 1}$$

$$y_p = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + 1}$$

Camera Calibration

$$x_p = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + 1}$$

$$y_p = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + 1}$$



$$\begin{bmatrix} X & Y & Z & 1 & 0 & 0 & 0 & 0 & -Xx_p & -Yx_p & -Zx_p \\ 0 & 0 & 0 & 0 & X & Y & Z & 1 & -Xy_p & -Yy_p & -Zy_p \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ \vdots \\ p_{33} \end{bmatrix} = \begin{bmatrix} x_p \\ y_p \end{bmatrix}$$

11 unknowns
2 eqs. Per point => minimum 6 points for the calibration

Estimation of P

For a set of N pairs of corresponding points we have the following system of linear equations:

$2N$ eqs

$$\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0 & -X_1x_{p1} & -Y_1x_{p1} & -Z_1x_{p1} \\ 0 & 0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & -X_1y_{p1} & -Y_1y_{p1} & -Z_1y_{p1} \\ X_2 & Y_2 & Z_2 & 1 & 0 & 0 & 0 & 0 & -X_2x_{p2} & -Y_2x_{p2} & -Z_2x_{p2} \\ 0 & 0 & 0 & 0 & X_2 & Y_2 & Z_2 & 1 & -X_2y_{p2} & -Y_2y_{p2} & -Z_2y_{p2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ X_N & Y_N & Z_N & 1 & 0 & 0 & 0 & 0 & -X_Nx_{pN} & -Y_Nx_{pN} & -Z_1x_{pN} \\ 0 & 0 & 0 & 0 & X_N & Y_N & Z_N & 1 & -X_Ny_{pN} & -Y_Ny_{pN} & -Z_1y_{pN} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{14} \\ p_{21} \\ p_{22} \\ p_{23} \\ p_{24} \\ p_{31} \\ p_{32} \\ p_{33} \end{bmatrix} = \begin{bmatrix} x_{p1} \\ y_{p1} \\ x_{p2} \\ y_{p2} \\ \dots \\ \dots \\ x_{pN} \\ y_{pN} \end{bmatrix}$$

11 unknowns.

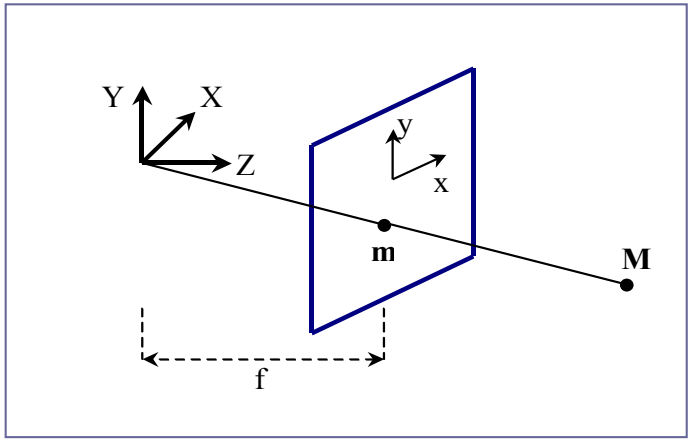
$$C \cdot p = c$$

Estimation of P

When the system is overdetermined (N>11) ,
we can use least-squares:

$$\hat{p}_{LS} = \arg\min_{\theta} \| C \cdot p - c \|^2$$

$$\hat{p}_{LS} = \left(C^T C\right)^{-1} C^T c$$



In Matlab:

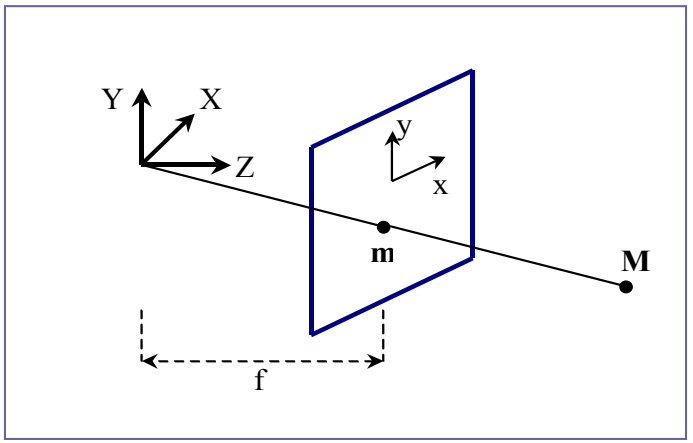
```
p_hat = inv(C'*C)*C'*c
```

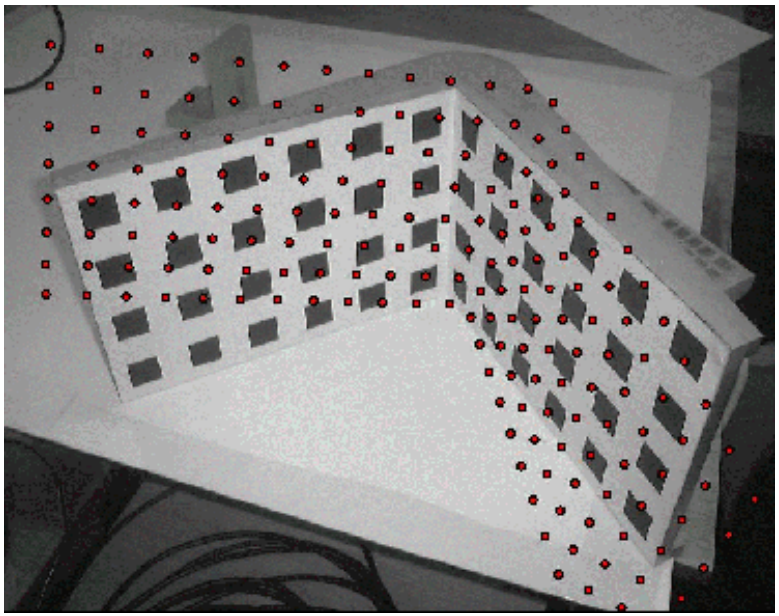
Institute for Systems and Robotics | LISBOA

Evaluation

The Reprojection Error:

$$E = \frac{\| C \cdot \hat{p}_{LS} - c \|^2}{N}$$





Stereo Vision

Stereo Vision is one of the main depth perception mechanisms, allowing the estimation of distances to objects in the environment.

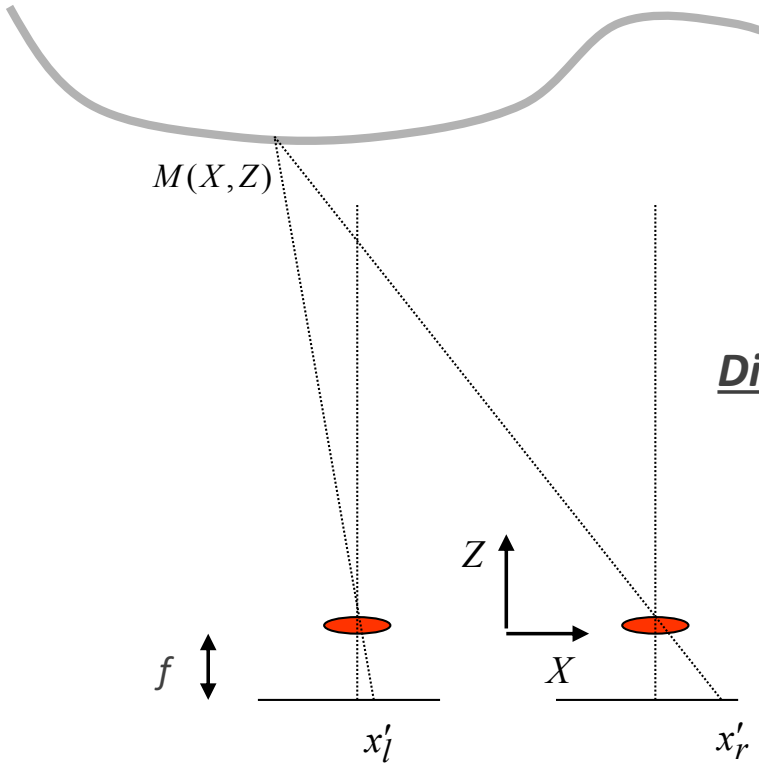
Steps :

MATCHING – Searches for points in a pair of images, resulting from the projection of the same 3D point.

RECONSTRUCTION – Computes the 3D point position from its projection in two cameras.

Parallel Stereo

Let us consider two cameras, aligned with parallel optical axes at distance *b* - *baseline*.



$$\frac{x'_l}{f} = -\frac{X+\frac{b}{2}}{Z}; \qquad \frac{x'_r}{f} = -\frac{X-\frac{b}{2}}{Z}$$

Disparity (d):

$$d \triangleq x'_l - x'_r = \frac{b\,f}{z}$$

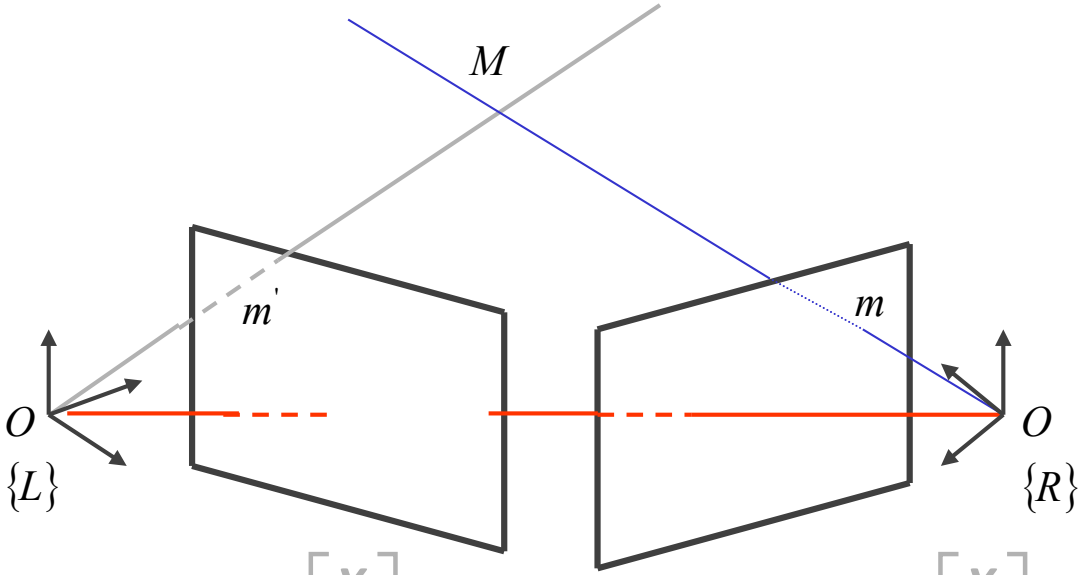
$$\boxed{z = \frac{b\,f}{d}}$$

Exercise

An object is observed by a parallel stereo setup, composed by cameras with a baseline b = 10 cm and with focal length f = 6mm. The object disparity in the image planes is 10 pixel. The CCD has 1024x1024 pixels and is 6mm side. At which distance from the cameras is the object located ?

Stereo General Case

When two cameras are not parallel, but are calibrated, we can reconstruct the 3D coordinates of world points, given their homologous projections in the stereo images.



$$\begin{bmatrix} \lambda_1 x_1 \\ \lambda_1 y_1 \\ \lambda_1 \end{bmatrix} = \tilde{P}_1 \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \qquad \begin{bmatrix} \lambda_2 x_2 \\ \lambda_2 y_2 \\ \lambda_2 \end{bmatrix} = \tilde{P}_2 \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

3D Reconstruction

Expanding the equations:

$$x_i = \frac{p^i_{11} X + p^i_{12} Y + p^i_{13} Z + p^i_{14}}{p^i_{31} X + p^i_{32} Y + p^i_{33} Z + 1} \qquad y_i = \frac{p^i_{21} X + p^i_{22} Y + p^i_{23} Z + p^i_{24}}{p^i_{31} X + p^i_{32} Y + p^i_{33} Z + 1}$$

In matrix form:

$$\begin{bmatrix} p^1_{11} - p^1_{31}x_1 & p^1_{12} - p^1_{32}x_1 & p^1_{13} - p^1_{33}x_1 \\ p^1_{21} - p^1_{31}y_1 & p^1_{22} - p^1_{32}y_1 & p^1_{23} - p^1_{33}y_1 \\ p^2_{11} - p^2_{31}x_2 & p^2_{12} - p^2_{32}x_2 & p^2_{13} - p^2_{33}x_2 \\ p^2_{21} - p^2_{31}y_2 & p^2_{22} - p^2_{32}y_2 & p^2_{23} - p^2_{33}y_2 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} x_1 - p^1_{14} \\ y_1 - p^1_{24} \\ x_2 - p^2_{14} \\ y_2 - p^2_{24} \end{bmatrix}$$

4 equations

3 unknowns

Computing the Solution

$$\begin{bmatrix} p_{11}^1 - p_{31}^1 x_1 & p_{12}^1 - p_{32}^1 x_1 & p_{13}^1 - p_{33}^1 x_1 \\ p_{21}^1 - p_{31}^1 y_1 & p_{22}^1 - p_{32}^1 y_1 & p_{23}^1 - p_{33}^1 y_1 \\ p_{11}^2 - p_{31}^2 x_2 & p_{12}^2 - p_{32}^2 x_2 & p_{13}^2 - p_{33}^2 x_2 \\ p_{21}^2 - p_{31}^2 y_2 & p_{22}^2 - p_{32}^2 y_2 & p_{23}^2 - p_{33}^2 y_2 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} x_1 - p_{14}^1 \\ y_1 - p_{24}^1 \\ x_2 - p_{14}^2 \\ y_2 - p_{24}^2 \end{bmatrix}$$

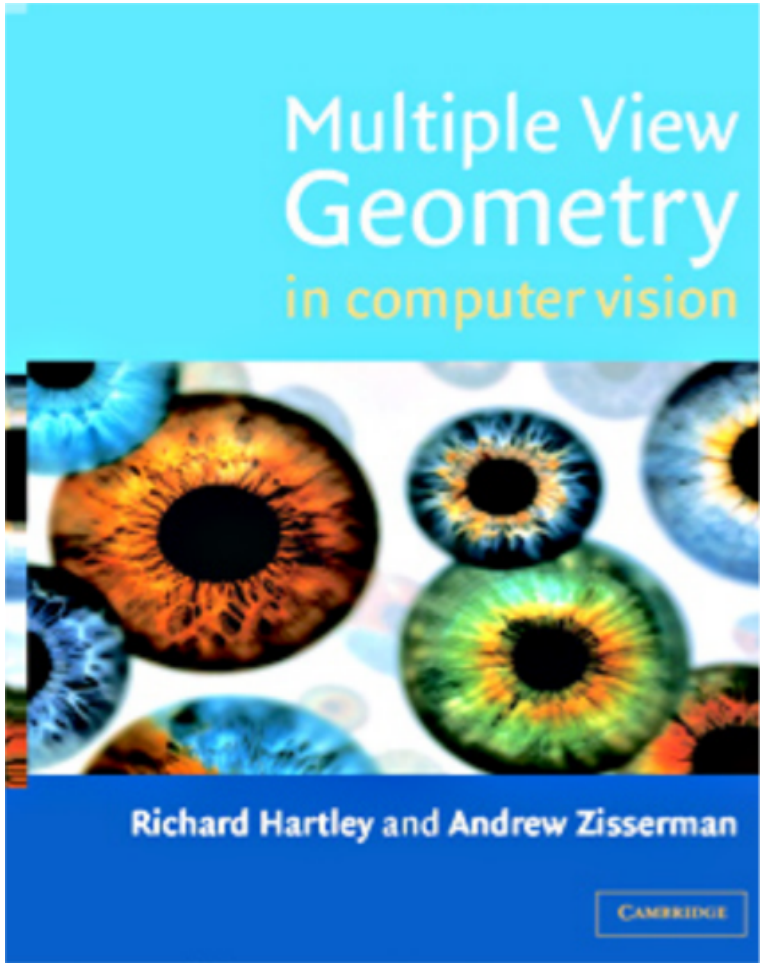
R M = r

Can be solved with least squares:

$$\hat{M}_{LS} = \arg\min_{P} \| RM - r \|^2$$

$$\hat{M}_{LS} = \left(R^T R\right)^{-1} R^T r$$

Multiview Geometry



What can we do with multi-view geometry?

$M_i, \Theta_j \Longrightarrow m_{ij}$, Image rendering

$m_{ij}, \Theta_j \Longrightarrow M_i$, 3D reconstruction, Structure from Motion

$m_{ij}, M_i \Longrightarrow \Theta_j$, Calibration, Egomotion

$m_{ij} \Longrightarrow \tilde{M}_i, \tilde{\Theta}_j$, Uncalibrated reconstruction, Partial Calibration

M_i, $i = 1, \cdots N$, 3D world points

Θ_j, $j = 1, \cdots K$, Camera parameters (intrinsic and extrinsic)

m_{ij}, Projection of point i in camera j

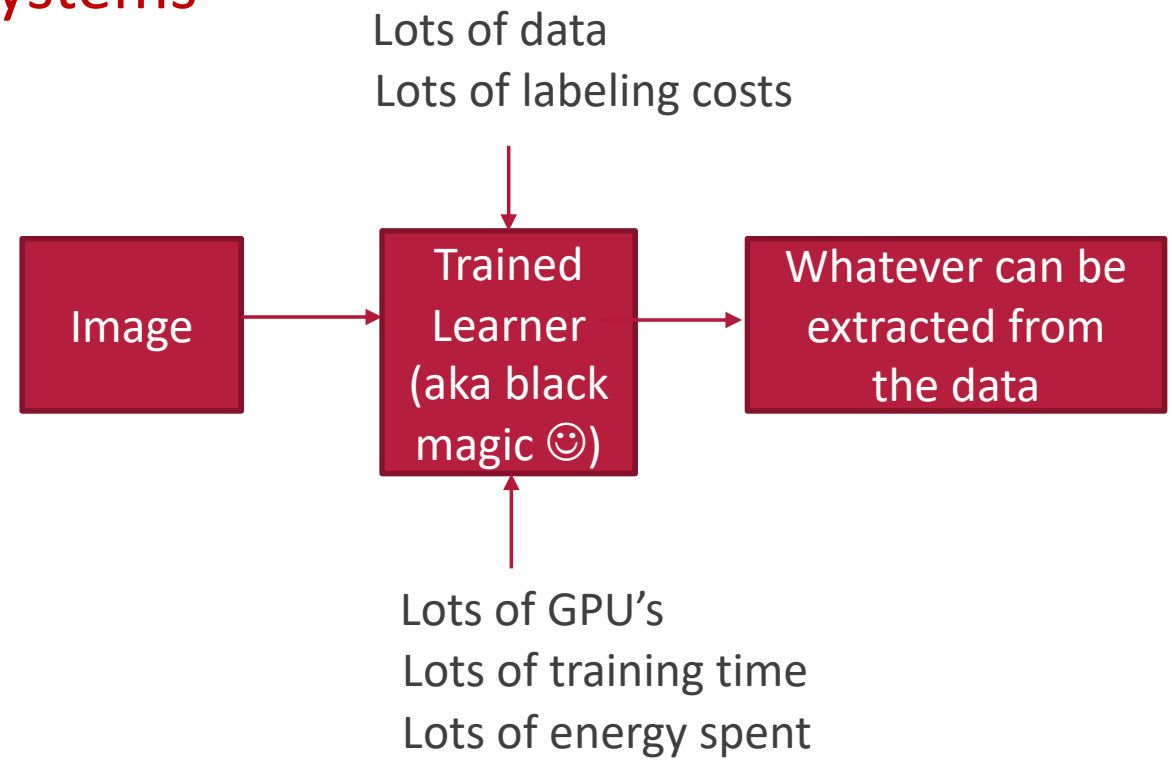
Main problem with multi-view geometry?

How to determine the point matches m_{ij} ?

- By hand, clicking in the images -> time consuming
- By tracking keypoints -> only for small baselines / high framerates
- By matching discriminant descriptors -> prone to false matches

Alternatives to multi-view geometry.

End-to-end Systems



Still lots to do

Humans still outperform machines in many tasks:

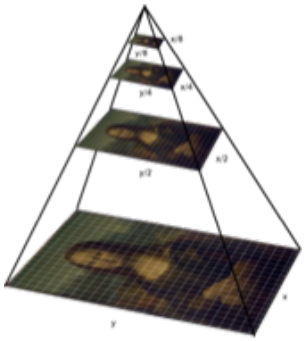
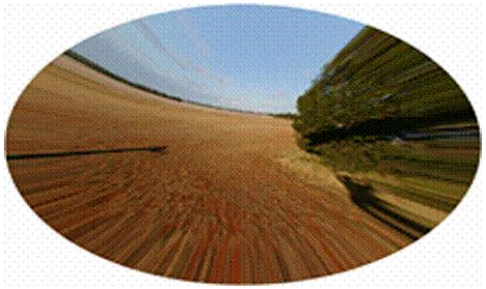
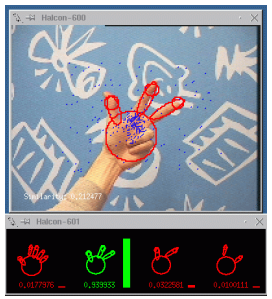
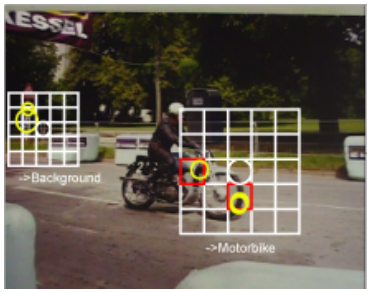
- More robust to tough conditions
- More energy efficient

Human Vision vs Machine Vision

Do it FASTER



Do it SMARTER

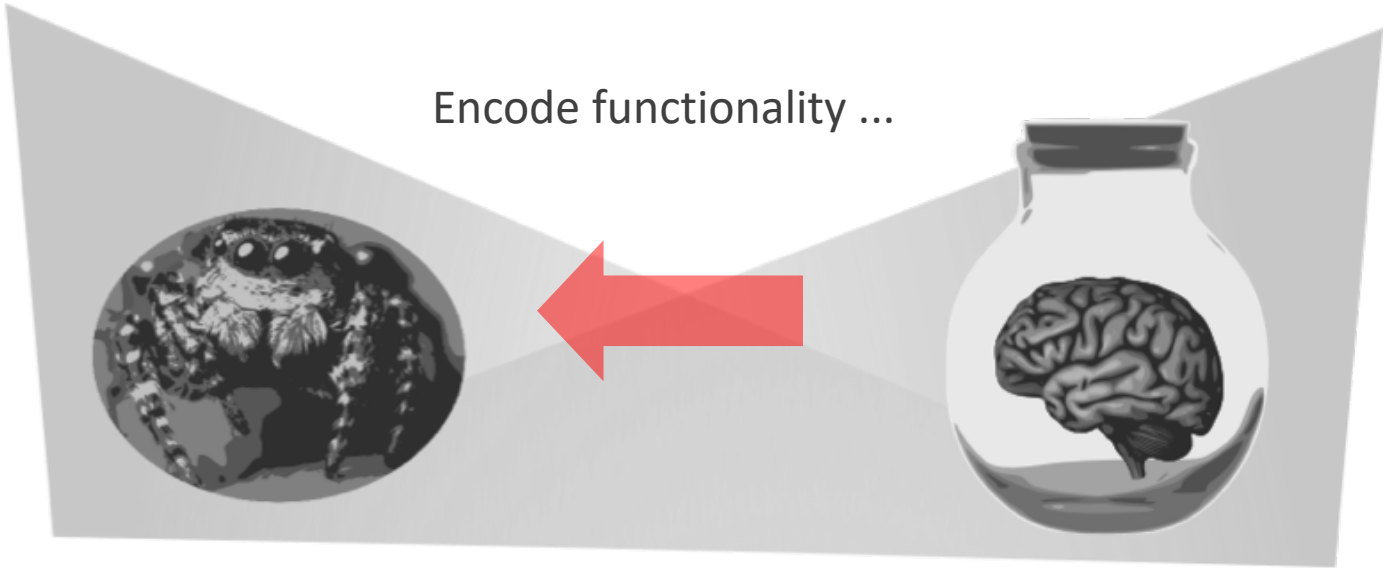





Hardware or software acceleration can speed up processing but not reduce intrinsic complexity. Only **algorithmic or architecture** changes can.

Institute for Systems and Robotics | LISBOA

In favour of simple brains

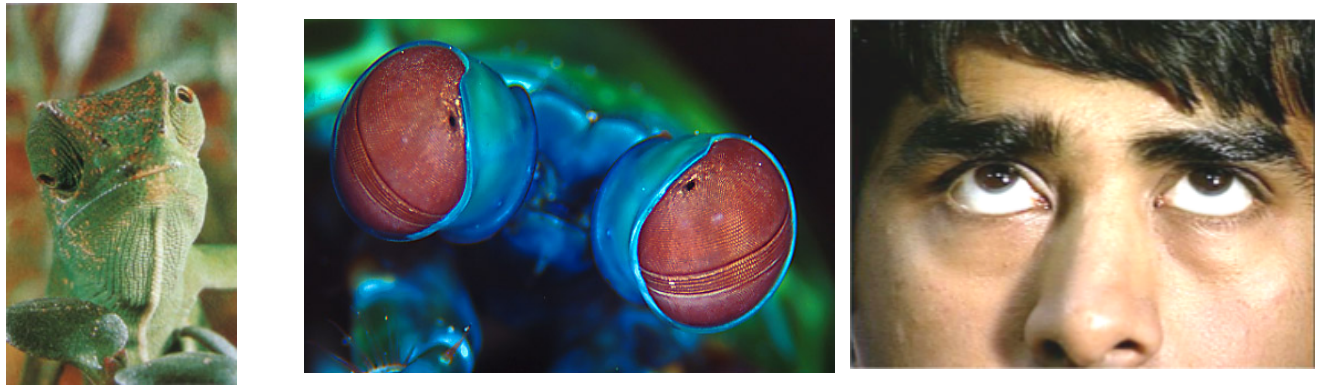
- Neural tissue is expensive (Niven & Laughlin 2008).
- Sparsity in neural circuits is important (Olshausen & Field 04).



Lessons From Nature

Biological Visual Systems are Space Variant

Biological Visual Systems have Moving Eyes



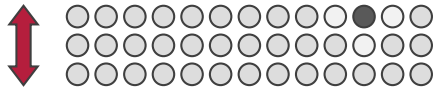
Some Biological Vision Systems

Plankton Copilia

- 7 receptors / eye
- horizontal scanning

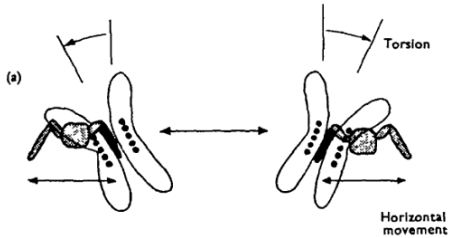
Sea Snail Oxygyrus

- 3x400 receptors / eye
- vertical scanning

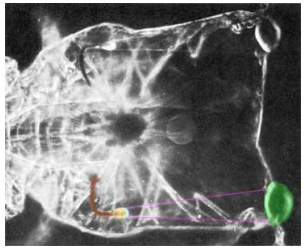


Jumping Spiders

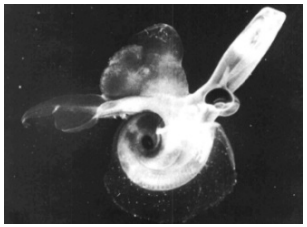
- scanning patterns



O. Drees 1952, M. F. Land 1969



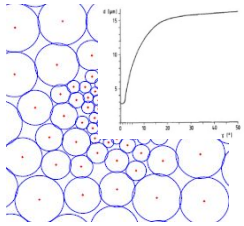
R. L. Gregory 1964



M.F. Land 1981

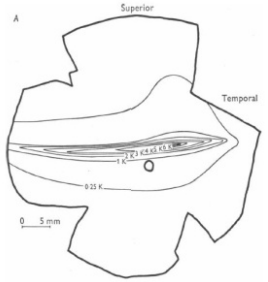
(c) L. Jonaitis 2011

- Humans / Primates
  - central fovea
  - radially symmetrical



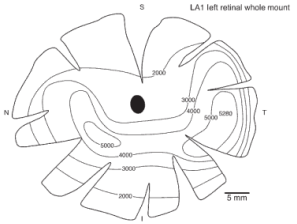
Schwartz 1980, Ruesch 2010

- Many Herbivores
  (Sheep, Pigs, Horses, Kangaroos)
  - elongated area with high receptor density "visual streak"



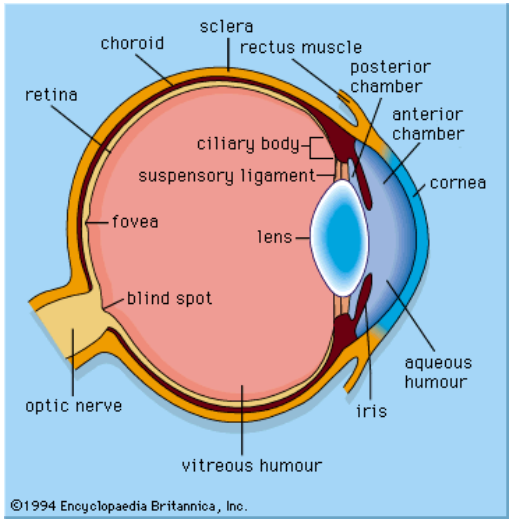
Hughes 1975

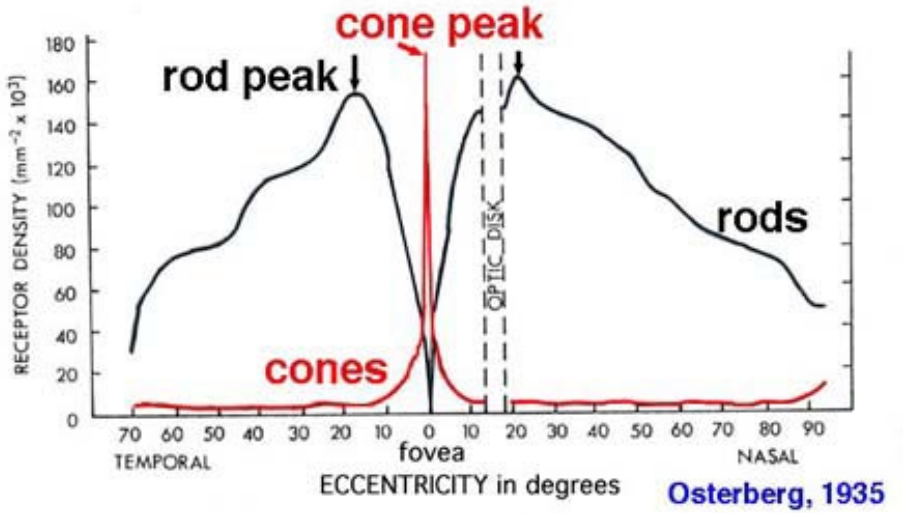
- Elephants
  - sensor topology which combines a fovea with a visual streak

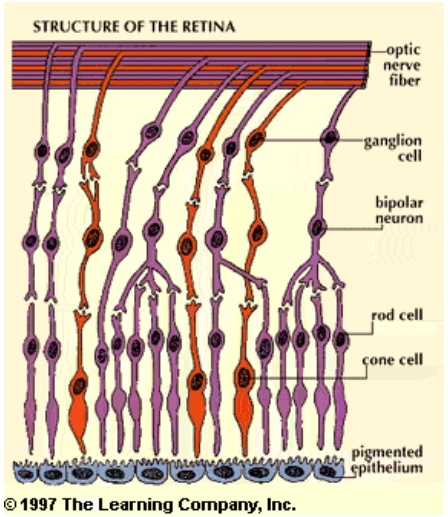


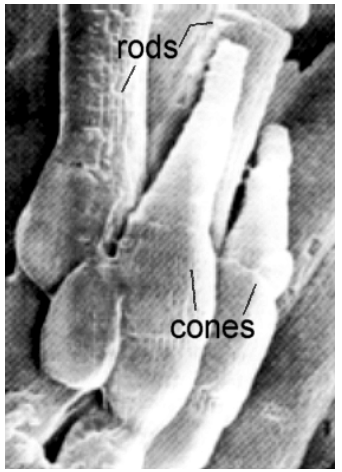
Pettigrew et al. 2010

The Human Eye





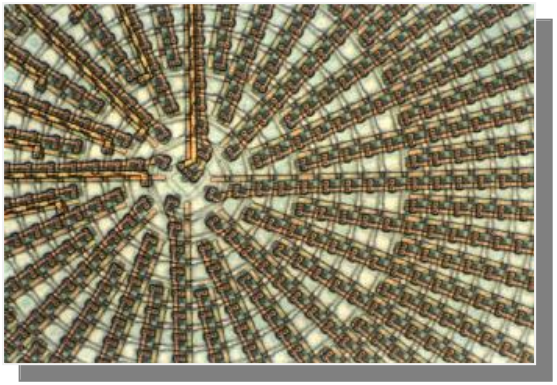


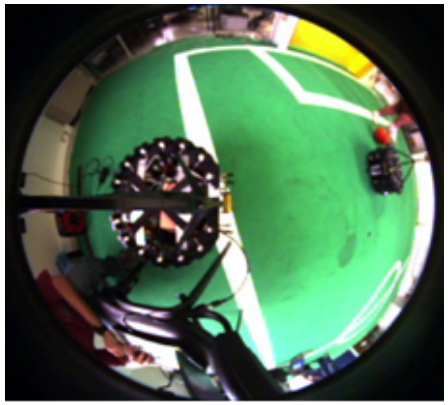


Institute for Systems and Robotics | LISBOA

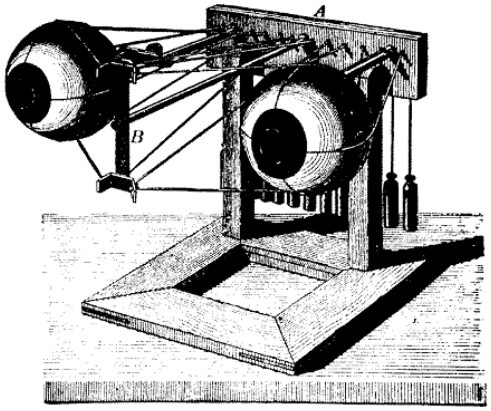
Engineering Human Vision

Space Variant Vision – Hardwired Computational Savings.



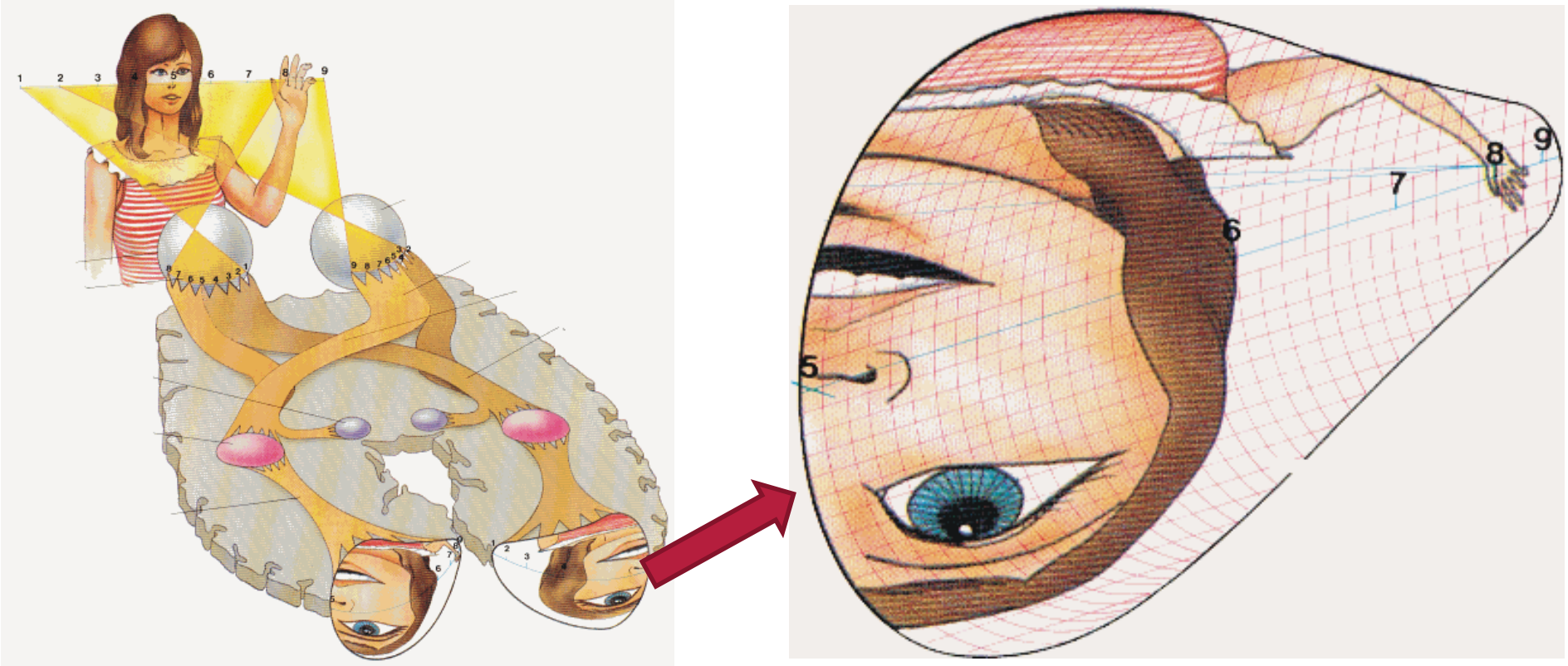


Selective Attention – Gaze Shifts, Dynamical Allocation



Foveal Vision

The Retino-Cortical Mapping

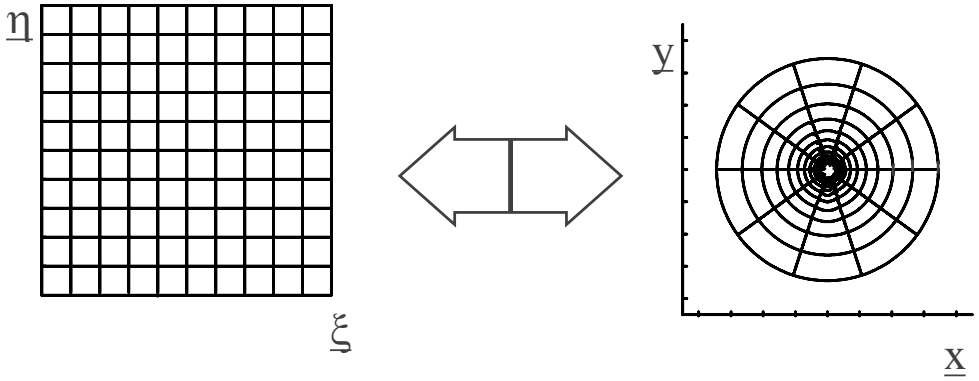


Institute for Systems and Robotics | LISBOA

Log-Polar Map

The log-polar map approximates retino-cortical mapping

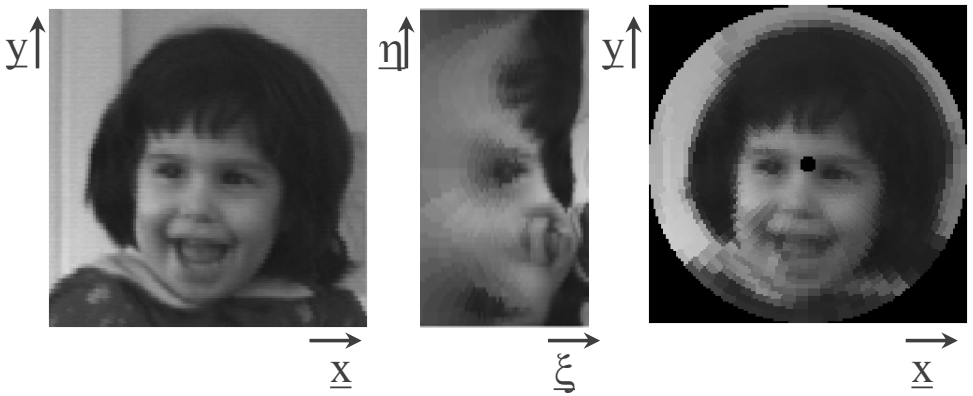
[Schwartz 77, Sandini 80, Wilson 83]



$$\begin{cases} \xi = \log\left(\sqrt{x^2 + y^2}\right) \\ \eta = \arctan\left(\frac{y}{x}\right) \end{cases}$$

Radial coordinate

Angular coordinate

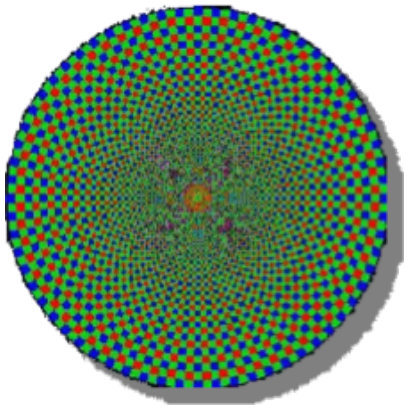


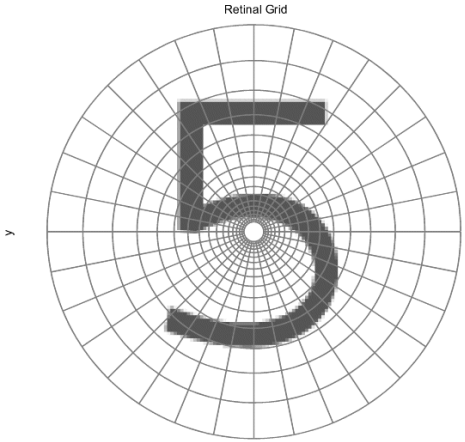
Some properties:

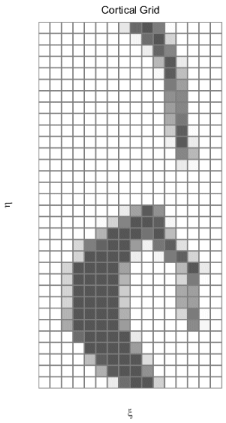
- Similarity to human retina
- Data compression
- Shape invariance to rotation
- Shape invariance to scale

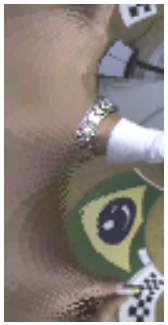
Institute for Systems and Robotics | LISBOA

Log-Polar Sensors



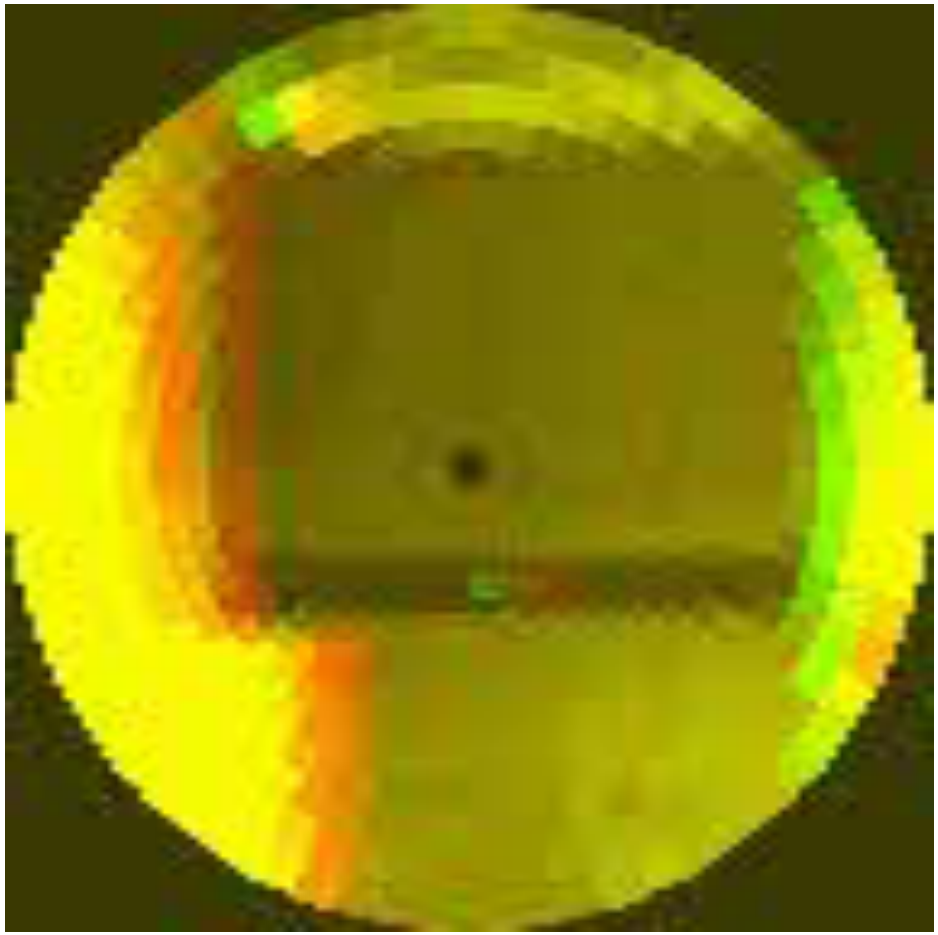


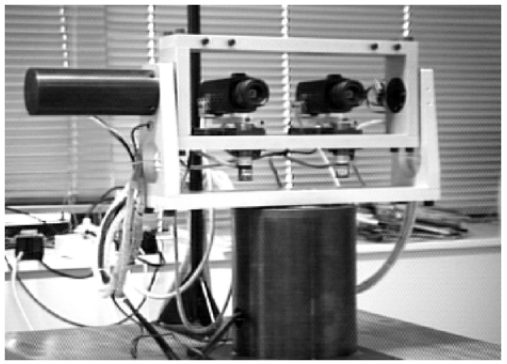


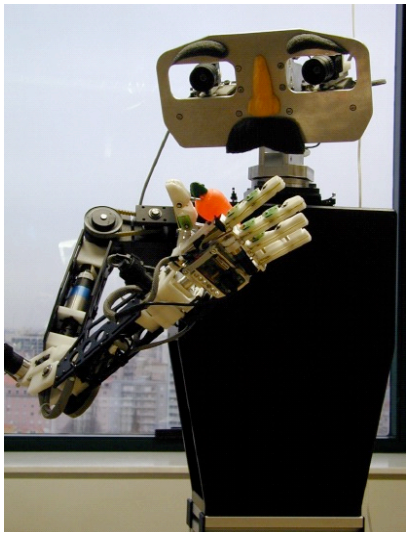


Institute for Systems and Robotics | LISBOA

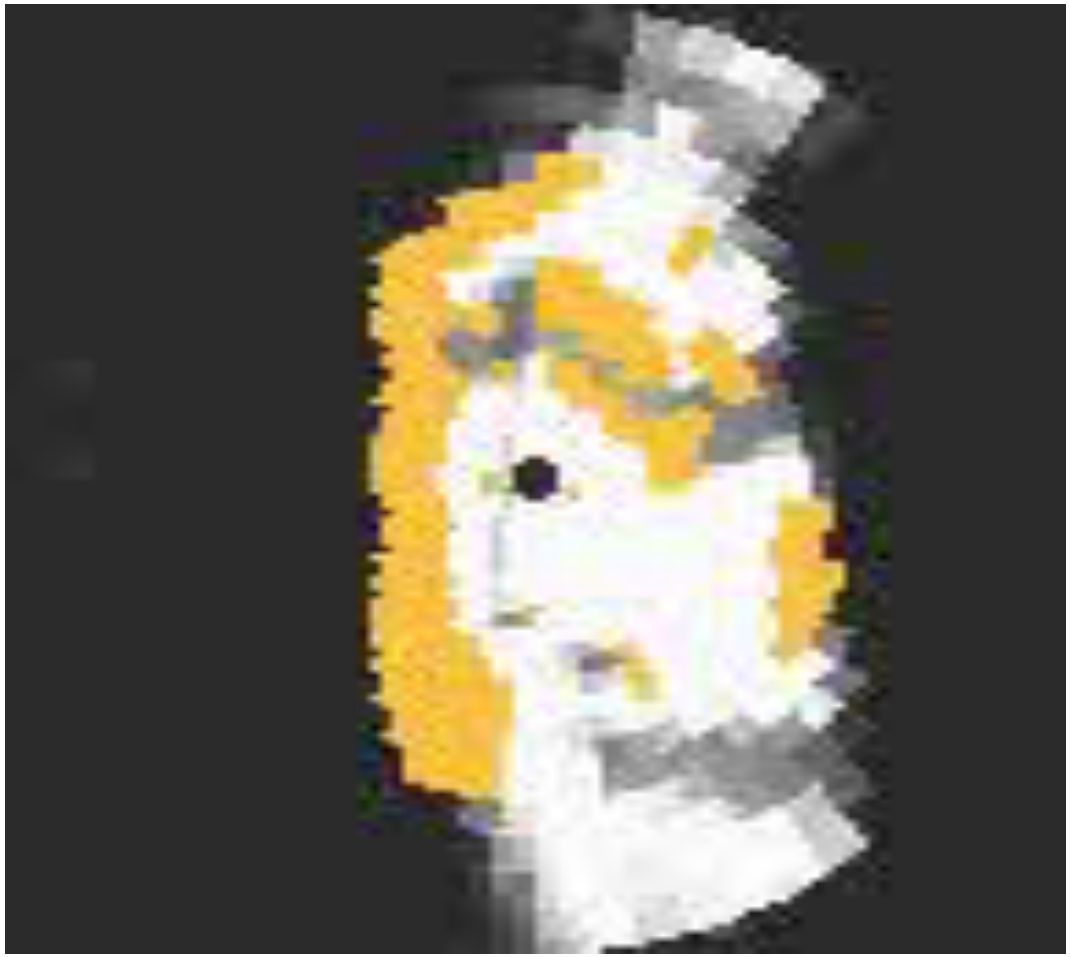
Vergence Control





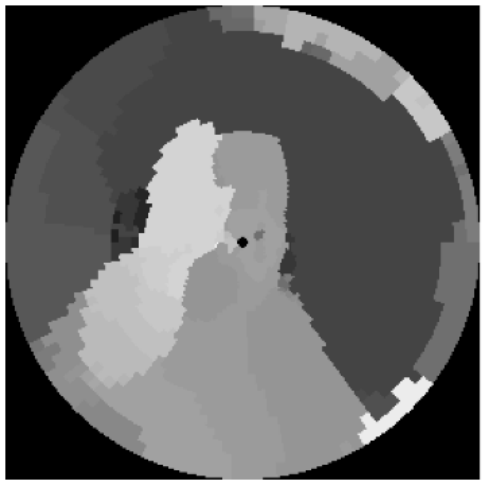


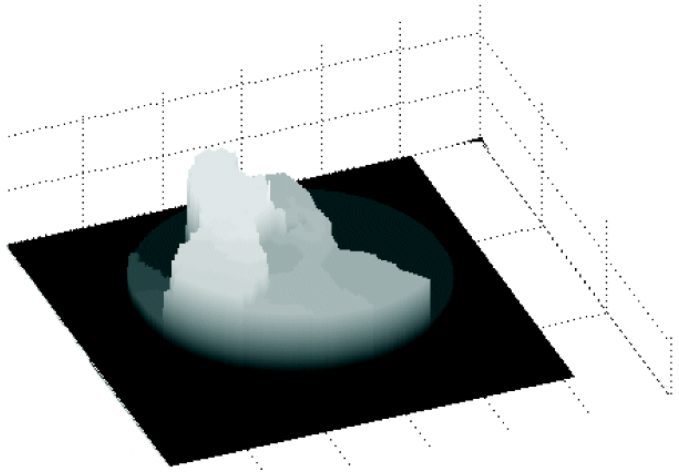
Binocular Tracking



Motion Estimation

Stereo Depth Maps





Challenge 1

Back in 1996 (during my MSc), I was able to track my Hand with an active stereo head at 50Hz.

Try to do the same with the iCub today at 60Hz!

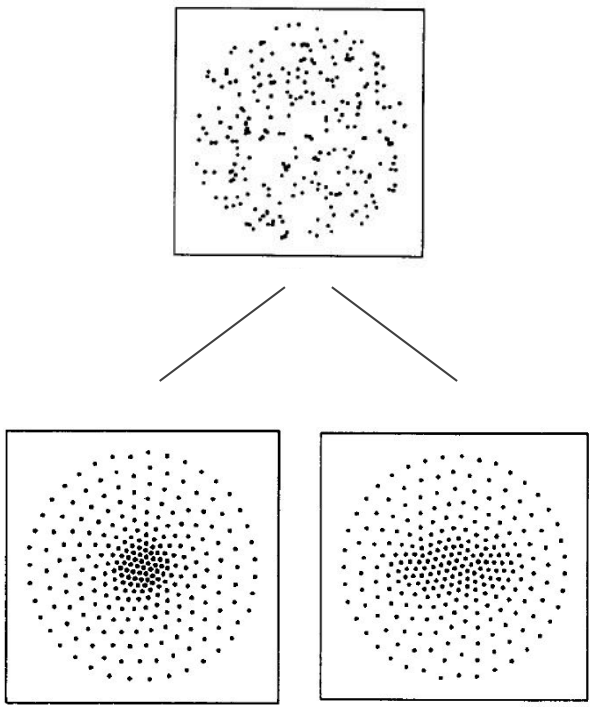
The self-organizing retina

How the define the optimal retina for an agent?

Institute for Systems and Robotics | LISBOA

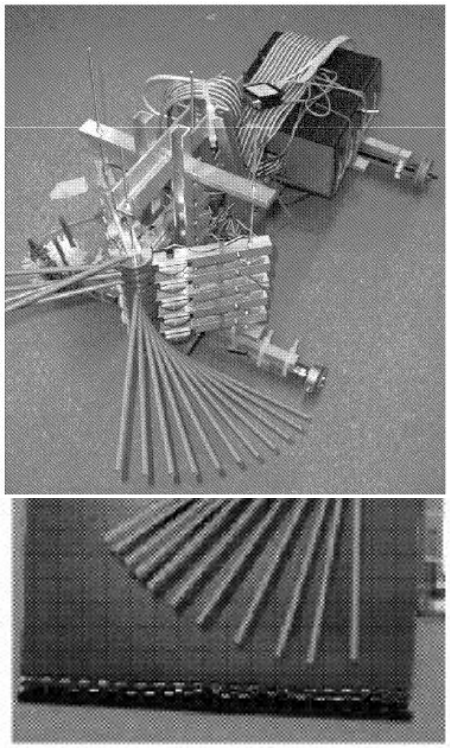
The self-organizing retina

Clippingdale et al. (1996):



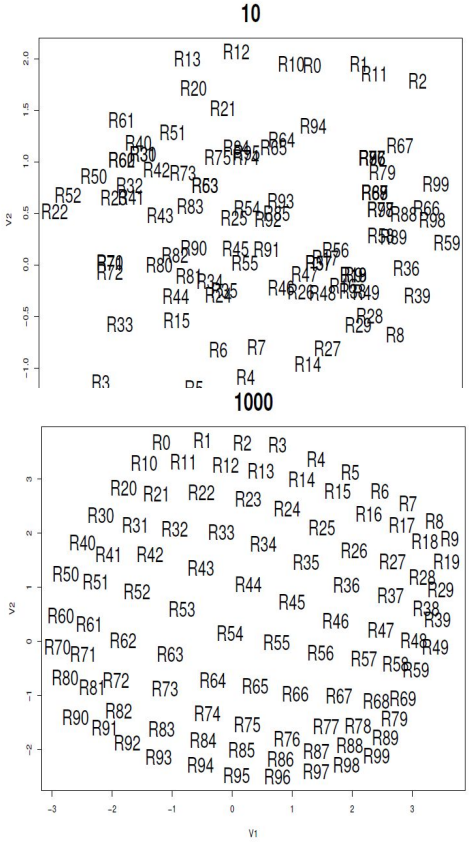
Synthesize abstract, behavior dependent sensor topologies.

Lichtensteiger et al. (1999)



Evolve a 1-dimensional receptor distribution which simplifies a task.

Olsson et al. (2006)



Reconstruct an unknown sensor topology.

Institute for Systems and Robotics | LISBOA

The self-organizing retina

How the perceptual and motor representations of an agent can be simultaneously optimized by self-experience ?

Jonas Ruesch
(and author of most slides)

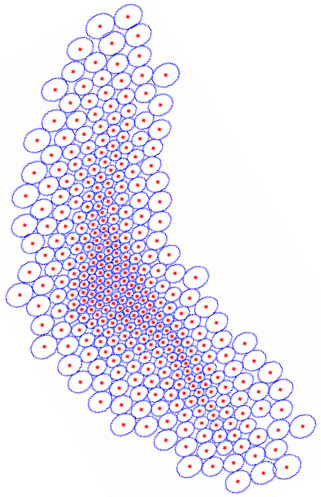
Ricardo Ferreira

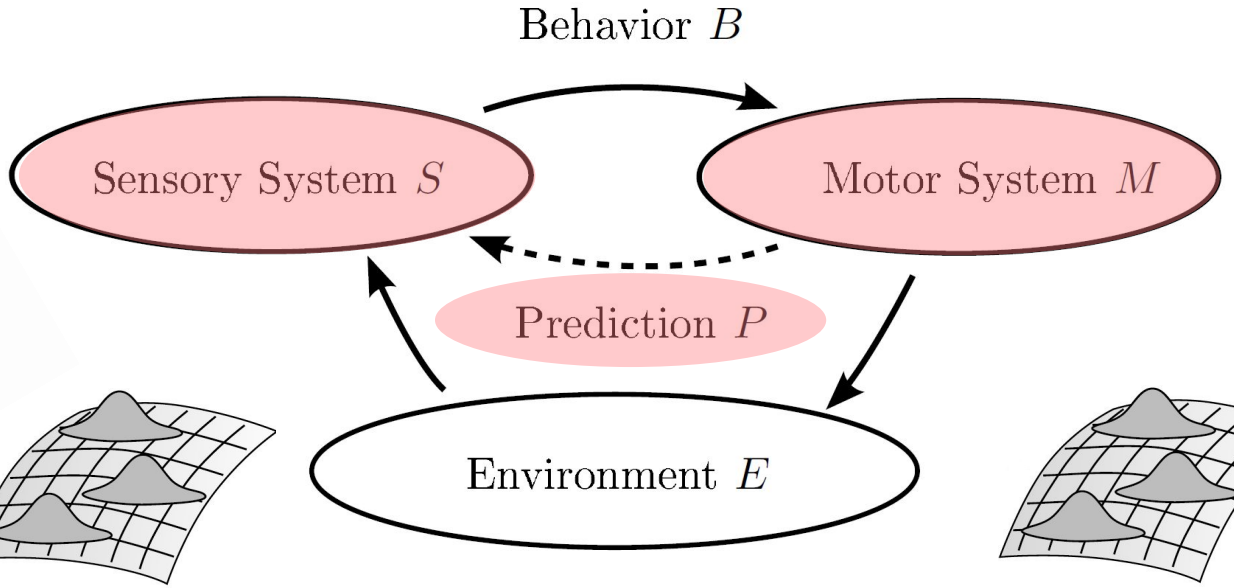
- A Computational Approach on the Co-Development of Visual Sensorimotor Structures. J. Ruesch, R. Ferreira, A. Bernardino. Adaptive Behavior, 21(6):452-464, December 2013.
- Predicting visual stimuli from self-induced actions: an adaptive model of a corollary discharge circuit. J. Ruesch, R. Ferreira, A. Bernardino, IEEE Transactions on Autonomous Mental Development, 4(4):290-304, Dec 2012.
- Self-organization of Visual Sensor Topologies Based on Spatio-temporal Cross-correlation, J. Ruesch, R. Ferreira and A. Bernardino. Simulation of Adaptive Behaviour 2012, Odense, Denmark.

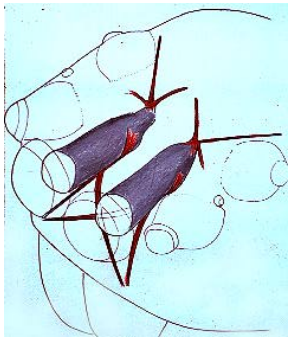
Institute for Systems and Robotics | LISBOA

Co-development

Sensory and motor systems are adapted to each other and to the environment.







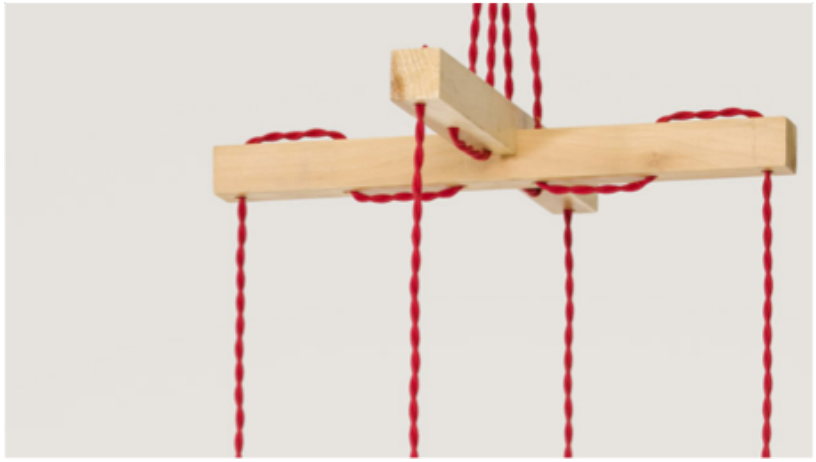
sensory receptive fields

sensorimotor loop

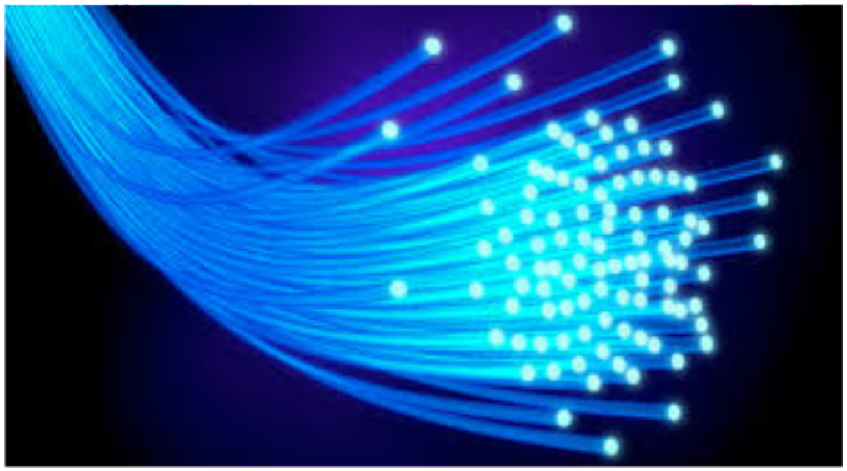
motor movement fields

Co-development

motor space



sensor space



? ⟷ ?

?

Basic Principles

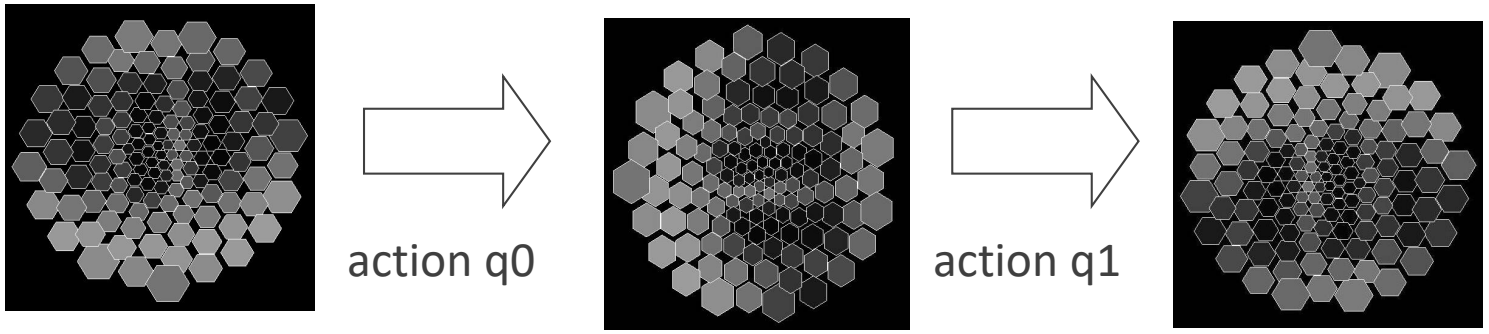
I – Self-Exploration

II – Receptive Fields

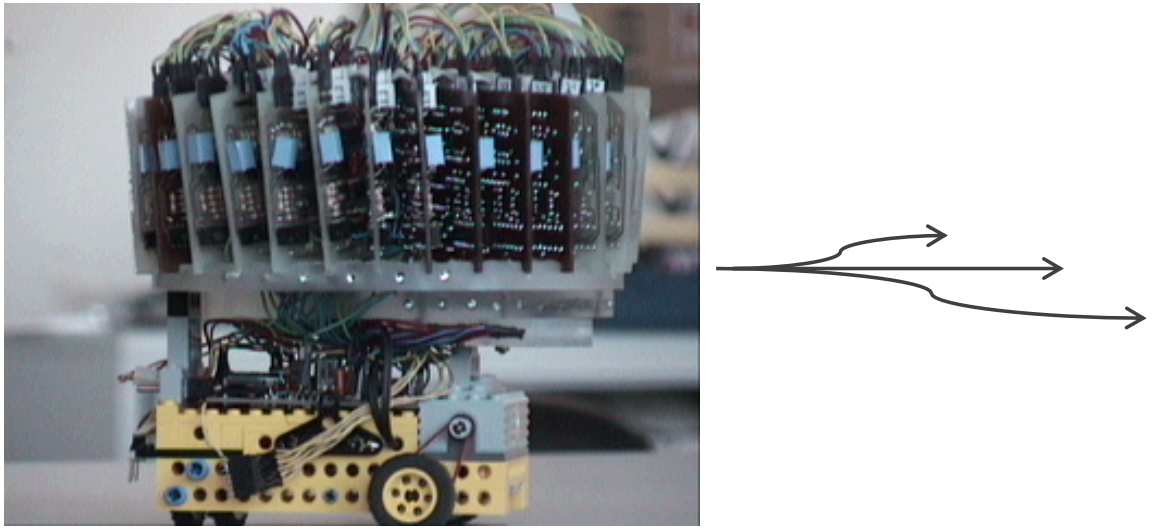
III – Prediction Error Minimization

IV - Simplicity

Self Exploration



stimulus i0 stimulus i1 stimulus i2

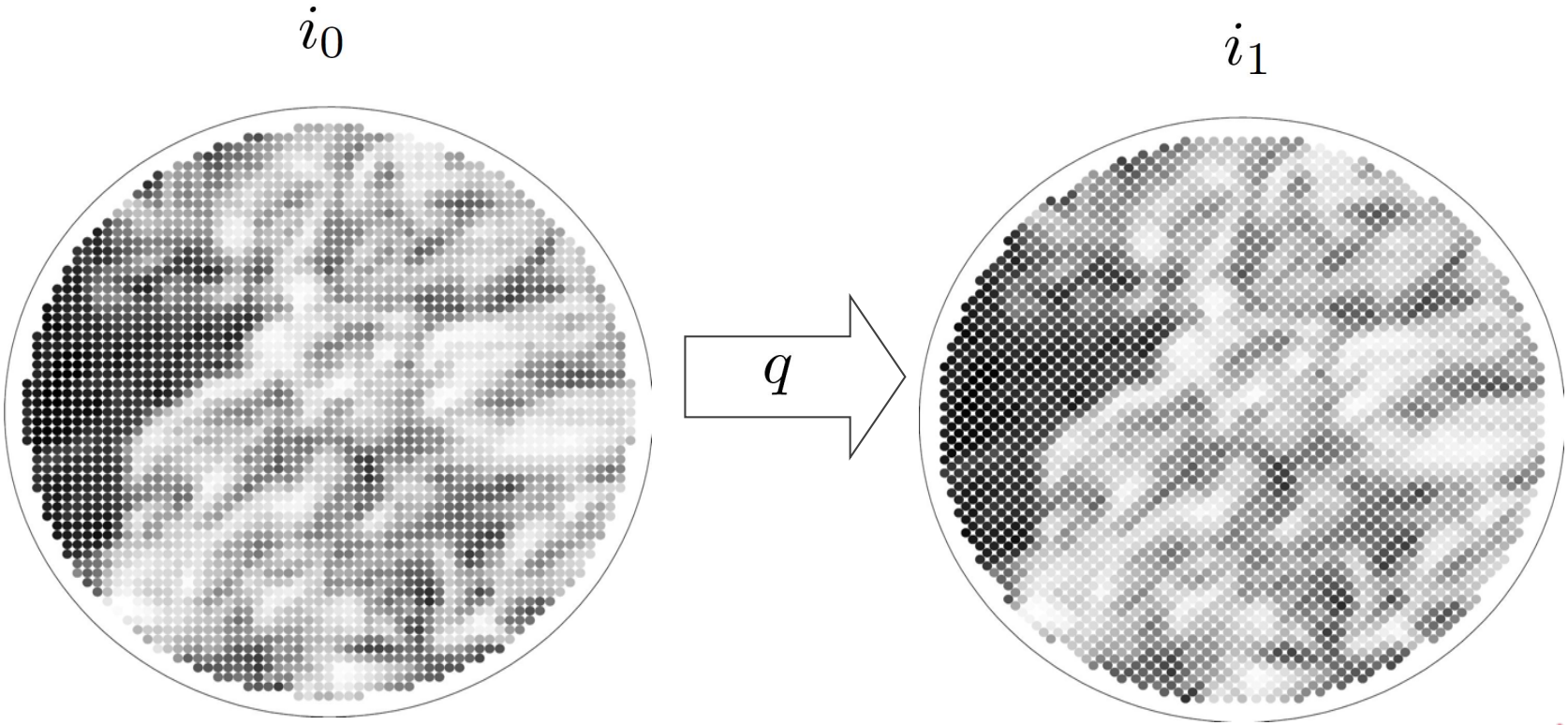


Alex Bernardino, VVV 2018

Experience Atoms

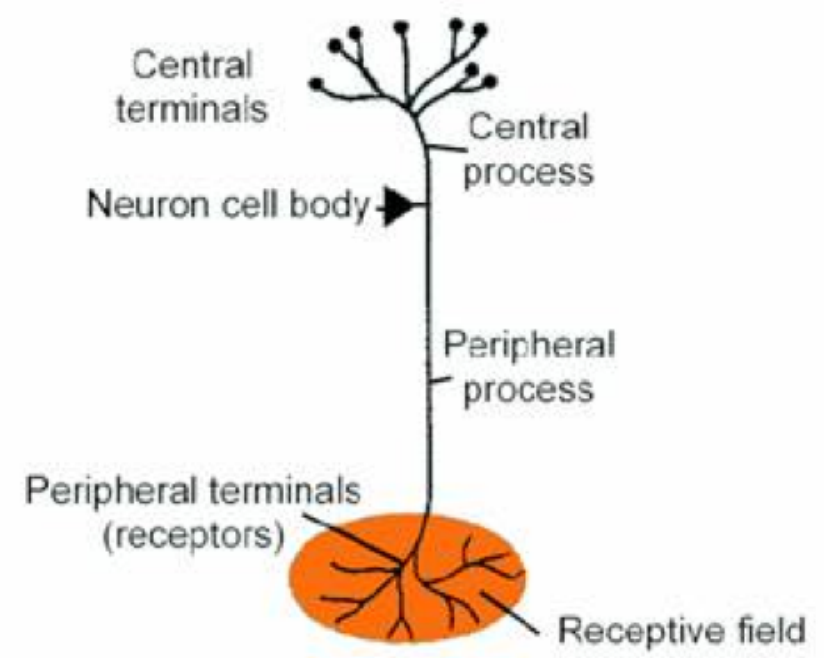
Sensorimotor experience is collected as a set of action-stimulus triplets while an agent explores the environment:

$$\{(i_0, i_1, q)\}$$



Receptive Fields

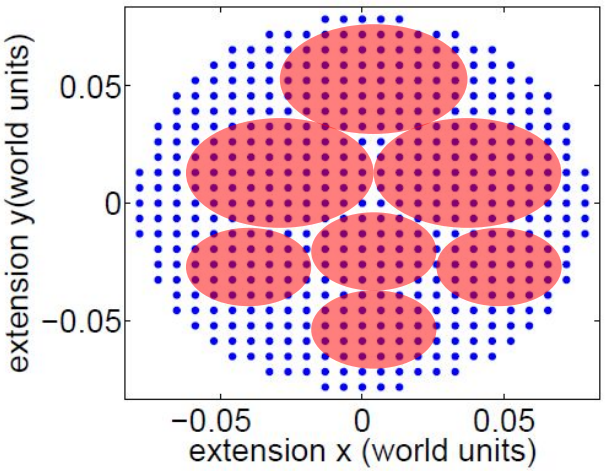
"Receptive Fields are probably the most prominent and ubiquitous computational mechanisms employed by biological information processing systems". (Weiss & Edelman (1993)).

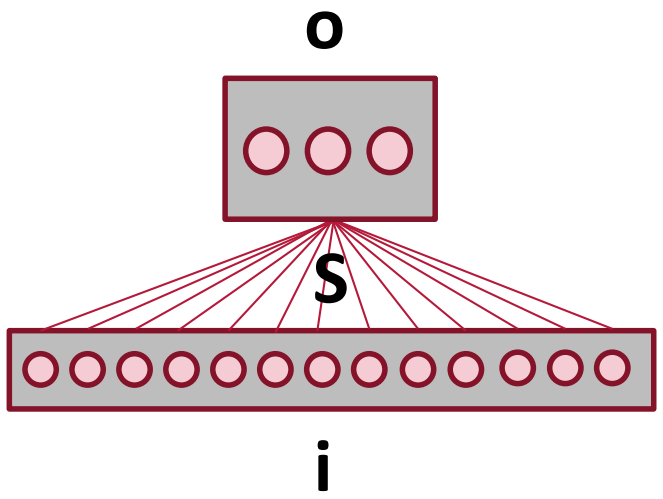


Institute for Systems and Robotics | LISBOA

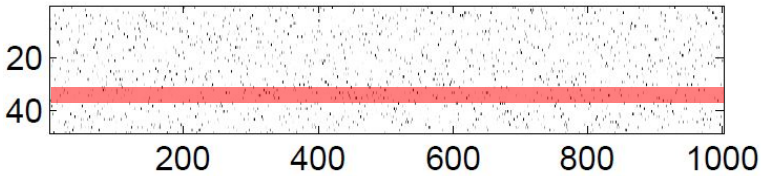
Sensor Receptive Field Model

Sensory System S





Discretized S encoded in a matrix **S**:

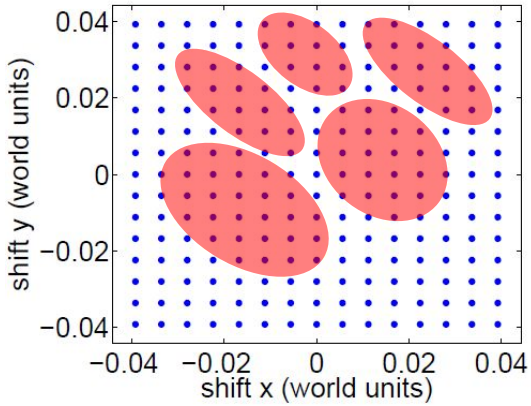


$$\mathbf{o} = \mathbf{Si}$$

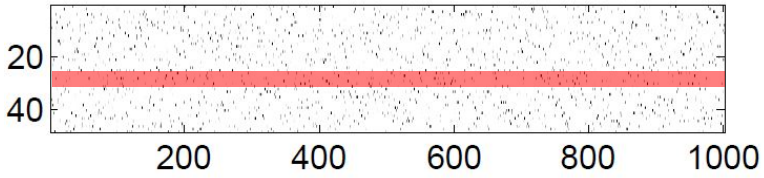
Institute for Systems and Robotics | LISBOA

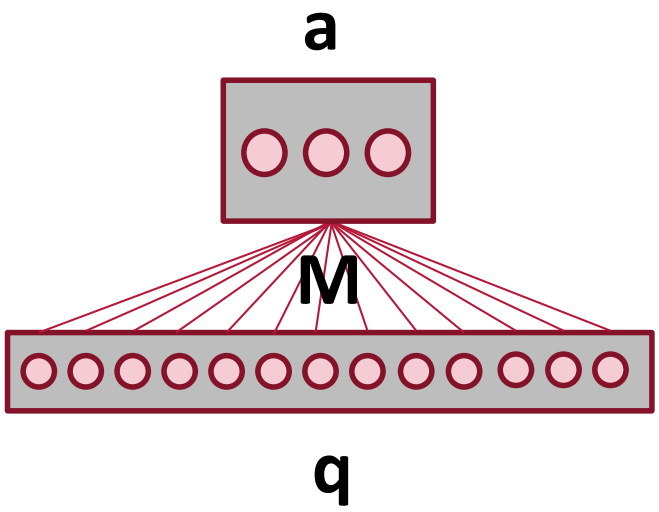
Motor Receptive Field Model

Motor System M



Discretized M encoded in a matrix **M**:



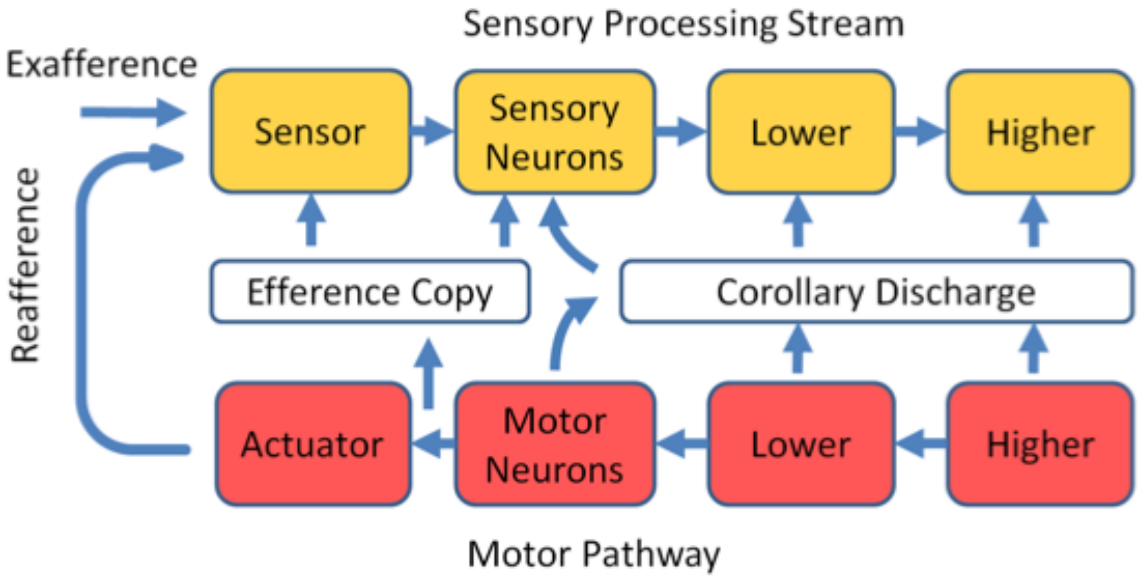


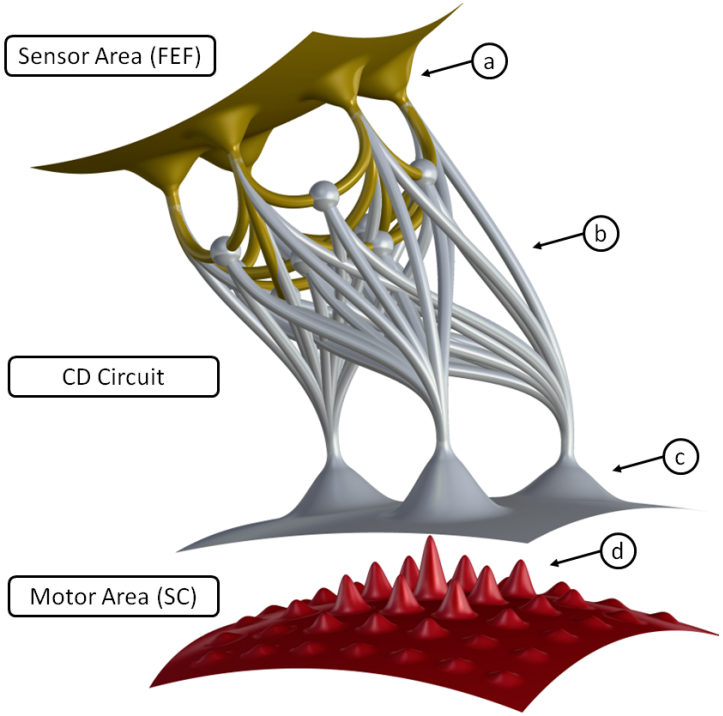
$$\mathbf{a} = \mathbf{M}\mathbf{q}$$

$$\hat{q} = \mathbf{E}_q \left\{ \mathbf{M}^T \mathbf{a} \right\}$$

Prediction

- Corollary discharge circuits which predict stimuli resulting from self-induced actions are ubiquitous in living organisms (Crapse & Sommer (2008)).

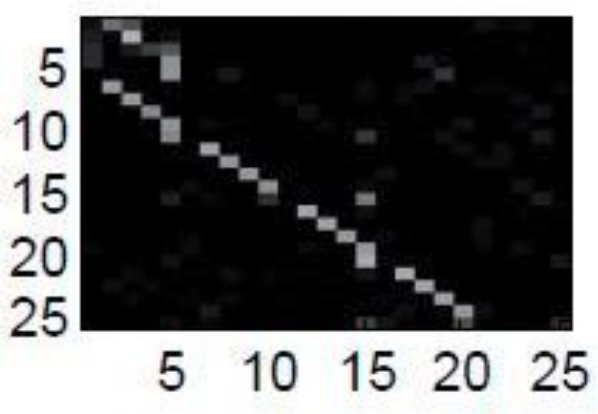




Ruesch J., Ferreira R., Bernardino A. Predicting visual stimuli from self-induced actions: an adaptive model of a corollary discharge circuit. Autonomous Mental Development (TAMD) (2012)

Prediction Model

Prediction Model P



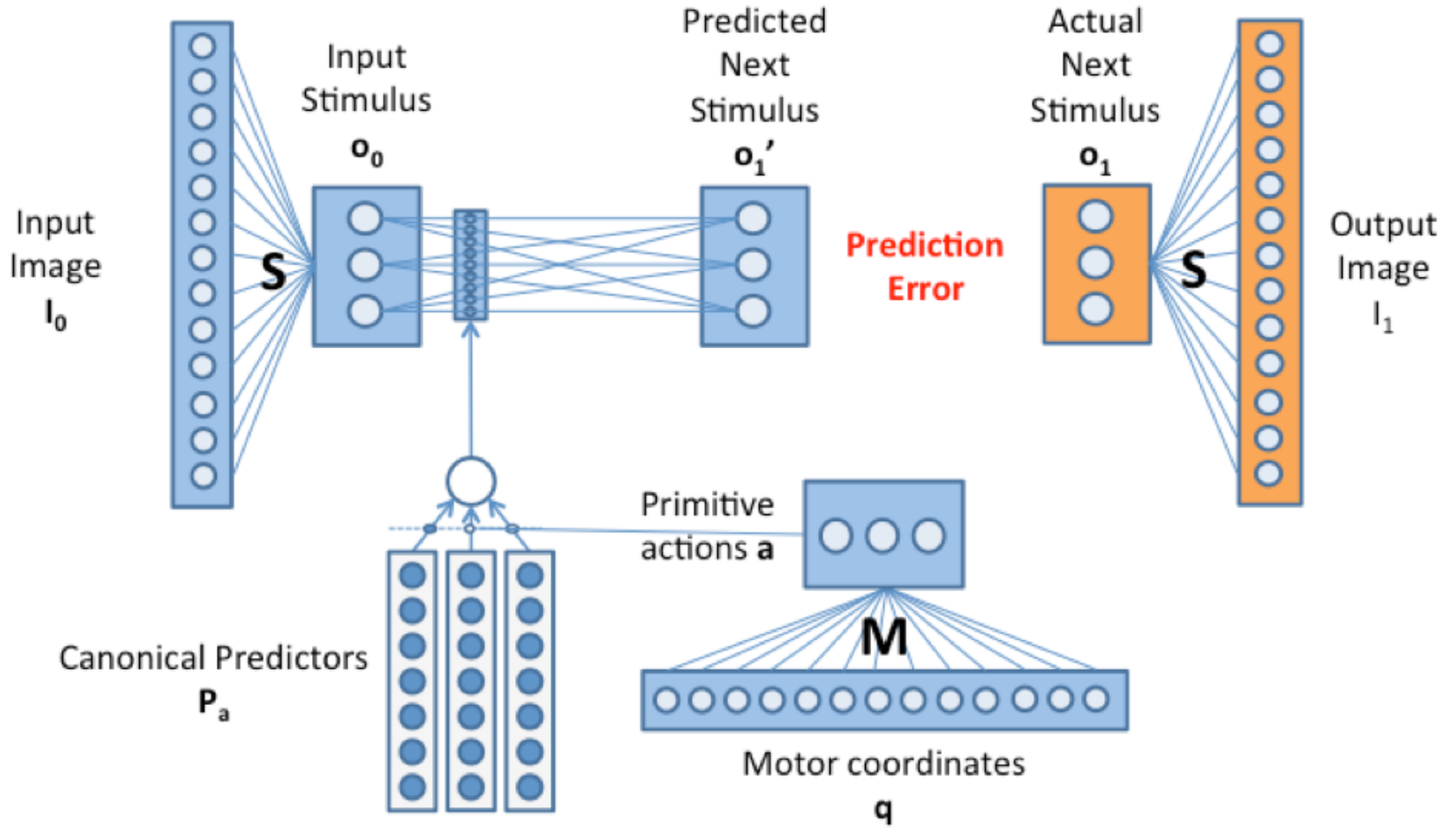
- Each motor field $\mathbf{m}_k$ has one canonical predictor matrix ($\mathbf{P}_k$).
- An arbitrary predictor can be expressed by linear combination of canonical predictors.

$$\mathbf{o}_1 = \mathbf{P}\mathbf{o}_0$$

$$\mathbf{P}^q = \sum_{k=1}^{n_m} \overbrace{\left(\mathbf{m}_k^T \mathbf{q}^q\right)}^{a_k} \mathbf{P}_k$$

Institute for Systems and Robotics | LISBOA

Prediction Error



Prediction Error Minimization

1. Use **lifelong sensorimotor experience.**
2. Use **receptive fields** as dimensionality reduction principle.
3. Minimize **stimulus prediction** error.
4. Enforce **sparsity** through non-negative constraints.

$$C = \sum_t || \overbrace{\mathbf{S}^T \sum_{k=1}^{n_m} (\mathbf{m}_k^T \mathbf{q}^q) \mathbf{P}_k \mathbf{S} \mathbf{i}_t}^{\hat{i}_{t+1}} - \mathbf{i}_{t+1} ||^2$$

$$s.t. \quad \mathbf{P} \geq 0, \mathbf{S} \geq 0, \mathbf{M} \geq 0$$

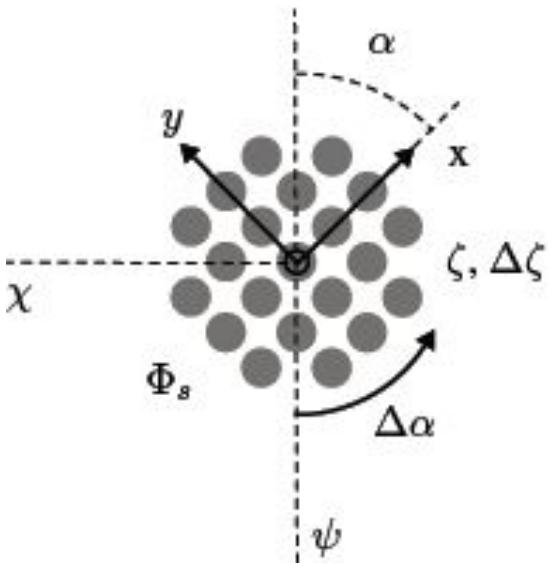
Optimization by Projected Gradient Descent (MATLAB)

Institute for Systems and Robotics | LISBOA

Experiments – Dataset

Actions:

- Dilate
- Rotate
- Shift horizontally
- Shift vertically



Experiment 1 - Sensor Design

2-dimensional motor space with 60 fixed motor fields

- Translation x,y,

  or

- Rotation and scale

Image space discretized in 2877 pixels

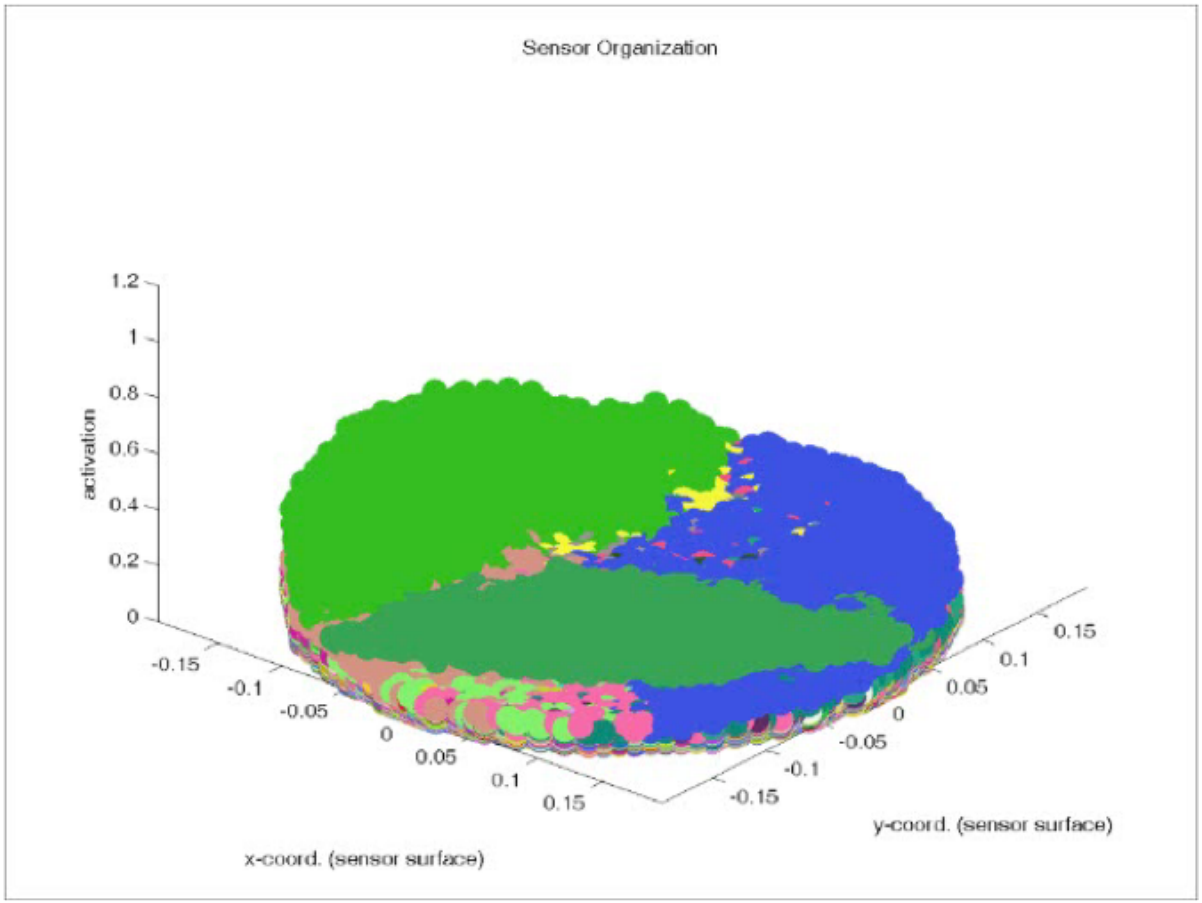
48 receptive fields

About 275’000 variables to optimize

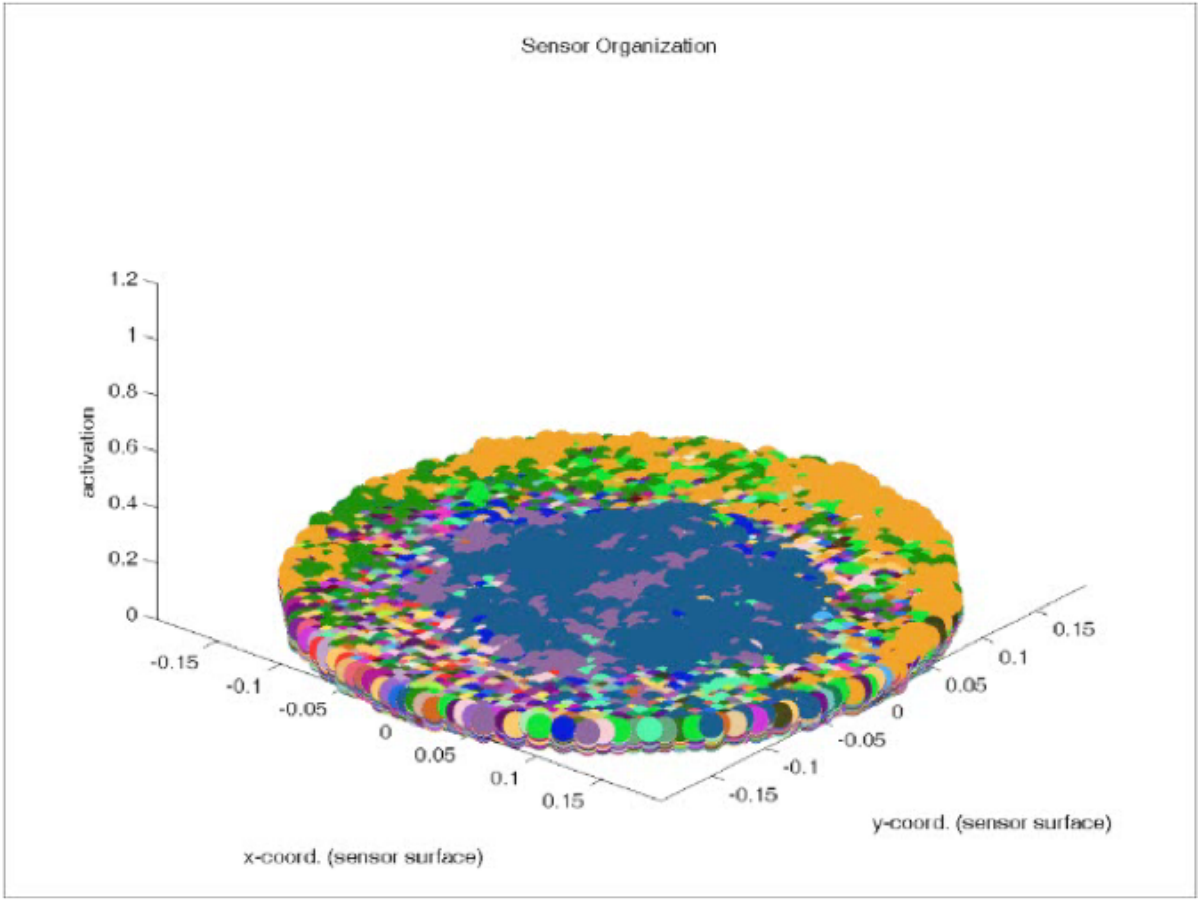
4080 experience triplets (68 per action).

Institute for Systems and Robotics | LISBOA

Translation Actions



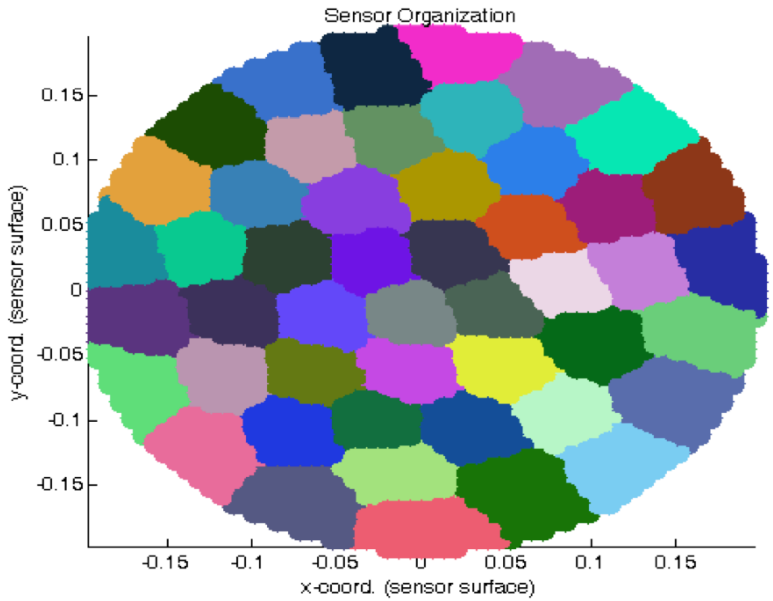
Rotation and Dilation Actions



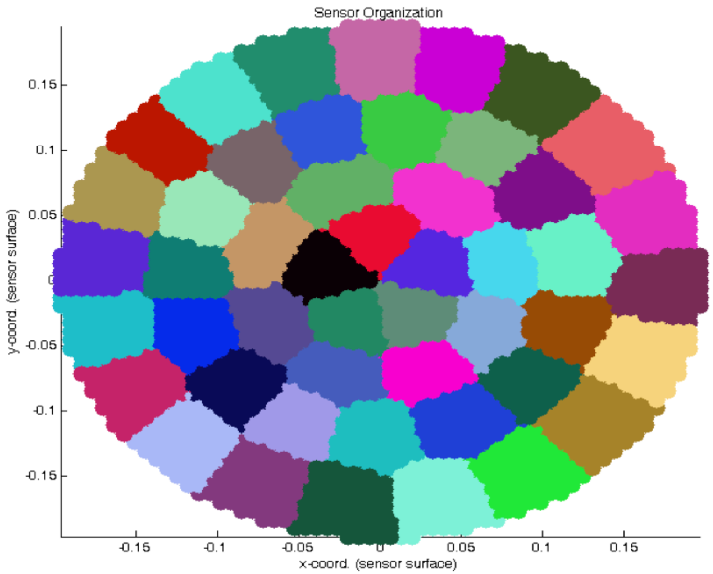
Institute for Systems and Robotics | LISBOA

Sensor Organization

Translation actions



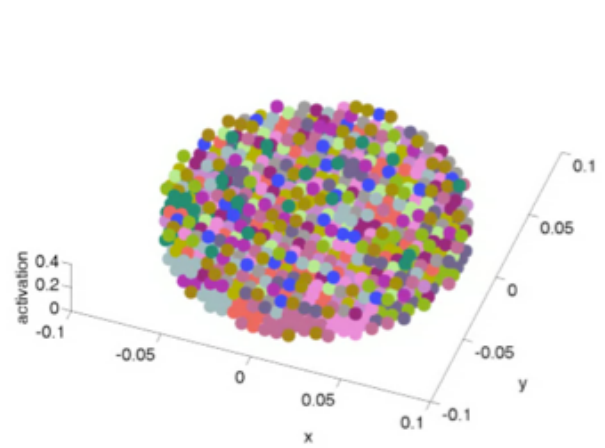
Rotation and Dilation Actions

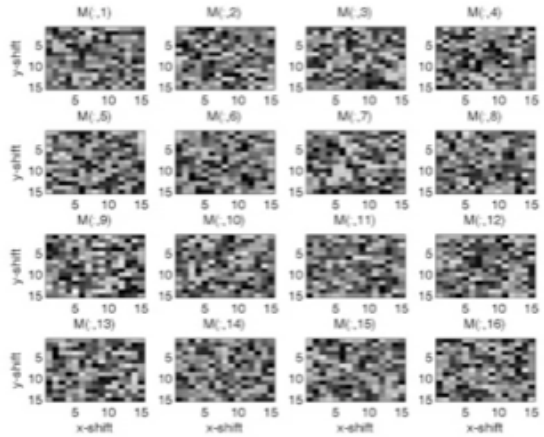


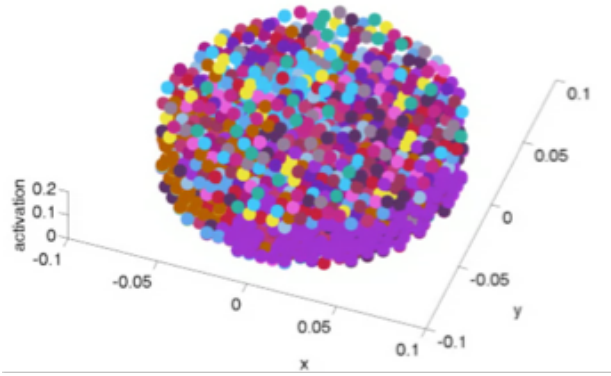
Institute for Systems and Robotics | LISBOA

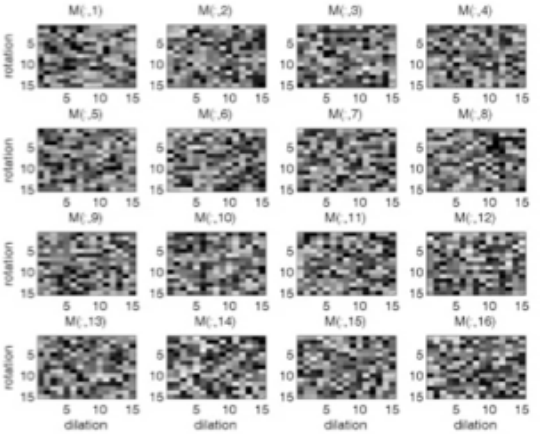
Experiment 2 – Co-development

- 2-D motor space with 225 cells.
  - Translation x,y
  - Rotation and scale
- Image space with 481 pixels
- 16 receptive fields
- 16 motor fields
- About **15’000** variables to optimize
- 22’500 experience triplets with random (uniform) actions.





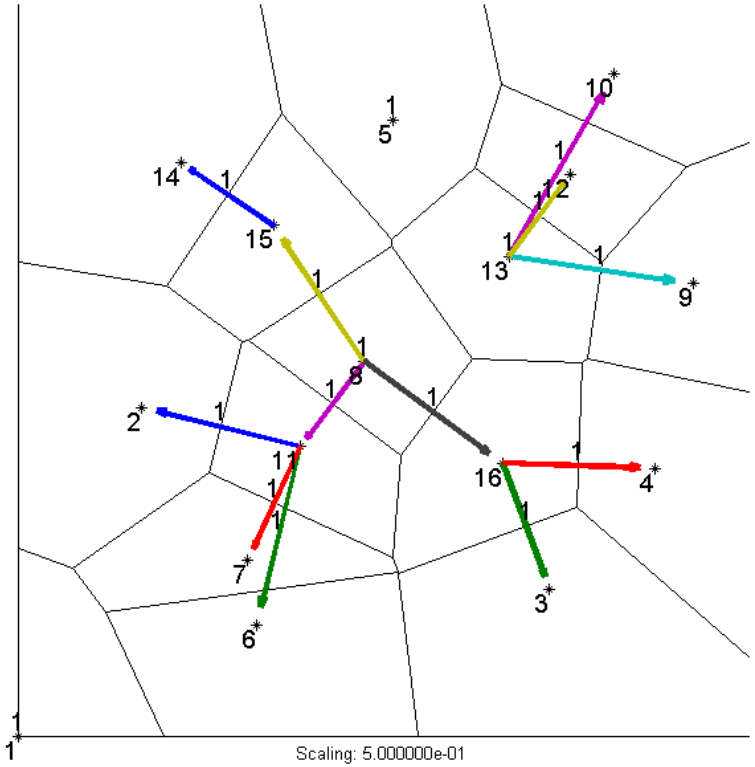




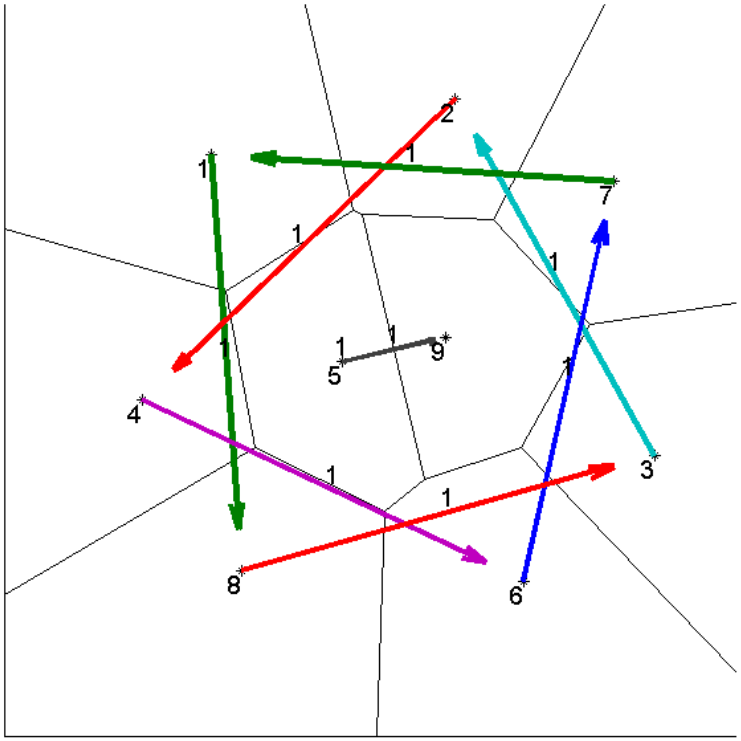
Institute for Systems and Robotics | LISBOA

Learned Canonical Predictors

Zoom-out action



Rotate right action



Challenge 2

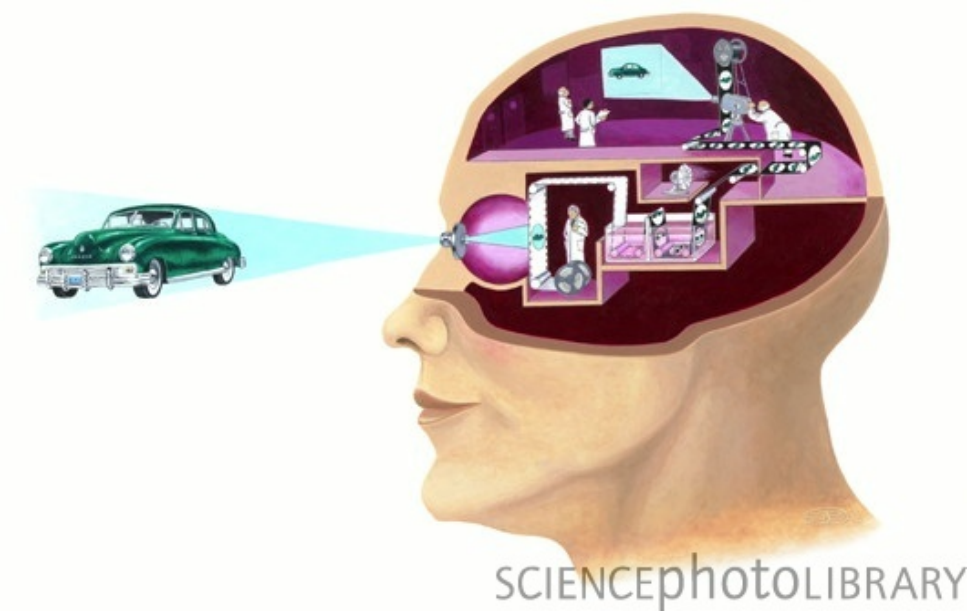
Use deep learning solvers to optimize the retinas!

Visual Attention

Visual Attention

Even though a smart hardwired sampling on the visual field may reduce image size and simplify computations for some visual tasks, this is not enough for human like perception.

Humans use the available resources (detection, memory, identification) in order to maximizing reliability and minimizing reaction time. How?



Visual Attention

Attention is the set of mechanisms that tune the search processes in vision to achieve their best performance to a given task.

(John Tsostos, 2009)

Psychophysical Studies

The study of human attention provides important examples of architectural, functional and resource management aspects of a “real-time” performing system.

Many studies on experimental psychology have evaluated the performance limitations of human subjects and are na important source of inspiration.

Institute for Systems and Robotics | LISBOA

Control of Attention

Voluntary ⟷ *Reflexive*

Overt ⟷ *Covert*

Covert Attention

Covert Attention

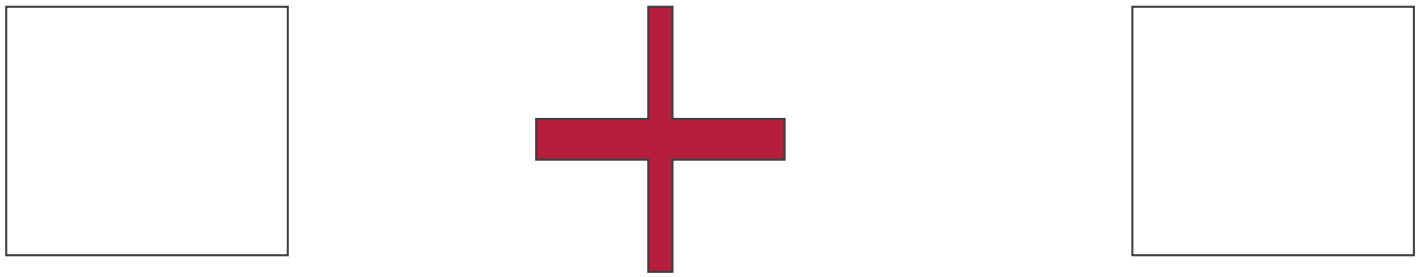
Attention can be oriented *covertly*

- a commonly used metaphor is “the spotlight of attention”

Institute for Systems and Robotics | LISBOA

Covert Attention

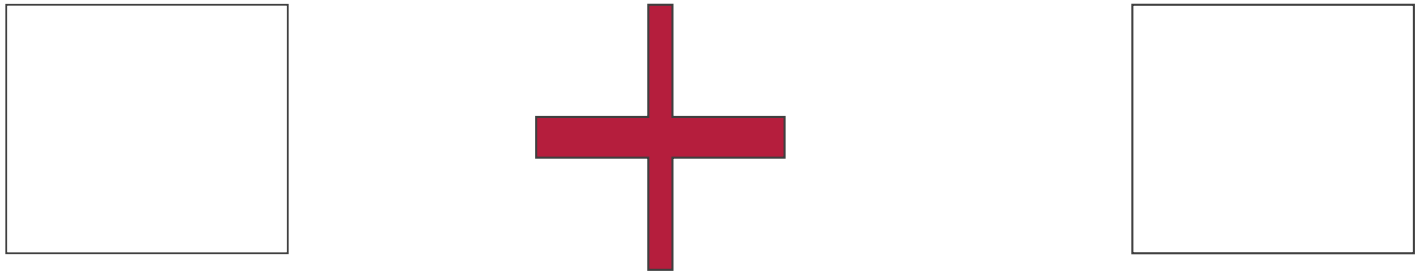
Posner Cue - Target Paradigm:



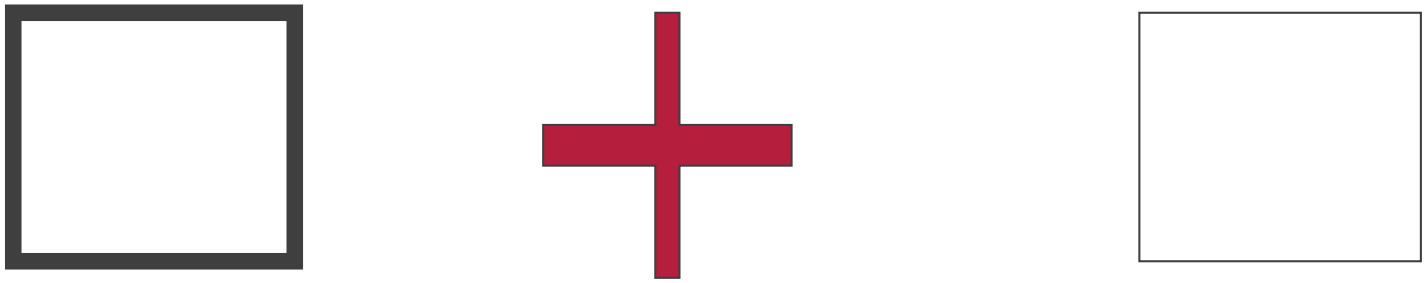
Subject presses a button as soon as x appears

Institute for Systems and Robotics | LISBOA

Covert Attention

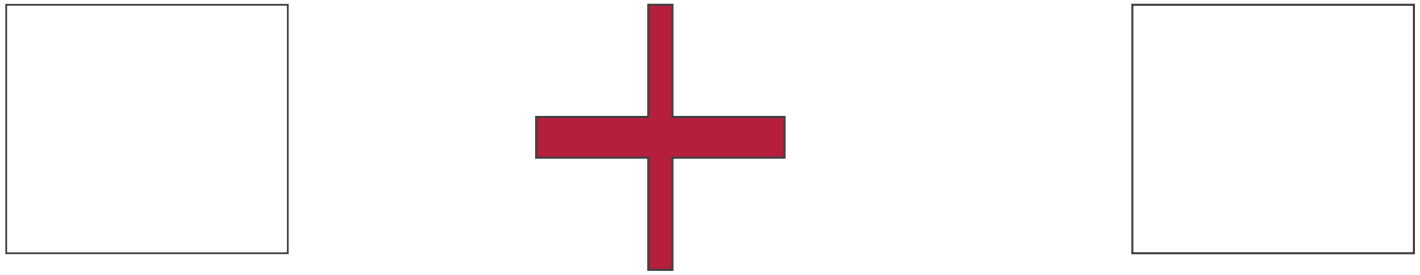


Covert Attention

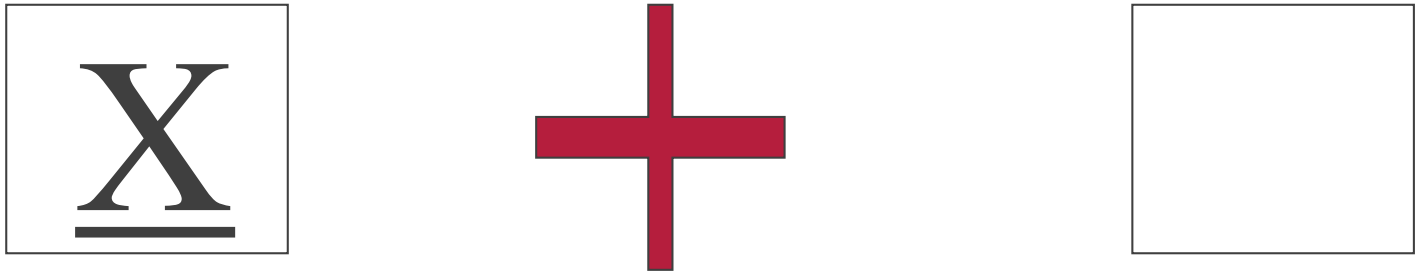


Institute for Systems and Robotics | LISBOA

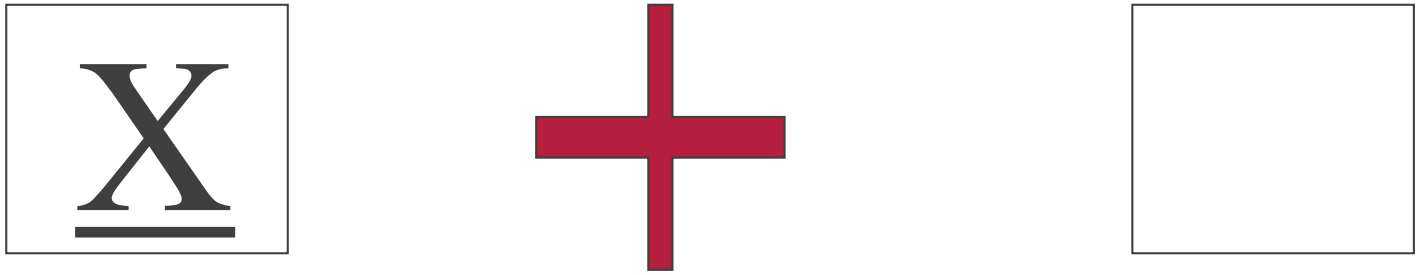
Covert Attention



Covert Attention



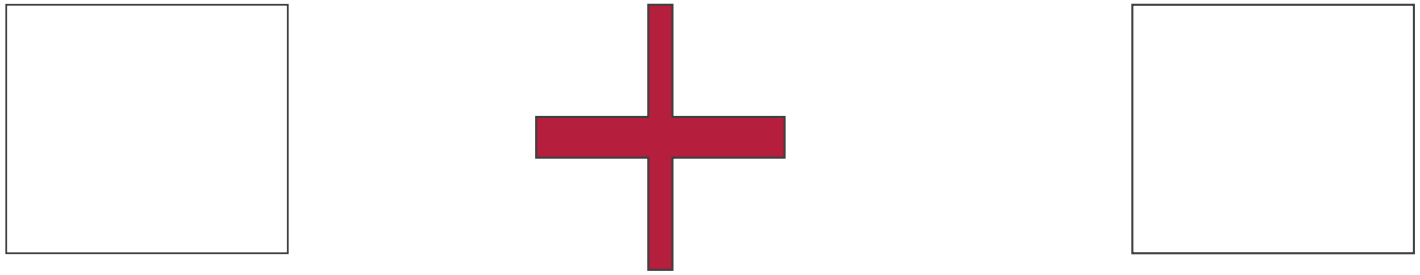
Covert Attention



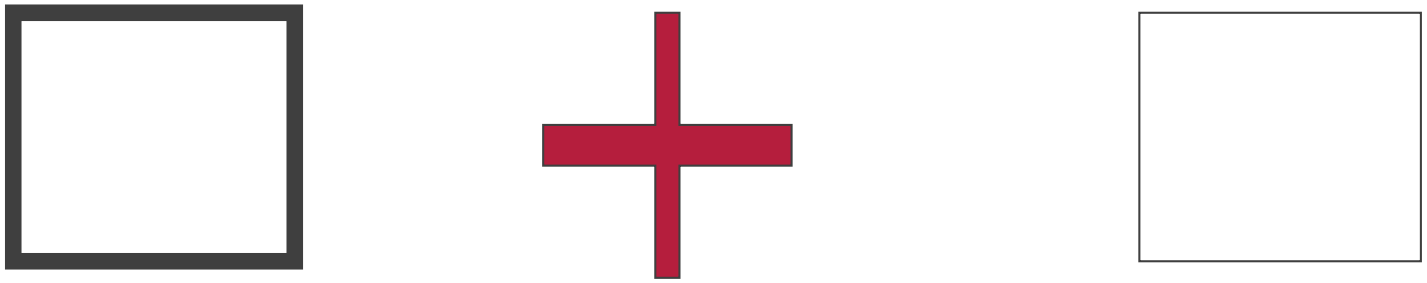
That was a validly cued trial because the x appeared in the box that flashed

Institute for Systems and Robotics | LISBOA

Covert Attention

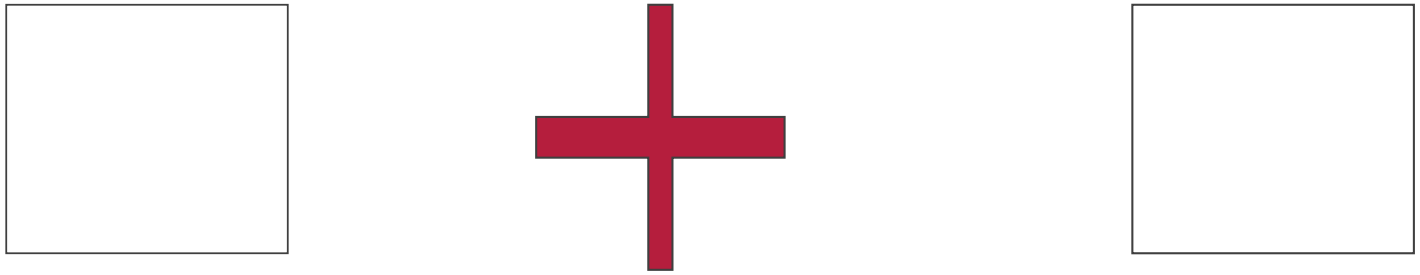


Covert Attention



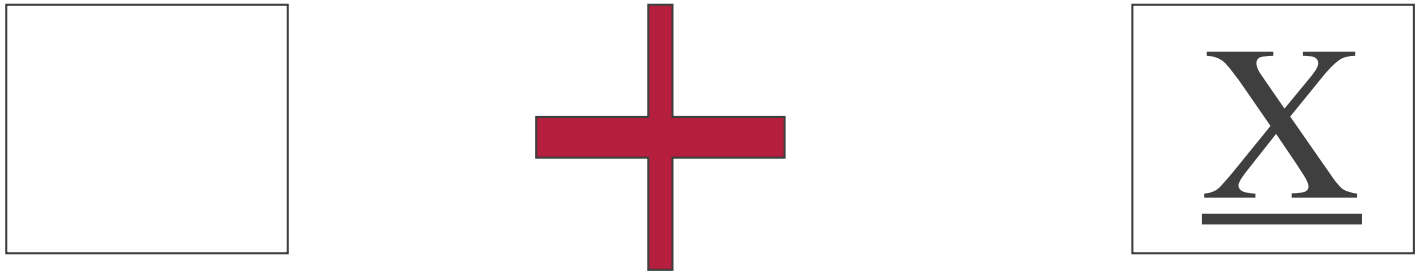
Institute for Systems and Robotics | LISBOA

Covert Attention



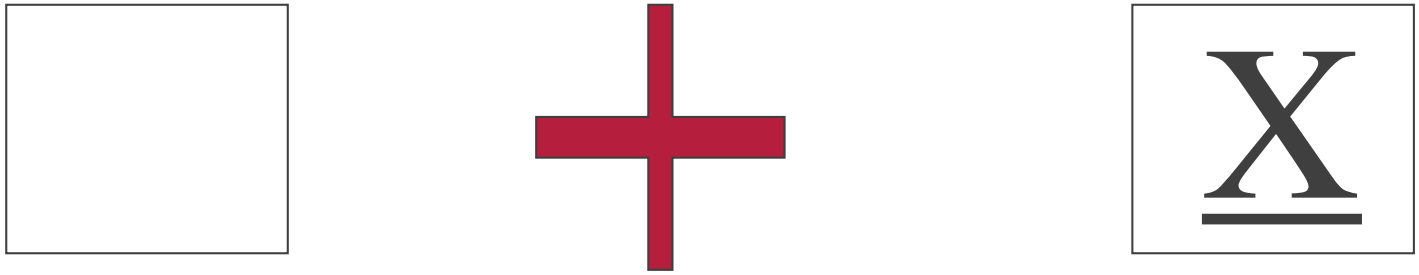
Institute for Systems and Robotics | LISBOA

Covert Attention



Institute for Systems and Robotics | LISBOA

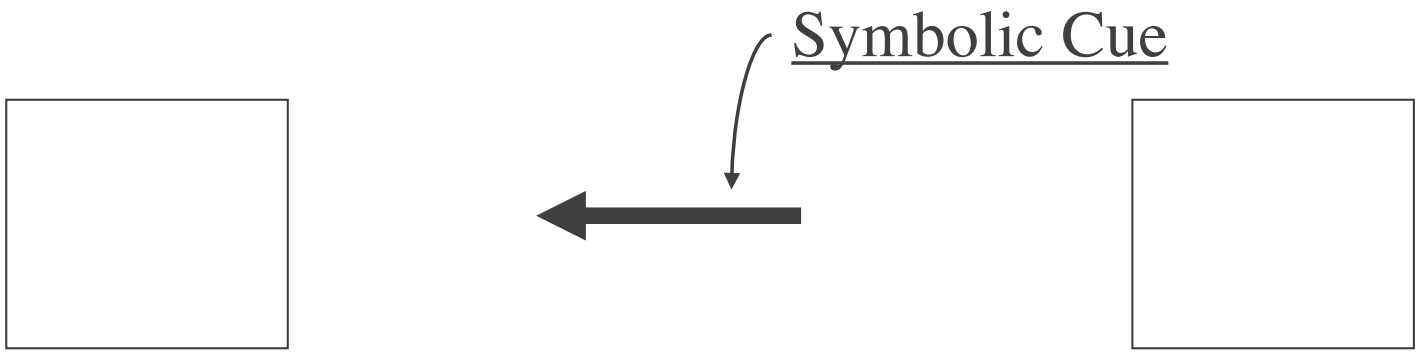
Covert Attention



That was an invalidly cued trial because the x appeared in the box that *didn't* flash

Covert Attention

What is another way to make this paradigm a voluntary orienting paradigm?

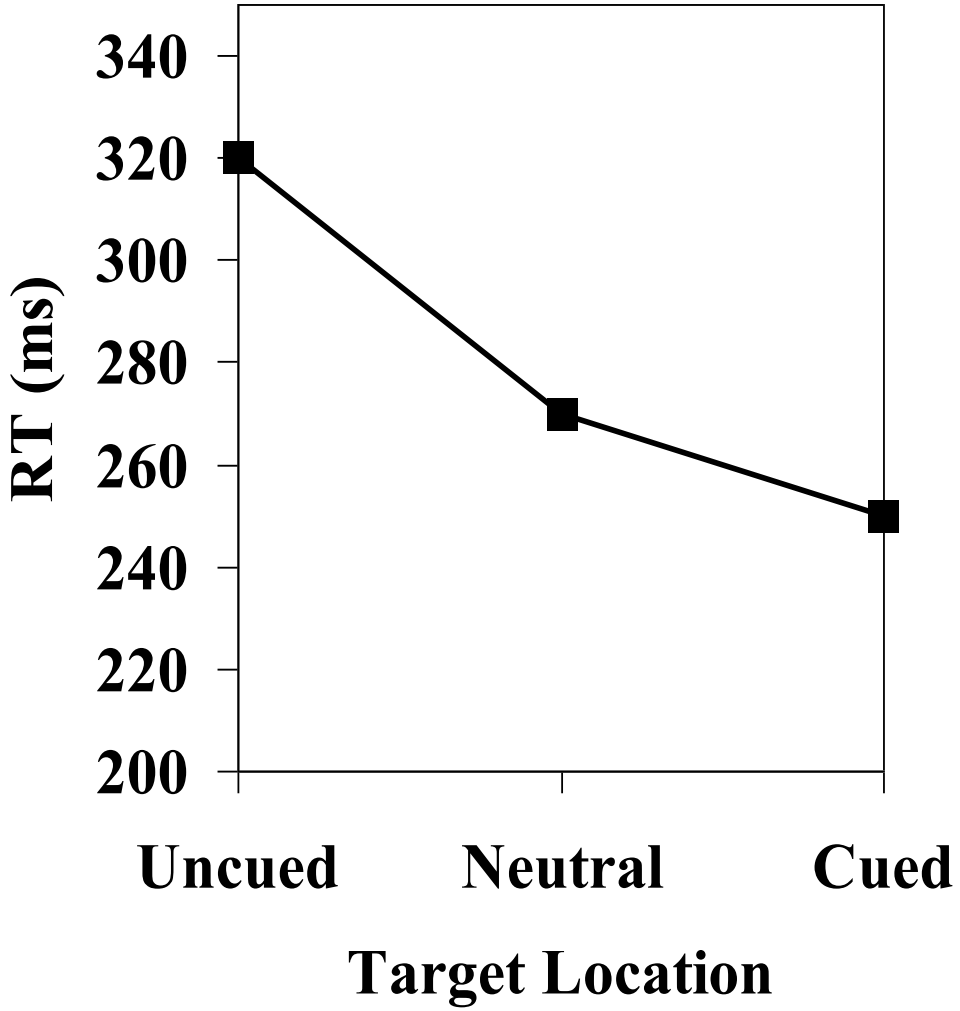


Symbolic cues may orient attention towards *another* location.
Stimulus cues orient attention to the *stimulated* location.

Institute for Systems and Robotics | LISBOA

Covert Attention

Reaction times



Covert Attention

What does Posner's experiments tell us:

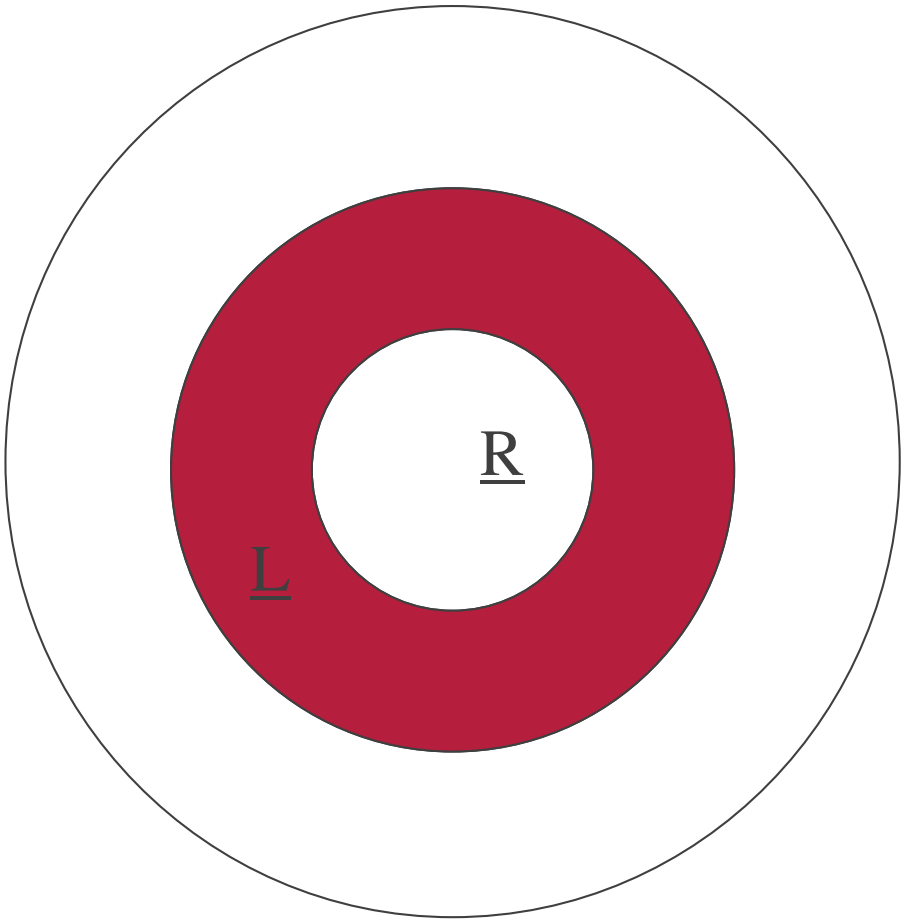
- We can move our "**spotlight of atention**" without moving our eyes.
- **Flicker** (motion) and symbolic cues attract attention automatically.
- We are faster to react when the stimulus is on the "spotlight". More "computational" **resources are allocated** to that location.

Spotlight Shapes

"Spotlights" can have different sizes, shapes, can be non compact, non connected…

The greater the extent over which attention is spread, the less efficient is the processing of information within that area.

The farther a stimulus is away from the center of an attended region, the less efficient is processing



Multiple Spotlights

- FINST – Fingers of INSTantiation: A limited number of objects (4-5) can be simultaneously indexed, independently of retinal location (Pylyshyn and Storm 88).

Canadian Journal of Experimental Psychology, 2000, 54(1), 1-14.

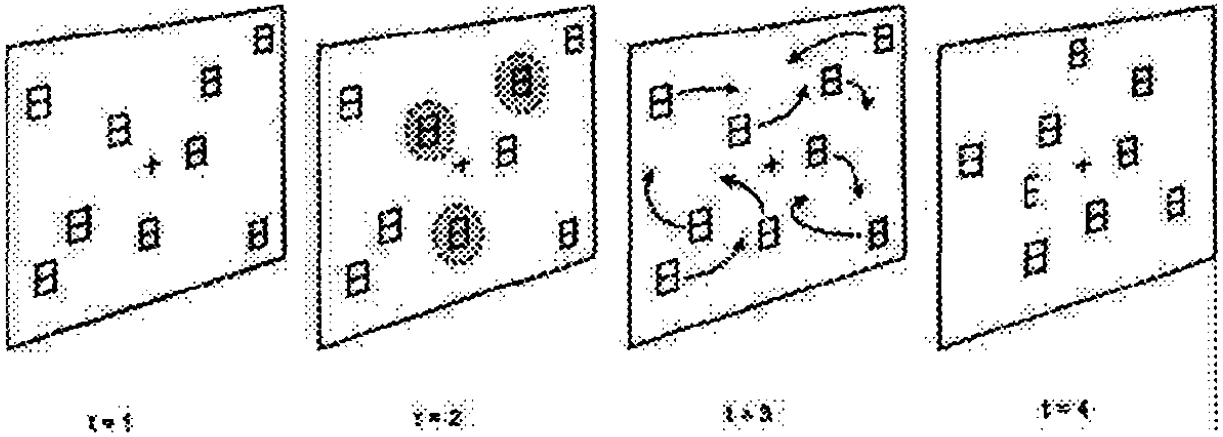


Figure 1: A schematic representation of a trial sequence. Participants view items on a video monitor. In the target designation phase (t=2) the target objects were flashed for 3 seconds (the selected targets are shown in circles in this illustration). The targets were then tracked for several seconds during the motion phase (t=3), and then a target or a distractor object underwent a form change by dropping two segments and ending up as an E or an H (t=4). The participants' task was to identify this form change as quickly and as accurately as possible. In this example, there are three target objects, and one of these target object undergoes a form change to an E shape.

Change Blindness

Change Blindness

Change Blindness

CB during Mudsplashes (O'Regan, Rensink & Clark, 1999)

Change Blindness

Without blank screen is easy to see

Change Blindness (using flicker)
(from J. Kevin O'Regan -- http://nivea.psycho.univ-paris5.fr)

Change Blindness

In the center is easy to see

Change Blindness (using flicker)
(from J. Kevin O'Regan -- http://nivea.psycho.univ-paris5.fr)

Change Blindness

What does **change blindness** tell us about the visual system ?

- Flicker of a single object is easy to spot but the blank screen or distractor flickers disturb the change detection system.
- With the blank screen or flicker distractors, we can just report changes on the objects that have been "attended".
- Objects that are more "salient" are "attended" faster than others.

Inattentional Blindness

Inattentional Blindness

In this video, two teams are playing basketball (the blacks and the whites). Look at the scene and count how many times the blacks pass the ball among themselves.

video

Inattentional Blindness

In this video, people have lerned to push doors not at the appropriate handle, but at the glass, because it is easier …

Inattentional Blindness

What does **inattentional blindness** tells us about the visual system ?

- Human perception allocates resources to the task at hand. Detecting the gorila would have been of no use for the assigned task.
- Human perception saves resources whenever possible. If the glass is usually in the door, why spending resources in trying to detect it.
- **A lot of learning is involved.**

Visual Search

Visual Search

Visual Search

Find the face

Visual Search

Unbounded Visual Search is NP-Complete

(John Tsotsos)

Complexity = $O(N2^{PM})$

N = number of prototypes

P = size of image in pixels

M = number of features computed at each pixel.

NP–completeness eliminates the possibility of developing a completely optimal and general algorithm.

Visual Search

Fortunatelly most objects in the world are not like this !

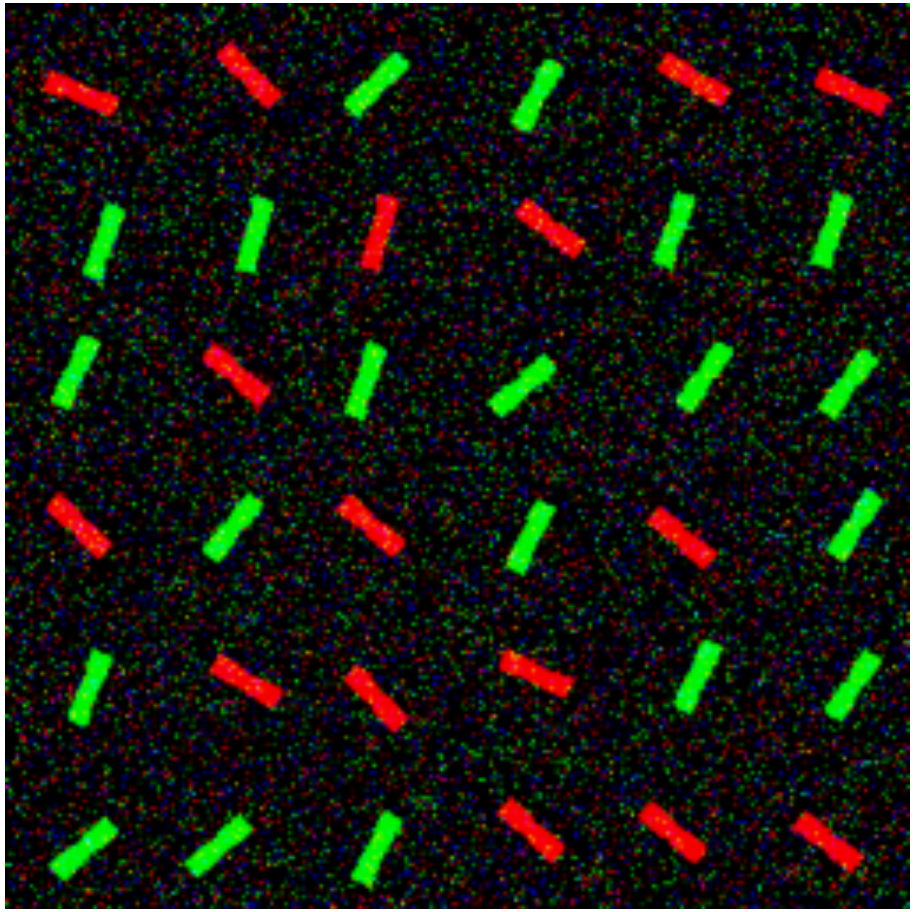
Bounded Visual Search

- Recognition with knowledge of a target and task in advance, and that knowledge is used to optimize the process.

Bounded Visual Search has time complexity **linear** in the number of test image pixel locations.

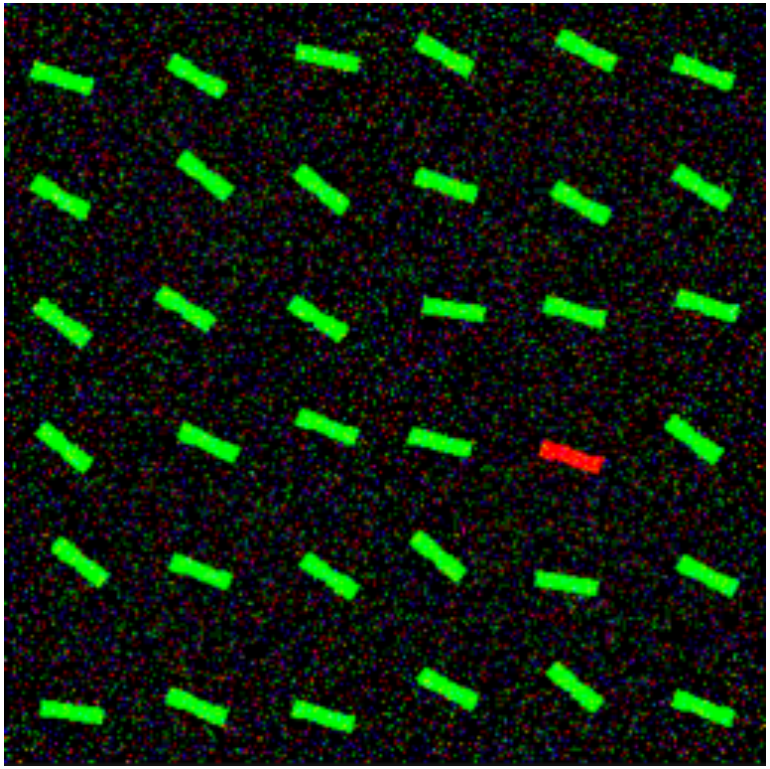
Visual Search

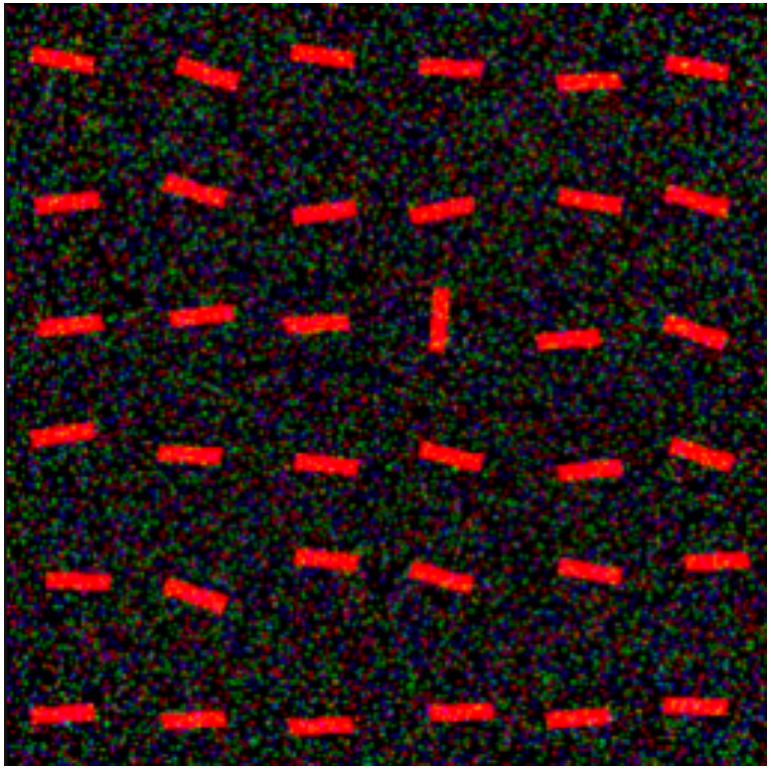
Find the distinct element



Visual Search

Find the distinct elements

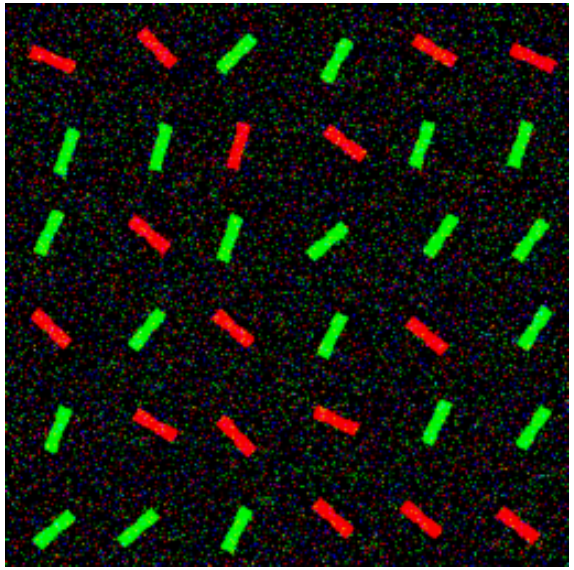




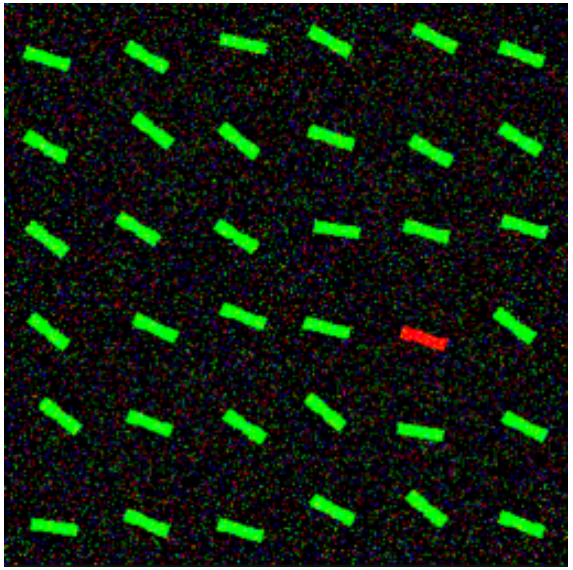
Visual Search

Why is the last example much easier than the others?

Conjunction Search

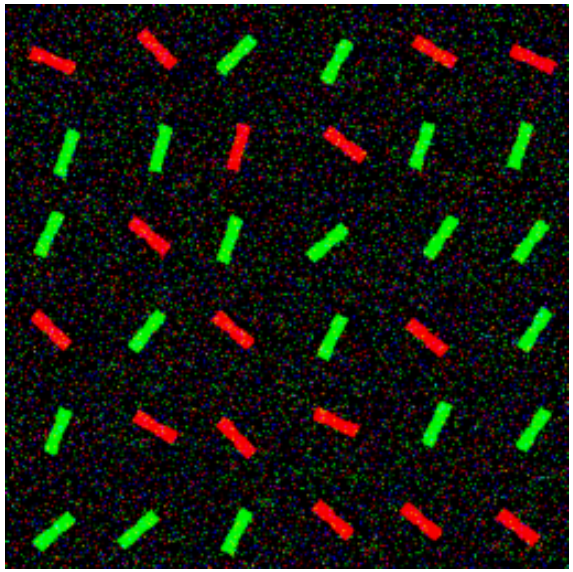


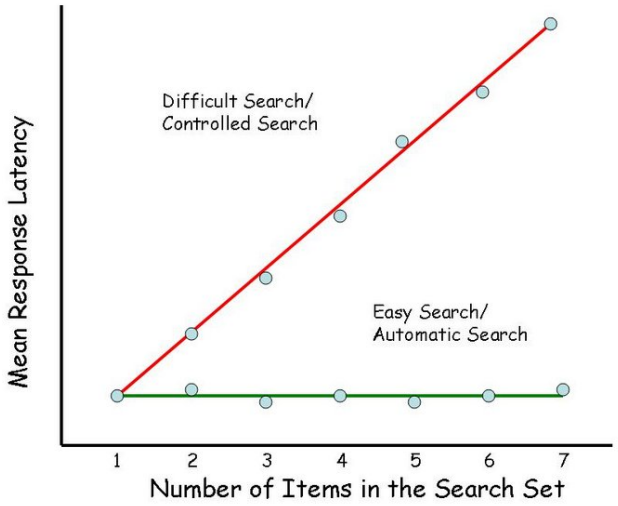
Pop-out



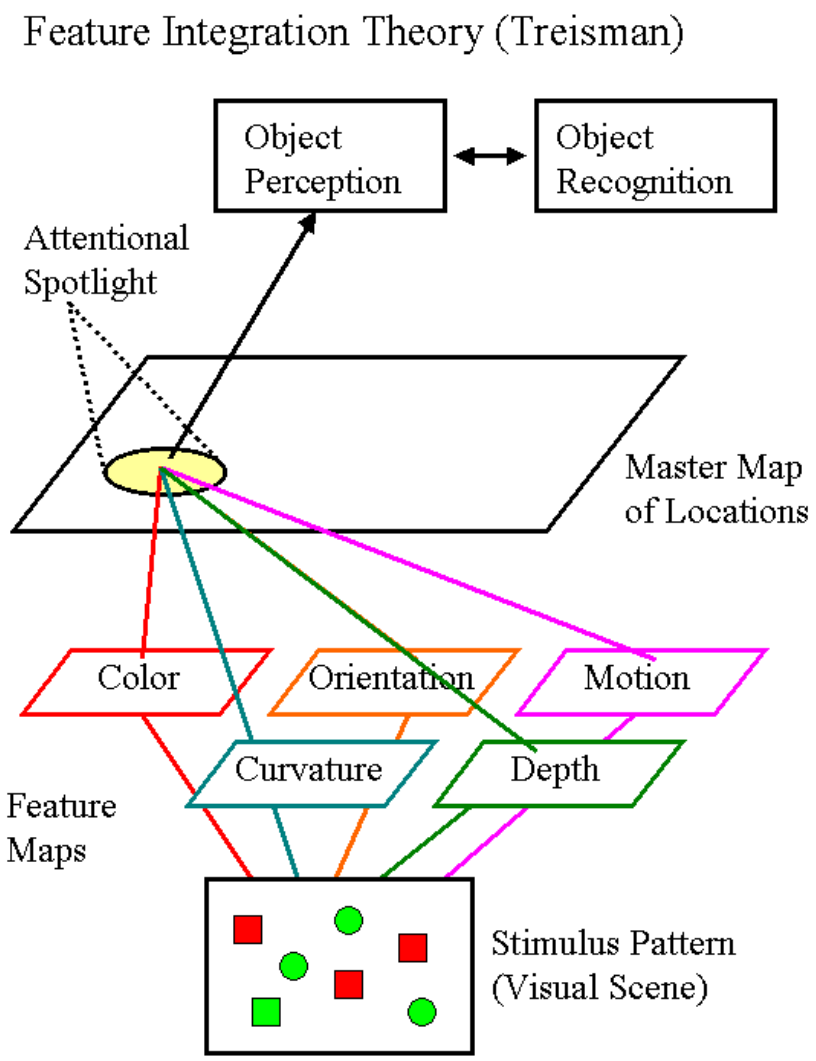
Visual Search

When a conjunction of features is required to detect the target, search times depend on the number of distracters.





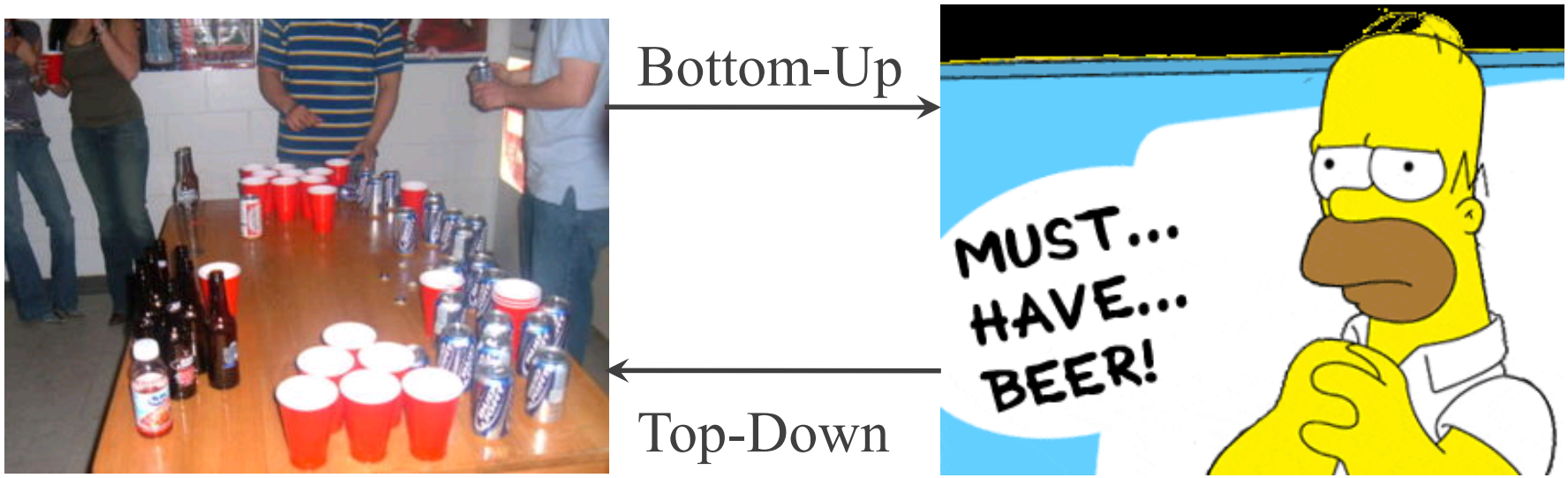
Visual Search



Pop-out items immediately attract the attentional spotlight.

Bottom-up and Top-down Attention

Attention can be focused volitionally by "top-down" signals derived from **task demands (context driven)**and automatically by "bottom-up" signals from **salient stimuli (data driven)**.



Note: This is not the whole story ...

Summary of Human Visual Attention

Bottom-up (data driven) and top-down (task based) cues attract attention to certain visual locations.

When a region is attended it can be recognized and stored in memory.

Locations that are more salient (data driven) or more important to the task atract more attentional resources.

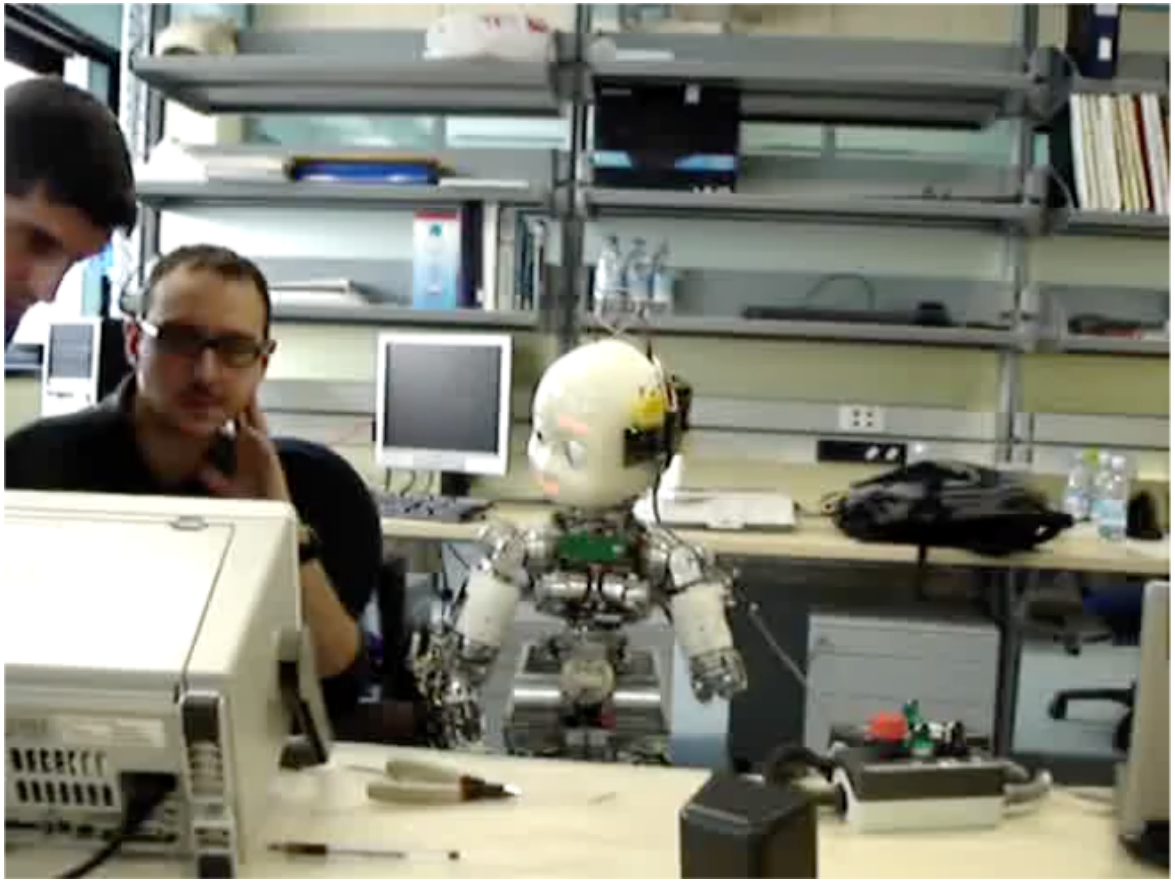
The brain is parsimonious in allocating attention. It may use previous knowledge to infer the state of the world instead of actively looking at it.

Using task knowledge is key in reducing complexity.

Bottom-Up Attention

Institute for Systems and Robotics | LISBOA

Bottom-Up Attention on the iCub



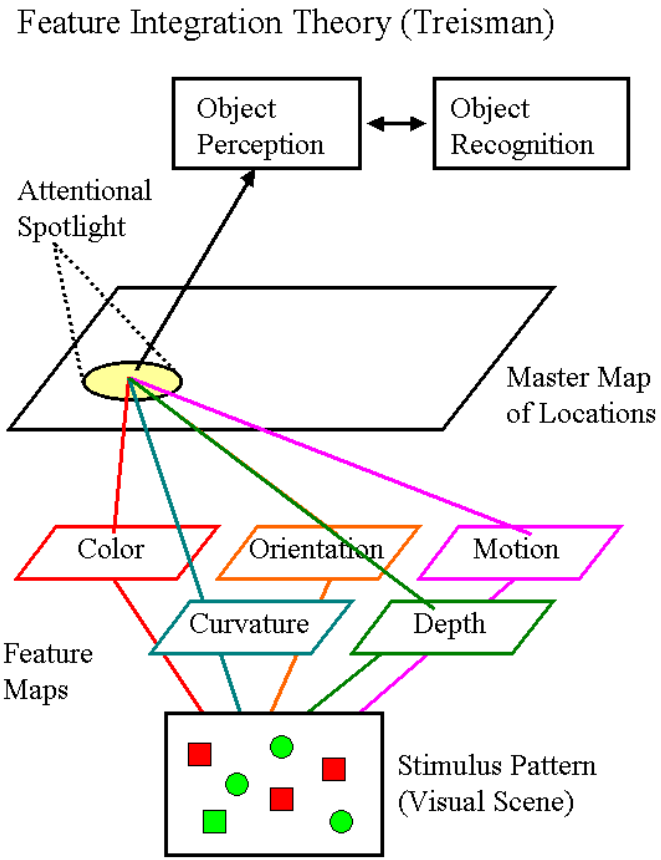
Multimodal Saliency-Based Bottom-Up Attention A Framework for the Humanoid Robot iCub, *J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, R. Pfeifer*, ICRA, 2008

Institute for Systems and Robotics | LISBOA

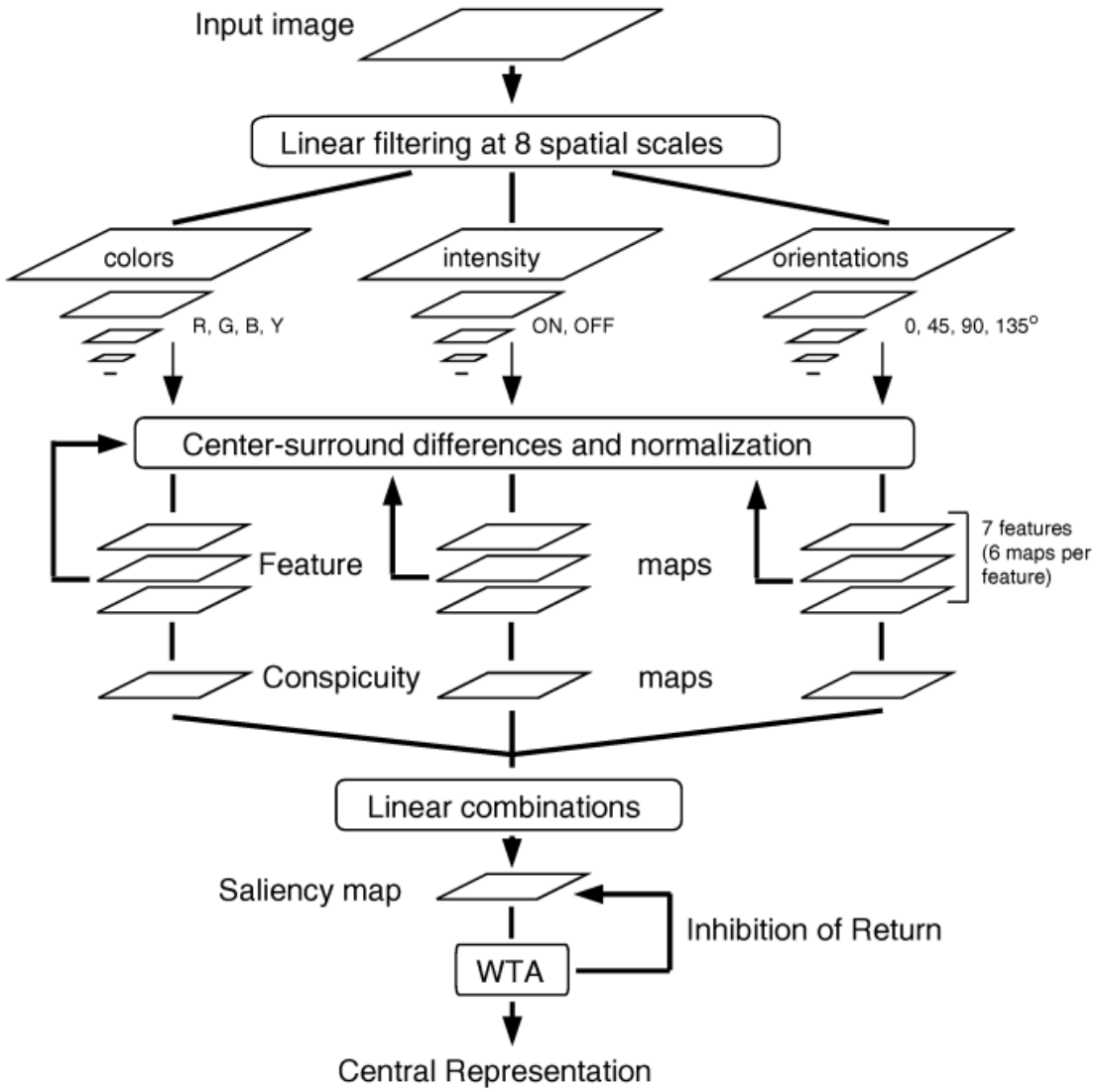
Bottom-up Attention on the iCub

We follow the most widely used model of attention based on salience, from Laurent Itti's work.

That work is related to the Feature Integration Theory of Treisman and Gelade.



Bottom-Up Attention

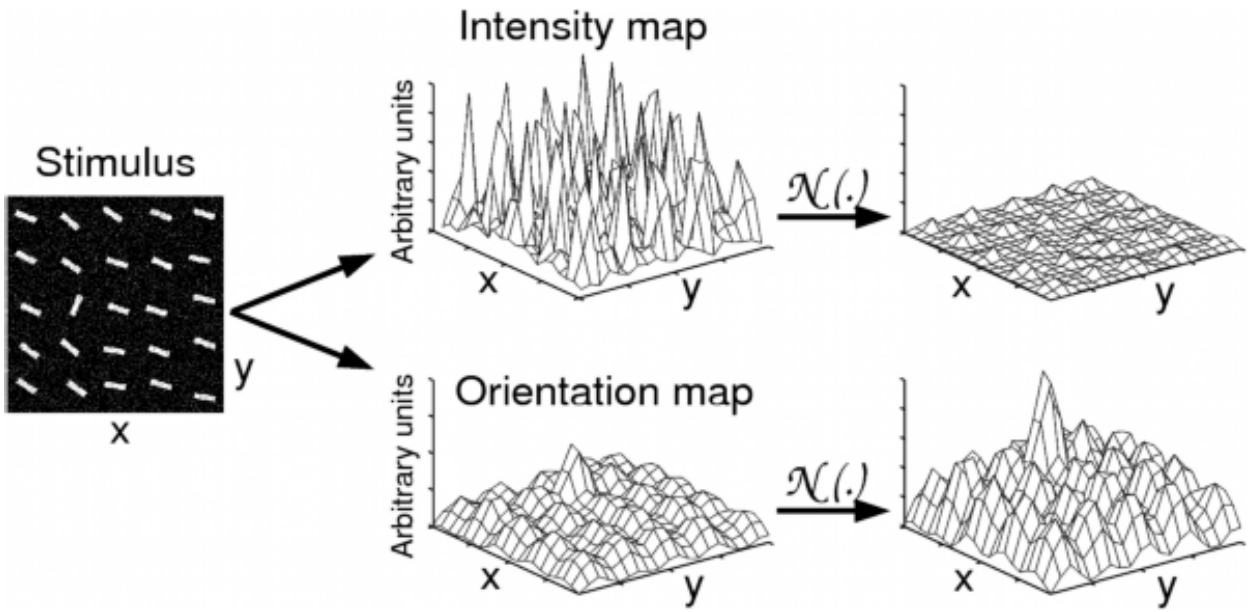


Itti et al, 98

Feature Maps

- Early stages of visual processing extract features from the image.

- Visual attention is attracted to image regions with features very distinct (salient) from the neighbors.

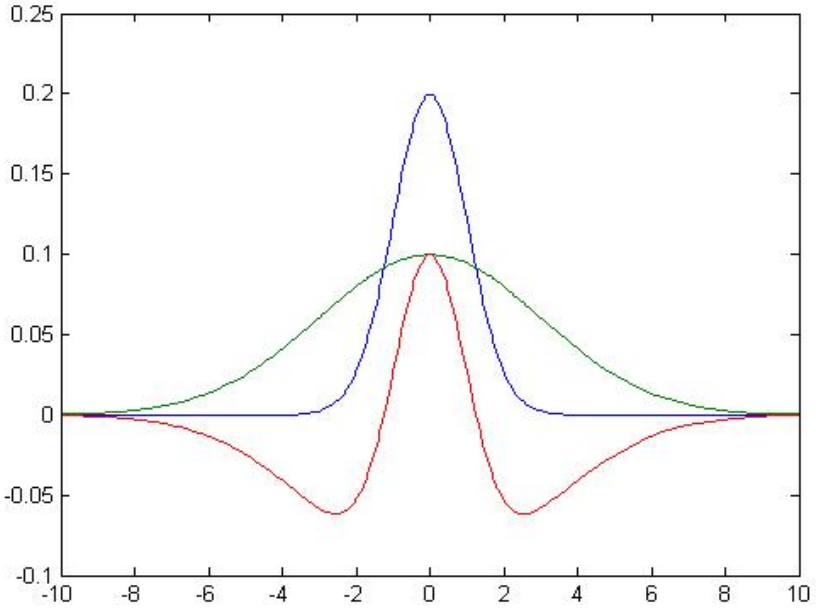


Institute for Systems and Robotics | LISBOA

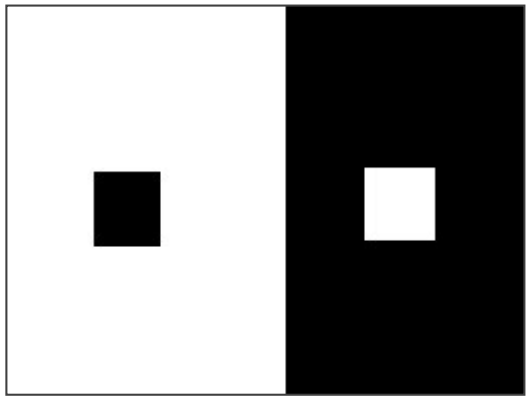
Saliency Maps

- Locations with high conspicuity: application of different scales of center-surround filters.

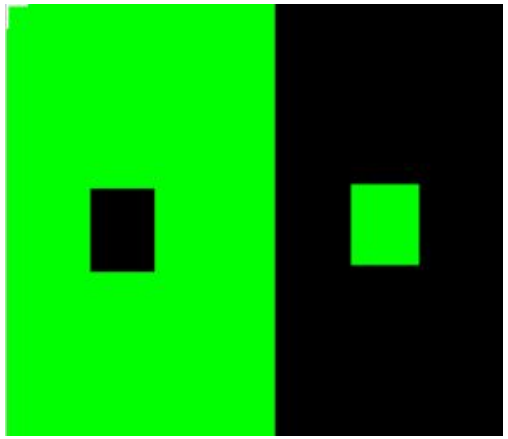
- Scales with clear peaks receive higher weights.



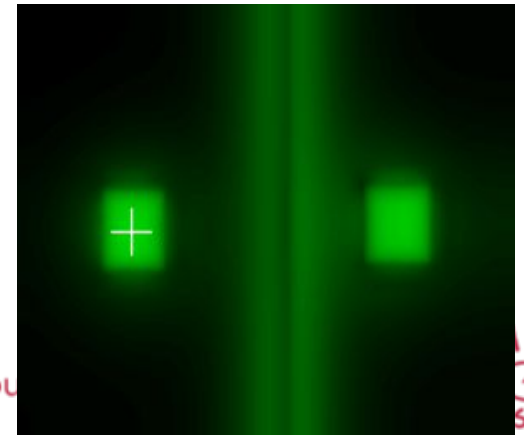
Input Image



Feature Map



Saliency Map

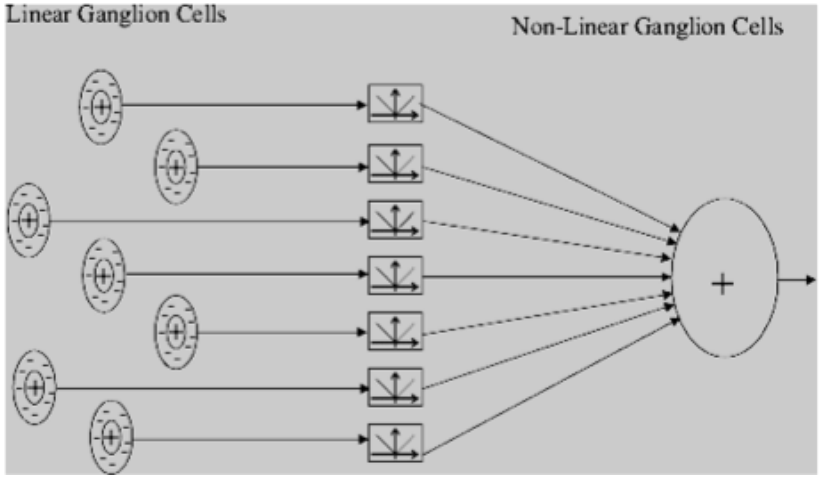


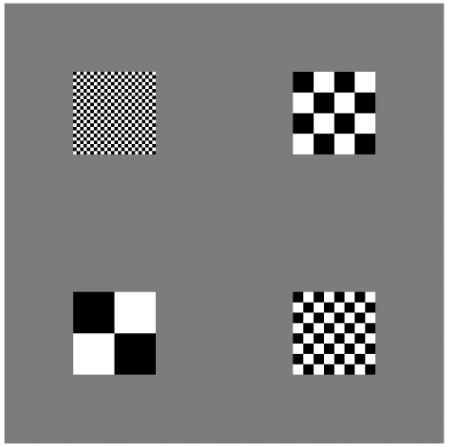
Institute for Systems and Robotics | LISBOA

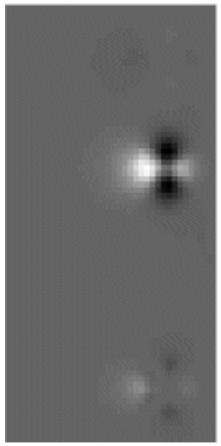
Texture Features

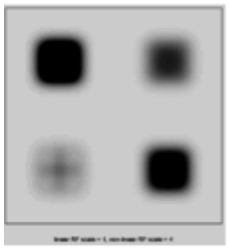
Spatial frequency features

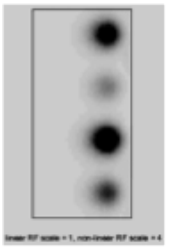
- Non-linear ganglion cells in the retina compute spatial frequency features [Demb et al 99].

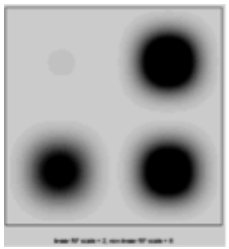


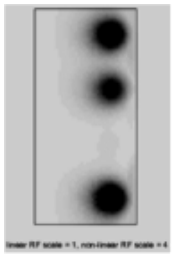


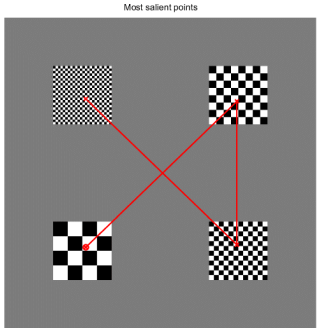


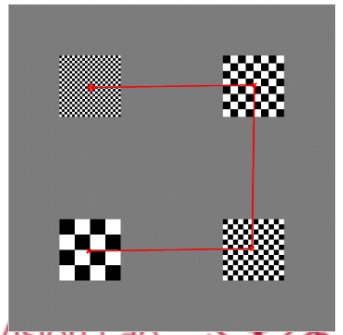








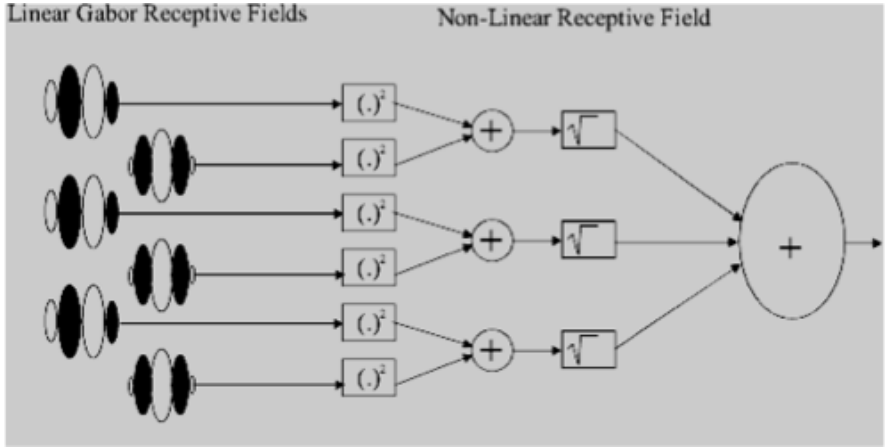




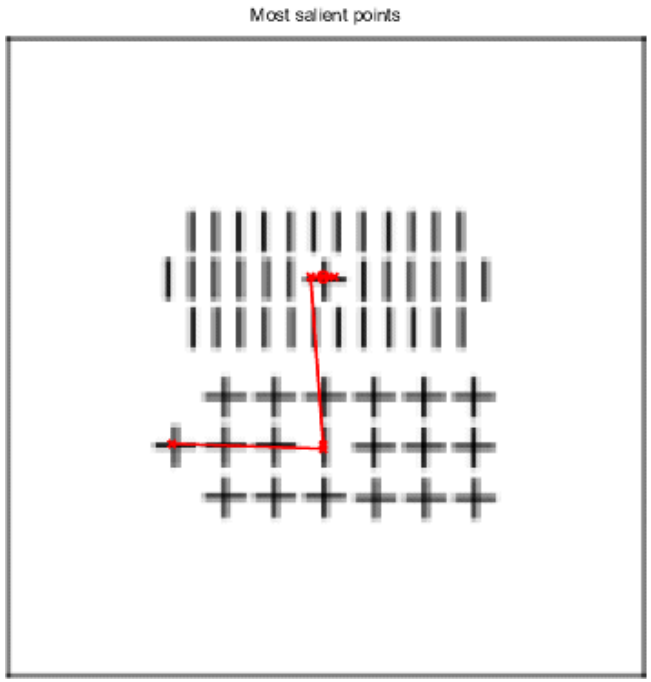
Institute for Systems and Robotics | LISBOA

Orientation Features

- Orientation
  - Visual Cortex Cells resemble Gabor Kernels and compute oriented features.



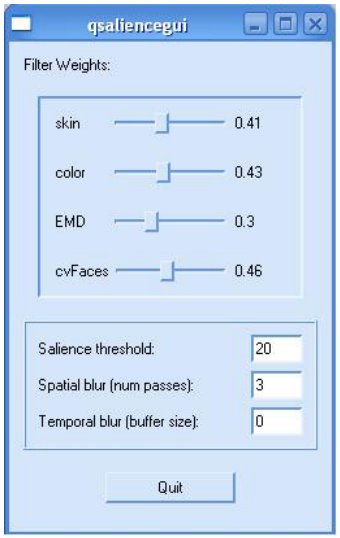
- We have developed:
  - a fast Gabor Decomposition
  - A new rule for orientation conspicuity

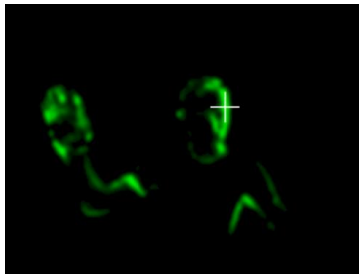


Institute for Systems and Robotics | LISBOA

Basic Stimulus Enhancement

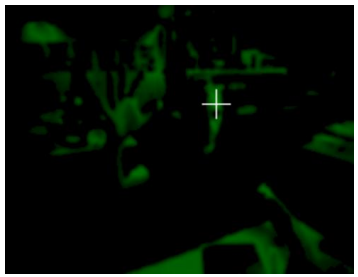
Top-Down Modulation Knobs





motion (waving)

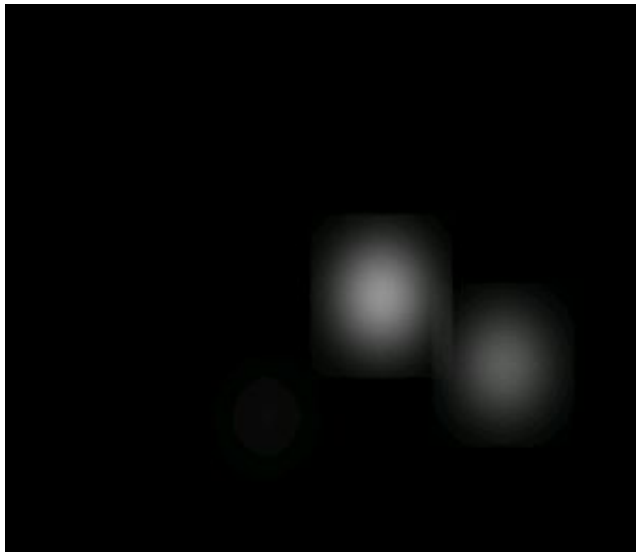
face salience



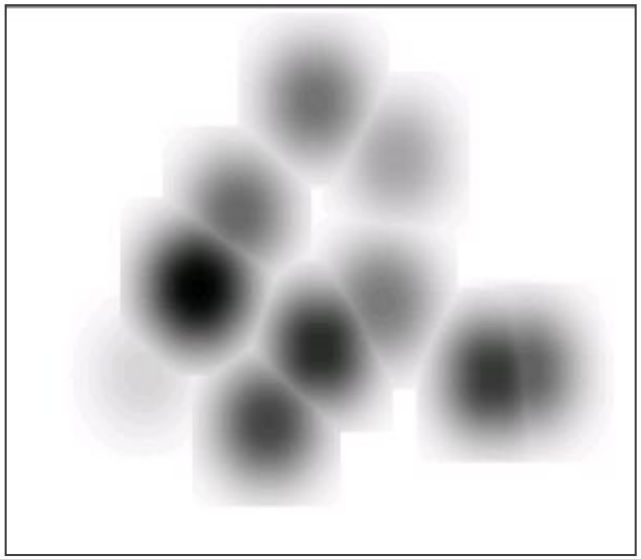
skin color

Temporal Integration

- An Habituation Map (left side) keeps track of attended locations
- New inhibition region (dark blob) added to IOR Map (right side) if habituation reaches a threshold.
- Values in both maps subject to decay.

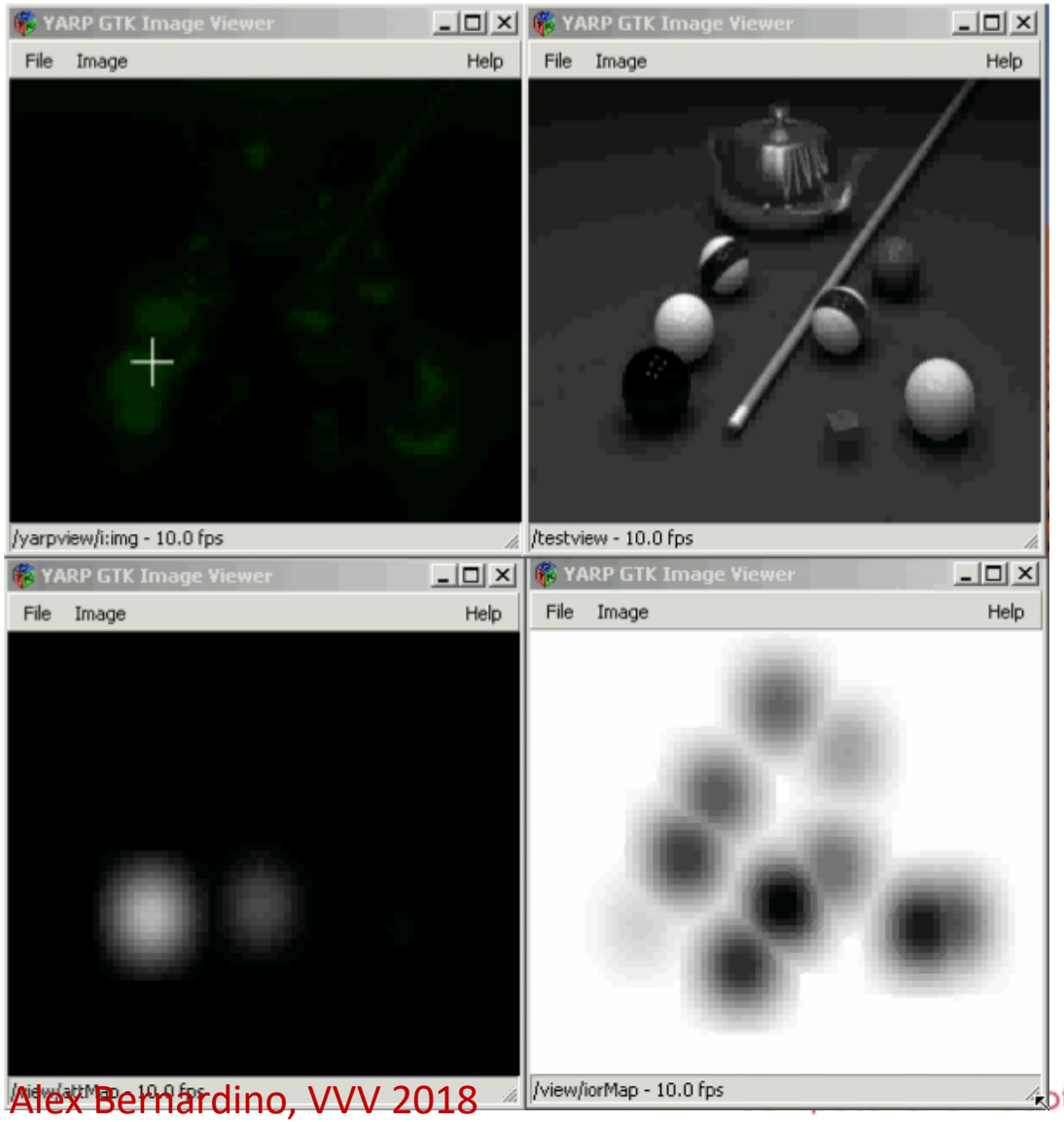


Habituation Map



Inhibition of Return Map
white=1, black=0

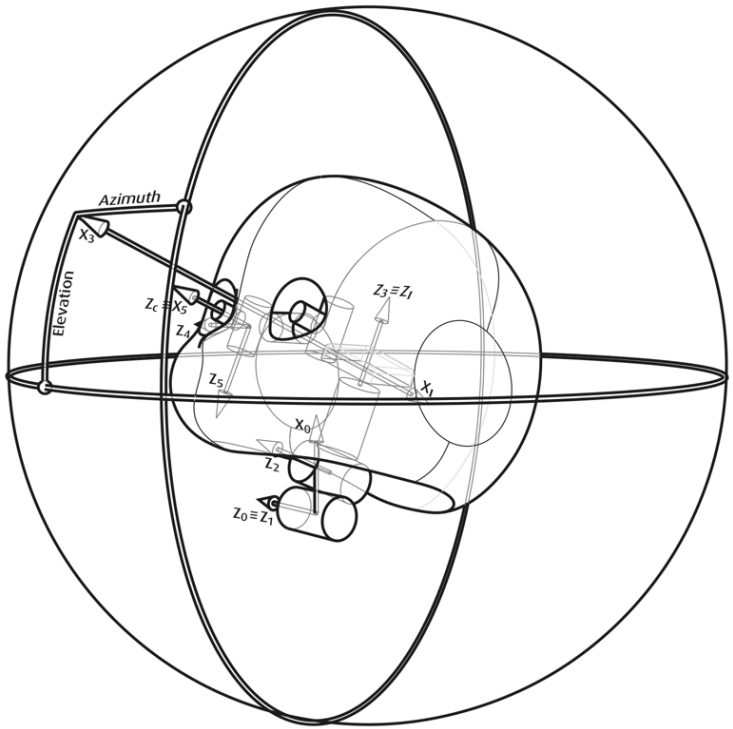
Saccade Dynamics

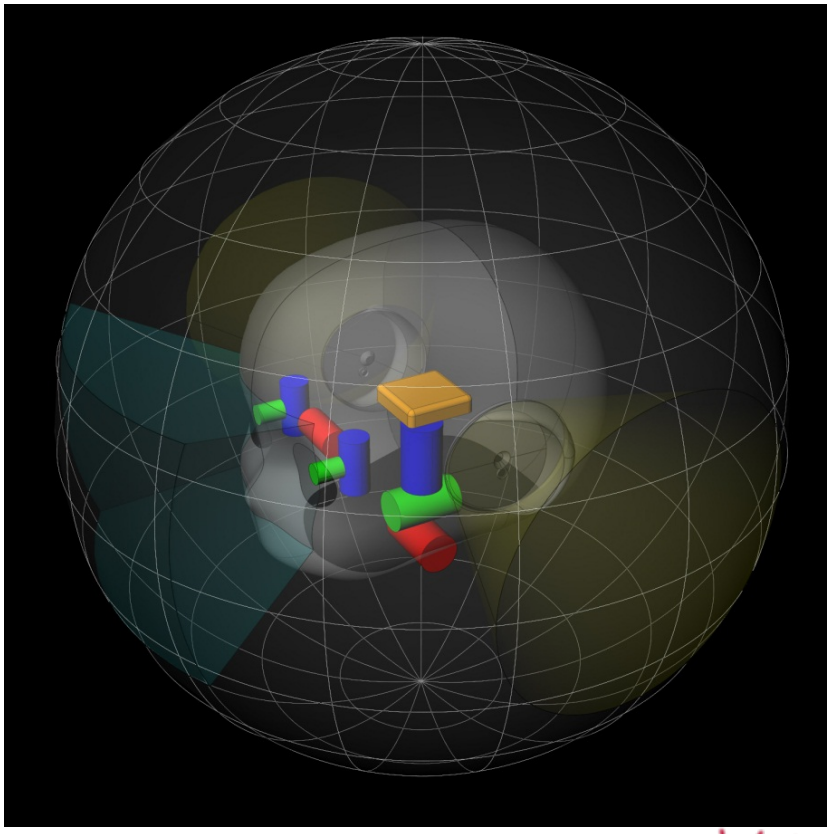


Institute for Systems and Robotics | LISBOA

Spatial Integration

The Ego-sphere: Aggregation of perceptions into one spherical saliency map in a neck based reference frame.

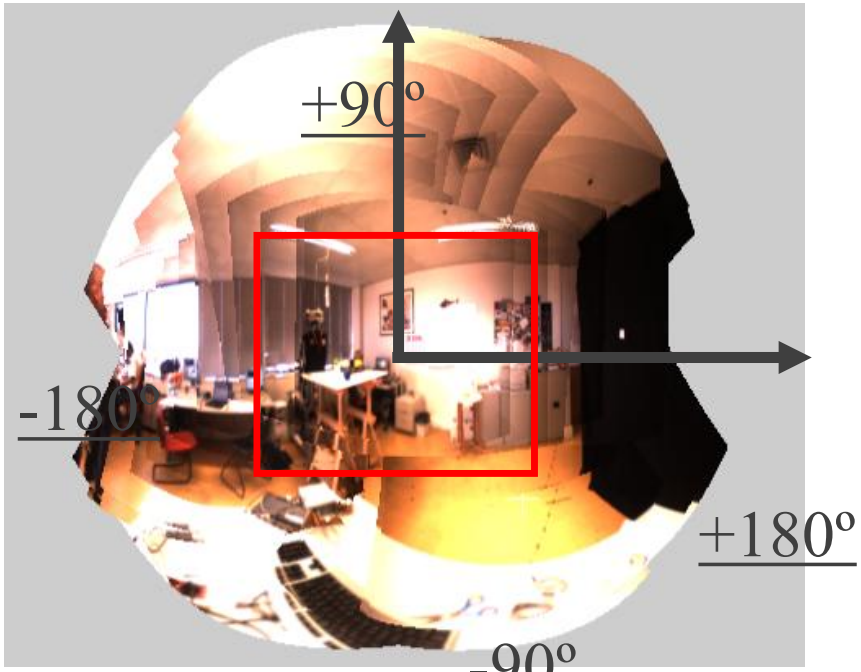




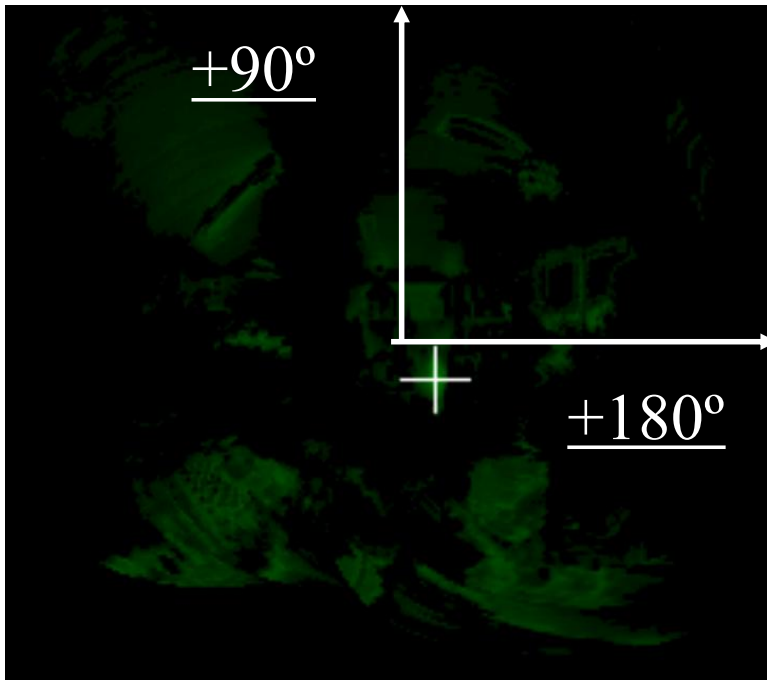
Institute for Systems and Robotics | LISBOA

Global Frame of Reference

Acquisition of coherent representation of its surroundings through head and eye movements



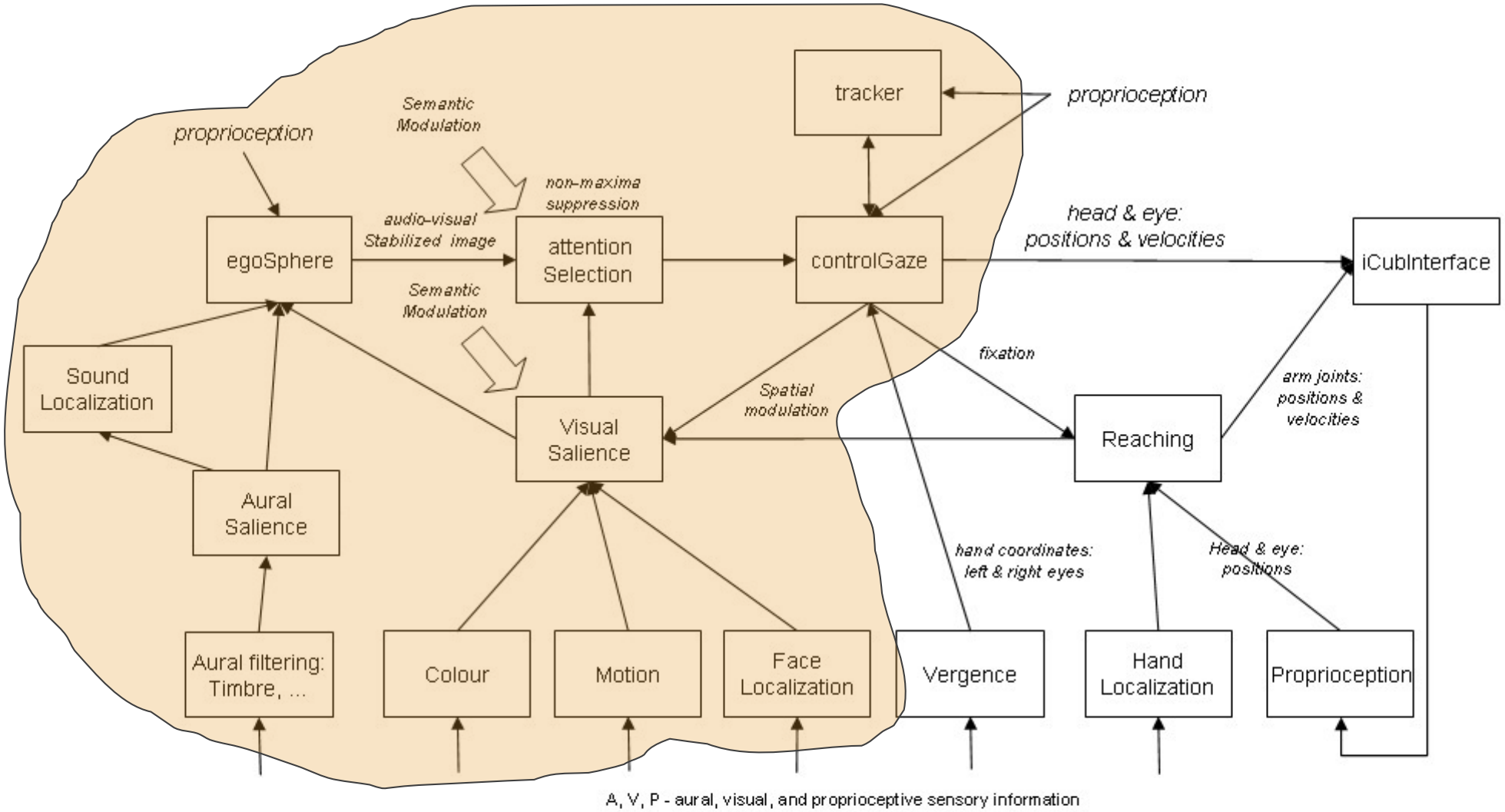
RGB Mosaic



Saliency Mosaic

Institute for Systems and Robotics | LISBOA

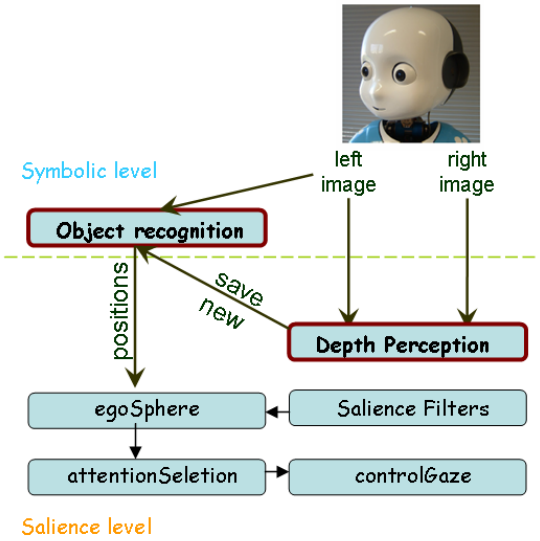
Software Architecture

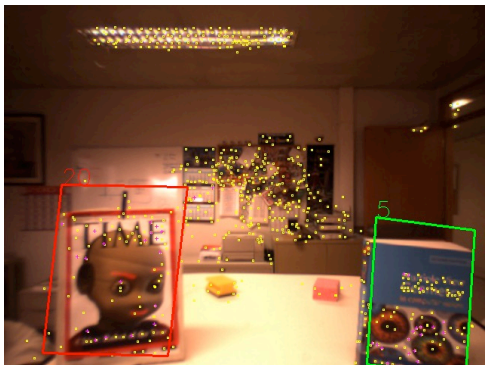


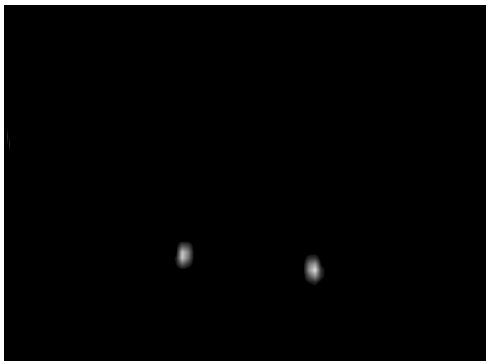
Institute for Systems and Robotics | LISBOA

Object Memory

- Representation and detection familiar objects.
- Mapping to the egosphere.
- Learning triggered by depth (proximity based) segmentation.







- **From Pixels to Objects: Enabling a spatial model for humanoid social robots.** Dario Figueira, *et al.* ICRA'09

Bottom-Up Attention

Institute for Systems and Robotics | LISBOA

Alex Bernardino, VVV 2018

Top-Down Attention

Institute for Systems and Robotics | LISBOA

Top-down attention as Visual Search

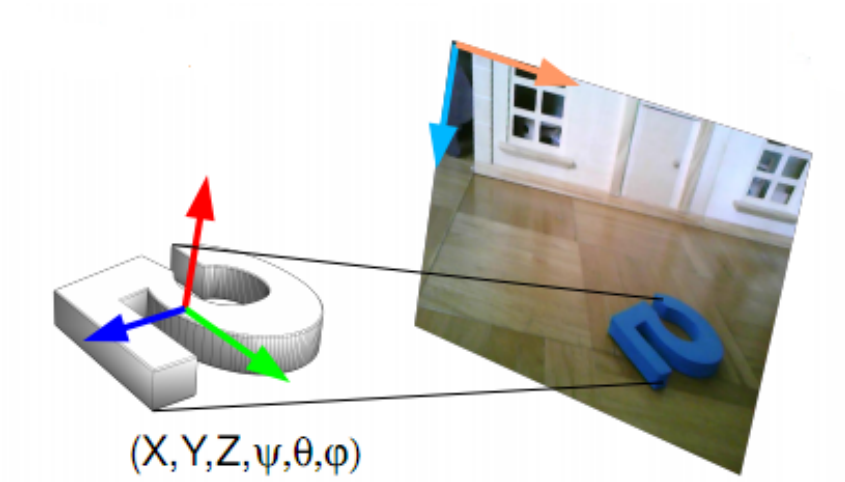
- **Task:** detect a known object in one image.
- **Straighforward Solution:** Exaustive Search

- Assume rigid object -> 6 D.O.F
  - 3D pos + 3D orientation

- Discretize the space of possible configurations in N slices at each dimension

  -> N^6 hypotheses

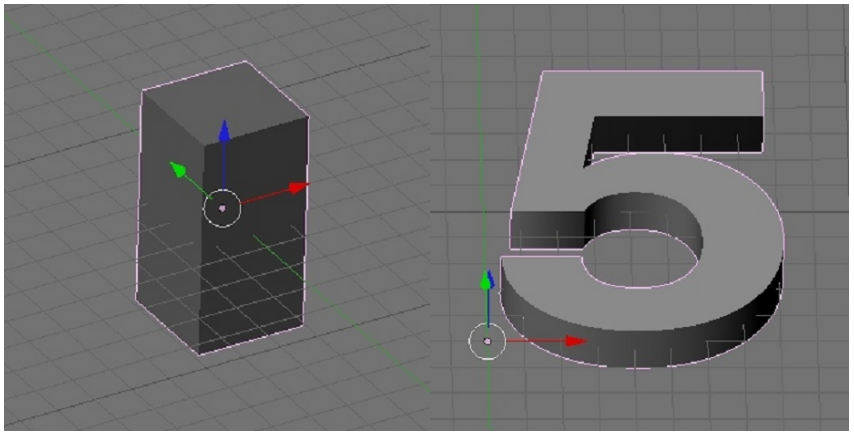
- Project the edges/texture of the object in the image and check if match.

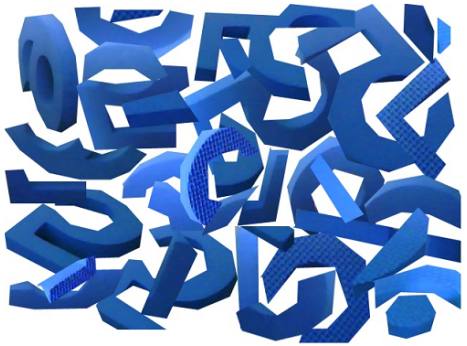


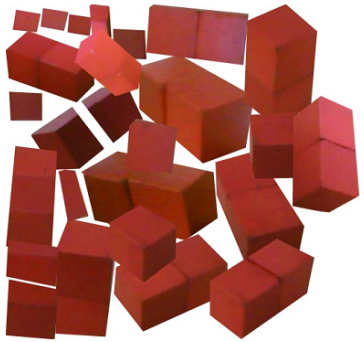
Institute for Systems and Robotics | LISBOA

Object Modeling

- Shape
  - 3D CAD Model
- Color
  - Color histogram
  - Can be learned from examples.



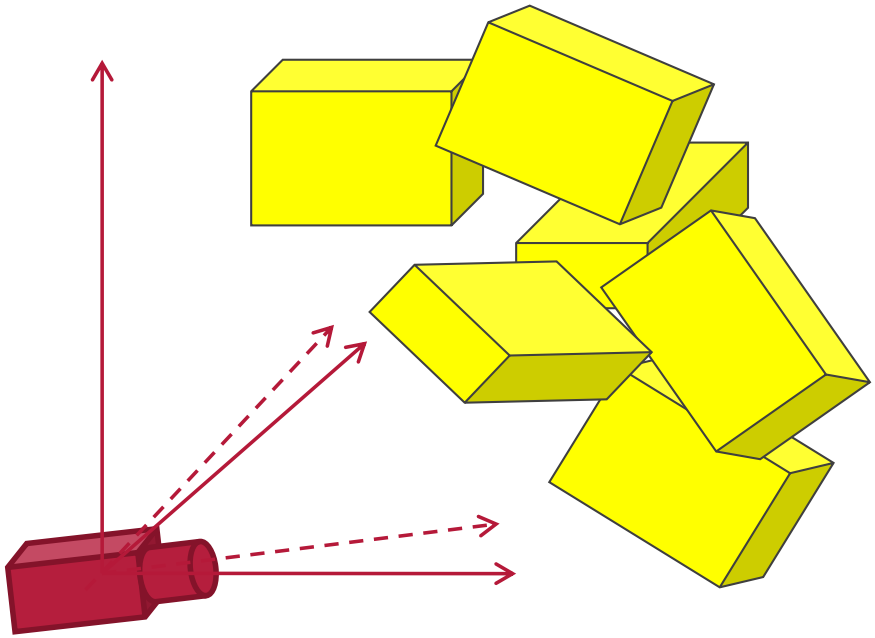


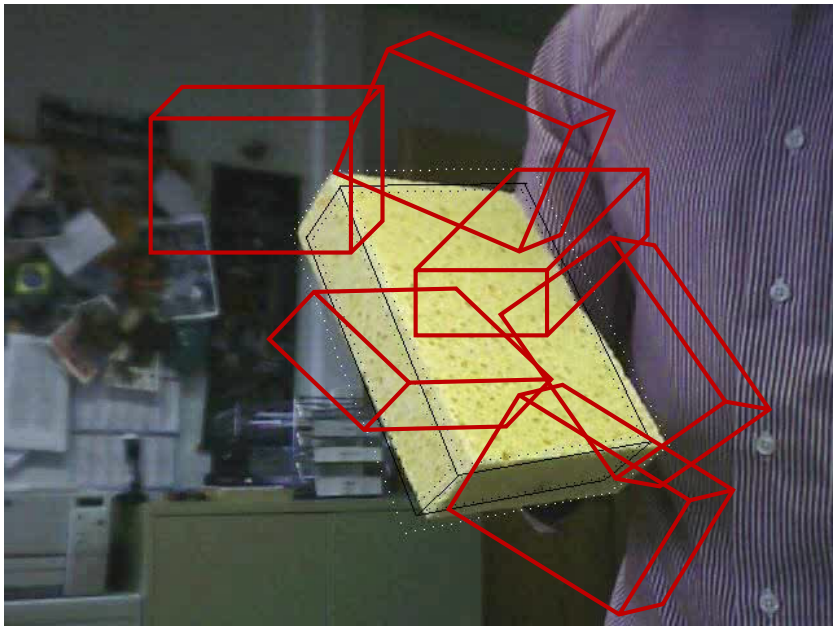


Institute for Systems and Robotics | LISBOA

Hypothesis Tests

$$X = [x, y, z, \theta_x, \theta_y, \theta_z, \ldots]$$





Maximize P(X|Image)

Complexity

- Divide each dimension in **10 slices**
- **10^6 hypotheses**

- Suppose at each frame we can only evaluate **10000** hypotheses
- need **100** frames (worst case) to detect the object.

- Suppose **frame period = 40ms**
- have to wait **4s** (worst case) to detect the object.

- **Can we reduce this?**

Search Space Reduction

- Use Feature Priors
  - Suppose we know the features of the object that best distingishes it from the surroundings. We can enhance these features (e.g. color) and inhibit others (e.g. orientation).

- Use Spatial priors:
  - Suppose we know the approximate location of the object (e.g. airplanes are usually up, my feet are usually down, when I see a computer, where is the mouse?). Hypothesis can be generated taking this into account.

- Use Temporal priors:
  - Suppose we what to track an object and approximatelly know its motion, so we can predict its future location. By the laws of inertia, hypothesis far from the prediction can be discarded.

Particle Filters

Fundamentals of Particle Filters

- Let $\mathbf{x}_k$ represent the state of the object at time k.
  - Position, orientation and derivatives.

- Let $\mathbf{z}_k$ represent the observations at time k.
  - Image, feature maps, …

- Want to compute
  - » $\mathbf{x}_k$

given the current and past observations

- » $\mathbf{z}_{1:k}=\{\mathbf{z}_i,\ i=1…k\}$

Fundamentals of Particle Filters

- Non-linear Bayesian Tracking
  - Compute $p(\mathbf{x}_k/\mathbf{z}_{1:k})$ in two stages:
  - Prediction

Motion model (prediction)

Posterior at k-1

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})\, p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})\, d\mathbf{x}_{k-1}$$

Prior at k

  - Update

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \;\alpha\; p(\mathbf{z}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$$

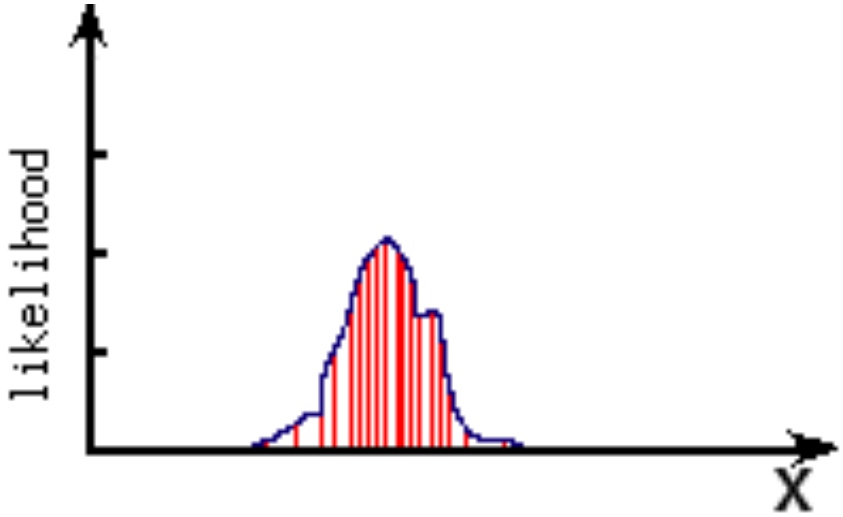
Posterior at k

Observation model (likelihood)

Fundamentals of Particle Filters

- Particle Filters
  - Approximate the posteriors by a set of random samples and associated weights: $\{\mathbf{x}_{0:k}^i, w_k^i\}_{i=1}^{N_s}$

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^i)$$



Note: We can control the used resources through N_s.

Fundamentals of Particle Filters

- Computing the weights

$$w_k^i \propto \frac{p(\mathbf{x}_{0:k}^i|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}^i|\mathbf{z}_{1:k})}$$

Trajectory evaluator

Trajectory generator

- In the sequential case

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i,\mathbf{z}_k)}$$

likelihood

prediction

Sample generator

Fundamentals of Particle Filters

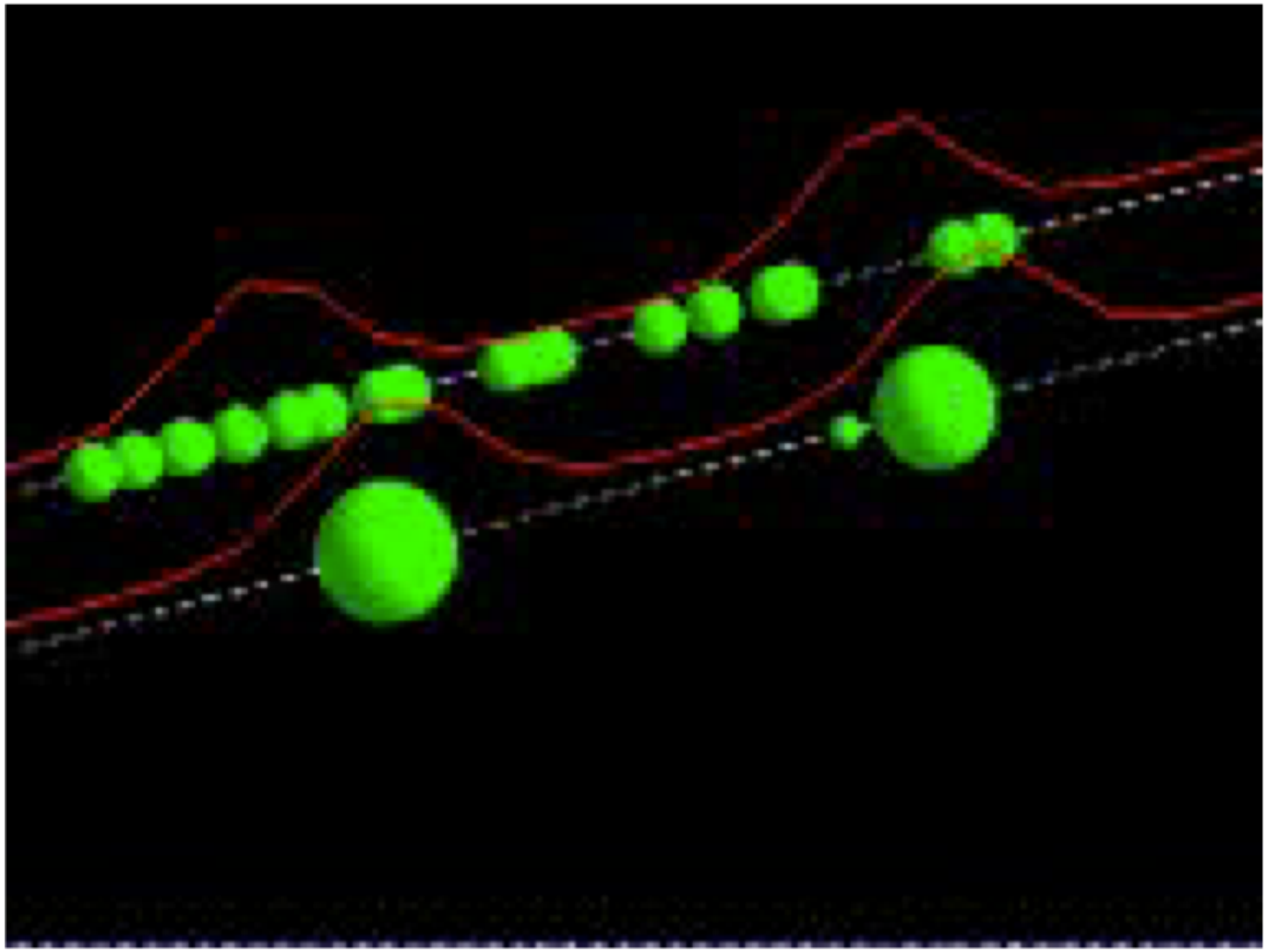
- Sequential Importance Sampling

ALGORITHM 1: SIS PARTICLE FILTER

$[\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}] = \text{SIS}\ [\{\mathbf{x}_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}, \mathbf{z}_k]$

- FOR $i = 1 : N_s$
  - Draw $\mathbf{x}_k^i \sim q(\mathbf{x}_k|\mathbf{x}_{k-1}^i, \mathbf{z}_k)$
  - Assign the particle a weight, w_k^i, according to $w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)}$
- END FOR

Fundamentals of Particle Filters



Fundamentals of Particle Filters

- The Sampling Importance Resampling (SIR) particle filter is less prone to degeneracy.

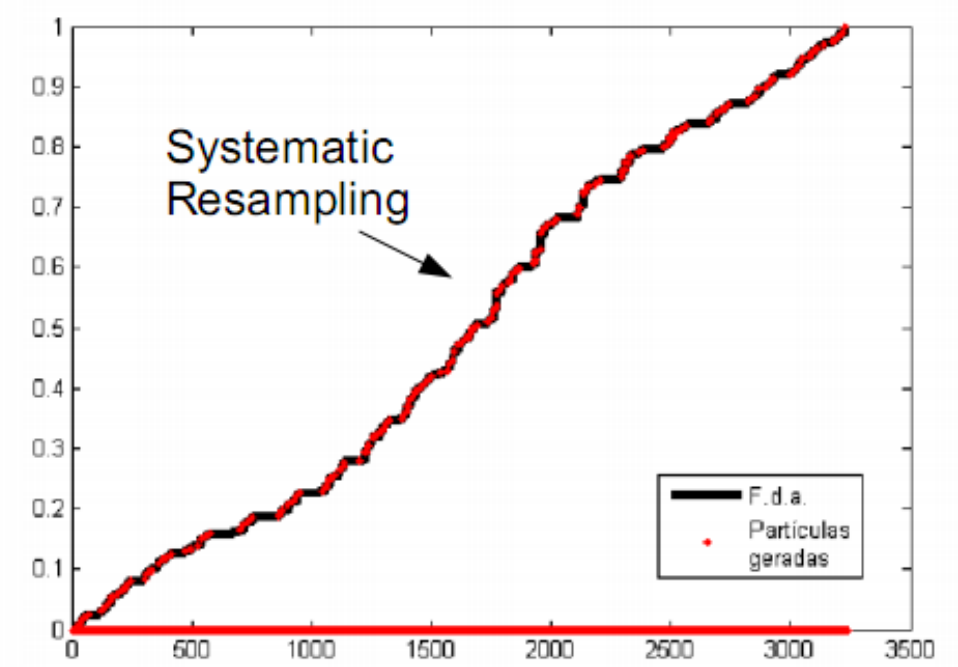
ALGORITHM 4: SIR PARTICLE FILTER

$[\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}] = \text{SIR}\ [\{\mathbf{x}_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}, \mathbf{z}_k]$

- FOR $i = 1 : N_s$
  - Draw $\mathbf{x}_k^i \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$
  - Calculate $w_k^i = p(\mathbf{z}_k|\mathbf{x}_k^i)$
- END FOR
- Calculate total weight: $t = \text{SUM}\ [\{w_k^i\}_{i=1}^{N_s}]$
- FOR $i = 1 : N_s$
  - Normalise: $w_k^i = t^{-1} w_k^i$
- END FOR
- Resample using algorithm 2:
  - $[\{\mathbf{x}_k^i, w_k^i, -\}_{i=1}^{N_s}] = \text{RESAMPLE}\ [\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}]$

Fundamentals of Particle Filters

- **Resamples from the distribution of the weights:**
  - Compute a 1D cumulative function of the weights (normalized between 0 and 1).
  - Generate N uniform random numbers.
  - Choose corresponding samples. High likelihood samples will be chosen more often.



ALGORITHM 2: RESAMPLING ALGORITHM

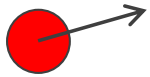
$[\{\mathbf{x}_k^{j*}, w_k^j, i^j\}_{j=1}^{N_s}] = \text{RESAMPLE}\ [\{\mathbf{x}_k^i, w_k^i\}_{i=1}^{N_s}]$

- Initialise the CDF: $c_1 = 0$
- FOR $i = 2 : N_s$
  - Construct CDF: $c_i = c_{i-1} + w_k^i$
- END FOR
- Start at the bottom of the CDF: $i = 1$
- Draw a starting point: $u_1 \sim \mathbb{U}\left[0, N_s^{-1}\right]$
- FOR $j = 1 : N_s$
  - Move along the CDF: $u_j = u_1 + N_s^{-1}(j-1)$
  - WHILE $u_j > c_i$
    - $i = i + 1$
  - END WHILE
  - Assign sample: $\mathbf{x}_k^{j*} = \mathbf{x}_k^i$
  - Assign weight: $w_k^j = N_s^{-1}$
  - Assign parent: $i^j = i$
- END FOR

Case Study – Ball tracking

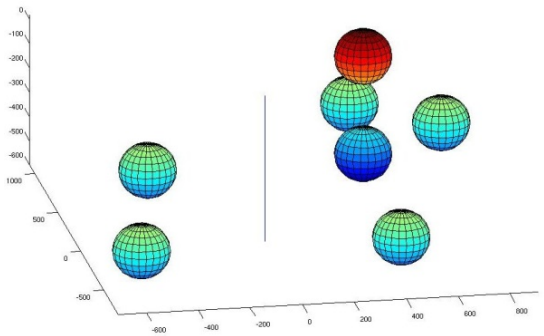
3D ball tracking

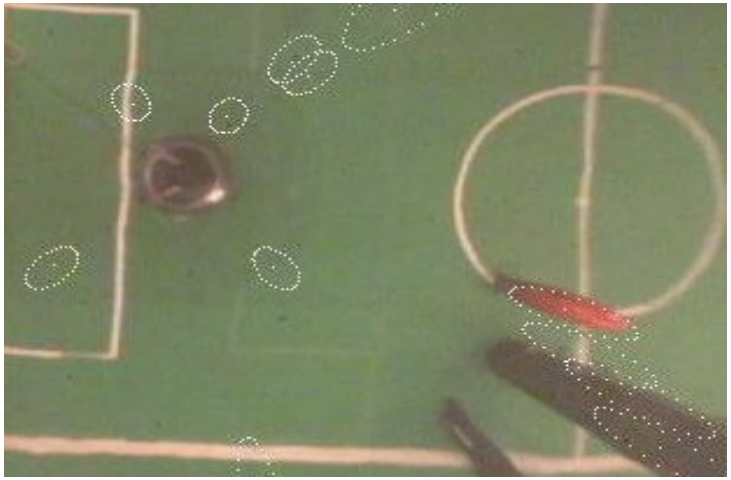
$$[x, y, z, \dot{x}, \dot{y}, \dot{z}]^T$$



Institute for Systems and Robotics | LISBOA

The Likelihood Function





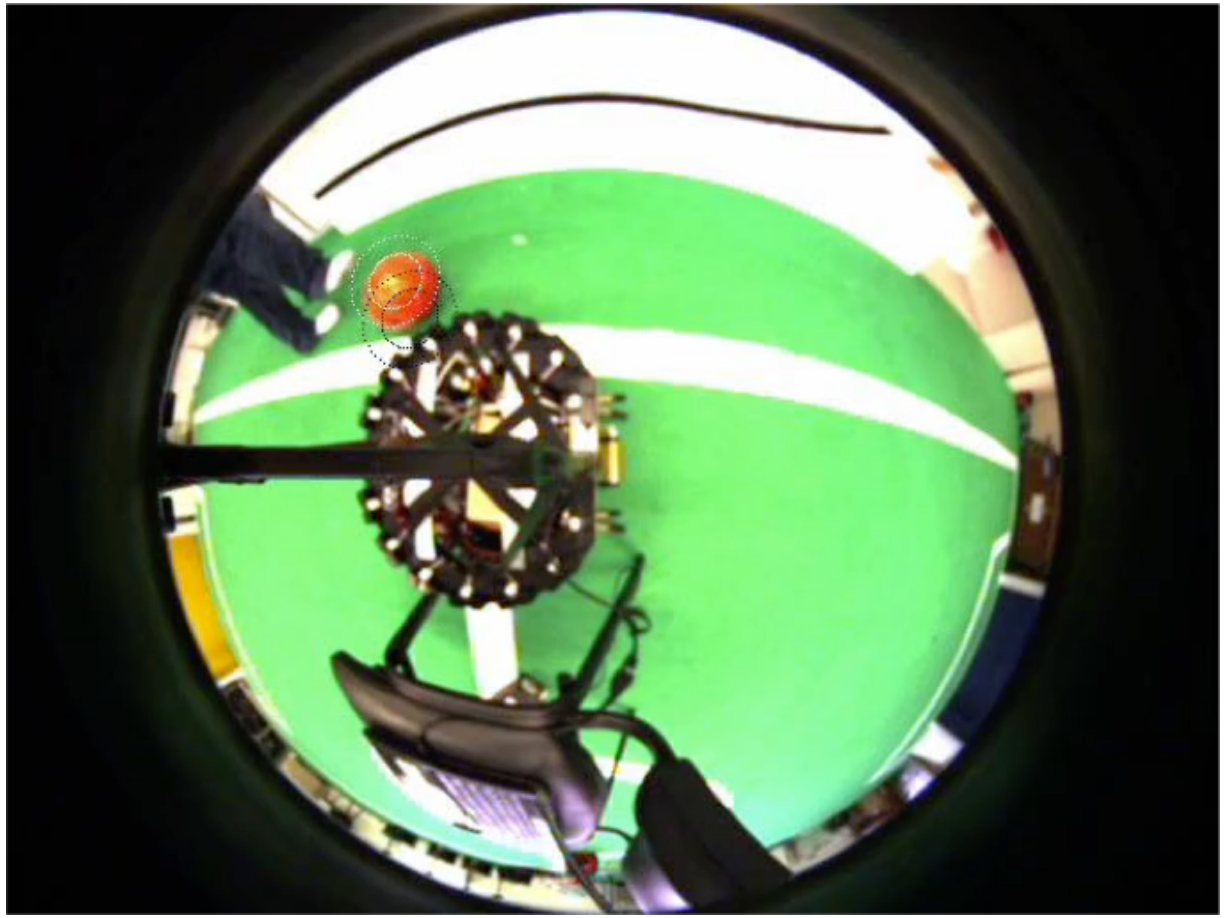
Likelihood P(X|Z) =

+ match(inner_hist, model_hist)

- match(inner_hist, outer_hist)

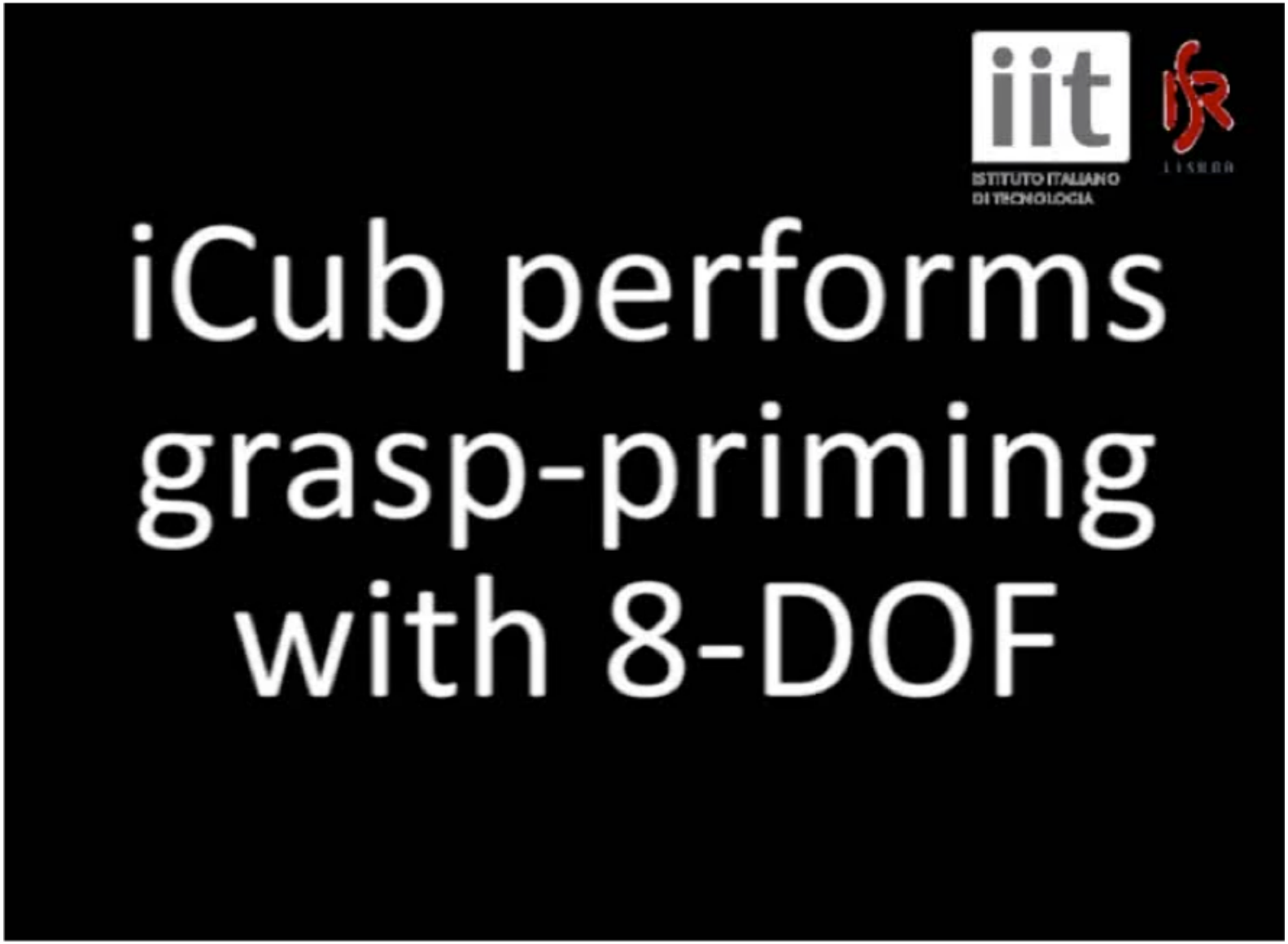
match = Bhattacharyya metric

Robust to (some) occlusion



Sample-Based 3D Tracking of Colored Objects: A Flexible Architecture,
M. Taiana, J. Nascimento, J, Gaspar and A. Bernardino, BMVC 2008

Top-Down Attention



A more complex case

Tracking the iCub Hand

The Problem(s):

1 – An adequate grasping system requires the precise estimate of the relative pose between hand and object.

2 – Object pose is often estimated in the camera reference frame but hand pose is computed via forward kinematics.

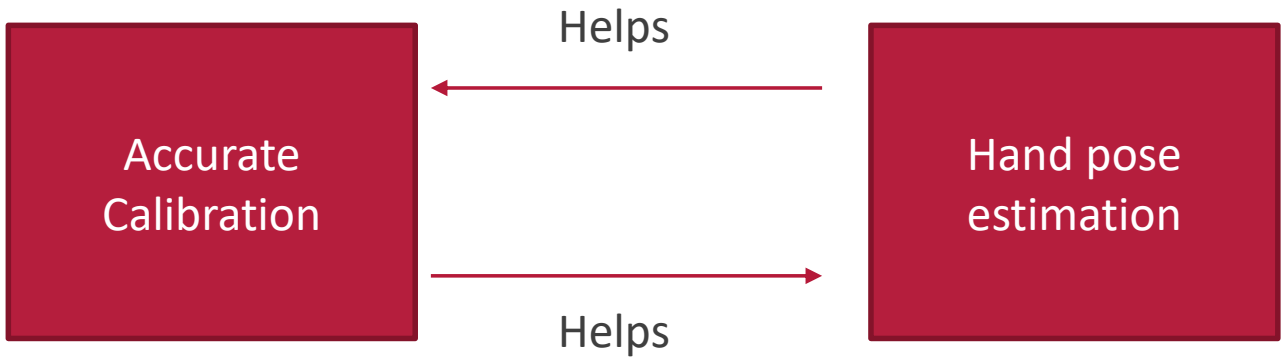
4 – There are errors in the kinematics chain between camera and hand: open-loop approaches will likely fail.

5 – Closed-loop control is rarely used in Humanoid Robots because it is difficult to visually estimate the pose of the hand.

Institute for Systems and Robotics | LISBOA

Simultaneous Calibration and Hand-Pose Estimation

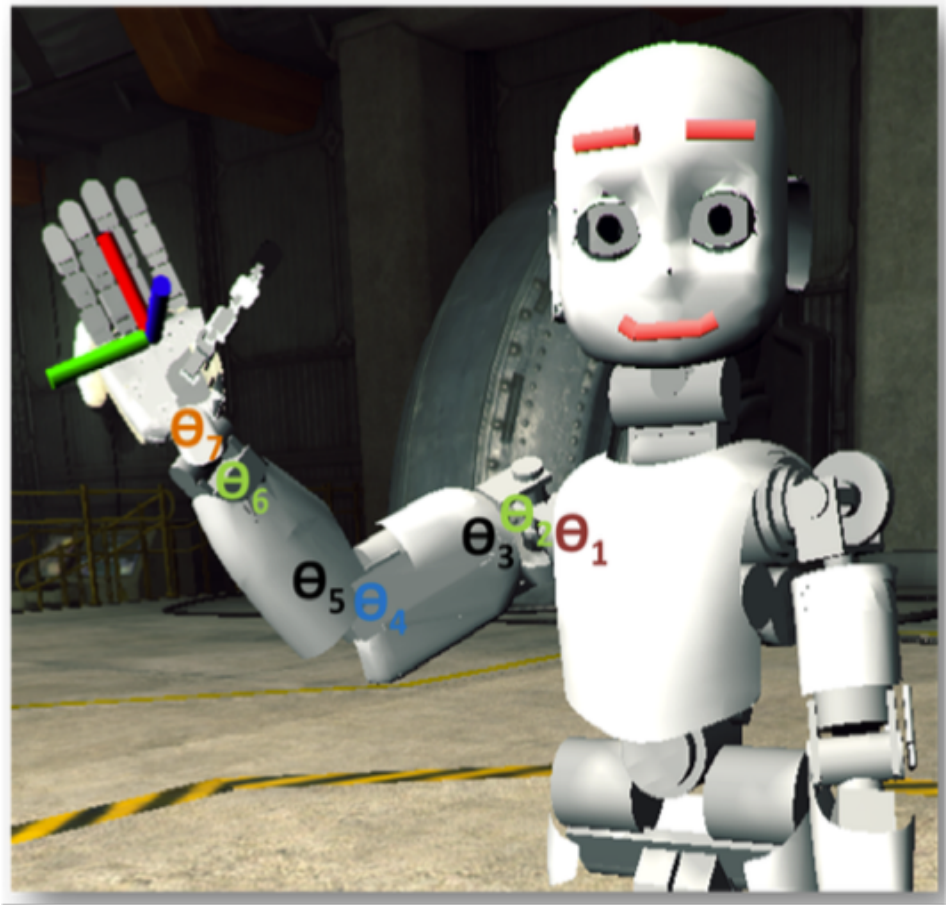
Towards markerless visual servoing of grasping tasks for humanoid robots, P. Vicente, L. Jamone and A. Bernardino, ICRA 2017.

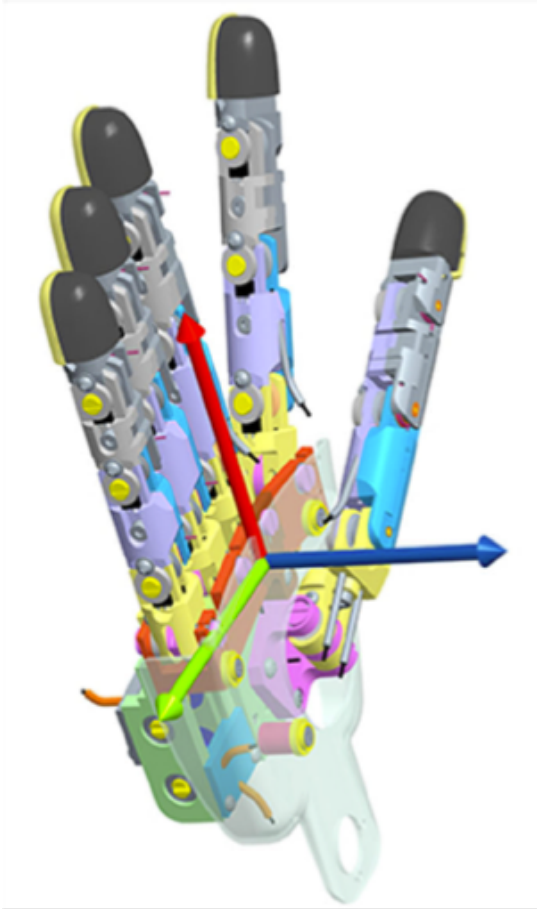


Institute for Systems and Robotics | LISBOA

Internal Model

Internal robot computer graphics model (geometry and appearance). Special care taken in modeling the hand. Implemented in Unity.

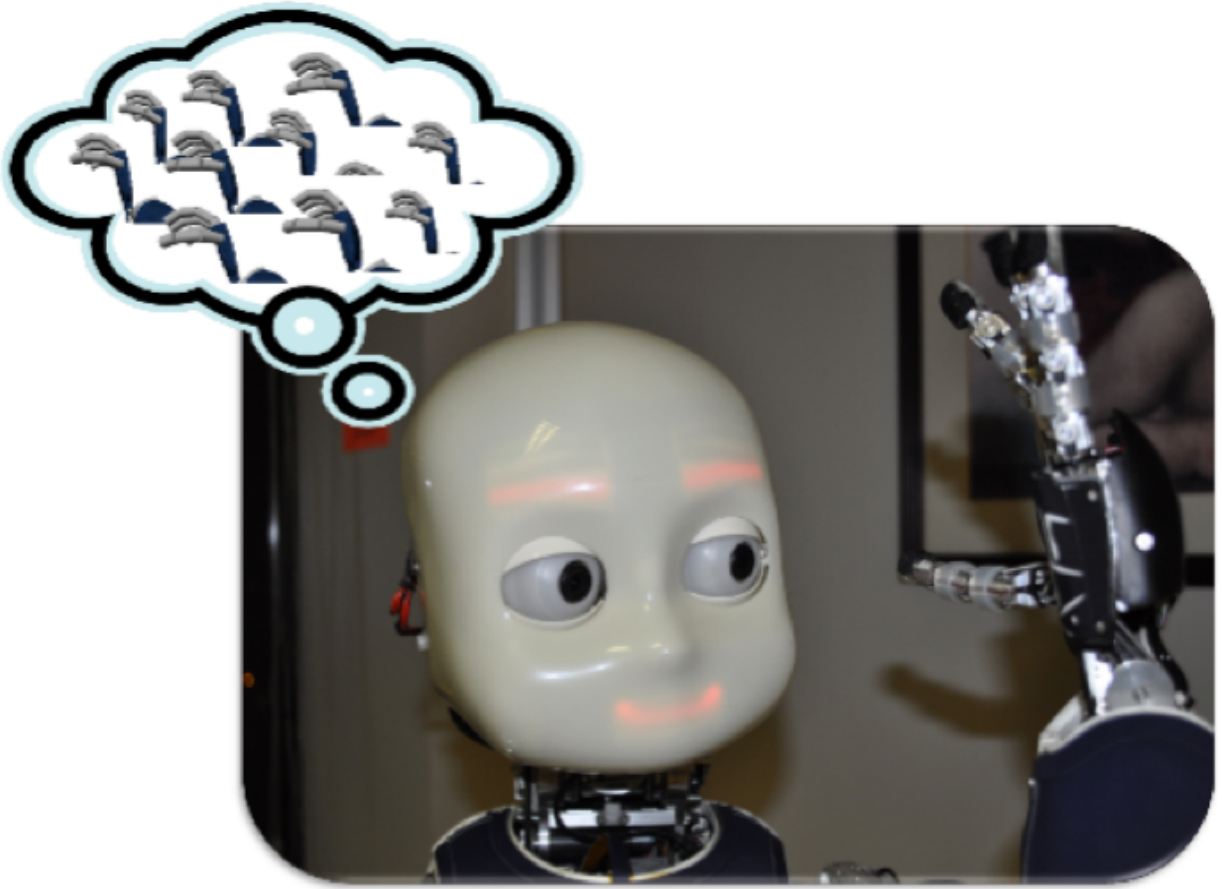




Institute for Systems and Robotics | LISBOA

Hypothesis Generation

An internal mental simulation of likely hand poses based on the current joint angles (random perturbations to nominal). Simulated images rendered in Unity.

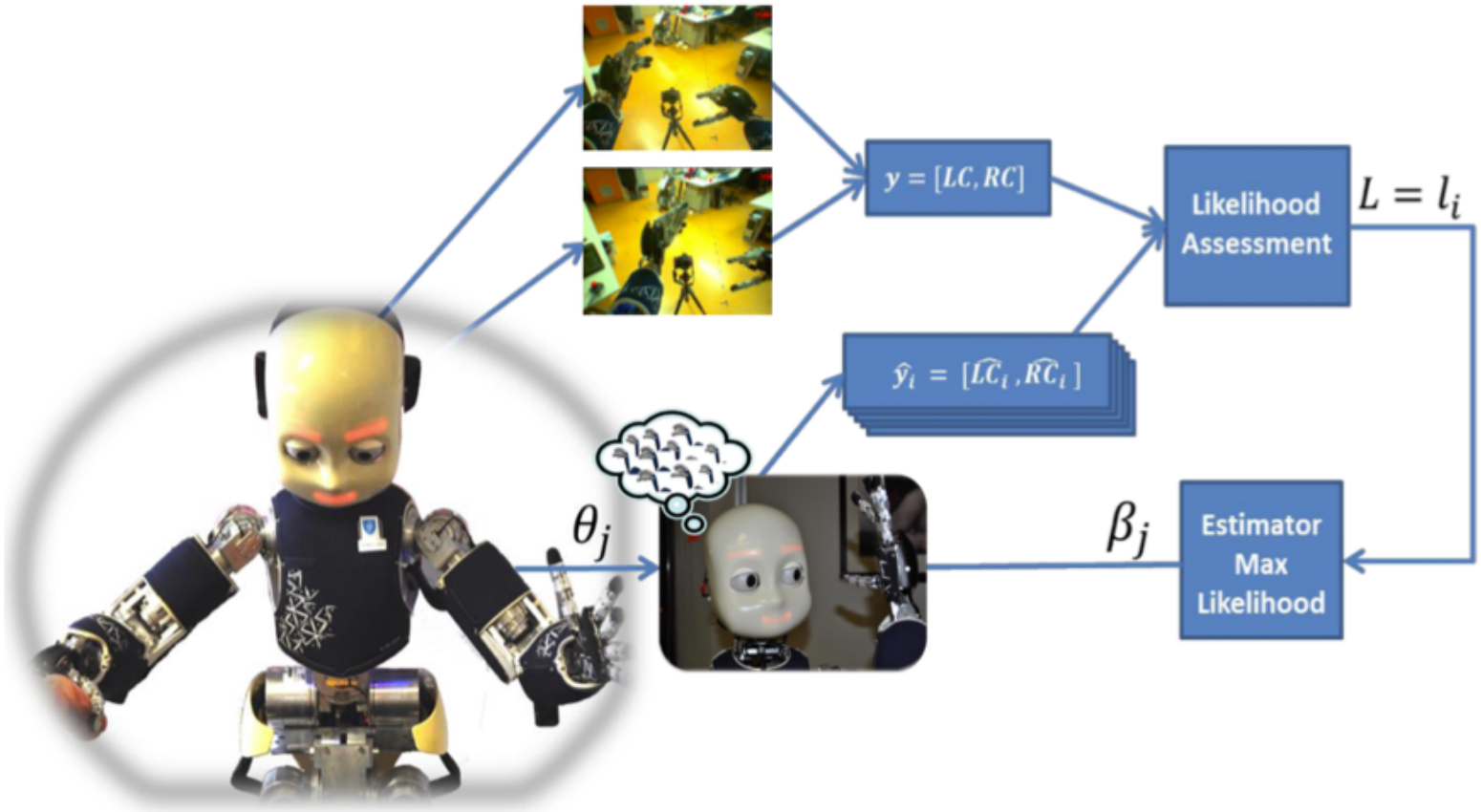


Likelihood Function

A comparison between the real and simulated images of the robot hand. Likelihood metric based on edges and distance transform, computed on the GPU (CUDA+OpenCV+OpenGL).

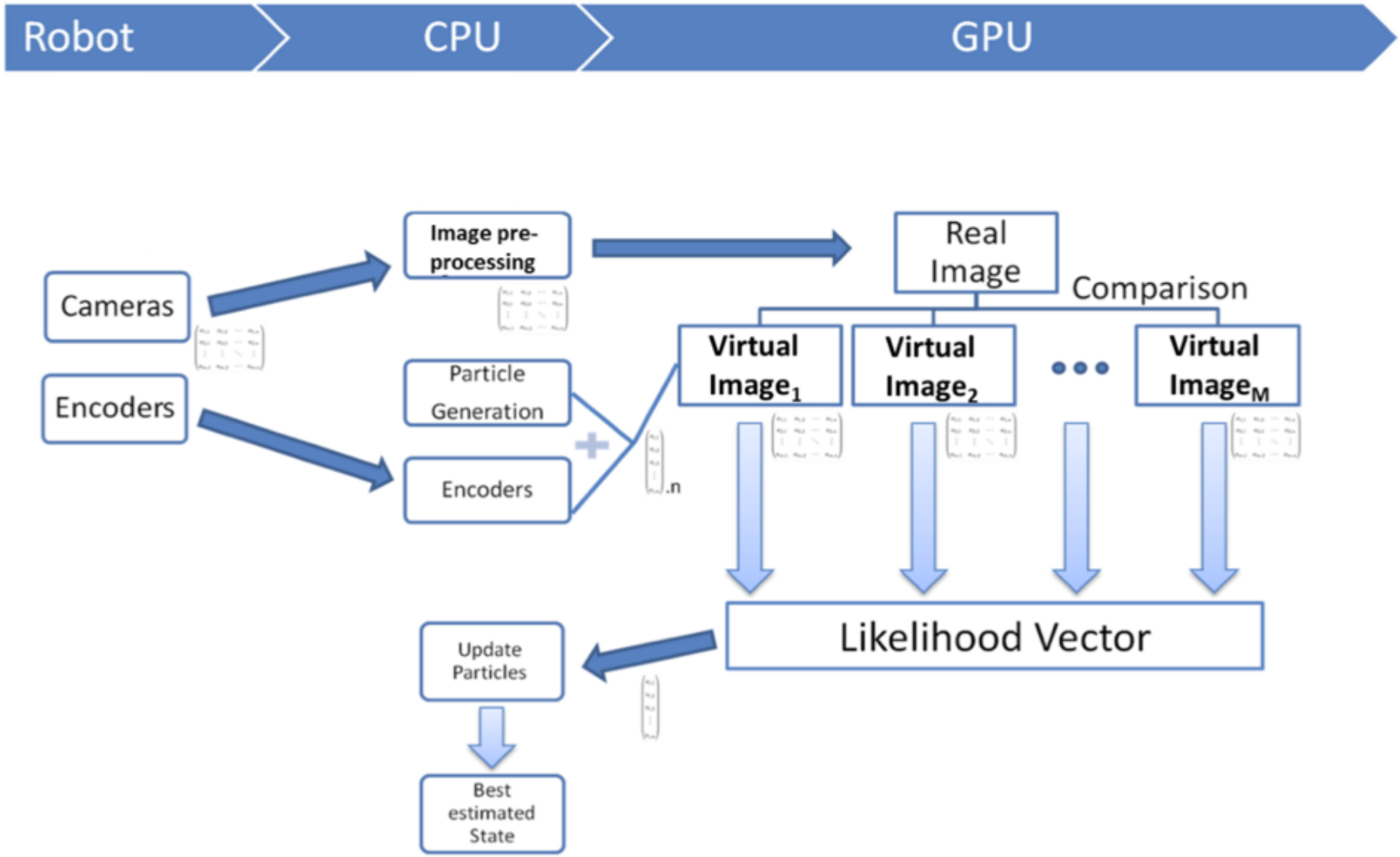
Particle Filter

The most likely hypothesis represents the current hand pose. A particle filter on joint offsets updates the body schema.



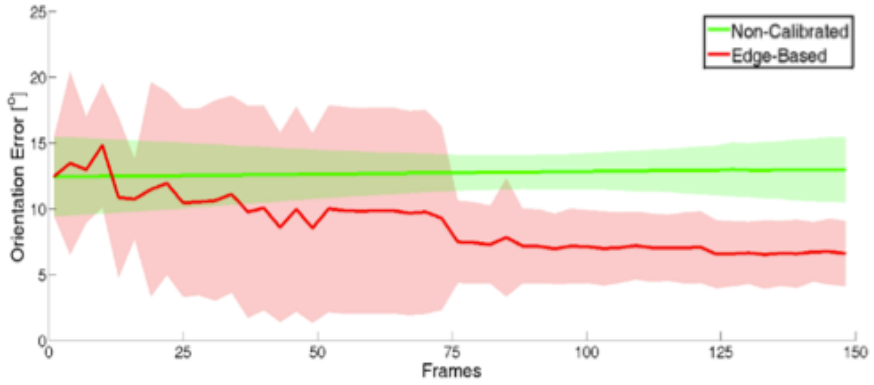
Institute for Systems and Robotics | LISBOA

Software Implementation

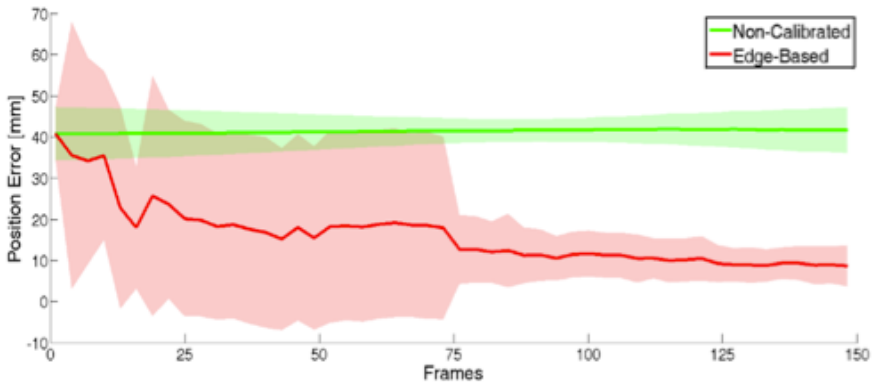


Results

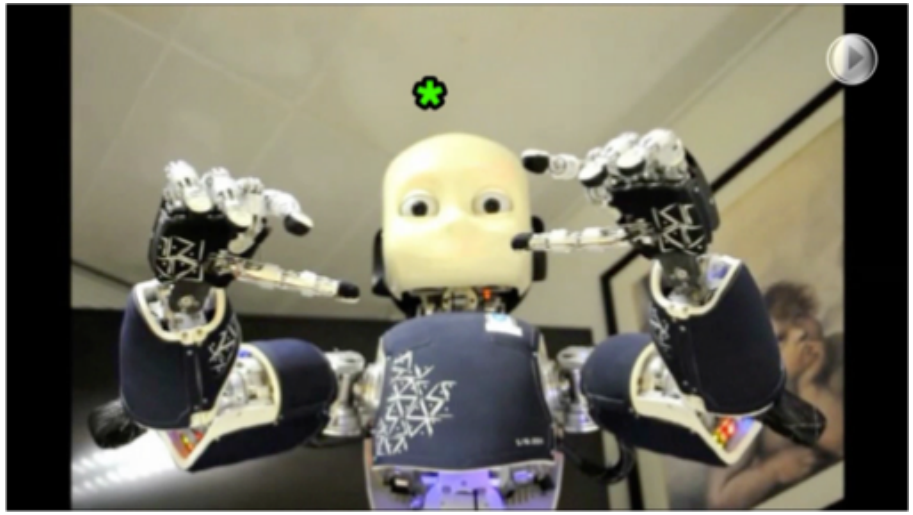
Calibration Error:

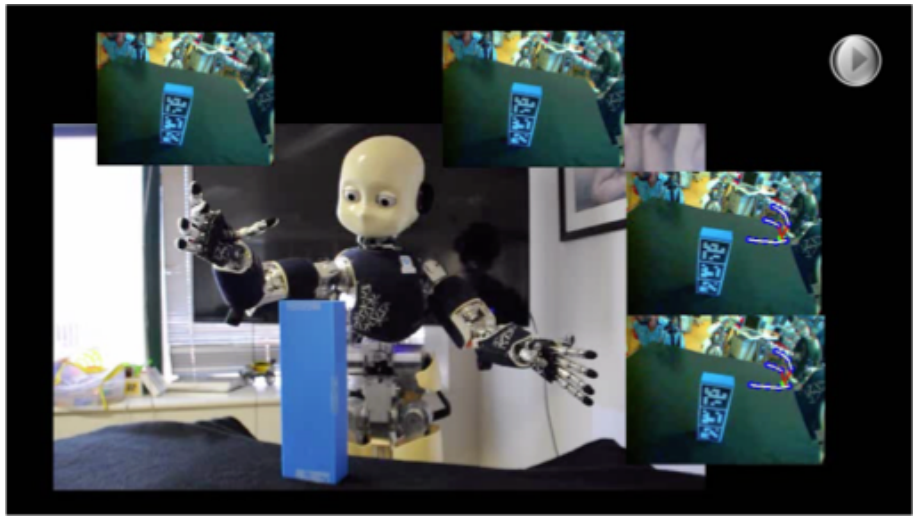


(a) Orientation Error



(b) Position Error



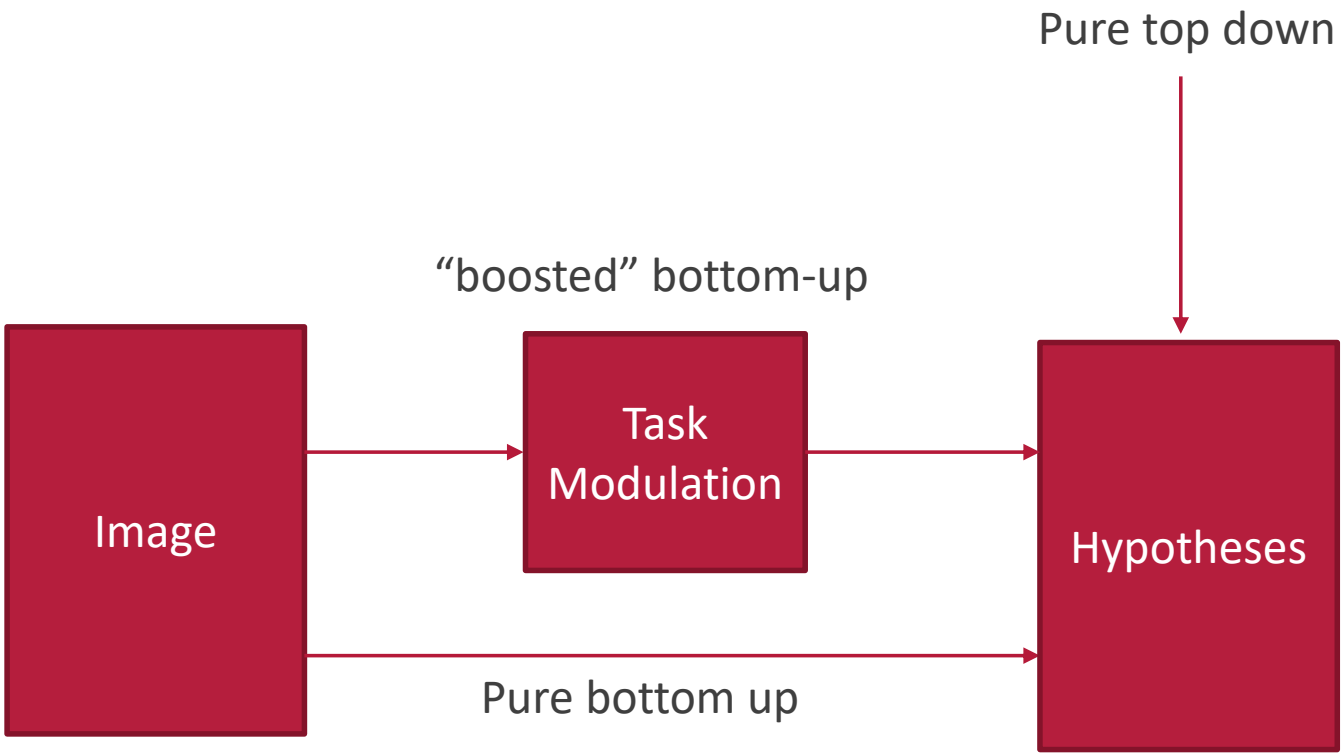


Alex Bernardino, VVV 2018

Institute for Systems and Robotics | LISBOA

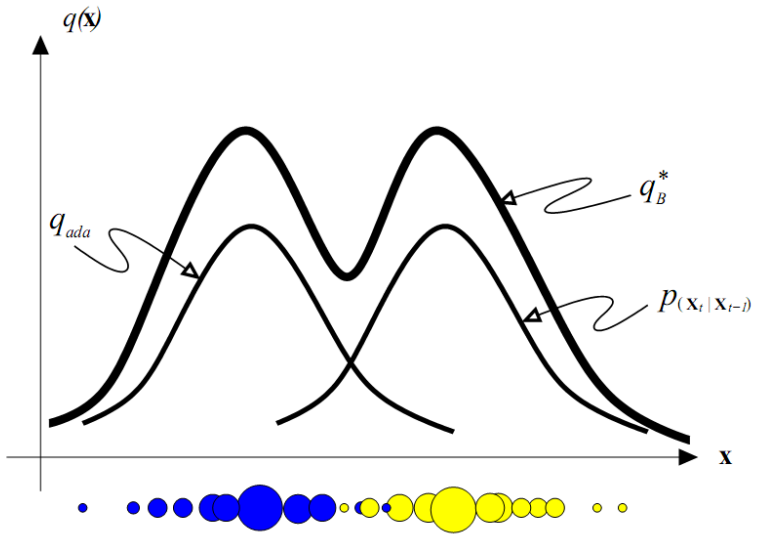
Mixed Bottom-Up Top-Down Attention

Mixed Bottom-Up Top-Down Attention



Boosted Particle Filter

- Make sampling depend on the image data.
  - Include particles based on image cues.



K. Ojuma et al, 2004

Time information: laws of physics, expectations, prior knowledge.

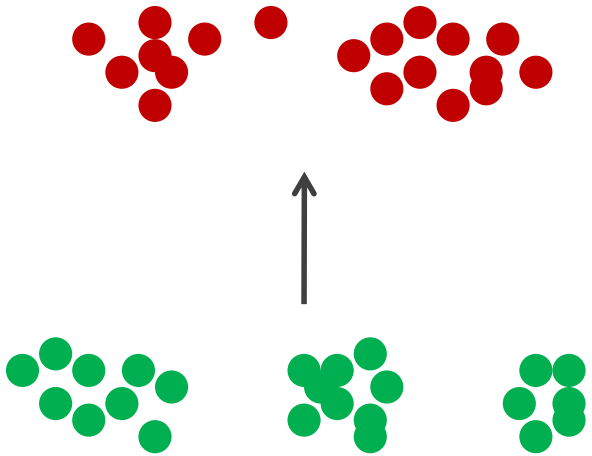
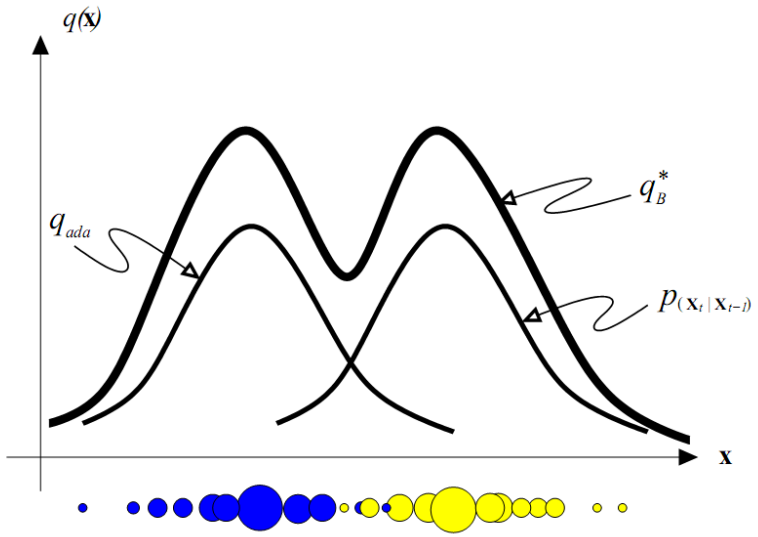


Image information: image cues, salience, visual attention.

Boosted Particle Filter

- Make sampling depend on the image data.
  - Include particles based on image cues.



K. Ojuma et al, 2004

Time Information: laws of physics, expectations, prior knowledge.

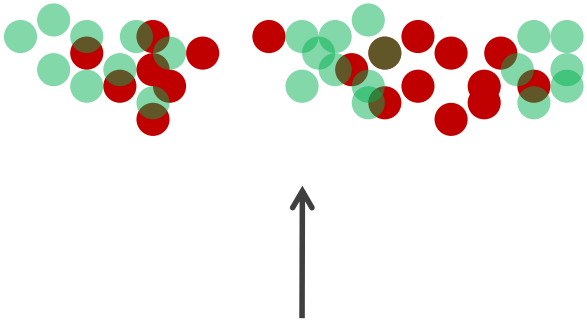


Image information: image cues, salience, visual attention.

Bottom-up Ball Detections

- Simple method based on color segmentation (known size of the ball makes ball depth related to blob area).

Institute for Systems and Robotics | LISBOA

"Boosted" Ball Tracking

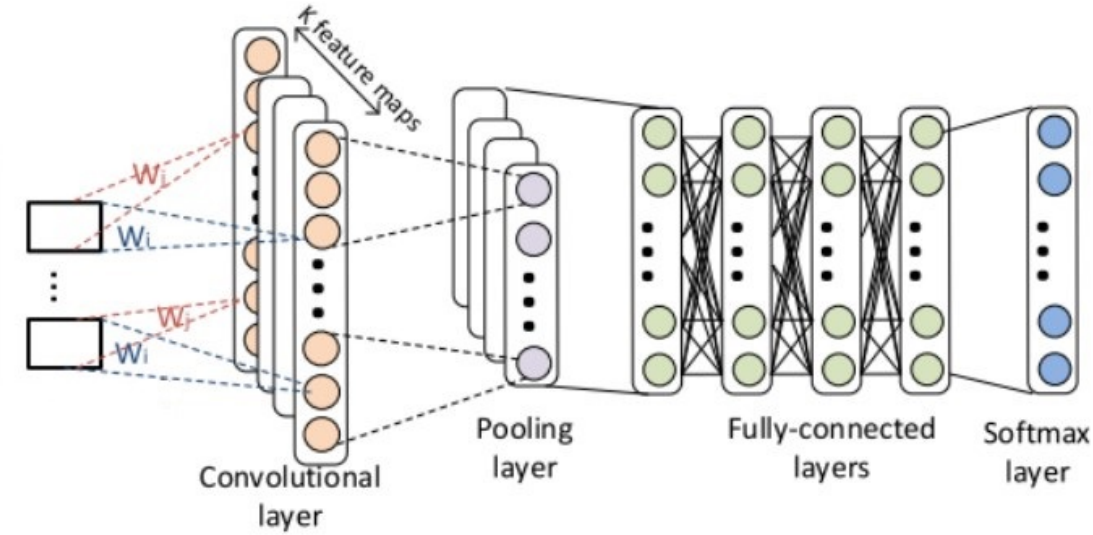
Top-down Tracker

Mixed Bottom-up/top-down Tracker

CNNs for Mixed Attention

- Deep CNN's Extract useful information without feature selection

- CaffeNet [1]
- GoogLeNet [2]
- VGGNet [3]



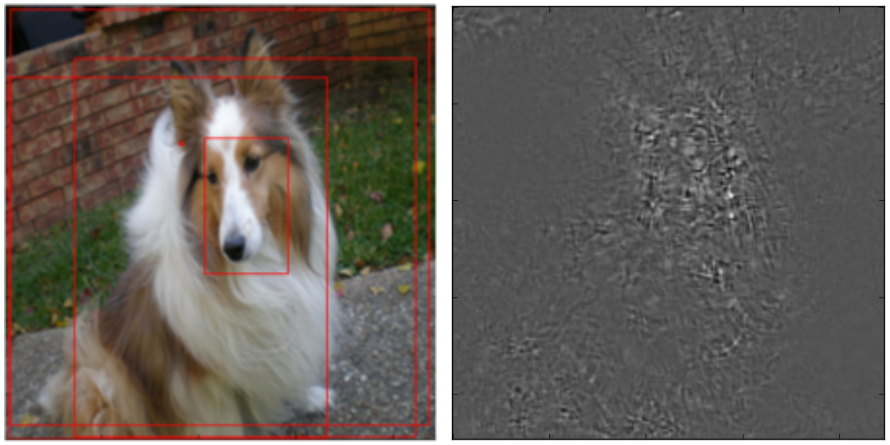
[1] A. Krizhevsky et a. "Imagenet classification with deep convolutional neural networks." In *NIPS 2012*
[2] C. Szegedy et al. "Going deeper with convolutions." In *CVPR 2015*
[3] K. Simonian et al. "Very deep convolutional networks for large-scale image recognition.." In *CVPR 2015*

CNNs for Mixed Attention

Image-Specific Class Saliency Extraction

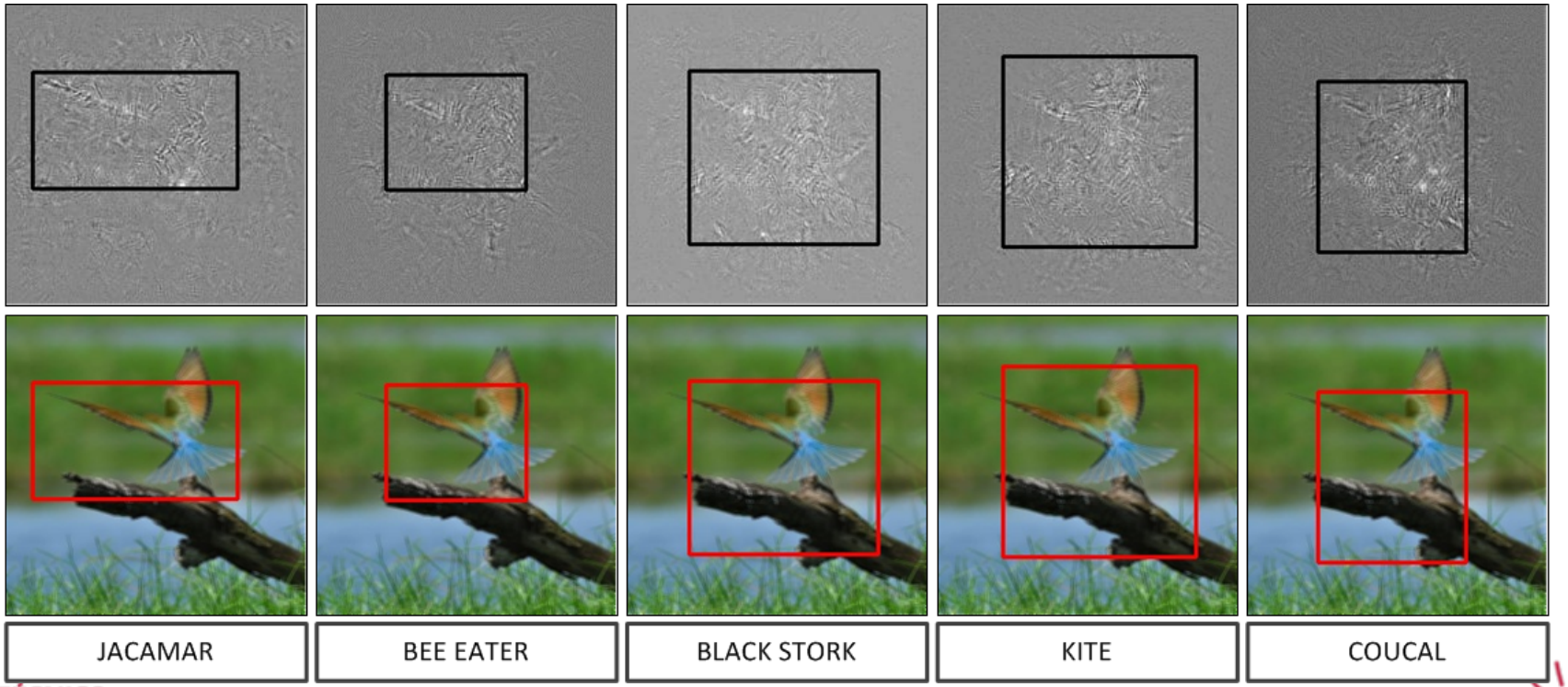
- The class score of an object class *c* in an image *I*, $S_c(I)$, is the output of the neural network for class *c* and G_c is the gradient of S_c with respect to *I*.

$$S_c(I) \approx G_c^T I + b \qquad G_c = \frac{\partial S_c}{\partial I}$$



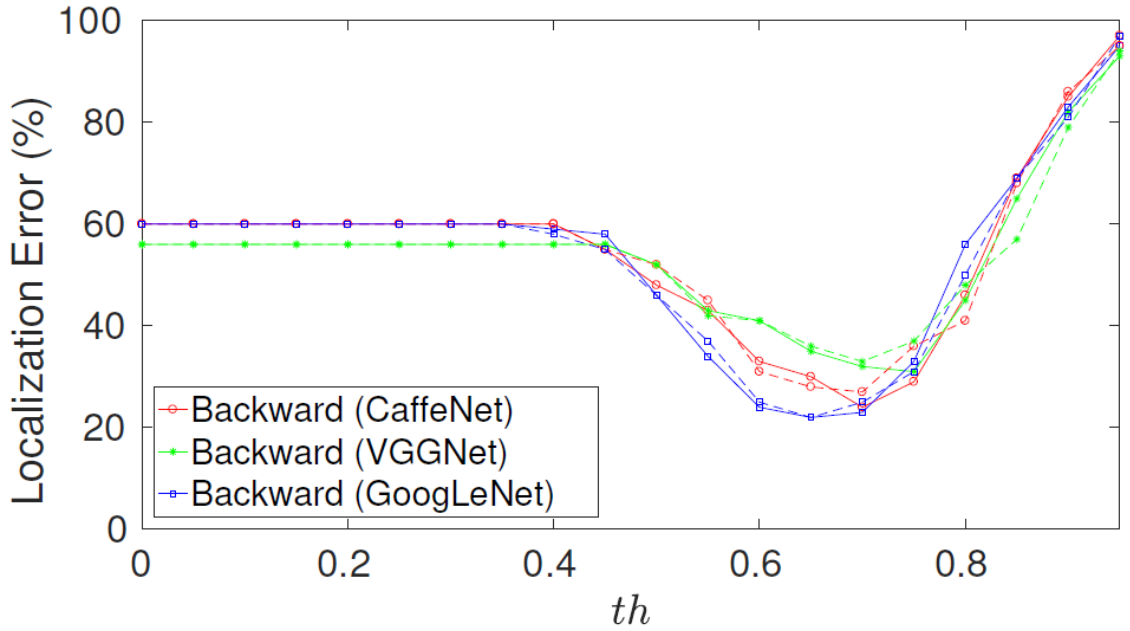
CNNs for Mixed Attention

- Class specific top down spatial localization.



CNNs for Mixed Attention

- Localization Performance



- GoogLeNet is deeper and hence can learn discriminant features at higher levels of abstraction.

Deep Networks for Human Visual Attention: A hybrid model using foveal vision, *A. F. Almeida, R. Figueiredo, A. Bernardino, J. Santos-Victor*, ROBOT 2017

Conclusions

- Real-Time Vision must be achieved by both speeding computations and reducing complexity.
- Several biological strategies are present in nature to reduce complexity:
  - Space Variant Vision
  - Visual Attention
- Foveal Sensors
  - Hardcoded reduction of the image content.
  - Fit nicely with systems with mobile cameras.
- Visual Attention
  - Concentrate computational resources on revelant items for the agent.
  - Hypotheses Testing and Particle Filtering methods are good computational paradigms to fuse bottom up with top down data.

Thank you!

Computer and Robot Vision Lab

alex@isr.tecnico.ulisboa.pt