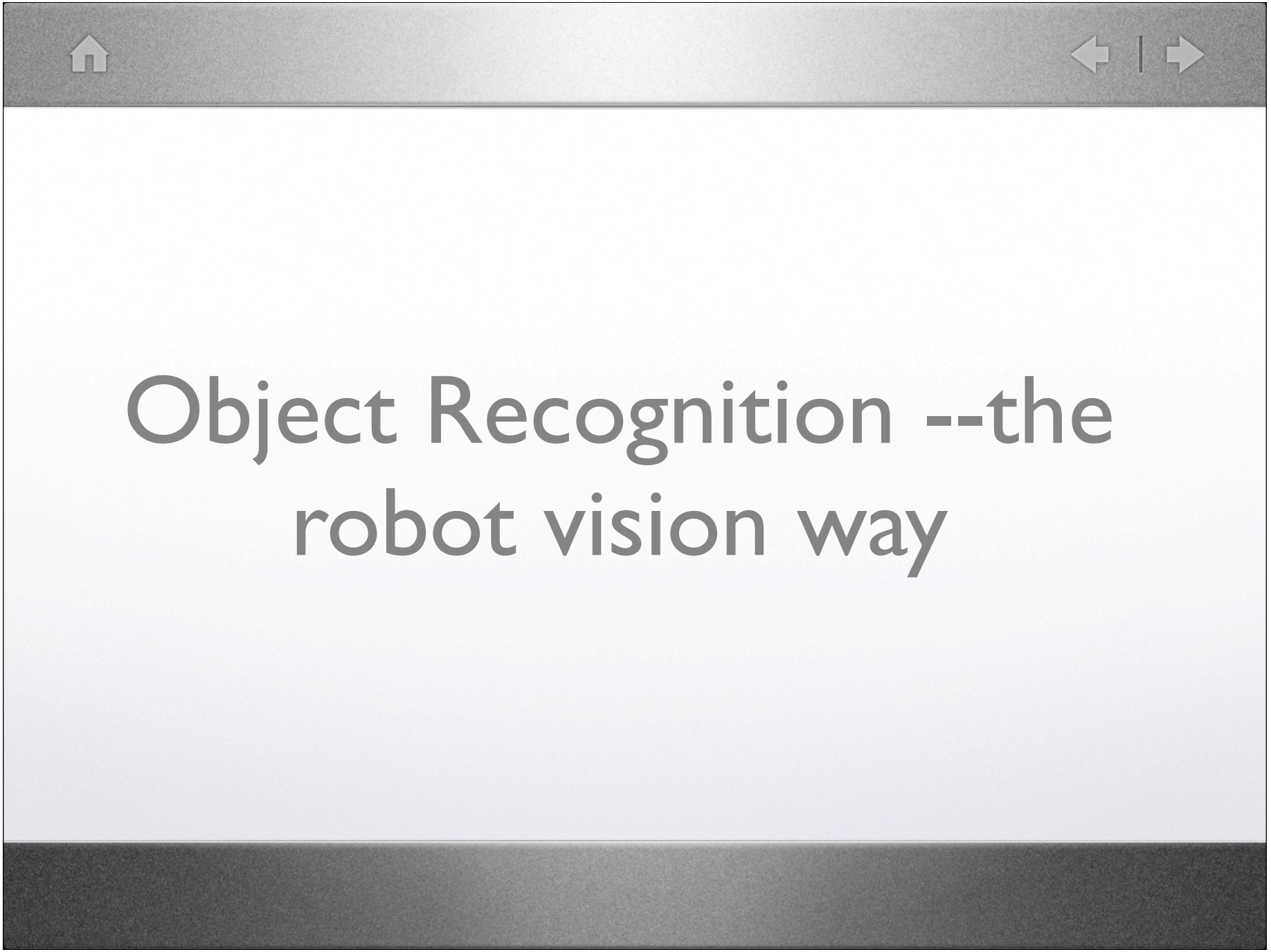


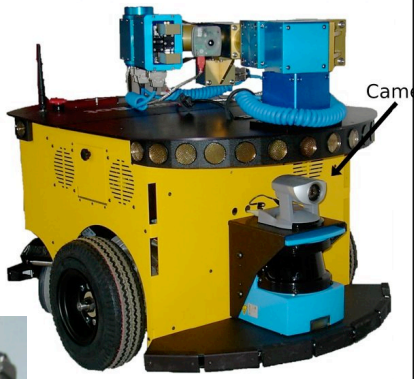

Cognitive Vision for Cognitive Systems

Barbara Caputo, Marco Fornoni
Idiap Research Institute
http://www.idiap.ch/~bcaputo
http://www.idiap.ch/~mfornoni
bcaputo@idiap.ch
mfornoni@idiap.ch

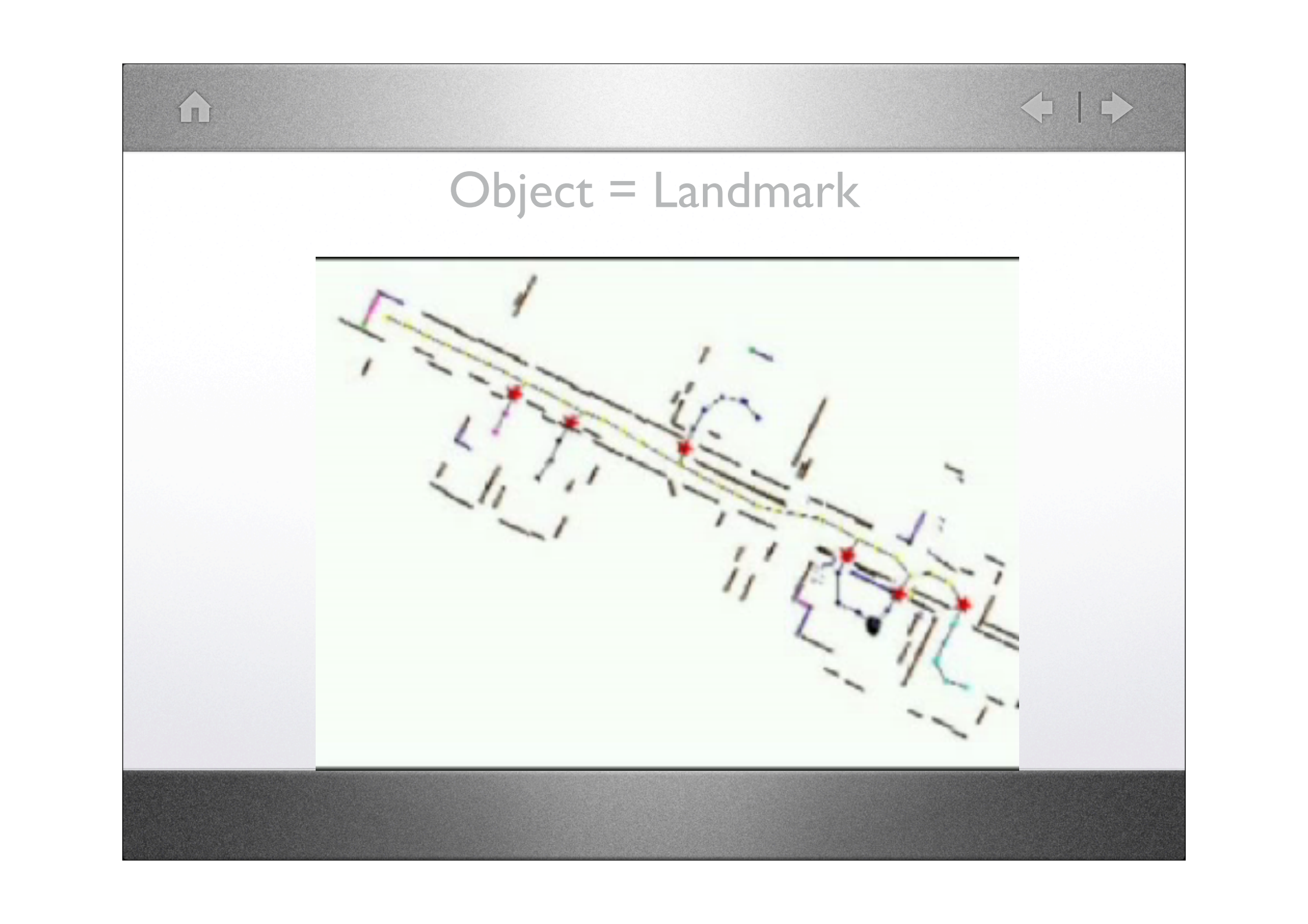

Object Recognition --the
robot vision way

What is an Object for a Robot?

- A visual landmark for helping localization, mapping and navigation
- An obstacle to be avoided
- Something to grasp/ manipulate




Objects as Visual Landmarks


Object = Landmark

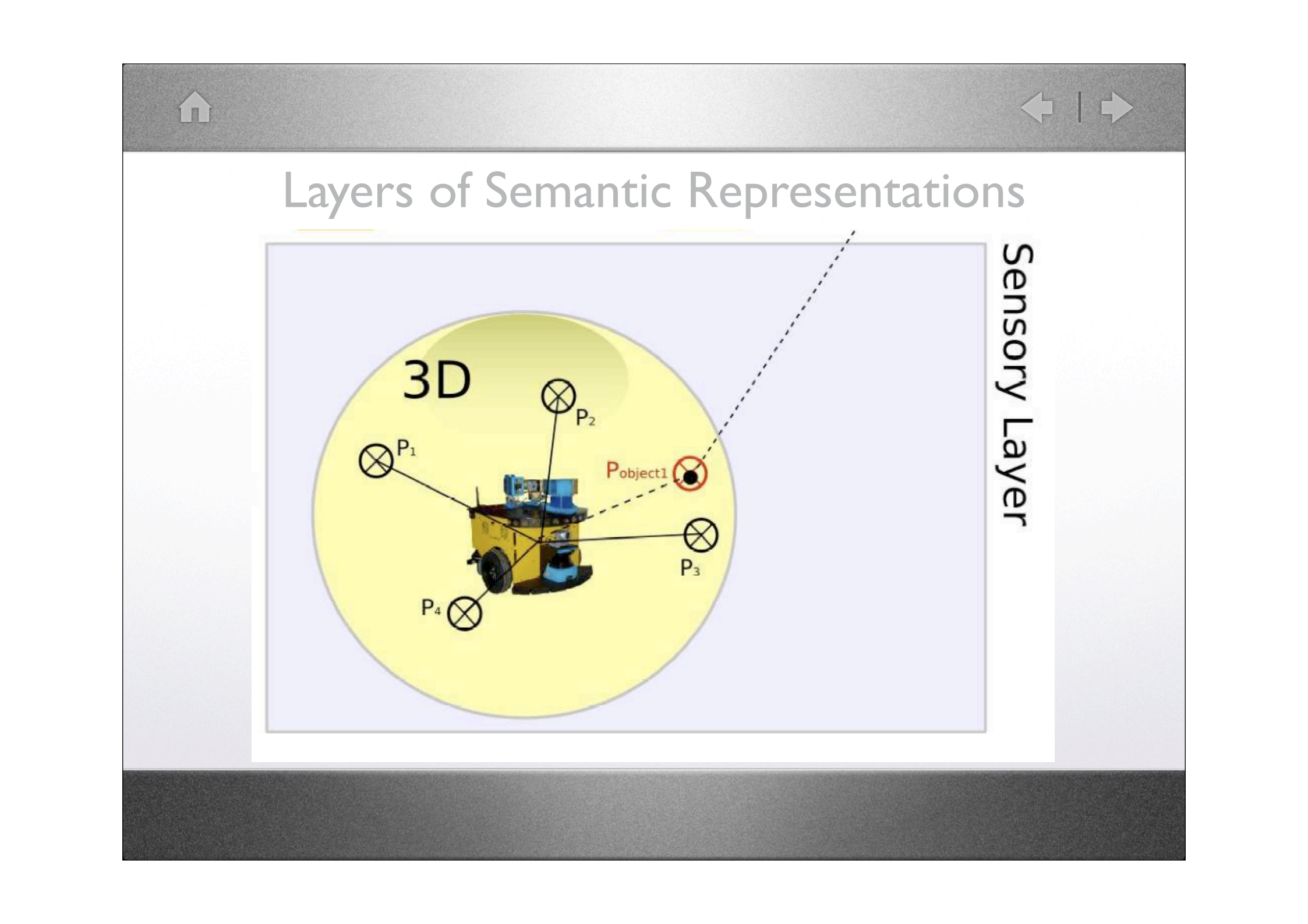

Objects as Visual Landmarks

- Objects are used to enrich the map of the environment
- Objects are used to facilitate localization in an environment
- Objects can be referred to/constitute a goal when navigating in an environment

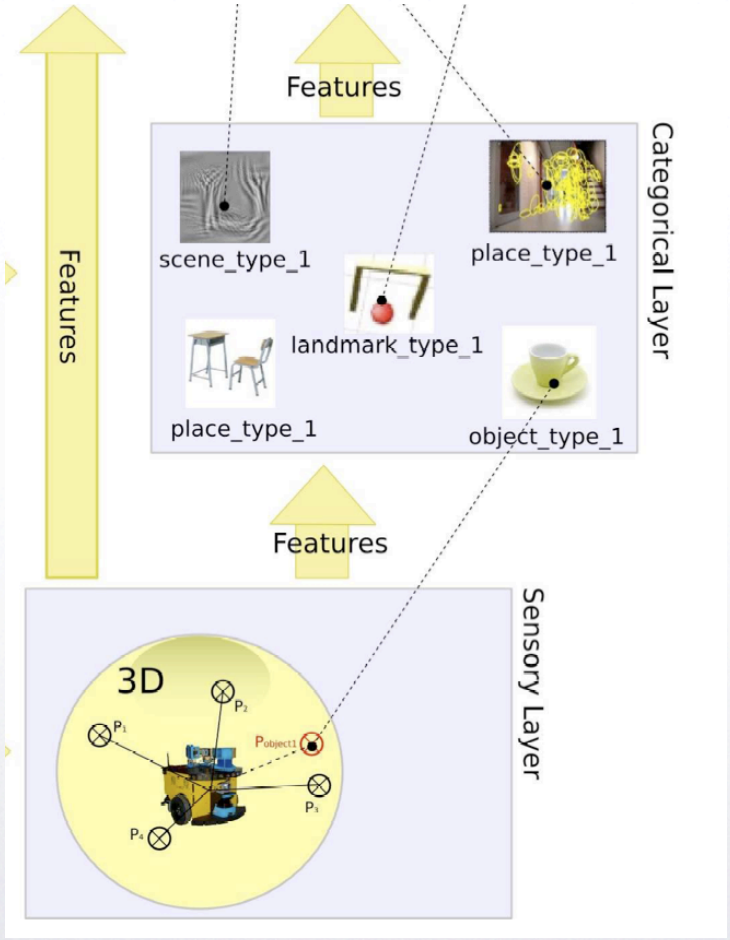
almost no need for 3D information....
we are back to appearance based!

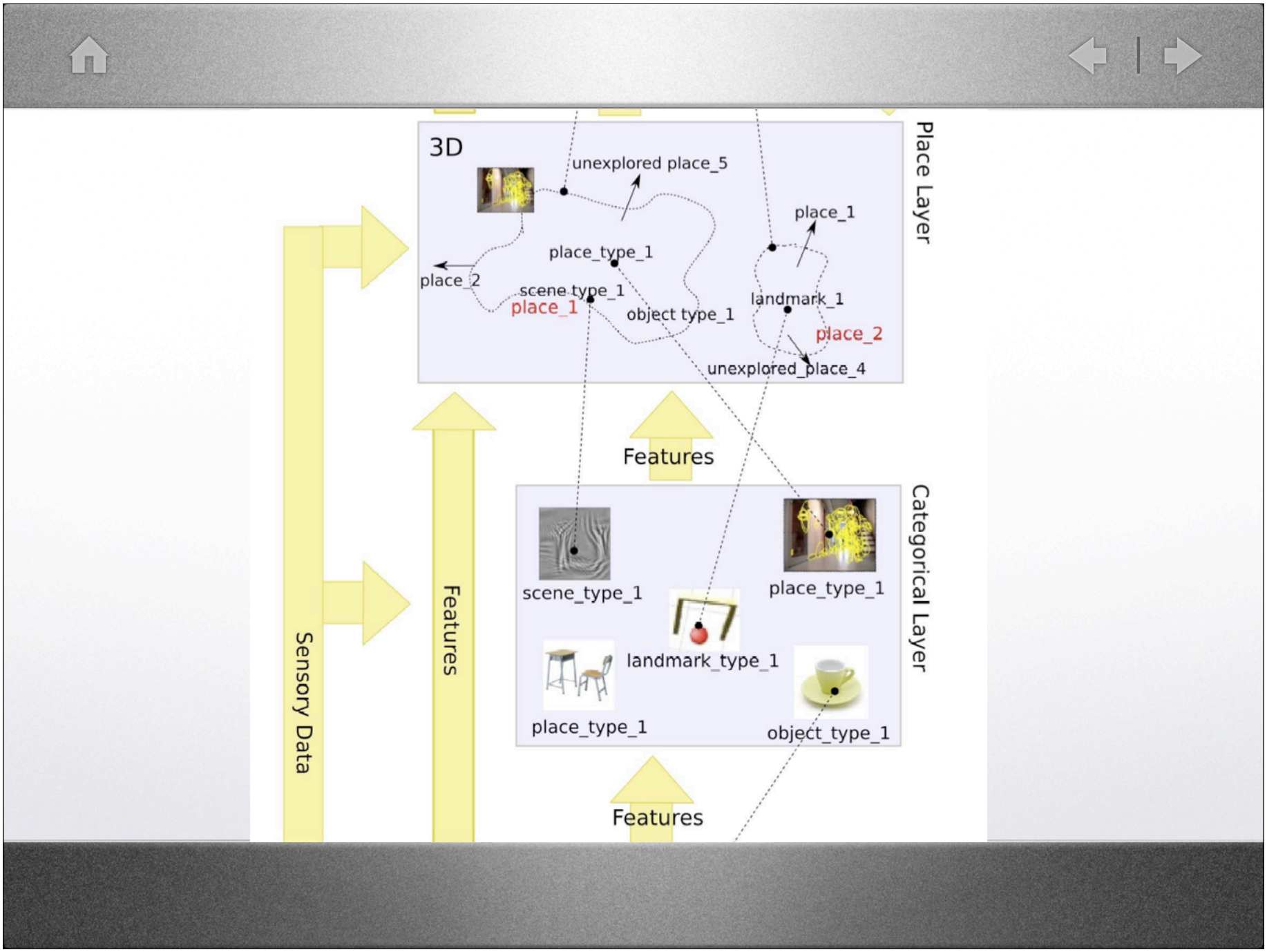

Key differences

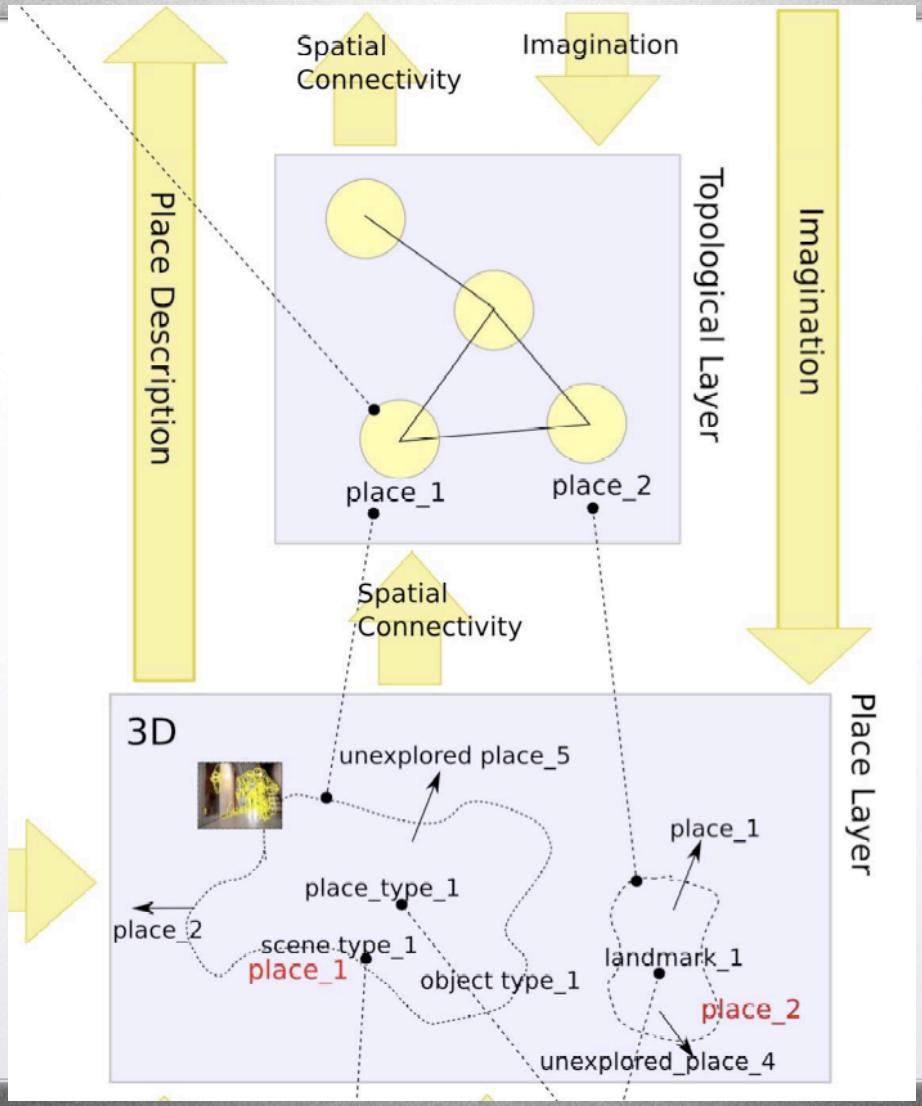
- Static images versus image sequences
- Training time and memory not relevant versus real time and memory bounded relevant
- Robustness to changes in illumination, scale, viewpoint is crucial

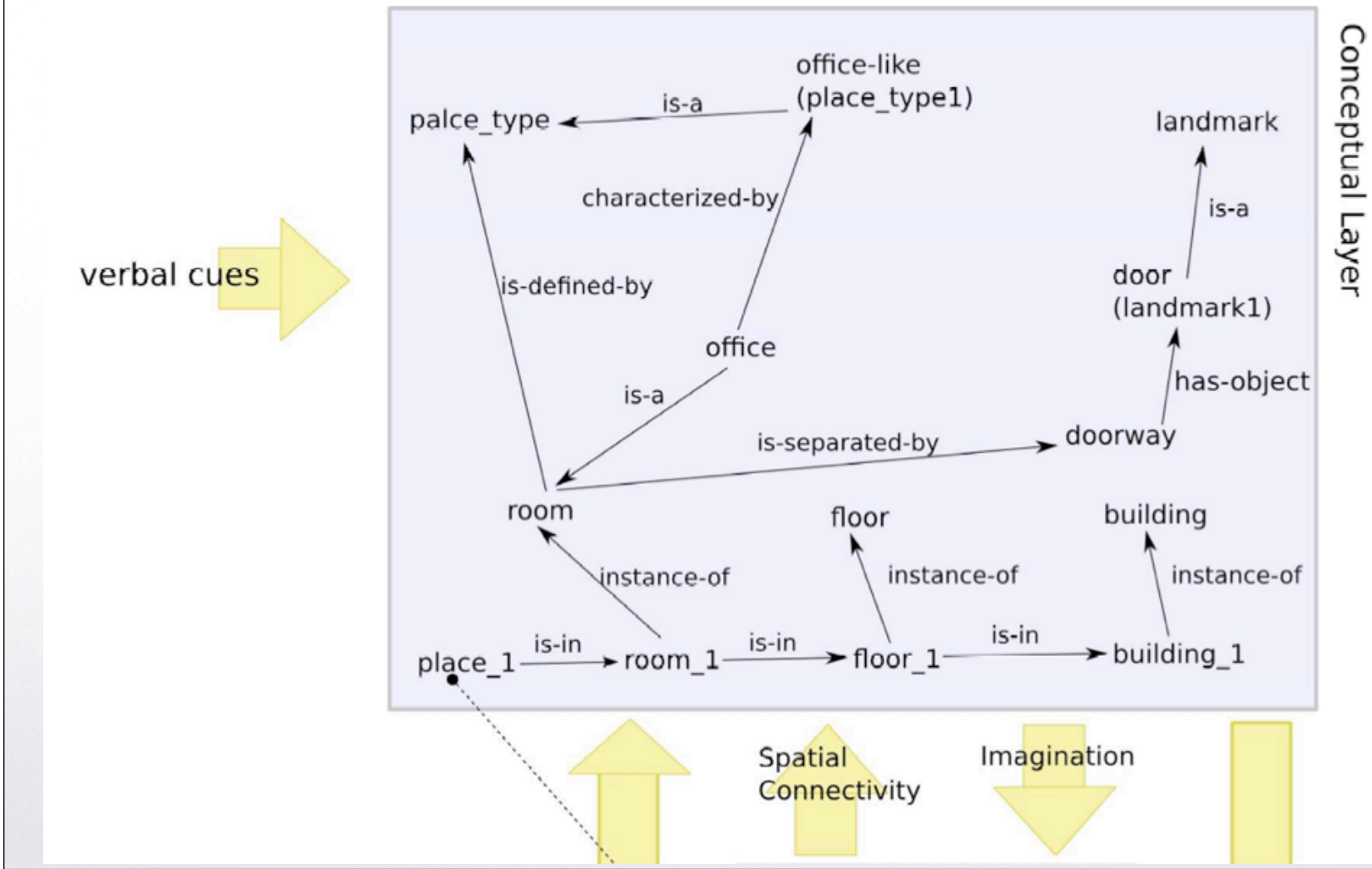

Layers of Semantic Representations
3D
P1
P2
P3
P4
Pobject1
Sensory Layer

Layers of Semantic Representations

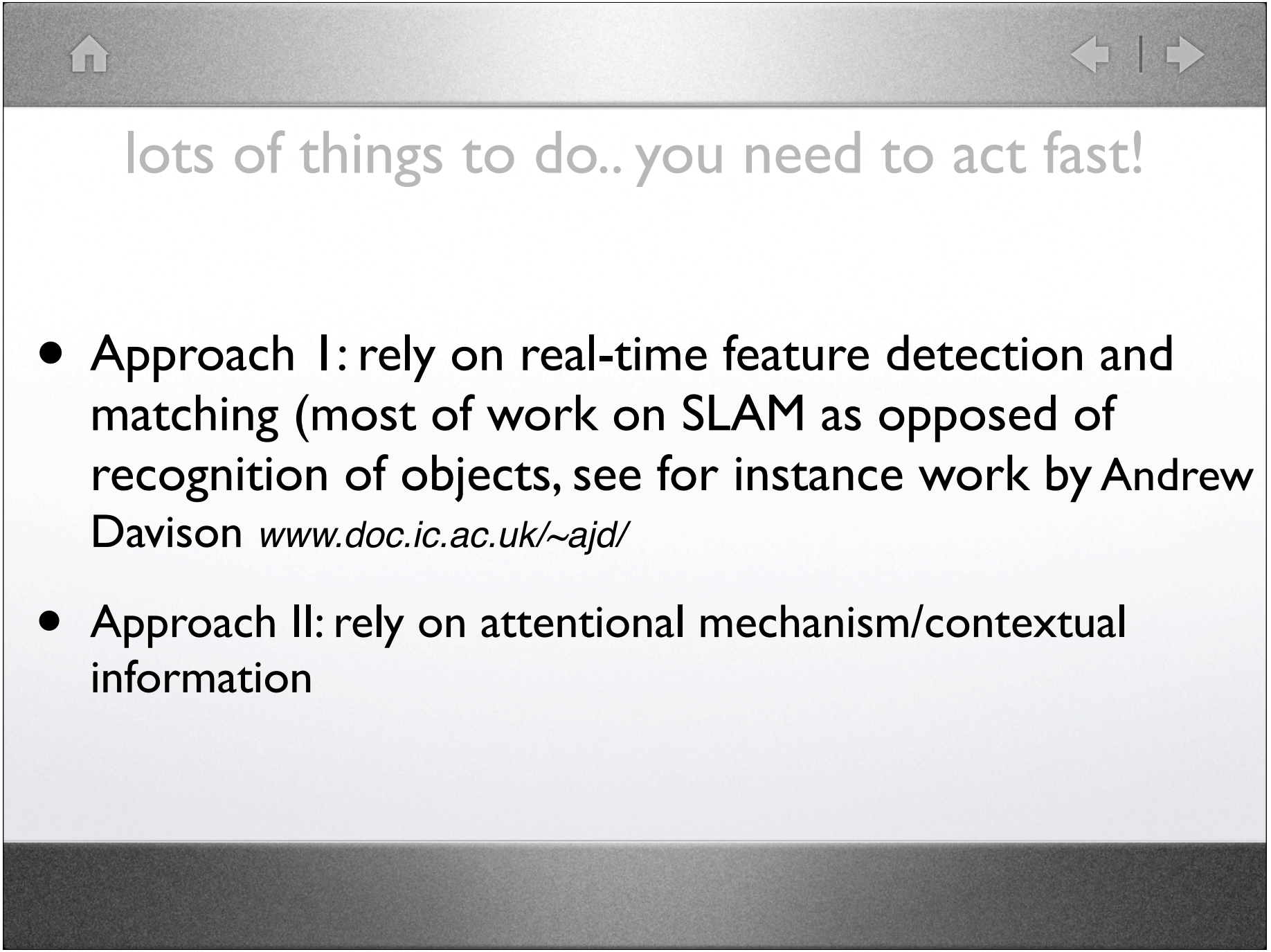



3D
unexplored place_5
place_1
place_type_1
place_2
scene type_1
place_1
object type_1
landmark_1
place_2
unexplored_place_4
Place Layer
Features
scene_type_1
place_type_1
landmark_type_1
place_type_1
object_type_1
Categorical Layer
Features
Sensory Data
Features

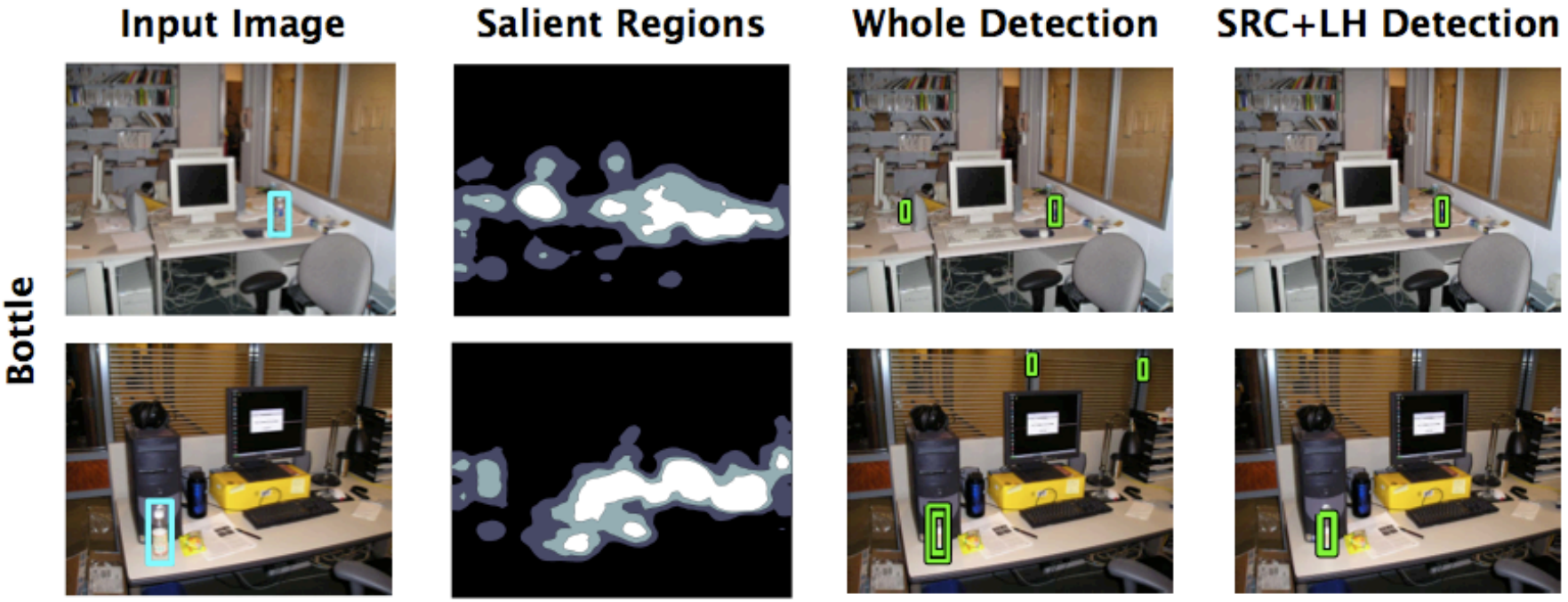

Spatial
Connectivity
Imagination
Place Description
Topological Layer
Imagination
place_1
place_2
Spatial
Connectivity
3D
Place Layer
unexplored place_5
place_1
place_type_1
place_2
scene type_1
place_1
object type_1
landmark_1
place_2
unexplored_place_4


verbal cues
office-like
(place_type1)
is-a
palce_type
landmark
characterized-by
is-a
is-defined-by
door
(landmark1)
office
has-object
is-a
is-separated-by
doorway
room
floor
building
instance-of
instance-of
instance-of
is-in
is-in
is-in
place_1
room_1
floor_1
building_1
Spatial
Connectivity
Imagination
Conceptual Layer

Conceptual Layer
landmark
is-a
door (landmark1)
has-object
doorway
building
instance-of
building_1
office like (place_type1)
is-a
floor
instance-of
is-in
floor_1
characterized-by
office
is-a
is-separated-by
is-in
room_1
is-defined-by
room
instance-of
is-in
place_1
place_type
Imagination
Topological Layer
Imagination
place_2
Spatial Connectivity
place_1
Spatial Connectivity
Place Description
Place Layer
place_1
landmark_1
place_2
unexplored place_4
unexplored place_5
object type_1
place type_1
scene type_1
place_1
3D
Features
Categorical Layer
place type_1
object type_1
landmark type_1
scene type_1
place type_1
Features
Features
Sensory Layer
3D
Sensory Data
verbal cues


lots of things to do.. you need to act fast!
• Approach I: rely on real-time feature detection and matching (most of work on SLAM as opposed of recognition of objects, see for instance work by Andrew Davison www.doc.ic.ac.uk/~ajd/
• Approach II: rely on attentional mechanism/contextual information


- Approach II: rely on attentional mechanism/ contextual information

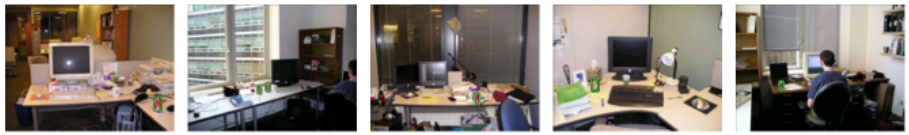


[Choi&Christensen, IROS 2009]

- Two stage object recognition approach

- Stage I: identify salient regions in the scene using multiple cues
- Stage II: identify objects inside the salient regions

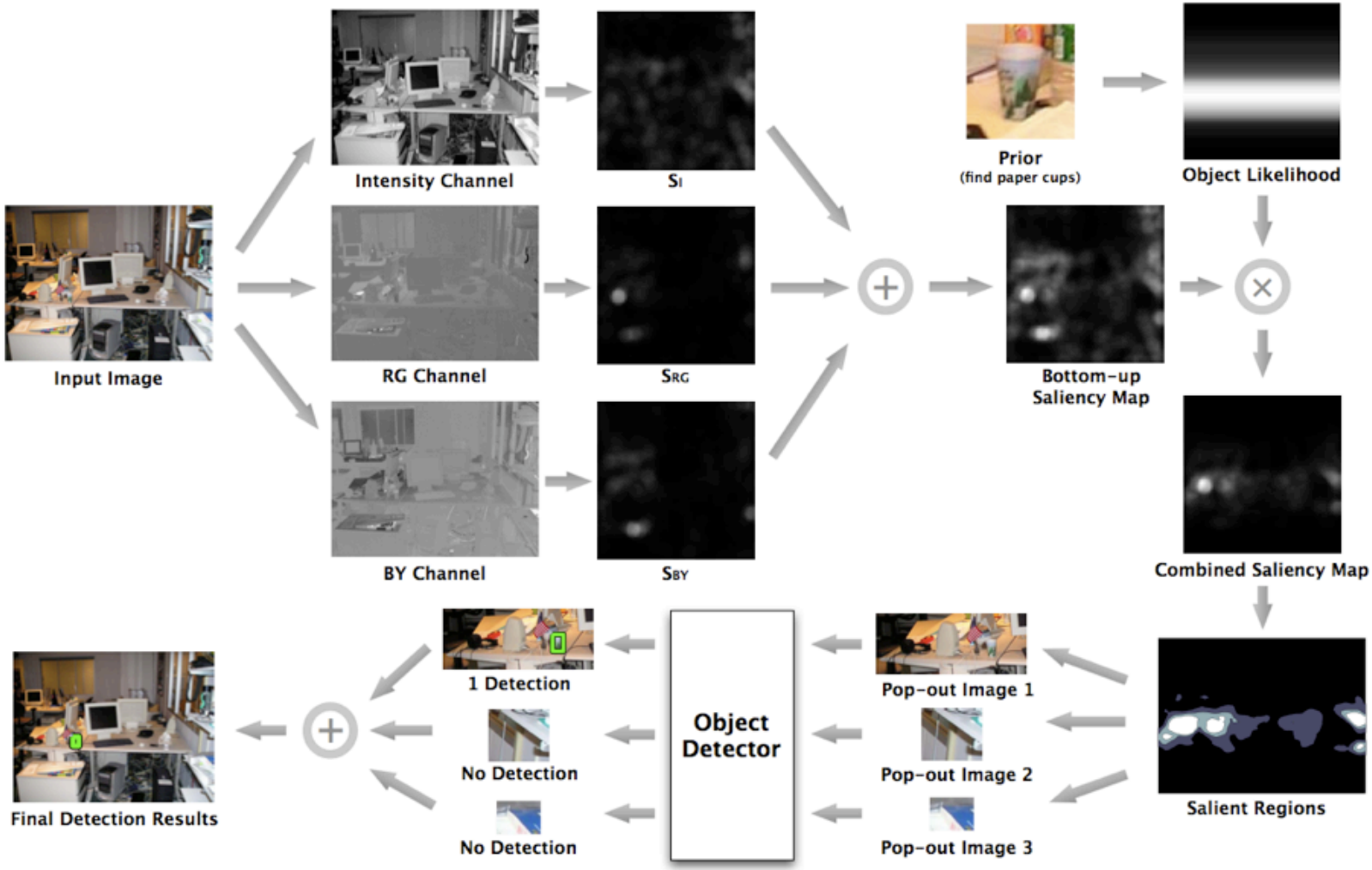
(a) Bottle

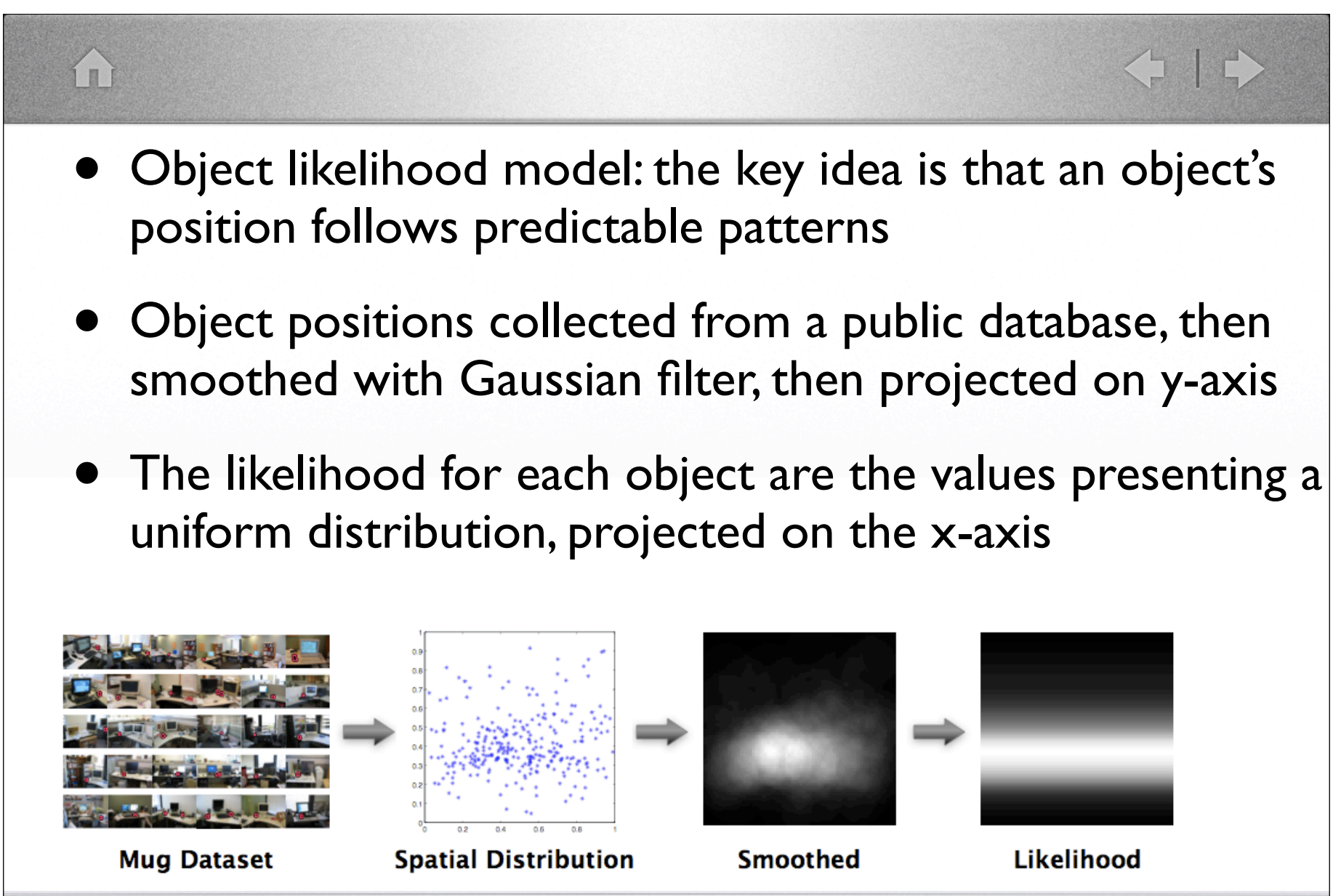


(b) Can

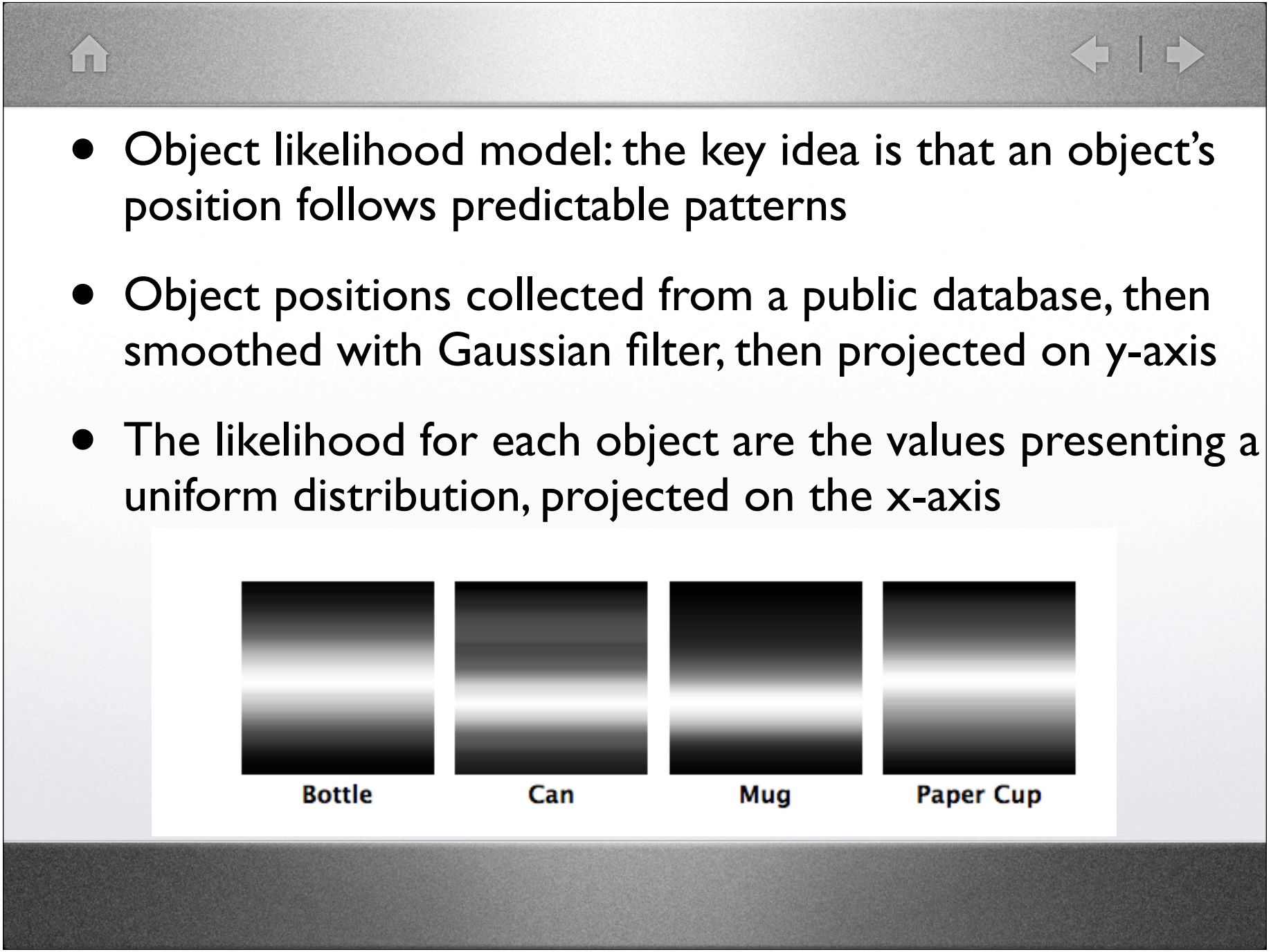
(c) Mug

(d) Paper cup

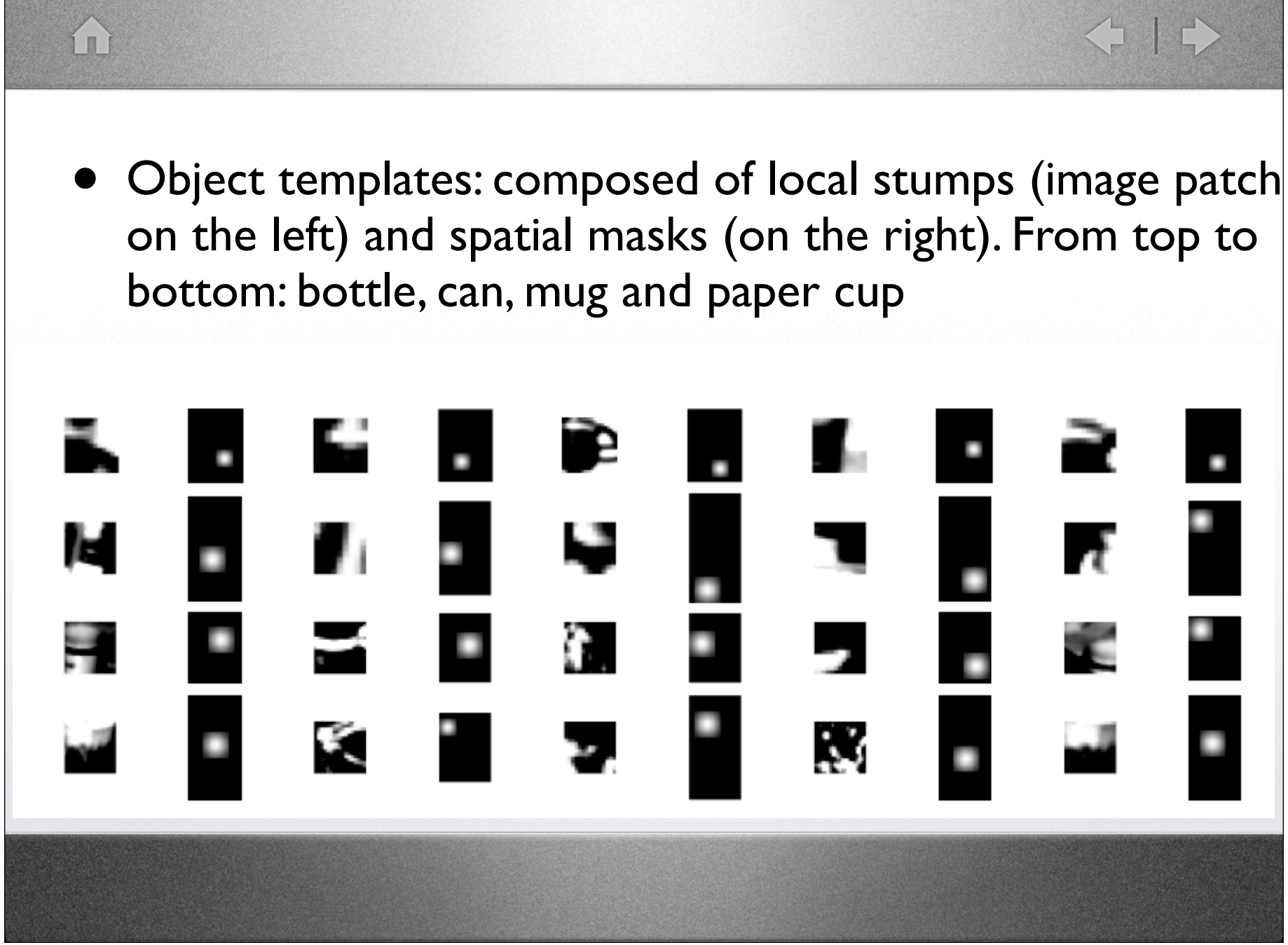

Intensity Channel
SI
Prior
(find paper cups)
Object Likelihood
Input Image
RG Channel
SRG
Bottom-up
Saliency Map
BY Channel
SBY
Combined Saliency Map
1 Detection
Pop-out Image 1
Object
Detector
Salient Regions
No Detection
Pop-out Image 2
Final Detection Results
No Detection
Pop-out Image 3



- Object likelihood model: the key idea is that an object's position follows predictable patterns
- Object positions collected from a public database, then smoothed with Gaussian filter, then projected on y-axis
- The likelihood for each object are the values presenting a uniform distribution, projected on the x-axis

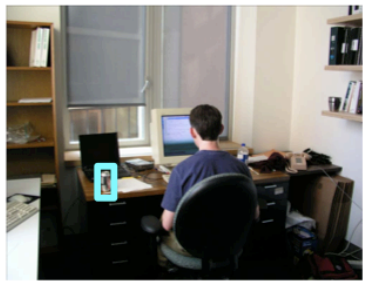


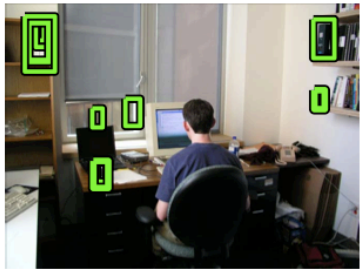
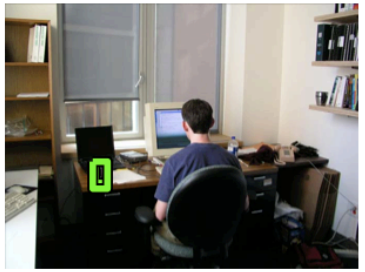
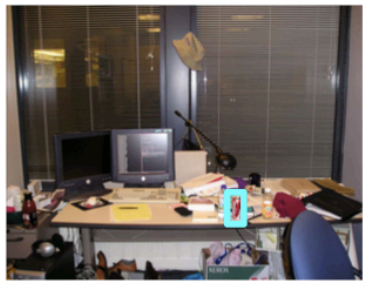
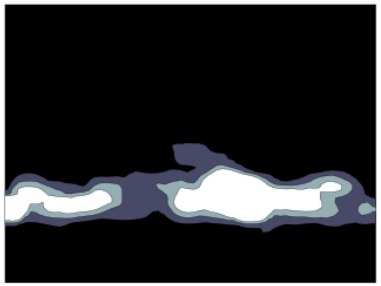
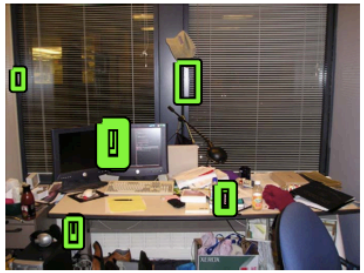
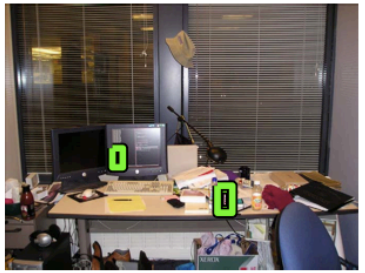
- Object likelihood model: the key idea is that an object's position follows predictable patterns
- Object positions collected from a public database, then smoothed with Gaussian filter, then projected on y-axis
- The likelihood for each object are the values presenting a uniform distribution, projected on the x-axis



- Object templates: composed of local stumps (image patch on the left) and spatial masks (on the right). From top to bottom: bottle, can, mug and paper cup


Input Image
Salient Regions
Whole Detection
SRC+LH Detection
Bottle



Salient Regions Whole Detection SRC+LH Detection

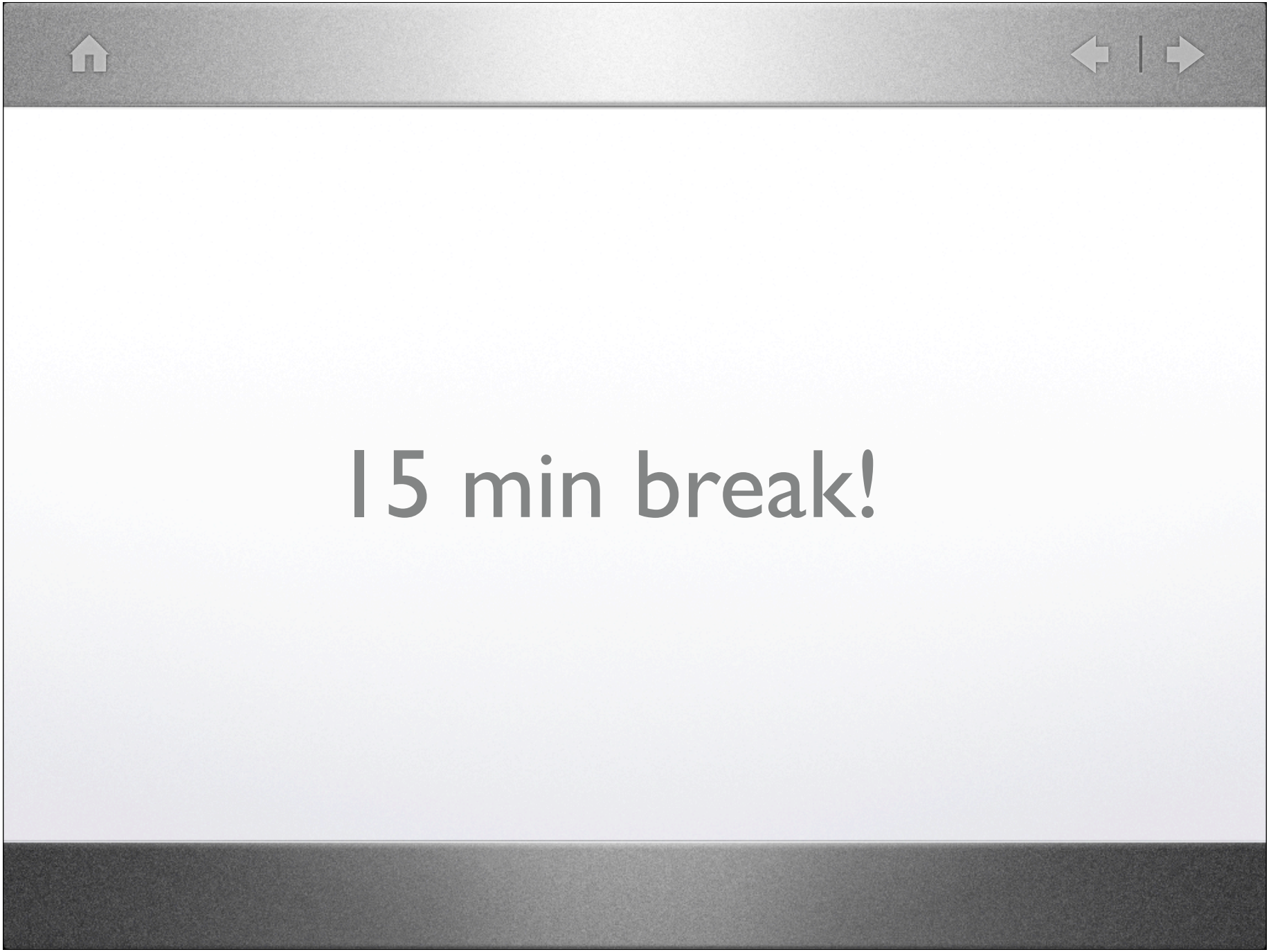
Can


Salient Regions
Whole Detection
SRC+LH Detection
Mug


Salient Regions
Whole Detection
SRC+LH Detection
Paper Cup


Take home message

- appearance-based methods are useful for detecting landmark objects
- the system resources are the key factor (speed, memory)
- contextual information/attention mechanisms/real-time visual recognition algorithms


15 min break!


Manipulable Objects
and Affordances

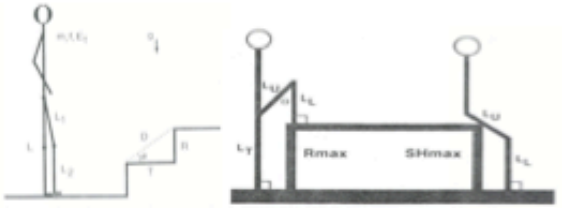
The Concept of Affordances

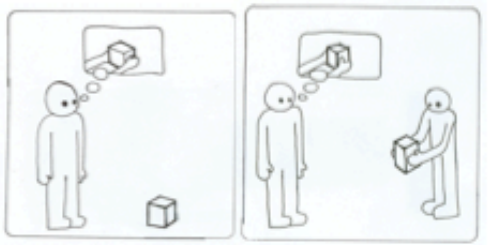
- Introduced J.J. Gibson to explain
  - how inherent "values" and "meanings" of things in the environment can be directly perceived, and
  - how this information can be linked to the action possibilities offered to the organism by the environment.
- Gibson argued that an organism and its environment complement each other and that studies on the organism should be conducted in its natural environment rather than in isolation
- An elusive, yet confusing notion that has influenced a wide range of files ranging from Human-Computer Interaction and Neuroscience, to Robotics.

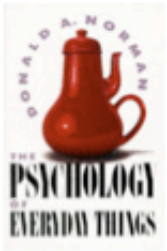
J.J. Gibson (1904–1979)

Affordance in Different Fields

- Ecological Psychology
  - Warren's (1984) stair-climbing experiments
    - affordances are perceived in body-scaled metrics.
- Neuroscience
  - Canonical and Mirror neurons, that are used in motor actions, are also observed to be active during perception.
- Human-Computer Interaction
  - How "everyday things" can be designed such that the user can easily infer what they afford. (D.Norman; 1998).
  - Identify the visual clues that make the affordances of the tools apparent.








Affordances in Robotics

- The concept of affordances were also used as guiding principles for the design of behaviors in robotic systems (Duchon et al; 1998, Murphy; 1999)
- Affordance learning is referred to as the learning of the consequences of a certain action in a given situation (Fitzpatrick et al.; 2003, Stoytchev; 2005a, 2005b).
- Learning of the invariant properties of environments that afford a certain behavior (Cos-Aguilera et al.;2003, 2004, MacDorman; 2000).
  - These studies also relate these properties to the consequences in terms of the internal values of the agent, rather than changes in the environment.

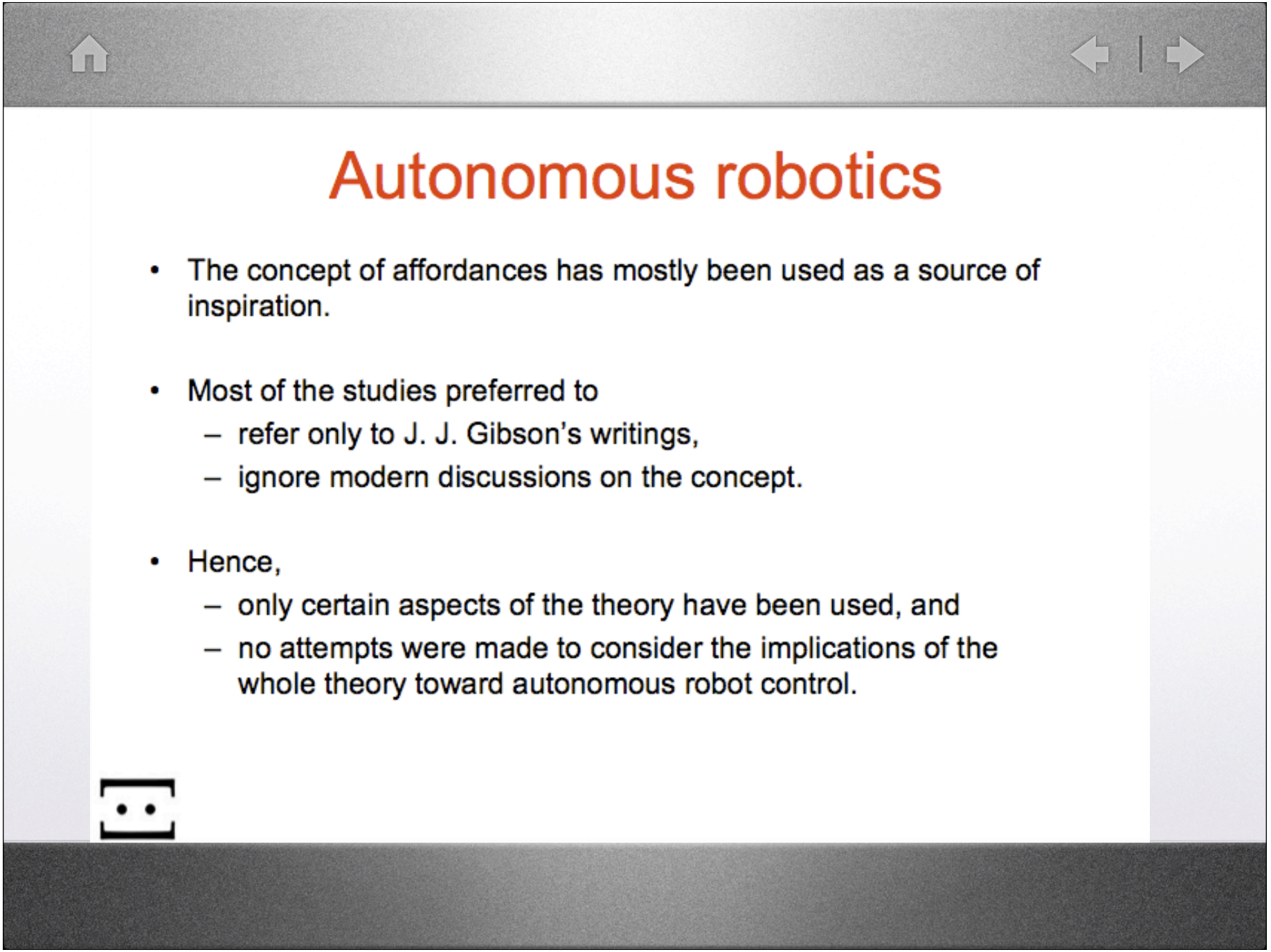
Autonomous Robotics

- The concept of affordances and behavior-based robotics have emerged in similar ways, objecting to the then dominant paradigms in their fields.

- Contemporary view: the meaning of objects are created internally with further "mental calculation" of the otherwise meaningless perceptual data.
- Gibson's view: affordances are directly perceivable (a.k.a. *direct perception*) by the organism, thus the meaning of the objects in the environment are directly apparent to the agent acting in it.

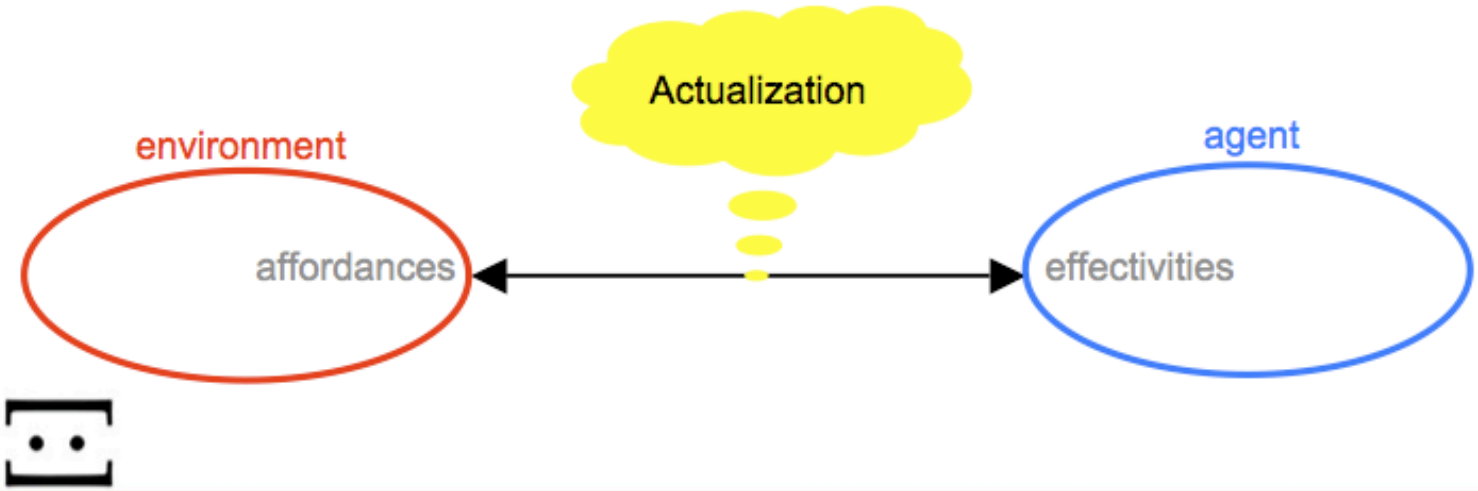
Contemporary view: Robot's perception should build and maintain a generic world model of the environment, over which the robot can make inferences.

- Brooks' view: "The world is its own best model" and there is no need for internal representations. The robot's behaviors should directly connect perception to action, and cognition will emerge.


Autonomous robotics
• The concept of affordances has mostly been used as a source of inspiration.
• Most of the studies preferred to
– refer only to J. J. Gibson's writings,
– ignore modern discussions on the concept.
• Hence,
– only certain aspects of the theory have been used, and
– no attempts were made to consider the implications of the whole theory toward autonomous robot control.

Turvey's formalization (1992)

- According to Turvey:
  - Affordances are dispositional properties of the environment
  - Effectivities are dispositional properties of the animal
  - When these two meet in space and time they get actualized

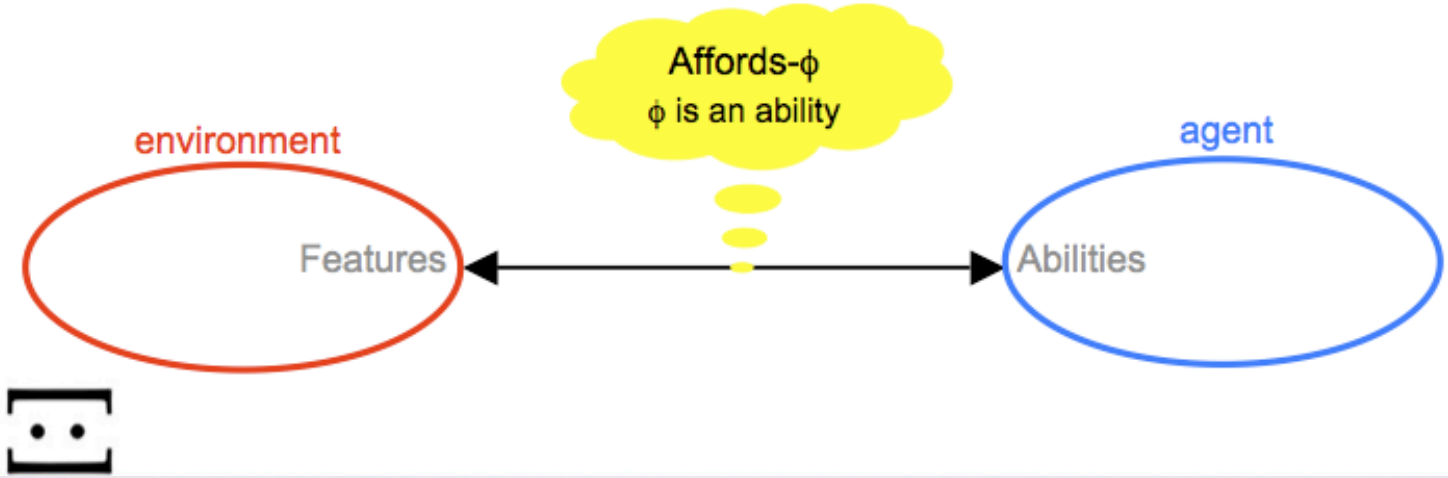


Stoffregen's formalization(2003)

- According to Stoffregen affordances *can not* be defined as properties of the environment only.
- He proposes that,
  - Affordances are properties of the animal-environment system
  - They are emergent properties that do not inhere in either the environment or the animal.

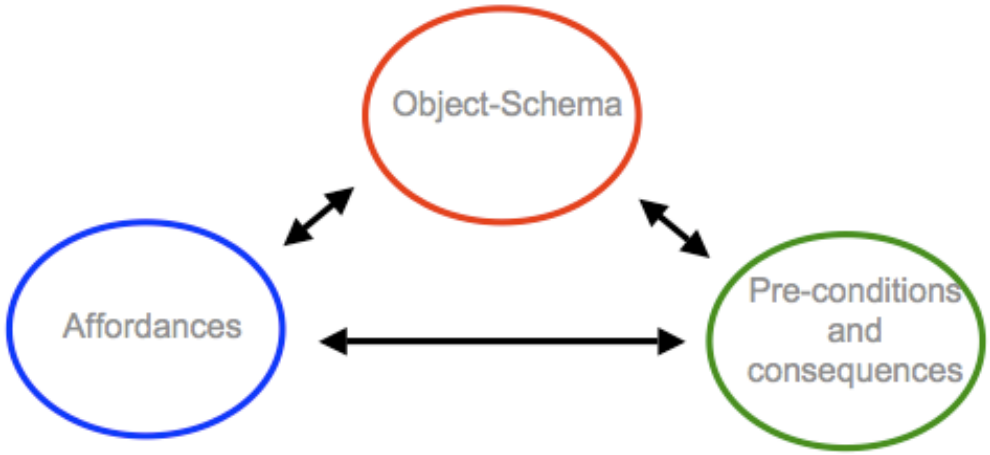
Chemero's formalization (2003)

- Chemero also claims that affordances must be defined at the animal-environment system scale
- He proposed that
  - Affordances are relations between the abilities of organisms and features of the environment.



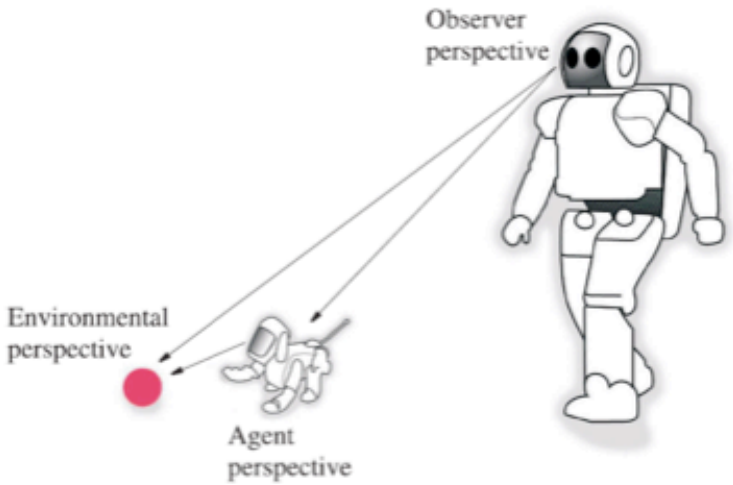

Steedman's formalization (2002)

- Steedman skipped the perceptual aspect of affordances and associated affordances with planning, and linguistic capabilities.
- He claimed that a door is linked with the actions of “pushing” and “going-through”, and the preconditions and consequences of applying these actions to the door.



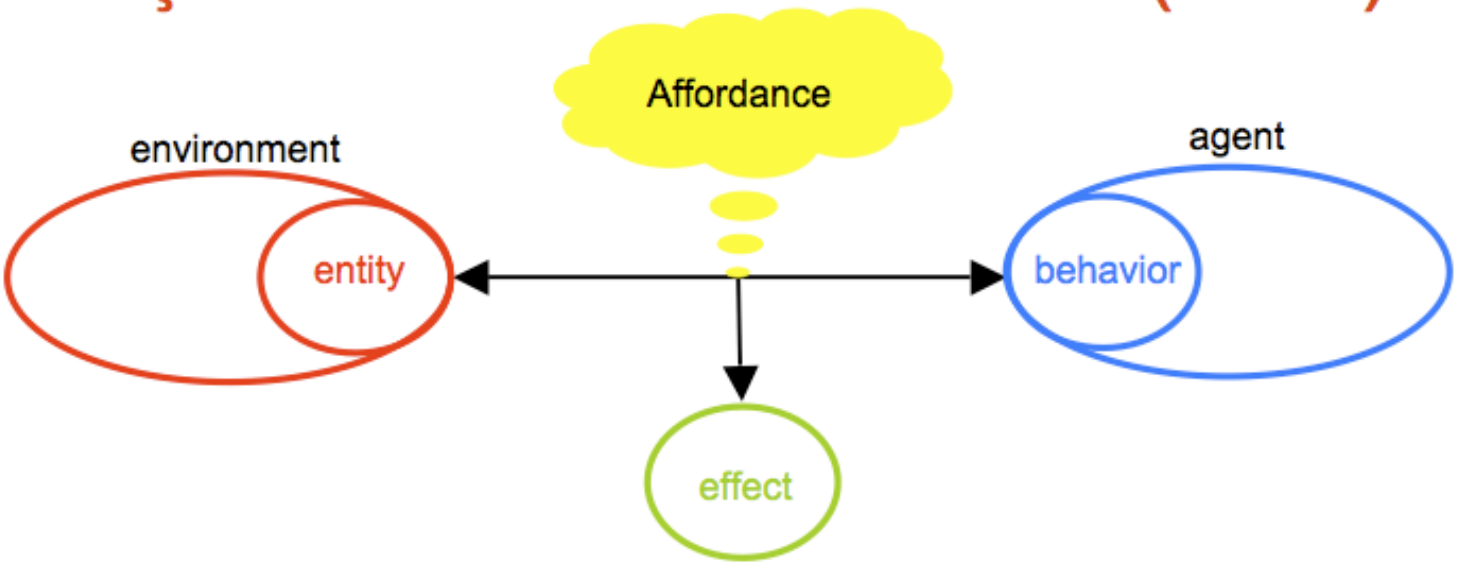
Three perspectives to view affordances

- Affordances are relations and can be viewed from three (not one!) perspectives.
- Agent perspective:
  - I perceive pushability affordance.
- Observer perspective:
  - There is pushability affordance in the dog-ball system.
- Environmental perspective:
  - I offer pushability (to a dog).



(adapted from Erich Rome's slide depicting a similar scene),

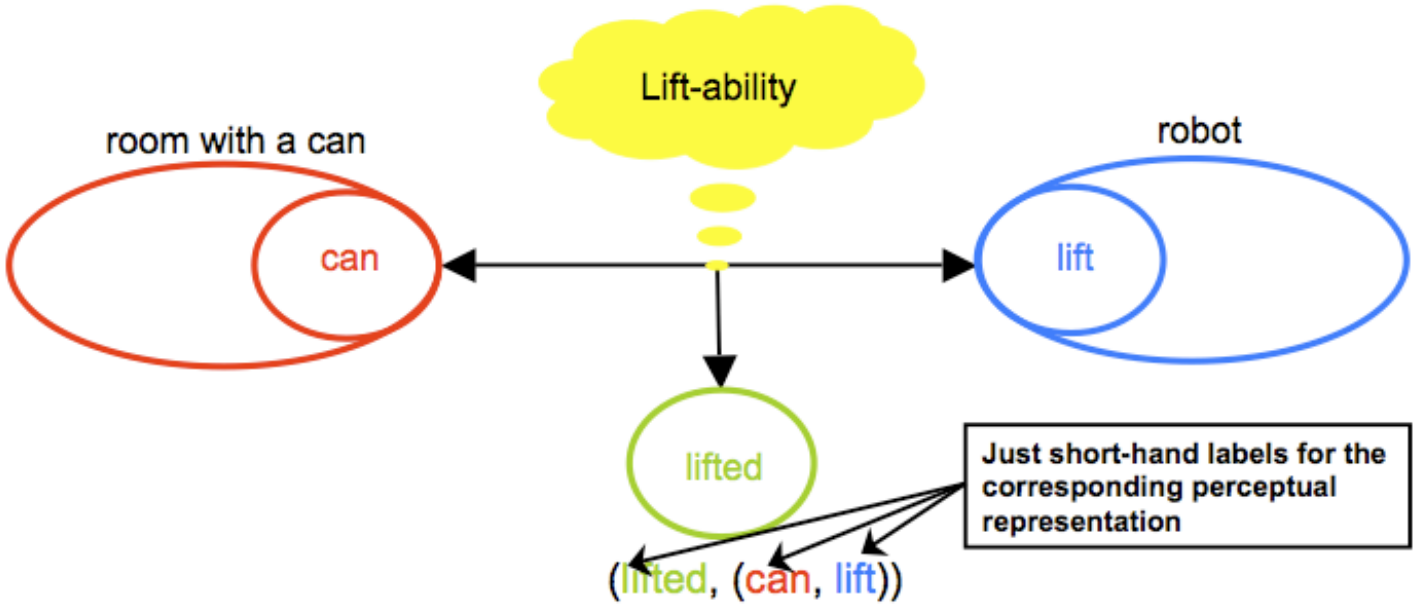
Şahin et al.'s formalization (2007)



(effect, (entity, behavior))

Definition: An affordance is an acquired relation between a behavior of an agent and an entity in the environment such that the application of the behavior on the entity generates a certain effect.

Cont'd



The robot applied its lift behavior on the can and obtained the lifted effect.
Can: The perceptual representation of the can as seen by the robot
Lift: The behavior executed by the robot
Lifted: The effect of the behavior on the environment as perceived by the robot.


Implications for Robot Control

- Affordances can be viewed from three different perspectives: namely, agent, observer and environmental.
  - Although only the agent and observer perspectives are relevant for robotics.
- Affordances (agent and observer) are relations that reside inside the agent.
  - Does not contradict the view that affordances are relation within the agent-environment system.
- Affordances encode “general relations” pertaining to the agent-environment interaction.
  - A relation such as “the-red-ball-on-my-table is not rollable (since it is glued to the table)” does not have any predictive value, and cannot be considered as an affordance.

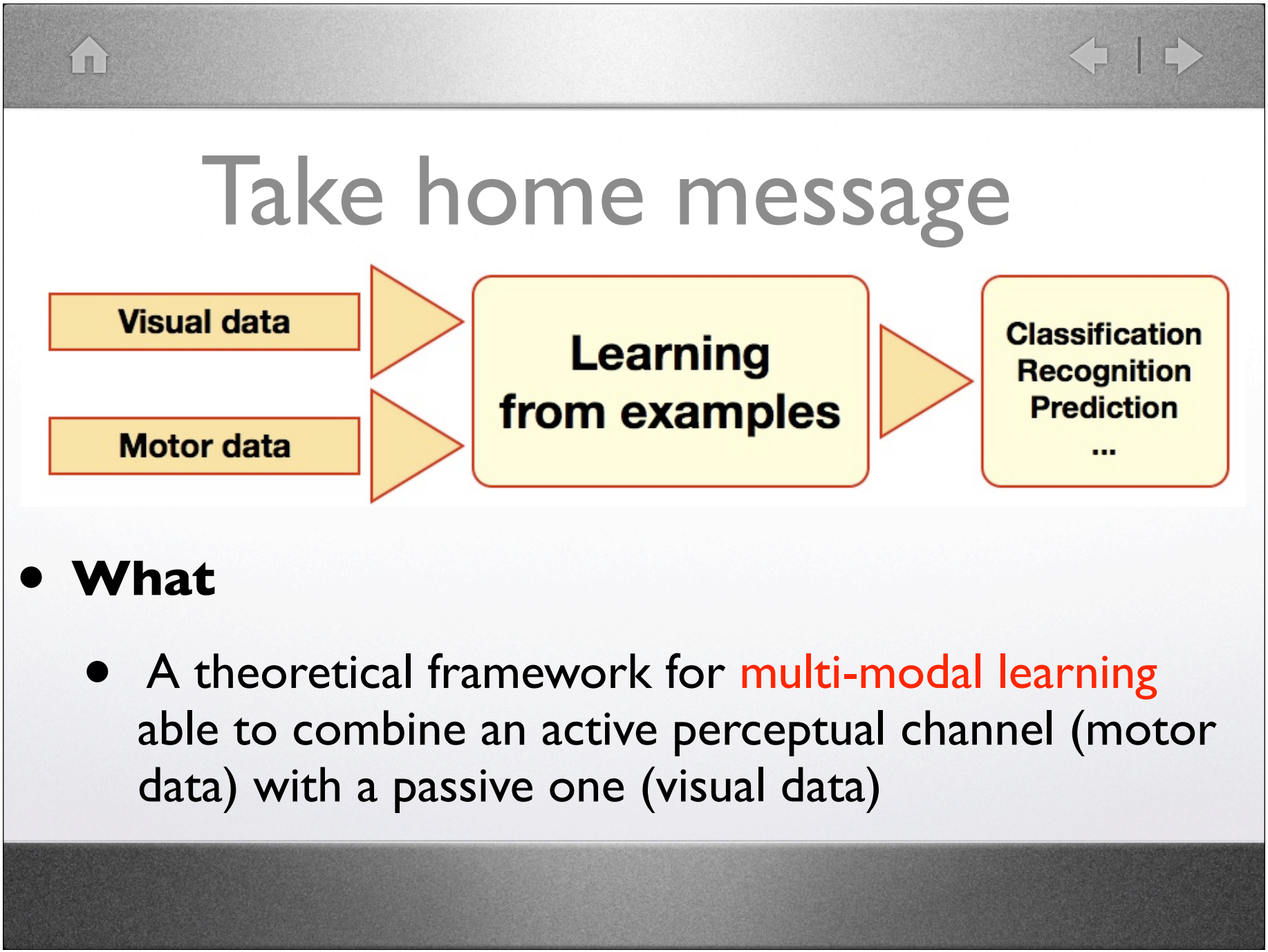
Affordances are acquired relations

- Affordances are acquired through the interaction of the organism with its environment:
  - Acquisition through evolution -> innate affordances (J.Norman; 2001)
  - Acquisition through learning -> learned affordances (E.J. Gibson; 2000)
  - Acquisition through design -> designed affordances (Murphy; 1999)
- Acquired relations are automatically in “body-scaled” metrics.


Affordances provide a framework for symbol formation

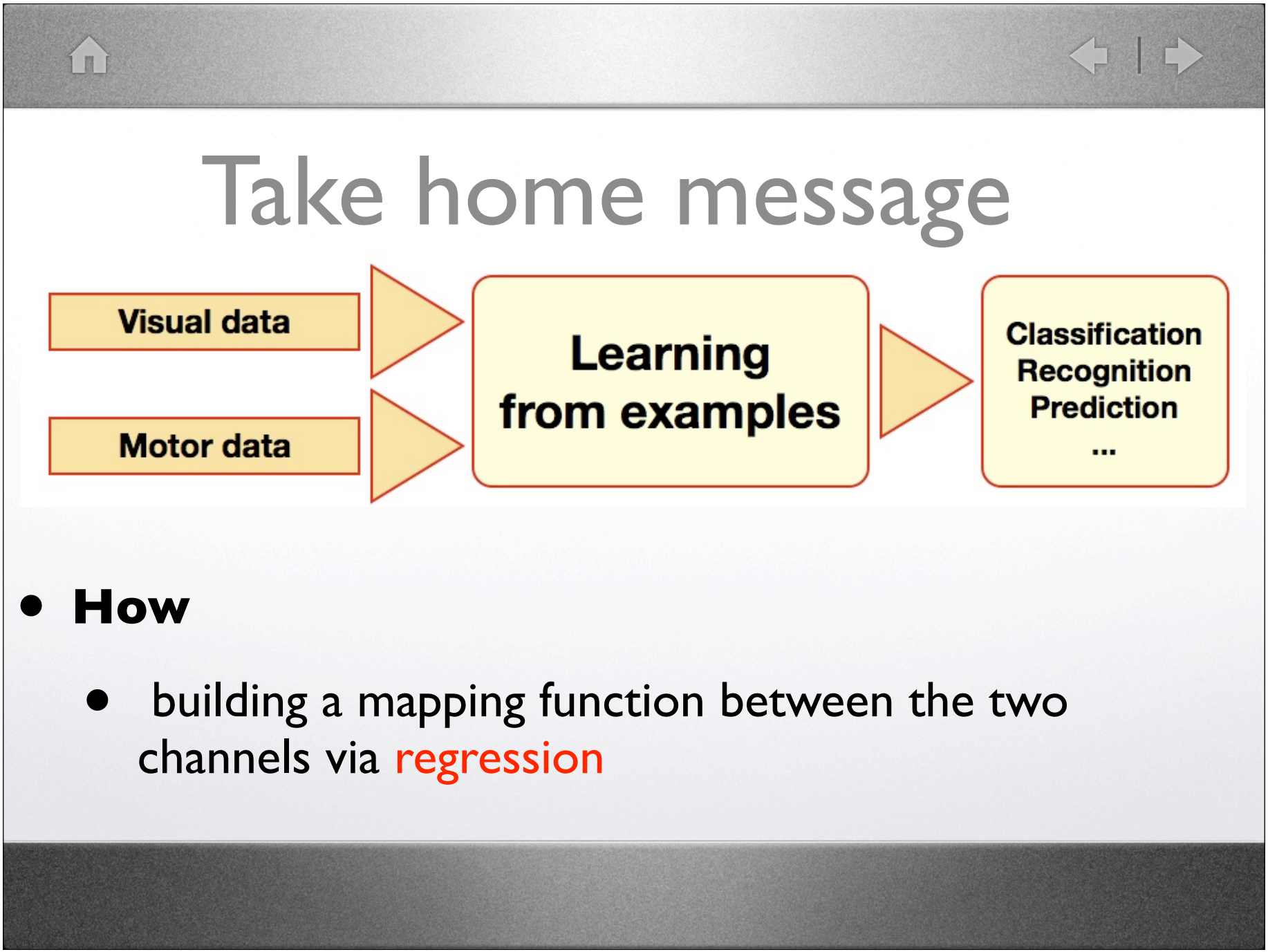
- The problem of how symbols are related to the raw sensory-motor perception of the robot is known as the symbol grounding problem (Harnad; 1990).
- Sun (2000) argued that symbols should be "formed in relation to the experience of agents, through their perceptual/motor apparatuses, in their world and linked to their goals and actions".
- The formation of equivalence classes are triggered by the formation of affordance relations. Hence symbol formation is not an isolated process.


Visuo-Motor Object
Modeling and Recognition



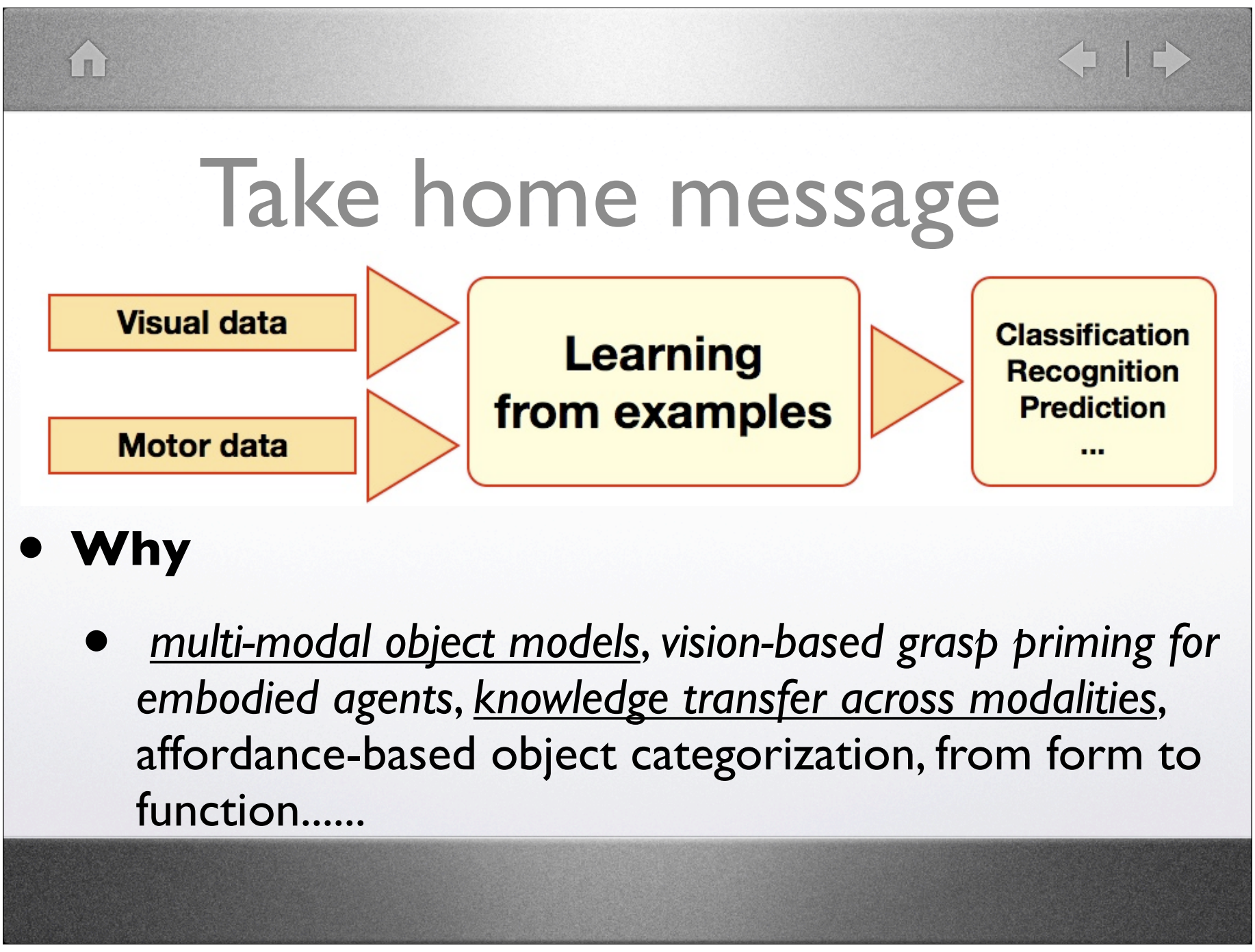
Take home message

- **What**
  - A theoretical framework for multi-modal learning able to combine an active perceptual channel (motor data) with a passive one (visual data)



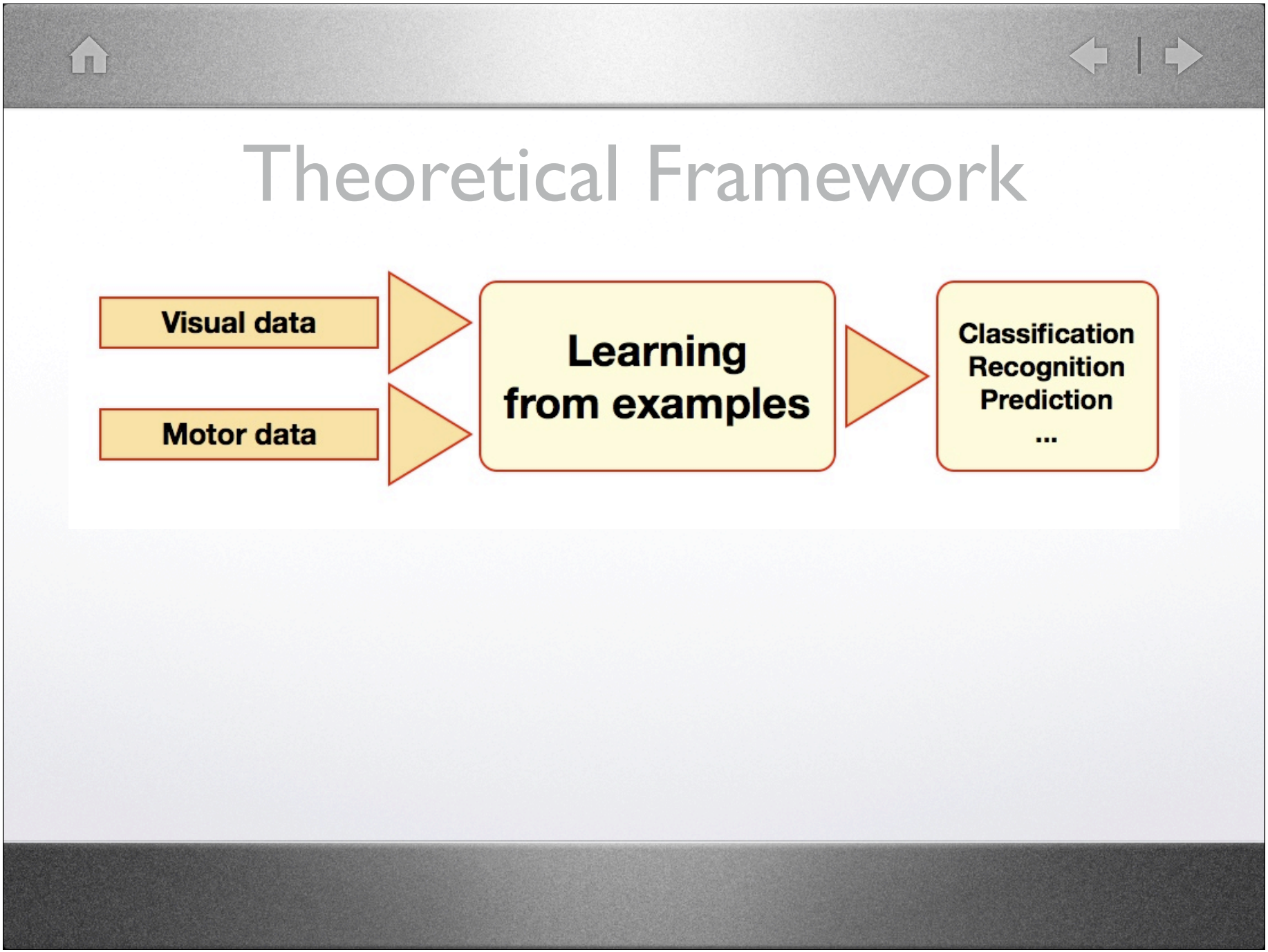
Take home message

- **How**
  - building a mapping function between the two channels via regression

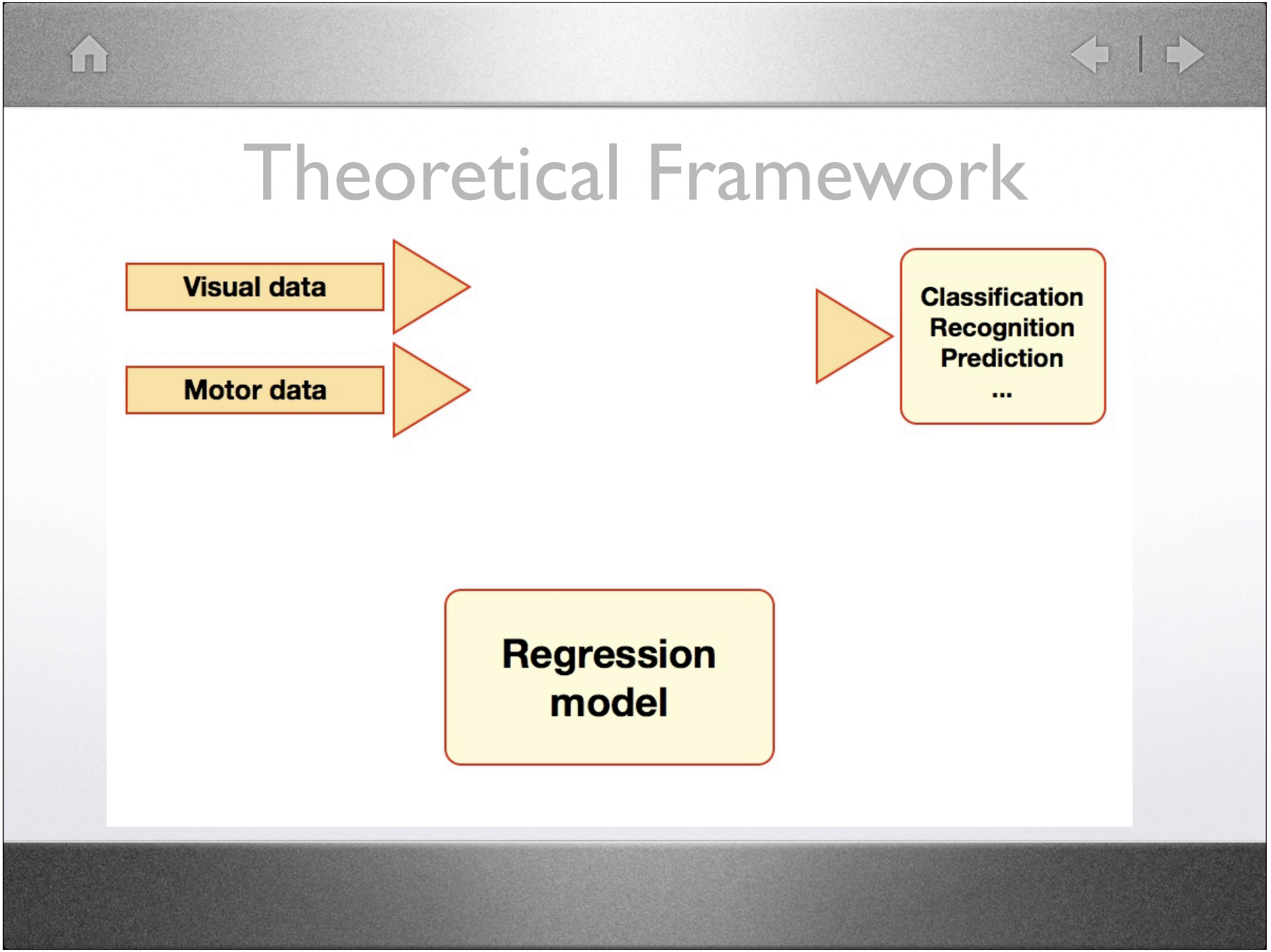

Take home message
Visual data
Motor data
Learning from examples
Classification
Recognition
Prediction
...
• Why
• multi-modal object models, vision-based grasp priming for embodied agents, knowledge transfer across modalities, affordance-based object categorization, from form to function.....


Motivations & Background

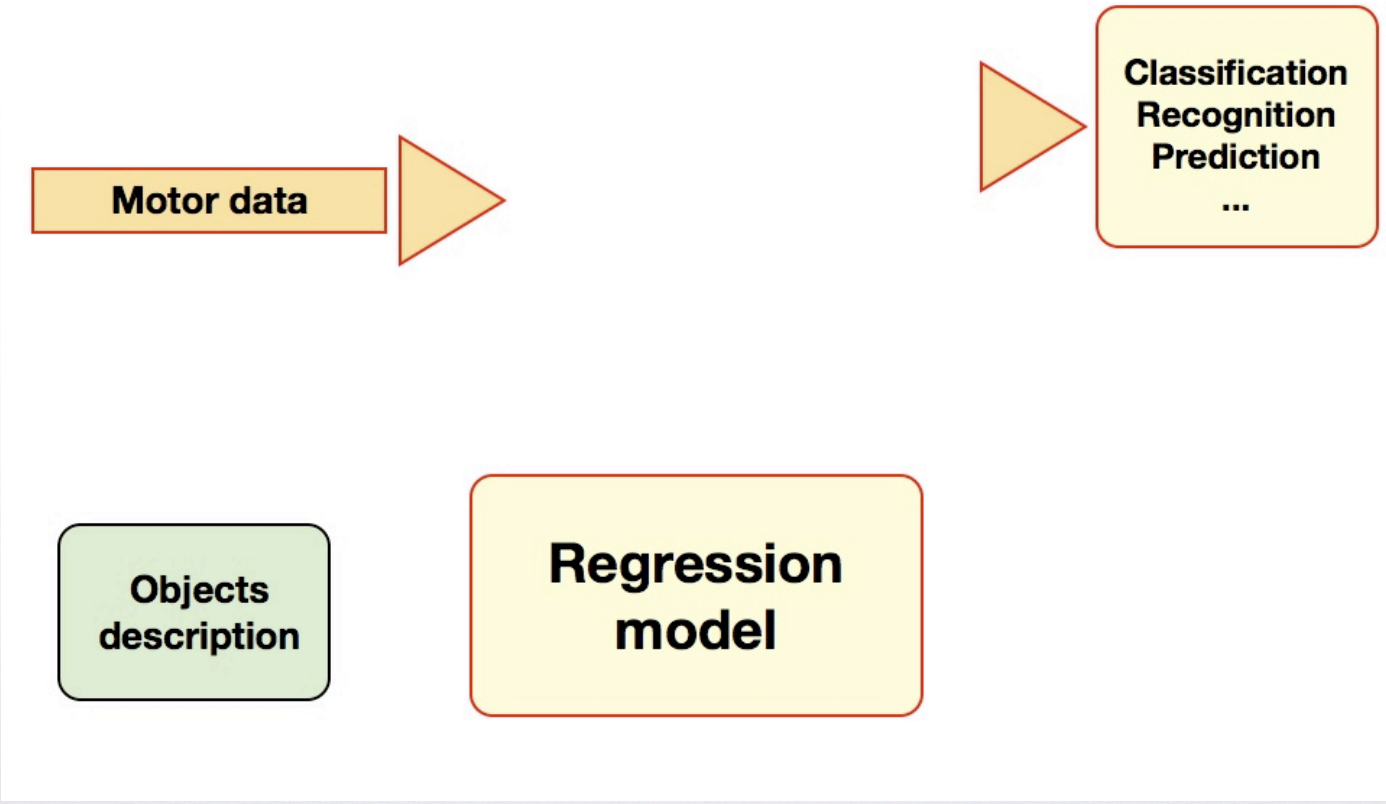
- We draw inspiration from the mirror neurons [Rizzolati96-04]
  - they are clusters of neural cells which fire iif an agent grasps an object, or sees that object, or sees another agent grasping the same object
- We follow a path similar to that laid out in [Metta06, Castellini07] based on a PAM (Perception to Action Map)



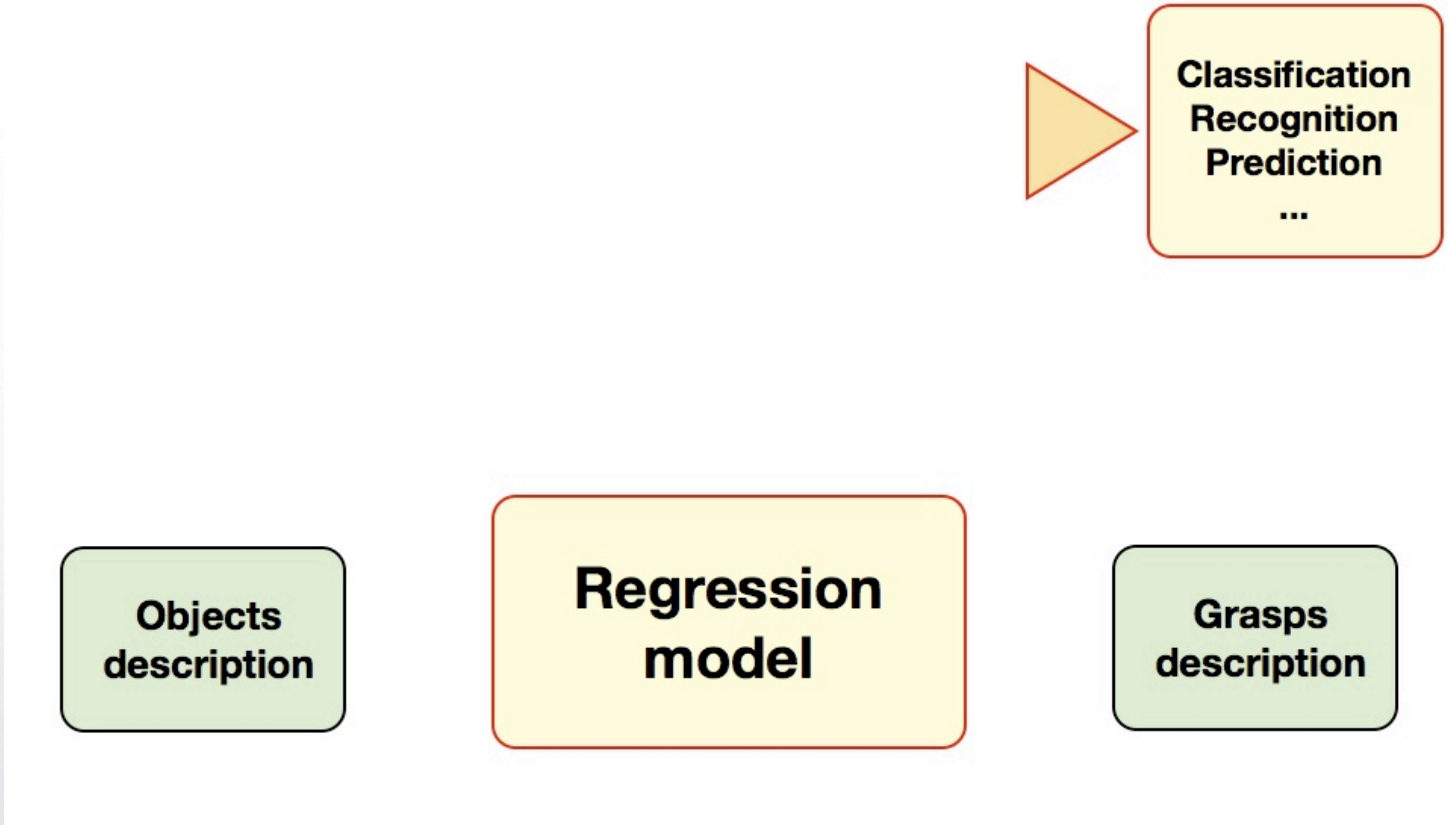
Theoretical Framework


Theoretical Framework
Visual data
Motor data
Classification
Recognition
Prediction
...
Regression
model

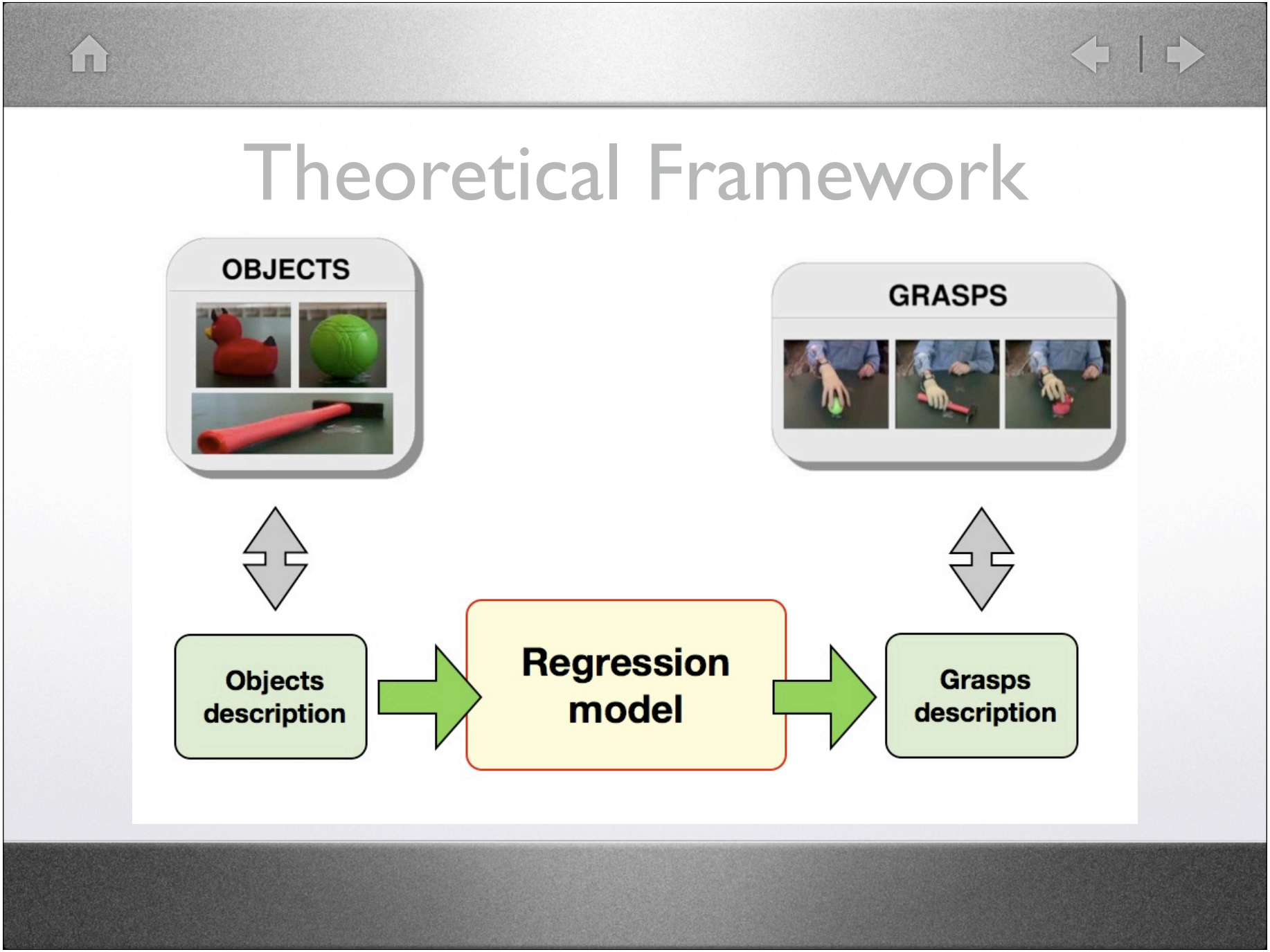
Theoretical Framework

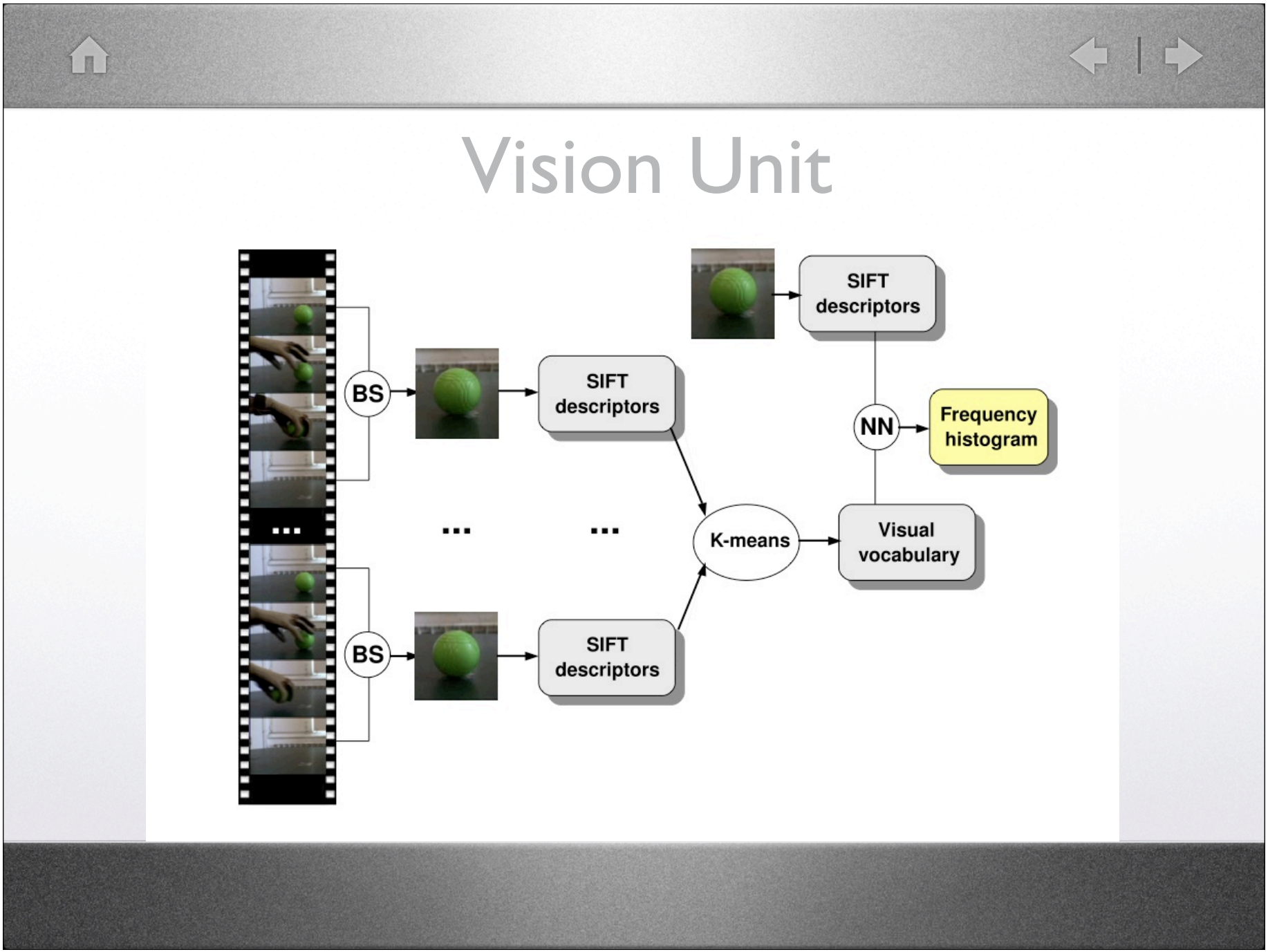



Theoretical Framework




Theoretical Framework


Theoretical Framework
OBJECTS
GRASPS
Objects description
Regression model
Grasps description



Vision Unit

BS
BS
SIFT descriptors
SIFT descriptors
SIFT descriptors
K-means
Visual vocabulary
NN
Frequency histogram


Learning the mapping

The goal is *not* to memorize but to **generalize**, i.e. to predict

Given a set of training data

$$\{(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)\}$$

find a function

$$f(x)\sim y$$

such that ***f* is a good predictor on new data** as well as on the given dataset


Learning the Mapping: Kernel Methods (fancy stuff)

$$\boxed{\mathbf{y}_i = (y_i^1, \ldots, y_i^d)}$$


Learning the Mapping: Kernel Methods (fancy stuff)

$$\mathbf{y}_i = (y_i^1, \ldots, y_i^d)$$

The kernel function defines similarity between input points and correlation among output components

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$


Learning the Mapping: Kernel Methods (fancy stuff)

$$\boxed{\mathbf{y}_i = (y_i^1, \ldots, y_i^d)}$$

The kernel function defines similarity between input points and correlation among output components

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

The estimator is a linear combination of the kernel function evaluated at the training points

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \quad c_i \in \mathbb{R}^T$$


Learning the Mapping: Kernel Methods (fancy stuff)

$$\boxed{\mathbf{y}_i = (y_i^1, \ldots, y_i^d)}$$

The kernel function defines similarity between input points and correlation among output components

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

The estimator is a linear combination of the kernel function evaluated at the training points

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \quad c_i \in \mathbb{R}^T$$

The estimator is found minimizing

$$\mathcal{E}[f] + \lambda ||f||_{\Gamma}^2$$


Learning the Mapping: Kernel Methods (fancy stuff)

$$\boxed{\mathbf{y}_i = (y_i^1, \ldots, y_i^d)}$$

The kernel function defines similarity between input points and correlation among output components

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d \times d}$$

The estimator is a linear combination of the kernel function evaluated at the training points

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \quad c_i \in \mathbb{R}^T$$

The estimator is found minimizing

$$\mathcal{E}[f] + \lambda ||f||_{\Gamma}^2$$

Empirical Risk
Error on the training data

$$\mathcal{E}_n[f] = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{y}_i - f(\mathbf{x}_i)||_d^2$$


Learning the Mapping: Kernel Methods (fancy stuff)

$$\boxed{\mathbf{y}_i = (y_i^1, \ldots, y_i^d)}$$

The kernel function defines similarity between input points and correlation among output components

$$\Gamma : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^{d \times d}$$

The estimator is a linear combination of the kernel function evaluated at the training points

$$f_{\mathbf{z}}^{\lambda}(x) = \sum_{i=1}^{n} \Gamma(x, x_i) c_i, \quad c_i \in \mathbb{R}^T$$

The estimator is found minimizing

$$\mathcal{E}[f] + \lambda ||f||_{\Gamma}^2$$

Smoothness term

Empirical Risk
Error on the training data

$$\mathcal{E}_n[f] = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{y}_i - f(\mathbf{x}_i)||_d^2$$


Learning the Mapping: Kernel Methods (fancy stuff)

- **Proof-of concept result:** assuming a 1-1 mapping between object and grasp posture, the 22-valued motor descriptors can be estimated accurately from the visual feature of the corresponding object

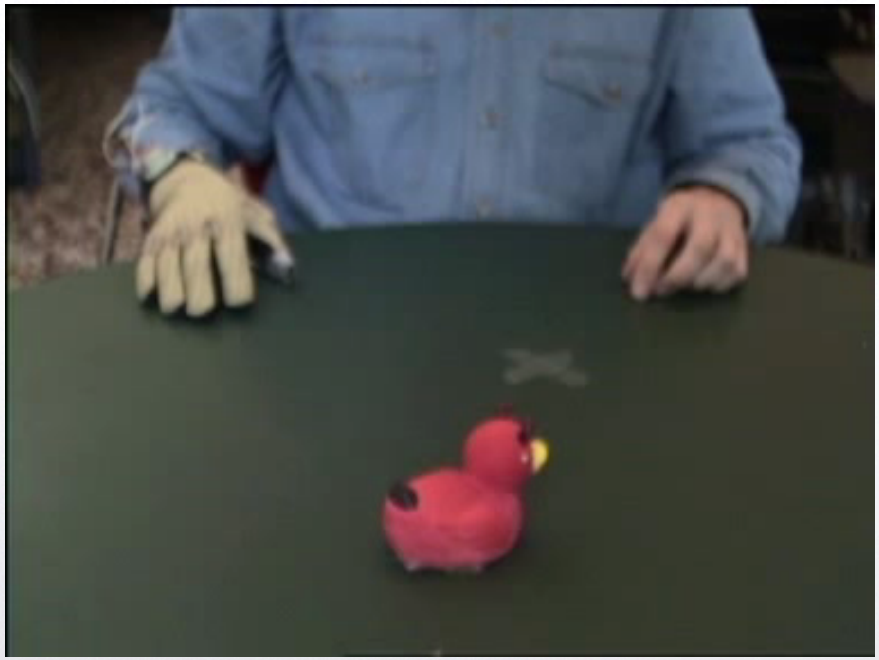
N. Noceti, B. Caputo, C. Castellini, L. Baldassarre, A. Barla, L. Rosasco, F. Odone, G. Sandini. *Towards a theoretical framework for learning multimodal patterns for embodied agents*. Proc ICIAP 2009

- **Extension to many-to-many:** .....mmmmmmm.....

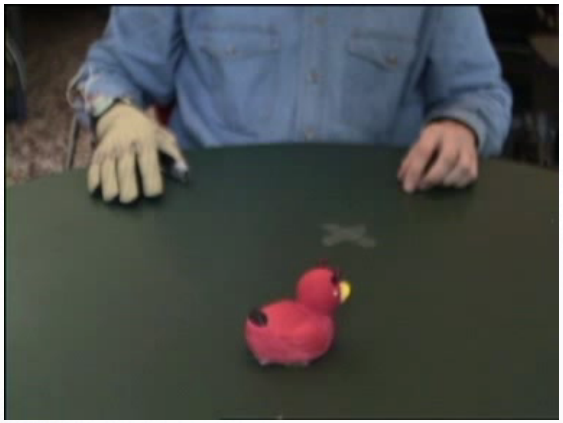

Learning the Mapping: ANN (not so fancy stuff)

- 200 inputs (dim of visual features), 22 output (dim of grasp posture descriptor
- 1 hidden layer (20 neurons, log-sigmoid transfer function, gradient backpropagation)
- instead of modeling a many-to-many correspondence, we define an archetypal grasp for each object, i.e. a mixture of the possible grasps
- amazingly it works!

The CONTACT Visuo-Motor Grasping DataBase

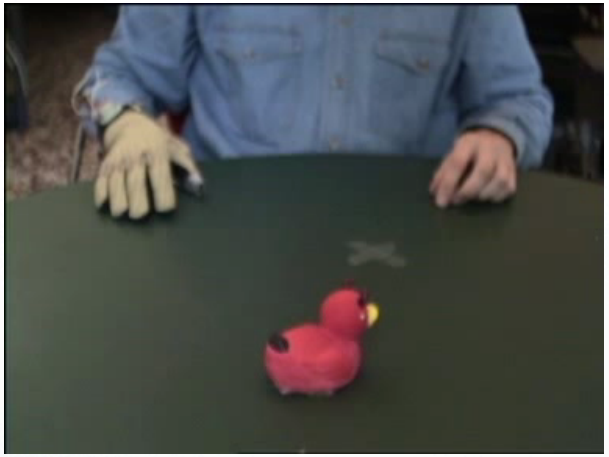


The CONTACT VMG DataBase



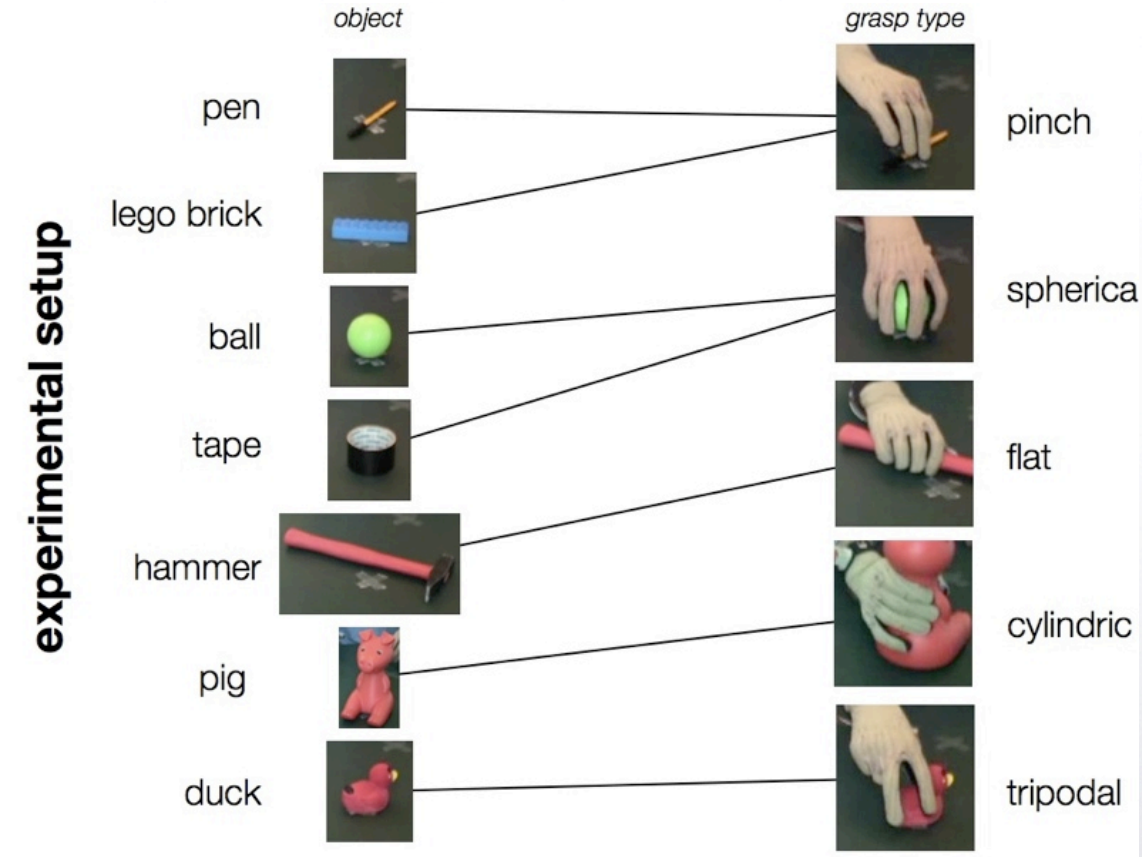
- Immersion *CyberGlove* with 22-sensors (**hand posture**), ascension *Flock-Of-Birds* magnetic tracker on the wrist (**position and speed**) and a force sensing resistor glued to the thumb (**instant of contact**)

The CONTACT VMG DataBase



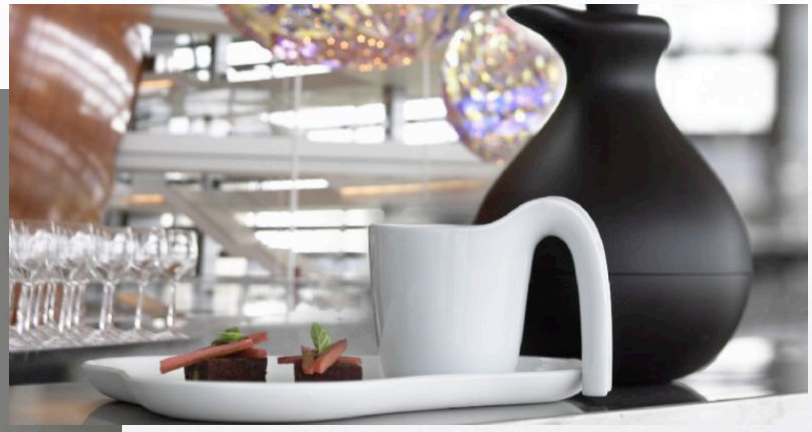
- 20 human subjects, 7 objects, 5 grasp types
- each subject asked to repeat the same grasp type 20 times
- data recorded with two cameras (visual data) and CyberGlove, for a total of 5200 grasping acts

The CONTACT VMG DataBase



App 1: Multi-Modal Object Recognition

- Goal: to augment visual information about an object with motor information about it, i.e. the way the object can be grasped by a human being

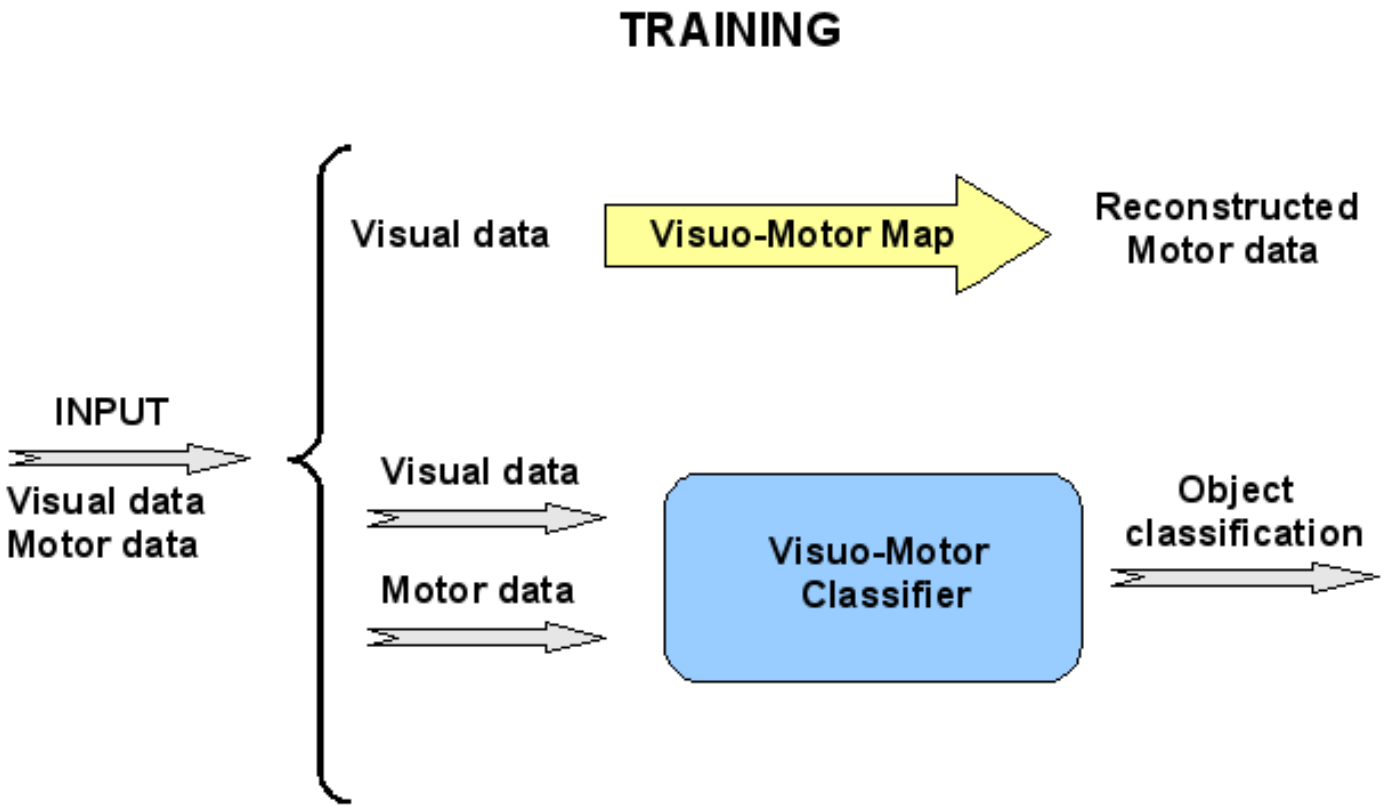





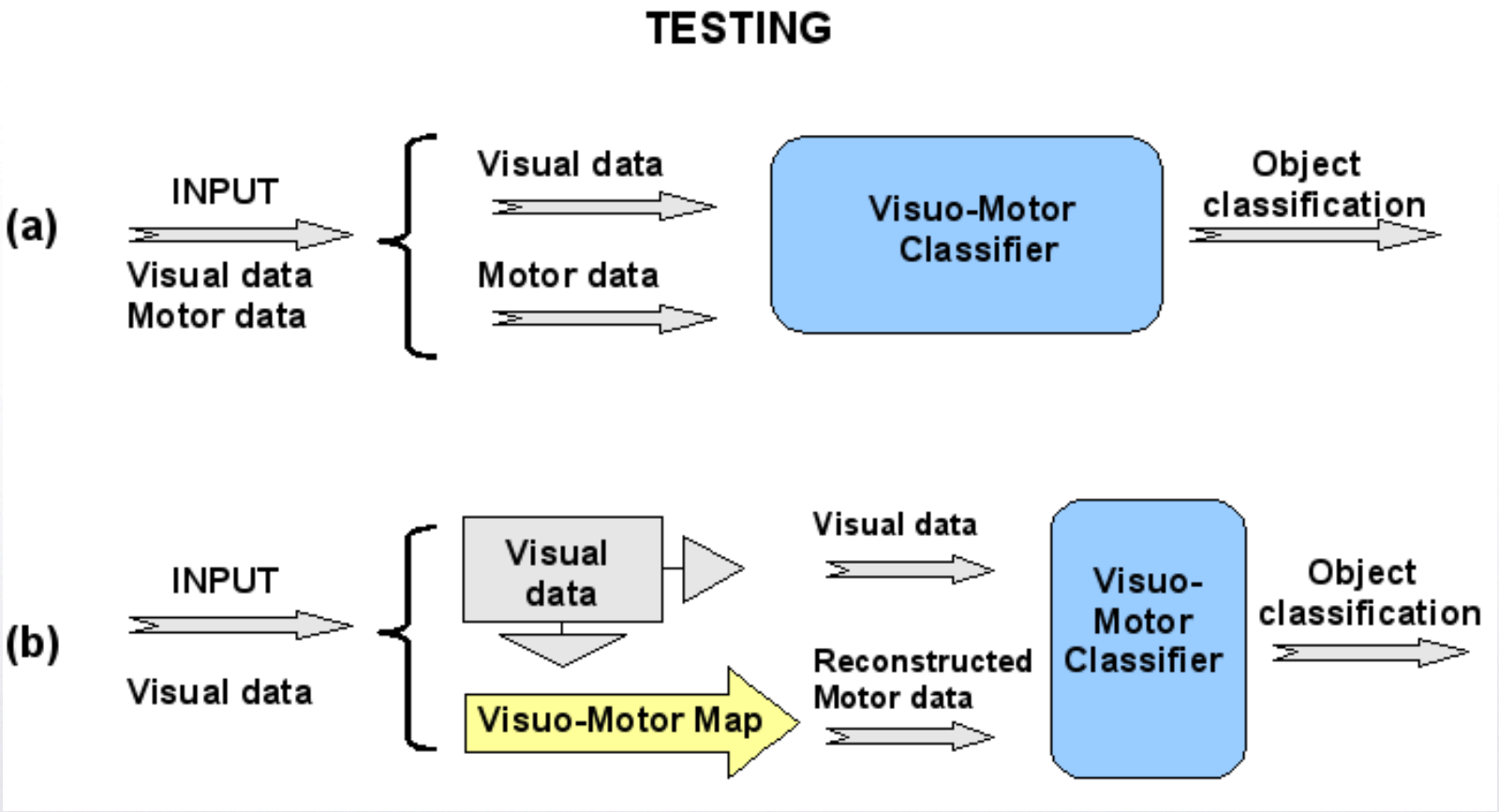
App 1: Multi-Modal Object Recognition

- We build an object recognition system on a set of visual and motor features
- Whenever the motor features are not perceived by the system (i.e. the agent is not grasping/manipulating the object in the field of view) we infer them from the visual input
- Motor features are derived from perceived visual features through the mapping function learned during training

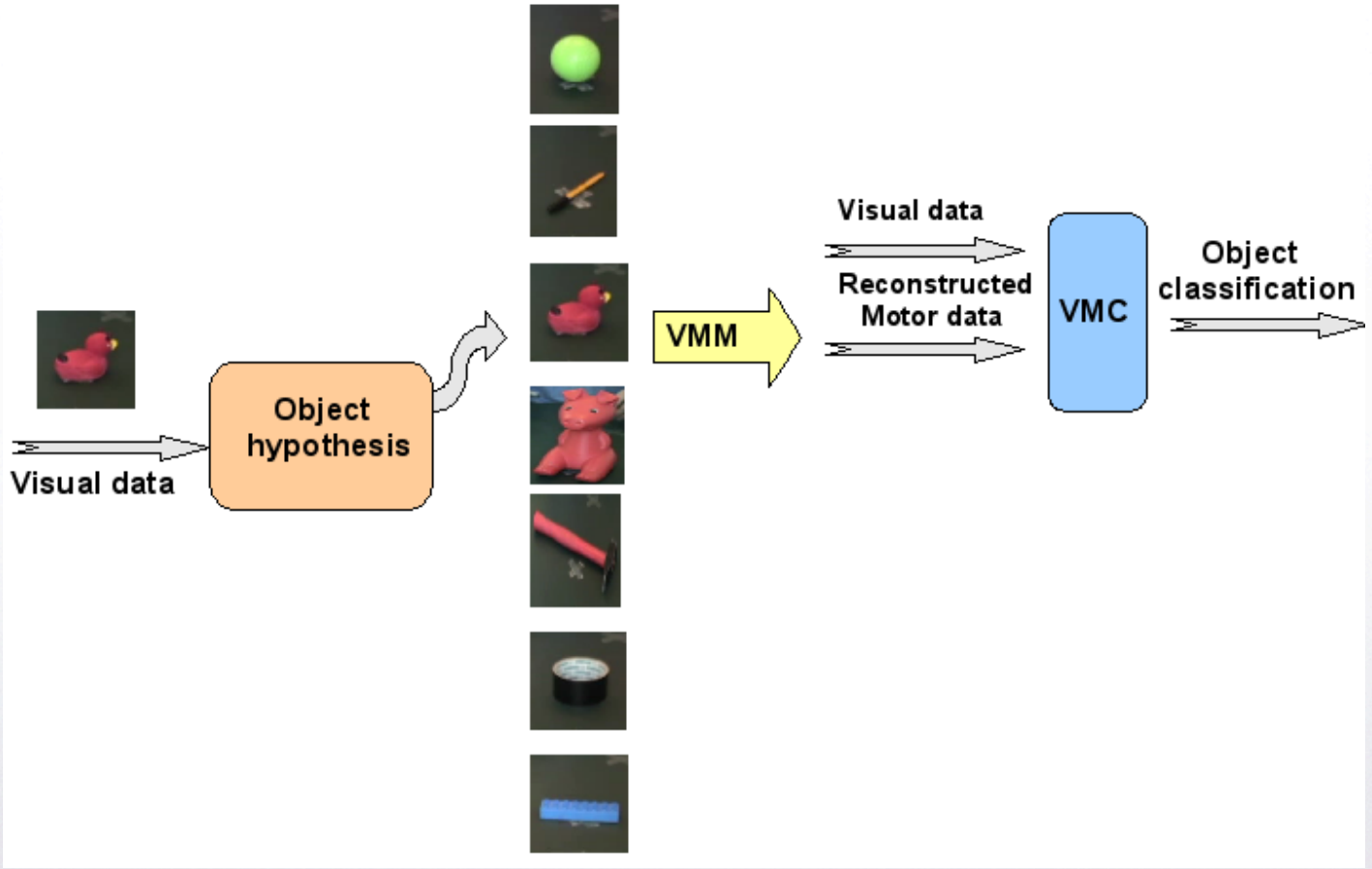

App 1: Multi-Modal Object Recognition




App 1: Multi-Modal Object Recognition

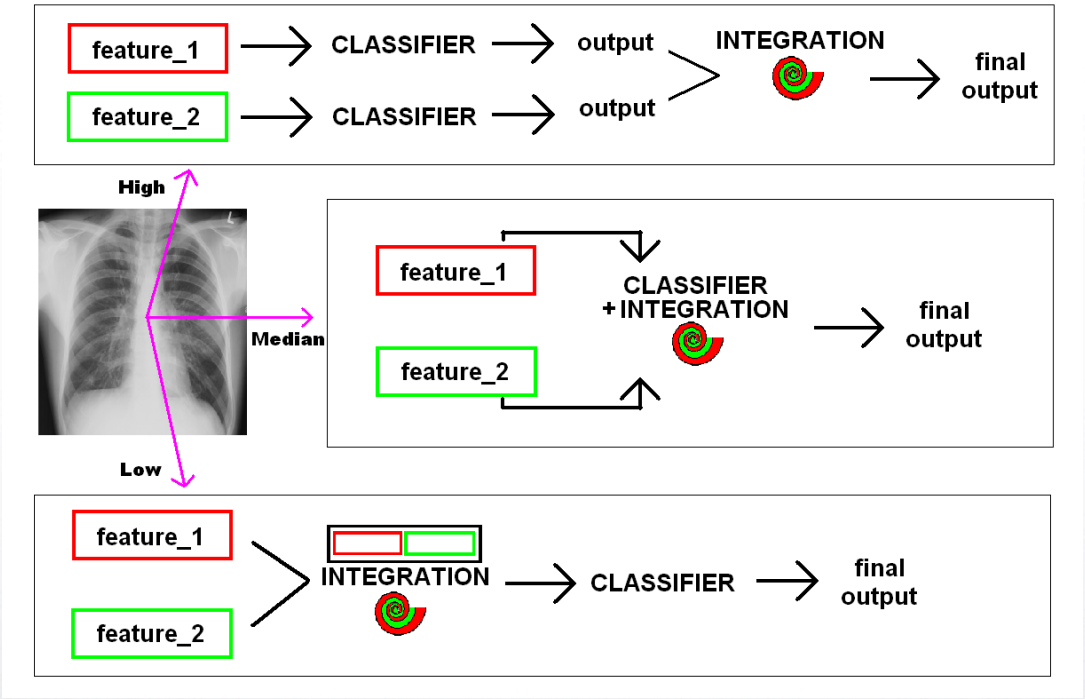



App 1: Multi-Modal Object Recognition

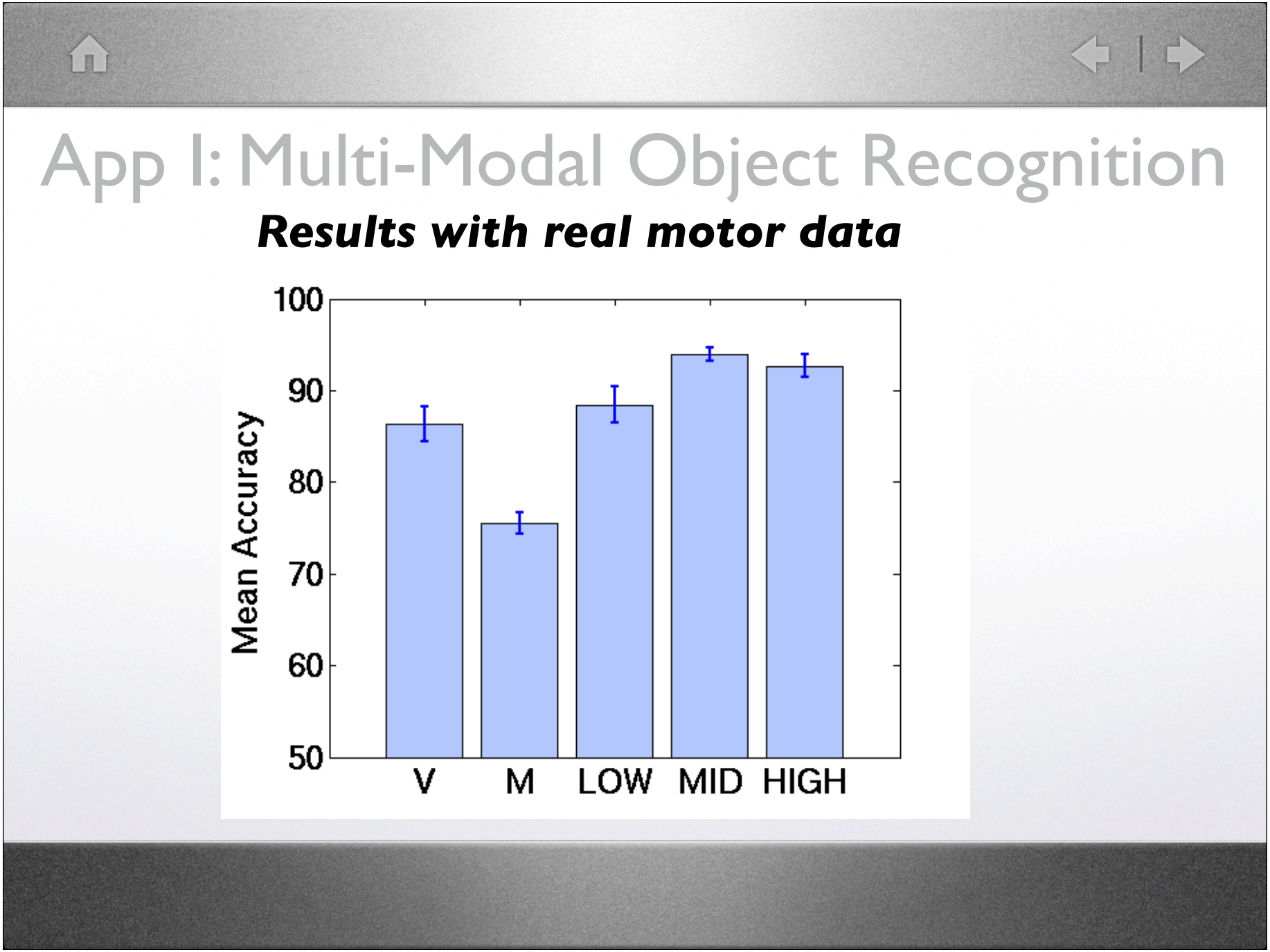


App I: Multi-Modal Object Recognition

- Visuo-motor classifier: low-level, mid-level or high-level?

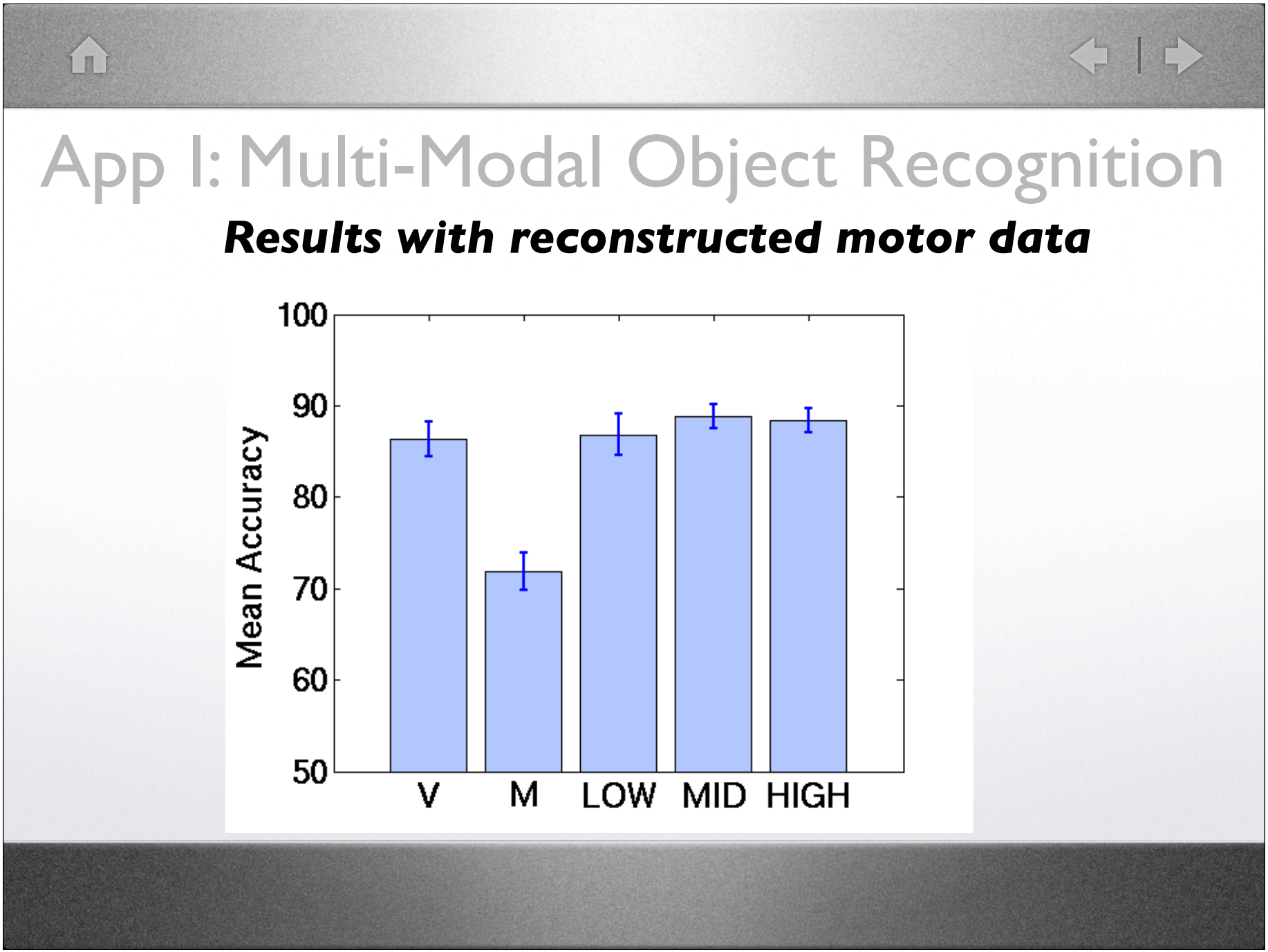


T. Tommasi, F. Orabona, B. Caputo. *Discriminative cue integration for medical image annotation.* Pattern Recognition Letters, 2008



App I: Multi-Modal Object Recognition

Results with real motor data



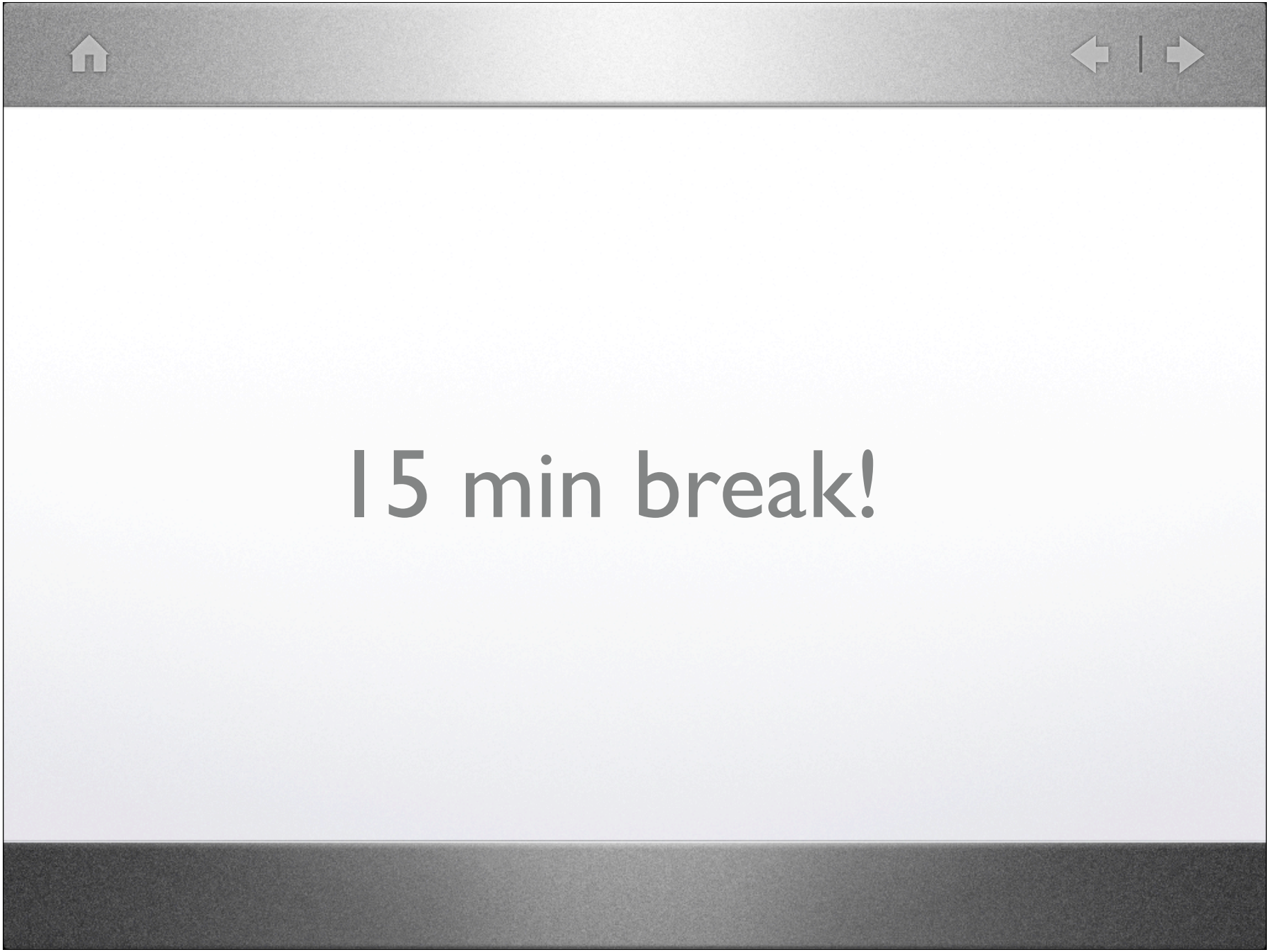
App 1: Multi-Modal Object Recognition

Results with reconstructed motor data

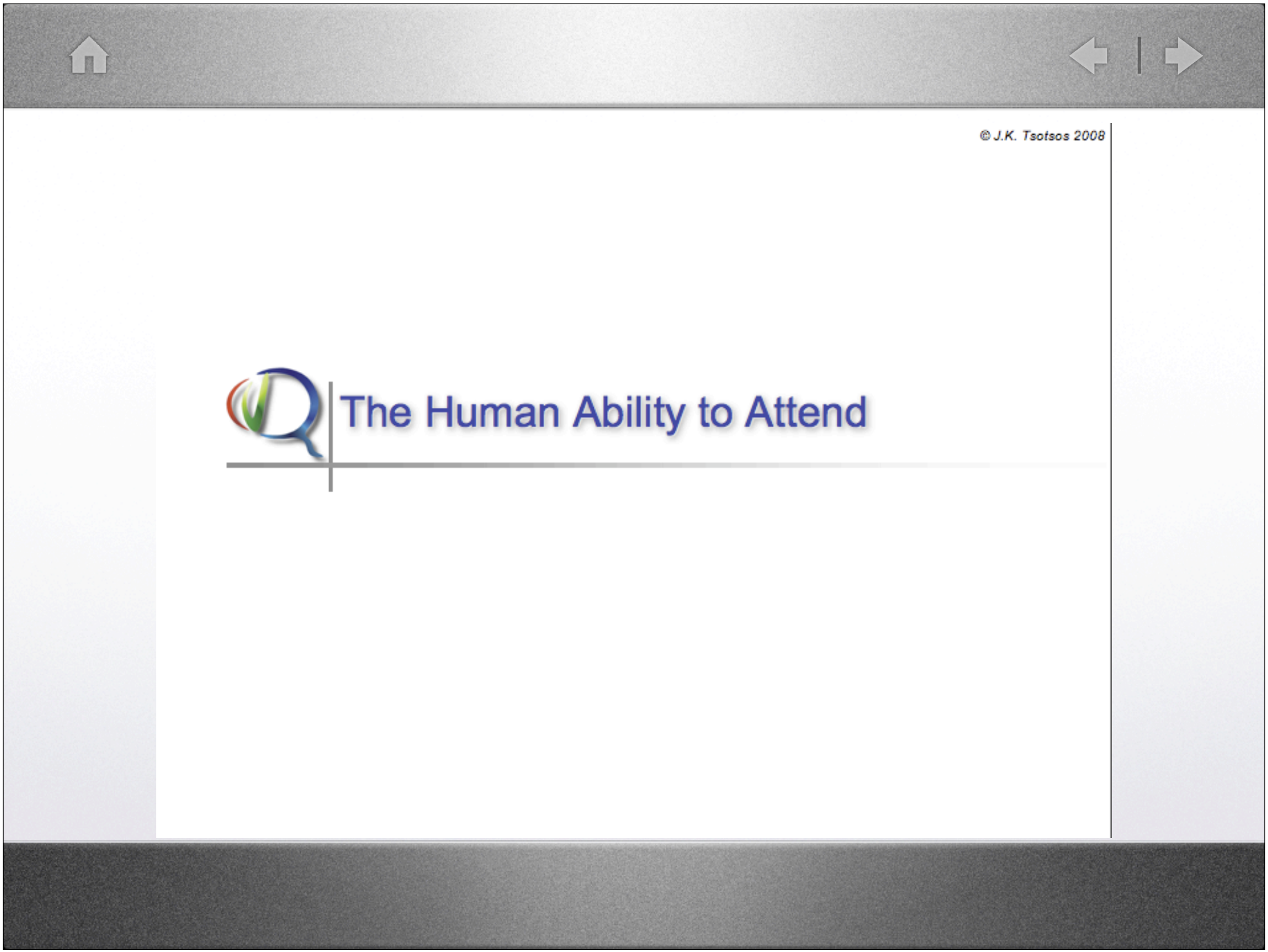

App 1: Multi-Modal Object Recognition

It Works!

- *Needs to be improved:* **visual features, mapping function, motor representation**
- *Needs to be added:* **dynamic of the grasp, reaction of the object, task, longer/more complex actions, not only manipulable objects**, ...(fill the dots as you wish)
- *Categorization and scaling:* not for google vision, but for robot vision perhaps....


15 min break!


Attention


© J.K. Tsotsos 2008
The Human Ability to Attend


Attention: Its Roots...

Attention : from the Latin “attenti”, from attentus, the past participle of attendere, meaning ‘to heed”

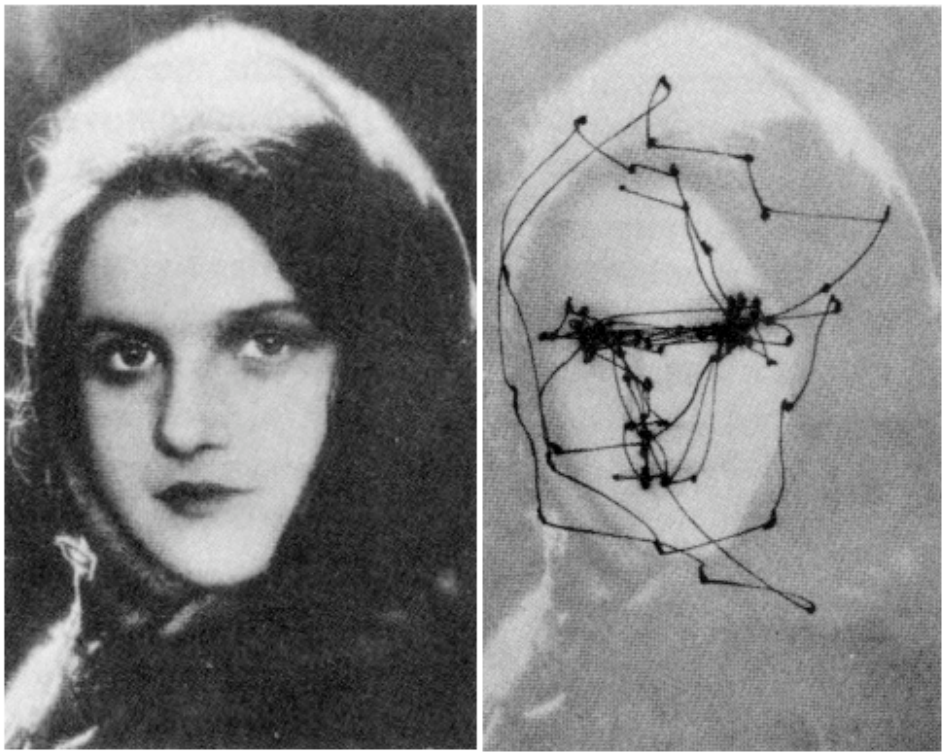
Descartes (1649): Thus when one wishes to arrest one’s attention so as to consider one object for a certain length of time, this volition keeps the pineal gland titled towards one side during that time.

Hobbes (1655): While the sense organs are occupied with one object, they cannot be simultaneously be moved by another so that an image of both arises. There cannot therefore be two images of two objects but one put together from the action of both.

Malebranche (1674): Attention is necessary for conserving evidence in our knowledge.

Leibnitz (1765): In order for the mind to become conscious of perceive objects, and therefore for the act of apperception, attention is required.

Typical Scanpath



regardless of the claims of biological plausibility or realism, none of the attention models can replicate such scanpaths

Scanpaths and Object Representation

Noton, D., Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception, *Science 171(3968)*, p308-311.

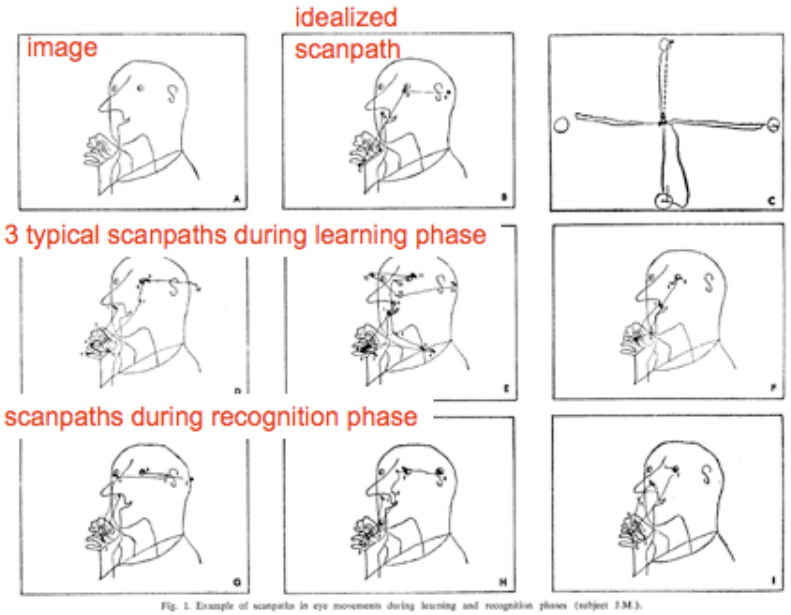
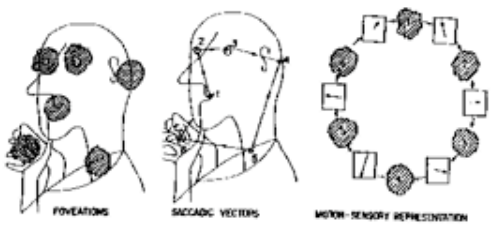


Fig. 1. Example of scanpaths in eye movements during learning and recognition phases (subject J.M.).

Reported a connection between the eye movement patterns observed during learning of a visual pattern and the subsequent viewing of that pattern.

During learning, subjects followed a characteristic scanpath. When later presented with the pattern again, subjects usually followed a very similar scanpath for at least the first few fixations.

This suggested that the internal representation of a pattern in memory is a network of features, and thus attention shifts move from feature to feature.



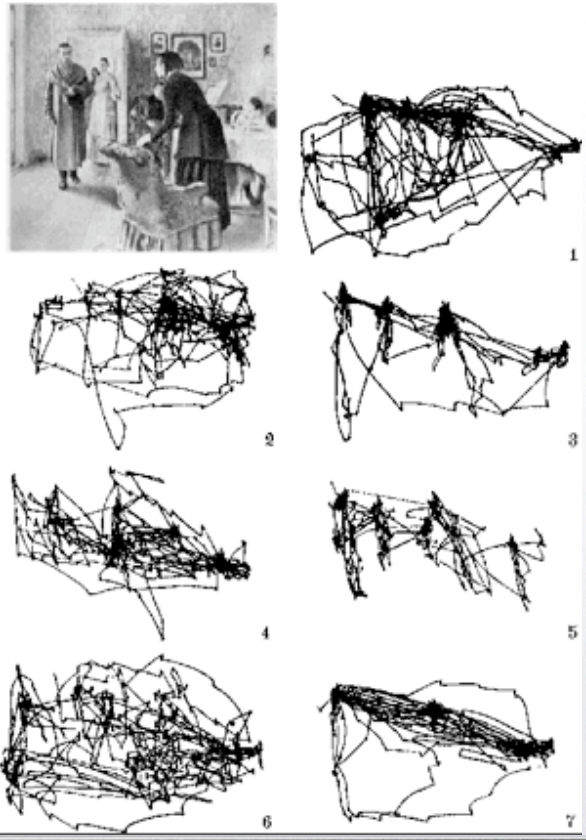
Task and Eye Movements

Yarbus, A. L. (1967). **Eye Movements and Vision**. New York: Plenum.

Yarbus demonstrated how eye movements changed depending on the question asked of the subject:

1. No question asked
2. Judge economic status
3. "What were they doing before the visitor arrived?"
4. "What clothes are they wearing?"
5. "Where are they?"
6. "How long is it since the visitor has seen the family?"
7. Estimate how long the "unexpected visitor" had been away from the family

Each recording lasted 3 minutes




Salience

What's a Feature? What Attracts Attention?
Master Map of Locations (Treisman 1985)
Saliency Map (Koch & Ullman 1985, Itti & Koch 2001)
Activation Map (Wolfe, Cave & Franzel 1989)
Priority Map (Fecteau & Munoz 2006)

What's a Feature?
What Attracts Attention?

for a nice summary, see Wolfe, J. (1998). Visual Search, in **Attention** (ed. Pashler, H.), 13–74, University College London, London.

Just about everything someone may have studied can be considered a feature or can capture attention

Wolfe presents the kinds of features that humans can detect 'efficiently':

- Color
- Orientation
- Curvature
- Texture
- Scale
- Vernier Offset
- Size, Spatial Frequency, and Scale
- Motion
- Shape
- Onset/Offset
- Pictorial Depth Cues
- Stereoscopic Depth

For most, subjects can 'select' feature or feature values to attend in advance

Saliency Map

Koch, C., Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry, *Human Neurobiology 4*, 219–227

Saliency map - a topographic representation that combines the information from the individual feature maps into one global measure of conspicuity

Point-wise mapping from one map to the other

Can be modulated by higher cortical centres

Master Map of Locations

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing, 31*, 156-177.

Attention selects one area at a time, within a master map of locations thereby retrieving the features linked to the corresponding locations in a number of separable feature maps

Activation Map

Wolfe, J., Cave, K., Franzel, S. (1989). Guided search: An alternative to the feature integration model for visual search, *J. Exp. Psychology: Human Perception and Performance 15*, 419-433.

A topographic representation of the weighted sums of feature map activations.
Feature map activations are based on local differences and task demands

What is Attention?

Attention is the set of mechanisms that optimize/control the search processes inherent in vision

- **select** spatial region of interest
  - temporal window of interest
  - world/task/object/event model
  - gaze/viewpoint
  - best interpretation/response
- **restrict** task relevant search space pruning
  - location cues
  - fixation points
  - search depth control
- **suppress** spatial/feature surround inhibition
  - inhibition of return


Points/Regions of Interest Detection

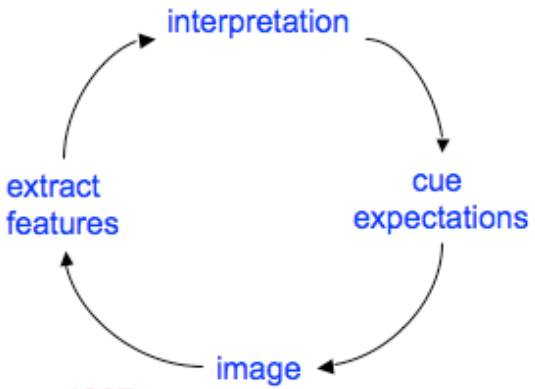
Used in image/object recognition to provide invariant descriptions of important features and in indexing to "summarize" images for fast querying

definition: a point in an image is interesting if it has two main properties: distinctiveness and invariance. This means that a point should be distinguishable from its immediate neighbors and the position as well and the selection of the interesting point should be invariant with respect to the expected geometric and radiometric distortions.

Moravec, H. *Rover visual obstacle avoidance*, IJCAI, Vancouver, BC, pp. 785-790, 1981

The classic interest point detector

Predictive Methods



Neisser 1967
Mackworth 1978

focuses system resources on image regions where analysis might be most profitable

U. Neisser, (1967). **Cognitive Psychology** Appleton-Century-Crofts New York

M. Kelly, Edge detection in pictures by computer using planning, *Machine Intell.* 6, 397-409 (1971).

Y. Shirai, A context-sensitive line finder for recognition of polyhedra, *Artif. Intell.* 4(2), 95-119 (1973).

E. Freuder, A Computer System for Visual Recognition Using Active Knowledge, *Proc. of the Fifth IJCAI*, Cambridge, MA, pp. 671-677, (1977).

A. Mackworth, Vision Research Strategy: Black Magic, Metaphors, Mechanisms, Miniworlds, and Maps, in A. Hanson and E. Riseman (eds.), *Computer Vision Systems*, Academic Press, New York, pp. 53-60, (1978).

J. Tsotsos, J. Mylopoulos, H. Covvey and S. Zucker, "A framework for visual motion understanding," *IEEE Patt. Anal. Machine Intell.* 2, 563-573 (1980).

C.Brown, Gaze Controls Cooperating Through Prediction, J. Image and Vision Computing,Vol. 8 #1, (1990) pp10 - 19.


Active Vision

In 1985 Ruzena Bajcsy wrote:

"Active sensing is the problem of intelligent control strategies applied to the data acquisition process which will depend on the current state of data interpretation including recognition."


Why Active?

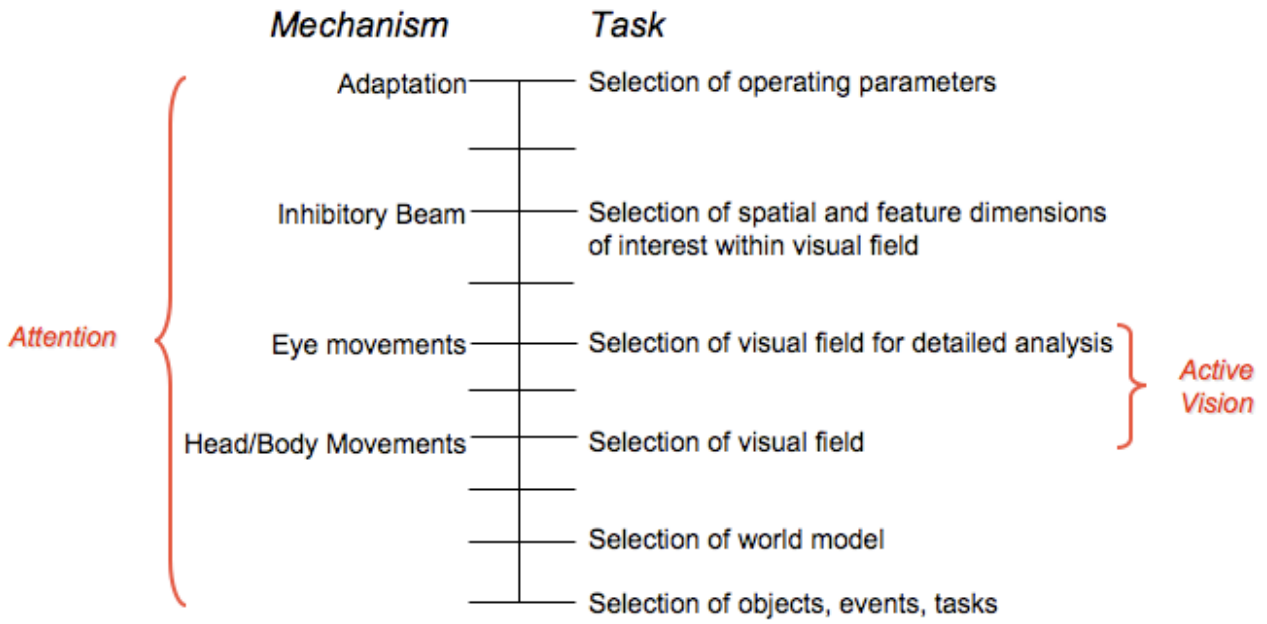
- to move to fixation point/plane or to track motion
- to see a portion of the visual field otherwise hidden due to occlusion
  - manipulation
  - viewpoint change
- to see a larger portion of the surrounding visual world
  - exploration
- to compensate for spatial non-uniformity of a processing mechanism
  - foveation
- to increase spatial resolution or to focus
  - sensor zoom or observer motion
  - adjust camera depth of field, stereo vergence
- to disambiguate or to eliminate degenerate views
  - induced motion (kinetic depth)
  - lighting changes (photometric stereo)
  - viewpoint change
- to achieve a "pathognomonic" view
  - viewpoint change
- to complete a task
  - multiple fixations

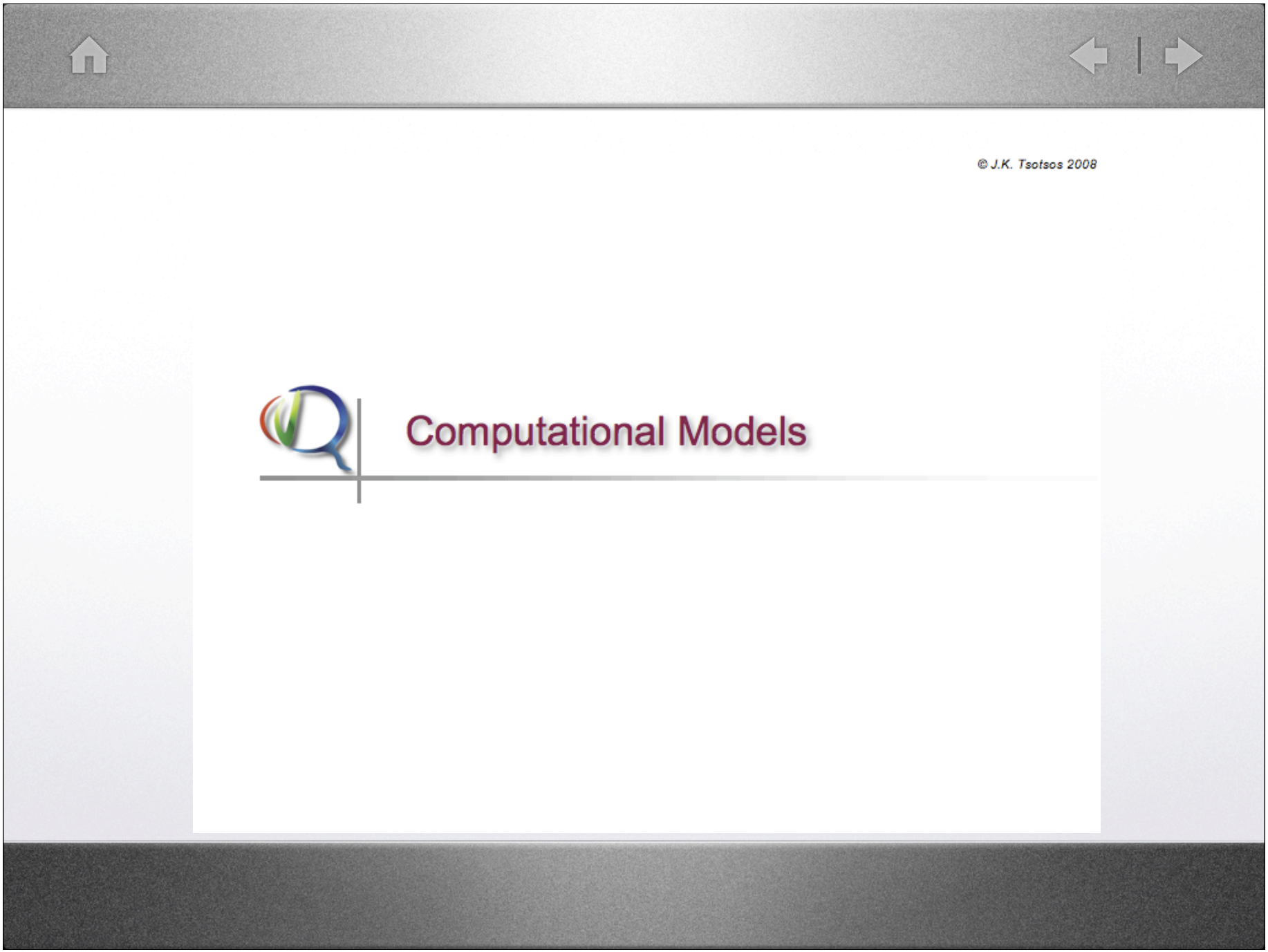

Active vision has cost

- decide that some action is needed
- decide which change to apply in priority sequence
- execute change
- adapt system to new viewpoint
- correspondence between old and new viewpoints

Benefit must outweigh cost (see Tsotsos IJCV 1992)

Active Vision ⊂ Attention



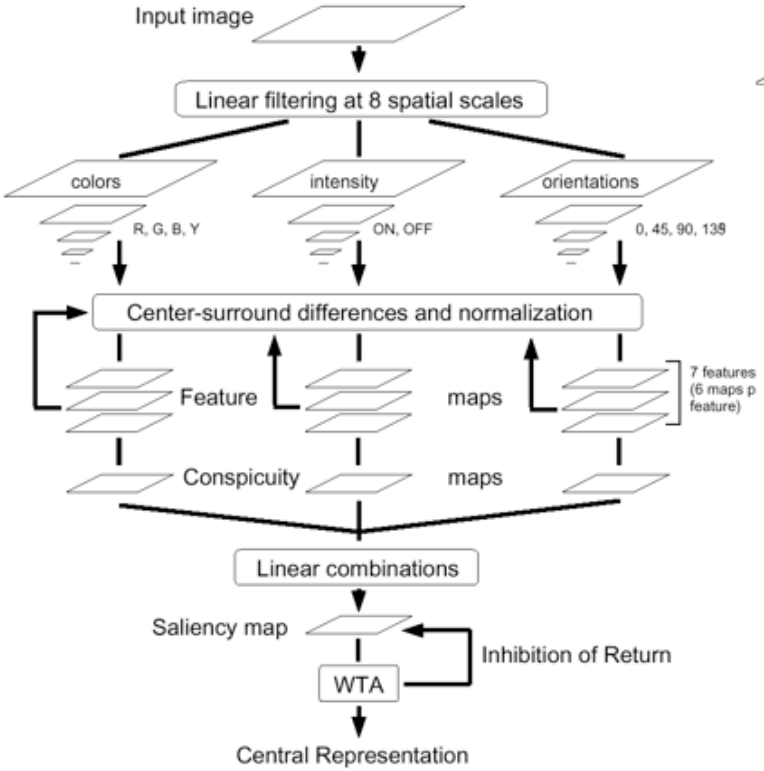


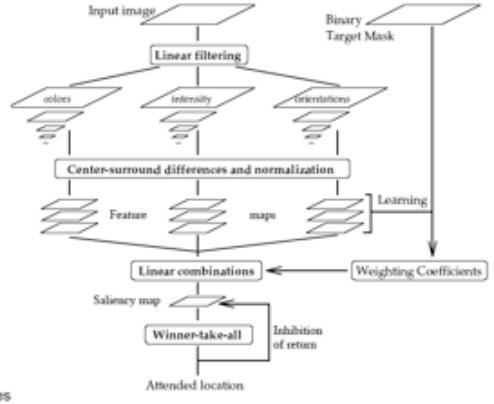
Itti, Koch & Niebur 1998+

Itti, L., Koch, C., Niebur, E. (1998). A model for saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence 20*, 1254-1259.
L Itti, C Koch. (2001). Computational modelling of visual attention. *Nat Rev Neurosci. 2(3)*:194-203.
Navalpakkam V, Itti L. (2005) Modeling the influence of task on attention, *Vision Res.* 45(2):205-31.

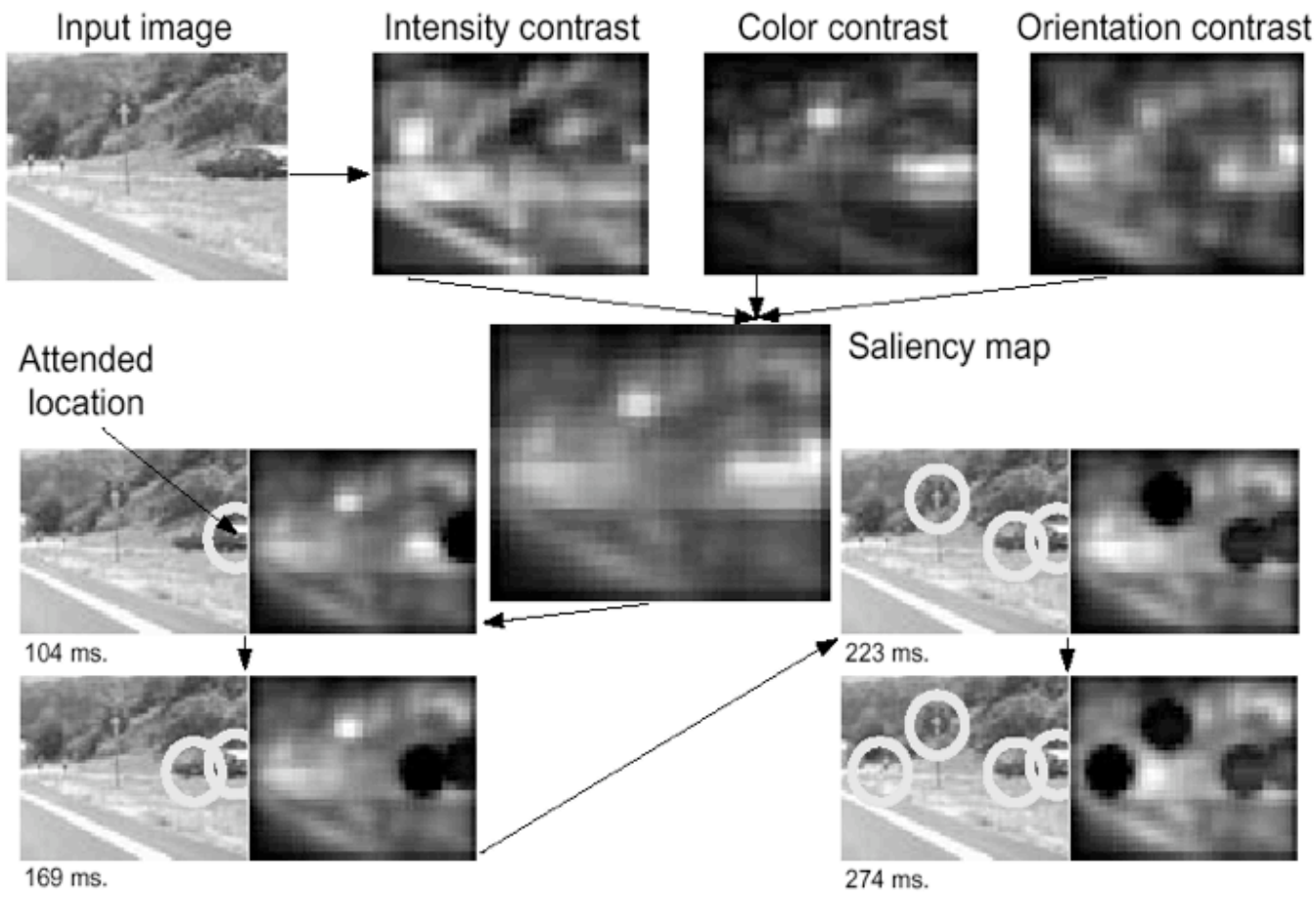
Key ideas:

- a newer implementation of Koch and Ullman's scheme
- fast and parallel pre-attentive extraction of visual features across 50 spatial maps (for orientation, intensity and color, at six spatial scales)
- features are computed using linear filtering and center-surround structures
- these features form a saliency map
- Winner-Take-All neural network to select the most conspicuous image location
- inhibition-of-return mechanism to generate attentional shifts
- saliency map topographically encodes for the local conspicuity in the visual scene, and controls where the focus of attention is currently deployed



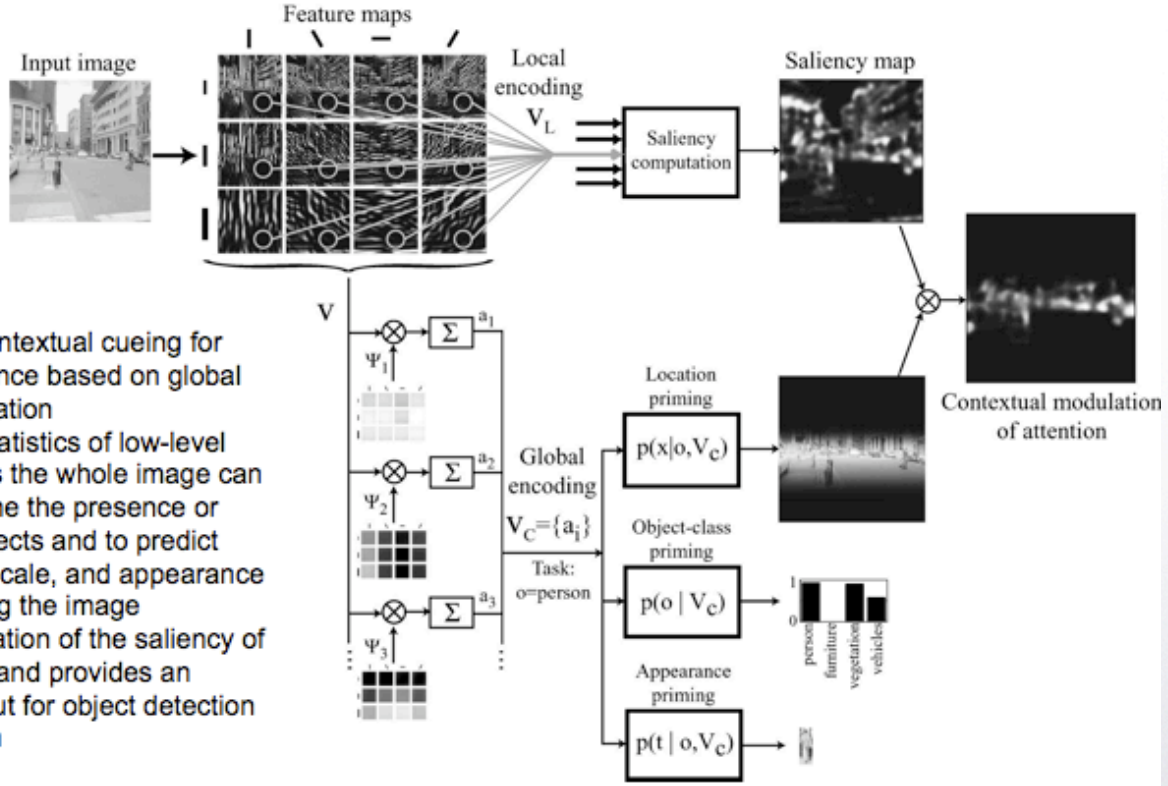


Navalpakkam & Itti - Given any new scene, our model uses the learnt representation of the target object to perform top-down biasing on the attention system such as to render this object more salient by enhancing those features which are characteristic of the object.



Torralba 2003

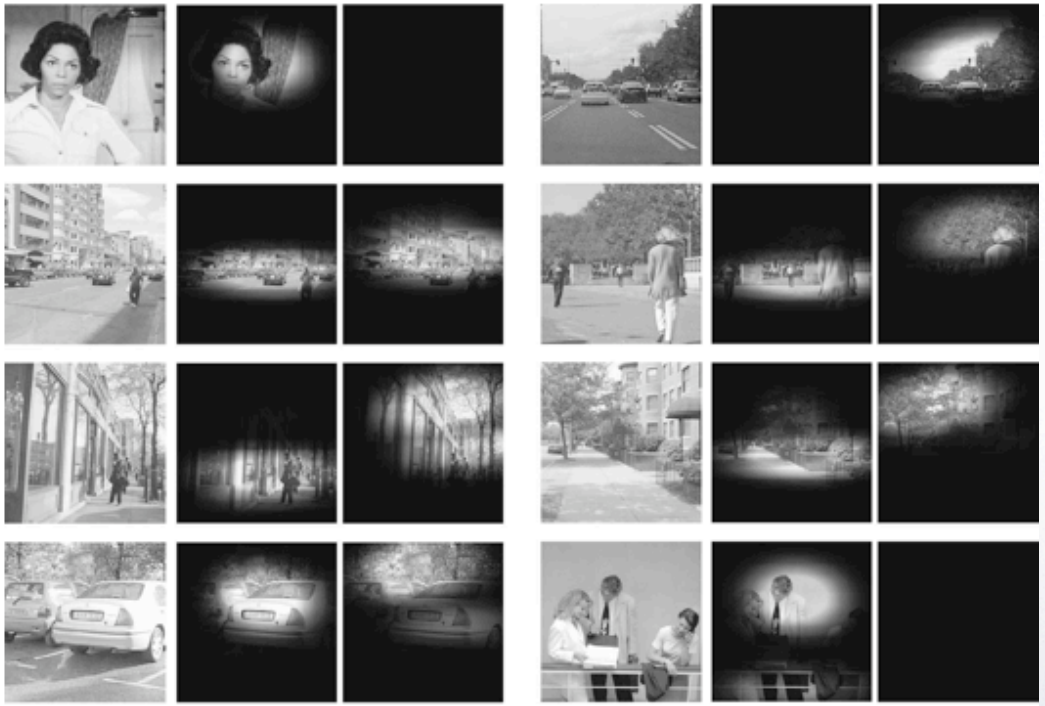
Torralba, A, (2003). Modeling global scene factors in attention, *J. Opt. Soc. Am. A*, Vol. 20, No. 7, 1407-1412.

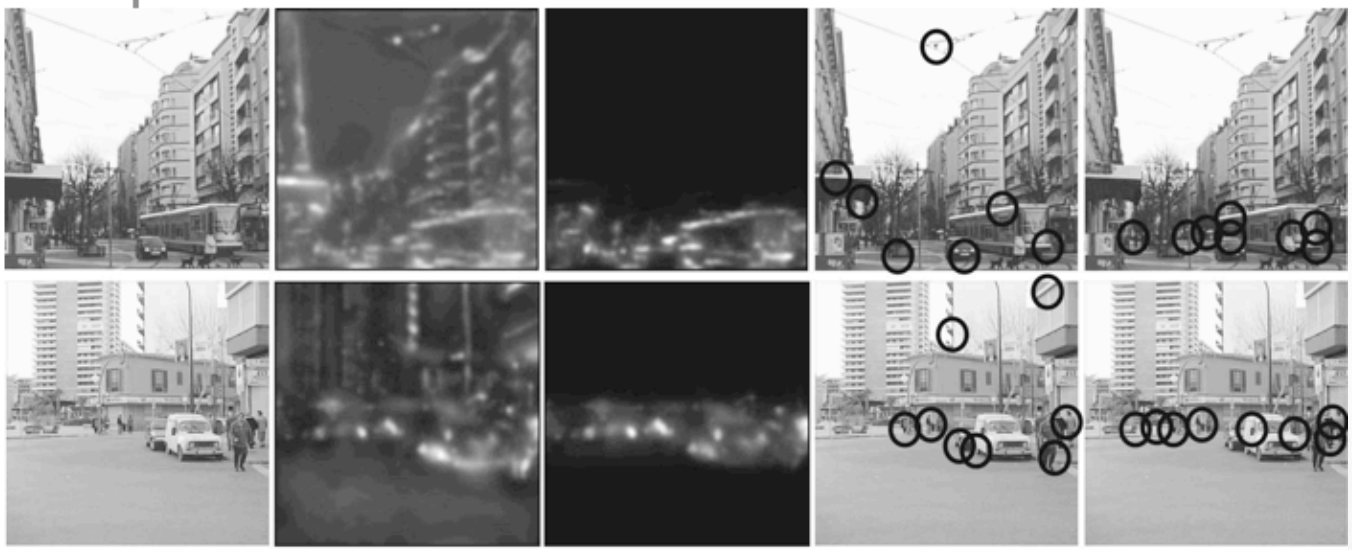


Key Ideas:

- a model of contextual cueing for attention guidance based on global scene configuration
- shows how statistics of low-level features across the whole image can be used to prime the presence or absence of objects and to predict their location, scale, and appearance before exploring the image
- allows modulation of the saliency of image regions and provides an efficient shortcut for object detection and recognition

Model results on context-driven focus of attention in the task of looking for faces (left) and vegetation (right).





(a) (b) (c) (d) (e)

(a) Input image (color is not taken into account). The task is to look for pedestrians.
(b) Bottom-up saliency map.
(c) Context-driven focus of attention. The image region in the shadow is not relevant for the task, and saliency is suppressed.
(d) Points that correspond to the largest salience.
(e) Image regions with the largest salience, including contextual priming

Lee, Buxton, Feng 2003

Lee, K.W., Buxton, H., Feng, J.F. Selective attention for cue-guided search using a spiking neural network, in Proc. International Workshop on Attention and Performance in Computer Vision, 2003, L. Paletta, G.W. Humphreys, and R.B. Fisher (eds.), pp. 55-63, 2003

Key ideas:

- a quick and dirty preprocessing primes the saliency map
- full resolution saliency map
- saliency is a combination of intermediate level bottom-up
- information (ellipses, symmetry, etc) and top down image
- based bias ("near red" "above blue" etc)

Fig. 4. Example of processing with a natural image containing faces. The model allocates focus of attention to possible target locations. The more task-relevant the target location with respect to the cue, the more likely the location is selected early in the attentional trajectory.

Fig. 5. Comparison with a saliency based model. See (left) trajectory of attention obtained from Itti's model and (right) trajectory of attention obtained from our model.

Fig. 7. The trajectories of attention guided by different colour cues. See (left) red colour cue "find a man who is wearing a red T-shirt" and (right) blue colour cue "find a man who is carrying a blue plastic bag".


The Basics...

1 Def'n: Attention is the set of mechanisms that optimize and control the inherent search processes in vision, sensory perception or cognition

2. The set of mechanisms may be summarized as:

Selection	spatio-temporal region of interest
	world/task/object/event model
	gaze/viewpoint
	best interpretation/response
Restriction	task relevant search space pruning
	location cues
	fixation points
	search depth control
Suppression	spatial/feature surround inhibition
	inhibition of return

3. If you wish to use biological motivation for a computational theory then you cannot ignore the task subjects performed, the class of images subjects viewed, and the experimental paradigm that lead to the results you choose to use