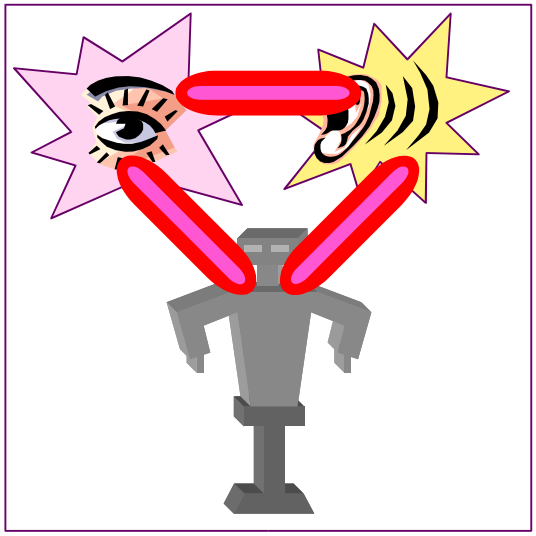
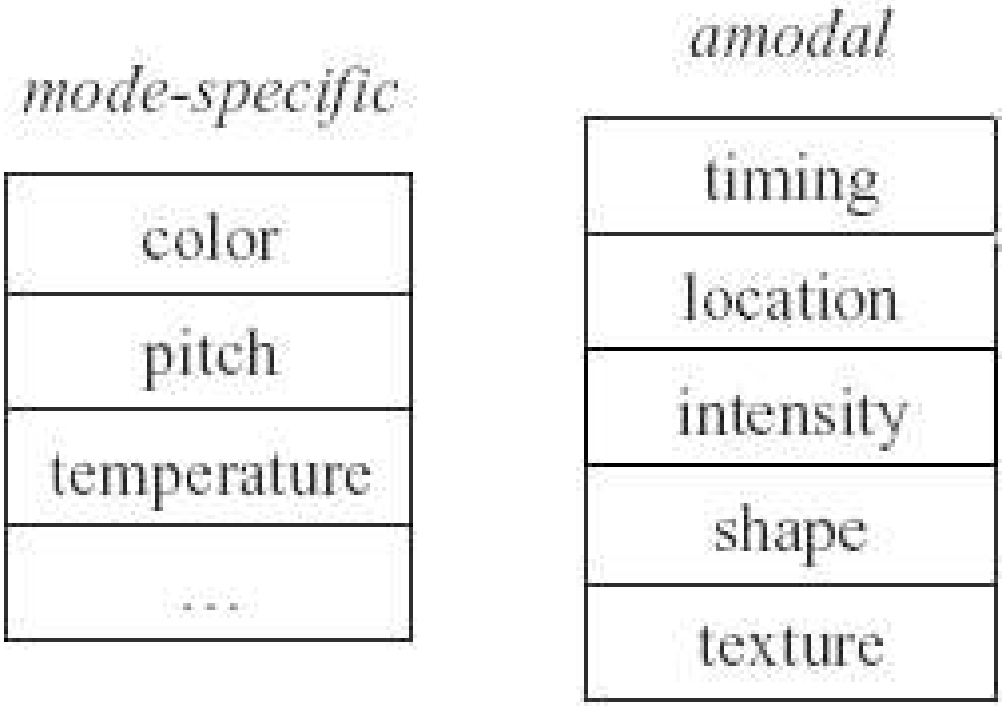


Feel the beat:
using cross-modal rhythm to integrate perception of objects, others, and self

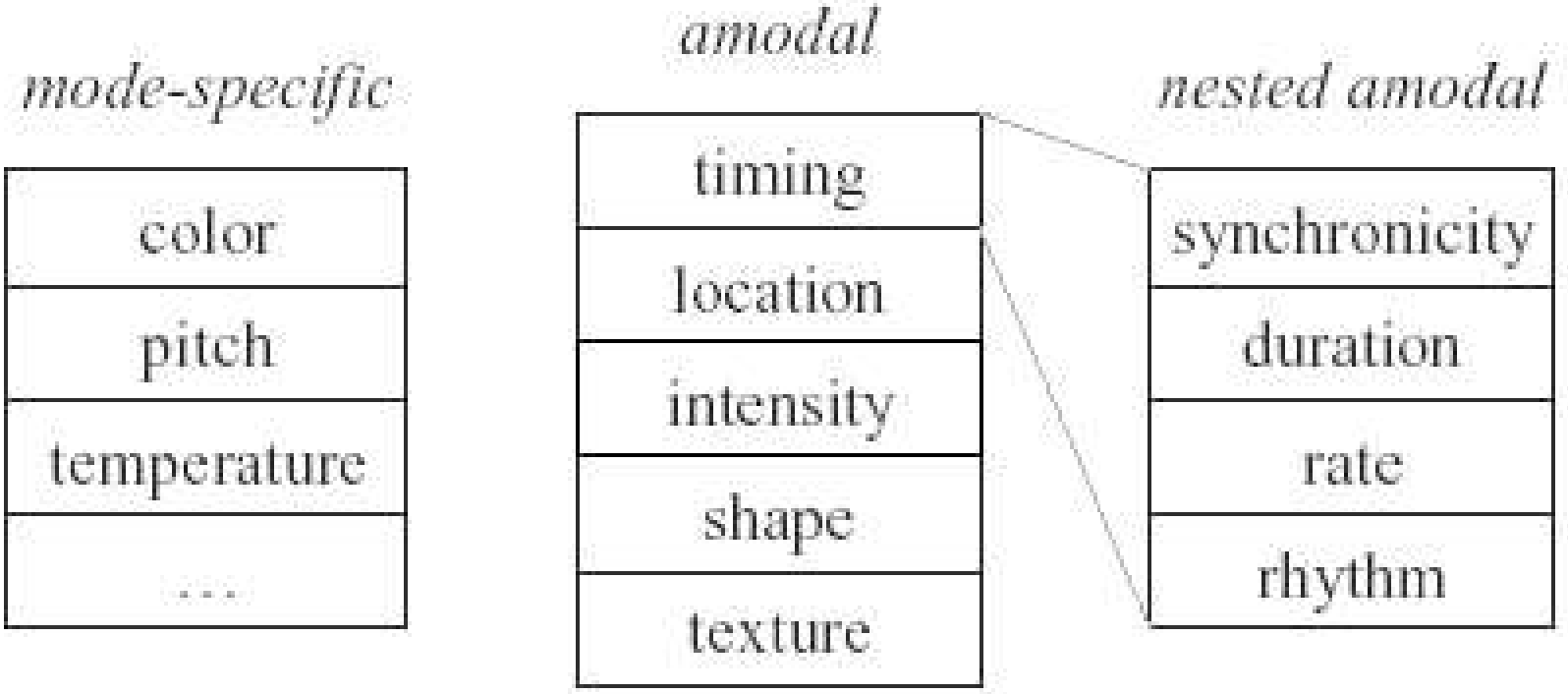
Paul Fitzpatrick and Artur M. Arsenio

CSAIL, MIT

Modal and amodal features



Modal and amodal features



(following Lewkowicz)

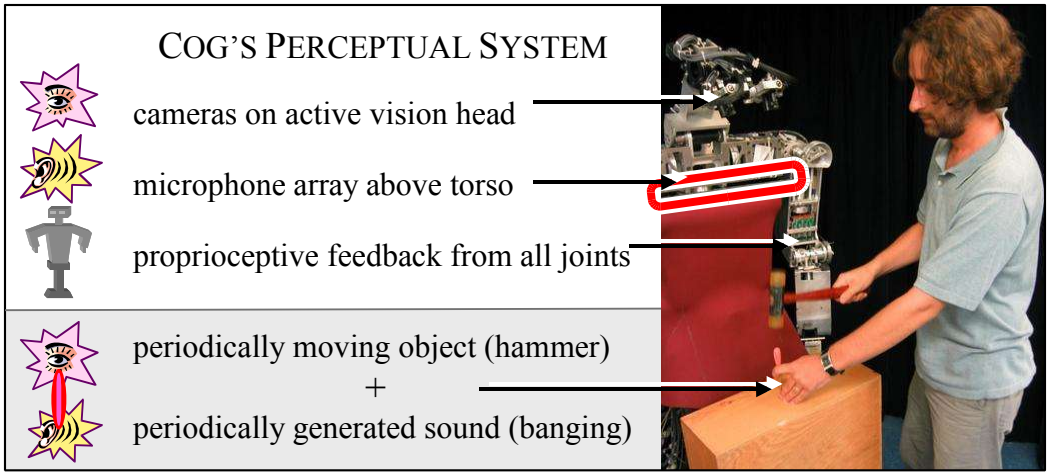
Motivation

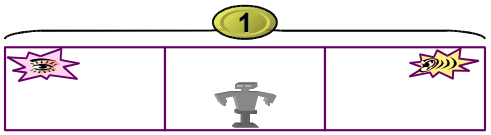
- Tools and toys are often used in a manner that is composed of some repeated motion - consider hammers, saws, brushes, files, ...
- Rhythmic information across the visual and acoustic sensory modalities have complementary properties
- Features extracted from visual and acoustic processing are what is needed to build an object recognition system

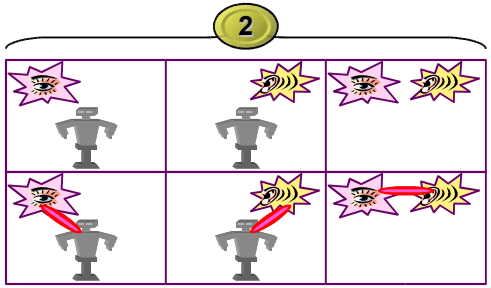
Talk Outline

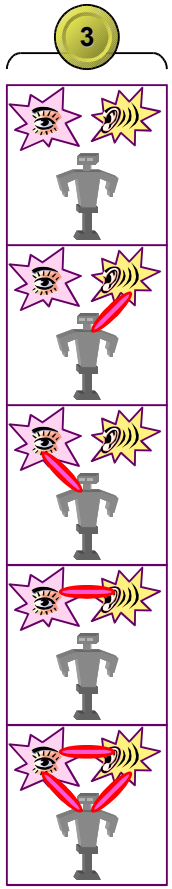
- Hardware
- Matching sound and vision
- Priming for attention
- Differentiation
- Integration
- The self and others

Cog's Perceptual System









Interacting with the robot

Making sense of the senses…

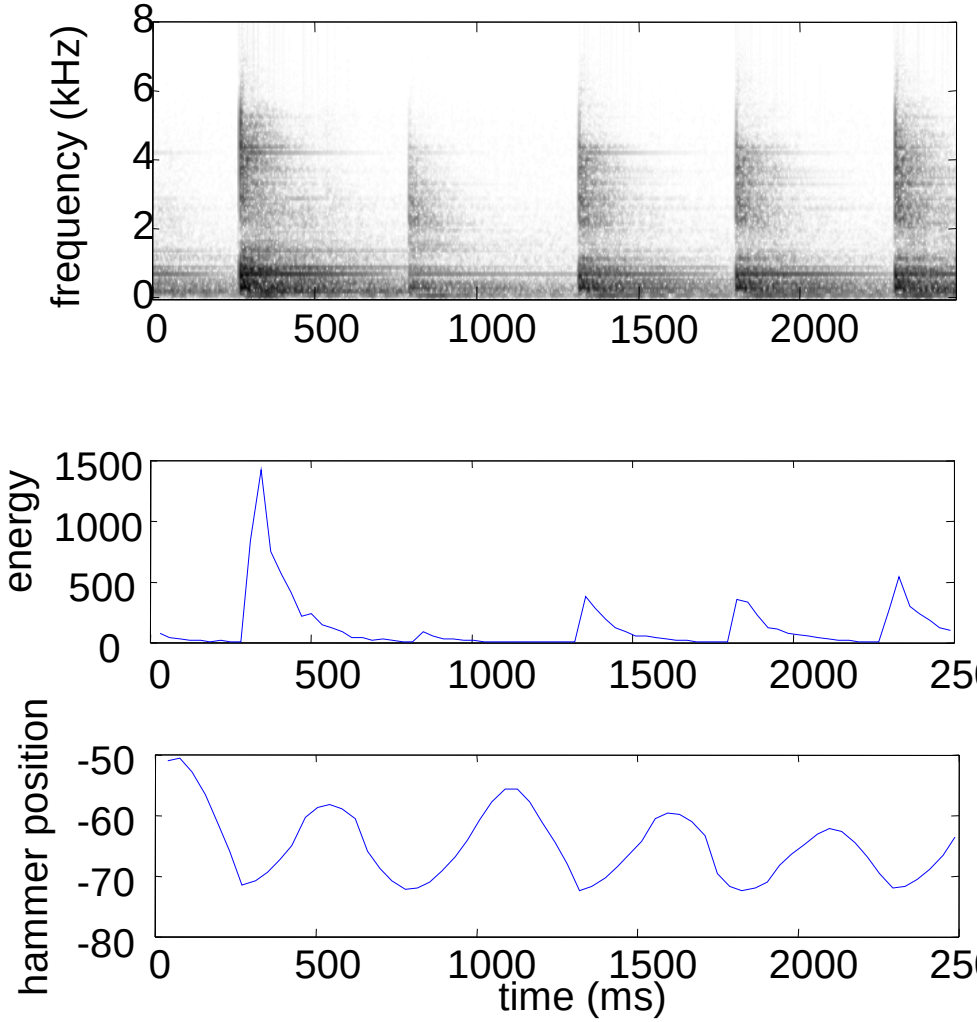
Bang, Bang !

Who is he?

Talk Outline

- Hardware
- Matching sound and vision
- Priming for attention
- Differentiation
- Integration
- The self and others

Matching sound and vision



The sound intensity peaks once per visual period of the hammer (CIRAS 2003)

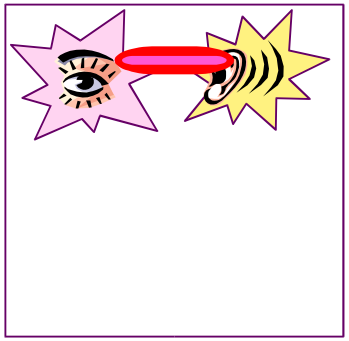
Matching algorithm

- Estimate signal period (histogram technique from CIRAS 2003)
- Cluster rising and falling intervals, guided by the scale of estimated period
- Merge sufficiently close clusters
- Segment full periods in the signal

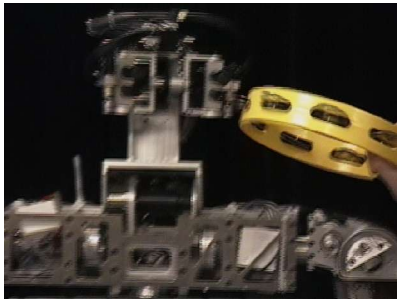

2.6 CU. FT.
LETTER

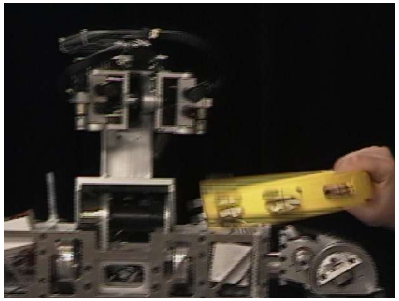
Playing a tambourine

Appearance and sound of tambourine are bound together

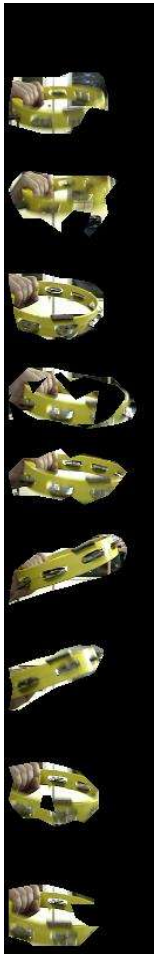


robot sees and hears a tambourine shaking



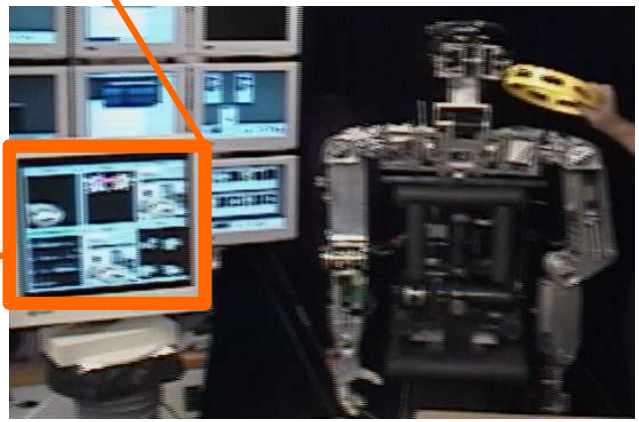


tambourine segmentations



Object Segmentation

Multiple Object Tracking



Sound Segmentation (window divided in 4x4 images)

Cog's view

Object Recognition (window divided in 2x2 images)

Robustness

to random visual disturbances

to auditory disturbances

Person talks – sound not matched to object!

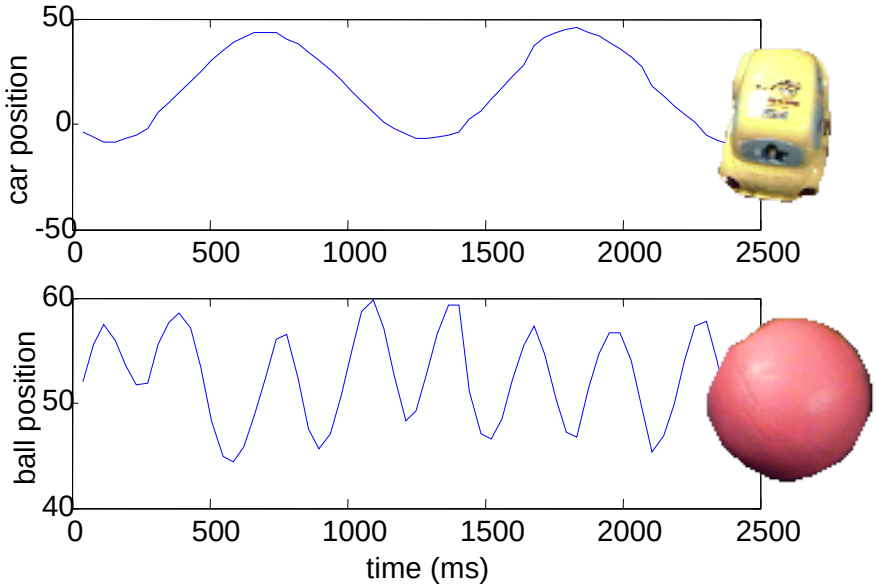
Talk Outline

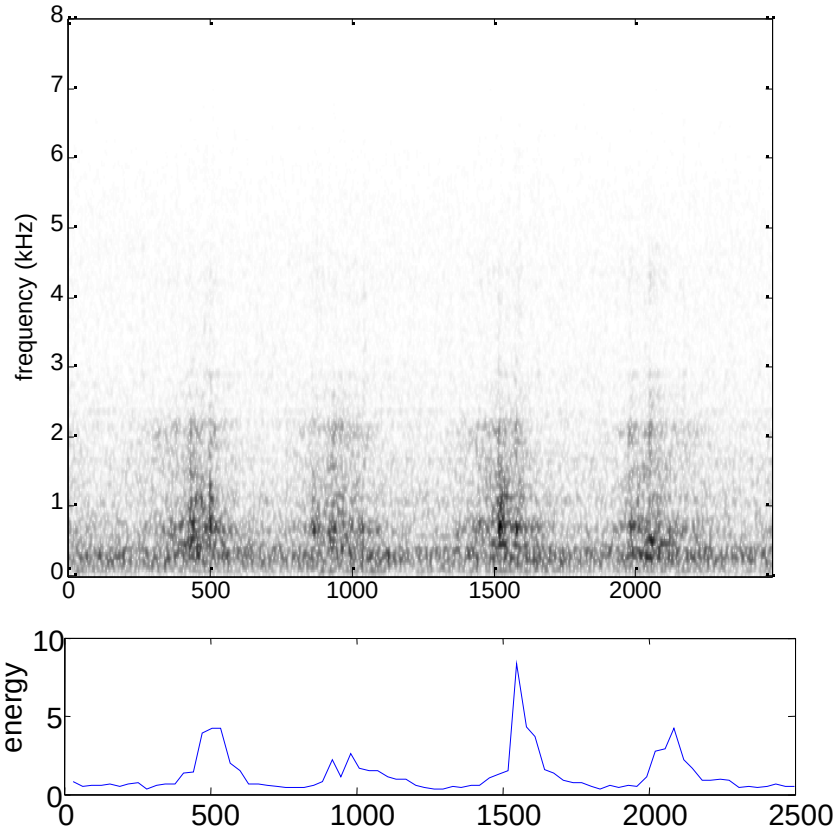
- Hardware
- Matching sound and vision
- Priming for attention
  - Priming visual foreground with sound
  - Priming acoustic foreground with vision
  - Matching multiple sources
- Differentiation
- Integration
- The self and others

Priming visual foreground with sound

- One object (the car) making noise
- Another object (the ball) in view
  - Problem: which object goes with the sound?
  - Solution: Match periods of motion and sound

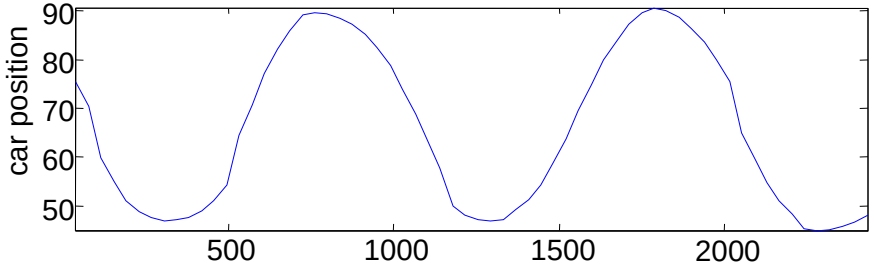
Comparing periods

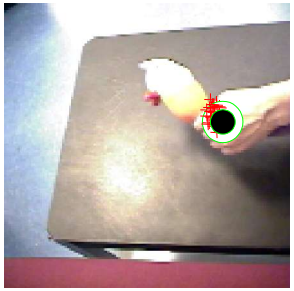


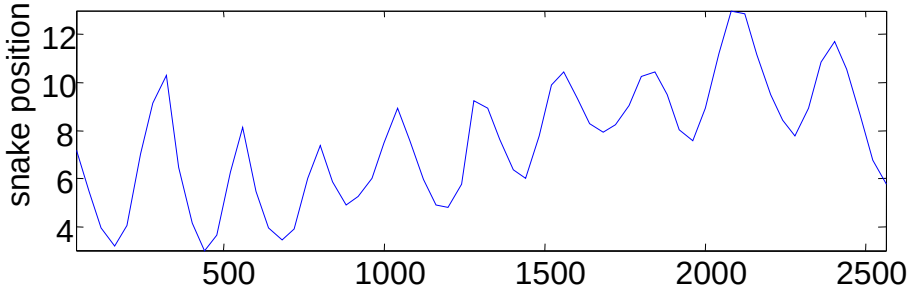


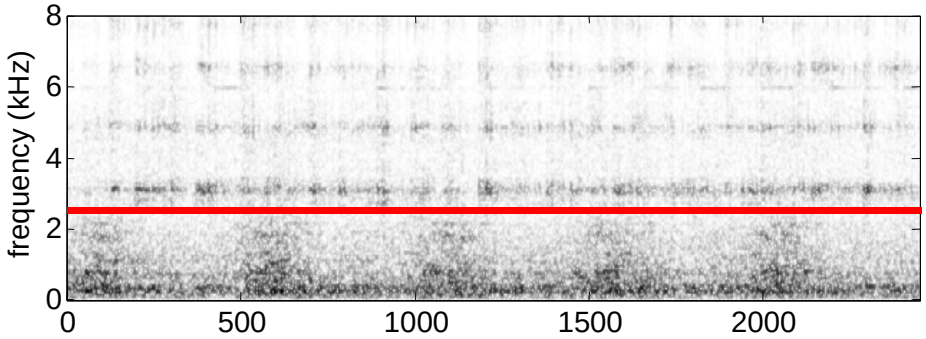
- The sound intensity peaks twice per visual period of the car

Matching with acoustic distraction





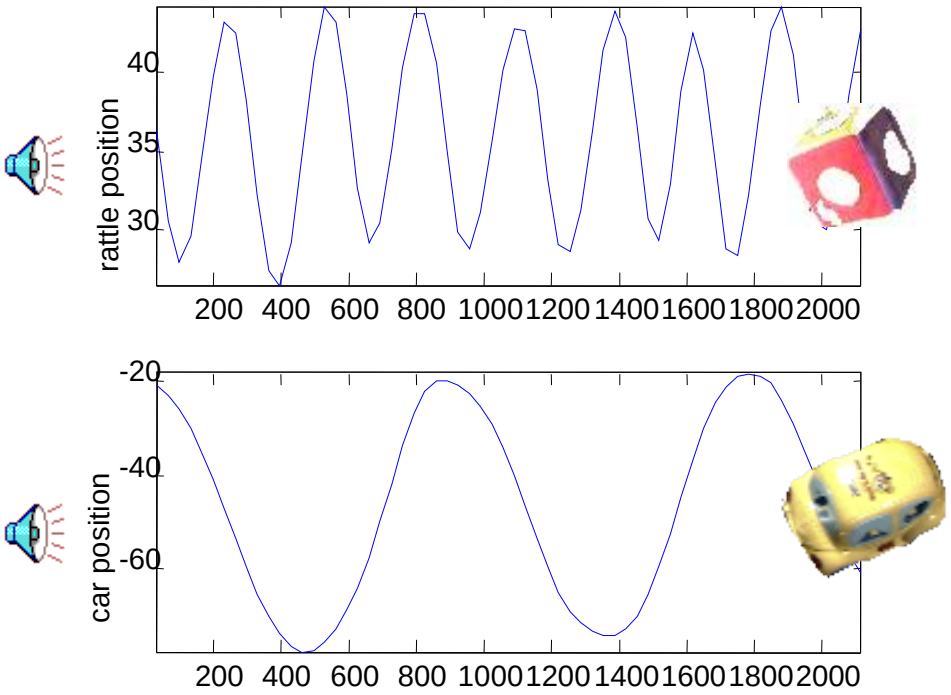


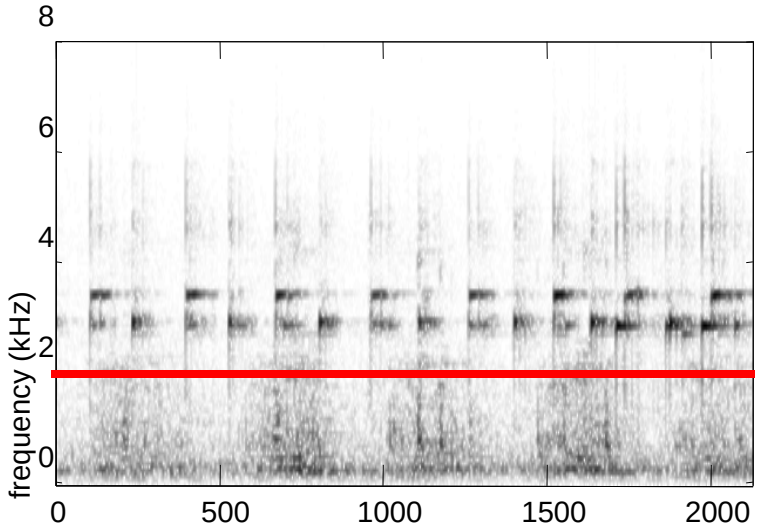


Matching multiple sources

- Two objects making sounds with distinct spectrums
  - Problem: which object goes with which sound?
  - Solution: Match periods of motion and sound

Binding periodicity features





- The sound intensity peaks twice per visual period of the car. For the cube rattle, the sound/visual signals have different ratios according to the frequency bands

Cross-modal association - errors

Experiment	visual period found	sound period found	bind sound, vision	candidate binds	correct binds	incorrect binds
hammer	8	8	8	8	8	0
car and ball	14	6	6	15	5	1
plane & mouse/remote	18	3	3	20	3	0
car (snake in backg'd)	5	1	1	20	1	0
snake (car in backg'd)	8	6	6	8	6	0
car & cube *car*	9	3	3	11	3	0
car & cube *cube*	10	8	8	11	8	0
car & snake *car*	8	0	0	8	0	0
car & snake *snake*	8	5	5	8	5	0

Talk Outline

- Hardware
- Matching sound and vision
- Priming for attention
- Differentiation
  - Visual Recognition
  - Sound Recognition
- Integration
- The self and others

Visual Object Segmentation/Recognition

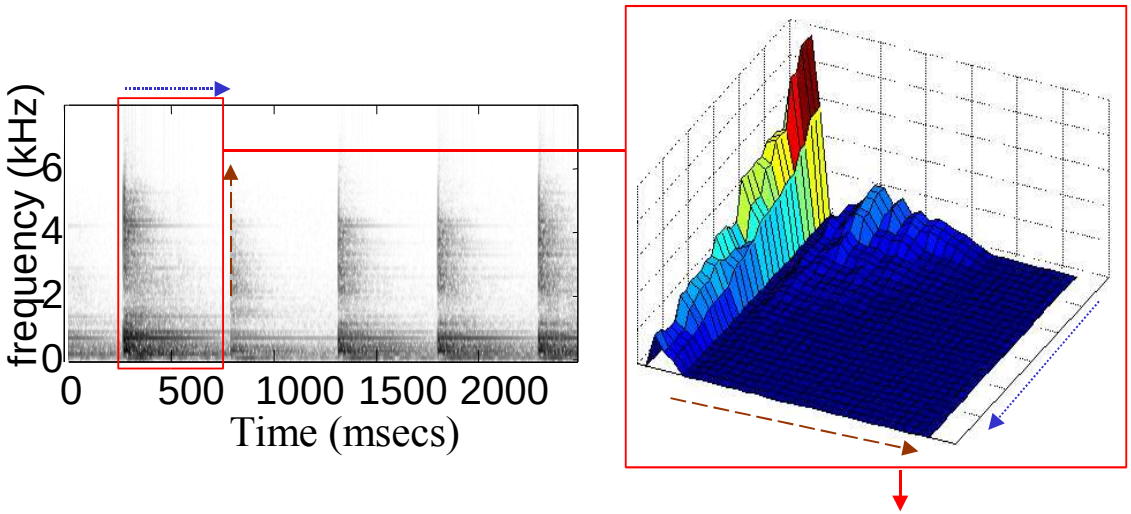
- Object Segmentation

- Object Recognition

see Arsenio, MIT PhD thesis, 2004 for visual objectsegmentation/recognition

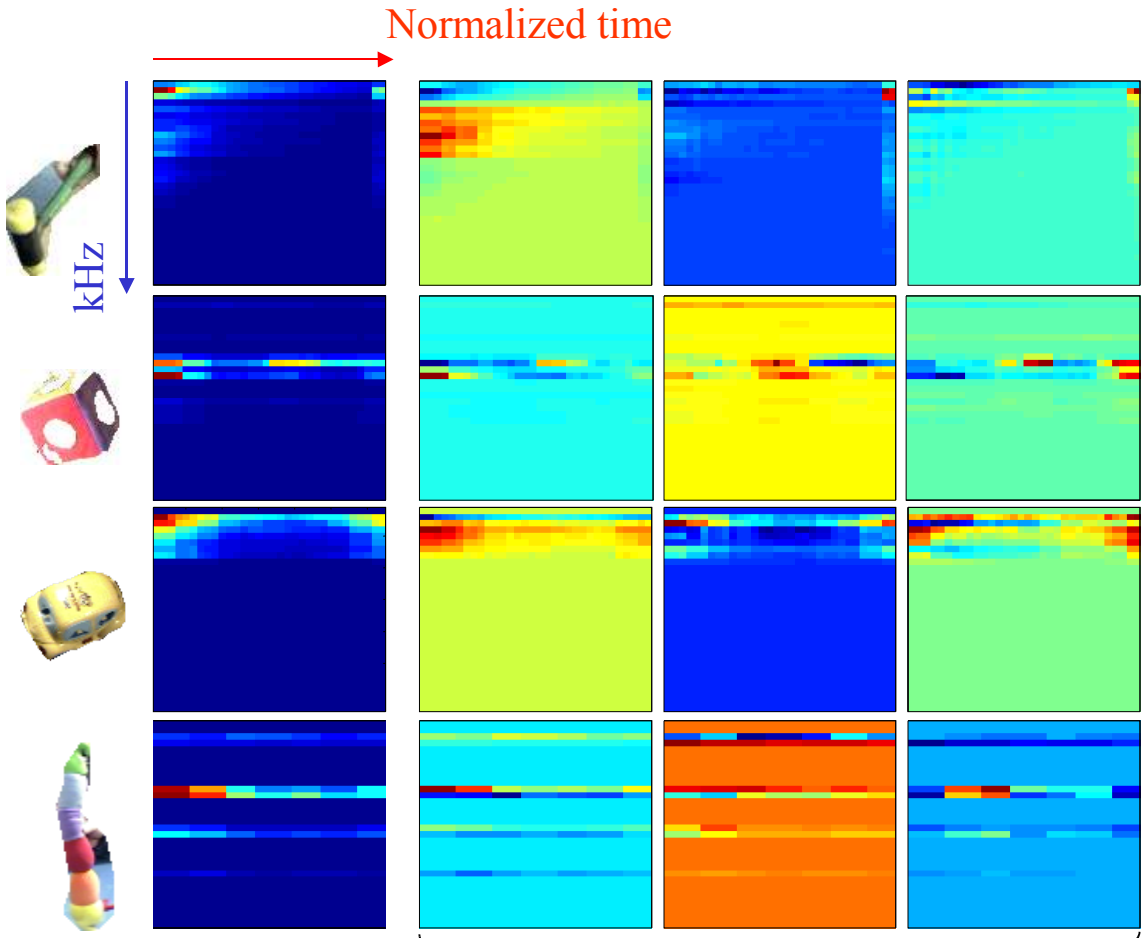
Sound Segmentation

Goal: Extract acoustic signatures from repetitive data
Problem: STFTs applied for spectral analysis, but not ideal for irregular signals
Solution: Build histograms of hypothesized periods



7 random sound samples for each of 4 objects. From top to bottom: hammer, cube rattle, car and snake rattle, respectively.

Sound Recognition



Recognition rate: 82%

Talk Outline

- Hardware
- Matching sound and vision
- Priming for attention
- Differentiation
- Integration
  - Cross-modal segmentation/recognition
  - Cross-modal enhancement of detection
- The self and others

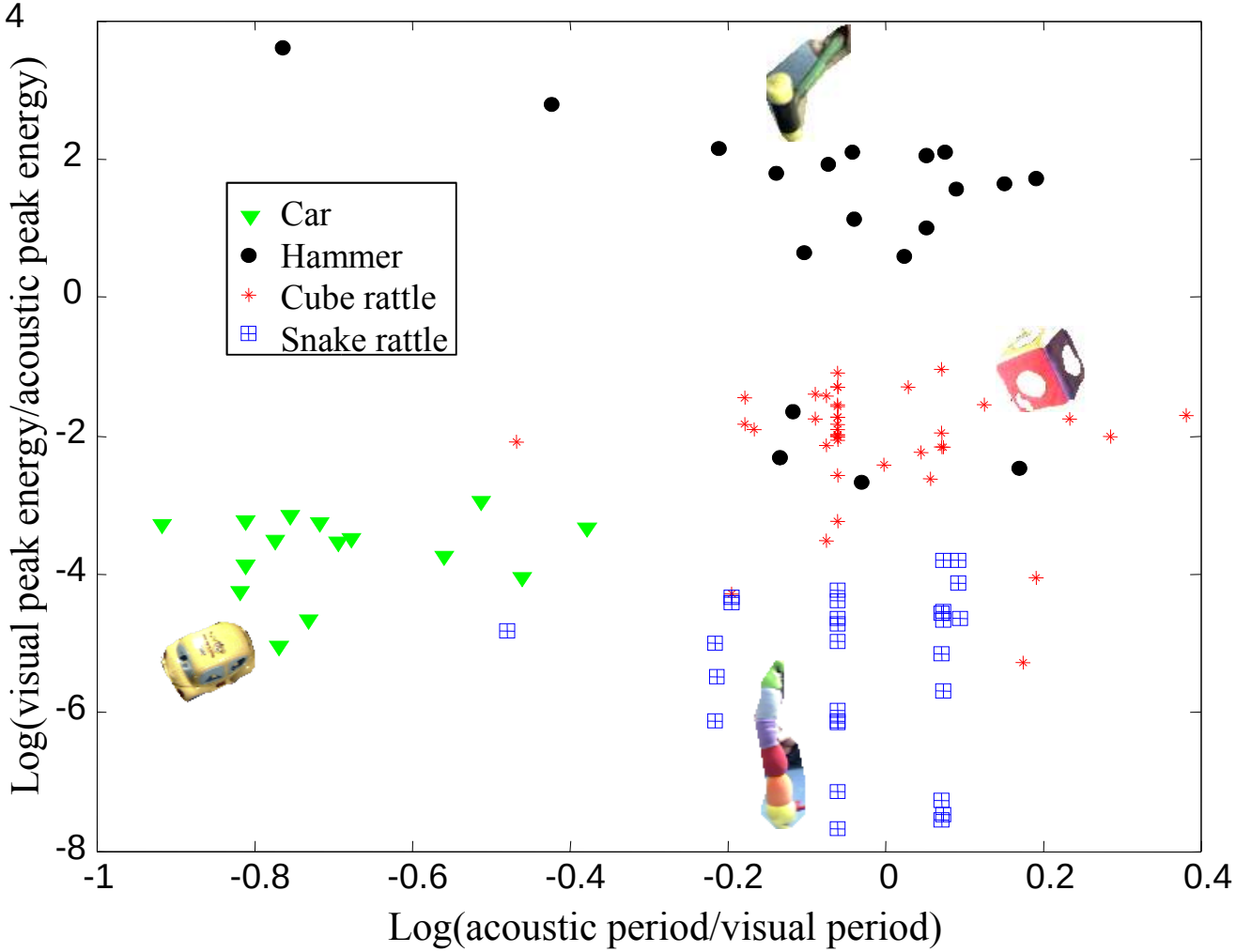
Cross-modal object recognition

Causes sound when changing direction after striking object; quiet when changing direction to strike again

Causes sound while moving rapidly with wheels spinning; quiet when changing direction

Causes sound when changing direction, often quiet during remainder of trajectory (although bells vary)

Cross-modal object recognition



Cross-modal recognition rate: 92%

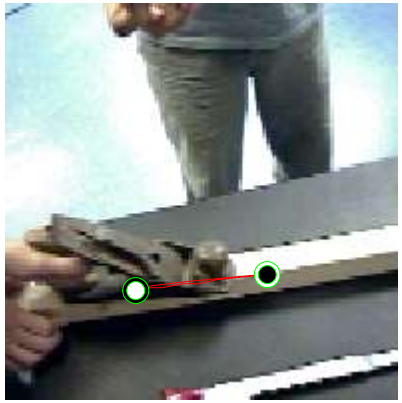
Dynamic Programming

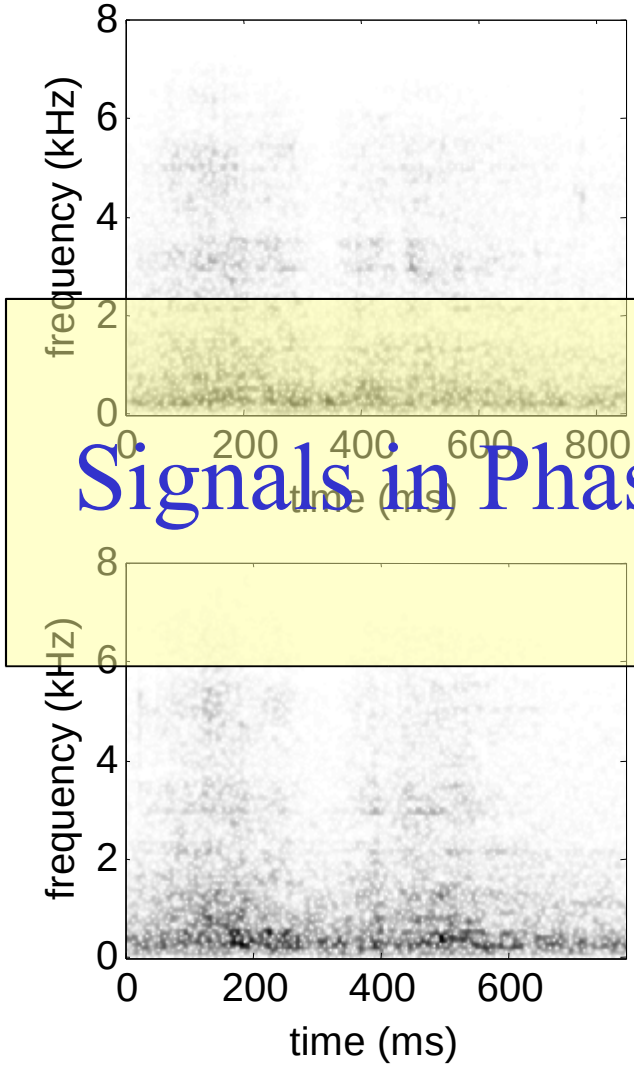
is applied to match previously segmented sensory signals: visual trajectories to the sound energy signal

Cross-modal recognition – confusion table

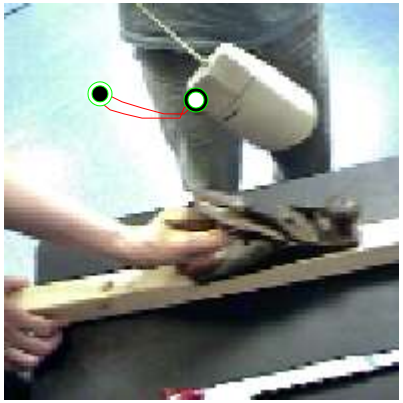
Confusion matrix	car	cube	snake	hammer
car	**30**	0	0	0
cube	0	**52**	7	1
snake	0	0	**45**	0
hammer	0	5	0	**25**

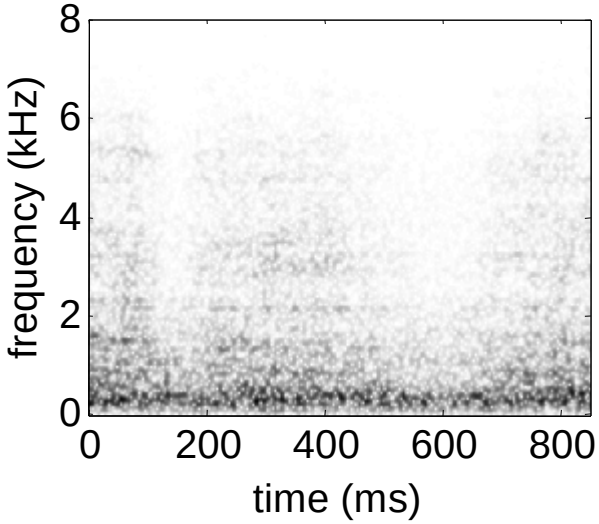
Cross-modal enhancement of detection

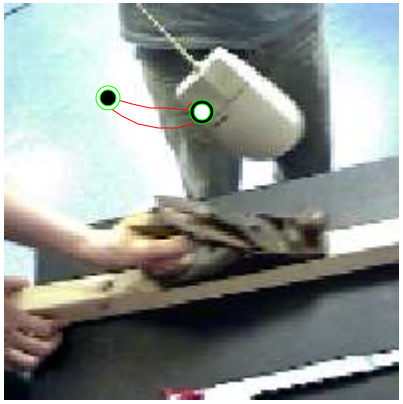


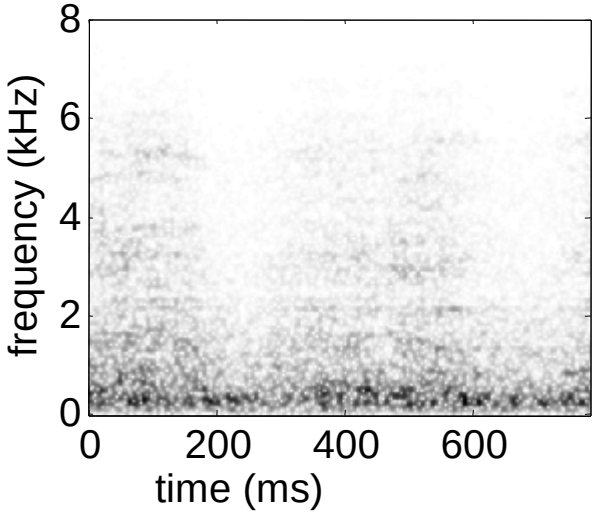


Signals out of phase!









Talk Outline

- Hardware
- Matching sound and vision
- Priming for attention
- Differentiation
- Integration
- The self and others
  - Learning about people
  - Learning about the self

Cross-modal rhythm

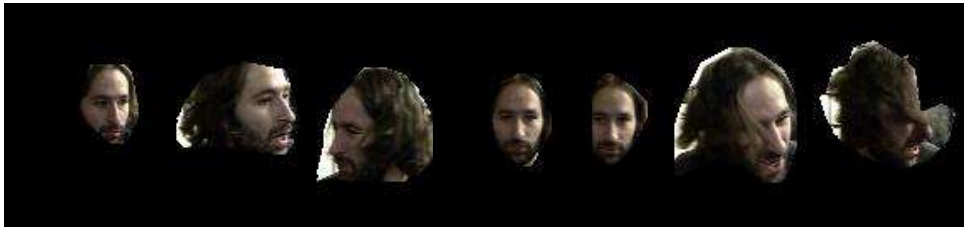
to integrate perception of others

Control Experiment 1

the robot sees a person shaking head – no periodic sound

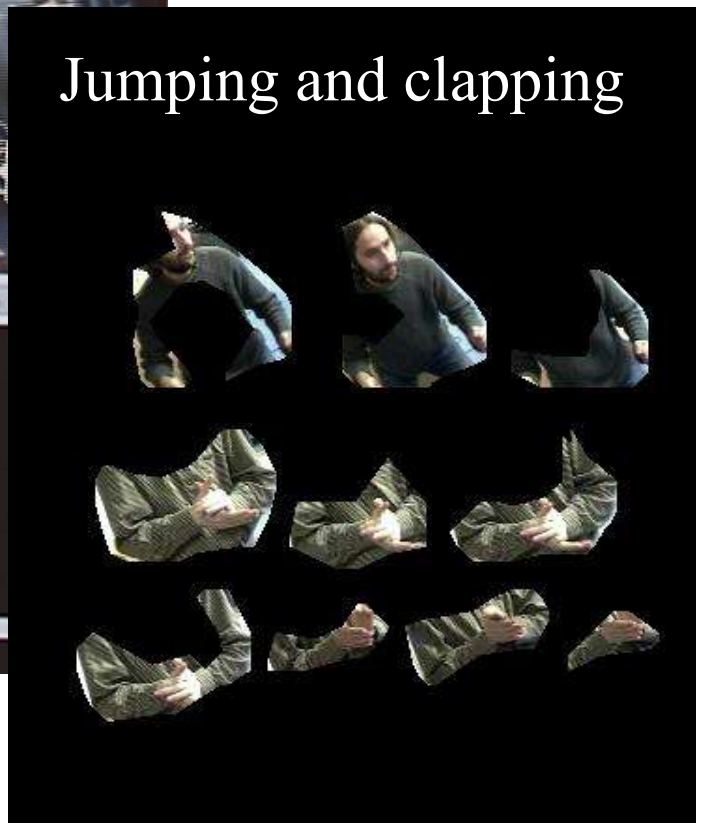
Experiment 2

the robot sees a person shaking head and saying “no”



Cross-modal rhythm

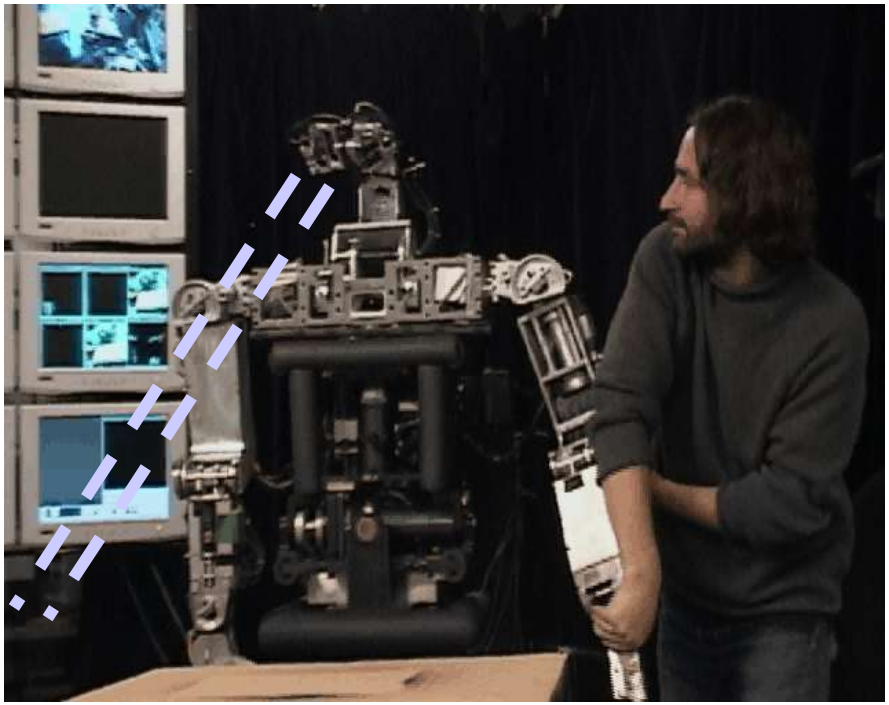
to integrate perception of others



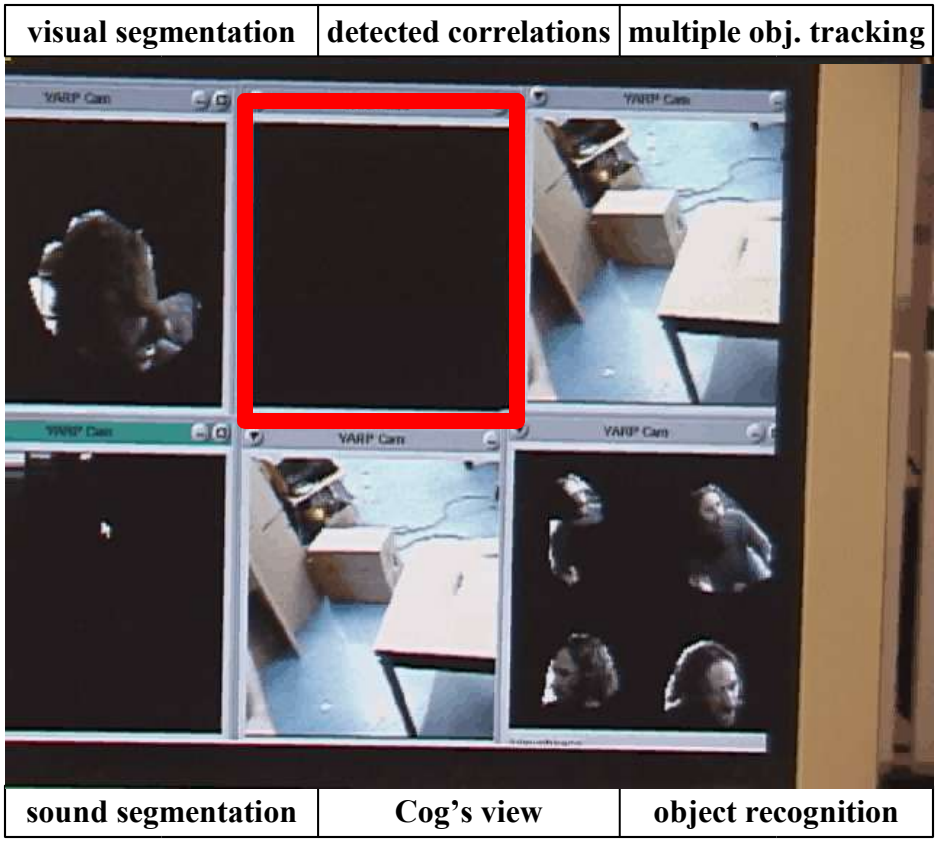
Small visual/sound delay gap – network delay

Binding

Sound and Proprioceptive Data



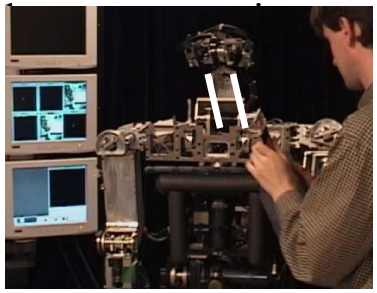
Detecting ones'own rhythms

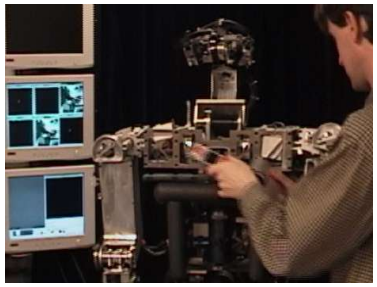


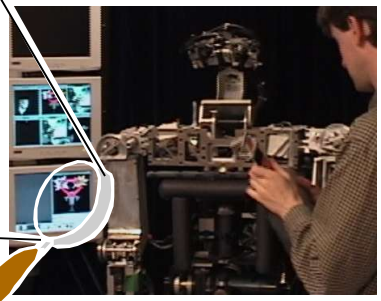
Binding Vision, Sound and Proprioceptive Data

Visual image segmented, sound detected, and all bounded to the motion of the arm

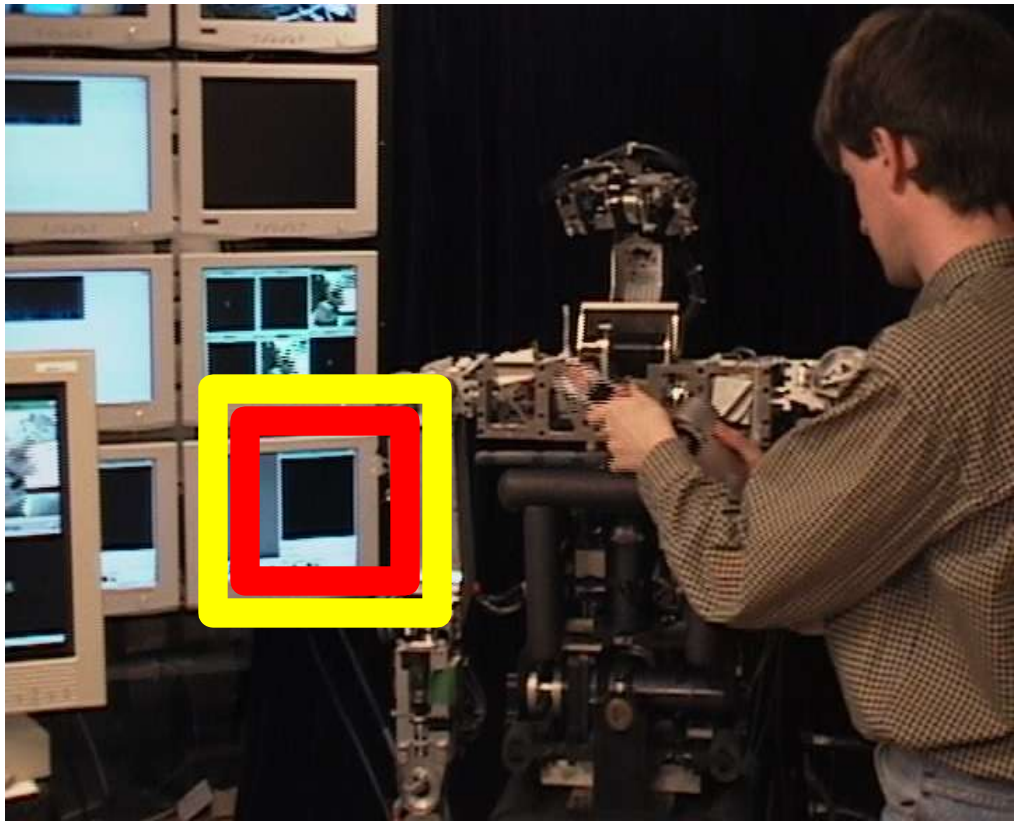
robot is looking towards its arm as





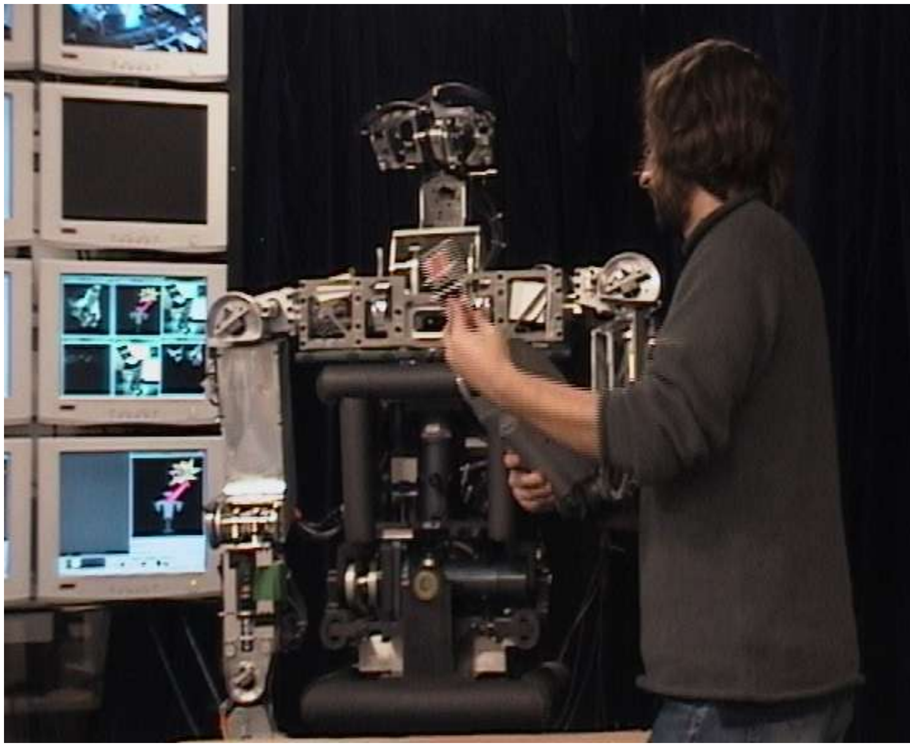


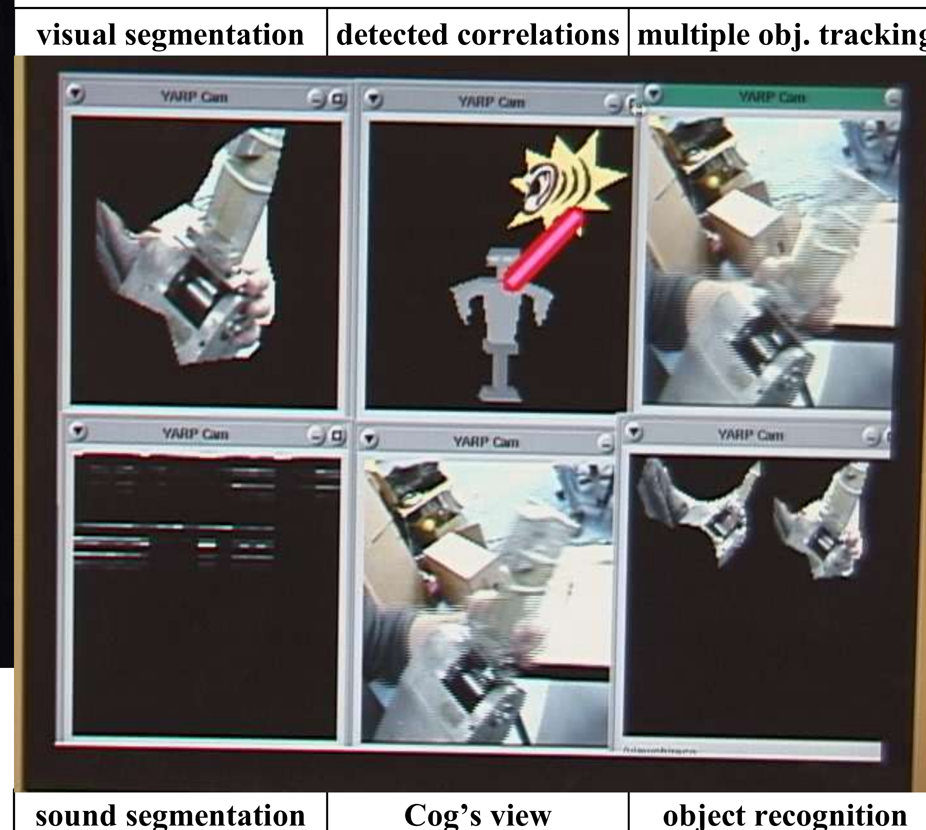
Video



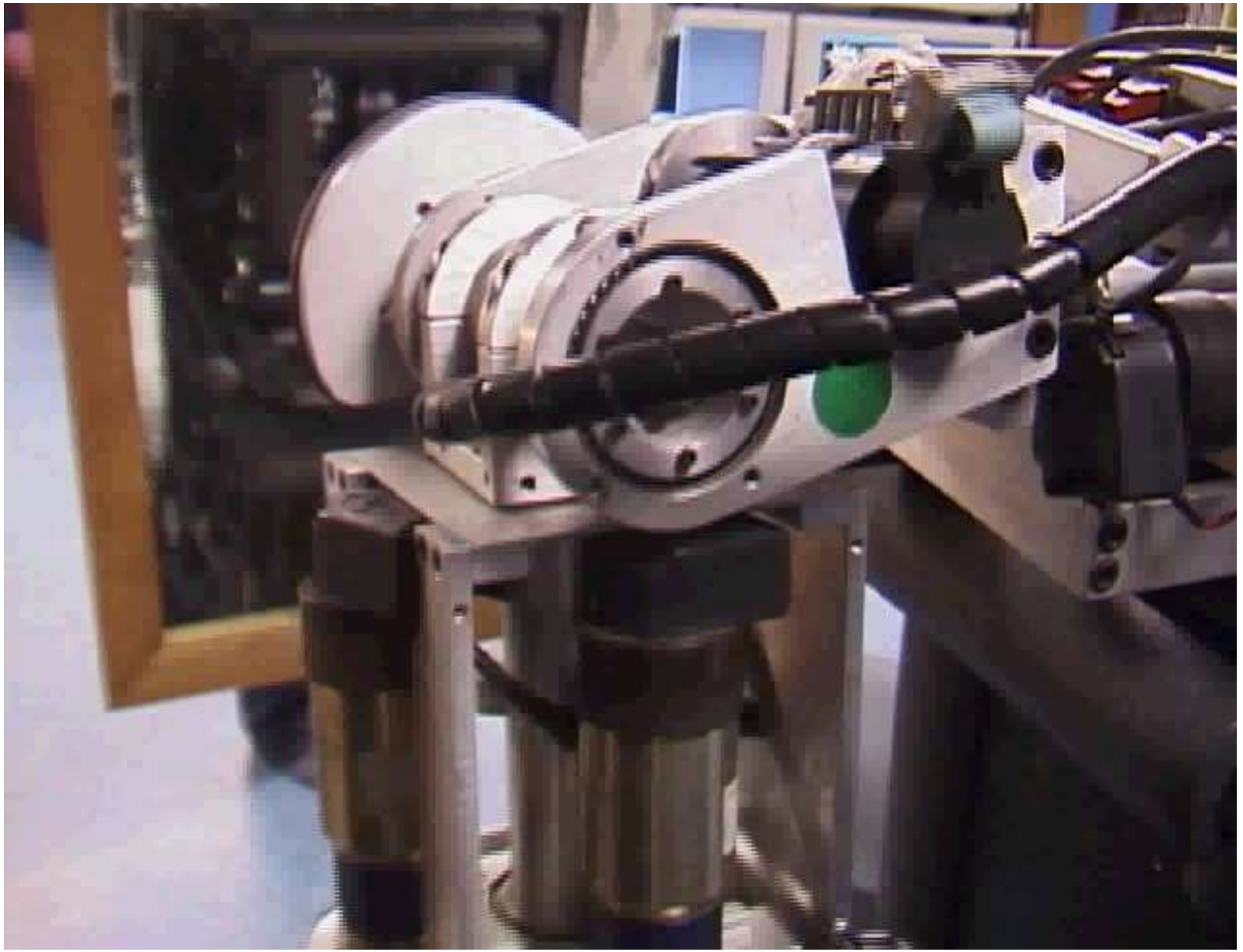
Binding

Vision, Sound and Proprioceptive Data

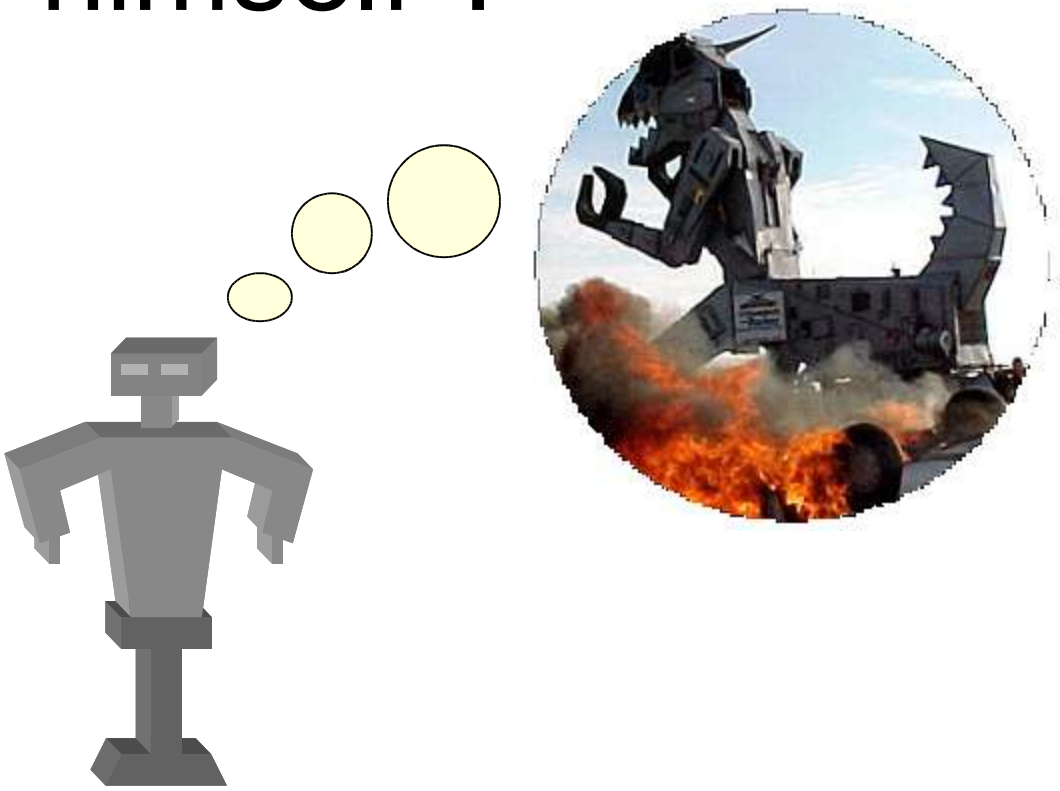


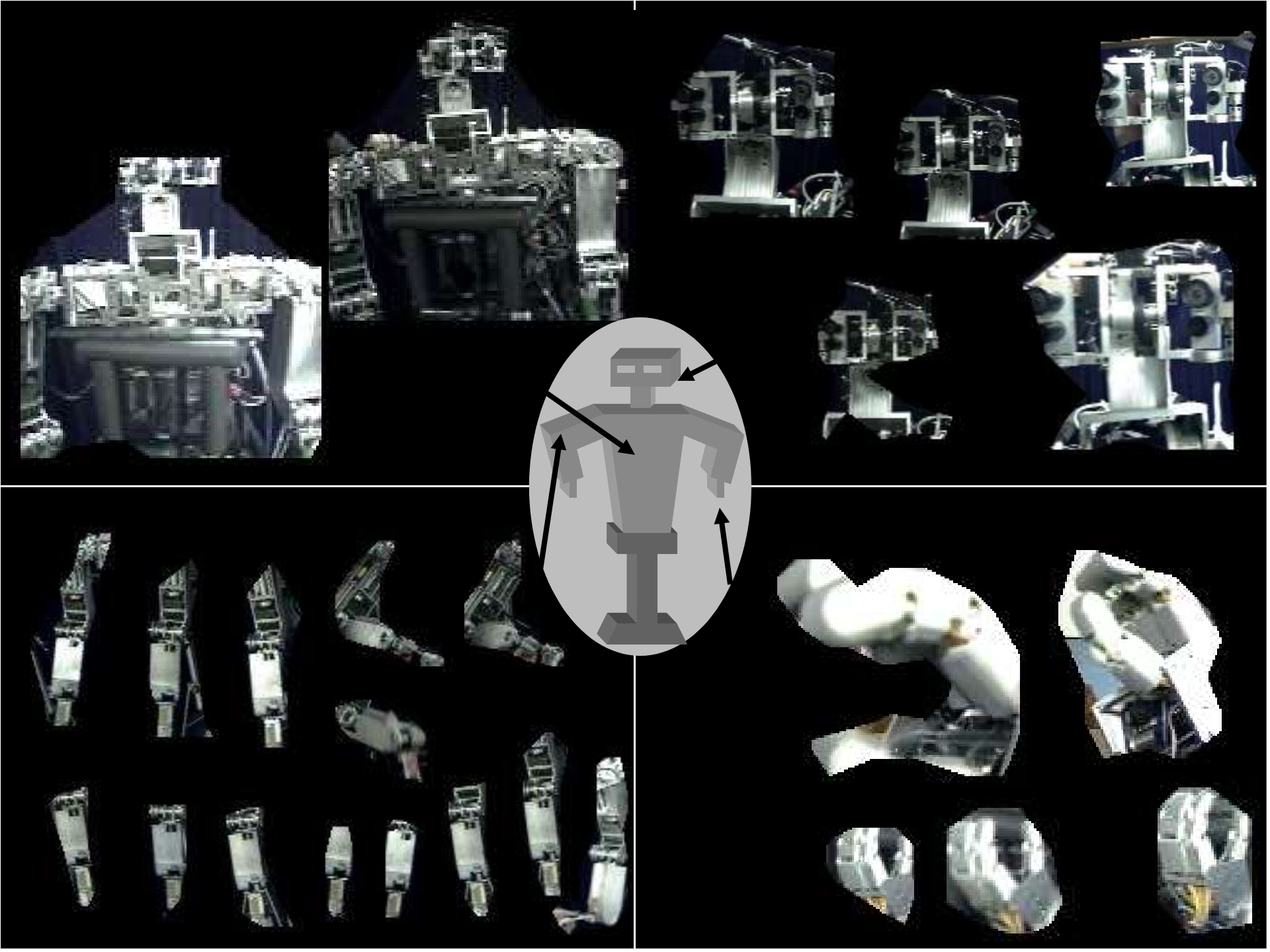


Cog’s mirror image

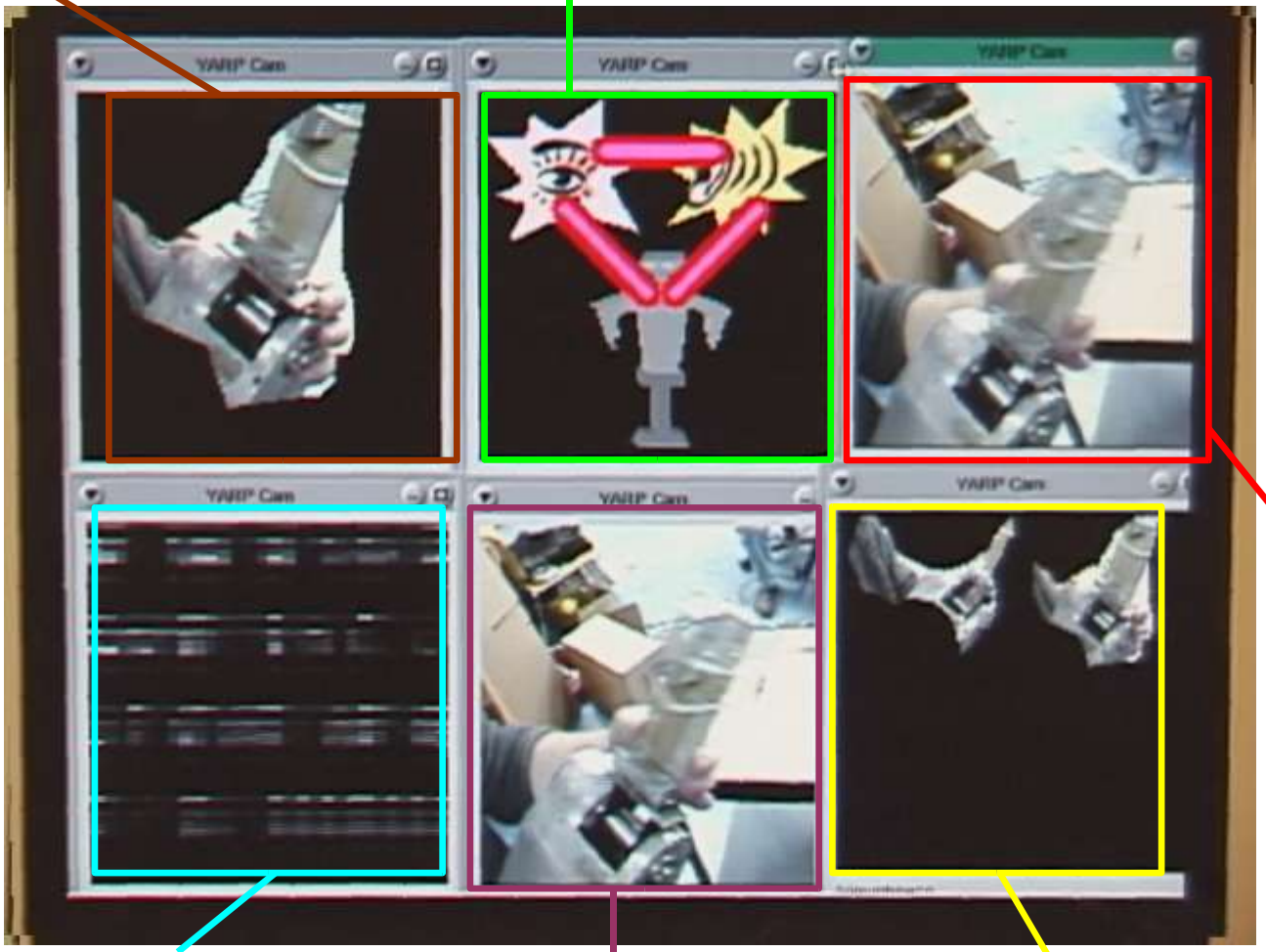


So, how does Cog perceive himself ?





The robot's experience of an event



Conclusions

- Amodal features are key to detecting relationships across senses
- Useful for learning to recognize an object in different senses (e.g. by its appearance or its sound)
- There are features for object recognition that exist only in relationships across senses and do not exist in any one sense
- Useful both for perception of external objects and robot's own body, by incorporating proprioception as another sense