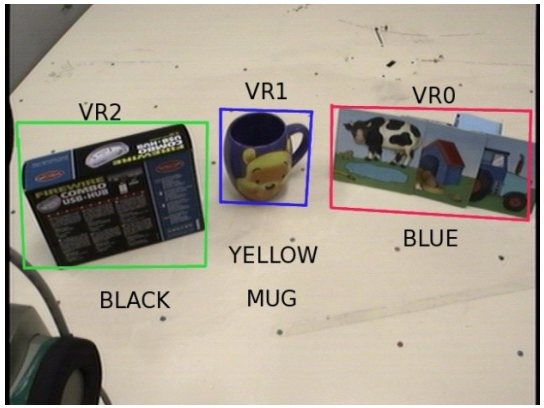
Grounding Object Semantics in Multi-modal Interactions

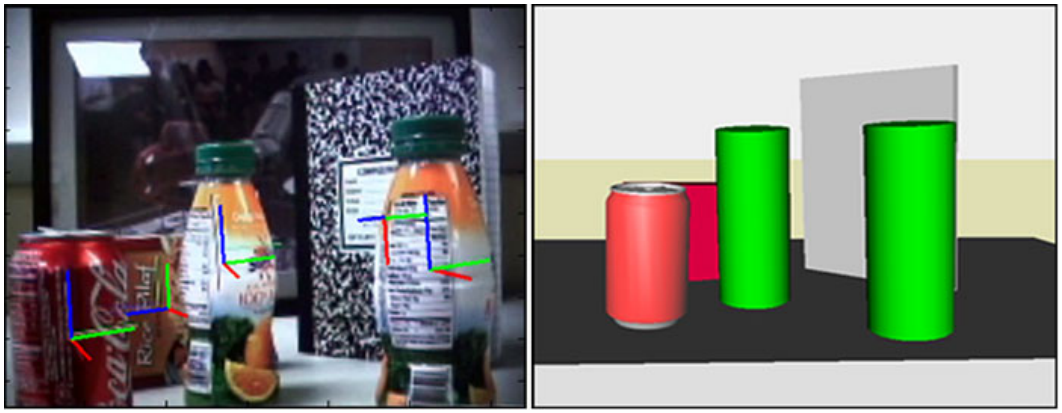
Jivko Sinapov
Assistant Professor
Tufts University
www.cs.tufts.edu/~jsinapov

Motivation: why multi-modal?

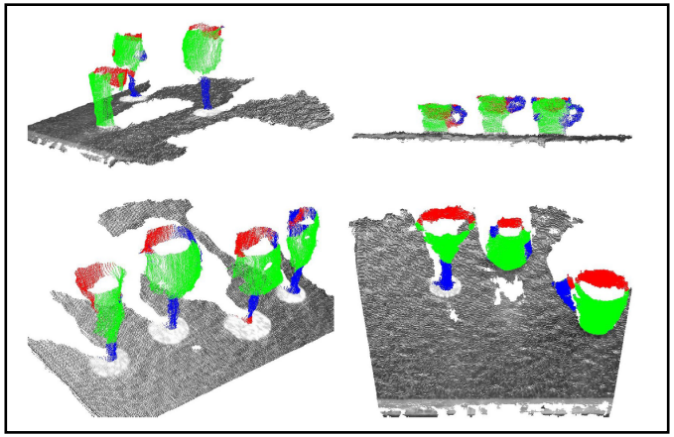
Visual Object Perception in Robotics



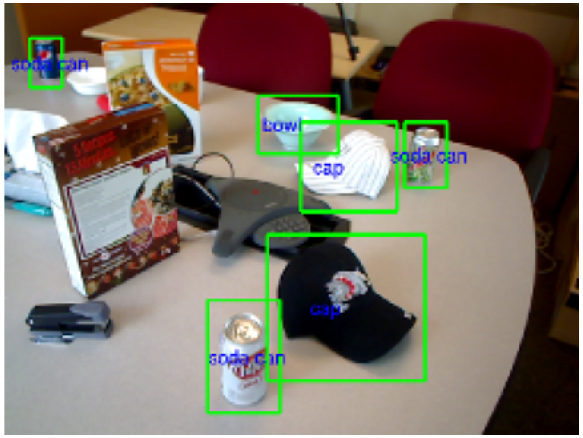
Sridharan *et al*. 2008



Collet *et al*. 2009



Rusu *et al*. 2009



Lai *et al*. 2011

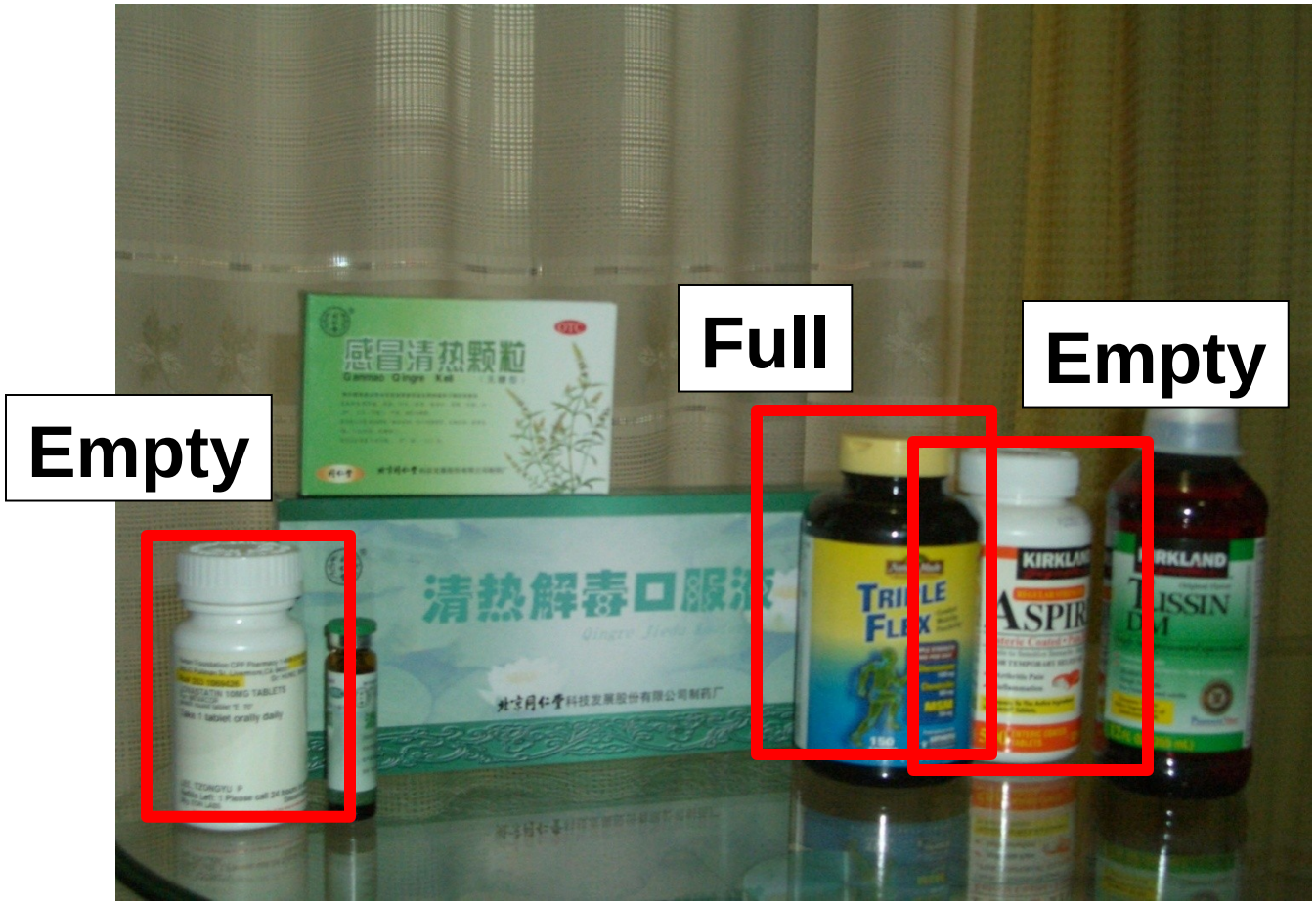
Why Exploratory Behaviors?

Why Exploratory Behaviors?

Solution: Lift the Object

Why Exploratory Behaviors?

Why Exploratory Behaviors?



Solution: Shake the Object

Solution: Shake the Object

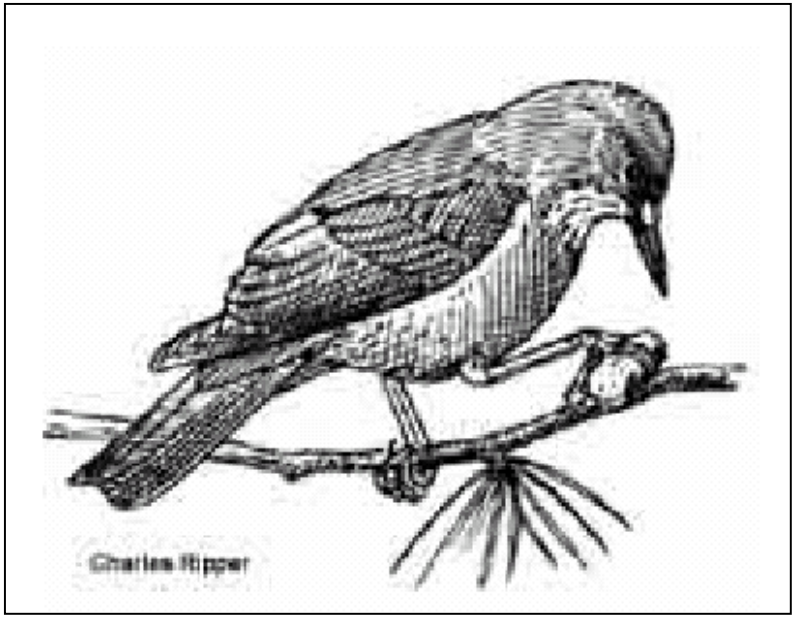
Exploratory behaviors give us information about objects that vision cannot!

KNOWLEDGE ABOUT OBJECT	EXPLORATORY PROCEDURE
Substance-related properties	
Texture	Lateral motion
Hardness	Pressure
Temperature	Static contact
Weight	Unsupported holding
Structure-related properties	
Weight	Unsupported holding
Volume	Enclosure, contour following
Global shape	Enclosure
Exact shape	Contour following
Functional properties	
Part motion	Part motion test
Specific motion	Function test

[Power, 2000]

[Lederman and Klatzky, 1987]

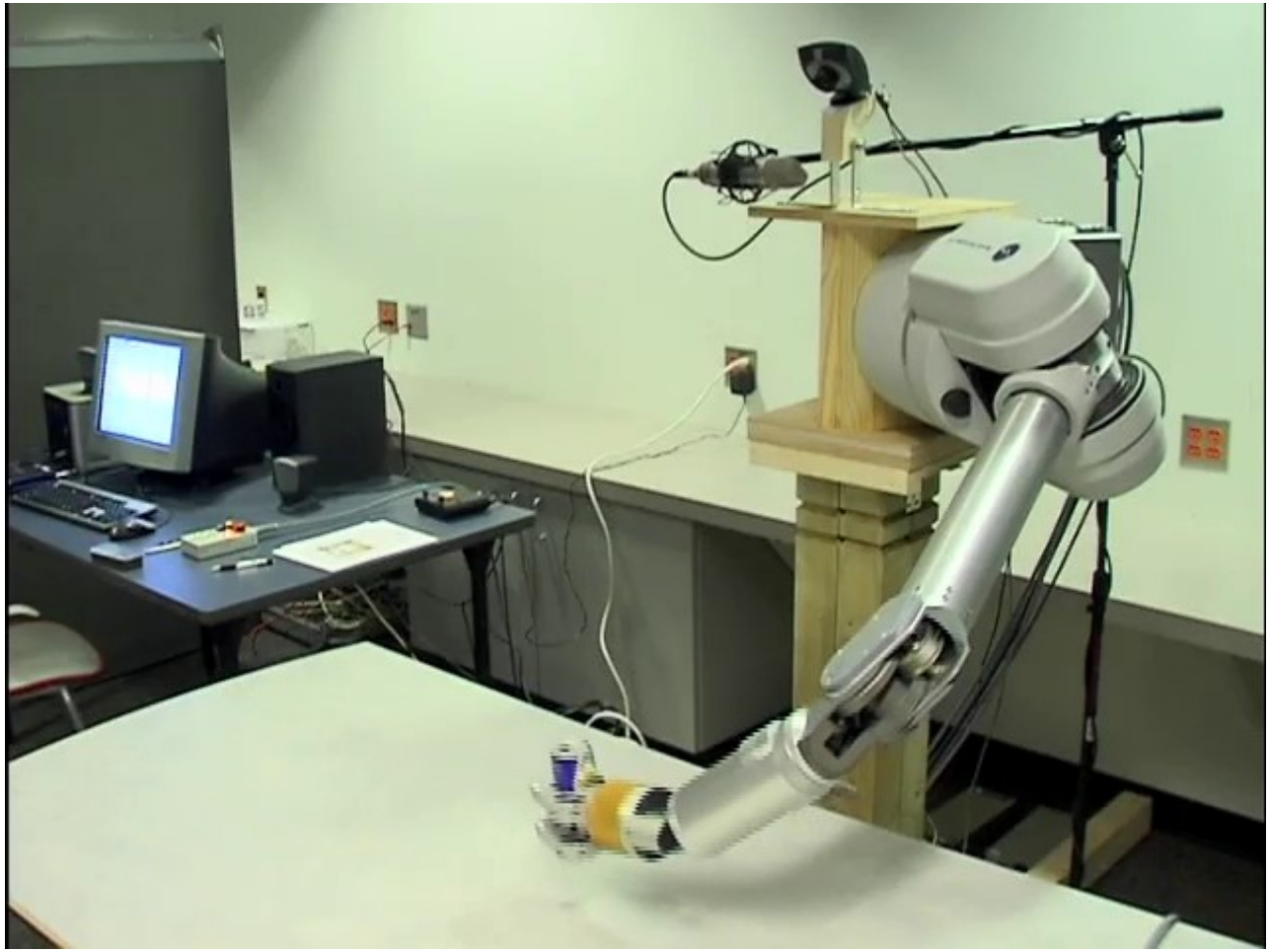
Object Exploration in the Wild



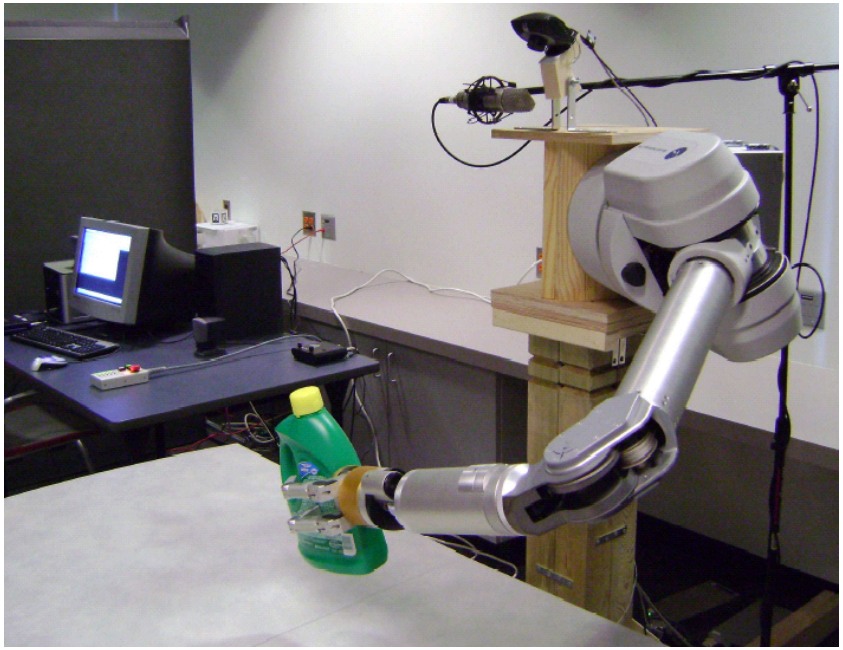
“A young corvide bird, confronted with an object it has never seen, runs through practically all of its behavioral patterns, except social and sexual ones.”

-Konrad Lorenz

Object Exploration in Infancy

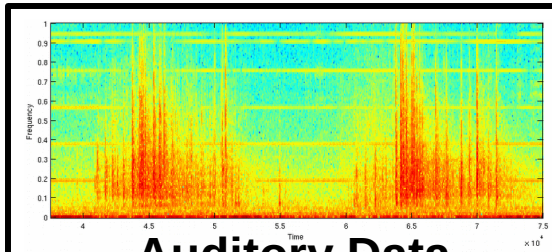


Can a robot use auditory perception to recognize objects?



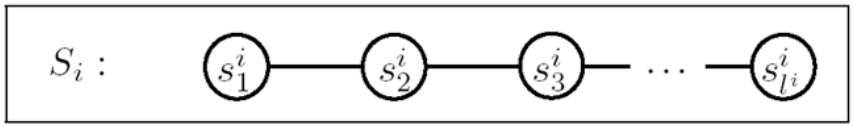
Sinapov *et al*. "Learning the Acoustic Properties of Household Objects"
2009 IEEE Int'l Conf. on Robotics and Automation

Acoustic Object Recognition



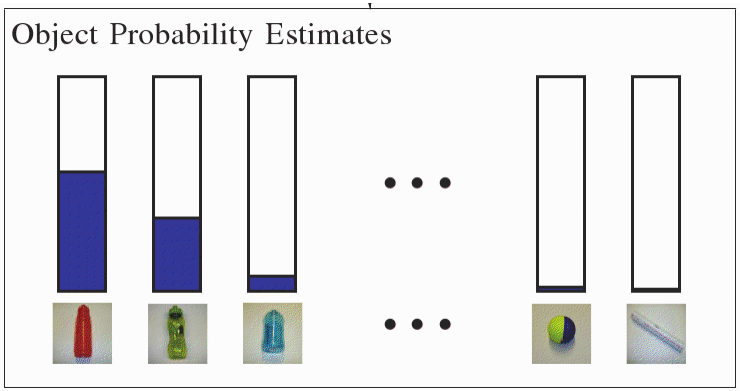
Auditory Data

Dimensionality Reduction

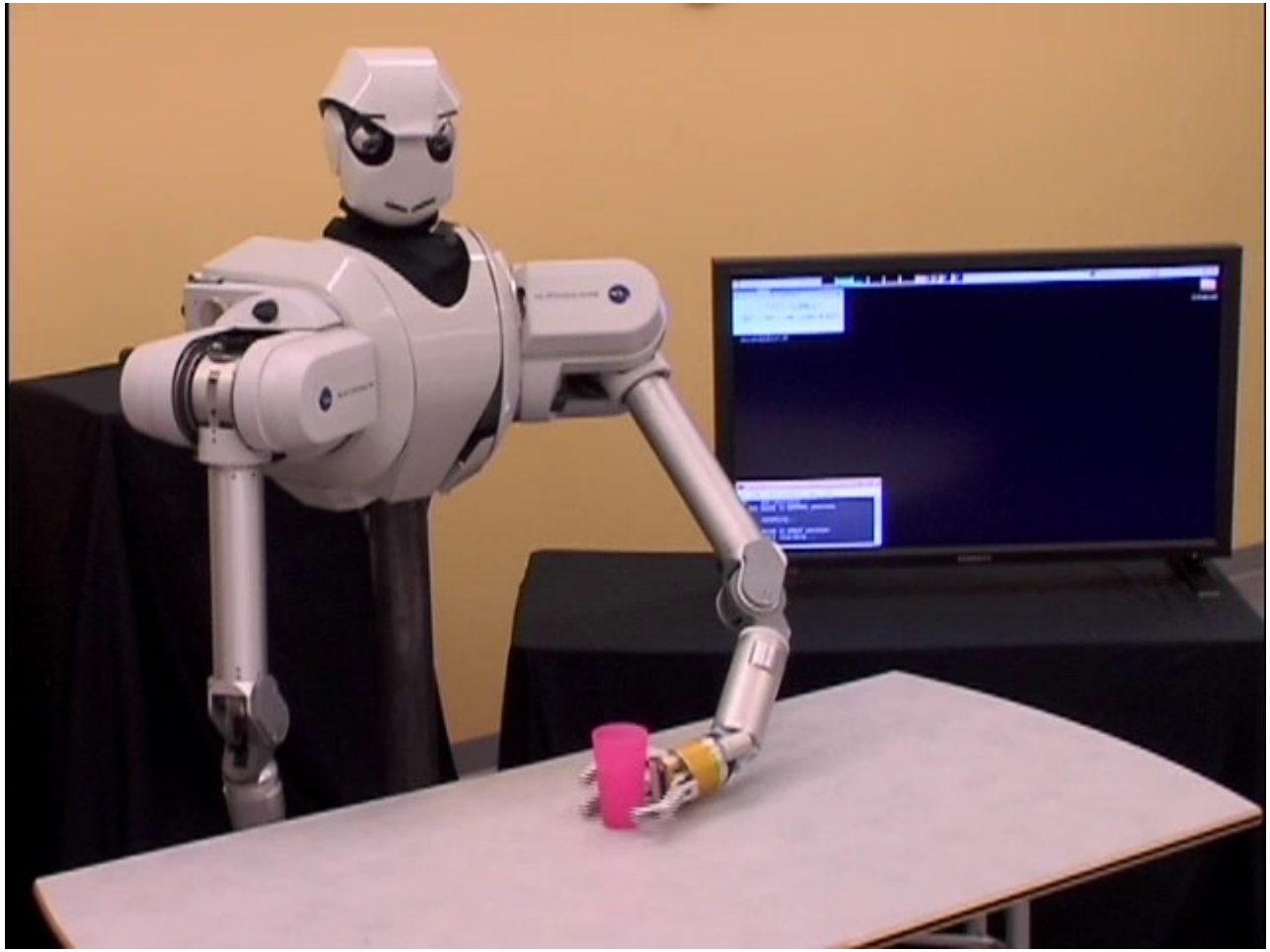


Discrete Auditory Sequence

Auditory Recognition Model



Object Probability Estimates



Object Recognition Results

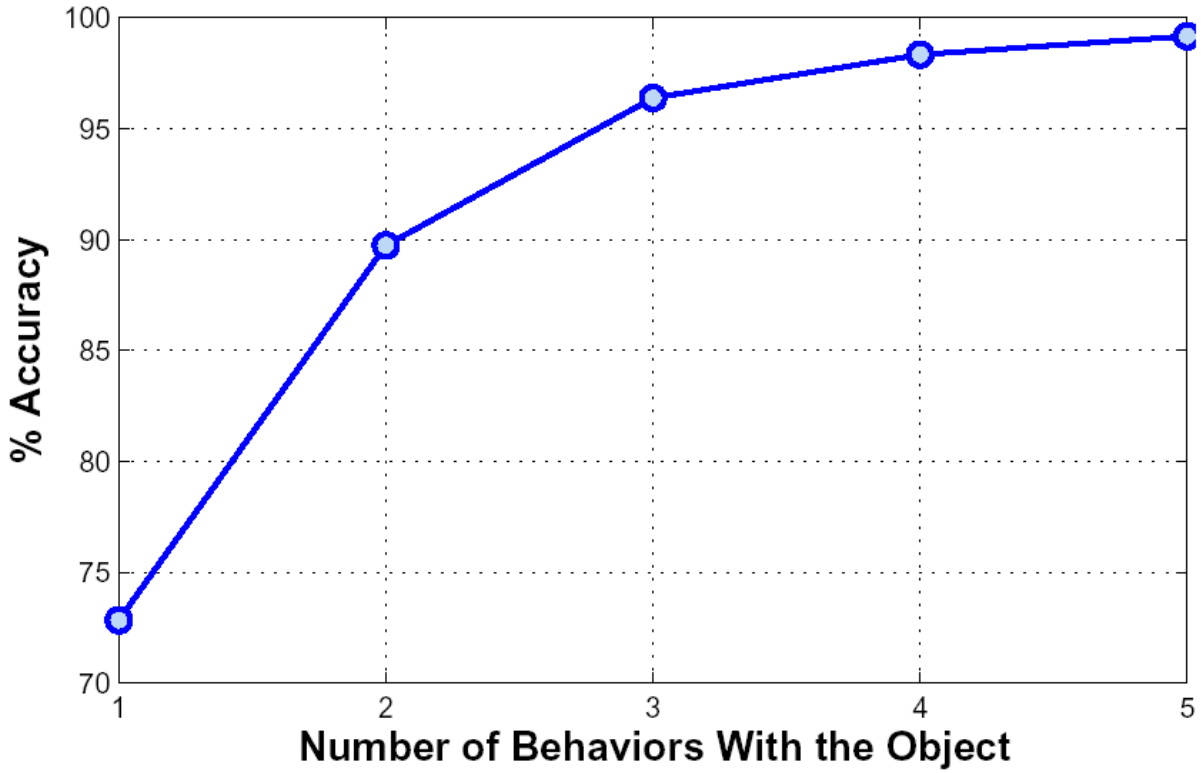
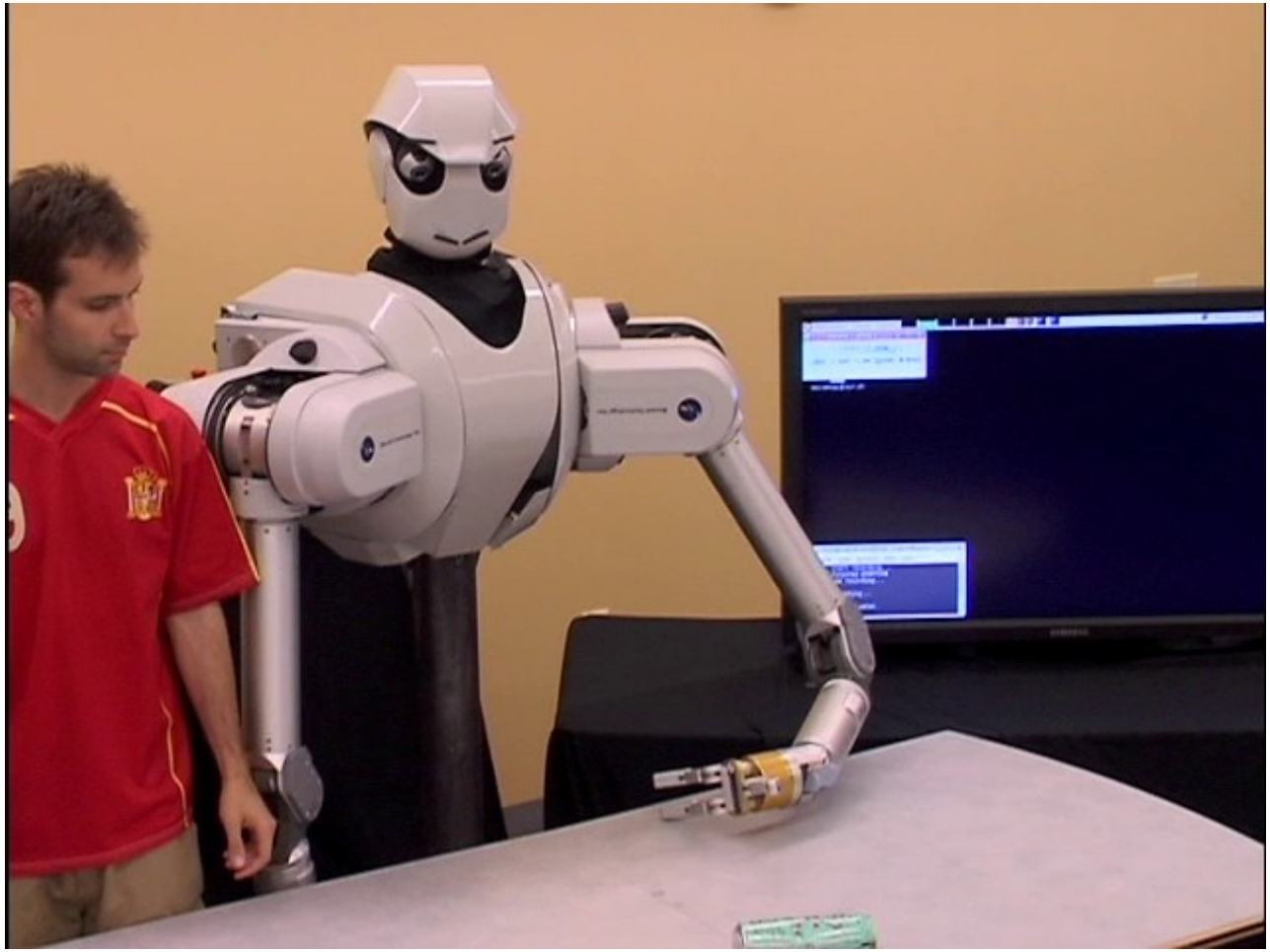
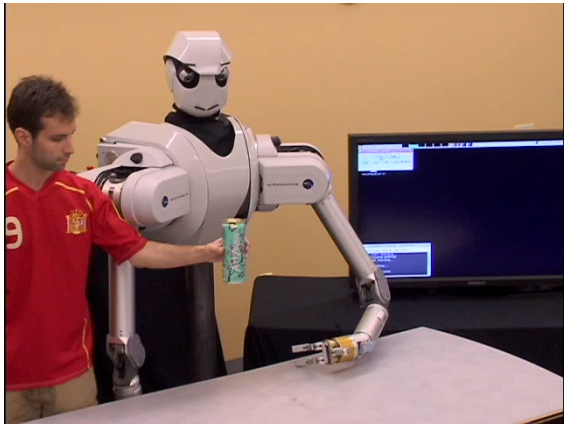
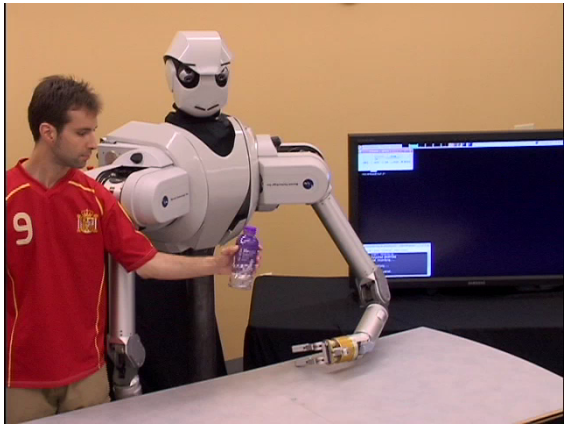
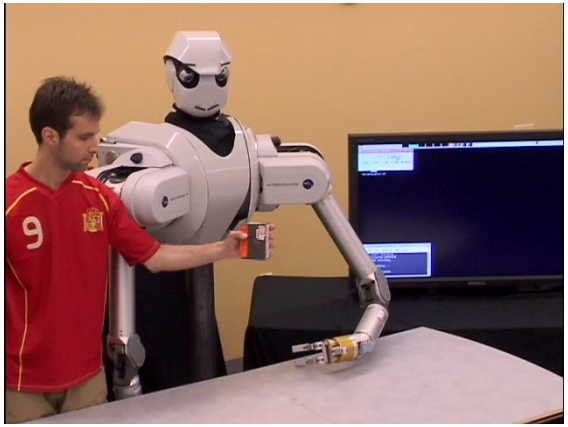


Fig. 6. Object recognition performance with k-Nearest Neighbor as the number of interactions with the object is varied from 1 (the default, used to generate Table I) to 5 (applying all five behaviors on the object).

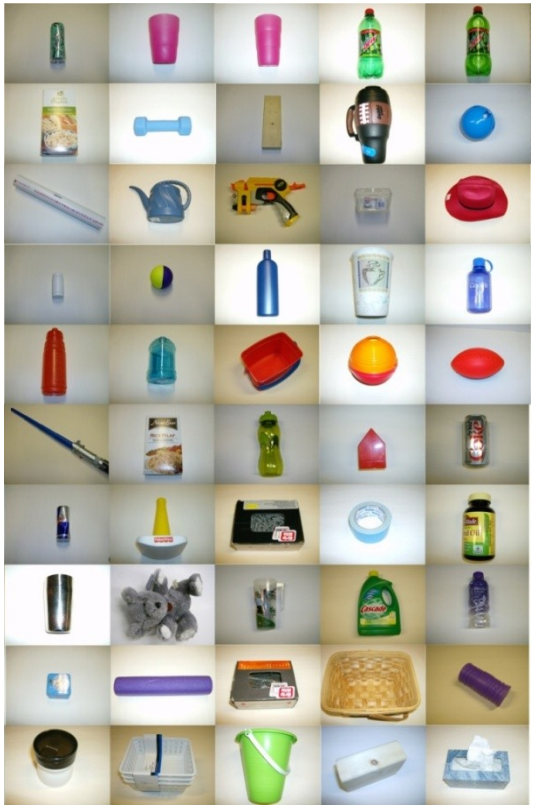


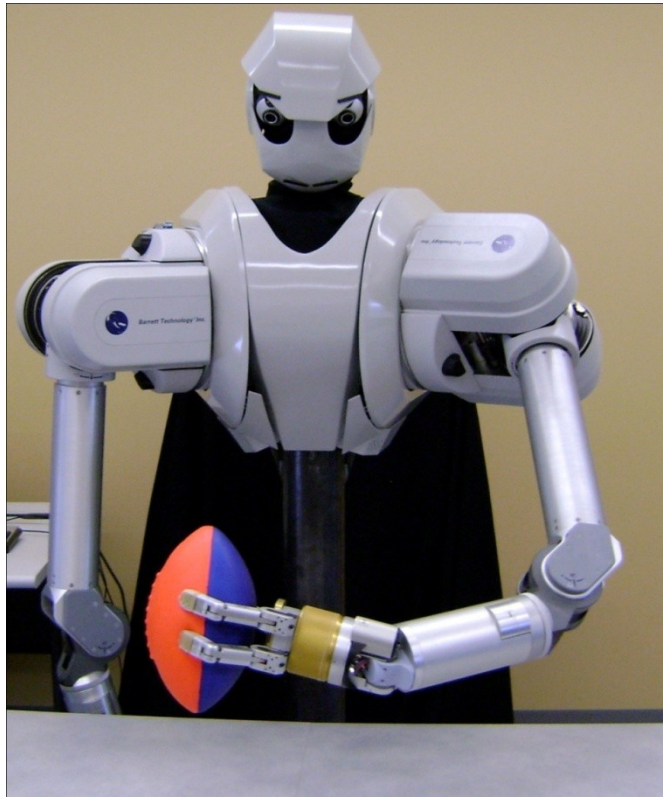




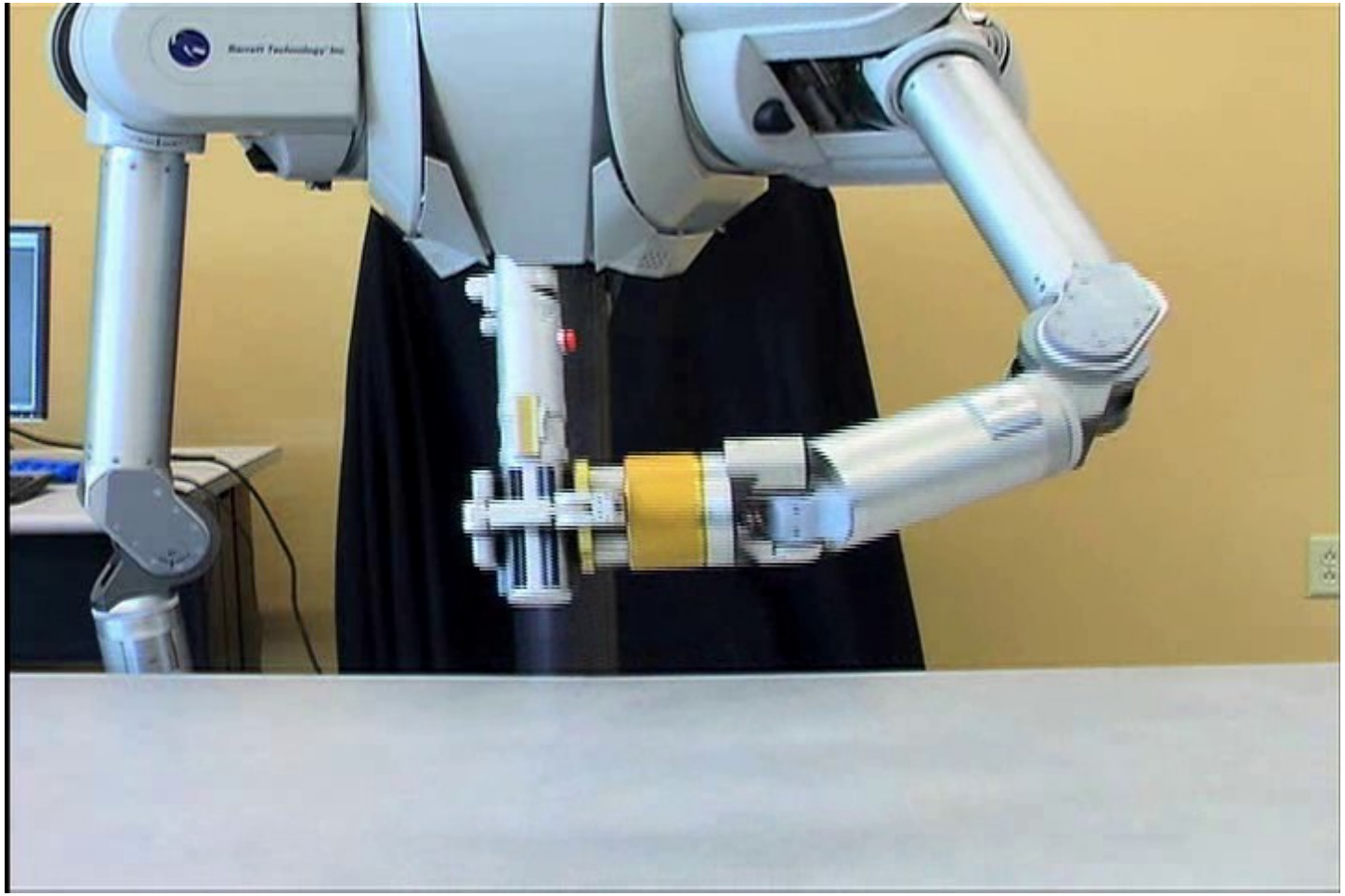


Scaling up: more objects + the proprioceptive sensory modality



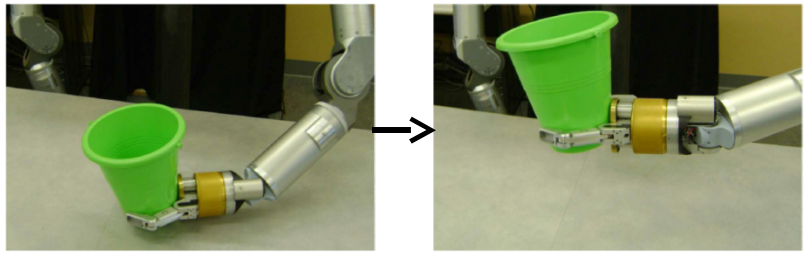


Sinapov *et al*. "Interactive Object Recognition using Proprioceptive and Auditory Feedback"
International Journal of Robotics Research, Vol. 30, No. 10, September 2011

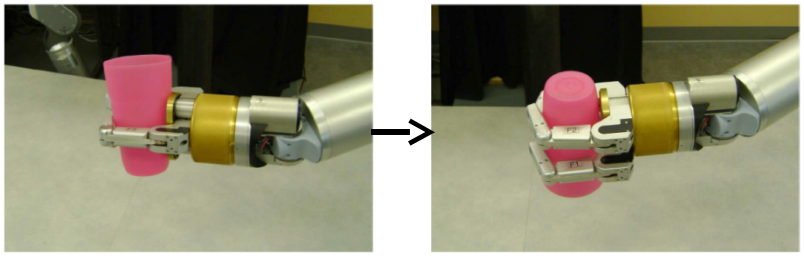

Barrett Technology Inc

Exploratory Behaviors

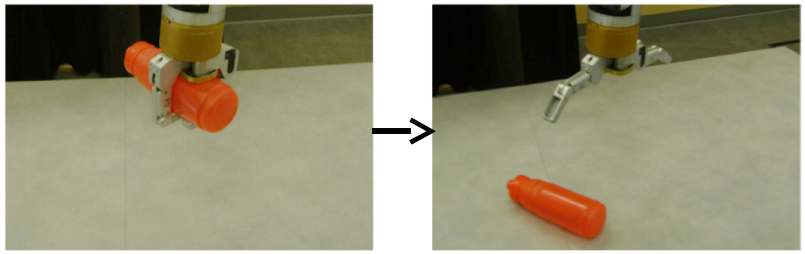
Lift:



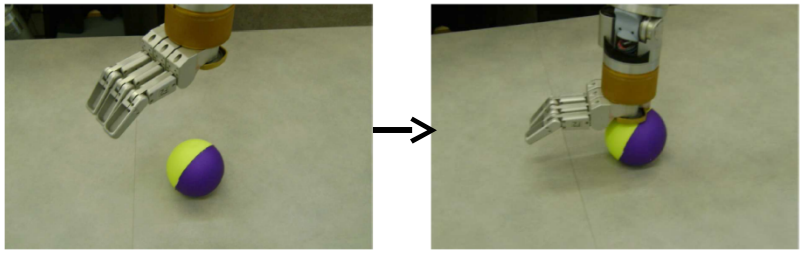
Shake:



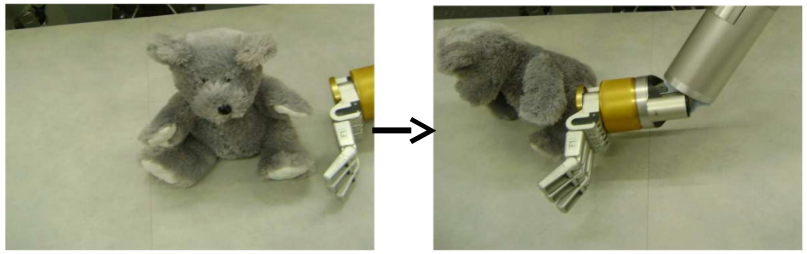
Drop:



Crush:

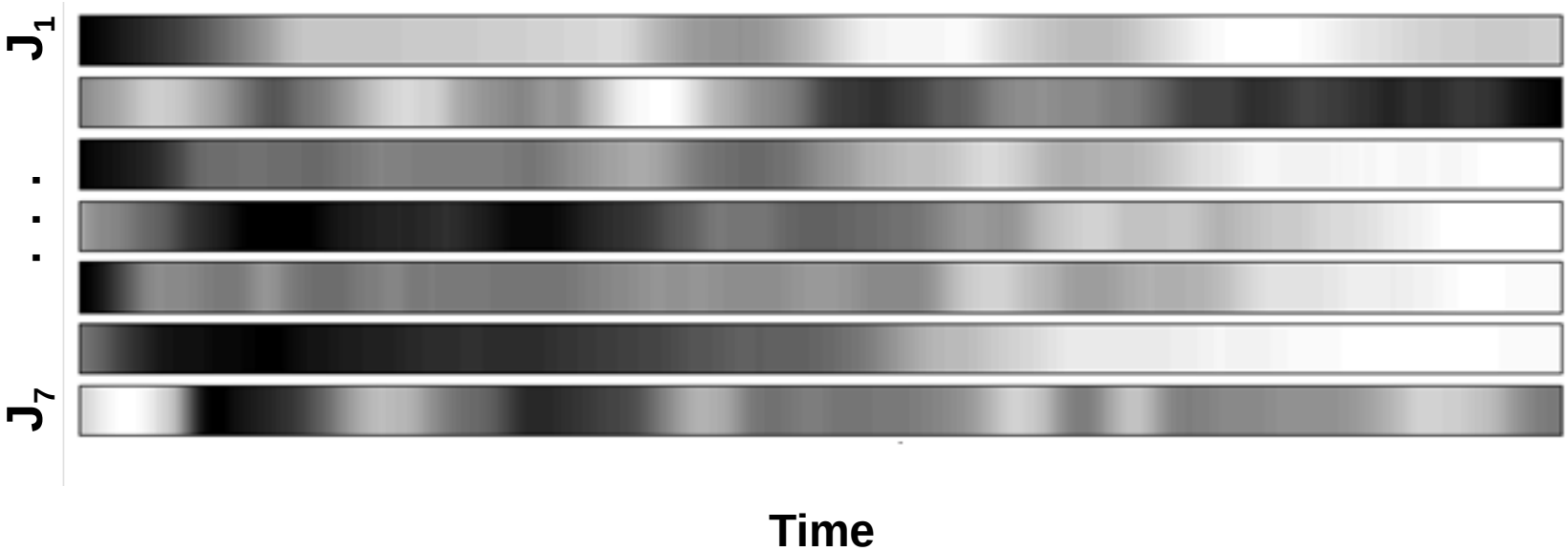


Push:



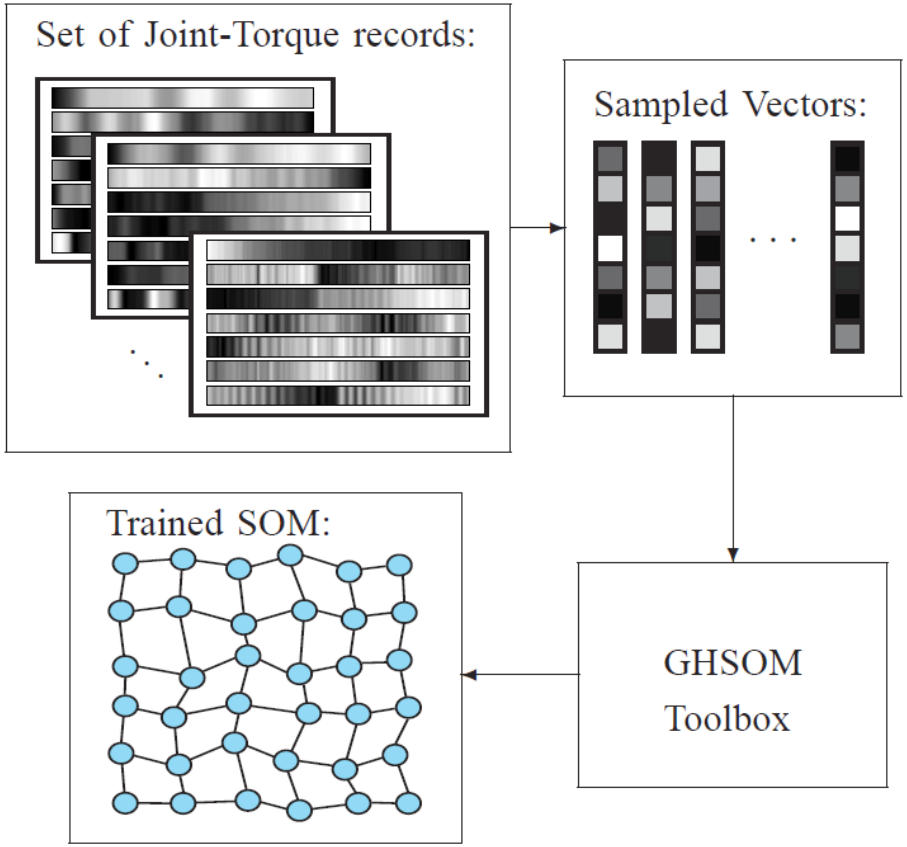
Objects

The Proprioceptive / Haptic Modality

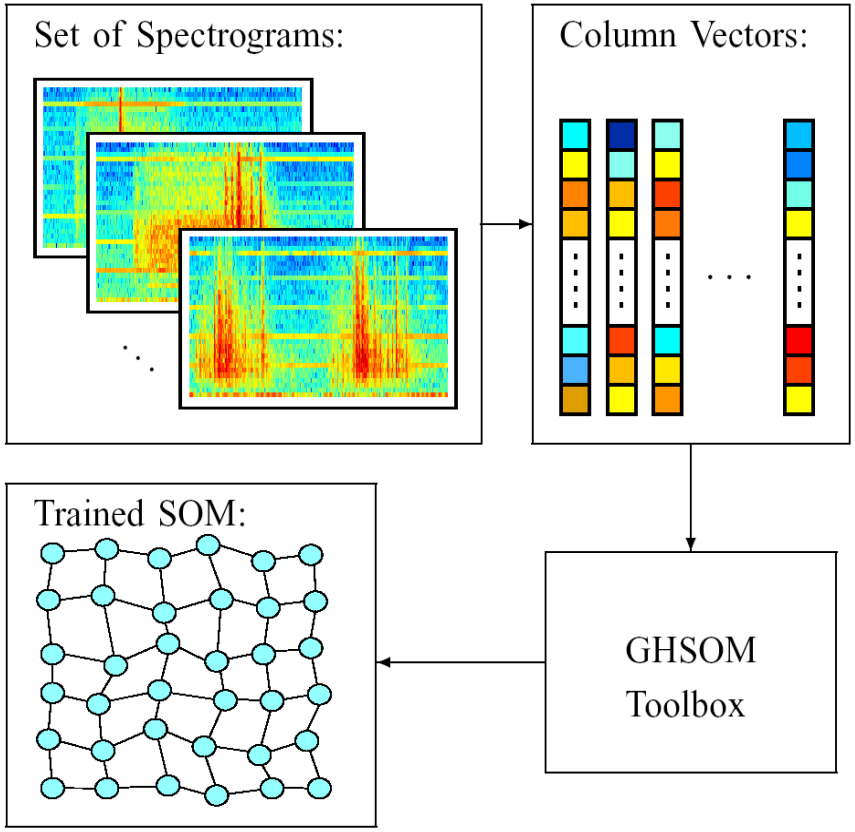


Feature Extraction

Training a self-organizing map (SOM) using sampled joint torques:

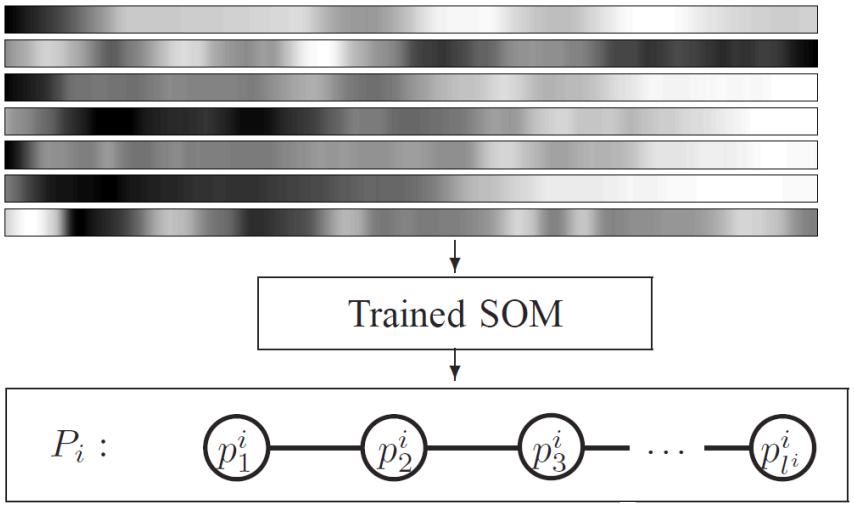


Training an SOM using sampled frequency distributions:

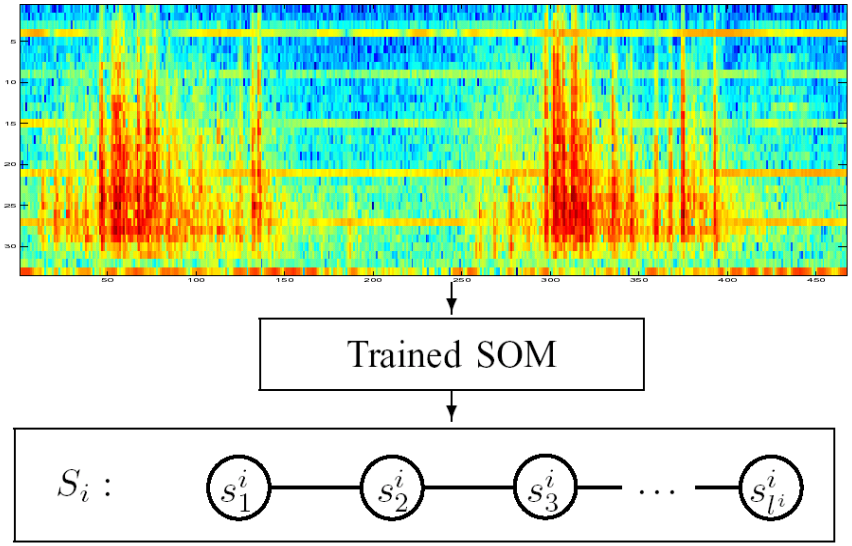


Feature Extraction

Discretization of joint-torque records using a trained SOM



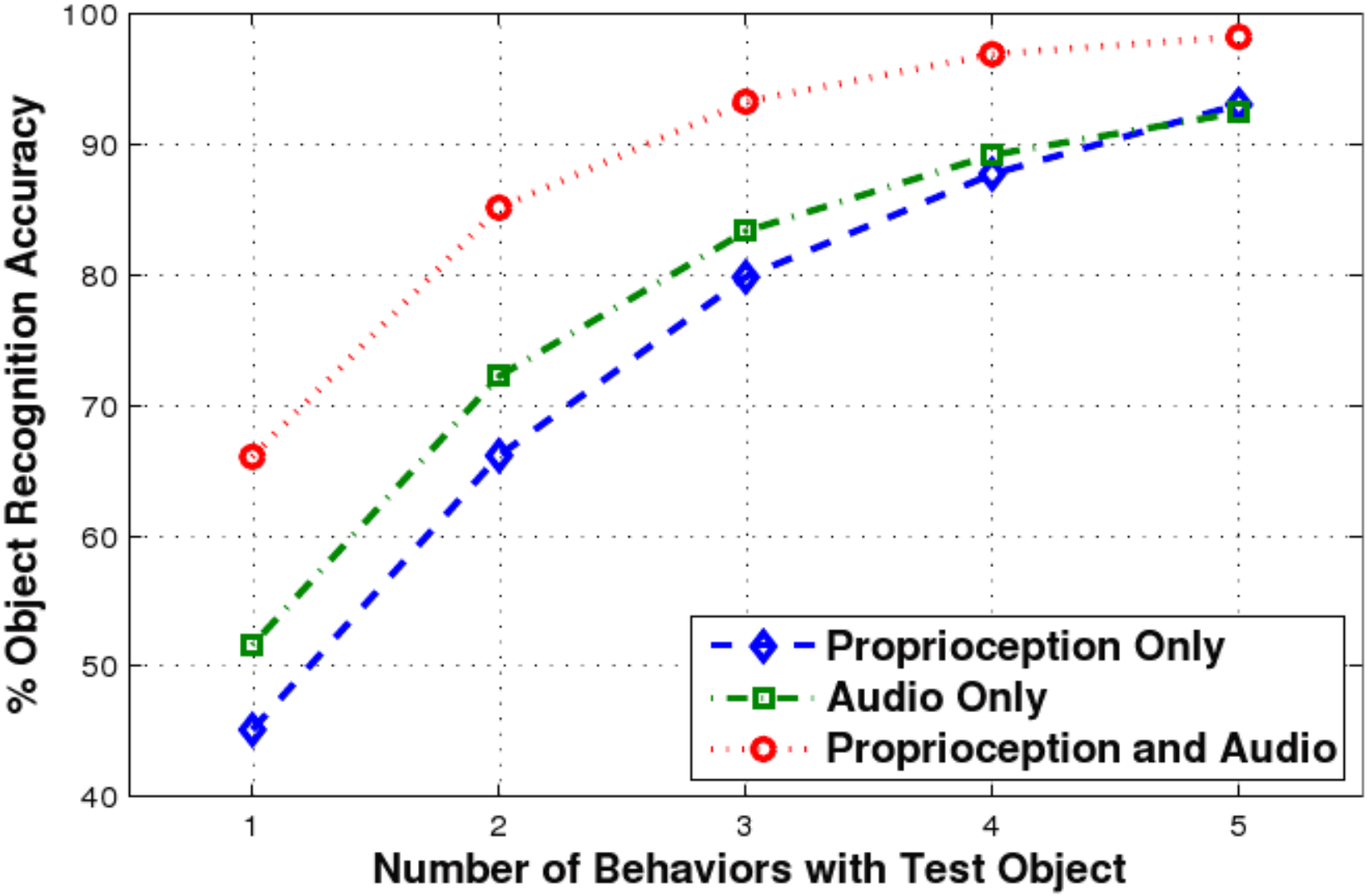
Discretization of the DFT of a sound using a trained SOM

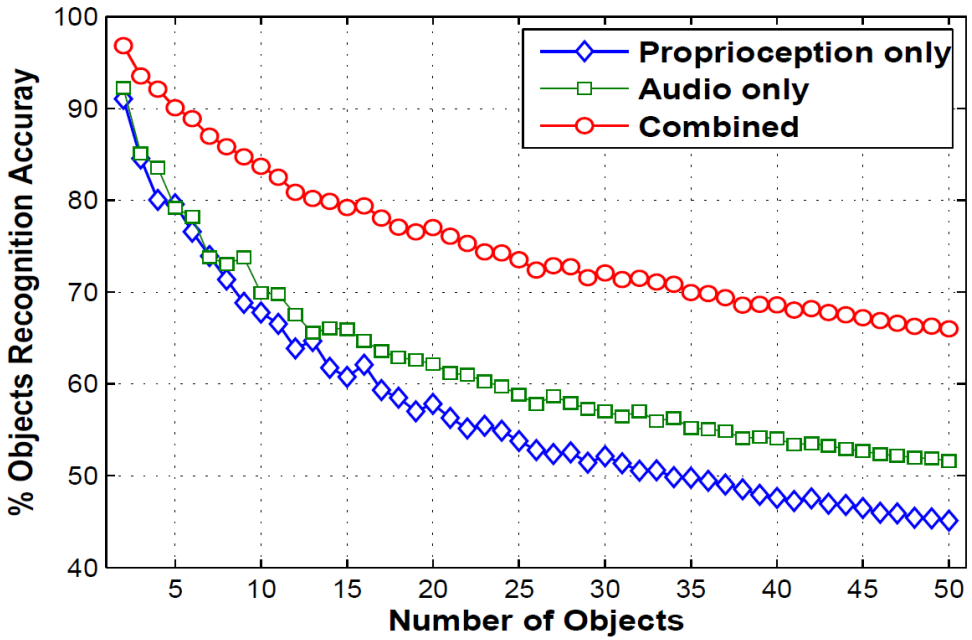


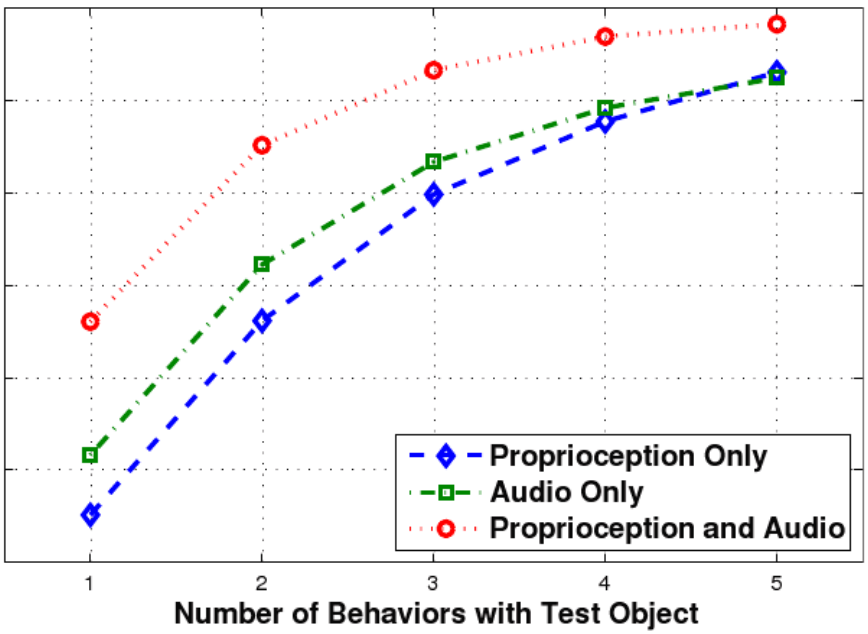
Sensorimotor Contexts

Behaviors \ Sensory Modalities	audio	proprioception
lift	✓	✓
shake	✓	✓
drop	✓	✓
press	✓	✓
push	✓	✓

Accuracy vs. Number of Behaviors

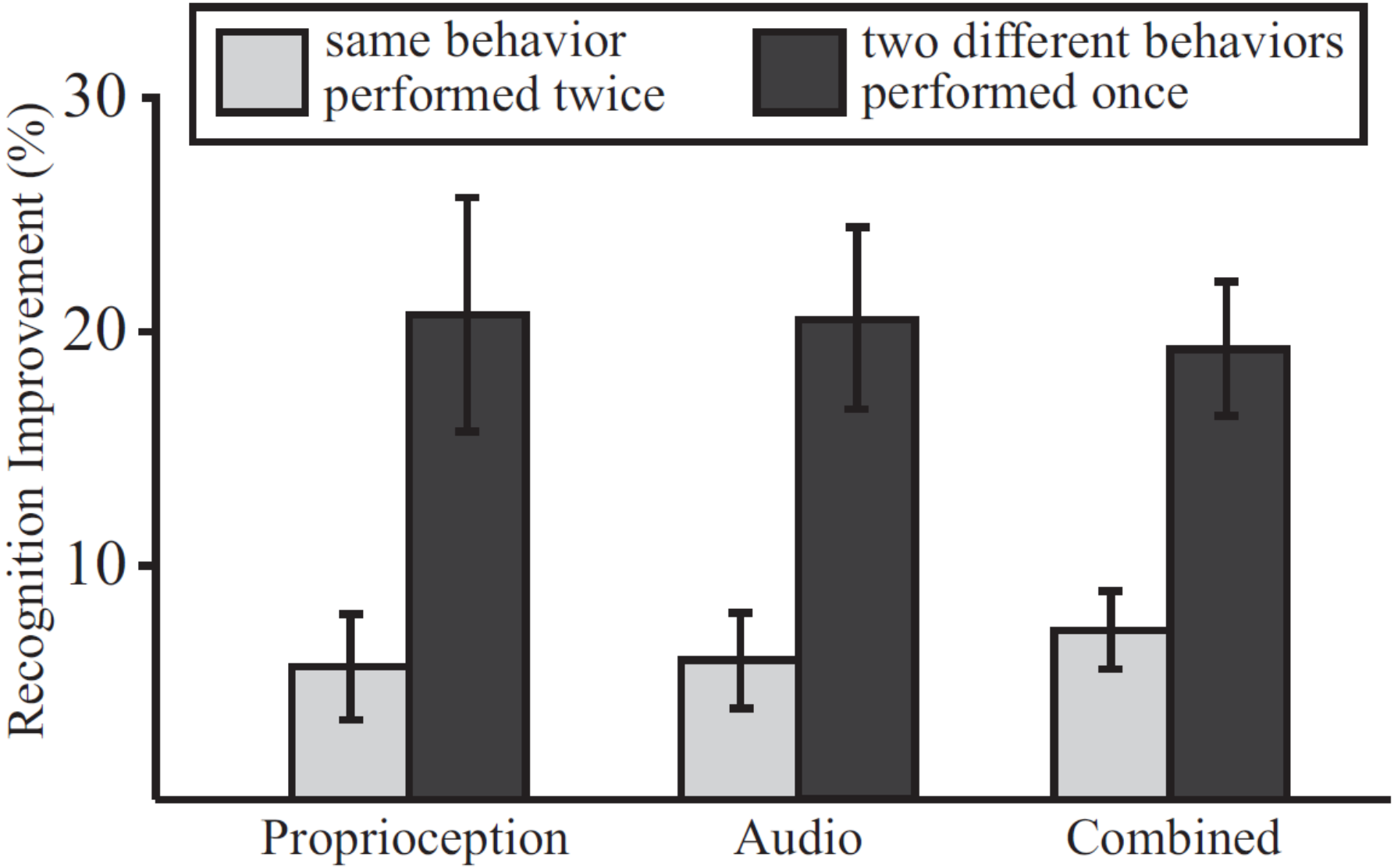






1 Behavior

Multiple Behaviors


same behavior performed twice
two different behaviors performed once
Recognition Improvement (%)
30
20
10
Proprioception
Audio
Combined

What can explain this boost?

- Machine Learning Theory:
  - Classification improvement is directly related to classifier diversity [Dietterich, 2000; Kuncheva *et al*, 2003]

- Our hypothesis:
  - The robot's collection of behavior-grounded recognition models acts as a classifier ensemble

Sinapov and Stoytchev, "The Boosting Effect of Exploratory Behaviors"
In Proceedings of AAAI, 2010

Take Home Message

- **Number of objects $\uparrow$ recognition accuracy $\downarrow$**
- **Number of behaviors $\uparrow$ recognition accuracy $\uparrow$**
- **Number of modalities $\uparrow$ recognition accuracy $\uparrow$**
- **Behaviors are classifiers that can be boosted**

Developmental Progression

Developmental Progression of Objects Knowledge

Categorization: infants can learn object categories based on perceptual object properties

Recognition: infants can perform visual object recognition.

Individuation: infants can individuate objects. Furthermore, naming objects enhances infant object individuation.

Birth → **4 mo** | **5 mo** | **6-10 mo** | **11-20 mo** →

Object Permanence: infants can learn that an object moving behind an occluder remains the same object after it reappears

Recognition: infants can discriminate textures by touch. In the absence of visual cues, infants use more efficient exploratory strategies.

Categorization: infants begin to form categories based on abstract properties such as the objects' function and their spatial relations.

Developmental Progression

Beyond Recognition: Category Acquisition

pots

pans

cups

silverware

glasses

Categorizing a novel object

pots

pans

cups

?

silverware

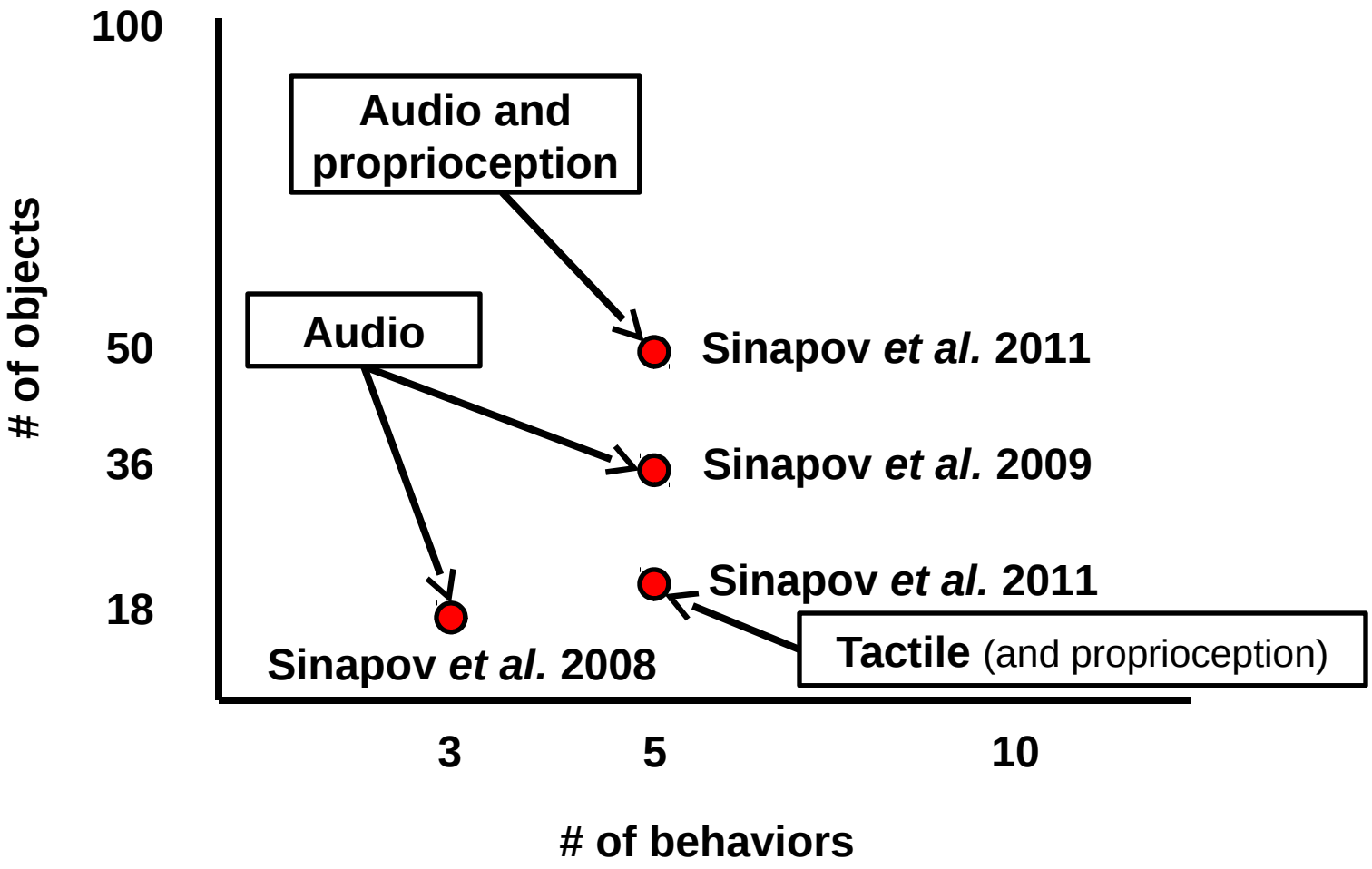
glasses

The Symbol Grounding Problem

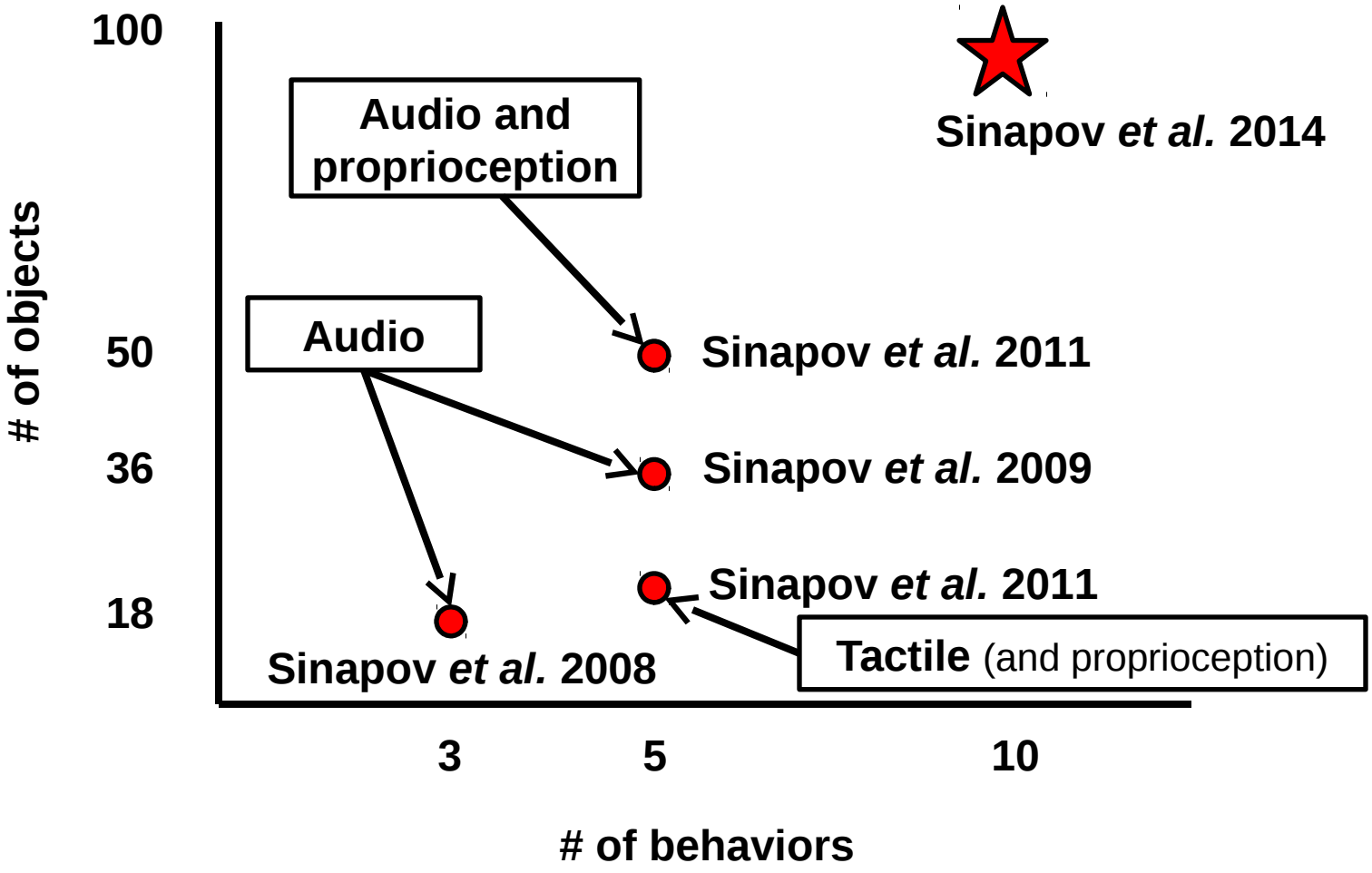
GOFAI was the dominant paradigm of AI research from the middle 1950s until the late 1980s. The Symbol Grounding Problem is related to the problem of how words (symbols) get their meanings, and hence to the problem of what meaning itself really is.

If symbols (words) always are explained with other symbols we get infinite regress. Somewhere symbols must be “grounded”! In what way does that grounding happen?

Progression of dataset sizes



Progression of dataset sizes



Experimental Setup

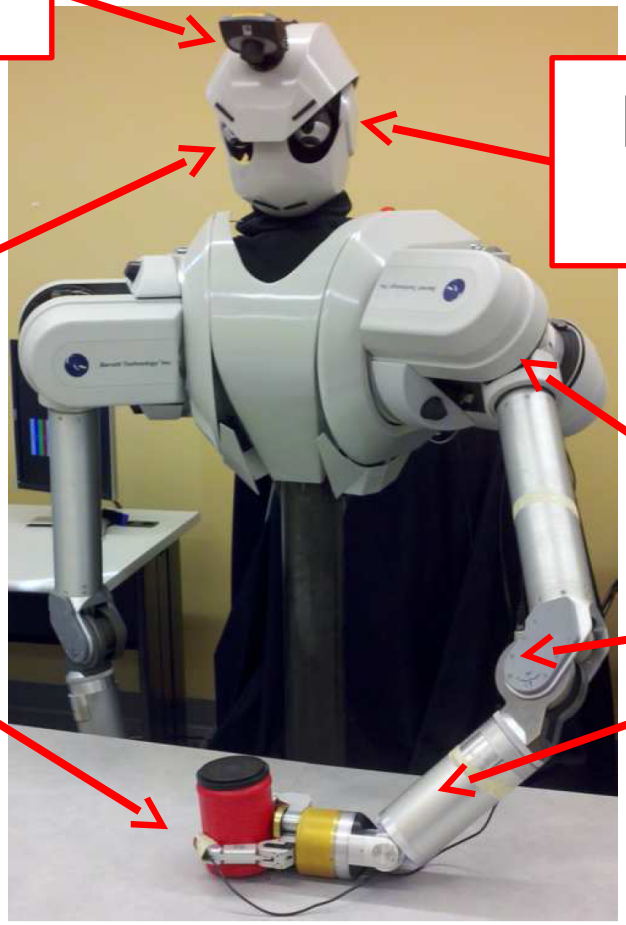
Sinapov *et al*, "Grounding Semantic Categories in Behavioral Interactions: Experiments with 100 Objects."
Robotics and Autonomous Systems, Vol. 62, No. 5, pp. 632-645, May 2014.

Experimental Platform

ZCam (RGB+D)

Logitech Webcam

3-axis accelerometer



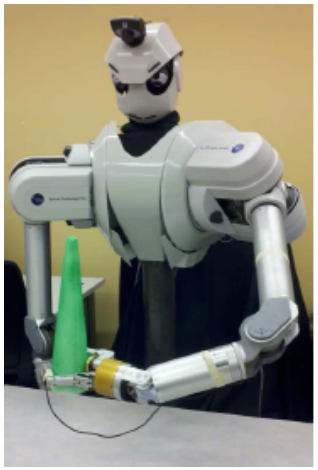
Microphones in the head

Torque sensors in the joints

100 objects

Exploratory Behaviors

grasp



lift

hold

shake

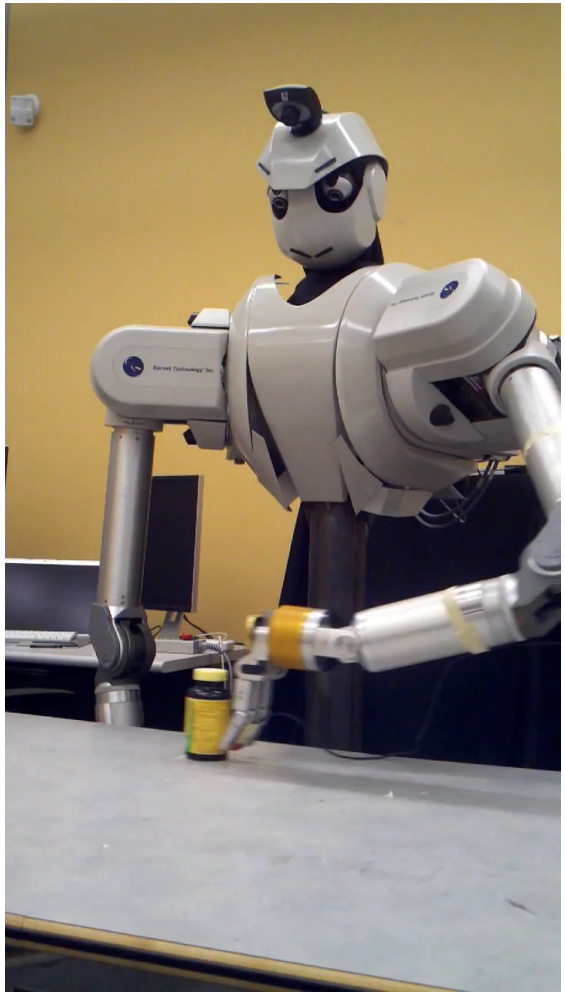
drop

tap

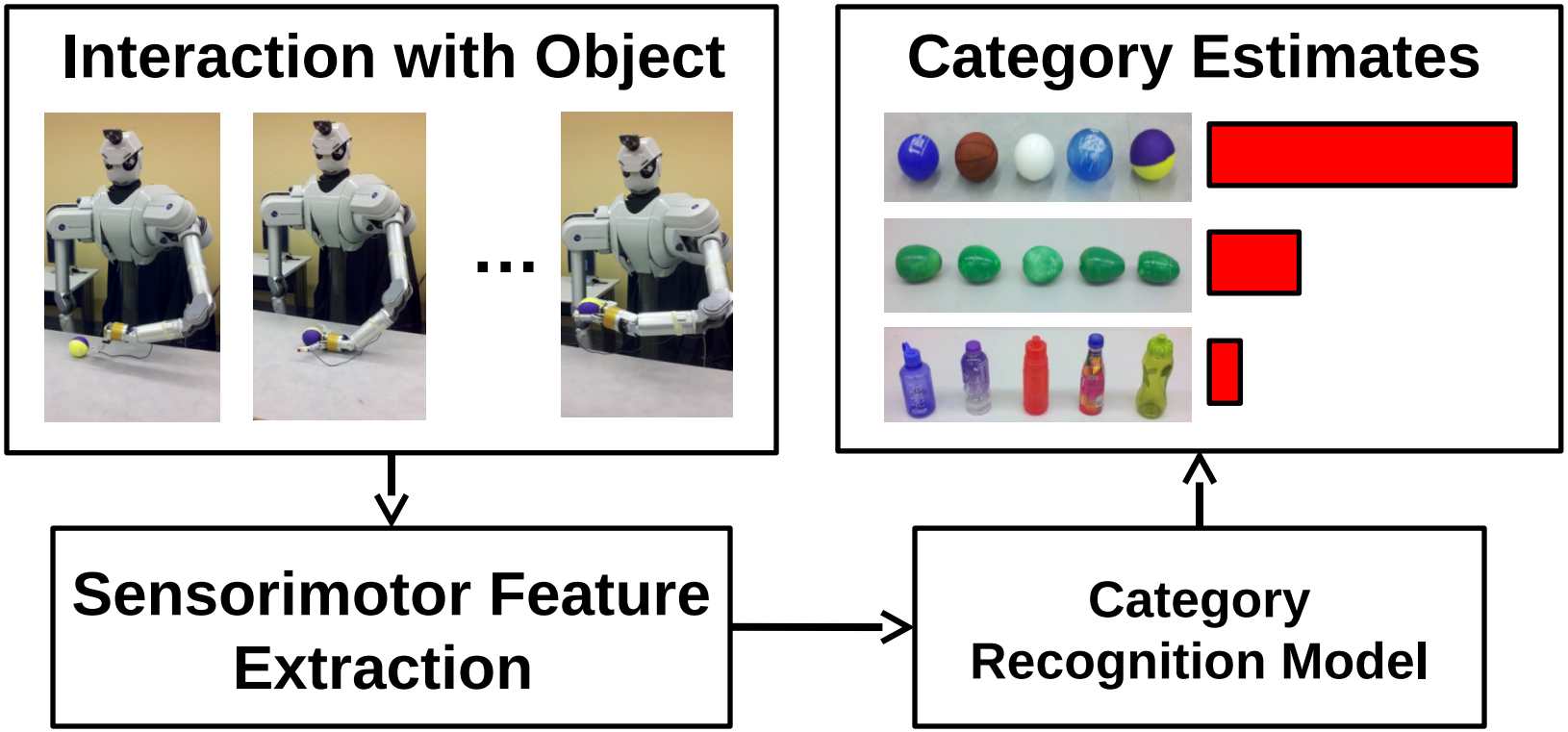
poke

push

press

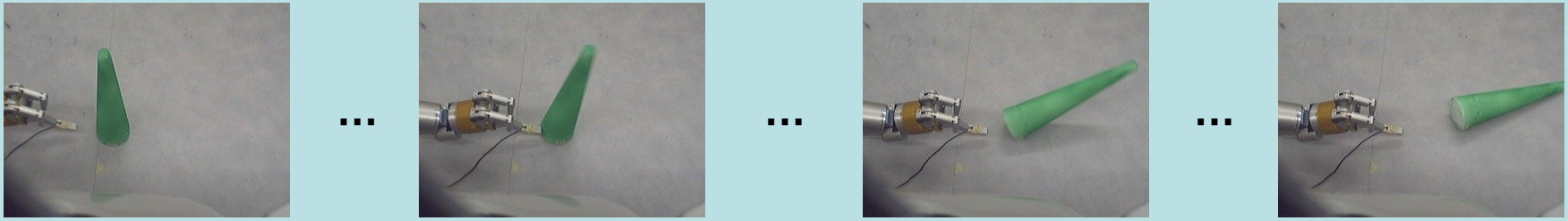


Overview

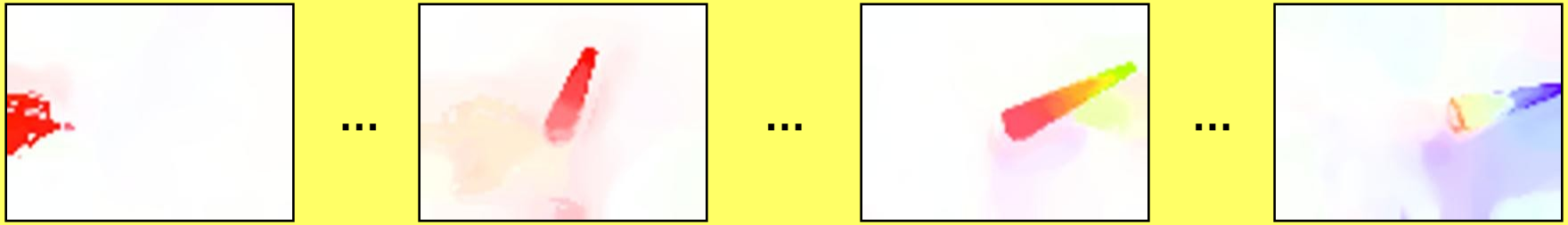

Interaction with Object
...
Category Estimates
Sensorimotor Feature Extraction
Category Recognition Model

Coupling Action and Perception

Action: poke



Perception: optical flow



Time

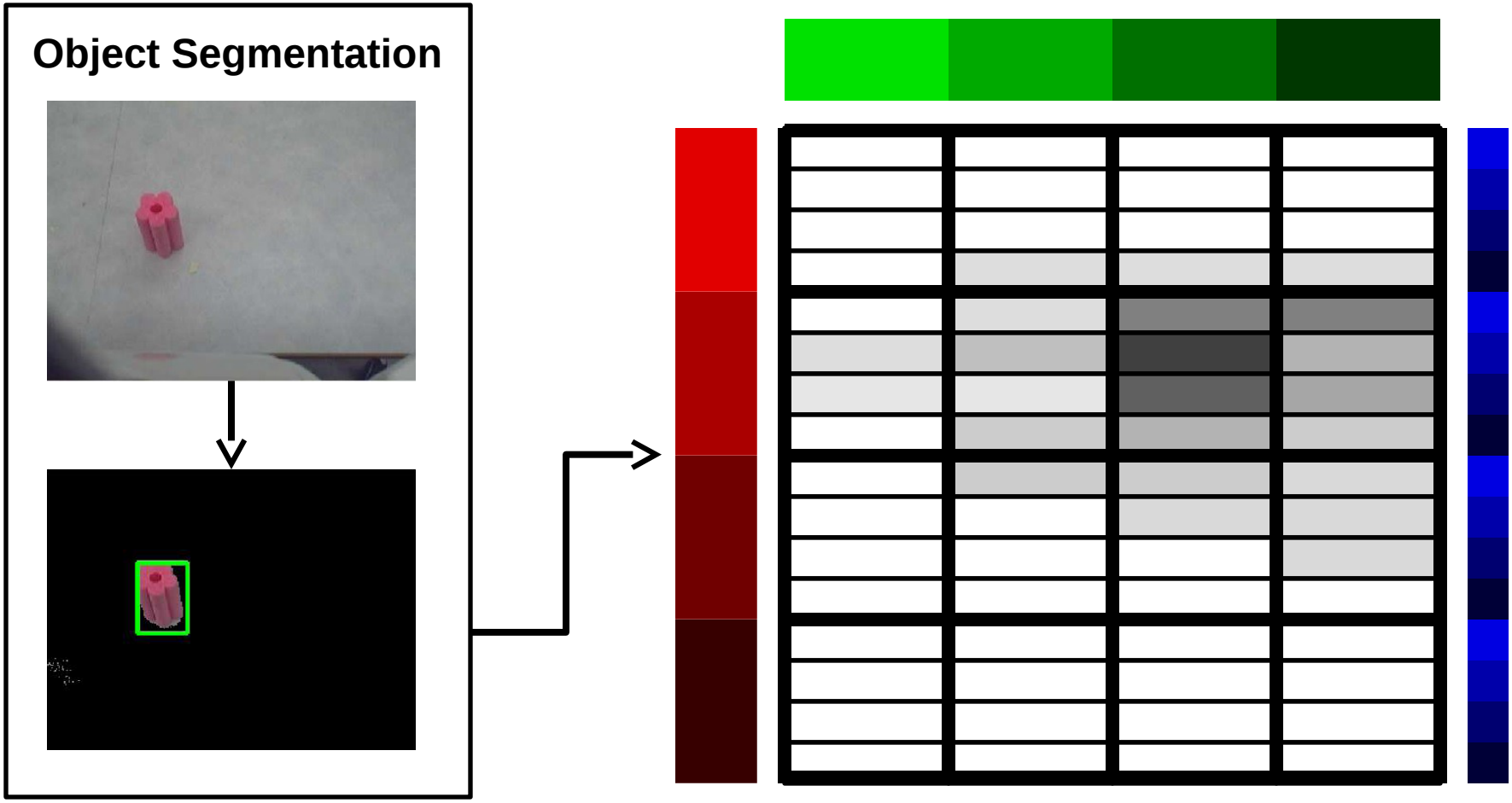
Sensorimotor Contexts

	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look						
grasp						
lift						
hold						
shake						
drop						
tap						
poke					✓	
push						
press						

Sensorimotor Contexts

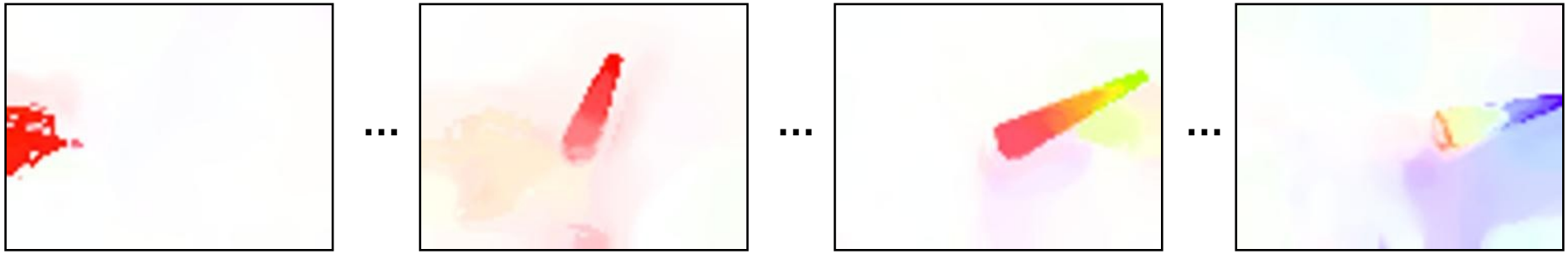
	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look				✓		✓
grasp	✓	✓	✓		✓	✓
lift	✓	✓			✓	✓
hold	✓	✓			✓	✓
shake	✓	✓			✓	✓
drop	✓	✓			✓	✓
tap	✓	✓			✓	✓
poke	✓	✓			✓	✓
push	✓	✓			✓	✓
press	✓	✓			✓	✓

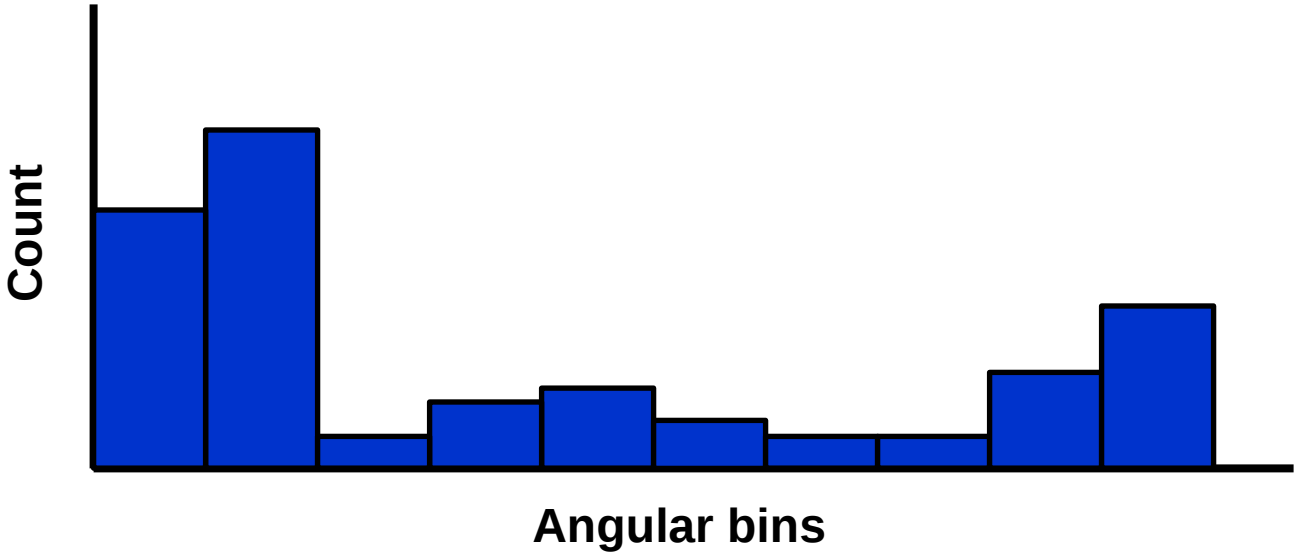
Feature Extraction: Color



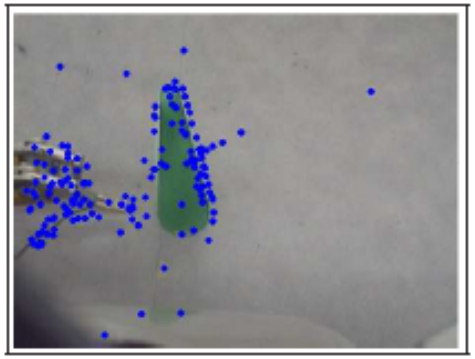
Color Histogram (4 x 4 x 4 = 64 bins)

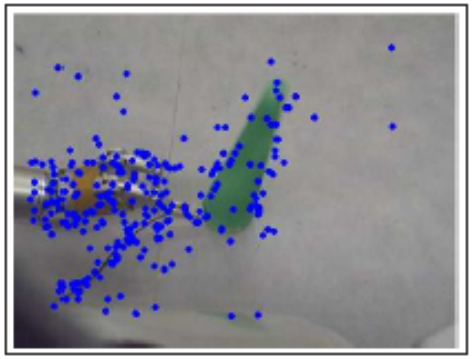
Feature Extraction: Optical Flow

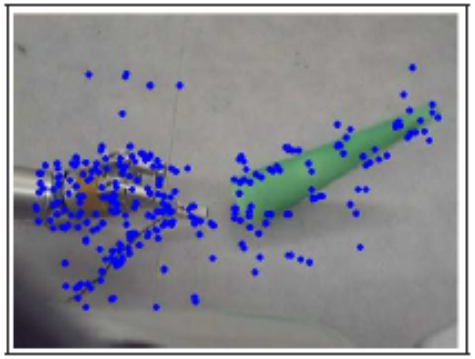




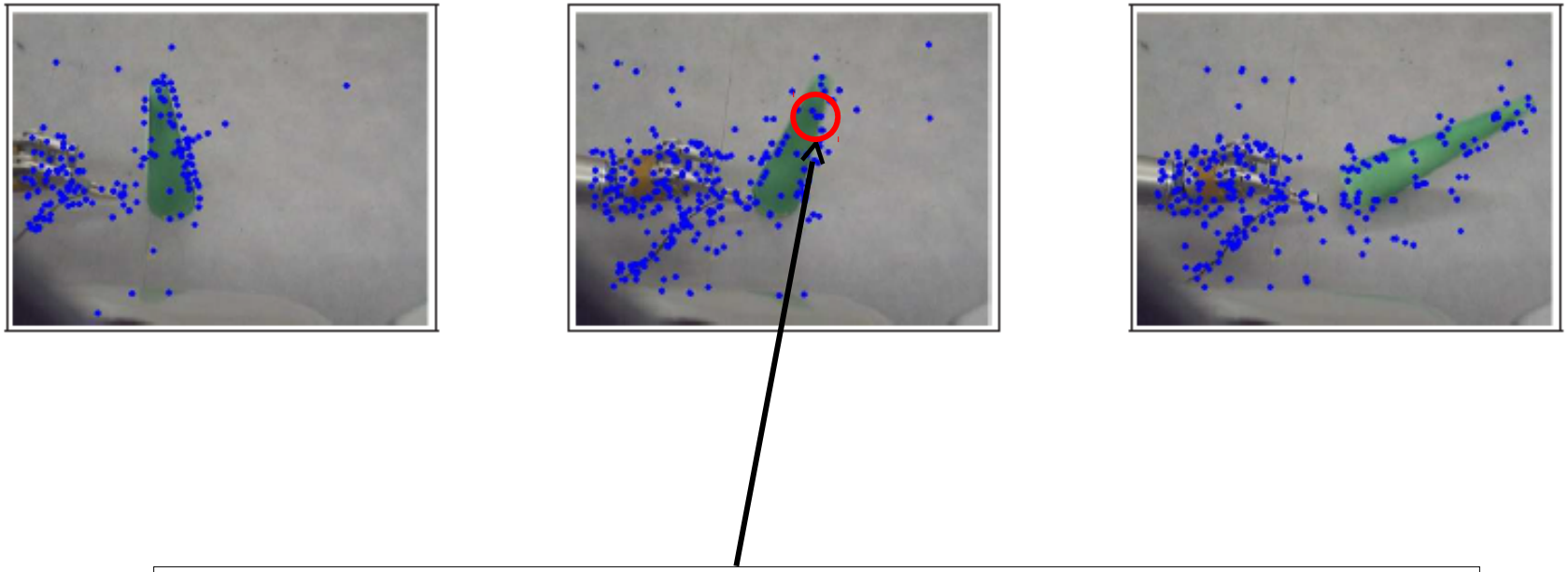
Feature Extraction: SURF





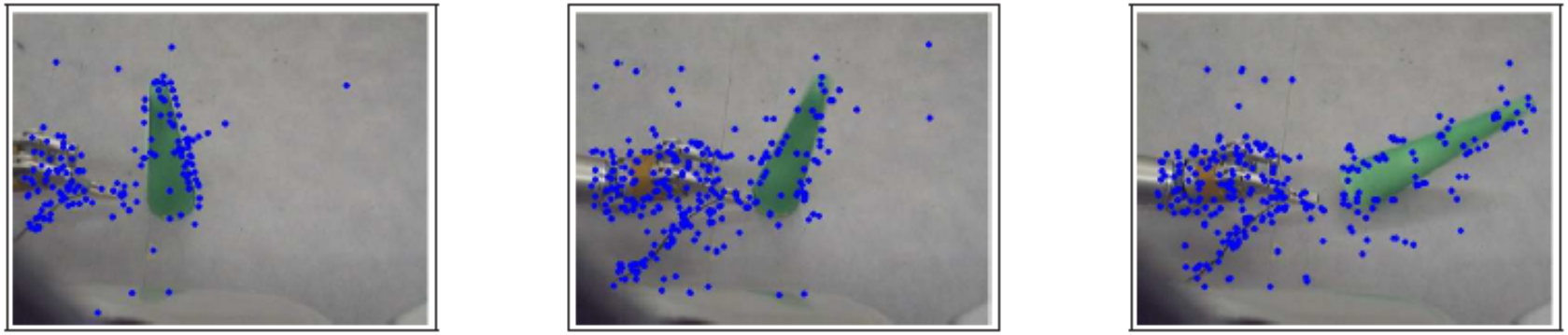


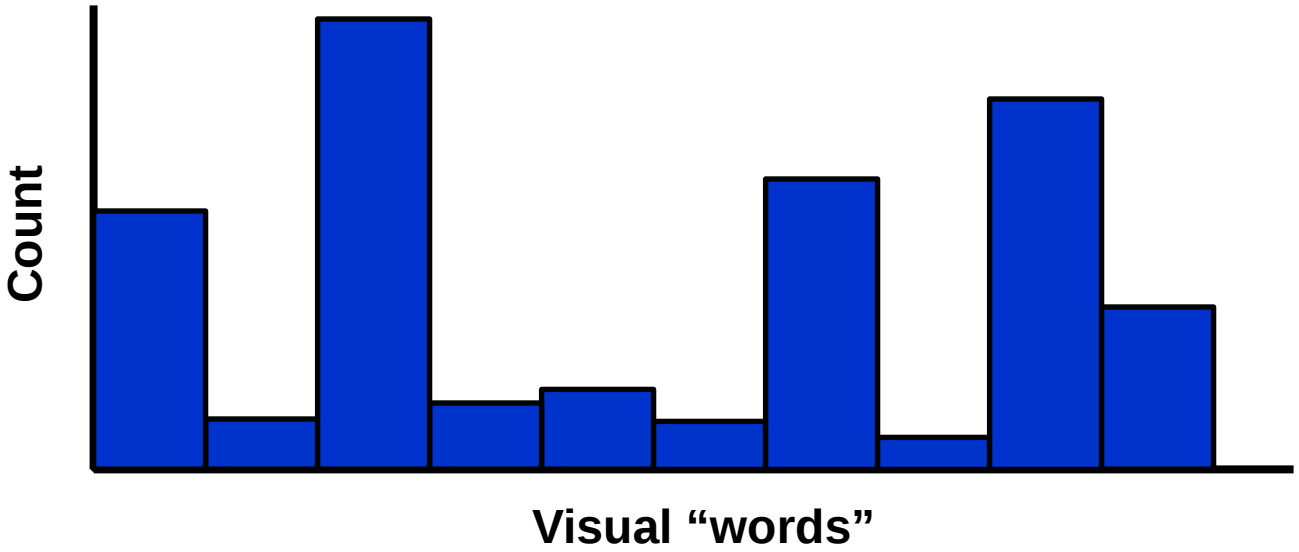
Feature Extraction: SURF



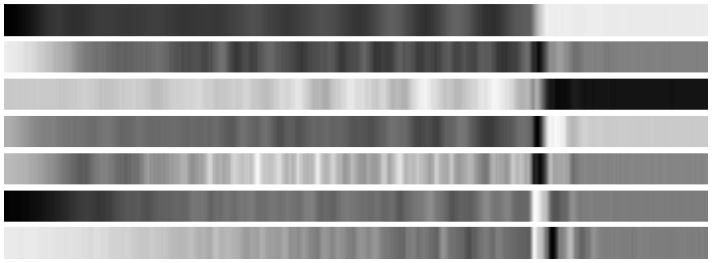
Each interest point is described by a 128-dimensional vector

Feature Extraction: SURF

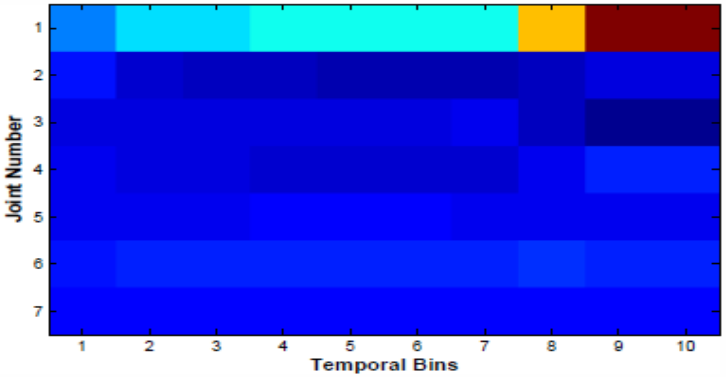




Feature Extraction: Proprioception

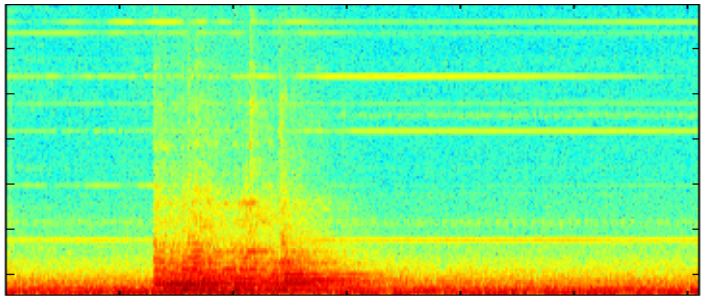


Joint-Torque values for all 7 Joints

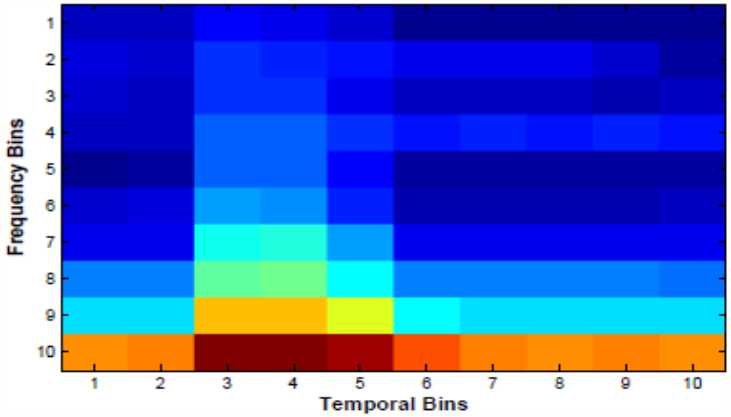


Joint-Torque Features

Feature Extraction: Audio



audio spectrogram



Spectro-temporal Features

Dimensionality of Data

audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
100	70	6	64	10	200

Data From a Single Exploratory Trial

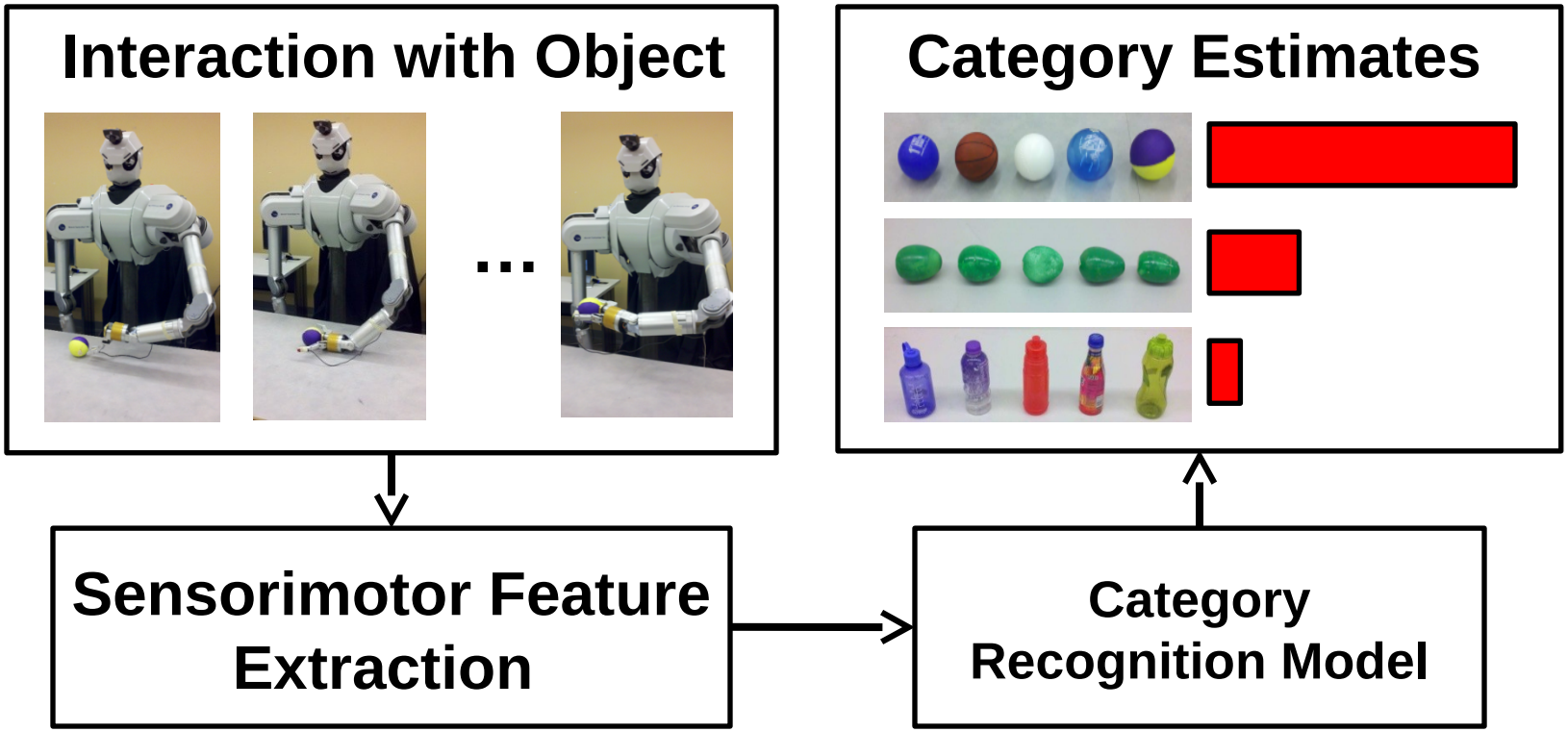
	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look				✓		✓
grasp	✓	✓	✓		✓	✓
lift	✓	✓			✓	✓
hold	✓	✓			✓	✓
shake	✓	✓			✓	✓
drop	✓	✓			✓	✓
tap	✓	✓			✓	✓
poke	✓	✓			✓	✓
push	✓	✓			✓	✓
press	✓	✓			✓	✓

Data From a Single Exploratory Trial

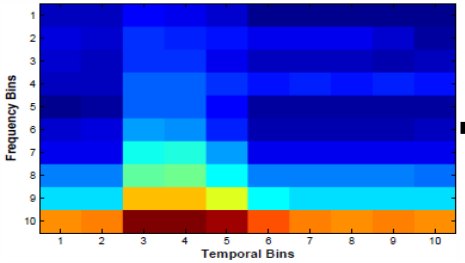
	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look				✓		✓
grasp	✓	✓	✓		✓	✓
lift	✓	✓			✓	✓
hold	✓	✓			✓	✓
shake	✓	✓			✓	✓
drop	✓	✓			✓	✓
tap	✓	✓			✓	✓
poke	✓	✓			✓	✓
push	✓	✓			✓	✓
press	✓	✓			✓	✓

x 5 per object

Overview


Interaction with Object
...
Category Estimates
Sensorimotor Feature Extraction
Category Recognition Model

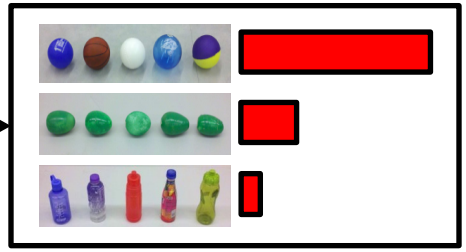
Context-specific Category Recognition



Observation from *poke-audio* context

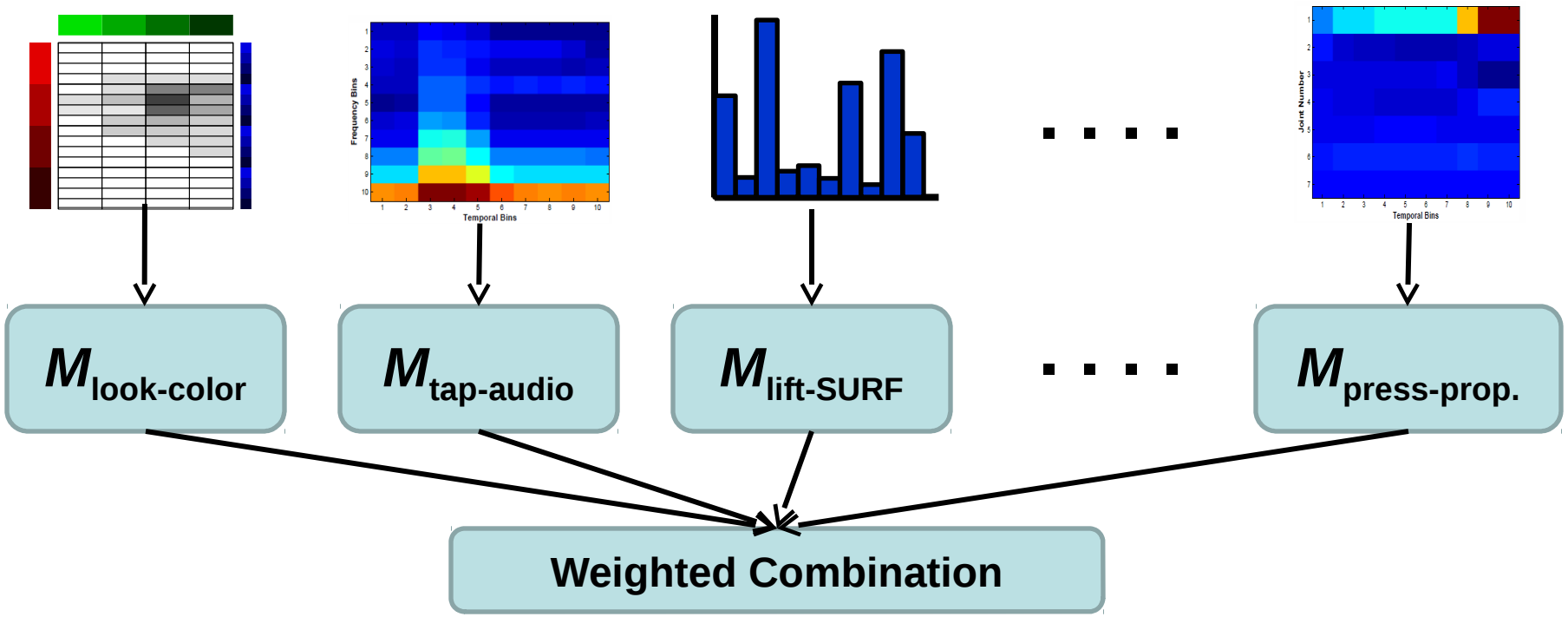
$M_{\text{poke-audio}}$

Recognition model for *poke-audio* context

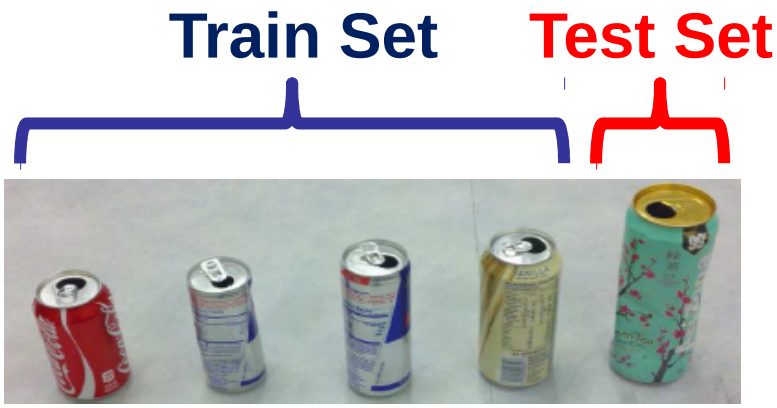


Distribution over category labels

Combining Model Outputs



Model Evaluation: 5 fold Cross-Validation



Recognition Rates (%) with SVM

	Audio	Proprioception	Color	Optical Flow	SURF	All
look			58.8		58.9	67.7
grasp	45.7	38.7		12.2	57.1	65.2
lift	48.1	63.7		5.0	65.9	79.0
hold	30.2	43.9		5.0	58.1	67.0
shake	49.3	57.7		32.8	75.6	76.8
drop	47.9	34.9		17.2	57.9	71.0
tap	63.3	50.7		26.0	77.3	82.4
push	72.8	69.6		26.4	76.8	88.8
poke	65.9	63.9		17.8	74.7	85.4
press	62.7	69.7		32.4	69.7	77.4

Recognition Rates (%) with SVM

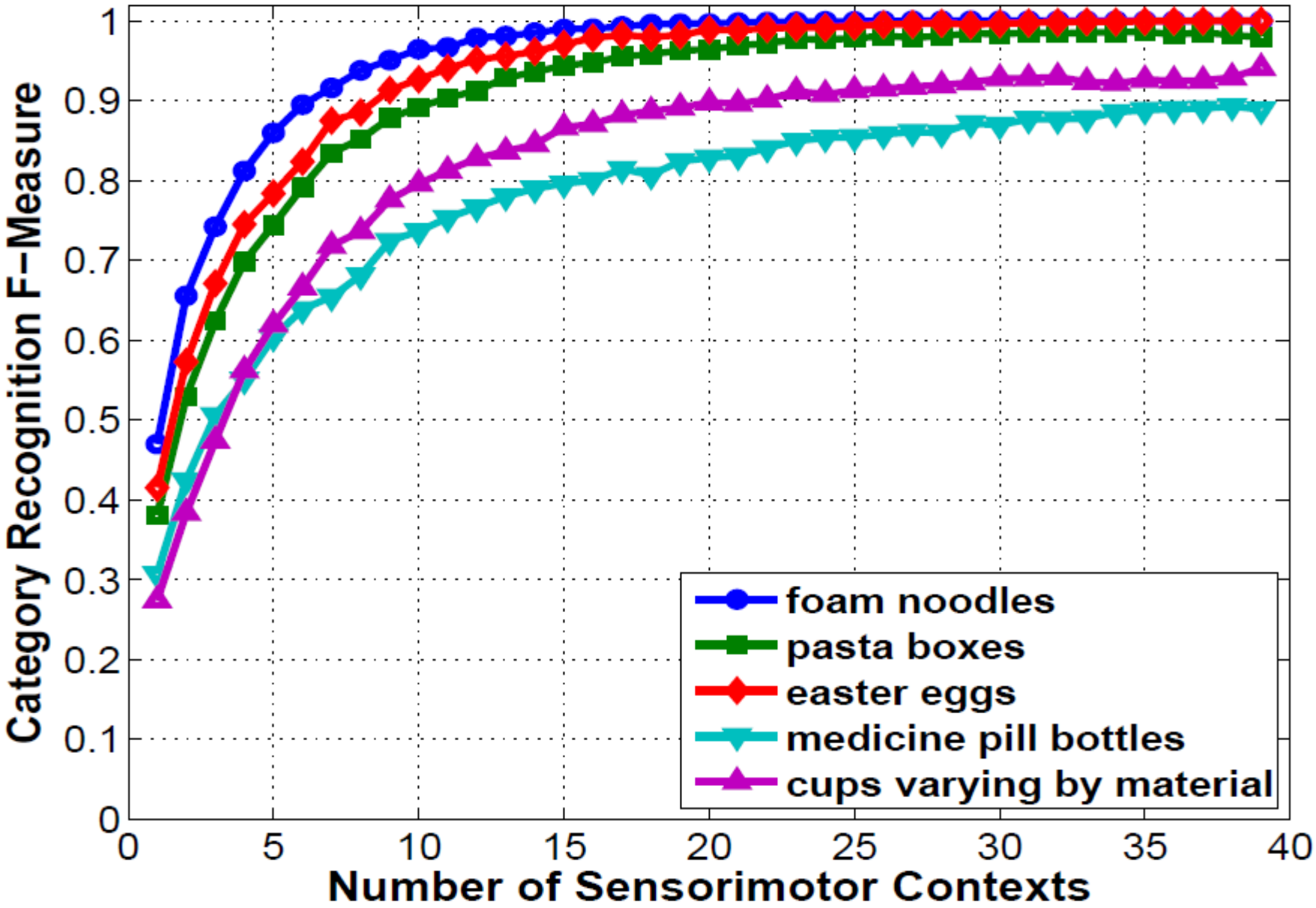
	Audio	Proprioception	Color	Optical Flow	SURF	All
look			58.8		58.9	67.7
grasp	45.7	38.7		12.2	57.1	65.2
lift	48.1	63.7		5.0	65.9	79.0
hold	30.2	43.9		5.0	58.1	67.0
shake	49.3	57.7		32.8	75.6	76.8
drop	47.9	34.9		17.2	57.9	71.0
tap	63.3	50.7		26.0	77.3	82.4
push	72.8	69.6		26.4	76.8	88.8
poke	65.9	63.9		17.8	74.7	85.4
press	62.7	69.7		32.4	69.7	77.4

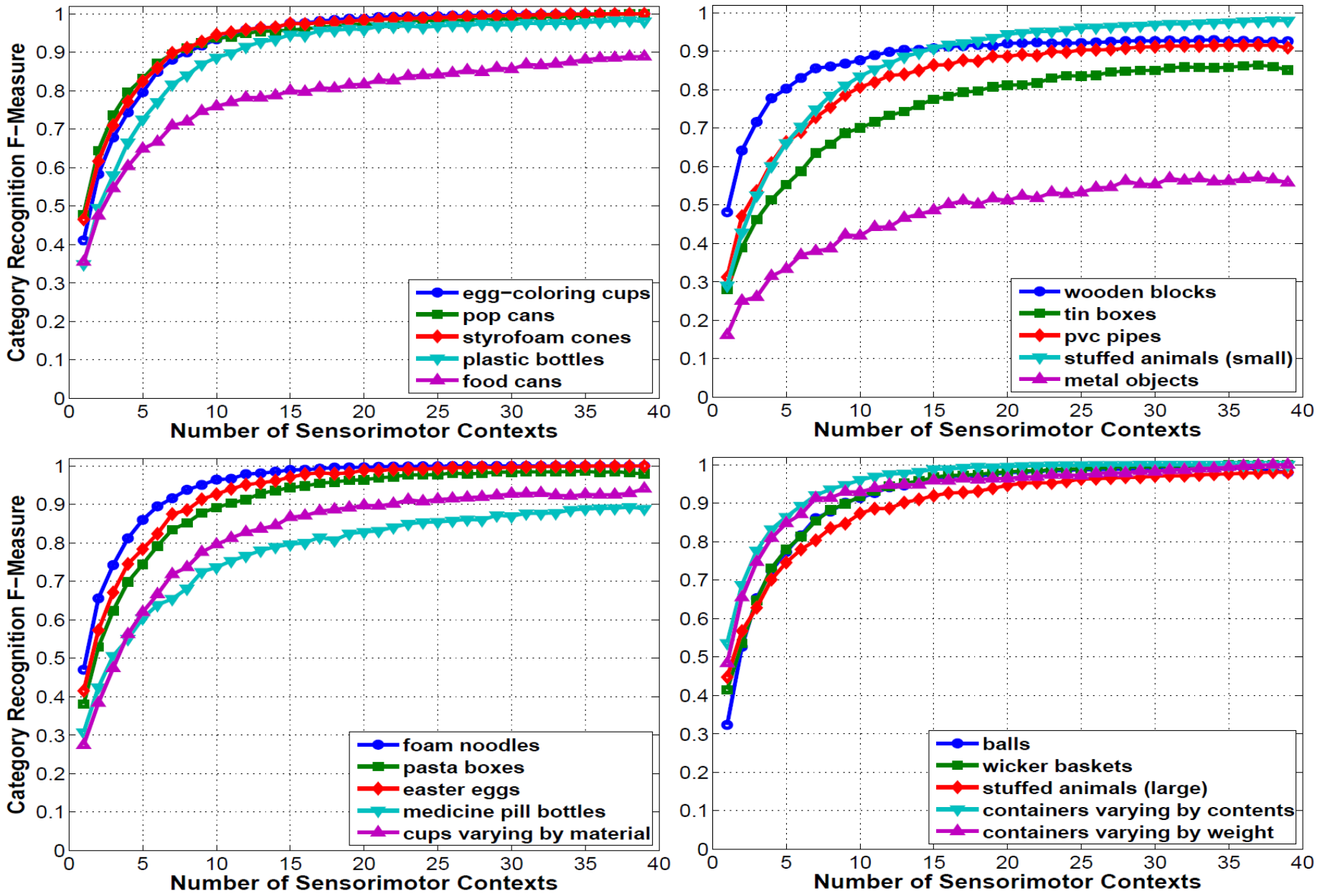
Recognition Rates (%) with SVM

	Audio	Proprioception	Color	Optical Flow	SURF	All
look			58.8		58.9	67.7
grasp	45.7	38.7		12.2	57.1	65.2
lift	48.1	63.7		5.0	65.9	79.0
hold	30.2	43.9		5.0	58.1	67.0
shake	49.3	57.7		32.8	75.6	76.8
drop	47.9	34.9		17.2	57.9	71.0
tap	63.3	50.7		26.0	77.3	82.4
push	72.8	69.6		26.4	76.8	88.8
poke	65.9	63.9		17.8	74.7	85.4
press	62.7	69.7		32.4	69.7	77.4

Recognition Rates (%) with SVM

	Audio	Proprioception	Color	Optical Flow	SURF	All
look			58.8		58.9	67.7
grasp	45.7	38.7		12.2	57.1	65.2
lift	48.1	63.7		5.0	65.9	79.0
hold	30.2	43.9		5.0	58.1	67.0
shake	49.3	57.7		32.8	75.6	76.8
drop	47.9	34.9		17.2	57.9	71.0
tap	63.3	50.7		26.0	77.3	82.4
push	72.8	69.6		26.4	76.8	88.8
poke	65.9	63.9		17.8	74.7	85.4
press	62.7	69.7		32.4	69.7	77.4


Category Recognition F−Measure
Number of Sensorimotor Contexts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
foam noodles
pasta boxes
easter eggs
medicine pill bottles
cups varying by material


Category Recognition F−Measure
Number of Sensorimotor Contexts
egg−coloring cups
pop cans
styrofoam cones
plastic bottles
food cans
wooden blocks
tin boxes
pvc pipes
stuffed animals (small)
metal objects
foam noodles
pasta boxes
easter eggs
medicine pill bottles
cups varying by material
balls
wicker baskets
stuffed animals (large)
containers varying by contents
containers varying by weight

Can behaviors be selected actively to minimize exploration time?

Active Behavior Selection

- For each behavior $b \in \mathcal{B}$, estimate C_{ij}^{b} such that

$$C_{ij}^{b} = \frac{Pr(\hat{y} = y_i | y = y_j) + Pr(\hat{y} = y_j | y = y_i)}{2}$$

- Let $\hat{\mathbf{p}} \in \mathbb{R}^{|\mathcal{Y}|}$ be the vector encoding the robot's current estimates over the category labels and let $\mathcal{B}_r$ be the remaining set of behaviors available to the robot

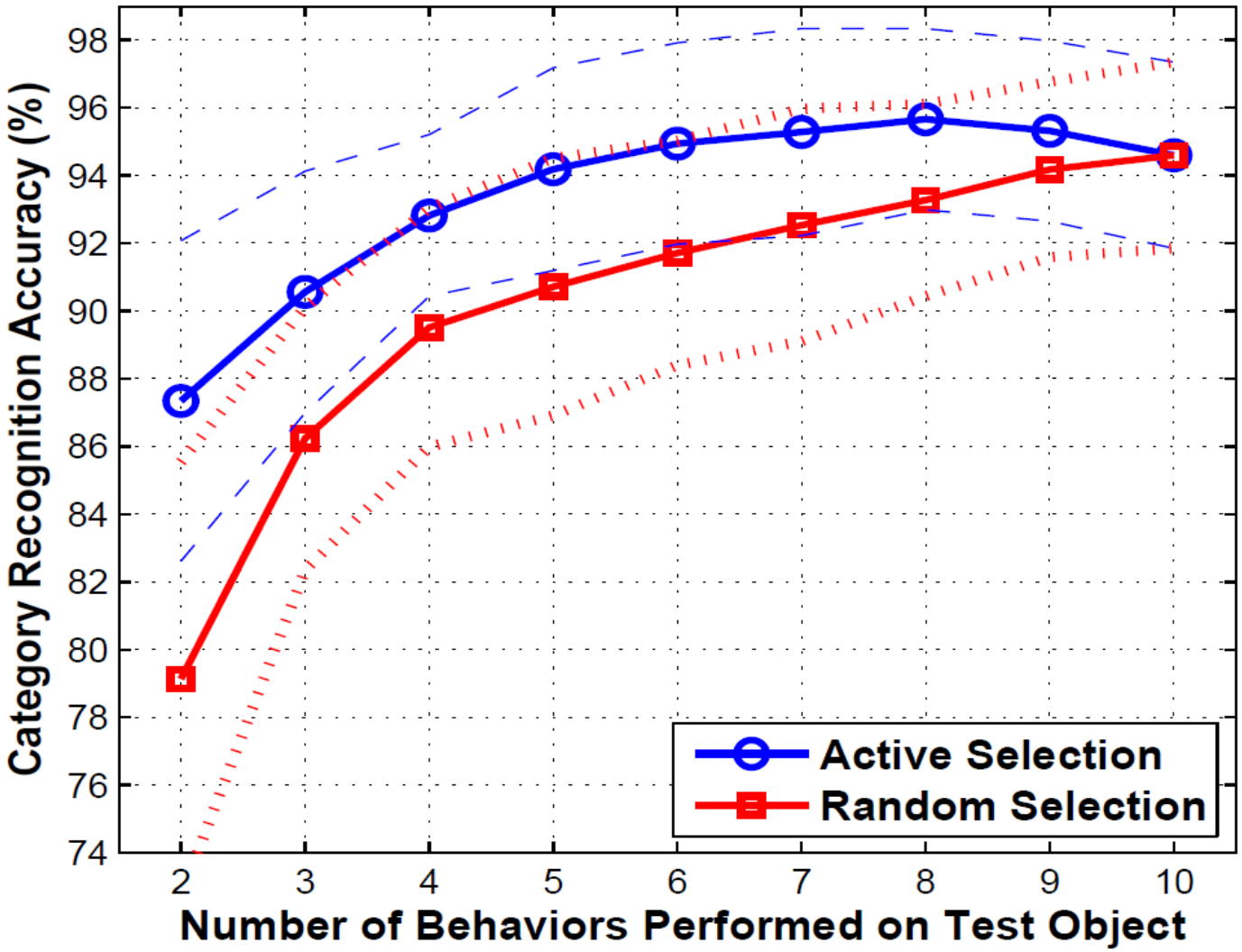
Active Behavior Selection

1) Compute the set $\mathcal{Y}_K \subset \mathcal{Y}$ such that it contains the K most likely object categories according to $\hat{\mathbf{p}}$.
2) Pick the next behavior b_{next} with an associated confusion matrix that is least likely to confuse the categories within the set $\mathcal{Y}_K$, i.e.,

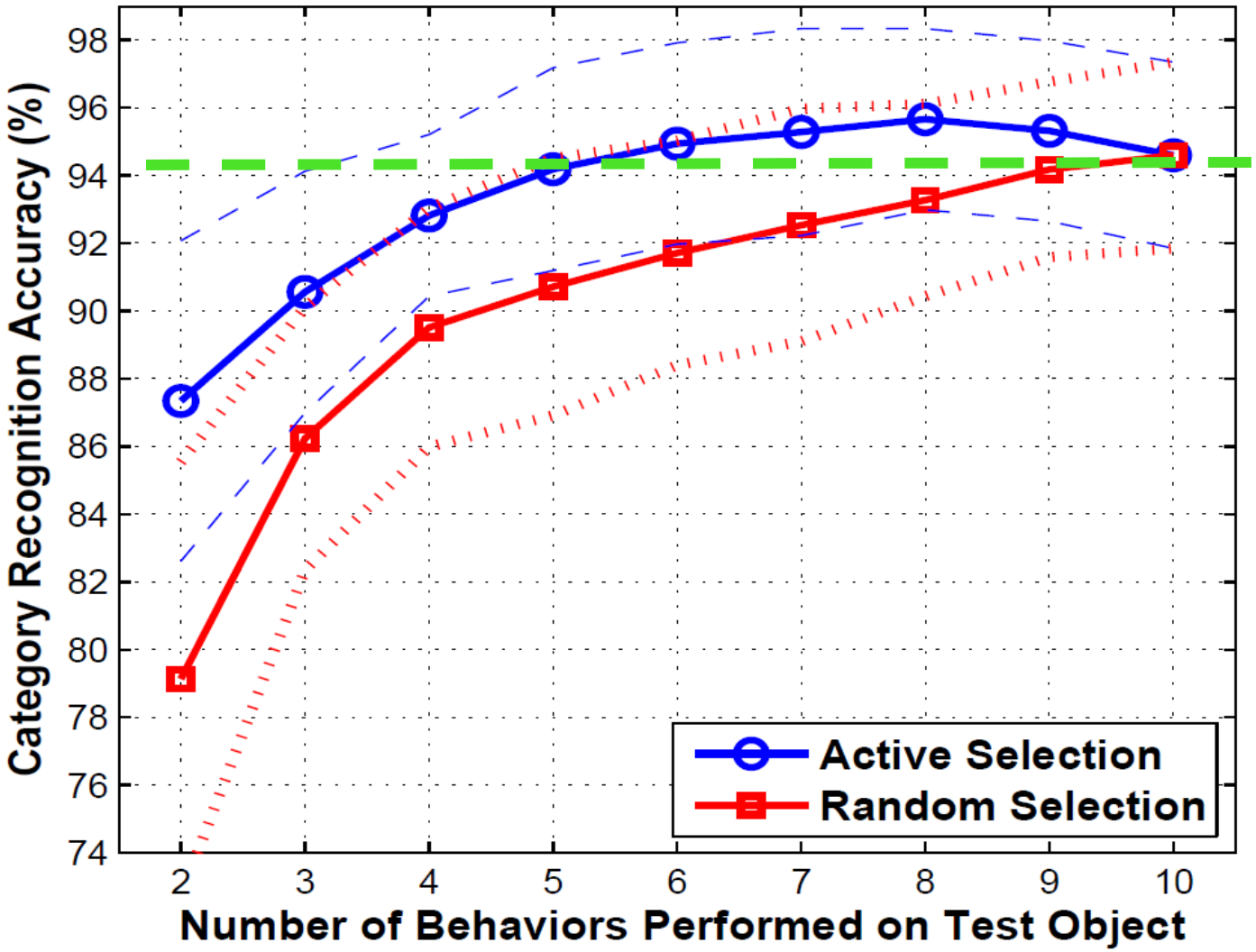
$$b_{next} = \arg\min_{b \in \mathcal{B}_r} \sum_{y_i \in \mathcal{Y}_K} \sum_{y_j \in \mathcal{Y}_K / y_i} C_{ij}^b$$

3) Update the estimate $\hat{\mathbf{p}}$ using the classifiers associated with the sensorimotor contexts of b_{next}.
4) Remove b_{next} from $\mathcal{B}_r$. If $|\mathcal{B}_r| \geq 1$, go back to step 1).

Active vs. Random Behavior Selection



Active vs. Random Behavior Selection



Summary of Experiment

- Using visual, proprioceptive and auditory sensory modalities, coupled with 10 different behaviors, the robot achieved high category recognition rates
- Sensorimotor contexts can be linked to specific categories enabling even better rates when looking for a specific object type
- Active behavior selection when classifying a novel object reduced exploration time by half

Developmental Progression

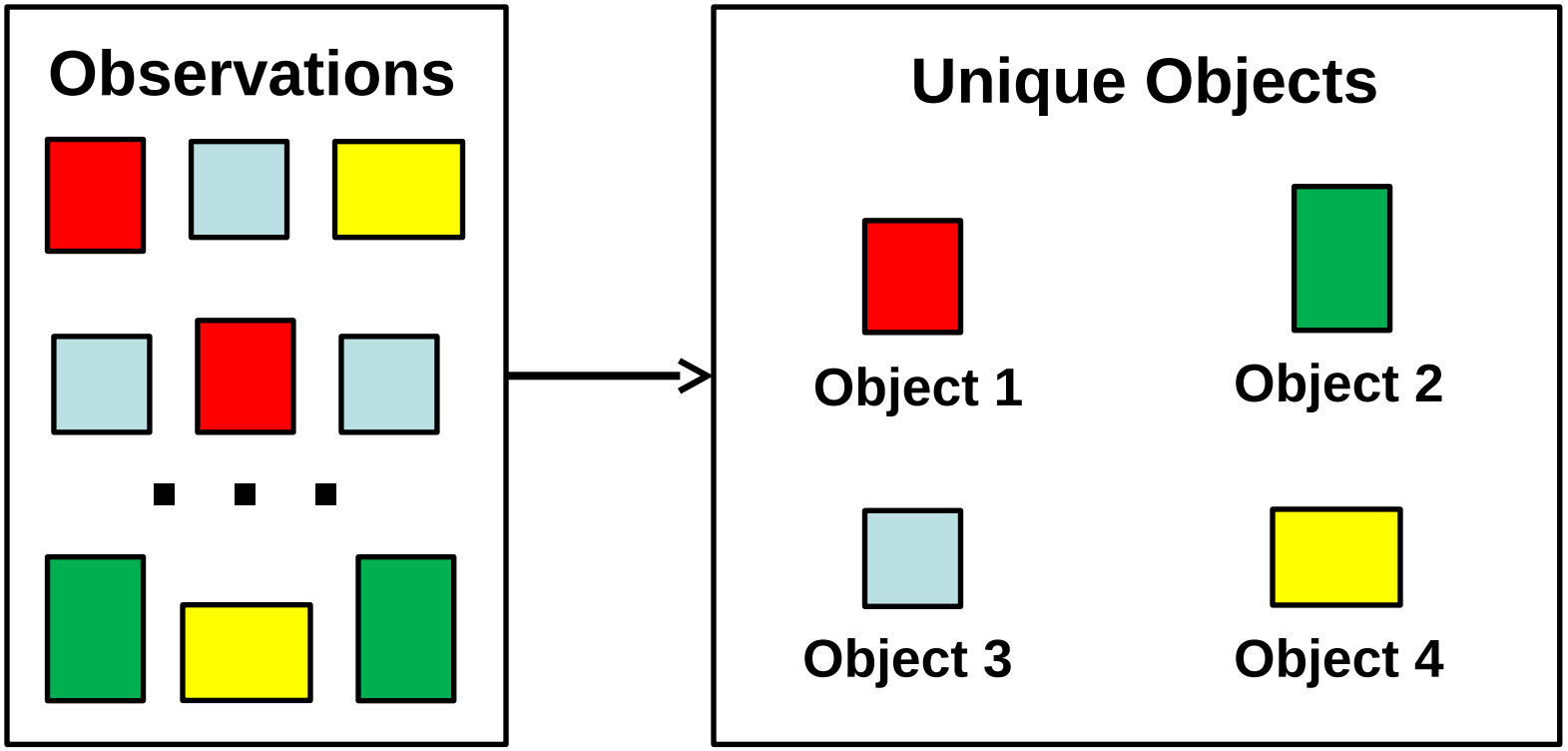
Developmental Progression

Object Individuation

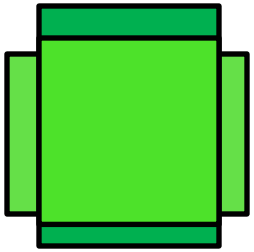
- Object individuation is the problem of inferring how many unique objects have been observed

- To solve it, a robot has to partition its sensorimotor experience according to the identity of the object that is present

Object Individuation



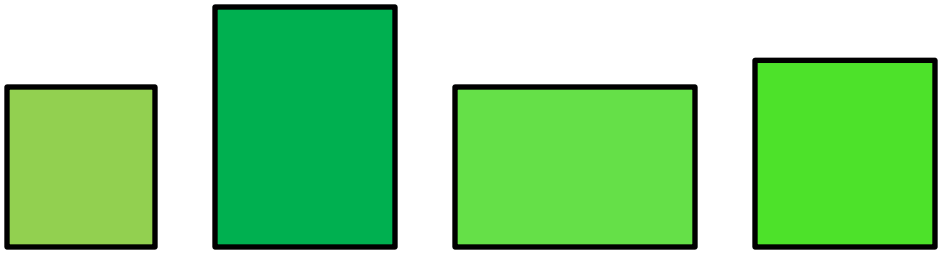
Sinapov and Stoytchev, "Grounded Object Individuation by a Humanoid Robot"
Proceedings of the 2013 IEEE ICRA



How many unique objects were there?

How many unique objects were there?

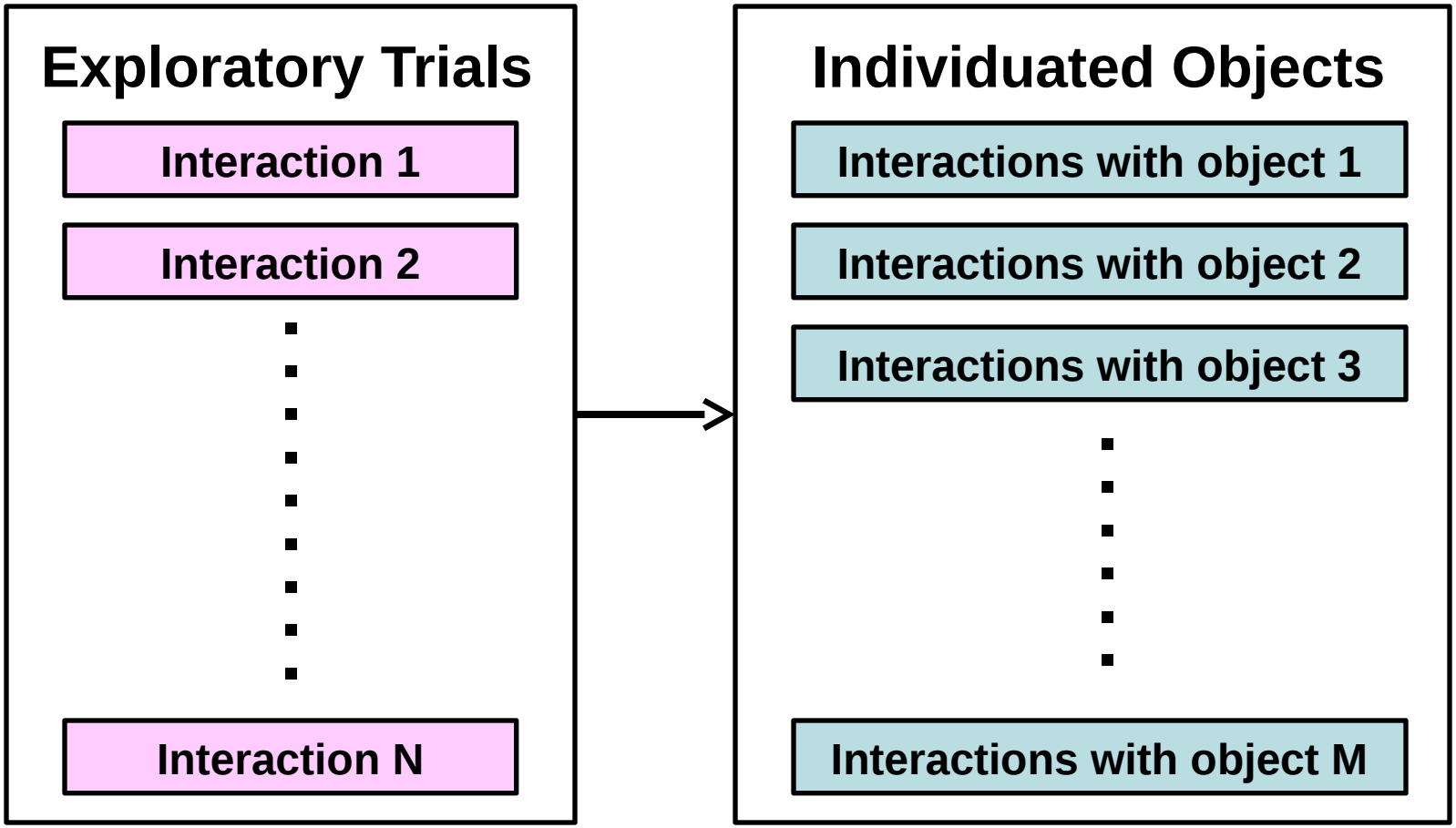
- Correct answer: 4



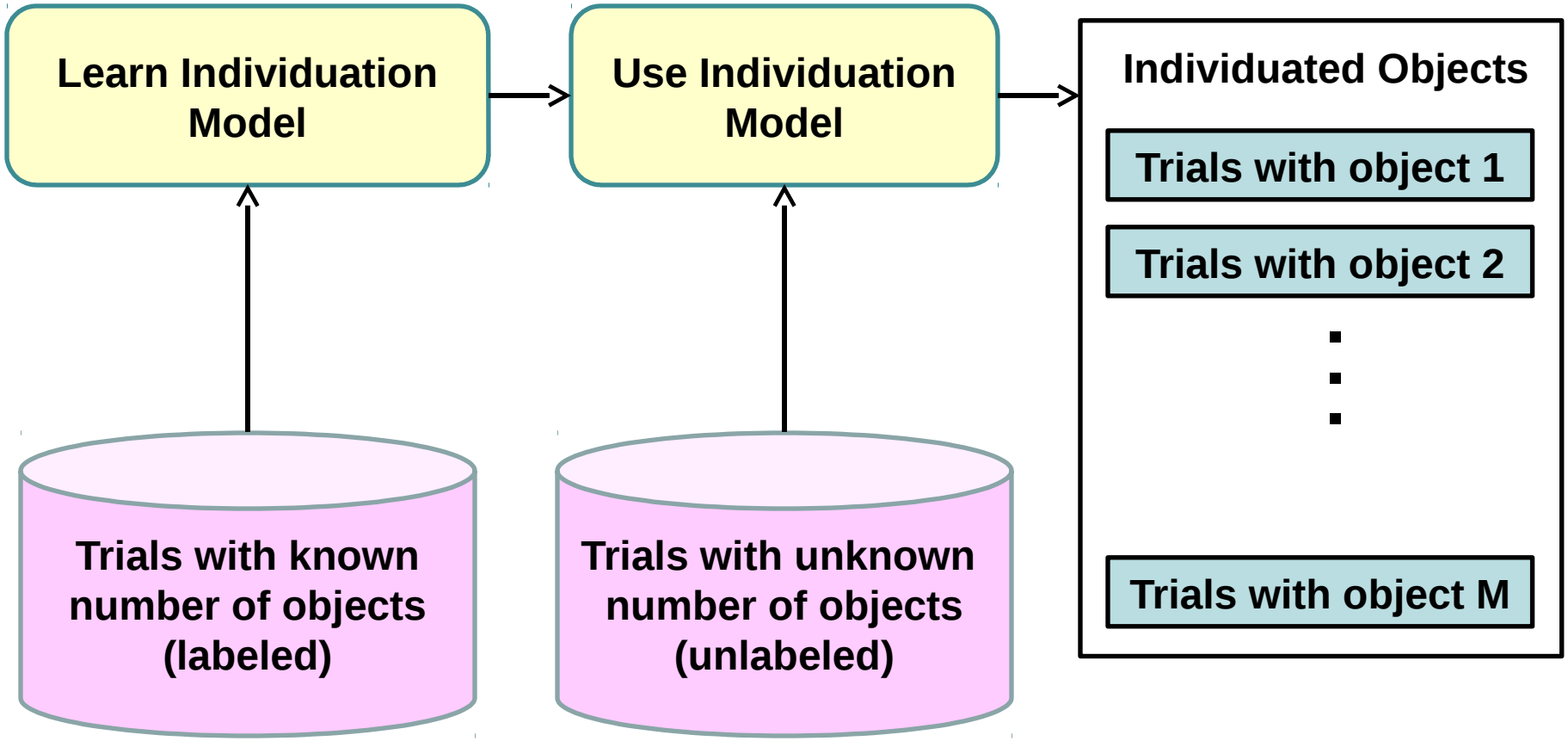
Related Work in 3D Object Discovery

Collet *et al*. 2013

Problem Formulation

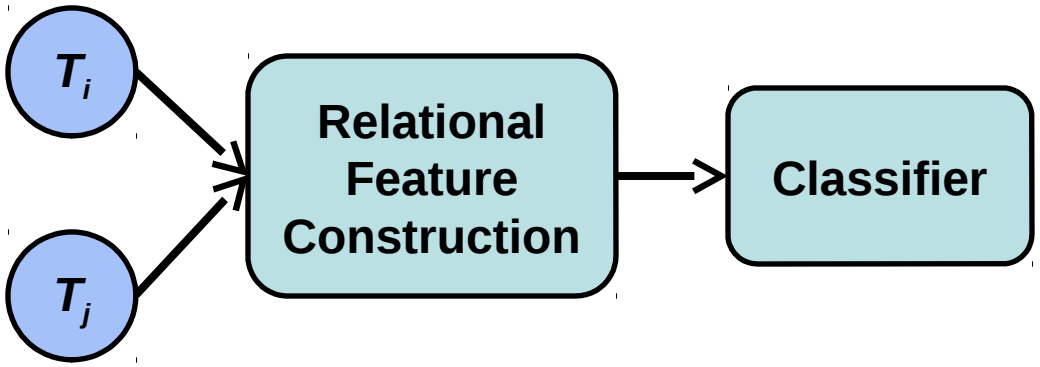


Approach

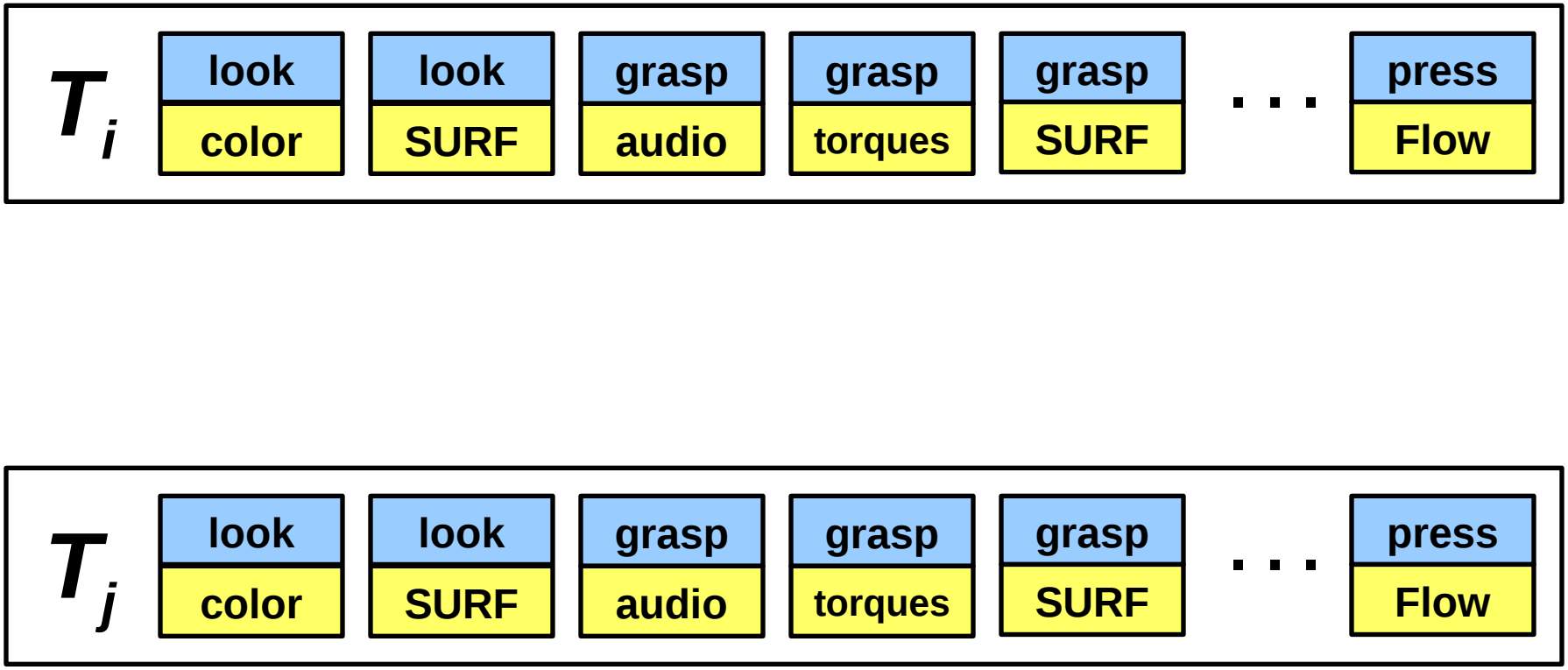

Learn Individuation Model
Use Individuation Model
Individuated Objects
Trials with object 1
Trials with object 2
Trials with object M
Trials with known number of objects (labeled)
Trials with unknown number of objects (unlabeled)

Learning an Individuation Model

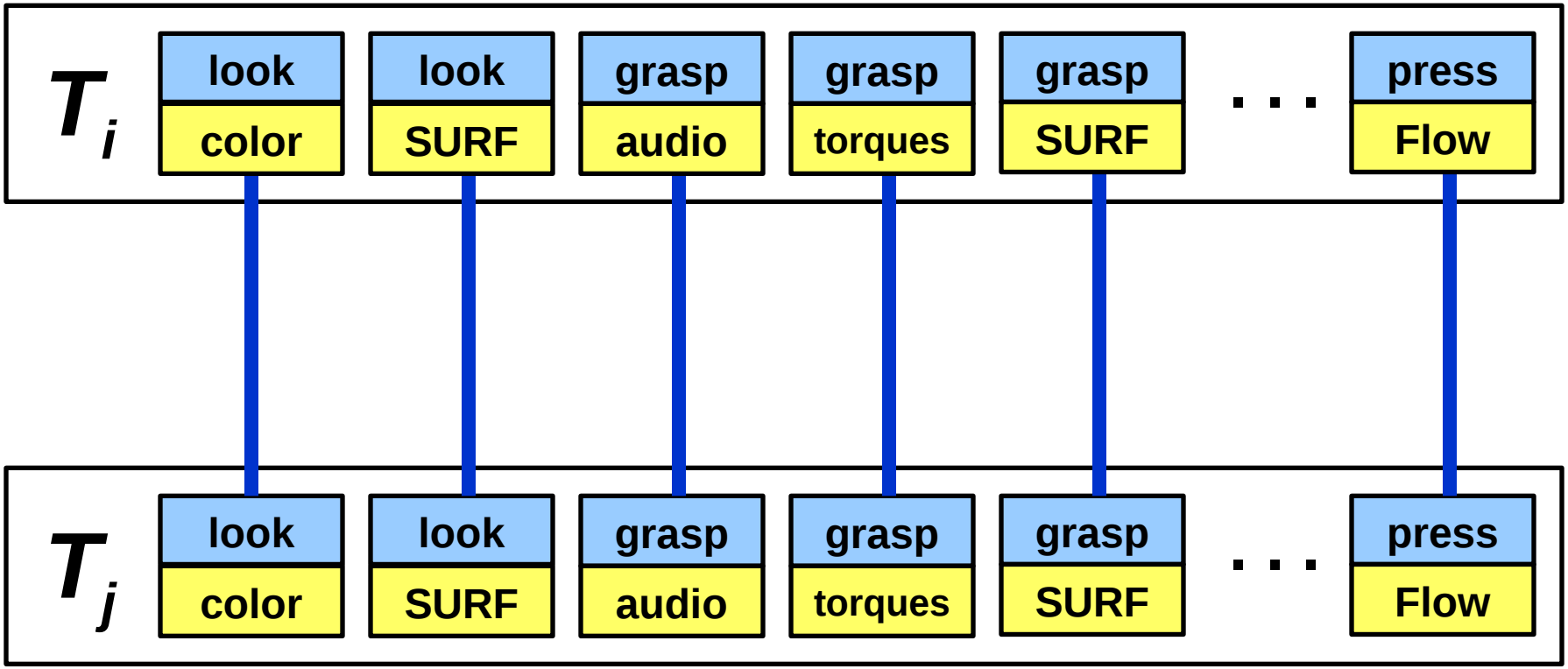
- Learn a model that classifies a trial pair (T_i, T_j) as either “same object” or “different object”



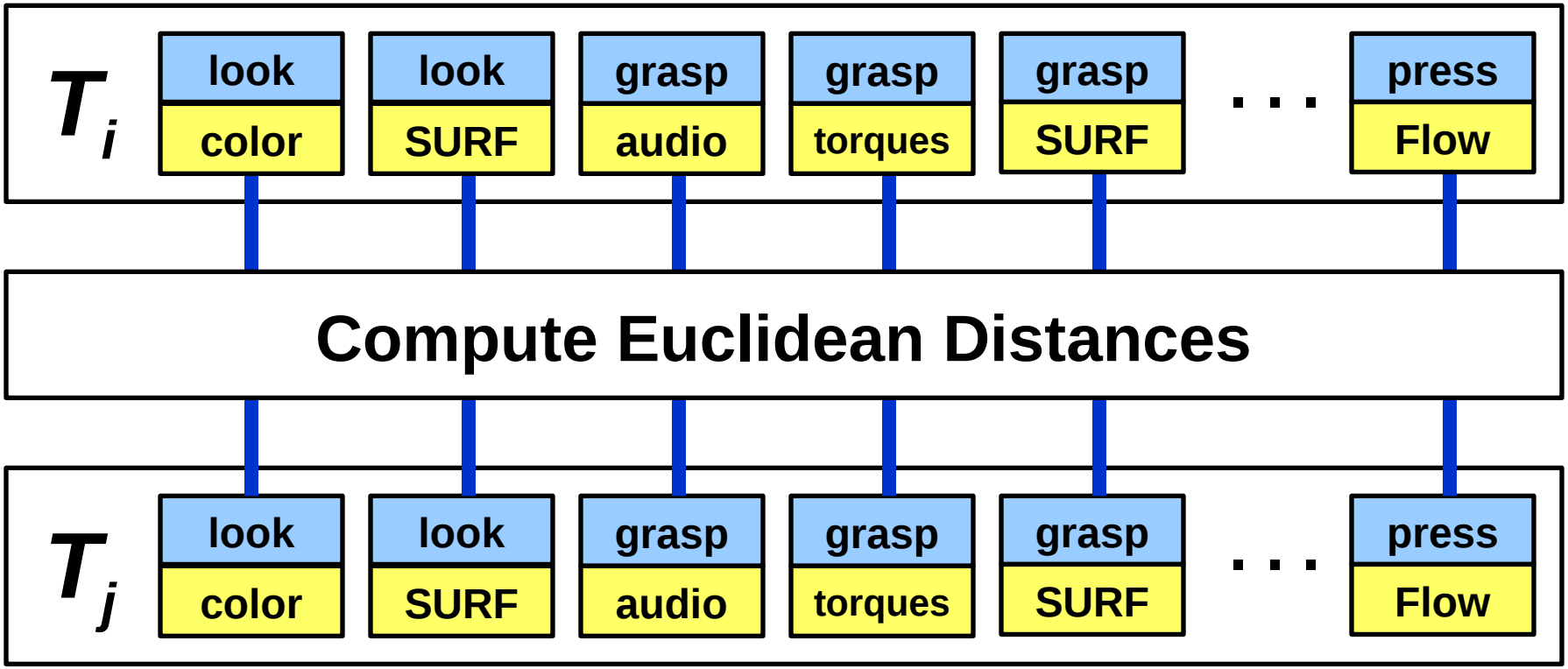
Learning an Individuation Model



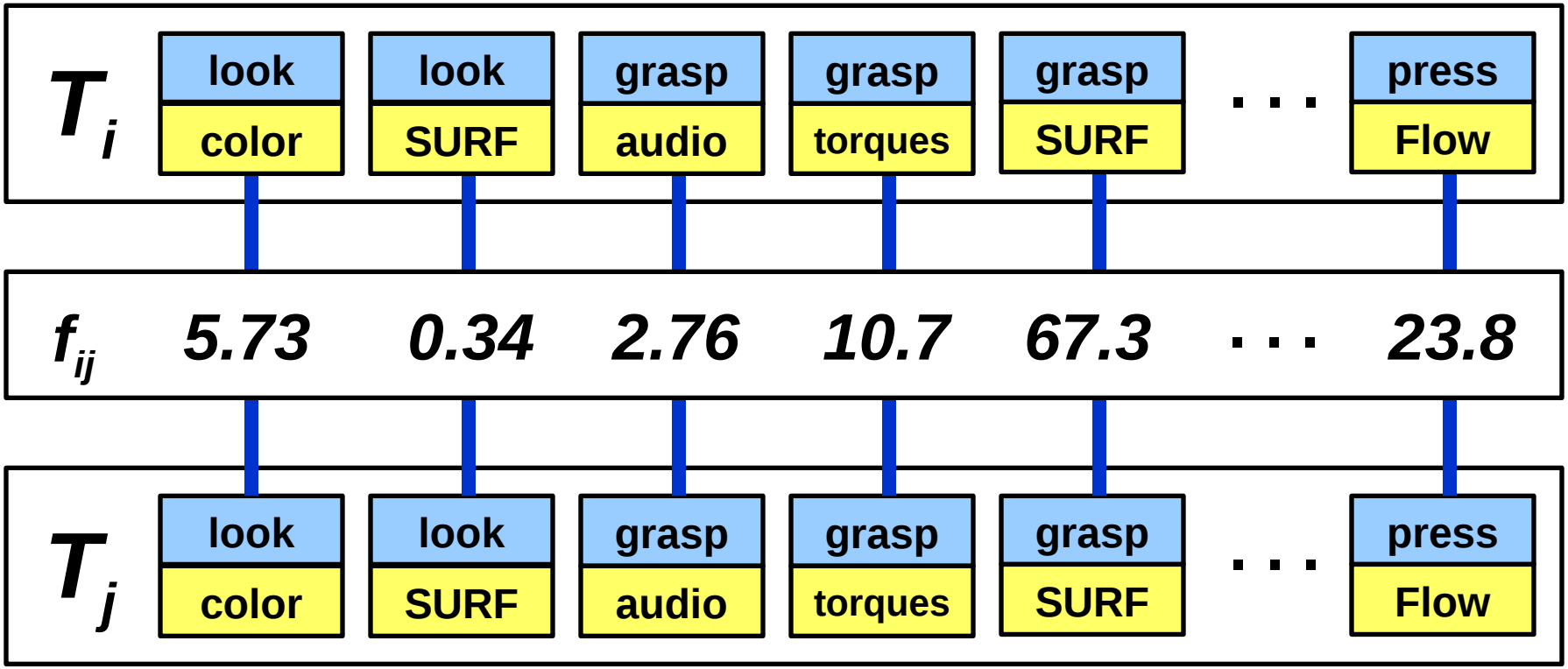
Learning an Individuation Model



Learning an Individuation Model



Learning an Individuation Model



Example Training Set
(2 objects and 4 trials)

Objects:

Trials:

T_1 

T_2 

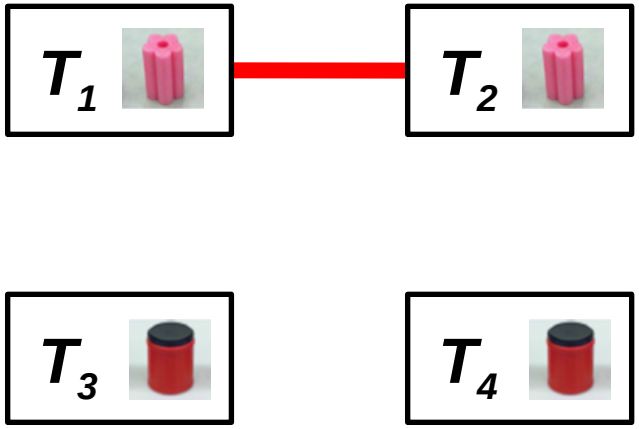
T_3 

T_4 

Learning an Individuation Model

Objects:

Trials:

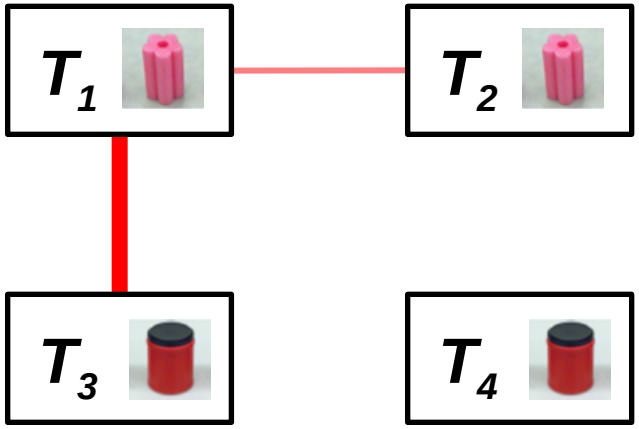


$\mathbf{f}_{12} \in \mathbb{R}^{39}$ **, "same"**

Learning an Individuation Model

Objects:

Trials:



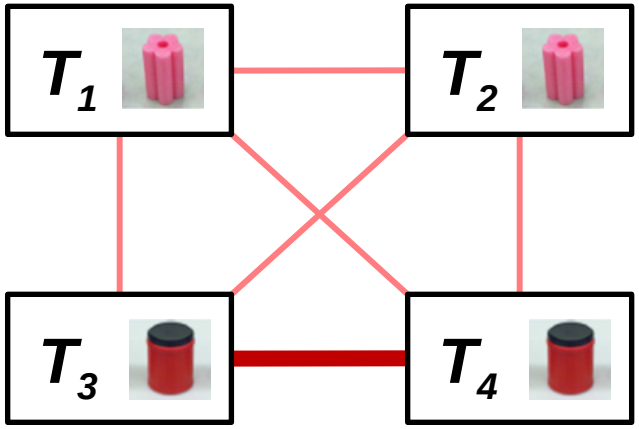
$\mathbf{f}_{12} \in \mathbb{R}^{39}$ **, "same"**

$\mathbf{f}_{13} \in \mathbb{R}^{39}$ **, "different"**

Learning an Individuation Model

Objects:

Trials:



$\mathbf{f}_{12} \in \mathbb{R}^{39}$ **, "same"**

$\mathbf{f}_{13} \in \mathbb{R}^{39}$ **, "different"**

$\mathbf{f}_{14} \in \mathbb{R}^{39}$ **, "different"**

$\mathbf{f}_{23} \in \mathbb{R}^{39}$ **, "different"**

$\mathbf{f}_{24} \in \mathbb{R}^{39}$ **, "different"**

$\mathbf{f}_{34} \in \mathbb{R}^{39}$ **, "same"**

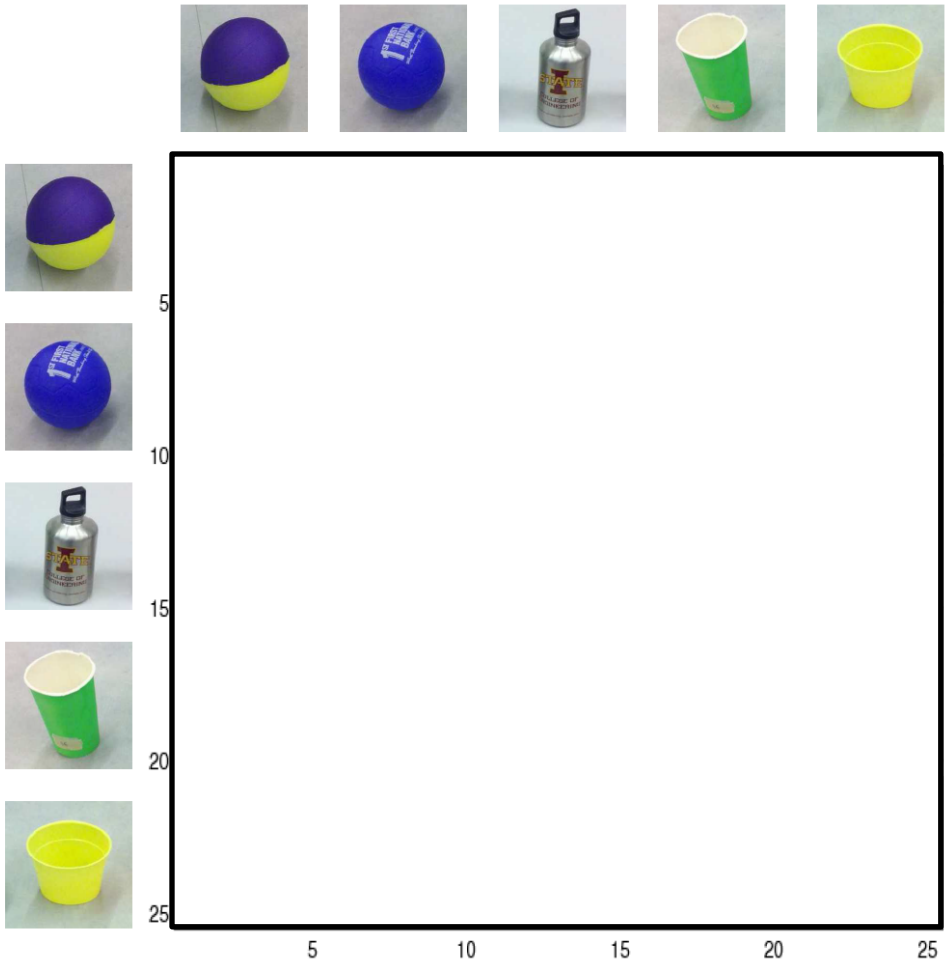
Individuating Novel Objects

- Let $\mathcal{T}_{test}$ be a test set of 25 trials with 5 different objects (5 trials per object)

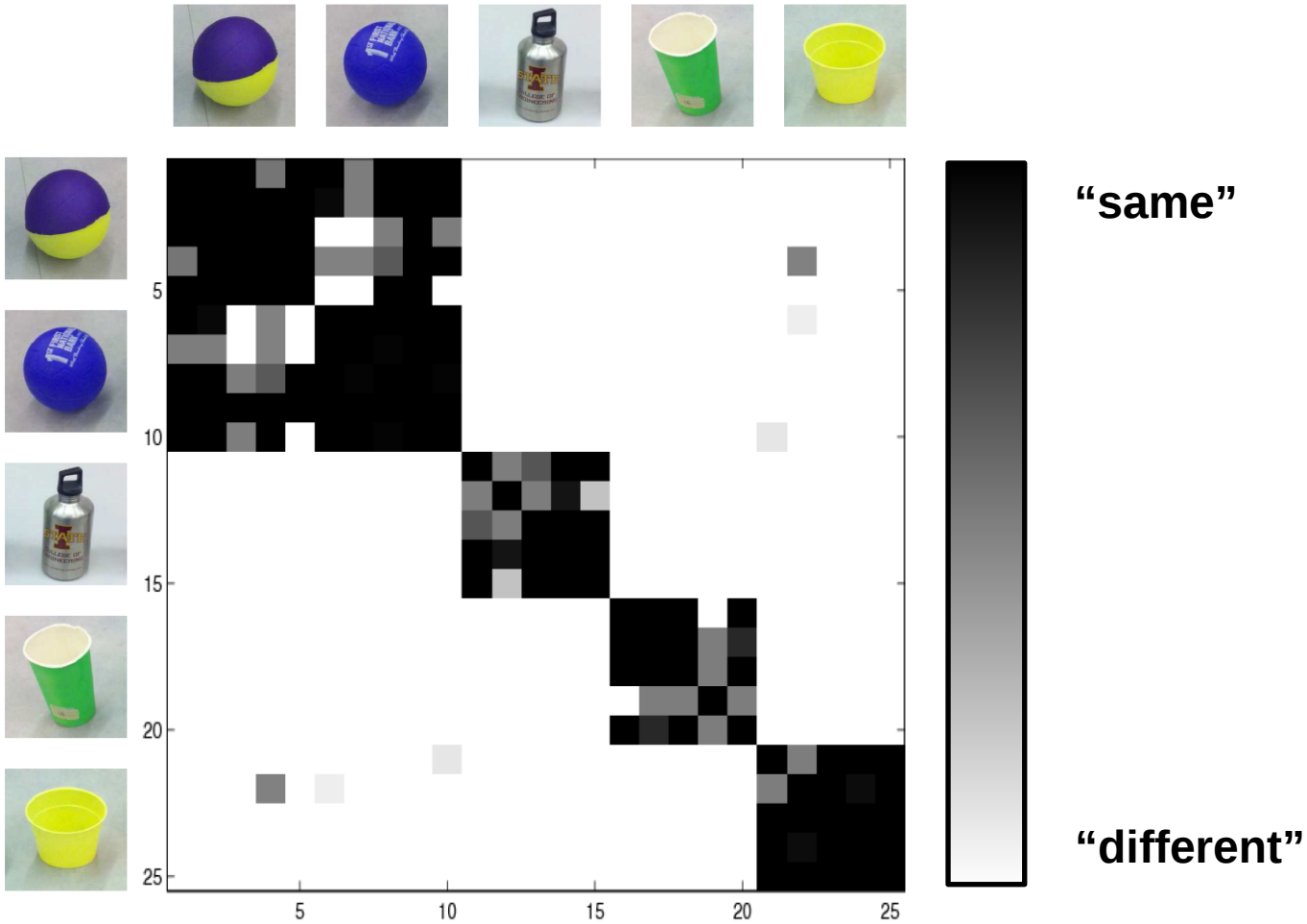




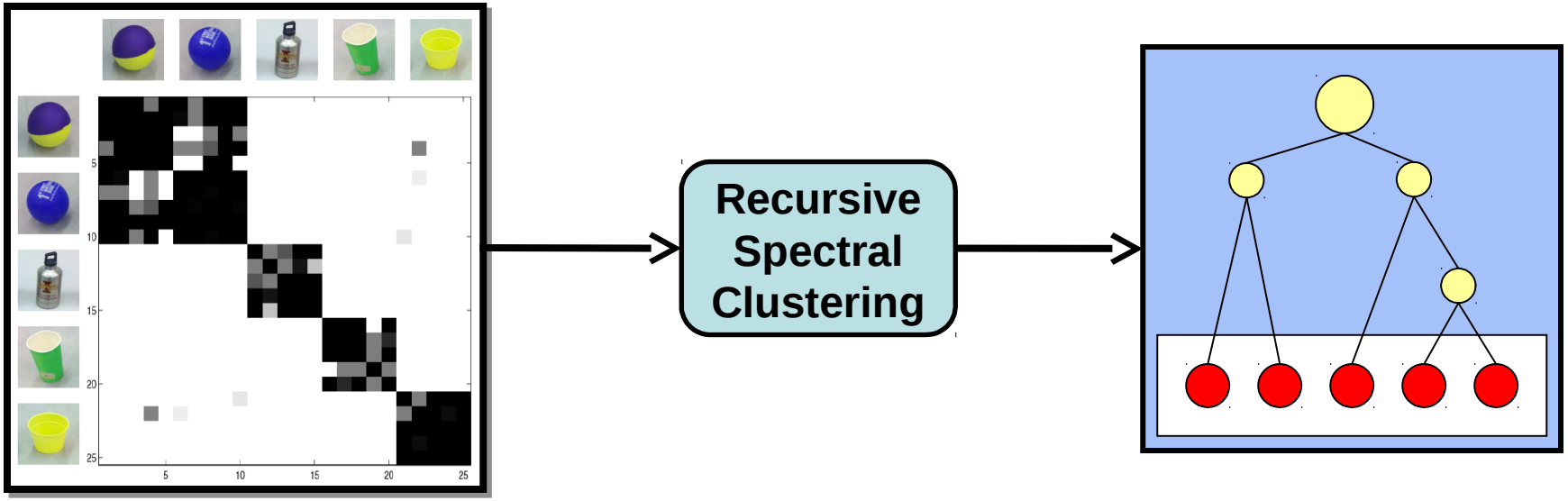
Output of Individuation Model



Output of Individuation Model



Individuating Novel Objects



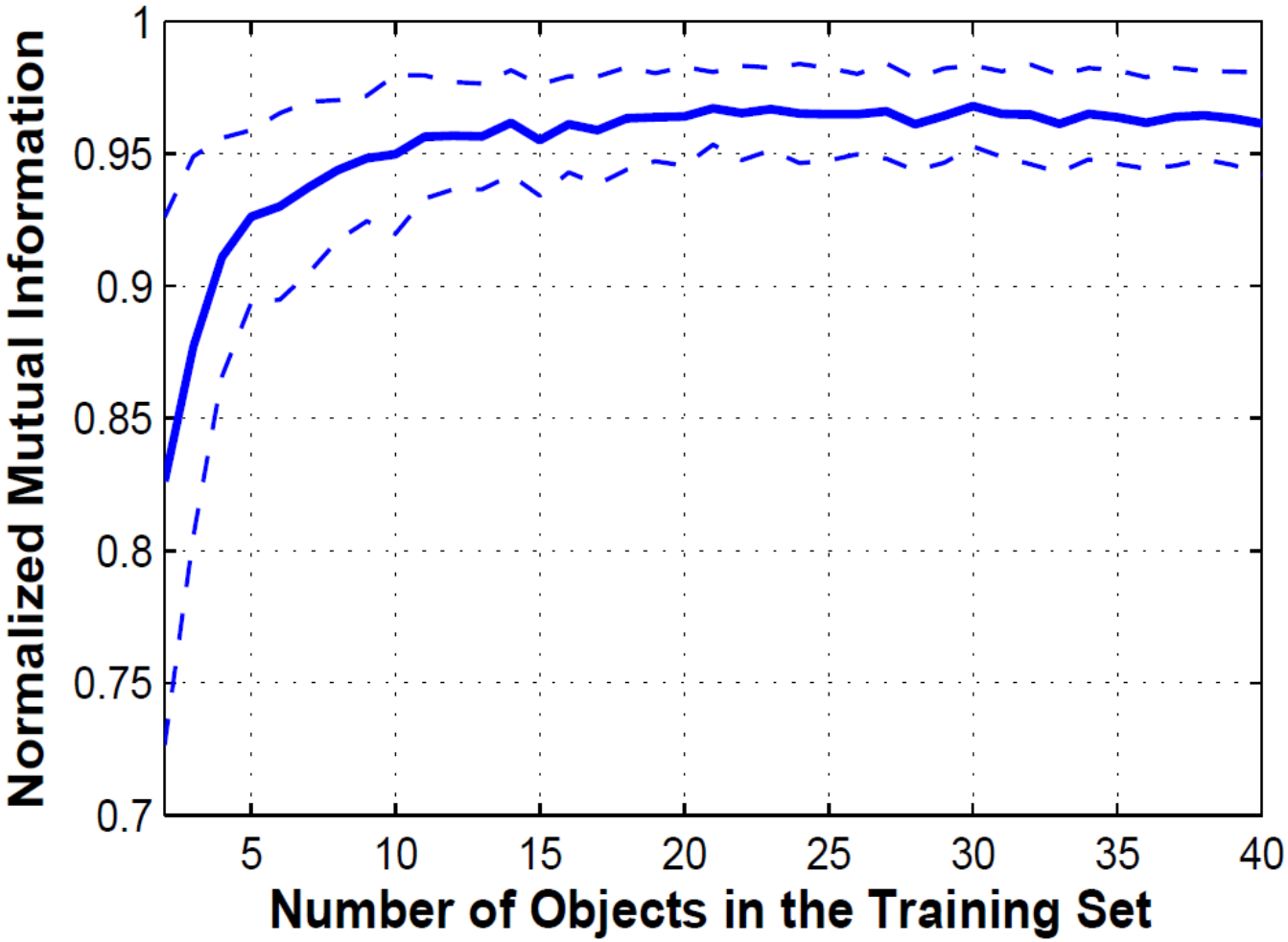
Individuating Objects

Evaluation Metrics

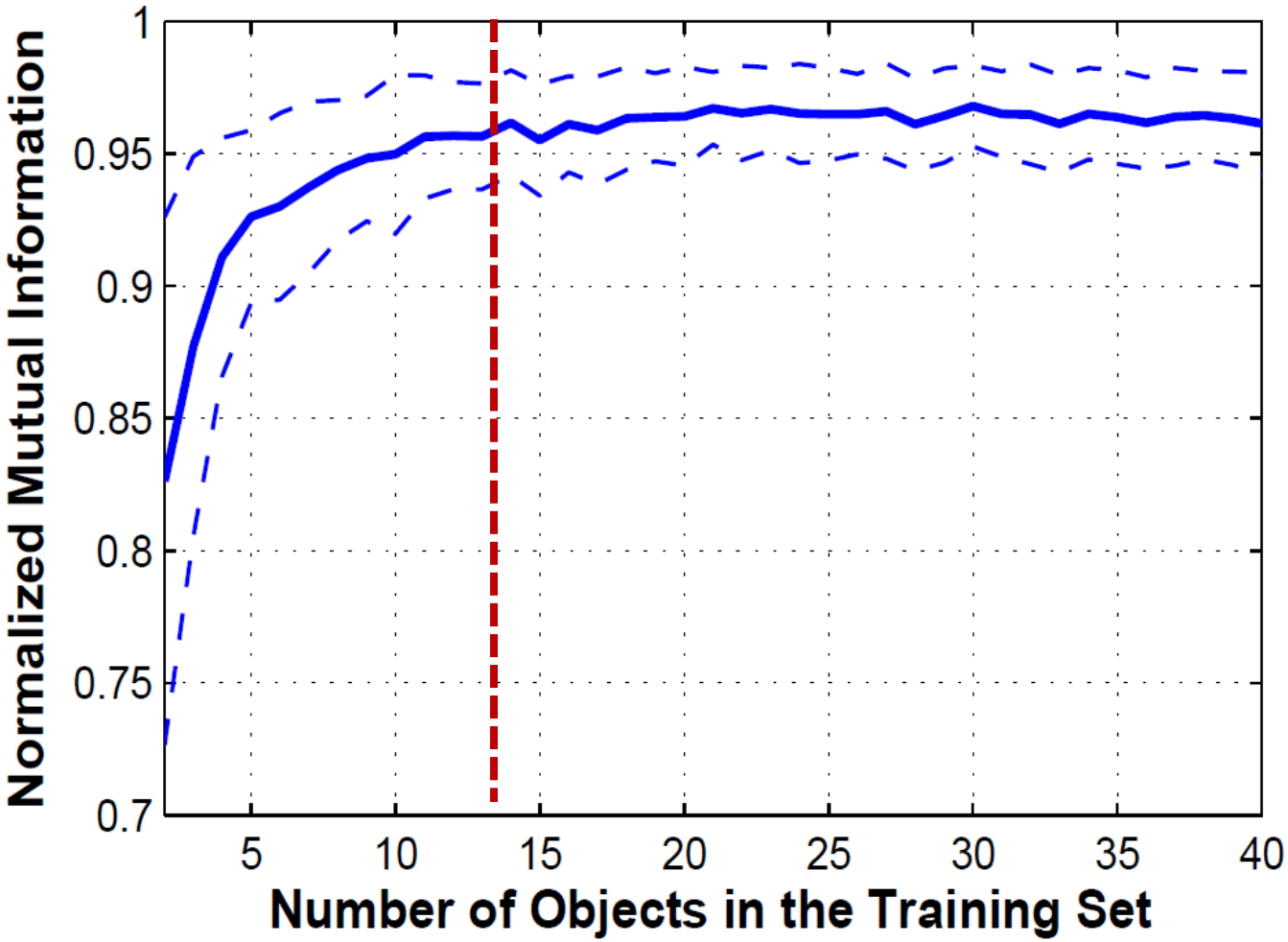
- Normalized Mutual Information:
  - captures the similarity between two different clusterings over the same dataset

$$\phi^{NMI}(\omega^a, \omega^b) = \frac{\sum_{h=1}^{k^a} \sum_{\ell=1}^{k^b} n_{h,\ell} \, log \left(\frac{n * n_{h,\ell}}{n_h^a * n_\ell^b} \right)}{\sqrt{\left(\sum_{h=1}^{k^a} n_h^a \, log(\frac{n_h^a}{n}) \right) \left(\sum_{\ell=1}^{k^b} n_\ell^b \, log(\frac{n_\ell^b}{n}) \right)}}$$

How many training objects do we need?

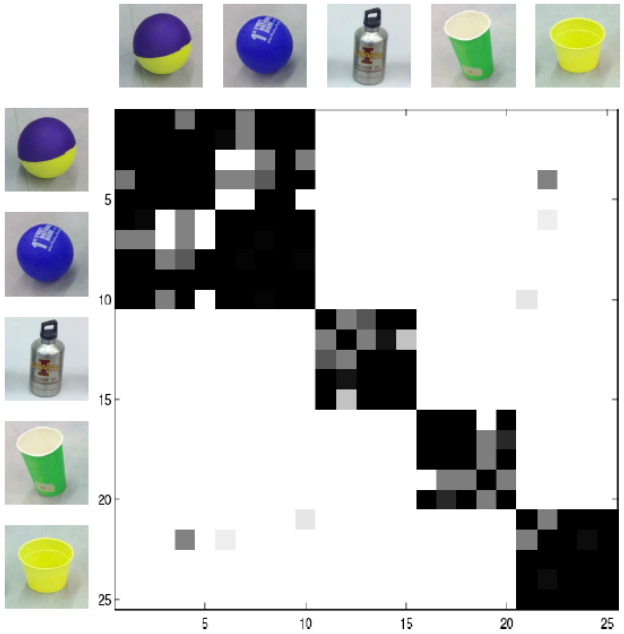


How many training objects do we need?

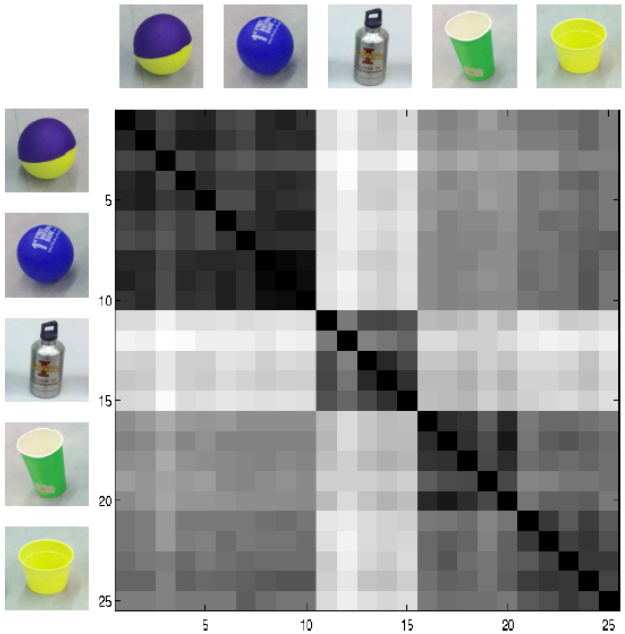


Baseline Comparison

- The approach is compared with unsupervised clustering over a matrix encoding raw perceptual similarity

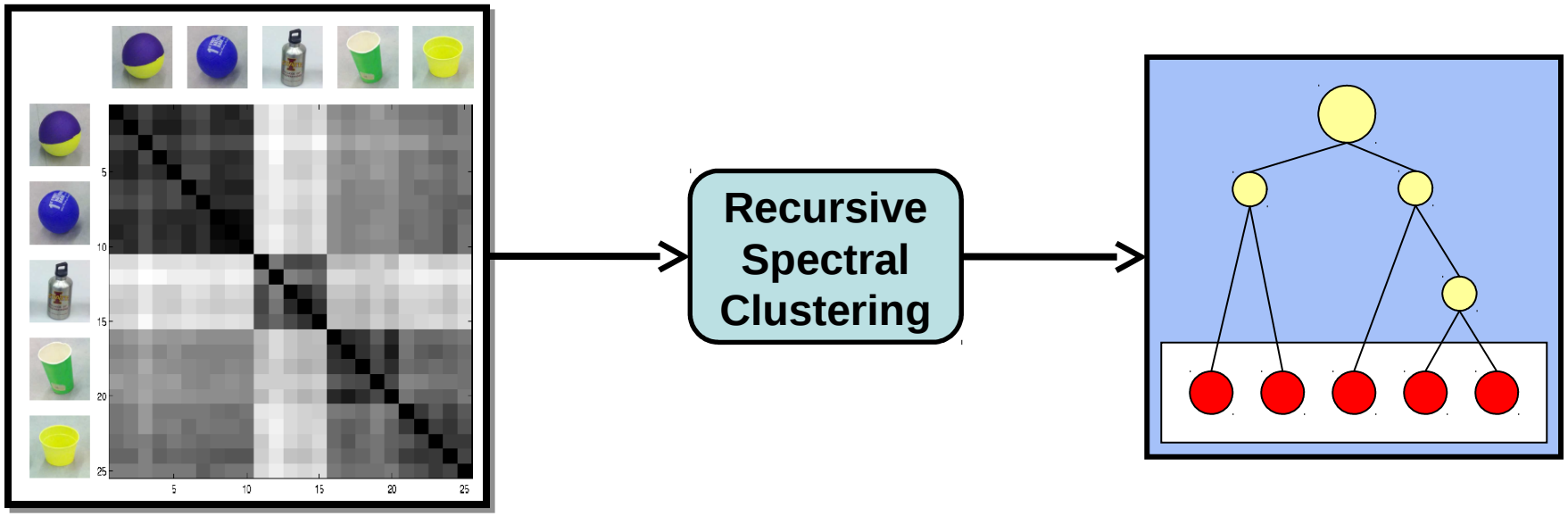


Learned

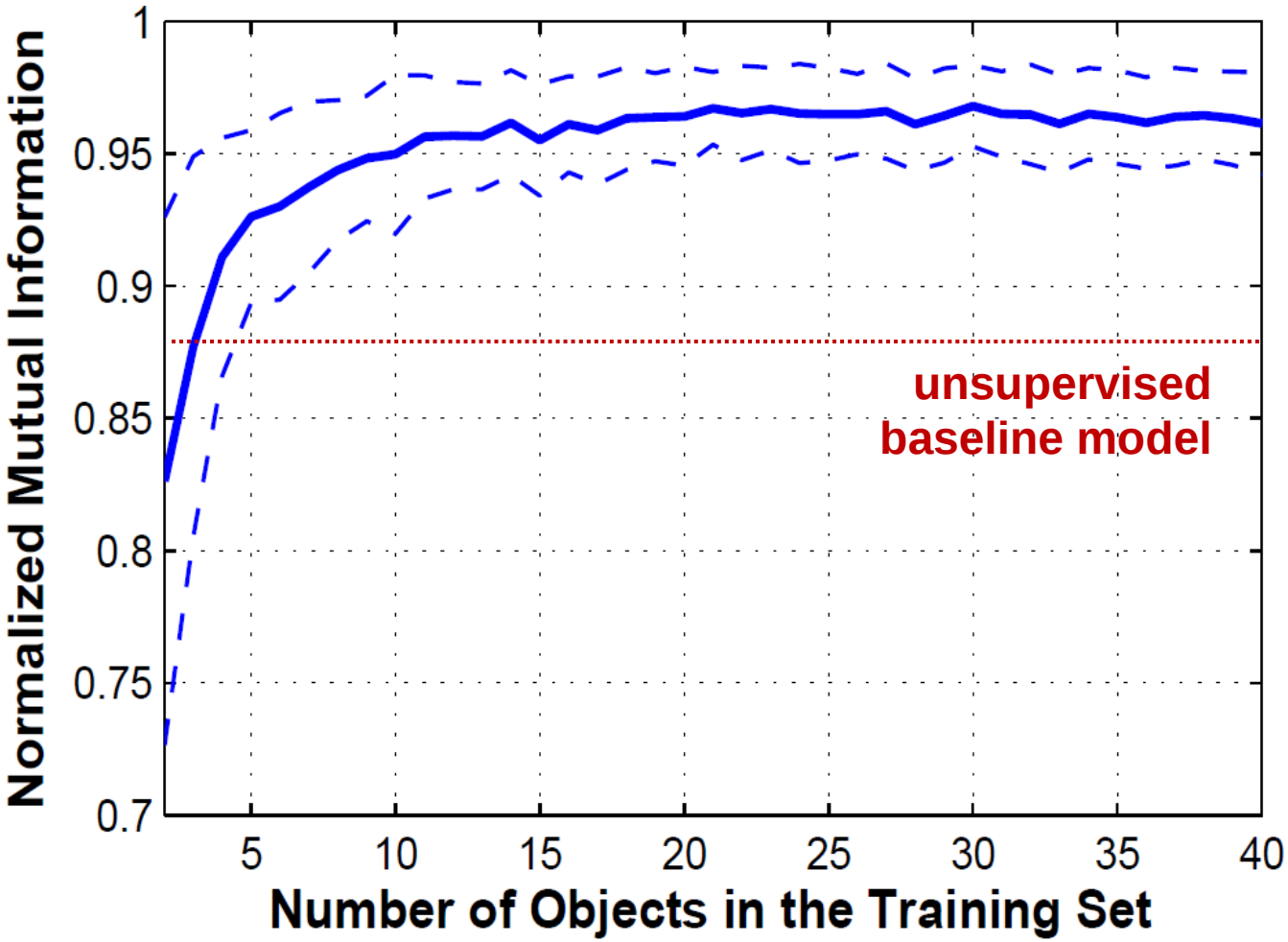


Unsupervised

Baseline Comparison



How many training objects do we need?



Summary

- The robot discovered novel objects using exploratory behaviors coupled with multiple sensory modalities
- Unsupervised clustering alone is insufficient for solving the object individuation problem
- Object discovery can be improved using small amounts of training data

Developmental Progression

Claims

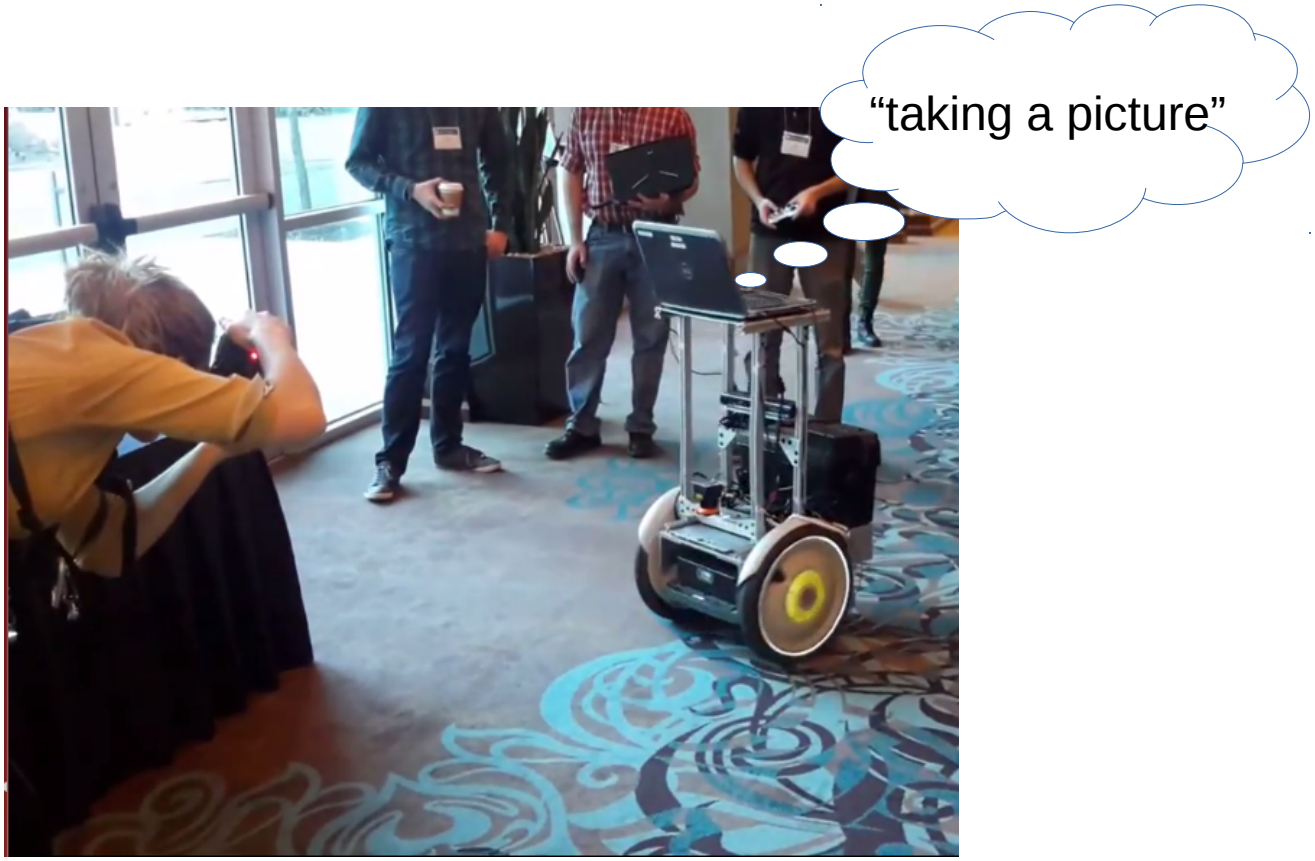
- Behaviors allow robots not only to affect the world, but also to perceive it
- Non-visual sensory feedback improves object classification and perception tasks that are typically solved using vision alone
- A diverse sensorimotor repertoire is necessary for scaling up object recognition, categorization, and individuation to a large number of objects

What's next?

Learning 'in the wild'

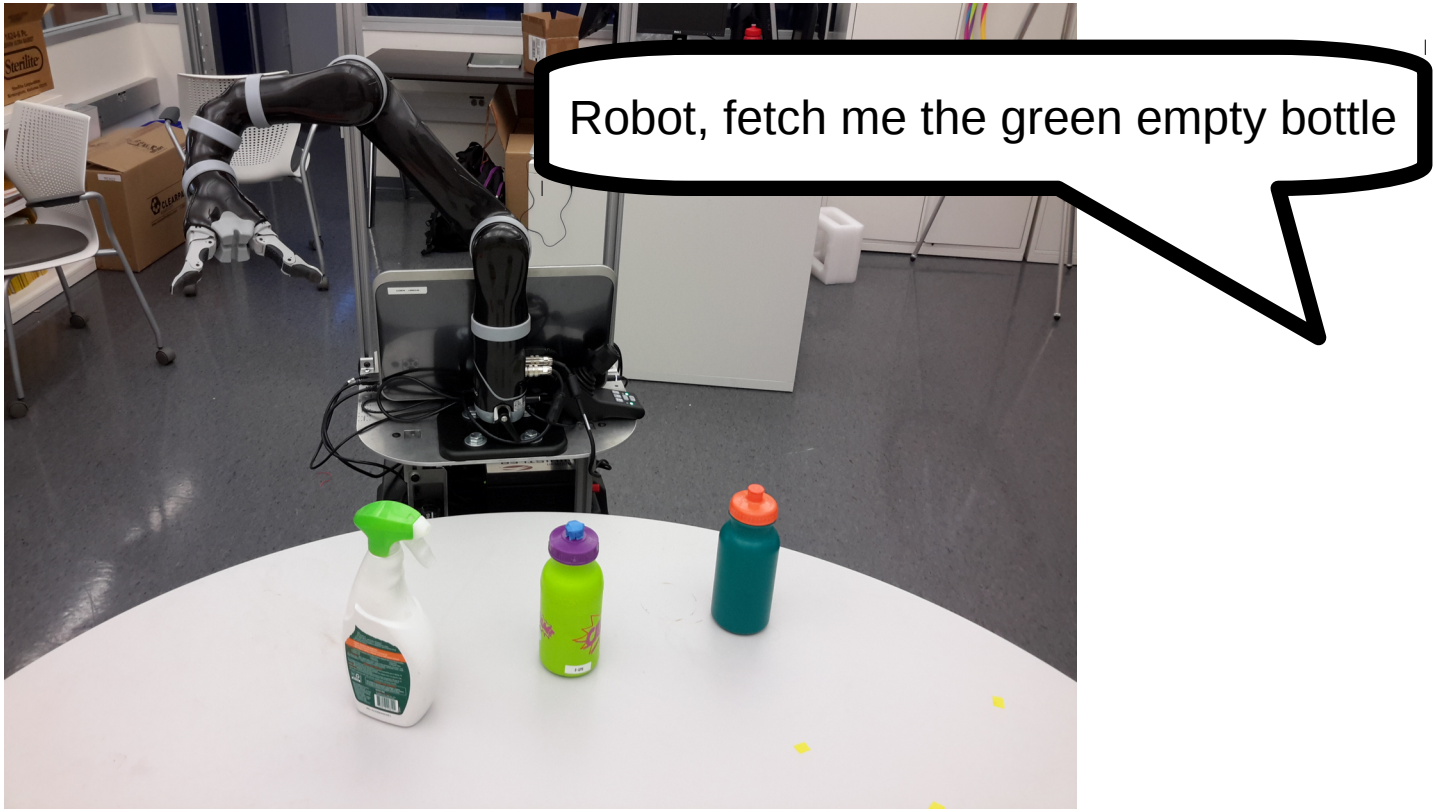
Gori *et al*, "Robot-Centric Activity Recognition 'in the Wild'"
In proceedings of the 7th International Conference on Social Robotics (ICSR 2015)

Learning 'in the wild'



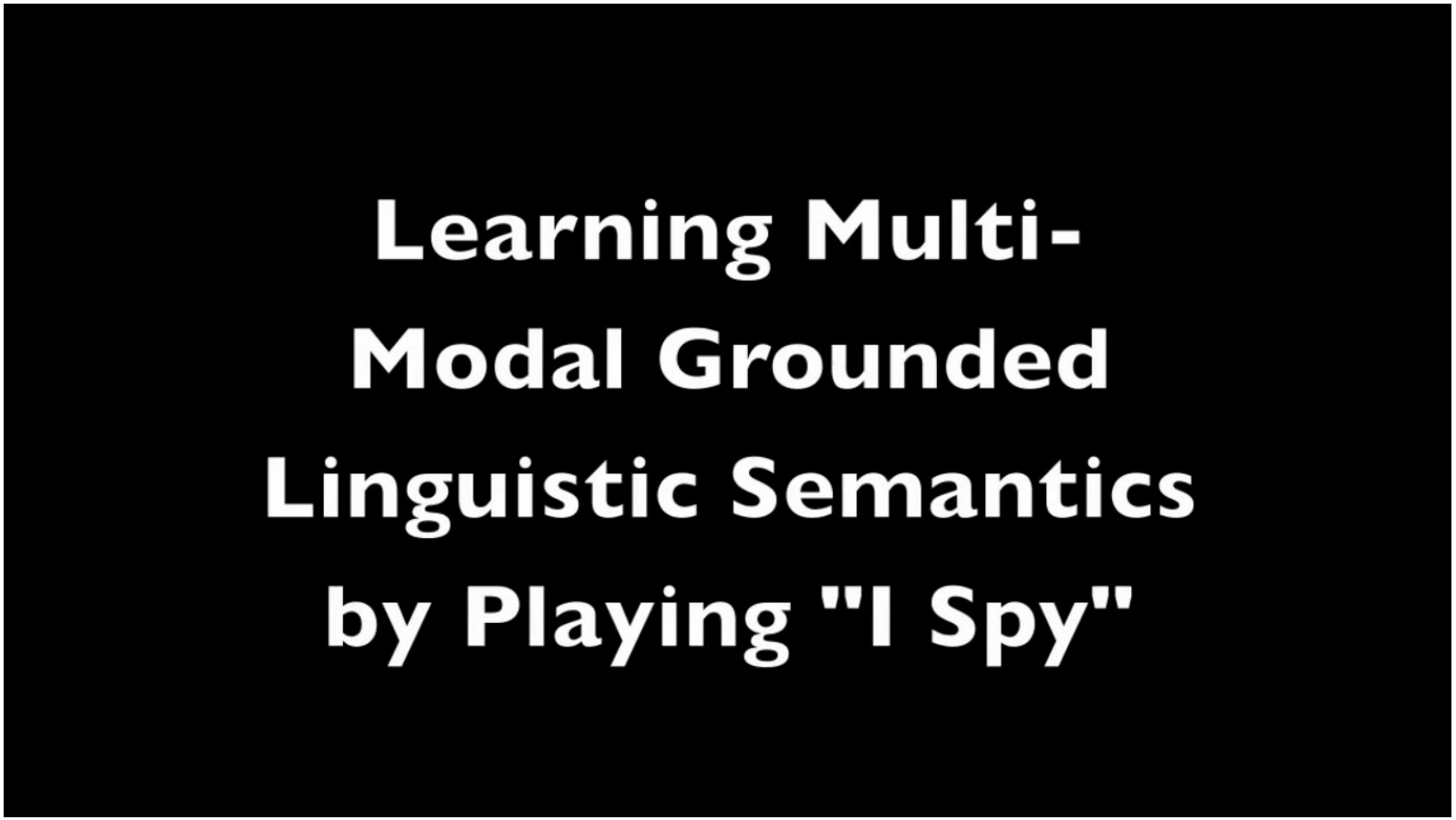
Gori *et al*, “Robot-Centric Activity Recognition 'in the Wild'”
In proceedings of the 7th International Conference on Social Robotics (ICSR 2015)

Language Learning



Thomason *et al*, "Learning Multi-Modal Grounded Linguistic Semantics by Playing I, Spy"
In proceedings of the 2016 International Joint Conference on Artificial Intelligence

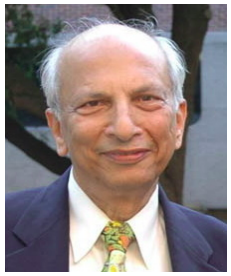
Language Learning



Thomason *et al*, "Learning Multi-Modal Grounded Linguistic Semantics by Playing I, Spy"
In proceedings of the 2016 International Joint Conference on Artificial Intelligence

Acknowledgements

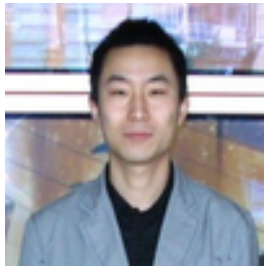

Peter Stone


J. K. Aggrawal

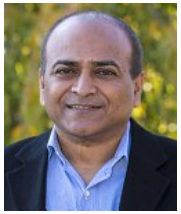

Ray Mooney


Ilaria Gori


Matteo Leonetti


Shiqi Zhang


Alex Stoytchev

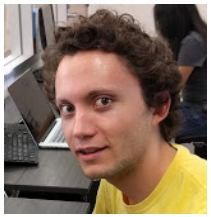

Vasant Honavar


Heiko Hoffmann


Cornelia Caragea


Sanmit Narvekar

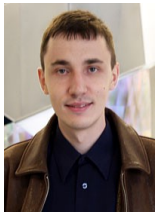

Proiyanka Khante

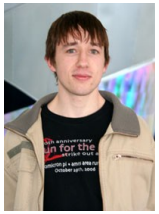

Maxwel Svetlik


Jesse Thomason


Piyush Khandelwal


Nick Walker


Vlad Sukhoy


Shane Griffith


Connor Schenck


Ritika Sahai


Taylor Bergquist


Adrian Silvescu

Acknowledgements

Tufts Computer Science

UT CS COMPUTER SCIENCE

IOWA STATE UNIVERSITY

VIRTUAL REALITY APPLICATIONS CENTER

JOHN DEERE

Willow Garage

HRL LABORATORIES

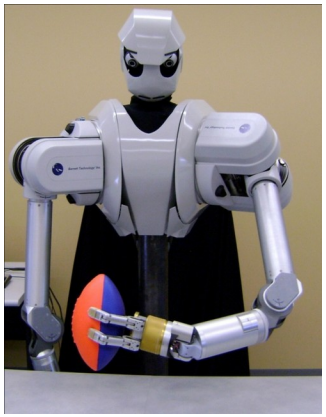
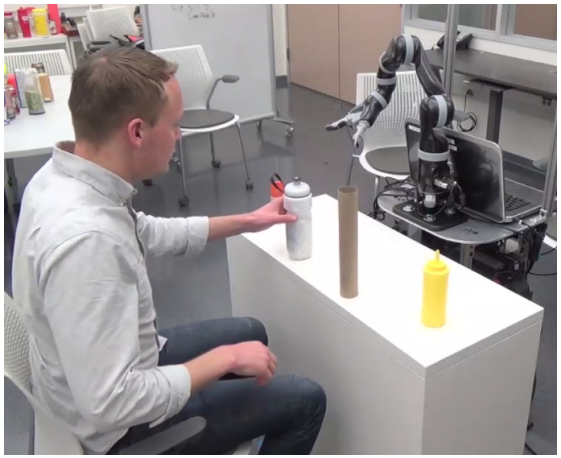
AFRL THE AIR FORCE RESEARCH LABORATORY LEAD | DISCOVER | DEVELOP | DELIVER

NSF

...and a big thanks to my undergrads students

Thanks! Questions?




Thank You