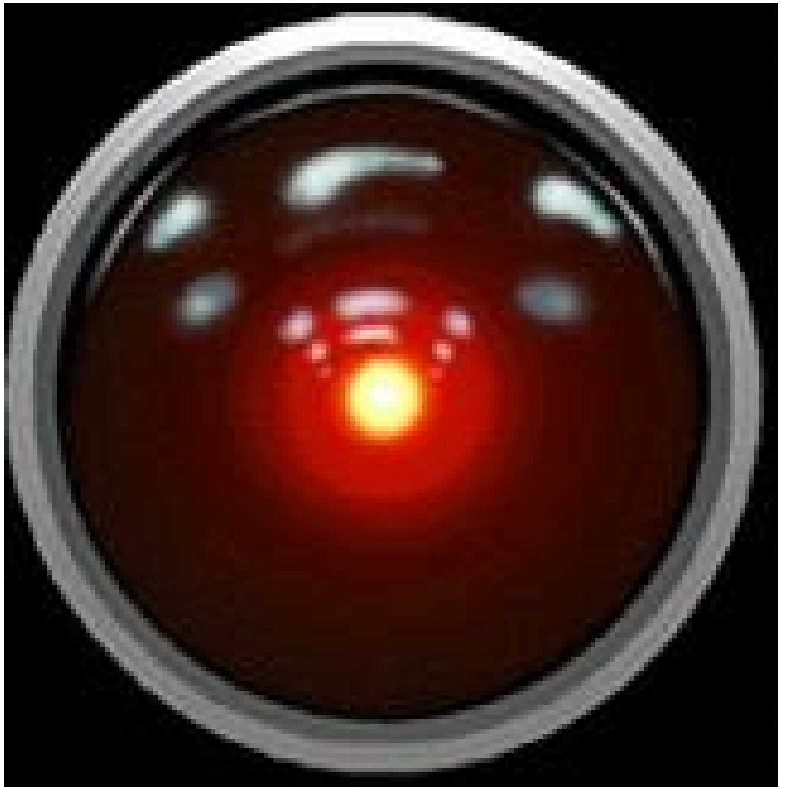
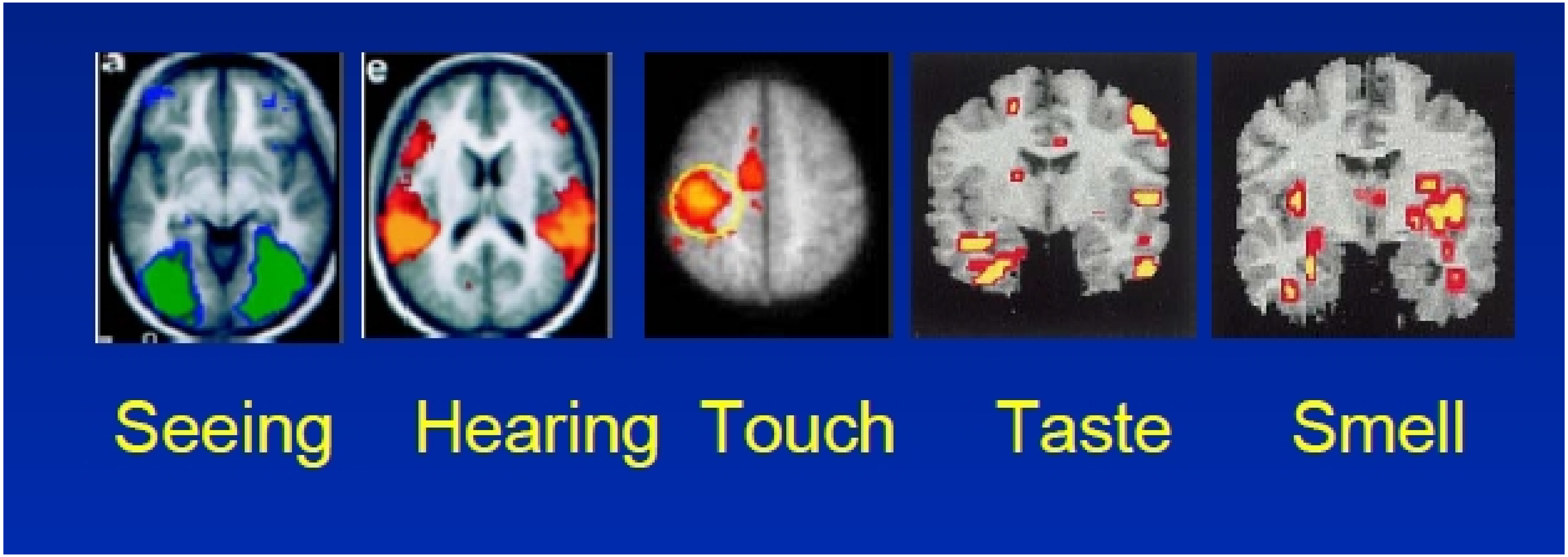


CS 378: Autonomous Intelligent Robotics

Instructor: Jivko Sinapov

http://www.cs.utexas.edu/~jsinapov/teaching/cs378/

Multimodal Perception



Announcements

Final Projects Presentation Date:

Thursday, May 12, 9:00-12:00 noon

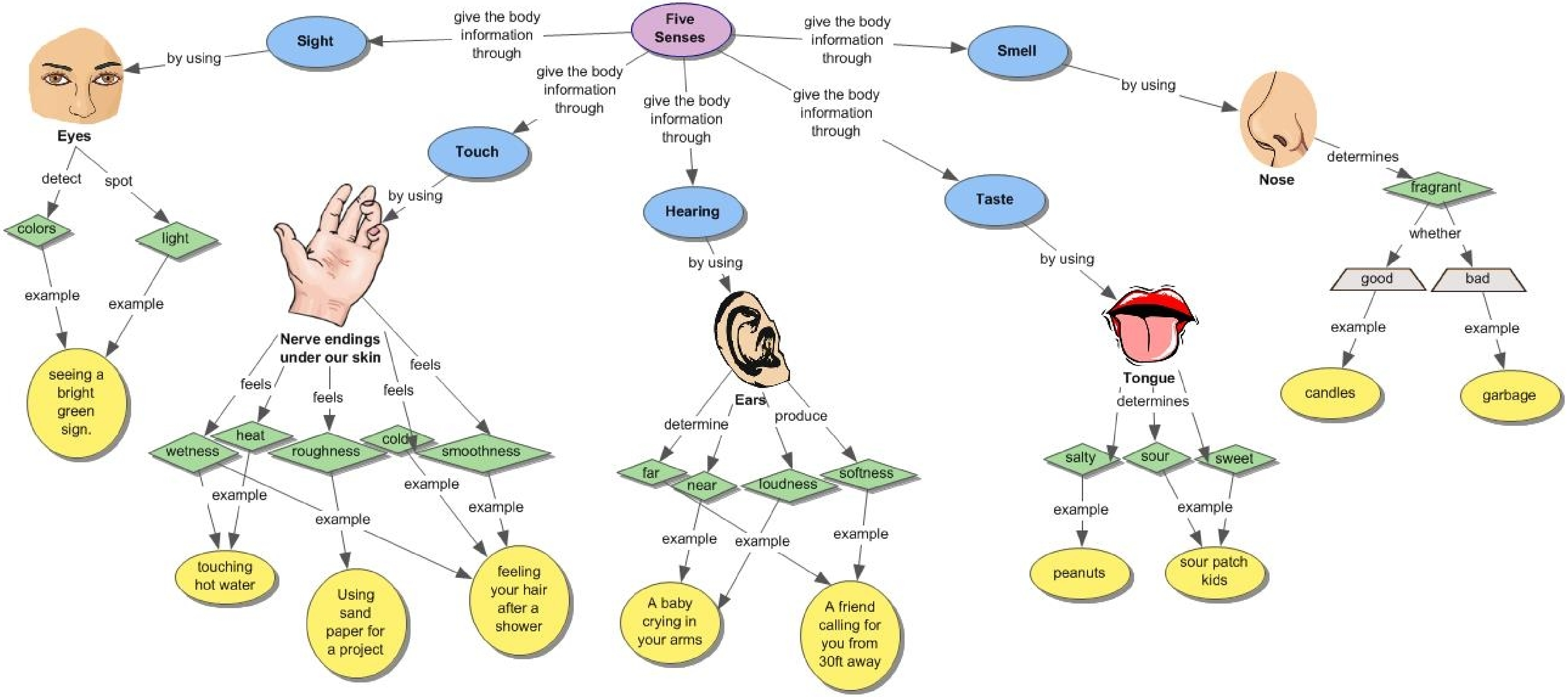
Project Deliverables

- Final Report (6+ pages in PDF)
- Code and Documentation (posted on github)
- Presentation including video and/or demo

Multi-modal Perception

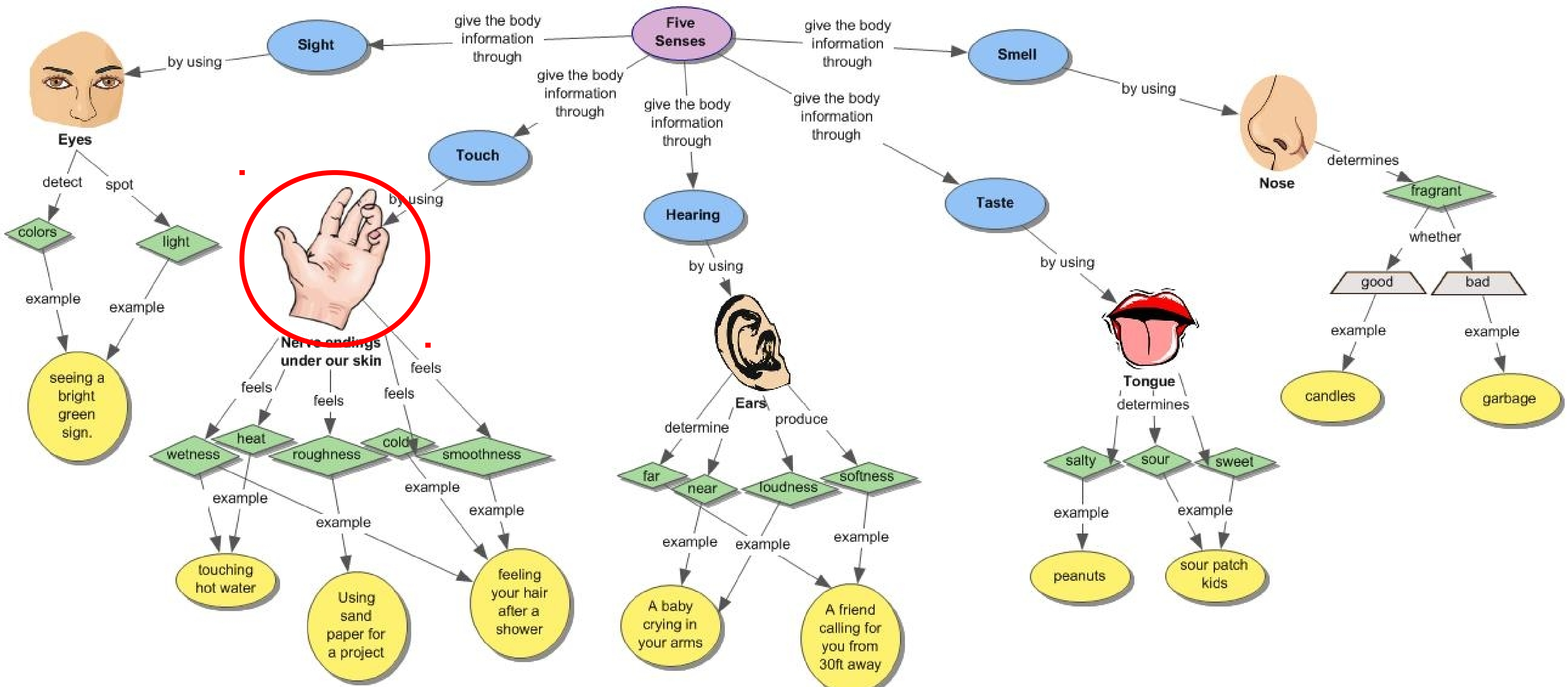
The “5” Senses

The "5" Senses

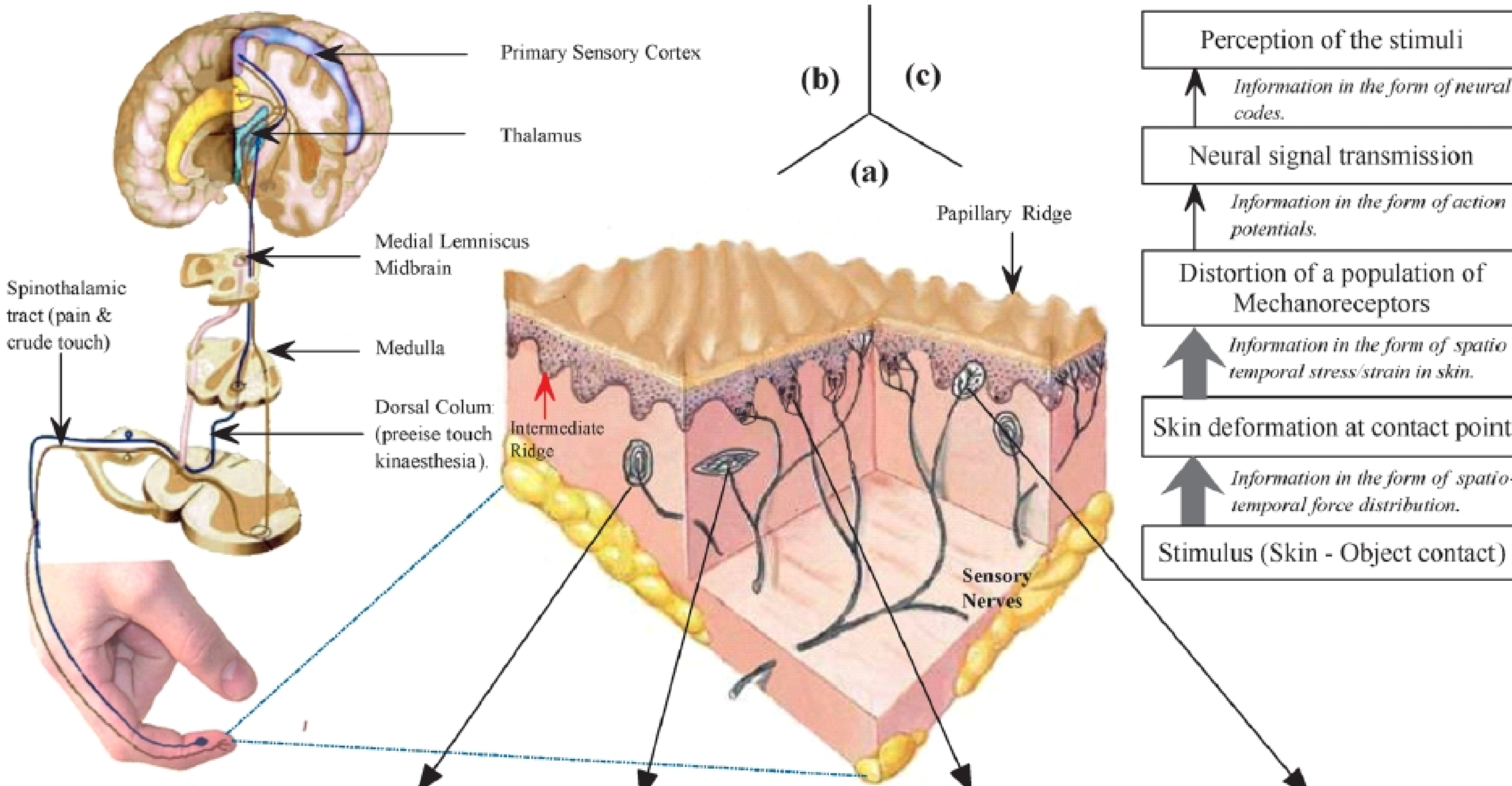


[http://edublog.cmich.edu/meado1bl/files/2013/03/Five-Senses2.jpg]

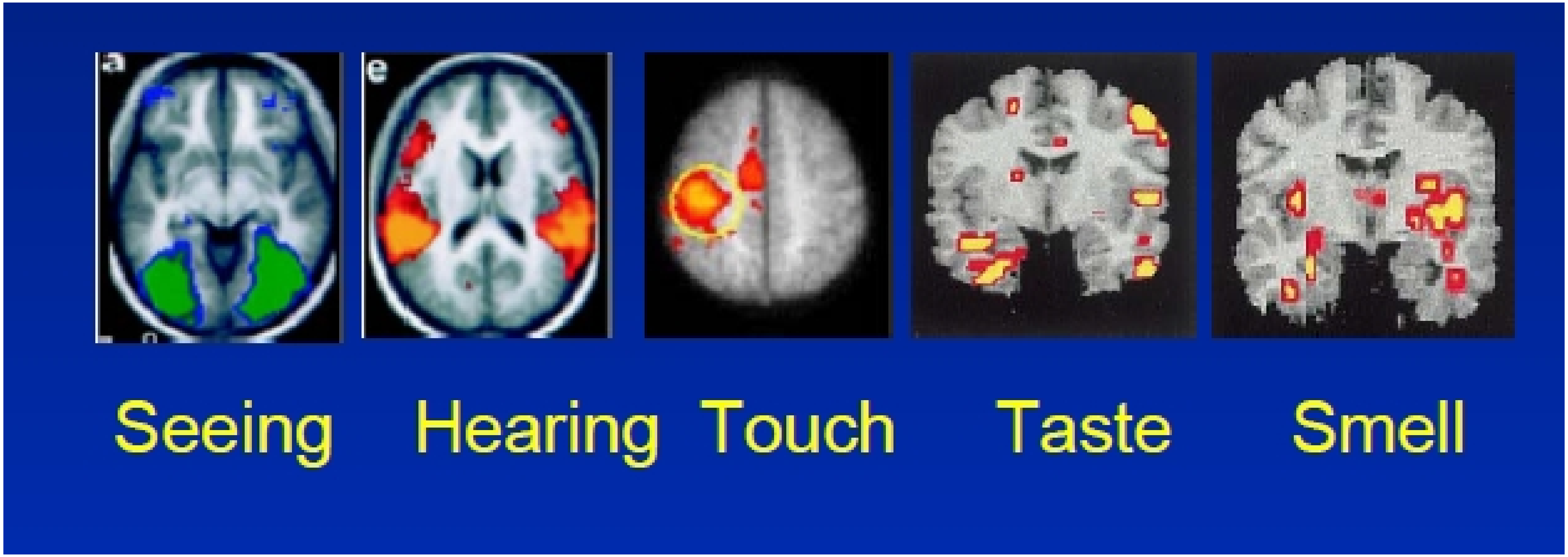
The “5” Senses



[http://edublog.cmich.edu/meado1bl/files/2013/03/Five-Senses2.jpg]



Classification Basis		Pacinian Corpuscle	Ruffini Corpuscle	Merkel Cells	Meissner's Corpuscle
Type		FA II	SA II	SA I	FA I
Adaptation Rate		Fast	Slow	Slow	Fast
Spatial Acuity (mm)		10+	7+	0.5	3-4
Vibration/rapid indent. threshold	Best(μm)	0.01	40	8	2
	Mean(μm)	0.08	300	30	6
Stimuli Frequency (Hz)		40-500+	100-500+	0.4-3	3-40
Conduction Velocity (m/s)		35-70	35-70	40-65	35-70
Effective Stimuli		Temporal changes in the skin deformation	Sustained downward Pressure; Lateral skin stretch; Skin slip.	Spatial deformation; Sustained pressure; Curvature, edge, corners.	Temporal changes in skin deformation
Sensory Function		High frequency vibration detection; Tool use.	Finger position; Stable grasp; Tangential Force; Motion direction	Pattern/form detection; texture perception; Tactile flow perception.	Low frequency vibration & motion detection; Grip control; Tactile flow perception.



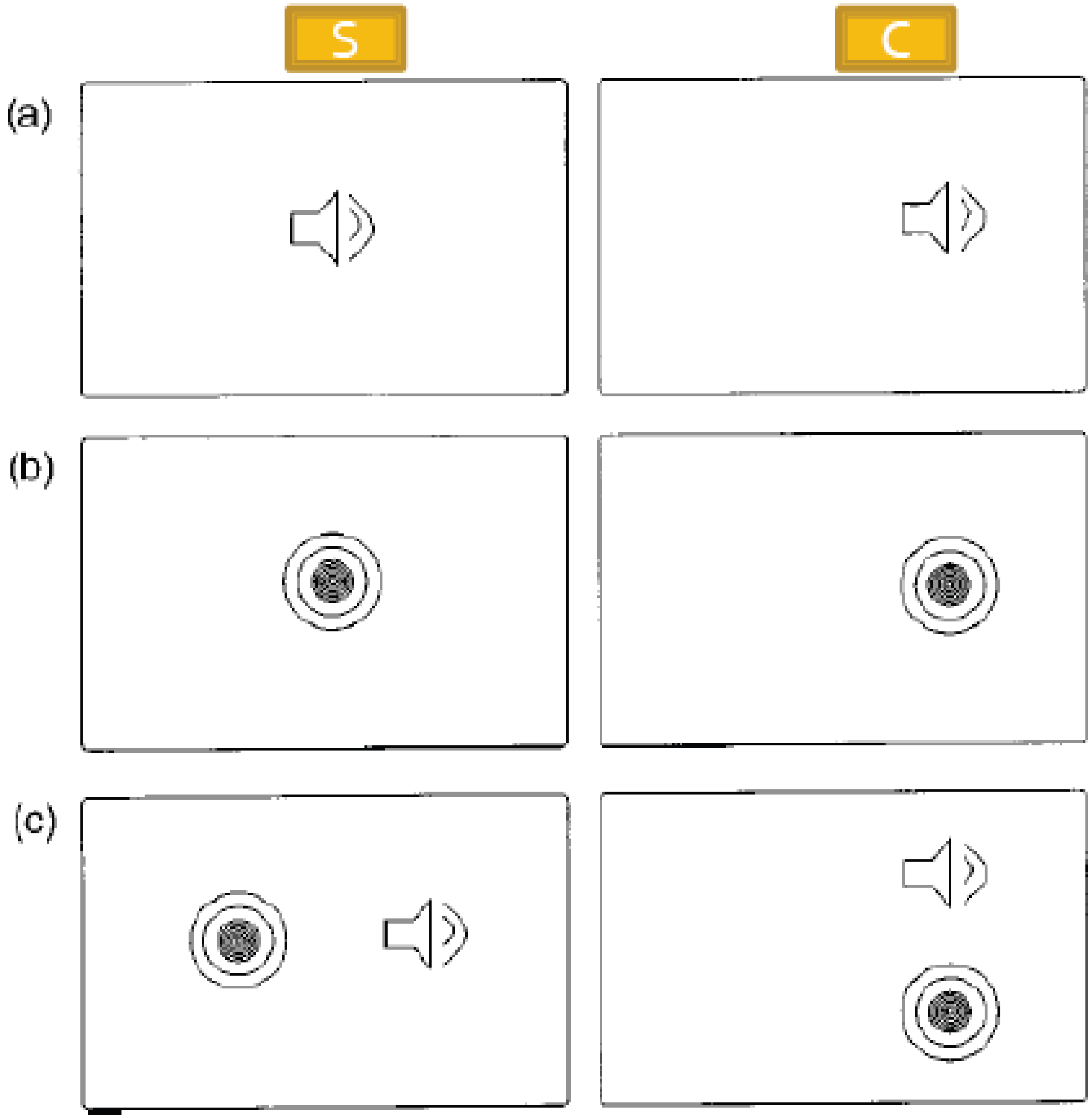
[http://neurolearning.com/sensoryslides.pdf]

MAKING SENSE OF THE SENSES

There are many opinions about how many senses we have

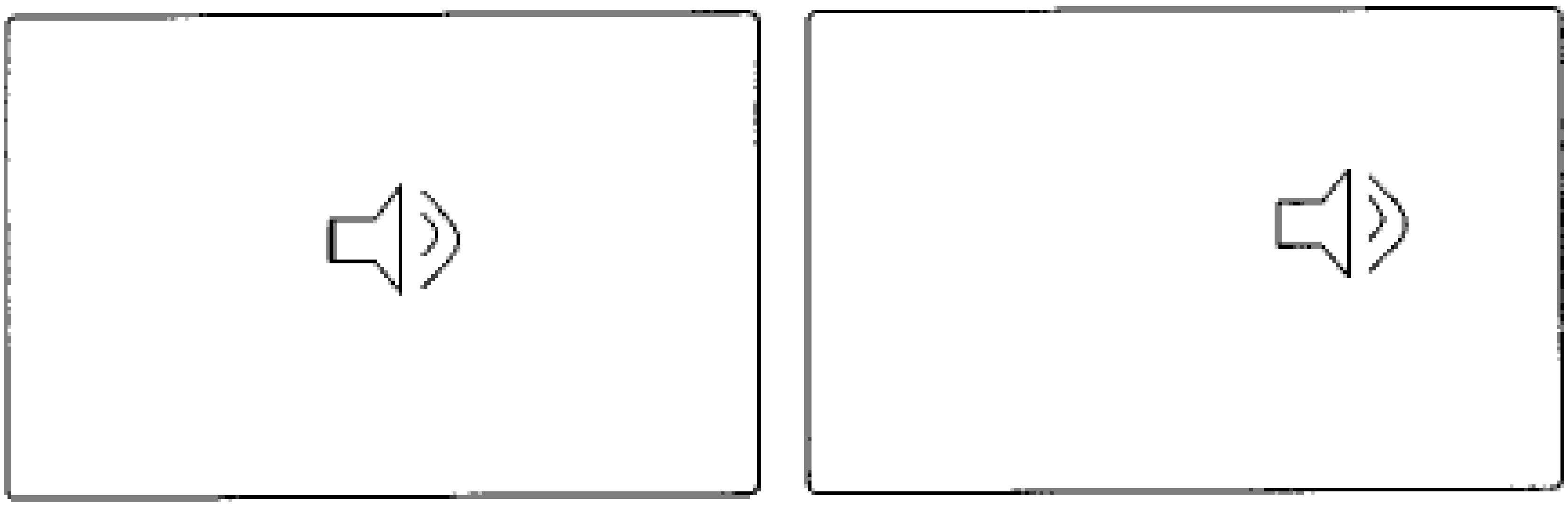
SENSORY MODALITY	Conservative	Accepted	Radical
Vision	■		
Light		■	■
Colour		■	
Red			■
Green			■
Blue			■
Hearing	■	■	■
Smell	■	■	
2000 or more receptor types			■
Taste	■		
Sweet		■	■
Salt		■	■
Sour		■	■
Bitter		■	■
Umami			■
Touch	■	■	
Light touch			■
Pressure			■
Pain	■	■	
Cutaneous			■
Somatic			■
Visceral			■
Mechanoreception	■		
Balance	■	■	
Rotational acceleration			■
Linear acceleration			■
Proprioception – joint position		■	■
Kinaesthesis		■	
Muscle stretch – Golgi tendon organs			■
Muscle stretch – muscle spindles			■
Temperature	■		
Heat		■	■
Cold		■	■
Interoceptors			
Blood pressure	■	■	
Arterial blood pressure			■
Central venous blood pressure			■
Head blood temperature			■
Blood oxygen content		■	■
Cerebrospinal fluid pH		■	■
Plasma osmotic pressure (thirst?)		■	■
Artery-vein blood glucose difference (hunger?)		■	■
Lung inflation		■	■
Bladder stretch			■
Full stomach			■
TOTAL	10	21	33

How are sensory signals from different modalities integrated?



[Battaglia et. al. 2003]

Locating the Stimulus Using a Single Modality

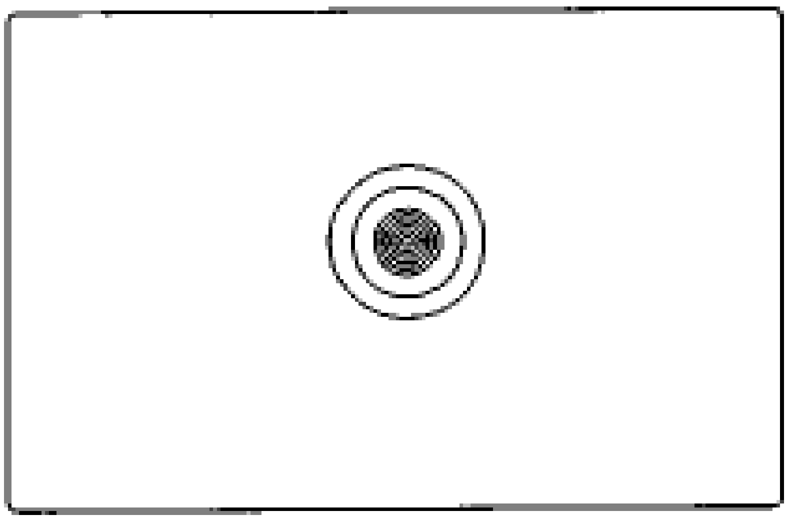


Standard Trial

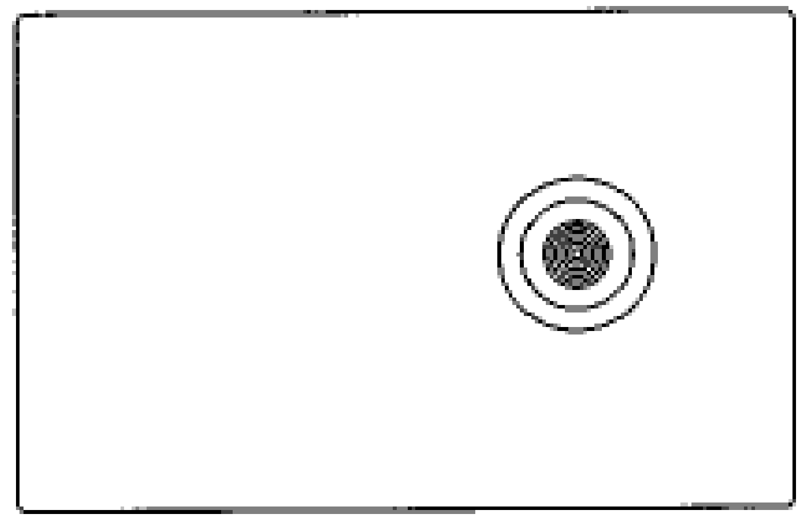
Comparison Trial

Is the stimulus in Trial 2 located to the left or to the right of the stimulus in Trial 1?

Locating the Stimulus Using a Single Modality



Standard
Trial



Comparison
Trial

Is the stimulus in Trial 2 located to the left or to the right of the stimulus in Trial 1?

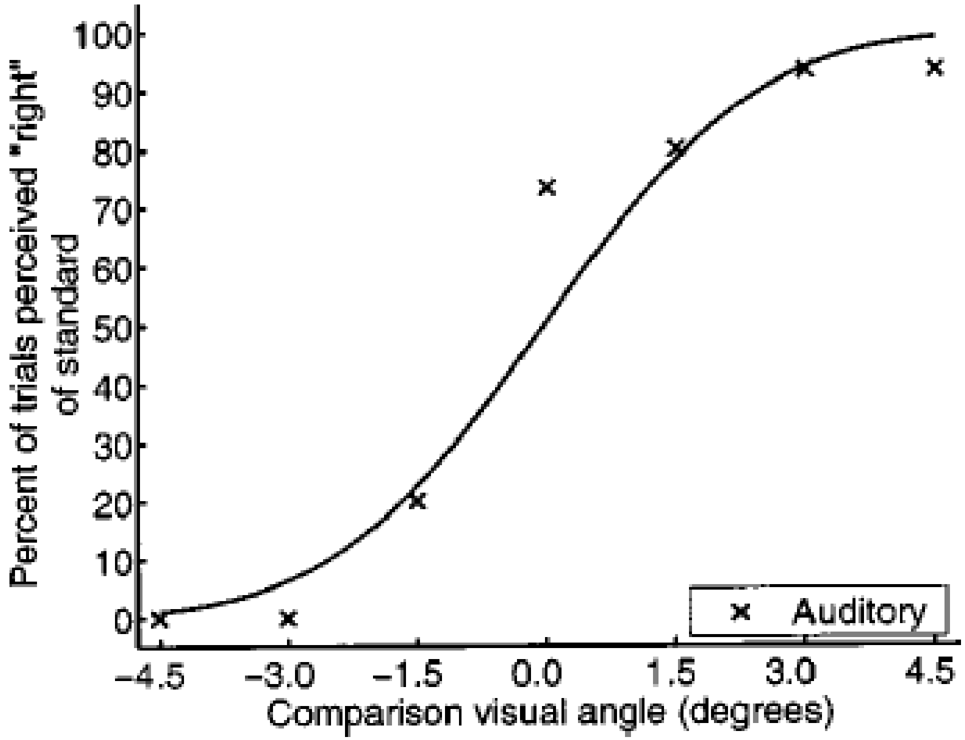


Fig. 3. Results for one subject on the auditory-only trials. The horizontal axis shows the comparison locations (in degrees of visual angle away from the center of the workspace), and the vertical axis shows the percentage of trials in which the subject judged the comparison stimulus as depicting an event located to the right of the event depicted in the standard stimulus. The curve fitted to the data points is a cumulative normal distribution.

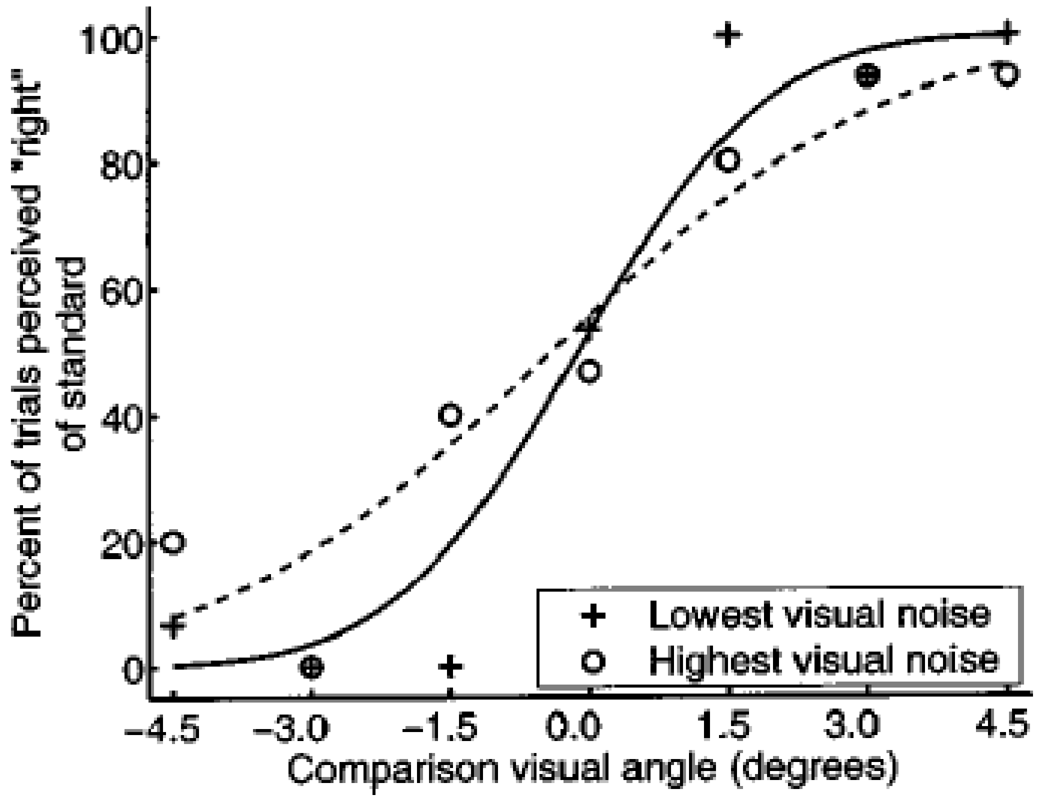
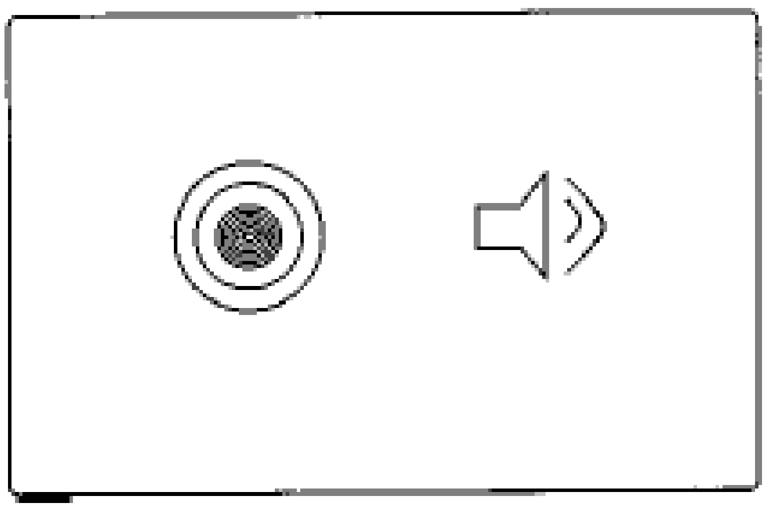
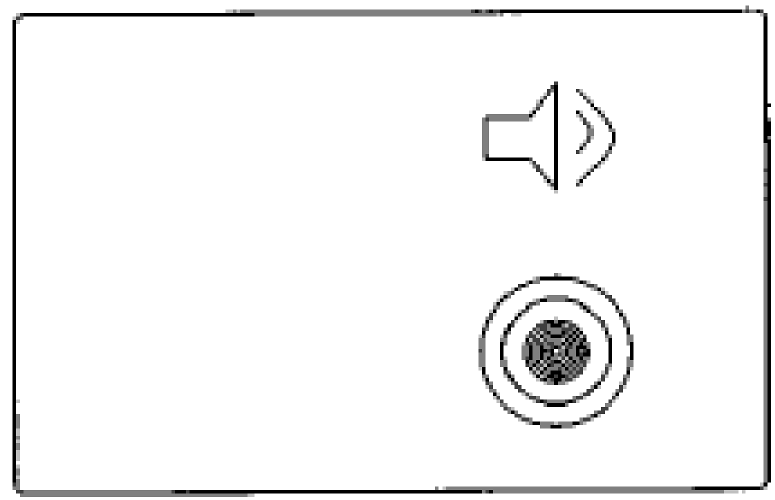


Fig. 4. Results for one subject on the visual-only trials. The solid and dashed curves are cumulative normal distributions fitted to the data points in the lowest-noise and highest-noise conditions, respectively.

Multimodal Condition



Standard
Trial



Comparison
Trial

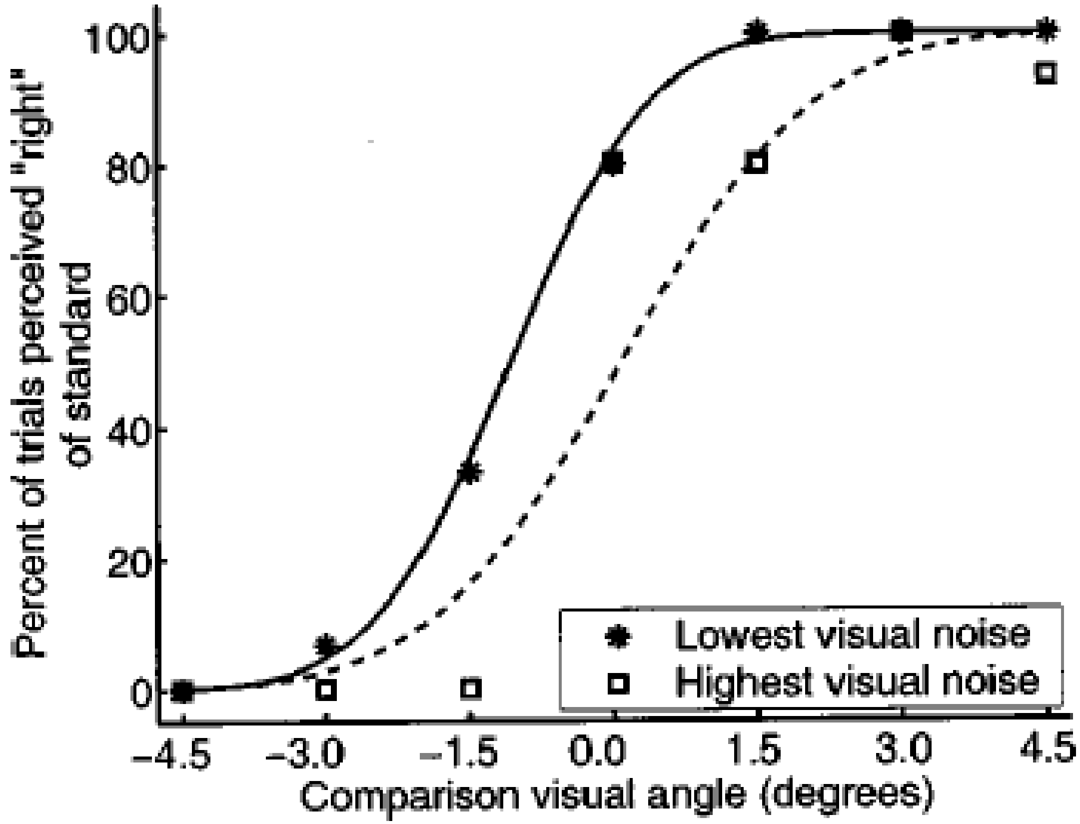
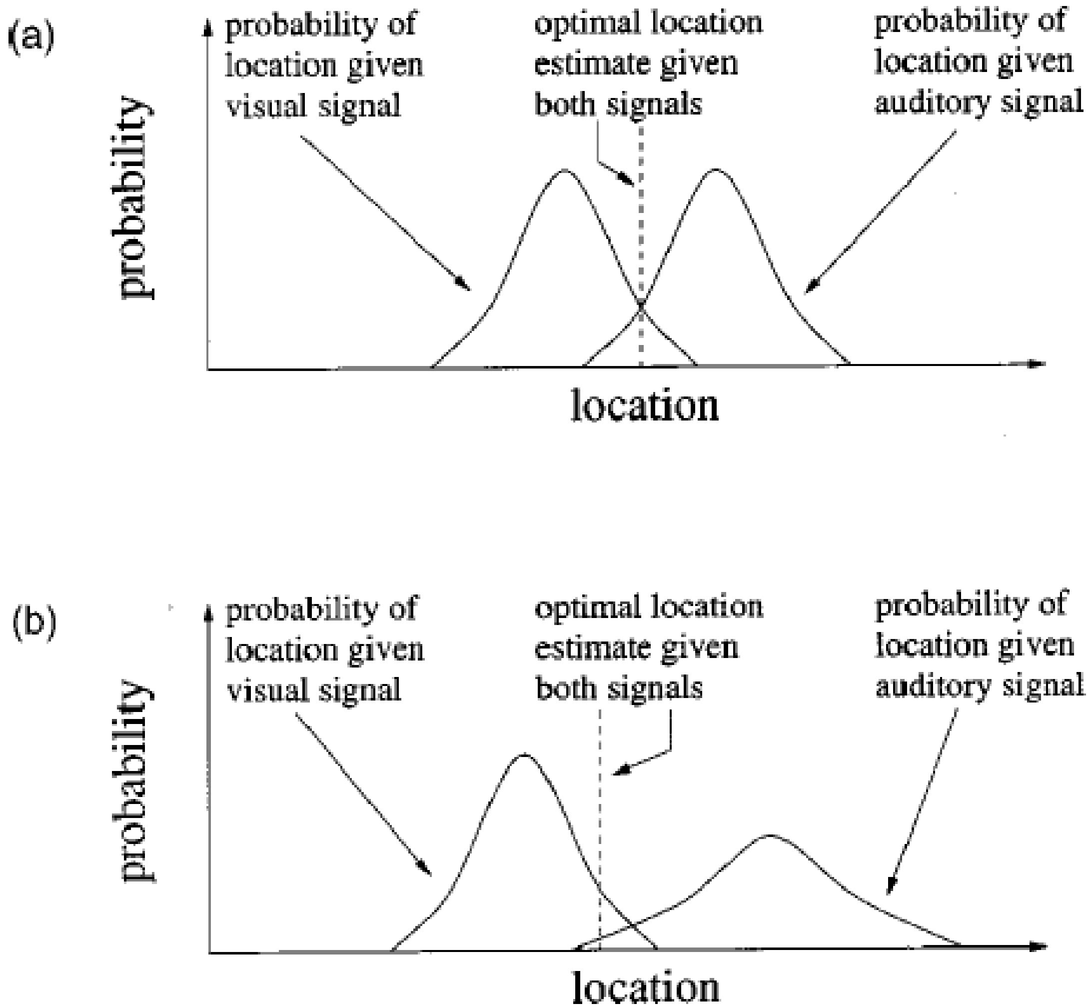


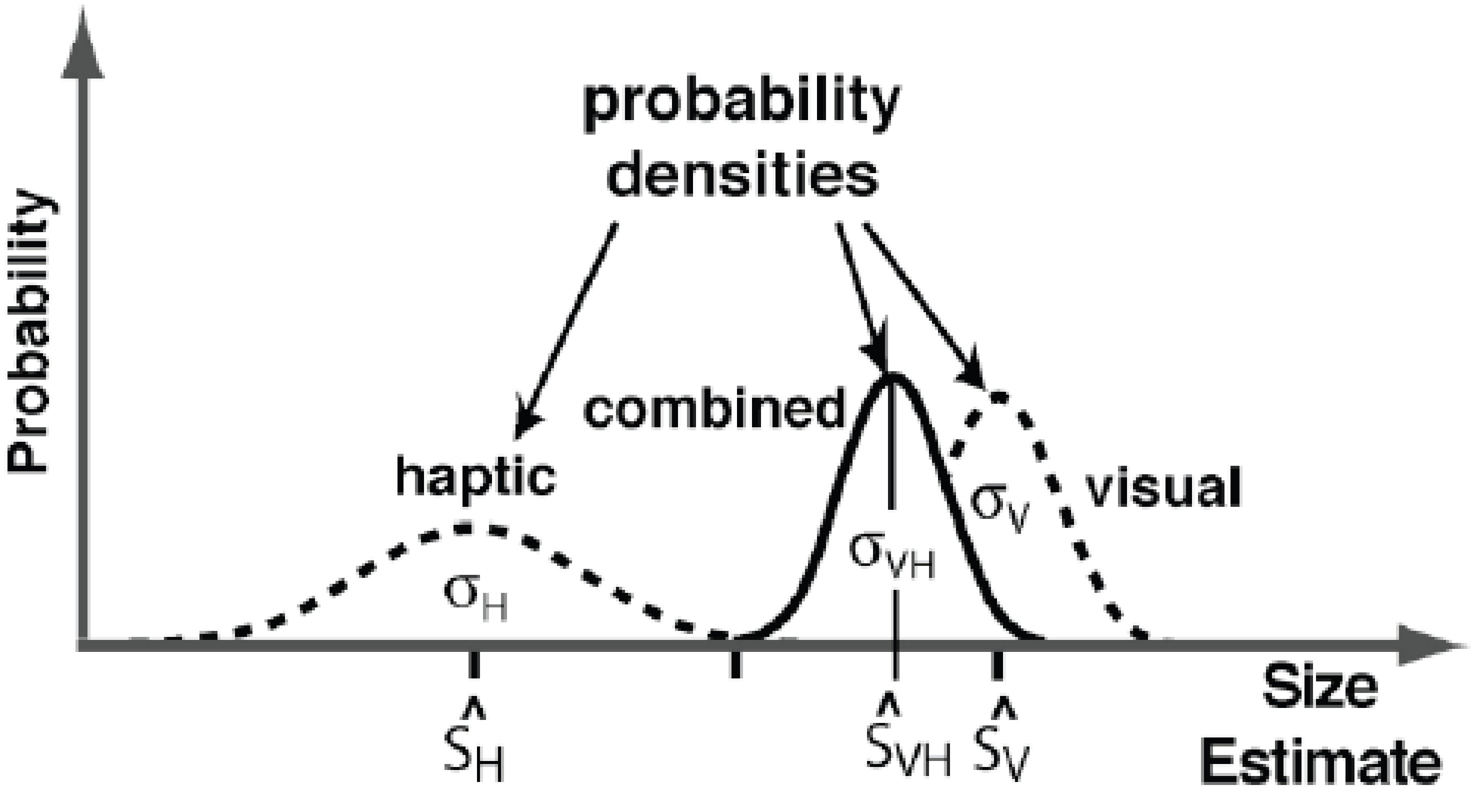
Fig. 5. Results for one subject on the visual–auditory trials. The solid and dashed curves are cumulative normal distributions fitted to the data points in the lowest-noise and highest-noise conditions, respectively.



$$L^* = w_v L_v^* + w_a L_a^* ,$$

$$w_v = \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2} \quad \text{and} \quad w_a = \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2}$$

Fig. 1. Optimal model of sensory integration based on MLE theory. (a) Visual and auditory signals are equally reliable indicators of event location. (b) Visual signal is a more reliable indicator of event location.



probability
densities
Probability
combined
haptic
visual
σ_V
σ_{VH}
σ_H
Size
Estimate
$\hat{S}_H$
$\hat{S}_{VH}$
$\hat{S}_V$

[Ernst, 2006]

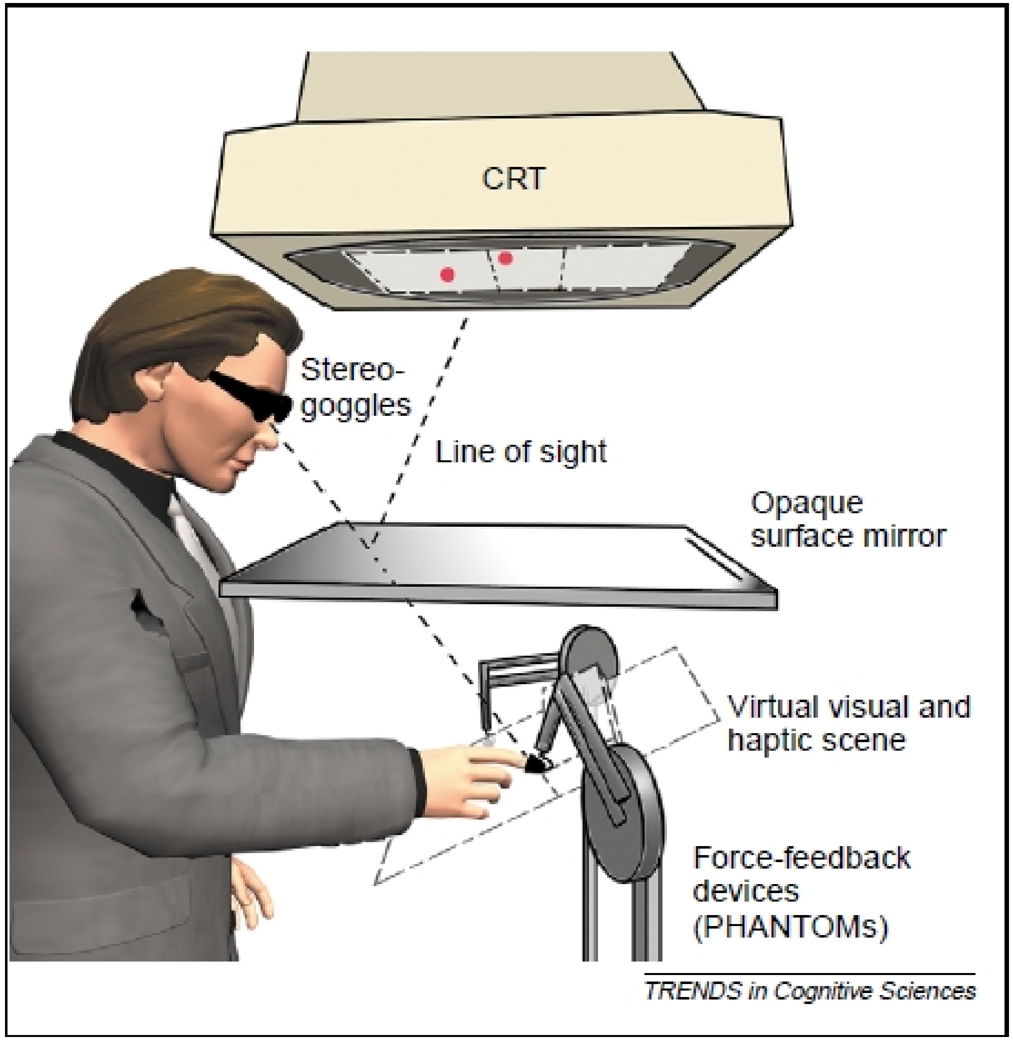


Figure 2. In this visual–haptic set-up used by Ernst and Banks [29] observers view the reflection of the visual stimulus binocularly in a mirror using stereo-goggles. The haptic stimulus is presented with two PHANTOM™ force-feedback devices, one each for the index finger and thumb of the right hand. With this arrangement the visual and the haptic virtual scenes can be independently manipulated.

Take-home Message

During integration, sensory modalities are weighted based on their individual reliability

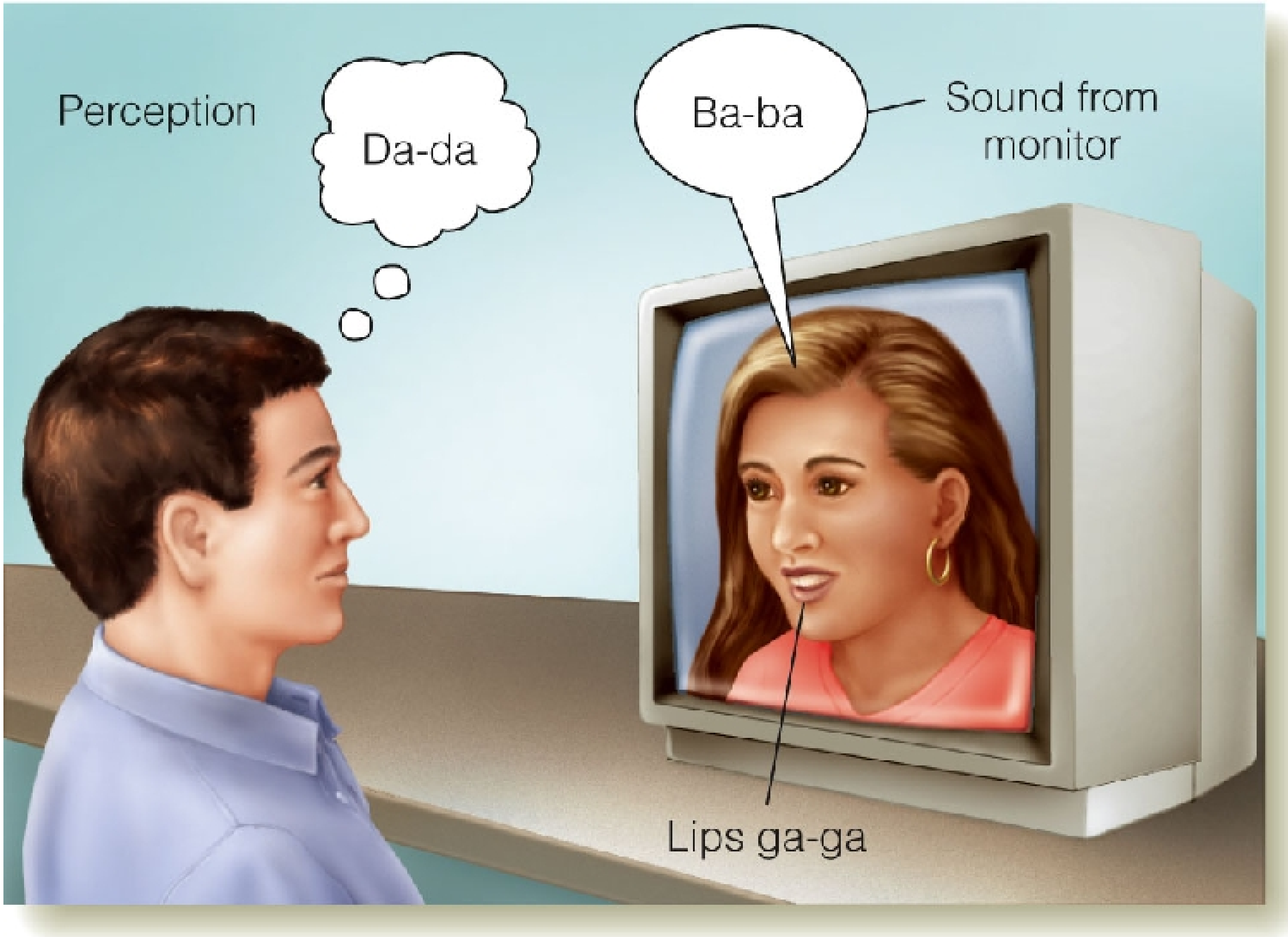
Further Reading

Ernst, Marc O., and Heinrich H. Bülthoff. "Merging the senses into a robust percept." *Trends in cognitive sciences* 8.4 (2004): 162-169.

Battaglia, Peter W., Robert A. Jacobs, and Richard N. Aslin. "Bayesian integration of visual and auditory signals for spatial localization." *JOSA A* 20.7 (2003): 1391-1397.

Sensory Integration During Speech Perception

McGurk Effect

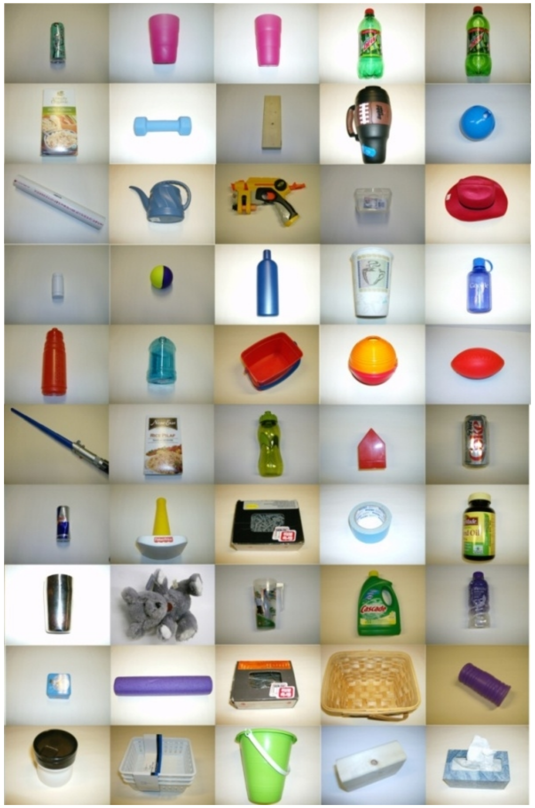


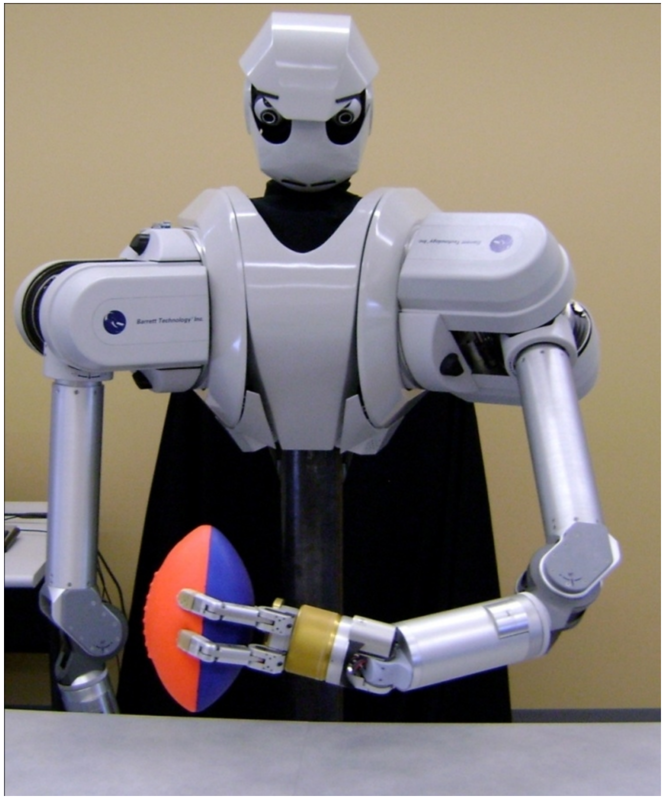
McGurk Effect

https://www.youtube.com/watch?v=G-lN8vWm3m0

https://vimeo.com/64888757

Object Recognition Using Auditory and Proprioceptive Feedback





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

What is Proprioception?

“It is the sense that indicates whether the body is moving with required effort, as well as where the various parts of the body are located in relation to each other.”

- Wikipedia

Why Proprioception?

Why Proprioception?

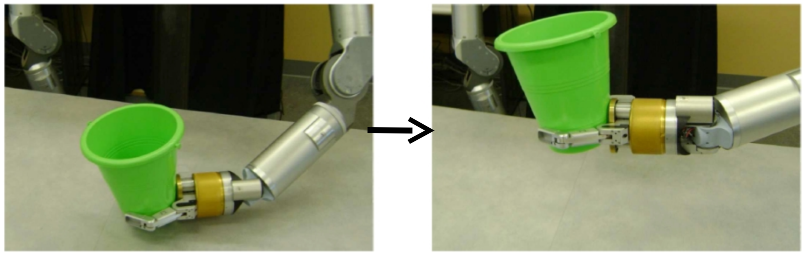
Why Proprioception?

Hard

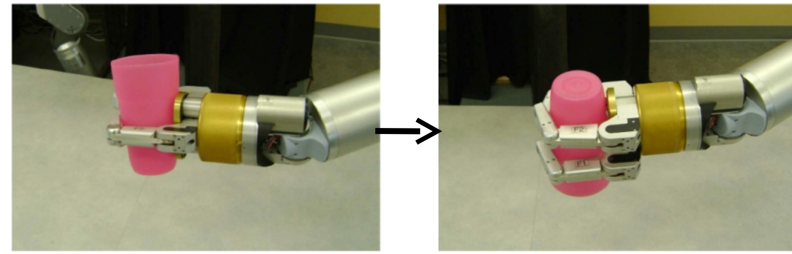
Soft

Exploratory Behaviors

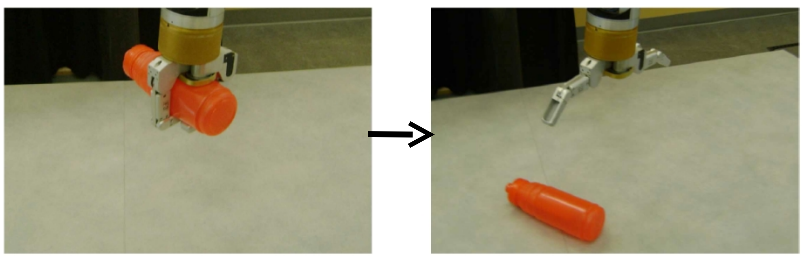
Lift:



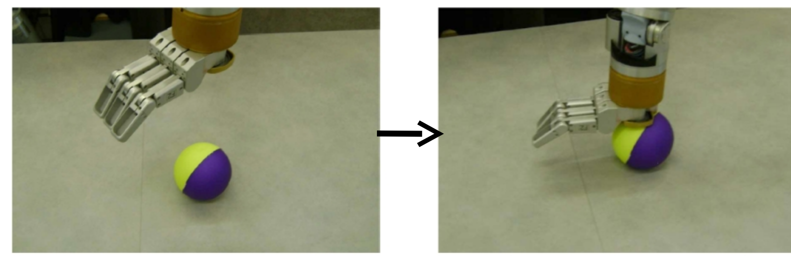
Shake:



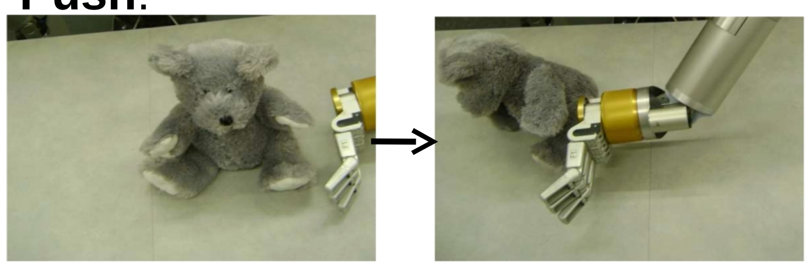
Drop:



Crush:



Push:



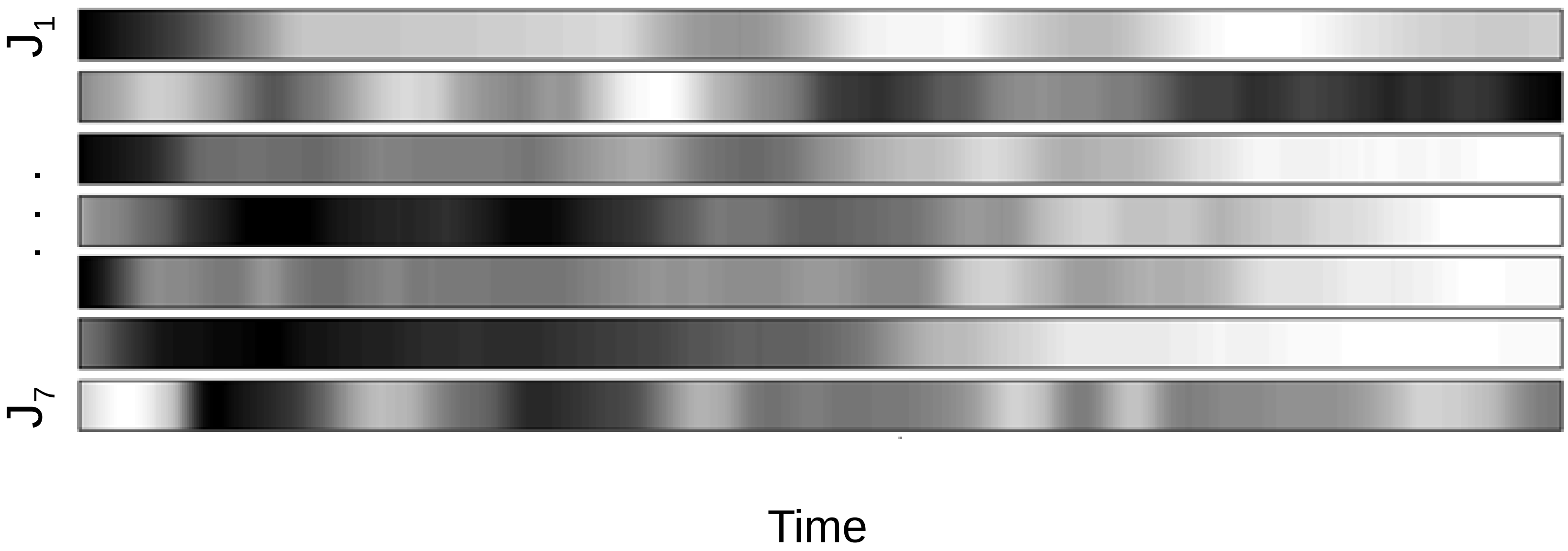
Objects

Sensorimotor Contexts

Behaviors	audio	proprioception
lift	✓	✓
shake	✓	✓
drop	✓	✓
press	✓	✓
push	✓	✓

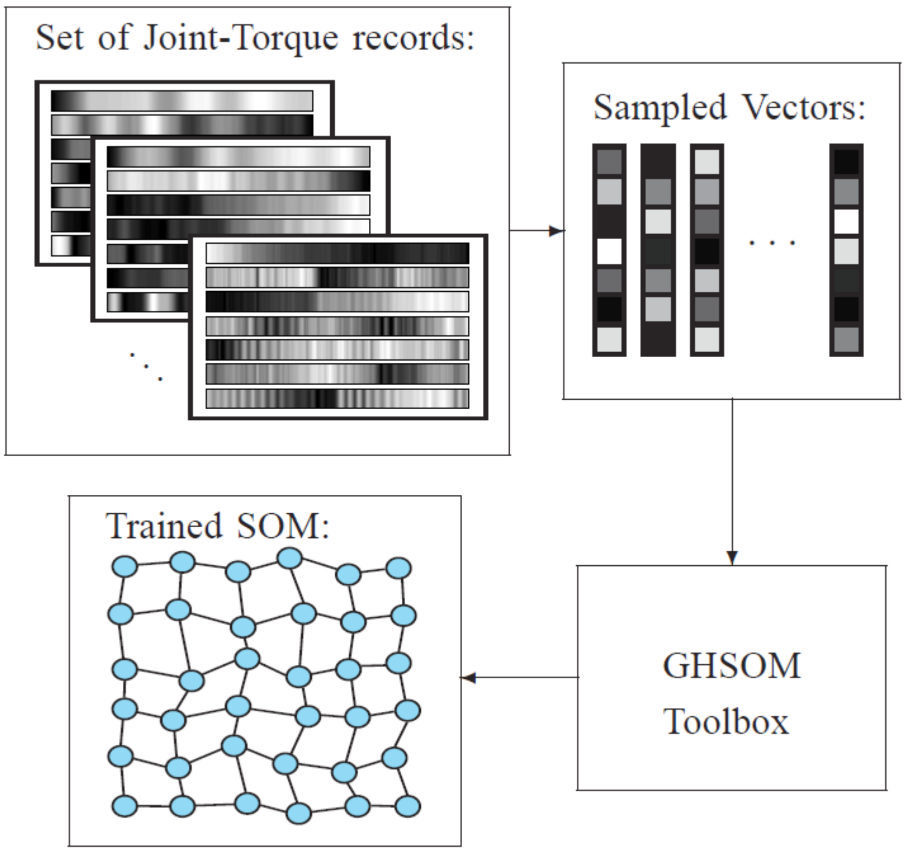
Sensory Modalities

Feature Extraction

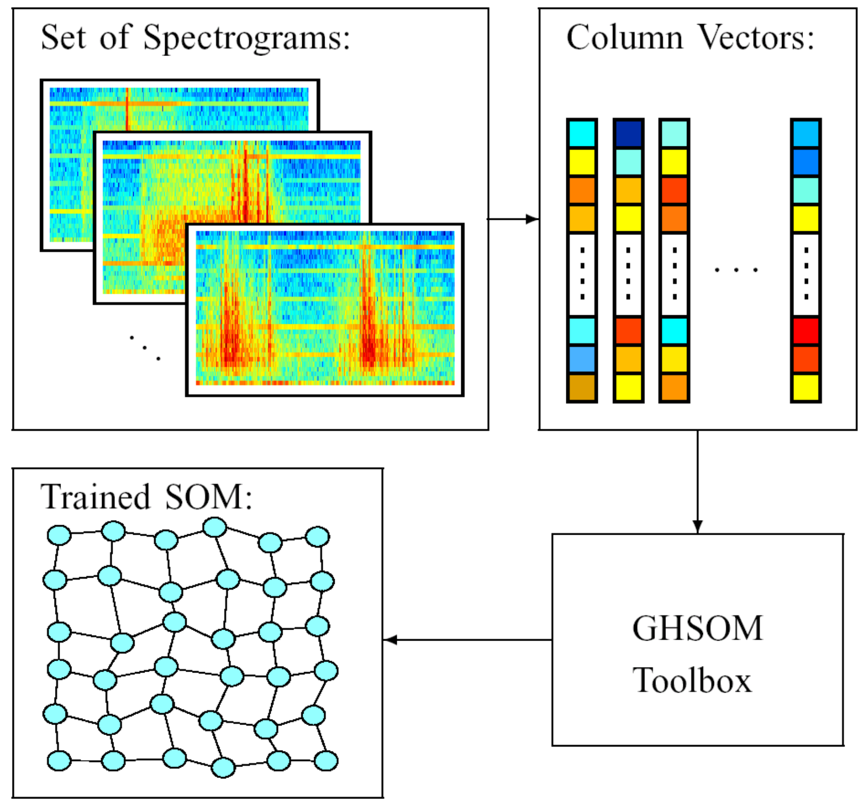


Feature Extraction

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

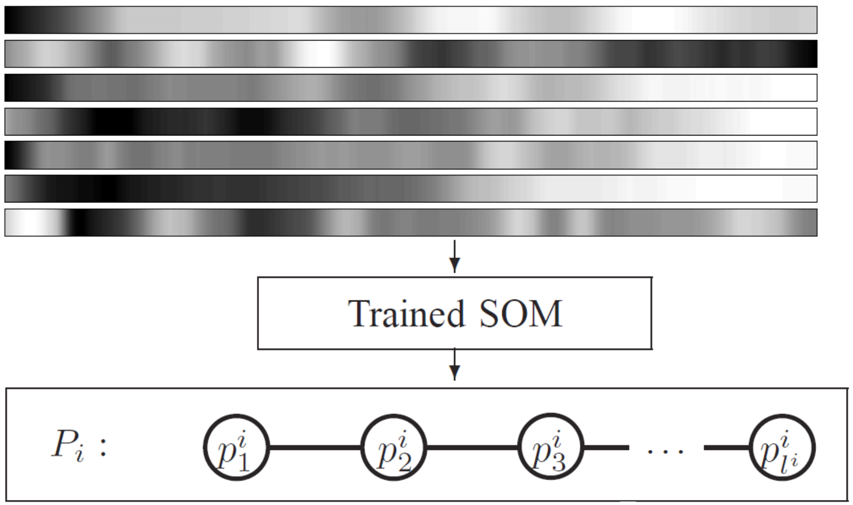


Training an SOM using sampled frequency distributions:



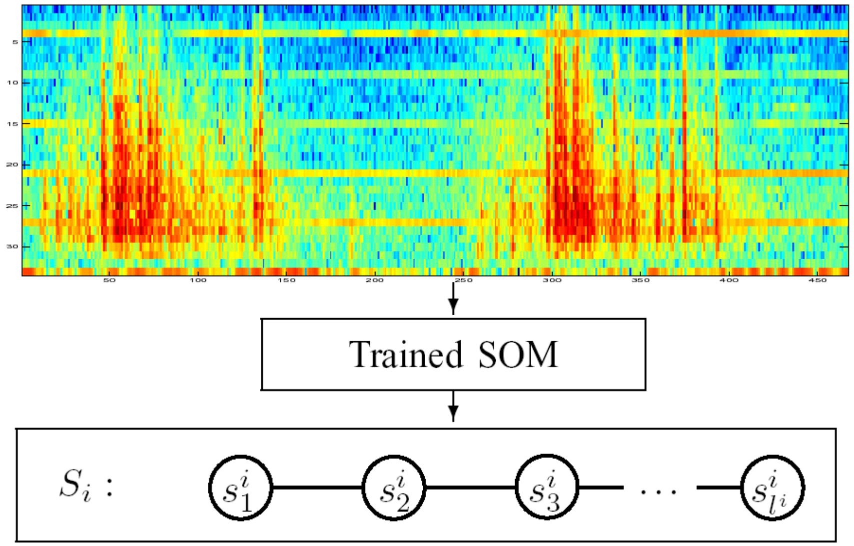
Feature Extraction

Discretization of joint-torque records using a trained SOM

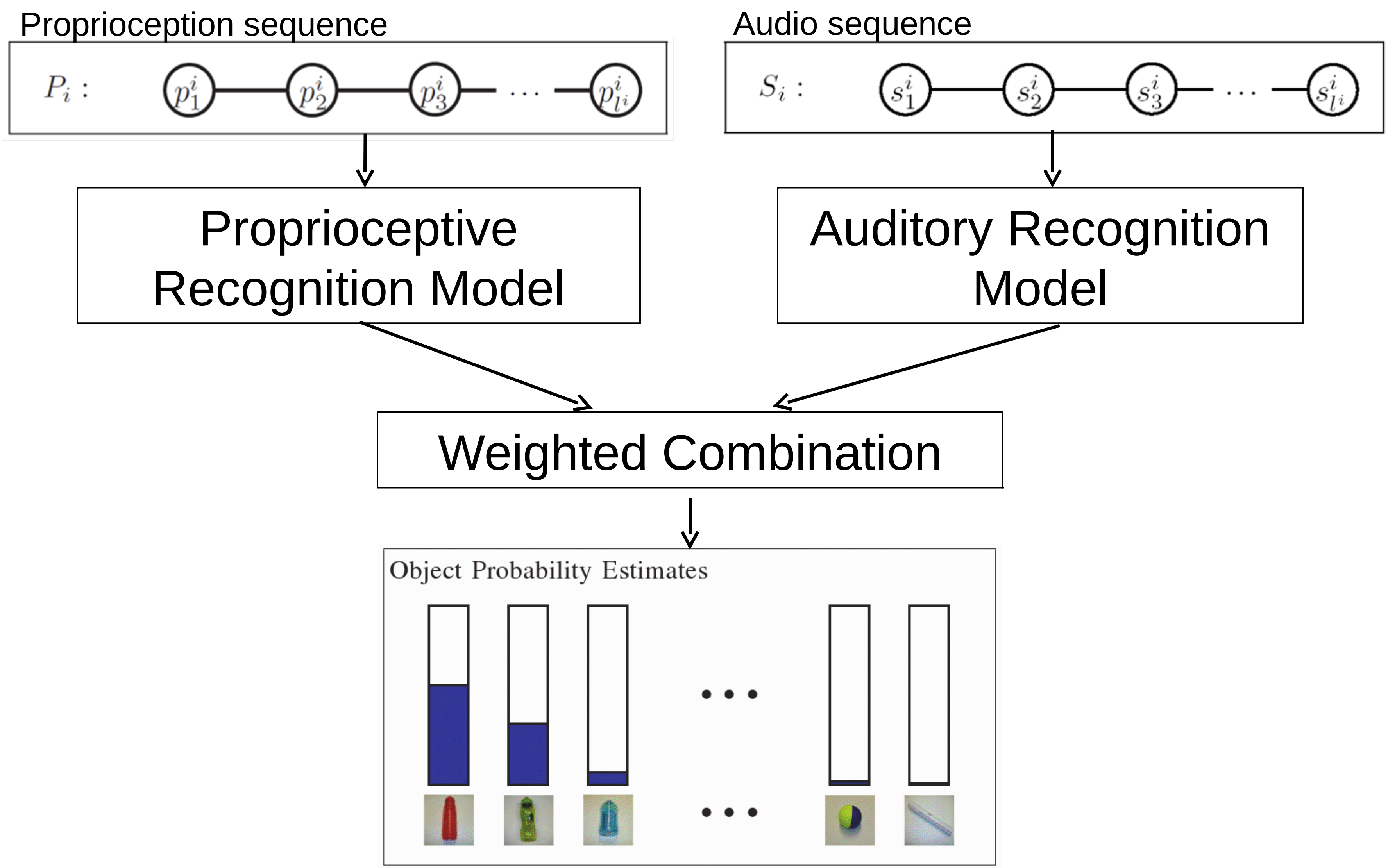


$P_i = p_1^i p_2^i \dots p_{l^{P_i}}^i$ is the sequence of activated SOM nodes over the duration of the interaction

Discretization of the DFT of a sound using a trained SOM



$S_i = s_1^i s_2^i \dots s_{l^{S_i}}^i$ is the sequence of activated SOM nodes over the duration of the sound

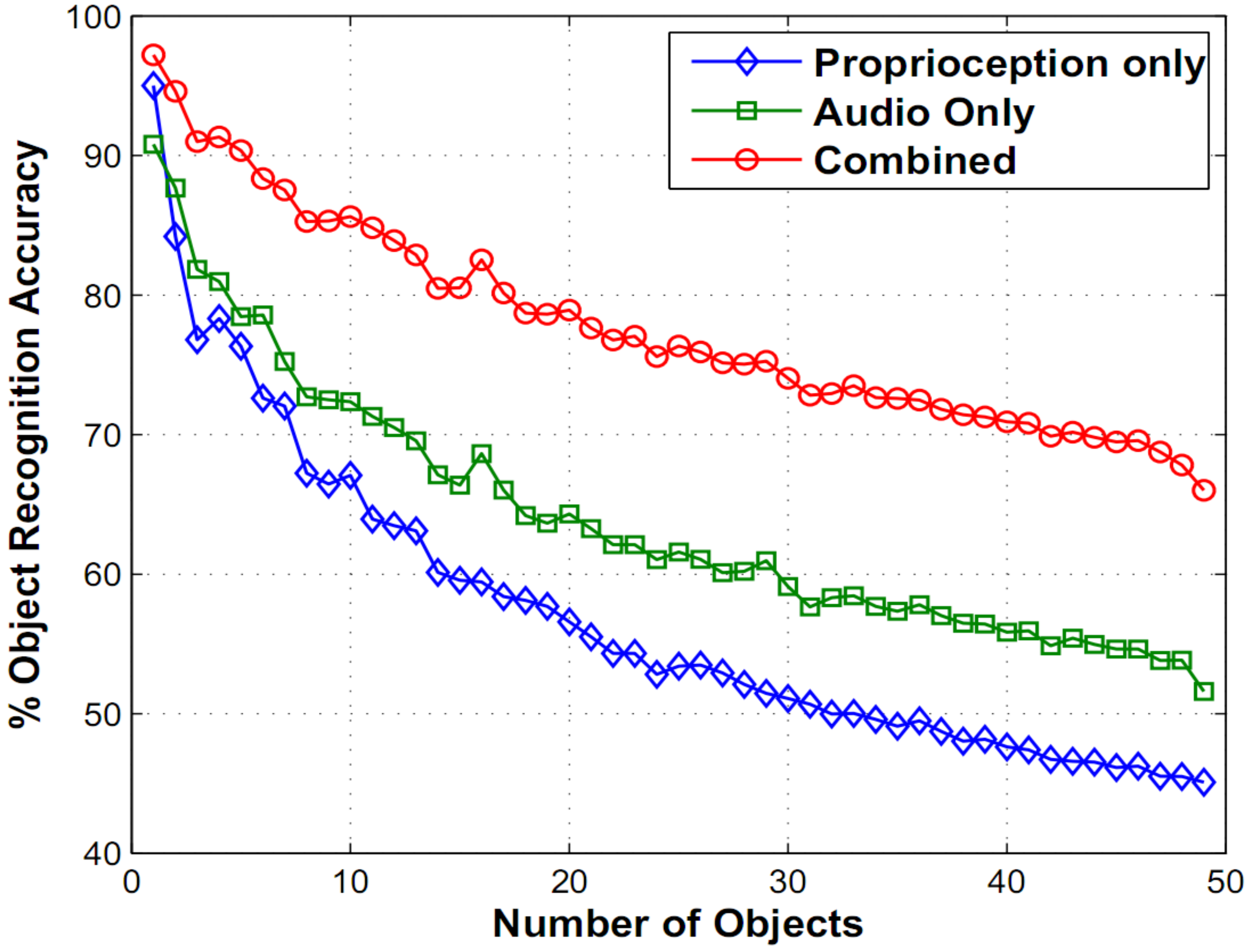


Proprioception sequence
P_i : p_1^i p_2^i p_3^i ... $p_{l^i}^i$
Audio sequence
S_i : s_1^i s_2^i s_3^i ... $s_{l^i}^i$
Proprioceptive Recognition Model
Auditory Recognition Model
Weighted Combination
Object Probability Estimates

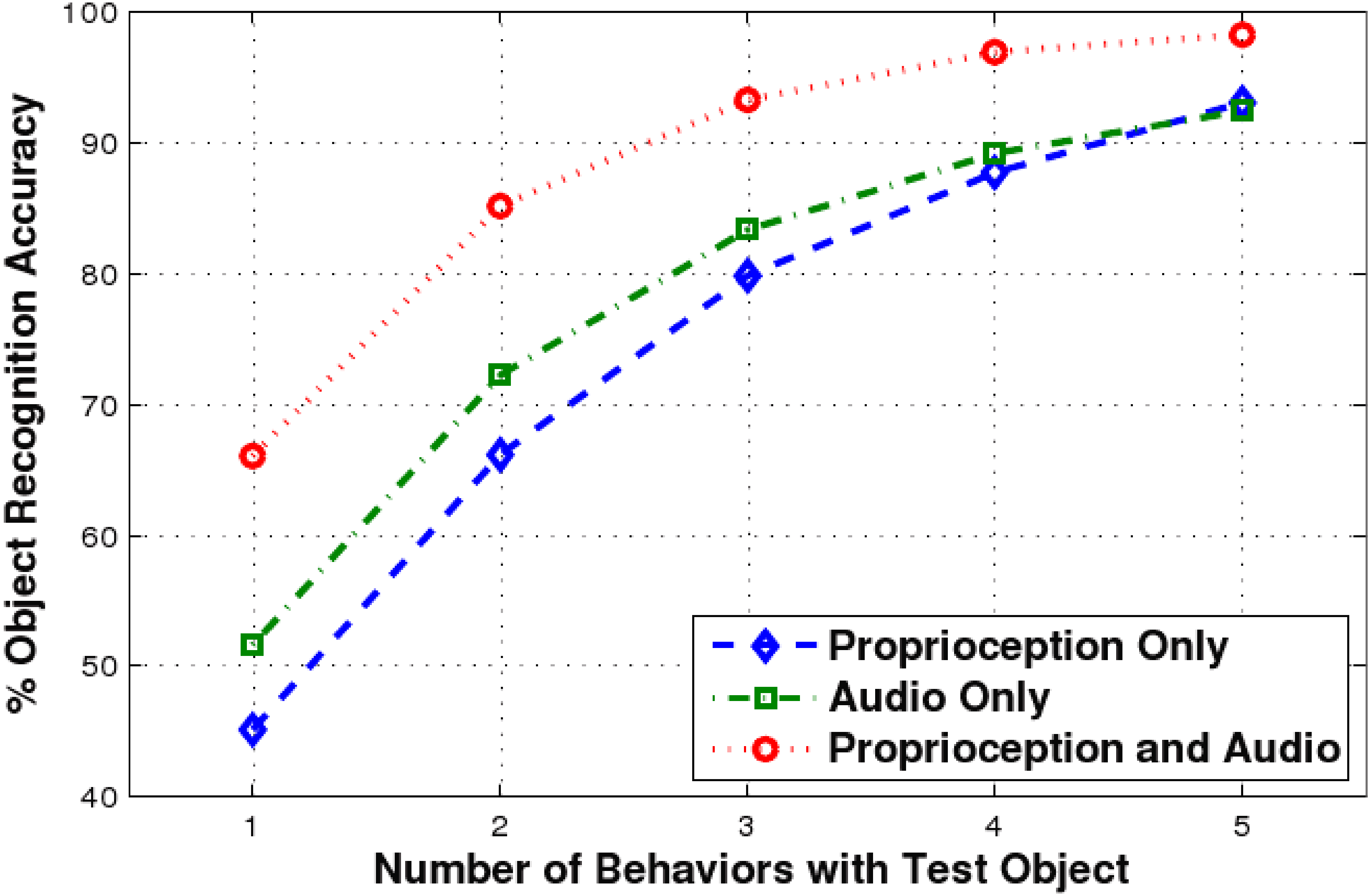
Object Recognition accuracy using k-NN model

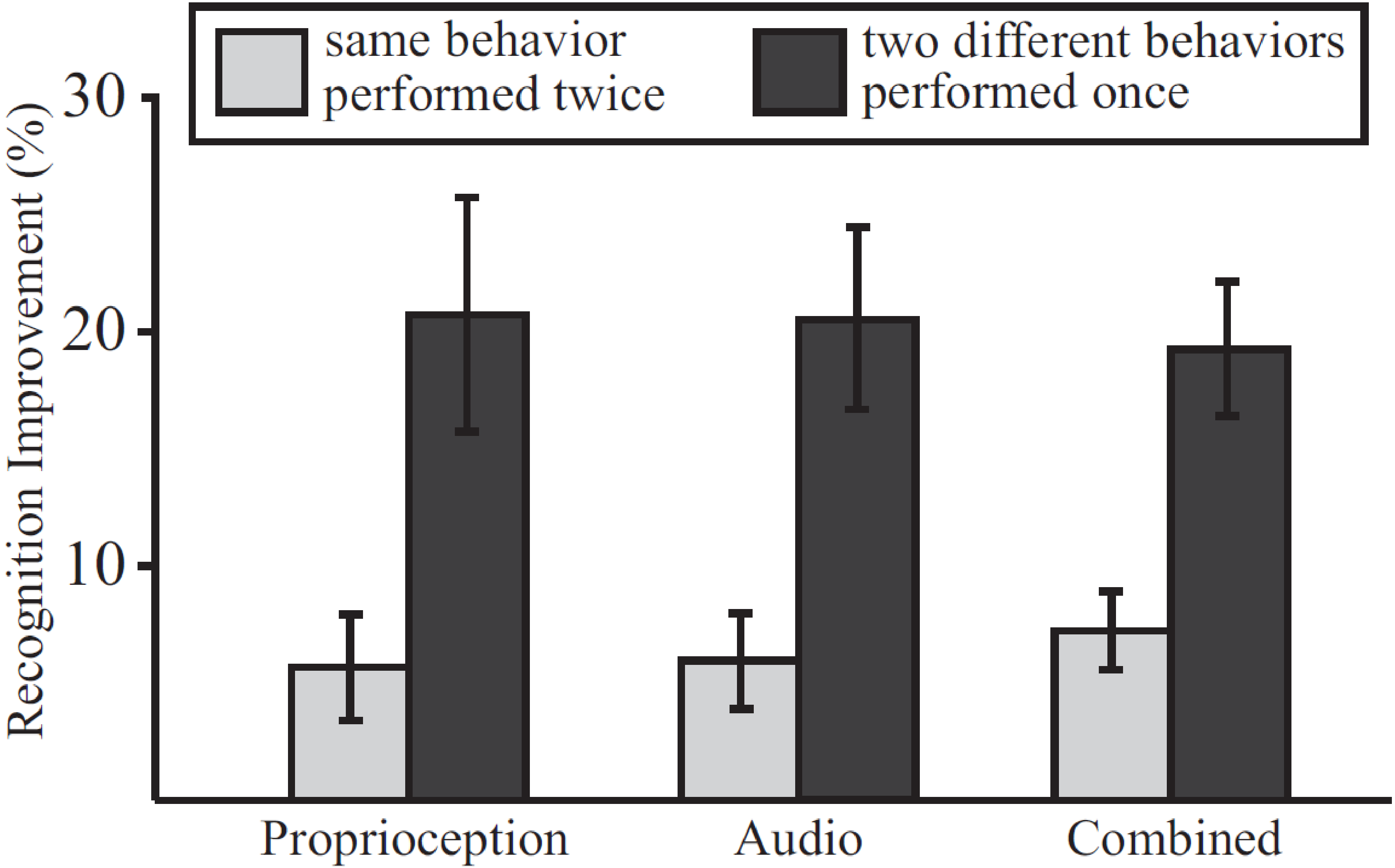
Behavior	Audio	Proprioception	Combined
Lift	17.4 %	64.8 %	66.4 %
Shake	27.0 %	15.2 %	29.4 %
Drop	76.4 %	45.6 %	80.8 %
Crush	73.4 %	84.6 %	88.6 %
Push	63.8 %	15.4 %	65.0 %
Average	51.6 %	45.1 %	66.0 %

Accuracy vs. Number of Objects



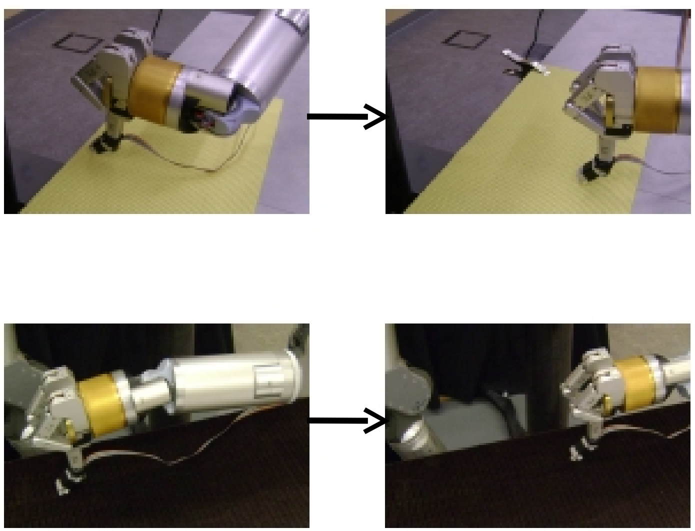
Accuracy vs. Number of Behaviors

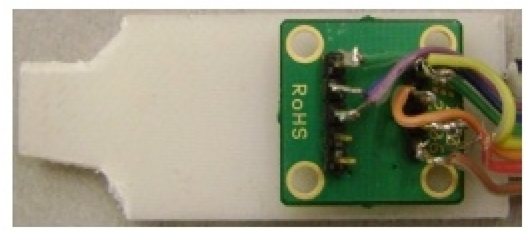



same behavior performed twice
two different behaviors performed once
Recognition Improvement (%)
30
20
10
Proprioception
Audio
Combined

Results with a Second Dataset

- Tactile Surface Recognition:
  - 5 scratching behaviors
  - 2 modalities: *vibrotactile* and *proprioceptive*

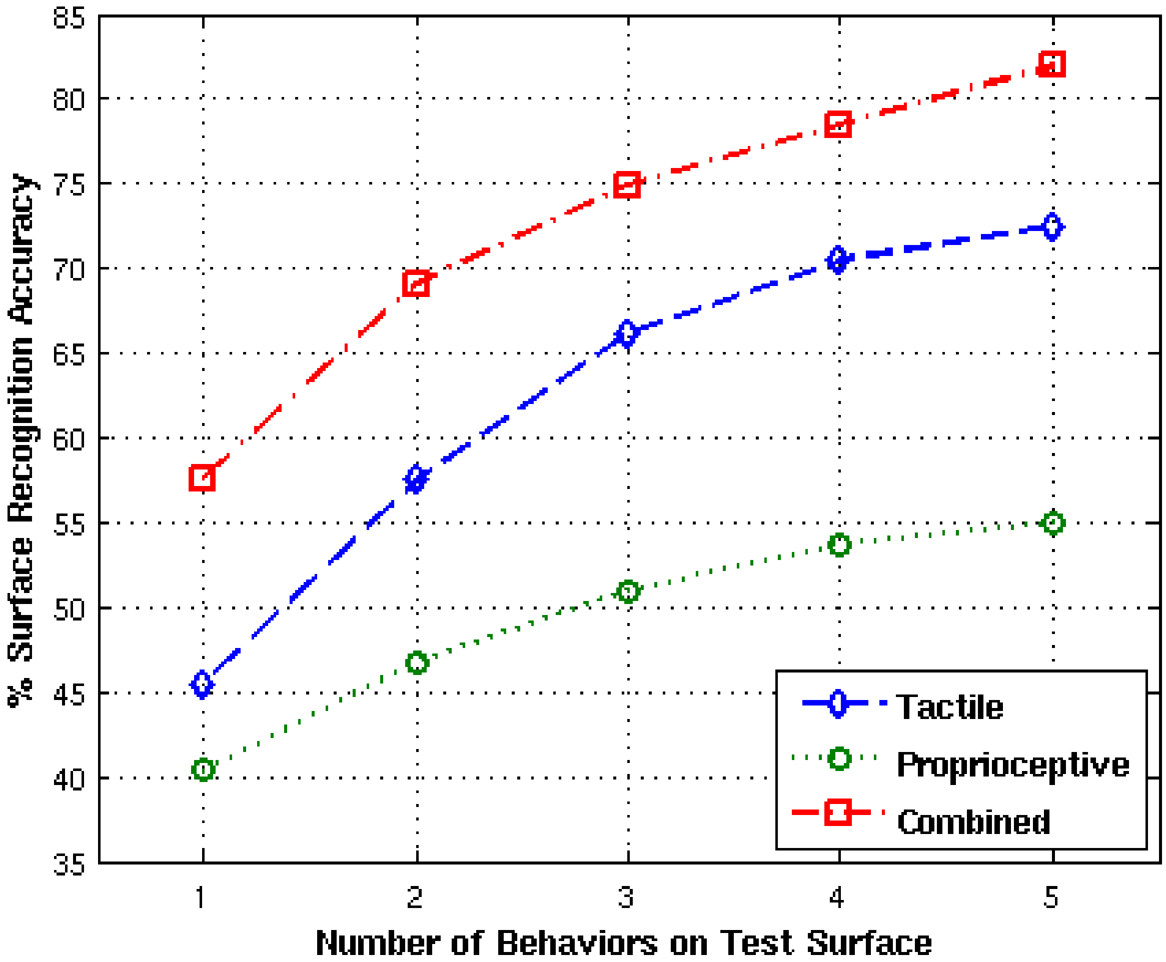




Artificial Finger Tip

Sinapov et al. "Vibrotactile Recognition and Categorization of Surfaces by a Humanoid Robot"
IEEE Transactions on Robotics, Vol. 27, No. 3, pp. 488-497, June 2011

Surface Recognition Results



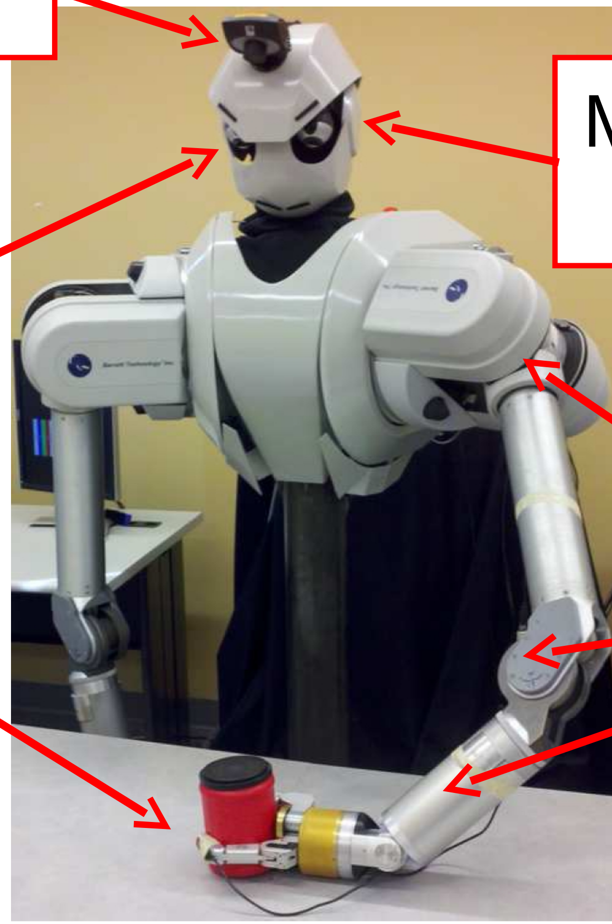
Chance accuracy = 1/20 = 5 %

Scaling up: more sensory modalities, objects and behaviors

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

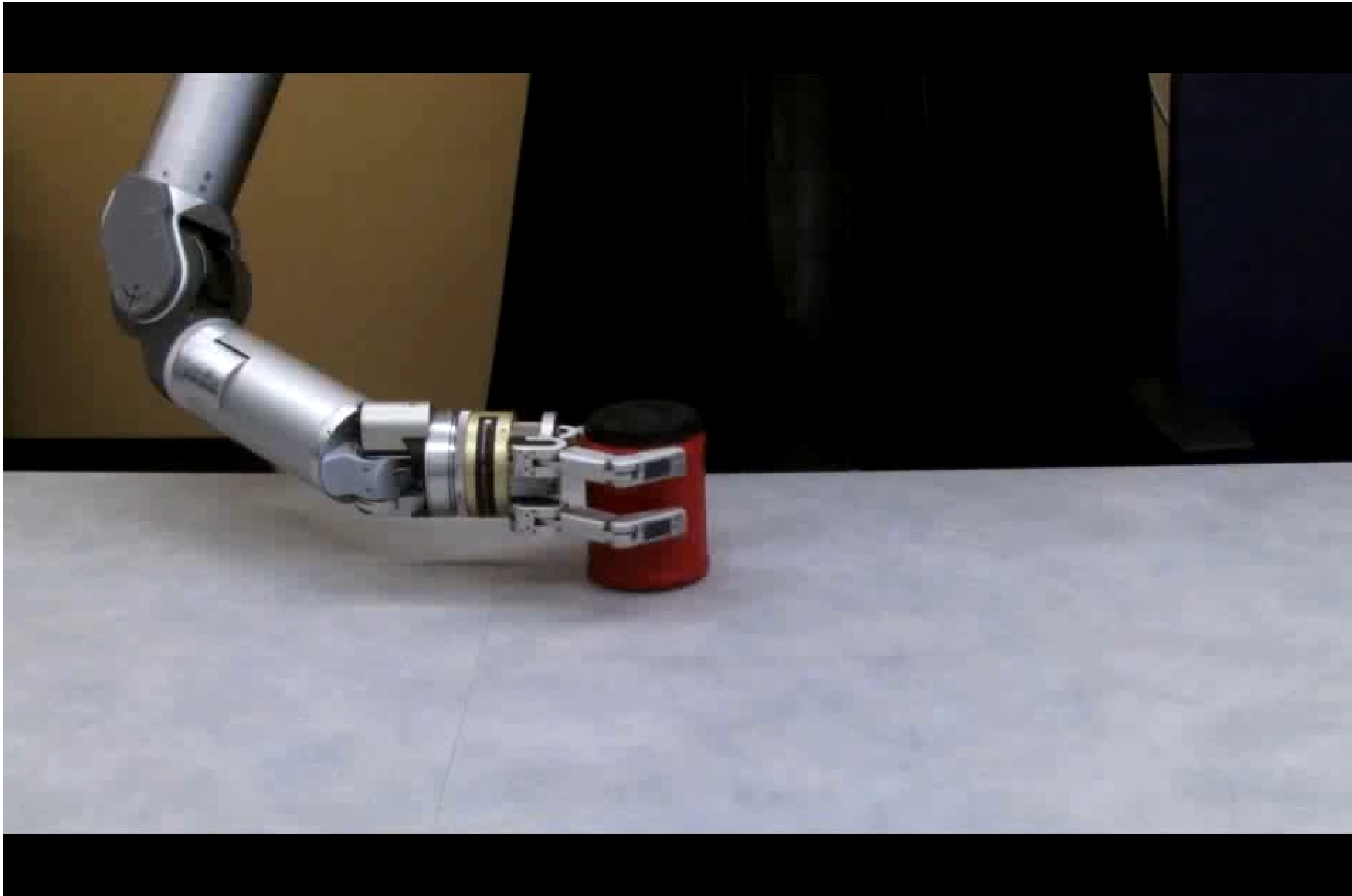

drop


tap


poke

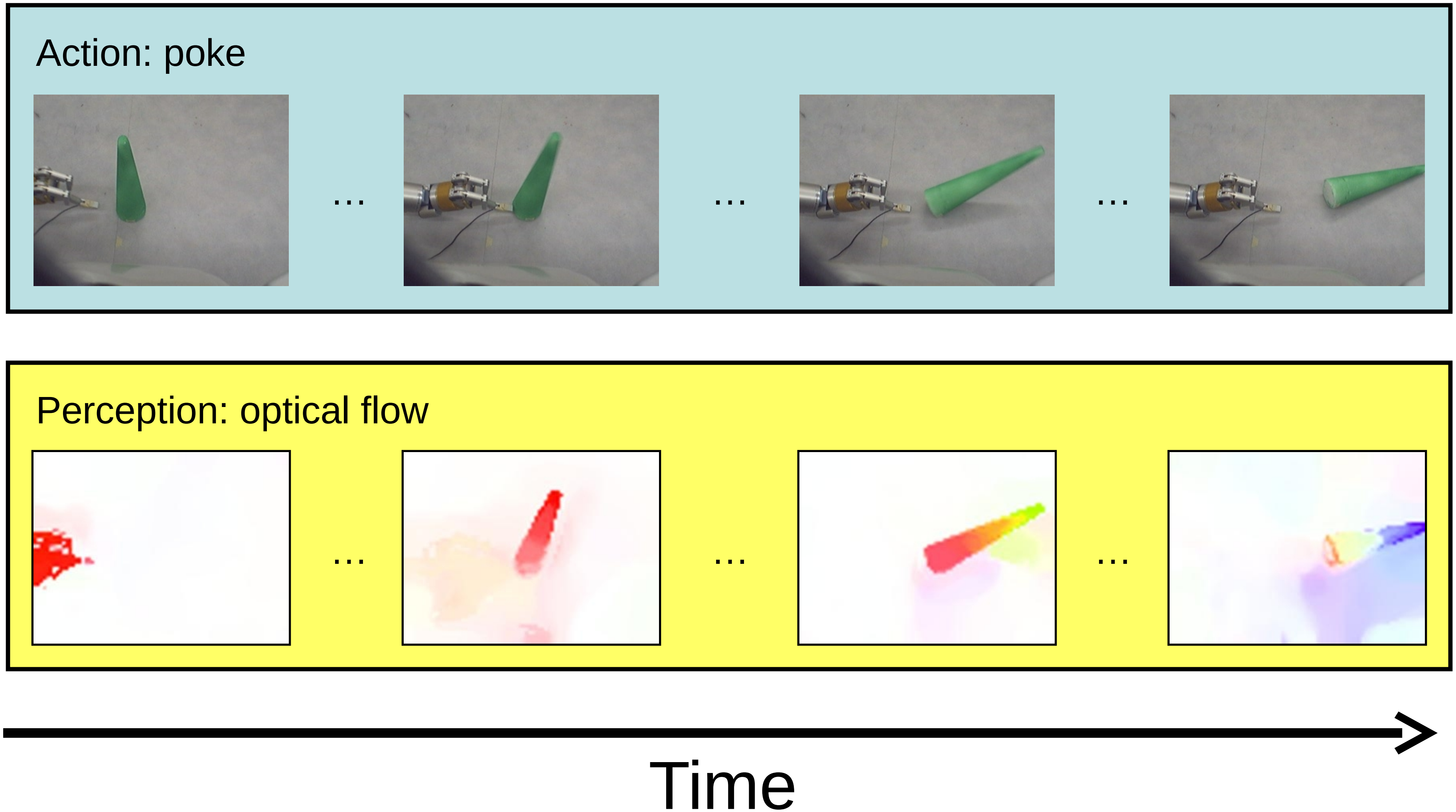

push


press



Object Exploration Video #2

Coupling Action and Perception



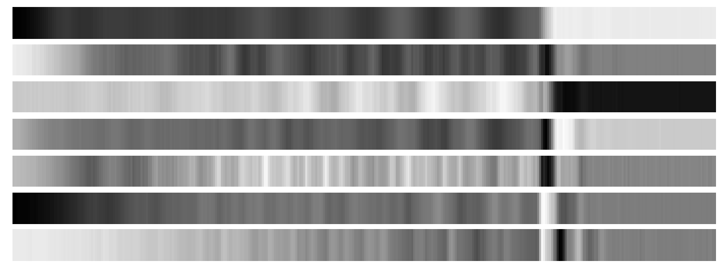
Sensorimotor Contexts

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

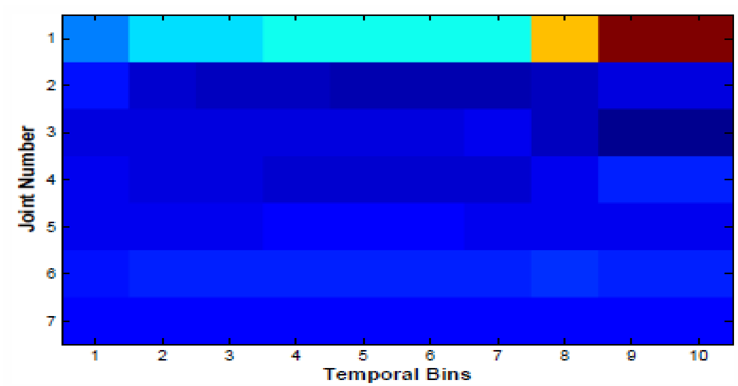
Sensorimotor Contexts

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

Feature Extraction: Proprioception

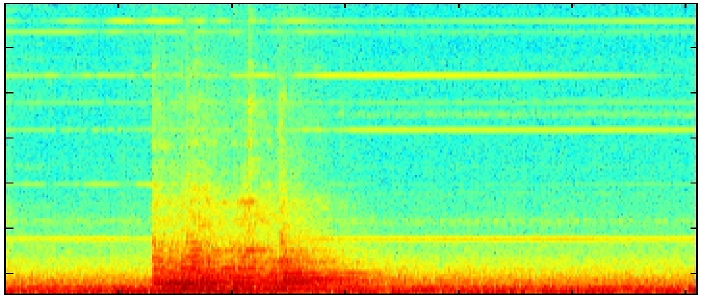


Joint-Torque values for all 7 Joints

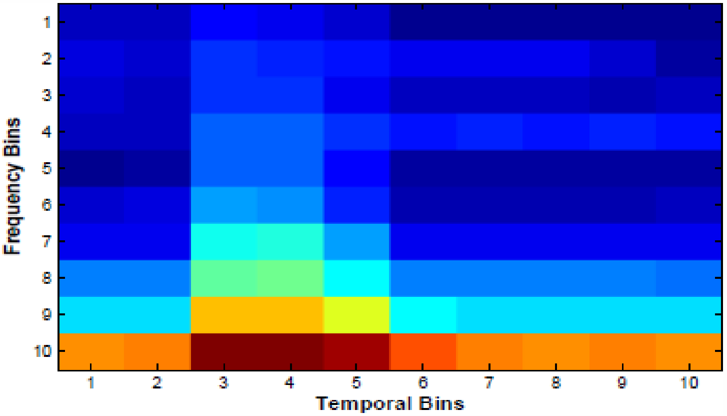


Joint-Torque Features

Feature Extraction: Audio

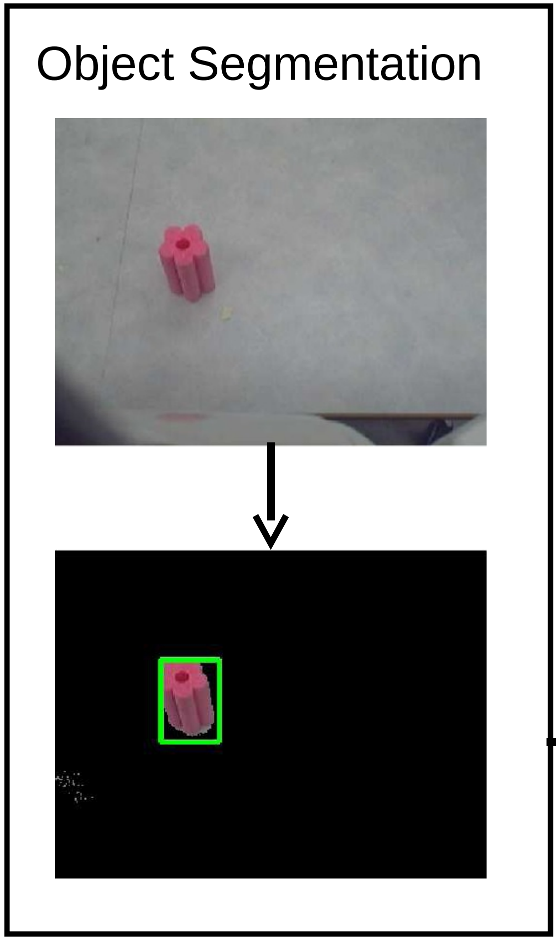


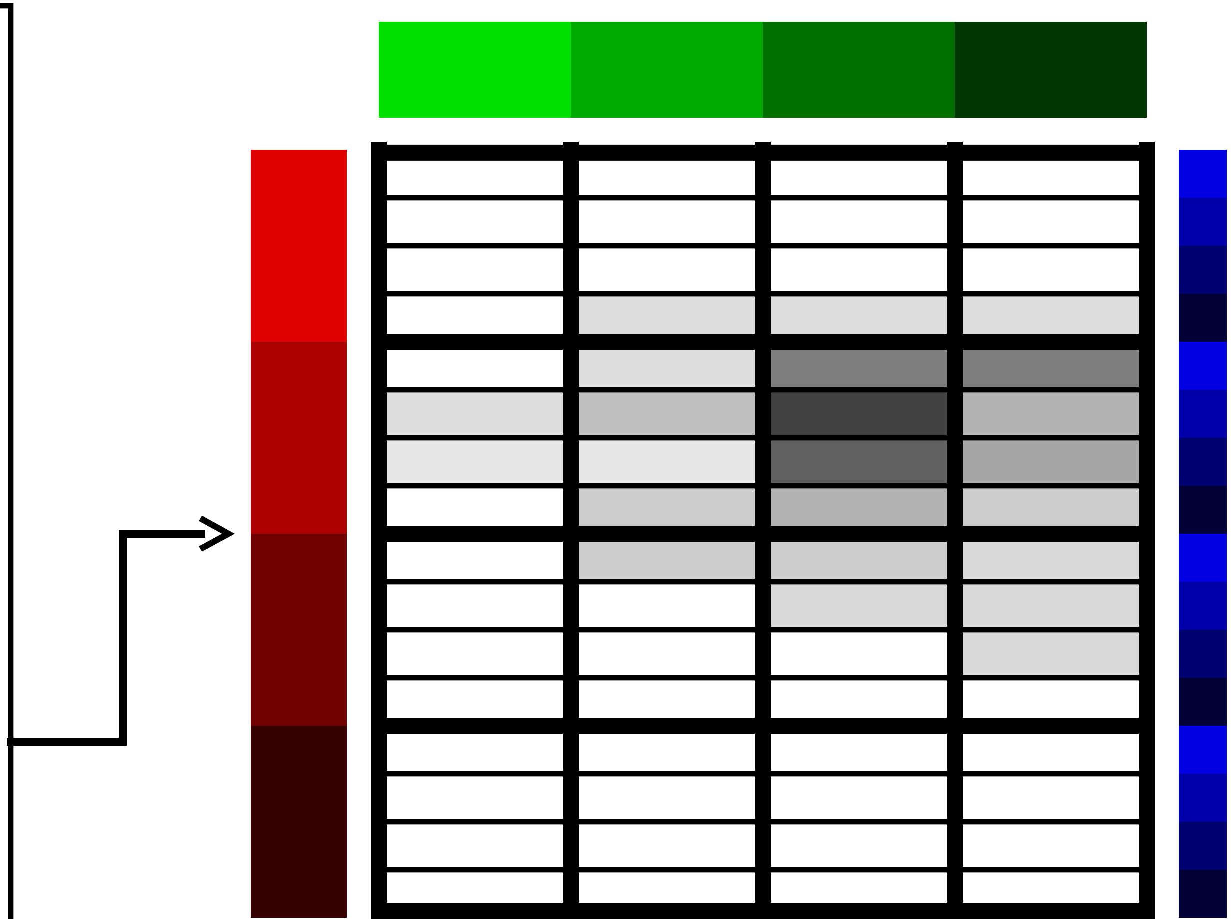
audio spectrogram



Spectro-temporal Features

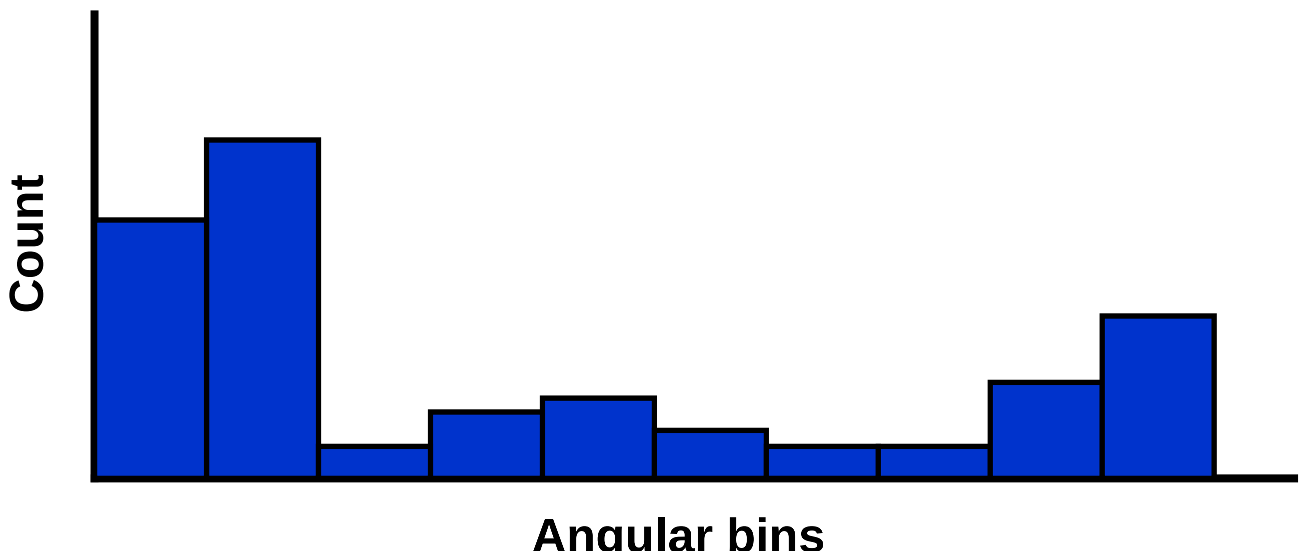
Feature Extraction: Color



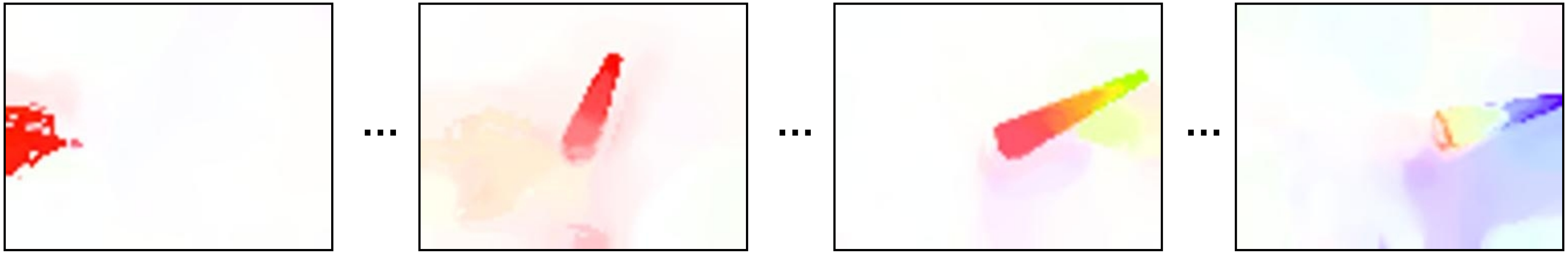


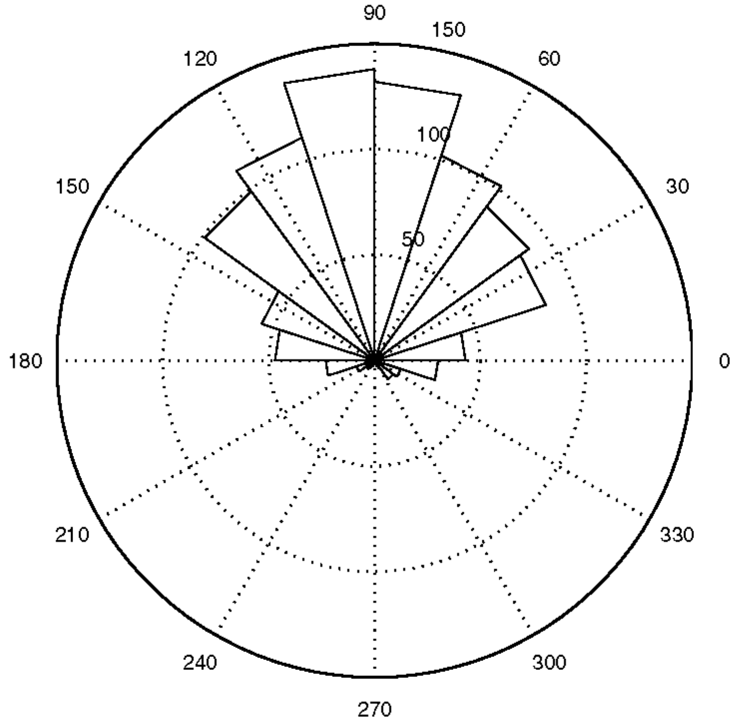
Color Histogram (4 x 4 x 4 = 64 bins)

Feature Extraction: Optical Flow

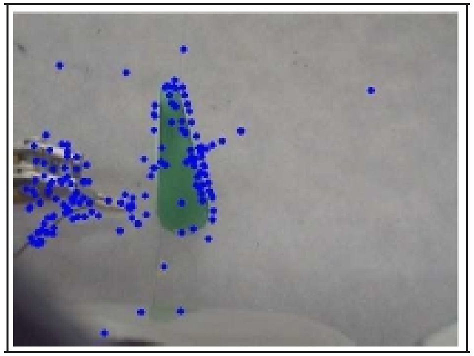
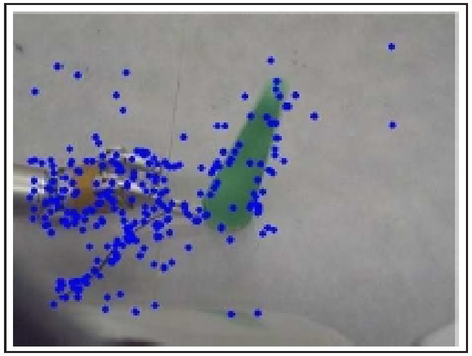
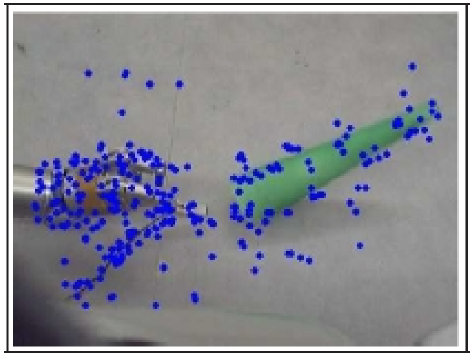


Feature Extraction: Optical Flow

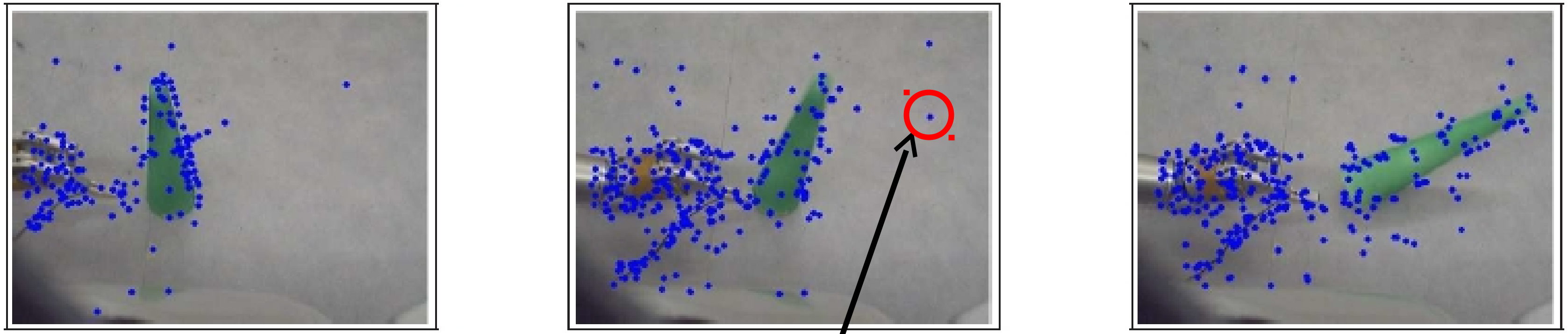




Feature Extraction: SURF

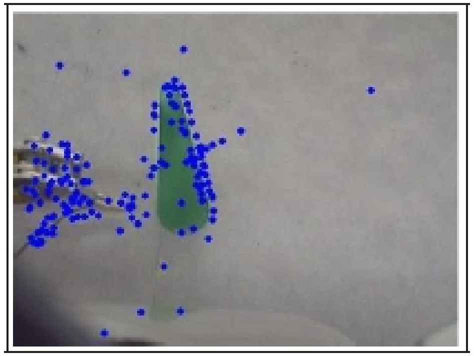




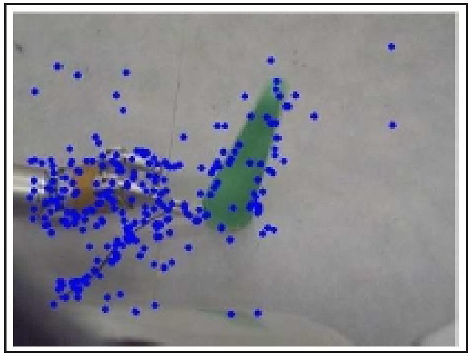
Feature Extraction: SURF

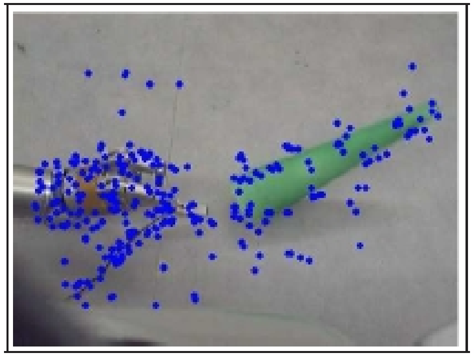


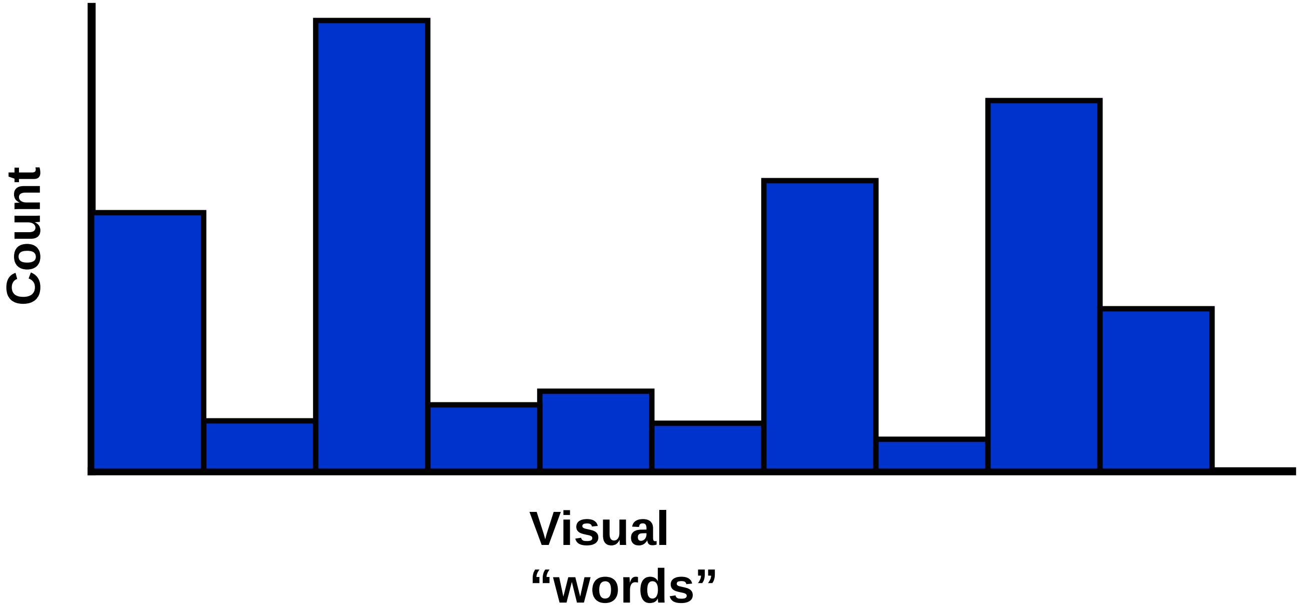
Each interest point is described by a 128-dimensional vector

Feature Extraction: SURF









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

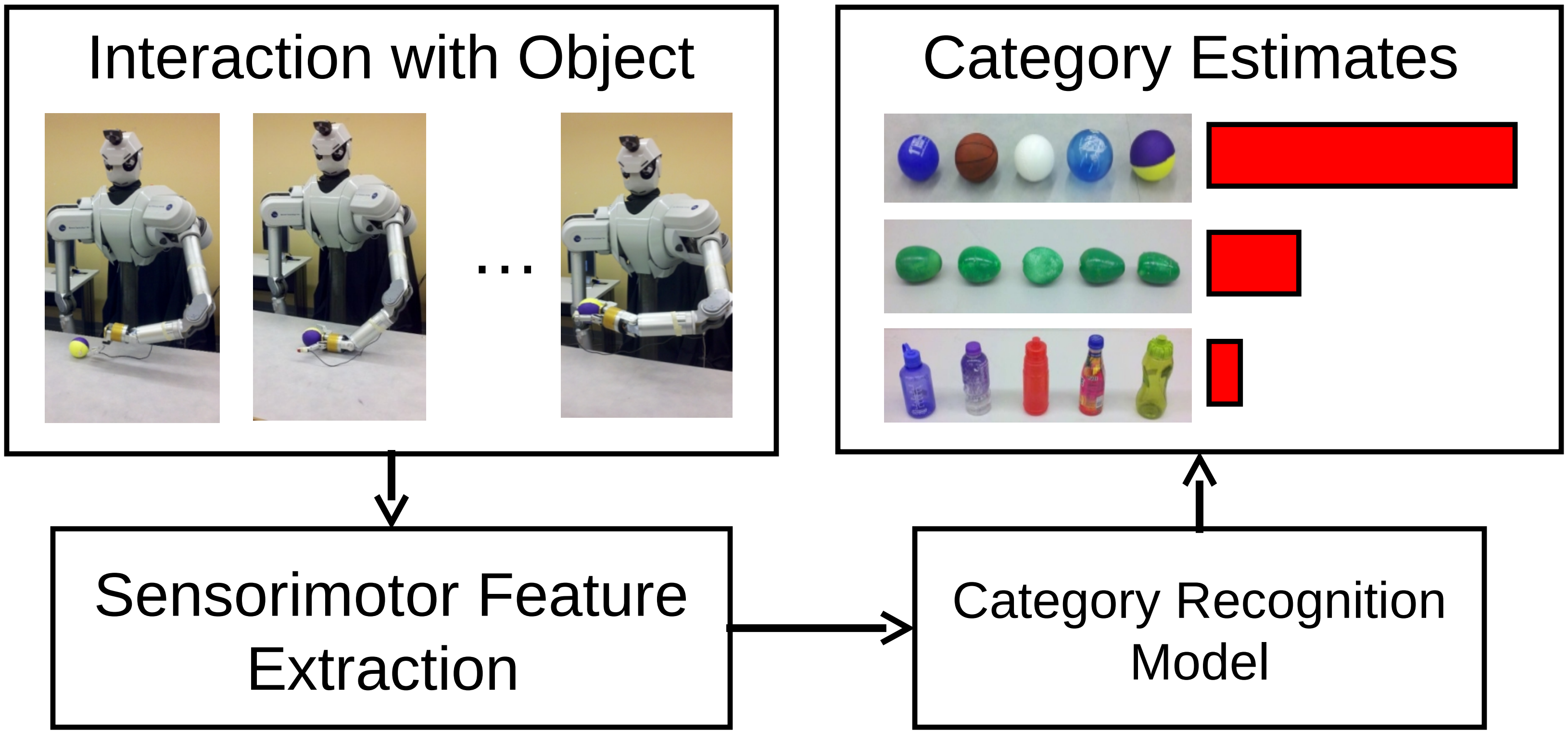
	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look				✓		✓
grasp	✓	✓	✓		✓	✓
lift	✓	✓			✓	✓
hold	✓	✓			✓	✓
shake	✓	✓			✓	✓
drop	✓	✓			✓	✓
tap	✓	✓			✓	✓
poke	✓	✓			✓	✓
push	✓	✓			✓	✓
press	✓	✓			✓	✓

Data From a Single Exploratory Trial

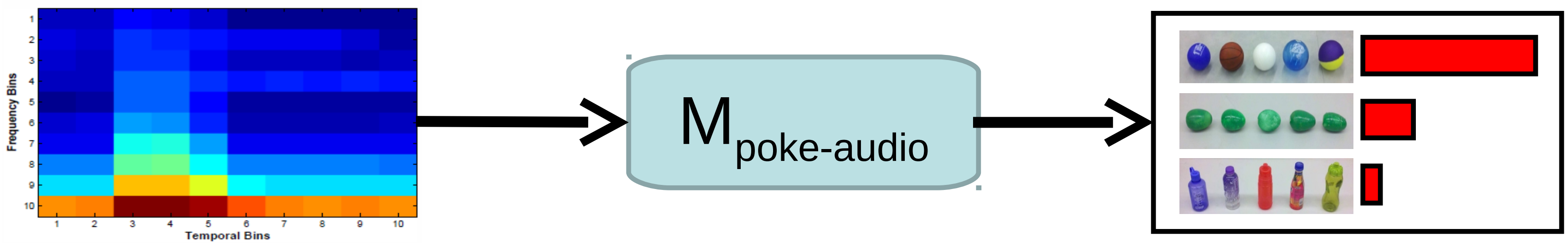
	audio (DFT)	proprioception (joint torques)	proprioception (finger pos.)	Color	Optical flow	SURF
look				✓		✓
grasp	✓	✓	✓		✓	✓
lift	✓	✓			✓	✓
hold	✓	✓			✓	✓
shake	✓	✓			✓	✓
drop	✓	✓			✓	✓
tap	✓	✓			✓	✓
poke	✓	✓			✓	✓
push	✓	✓			✓	✓
press	✓	✓			✓	✓

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

Recognition model for poke-audio context

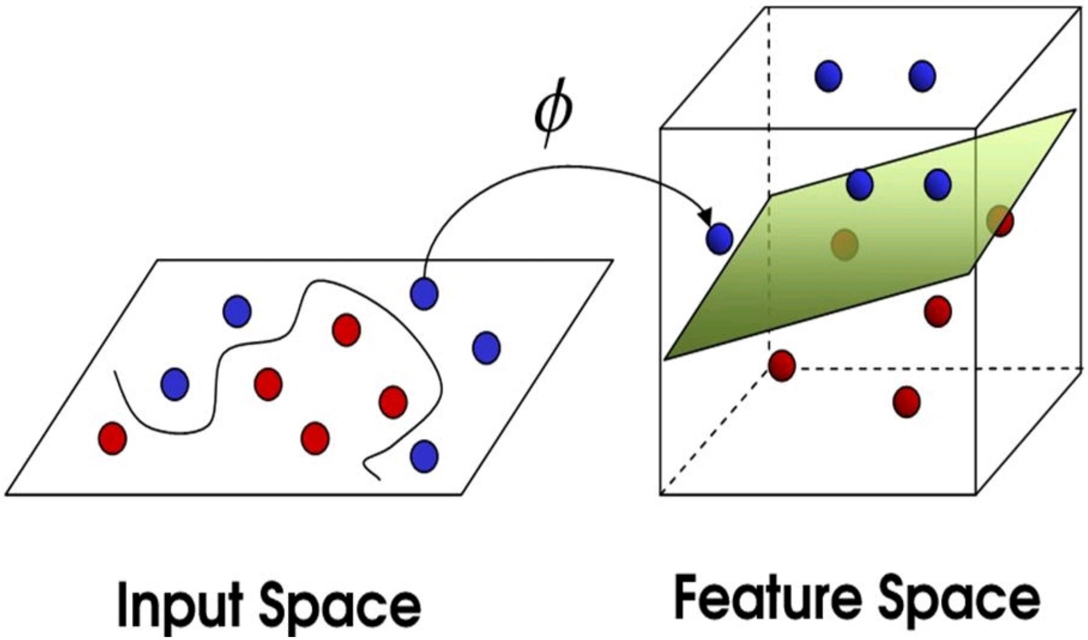
Distribution over category labels

Context-specific Category Recognition

- The models were implemented by two machine learning algorithms:
  - K-Nearest Neighbors (k = 3)
  - Support Vector Machine

Support Vector Machine

- <u>Support Vector Machine</u>: a discriminative learning algorithm

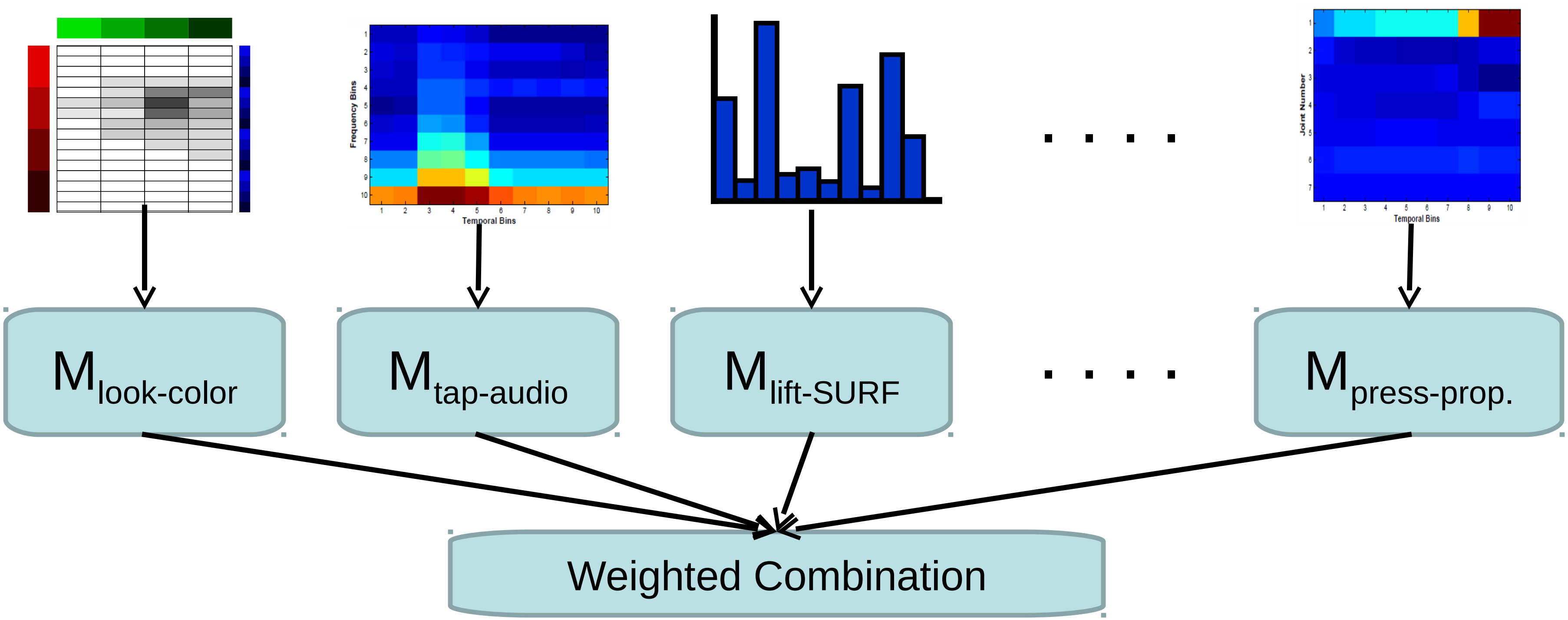


[http://www.imtech.res.in/raghava/rbpred/svm.jpg]

1. Finds maximum margin hyperplane that separates two classes

2. Uses Kernel function to map data points into a feature space in which such a hyperplane exists

Combining Model Outputs

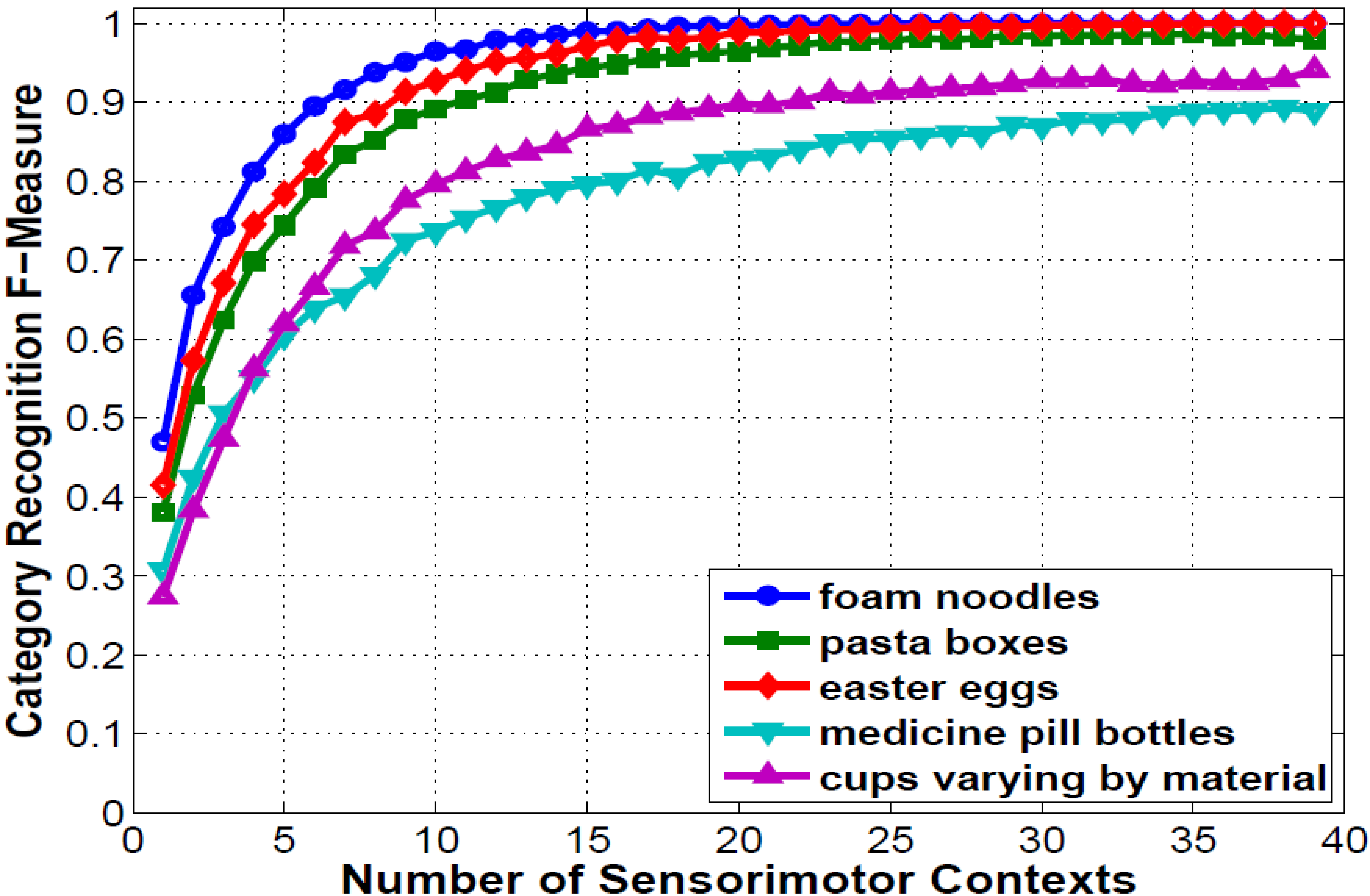


Model Evaluation: 5 fold Cross-Validation

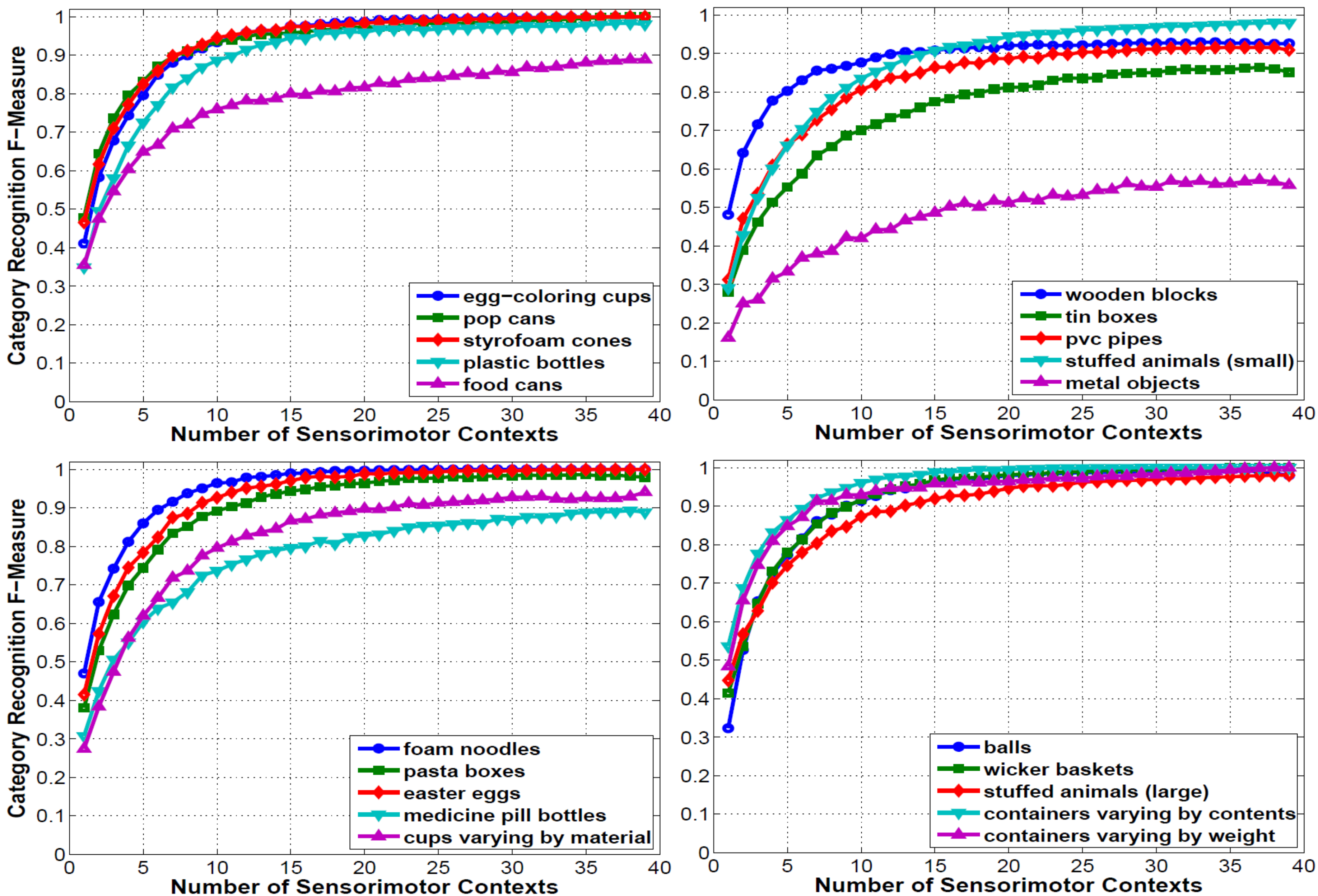
Train Set **Test Set**

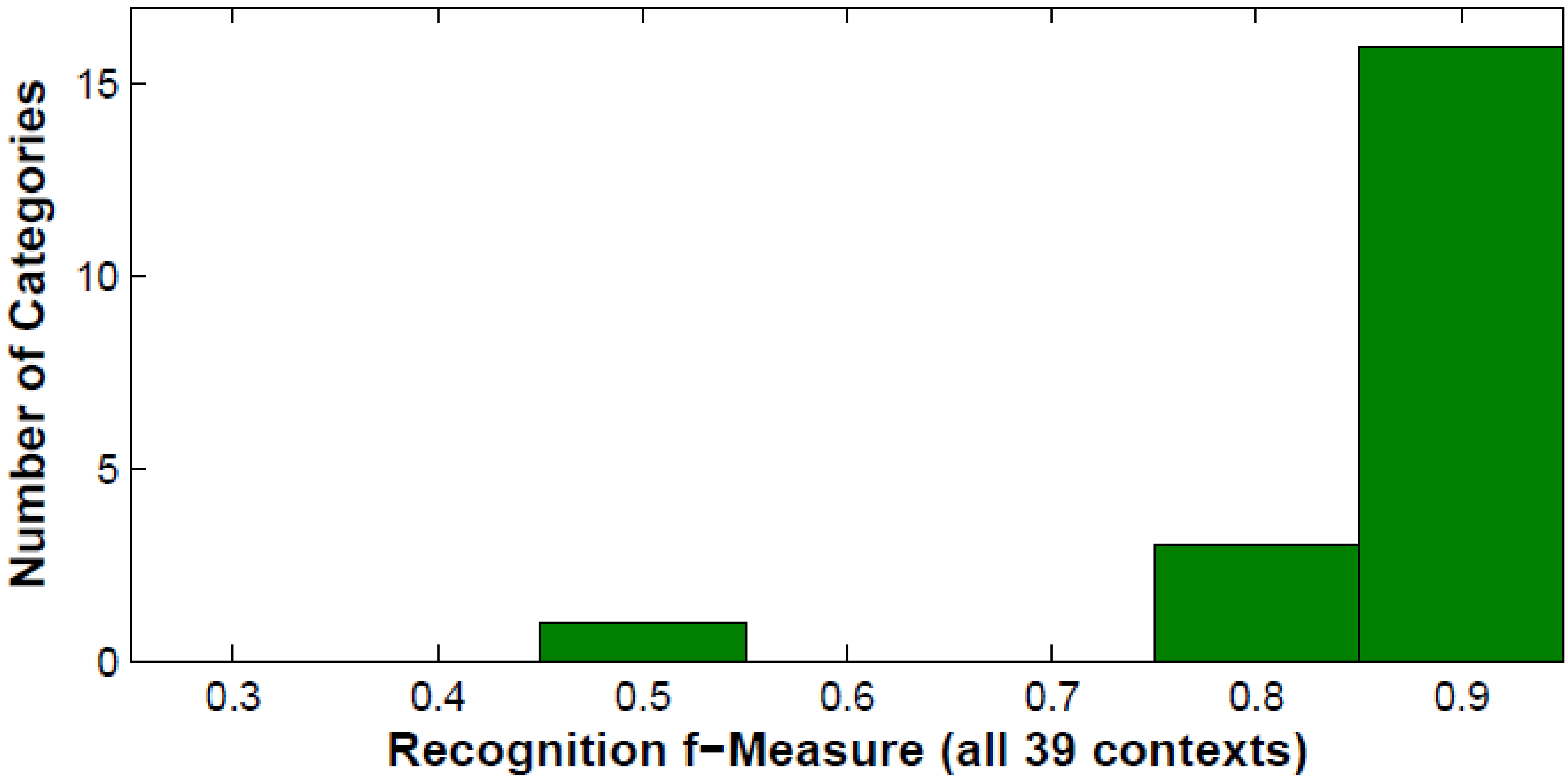
Recognition Rates (%) with SVM

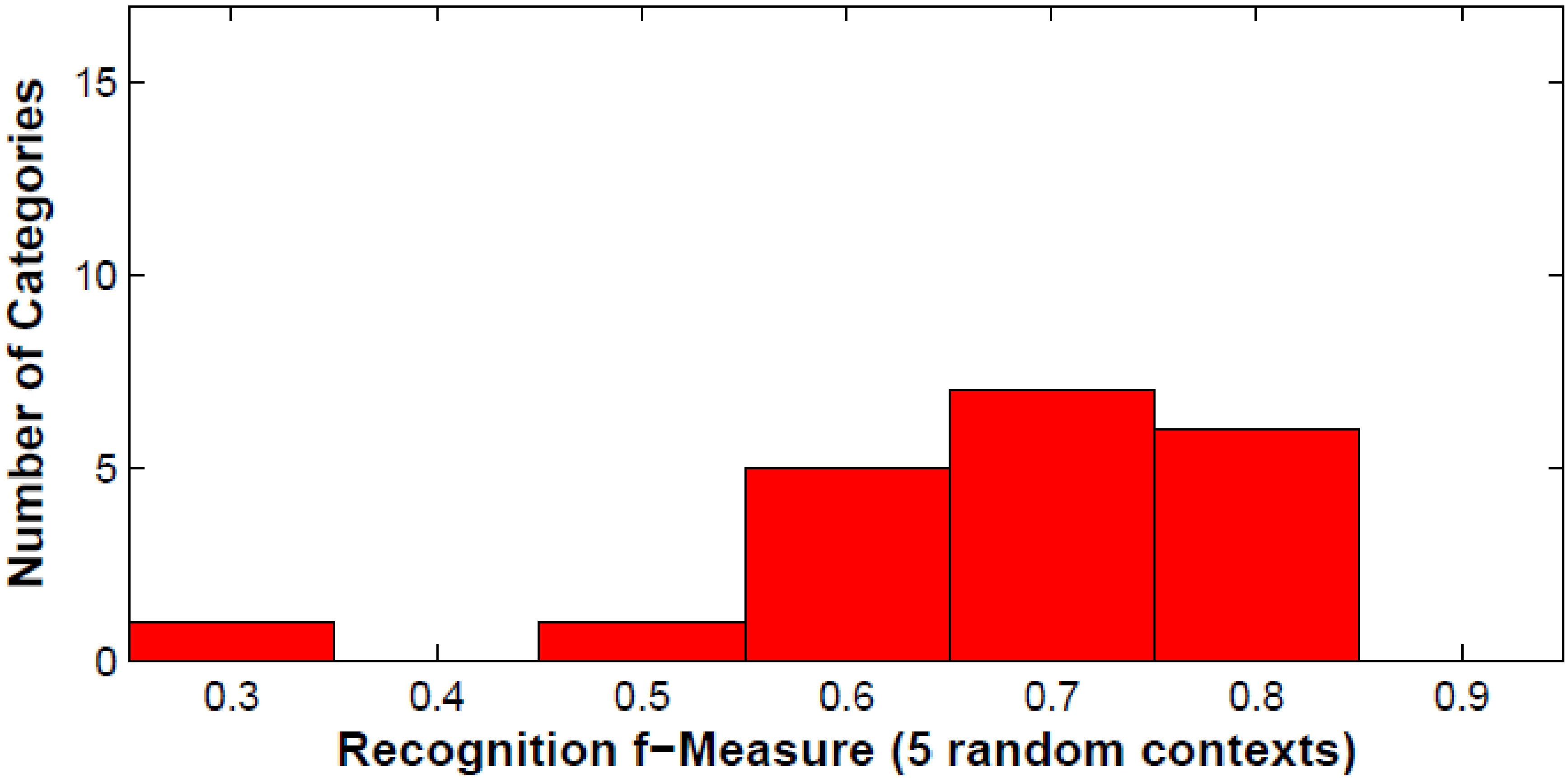
	Audio	Proprioception	Color	Optical Flow	SURF	All
look			58.8		58.9	67.7
grasp	45.7	38.7		12.2	57.1	65.2
lift	48.1	63.7		5.0	65.9	79.0
hold	30.2	43.9		5.0	58.1	67.0
shake	49.3	57.7		32.8	75.6	76.8
drop	47.9	34.9		17.2	57.9	71.0
tap	63.3	50.7		26.0	77.3	82.4
push	72.8	69.6		26.4	76.8	88.8
poke	65.9	63.9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



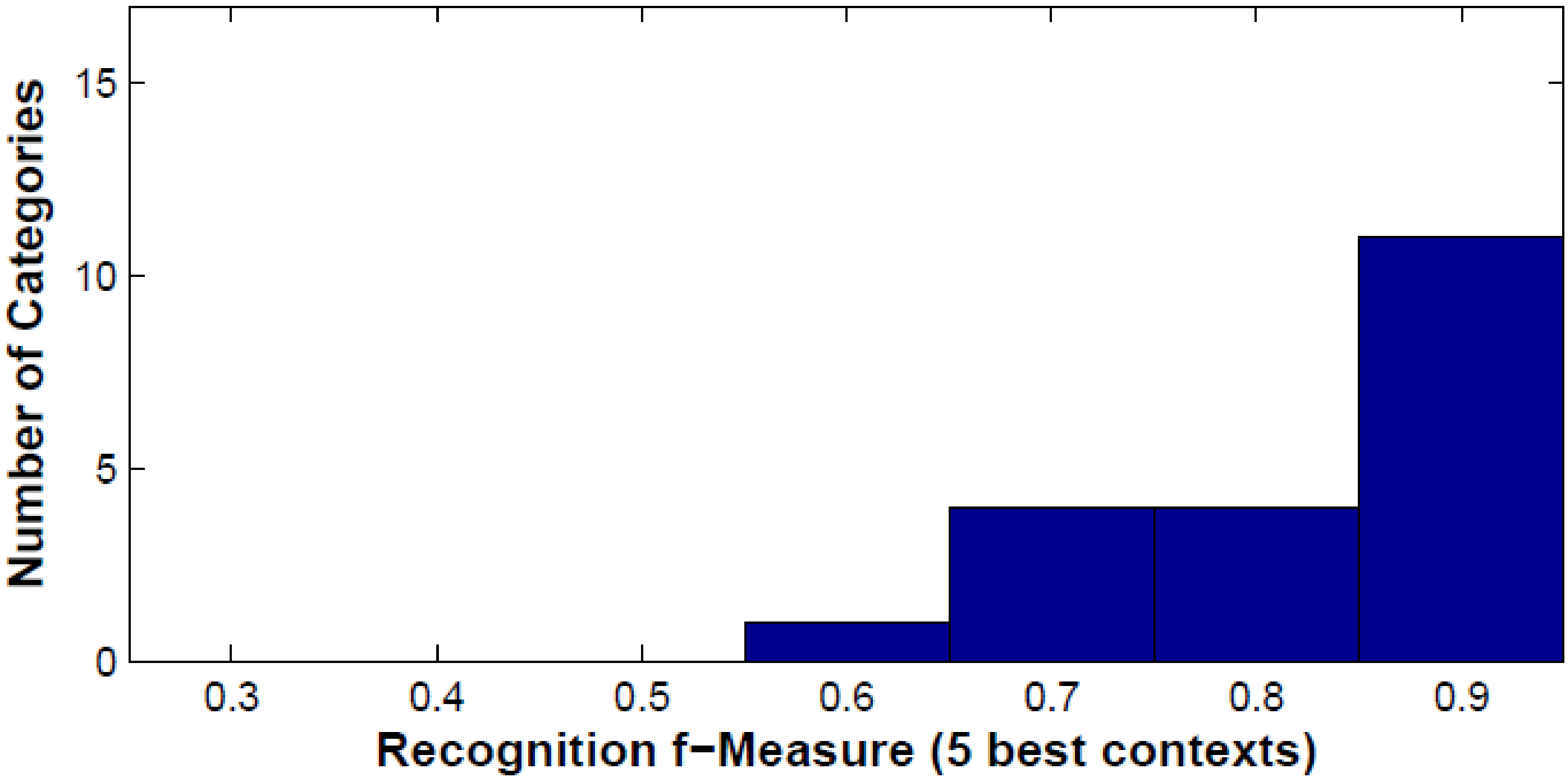
Category Recognition F−Measure
Number of Sensorimotor Contexts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
foam noodles
pasta boxes
easter eggs
medicine pill bottles
cups varying by material


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

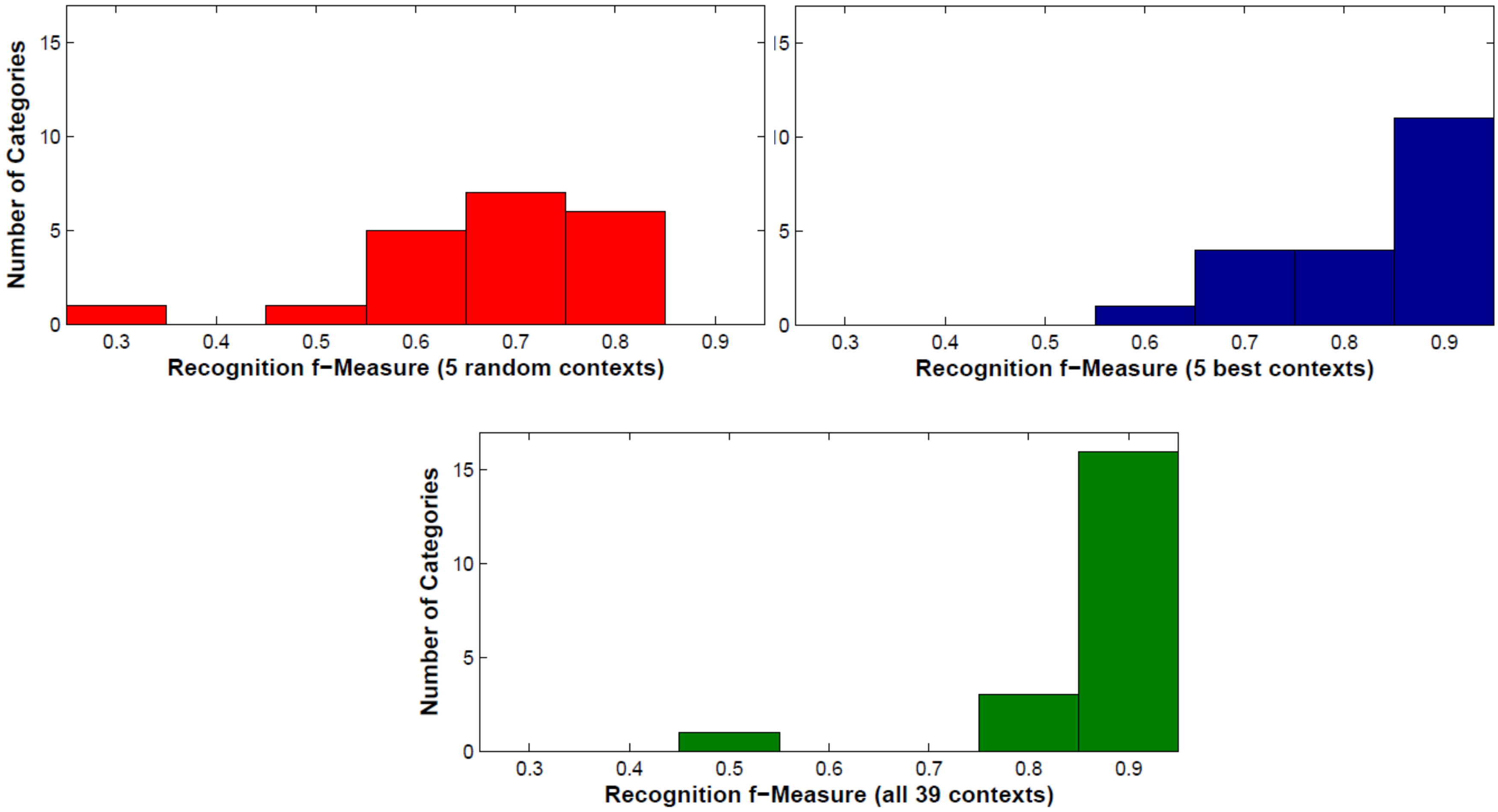

Number of Categories
0
5
10
15
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Recognition f−Measure (all 39 contexts)



Number of Categories
0
5
10
15
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Recognition f−Measure (5 random contexts)


Number of Categories
15
10
5
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Recognition f−Measure (5 best contexts)

Distribution of rates over categories



Can behaviors be selected actively to minimize exploration time?

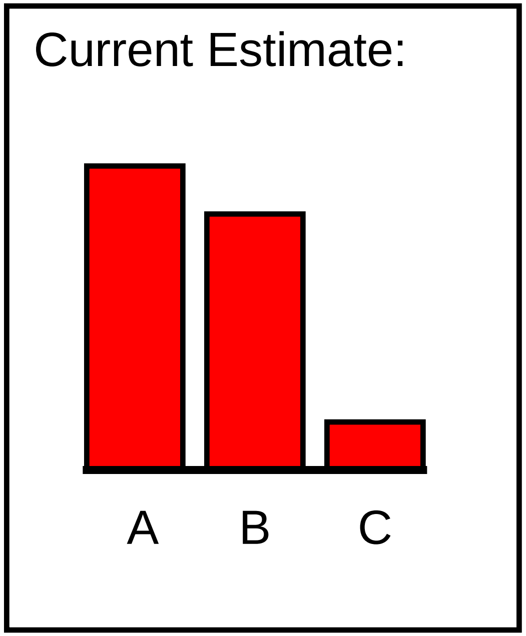
Active Behavior Selection

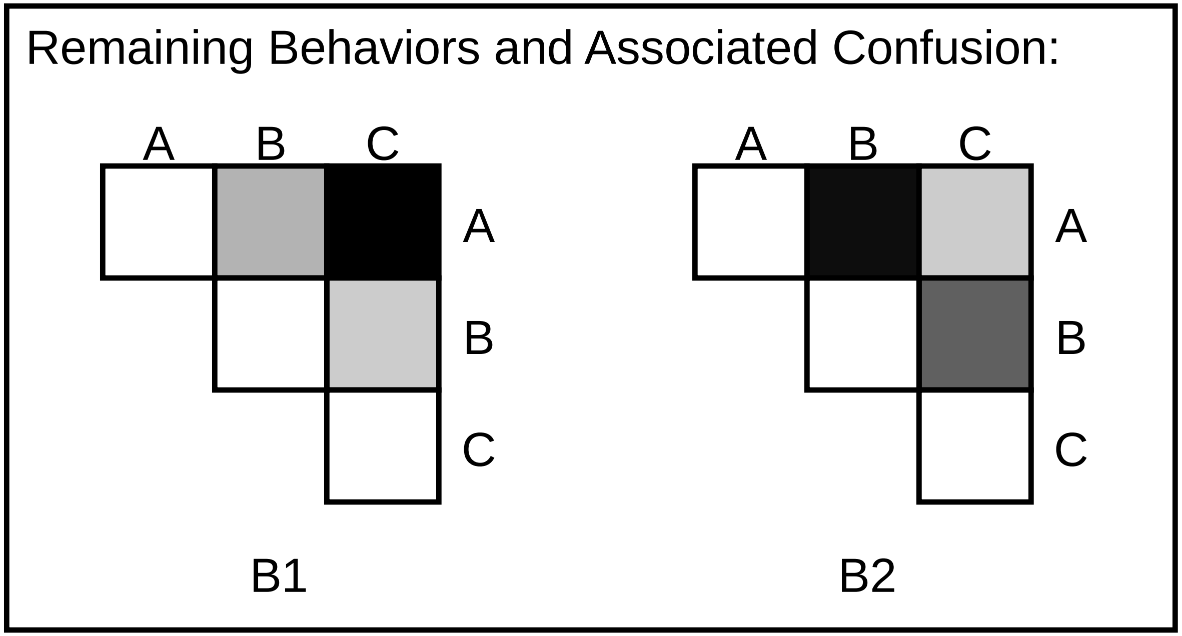
- For each behavior $b \in \mathcal{B}$, estimate C_{ij}^{b} such that

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

- 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

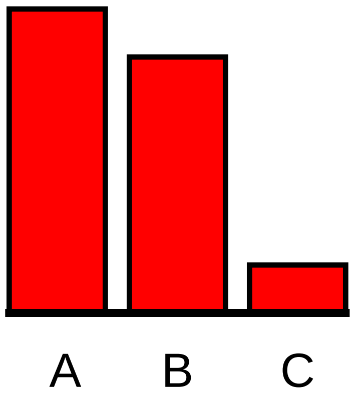
Example with 3 Categories and 2 Behaviors



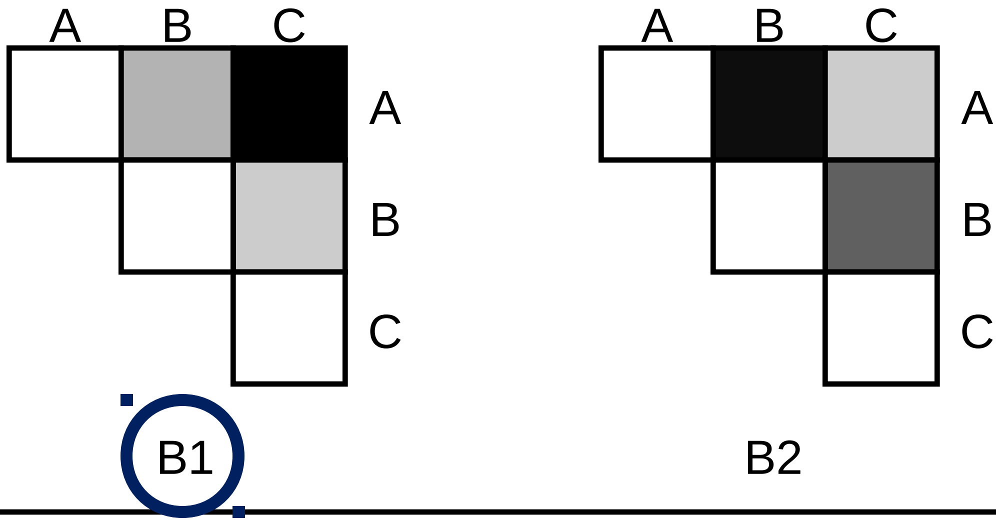


Active Behavior Selection: Example

Current Estimate:



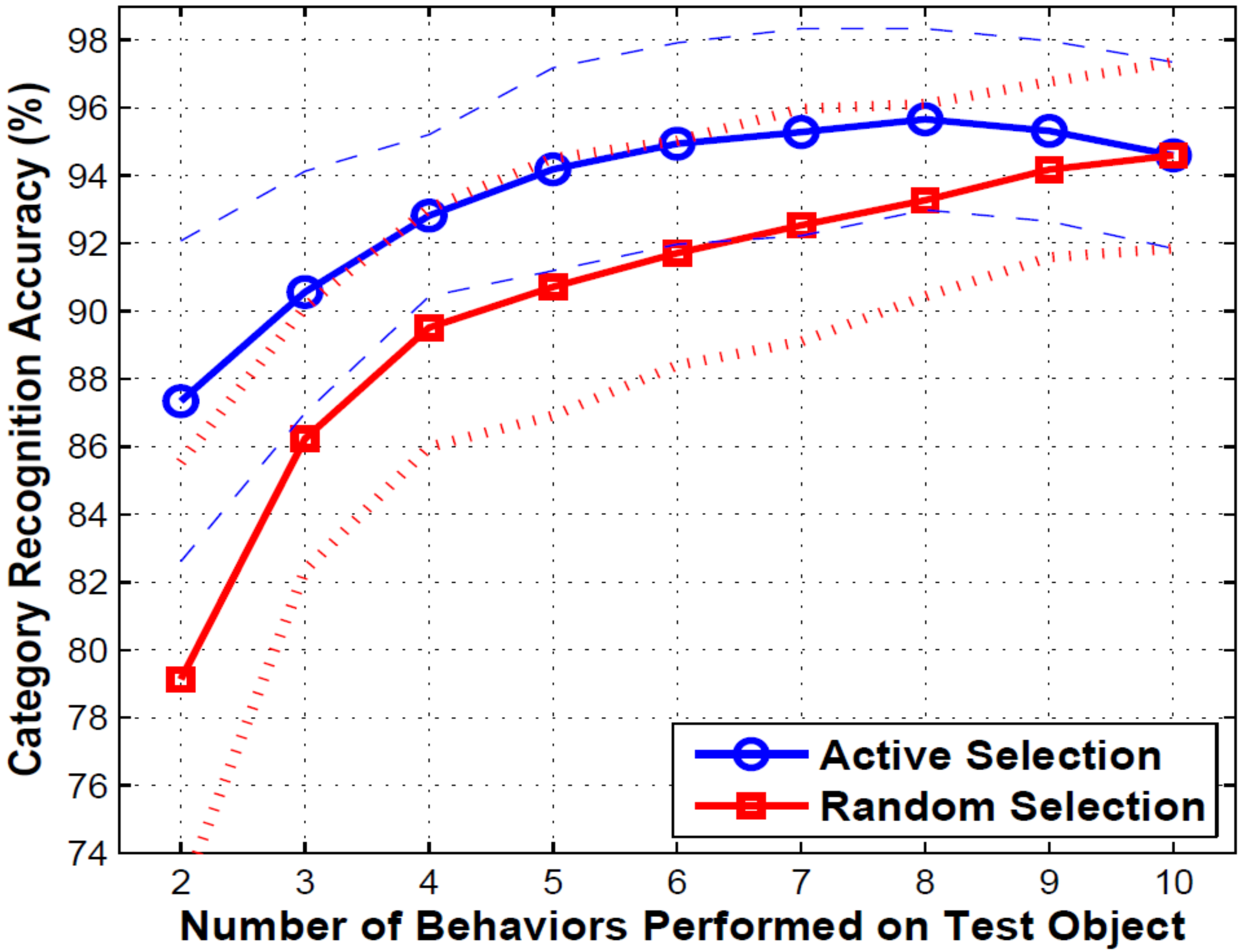
Remaining Behaviors and Associated Confusion:



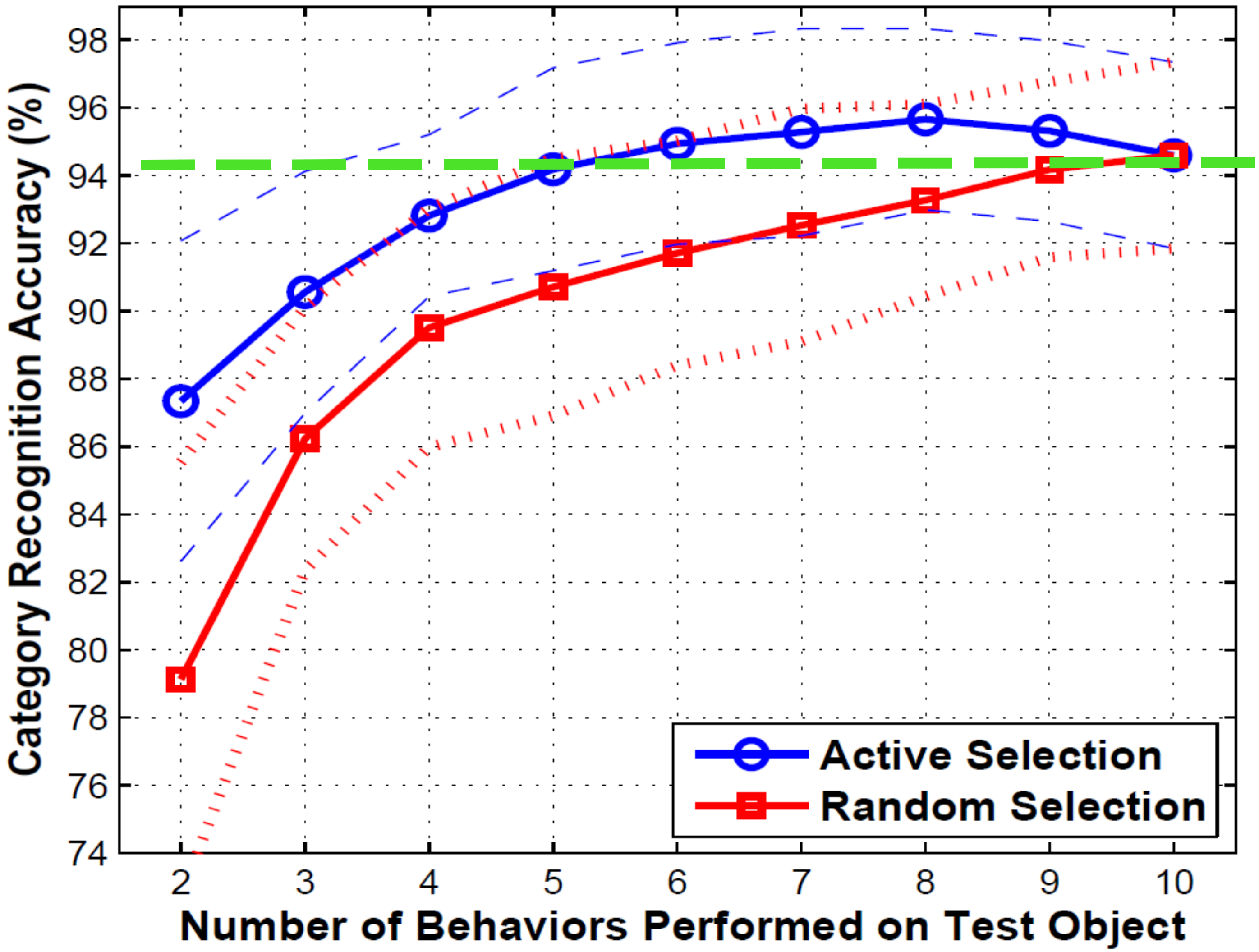
Active Behavior Selection

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



Discussion

What are some of the limitations of the experiment?

What are some ways to address them?

What other possible senses can you think of that would be useful to a robot?

References

Sinapov, J., Bergquist, T., Schenck, C., Ohiri, U., Griffith, S., and Stoytchev, A. (2011) *Interactive Object Recognition Using Proprioceptive and Auditory Feedback*. International Journal of Robotics Research, Vol. 30, No. 10, pp. 1250-1262

Sinapov, J., Schenck, C., Staley, K., Sukhoy, V., and Stoytchev, A. (2014) *Grounding Semantic Categories in Behavioral Interactions: Experiments with 100 Objects.* Robotics and Autonomous Systems, Vol. 62, No. 5, pp. 632-645

THE END