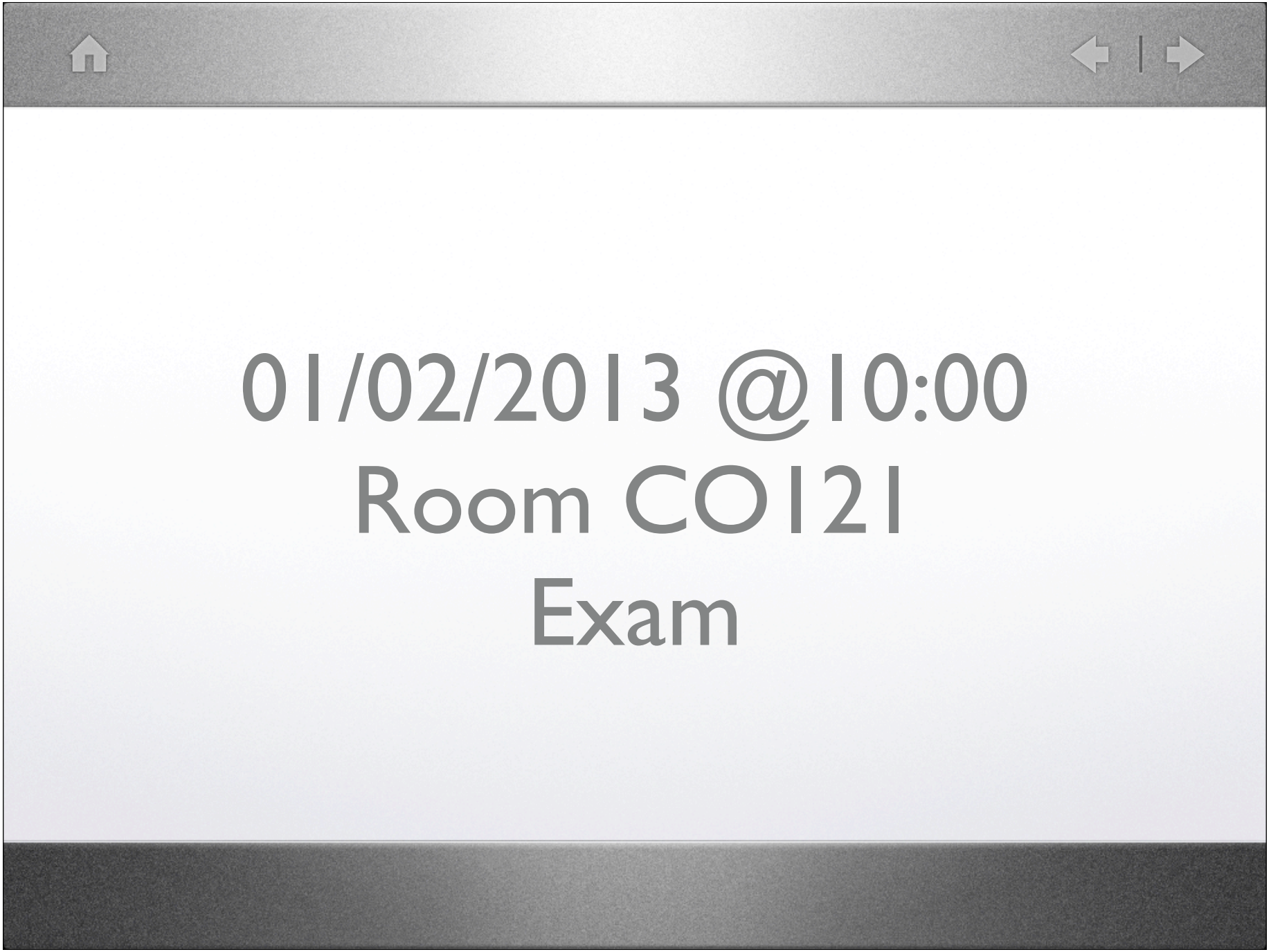



Cognitive Vision for Cognitive Systems

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


18/12/2012 @14:15-17:00
Room MAB1486
Mandatory Experience 3


19/12/2012 @12:15-15:00
Room MAB1486
Extra Experiences

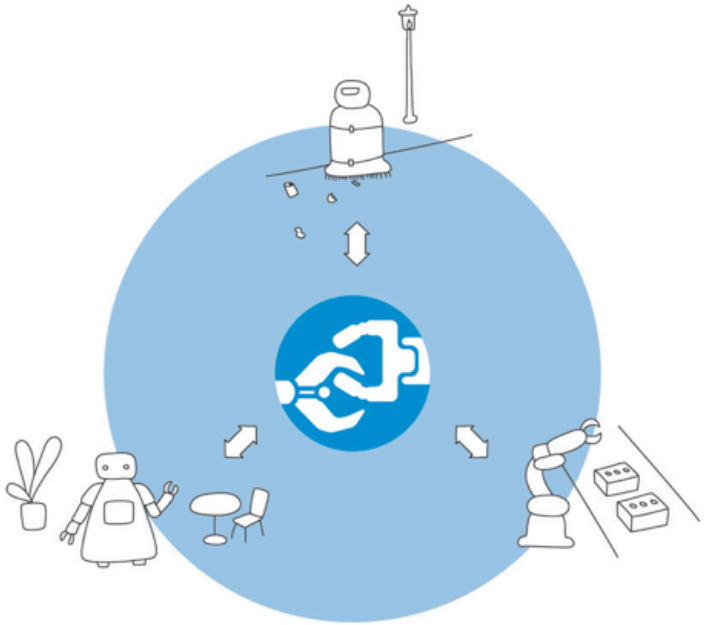

01/02/2013 @10:00
Room CO121
Exam


Transfer Learning,
The Robotics Perspective:
Web robotics

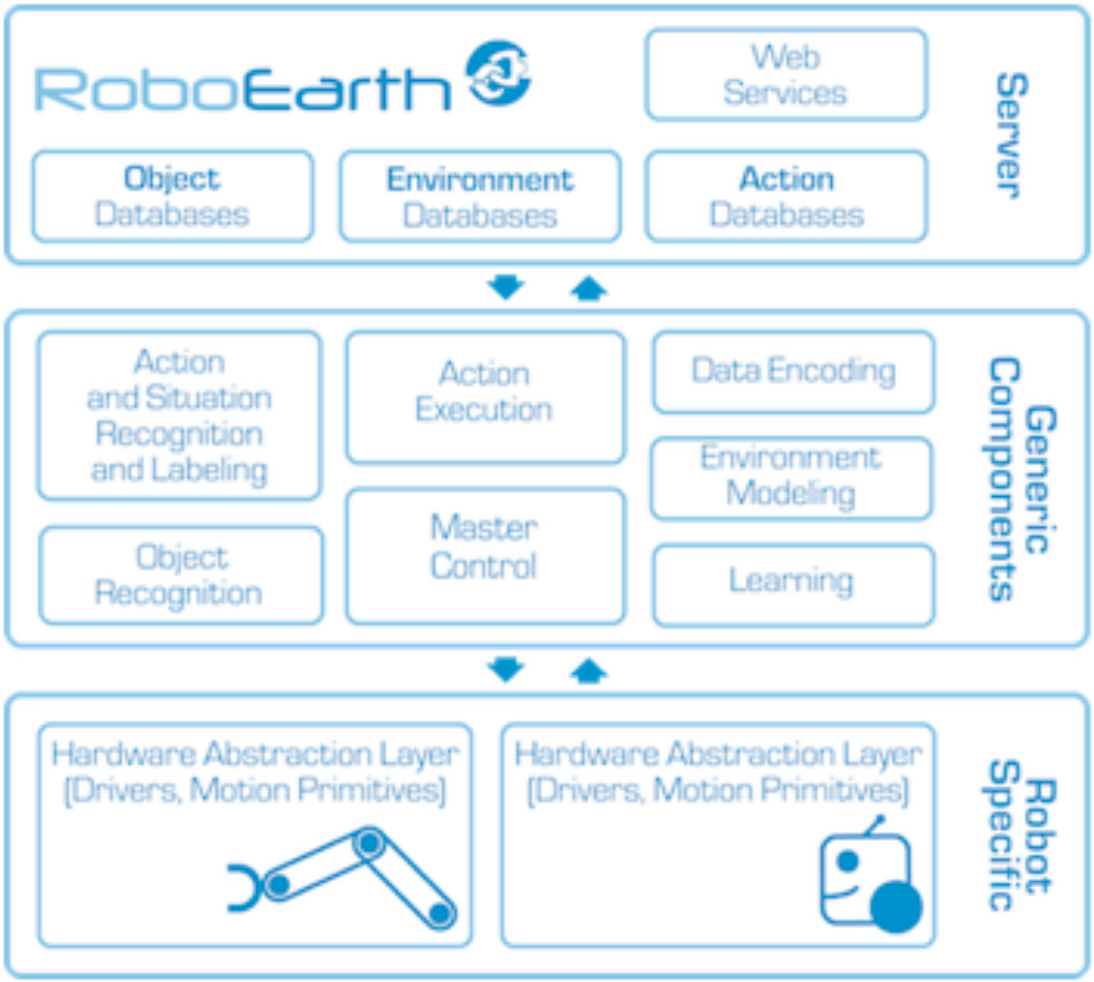
RoboEarth

What is RoboEarth?

At its core, RoboEarth is a World Wide Web for robots: a giant network and database repository where robots can share information and learn from each other about their behavior and their environment. Bringing a new meaning to the phrase "experience is the best teacher", the goal of RoboEarth is to allow robotic systems to benefit from the experience of other robots, paving the way for rapid advances in machine cognition and behaviour, and ultimately, for more subtle and sophisticated human-machine interaction.

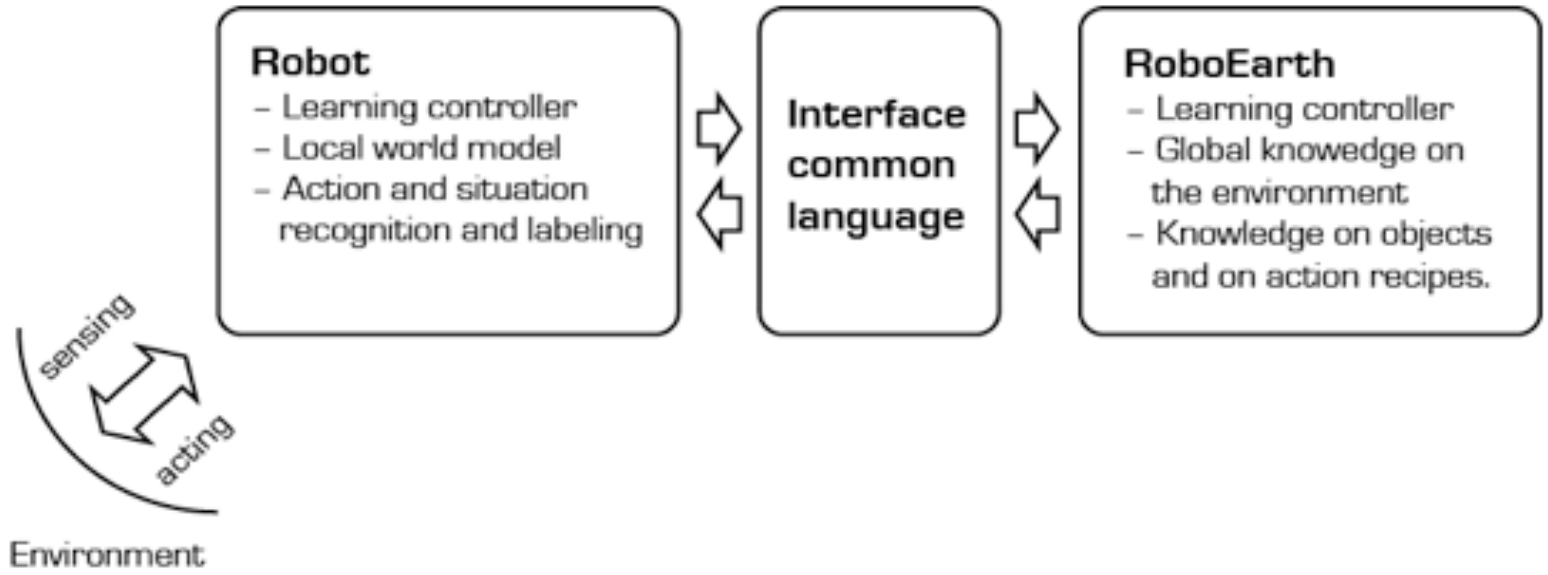


RoboEarth offers a complete Cloud Robotics infrastructure, which includes everything needed to close the loop from robot to RoboEarth to robot. The RoboEarth World-Wide-Web style database is implemented on a server with Internet and Intranet functionality, making it attractive for both research and business applications. It stores information required for object recognition (e.g., images, object models), navigation (e.g., maps, world models), tasks (e.g., action recipes, manipulation strategies) and hosts intelligent services (e.g., image annotation, offline learning).


RoboEarth
Web Services
Object Databases
Environment Databases
Action Databases
Server
Action and Situation Recognition and Labeling
Action Execution
Data Encoding
Environment Modeling
Object Recognition
Master Control
Learning
Generic Components
Hardware Abstraction Layer (Drivers, Motion Primitives)
Hardware Abstraction Layer (Drivers, Motion Primitives)
Robot Specific

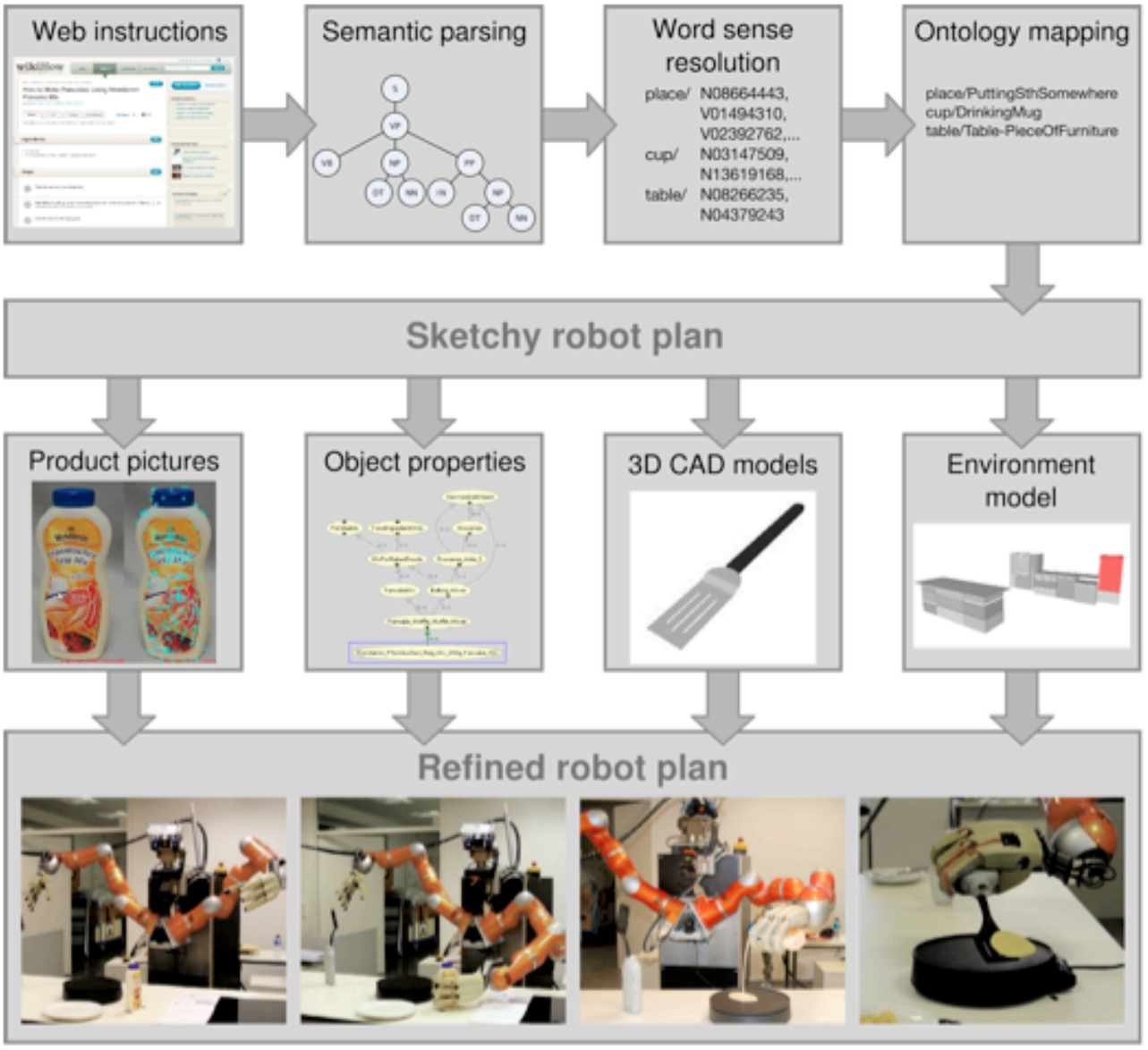
Motivation

Until recently, robots have not been capable of understanding and coping with unstructured environments (like the ones humans work in) because their systems have relied on knowing in advance the specifics of every possible situation they might encounter. Each response to a contingency has had to be programmed in advance, and systems have had to rebuild their world model from sensor data each time they had to perform a new task.


Robot
– Learning controller
– Local world model
– Action and situation recognition and labeling
Interface common language
RoboEarth
– Learning controller
– Global knowedge on the environment
– Knowledge on objects and on action recipes.
sensing
acting
Environment


http://www.youtube.com/watch?
feature=player_embedded&v=o3oi7PZkFII

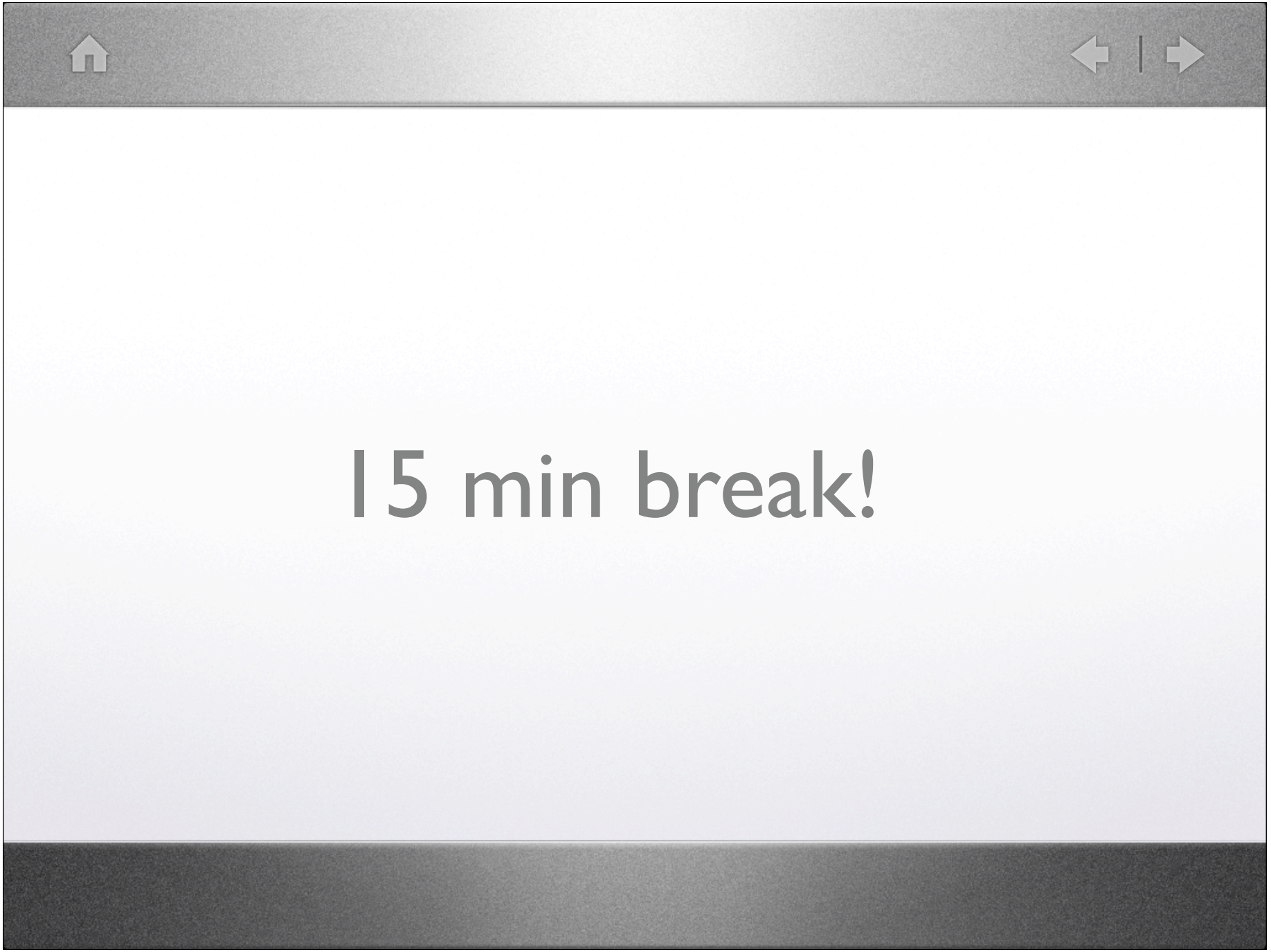
RoboHow.cog


Web instructions
Semantic parsing
Word sense resolution
place/ N08664443, V01494310, V02392762,...
cup/ N03147509, N13619168,...
table/ N08266235, N04379243
Ontology mapping
place/PuttingSthSomewhere
cup/DrinkingMug
table/Table-PieceOfFurniture
Sketchy robot plan
Product pictures
Object properties
3D CAD models
Environment model
Refined robot plan

http://www.youtube.com/watch?feature=player_embedded&v=4usoE981e7I


http://www.youtube.com/watch?
feature=player_embedded&v=BVAIt0FYmiI


http://www.youtube.com/watch?
feature=player_embedded&v=_SIUCrmE8J0


15 min break!

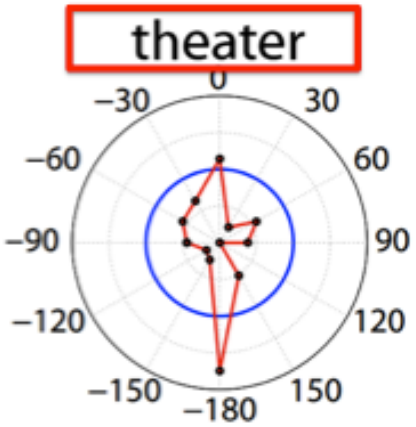

Spotlight presentations:
computer vision


Spotlight -CV1
Recognizing scene viewpoint using panoramic place representation
J. Xiao, K. A. Ehinger, A. Oliva, A. Torralba
Proc CVPR 2012

Scene Viewpoint Recognition

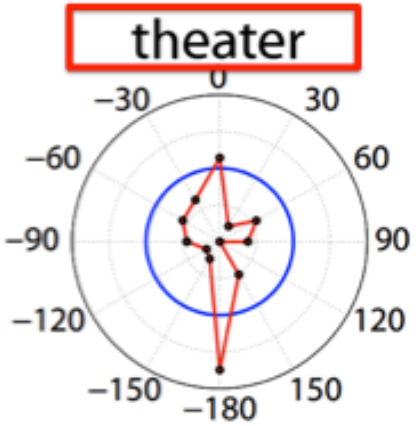

Problem Definition

- Place Categorization
- Viewpoint Recognition

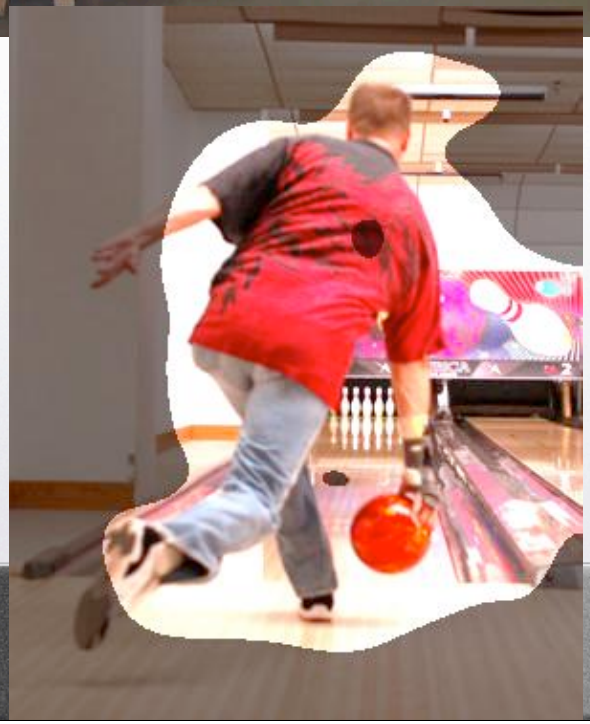


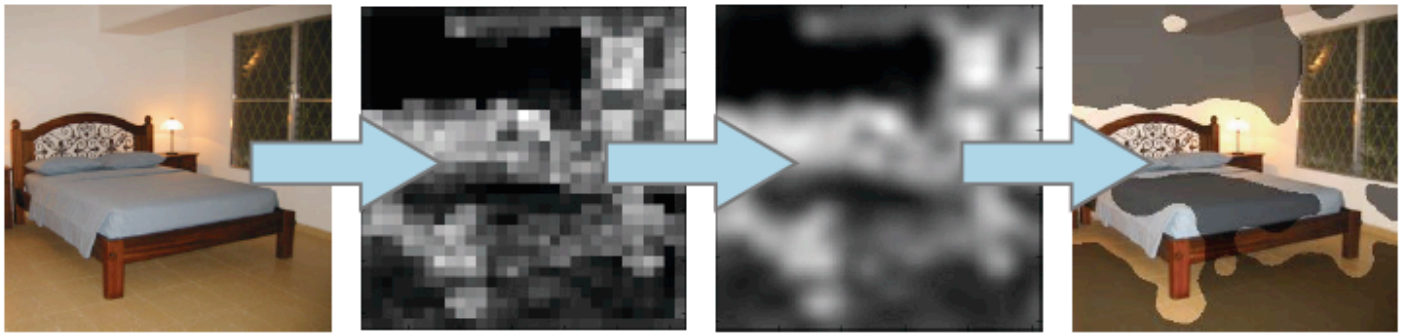

Algorithm Pipeline

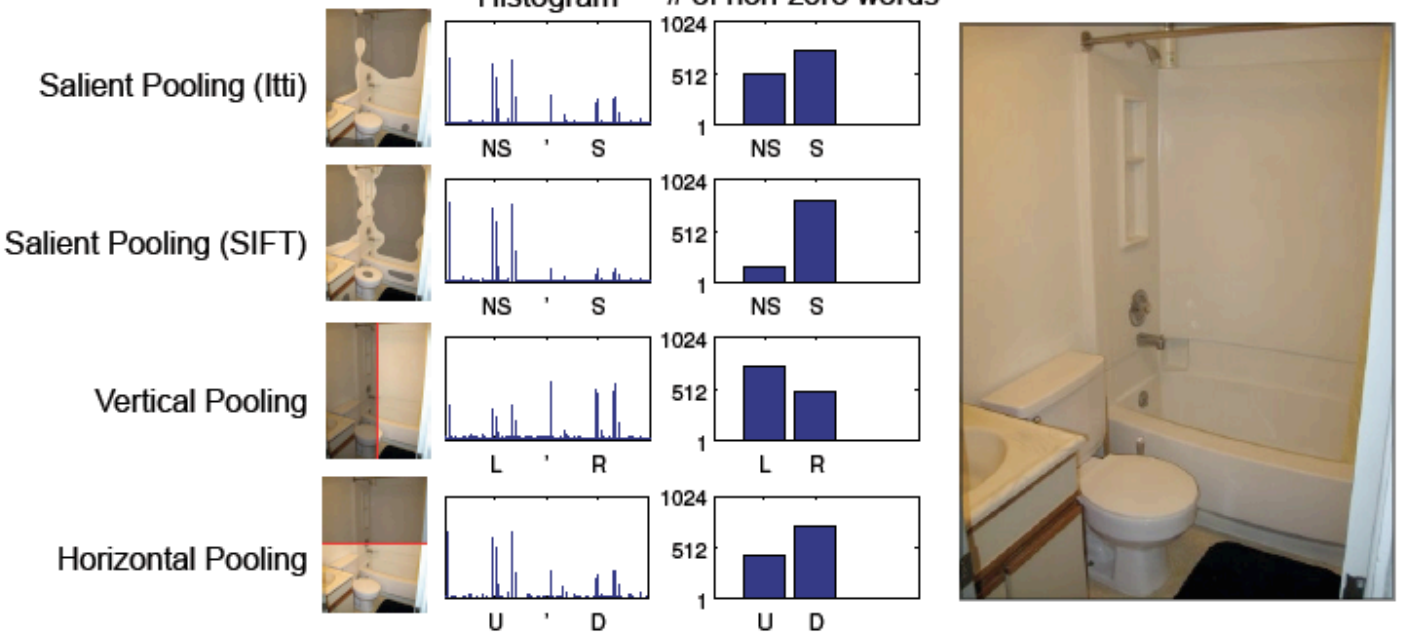
- Step 1: Place Categorization
- Step 2: Panorama alignment & viewpoint classifier

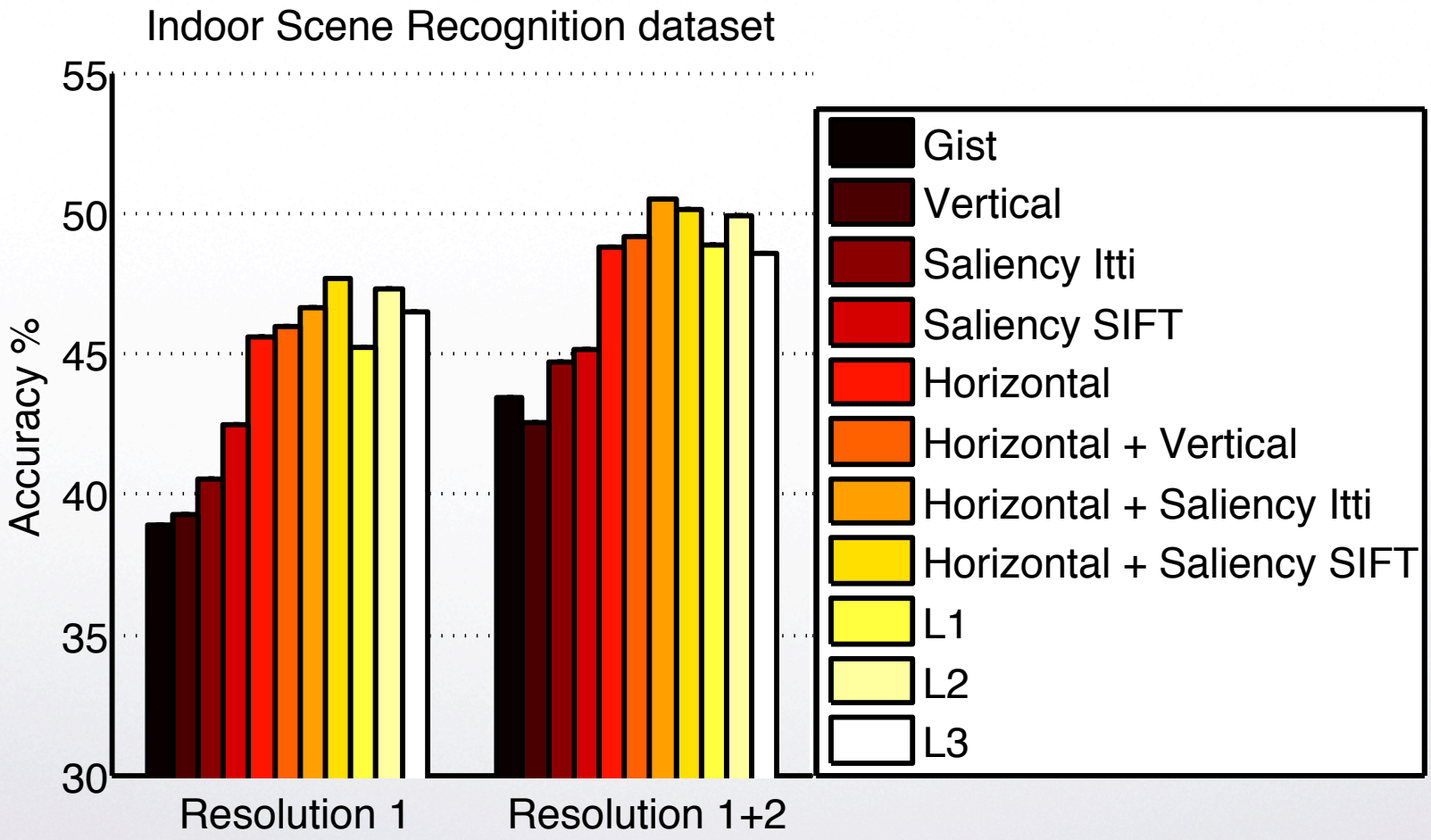


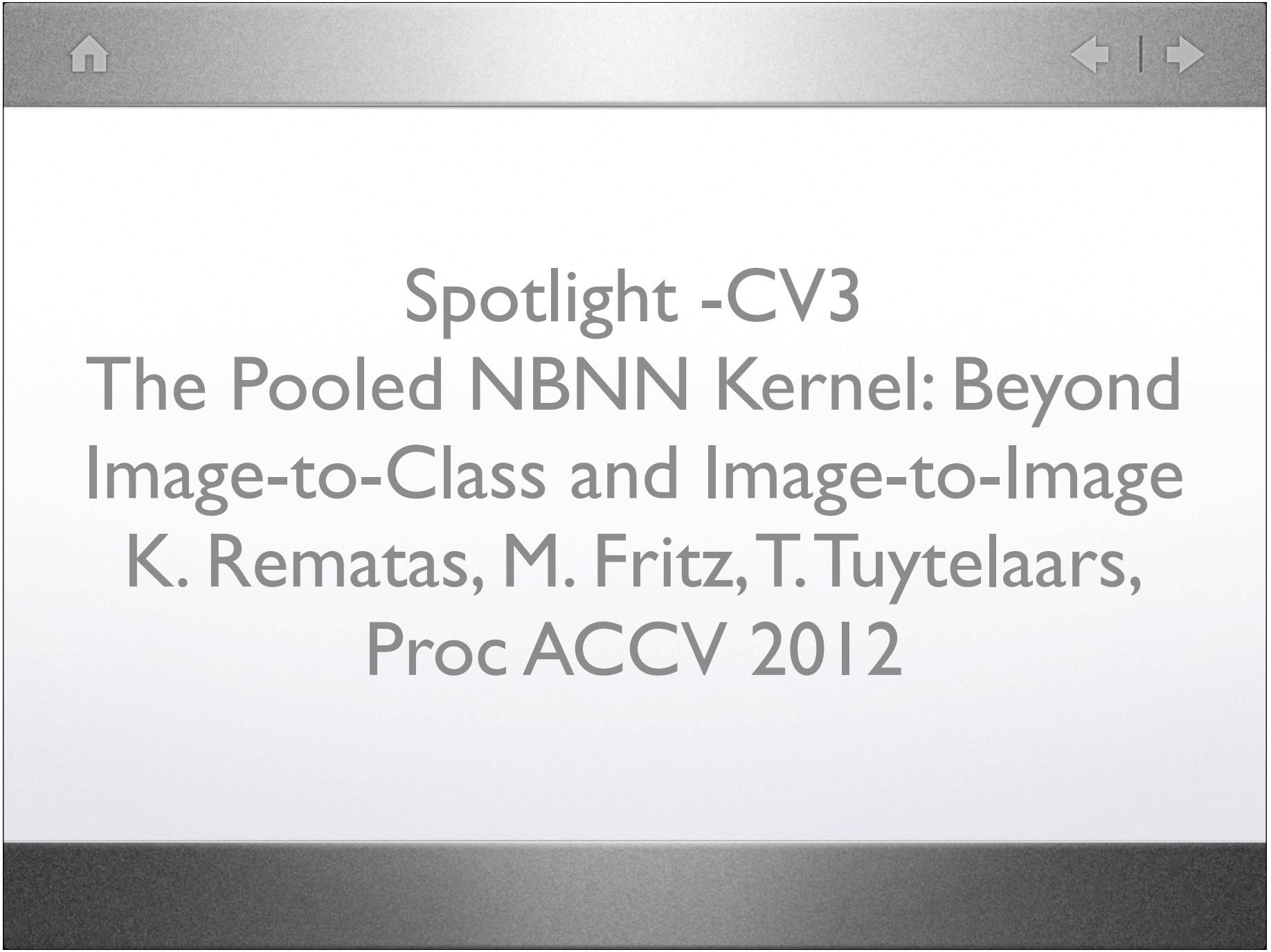

Spotlight -CV2
Indoor scene recognition using task and saliency-driven feature pooling.
M. Fornoni, B. Caputo
Proc BMVC 2012

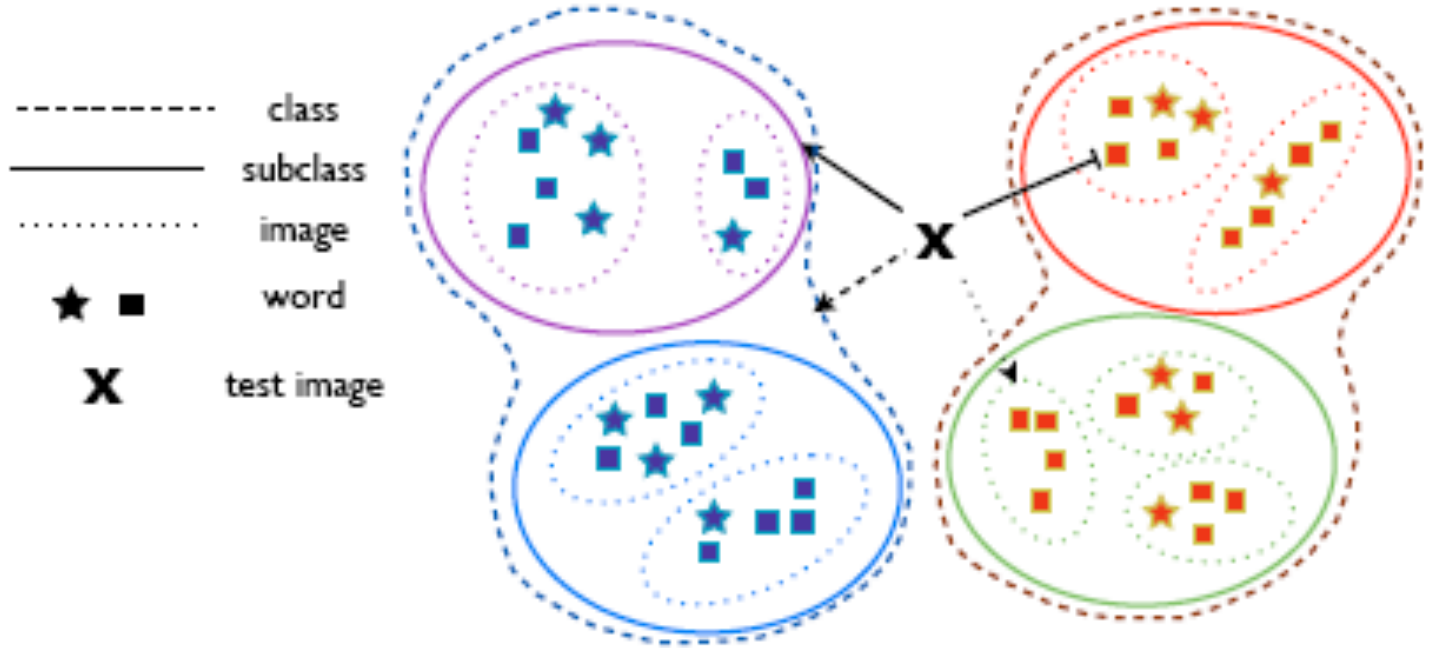





Histogram
# of non-zero words
Salient Pooling (Itti)
Salient Pooling (SIFT)
Vertical Pooling
Horizontal Pooling
NS ' S
L ' R
U ' D
1024
512
1
NS S
L R
U D


Indoor Scene Recognition dataset
Accuracy %
55
50
45
40
35
30
Resolution 1
Resolution 1+2
Gist
Vertical
Saliency Itti
Saliency SIFT
Horizontal
Horizontal + Vertical
Horizontal + Saliency Itti
Horizontal + Saliency SIFT
L1
L2
L3


Spotlight -CV3
The Pooled NBNN Kernel: Beyond Image-to-Class and Image-to-Image
K. Rematas, M. Fritz, T. Tuytelaars,
Proc ACCV 2012


class
subclass
image
word
test image
X

Method	NBNN	NBNN kernel
I2C (baseline)	72.0±0.3	76.0 ± 0.8
I2Sub-2	72.3±1.5	76.4 ± 0.4
I2Sub-4	68.9±1	76.3 ± 2.3
I2Sub-6	68.1±3.1	76.6 ± 1.0
I2Sub-10	66.1±8	77.0 ± 0.5
I2Sub-20	66.2±1.7	76.4 ± 0.4
I2Word-2	70.7±0.8	74.8 ± 1.3
I2Word-4	67.0±0.6	74.8 ± 0.5
I2Word-8	63.5±1.3	75.9 ± 0.3
I2Word-16	61.0±0.6	74.4 ± 0.7
I2Word-32	60.2±3.2	71 ± 3
Average Kernel	-	**79.7±1.5**

Fig. 3. Classification accuracy of NBNN and NBNN kernel for different pooling strategies in 15 Scenes.

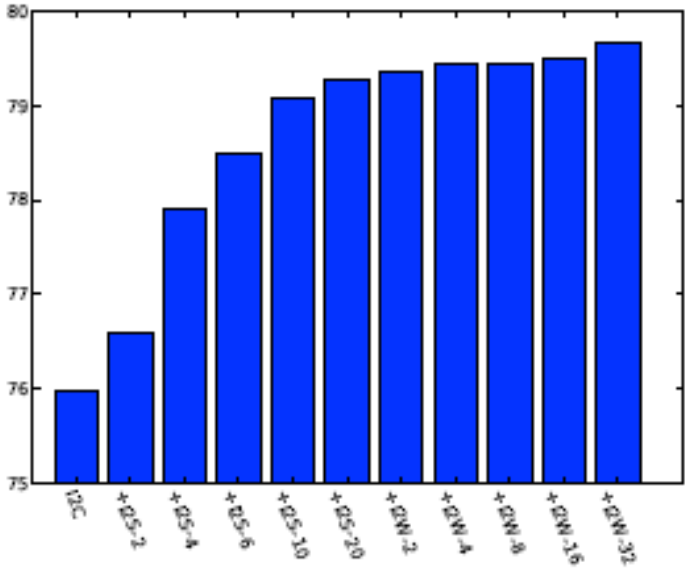
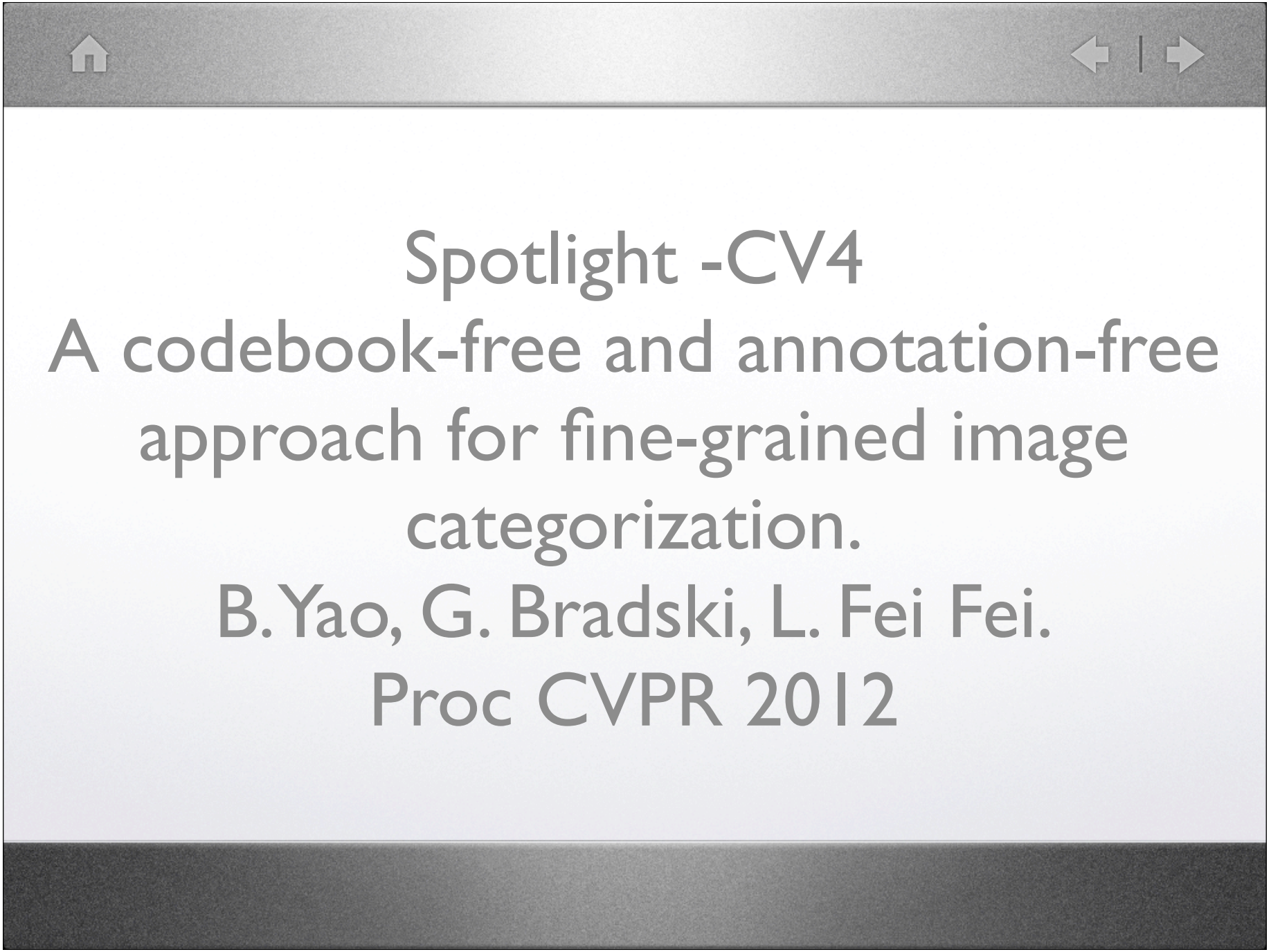
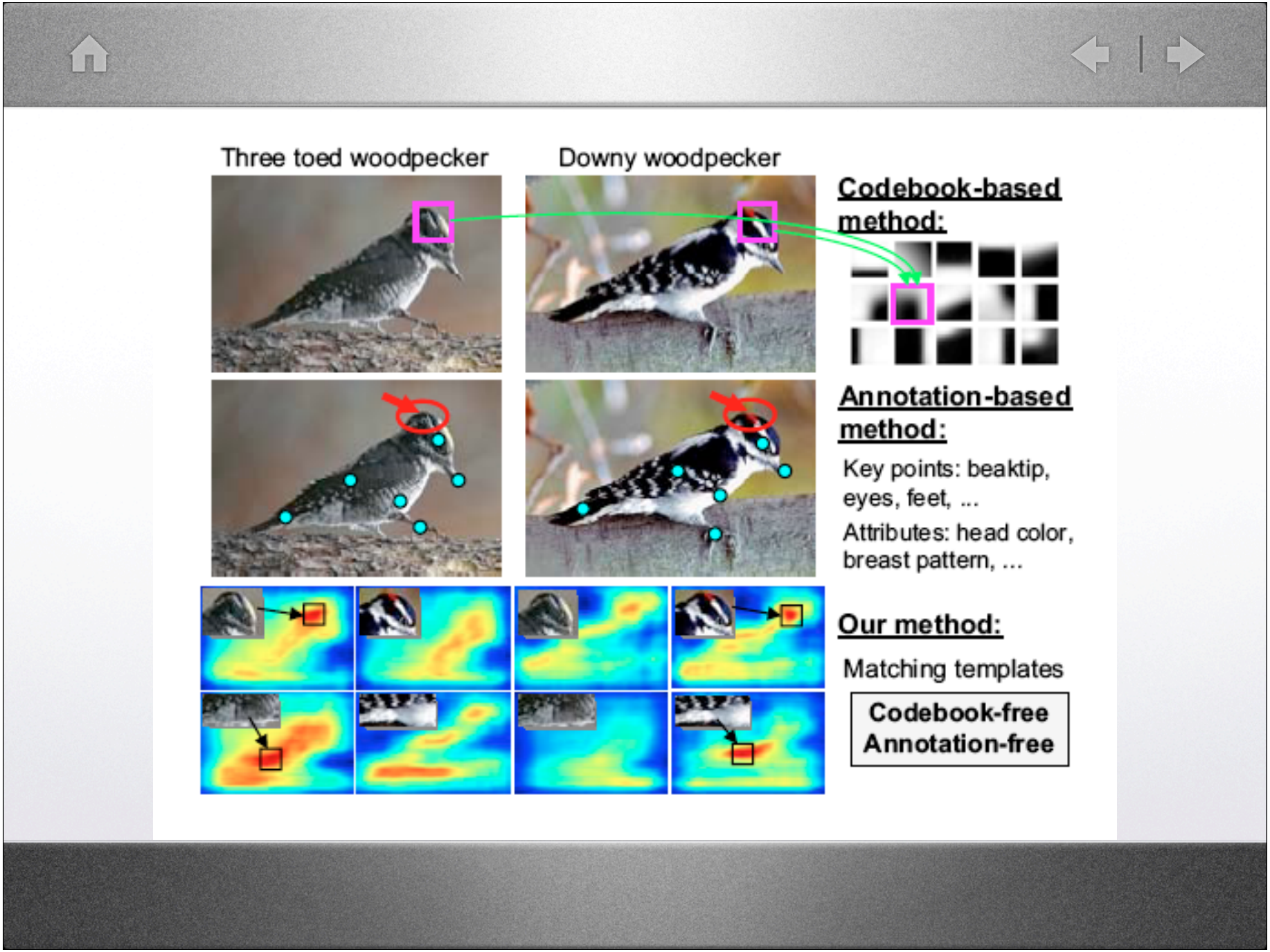
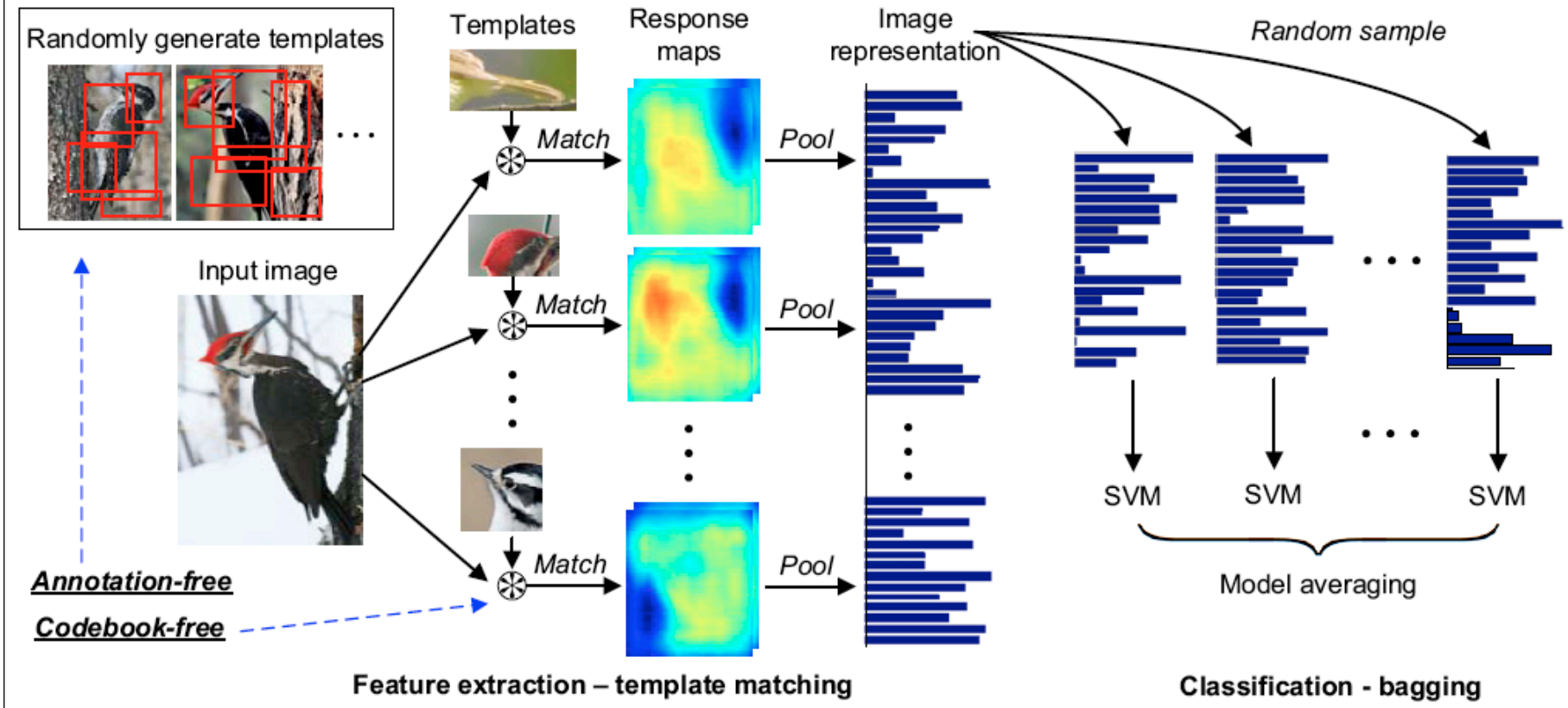
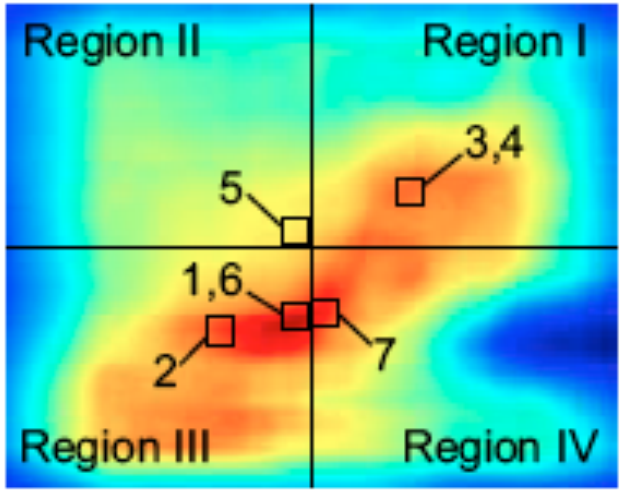


Fig. 4. Performance gain for 15 Scenes when adding kernels. The order of addition is the same order as in Table 3.

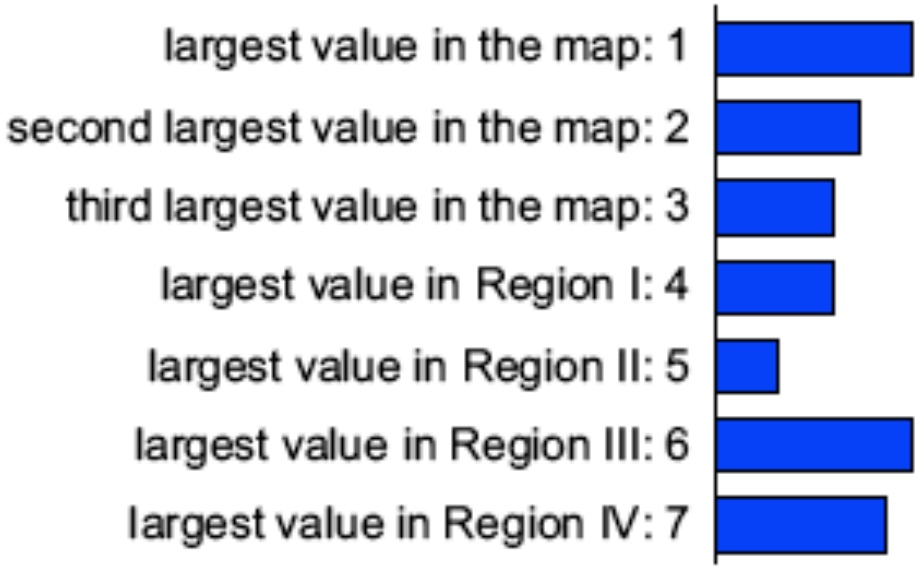

Spotlight -CV4
A codebook-free and annotation-free approach for fine-grained image categorization.
B. Yao, G. Bradski, L. Fei Fei.
Proc CVPR 2012


Three toed woodpecker
Downy woodpecker
Codebook-based method:
Annotation-based method:
Key points: beaktip, eyes, feet, ...
Attributes: head color, breast pattern, ...
Our method:
Matching templates
Codebook-free
Annotation-free

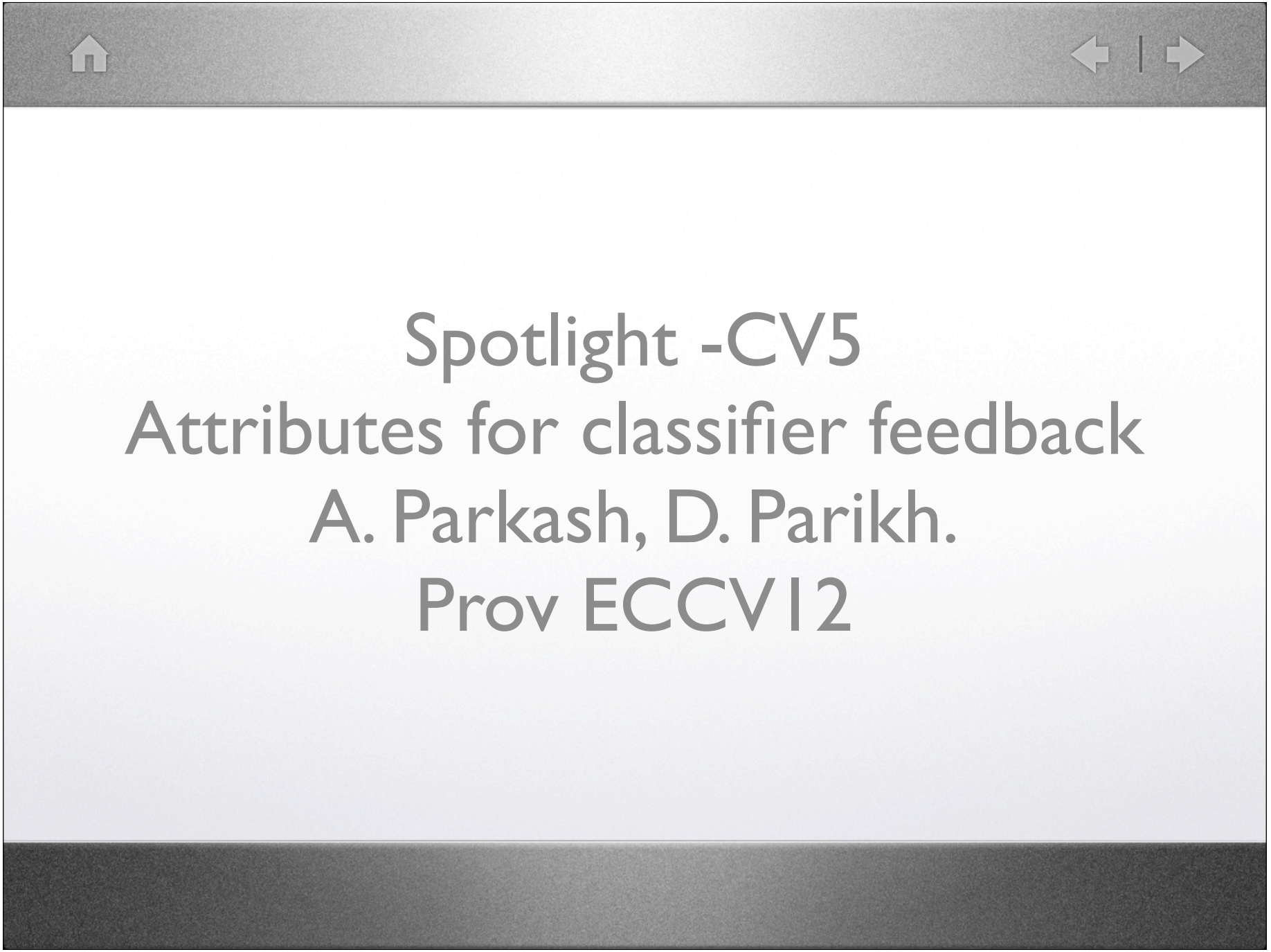

Randomly generate templates
Templates
Response maps
Image representation
Random sample
Match
Pool
Input image
Annotation-free
Codebook-free
SVM
SVM
SVM
Model averaging
Feature extraction – template matching
Classification - bagging

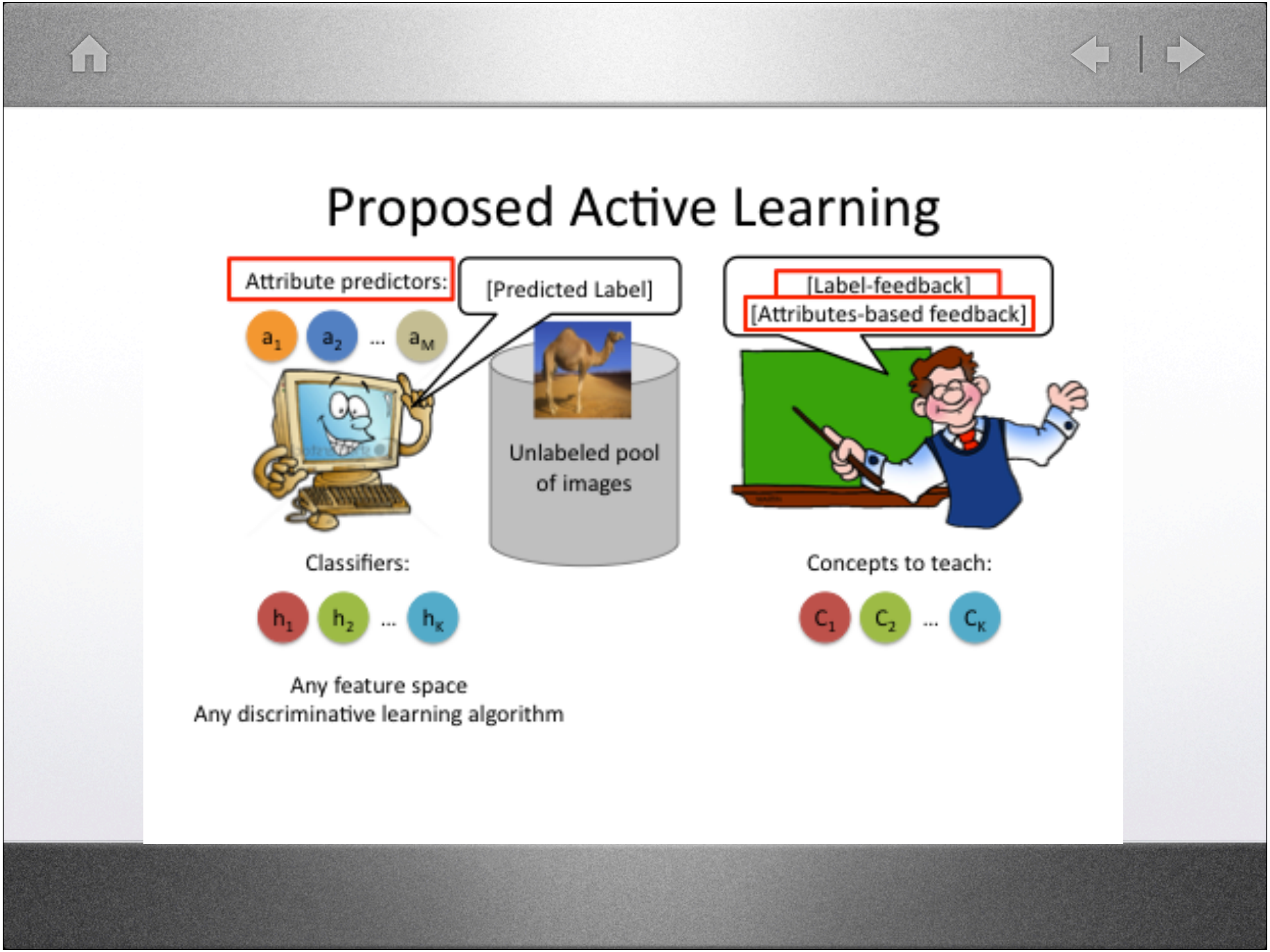


Response map

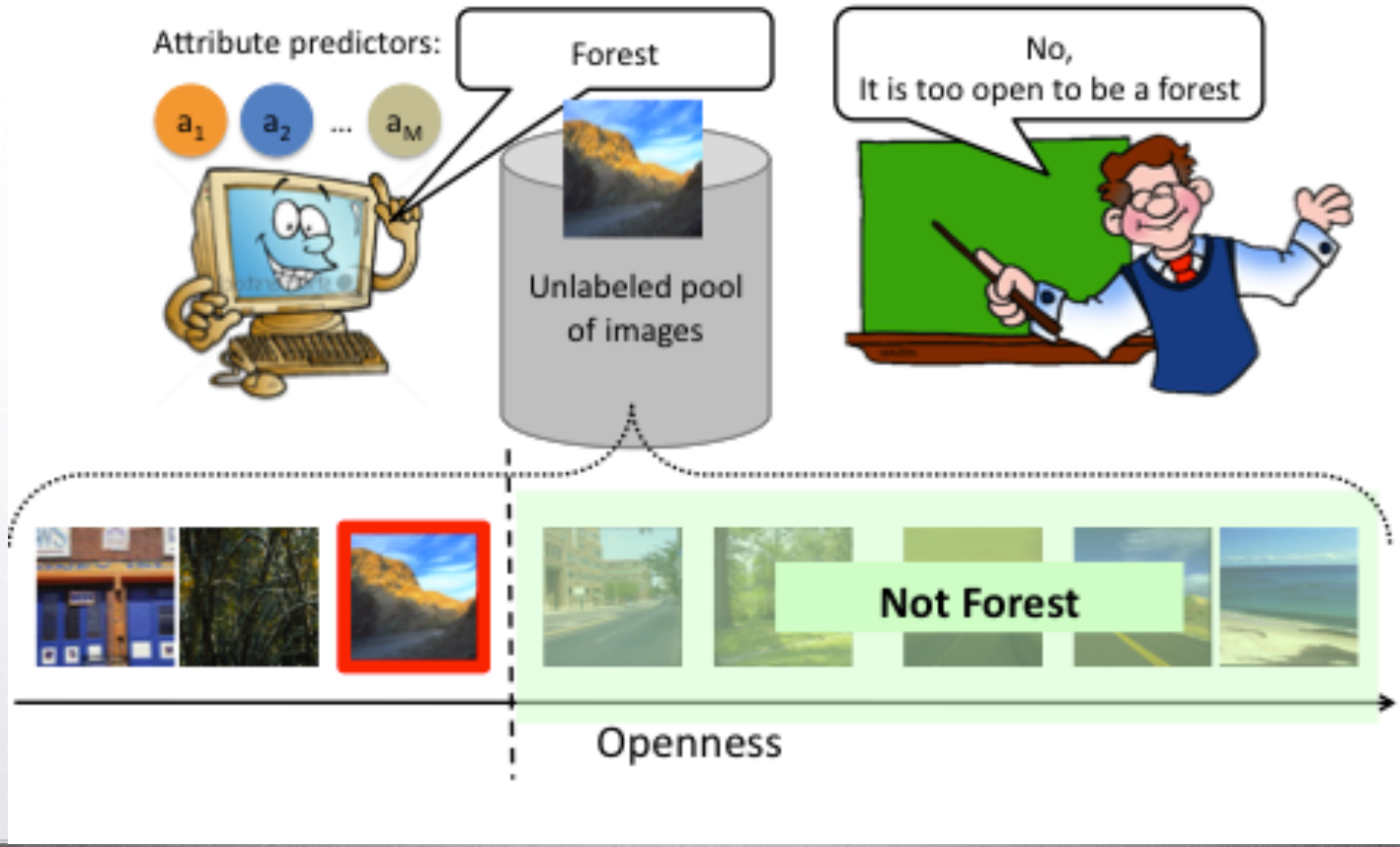


Pooling result

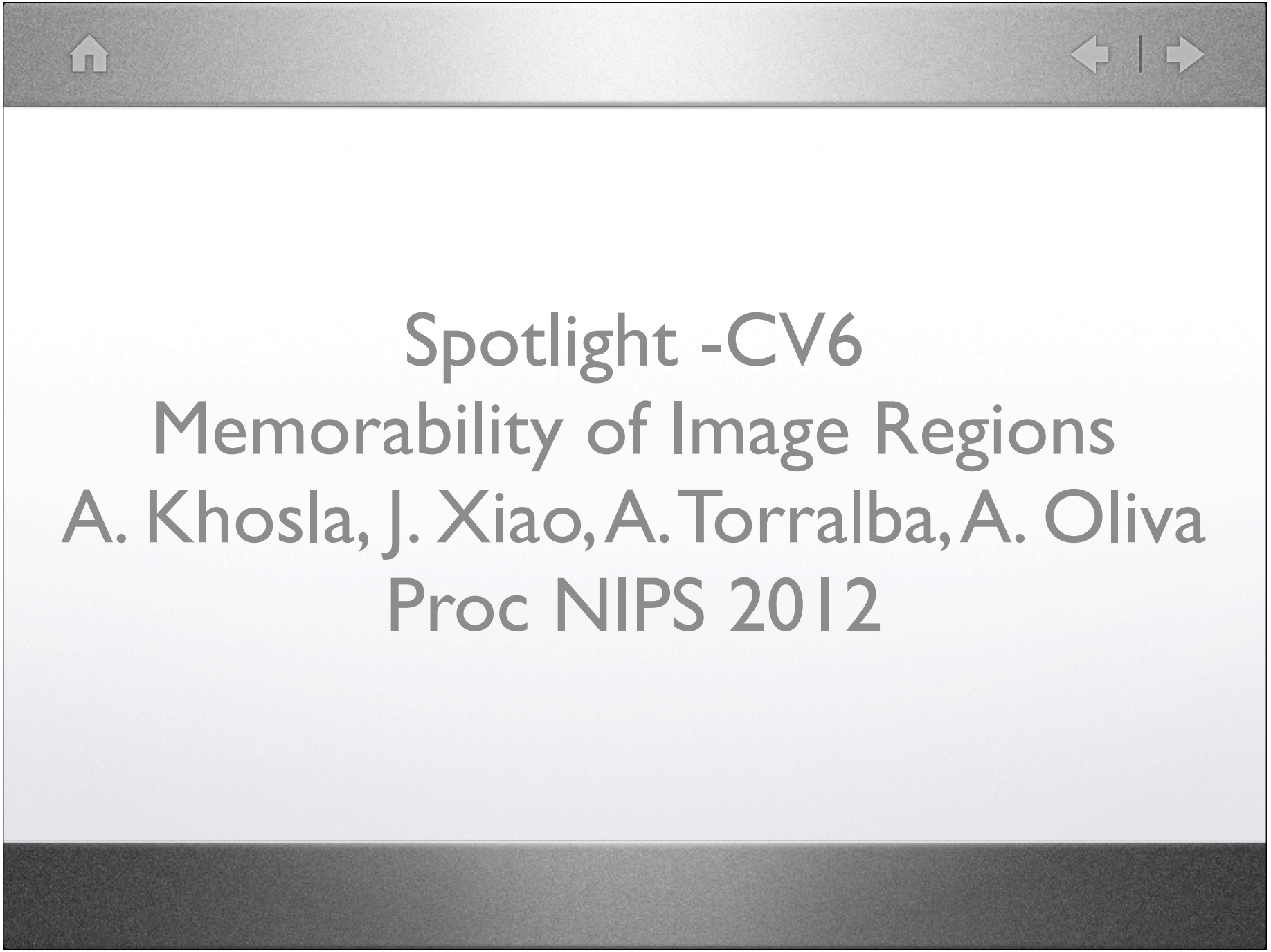

Spotlight -CV5
Attributes for classifier feedback
A. Parkash, D. Parikh.
Prov ECCV12

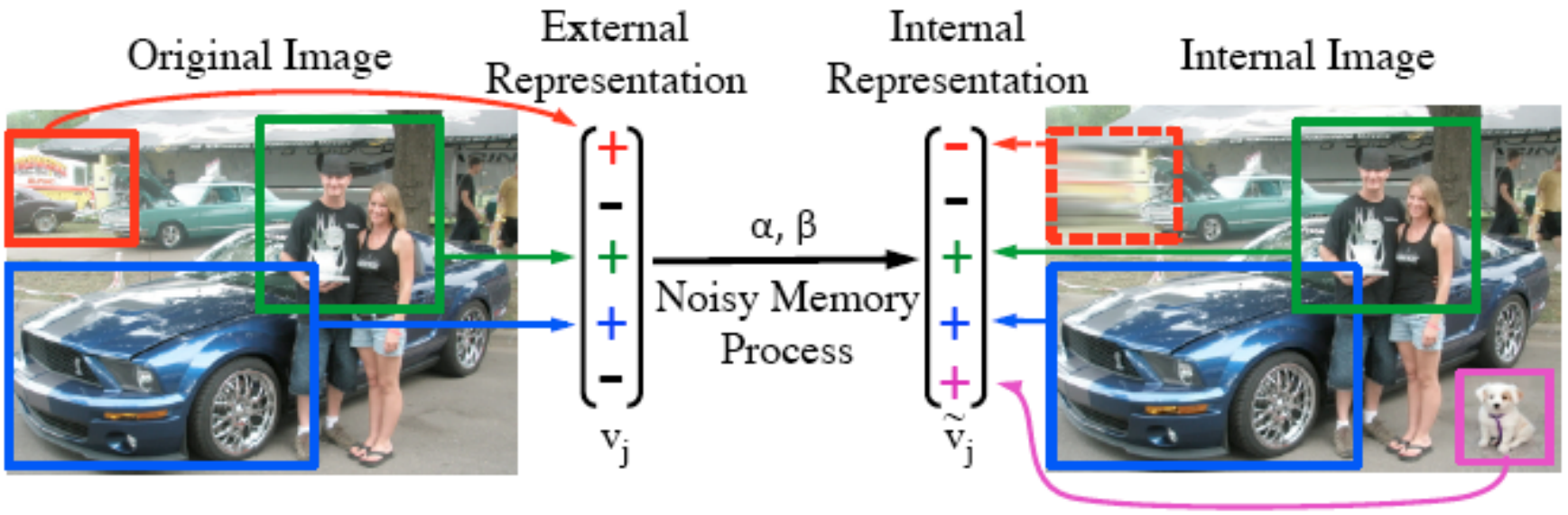

Proposed Active Learning
Attribute predictors:
a_1 a_2 ... a_M
[Predicted Label]
Unlabeled pool of images
[Label-feedback]
[Attributes-based feedback]
Classifiers:
h_1 h_2 ... h_K
Concepts to teach:
C_1 C_2 ... C_K
Any feature space
Any discriminative learning algorithm

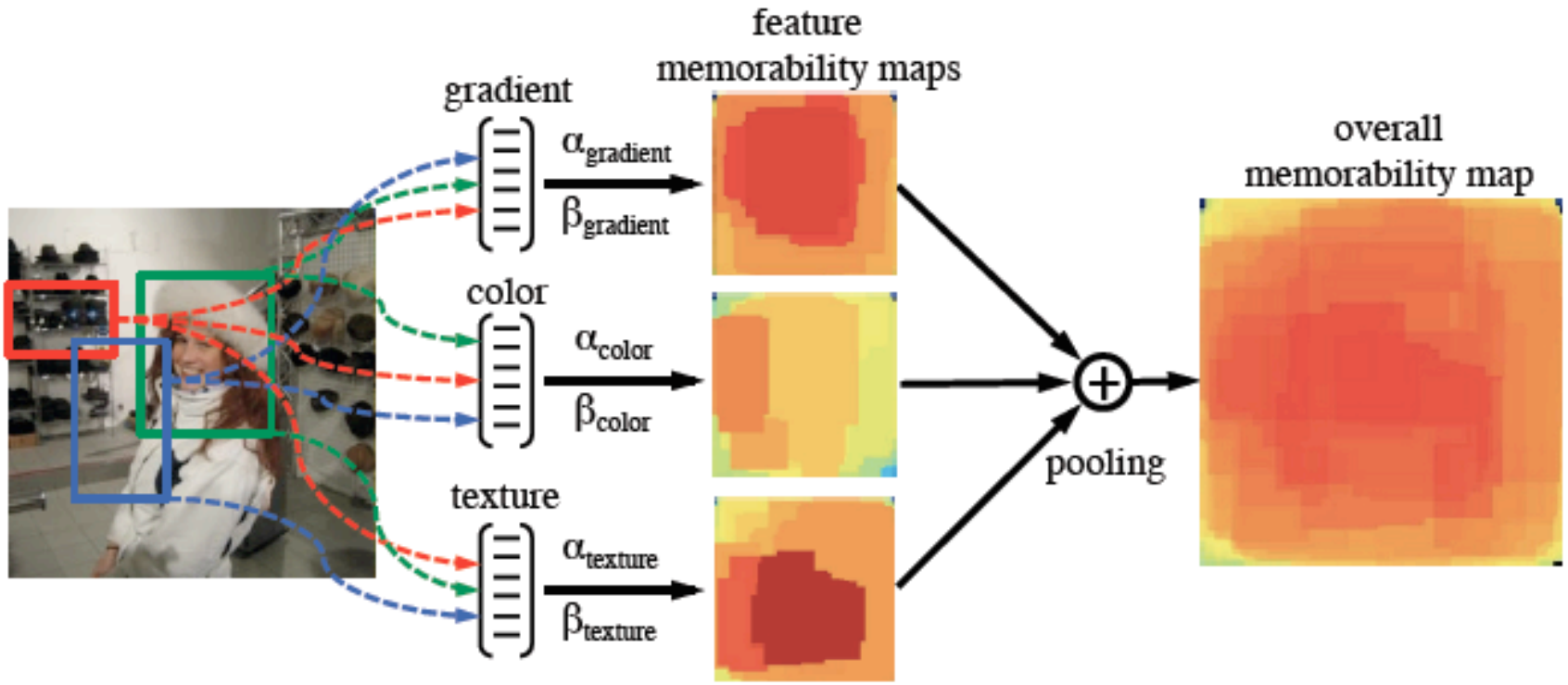
Attributes-based Feedback

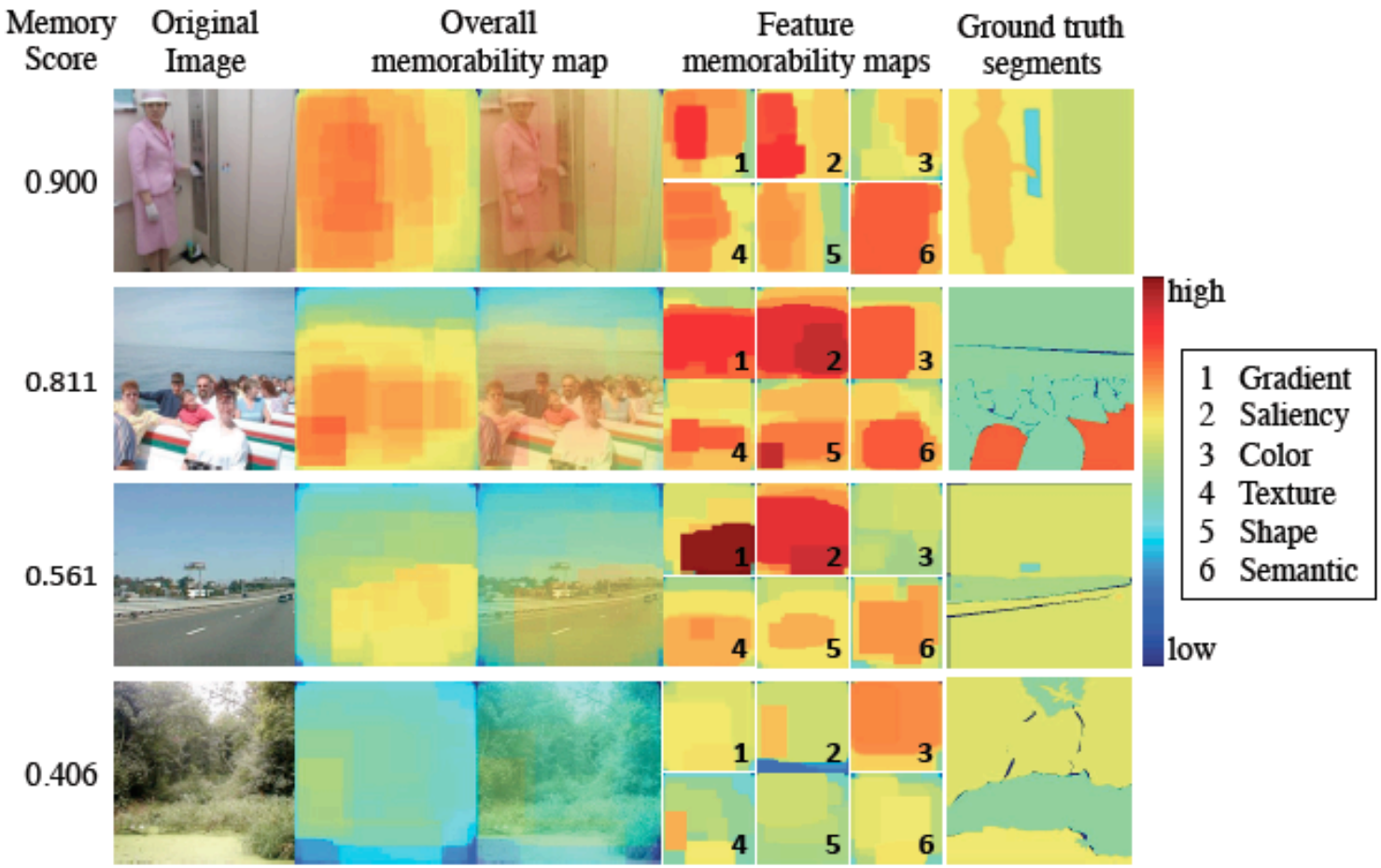


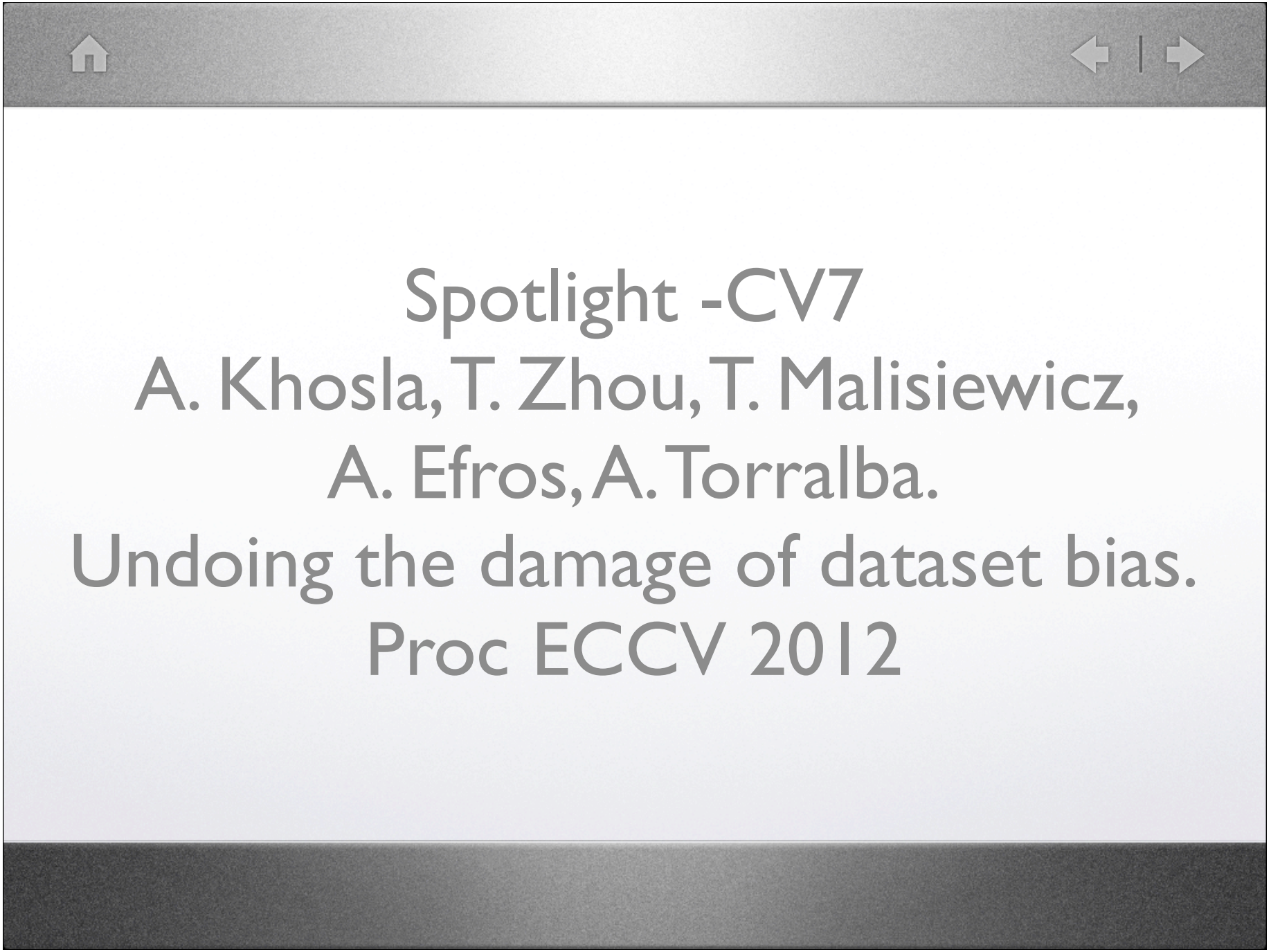

Label-based Feedback
• Accept: Yes, this is a forest.
– Strong: It is not anything else
Example: Classification
– Weak: It can be other things
Example: Annotation
Forest
Unlabeled pool
of images

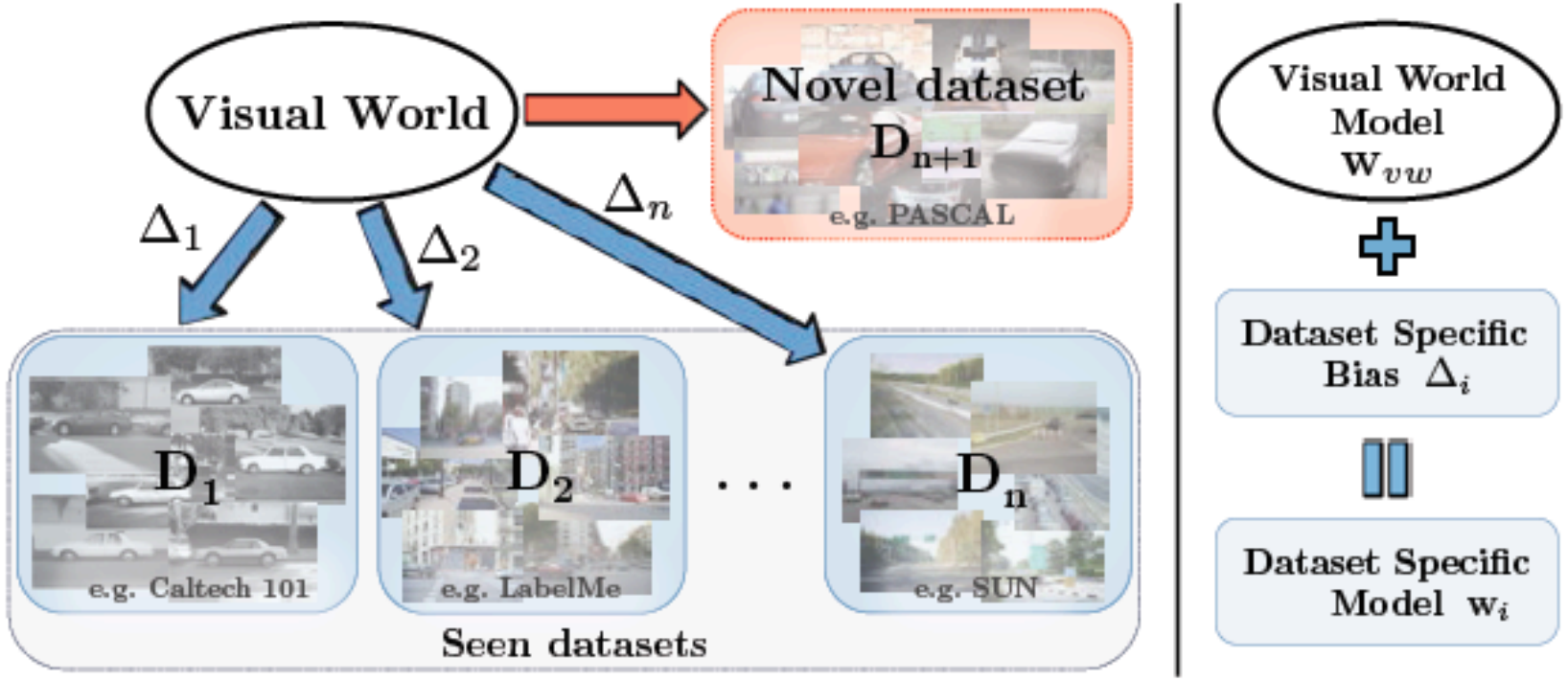

Spotlight -CV6
Memorability of Image Regions
A. Khosla, J. Xiao, A. Torralba, A. Oliva
Proc NIPS 2012


Original Image
External Representation
Internal Representation
Internal Image
α, β
Noisy Memory Process
v_j
$\tilde{v}_j$


feature
memorability maps
gradient
$\alpha_{gradient}$
$\beta_{gradient}$
color
α_{color}
β_{color}
texture
$\alpha_{texture}$
$\beta_{texture}$
pooling
overall
memorability map


Memory Score
Original Image
Overall memorability map
Feature memorability maps
Ground truth segments
0.900
0.811
0.561
0.406
1
2
3
4
5
6
high
low
1 Gradient
2 Saliency
3 Color
4 Texture
5 Shape
6 Semantic


Spotlight -CV7
A. Khosla, T. Zhou, T. Malisiewicz,
A. Efros, A. Torralba.
Undoing the damage of dataset bias.
Proc ECCV 2012

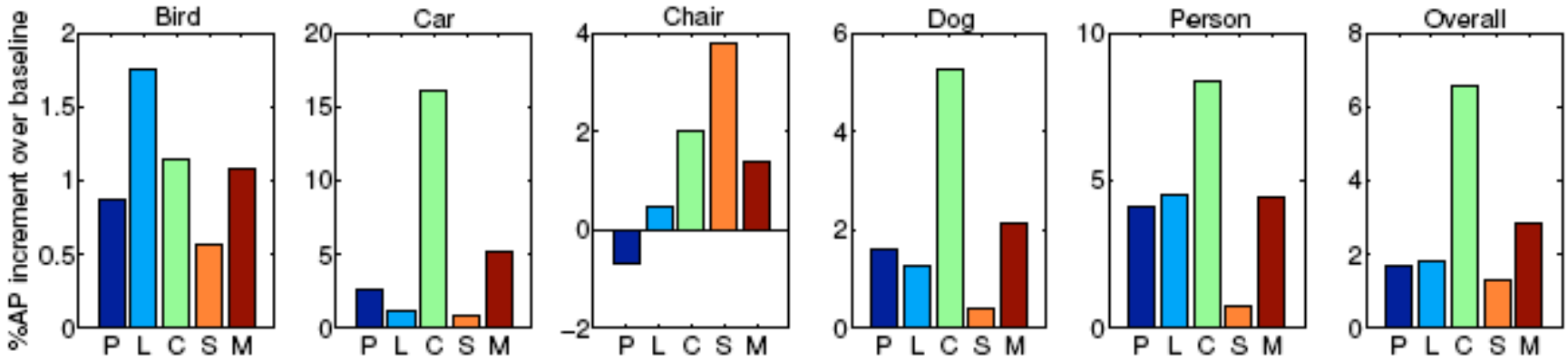

Visual World
Novel dataset
D_{n+1}
e.g. PASCAL
Δ_1
Δ_2
Δ_n
D_1
e.g. Caltech 101
D_2
e.g. LabelMe
. . .
D_n
e.g. SUN
Seen datasets
Visual World Model W_{vw}
Dataset Specific Bias Δ_i
Dataset Specific Model w_i


(a) Train on all, test on one at a time

Train	Test	$\mathbf{w}_{Pas}$	$\mathbf{w}_{Lab}$	$\mathbf{w}_{Cal}$	$\mathbf{w}_{SUN}$	$\mathbf{w}_{vw}$	SVM_{all}
All	Pas	**0.638**	0.511	0.548	0.495	0.558	0.590
All	Lab	0.690	**0.729**	0.719	0.733	0.729	0.722
All	Cal	0.894	0.928	**0.998**	0.918	0.979	0.936
All	SUN	0.427	0.515	0.530	**0.603**	0.568	0.549
Average		0.662	0.671	0.698	0.687	**0.709**	0.699

(b) Train + test on one

Train	Test	SVM_{one}
Pas	Pas	0.650
Lab	Lab	0.731
Cal	Cal	0.995
SUN	SUN	0.597
Average		0.743


%AP increment over baseline
Bird
Car
Chair
Dog
Person
Overall
P L C S M


Spotlight -CV8
Geodesic Flow Kernel for unsupervised domain adaptation.
B. Gong, Y. Shi, F. Sha, K. Grauman.
Proc CVPR 2012


Motivation

TRAIN

TEST

Mismatch between different domains/datasets

- Object recognition
  - *Ex. [Torralba & Efros'11, Perronnin et al.'10]*
- Video analysis
  - *Ex. [Duan et al.'09, 10]*
- Pedestrian detection
  - *Ex. [Dollár et al.'09]*
- Other vision tasks

Performance degrades significantly!

Images from [Saenko et al.'10].


Unsupervised domain adaptation

- Source domain (labeled)

$$\mathbf{D}_S = \{(x_i, y_i), i = 1, 2, \cdots, N\} \sim P_S(X, Y)$$

- Target domain (unlabeled)

$$\mathbf{D}_T = \{(x_i, \mathbf{?}), i = 1, 2, \cdots, M\} \sim P_T(X, Y)$$

*The two distributions are **not** the same!*

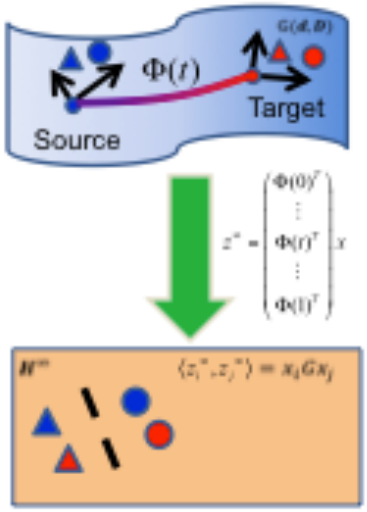
- **Objective**

  Train classification model to work well on the target

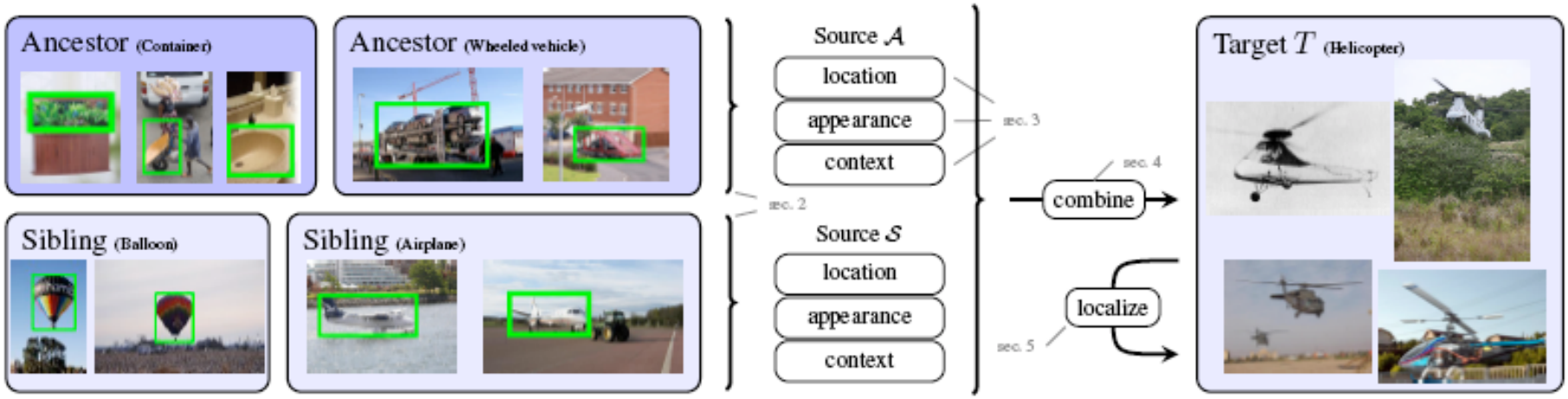
Our approach:
learning a shared representation

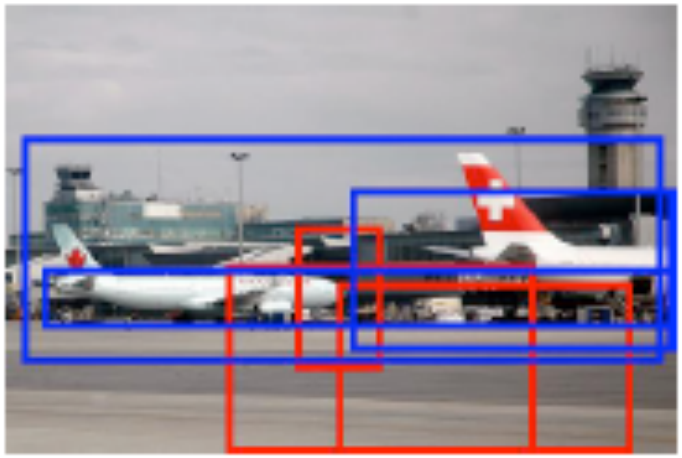
Key insight: bridging the gap

- Fantasize infinite number of domains
- Integrate out analytically idiosyncrasies in domains
- Learn invariant features by constructing kernel




Spotlight -CV9
Large-scale knowledge transfer
for object localization in ImageNet.
M. Guillaumin, V. Ferrari.
Proc CVPR 2012

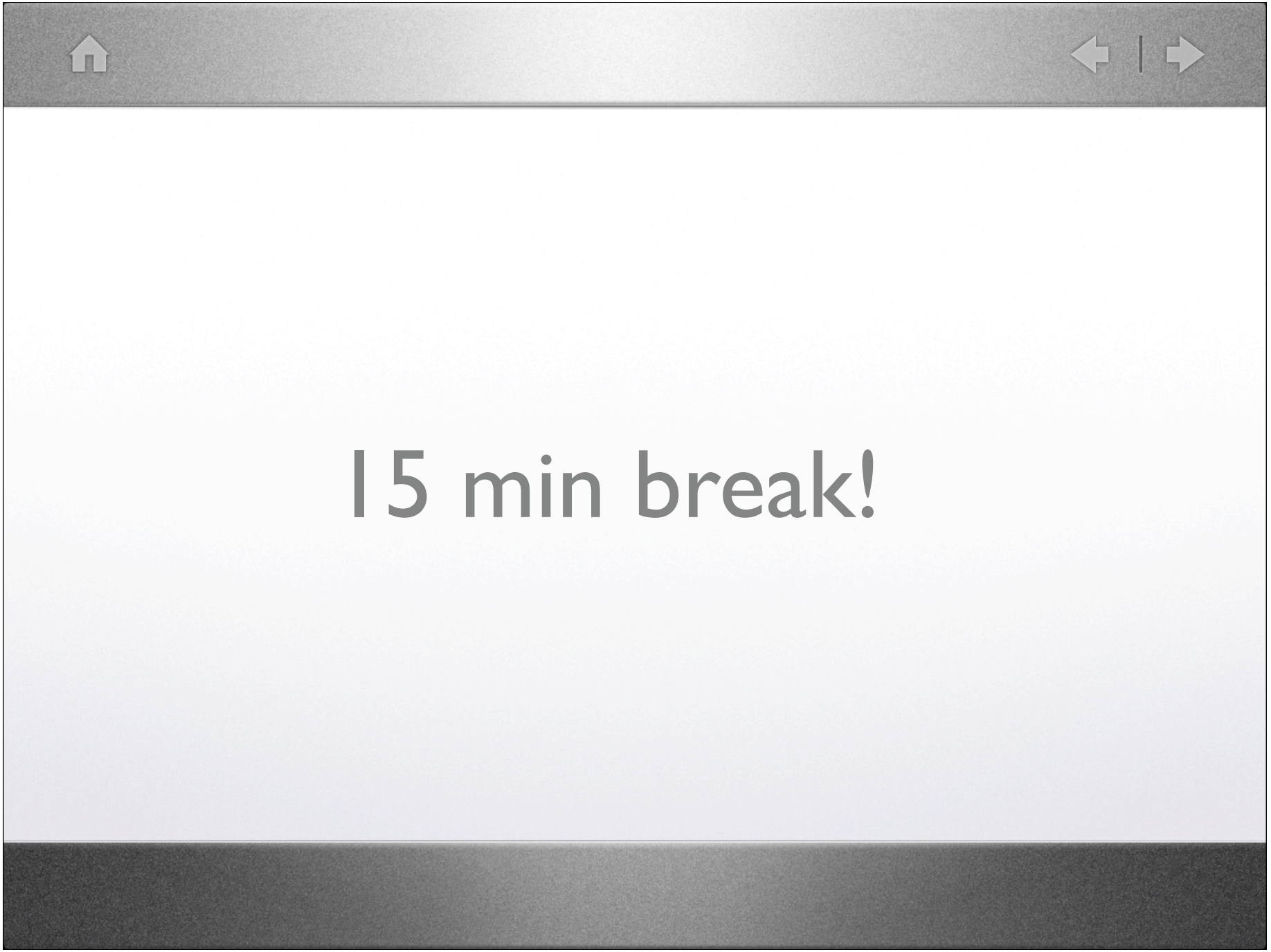

Ancestor (Container)
Ancestor (Wheeled vehicle)
Sibling (Balloon)
Sibling (Airplane)
sec. 2
Source $\mathcal{A}$
location
appearance
context
Source $\mathcal{S}$
location
appearance
context
sec. 3
sec. 4
combine
localize
sec. 5
Target T (Helicopter)



(a) Objectness windows: the three windows with the highest (blue) and lowest (red) objectness sampled in an image of *airliner*.

(b) Context knowledge: background prototypes of siblings are shown surrounded by positive and negative windows.

Knowledge types	Location						Appearance						Context		All types		Sec. 5
	aspect-ratio	scale	x	y	*all subtypes*	*all sources*	SURF-χ^2	Lab-χ^2	HOG	Objectness	*all subtypes*	*all sources*	SURF-χ^2	*all sources*	*one source*	*all sources*	*with target*
Siblings	45.0	**49.8**	43.3	45.1	52.6	**52.7**	**51.2**	36.4	39.0	-	48.6	**53.1**	**49.9**	**50.7**	53.9	**54.1**	**55.2**
Ancestor	44.4	44.5	43.6	45.2	50.8		-	-	-	50.5	50.5		49.8		53.3		

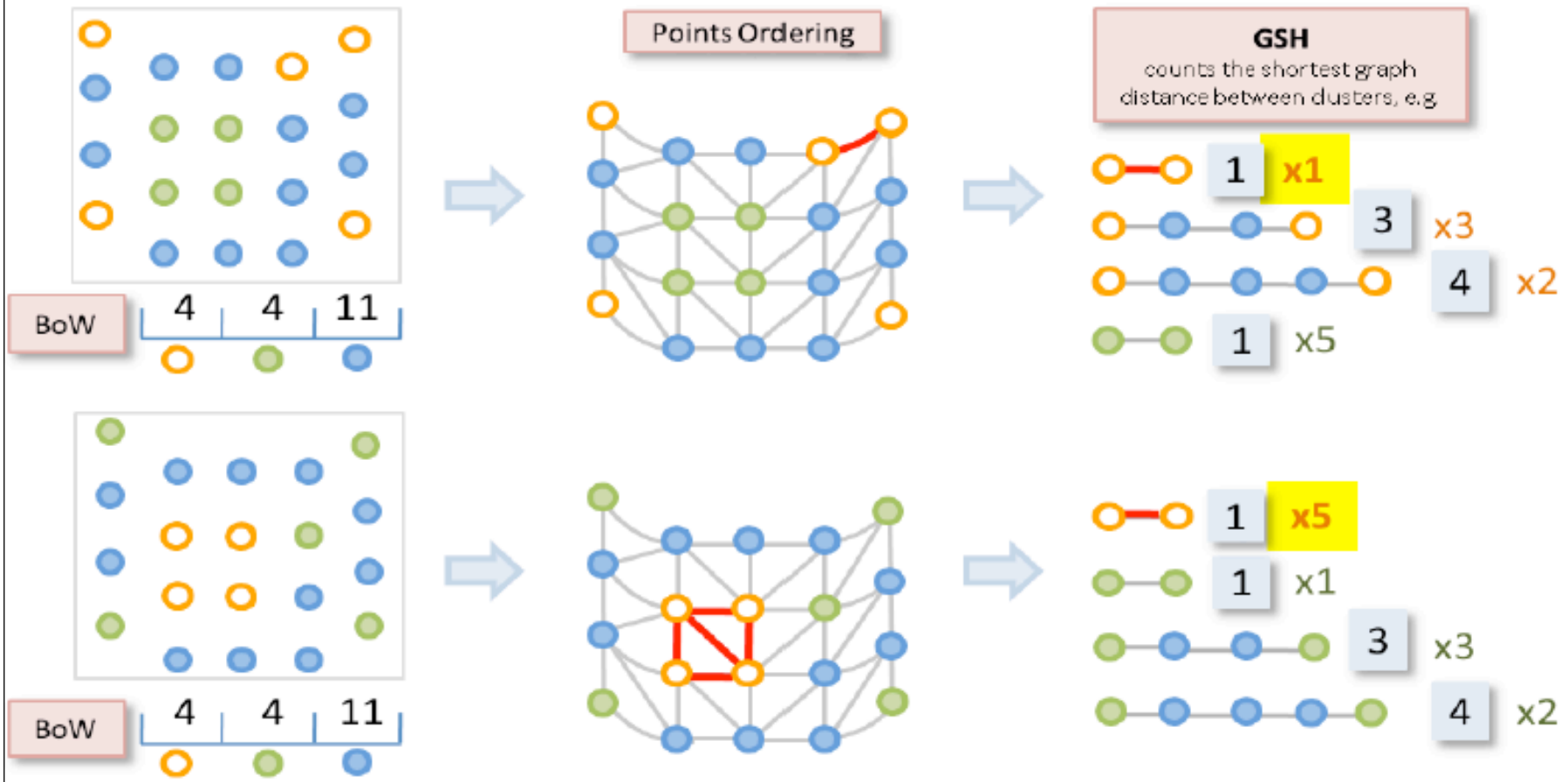

15 min break!

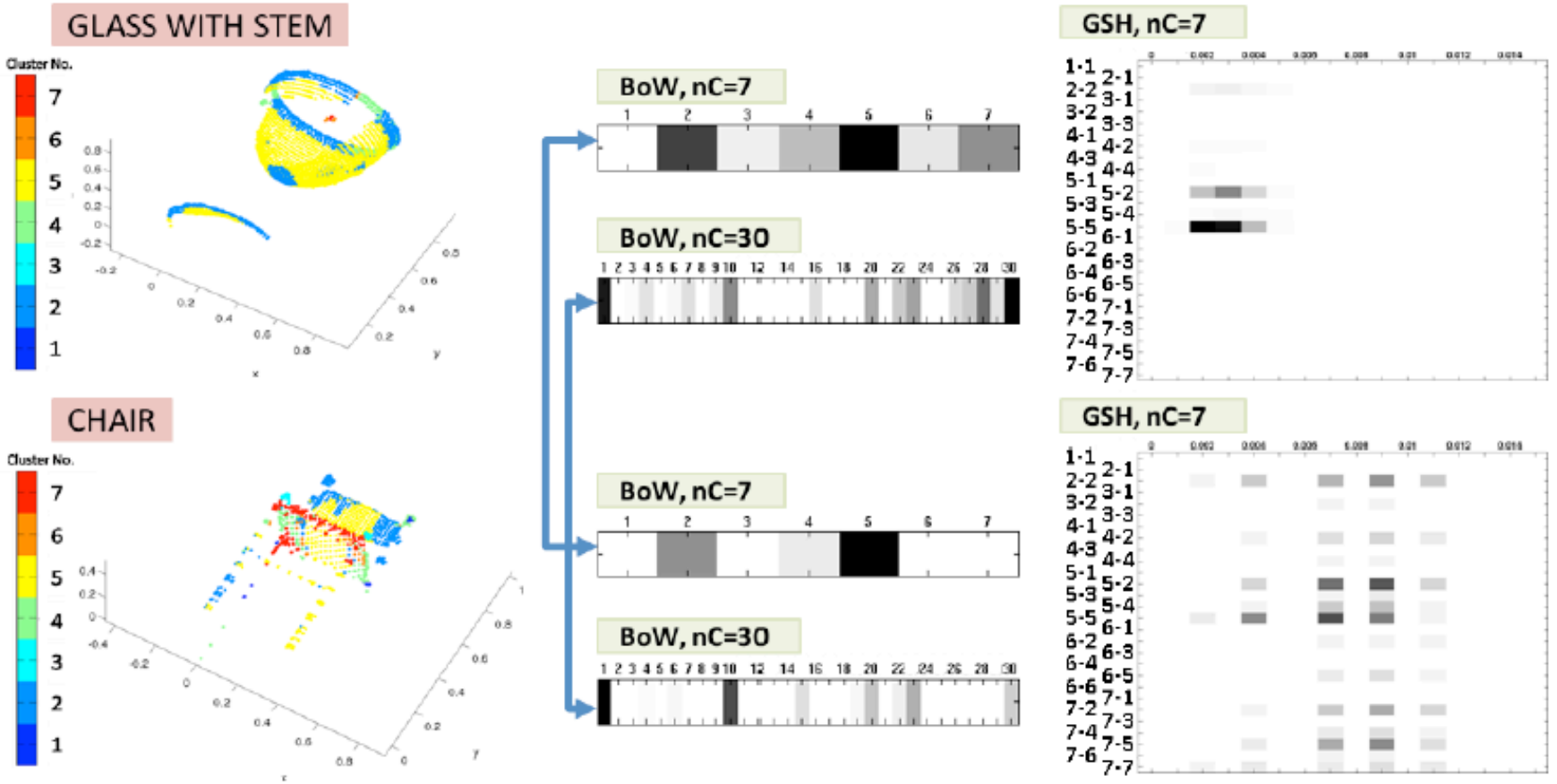

Spotlight presentations:
robot vision

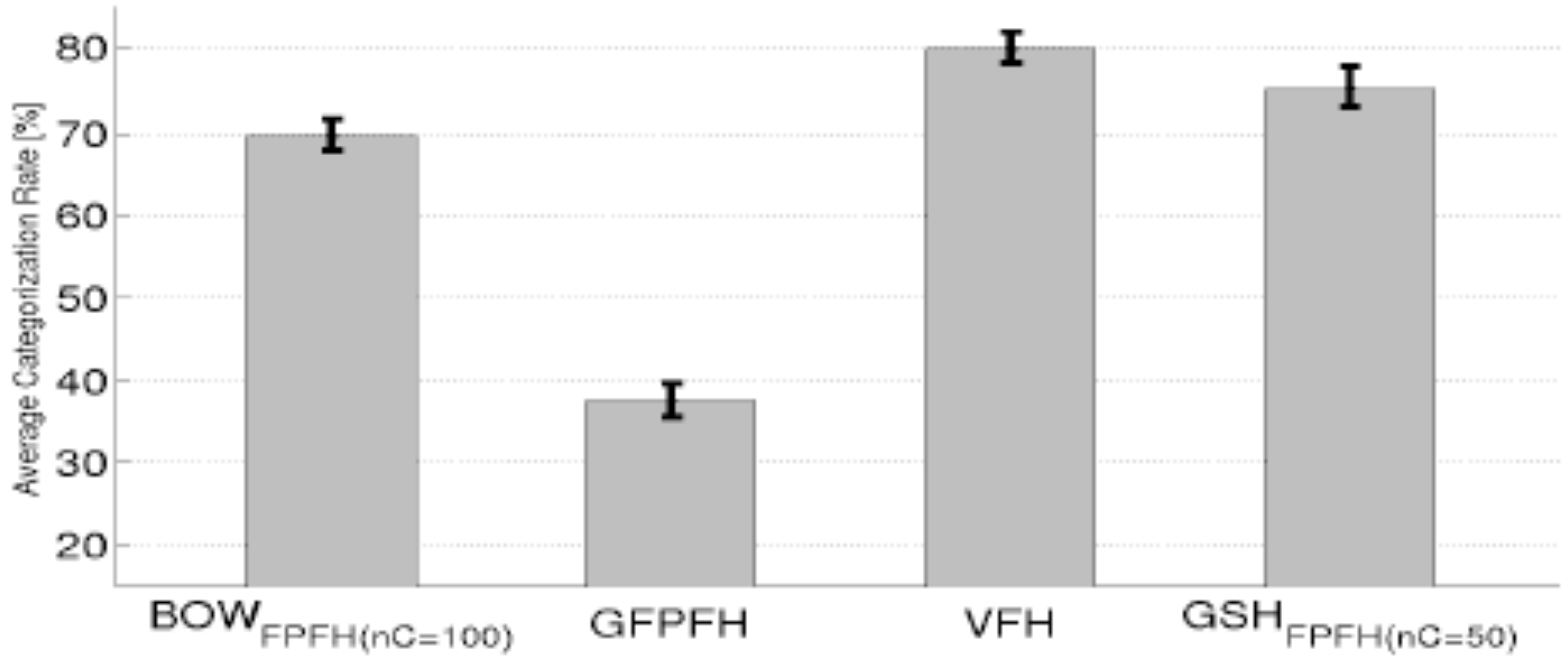

Spotlight -RV1
Improving generalization for 3D object categorization with global structure histograms.

M. Madry, C. H. Ek, R. Deutry,
K. Hang,D. Kragic

Proc IROS 2012


Points Ordering
GSH
counts the shortest graph distance between clusters, e.g.
BoW
4
4
11
1
x1
3
x3
4
x2
1
x5
BoW
4
4
11
1
x5
1
x1
3
x3
4
x2


GLASS WITH STEM
Cluster No.
7
6
5
4
3
2
1
CHAIR
Cluster No.
7
6
5
4
3
2
1
BoW, nC=7
BoW, nC=30
BoW, nC=7
BoW, nC=30
GSH, nC=7
GSH, nC=7



(b) *Results for single objects from SOC database (real stereo data) where training and test data differ in an object rotation. Experimental data protocol is illustrated in Fig. 4(a).*


Spotlight -RV2
Active object recognition on
a humanoid robot.
B. Browatzki, V. Tikhanoff, G. Metta, H.
Bulthoff, C. Wallraven.
Proc ICRA 2012

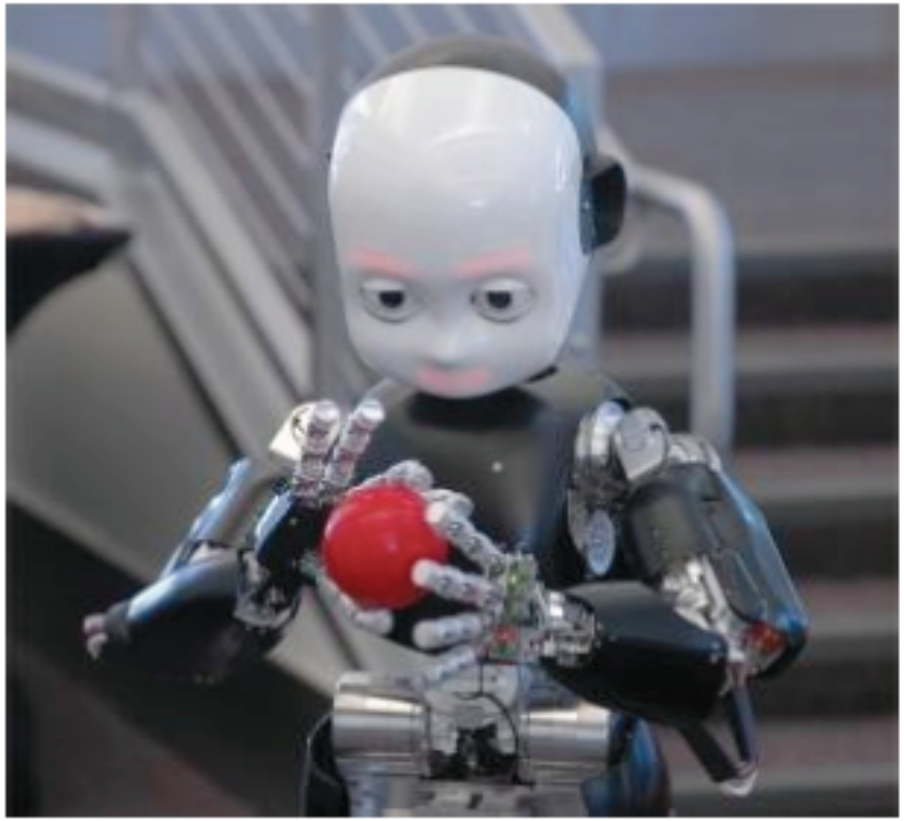


Fig. 1: The iCub humanoid robot. Implementation and evaluation platform of the presented object recognition method.

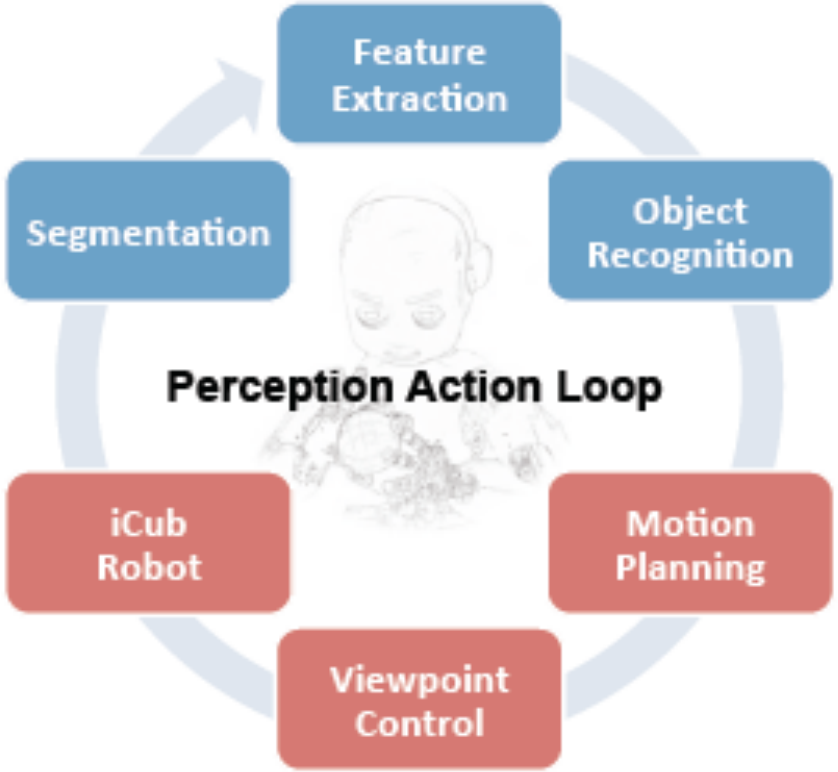


Fig. 2: System components forming a perception action loop.

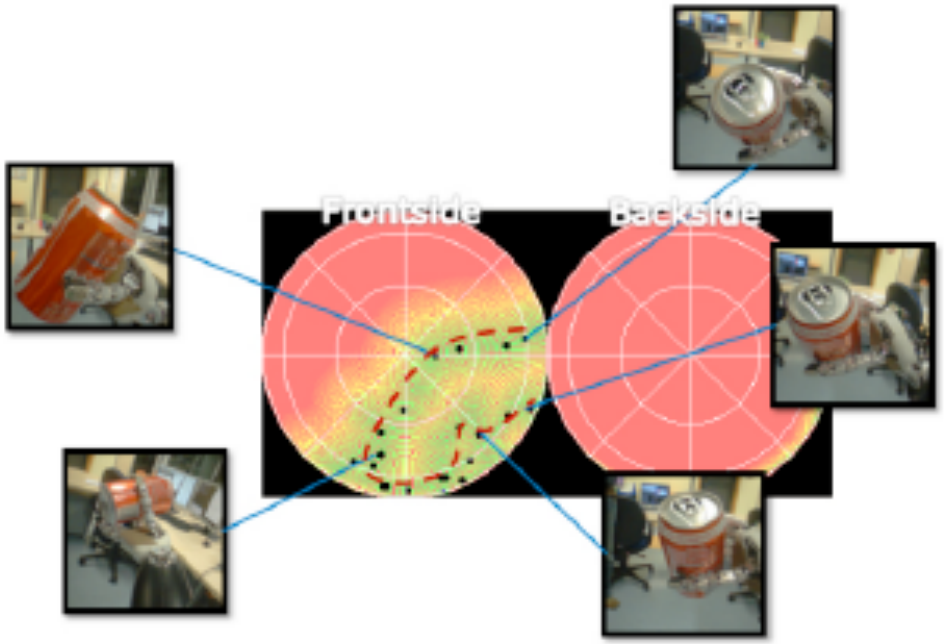
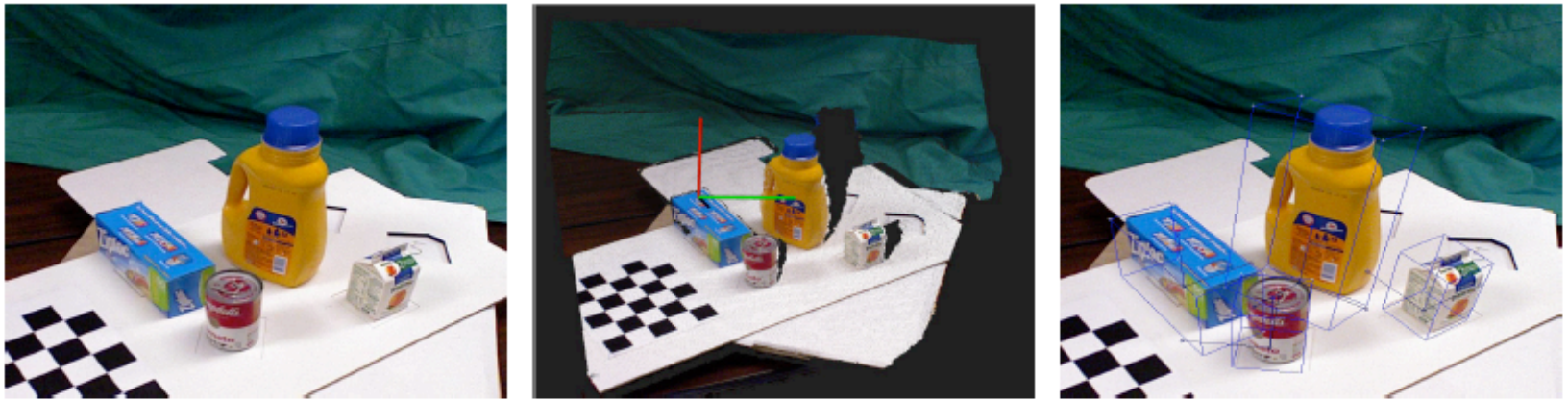


Fig. 4: Example of an exploration sequence executed to learn a new object. The dashed red line shows the exploration path on the view sphere. Keyframes are marked by black dots.


Spotlight -RV3
A textured object recognition pipeline
for color and depth image data.
J. Tang, S. Miller, A. Singh, P. Abbeel.
Proc ICRA 2012



(a) (b) (c)

Fig. 1. (a) Example test image (b) Example test point cloud (c) Sample object detections.

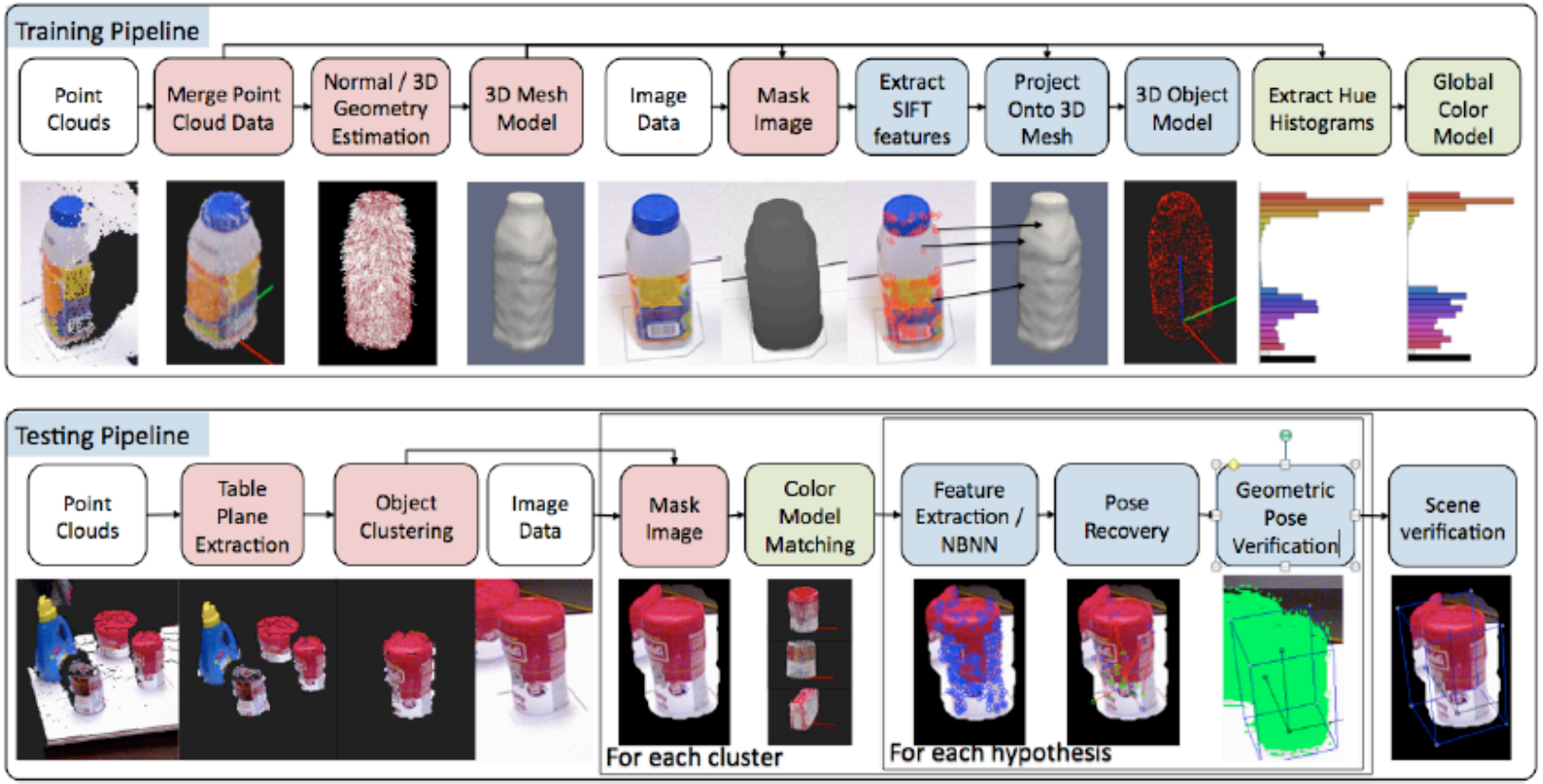
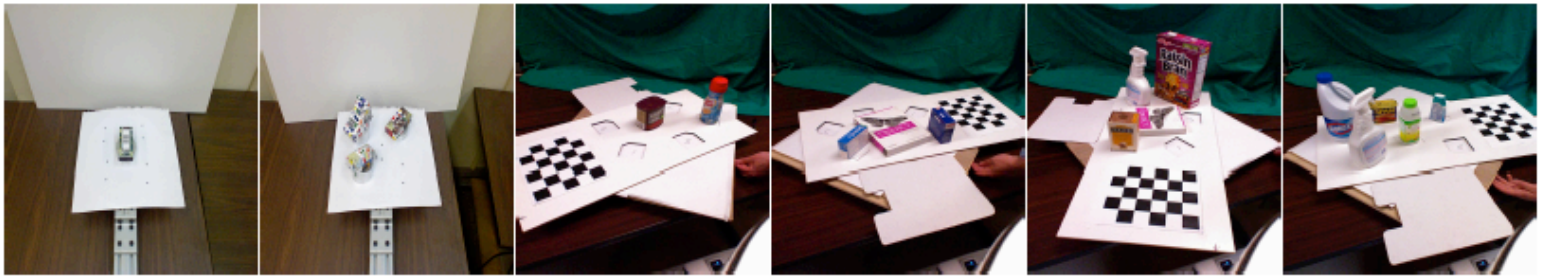


Fig. 2. Overview of our training and testing pipelines. Red boxes correspond to steps which make use of point cloud and depth information. Green boxes correspond to steps which make use of global color information. Blue boxes correspond to steps which make use of local SIFT feature information.

TABLE II

CONFUSION MATRIX FOR THE SINGLE OBJECT INSTANCE RECOGNITION EXPERIMENT. RESULTS ARE REPORTED FOR THE THIRTY-FIVE WILLOW TEST OBJECTS. THE VERTICAL AXIS SHOWS THE TRUE OBJECT LABEL, AND THE HORIZONTAL AXIS SHOWS THE LABEL OUR ALGORITHM REPORTED. MISSING OBJECT LABELS WERE CORRECTLY DETECTED 100% OF THE TIME.

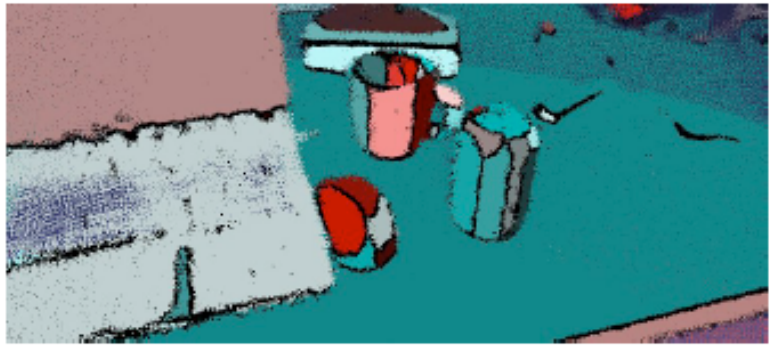
	1	3	4	5	6	7	10	11	13	16	17	20	23	24	25	27	29	30	31	32	33	34	35
1	98.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.4	0	0
3	0	98.6	0	0	1.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	1.4	98.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	98.4	0	0	0	0	0	0	0	0	0	0	0	0	0	1.6
17	0	0	0	2.8	0	0	0	0	0	2.8	87.8	1.9	0	0	0	0	0	1.9	0	0	0	0	2.8
20	0	0	0	0	0	1.7	0	0	0	0	0	93.2	0	0	0	0	0	0	1.7	0	3.4	0	0
24	0	0	0	0	0	1.4	0	0	1.4	0	0	0	0	94.4	0	0	0	1.4	0	0	0	0	1.4
25	0	0	0	0	0	0	0	0	0	0	0	0	2.7	0	97.3	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	1.4	0	0	0	94.3	0	0	0	0	4.3	0
30	0	0	3.3	0	0	0	0	0	3.3	0	0	0	0	0	0	0	0	93.4	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.4	0	0	0	97.2	0	1.4	0
33	1.4	0	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0	0	0	0	0	96.8	0.9	0



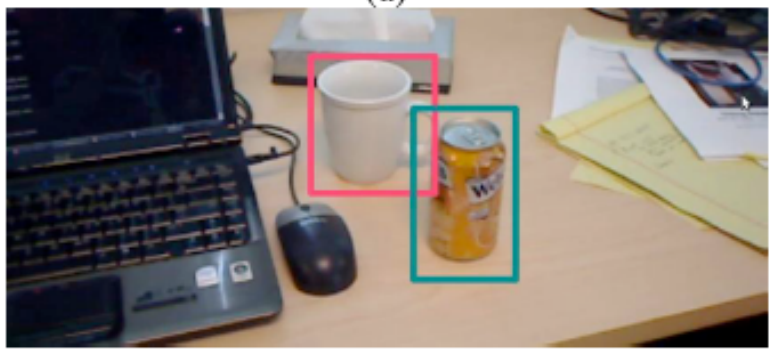
(a) (b) (c) (d) (e) (f)

Fig. 6. Sample test data for the NIST (a), (b) and Willow (c), (d), (e), (f) data sets.


Spotlight -RV4
Detection-based object labeling in 3D scenes.
K. Lai, L. Bo, X. Ren, D. Fox
Proc ICRA 2012.



(a)



(b)

Fig. 1. An illustration of our detection-based approach versus a 3D segmentation approach. (a) A typical approach to 3D labeling segments the scene into components, inevitably causing some objects to be oversegmented. (b) We train object detectors and use them to localize entire objects in individual frames, taking advantage of rich visual information.

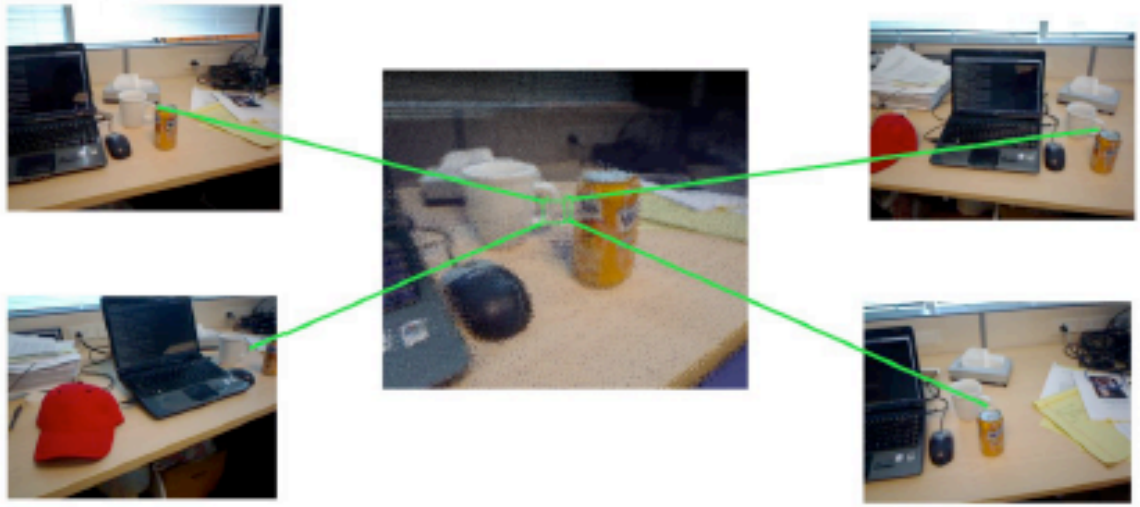


Fig. 2. Each voxel in the scene (center) contains 3D points projected from multiple RGB-D frames. The points and their projections are known, so we can compute average likelihoods for each voxel based on the likelihoods of constituent points. Combining detections from multiple frames greatly improves the robustness of object detection.

Technique	Precision/Recall						
	Bowl	Cap	Cereal Box	Coffee Mug	Soda Can	Background	Overall
DetOnly	46.9/90.7	54.1/90.5	76.1/90.7	42.7/74.1	51.6/87.4	98.8/93.9	61.7/87.9
PottsMRF	84.4/90.7	74.8/91.8	88.8/94.1	87.2/73.4	87.6/81.9	99.0/98.3	86.9/88.4
ColMRF	93.7/86.9	81.3/92.2	91.2/89.0	88.3/73.6	83.5/86.5	98.7/98.8	89.4/87.8
Det3DMRF	91.5/85.1	90.5/91.4	93.6/94.9	90.0/75.1	81.5/87.4	99.0/99.1	91.0/88.8

Fig. 5. Per-category and overall (macro-averaged across categories) precisions and recalls for the proposed detection-based 3D scene labeling approach and its variants. Our approach works very well for all object categories in the RGB-D Scene Dataset.

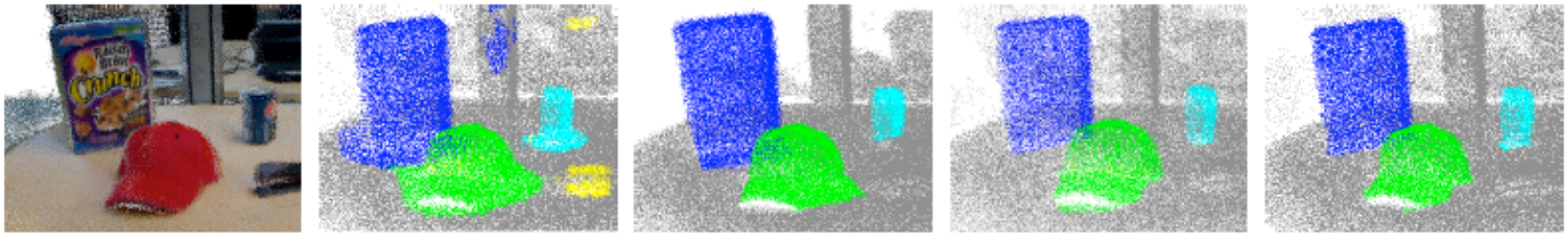
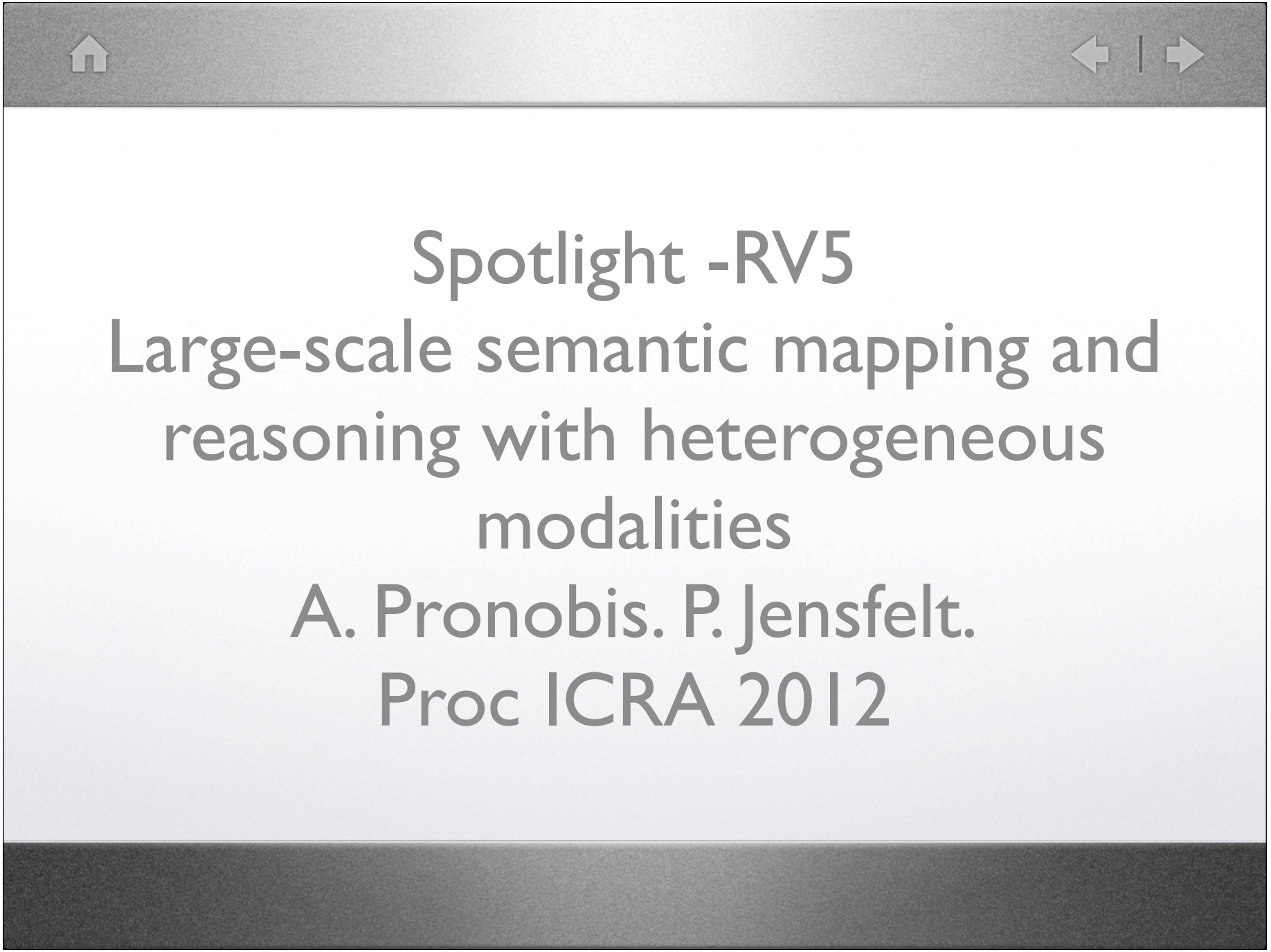


Fig. 6. Close-up of a scene containing a cap (green), a cereal box (blue), and a soda can (cyan). From left to right, the 3D scene; Detection-only leaves patches of false positives; Potts MRF removes isolated patches but cannot cleanly segment objects from the table; Color MRF includes part of the table with the cap because it is similar in color due to shadows; the proposed detection-based scene labeling obtains clean segmentations. Best viewed in color.


Spotlight -RV5
Large-scale semantic mapping and reasoning with heterogeneous modalities
A. Pronobis. P. Jensfelt.
Proc ICRA 2012

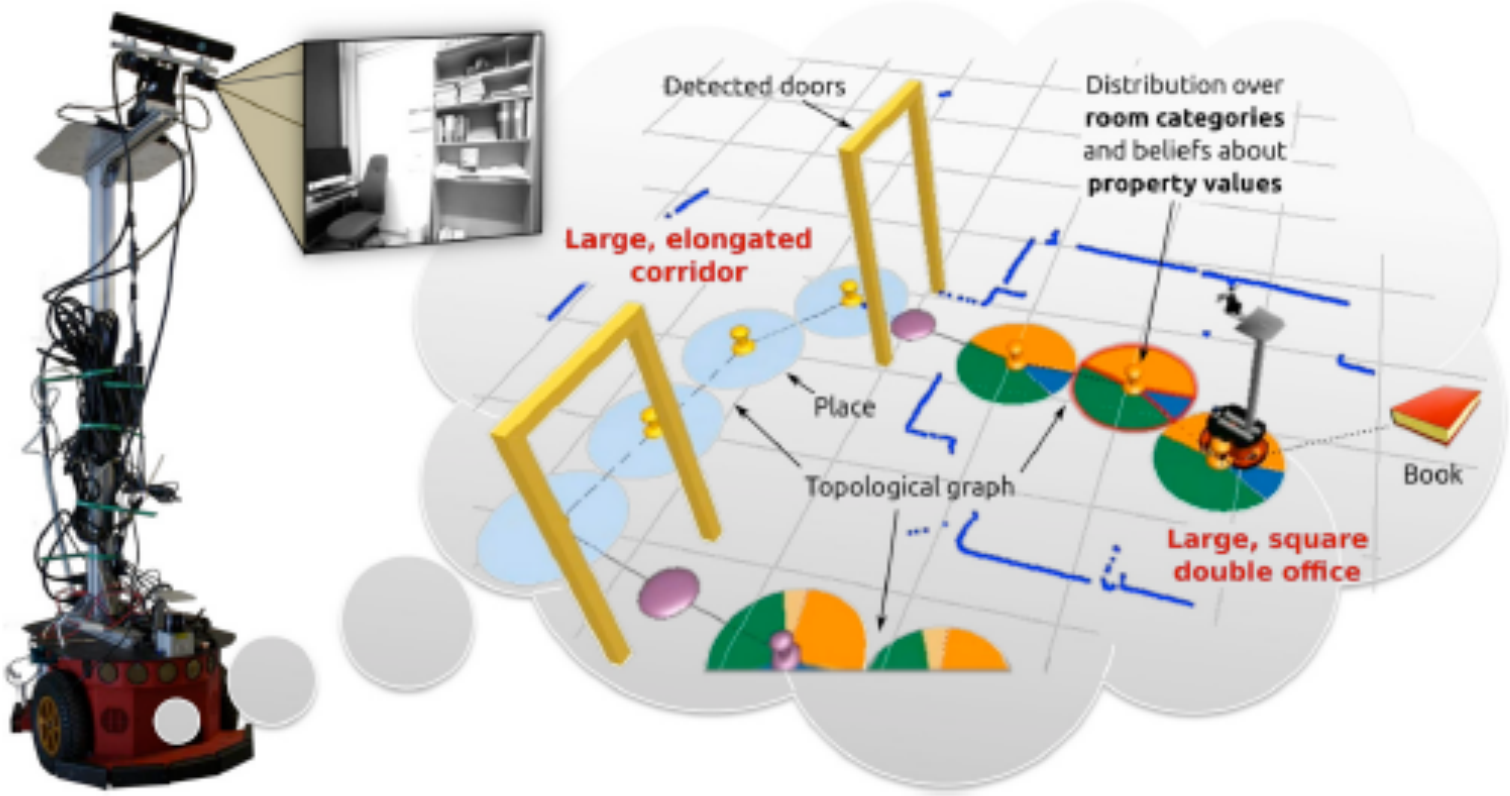


Fig. 1. Our robot platform and an illustration of a semantic map.

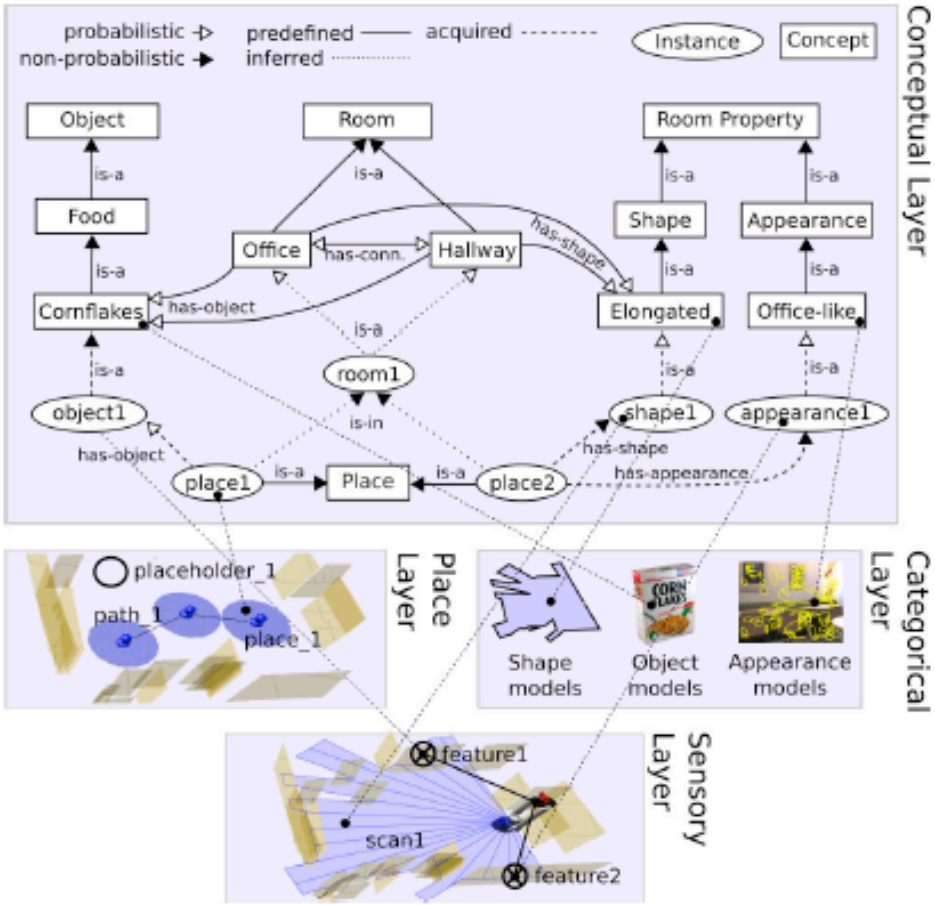


Fig. 2. The layered structure of the spatial representation and a visualization of an excerpt of the ontology of the conceptual layer. The conceptual layer comprises knowledge about concepts (rectangles), relations between those concepts and instances of spatial entities (ellipses).

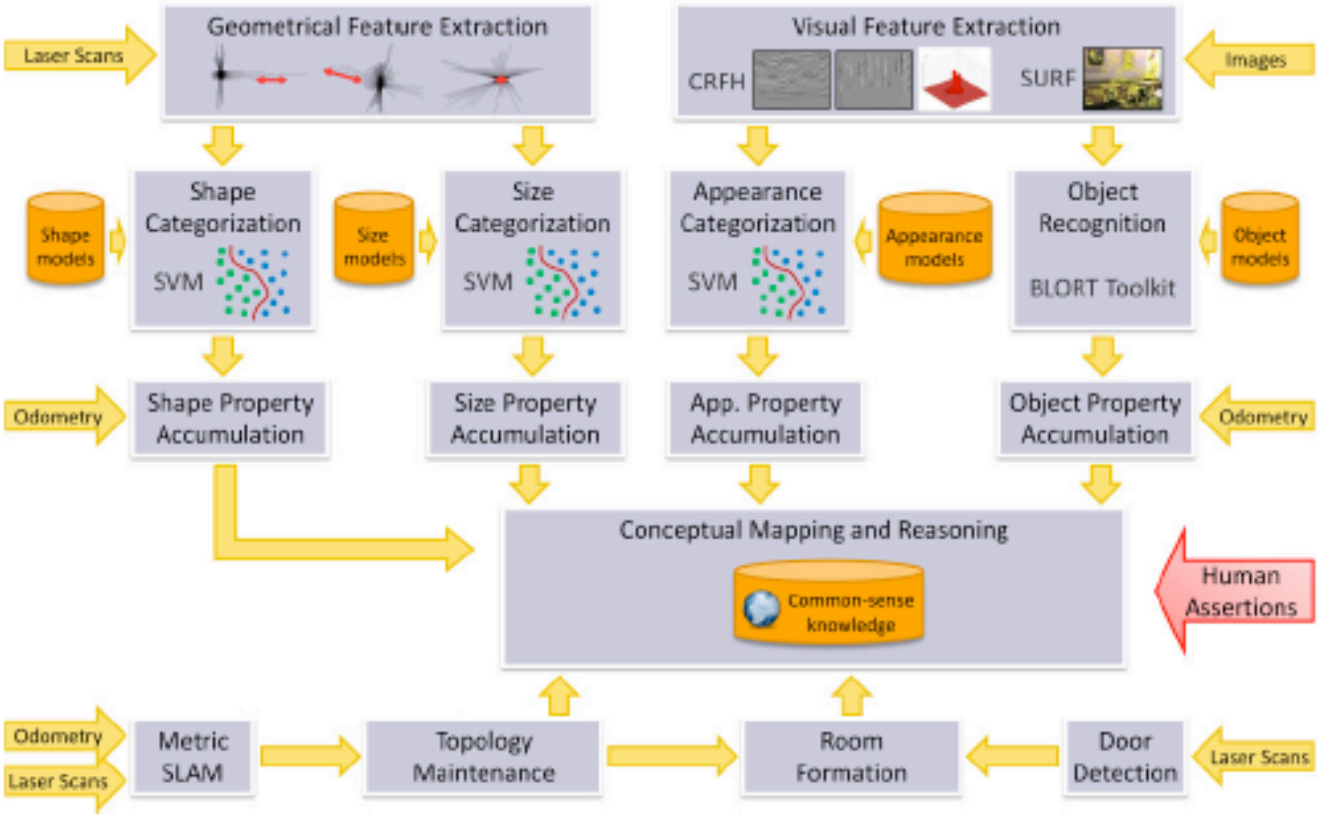


Fig. 3. Structure of the system and data flow between its main components.


Spotlight -RV6
Semantic object maps for robotics housework --representation, acquisition and use.
D. Pangercic, B. Pitzer, M. Tenorth, M. Beetz.
Proc ICRA 2012

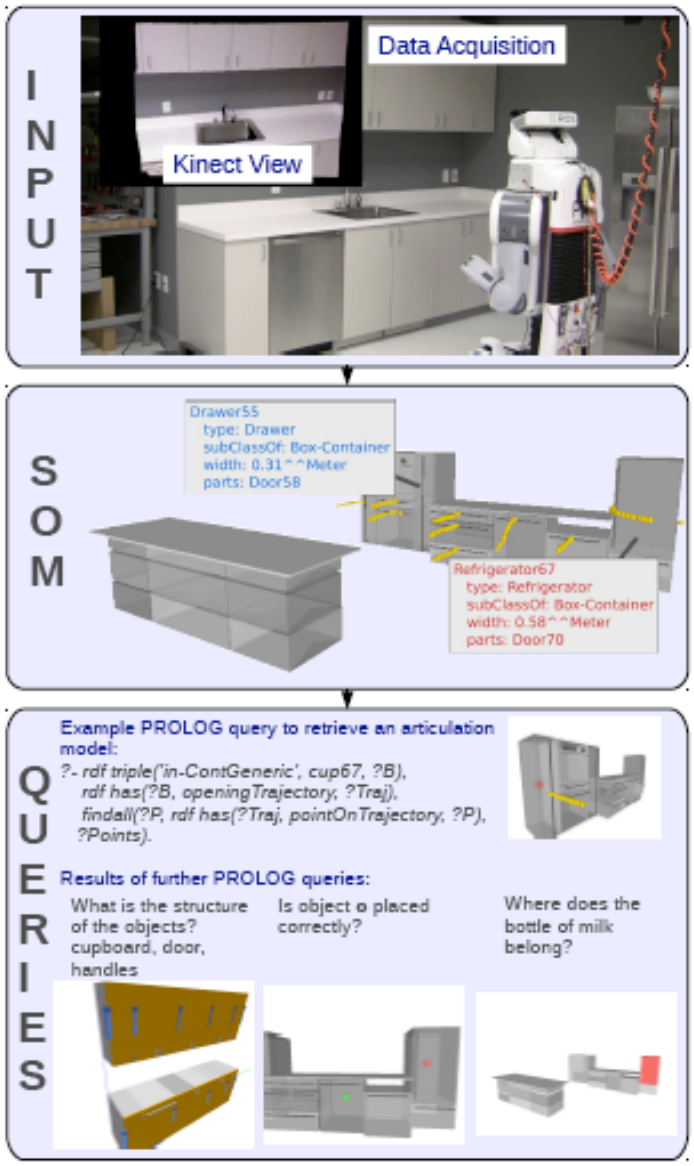


Fig. 1. Building of a SOM$^+$ map in a kitchen environment (top), SOM$^+$ map representation (middle) and a set of robot queries made possible due to such powerful representation (bottom).

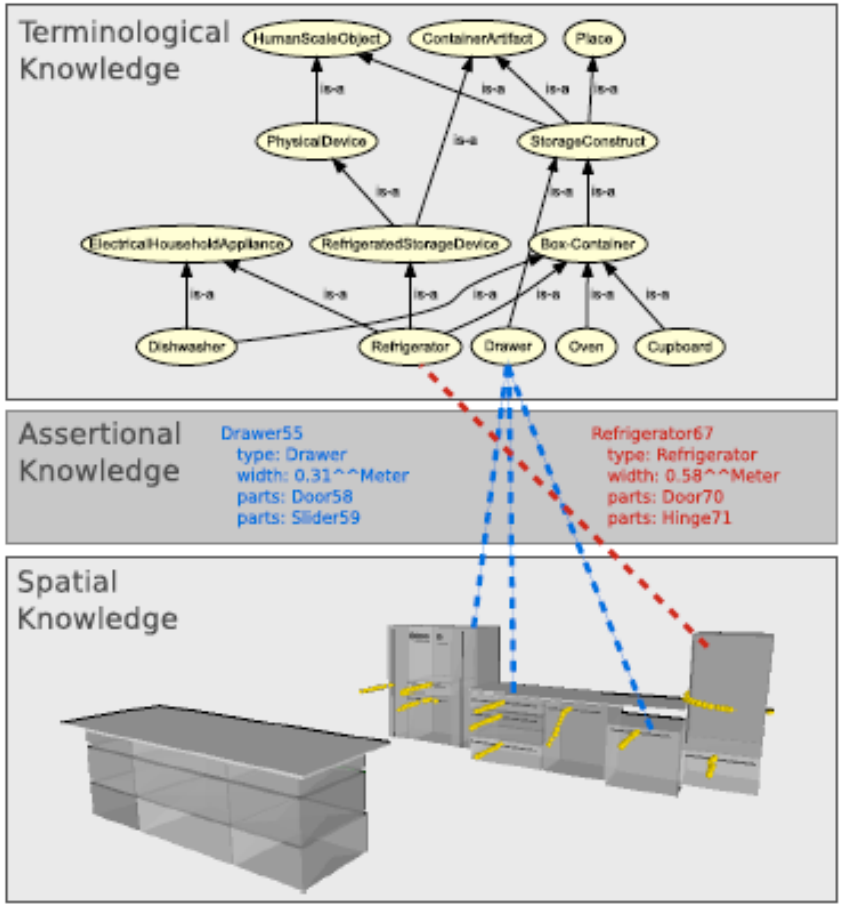


Fig. 2. Part of the ontology of household appliances and entities of furniture. Super-classes of e.g. *HumanScaleObject* have been omitted for better readability.

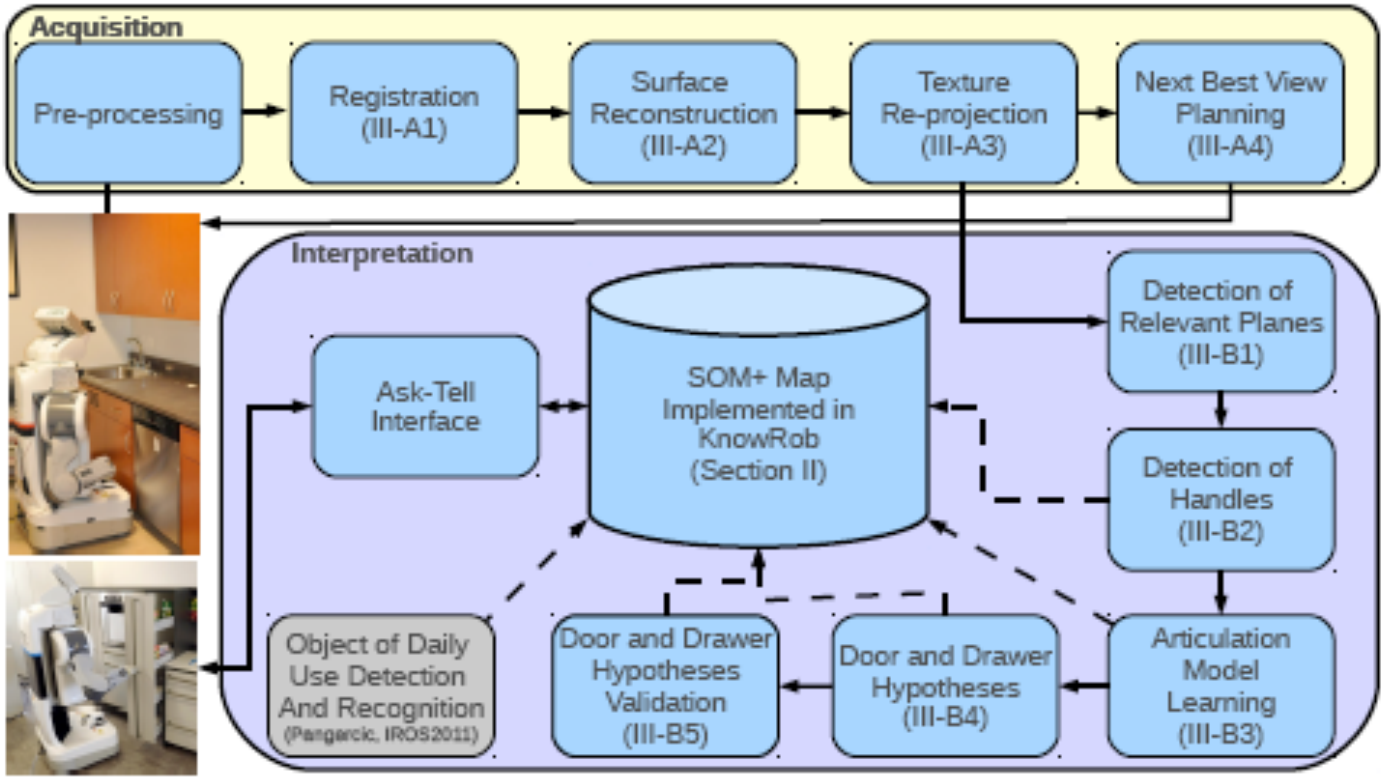


Fig. 4. System integration described in Sec. III. Module for objects of daily use detection and recognition is part of the system but not discussed in this paper due to space constraints.


Spotlight -RV7
The RoboEarth language: representing
and exchanging knowledge about actions
objects and environments.
M. Tenorth, A. Perzylo, R. Lafrenz, M.
Beetz.
Proc IROS 2012

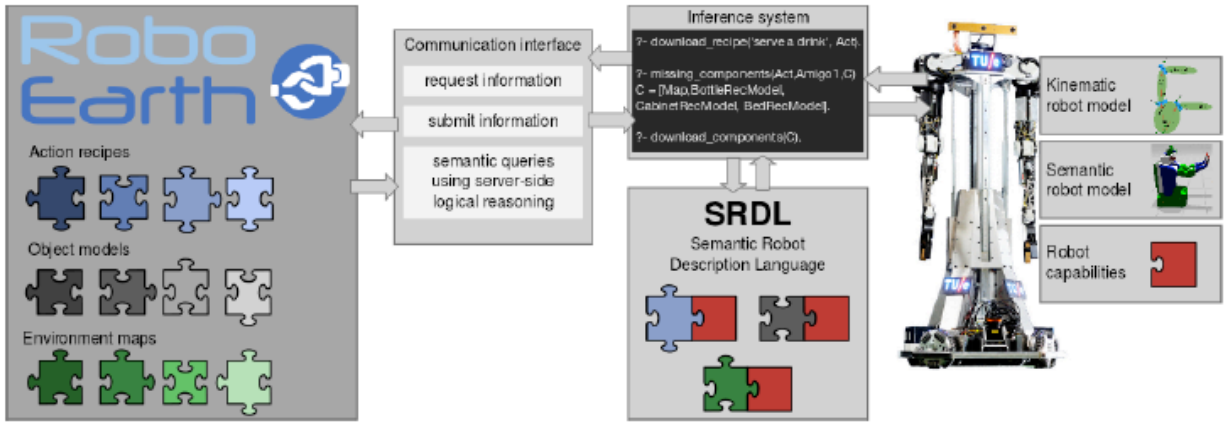


Fig. 2. Overview of the RoboEarth system: A central database provides information about actions, objects, and environments. The robot can up- and download information and determine if it can use it based on a semantic model of its own capabilities.

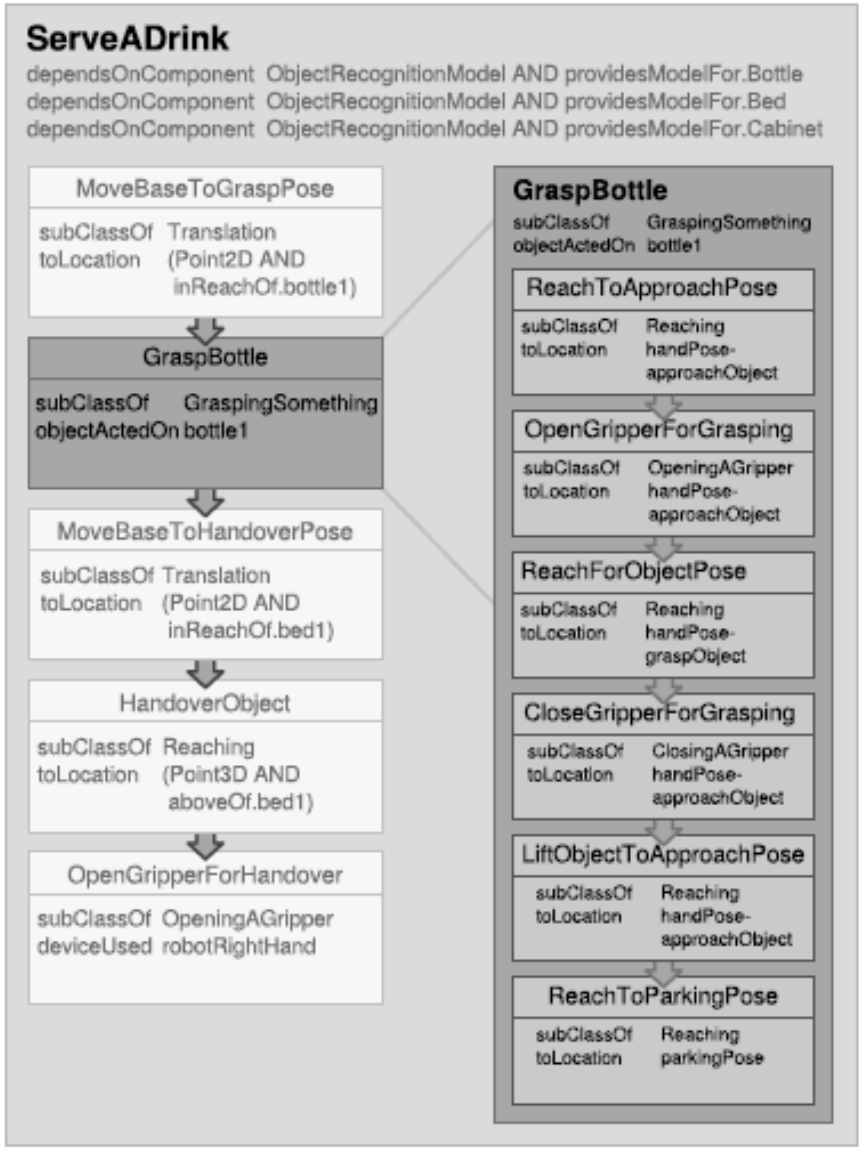


Fig. 4. Representation of a "serving a drink" task, called "action recipe" in the RoboEarth terminology, which is composed of five sub-actions that themselves can be described by another action recipe.

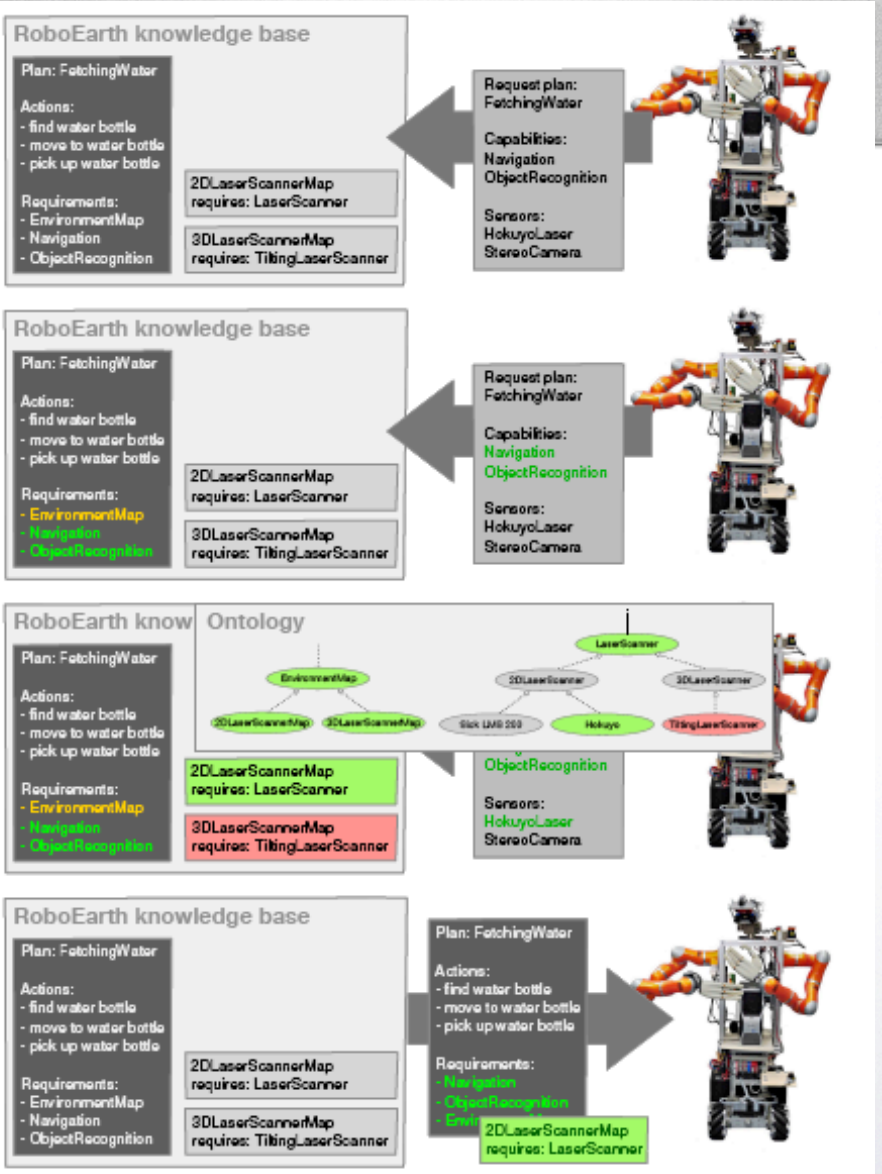


Fig. 7. Matching requirements of action recipes against robot capabilities to determine which further information is still missing and has to be downloaded. The matching becomes more flexible by taking the robot's knowledge into account and selecting the *2DLaserScannerMap* to fulfill the general requirement on an *EnvironmentMap*.


thanks!