



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





Useful Info

- **56 hours course** (28 teaching, 28 laboratory)
- **4 credits**
- **Topics**
 - *Scene Recognition and Understanding*
 - *Object Recognition and Categorization*
 - *Action Recognition and Understanding*
 - *Life Long Learning of Concepts*



Useful Info

- **web-page course:**<http://www.idiap.ch/ftp/courses/EE-700/CogVisCogSys.html>
- **how to reach me/Marco: email**
{bcaputo,mfornoni}@idiap.ch
- **Exam:**
 - *Report on laboratory experiences, with discussion*
 - *Oral presentation of research paper*
 - *Date: ?????*



● **Exam: Report on laboratory experiences**

- *For each topic, there will be a corresponding laboratory experience*
- *It will consist of replicating the experiments of a seminal paper in the field, on the same data presented in the paper and on different data collections (mandatory)*
- *For the mandatory part of the work, we provide software and data, you develop the tools for the analysis of the experimental results*



● **Exam: Report on laboratory experiences**

- *Optional: more exciting, research-like stuff (will require some coding)*
- *Once all the experiences are done, you write a report with one chapter for each experience, and you send it to bcaputo@idiap.ch*
- *Minimum for passing the exam: all experiences done and well reported, plus at least for one experience some optional work done*
- *No special requirements on length, template, etc*
- *To be submitted at the very latest 15 days before the day of the exam!!*



● **Exam: Oral Presentation of Research Paper**

- *For each topic, I will present the most recent trends in the research field, i.e. papers presented during the last 6-9 months at the top conferences in the field (acceptance rate 40-20%)*
- *Between the papers presented in this lecture, you pick one by sending me an email (first come, first serve)*
- *The day of the exam you make a 30m presentation of the paper, putting it into the context of what was discussed during lectures*
- *Exam consists of: (1) doing lab experiences and reporting on them (2) discussion of the lab experience report (3) 30m presentation of paper chosen by you*



Basic Definitions

- **Computer Vision:** *the science and technology of machines that see [...] concerned with the theory for building artificial systems that obtain information from images (video sequences, views from multiple cameras, multi-dimensional data from a medical scanner)*

source: wikipedia



..that obtain information from images..

- *Place Recognition*





..that obtain information from images..

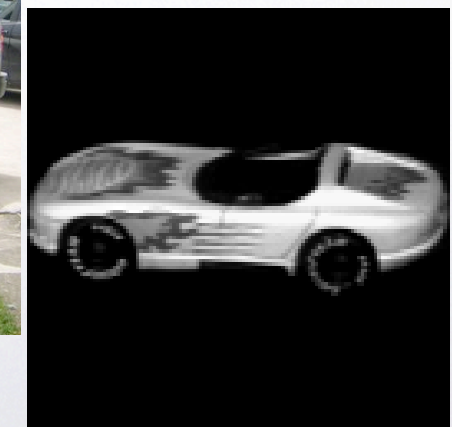
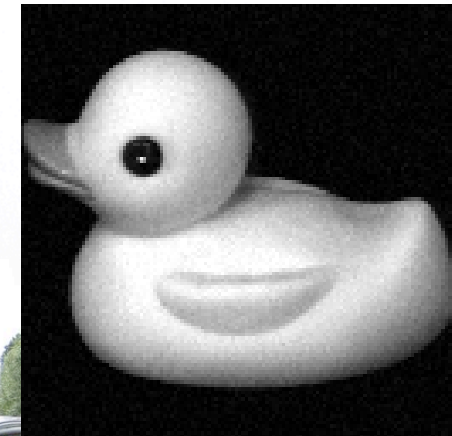
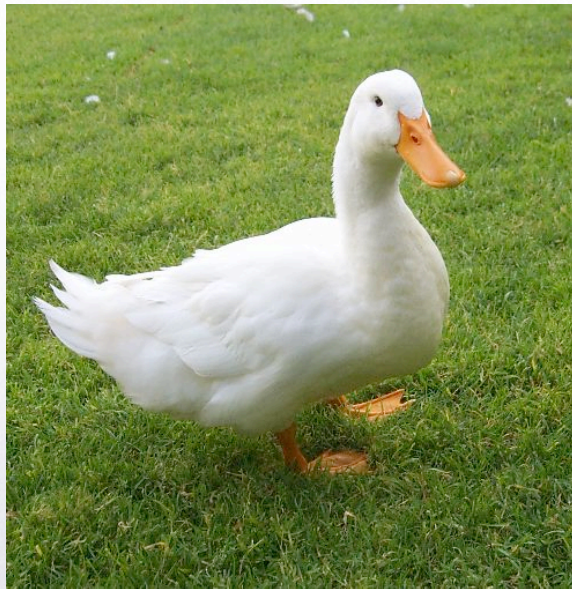
- *Scene Understanding*





..that obtain information from images..

- *Object Recognition*





..that obtain information from images..

- *Object Categorization*





..that obtain information from images..

- *Action Recognition*





Basic Definitions

- **Cognitive Systems:** *systems that have cognitive functions normally associated with people or animals and which exhibit a high degree of robustness in coping with unpredictable situations [...] act purposefully and autonomously towards achieving goals*



..act purposefully towards a goal..

- *Go to the kitchen*

Current position: Printer area

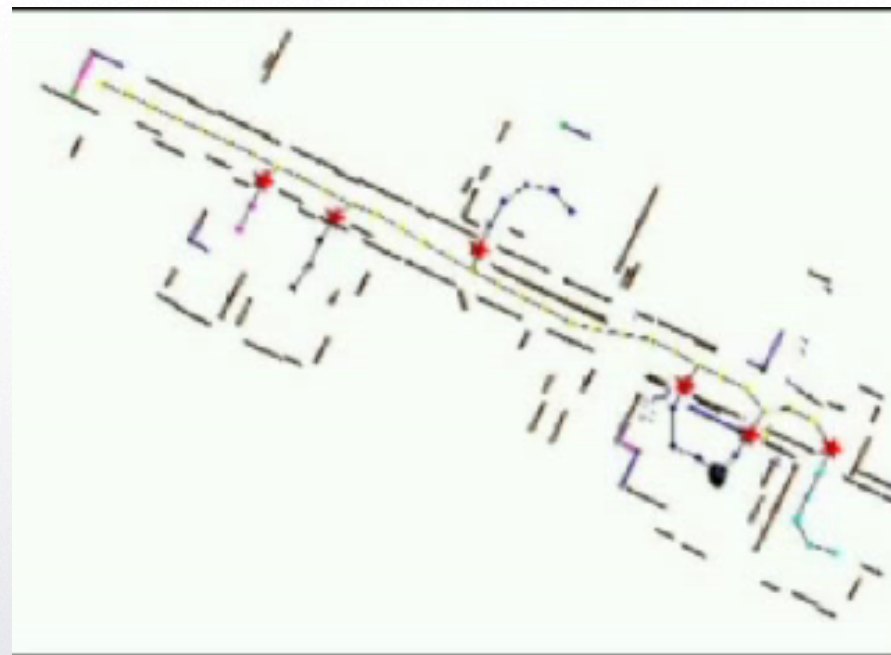
Color codes indicating the recognition results:

■ O-p office	■ Kitchen	■ Printer
■ T-p office	■ Corridor	



..act purposefully towards a goal..

- *Find the cereal box in the living room*





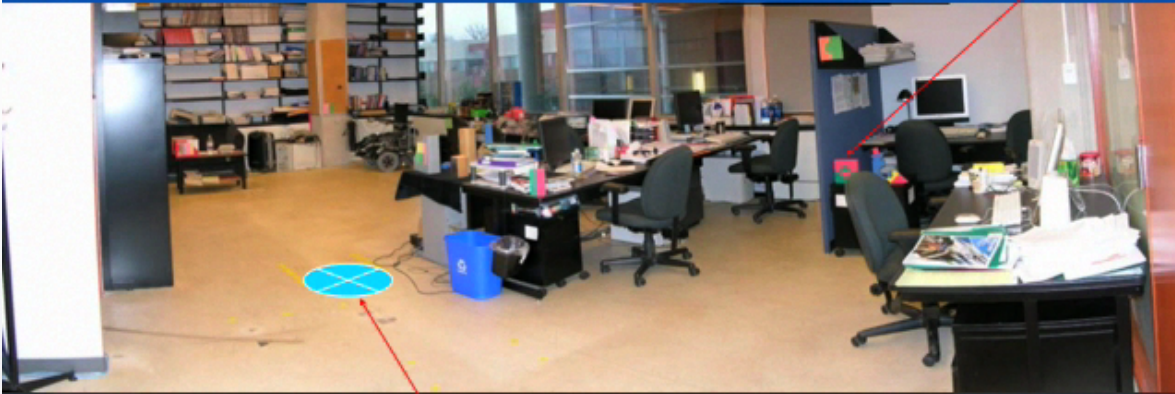
..act purposefully towards a goal..

- *Find my cup*

Simple Example of the Shubina-Ye-Tsotsos (SYT) Strategy

search space is rectangular, 20' x 30'

target



robot starting position, eyes face bookshelves

February 2008



..act purposefully towards a goal..

- *Where is the cup?*



Learn

Attend

Find

Clear Inhibitions

Use Context

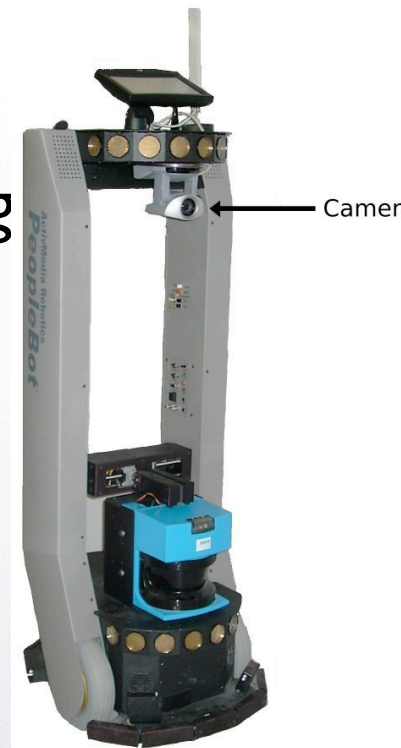
Auto Sift

Object	# matches



the key link

- *Place Recognition*
- *Scene Understanding*
- *Object Recognition*
- *Object Categorization*
- *Object Localization*



- *Go to the kitchen*
- *Find the cereal box in the living room*
- *Find my cup*
- *Find a cup*
- *where is my cup?*



Robots...





Robot Vision





Robot Vision





Robot Vision vs Cognitive Vision



- robot vision needs: 3D description of objects, compact representation of places, fast localization of objects in scenes
- cognitive vision needs: scene understanding in space and time, continuous learning of object categories, attention mechanism for scene interpretation



Wrapping Up

- **Our Working Definition of Cognitive Vision:**
theory and algorithms for building autonomous artificial systems that are able to obtain information from, and to understand, visual data in space and time (video sequences, views from multiple cameras)



Some Reading

- *Unifying perspectives in computational and robot vision.* D. Kragic, V. Kyrki, *Lecture Notes in Electrical Engineering*, Springer
- *How the body shapes the way we think.* R. Pfeifer, J. C. Bongard. *MIT Press*, 2007.
- *Active vision.* A. Blake, A. Yuille. *MIT Press*, 1992
- *Dynamic vision for perception and control of motion.* E. D. Dickmanns, Springer, 2007.



15 min break!



Scene Recognition



What do you see?





What do you see?





Some useful thoughts

- We easily (= quickly) distinguish between indoor and outdoor scenes





Some useful thoughts

- We are able to identify easily (= quickly) few landmark objects in a scene





Some useful thoughts

- We expect to find some objects only in certain parts of the scene





Human visual perception

- **What do we remember and what do we forget when we recall a scene?**
 - ***WE DO REMEMBER:*** the gist of a scene, 4-5- landmark objects and their spatial configuration
 - ***WE DO NOT REMEMBER:*** all the objects in the scene, mid- to fine details

J. M. Wolfe. *Visual memory: what do you know about what you saw?*
Current Biology, 1998, 8: R303-R304



Computer Vision

- Most of work on **outdoor** place recognition, only recently (2009) attention shifted on indoor place recognition
- Gist of a scene = holistic representation
- Applications: image retrieval, context priming

A. Oliva, A. Torralba. *Modeling the shape of the scene: a holistic representation of the spatial envelope*. International Journal of Computer Vision, 42(3), 145-175, 2001



Holistic Scene Recognition

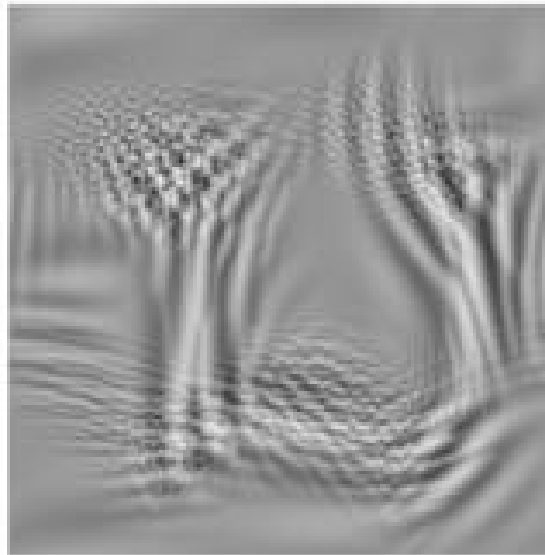
- **Key idea:** to represent the dominant spatial structure of a scene with a global low dimension representation





Holistic Scene Recognition

- **Key idea:** to represent the dominant spatial structure of a scene with a global low dimension representation

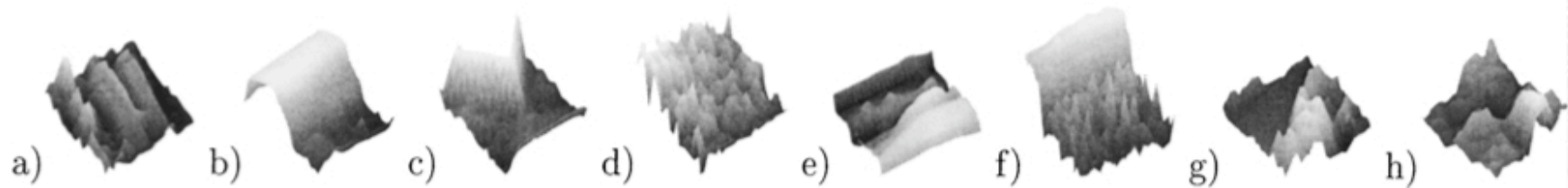


Computationally, it translates into using spectral and coarsely localized information



Holistic Scene Recognition

- **Concretely:** look at a scene as an individual object, with a unitary shape

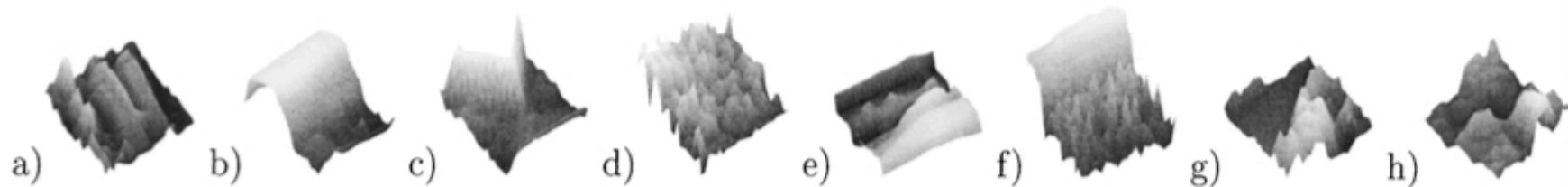


Scenes belonging to the same category share a similar, stable spatial structure that can be extracted with an intermediate, global representation



Holistic Scene Recognition

- **Spatial envelope:** a composite set of boundaries -like walls, sections, ground, elevation -that define the shape of a scene. It is represented by the relationship between the outlines of the surfaces and their properties including the inner textured pattern generated by windows, trees, cars, etc.





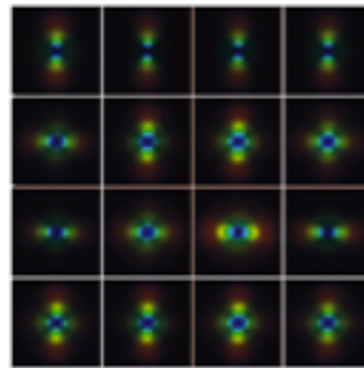
- **Five spatial envelope properties:**

- *Degree of Naturalness*: scenes having a distribution of edges biased toward vertical and horizontal orientations would have a low degree of naturalness, and vice-versa
- *Degree of Openness*: the existence of a horizon line and the lack of visual references confer to the scene a high degree of openness
- *Degree of Roughness*: it is correlated with the fractal dimension of the scene and thus its complexity
- *Degree of Expansion*: the convergence of parallel lines gives the perception of the depth gradient of the space
- *Degree of Ruggedness*: it refers to deviation of the ground w.r.t the horizon --mostly natural

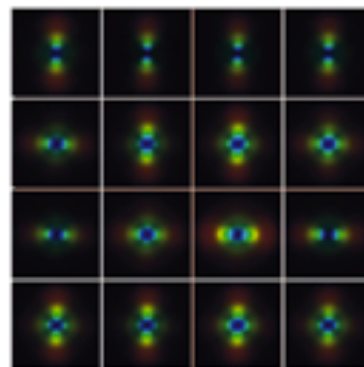


Gist descriptor

Oliva and Torralba, 2001



- Apply oriented Gabor filters over different scales
- Average filter energy in each bin



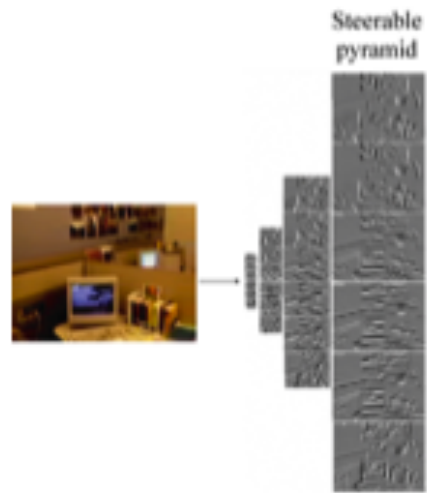
8 orientations
4 scales
x 16 bins
512 dimensions

Similar to SIFT (Lowe 1999) applied to the entire image

M. Gorkani, R. Picard, ICPR 1994; Walker, Malik. Vision Research 2004; Vogel et al. 2004;
Fei-Fei and Perona, CVPR 2005; S. Lazebnik, et al, CVPR 2006; ...

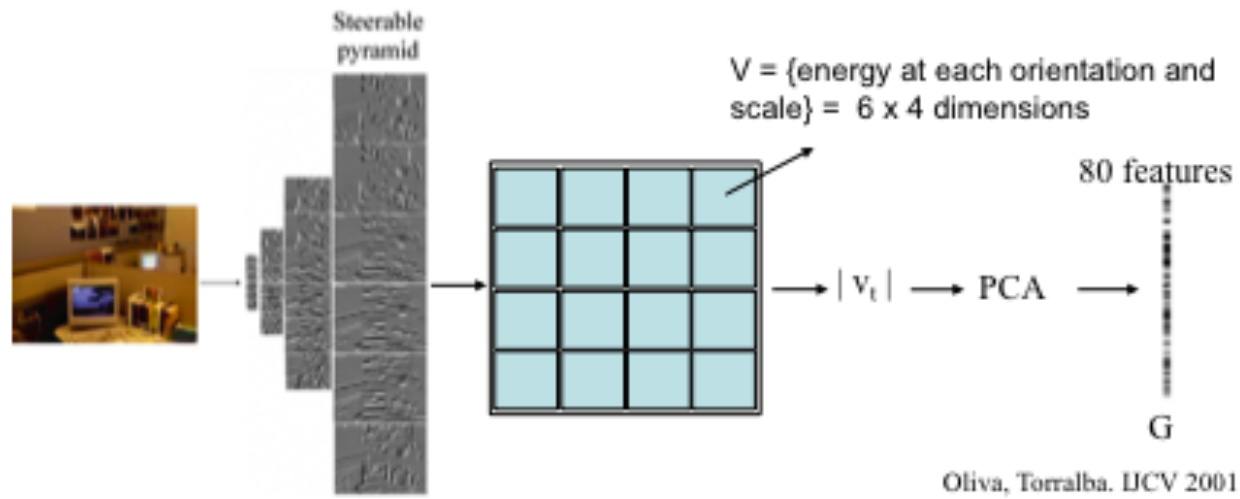


Gist descriptor



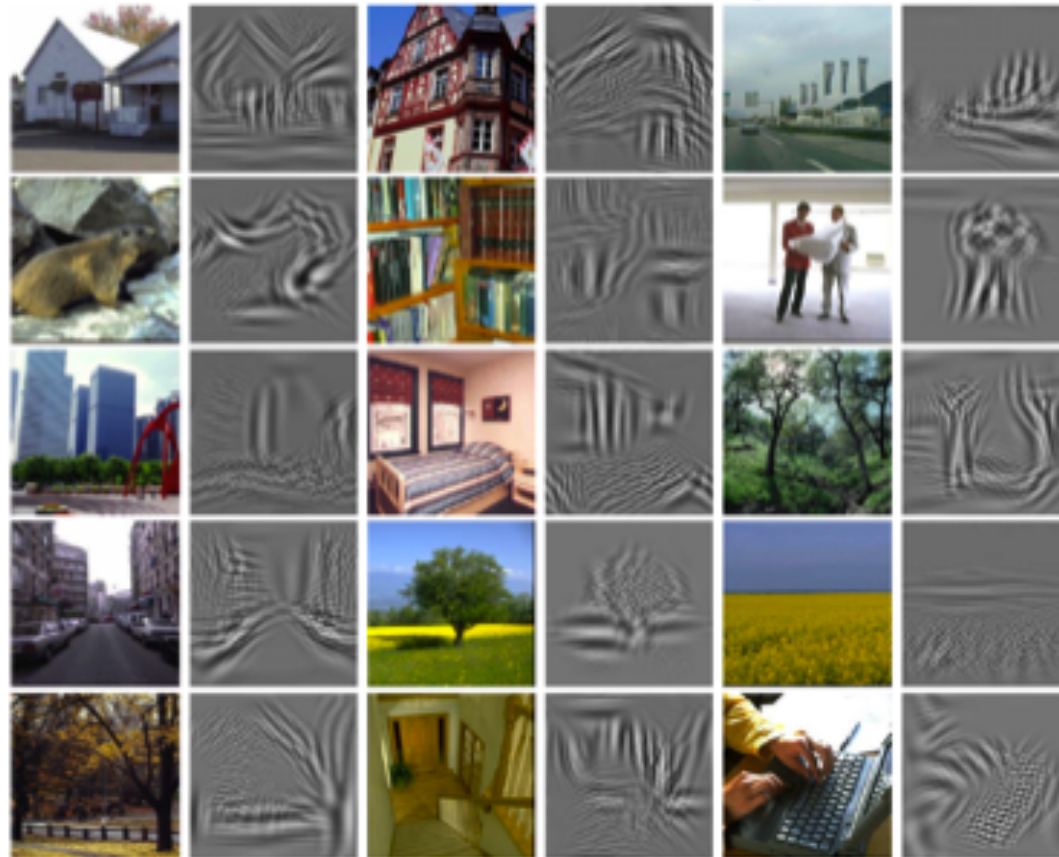


Gist descriptor





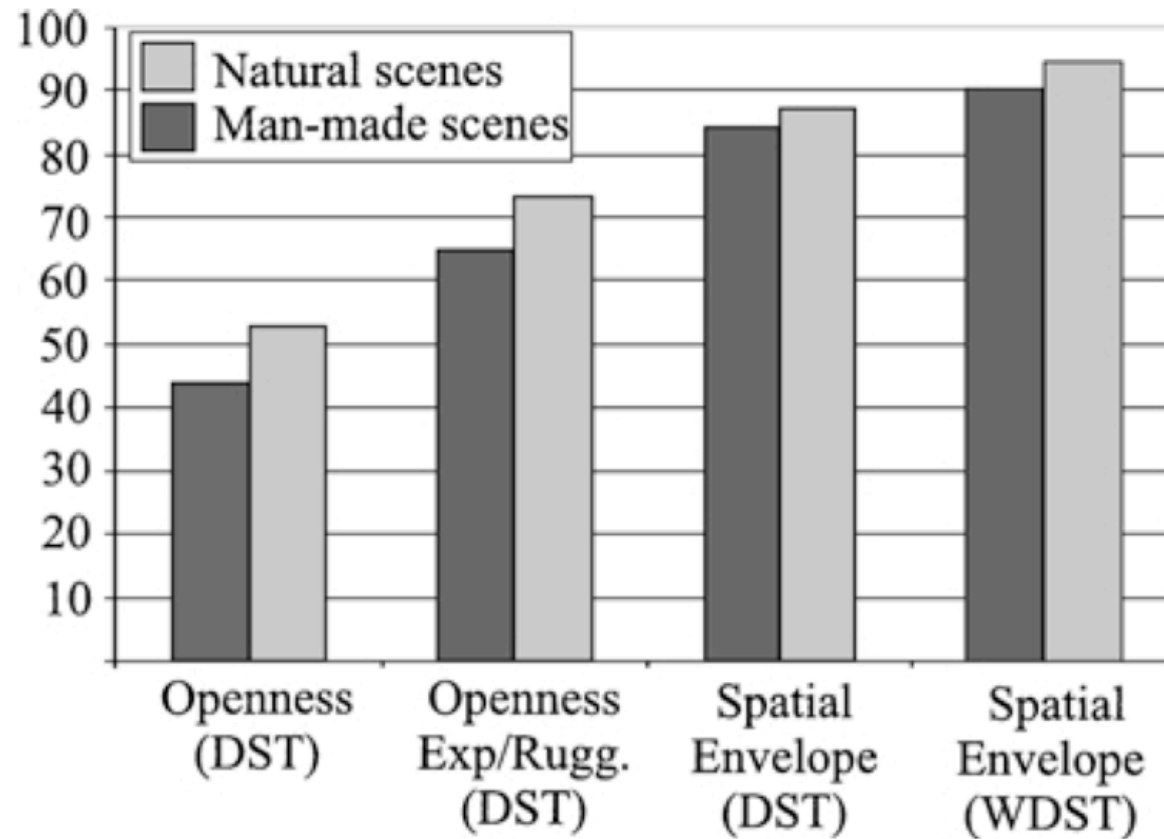
Example visual gists



Global features (I) ~ global features (I')



Are these measures useful for scene recognition?





Confusion matrix

experiments done using K-NN as classifier

	Coast	Country	Forest	Mountain
Coast	88.6	8.9	1.2	1.3
Country	9.8	85.2	3.7	1.3
Forest	0.4	3.6	91.5	4.5
Mountain	0.4	4.6	3.8	91.2



Confusion matrix

experiments done using K-NN as classifier

	Highway	Street	Close-up	Tall building
Highway	91.6	4.8	2.7	0.9
Street	4.7	89.6	1.8	3.9
Close-up	2.5	2.3	87.8	7.4
Tall building	0.1	3.4	8.5	88



Take Home Message

- you can recognize an **outdoor** scene with very simple global features
- frequency-based features seem to work very well
- adding spatial information seems to increase the effectiveness of the description



What happened next?

- Holistic Representation = Global Feature Representation
- Put some locality into the features/somewhere in the overall algorithm to preserve some spatial information
- Focus on the classification component



L. Fei Fei & P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. Proc CVPR 2005.

- Follow the idea of using feature representations before classifying scenes
- Oliva and Torralba needed to annotate manually the holistic properties --expensive!
- Contribution: automatic learning of relevant intermediate representations of scenes, using only the category label attached to the scene.



o. country
coast
suburb
tall bldg
highway
livingroom
bedroom



mountain
forest
streets
ins. city
office
kitchen

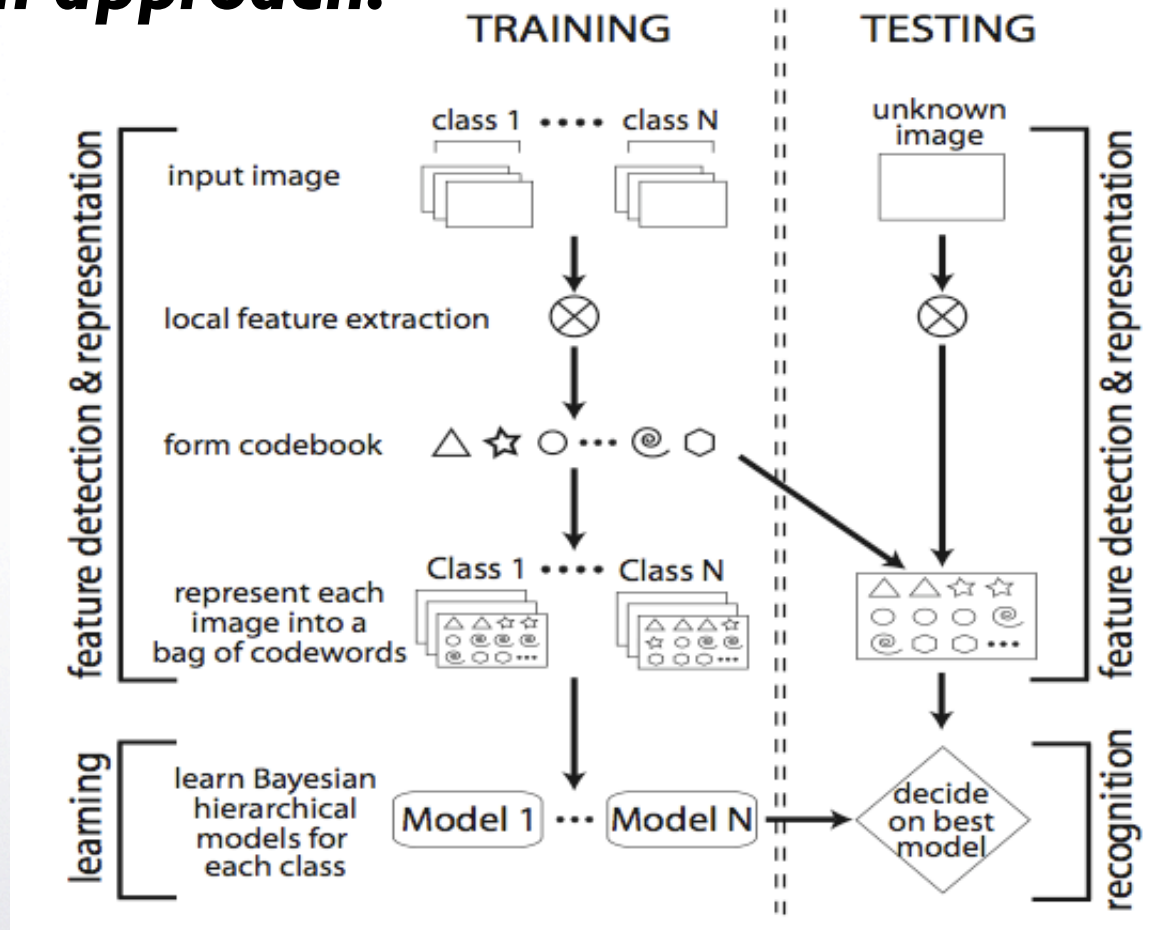


Contribution: automatic learning of relevant intermediate representations of scenes, using only the category label attached to the scene.

Further Contribution: database of 13 scene categories



The overall approach:





**Parenthesis: what is a codebook
representation?**



Object

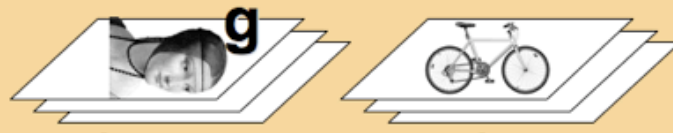


Bag of 'words'





learnin



feature detection
& representation

codewords dictionary



image representation

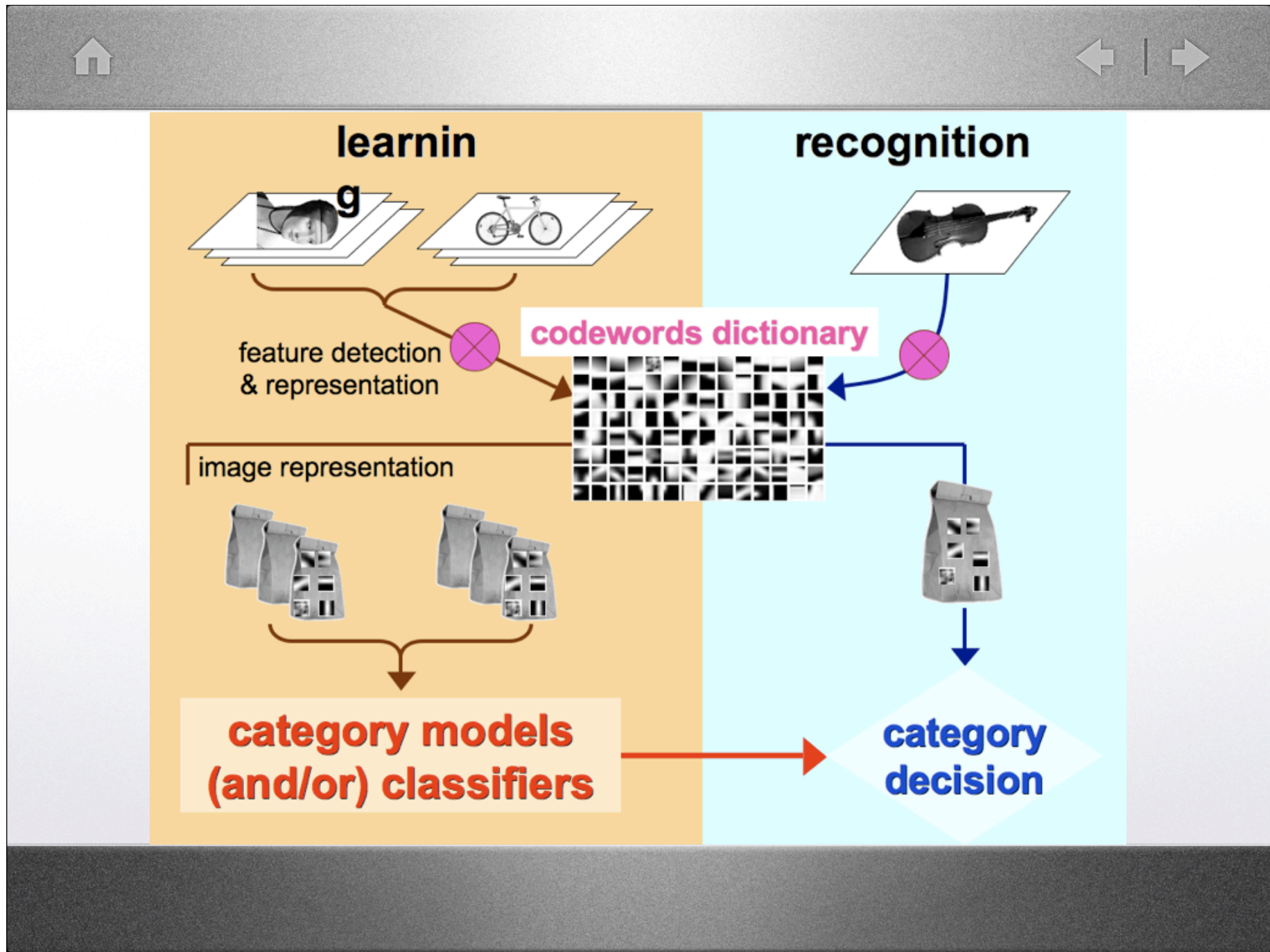


**category models
(and/or) classifiers**

recognition

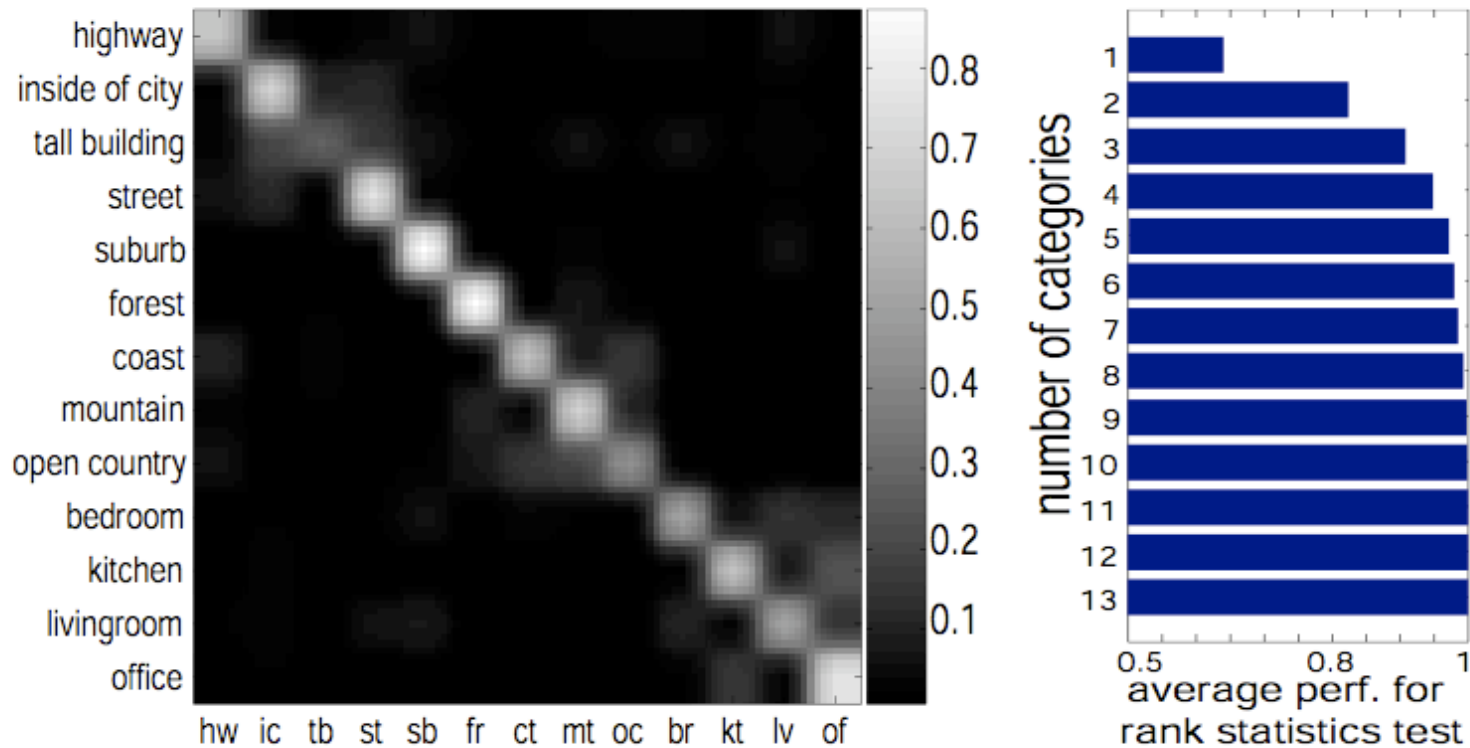


**category
decision**





Results: training 100 images per class, testing 50 images per class.
Performance: 64%





S. Lazebnik, C. Schmid, J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proc CVPR 2006.

- Contribution 1: the spatial information is preserved by the similarity measure between feature representations
- Contribution 2: a discriminative classification scheme (SVM).

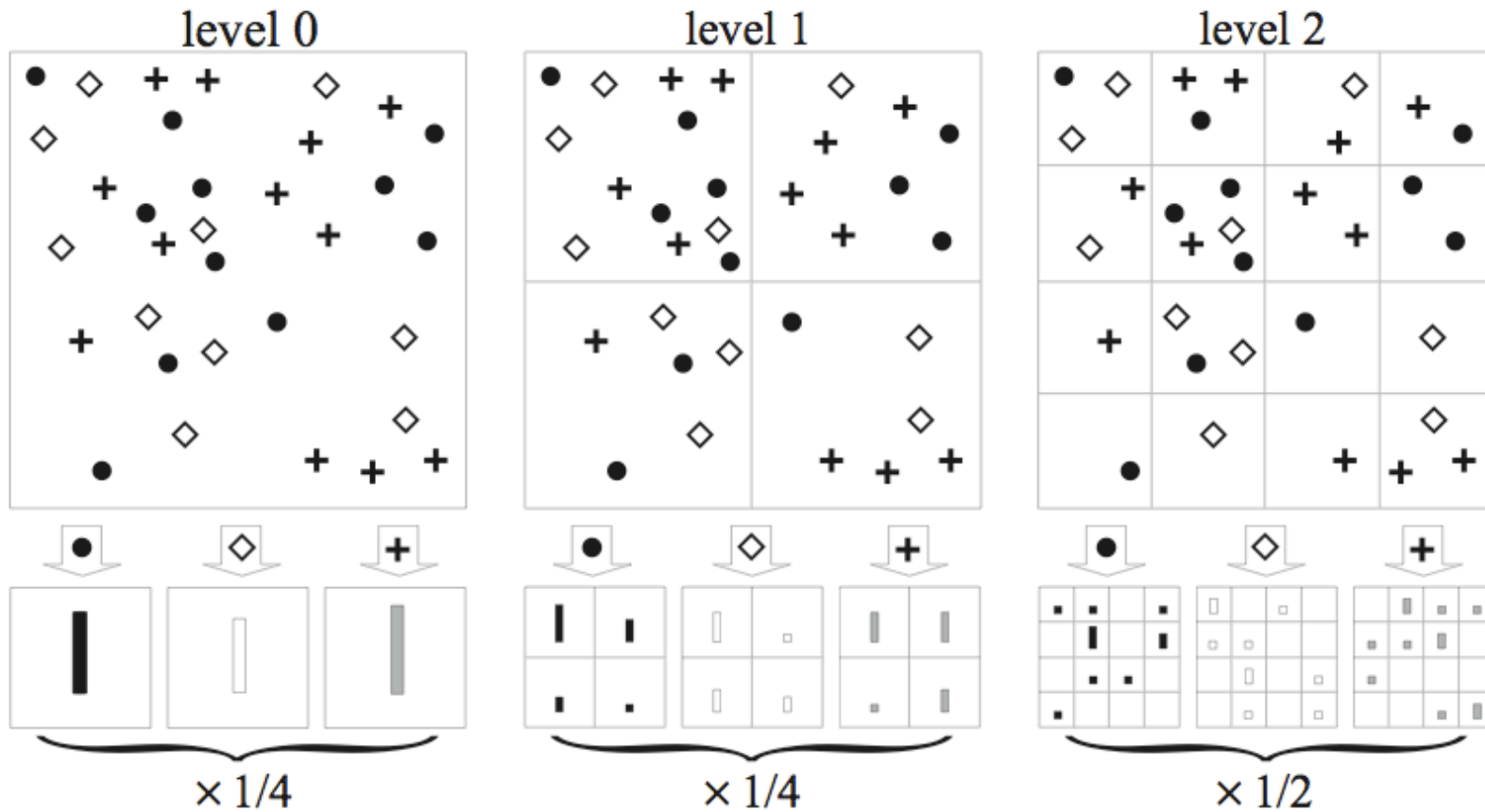


Similarity measure: Spatial Pyramid Matching

- At each level, build a codebook and generate a histogram representation
- Measure the similarity of each spatial area with an intersection measure
- Weighted sum (weights are a normalizing factor)



Similarity measure: *Spatial Pyramid Matching*





Experiments: The data



office



kitchen



living room



bedroom



store



industrial



tall building*



inside city*



street*



highway*



coast*



open country*



mountain*



forest*



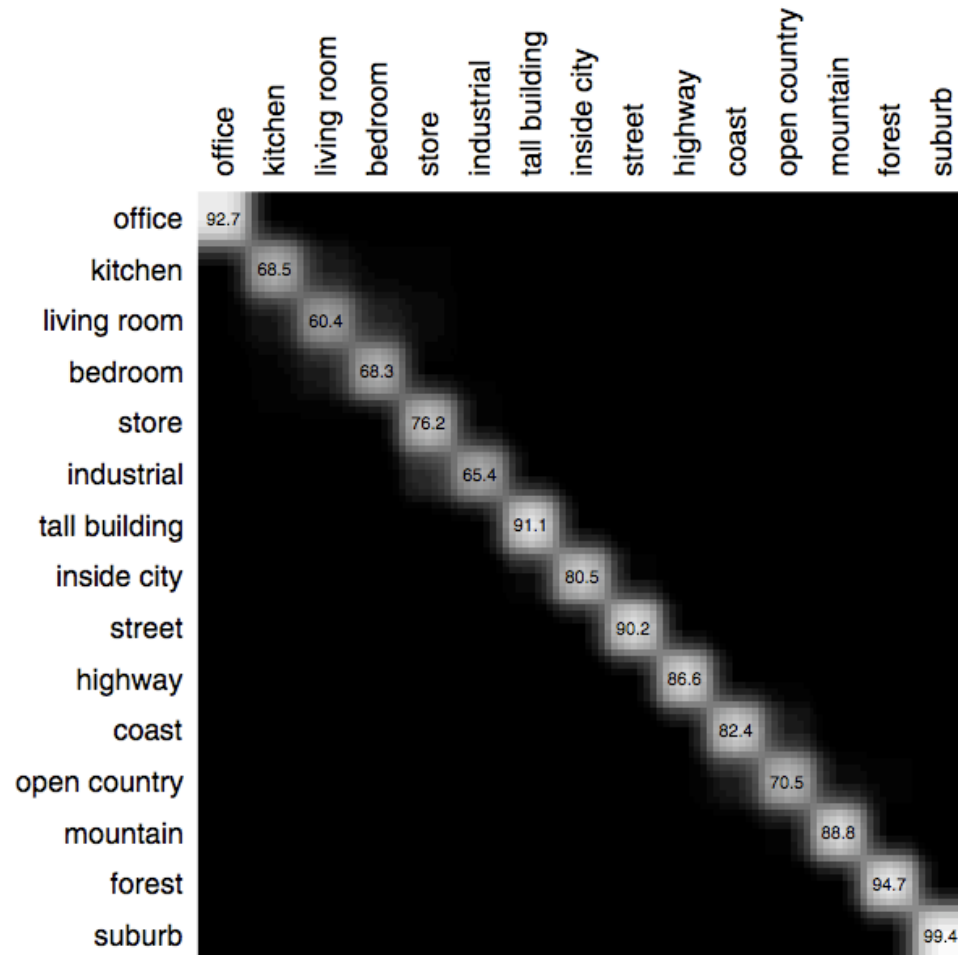
suburb



Results

L	Weak features ($M = 16$)		Strong features ($M = 200$)		Strong features ($M = 400$)	
	Single-level	Pyramid	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ± 0.5		72.2 ± 0.6		74.8 ± 0.3	
1 (2×2)	53.6 ± 0.3	56.2 ± 0.6	77.9 ± 0.6	79.0 ± 0.5	78.8 ± 0.4	80.1 ± 0.5
2 (4×4)	61.7 ± 0.6	64.7 ± 0.7	79.4 ± 0.3	81.1 ± 0.3	79.7 ± 0.5	81.4 ± 0.5
3 (8×8)	63.3 ± 0.8	66.8 ± 0.6	77.2 ± 0.4	80.7 ± 0.3	77.2 ± 0.5	81.1 ± 0.6

- Weak features: Oliva&Torralba-like;
- Strong features: SIFT
- The database is the Fei-Fei&Perona05, plus two new category



Confusion matrix:
the confusions occur
between the indoor
classes!



15 min break!



J.Vogel, B. Schiele. *Semantic Scene modeling and retrieval for content-based image retrieval*. IJCV, Vol 72 (2), 133-157, 2007.

- Contribution 1: definition of local semantic descriptor for scene representations --a semantic vocabulary
- Contribution 2: ranking of natural scenes according to semantic similarity to chosen scene category
- Contribution 3: a perceptually plausible similarity measure highly correlated to human rankings

(Slides credit B. Schiele)



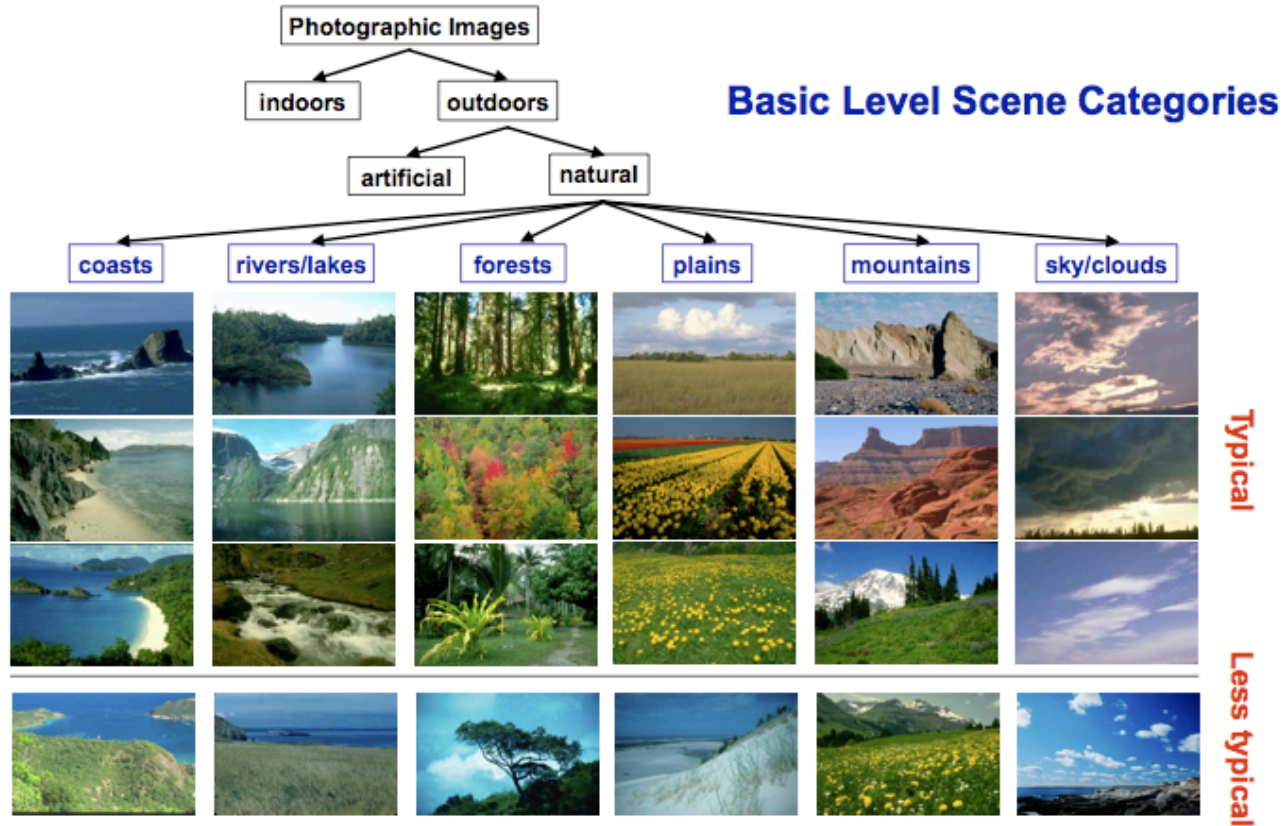
J.Vogel, B. Schiele. *Semantic Scene modeling and retrieval for content-based image retrieval*. IJCV, Vol 72 (2), 133-157, 2007.

- Contribution 1: definition of local semantic descriptor for scene representations --a semantic vocabulary
- Contribution 2: ranking of natural scenes according to semantic similarity to chosen scene category
- Contribution 3: a perceptually plausible similarity measure highly correlated to human rankings

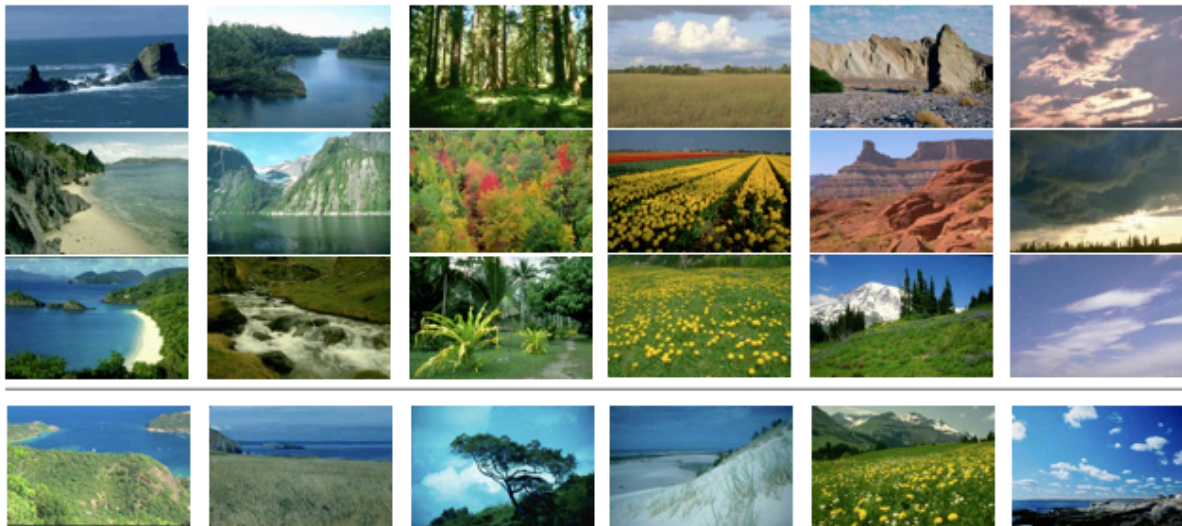
(Slides credit B. Schiele)



Natural Scene Modeling



Basic Level Scene Categories



Typical

Less typical



Semantic Modeling

Local Semantic Concepts*

*inspired by [Mojsilovic et al., 2004]



Global image representation, e.g. for categorization or ranking

Database Images



10x10 Grid



Semantic Labeling

sky	sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	sky	sky	sky	sky	sky	sky	sky	sky	sky
sky	sky	rocks	rocks	rocks	sky	sky	sky	sky	sky	sky
rocks	rocks	water	rocks	rocks	rocks	rocks	rocks	rocks	water	water
sand	sand	sand	sand	sand	sand	sand	sand	sand	water	water
sand	sand	water	water	water	water	water	water	water	water	water
sand	sand	sand	water	water	water	water	water	water	water	water

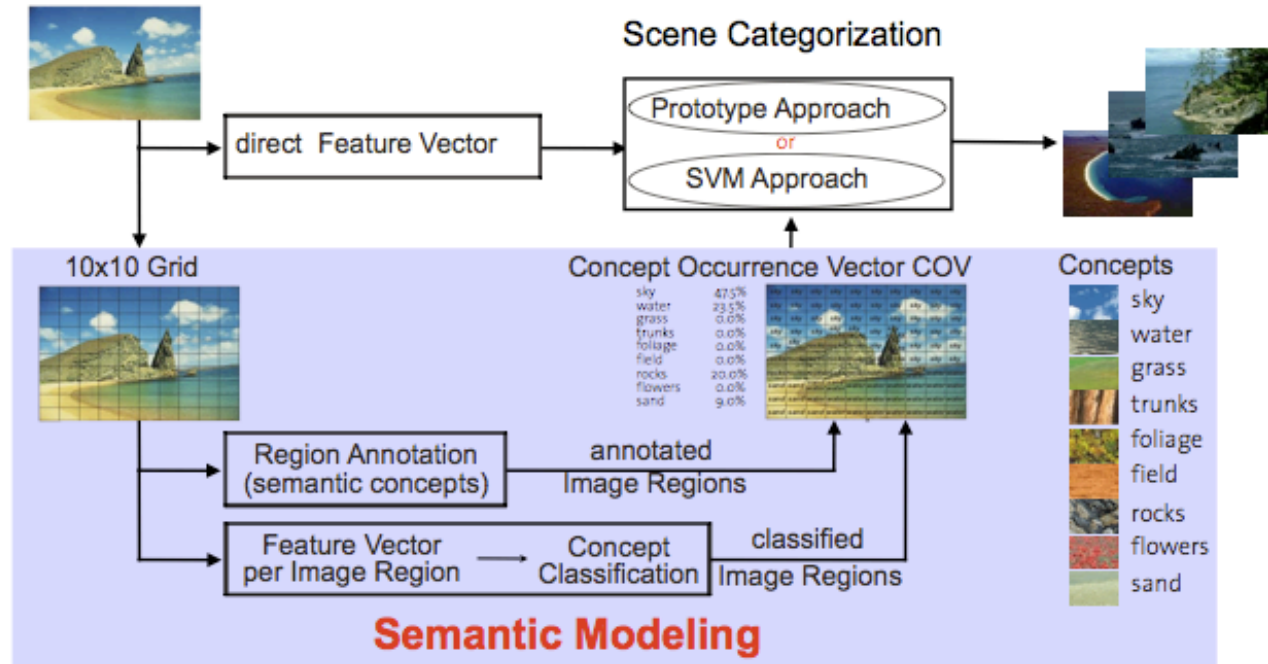
Concept Occurrence Vector

sky	47.5%
water	23.5%
grass	0.0%
trunks	0.0%
foliage	0.0%
field	0.0%
rocks	20.0%
flowers	0.0%
sand	9.0%



Categorization Experiments

Database Images

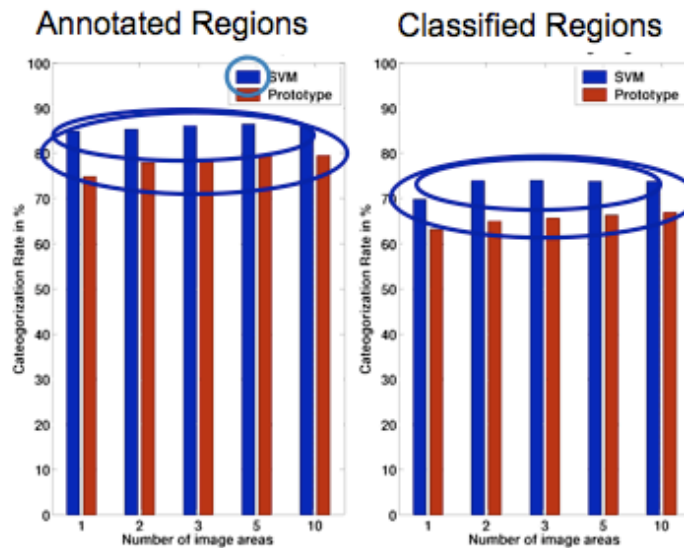


1. Semantic Modeling vs. Direct Feature Extraction
2. Annotated vs. Classified Image Regions
3. Prototype vs. SVM Classifier

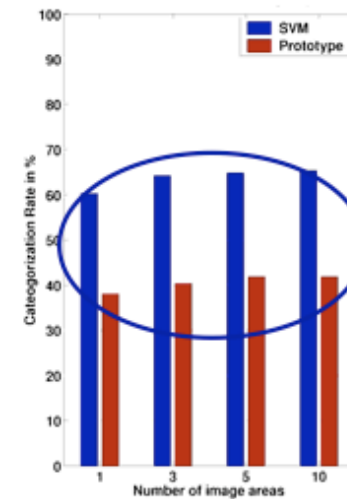


Categorization Results

Semantic Modeling



No Semantic Modeling



1. **Support-Vector Machines outperform Prototypes.**
2. **Semantic Modeling improves results considerably.**
3. **Fully automatic categorization at 74% categorization rate**

But: Benchmark (annotated regions) at only 86.4% categorization rate.



Wrapping Up

- Global descriptors combined with some local information, plus your pet learning algorithm will classify decently well 10-15 natural scenes --you might add few indoor scenes and still be fine
- Not clear how this approach would scale to 50-100-more scene categories
- Not clear how would it work with only indoor scenes
- Not clear if this approach is transferrable to an autonomous system



What about *indoor* scenes?





What about indoor scenes?

- **spatial properties:**
 - *Degree of Naturalness*
 - *Degree of Openness*
 - *Degree of Roughness*
 - *Degree of Expansion*
 - *Degree of Ruggedness*





What about indoor scenes?

- **spatial properties:**

- *Degree of Naturalness*
- *Degree of Openness*
- *Degree of Roughness*
- *Degree of Expansion*
- *Degree of Ruggedness*





What about indoor scenes?

- The properties of the spatial envelope defined in Oliva&Torralba do not seem to make much sense for indoor scenes
- Indoor environments are usually images at a much closer distance than outdoor scenes, therefore they presents a much higher variability in their visual appearance as the imaging viewpoint changes
- For these reasons it is also not obvious that a codebook/BOW representation would work



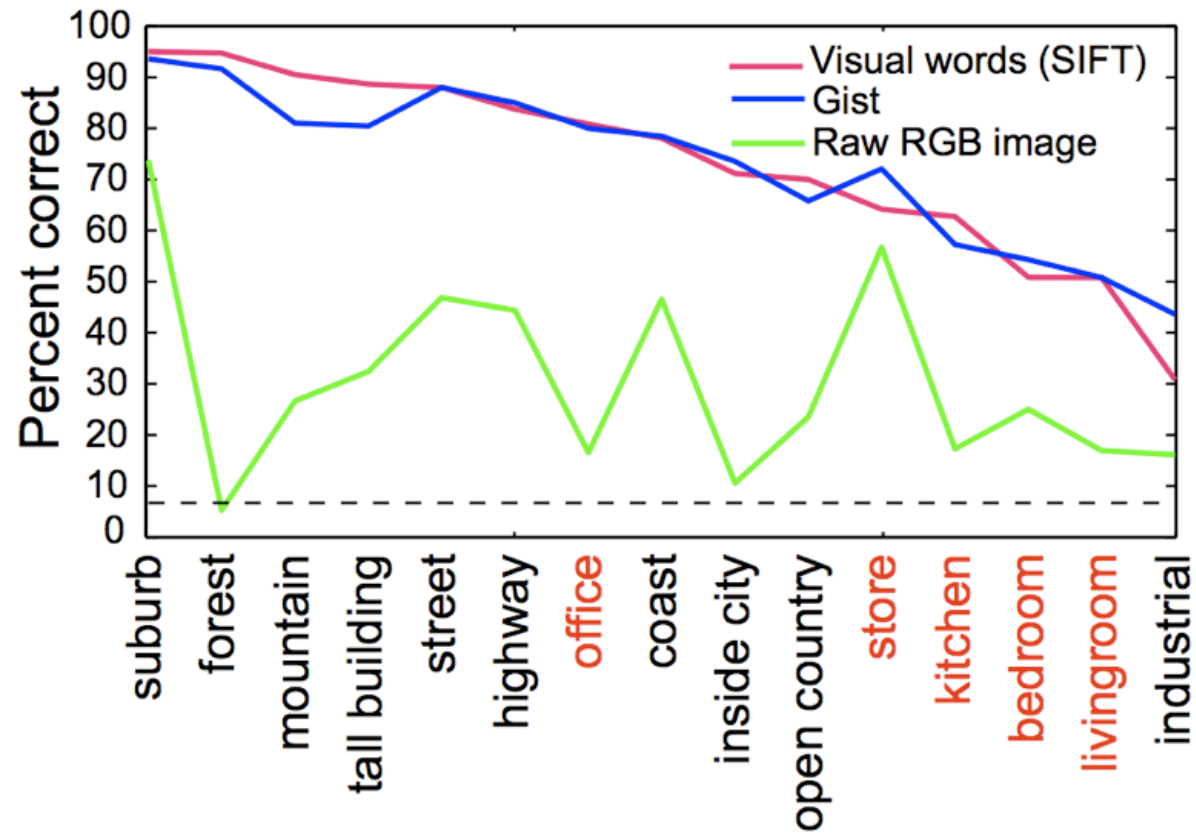
Let's check this all out computationally

A. Quattoni, A. Torralba. *Recognizing indoor scenes*. Proc International Conference on Computer Vision and Pattern Recognition, 2009

- Contribution 1: experimental evaluation of several methods for outdoor recognition on Lazebnik et al 2006 database, outlining current limitations
- Contribution 2: a database of 67 indoor categories, publicly available
- Contribution 3: a new computational model for tackling the indoor scene recognition problem



Contribution I: a proof-of-concept experiment





Contribution II: a new dataset



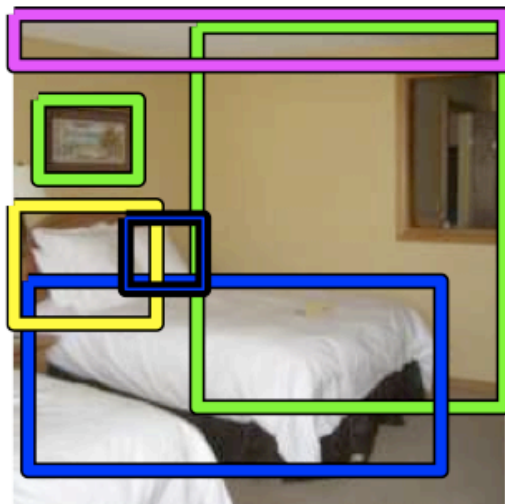


Contribution II: a new dataset

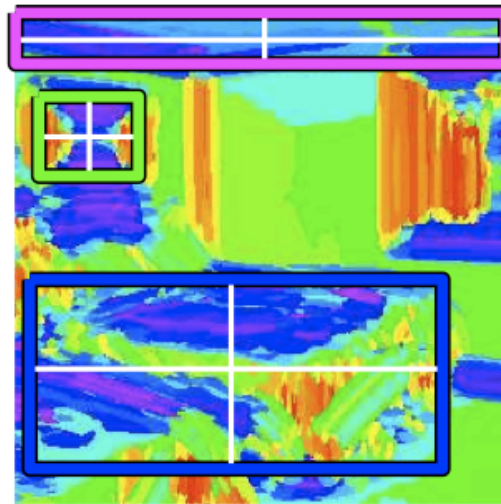
Public spaces				Leisure			Working place			
prison cell	library	cloister	church	buffet	fastfood	concert hall	hospital room	kinder garden	restaurant kitchen	artstudio
waiting room	museum	elevator		restaurant	bar	movie theater	classroom	laboratory wet	studio music	operating room
pool inside	inside bus	inside subway	subway	gameroom	casino	bowling	office	computer room	warehouse	green house
locker room	trainstation	airport inside		gym	hair salon		dental office	tv studio	meeting room	



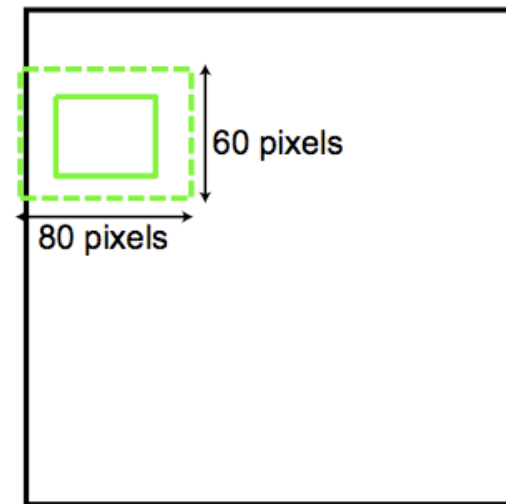
Contribution III: A New Computational Approach



a) Prototype and candidate ROI



b) ROI descriptors



c) Search region

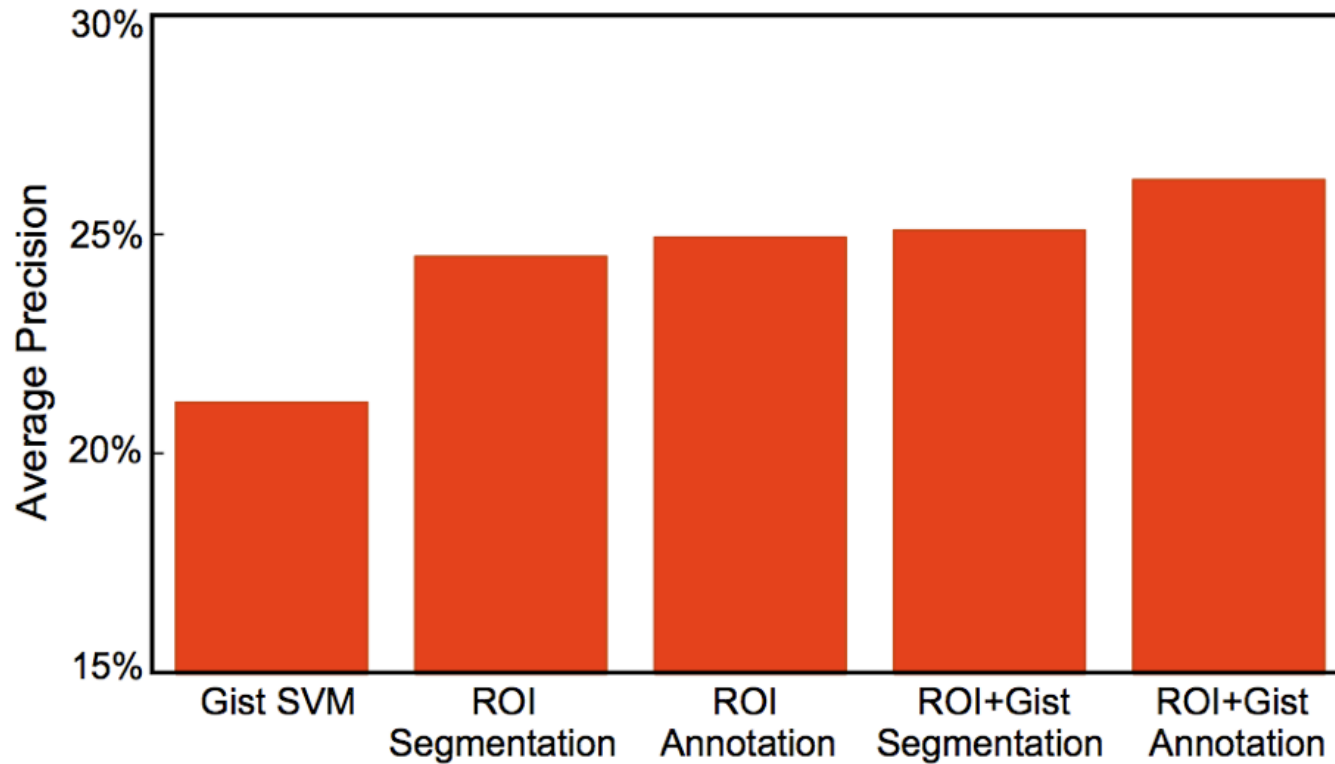


Indoor Place Recognition

- Each scene class described by a set of prototypes
- Each scene prototype is defined by a set of Region of Interests (ROI) and their loose relative position. ROIs represented via BOWs
- The ROI may or may not correspond to a specific object
- ROIs, the relative positions and the prototypes are learned in a supervised fashion during training



Indoor Place Recognition





Indoor Place Recognition

church inside 63.2%
elevator 61.9%
auditorium 55.6%
buffet 55.0%
classroom 50.0%
greenhouse 50.0%
bowling 45.0%
cloister 45.0%
concert hall 45.0%
computerroom 44.4%
dentaloffice 42.9%
library 40.0%
inside bus 39.1%
closet 38.9%
corridor 38.1%
grocerystore 38.1%
locker room 38.1%
florist 36.8%

studiomusic 36.8%
hospitalroom 35.0%
nursery 35.0%
trainstation 35.0%
bathroom 33.3%
laundromat 31.8%
stairscase 30.0%
garage 27.8%
gym 27.8%
tv studio 27.8%
videostore 27.3%
gameroom 25.0%
pantry 25.0%
poolinside 25.0%
inside subway 23.8%
kitchen 23.8%
winecellar 23.8%

fastfood restaurant 23.5%
bar 22.2%
clothingstore 22.2%
casino 21.1%
deli 21.1%
bookstore 20.0%
waitingroom 19.0%
dining room 16.7%
bakery 15.8%
livingroom 15.0%
movietheater 15.0%
bedroom 14.3%
toystore 13.6%
operating room 10.5%
airport inside 10.0%
artstudio 10.0%
lobby 10.0%
prison cell 10.0%

hairsalon 9.5%
subway 9.5%
warehouse 9.5%
meeting room 9.1%
children room 5.6%
shoeshop 5.3%
kindergarden 5.0%
restaurant 5.0%
museum 4.3%
restaurant kitchen 4.3%
jewelleryshop 0.0%
laboratorywet 0.0%
mall 0.0%
office 0.0%



Take home Message

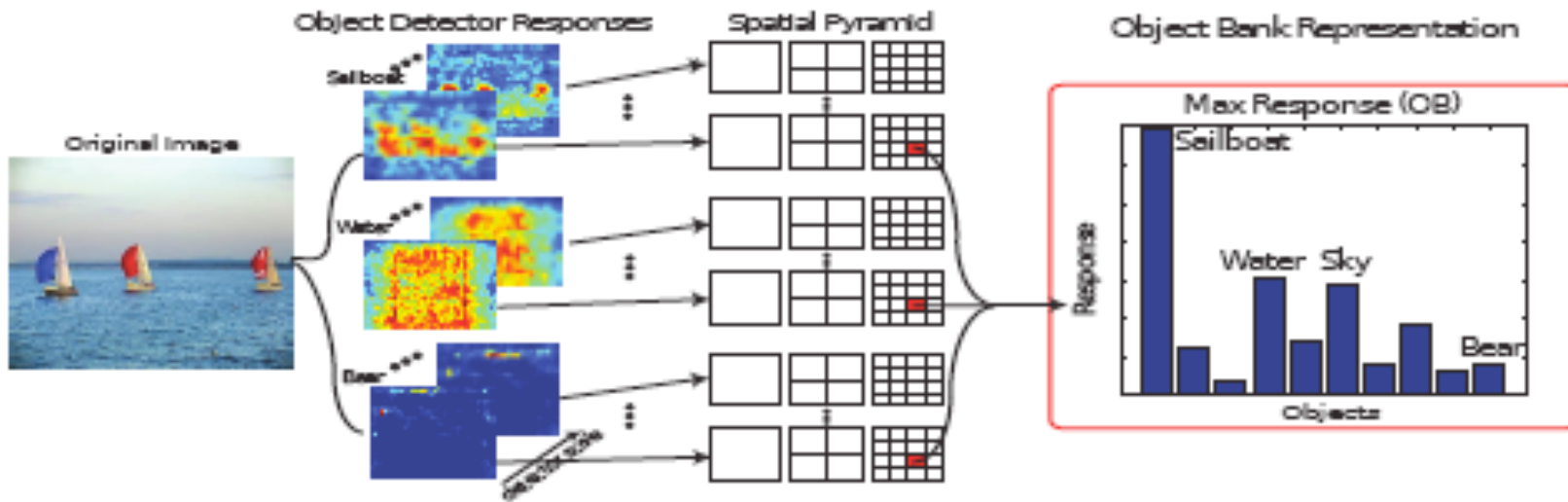
- You can represent and recognize well outdoor scene images with a global description
- The same approach fails for indoor scene images
- Adding object description seem to help but not so much



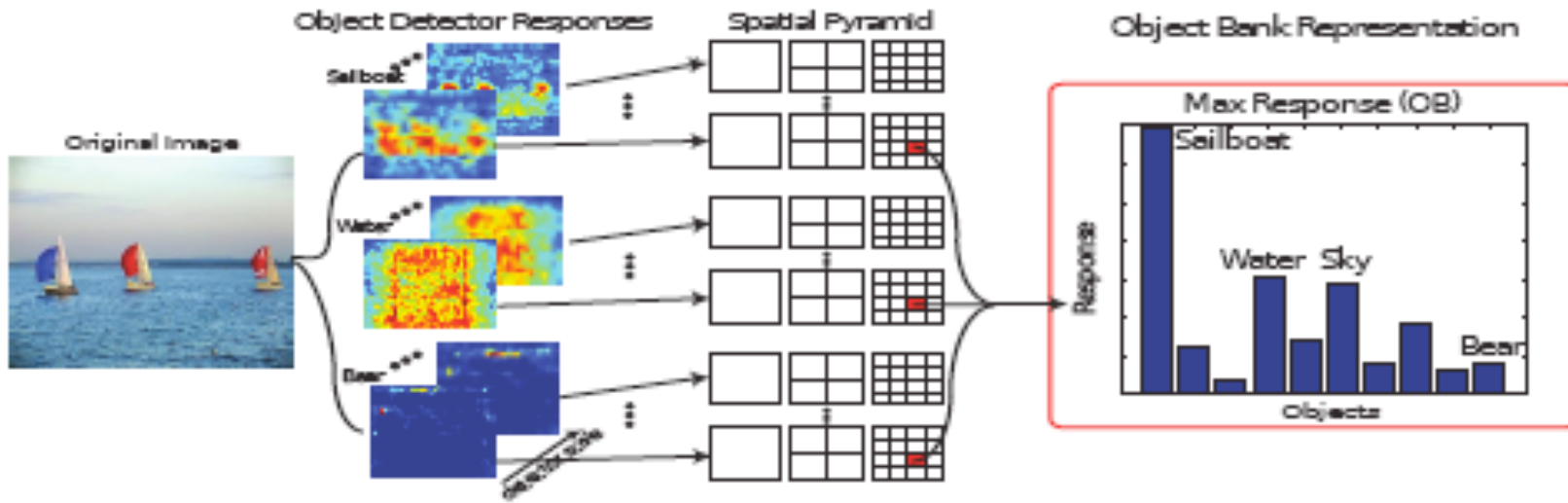
How about using only objects?

Li-Jia Li, Hao Su, Eric P. Xing, Li Fei Fei. *Object Bank: a high-level image representation for scene classification and semantic feature sparsification*. Proc NIPS, 2010

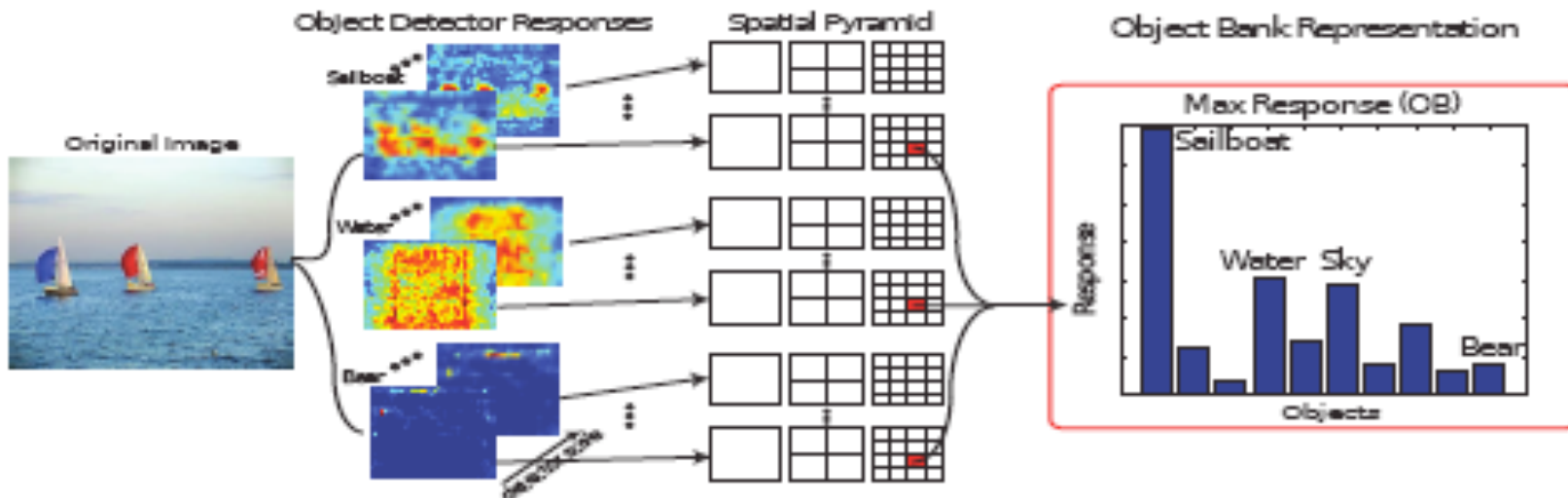
- Contribution 1: an image representation based on the scale invariant response map of a large number of pre-trained object detectors
- Contribution 2: state of the art results on ISR dataset



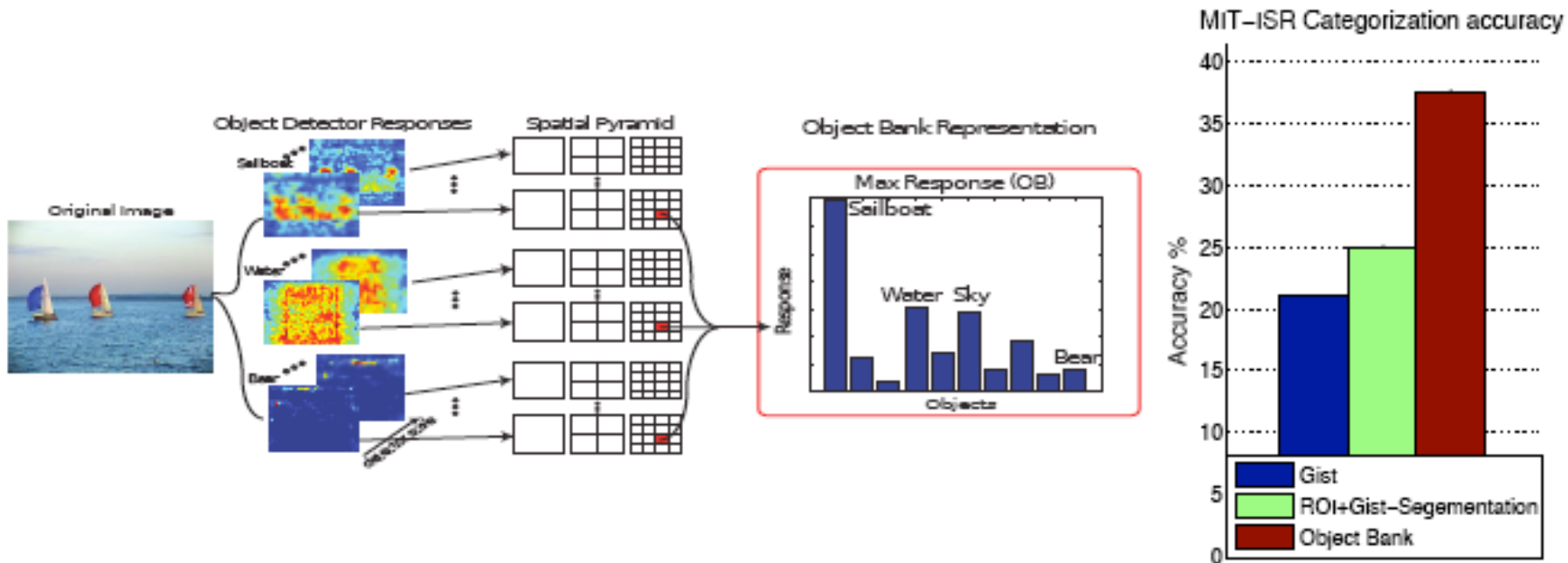
- **Step 1** apply object detectors to an input image at different scales



- **Step2** for each object at each scale, use a three-level spatial pyramid representation of the resulting object filter map



- **Step3** compute the maximum response for each object in each grid. It results in a feature vector of length $\#_of_objects$. **Their concatenation is the object bank**



- **Outperforms Gist, gist+ROI**



what does it measure, really?



TV Studio



Bathroom



what does it measure, really?



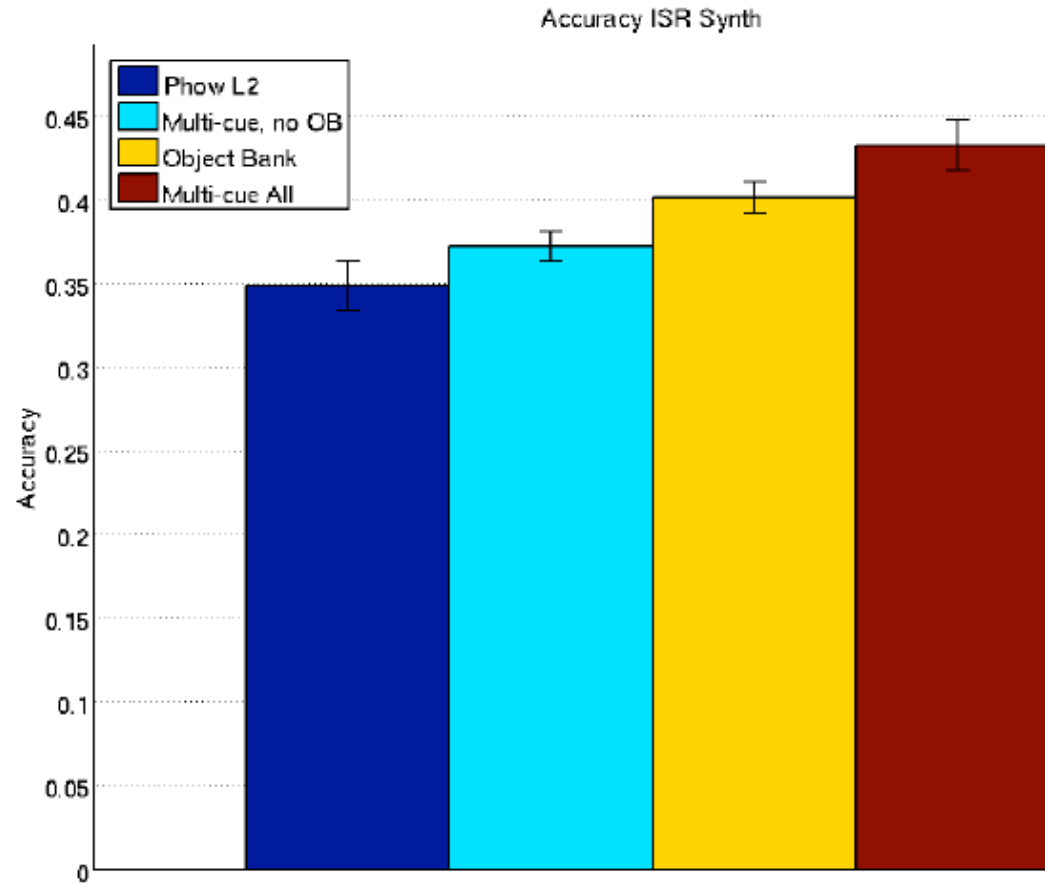
Airport Inside



Laundry



still, it seems to provide complementary information wrt global descriptors





Scene Recognition (to be continued...)