MIT CSAIL

6.869: Advances in Computer Vision

MIT COMPUTER VISION

Lecture 25

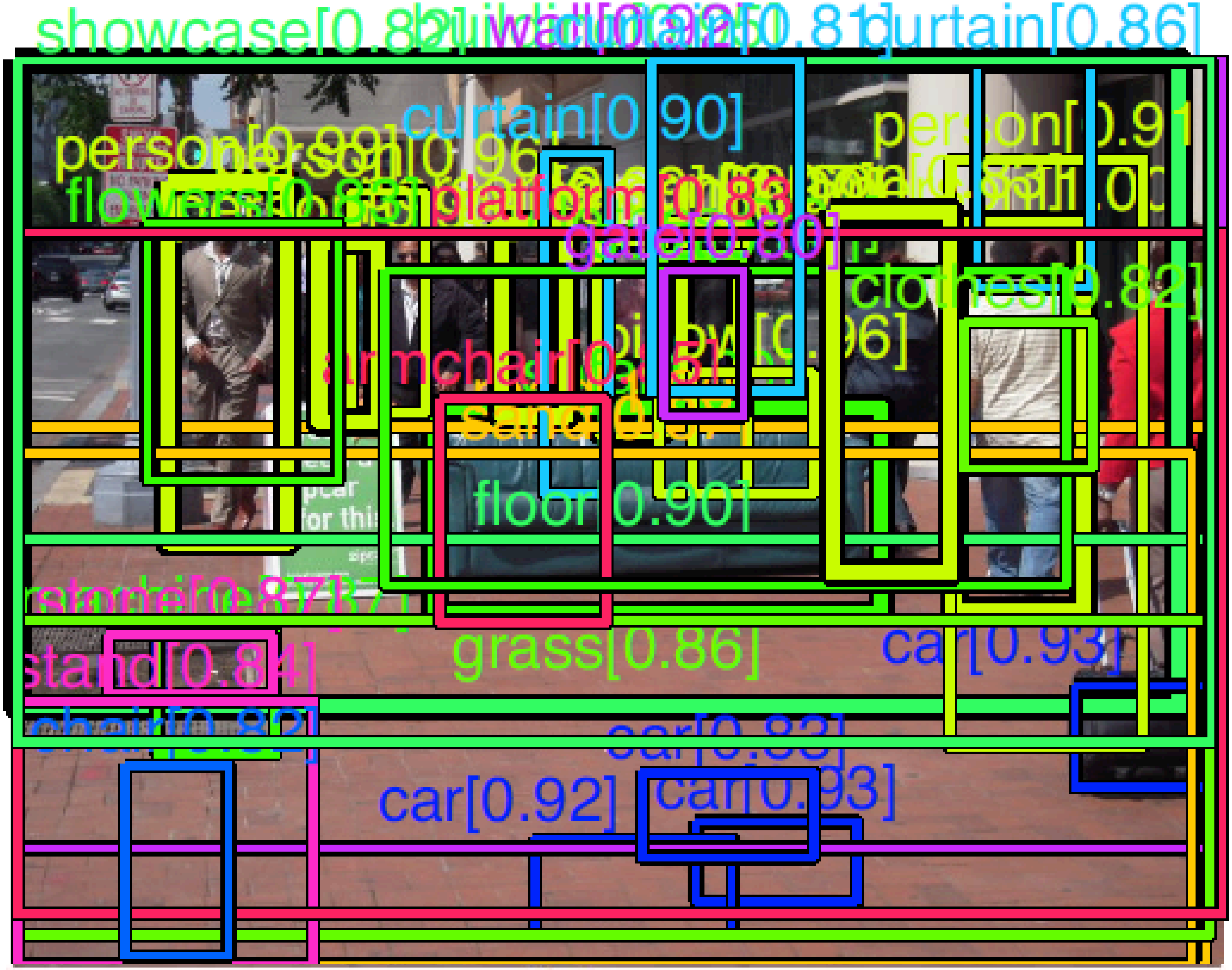
Scene recognition

The texture

The object

The scene


you
need a
zipcar
for this.
zipcar


curtain[0.86]
curtain[0.90]
person[0.99]
platform[0.83]
gate[0.80]
clothes[0.82]
floor[0.90]
grass[0.86]
car[0.93]
stand[0.84]
chair[0.82]
car[0.83]
car[0.92]
car[0.93]

The detector challenge

By looking at the output of a detector on a random set
of images, can you guess which object is it trying to detect?

What object is the detector trying to detect?

By looking at the output of a detector on a random set
of images, can you guess which object is it trying to detect?

What object is the detector trying to detect?

By looking at the output of a detector on a random set
of images, can you guess which object is it trying to detect?

Microwave

Search

Top 8 out of 4317 images

P. Felzenszwalb, D. McAllester, and D. Ramanan. *CVPR, 2008*

Microwave & refrigerator | Search

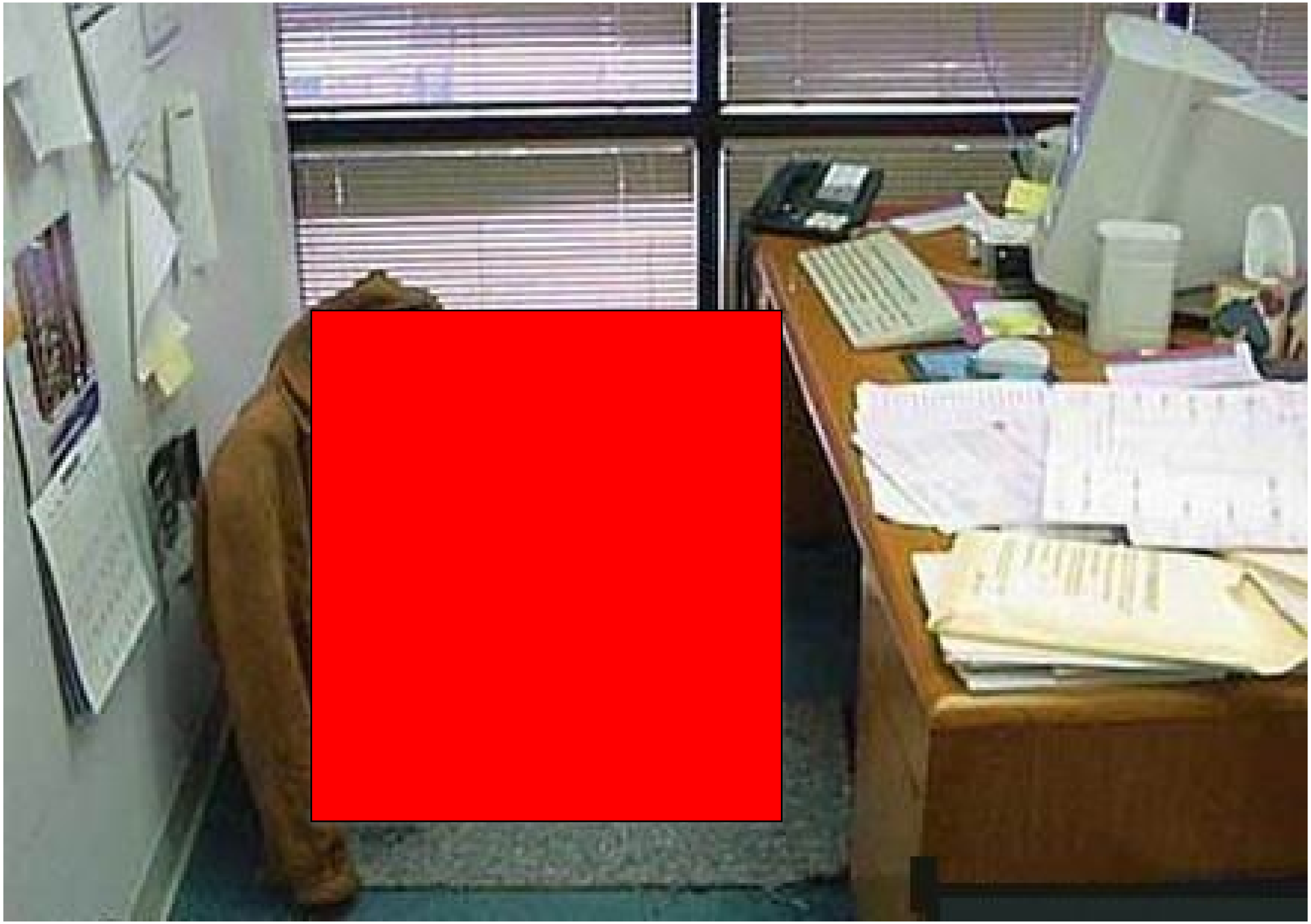







Top 8 out of 4317 images

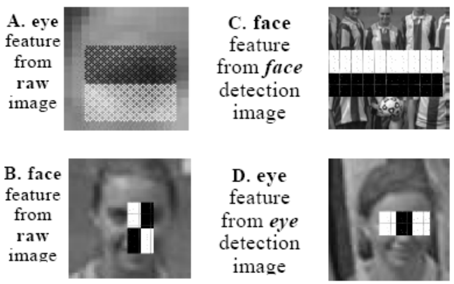
What object is hidden behind the red box?



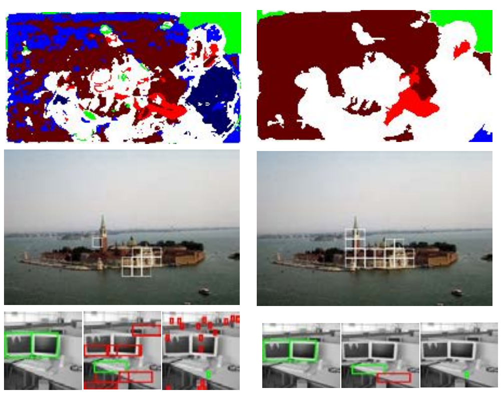
Objects in context

Torralba, Sinha (2001)

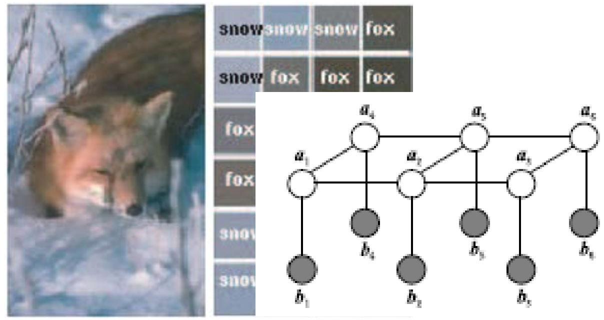
Fink & Perona (2003)



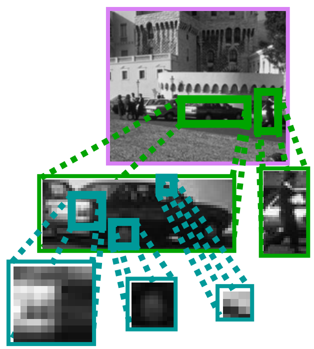
Kumar, Hebert (2005)



Carbonetto, de Freitas & Barnard (2004)

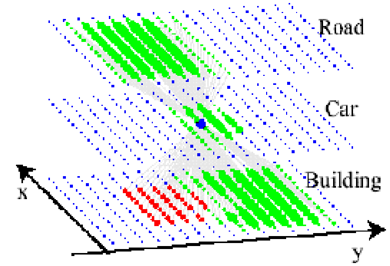


Sudderth, Torralba, Wilsky, Freeman (2005)

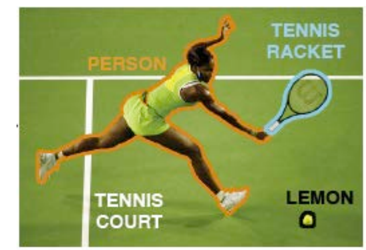


Heitz and Koller (2008)

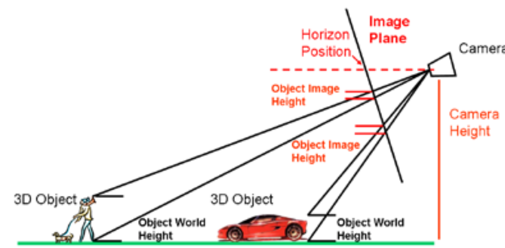
Torralba Murphy Freeman (2004)



Rabinovich et al (2007)

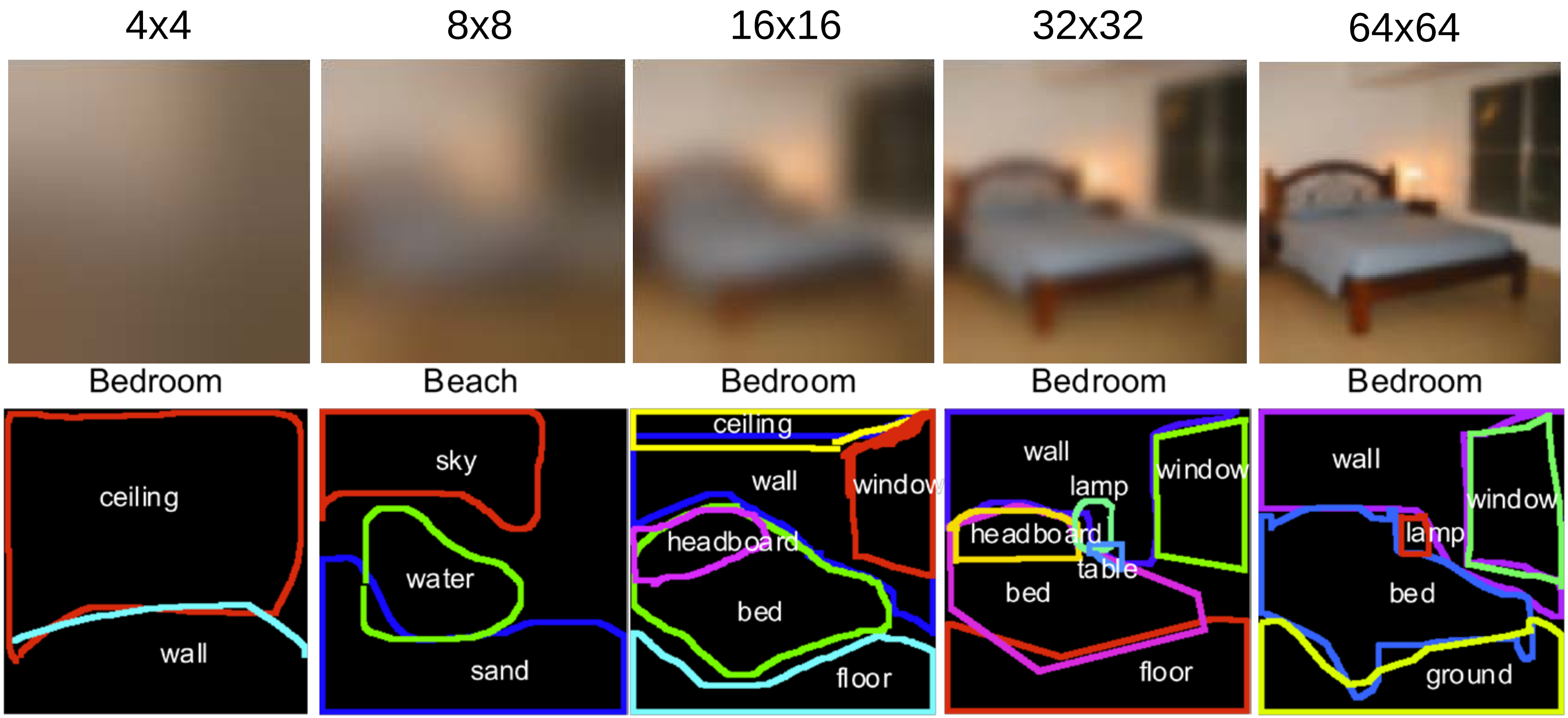


Hoiem, Efros, Hebert (2005)

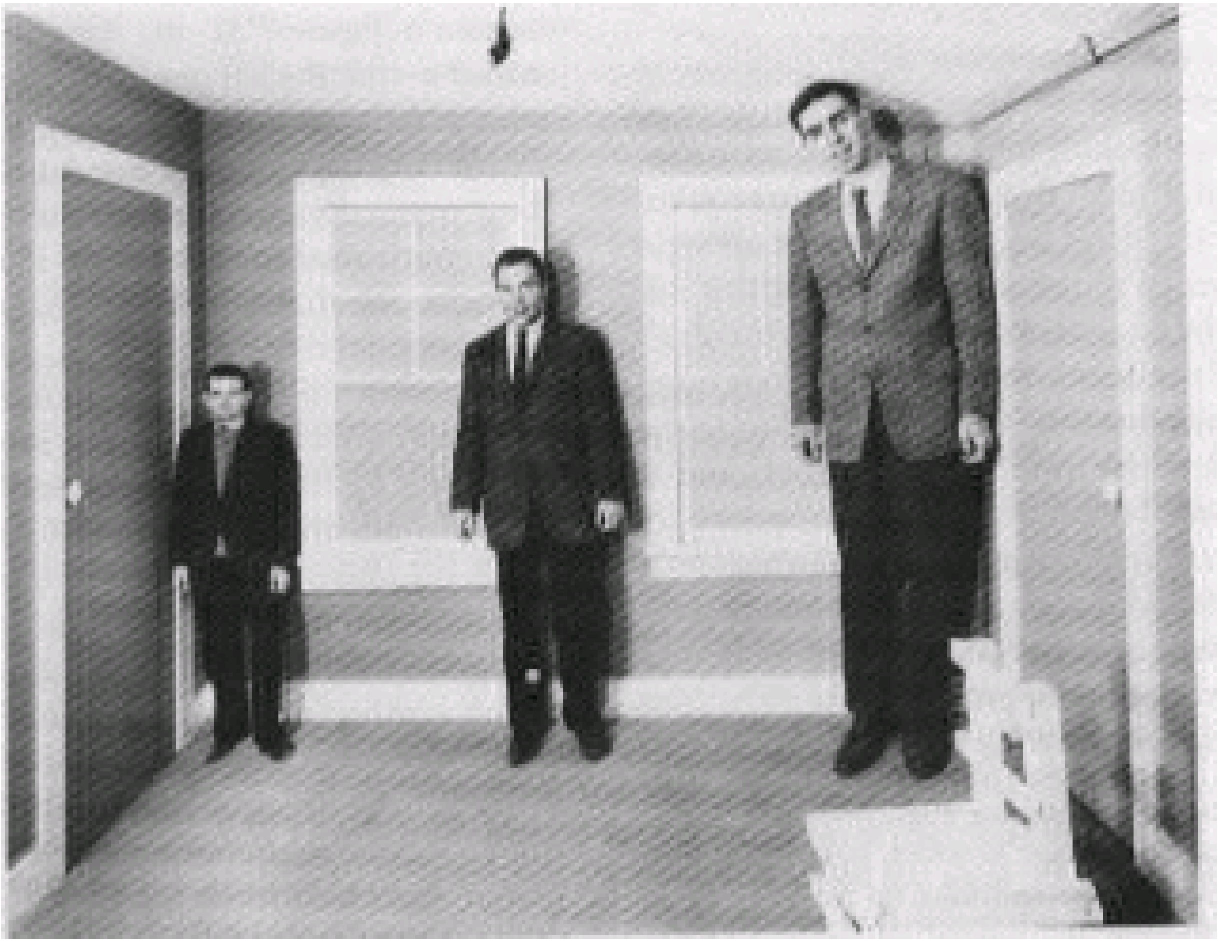


Desai, Ramanan, and Fowlkes (2009)

Increasing the context strength



Scenes rule over objects



3D percept is driven by the scene, which imposes its ruling to the objects

Scene views vs. objects

By scene we mean a place in which a human can act within, or a place to which a human being could navigate. Scenes are a lot more than just a combination of objects (just as objects are more than the combinations of their parts). Like objects, scenes are associated with specific functions and behaviors, such as eating in a restaurant, drinking in a pub, reading in a library, and sleeping in a bedroom.

Scene views vs. objects

A photograph of a firehydrant

A photograph of a street

Scene Categorization

Oliva and Torralba, 2001


Coast


Forest


Highway


Inside City


Mountain


Open Country


Street


Tall Building

Fei Fei and Perona, 2005

\+


Bedroom


Kitchen


Living Room


Office


Suburb

Lazebnik, Schmid, and Ponce, 2006

\+


Industrial


Store

15 Scene Database

Mary Potter (1976)

Mary Potter (1975, 1976) demonstrated that during a rapid sequential visual presentation (100 msec per image), a novel picture is instantly **understood** and observers seem to comprehend a lot of visual information

Demo : Rapid image understanding

By Aude Oliva

Instructions: 9 photographs will be shown for half a second each. Your task is to **memorize these pictures**

Memory Test

Which of the following pictures have you seen ?

If you have seen the image
clap your hands once

If you have not seen the image
do nothing

Have you seen this picture ?


NO

Have you seen this picture ?


NO

Have you seen this picture ?


NO

Have you seen this picture ?


NO

Have you seen this picture ?


Yes

Have you seen this picture ?


NO

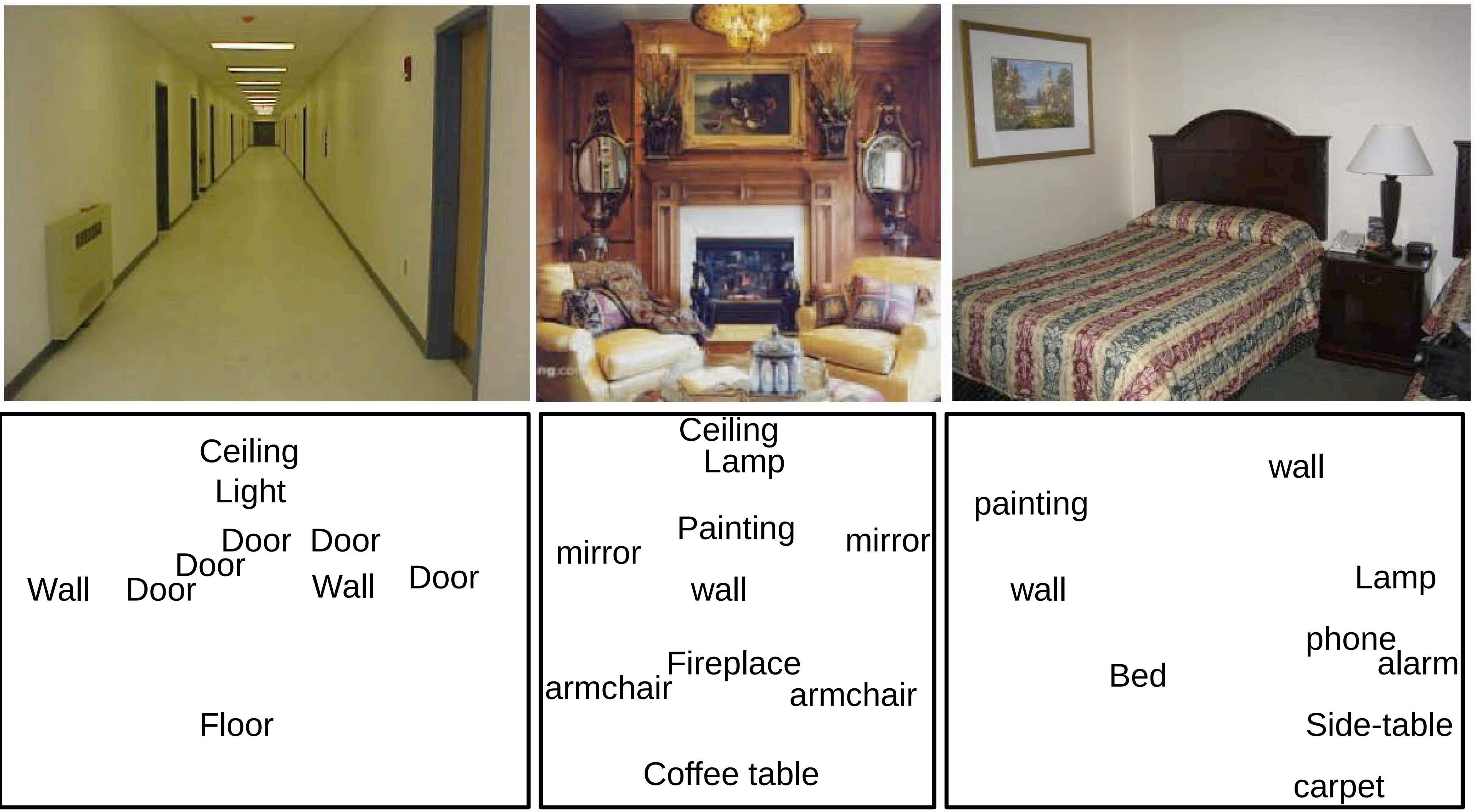
You have seen these pictures

You were tested with these pictures

The gist of the scene

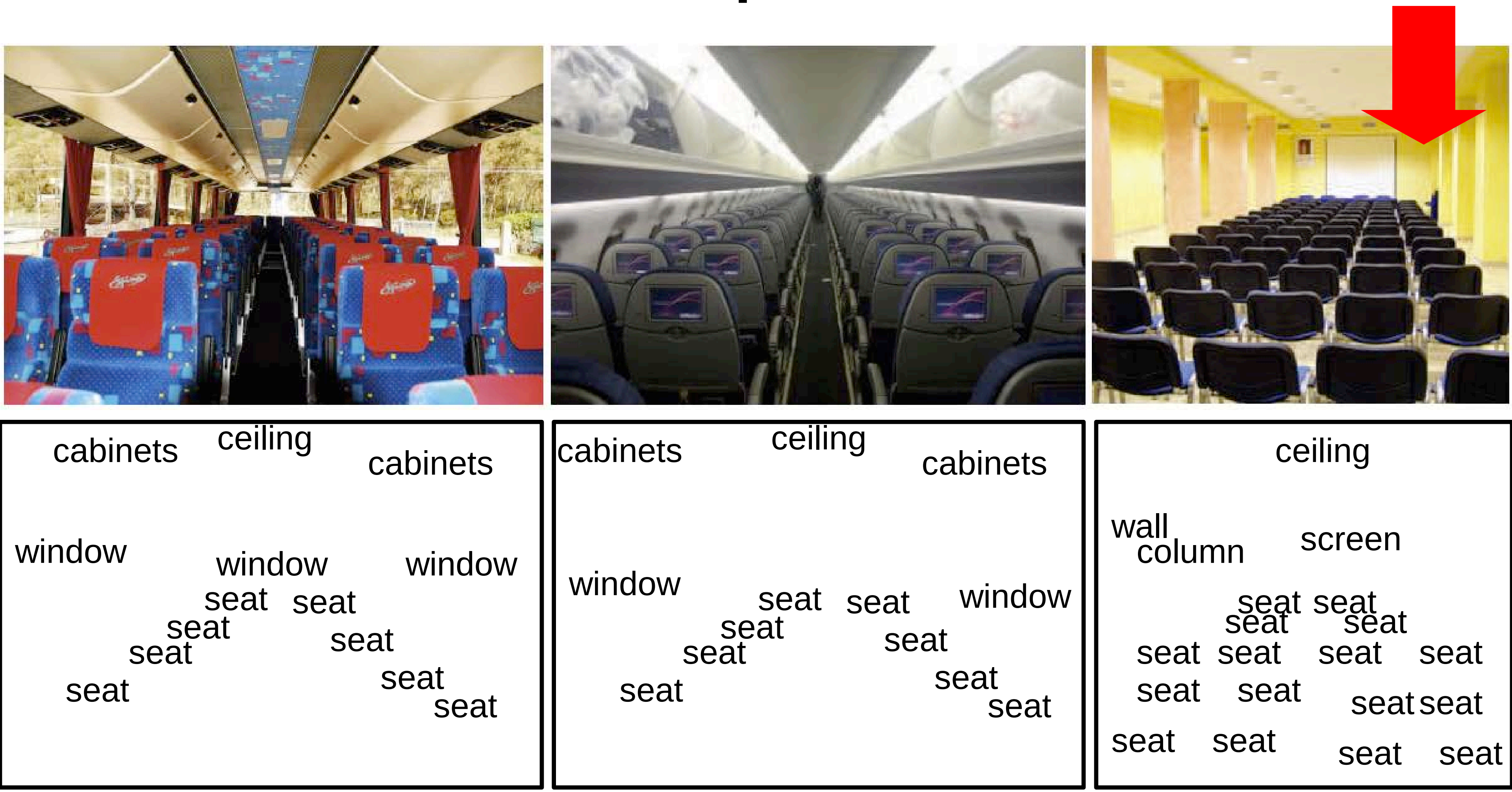
In a glance, we remember the meaning of an image and its global layout but some objects and details are forgotten

Which are the important elements?



Different content (i.e. objects), different spatial layout

Which are the important elements?



Similar objects, and similar spatial layout

Different lighting, different materials, different “stuff”

What can be an alternative to objects?

Scene emergent features

“Recognition via features that are not those of individual objects but “emerge” as objects are brought into relation to each other to form a scene.” – Biederman 81

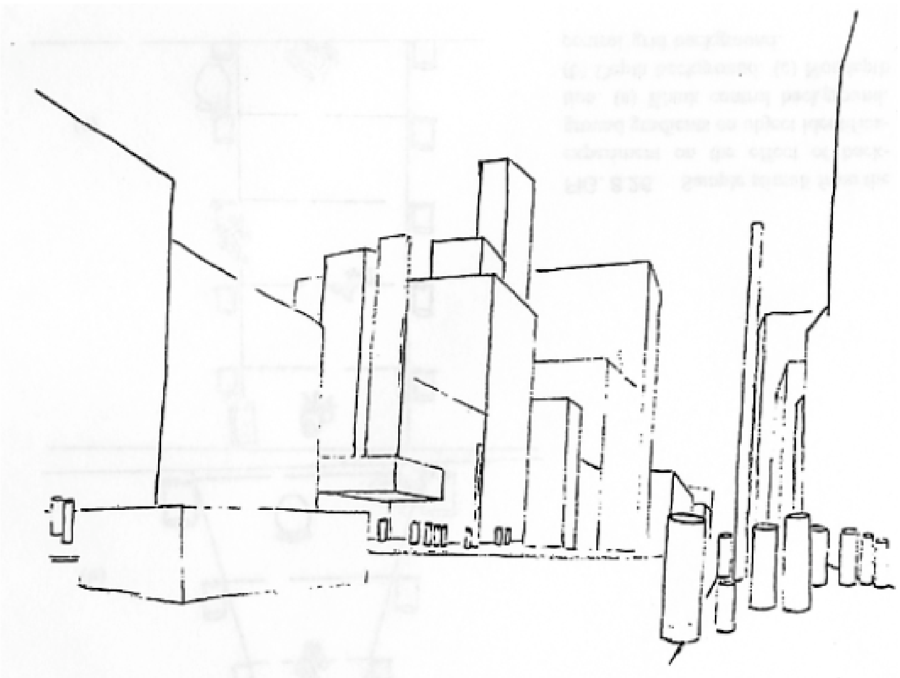


FIG. 8.23. *Downtown Buffalo*. Drawn by Robert Mezzanotte by converting objects in a photograph to basic rectilinear or cylindrical bodies.

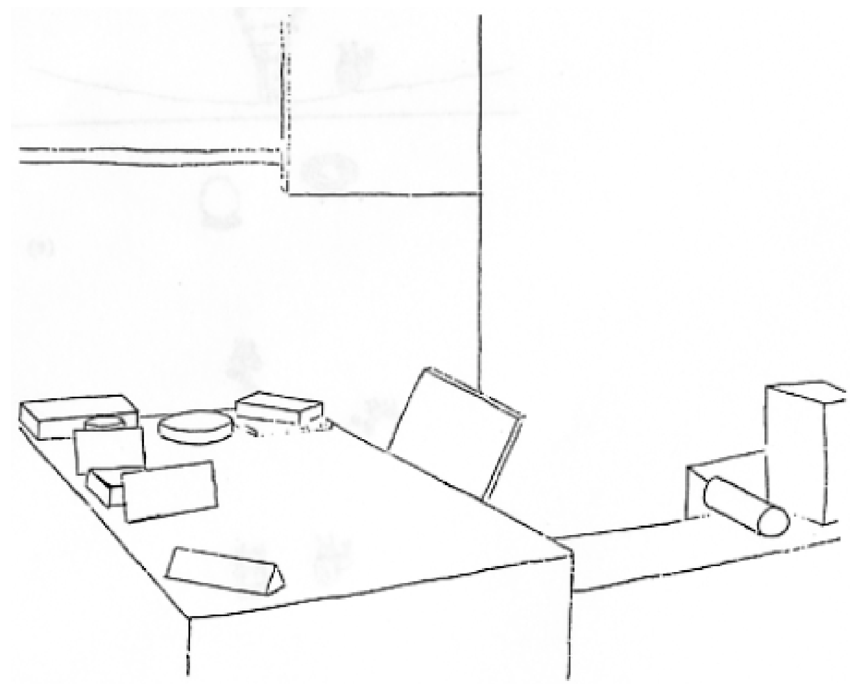
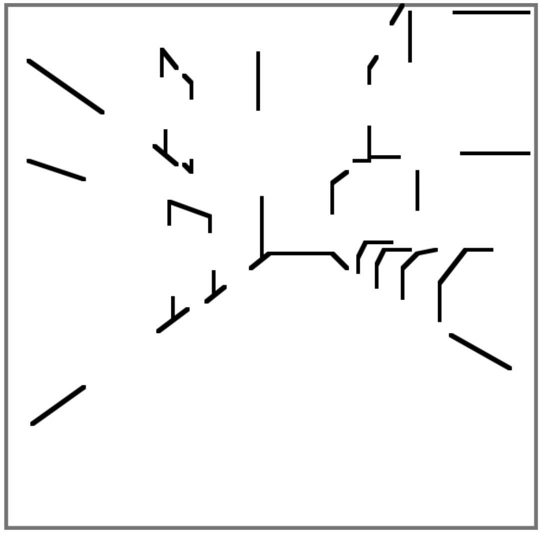


FIG. 8.24. *Office*, drawn by Robert Mezzanotte.

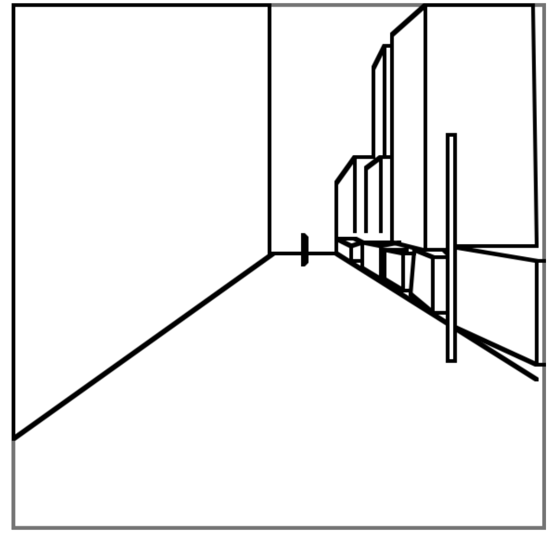
From “on the semantics of a glance at a scene”, Biederman, 1981

Examples of scene emergent features



Biederman, 1981

Suggestive edges and junctions



Biederman, 1981

Simple geometric forms

Brunet & Potter, 1969

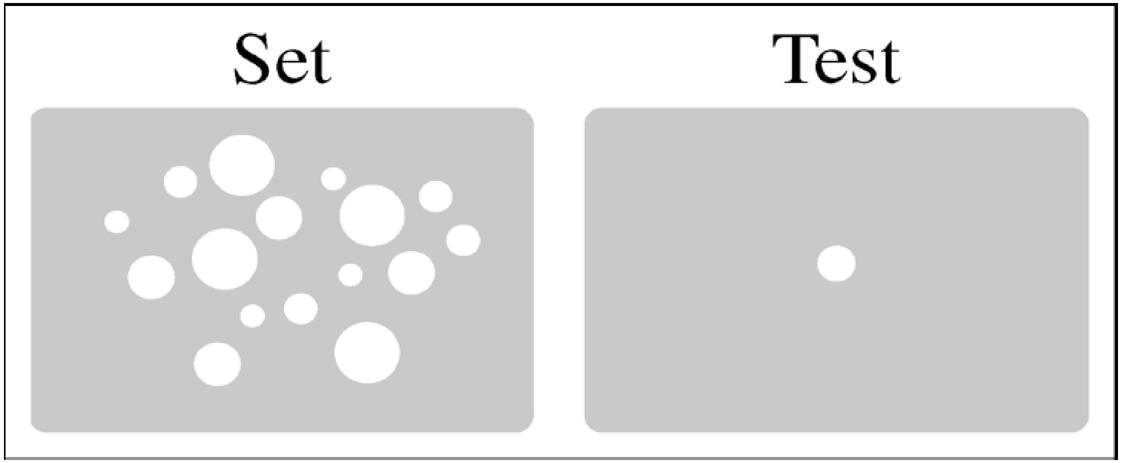
Blobs

Oliva & Torralba, 2001

Textures ~ Sketch

Ensemble statistics

Ariely, 2001, Seeing sets: Representation by statistical properties
Chong, Treisman, 2003, Representation of statistical properties
Alvarez, Oliva, 2008, 2009, Spatial ensemble statistics

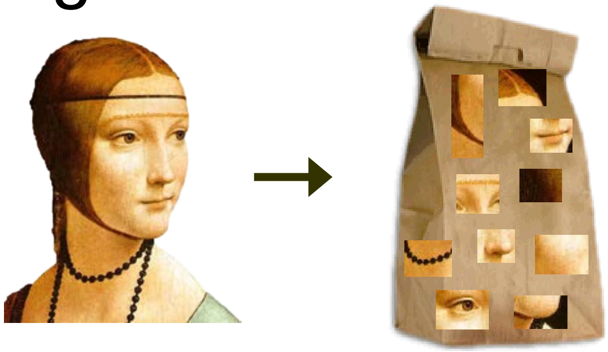


Conclusion: observers had more accurate representation of the mean than of the individual members of the set.

Global image descriptors

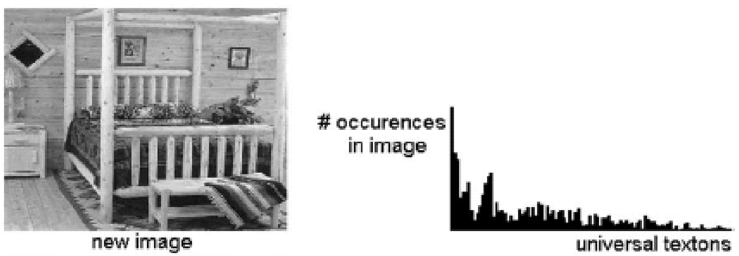
Global image descriptors

Bag of words



Sivic et. al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

Non localized textons

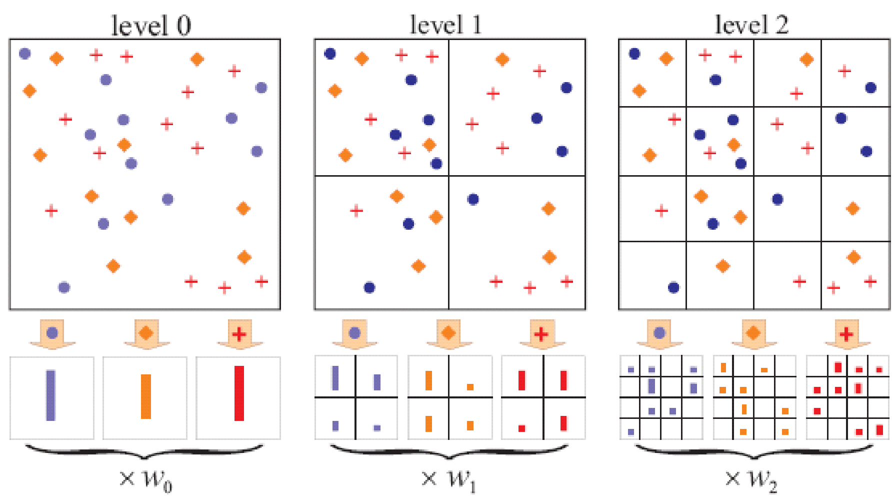


Walker, Malik. Vision Research 2004

. . .

Spatially organized textures

M. Gorkani, R. Picard, ICPR 1994
A. Oliva, A. Torralba, IJCV 2001



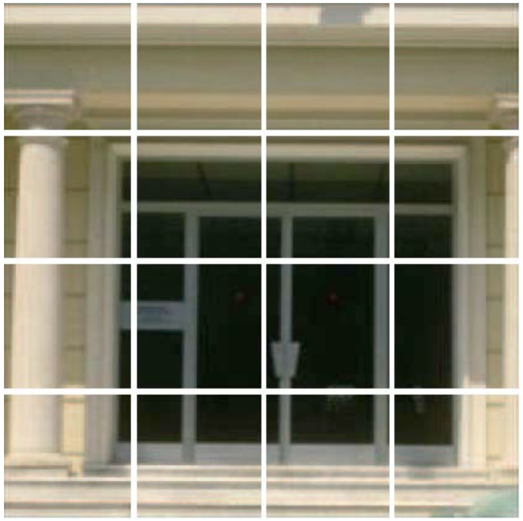
S. Lazebnik, et al, CVPR 2006

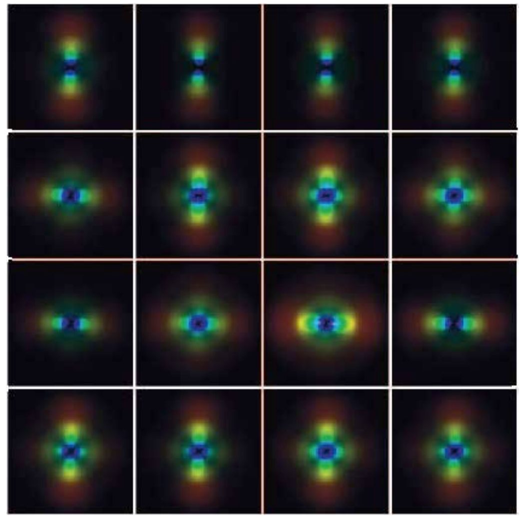
. . .

R. Datta, D. Joshi, J. Li, and J. Z. Wang, **Image Retrieval: Ideas, Influences, and Trends of the New Age**, *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1-60, 2008.

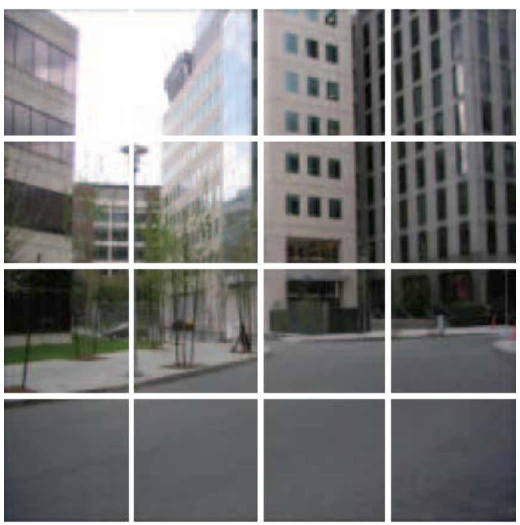
Gist descriptor

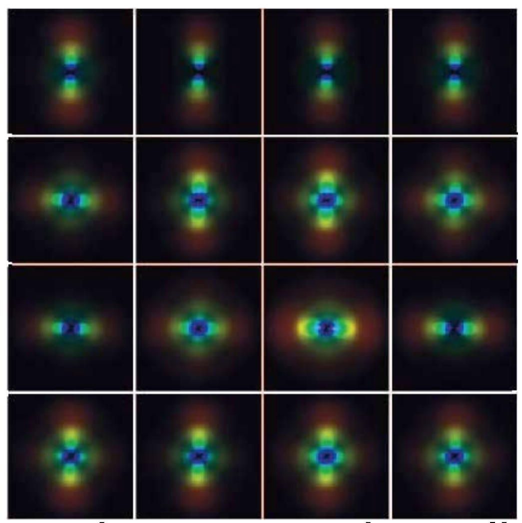
Oliva and Torralba, 2001





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



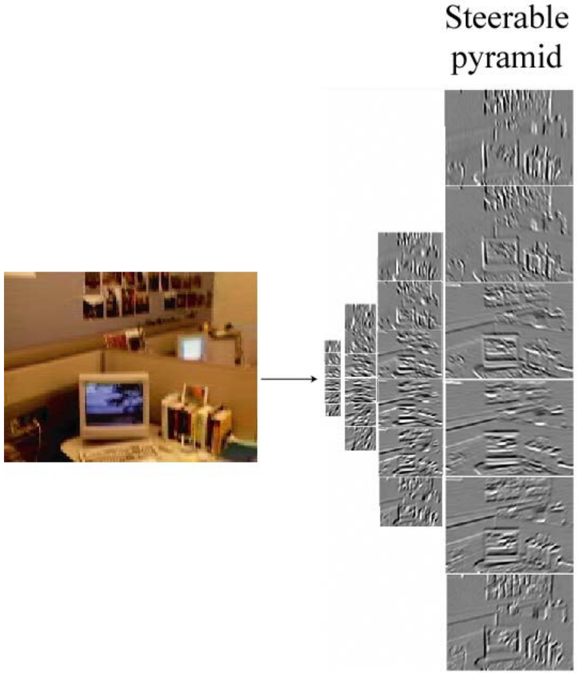


8 orientations
4 scales
x 16 bins
512 dimensions

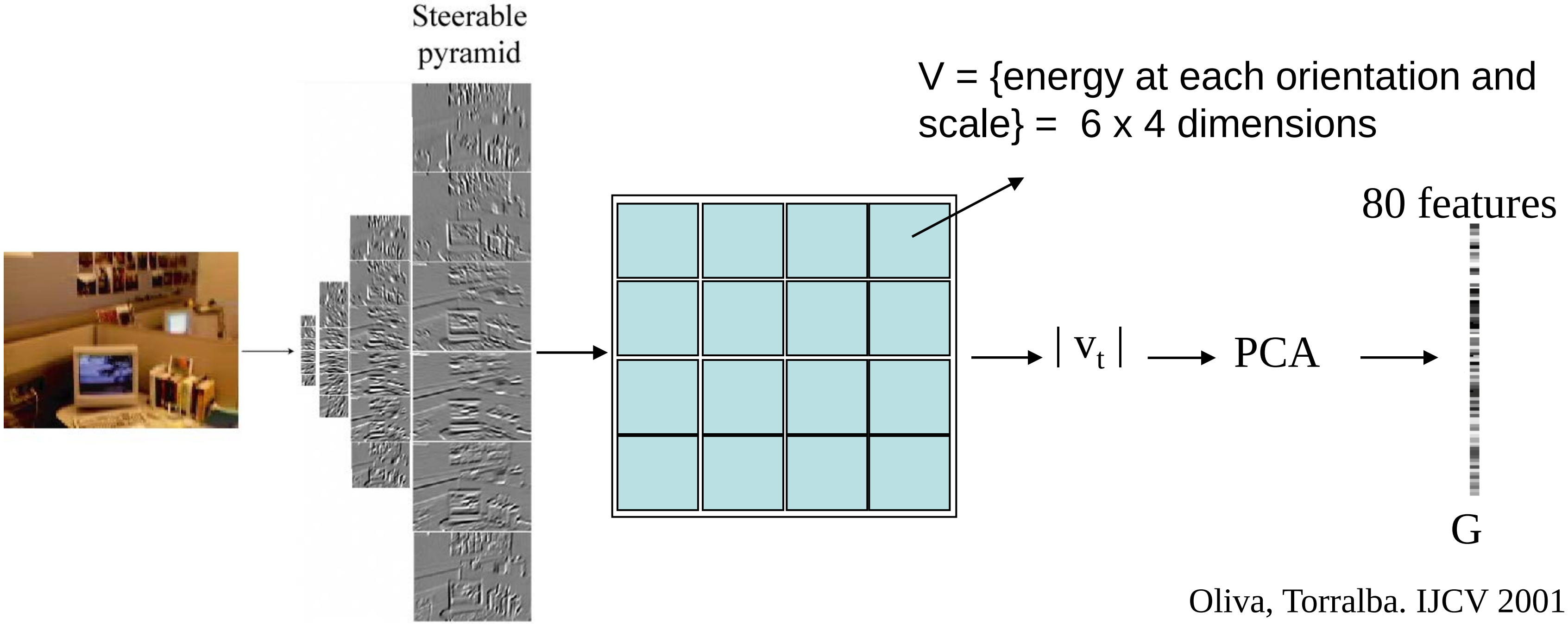
Similar to SIFT (Lowe 1999) applied to the entire image

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

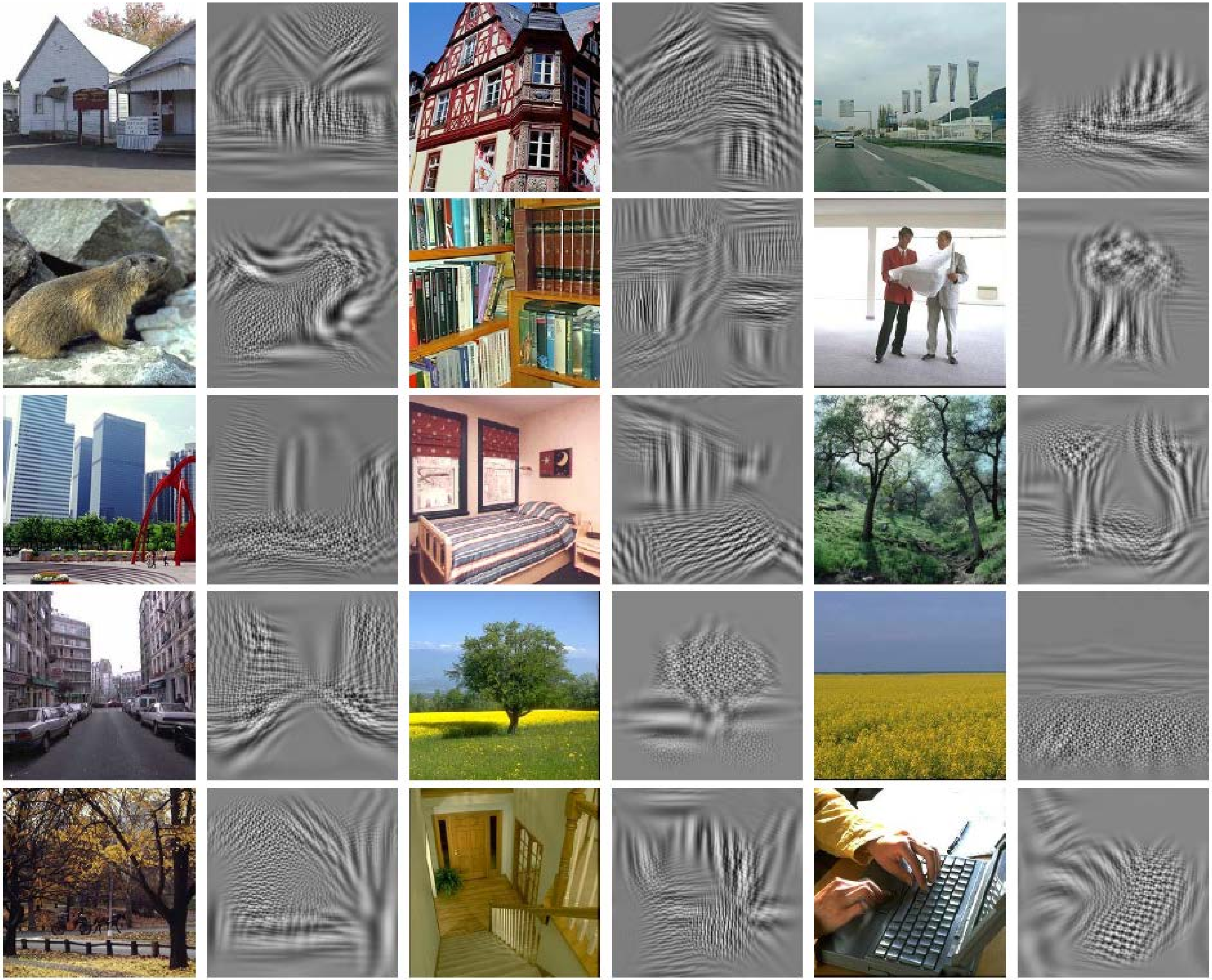


Gist descriptor



Oliva, Torralba. IJCV 2001

Example visual gists



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

Oliva & Torralba (2001)

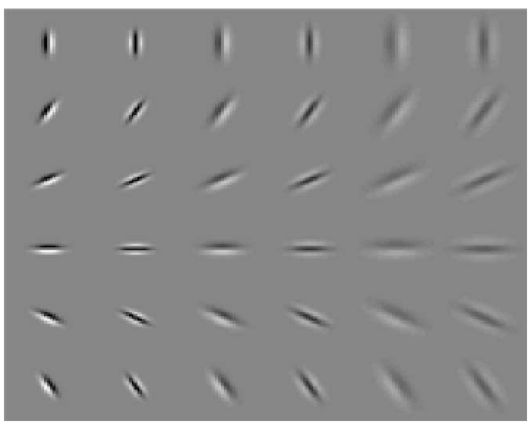
Global features

"The viewer is presented with a 'potential image', that is, a complex multiplicity of possible images, none of which ever finally resolves".

Textons

Vector of filter responses at each pixel

Kmeans over a set of vectors on a collection of images

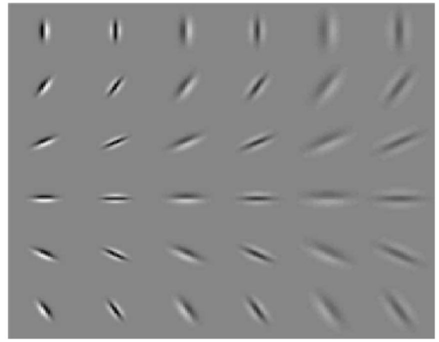


Filter bank

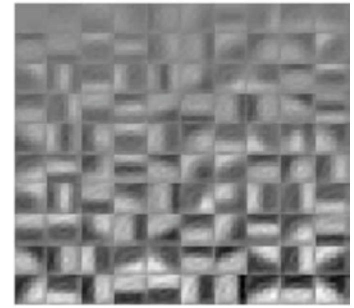
Malik, Belongie, Shi, Leung, 1999

Textons

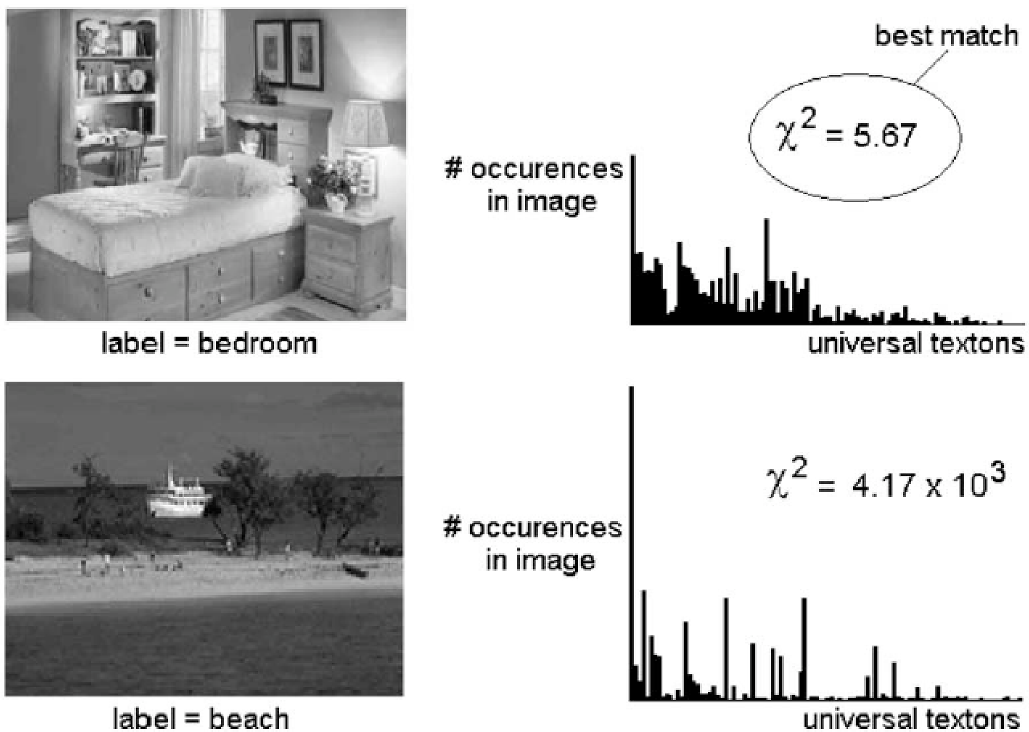
Filter bank



K-means (100 clusters)



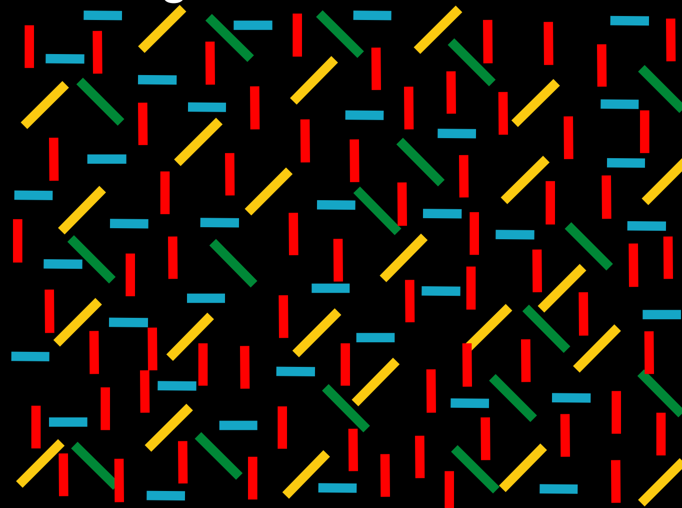
Malik, Belongie, Shi, Leung, 1999



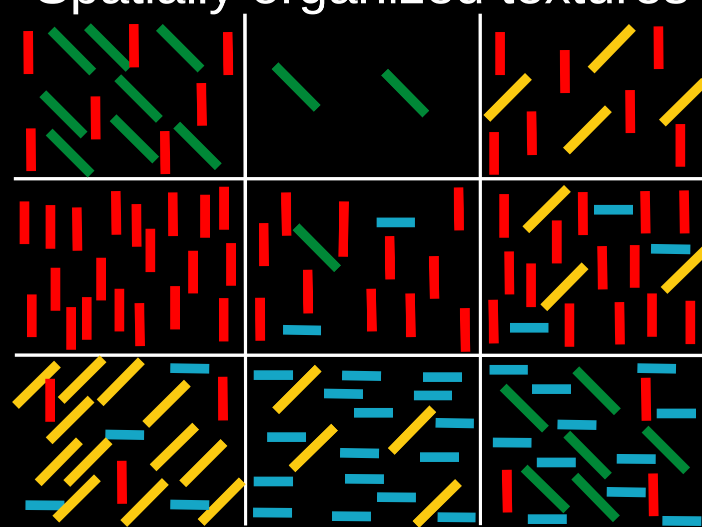
Walker, Malik, 2004

Bag of words

Bag of words model

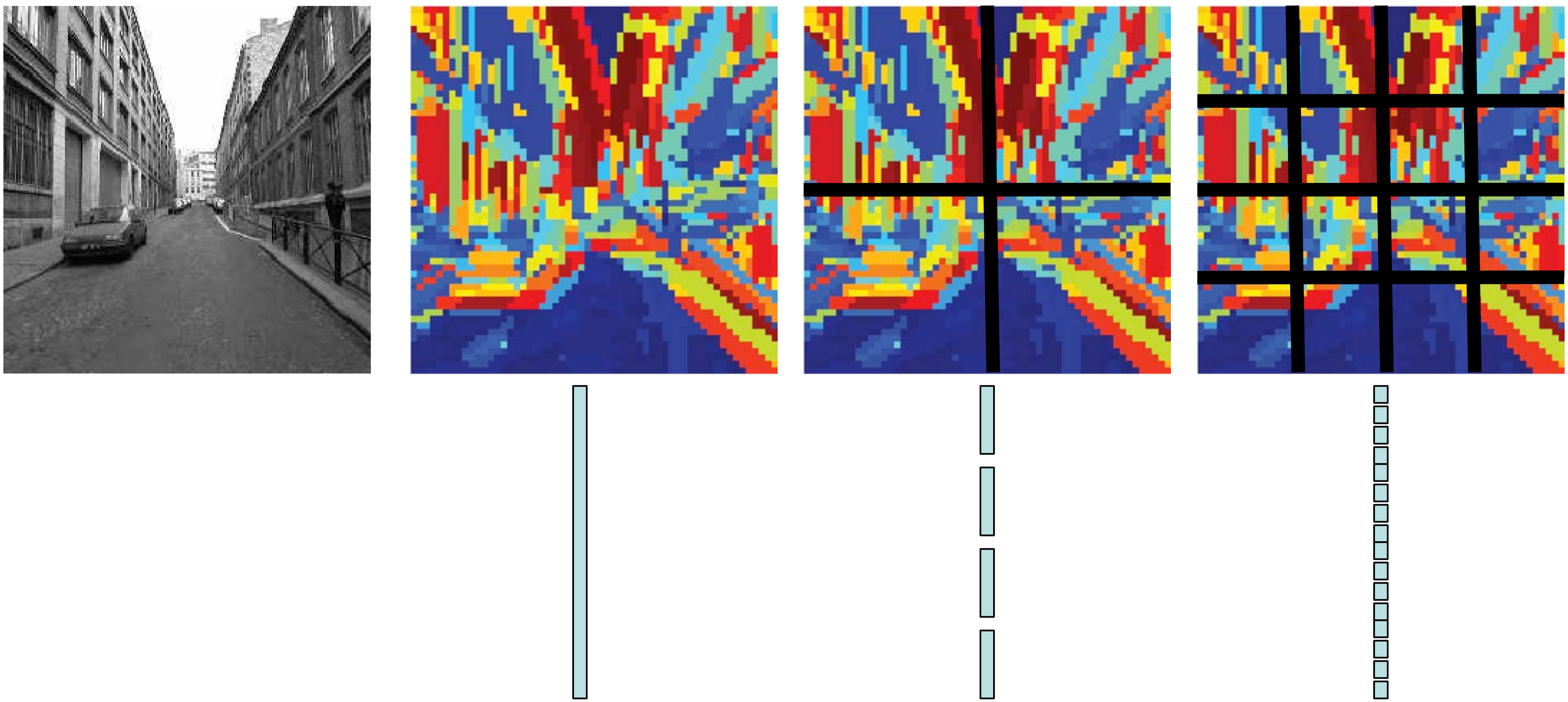


Spatially organized textures



Bag of words &
spatial pyramid matching

Sivic, Zisserman, 2003. Visual words = Kmeans of SIFT descriptors

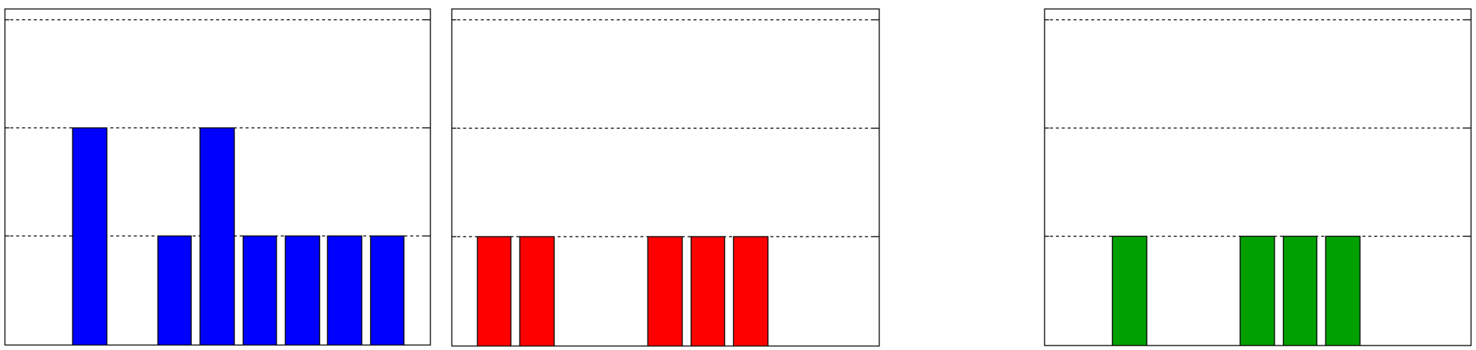


S. Lazebnik, et al, CVPR 2006

Histogram Intersection

Histogram intersection

$$\mathcal{I}\left(H(\mathbf{X}), H(\mathbf{Y})\right) = \sum_{j=1}^{r} \min\left(H(\mathbf{X})_j, H(\mathbf{Y})_j\right)$$



$H(\mathbf{X})$ $\quad$ $H(\mathbf{Y})$ $\quad$ $\mathcal{I}\left(H(\mathbf{X}), H(\mathbf{Y})\right) = 4$

Adapted from Kristen Grauman

SVM

A Support Vector Machine (SVM) learns a classifier with the form:

$$H(x) = \sum_{m=1}^{M} a_m y_m k(x, x_m)$$

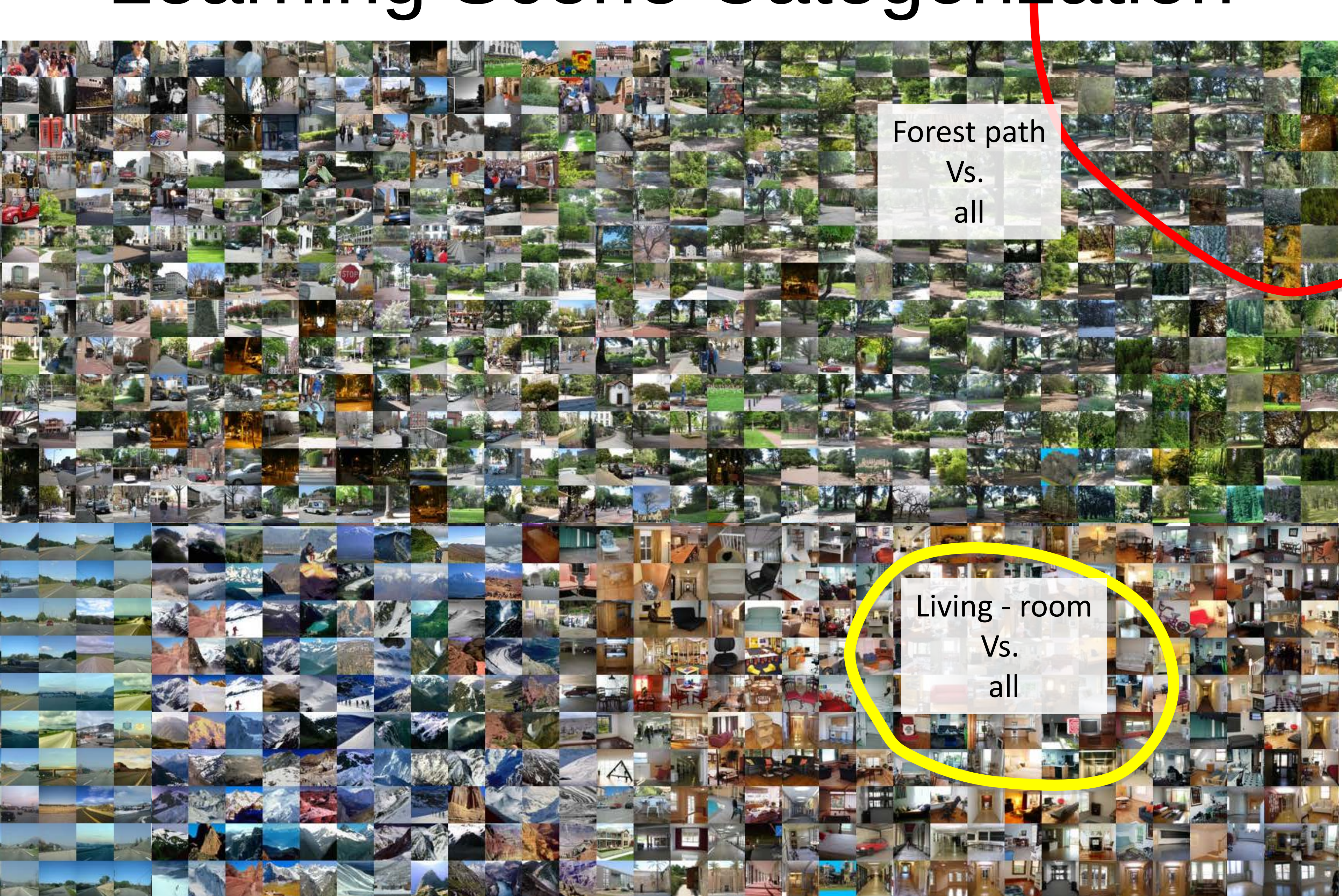
Where $\{x_m, y_m\}$, for m = 1 . . .M, are the training data with x_m being the input feature vector and y_m = +1,-1 the class label. $k(x, x_m)$ is the kernel and it can be any symmetric function satisfying the Mercer Theorem.

The classification is obtained by thresholding the value of H(x).

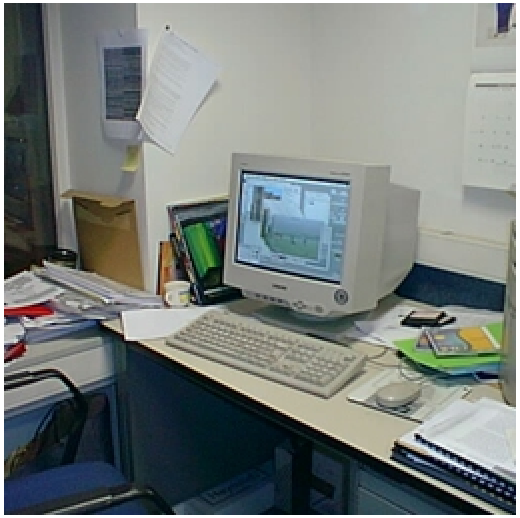
There is a large number of possible kernels, each yielding a different family of decision boundaries:

- Linear kernel: $k(x, x_m) = x^T x_m$
- Radial basis function: $k(x, x_m) = \exp(-|x - x_m|^2/\sigma^2)$.
- Histogram intersection: $k(x,x_m) = \mathrm{sum}_i(\min(x(i), x_m(i)))$

Learning Scene Categorization



The 15-scenes benchmark



Oliva & Torralba, 2001
Fei Fei & Perona, 2005
Lazebnik, et al 2006

Office

Skyscrapers

Suburb

Building facade

Coast

Forest

Bedroom

Living room

Industrial

Street

Highway

Mountain

Open country

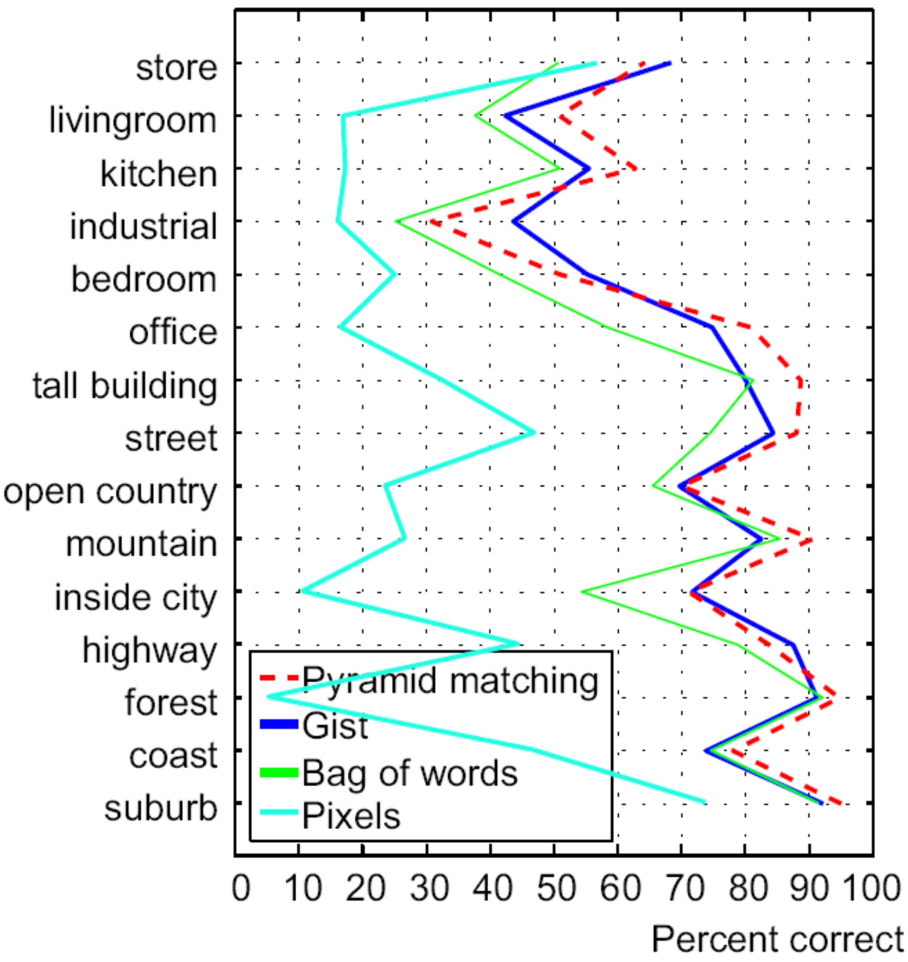
Kitchen

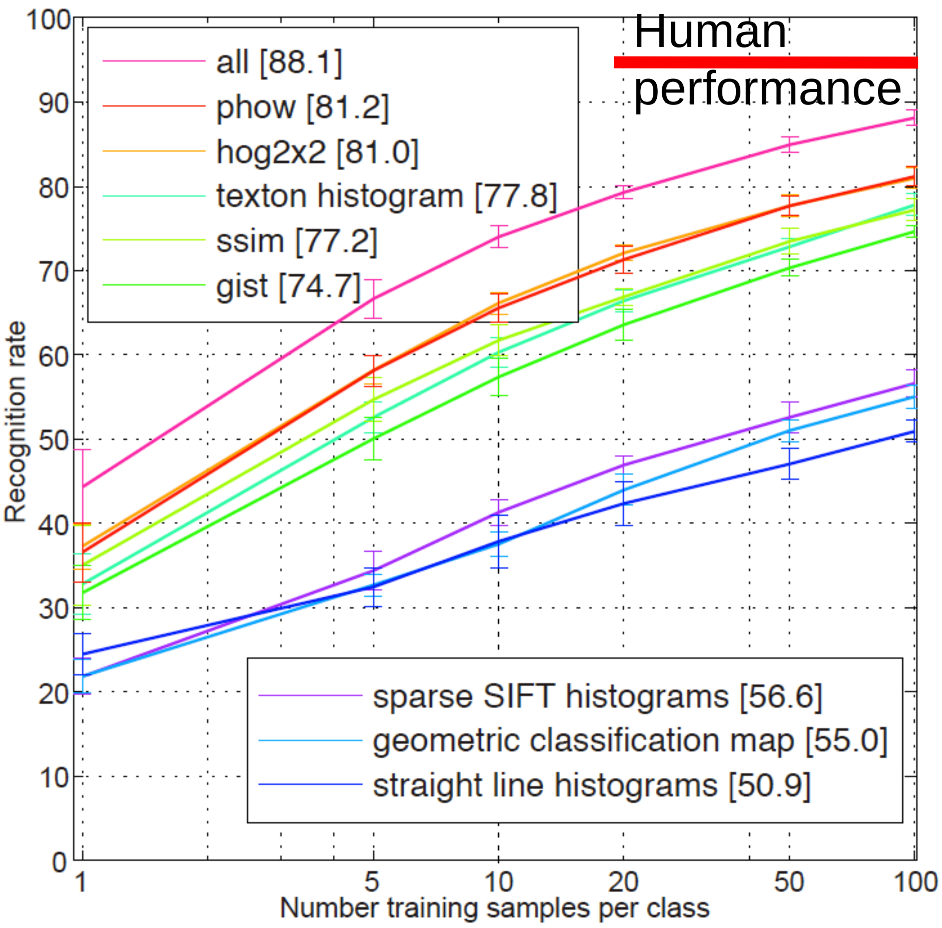
Store

Scene recognition

100 training samples per class

SVM classifier in both cases

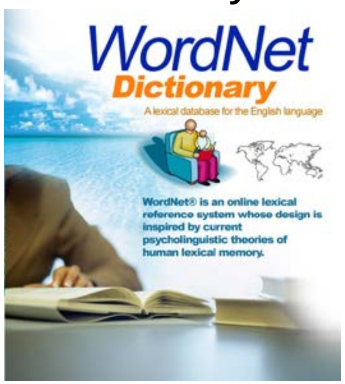




SUN Dataset Project

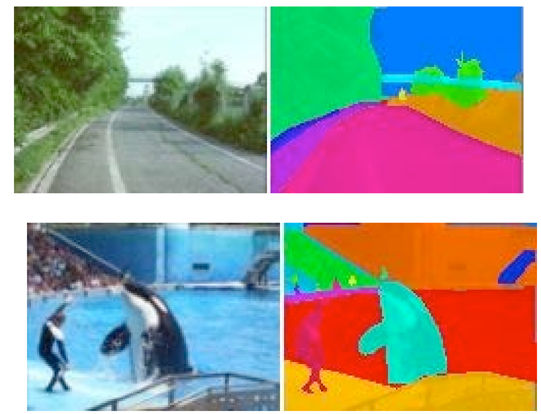
We want:
- Large variety of scene categories (we want them all)
- Lots of objects categories
- Multi-object scenes

1. We take all scene words from a dictionary



2. We download images and clean the categories

3. We segment all the images

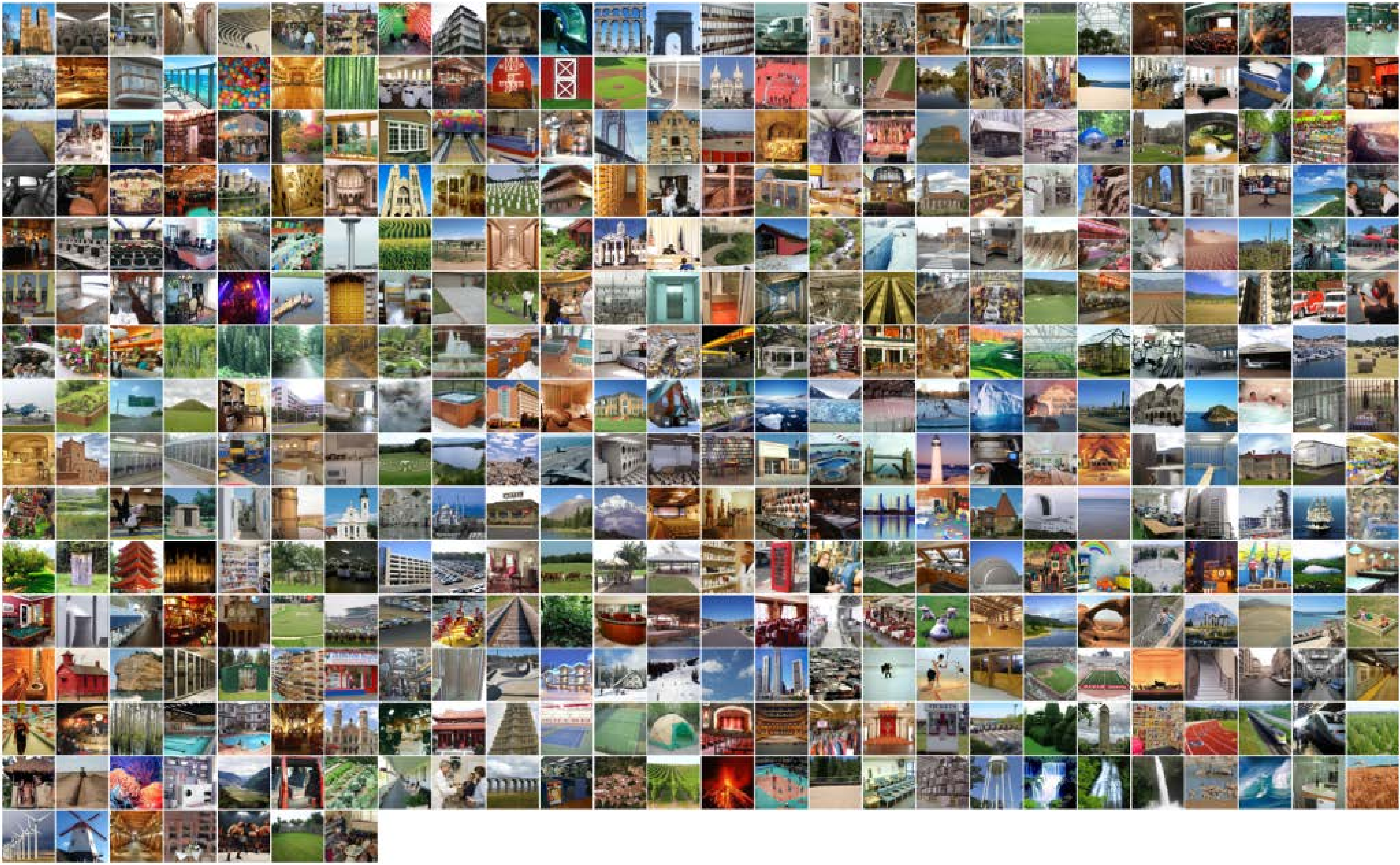


Krista Ehinger

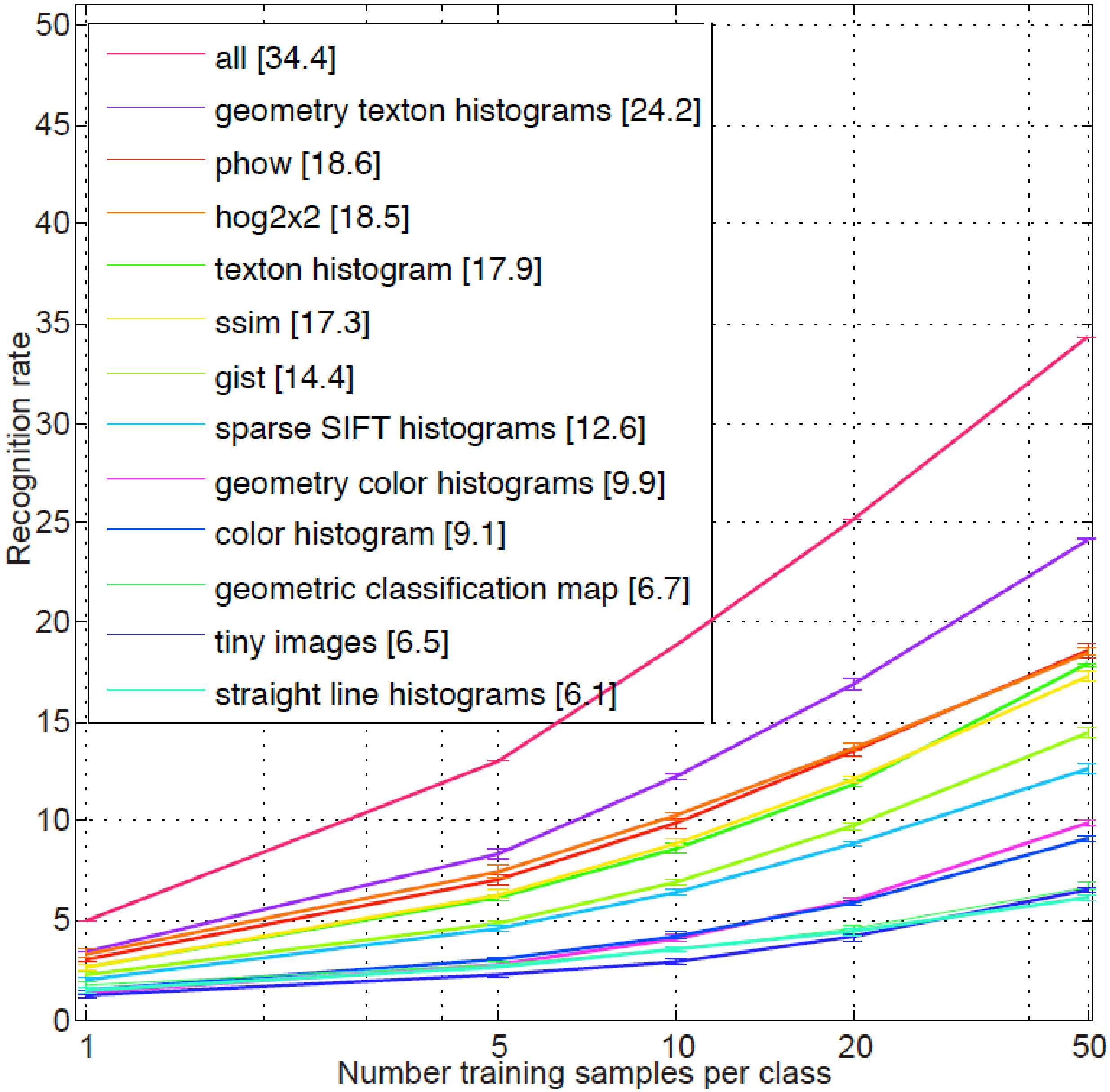
Jianxiong Xiao

Xiao, Hays, Ehinger, Oliva, Torralba; CVPR 2010

397 Well-sampled Categories



Performance with 400 categories



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010

Training images

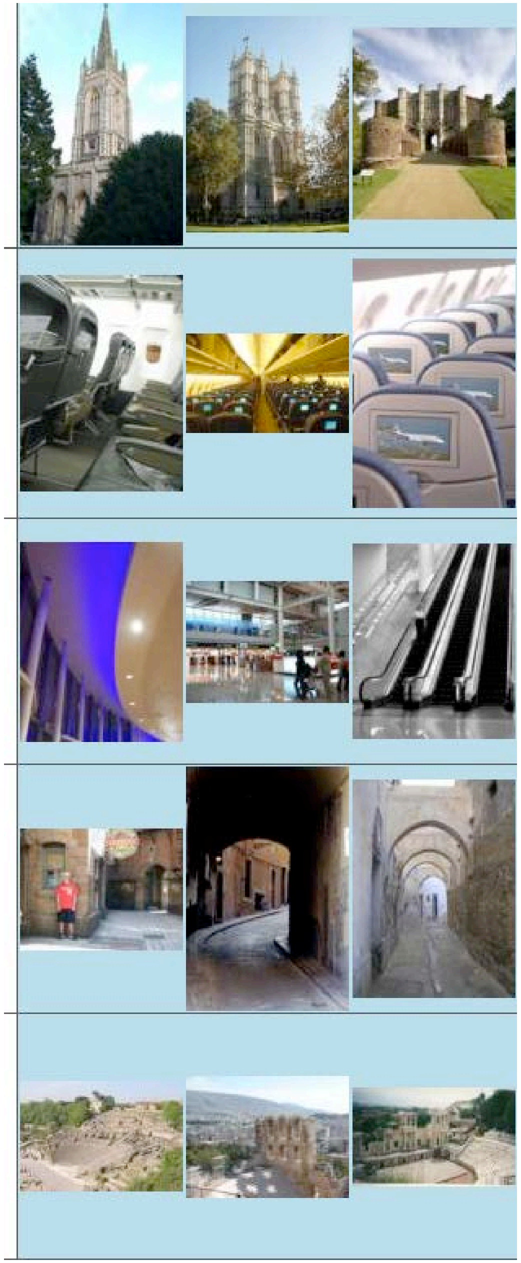
Abbey

Airplane cabin

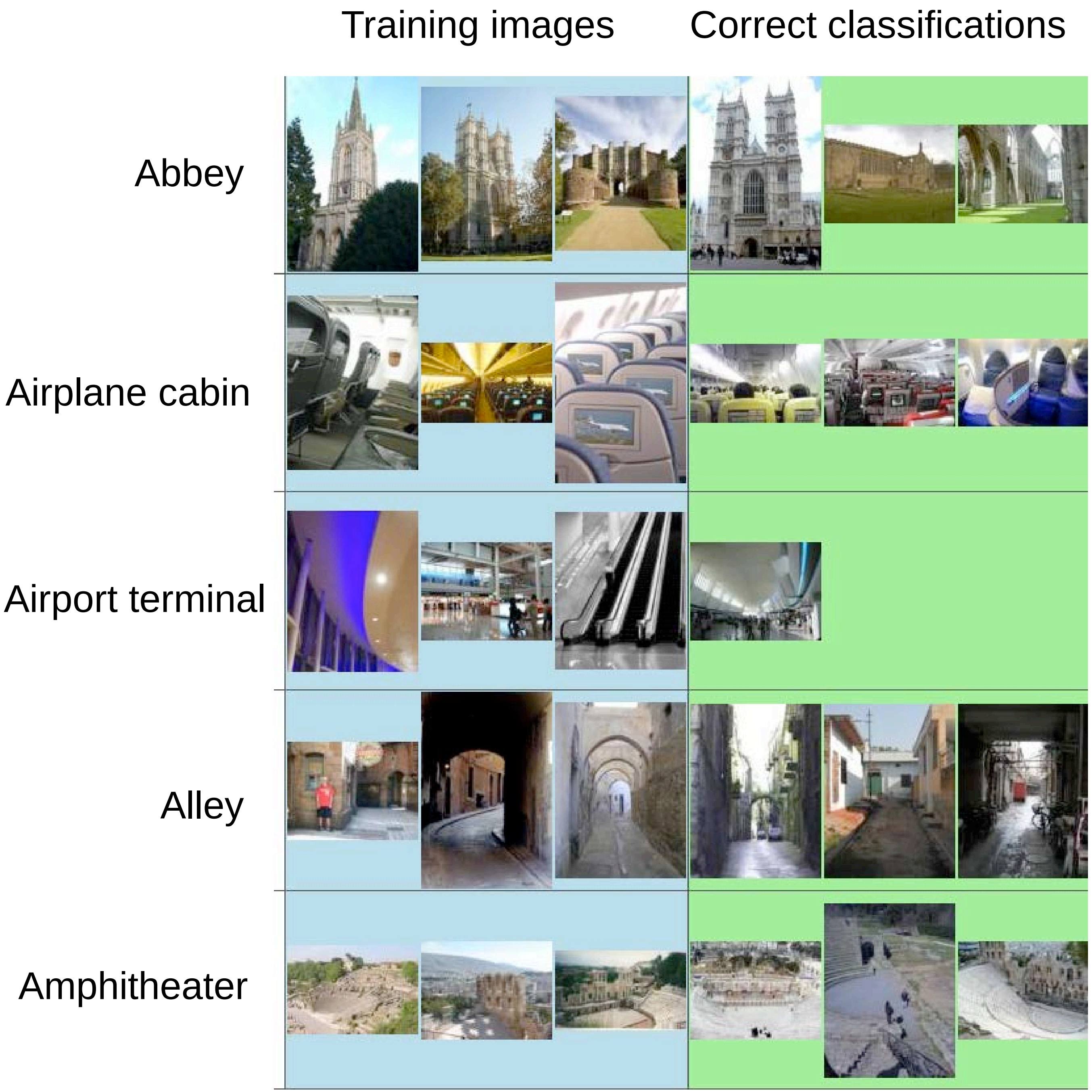
Airport terminal

Alley

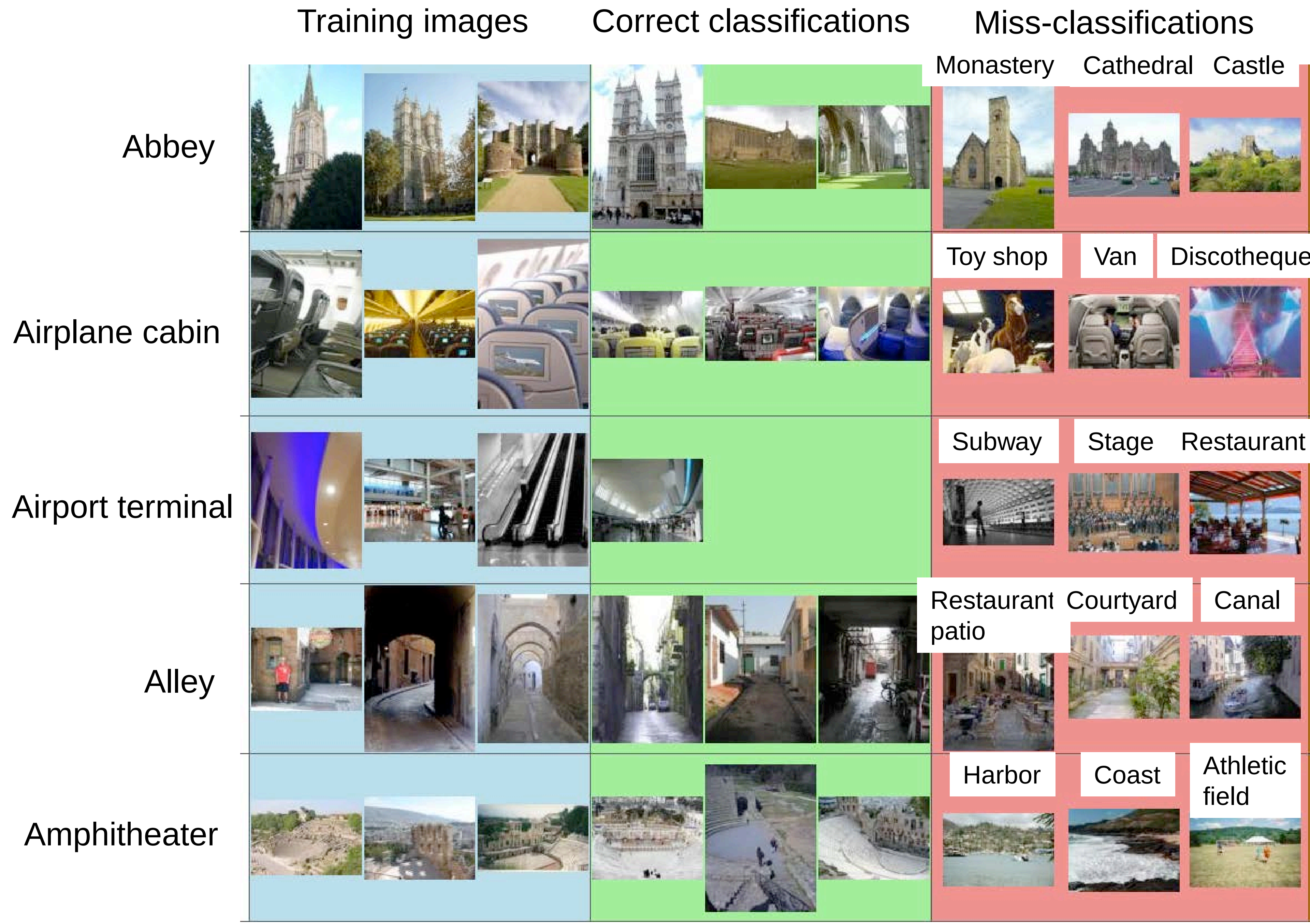
Amphitheater



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010



Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010

Categories or a continuous space?

Check poster by Malisiewicz, Efros

Categories or a continuous space?

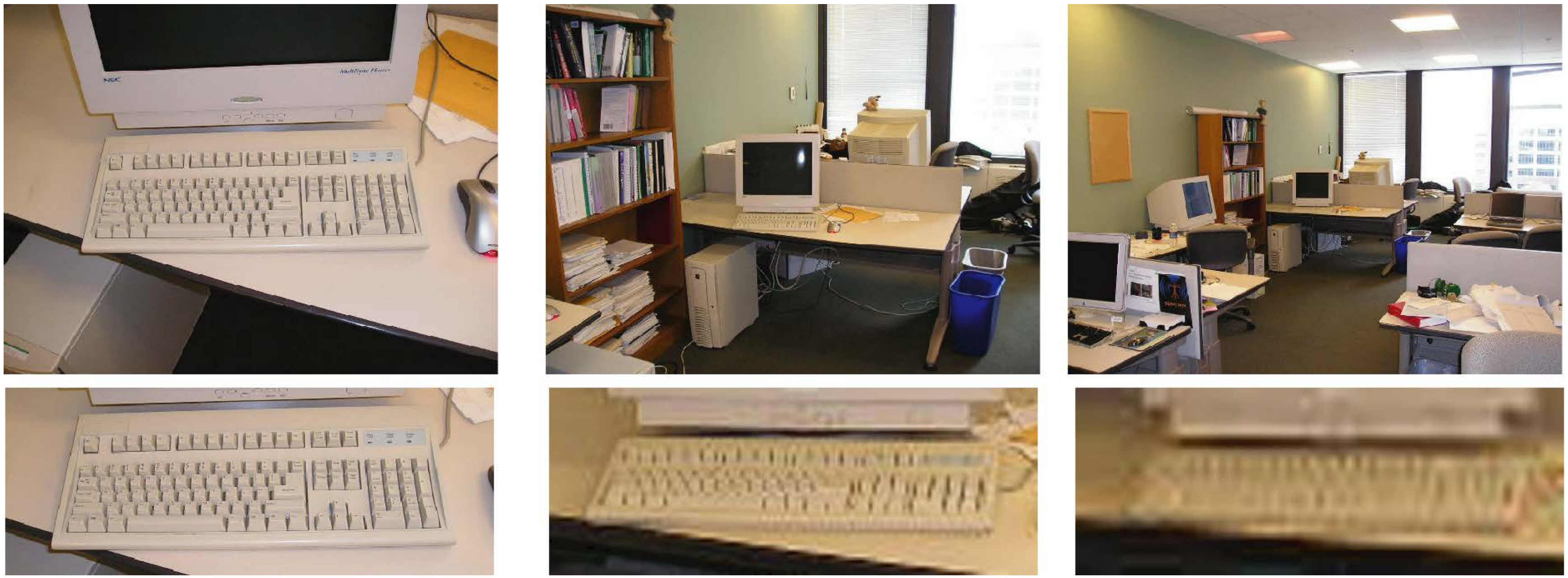
From the city to the mountains in 10 steps

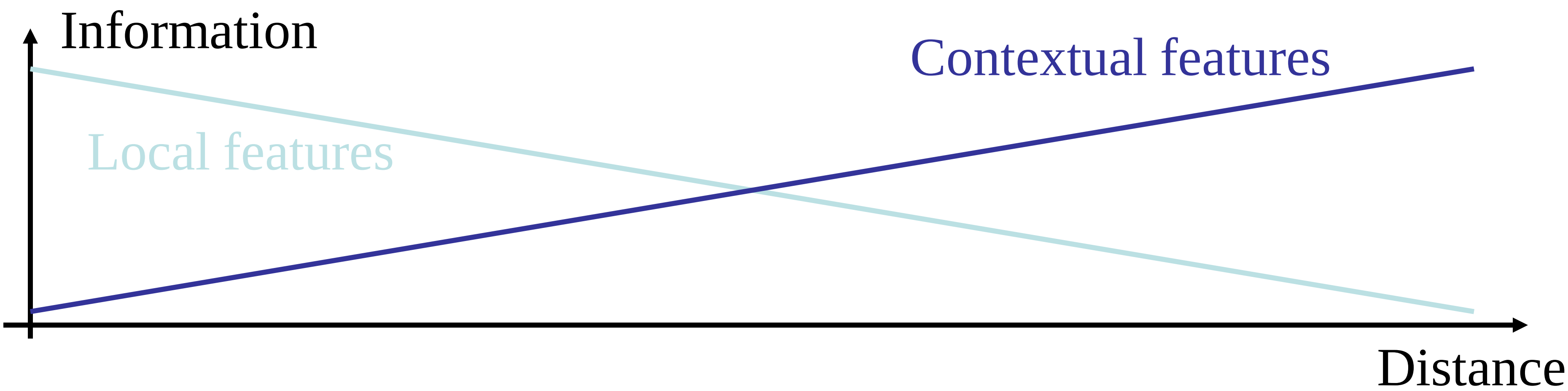
Objects in context

Is local information enough?

Is local information even enough?

Is local information even enough?





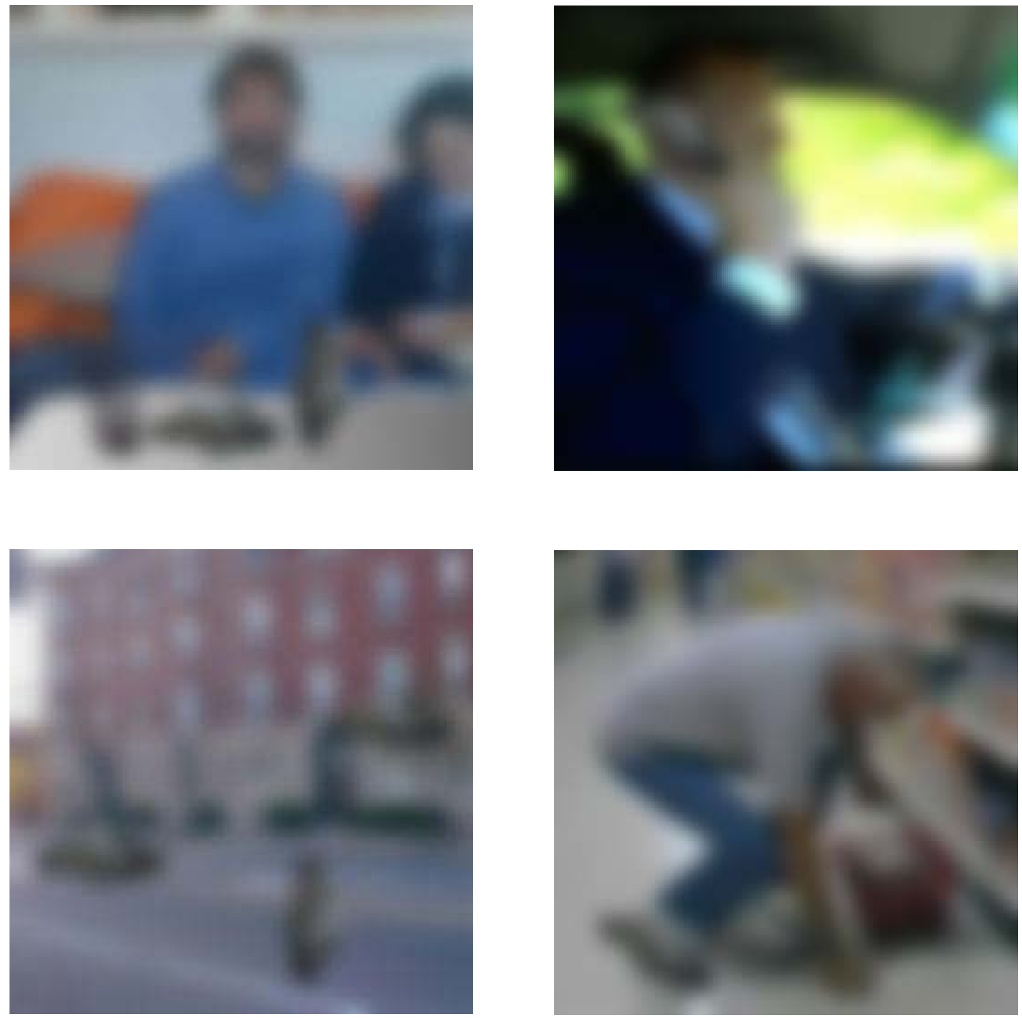
The system does not care about the scene, but we do…

We know there is a keyboard present in this scene even if we cannot see it clearly.

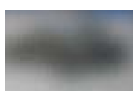
We know there is no keyboard present in this scene

… even if there is one indeed.

The multiple personalities of a blob

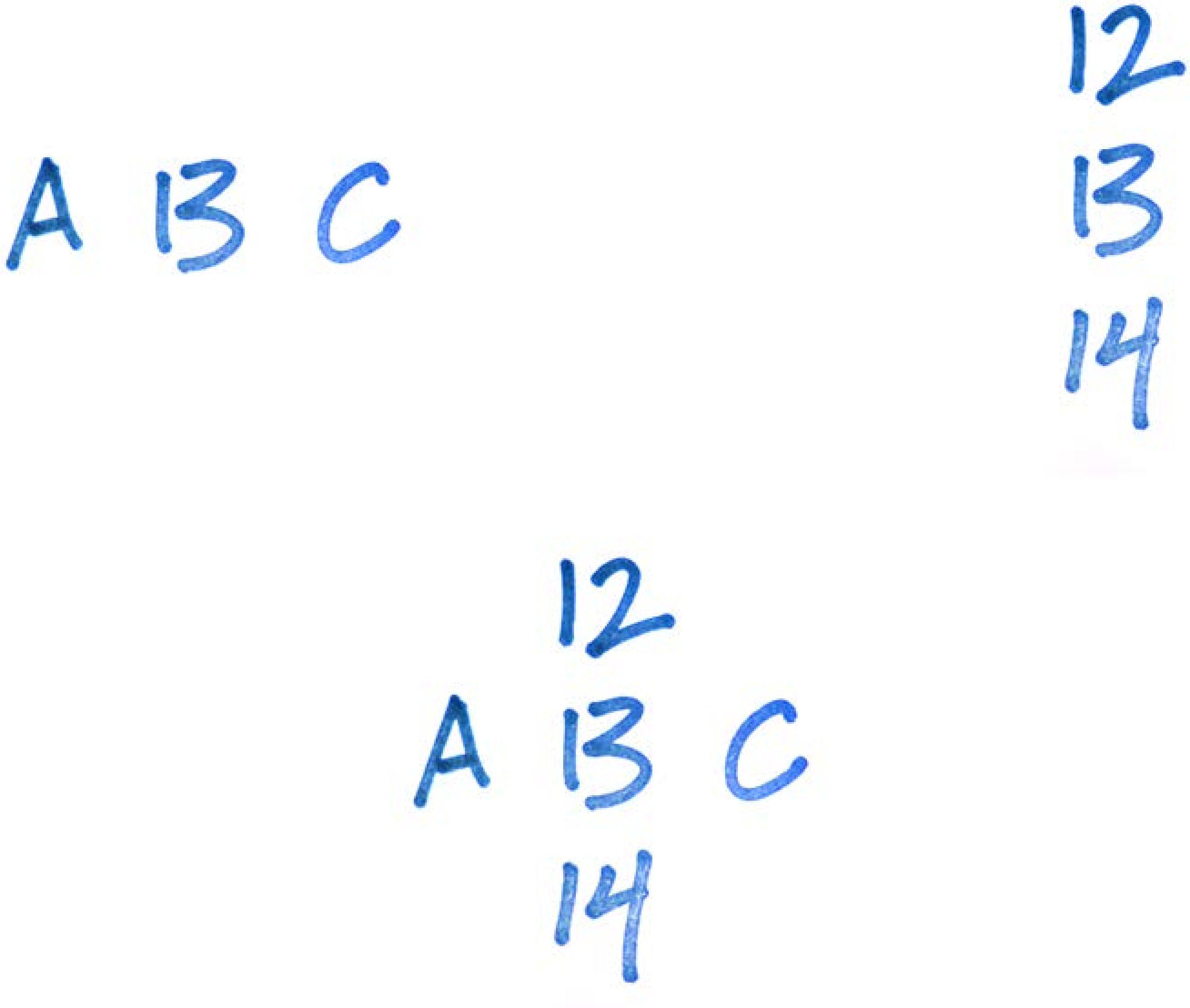


The multiple personalities of a blob



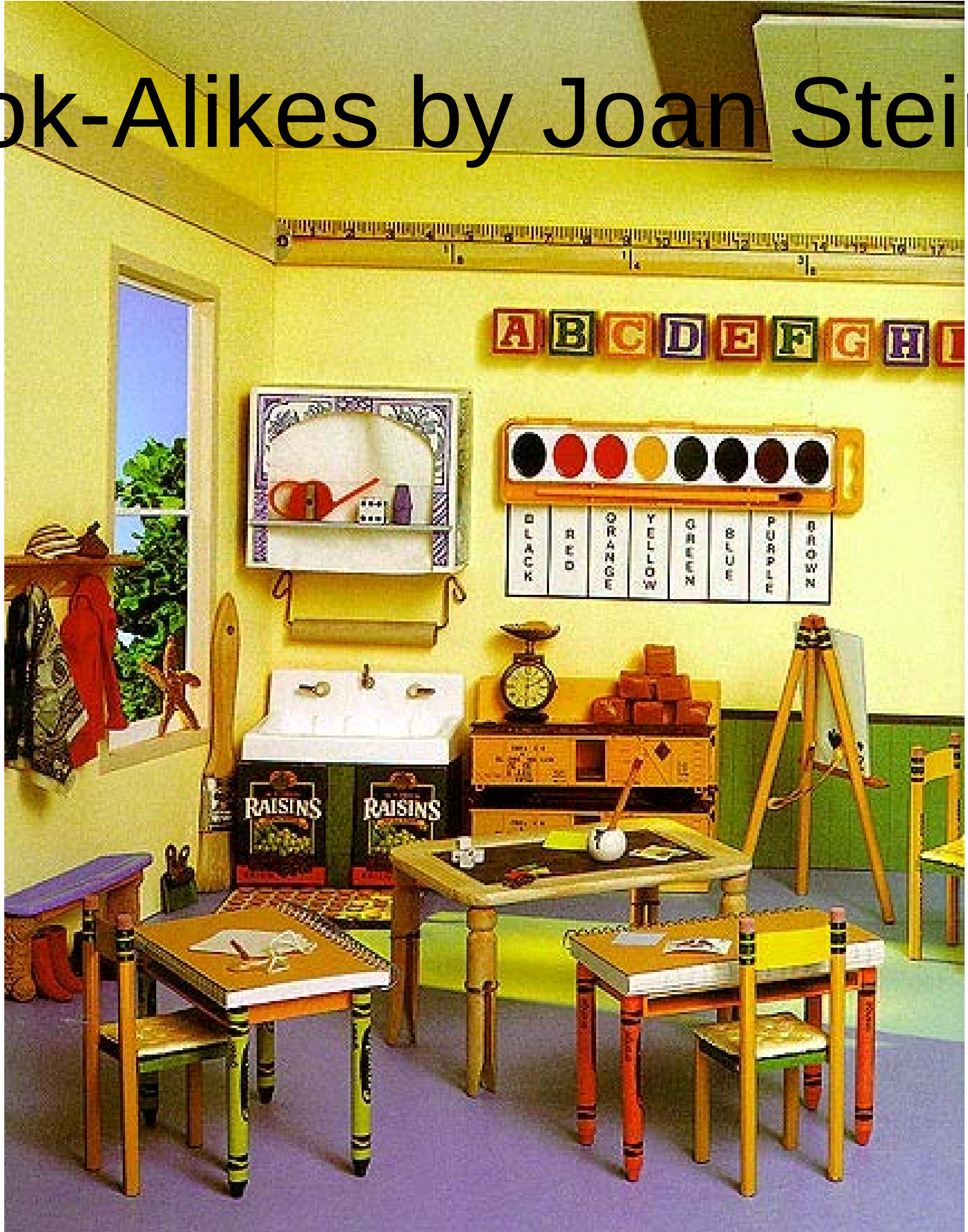
A B C

12
13
14


A B C
12
13
14
12
A B C
14

Look-Alikes by Joan Steiner

Look-Alikes by Joan Steiner



Look-Alikes by Joan Steiner

The importance of context

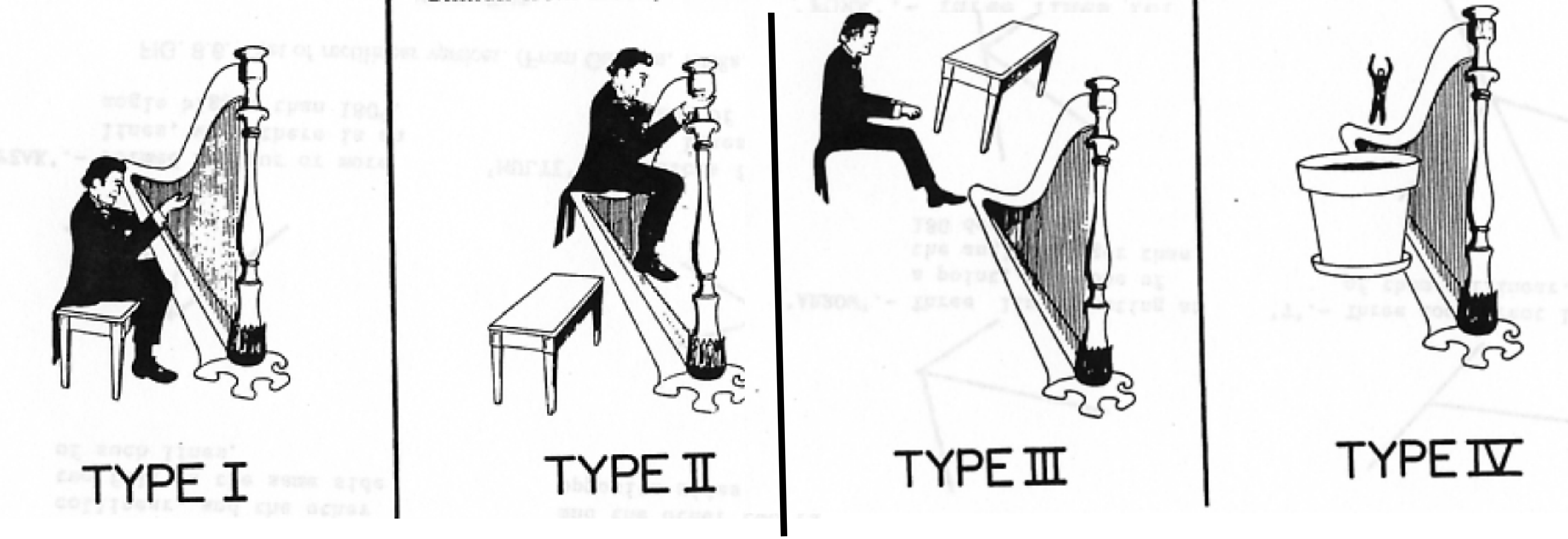
- Cognitive psychology
  - Palmer 1975
  - Biederman 1981
  - …

- Computer vision
  - Noton and Stark (1971)
  - Hanson and Riseman (1978)
  - Barrow & Tenenbaum (1978)
  - Ohta, kanade, Skai (1978)
  - Haralick (1983)
  - Strat and Fischler (1991)
  - Bobick and Pinhanez (1995)
  - Campbell et al (1997)

Class	Context elements	Operator
SKY	ALWAYS	ABOVE-HORIZON
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	WHITE
SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
SKY	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
SKY	RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)	SIMILAR-COLOR
GROUND	CAMERA-IS-HORIZONTAL	HORIZONTALLY-STRIATED
GROUND	CAMERA-IS-HORIZONTAL	NEAR-BOTTOM
GROUND	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGES-FORM-HORIZONTA
GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTA
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(geometric-horizon) ∧ ¬ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
GROUND	TIME-IS-DAY	DARK

Objects and Scenes

Stimuli from Hock, Romanski, Galie, and Williams (1978).



Biederman's violations (1981):

1. *Support* (e.g., a floating fire hydrant). The object does not appear to be resting on a surface.
2. *Interposition* (e.g., the background appearing through the hydrant). The objects undergoing this violation appear to be transparent or passing through another object.
3. *Probability* (e.g., the hydrant in a kitchen). The object is unlikely to appear in the scene.
4. *Position* (e.g., the fire hydrant on top of a mailbox in a street scene). The object is likely to occur in that scene, but it is unlikely to be in that particular position.
5. *Size* (e.g., the fire hydrant appearing larger than a building). The object appears to be too large or too small relative to the other objects in the scene.

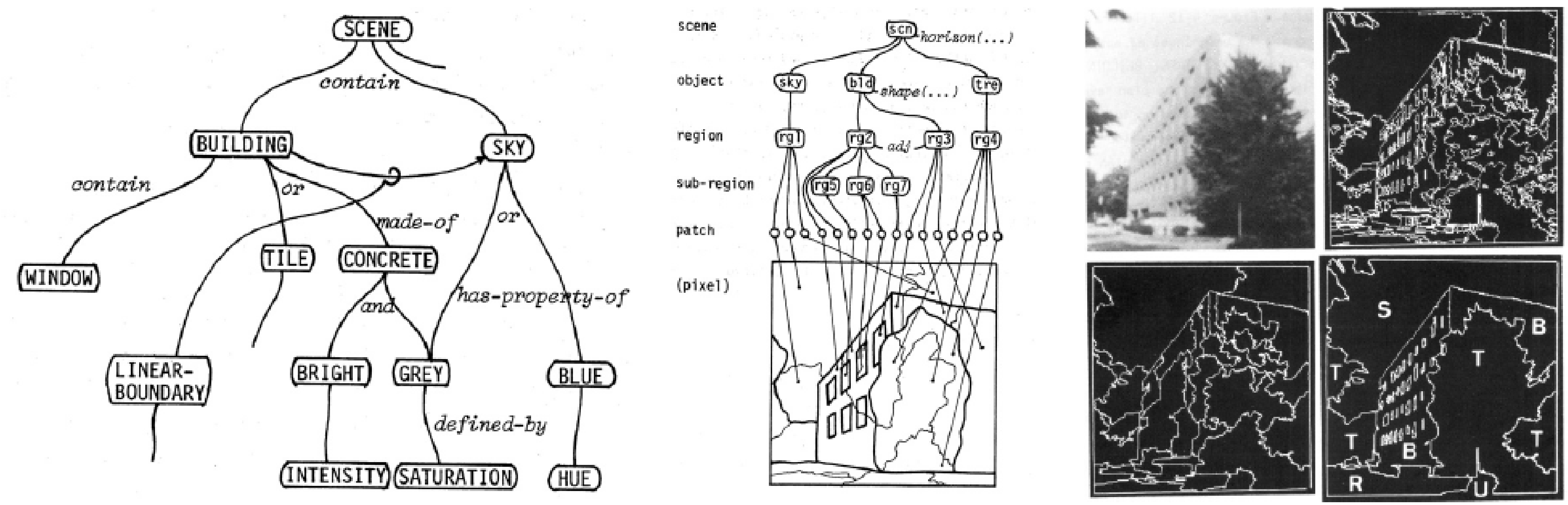
CONDOR system

Strat and Fischler (1991)

Class	Context elements	Operator
SKY	ALWAYS	ABOVE-HORIZON
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	WHITE
SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
SKY	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
SKY	RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)	SIMILAR-COLOR
GROUND	CAMERA-IS-HORIZONTAL	HORIZONTALLY-STRIATED
GROUND	CAMERA-IS-HORIZONTAL	NEAR-BOTTOM
GROUND	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGES-FORM-HORIZONT/
GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTA
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(geometric-horizon) ∧ ¬ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
GROUND	TIME-IS-DAY	DARK

- Guzman (*SEE*), 1968
- Noton and Stark 1971
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Marr, 1982
- Ohta & Kanade, 1978
- Yakimovsky & Feldman, 1973

An Age of Scene Understanding



(a) Bottom-up process (b) Top-down process (c) Result

[Ohta & Kanade 1978]

- Guzman (*SEE*), 1968
- Noton and Stark 1971
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Marr, 1982
- Ohta & Kanade, 1978
- Yakimovsky & Feldman, 1973

objects image

$$p(O \mid I) \; \alpha \; p(I|O) \; p(O)$$

Object model

Context model

$$p(O \mid I) \ \alpha \ p(I|O) \ p(O)$$

Object model

Context model

Full joint

Scene model

Approx. joint

$$p(O \mid I)\ \alpha\ p(I|O)\ p(O)$$

Object model

Context model

Full joint

Scene model

Approx. joint

$$p(O \mid I) \; \alpha \; p(I|O) \, p(O)$$

Object model

Context model

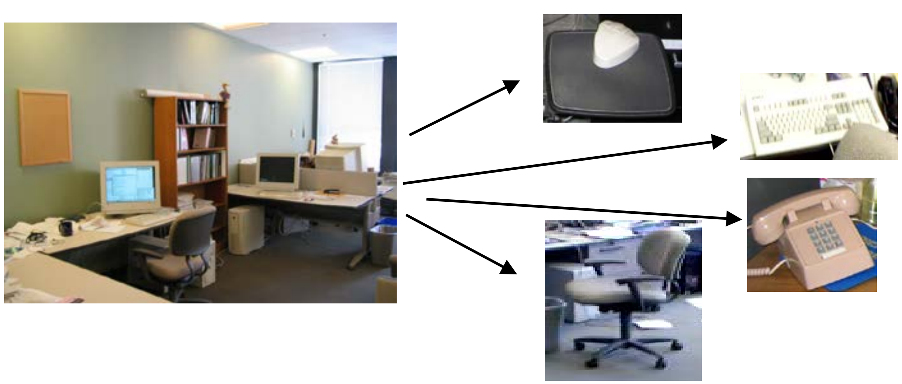
Full joint

Scene model

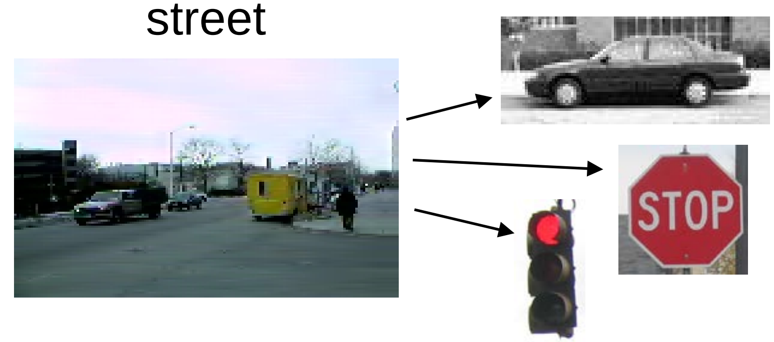
Approx. joint

$$p(O) = \sum_s \prod_i p(Oi|S=s) \, p(S=s)$$

office



street



$$p(O \mid I) \ \alpha \ p(I|O) \ p(O)$$

Object model

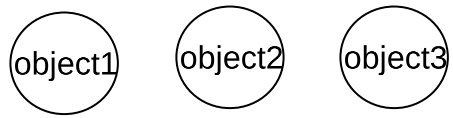
Context model

Full joint

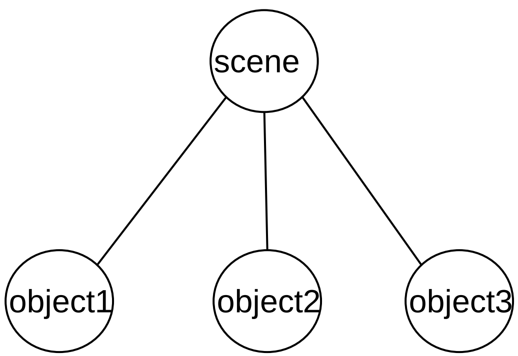
Scene model

Approx. joint

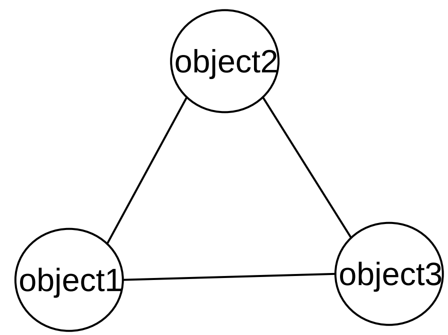
Context models



Independent model

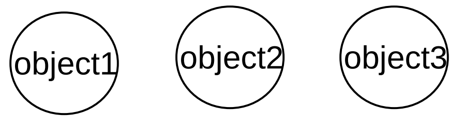


Objects are correlated via the scene

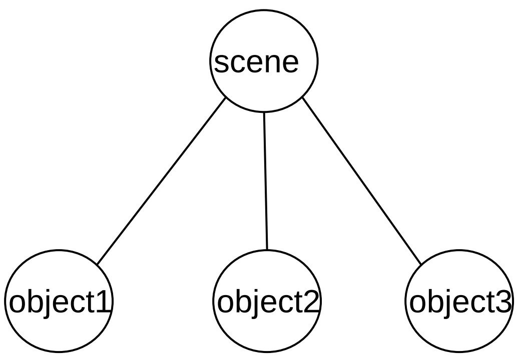


Dependencies among objects

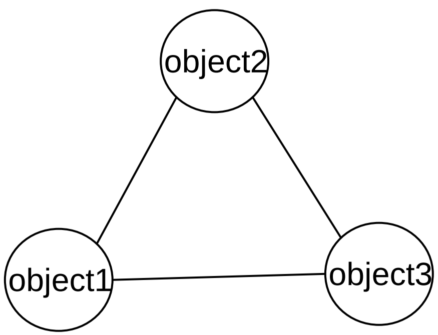
Context models



Independent model



Objects are correlated via the scene

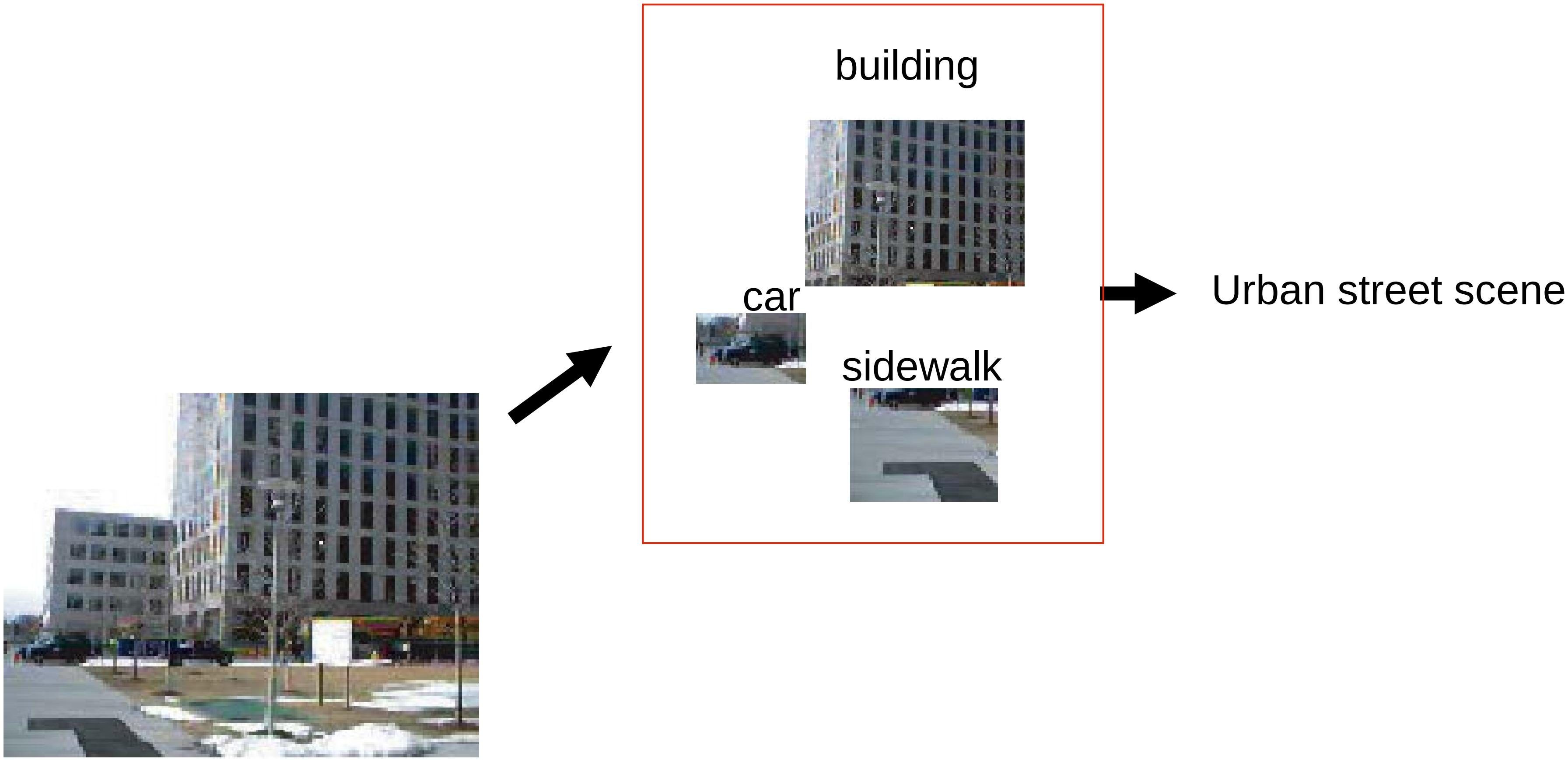


Dependencies among objects

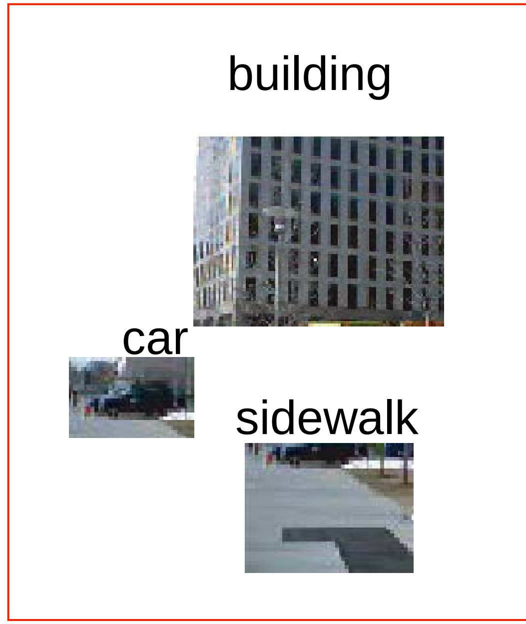
Global precedence

Forest Before Trees: The Precedence of Global Features in Visual Perception
Navon (1977)

Global and local representations

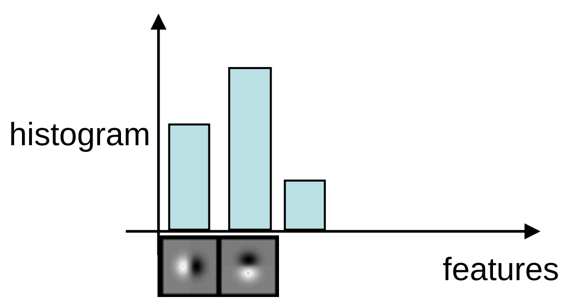


Global and local representations



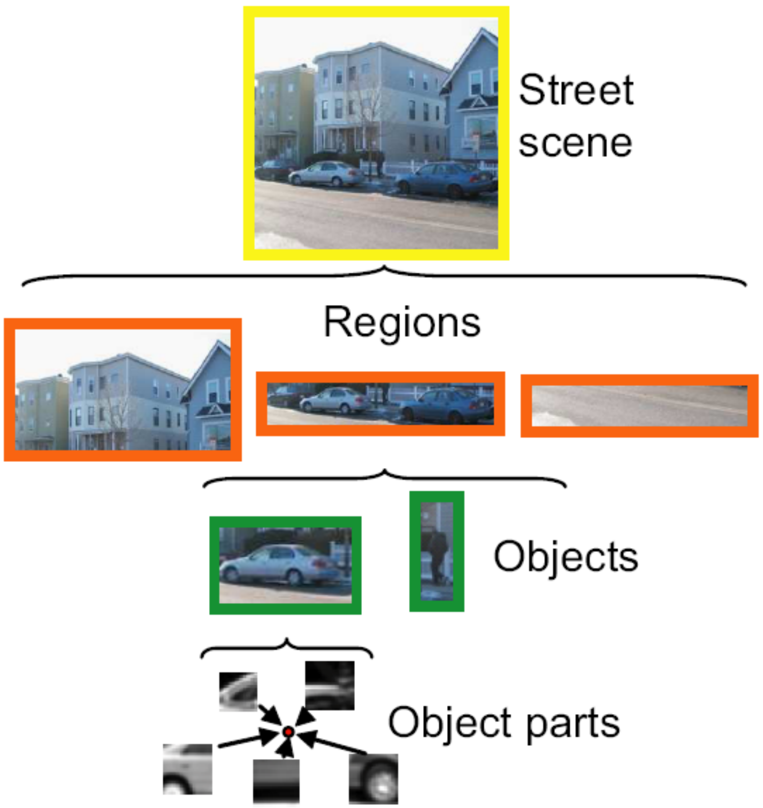
Urban street scene

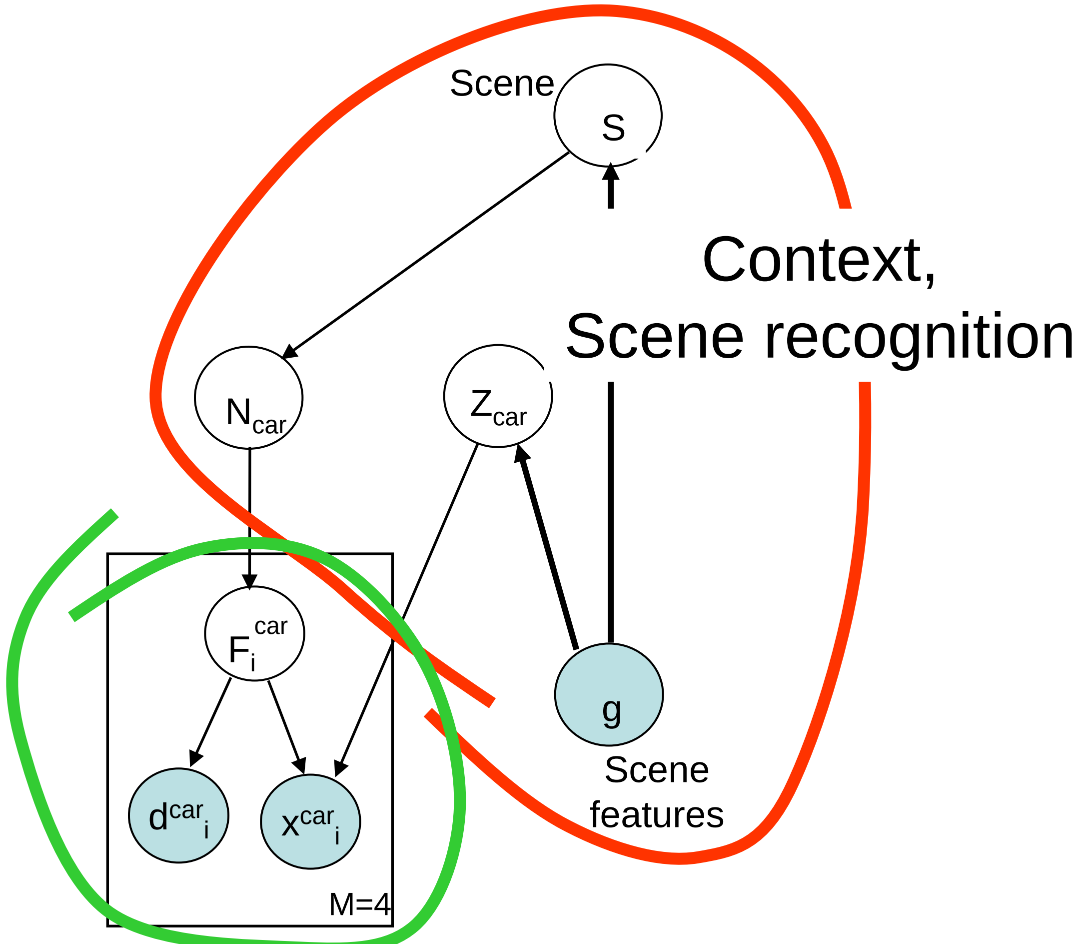
Image index: Summary statistics,
configuration of textures



Urban street scene

An integrated model of Scenes, Objects, and Parts

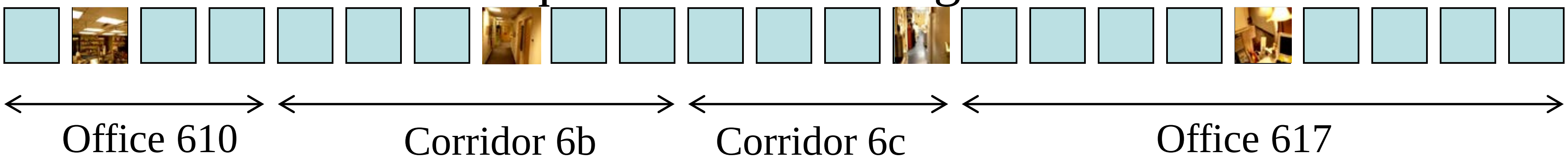




Multiclass and pose invariant
object detection,

Context-based vision system for place and object recognition

We use 17 annotated sequences for training



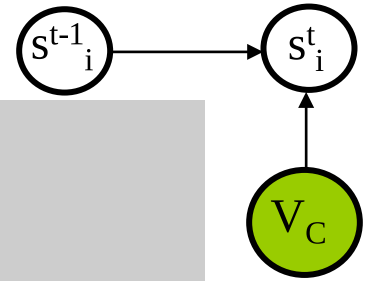
- Hidden states = location (63 values)
- Observations = v^G_t (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)

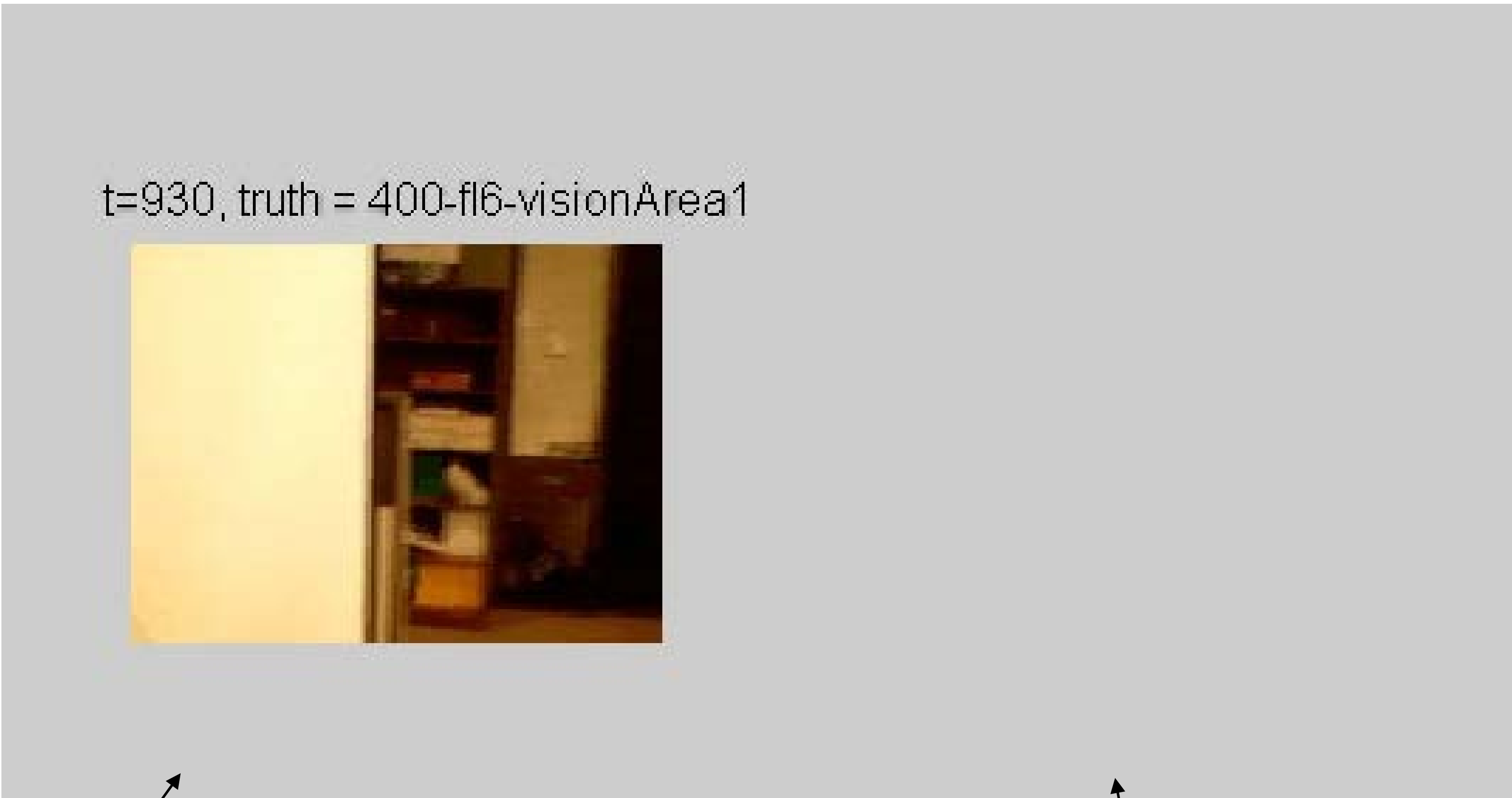
Torralba, Murphy, Freeman and Rubin. ICCV 2003

Our mobile rig

Torralba, Murphy, Freeman, Rubin. 2003

Place recognition demo

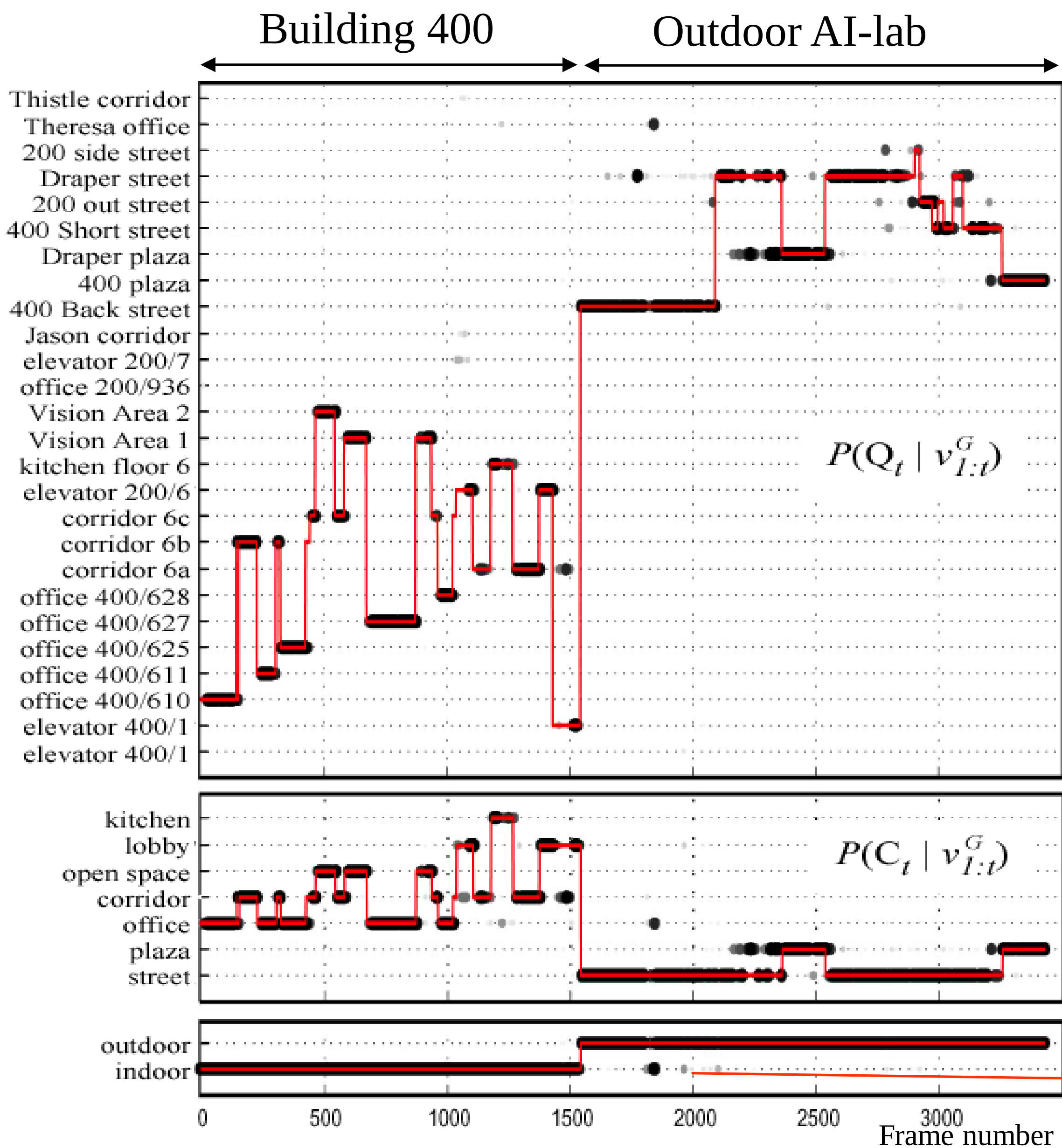




Input image (120x160)

Shows the category and the identity of
The place when the system is confident.
Runs at 4 fps on Matlab.

Identification and categorization of known places



Ground truth

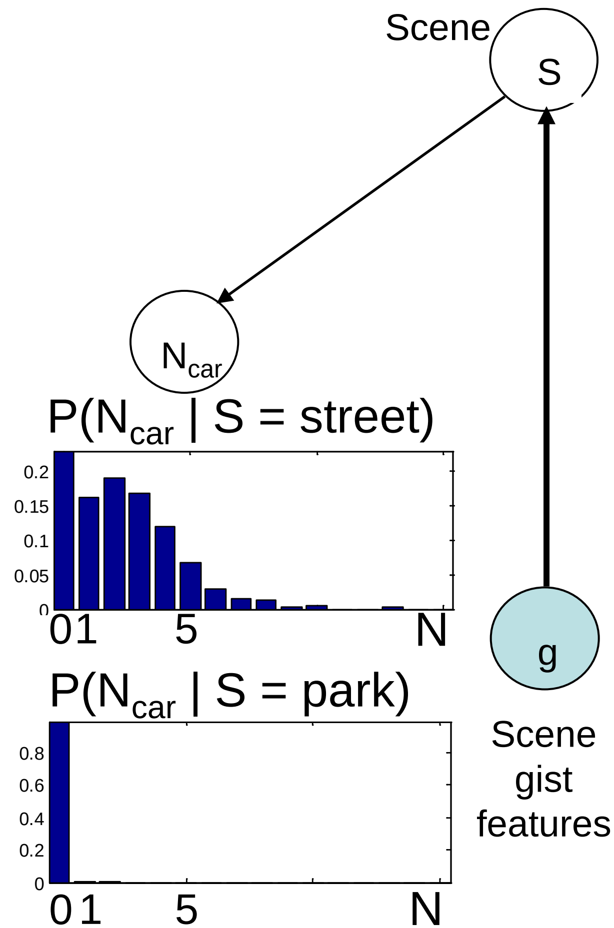
System estimate

Specific location

Location category

Indoor/outdoor

An integrated model of Scenes, Objects, and Parts



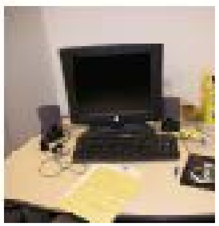
Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.

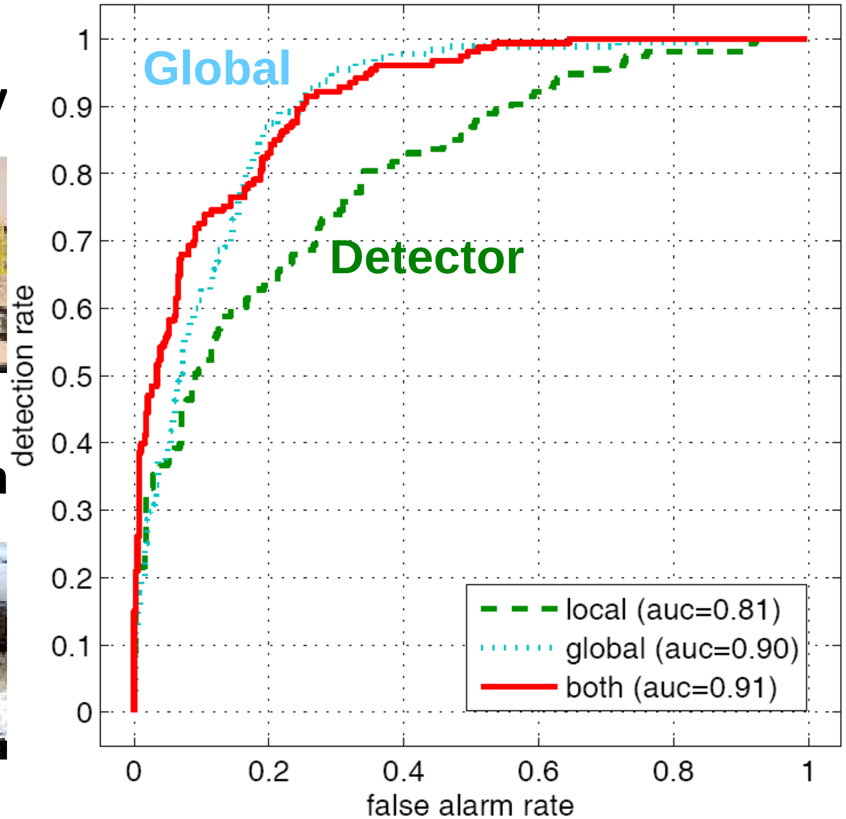
Application of object detection for image retrieval

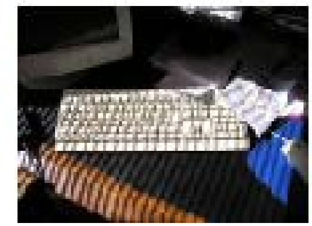
Results using the keyboard detector alone

Application of object detection for image retrieval

Results using the key

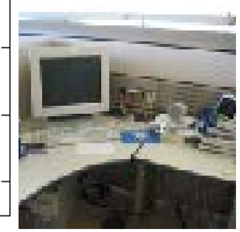





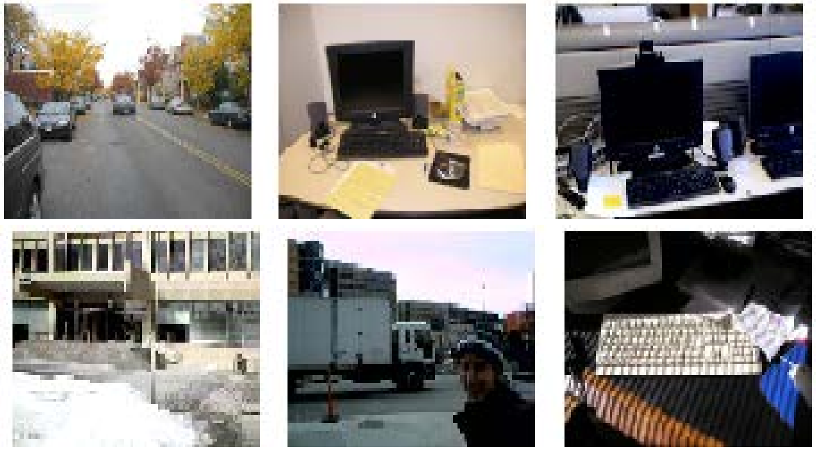
Results using both th **cene features**



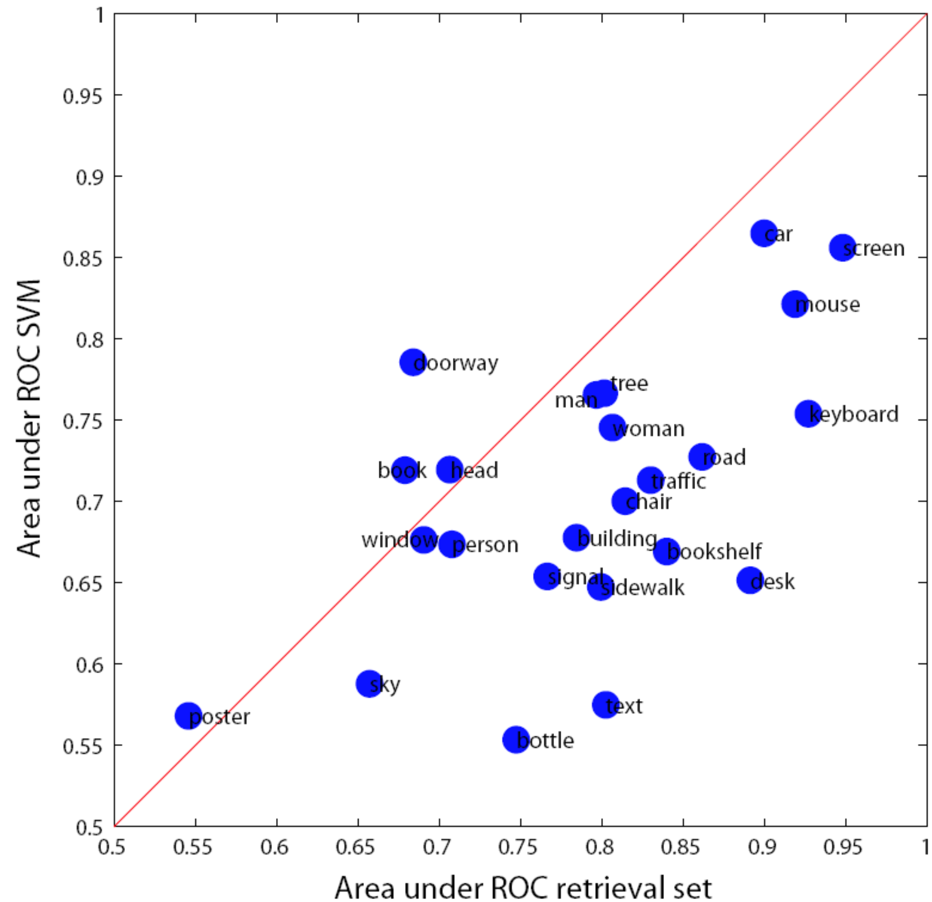



Object retrieval: scene features vs. detector

Results using the keyboard detector alone



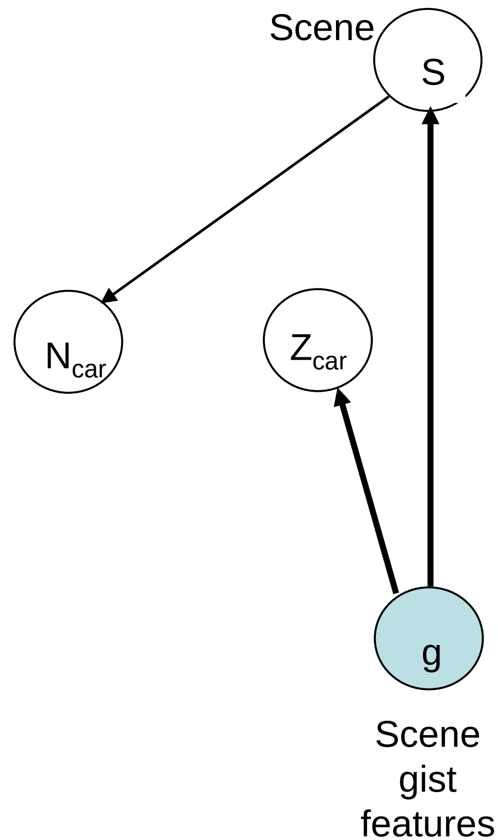
Results using both the detector
and the global scene features



Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.

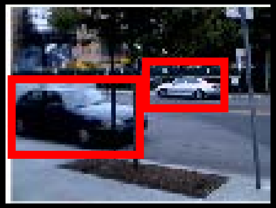
Localizing the object

An integrated model of Scenes, Objects, and Parts



Predicting object location

Training set (cars)



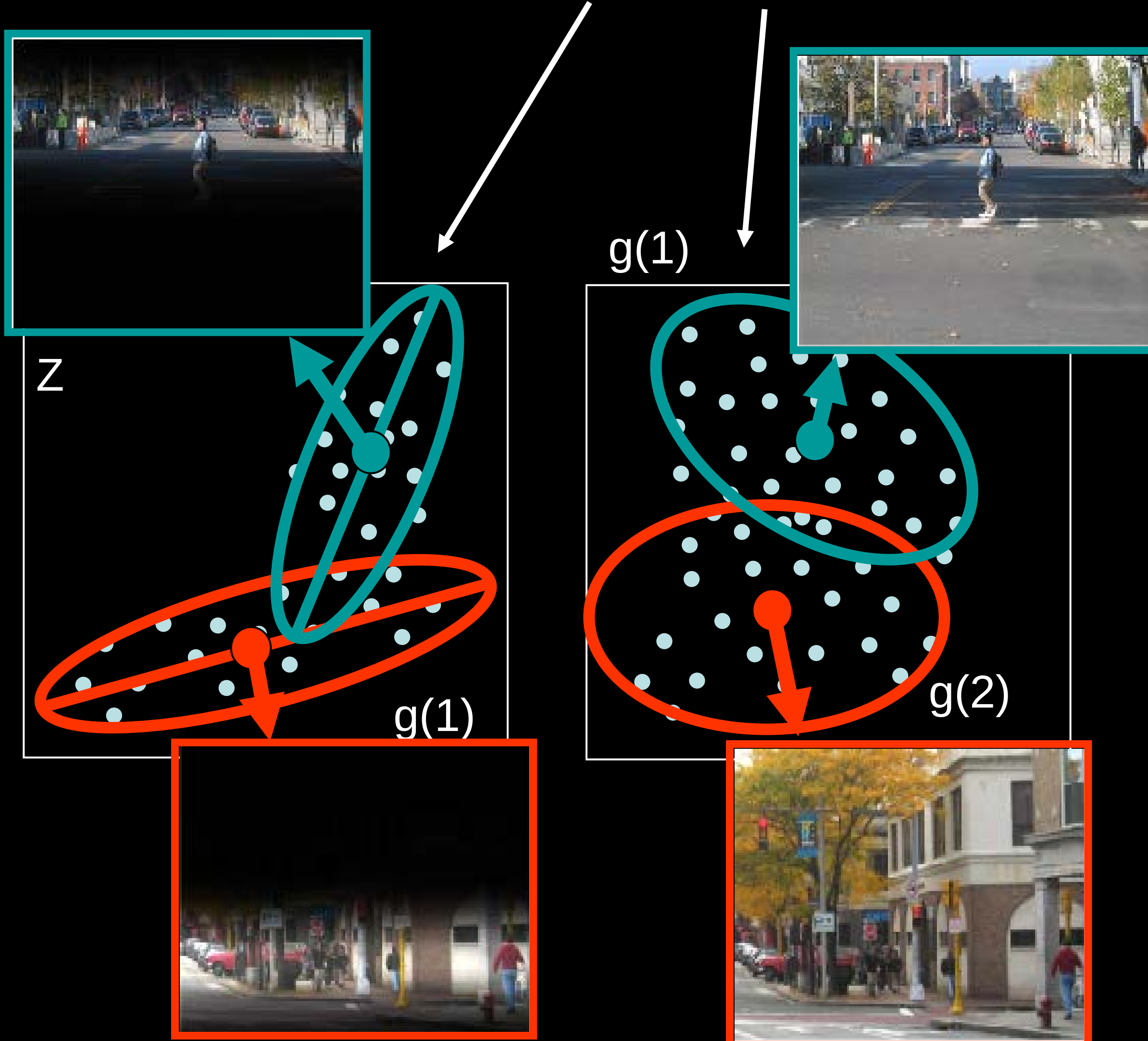
$\longrightarrow \{g^1, z^1\}$

$\longrightarrow \{g^2, z^2\}$

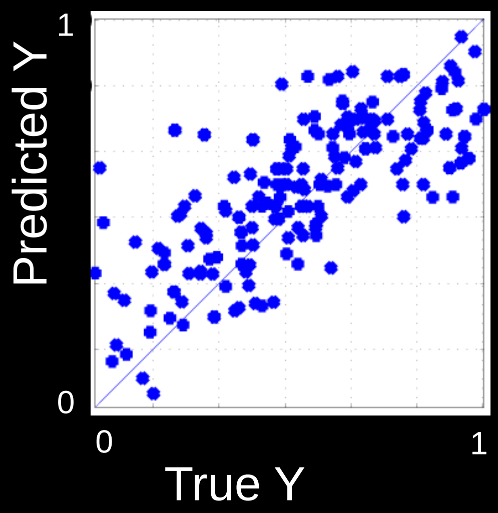
$\longrightarrow \{g^3, z^3\}$

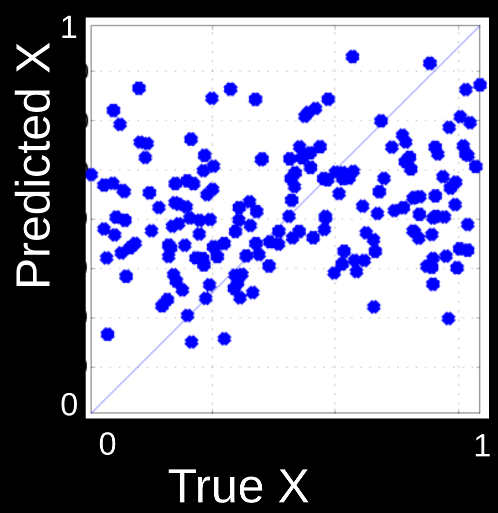
⋮

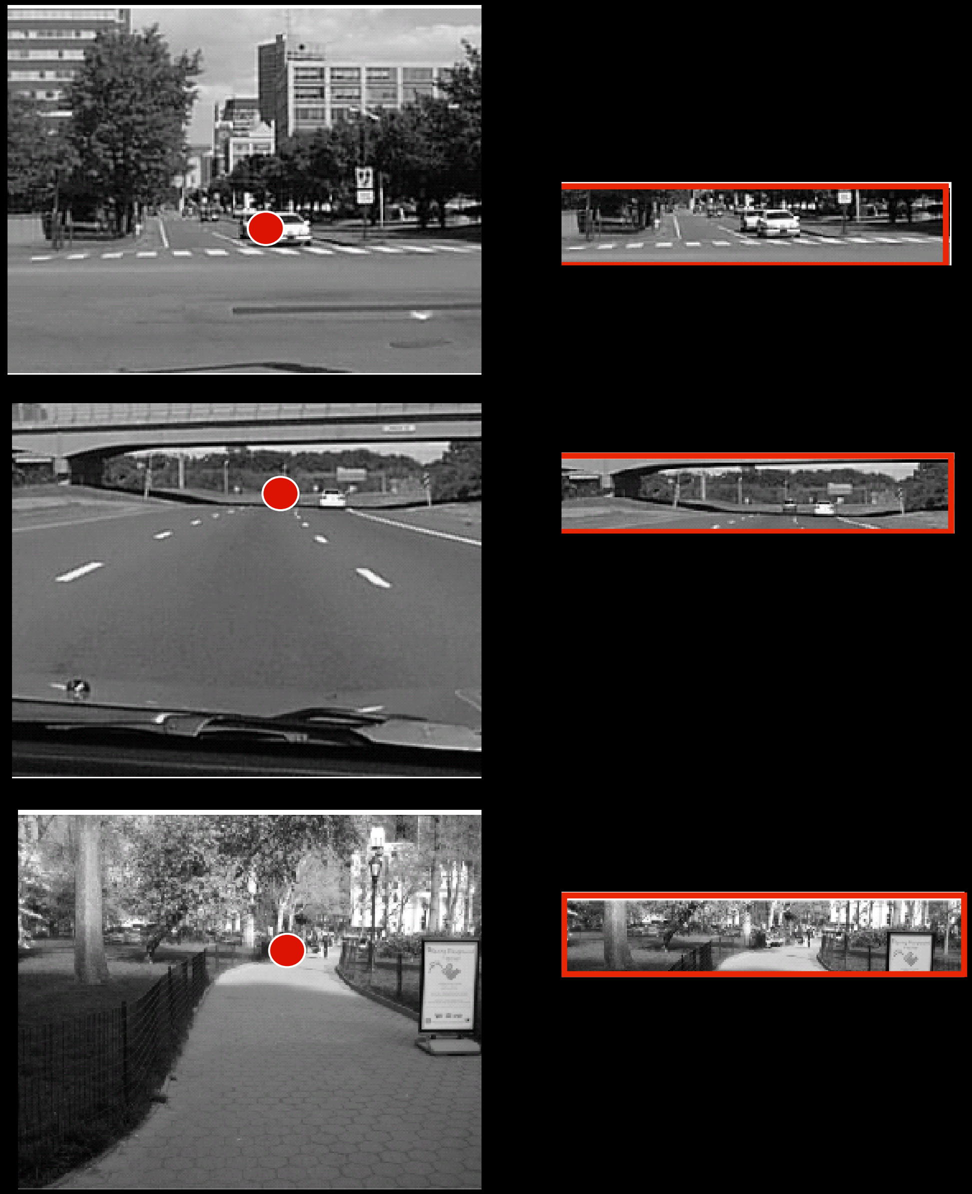
$$Z|g = \sum (A_n\, g + b_n)\; w_n(g)$$



Predicting location







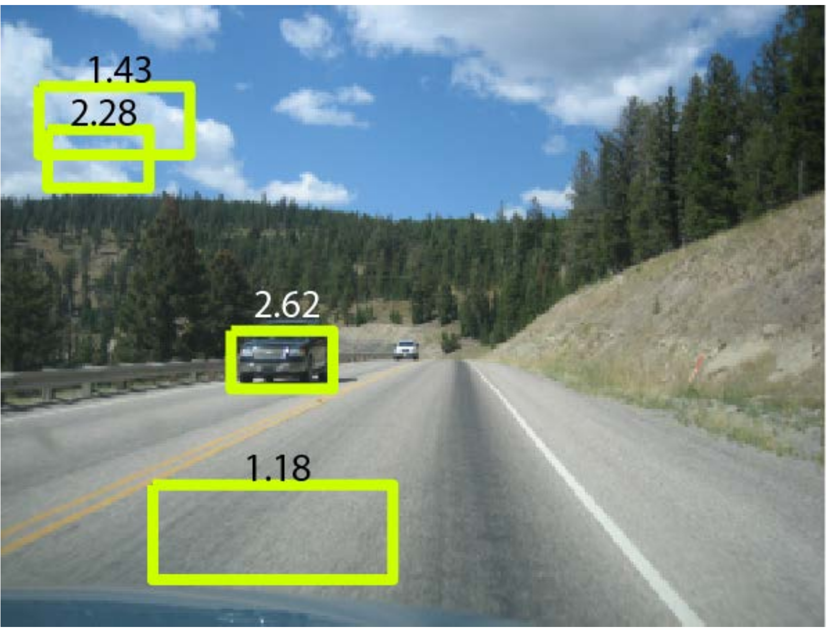
Torralba & Sinha, 2001; Murphy, Torralba, Freeman, 2003; Hoeim, Efros, Hebert. 2006 screens

keyboard

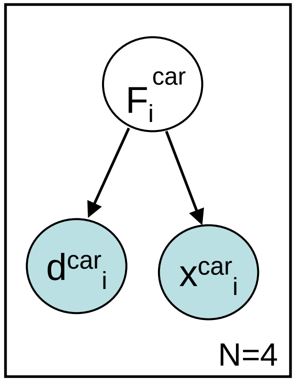
car

pedestrian

An integrated model of Scenes, Objects, and Parts

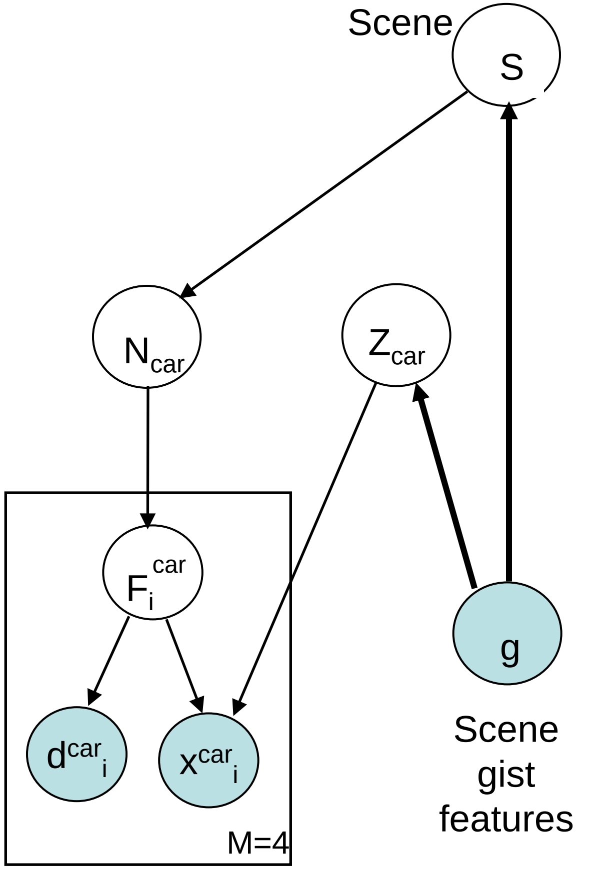


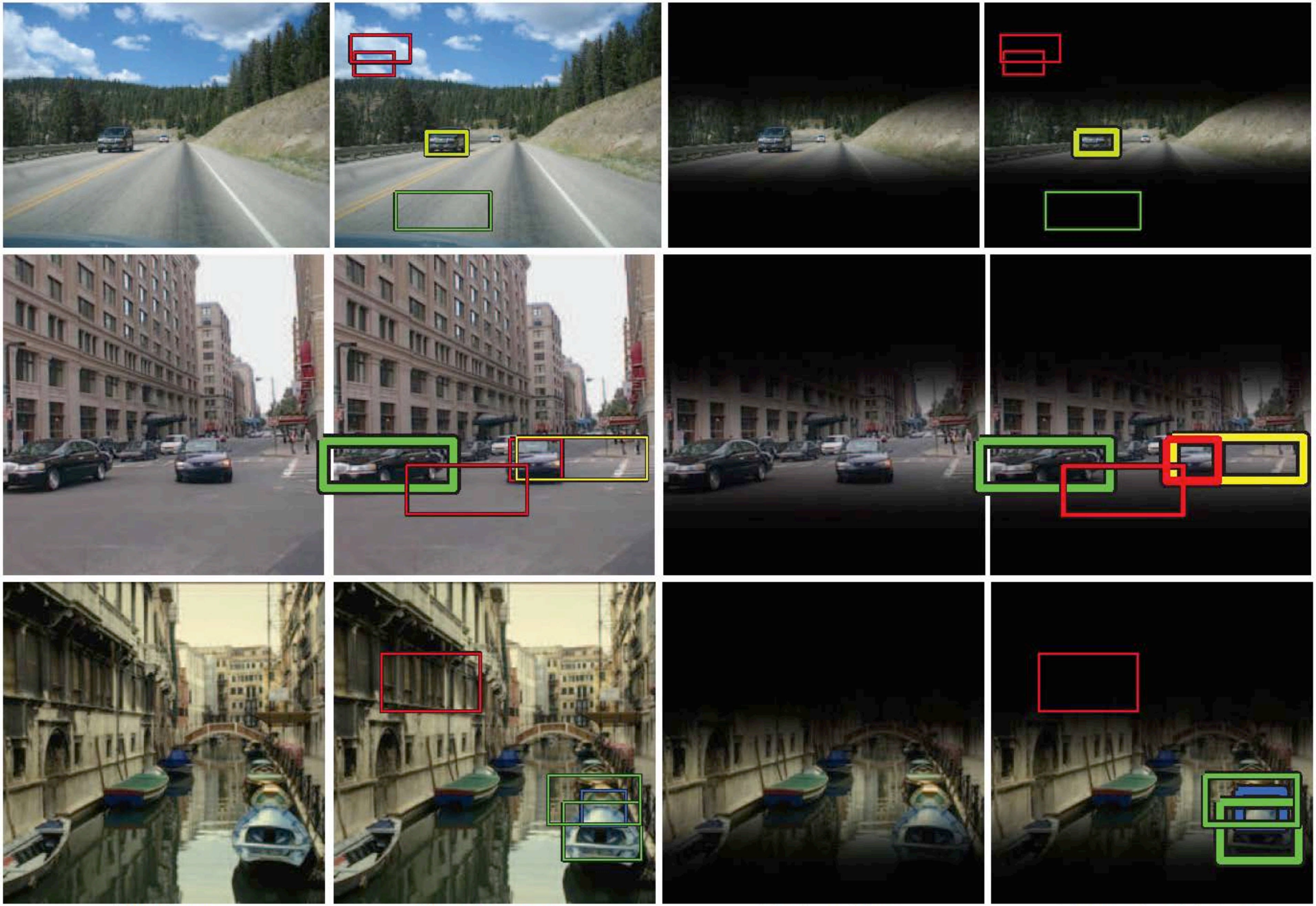
We train a multiview car detector.



$$p(d \mid F=1) = \mathcal{N}(d \mid \mu_1, \sigma_1)$$
$$p(d \mid F=0) = \mathcal{N}(d \mid \mu_0, \sigma_0)$$

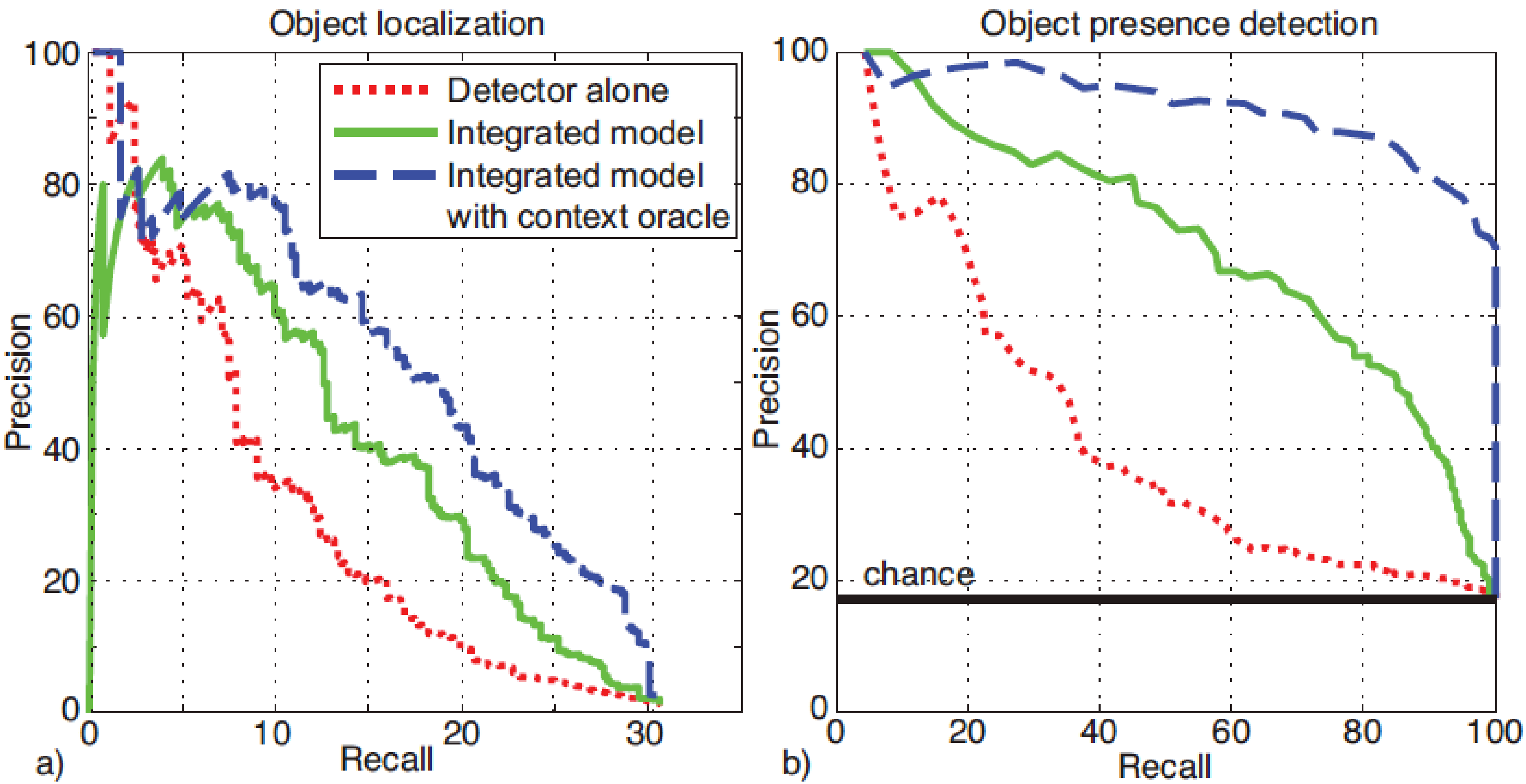
An integrated model of Scenes, Objects, and Parts





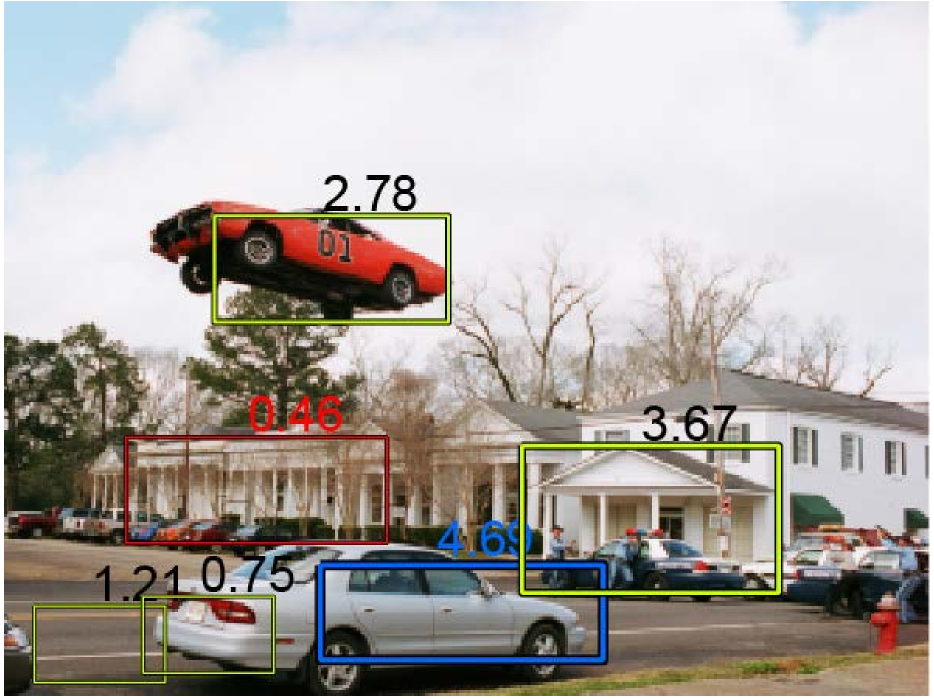
a) input image b) car detector output c) location priming c) integrated model output

Two tasks



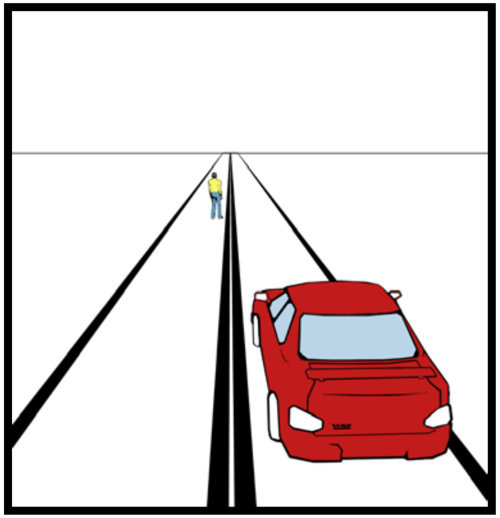
Object localization
Detector alone
Integrated model
Integrated model with context oracle
Precision
Recall
100
80
60
40
20
0
0
10
20
30
a)
Object presence detection
chance
Precision
Recall
100
80
60
40
20
0
0
20
40
60
80
100
b)

A car out of context …

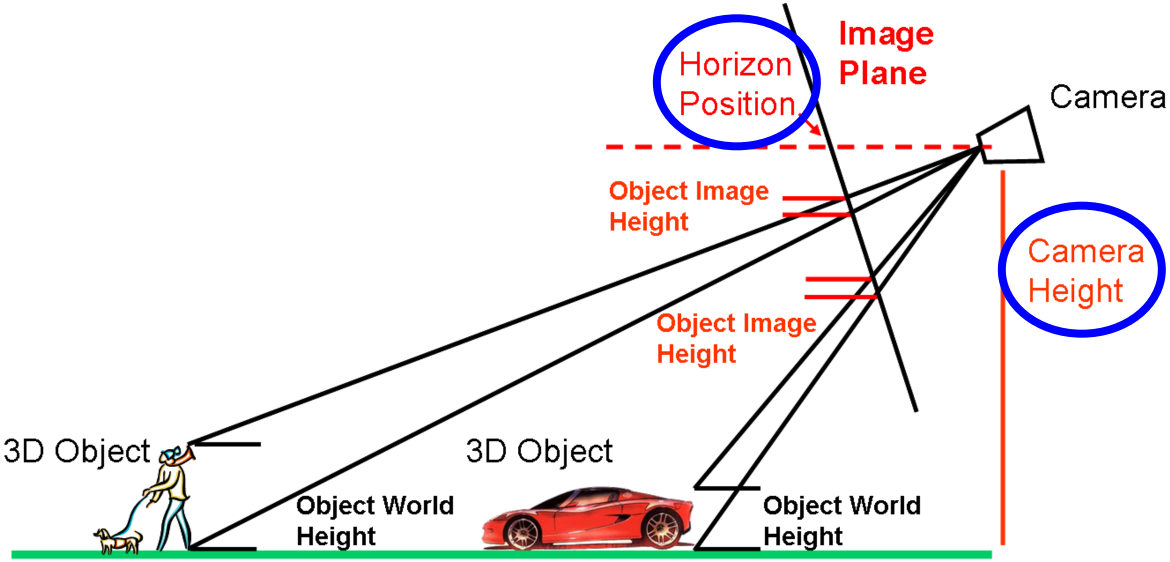


A car out of context …

3d Scene Context



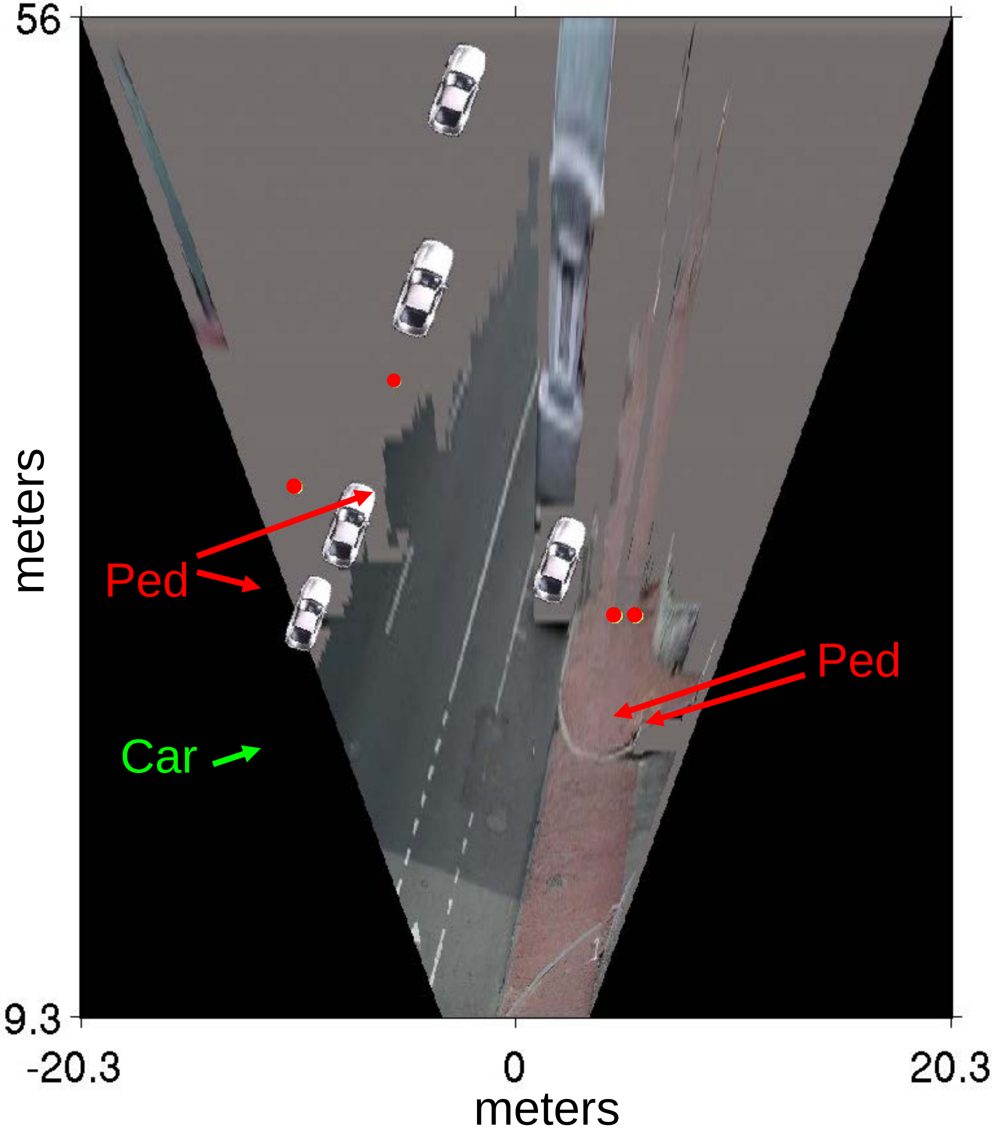
Image



World

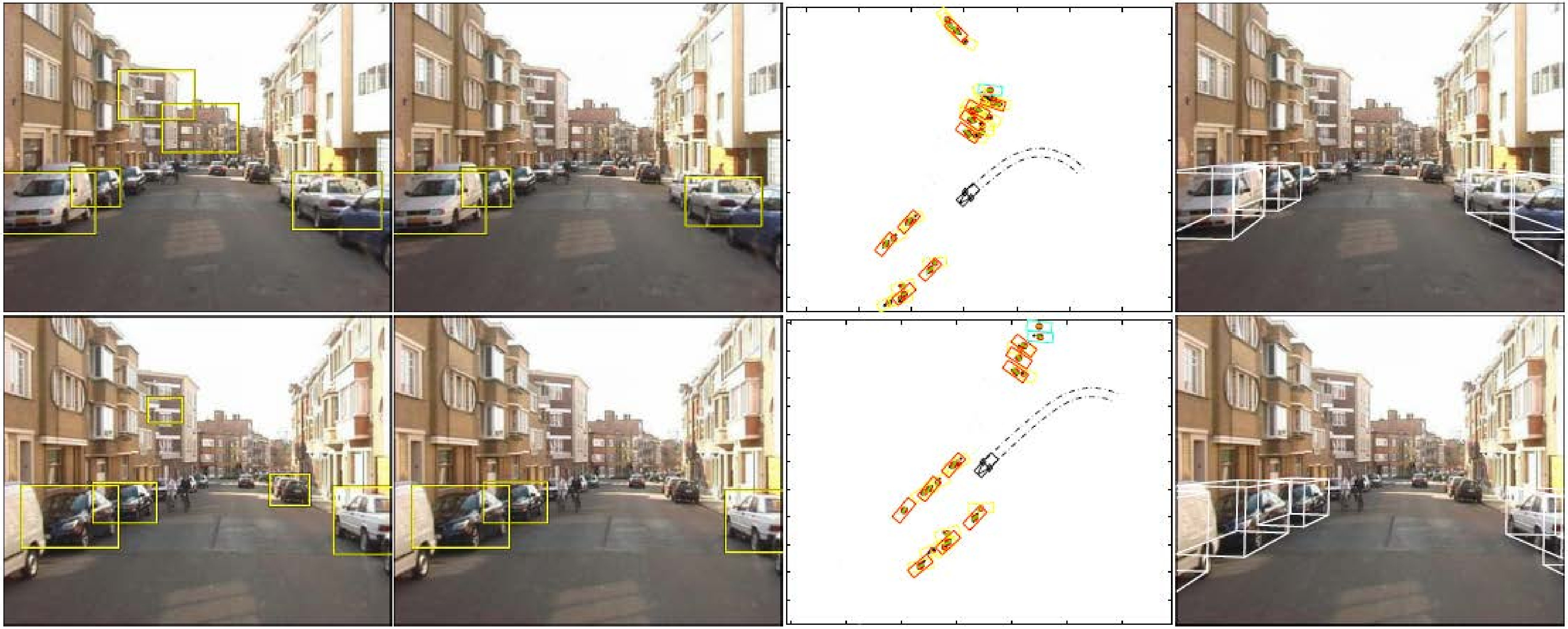
Hoiem, Efros, Hebert ICCV 2005

3d Scene Context



Hoiem, Efros, Hebert ICCV 2005

3D City Modeling using Cognitive Loops

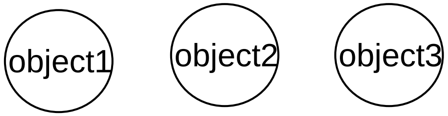


(a) (b) (c) (d)

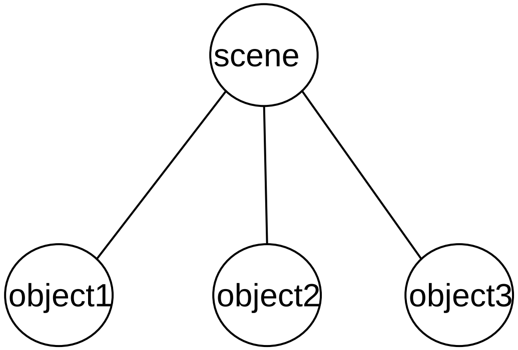
Figure 6. Stages of the recognition system: (a) initial detections before and (b) after applying ground plane constraints, (c) temporal integration on reconstructed map, (d) estimated 3D car locations, rendered back into the original image.

N. Cornelis, B. Leibe, K. Cornelis, L. Van Gool. CVPR'06

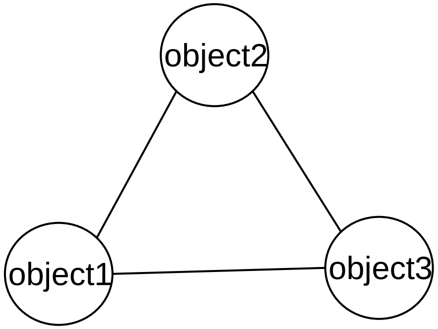
Context models



Independent model



Objects are correlated via the scene



Dependencies among objects

1) Generate candidate objects (run a detector, or segmentation)

M possible object labels
N regions

Label: $c_k = [1 \ldots M]$ with $k = [1 \ldots N]$
Scores: s_k = vector length M

2) For each candidate, get a list of possible interpretations with their probabilities

$$p(c_k = m \mid s_k)$$

3) Goal: to assign labels c_k to each candidate so that they are in contextual agreement. We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie. Objects in Context. ICCV 2007

Goal: to assign labels c_k to each candidate so that they are in contextual agreement.

M possible object labels
N regions

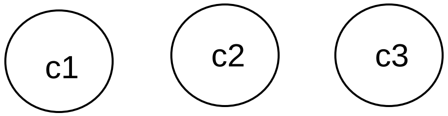
Label: $c_k = [1\ldots M]$ with $k = [1\ldots N]$
Scores: s_k = vector length M

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

Solution 1: Assume objects are independent:

$$p(c_1=m_1,\ldots, c_N=m_N|s_1,\ldots, s_N) = \prod_{i=1\ldots N} p(c_i=m_i|s_i)$$



Independent model

Problem: it does not makes use of the correlation between objects in the world. This is fine if the detectors are perfect.

Goal: to assign labels c_k to each candidate so that they are in contextual agreement.

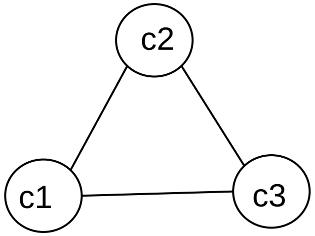
M possible object labels
N regions

Label: $c_k = [1\ldots M]$ with $k = [1\ldots N]$
Scores: s_k = vector length M

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

Solution 2: Assume objects are fully dependent:



$$p(c_1=m_1,\ldots, c_N=m_N|s_1,\ldots, s_N) =$$

$$= \frac{p(s_1,\ldots,s_N|c_1=m_1,\ldots,c_N=m_N)\; p(c_1=m_1,\ldots,c_N=m_N)}{Z(s_1,\ldots s_N)}$$

$$= \frac{\prod_{i=1\ldots N} p(s_i|c_i=m_i)\; p(c_1=m_1,\ldots,c_N=m_N)}{Z(s_1,\ldots s_N)}$$

$$Z(s_1,\ldots s_N) = \sum_{\text{All } [c_1,\ldots,c_N] \text{ assignments}} \prod p(s_i|c_i=m_i)\; p(c_1=m_1,\ldots,c_N=m_N)$$

Problem: learning $p(c_1=m_1,\ldots,c_N=m_N)$ will need a lot of data. Recognition can be slow.

Goal: to assign labels c_k to each candidate so that they are in contextual agreement.

M possible object labels
N regions

Label: $c_k = [1\ldots M]$ with $k = [1\ldots N]$
Scores: s_k = vector length M

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

Solution 3: Approximated model of dependencies:

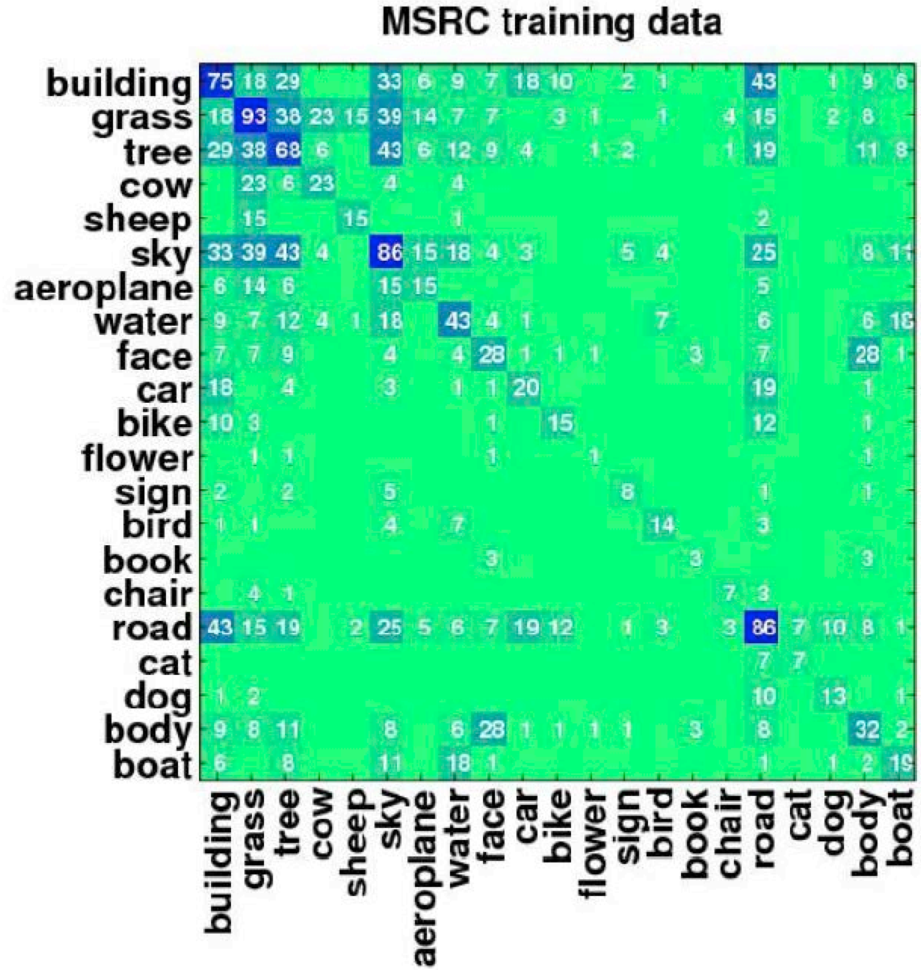
$$p(c_1=m_1,\ldots, c_N=m_N|s_1,\ldots, s_N) = \frac{\prod_{i=1\ldots N} p(s_i|c_i=m_i)\, p(c_1=m_1,\ldots,c_N=m_N)}{Z(s_1,\ldots s_N)}$$

$$p(c_1=m_1,\ldots,c_N=m_N) = \exp\Big(\sum_{i,j=1\ldots N} \Phi(c_i=m_i, c_j=m_j)\Big)$$

$\Phi(c_i=m_i, c_j=m_j)$ = co-ocurrence matrix on training set (count how many times two objects appear together).

Problem: learning $p(c_1=m_1,\ldots,c_N=m_N)$ will be easier, but recognition may still be slow.

$\Phi(c_i=m_i, c_j=m_j)$ = co-ocurrence matrix on training set (count how many times two objects appear together).



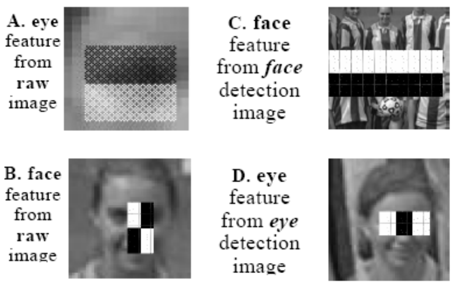
A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie. Objects in Context. ICCV 2007

A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora and S. Belongie. Objects in Context. ICCV 2007

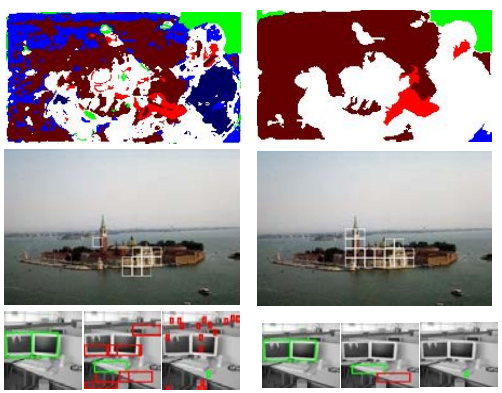
Objects in context

Torralba, Sinha (2001)

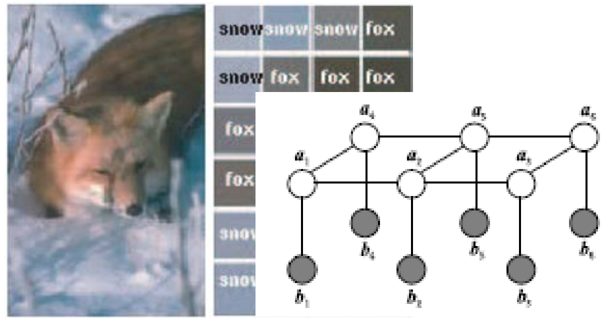
Fink & Perona (2003)



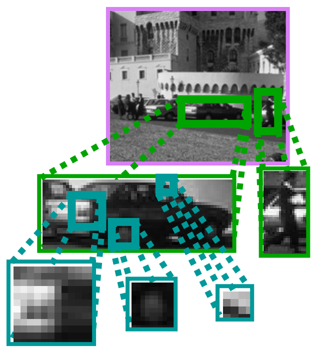
Kumar, Hebert (2005)



Carbonetto, de Freitas & Barnard (2004)

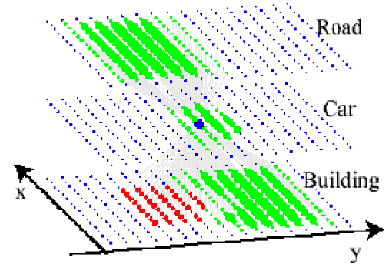


Sudderth, Torralba, Wilsky, Freeman (2005)



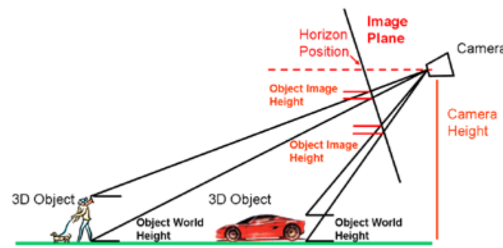
Heitz and Koller (2008)

Torralba Murphy Freeman (2004)



Rabinovich et al (2007)

Hoiem, Efros, Hebert (2005)



Desai, Ramanan, and Fowlkes (2009)

Object-Object Relationships

- Fink & Perona (NIPS 03)

Use output of boosting from other objects at previous iterations as input into boosting for this iteration

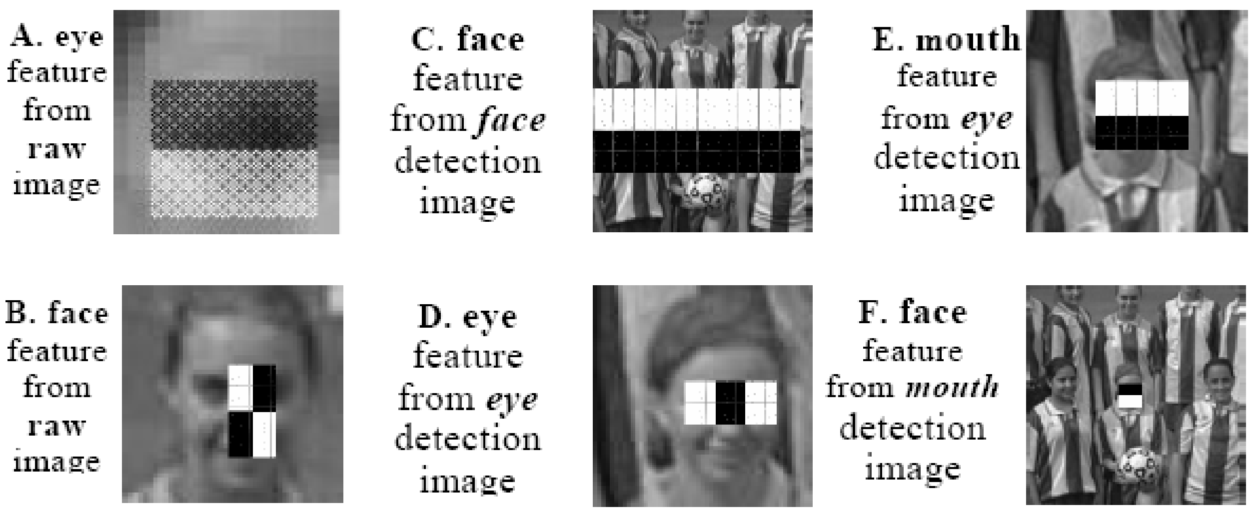
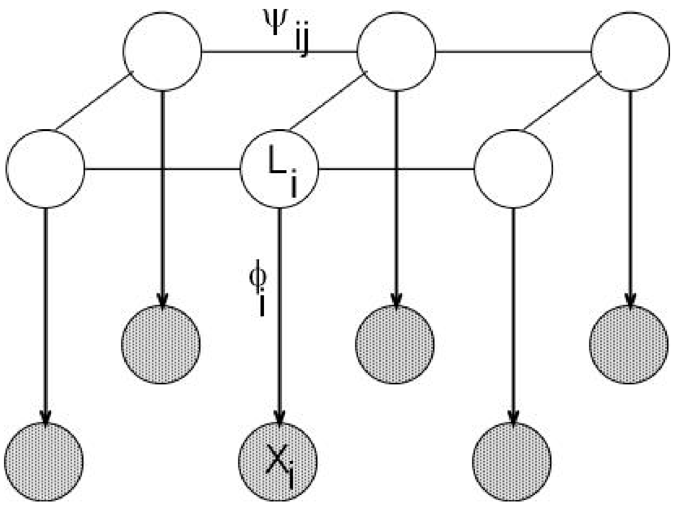


Figure 5: **A-E.** Emerging features of eyes, mouths and faces (presented on windows of raw images for legibility). The windows' scale is defined by the detected object size and by the map mode (local or contextual). **C.** faces are detected using face detection maps H^{Face}, exploiting the fact that faces tend to be horizontally aligned.

Pixel labeling using MRFs

Enforce consistency between neighboring labels, and between labels and pixels

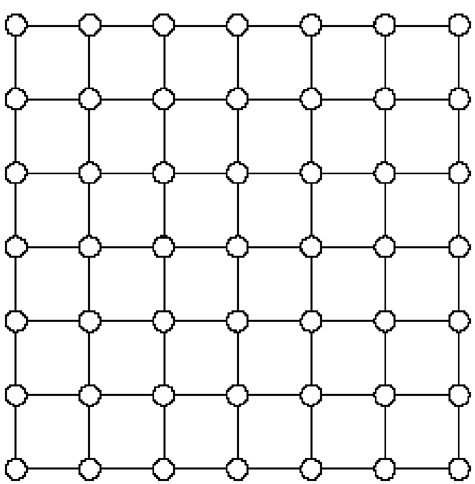
$$P(L,x) = P(L)P(x|L) = [\frac{1}{Z}\prod_i \prod_{j\in N_i} \psi_{ij}(L_i,L_j)][\prod_i P(x_i|L_i)]$$

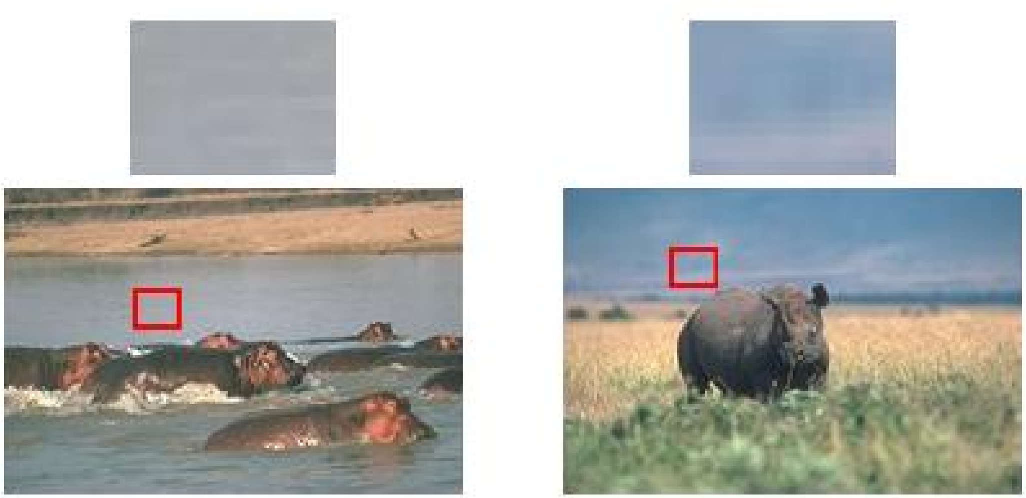


Carbonetto, de Freitas & Barnard, ECCV'04

Beyond nearest-neighbor grids

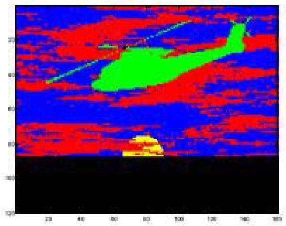
- Most MRF/CRF models assume nearest-neighbor graph topology
- This cannot capture long-distance correlations

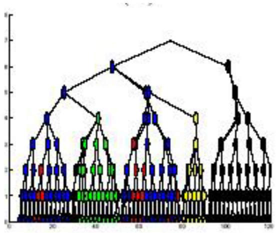


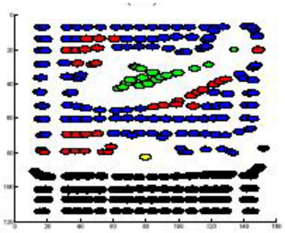


Dynamically structured trees

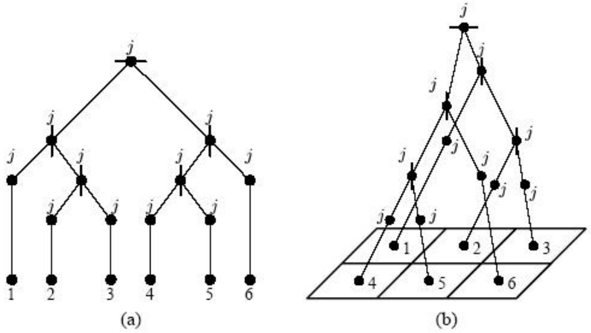
- Each node pick its parents
  (Storkey& Williams, PAMI'03)





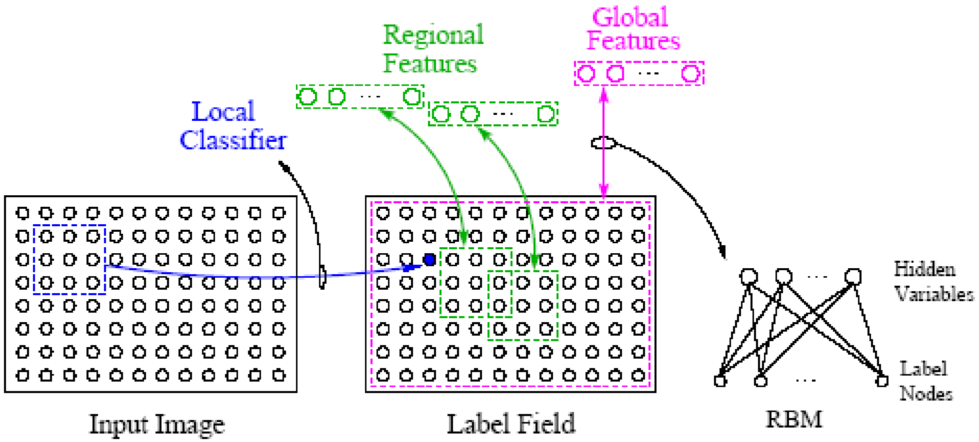


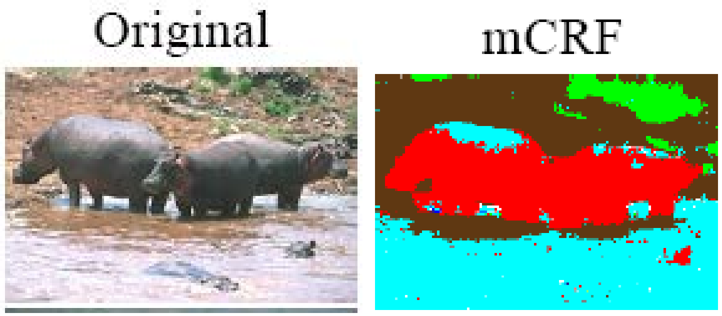
- 2D SCFGs
  (Pollak, Siskind, Harper & Bouman ICASSP'03)



Object-Object Relationships

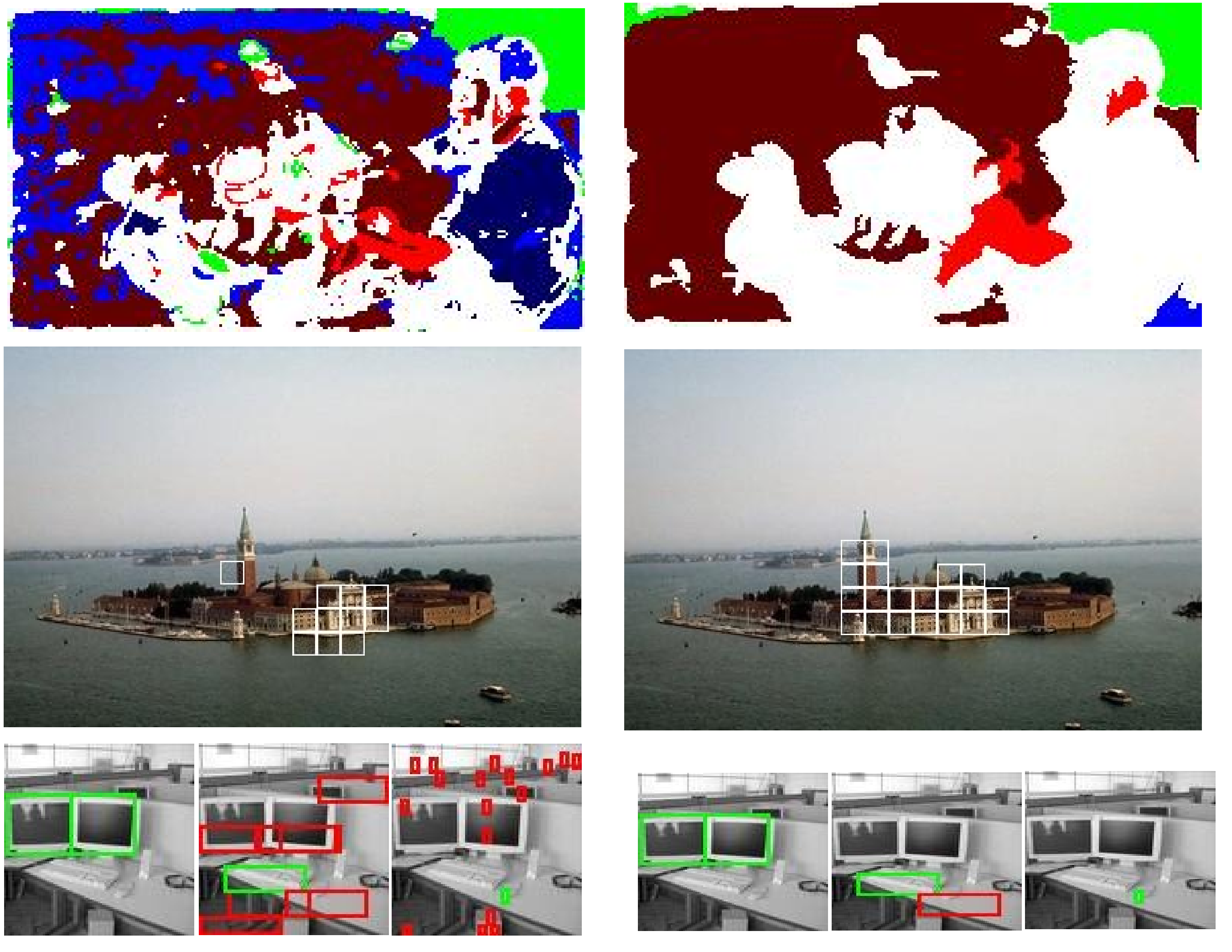
Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)





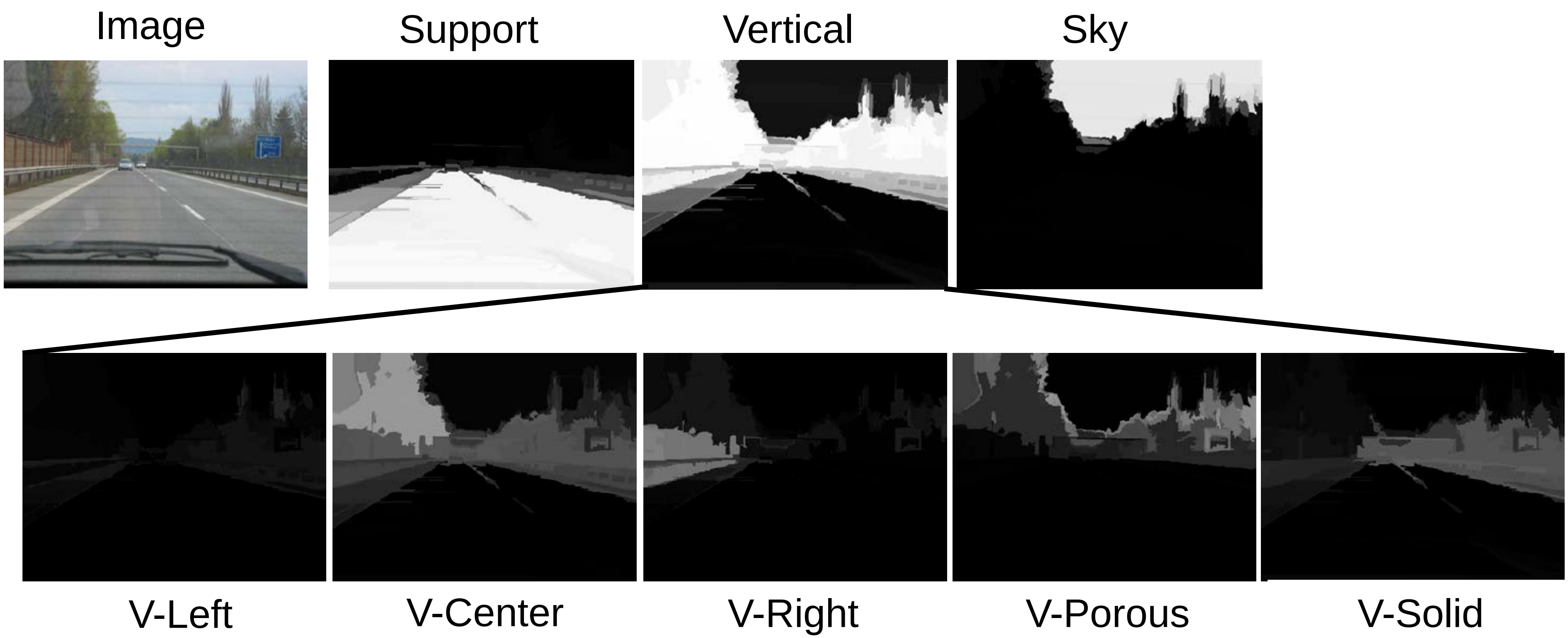
He, Zemel & Carreira-Perpinan (04)

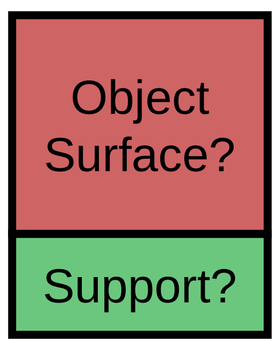
Object-Object Relationships



[Kumar Hebert 2005]

3d Scene Context





[Hoiem, Efros, Hebert ICCV 2005]

Using stuff to find things

Heitz and Koller, ECCV 2008

In this work, there is not labeling for stuff. Instead, they look for clusters of textures and model how each cluster correlates with the target object.

What, where and who? Classifying events by scene and object recognition

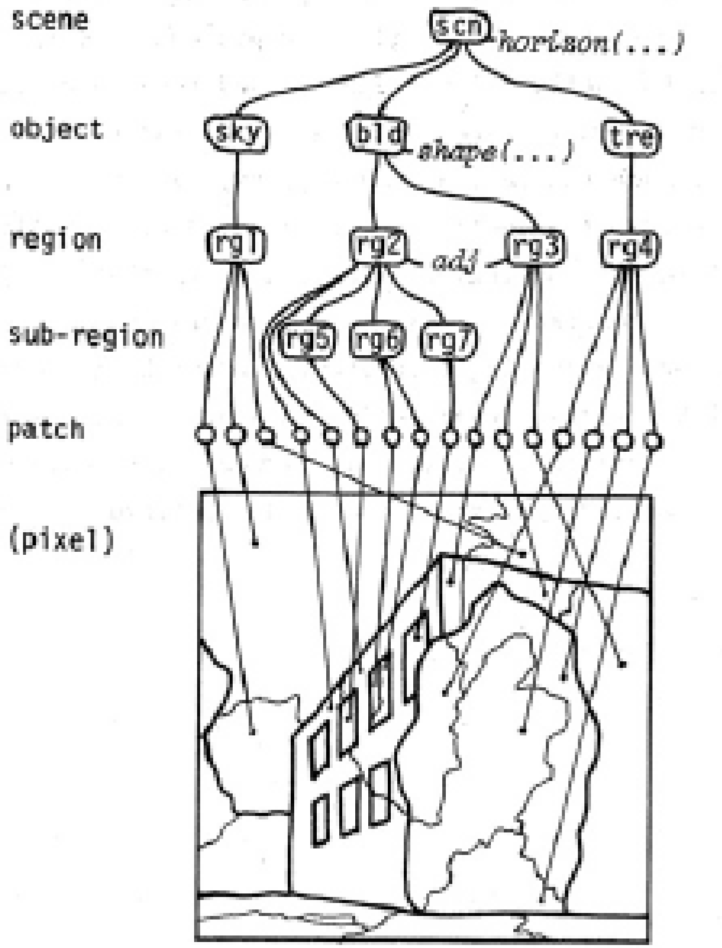
L-J Li & L. Fei-Fei, ICCV 2007

what who where

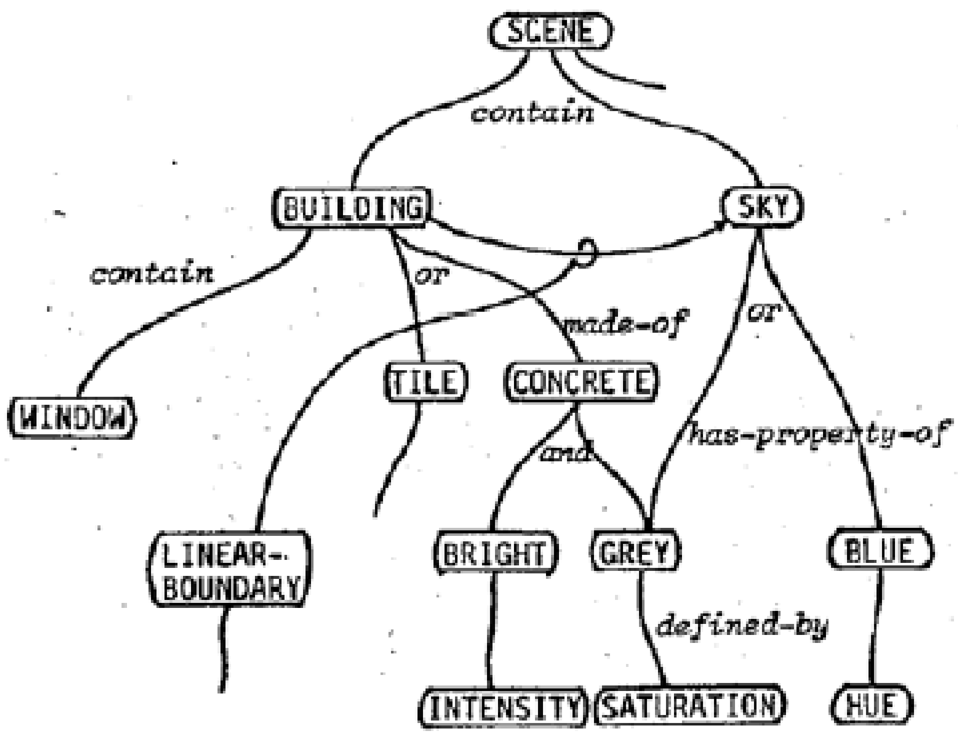
Slide by Fei-fei

L.-J. Li & L. Fei-Fei ICCV 2007

Grammars

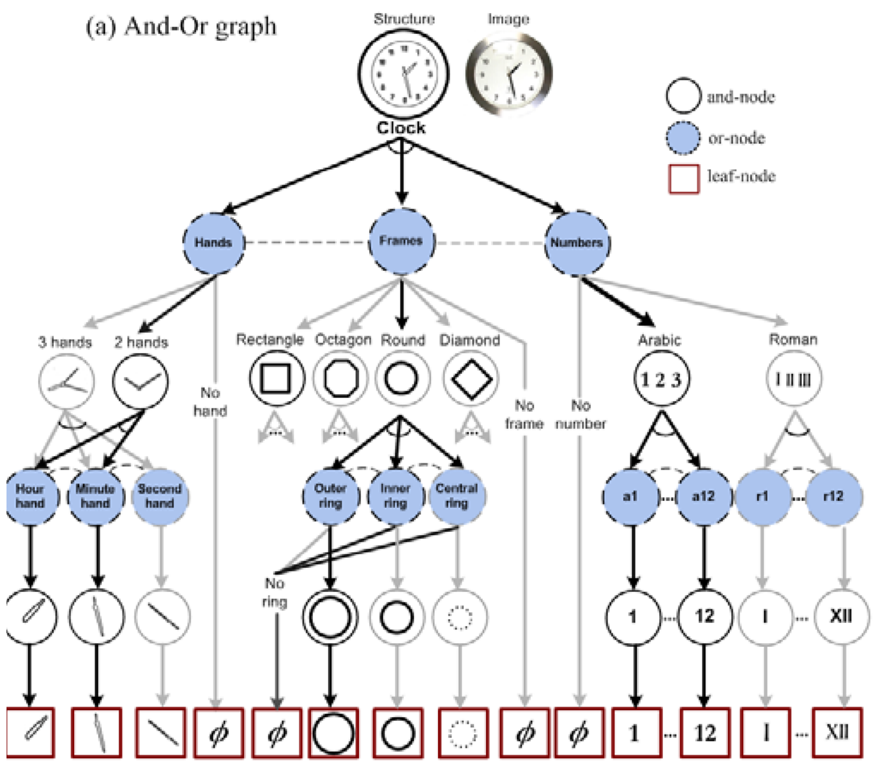


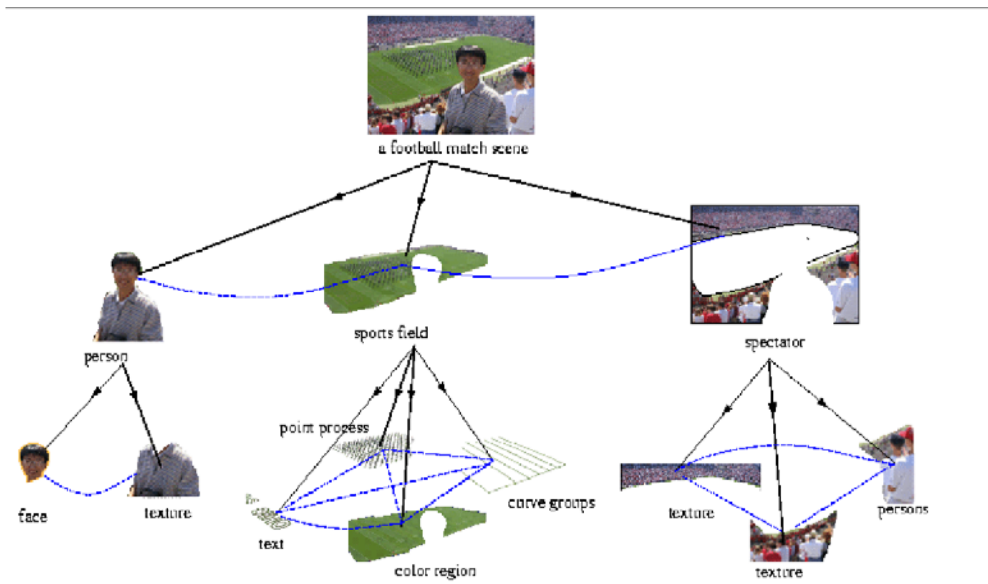
[Ohta & Kanade 1978]



- Guzman (*SEE*), 1968
- Noton and Stark 1971
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Marr, 1982
- Yakimovsky & Feldman, 1973

Grammars for objects and scenes





Example: parsing (Tu et al, 2000-2004)

S.C. Zhu and D. Mumford. A Stochastic Grammar of Images. Foundations and Trends in Computer Graphics and Vision, 2006.

Who needs context anyway?

We can recognize objects even out of context

Banksy