Scene Classification

BIL719 – Computer Vision

Pinar Duygulu
Hacettepe University

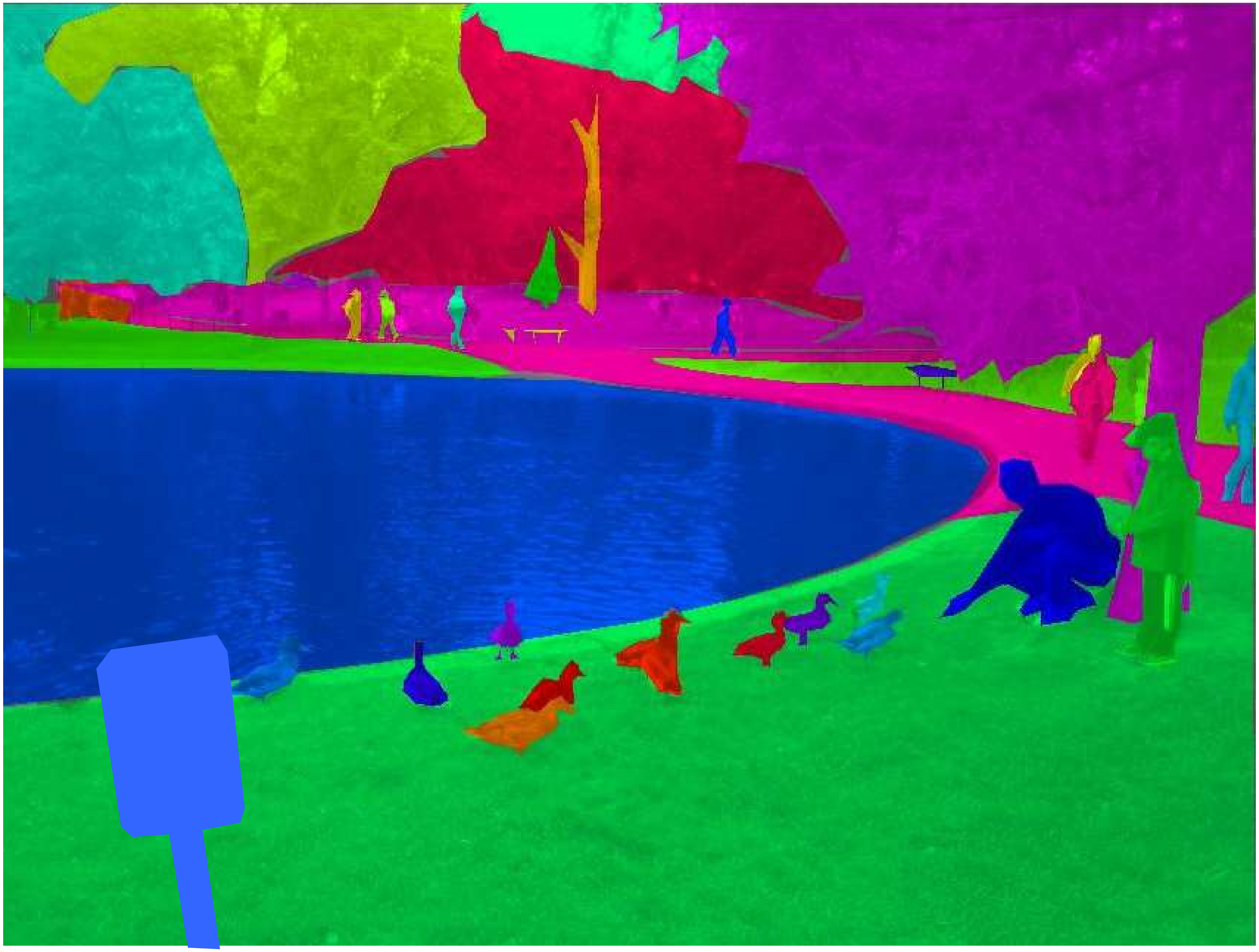
(Source:Antonio Torralba)

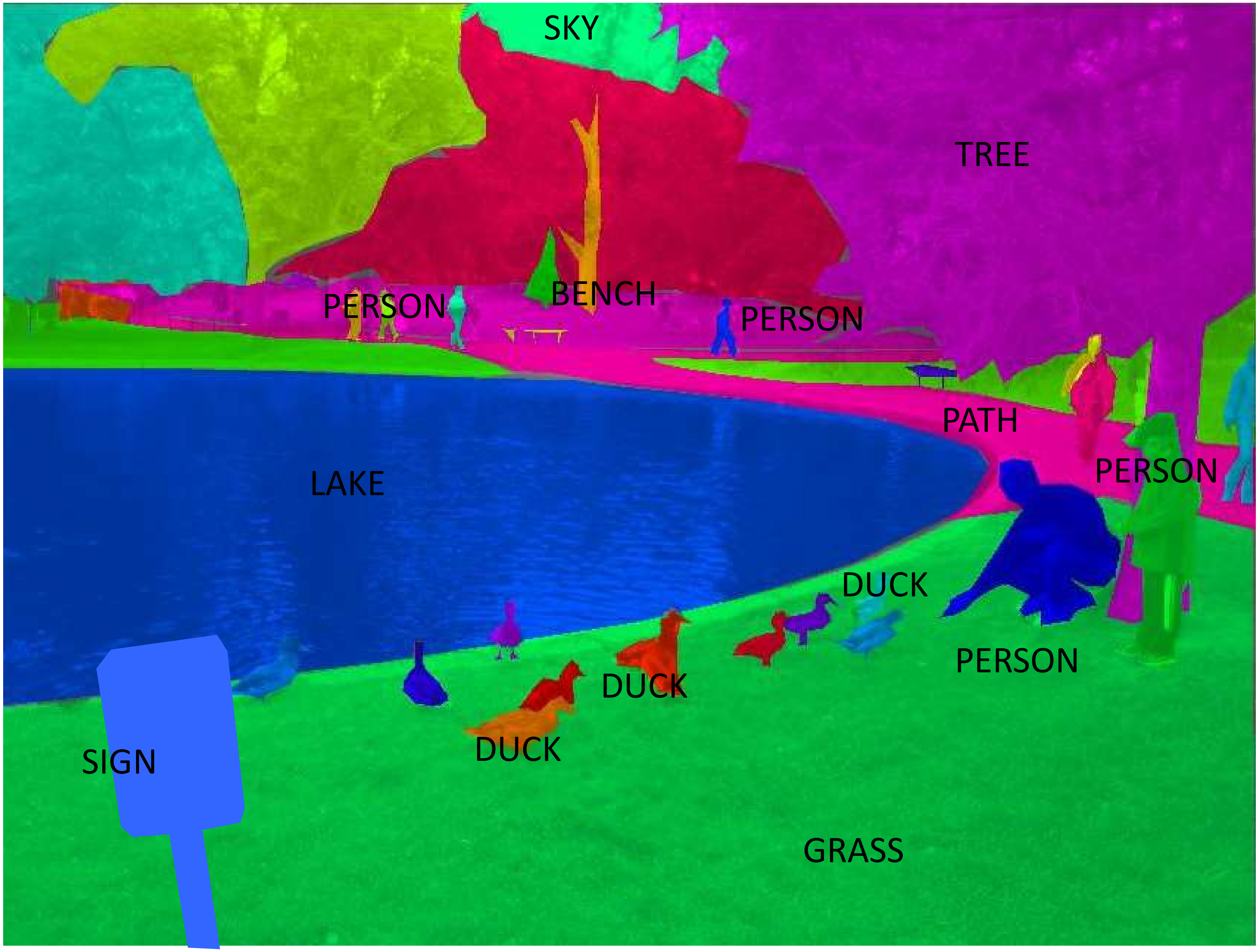
The texture

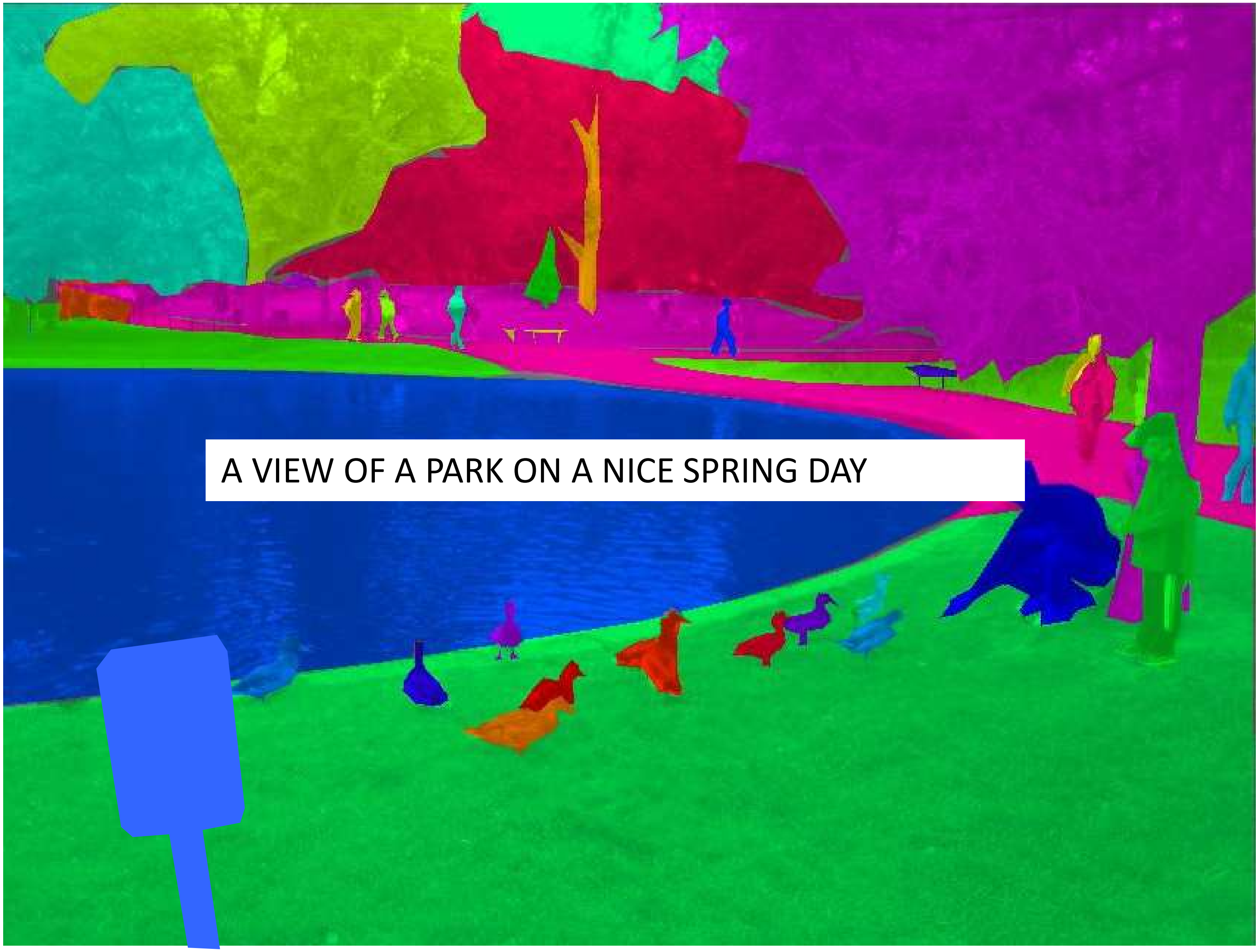
The object

The scene


PLEASE
DO NOT
FEED
THE
DUCKS




SKY
TREE
PERSON
BENCH
PERSON
PATH
PERSON
LAKE
DUCK
DUCK
PERSON
DUCK
SIGN
GRASS


A VIEW OF A PARK ON A NICE SPRING DAY


PEOPLE WALKING IN THE PARK
Do not feed the ducks sign
DUCKS LOOKING FOR FOOD
PERSON FEEDING DUCKS IN THE PARK


PEOPLE UNDER THE
SHADOW OF THE TREES
PLEASE
DO NOT
FEED
THE
DUCKS
DUCKS ON TOP
OF THE GRASS

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

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

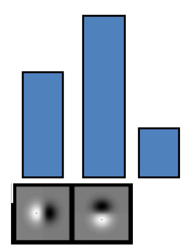
In a glance, we remember the meaning of an image and its global layout but some objects and details are forgotten

What can be an alternative to objects?

- **An alternative to objects: scene emergent features**

Global and local representations

Global and local representations



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

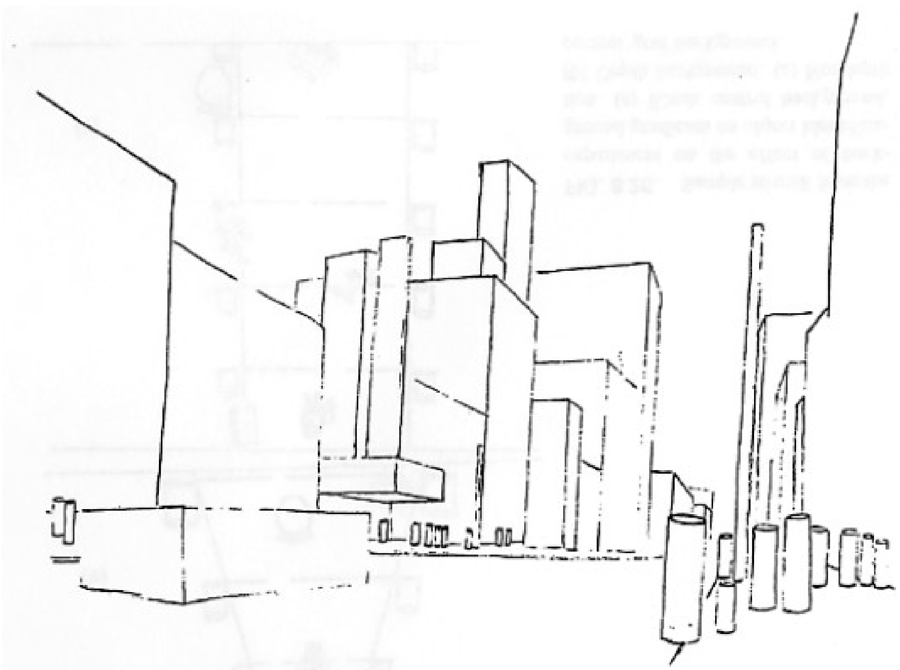


FIG. 8.23. *Downtown Buffalo*. Drawn by Robert Mezzanotte by converting objects in a photograph to basic rectilinear or cylindrical bodies.

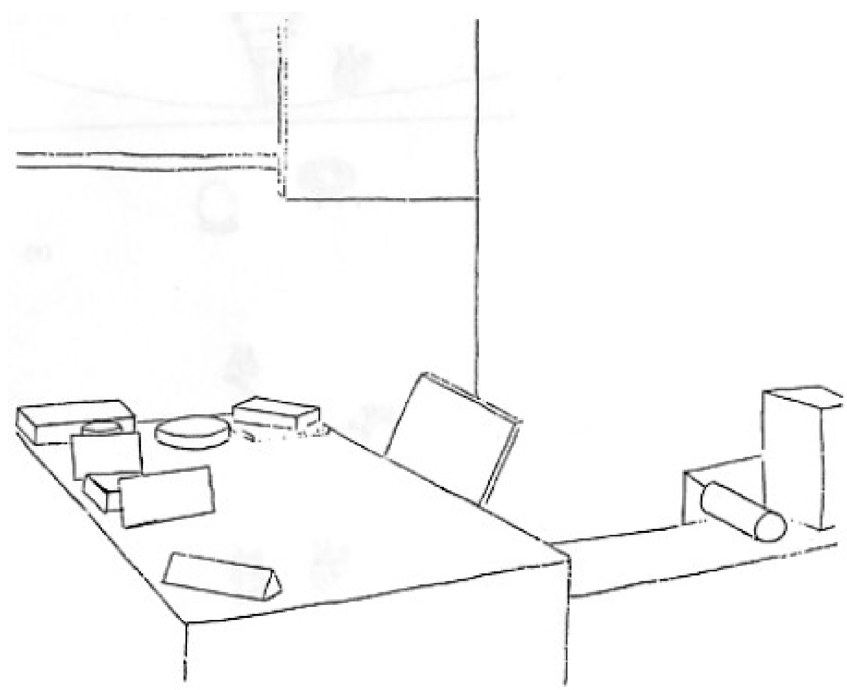
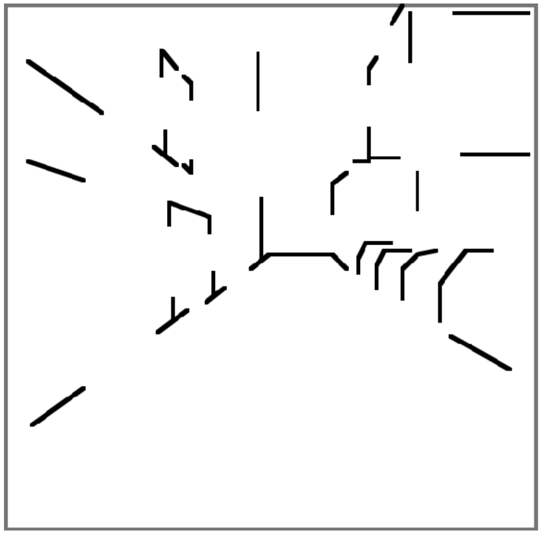


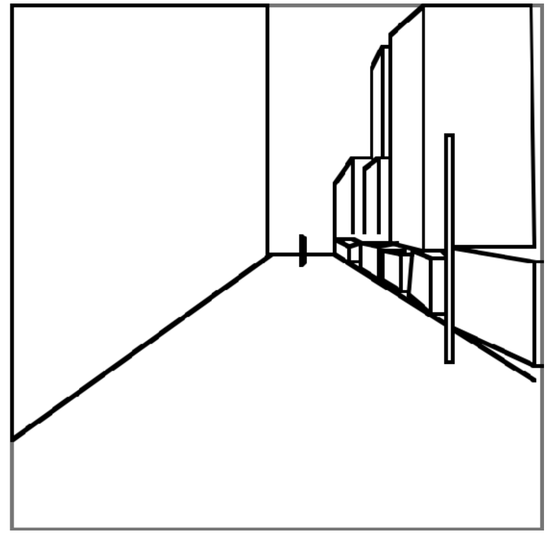
FIG. 8.24. *Office*, drawn by Robert Mezzanotte.

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

Examples of scene emergent features



Suggestive edges and junctions



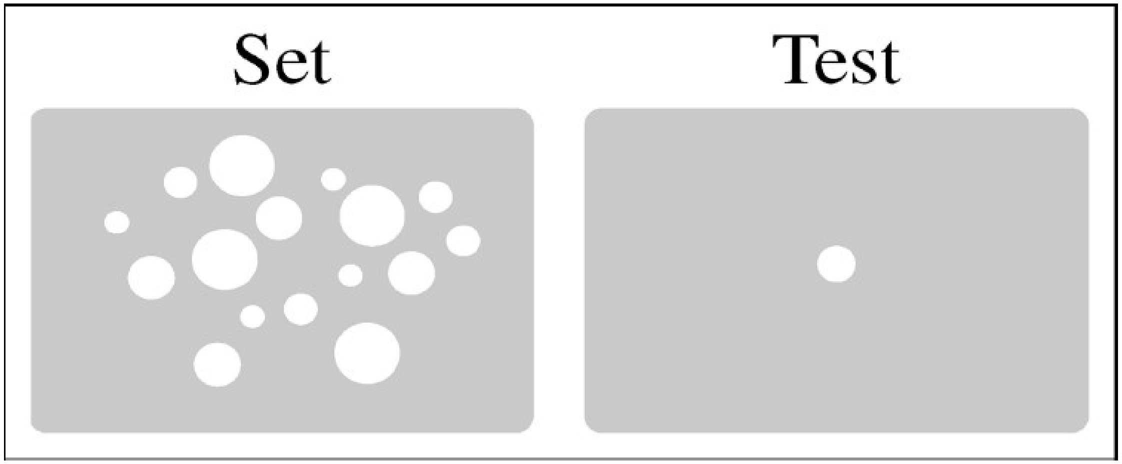
Simple geometric forms

Blobs

Textures

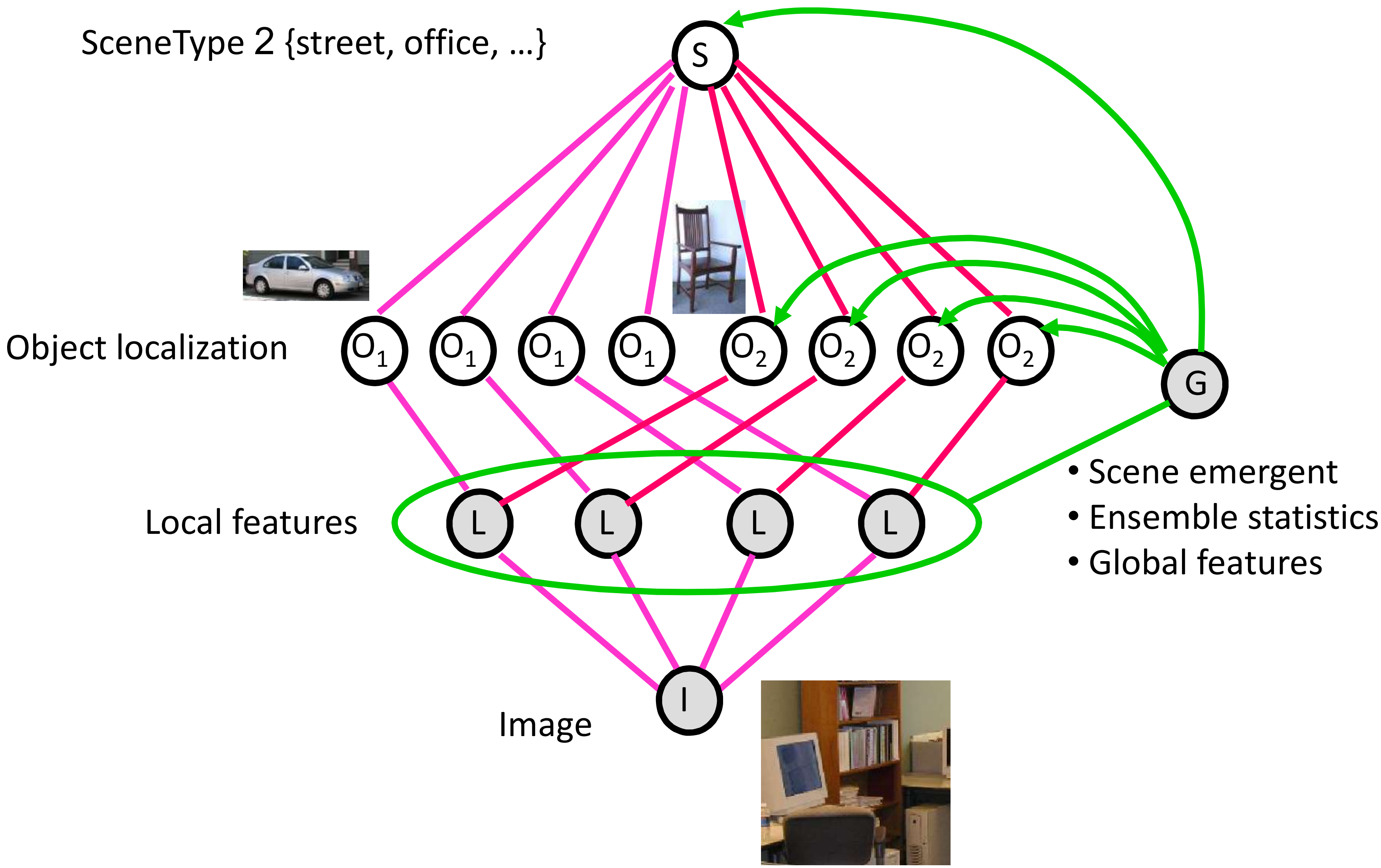
Ensemble statistics

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

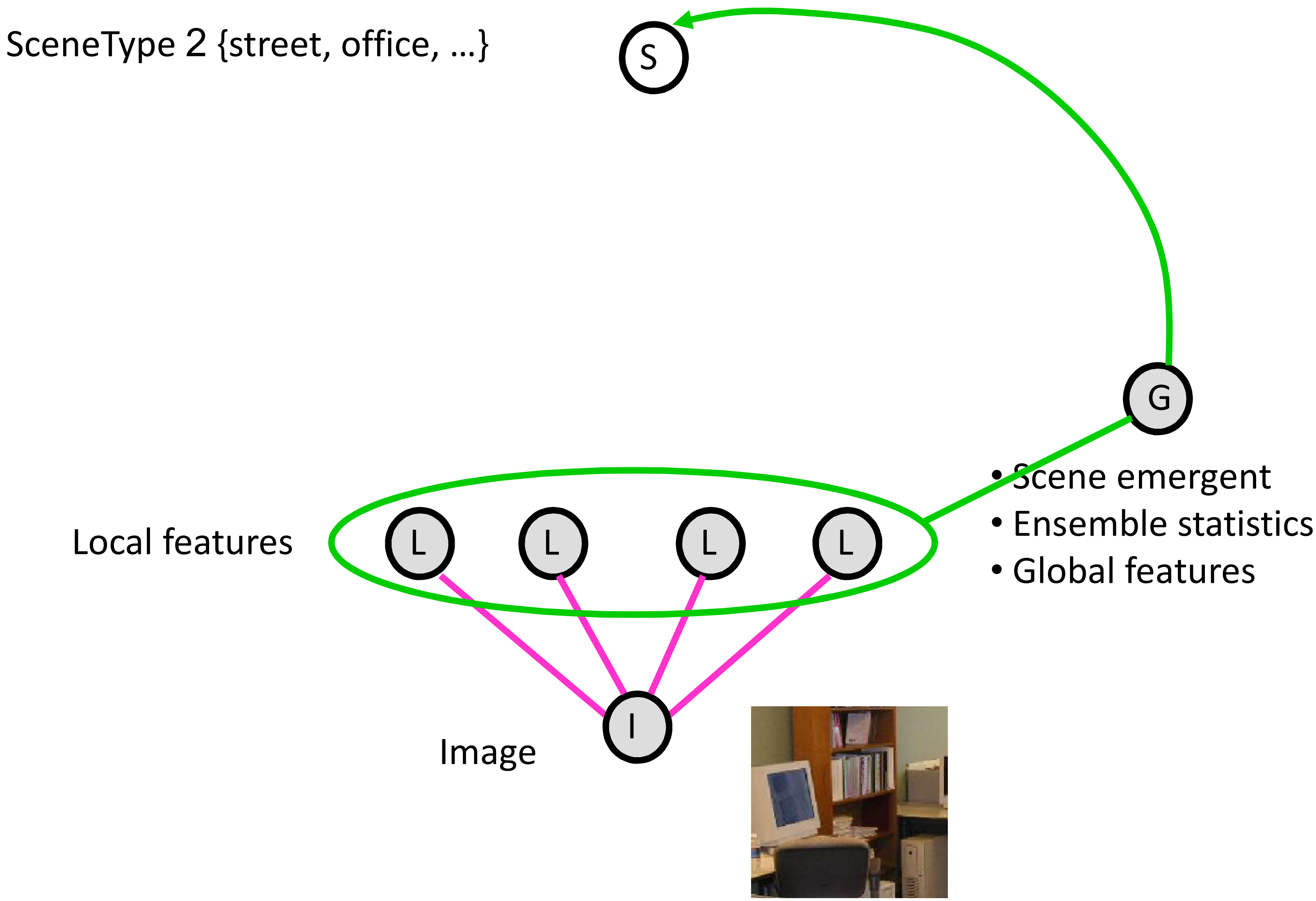


Conclusion: observers had more accurate representation of the mean than of the individual members of the set.

From scenes to objects



How far can we go without objects?

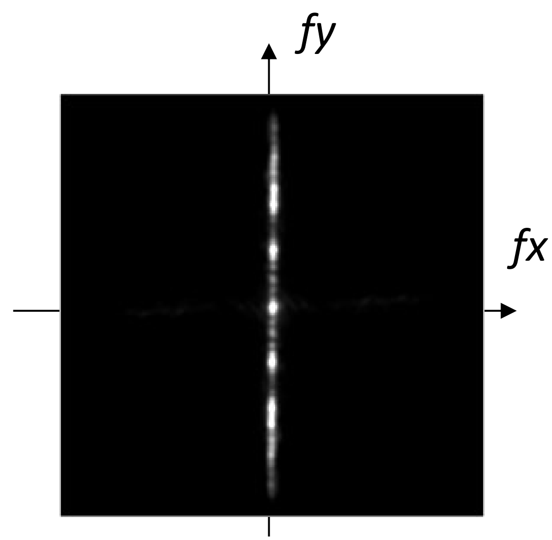


- **Scenes as textures**

A simple texture descriptor

Magnitude of the Fourier Transform

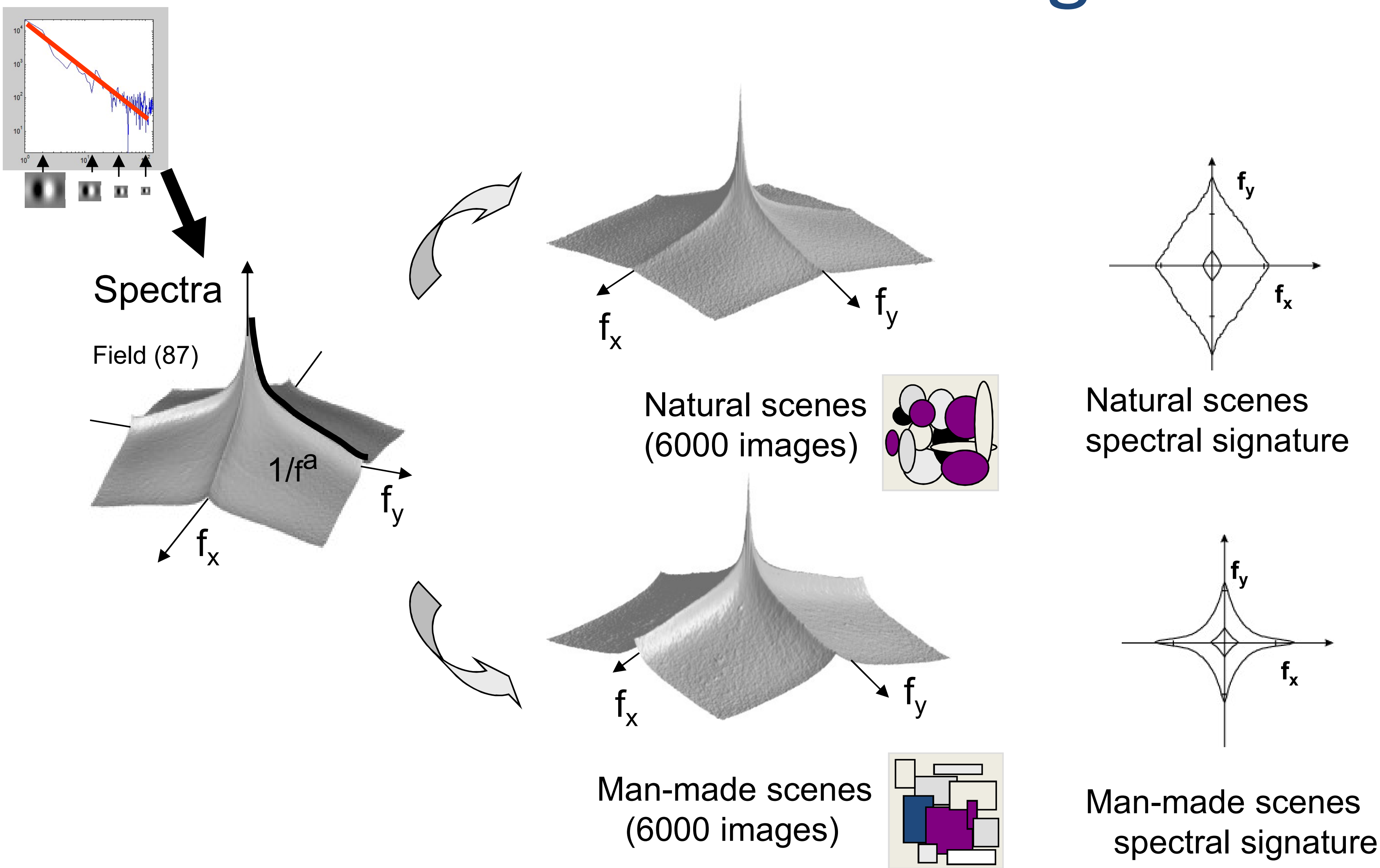
$$A(f_x, f_y) = \left| \sum_{x,y} i(x,y) e^{-2\pi j (f_x x + f_y y)} \right|$$



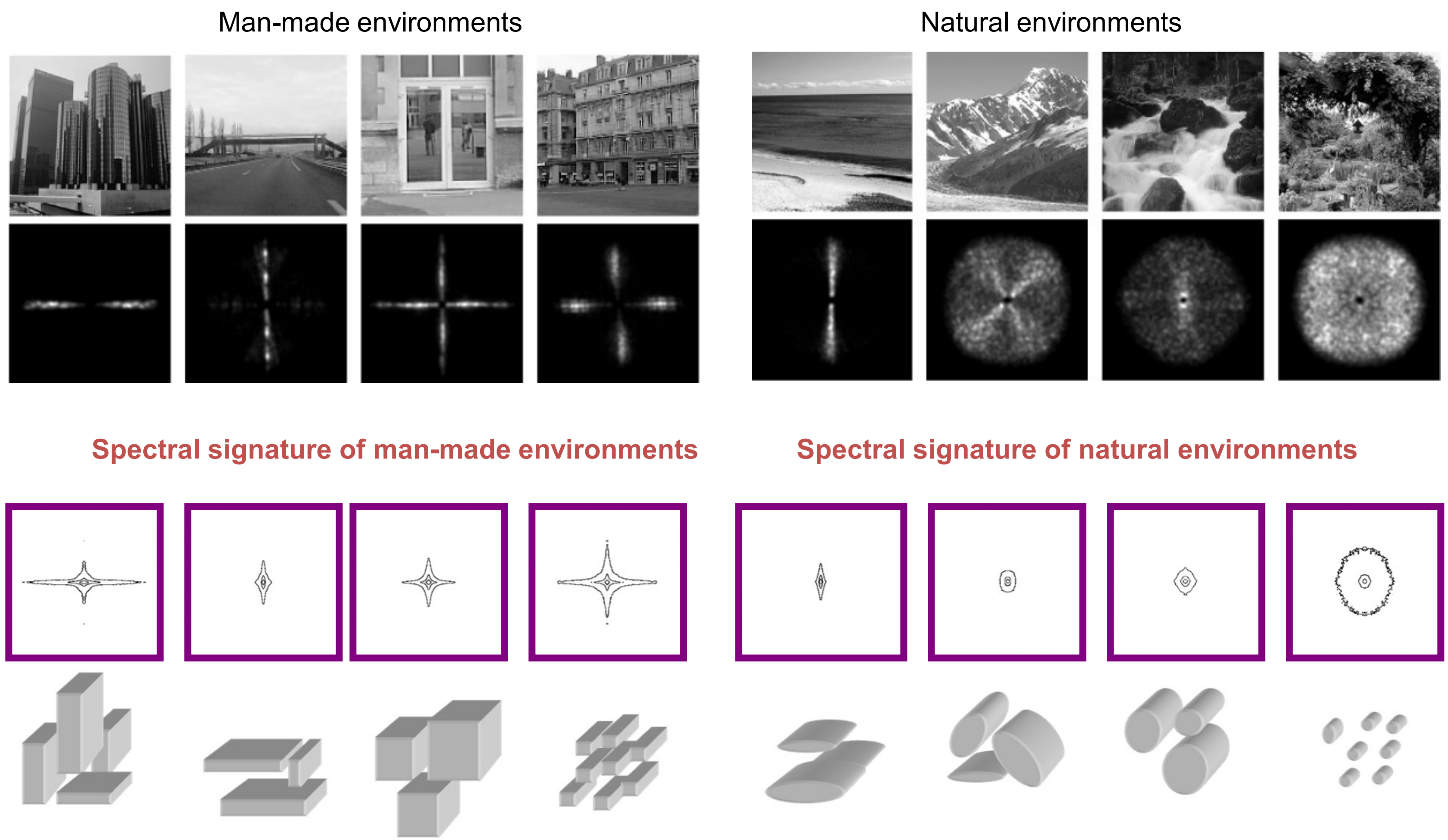
Magnitude of the Fourier Transform encodes unlocalized information about dominant orientations and scales in the image.

The magnitude of the Fourier transform does not contain information about object identities and spatial arrangements.

Statistics of Scene Categories



Statistics of Scene Categories



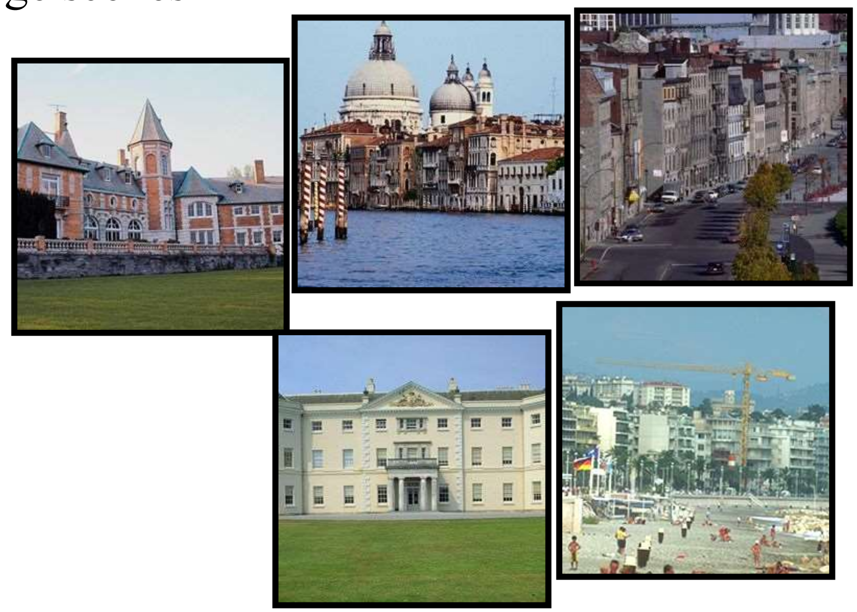
Oliva et al (99), Oliva & Torralba (01)

Look at Mumford's work for models…

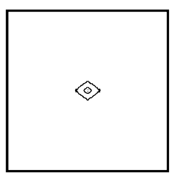
Image Statistics and Scene Scale

Close-up views

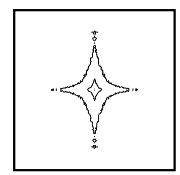
Large scenes



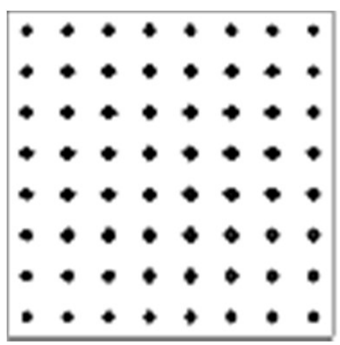
On average, low clutter



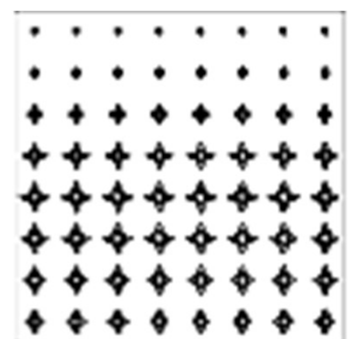
On average, highly cluttered



Point view is unconstrained

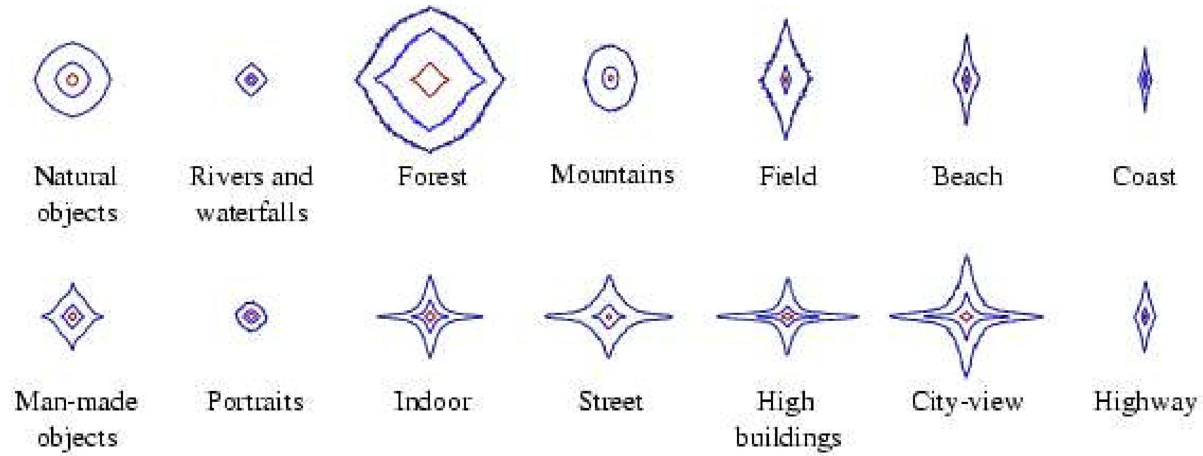


Point view is strongly constrained

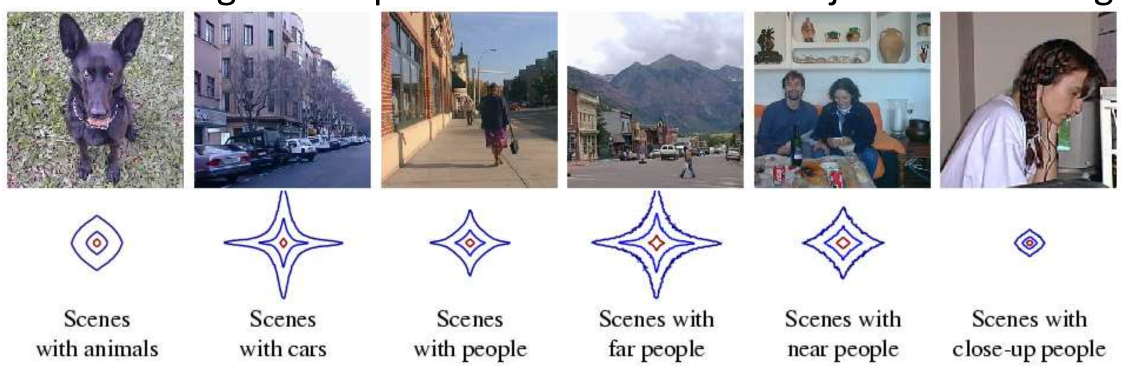


Statistics of Scene Categories

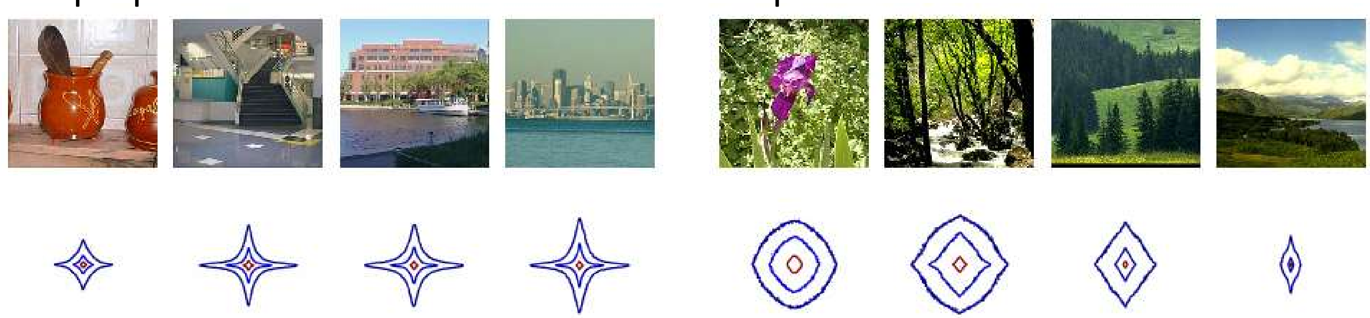
- The statistics of orientations and scales across the image differ between scene categories:



- also differ when conditioning for the presence or absence of objects in the image:



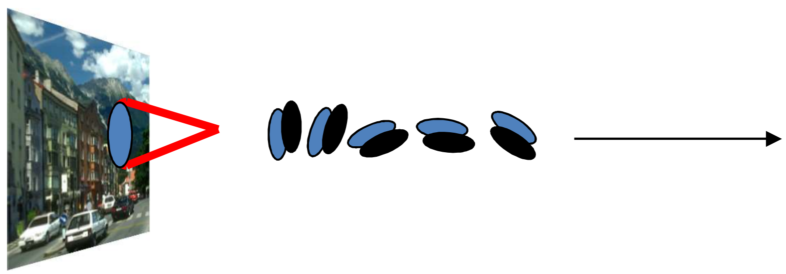
- or for different properties of the scene like the mean depth:



- **Gist**
  - **Spatial envelope**
  - **Depth**

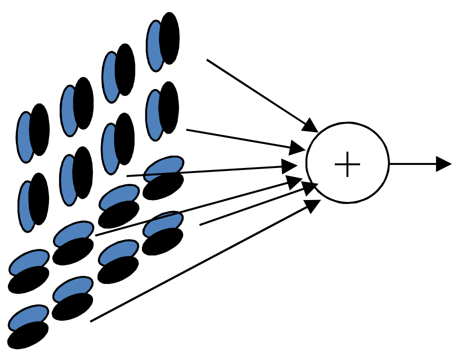
Local and Global features

A set of local features describes image properties at one particular location in the image:



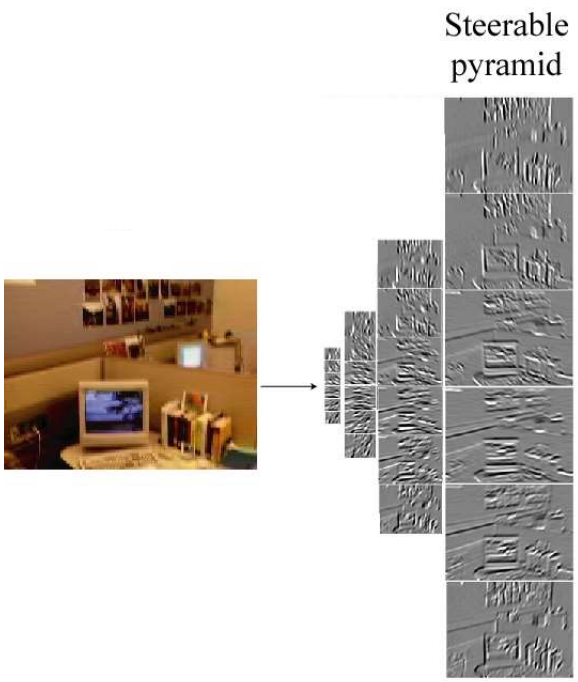
Jet of local orientations and scales

A set of global features provides information about the global image structure without encoding specific objects

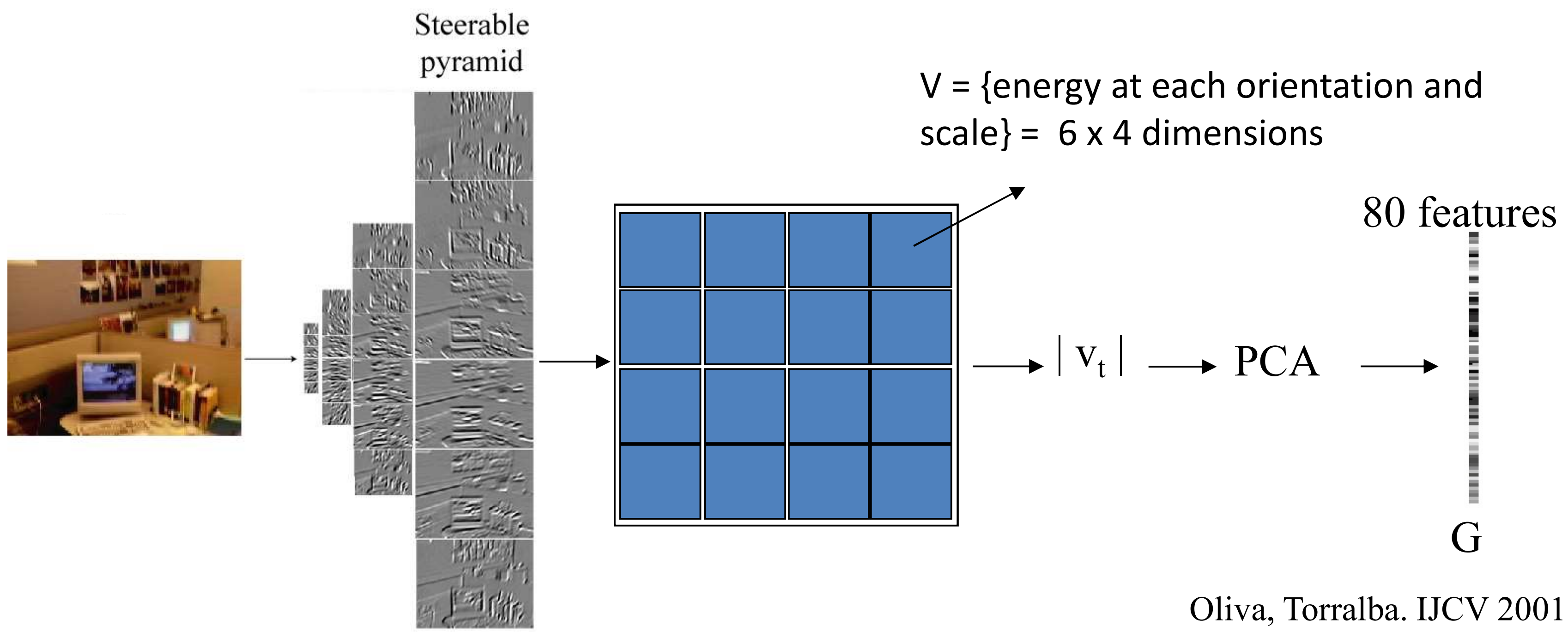


This feature likes images with vertical structures at the top part and horizontal texture at the bottom part (this is a typical composition of an empty street)

Gist descriptor



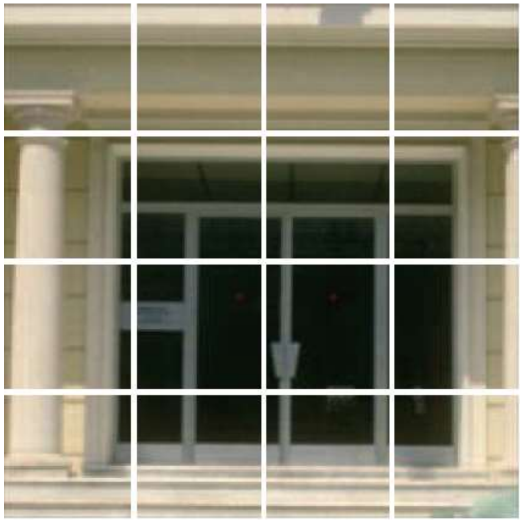
Gist descriptor

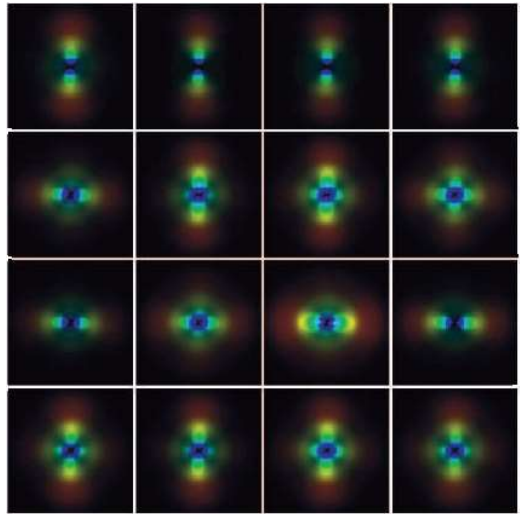


Oliva, Torralba. IJCV 2001

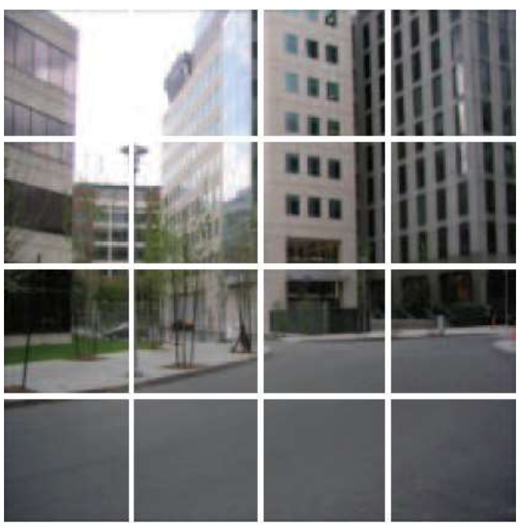
Gist descriptor

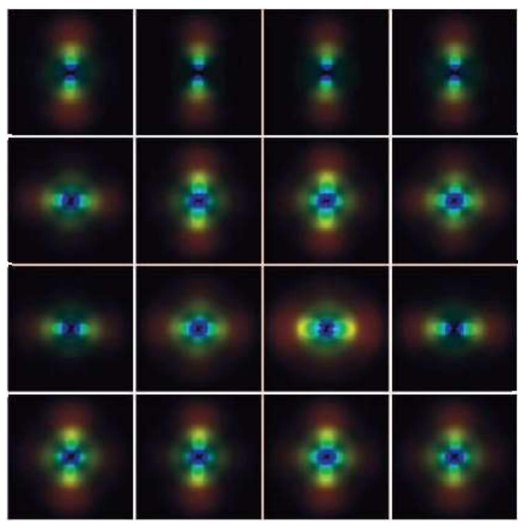
Oliva and Torralba, 2001





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

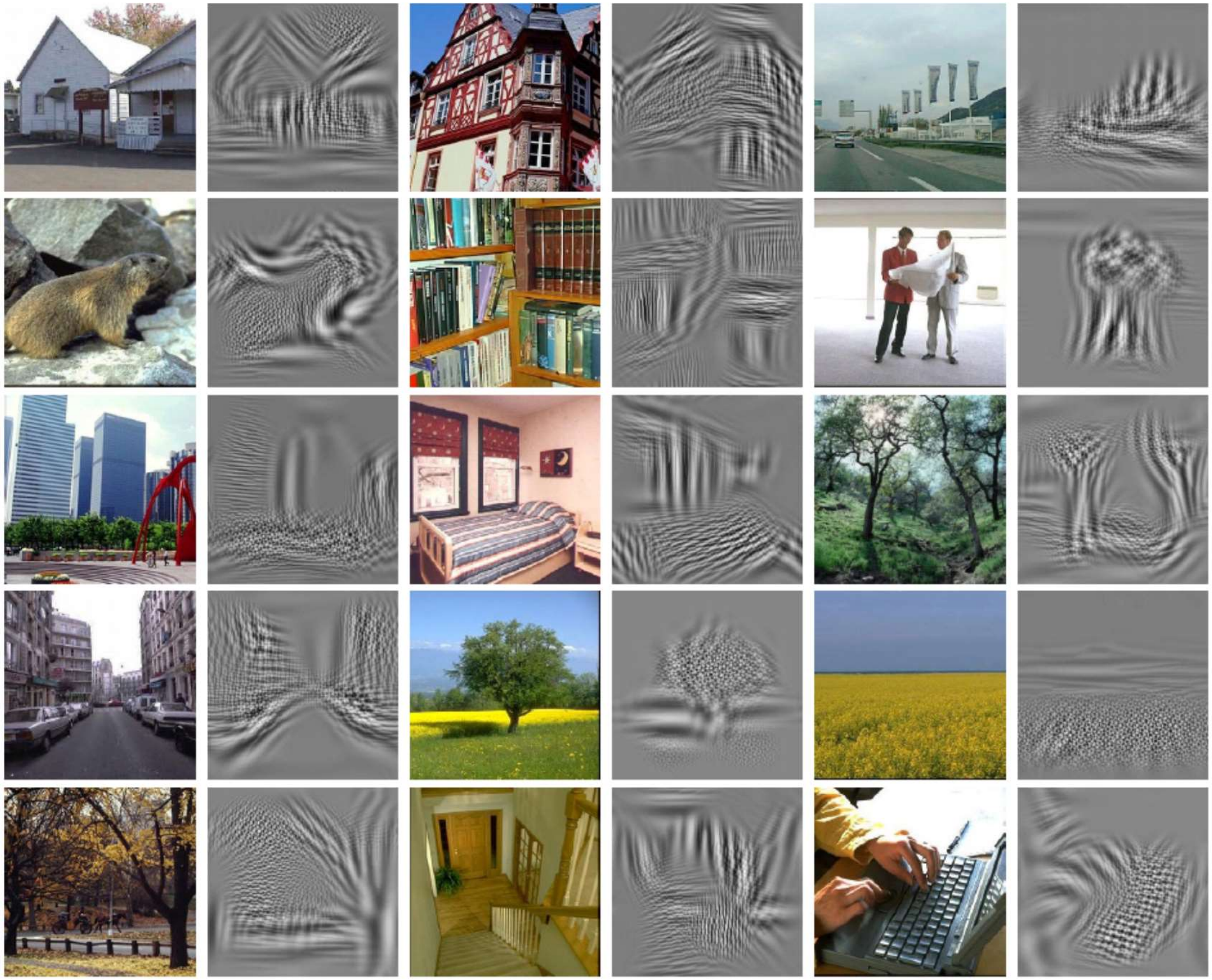




8 orientations
4 scales
x 16 bins
512 dimensions

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; …

Example visual gists



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

Oliva & Torralba (2001)

Scene Perceptual Dimensions

Like a *texture*, a scene could be represented by a set of structural dimensions, but describing surface properties of a *space*.

<u>We use a classification task:</u> observers were given a set of scene pictures and were asked to organize them into groups of similar shape, similar global aspect, similar spatial structure.

They were explicitly told to not use a criteria related to the objects or a scene semantic group.

Oliva et al (99), Oliva & Torralba (01)

Scene Perceptual Dimensions

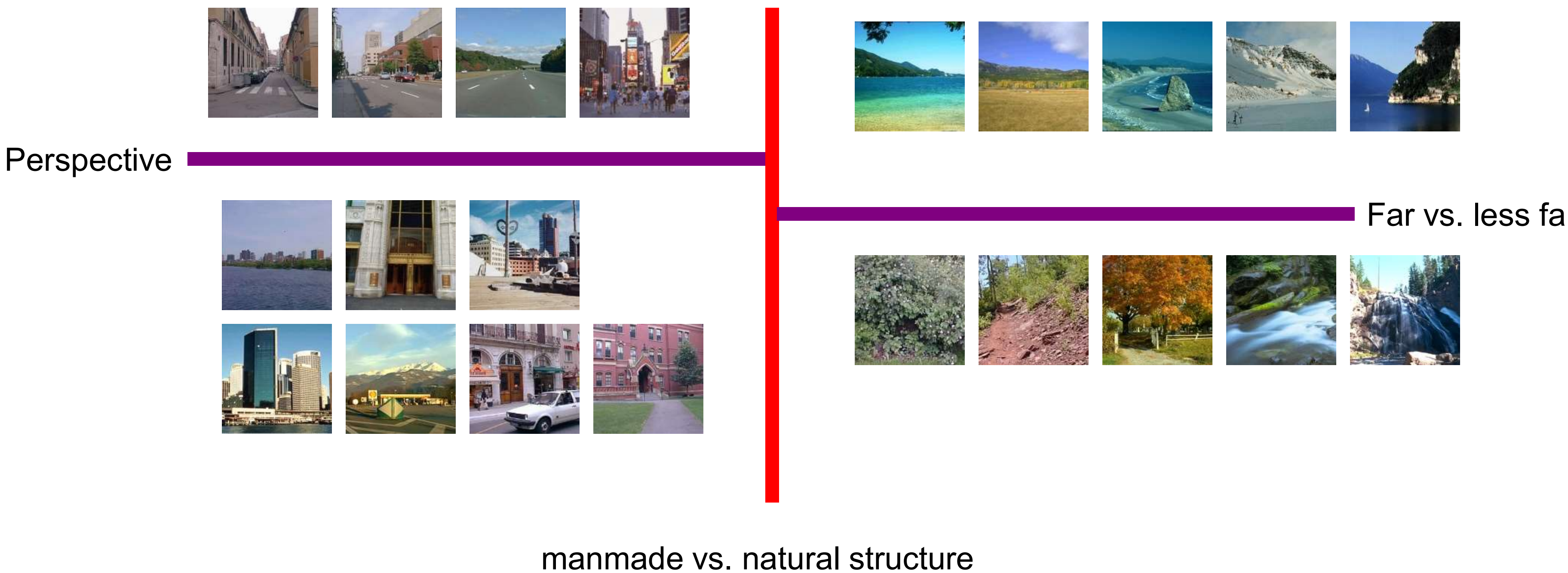
<u>Task:</u> The task consisted in 3 steps: the first step was to divide the pictures into 2 groups of similar shape.

Example: manmade vs. natural structure

Oliva et al (99), Oliva & Torralba (01)

Scene Perceptual Dimensions

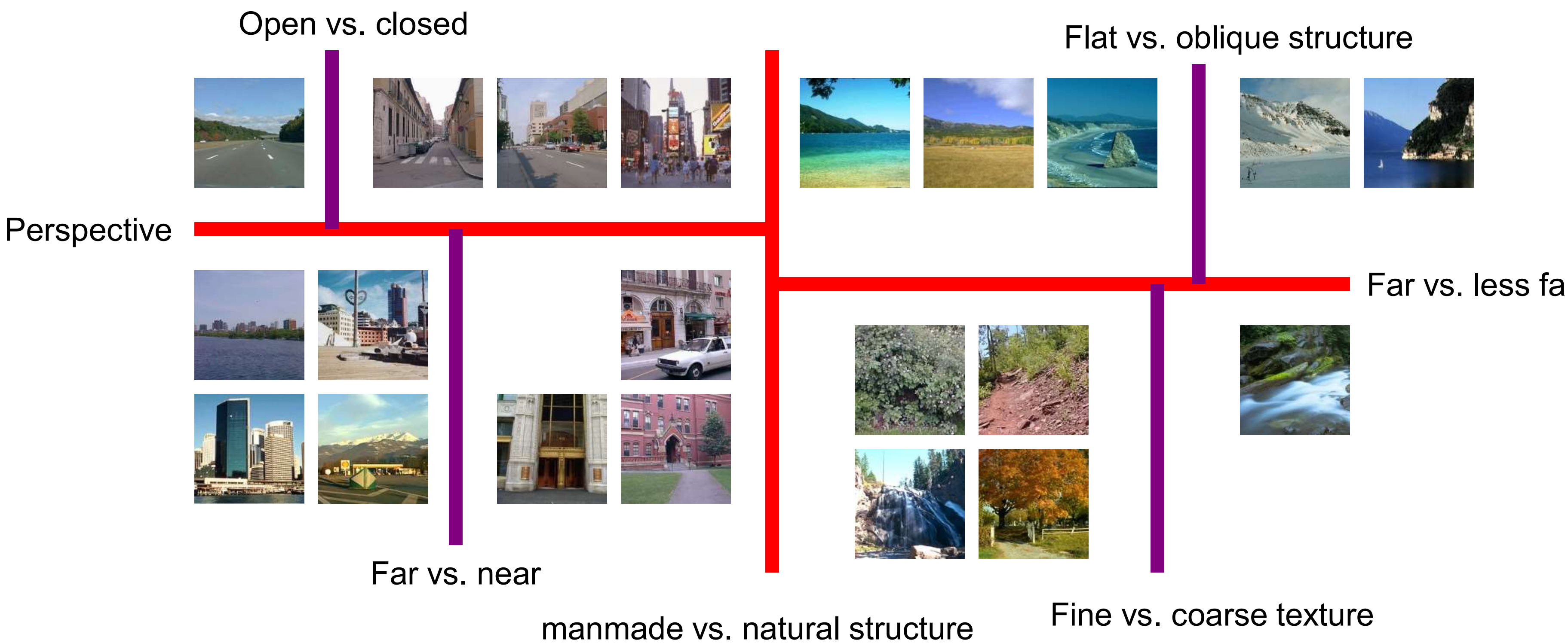
<u>Task:</u> The second step was to split each of the 2 groups in two more subdivisions.



Oliva et al (99), Oliva & Torralba (01)

Scene Perceptual Dimensions

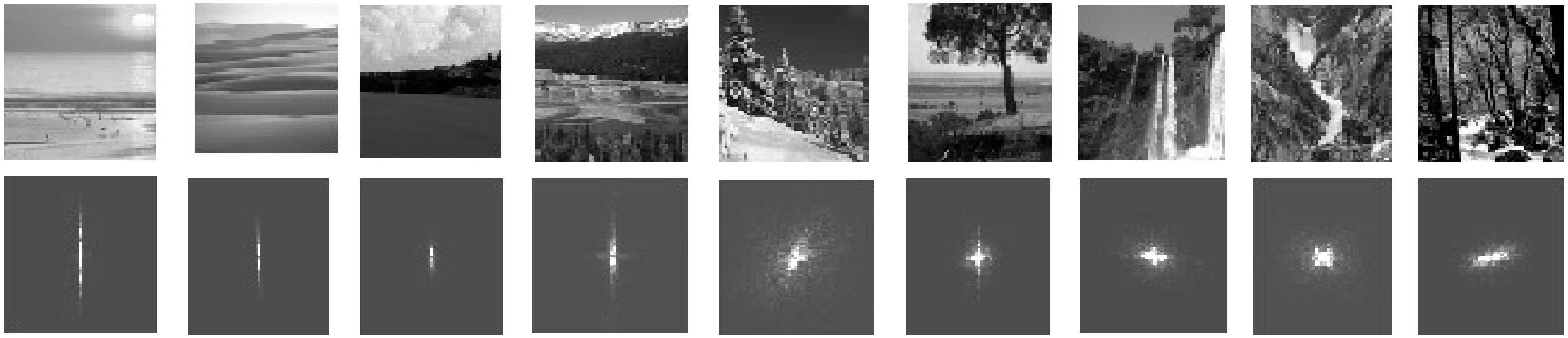
Task: In the third step, participants split the 4 groups in two more groups.



Oliva et al (99), Oliva & Torralba (01)

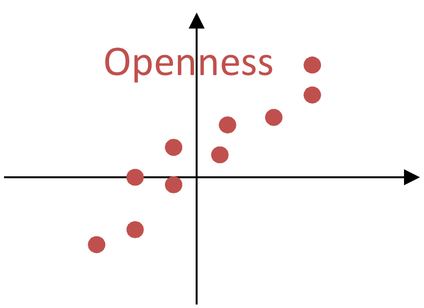
Estimation of a space descriptor: *openness*

From open scenes.... ________________________________ to closed scenes.

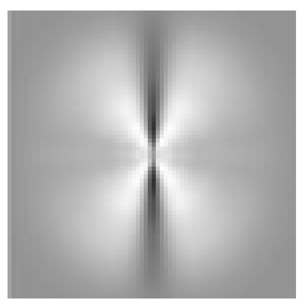


From vertical components ____________________________ to isotropic components.

Regression: we look for a weighting of the spectral components so that we can reproduce the same ordinal ranking as the subjects.

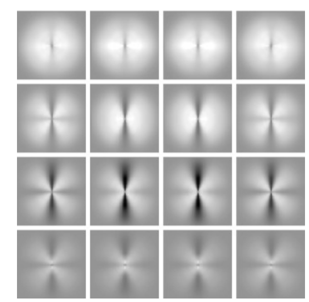


Oliva et al (99), Oliva & Torralba (01)



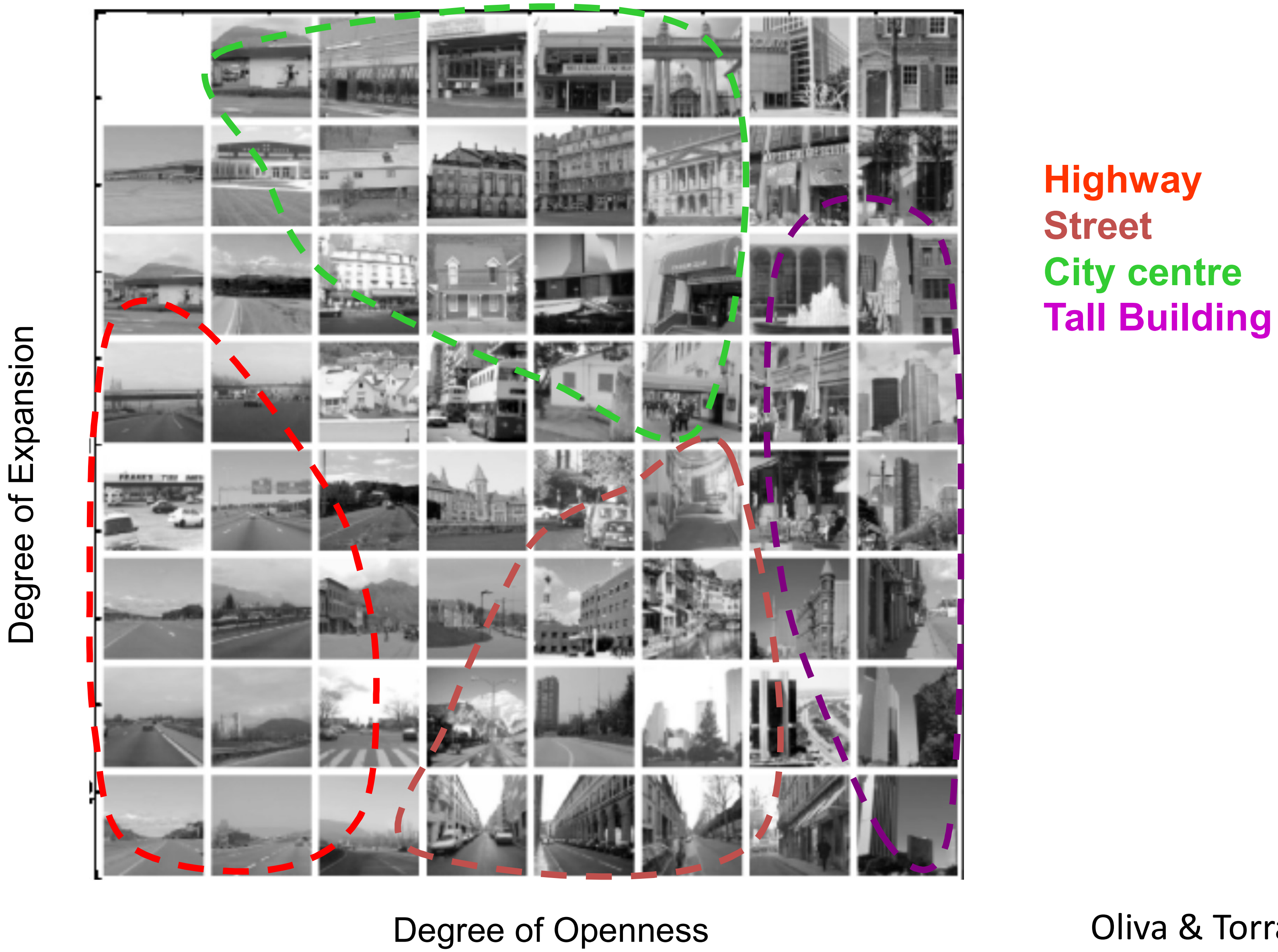
Weighting of the spectral features

The template represents
the best weighting of the
spectral components in order
to estimate the degree of *openness*



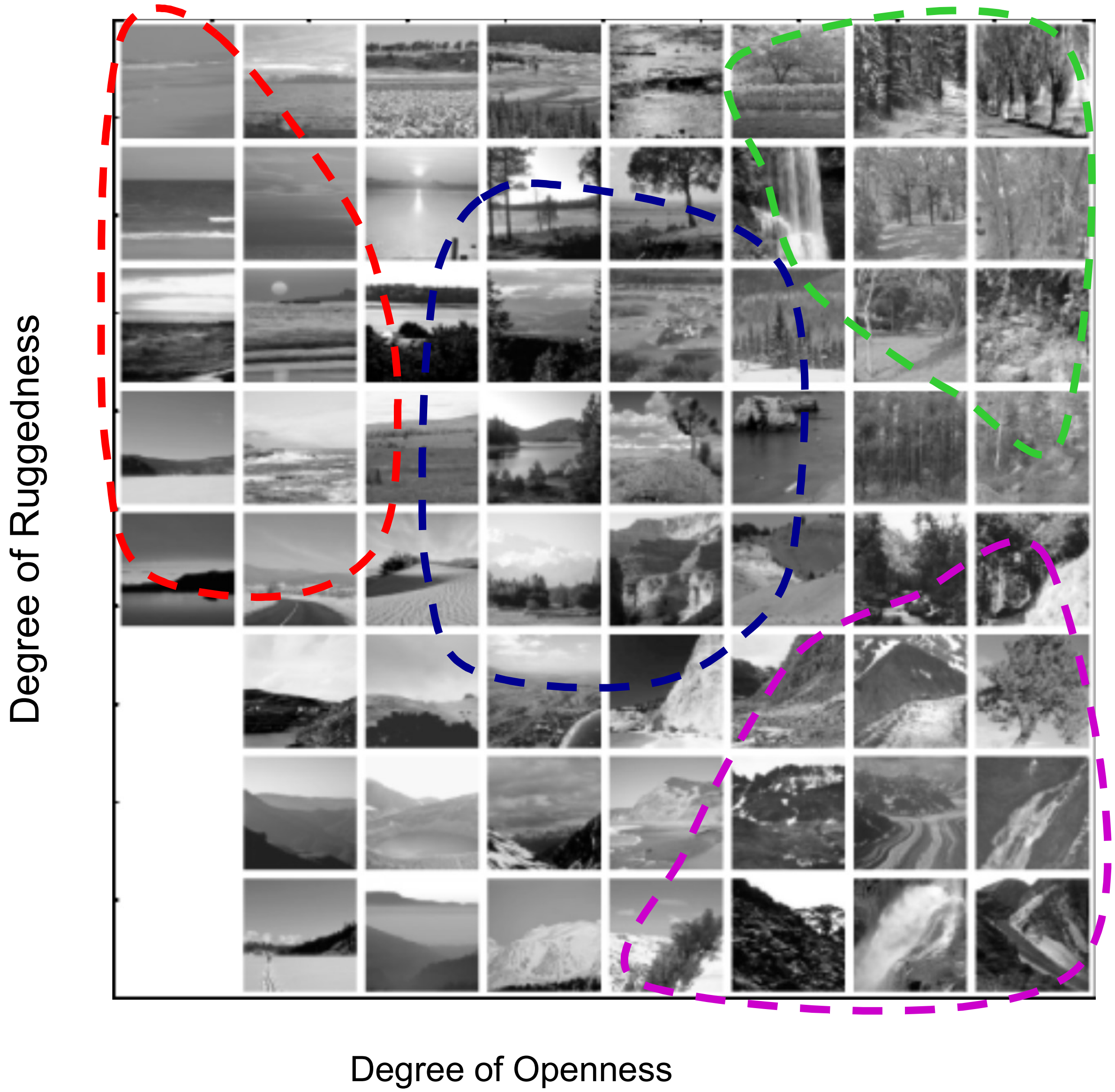
Layout of weighted spectral features

Spatial envelope:
a continuous space of scenes



Oliva & Torralba, 2001

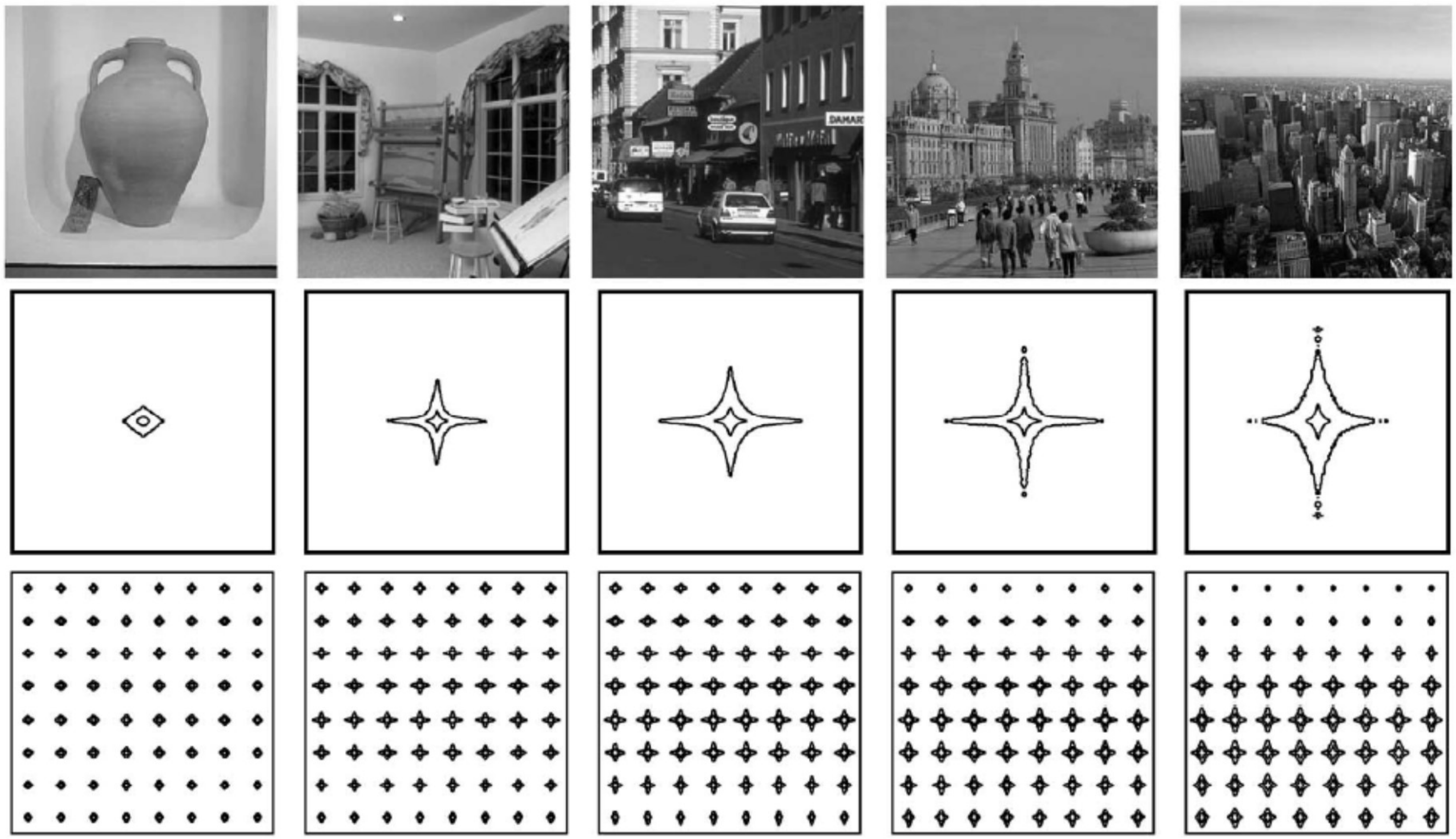
Spatial envelope:
a continuous space of scenes



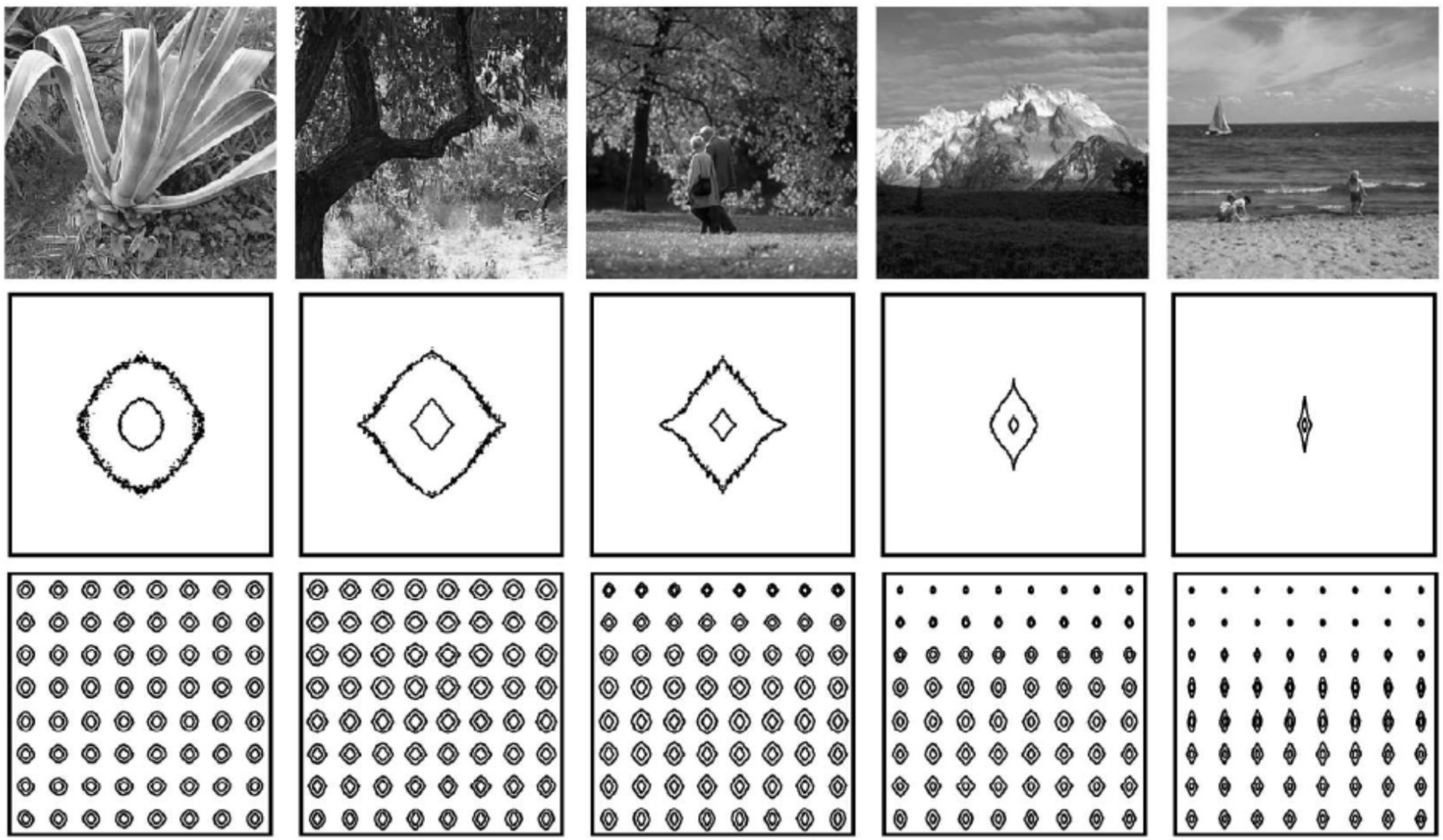
Coast
Countryside
Forest
Mountain

Oliva & Torralba, 2001

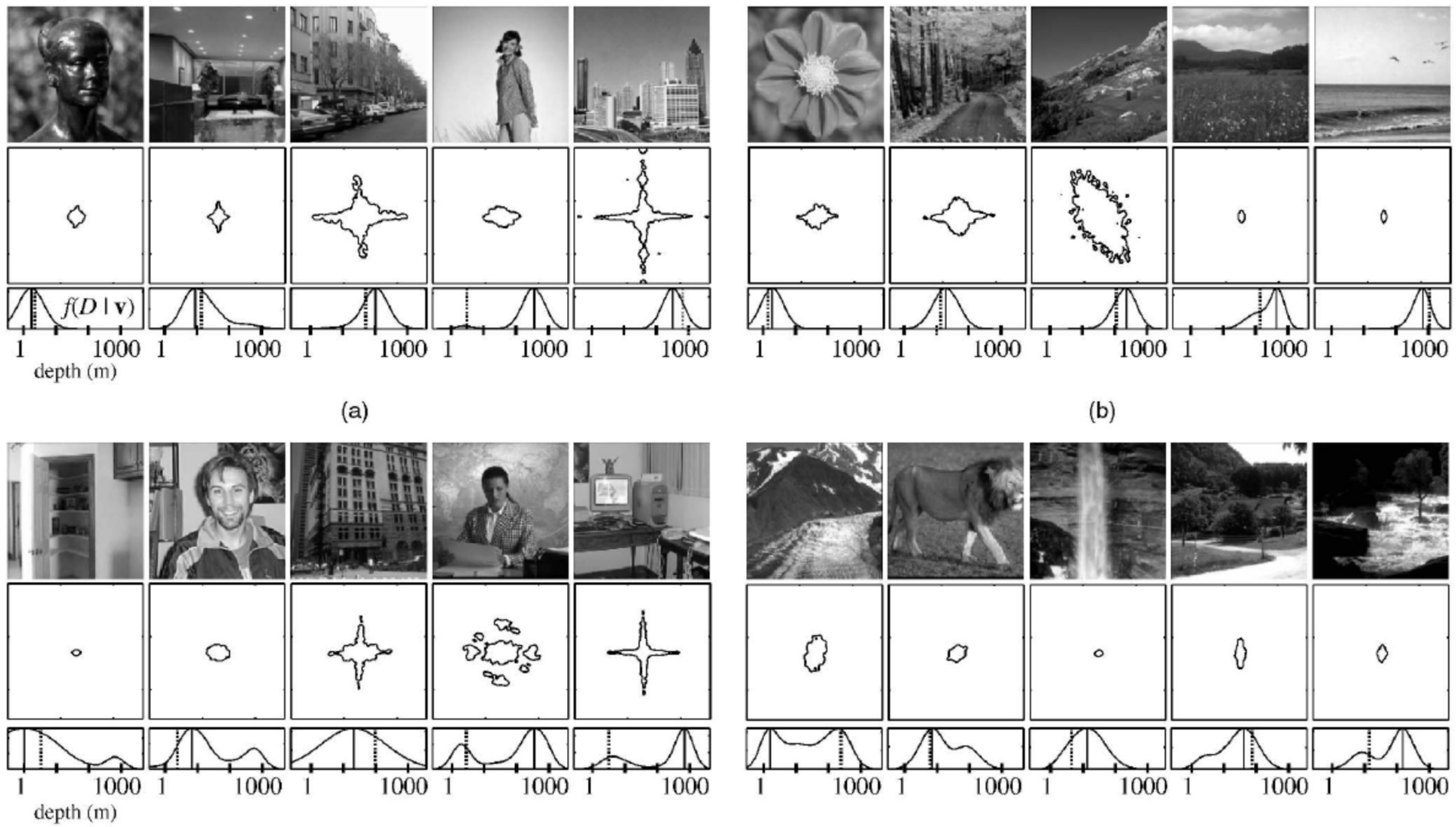
Examples (man-made)



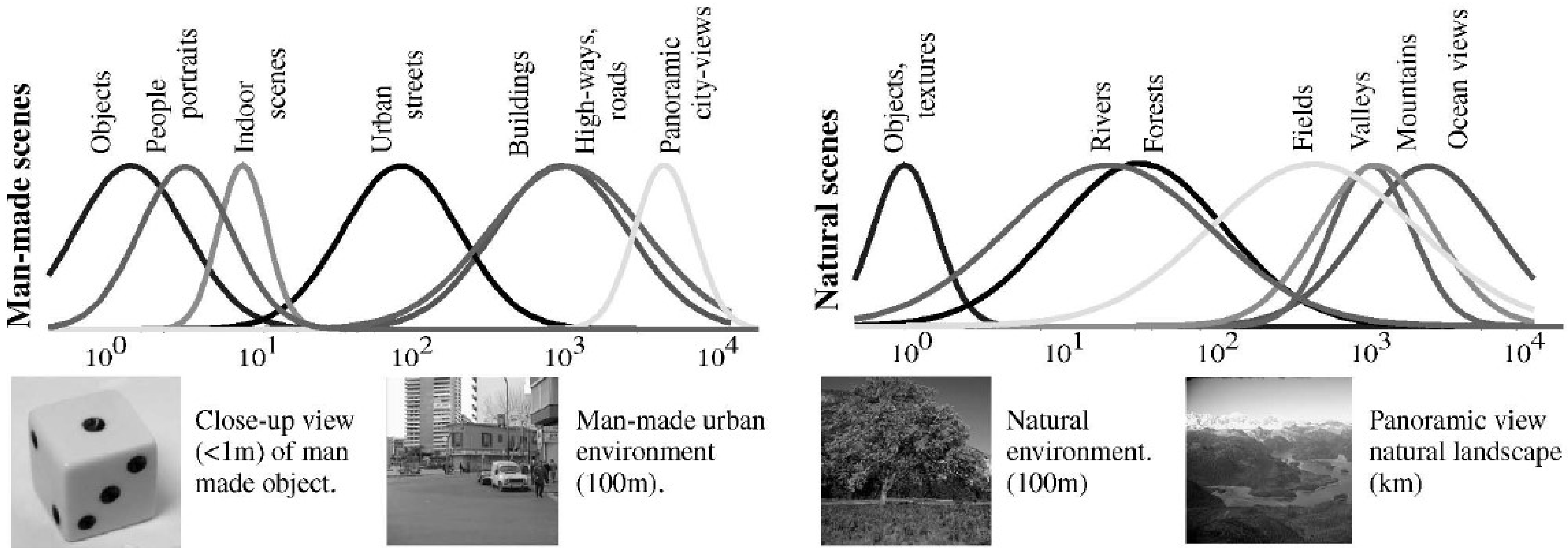
Examples (Natural)



Some Results



Distribution of Scene Categories as a function of mean depth.



Multiple-Level Categorization

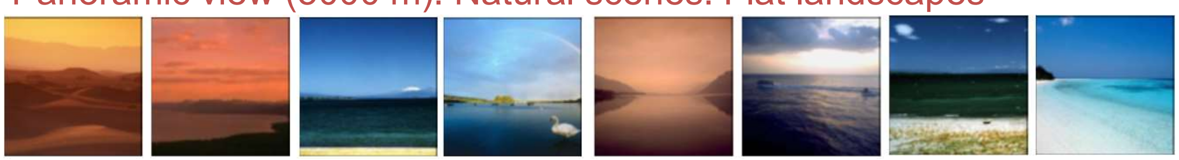
Panoramic view (5000 m)

From superordinate category to ….

Panoramic view (5000 m). Manmade scenes.

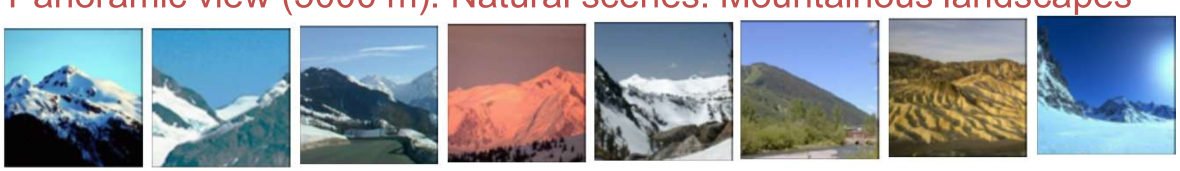
Panoramic view (5000 m). Natural scenes.

Panoramic view (5000 m). Natural scenes. Flat landscapes



…. Basic-level category coast …..

Panoramic view (5000 m). Natural scenes. Mountainous landscapes



…. Basic-level category mountain …..

Oliva et al (99), Oliva & Torralba (01)

Oliva & Torralba (2003)

Space-centered description


Close-up view (1m)


Close-up view (1m)
Natural scene.


Close-up view (1m)
Natural scene.


Natural scene.
Close-up view (1m)


Close-up view (1m)
Man-made object.


Small space (6m)
Man-made scene.
Closed environment.


Small space (3m)
Man-made scene.
Enclosed environment.


Small space (9m)
Man-made scene.
Closed environment.
Empty space.


Small space (10m)
Man-made scene.
Closed environment.
Empty space.


Large space (140m)
Man-made scene.
Semiclose environment.


Large space (120m)
Natural scene.
Closed environment.


Large space (80m)
Man-made scene.
Semiopen environment.
Space in perspective.


Panoramic view (3500m)
Man-made scene.
Open environment.
Space in perspective.
Empty space.


Large space (200m)
Natural scene.
Semiopen environment.

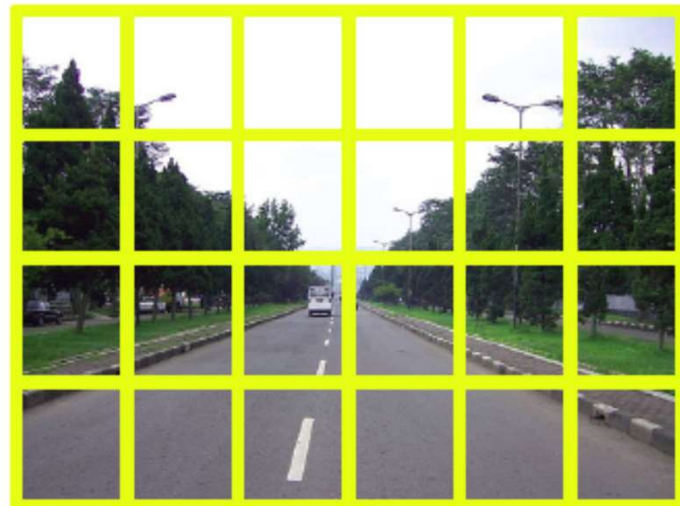

Panoramic view (4000m)
Natural scene.
Open environment.
Flat view.

Oliva et al (99), Oliva & Torralba (01)

Scene matching

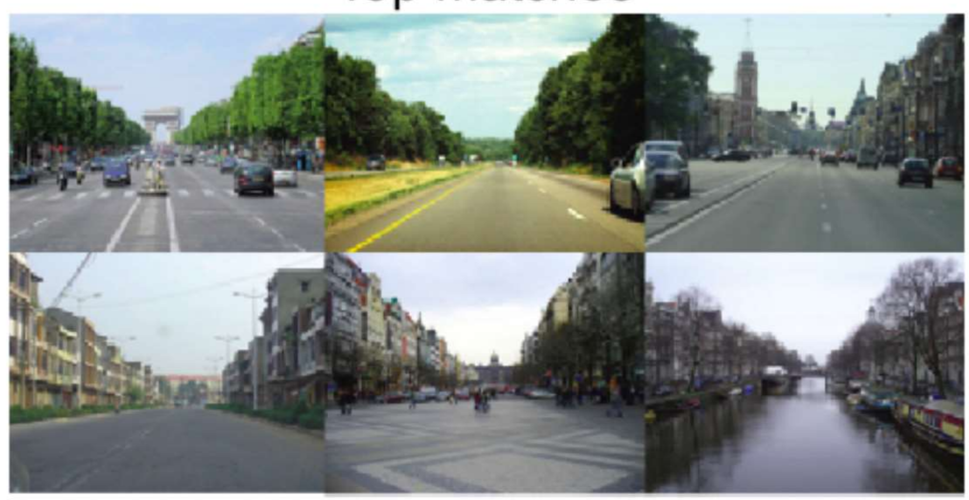
Query image

GIST



Best match

Top matches



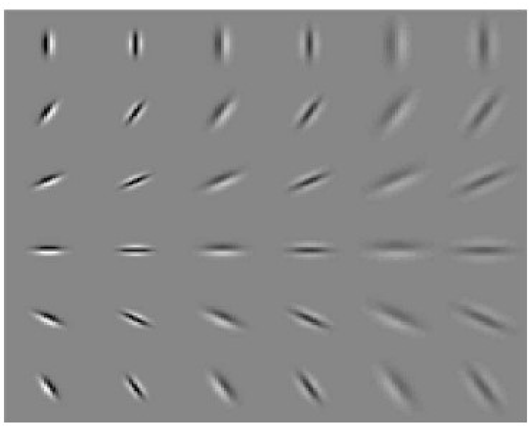
- Bag of words
  - Sift
  - Visual words
  - Pyramid matching
  - SVM

Scene → Bag of ‘words’

Textons

Vector of filter responses at each pixel

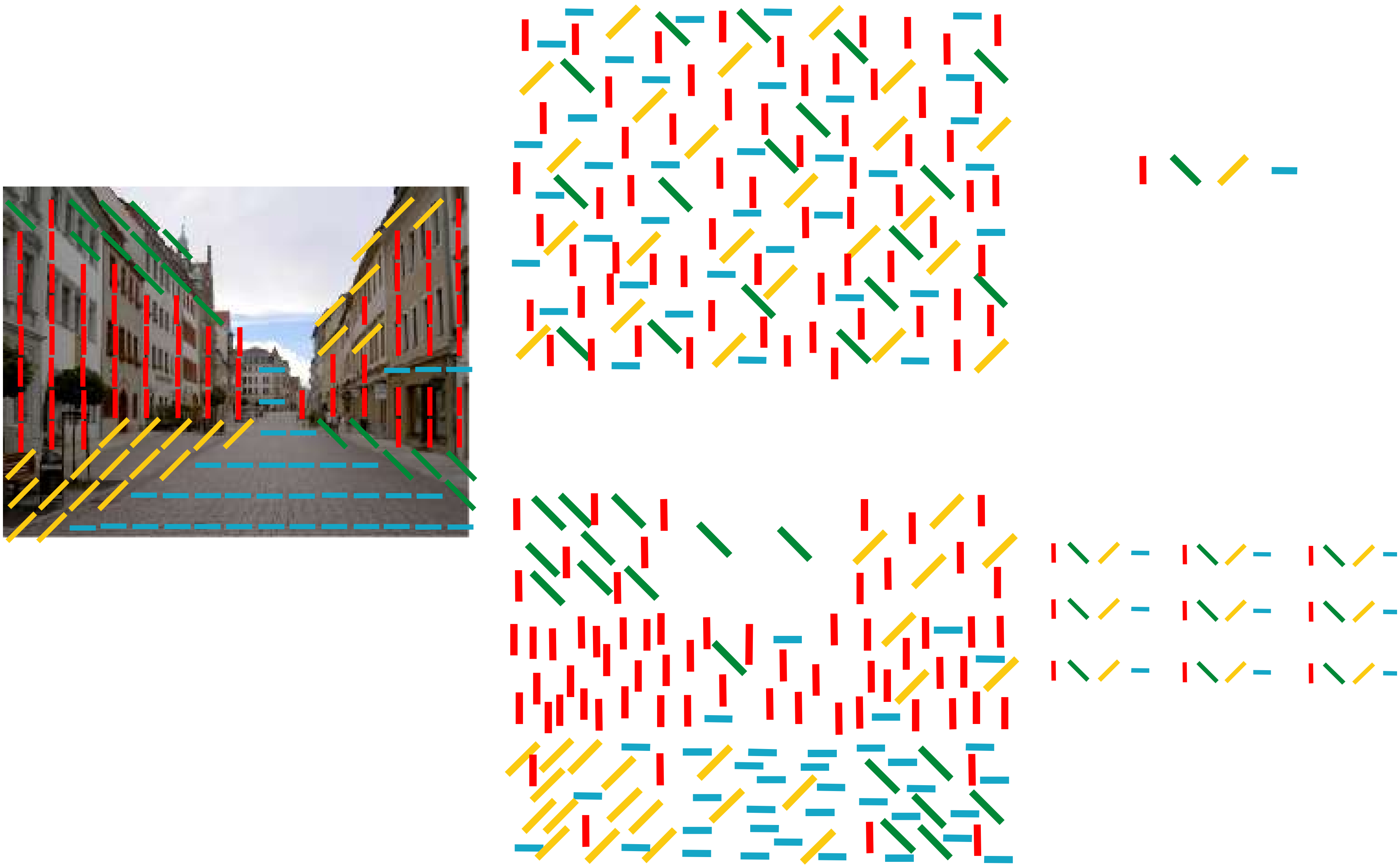
Kmeans over a set of vectors on a collection of images



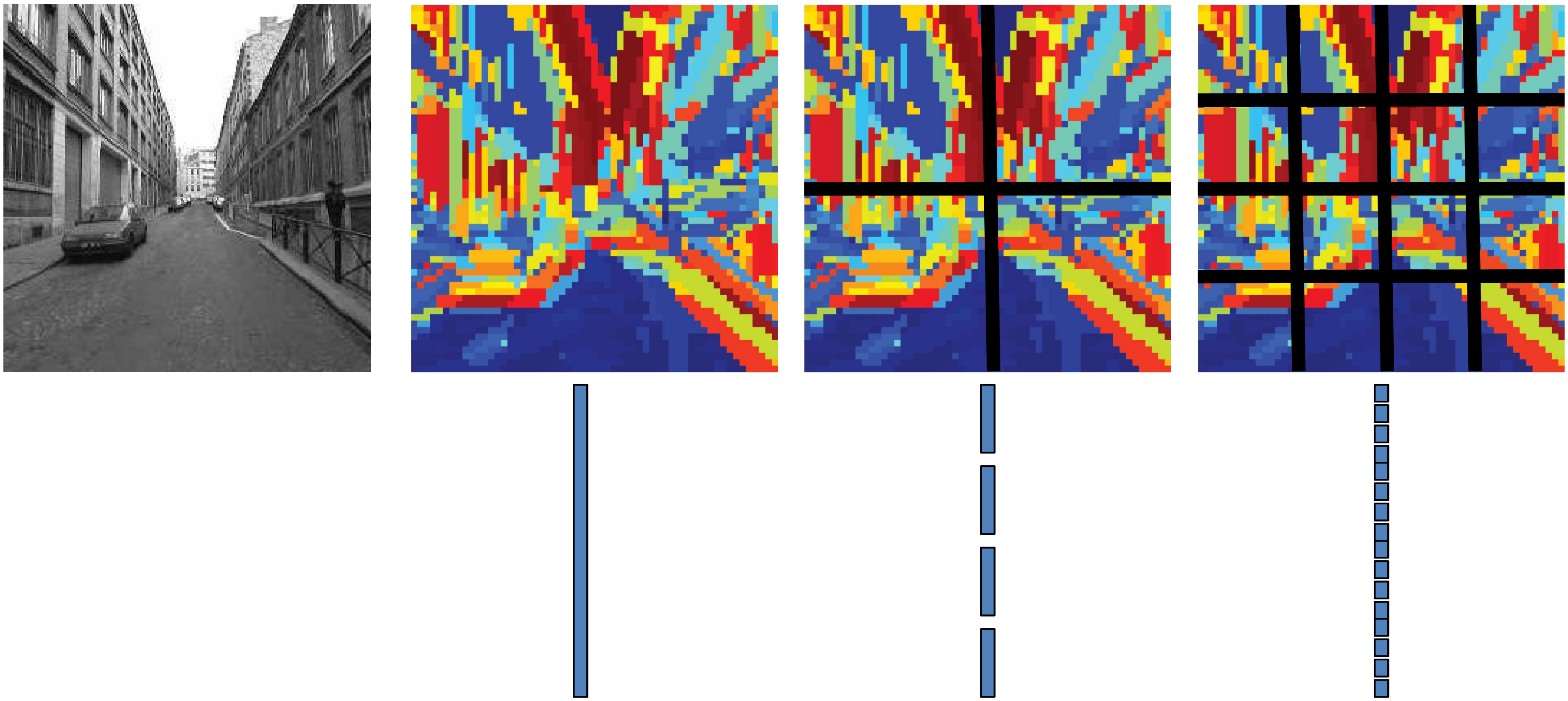
Filter bank

Malik, Belongie, Shi, Leung, 1999

Bag of words

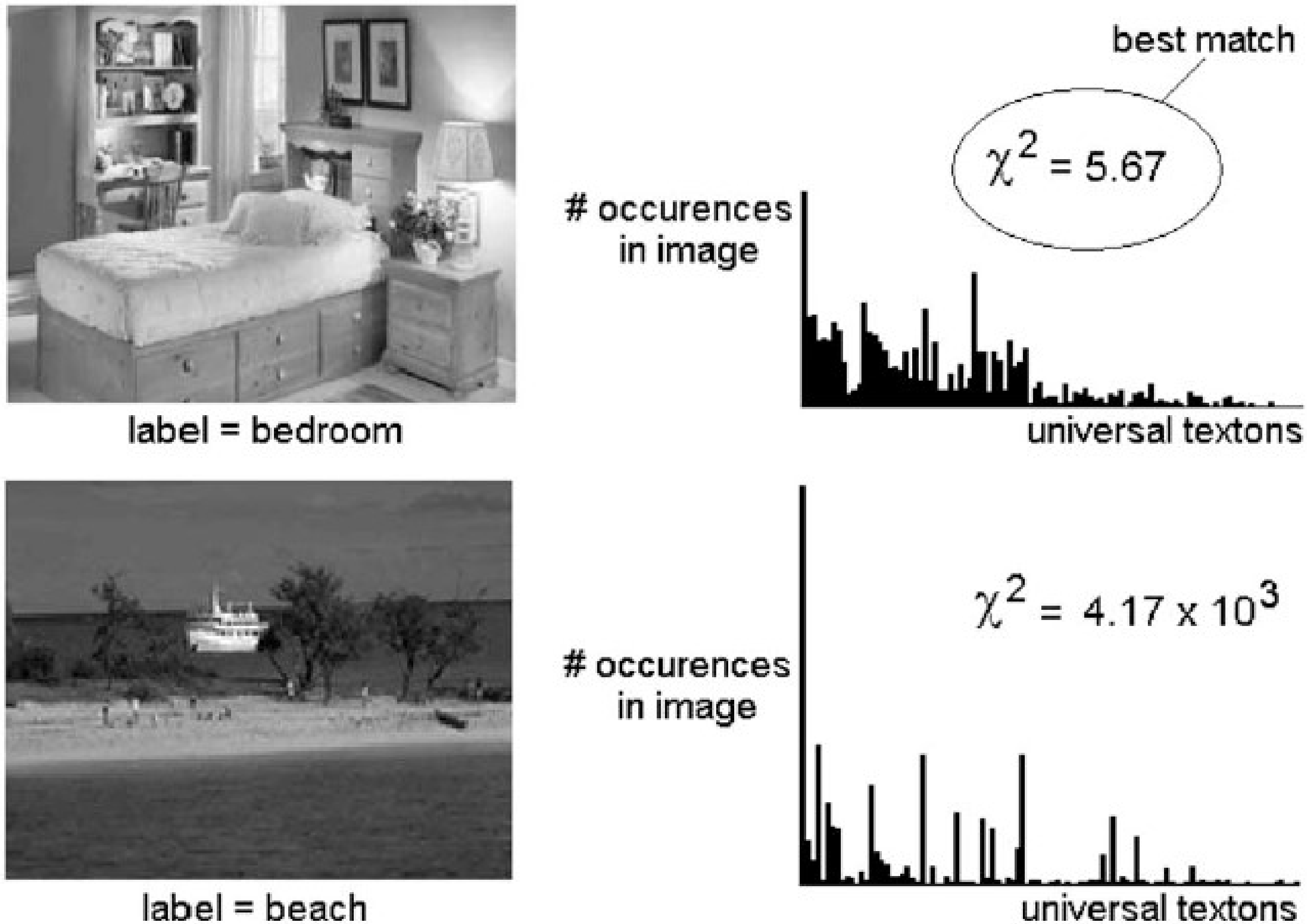


Bag of words & spatial pyramid matching



Grauman & Darell,
S. Lazebnik, et al, CVPR 2006

Textons



Walker, Malik, 2004

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

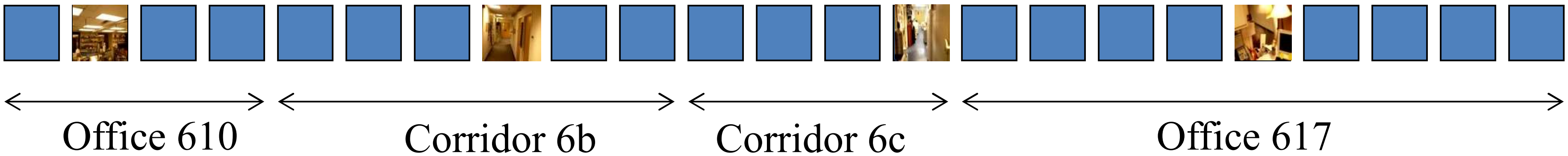
Open country

Kitchen

Store

- **Classification results and applications**
  - **Categorization**
  - **Computing image similarities**
  - **Place recognition**

Training for scene recognition



Scene categorization:

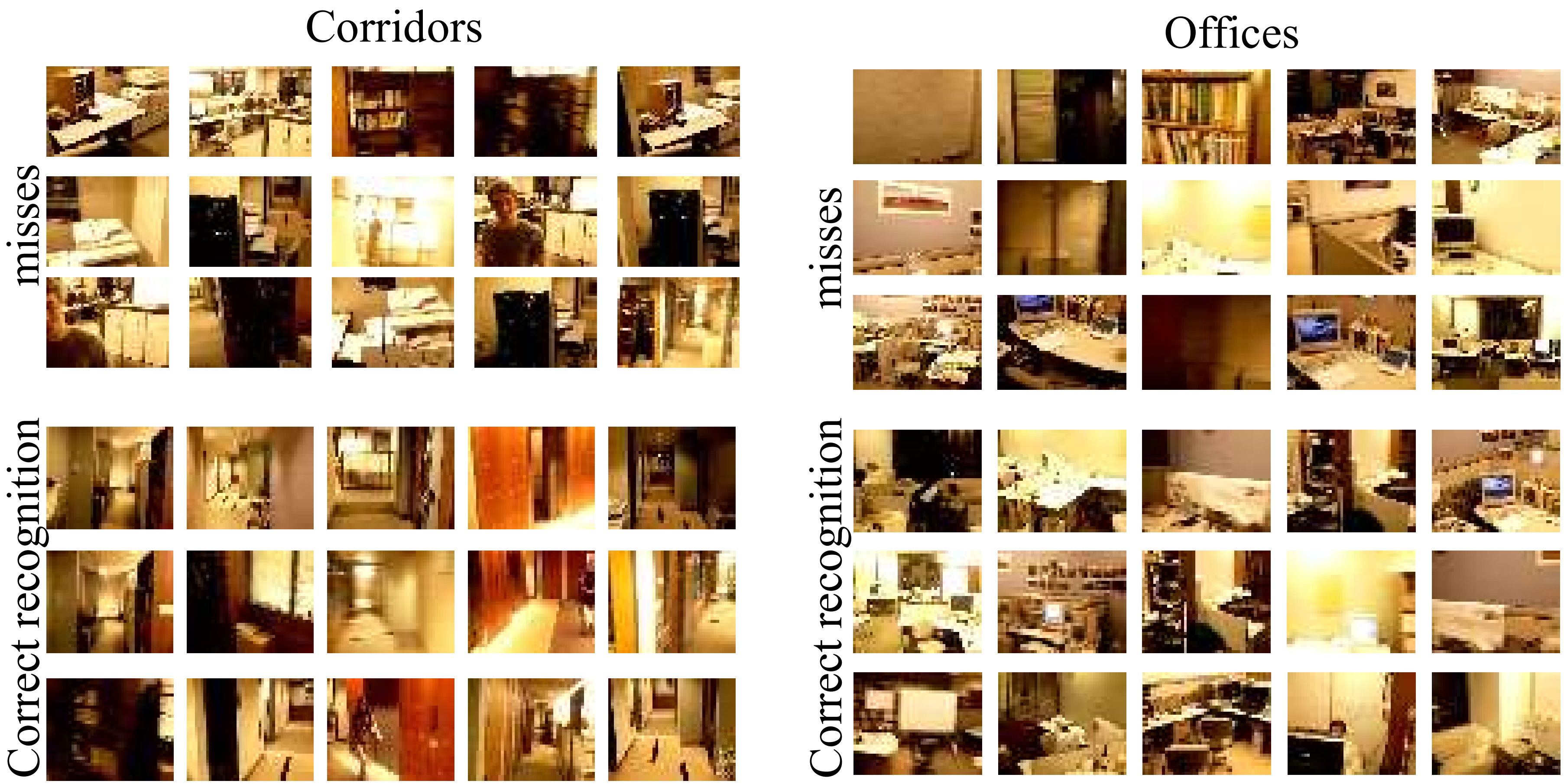
3 categories

Place identification:

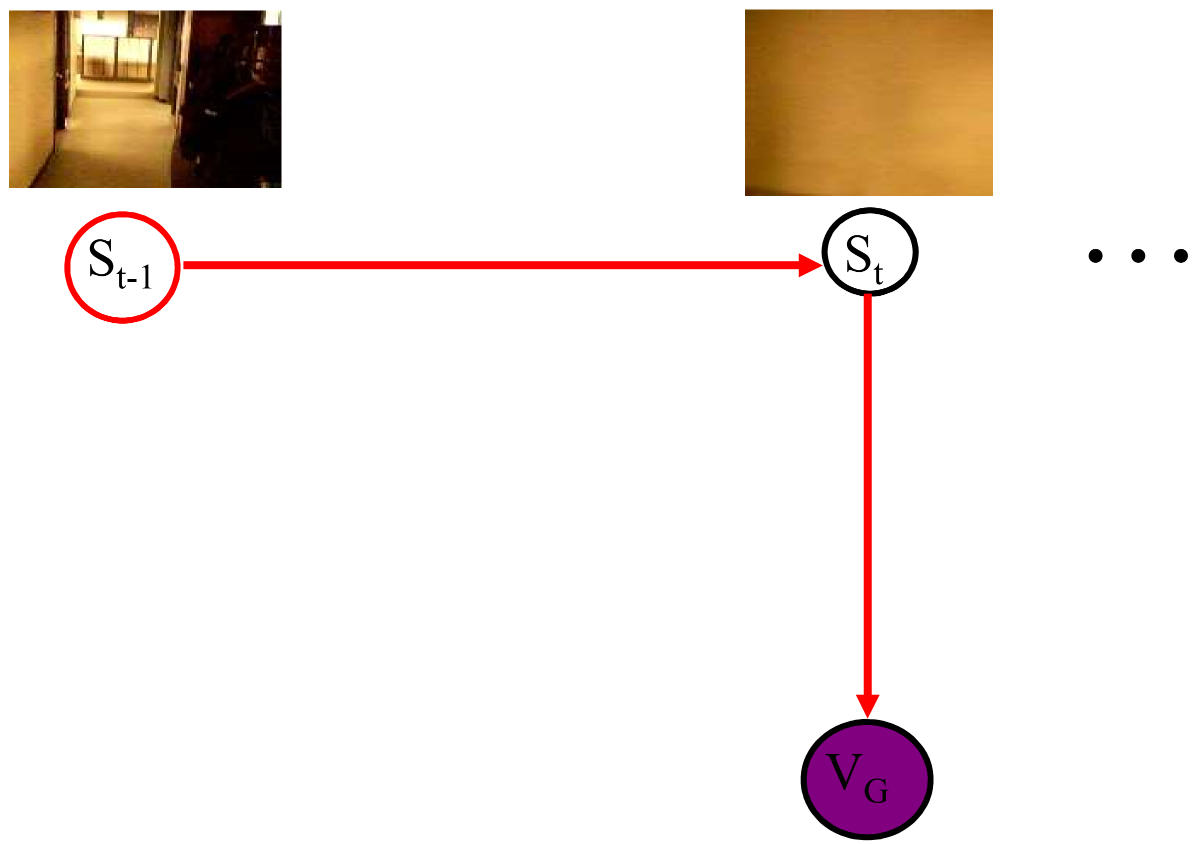
...

62 places

Classifying isolated scene views can be hard



Scene recognition over time



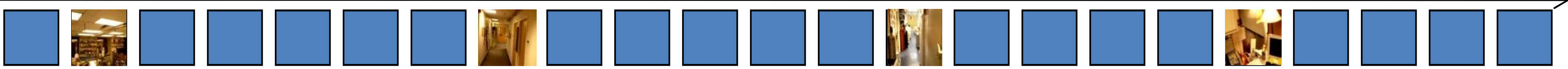
Cf. topological localization in robotics

Torralba, Murphy, Freeman, Rubin, ICCV 2003

Input

Hidden Markov Model

t=0 1 2 3 ....



Gist: v_t

Place: S_t

Output: estimation S_t

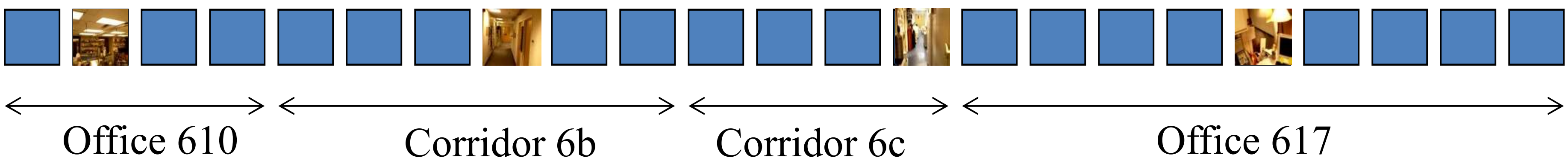
$P(\underbrace{S_t}_{\text{Location}} \mid \underbrace{v_{1:t}}_{\text{Sequence gist features}})$

We use a HMM to estimate the location recursively:

$$\underbrace{P(q_t \mid v_{1:t})}_{\text{Probability for each location}} \;\alpha\; \underbrace{p(v_t \mid q_t)}_{\text{Observation likelihood for frame } t} \sum_{q'} \underbrace{P(q_t \mid q'_{t-1})}_{\text{Transition matrix (encodes topology)}} \underbrace{P(q'_{t-1} \mid v_{1:t-1})}_{\text{Previous estimation}}$$

Learning to recognize places

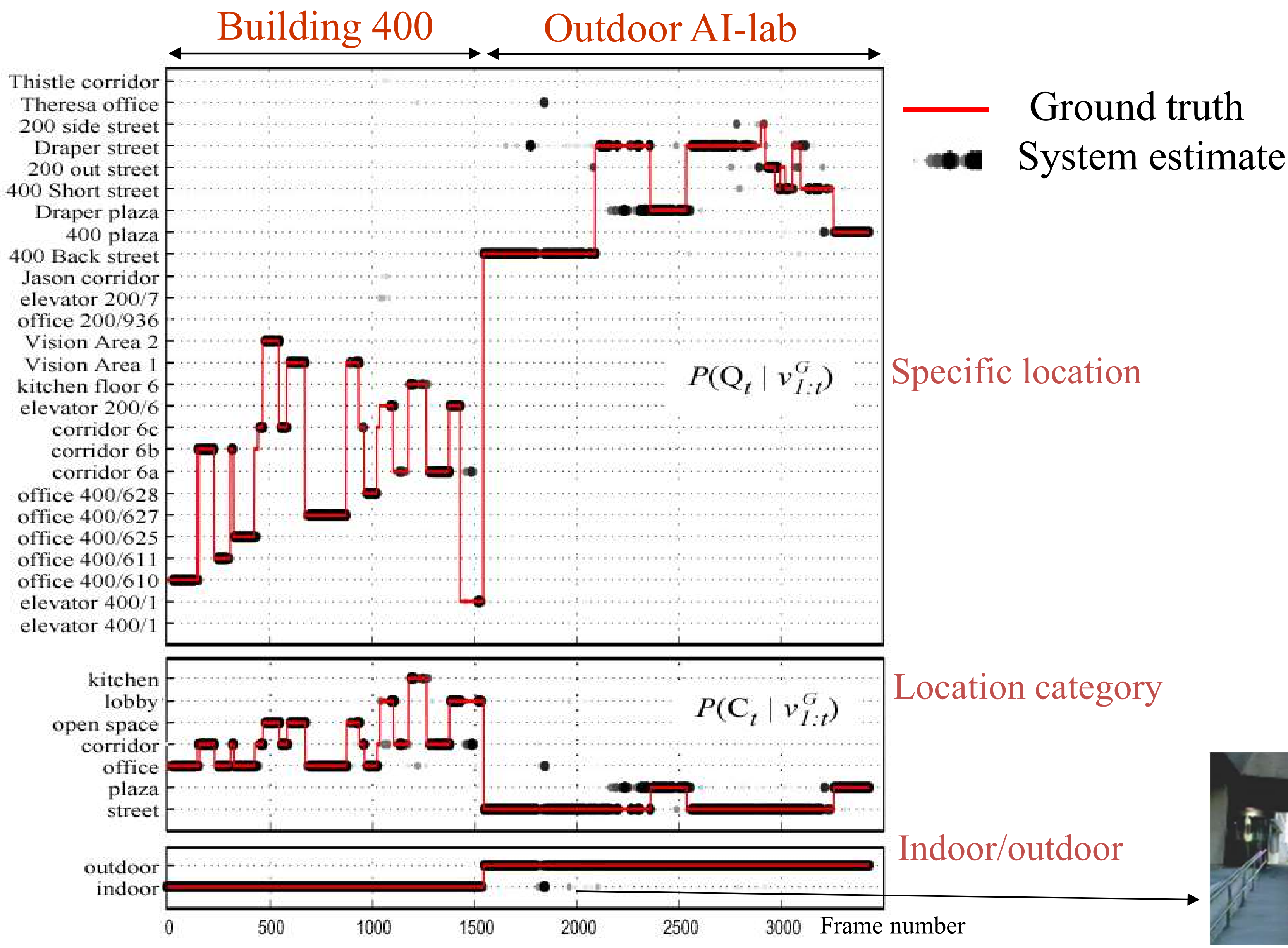
We use annotated sequences for training



- 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)

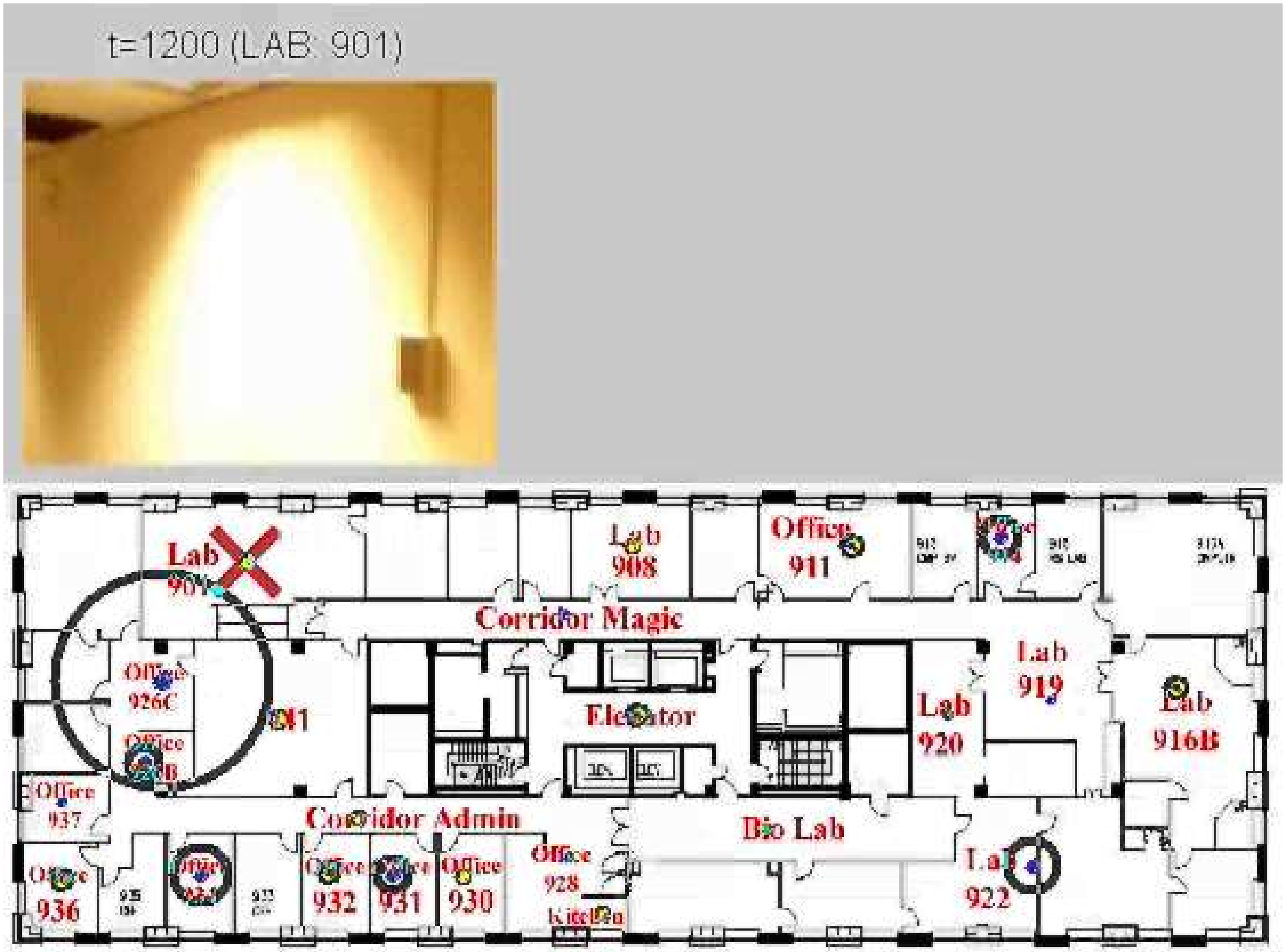
Oliva & Torralba, IJCV’01; Torralba, Murphy, Freeman, Mark, CVPR 03.

Place and scene recognition using gist



Place recognition demo

$p(q_t \mid v_t)$ $P(q_t \mid v_{1:t})$

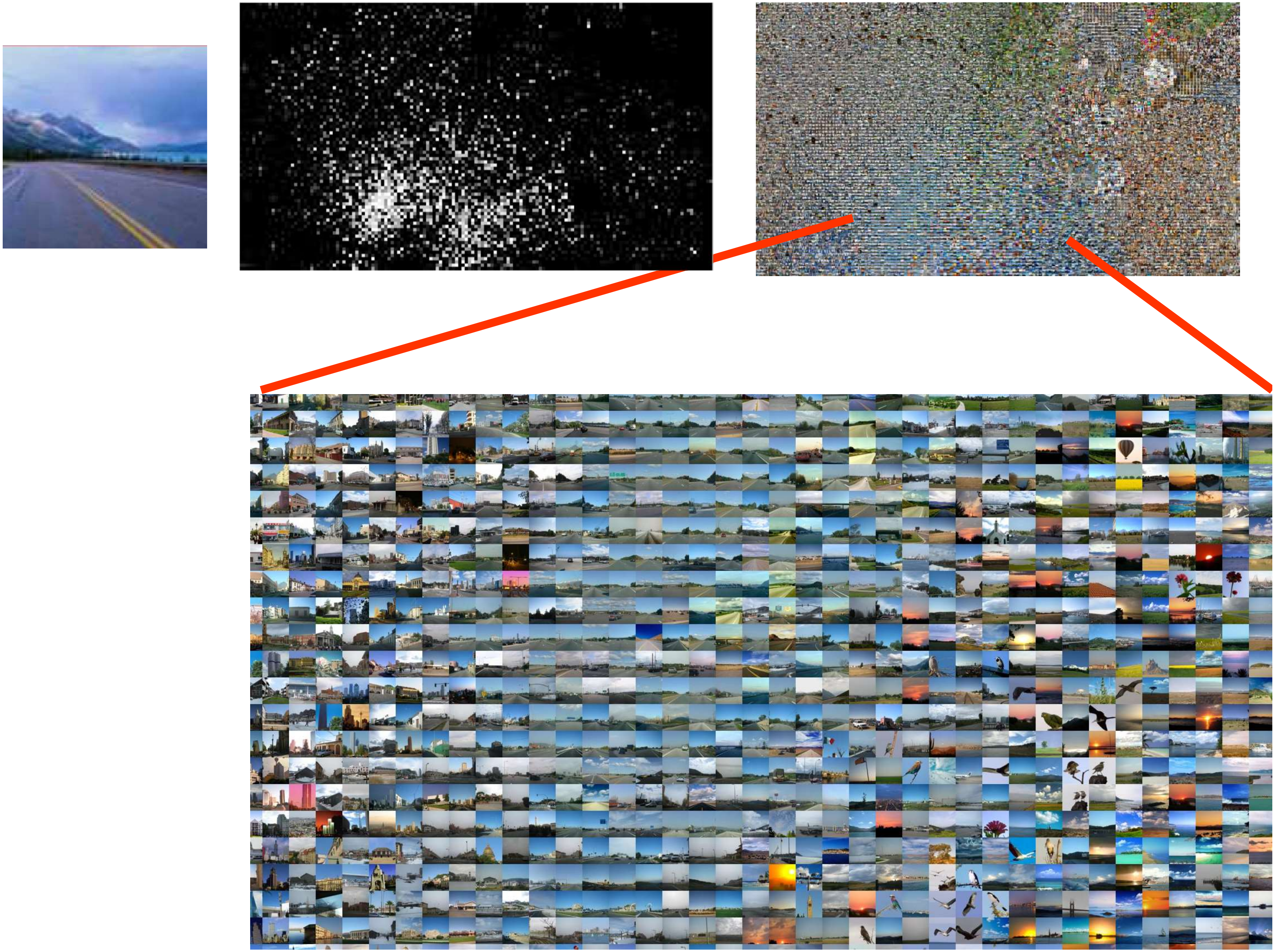


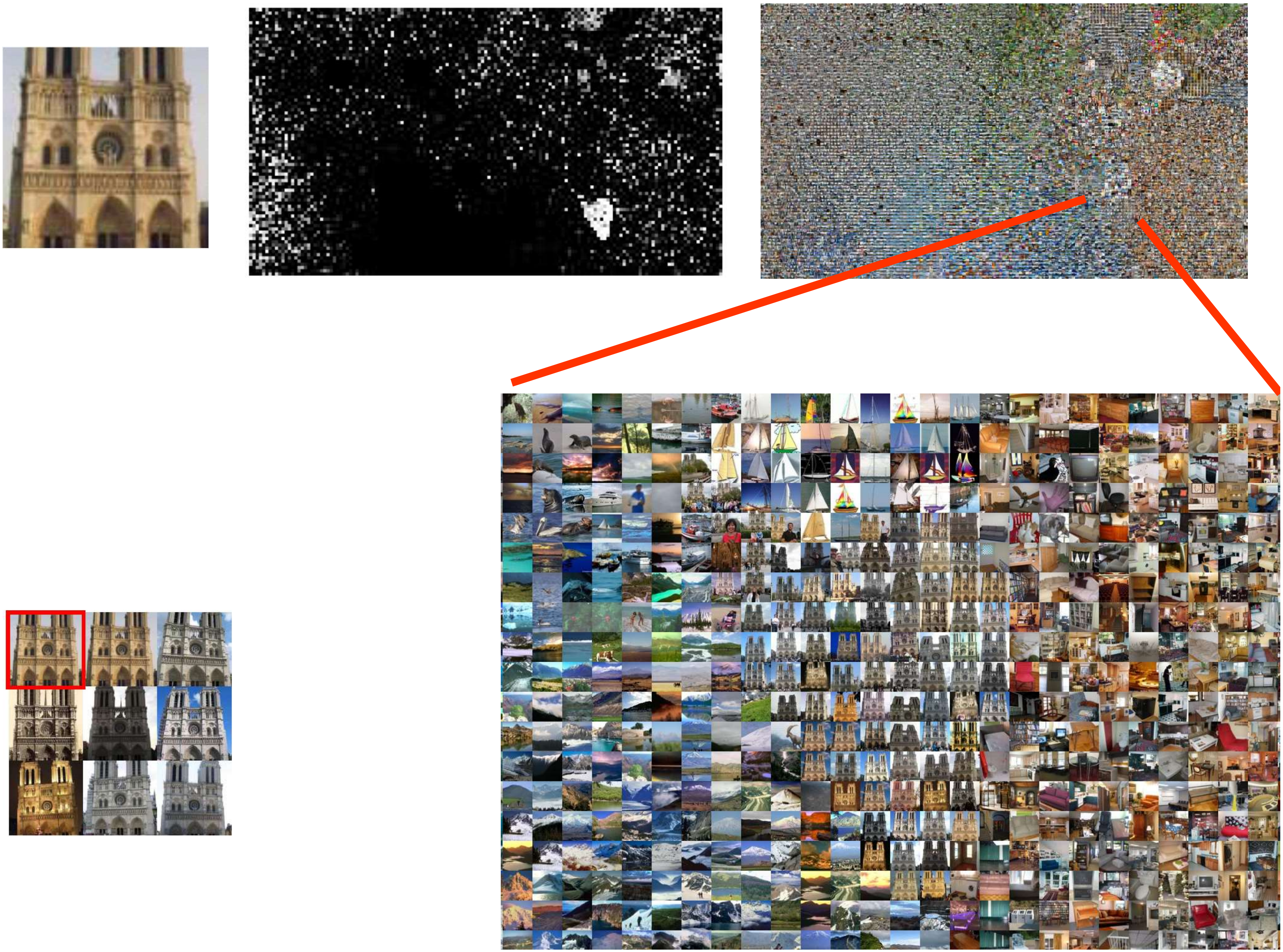
Categories or a continuous space?

From the city to the mountains in 10 steps

Mosaic using 12,000 images

Interactive version at: http://people.csail.mit.edu/torralba/research/LabelMe/labelmeMap/





im2gps

Instead of using objects labels, the web provides other kinds of metadata associate to large collections of images

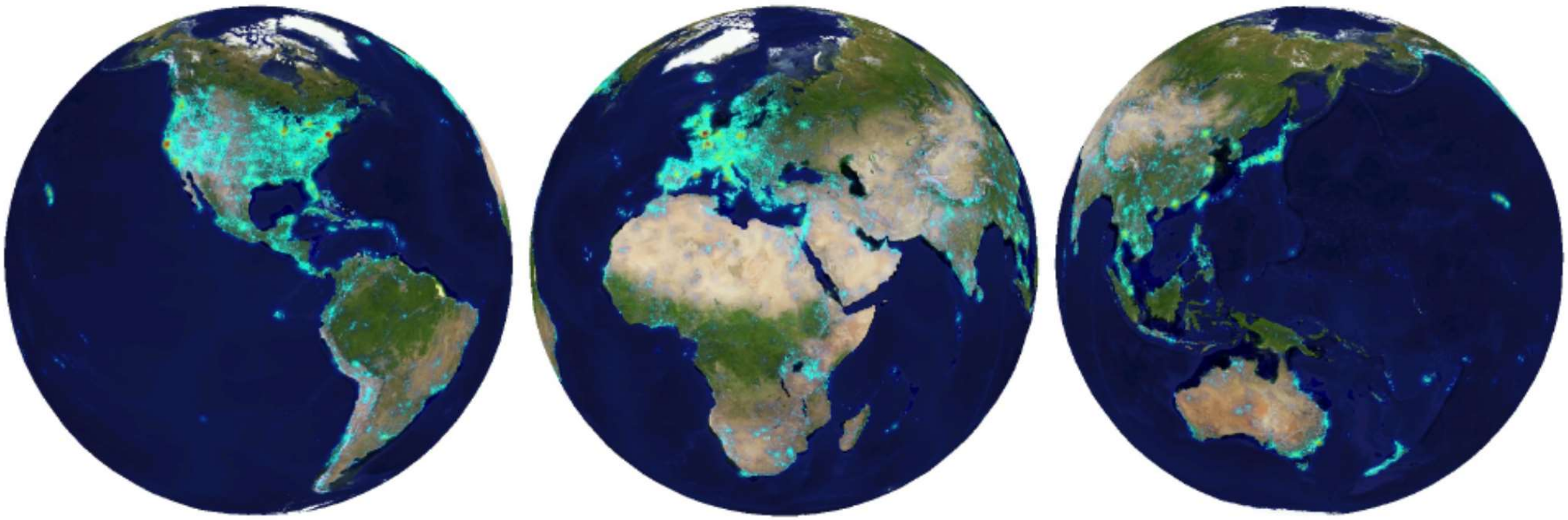


Figure 2. The distribution of photos in our database. Photo locations are cyan. Density is overlaid with the jet colormap (log scale).

20 million geotagged and geographic text-labeled images

Hays & Efros. CVPR 2008

Hays & Efros. CVPR 2008

im2gps

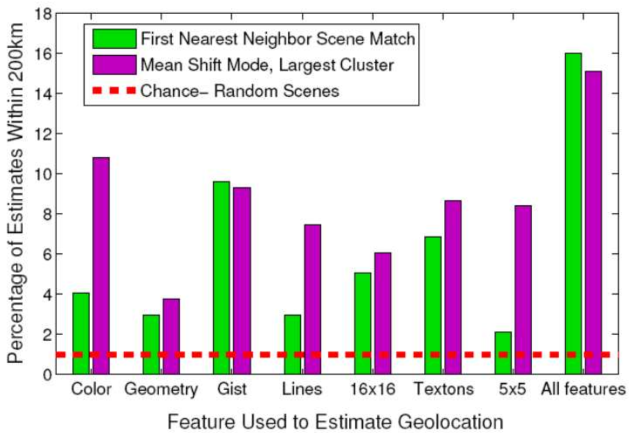


Figure 5. *Geolocation performance across features.* Percentage of test cases geolocated to within $200km$ for each feature. We compare geolocation by 1-NN vs. largest mean-shift mode.

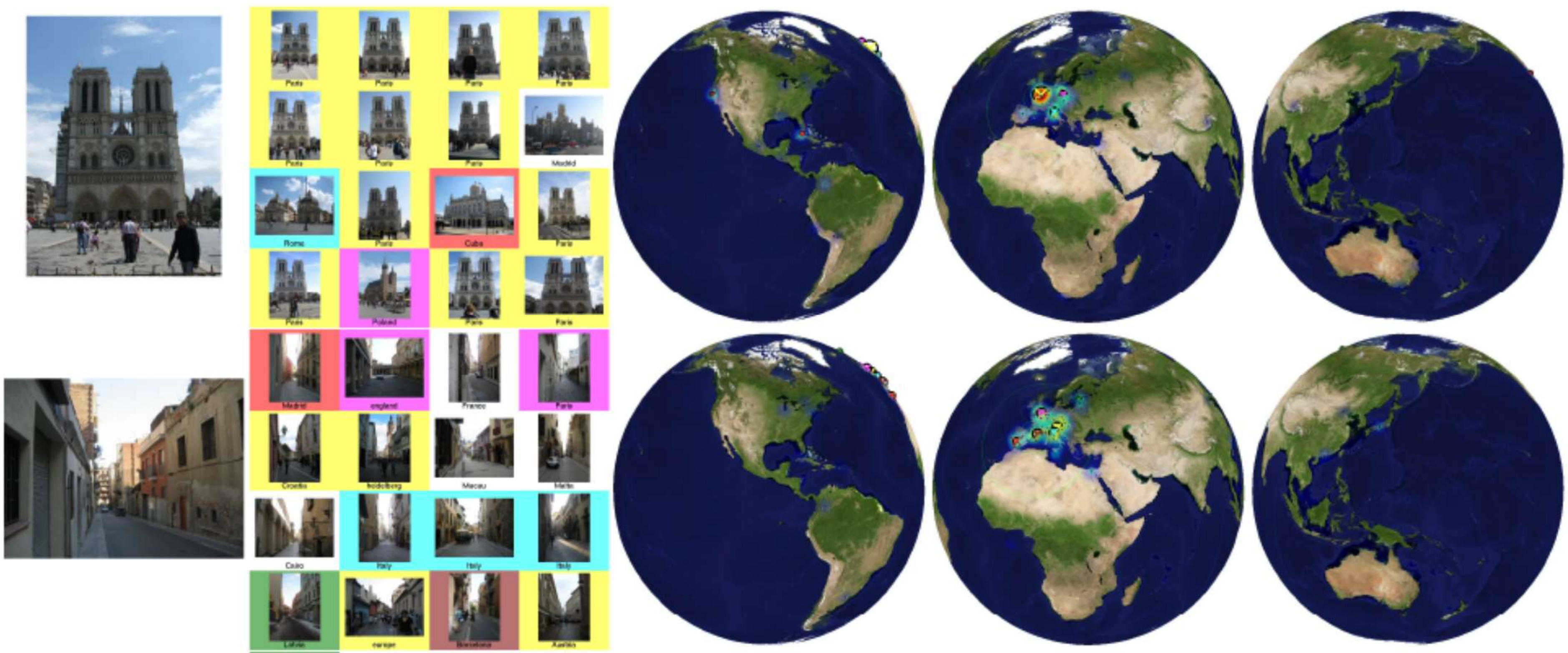
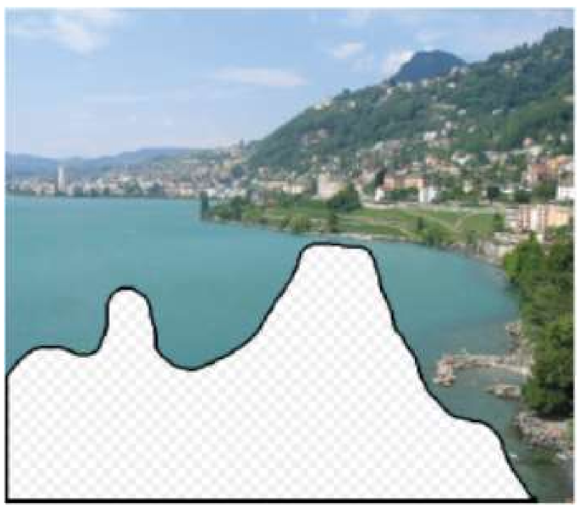


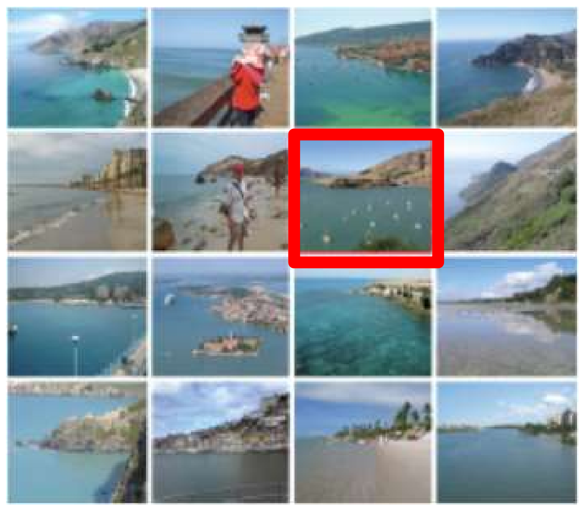
Image completion

Original Image | Input | Criminisi et al. | MS *Smart Erase*

Instead, generate proposals using millions of images



Input



16 nearest neighbors
(gist+color matching)

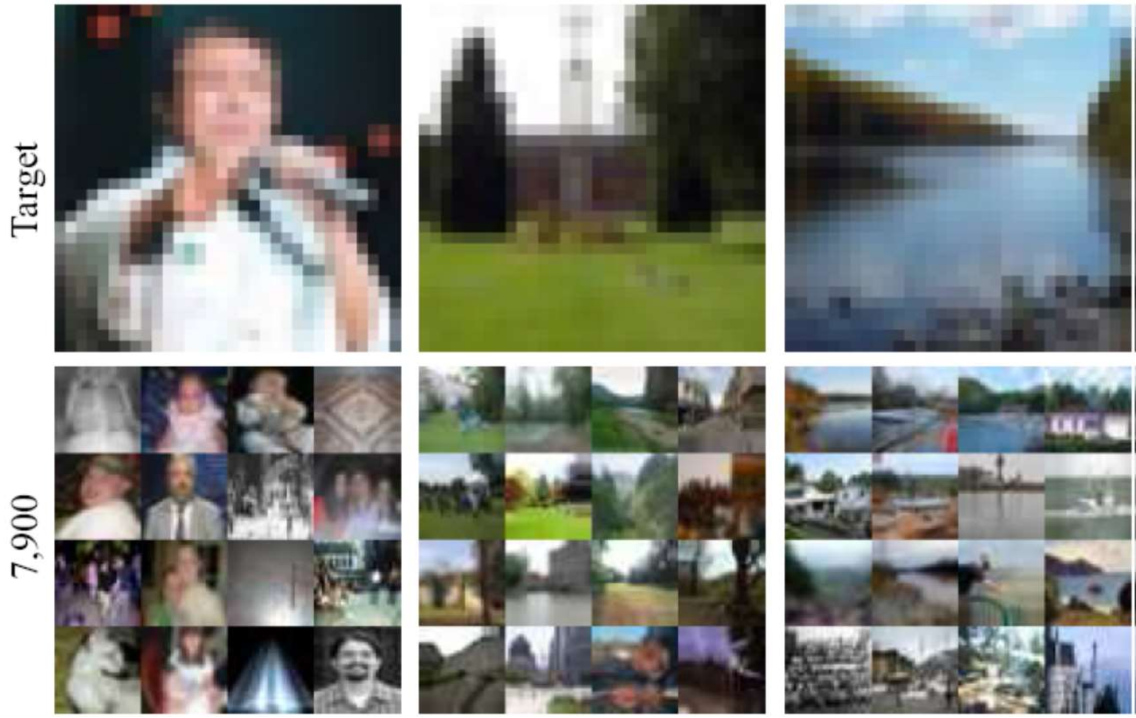
output

Hays, Efros, 2007

Lots

Of

Images

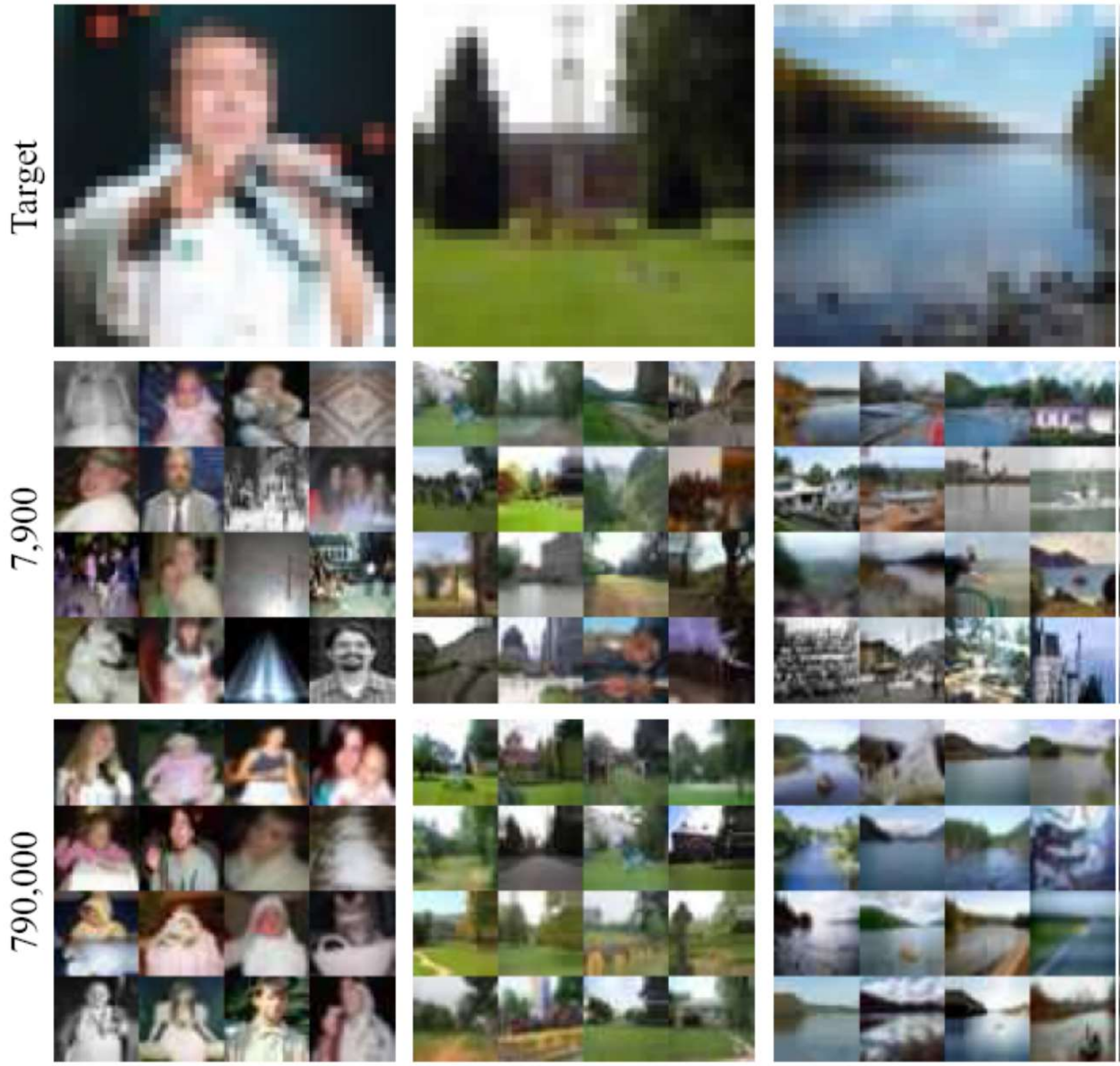


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

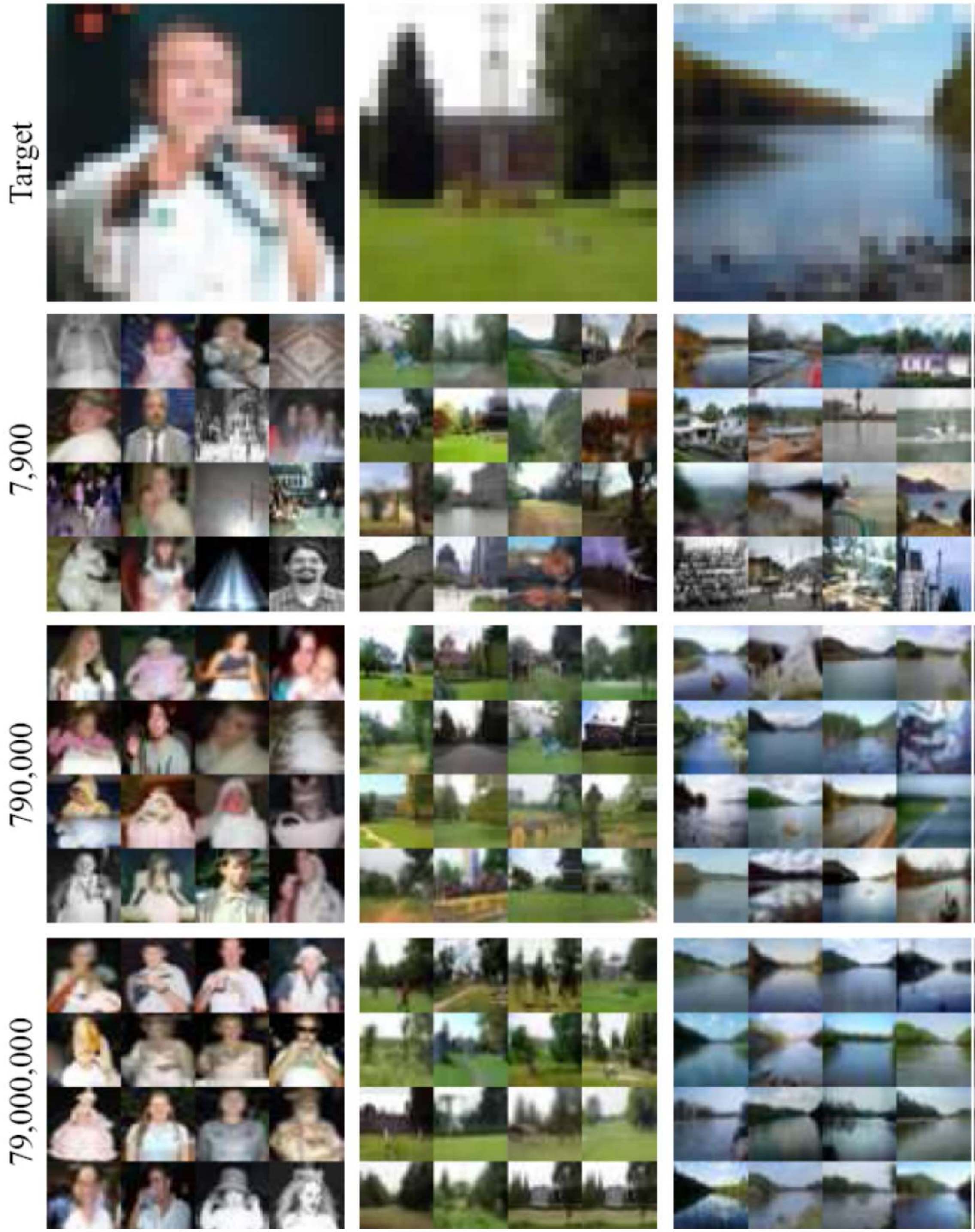


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images



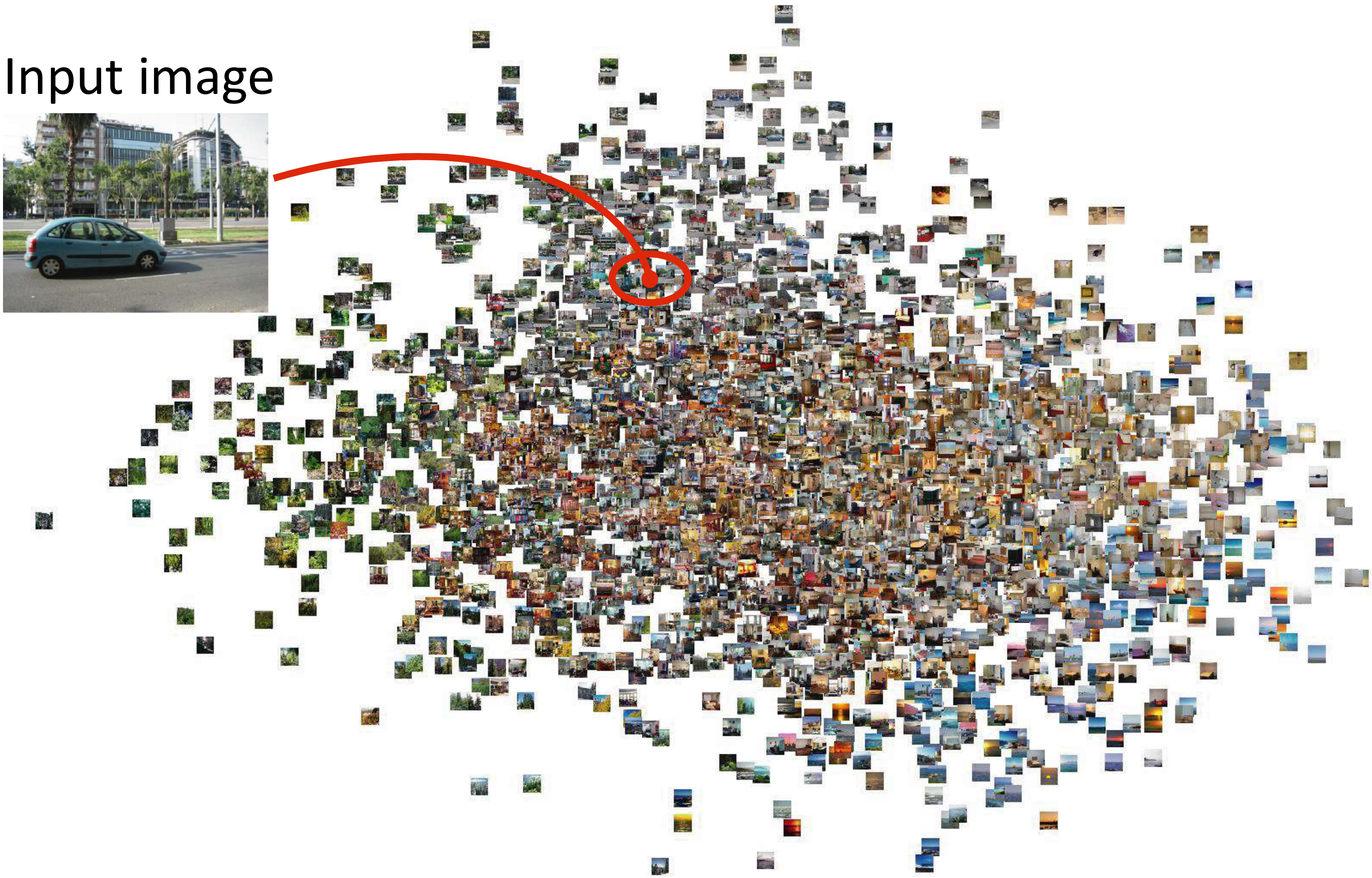
Automatic Colorization Result

Grayscale input High resolution

Colorization of input using average

A. Torralba, R. Fergus, W.T.Freeman. 2008

Nearest neighbors classification




Target
Neighbors (SSD + warping)
Average

Predicting events

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Predicting events

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Query

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Query

Retrieved video

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Query

Retrieved video

Synthesized video

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Query

Retrieved video

Synthesized video

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Query

Synthesized video

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

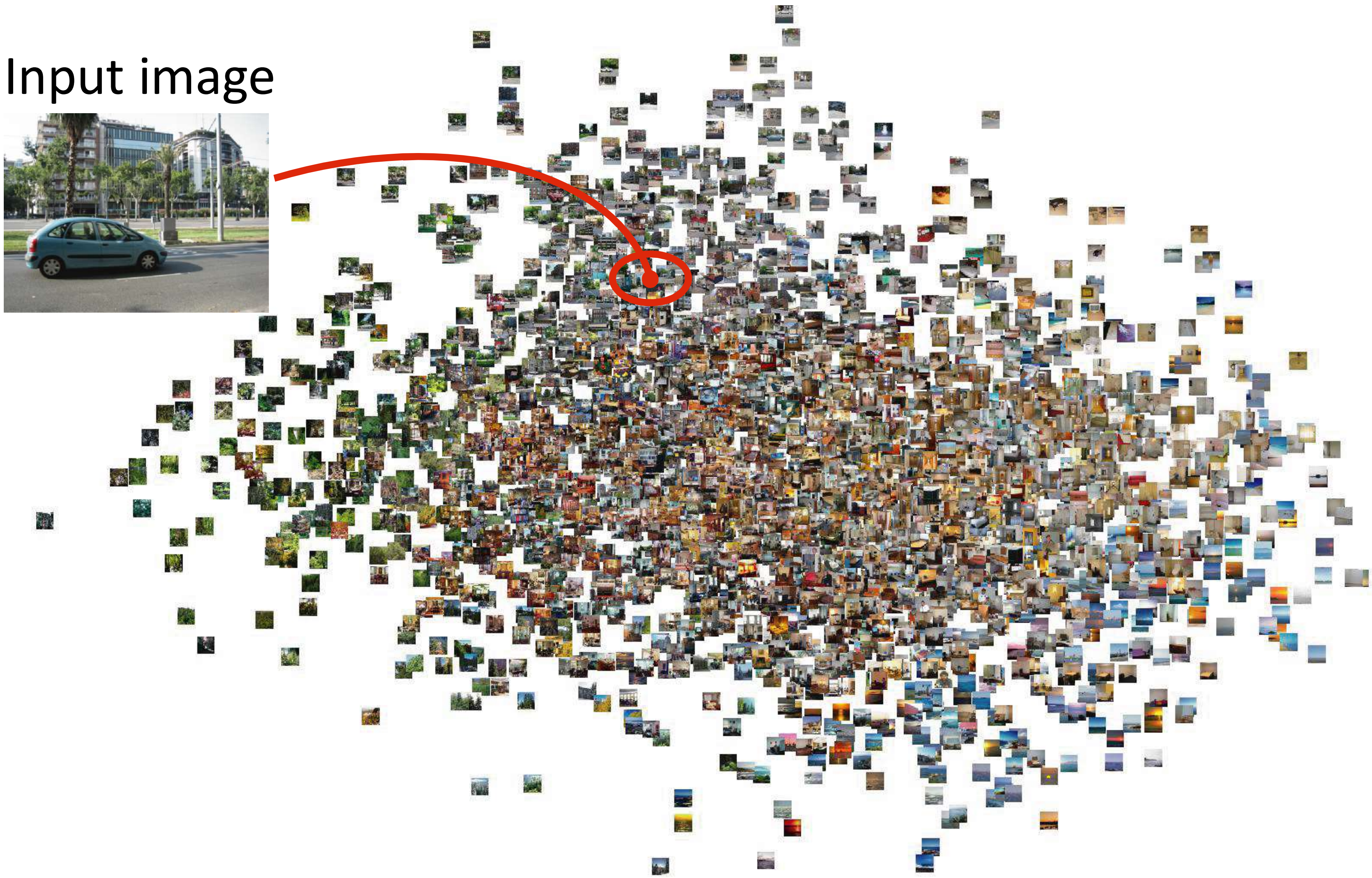
Query

Retrieved video

Synthesized video

C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, ECCV 2008

Dealing with millions of images



Powers of 10

Number of images on my hard drive: 10^4

Number of images seen during my first 10 years: 10^8

(3 images/second * 60 * 60 * 16 * 365 * 10 = 630720000)

Number of images seen by all humanity: 10^{20}

106,456,367,669 humans[1] * 60 years * 3 images/second * 60 * 60 * 16 * 365 =

1 from http://www.prb.org/Articles/2002/HowManyPeopleHaveEverLivedonEarth.aspx

Number of all images in the universe: 10^{243}

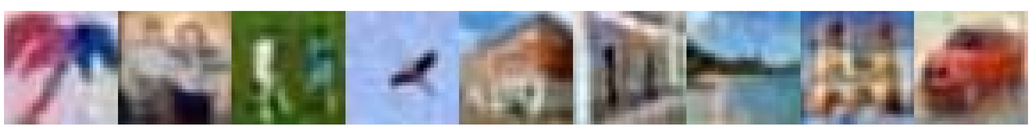
10^{81} atoms * 10^{81} * 10^{81} =

Number of all 32x32 images: 10^{7373}

$256^{32*32*3} \sim 10^{7373}$

Binary codes for global scene representation

- Short codes allow for storing millions of images
- Efficient search: hamming distance (search millions of images in few microseconds)
- Internet scale experiments: compute nearest neighbors between all images in the internet



512 bits

Binary codes for images

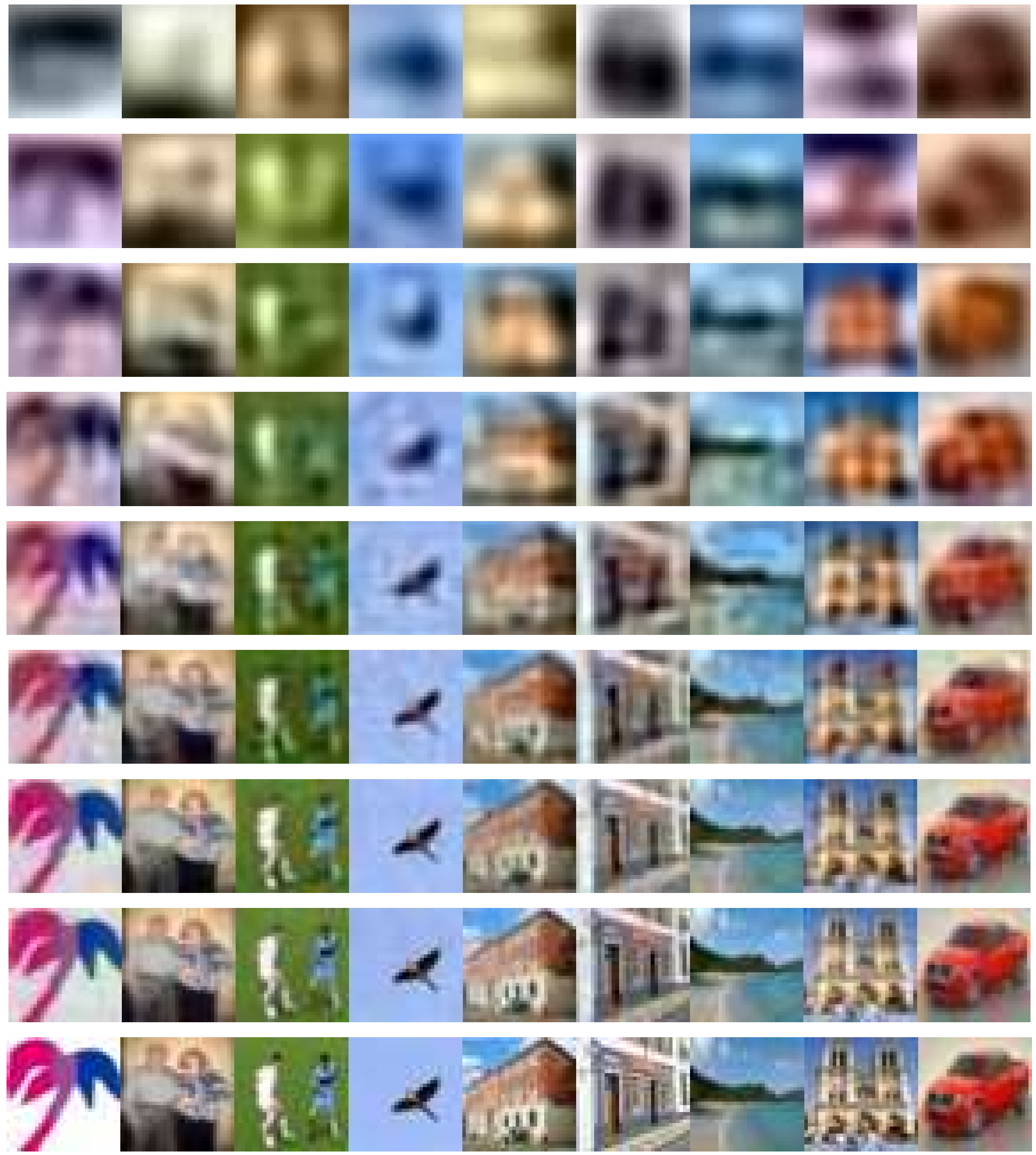
- Want images with similar content to have similar binary codes
- Use Hamming distance between codes
  - Number of bit flips
  - E.g.:

```
Ham_Dist(10001010,10001110)=1
Ham_Dist(10001010,11101110)=3
```

- Semantic Hashing [Salakhutdinov & Hinton, 2007]
  - Text documents

Slide Rob Fergus

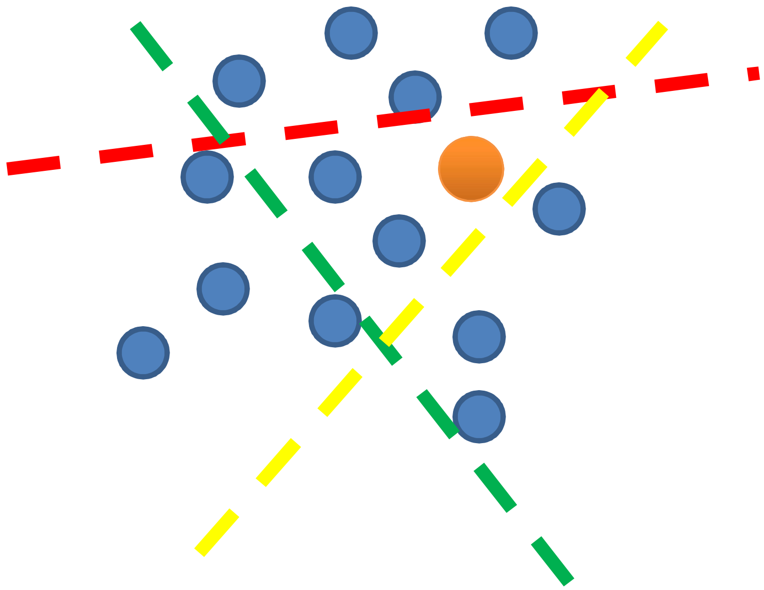
How many bits do we need?



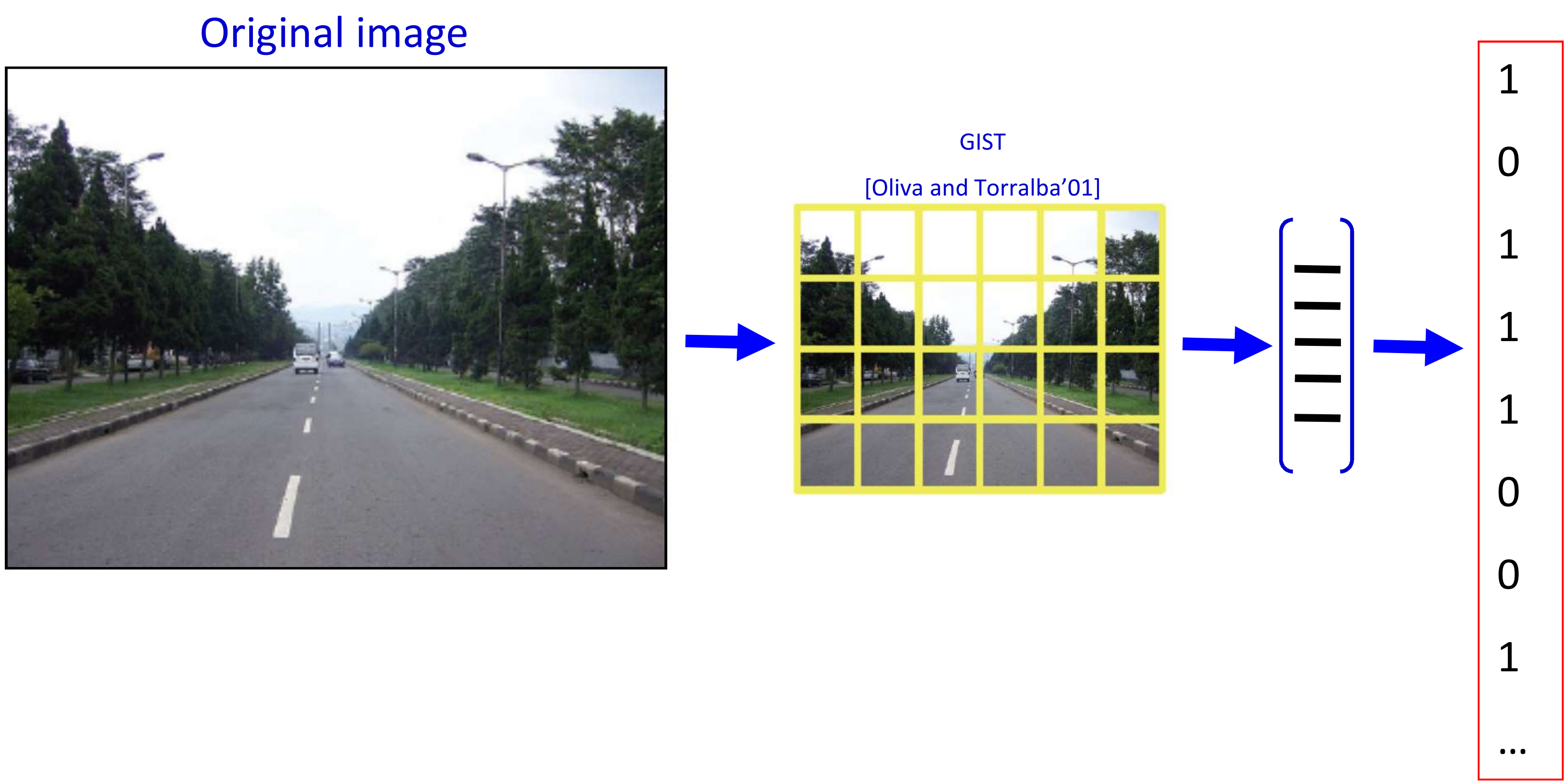
16 bits
32 bits
64 bits
128 bits
256 bits
512 bits
1024 bits
2048 bits
24576 bits

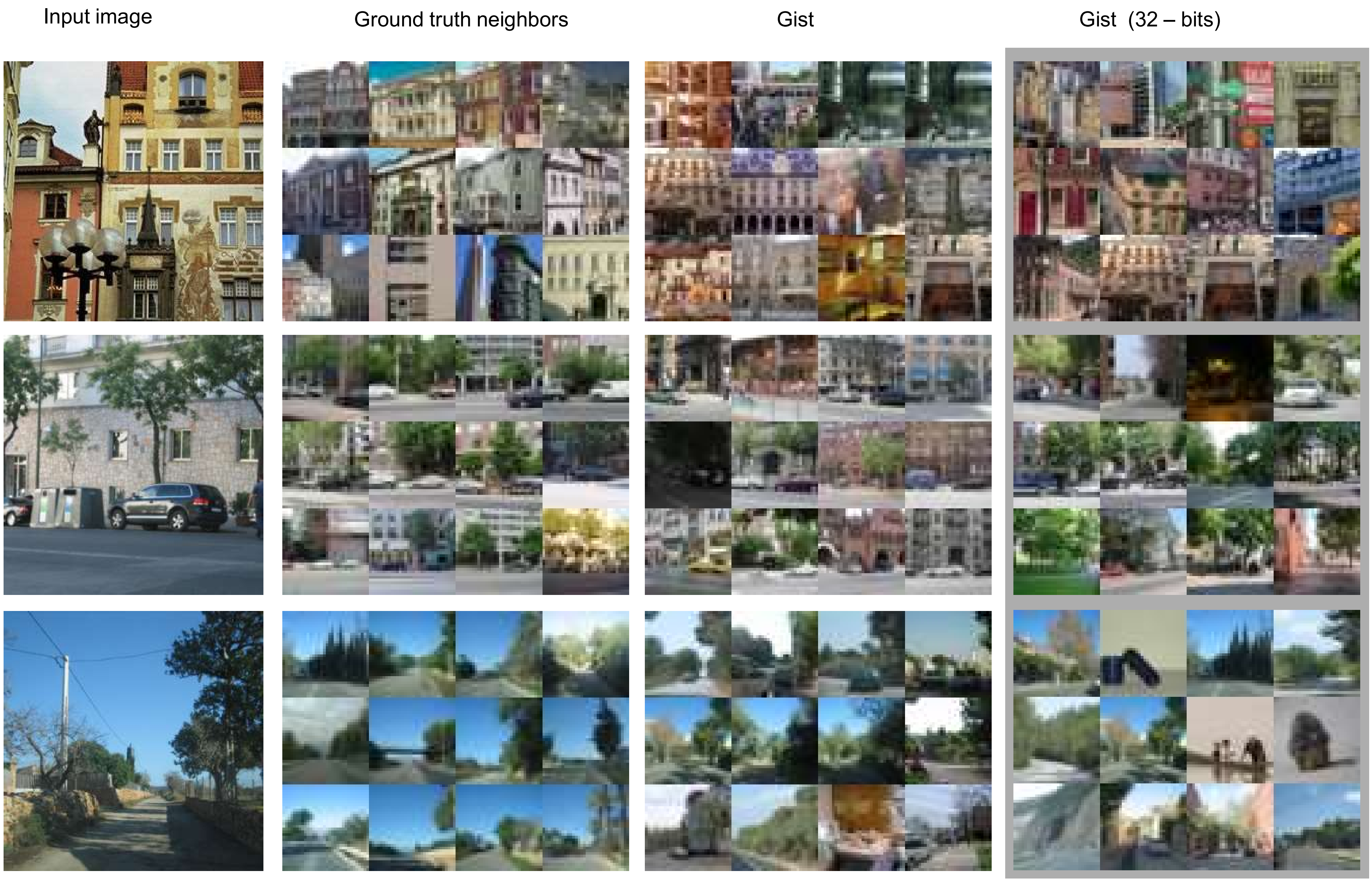
Locality Sensitive Hashing

- Gionis, A. & Indyk, P. & Motwani, R. (1999)
- Take random projections of data
- Quantize each projection with few bits



Compressing the gist descriptor




Input image
Ground truth neighbors
Gist
Gist (32 – bits)

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

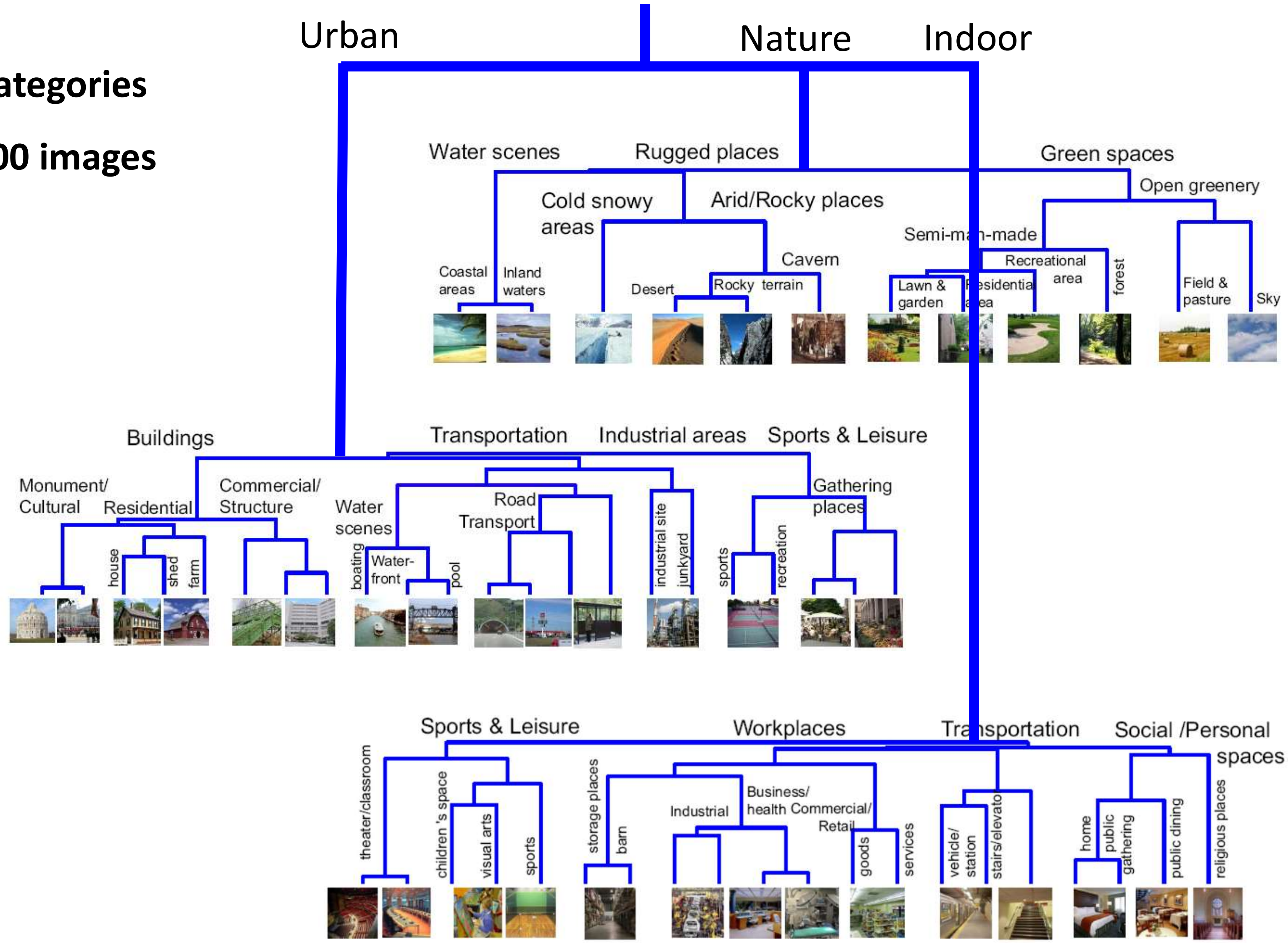
Kitchen

Store

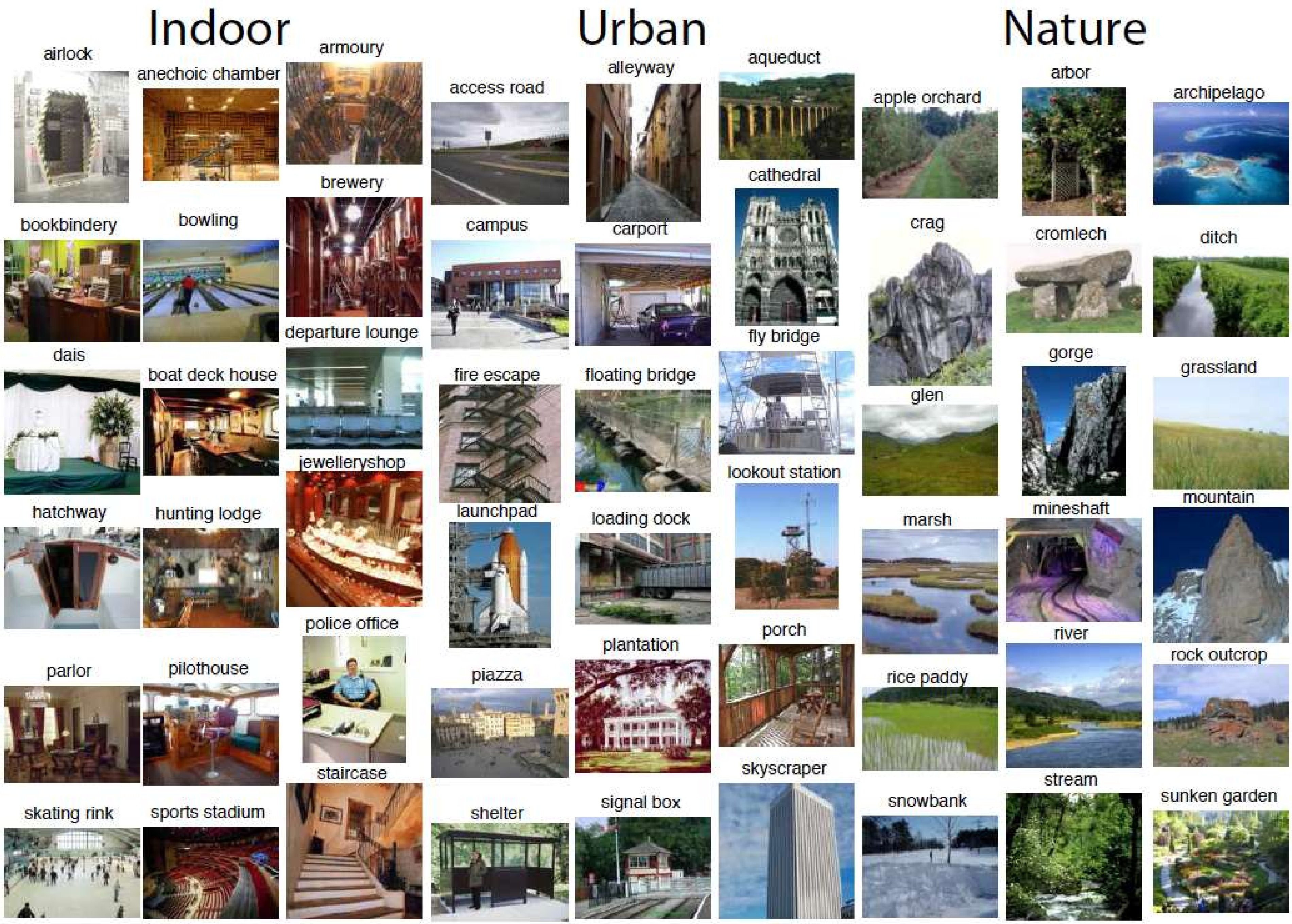
Large Scale Scene Recognition

> 400 categories

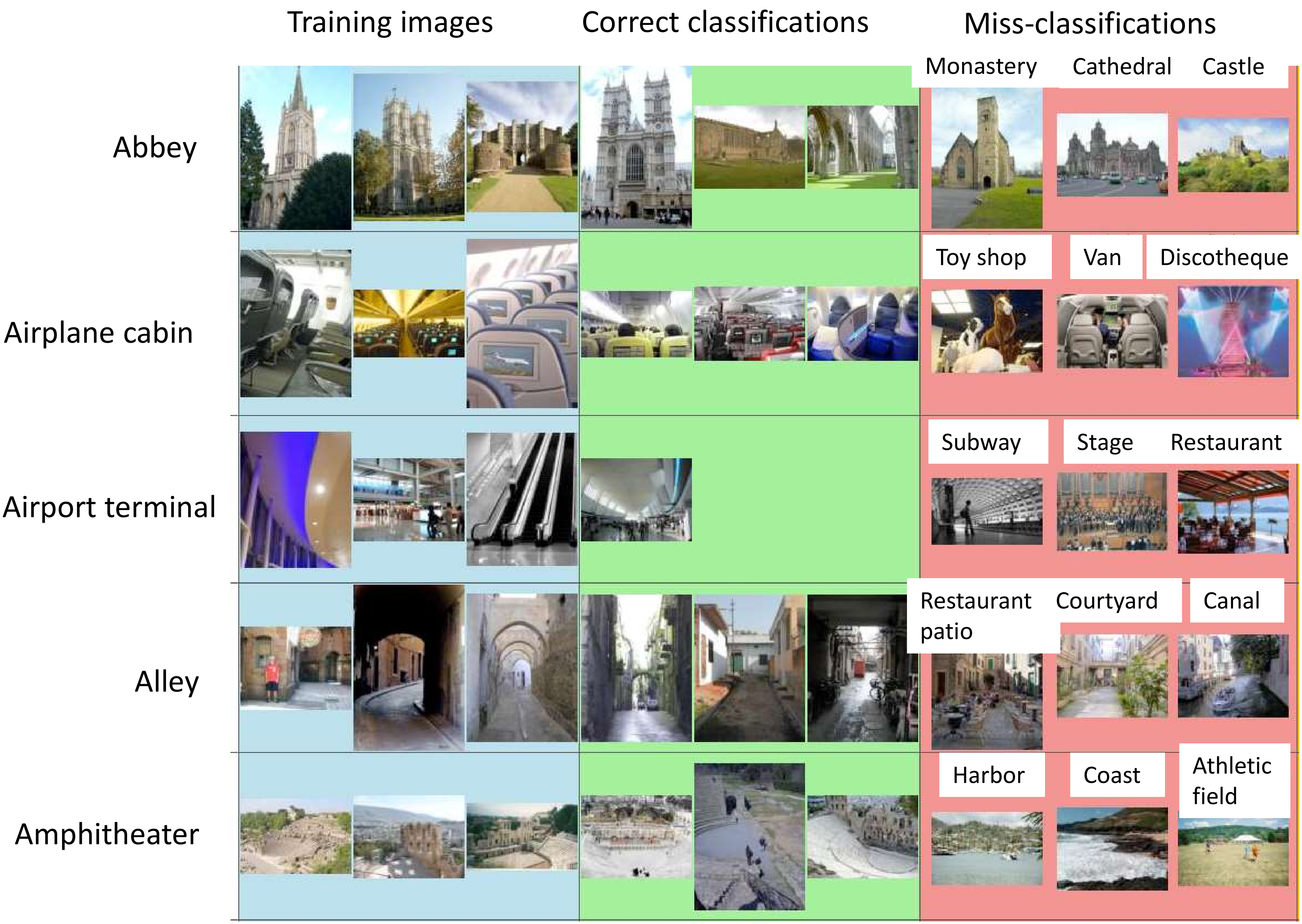
>140,000 images



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



Indoor
Urban
Nature
airlock
bookbindery
dais
hatchway
parlor
skating rink
anechoic chamber
bowling
boat deck house
hunting lodge
pilothouse
sports stadium
armoury
brewery
departure lounge
jewelleryshop
police office
staircase
access road
campus
fire escape
launchpad
piazza
shelter
alleyway
carport
floating bridge
loading dock
plantation
signal box
aqueduct
cathedral
fly bridge
lookout station
porch
skyscraper
apple orchard
crag
glen
marsh
rice paddy
snowbank
arbor
cromlech
gorge
mineshaft
river
stream
archipelago
ditch
grassland
mountain
rock outcrop
sunken garden



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