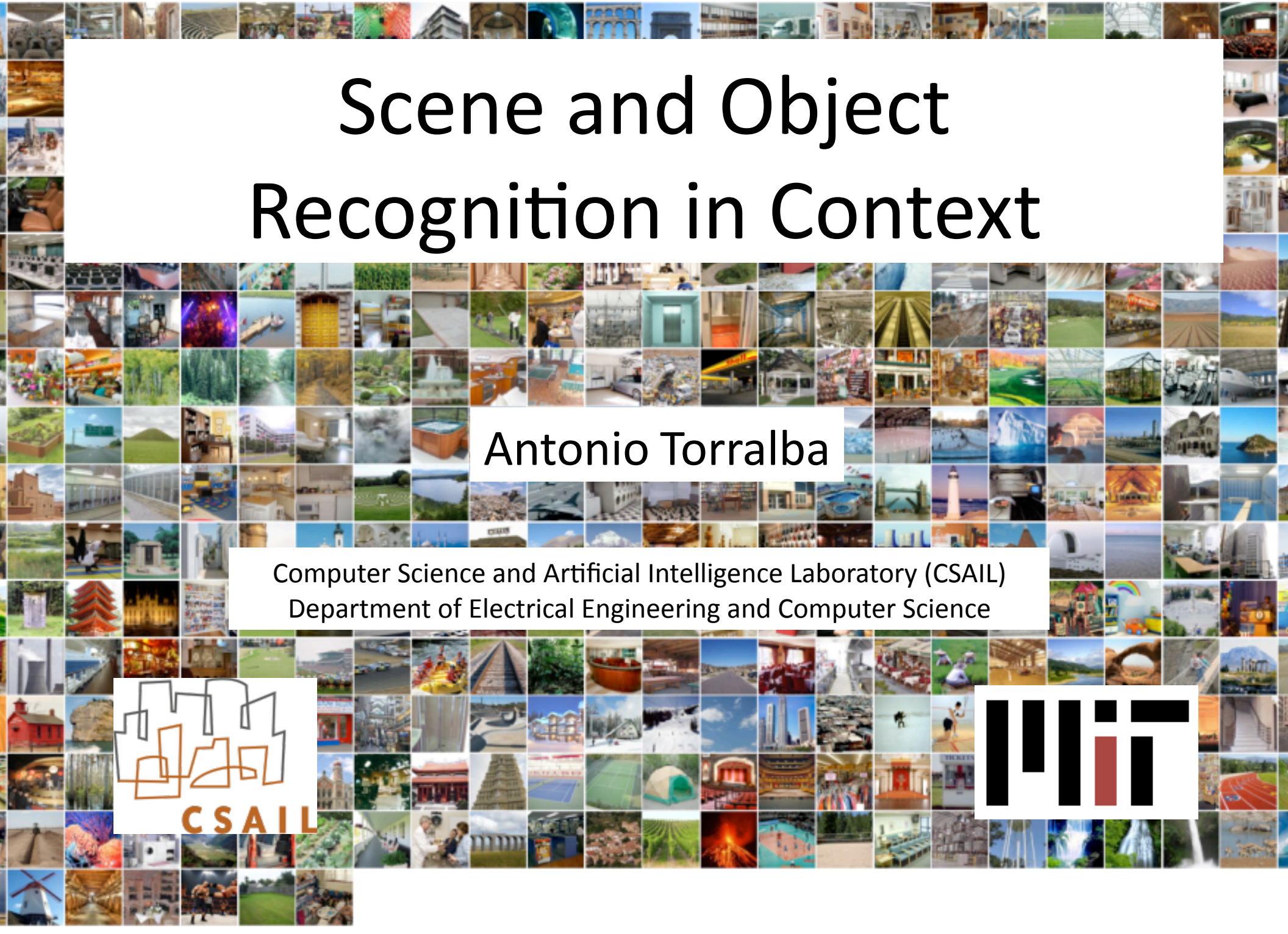
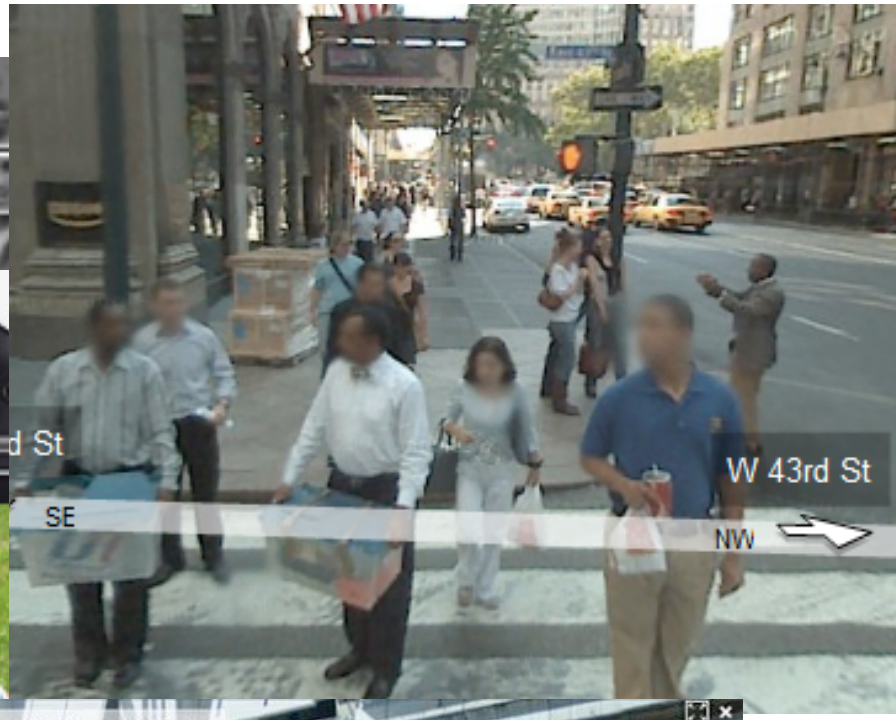


Scene and Object Recognition in Context

Antonio Torralba

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Department of Electrical Engineering and Computer Science

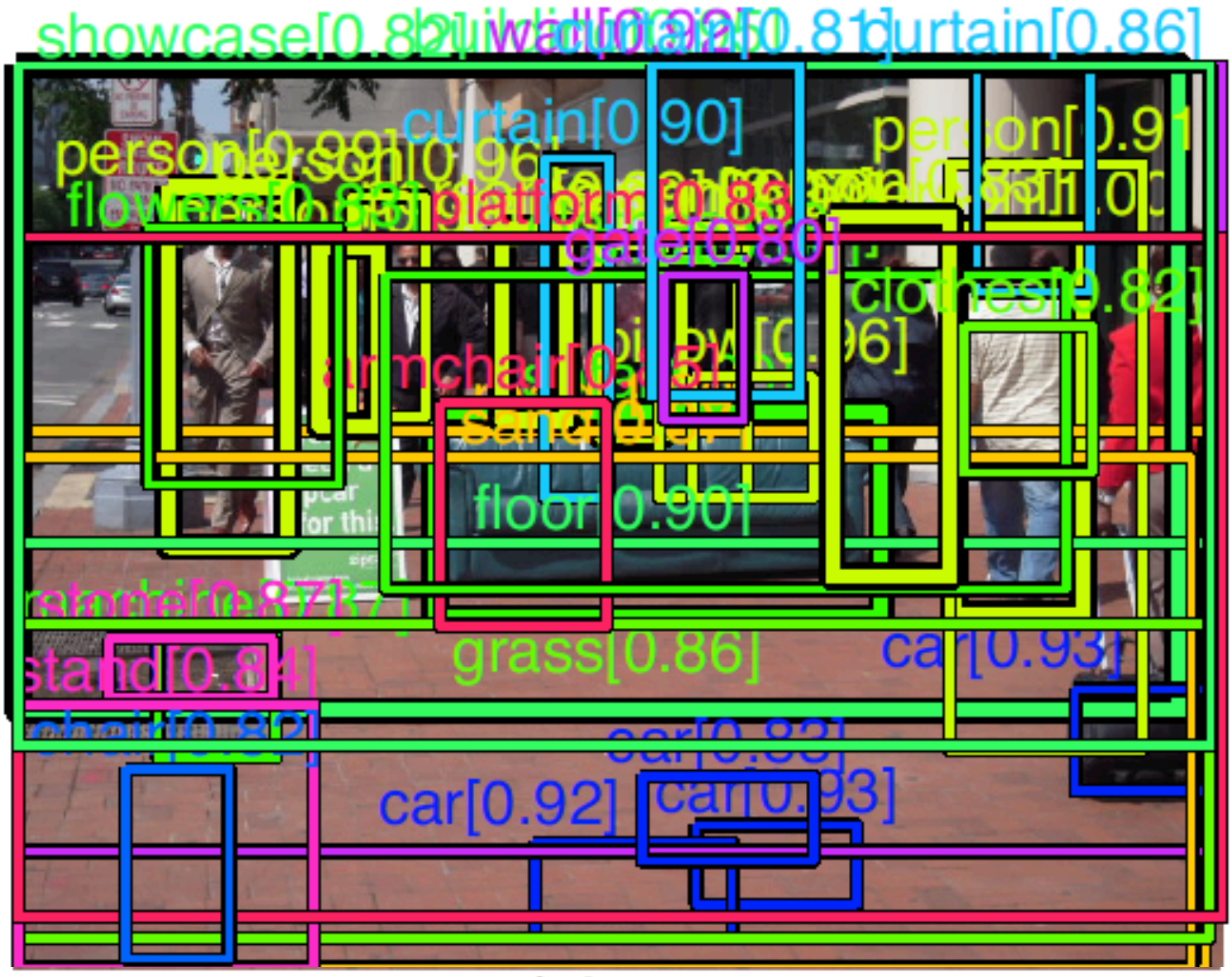






- The representation and matching of pictorial structures Fischler, Elschlager (1973).
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….


SNOW
ROUTE
you
need a
zipcar
for this.
zipcar


showcase[0.82]
curtain[0.86]
curtain[0.90]
person[0.88]
person[0.96]
flower[0.85]
platform[0.83]
gate[0.80]
clothes[0.82]
armchair[0.81]
floor[0.90]
grass[0.86]
car[0.93]
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 detector trying to detect?

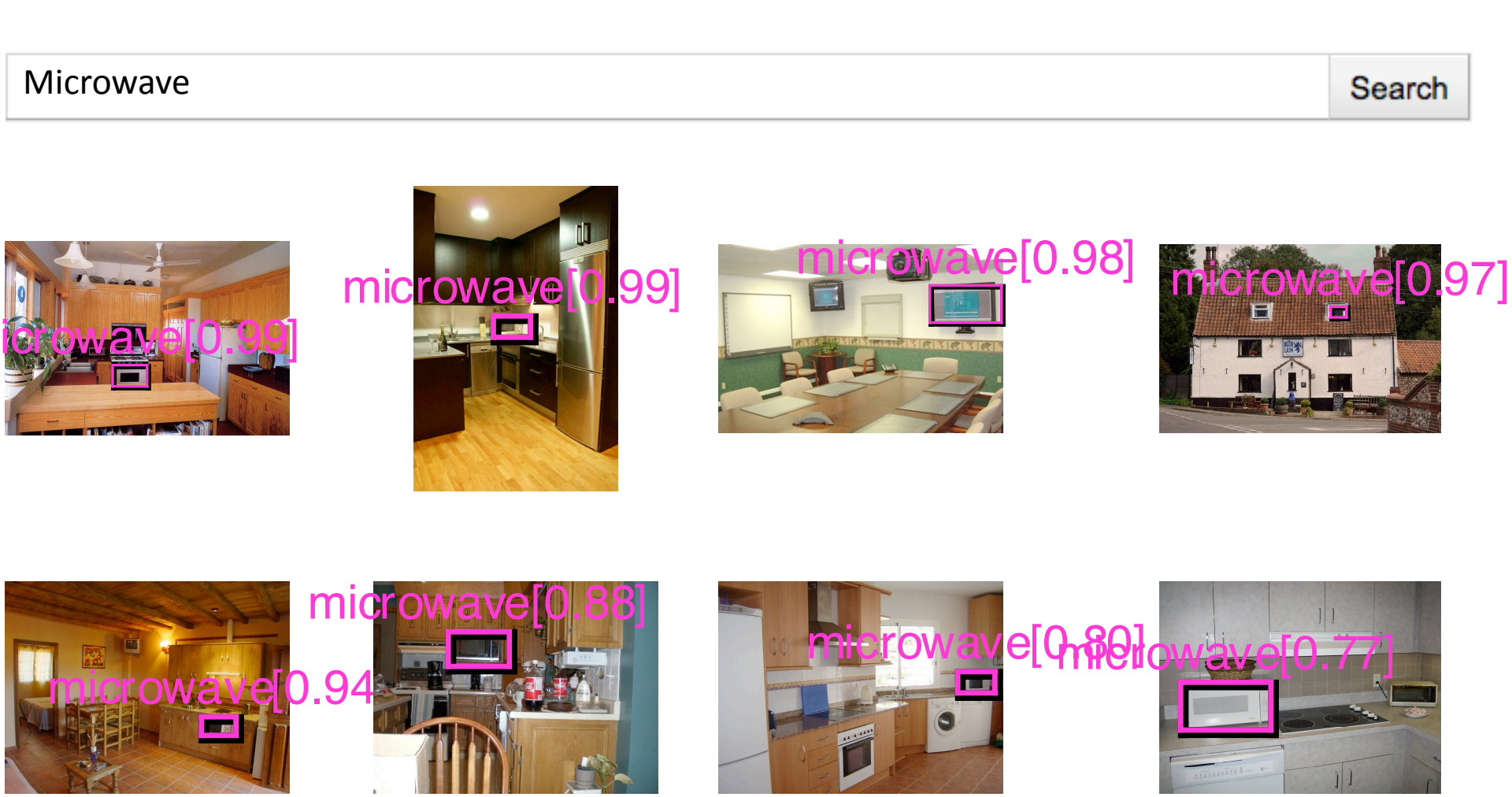


By looking at the output of a detector on a random set of images, can you guess which object is it trying to detect?

What object is 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?



Top 8 out of 4317 images

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

Microwave & refrigerator Search

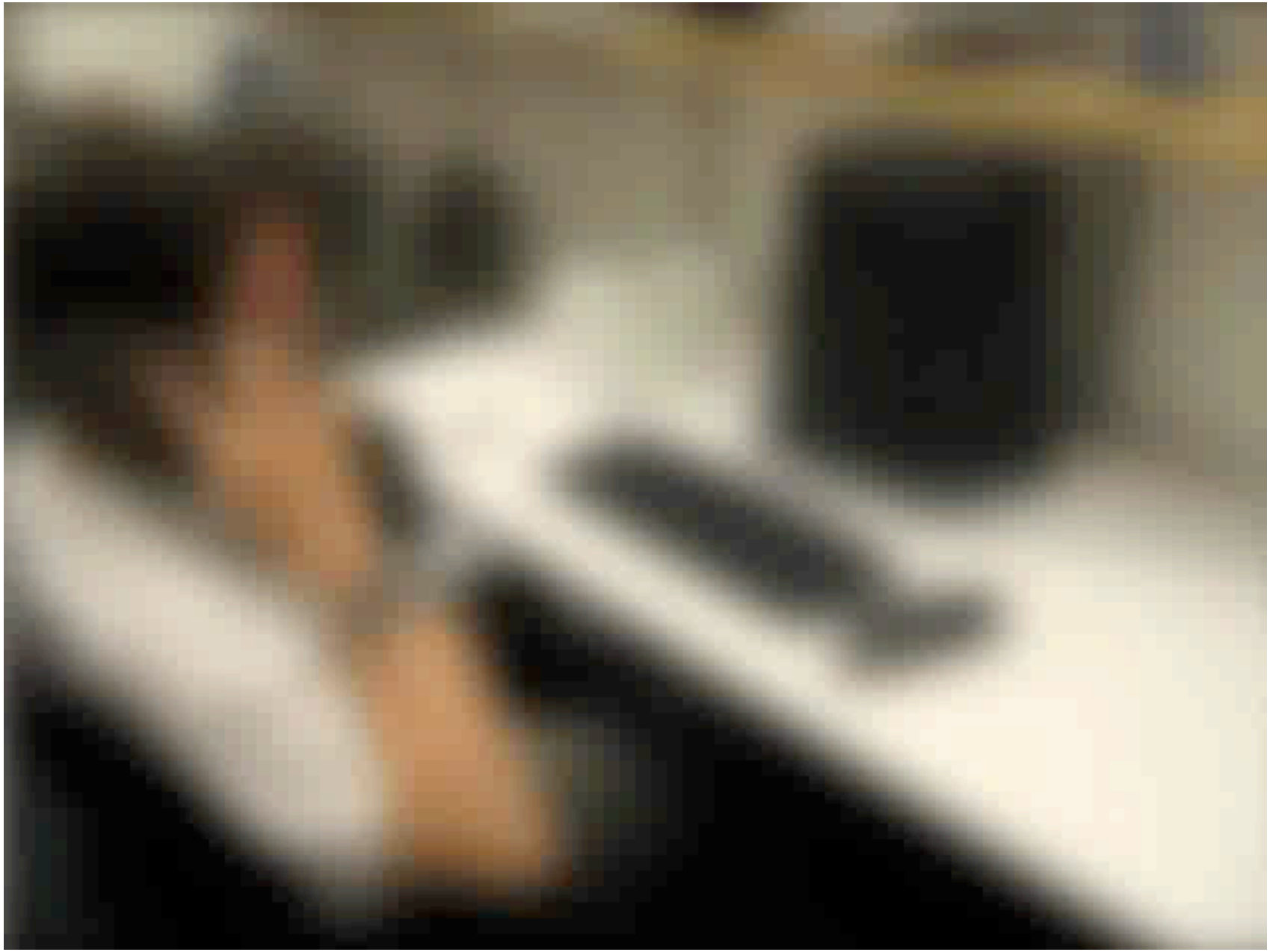









Top 8 out of 4317 images





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

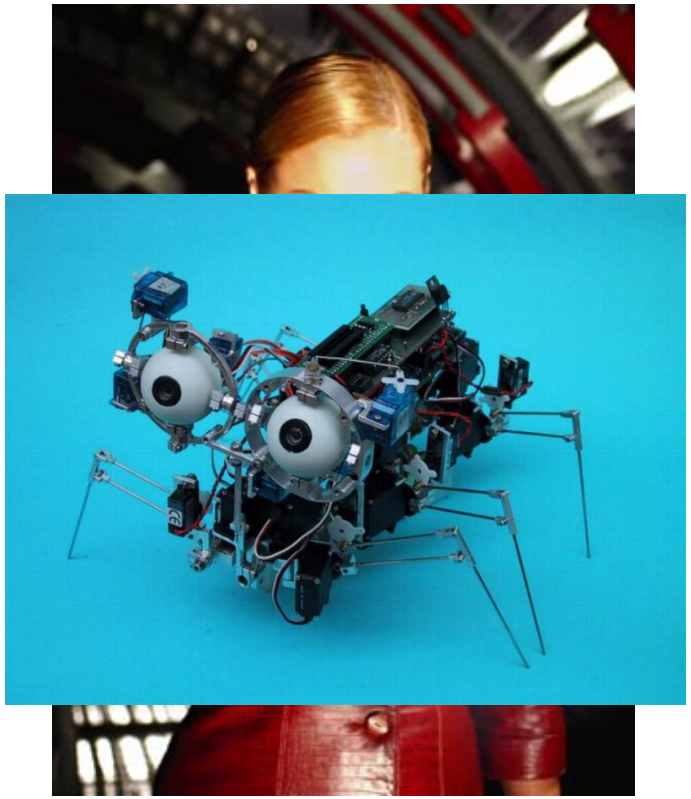
Data

Human vision

•Many input modalities

•Active

•Supervised, unsupervised, semi supervised learning. It can look for supervision.



Robot vision

•Many poor input modalities

•Active, but it does not go far

Internet vision

•Many input modalities

•It can reach everywhere

•Tons of data

Collecting datasets (towards 10^{6-7} examples)

- ESP game (CMU)

Luis Von Ahn and Laura Dabbish 2004



- LabelMe (MIT)

Russell, Torralba, Freeman, 2005

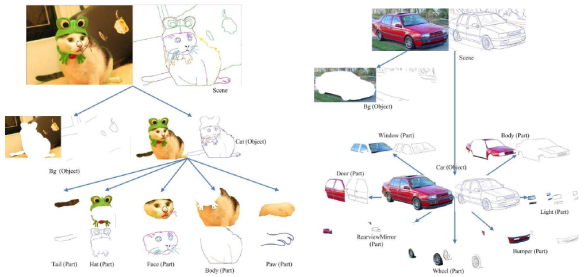


- StreetScenes (CBCL-MIT)

Bileschi, Poggio, 2006

- WhatWhere (Caltech)

Perona et al, 2007



- PASCAL challenge

2006, 2007

- Lotus Hill Institute

Song-Chun Zhu et al, 2007

- 80 million images

Torralba, Fergus, Freeman, 2007





B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

Extreme labeling



B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

The other extreme of extreme labeling



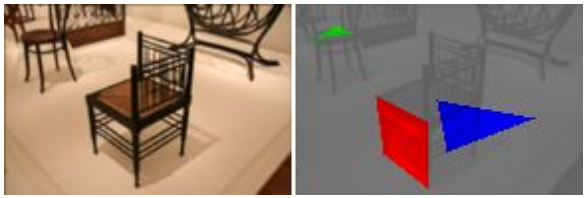
B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

Testing















Most common labels:

test

adksdsa

woiieiie

…

B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

Sophisticated testing





Most common labels:

Star

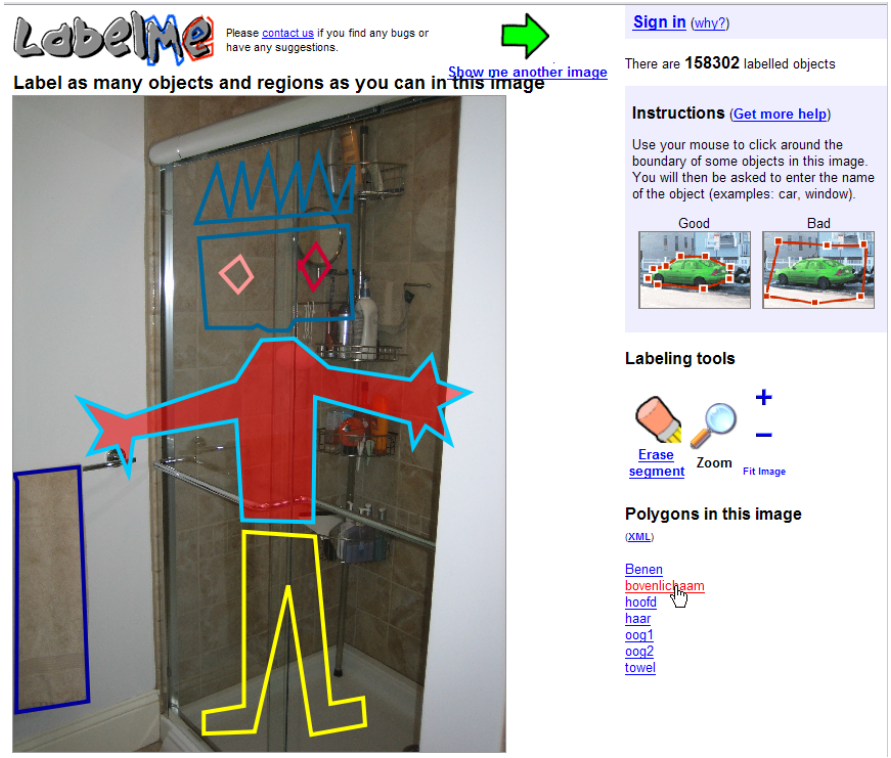
Square

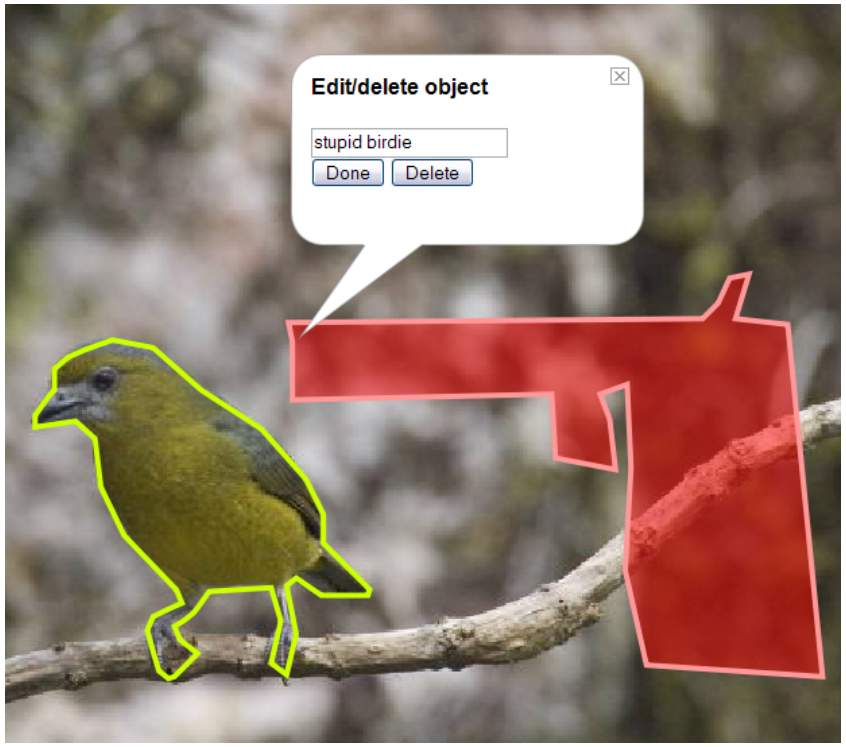
Nothing

…

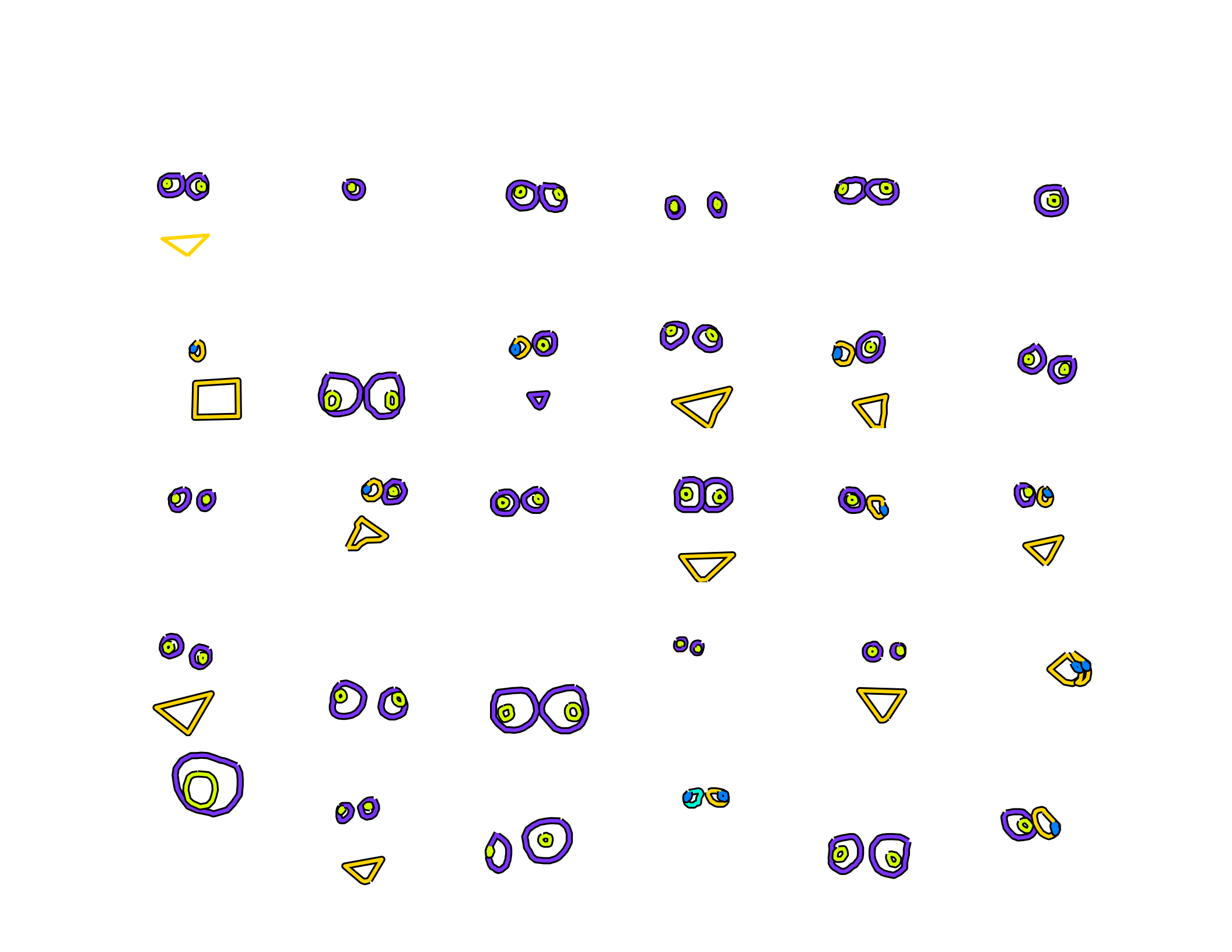
Creative testing

Do not try this at home

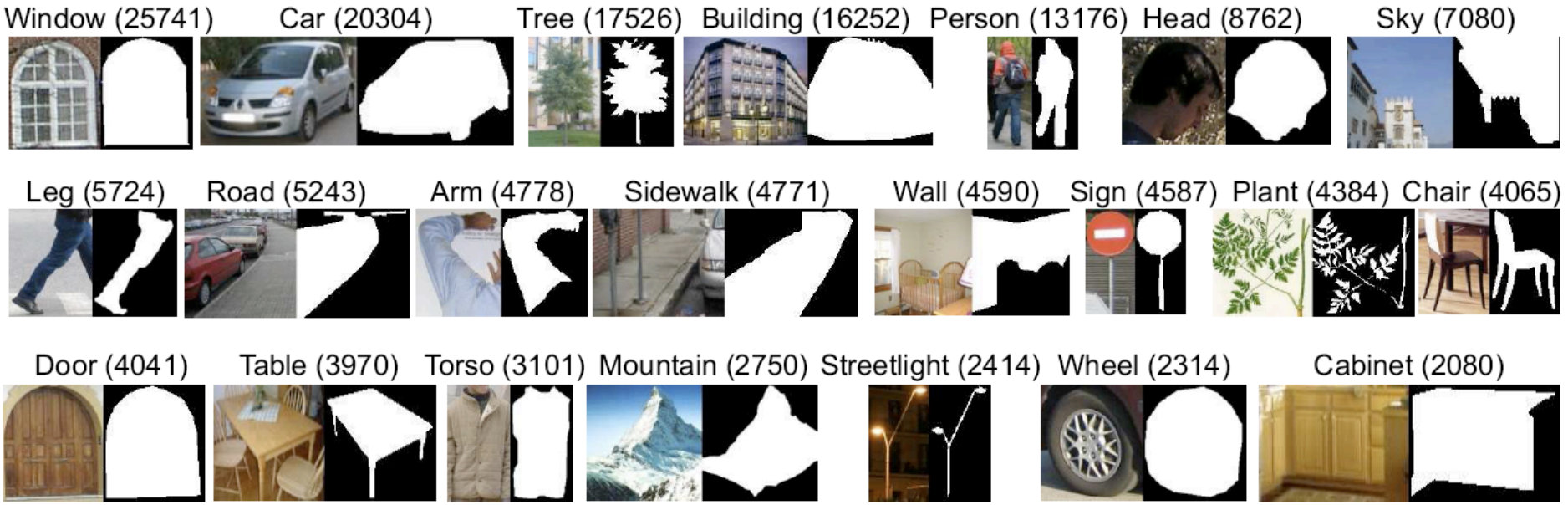




B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

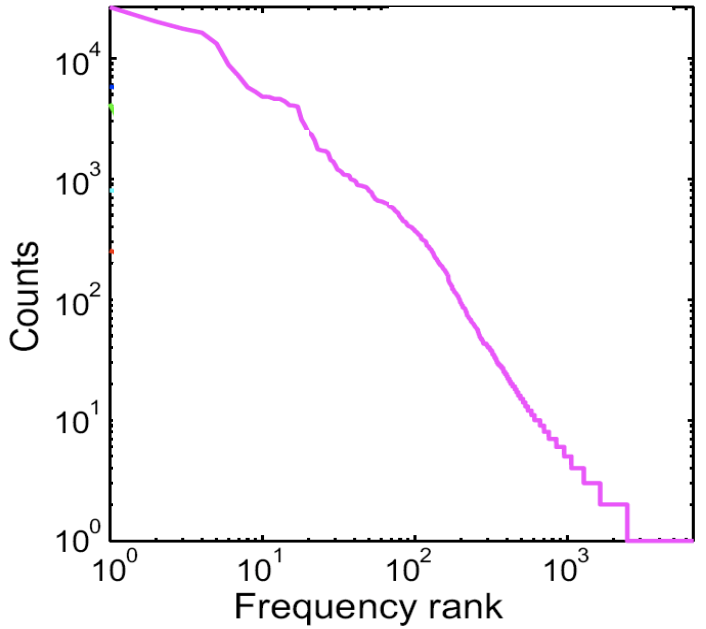


Object statistics



Stats:

- 530,000 polygons
- 8500 different object descriptions
- 265 descriptions with more than 100 instances



B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

LabelMe now in 3D!

Written by Bryan Russell and Antonio Torralba

1 Upload your picture



2 Label the objects



3 Build your own pop-up book!



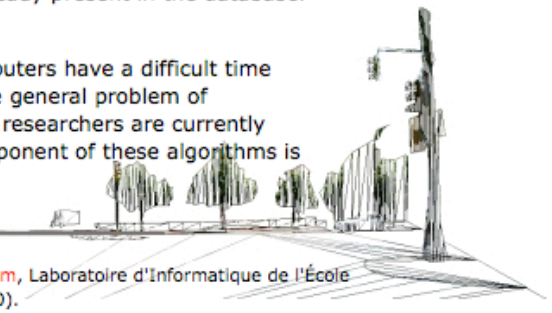
Annotation tool

Upload image

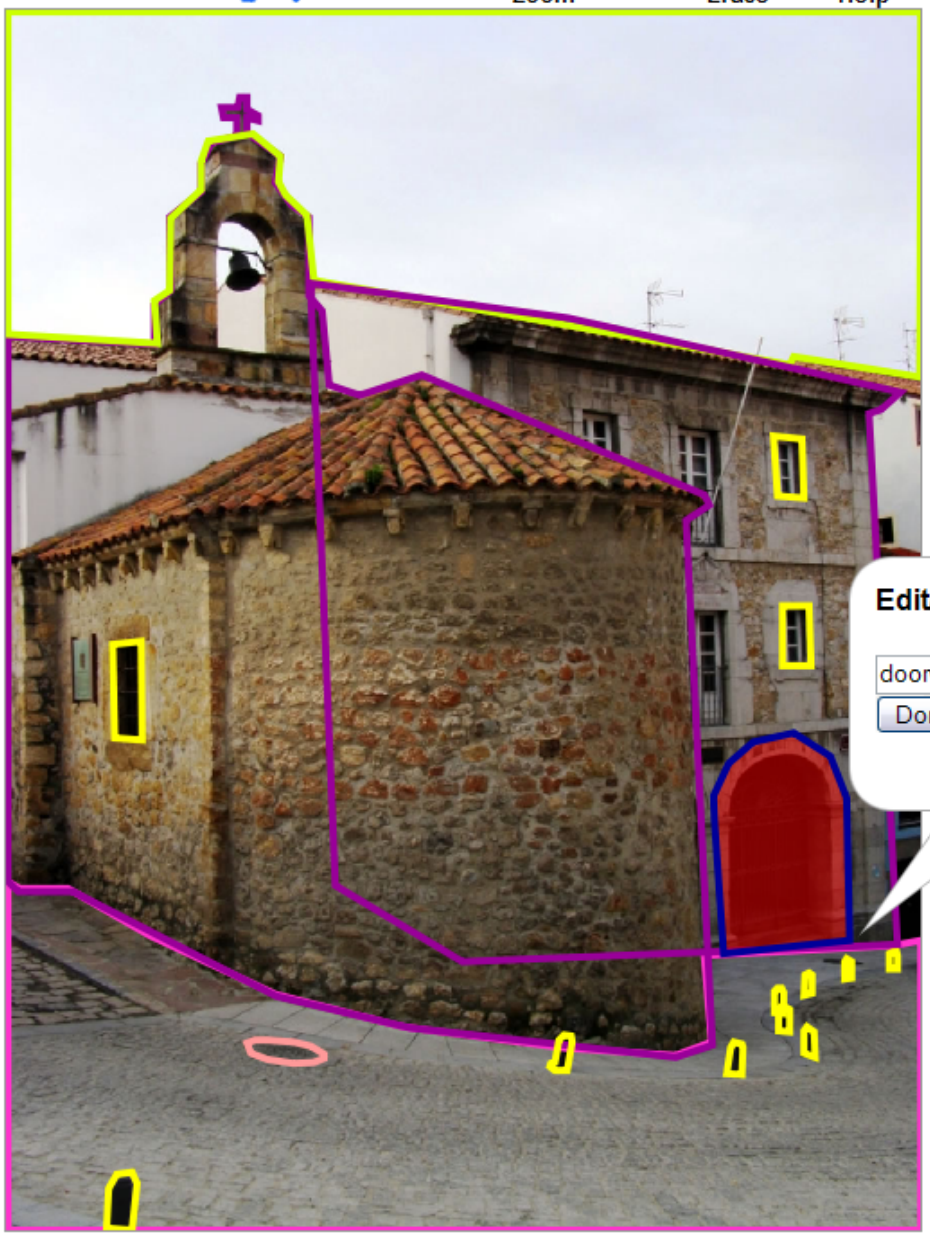
How it works?

When you label objects and their location in an image, the tool uses the labels to build a 3D model of the scene. The tool does not require from you any knowledge about geometry, as all of the 3D information is automatically inferred from the annotations. For instance, the tool will know that a 'road' is a horizontal surface and that a 'car' is supported by the road. The tool learns to go from 2D to 3D using all the other labels already present in the database. The more images that are labeled, the better models the tool will learn.

By using this tool, you can help us build a large collection of annotated images for computer vision research. Currently, computers have a difficult time recognizing objects in images. While practical solutions exist for a few simple classes, such as human faces or cars, the more general problem of recognizing all different classes of objects in the world (e.g. guitars, bottles, telephones) remains unsolved. Computer vision researchers are currently investigating methods that can recognize and localize thousands of different object categories in complex scenes. A key component of these algorithms is the data used to train the computer's model of each object.

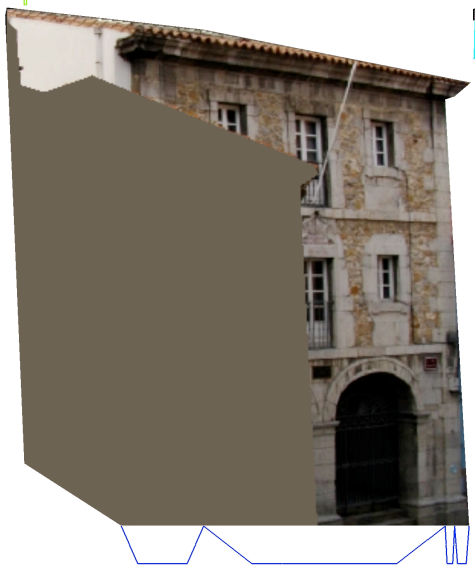


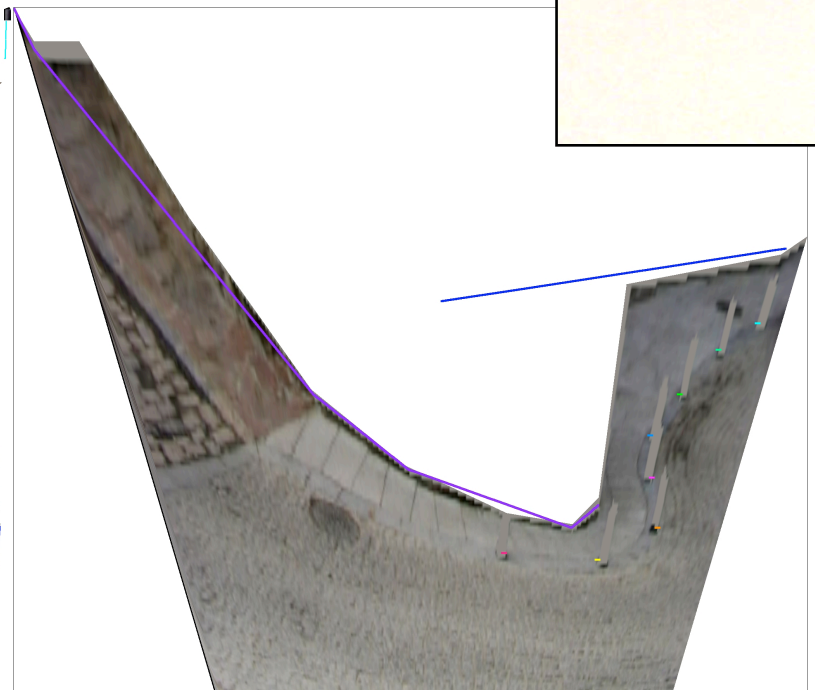
This work is the result of a collaboration between the Computer Science and Artificial Intelligence Laboratory at MIT and the INRIA (Willow project-team, Laboratoire d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548). Funding for this research was provided by National Science Foundation Career award (IIS 0747120).


LabelMe
Zoom
Erase
Help
Sign in (why?)
There are 230642 labelled objects
Polygons in this image (IMG, XML)
road
building
sky
pole
pole
pole
window
window
window
pole
pole
pole
pole
pole
Edit/delete object
doorway
Done
Delete

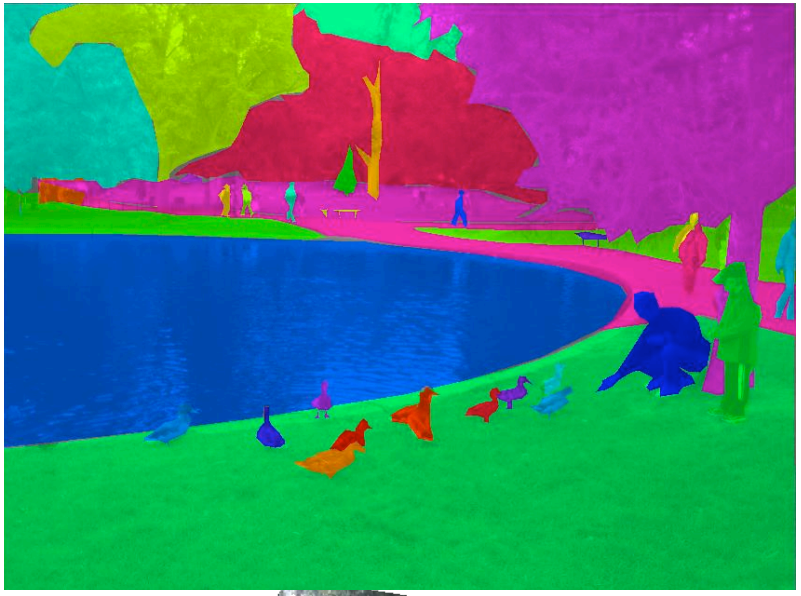






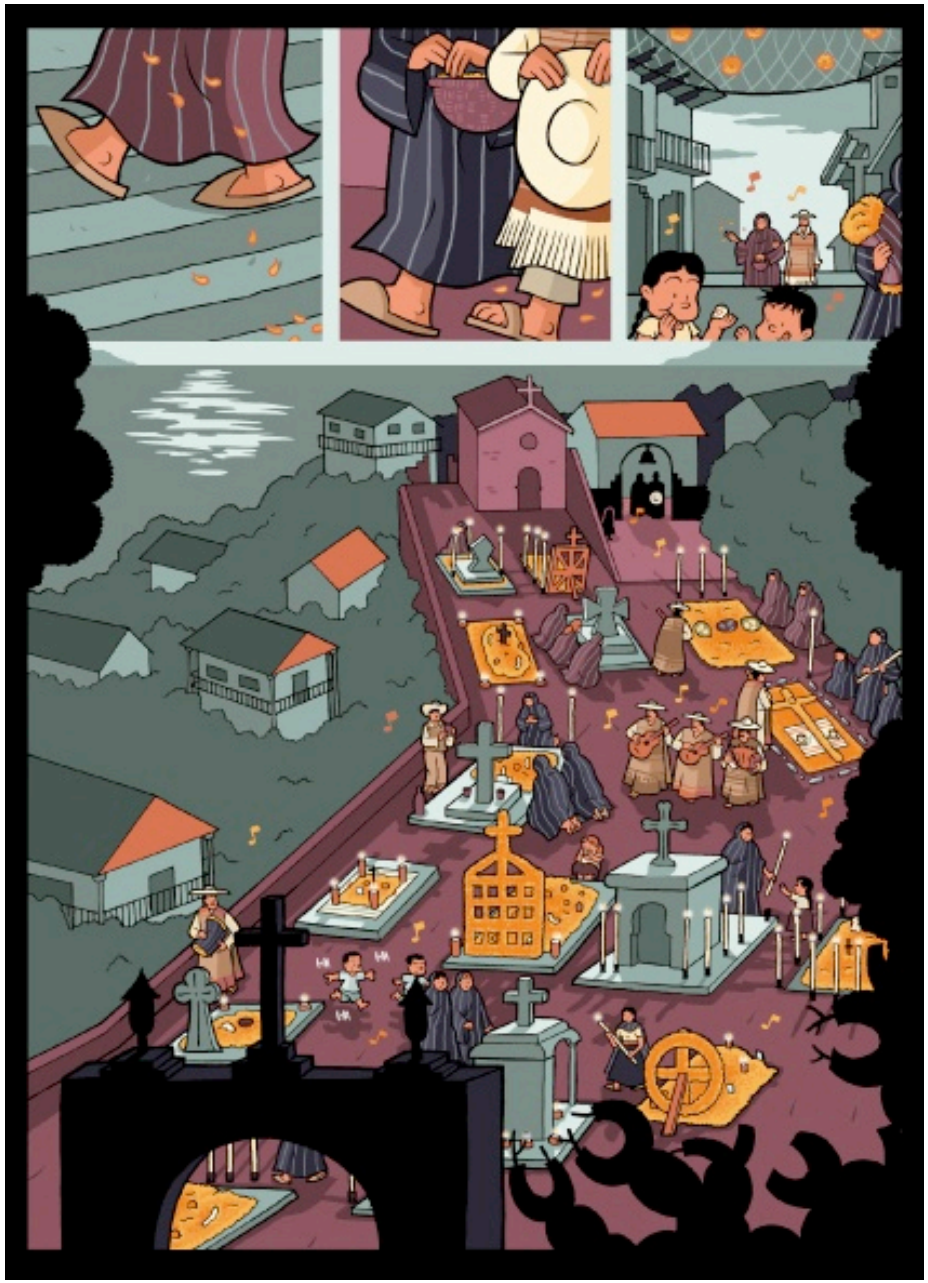








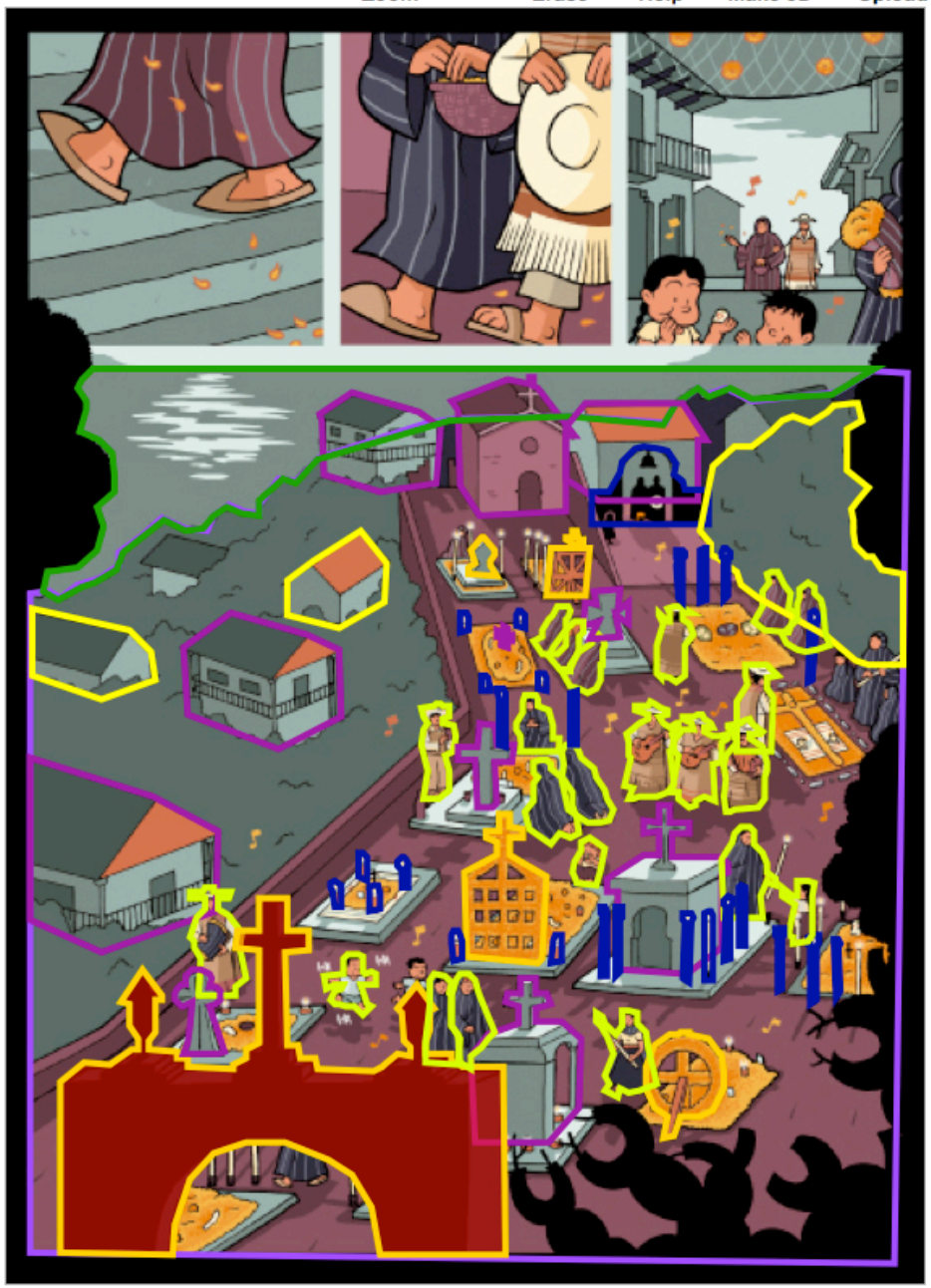
B. Russell, A. Torralba. CVPR 2009



LabelMe

Zoom Erase Help Make 3D Upload im

You are: **admin**
(sign out)

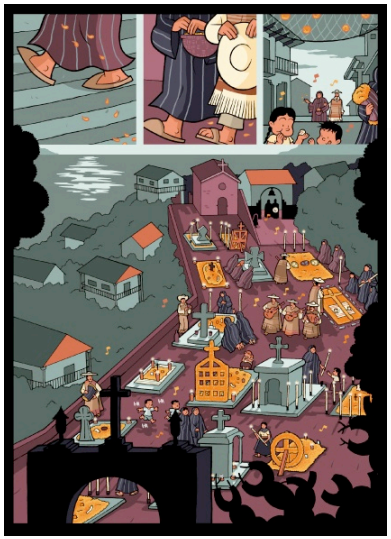


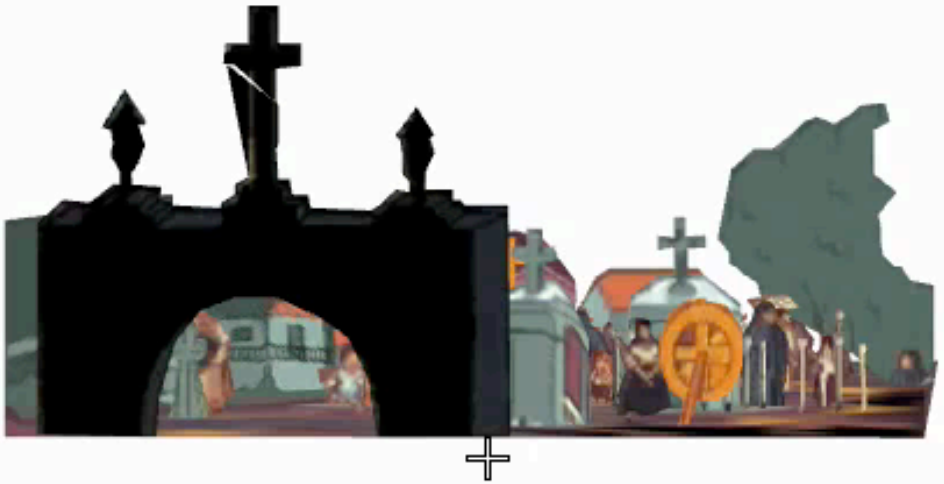
There are **622136** labelled objects

Polygons in this image

(IMG, XML)

ground
building
person
person
building
building
sea
building
building
person
person
building
fence
person
person
person
building
person
sculpture
cross
candle
person
person
person
candle
candle
candle
candle
person
person
sculpture
person
sculpture
person
cross
person
candle
candle
house
person
sculpture
candle
candle
candle
candle
candle





3D measuring tool

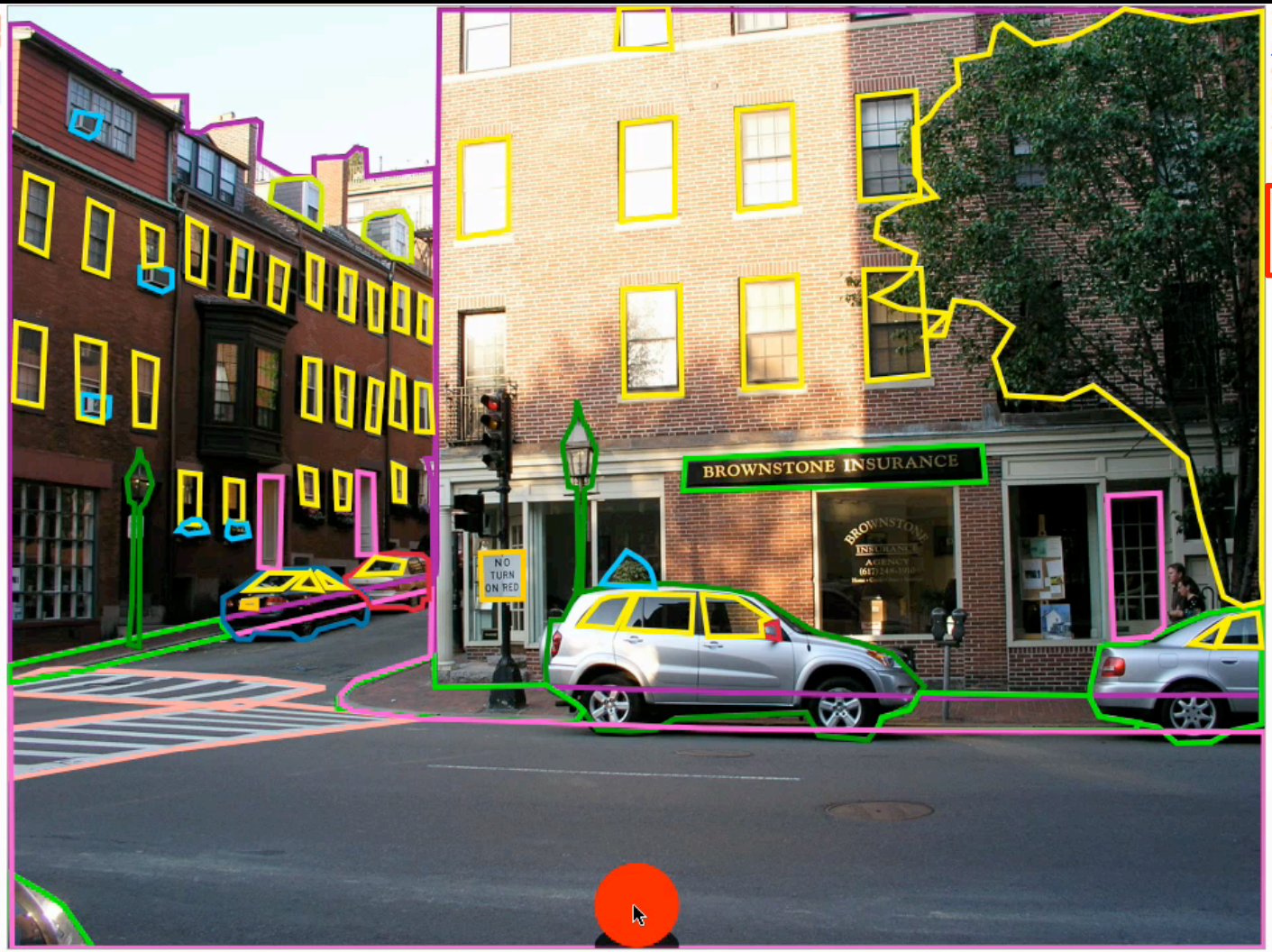


Car is 13.68 meters away

LabelMe

Sign in (why?)

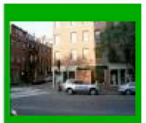
Help
Annotation tool
Zoom: + - fit
Delete
Make scene 3D
Navigate in 3D
Measure in 3D
3D visualization
Anaglyph
Pop-up book
Files
Image
XML



3D Information
meters | feet
ground (no object labeled here): 0
2D: (472,707)
Norm 2D: (427,639)
3D: (1,0,473)
Depth: 4.73 meters away

Your gallery

Upload image


p1010923.jpg


p1010919.jpg

LabelMe Video



J. Yuen, et al ICCV 09

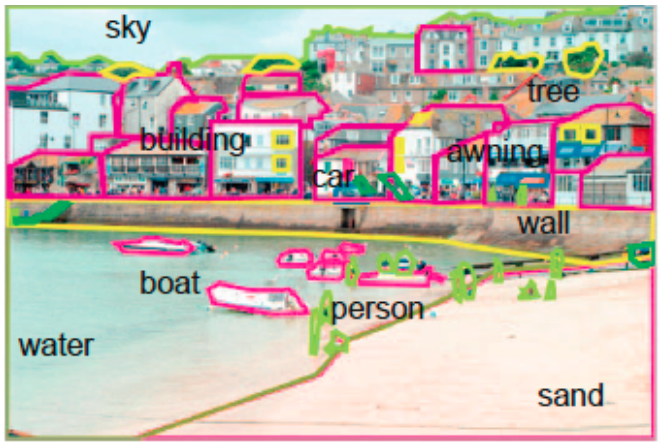


Levels of description

The scene



The objects



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 ?


Yes



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

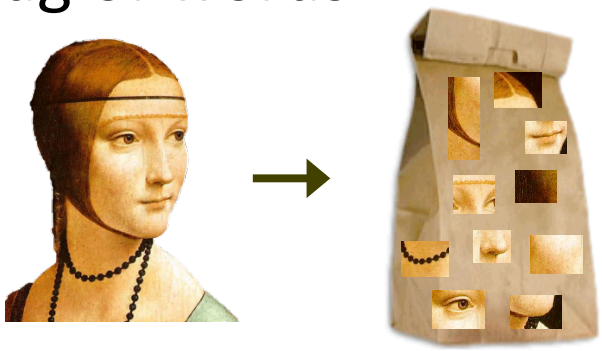
In a glance, we remember the meaning of an image and its global layout but some objects and details are forgotten





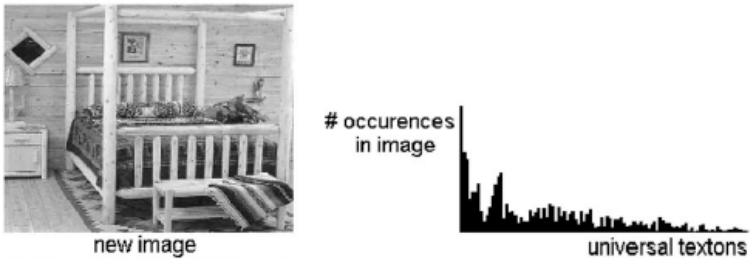
Global Texture Descriptors

Bag of words



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

Non localized textons



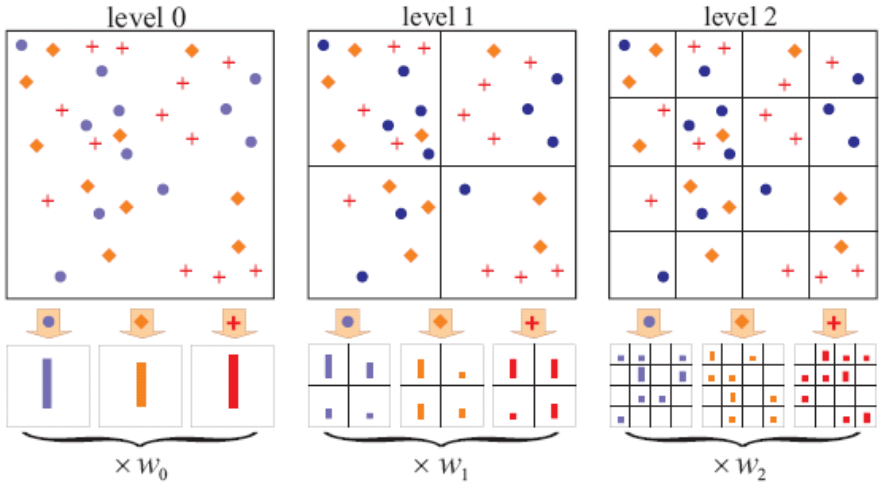
Walker, Malik. Vision Research 2004

...

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.

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

How many object categories are there?



Biederman 1987

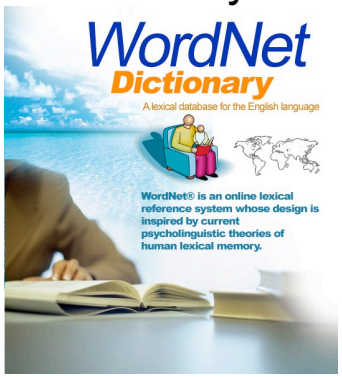
SUN Dataset Project

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

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



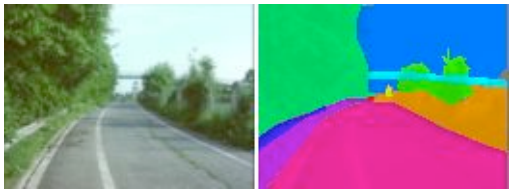
2. We download images and clean the categories

Google Image Search

altavista

flickr

3. We segment all the images





Jianxiong Xiao



James hays

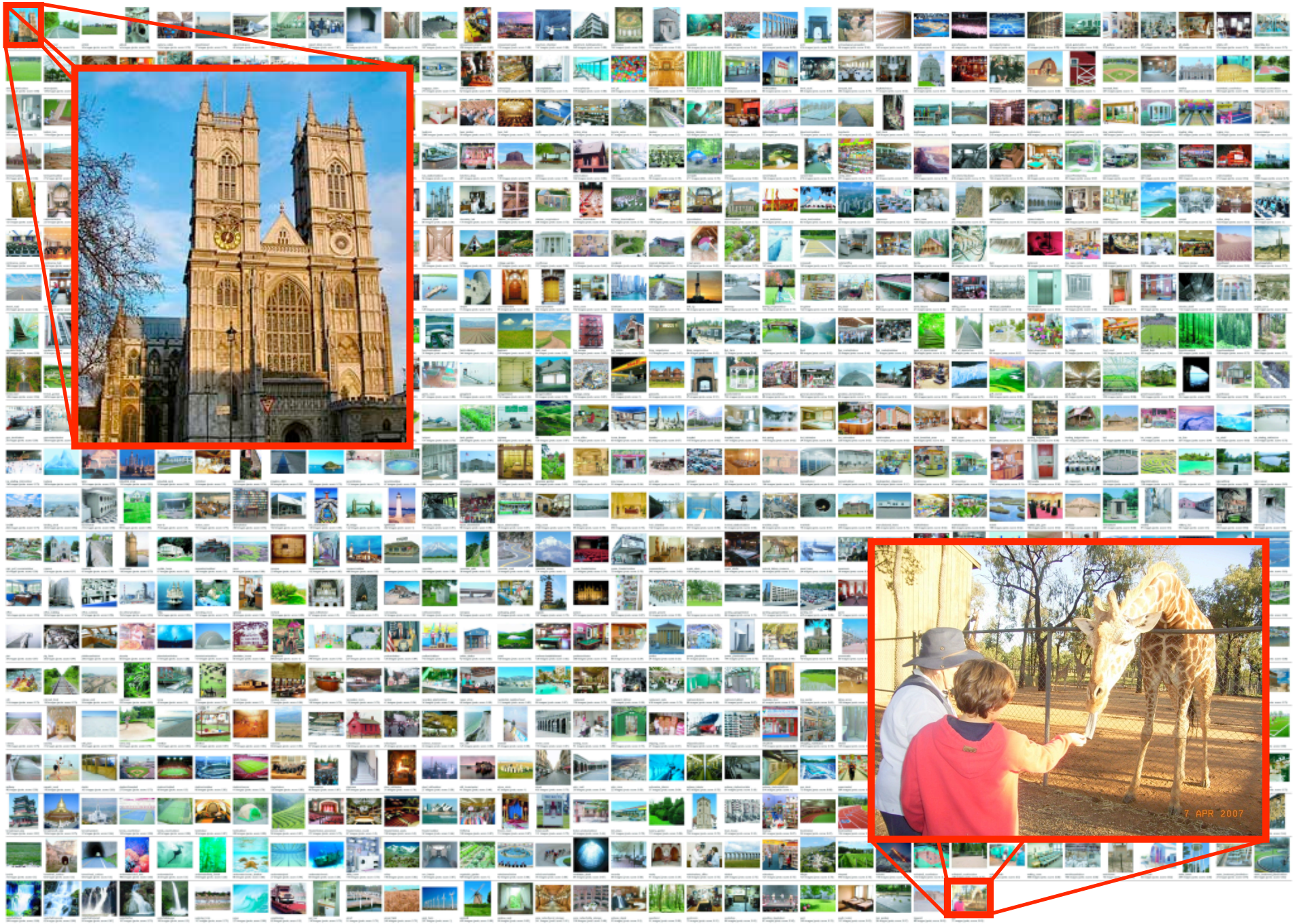


Krista Ehinger



Aude Oliva



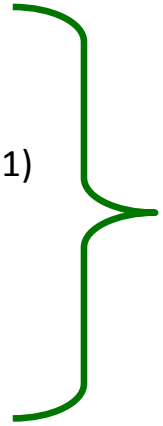


Xiao, Hays, Ehinger, Oliva, Torralba; CVPR 2010

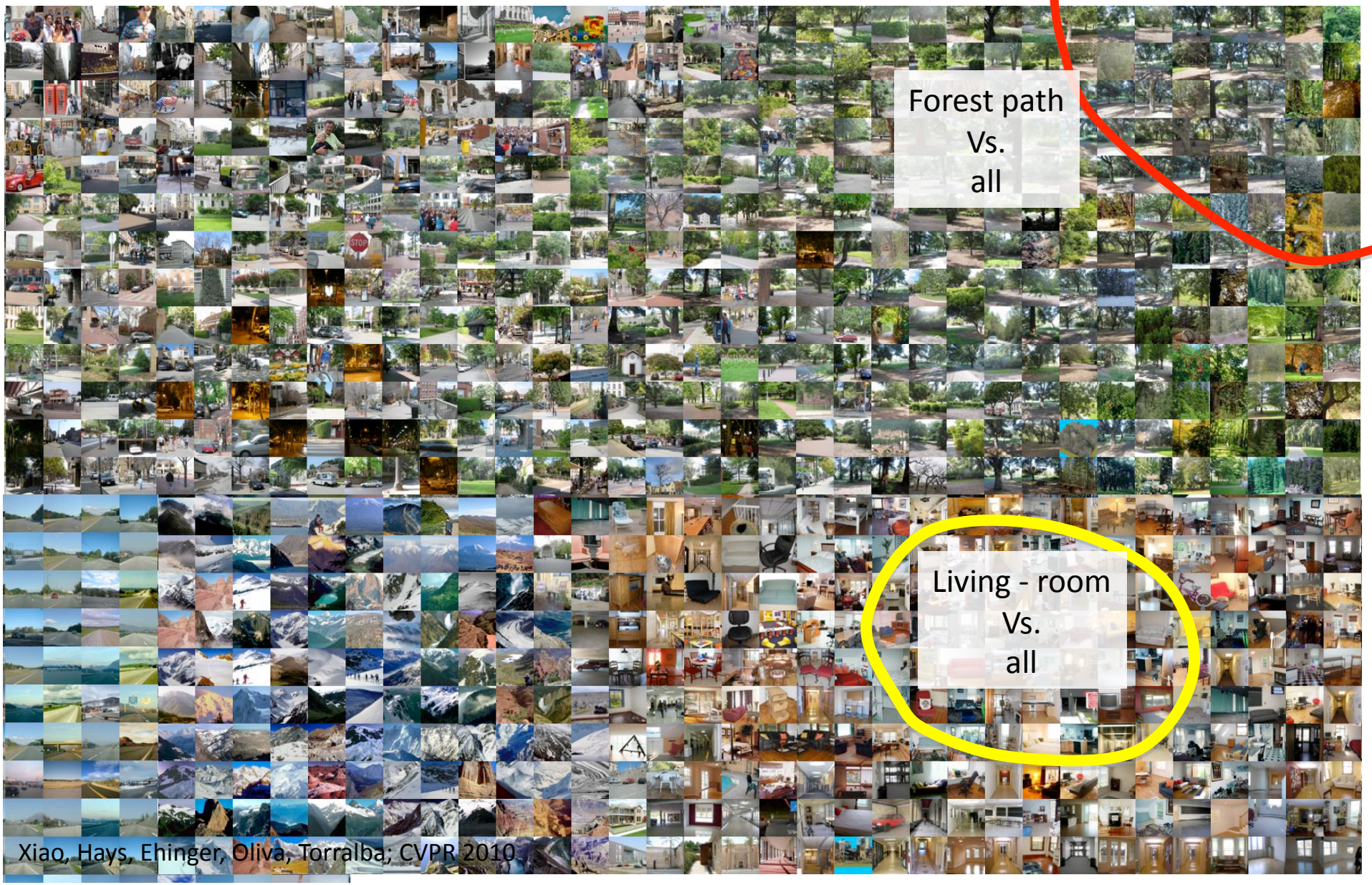
397 Well-sampled Categories

Global Image Descriptors

- Tiny images (Torralba et al, 2008)
- Color histograms
- Self-similarity (Shechtman and Irani, 2007)
- Geometric class layout (Hoiem et al, 2005)
- Geometry-specific histograms (Lalonde et al, 2007)
- Dense and Sparse SIFT histograms
- Berkeley texton histograms (Martin et al, 2001)
- HoG 2x2 spatial pyramids
- Gist scene descriptor (Oliva and Torralba, 2008)

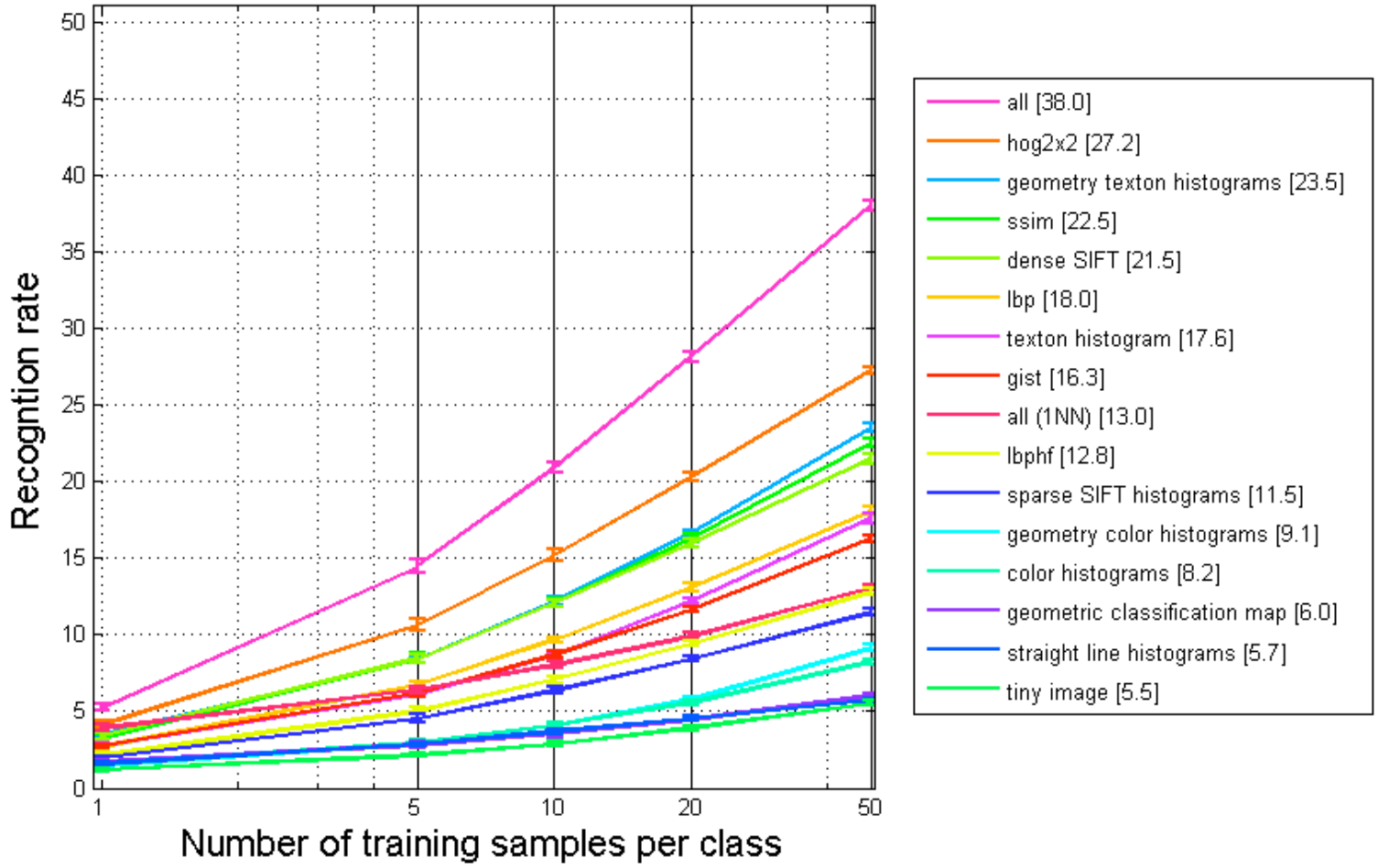


Learning Scene Categorization



397-Scene recognition

Humans [68.5]

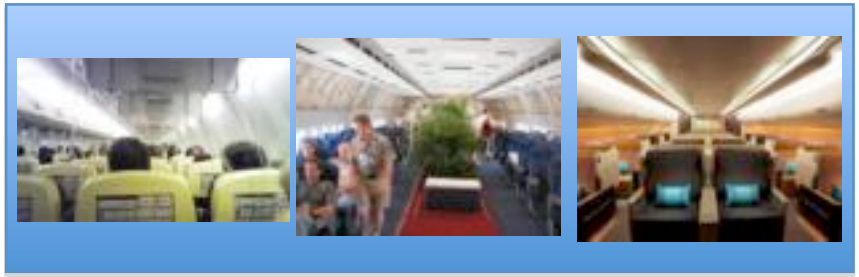


Classifier: 1-vs-all SVM with histogram intersection, chi squared, or RBF kernel.

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

A look into the results

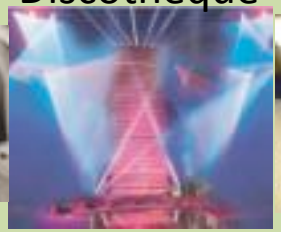
Airplane cabin (64%)



Van interior



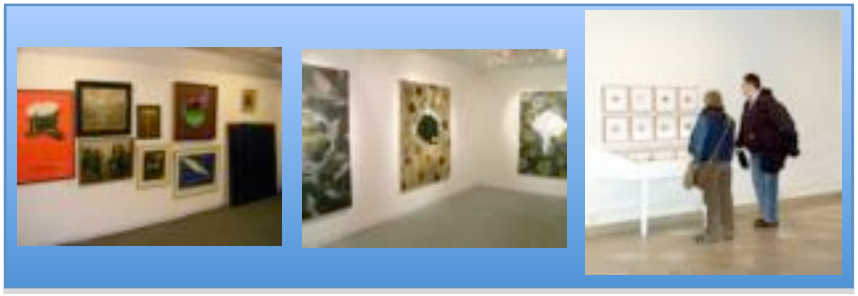
Discotheque



Toyshop



Art gallery (38%)



Iceberg



Hotel room



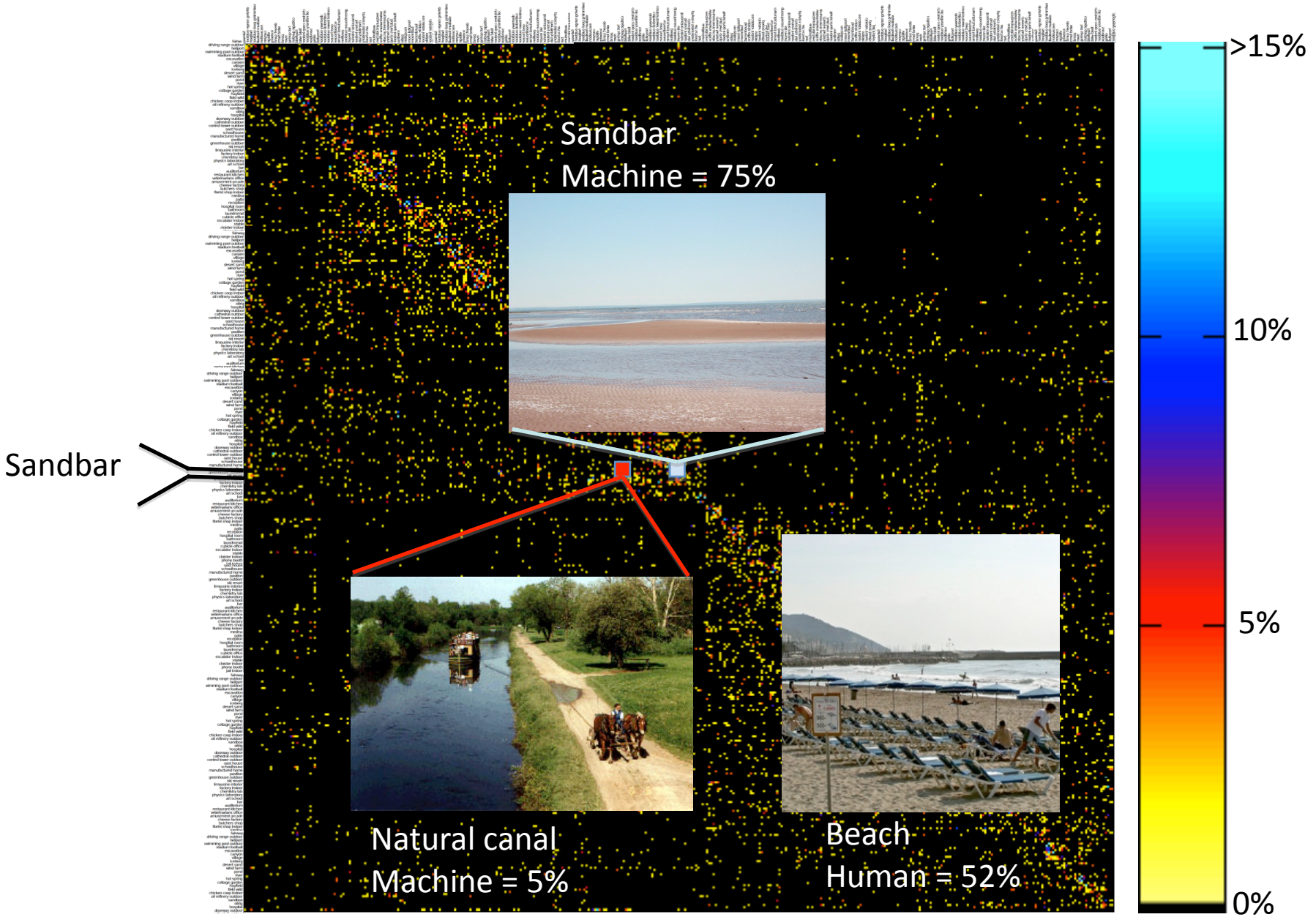
Kitchenette



...

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

A 397 x 397 confusion matrix

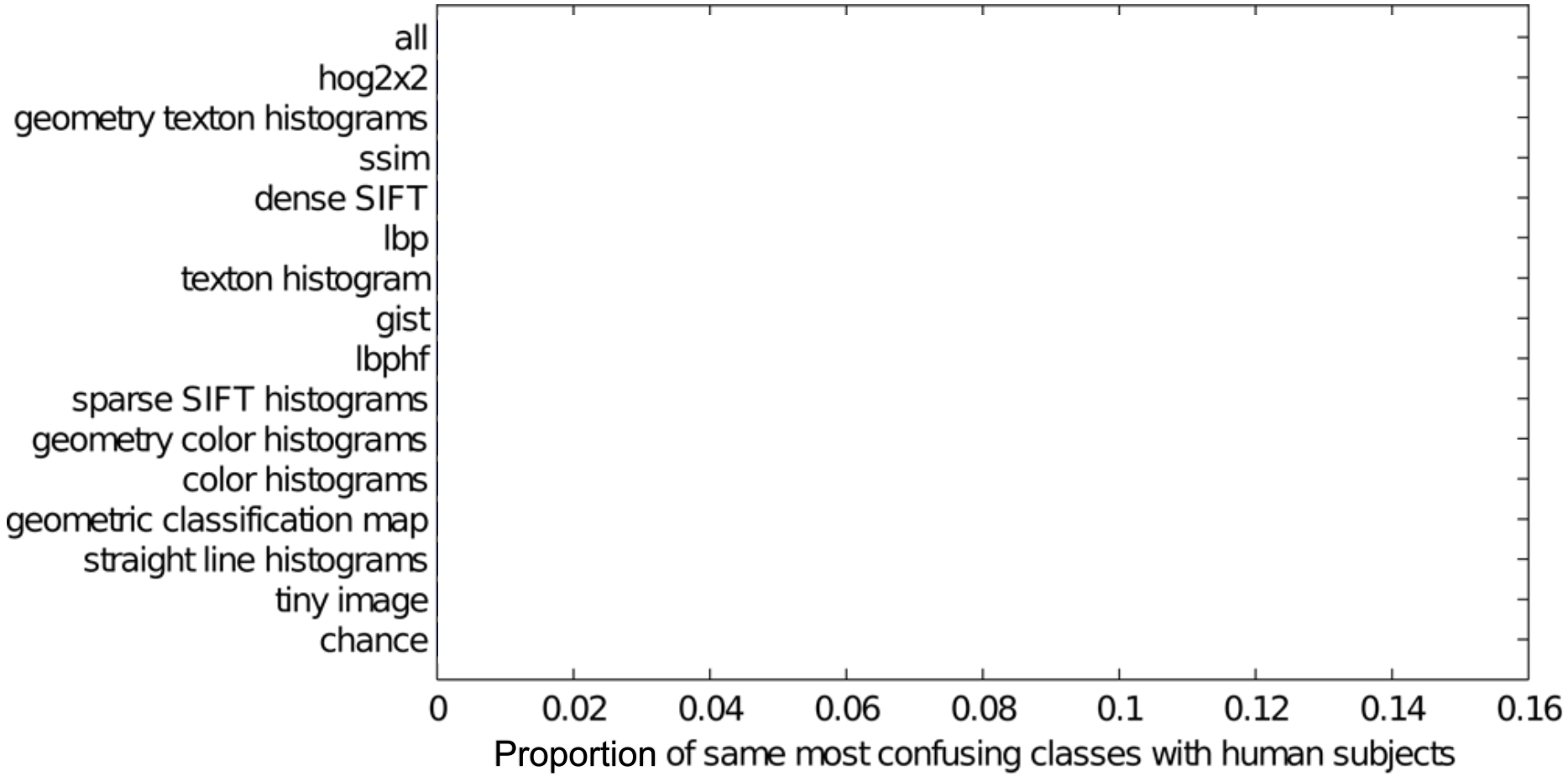


Xiao, Hays, Ehinger, Oliva, Torralba; CVPR 2010

True category	Humans	Tiny image	Gist	All features
bus_interior	subway_interior (24%)	podium indoor (10%)	jail indoor (10%)	subway_interior (15%)
				
general_store indoor	bazaar indoor (16%)	apse indoor (5%)	amusement_arcade (10%)	florist_shop/indoor (10%)
				
pasture	field wild (19%)	athletic_field/outdoor (10%)	corral (10%)	desert/vegetation (10%)
				

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

Human and Computer Confusion Overlap



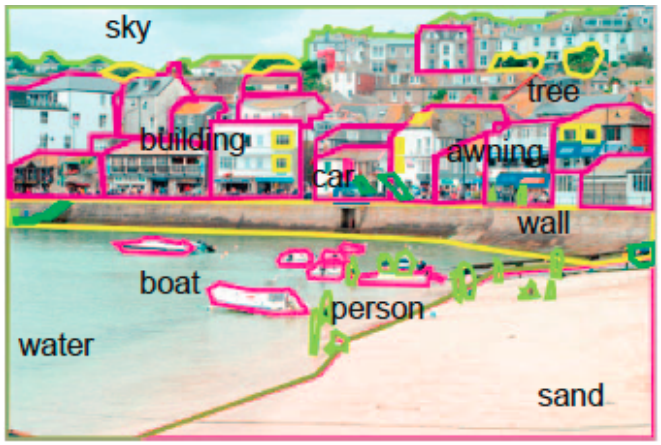
Xiao, Hays, Ehinger, Oliva, Torralba; CVPR 2010

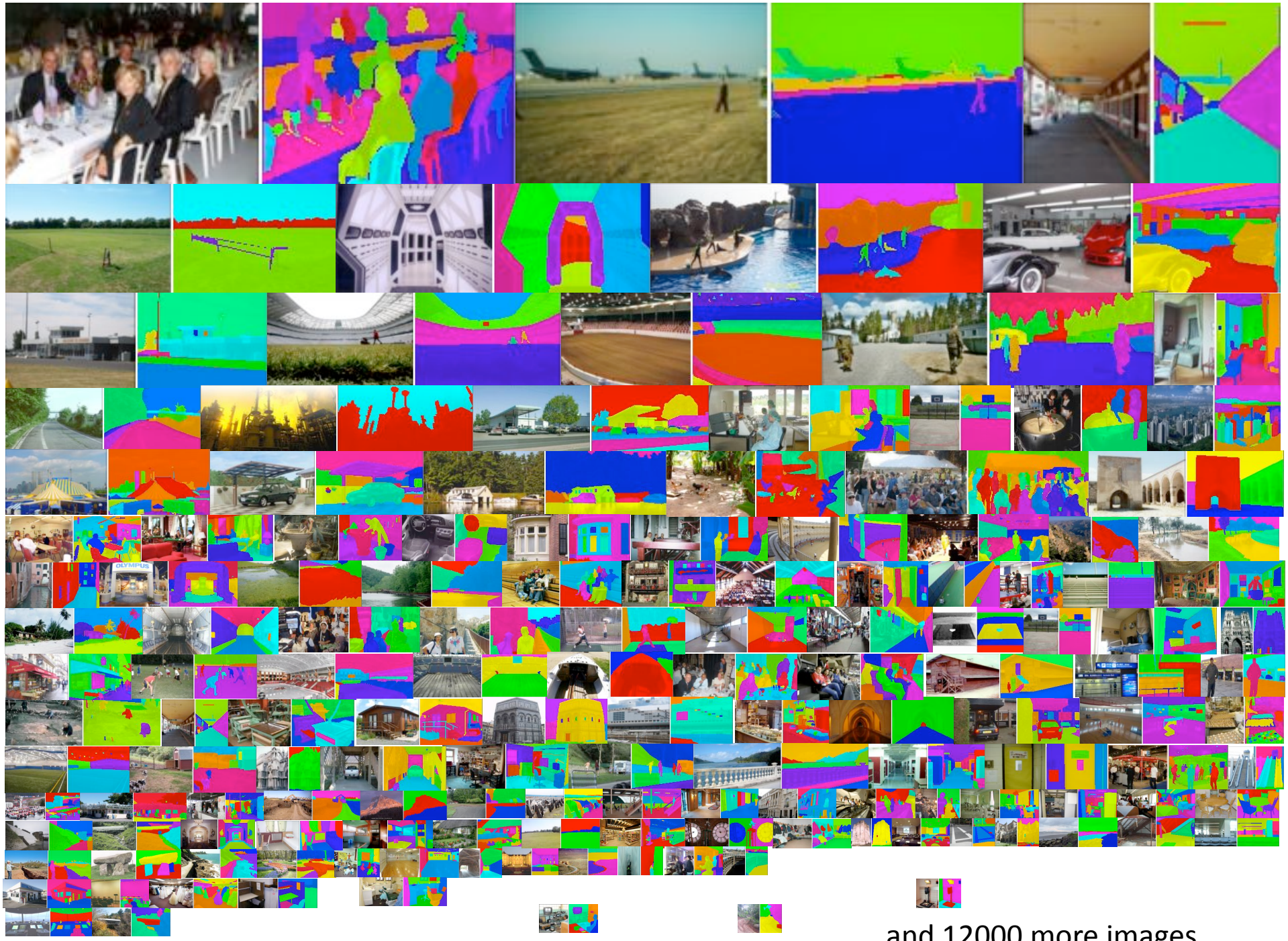
Levels of description

The scene



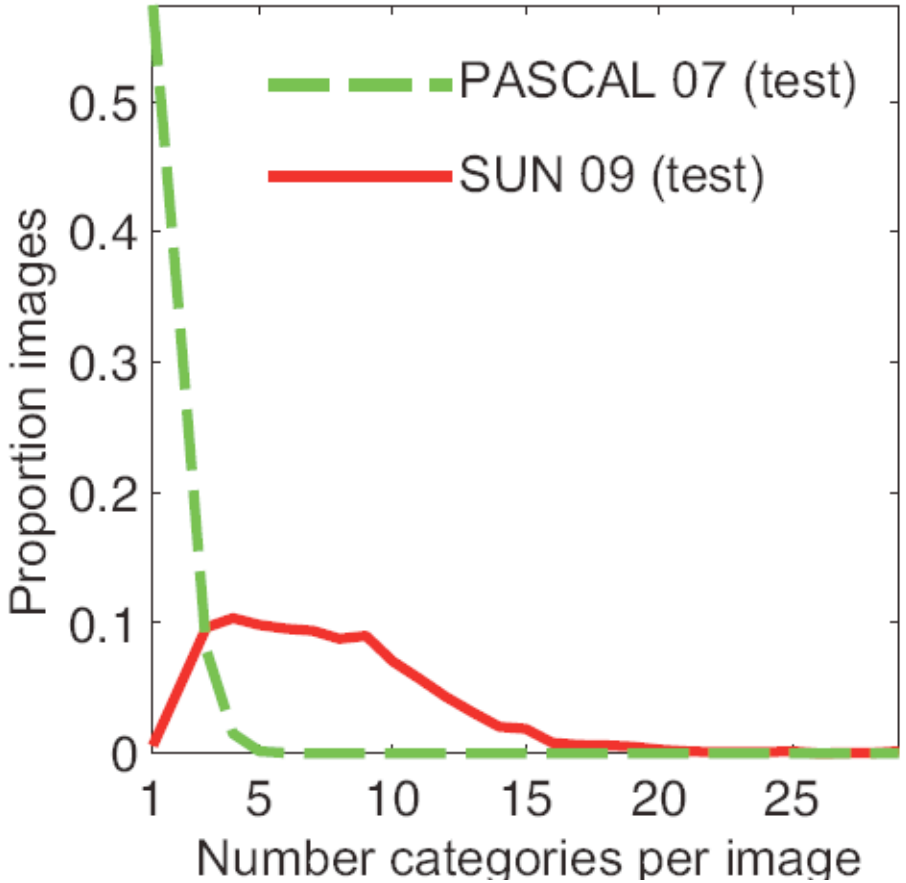
The objects




…and 12000 more images

SUN database





12,000 annotated images

107 object categories

152,000 annotated object instances

The best objects/regions



The worst objects/regions

Can you guess what are they trying to detect?

Bread







Bench







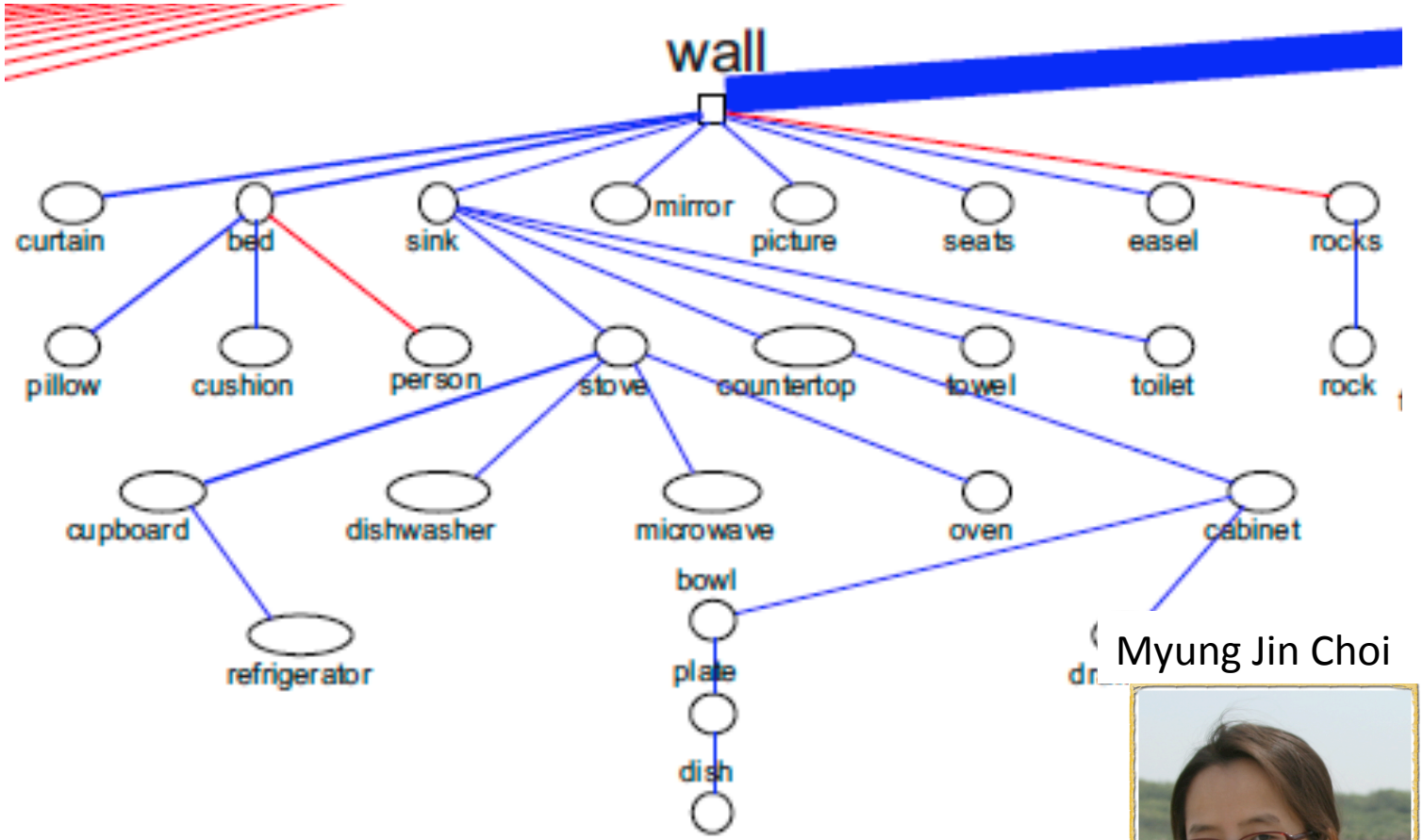

Vase








Exploiting Hierarchical Context on a Large Database of Object Categories



Myung Jin Choi, Joseph Lim, Antonio Torralba, and Alan S. Willsky. CVPR 2010

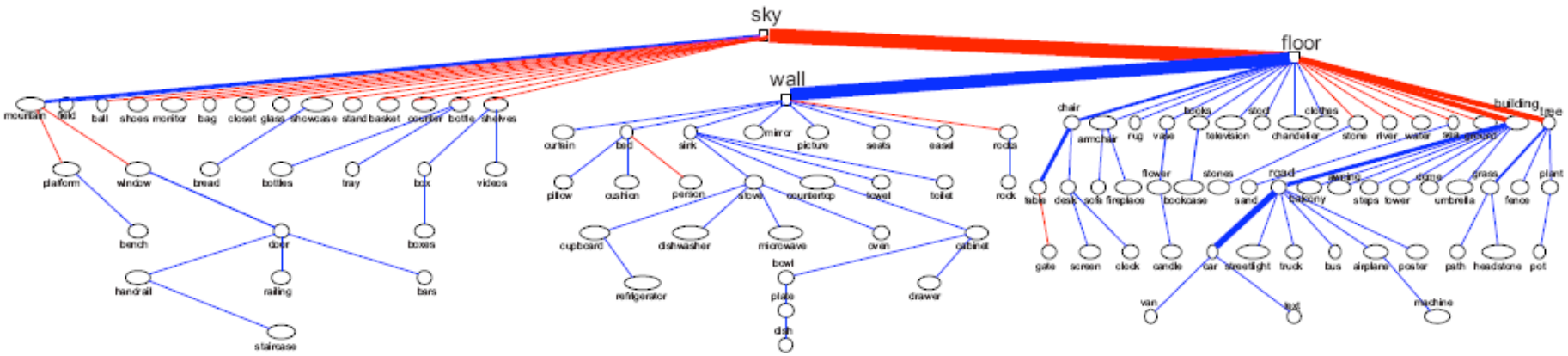
Myung Jin Choi



Joseph Lim



Tree learned from SUN 09



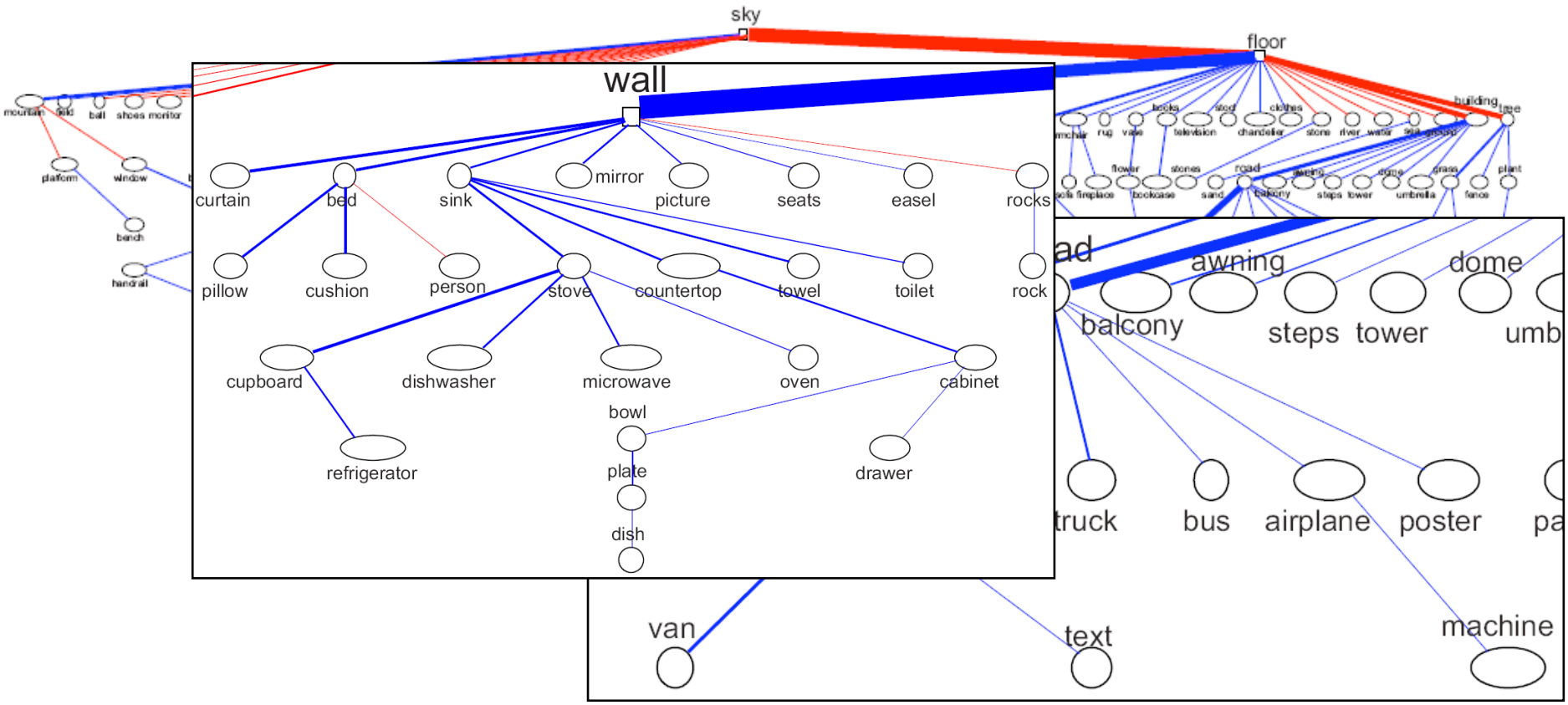


107 object categories

4317 training images

25/106 edges and 7/top-53 edges (≈13%) negative

Learning object dependencies



107 object categories

4317 training images

25/106 edges and 7/top-53 edges (≈13%) negative

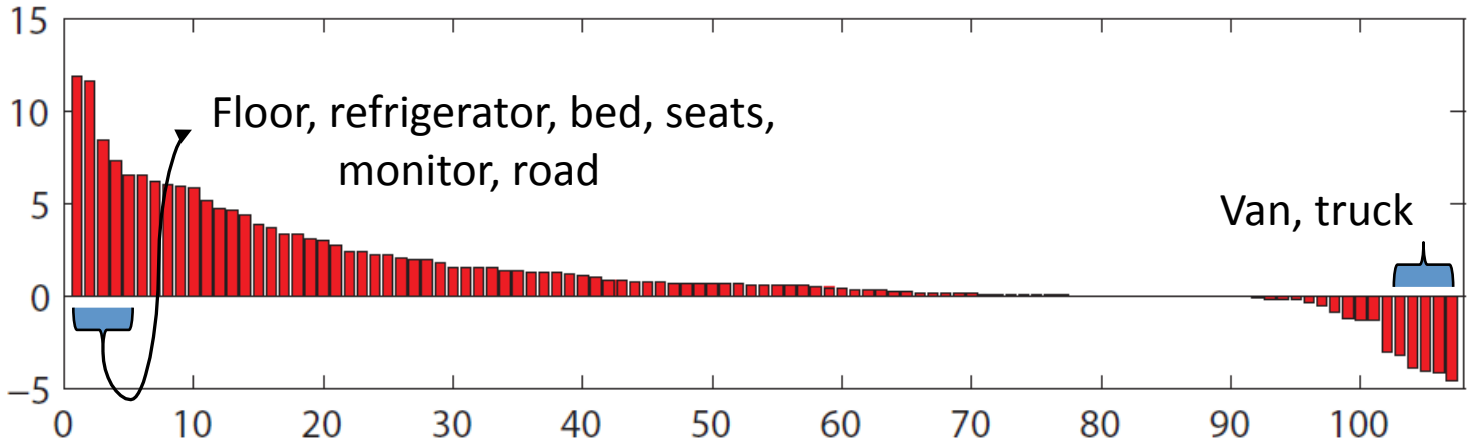

Baseline
sky
wall
cabinet
building
floor
road
building
sky
sky
person
person
floor
wall
sky
truck
cupboard
building
dishwasher
With Context
sky
building
tree
road
grass
car
shelves
bookcase
person
person
desk
floor
wall
cupboard
refrigerato
stove
dishwasher
cupboa

Baseline

With Context

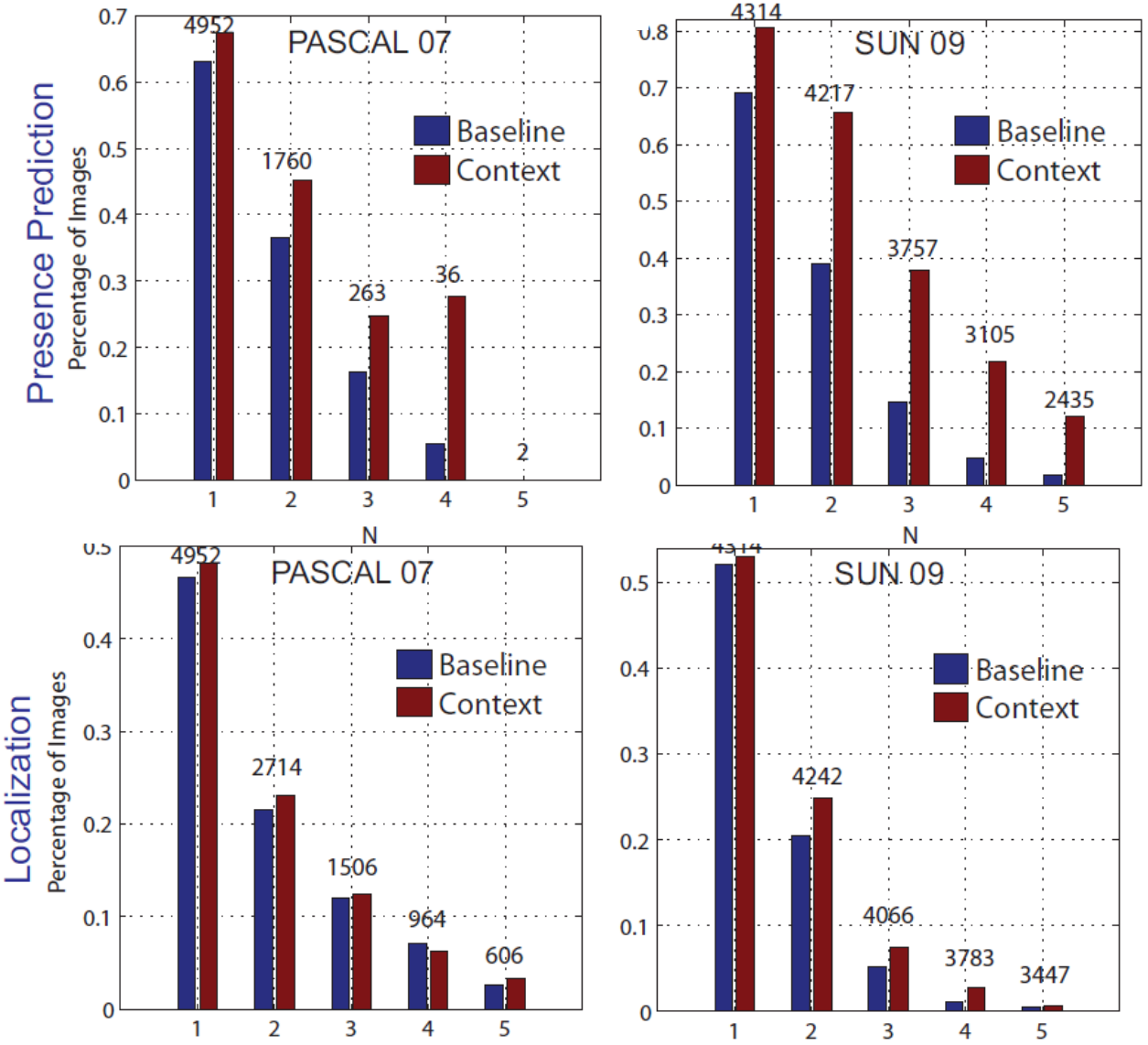


Localization improvement with respect to baseline



Scene Interpretation

How many correctly detected objects are there per image?



Microwave & refrigerator Search







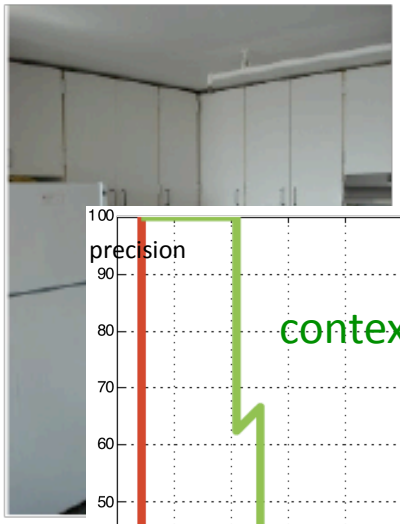


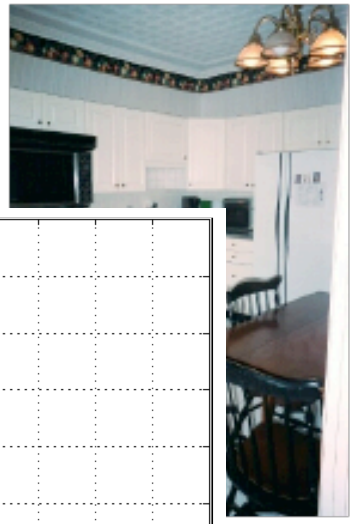

24 targets out of 4317 images

Microwave & refrigerator Search

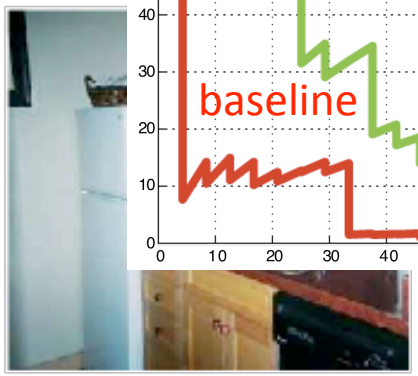










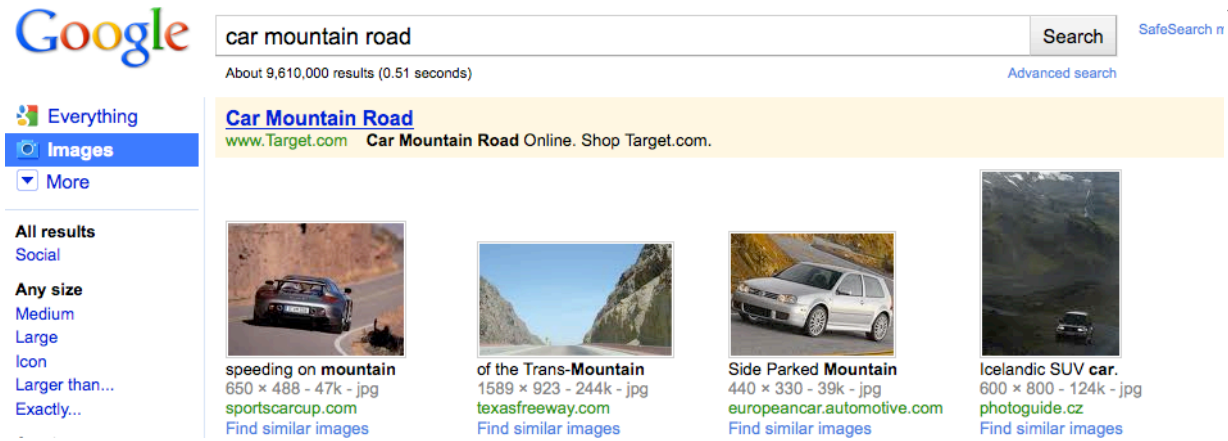


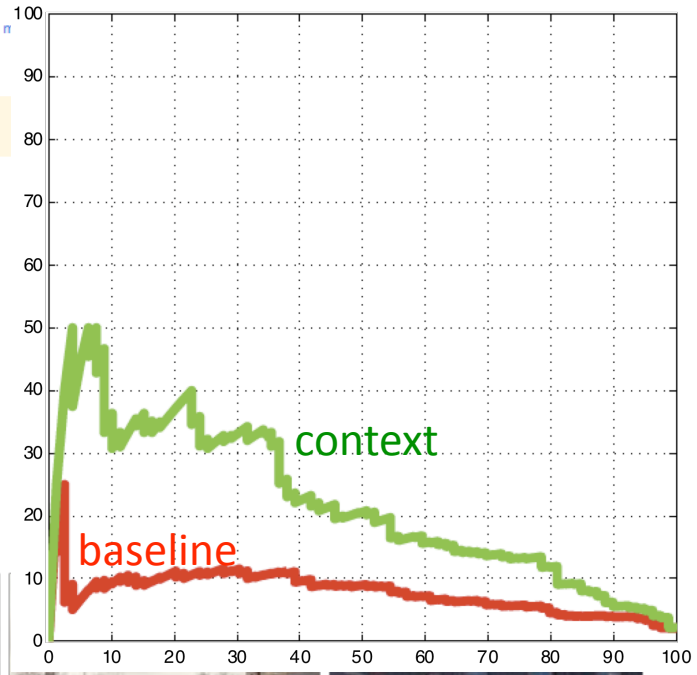




24 targets out of 4317 images

The context model benefits from the 107 object classes,
and not just from the 2 searched classes





Baseline



Context



79 targets out of 4317 images

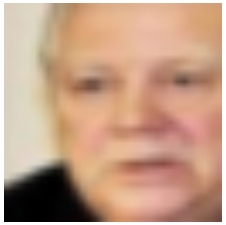
Who needs context anyway?

We can recognize objects even out of context



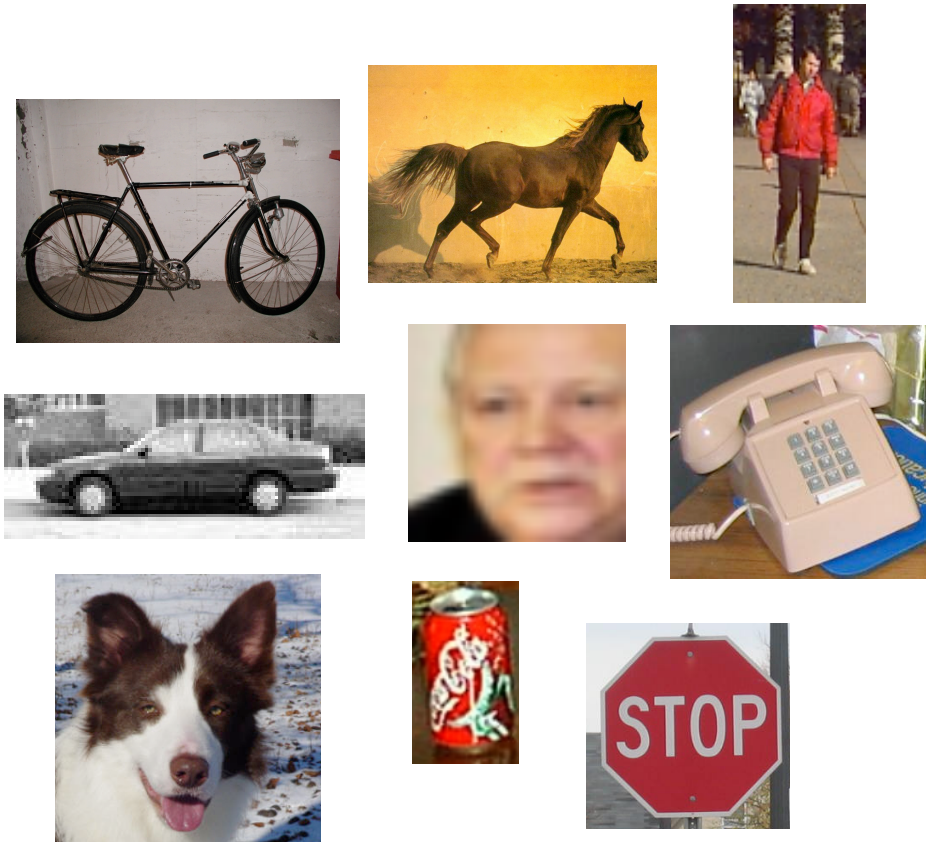
Banksy

Is context really needed?

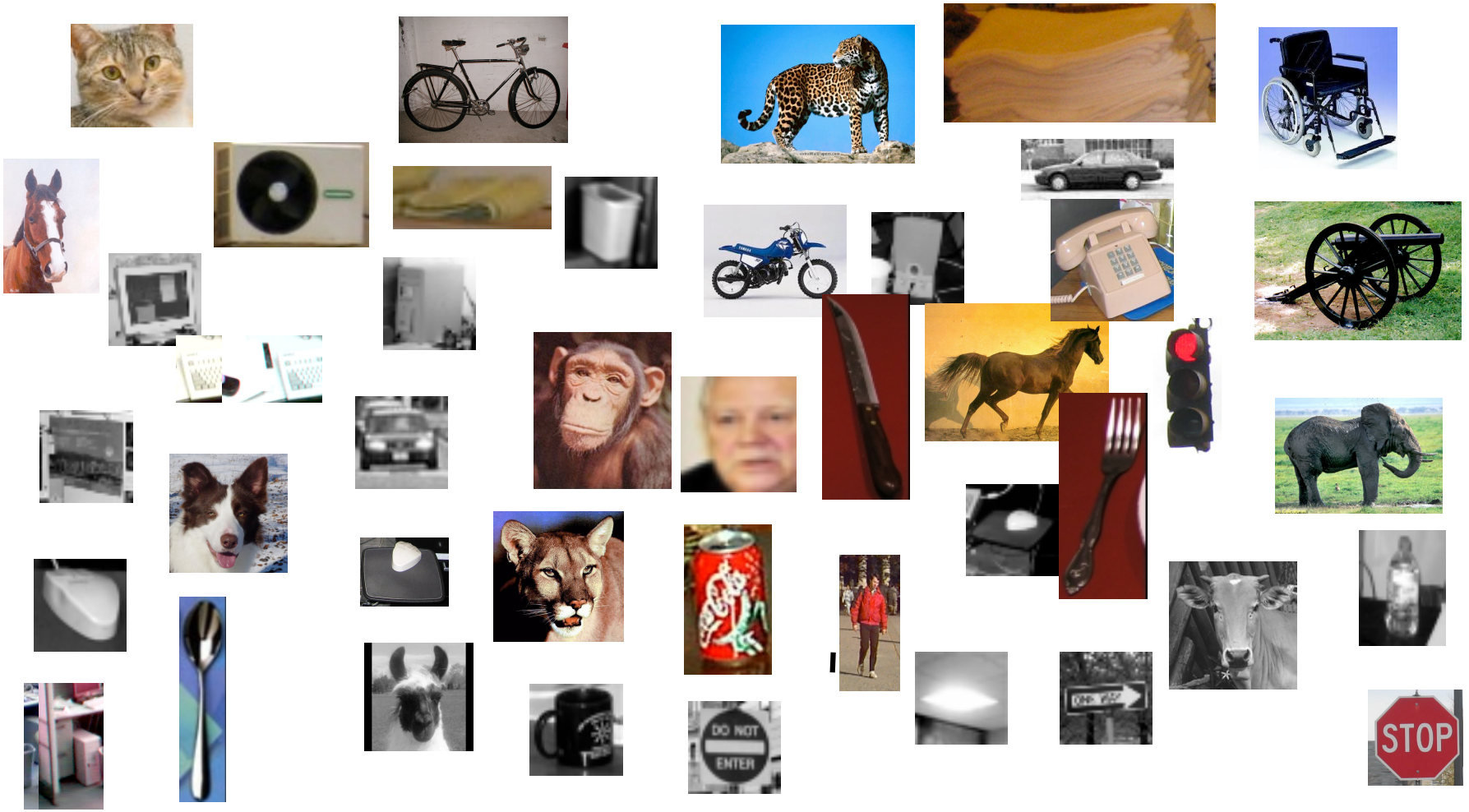

Objects are clearly defined by their local appearance

Is context really needed?



Objects are clearly defined by their local appearance

…Things get a lot more interesting when considering many classes



Objects are clearly defined by their local appearance.

Is context really needed?



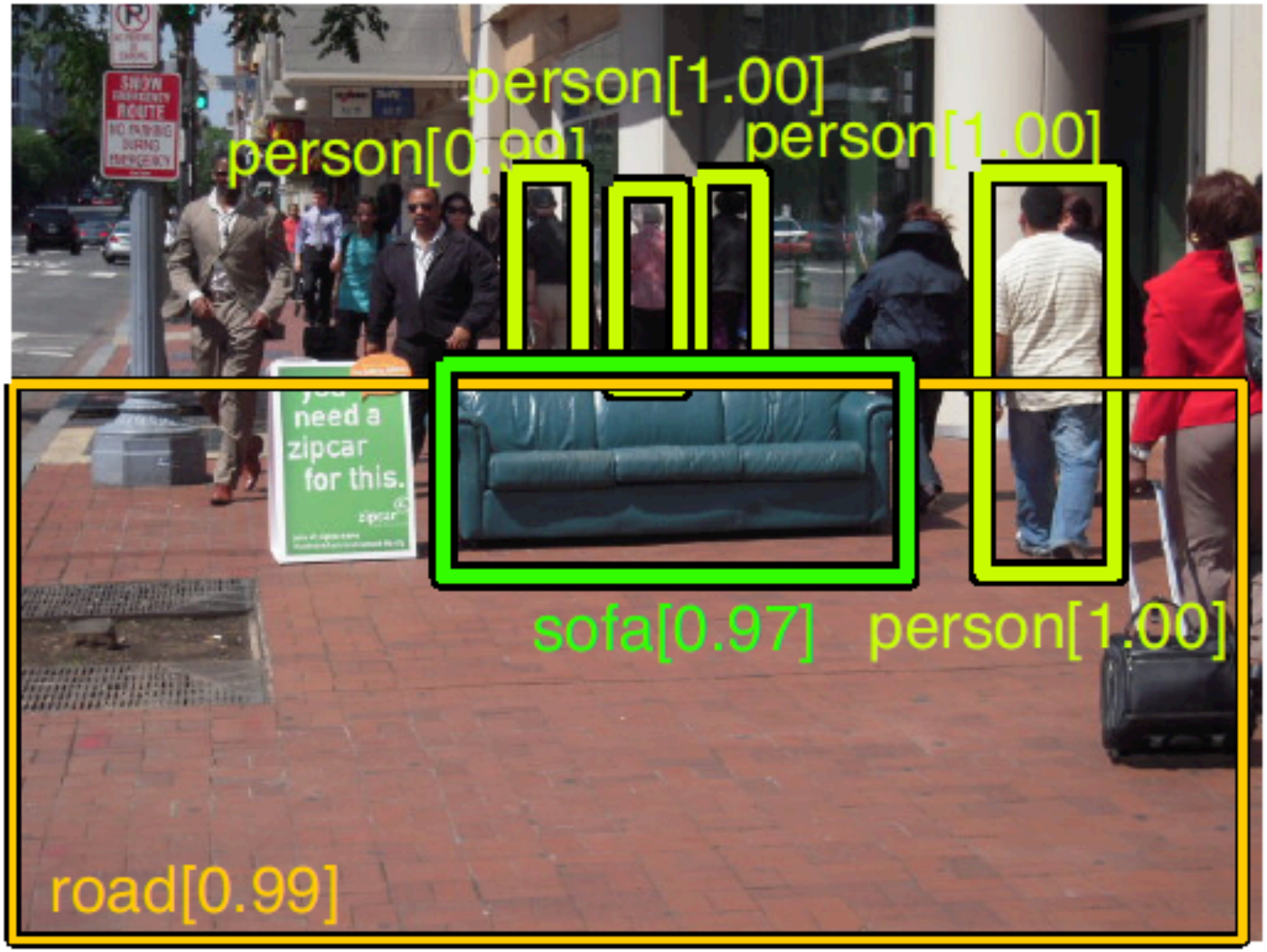






We need context to know what is not supposed to be there




person[1.00]
person[1.00]
person[1.00]
need a
zipcar
for this.
sofa[0.97]
person[1.00]
road[0.99]


you need a zipcar for this.
zipcar

Out of context objects in the real world



Out of context objects in the real world



Out of context objects in the real world



Detecting out of context objects



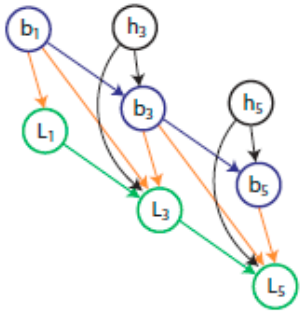
Detecting out of context objects



Detecting out of context objects

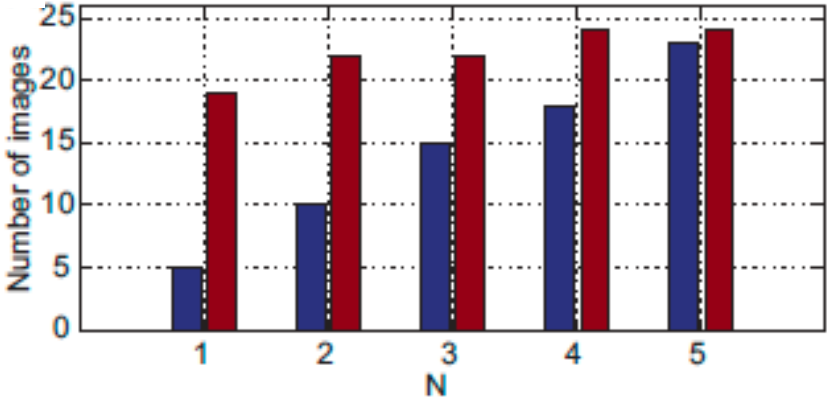


Test on 26 images with 1 or few out of context objects

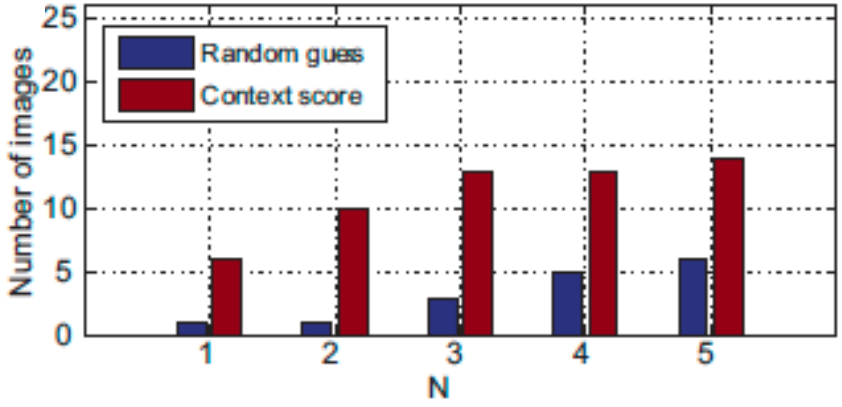




With groundtruth



With detector outputs



Challenge: telling apart false alarms from out of context objects

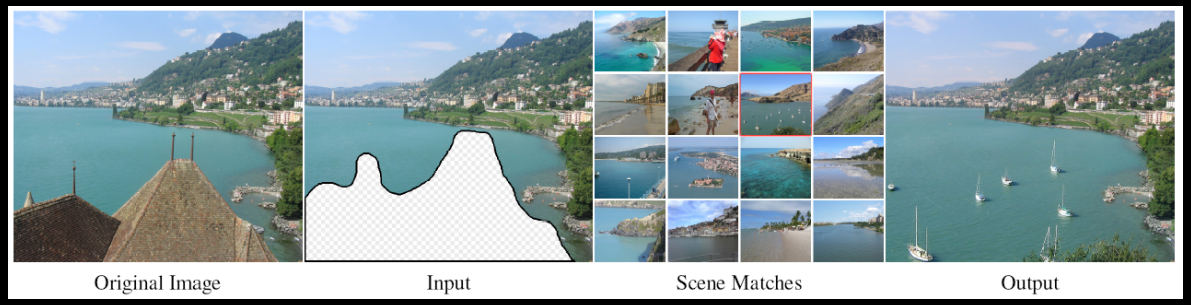
Large databases

PhotoSynth, Snavely et al. 2006



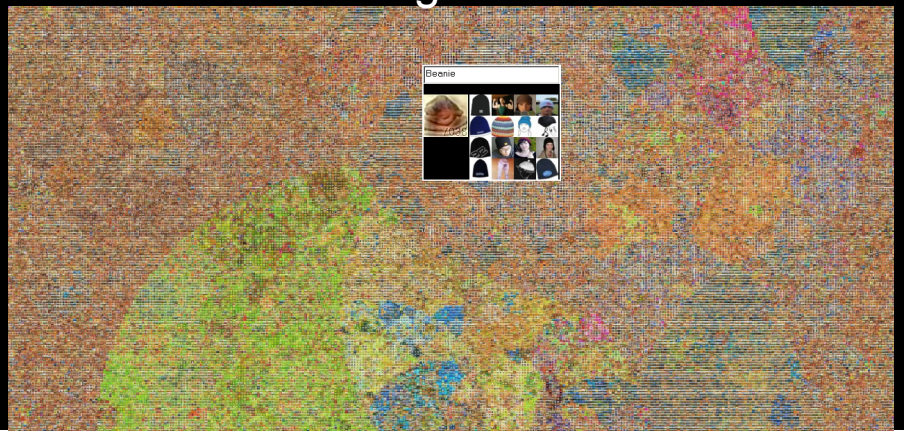
Image completion using Flickr images

Hays and Efros, 2007

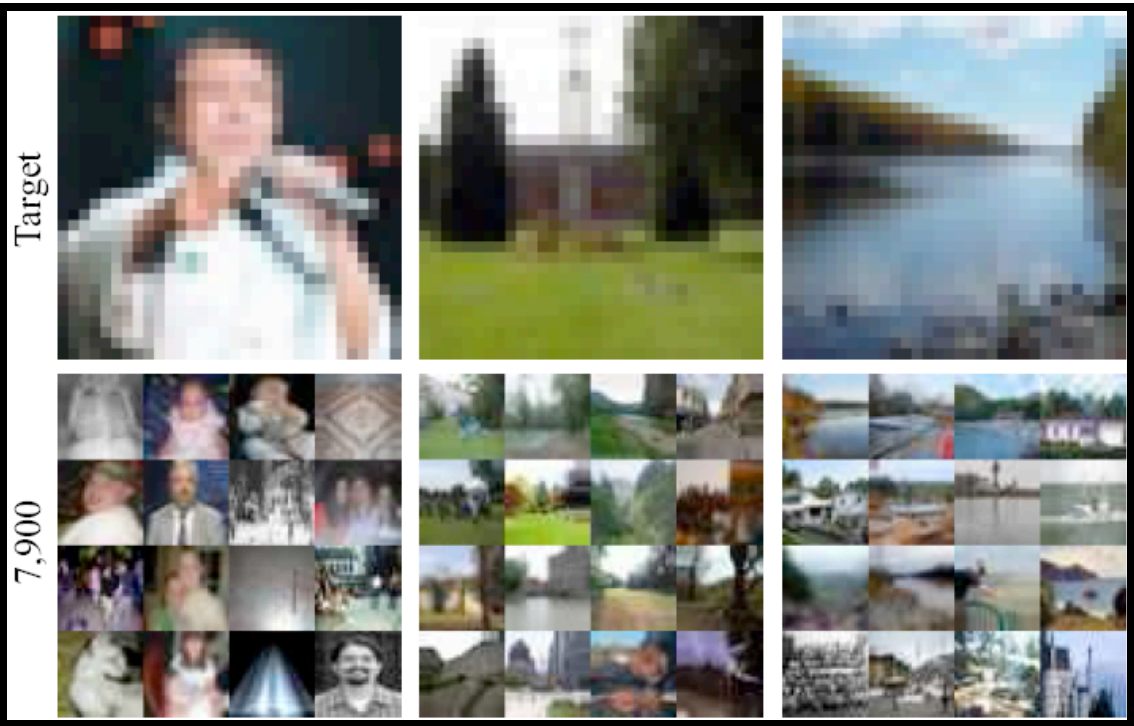


Recognition: 80 million images

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

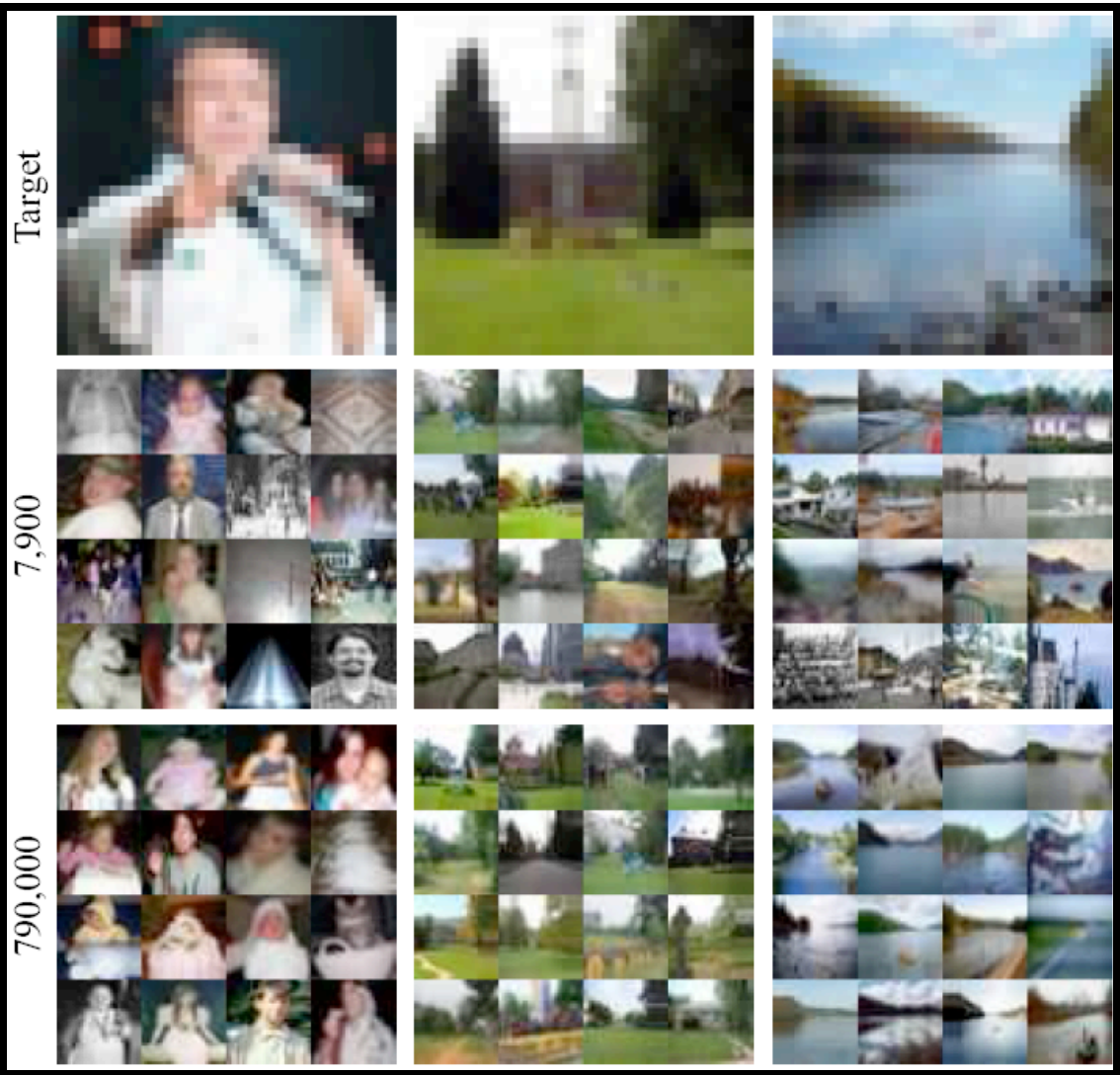


The Power Of Lots Of Images



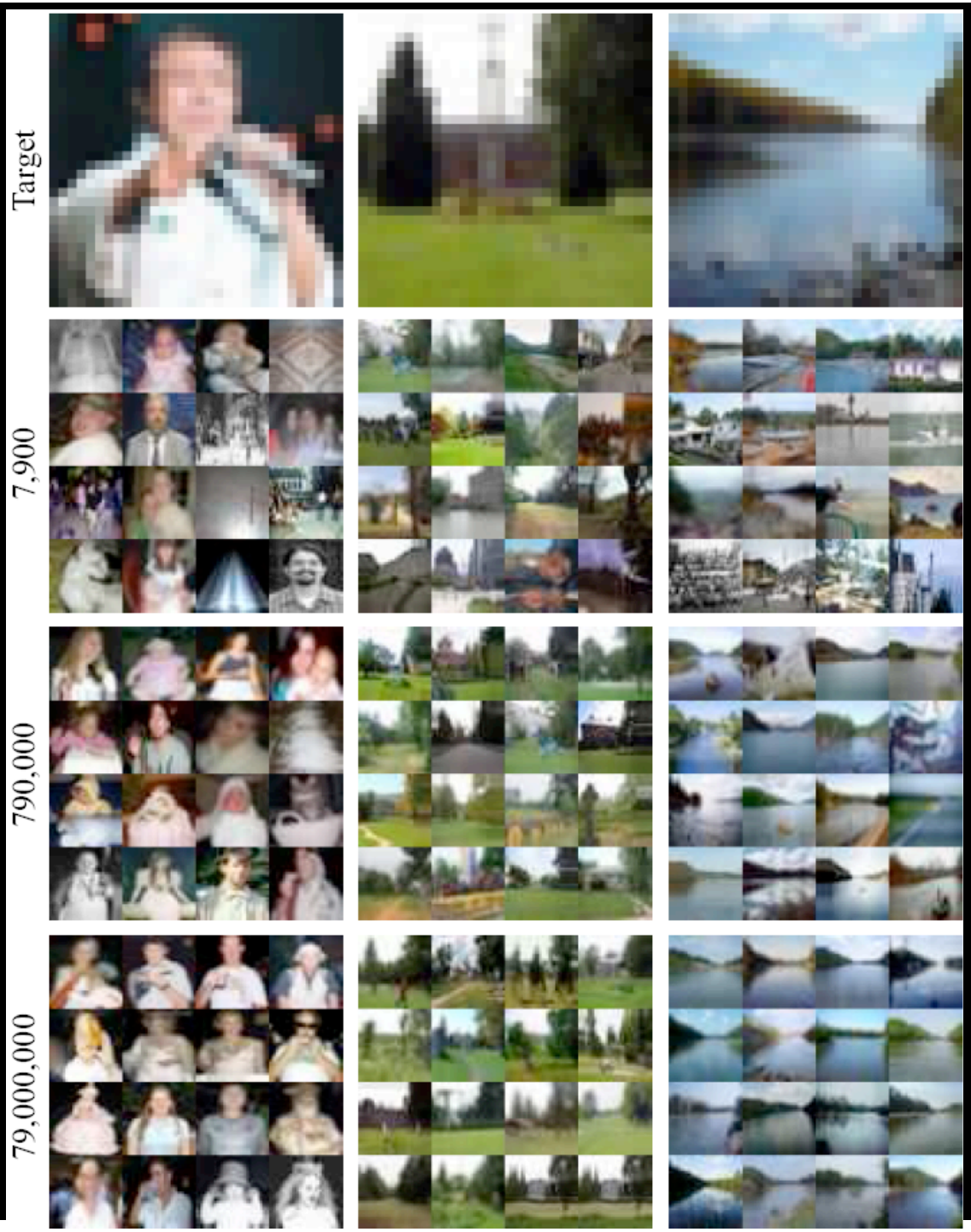
A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

The Power Of Lots Of Images



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

The Power Of Lots Of Images



What can we do with a good similarity metric and **a lot of data**?

Input image

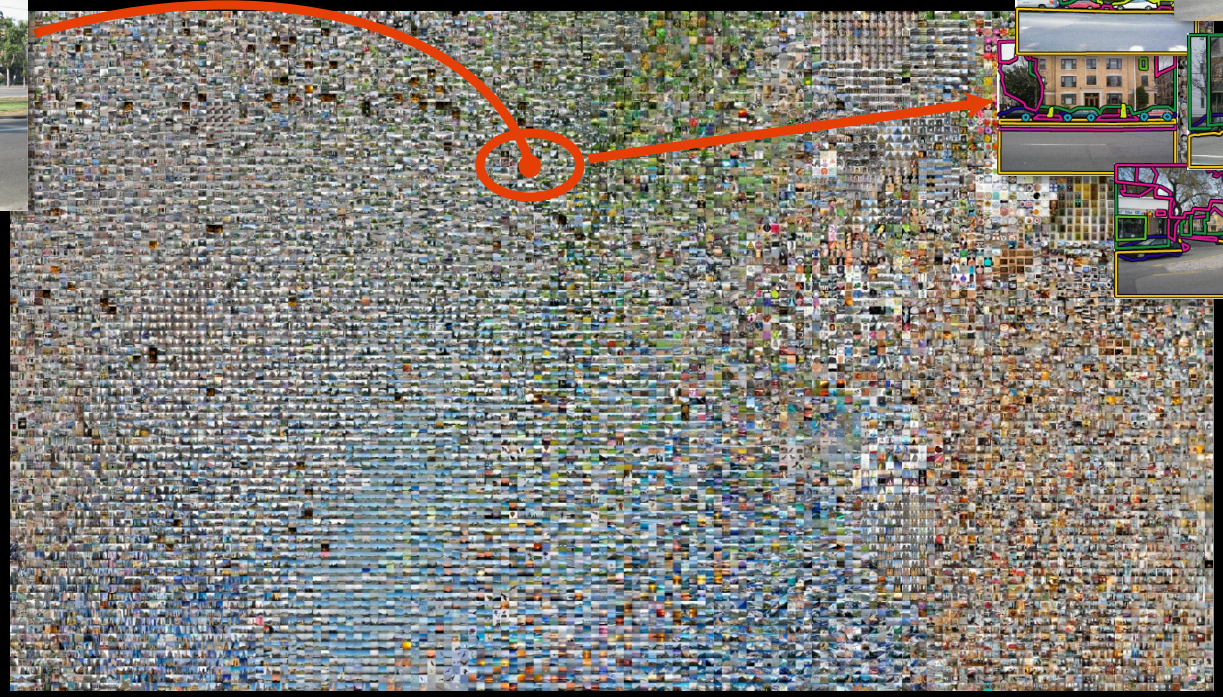


- Labels
- Motion
- Depth
- …

Nearest neighbors



- Labels
- Motion
- Depth
- …



The space of images

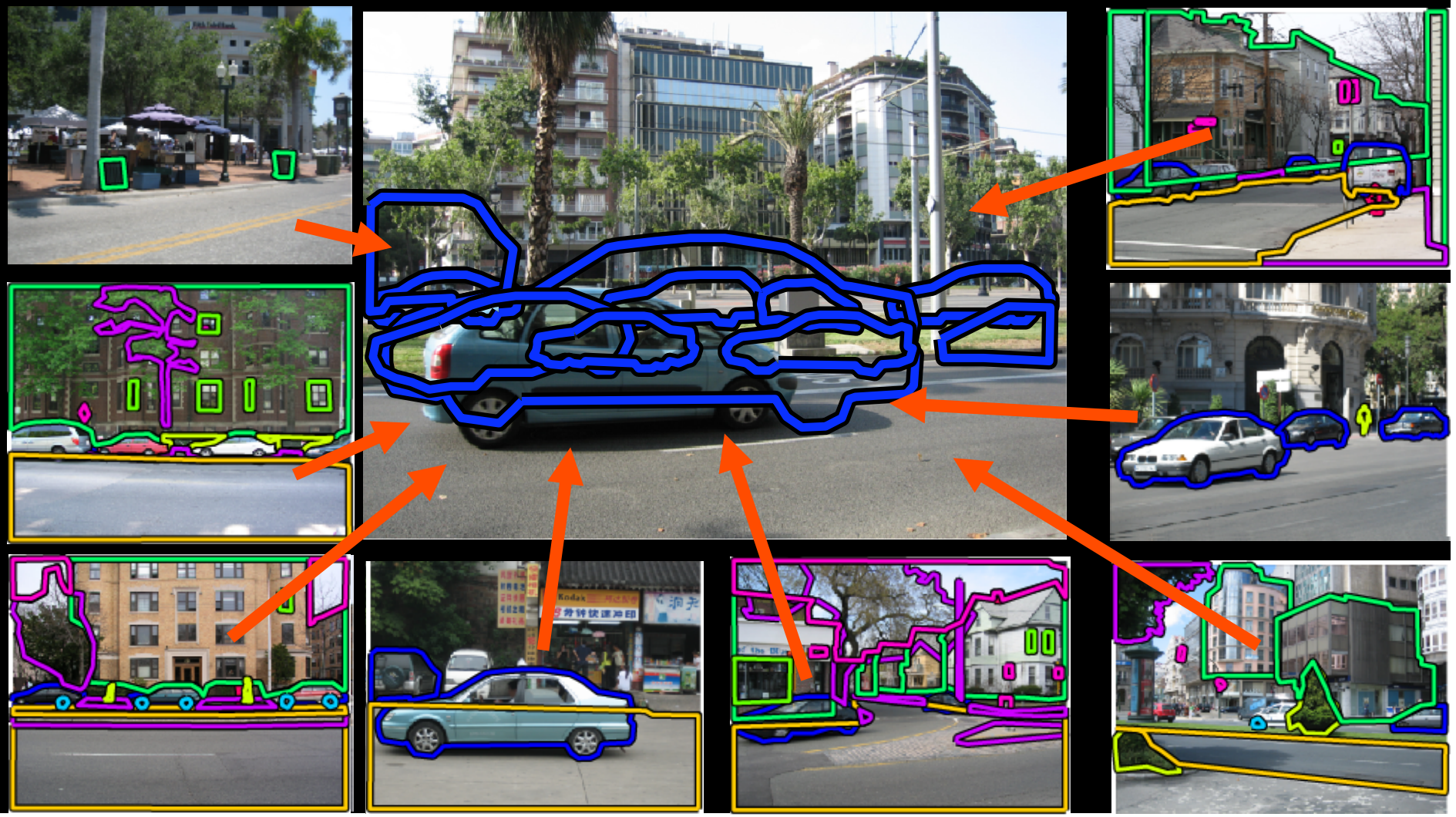
C. Liu, J. Yuen, A. Torralba. CVPR 2009.

With a good image similarity and a lot of data…

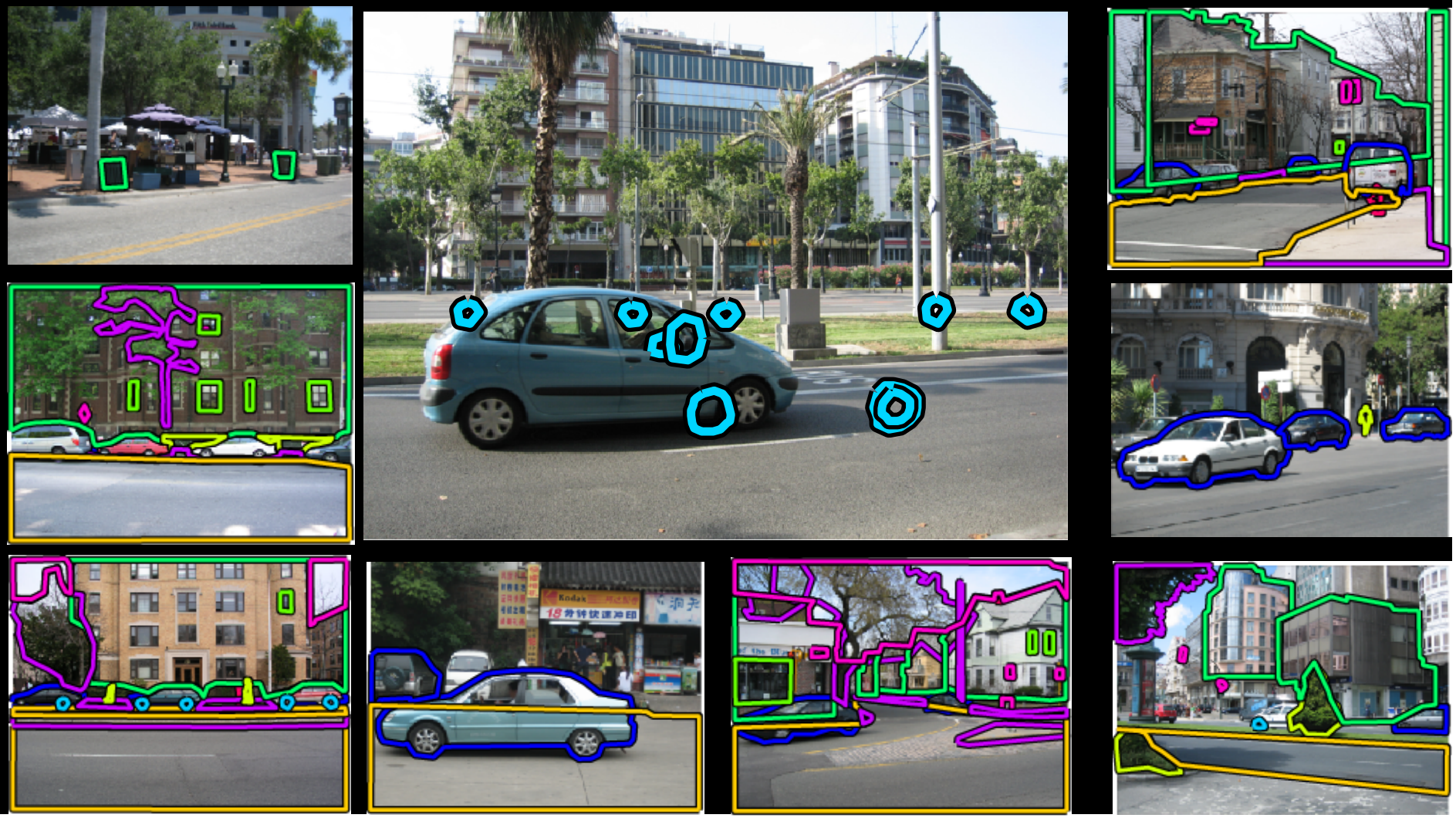


C. Liu, J. Yuen, A. Torralba. CVPR 2009.

With a good image similarity and a lot of data…



With a good image similarity and a lot of data…



With a good image similarity and a lot of data…

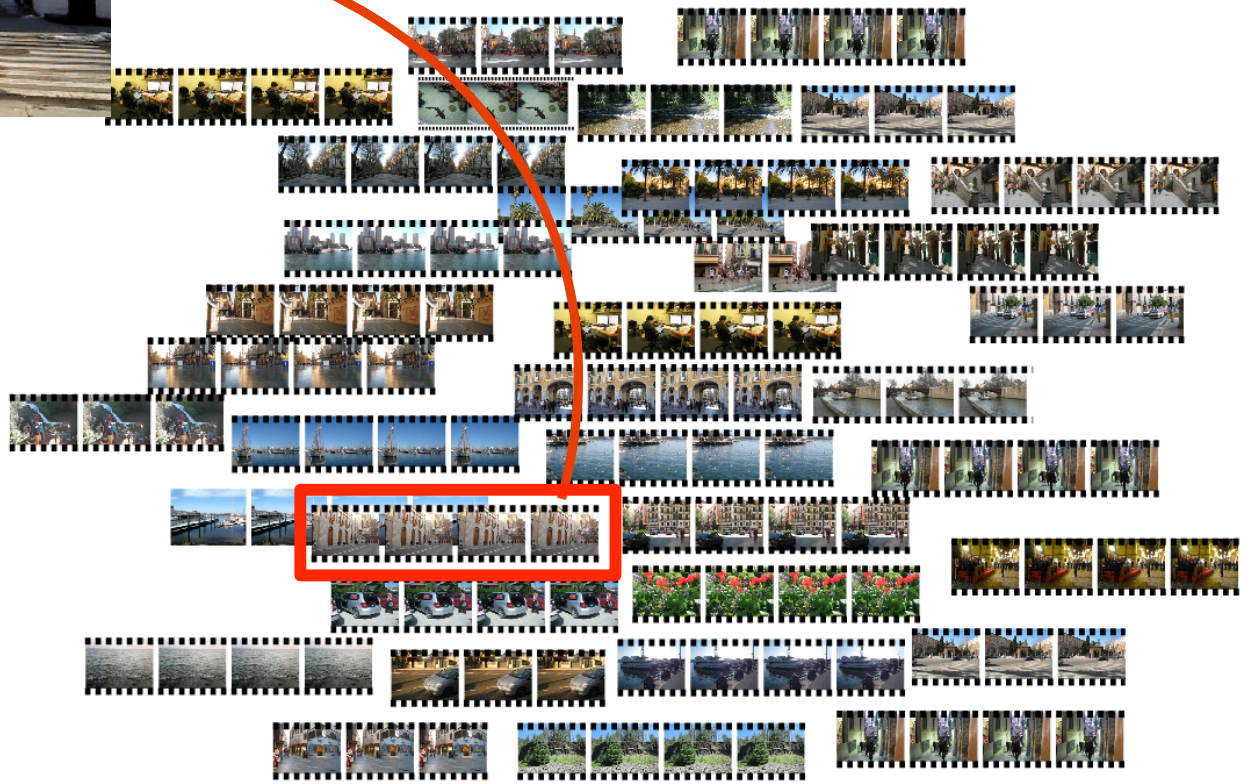


Event prediction

Input image



Video database



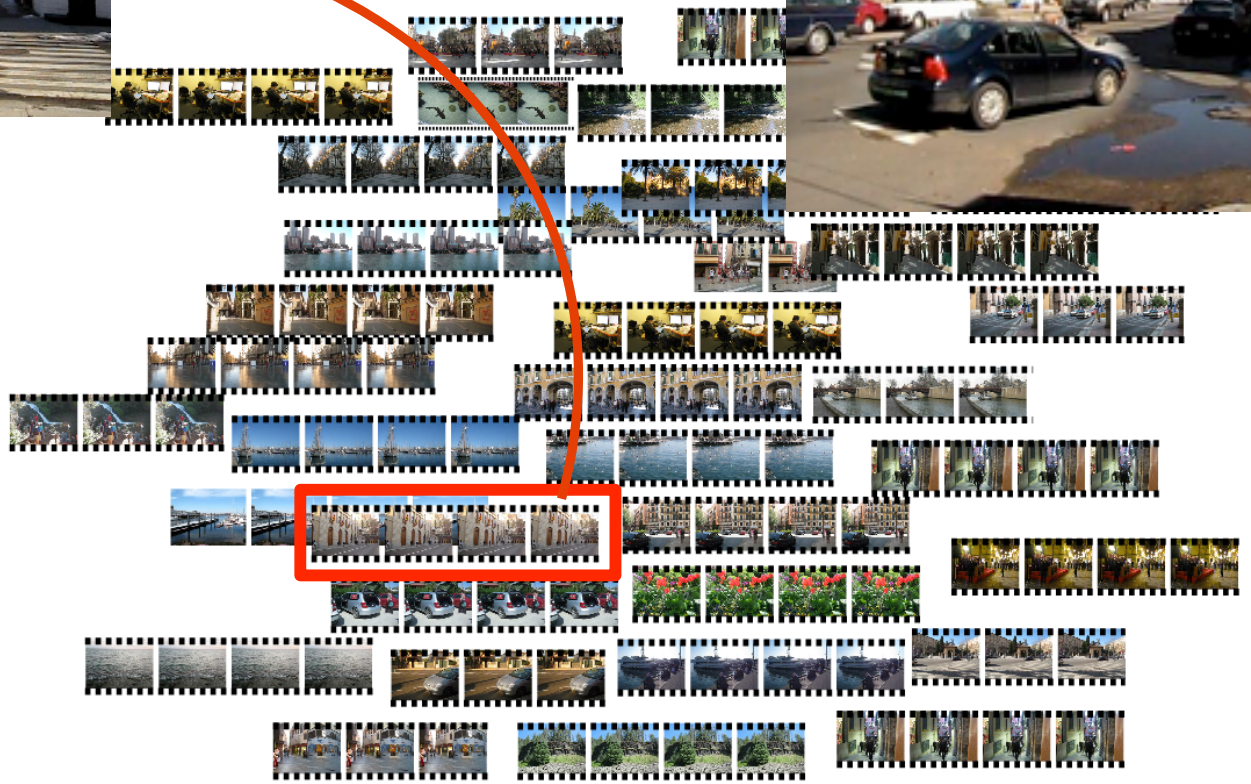
Liu, Yuen, Torralba. CVPR 2009. Yuen, Torralba. ECCV 2010

Event prediction

Input image



Video database





Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010







Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010



Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010





Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010







Liu, Yuen, Torralba. CVPR 2009; Yuen, Torralba. ECCV 2010

A data-driven approach for event prediction

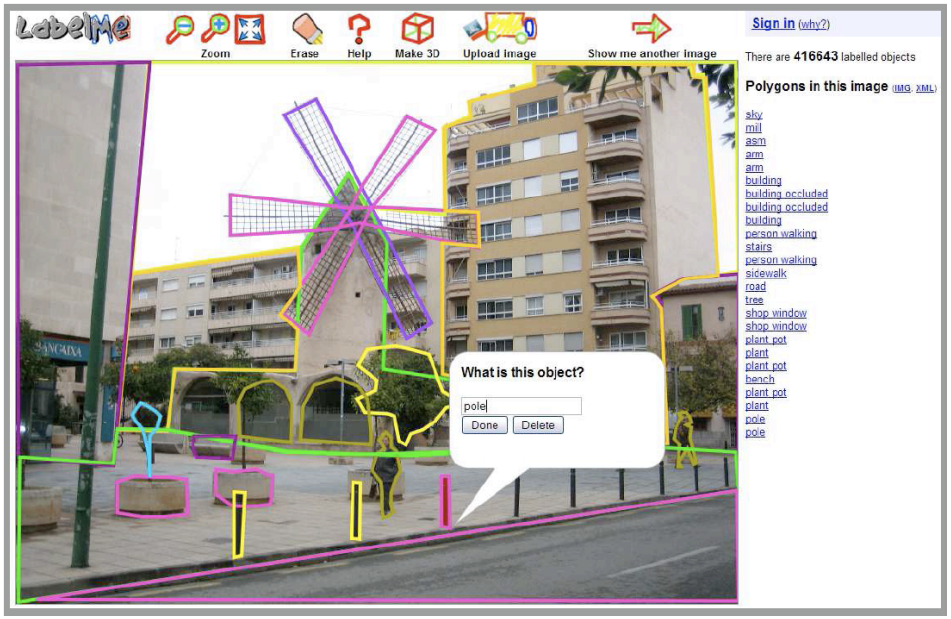
Jenny Yuen and Antonio Torralba



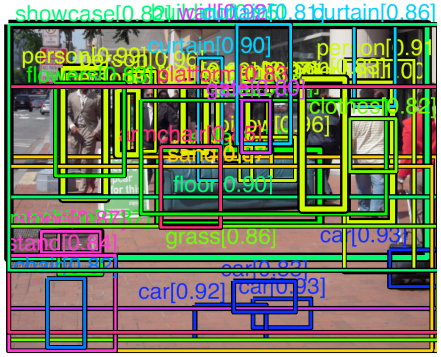
Conclusion

We need large and realistic datasets (images and video)





We need models that reason about relationships between objects, scenes, events, ...







Raw detector output Output context model Most likely outlier

This work is partially funded by NSF Career 0747120

Context

Contextual priming for object detection

A. Torralba

International Journal of Computer Vision, Vol. 53(2), 169-191, 2003.

Context-based vision system for place and object recognition

A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin

IEEE Intl. Conference on Computer Vision (ICCV), Nice, France, October 2003.

Exploiting Hierarchical Context on a Large Database of Object Categories

M. J. Choi, J. Lim, A. Torralba, and A. S. Willsky

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, June 2010.

Using the forest to see the trees: object recognition in context

A. Torralba, K. Murphy, W. T. Freeman

Communications of the ACM, Research Highlights, 53(3): 107-114, 2010.

Large databases

A data-driven approach for event prediction
J. Yuen, A. Torralba
European Conference on Computer Vision (ECCV), 2010.

80 million tiny images: a large dataset for non-parametric object and scene recognition
A. Torralba, R. Fergus, W. T. Freeman
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30(11), pp. 1958-1970, 2008.

SIFT flow: dense correspondence across different scenes
C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman
European Conference on Computer Vision (ECCV), 2008.

LabelMe: a database and web-based tool for image annotation
B. Russell, A. Torralba, K. Murphy, W. T. Freeman
International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, May, 2008.

Connections with human vision

The role of context in object recognition
A. Oliva, A. Torralba
Trends in Cognitive Sciences, vol. 11(12), pp. 520-527. December 2007.

How many pixels make an image?
A. Torralba
Visual Neuroscience, volume 26, issue 01, pp. 123-131, 2009.

Contextual Guidance of Attention in Natural scenes: The role of Global features on object search
A. Torralba, A. Oliva, M. Castelhano and J. M. Henderson
Psychological Review. Vol 113(4) 766-786, Oct, 2006.

To know more

Websites:

Home Page: http://people.csail.mit.edu/torralba

LabelMe: http://labelme.csail.mit.edu/

Visual Dictionary: http://groups.csail.mit.edu/vision/TinyImages/