Scenes and objects

Ivan Laptev and Josef Sivic
http://www.di.ens.fr/~josef
INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

With most slides from: A. Torralba
And also: L. Fei Fei, W. Freeman, D. Hoiem, R. Fergus, A. Gupta, A. Efros

Announcements

- Final project presentations on
  - **Tuesday Dec 11 16:15-19:15 Salle U/V** (the standard class time and location)
  - **Wednesday Dec 12 15:00-18:00 Salle Verte1**

  INRIA, 23 Av. d'Italie, 75013. See the the schedule linked from the class webpage for instructions on how to get there.

- See the schedule linked from:

http://www.di.ens.fr/willow/teaching/recvis12/

- If you need a swap arrange it with someone today in the class and email me.

- **Final project report deadline is on December 23rd.**

MVA Itnernships (MVA stage)

- MVA internships in computer vision at Willow are listed here:

http://www.di.ens.fr/willow/teaching/recvis12/internships/

Talk to us, if you are interested in an internship.

The internship can lead to a PhD in computer vision in willow.

http://www.di.ens.fr/willow/teaching/recvis12/internships/

2013 Computer Vision Internships in the Willow Group

We are looking for strongly motivated candidates with interest in computer vision and applications of machine learning to computer vision problems. Good background in applied mathematics, strong programming skills and prior experience with Matlab are required. The internships can lead to a PhD in the Willow Group.

Proposed internship topics:

1. Large-scale image classification and object detection with Deep Convolutional Neural Networks
2. Predicting actions in places
3. Triangulation de nuages de points

We will assign topics to qualified students in the first-come, first-served basis. To apply, please send us your CV and come to visit us in the lab to discuss the topics.

1. Large-scale image classification and object detection with Deep Convolutional Neural Networks

Project supervisors: Leon Bottou <leon@bottou.org>, Ivan Laptev <Ivan.Laptev@ens.fr> and Josef Sivic <Josef.Sivic@ens.fr>
Location: Willow Group, Laboratoire d'Informatique de l'École Normale Supérieure

How to give a talk and write a paper

Slides by Bill Freeman, MIT:

http://www.di.ens.fr/willow/teaching/recvis12/slides/lecture23TalksAndPapers.pdf

Lecture notes by Bill Freeman, MIT:

http://www.di.ens.fr/willow/teaching/recvis12/slides/slideNotes23TalksPapers.pdf

Other sources:

http://www.cs.berkeley.edu/~messer/Bad_talk.html

http://www-psych.stanford.edu/~lera/talk.html

High order bit: prepare

- Practice by yourself.
- Give practice versions to your friends.
- Think through your talk.
- You can write out verbatim what you want to say in the difficult parts.
- Ahead of time, visit where you'll be giving the talk and identify any issues that may come up.
- Preparation is a great cure for nervousness.

http://www.di.ens.fr/willow/teaching/recvis12/slides/lecture23TalksAndPapers.pdf

Slides by W. Freeman

Some bad news…

The more you work on a talk, the better it gets: if you work on it for 3 hours, the talk you give will be better than if you had only worked on it for 2 hours. If you work on it for 5 hours, it will be better still. 7 hours, better yet…

Slides by A. Torralba

All talks are important

There are no unimportant talks.

There are no big or small audiences.

Prepare each talk with the same enthusiasm.

Slides by A. Torralba

How to give a talk

Delivering:

Look at the audience! Try not to talk to your laptop or to the screen. Instead, look at the other humans in the room.

You have to believe in what you present, be confident… even if it only lasts for the time of your presentation.

Do not be afraid to acknowledge limitations of whatever you are presenting. Limitations are good. They leave job for the people to come. Trying to hide the problems in your work will make the preparation of the talk a lot harder and your self confidence will be hurt.

The different kinds of talks you'll have to give as a researcher

- 2-5 minute talks
- 20 -30 minute conference presentations
- 30-60 minute colloquia

Very short talks

- Rehearse it.
- Cut things out that aren't essential. You can refer to them at a high level.
- You might focus on answering just a few questions, eg: what is the problem? Why is it interesting? Why is it hard?
- Typically these talks are just little advertisements for a poster or for some other (longer) talk. So you just need to show people that the problem is interesting and that you're fun to talk with.
- These talks can convey important info--note popularity of SIGGRAPH fast forward session.

http://groups.csail.mit.edu/vision/courses/6.869/lectures/lecture23TalksAndPapers.pdf

Slides by W. Freeman

In your talk try answering the following questions

- What problem did you address?
- Why is it interesting?
- Why is it hard?
- What was the key to your approach?
- How well did it work?

http://www.di.ens.fr/willow/teaching/recvis12/slides/lecture23TalksAndPapers.pdf

Slides by W. Freeman

See more at:

Writing papers and giving talks

Bill Freeman
MIT CSAIL
May 2, 2011

http://www.di.ens.fr/willow/teaching/recvis12/slides/lecture23TalksAndPapers.pdf

Sources on writing technical papers

- How to Get Your SIGGRAPH Paper Rejected, Jim Kajiya, SIGGRAPH 1993 Papers Chair, http://www.siggraph.org/publications/instructions/rejected.html
- Ted Adelson's Informal guidelines for writing a paper, 1991. http://www.ai.mit.edu/courses/6.899/papers/ted.htm
- Notes on technical writing, Don Knuth, 1989.

  http://www.ai.mit.edu/courses/6.899/papers/knuthAll.pdf

- What's wrong with these equations, David Mermin, Physics Today, Oct., 1989. http://www.ai.mit.edu/courses/6.899/papers/mermin.pdf
- Ten Simple Rules for Mathematical Writing, Dimitri P. Bertsekas http://www.mit.edu:8001/people/dimitrib/Ten_Rules.html

Slides by A. Torralba

Today: Scenes and objects

1. Scenes as textures (without modeling objects and their relations)

2. Objects within a scene

3. Recognizing multiple objects in an image.

What is a scene?

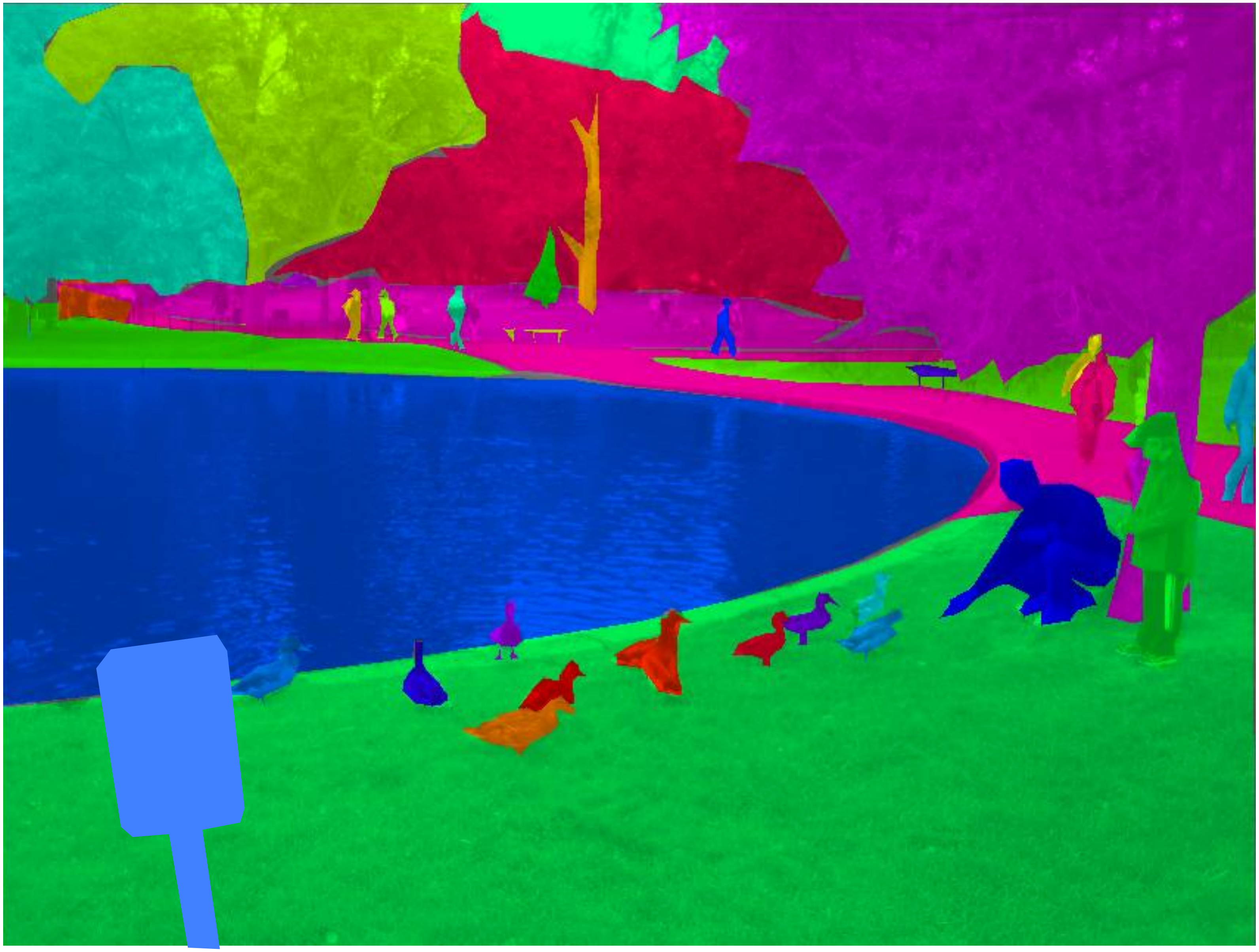
The texture

The object

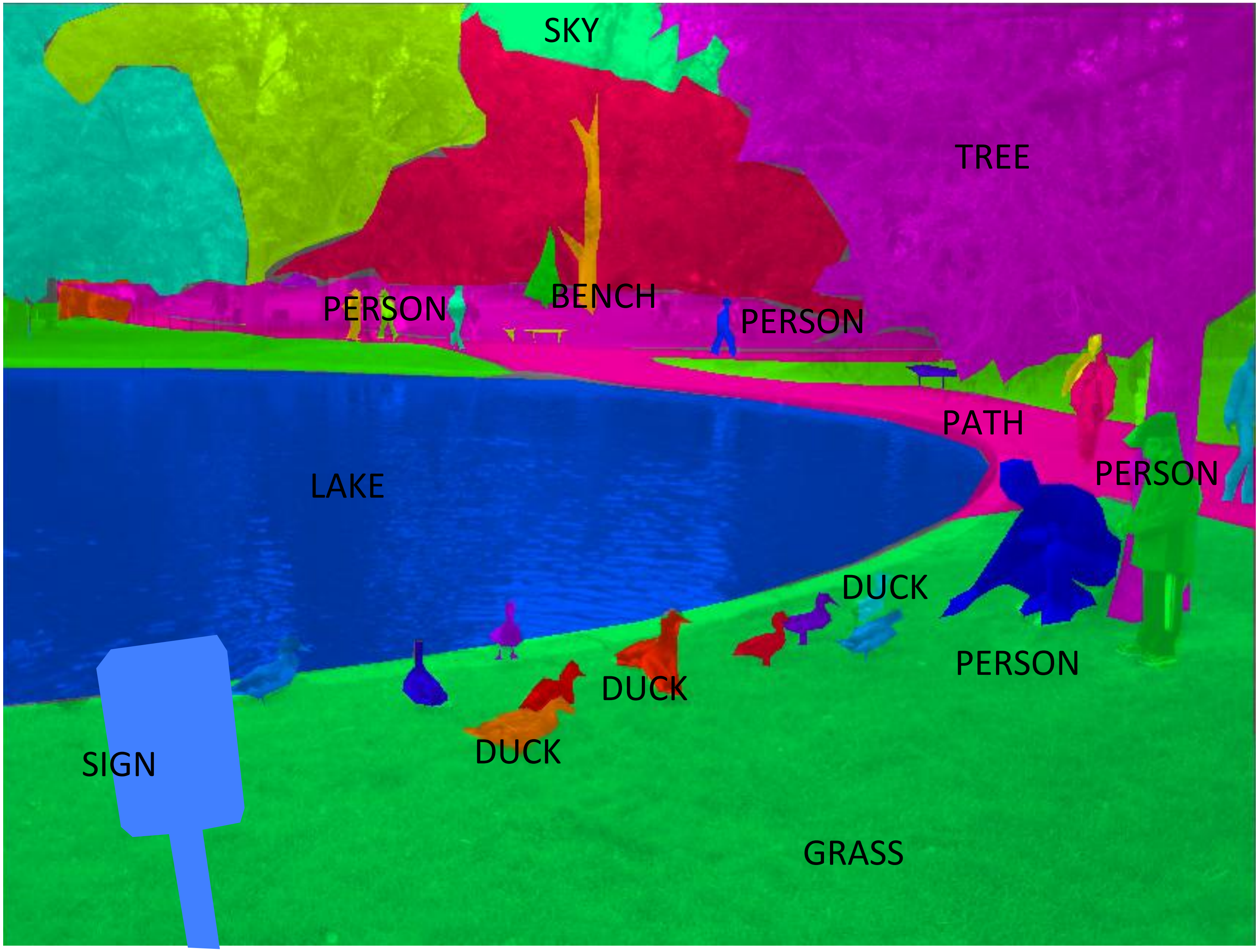
The scene

Slides by A. Torralba

Slides by A. Torralba

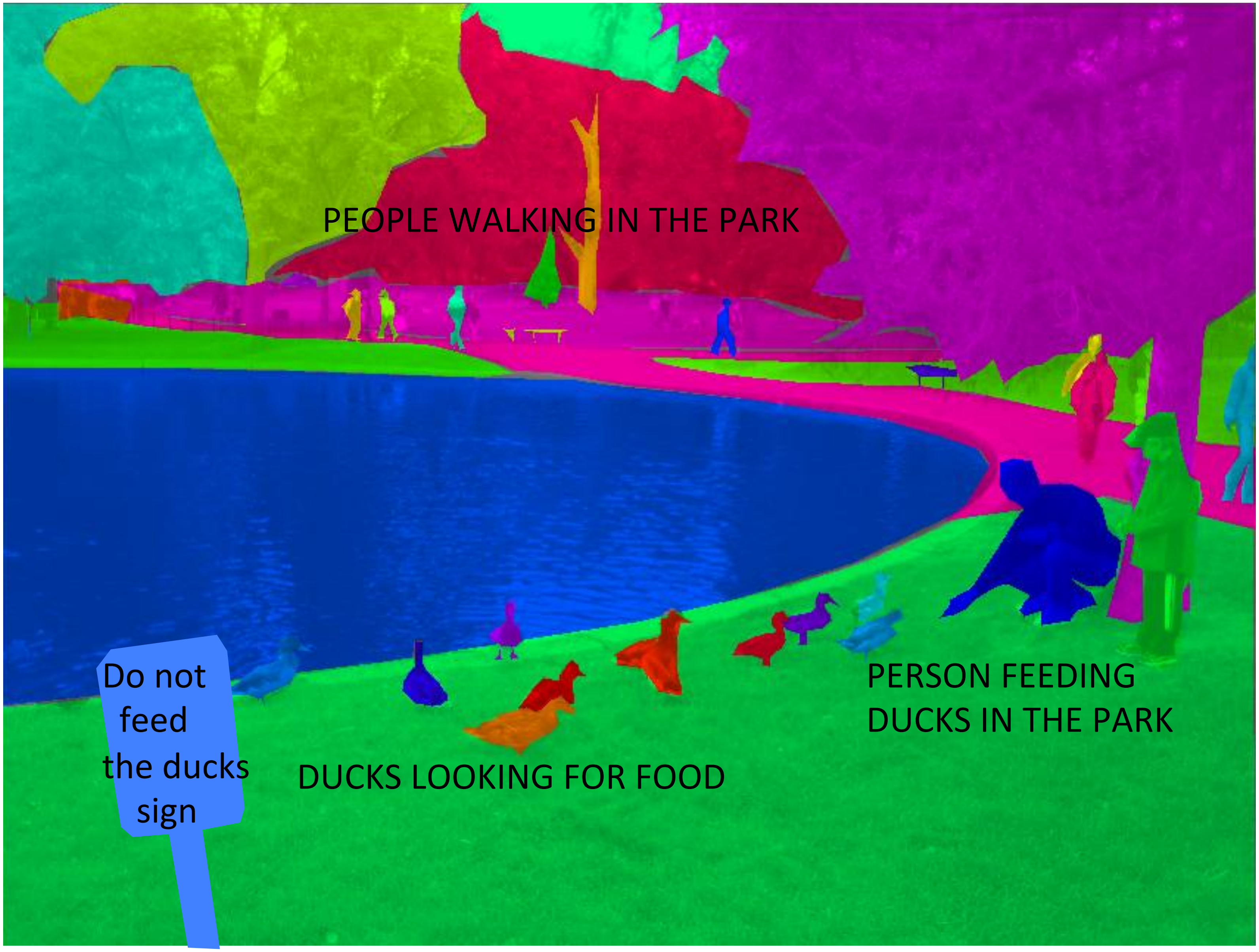


Slides by A. Torralba



Slides by A. Torralba

Slides by A. Torralba



Slides by A. Torralba

Slides by A. Torralba

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." – A. Torralba

Slides by A. Torralba

Scene views vs. objects

A photograph of a firehydrant

A photograph of a street

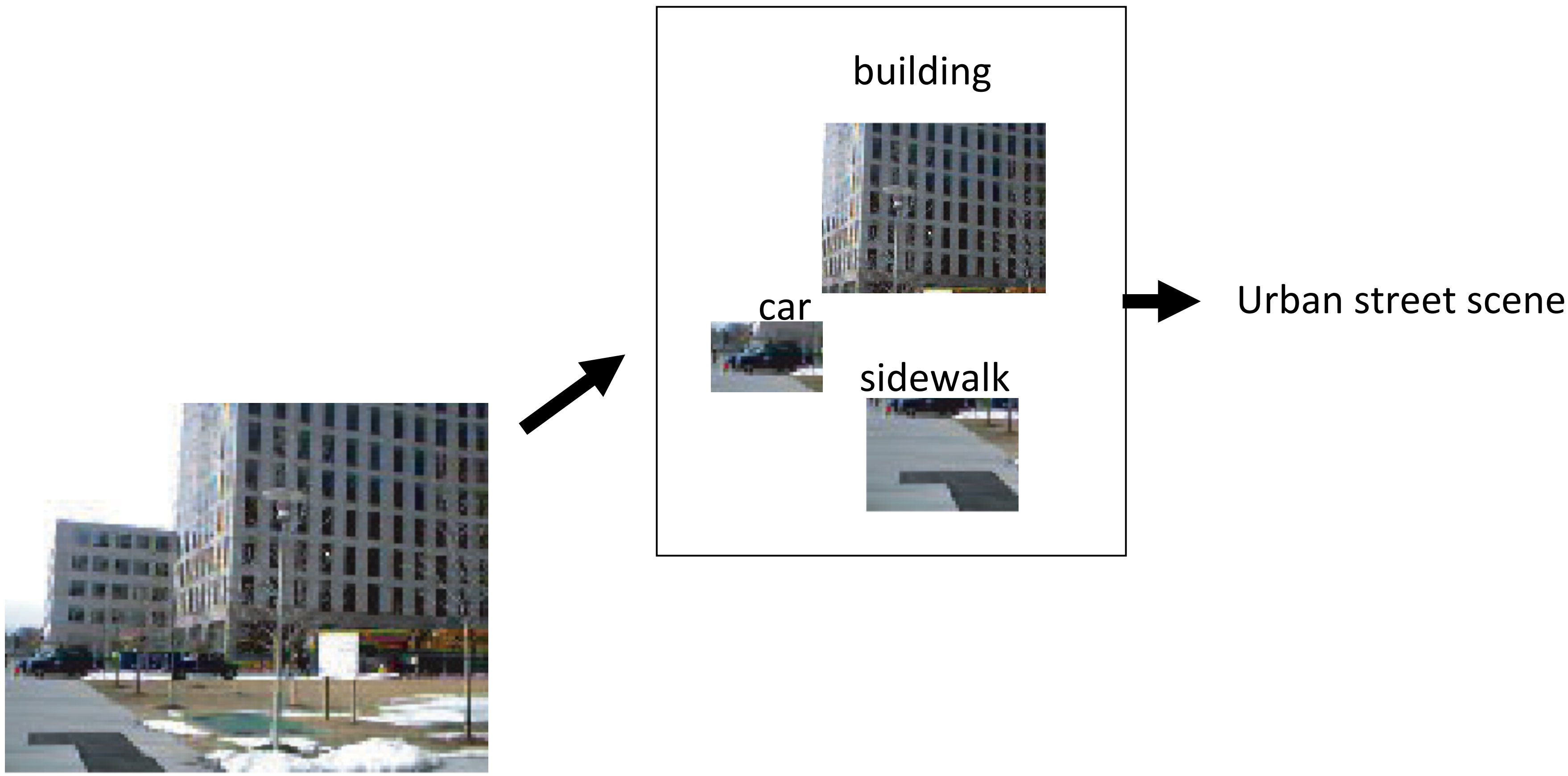
Slides by A. Torralba

Part I: Scenes as textures

(No explicit modeling of objects and their relations)

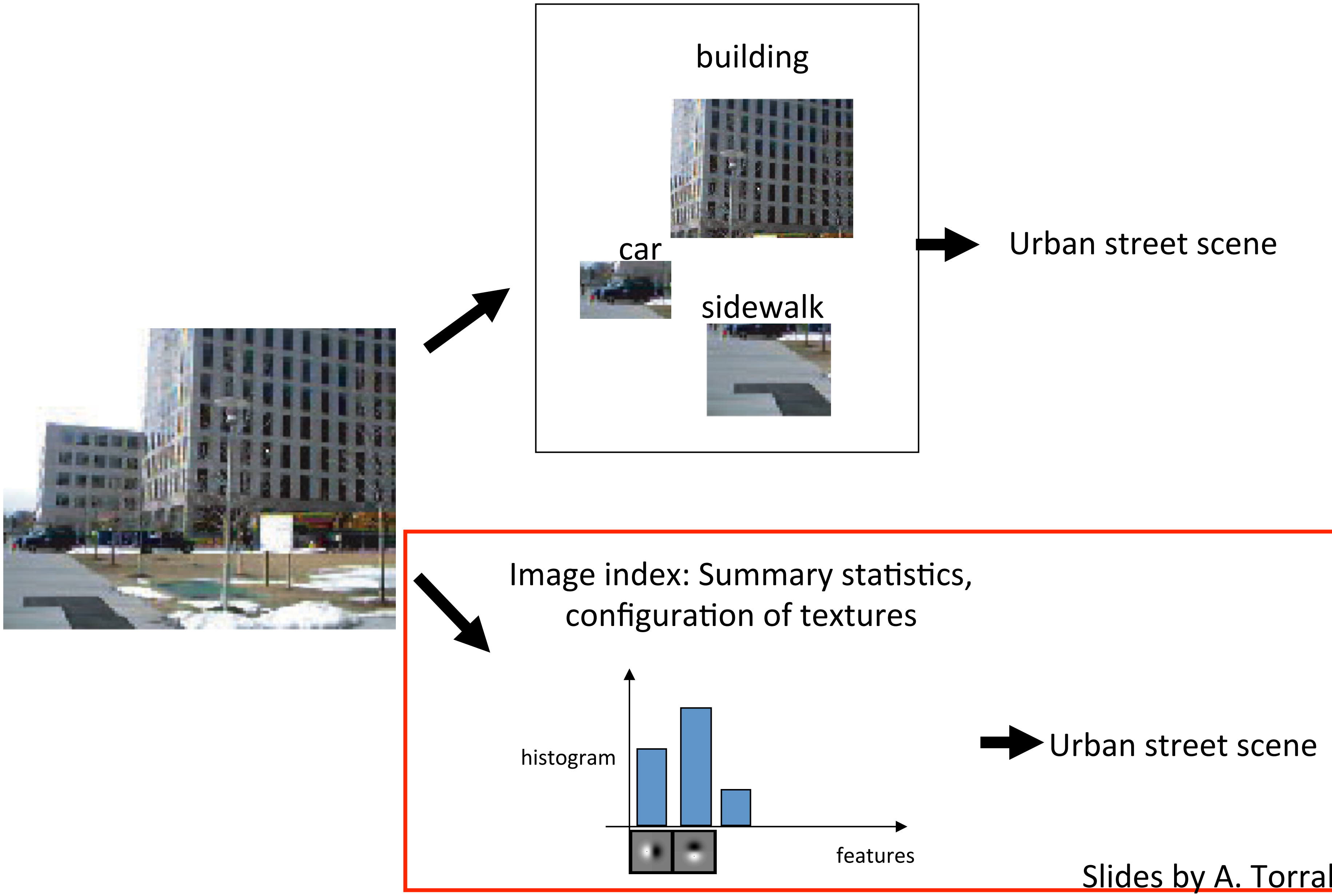
Slides by A. Torralba

Global and local representations



Slides by A. Torralba

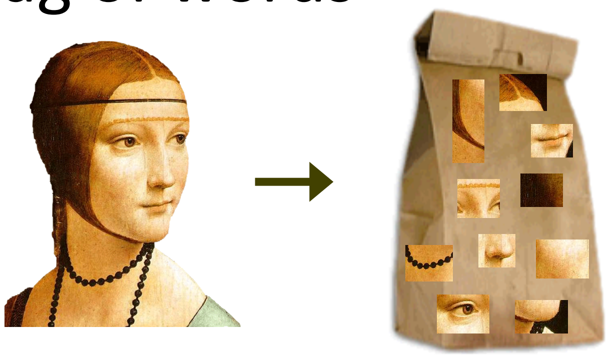
Global and local representations



Slides by A. Torralba

Global scene representations

Bag of words



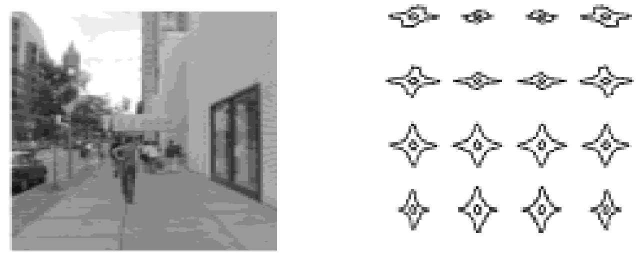
Sivic et. al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

Non localized textons

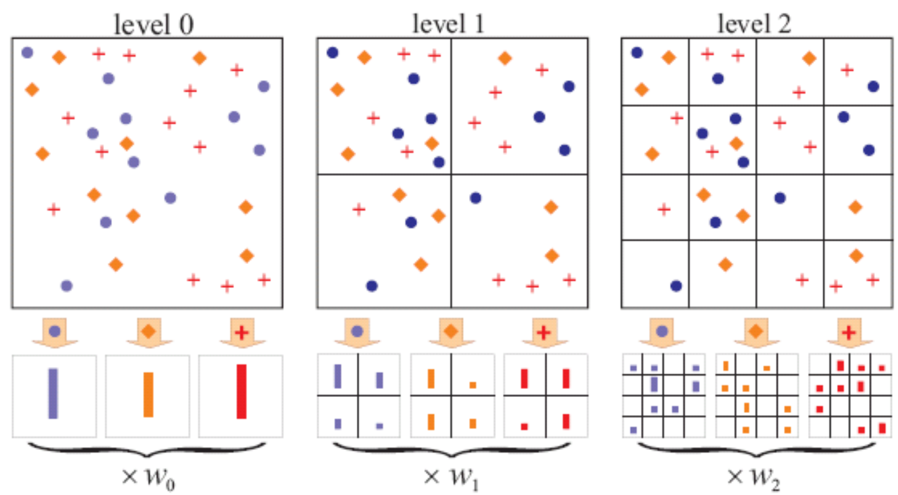
Walker, Malik. Vision Research 2004

…

Spatially organized textures



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



S. Lazebnik, et al, CVPR 2006

…

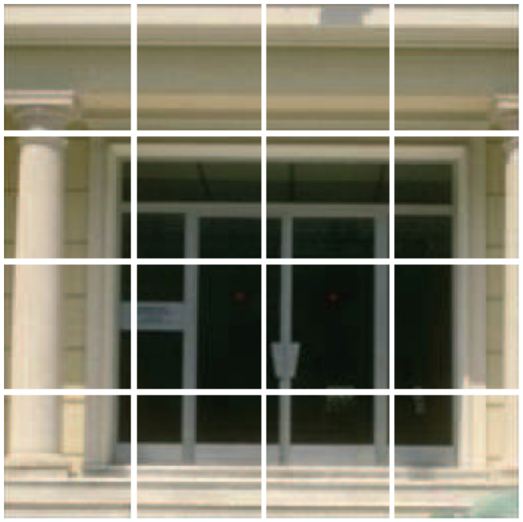
Spatial structure is important in order to provide context for object localization

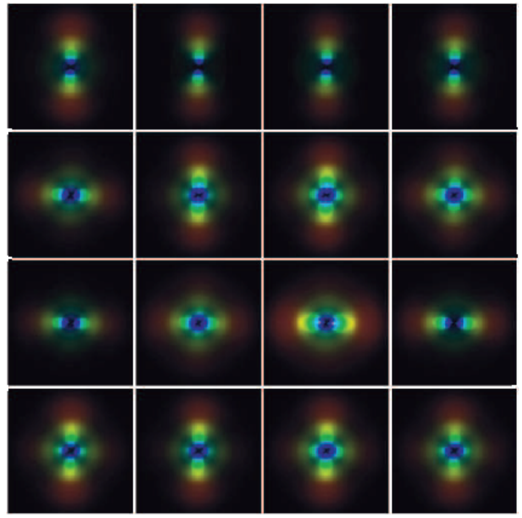
R. Datta, D. Joshi, J. Li, and J. Z. Wang, **Image Retrieval: Ideas, Influences, and Trends of the New Age**, *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1-60, 2008.

Slides by A. Torralba

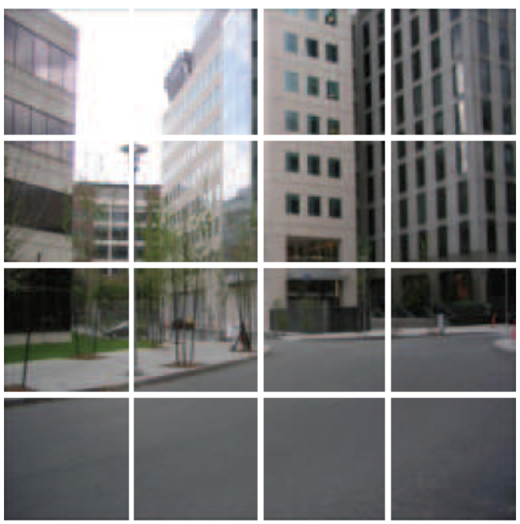
Gist descriptor

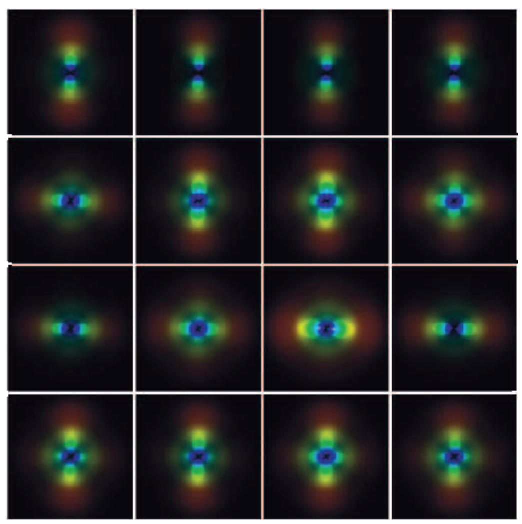
Oliva and Torralba, 2001





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





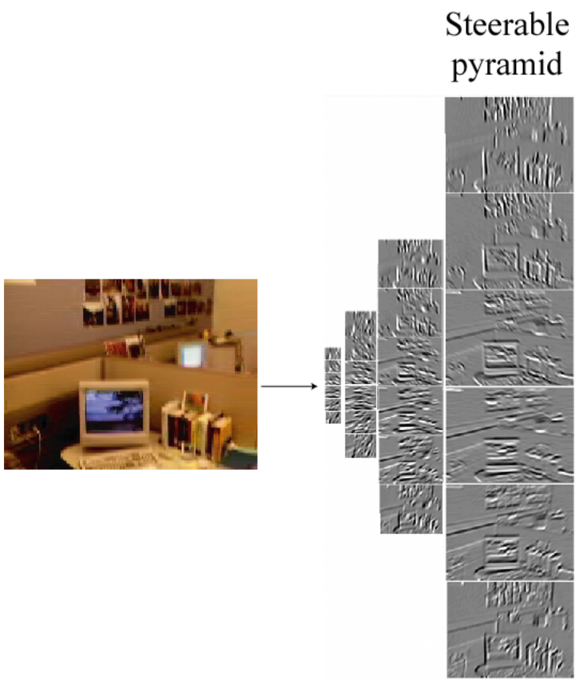
8	orientations
4	scales
x 16	bins
512	dimensions

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; …

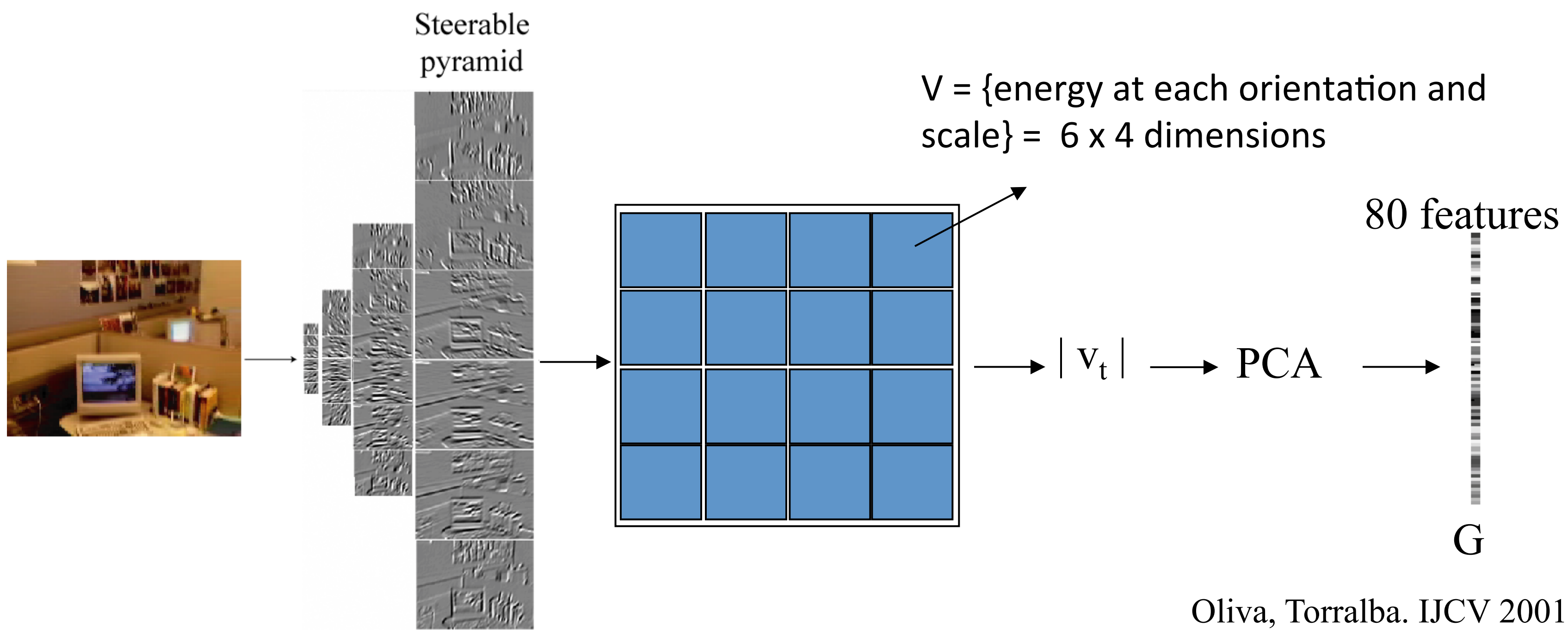
Slides by A. Torralba

Gist descriptor



Slides by A. Torralba

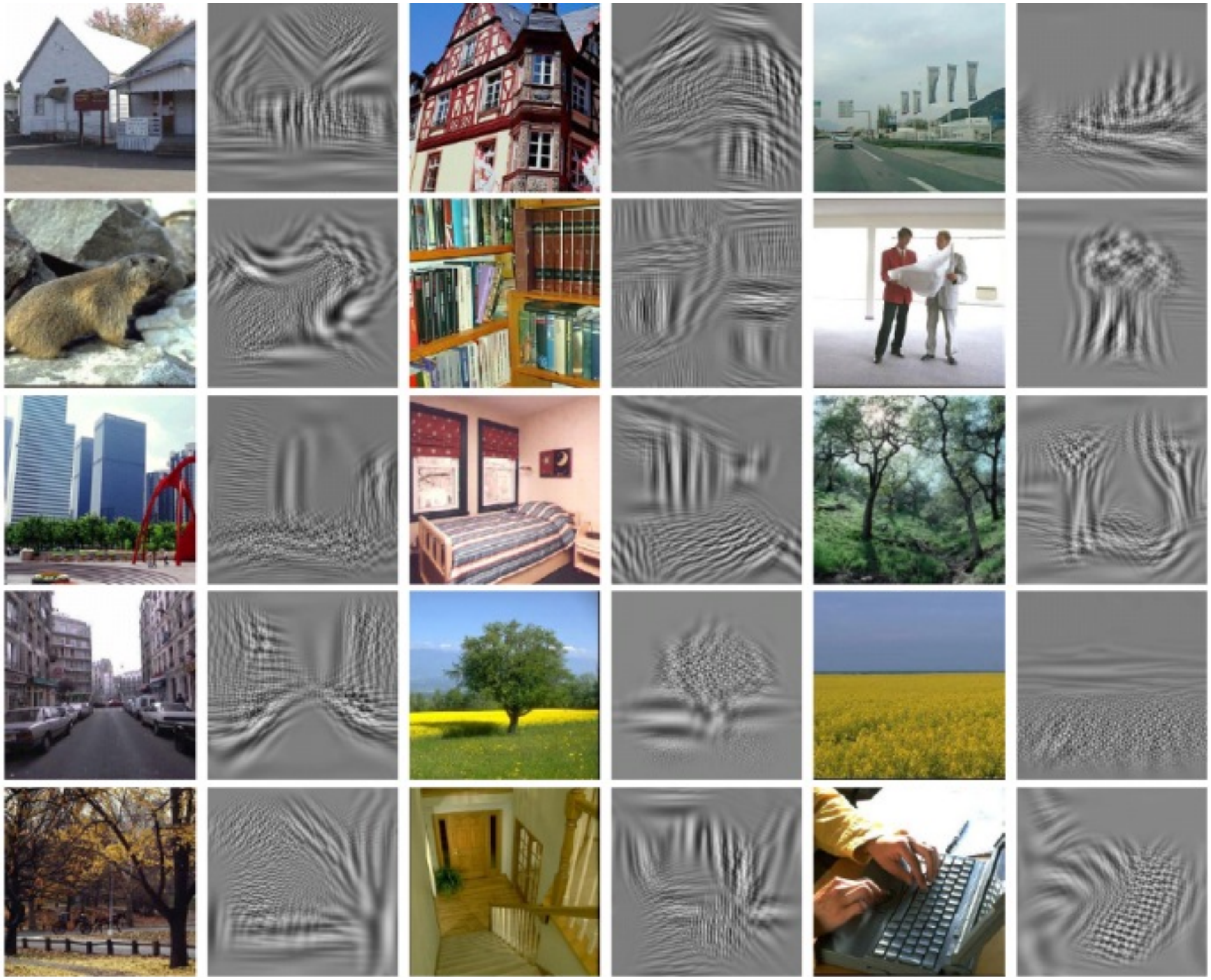
Gist descriptor



Oliva, Torralba. IJCV 2001

Slides by A. Torralba

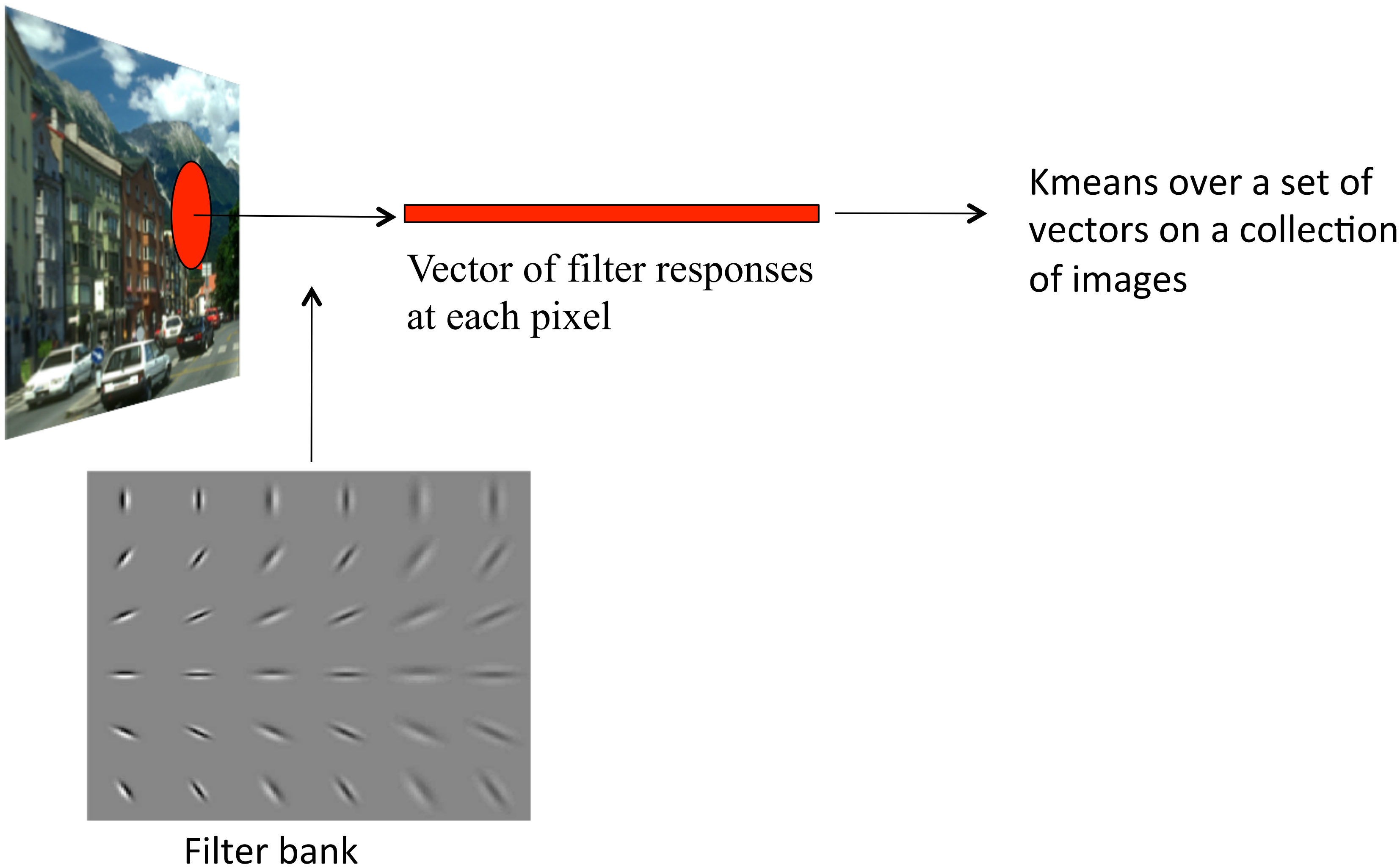
Example visual gists



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

Oliva & Torralba (2001)

Textons (review)

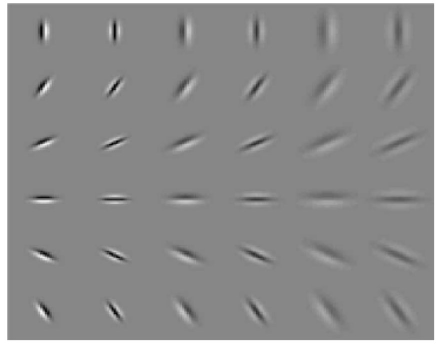


Slides by A. Torralba

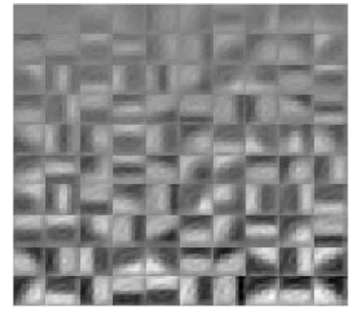
Malik, Belongie, Shi, Leung, 1999

Textons (review)

Filter bank

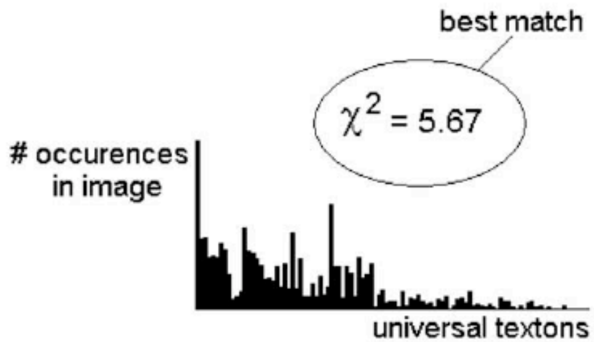


K-means (100 clusters)

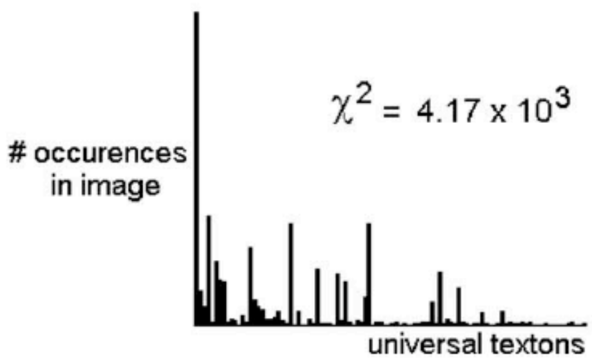


Malik, Belongie, Shi, Leung, 1999

label = bedroom



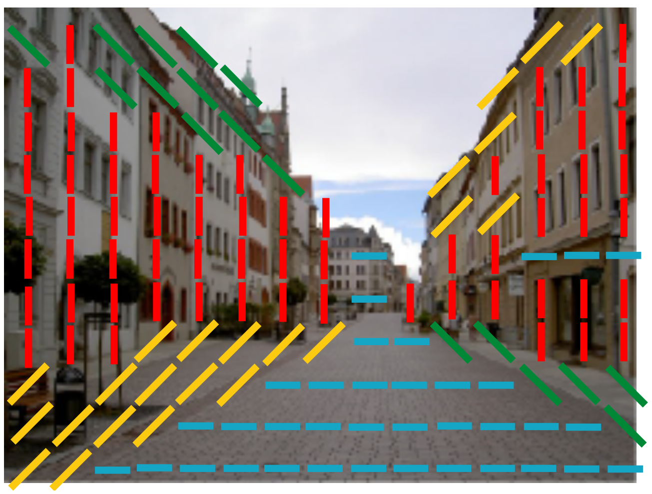
label = beach

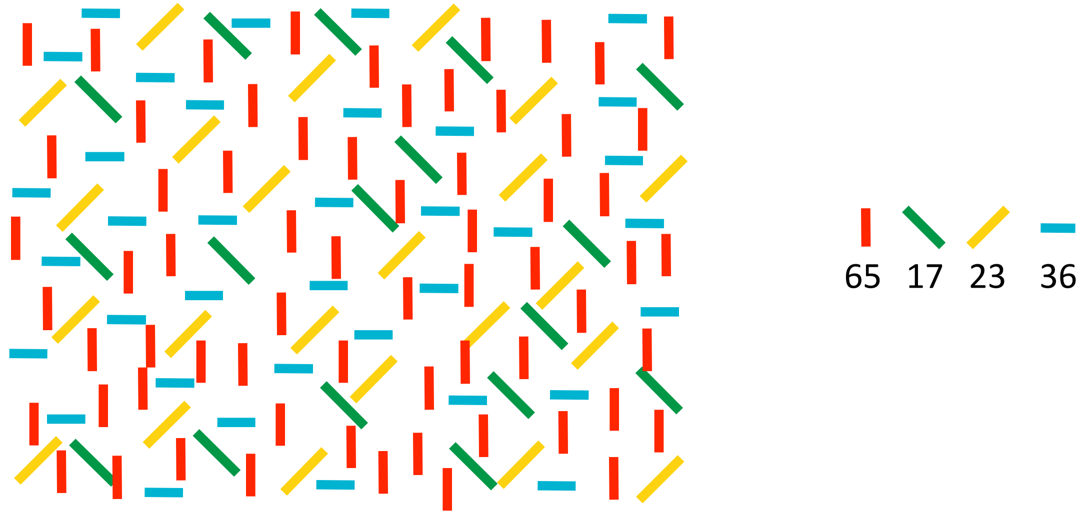


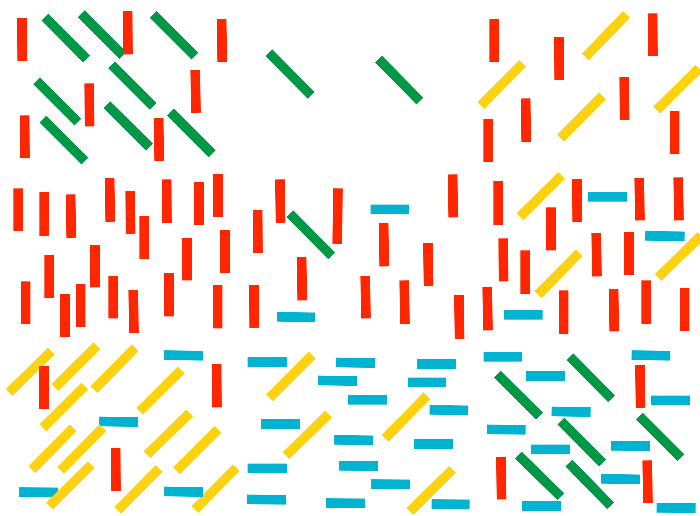
Walker, Malik, 2004

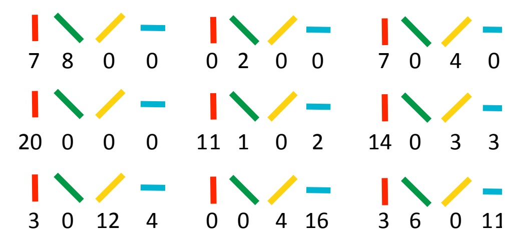
Slides by A. Torralba

Bag of words (review)





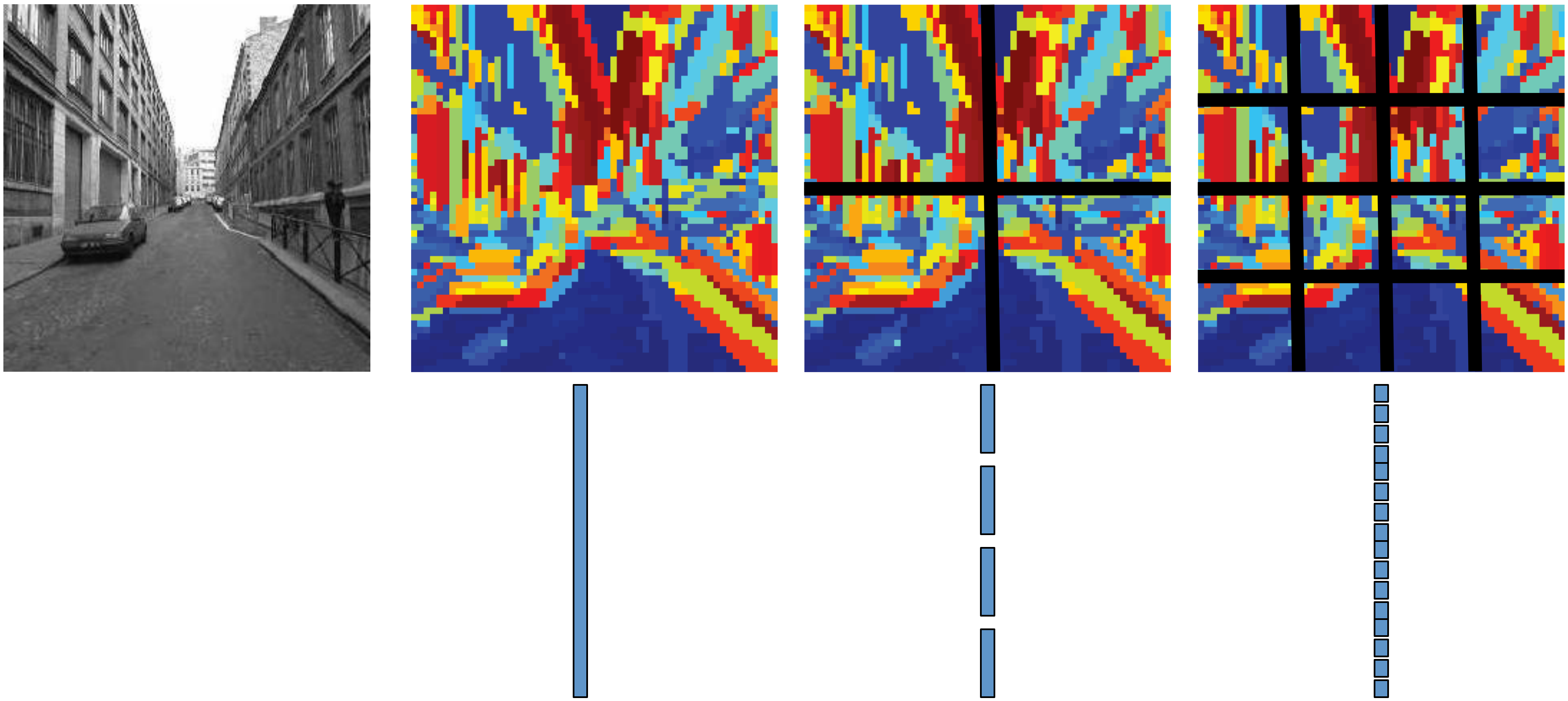




Slides by A. Torralba

Bag of words & spatial pyramid matching (review)

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



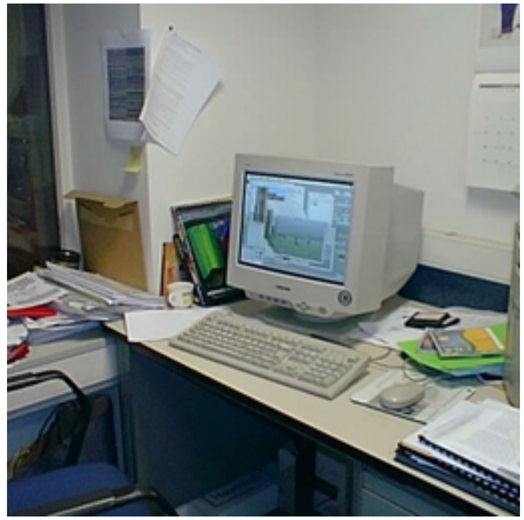
Slides by A. Torralba

S. Lazebnik, et al, CVPR 2006

Scene categorization

Can we use this representation to categorize scenes?

The 15-scenes benchmark



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

Office

Skyscrapers

Suburb

Building facade

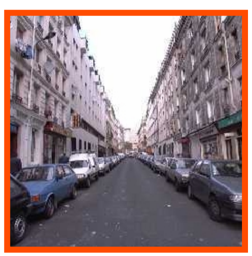
Coast

Forest

Bedroom

Living room

Industrial



Street

Highway

Mountain

Open country

Kitchen

Store

Slides by A. Torralba

SVM (review)

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: $k(x,x_m) = \mathrm{sum}_i(\min(x(i), x_m(i)))$

Slides by A. Torralba

Scene recognition

100 training samples per class

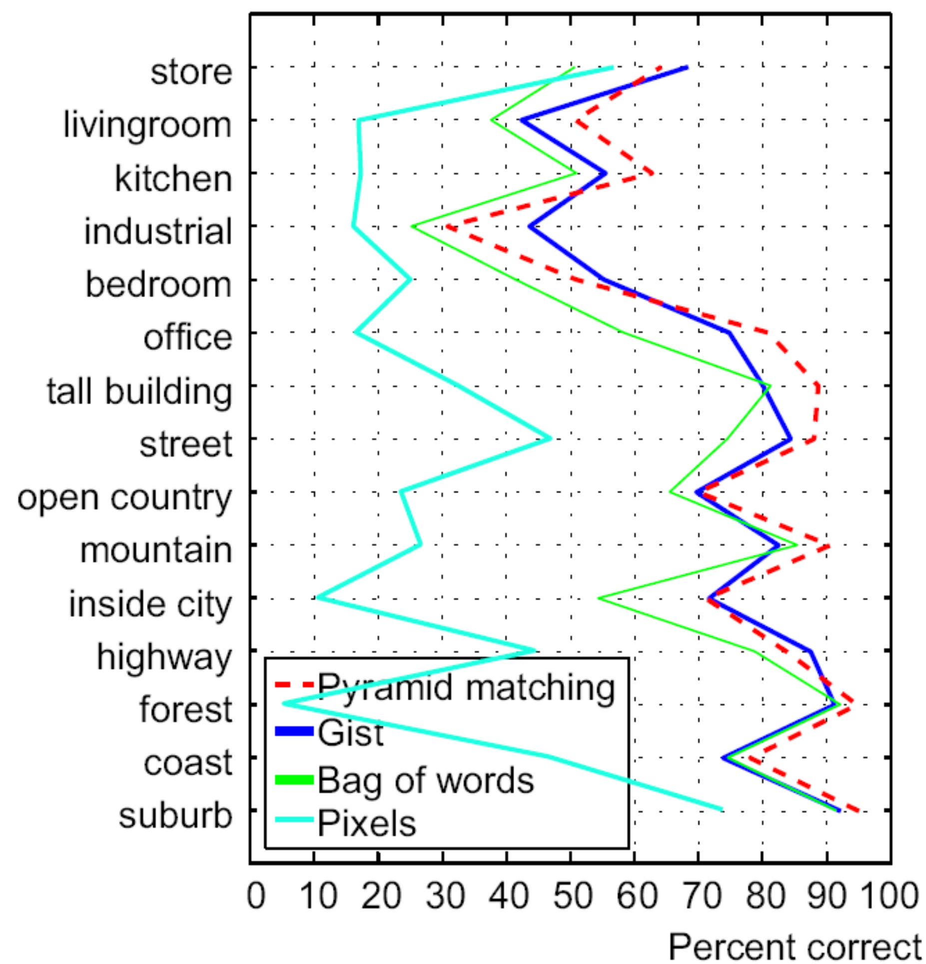
SVM classifier in all cases

Pixels: Gaussian kernel

Gist: Gaussian kernel

Bag of words: Histogram intersection

Pyr: Pyramid matching kernel

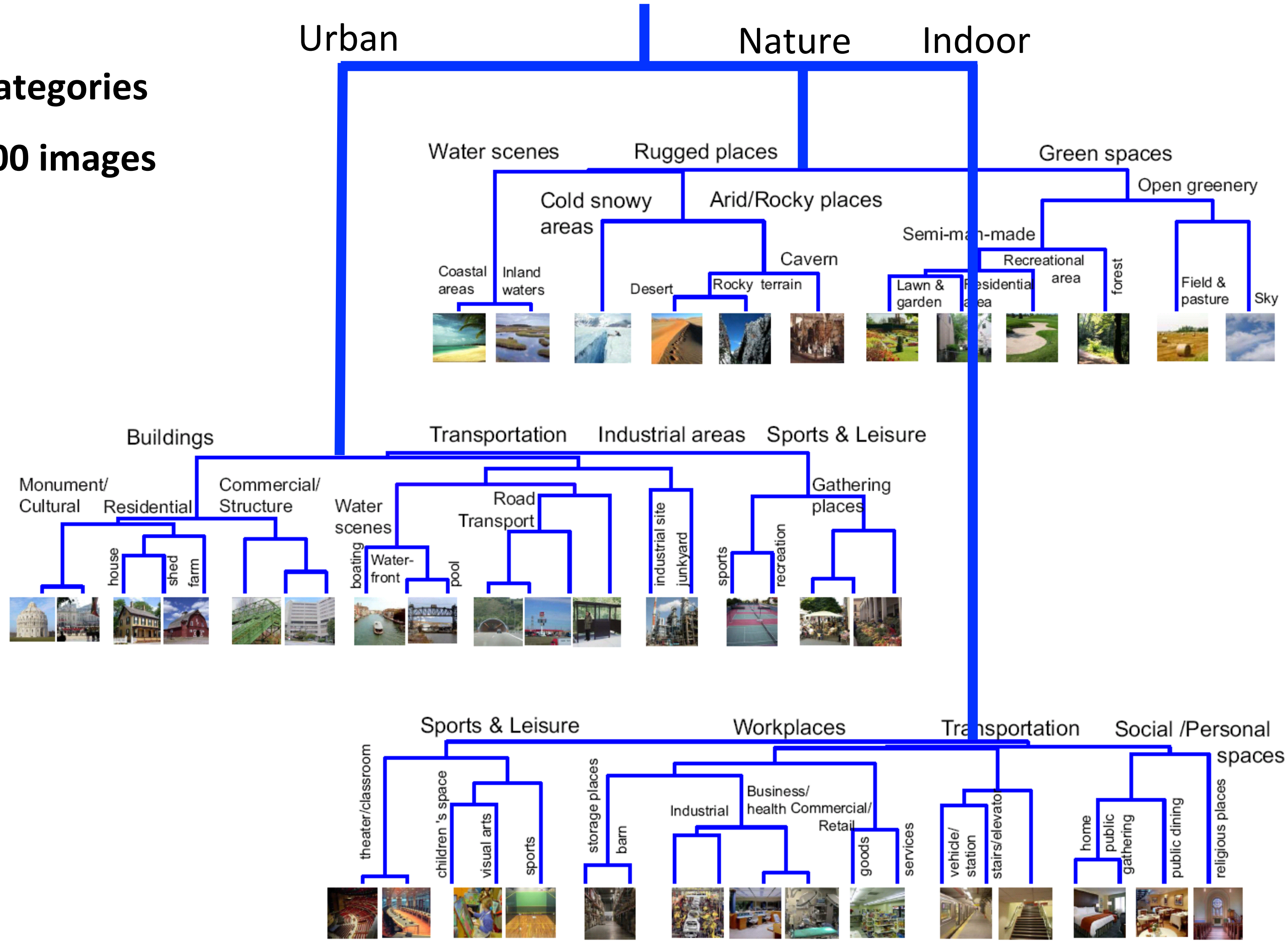


Slides by A. Torralba

Large Scale Scene Recognition

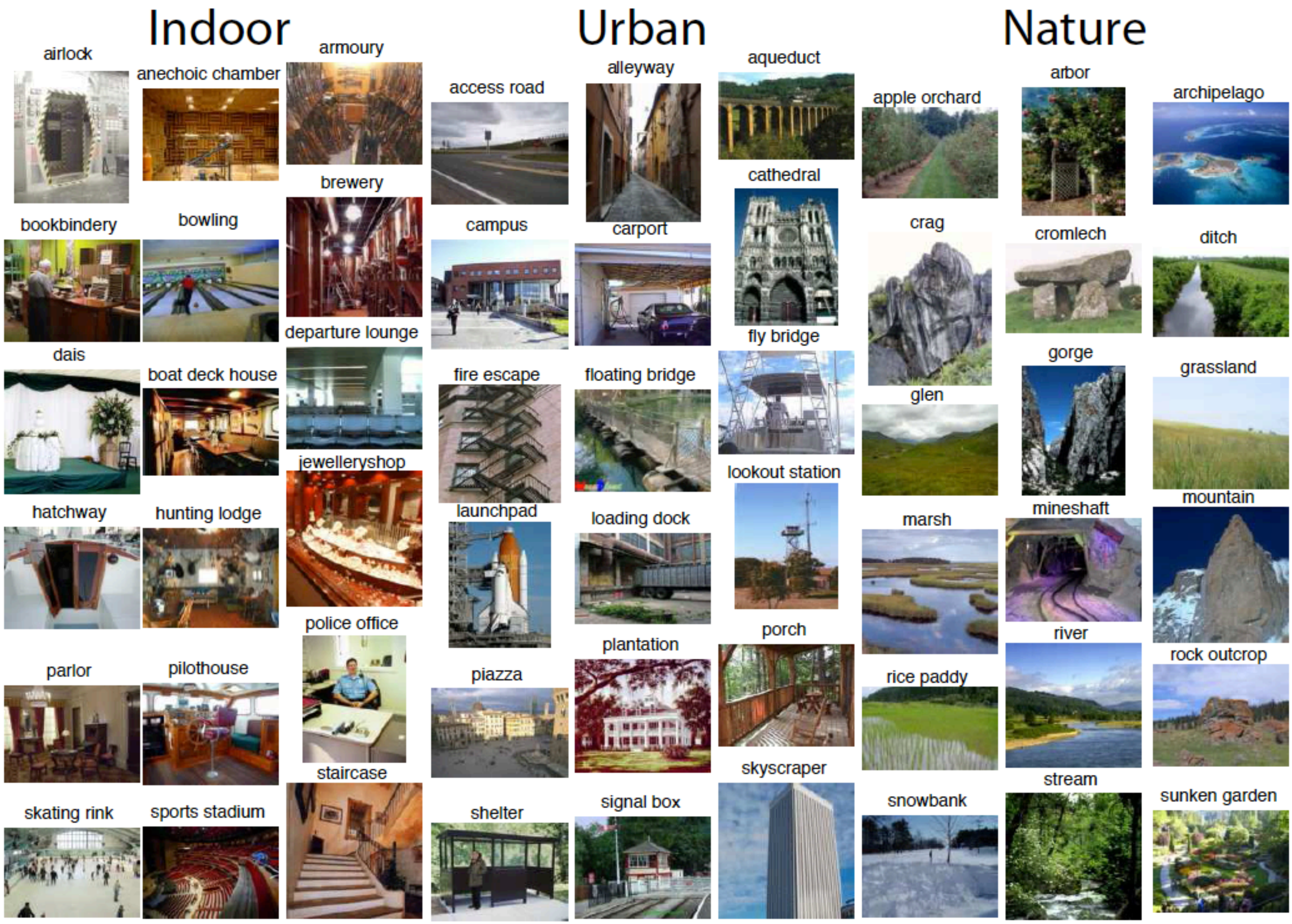
> 400 categories

>140,000 images



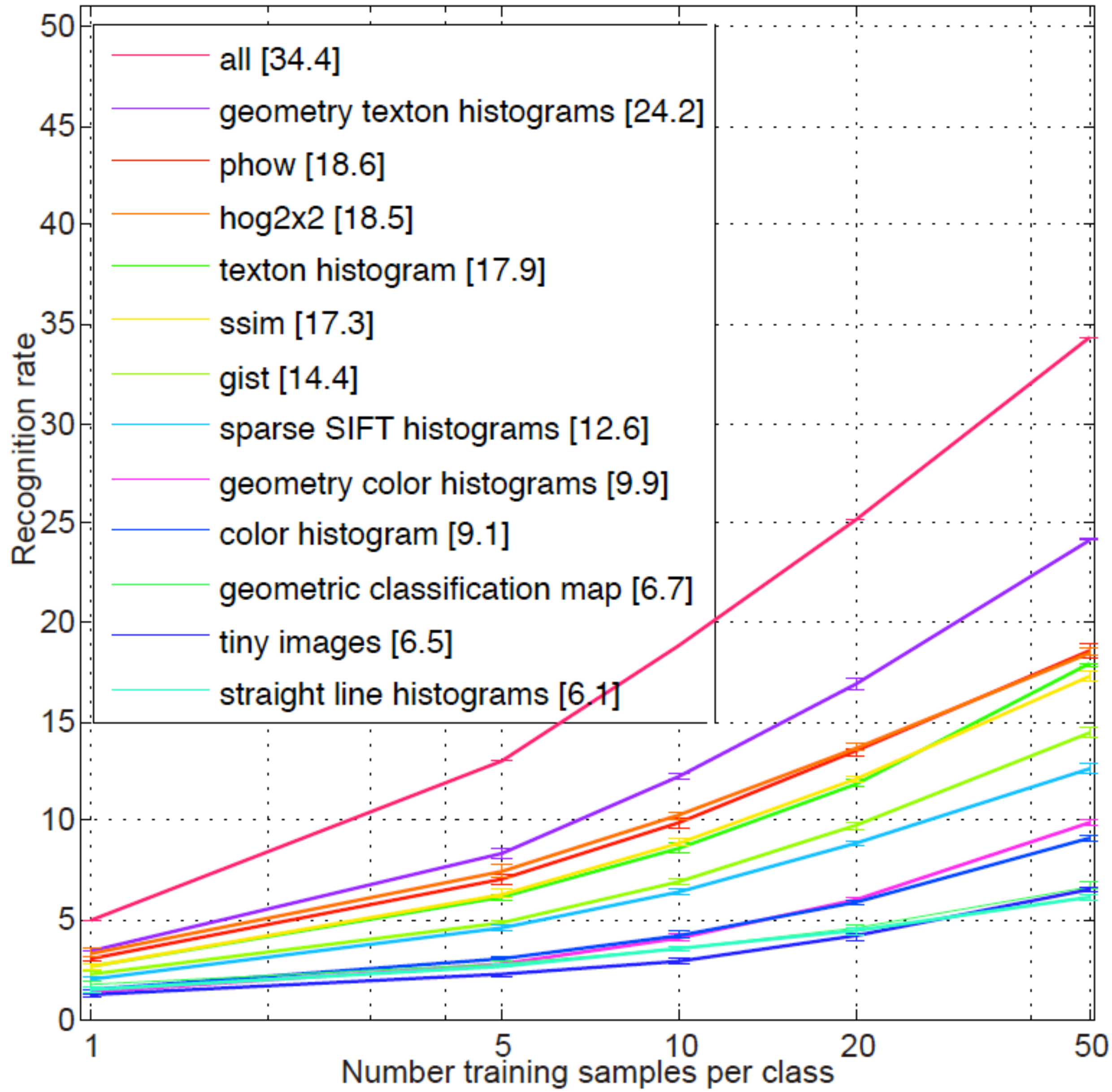
Slides by A. Torralba

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



Slides by A. Torralba

Performance with 400 categories



Slides by A. Torralba

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

Training images

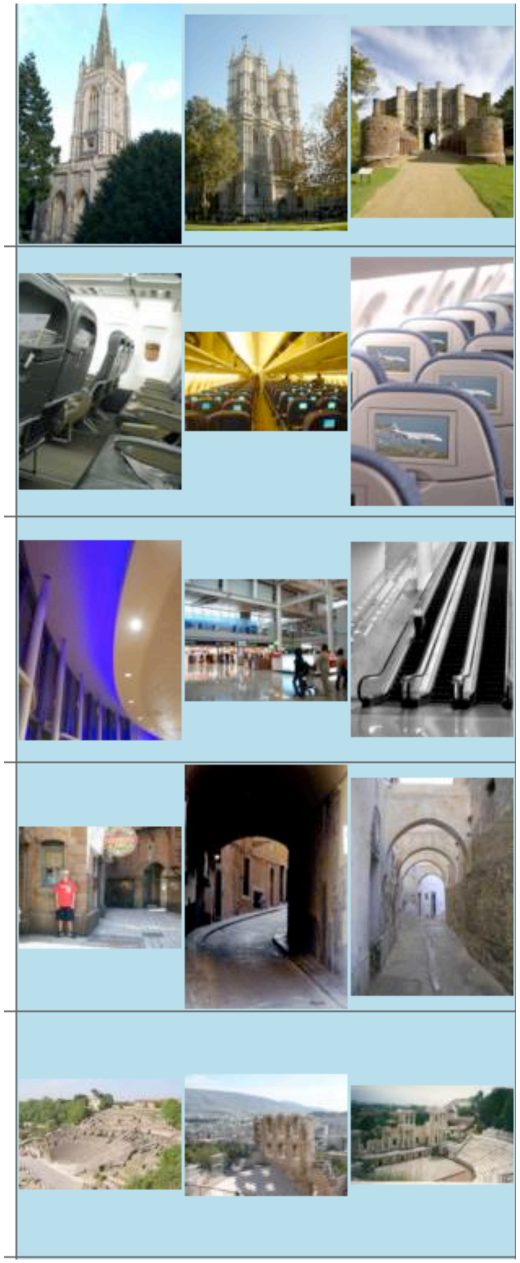
Abbey

Airplane cabin

Airport terminal

Alley

Amphitheater

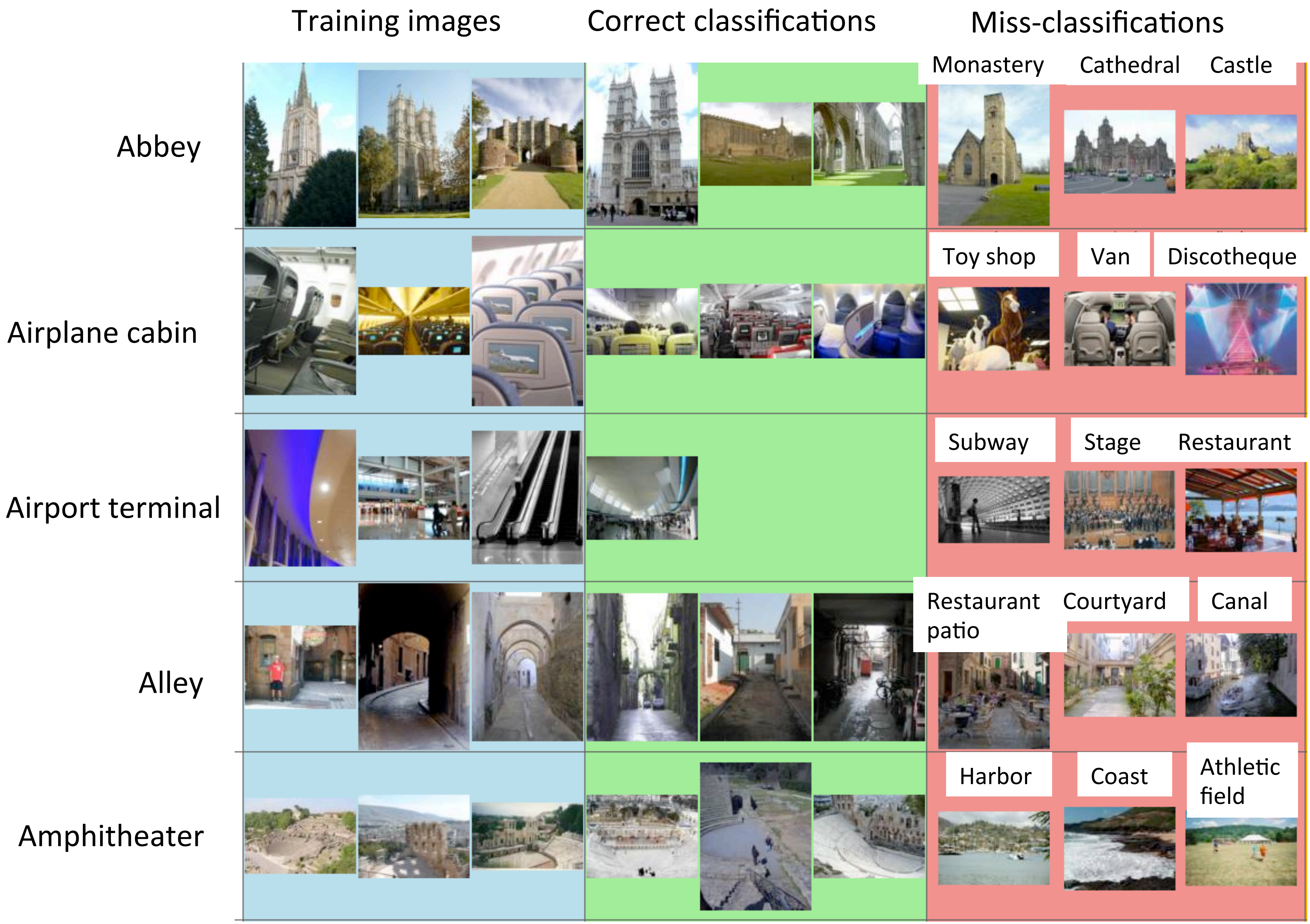


Slides by A. Torralba

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

Slides by A. Torralba

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



Slides by A. Torralba

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

Categories or a continuous space?

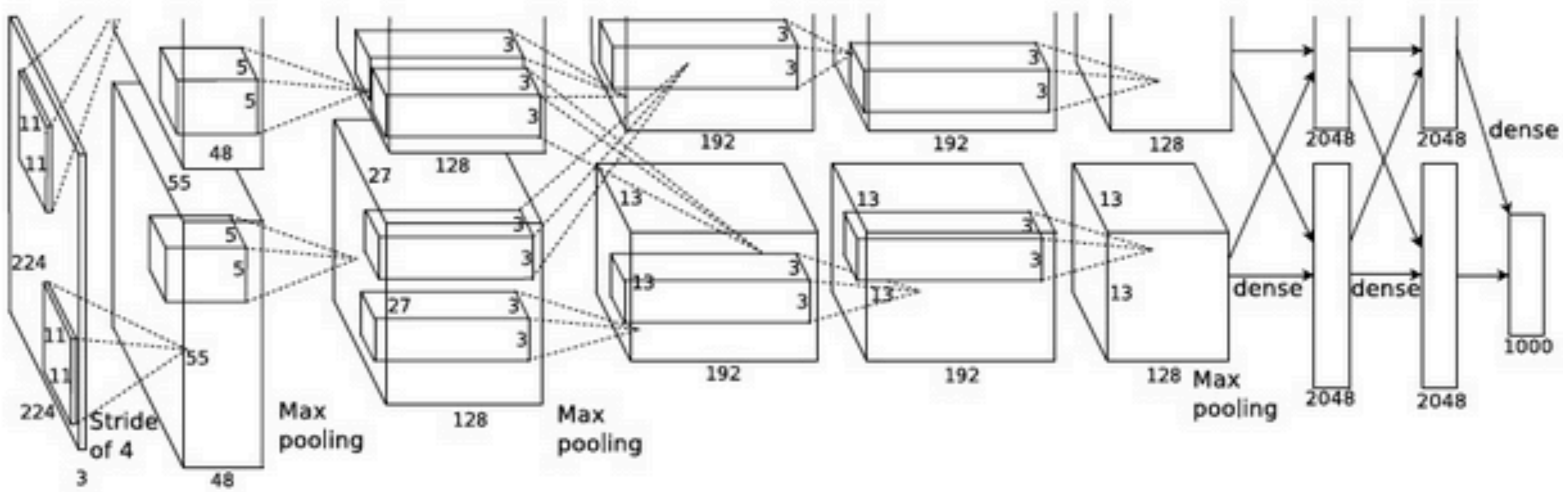
From the city to the mountains in 10 steps

Slides by A. Torralba

MVA internship topic:

1. Large-scale image classification and object detection with Deep Convolutional Neural Networks

Project supervisors: Leon Bottou <leon@bottou.org>, Ivan Laptev <Ivan.Laptev@ens.fr> and Josef Sivic <Josef.Sivic@ens.fr>
Location: Willow Group, Laboratoire d'Informatique de l'École Normale Supérieure



Recall: Similar classification pipeline has been used for object classification (lecture 4, C. Schmid)

Experimental results

- Datasets
  - ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC)
    - 1000 classes and 1.4M images
  - ImageNet10K dataset
    - 10184 classes and ~ 9 M images

(a) Star Anise (92.45%) (b) Geyser (85.45%) (c) Pulp Magazine (83.01%) (d) Carrycot (81.48%)

(e) European gallinule (15.00%) (f) Sea Snake (10.00 %) (g) Paintbrush (4.68 %) (h) Mountain Tent (0.00%)

Recall: Similar classification pipeline has been used for object classification (lecture 4, C. Schmid)

Experimental results

- Features: dense SIFT, reduced to 64 dim with PCA

- Fisher vectors
  - 256 Gaussians, using mean and variance
  - Spatial pyramid with 4 regions
  - Approx. 130K dimensions (4x [2x64x256])
  - Normalization: square-rooting and L2 norm

- BOF: dim 1024 + R=4
  - 4960 dimensions
  - Normalization: square-rooting and L2 norm

Recent break-through by neural networks?

Recent work [Krizhevsky12] (to appear at NIPS'12) have shown that significant performance gains on the ImageNet benchmark can be obtained by a vastly different architecture based on a **deep convolutional neural network** (Recall lecture 7, by N. Le Roux)

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The

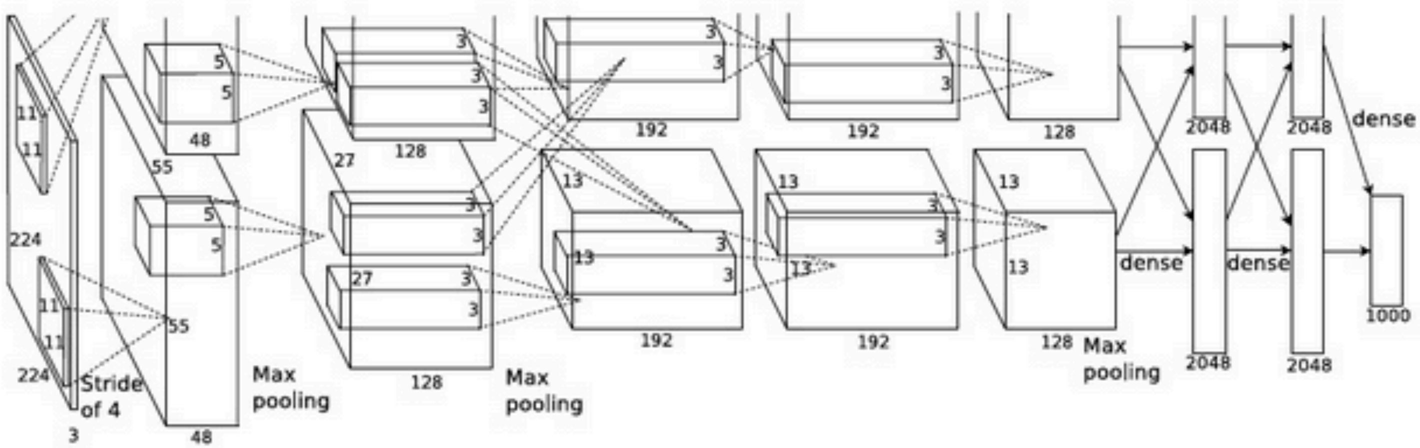
Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

MVA internship topic:

1. Large-scale image classification and object detection with Deep Convolutional Neural Networks

Project supervisors: Leon Bottou <leon@bottou.org>, Ivan Laptev <Ivan.Laptev@ens.fr> and Josef Sivic <Josef.Sivic@ens.fr>
Location: Willow Group, Laboratoire d'Informatique de l'École Normale Supérieure



Goal

You will experiment with a very recent and, as it appears, groundbreaking approach to image classification based on deep convolutional neural networks in [Krizhevsky12]. The goals are to replicate the state-of-the-art results in [Krizhevsky12] and to extend the method to object detection.

Motivation

Recognizing thouthands of object categories from images is a long-standing goal of computer vision. In recent years the research on large-scale image classification, e.g., [Sanchez11] has been sparked by the large amounts of now available image data and large-scale datasets such as ImageNet. Convolutional Neural Network (CNN) based methods exist for several decades, however, until recently the successful applications of CNNs have been only shown for relatively limited problems such as handwritten digit recognition [LeCun90] and face detection [Rowley98]. The groundbreaking results of [Krizhevsky12] presented in the Large Scale Visual Recognition Challenge 2012 Workshop (ILSVRC2012) now indicate that CNN is a highly competitive tool when powered with lots of image data. It might be that the necessary amount of image data and the critical processing power of modern GPUs sufficient to train successful CNN classifiers has just been reached and we are in front of many exciting new applications of CNNs. This internship will investigate this very timely topic by first re-producing results in [Krizhevsky12], investigating the performance and properties of this method when applied to other classification tasks, such as PASCCAL VOC, and then extending the classification method in [Krizhevsky12] to a more challenging task of object detection. This is an exploratory internship topic in an exciting and emerging area, which may have a significant impact on the current state of visual recognition.

Project description

The project will build on an existing publically available codebase available from [Krizhevsky12] and will proceed in the following three steps:

1. Understand the approach and the existing code of [Krizhevsky12]. Reproduce their quantitative and qualitative image classification results on the ImageNet database.

2. Improve image classification accuracy of [Krizhevsky12] by extending their work. The project will consider different extensions such as enlarging the class of image transformations when “jittering” the training data.

3. Apply and extend the deep convolutional neural network approach to object detection/localization on the Pascal VOC dataset.

The project will be co-supervised by Leon Bottou who is one of the world leading experts on neural networks and large-scale learning.

Project description

The project will build on an existing publically available codebase available from [Krizhevsky12] and will proceed in the following three steps:

1. Understand the approach and the existing code of [Krizhevsky12]. Reproduce their quantitative and qualitative image classification results on the ImageNet database.

2. Improve image classification accuracy of [Krizhevsky12] by extending their work. The project will consider different extensions such as enlarging the class of image transformations when "jittering" the training data.

3. Apply and extend the deep convolutional neural network approach to object detection/localization on the Pascal VOC dataset.

The project will be co-supervised by Leon Bottou who is one of the world leading experts on neural networks and large-scale learning.

Requirements

We are looking for strongly motivated candidates with an interest in computer vision and machine learning. The project requires strong background in applied mathematics and excellent programming skills. The project will also involve using and possibly programming GPUs. Prior experience with GPUs will be also useful, but not required. If we find a mutual, match the project can lead to a Phd at the Willow group.

References

[Krizhevsky09] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks (2012), In Proc. NIPS 2012.

[Sanchez11] J. Sanchez, F. Perronnin. High-dimensional signature compression for large-scale image classification, In Proc. CVPR 2011.

[LeCun90] Y. Le Cun, B. Boser, J.S. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel. Handwritten digit recognition with a back-propagation network. in Proc. NIPS 1990.

Exploiting regularities in real-world scenes

Scenes are unique

Slides by A. Torralba

But not all scenes are so original

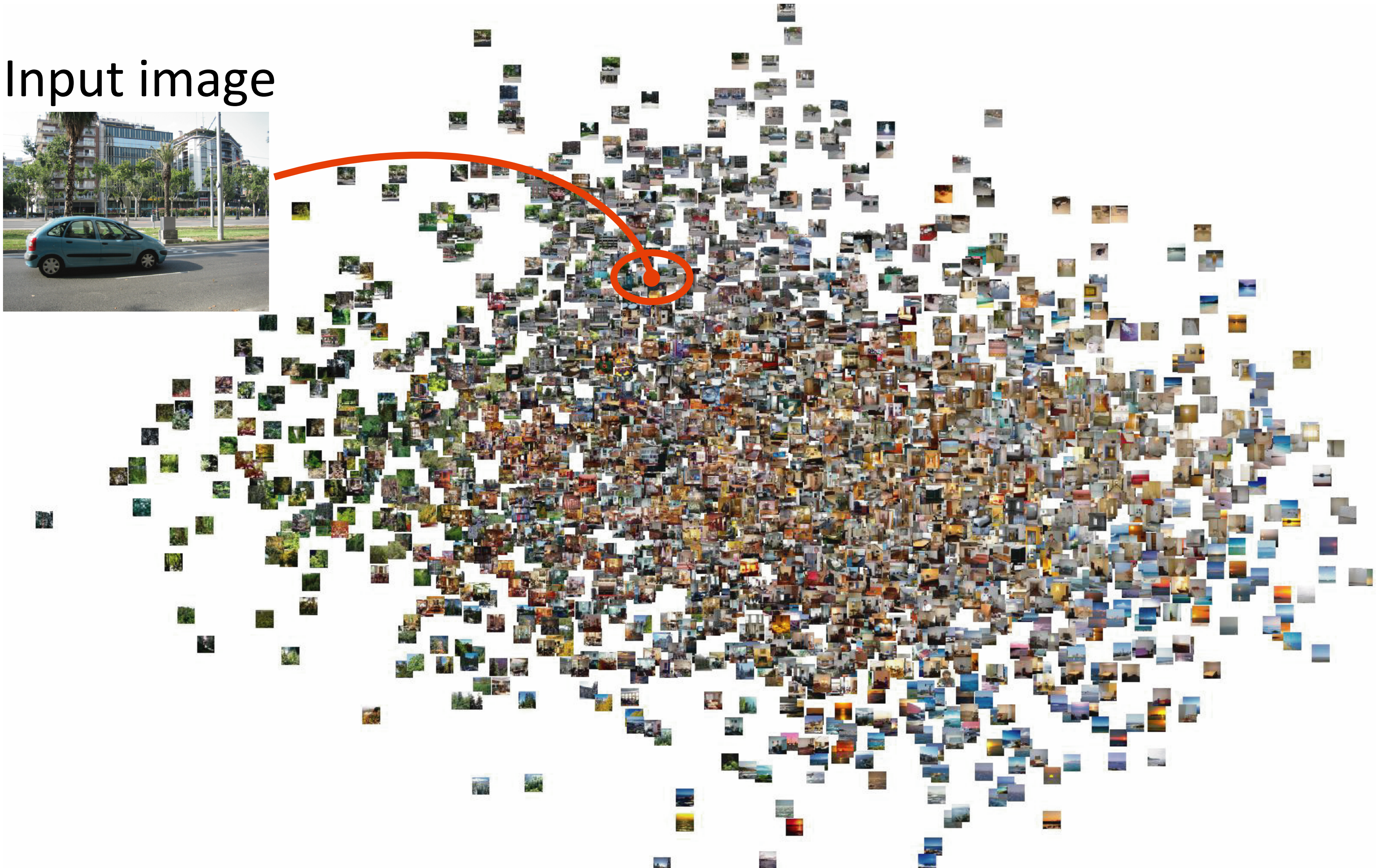
Slides by A. Torralba

But not all scenes are so original

Slides by A. Torralba

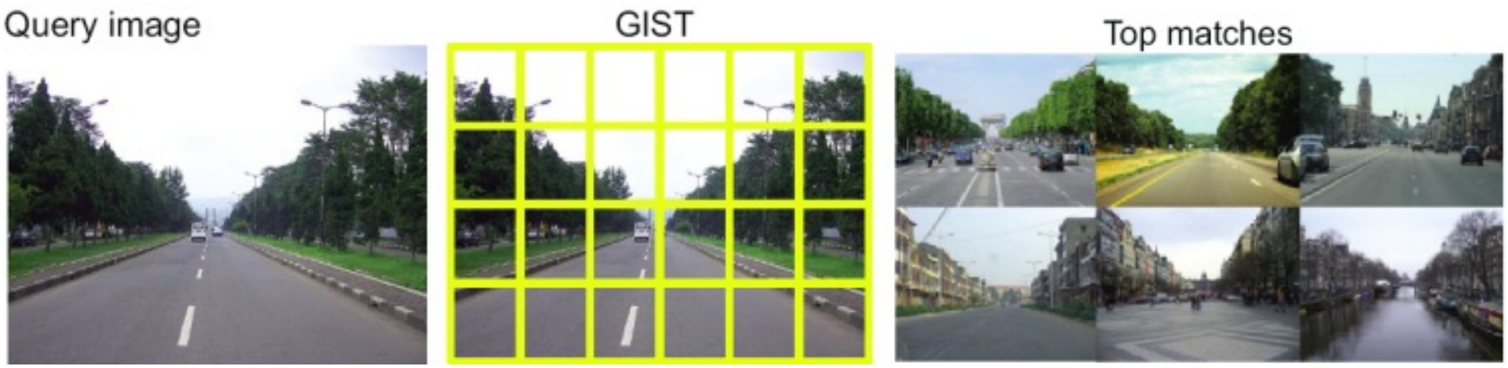
Find similar scenes by matching image descriptors

Input image



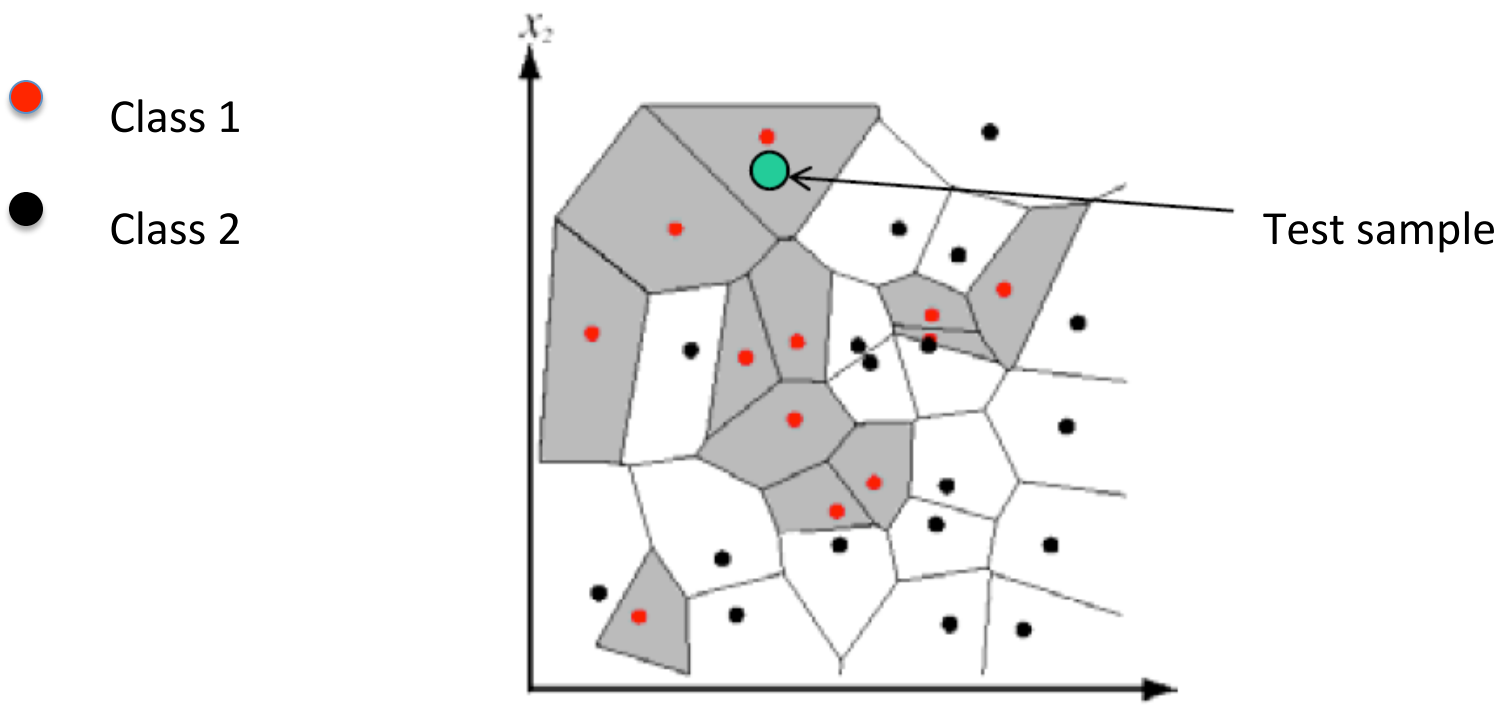
Slides by A. Torralba

Find similar scenes by matching image descriptors



Nearest neighbors classification

- Given a new test sample, assign the label of the nearest neighbor



from Duda *et al.*

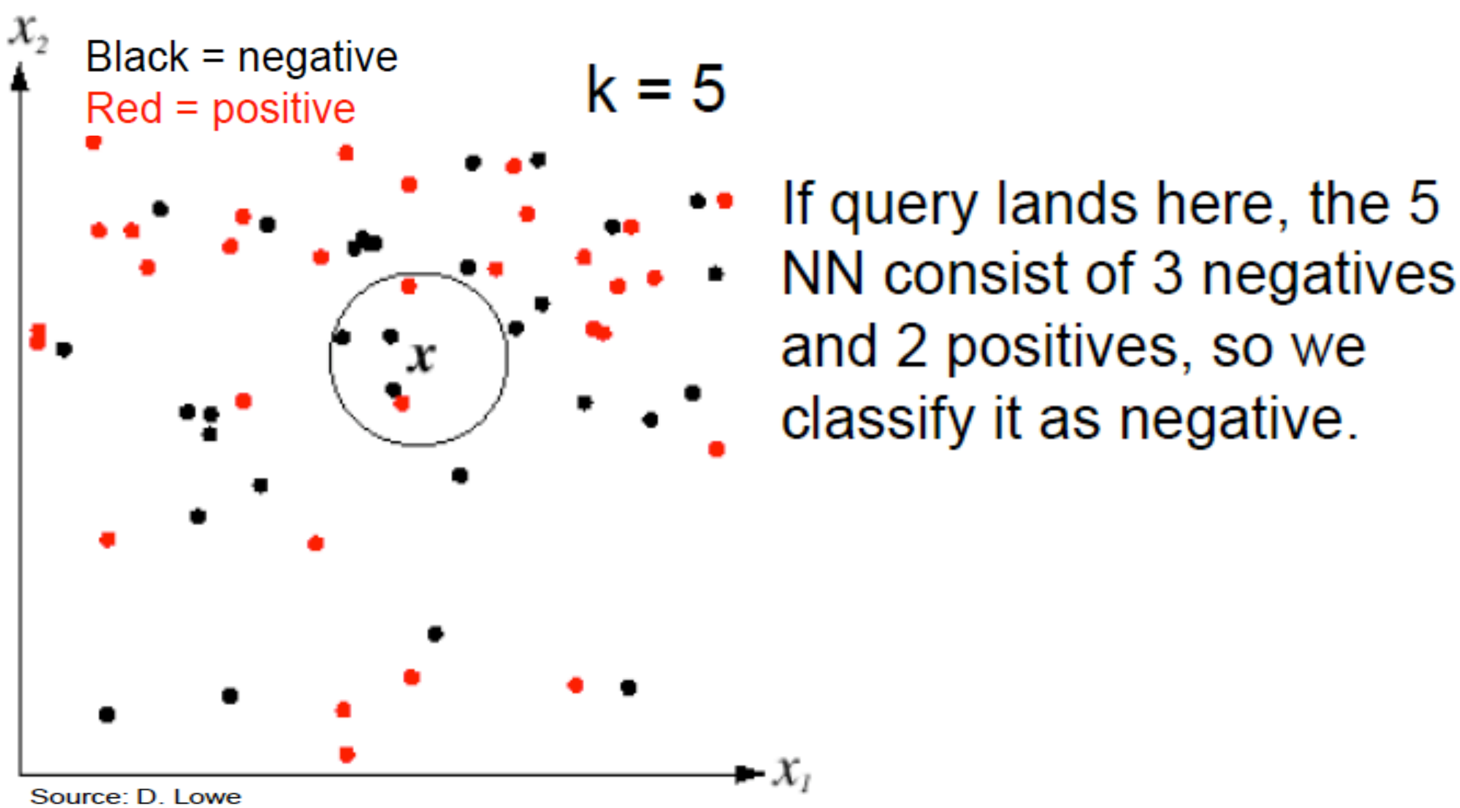
Voronoi partitioning of feature space

Slides by A. Torralba

K-Nearest neighbors classification

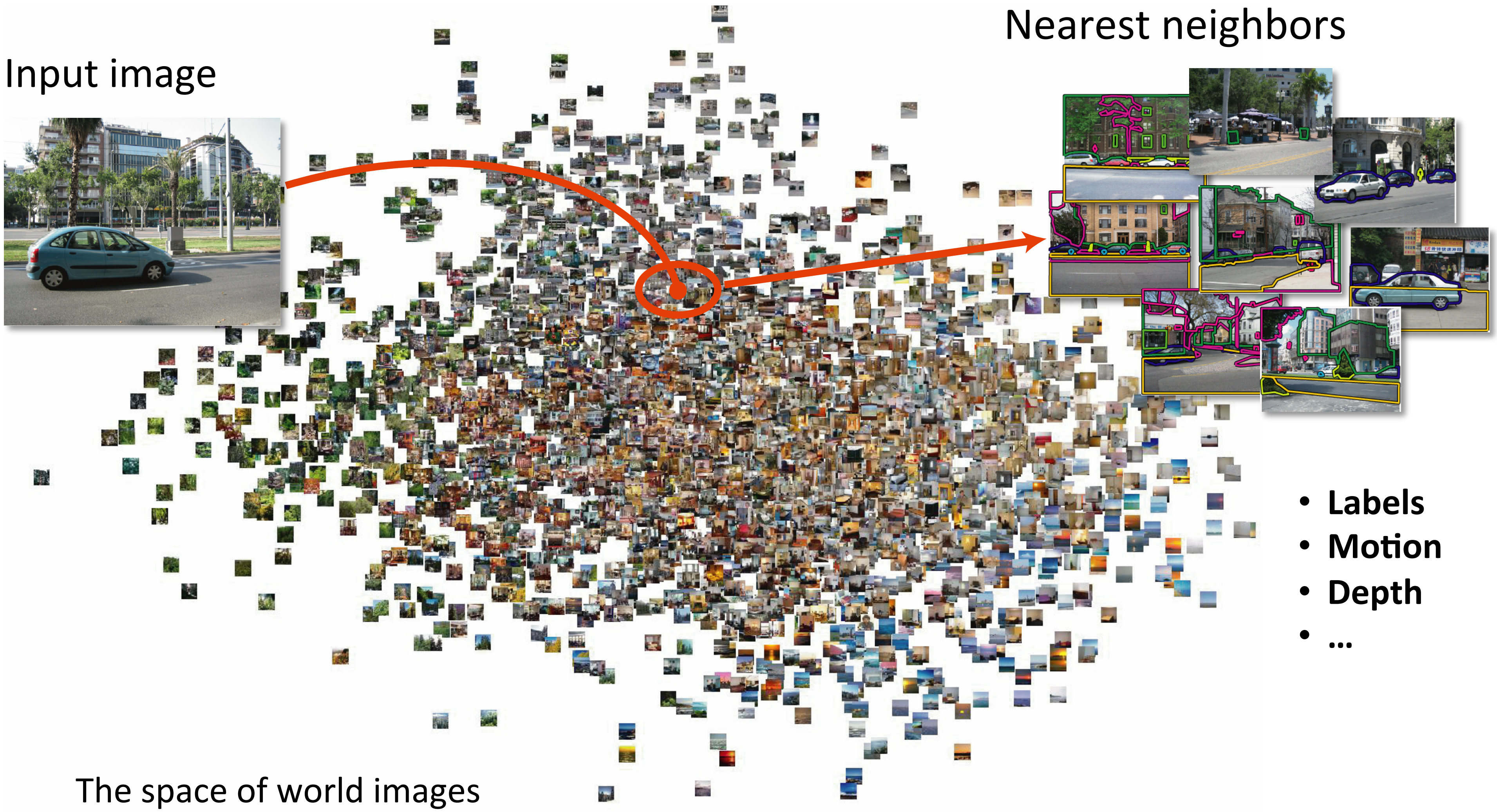
Find the K closest points to the test sample

Use labels of the K neighbors to vote



Slides by A. Torralba

Transfer information to the input image from the nearest neighbors



- **Labels**
- **Motion**
- **Depth**
- **…**

Hays, Efros, Siggraph 2006
Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007

Slides by A. Torralba

im2gps

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

20 million geotagged and geographic text-labeled images

Slides by A. Torralba

Hays & Efros. CVPR 2008

Hays & Efros. CVPR 2008

im2gps

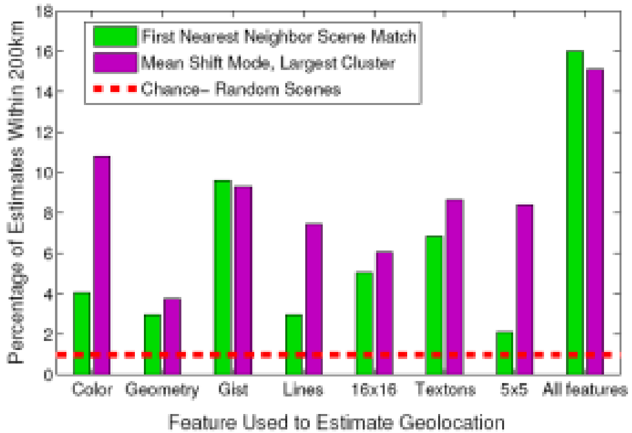


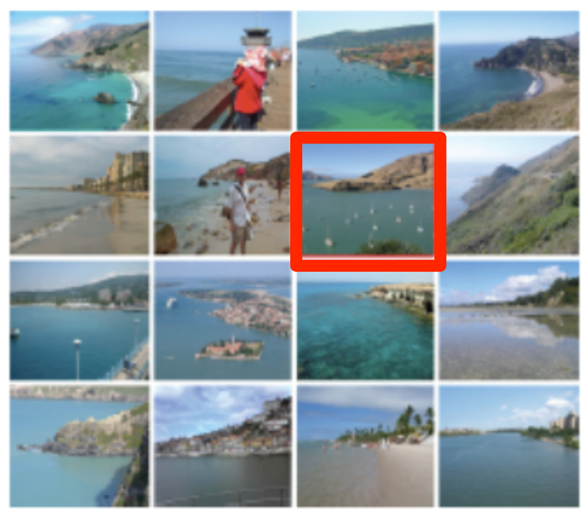
Figure 5. *Geolocation performance across features.* Percentage of test cases geolocated to within $200km$ for each feature. We compare geolocation by 1-NN vs. largest mean-shift mode.

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

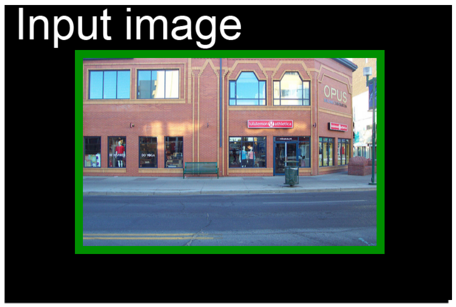
Slides by A. Torralba

Hays, Efros, 2007

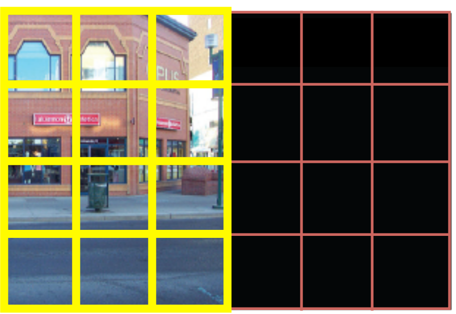
Scene matching with camera view transformations to predict scene outside of the image boundaries

[Sivic, Kaneva, Torralba, Avidan, Freeman, PIEEE 2009]
http://www.di.ens.fr/~josef/publications/kaneva09b.pdf

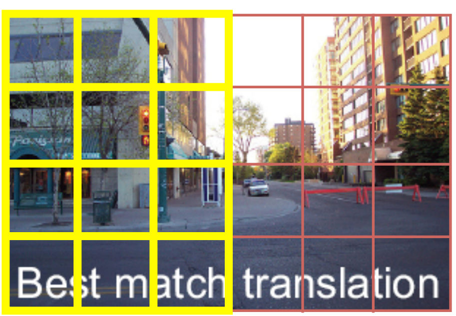
Scene matching with camera view transformations: Translation



1. Move camera



2. View from the virtual camera



3. Find a match to fill the missing pixels

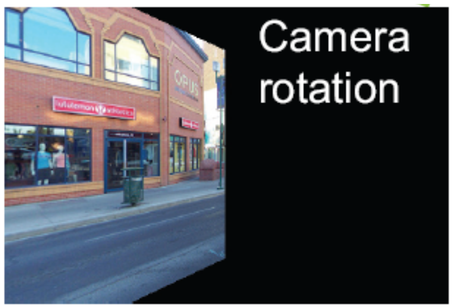
4. Locally align images

5. Find a seam

6. Blend in the gradient domain

Scene matching with camera view transformations: Camera rotation

1. Rotate camera



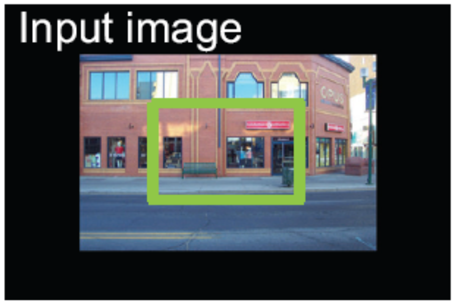
2. View from the virtual camera

3. Find a match to fill-in the missing pixels

4. Stitched rotation

5. Display on a cylinder

Scene matching with camera view transformations: Forward motion



1. Move camera

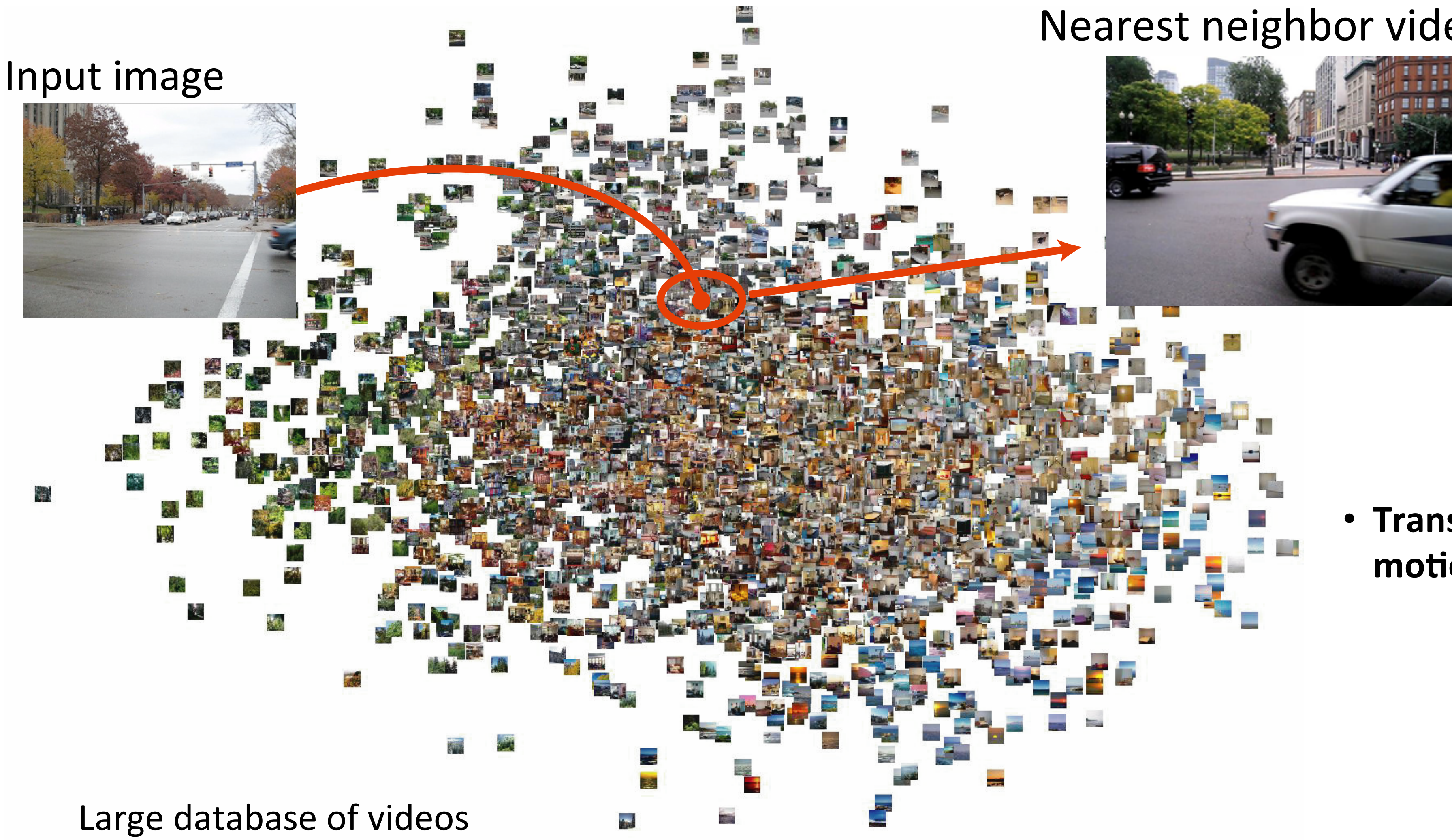
2. View from the virtual camera

3. Find a match to replace pixels

Basic camera motions

Exploring famous sites

Predict events



- **Transfer motion**

[Liu, Yuen, Torralba, Sivic and Freeman, ECCV 2008]

Adapted from A. Torralba

Motion synthesis results

Still image

Video of the best match

Motion synthesis results

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

Motion synthesis results

Still image

Video of the best match

Motion synthesis results

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

Discussion

- Regularities in scene appearance can be used for a number of applications (label transfer - recognition, scene completion, gps location prediction, event prediction…)

- Performance depends on the quality of the matches, i.e. is the particular scene represented in the database?

    - Increase database size [Torralba, PAMI 2008].

    - Combine multiple database images [Russell et al. NIPS 2009]

    - Object-level labeling [Liu et al. CVPR 2009]

However, some “atypical” scenes might still not be represented well.

Today: Scenes and objects

1. Scenes as textures (without modeling objects and their relations)

2. Objects within a scene

3. Recognizing multiple objects in an image.

Part II: Objects within a scene (context)

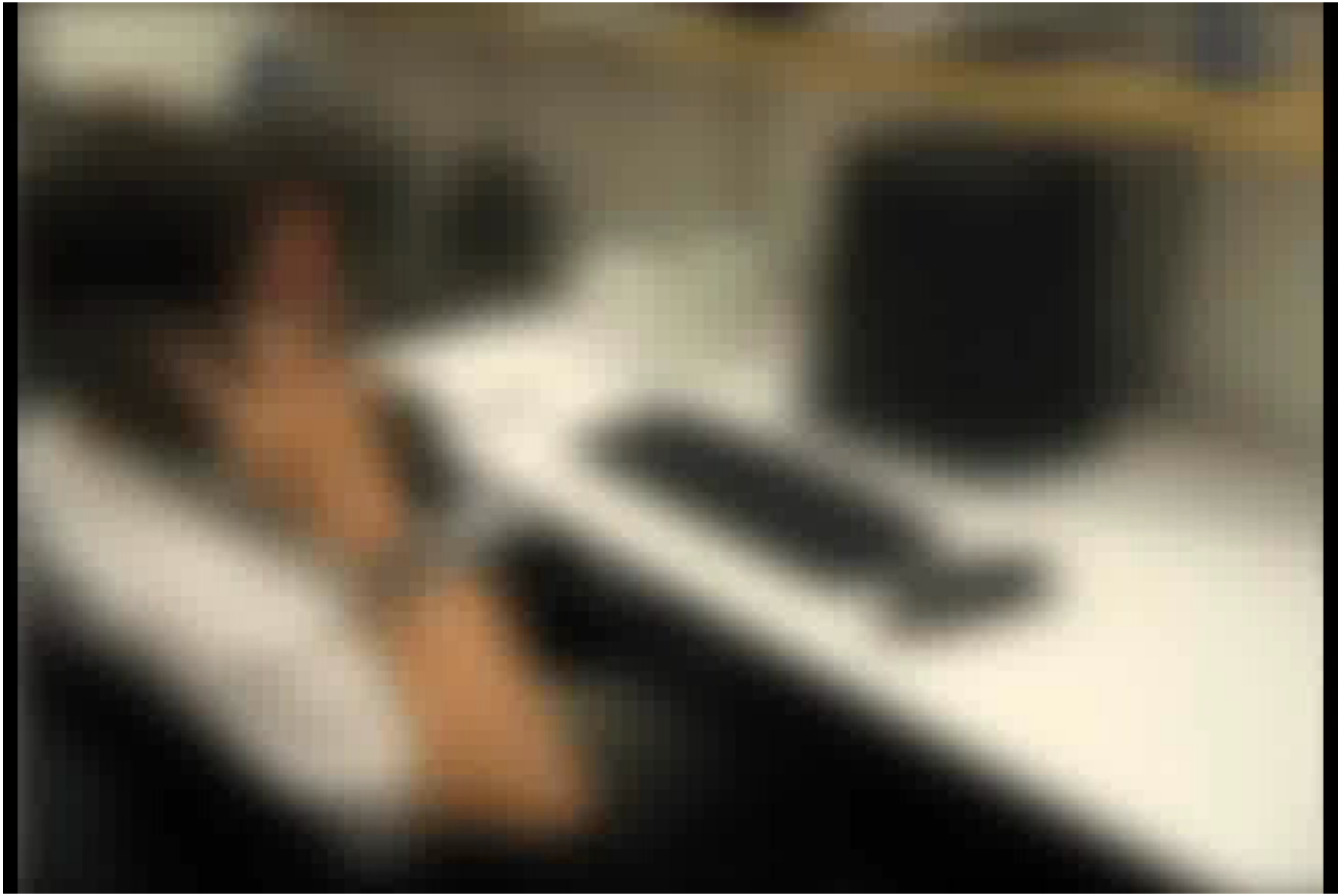
Figure from A. Torralba

Why is context important?

- Changes the interpretation of an object (or its function)

- Context defines what an unexpected event is

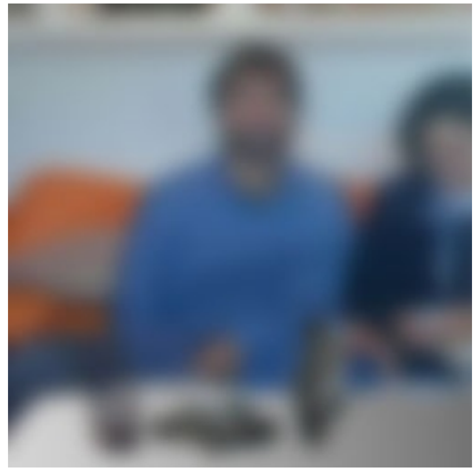
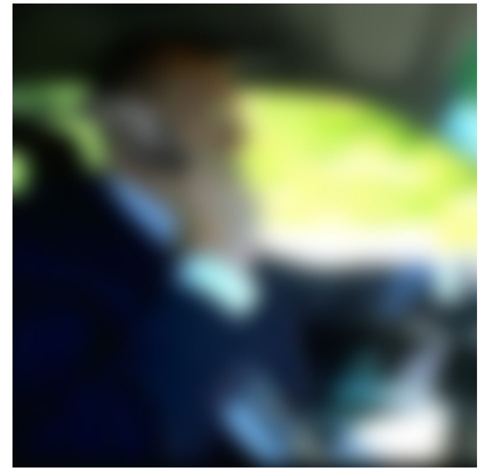
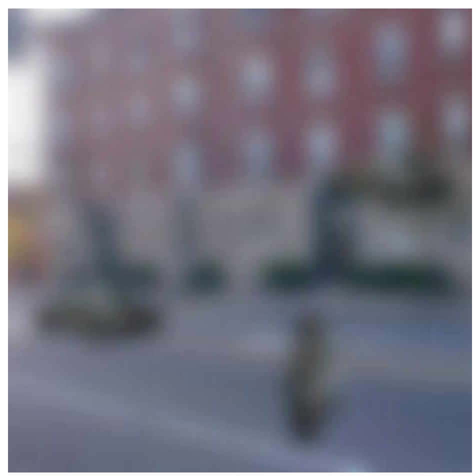
Slides by A. Torralba



Slides by A. Torralba

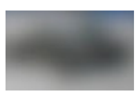
Slides by A. Torralba

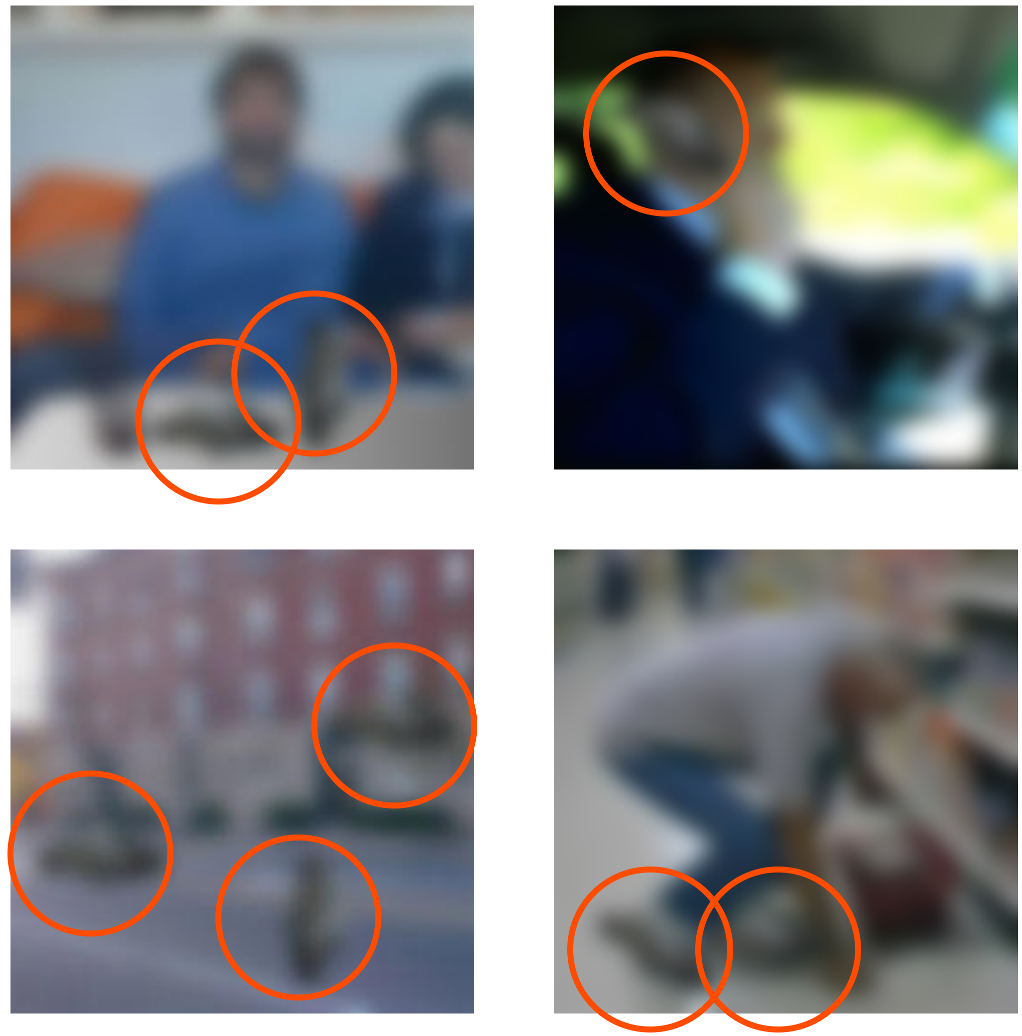
The multiple personalities of a blob


Slides by A. Torralba

The multiple personalities of a blob

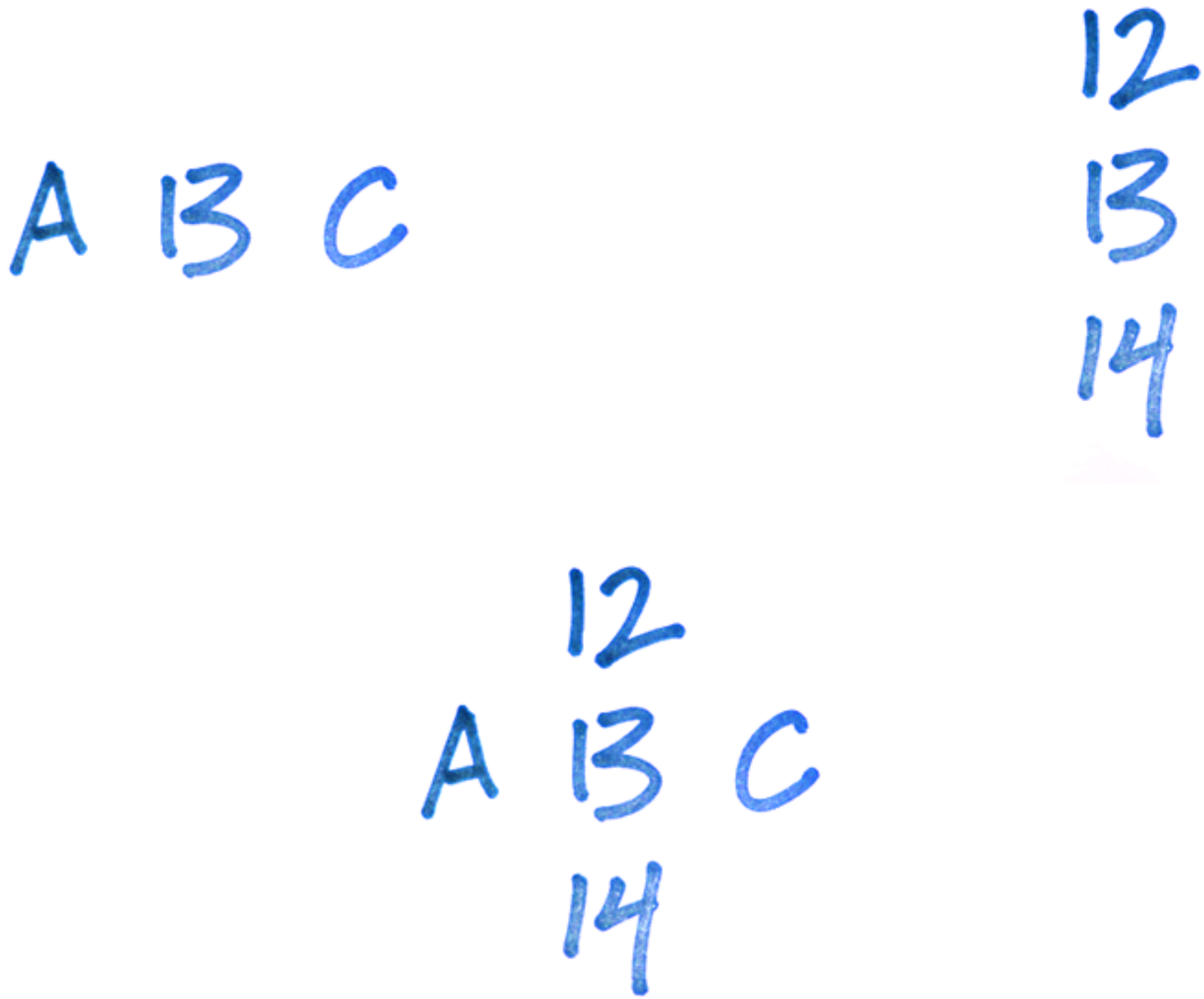




Slides by A. Torralba

A B C

12
13
14



Slides by A. Torralba

Look-Alikes by Joan Steiner

Even in high resolution, we can not shut down contextual processing and it is hard to recognize the true identities of the elements that compose this scene.

Slides by A. Torralba

Who needs context anyway?

We can recognize objects even out of context

Banksy

Slides by A. Torralba

The importance of context

- Cognitive psychology
  - Palmer 1975
  - Biederman 1981
  - …

Figure 3. *An example of a triple violation. The taxi is violating the Probability, Support, and Size relations.*

Biederman et al. 81

- 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)

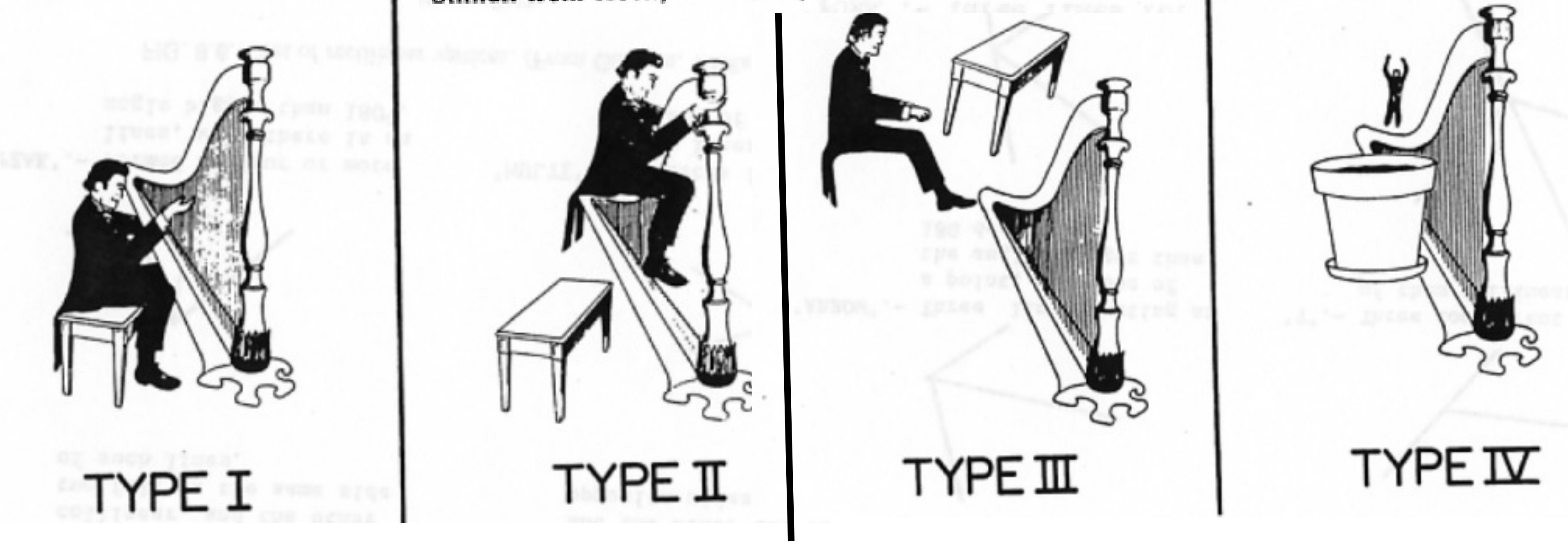
Class	Context elements	Operator
SKY	ALWAYS	ABOVE-HORIZON
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ 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

Slides by A. Torralba

[Strat and Fischler 1989]

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

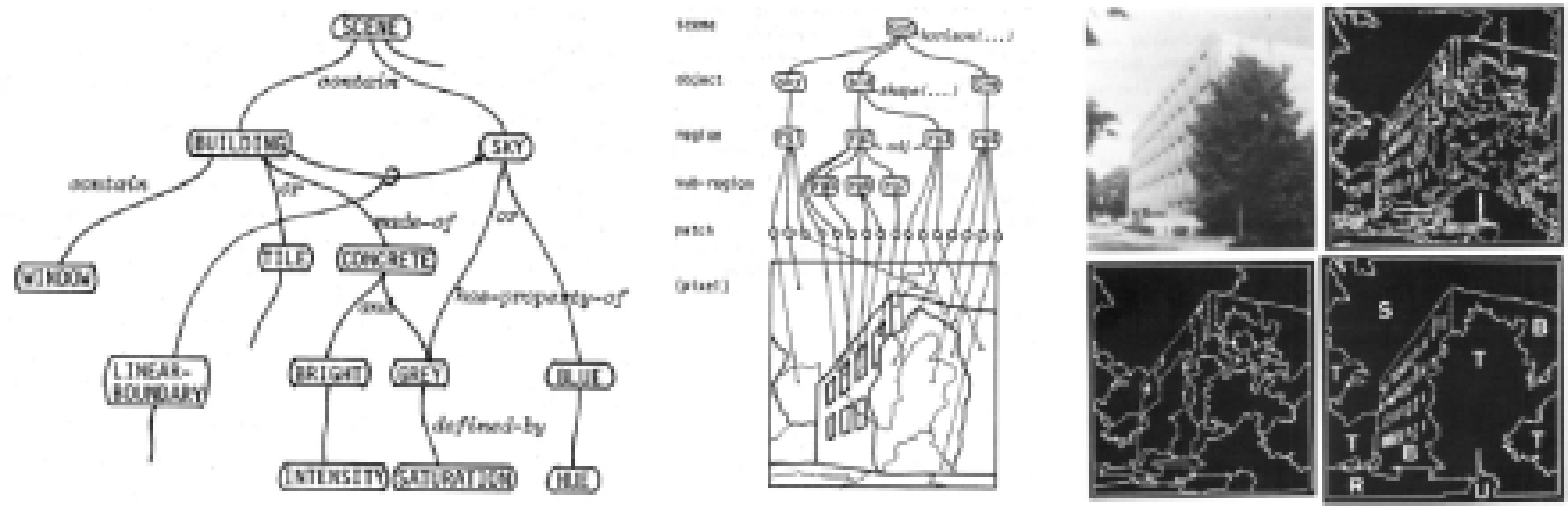
Strat and Fischler (1991)

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

Slides by A. Torralba

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

Slides by A. Torralba

What is the context for a single object category?

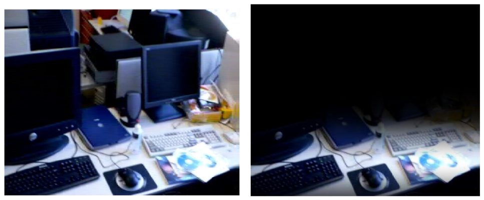
Slides by A. Torralba

The influence of an object extends beyond its physical boundaries

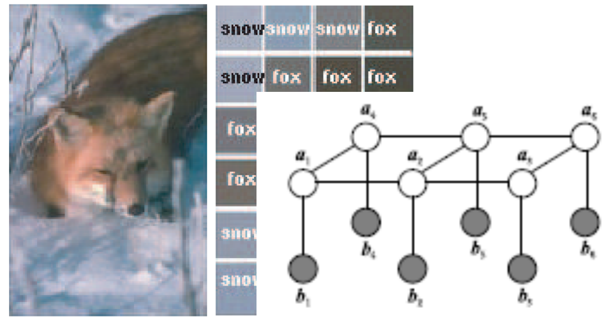
Slides by A. Torralba

Objects in context

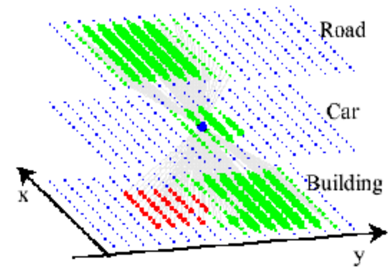
Torralba, Sinha (2001)



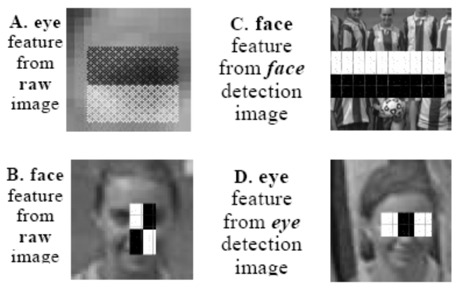
Carbonetto, de Freitas & Barnard (2004)



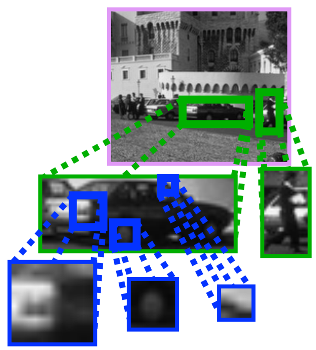
Torralba Murphy Freeman (2004)



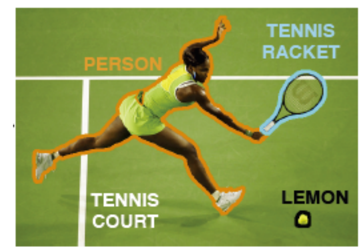
Fink & Perona (2003)



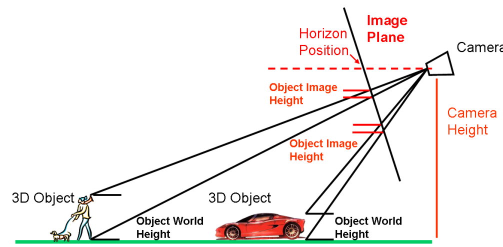
Sudderth, Torralba, Wilsky, Freeman (2005)



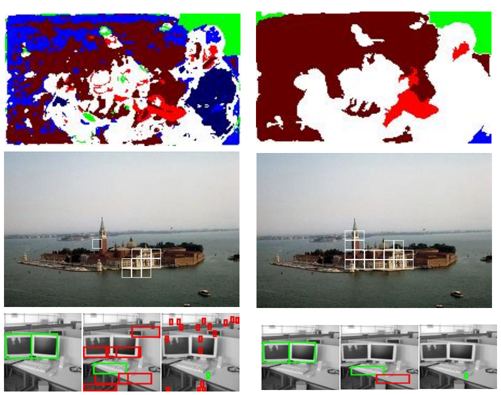
Rabinovich et al (2007)



Hoiem, Efros, Hebert (2005)



Kumar, Hebert (2005)



Heitz and Koller (2008)

Desai, Ramanan, and Fowlkes (2009)

Slides by A. Torralba

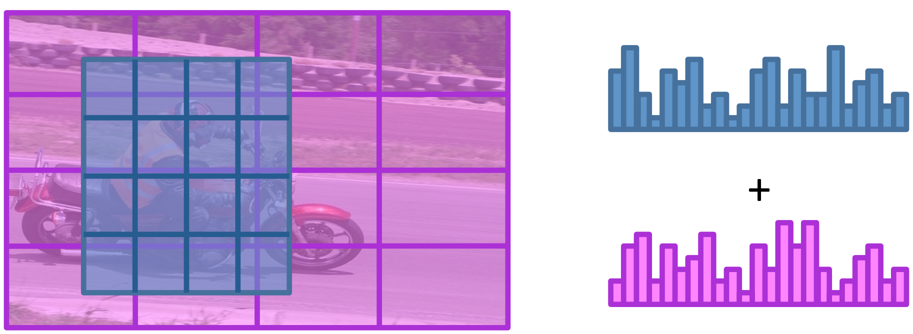
See also…

Choi, Torralba and Wilsky, PAMI 2012
Tree-based context model for object recognition

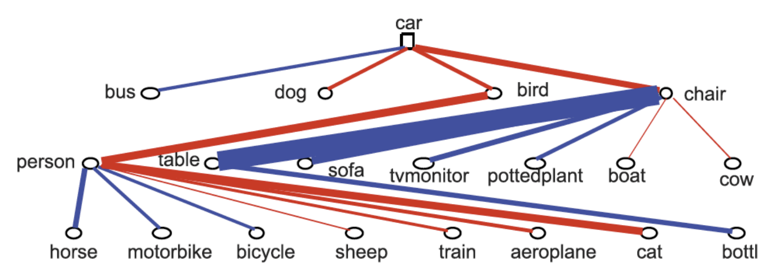
Localization++ Classification--

Localization-- Classification++

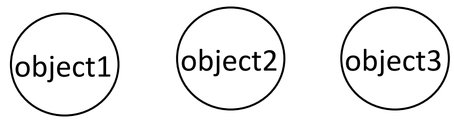
V. Delaitre, I. Laptev and J. Sivic
Action recognition in still images… , BMVC 2010



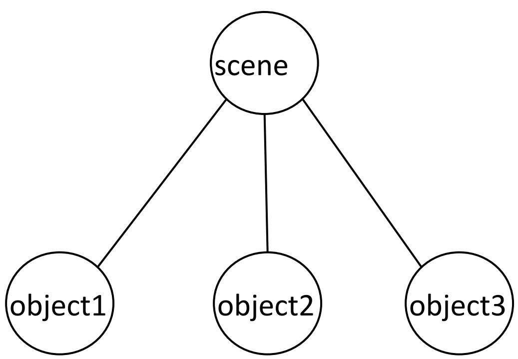
Choi, Torralba and Wilsky, PAMI 2012
Tree-based context model for object recognition



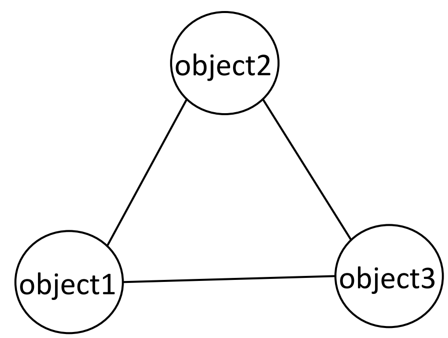
Context models



Independent model



Objects are correlated / constrained via the scene

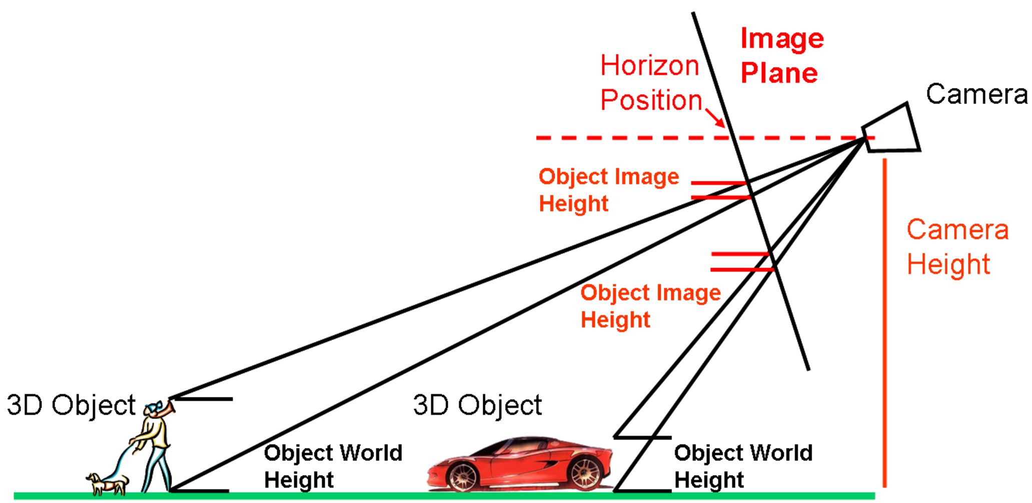


Dependencies among objects

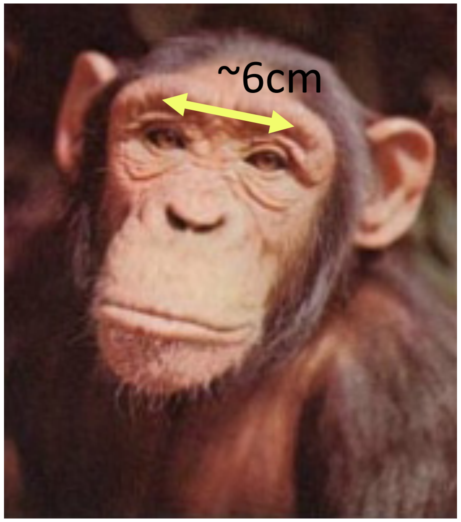
Slides by A. Torralba

Example: 3D scene context

[Hoiem, Efros, Hebert (2005)]

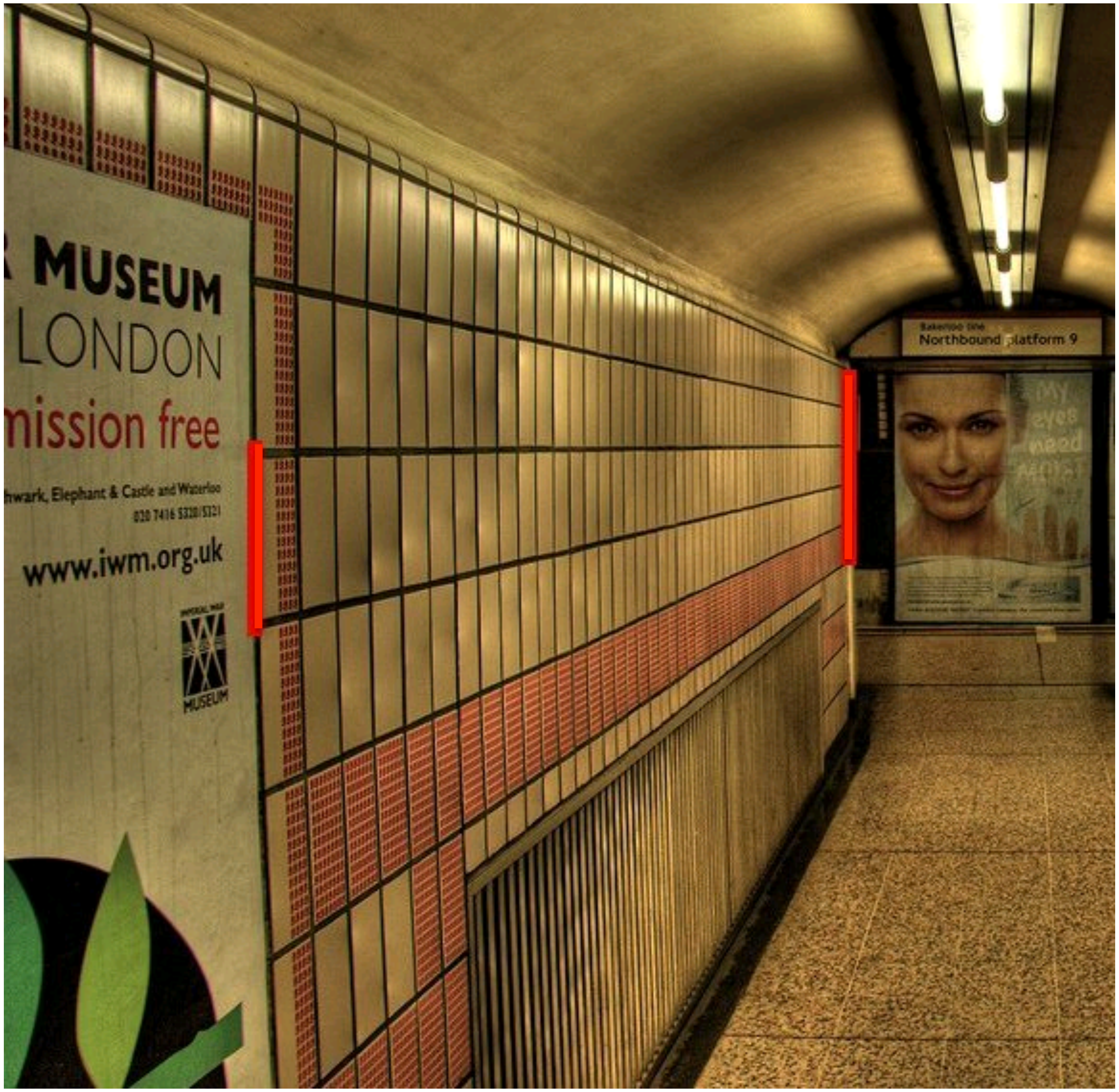


We are wired for 3D



Slides by A. Torralba

We can not shut down 3D perception



(c) 2006 Walt Anthony

Slides by A. Torralba

3D from pixel values (single view)

D. Hoiem, A.A. Efros, and M. Hebert, "Automatic Photo Pop-up”. SIGGRAPH 2005.

A. Saxena, M. Sun, A. Y. Ng. "Learning 3-D Scene Structure from a Single Still Image"
In ICCV workshop on 3D Representation for Recognition (3dRR-07), 2007.

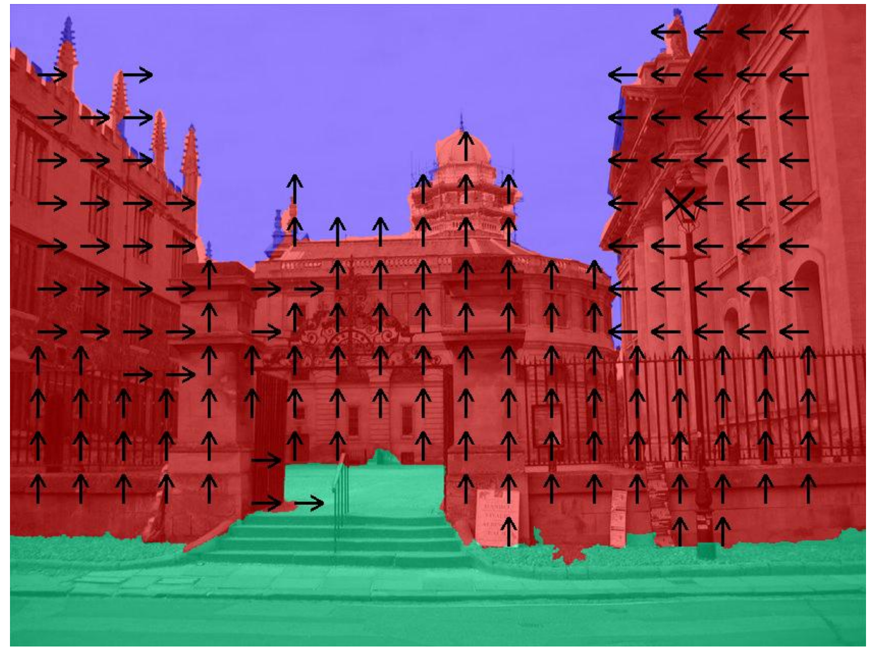
Slides by A. Torralba

Learn Surface Orientations

- User recognition to learn structure of the world from labeled examples

Slides by Efros

Label Geometric Classes



- **Goal:** learn labeling of image into 7 Geometric Classes:
- **Support (ground)**
- **Vertical**
  - Planar: facing **Left** **(←)**, **Center** **(↑)**, **Right** **(→)**
  - Non-planar: **Solid** **(X)**, **Porous** or wiry **(O)**
- **Sky**

Slides by Efros

What cues to use?

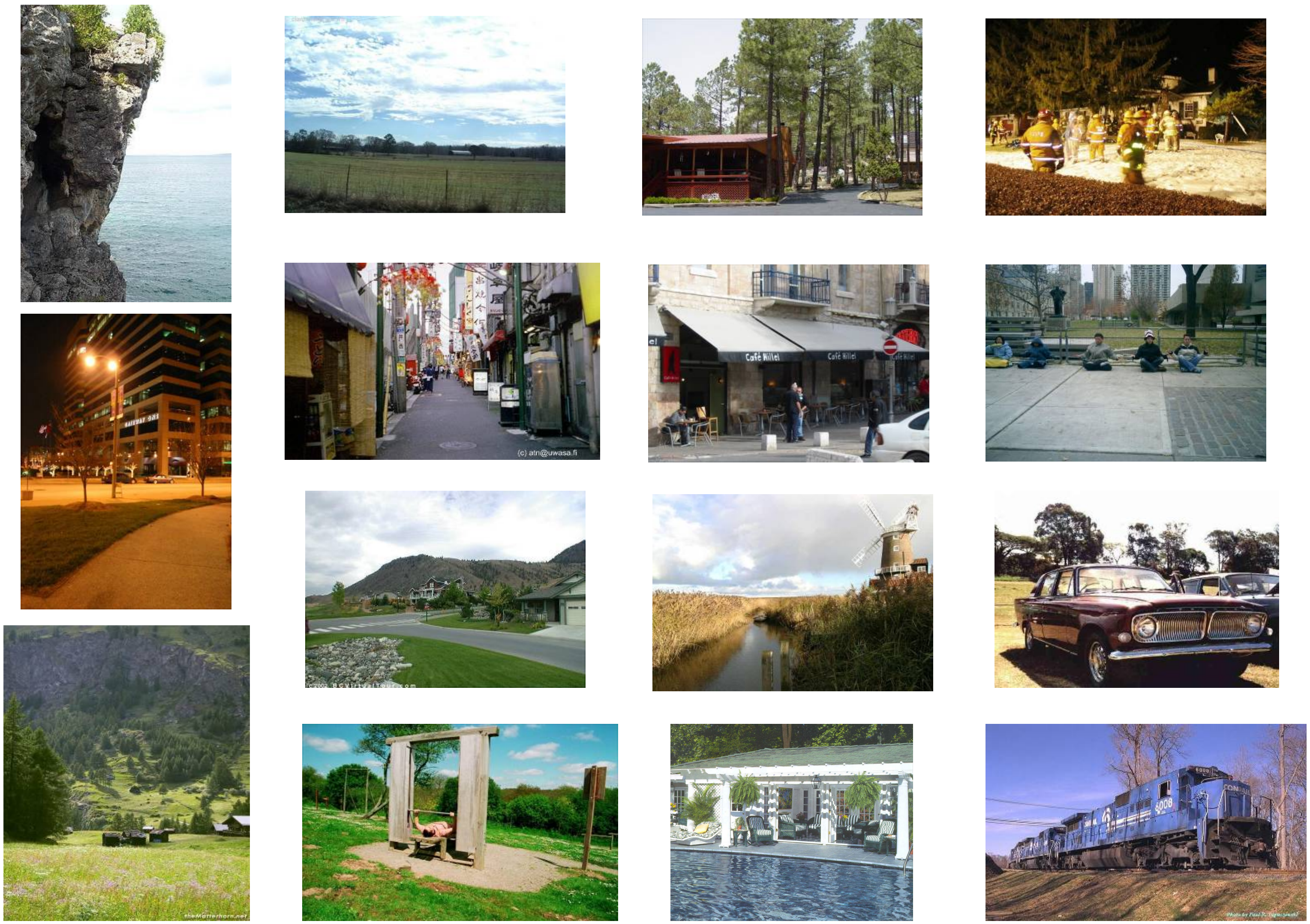
Vanishing points, lines

Color, texture, image location

Texture gradient

Slides by Efros

Dataset very general



Slides by Efros

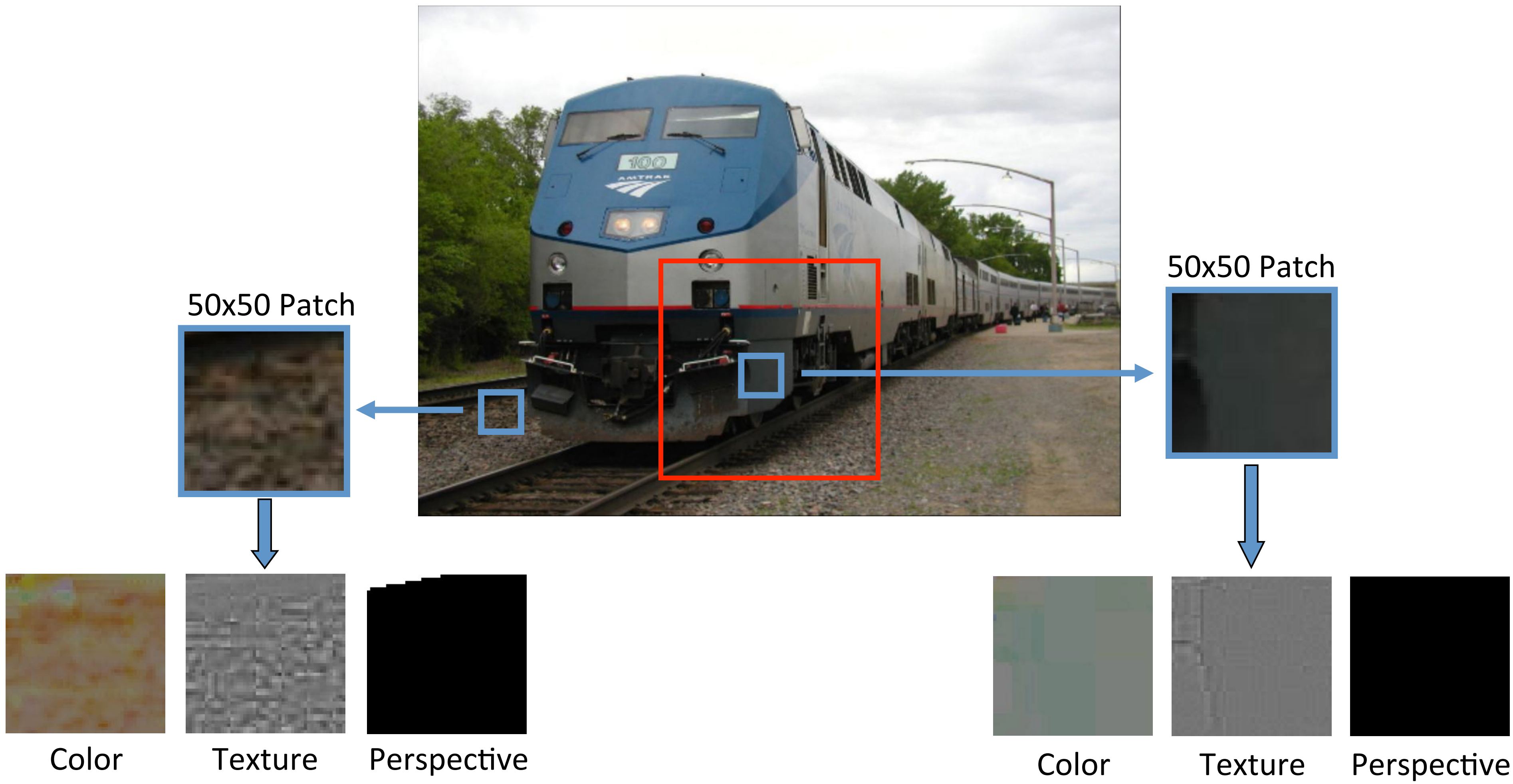
Let's use many <u>weak</u> cues

- Material
- Image Location
- Perspective

SURFACE CUES
Location and Shape
L1. Location: normalized x and y, mean
L2. Location: norm. x and y, 10^{th} and 90^{th} pctl
L3. Location: norm. y wrt estimated horizon, 10^{th}, 90^{th} pctl
L4. Location: whether segment is above, below, or straddles estimated horizon
L5. Shape: number of superpixels in segment
L6. Shape: normalized area in image
Color
C1. RGB values: mean
C2. HSV values: C1 in HSV space
C3. Hue: histogram (5 bins)
C4. Saturation: histogram (3 bins)
Texture
T1. LM filters: mean abs response (15 filters)
T2. LM filters: hist. of maximum responses (15 bins)
Perspective
P1. Long Lines: (num line pixels)/sqrt(area)
P2. Long Lines: % of nearly parallel pairs of lines
P3. Line Intersections: hist. over 8 orientations, entropy
P4. Line Intersections: % right of center
P5. Line Intersections: % above center
P6. Line Intersections: % far from center at 8 orientations
P7. Line Intersections: % very far from center at 8 orientations
P8. Vanishing Points: (num line pixels with vertical VP membership)/sqrt(area)
P9. Vanishing Points: (num line pixels with horizontal VP membership)/sqrt(area)
P10. Vanishing Points: percent of total line pixels with vertical VP membership
P11. Vanishing Points: x-pos of horizontal VP - segment center (0 if none)
P12. Vanishing Points: y-pos of highest/lowest vertical VP wrt segment center
P13. Vanishing Points: segment bounds wrt horizontal VP
P14. Gradient: x, y center of gradient mag. wrt. image center

Slides by Efros

Need Spatial Support



Slides by Efros

Image Segmentation

- Naïve Idea #1: segment the image

  – Chicken & Egg problem

- Naïve Idea #2: <u>multiple</u> segmentations

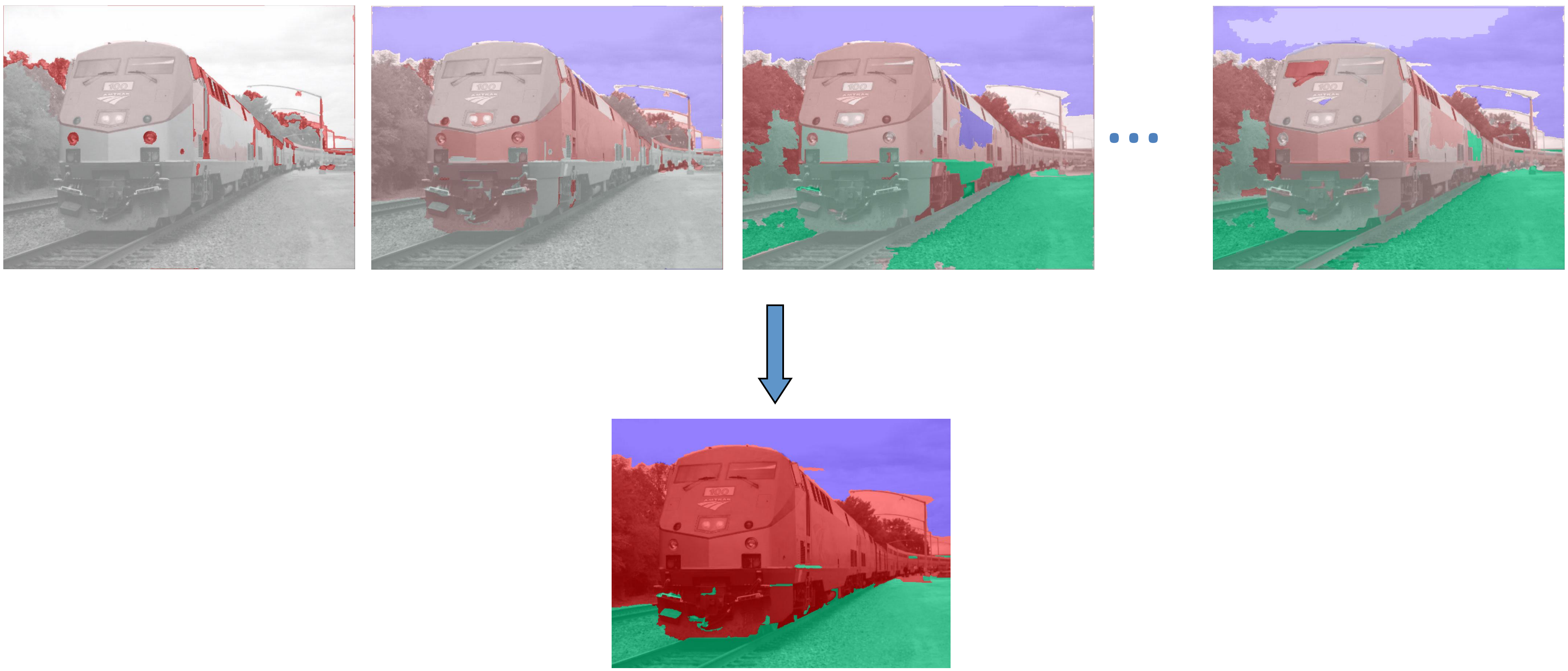
…

  – Decide later which segments are good

Slides by Efros

Image Labeling

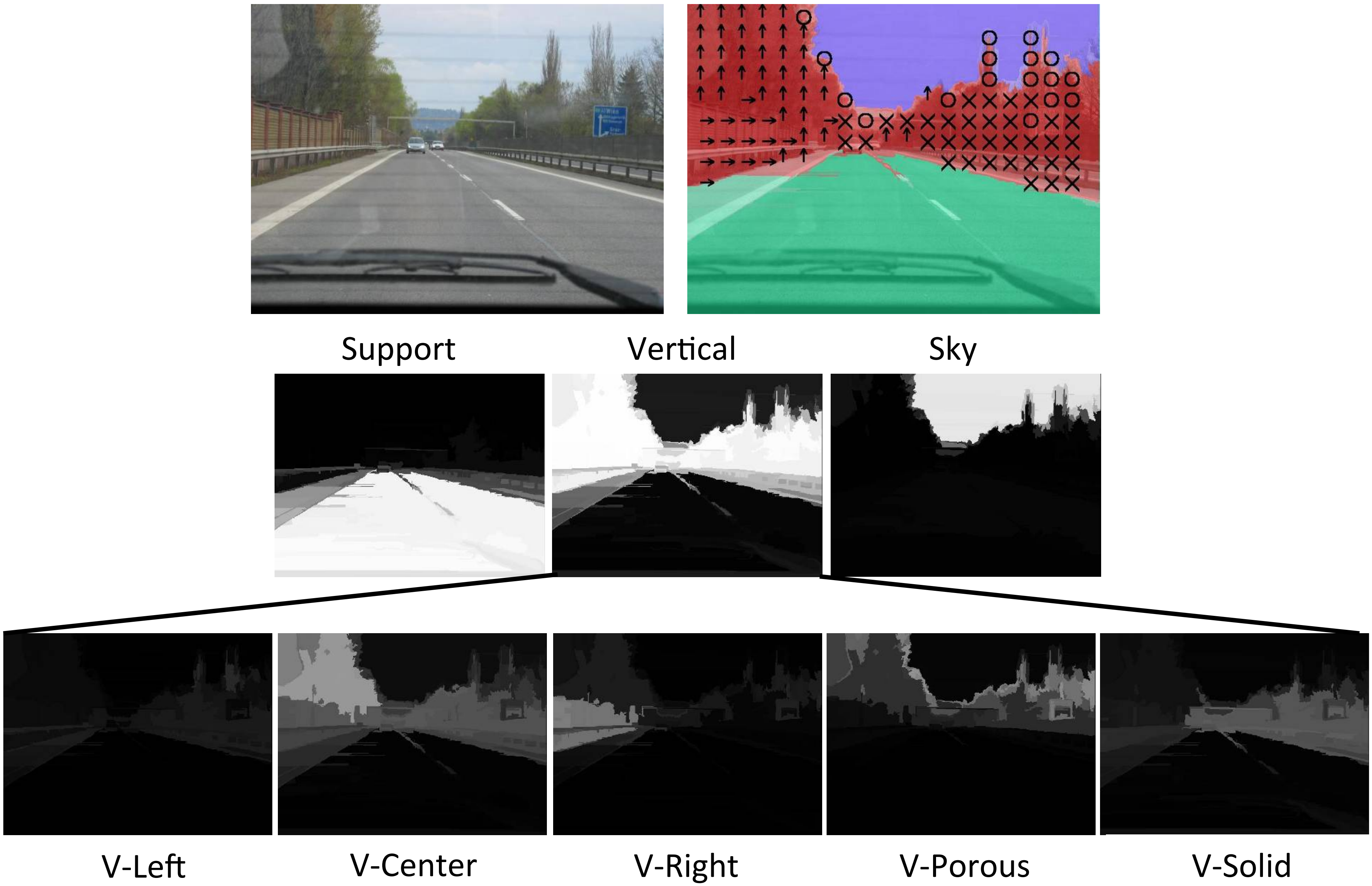
Labeled Segmentations



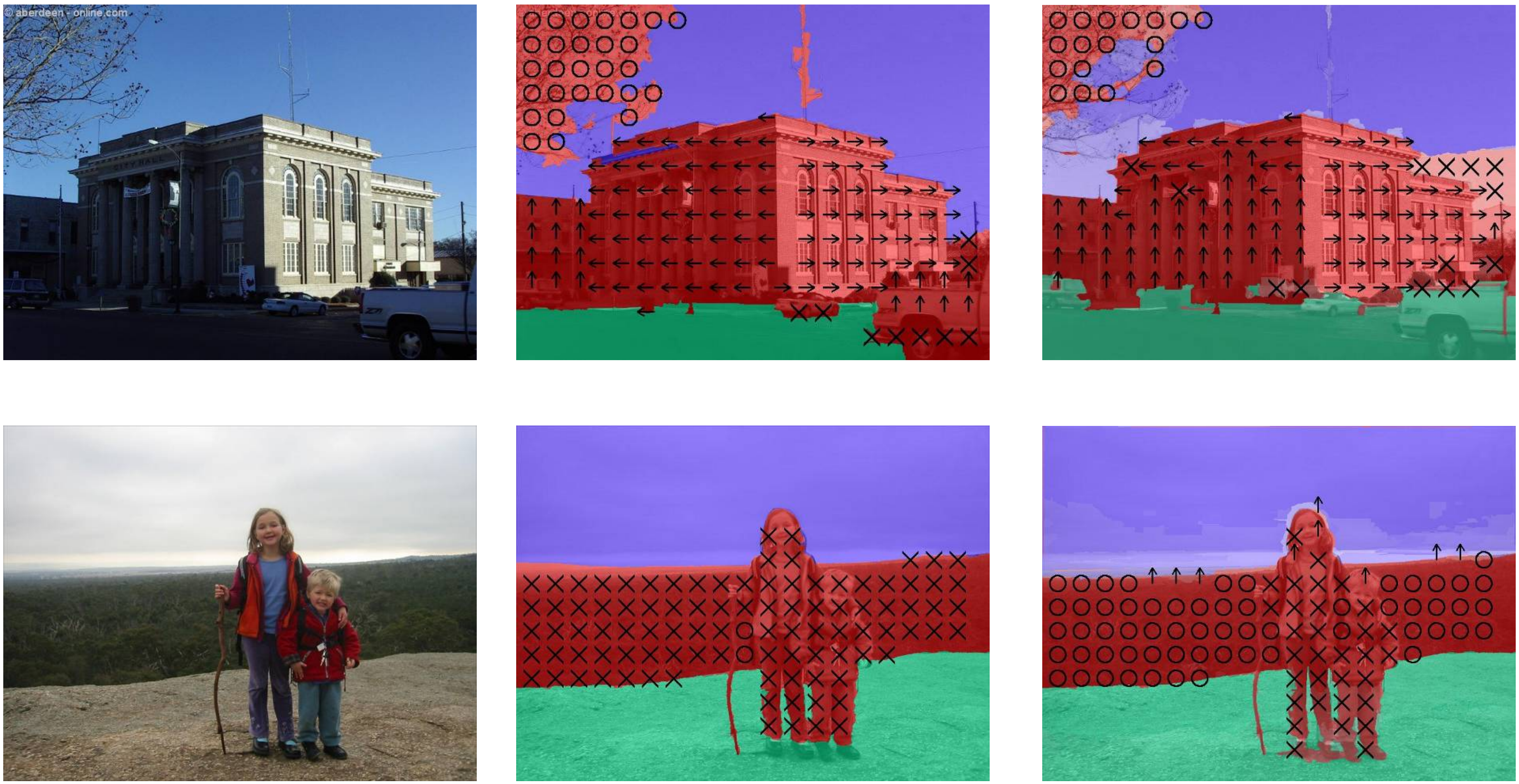
Labeled Pixels

Slides by Efros

No Hard Decisions



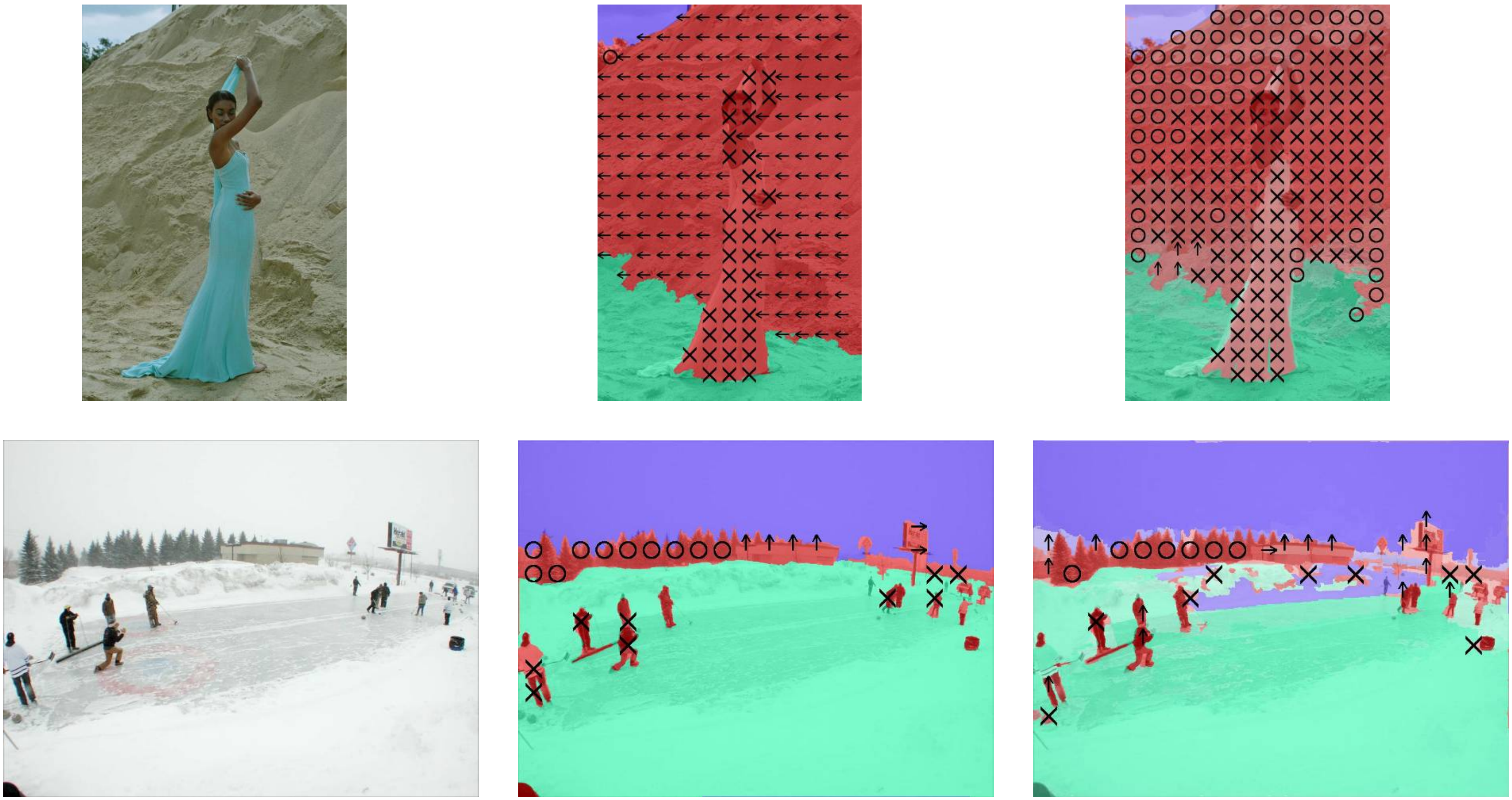
Labeling Results



Input image Ground Truth Our Result

Slides by Efros

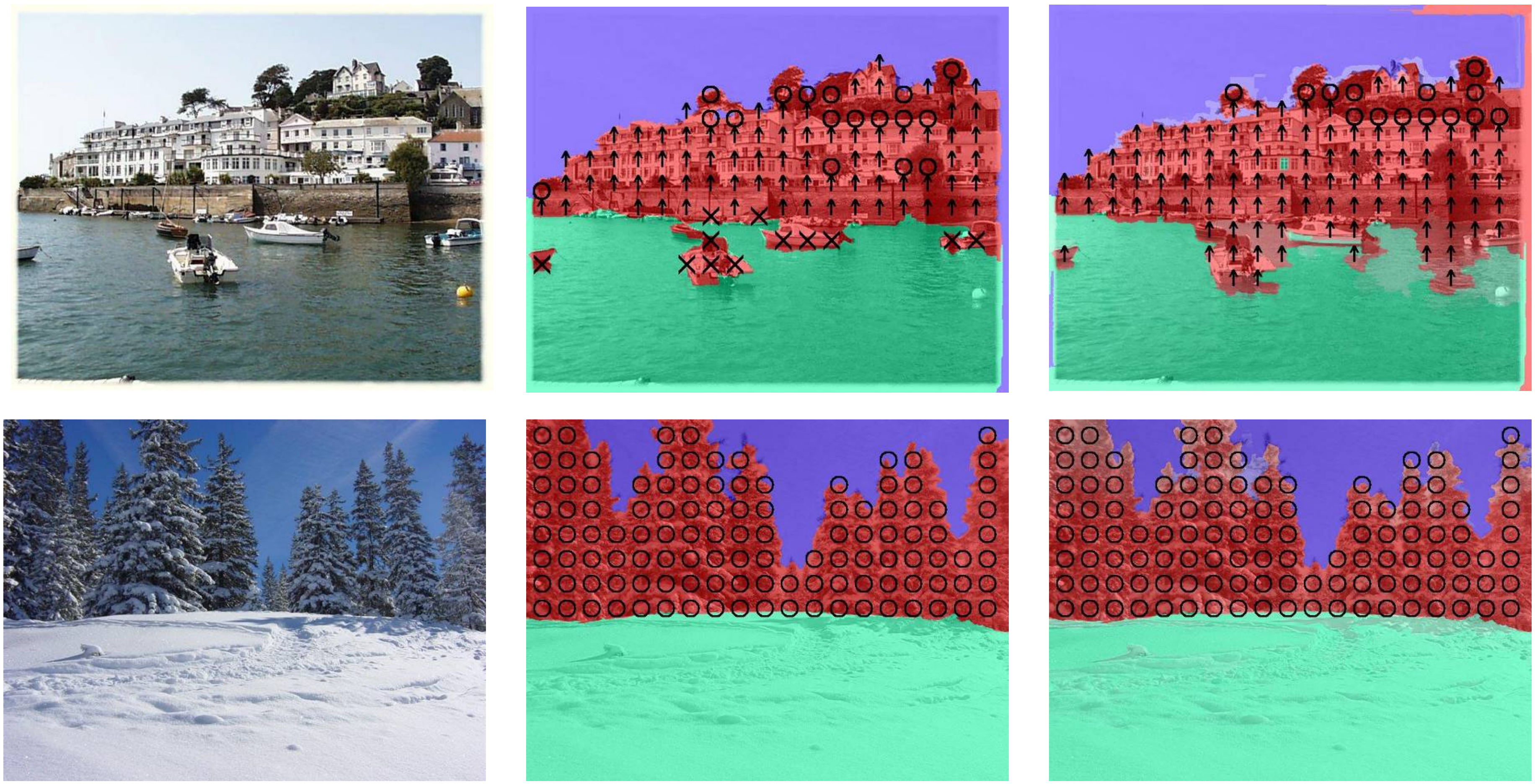
Labeling Results



Input image | Ground Truth | Our Result

Slides by Efros

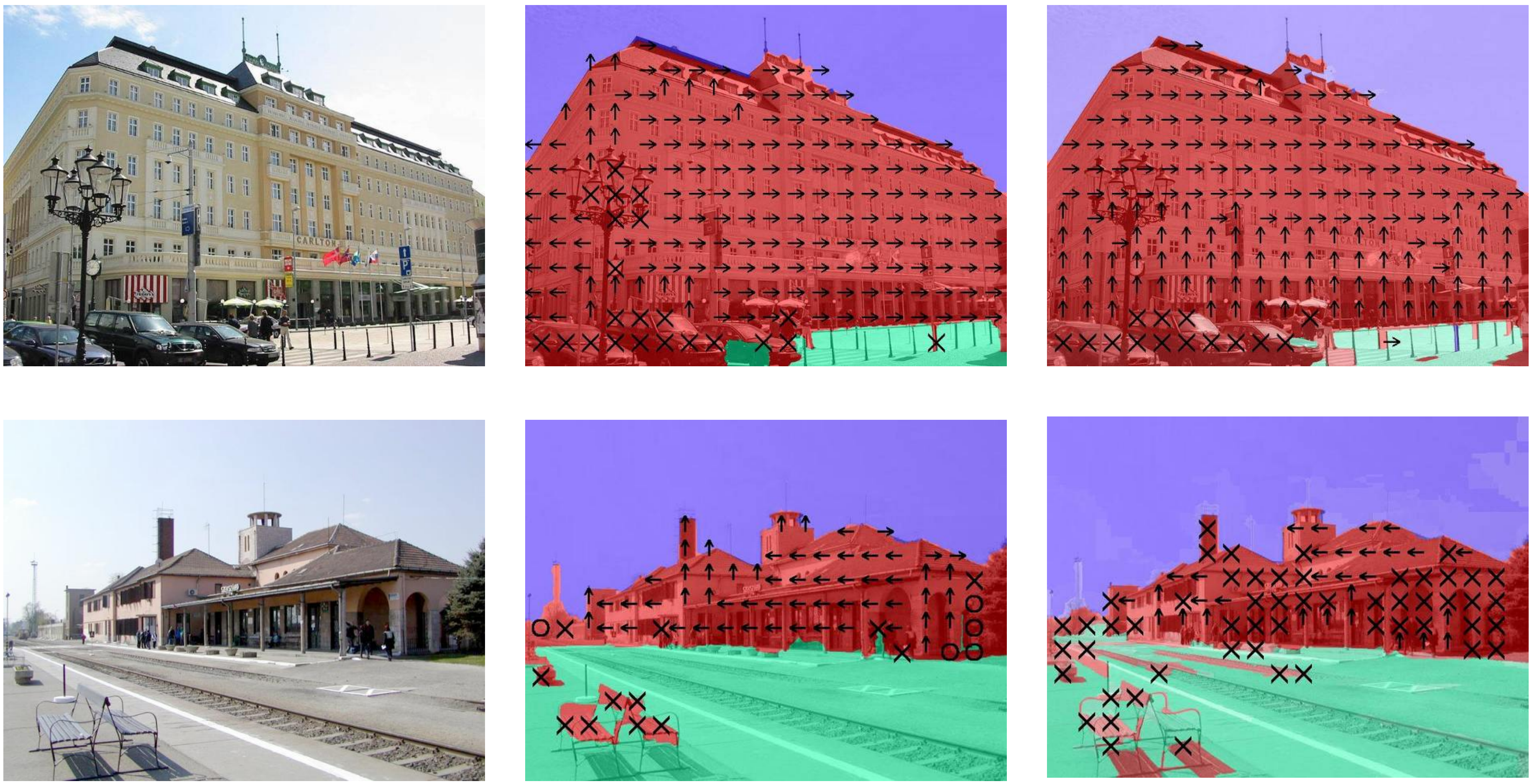
Labeling Results



Input image Ground Truth Our Result

Slides by Efros

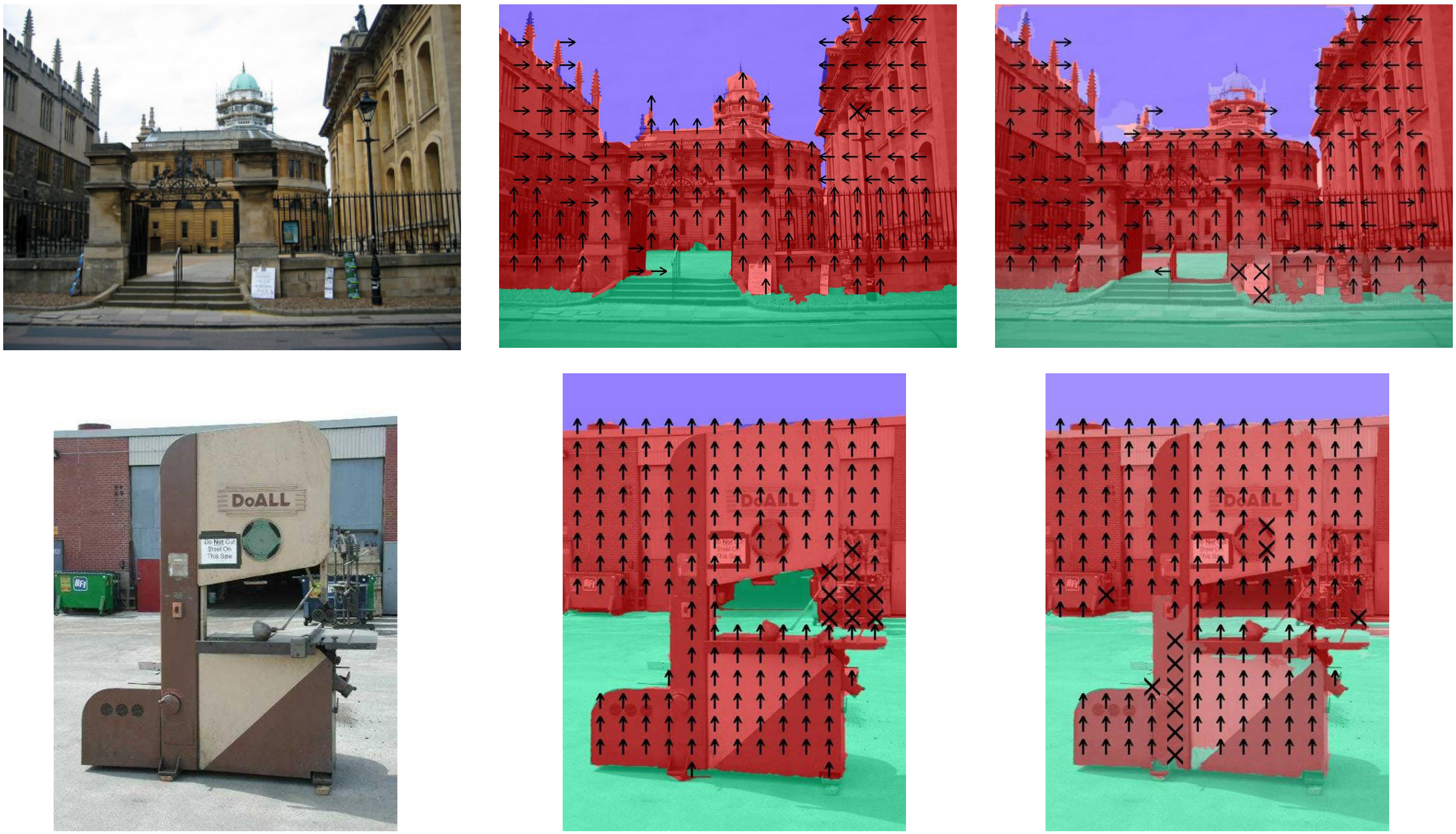
Labeling Results



Input image　　Ground Truth　　Our Result

Slides by Efros

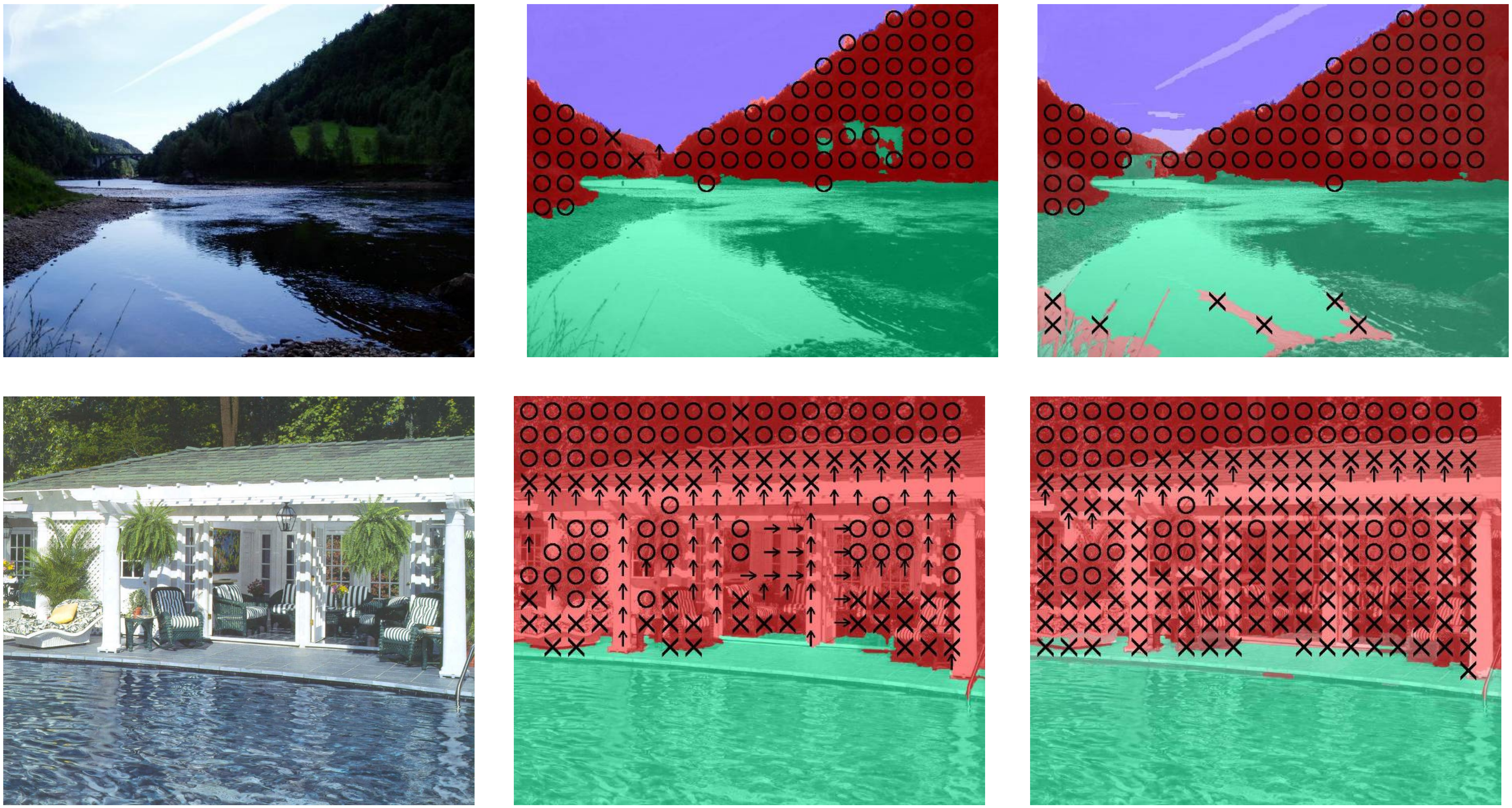
Labeling Results



Input image Ground Truth Our Result

Slides by Efros

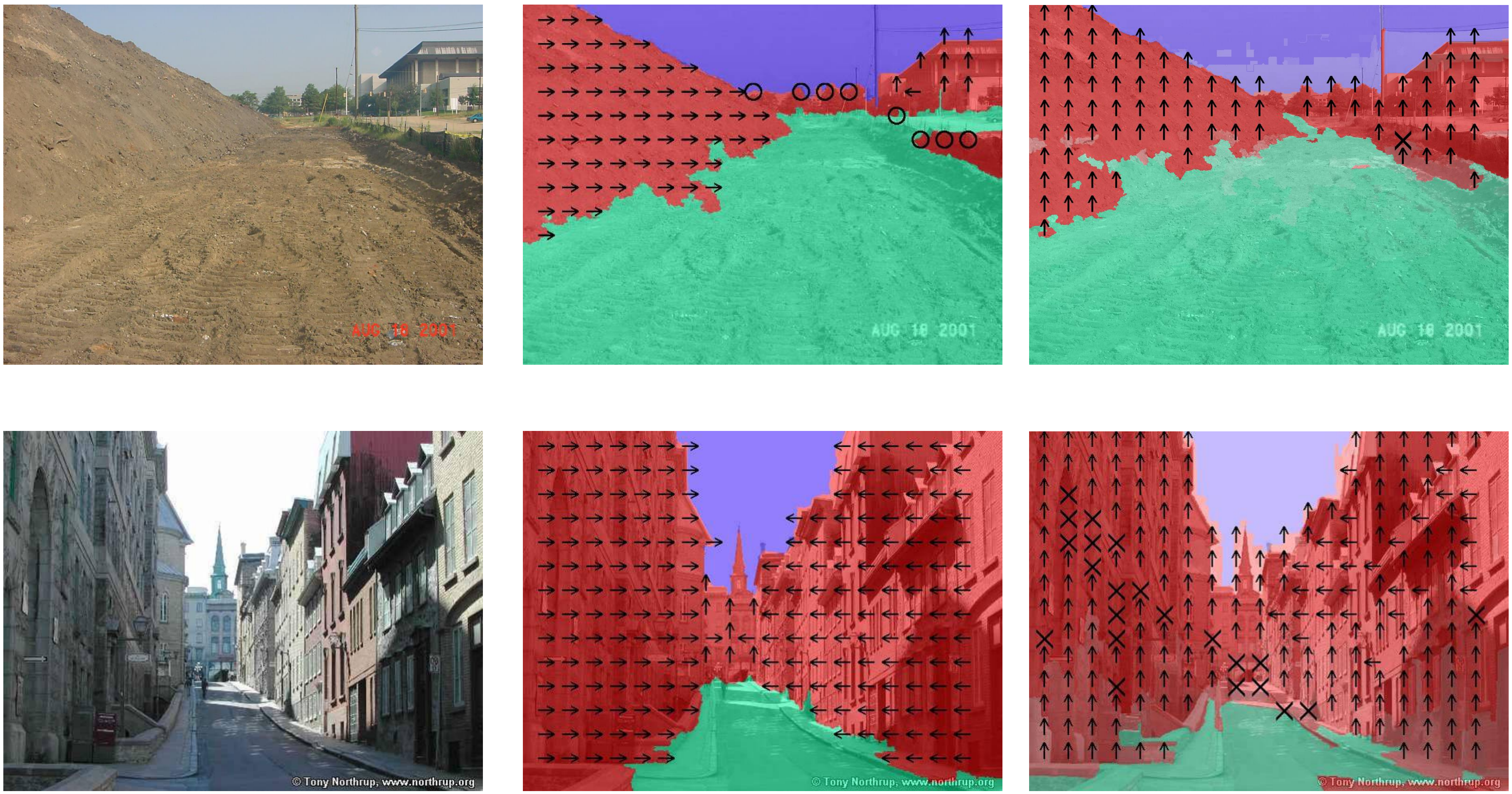
Labeling Results



Input image Ground Truth Our Result

Slides by Efros

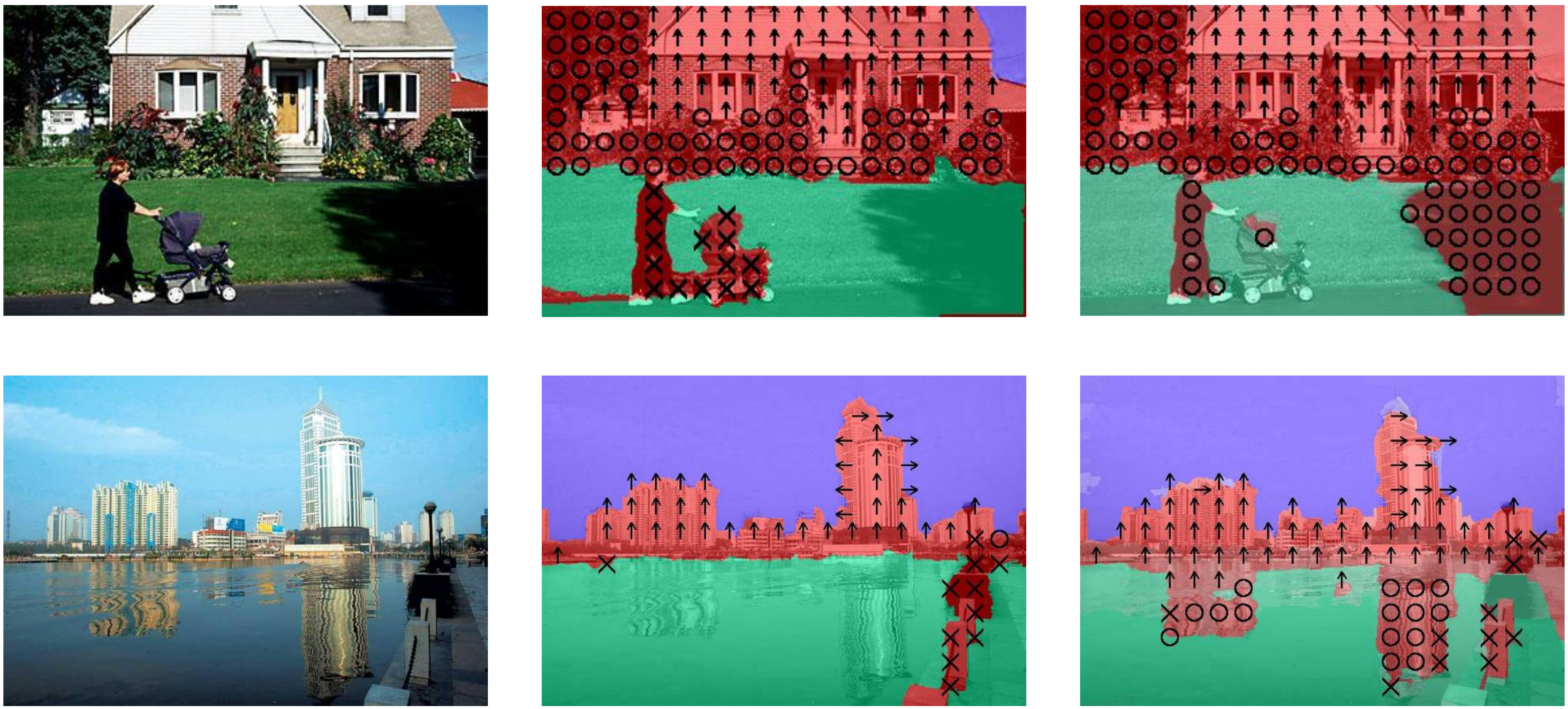
Labeling Results



Input image　　　Ground Truth　　　Our Result

Slides by Efros

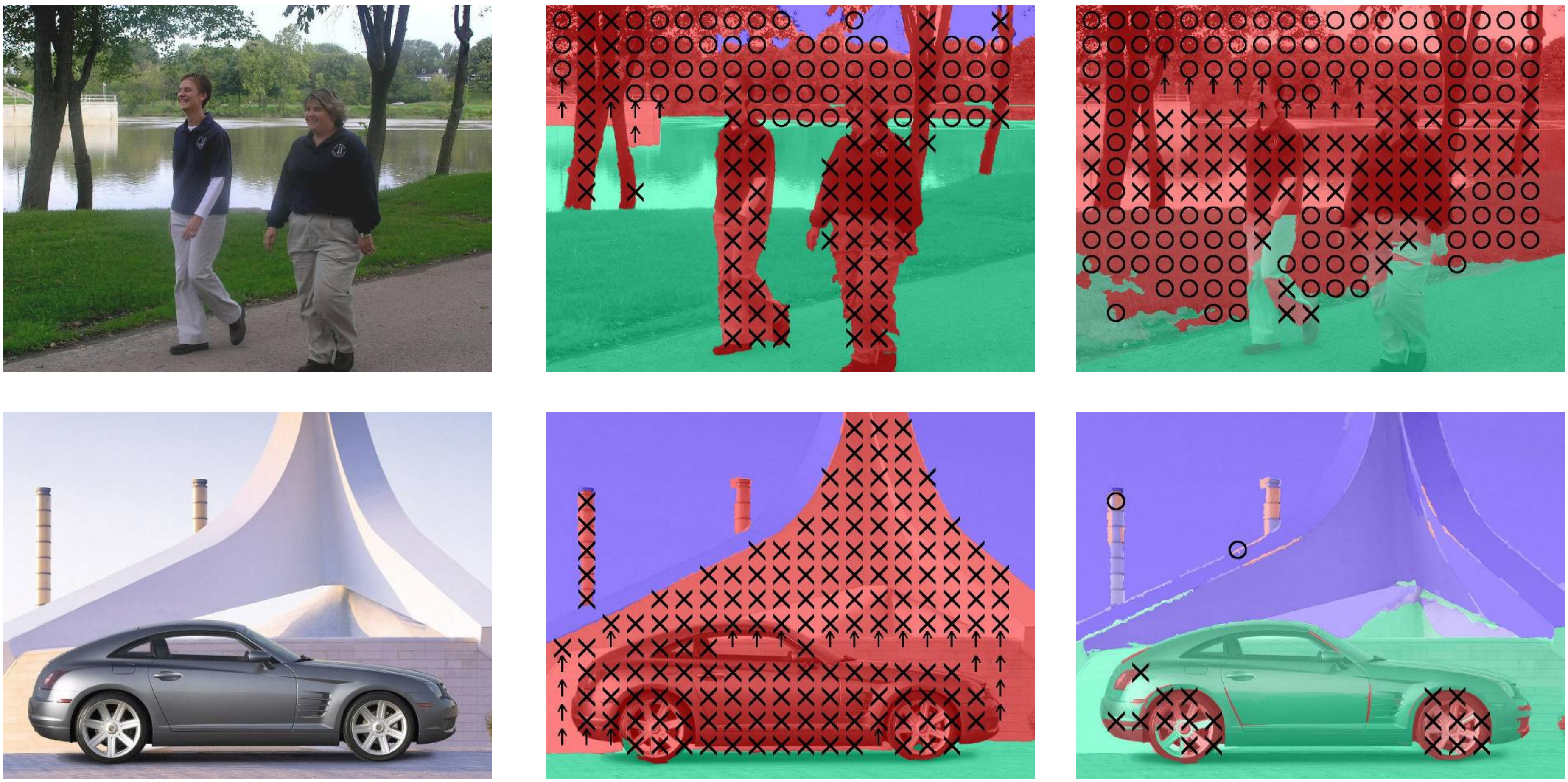
Some Failures



Input image Ground Truth Our Result

Slides by Efros

Catastrophic Failures



Input image Ground Truth Our Result

Slides by Efros

Average Accuracy

Main Class: 88.1%

Subclasses: 61.5%

Main Class			
	Support	Vertical	Sky
Support	**0.84**	0.15	0.00
Vertical	0.09	**0.90**	0.02
Sky	0.00	0.10	**0.90**

Vertical Subclass					
	Left	Center	Right	Porous	Solid
Left	**0.37**	0.32	0.08	0.09	0.13
Center	0.05	**0.56**	0.12	0.16	0.12
Right	0.02	0.28	**0.47**	0.13	0.10
Porous	0.01	0.07	0.03	**0.84**	0.06
Solid	0.04	0.20	0.04	0.17	**0.55**

Better Spatial Support Useful?

Method	Main	Sub
Pixels	82.1	44.3
Superpixels	86.2	53.5
Single Segmentation	86.2	56.6
Multiple Segmentations	**88.1**	**61.5**
Ground Truth Segmentation	95.1	71.5

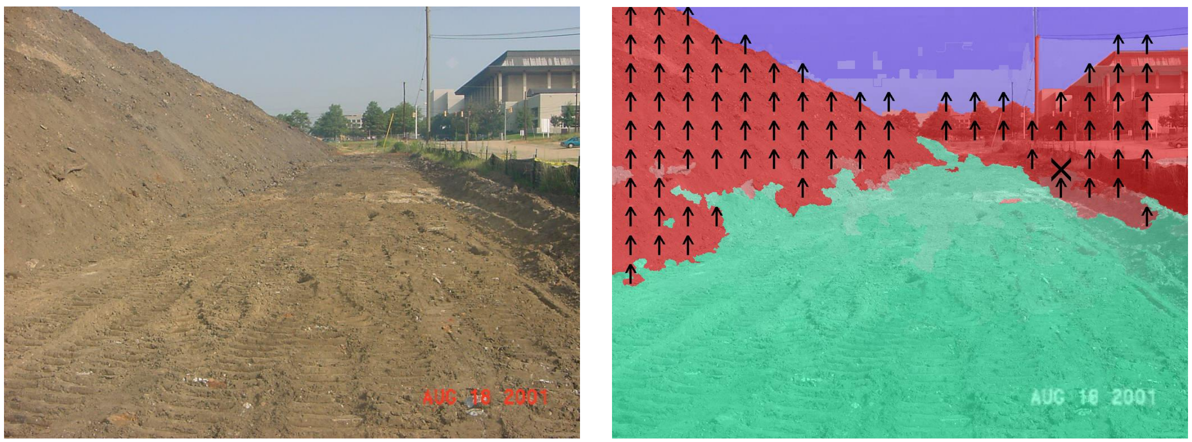
Table 4. Average accuracy (percent of correctly labeled image pixels) of methods using varying levels of spatial support.

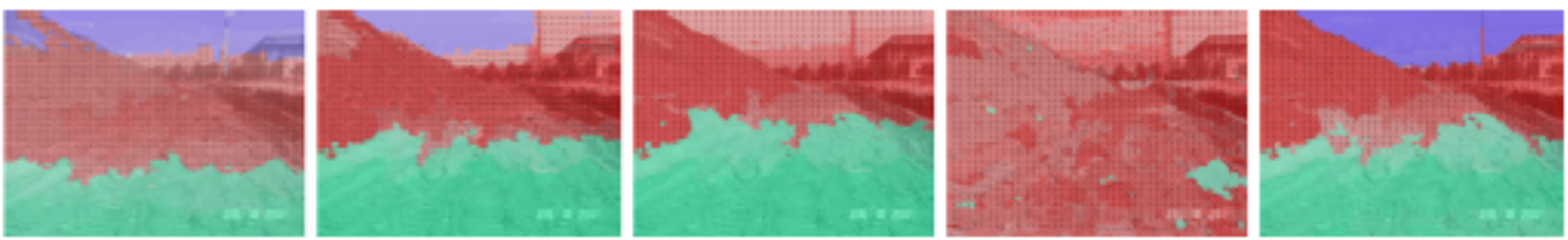
Slides by Efros

Do all features help?

Importance of Different Feature Types				
	Color	Texture	Loc/Shape	Geometry
Main	6%	2%	16%	2%
Sub	6%	2%	8%	7%

Drop in accuracy due to remove of each type of feature





(c) Loc Only (d) No Color (e) No Texture (f) No Loc/Shp (g) No Geom

Slides by Efros

How robust is it?

Input Labels Input Labels Input Labels

Figure 20. Results on paintings of outdoor scenes. Although the system is trained only on real images, it can often generalize to very different settings.

Input Ground Truth Labels Input Ground Truth Labels

Figure 21. Results on indoor scenes.

Slides by Efros

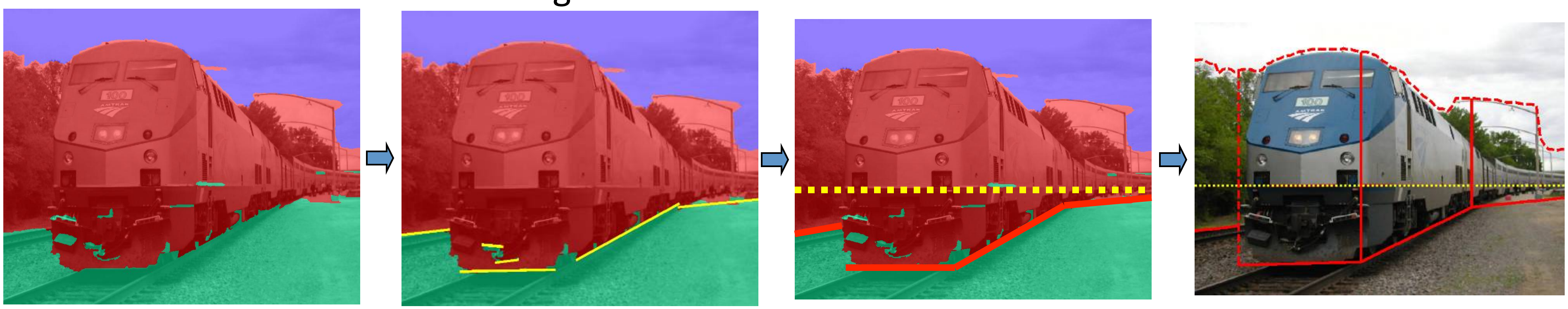
Automatic Photo Popup

Labeled Image

Fit Ground-Vertical Boundary with Line Segments

Form Segments into Polylines

Cut and Fold

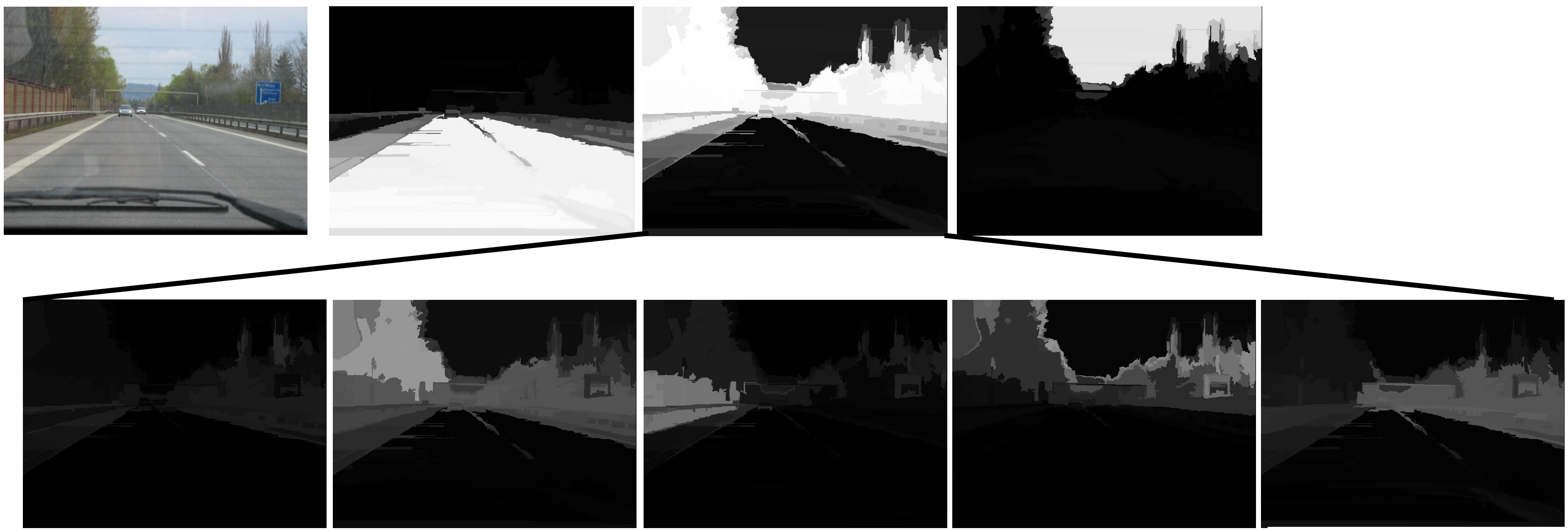


Final Pop-up Model

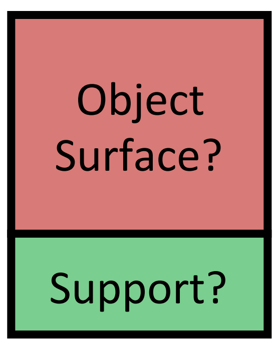
[Hoiem Efros Hebert 2005]

Surface Estimation

Image Support Vertical Sky



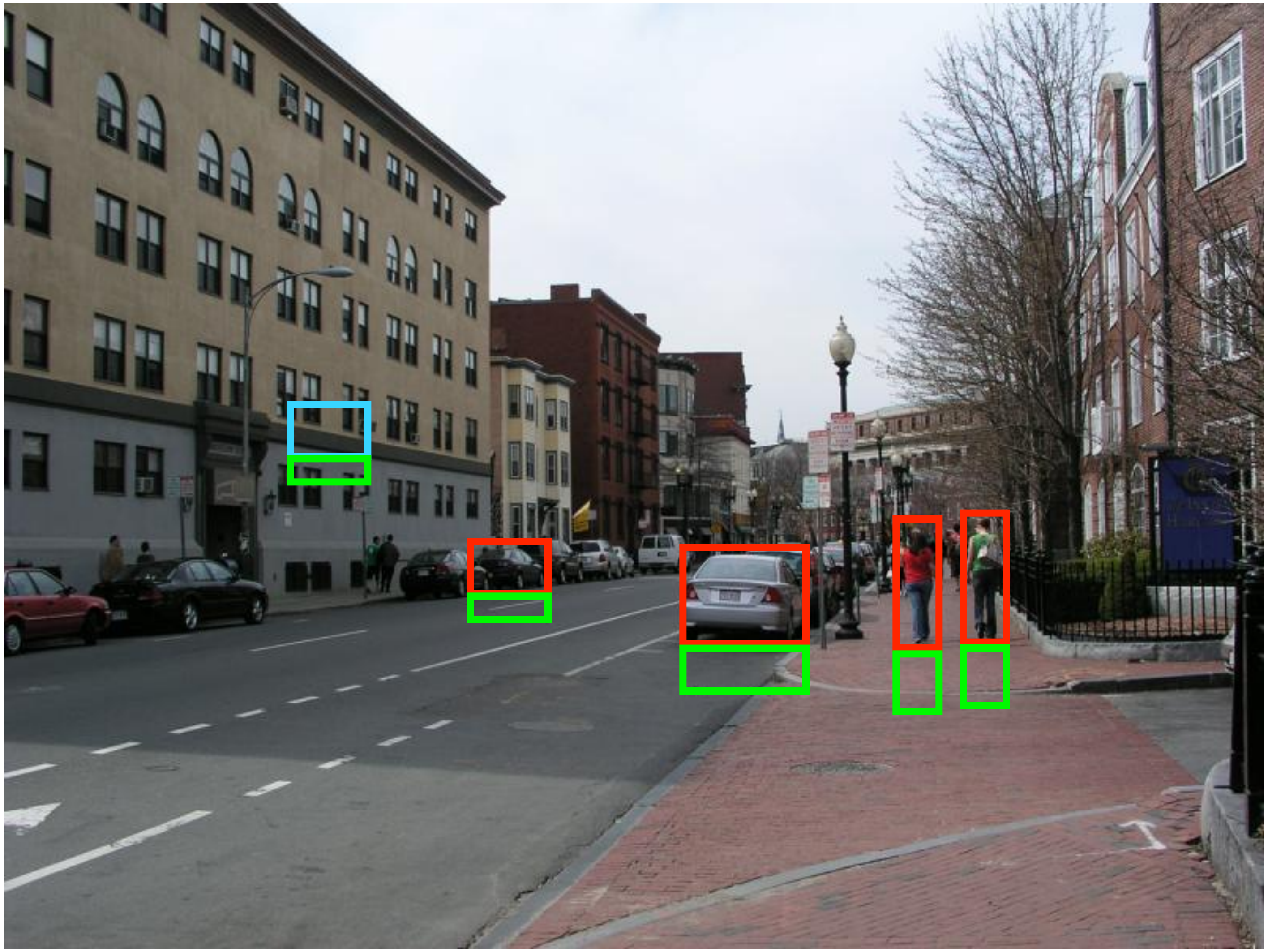
V-Left V-Center V-Right V-Porous V-Solid



[Hoiem, Efros, Hebert ICCV 2005]

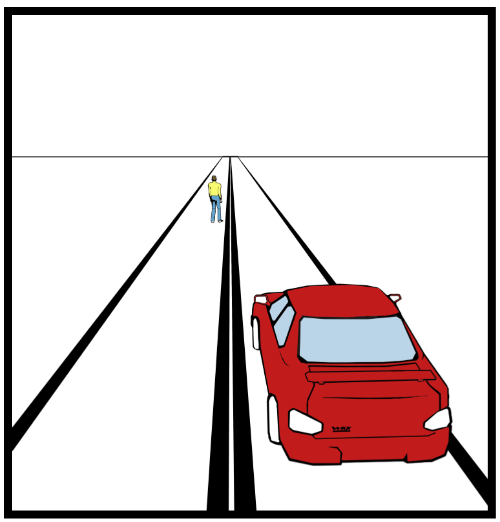
Slide by Derek Hoiem

Object Support

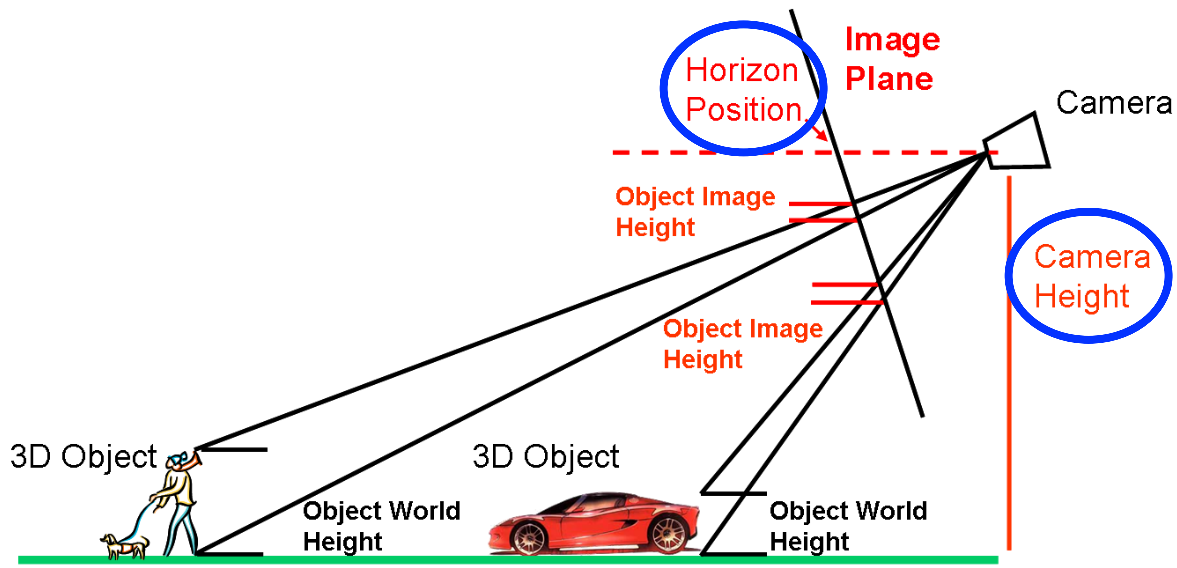


Slide by Derek Hoiem

3d Scene Context



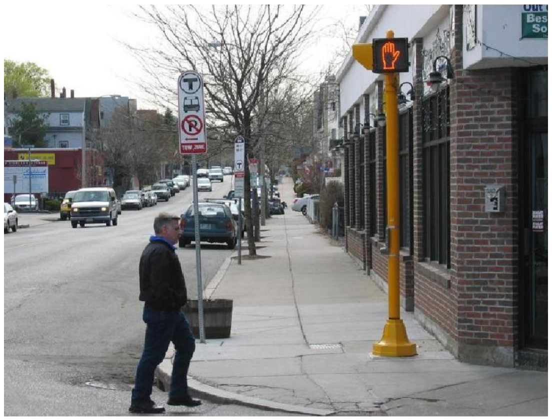
Image



World

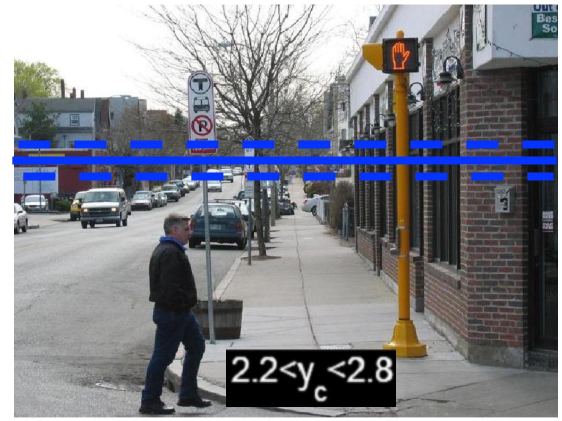
Hoiem, Efros, Hebert ICCV 2005

What does surface and viewpoint say about objects?

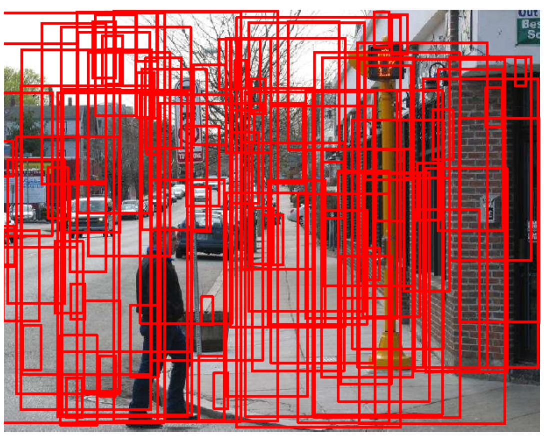


Image

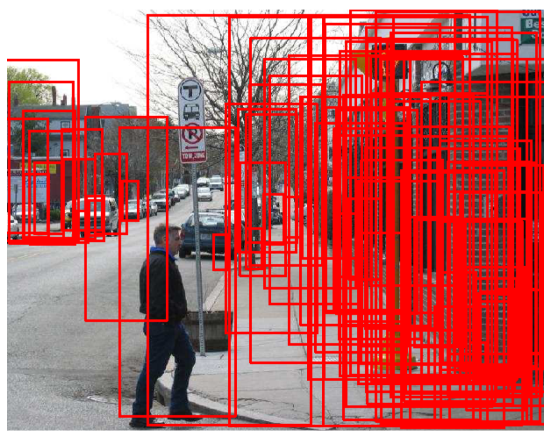
P(surfaces)



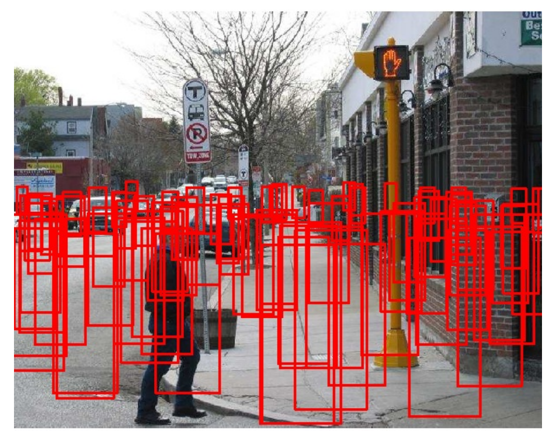
P(viewpoint)



P(object)



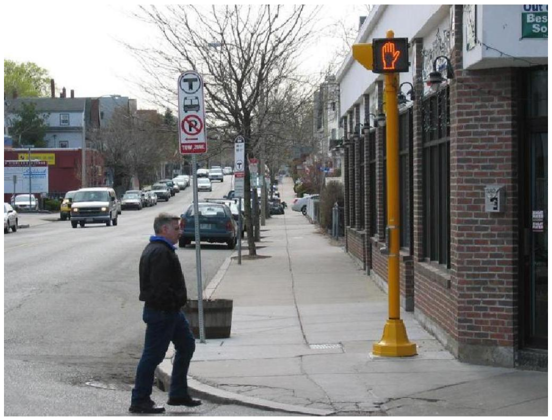
P(object | surfaces)



P(object | viewpoint)

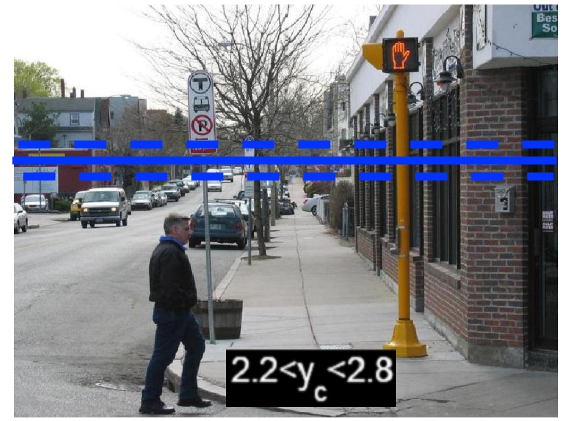
Slide by D. Hoiem

What does surface and viewpoint say about objects?

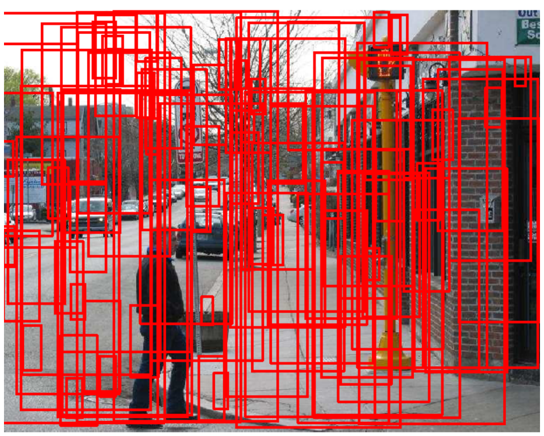


Image

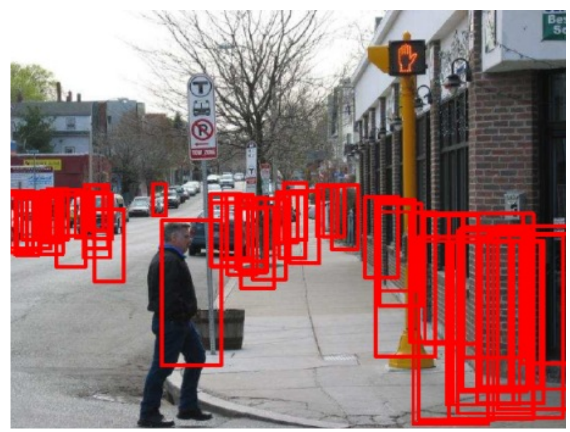
P(surfaces)



P(viewpoint)



P(object)

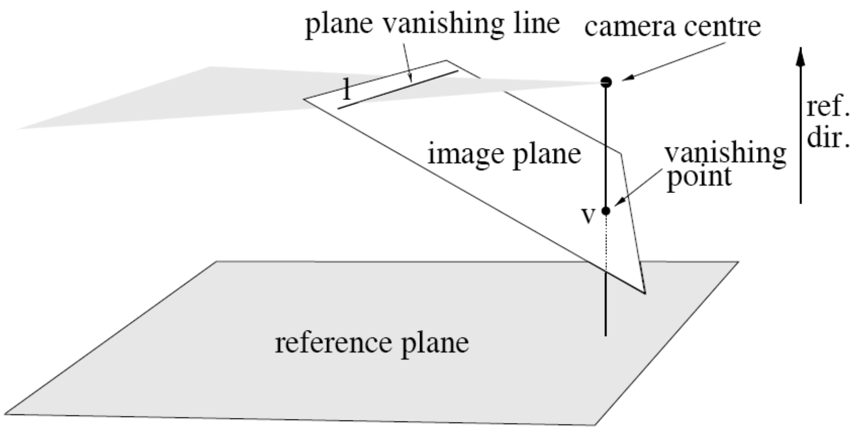


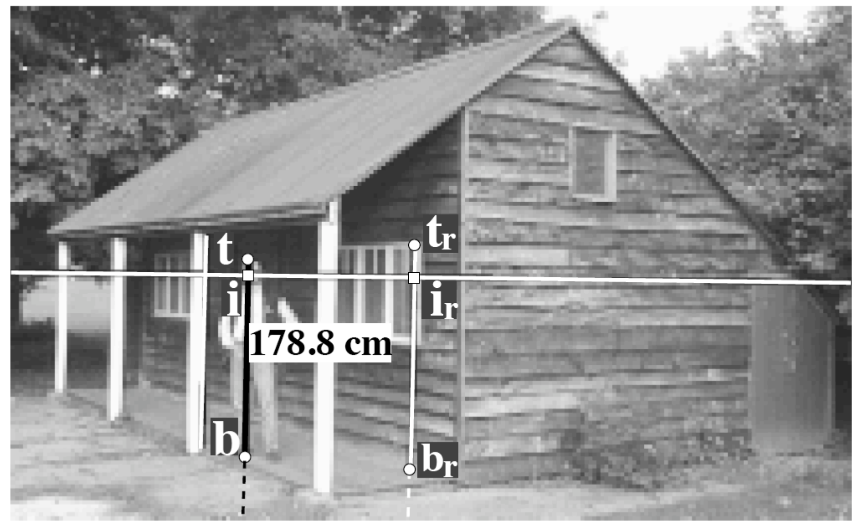
P(object | surfaces, viewpoint)

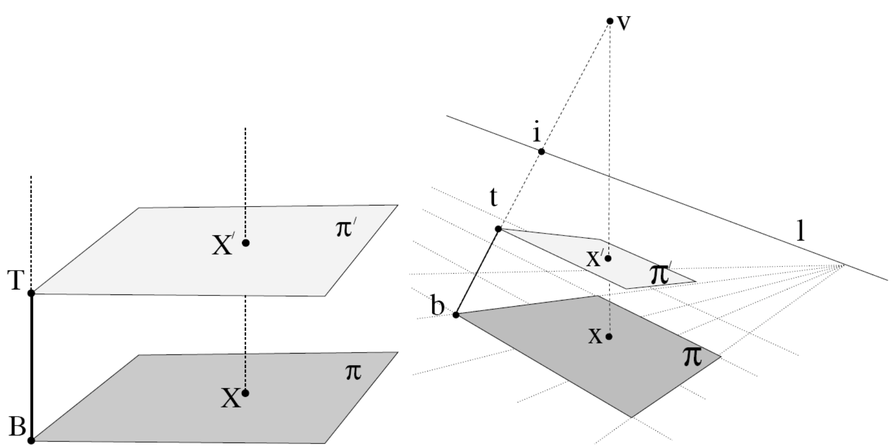
Slide by Derek Hoiem

Single view metrology

Criminisi, et al. 1999







Need to recover:

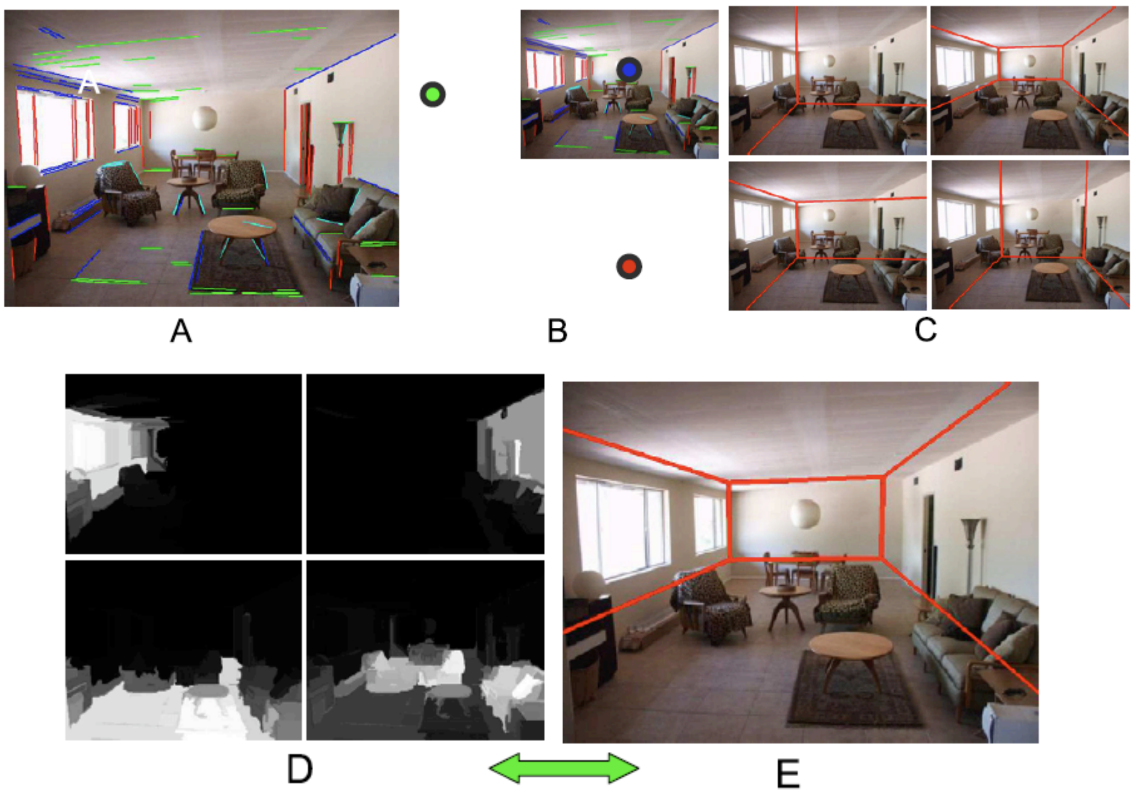
- Ground plane
- Reference height
- Horizon line
- Where objects contact the ground

See also the book by Hartley and Zisserman, 2004

Recovering spatial layout of indoor rooms from a single image

- Recover approximate camera calibration and orientation from three orthogonal directions.
- Assume a room can be modeled as a single 3D box.

Overview



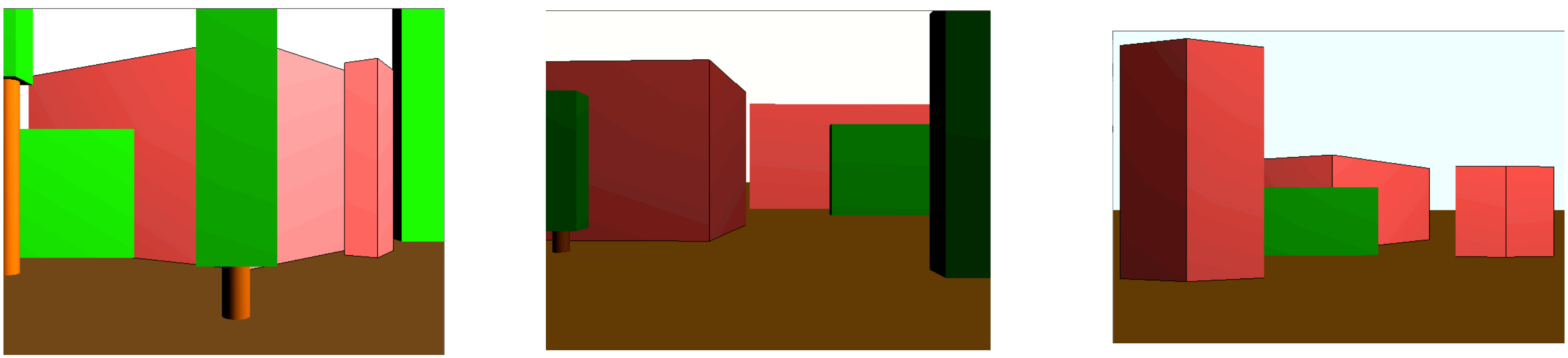
Example output

Varsha Hedau, Derek Hoiem, David Forsyth, "Recovering the Spatial Layout of Cluttered Rooms," in the Twelfth IEEE International Conference on Computer Vision, 2009.

Modeling outdoor scenes as blocks

- Model an outdoor scene from a single image as a collection of blocks (cuboids)
- Include physical constraints (support, stability, materials)

Input Images



Toy Blocks World Rendering

Gupta et al. , “Blocks world revisited: image understanding using qualitative geometry and machanics”, ECCV 2010