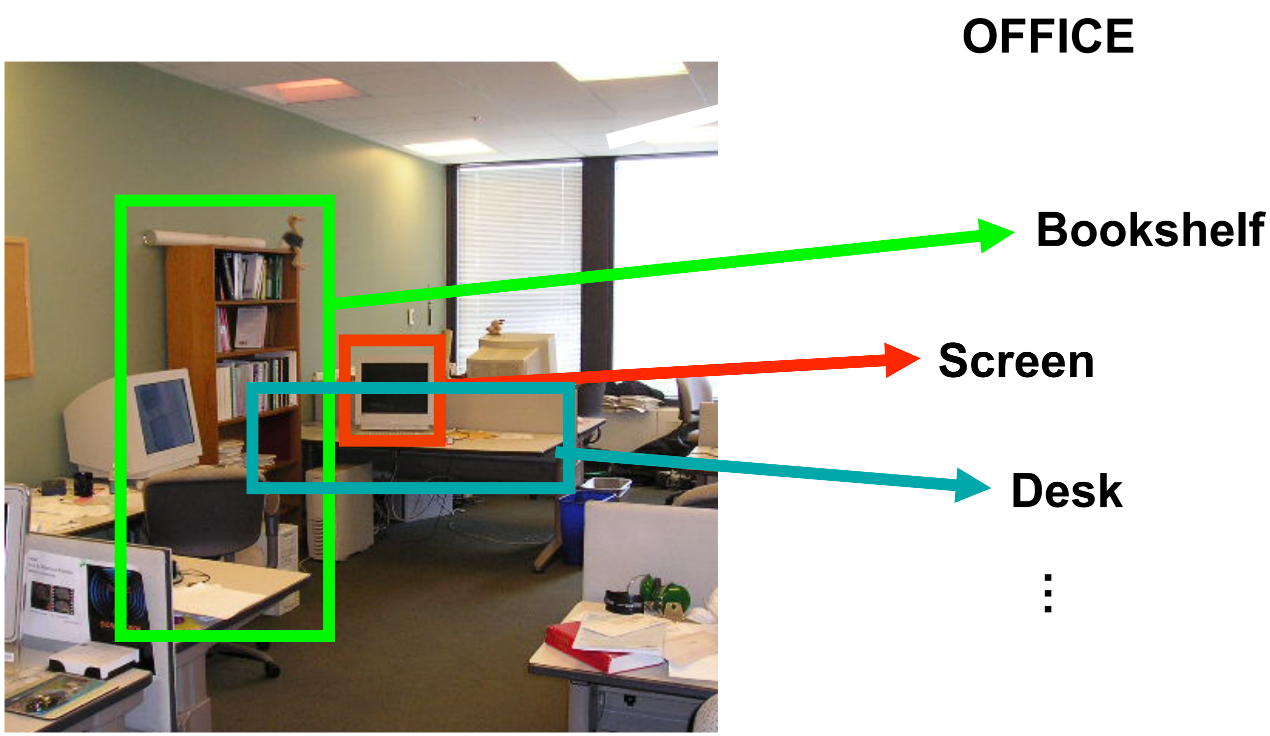
# Understanding Visual Scenes


Antonio Torralba

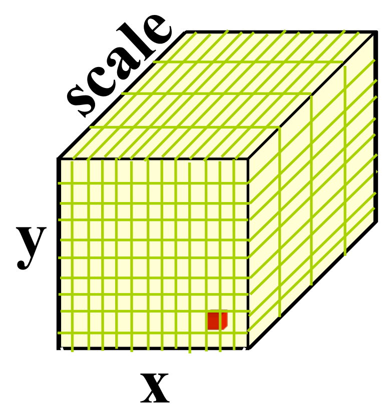
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Department of Electrical Engineering and Computer Science

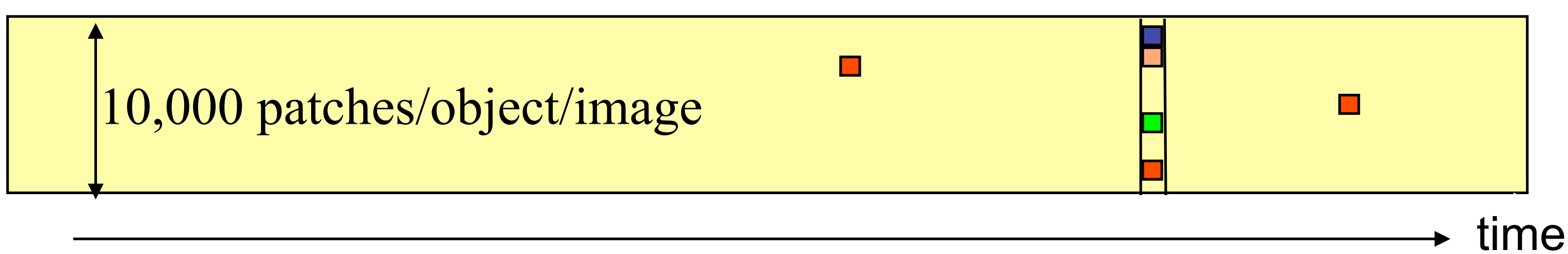
# A computer vision goal

Recognize many different objects under many viewing conditions in unconstrained settings.



# Why is this hard?

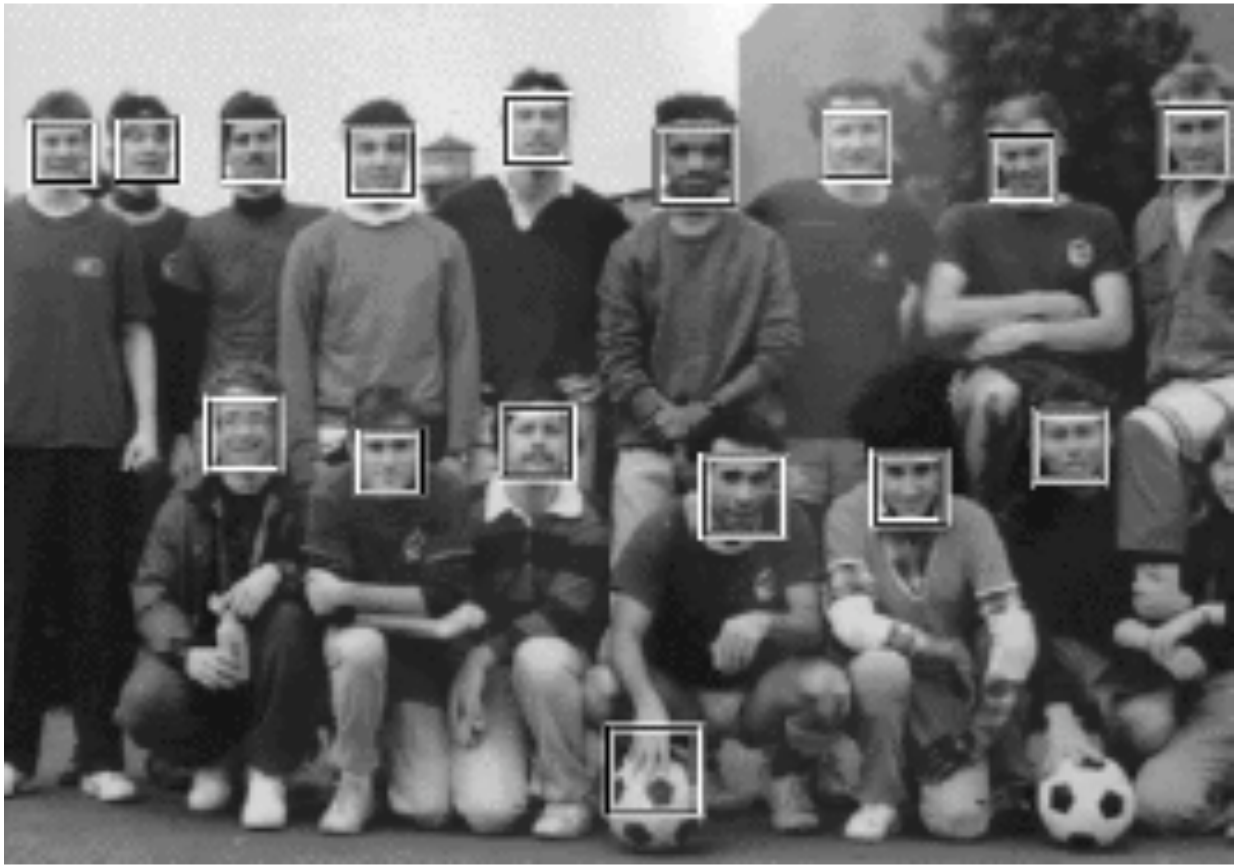



Plus, we want to do this for ~ 1000 objects

1,000,000 images/day

# The face detection age



- 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)
- ….

# “Head in the coffee beans problem”

Can you find the head in this image?

# “Head in the coffee beans problem”

Can you find the head in this image?

# “Head in the coffee beans problem”

Can you find the head in this image?

# Some symptoms of standard approaches

# Just objects is not enough

**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?

**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?


1. chair, 2. table, 3. road, 4. road, 5. table, 6. car, 7. keyboard.

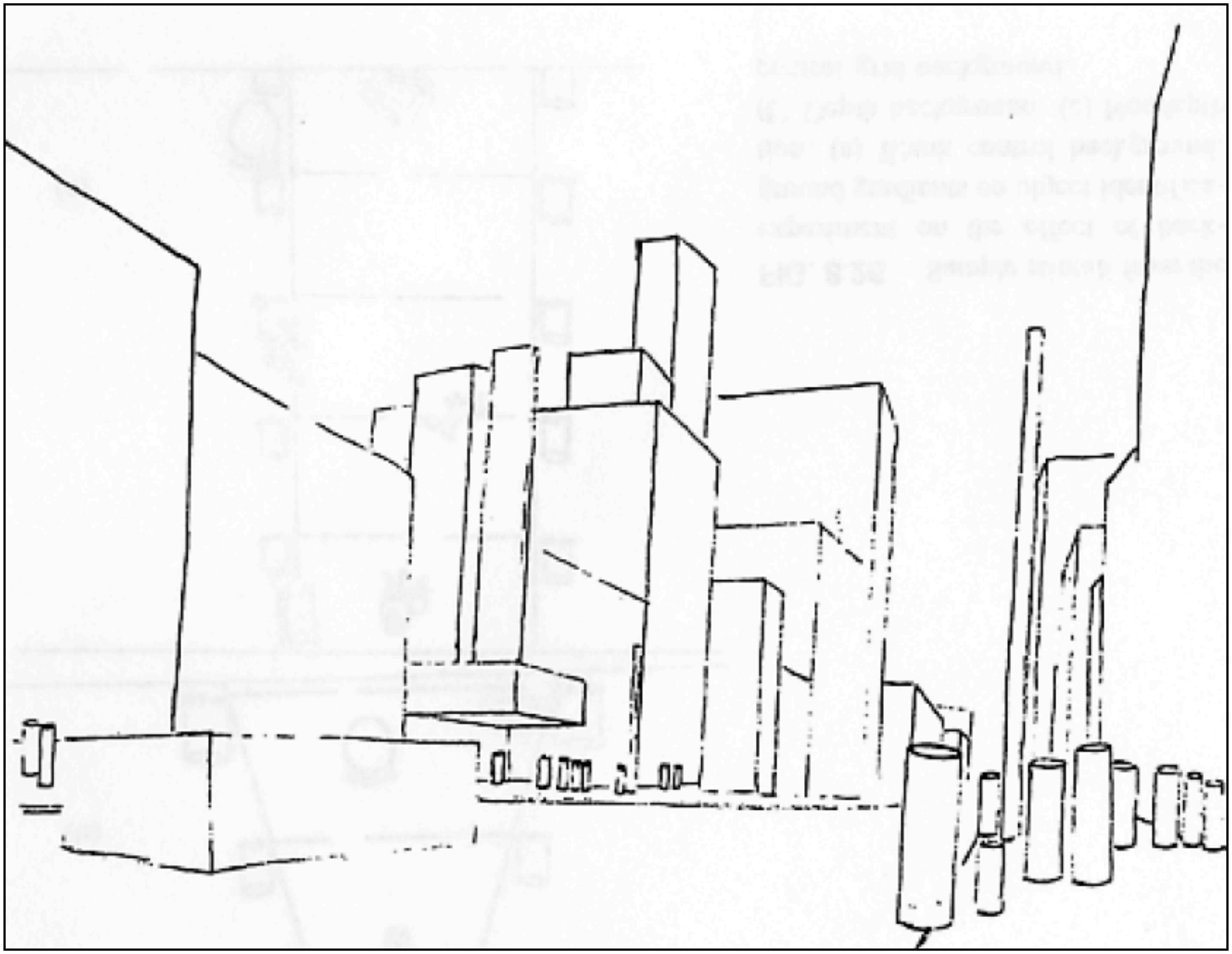
# The importance of context

- 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 |

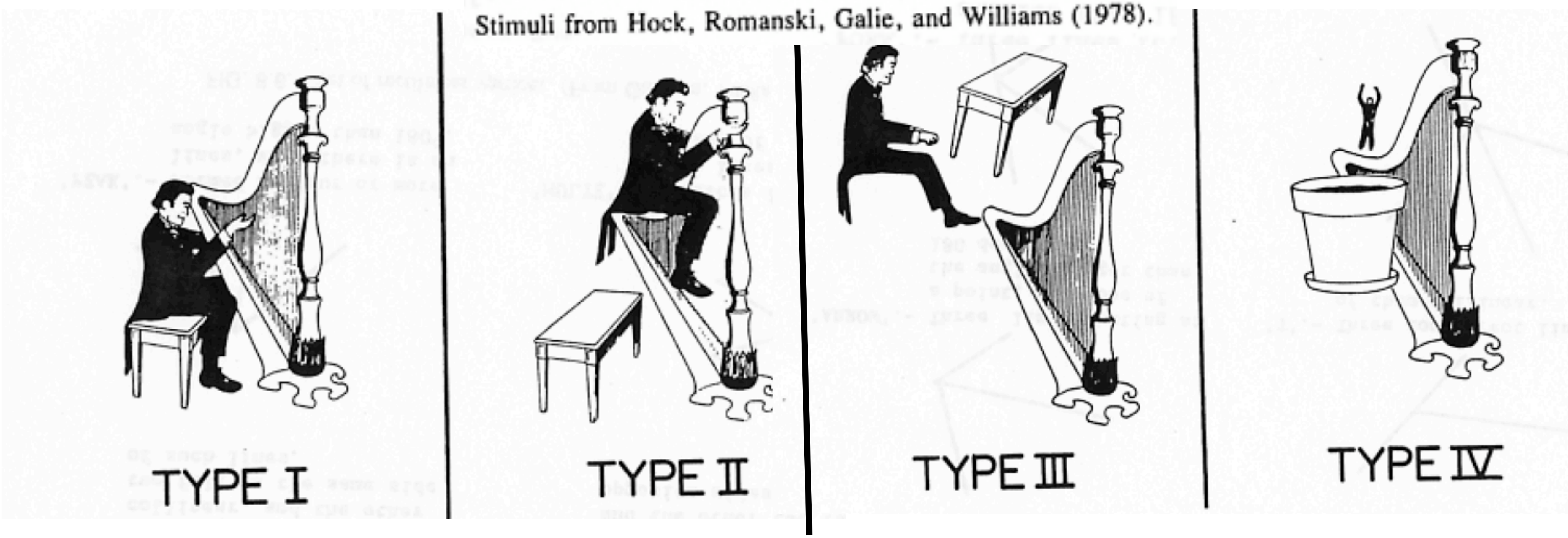
# Humans make extensive use of contextual visual information



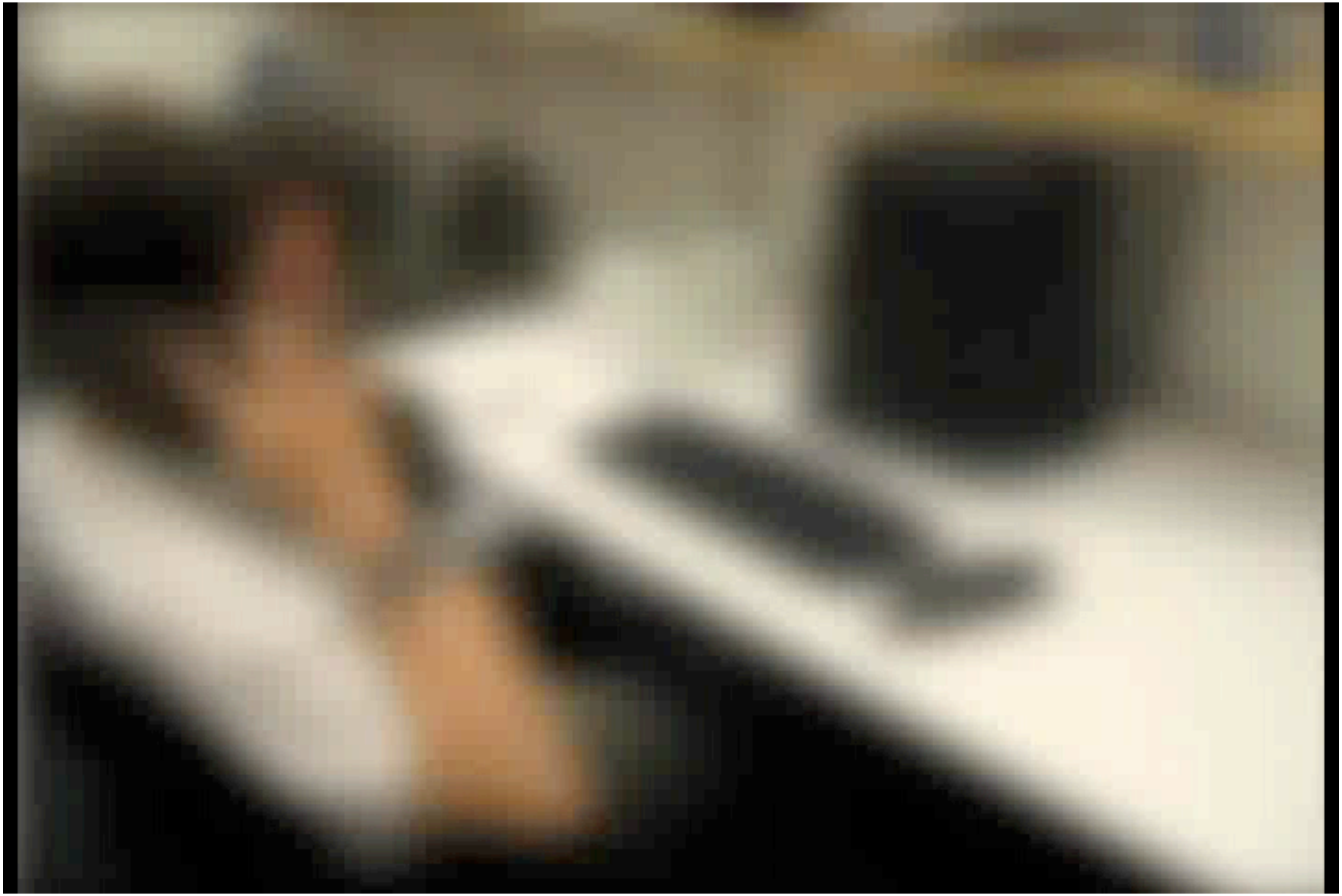
Mezzanotte & Biederman, 1980

# Objects and Scenes

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




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.



# Collecting datasets

# Human vision

•Many input modalities

•Active

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

# Robot vision

•Many poor input modalities

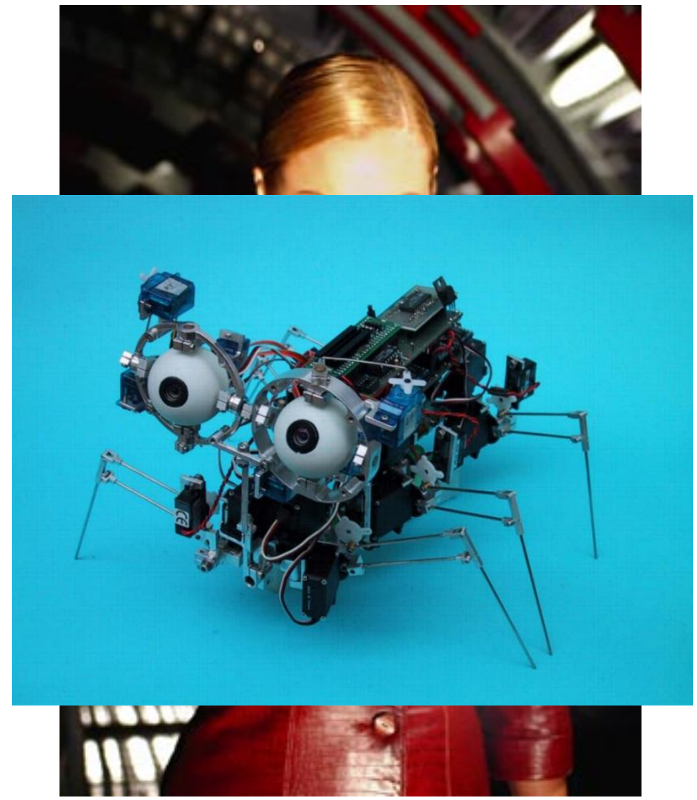
•Active, but it does not go far

# Internet vision

•Many input modalities

•It can reach everywhere

•Tons of data



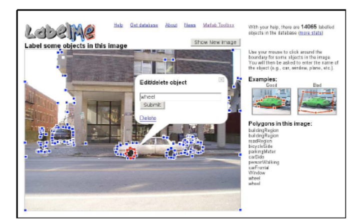
# Collecting datasets (towards $10^{6-7}$ examples)

- ESP game (CMU)

Luis Von Ahn and Laura Dabbish 2004


- LabelMe (MIT)

Russell, Torralba, Freeman, 2005

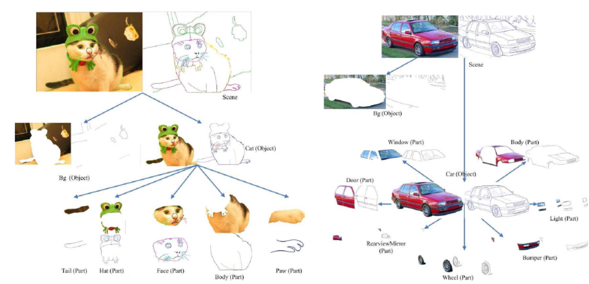


- StreetScenes (CBCL-MIT)

Bileschi, Poggio, 2006

- WhatWhere (Caltech)

Perona et al, 2007



- PASCAL challenge

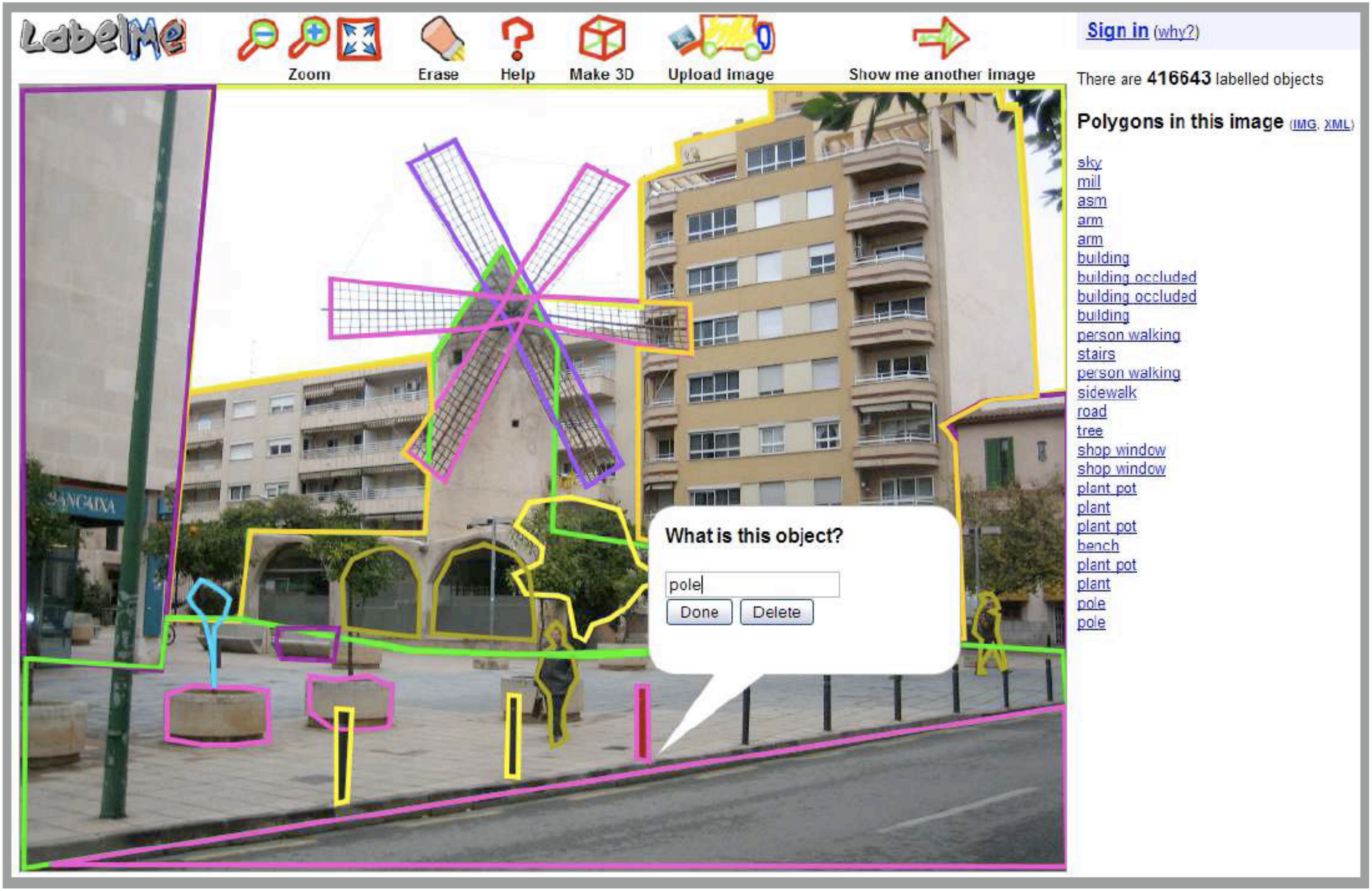
2006, 2007

- Lotus Hill Institute

Song-Chun Zhu et al, 2007

- 80 million images

Torralba, Fergus, Freeman, 2007




http://labelme.csail.mit.edu

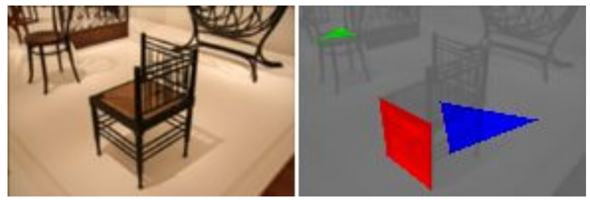
B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, IJCV 2008

# Extreme labeling

# The other extreme of extreme labeling

… things do not always look good…

# Testing



**Most common labels:**

test

adksdsa

woiieiie

…

# Sophisticated testing

**Most common labels:**

Star

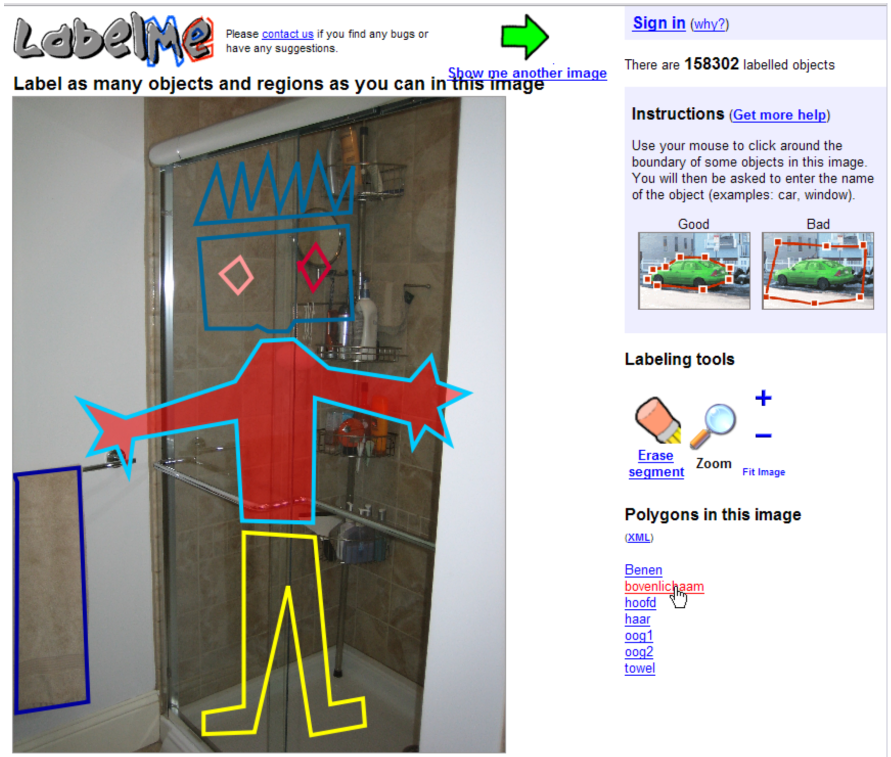
Square

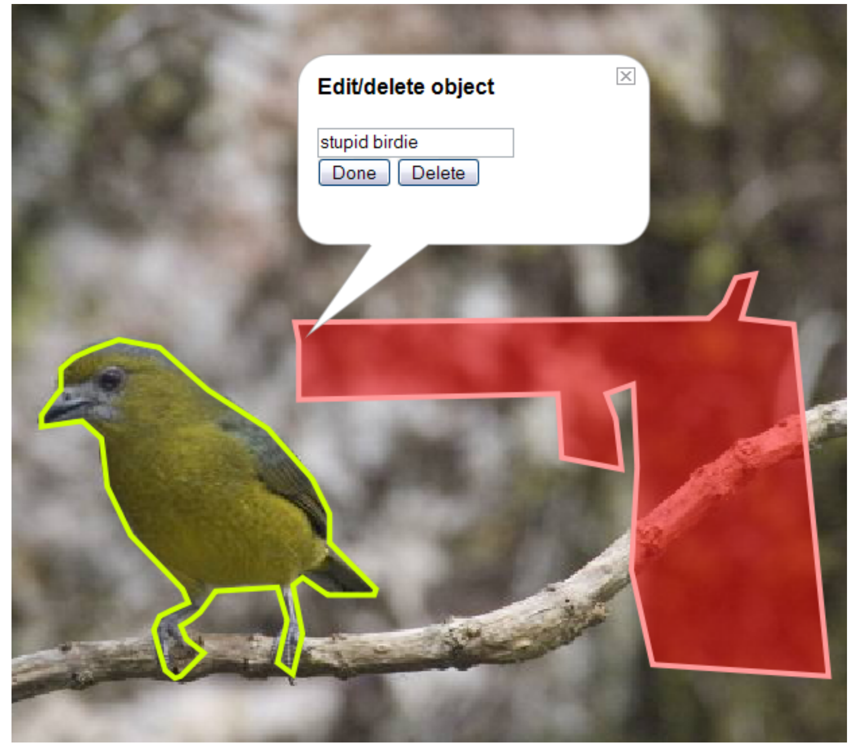
Nothing

…

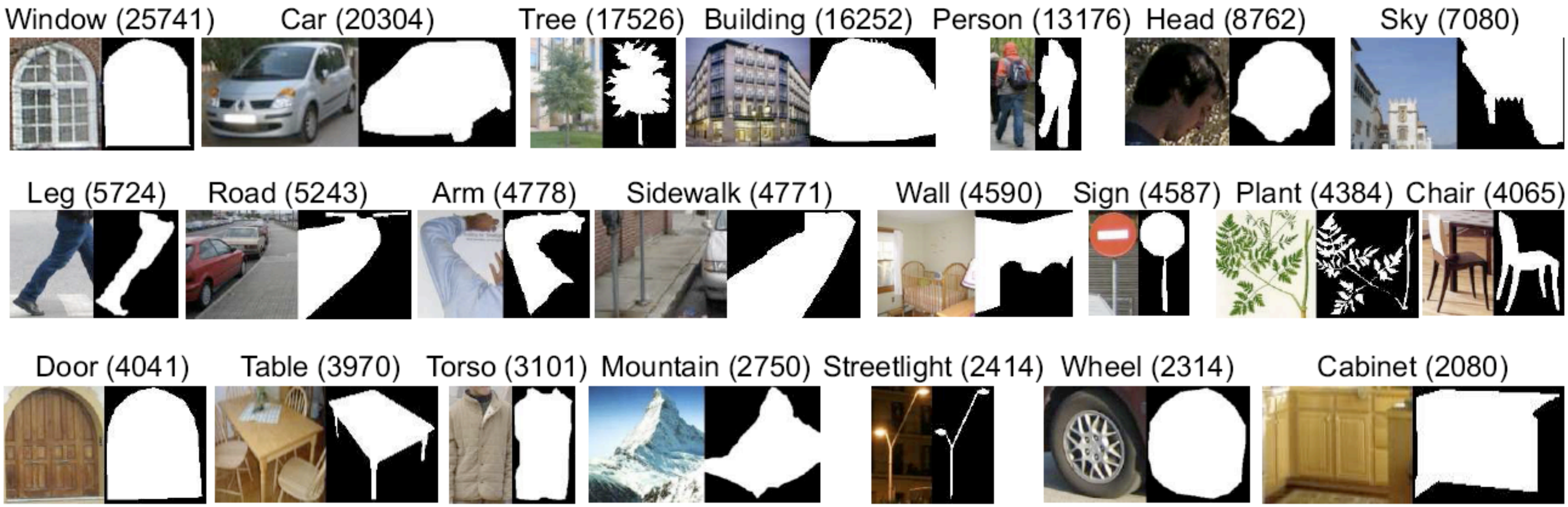
# Creative testing

Do not try this at home

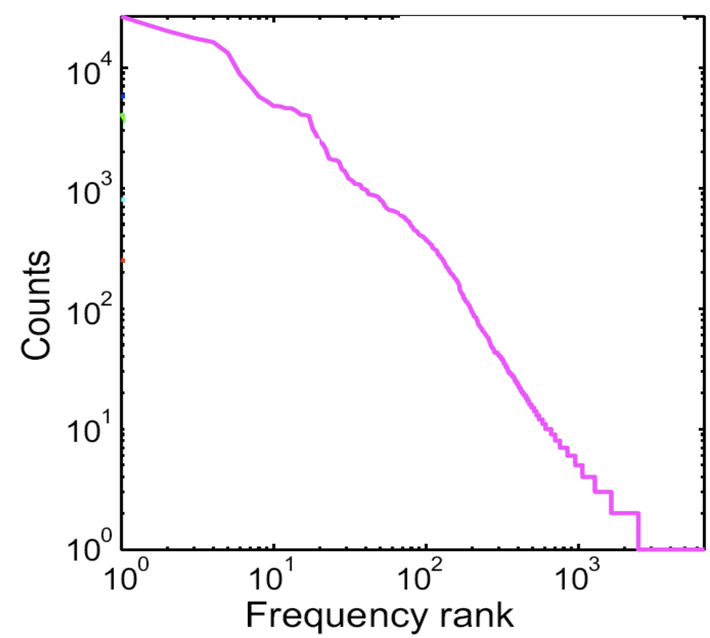


# Object statistics




**Stats:**
- 430,000 polygons
- 8500 different object descriptions
- 265 descriptions with more than 100 instances

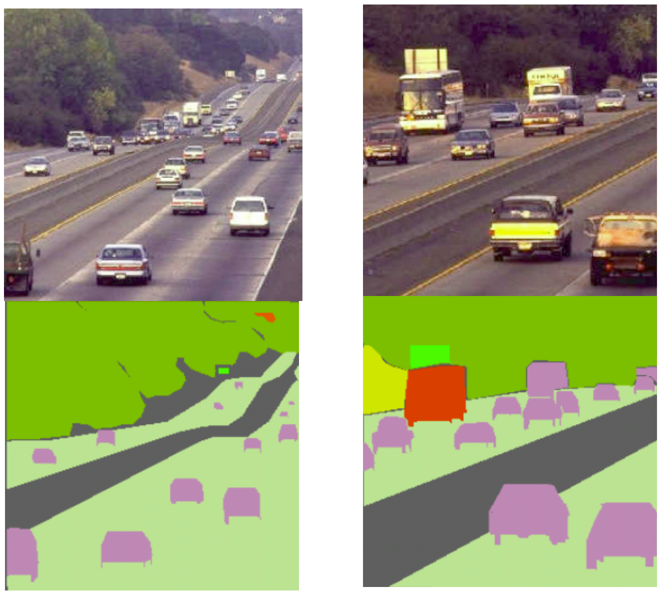


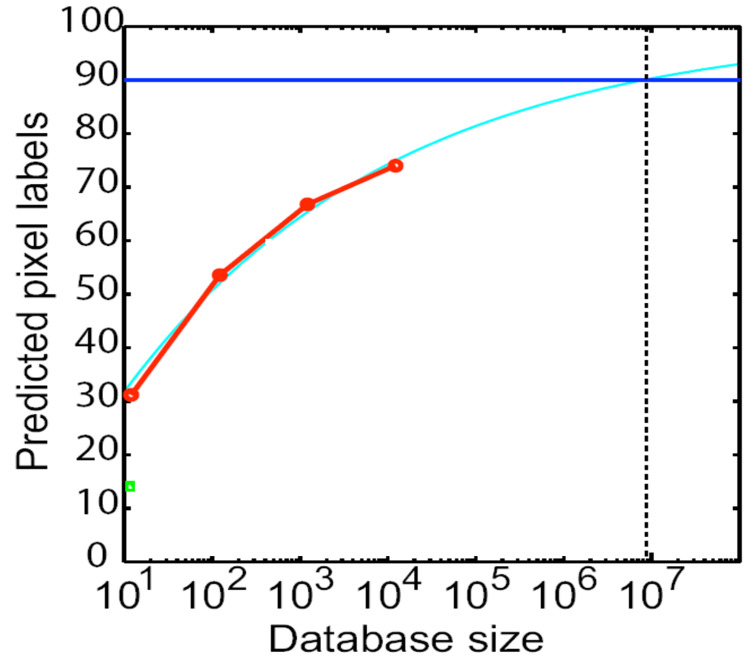
# How many more images do we need label?

Mosaic showing 12,000 fully annotated images

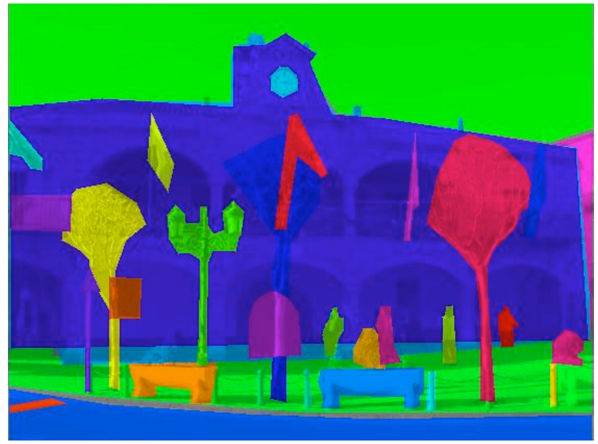
Interactive version at: http://people.csail.mit.edu/torralba/research/LabelMe/labelmeMap/

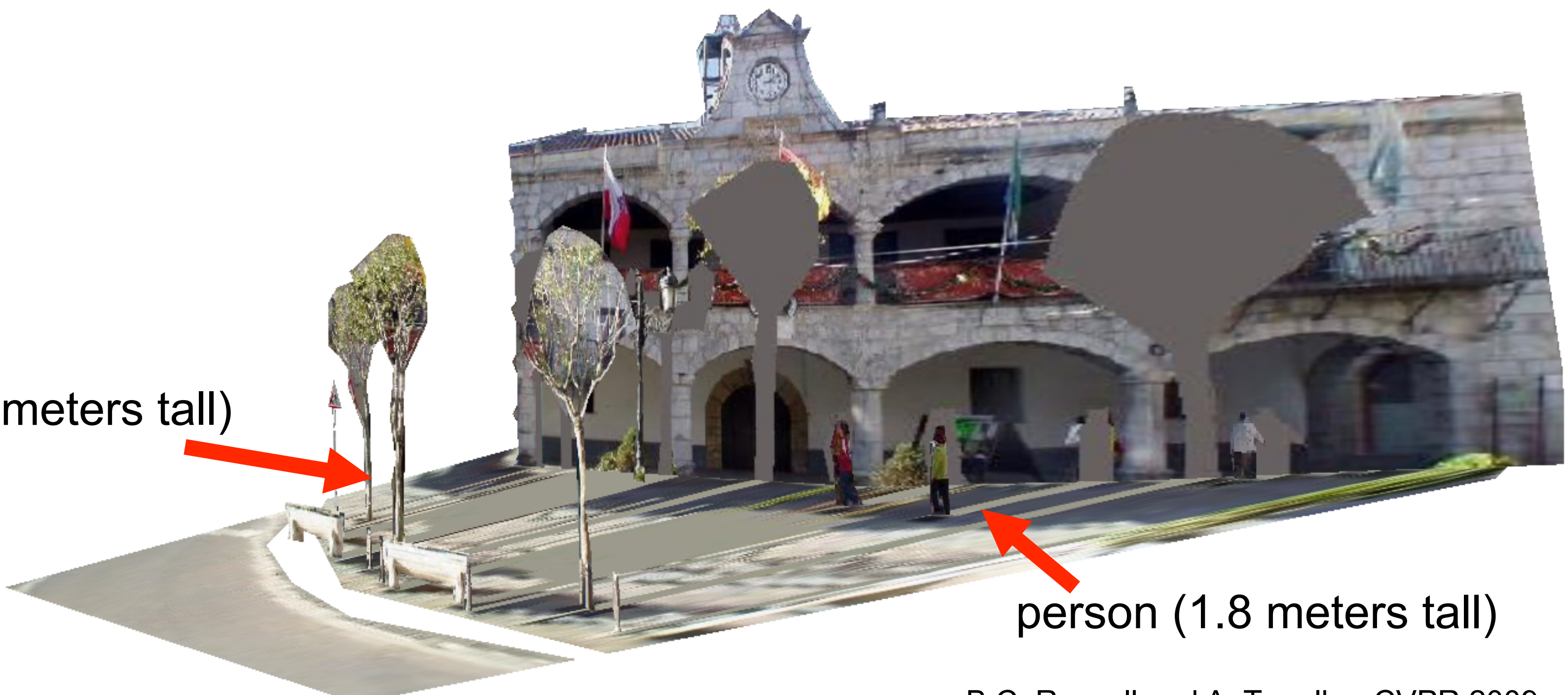
# How many images do we need to label?





# Beyond object annotation
# Building a database of 3D scenes

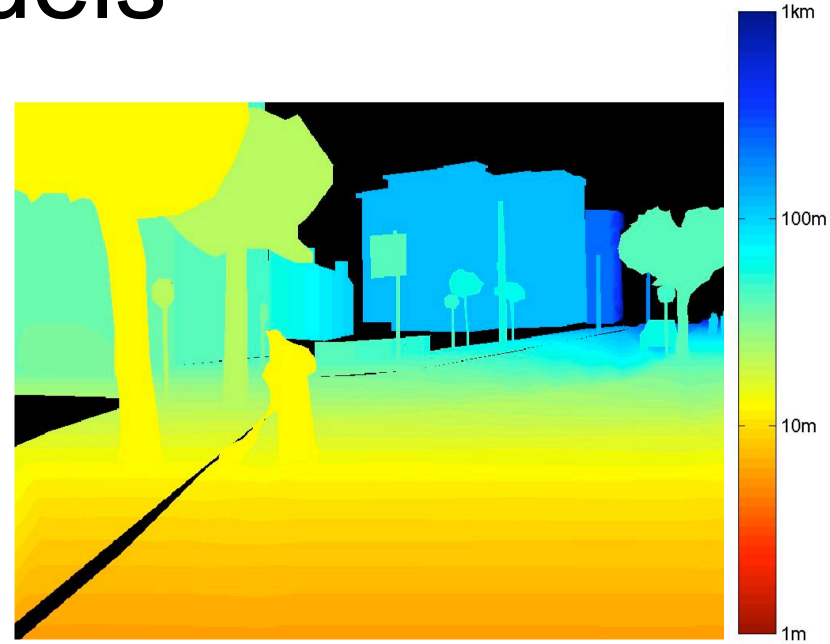





B.C. Russell and A. Torralba. CVPR 2009.

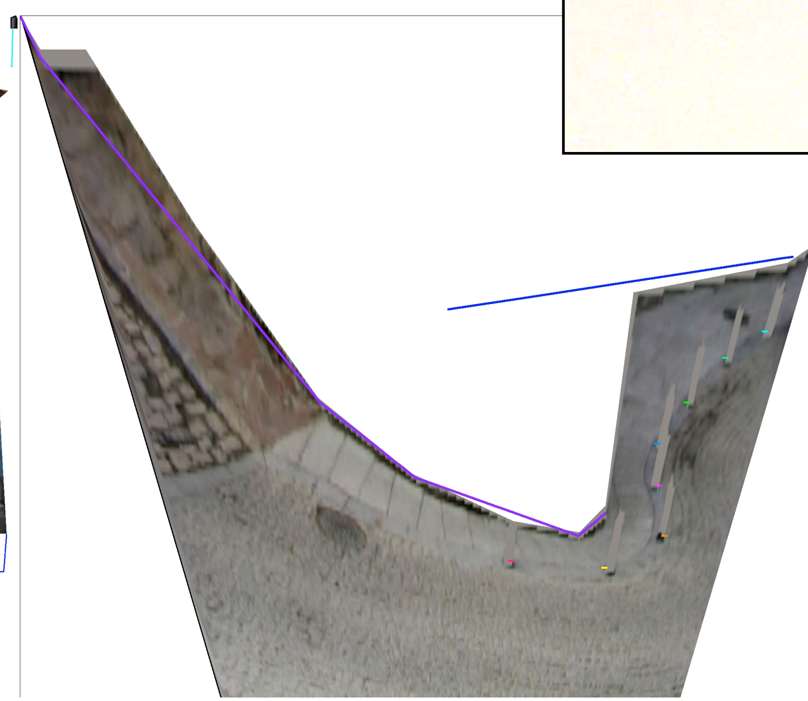
# 3D models

Depth map





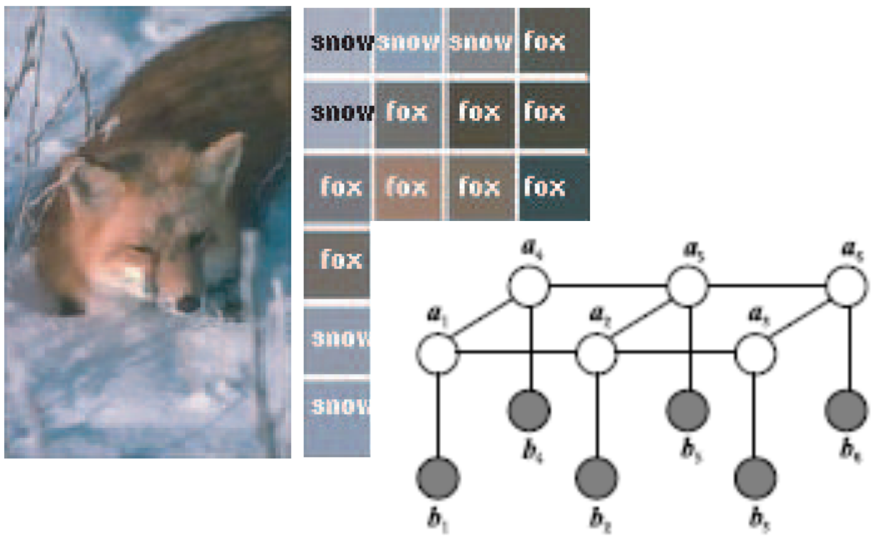
B. Russell, A. Torralba, J. Schwartz J. Ponce. 2008



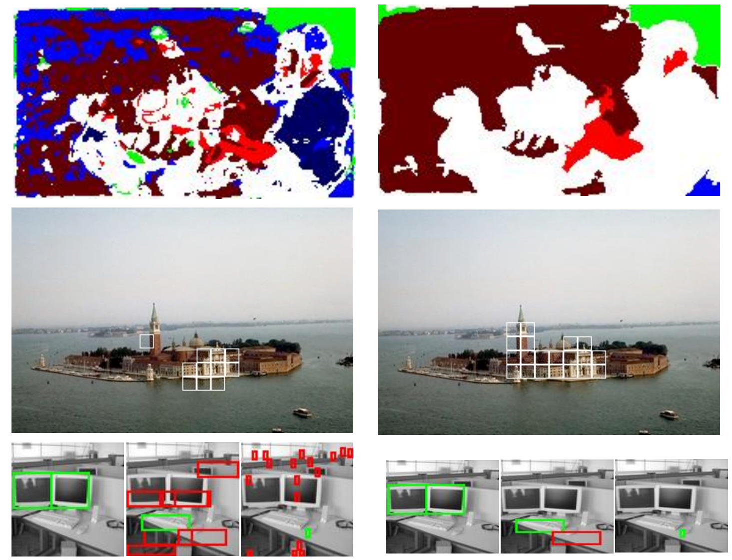
# Objects in context

# Contextual object relationships

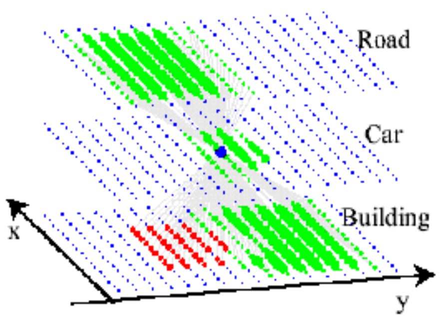
Carbonetto, de Freitas & Barnard (2004)



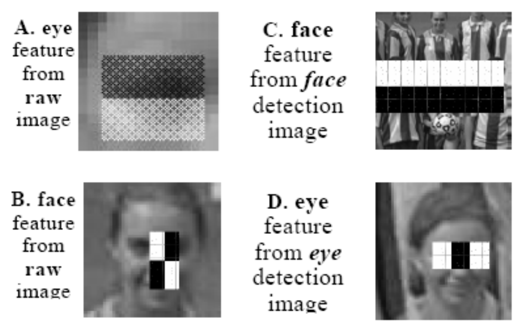

Kumar, Hebert (2005)



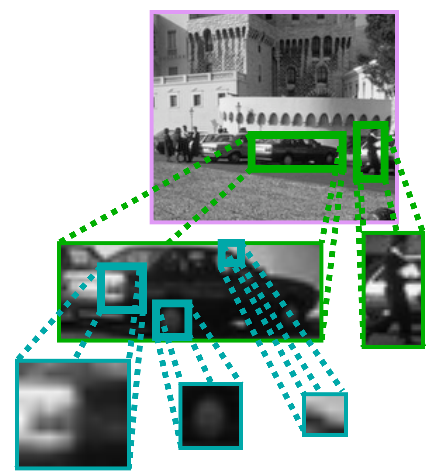
Torralba Murphy Freeman (2004)




Fink & Perona (2003)




E. Sudderth et al (2005)



# The context challenge

How far can you go without using an object detector?

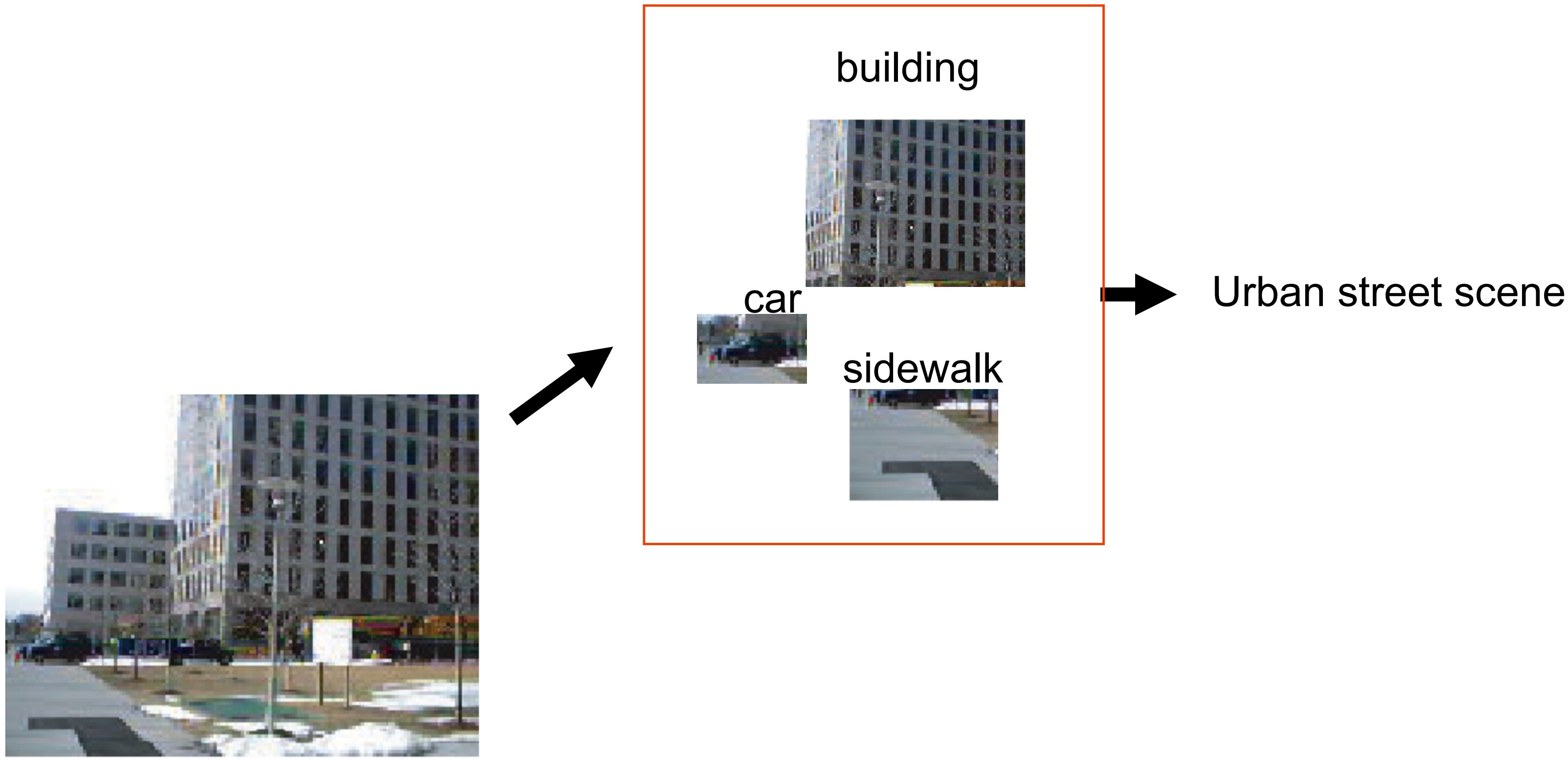
# What are the hidden objects?


# What are the hidden objects?

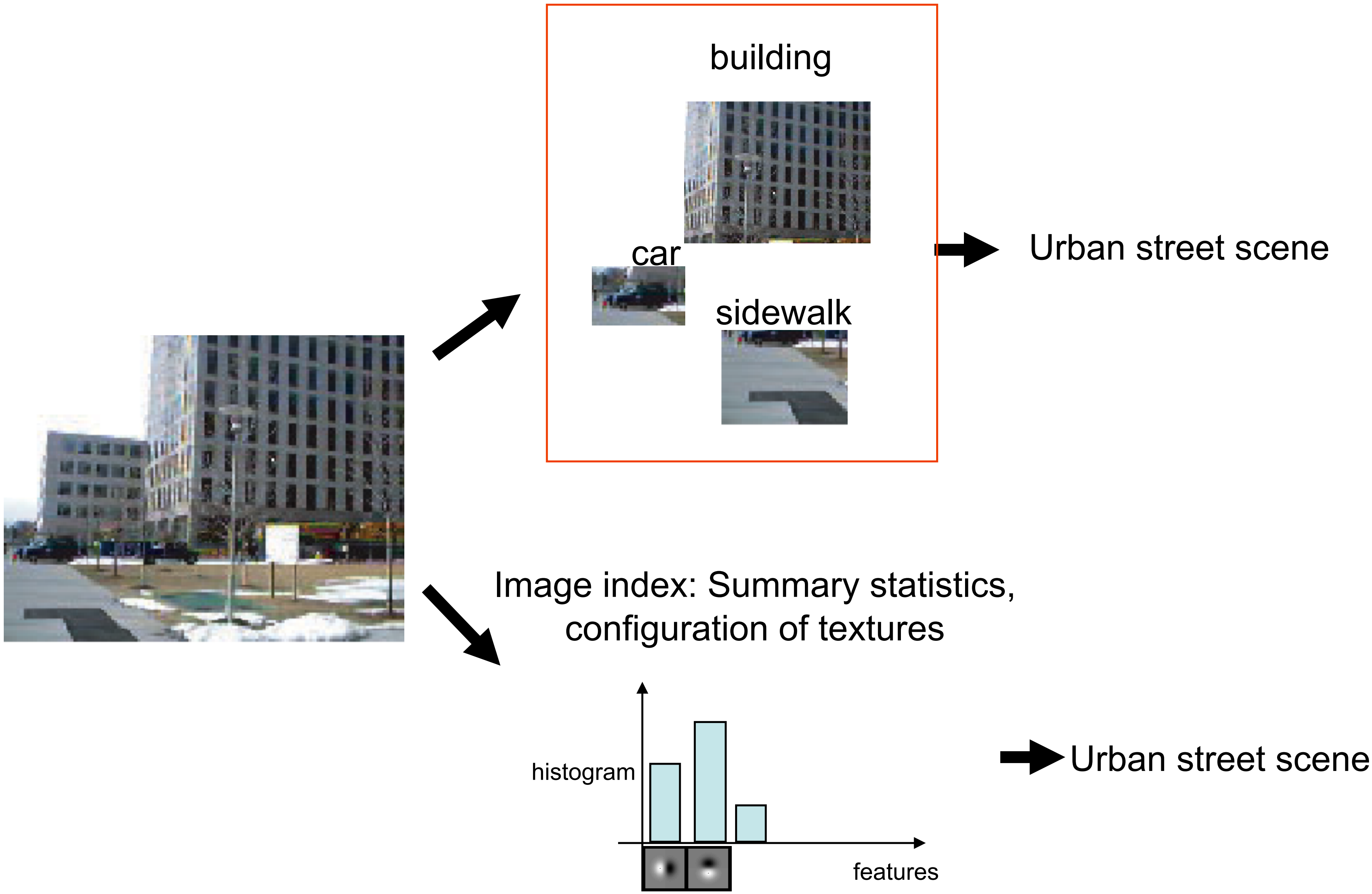
# What are the hidden objects?

Chance ~ 1/30000

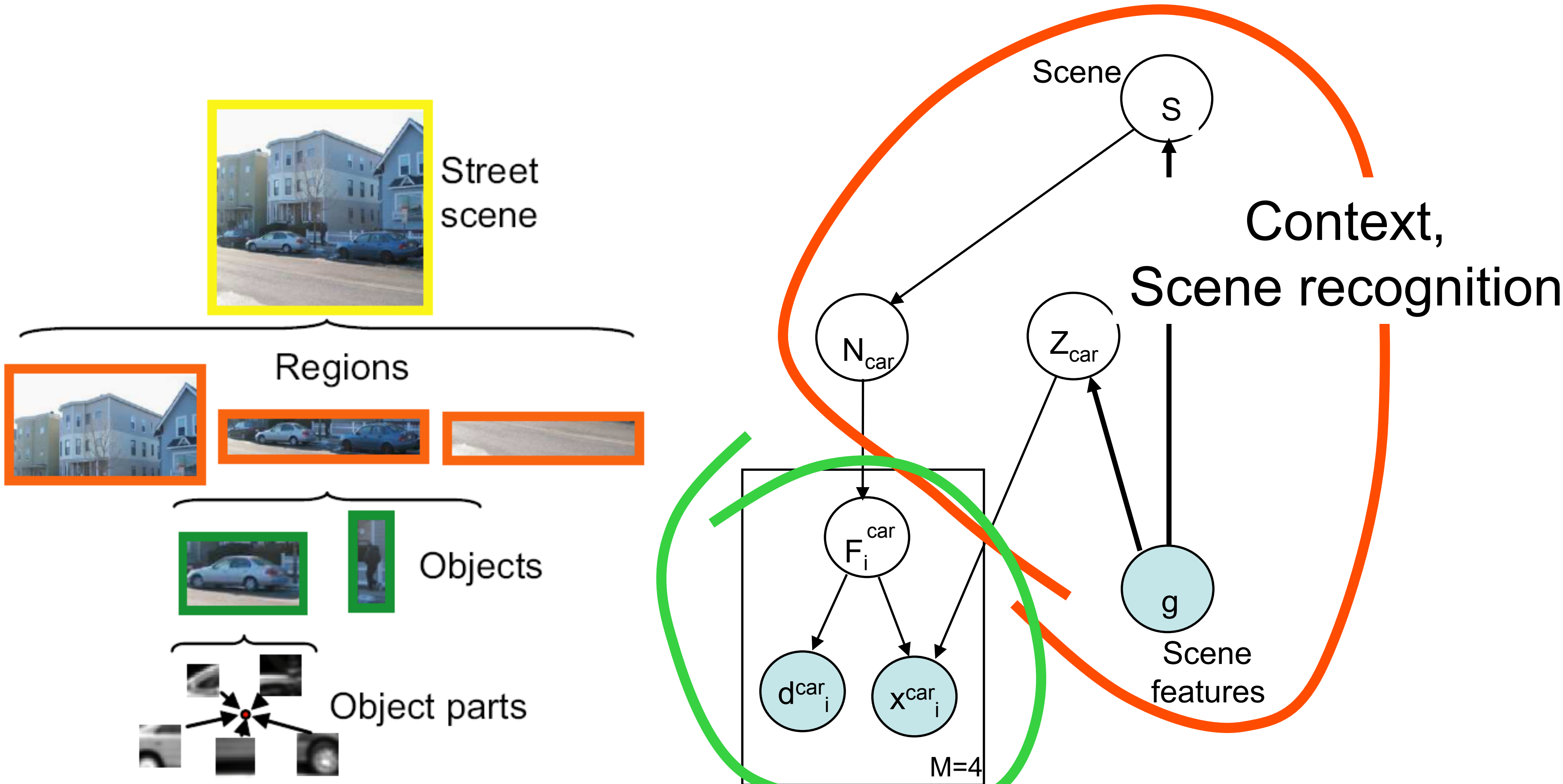
# Global and local representations



# Global and local representations

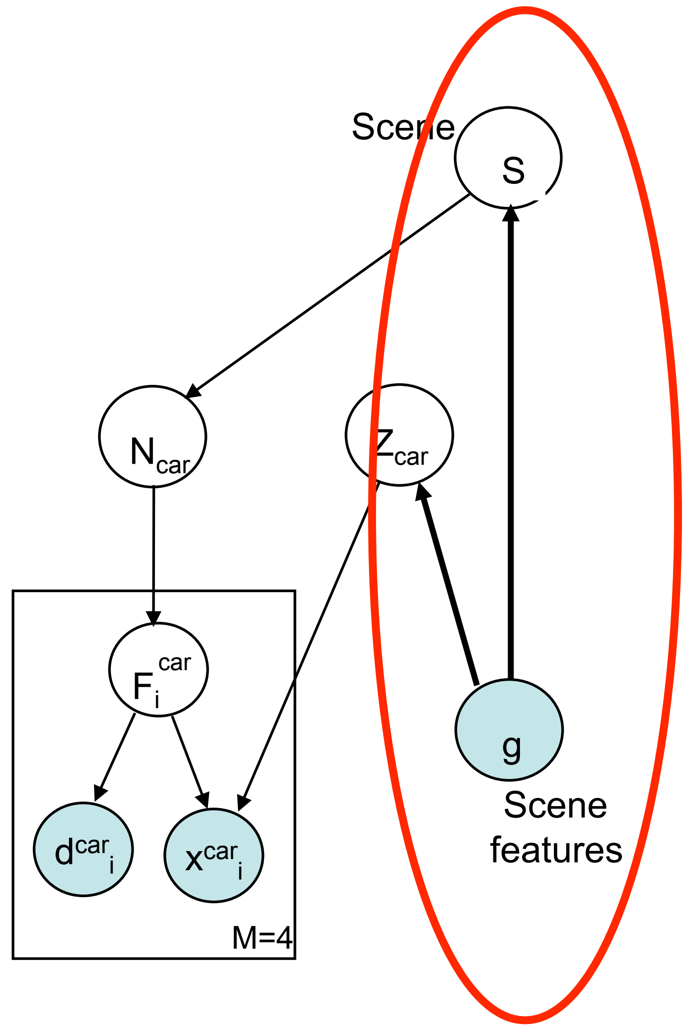


# An integrated model of Scenes, Objects, and Parts



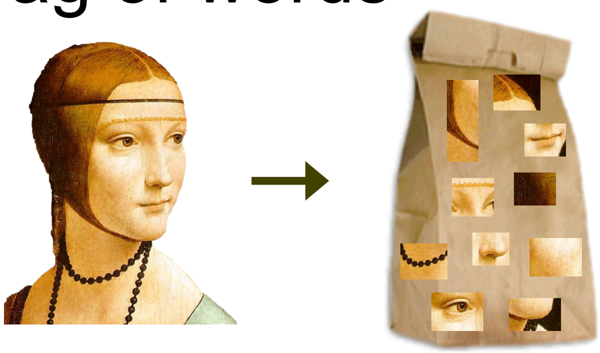

Multiclass and pose invariant object detection,

# Scene recognition



# Global scene representations

## Bag of words



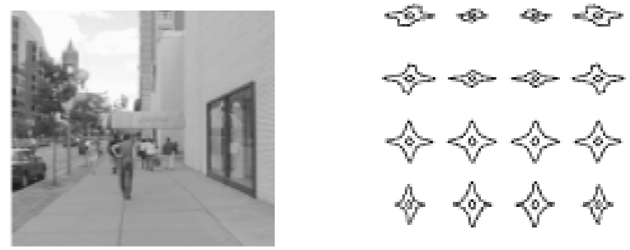
Sivic et. al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

## Non localized textons

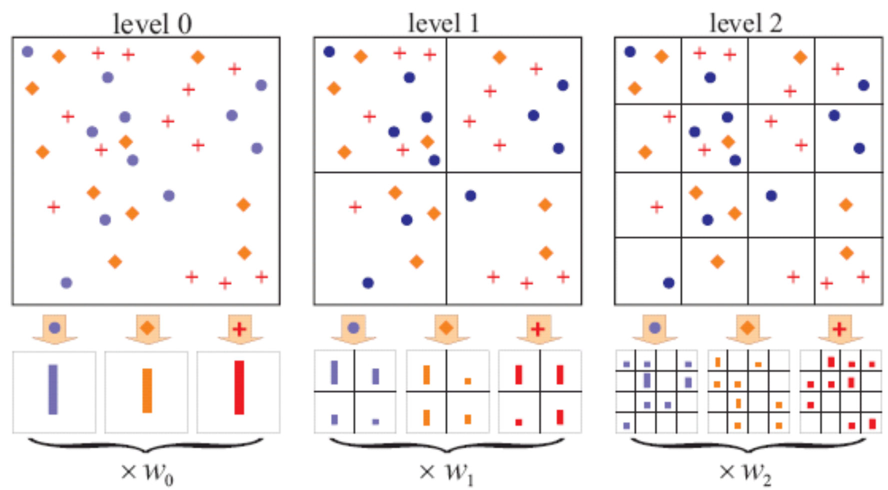

Walker, Malik. Vision Research 2004

...

## Spatially organized textures



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



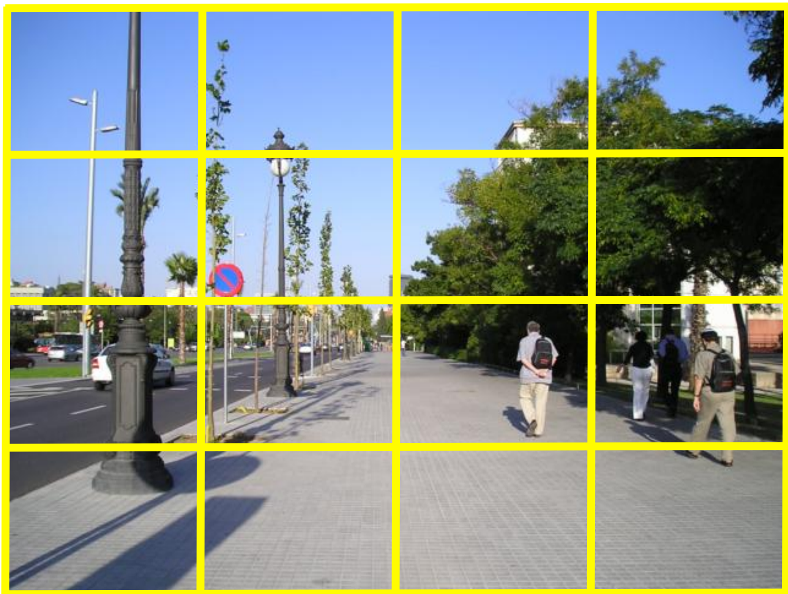

S. Lazebnik, et al, CVPR 2006

...

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

# Features for matching images: Gist

Oliva and Torralba, 2001

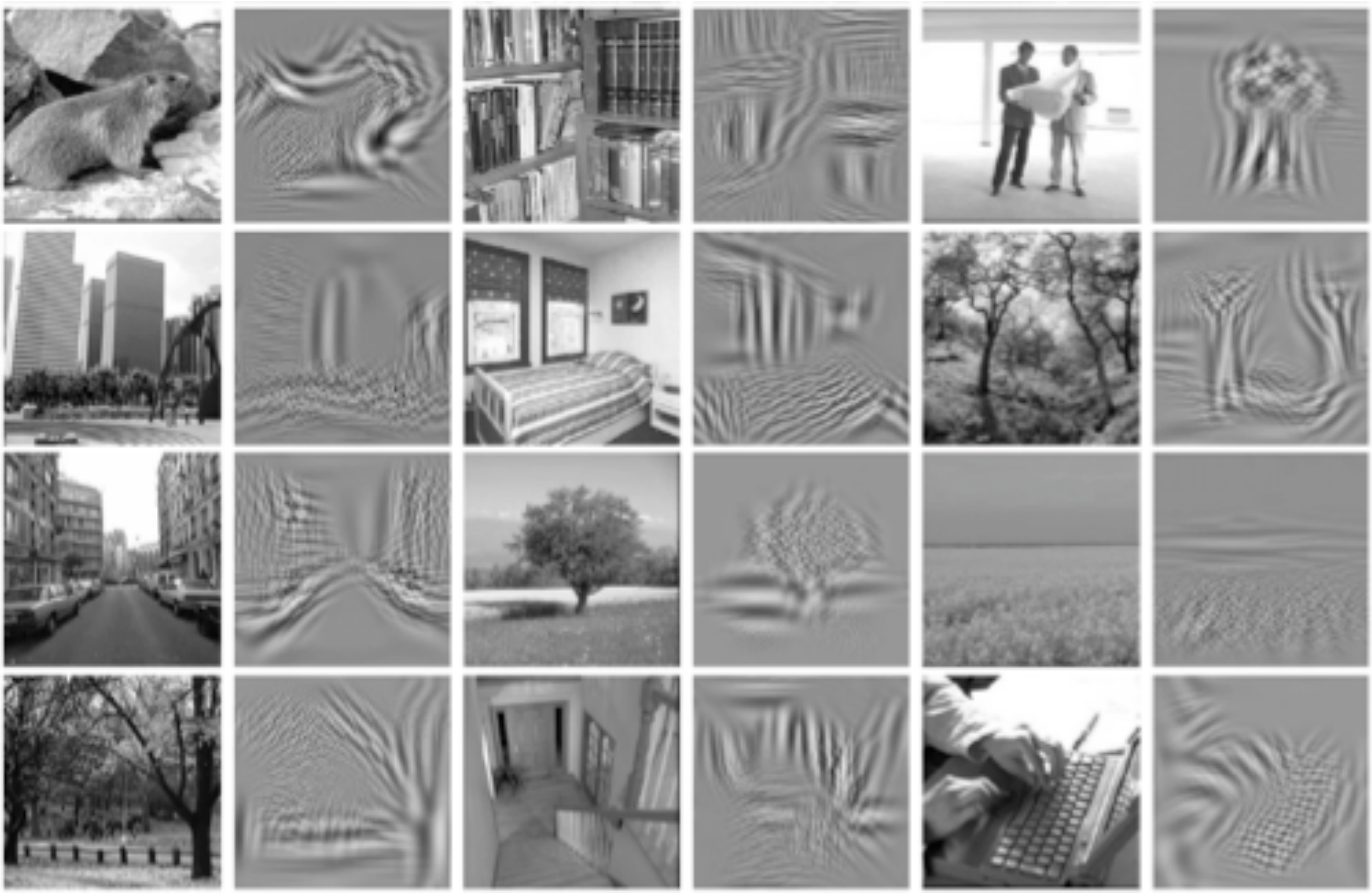


- Apply oriented Gabor filters over different scales
- Average filter energy in each bin

|  |  |
|---|---|
| 8 | orientations |
| 4 | scales |
| x 16 | bins |
| 512 | dimensions |

- Used for scene recognition
- Similar to SIFT (Lowe 1999)

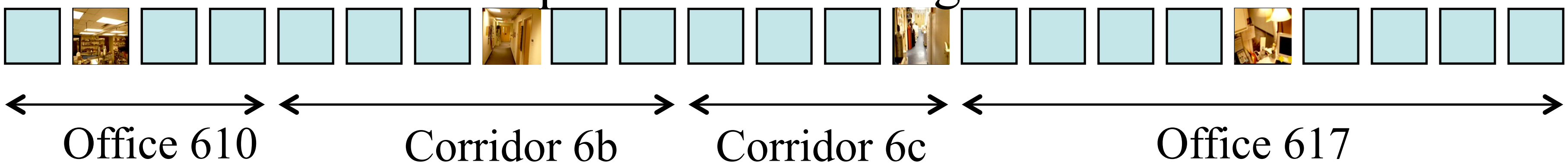
# Example visual gists



Global features (I) ~ global features (I') Oliva & Torralba (2001)

# Context-based vision system for place and object recognition

We use 17 annotated sequences for training



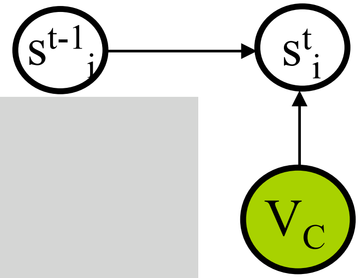

- Hidden states = location (63 values)
- Observations = $v^G_t$ (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)

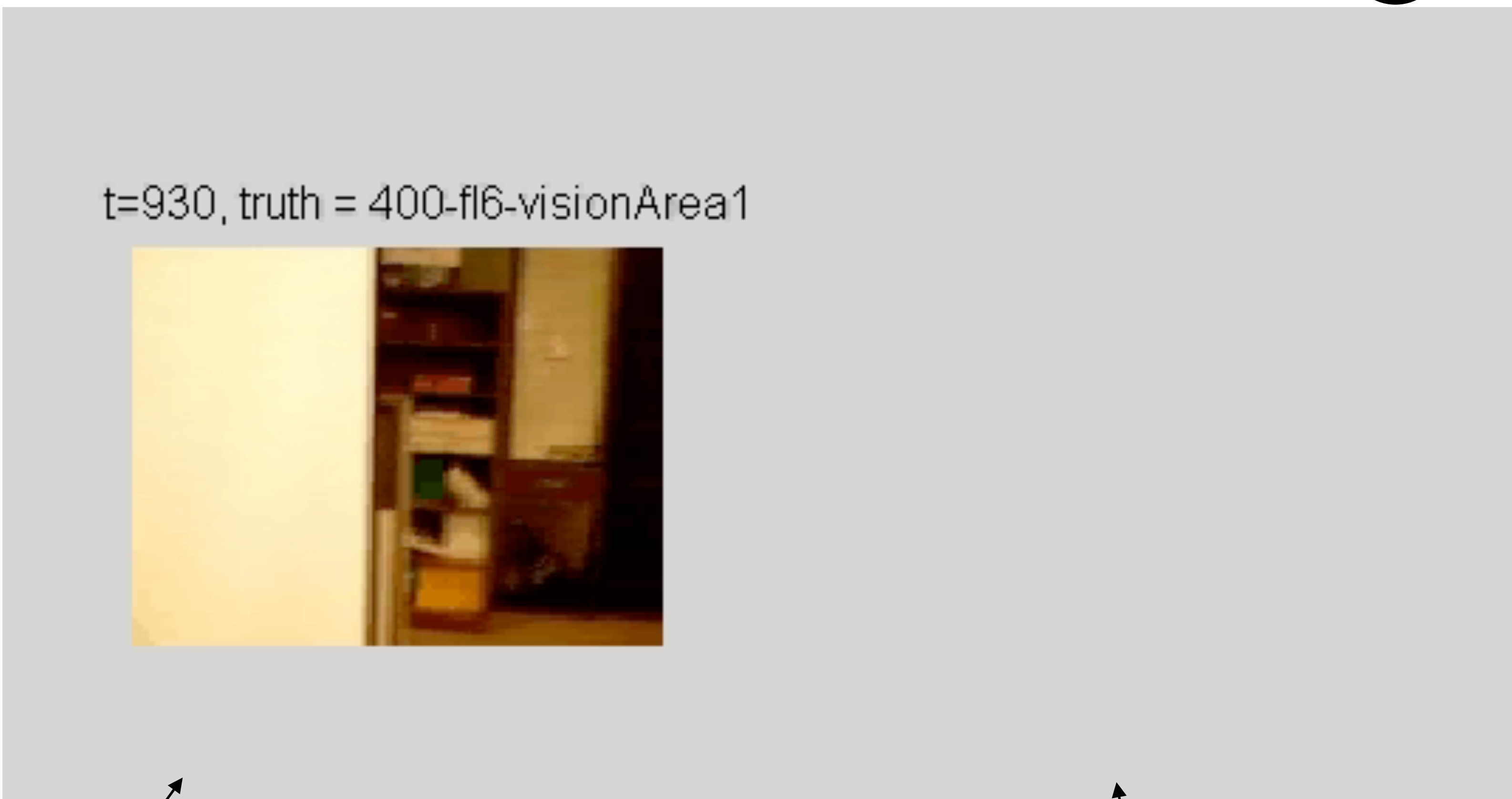
Torralba, Murphy, Freeman and Rubin. ICCV 2003

# Our mobile rig


Torralba, Murphy, Freeman, Rubin. 2003

# Place recognition demo

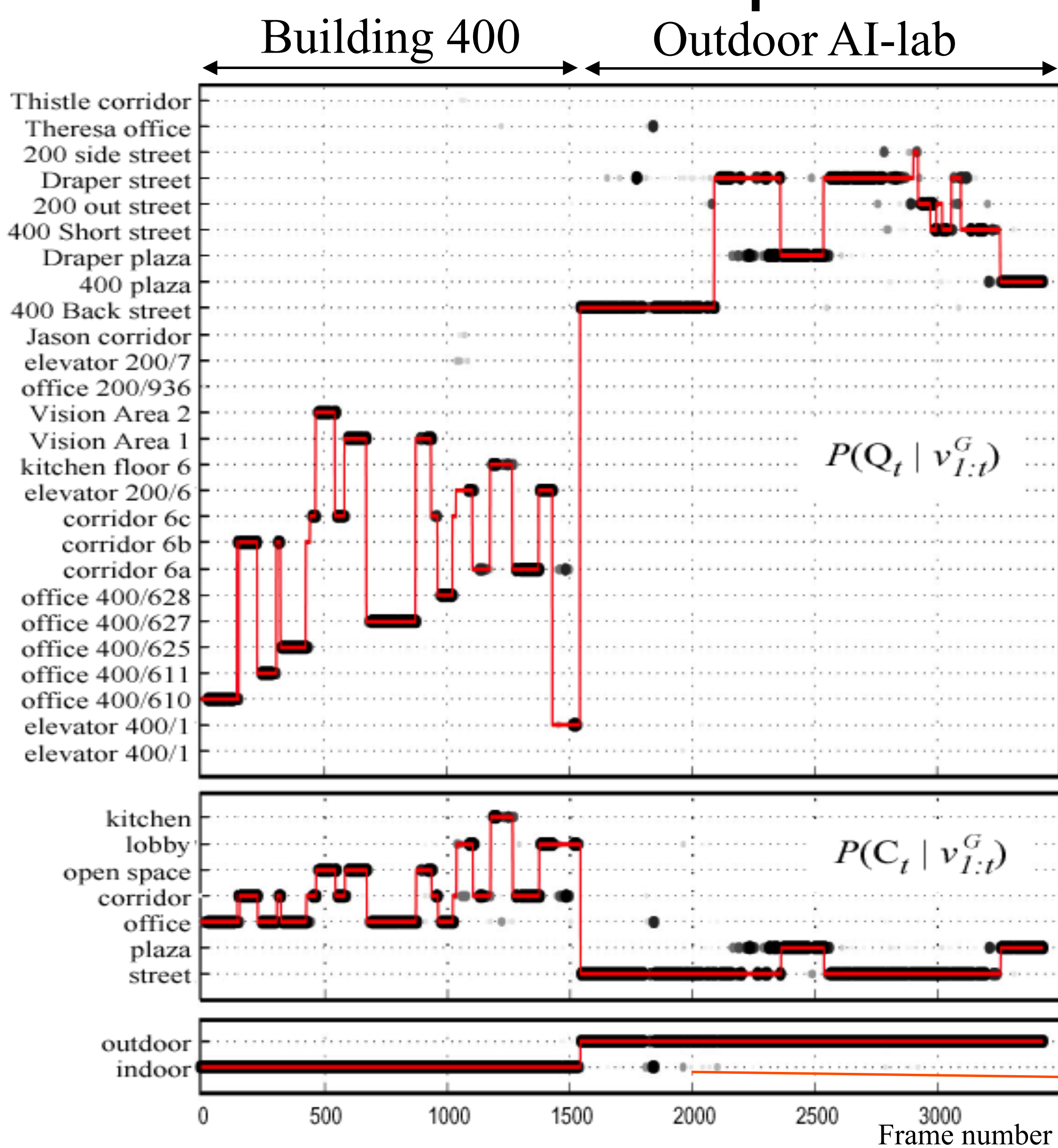



**Input image (120x160)**

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

# Identification and categorization of known places



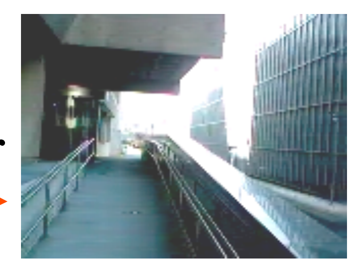

Ground truth

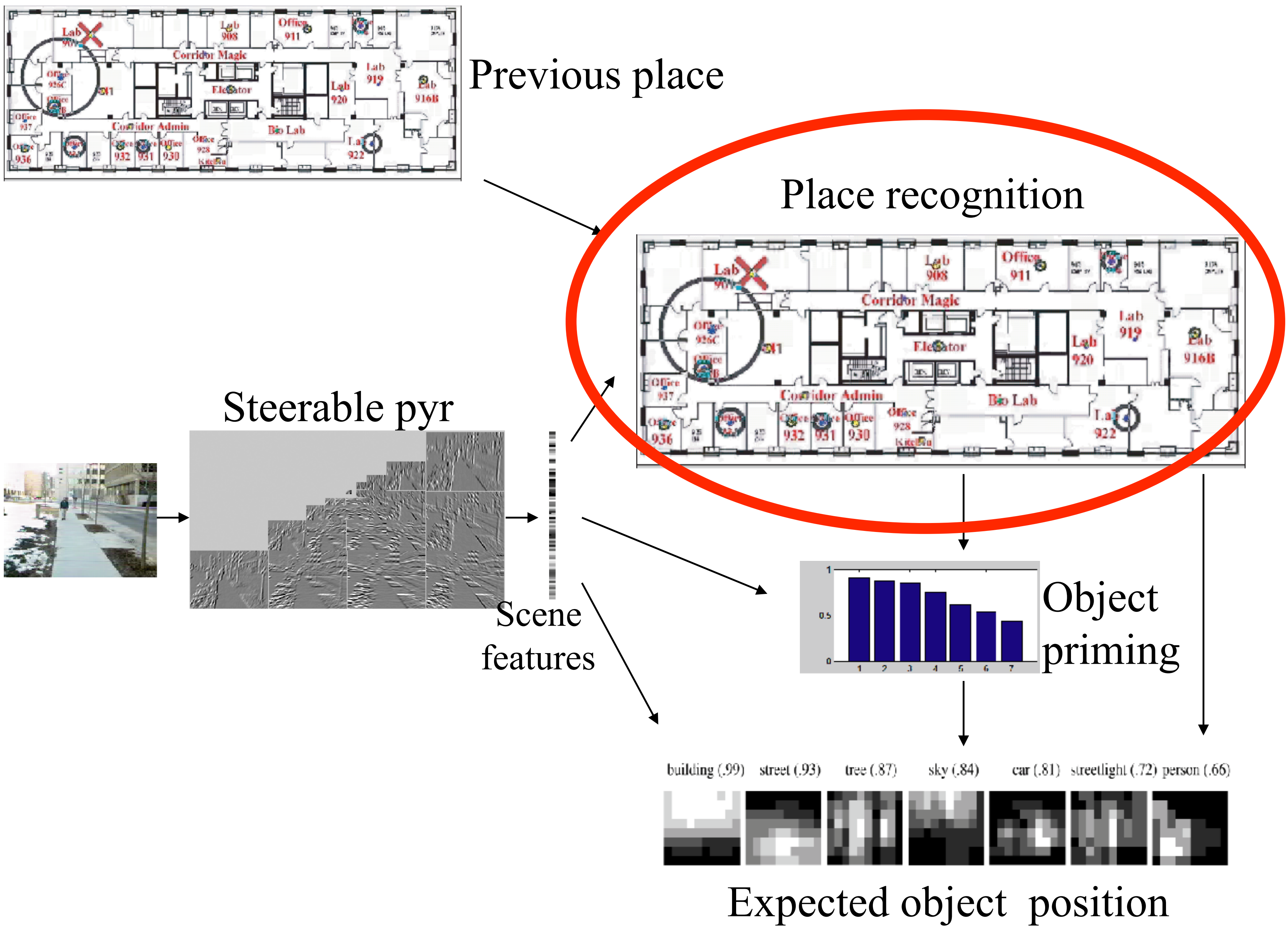
System estimate

Specific location

Location category

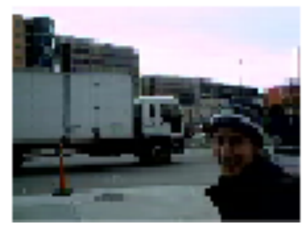
Indoor/outdoor




Previous place
Place recognition
Steerable pyr
Scene
features
Object
priming
building (.99)
street (.93)
tree (.87)
sky (.84)
car (.81)
streetlight (.72)
person (.66)
Expected object position

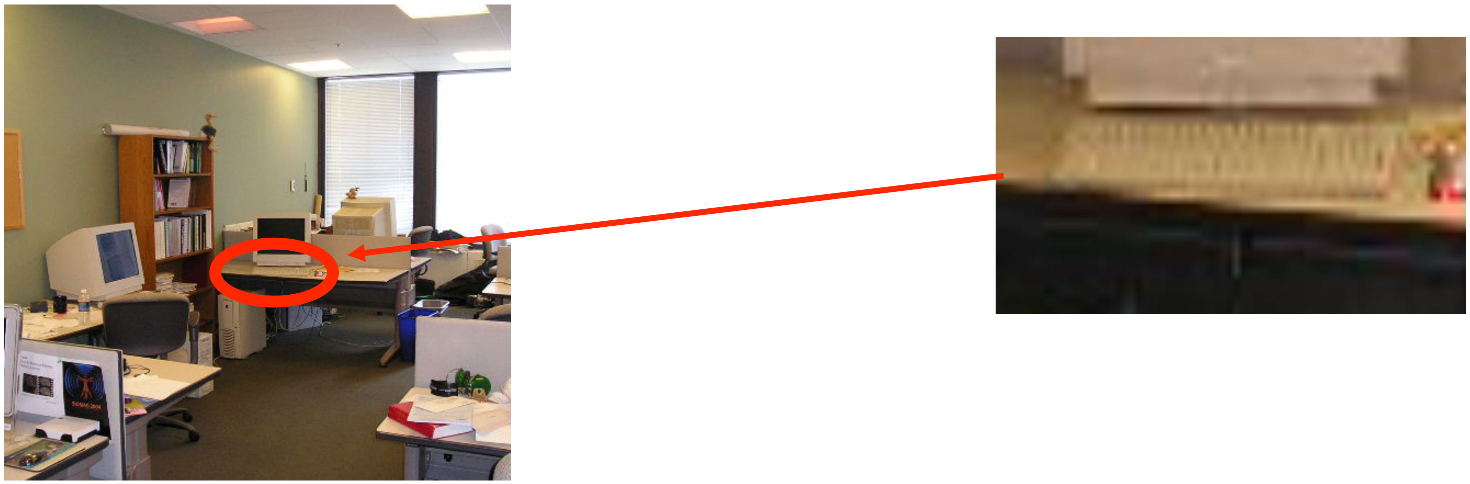
# Application of object detection for image retrieval

**Results using the keyboard detector alone**

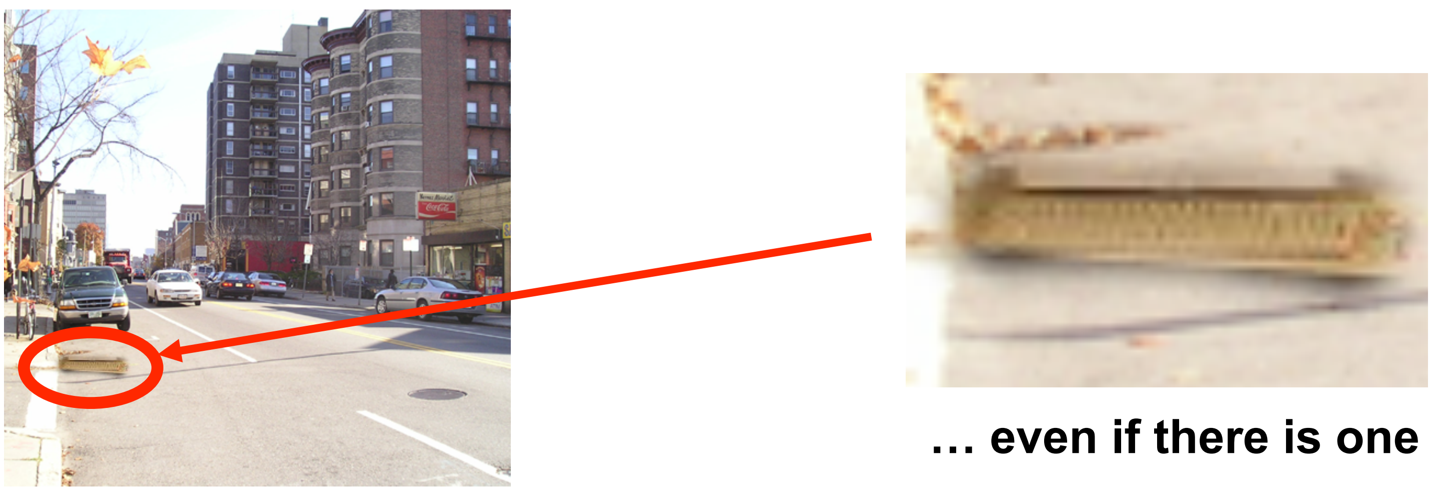
     

# The system does not care about the scene, but we do…

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

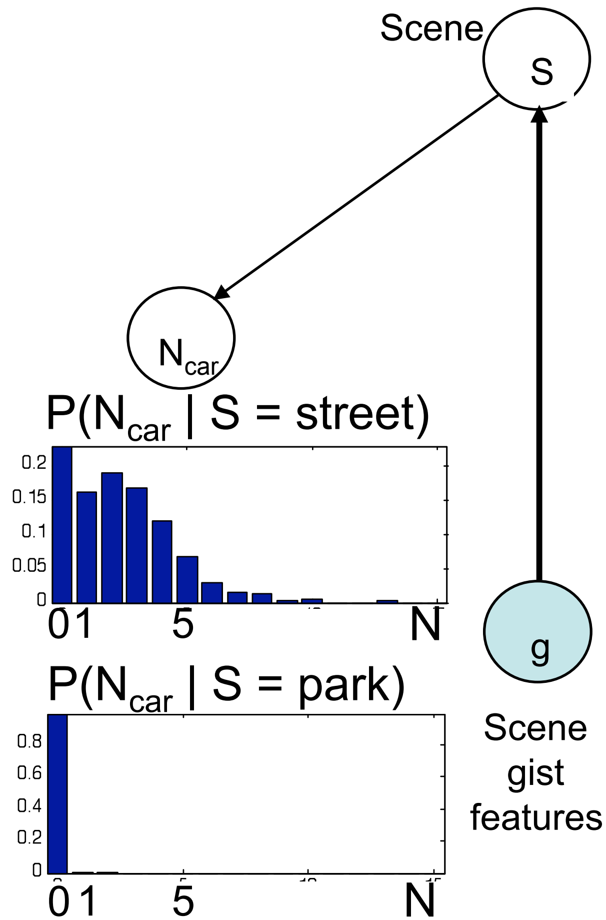


We know there is no keyboard present in this scene



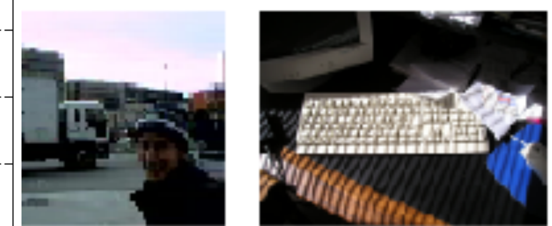
**… even if there is one indeed.**

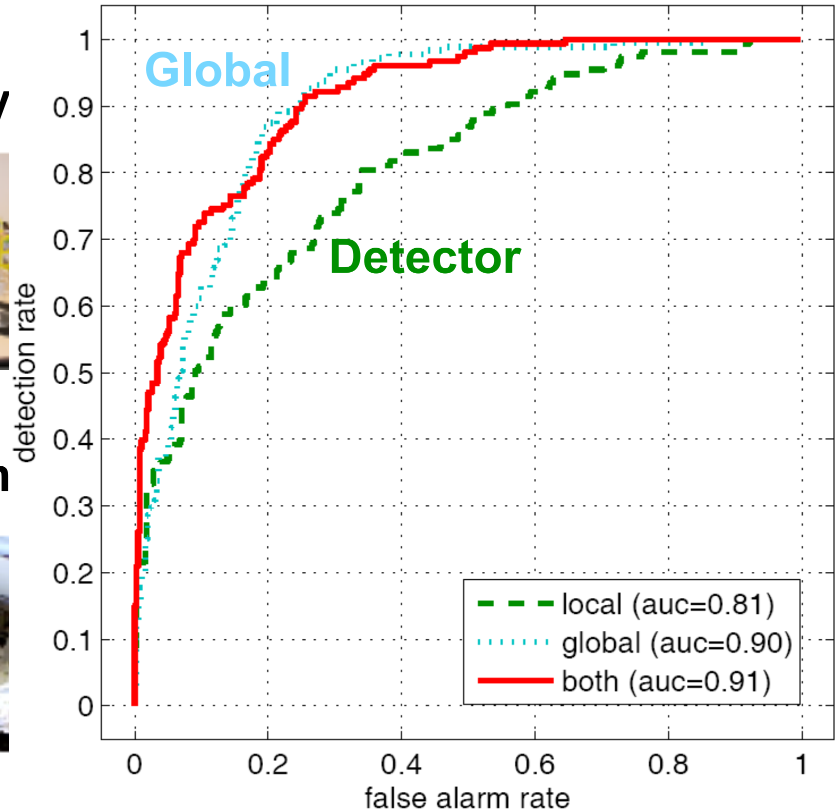
# An integrated model of Scenes, Objects, and Parts



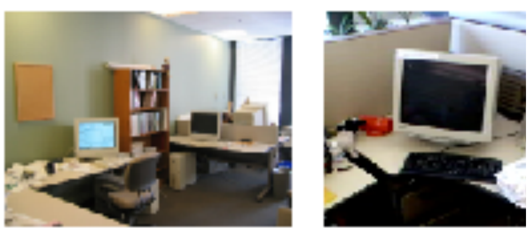
# Application of object detection for image retrieval

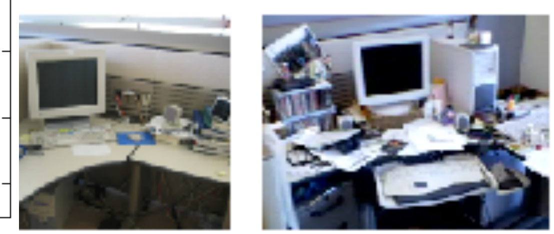
Results using the key



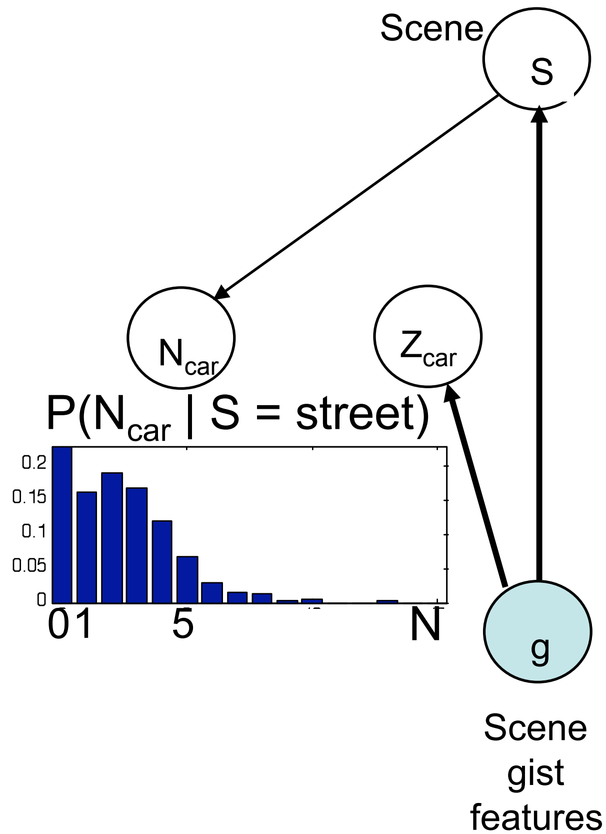



Results using both th cene features

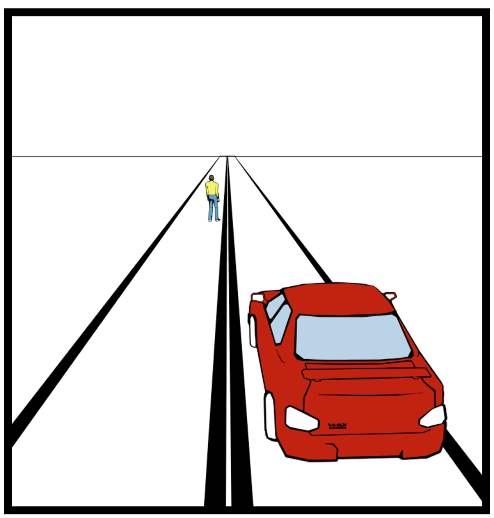




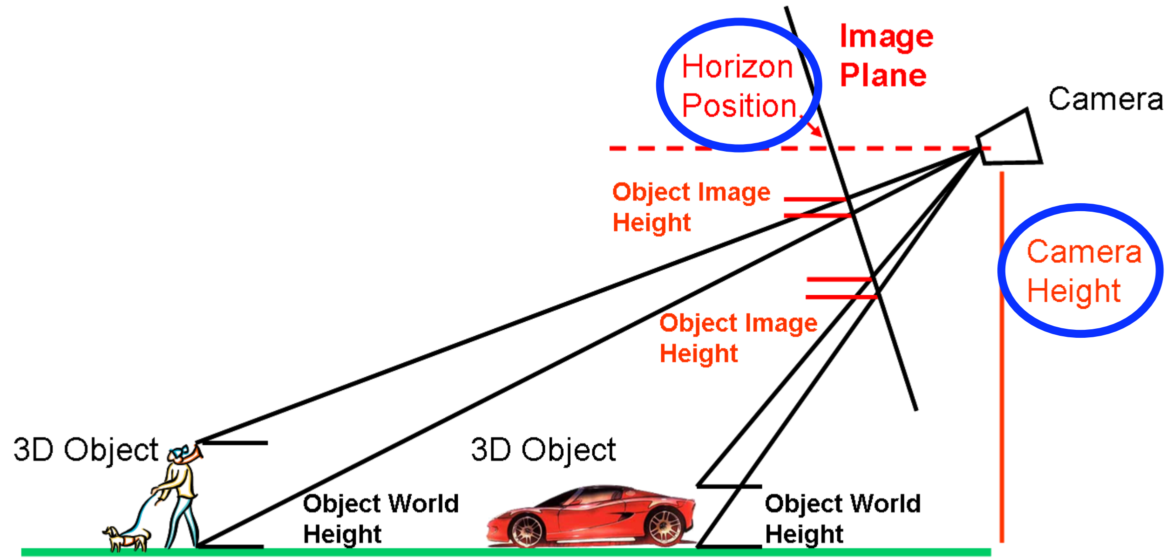
# Context driven object detection



# 3d Scene Context



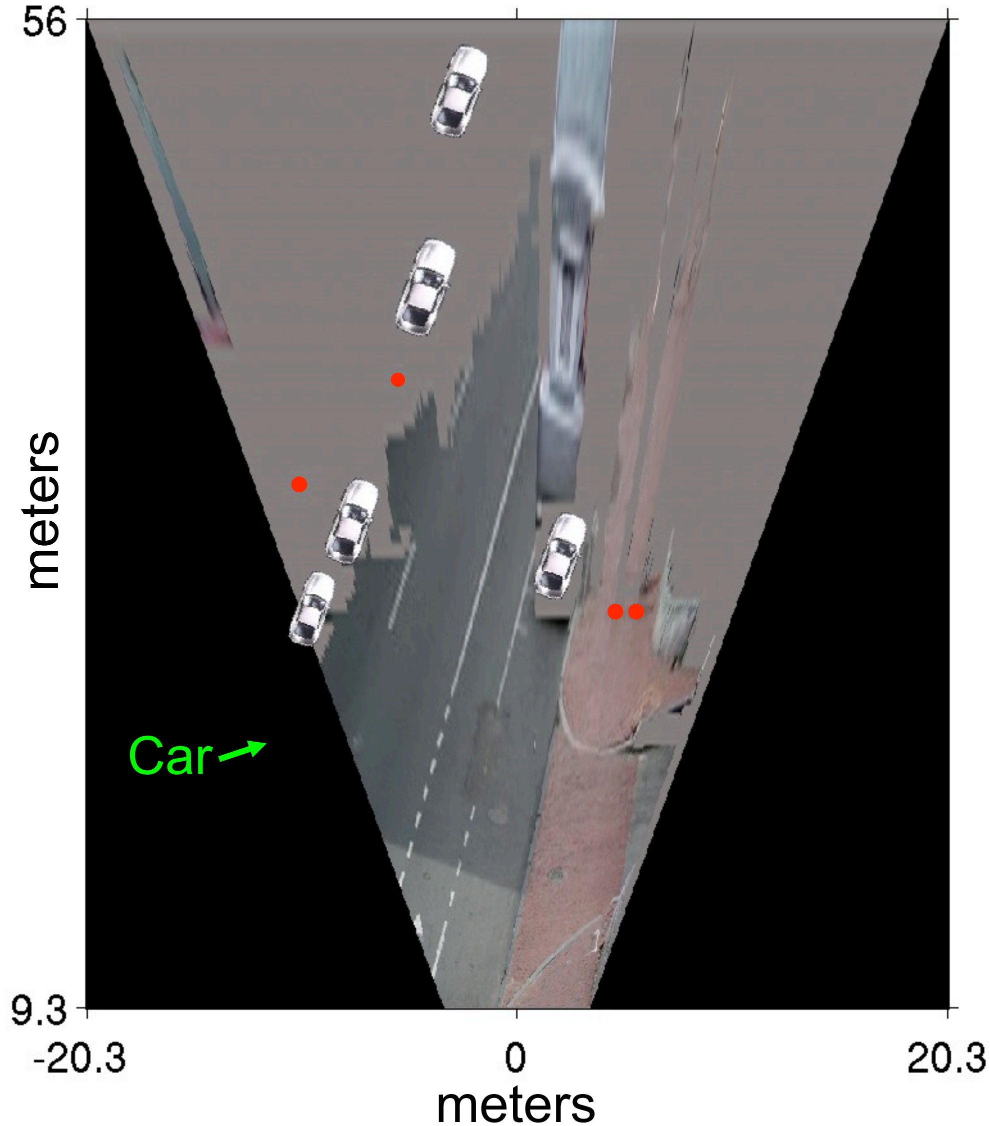
Image




World

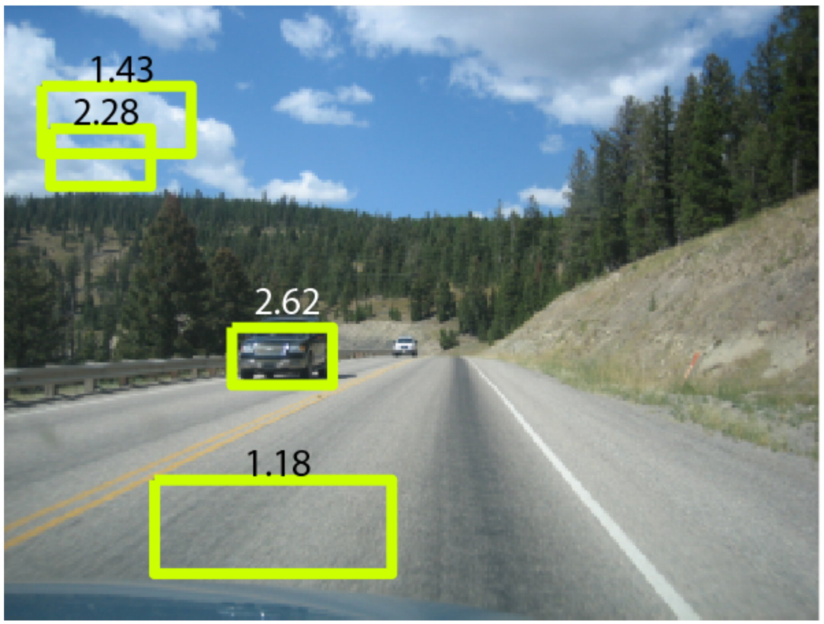
Hoiem, Efros, Hebert ICCV 2005

# 3d Scene Context

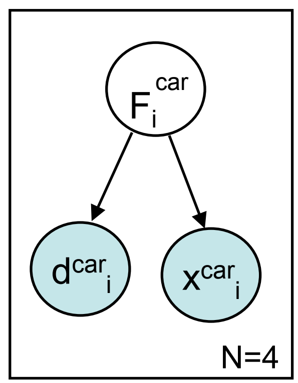



Hoiem, Efros, Hebert ICCV 2005

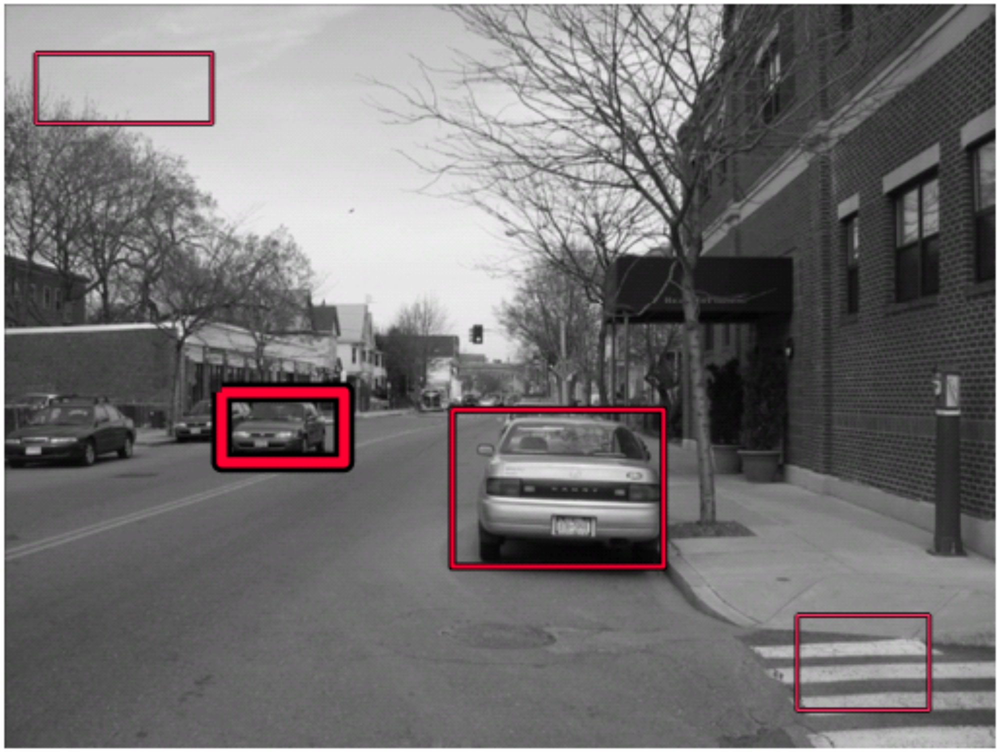
# An integrated model of Scenes, Objects, and Parts

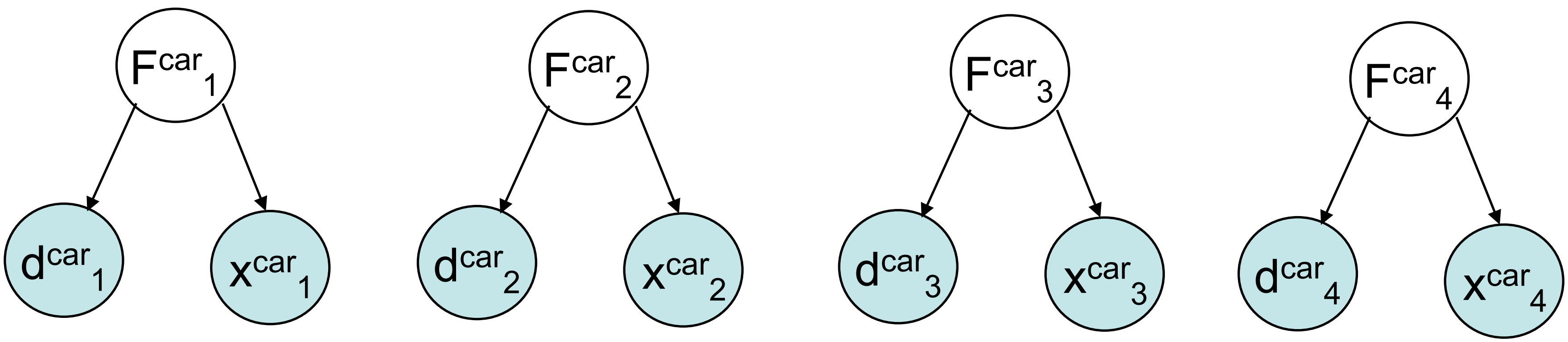



We train a multiview car detector.

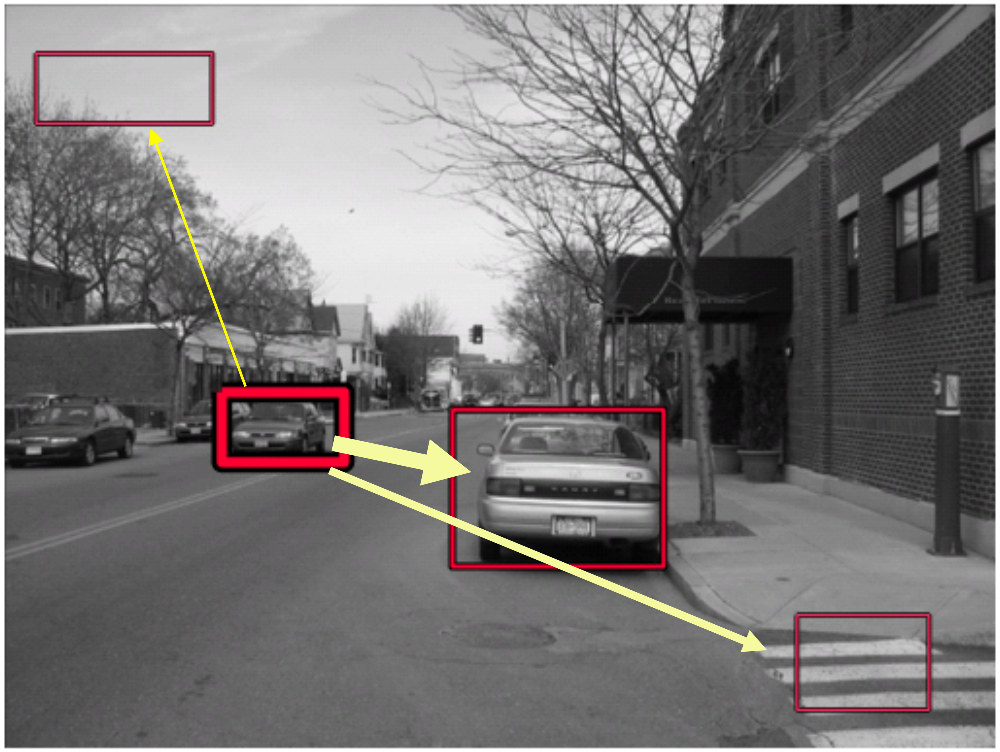

$$p(d \mid F=1) = N(d \mid \mu_1, \sigma_1)$$
$$p(d \mid F=0) = N(d \mid \mu_0, \sigma_0)$$





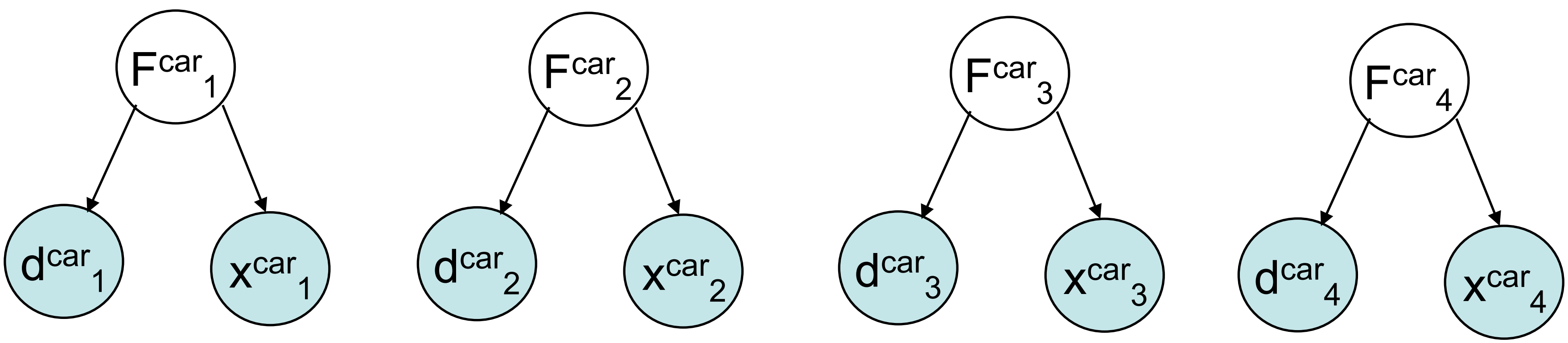

For each detected region we have to decide if the target is present

$$p(F_k^{car} = 1 | d_k^{car}, x_k^{car})$$

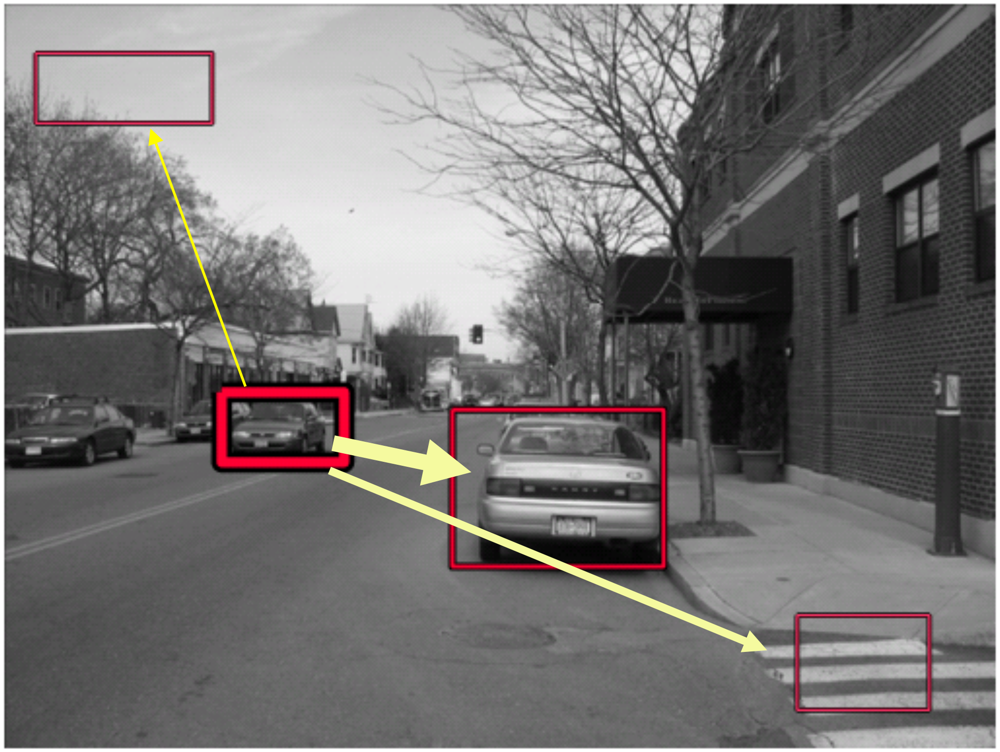


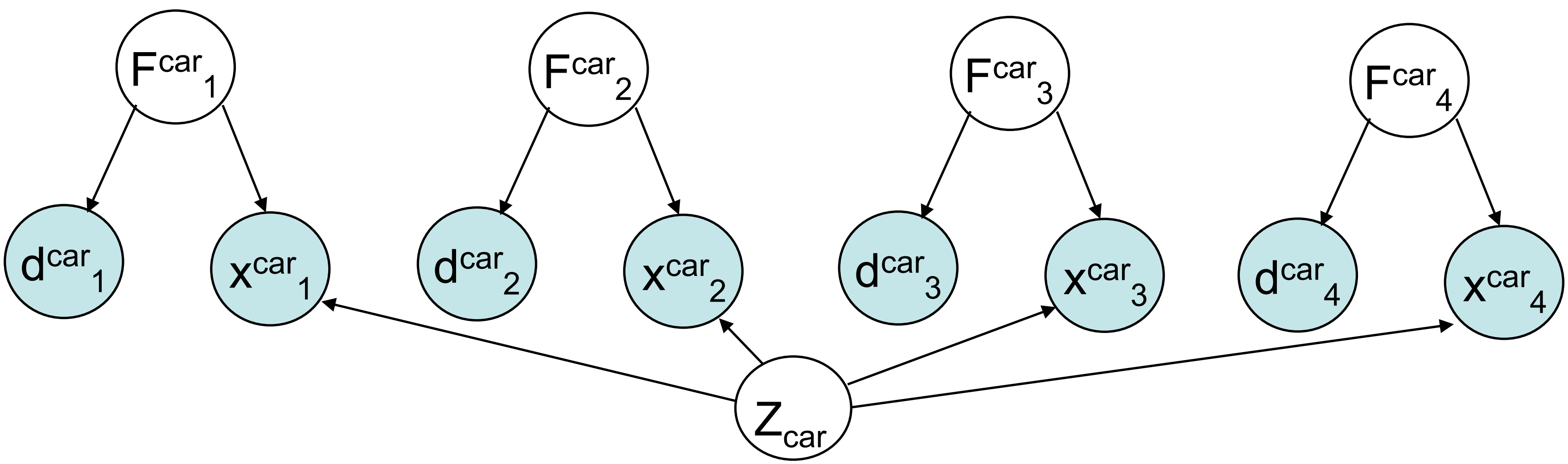
Object locations within the image are not independent.

All vehicles share the same ground plane.



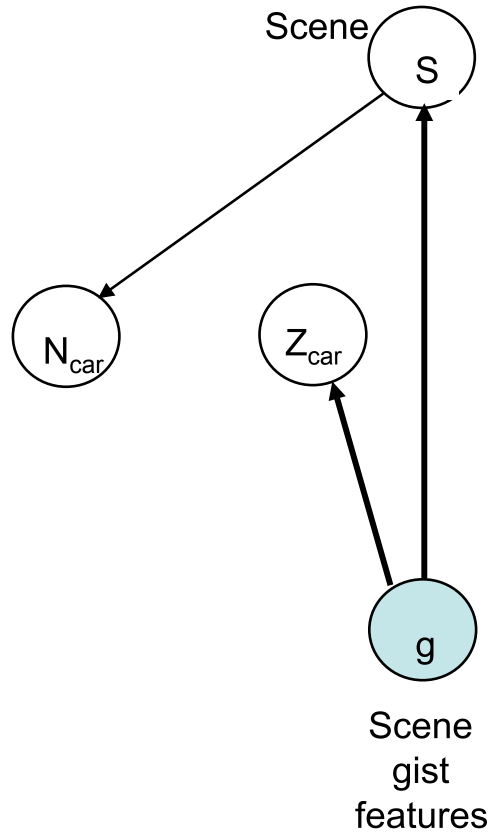

The graph is fully connected





$F^{car}{}_1$
$F^{car}{}_2$
$F^{car}{}_3$
$F^{car}{}_4$
$d^{car}{}_1$
$x^{car}{}_1$
$d^{car}{}_2$
$x^{car}{}_2$
$d^{car}{}_3$
$x^{car}{}_3$
$d^{car}{}_4$
$x^{car}{}_4$
$Z_{car}$

# An integrated model of Scenes, Objects, and Parts



# Predicting object location

**Training set (cars)**

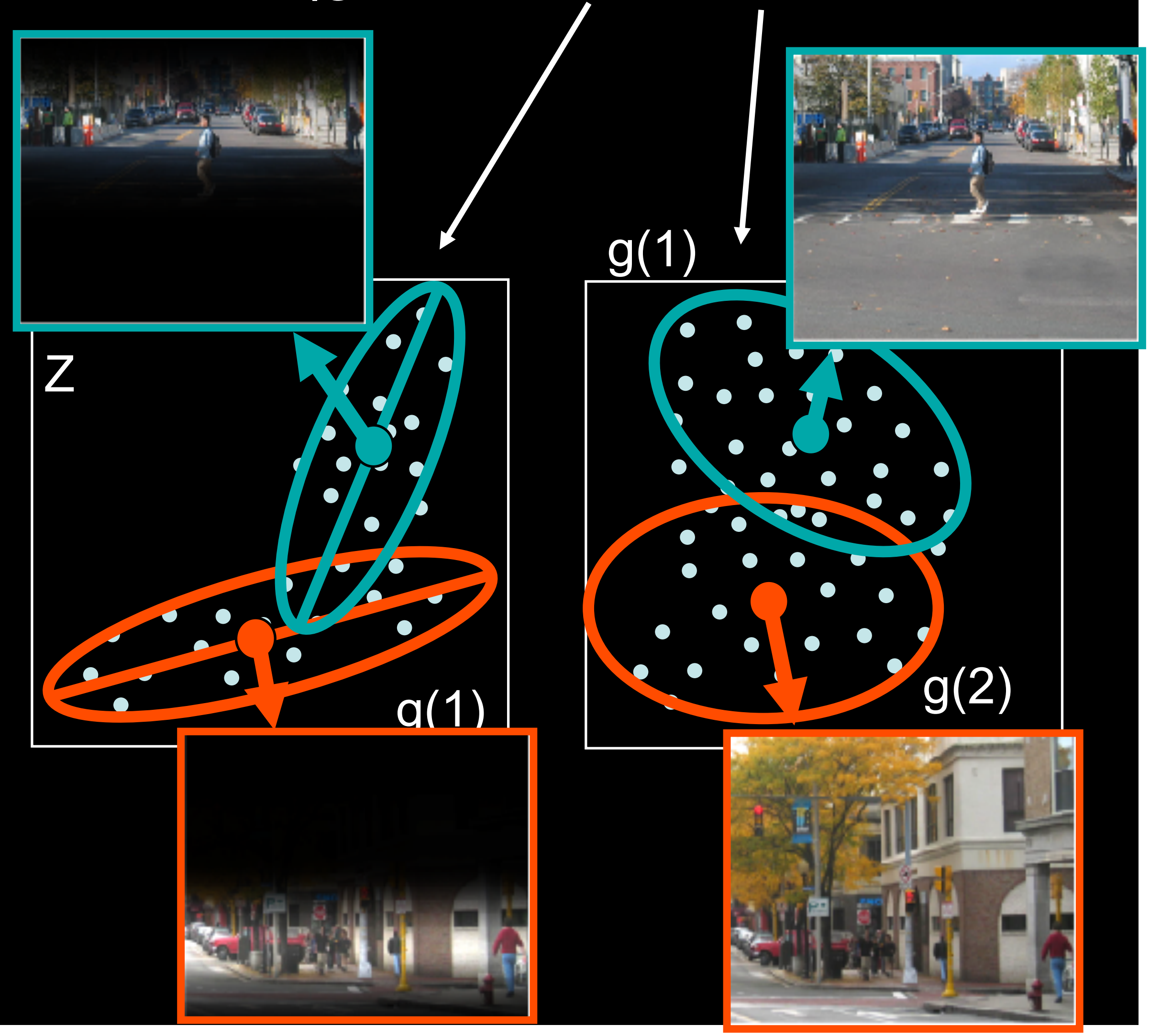
$\longrightarrow \{g^1, z^1\}$

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

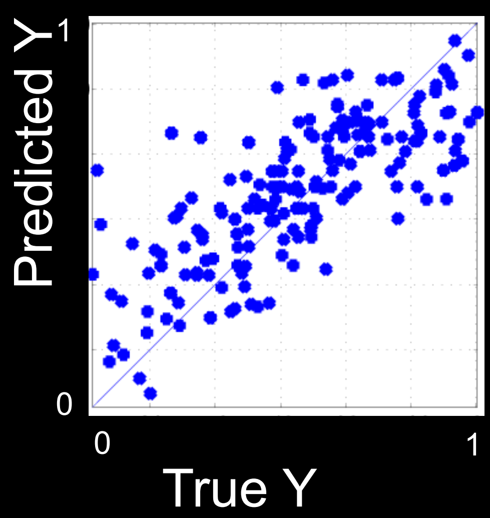
$\longrightarrow \{g^3, z^3\}$

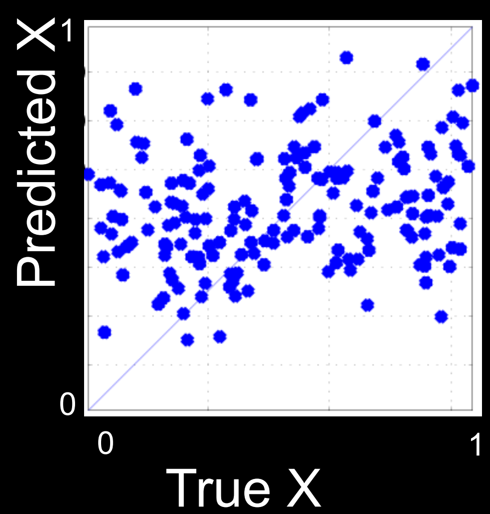
$\vdots$

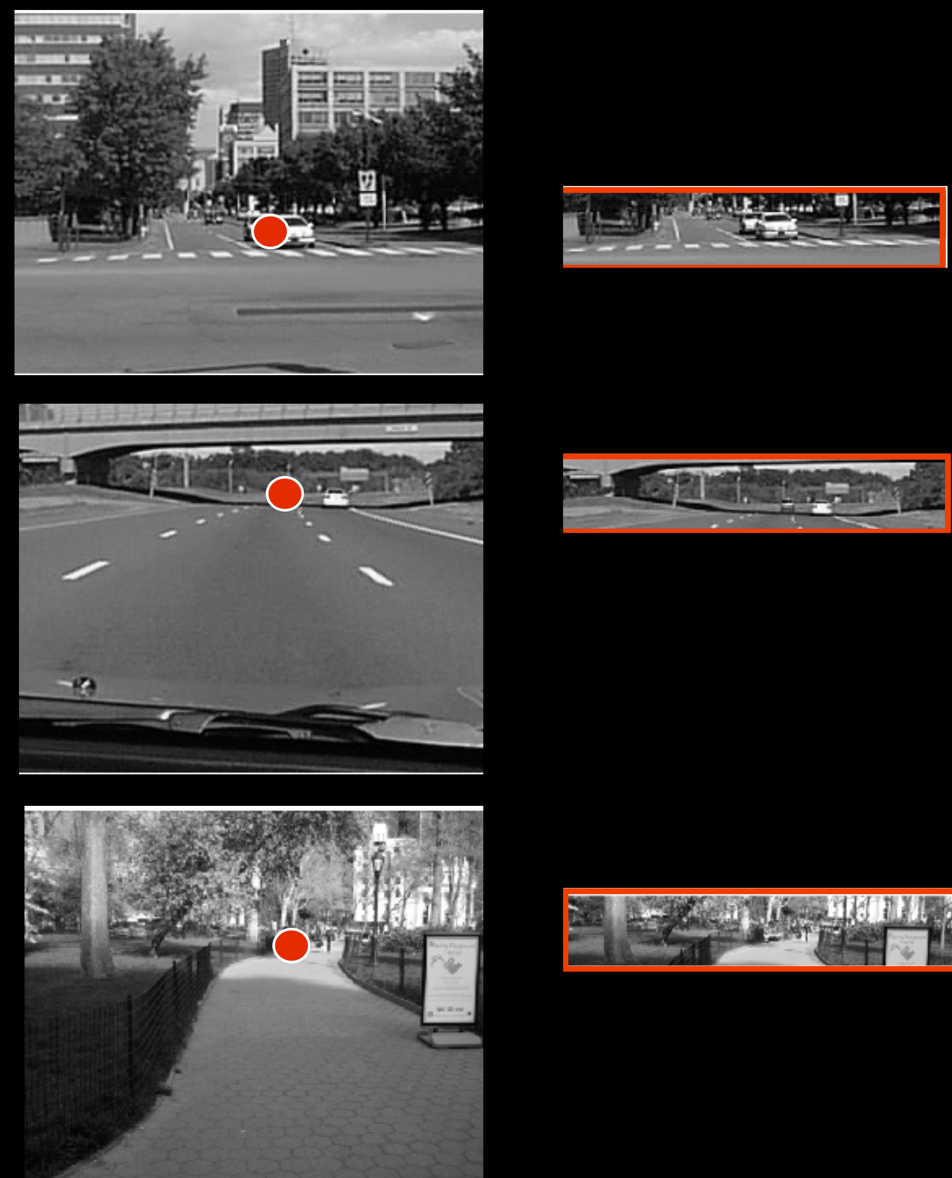
$$Z|g = \sum (A_n\, g + b_n)\, w_n(g)$$



# Predicting location



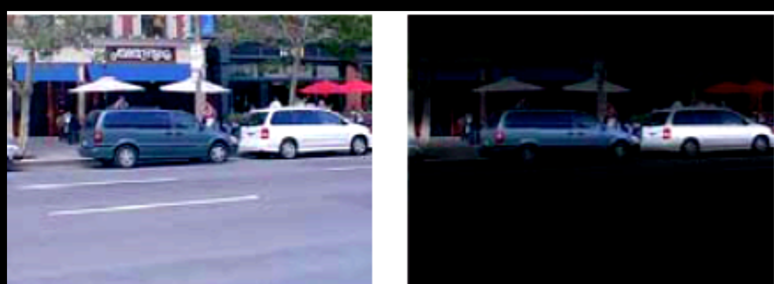



Torralba & Sinha, 2001; Murphy, Torralba, Freeman, 2003; Hoeim, Efros, Hebert. 2006

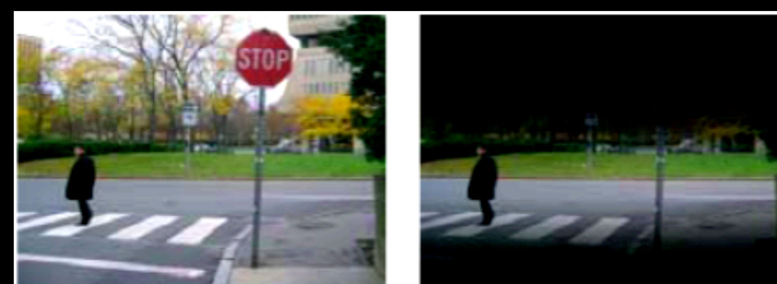
# Car detection without a car detector

screens

keyboard



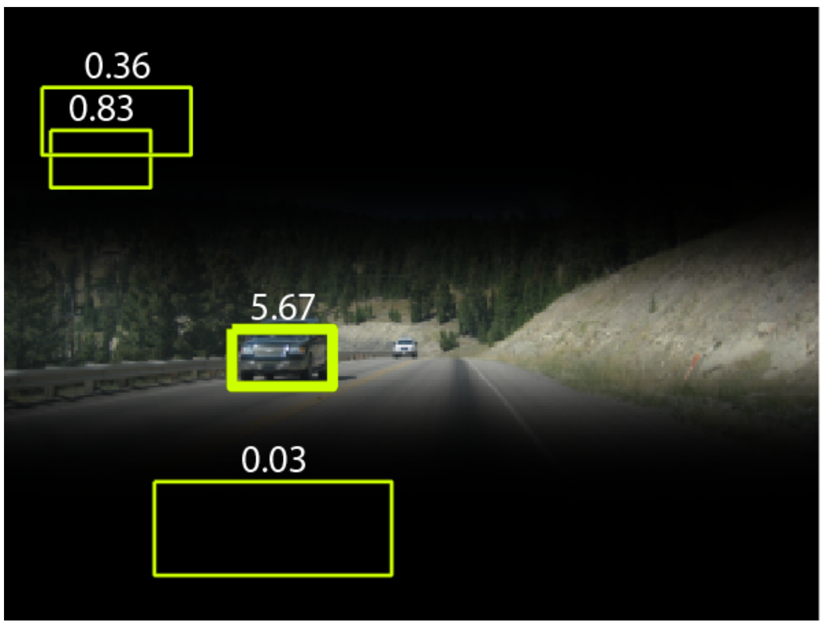
car

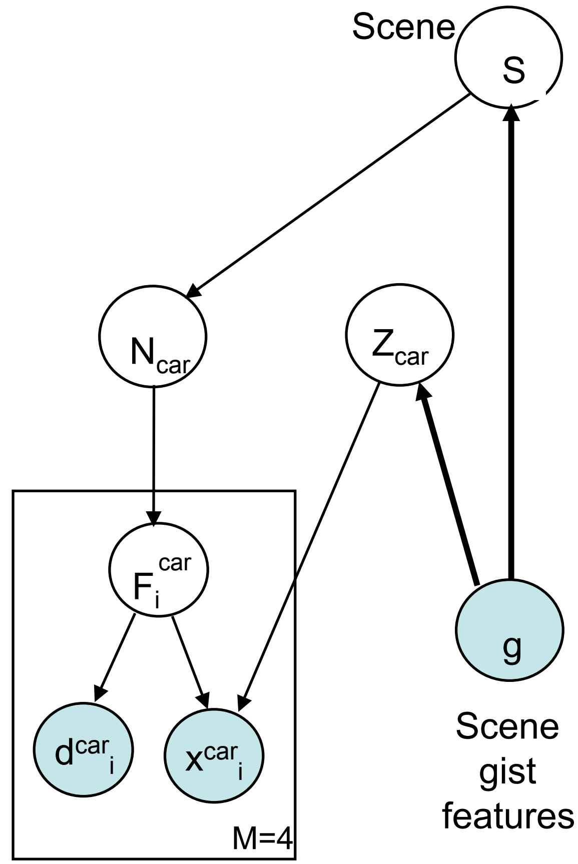


pedestrian

# Detecting faces without a face detector

# An integrated model of Scenes, Objects, and Parts




$$P(F,S \mid x,d,g) \;\alpha\; p(F \mid S)p(S \mid g)\;\; p(x_i \mid g) \prod_{i:F_i=0} N(x_i;\, \mu_b,\, \sigma_b^2) \prod_{i:F_i=1} N(d_i;\, \mu_{tp},\, \sigma_{tp}^2) \prod_{i:F_i=0} N(d_i;\, \mu_{tn},\, \sigma_{tn}^2)$$

# A car out of context …

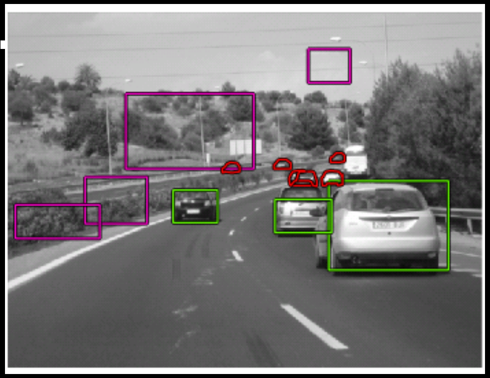
# Failures

- If the detector fails… context can not help

# Failures

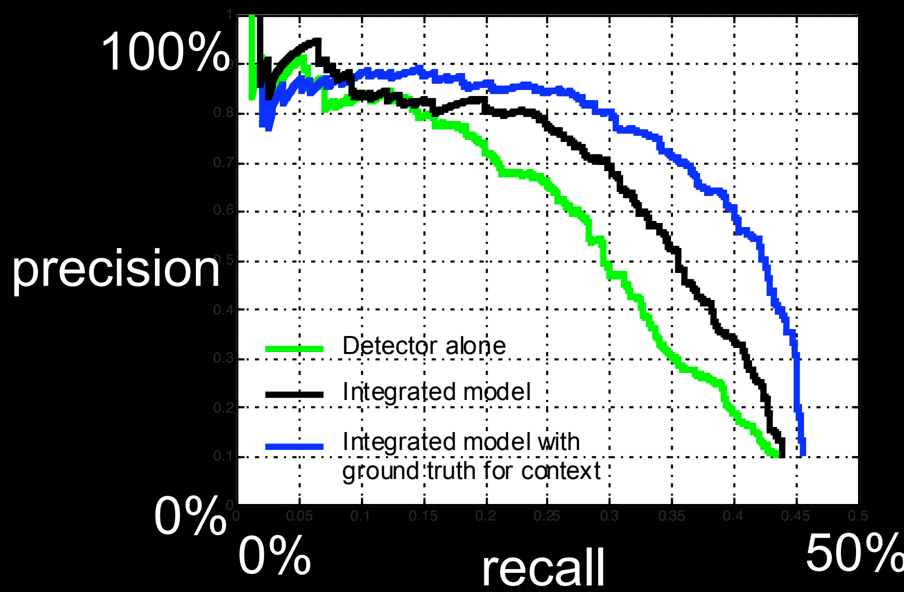
- If the detector fails… context can not help


- If the detector produces a contextually coherent false alarm, context will increase the error.

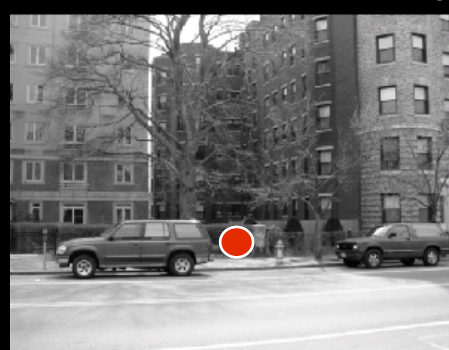



# Benefits of context

- Increases performances




- Increases efficiency



Reduced search space

# 3D City Modeling using Cognitive Loops

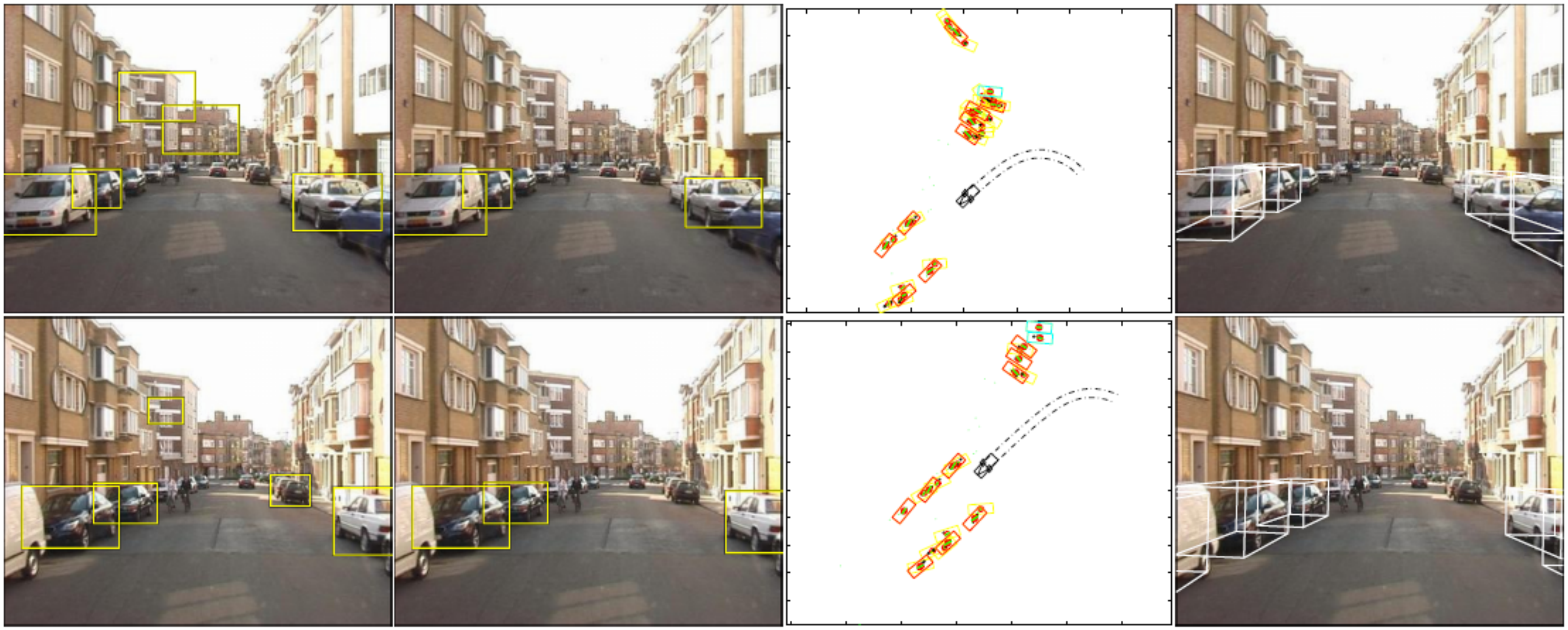


Figure 6. Stages of the recognition system: (a) initial detections before and (b) after applying ground plane constraints, (c) temporal integration on reconstructed map, (d) estimated 3D car locations, rendered back into the original image.

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

# Large databases

# Why is scene understanding hard? Scenes are unique



# But not all scenes are so original








# But not all scenes are so original





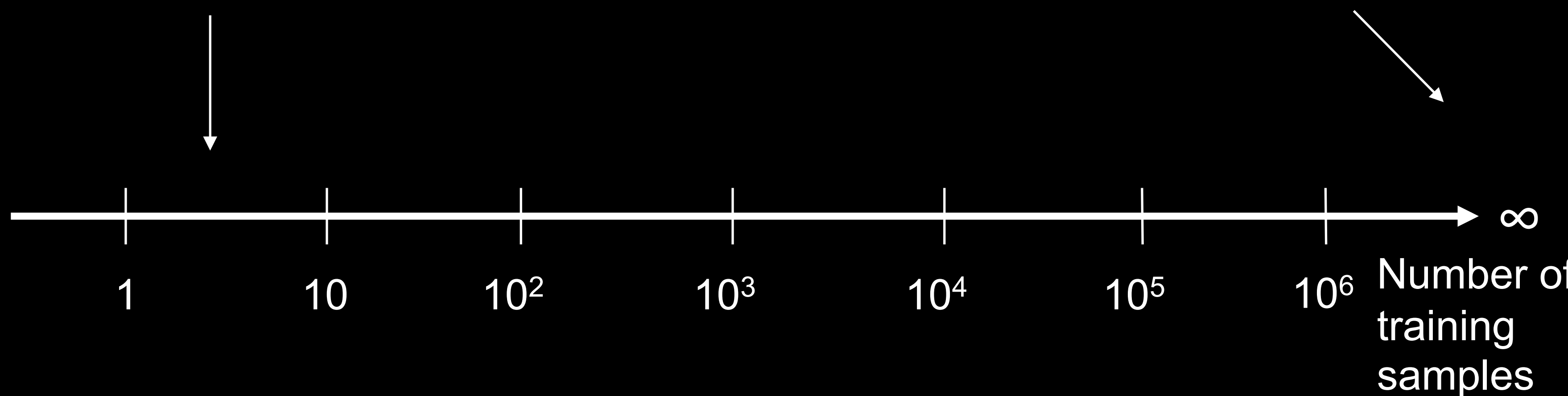


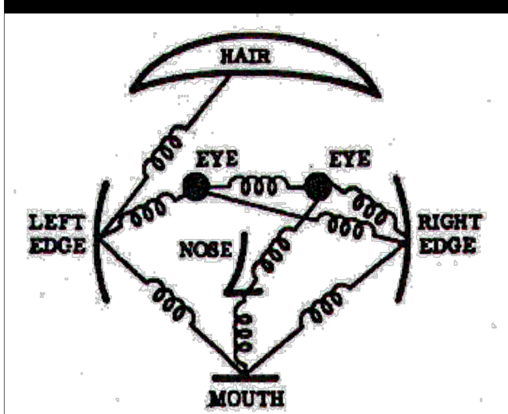


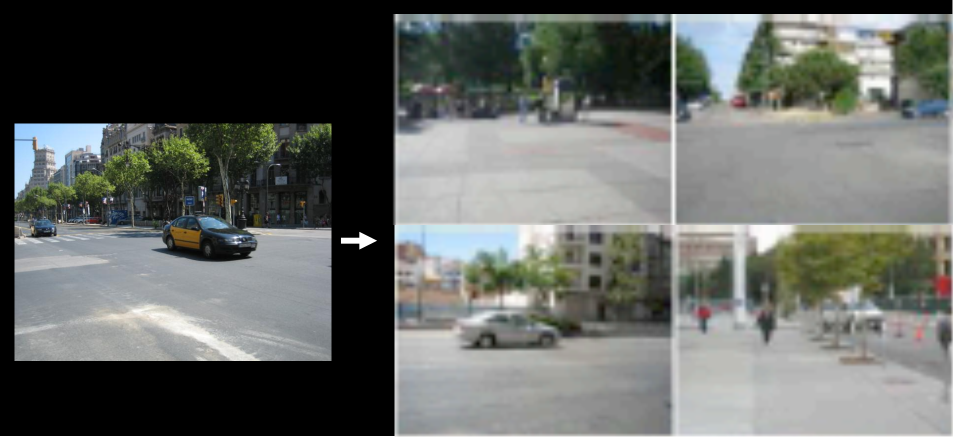
# The two extremes of learning

**Extrapolation problem**
Generalization
Transfer learning

**Interpolation problem**
Correspondence
Finding the differences

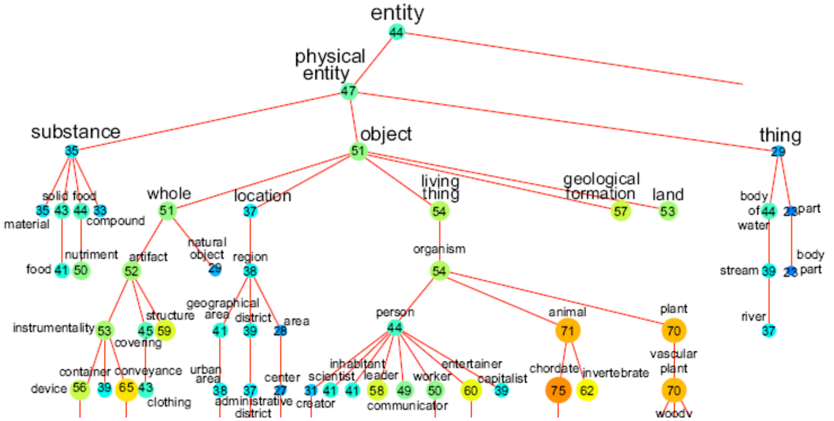


# 80.000.000 images

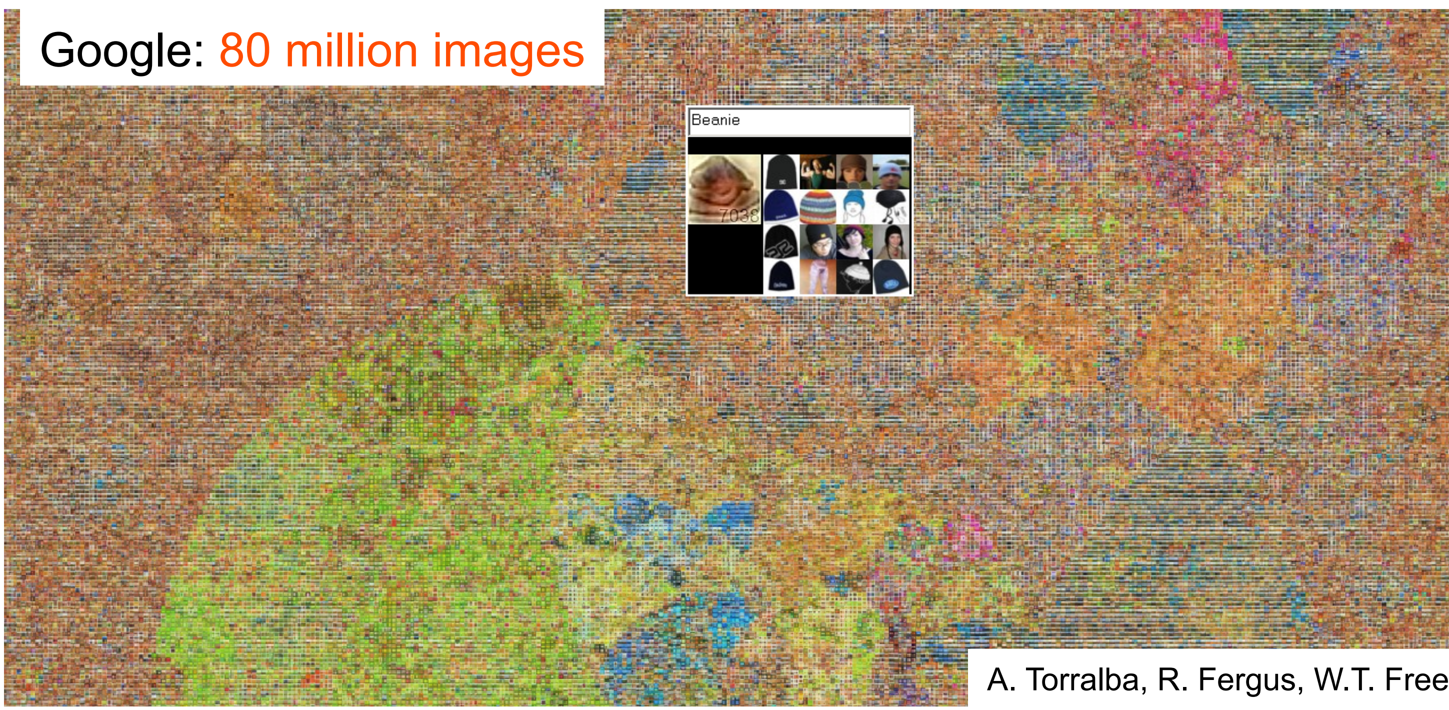
75.000 non-abstract nouns from WordNet




7 Online image search engines

Ask Images
Cydral
Google Image Search
webshots
picsearch
altavista
flickr

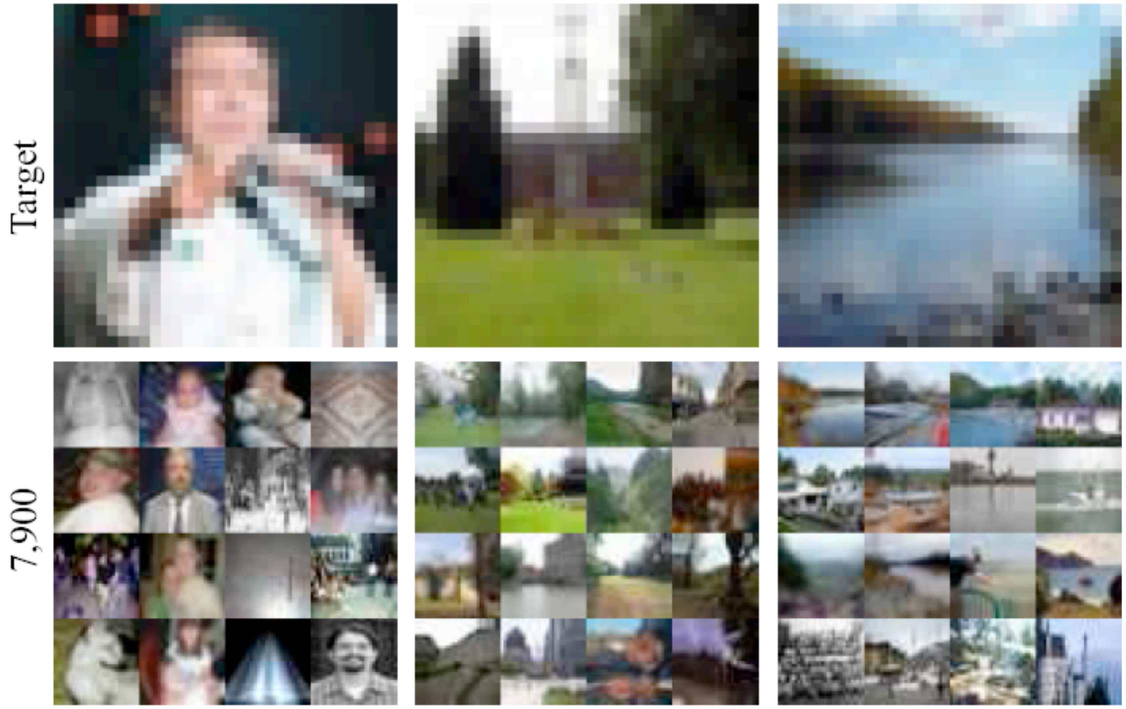
And after 1 year downloading images




Google: 80 million images

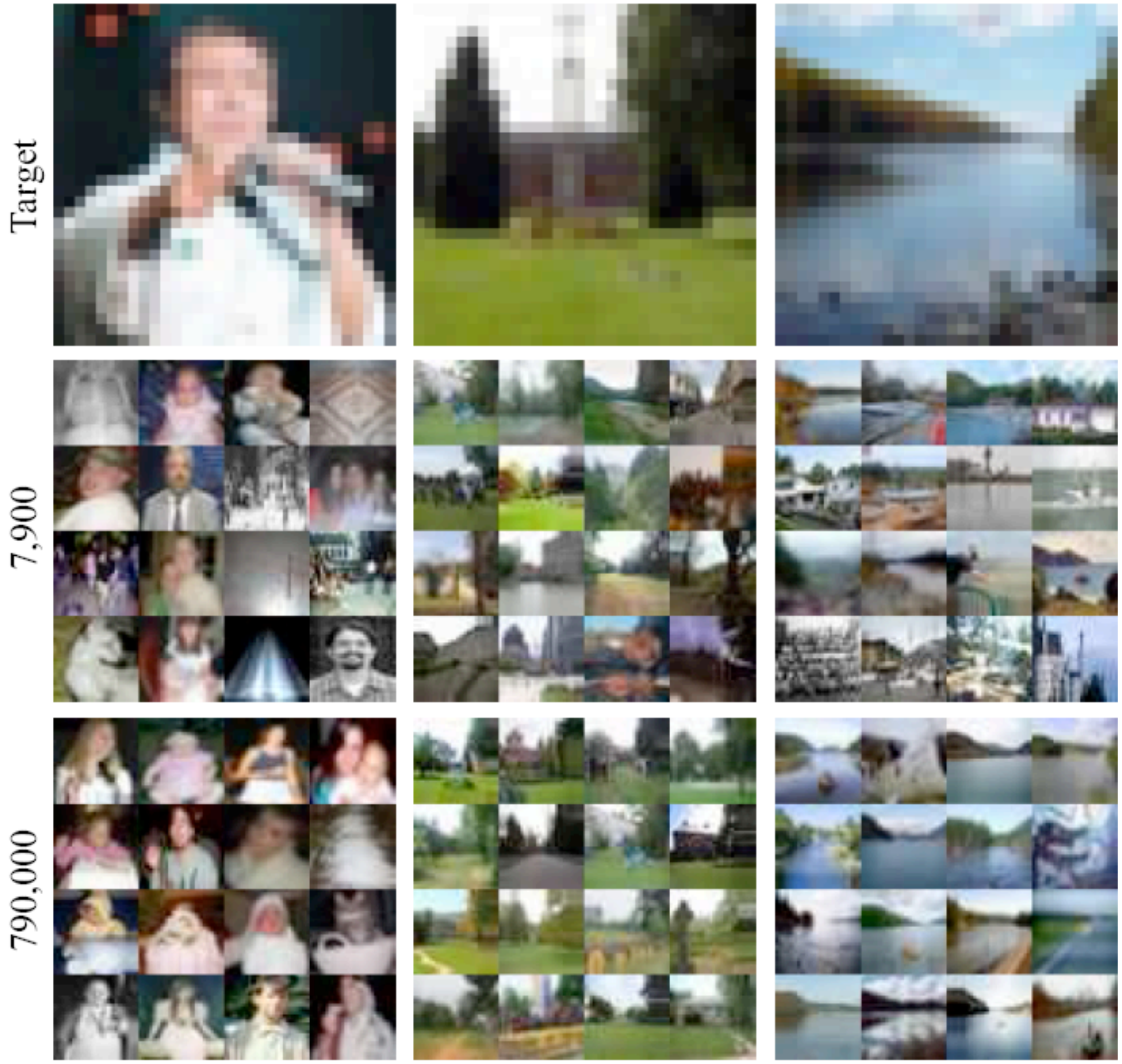
A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

# The Power Of Lots Of Images



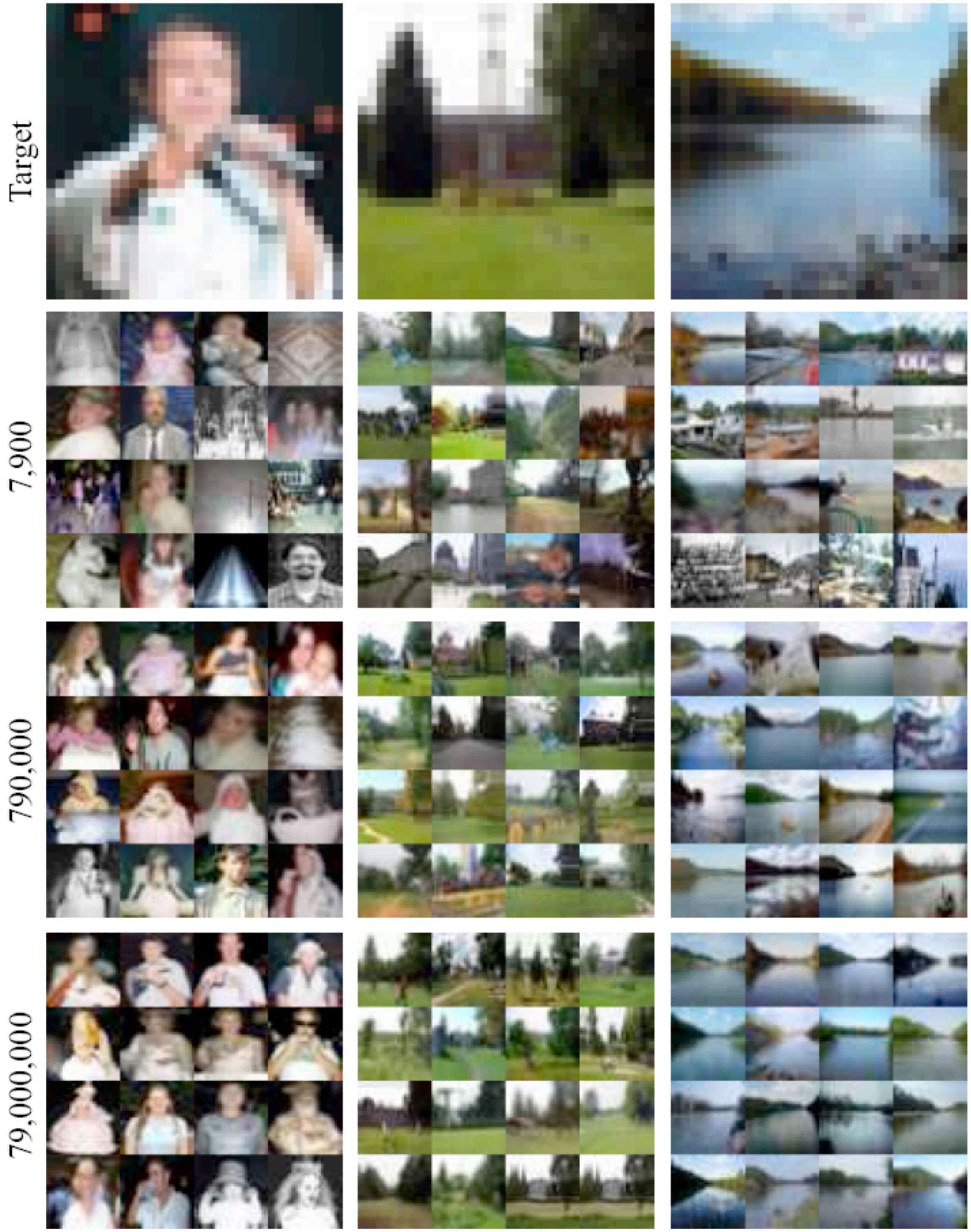

A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

# The Power Of Lots Of Images




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
- …

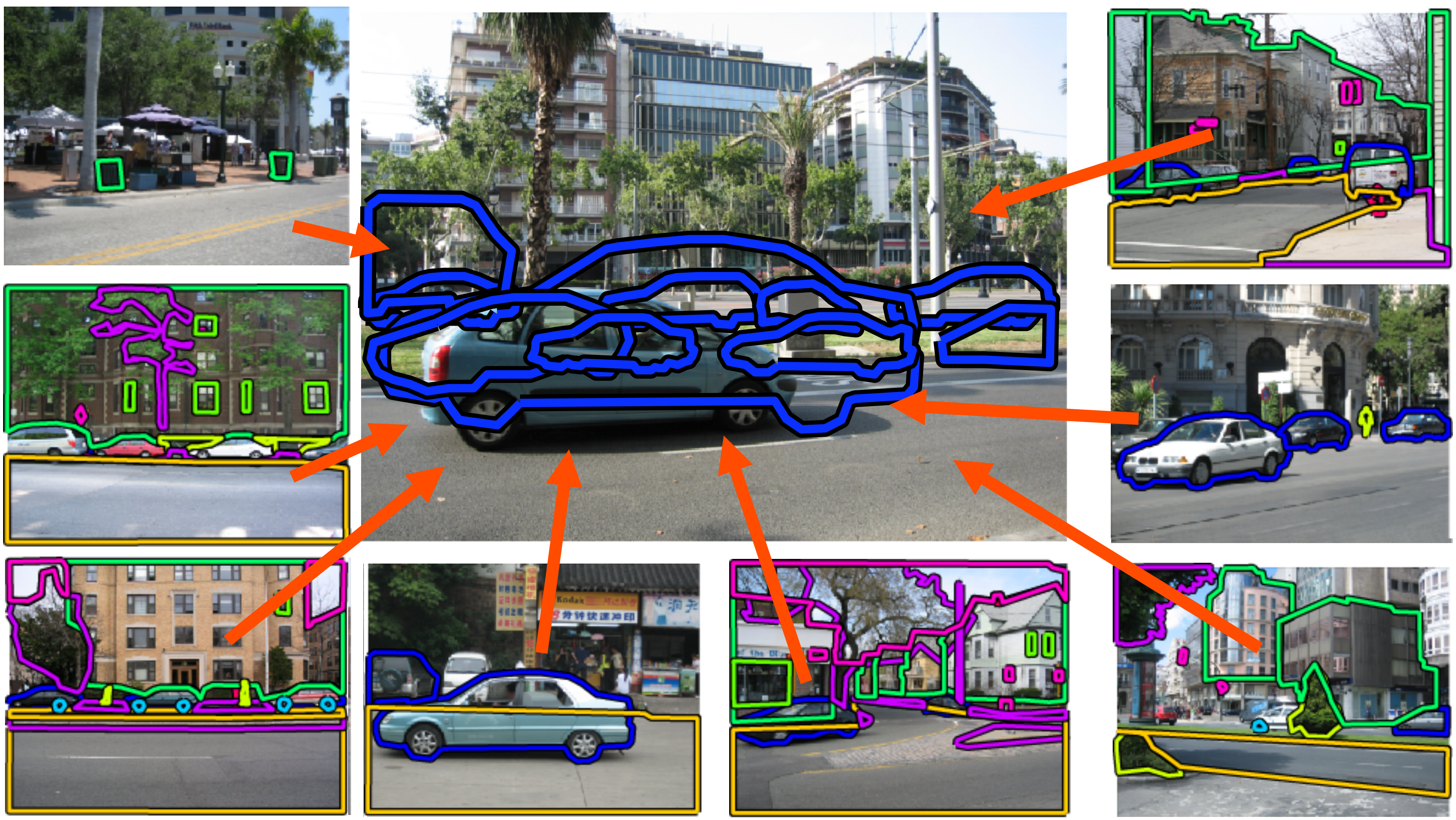
Nearest neighbors

- Labels
- Motion
- Depth
- …

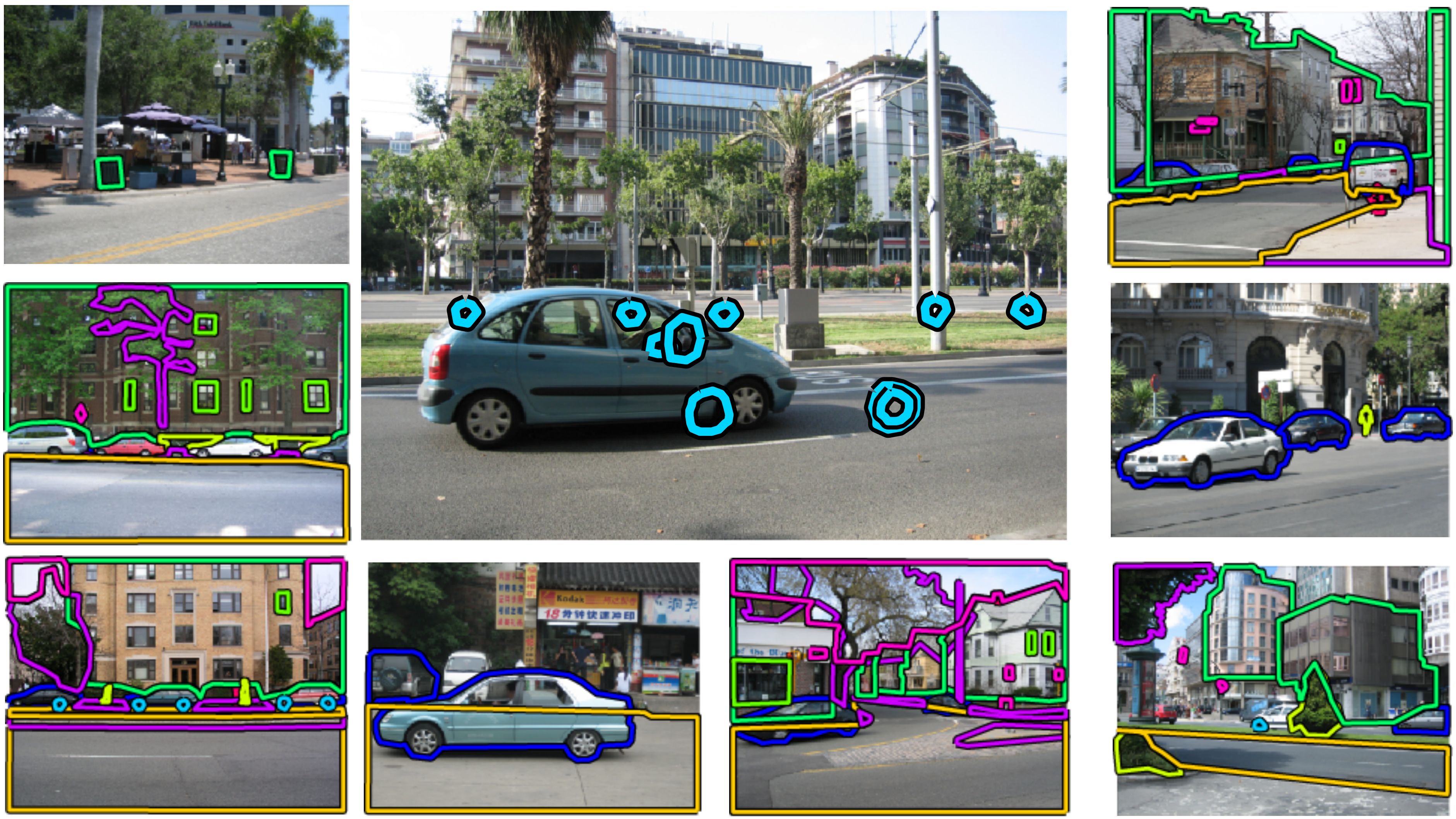
The space of world images

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

# With a good image similarity and a lot of data…



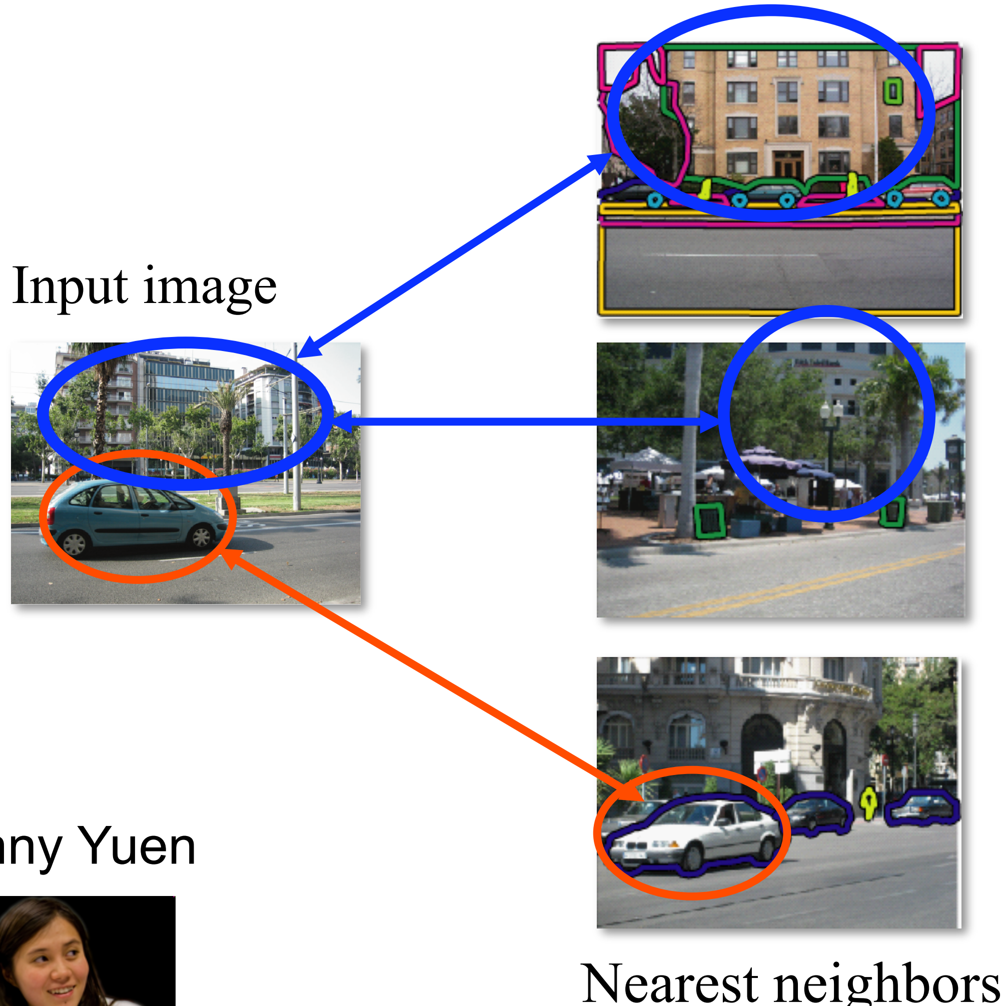
# With a good image similarity and a lot of data…



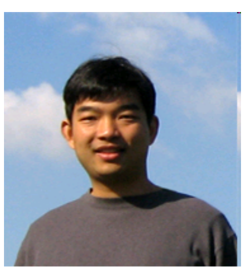
# With a good image similarity and a lot of data…

# SIFT flow:

## dense correspondence across different scenes



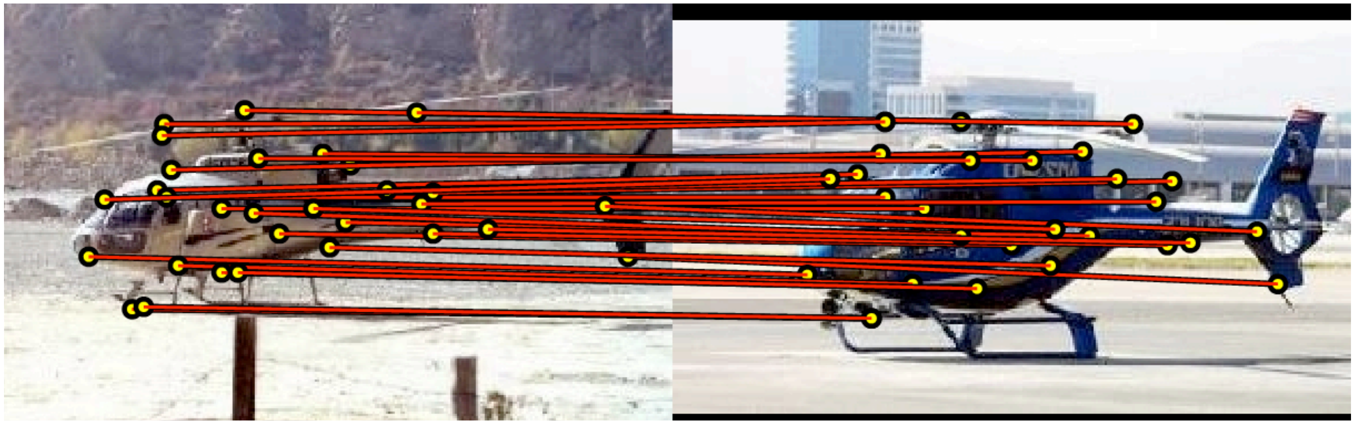

Ce Liu



Jenny Yuen

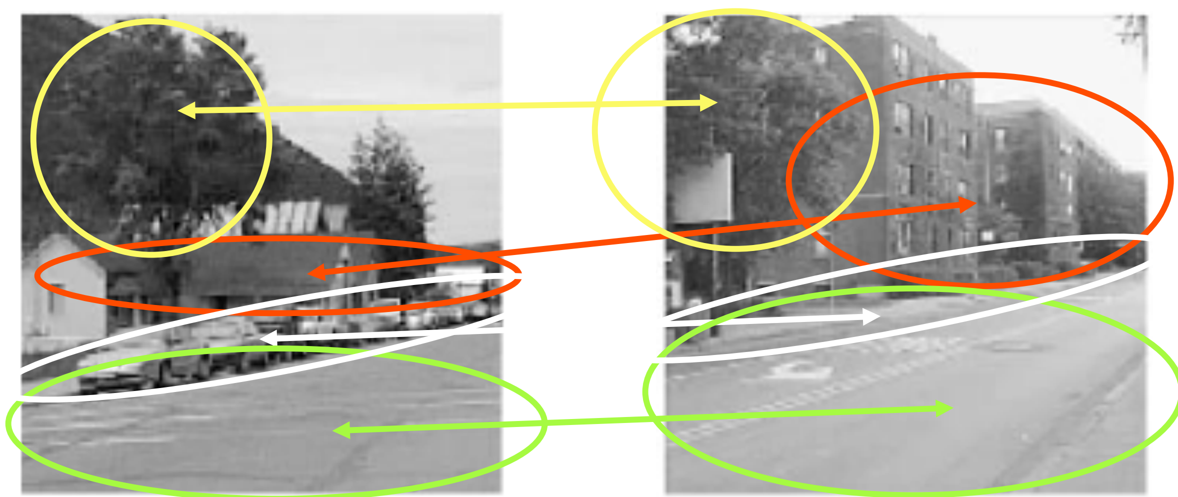
Liu, Yuen, Torralba. CVPR 2009.

Berg, Berg, Malik CVPR 2005



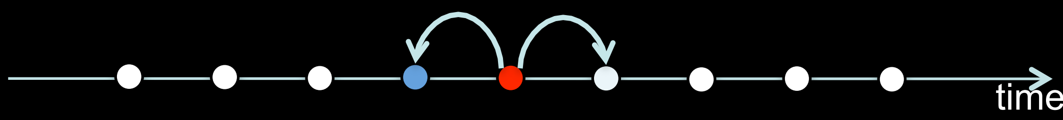
Yuille ’91; Brunelli & Poggio ’93; Lades, v.d. Malsburg et al. ’93; Cootes, Lanitis, Taylor et al. ’95; Amit & Geman ‘95, ‘99 ; Perona et al. ‘95, ‘96, ’98, ’00; Felzenszwalb & Huttenlocher ’00

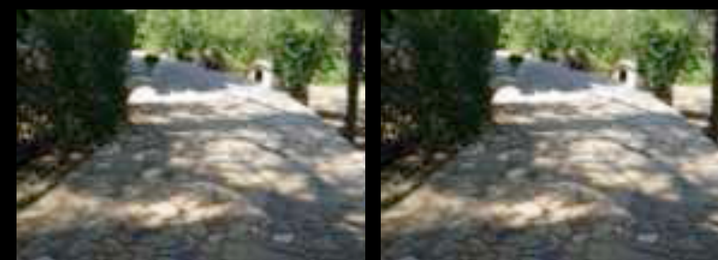
Liu, Yuen, Torralba CVPR 2009



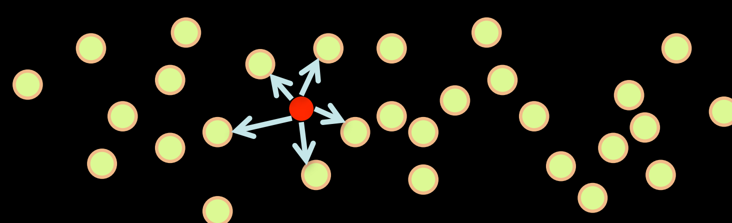
Object recognition by scene alignment

The simplest alignment problem: matching two consecutive frames



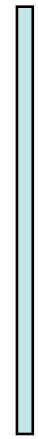



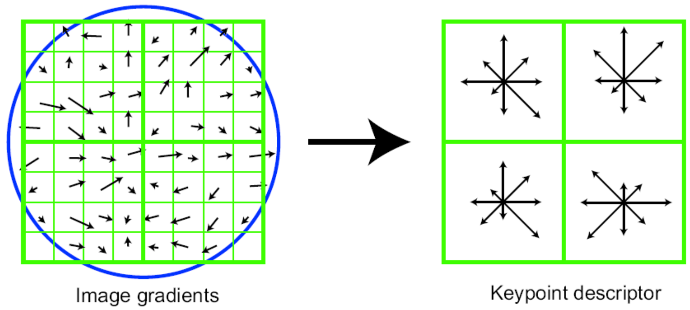
Hypothesis: if we have a dataset that is large enough, we can find an image that is close enough to our input.



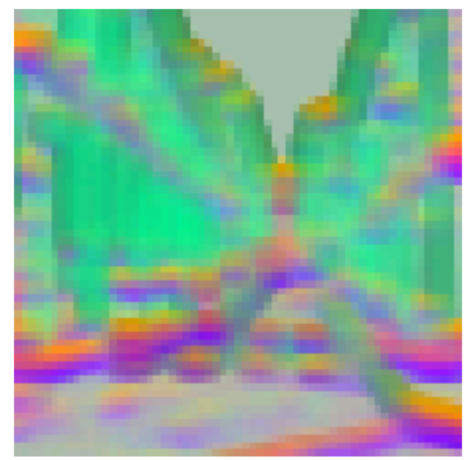
# Dense SIFT descriptor

128 dimensions/pixel






SIFT (scale-invariant feature transform)
- 8 orientations, 4×4 cell grid
- Characterize local image gradient

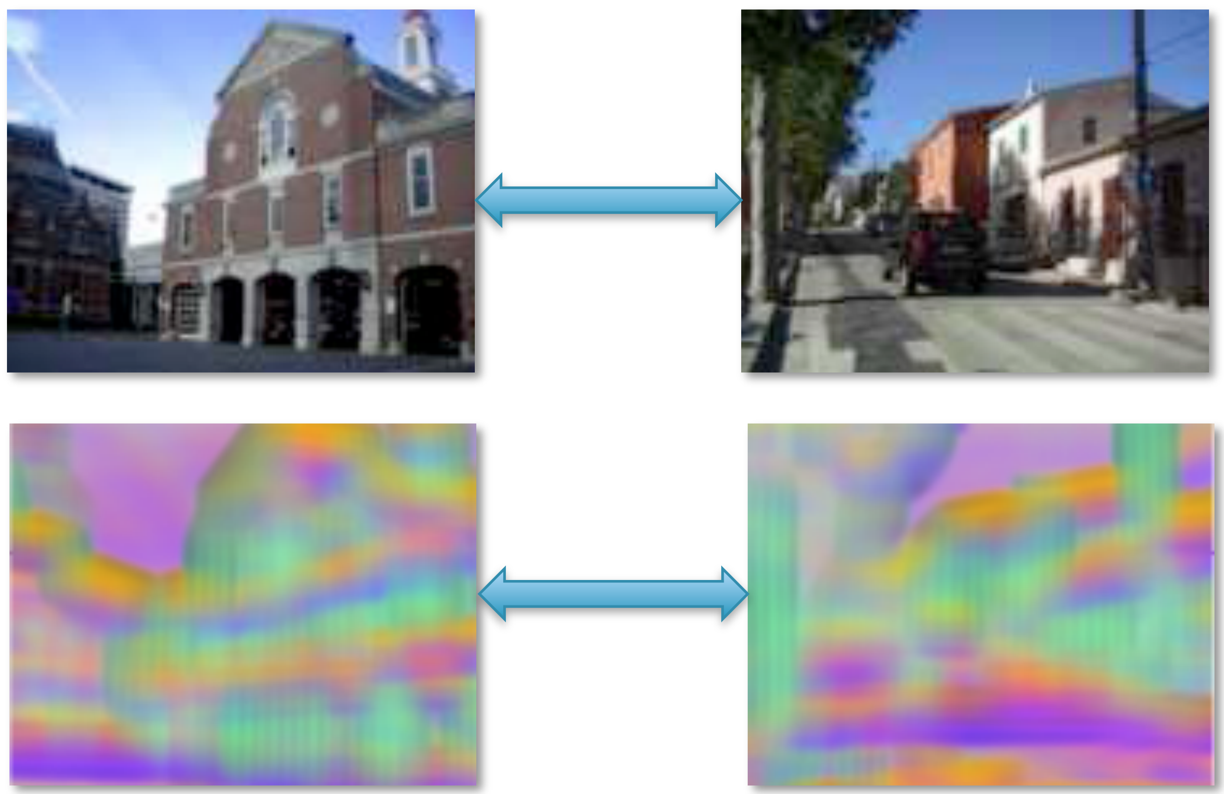


SIFT Visualization: map 128 dimensions in 3D color space

# Matching dense SIFT descriptors

RGB images

SIFT images



# Objective function of SIFT flow

- The energy function is similar to that of optical flow:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min\Big(\big\|s_1(\mathbf{p}) - s_2(\mathbf{p}+\mathbf{w}(\mathbf{p}))\big\|_1, t\Big) +$$

Data term (reconstruction)

$$\sum_{\mathbf{p}} \eta\Big(|u(\mathbf{p})| + |v(\mathbf{p})|\Big) +$$

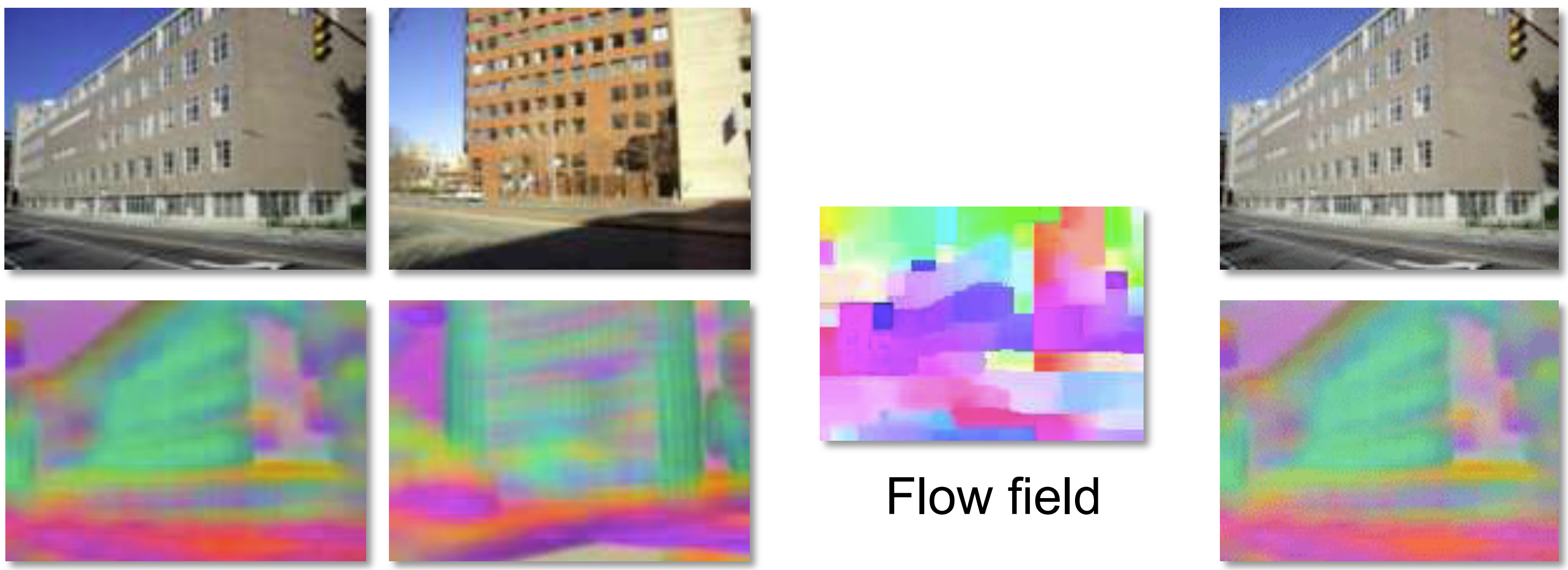
Small displacement bias

$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} \min\Big(\alpha|u(\mathbf{p}) - u(\mathbf{q})|, d\Big) + \min\Big(\alpha|v(\mathbf{p}) - v(\mathbf{q})|, d\Big)$$

Smoothness term

- **p**, **q**: grid coordinate, **w**: flow vector, *u*, *v*: *x*- and *y*-components, $s_1$, $s_2$: SIFT descriptors

# Retrieval results



RGB

SIFT

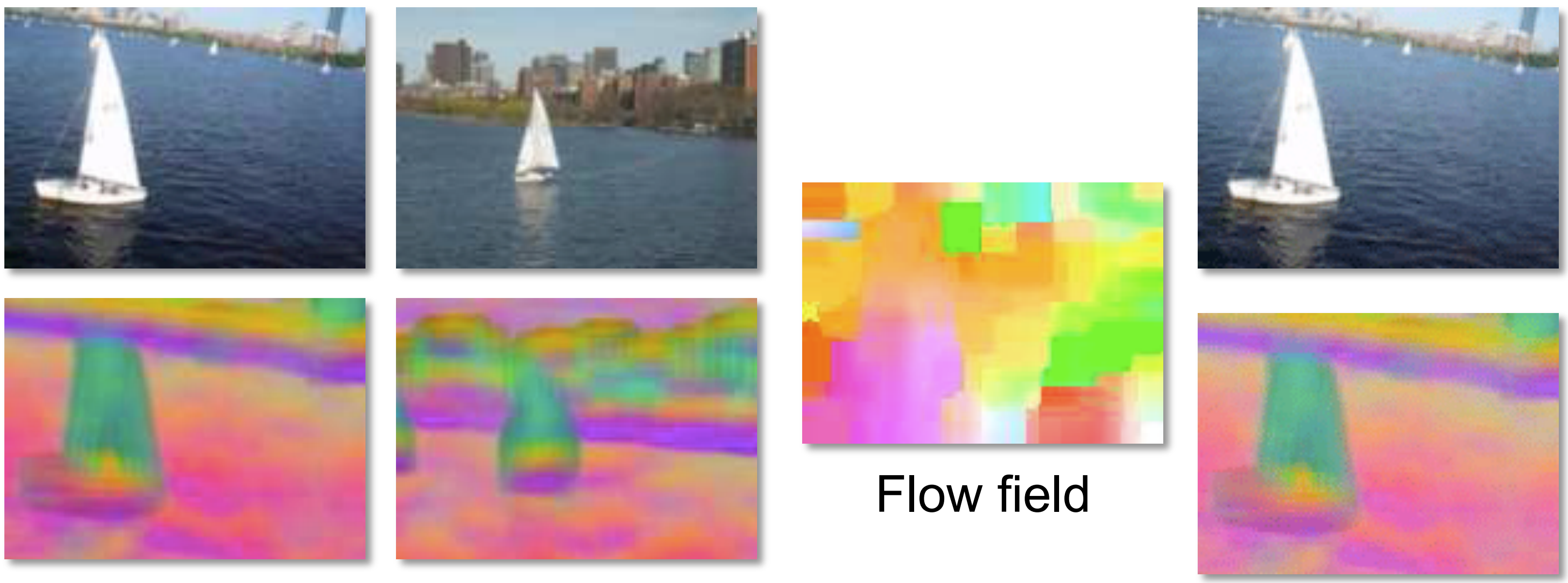
Query

Best match

Flow field

Query & warped
Best match

# Retrieval results



RGB

SIFT

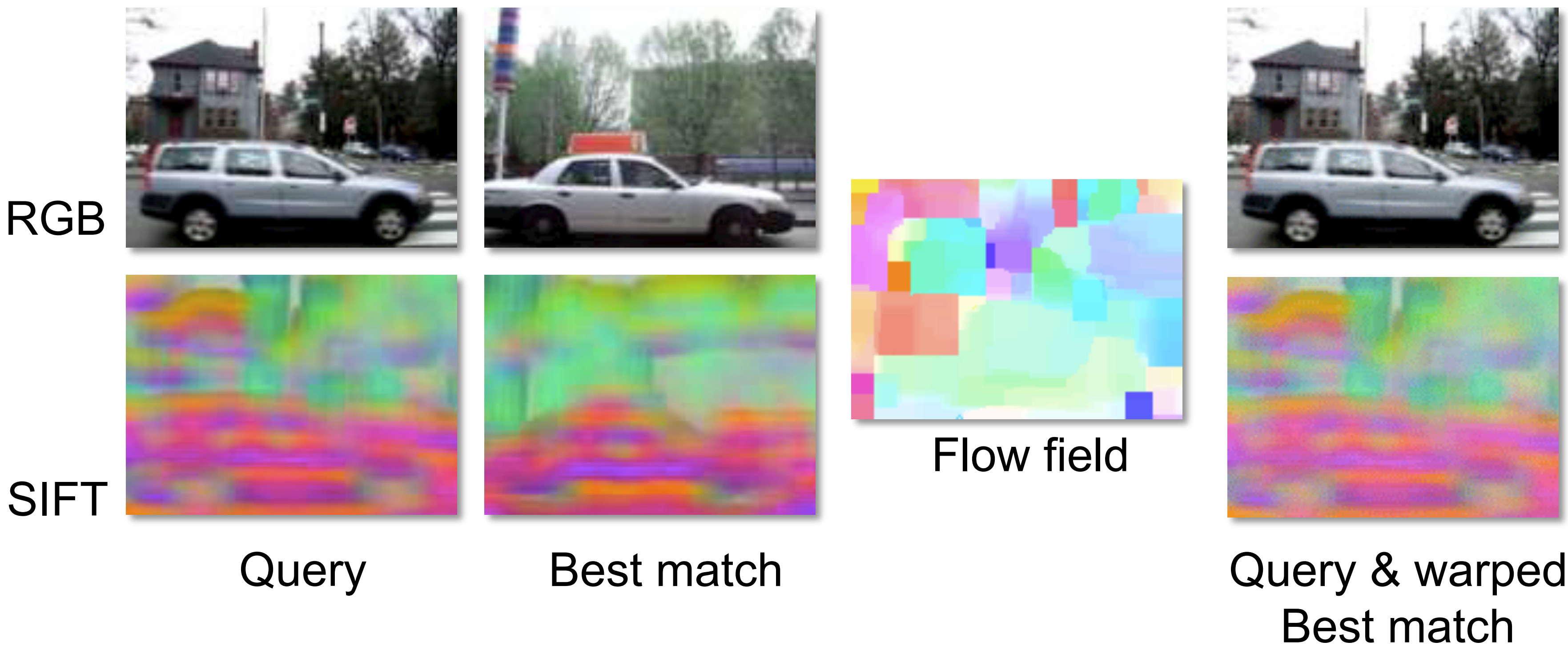
Query

Best match

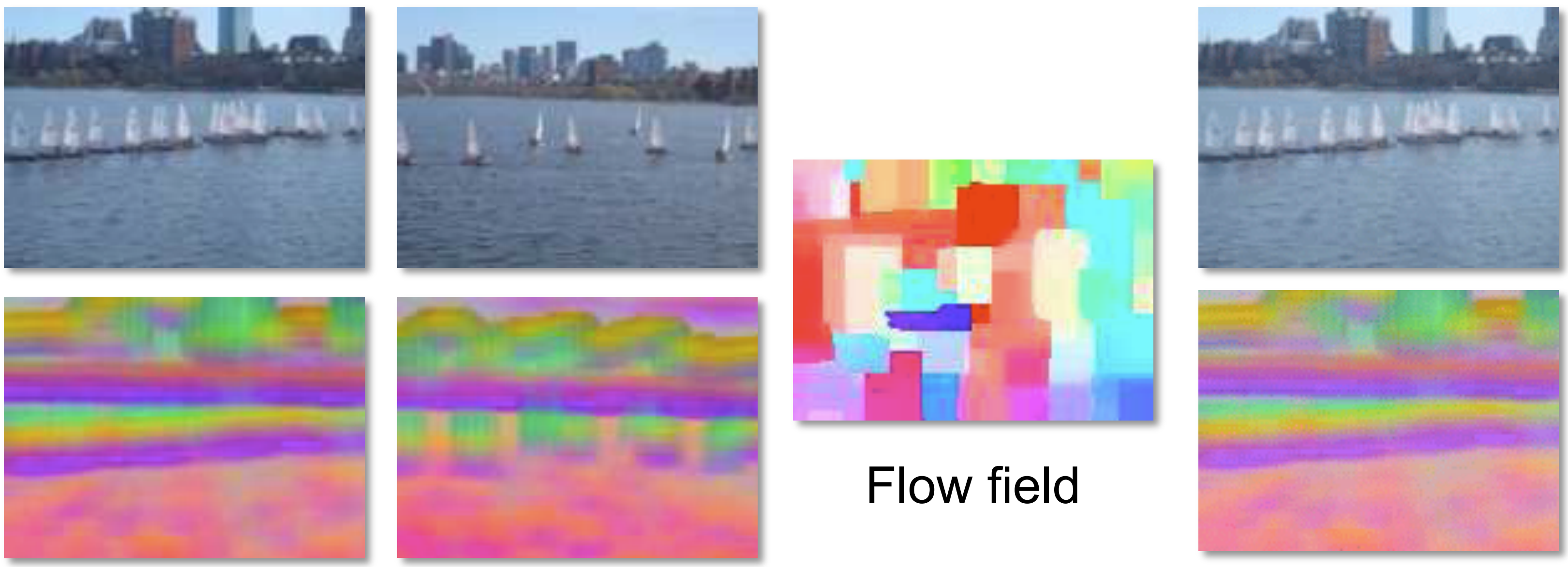
Flow field

Query & warped Best match

# Retrieval results



# Retrieval results



RGB

SIFT

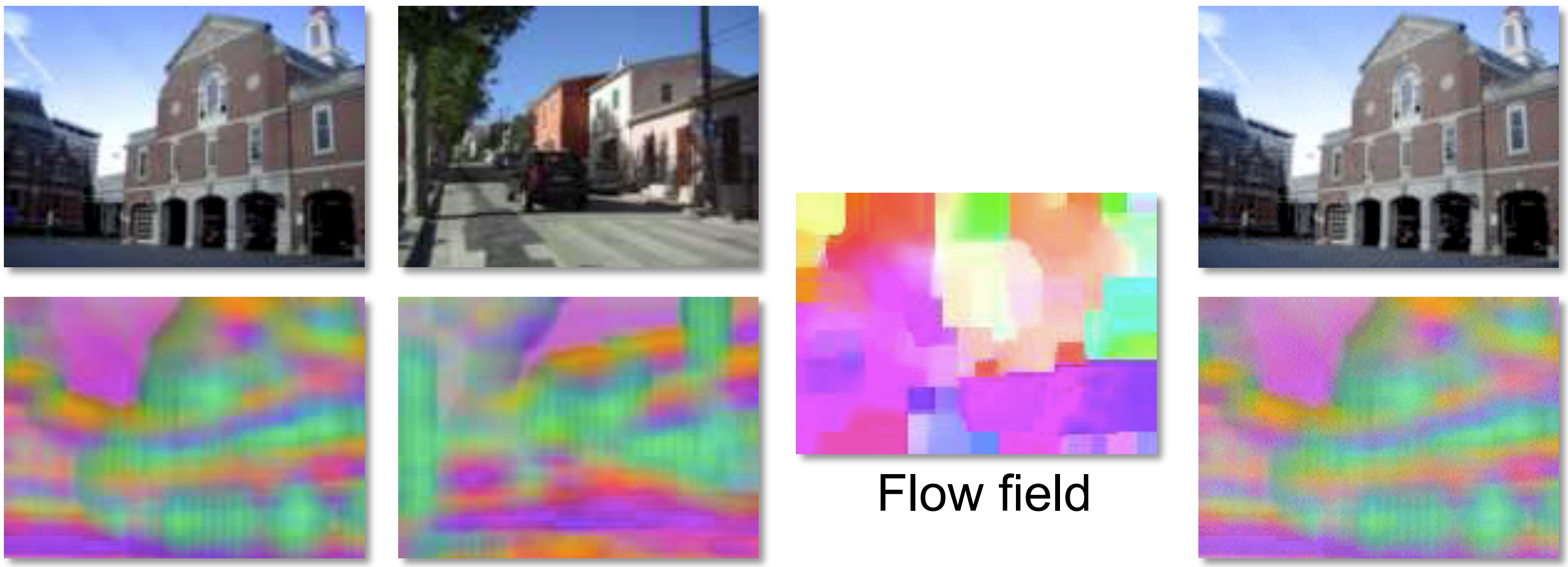
Query

Best match

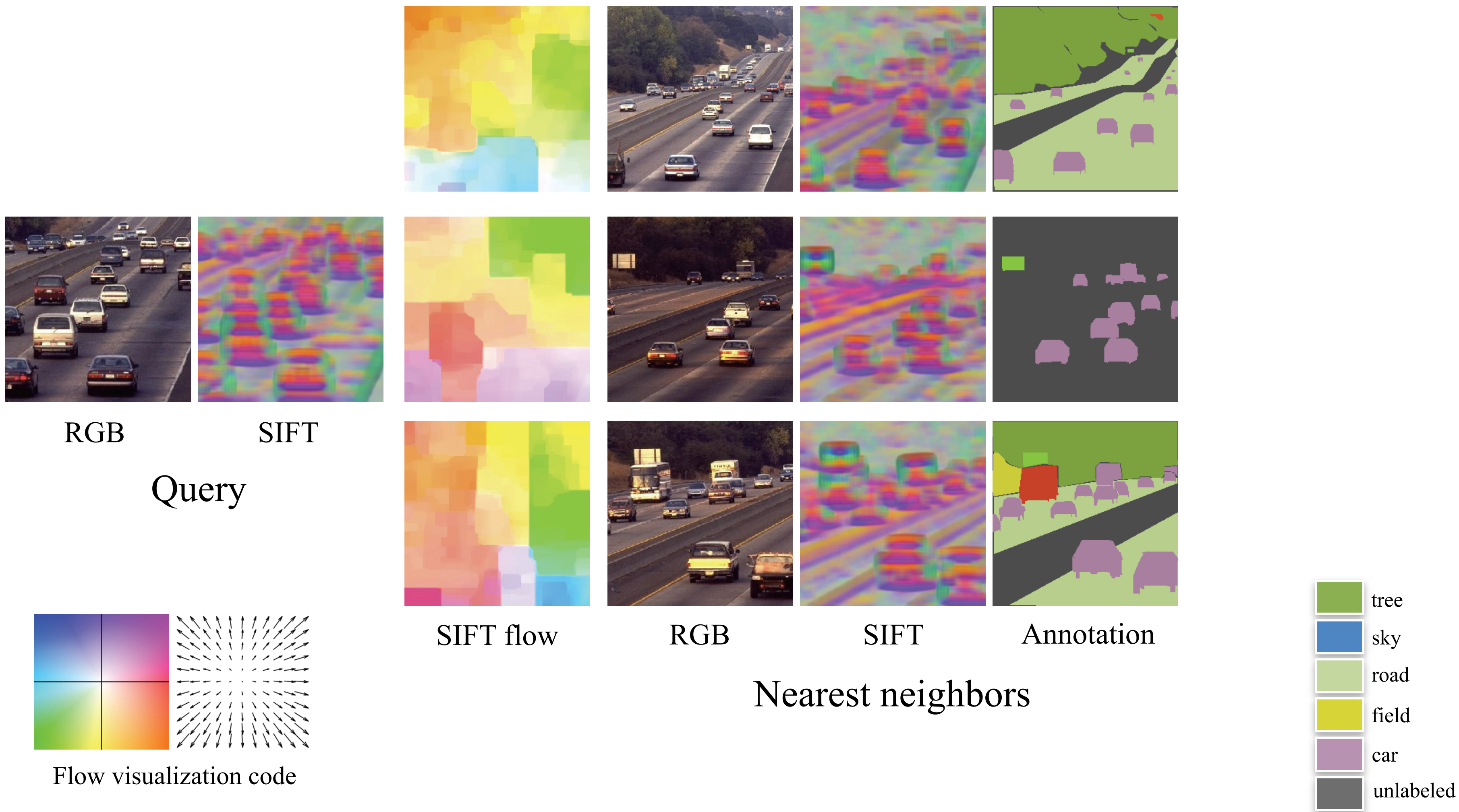
Flow field

Query & warped
Best match

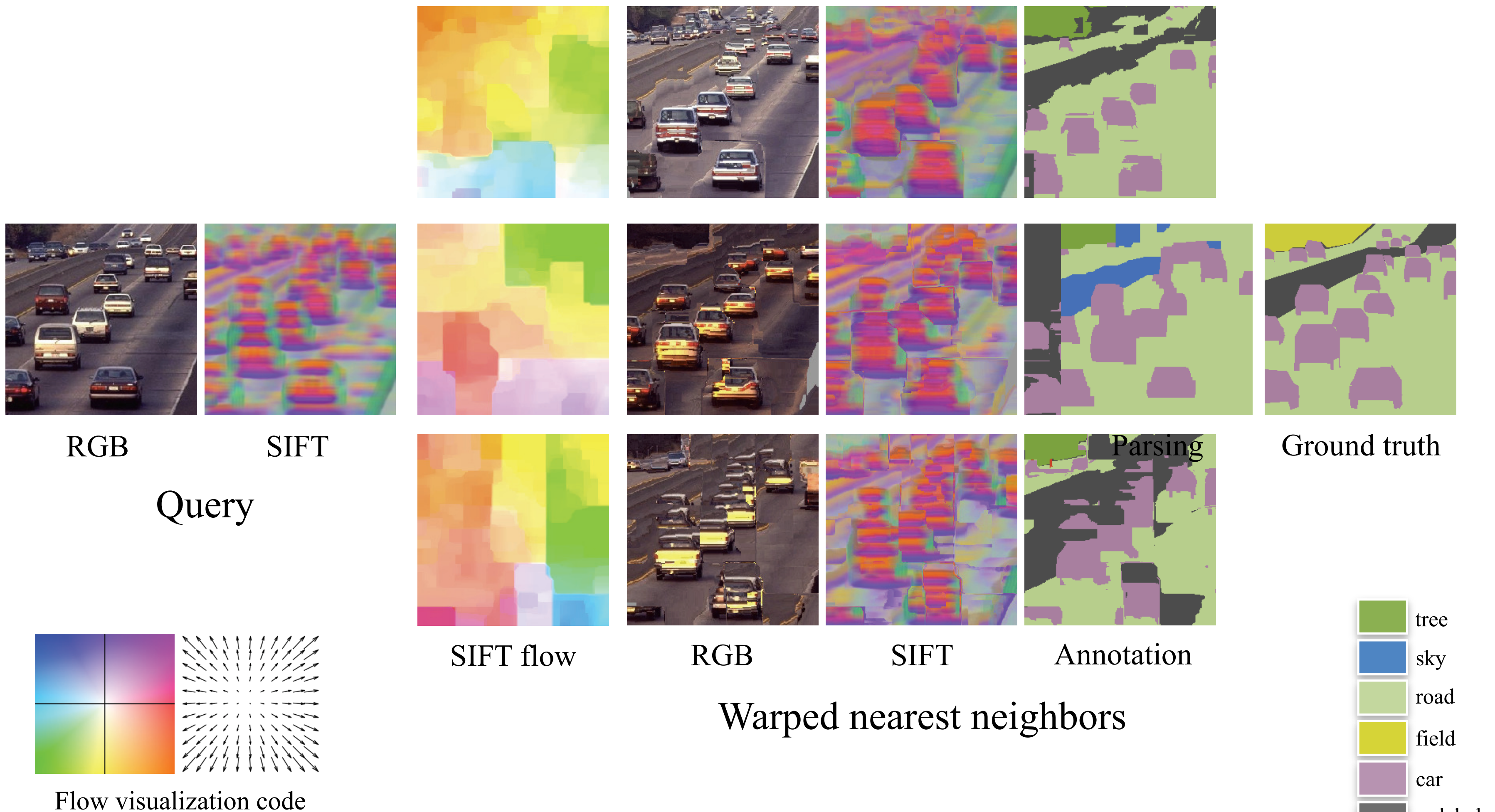
# Retrieval results



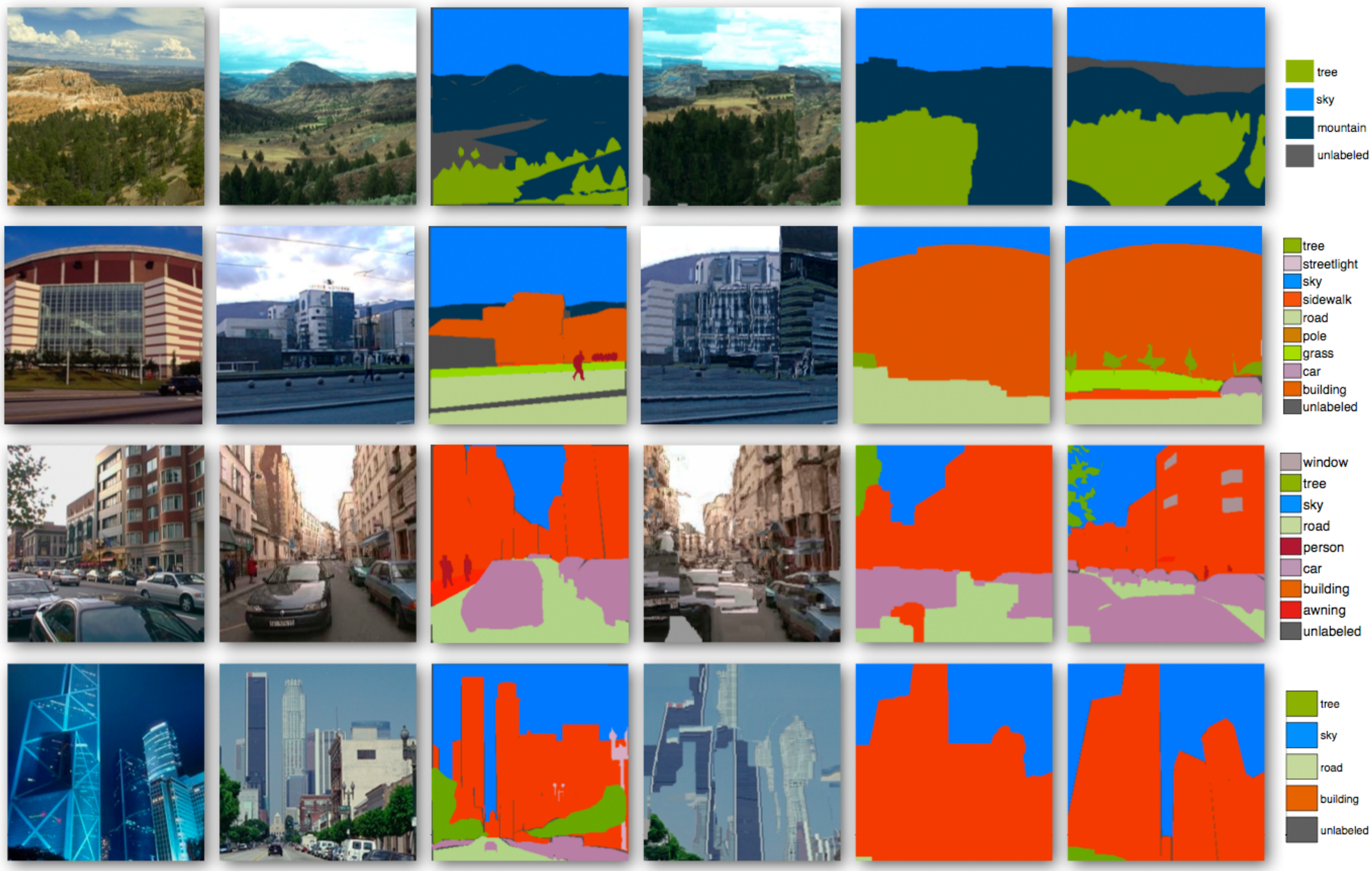
# System overview



# System overview



# Scene parsing results (2)




Query | Best match | Annotation of best match | Warped best match to query | Parsing result | Ground truth

# Predicting events

# Predicting events

Query

Query

Retrieved video

Query

Retrieved video

Synthesized video

Query

Retrieved video

Synthesized video

Query

Retrieved video

Synthesized video

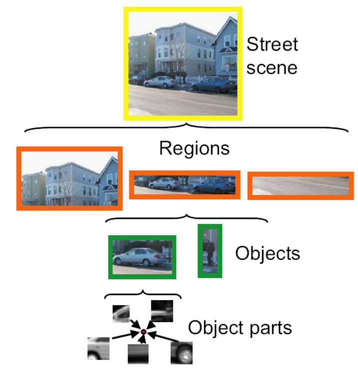
Query

Retrieved video

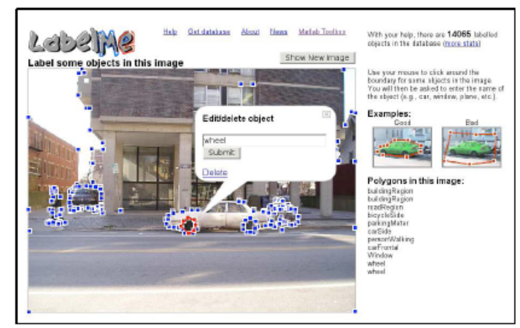
Synthesized video

# Summary

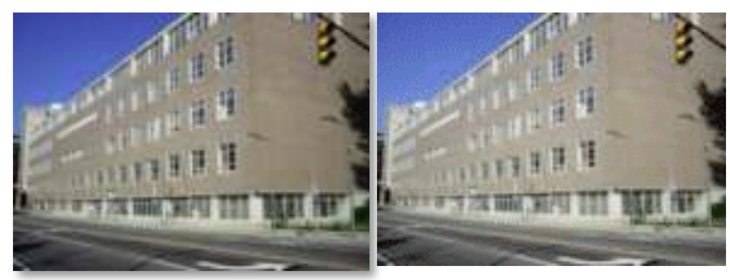
- Gist of the scene & context models for object and scene recognition




- Building datasets for computer vision




- Exploiting large databases and non-parametric methods for scene understanding



We have better low and mid-level vision
Better learning algorithms
Lot’s of computational power
And lot’s of data

…

We are running out of excuses