Recognizing and Learning Object Categories

Li Fei-Fei, Stanford

Rob Fergus, NYU

Antonio Torralba, MIT

Outline

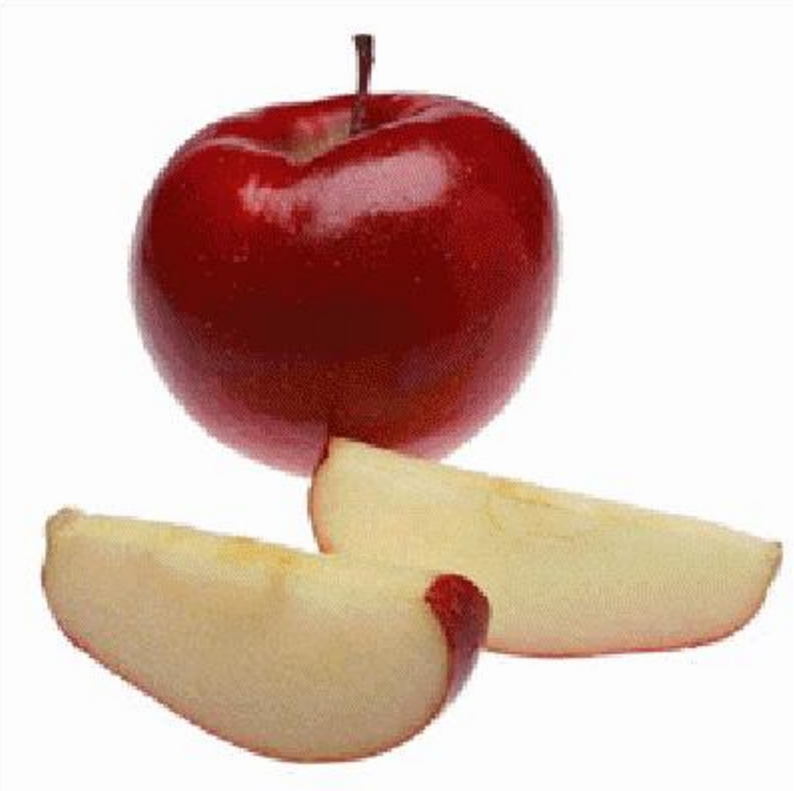
2. Single object categories

  - Bag of words
  - Part-based
  - Discriminative
  - Detecting single objects in contexts
  - 3D object classes

3. Multiple object categories

  - Recognizing a large number of objects
  - Recognizing multiple objects in an image
  - Objects and annotations

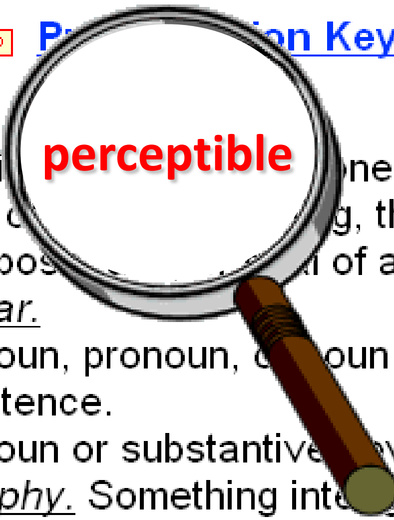
4. Object-related datasets and challenges



Dictionary.com

object Search

Dictionary Thesaurus Encyclopedia Web

ob·ject P 

on Key (ŏbʹjĭkt, -jĕktʹ)

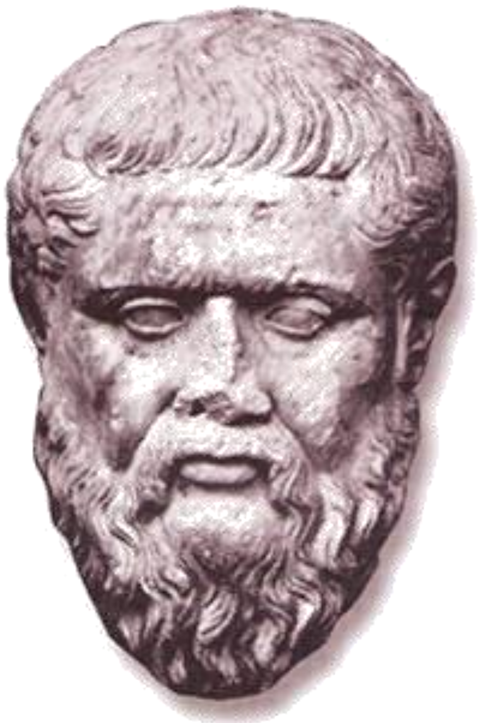
n.

1. Somethi ne or more of the senses, especia 

or touch; a 

2. A focus o g, thought, or action: *an object of co*
3. The purpos l of a specific action or effort: *the object o game.*
4. *Grammar.*
   a. A noun, pronoun, o oun phrase that receives or is affected by the a ion of a ve within a sentence.
   b. A noun or substantive verned by a preposition.
5. *Philosophy.* Something int gible or perceptible by the mind.
6. *Computer Science.* A discrete item that can be selected and maneuvered, such as an onscreen graphic. In object-oriented programming, objects include data and the procedures necessary to operate on that data.

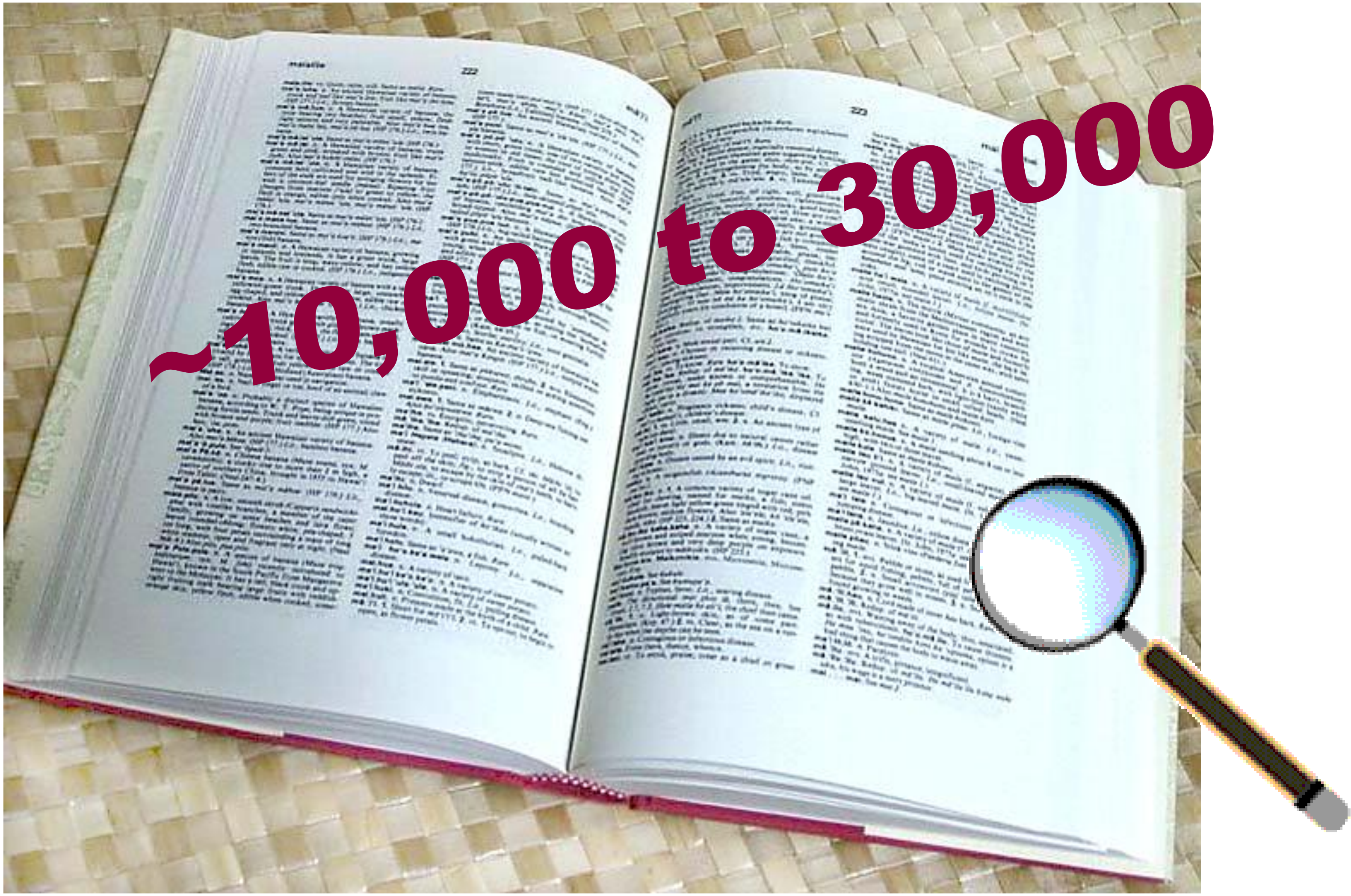
Plato said…



- Ordinary objects are classified together if they `participate' in the same abstract Form, such as the Form of a Human or the Form of Quartz.
- Forms are proper subjects of philosophical investigation, for they have the highest degree of reality.
- Ordinary objects, such as humans, trees, and stones, have a lower degree of reality than the Forms.
- Fictions, shadows, and the like have a still lower degree of reality than ordinary objects and so are not proper subjects of philosophical enquiry.

Bruegel, 1564

How many object categories are there?



Biederman 1987

Why do we care about recognition?

Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

"We therefore include the perception of function as a proper –indeed, crucial- subject for vision science", *from Vision Science, chapter 9, Palmer*.

The perception of function

- Direct perception (affordances): Gibson

Flat surface
Horizontal
Knee-high
… → Sittable upon

- Mediated perception (Categorization)

Flat surface
Horizontal
Knee-high
… → Chair → Sittable upon

Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)

It does not seem easy
to sit-upon this…

Direct perception

Some aspects of an object function can be perceived directly

- Observer relativity: Function is observer dependent

From http://lastchancerescueflint.org

Limitations of Direct Perception

Objects of similar structure might have very different functions

Figure 9.1.2 Objects with similar structure but different functions. Mailboxes afford letter mailing, whereas trash cans do not, even though they have many similar physical features, such as size, location, and presence of an opening large enough to insert letters and medium-sized packages.

Not all functions seem to be available from direct visual information only.

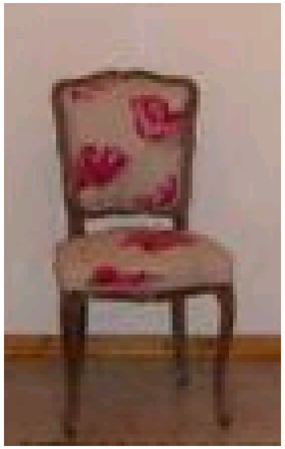
The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…

How do we achieve Mediated perception?

Well… this requires object recognition (for more details, see entire course)

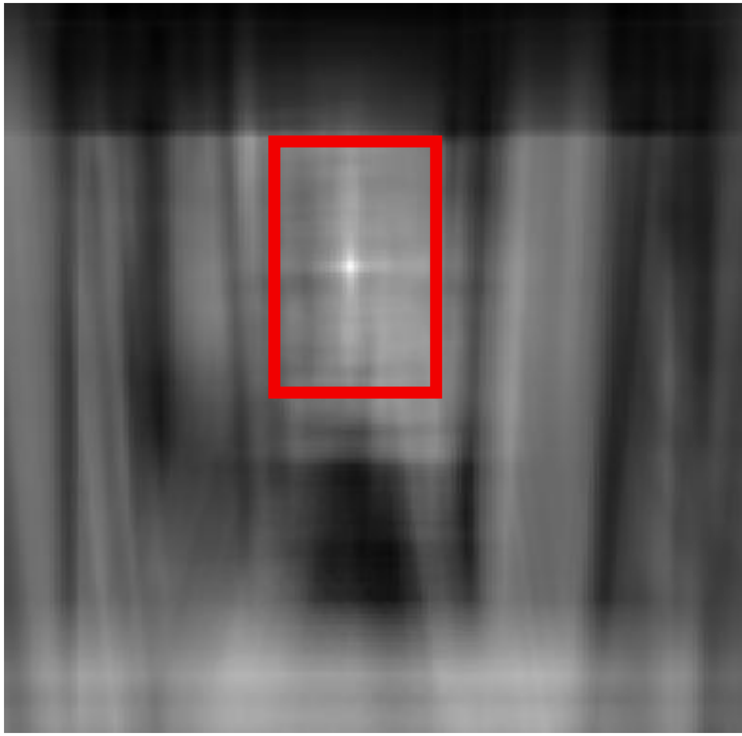
Object recognition
Is it really so hard?

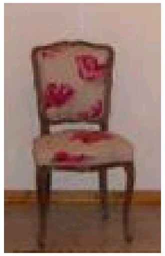
This is a chair



Find the chair in this image

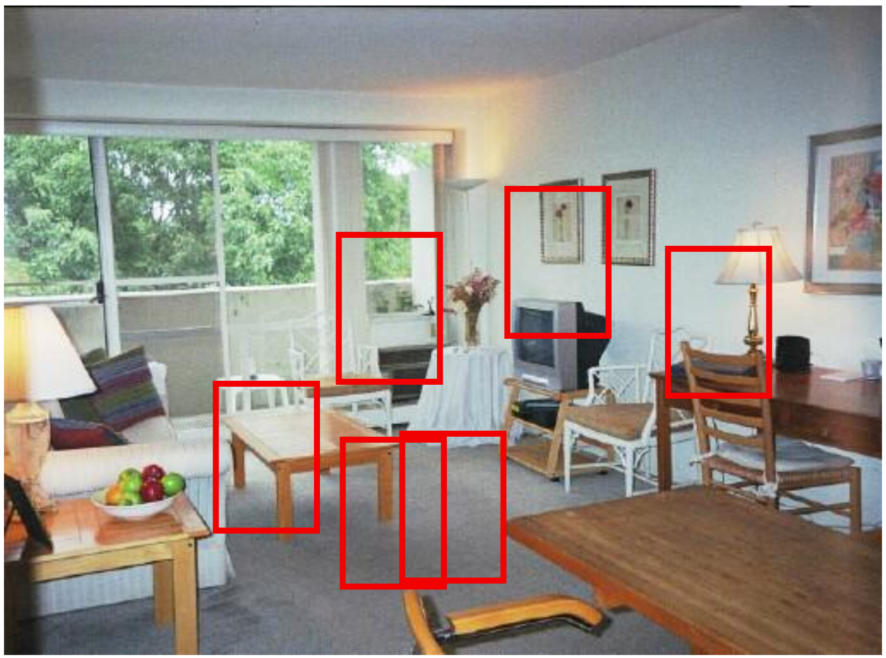
Output of normalized correlation

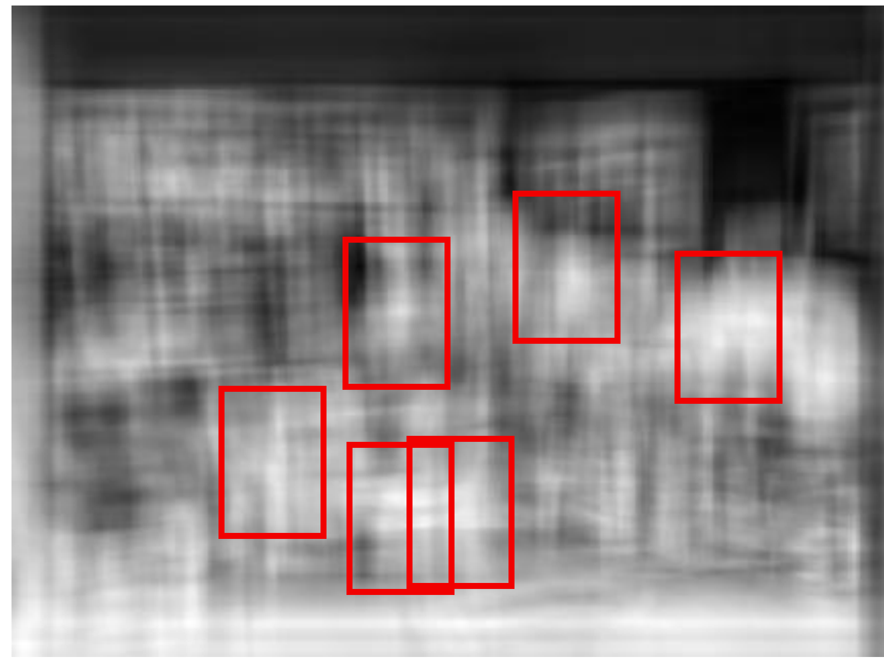




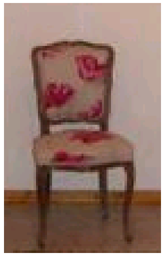
Object recognition
Is it really so hard?

Find the chair in this image





Pretty much garbage
Simple template matching is not going to make it



Object recognition
Is it really so hard?

Find the chair in this image

A “popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts.” Nivatia & Binford, 1977.

And it can get a lot harder

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

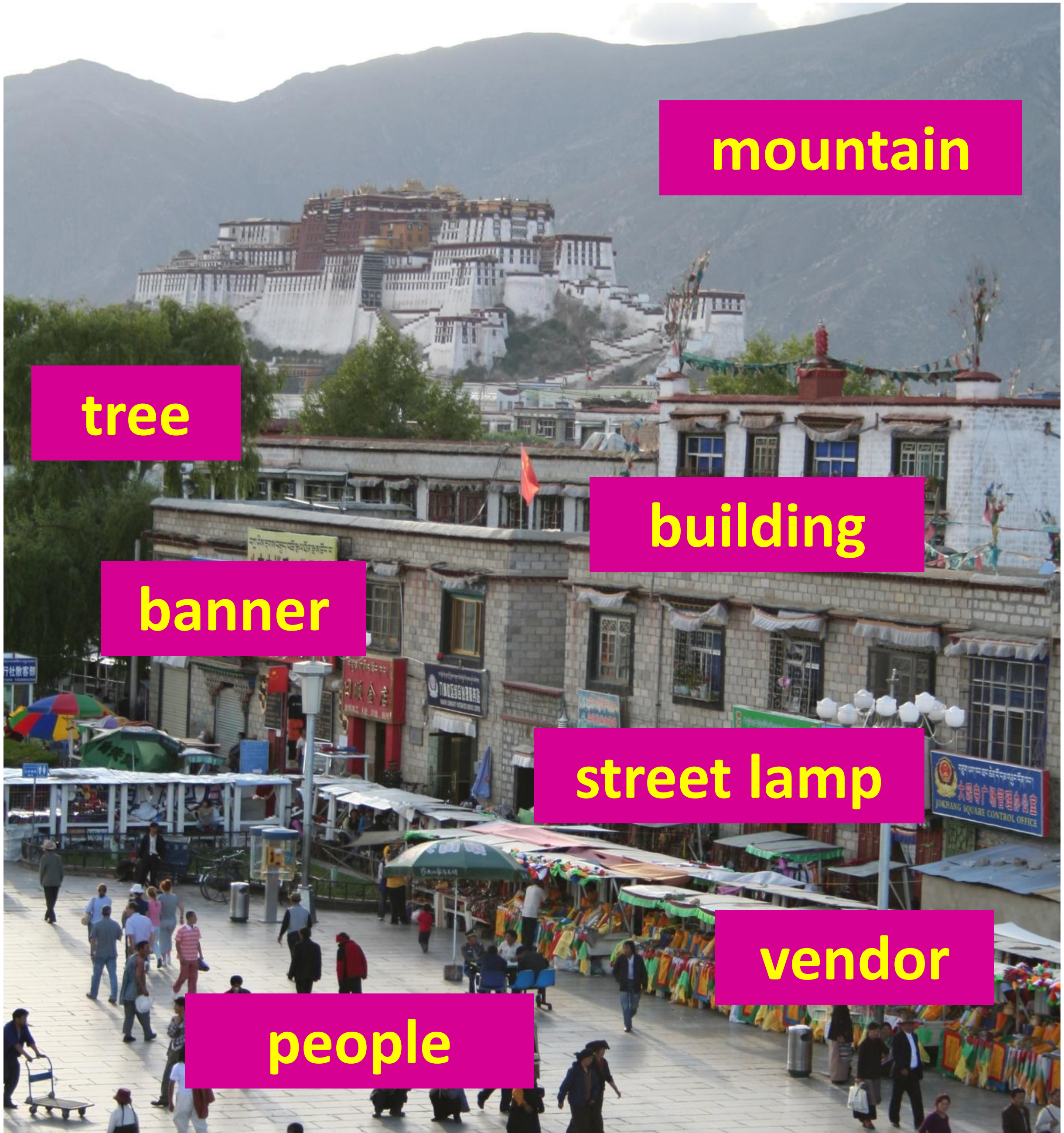
So what does object recognition involve?

Verification: is that a lamp?

Detection: are there people?

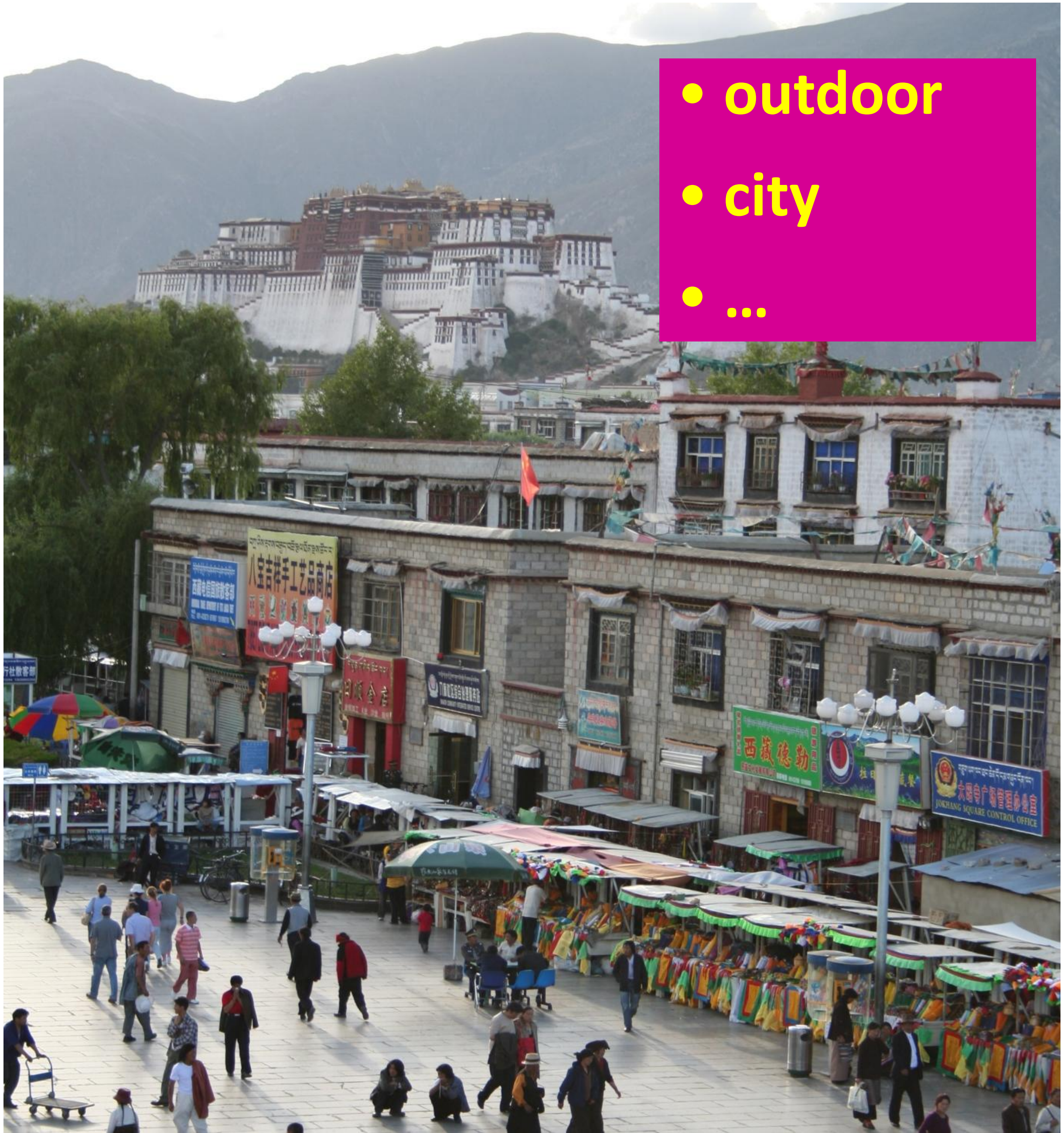
Identification: is that Potala Palace?

Object categorization



Scene and context categorization

- **outdoor**
- **city**
- **…**

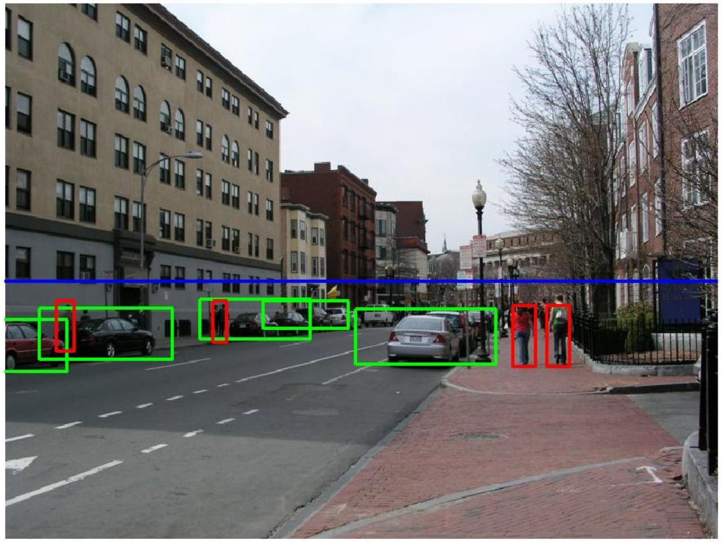


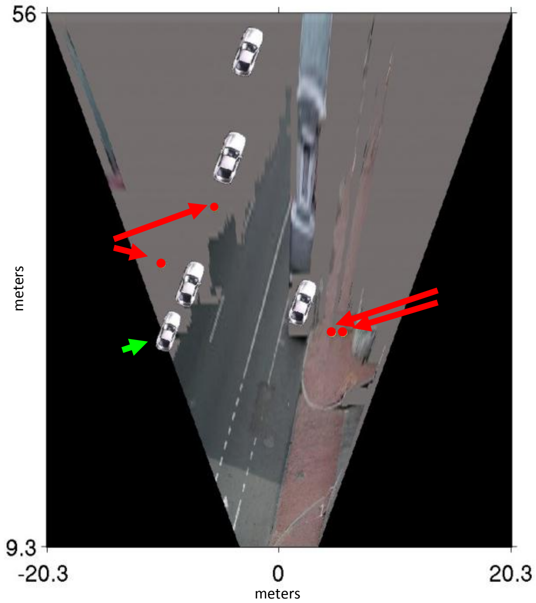
Computational photography

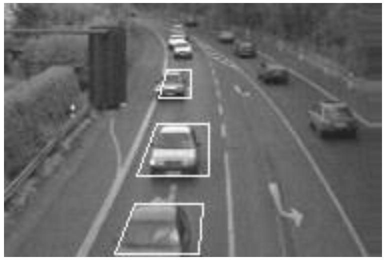
[Face priority AE] When a bright part of the face is too bright

Assisted driving

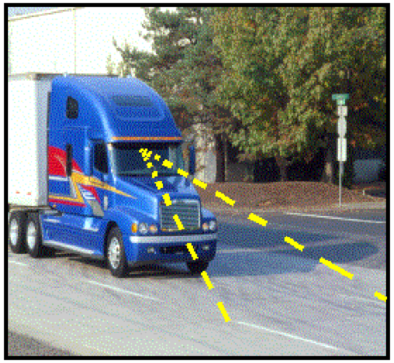
Pedestrian and car detection







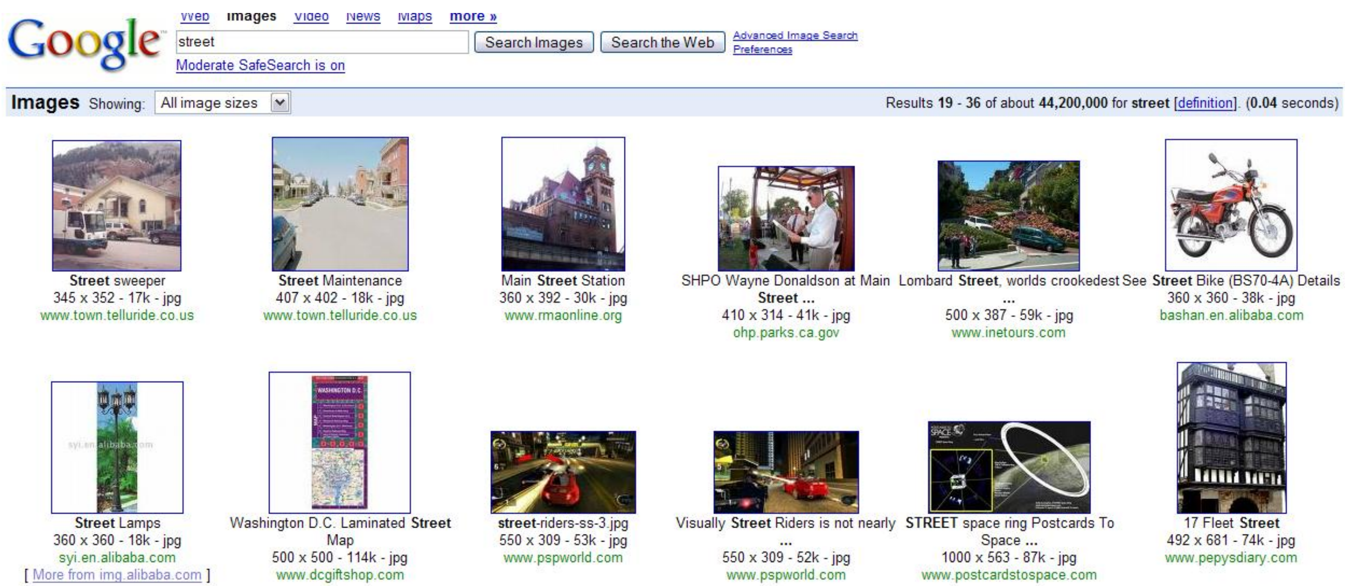
Lane detection



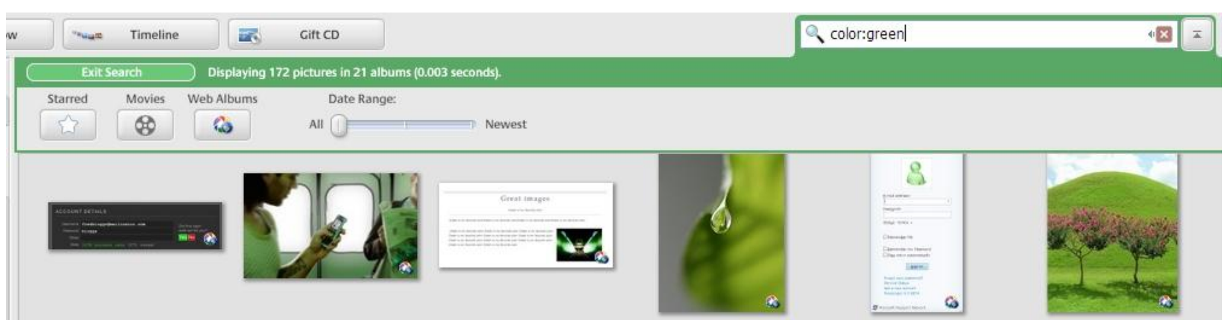
- Collision warning systems with adaptive cruise control,
- Lane departure warning systems,
- Rear object detection systems,

Improving online search

Query:
STREET



Organizing photo collections



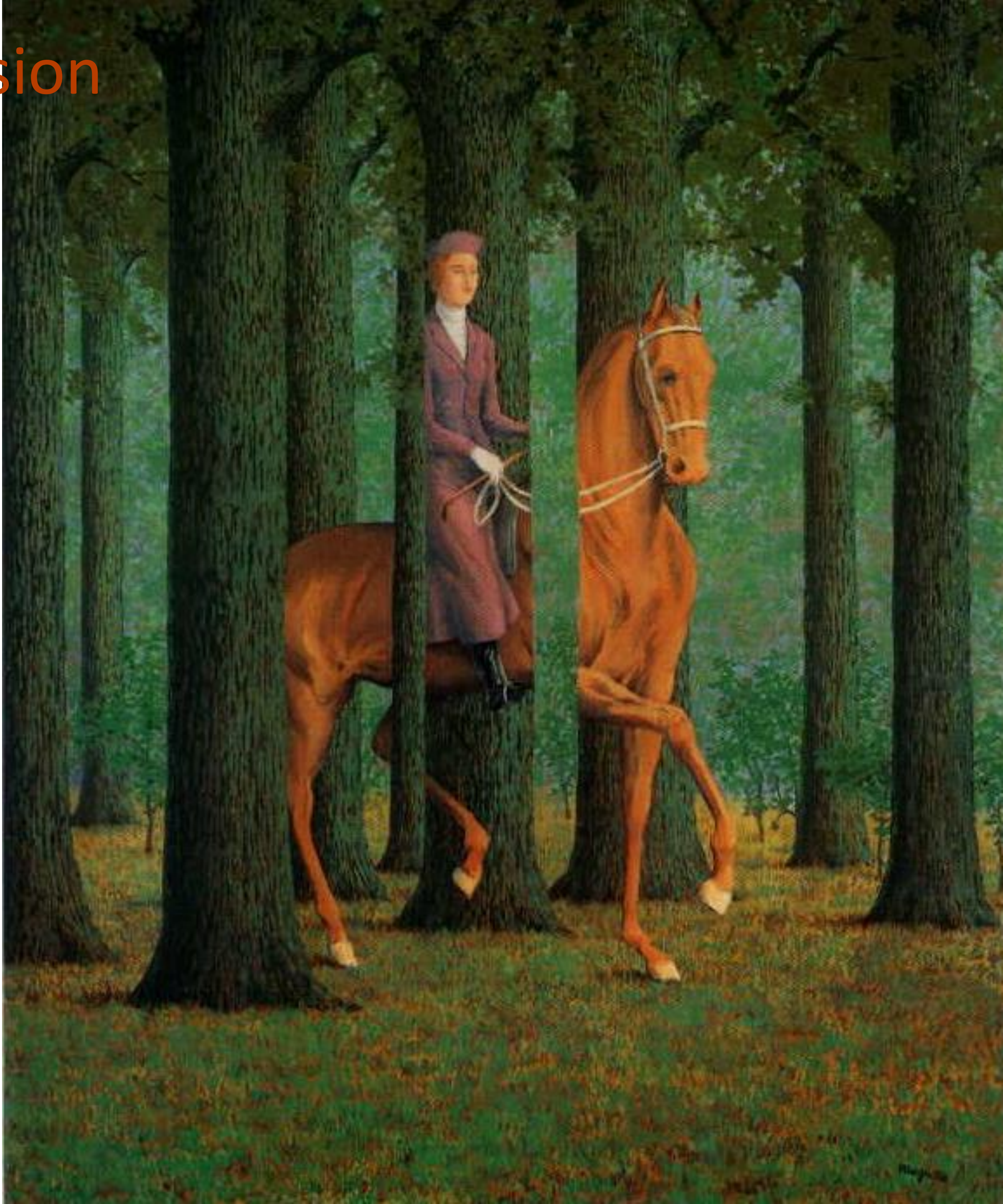
Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

slide credit: S. Ullman

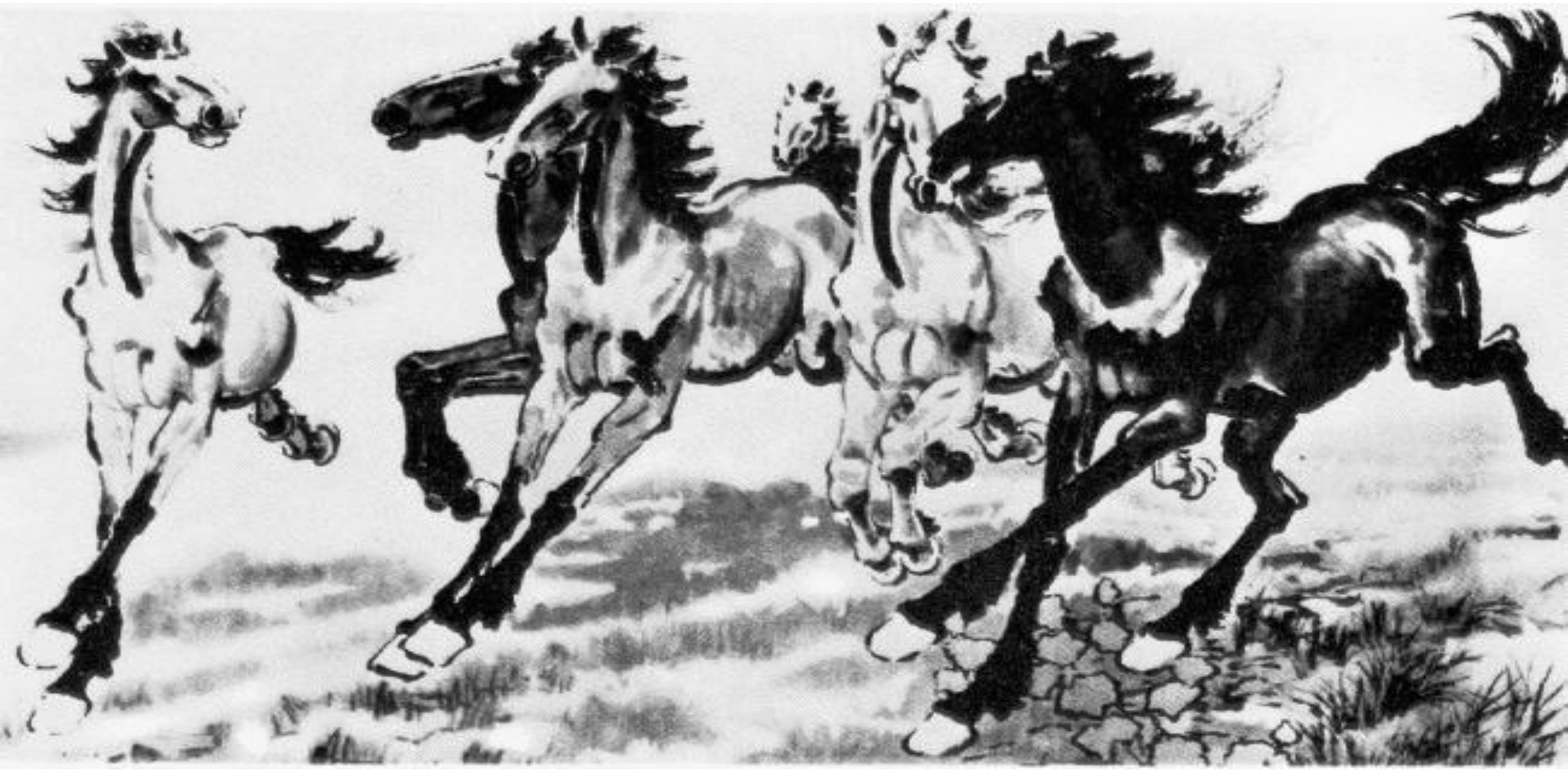
Challenges 3: occlusion



Magritte, 1957

Challenges 4: scale

Challenges 5: deformation

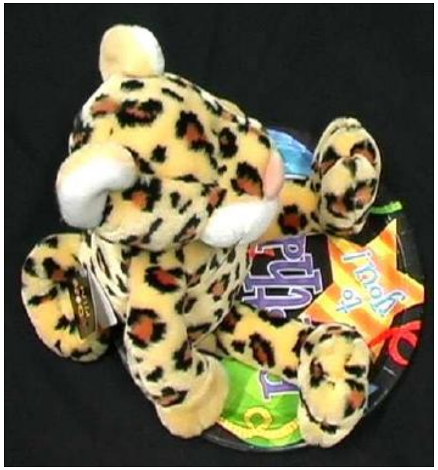


Xu, Beihong 1943

Challenges 6: background clutter

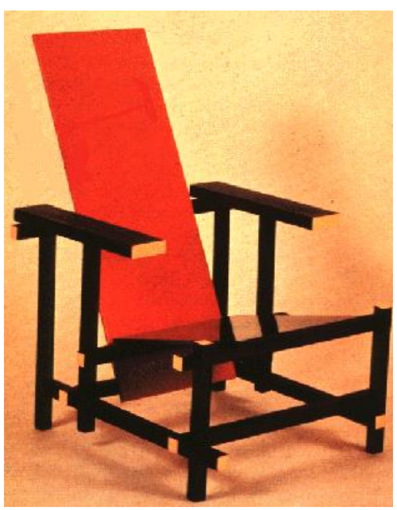
Klimt, 1913


Kellogg's
Crunchy
Nut
CORN FLAKES
Das Beste
jeden Morgen
Le meilleur
chaque matin
375g




Golden
Grahams
Crispix
Trix
French
Toast
Crunch
Cinnamon
Toast
Crunch
Kix

Challenges 7: intra-class variation




~10,000 to 30,000

Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

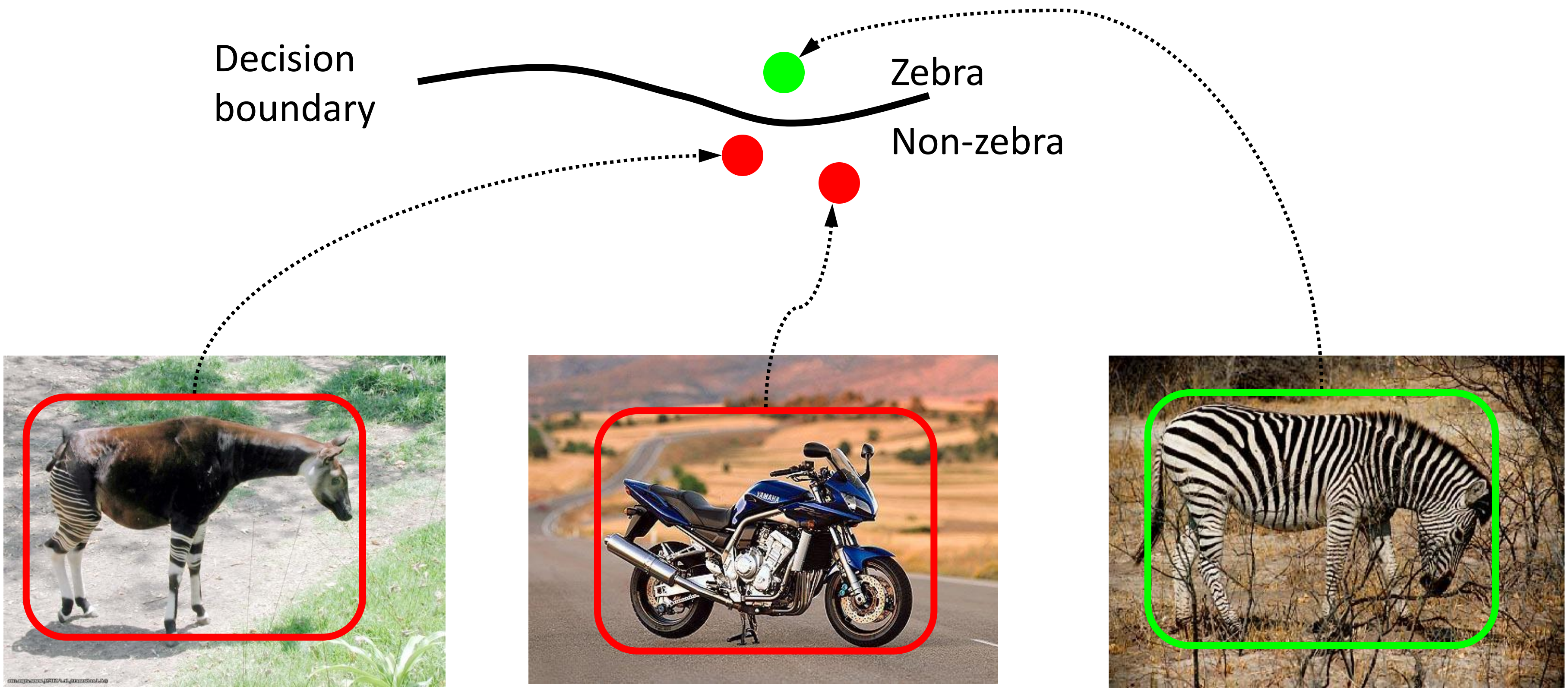
Object categorization: the statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

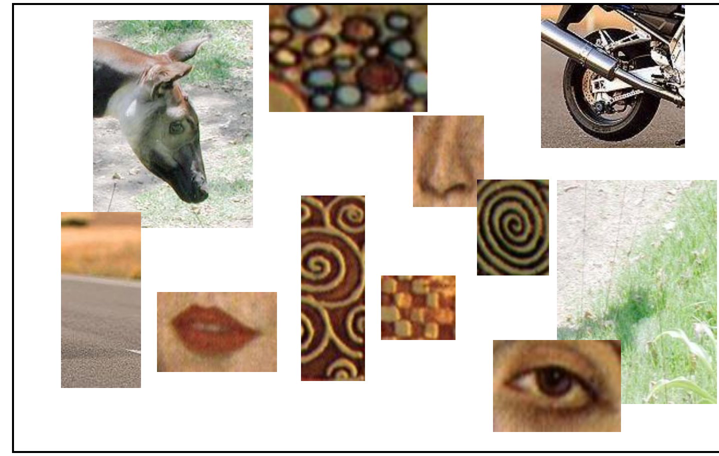
Discriminative

- Direct modeling of $\dfrac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



Generative

- Model $p(image|\, zebra)$ and $p(image|\, no\, zebra)$



	$p(image|\, zebra)$	$p(image|\, no\, zebra)$
	Low	Middle
	High	Middle→Low

Three main issues

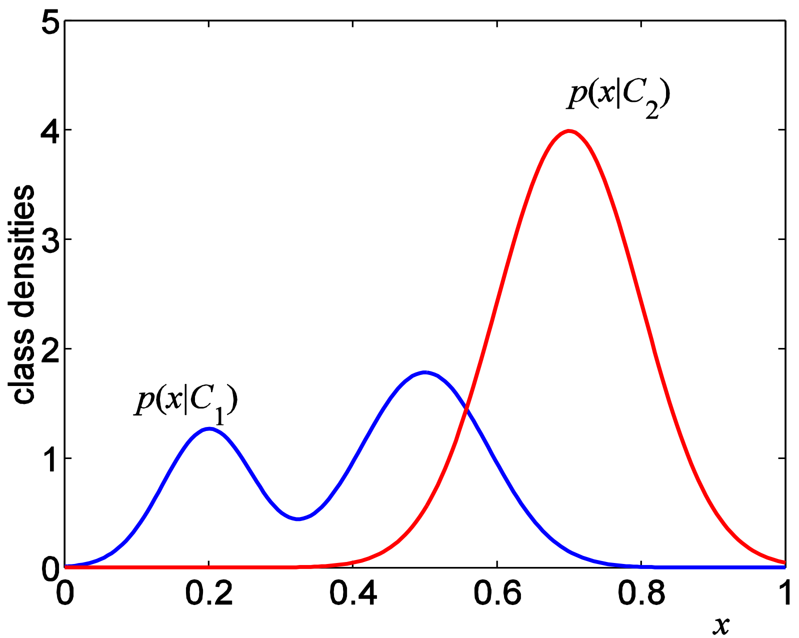
- Representation
  - How to represent an object category
- Learning
  - How to form the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

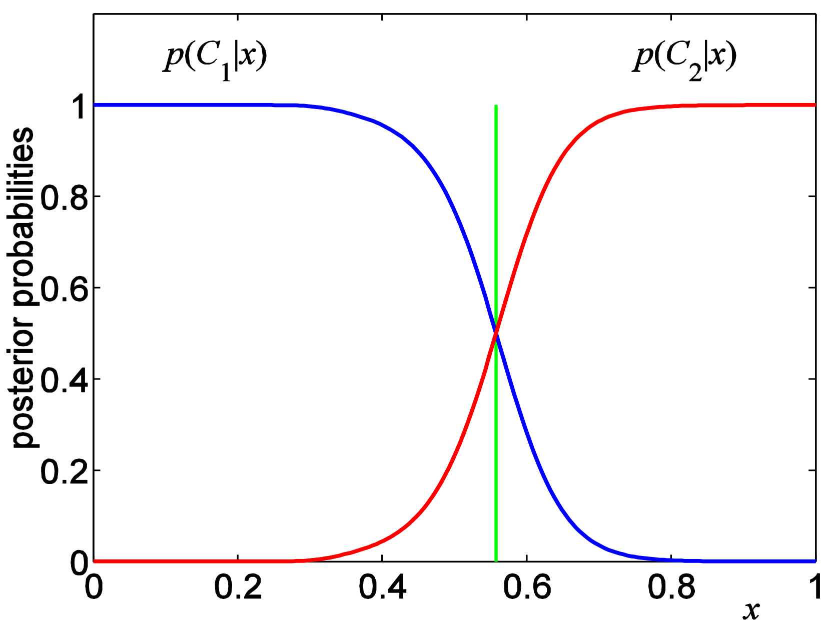
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
– Methods of training: generative vs. discriminative

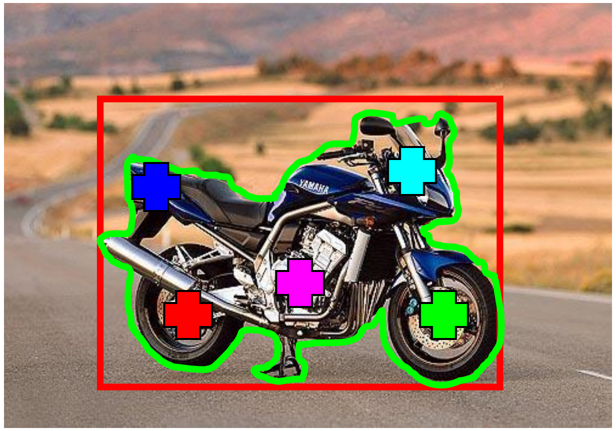




Learning

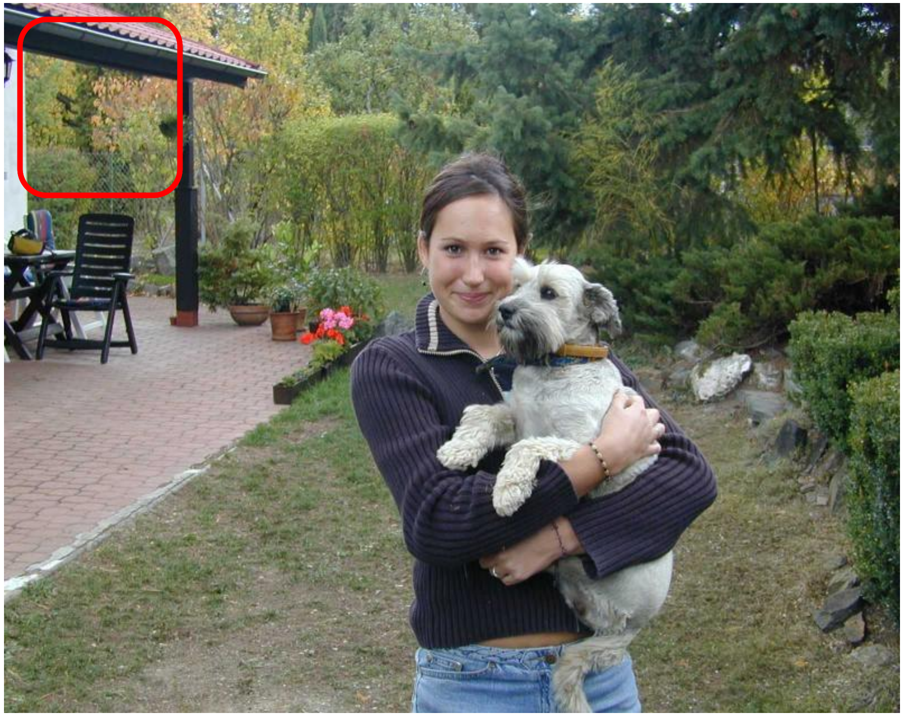
– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
– Level of supervision
  • Manual segmentation; bounding box; image labels; noisy labels

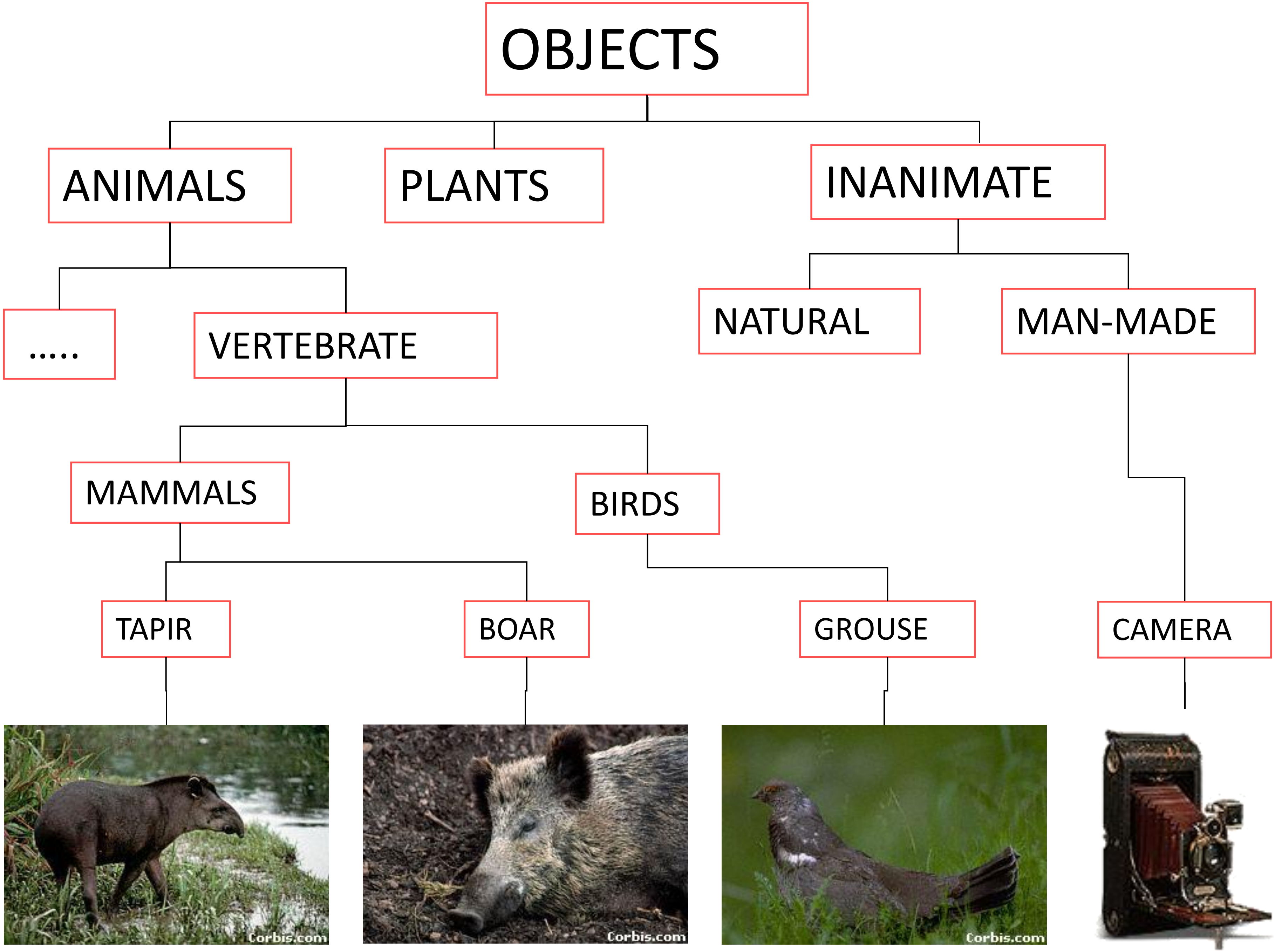
Contains a motorbike



Recognition

- Scale / orientation range to search over
- Speed
- Context




OBJECTS
ANIMALS
PLANTS
INANIMATE
.....
VERTEBRATE
NATURAL
MAN-MADE
MAMMALS
BIRDS
TAPIR
BOAR
GROUSE
CAMERA
Corbis.com
Corbis.com
Corbis.com

Classical Methods

1. Bag of words approaches
2. Parts and structure approaches
3. Discriminative
   methods

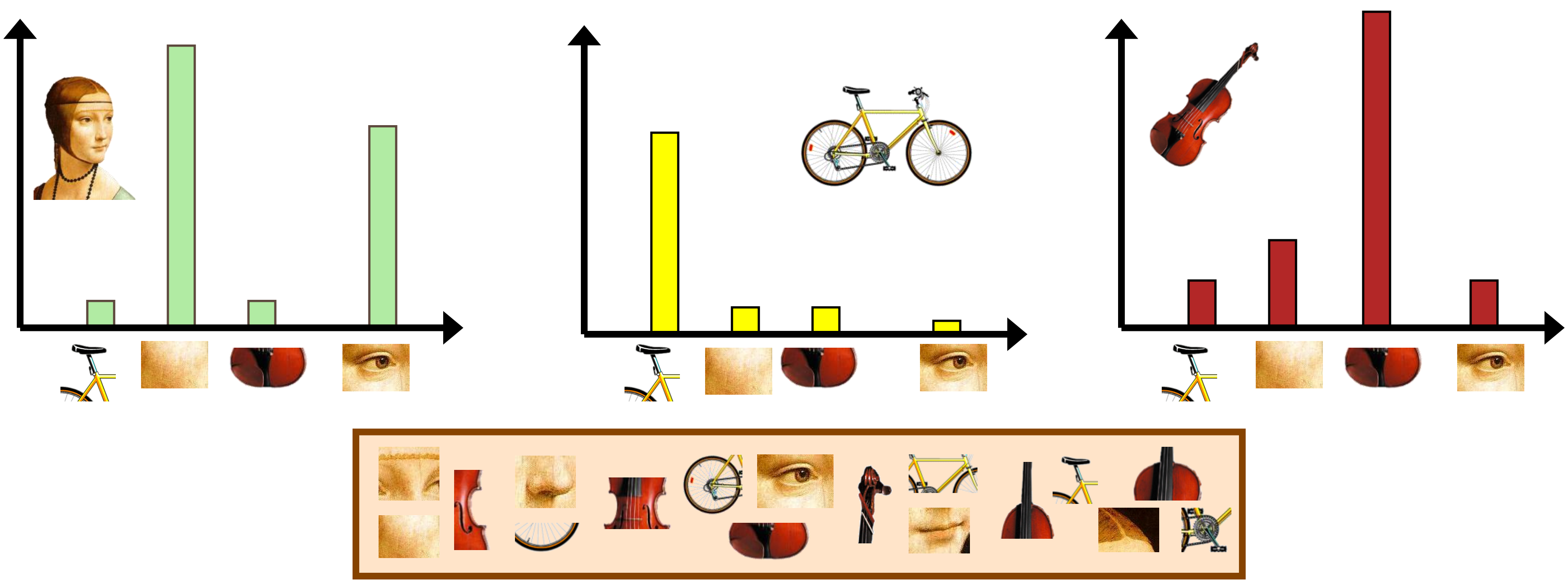
Bag of Words Models

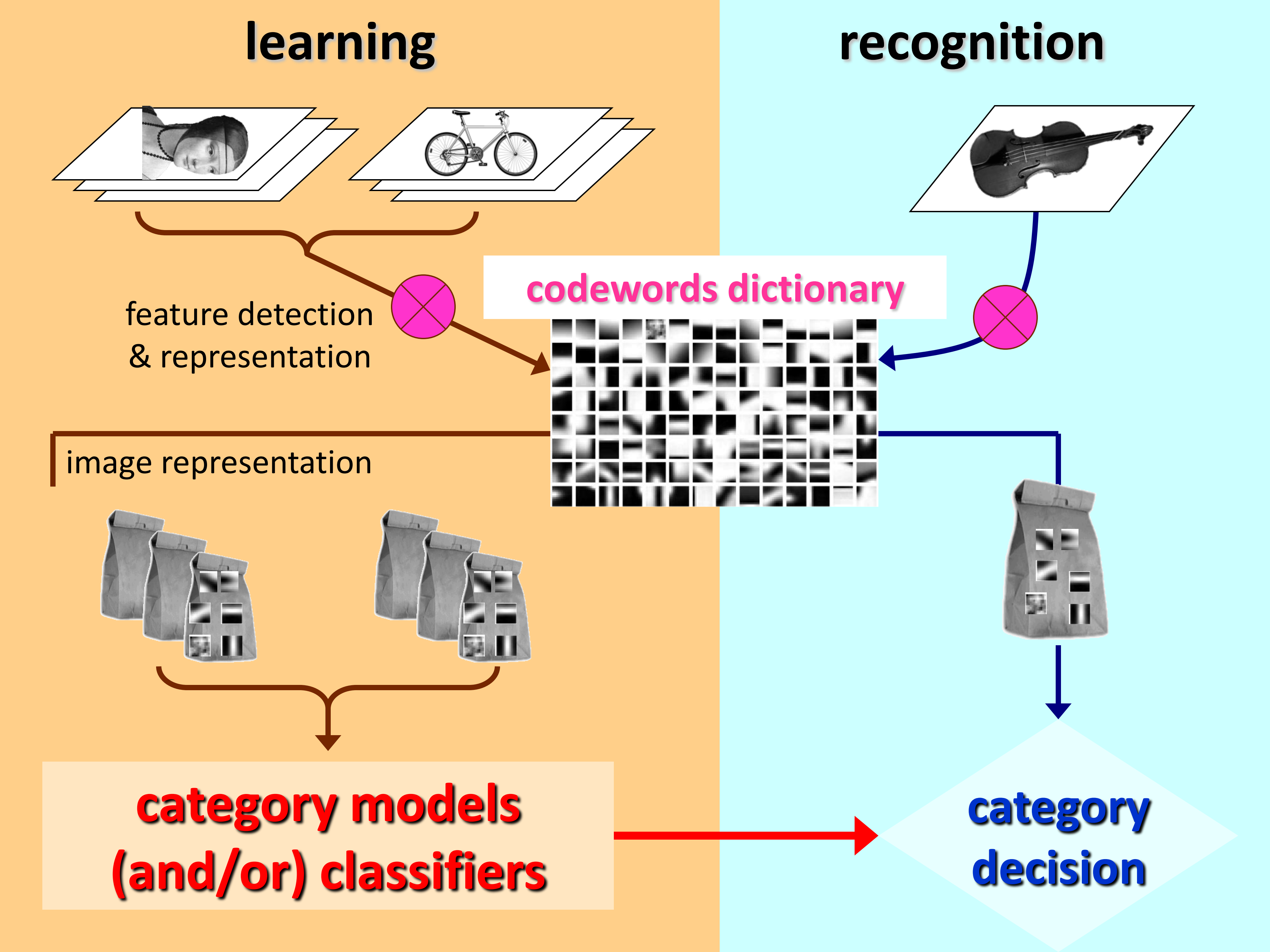
Object → **Bag of 'words'**

Bag of Words

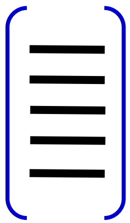
- Independent features

- Histogram representation



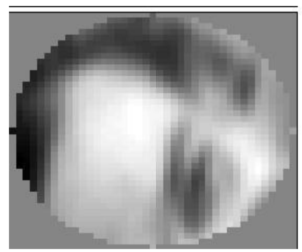

learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

1.Feature detection and representation

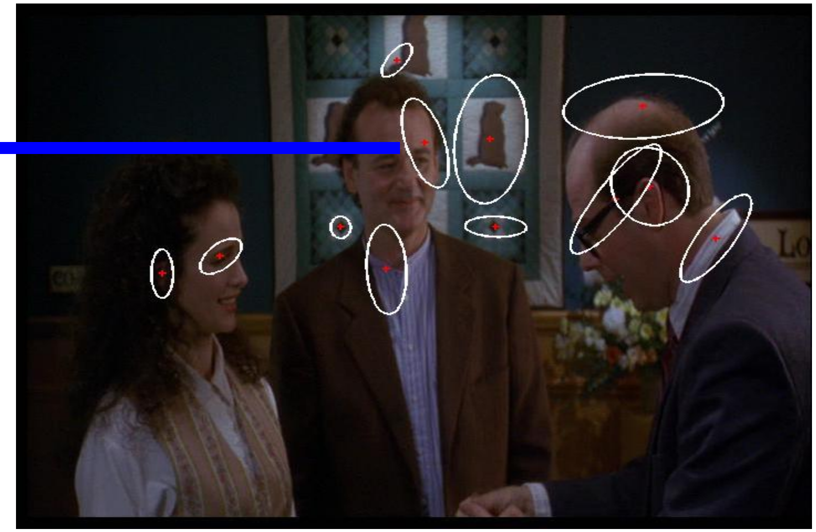


Compute descriptor

e.g. **SIFT** [Lowe'99]



Normalize patch



Detect patches

[Mikojaczyk and Schmid '02]

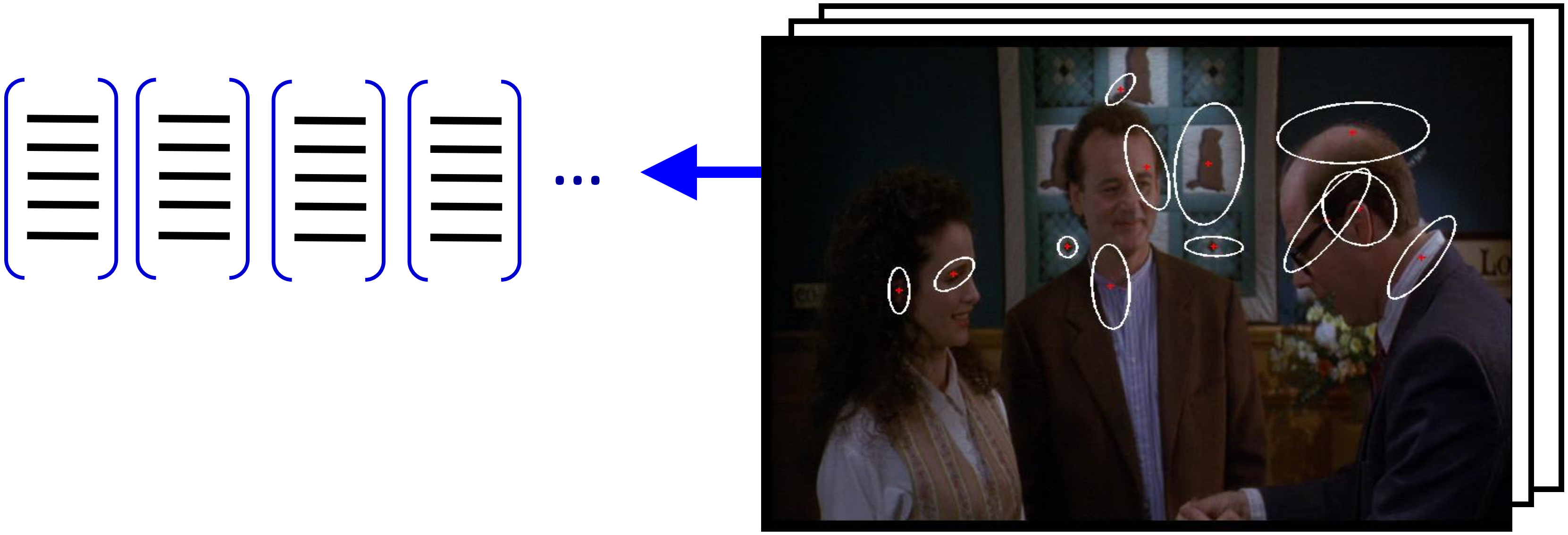
[Mata, Chum, Urban & Pajdla, '02]

[Sivic & Zisserman, '03]

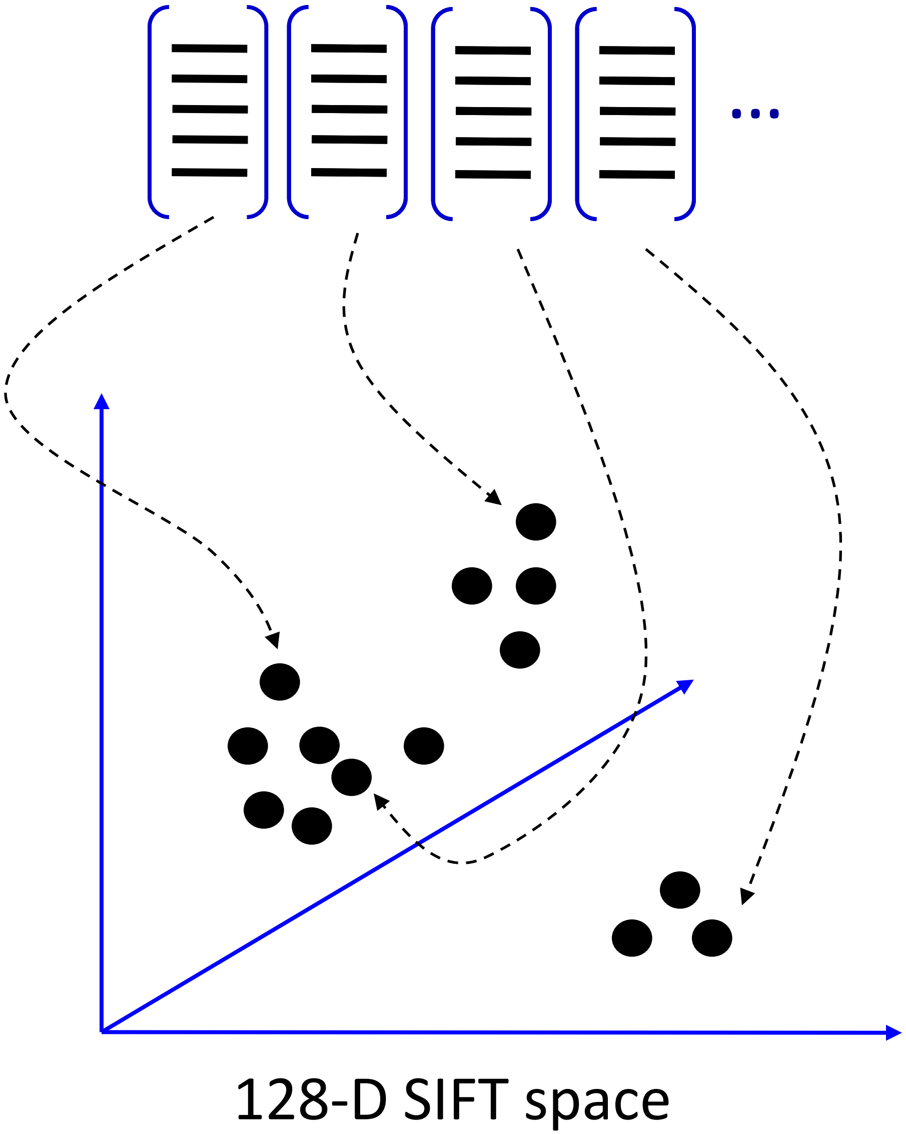
Local interest operator
or
Regular grid

Slide credit: Josef Sivic

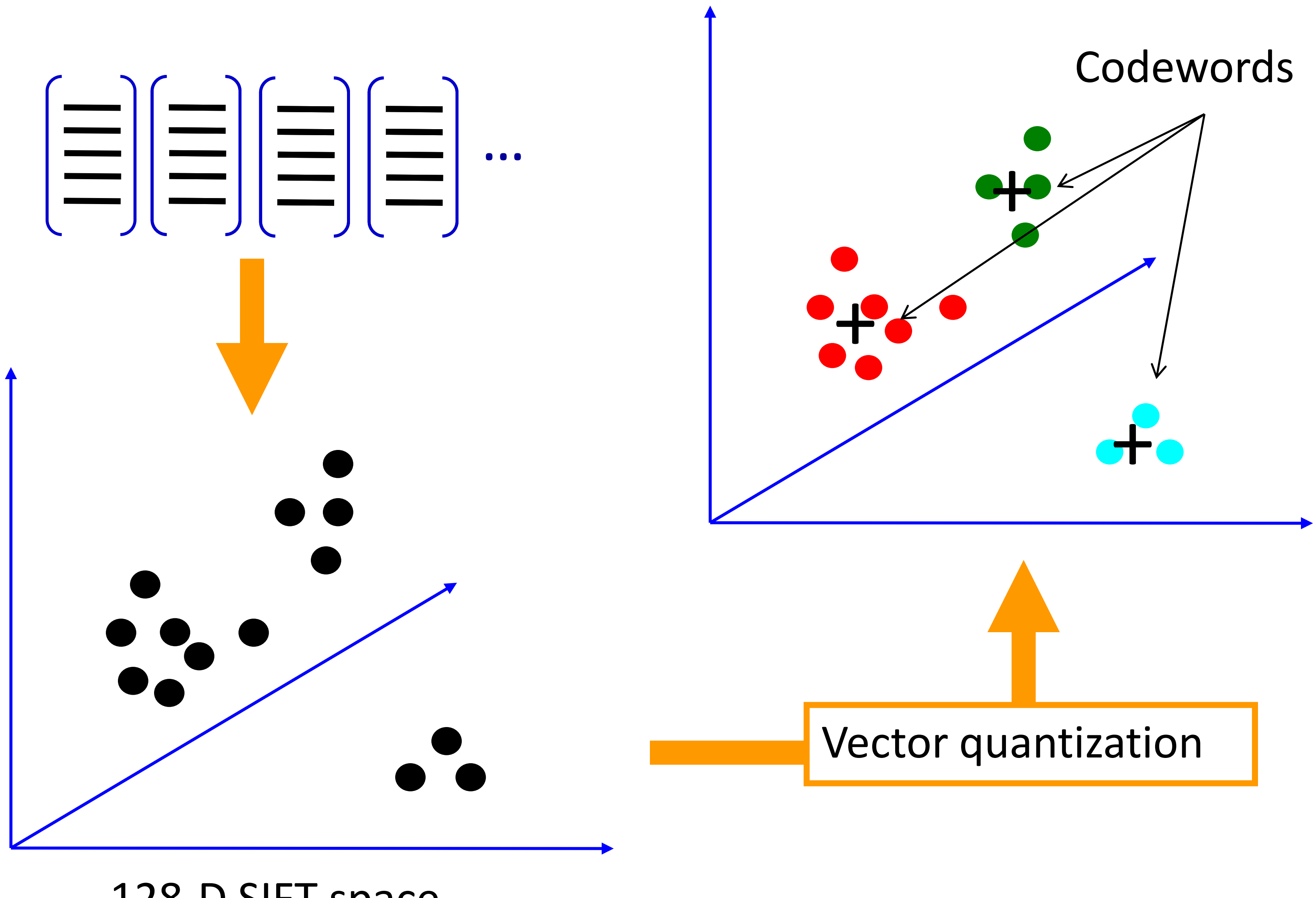
1.Feature detection and representation



2. Codewords dictionary formation

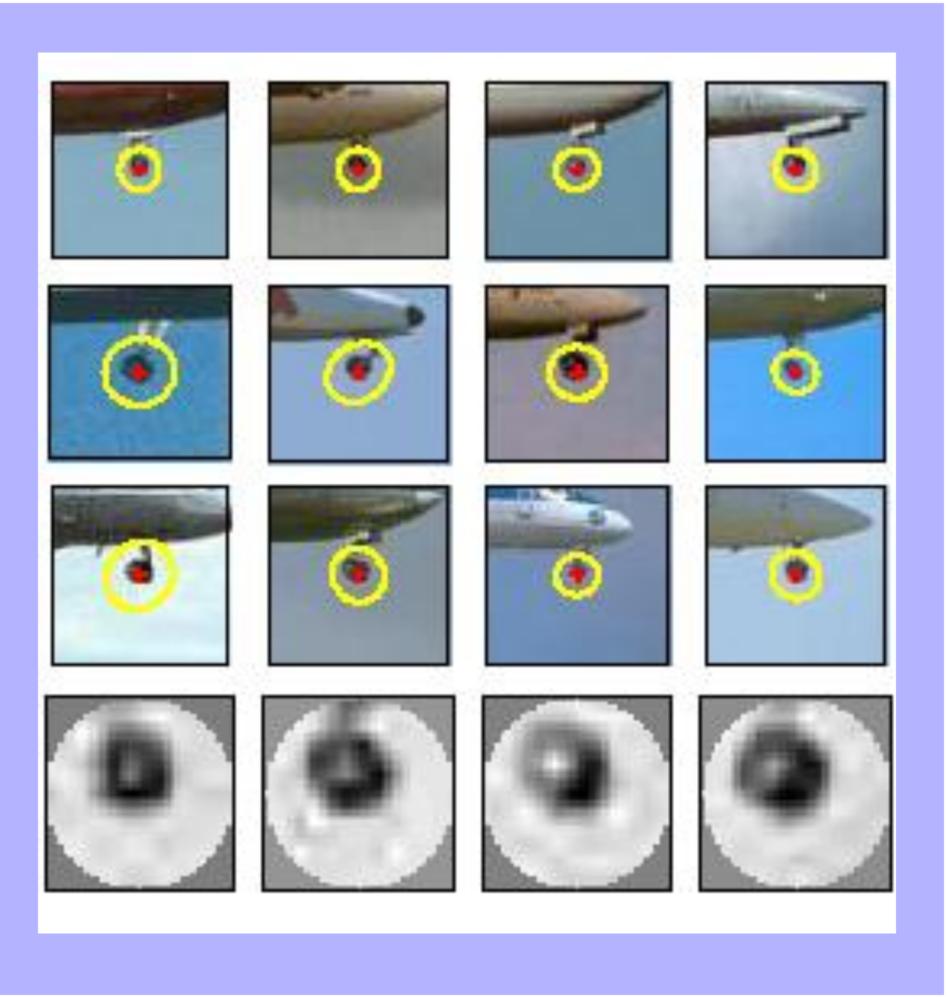


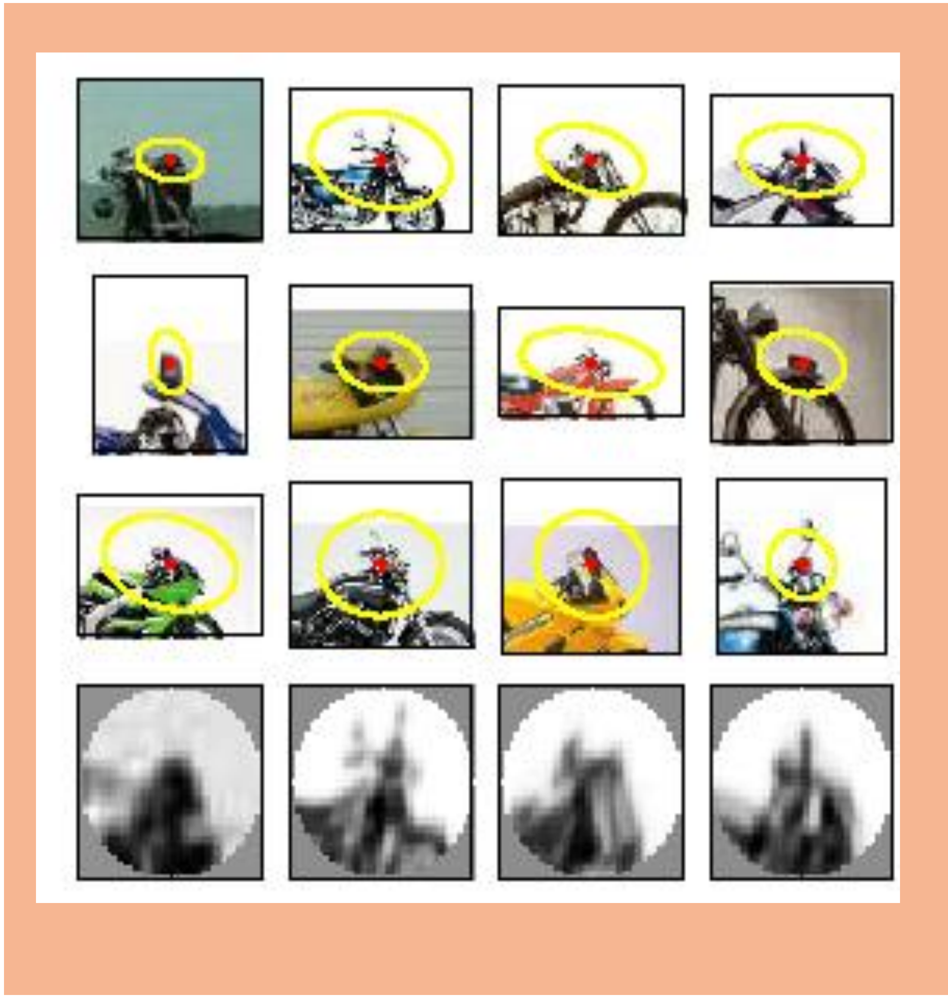
2. Codewords dictionary formation



Slide credit: Josef Sivic

Image patch examples of codewords

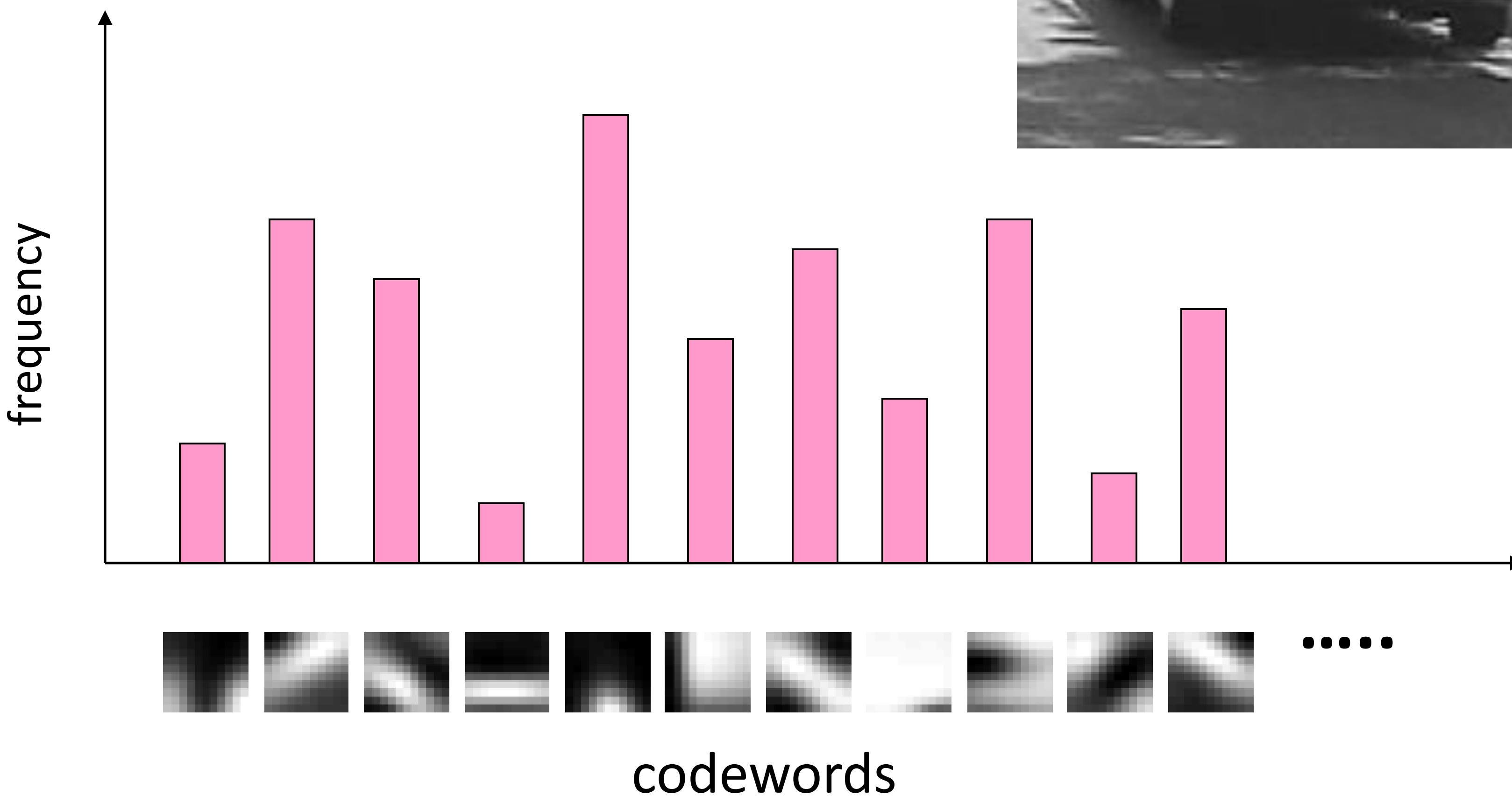




Sivic et al. 2005

Image representation

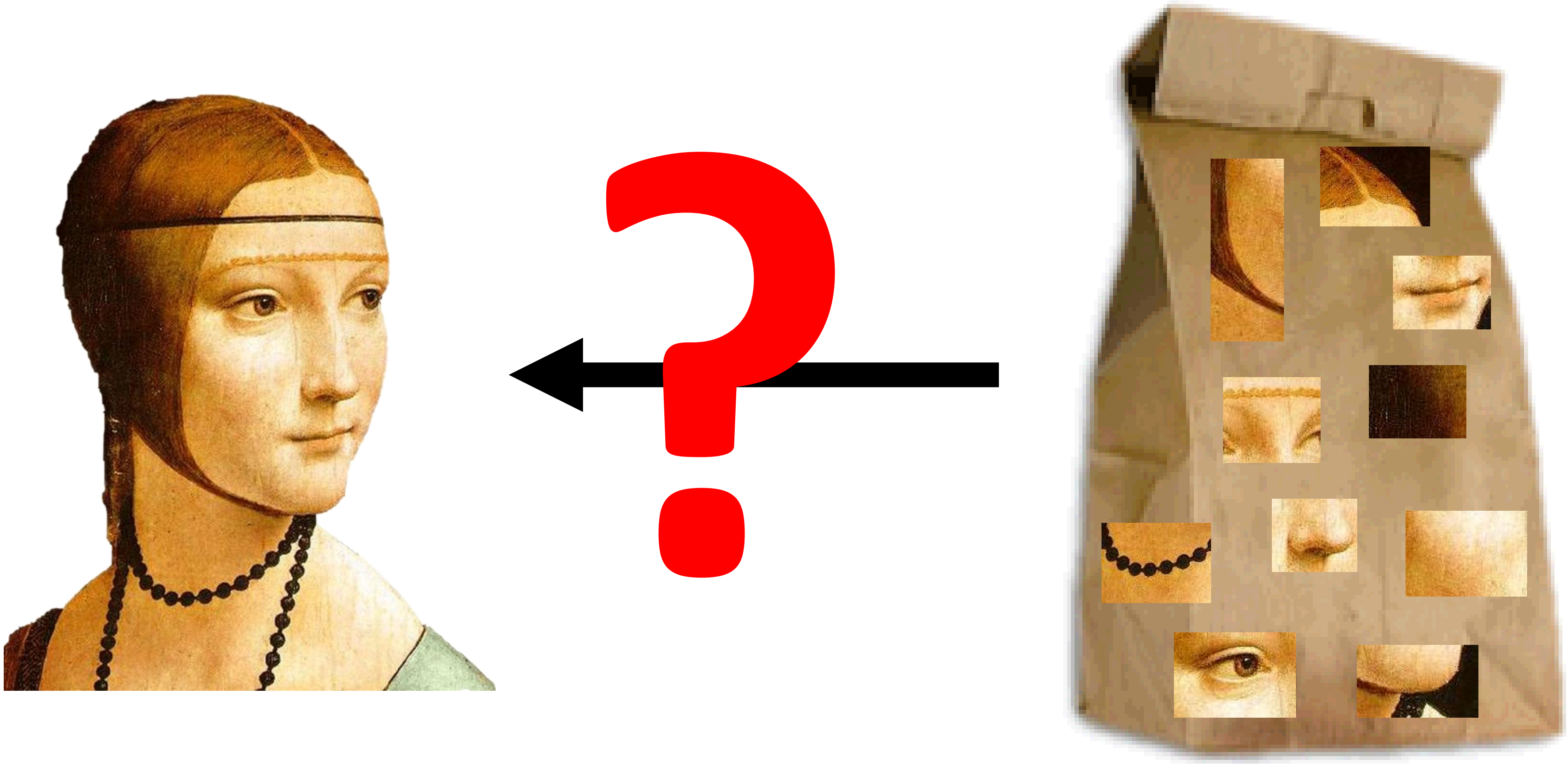
Histogram of features
assigned to each cluster



Uses of BoW representation

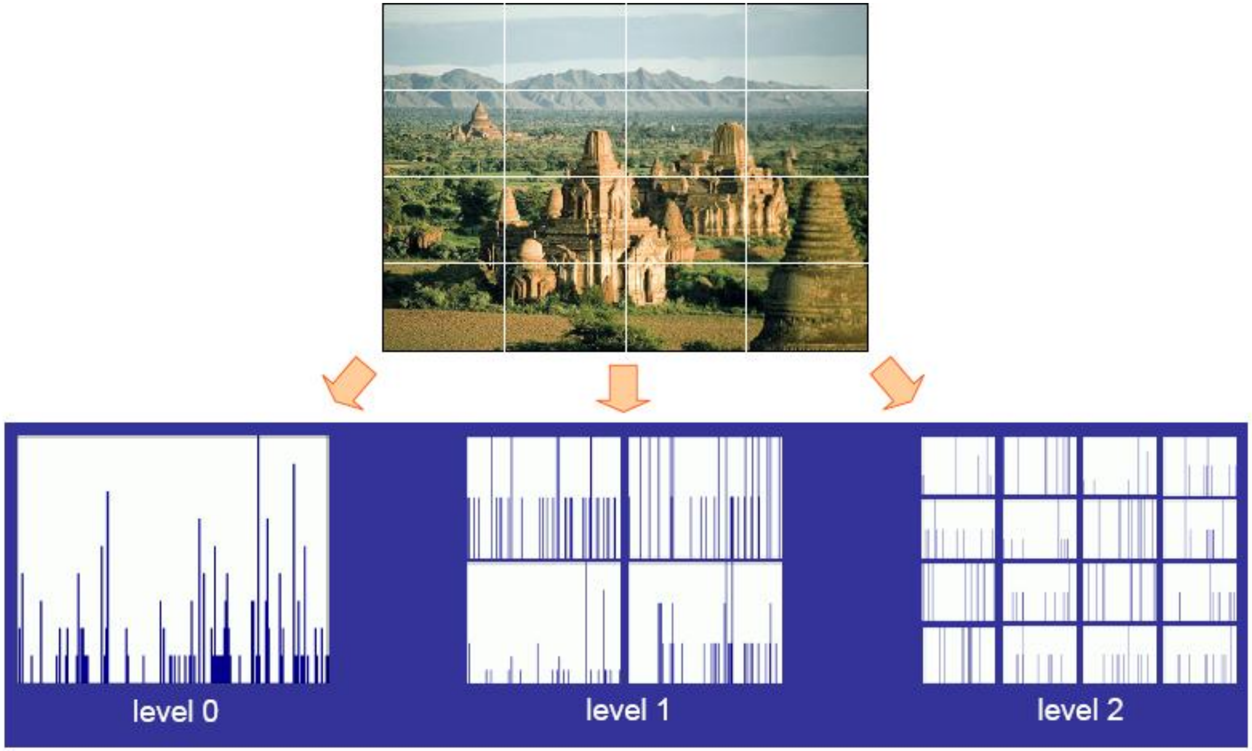
- Treat as feature vector for standard classifier
  - e.g SVM
- Cluster BoW vectors over image collection
  - Discover visual themes
- Hierarchical models
  - Decompose scene/object
- Scene

What about spatial info?

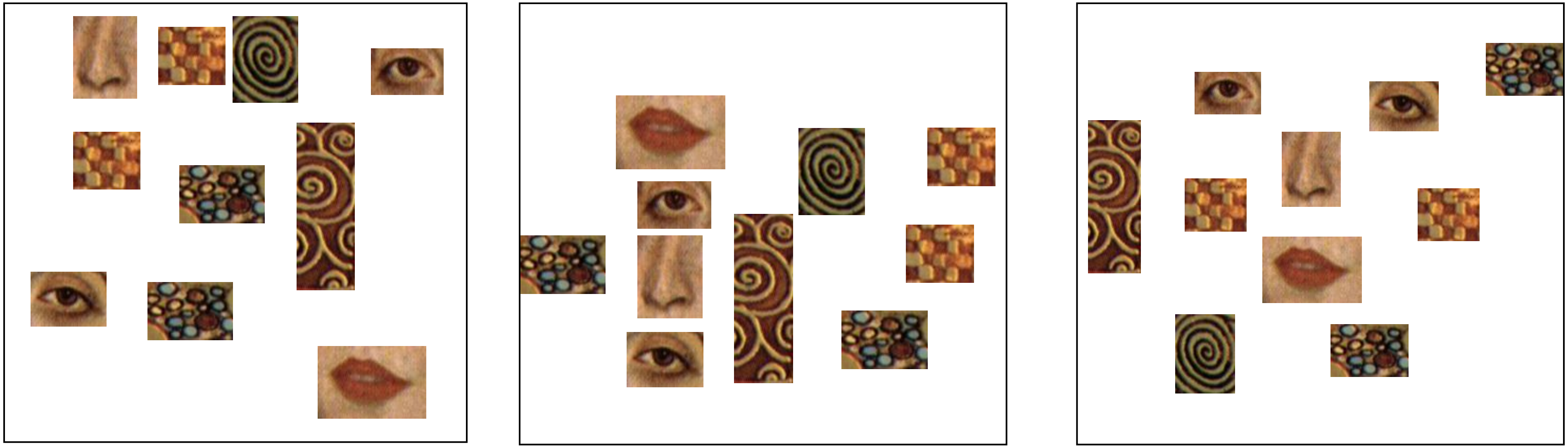


Adding spatial info. to BoW

- Feature level
- Generative models
- Discriminative methods
  - Lazebnik, Schmid & Ponce, 2006



Problem with bag-of-words



- All have equal probability for bag-of-words methods
- Location information is important
- BoW + location still doesn't give correspondence

Model: Parts and Structure

Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - How to handle occlusion/clutter

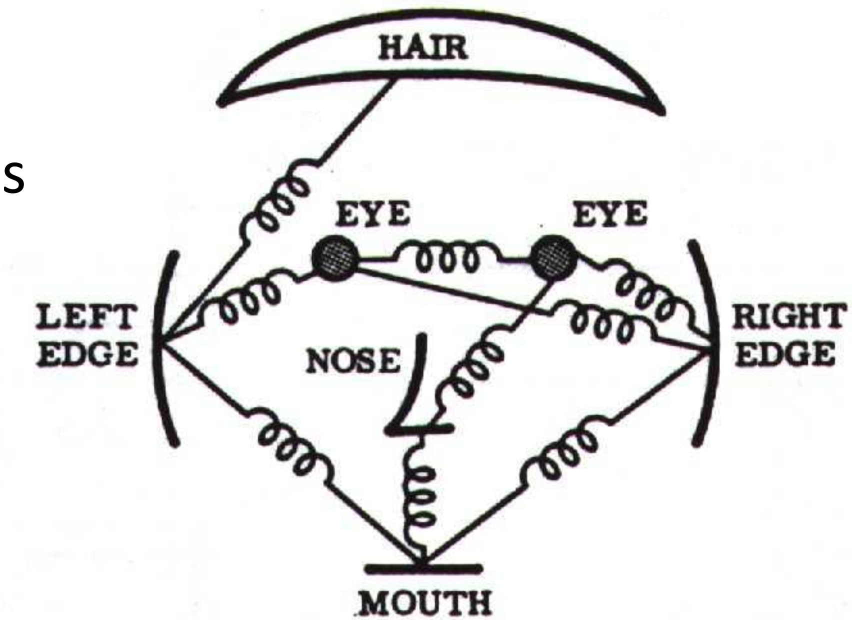
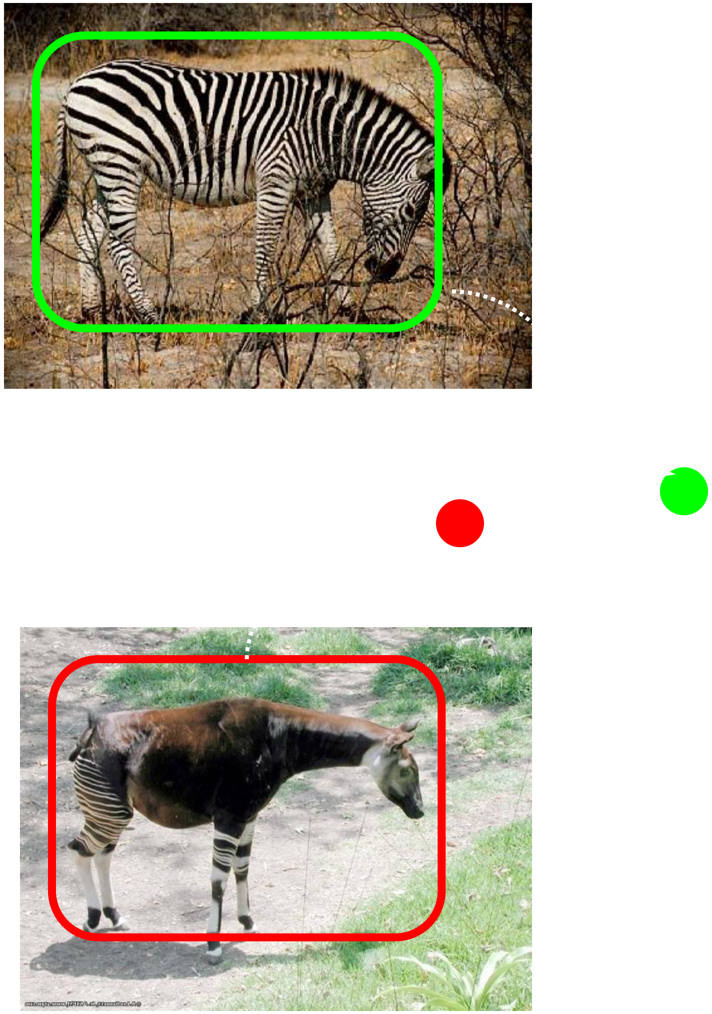


Figure from [Fischler & Elschlager 73]



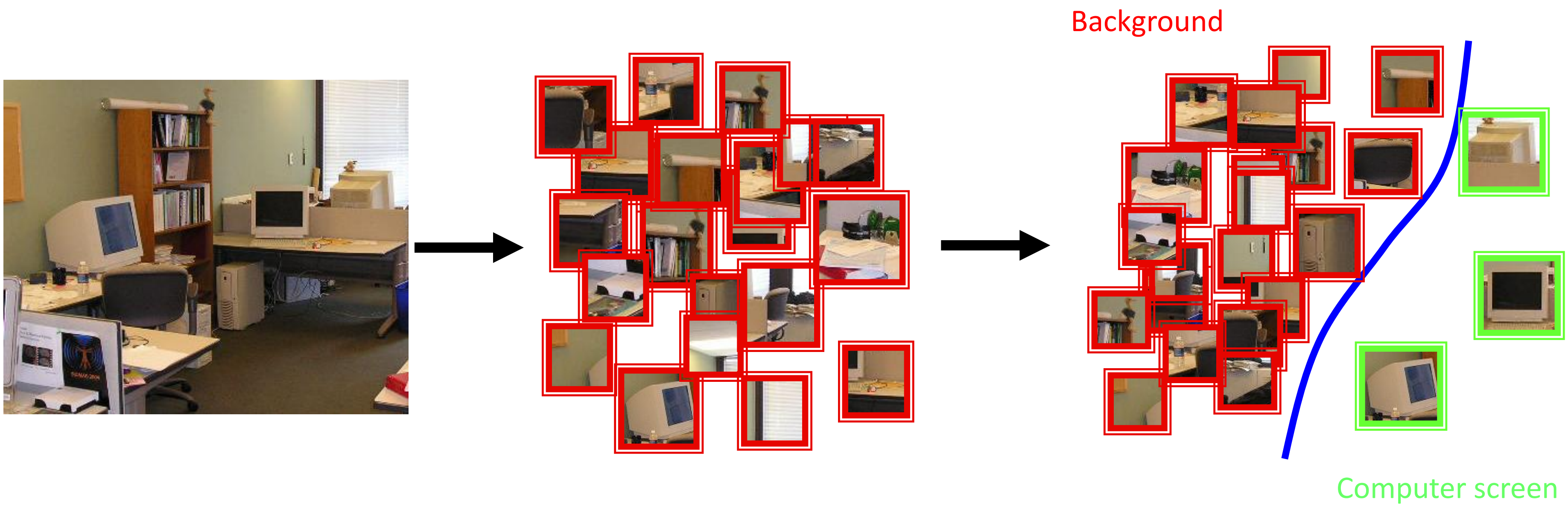
Classifier-based methods

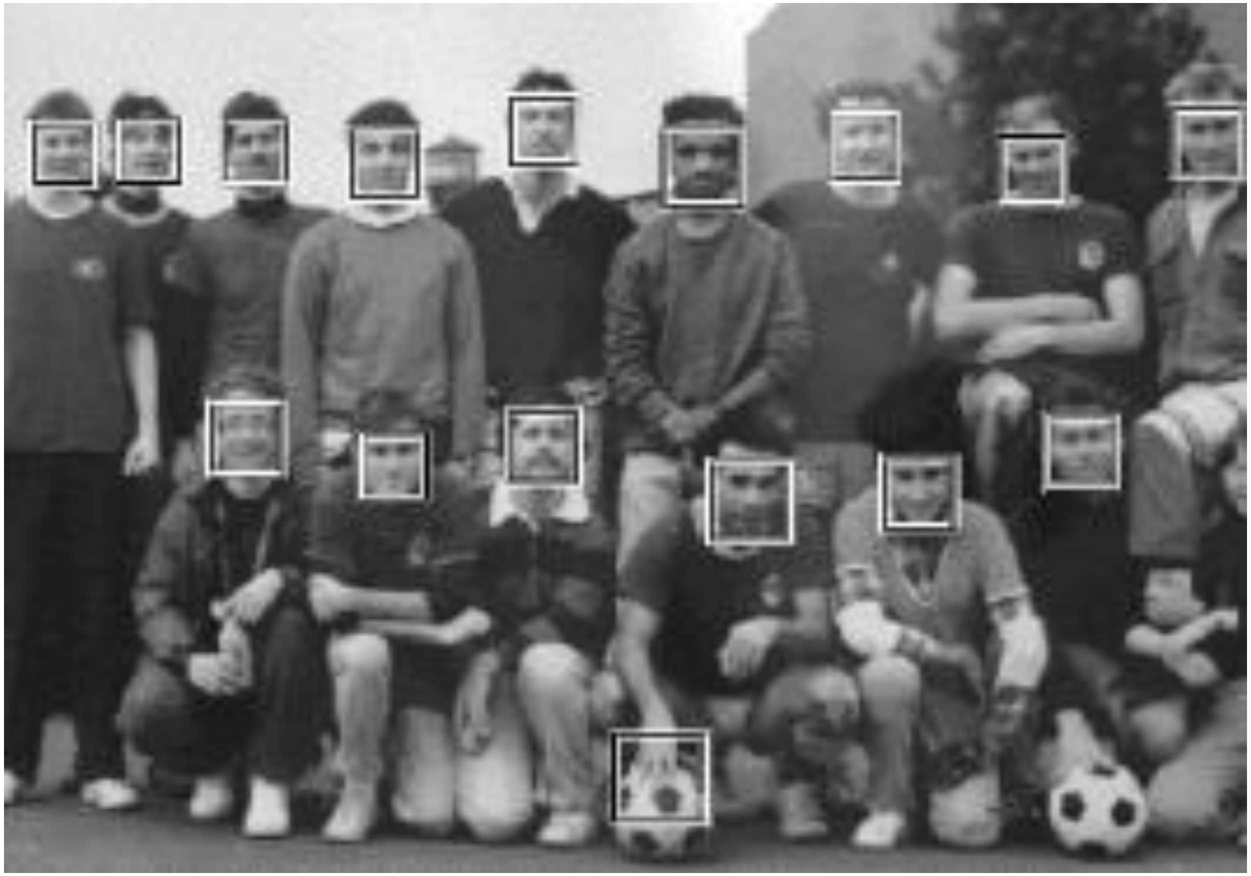
Classifier based methods

Object detection and recognition is formulated as a classification problem.

The image is partitioned into a set of overlapping windows

… and a decision is taken at each window about if it contains a target object or not.





Context for single object classes

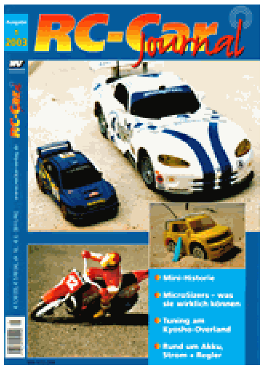
Who needs context anyway?

We can recognize objects even out of context

Banksy

Why is context important?

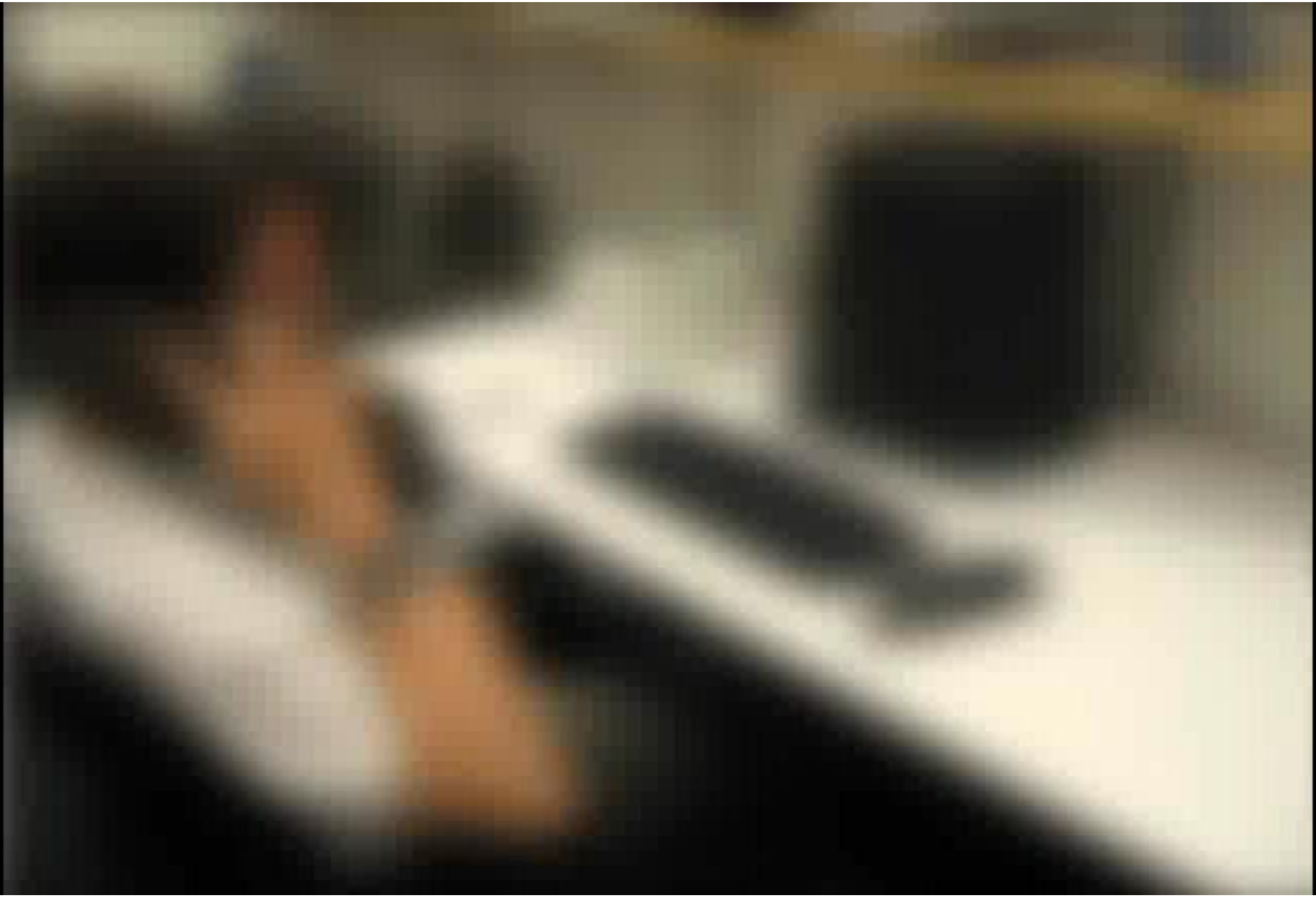
- Changes the interpretation of an object (or its function)



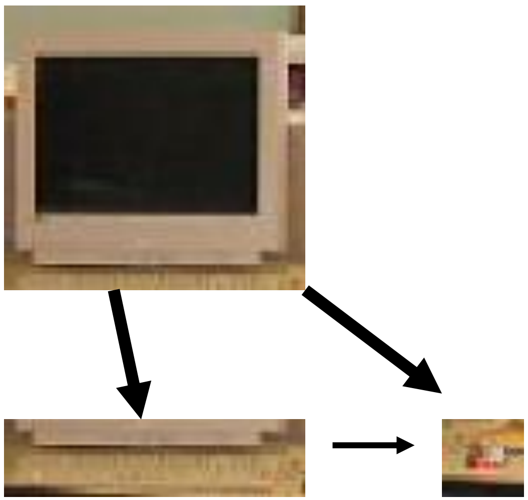
- Context defines what an unexpected event is

Look-Alikes by Joan Steiner

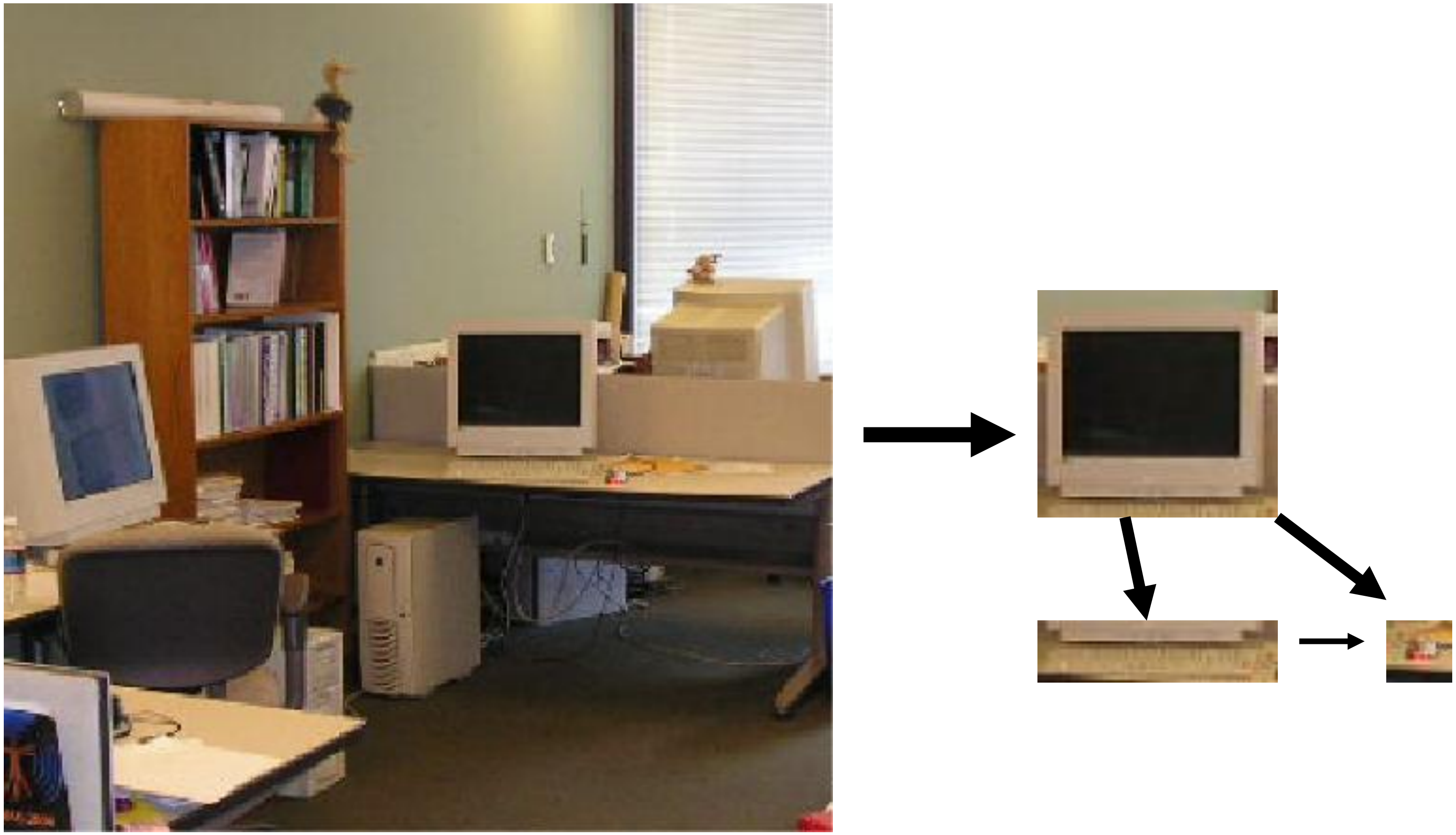
Even in high resolution, we can not shut down contextual processing and it is hard to recognize the true identities of the elements that compose this scene.



Modeling object relationships



Detecting difficult objects



Detect first simple objects (reliable detectors) that provide strong contextual constraints to the target (screen -> keyboard -> mouse)

Torralba, Murphy, Freeman. NIPS 2004.

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

Object model

Context model

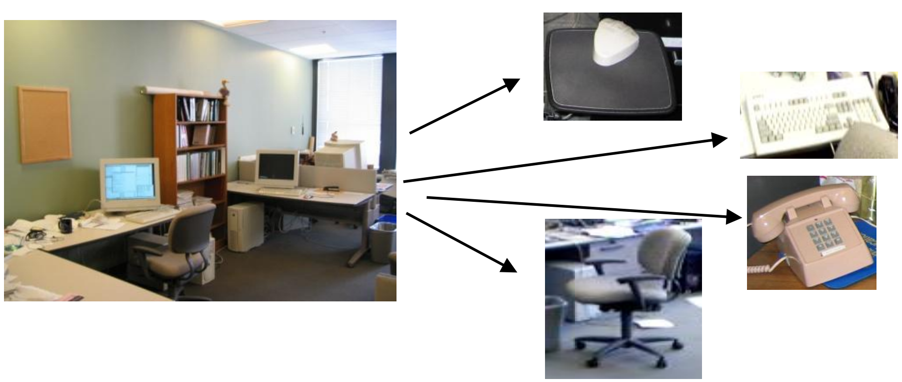
Full joint

Scene model

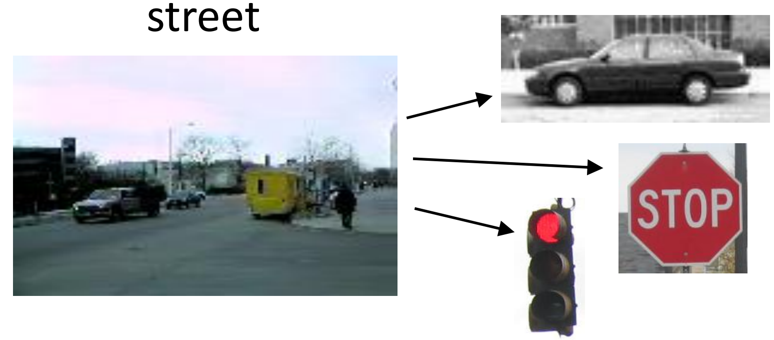
Approx. joint

$$p(O) = \sum_{s} \prod_{i} p(Oi|S=s)\ p(S=s)$$

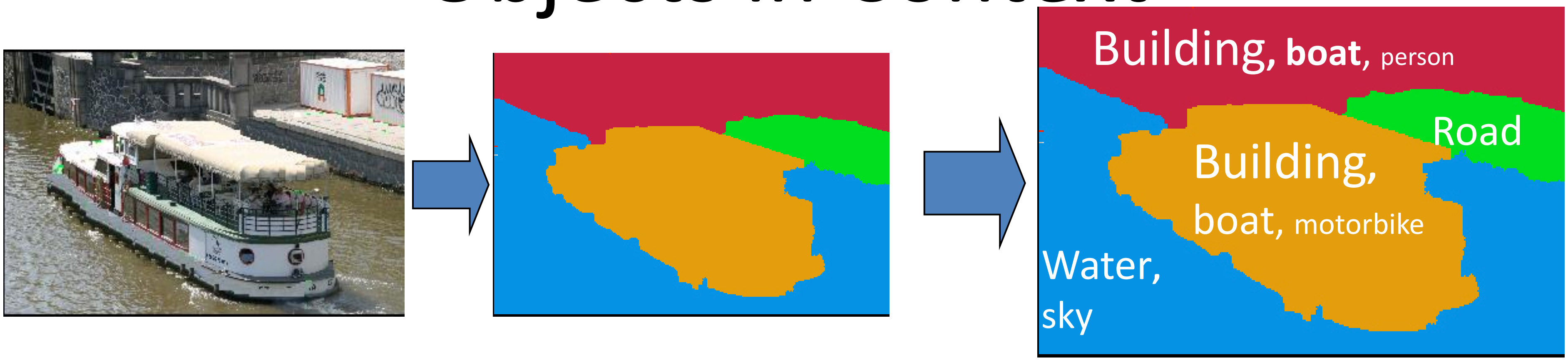
office

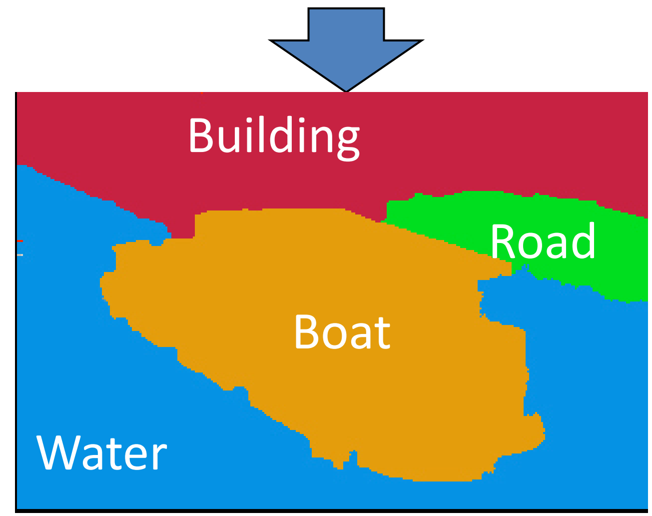


street



Objects in Context

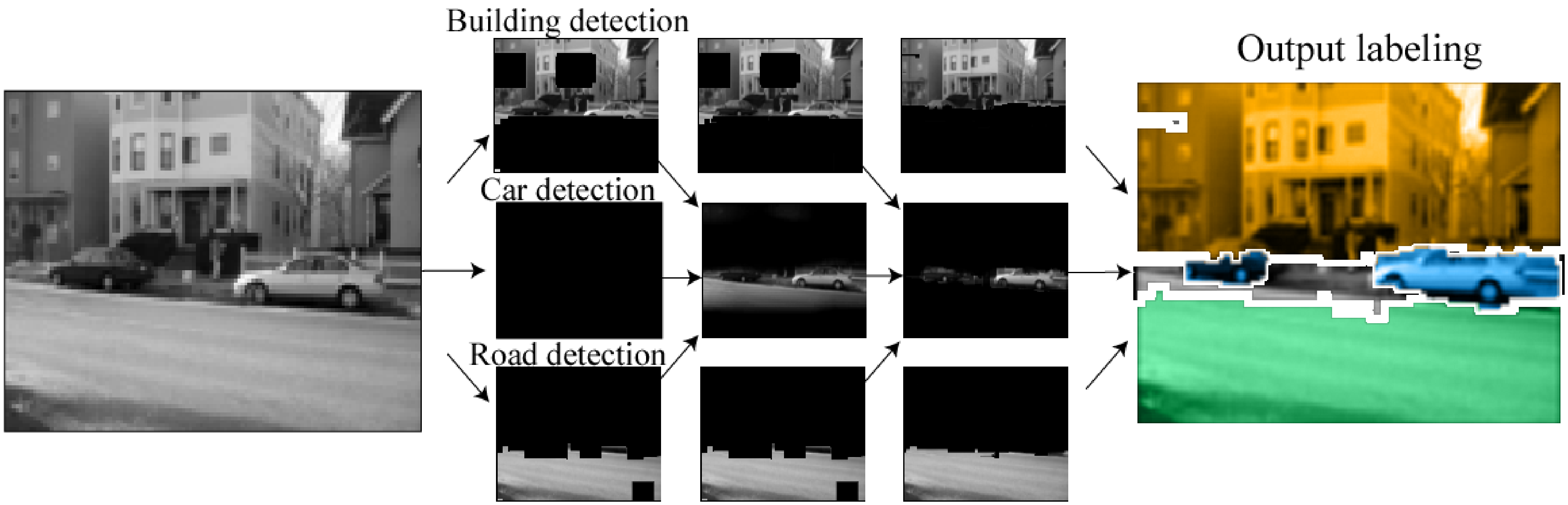


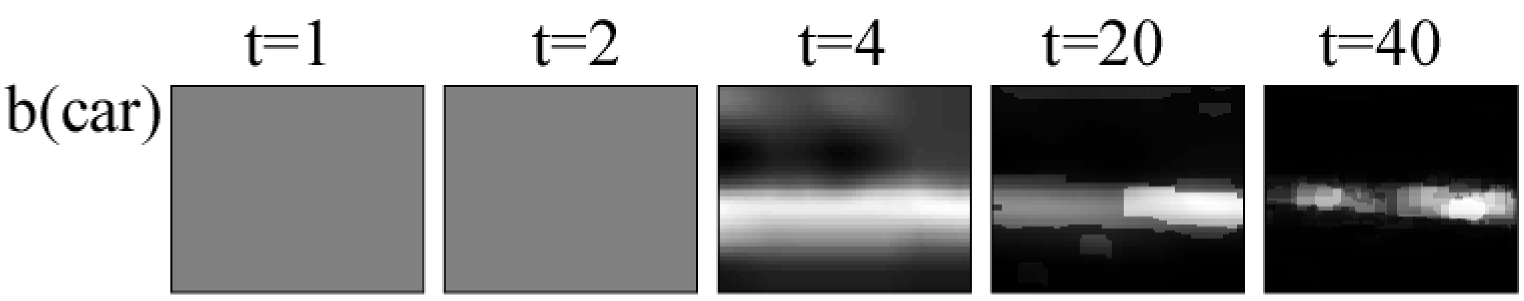


Most consistent labeling according to *object co-occurrences*& locallabel probabilities.

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

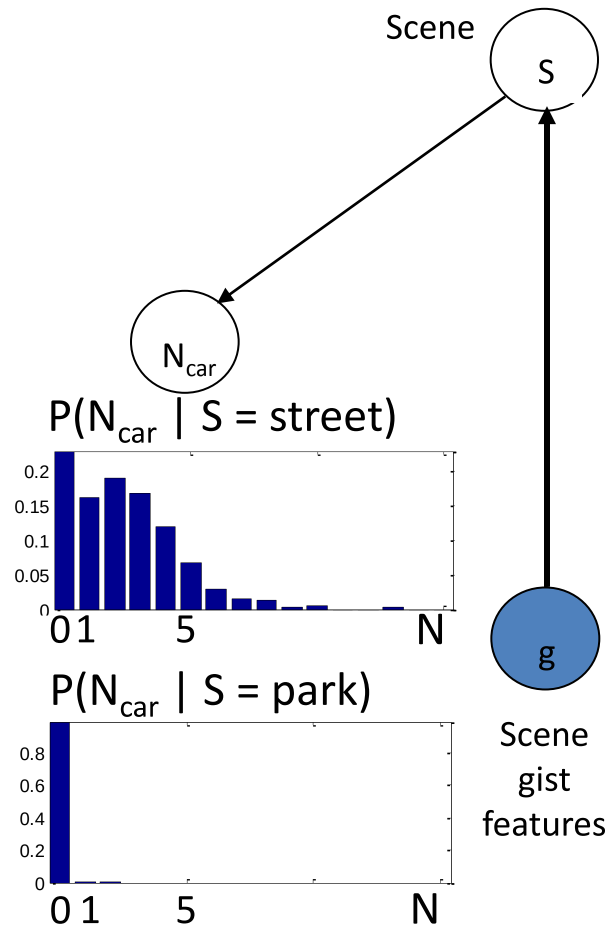
BRF for car detection: results



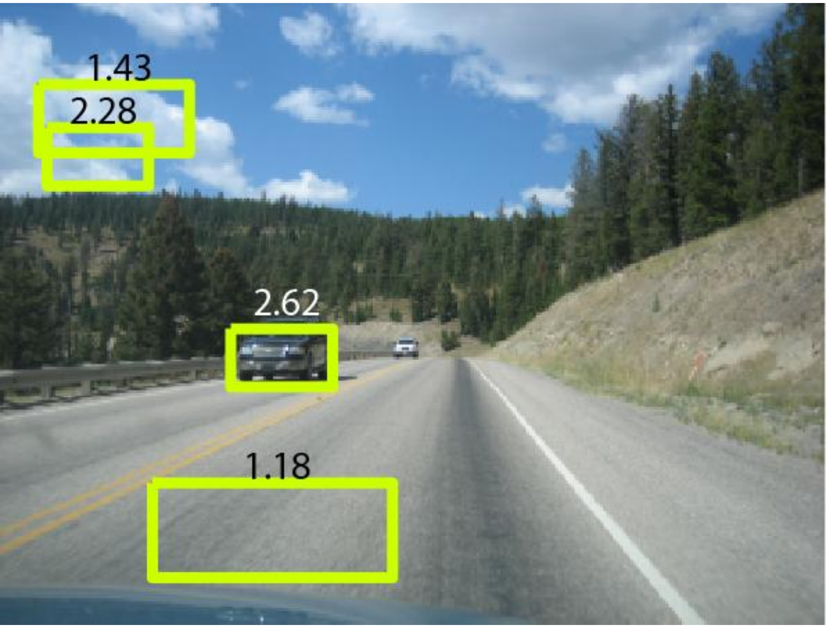


Torralba Murphy Freeman (2004)

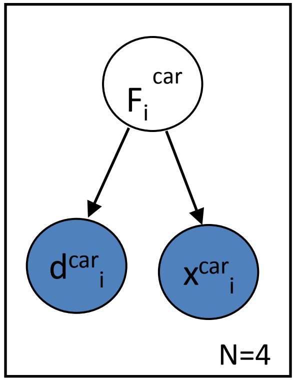
An integrated model of Scenes, Objects, and Parts



An integrated model of Scenes, Objects, and Parts



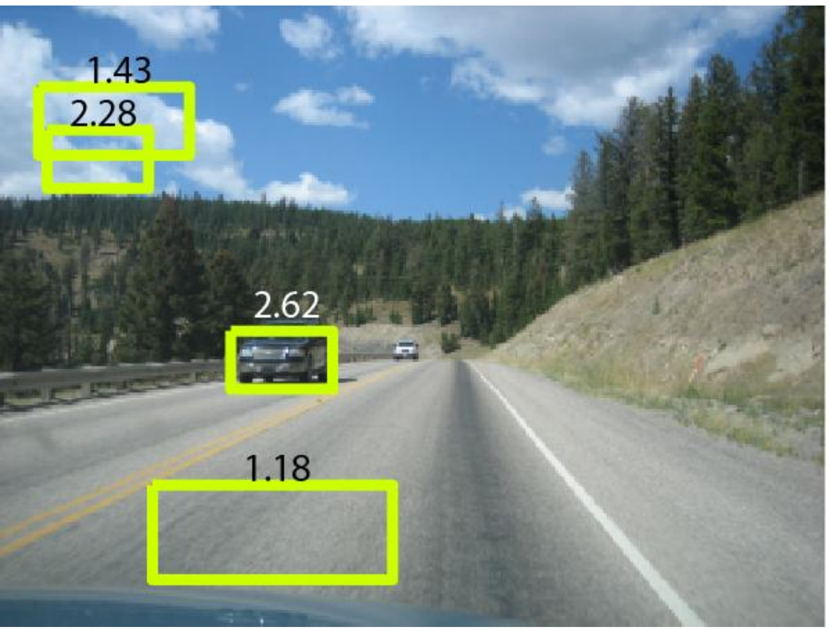
We train a multiview car detector.



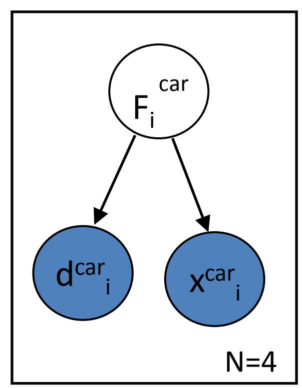
$p(d \mid F=1) = \mathcal{N}(d \mid \mu_1, \sigma_1)$

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

An integrated model of Scenes, Objects, and Parts

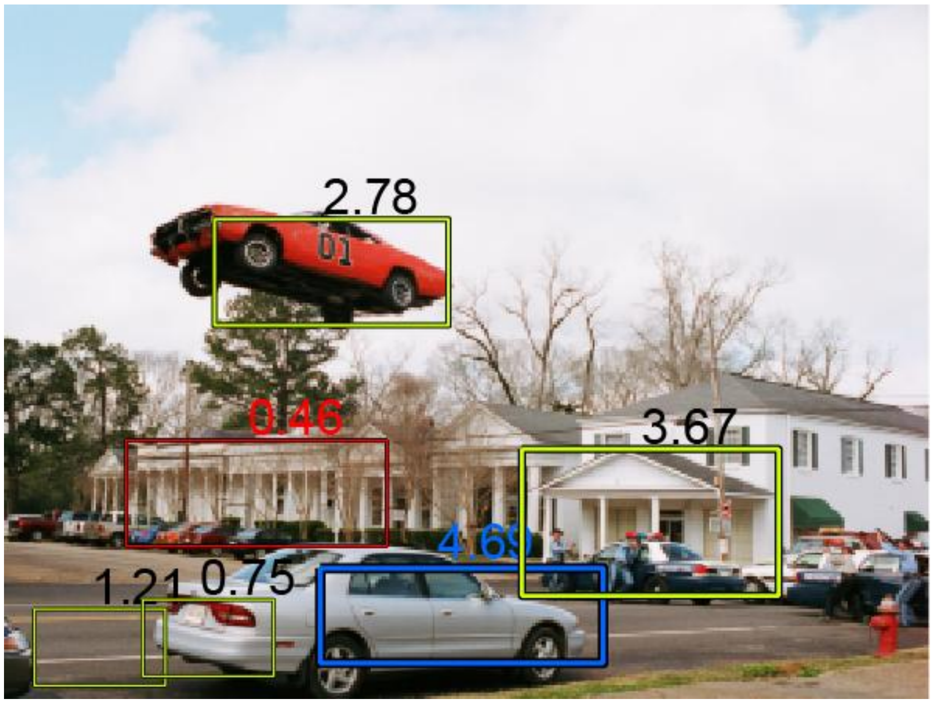


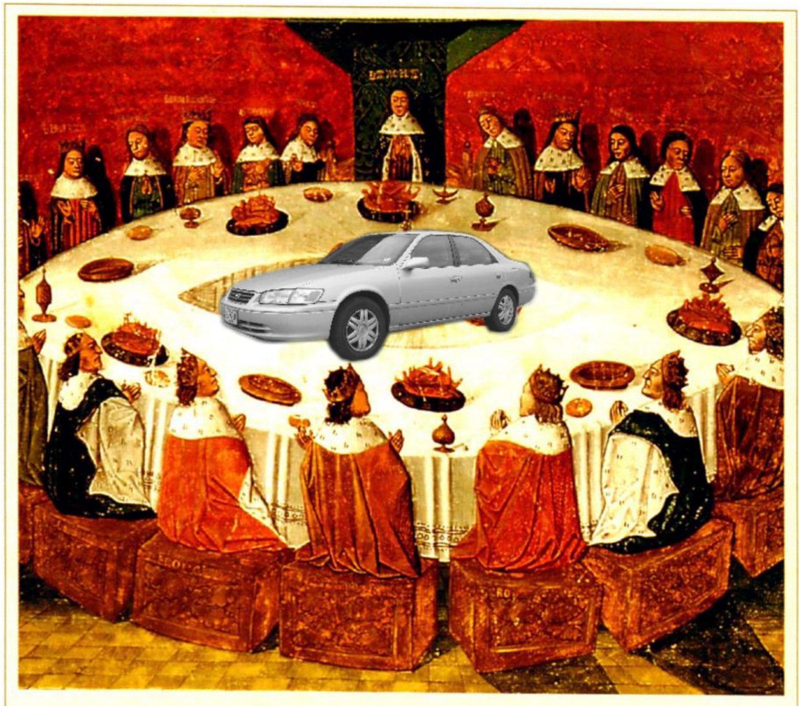
We train a multiview car detector.



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

A car out of context …

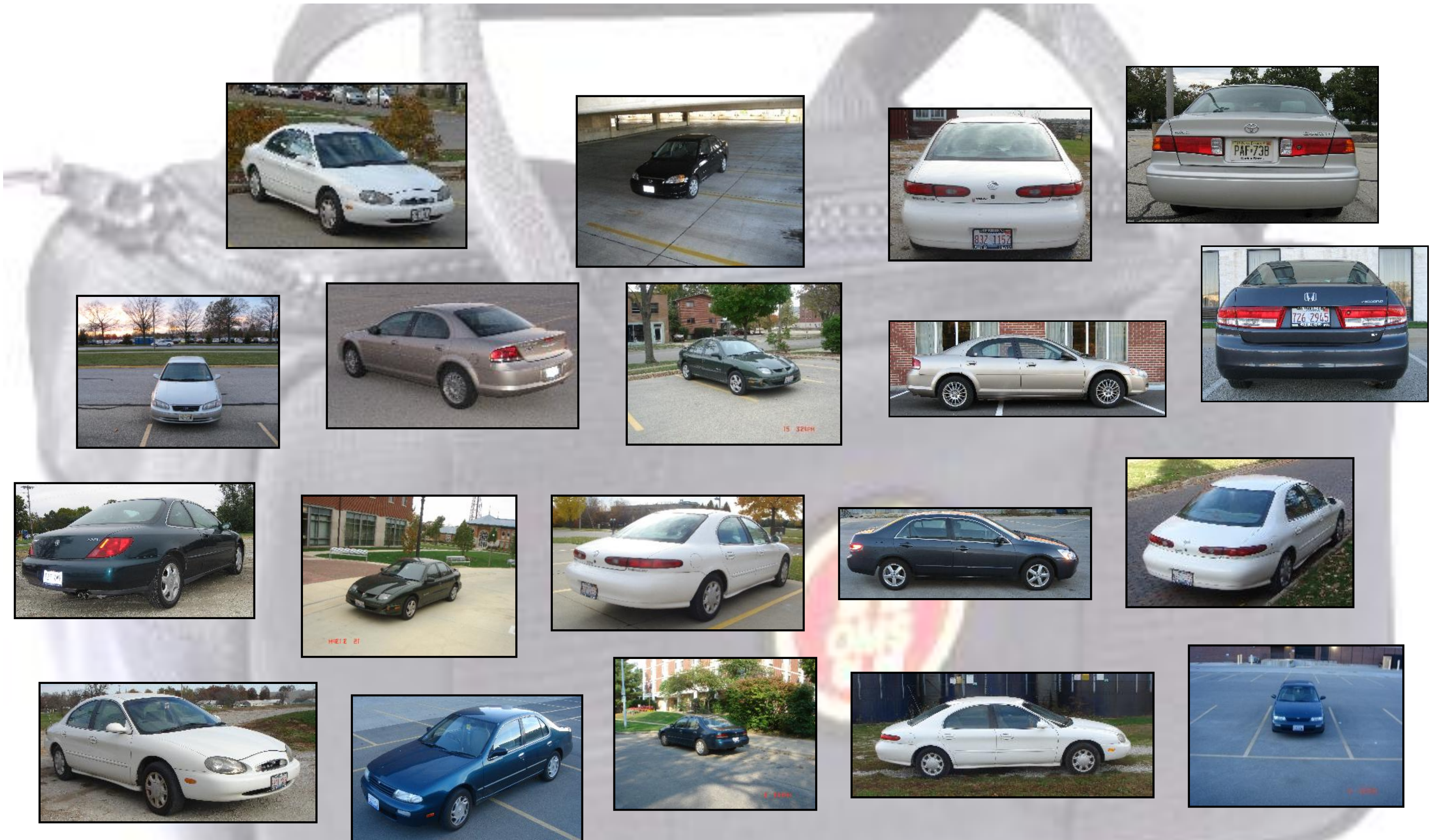




3D object categorization

Courtesy of Prof. Silvio Savarese (U. Michigan, Ann-Arbor)

3D Object Categorization



- Weber et al. ‘00
- Schneiderman et al. ’01
- •Capel et al ’02
- •Johnson & Herbert ‘99

- •Bronstein et al, ‘03
- •Ruiz-Correa et al. ’03,
- •Funkhouser et al ’03
- •Bart et al ‘04

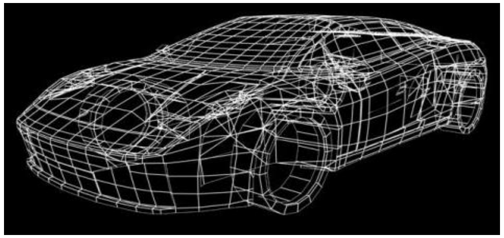
- •Thomas et al. ‘06
- Kushal, et al., ’07
- Savarese et al, 07, 08

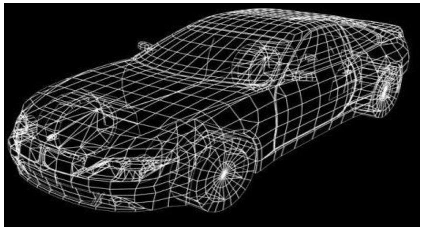
- Chiu et al. ‘07
- Hoiem, et al., ’07
- Yan, et al. ’07

3D Object Categorization

Challenges

- how to model 3D shape variability?



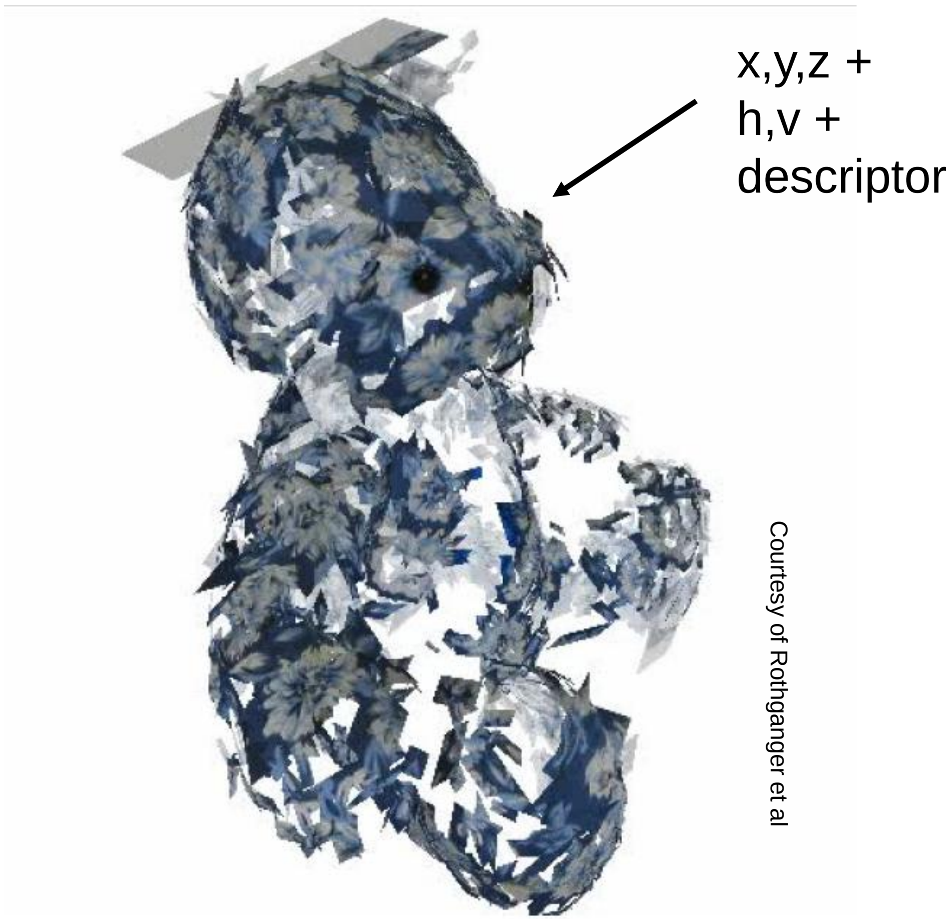


- How to model texture (appearance) variability?

- How to link texture (appearance) across views?

Object representation: Collection of patches in 3D

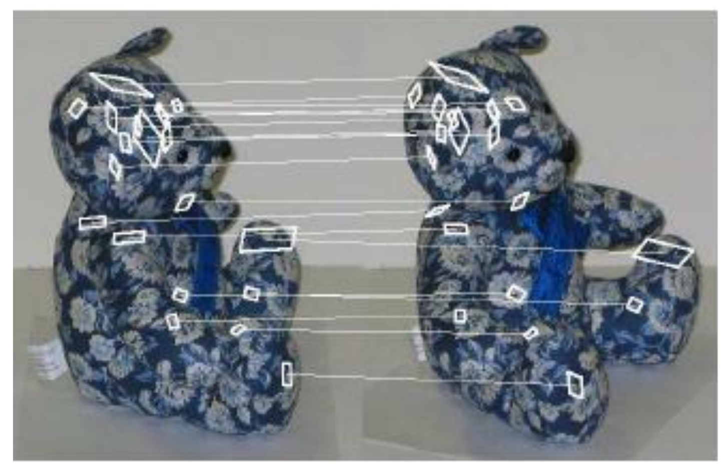
Rothganger et al. ’06



Model learning Rothganger et al. ‘03 ’06

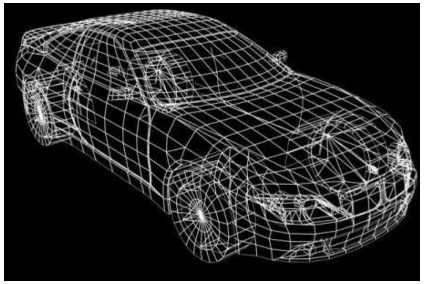
Build a 3D model:

- N images of object from N different view points
- Match key points between consecutive views [create sample set]
- Use affine structure from motion to compute 3D location and orientation + camera locations [RANSAC]
- Find connected components
- Use bundle adjustment to refine model
- Upgrade model to Euclidean assuming zero skew and square pixels

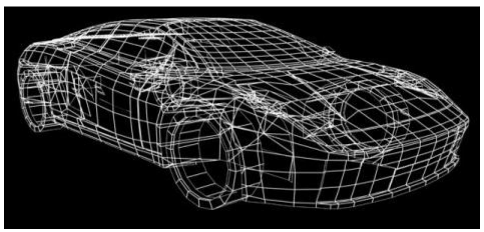


Full 3D models

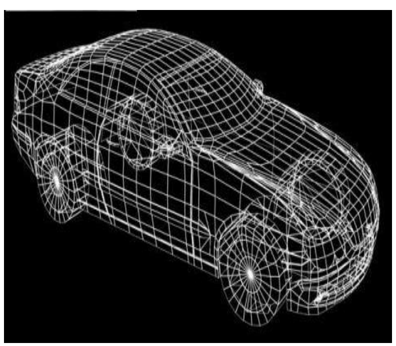
- Bronstein et al, ‘03
- Ruiz-Correa et al. ’03,
- Funkhouser et al ’03
- Kazhdan et al.03
- Osada et al ‘02
- Capel et al ’02
- Johnson & Herbert ’99
- Amberg et al ‘08



3D model instance



⋮



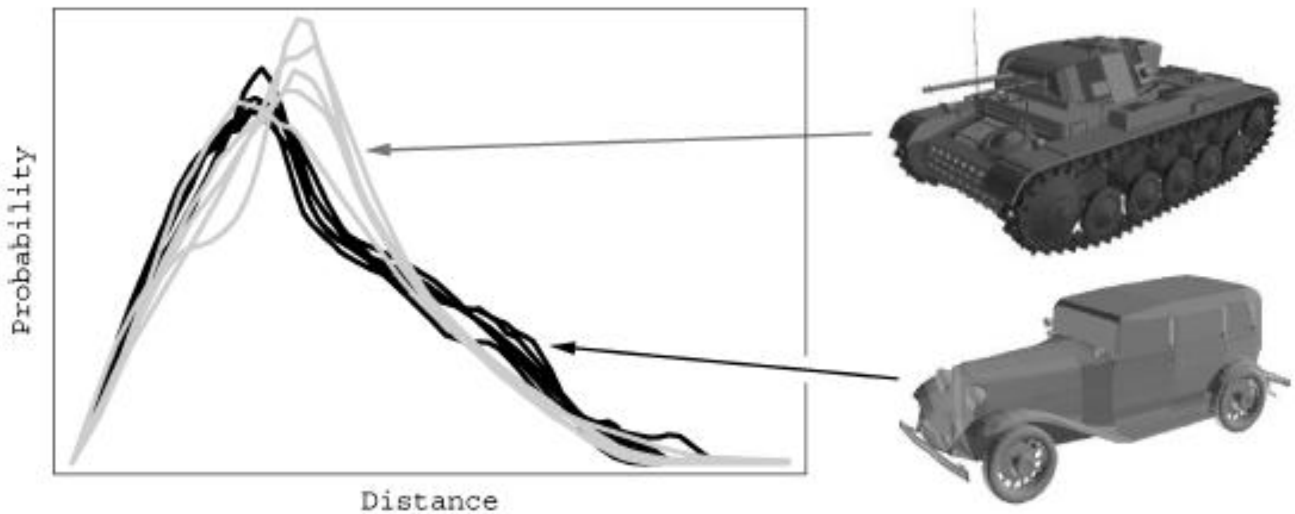
3D model instance

3D category model

A 3D model category is built from a collection of 3D range data or CAD models

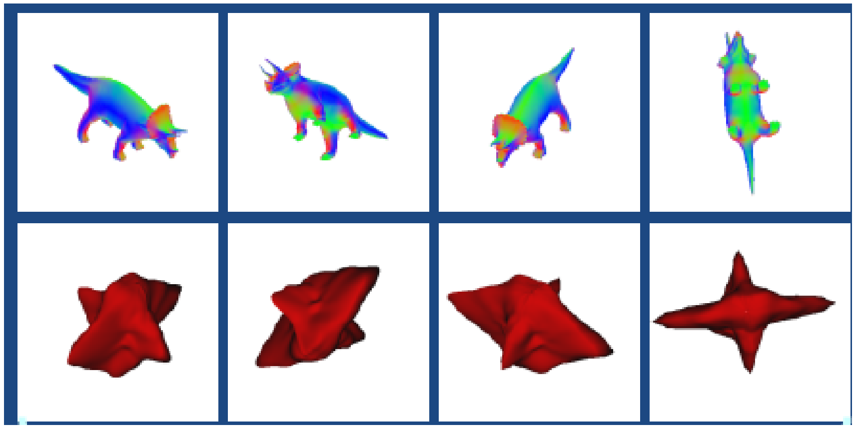
Shape distributions

Osada et al 02

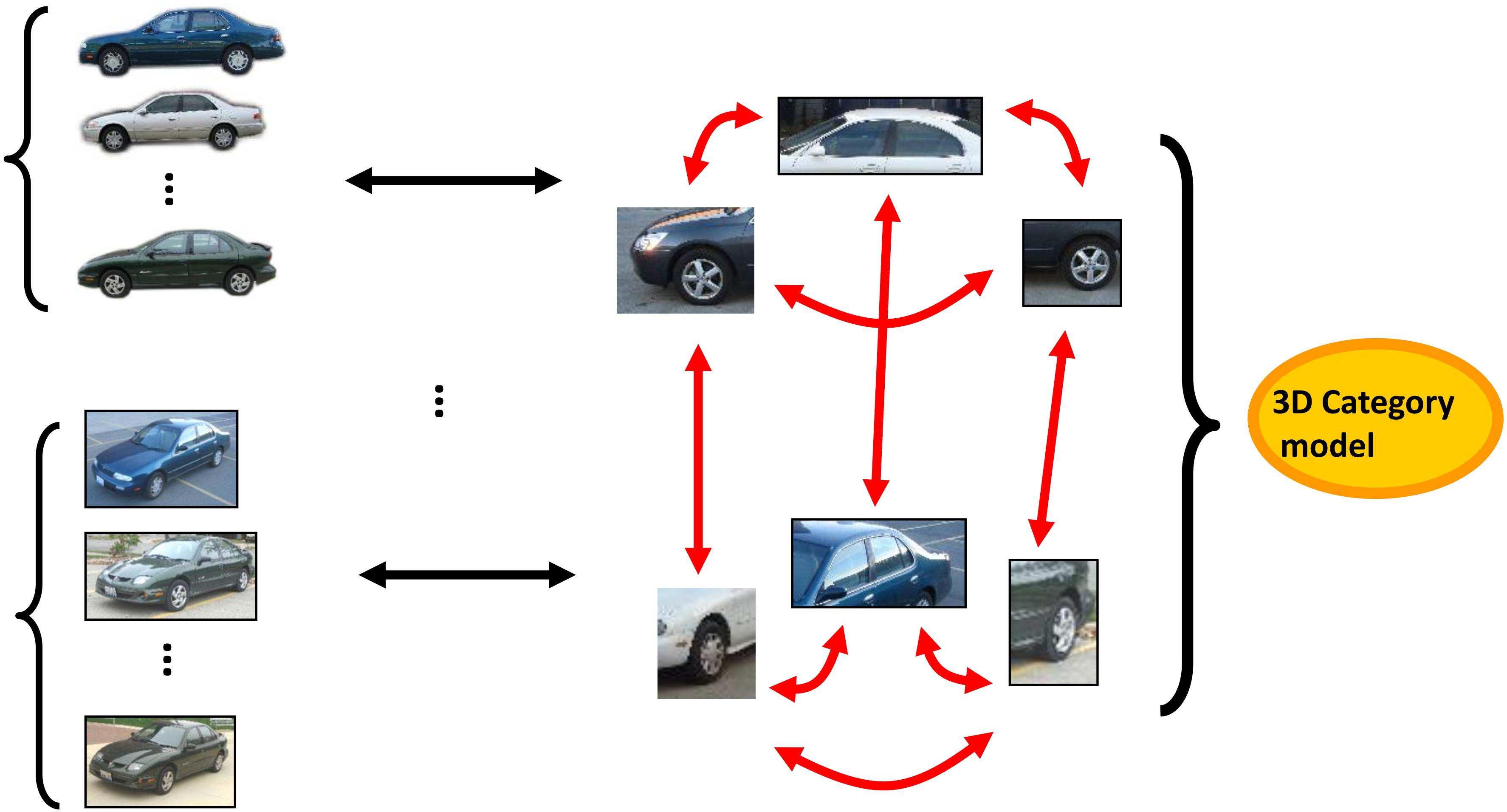


Spherical harmonics

Kazhdan et al. 03



Multi-view models



Sparse set of interest points or parts of the objects are linked across views.

Multi-view models by rough 3d shapes

Yan, et al. '07

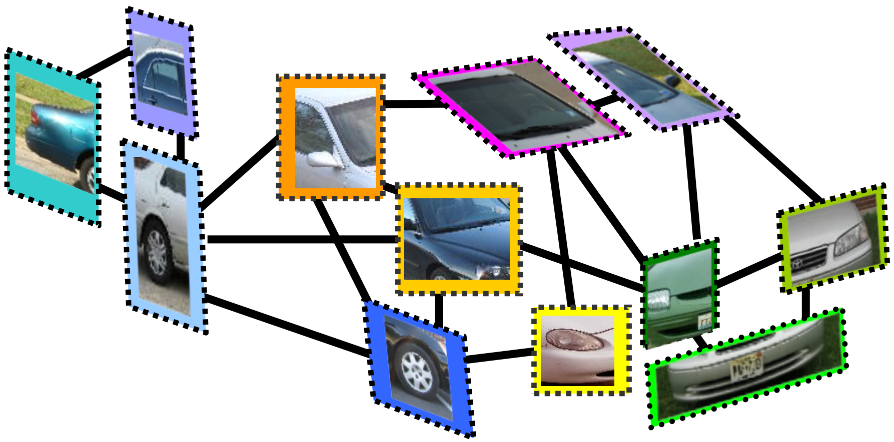
A **unified framework** for 3D object detection, pose classification, pose synthesis

Savarese, Fei-Fei, ICCV 07

Savarese, Fei-Fei, ECCV 08

Sun, Su, Savarese, Fei-Fei, CVPR 09

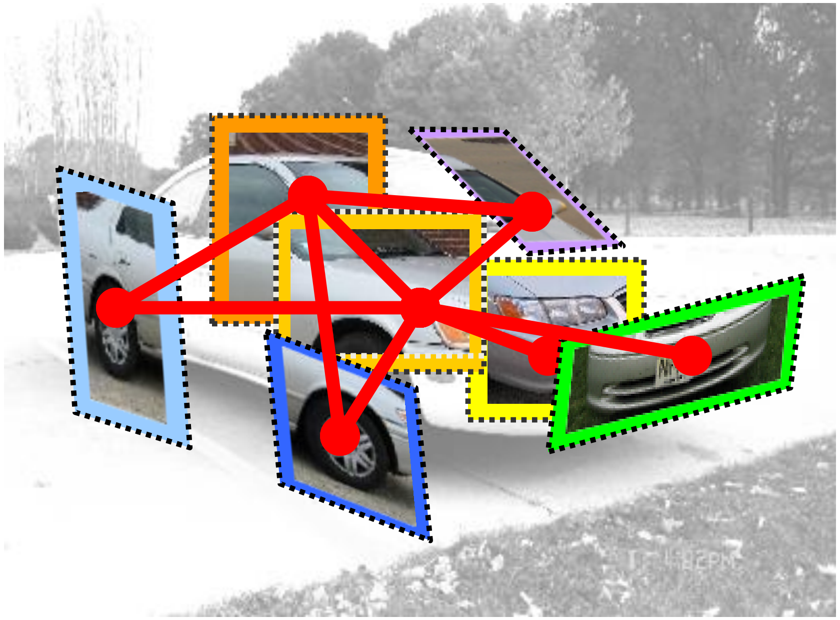
Su, Sun, Fei-Fei, Savarese, ICCV 09



- Canonical parts captures diagnostic appearance information
- 2d ½ structure linking parts via weak geometry

Object Recognition

Query image



model

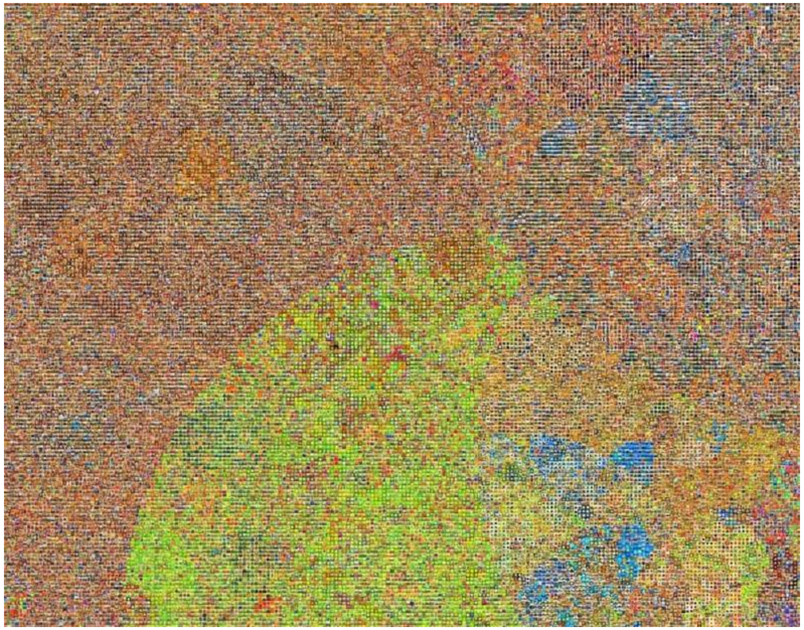
Algorithm

1. **Find hypotheses of canonical parts consistent with a given pose**

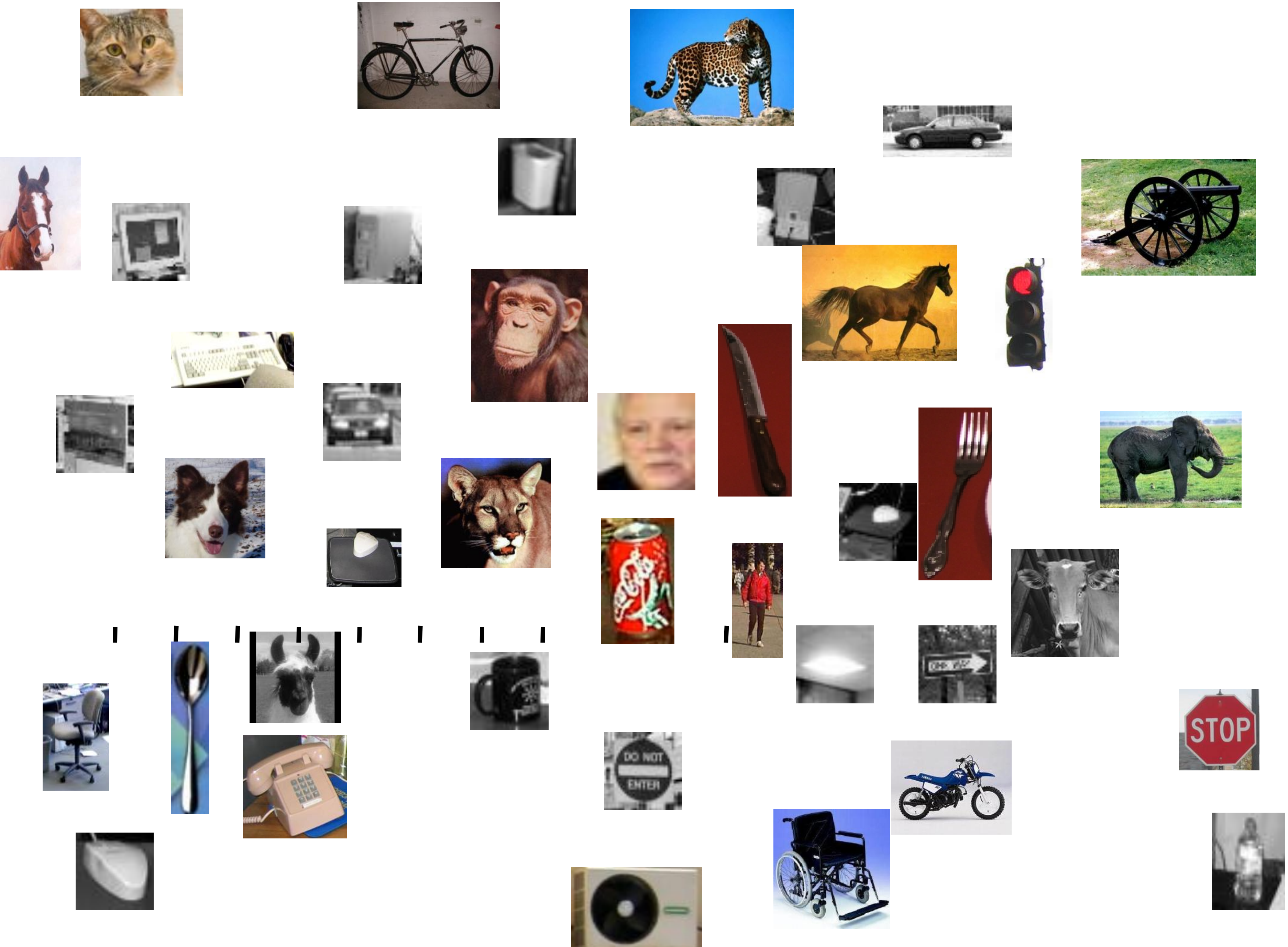
2. **Infer position and pose of other canonical parts**

3. **Optimize over $\mathbf{E}$, $\mathbf{G}$ and $\mathbf{s}$ to find best combination of hypothesis**

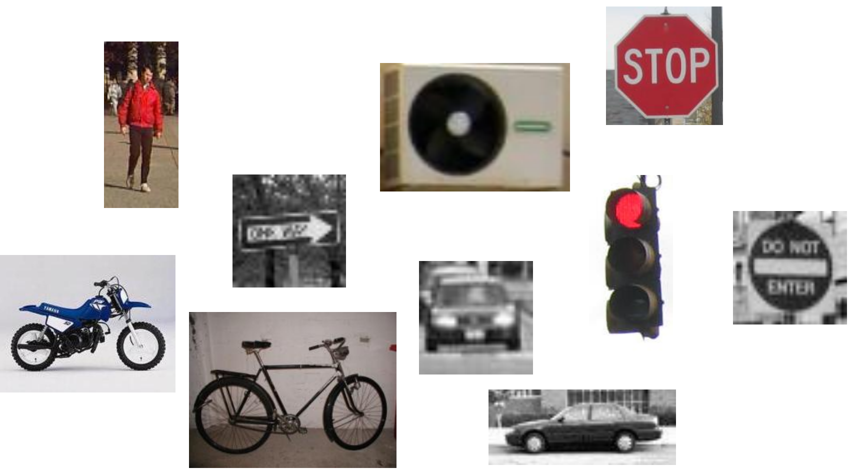
$\rightarrow$ **error**

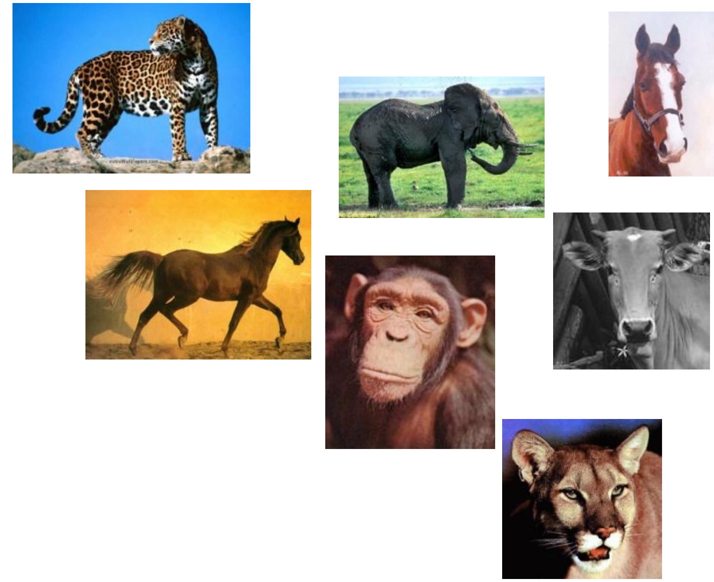


Multiclass object detection

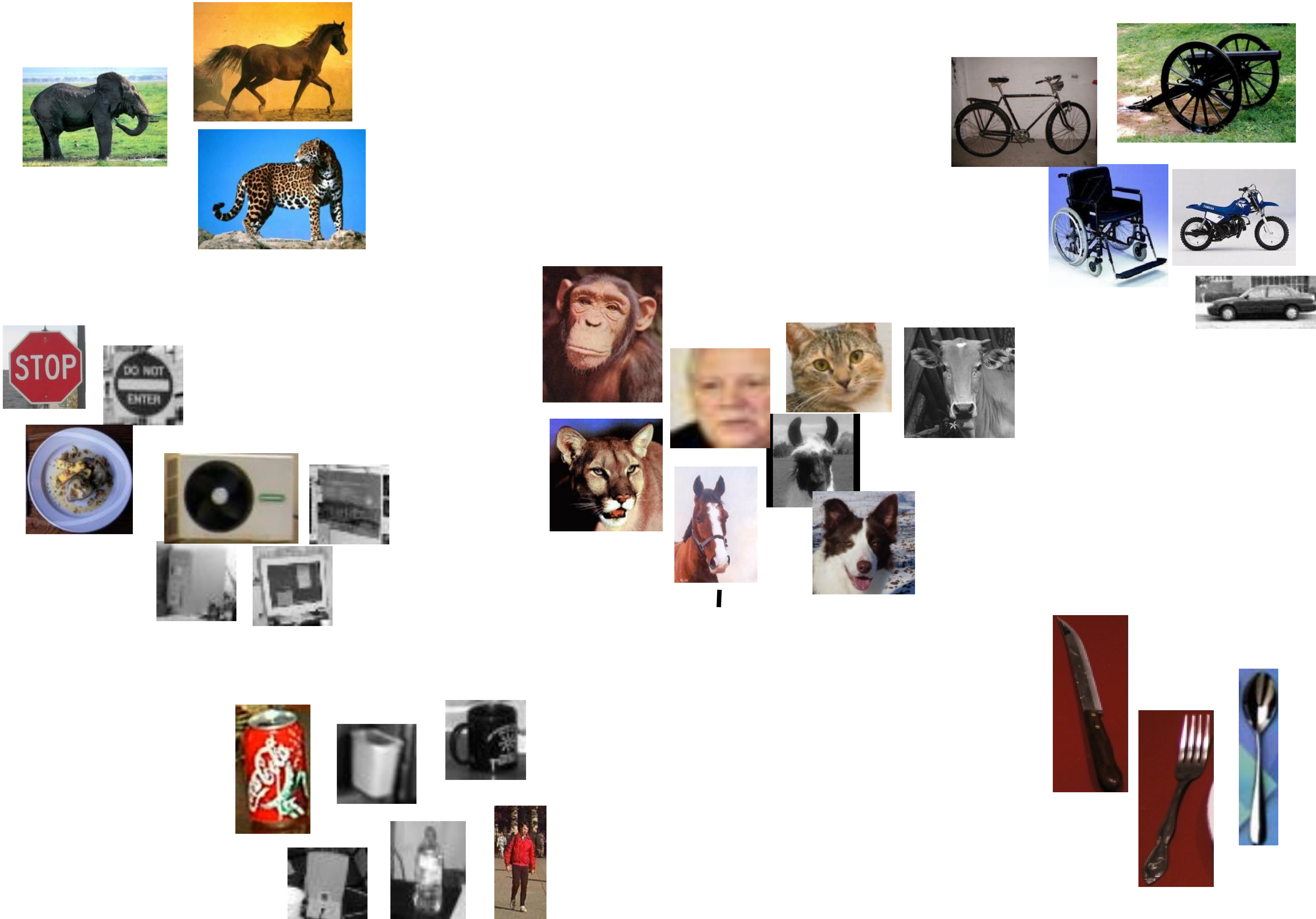

STOP
DO NOT
ENTER

Context: objects appear in configurations





Generalization: objects share parts



How many object categories are there?

Slide by Aude Oliva

We do not need to recognize the exact category

A new class can borrow information from similar categories

A bird

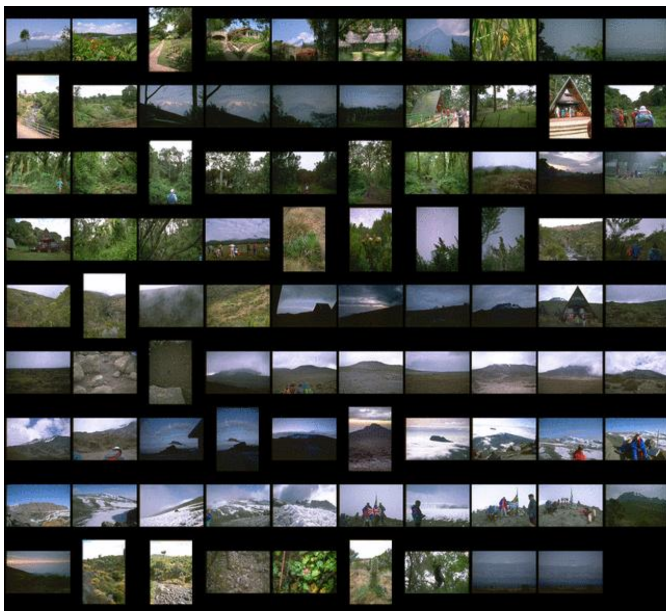
An ostrich

Large Scale Recognition and Retrieval

Scaling to billions of images

Object Recognition for large-scale search

Focus on scaling rather than understanding image



Google Image Search

Content-Based Image Retrieval

- Variety of simple/hand-designed cues:
  - Color and/or Texture histograms, Shape, PCA, etc.
- Various distance metrics
  - Earth Movers Distance (Rubner et al. ‘98)

- ~~QBIC from IBM (1999)~~
- Blobworld, Carson et al. 2002

flickr

Here is a shot of me and my brothers at my brother Jon's wedding to his first wife. I was 17, Garth was 19 and Jon was 21.

This photo has notes. Move your mouse over the photo to see them.

Page last updated at 14:45 GMT, Friday, 11 September 2009 15:45 UK

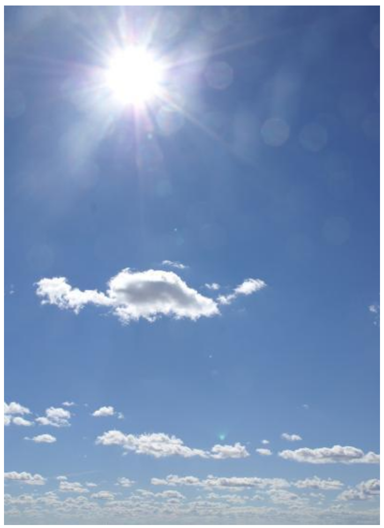
E-mail this to a friend

Day in pictures

A Thai government employee risks a close up shot of a captive tiger in Ratchaburi province as part of a scheme to tackle illegal trading by creating a database of the animals.

AP

Vision and language in human brain



Sentence:
The sky is blue.

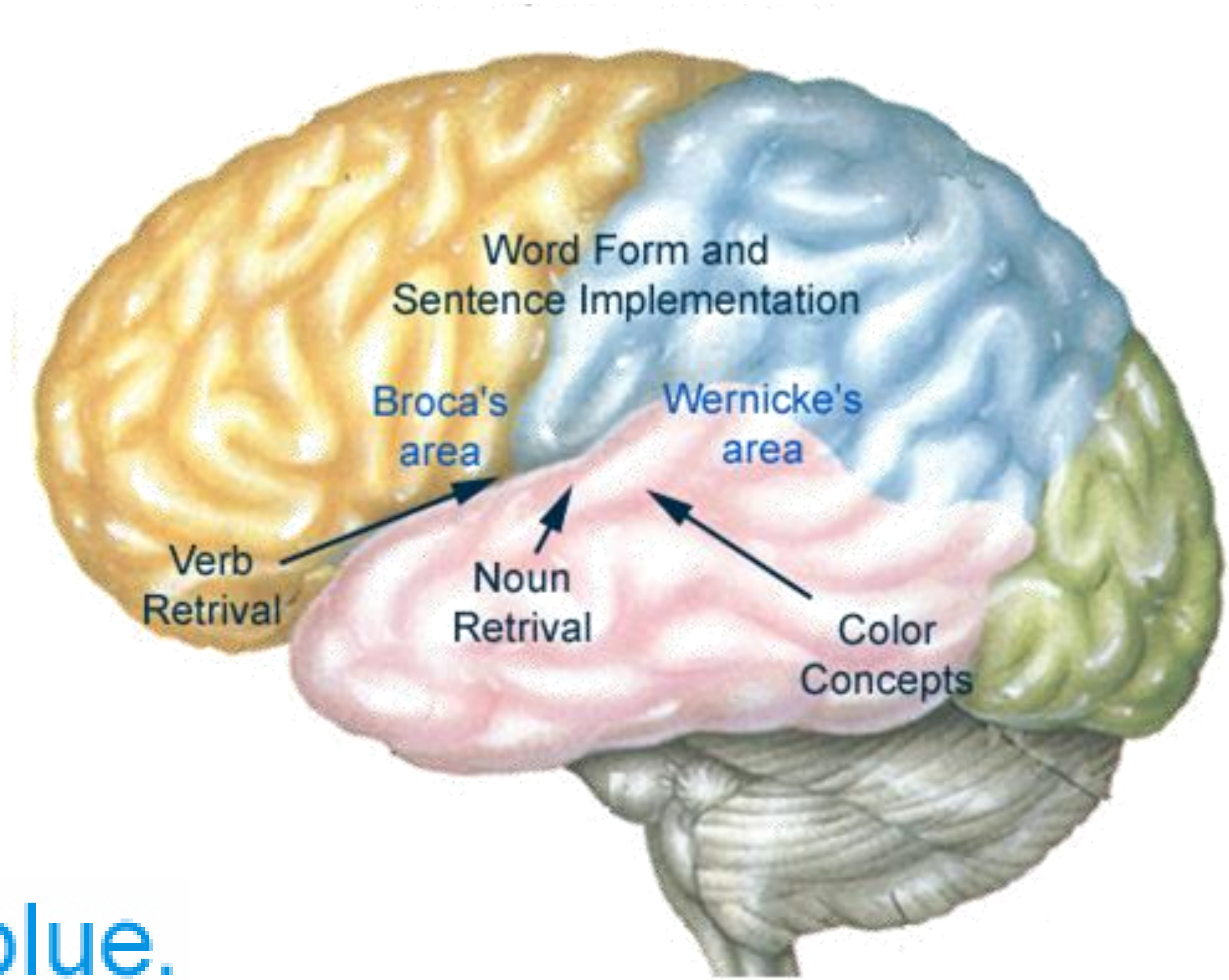


figure modified from: http://www.colorado.edu/intphys/Class/IPHY3730

Barnard et al. JMLR, 2005

Automatic Image Annotation: ALIP

Annotation Process

- Classification results form the basis
- Salient words appearing in the classification favored more

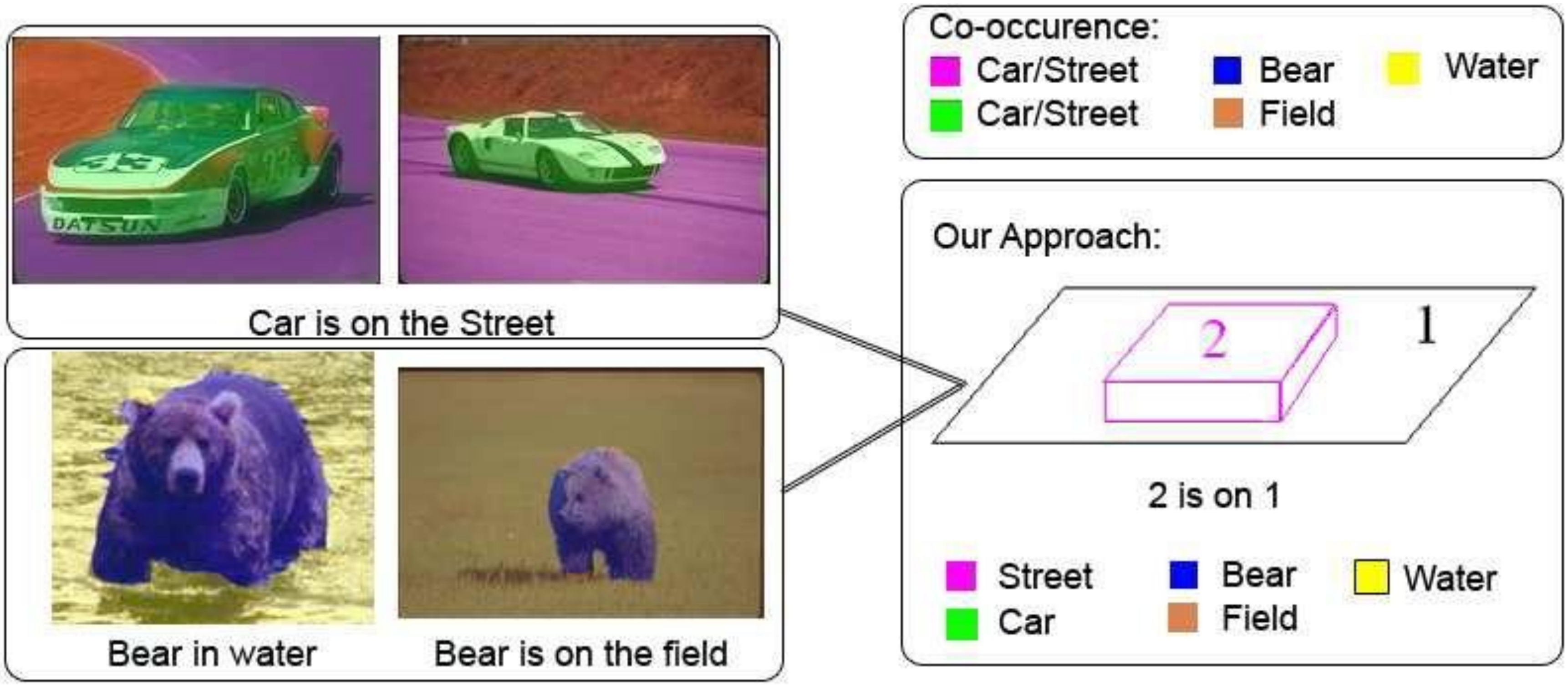
Snow, animal, wildlife, sky, cloth, ice, people

Building, sky, lake, landscape, Europe, tree

Food, indoor, cuisine, dessert

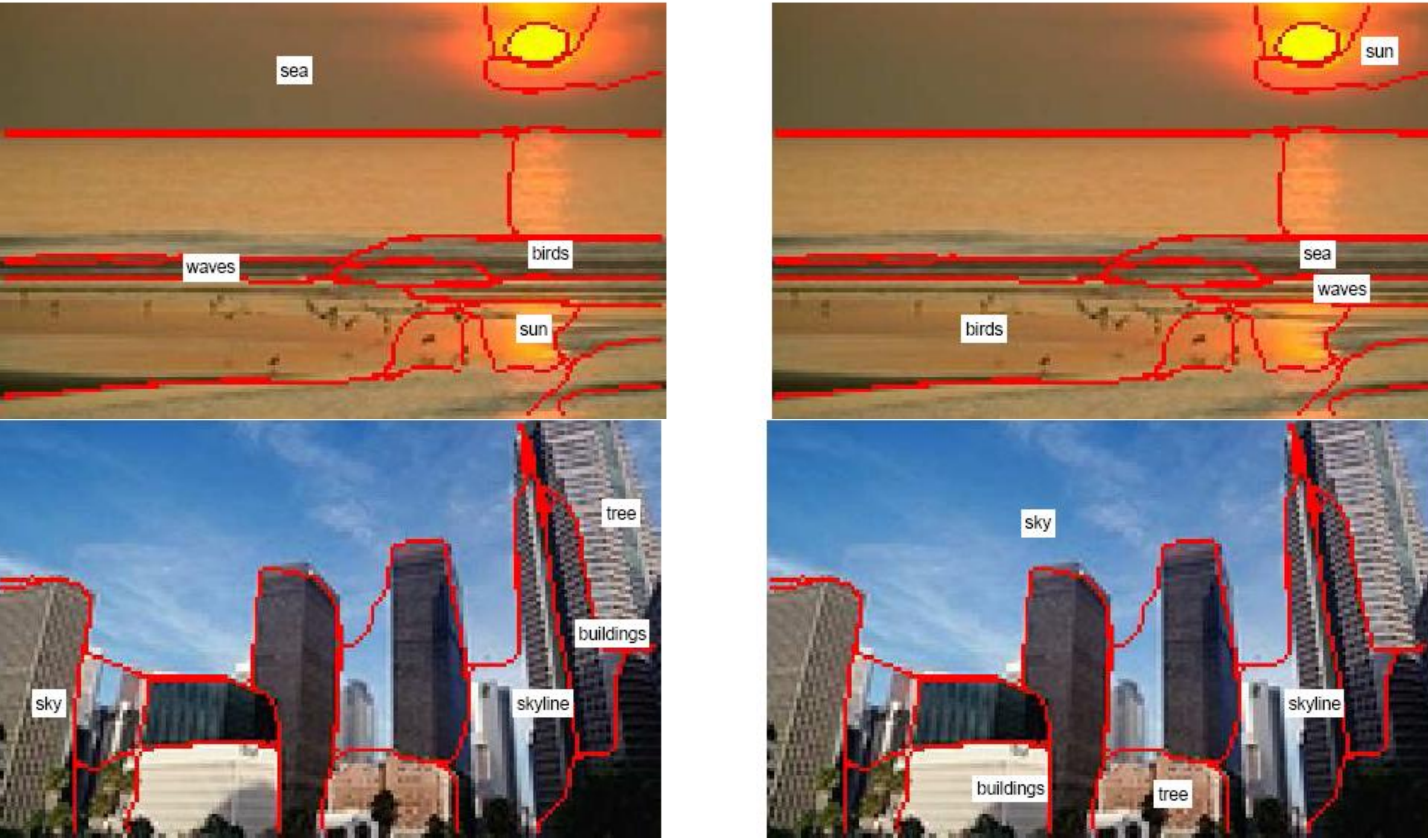
Slide courtesy of Ritendra Datta, Jia Li, James Z. Wang

“Beyond nouns”



Gupta & Davis, EECV, 2008

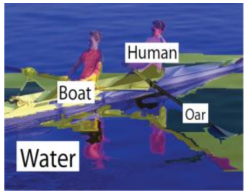
(i) Duygulu et. al (2002) | (ii) Our Approach



Gupta & Davis, EECV, 2008

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

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

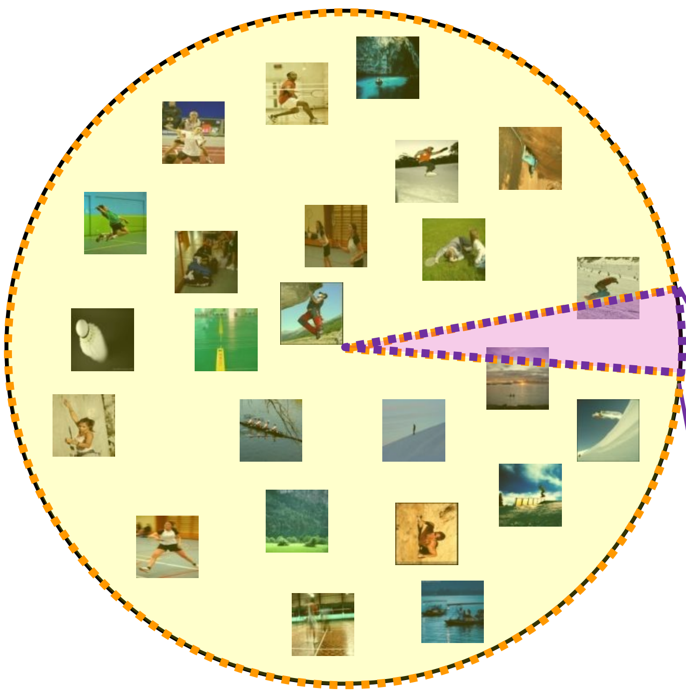


Total Scene

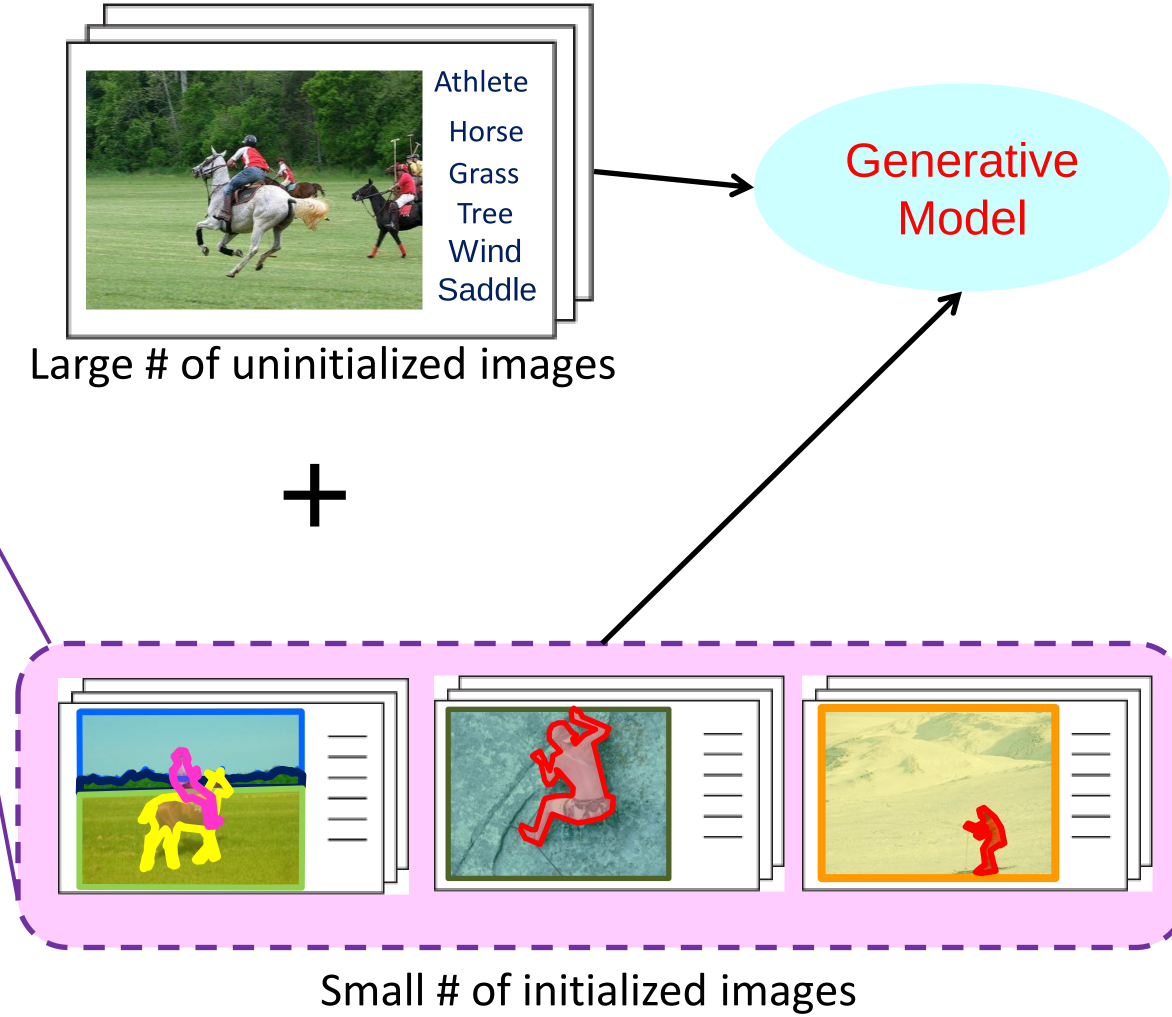
Auto-semi-supervised learning:

Small # of initialized images + Large # of uninitialized images

Scene/Event images from the Internet



flickr



L-J Li , R. Socher & L. Fei-Fei, CVPR, 2009

Datasets

10^0
images

1972

10^0
images

1972

LabelMe

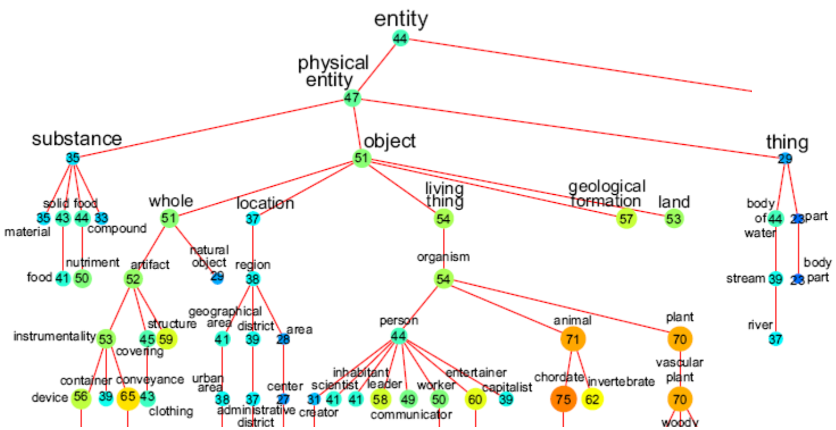
10^5 images

Russell, Torralba, Freman, 2005

80.000.000 images

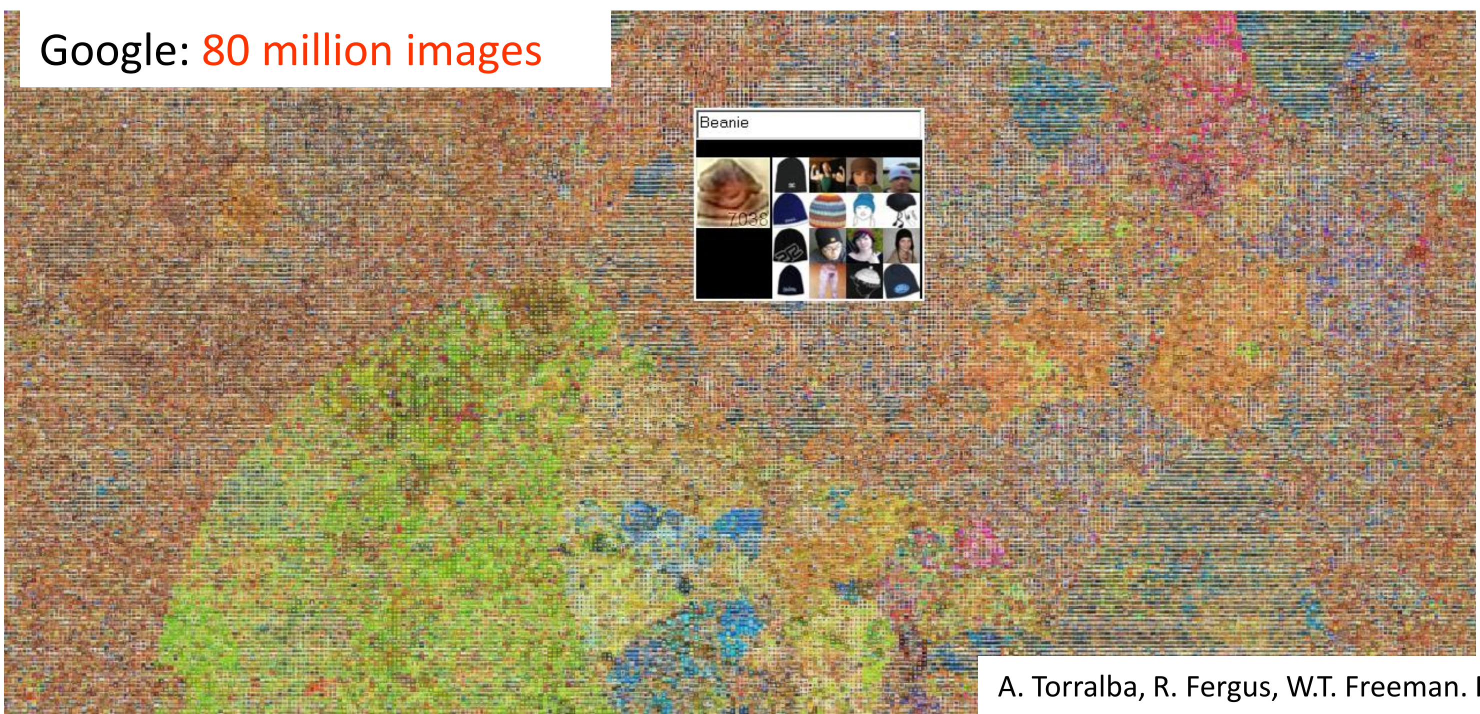
$10^{6\text{-}7}$ images

75.000 non-abstract nouns from WordNet



7 Online image search engines

And after 1 year downloading images

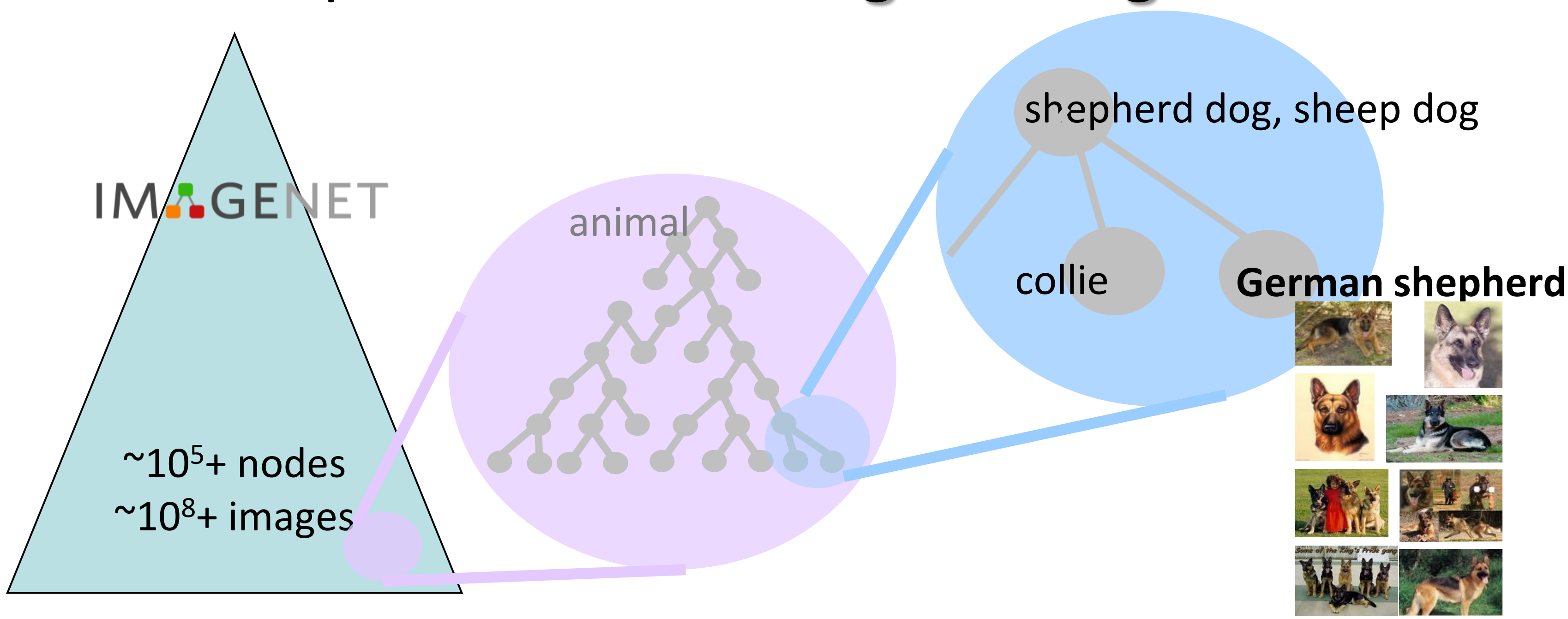


A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008

IMAGENET

$10^{6\text{-}7}$
images

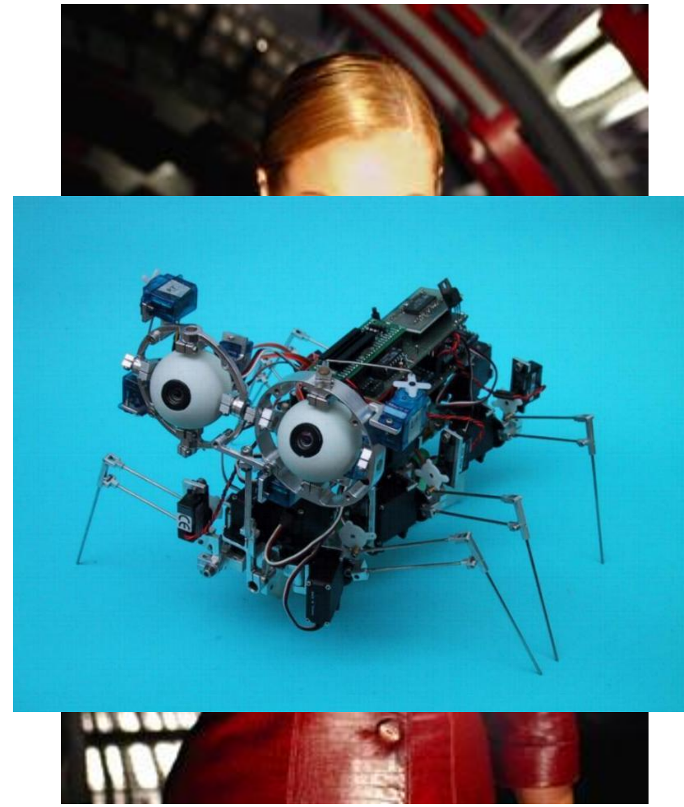
- An ontology of images based on WordNet
- ImageNet currently has
  - ~15,000 categories of visual concepts
  - 10 million human-cleaned images (~700im/categ)
  - Free to public @ **www.image-net.org**



Deng, Dong, Socher, Li & Fei-Fei, CVPR 2009

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

Labeling to get a Ph.D.

Labeling for fun

Labeling for money

amazonmechanical turk
beta
Artificial Artificial Intelligence

Labeling because it gives you added value

Visipedia

Just labeling

LabelMe

Dataset labeling by crowd sourcing

A word of warning of crowd sourcing

“We've heard that a million monkeys at a million keyboards could produce the complete works of Shakespeare; now, thanks to the Internet, we know that is not true.”

-- Robert Wilensky, 1996