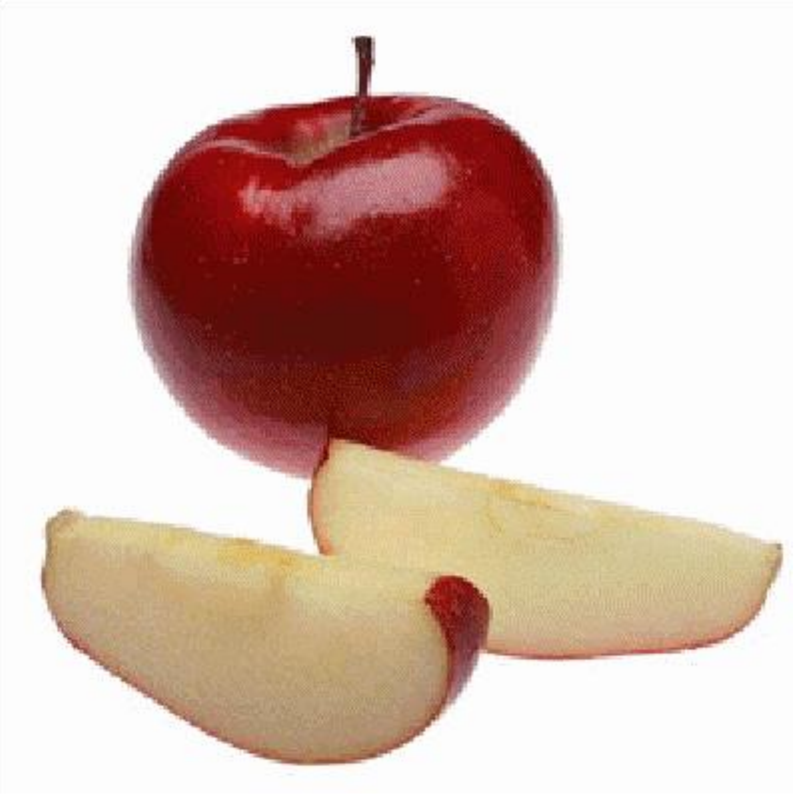
Object Recognition

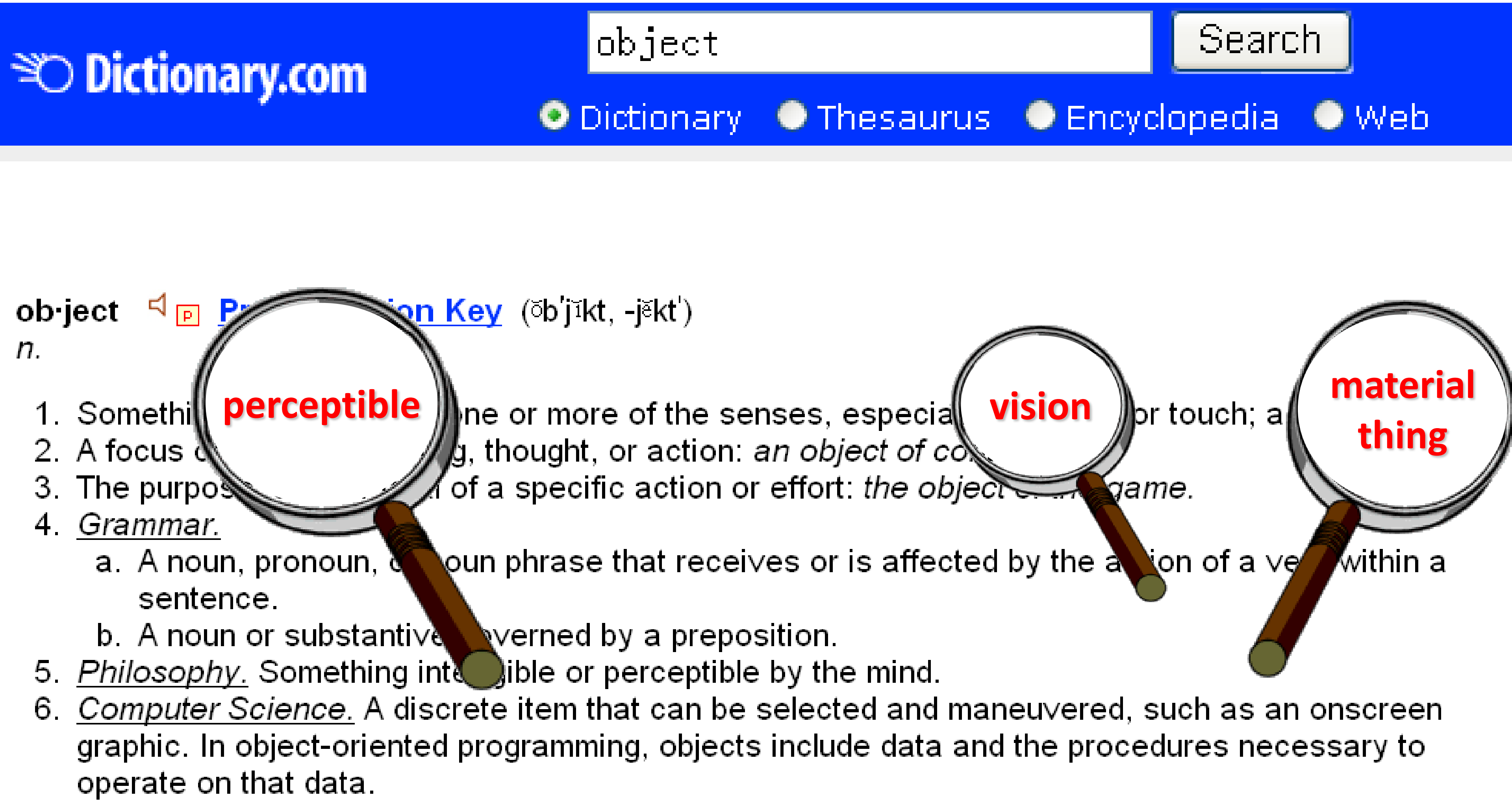
CMP719 – Computer Vision

Pinar Duygulu

(Slide credits:

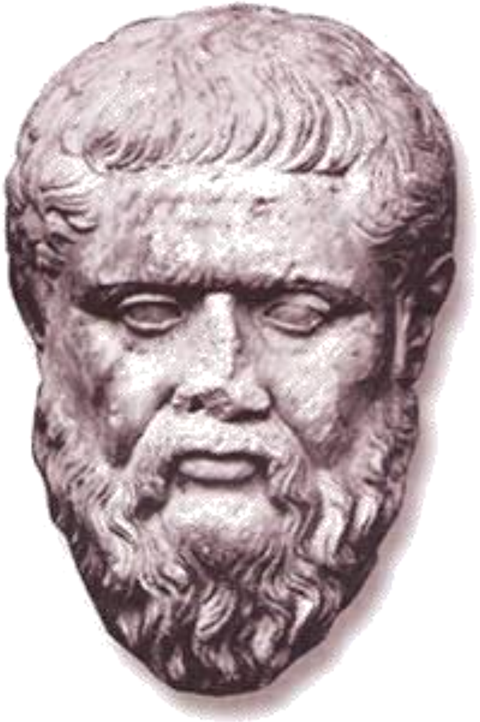
Kristen Grauman, Fei fei Li, Antonio Torralba, Hames Hays)




Dictionary.com
object
Search
Dictionary
Thesaurus
Encyclopedia
Web
ob·ject
Pr on Key (ŏb'jĭkt, -jĕkt')
n.
perceptible
vision
material thing
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 of a specific action or effort: the object 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 overned by a preposition.
5. Philosophy. Something inte ible 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.

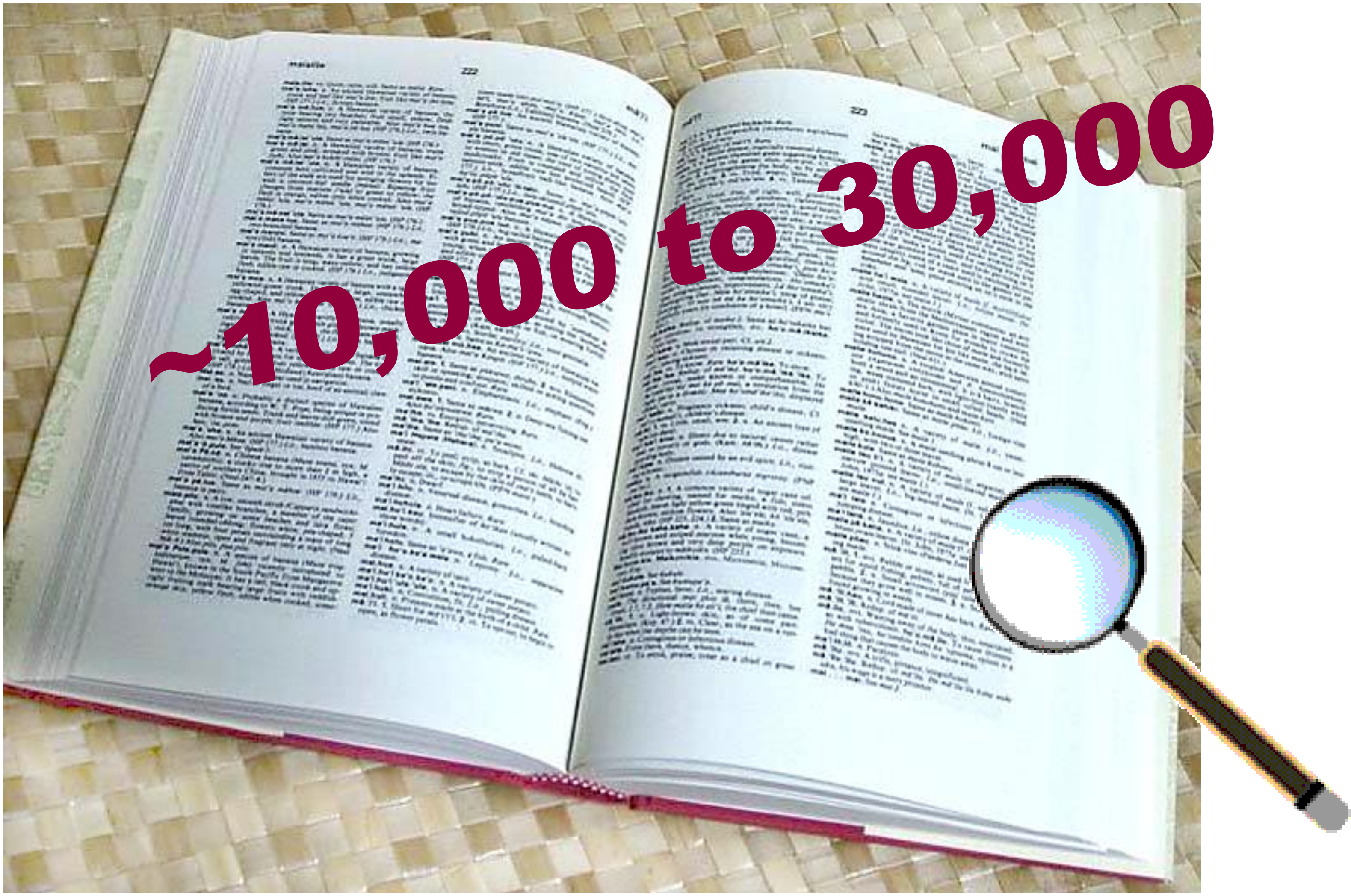
Plato said…

- 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

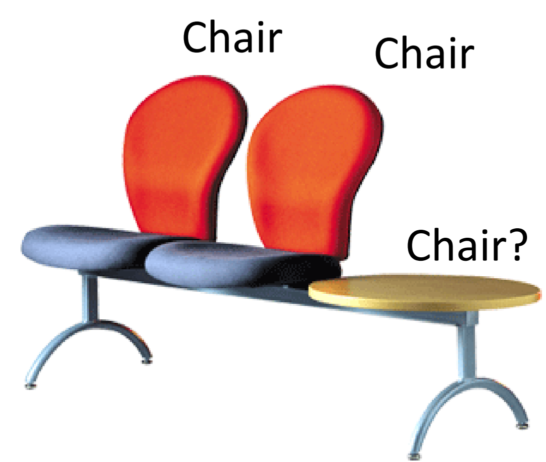
Flat surface
Horizontal
Knee-high
…

→ 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, …)

Direct perception

Some aspects of an object function can be perceived directly

- Observer relativity: Function is observer dependent

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.

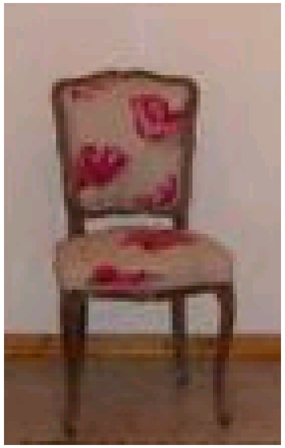
The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…

How do we achieve Mediated perception?

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

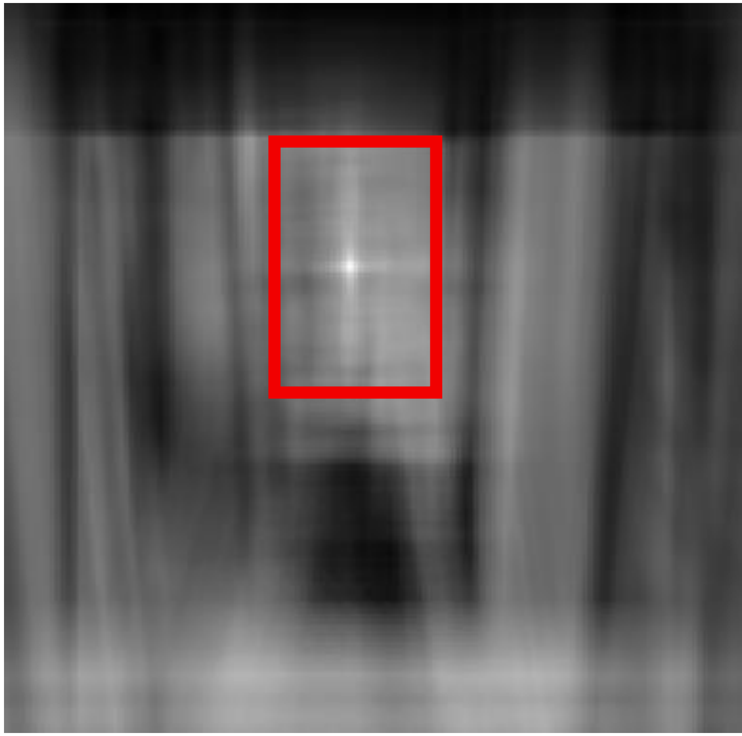
Object recognition
Is it really so hard?

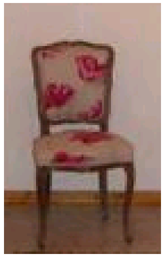
This is a chair



Find the chair in this image

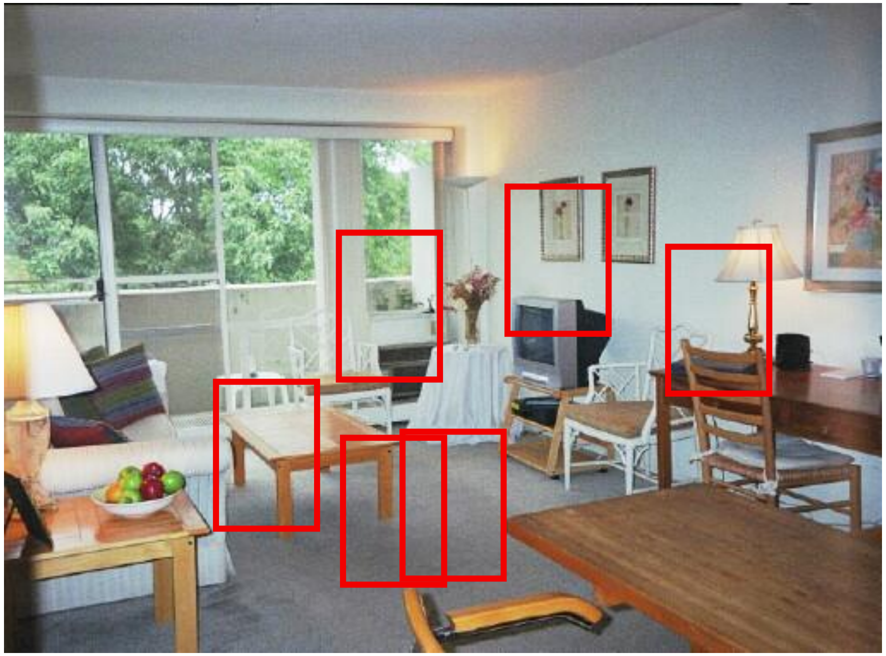
Output of normalized correlation

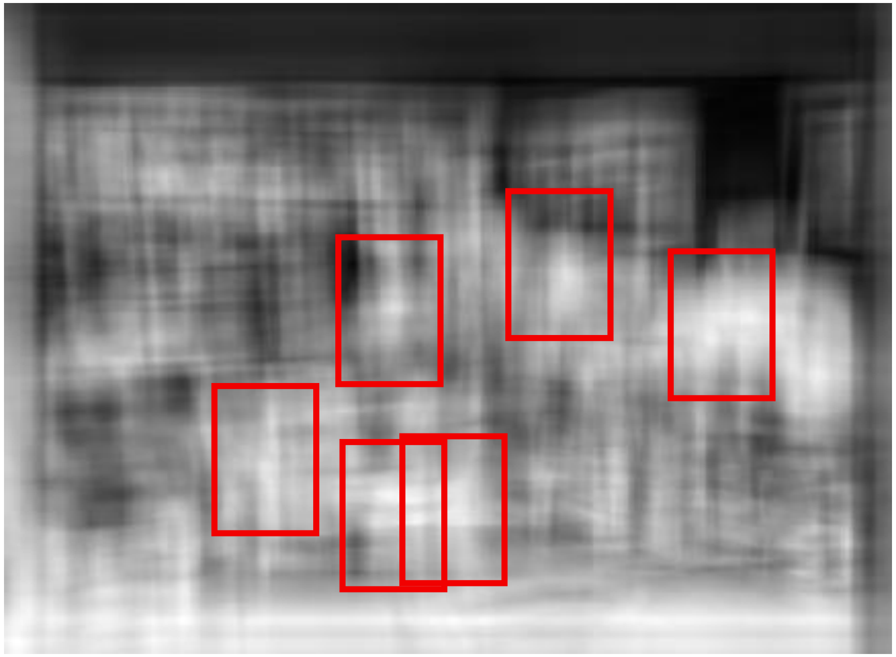




Object recognition
Is it really so hard?

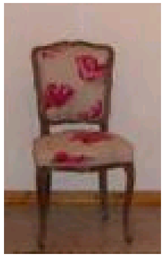
Find the chair in this image





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

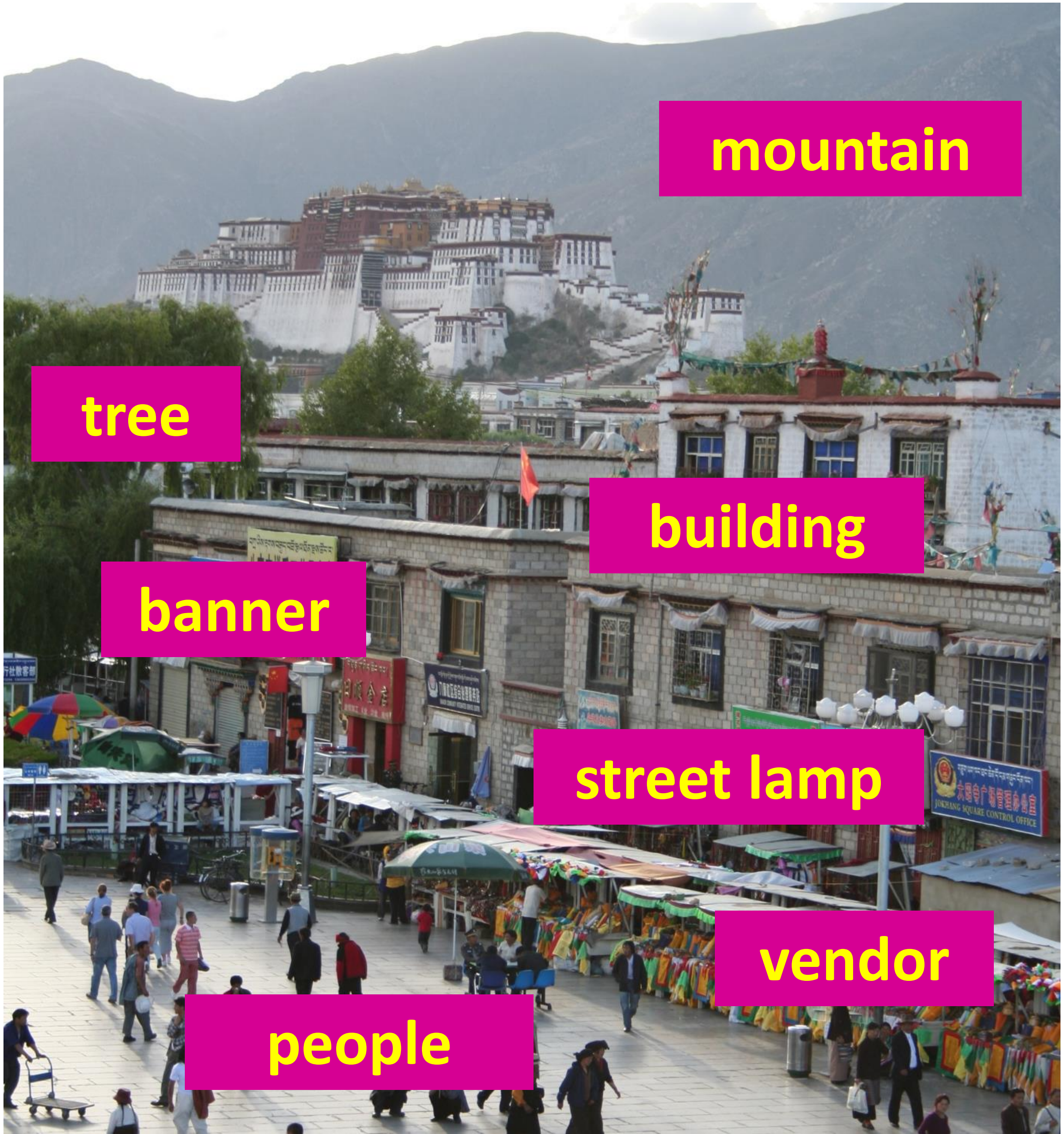
So what does object recognition involve?

Verification: is that a lamp?

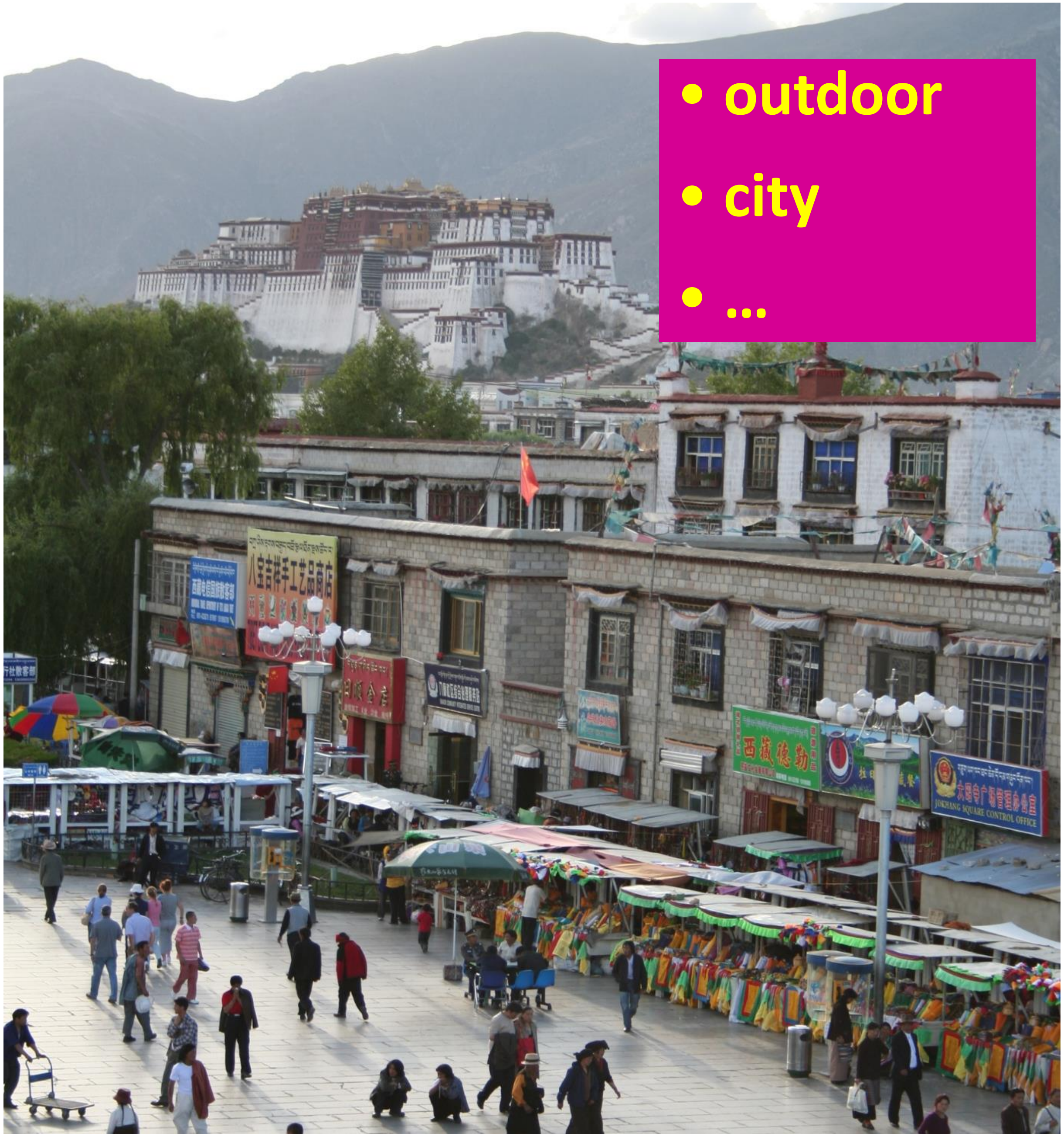
Detection: are there people?

Identification: is that Potala Palace?

Object categorization



Scene and context categorization



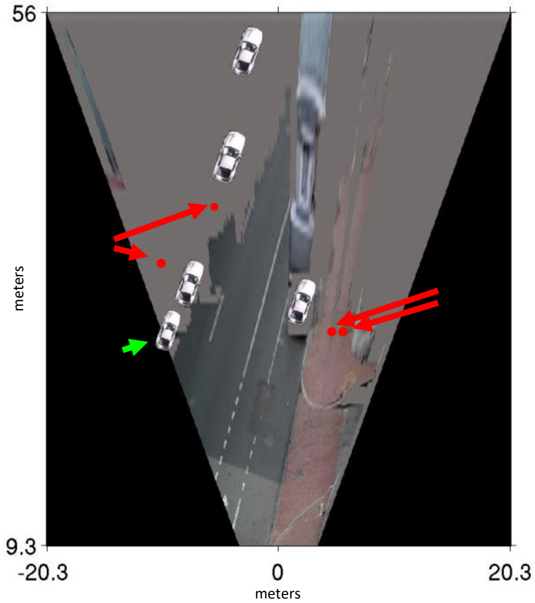
- **outdoor**
- **city**
- **…**

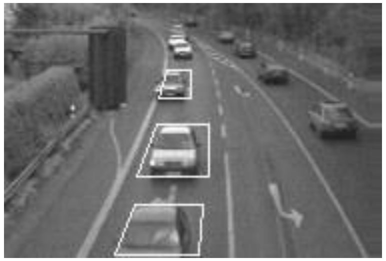
Computational photography

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

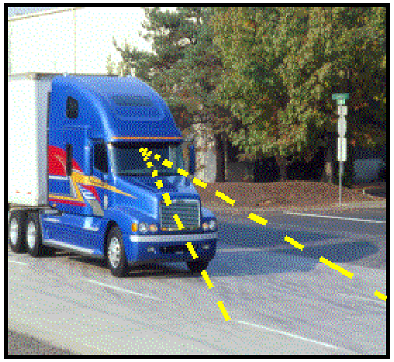
Assisted driving

Pedestrian and car detection





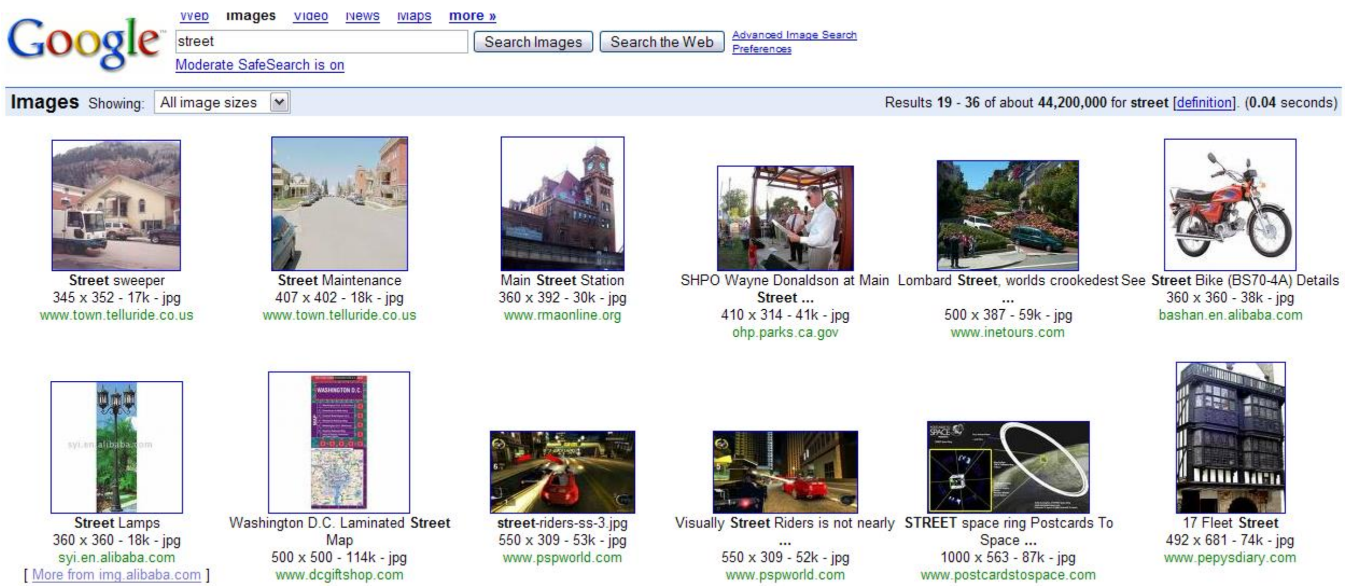
Lane detection



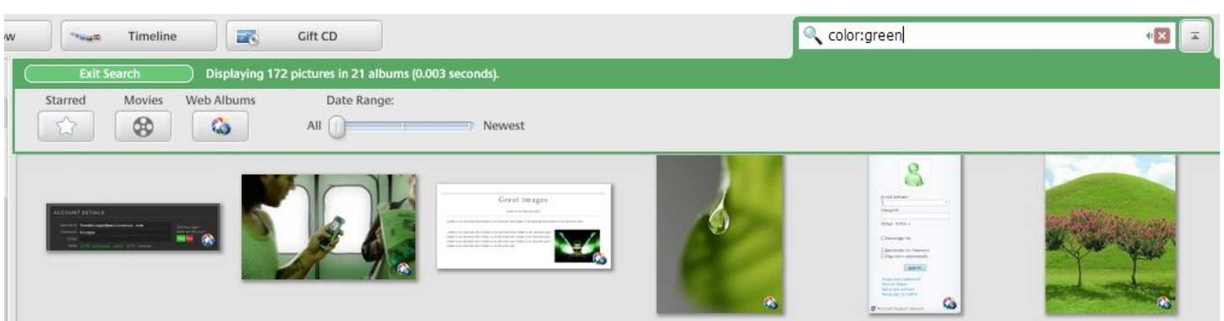
- Collision warning systems with adaptive cruise control,
- Lane departure warning systems,
- Rear object detection systems,

Improving online search

Query:
STREET



Organizing photo collections



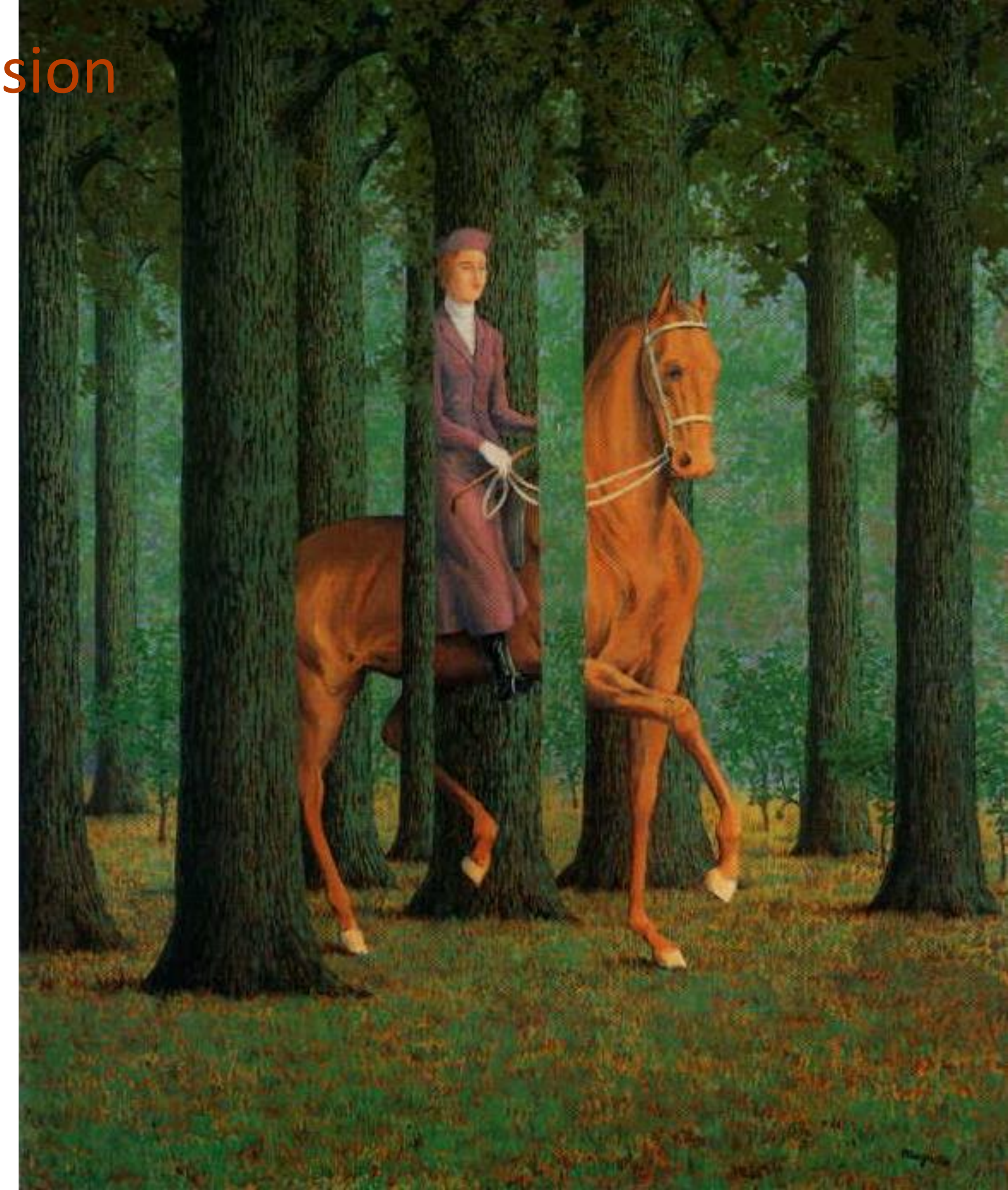
Challenges 1: view point variation

Michelangelo 1475-1564

Challenges 2: illumination

slide credit: S. Ullman

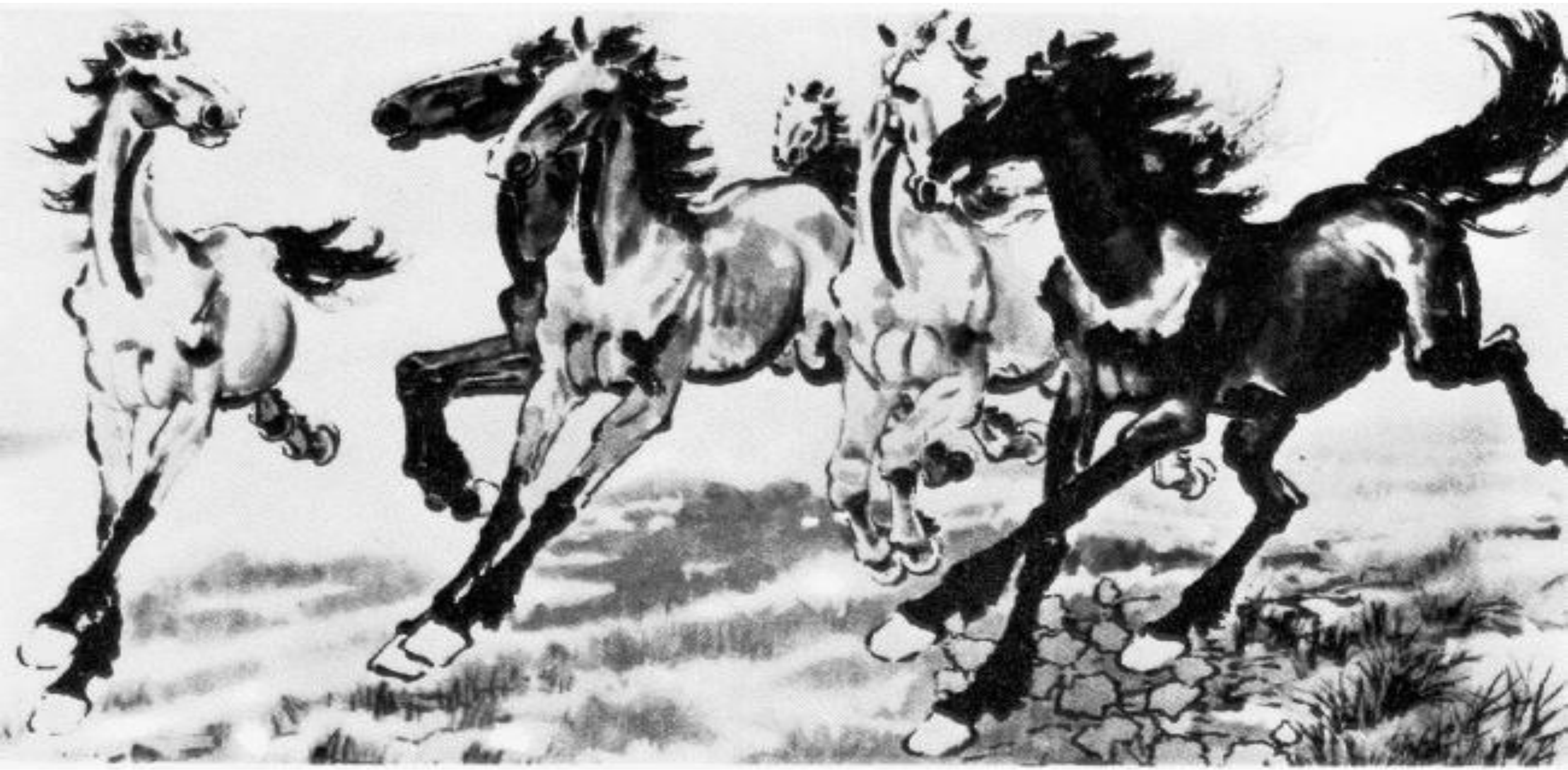
Challenges 3: occlusion



Magritte, 1957

Challenges 4: scale

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter

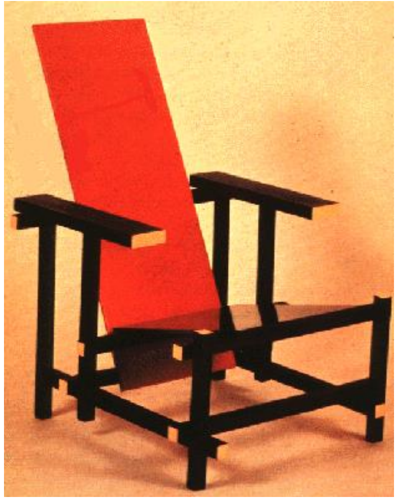
Klimt, 1913


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
Cheerios
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:

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

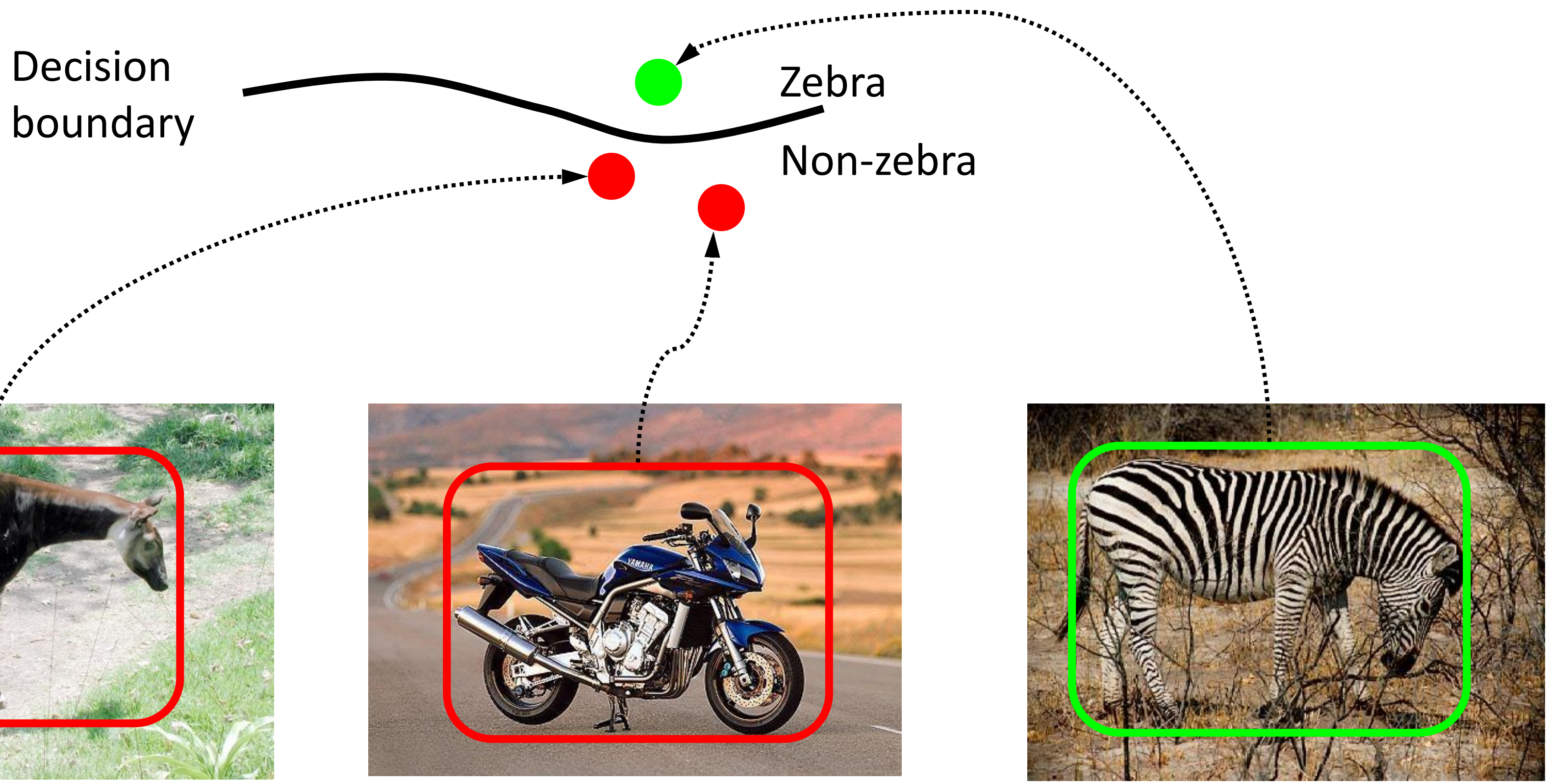
Object categorization: the statistical viewpoint

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

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

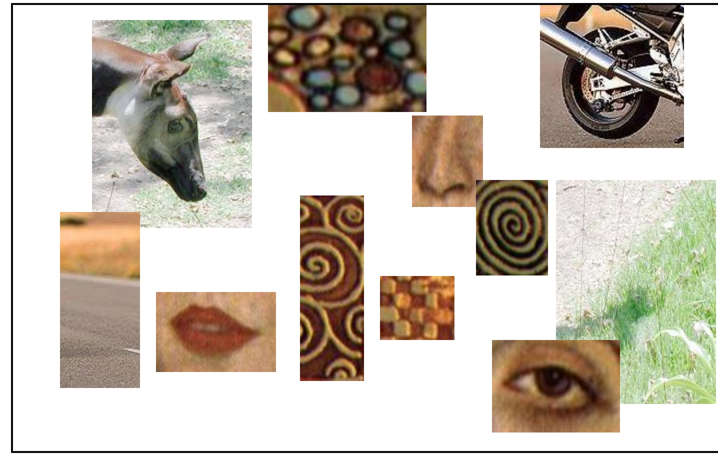
Discriminative

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



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$



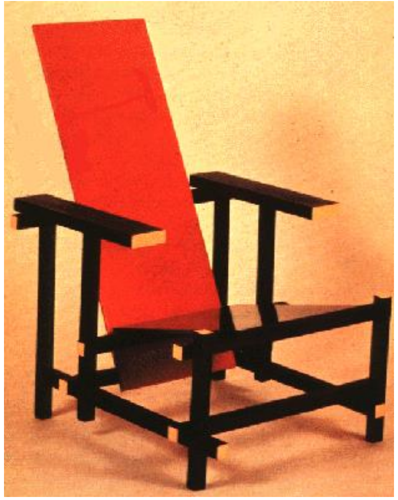
	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

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

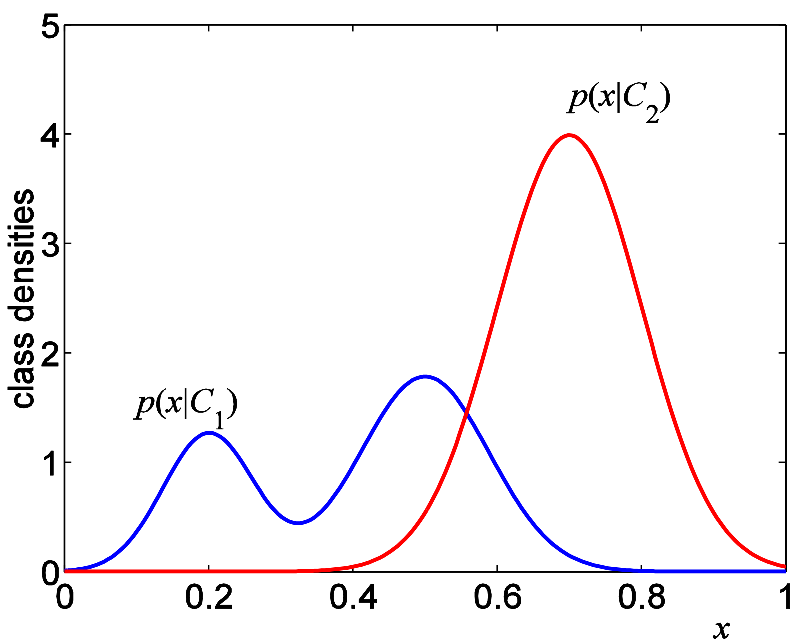
– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

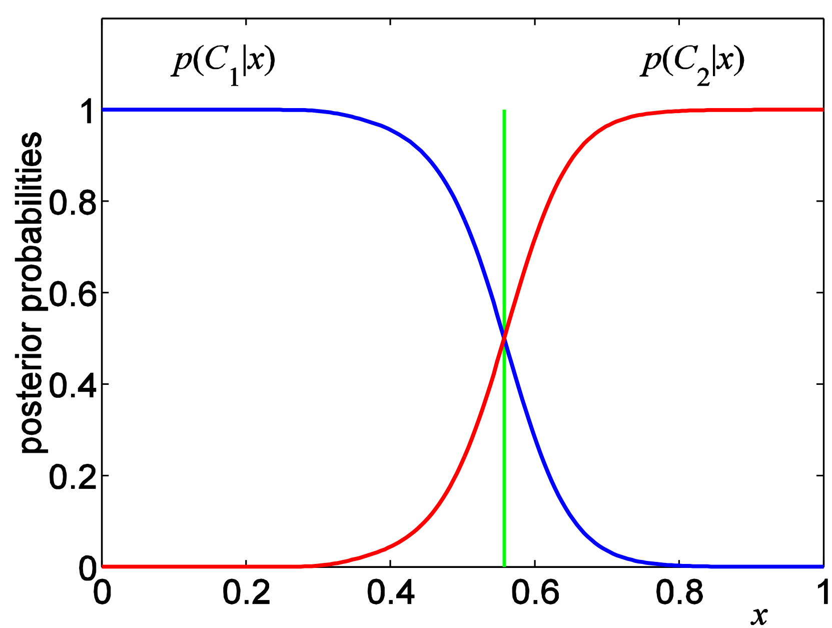


Learning

– 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

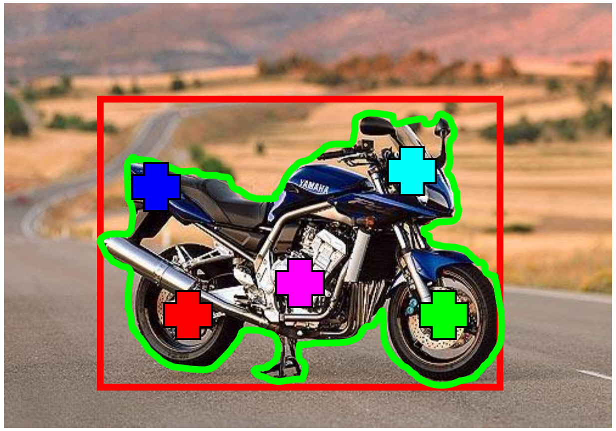




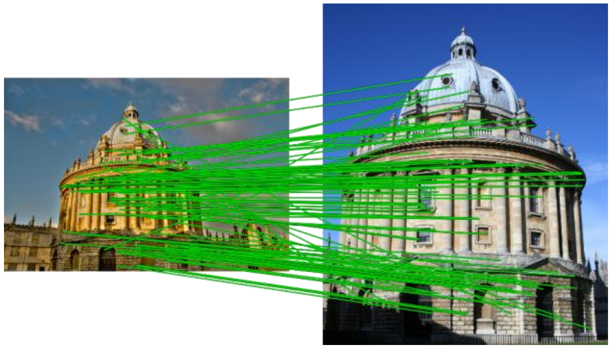
Learning

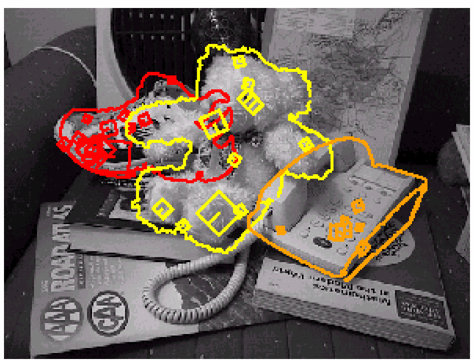
- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation;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

Contains a motorbike

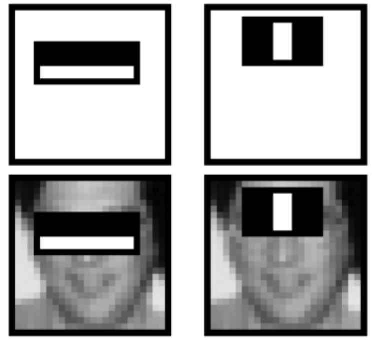


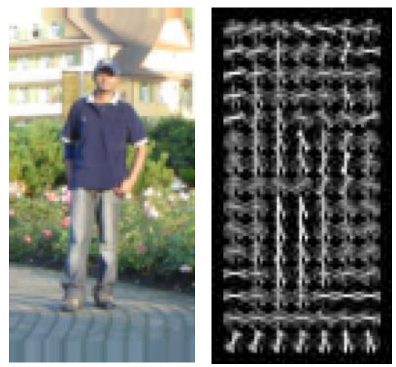
Recognition models



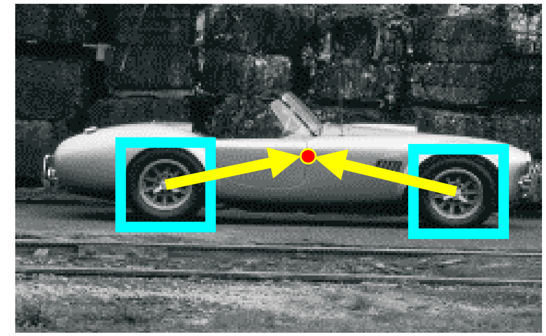


Instances:
recognition by
alignment





Categories:
Holistic appearance
models (and sliding
window detection)

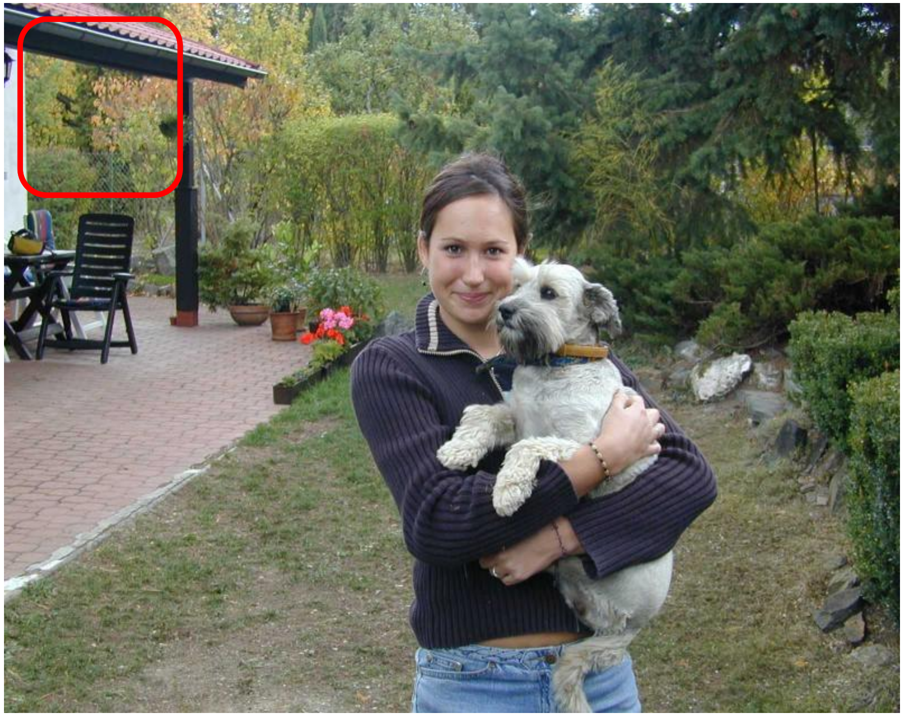


Categories:
Local feature and
part-based models

Kristen Grauman

Recognition

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



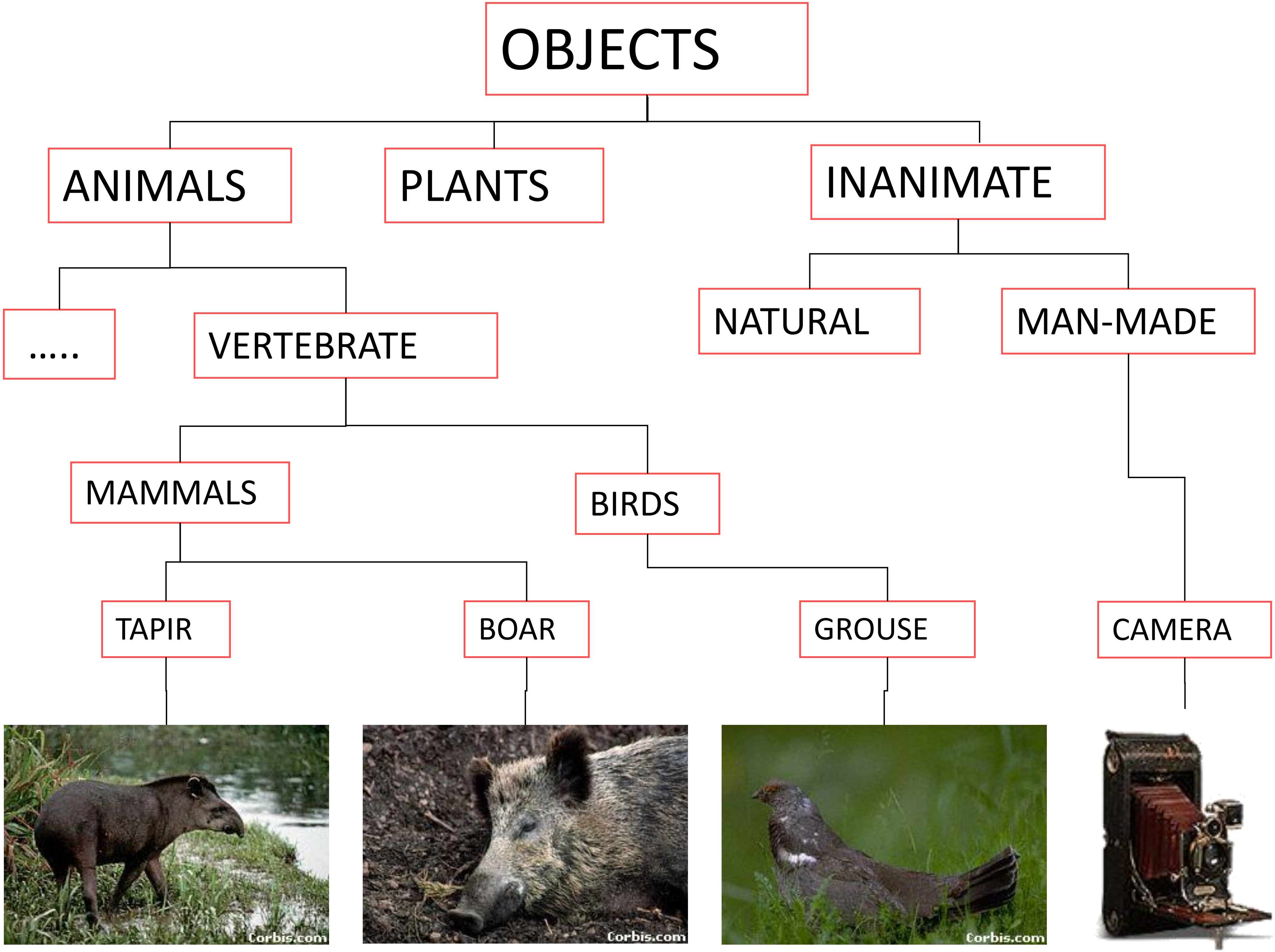
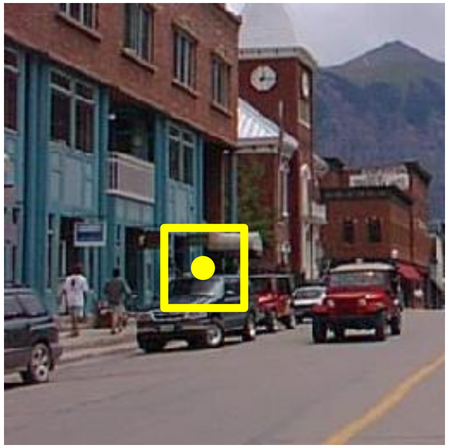

OBJECTS
ANIMALS
PLANTS
INANIMATE
.....
VERTEBRATE
NATURAL
MAN-MADE
MAMMALS
BIRDS
TAPIR
BOAR
GROUSE
CAMERA
Corbis.com
Corbis.com
Corbis.com

Image features



Pixel or local patch

Segmentation region

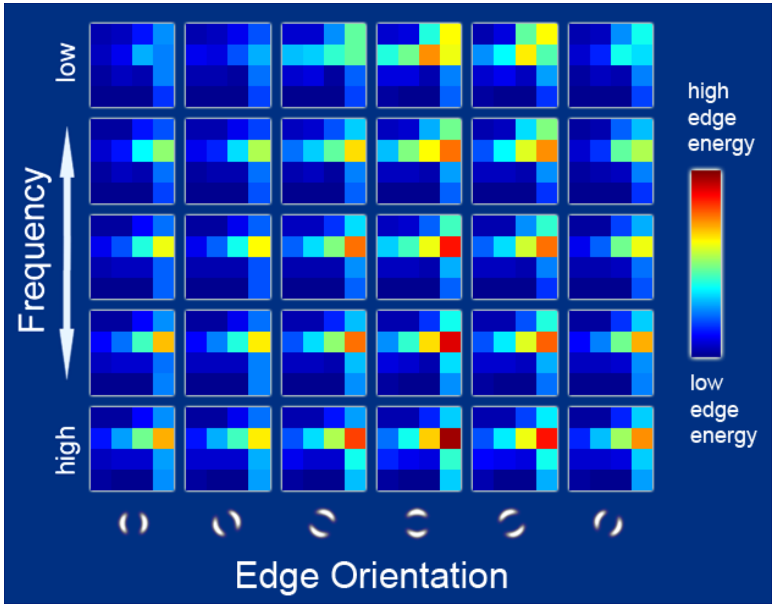
Bounding box

Whole image

Slide credit: Svetlana Lazebnik

GIST features

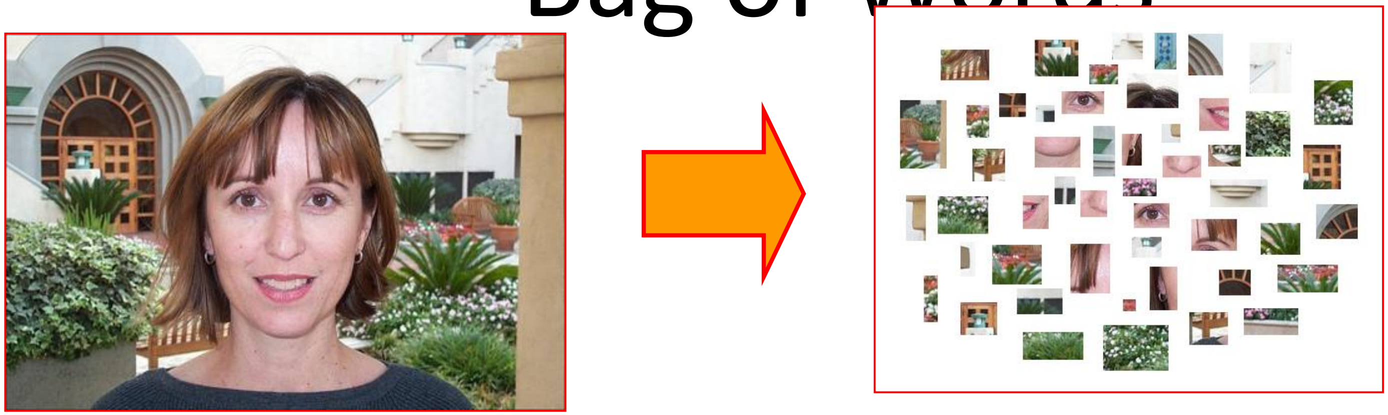
- Oliva & Torralba (2001)

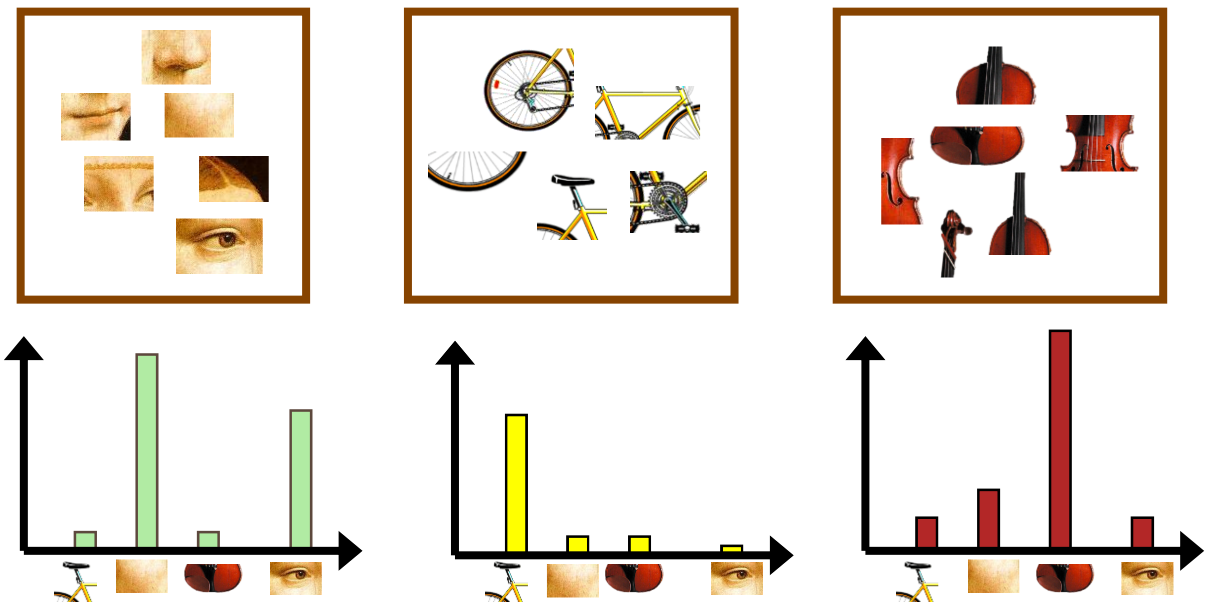


Spatial envelope
naturalness, openness, roughness, expansion, ruggedness

Slide credit: Svetlana Lazebnik

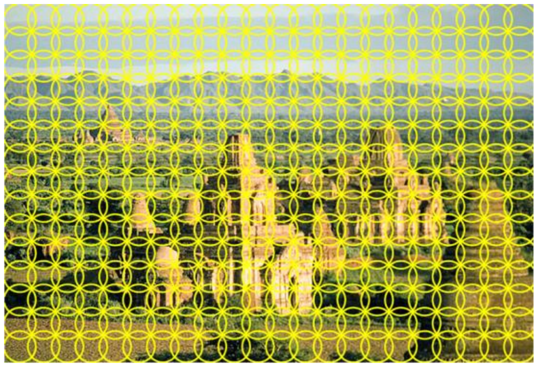
Bag of Words

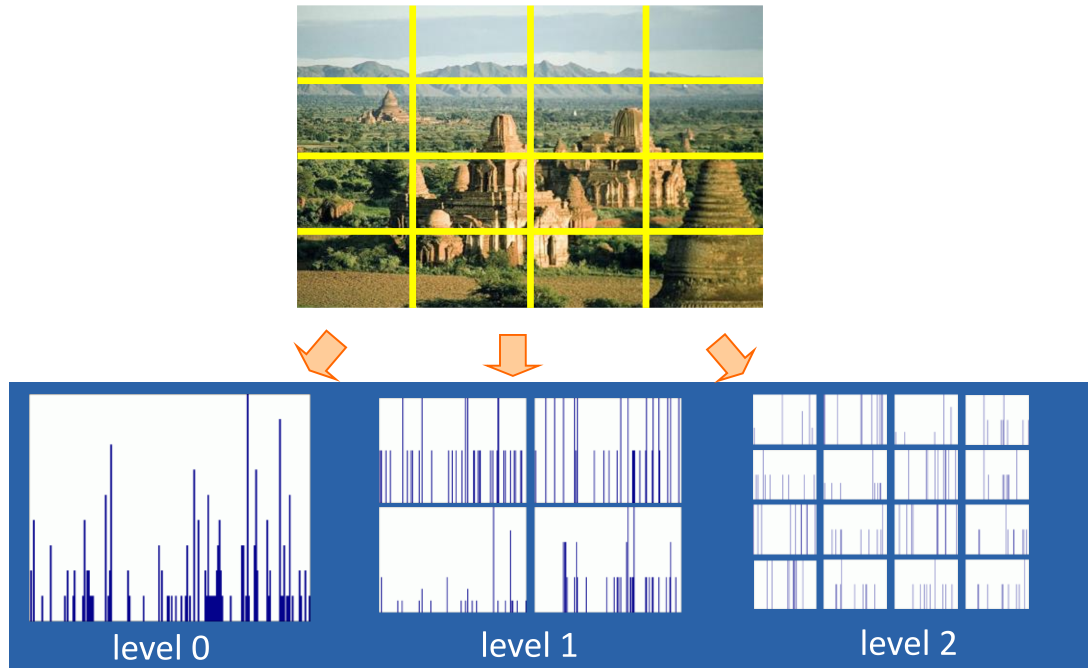




Slide credit: Svetlana Lazebnik

Local Feature Extraction

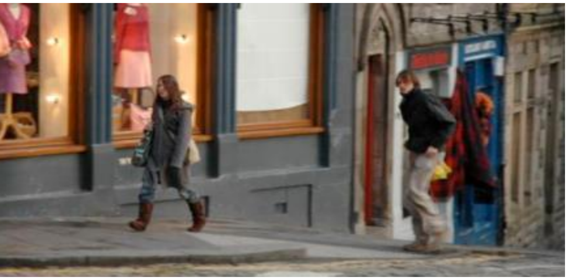




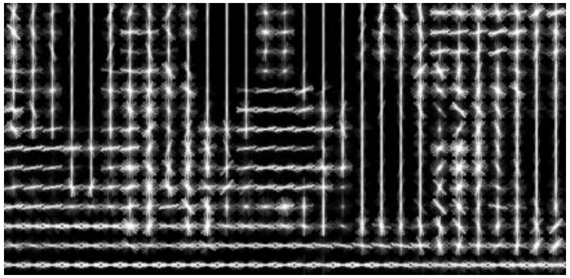
Lazebnik, Schmid & Ponce (CVPR 2006)

Slide credit: Svetlana Lazebnik

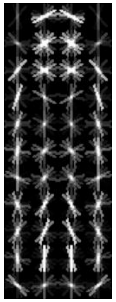
Histogram of Oriented Gradients
Part based models



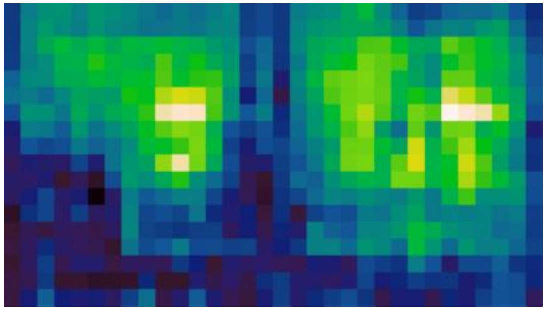
HOG feature map



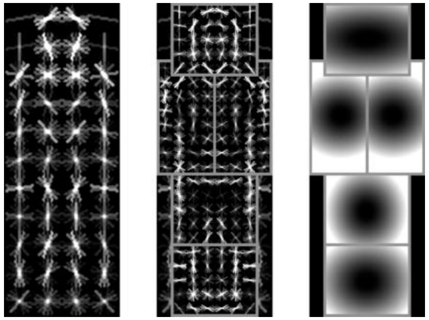
Template

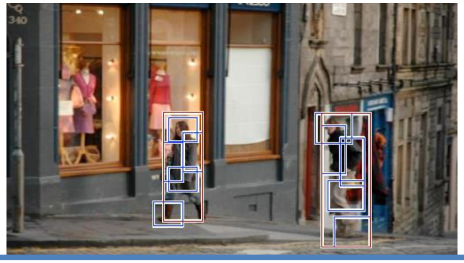


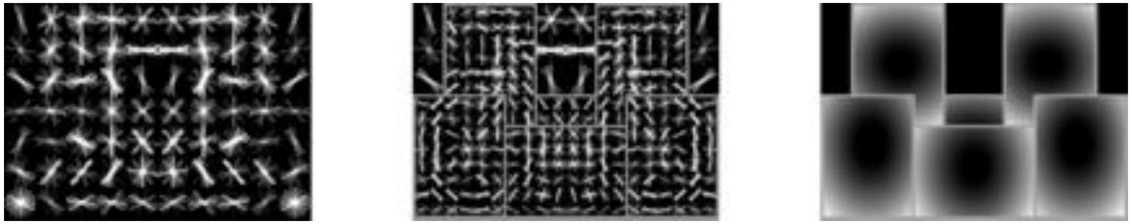
Detector response map

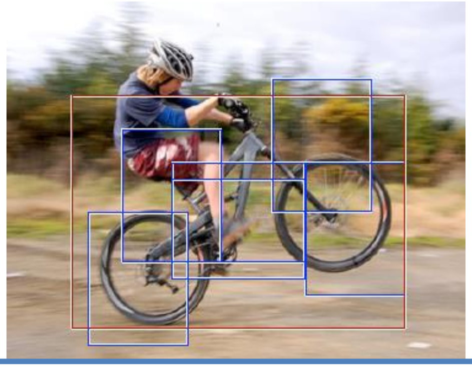


N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005









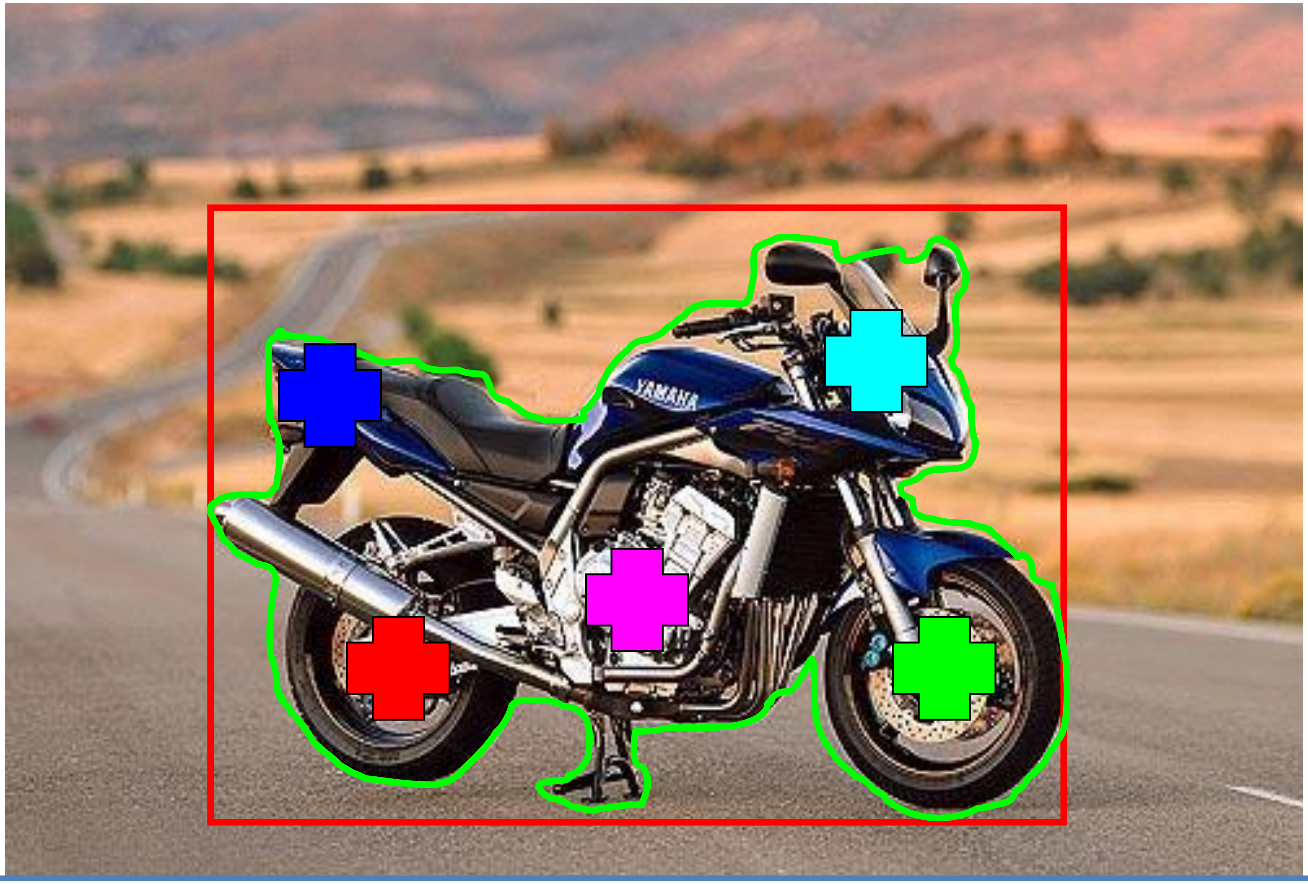
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010

Slide credit: Svetlana Lazebnik

Labeling required for supervision

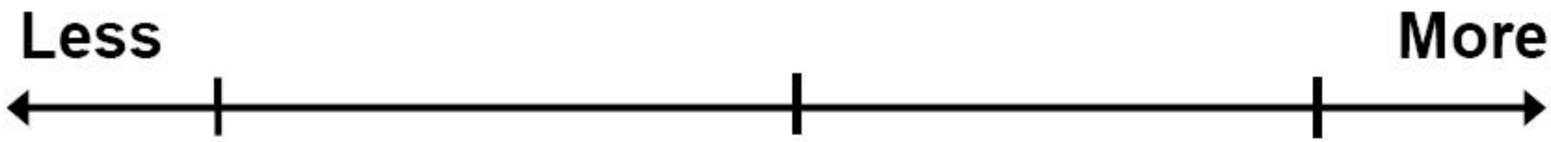
Images in the training set must be annotated with the “correct answer” that the model is expected to produce

Contains a motorbike



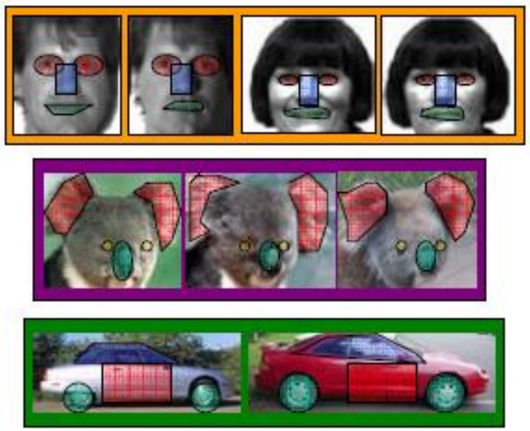
Slide credit: Svetlana Lazebnik

Spectrum of supervision



Unsupervised

"Weakly" supervised

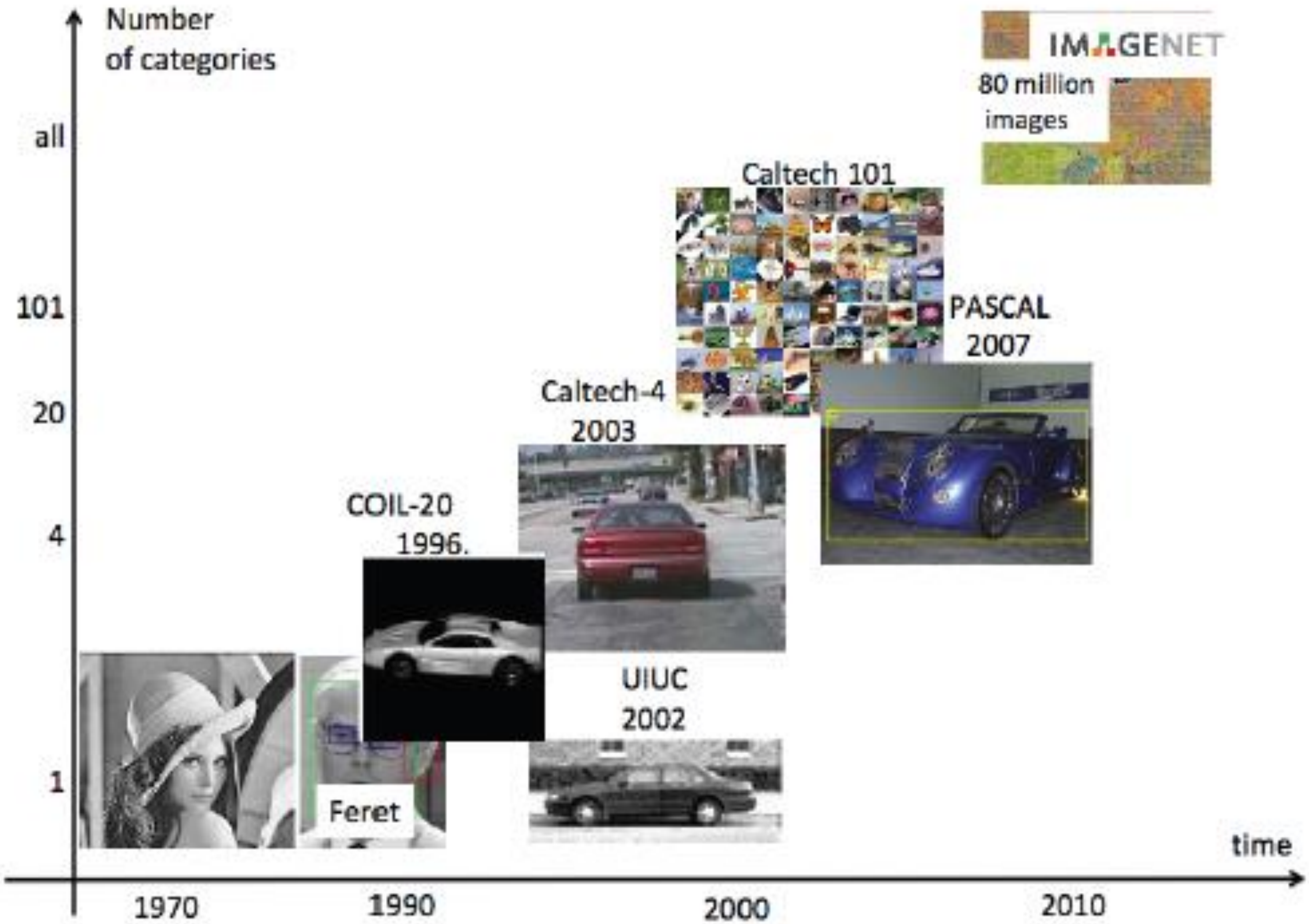


Fully supervised

Definition depends on task

Slide credit: Svetlana Lazebnik

Available datasets

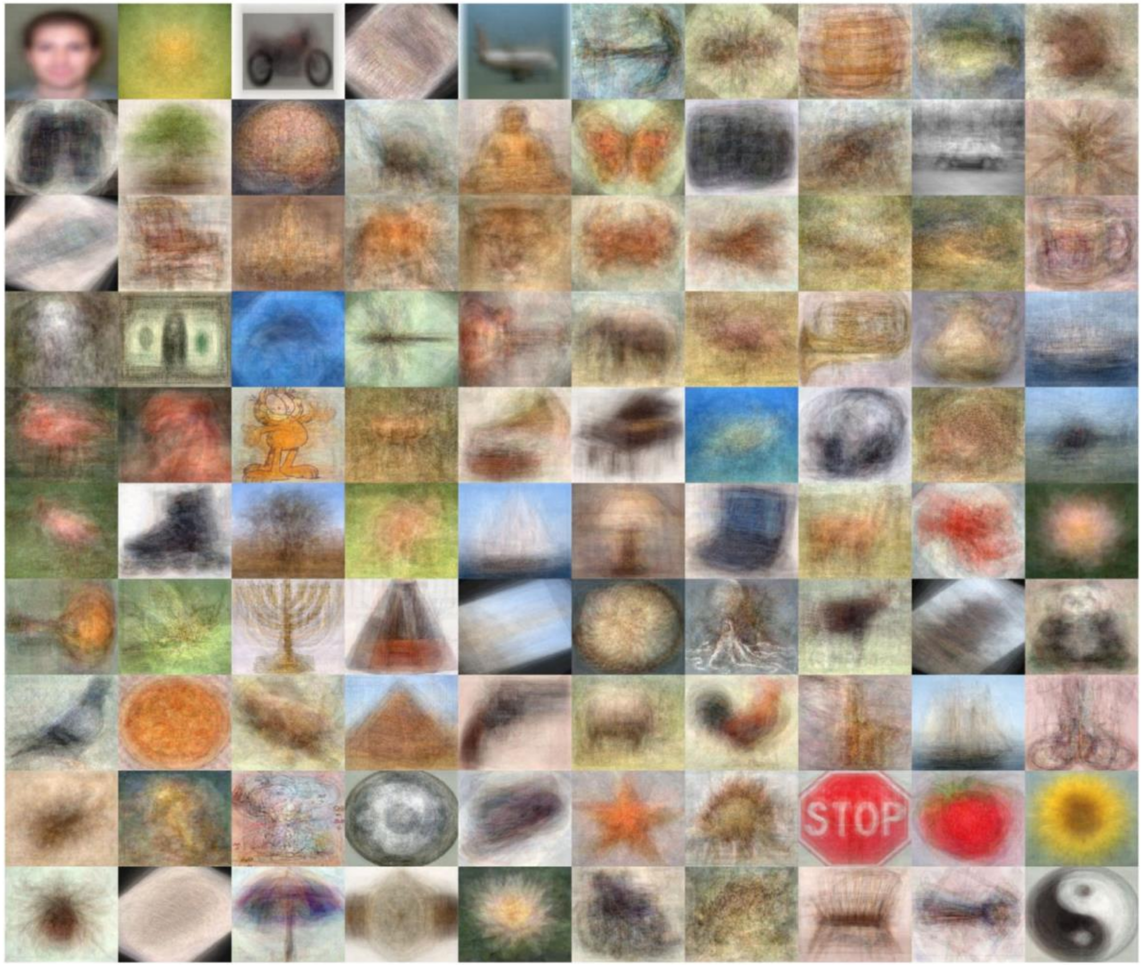


From "The Promise and Perils of Benchmark Datasets and Challenges", D. Forsyth, A. Efros, F.-F. Li, A. Torralba and A. Zisserman, Talk at "Frontiers of Computer Vision"

Caltech 101 and 256

Caltech-101: Intra-class variability

Fei-Fei, Fergus, Perona, 2004



Griffin, Holub, Perona, 2007

Slide credit: Svetlana Lazebnik

The PASCAL Visual Object Classes Challenge (2005-2012)

- **Challenge classes:**
  Person: person
  Animal: bird, cat, cow, dog, horse, sheep
  Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
  Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

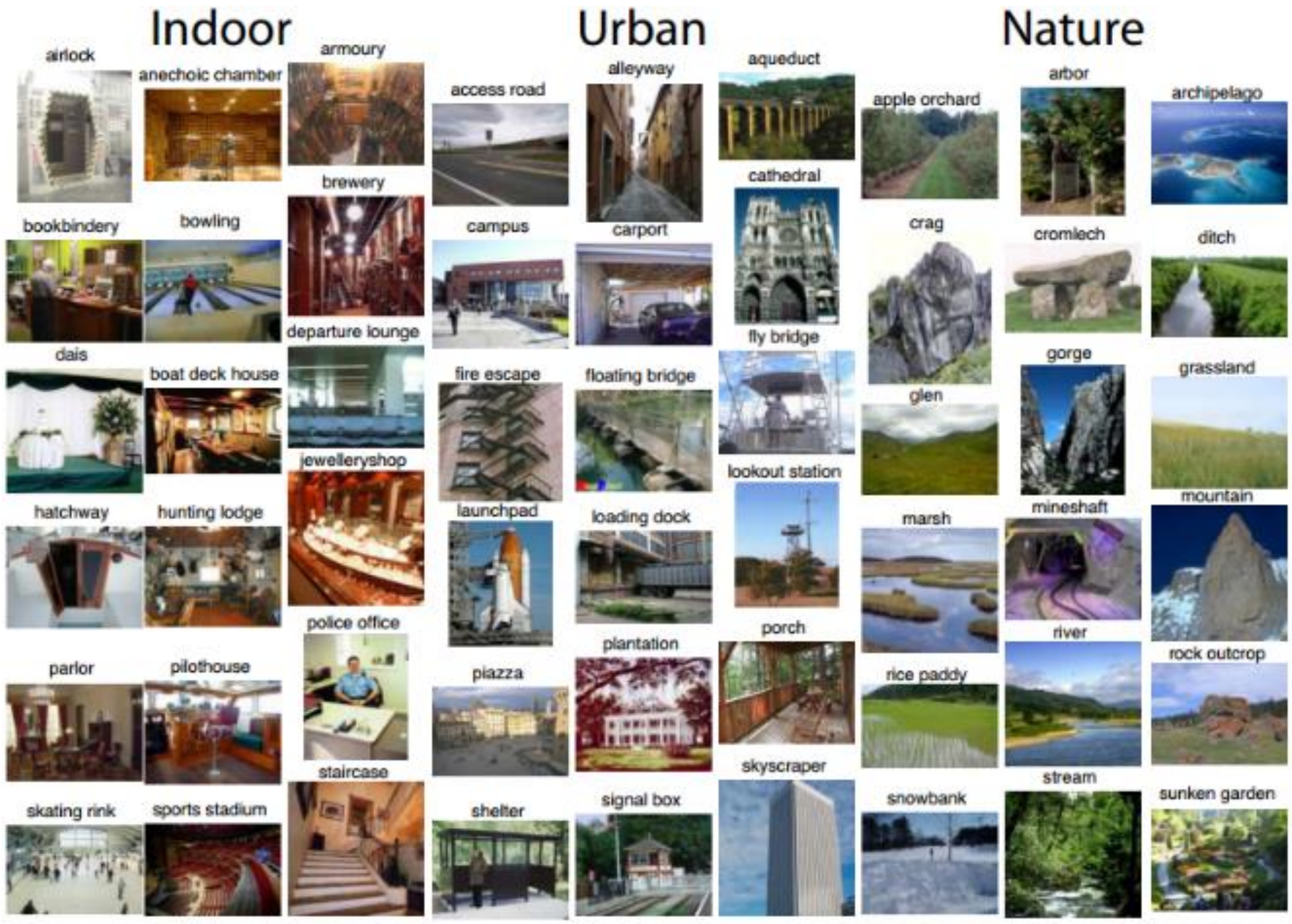
- **Dataset size (by 2012):**
  11.5K training/validation images, 27K bounding boxes, 7K segmentations

- Classification, detection, segmentation, person layout

Slide credit: Svetlana Lazebnik

Sun Dataset

~900 scene categories (~400 well-sampled), 130K images

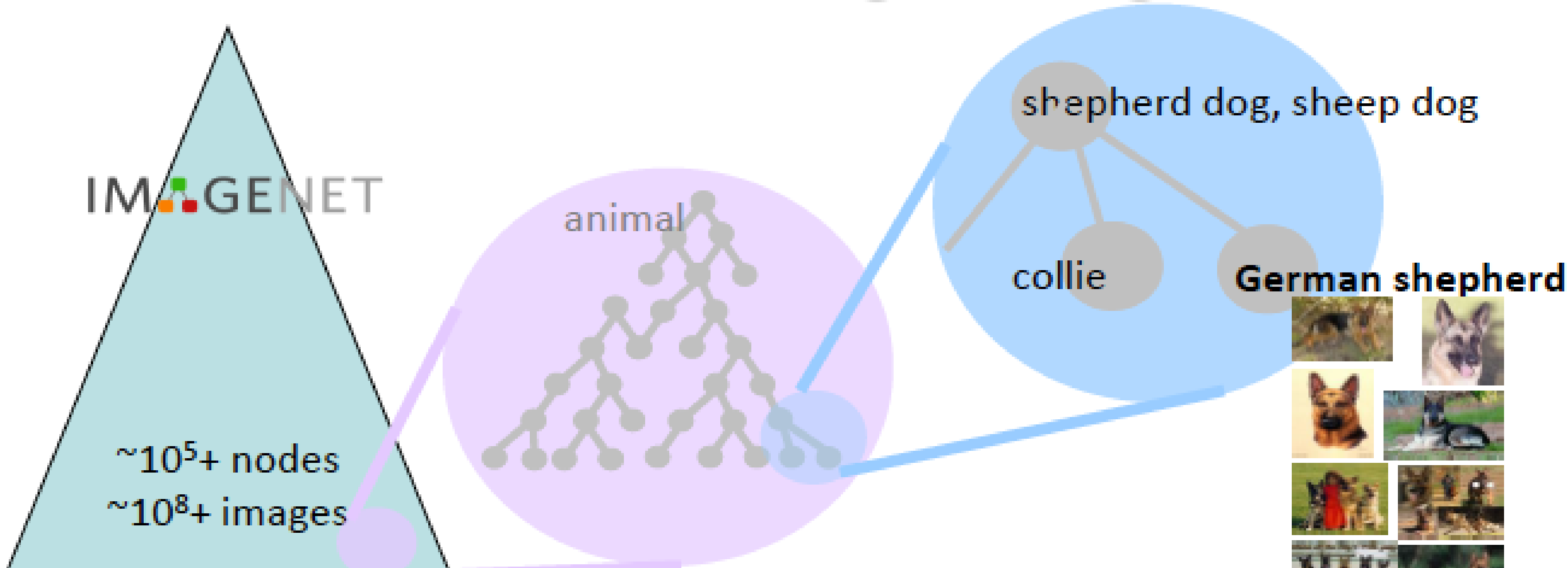


J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo," CVPR 2010

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**



Slide credit: Fei-fei Li

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

MS COCO

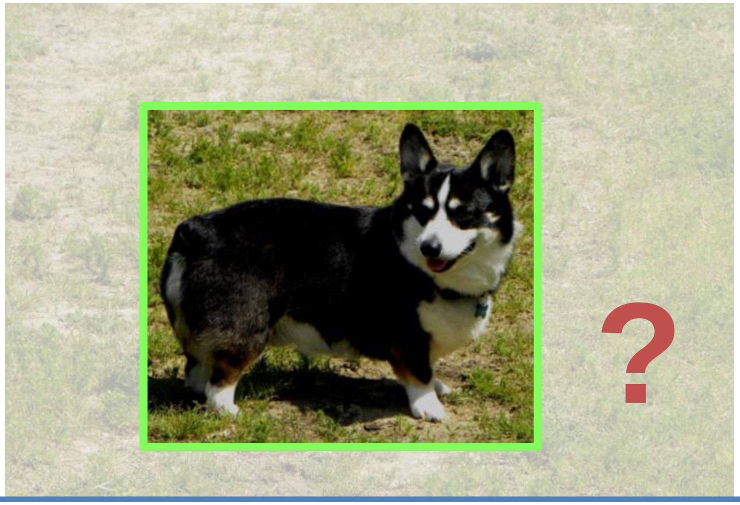
Over 77,000 worker hours (8+ years)

- 70-100 object categories (things not stuff)
- 330,000 images (~150k first release)
- 2 million instances (400k people)
- Every instance is segmented
- 7.7 instances per image (3.5 categories)
- Key points
- 5 sentences per image

http://mscoco.org

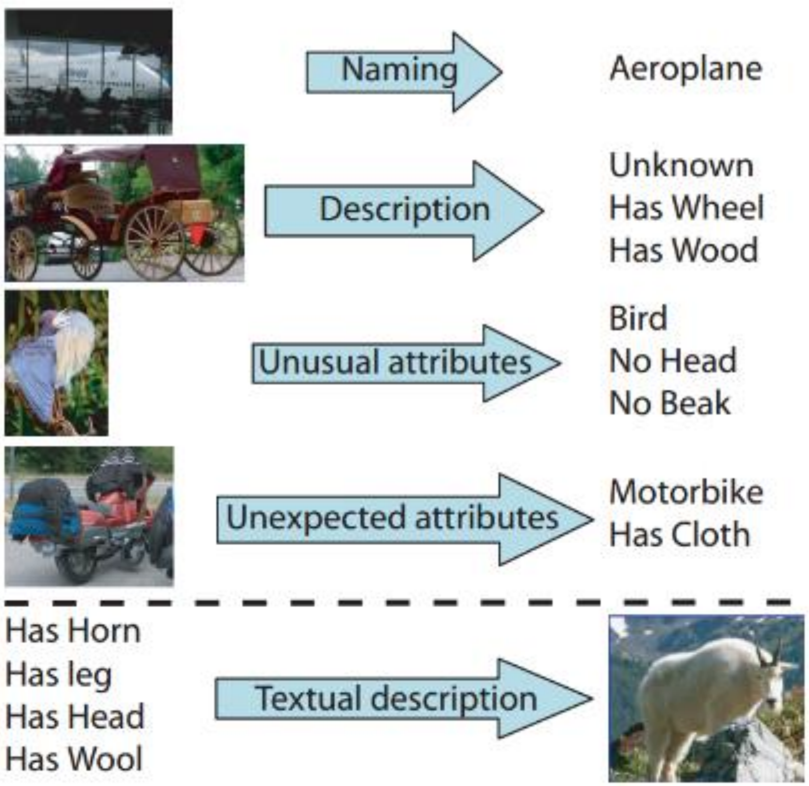
Fine grained recognition

What breed is this dog?

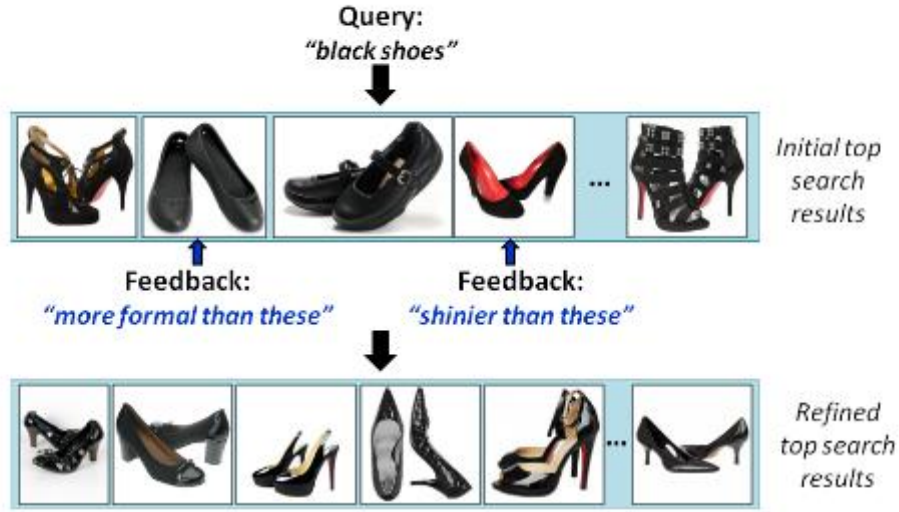


Slide credit: Svetlana Lazebnik

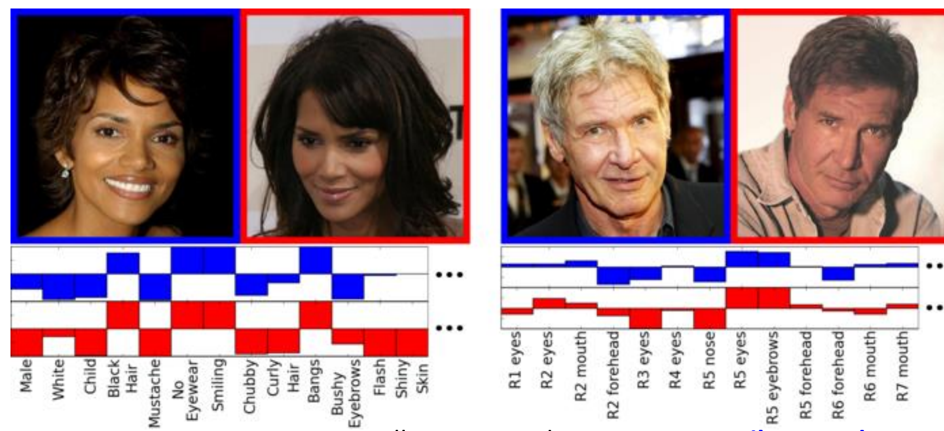
Attribute based recognition



A. Farhadi, I. Endres, D. Hoiem, and D Forsyth, [Describing Objects by their Attributes](), CVPR 2009



A. Kovashka, D. Parikh and K. Grauman, **[WhittleSearch: Image Search with Relative Attribute Feedback]()**, CVPR 2012



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, **[Attribute and Simile Classifiers for Face Verification]()**, ICCV 2009

Slide credit: Svetlana Lazebnik