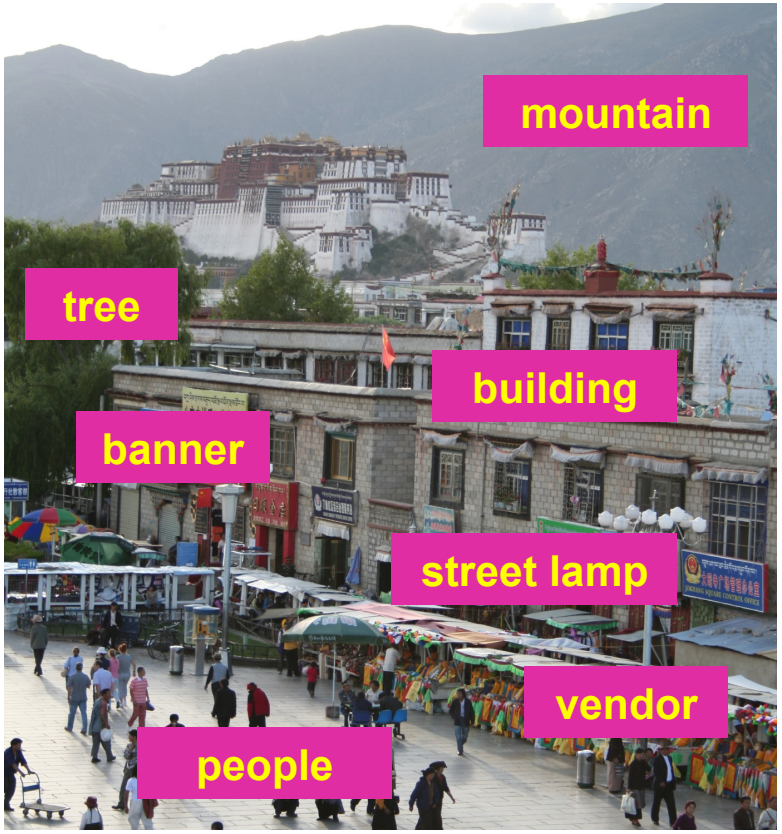
CS4670/5670: Computer Vision

Kavita Bala

Lecture 26: Recognition



Announcements

- PA 4 out. Due Apr 22.

Recognition: Overview and History

Slides from James Hays, Lana Lazebnik, Fei-Fei Li, Rob Fergus, Antonio Torralba, and Jean Ponce

OBJECTS

- ANIMALS
  - …..
  - VERTEBRATE
    - MAMMALS
      - TAPIR
      - BOAR
    - BIRDS
      - GROUSE
- PLANTS
- INANIMATE
  - NATURAL
  - MAN-MADE
    - CAMERA

Recognition Tasks

Fei-Fei Li

Classification: Does image have X? [Y/N]

Fei-Fei Li

Scene Categorization

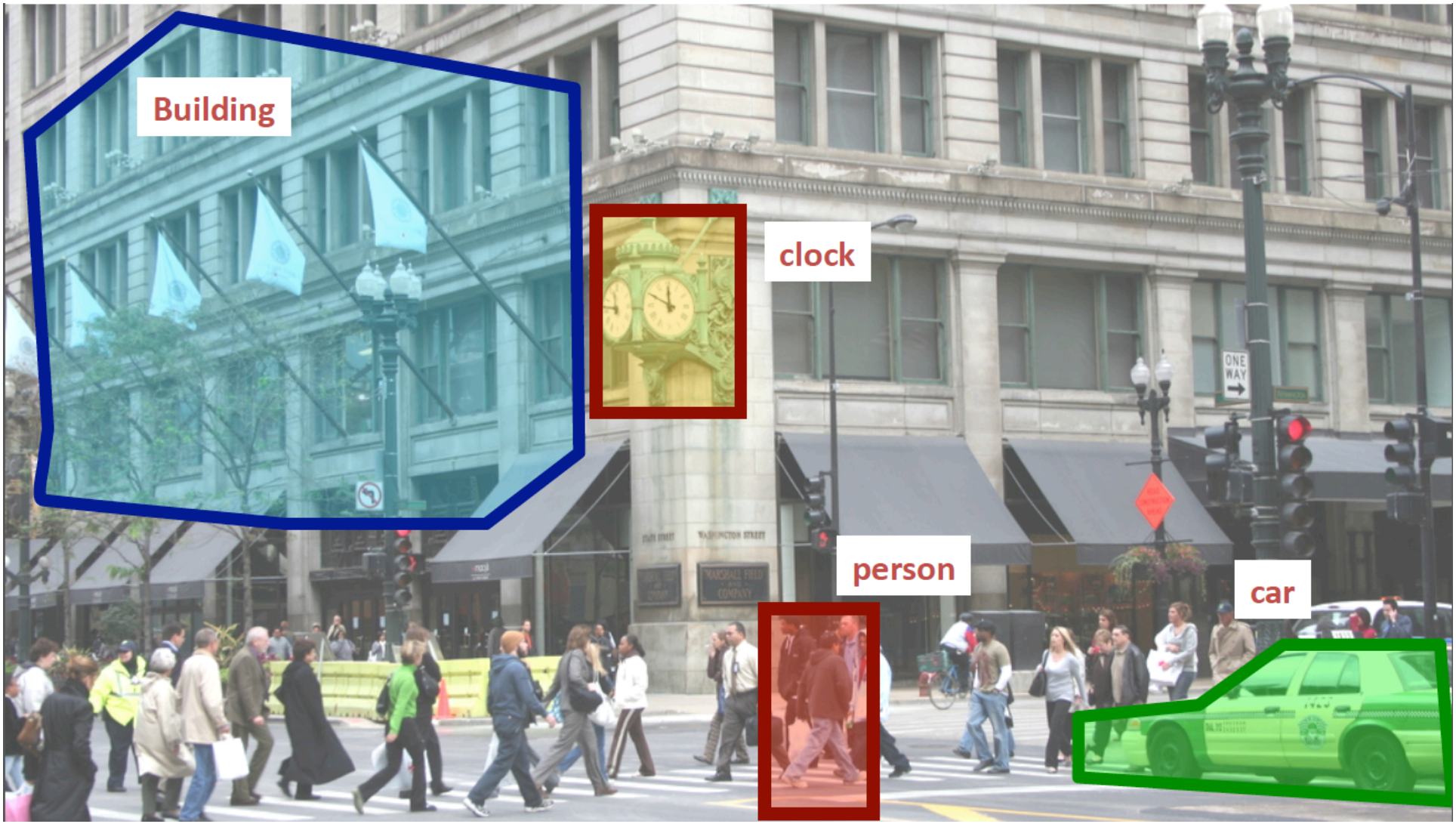
Fei-Fei Li

Object Detection:

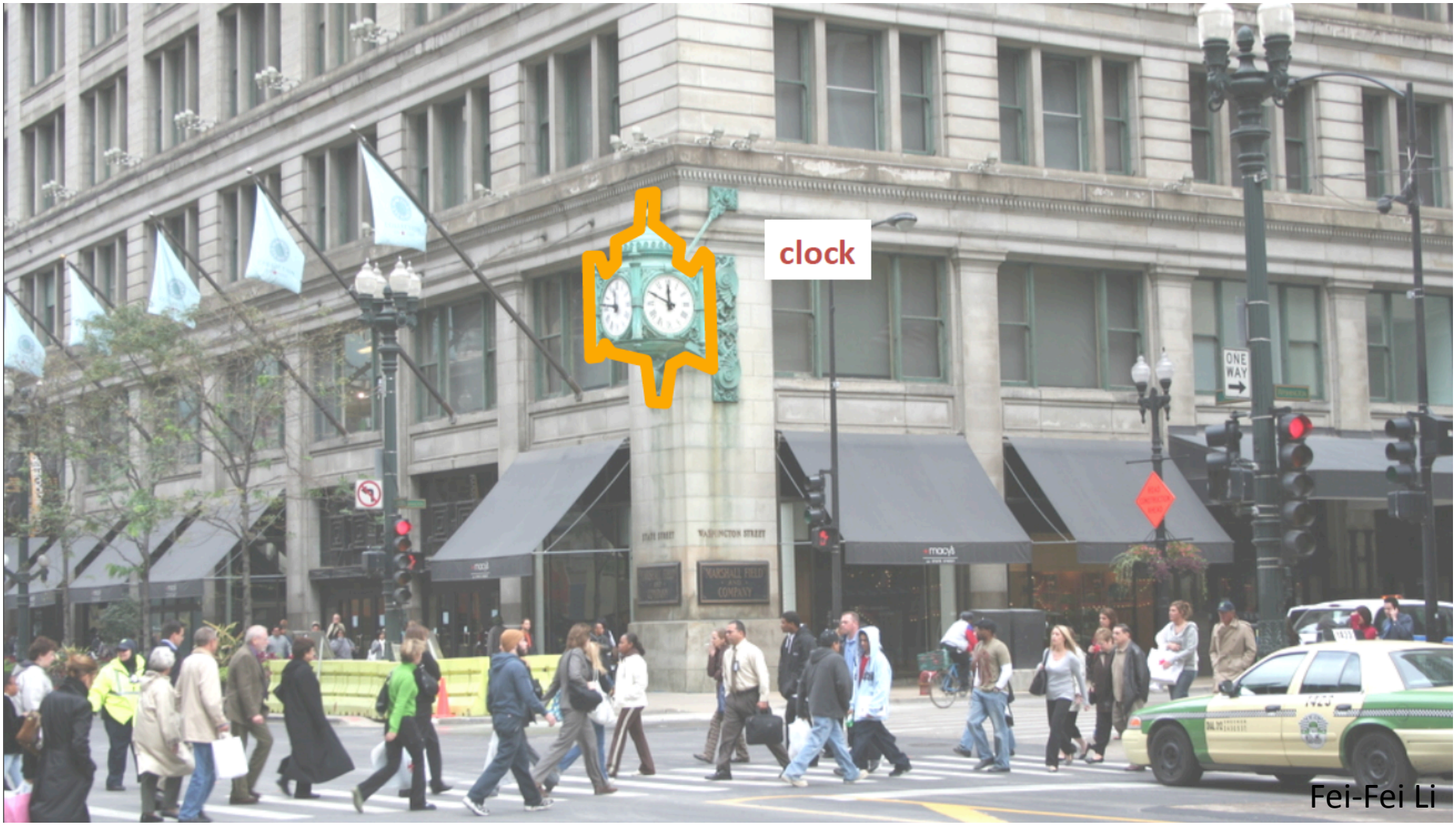
Does this image contain a car? [Y/N] where?

Object Detection:

Which objects does this image contain? where?

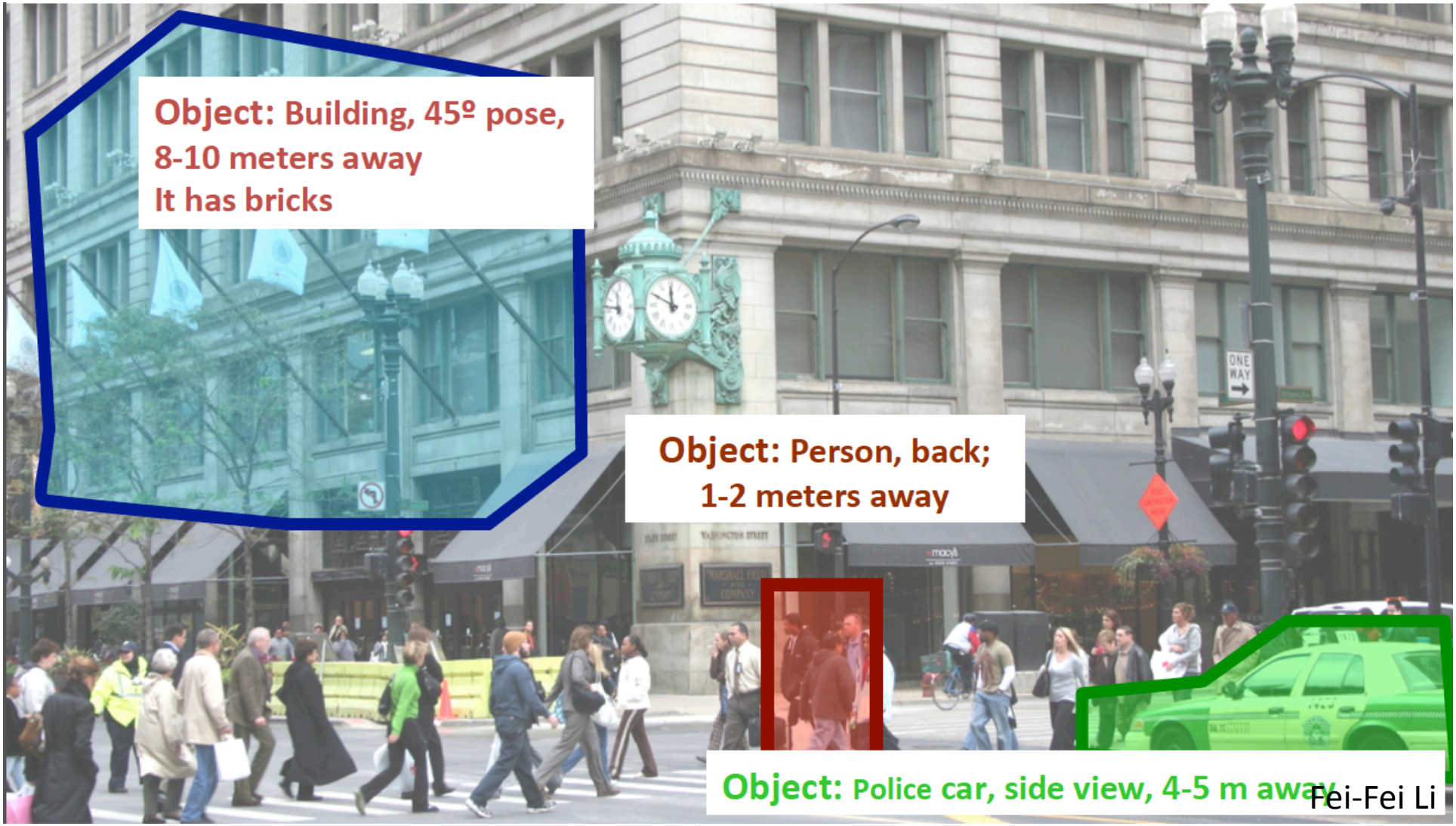


Object Detection:
Accurate localization (segmentation)



Detection:

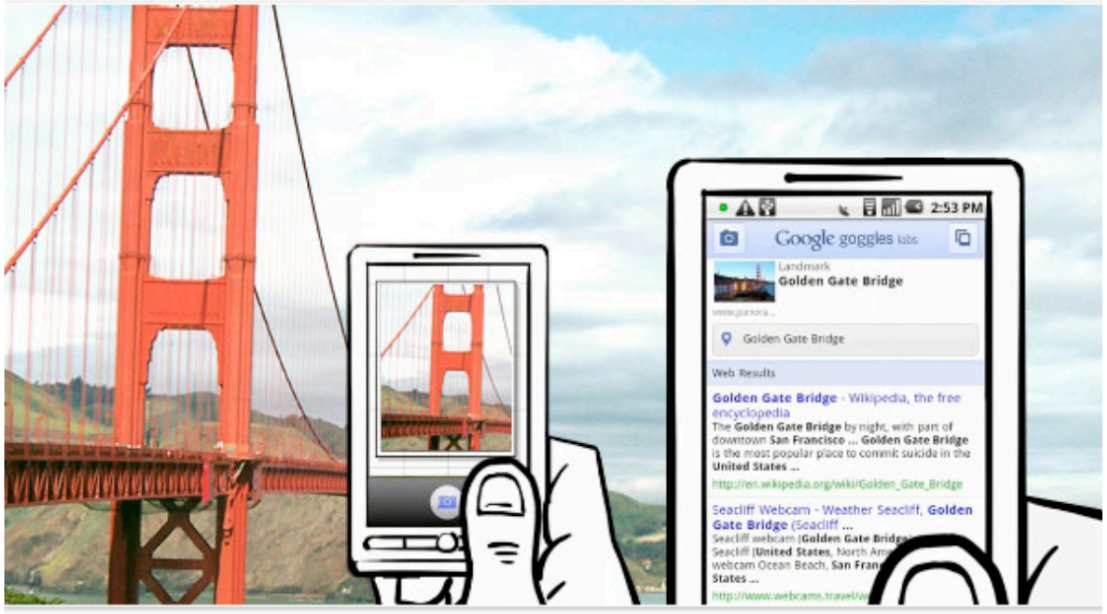
Object semantics & geometric attributes

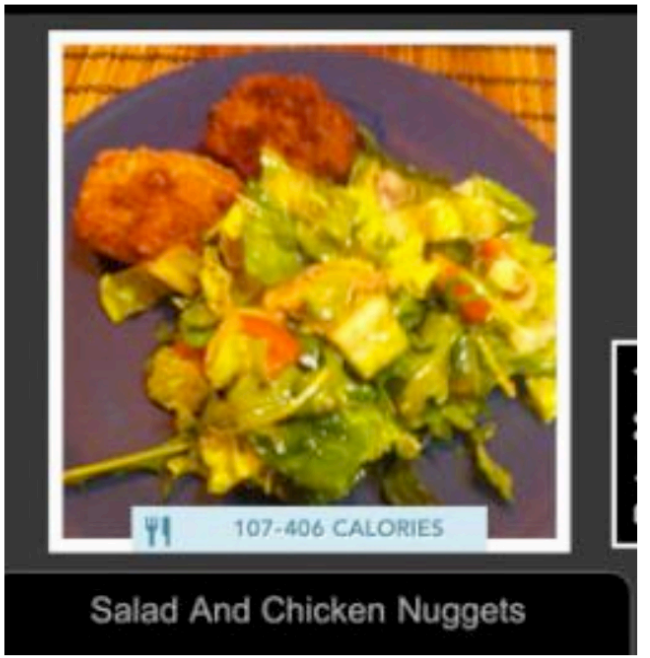


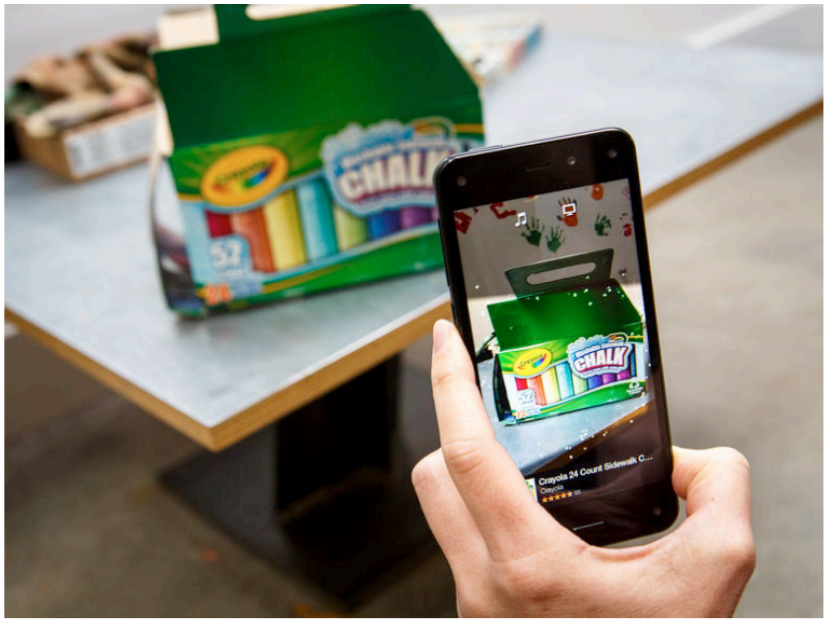
Single instance recognition

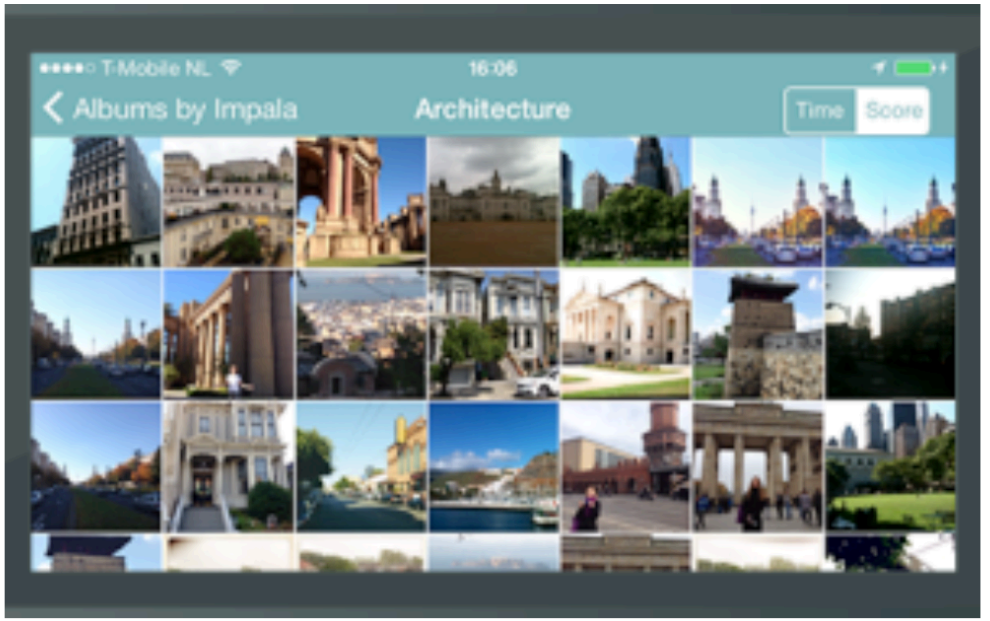
Activity/Event recognition
What are these people doing?

Fei-Fei Li


2:53 PM
Google goggles
Landmark
Golden Gate Bridge
Golden Gate Bridge
Web Results
Golden Gate Bridge - Wikipedia, the free encyclopedia
The Golden Gate Bridge by night, with part of downtown San Francisco ... Golden Gate Bridge is the most popular place to commit suicide in the United States ...
http://en.wikipedia.org/wiki/Golden_Gate_Bridge
Seacliff Webcam - Weather Seacliff, Golden Gate Bridge (Seacliff ...


107-406 CALORIES
Salad And Chicken Nuggets

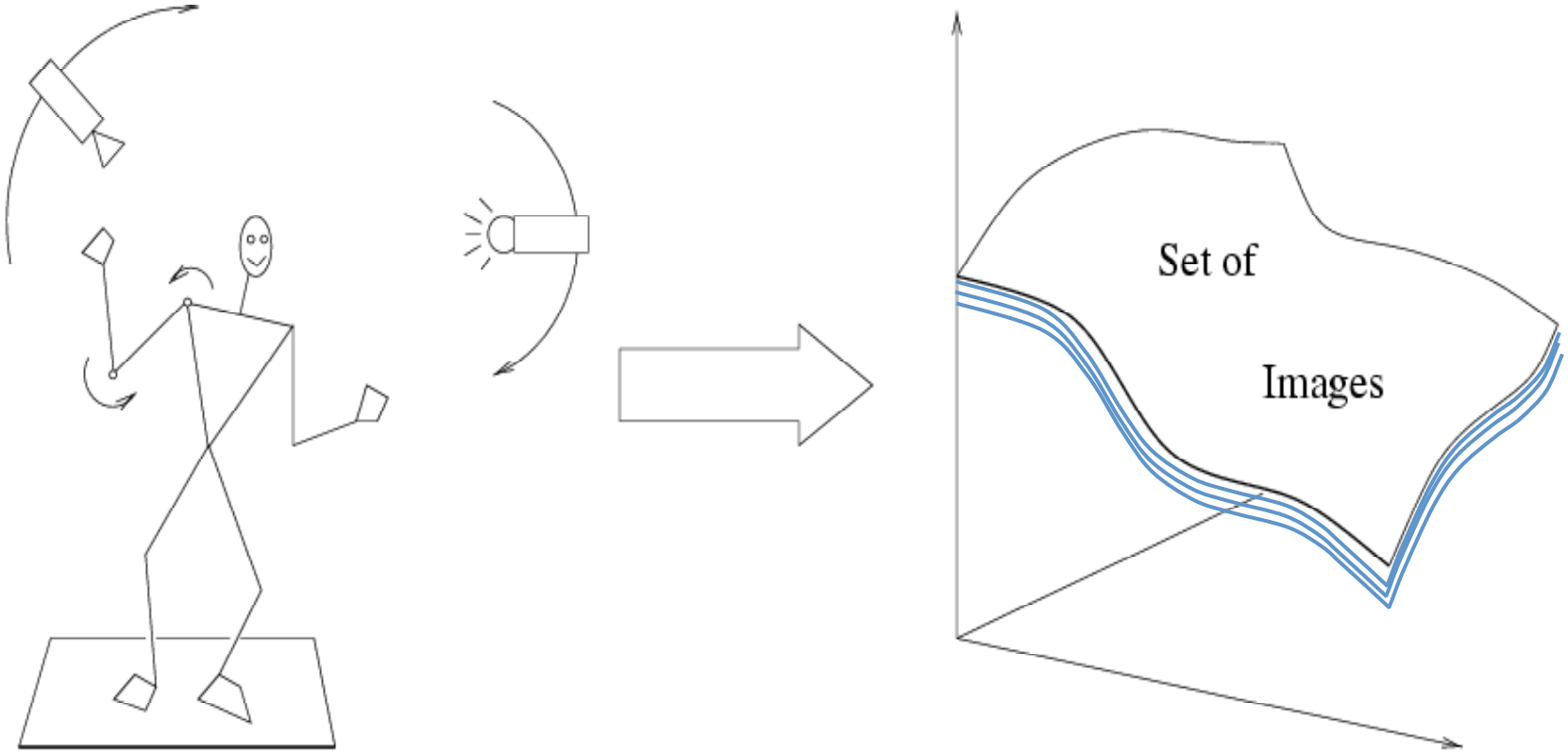

CHALK
Crayola 24 Count Sidewalk C...


T-Mobile NL
16:06
Albums by Impala
Architecture
Time
Score

Visual Recognition

- Need to
  - Classify images
  - Detect and localize objects
  - Estimate semantic and geometric attributes
  - Classify human activities

Why is this hard?



Variability: Camera position
Illumination
Shape parameters


How many object categories are there?
~10,000 to 30,000

Challenge: variable viewpoint

Michelangelo 1475-1564

Challenge: variable illumination

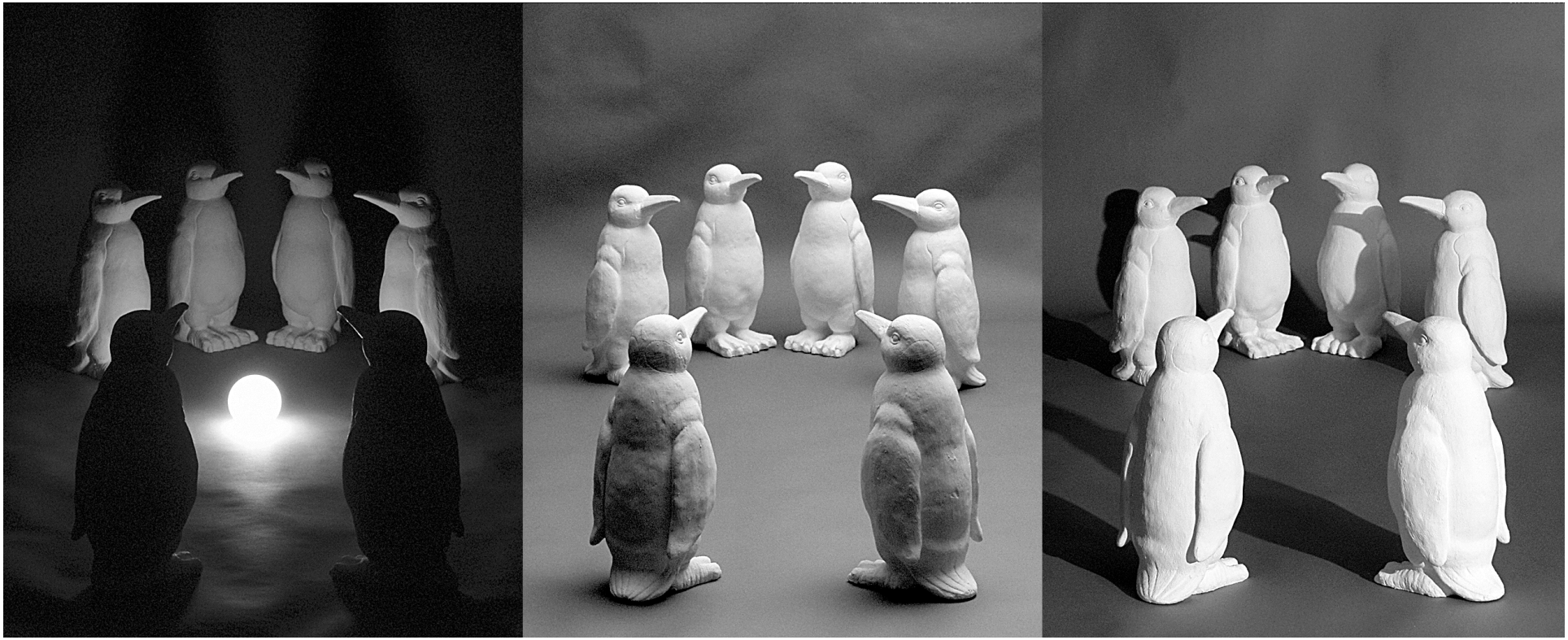
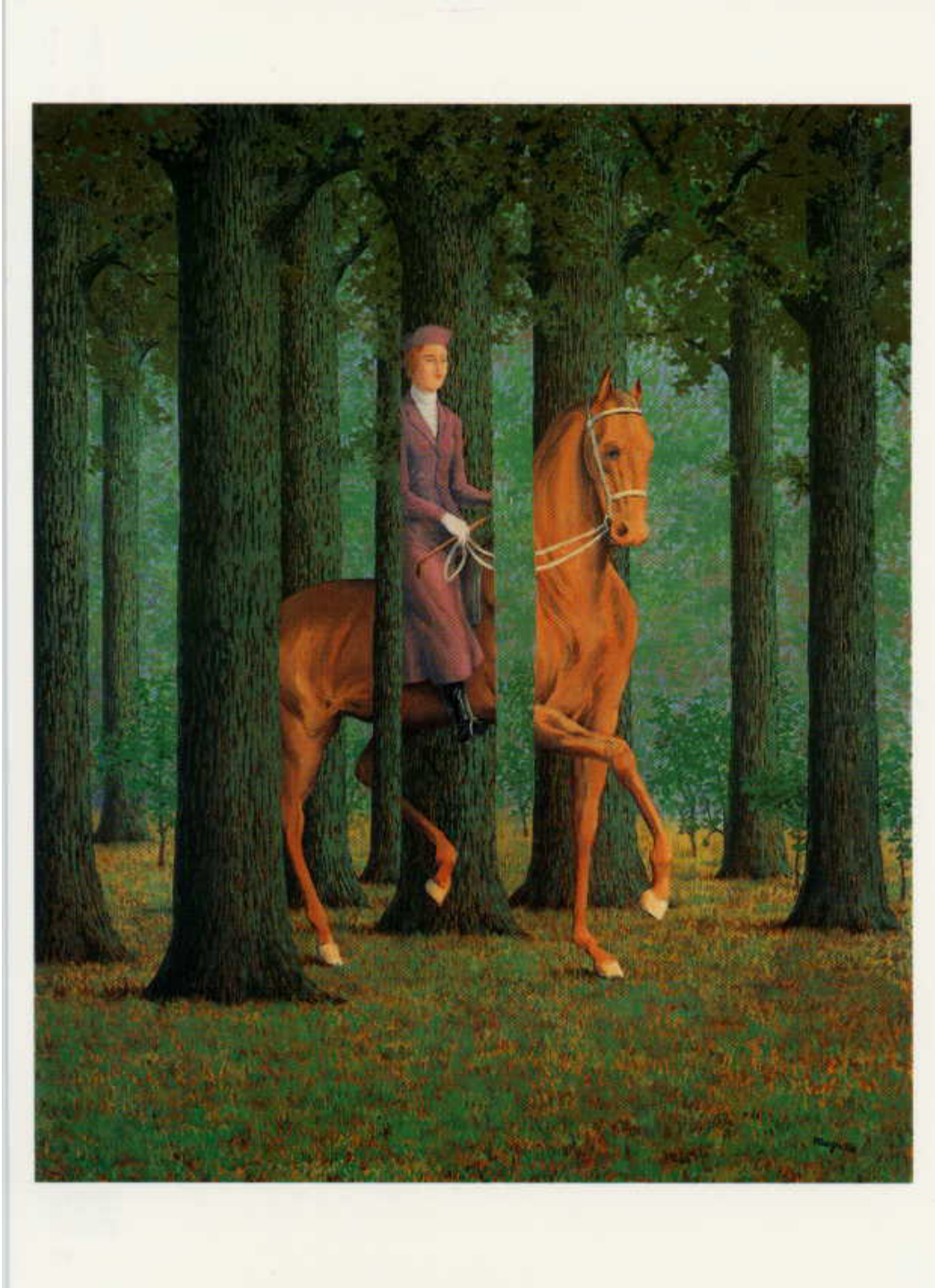


image credit: J. Koenderink

Challenge: scale

Challenge: deformation

Challenge: Occlusion



Magritte, 1957

Challenge: background clutter

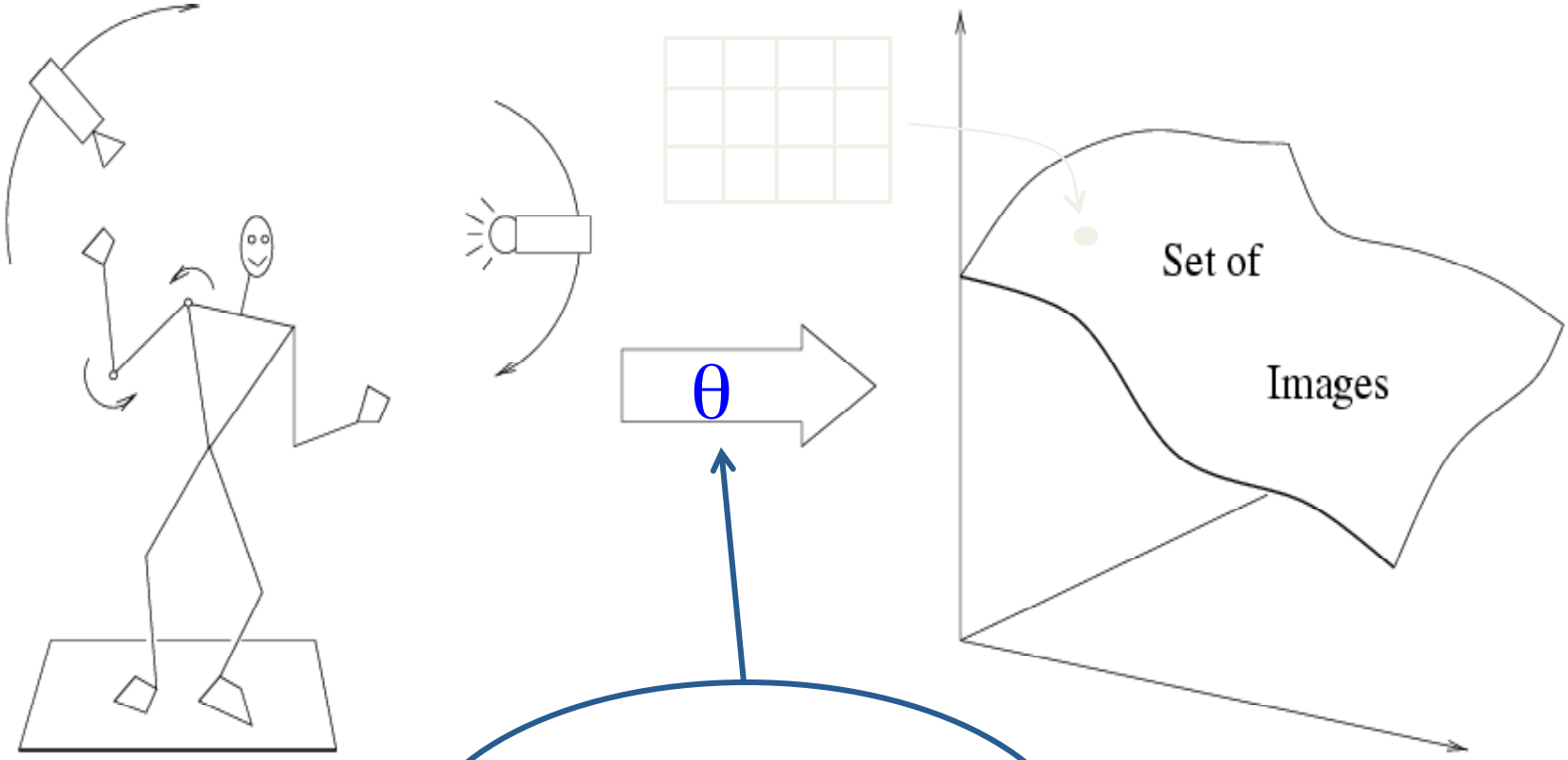
Kilmeny Niland. 1995

Challenge: intra-class variations

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era



Variability:

Camera position
Illumination

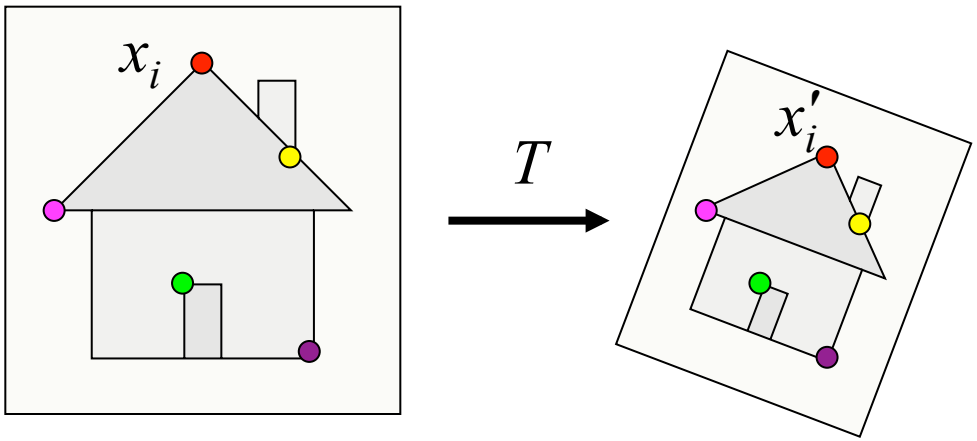
Alignment

Shape: assumed known

Roberts (1965); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

Instance Recognition

- Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images



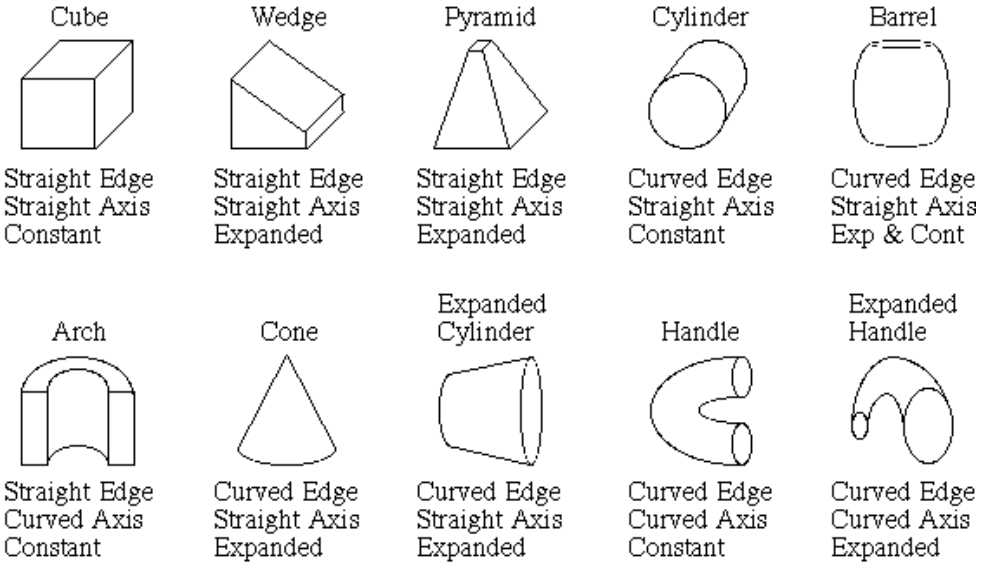
Find transformation T that minimizes

$$\sum_i \text{residual}(T(x_i), x_i')$$

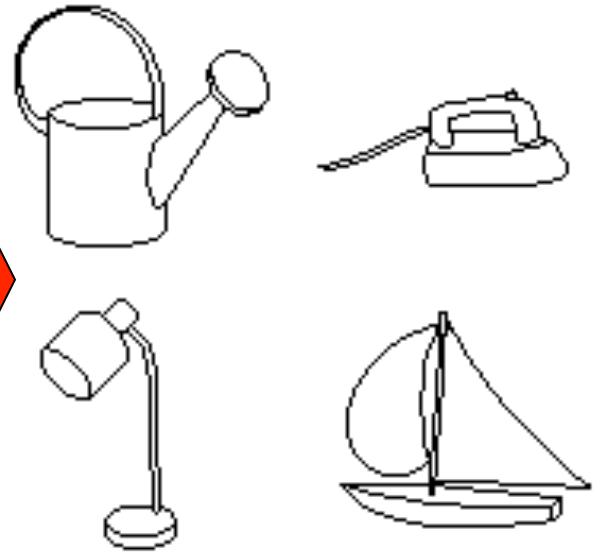
Recognition by components

Biederman (1987)

Primitives (geons)



Objects



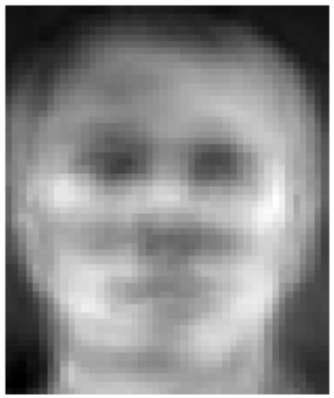
http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models

Eigenfaces (Turk & Pentland, 1991)



Experimental	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Svetlana Lazebnik

Limitations of global appearance models

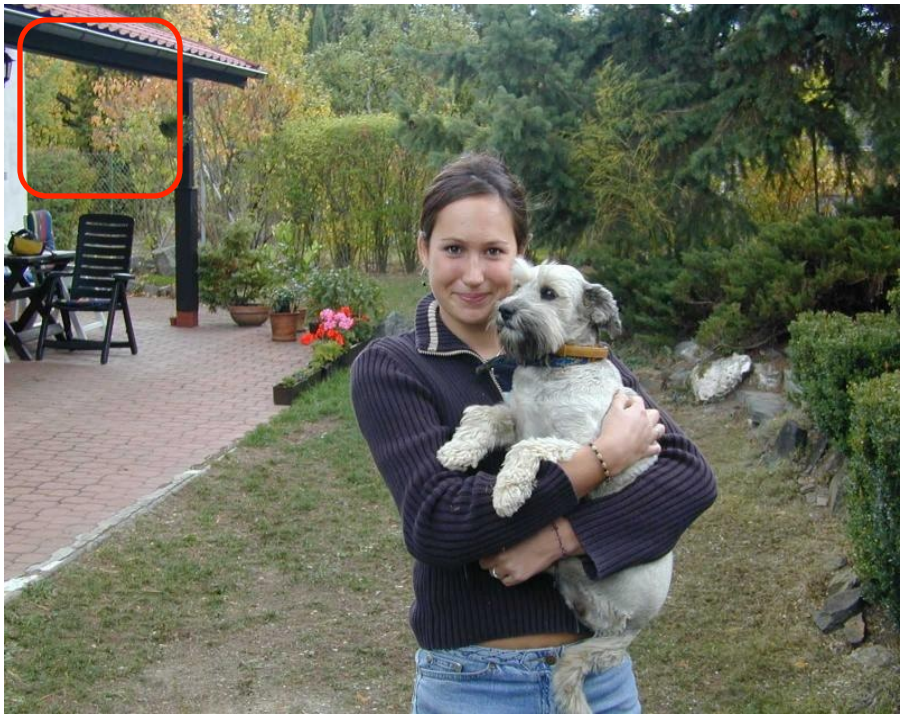
- Requires global registration of patterns
- Not robust to clutter, occlusion, geometric transformations

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- 1990s – present: sliding window approaches

Sliding window approaches



Sliding window approaches

- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade 2004
- Viola and Jones, 2000

- Schneiderman & Kanade, 2004
- Agrawal and Roth, 2002
- Poggio et al. 1993

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features

Local features for object instance recognition

D. Lowe (1999, 2004)

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models

Parts-and-shape models

- Model:
  - Object as a set of parts
  - Relative locations between parts
  - Appearance of part

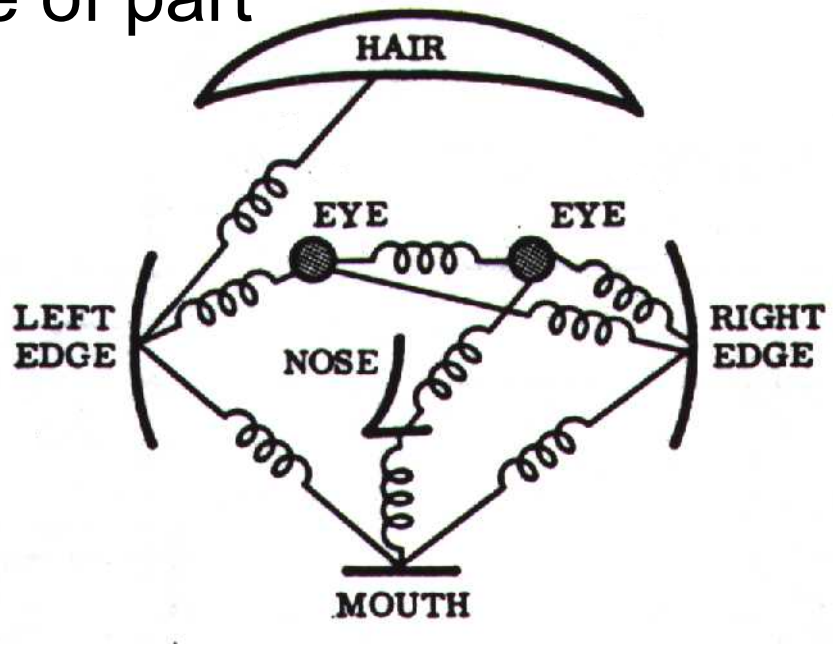
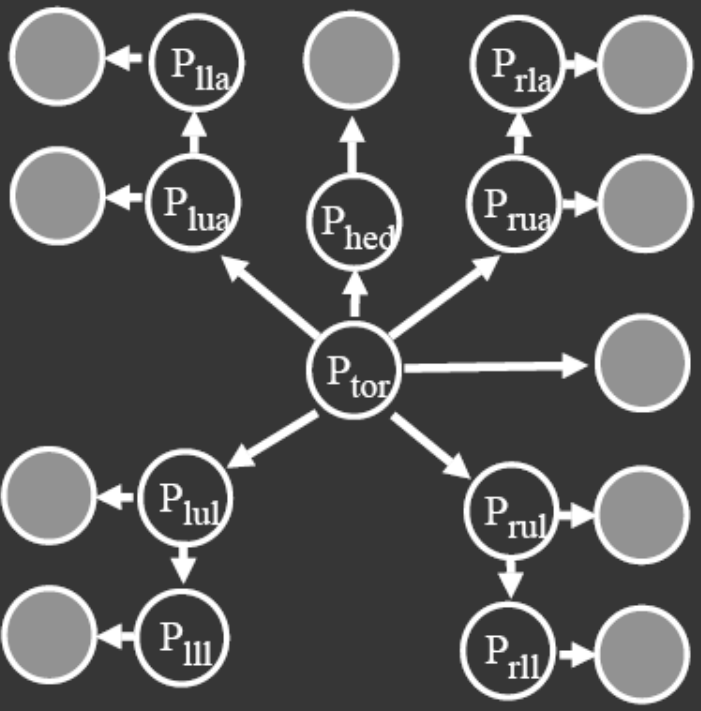
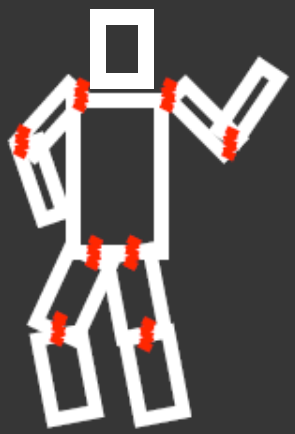


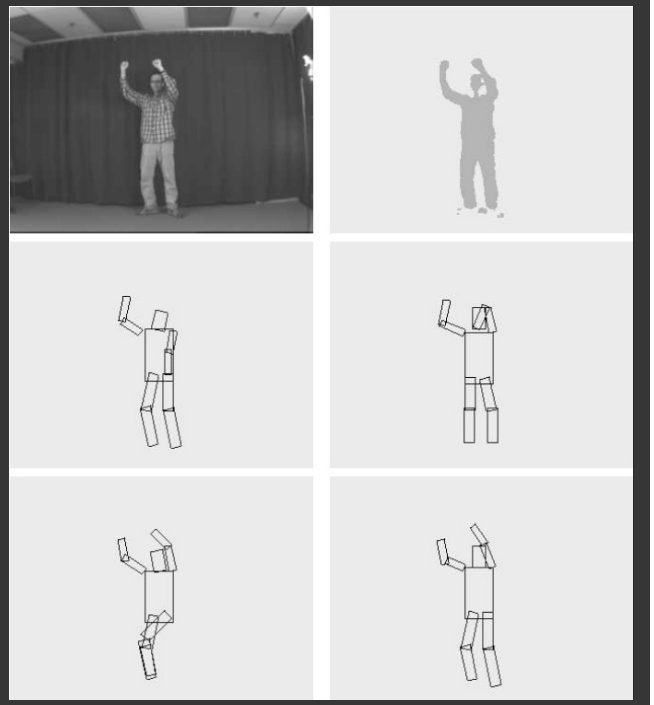
Figure from [Fischler & Elschlager 73]

Pictorial structure model

Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)





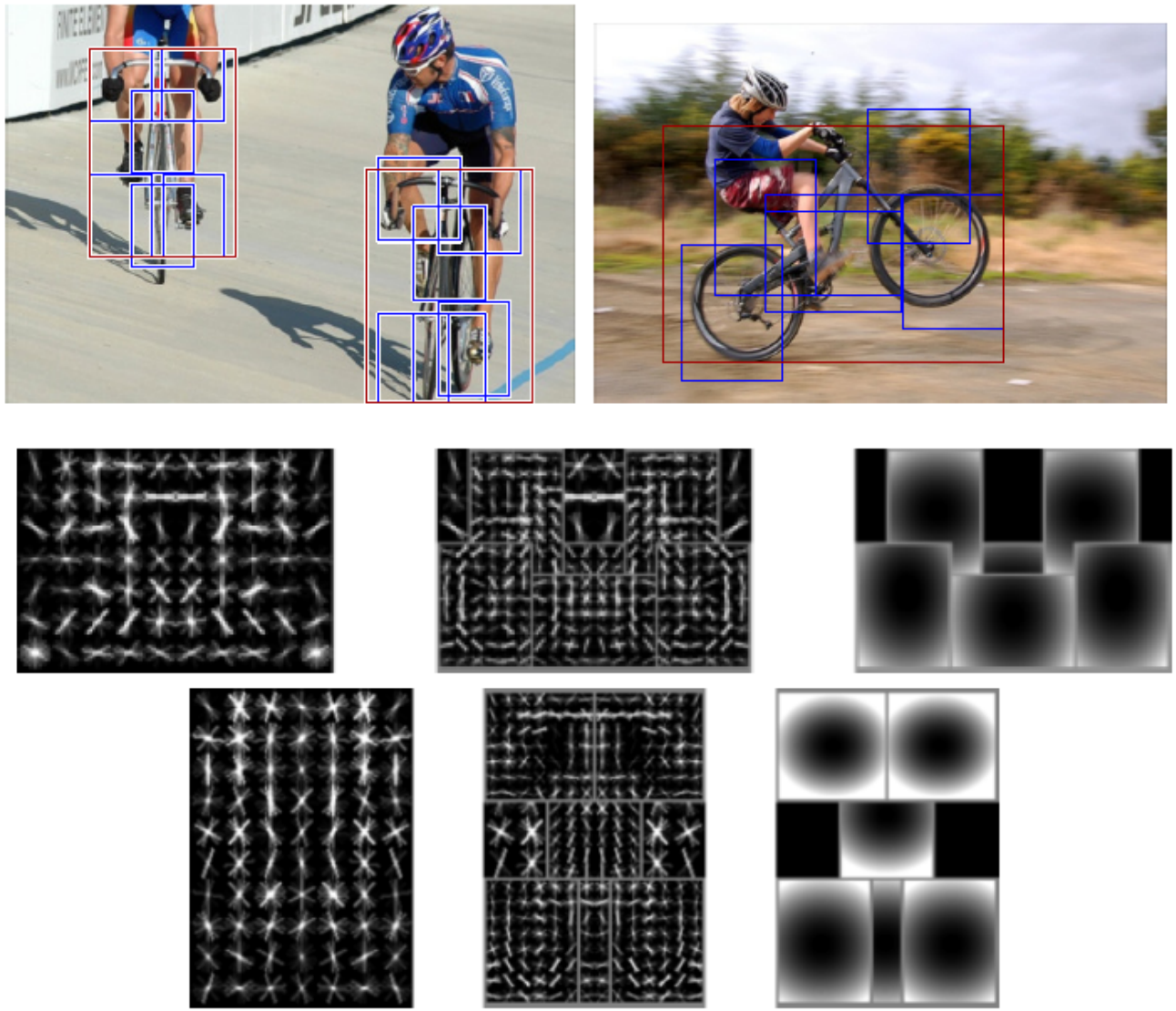


$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\mathrm{Im}(P_i))$$

part geometry

part appearance

Discriminatively trained part-based models



P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan,
"Object Detection with Discriminatively Trained Part-Based Models," PAMI 2009

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features

Bag-of-features models

Bag-of-features models

Object → **Bag of ‘words’**

Svetlana Lazebnik

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- Present trends: data-driven methods, context

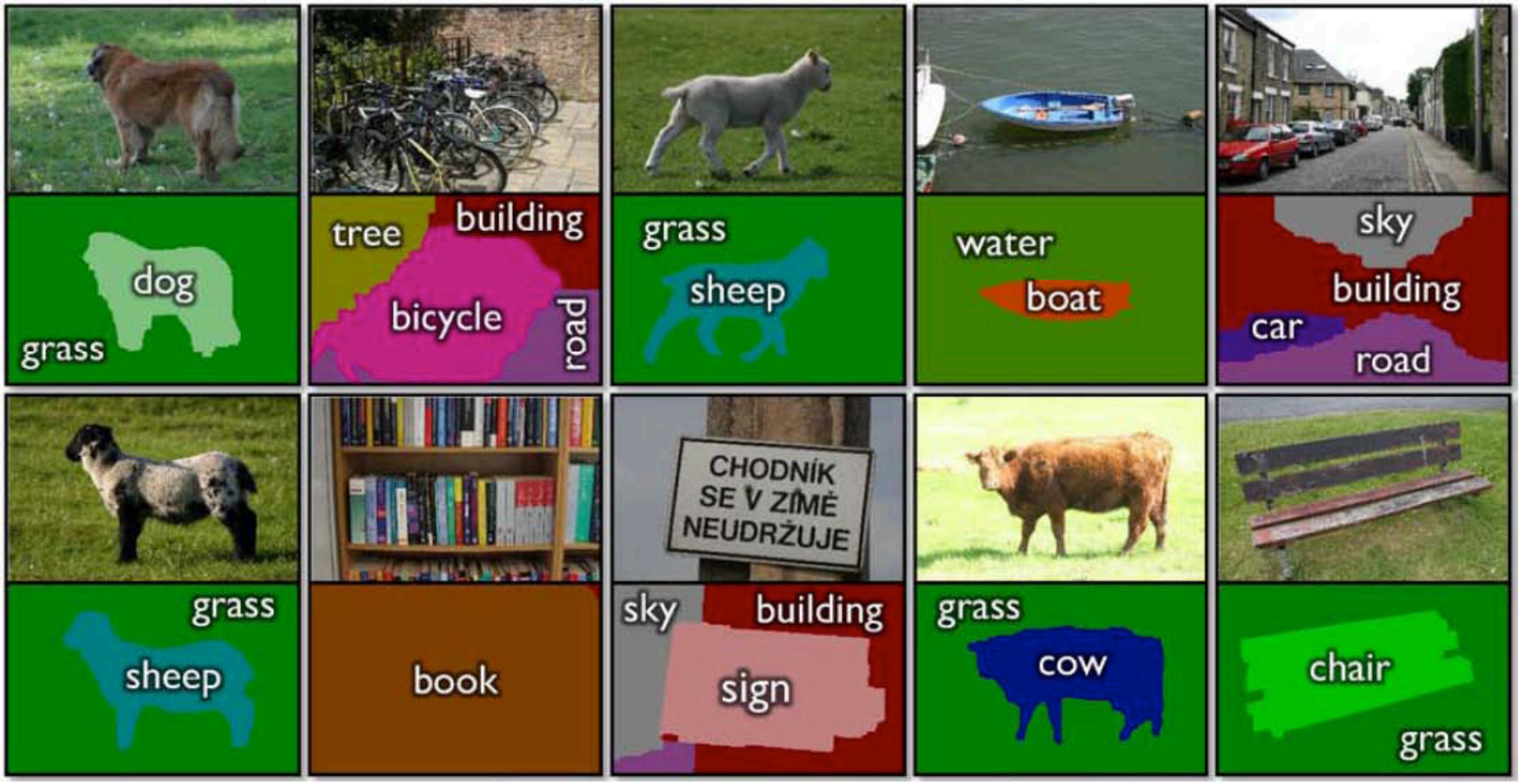
What Matters in Recognition?

- Learning Techniques
  - E.g. choice of classifier or inference method
- Representation
  - Low level: SIFT, HoG, GIST, edges
  - Mid level: Bag of words, sliding window, deformable model
  - High level: Contextual dependence
- Data
  - More is always better
  - Annotation is the hard part

Types of Recognition

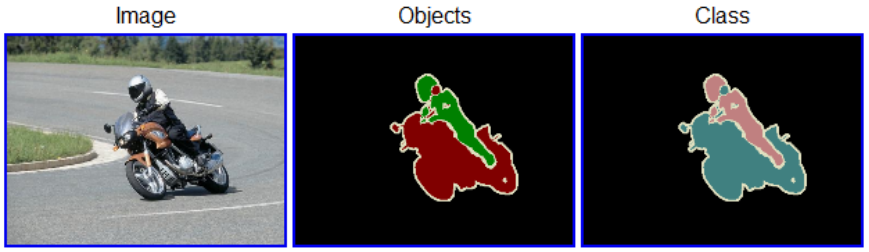
- Instance recognition
  - Recognizing a known object but in a new viewpoint, with clutter and occlusion
  - Location/Landmark Recognition
    - Recognize Paris, Rome, ... in photographs
    - Ideas from information retrieval
- Category recognition
  - Harder problem, even for humans
  - Bag of words, part-based, recognition and segmentation

Simultaneous recognition and detection



The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

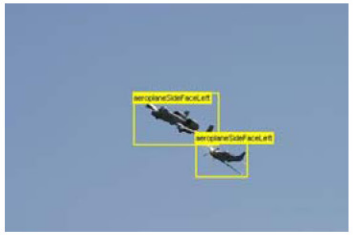
- Twenty object categories (aeroplane to TV/ monitor)
- Three (+2) challenges:
  - Classification challenge (is there an X in this image?)
  - Detection challenge (draw a box around every X)
  - Segmentation challenge (which class is each pixel?)



Slides from Noah Snavely

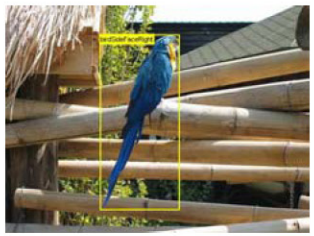
Examples

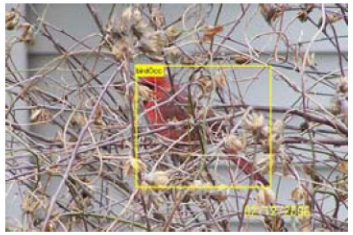
Aeroplane



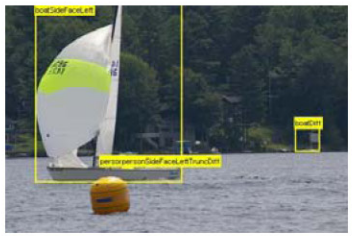
Bicycle

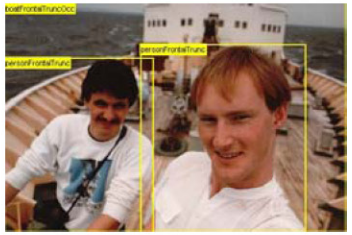
Bird





Boat





Bottle

Bus

Car

Cat

Chair

Cow

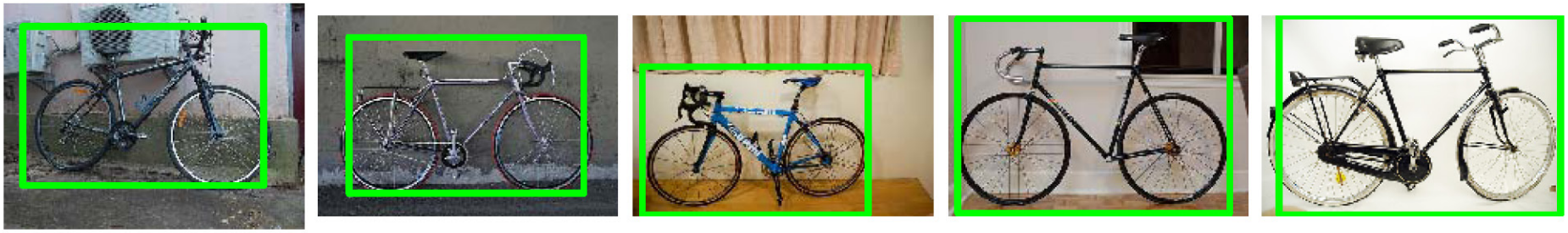
Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)

True Positives - Bicycle

UoCTTI_LSVM-MDPM

OXFORD_MKL



NECUIUC_CLS-DTCT

False Positives - Bicycle

UoCTTI_LSVM-MDPM

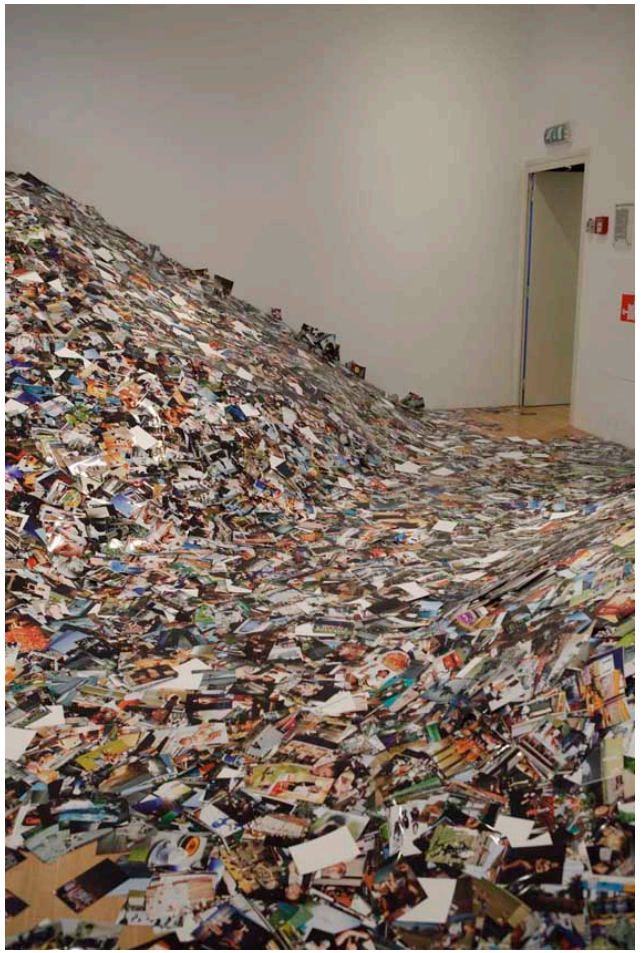
OXFORD_MKL

NECUIUC_CLS-DTCT

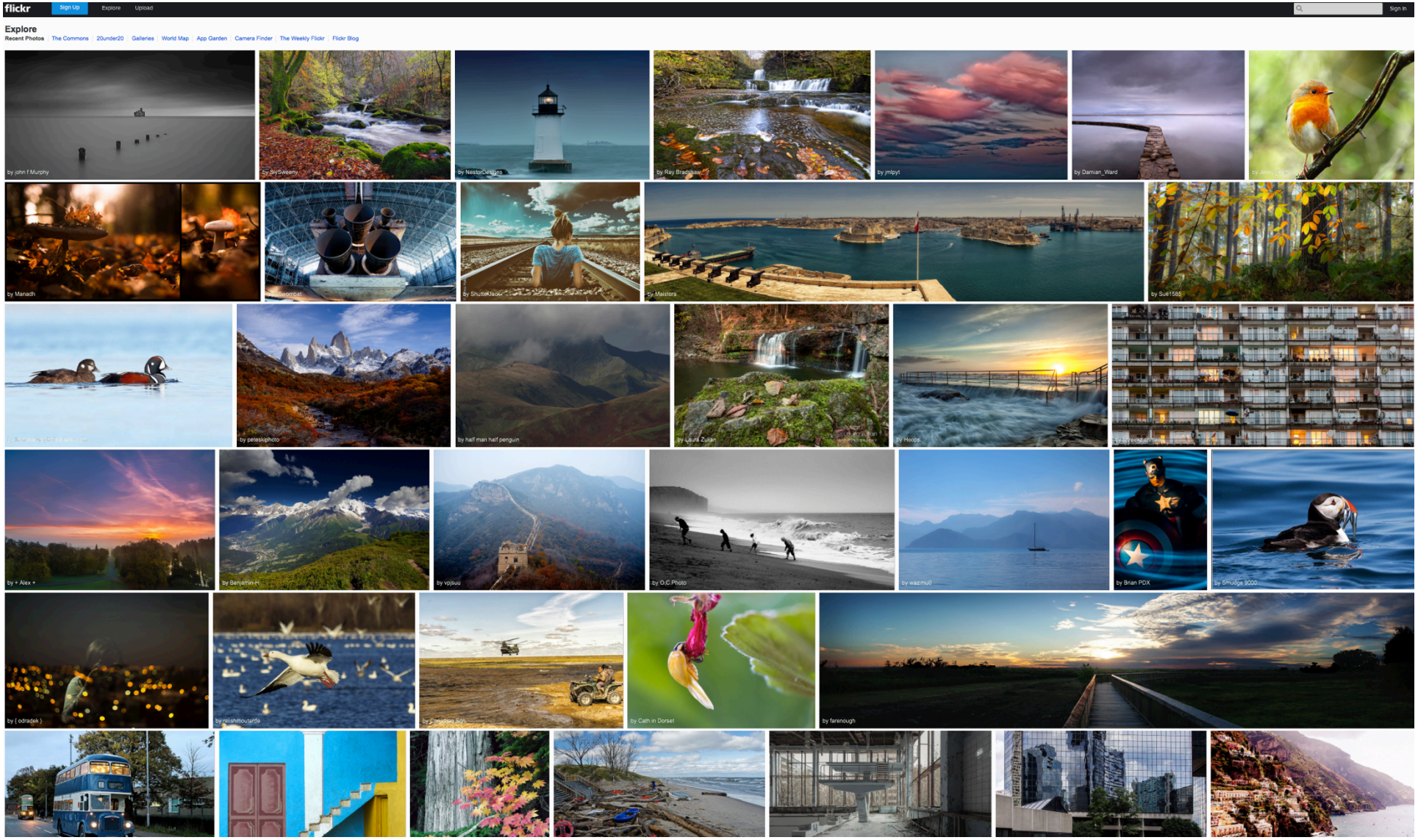
Where to from here?

- Scene Understanding
  - Big data – lots of images
  - Crowd sourcing – lots of people
  - Deep Learning – lots of compute

24 hours of Photo Sharing



installation by Erik Kessels


flickr
Sign Up
Explore
Upload
Sign in
Explore
Recent Photos | The Commons | 20under20 | Galleries | World Map | App Garden | Camera Finder | The Weekly Flickr | Flickr Blog
by john f Murphy
by SlySweeny
by NestorDesigns
by Ray Bradshaw
by jmlpyt
by Damian_Ward
by Manadh
by ShutterJack
by Malstora
by Sue1585
by peteskiphoto
by half man half penguin
by Laura Zuran
by Hoops
by + Alex +
by Benjamin-H
by vpjsuu
by O.C.Photo
by wazmu0
by Brian PDX
by Smudge 9000
by { odradek }
by relishmoutarde
by Cath in Dorset
by farenough

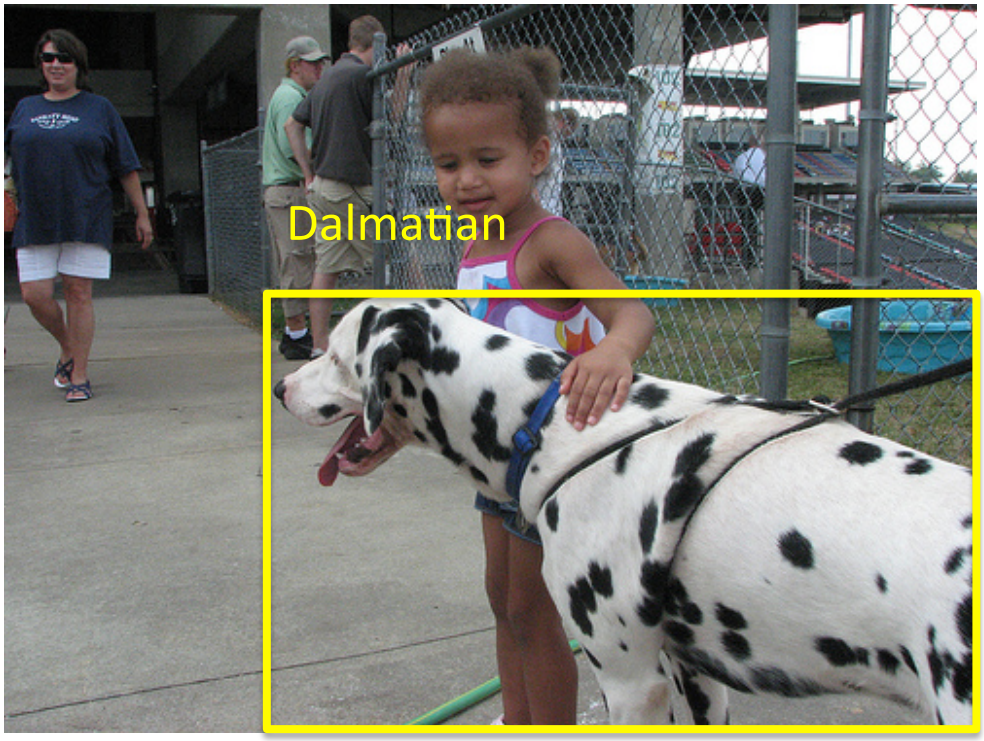
Data Sets

- ImageNet
  - Huge, Crowdsourced, Hierarchical, *Iconic* objects
- PASCAL VOC
  - *Not* Crowdsourced, bounding boxes, 20 categories.
- SUN Scene Database, Places
  - *Not* Crowdsourced, 397 (or 720) scene categories
- LabelMe (Overlaps with SUN)
  - Sort of Crowdsourced, Segmentations, Open ended
- SUN *Attribute* database (Overlaps with SUN)
  - Crowdsourced, 102 attributes for every scene
- OpenSurfaces
  - Crowdsourced, materials

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

~~20 object classes~~ ~~22,591 images~~

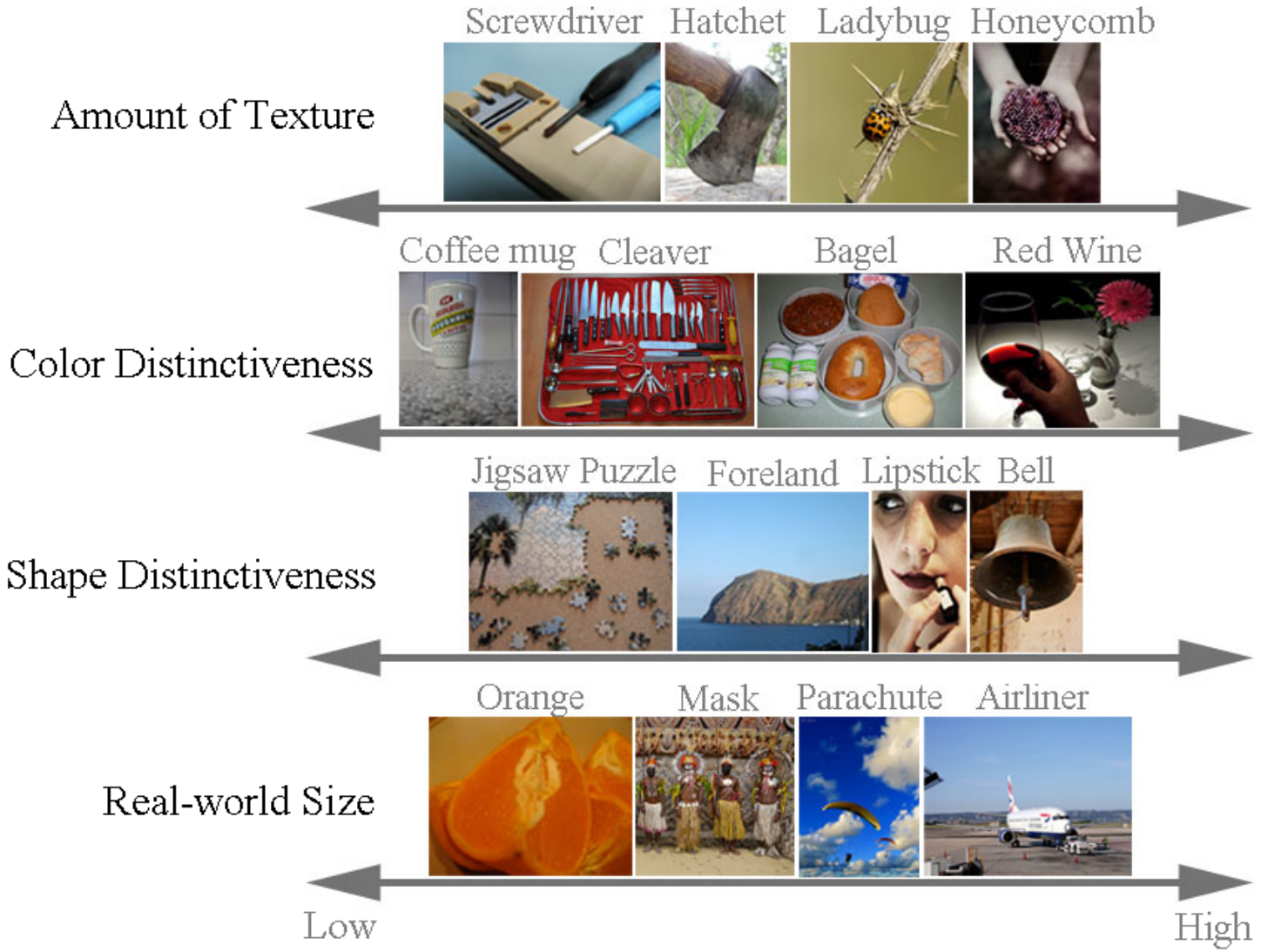
1000 object classes **1,431,167 images**

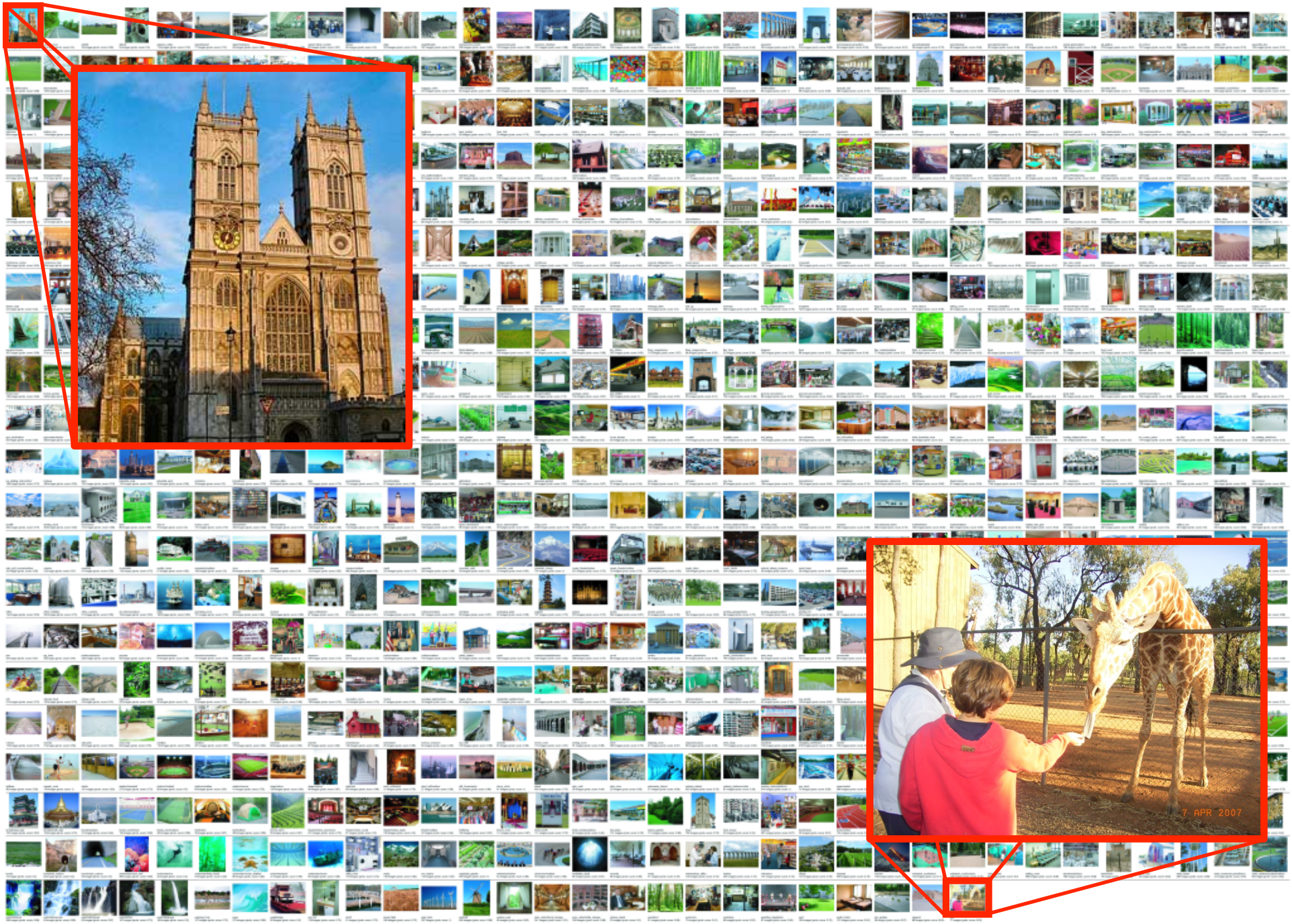


http://image-net.org/challenges/LSVRC/{2010,2011,2012}

Variety of object classes in ILSVRC

Variety of object classes in ILSVRC




7 APR 2007

Our Goal: Infer object properties

Can I **poke with it**?

Is it **alive**?

What **shape** is it?

Does it have a **tail**?

Can I **put stuff in it**?

Is it **soft**?

Will it **blend**?

What are attributes?

What do we want to know about this object?

Object recognition expert:
“Dog”

Person in the Scene:
“Big pointy teeth”, “Can move fast”, “Looks angry”

Why infer properties

1. We want detailed information about objects

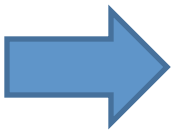
"Dog"
vs.
"Large, angry animal with pointy teeth"

Why infer properties

2. We want to be able to infer something about unfamiliar objects

Familiar Objects





New Object

Why infer properties

2. We want to be able to infer something about unfamiliar objects

If we can infer properties…

Familiar Objects

Has Stripes
Has Ears
Has Eyes
….

Has Four Legs
Has Mane
Has Tail
Has Snout
….

Brown
Muscular
Has Snout
….

New Object

Has Stripes (like cat)
Has Mane and Tail (like horse)
Has Snout (like horse and dog)

Why infer properties

3. We want to make comparisons between objects or categories

What is unusual about this dog?

What is the difference between horses and zebras?

Where to from here?

- Scene Understanding
  - Big data – lots of images
  - Crowd sourcing – lots of people
  - Deep Learning – lots of compute

Image categorization

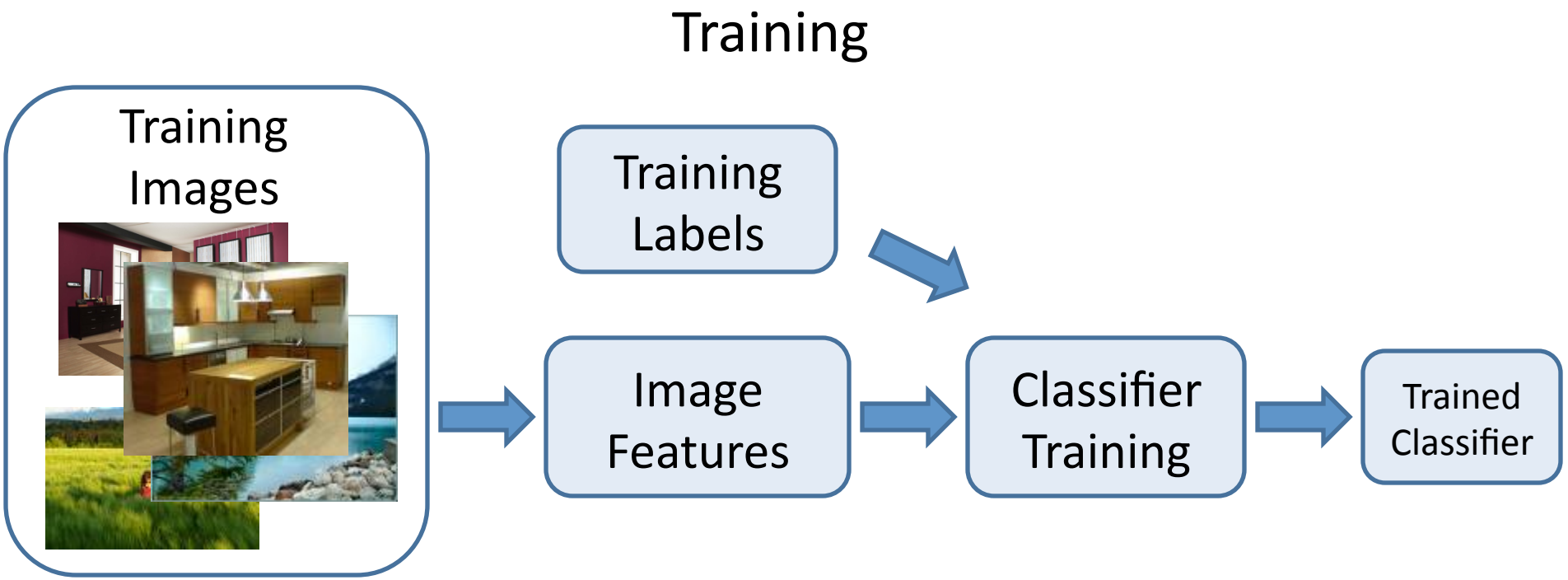
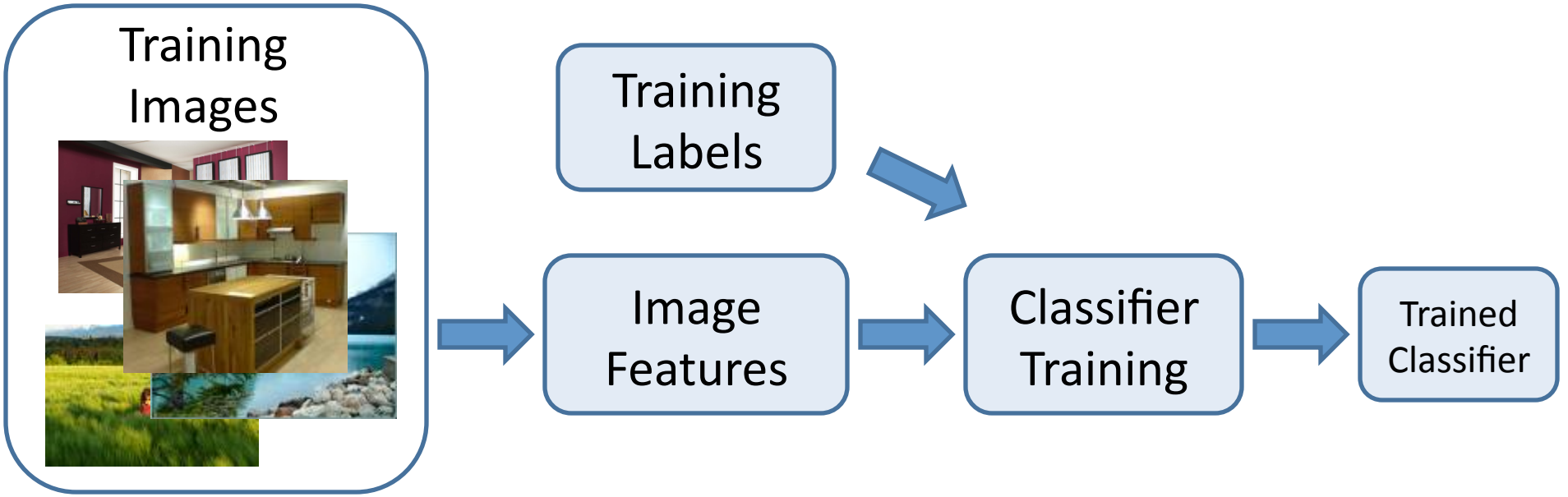
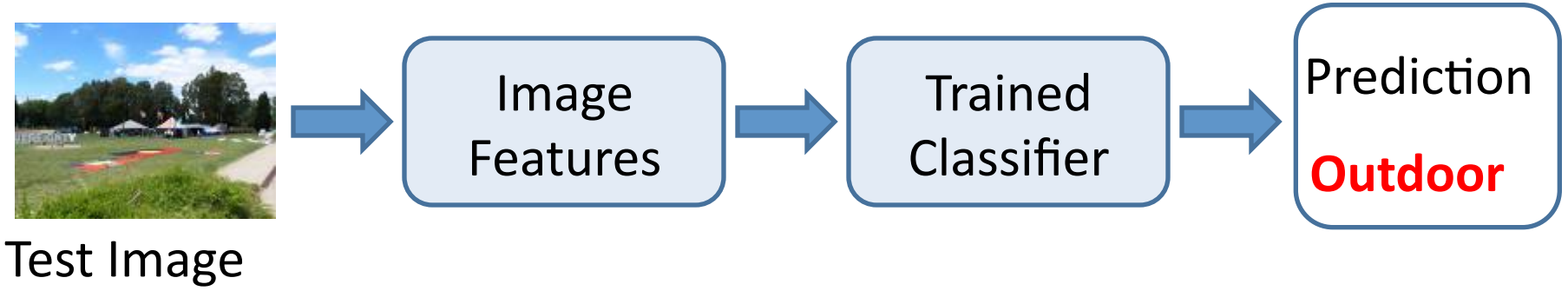


Image Categorization

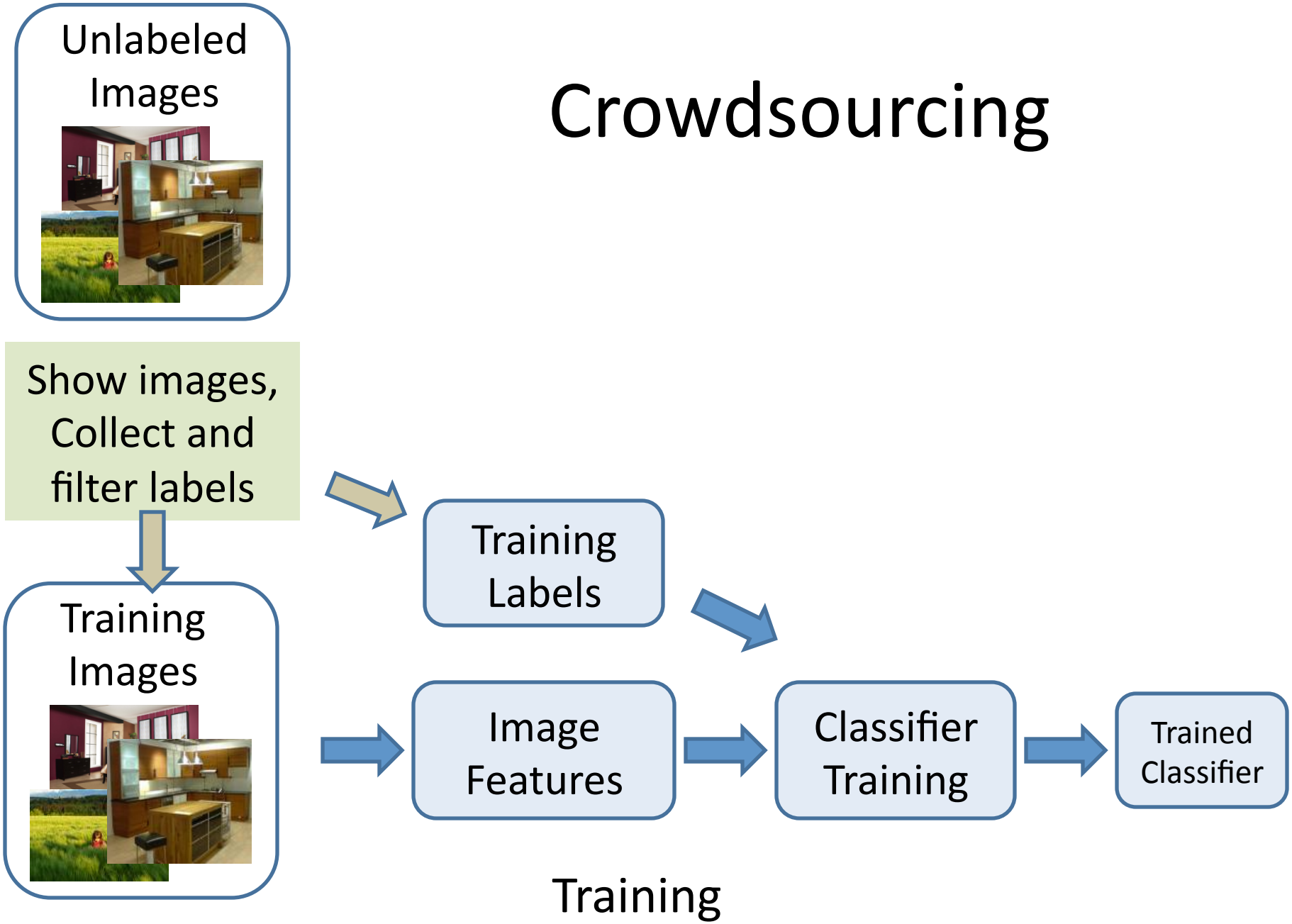
Training



Testing



Crowdsourcing



Where to from here?

- Scene Understanding
  - Big data – lots of images
  - Crowd sourcing – lots of people
  - Deep Learning – lots of compute

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC

Dense grid descriptor: HOG, LBP

Coding: local coordinate, super-vector

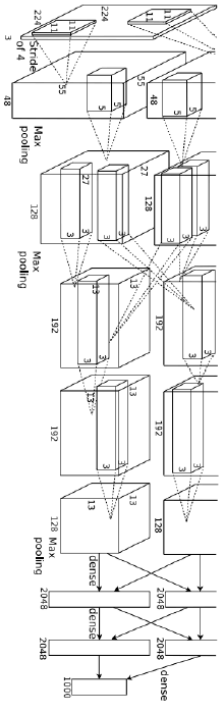
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

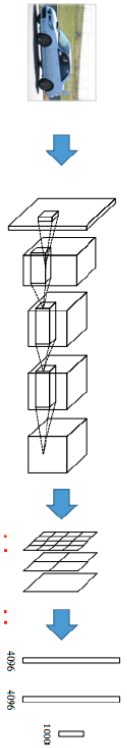
GoogLeNet

[Szegedy arxiv 2014]

VGG

[Simonyan arxiv 2014]

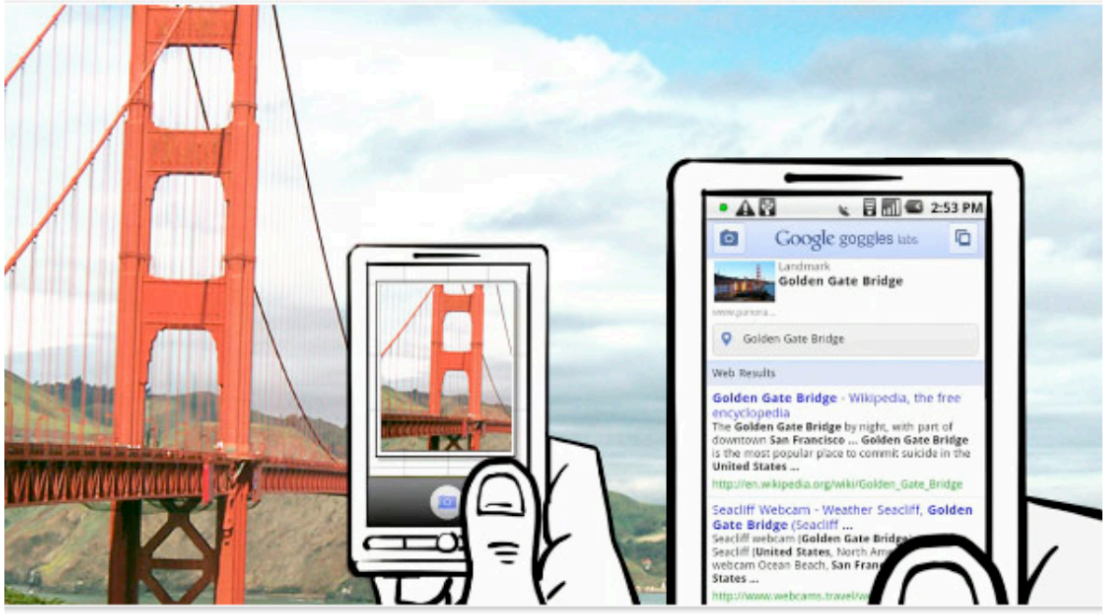
MSRA

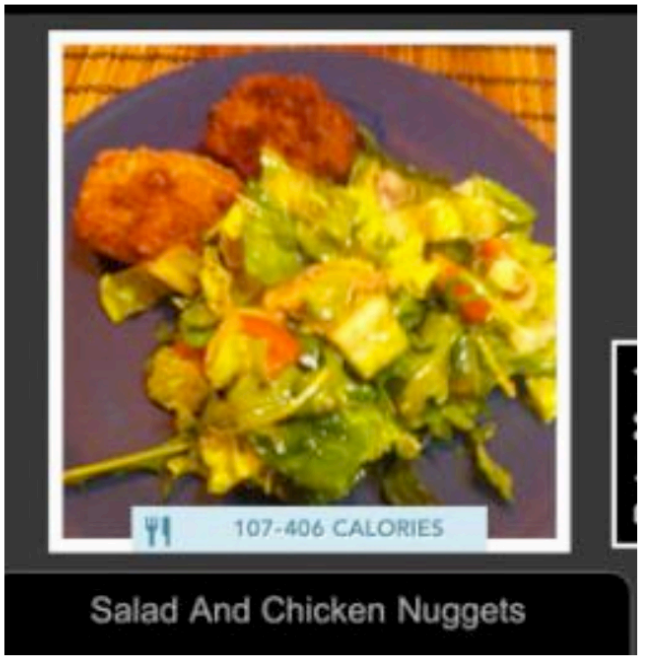


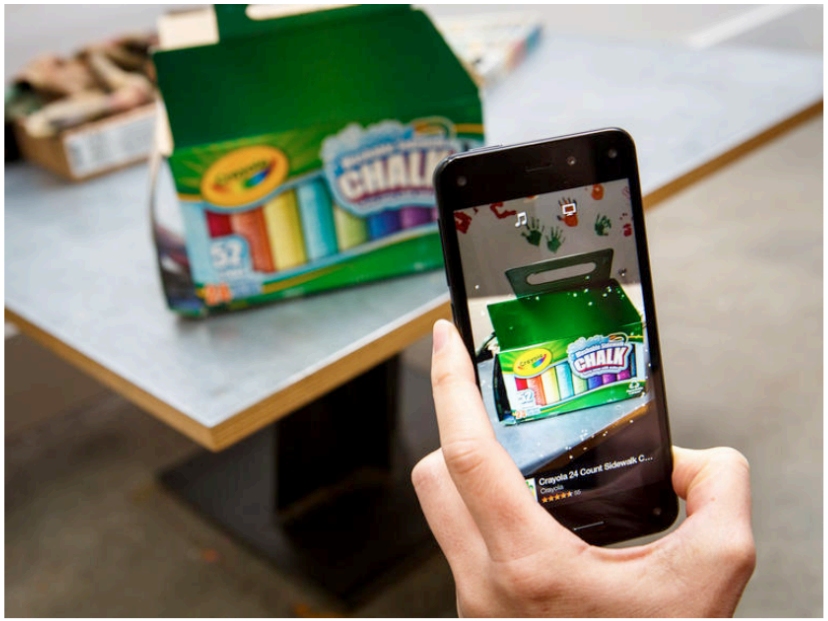
[He arxiv 2014]

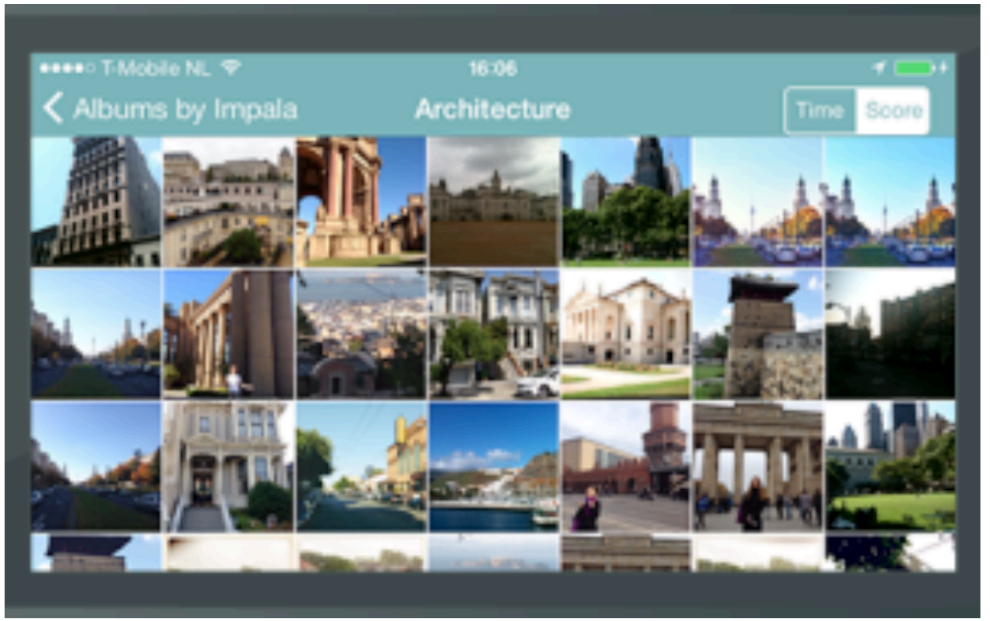
Deep Learning or CNNs

- Since 2012, huge impact…, best results
- Can soak up all the data for better prediction


2:53 PM
Google goggles
Landmark
Golden Gate Bridge
Golden Gate Bridge
Web Results
Golden Gate Bridge - Wikipedia, the free encyclopedia
The Golden Gate Bridge by night, with part of downtown San Francisco ... Golden Gate Bridge is the most popular place to commit suicide in the United States ...
http://en.wikipedia.org/wiki/Golden_Gate_Bridge
Seacliff Webcam - Weather Seacliff, Golden Gate Bridge (Seacliff ...


107-406 CALORIES
Salad And Chicken Nuggets


CHALK
Crayola 24 Count Sidewalk C...


T-Mobile NL
16:06
Albums by Impala
Architecture
Time
Score

Karpathy blog

- You recognize it is an image of a bunch of people and you understand they are in a hallway
- You recognize that there are 3 mirrors in the scene so some of those people are "fake" replicas from different viewpoints.
- You recognize Obama from the few pixels that make up his face. It helps that he is in his suit and that he is surrounded by other people with suits.
- You recognize that there's a person standing on a scale, even though the scale occupies only very few white pixels that blend with the background. But, you've used the person's pose and knowledge of how people interact with objects to figure it out.
- You recognize that Obama has his foot positioned just slightly on top of the scale. Notice the language I'm using: It is in terms of the 3D structure of the scene, not the position of the leg in the 2D coordinate system of the image.
- You know how physics works: Obama is leaning in on the scale, which applies a force on it. Scale measures force that is applied on it, that's how it works => it will over-estimate the weight of the person standing on it.
- The person measuring his weight is not aware of Obama doing this. You derive this because you know his pose, you understand that the field of view of a person is finite, and you understand that he is not very likely to sense the slight push of Obama's foot.
- You understand that people are self-conscious about their weight. You also understand that he is reading off the scale measurement, and that shortly the over-estimated weight will confuse him because it will probably be much higher than what he expects. In other words, you reason about implications of the events that are about to unfold seconds after this photo was taken, and especially about the thoughts and how they will develop inside people's heads. You also reason about what pieces of information are available to people.
- There are people in the back who find the person's imminent confusion funny. In other words you are reasoning about state of mind of people, and their view of the state of mind of another person. That's getting frighteningly meta.
- Finally, the fact that the perpetrator here is the president makes it maybe even a little more funnier. You understand what actions are more or less likely to be undertaken by different people based on their status and identity.

Karpathy blog