Visual Recognition: Introduction

Raquel Urtasun

TTI Chicago

Jan 3, 2012

Today's lecture ...

- Intro to what is computer vision
- Description of the class

What does visual recognition involve?

[Source: R. Fergus]

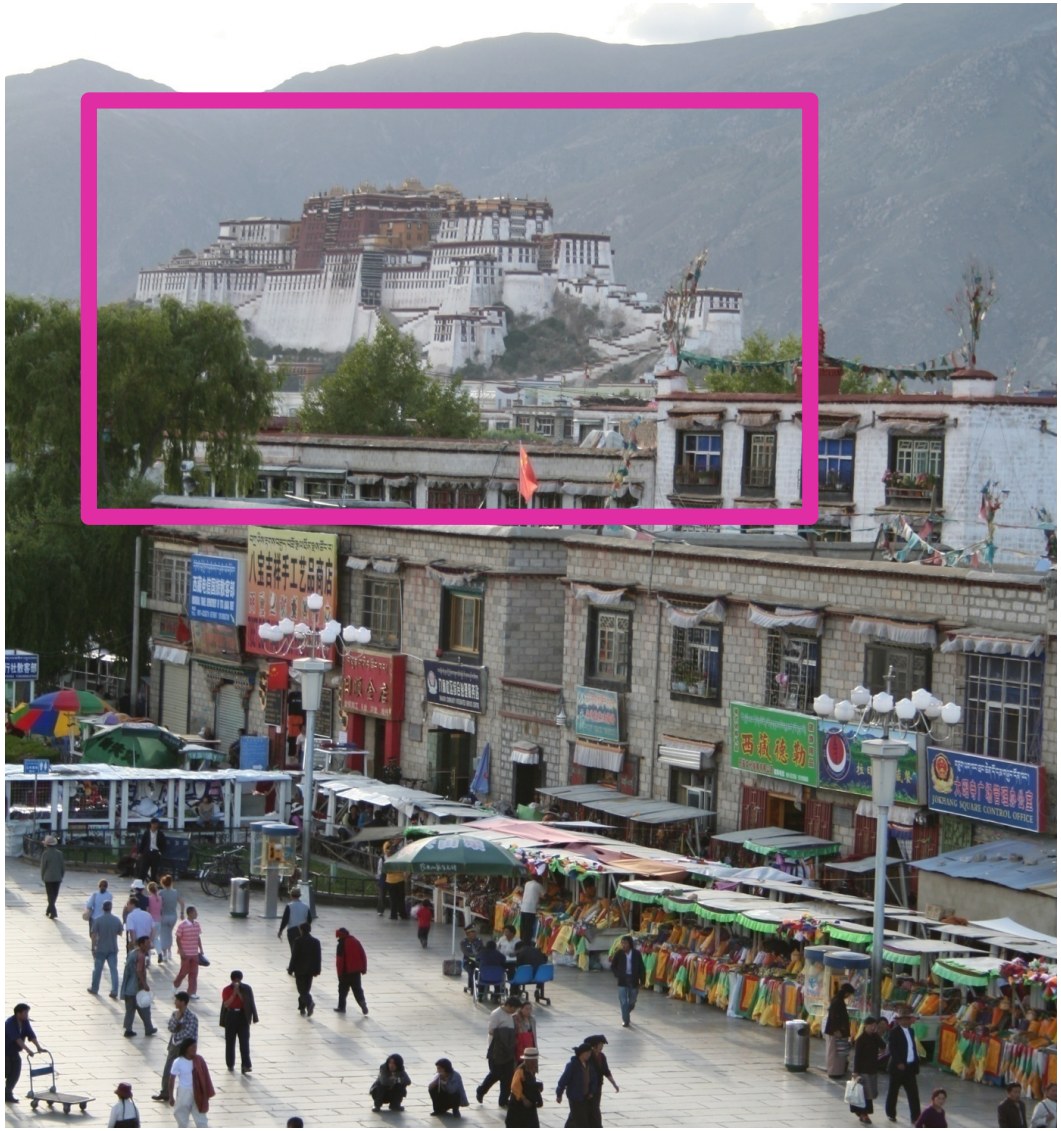
Verification: Is that a lamp?

Detection: Where are the people?

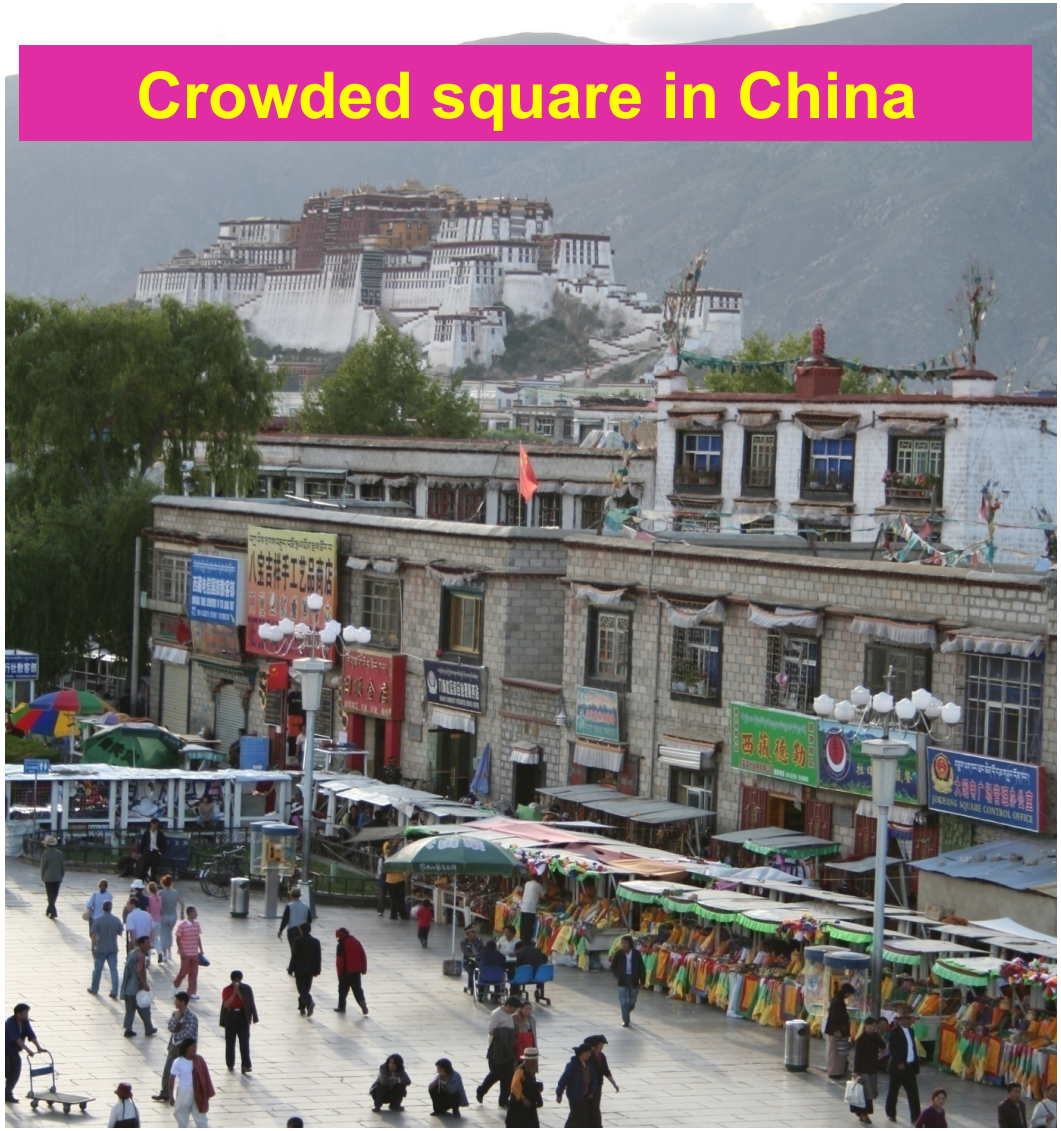
Activity Recognition: What are they doing?

Pose: Which pose do they have?

Identification: Is that the great wall?



Description: Attributes and relations



Is computer vision hard?

Is computer vision hard?

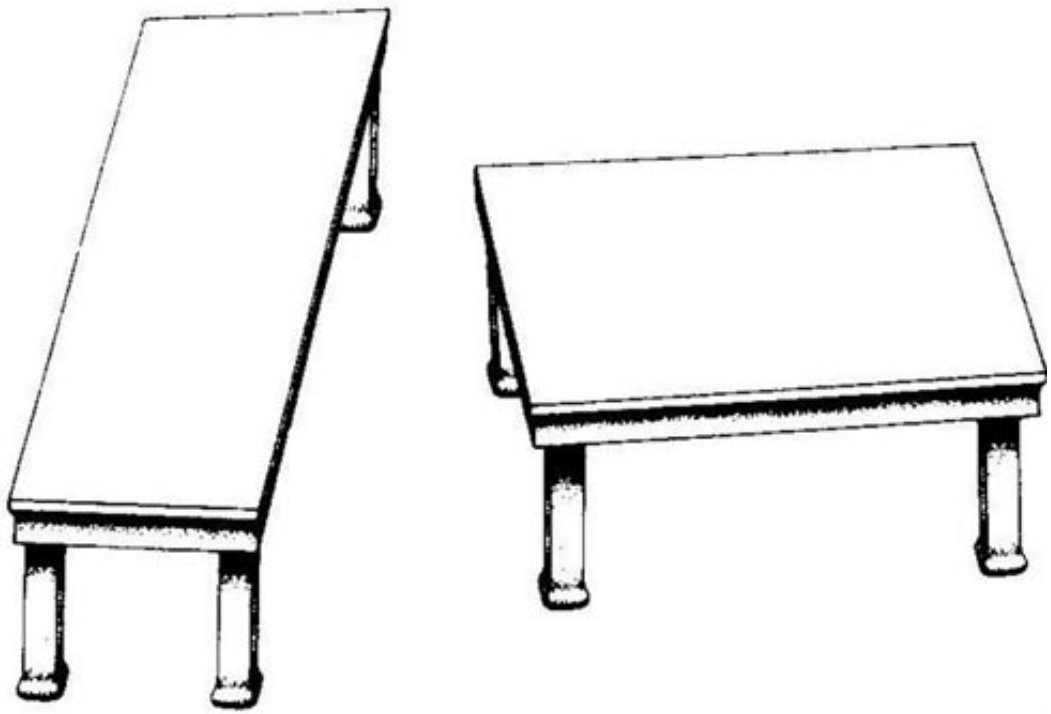
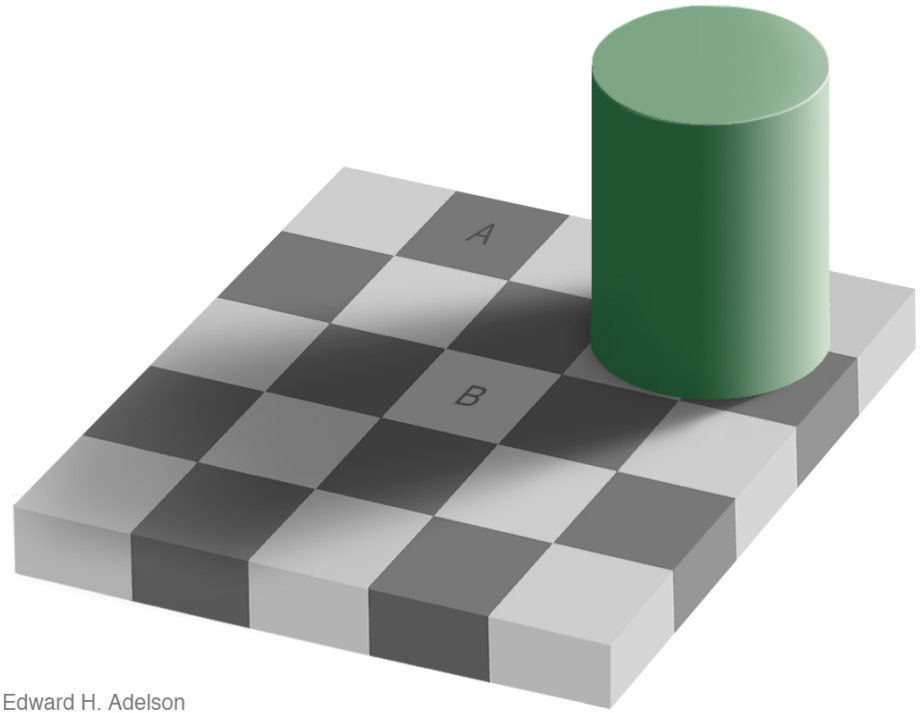


Figure: Turning the Tables by Roger Shepard

- Depth processing is automatic, and we can not shut it down

[Souce: A. Torralba]

Is computer vision hard?

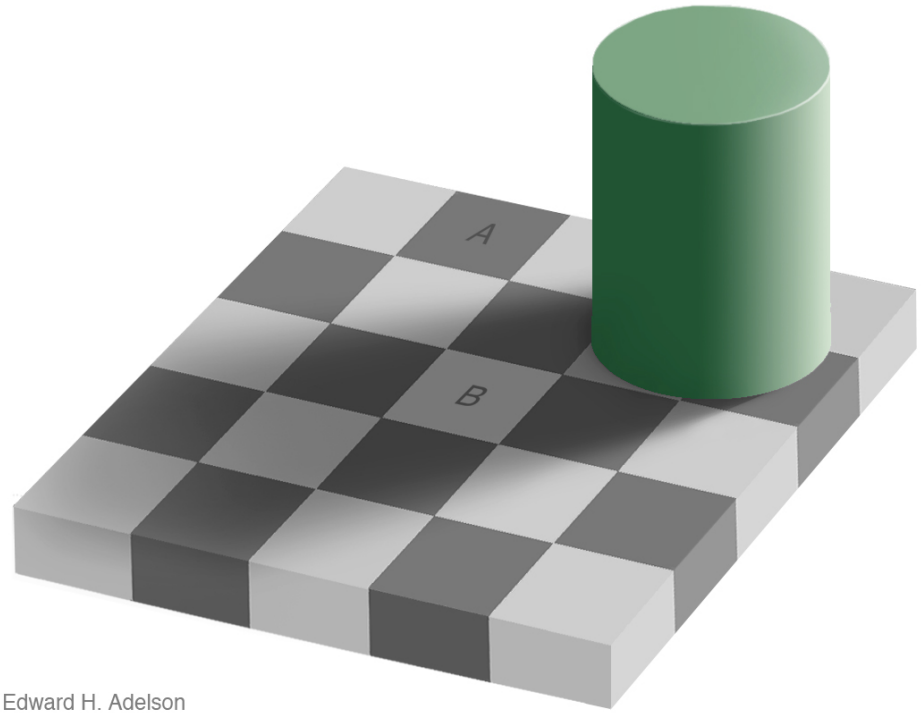


Edward H. Adelson

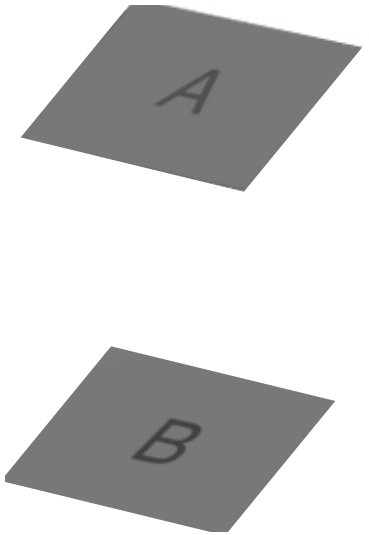
- Do A and B have the same gray level?

[Souce: A. Torralba]

Is computer vision hard?



Edward H. Adelson



- Do A and B have the same gray level?

[Souce: A. Torralba]

Is computer vision hard?

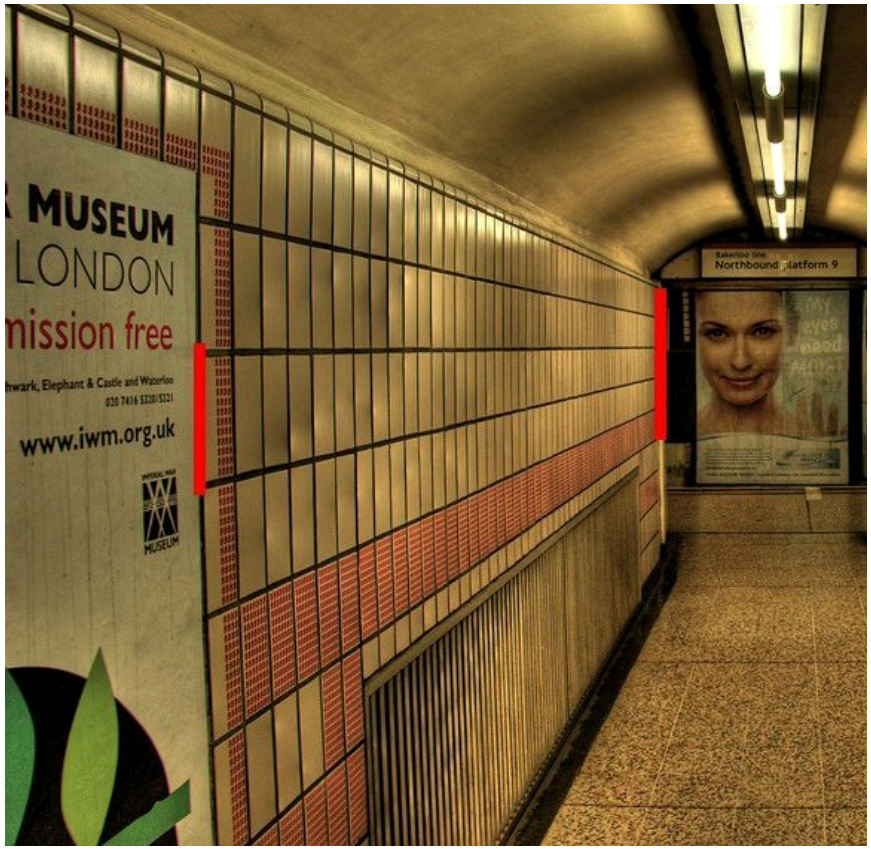


Figure: 2006 Walt Anthony

- Do they have the same length?

[Souce: A. Torralba]

Is computer vision hard?

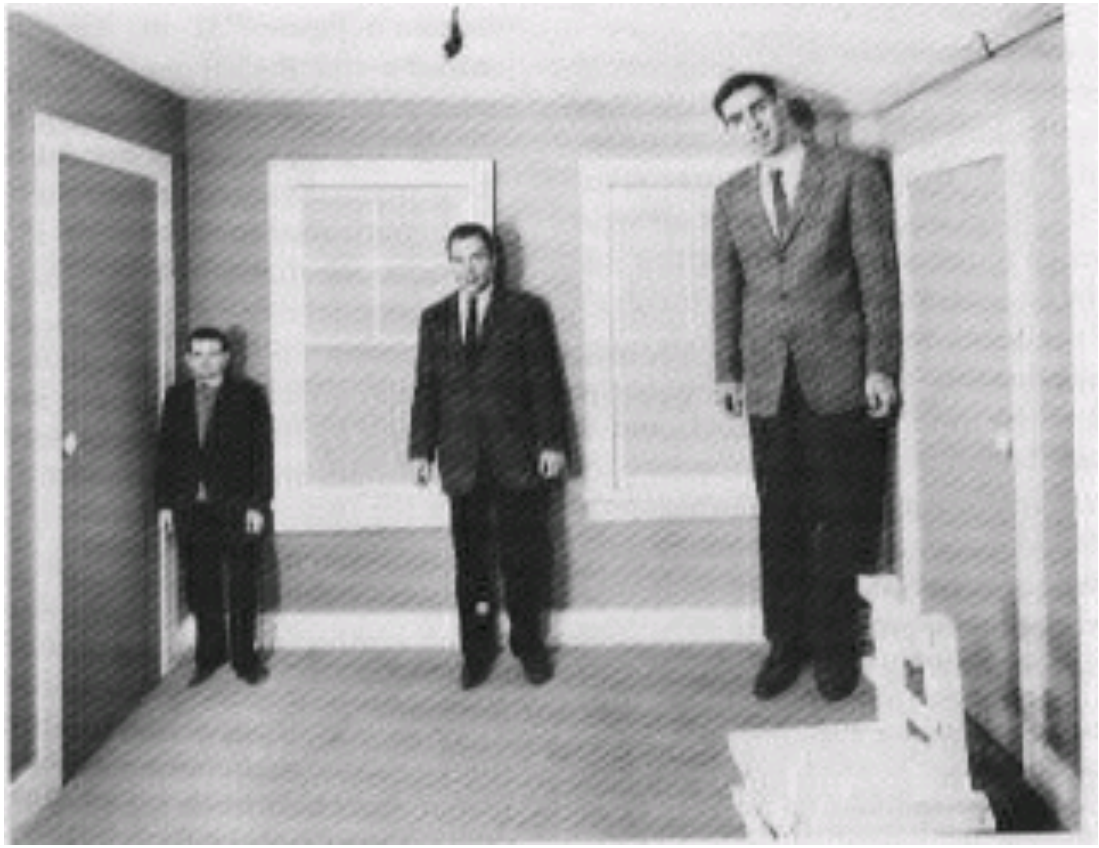


Figure: Ames room

- Assumptions can be wrong

[Souce: A. Torralba]

Is computer vision hard?

Figure: Chabris & Simons

- Count number of times the white team pass the ball
- Concentrate, difficult task!

Is computer vision hard?

Figure: Simons et al.

- Is something happening in the picture?

A bit of history ...

The beginning of Computer Vision ...

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

[Source: A. Torralba]

Vision is hard ...

So let's make the problem more simple

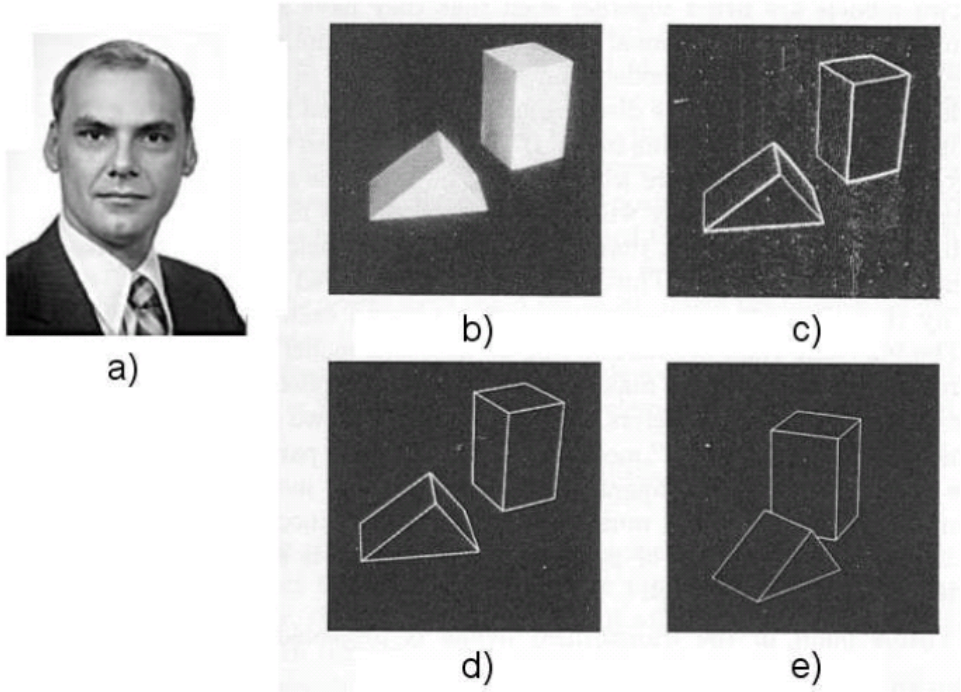


Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

[Source: A. Torralba]

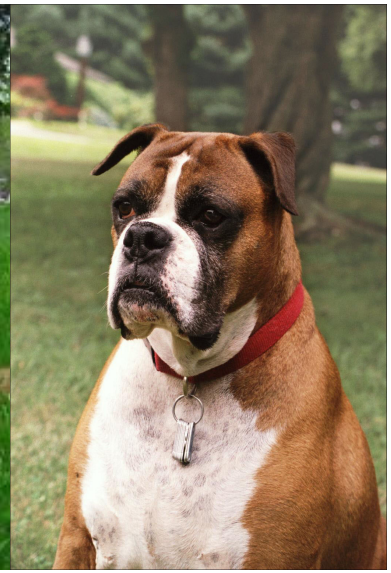
Vision is hard ...

- But, despite promising initial results, things did not work out so well (lack of data, processing power, lack of reliable methods for low-level and mid-level vision)
- Instead, a different way of thinking about object detection started making some progress: learning based approaches and classifiers, which ignored low and mid-level vision.
- Maybe the time is here to come back to some of the earlier models, more grounded in intuitions about visual perception

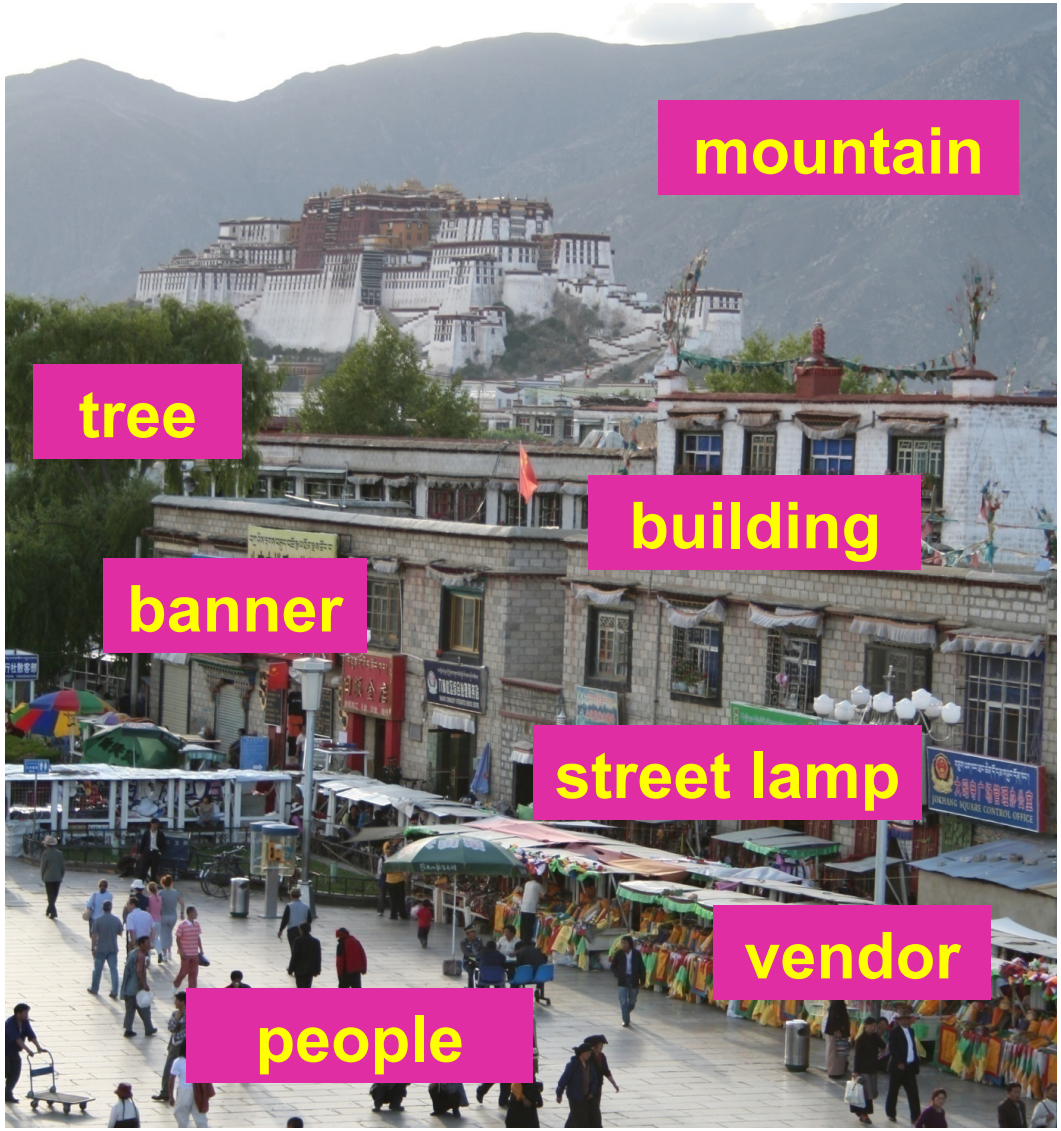
[Source: A. Torralba]

Object Recognition

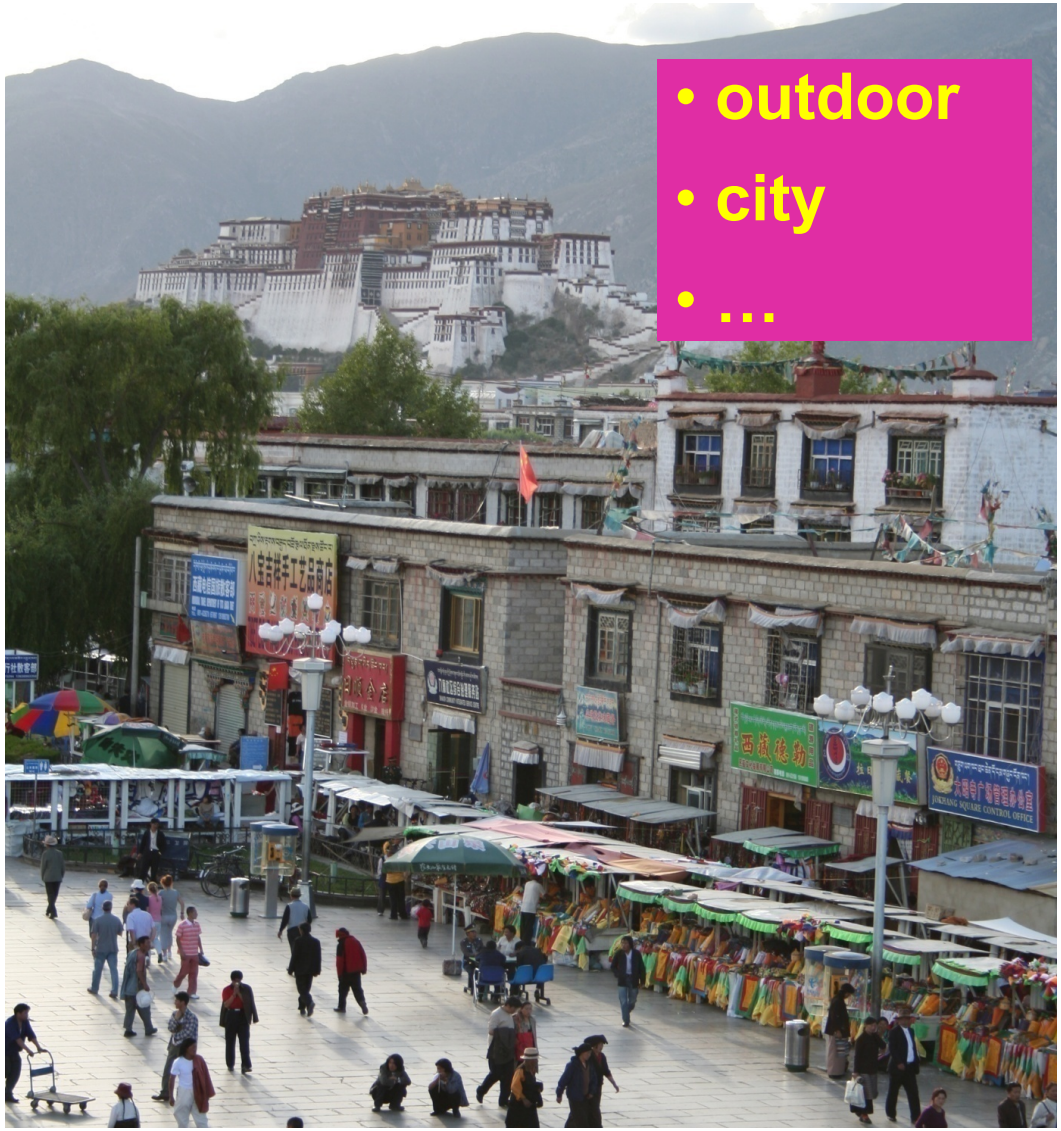
Instance vs category level recognition



Object categorization

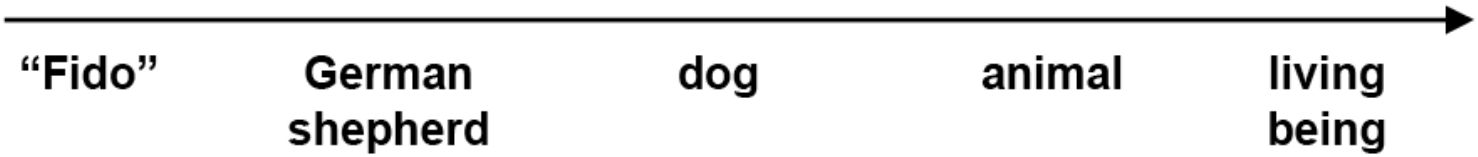


Scene categorization



Object categorization

- Given training examples of a category, recognize instances that you haven't seen before, assigning the correct category level.
- Categories are typically organized hierarchically
- Which ones can we identify visually?
- Basic level (i.e., dog) before identification



[Source: K. Grauman]

Challenges

1) Intra-class variation

[Source: A. Torralba]

Challenges

2) View point variation

Challenges

3) Scale

[Slide credit: R. Fergus]

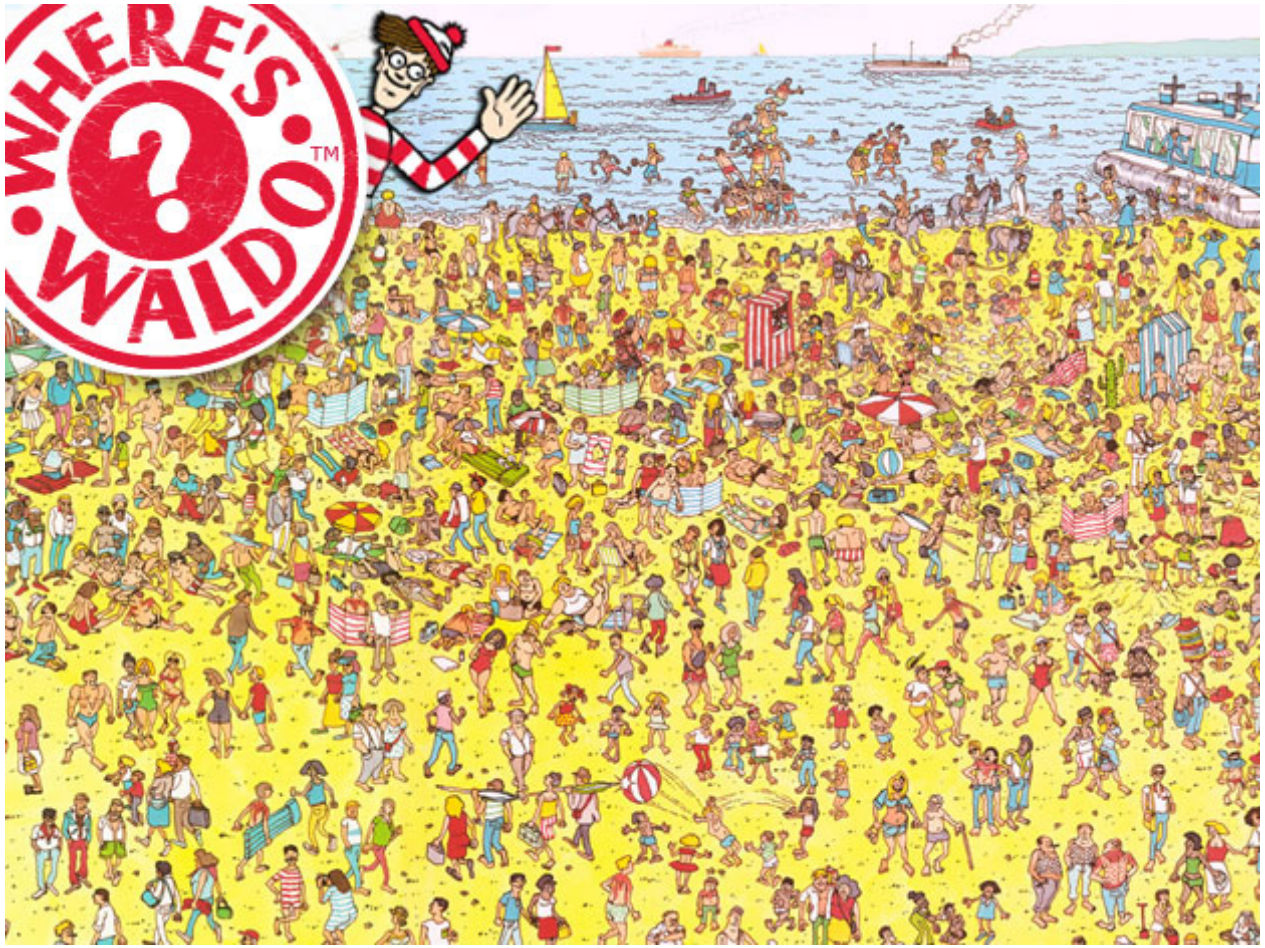
Challenges

4) Illumination

[Slide credit: S. Ullman]

Challenges

5) Background clutter

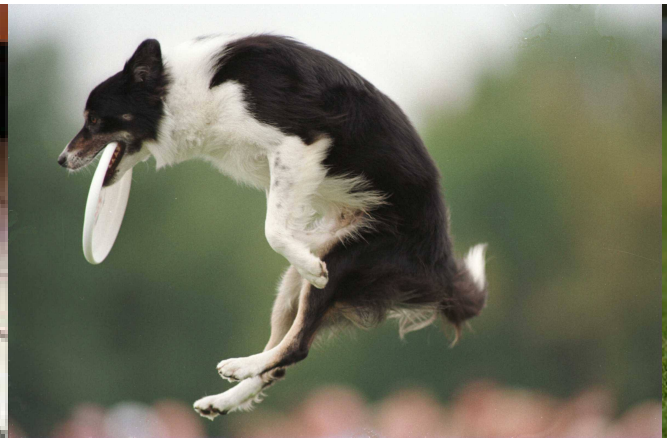


Challenges

6) Occlusion

Challenges

7) Deformation and pose


Challenges

8) Large number of categories

[Biderman 87]

Challenges

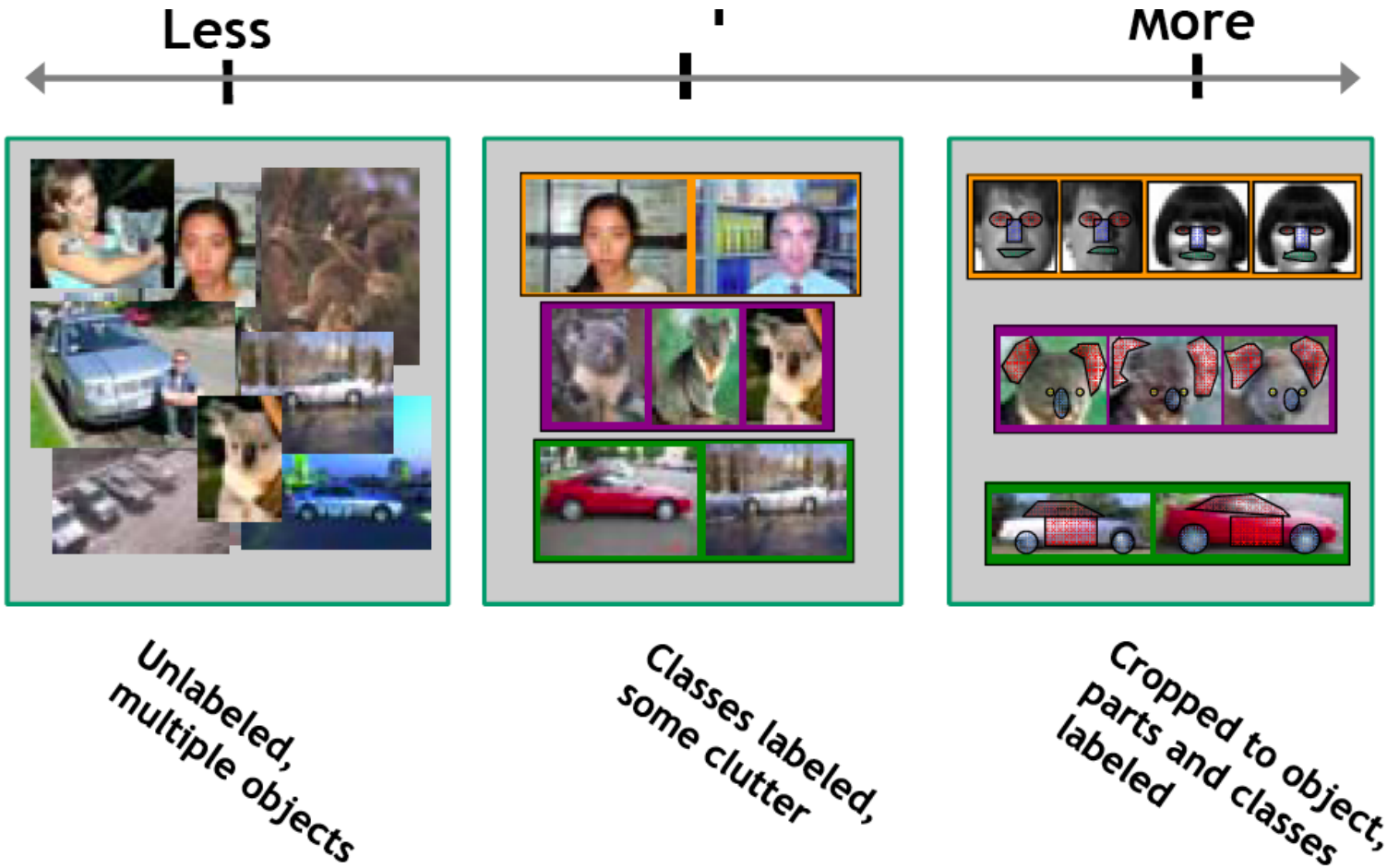
9) Scale and Efficiency

- Half of the cerebral cortex in primates is devoted to processing visual information
- $\sim$ 20 hours of video added to YouTube per minute
- $\sim$ 5000 new tagged photos added to Flickr per minute
- Thousands to millions of pixels in an image
- 30+ degrees of freedom in 2D and 80+ in 3D articulated pose
- 3,000-30,000 human recognizable object categories

[Source: K. Grauman]

Challenges

10) Learning with minimal supervision

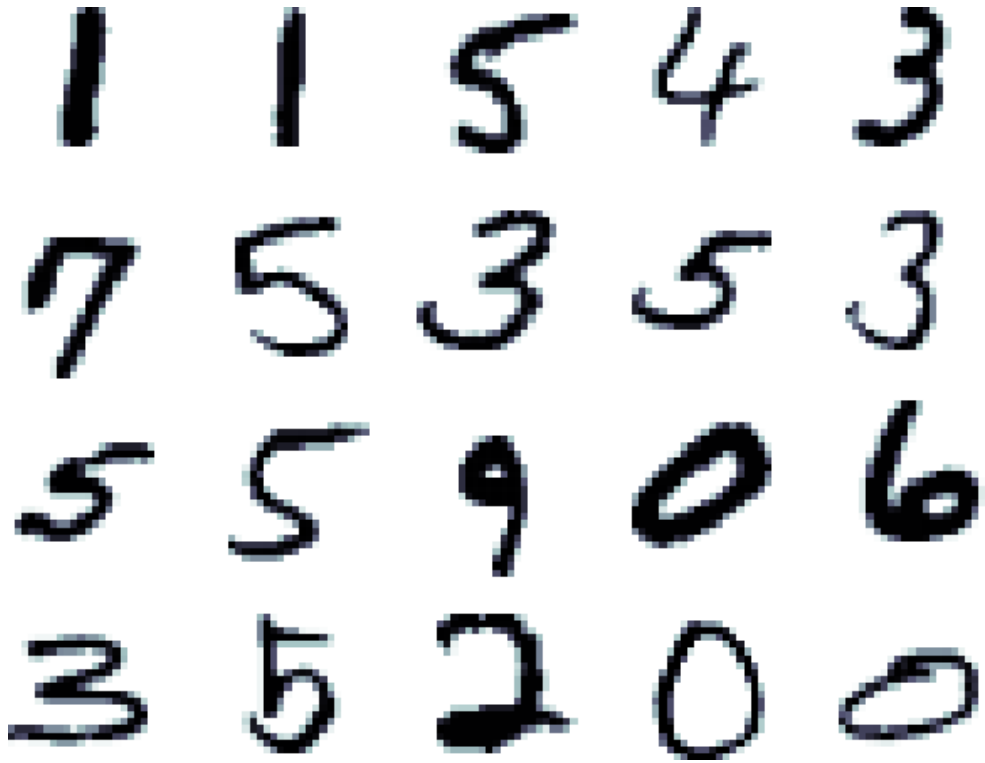


[Source: K. Grauman]

When does computer vision work?

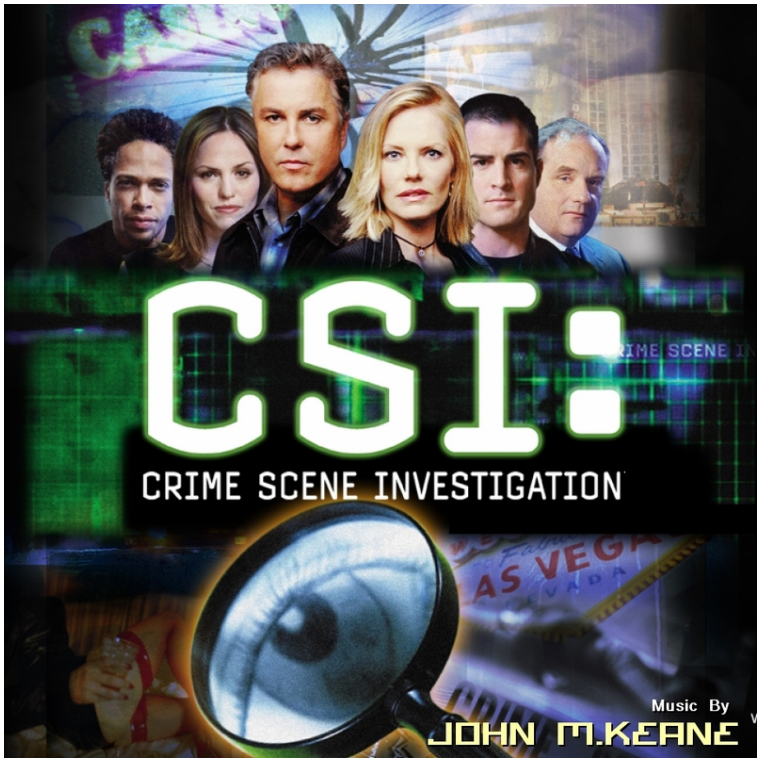
Applications: Digit recognition

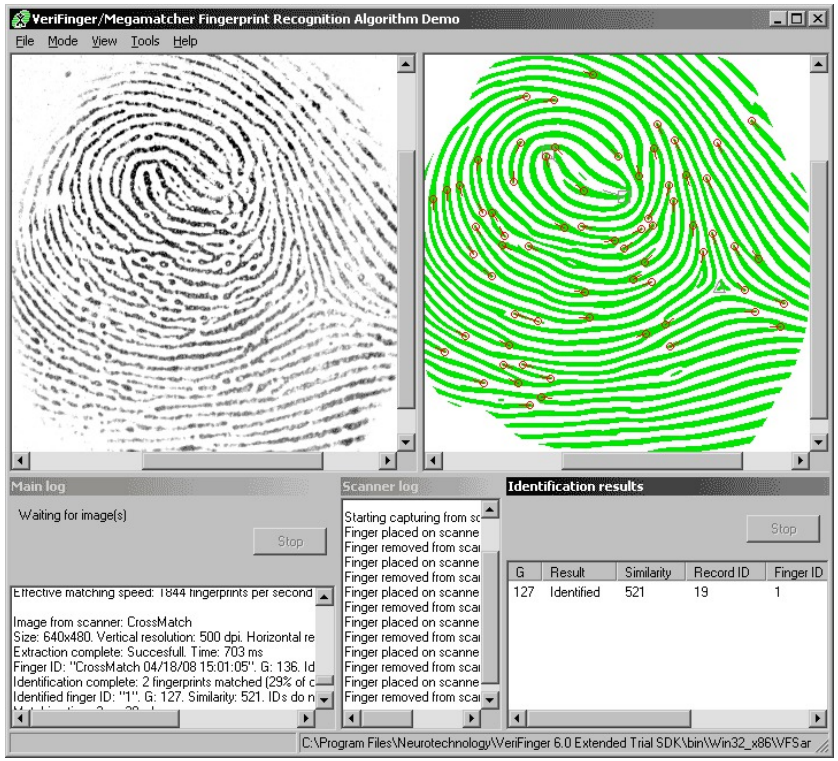
- Reading license plates, zip codes, checks, etc



[Source: S. Lazebnik]

Applications: Fingerprint Recognition

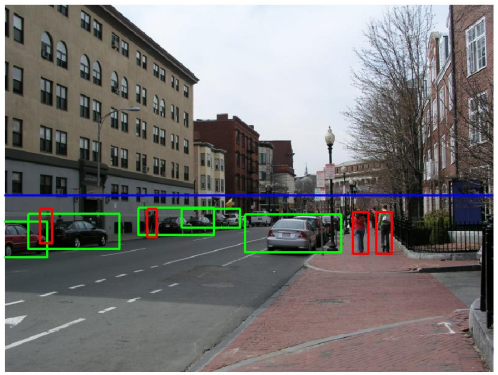
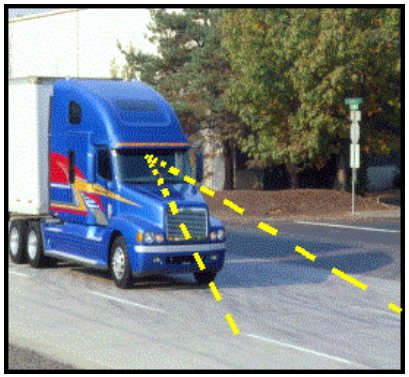




[Source: S. Lazebnik]

Applications: Assisted Driving

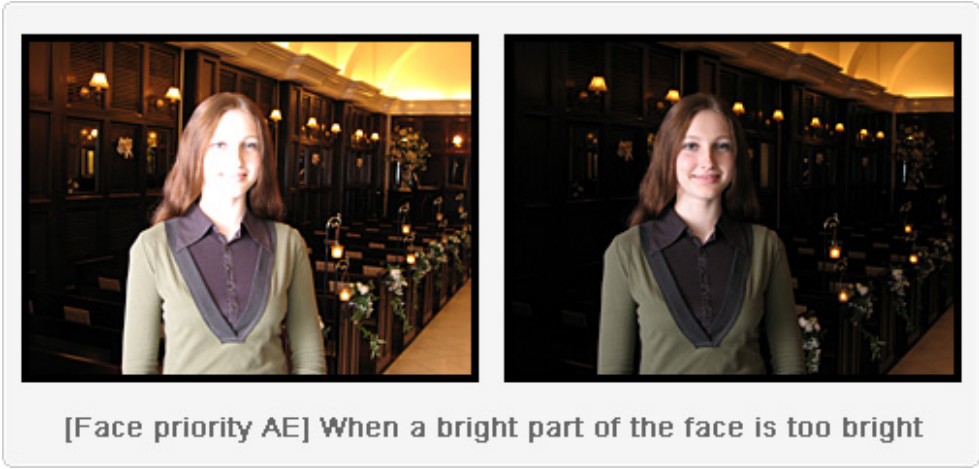
- Pedestrian and car detection
- Lane detection and lane departure warning
- Collision warning systems with adaptive cruise control


[Source: R. Fergus]

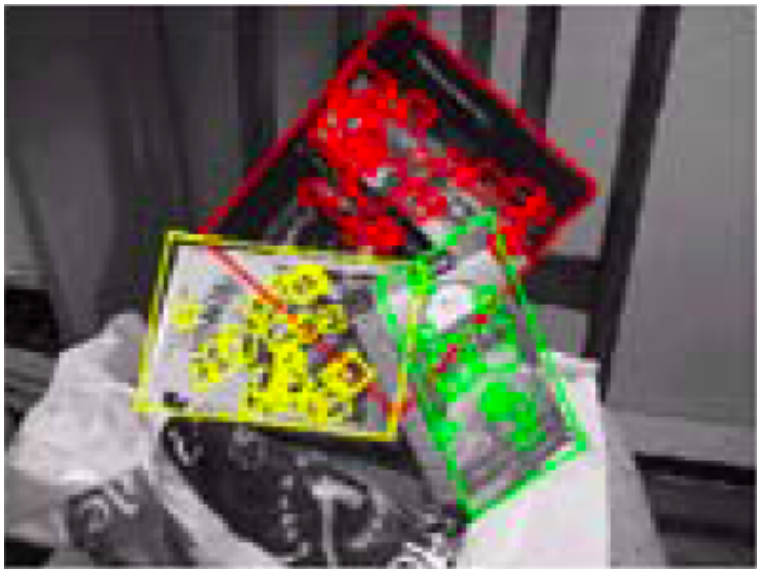
Applications: Computational Photography

Face recognition and image editing



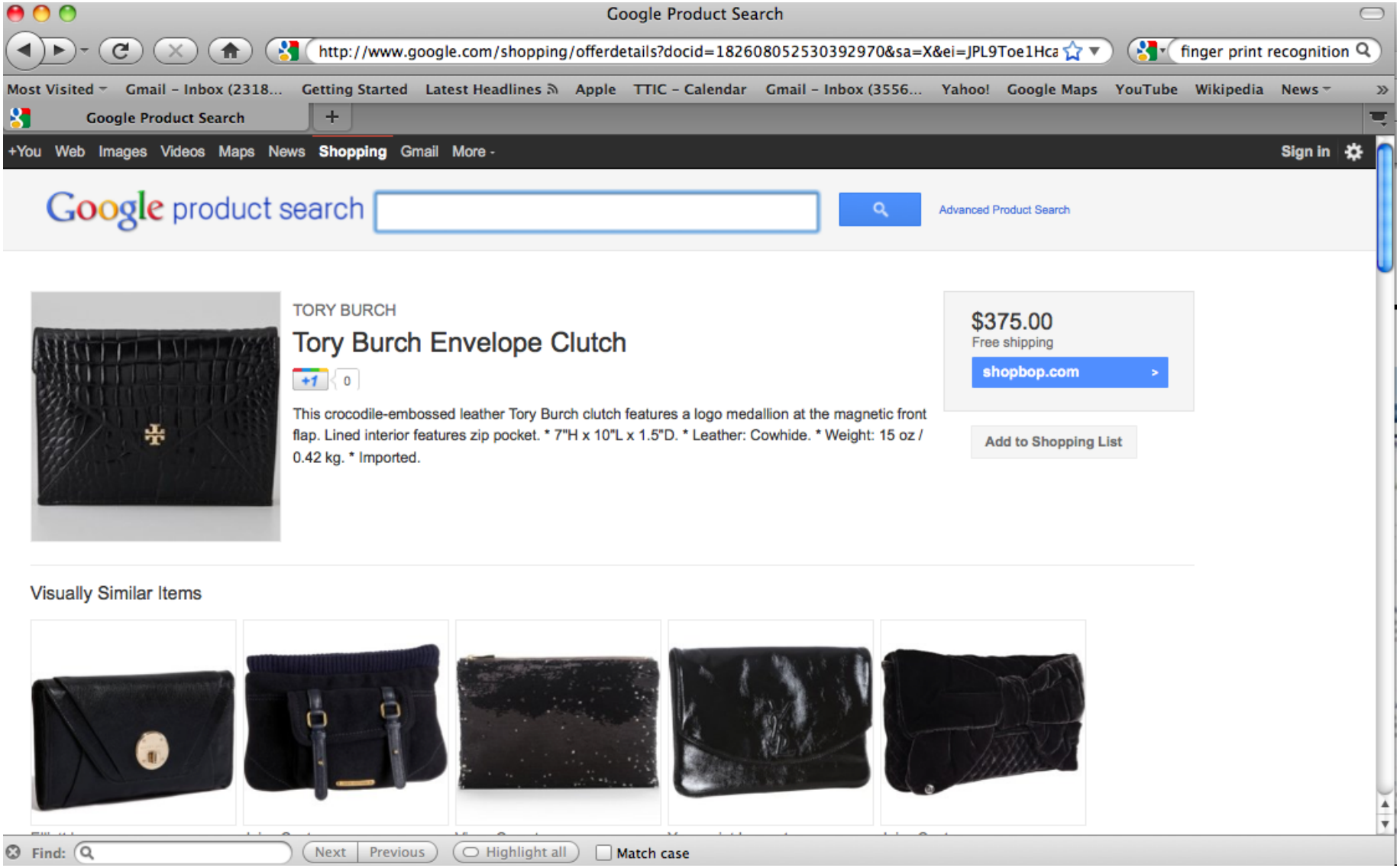
[Source: R. Fergus]

Applications: Recognizing CD covers

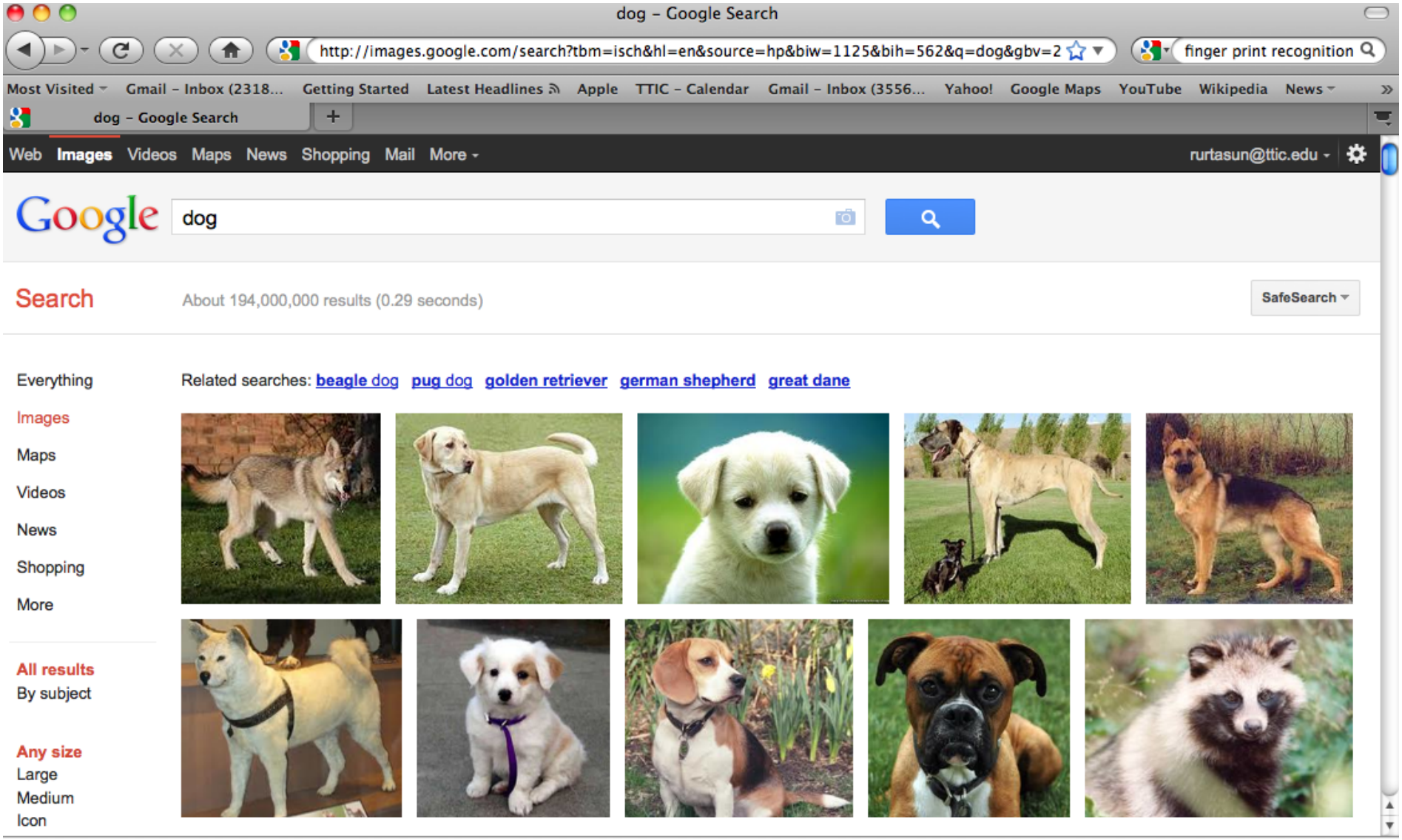


[Source: S. Lazebnik]

Applications: Finding Similar Products



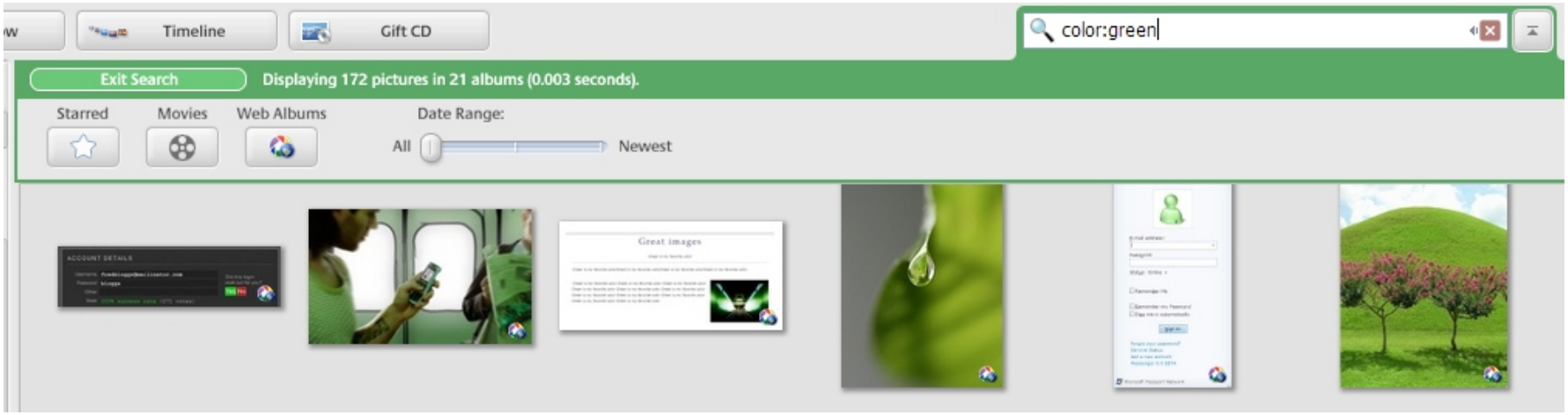
Applications: Search



Applications: Localization and Identification

[Source: Google Googles]

Applications: Organizing Photo Collections

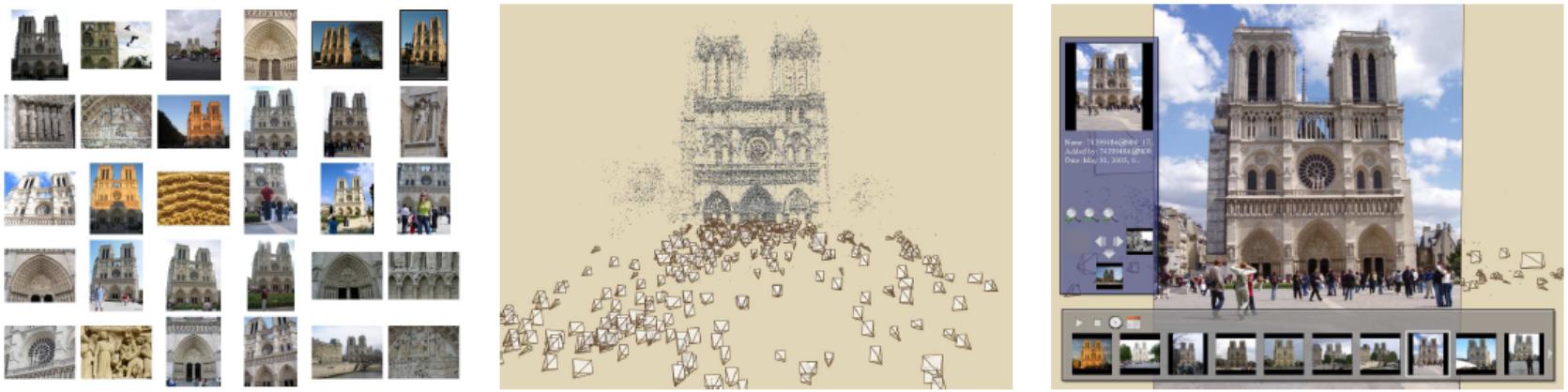


[Source: R. Fergus]

Applications: 3D Pose Estimation with Depth Sensors

[Source: Microsoft Kinect]

Applications: 3D Reconstruction from Photo Collections

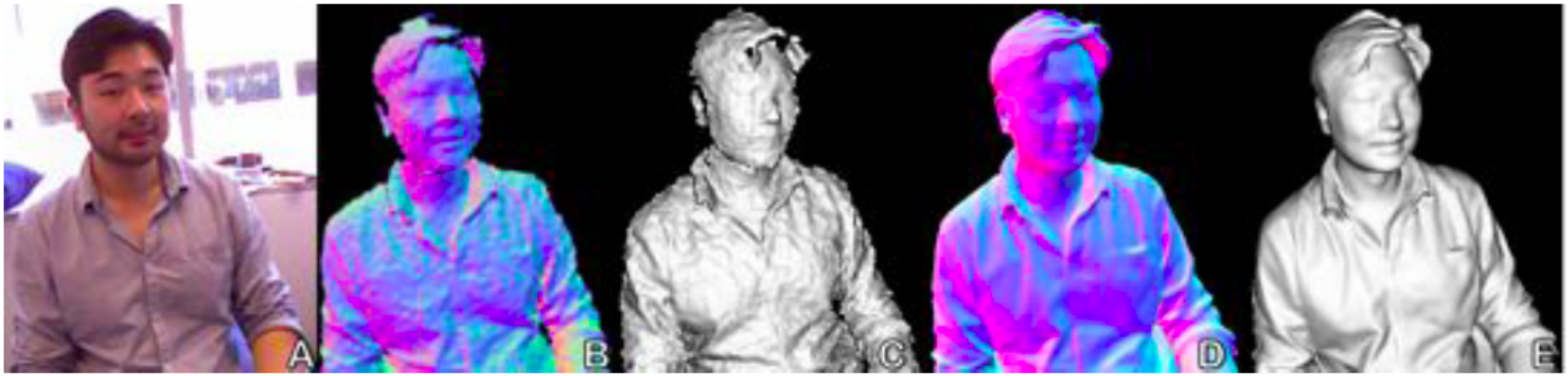


(a) (b) (c)

Figure 1: Our system takes unstructured collections of photographs such as those from online image searches (a) and reconstructs 3D points and viewpoints (b) to enable novel ways of browsing the photos (c).

[N. Snavely et al. Siggraph 2006]

Applications: 3D Reconstruction from Depth Cameras



[Izadi et al. Siggraph 2011]

What would we like to do with computer vision?

Future Applications

- Visual recognition is an important part of perception, 33% of the cortex is devoted to vision in humans and 50% in monkeys!
- Organize and access visual information
- Discover something about how humans do vision

What about this class ...?

Class Organization

The class consists in ...

- 3h/week of theoretical aspects
- Research and practical work

Grading

- Exam: 35%
- Research Project: 65%

Research Project

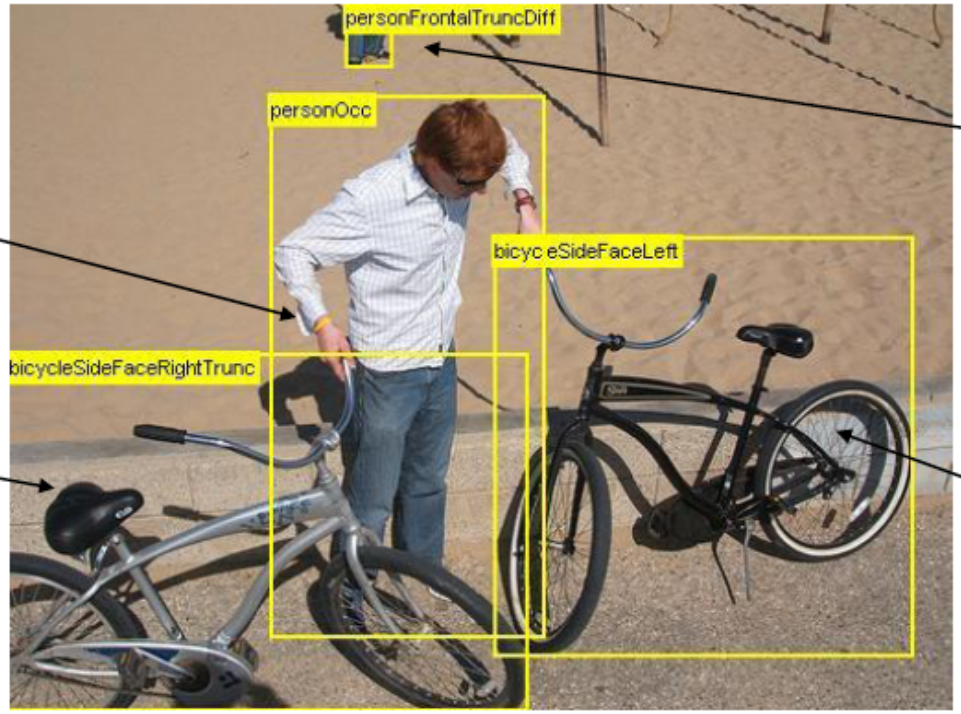
- Can be done individually or in groups of two.
- Can be related to your research.
- Novel research in the field: new idea, or extending existing papers.
- Ask a time to meet with me (this week!) so that we decide what would you do.

In this class ...

We will "hopefully" cover the following topics:

- Object Recognition and Scene Categorization
- Object Detection
- Segmentation
- 2D and 3D Pose Estimation
- Modern 3D Reconstruction

Classification



Occluded
Object is significantly occluded within BB

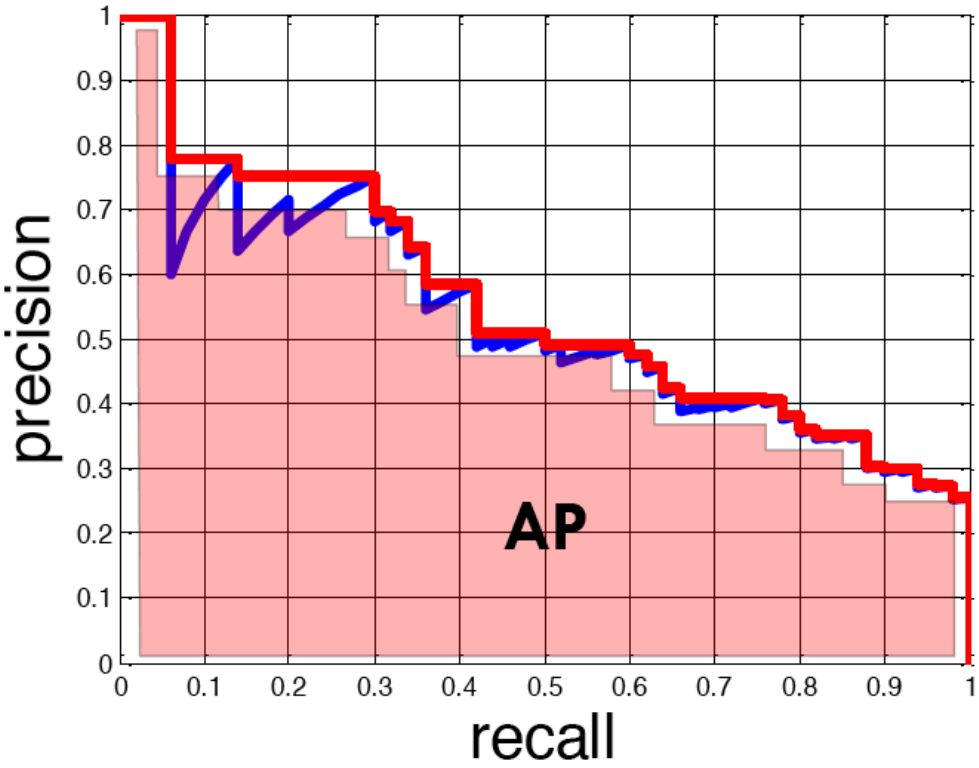
Difficult
Not scored in evaluation

Truncated
Object extends beyond BB

Pose
Facing left

- PASCAL VOC is main challenge: 10,000 training and 10,000 testing
- 20 categories
- Say if that image has a particular object

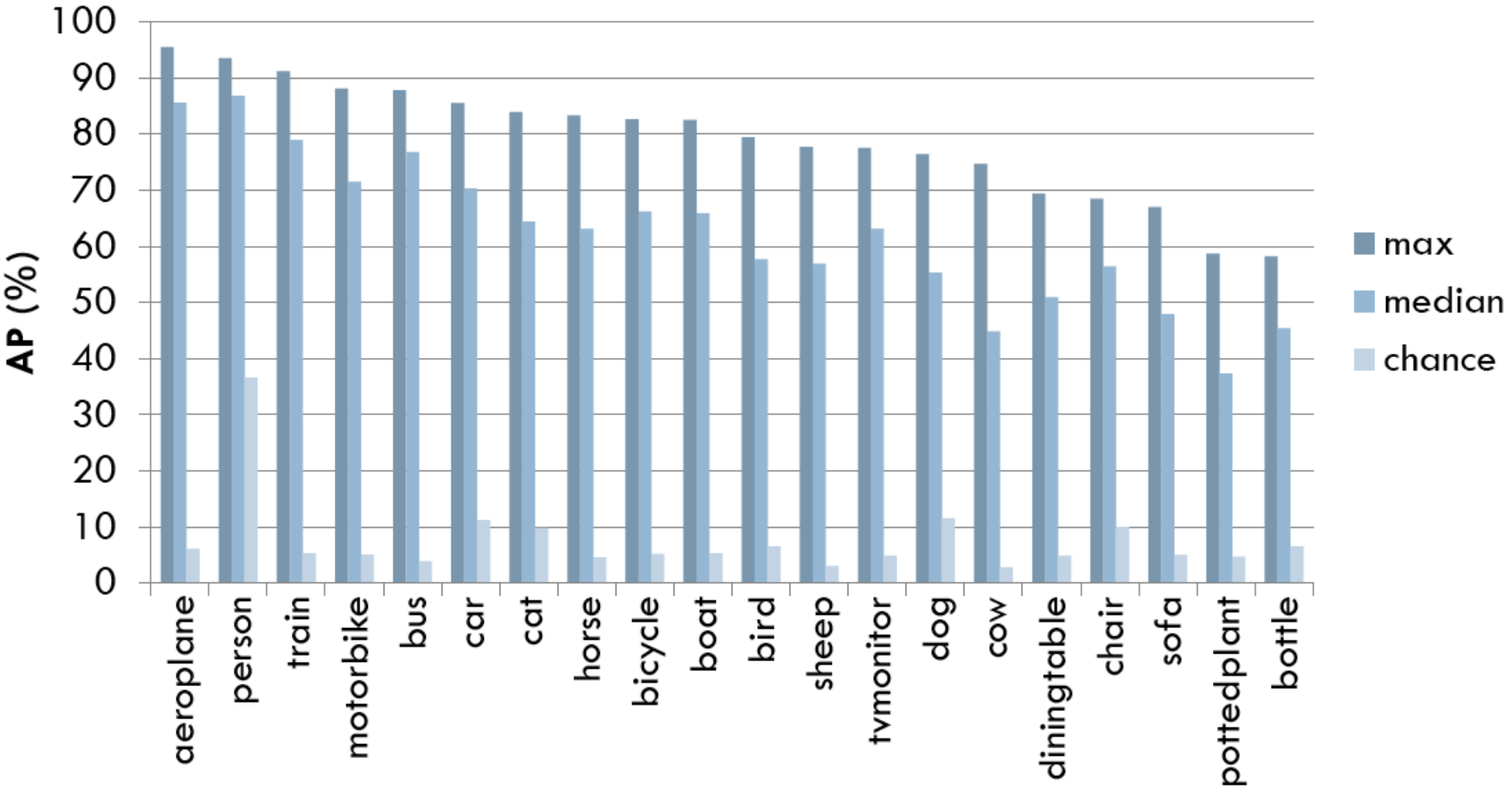
Classification



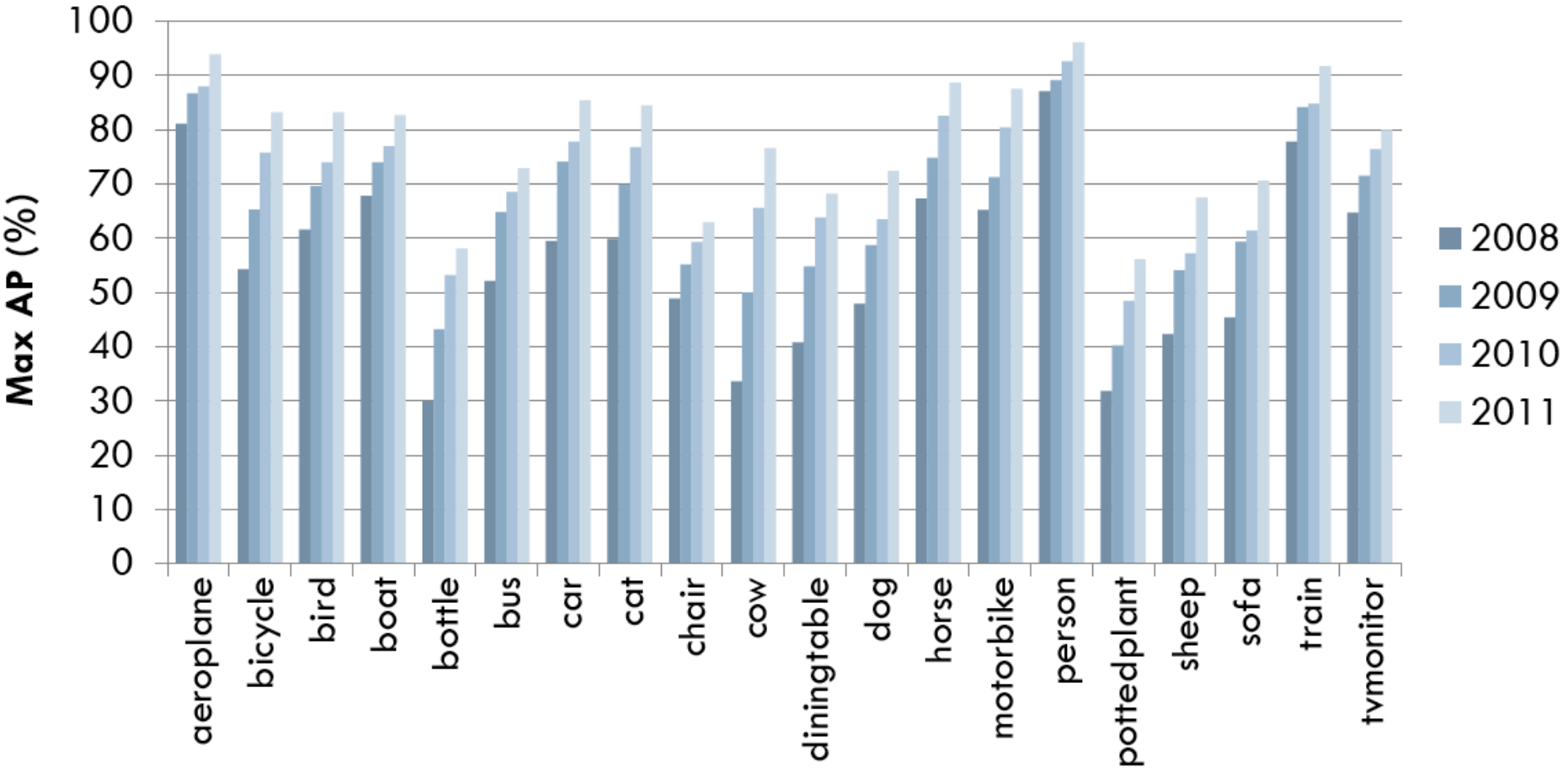
- Evaluated with Average Precision

$$P = \frac{tp}{tp + fp} \qquad R = \frac{tp}{tp + fn}$$

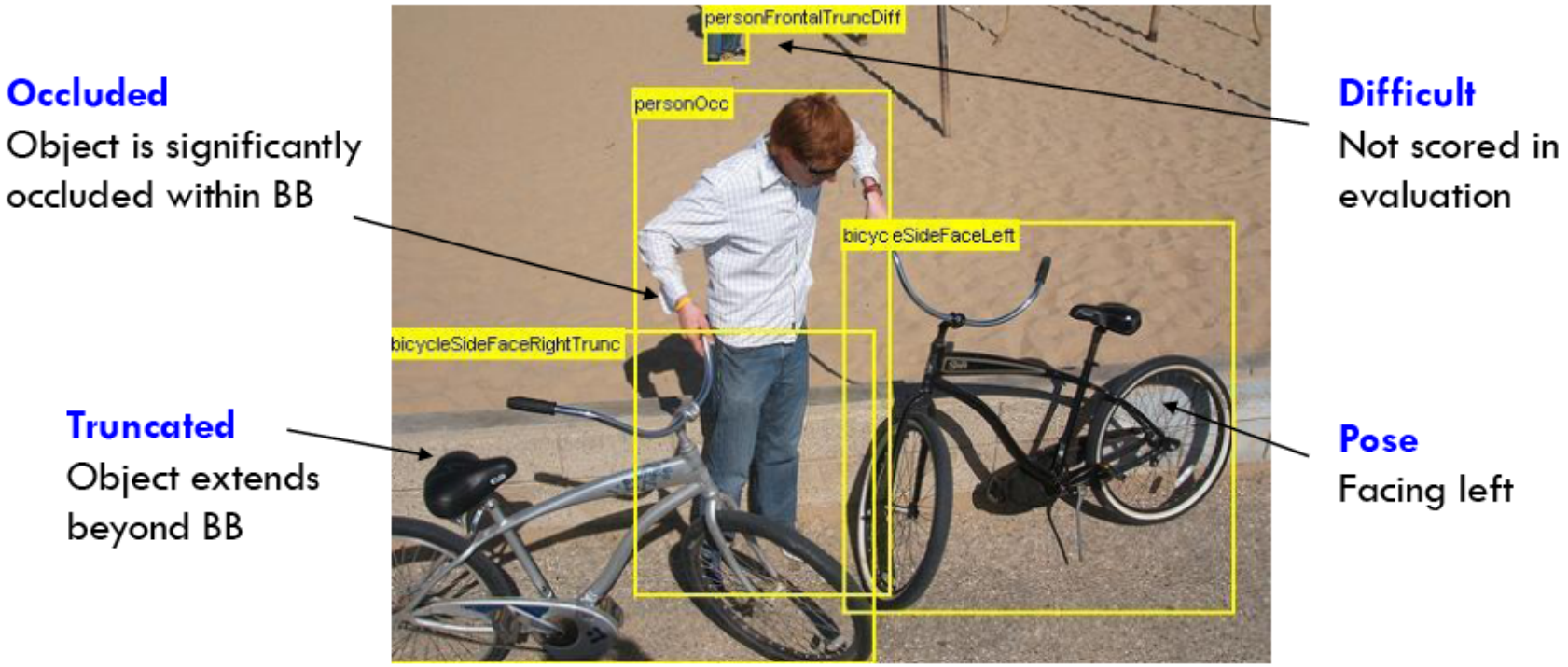
How good are we?



But we are improving...

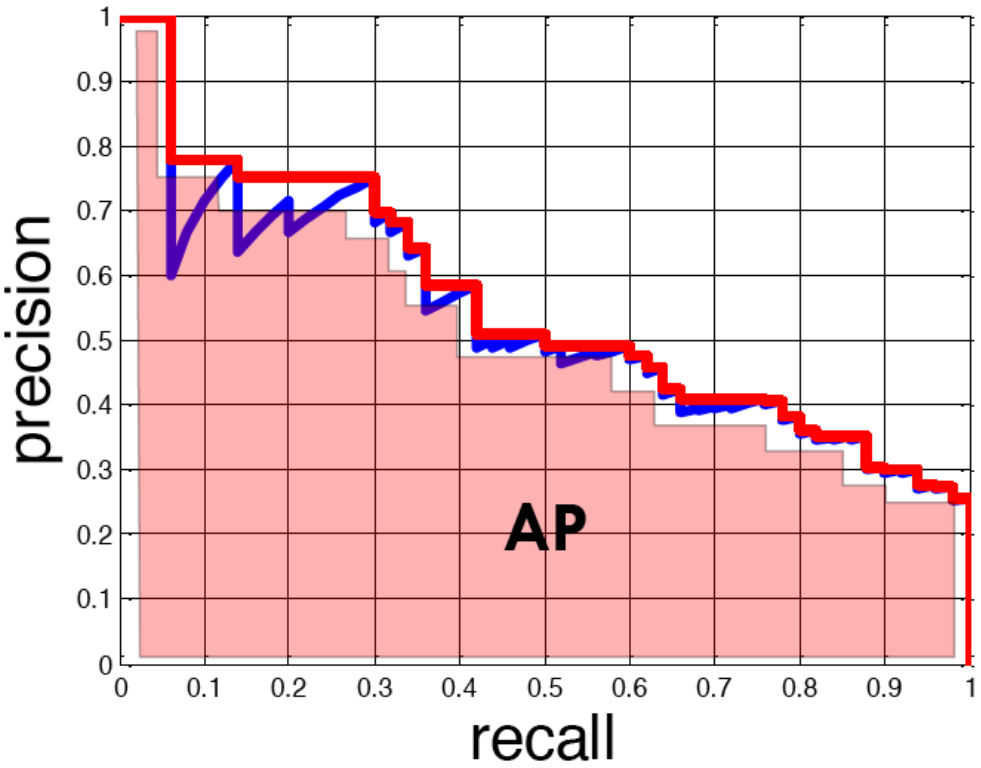


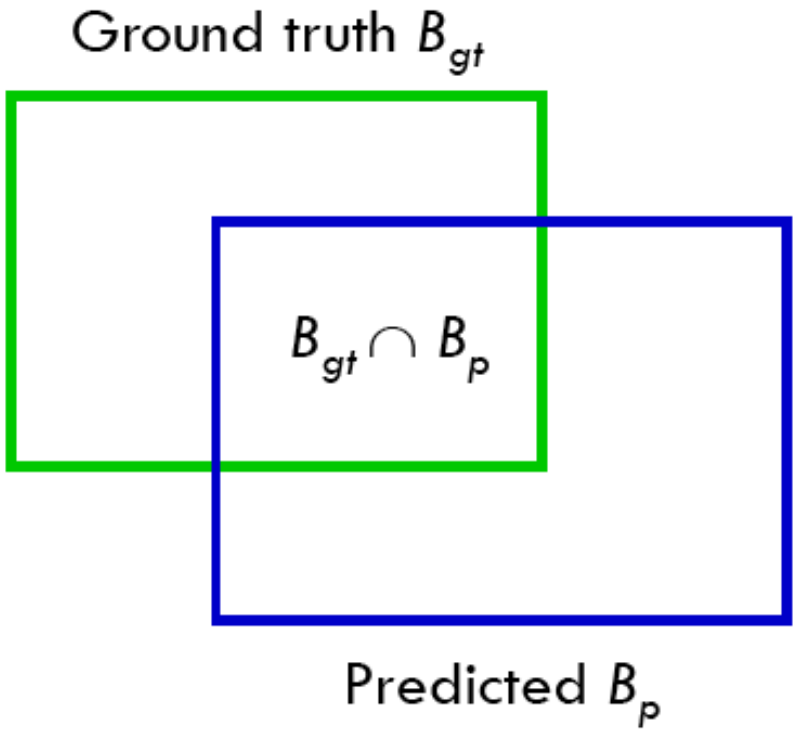
Detection



- PASCAL VOC is main challenge: 10,000 training and 10,000 testing
- Say where is the object
- 20 categories

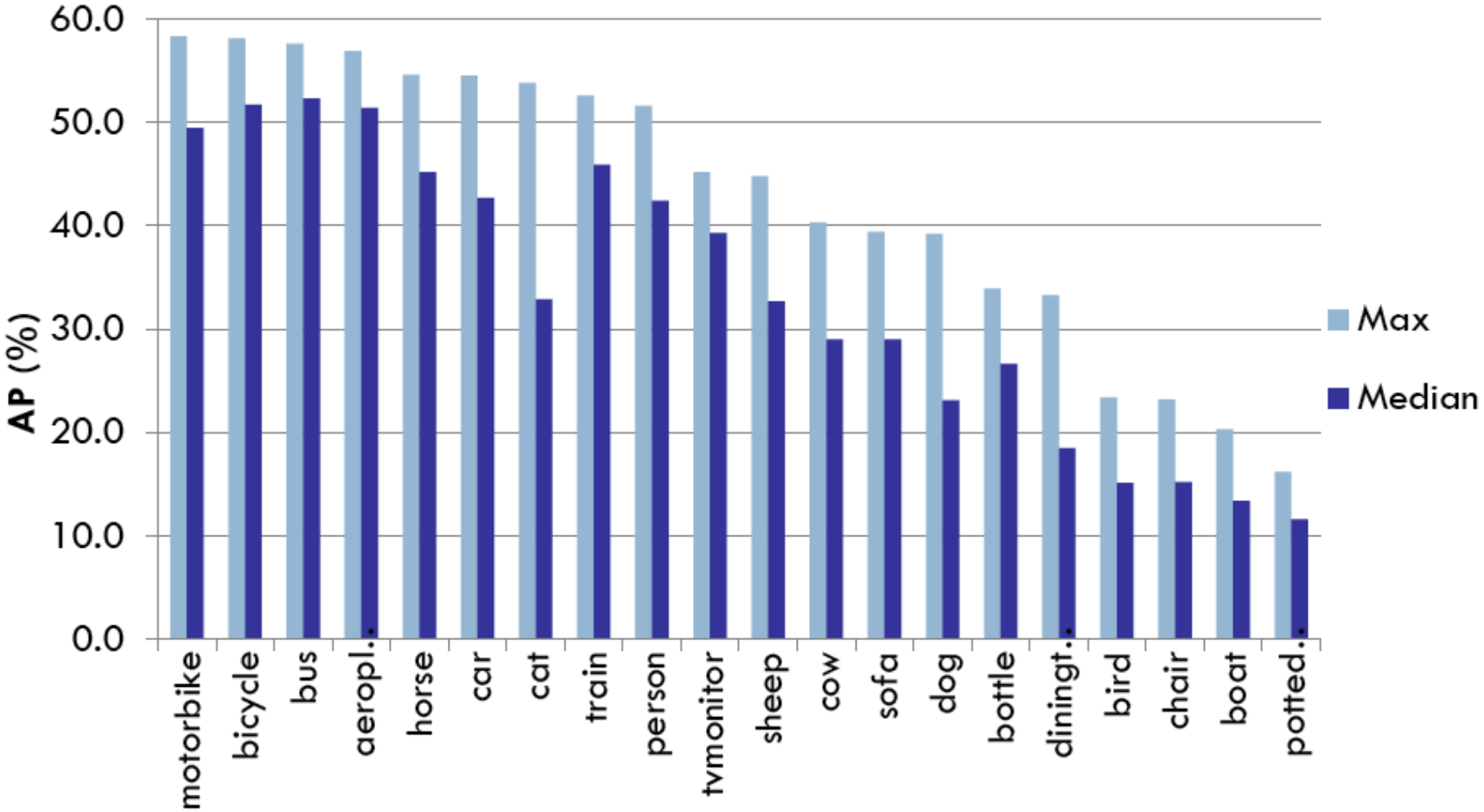
Detection



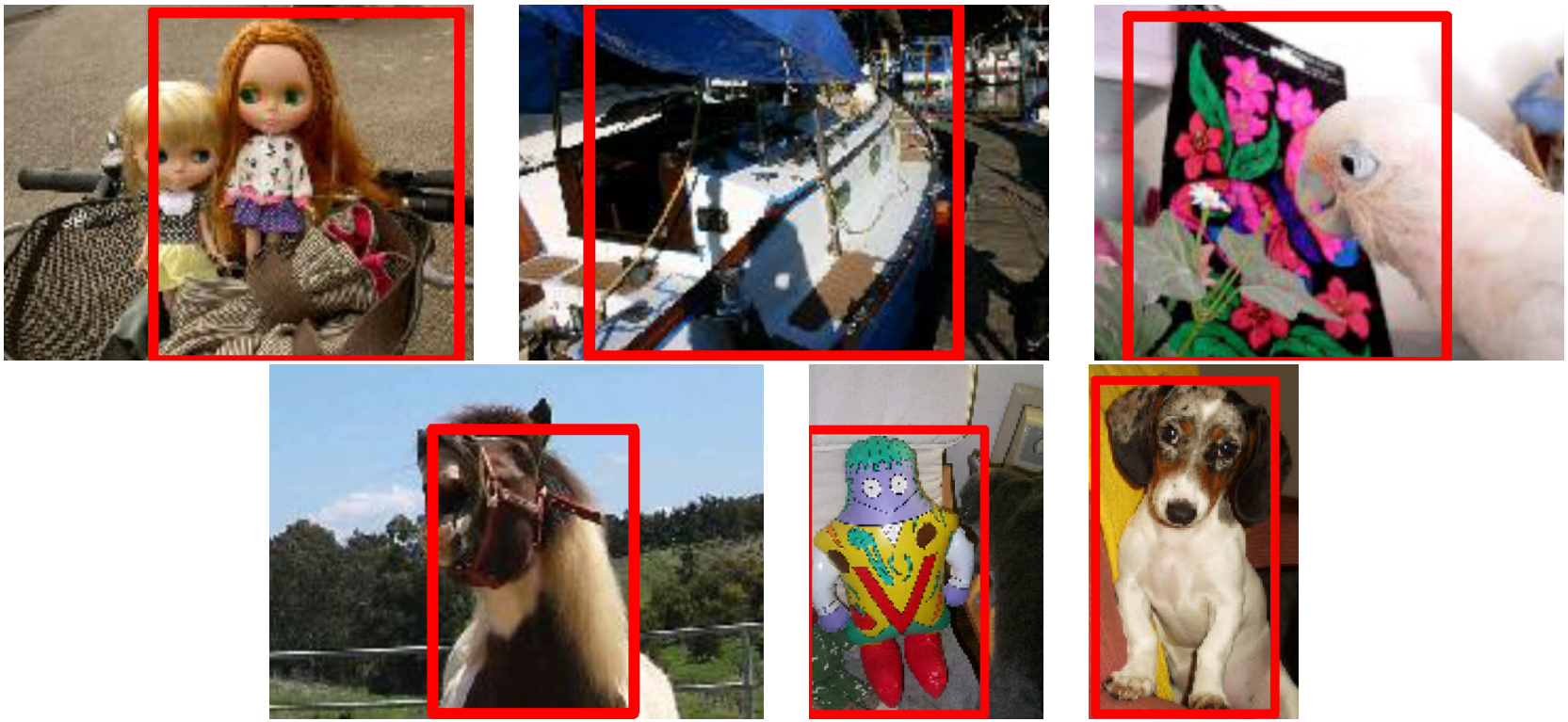


- Evaluated with Average Precision
- Detected if intersection over union is more than 50%.

How good are we?

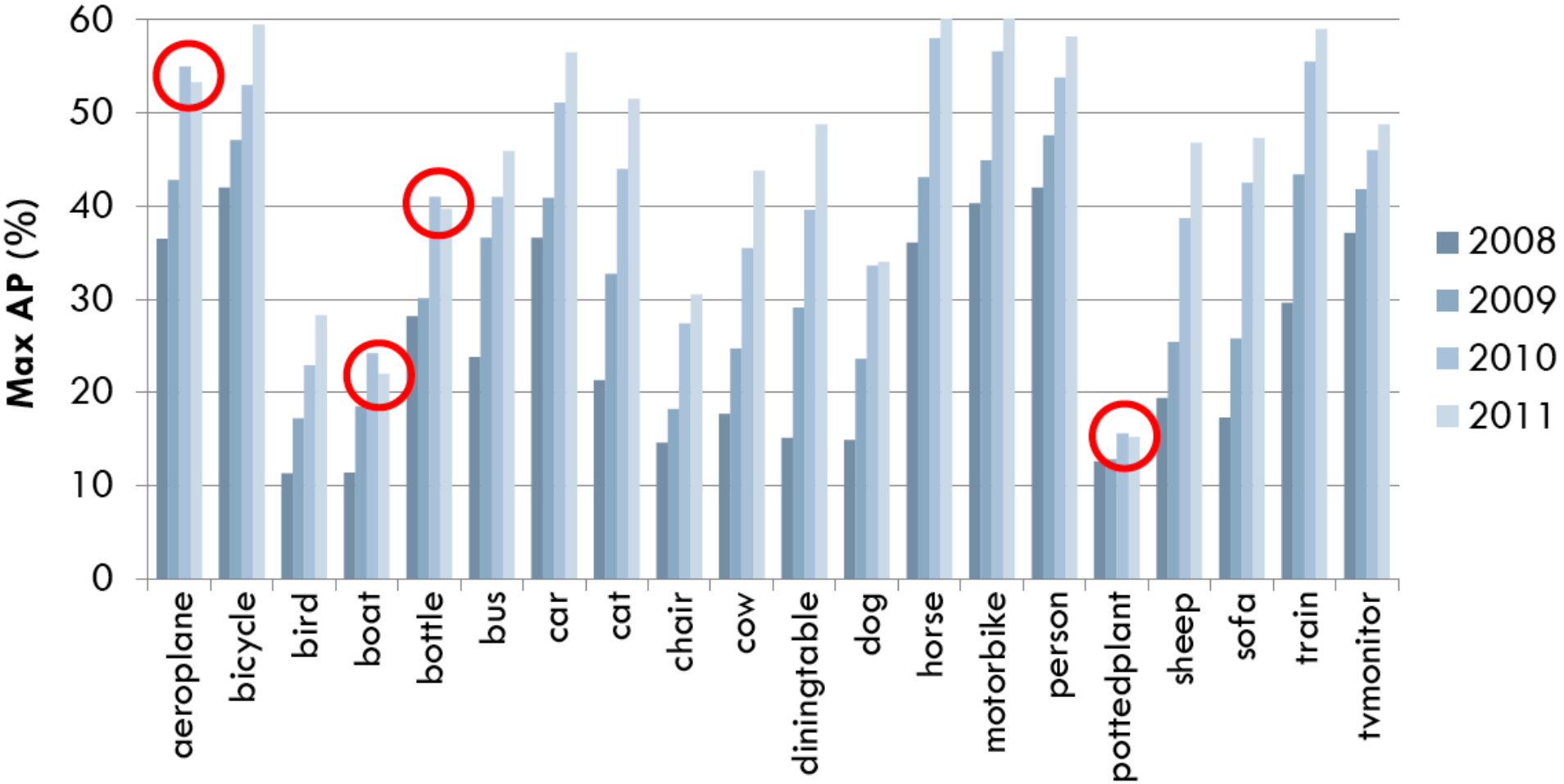


Still room for improvement ...

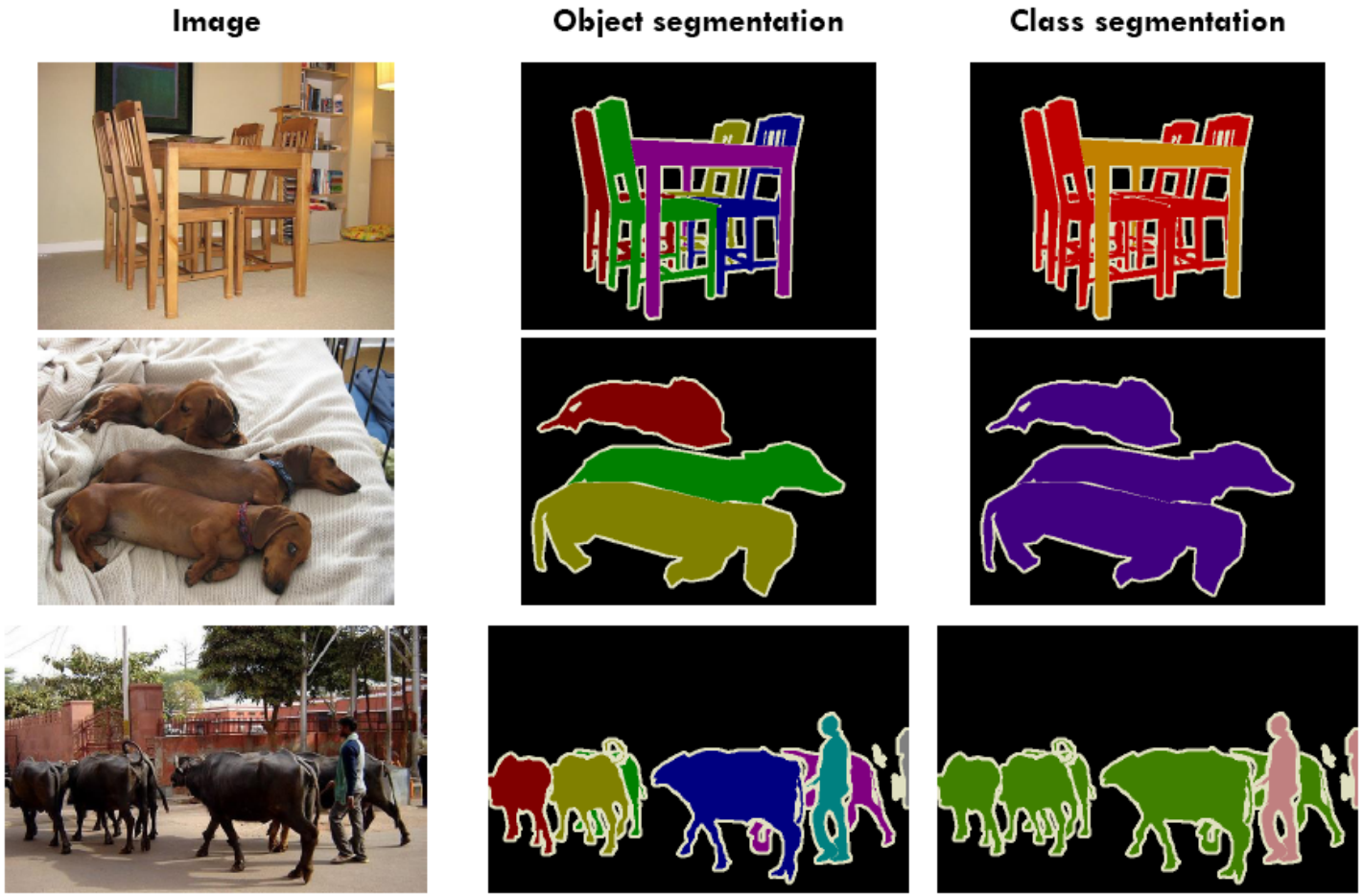


- Which category is this?

But we are improving...



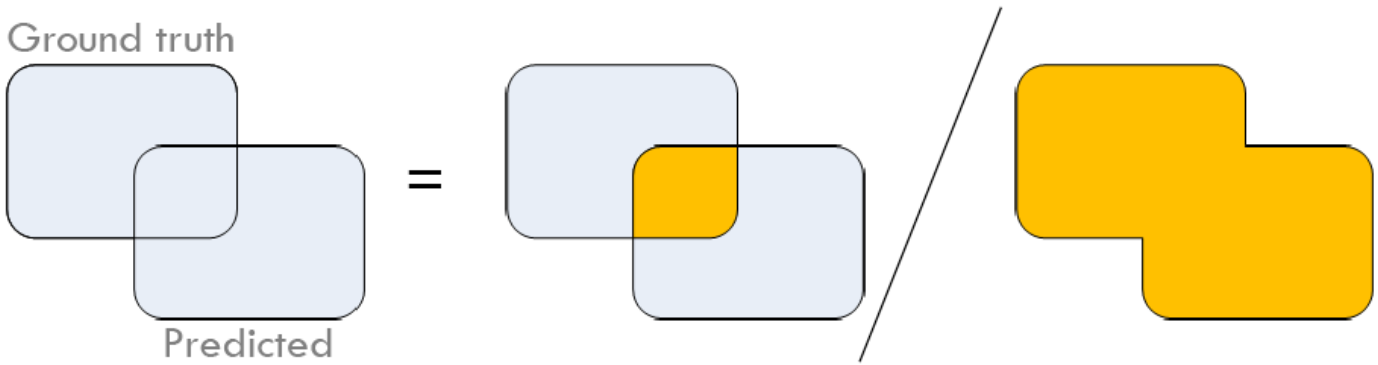
Segmentation



- PASCAL VOC is main challenge: 2000 training, 1000 testing
- Precise delimitation of the pixels
- 20 categories

Segmentation

$$\text{Intersection/union of class labels} = \frac{\text{true pos.class}}{\text{true pos.} + \text{false pos.} + \text{false neg.}}$$



- Evaluated with Intersection over the union.
- Penalizes under and over segmentations

How good are we?

	[mean]	back ground	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ Monitor
BONN_FGT_SEGM	41.4	83.4	51.7	23.7	46.0	33.9	49.4	66.2	56.2	41.7	10.4	41.9	29.6	24.4	49.1	50.5	39.6	19.9	44.9	26.1	40.0	41.6
BONN_SVR_SEGM	43.3	84.9	54.3	23.9	39.5	35.3	42.6	65.4	53.5	46.1	15.0	47.4	30.1	33.9	48.8	54.4	46.4	28.8	51.3	26.2	44.9	37.2
BROOKES_STRUCT _DET_CRT	31.3	79.4	36.6	18.6	9.2	11.0	29.8	59.0	50.3	25.5	11.8	29.0	24.8	16.0	29.1	47.9	41.9	16.1	34.0	11.6	43.3	31.7
NUS_CONTEXT_SVM	35.1	77.2	40.5	19.0	28.4	27.8	40.7	56.4	45.0	33.1	7.2	37.4	17.4	26.8	33.7	46.6	40.6	23.3	33.4	23.9	41.2	38.6
NUS_SEG_DET_MASK _CLS_CRF	37.7	79.8	41.5	20.2	30.4	29.1	47.4	61.2	47.7	35.0	8.5	38.3	14.5	28.6	36.5	47.8	42.5	28.5	37.8	26.4	43.5	45.8

Some good ...

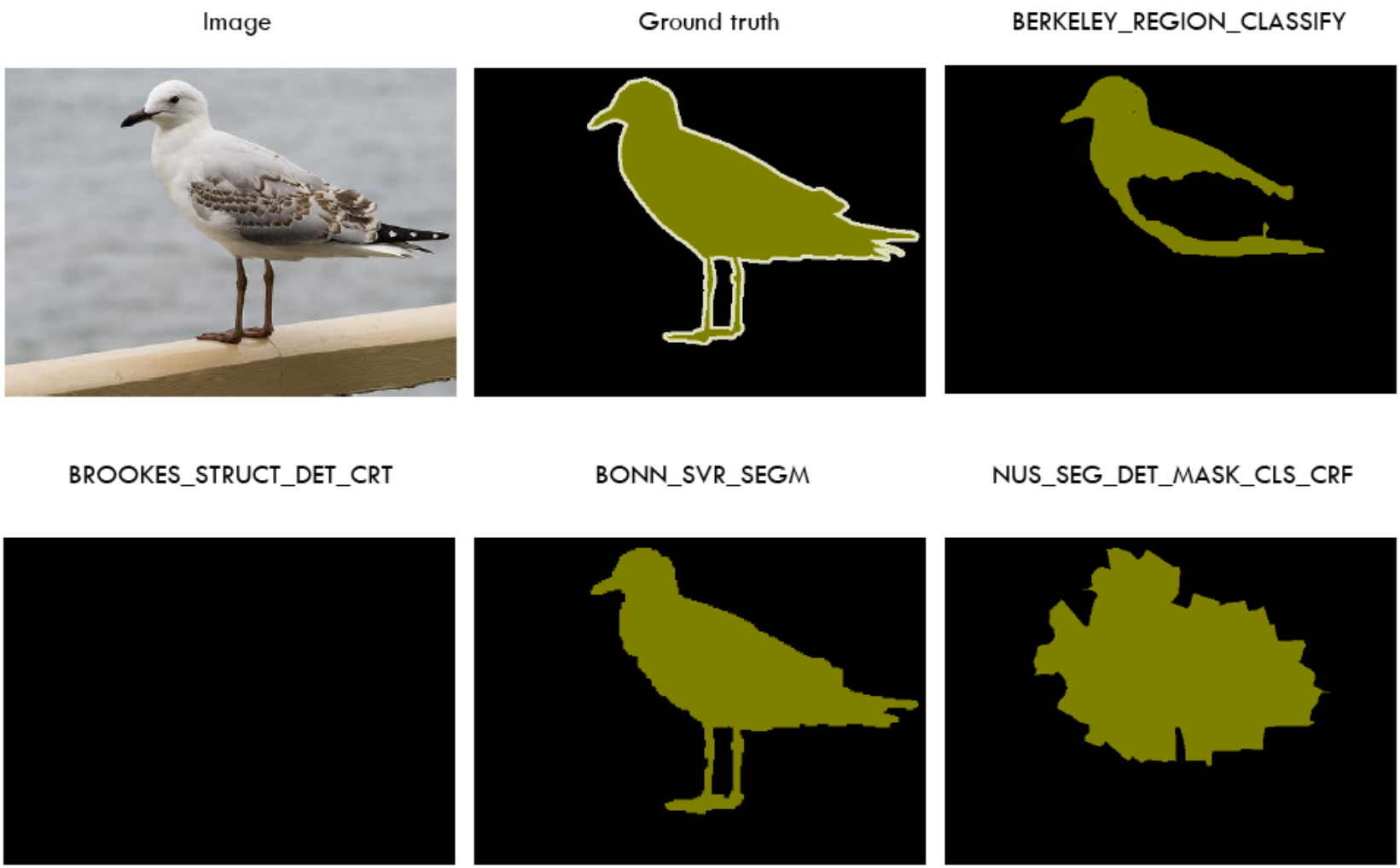


Image
Ground truth
BERKELEY_REGION_CLASSIFY
BROOKES_STRUCT_DET_CRT
BONN_SVR_SEGM
NUS_SEG_DET_MASK_CLS_CRF

Most bad ...

Image

Ground truth

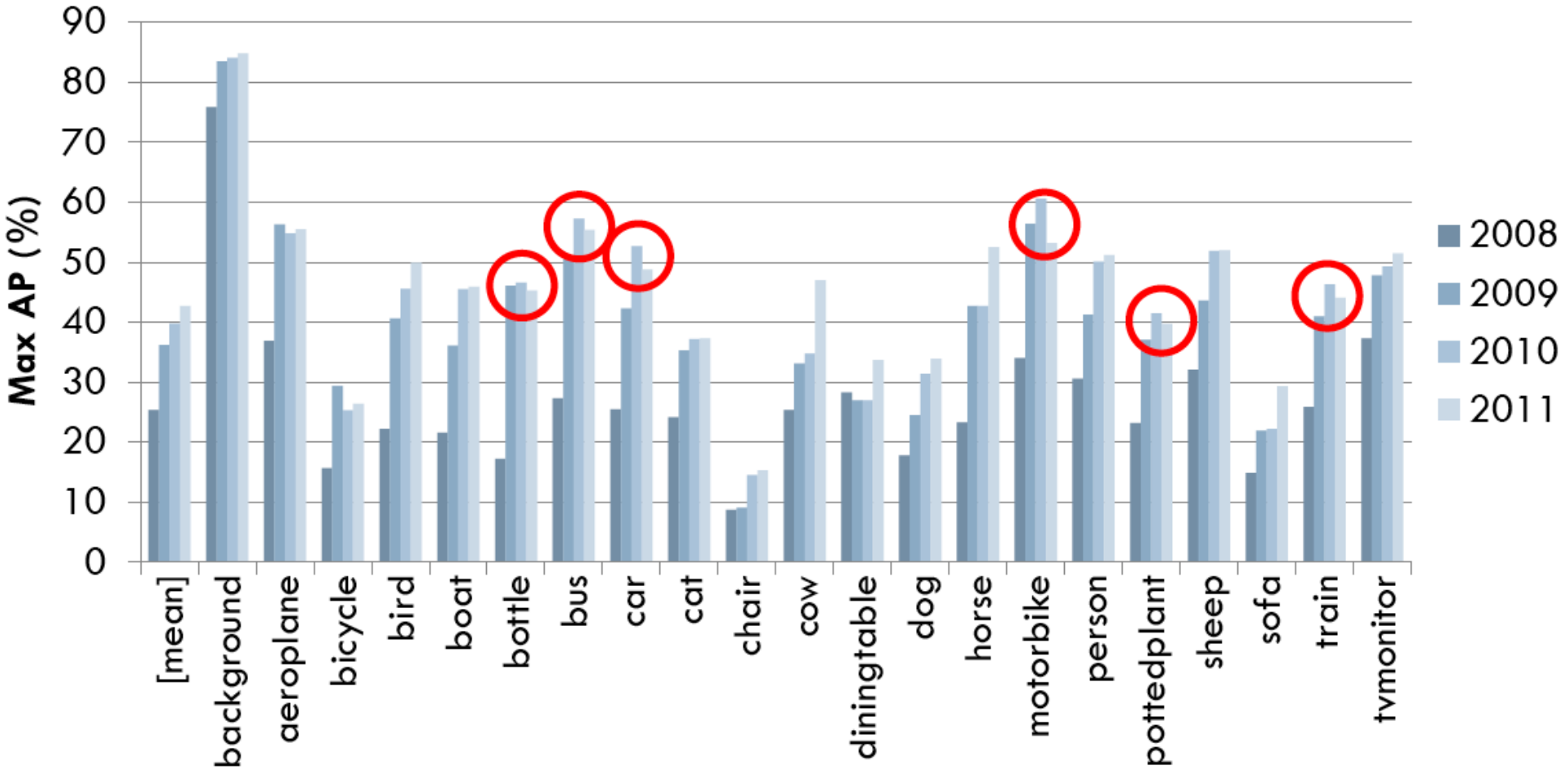
BERKELEY_REGION_CLASSIFY

BROOKES_STRUCT_DET_CRT

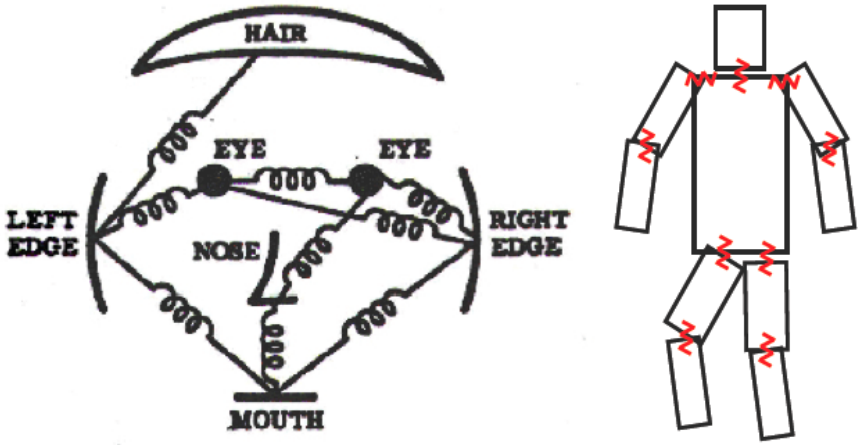
BONN_SVR_SEGM

NUS_SEG_DET_MASK_CLS_CRF

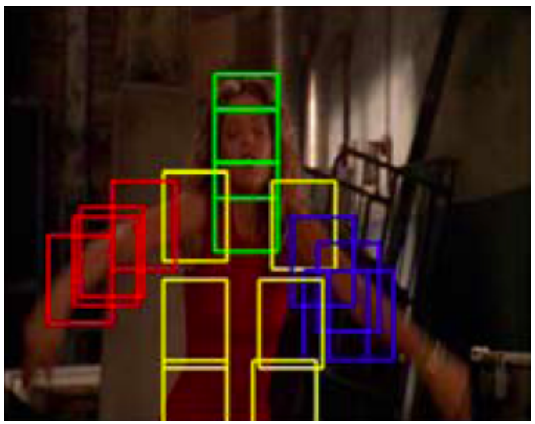
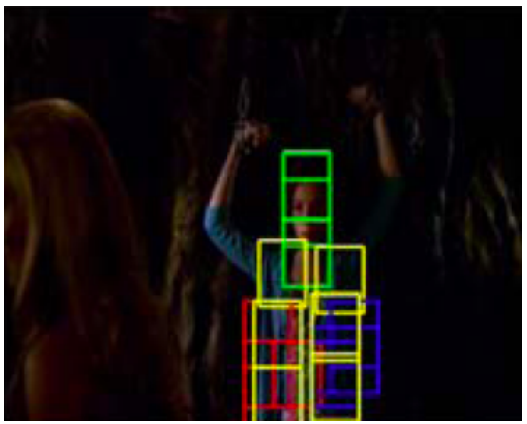
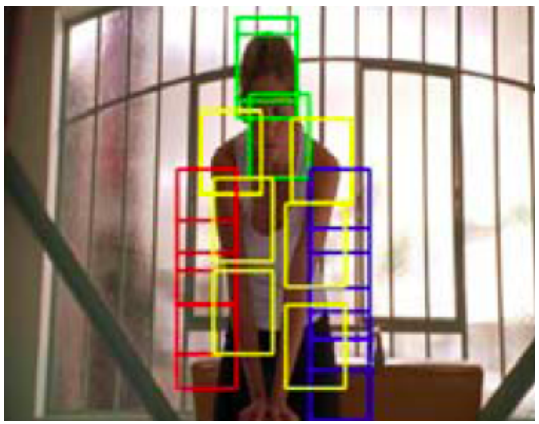
But we are improving...



2D pose estimation



- Pictorial Structures are the most employed methods
- Works in controlled environaments

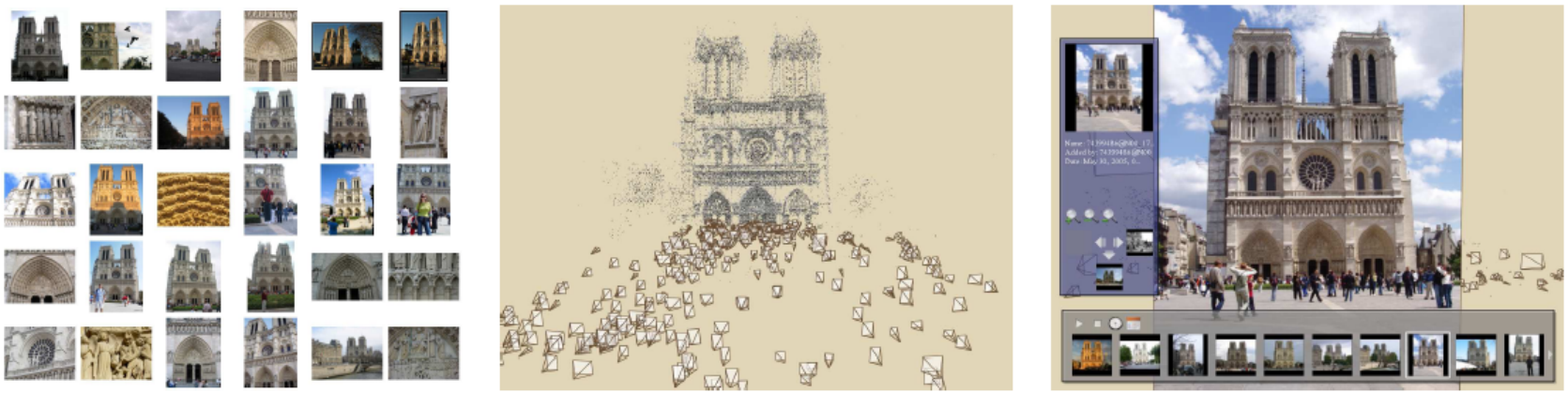




3D pose estimation

- Works with multiple cameras in control environments
- Or with depth sensors as kinect
- Monocular very difficult!
- Use motion, pose and physics priors

3D reconstruction

- Works well with multiple cameras
- or with multiple pictures of the same urban landmark
- Monocular very difficult, and only control environments works well
- If deformable objects, even more difficult.



(a) (b) (c)

Figure 1: Our system takes unstructured collections of photographs such as those from online image searches (a) and reconstructs 3D points and viewpoints (b) to enable novel ways of browsing the photos (c).

Next class ... a review on image formation