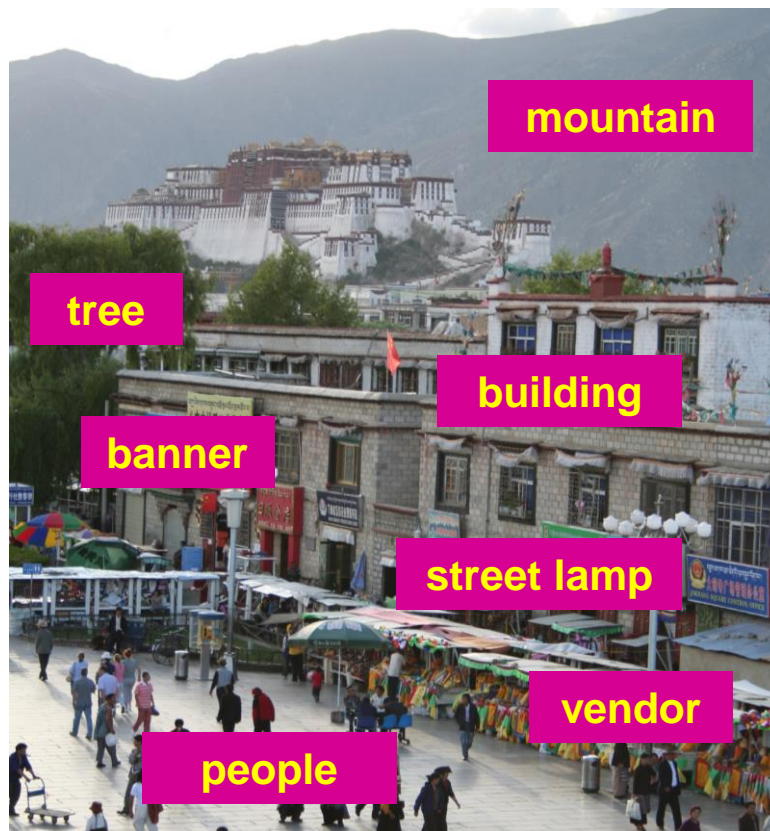


CS5670: Intro to Computer Vision

Noah Snavely

Introduction to Recognition, Part 2



Announcements

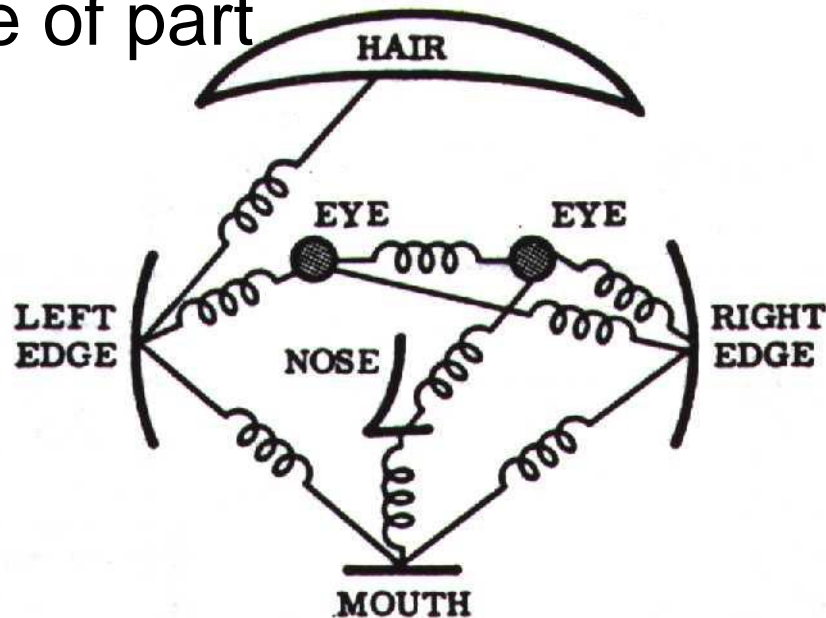
- Project 4 (Stereo)
 - Due this Friday, April 28, by 11:59pm
 - To be done in pairs
- Voting on Project 3 artifacts
- Quiz in class Thursday
- Final will be take-home, details to be announced next time

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

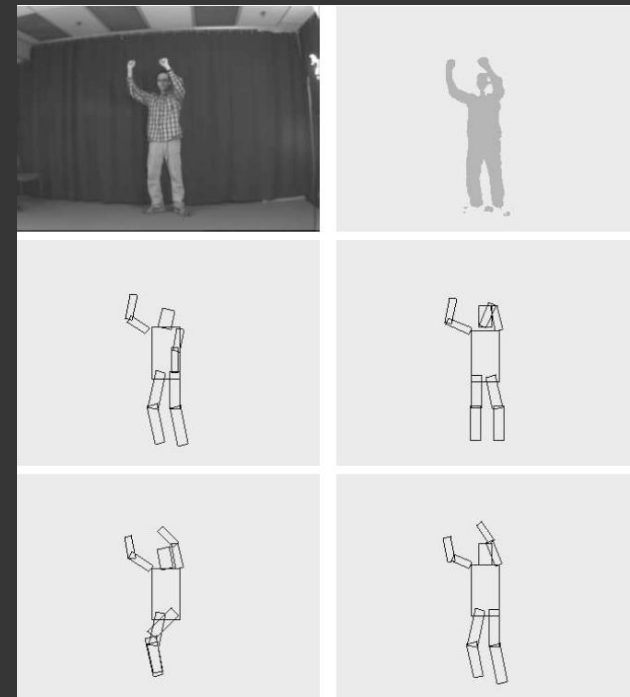
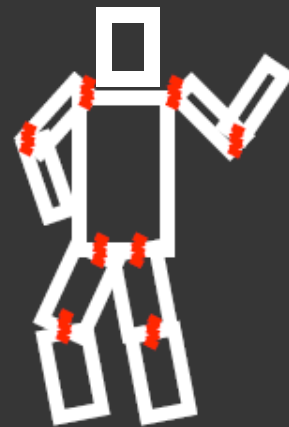
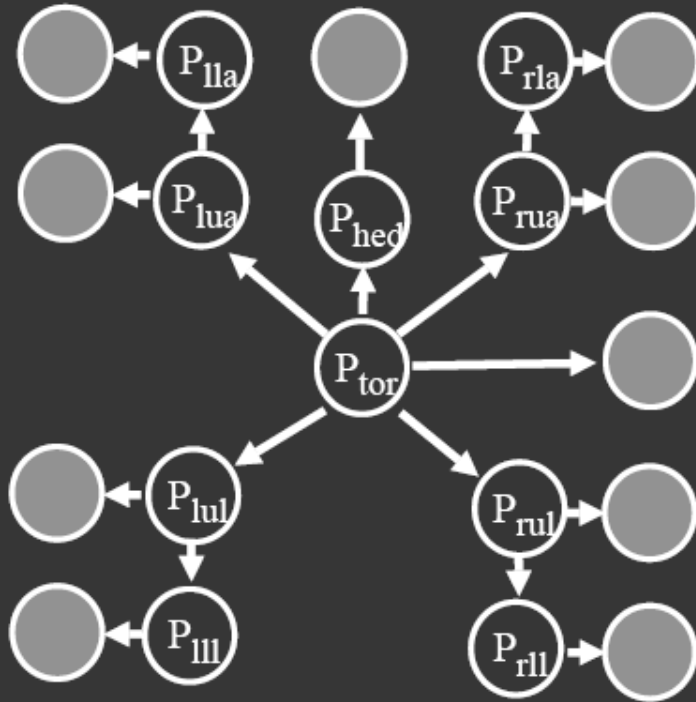
Parts-and-shape models

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



Pictorial structure model

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

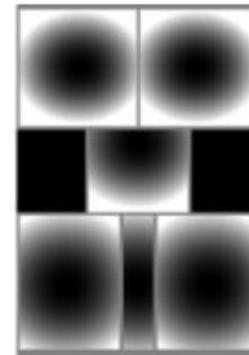
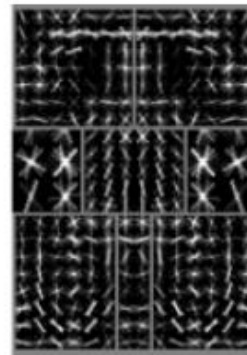
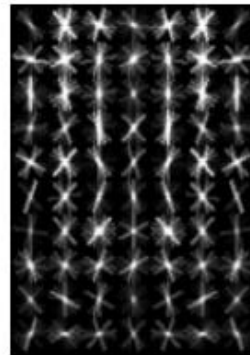
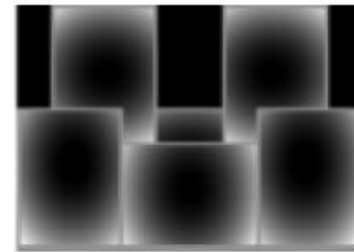
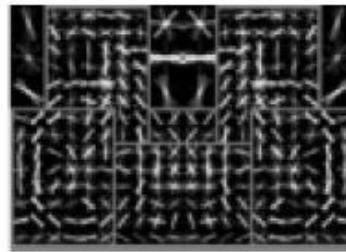
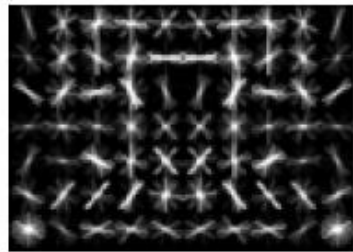
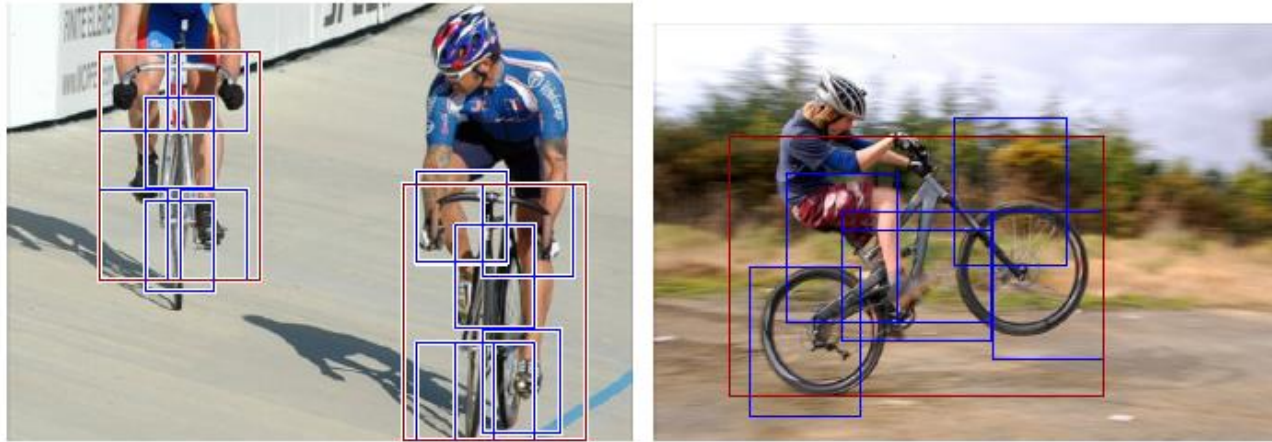


$$\Pr(P_{\text{tor}}, P_{\text{arm}}, \dots | \text{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\text{Im}(P_i))$$

↑
↑

part geometry
part appearance

Discriminatively trained part-based models

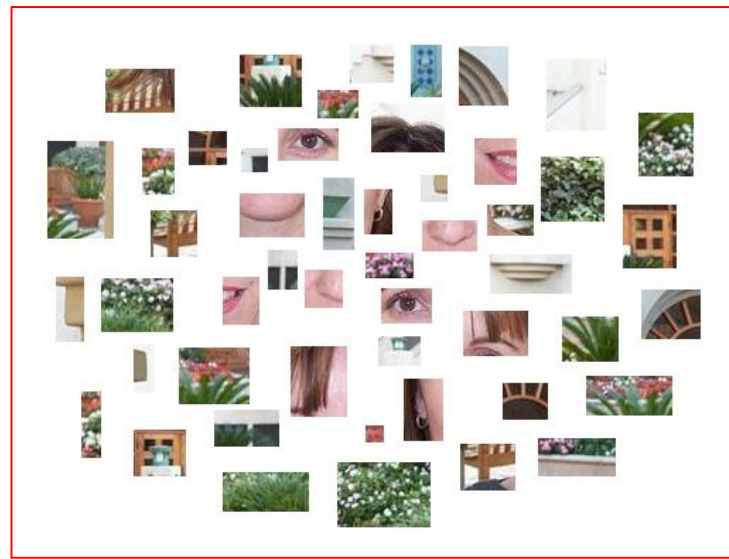
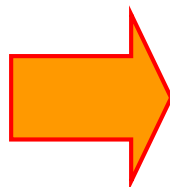


P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, "[Object Detection with Discriminatively Trained Part-Based Models,](#)" PAMI 2009

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'**



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,
deep learning

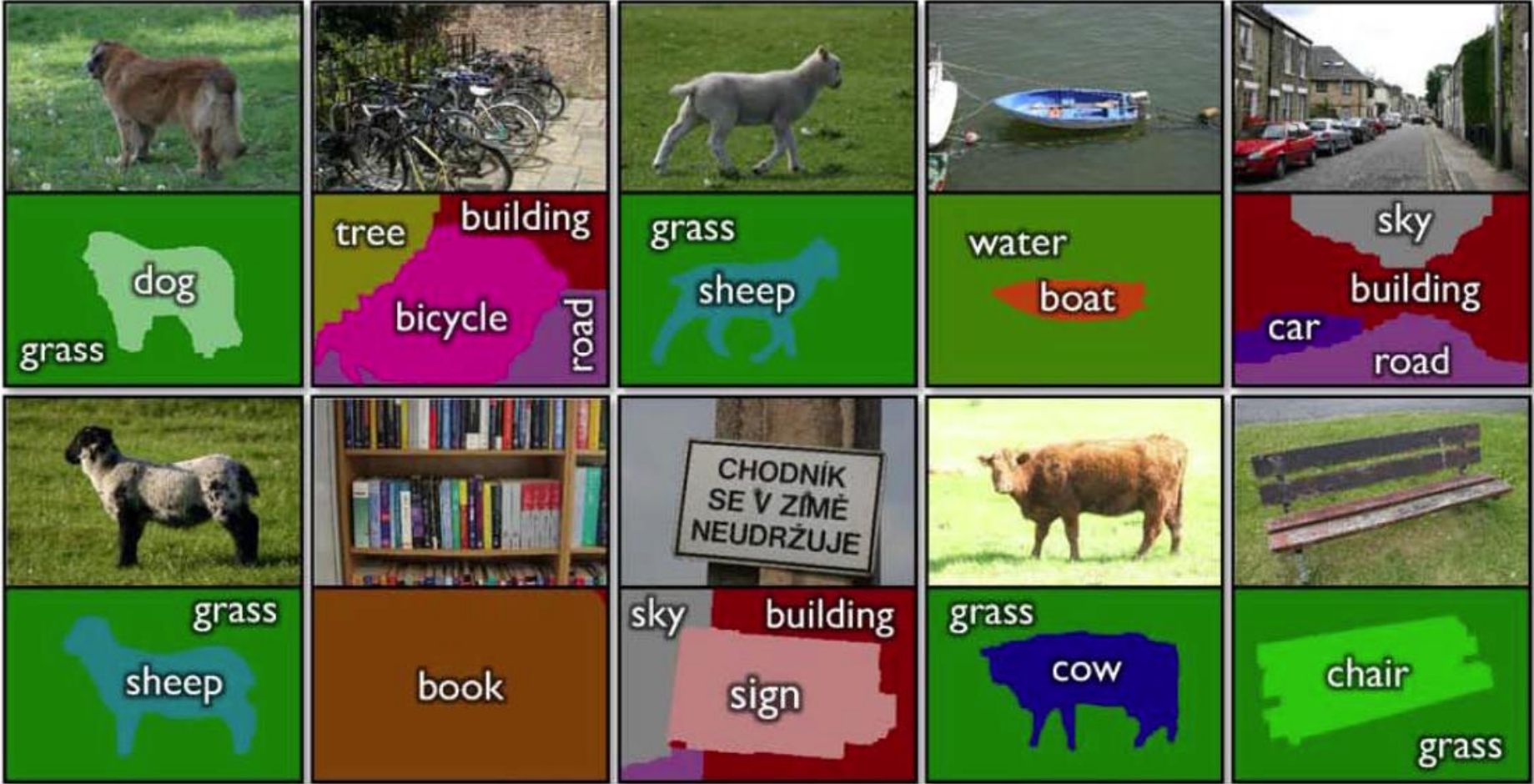
What Matters in Recognition?

- 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
 - Deep features
- 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



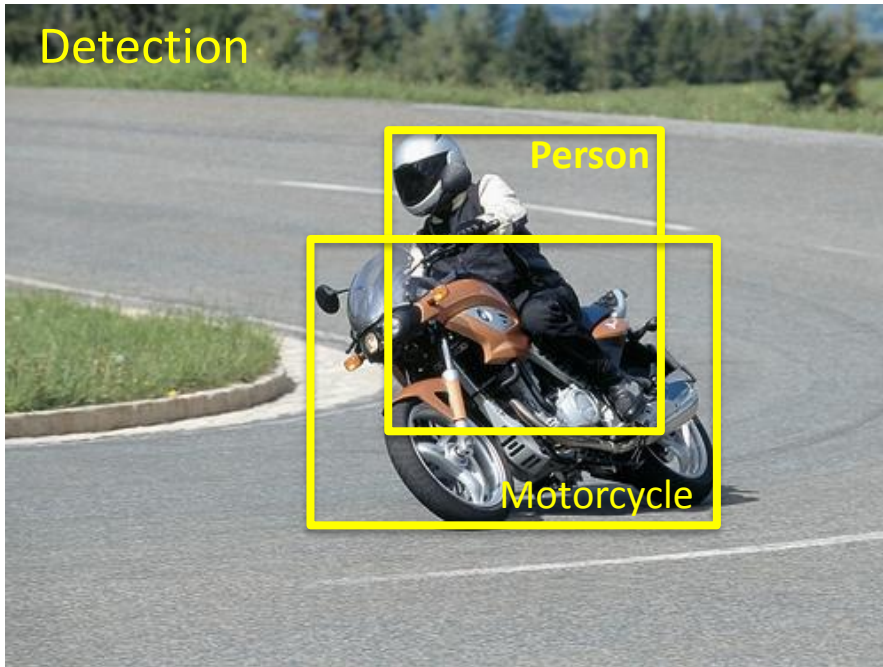
PASCAL VOC 2005-2012

20 object classes

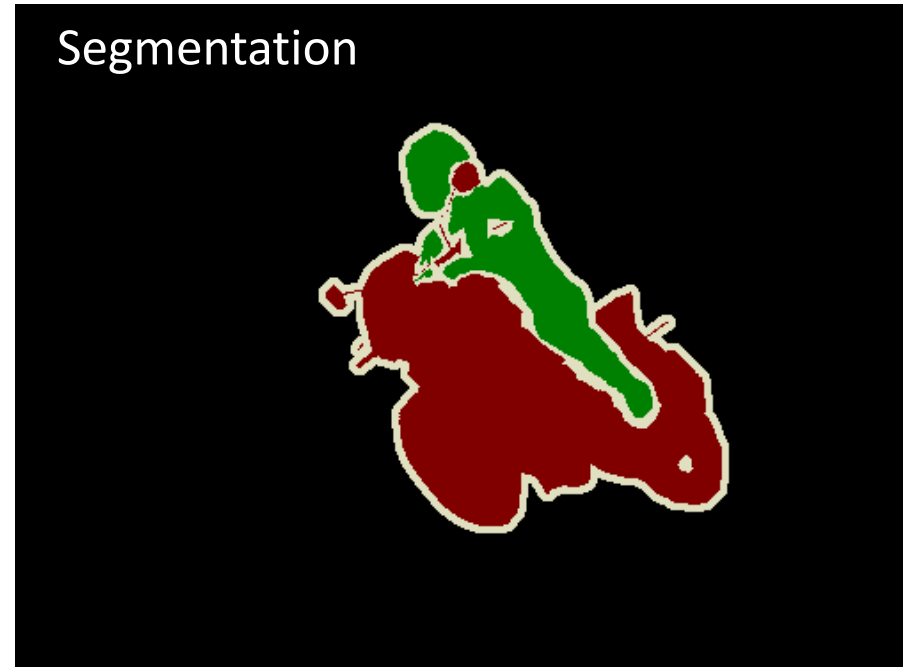
22,591 images

Classification: person, motorcycle

Detection



Segmentation

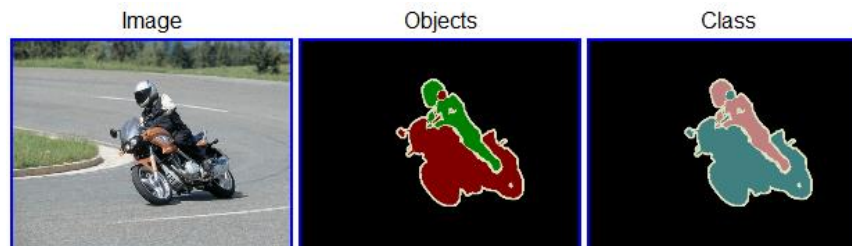


Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- 20 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?)



Examples

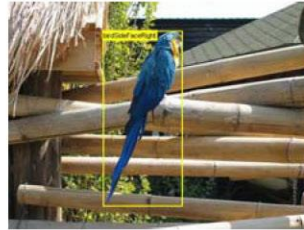
Aeroplane



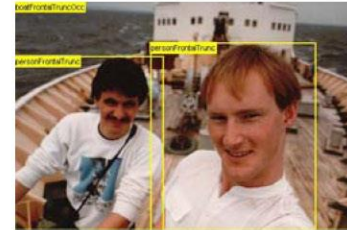
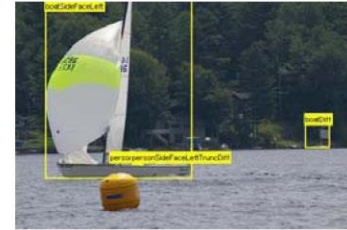
Bicycle



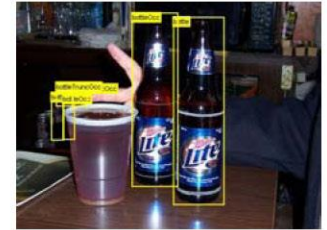
Bird



Boat



Bottle



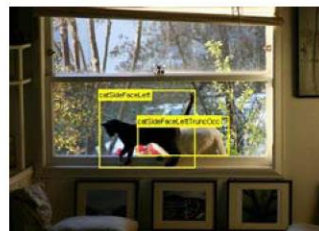
Bus



Car



Cat



Chair



Cow



Classification Challenge

- Predict whether at least one object of a given class is present in an image



is there a cat?

Precision / Recall for a Category X

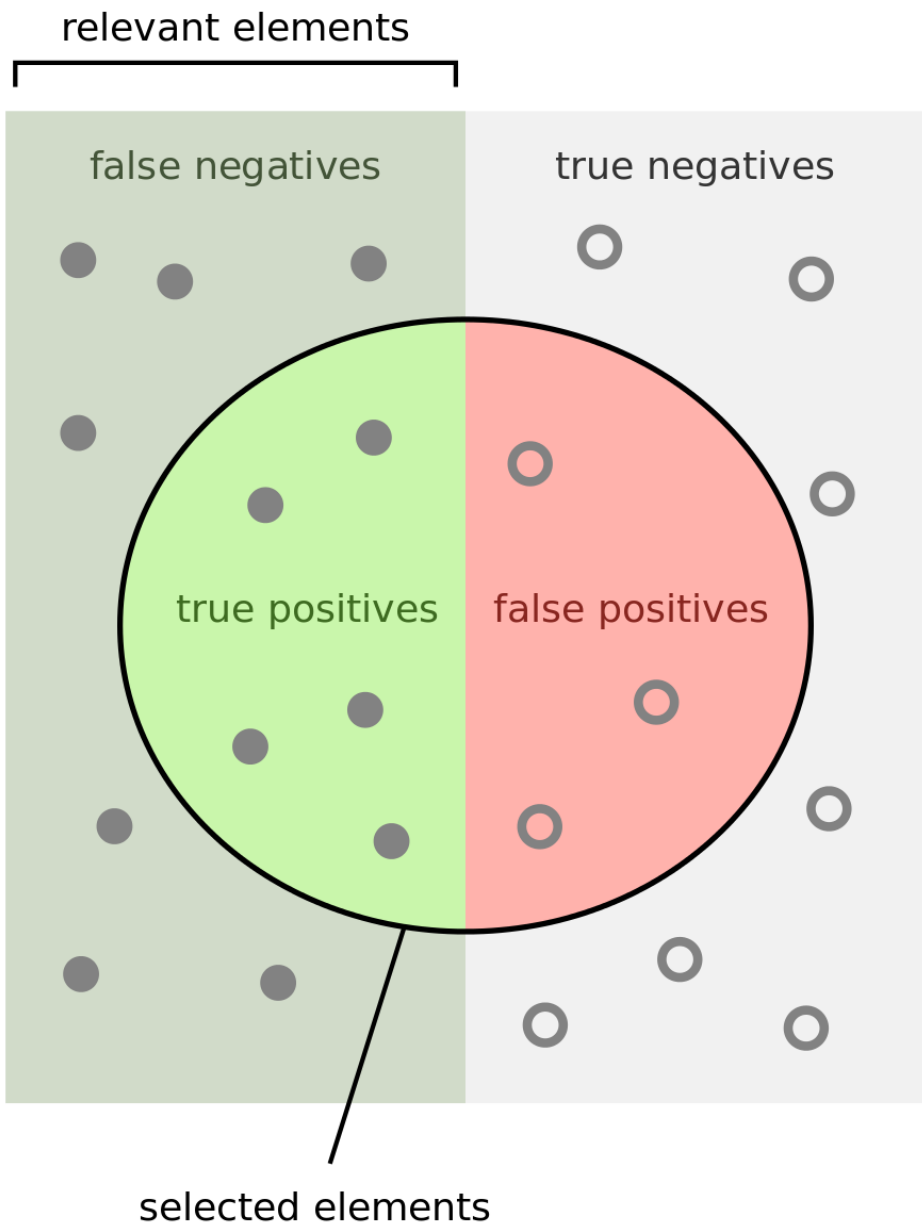
- Precision:

$$\frac{|\{\text{images that contain an X}\} \cap |\{\text{images classified as X}\}|}{|\{\text{images classified as X}\}|}$$

- Recall:

$$\frac{|\{\text{images that contain an X}\} \cap |\{\text{images classified as X}\}|}{|\{\text{images that contain an X}\}|}$$

- In reality, methods give a continuous-valued score for each image / category → PR curve



How many selected items are relevant?

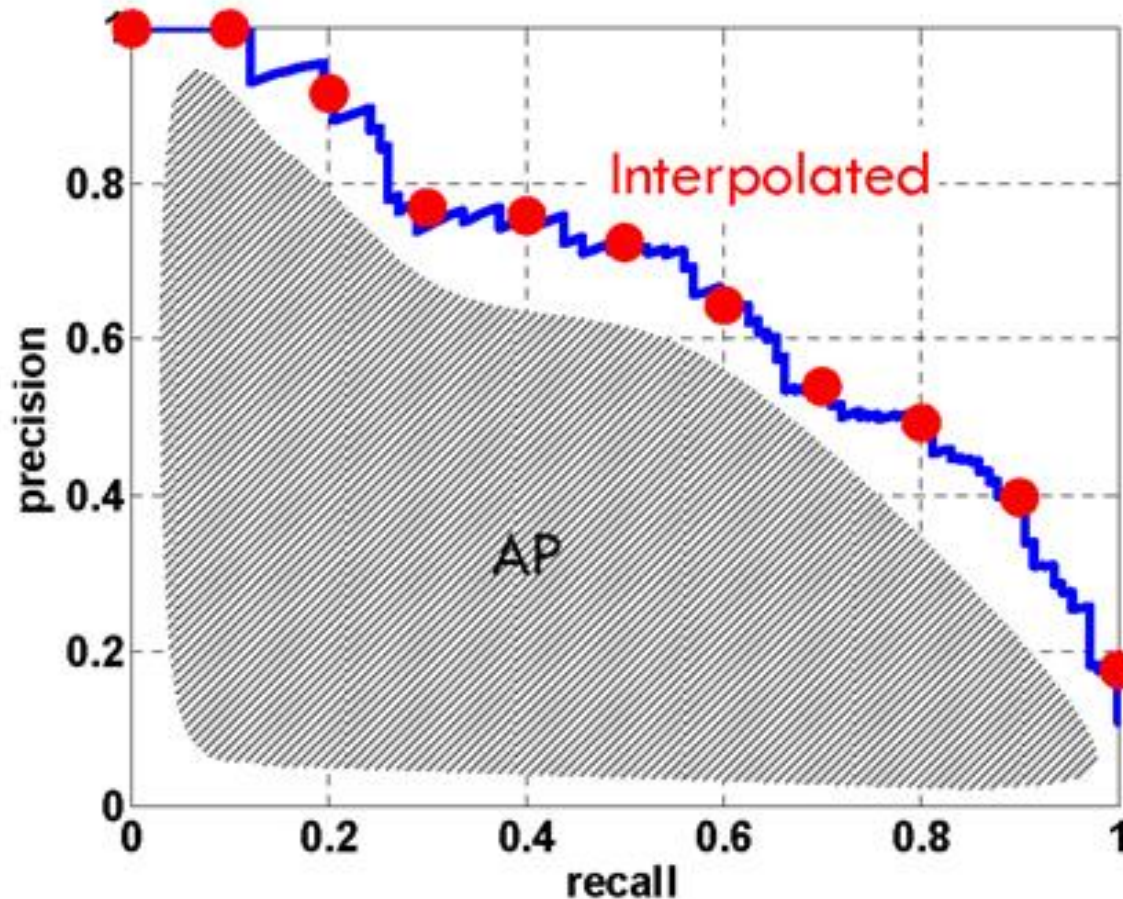
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

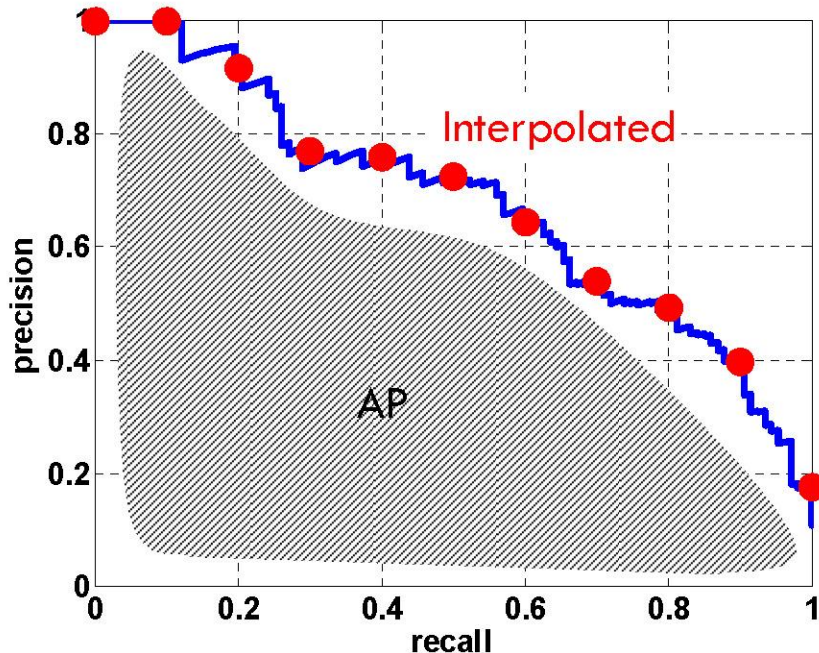
Precision / Recall Curve

- Similar to the ROC curves you saw in Project 2



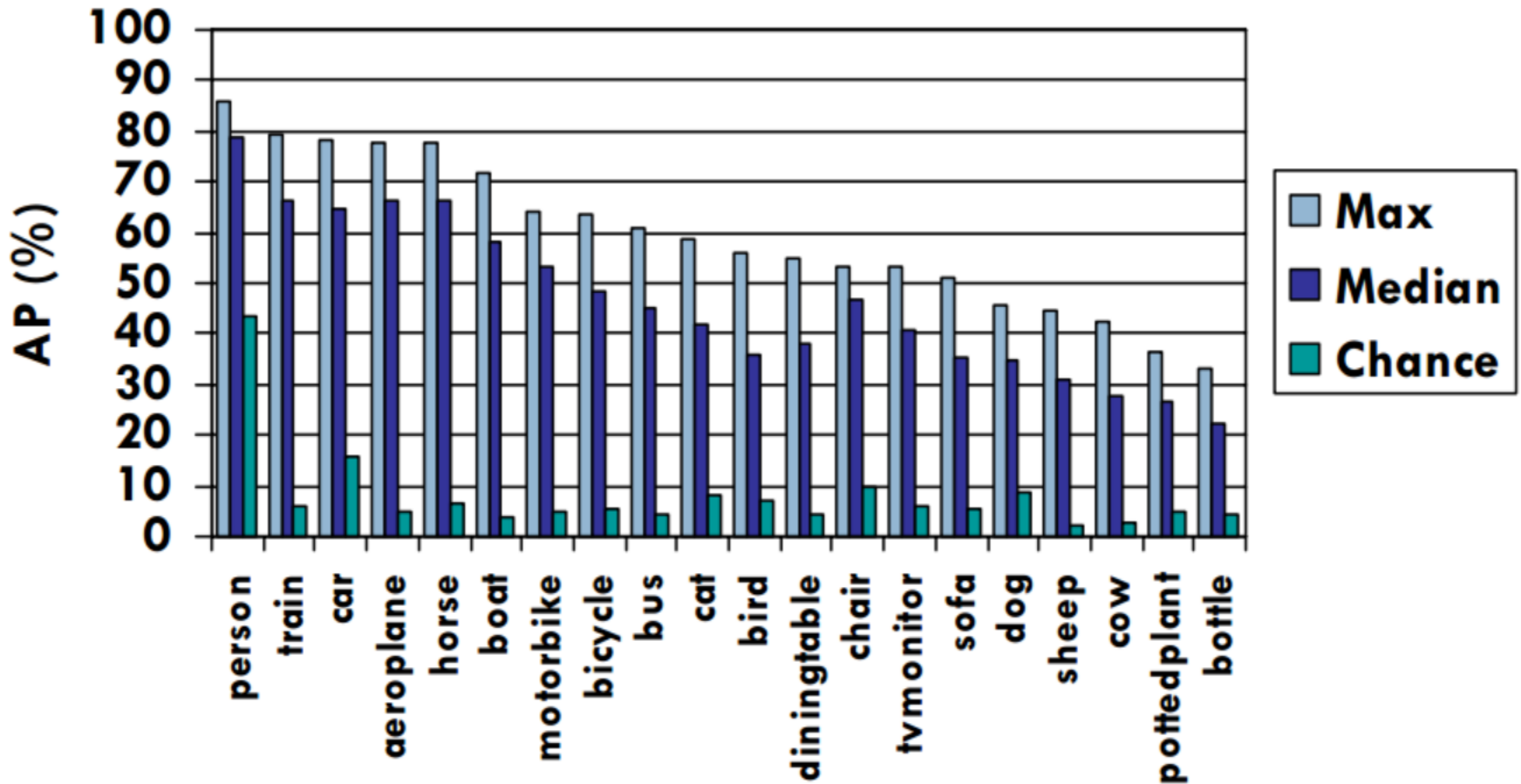
Evaluation

- Average Precision [TREC] averages precision over the entire range of recall
 - Curve interpolated to reduce influence of “outliers”

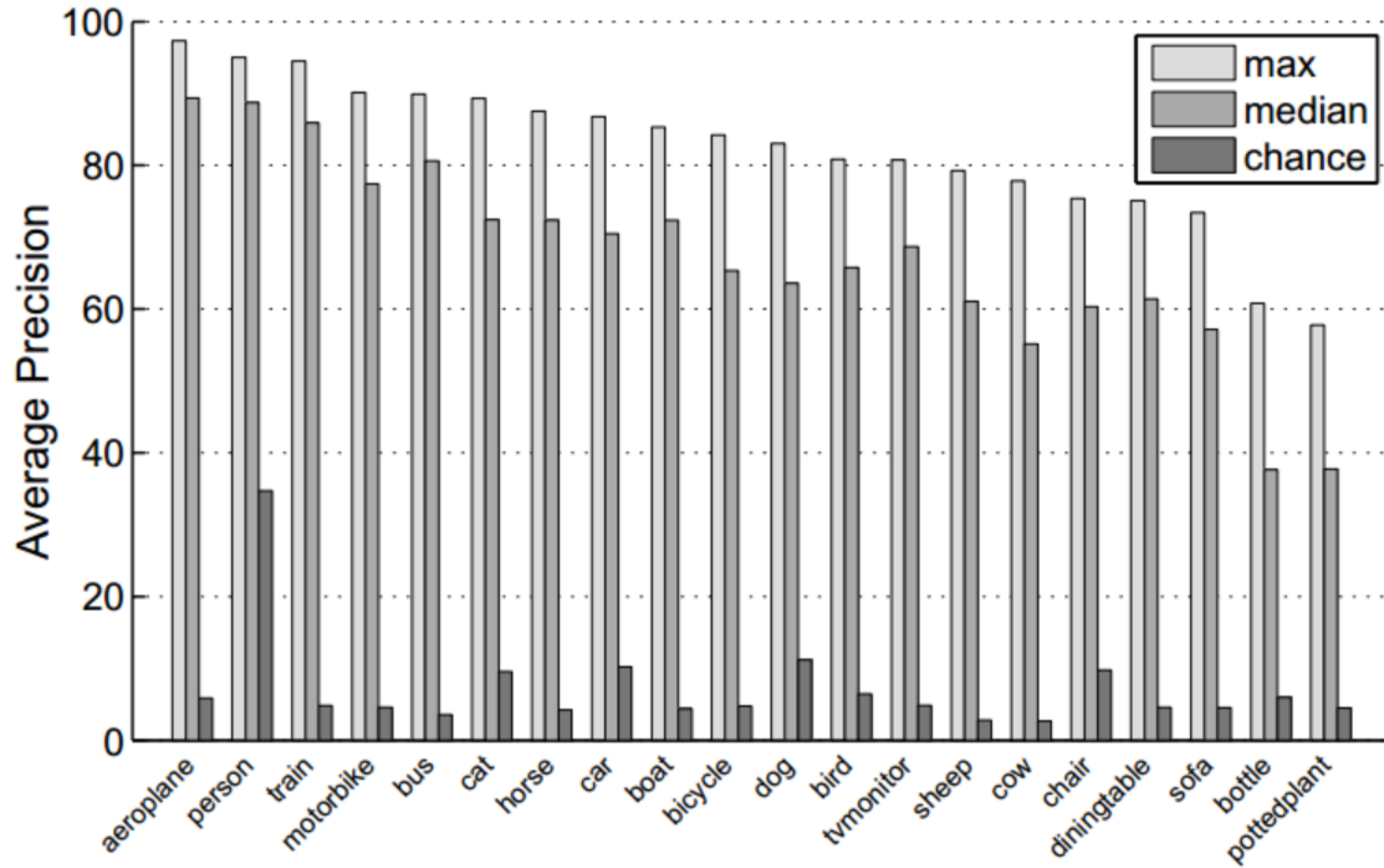


- A good score requires both high recall **and** high precision
- Application-independent
- Penalizes methods giving high precision but low recall

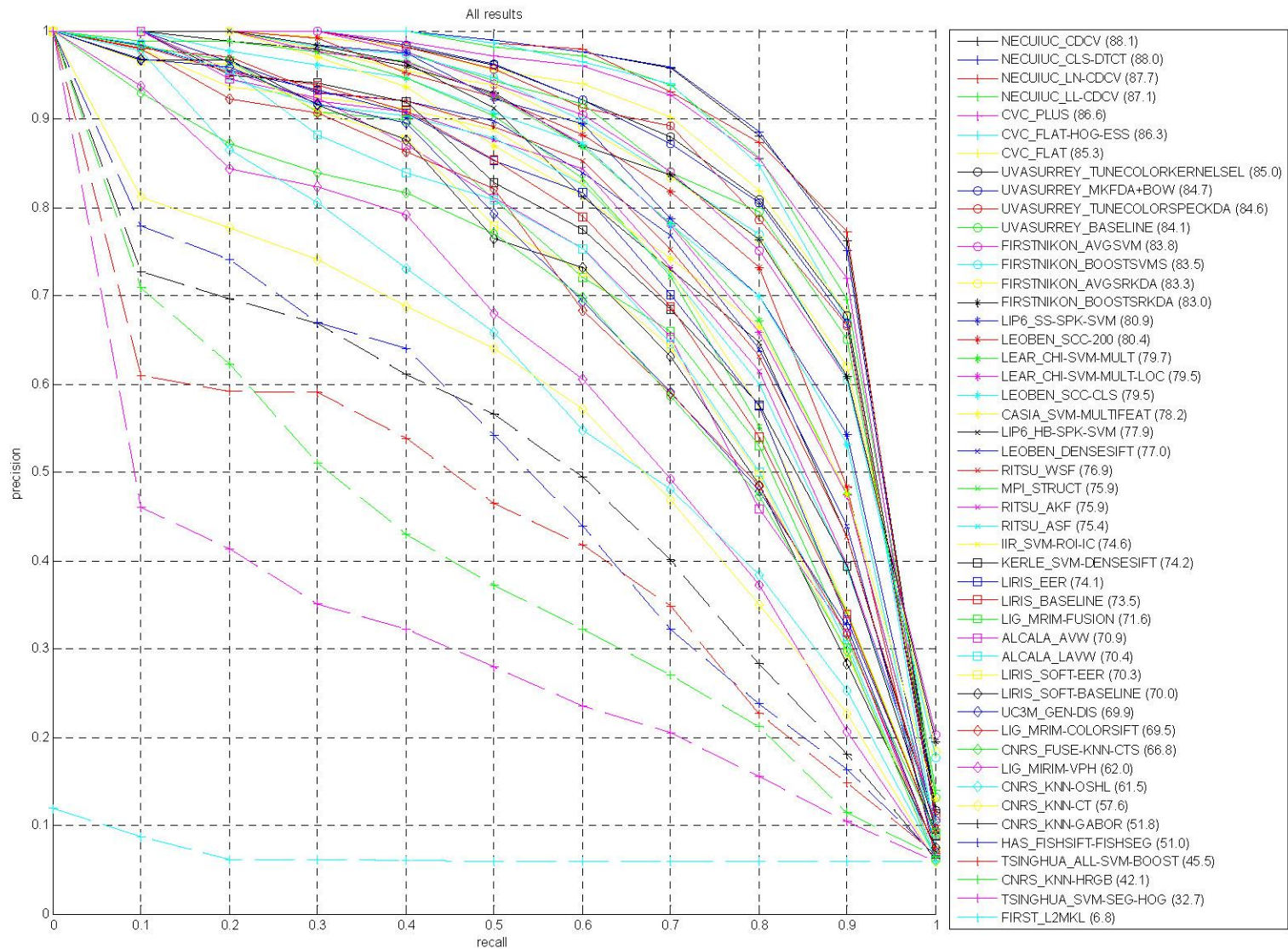
Pascal VOC 2007 Average Precision



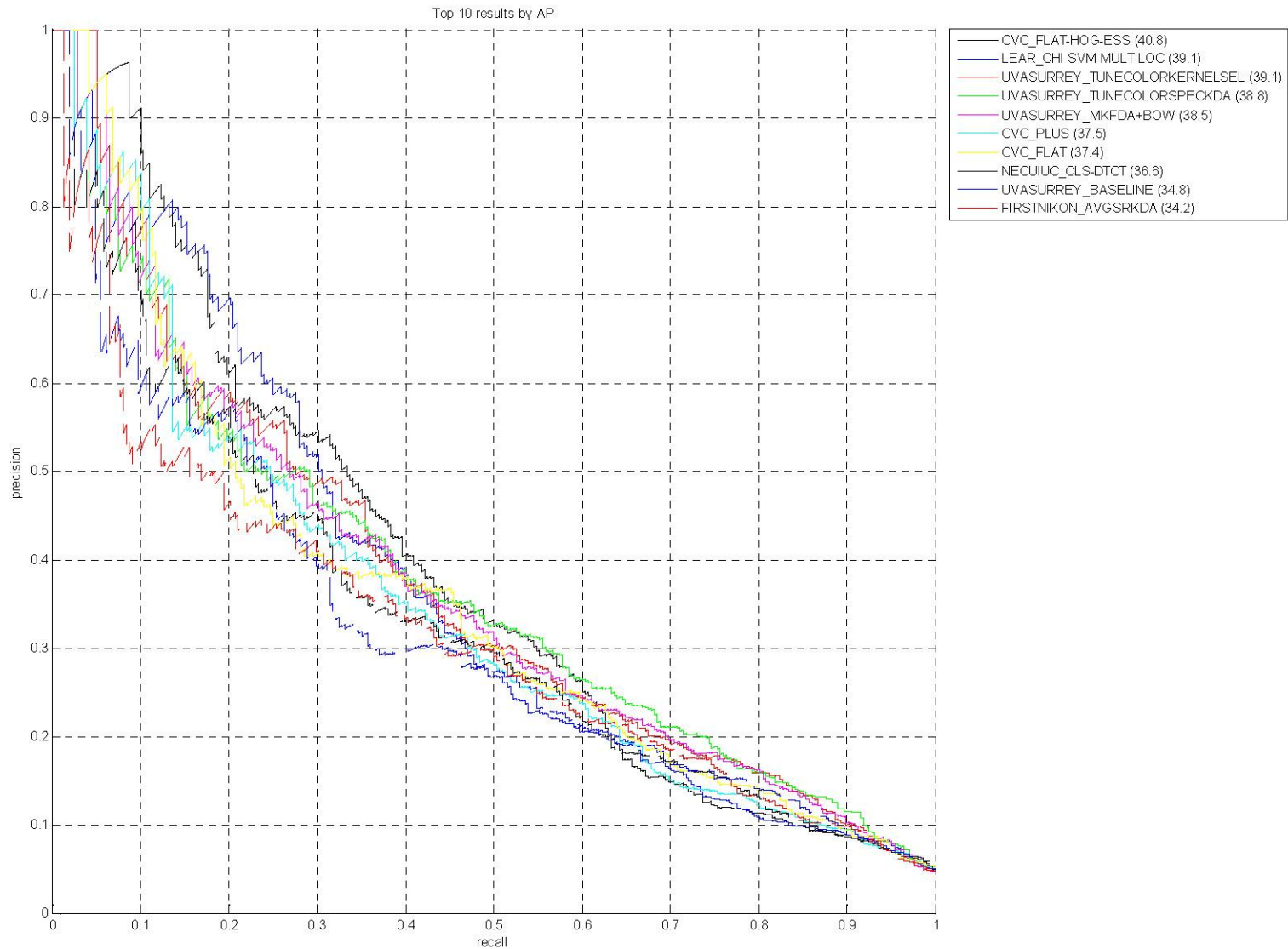
Pascal VOC 2012 Average Precision



Precision/Recall: Aeroplane (All)



Precision/Recall: Potted plant (Top 10 by AP)



Detection Challenge

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



True Positives - Person

UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP



NECUIUC_CLS-DTCT

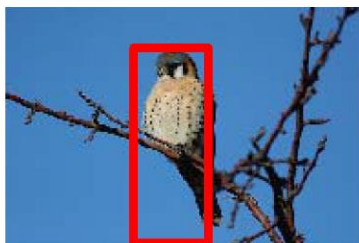


False Positives - Person

UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP



NECUIUC_CLS-DTCT



“Near Misses” - Person

UoCTTI_LSVM-MDPM



MIZZOU_DEF-HOG-LBP



NECUIUC_CLS-DTCT



True Positives - Bicycle

UoCTTI_LSVM-MDPM



OXFORD_MKL



NECUIUC_CLS-DTCT



False Positives - Bicycle

UoCTTI_LSVM-MDPM



OXFORD_MKL



NECUIUC_CLS-DTCT



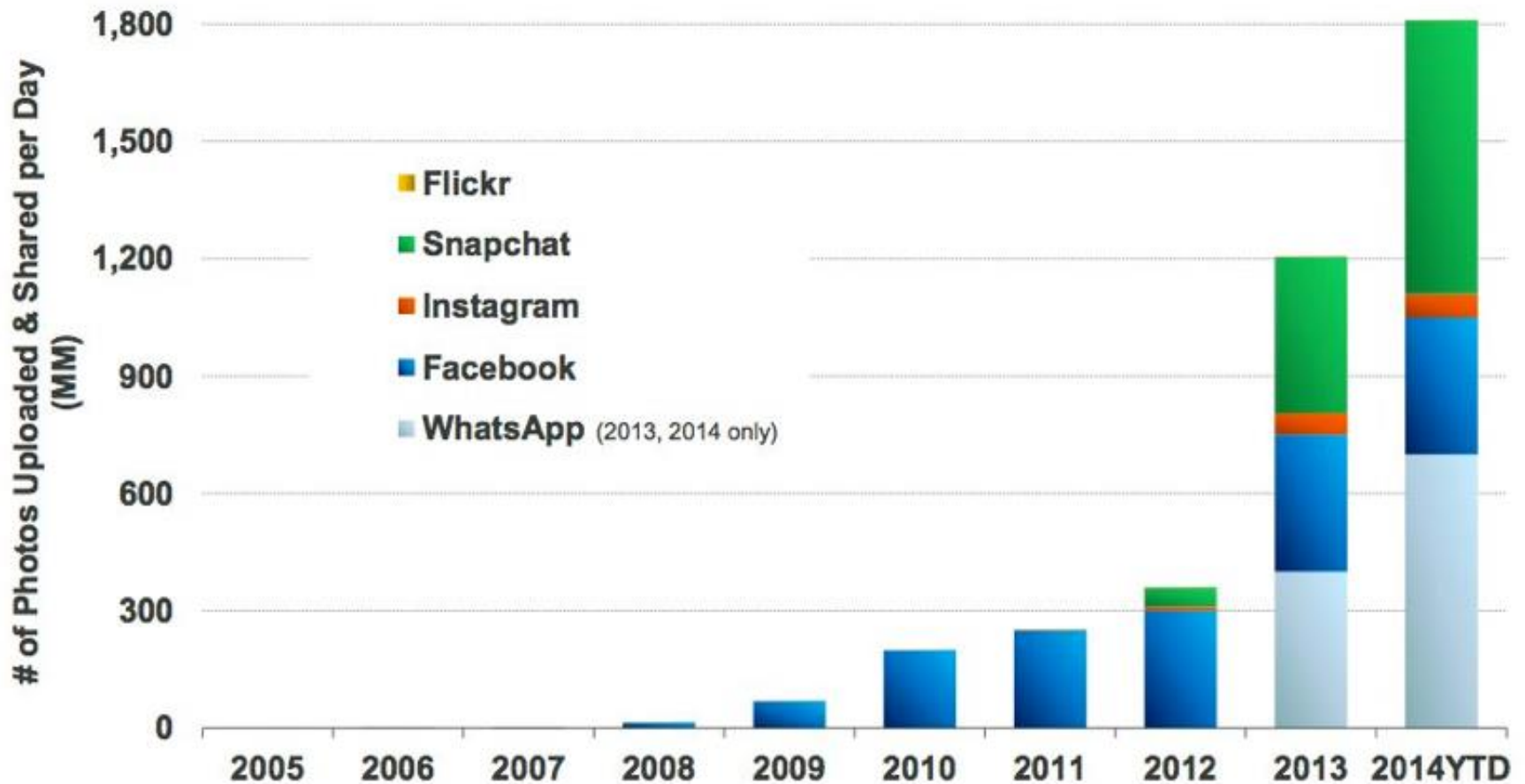
Where to from here?

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

24 Hrs in Photos



Daily Number of Photos Uploaded & Shared on Select Platforms, 2005 – 2014YTD





by John F Murphy



by SkySwarm



by NestorDesigns



by Ray Bradshaw



by jmtpt



by Damian_Ward



by Johnny B



by Manadh



by gerasphoto



by ShuMaJiao



by Maitora



by Suk1588



by Karamina166804600



by gerasphoto



by half man half penguin



by Laura Zupan



by Hoopa



by joshua1988



by Alex +



by Benjamin H



by vjjuu



by O.C Photo



by wazmull



by Brian POX



by Shudge 9000



by [odraze]



by cellinboards



by J. Wilson 500



by Cam in Dorset



by firetrough



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
- Microsoft COCO
 - Crowdsourced, large-scale objects

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

PASCAL

ILSVRC

birds



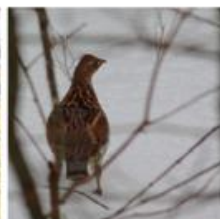
bird



flamingo



cock



ruffed grouse



quail



partridge . . .

bottles



bottle



pill bottle



beer bottle



wine bottle



water bottle



pop bottle . . .

cars



car



race car



wagon



minivan

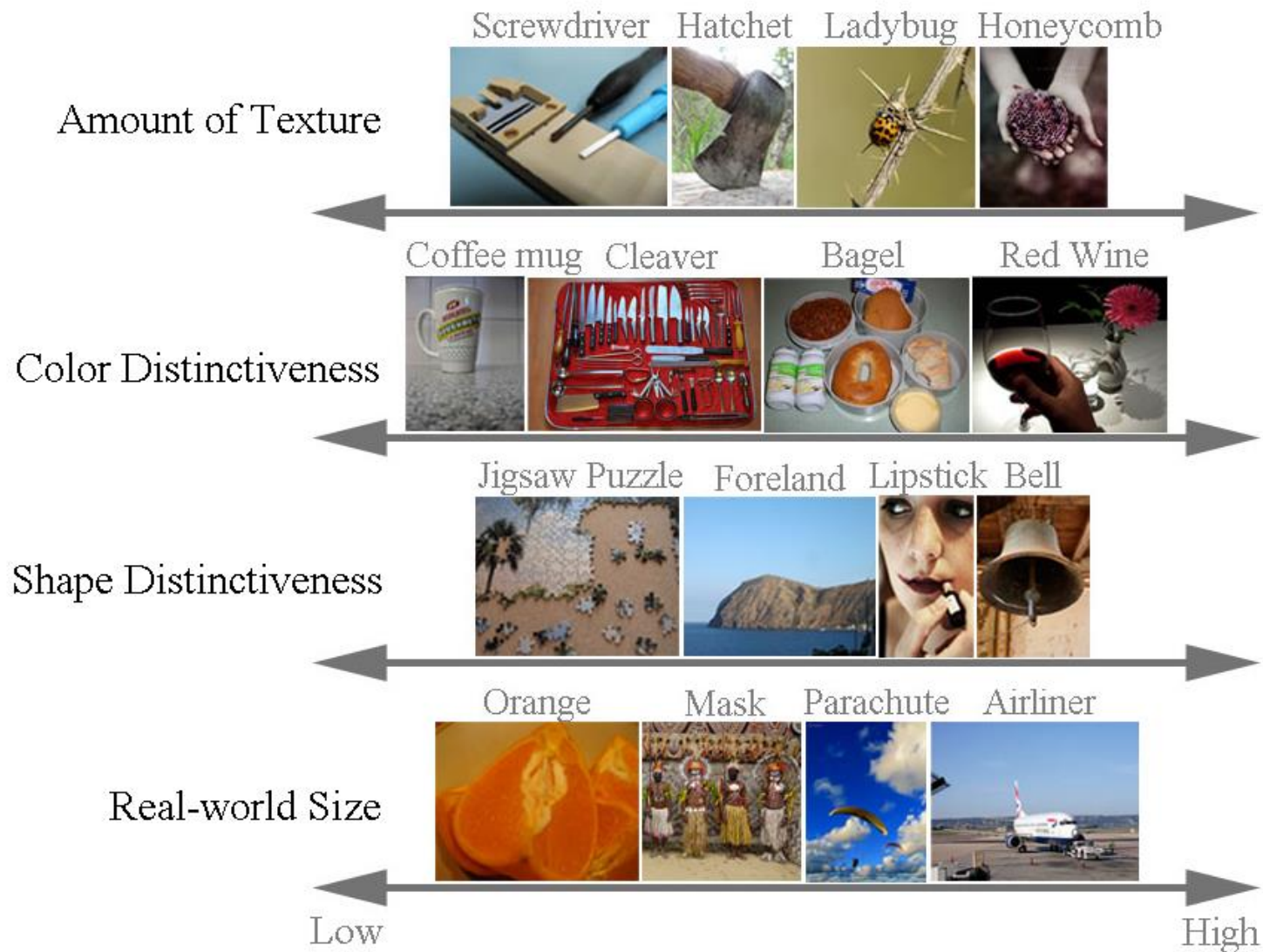


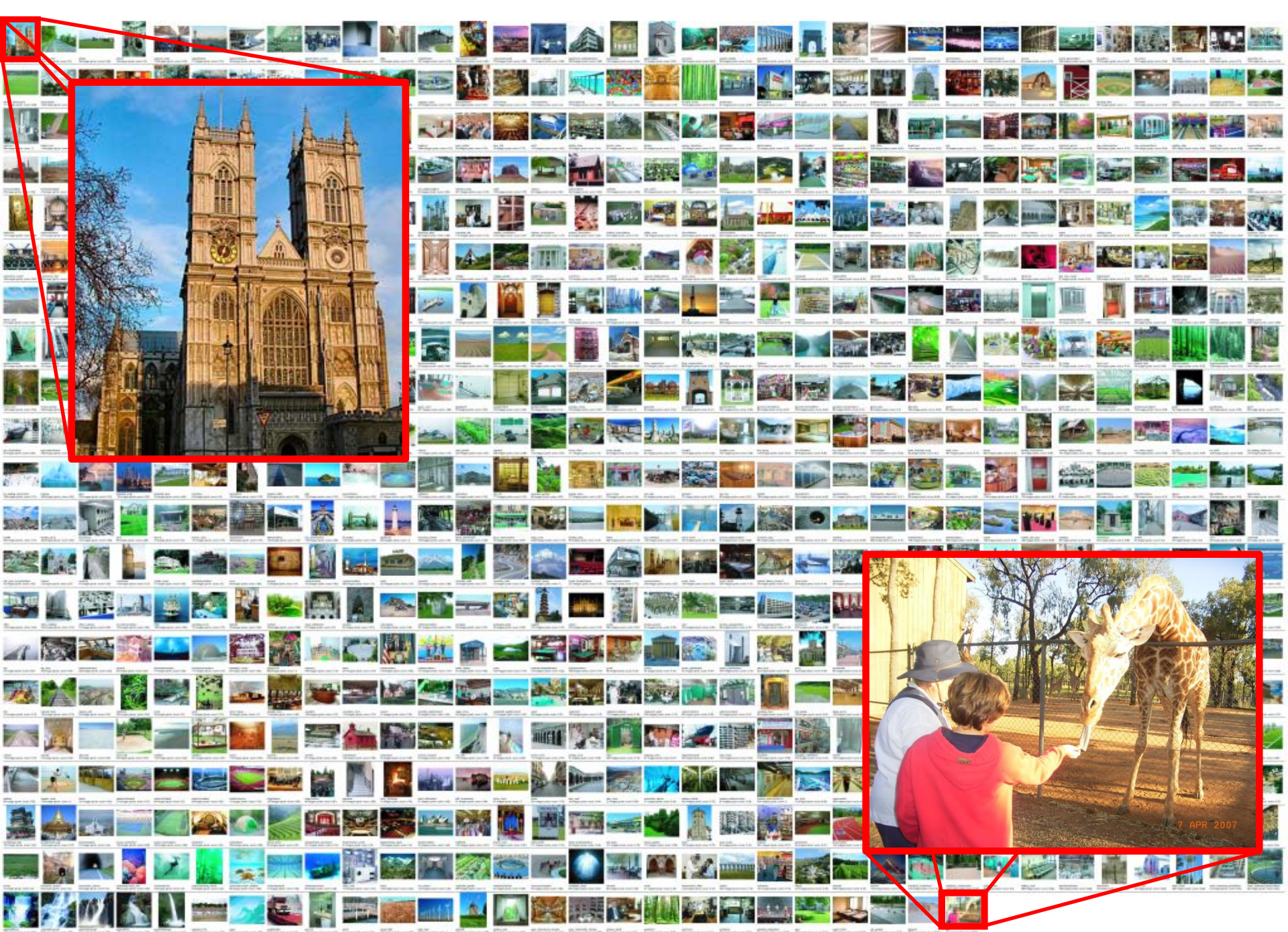
jeep



cab . . .

Variety of object classes in ILSVRC





What are attributes?



What do we want to know about this object?

Object recognition expert:
“Dog”

Next step: Infer object properties



Can I **poke with it**?

Can I **put stuff in it**?

What **shape** is it?

Is it **alive**?

Is it **soft**?

Does it have a **tail**?

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”

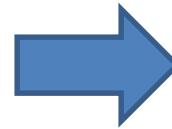
vs.

“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



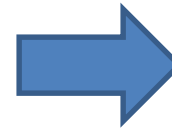
Has Stripes
Has Ears
Has Eyes
....



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?

Questions?

Where to from here?

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

Image categorization

Training

Training
Images



Training
Labels


Image
Features

Classifier
Training


Trained
Classifier



Categorization

$f(\text{}) = \text{“apple”}$

$f(\text{}) = \text{“tomato”}$

$f(\text{}) = \text{“cow”}$

Training

Training Images



Image Features

Training Labels

Training

Learned Classifier

Testing

Test Image



Image Features

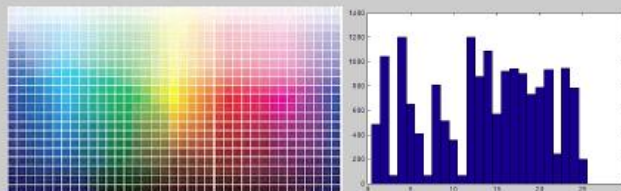
Learned Classifier

Prediction

Input image



Color: Quantize RGB values



Invariance?

- 😊 Translation
- 😊 Scale
- 😊 Rotation
- 😞 Occlusion

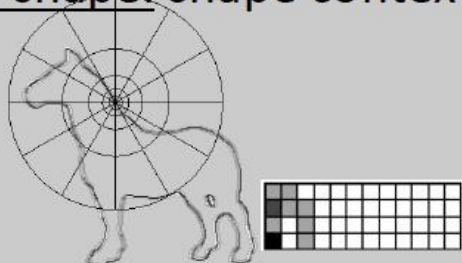
Global shape: PCA space



Invariance?

- 😊 Translation
- ? Scale
- 😊 Rotation
- 😞 Occlusion

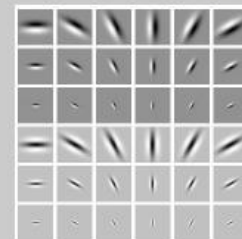
Local shape: shape context



Invariance?

- 😊 Translation
- 😊 Scale
- ? Rotation (in-planar)
- 😞 Occlusion

Texture: Filter banks



Invariance?

- 😊 Translation
- ? Scale
- ? Rotation (in-planar)
- 😞 Occlusion

Results



	Color	$D_x D_y$	Mag-Lap	PCA Masks	PCA Gray	Cont. Greedy	Cont. DynProg	Avg.
apple	57.56%	85.37%	80.24%	78.78%	88.29%	77.07%	76.34%	77.66%
pear	66.10%	90.00%	85.37%	99.51%	99.76%	90.73%	91.71%	89.03%
tomato	98.54%	94.63%	97.07%	67.80%	76.59%	70.73%	70.24%	82.23%
cow	86.59%	82.68%	94.39%	75.12%	62.44%	86.83%	86.34%	82.06%
dog	34.63%	62.44%	74.39%	72.20%	66.34%	81.95%	82.93%	67.84%
horse	32.68%	58.78%	70.98%	77.80%	77.32%	84.63%	84.63%	69.55%
cup	79.76%	66.10%	77.80%	96.10%	96.10%	99.76%	99.02%	87.81%
car	62.93%	98.29%	77.56%	100.0%	97.07%	99.51%	100.0%	90.77%
total	64.85%	79.79%	82.23%	83.41%	82.99%	86.40%	86.40%	80.87%

Crowdsourcing

Unlabeled
Images



Show images,
Collect and
filter labels

Training
Images



Training
Labels

Image
Features

Classifier
Training

Trained
Classifier

Training



IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC



Dense grid descriptor:
HOG, LBP

Coding: local coordinate,
super-vector

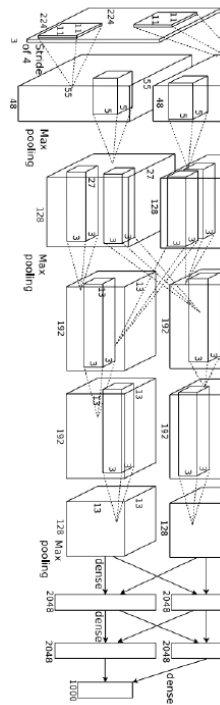
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet



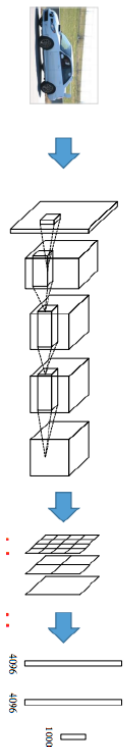
[Szegedy arxiv 2014]

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