Lecture 10: Object detection

COS 429: Computer Vision

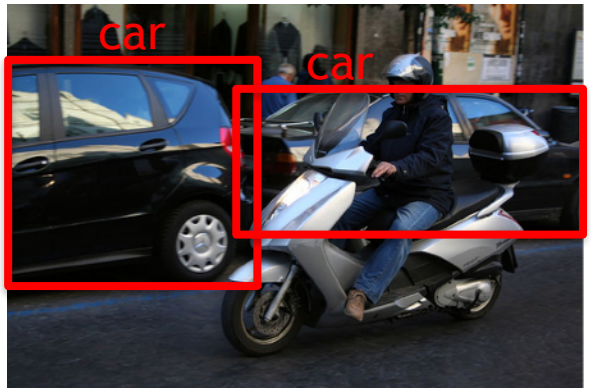
Recognition: classification vs detection

Classification

car

- Image-level label
- Doesn't require/assume object position in the image (blessing and curse)
- Frequently relies on context
- Doesn't require counting
- Doesn't require delineating multiple instances

Detection



- Box-level label
- Box tight around the object instance (blessing and curse)
- Requires counting and delineating nearby instances
- Requires finding all instances
- May require non-max suppression

Annotation costs

Draw a tight bounding box around the moped

Annotation costs

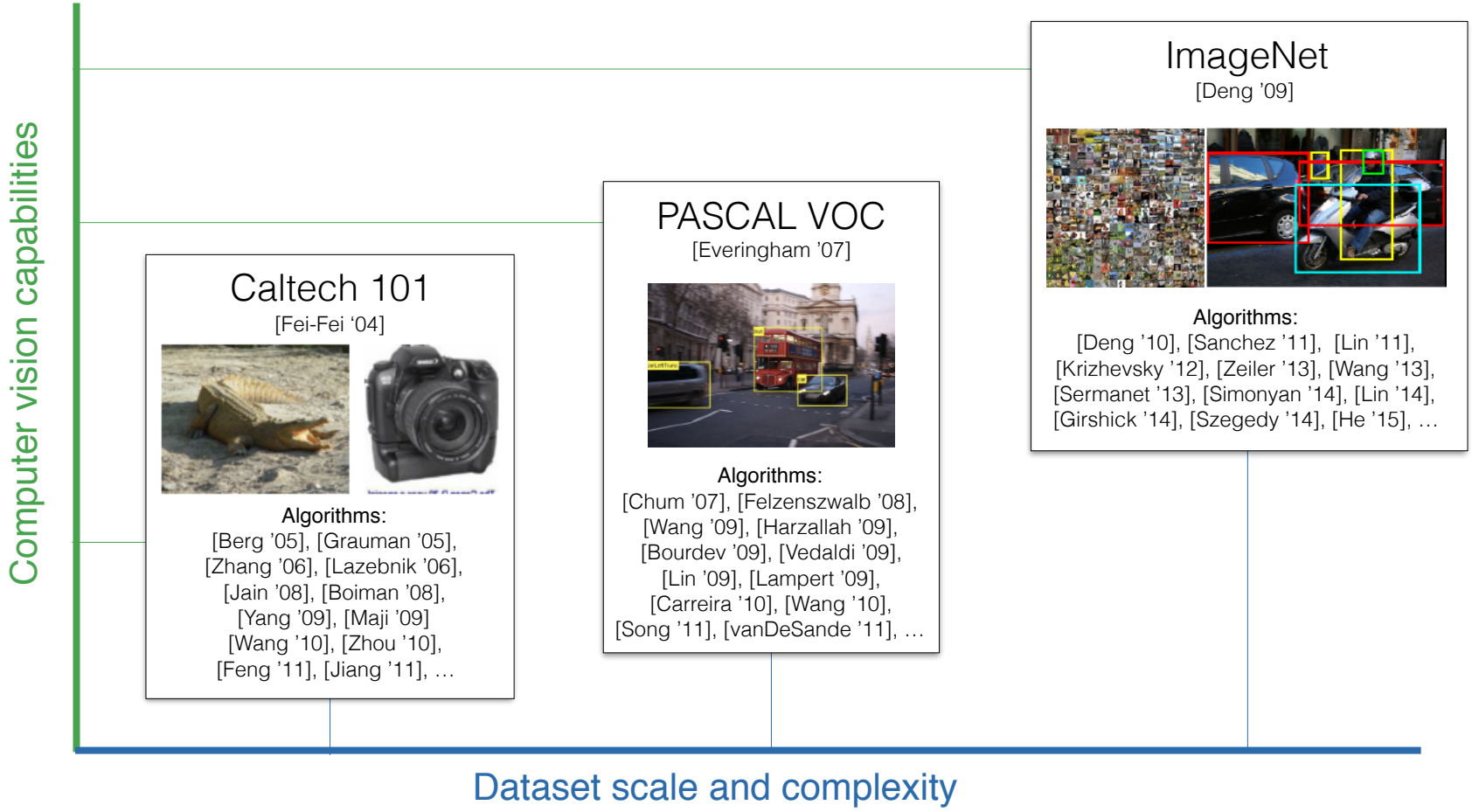
Draw a tight bounding box around the moped

Advanced topics in vision seminar (spring semester)

This took **14.5 seconds**

(**7 sec** [Jain&Grauman ICCV'13],
10.2 sec [Russakovsky, Li, Fei-Fei CVPR'15],
25.5 sec [Su, Deng, Fei-Fei AAAIW'12])

Datasets drive computer vision progress



Caltech 101

COMPUTATIONAL VISION AT CALTECH

Caltech 101

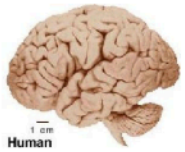
New Caltech256 New

[Description][Download][Discussion [Other Datasets]









Description

Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels.
We have carefully clicked outlines of each object in these pictures, these are included under the 'Annotations.tar'. There is also a matlab script to view the annotaitons, 'show_annotations.m'.

http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Classification evaluation: option 1 — accuracy


- Is this an umbrella or a strawberry?
- assign 1 of N class labels to an image (N = 101 in Caltech 101)

$$Accuracy = \frac{1}{NumImages} \sum_{i=1}^{NumImages} 1[correct\ on\ image\ i]$$

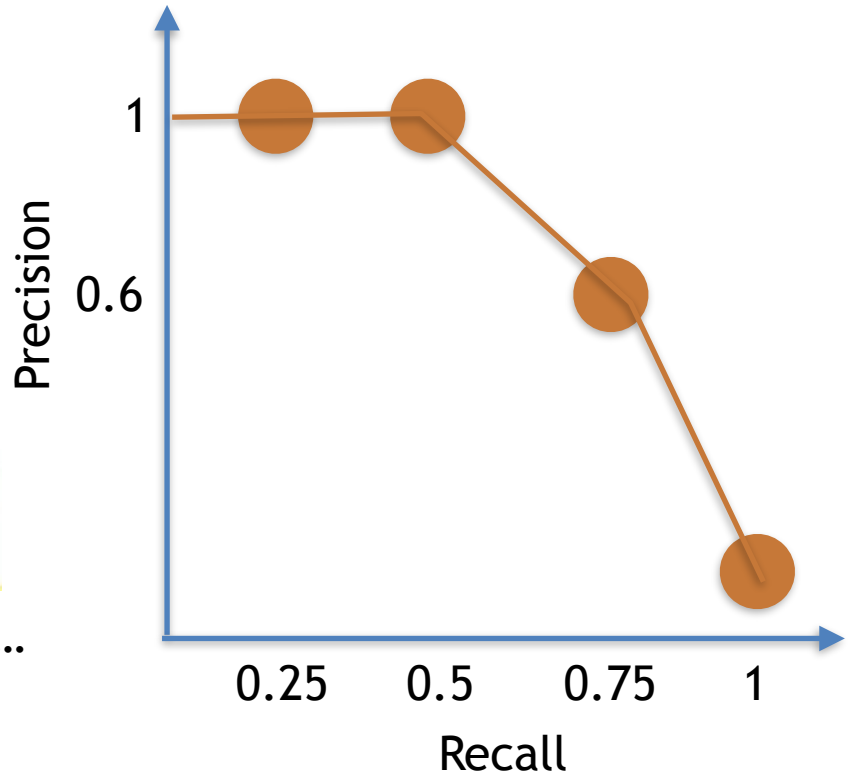
Good:
- clean and simple

Bad:
- What if there are multiple objects in the image?
- Sensitive to class priors

Classification evaluation: option 2 — area under the curve (AUC)

- give an “umbrellaness” score

0.95 0.92 0.85 0.74 0.55 …



$$Recall = \frac{NumTruePositives}{NumPositives}$$

$$Precision = \frac{NumTruePositives}{NumPredictions}$$

(supposing there are 4 umbrellas in the test set in this example)

Coming back to Caltech 101

COMPUTATIONAL VISION AT CALTECH

Caltech 101

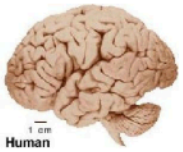
New Caltech256 New

[Description][Download][Discussion [Other Datasets]









Description

Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels.
We have carefully clicked outlines of each object in these pictures, these are included under the 'Annotations.tar'. There is also a matlab script to view the annotaitons, 'show_annotations.m'.

http://www.vision.caltech.edu/Image_Datasets/Caltech101/

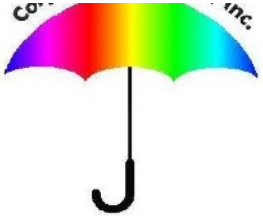
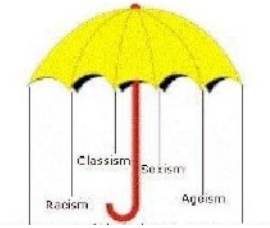
Caltech 101: umbrellas















Haar features on Caltech 101

How well would Haar features work?

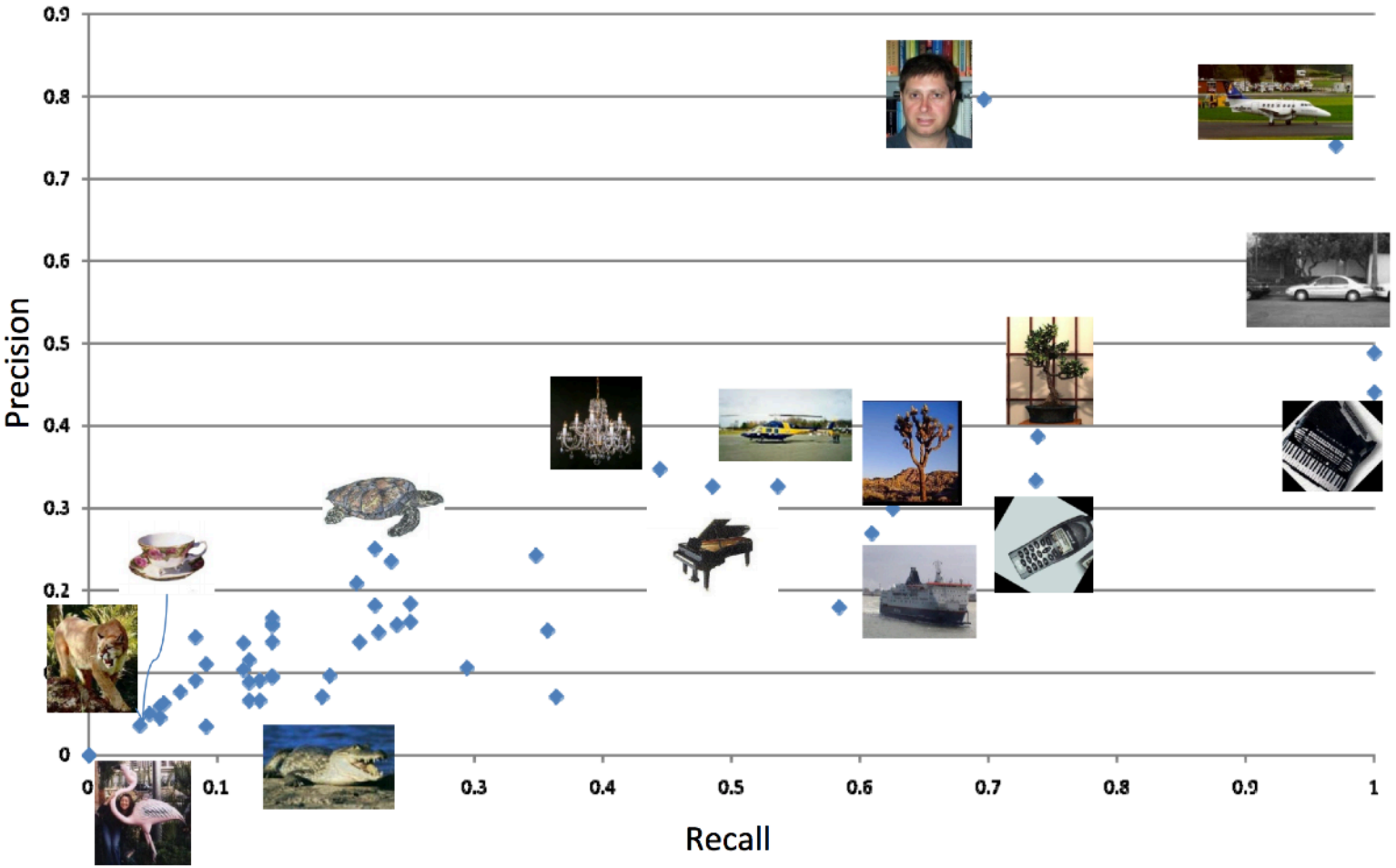
Some example object classes







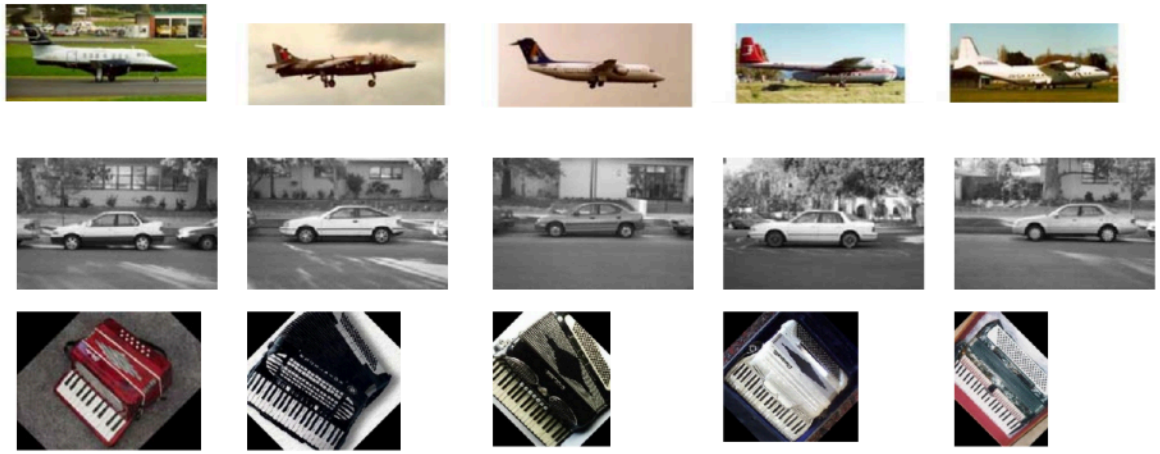

Haar features on Caltech 101



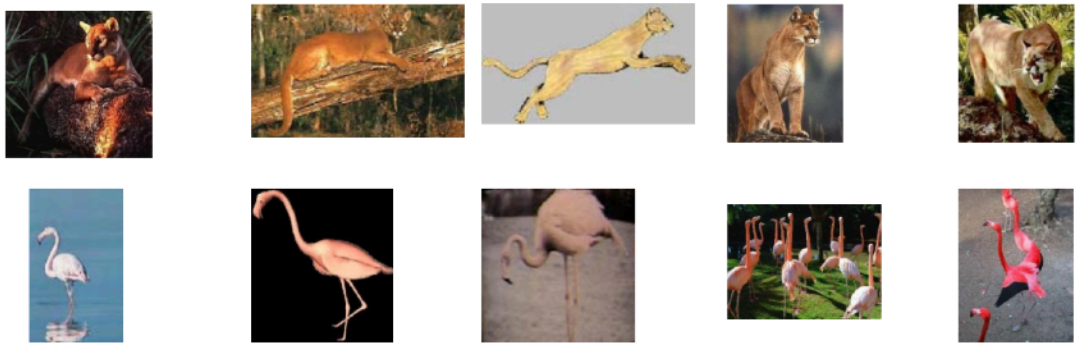
Credit: Kristen Grauman

Variation in training images

High accuracy categories



Low accuracy categories

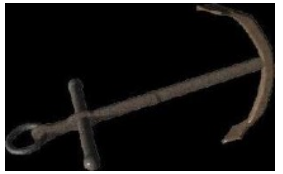


Credit: Kristen Grauman

Caltech 101: algorithms

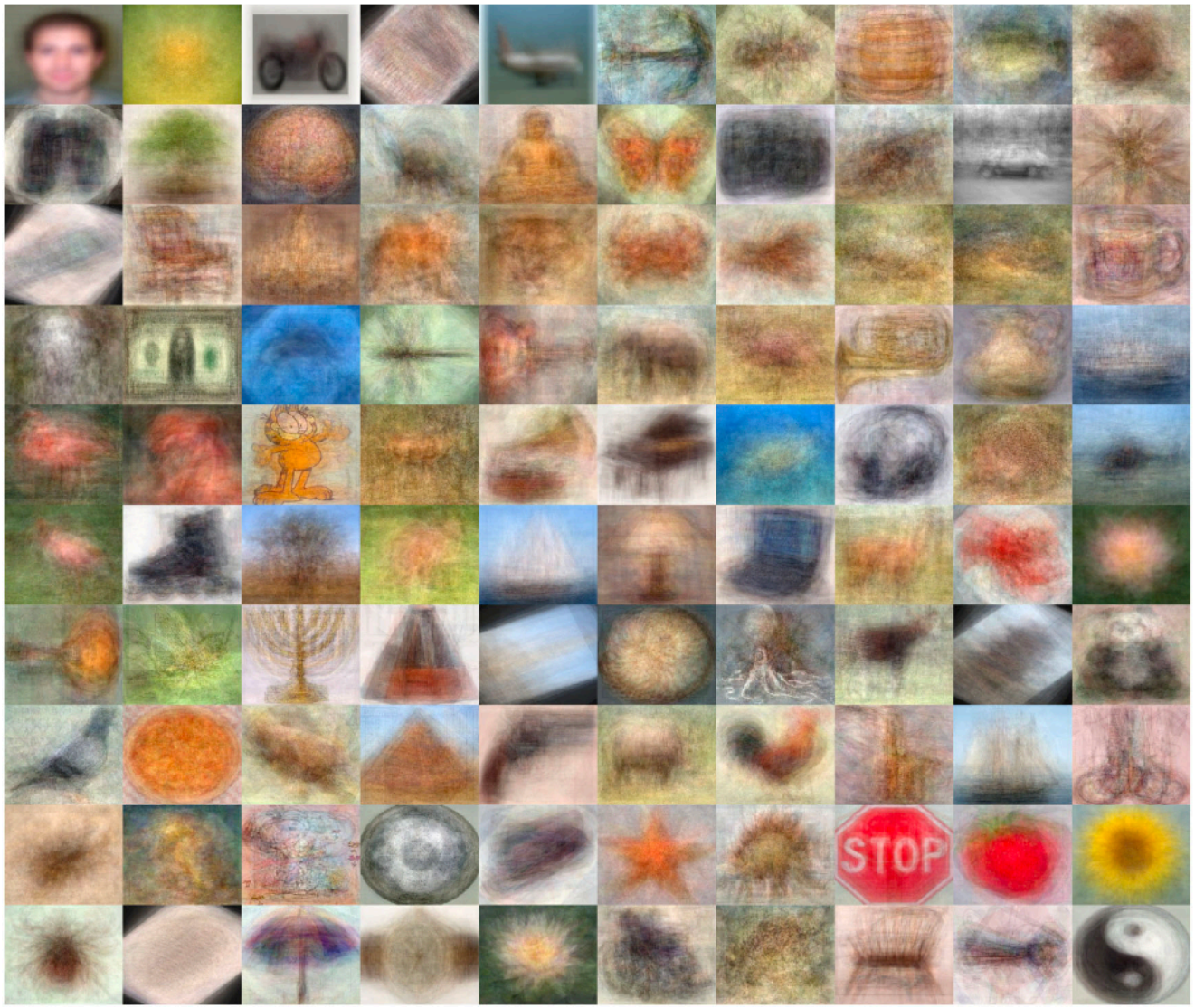
- How well would Haar features work?
- How well would HOG features work?
- How well would a SIFT bag of words model work?
- How well would a SIFT Spatial Pyramid Model work?



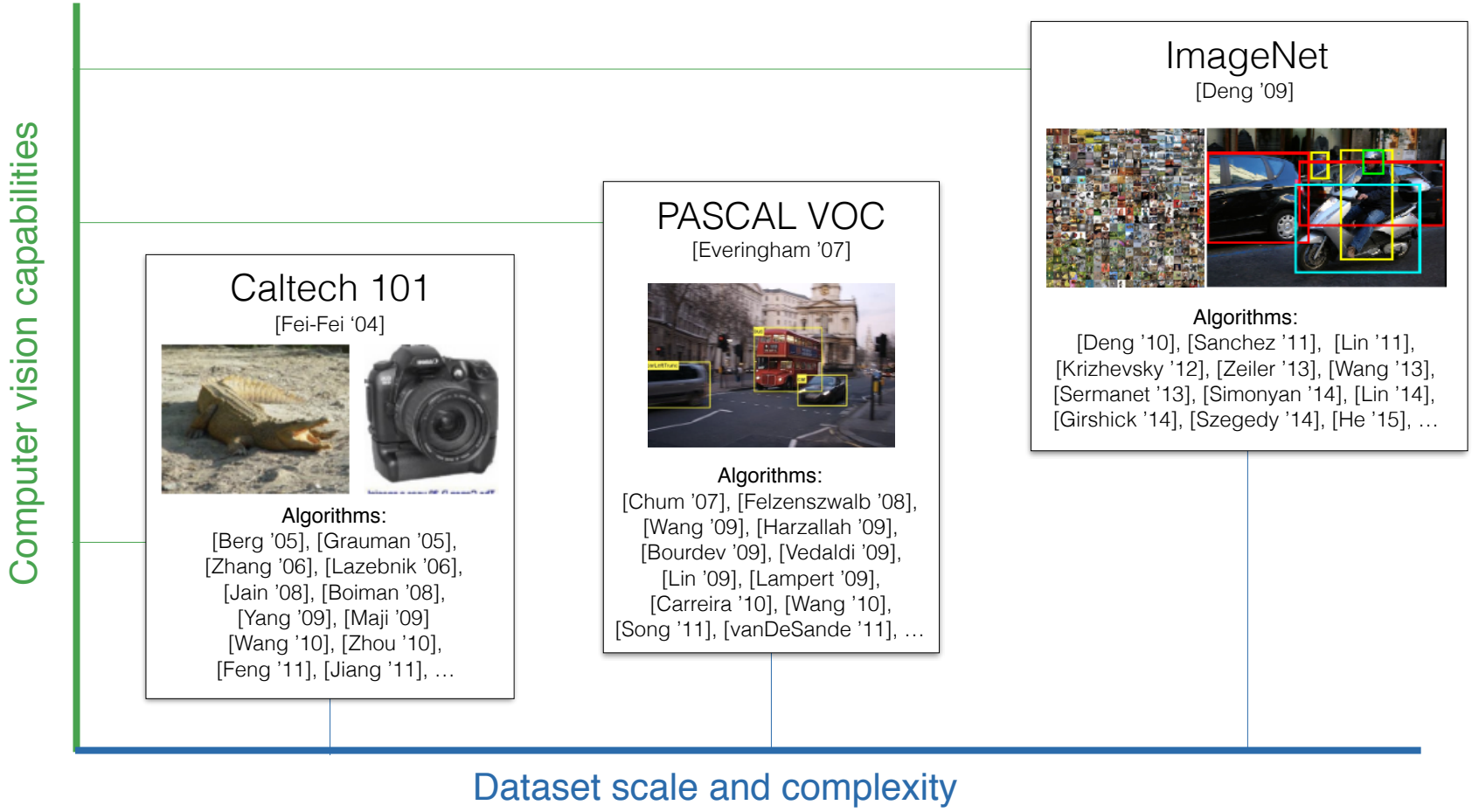





Caltech 101: average images



Datasets drive computer vision progress



PASCAL VOC benchmark

The PASCAL Visual Object Classes Homepage

The PASCAL VOC project:

- Provides standardised image data sets for object class recognition
- Provides a common set of tools for accessing the data sets and annotations
- Enables evaluation and comparison of different methods
- Ran challenges evaluating performance on object class recognition (from 2005-2012, now finished)

Pascal VOC data sets

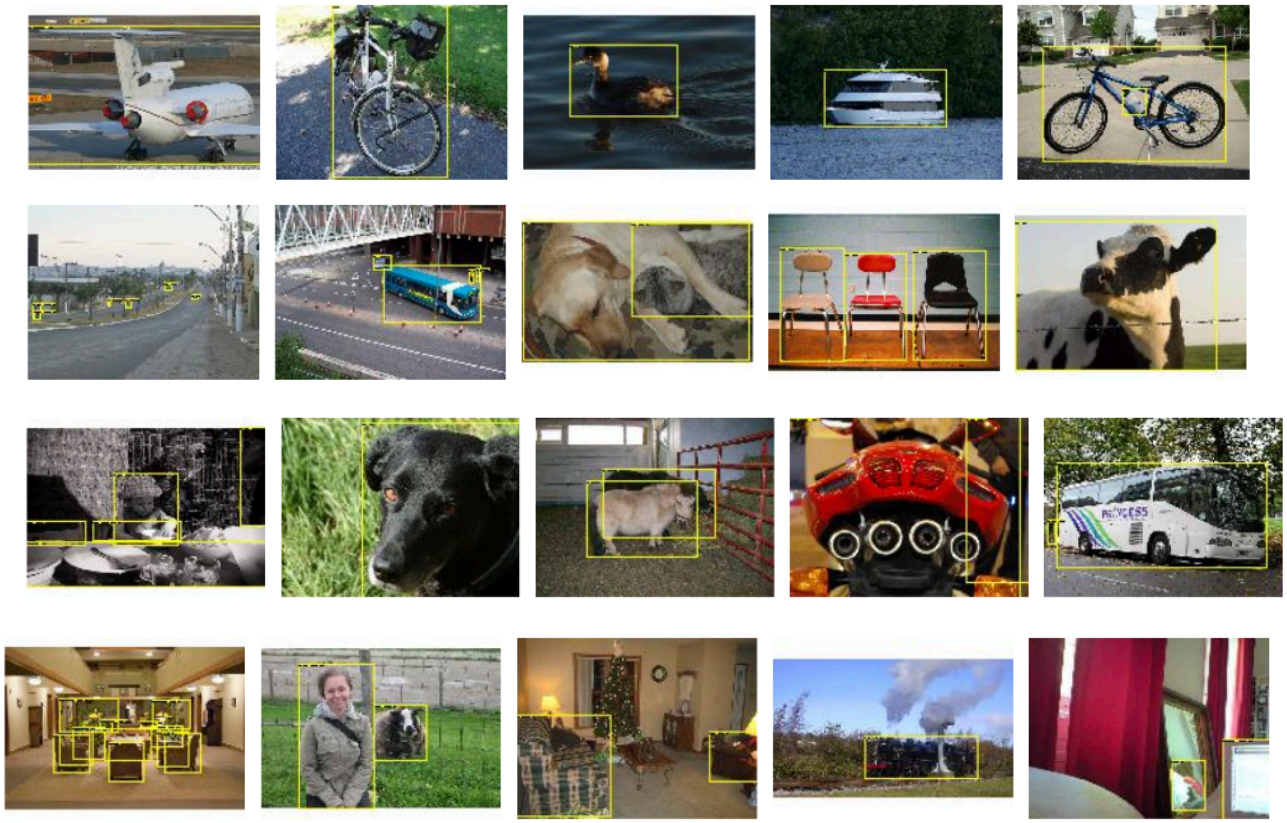
Data sets from the VOC challenges are available through the challenge links below, and evalution of new methods on these data sets can be achieved through the PASCAL VOC Evaluation Server. The evaluation server will remain active even though the challenges have now finished.

News

- **Nov-2014**: A new feature for the Leaderboards of the PASCAL VOC evaluation server has been added, indicating if the differences between a selected submission and others are statistically significant or not.
- **May-2014**: A new paper covering the 2008-12 years of the challenge, and lessons learnt, is now available:

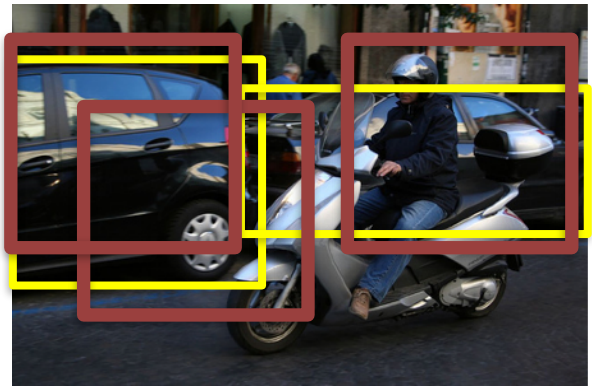
  The PASCAL Visual Object Classes Challenge: A Retrospective
  Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.
  International Journal of Computer Vision, 111(1), 98-136, 2015
  Bibtex source | Abstract | PDF

PASCAL VOC benchmark



- *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

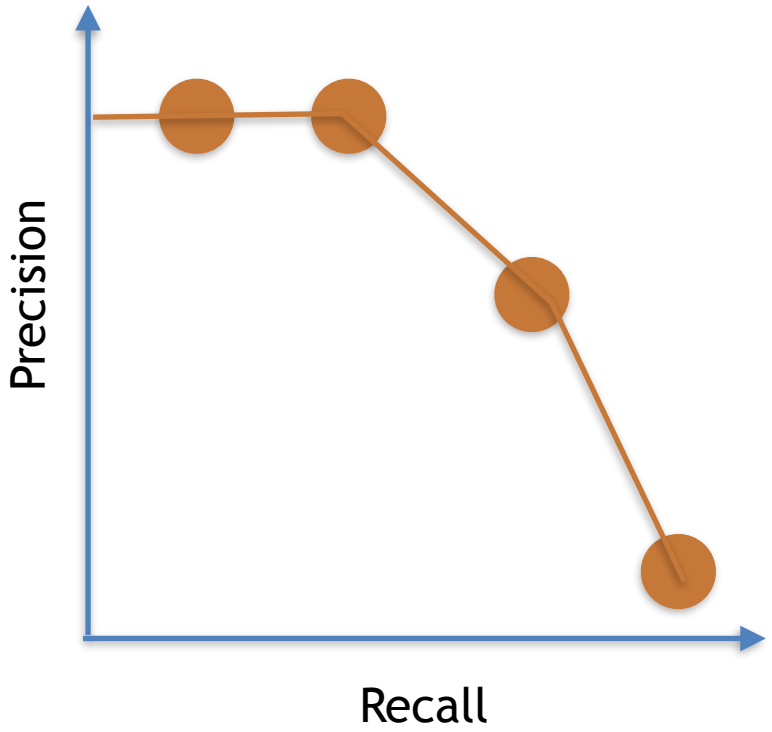
Object detection evaluation: average precision



- give a “car” score to each window

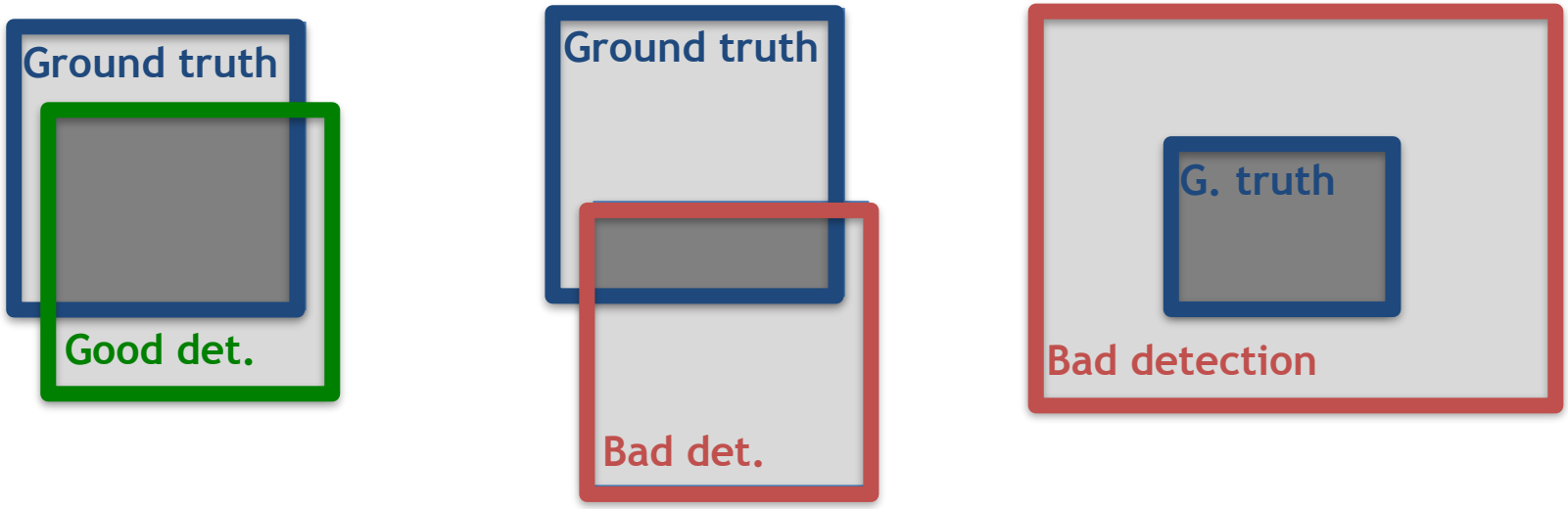
$$Recall = \frac{NumTruePositives}{NumPositives}$$

$$Precision = \frac{NumTruePositives}{NumPredictions}$$



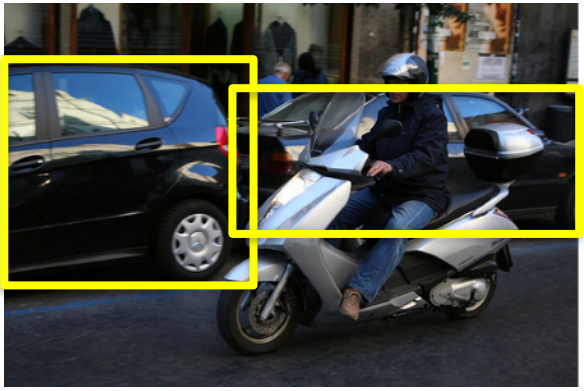
Threshold for Correct Detection

$$\frac{\text{Intersection}}{\text{Union}} \geq 0.5$$



Object detection evaluation

All instances of all target object classes expected to be localized on all test images



- Algorithm outputs a list of bounding box detections with confidences
- A detection is considered correct if overlap with ground truth is big enough
  - duplicate detections are penalized
- Evaluated by average precision (AP) per object class
- Overall evaluated usually by mAP
  - In competitions, also by number of classes won

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

Object detection is a collection of problems

Intra-class Variation for “Airplane”

Occlusion




Shape




Viewpoint




Distance



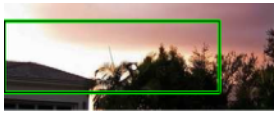


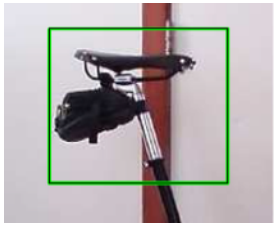
Credit: Derek Hoiem

Object detection is a collection of problems

Confusing Distractors for “Airplane”

Background



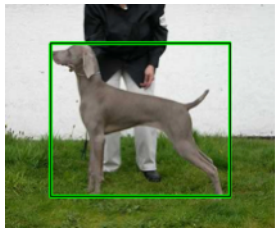
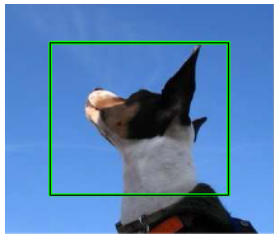


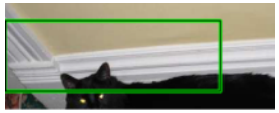
Similar Categories



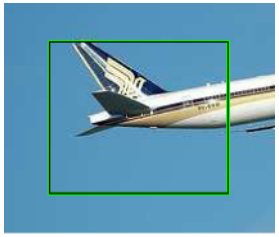


Dissimilar Categories




Localization Error





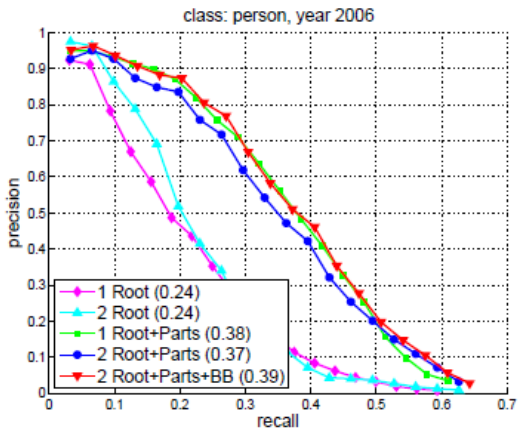

Credit: Derek Hoiem

Average precision evaluation is not enough

- Average Precision (AP)
  - Good summary statistic for quick comparison
  - Not a good driver of research

	aero	bike	bird	boat	bottle	bus
a) base	.290	.546	.006	.134	.262	.394
b) BB	.287	.551	.006	.145	.265	.397
c) context	.328	.568	.025	.168	.285	.397

Typical evaluation through comparison of AP numbers



- Need tools to determine
  - where detectors fail
  - potential impact of particular improvements

Credit: Derek Hoiem

figs from Felzenszwalb et al. 2010

Tool for object detection analysis

Diagnosing Error in Object Detectors

Derek Hoiem and Qieyun Dai and Yodsawalai Chodpathumwan
Computer Vision Group
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

This work provides a set of tools for analyzing object detector performance.

Note: (11/12/14) The summary plots (e.g., "animal" or "vehicle") for displayDetectionTrend were computed incorrectly. The revised code is now in the .tar.gz file, but the pdfs have not been updated. Thanks to Shaoqing Ren for noticing the bug and providing the fix. Another method displayDetectionTrend2.m is also provided, which averages across tic marks to summarize several categories.

Downloads

The following resources are available:

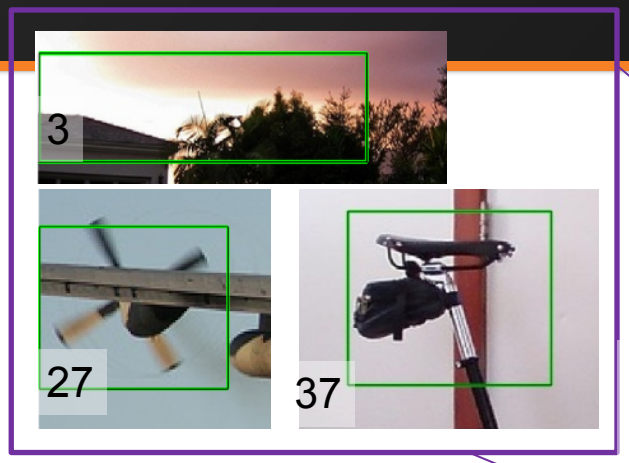
- An updated version (v2) of the code/annotations: [src/data (84.5MB)]
- Description of updates: [pdf]
- Examples of automatic analysis reports: [dpm_v4.pdf] [vedaldi2009.pdf] [cnn7_bb.pdf]
- Original version (v1) of the code/annotations (in case you have trouble with the new version): [src/data (69MB)]

Publications

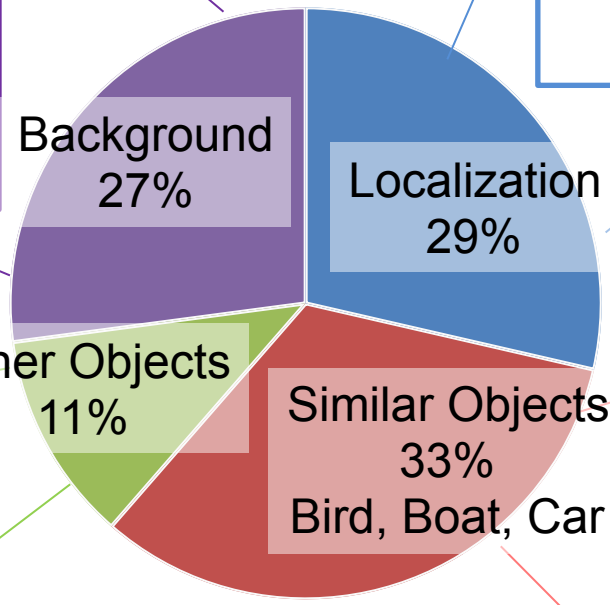
Diagnosing Error in Object Detectors
Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai
ECCV, 2012. [pdf] [slides]

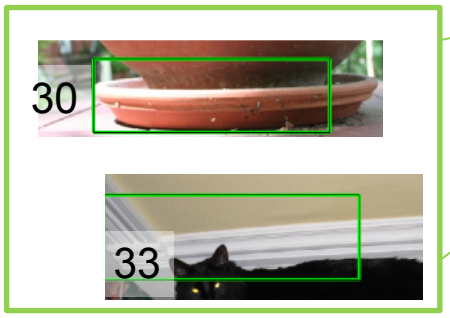
http://dhoiem.web.engr.illinois.edu/projects/detectionAnalysis/

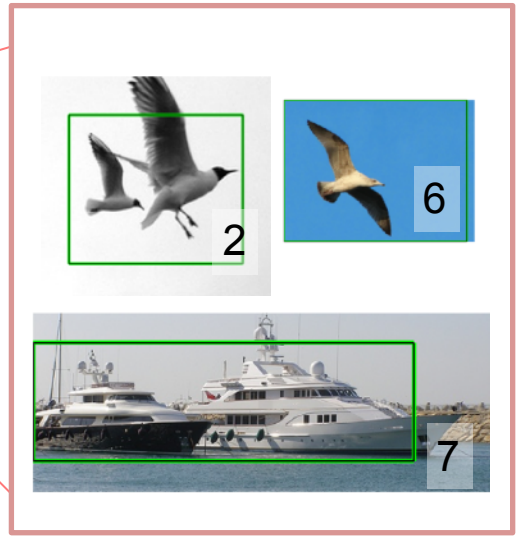
Top false positives: Airplane (DPM)



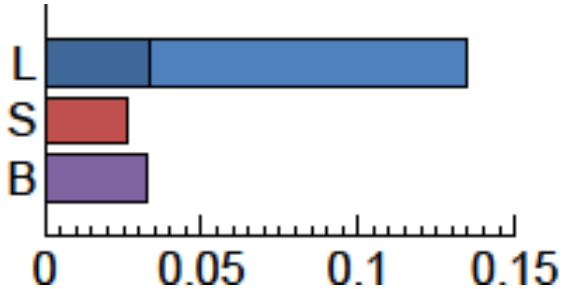
AP = 0.36





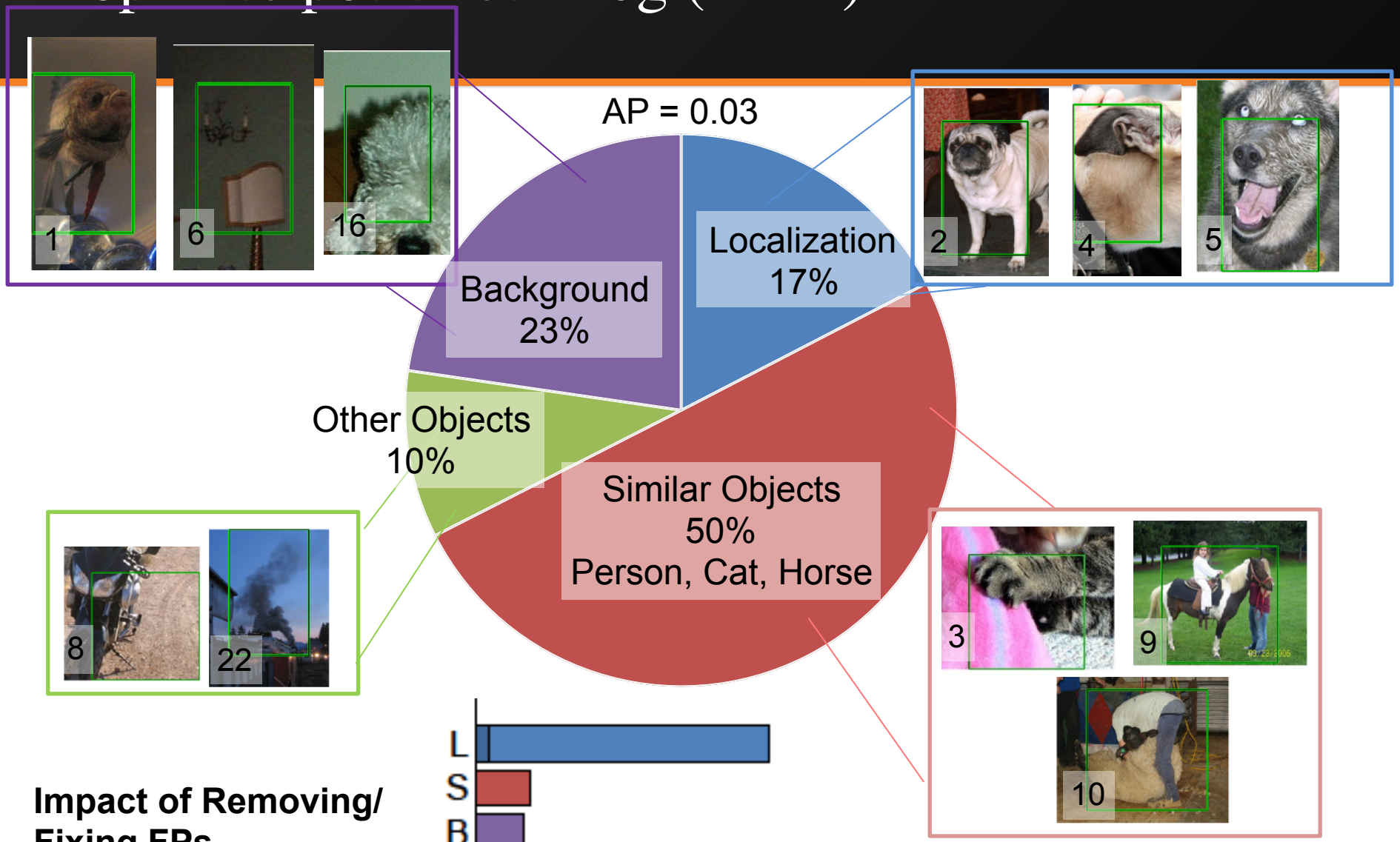


Impact of Removing/ Fixing FPs



Credit: Derek Hoiem

Top false positives: Dog (DPM)

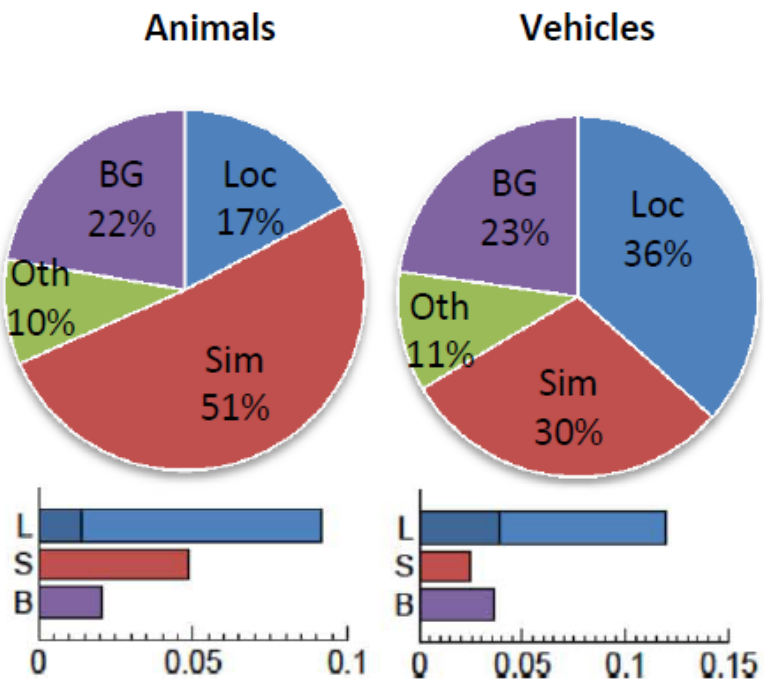


Impact of Removing/ Fixing FPs

Credit: Derek Hoiem

Summary of False Positive Analysis

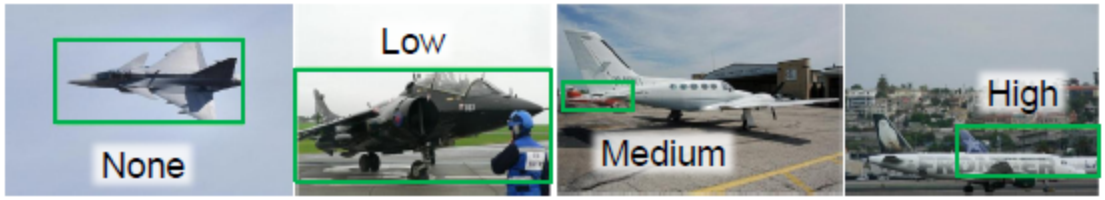
DPM v4
(FGMR 2010)



http://dhoiem.web.engr.illinois.edu/projects/detectionAnalysis/ Credit: Derek Hoiem

Analysis of object characteristics

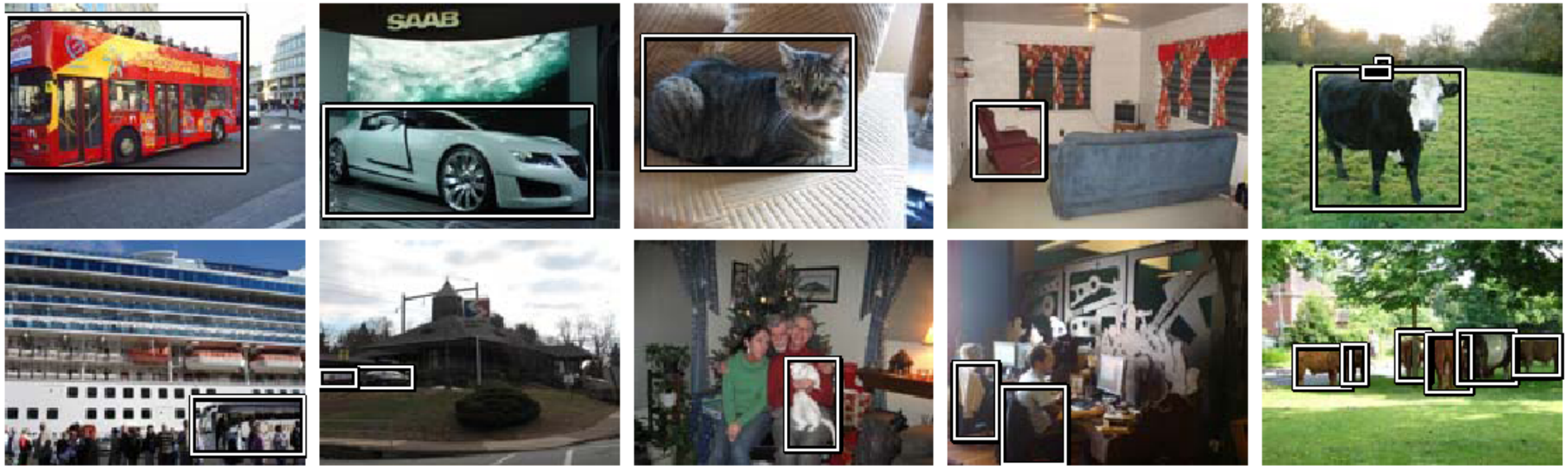
Level of occlusion: 2 (moderate)
Parts visible: bike body, handlebars, wheel
Parts not visible: seat
View: side visible (front, top, etc., not visible)
Area: 3233 pixels
Aspect Ratio (w/h): 1.24



Occlusion Level

http://dhoiem.web.engr.illinois.edu/projects/detectionAnalysis/ Credit: Derek Hoiem

PASCAL VOC challenge



- 20 categories
- Annual (2005-2012) classification, detection, segmentation, … challenges

Credit: Noah Snavely

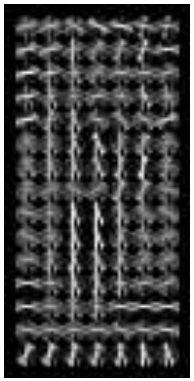
Machine learning for object detection

- What features do we use?
  - intensity, color, gradient information, …
- Which machine learning methods?
  - generative vs. discriminative
  - k-nearest neighbors, boosting, SVMs, …
- What hacks do we need to get things working?

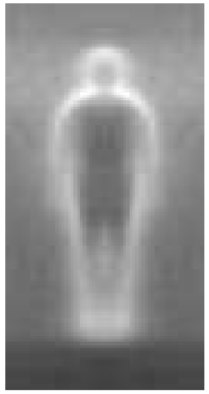
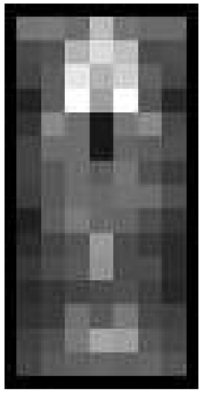
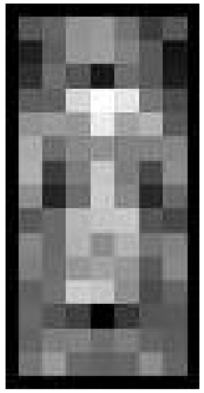
Credit: Noah Snavely

Person detection, ca. 2005 (Dalal Triggs)

1. Represent each example with a single, fixed HoG template

2. Learn a single [linear] SVM as a detector

Credit: Noah Snavely

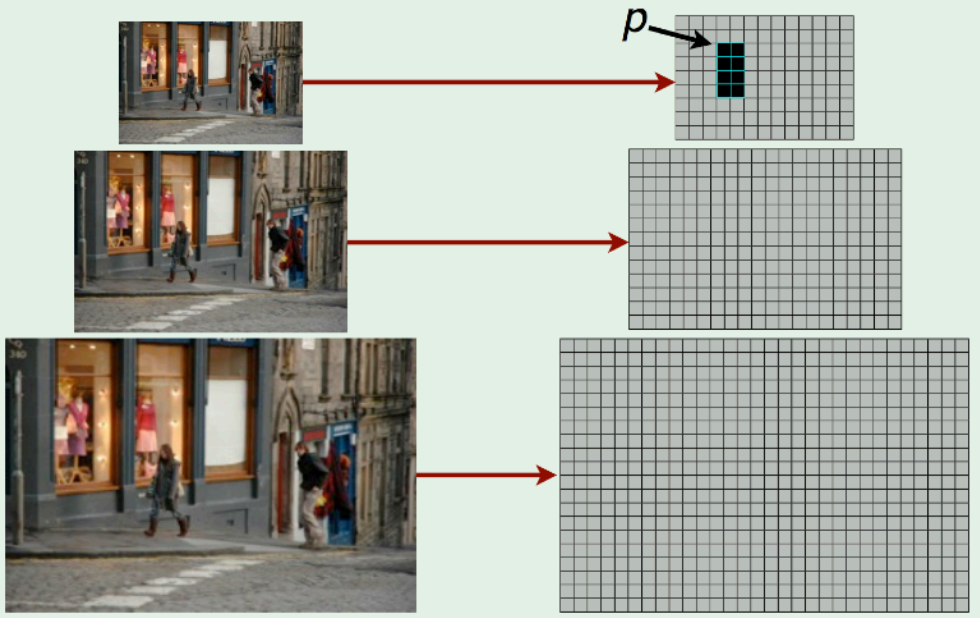
Positive and negative examples

+ thousands more…

+ millions more…

Credit: Noah Snavely

Sliding window detection

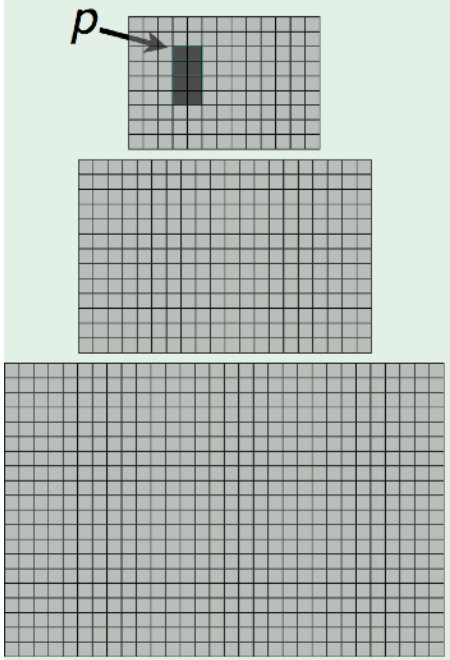


$$score(I, p) = \mathbf{w} \cdot \phi(I, p)$$

- Compute HOG of the whole image at multiple resolutions
- Score every subwindow of the feature pyramid
- Apply non-maxima suppression

Credit: Ross Girshick

Detection



number of locations p ~ 250,000 per image

test set has ~ 5000 images

>> 1.3×10^9 windows to classify

typically only ~ 1,000 true positive locations

Extremely unbalanced binary classification

Credit: Ross Girshick

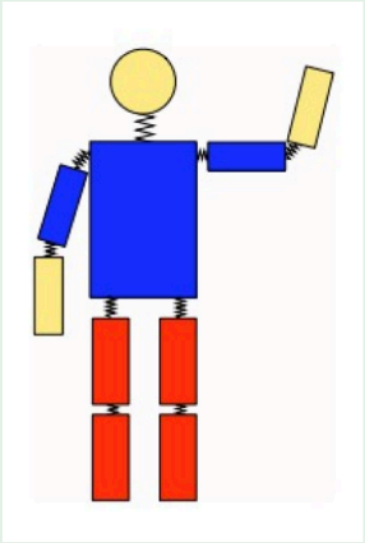
Dalal&Triggs on PASCAL VOC 2007

AP = 12%

Credit: Ross Girshick

Part-based models

- Parts — local appearance templates
- "Springs" — spatial connections between parts (geom. prior)



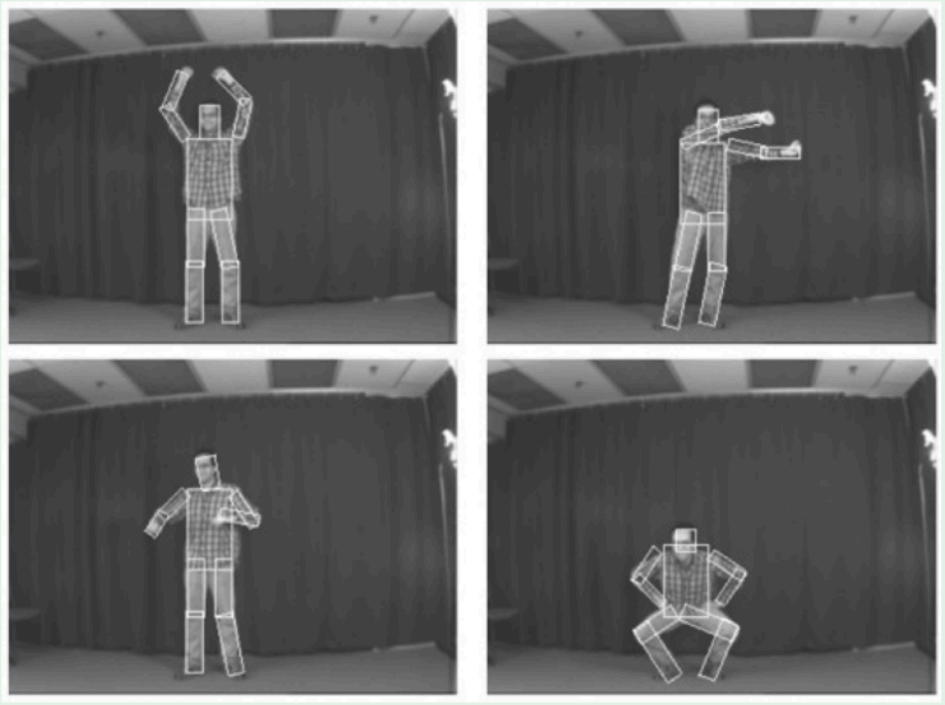
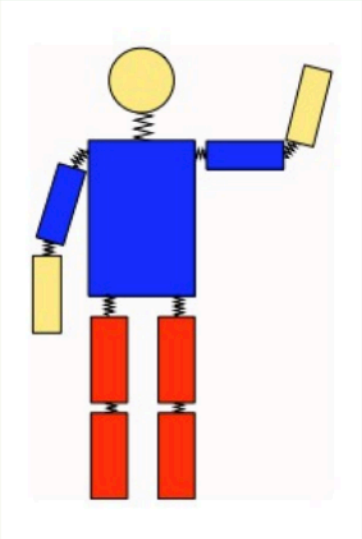


Image: [Felzenszwalb and Huttenlocher 05]

Credit: Ross Girshick

Part-based models

- Local appearance is easier to model than the global appearance
  - Training data shared across deformations
  - "part" can be local or global depending on resolution
- Generalizes to previously unseen configurations



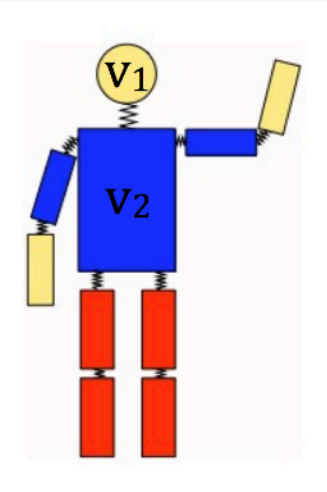
Credit: Ross Girshick

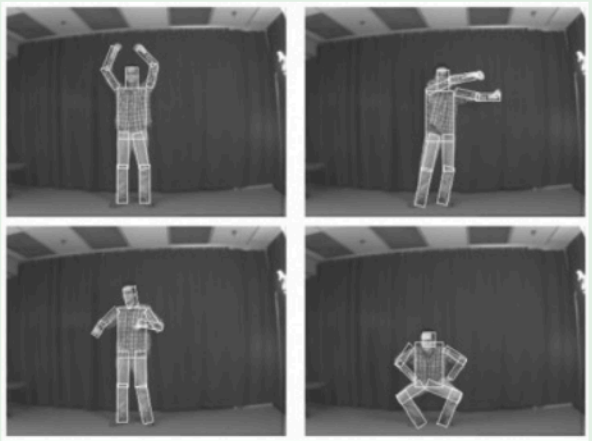
Part configuration score function

spring costs

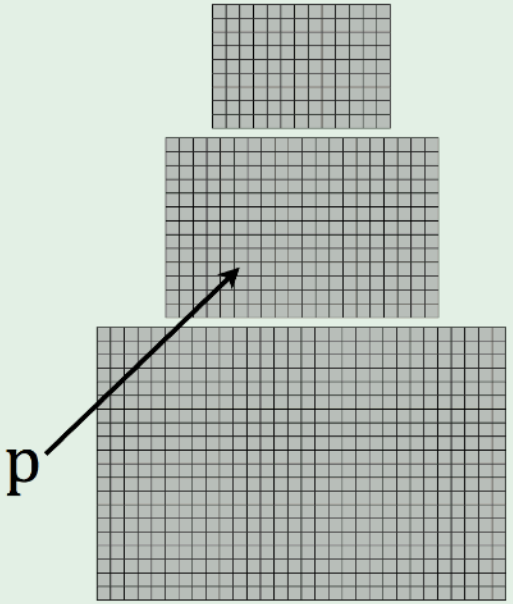
$$\text{score}(p_1, \ldots, p_n) = \sum_{i=1}^{n} m_i(p_i) - \sum_{(i,j) \in E} d_{ij}(p_i, p_j)$$

Part match scores





Highest scoring configurations



Credit: Ross Girshick

Part configuration score function

$$\text{score}(p_1, \ldots, p_n) = \sum_{i=1}^{n} m_i(p_i) - \sum_{(i,j) \in E} d_{ij}(p_i, p_j)$$

Part match scores

spring costs

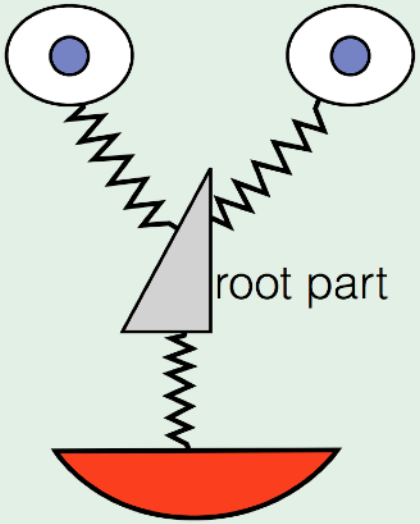
- Objective: maximize score over $p_1,\ldots,p_n$
- h^n configurations! ($h = |P|$, about 250,000)
- Dynamic programming

Object Detection with Discriminatively Trained Part Based Models Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan. PAMI 2010

Credit: Ross Girshick

Star-structured deformable part models

test image

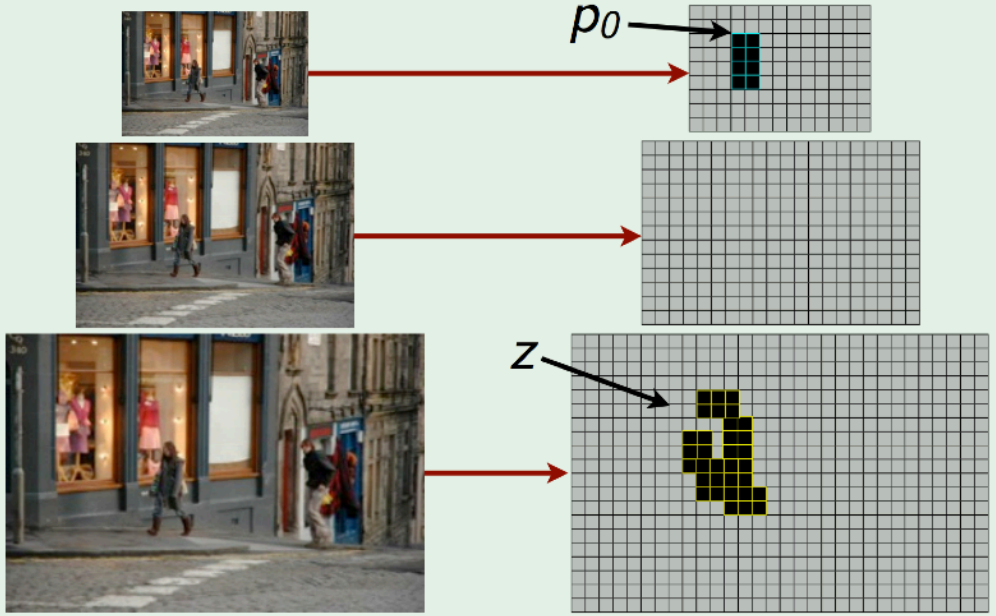


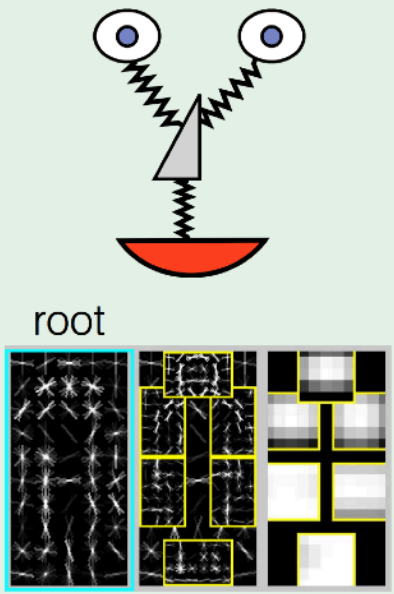
"star" model

detection

Credit: Ross Girshick

Dalal-Triggs + parts



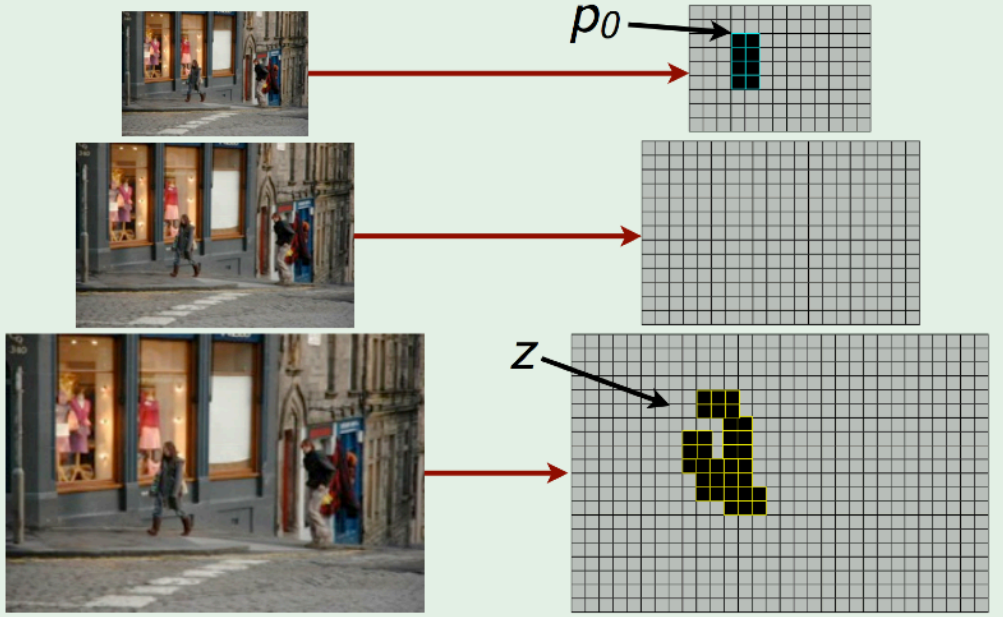


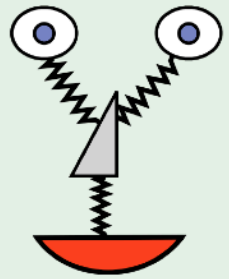
[FMR CVPR'08]
[FGMR PAMI'10]

- Add parts to the Dalal & Triggs detector
  - HOG features
  - Linear filters / sliding-window detector
  - Discriminative training

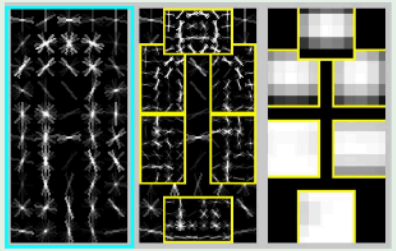
Credit: Ross Girshick

Sliding window DPM score function





root



[FMR CVPR'08]
[FGMR PAMI'10]

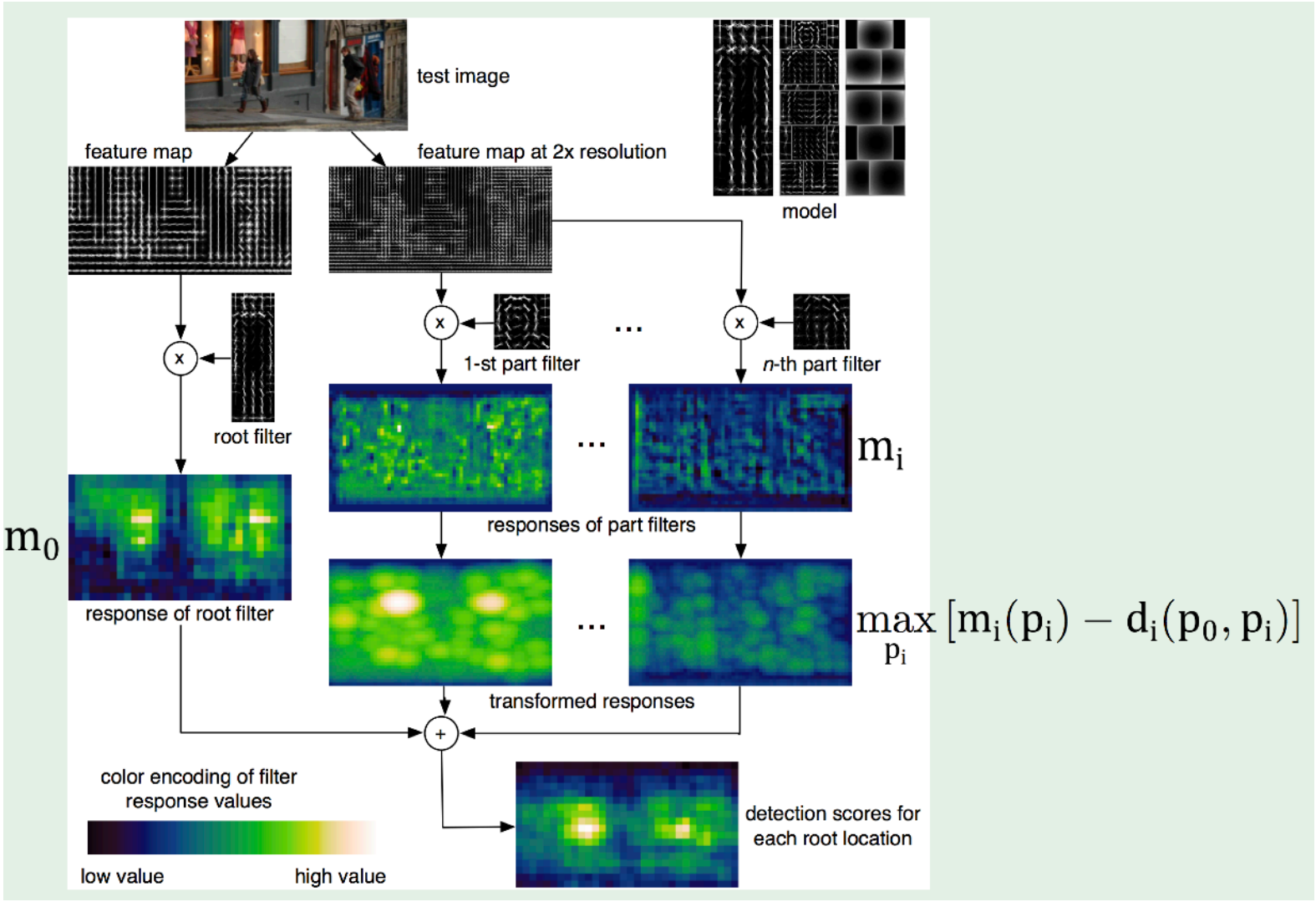
$$z = (p_1, \ldots, p_n)$$

$$\text{score}(I, p_0) = \max_{p_1,\ldots,p_n} \sum_{i=0}^{n} m_i(I, p_i) - \sum_{i=1}^{n} d_i(p_0, p_i)$$

Filter scores Spring costs

Credit: Ross Girshick

DPM detection in a slide

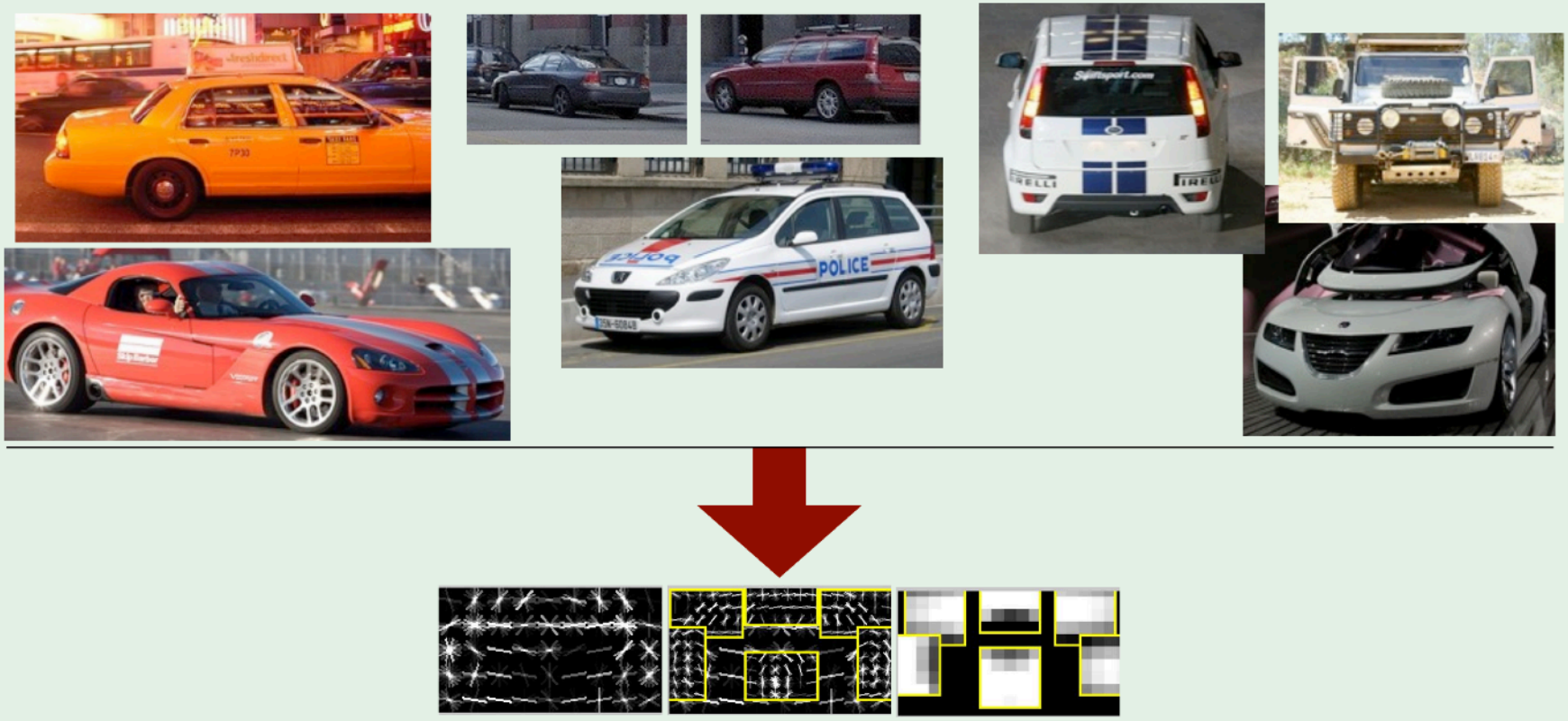


Credit: Ross Girshick

What are the parts?

Credit: Ross Girshick

Clustering by viewpoints (aspect ratios as proxy)

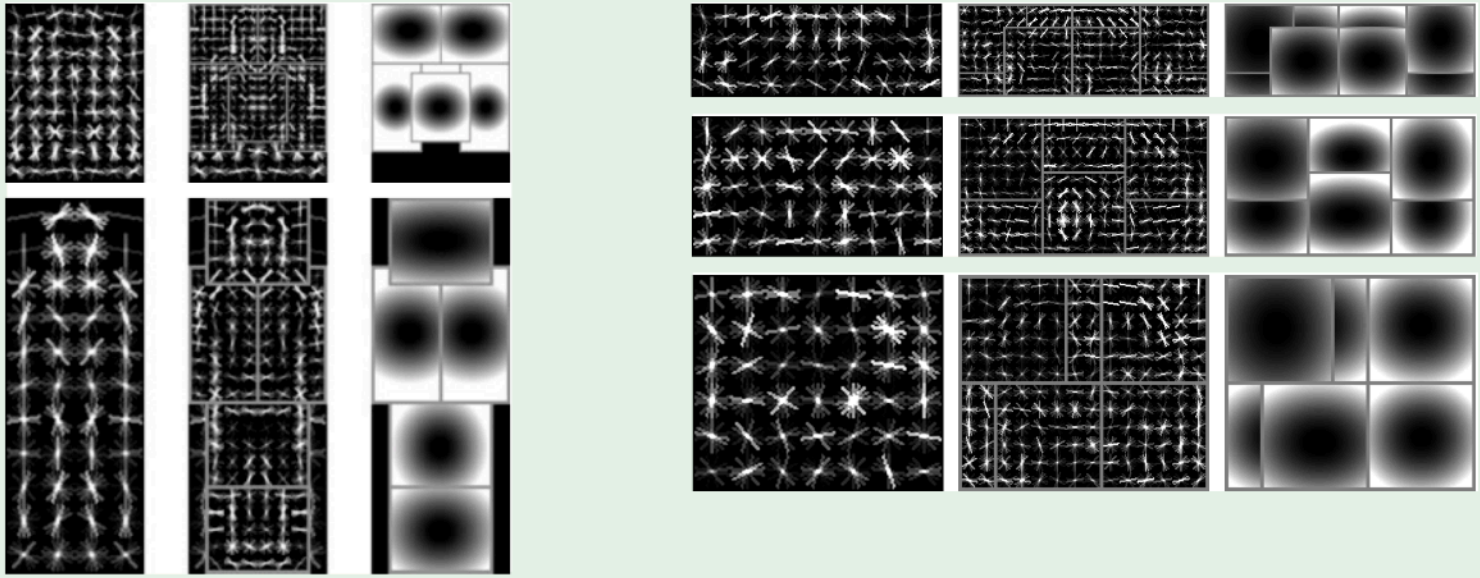


General philosophy: enrich models to better represent the data

Credit: Ross Girshick

DPM with mixture models

Data driven: aspect, occlusion modes, subclasses



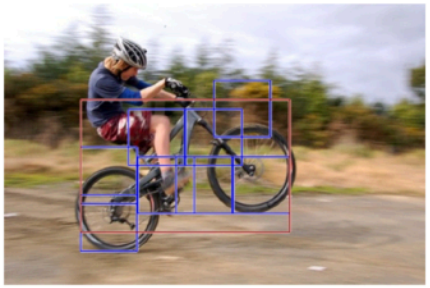
Person detection
Without parts: AP = 0.12
Parts but no mixtures: AP = 0.27
Parts + mixtures: AP = 0.36

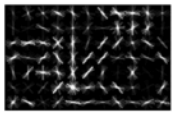
Credit: Ross Girshick

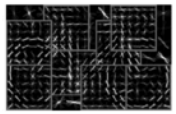
Discriminatively trained deformable parts model

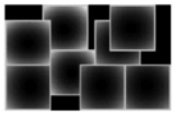
Discriminatively trained deformable part models

Version 5 (Sept. 5, 2012)









Introduction

Over the past few years we have developed a complete learning-based system for detecting and localizing objects in images. Our system represents objects using mixtures of deformable part models. These models are trained using a discriminative method that only requires bounding boxes for the objects in an image. The approach leads to efficient object detectors that achieve state of the art results on the PASCAL and INRIA person datasets.

At a high level our system can be characterized by the combination of

1. Strong low-level features based on histograms of oriented gradients (HOG)
2. Efficient matching algorithms for deformable part-based models (pictorial structures)
3. Discriminative learning with latent variables (latent SVM)

This work was awarded the PASCAL VOC "Lifetime Achievement" Prize in 2010.

Code: http://www.rossgirshick.info/latent/
Slides: http://vision.stanford.edu/teaching/cs231b_spring1213/slides/dpm-slides-ross-girshick.pdf
Paper: Object Detection with Discriminatively Trained Part Based Models
Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan. PAMI 2010

Where would DPM succeed and fail?

- *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

aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	*mean*
33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	*33.7*

http://www.rossgirshick.info/latent/

How would DPM compare to SIFT+SPM?

(not a perfect comparison by any means, but an attempt)

	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
UOC_OXFORD_DPM_MKL	59.6	54.5	21.9	21.6	32.1	52.5	49.3	40.8	19.1	35.2	28.9	37.2	50.9	49.9	46.1	15.6	39.3	35.6	48.9	42.8
NEC_STANFORD_OCP	65.1	46.8	25.0	24.6	16.0	51.0	44.9	51.5	13.0	26.6	31.0	40.2	39.7	51.5	32.8	12.6	35.7	33.5	48.0	44.8

UOC_OXFORD_DPM_MKL

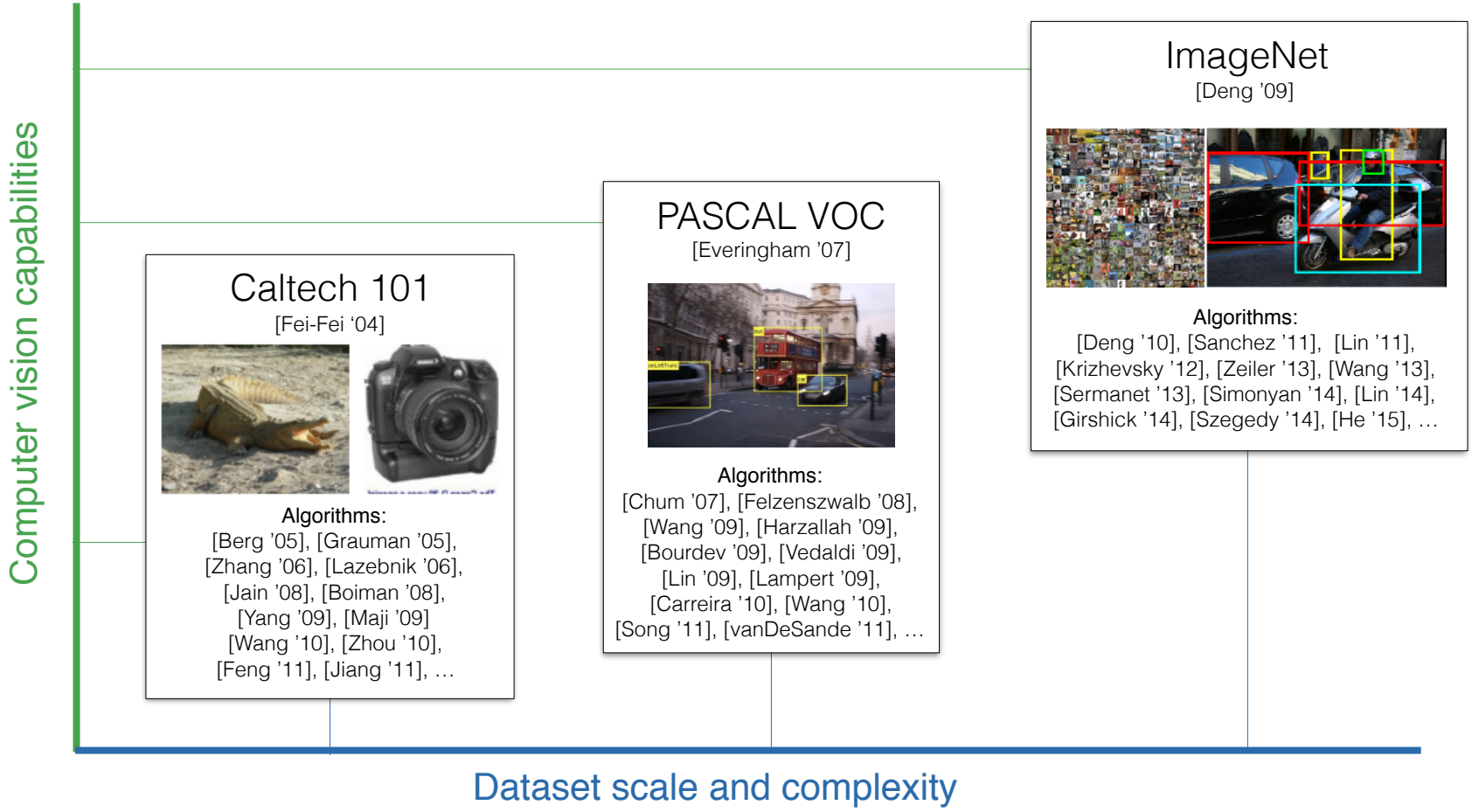
This method is similar to last year DPM-MKL entry. We updated several aspects of the implementation (e.g. the type of features).

NEC_STANFORD_OCP

Object-centric pooling (OCP) is a method which represents a bounding box by pooling the coded low-level descriptors on the foreground and background separately and then concatenating them (Russakovsky et al. ECCV 2012). This method exploits powerful classification features that have been developed in the past years. In this system, we used DHOG and LBP as low-level descriptors. We developed a discriminative LCC coding scheme in addition to traditional LCC coding. We make use of candidate bounding boxes (van de Sande et al. ICCV 2011).

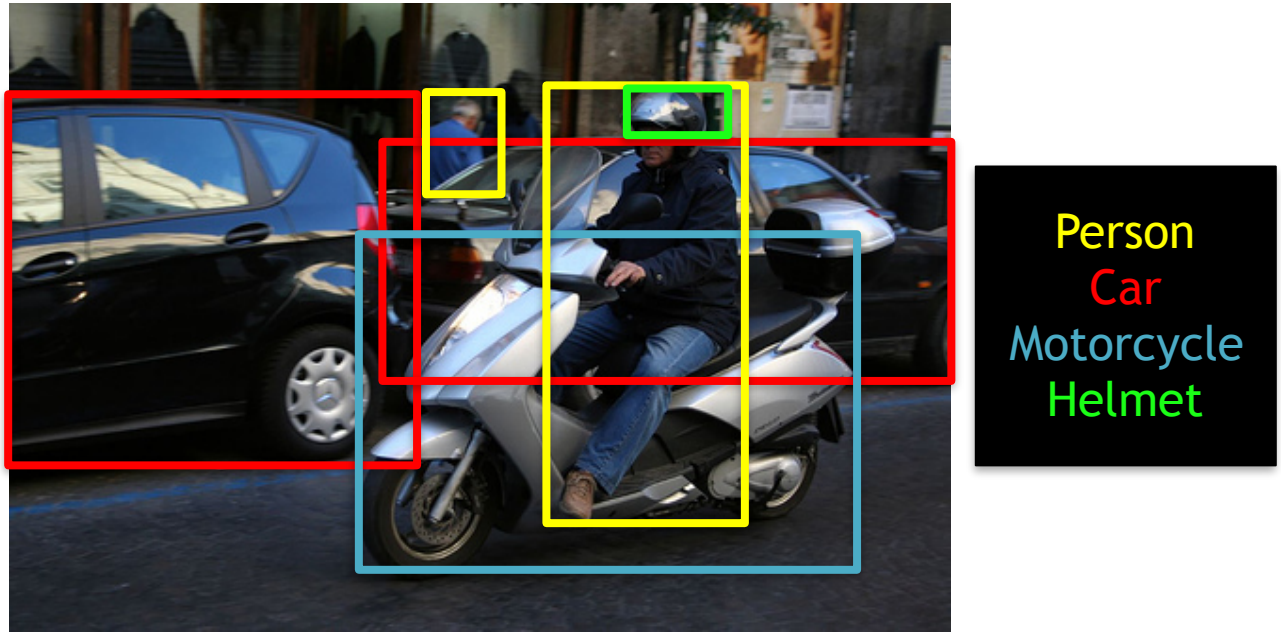
http://host.robots.ox.ac.uk/pascal/VOC/voc2012/results/index.html

Datasets drive computer vision progress

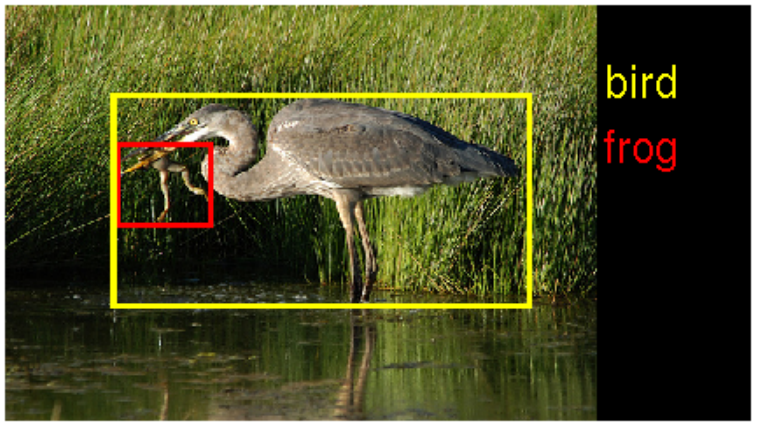


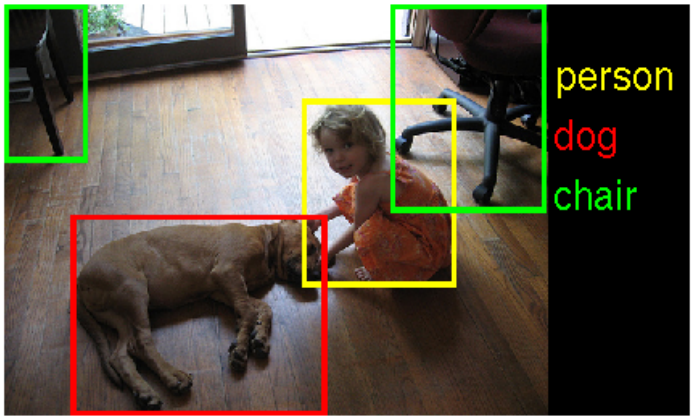
ILSVRC object detection task

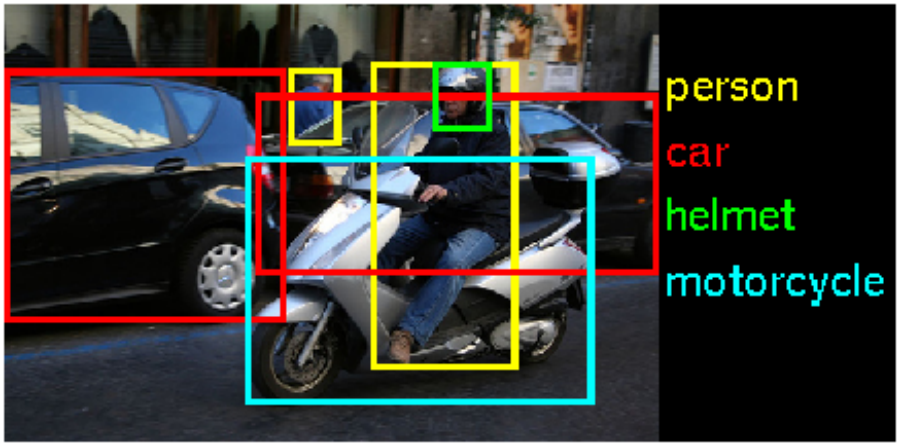
Allows evaluation of generic object detection in cluttered scenes at scale



ILSVRC object detection data







ILSVRC object detection data

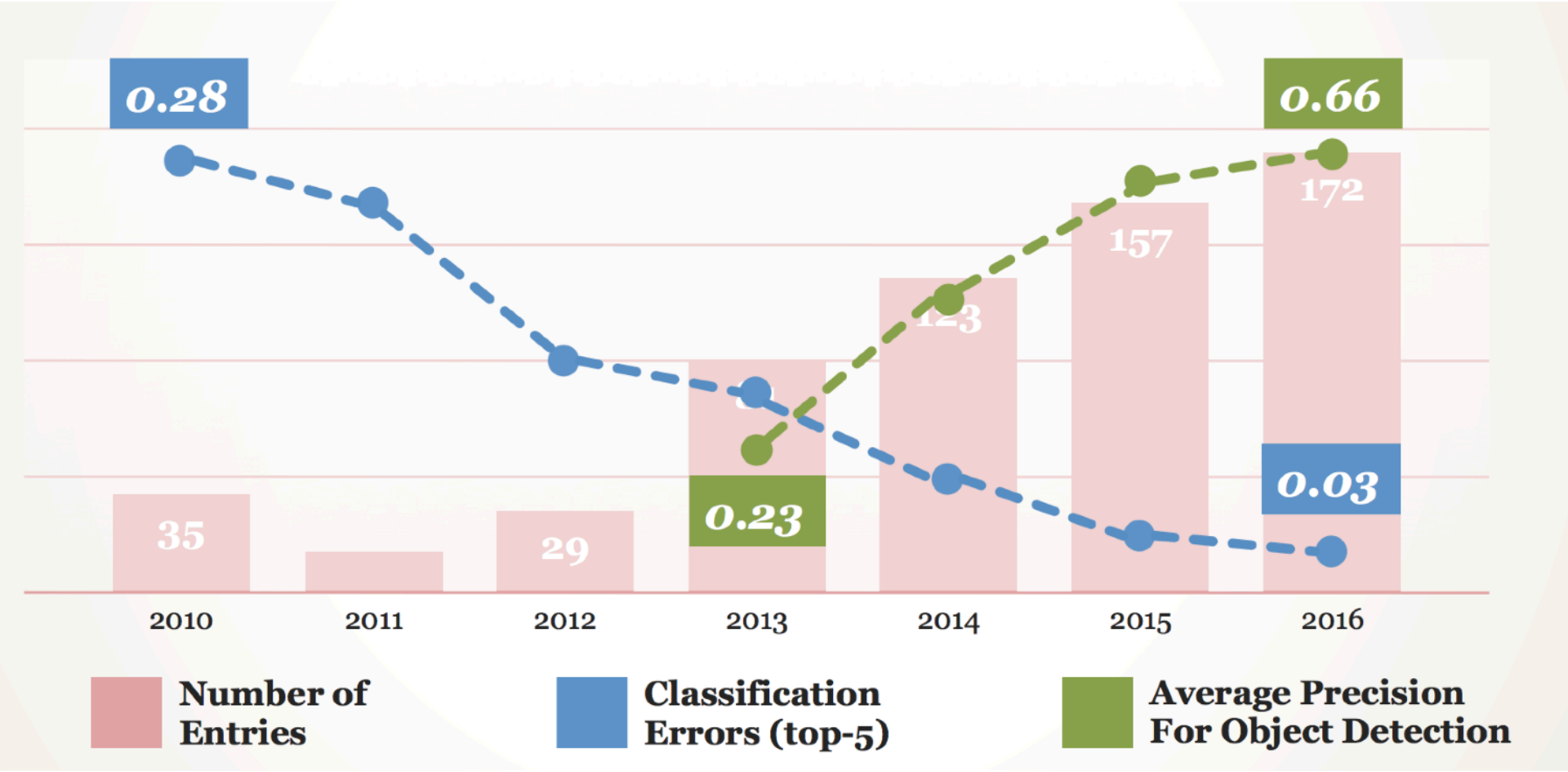
Comparative scale

		PASCAL VOC 2012	ILSVRC 2014
Number of object classes		20	**200**
Training	Num images	5717	456567
	Num objects	13609	478807
Validation	Num images	5823	20121
	Num objects	13841	55502
Testing	Num images	10991	40152
	Num objects	---	---

Comparative statistics (on validation set)

	PASCAL VOC 2012	ILSVRC 2013
Average image resolution	469x387 pixels	482x415 pixels
Average object classes per image	1.521	1.534
Average object instances per image	2.711	2.758
Average object scale (bounding box area as fraction of image area)	0.207	0.170

ImageNet challenge: participation and performance



Credit: Fei-Fei Li & Jia Deng

Easiest and hardest categories

Olga Russakovsky*, Jia Deng*, et al. **ImageNet Large Scale Visual Recognition Challenge**. *IJCV,* 2015.

What are the remaining challenges?

Number of concepts

Faces

1

PASCAL VOC

[Everingham et al. IJCV10]

20

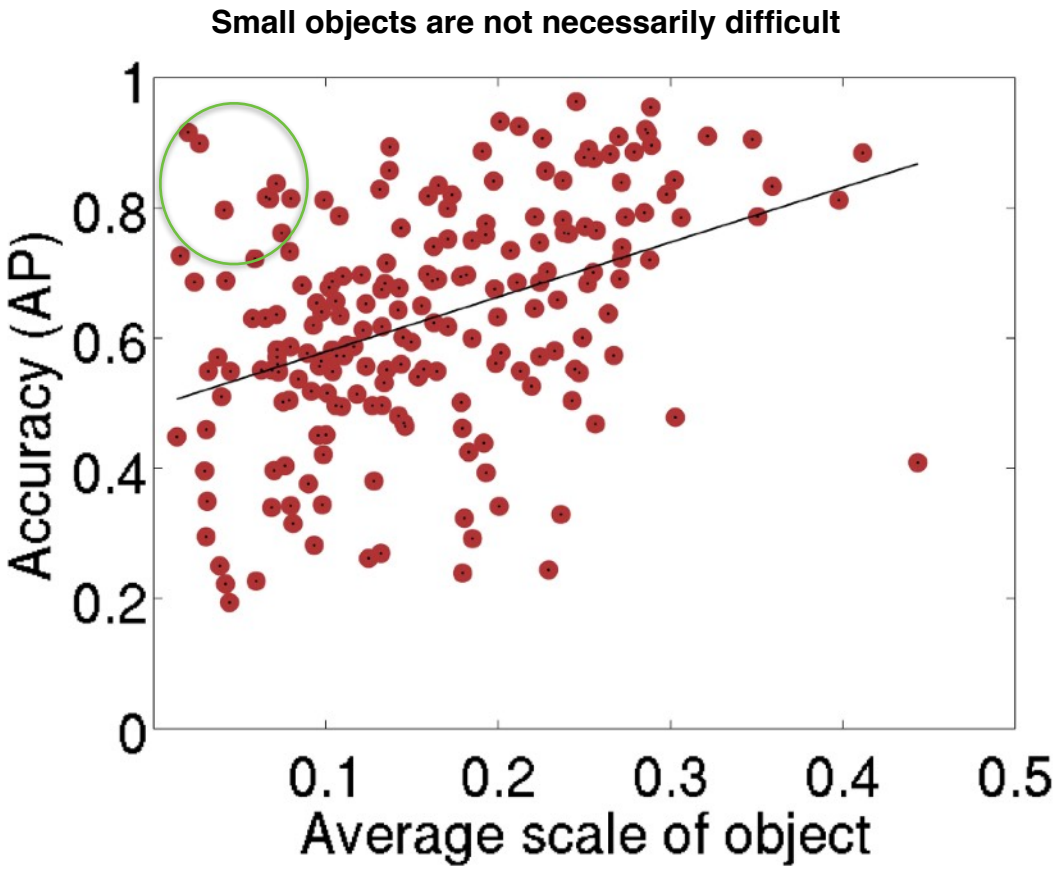
ImageNet challenge

[Russakovsky et al. IJCV15]

200

Impact of object scale on detection accuracy

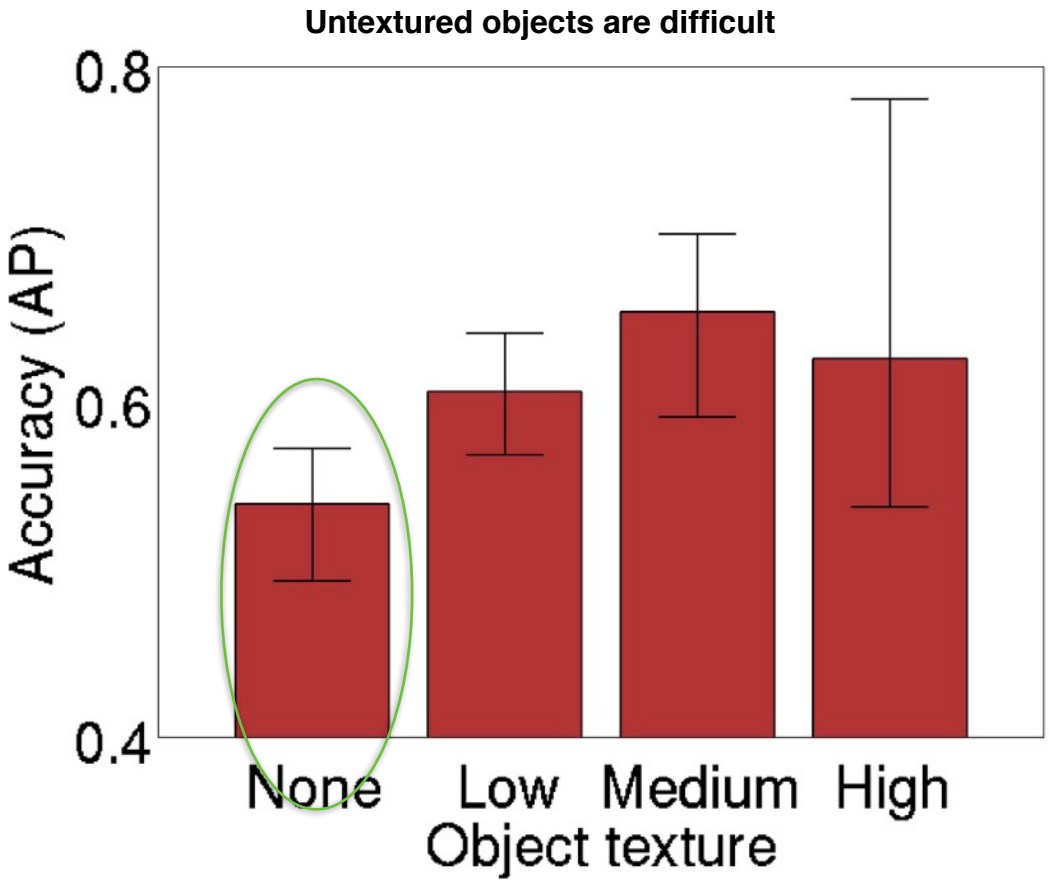
(ImageNet challenge 2013-2015 winning object detection entries)



[Russakovsky et al. IJCV 2015]

Impact of object scale on detection accuracy

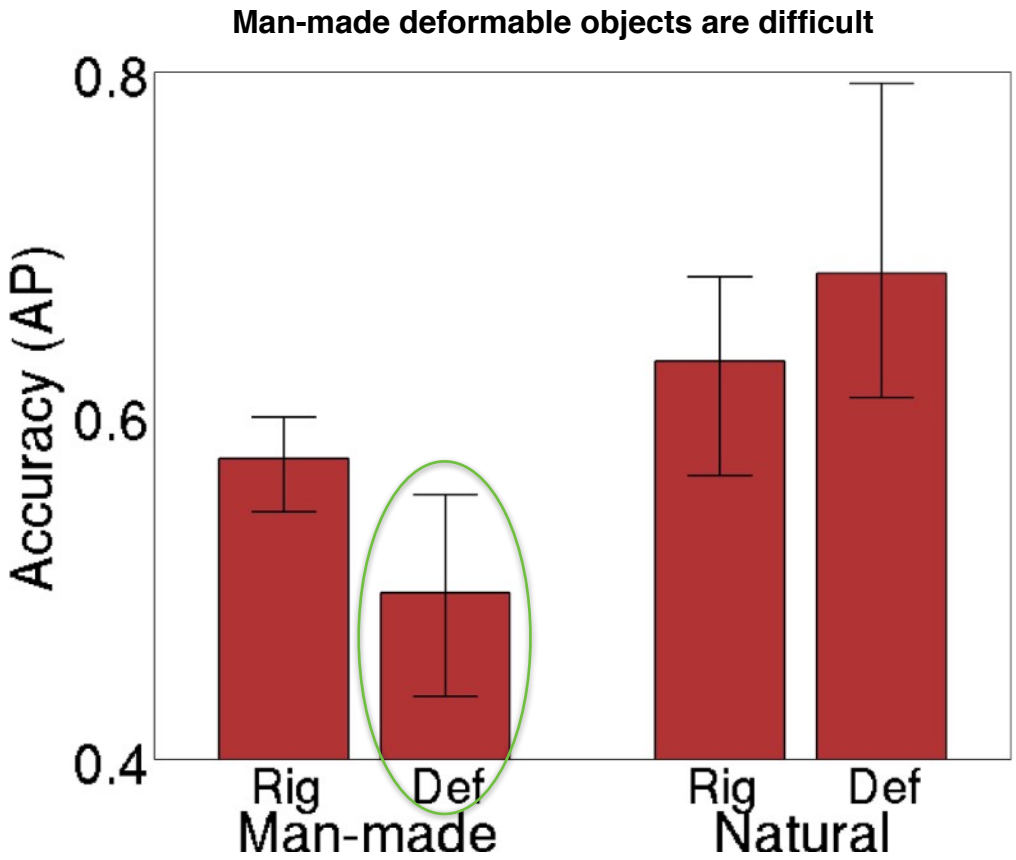
(ImageNet challenge 2013-2015 winning object detection entries)



[Russakovsky et al. IJCV 2015]

Impact of object scale on detection accuracy

(ImageNet challenge 2013-2015 winning object detection entries)



[Russakovsky et al. IJCV 2015]

ILSVRC data, challenge, algorithms

http://image-net.org/challenges/LSVRC/

Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) **ImageNet Large Scale Visual Recognition Challenge**. *IJCV,* 2015.
https://arxiv.org/abs/1409.0575

We'll come back to this in the deep learning section of the course

Next class: texture