The DPM Detector

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan

Object Detection with Discriminatively Trained Part Based Models

T-PAMI, 2010

Paper: `http://cs.brown.edu/~pff/papers/lsvm-pami.pdf`
Code: `http://www.cs.berkeley.edu/~rbg/latent/`

The HOG Detector

- The HOG detector models an object class as a single rigid template

Figure: Single HOG template models people in upright pose.

But Objects Are Composed of Parts

Even Rigid Objects Are Composed of Parts

Objects Are Composed of Deformable Parts

- Revisit the old idea by Fischler & Elschlager 1973
- Objects are composed of parts at specific **relative locations**. Our model should probably also model object parts.
- Different instances of the same object class have parts in slightly different locations. Our object model should thus allow slight **slack** in part position.

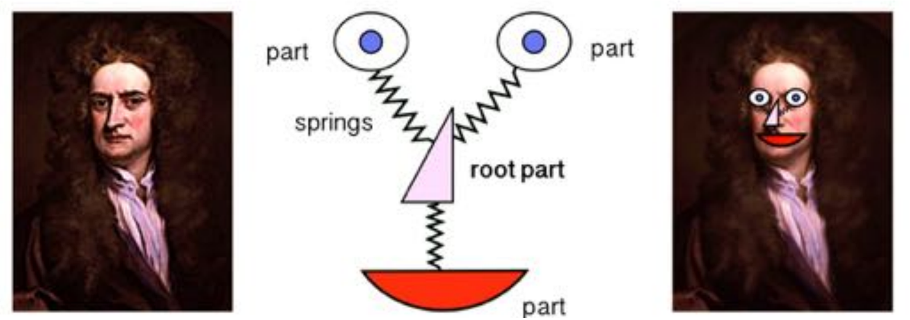
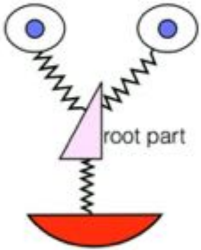


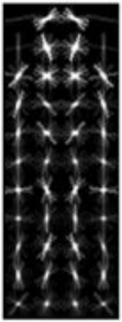
Figure: Objects are a collection of deformable parts

[Pic from: R. Girshik]

The DPM Model

- The DPM model starts by borrowing the idea of the HOG detector. It takes a HOG template for the full object. (If you take something that works, things can only get better, right?)

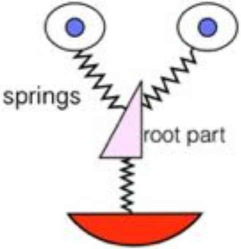


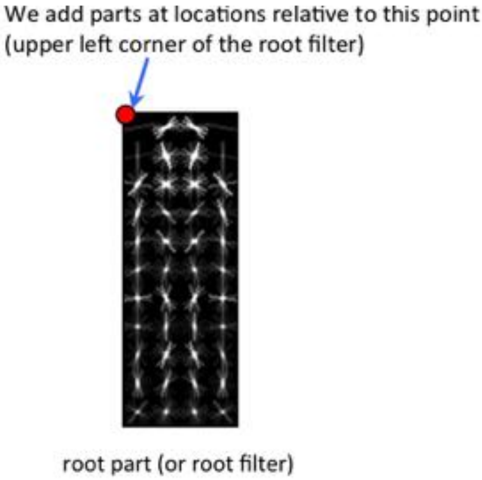


root part (or root filter)

The DPM Model

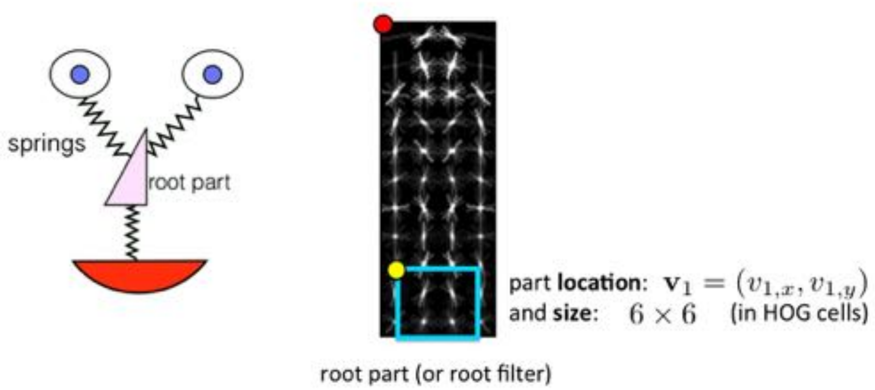
- DPM now wants to add parts. It wants to add them at locations **relative** to the location of the root filter. Relative makes sense: if we move, we take our parts with us.





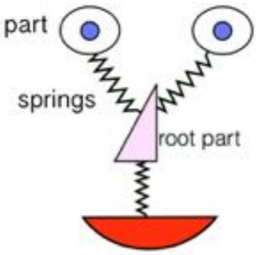
The DPM Model

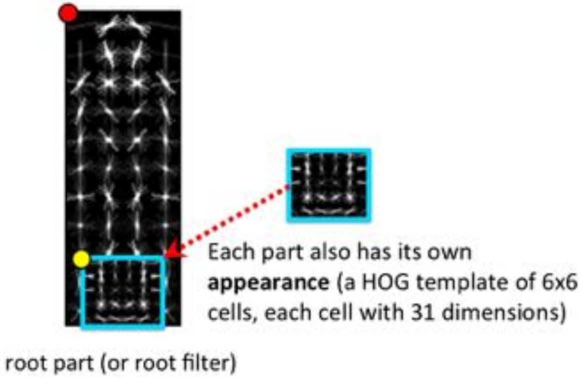
- Add a part at a relative location and scale.



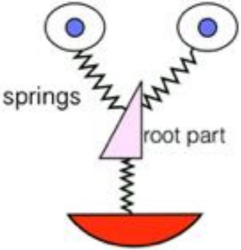
The DPM Model

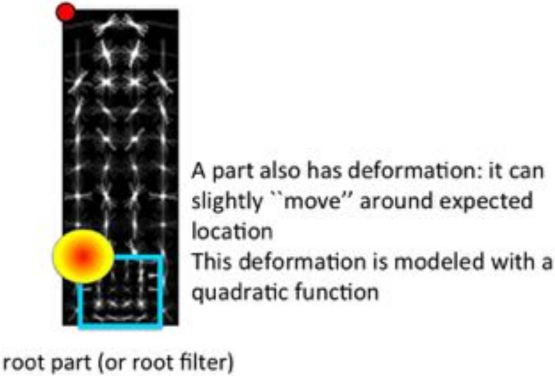
- Each part has an **appearance**, which is modeled with a HOG template
- Each part's template is at **twice the resolution** as the root filter





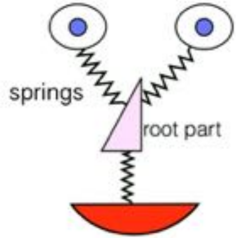
- Give some slack to the location of the part. Why is this a good idea?





The DPM Model

- People are of different heights, thus have feet at different locations relative to the head. And we want to detect all people, not just the average ones.

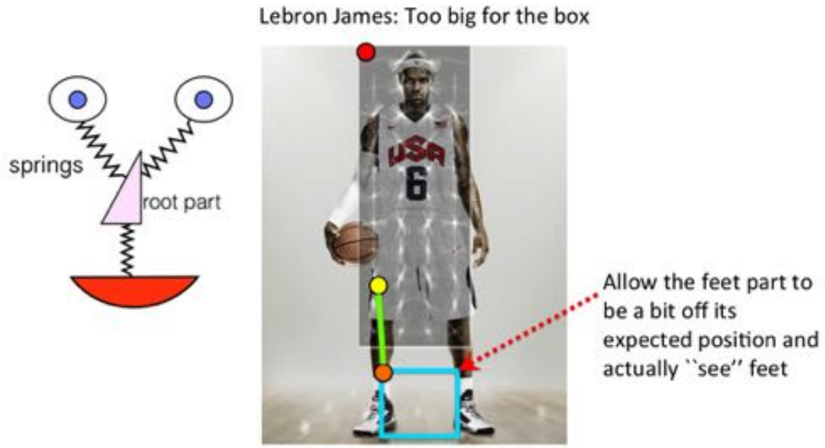


Lebron James: Too big for the box

If no deformation:
Feet part will ``see''
knees instead of feet!

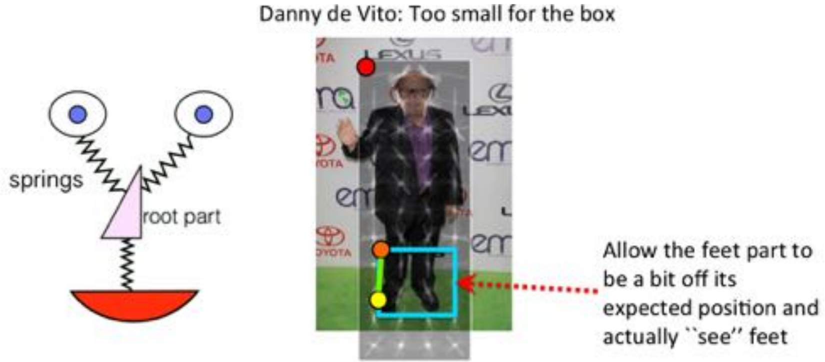
The DPM Model

- People are of different heights, thus have feet at different locations relative to the head. And we want to detect all people, not just the average ones.



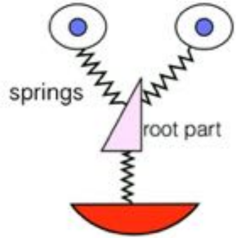
The DPM Model

- People are of different heights, thus have feet at different locations relative to the head. And we want to detect all people, not just the average ones.



The DPM Model

- People are of different heights, thus have feet at different locations relative to the head. And we want to detect all people, not just the average ones.



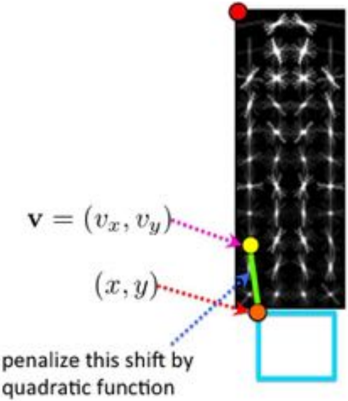
Brad Pitt: Fits perfectly

The DPM Model

- We will, however, trust less detections where parts are not exactly in their expected location. DPM penalizes part shifts with a quadratic function:

$$a(x - v_x)^2 + b(x - v_x) + c(y - v_y)^2 + d(y - v_y)$$

(here a, b, c, d are weights that are used to penalize different terms)

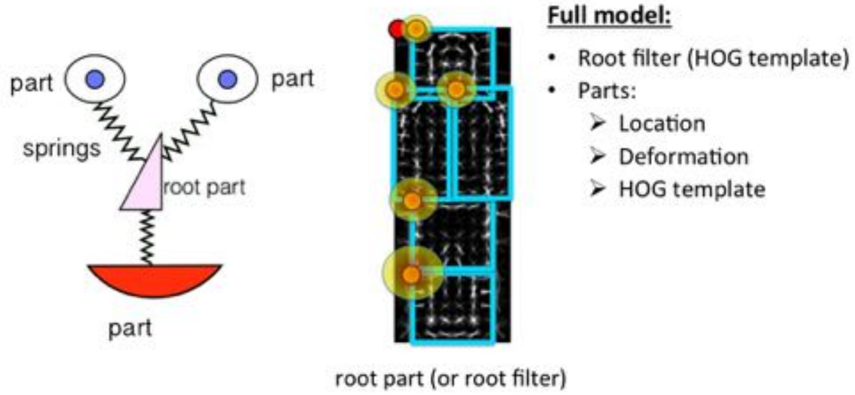


For example, a very tall person may have feet way lower. We want our model to detect also tall people.

But since there are less really tall people, we want to penalize such detections a little bit (we will trust it less – how many images do actually have NBA players, afterall?).

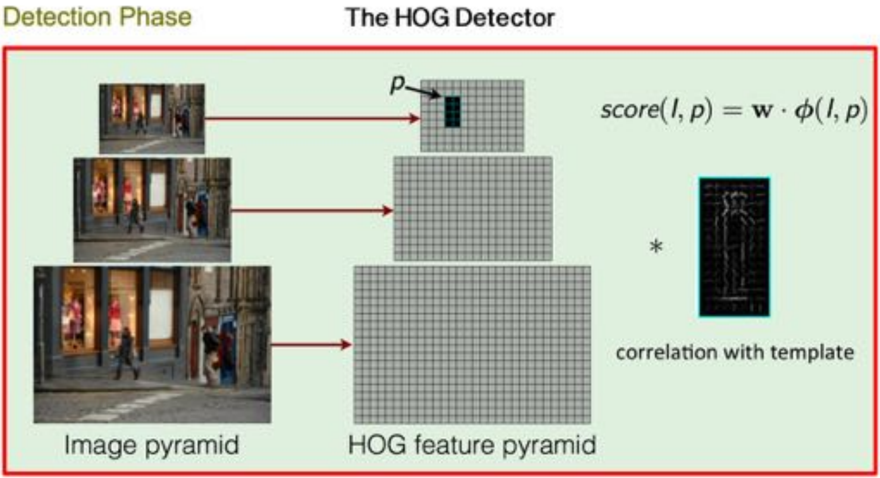
The DPM Model

- And finally, DPM has a few parts. Typically 6 (but it's a parameter you can play with). How many weights does a 6-part DPM model have?
- How shall we score this part-model guy in an image (how to do detection)?



Remember the HOG Detector

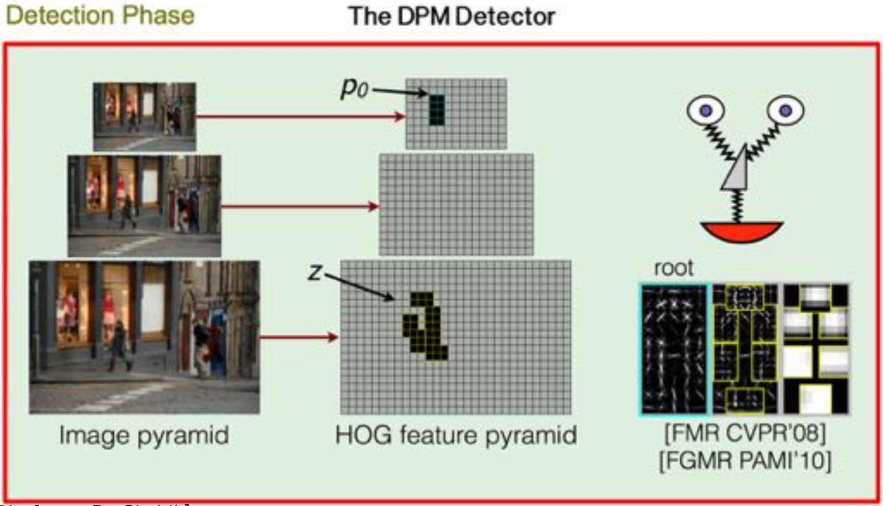
- The HOG detector computes image pyramid, HOG features, and scores each window with a learned linear classifier



[Pic from: R. Girshik]

DPM Detector

- For DPM the story is quite similar (pyramid, HOG, score window with a learned linear classifier), but now we also need to score the parts.



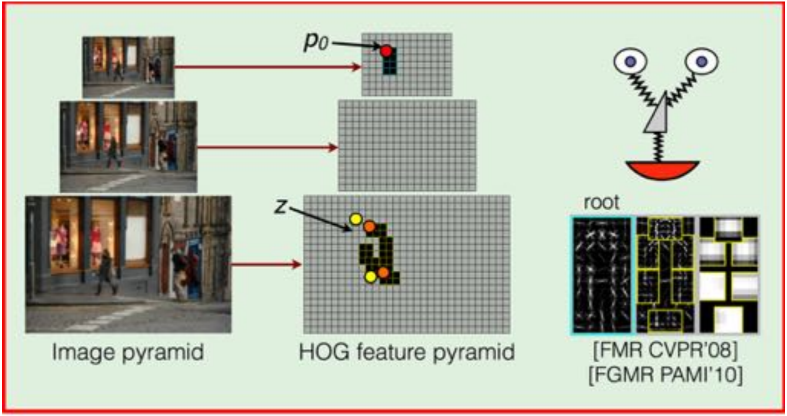
[Pic from: R. Girshik]

Scoring

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

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

Filter scores　　Spring costs



Scoring

- More specifically, we will score a location (window) in the image as follows:

$$\text{score}(l, p_0) = \max_{p_1,\ldots,p_n} \left(\sum_{i=0}^{n} F_i \cdot HOG(l, p_i) - \sum_{i=1}^{n} \mathbf{w_{def}}^i \cdot (dx, dy, dx^2, dy^2) \right)$$

where

  - F_0 is the (learned) HOG template for root filter
  - F_i is the (learned) HOG template for part i
  - $HOG(l, p_i)$ means a HOG feature cropped in window defined by part location p_i at level l of the HOG pyramid
  - $\mathbf{w_{def}}^i$ are (learned) weights for the deformation penalty
  - (dx, dy, dx^2, dy^2) with $(dx, dy) = (x_i, y_i) - ((x_0, y_0) + \mathbf{v_i})$ tell us how far the part i is from its expected position $(x_0, y_0) + \mathbf{v_i}$

- **Main question:** How shall we compute that nasty $\max_{p_1,\ldots,p_n}$?

Scoring

- More specifically, we will score a location (window) in the image as follows:

$$\text{score}(l, p_0) = \max_{p_1,\ldots,p_n} \Big(\sum_{i=0}^{n} F_i \cdot HOG(l, p_i) - \sum_{i=1}^{n} \mathbf{w_{def}}^i \cdot (dx, dy, dx^2, dy^2) \Big)$$

  where

  - F_0 is the (learned) HOG template for root filter
  - F_i is the (learned) HOG template for part i
  - $HOG(l, p_i)$ means a HOG feature cropped in window defined by part location p_i at level l of the HOG pyramid
  - $\mathbf{w_{def}}^i$ are (learned) weights for the deformation penalty
  - (dx, dy, dx^2, dy^2) with $(dx, dy) = (x_i, y_i) - ((x_0, y_0) + \mathbf{v_i})$ tell us how far the part i is from its expected position $(x_0, y_0) + \mathbf{v_i})$

- **Main question:** How shall we compute that nasty $\max_{p_1,\ldots,p_n}$?

Scoring

- Push the max inside (why can we do that?):

$$\text{score}(I, p_0) = F_0 \cdot HOG(I, p_0) + \sum_{i=1}^{n} \max_{p_i} \Big(F_i \cdot HOG(I, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \Big)$$

Scoring

- Push the max inside:

$$\mathrm{score}(I, p_0) = F_0 \cdot HOG(I, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(I, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$

- We can compute this with **dynamic programming**. Any idea how?

Computing the Score with Dynamic Programming

$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \Big(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \Big)$$

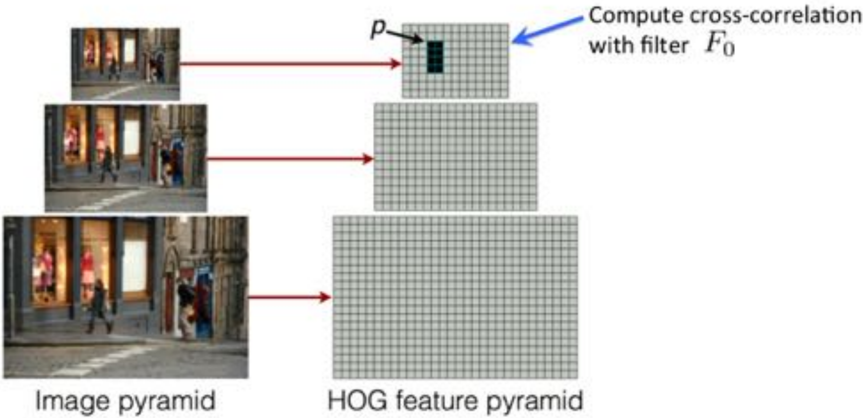
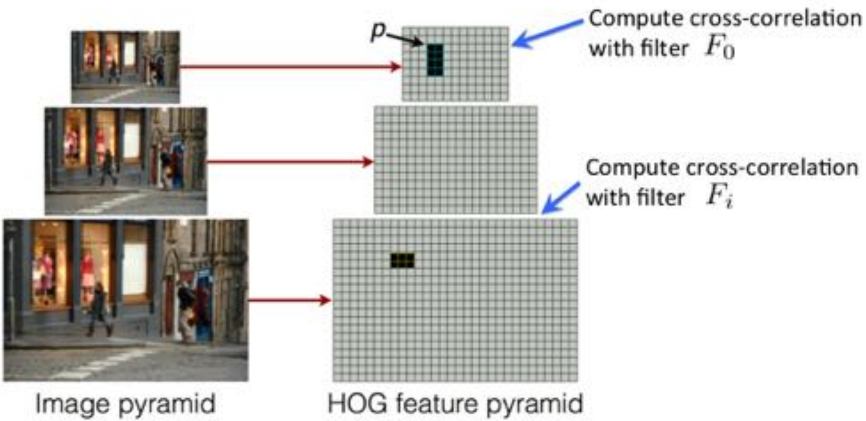


Figure: We can compute $F_i \cdot HOG(l, p_i)$ for the full level l via cross-correlation of the HOG feature matrix at level l with the template (filter) F_i

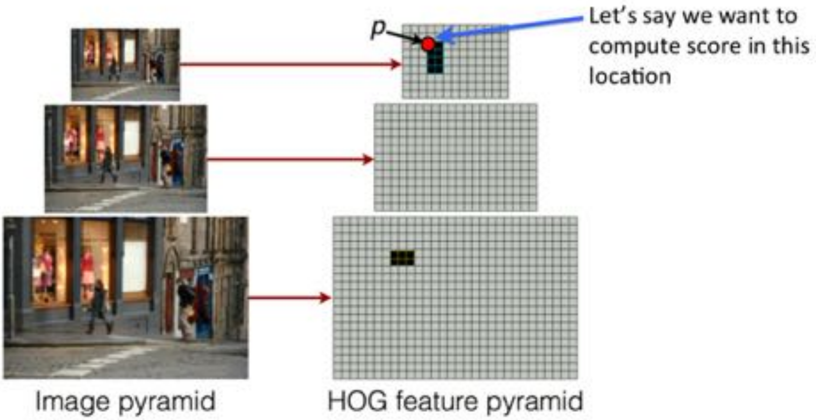
Computing the Score with Dynamic Programming

$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$



Computing the Score with Dynamic Programming

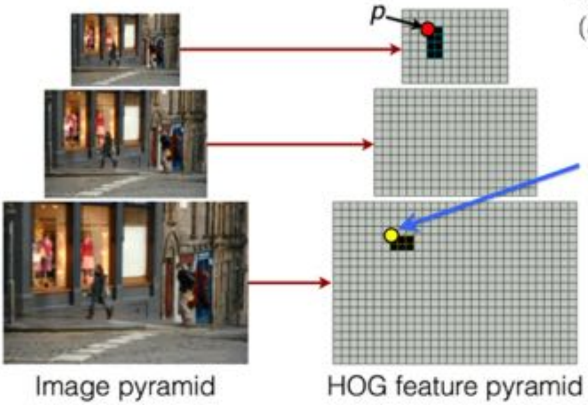
$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$



Computing the Score with Dynamic Programming

$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$

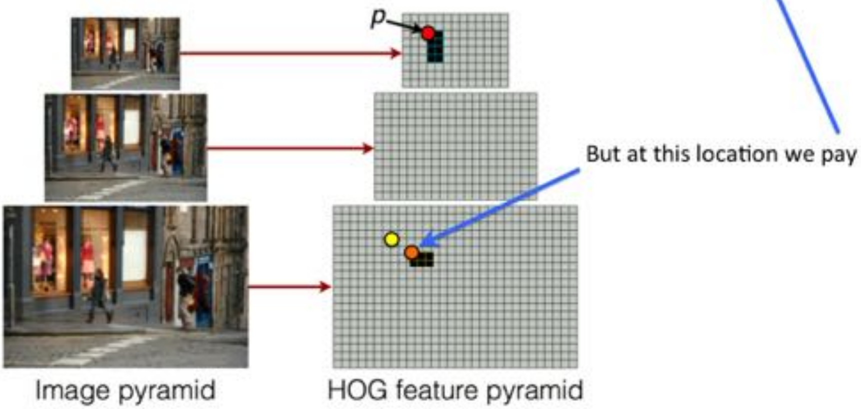
This is 0 in yellow point, because $(dx, dy, dx^2, dy^2) = (0, 0, 0, 0)$



There is no penalty for placing the part in the yellow location (the part is at expected location relative to the location of the root filter)

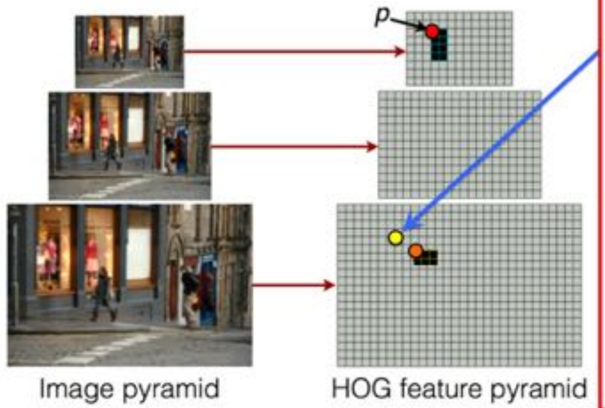
Computing the Score with Dynamic Programming

$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$



Computing the Score with Dynamic Programming

We are computing this: $\max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$



We need to loop over all possible placements of the part. For each placement we need to:

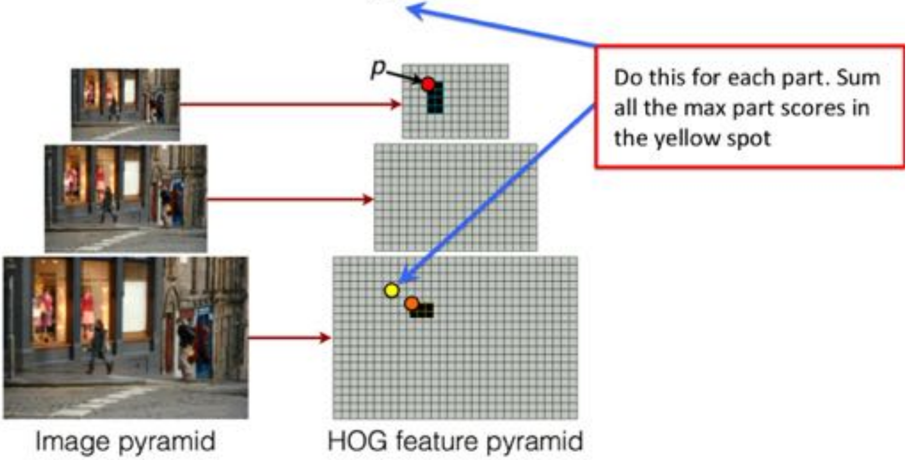
- Compute deformation cost
- Read out the correlation value
- Subtract deformation from corr value

Find the max of these scores across all placements. Store the max in the yellow spot.

Figure: We can compute these scores efficiently with something called **distance transforms** (this is exact). But works equally well: Simply limit the scope of where each part could be to a small area, e.g., a few HOG cells up,down,left,right relative to yellow spot (this is approx).

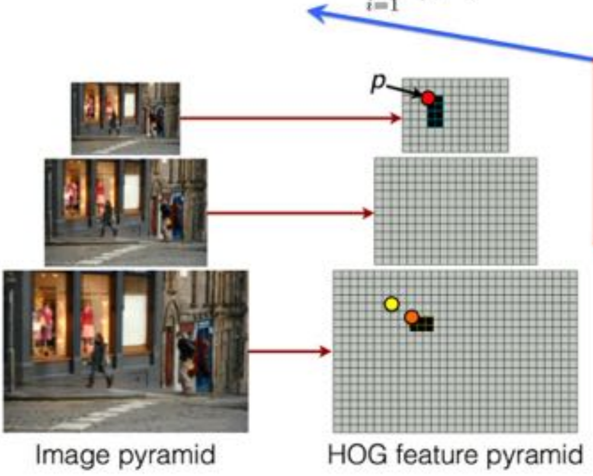
Computing the Score with Dynamic Programming

$$\sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$



Computing the Score with Dynamic Programming

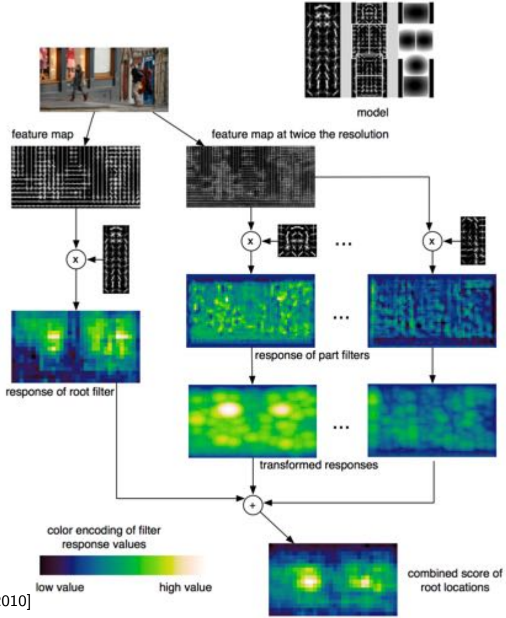
$$\text{score}(l, p_0) = F_0 \cdot HOG(l, p_0) + \sum_{i=1}^{n} \max_{p_i} \left(F_i \cdot HOG(l, p_i) - \mathbf{w_{def}}^i \cdot \phi_{def}(x_i, y_i) \right)$$



Add the value in the yellow location to the value in the red location.

Done!

Detection



[Pic from: Felzenswalb et al., 2010]

Training

- You can't train this model as simple as the HOG detector, via SVM. For those taking CSC411: Why not?

Training

- You can't train this model as simple as the HOG detector, via SVM. For those taking CSC411: Why not?
- Because the part positions are not annotated (we don't have ground-truth, and SVM needs ground-truth). We say that the parts are **latent**.
- You can train the model with something called **latent SVM**. For ML buffs:
  - Check the Felzenswalb paper
  - For those with even stronger ML stomach: Yu, Joachims, Learning Structural SVMs with Latent Variables, ICML'09.

Results

Figure: Performance of the HOG detector on person class on PASCAL VOC

[Pic from: R. Girshik]

Results

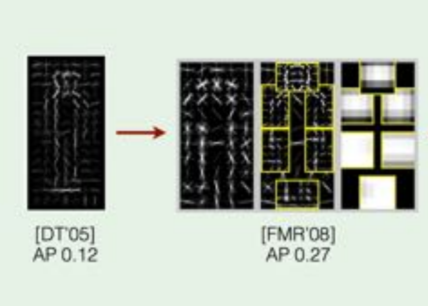


Figure: DPM version 1: adds the parts

[Pic from: R. Girshik]

Results

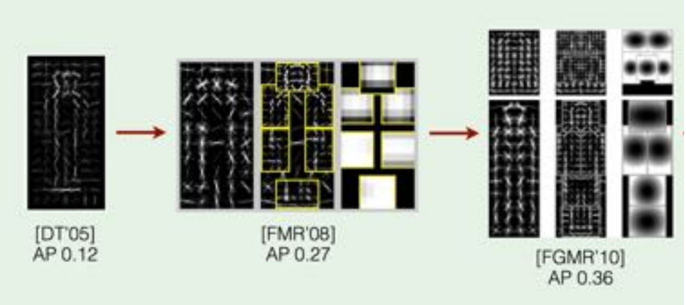


Figure: DPM version 2: adds another template (called mixture or component). Supposed to detect also people sitting down (e.g., occluded by desk).

[Pic from: R. Girshik]

Results

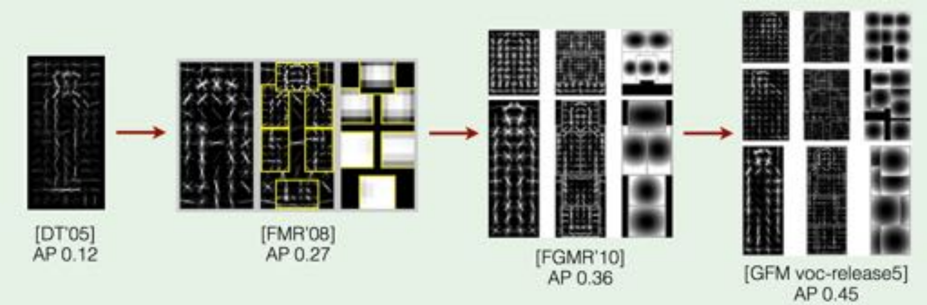
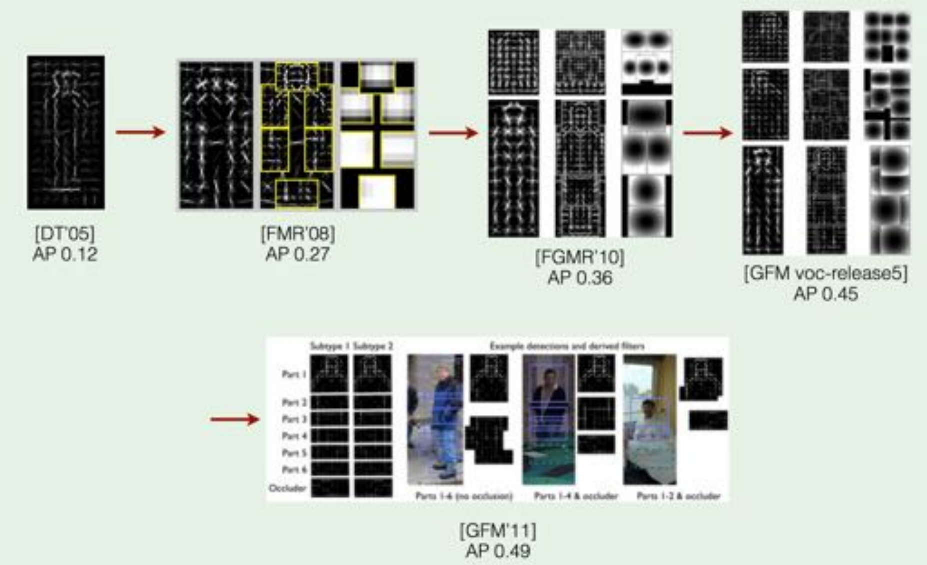


Figure: DPM version 3: adds multiple mixtures (components)

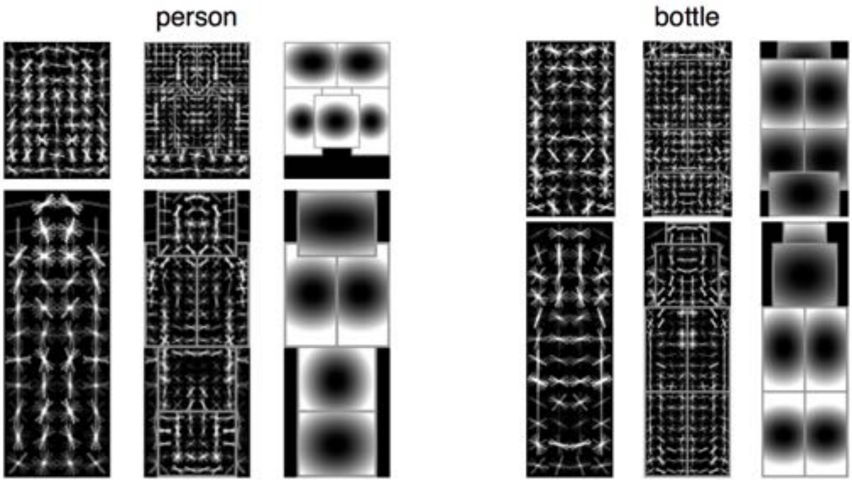
[Pic from: R. Girshik]

Results



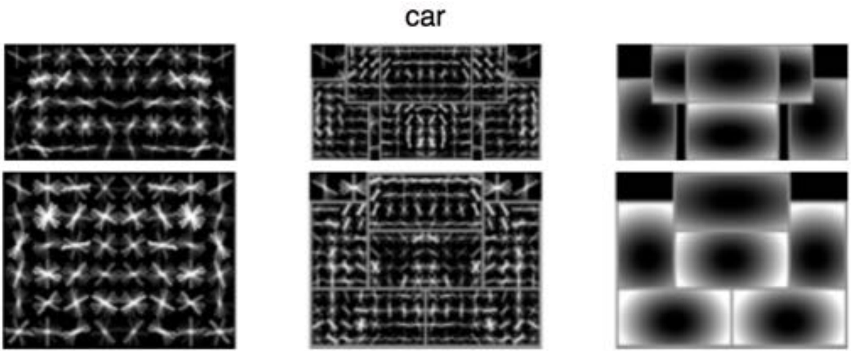
[Pic from: R. Girshik]

Learned Models



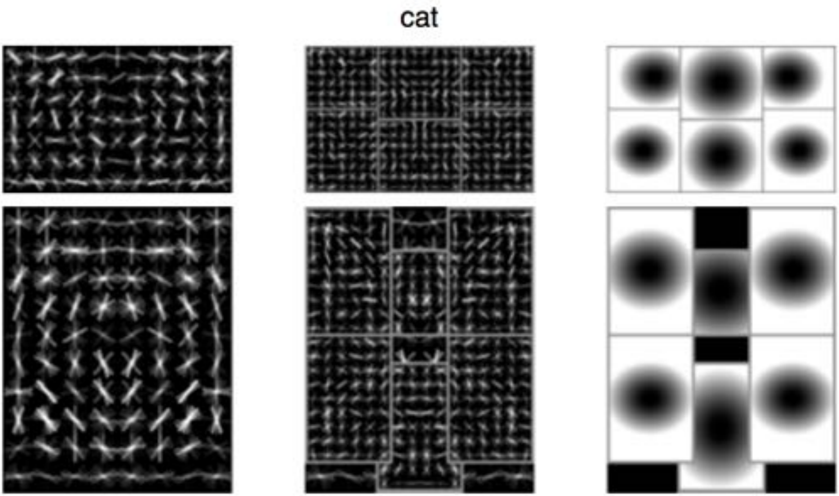
[Pic from: Felzenswalb et al., 2010]

Learned Models



[Pic from: Felzenswalb et al., 2010]

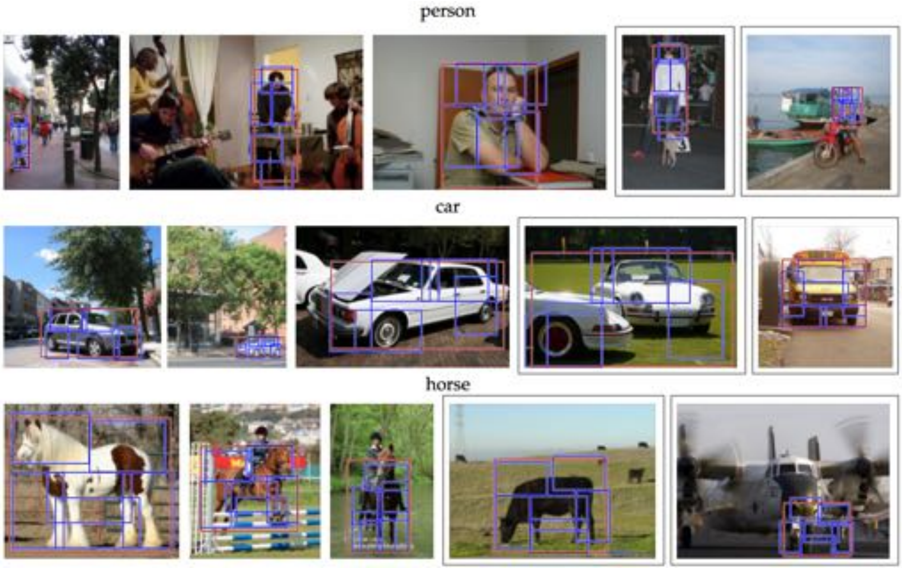
Learned Models



(Takes some imagination to see a cat...)

[Pic from: Felzenswalb et al., 2010]

Results



[Pic from: Felzenswalb et al., 2010]

Results

[Pic from: Felzenswalb et al., 2010]

DPM

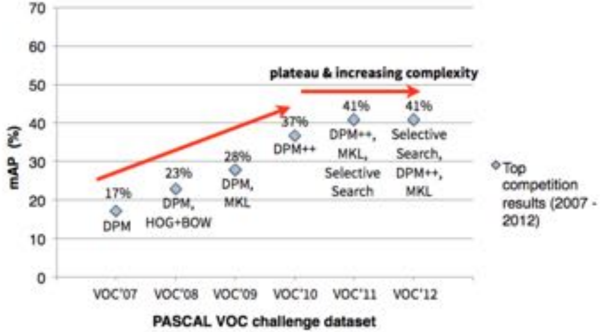
- As you already know, the code is available:

  `http://www.cs.berkeley.edu/~rbg/latent/`

- Trivia:
  - Takes about 20-30 seconds per image per class. Speed-ups exist.
  - Depending on the size of the dataset, training takes around 12 hours (for most PASCAL classes).
  - Has some cool post-processing tricks: bounding box prediction and context re-scoring. Each typically results in around 2% improvement in AP.
  - In the code, if you switch off the parts, you get the Dalal & Triggs' HOG detector.

Results

DPM on PASCAL VOC



[Source: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc20(07,08,09,10,11,12)/results/index.html]

Ross Girshick

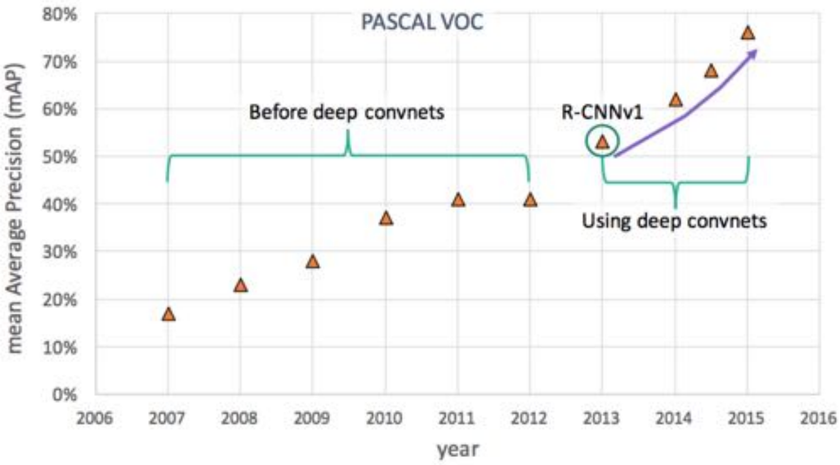
Lifetime Achievement Award by PASCAL VOC

Object Class Detection

- Pre 2014
  - HOG detector
  - Deformable Part-based Model
- Post 2014 (neural networks)
  - R-CNN
  - Fast(er) R-CNN
  - Yolo, SSD

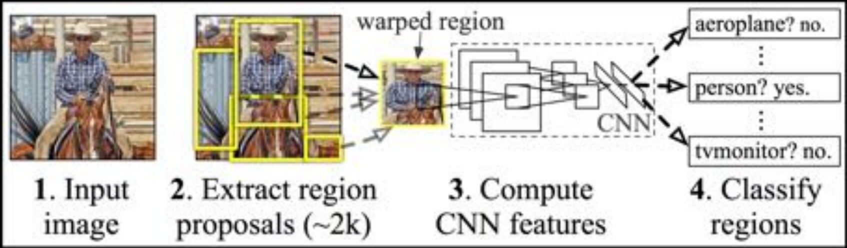
[Credit for the slides to follow: Bin Yang]

The CNN Era



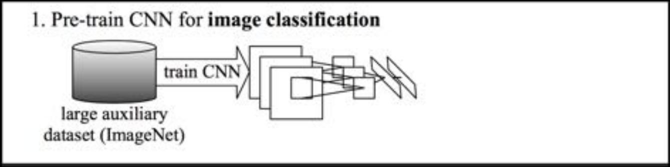
[Slide credit: Renjie Liao]

RCNN: Regions with CNN Features

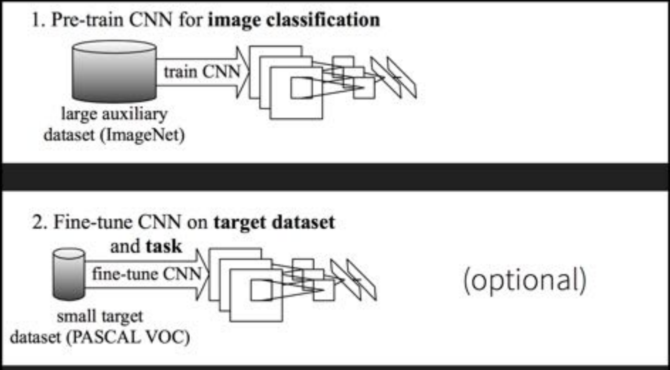


[Slide credit: Ross Girshick]

Training

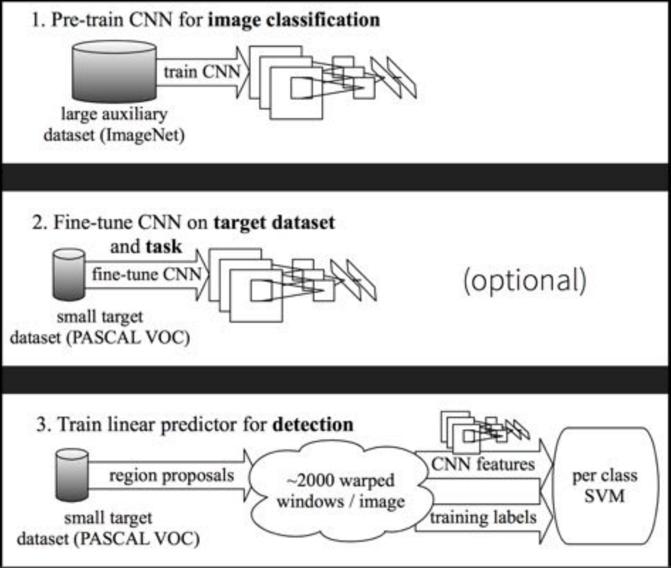

1. Pre-train CNN for **image classification**
train CNN
large auxiliary
dataset (ImageNet)

Training



(optional)

Training


1. Pre-train CNN for **image classification**
train CNN
large auxiliary dataset (ImageNet)
2. Fine-tune CNN on **target dataset** and **task**
fine-tune CNN
small target dataset (PASCAL VOC)
(optional)
3. Train linear predictor for **detection**
region proposals
small target dataset (PASCAL VOC)
~2000 warped windows / image
CNN features
training labels
per class SVM

RCNN: Performance

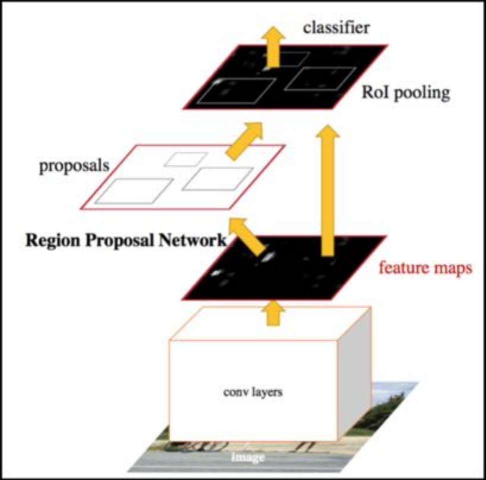
	VOC2007
DPM v5 (Girshick et al. 2011)	33.7%
Regionlets (Wang et al. 2013)	41.7%
R-CNN (AlexNet)	54.2%
R-CNN (AlexNet) + BB	58.5%
R-CNN (VGGNet)	62.2%
R-CNN (VGGNet) + BB	66.0%

RCNN: Performance

	VOC2007
DPM v5 (Girshick et al. 2011)	33.7%
Regionlets (Wang et al. 2013)	41.7%
R-CNN (AlexNet)	54.2%
R-CNN (AlexNet) + BB	58.5%
R-CNN (VGGNet)	62.2%
R-CNN (VGGNet) + BB	66.0%

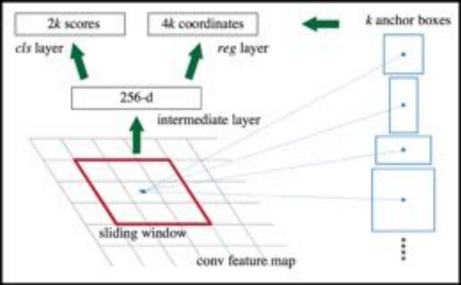
R-CNN (VGGNet)	Time
Train	84 hours
Test	47 s/im

Faster R-CNN



Ren S, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS2015

Region Proposal Network (RPN)



- Sliding window style
- Multi-scale predictions on fix-sized window for efficiency (take advantage of the large receptive field of CNN features)
- Same loss as R-CNN (cls+bbox)

anchor	128^2, 2:1	128^2, 1:1	128^2, 1:2	256^2, 2:1	256^2, 1:1	256^2, 1:2	512^2, 2:1	512^2, 1:1	512^2, 1:2
proposal	188×111	113×114	70×92	416×229	261×284	174×332	768×437	499×501	355×715

Region Proposal Network (RPN)

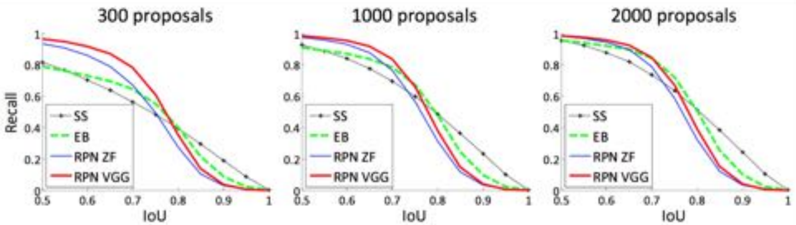


Figure 2: Recall *vs.* IoU overlap ratio on the PASCAL VOC 2007 test set.

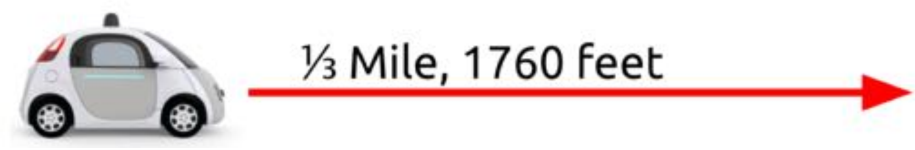
Faster R-CNN: Performance

- Fewer and better proposals not only bring speed-up, but also detection performance boost.

method	# proposals	data	mAP (%)	time (ms)
SS	2k	07	66.9	1830
SS	2k	07+12	70.0	1830
RPN+VGG, unshared	300	07	68.5	342
RPN+VGG, shared	300	07	69.9	**196**
RPN+VGG, shared	300	07+12	**73.2**	**196**

Car Example

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img



[Slide credit: Joseph Chet Redmon]

Car Example

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img

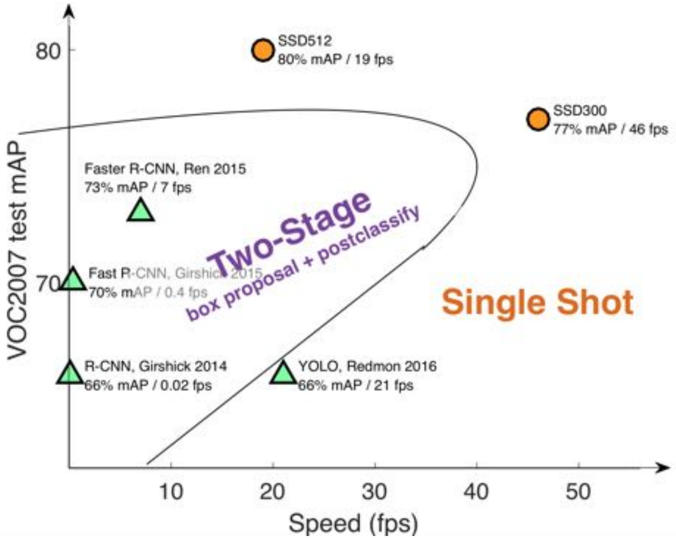
[Slide credit: Joseph Chet Redmon]

Car Example

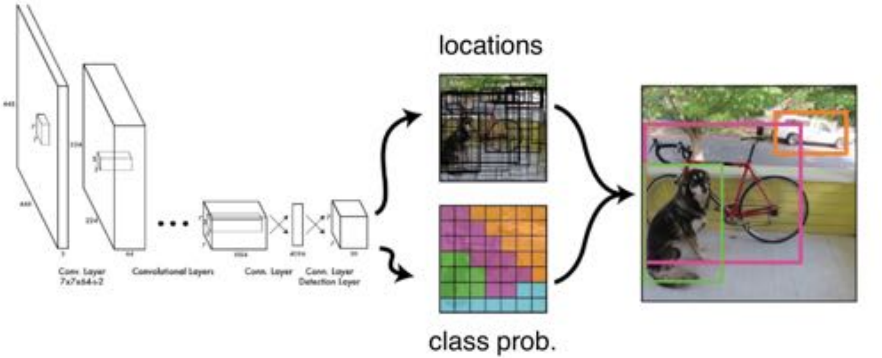
	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img

[Slide credit: Joseph Chet Redmon]

Real Time Object Detection?



YOLO: You Only Look Once

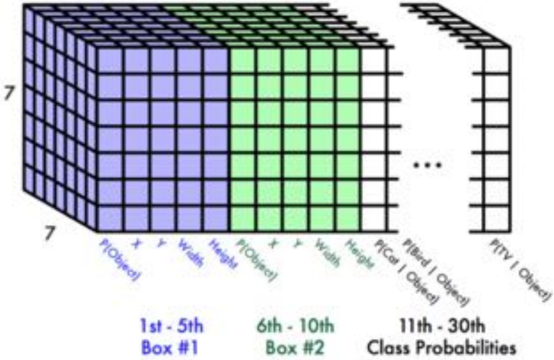


[Slide credit: Redmon J et al. You only look once: Unified, real-time object detection. CVPR'16]

YOLO: Output Parametrization

Each cell predicts:

- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities



For Pascal VOC:

- 7x7 grid
- 2 bounding boxes / cell
- 20 classes

7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

[Slide credit: Redmon J et al. You only look once: Unified, real-time object detection. CVPR'16]

SSD: Single Shot MultiBox Detector

[Slide credit: Wei L, et al. SSD: Single Shot MultiBox Detector. ECCV'16]

That's It For CSC420... But There Is Much More of Computer Vision For Those Interested!