04/11/17

Object Category Detection

Computer Vision

CS 543 / ECE 549

University of Illinois

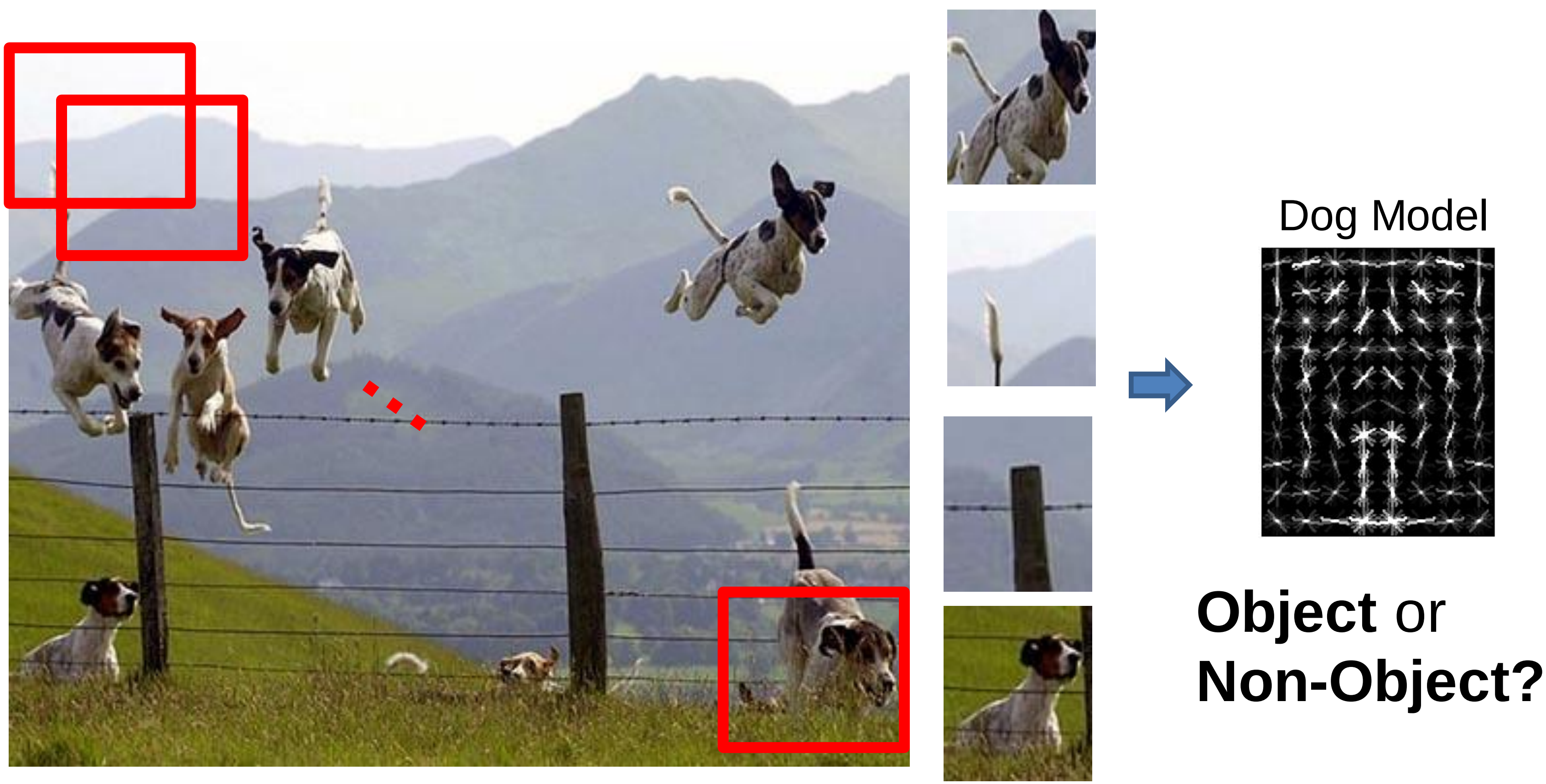
Derek Hoiem

Today’s class: Object Category Detection

- Overview of object category detection
- Detection methods
  - Dalal-Triggs pedestrian detector (basic concept)
  - Viola-Jones detector (cascades, integral images)
  - R-CNN line of detectors (CNN)
  - YOLO (refinement/simplification of R-CNN)

Object Category Detection

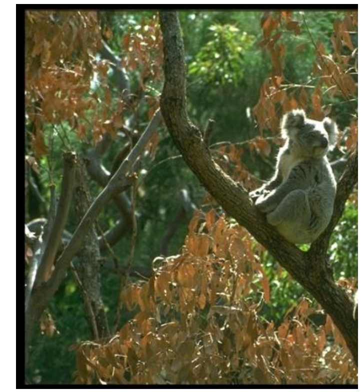
- Focus on object search: “Where is it?”
- Build templates that quickly differentiate object patch from background patch



Challenges in modeling the object class

Illumination

Object pose



Clutter

Occlusions

Intra-class appearance

Viewpoint

Slide from K. Grauman, B. Leibe

Challenges in modeling the non-object class

True Detections

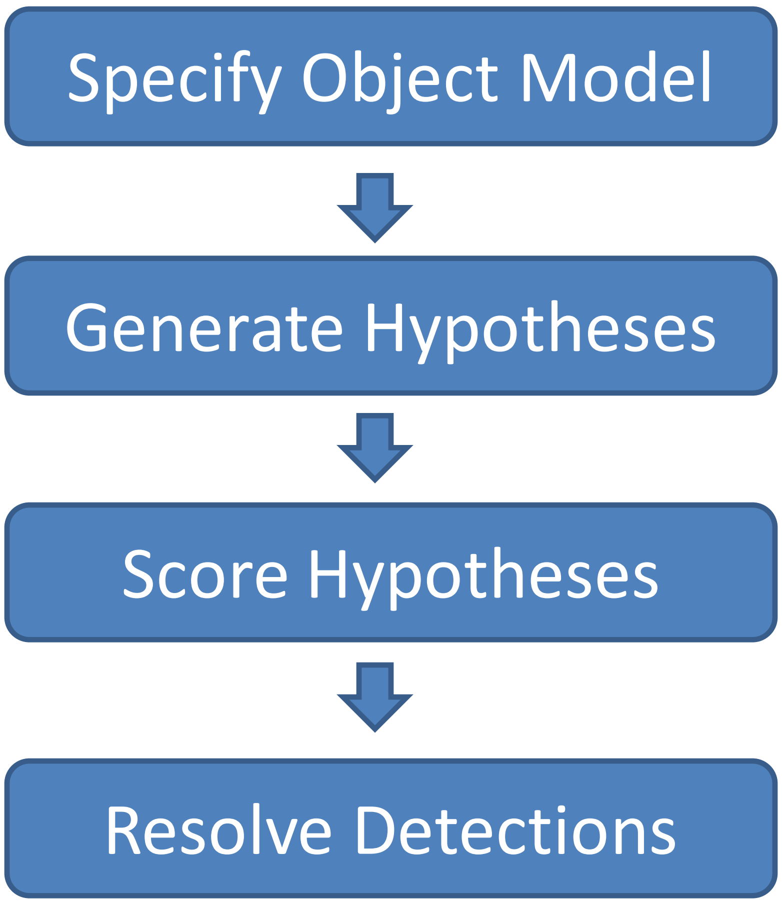
Bad Localization

Confused with Similar Object

Misc. Background

Confused with Dissimilar Objects

General Process of Object Recognition

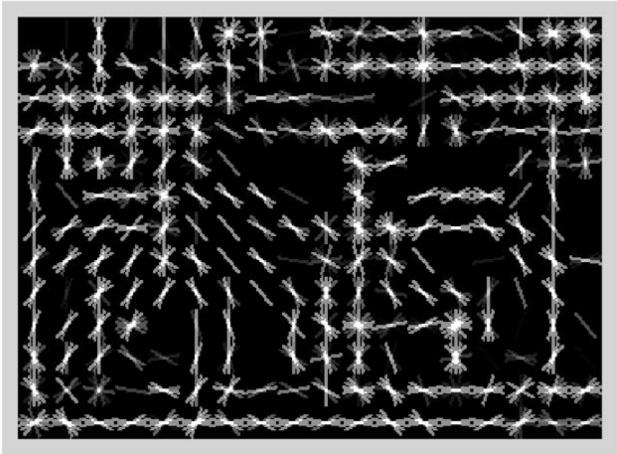


What are the object parameters?

Specifying an object model

1. Statistical Template in Bounding Box
   - Object is some (x,y,w,h) in image
   - Features defined wrt bounding box coordinates

Image

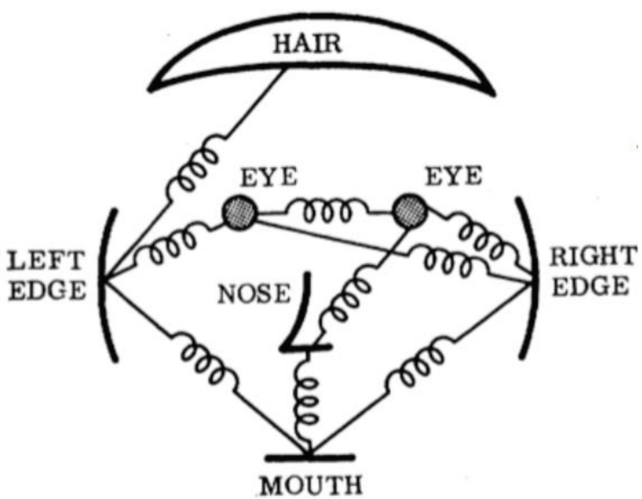


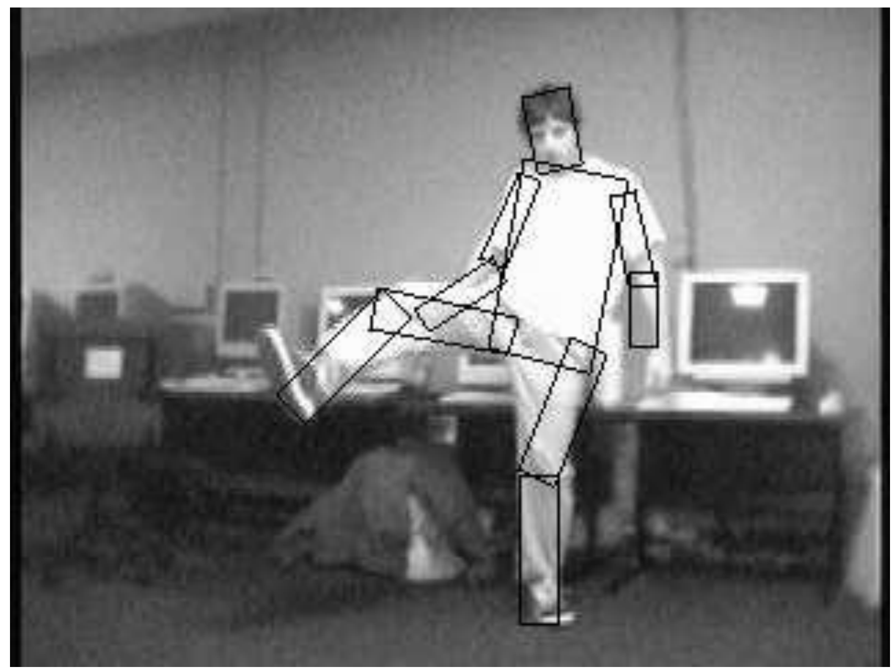
Template Visualization

Images from Felzenszwalb

Specifying an object model

2. Articulated parts model
   - Object is configuration of parts
   - Each part is detectable



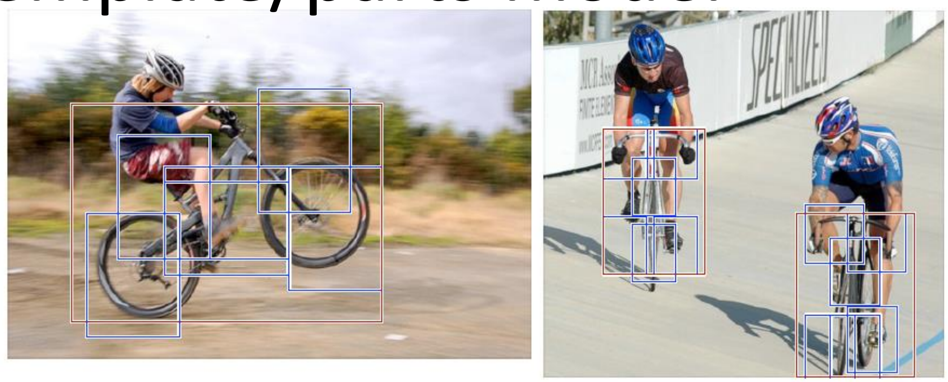


Images from Felzenszwalb

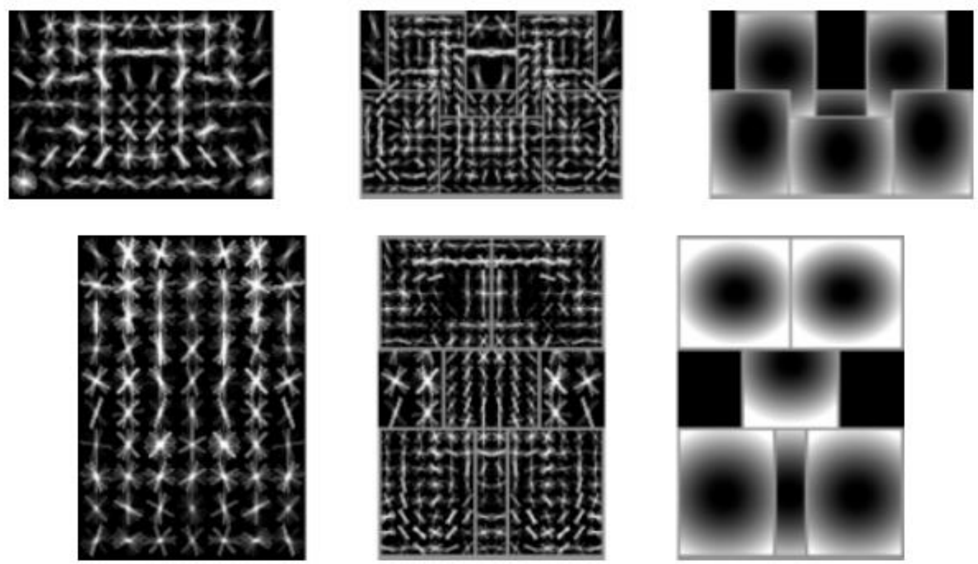
Specifying an object model

3. Hybrid template/parts model

Detections



Template Visualization



root filters
coarse resolution

part filters
finer resolution

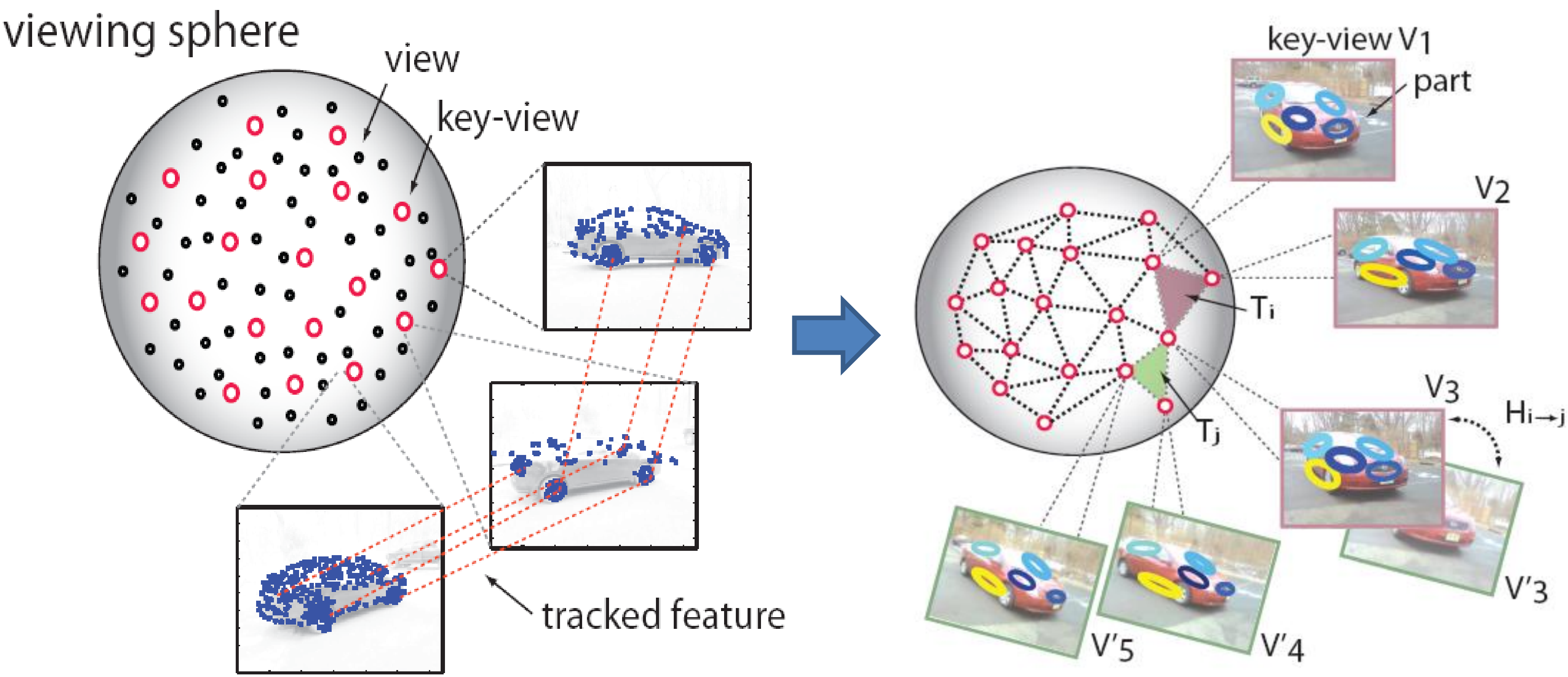
deformation
models

Felzenszwalb et al. 2008

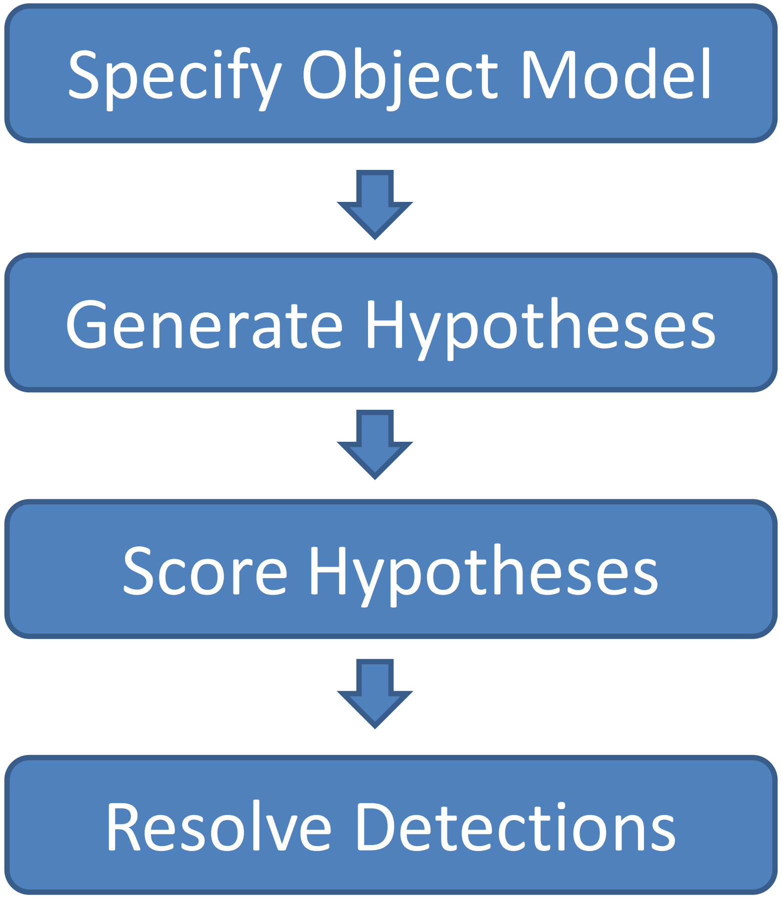
Specifying an object model

4. 3D-ish model

- Object is collection of 3D planar patches under affine transformation



General Process of Object Recognition



Propose an alignment of the model to the image

Generating hypotheses

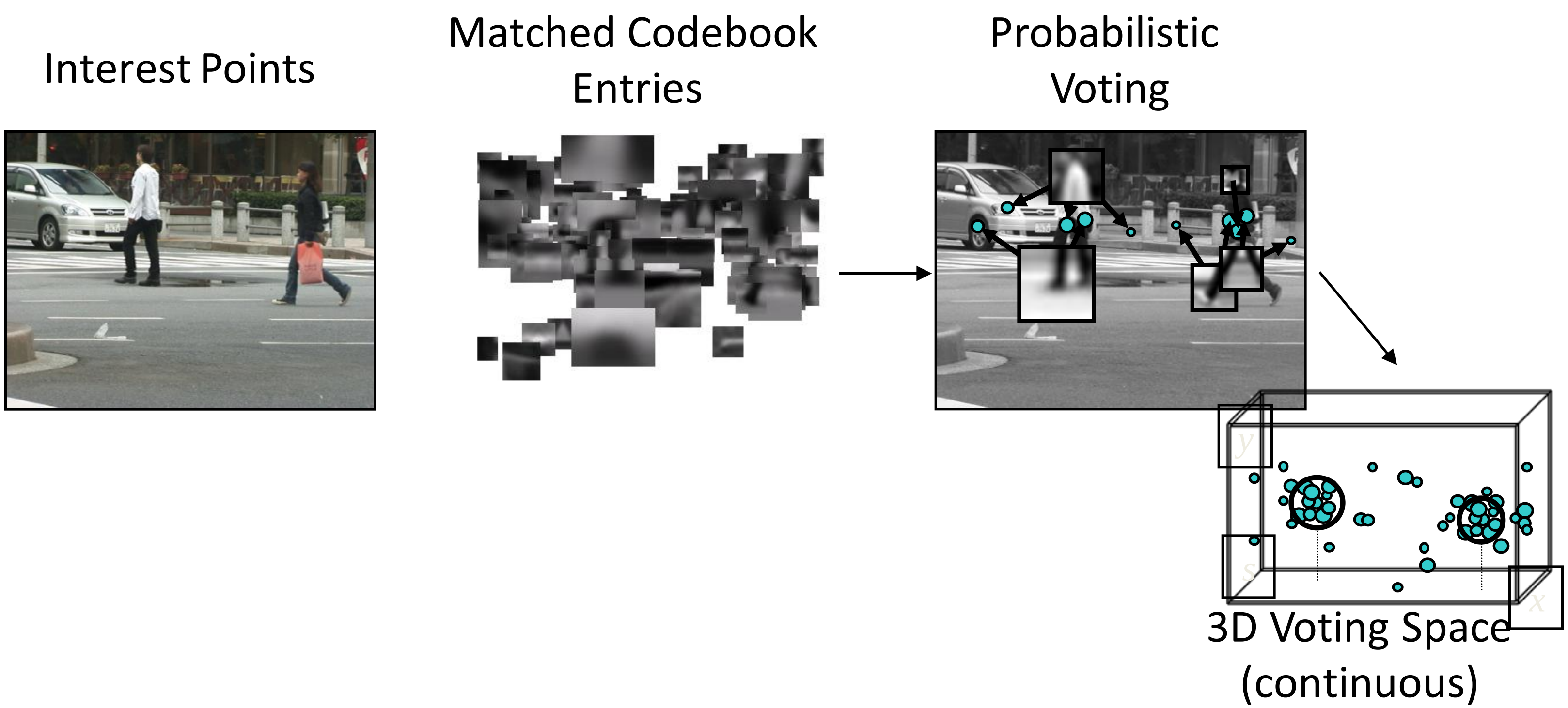
1. Sliding window
   – Test patch at each location and scale

Generating hypotheses

1. Sliding window
   – Test patch at each location and scale

Generating hypotheses

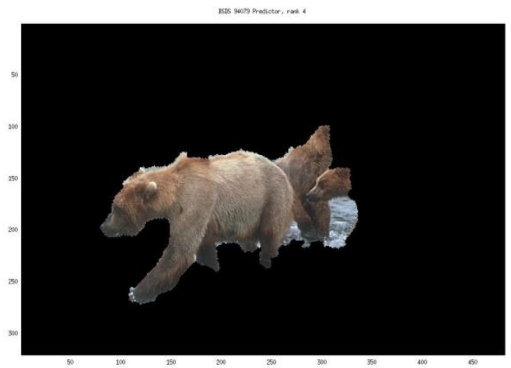
2. Voting from patches/keypoints

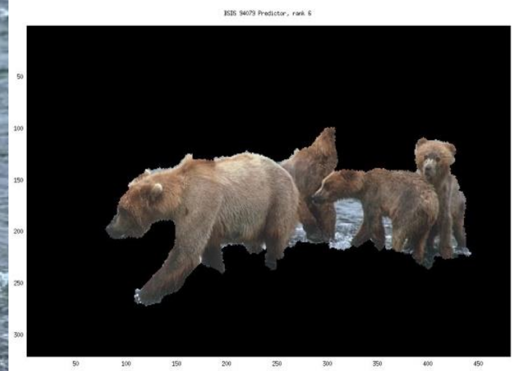


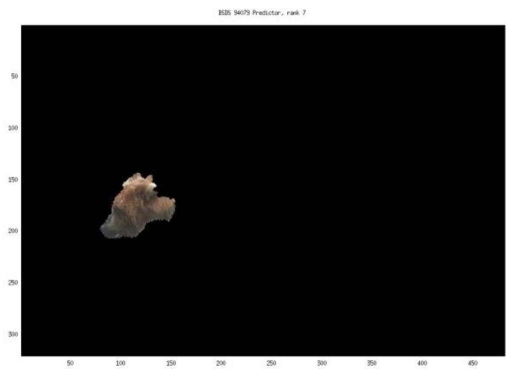
ISM model by Leibe et al.

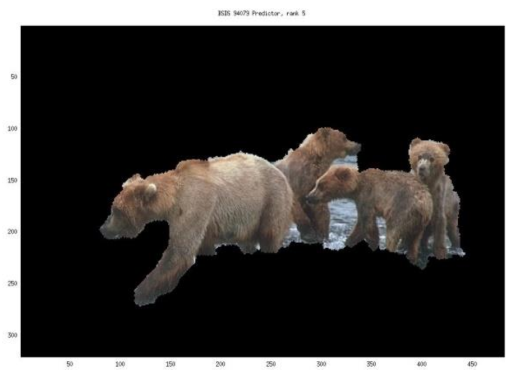
Generating hypotheses

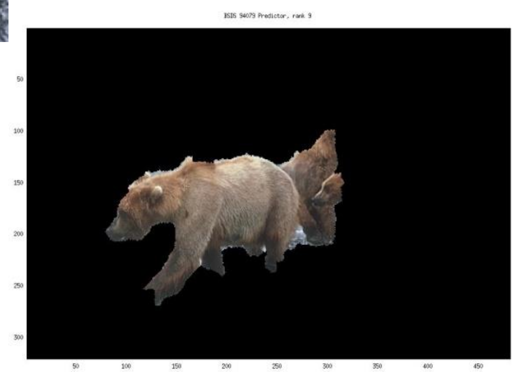
3. Region-based proposal





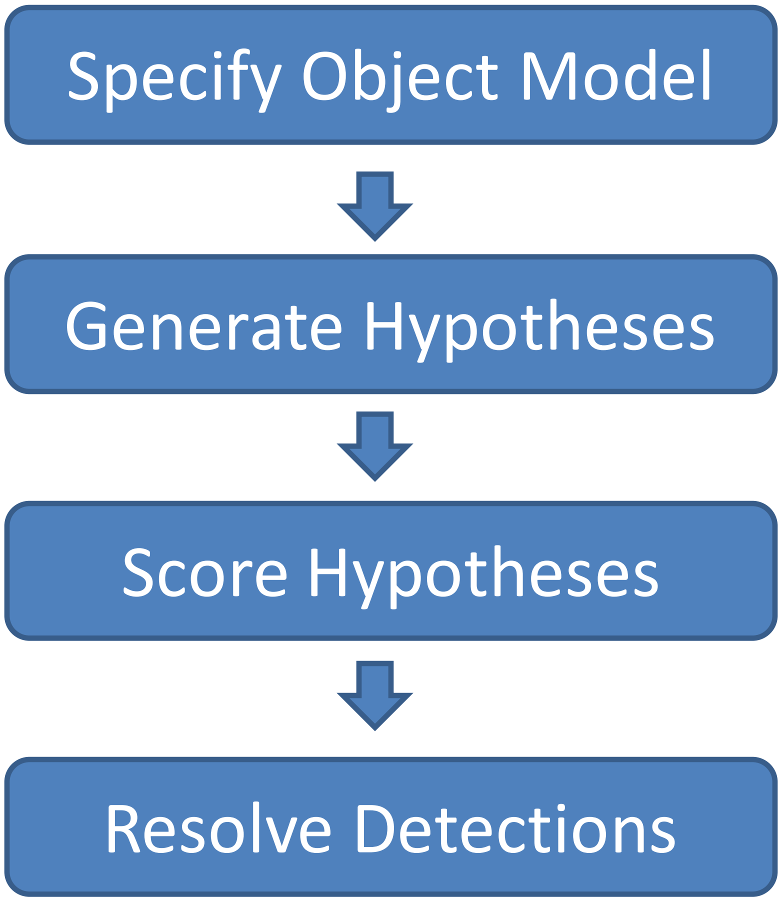






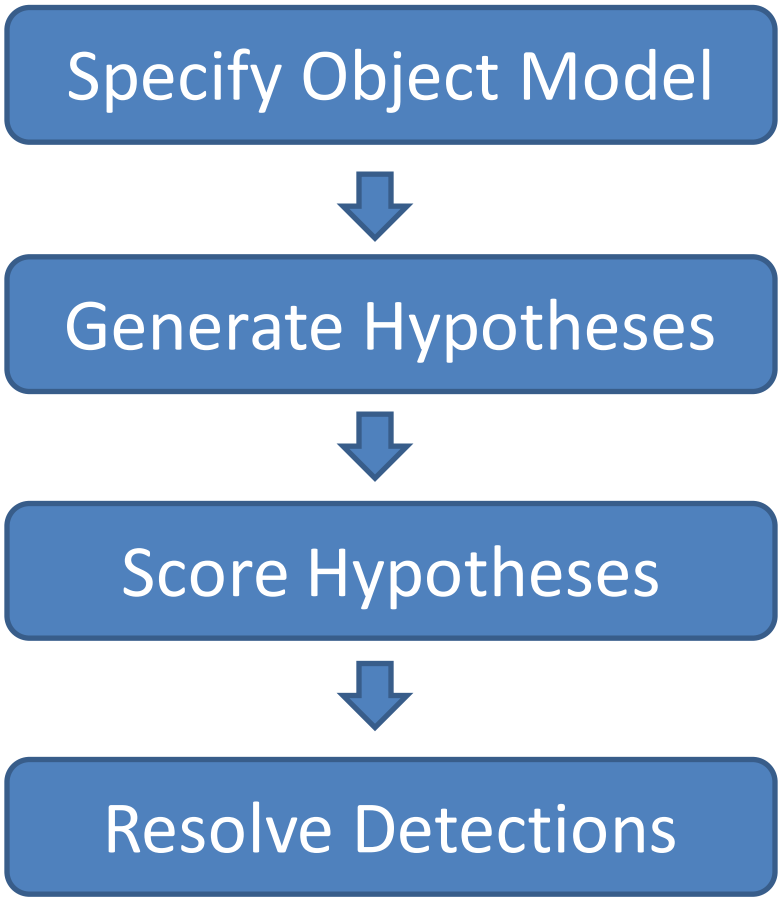
Endres Hoiem 2010

General Process of Object Recognition



Currently CNN features and classifiers

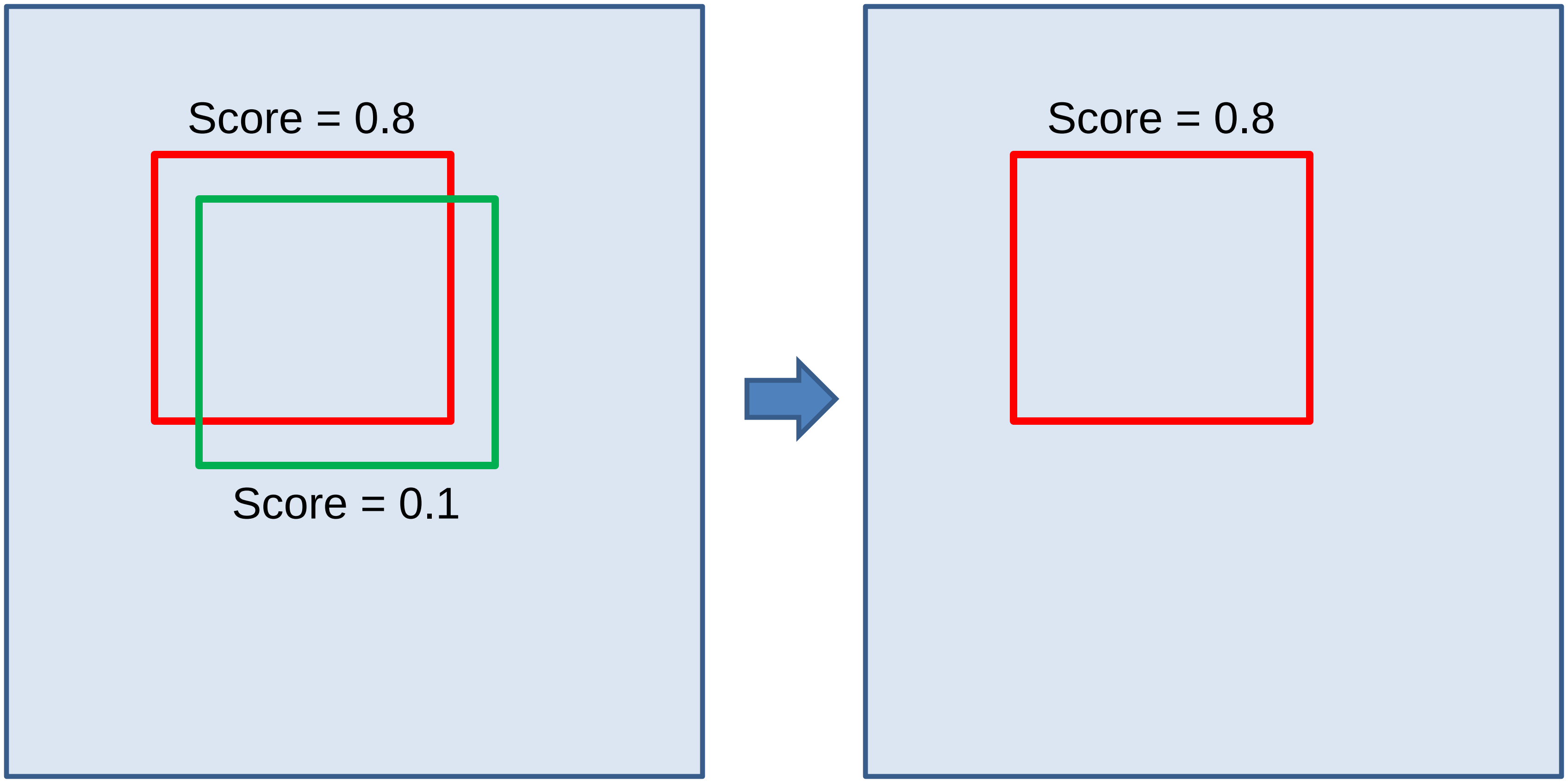
General Process of Object Recognition



Optionally, rescore each proposed object based on whole set

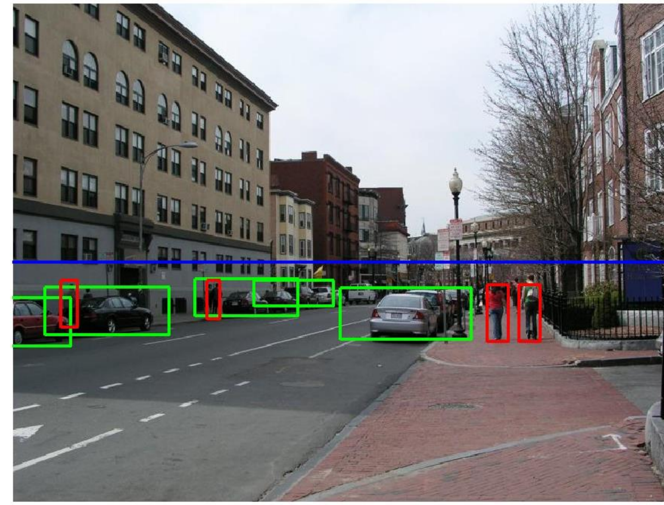
Resolving detection scores

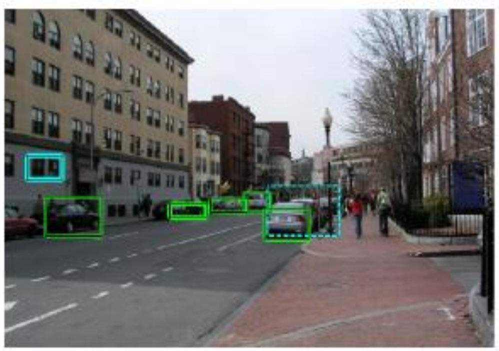
1. Non-max suppression



Resolving detection scores

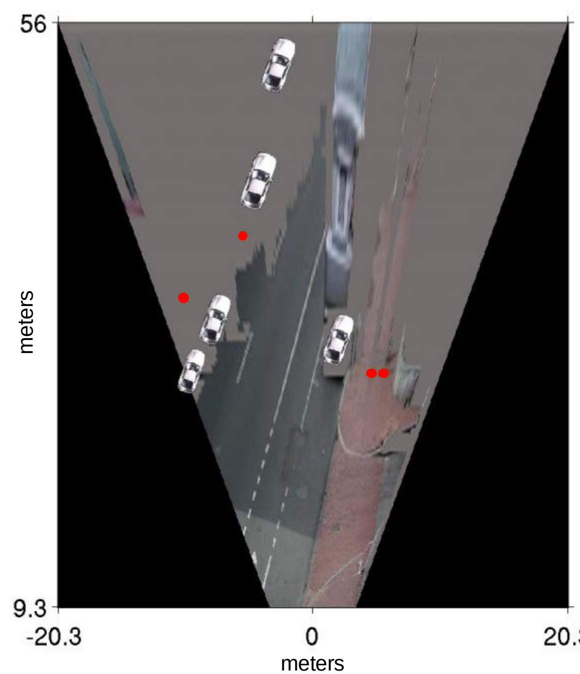
2. Context/reasoning





(g) Car Detections: Local

(h) Ped Detections: Local



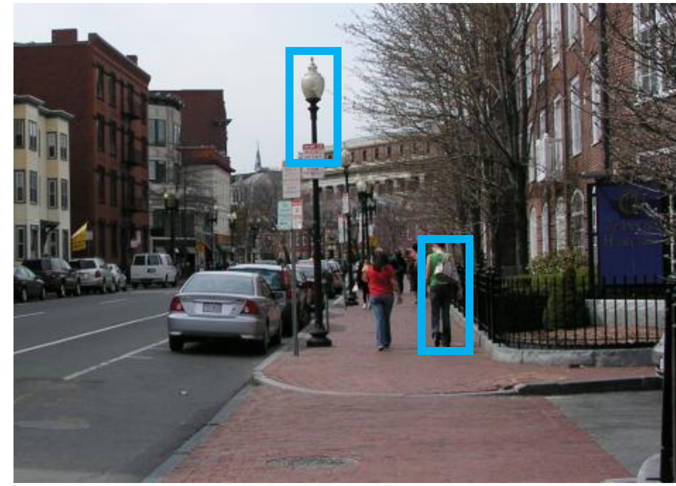
Hoiem et al. 2006

Object category detection in computer vision

Goal: detect all pedestrians, cars, monkeys, etc in image

Basic Steps of Category Detection

1. Align
   - E.g., choose position, scale orientation
   - How to make this tractable?



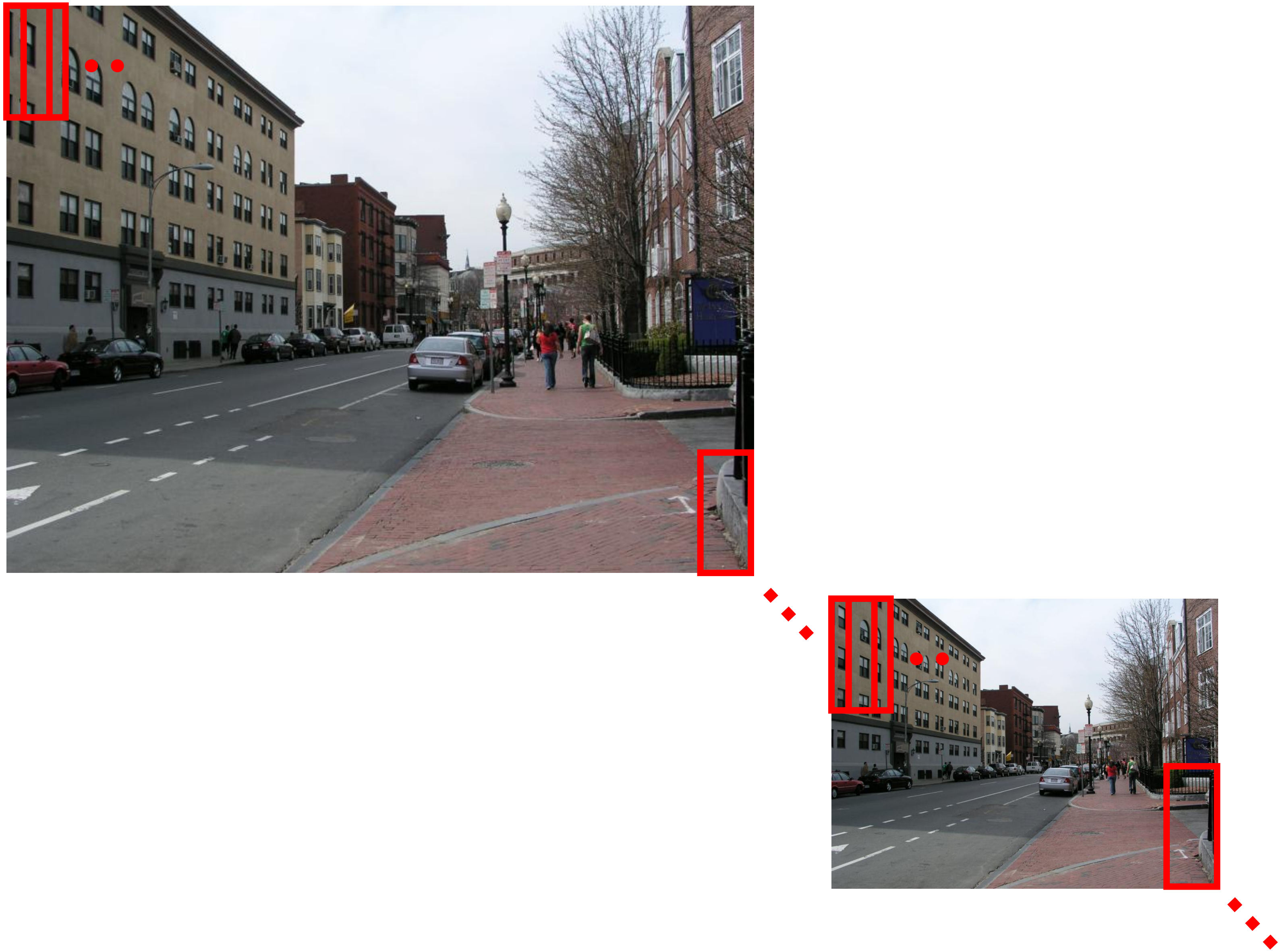
2. Compare
   - Compute similarity to an example object or to a summary representation
   - Which differences in appearance are important?

Aligned Possible Objects

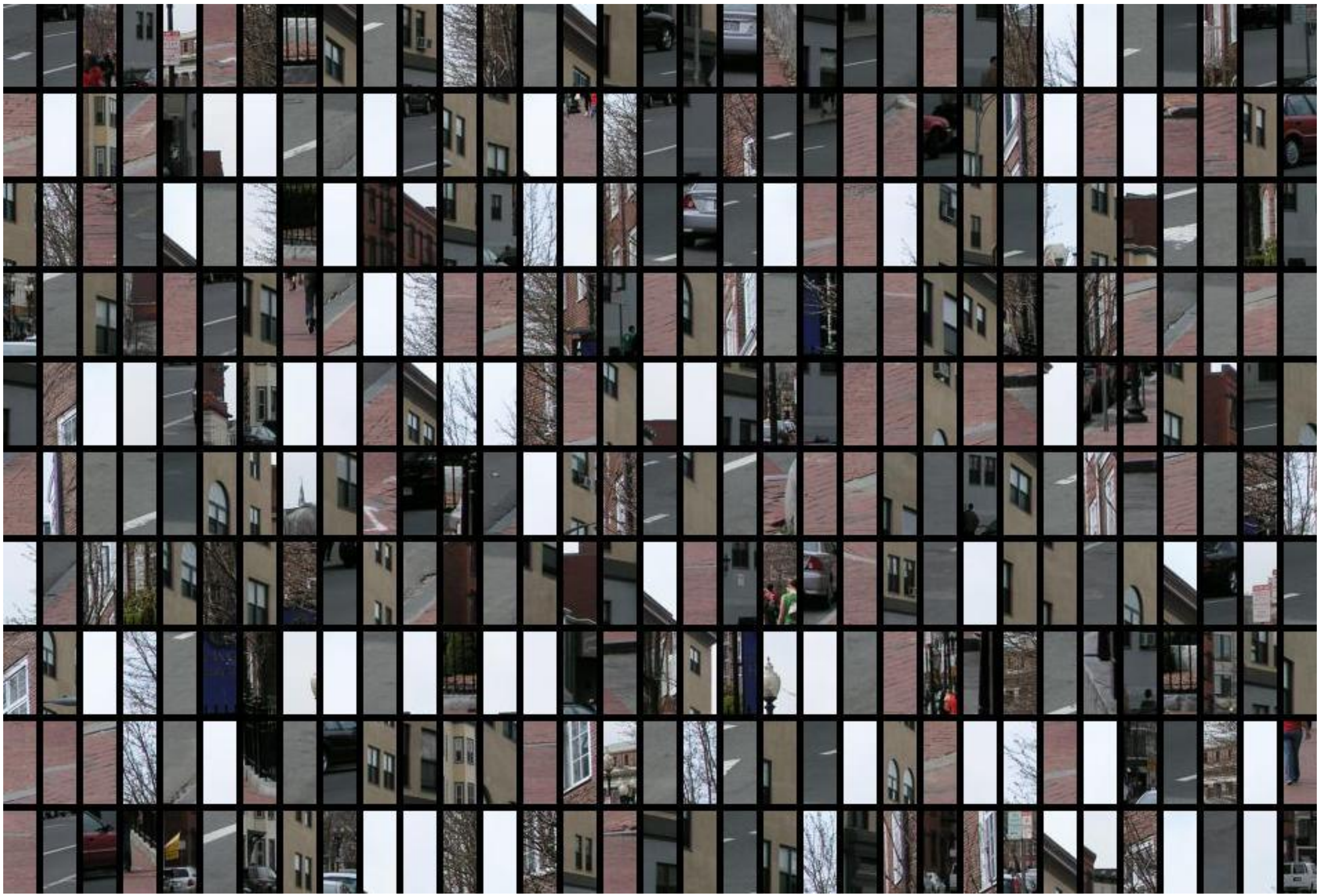
Exemplar

Summary

Sliding window: a simple alignment solution



Each window is separately classified

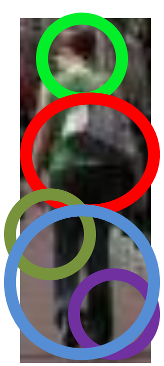


Statistical Template

- Object model = sum of scores of features at fixed positions

+3 +2 -2 -1 -2.5 = -0.5 $\overset{?}{>}$ **7.5**

Non-object



+4 +1 +0.5 +3 +0.5 = 10.5 $\overset{?}{>}$ **7.5**

Object

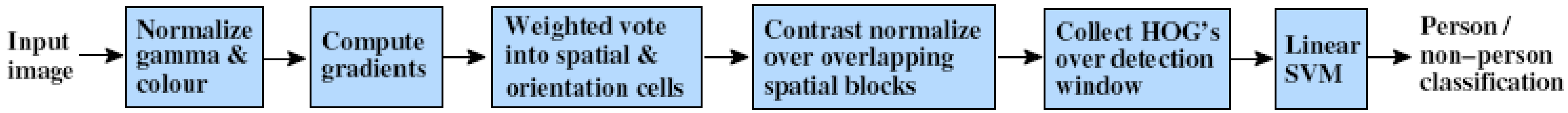
Design challenges

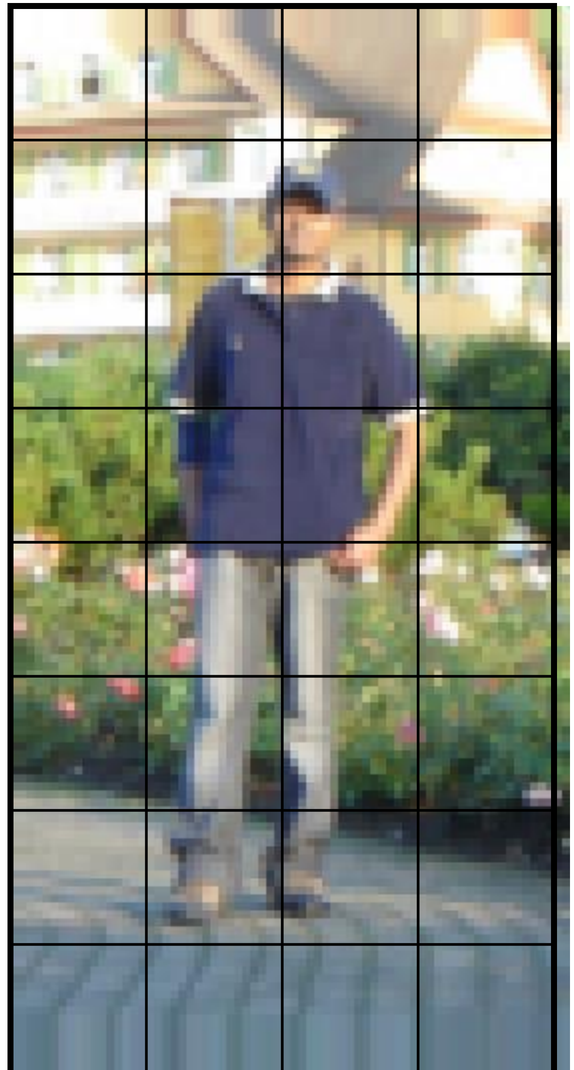
- How to efficiently search for likely objects
  - Sliding windows require searching hundreds of thousands of positions and scales
- Feature design and scoring
  - How should appearance be modeled? What features correspond to the object?
- How to deal with different viewpoints?
  - Often train different models for a few different viewpoints
- Implementation details
  - Window size
  - Aspect ratio
  - Translation/scale step size
  - Non-maxima suppression

Example: Dalal-Triggs pedestrian detector

1. Extract fixed-sized (64x128 pixel) window at each position and scale
2. Compute HOG (histogram of gradient) features within each window
3. Score the window with a linear SVM classifier
4. Perform non-maxima suppression to remove overlapping detections with lower scores

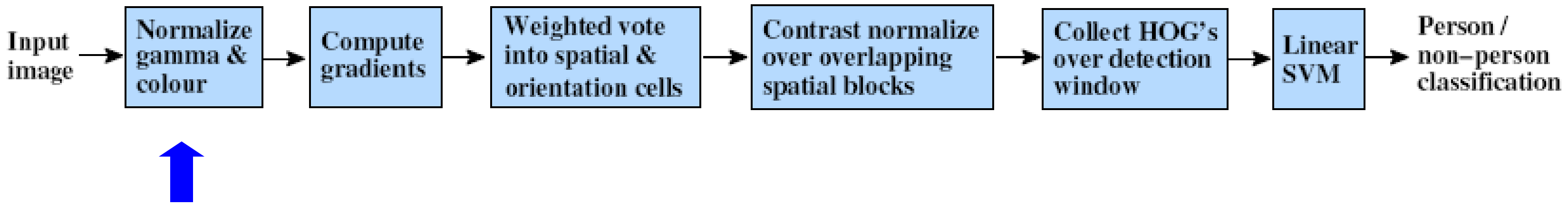
Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05





Slides by Pete Barnum

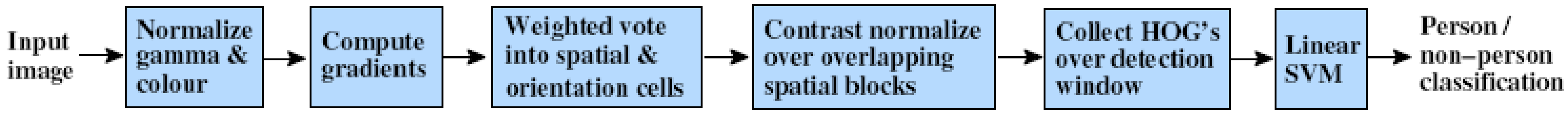
Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05



- Tested with
  - RGB
  - LAB
  - Grayscale

  RGB, LAB: Slightly better performance vs. grayscale
- Gamma Normalization and Compression
  - Square root
  - Log

  Square root: Very slightly better performance vs. no adjustment



Outperforms

-1	0	1

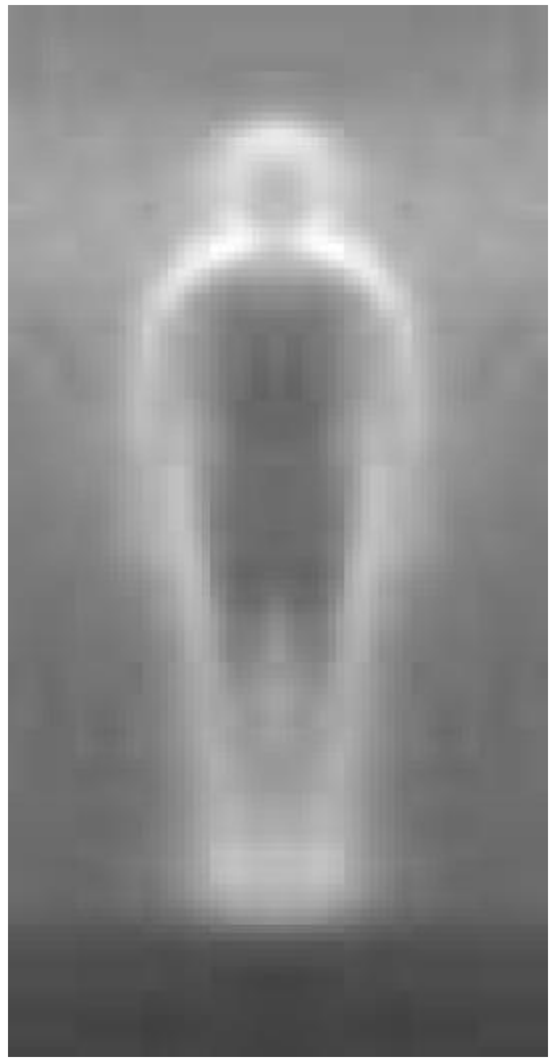
centered

-1	1

uncentered

1	-8	0	8	-1

cubic-corrected



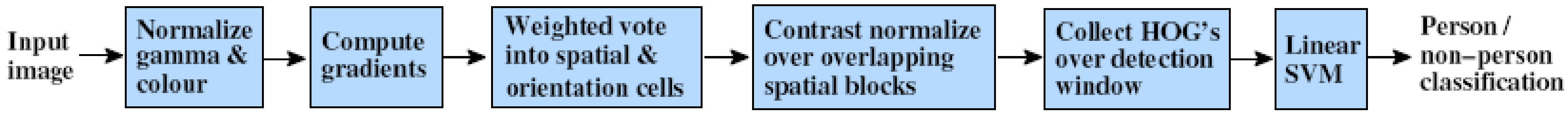
0	1
-1	0

diagonal

-1	0	1
-2	0	2
-1	0	1

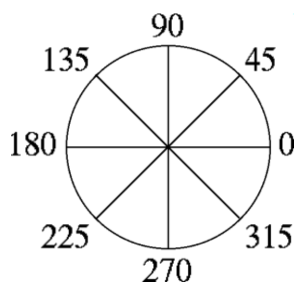
Sobel

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05

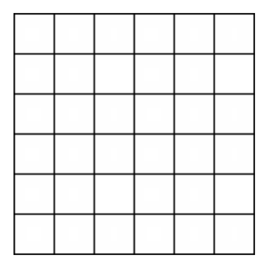


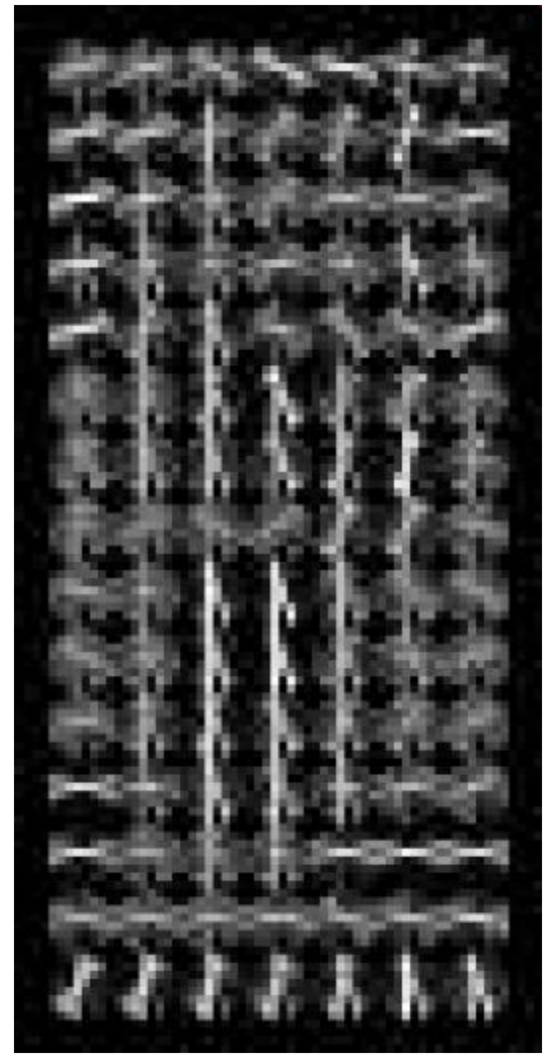
- Histogram of gradient orientations

Orientation: 9 bins (for unsigned angles)



Histograms in 8x8 pixel cells

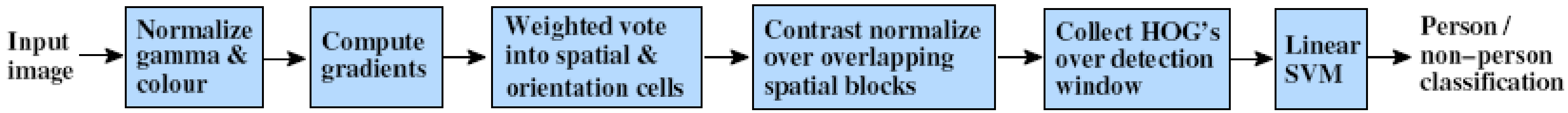




  – Votes weighted by magnitude
  – Bilinear interpolation between cells

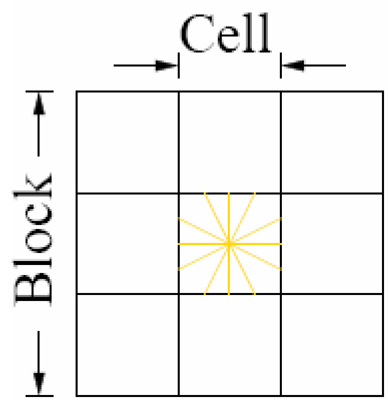
Slides by Pete Barnum

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05

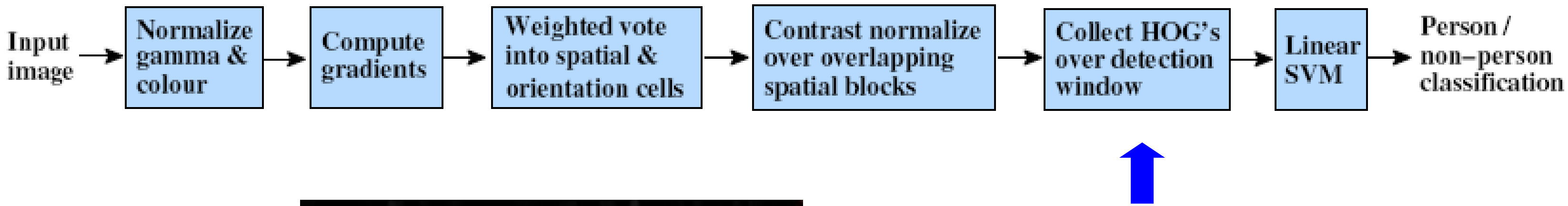


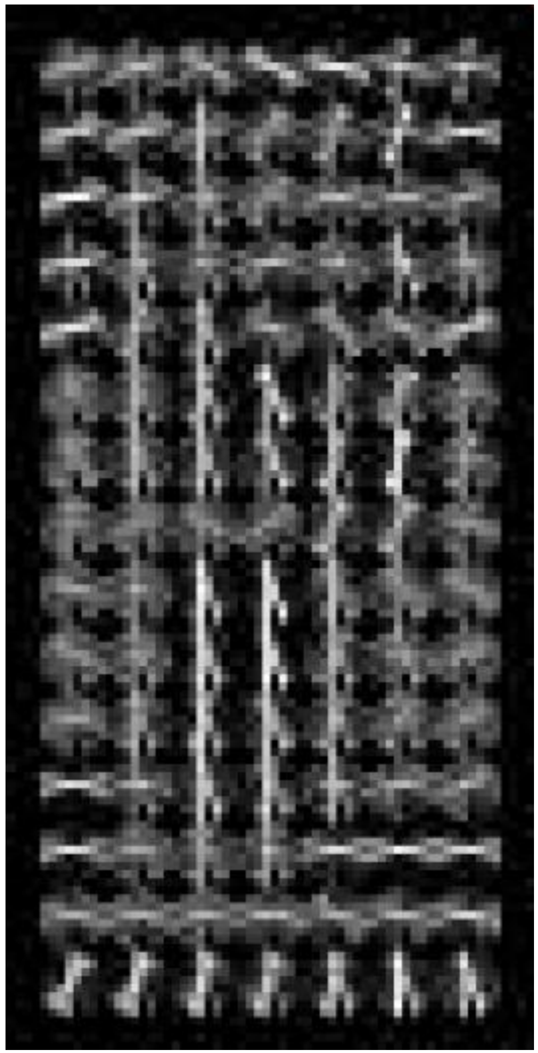
R-HOG

Normalize with respect to surrounding cells



$$L2-norm : v \longrightarrow v / \sqrt{||v||_2^2 + \epsilon^2}$$



X= 

orientations

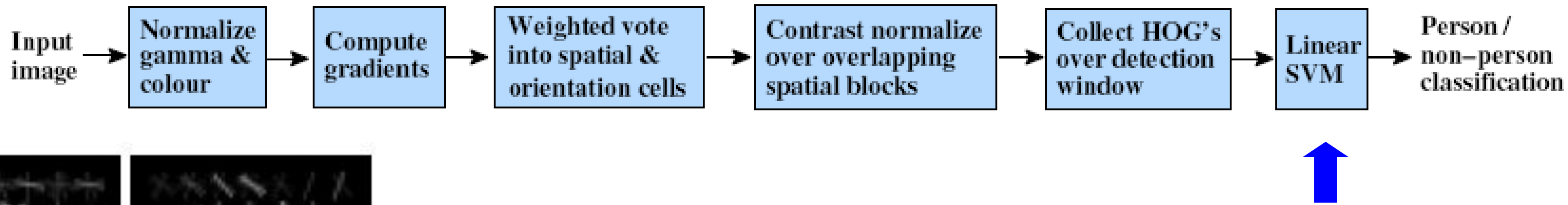
features = 15 x 7 x 9 x 4 = 3780

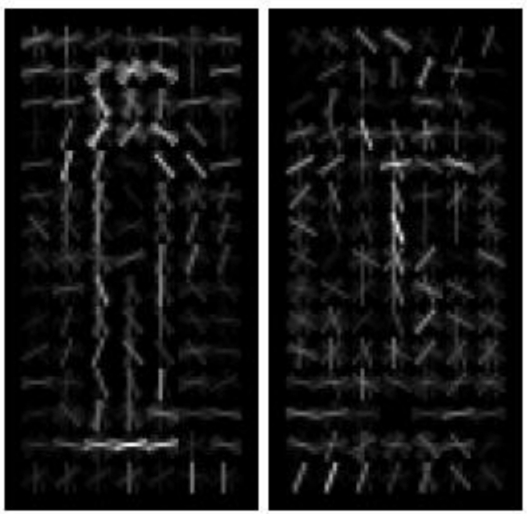
cells

normalizations by neighboring cells

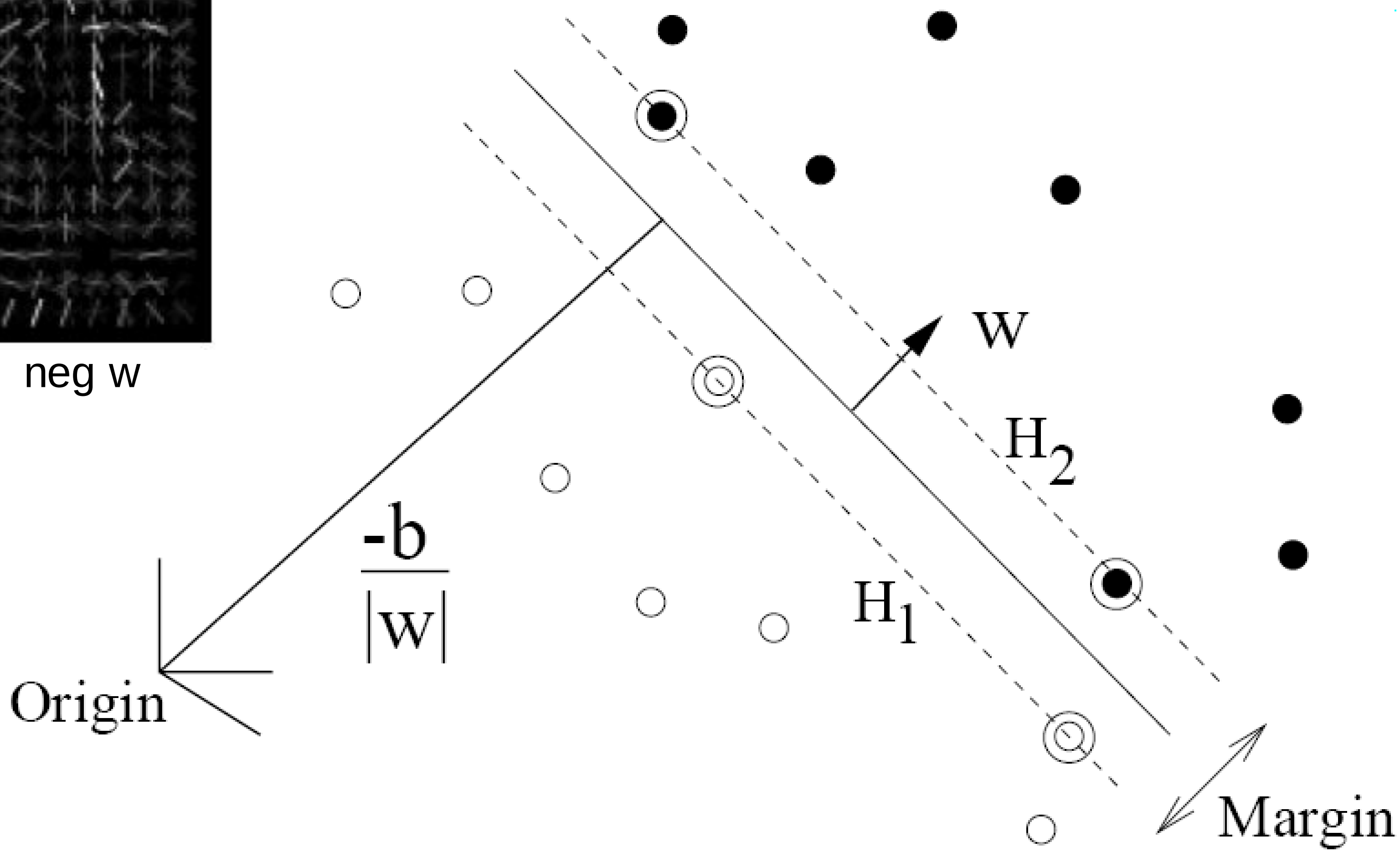
Slides by Pete Barnum

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05



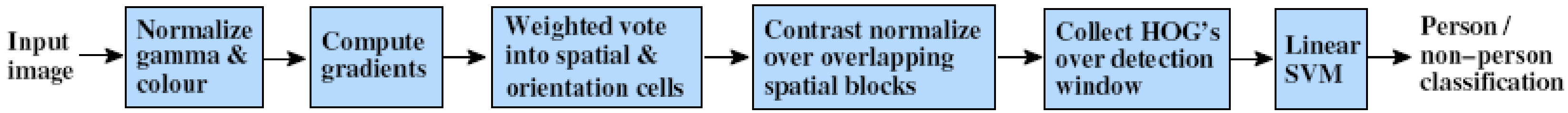


pos w neg w



Slides by Pete Barnum

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05



$$0.16 = w^T x - b$$

$$sign(0.16) = 1$$

$=>$ pedestrian

Slides by Pete Barnum

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR05

Detection examples

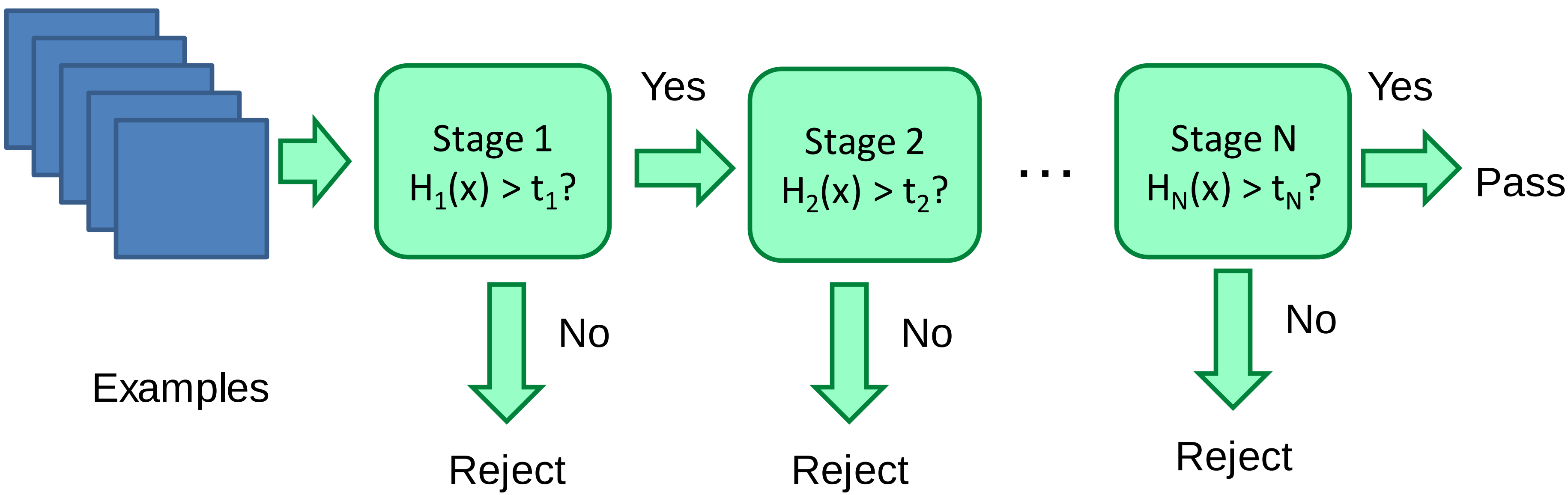
Viola-Jones sliding window detector

Fast detection through two mechanisms

- Quickly eliminate unlikely windows
- Use features that are fast to compute

Viola and Jones. Rapid Object Detection using a Boosted Cascade of Simple Features (2001).

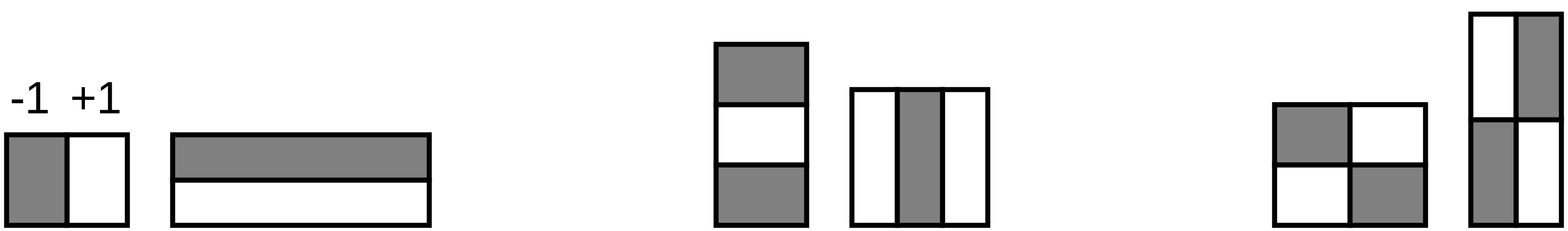
Cascade for Fast Detection



- Choose threshold for low false negative rate
- Fast classifiers early in cascade
- Slow classifiers later, but most examples don't get there

Features that are fast to compute

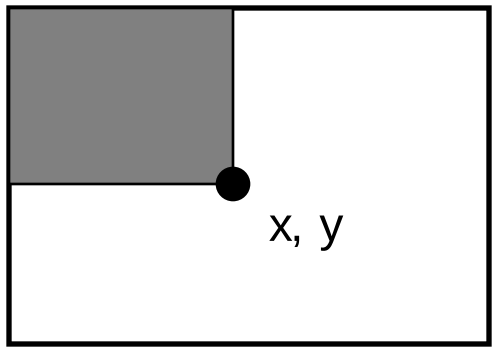
- “Haar-like features”
  - Differences of sums of intensity
  - Thousands, computed at various positions and scales within detection window



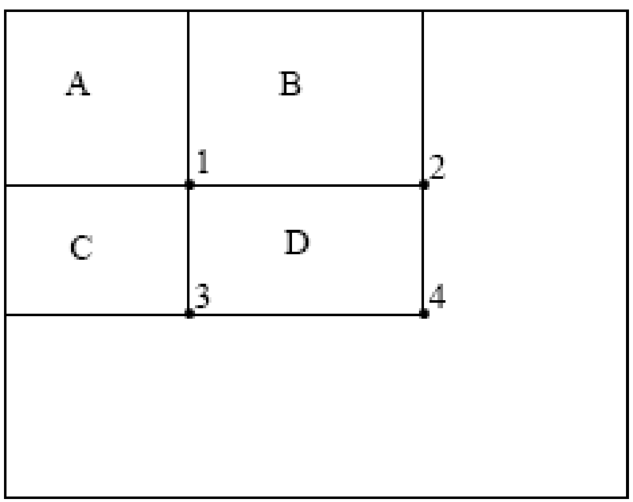
Two-rectangle features Three-rectangle features Etc.

Integral Images

- `ii = cumsum(cumsum(im, 1), 2)`



ii(x,y) = Sum of the values in the grey region



How to compute B-A?

How to compute A+D-B-C?

Feature selection with Adaboost

- Create a large pool of features (180K)
- Select features that are discriminative and work well together
  – “Weak learner” = feature + threshold + parity

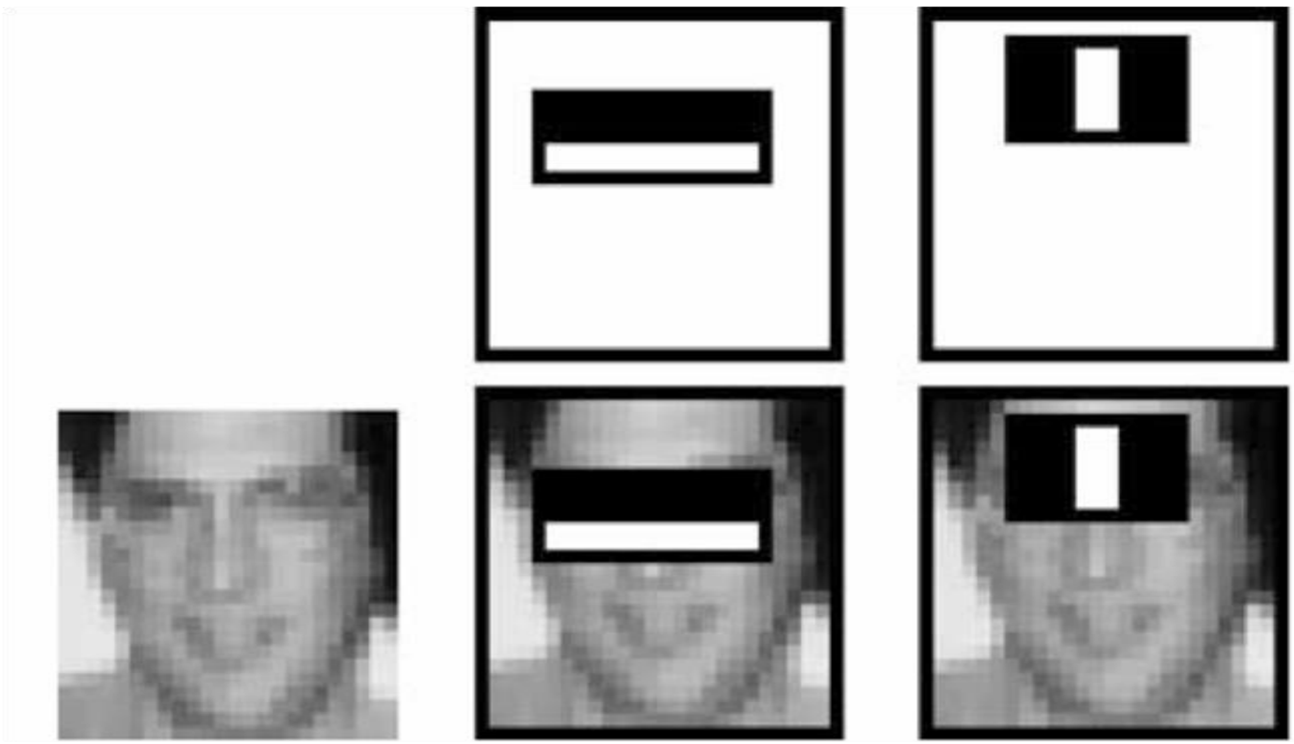
$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j \theta_j \\ 0 & \text{otherwise} \end{cases}$$

  – Choose weak learner that minimizes error on the weighted training set
  – Reweight

Adaboost

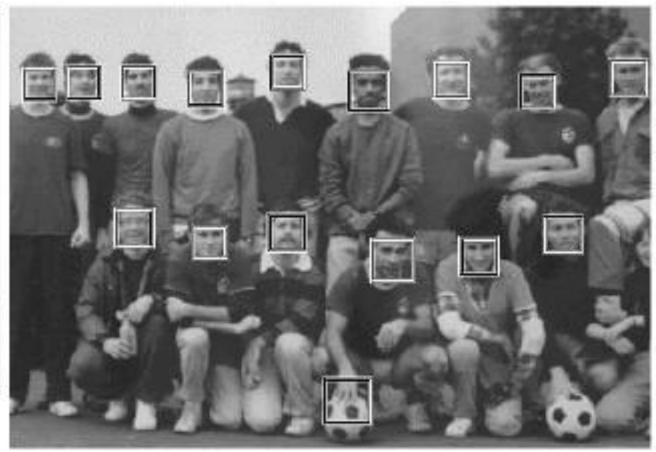
- Given example images $(x_1, y_1), \dots, (x_n, y_n)$ where $y_i = 0, 1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$ respectively, where m and l are the number of negatives and positives respectively.
- For $t = 1, \dots, T$:
  1. Normalize the weights,
  $$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$
  so that w_t is a probability distribution.
  2. For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to w_t, $\epsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$.
  3. Choose the classifier, h_t, with the lowest error ϵ_t.
  4. Update the weights:
  $$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$$
  where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.
- The final strong classifier is:
$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & \text{otherwise} \end{cases}$$
where $\alpha_t = \log \frac{1}{\beta_t}$

Top 2 selected features



Viola Jones Results

Speed = 15 FPS (in 2001)

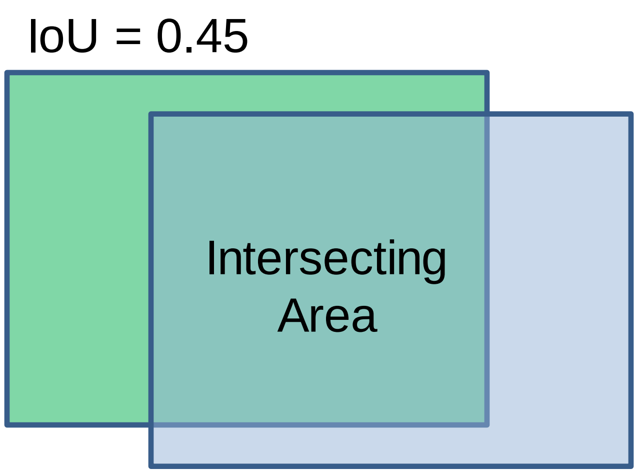

Detector \ False detections	10	31	50	65	78	95	167
Viola-Jones	76.1%	88.4%	91.4%	92.0%	92.1%	92.9%	93.9%
Viola-Jones (voting)	81.1%	89.7%	92.1%	93.1%	93.1%	93.2 %	93.7%
Rowley-Baluja-Kanade	83.2%	86.0%	-	-	-	89.2%	90.1%
Schneiderman-Kanade	-	-	-	94.4%	-	-	-
Roth-Yang-Ahuja	-	-	-	-	(94.8%)	-	-

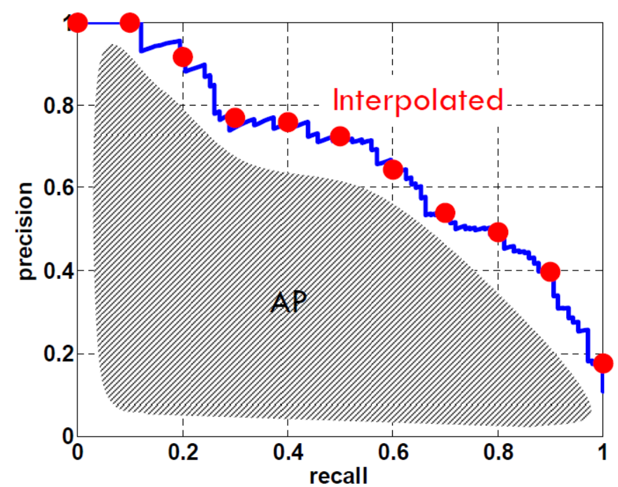
MIT + CMU face dataset

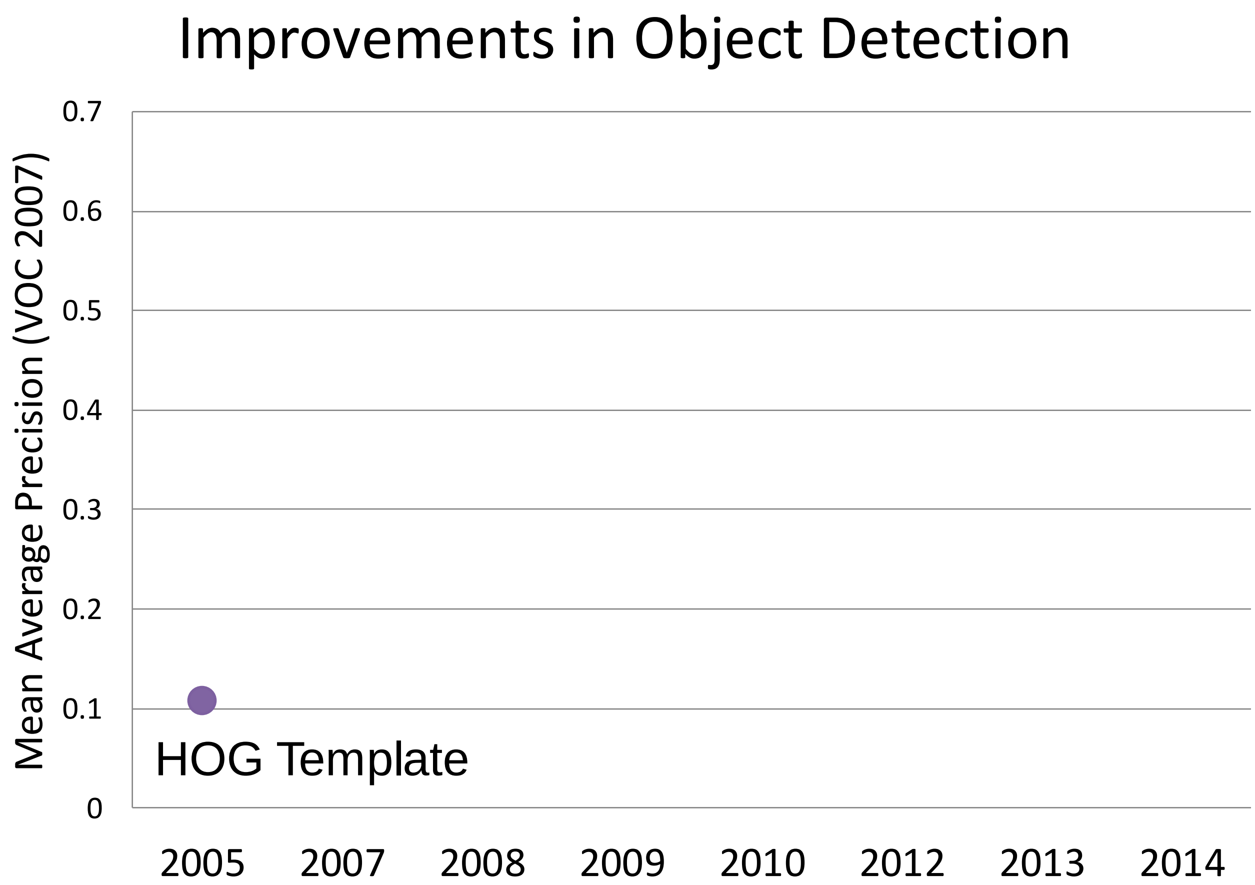
Object Detection Evaluation

- Datasets
  - PASCAL VOC (2005-2012): 20 classes, ~20,000 images
  - MS COCO (2014-?): 60 classes, ~300,000 images

- Evaluation
  - Output: for each class, predict bounding boxes (x1, y1, x2, y2) with confidences
  - Metric:
    - True detection: >= 0.5 Intersection over Union (IoU), not a duplicate
    - Precision: $\frac{\#\ true\ detections}{\#\ detections}$ Recall: $\frac{\#\ true\ detections}{\#\ positive\ examples}$
    - AP: area under the interpolated curve

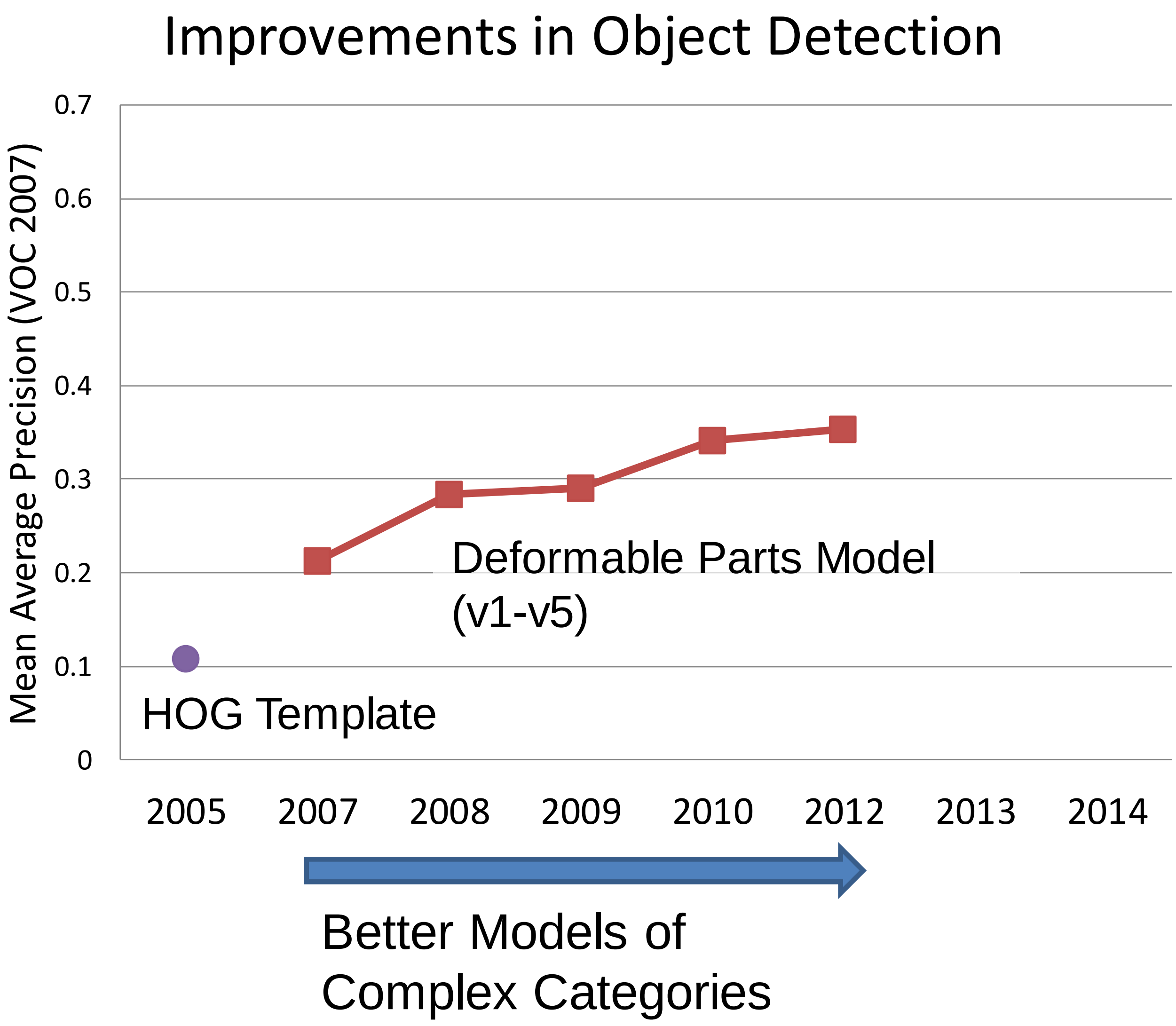




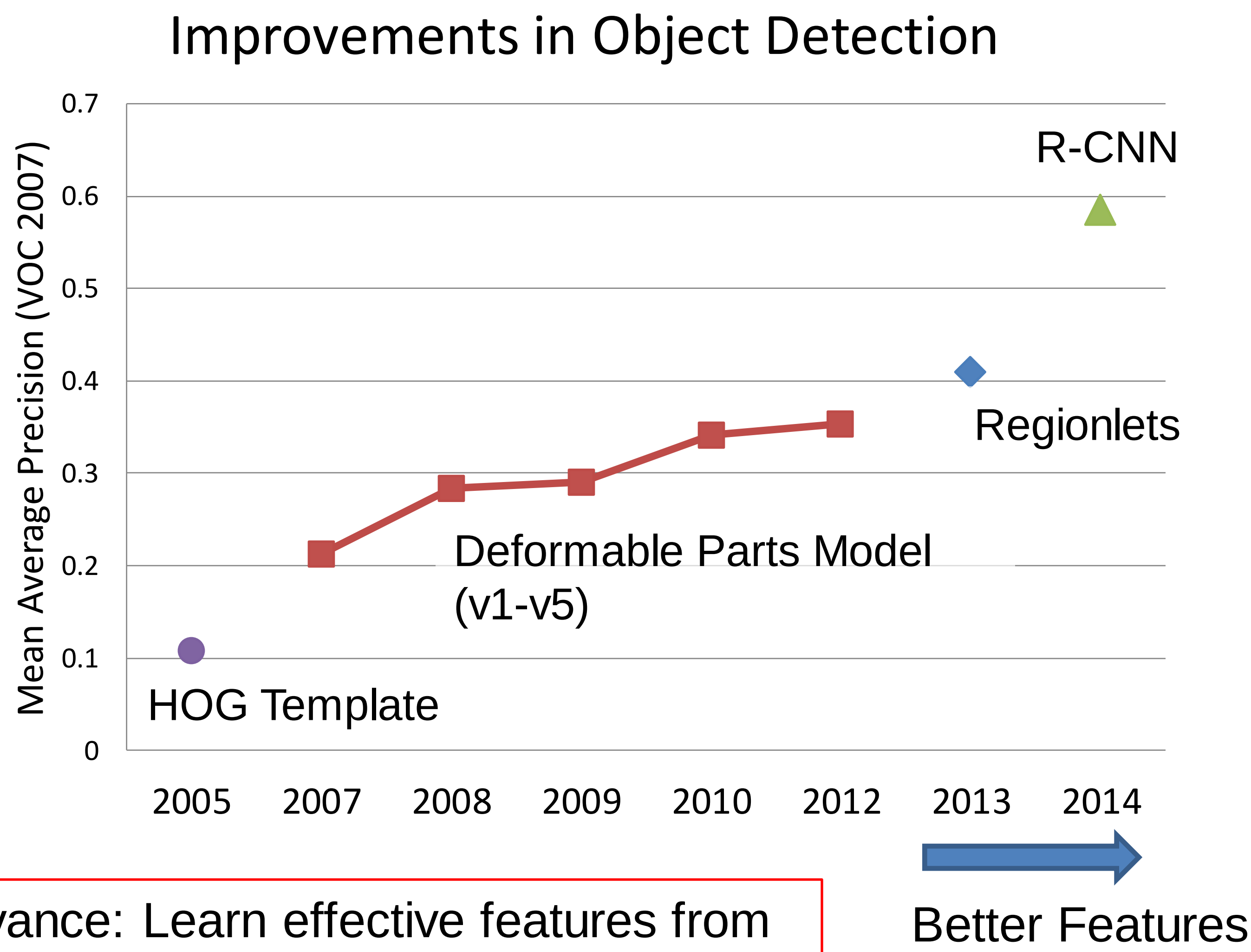


Statistical Template Matching

HOG: Dalal-Triggs 2005



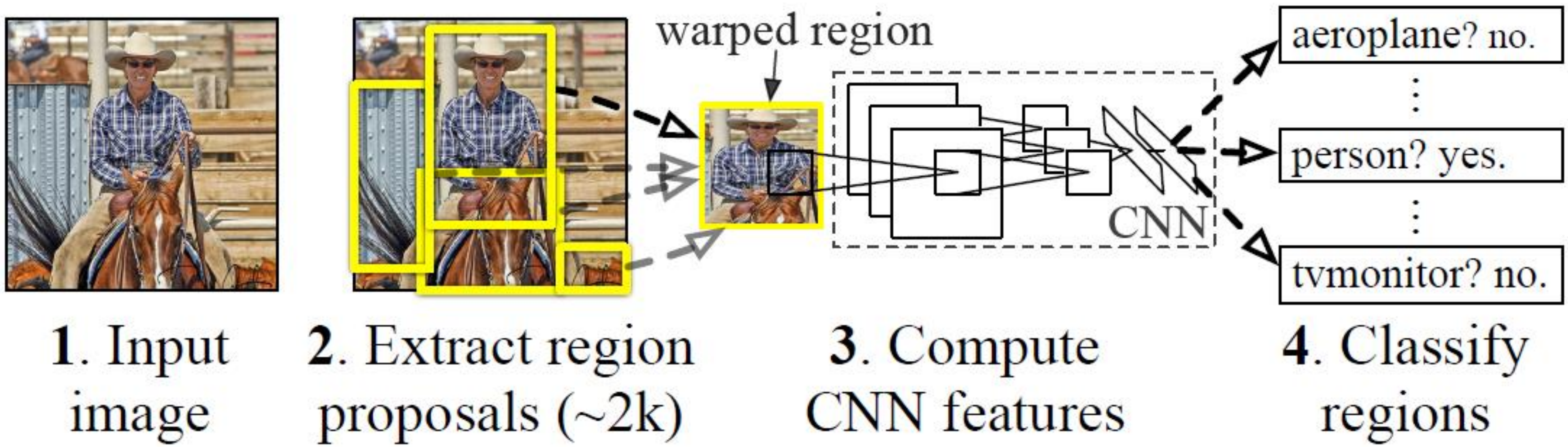
HOG: Dalal-Triggs 2005 DPM: Felzenszwalb et al. 2008-2012



Key Advance: Learn effective features from massive amounts of labeled data *and* adapt to new tasks with less data

HOG: Dalal-Triggs 2005 DPM: Felzenszwalb et al. 2008-2012 Regionlets: Wang et al. 2013 R-CNN: Girshick et al. 2014

R-CNN (Girshick et al. CVPR 2014)



- Replace sliding windows with "selective search" region proposals (Uijilings et al. IJCV 2013)
- Extract rectangles around regions and resize to 227x227
- Extract features with fine-tuned CNN (that was initialized with network trained on ImageNet before training)
- Classify last layer of network features with SVM

http://arxiv.org/pdf/1311.2524.pdf

Fine-tuning example: ImageNet->VOC

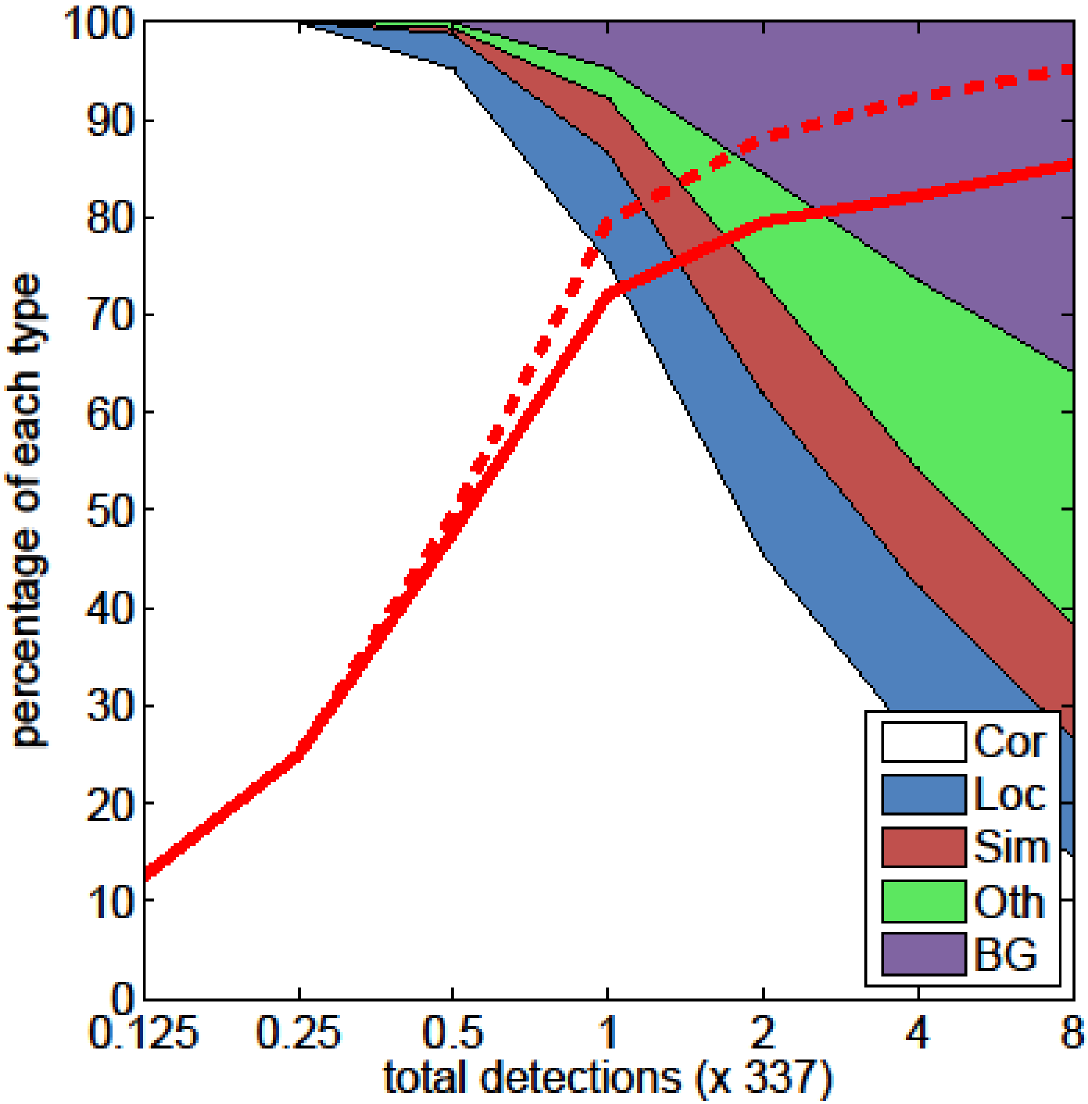
1. Train full network on ImageNet 1000-class classification
2. Replace classification layer with output layer for VOC (e.g. confidences for 20 classes)
3. Train on VOC pos/neg examples with low initial learning rate (1/10th what is used for new network)

Notes

- This usually works well if the "big data" task and target task are similar (object classification vs detection)
  - 0.45 AP without fine-tuning → 0.54 AP with fine tuning; training only on VOC does much worse
- Not necessary if target task is also very big

Mistakes are often reasonable

Bicycle: AP = 0.73



Confident Mistakes


bicycle (loc): ov=0.44 1−r=0.70

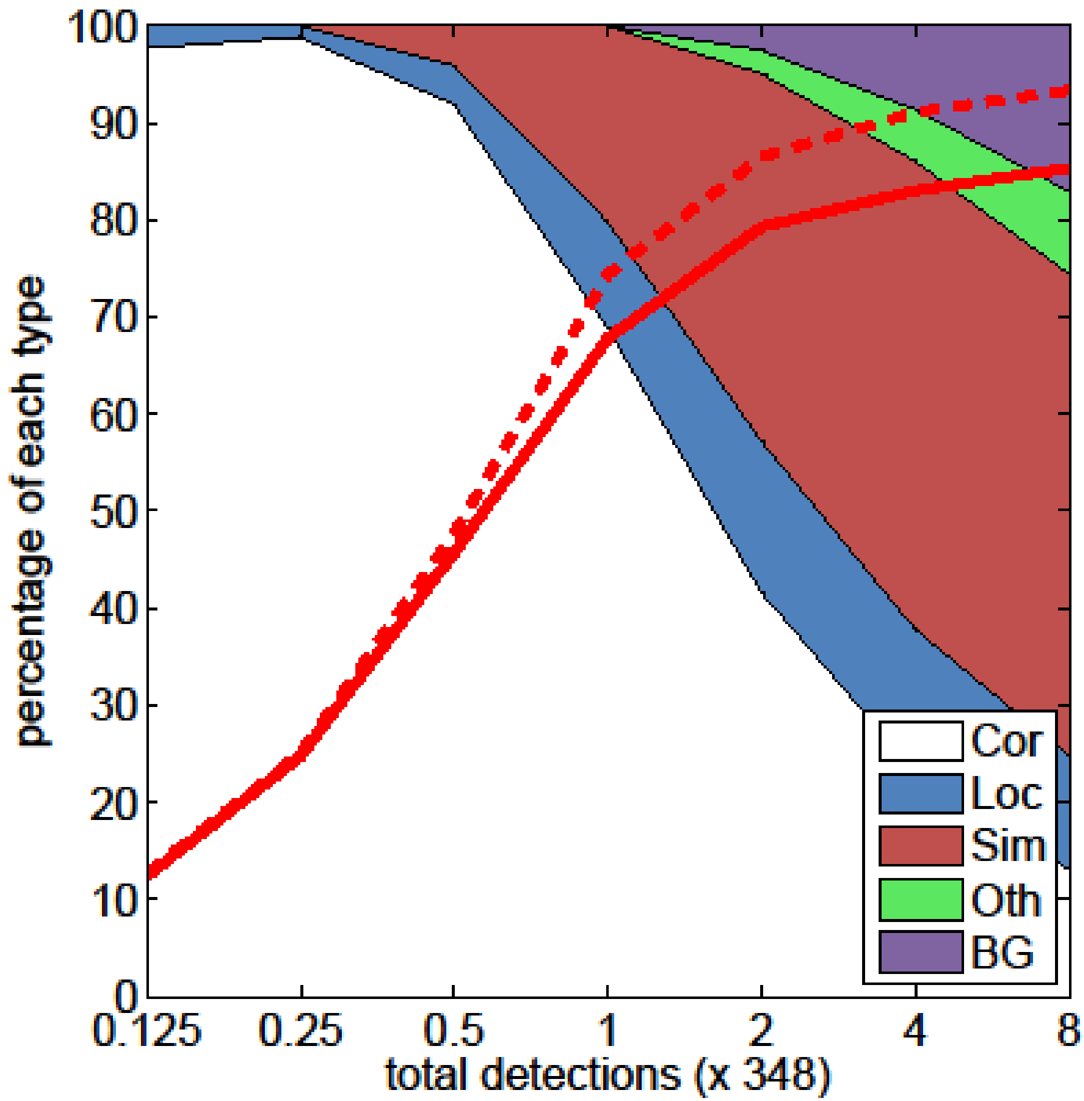

bicycle (sim): ov=0.00 1−r=0.56


bicycle (bg): ov=0.00 1−r=0.47

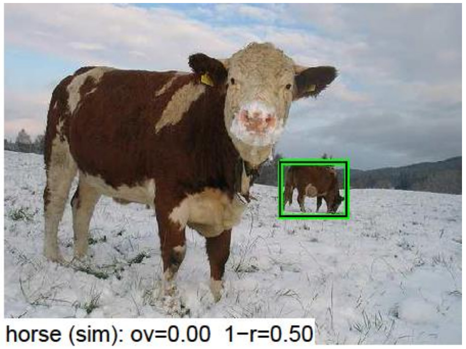
R-CNN results

Mistakes are often reasonable

Horse: AP = 0.69

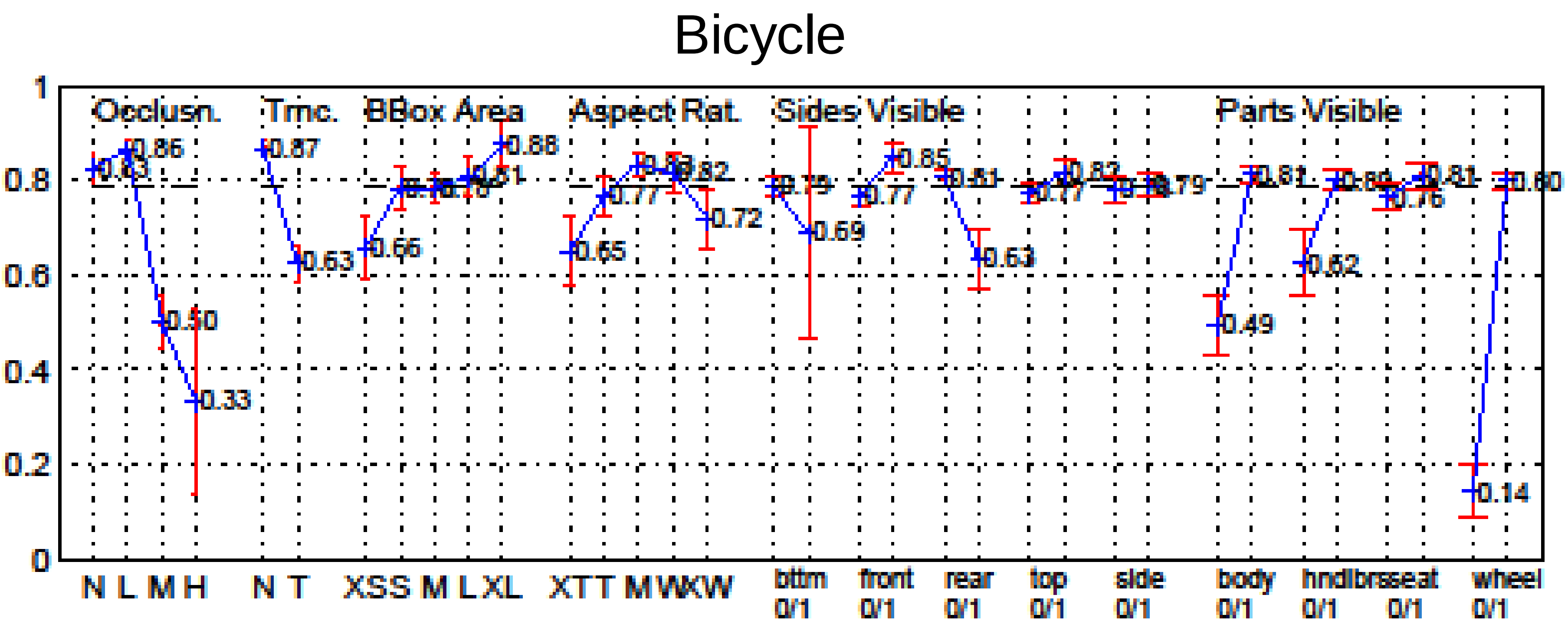


Confident Mistakes



R-CNN results

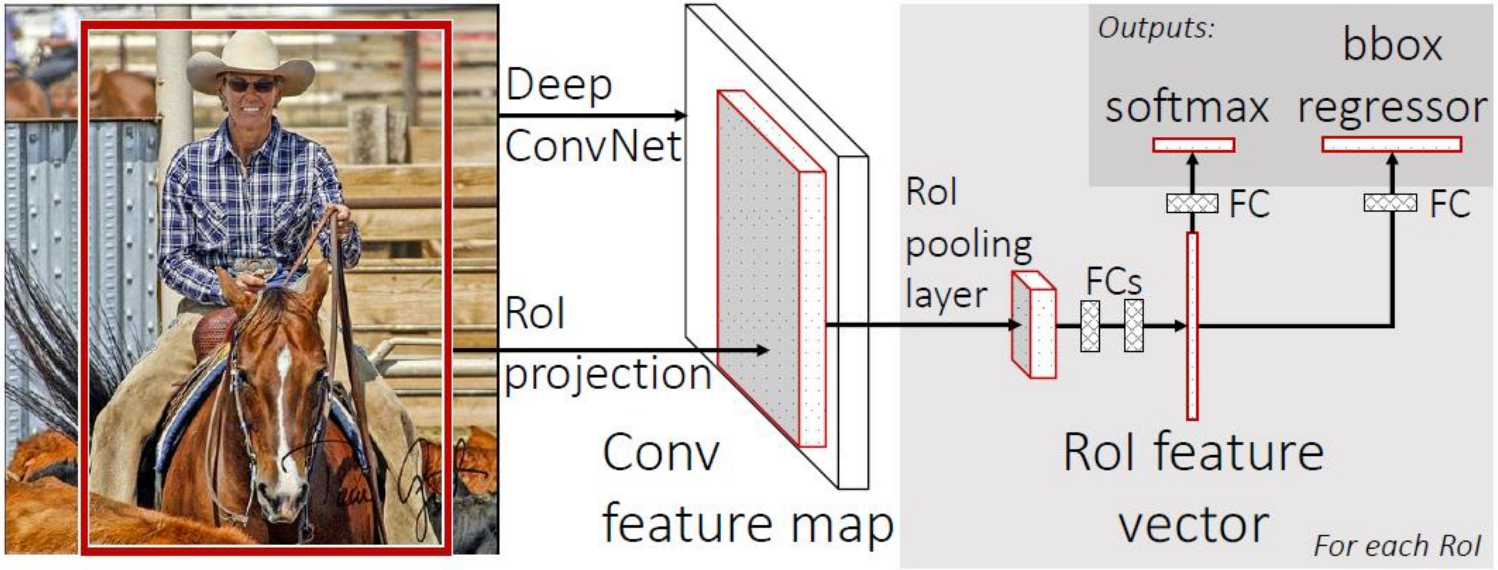
Misses are often predictable



Small objects, distinctive parts absent
or occluded, unusual views

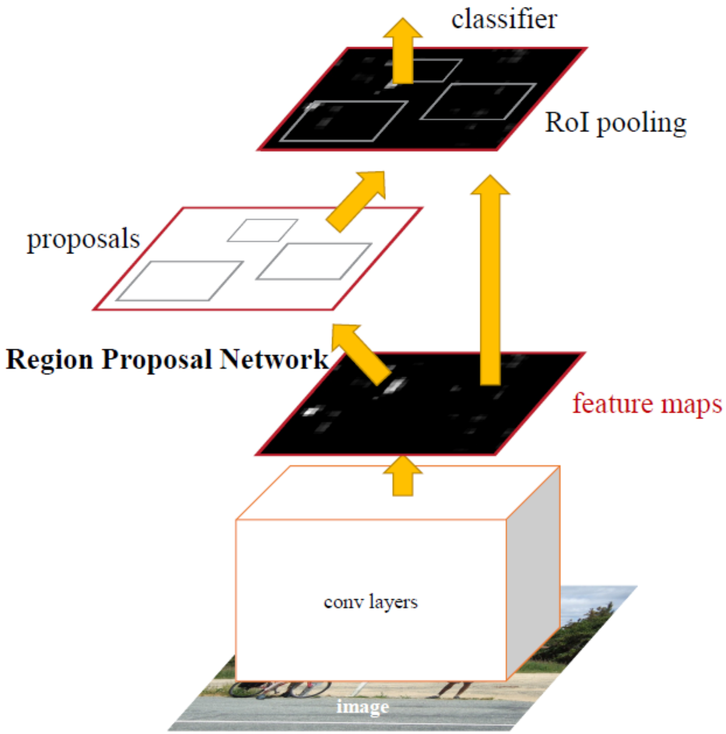
R-CNN results

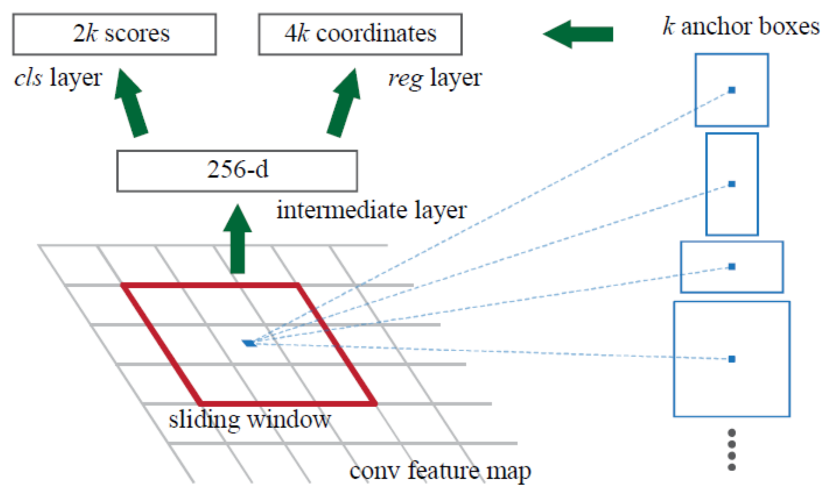
Fast R-CNN – Girshick 2015



- Compute CNN features for image once
- Pool into 7x7 spatial bins for each region proposal, output class scores and regressed bboxes
- 100x speed up of R-CNN (0.02 – 0.1 FPS → 0.5-20 FPS) with similar accuracy

Faster R-CNN – Ren et al. 2016





- Convolutional features used for generating proposals and scoring
  - Generate proposals with “objectness” scores and refined bboxes for each of k “anchors”
  - Score proposals in same way as Fast R-CNN
- Similar accuracy to Fast R-CNN with 10x speedup

- Faster R-CNN slightly better accuracy than Fast R-CNN
- More data improves results considerably

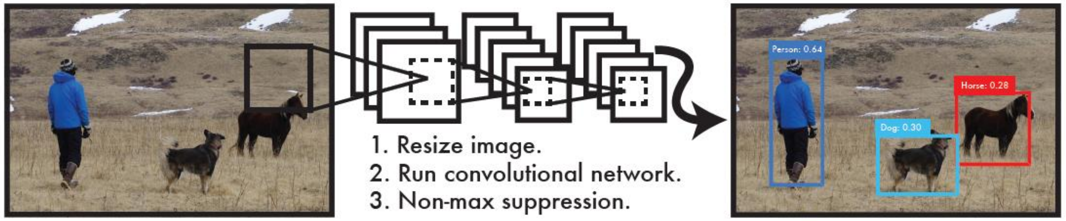
Table 6: Results on PASCAL VOC 2007 test set with Fast R-CNN detectors and VGG-16. For RPN, the train-time proposals for Fast R-CNN are 2000. RPN* denotes the unsharing feature version.

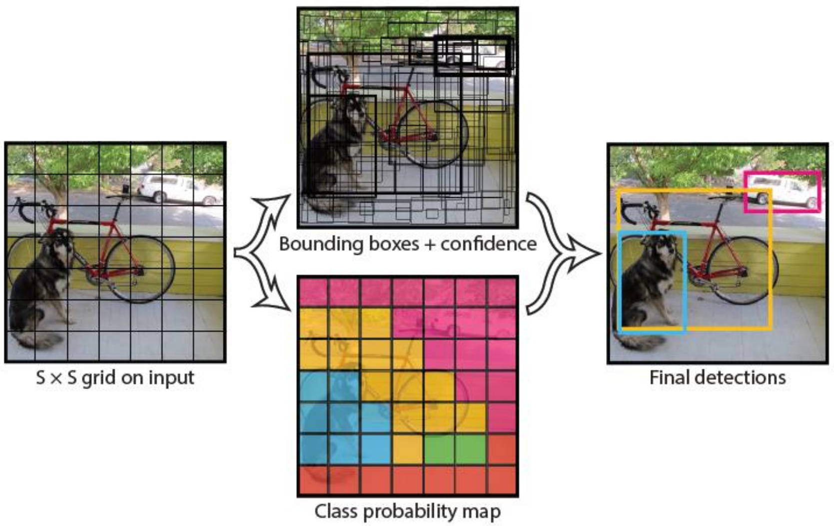
method	# box	data	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
SS	2000	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
SS	2000	07+12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
RPN*	300	07	68.5	74.1	77.2	67.7	53.9	51.0	75.1	79.2	78.9	50.7	78.0	61.1	79.1	81.9	72.2	75.9	37.2	71.4	62.5	77.4	66.4
RPN	300	07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
RPN	300	07+12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
RPN	300	COCO+07+12	**78.8**	**84.3**	**82.0**	**77.7**	**68.9**	**65.7**	**88.1**	**88.4**	**88.9**	**63.6**	**86.3**	**70.8**	**85.9**	**87.6**	**80.1**	**82.3**	**53.6**	**80.4**	**75.8**	**86.6**	**78.9**

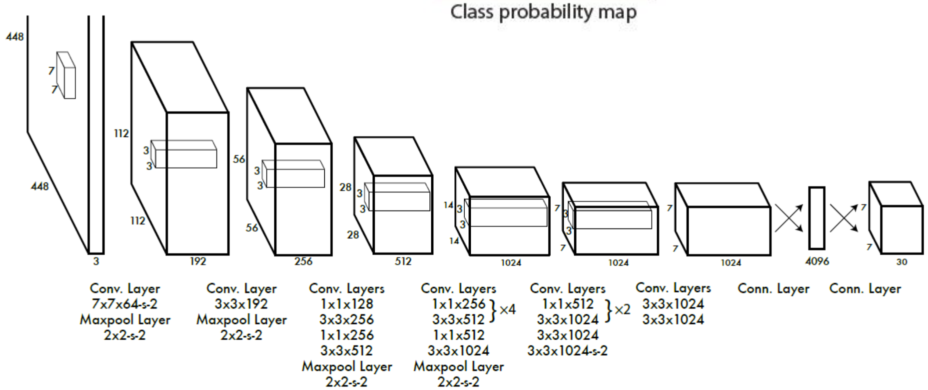
[YOLO](#) – Redmon et al. 2016

1. CNN produces 4096 features for 7x7 grid on image (fully conv.)
2. Each cell produces a score for each object and 2 bboxes w/ conf
3. Non-max suppression

- 7x speedup over Faster RCNN (45-155 FPS vs. 7-18 FPS)
- Some loss of accuracy due to lower recall, poor localization







Yolo v2 – Redmon et al. 2017

- Batch normalization
- Pre-train on higher resolution ImageNet
- Use and improve anchor box idea from Faster RCNN
- Train at multiple resolutions
- Very good accuracy, very fast

	YOLO								YOLOv2
batch norm?		✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?			✓	✓	✓	✓	✓	✓	✓
convolutional?				✓	✓	✓	✓	✓	✓
anchor boxes?				✓	✓				
new network?					✓	✓	✓	✓	✓
dimension priors?						✓	✓	✓	✓
location prediction?						✓	✓	✓	✓
passthrough?							✓	✓	✓
multi-scale?								✓	✓
hi-res detector?									✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	**78.6**

Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19
YOLOv2 288×288	2007+2012	69.0	91
YOLOv2 352×352	2007+2012	73.7	81
YOLOv2 416×416	2007+2012	76.8	67
YOLOv2 480×480	2007+2012	77.8	59
YOLOv2 544×544	2007+2012	**78.6**	40

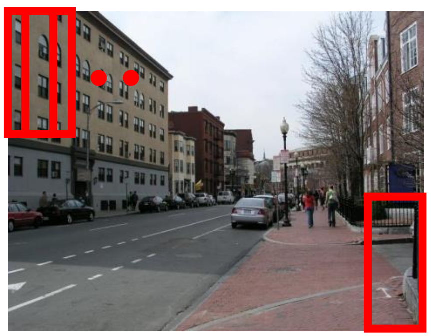
https://youtu.be/VOC3huqHrss

Influential Works in Detection

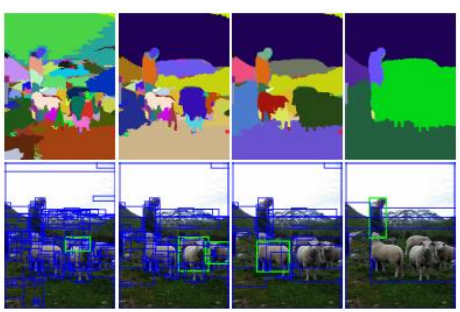
- Sung-Poggio (1994, 1998) : ~2412 citations
  - Basic idea of statistical template detection (I think), bootstrapping to get “face-like” negative examples, multiple whole-face prototypes (in 1994)
- Rowley-Baluja-Kanade (1996-1998) : ~4953
  - “Parts” at fixed position, non-maxima suppression, simple cascade, rotation, pretty good accuracy, fast
- Schneiderman-Kanade (1998-2000,2004) : ~2600
  - Careful feature/classifier engineering, excellent results, cascade
- Viola-Jones (2001, 2004) : ~27,000
  - Haar-like features, Adaboost as feature selection, hyper-cascade, very fast, easy to implement
- Dalal-Triggs (2005) : ~18000
  - Careful feature engineering, excellent results, HOG feature, online code
- Felzenszwalb-Huttenlocher (2000): ~2100
  - Efficient way to solve part-based detectors
- Felzenszwalb-McAllester-Ramanan (2008,2010): ~7200
  - Excellent template/parts-based blend
- Girshick-Donahue-Darrell-Malik (2014): ~4700
  - Region proposals + fine-tuned CNN features (marks significant advance in accuracy over hog-based methods)
- Redmon, Divvala, Girshick, Farhadi (2016): ~210
  - Refine and simplify RCNN++ approach to predict directly from last conv layer

Summary: statistical templates

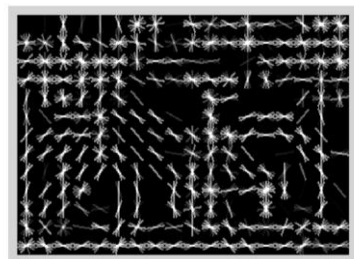
Propose Window → Extract Features → Classify → Post-process



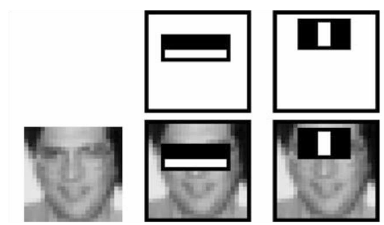
Sliding window: scan image pyramid



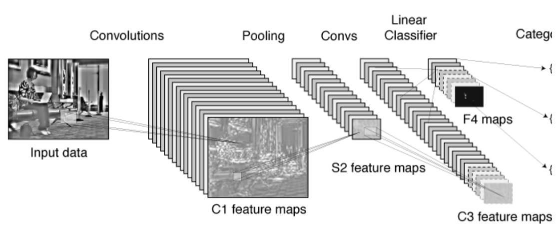
Region proposals: edge/region-based, resize to fixed window



HOG



Fast randomized features



CNN features

SVM

Boosted stubs

Neural network

Non-max suppression

Segment or refine localization

Next class

- Pixel/part labeling