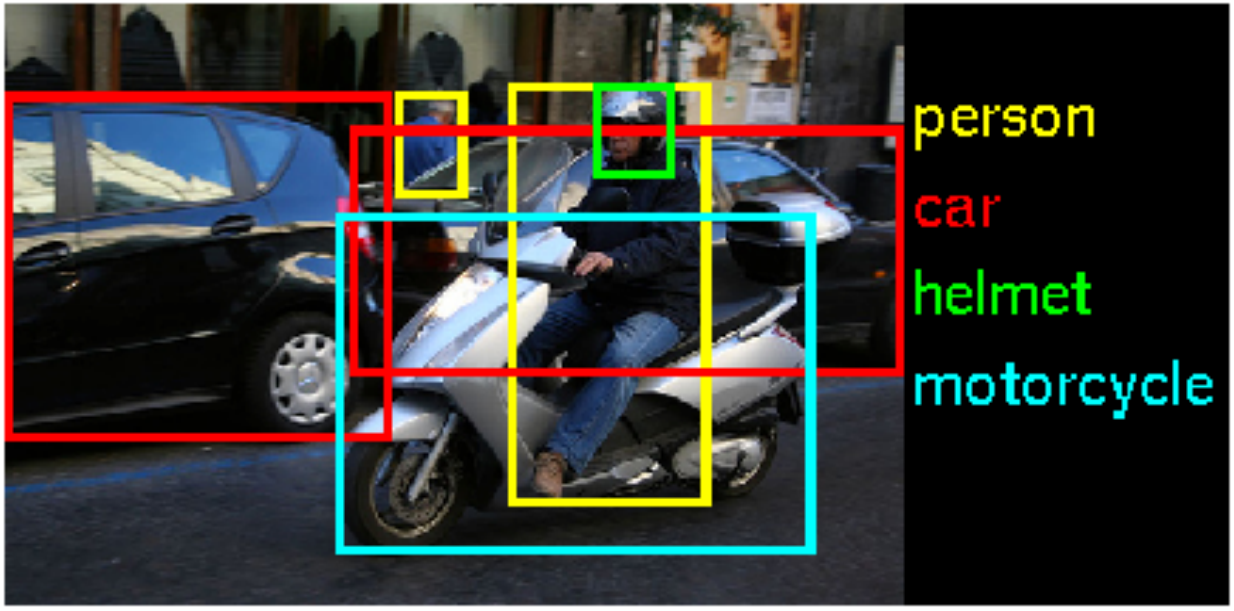
Lecture 6: Introduction to Detection

Jonathan Krause

Goal

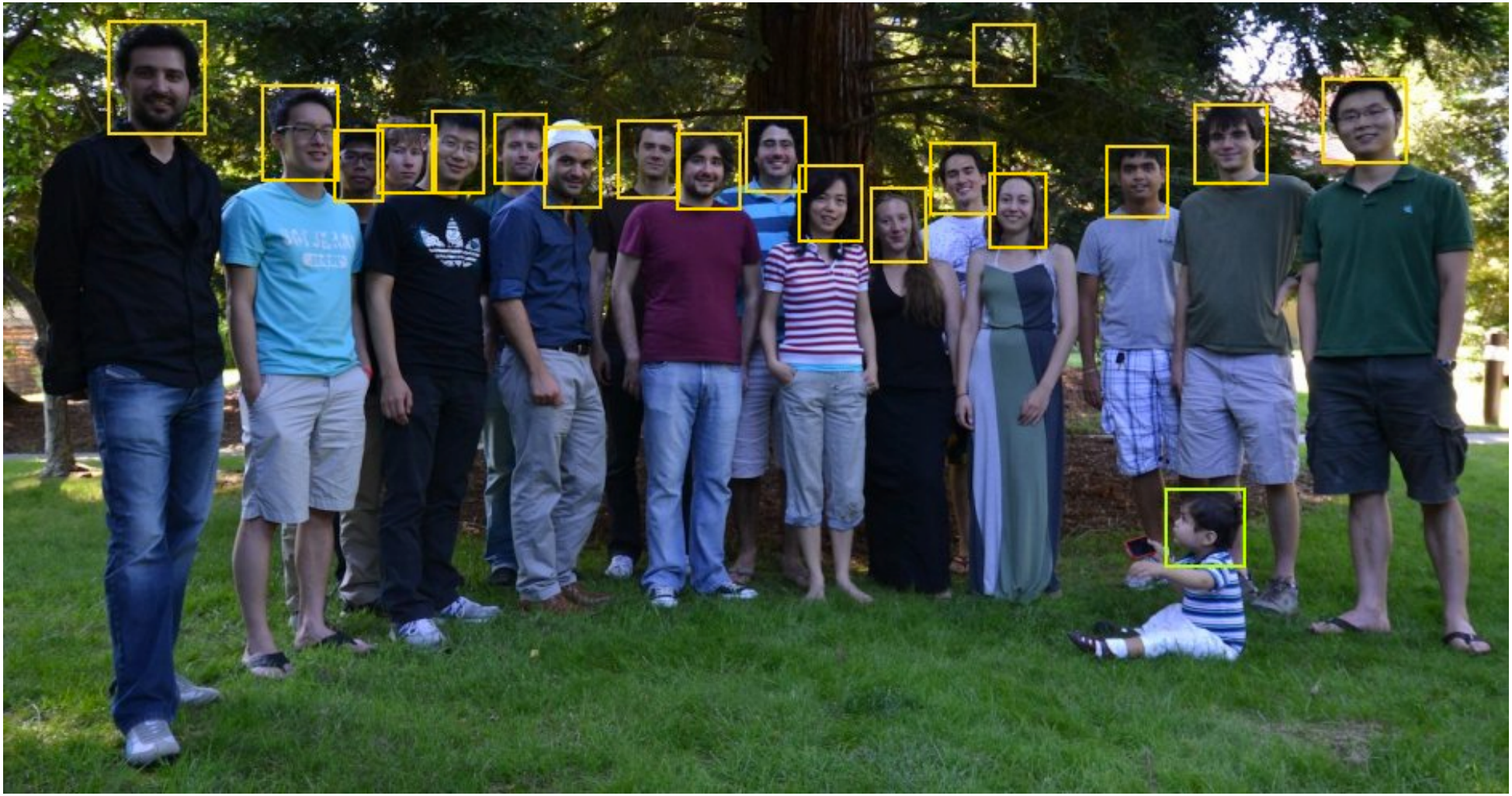
- Locate objects in images



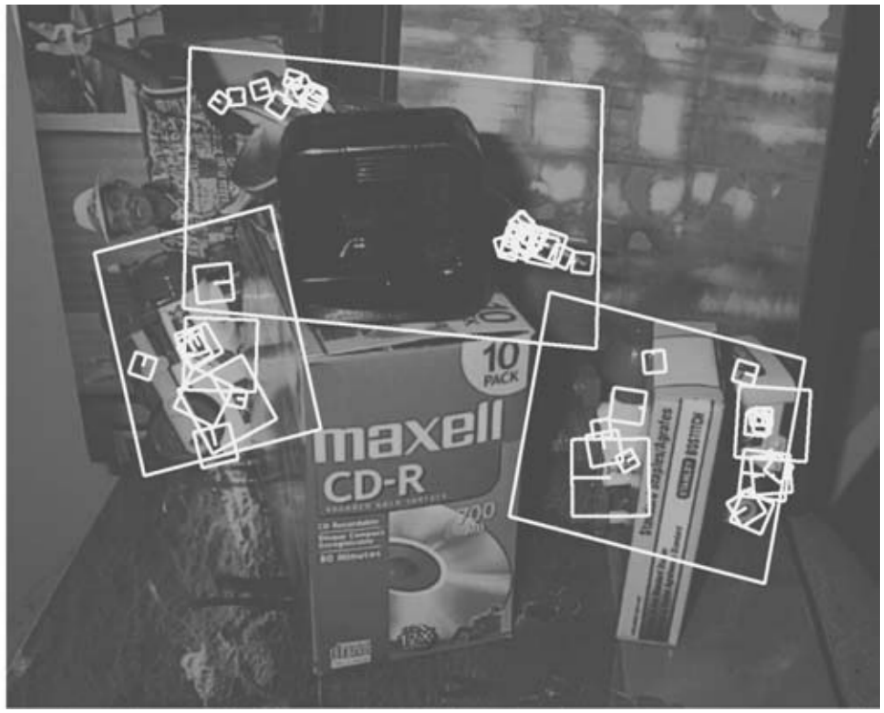
Variants: Pedestrian Detection

Leibe et al., 2005

Variants: Face Detection



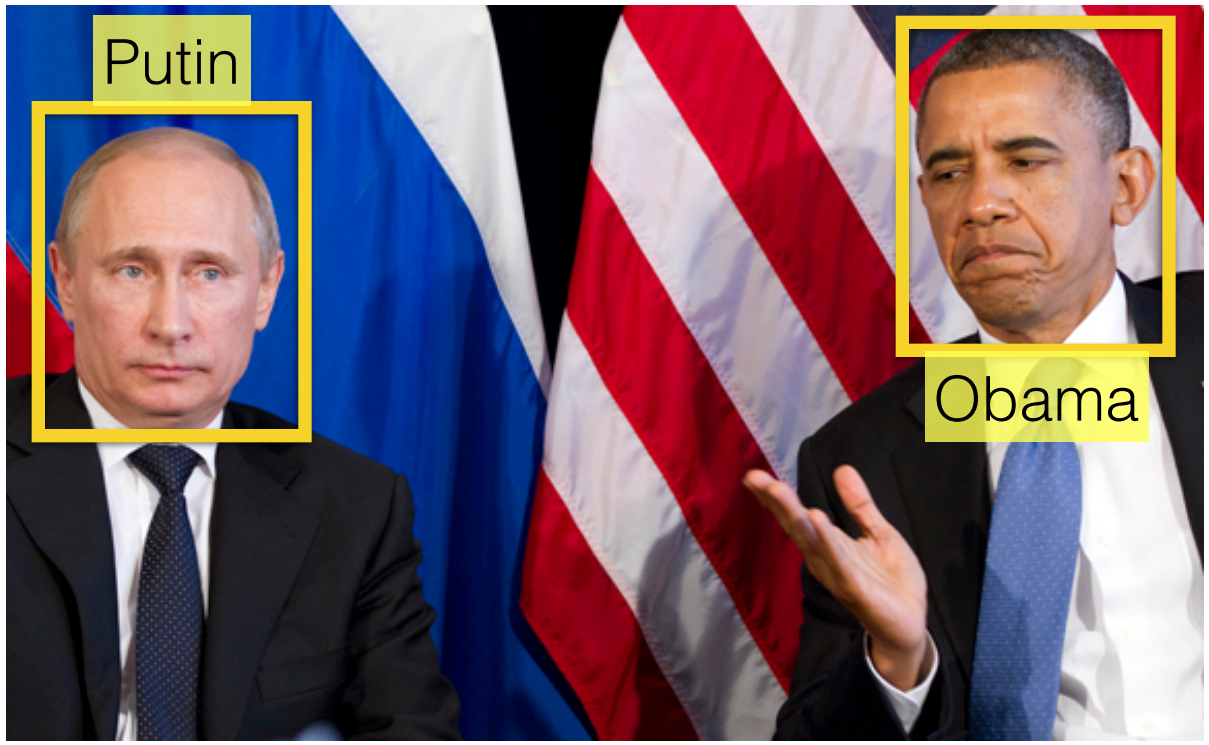
Variants: Instance Detection



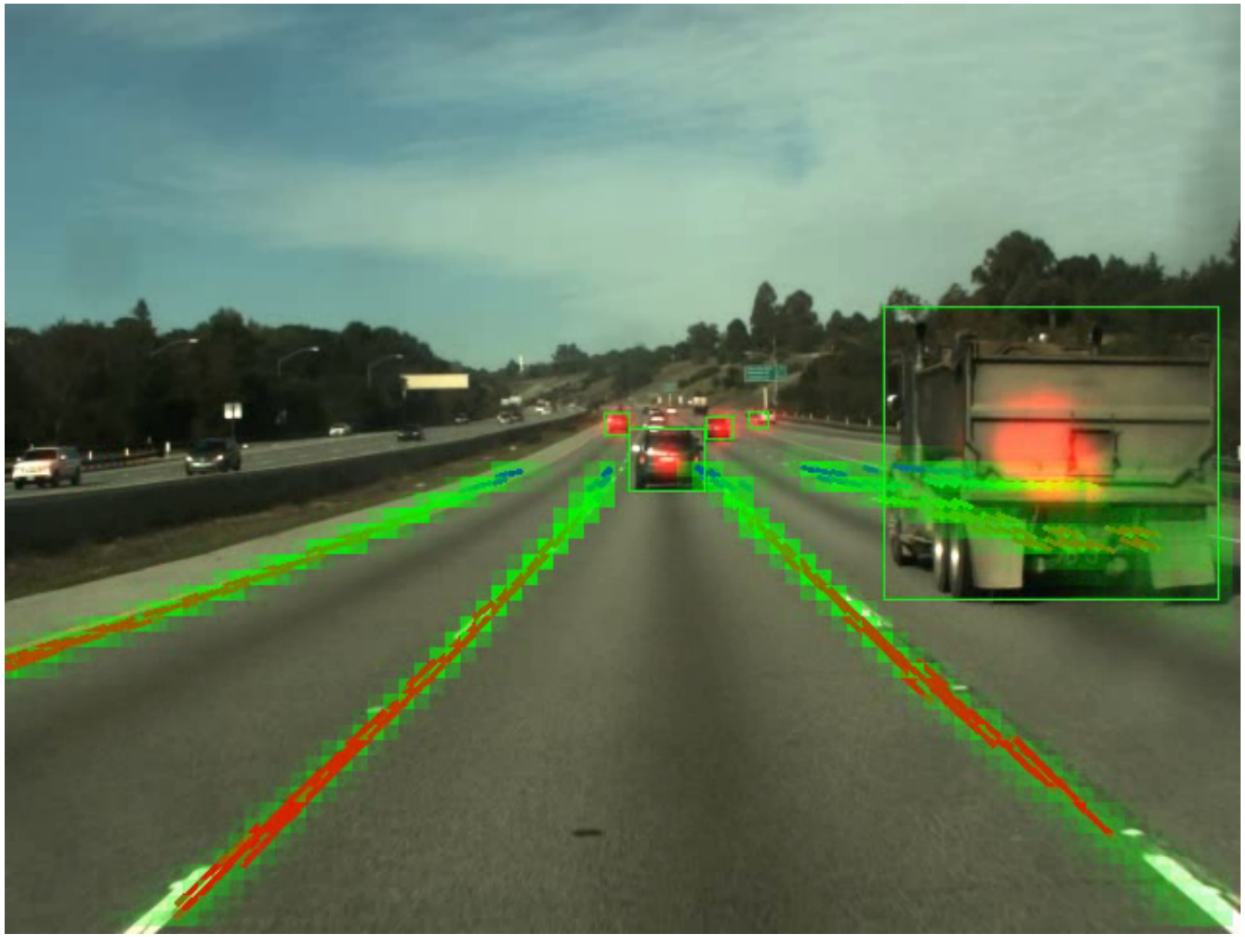
Lowe 2004

Variants: Multi-Class Detection

Application: Tagging People

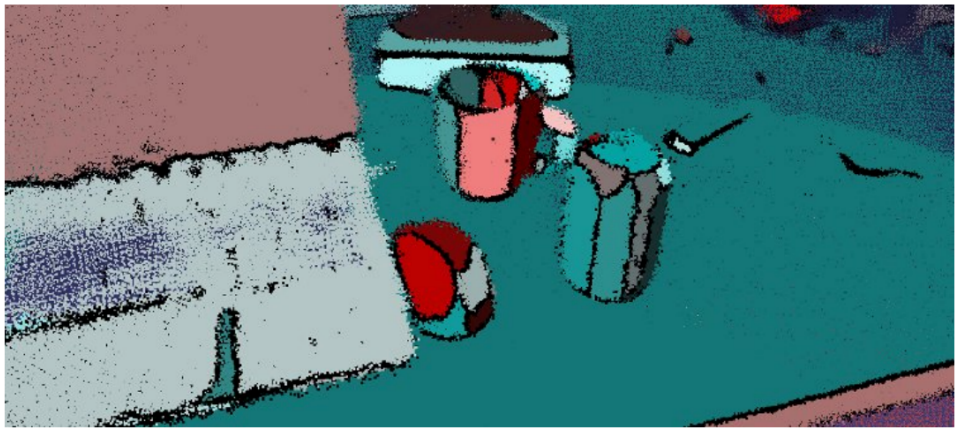


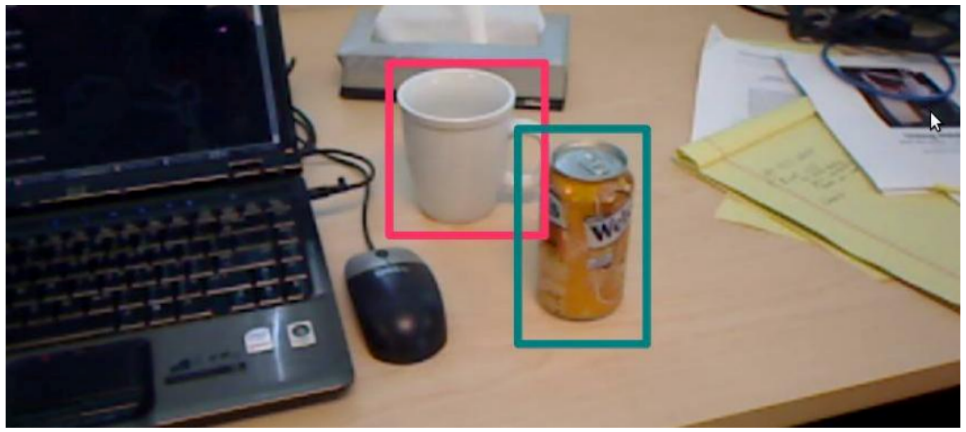
Application: Autonomous Driving



Huval et al., 2015

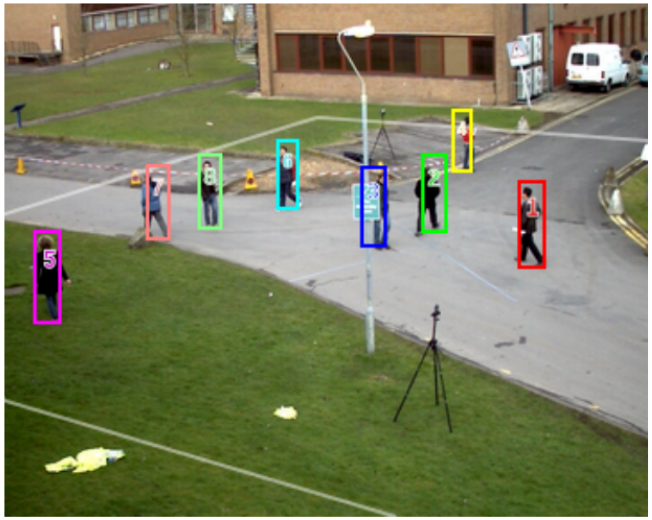
Application: Robotics

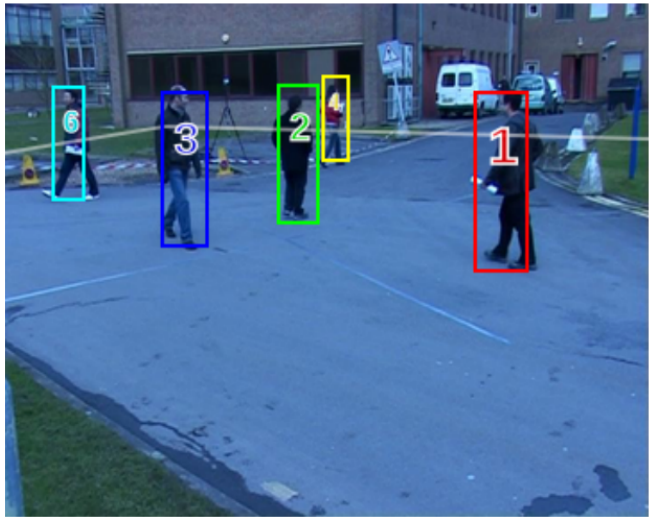




Lai et al., 2012

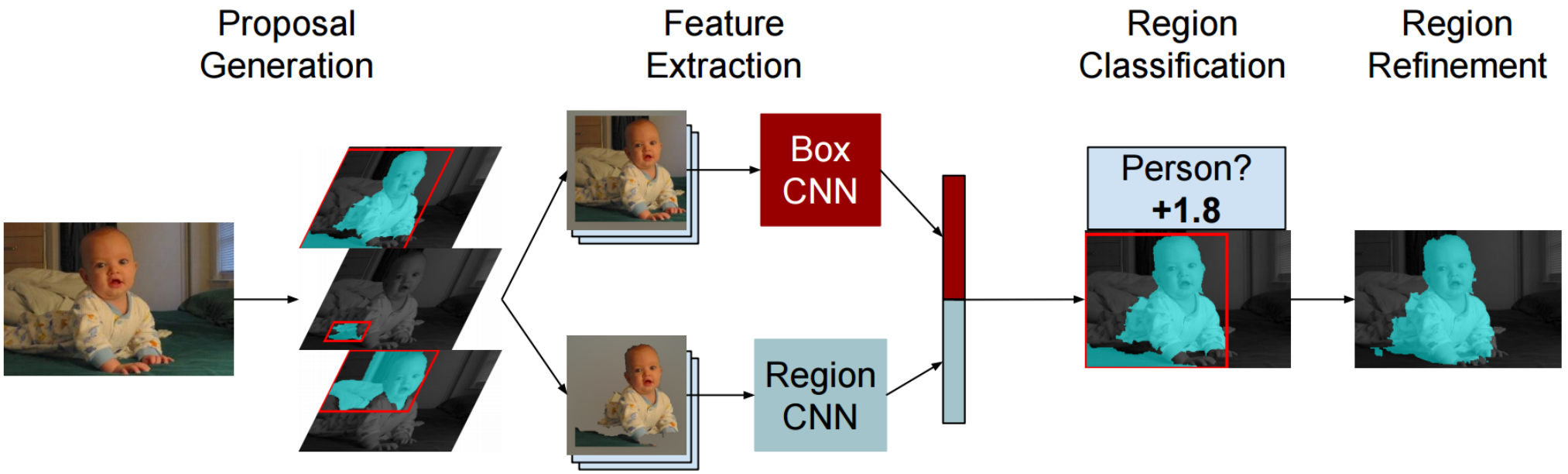
Application: Tracking





Berclaz et al., 2011

Application: Segmentation



Hariharan et al., 2014

Outline

1. Sliding Window Methods
2. Region-based Methods
3. Extra Topics

Outline

1. Sliding Window Methods
    1. Overview
    2. Viola-Jones Face Detection
    3. HOG
    4. Exemplar SVM
    5. DPM
2. Region-based Methods
3. Extra Topics

Getting Started: Kitten Detection

Goal: Detect all kittens

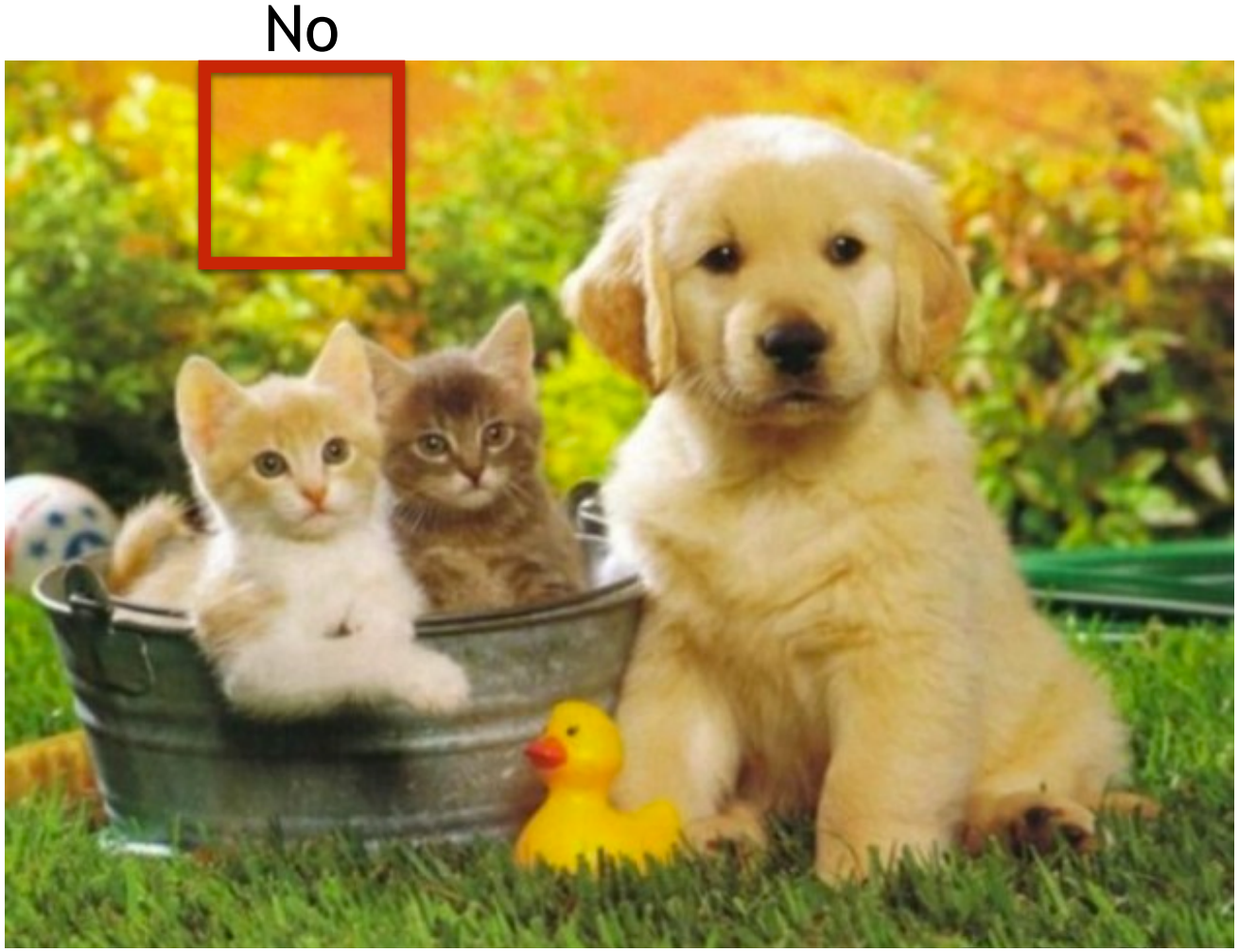
Checking Windows for Kittens

No

Checking Windows for Kittens

Checking Windows for Kittens

Checking Windows for Kittens

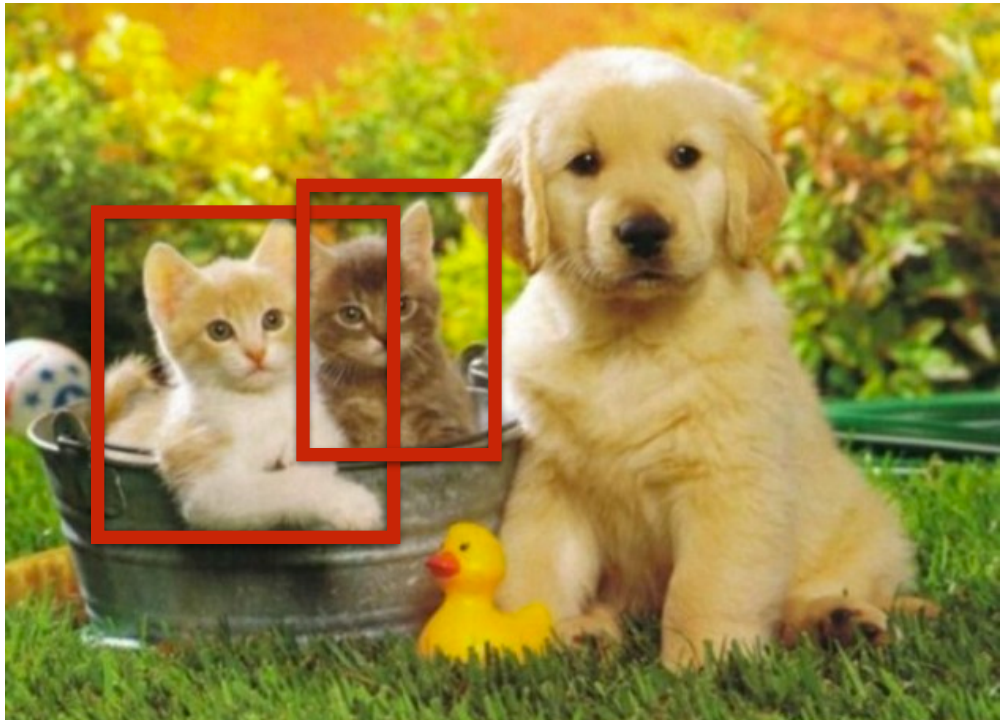


Sliding Windows

Evaluate every bounding box position

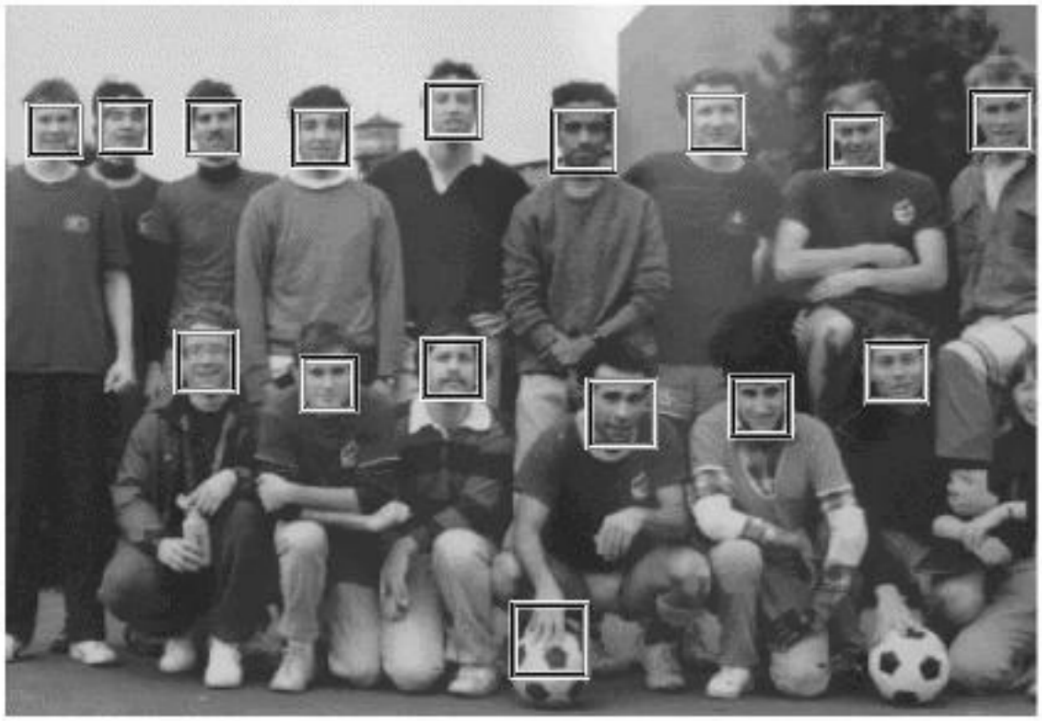
Aspect Ratio and Scale

- Even if we search all 2d positions, still don't know *aspect ratio* or *scale*.



- Solution: Multiple aspect ratios and multi-scale

Viola Jones Face Detector

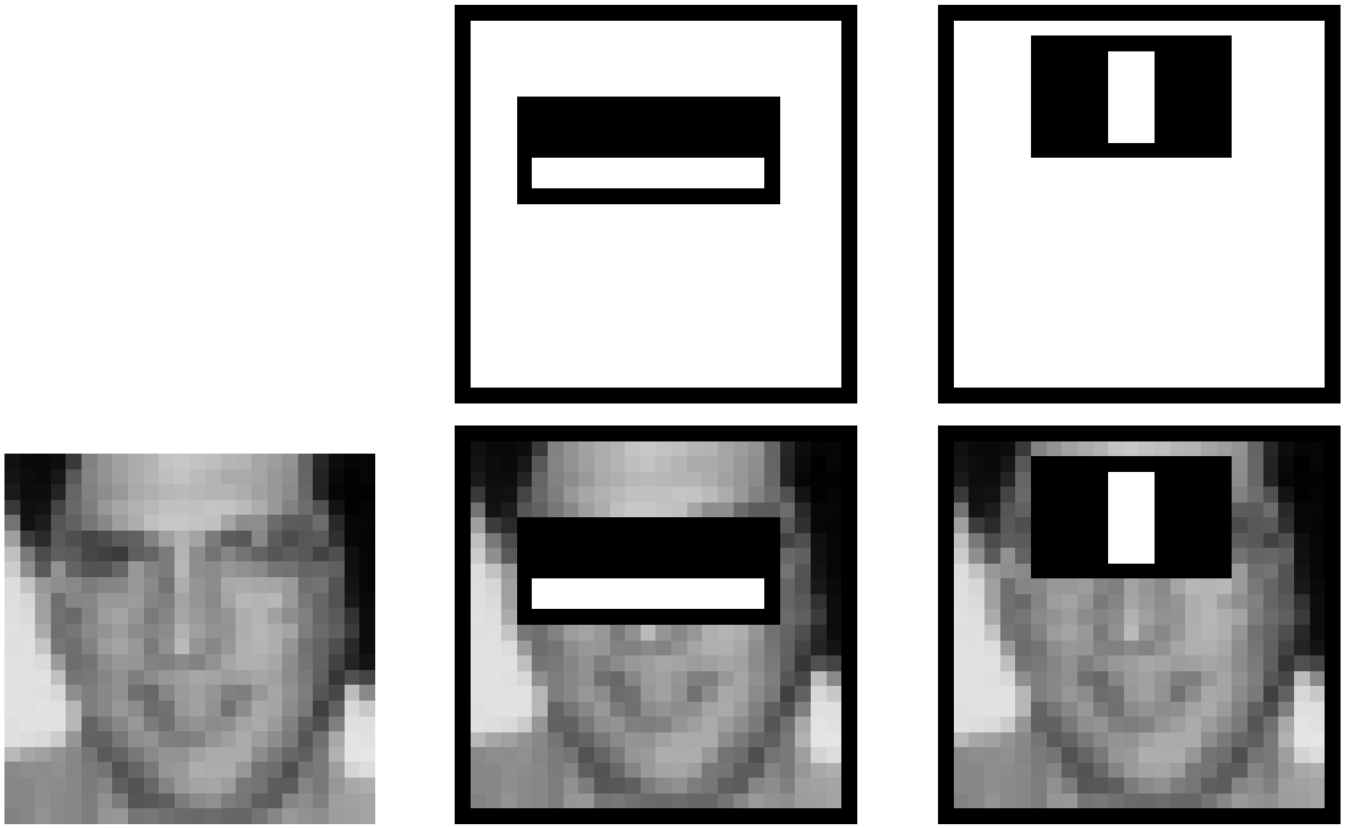


- Extremely fast
- Very accurate (at the time)

Viola, Jones. 2001

Viola Jones

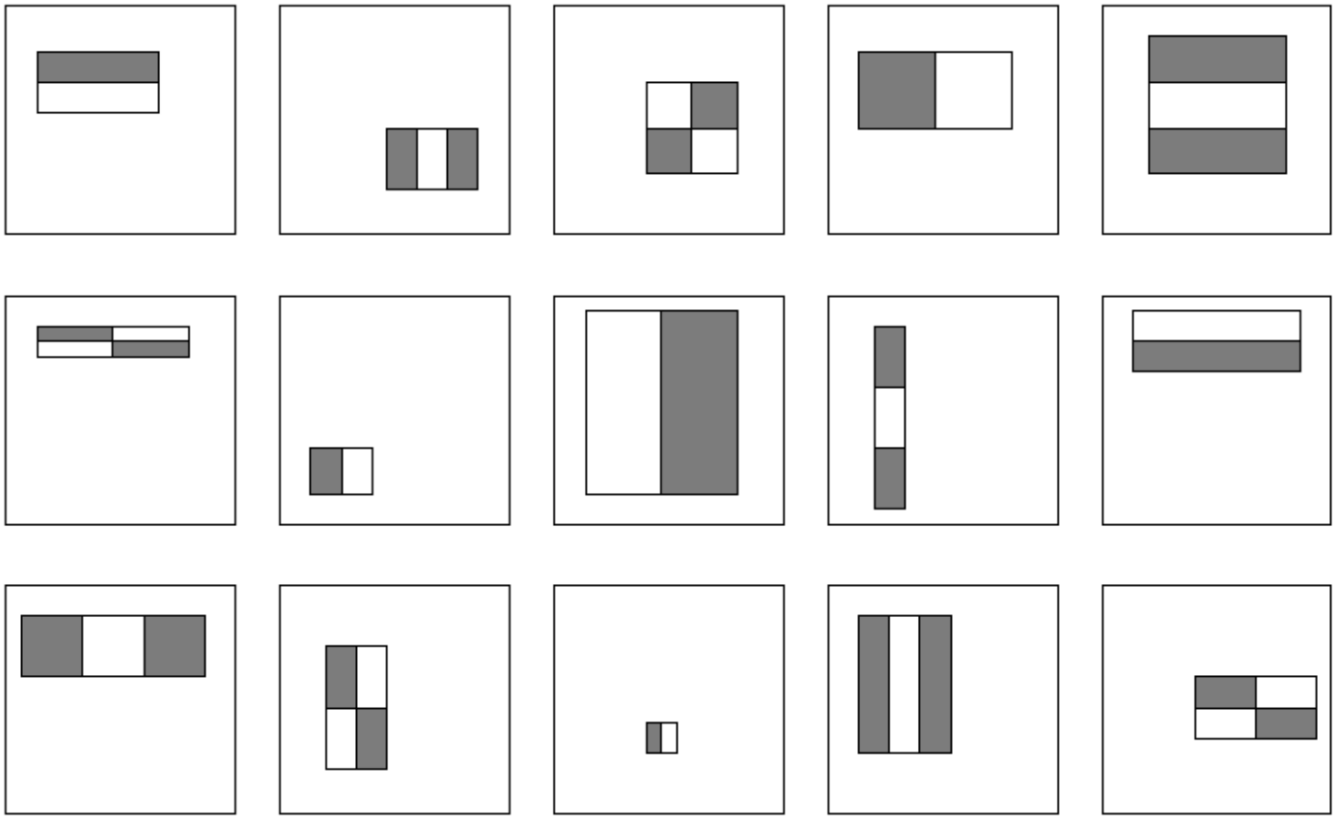
Key Idea: Boosting on weak classifiers



Viola, Jones. 2001

Haar Filters

Simple patterns of lightness and darkness



Viola, Jones. 2001

Haar Filters w/Integral Images

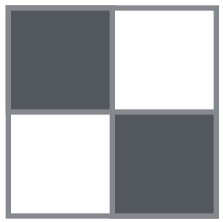
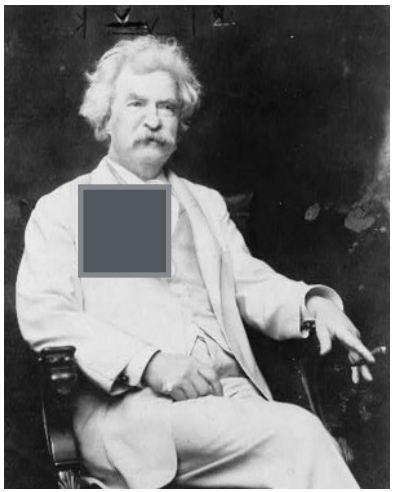
Filter: 

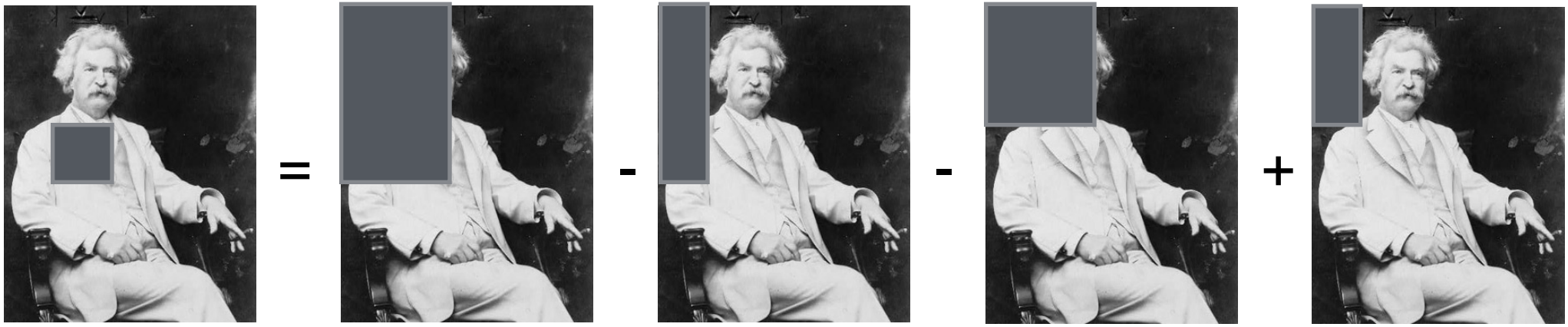
Image: 

Decomposition: smaller filters 

Haar Filters w/Integral Images

Response at a single location:

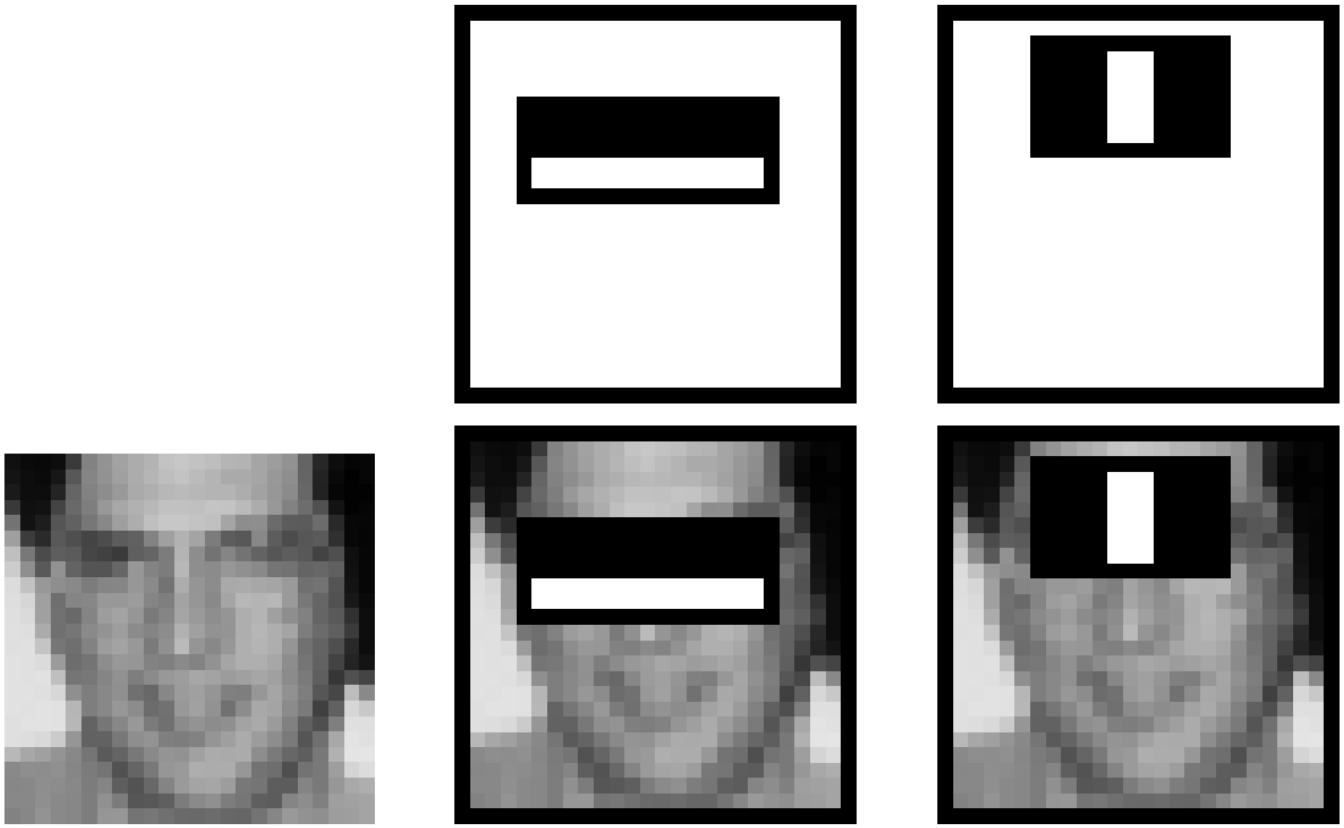




Only need to compute sum of top-left responses (DP)!

Viola Jones: Weak Classifiers

Each Haar filter is a weak classifier



Top classifier

Second best

Viola, Jones. 2001

Combining Weak Classifiers

AdaBoost:

$h_t(x)$: binary classifier on Haar filter *t*

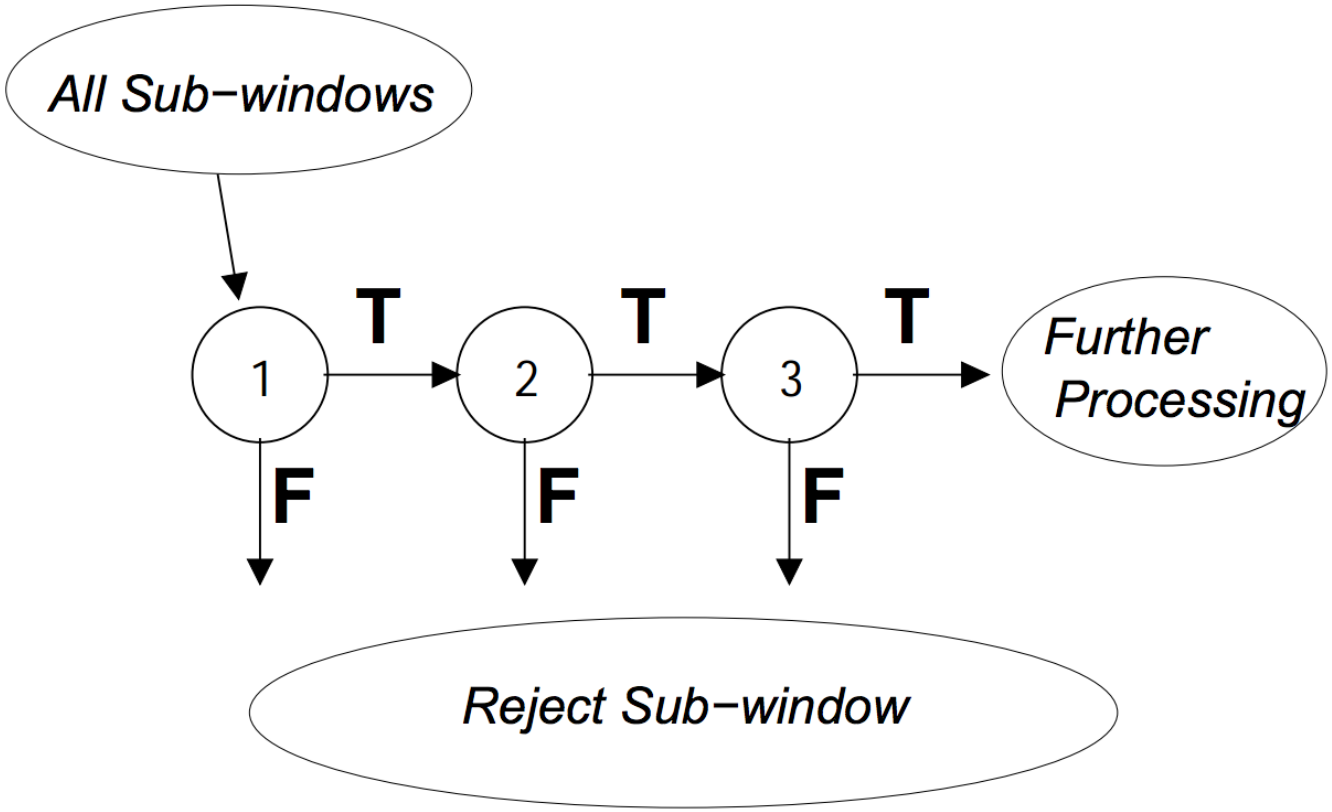
α_t : learned weight on classifier *t*

AdaBoost classifier: $h(x) = \left[\sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T} \alpha_t\right]$

minimizes loss: $\sum_{i=1}^{N} e^{-y_i h(x_i)}$

Viola, Jones. 2001

Cascade

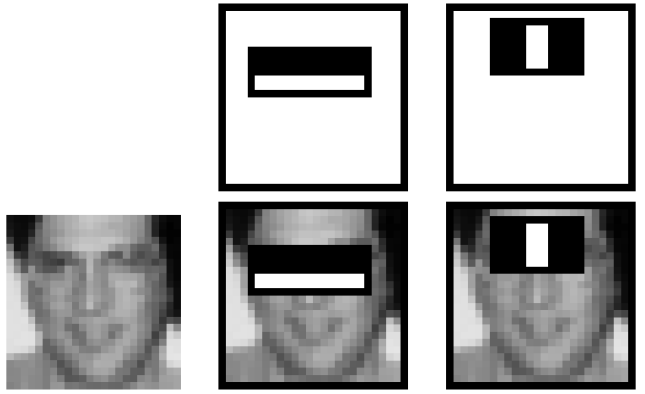


Reject negatives quickly

Viola, Jones. 2001

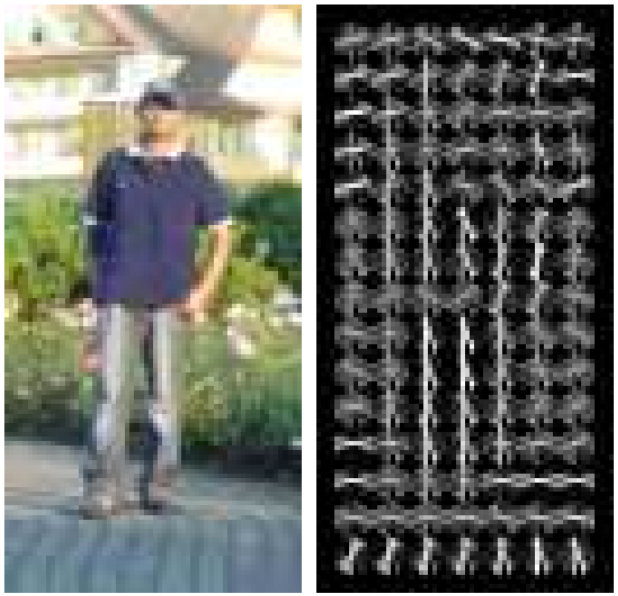
Viola Jones Summary

- Fast at runtime
- Takes a long time to train
- Very accurate (at the time)
- Inspired other detection methods



HOG

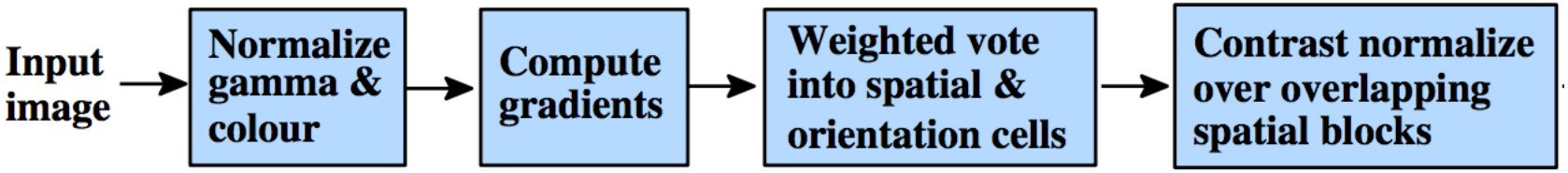
- Histograms of Oriented Gradients
- Designed for Pedestrian Detection
- Really just good feature engineering

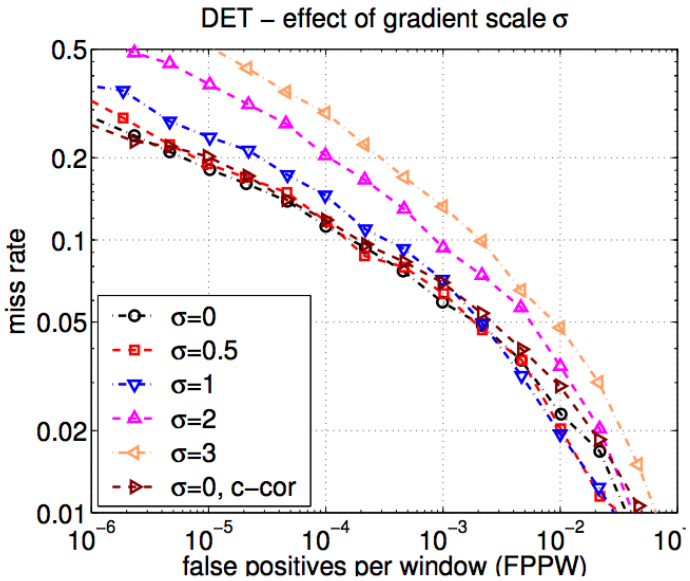


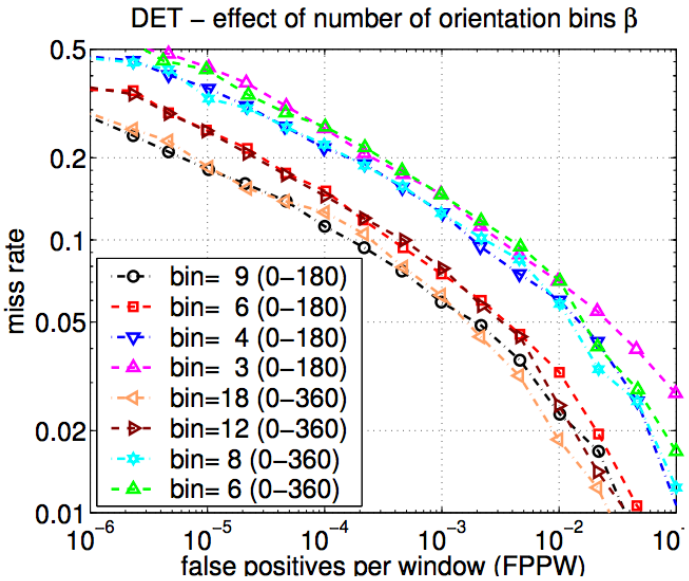
Dalal, Triggs. 2005

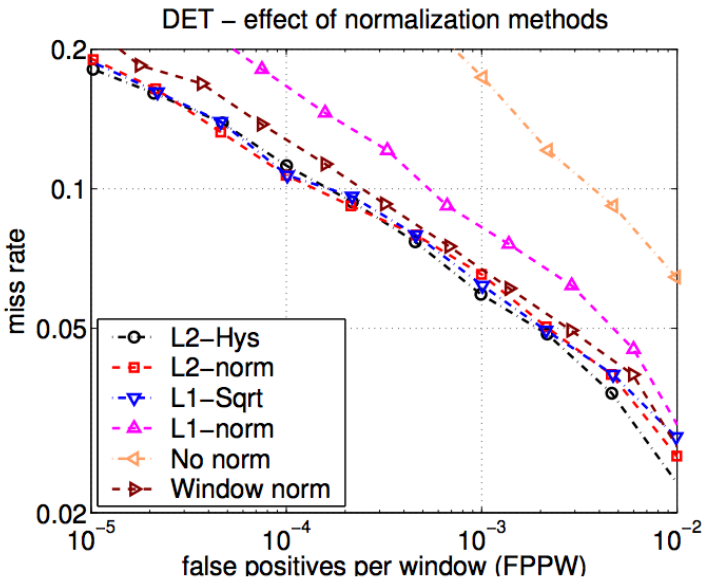
HOG

- Lots of feature engineering…



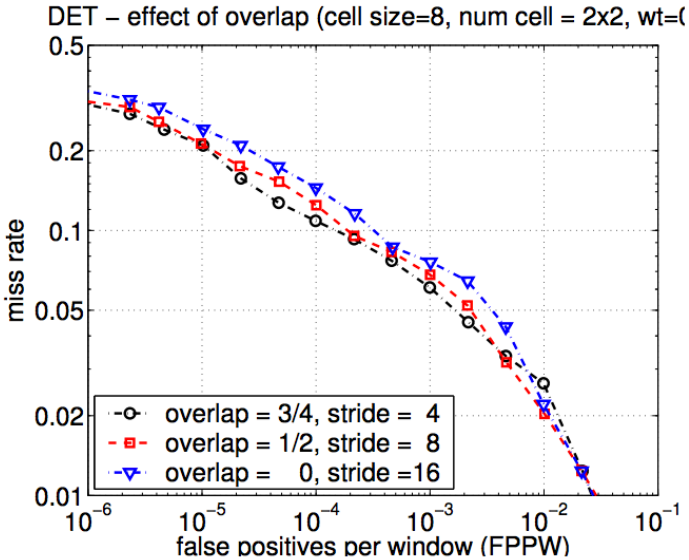


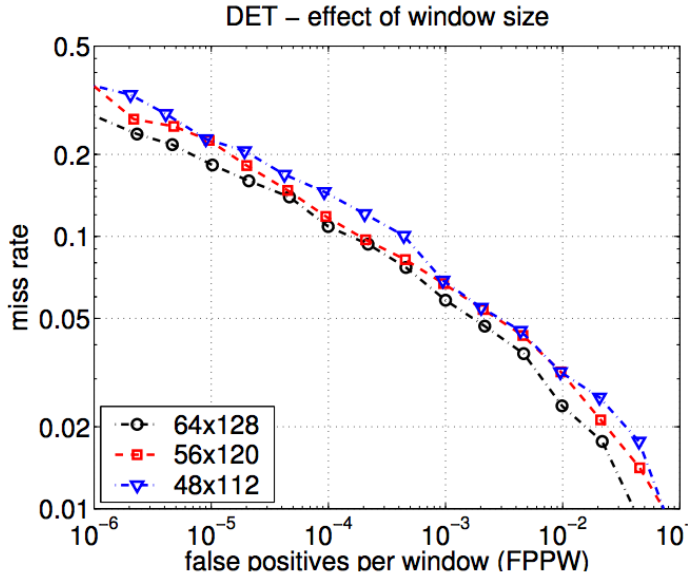


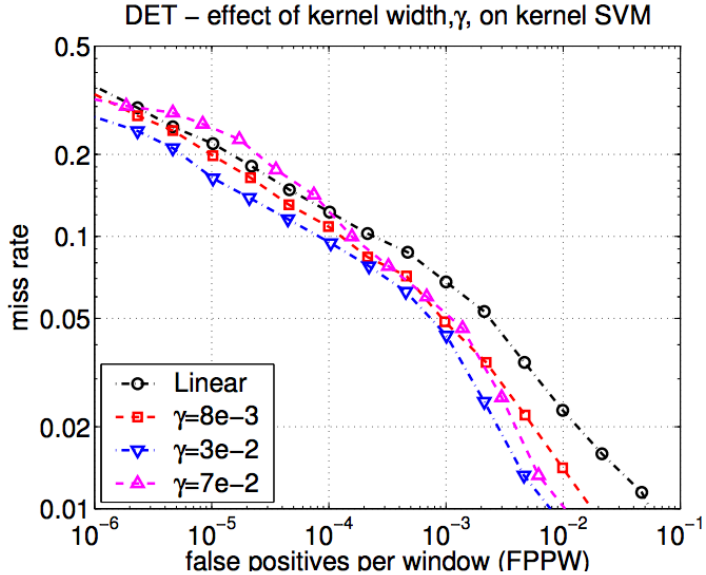


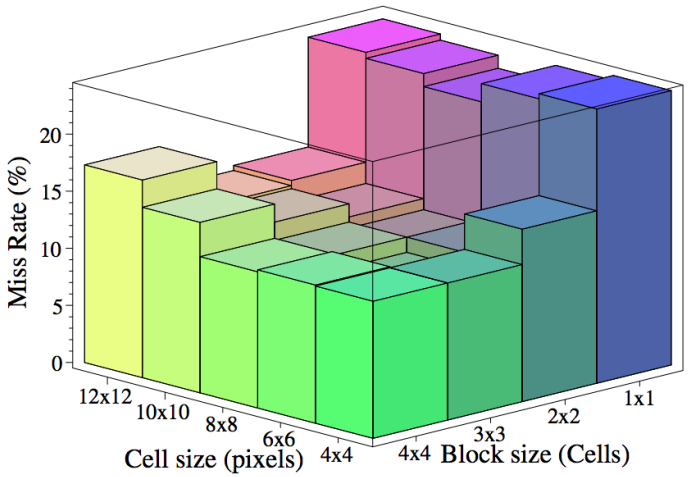
Dalal, Triggs. 2005

More feature engineering



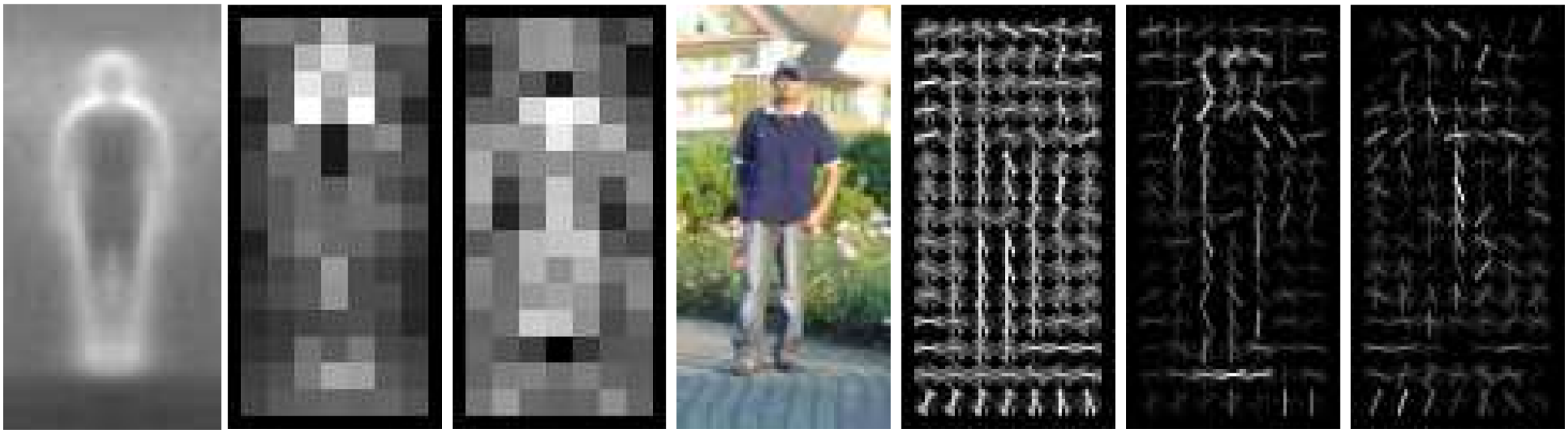






Dalal, Triggs. 2005

But it works

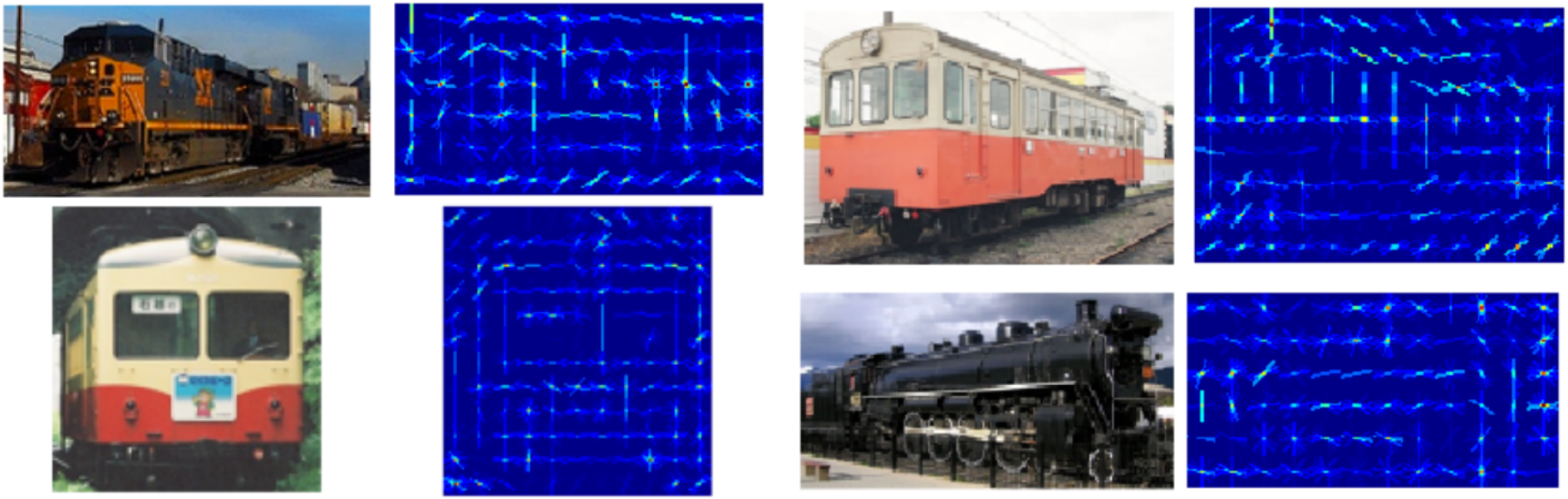


avg. gradient | max pos. SVM weight | min neg. SVM weight | HOG | pos SVM weights | neg SVM weights

Dalal, Triggs. 2005

Exemplar SVM

- Key idea: Train a separate SVM for each positive training example (on HOG features!).



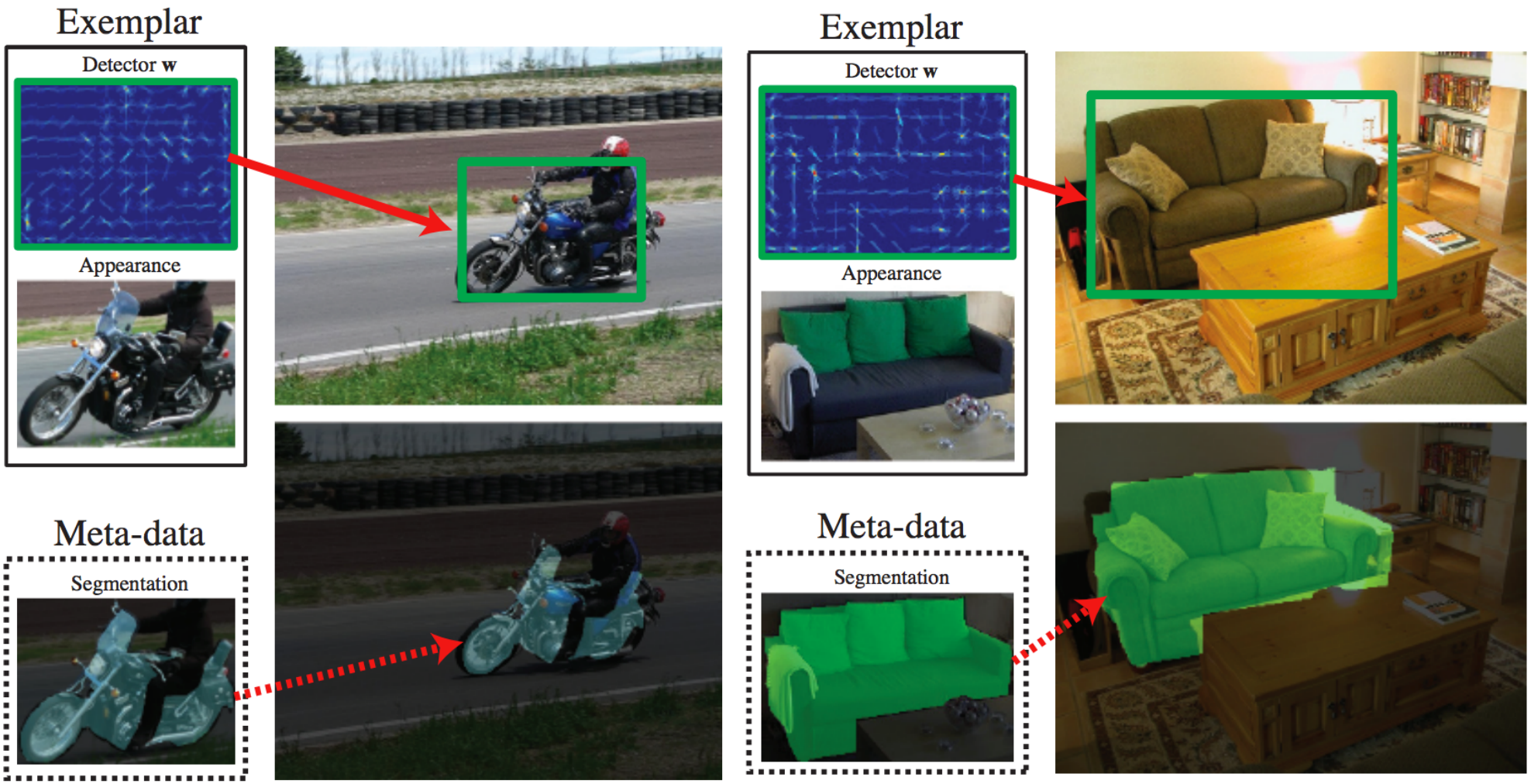
Malisiewicz et al. 2011

Exemplar SVM

- Q: But wait, isn't that going to be horribly slow?
- A: Yep! Much slower than a single SVM. No one I know of actually uses this. However….
- Can transfer metadata (segmentations!)

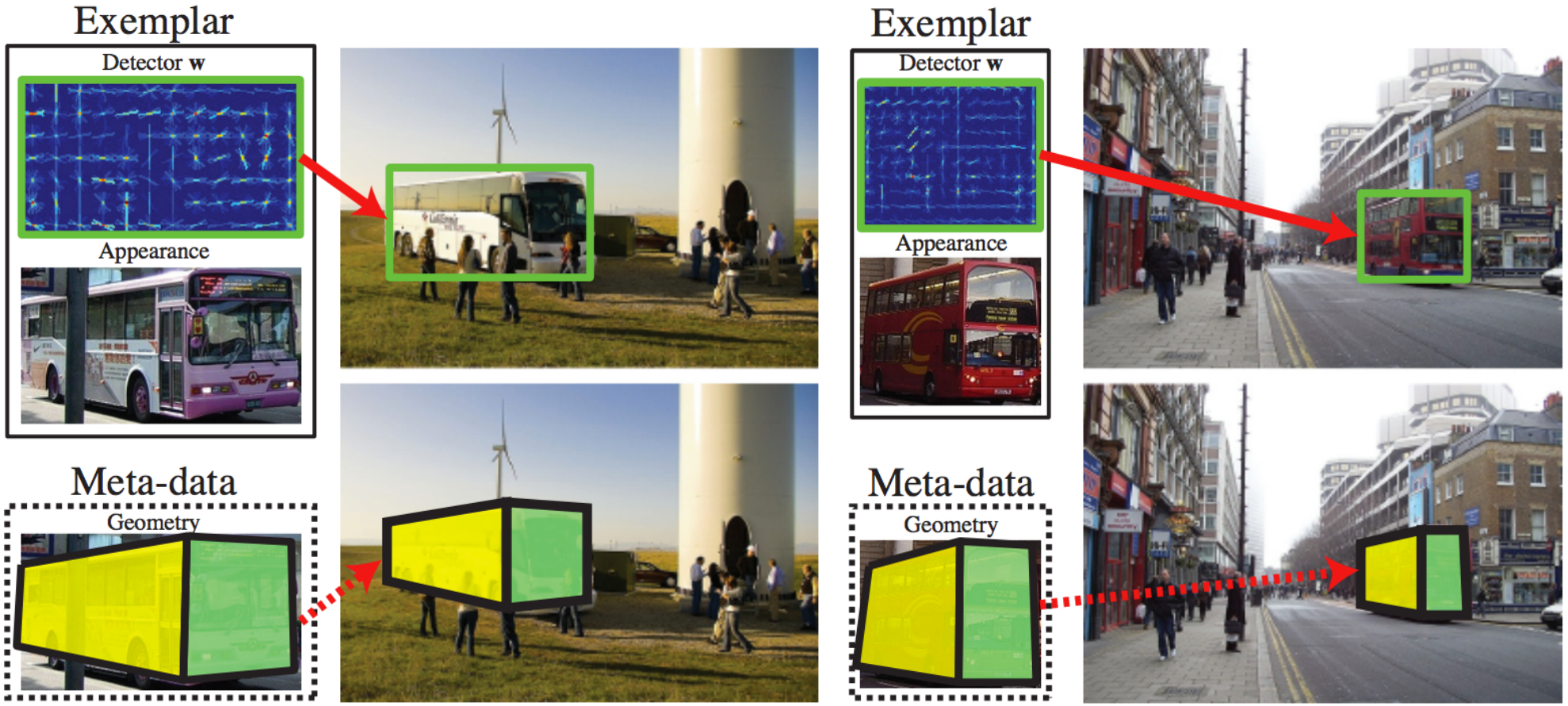
Malisiewicz et al. 2011

Exemplar SVM Examples



Malisiewicz et al. 2011

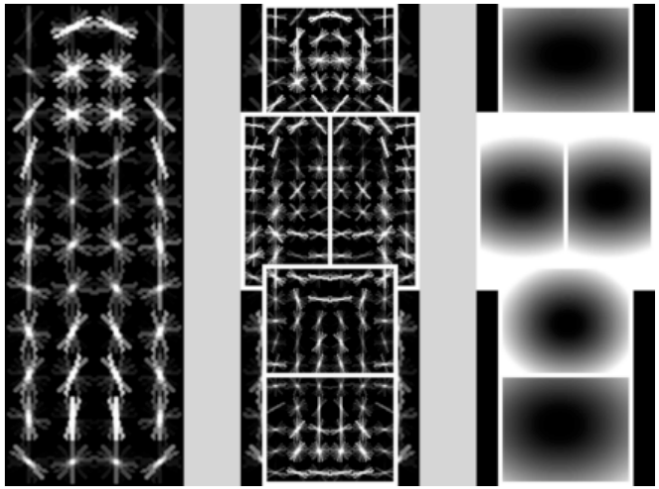
Exemplar SVM Examples



Malisiewicz et al. 2011

Deformable Part Models

- (sneak preview of student presentation)
- Similar to SVM on HOG, but also with parts (latent SVM)
- State of the art for several years



Sliding Window Summary

- Evaluate classifier at many positions
- Dominant detection paradigm until ~2 years ago
- Boosting, SVM, and DPM

Outline

1. Sliding Window Methods
2. Region-based Methods
    1. Motivation
    2. Region Proposals
    3. R-CNN
3. Extra Topics

Sliding Window Problem: Efficiency

Q: How many bounding boxes in this 482 x 348 image?

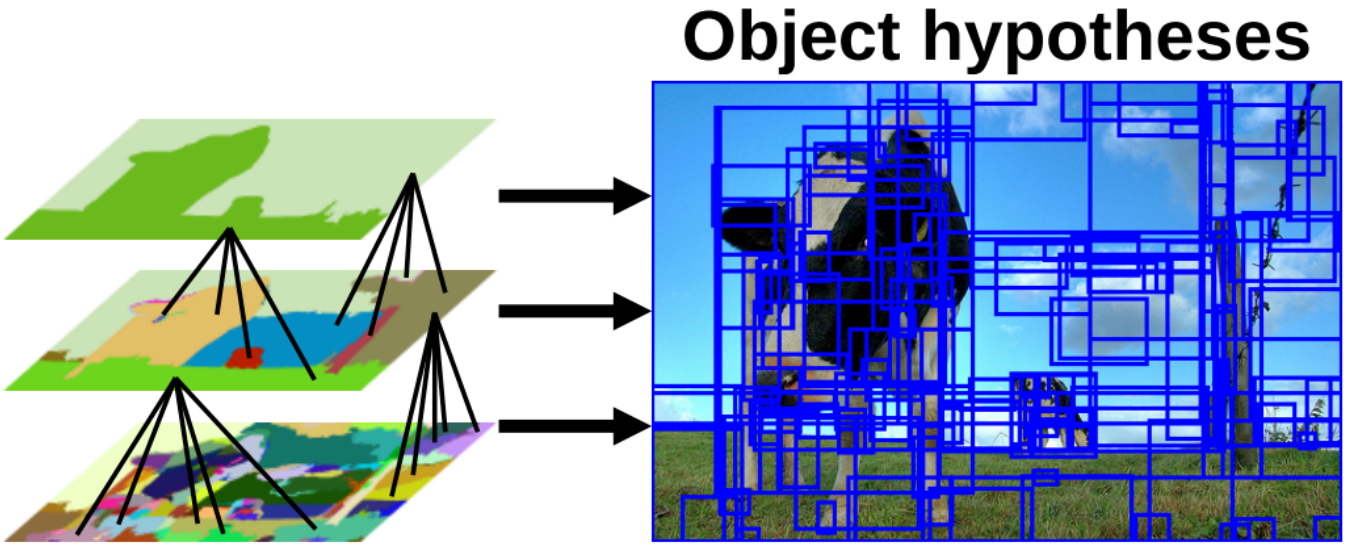
A: 6,999,078,138 (7 trillion)

Sliding Window Problem: Efficiency

Can’t classify 7 trillion windows, even millions is slow.

Can we massively cut down this number (e.g. 1000s)?

Detection on Regions



- Generate detection *proposals* (typically ~2000)
- Classify each region with a much stronger classifier
- More or less taken over modern detection

van de Sande et al., 2011

Region Proposals

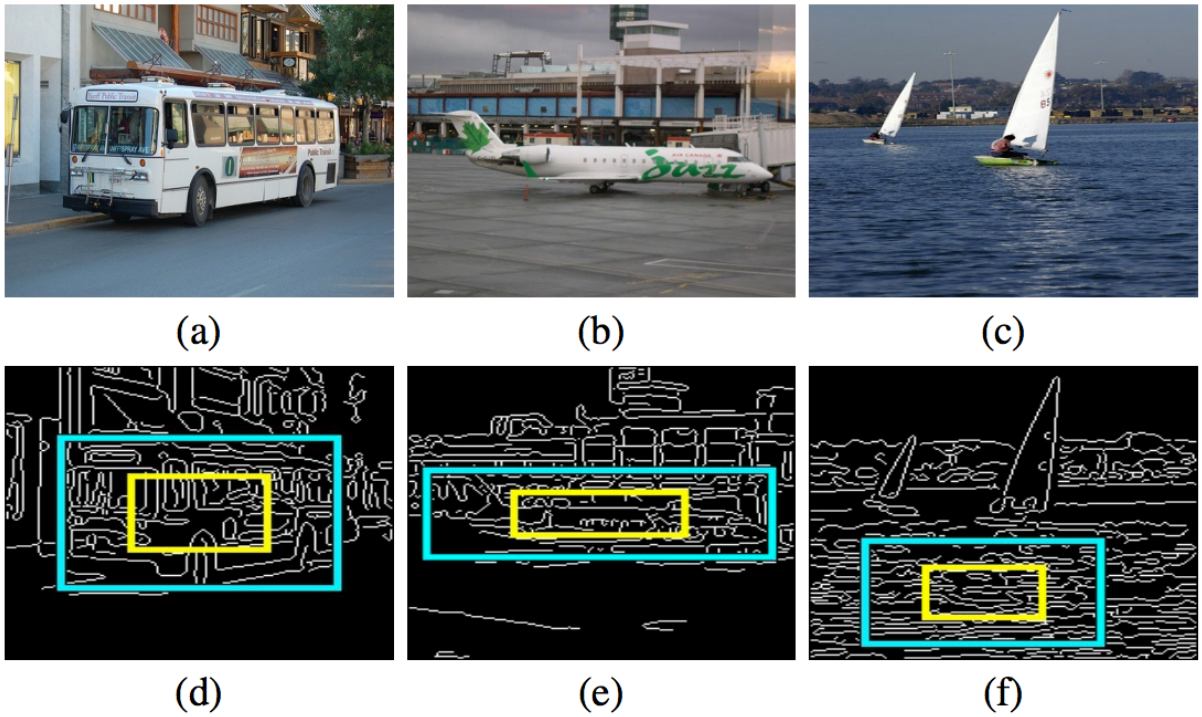
- Sliding window or grouping pixels
- May or may not output score
- Varying amount of control over number of regions

Method	Approach	Outputs Segments	Outputs Score	Control #proposals	Time (sec.)	Repea-tability	Recall Results	Detection Results
Bing [16]	Window scoring		✓	✓	0.2	⋆⋆⋆	⋆	·
CPMC [17]	Grouping	✓	✓		250	-	⋆⋆	⋆
EdgeBoxes [18]	Window scoring		✓	✓	0.3	⋆⋆	⋆⋆⋆	⋆⋆
Endres [19]	Grouping	✓	✓	✓	100	-	⋆⋆	⋆⋆
Geodesic [20]	Grouping	✓		✓	1	⋆	⋆⋆⋆	⋆⋆
MCG [21]	Grouping	✓	✓		30	⋆	⋆⋆⋆	⋆⋆
Objectness [22]	Window scoring		✓	✓	3	·	⋆	·
Rahtu [23]	Window scoring		✓	✓	3	·	·	⋆
RandomizedPrim's [24]	Grouping	✓		✓	1	⋆	⋆	⋆
Rantalankila [25]	Grouping	✓		✓	10	⋆⋆	·	⋆
Rigor [26]	Grouping	✓		✓	10	⋆	⋆⋆	⋆⋆
SelectiveSearch [27]	Grouping	✓	✓	✓	10	⋆⋆	⋆⋆⋆	⋆⋆
Gaussian				✓	0	·	·	⋆
SlidingWindow				✓	0	⋆⋆⋆	·	·
Superpixels		✓			1	⋆	·	·
Uniform				✓	0	·	·	·

"What makes for effective detection proposals?". Hosang, Benenson, Dollar, Schiele. 2015

Objectness

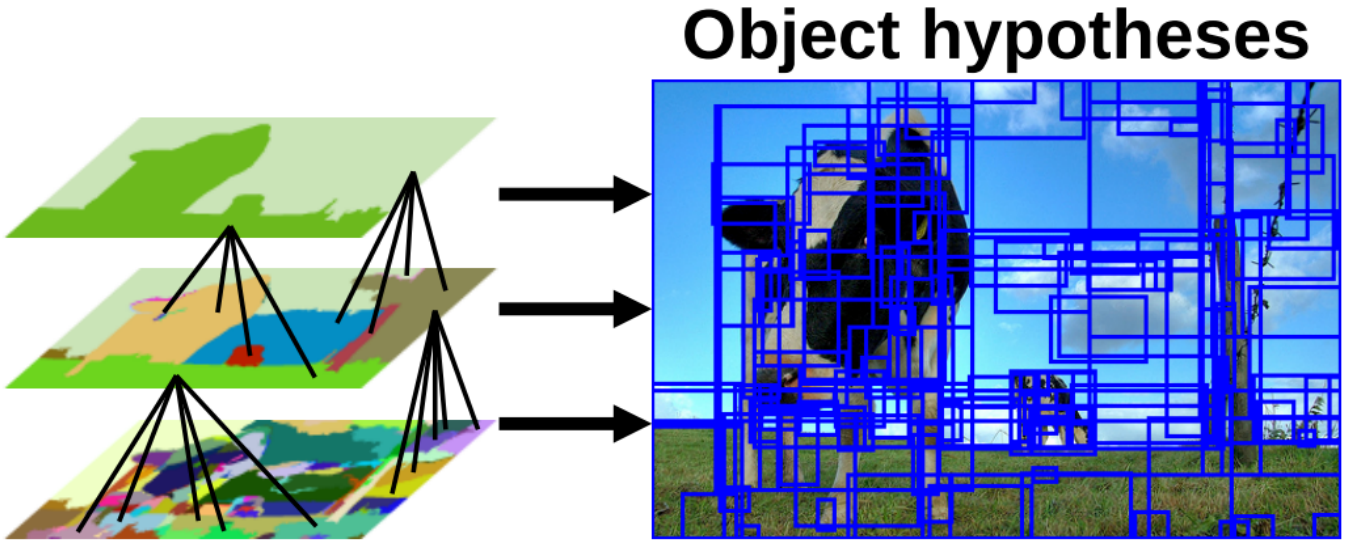
- Sliding window
- Score based on a bunch of heuristic features



Alexe, Deselares, Ferrari. 2010

Selective Search

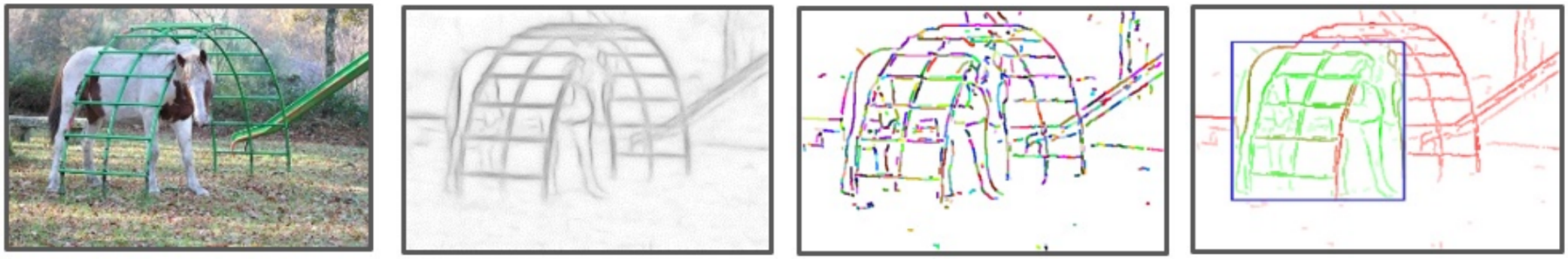
- Felzenszwalb superpixels
- Merge based on color features
- Most common method in use



van de Sande et al., 2011

Edge Boxes

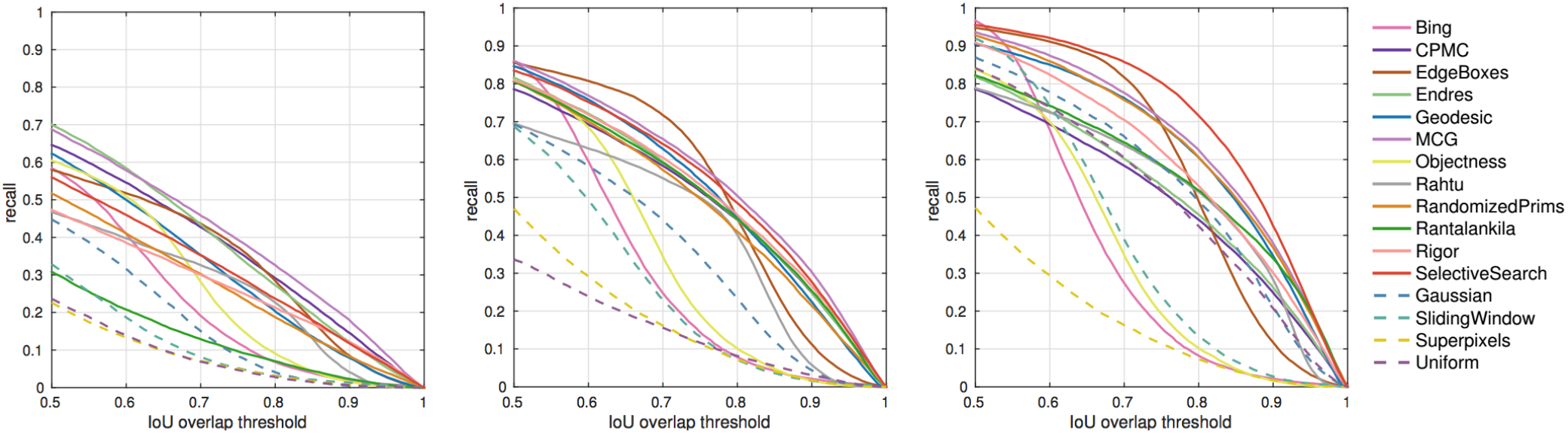
- Structured decision forest for object boundaries
- Coarse sliding windows with location refinement
- Seems fast and accurate, but time will tell



Zitnick, Dollar. 2014

Evaluating Region Proposals

- What fraction of ground truth bounding boxes do they recover?
- How many proposals does it take?
- At what IoU overlap threshold?



(a) 100 proposals per image. (b) 1 000 proposals per image. (c) 10 000 proposals per image.

"What makes for effective detection proposals?". Hosang, Benenson, Dollar, Schiele. 2015

In Practice

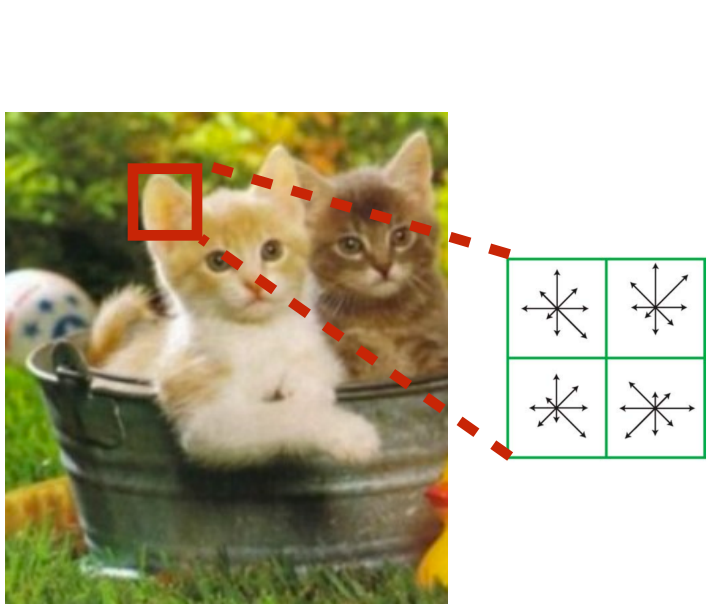
- Recall at IoU threshold=0.7 predicts detection performance well
- Most people use ~2000 regions produced with Selective Search (a few seconds/image)
- Edge Boxes looks promising

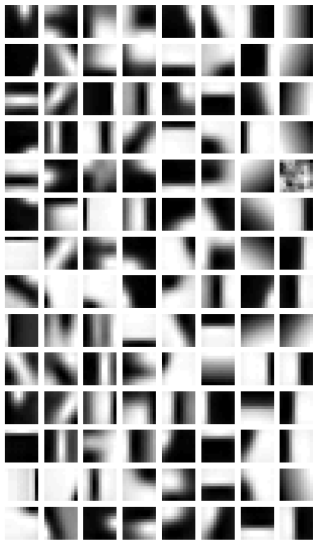
Aside: Classification

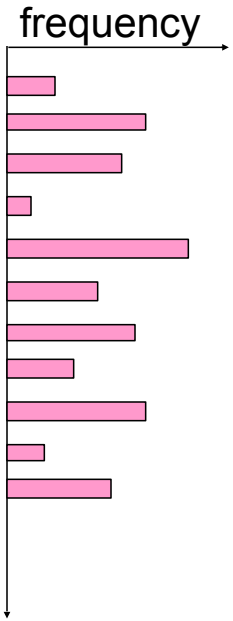
- Most detectors, region proposal methods in particular, reduce detection to repeated classification

- Let's take a look at a few key ideas in classification

Classification: Bag of Words





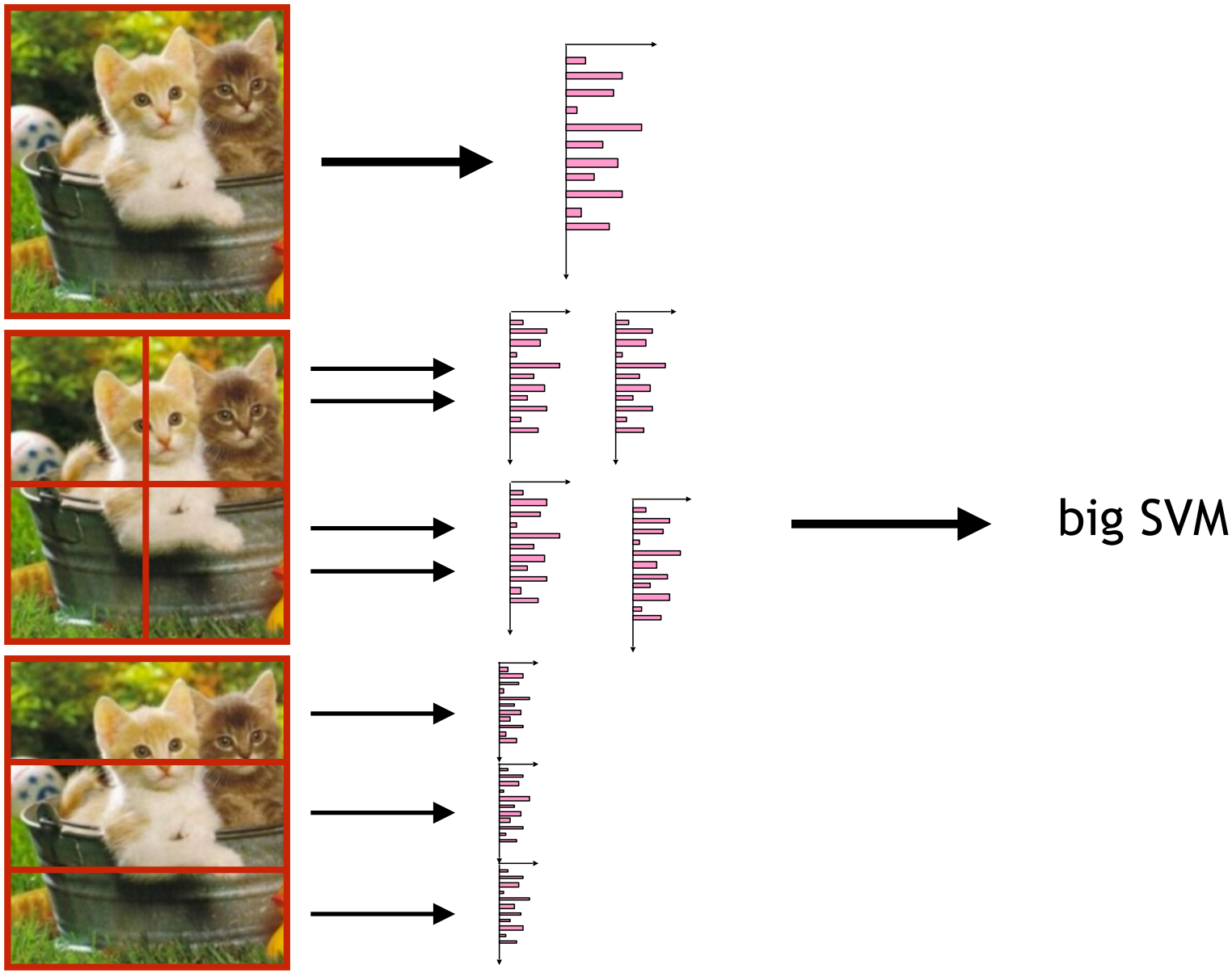


Descriptors → Codebook → Histogram → SVM

Offline: Cluster descriptors in training images

Note: No spatial information

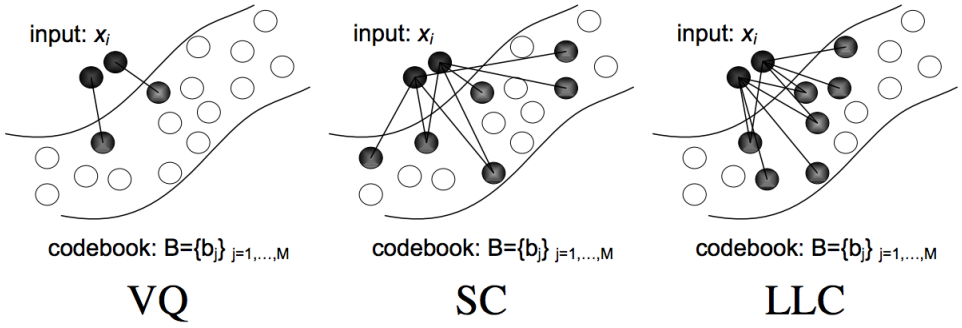
Classification: Spatial Pyramid



Lazebnik et al. 2006

Classification

- Sparse Coding (LLC: Locality constrained Linear Coding)
  - Represent descriptor with more than one codeword



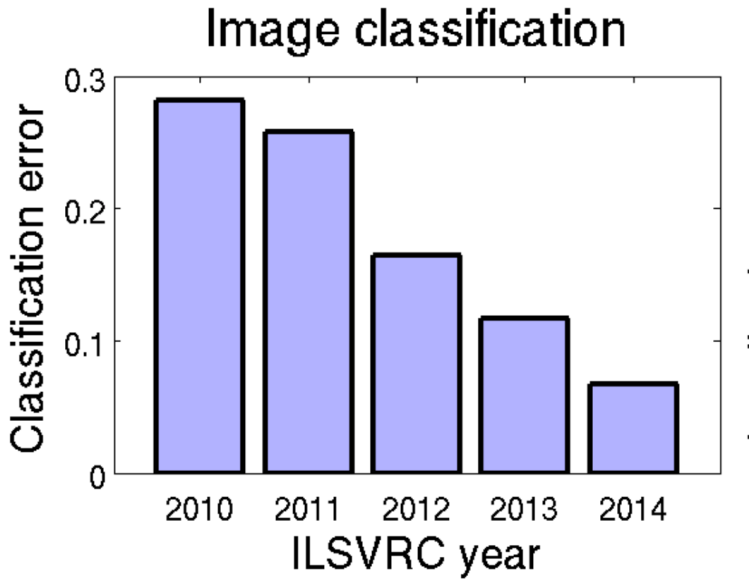
Wang et al. 2010

- Fisher Vectors
  - Represent difference between descriptor and codewords (very roughly)
  - A little better, still used sometimes

Perronnin et al. 2010

2012

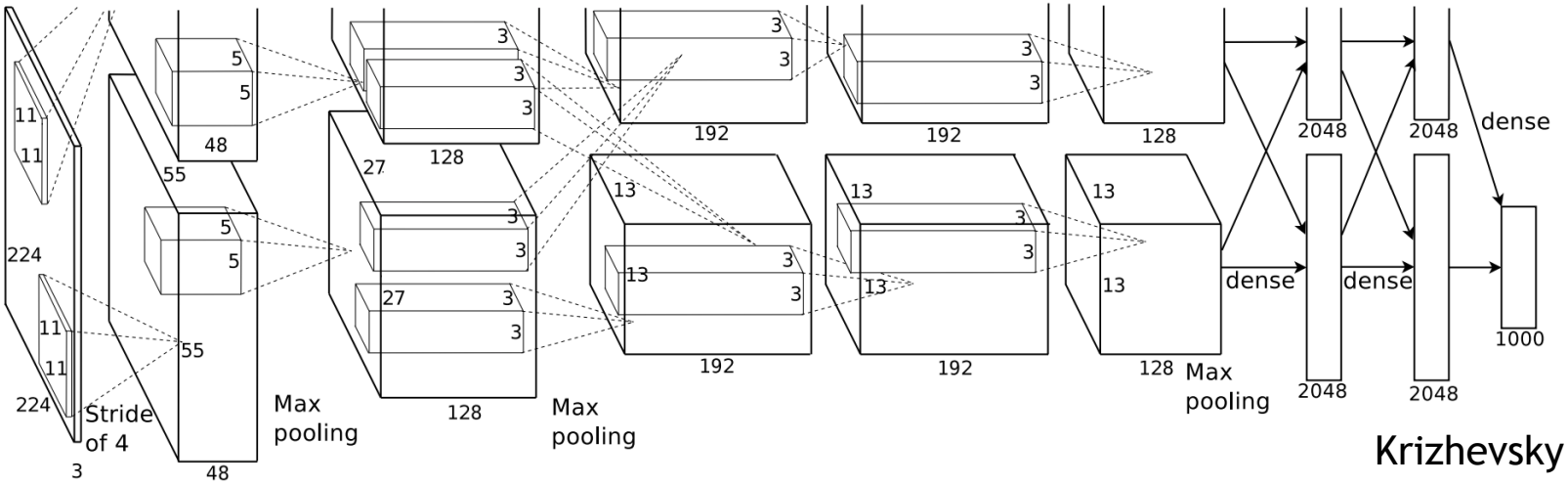
- In 2012 neural networks started working [Krizhevsky et al. 2012]



Russakovsky et al. 2015

Neural Nets

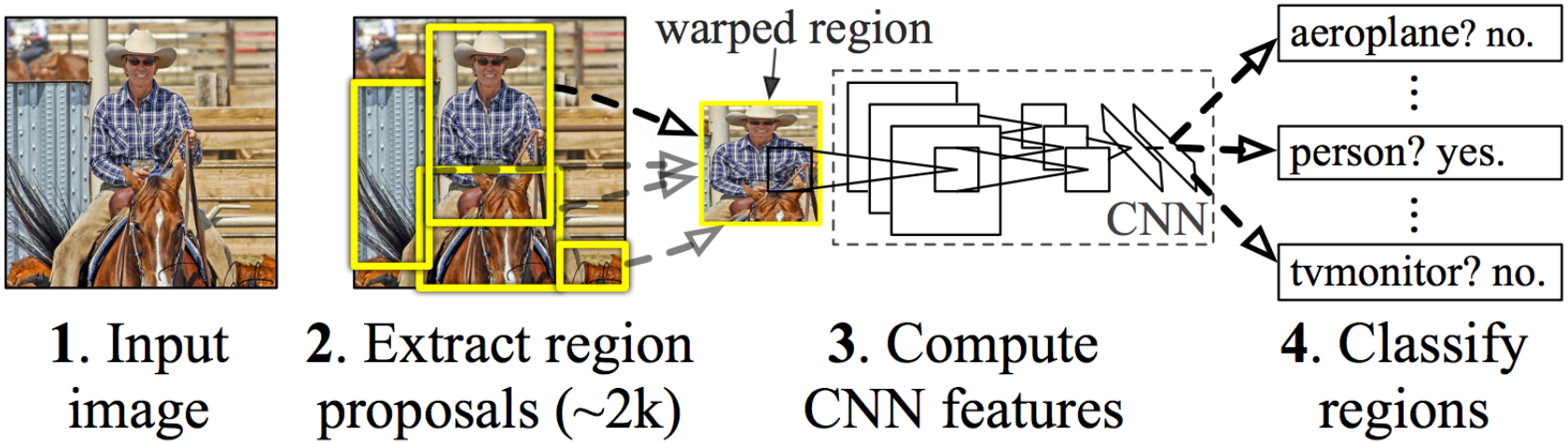
- Learn the whole pipeline (pixels to classes) from scratch.
- Many layers of (learned) intermediate features
- Will see more in student presentation



Krizhevsky et al. 2012

R-CNN

- R-CNN = Selective Search + CNN
- That's it.



Girshick et al. 2014

R-CNN Details

- Need region to fit input size of CNN
- Region warping method:

Girshick et al. 2014

R-CNN Details

- Context around region
- 0 or 16 pixels (in CNN reference frame)

Girshick et al. 2014

R-CNN Details

- CNN Layer is important
- fc_6 best?

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7

Girshick et al. 2014

R-CNN Details

- fine-tuning on PASCAL (CNN trained on ILSVRC)
- It helps, and may make another layer better

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT $pool_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc_6	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc_7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2

Girshick et al. 2014

R-CNN Details

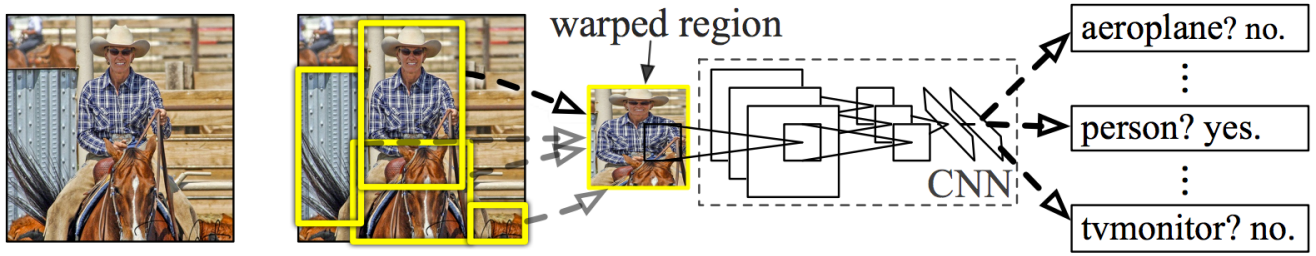
- Bounding box regression
- Regress from CNN features to bounding box
- Helps quite a bit

VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN $pool_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc_6	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc_7	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT $pool_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc_6	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc_7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc_7 BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**

Girshick et al. 2014

R-CNN Details

- Train SVM on top of CNN features
- Be careful about which are positives and which are negatives (use the IoU overlap!)
- Hard negative mining for efficiency.



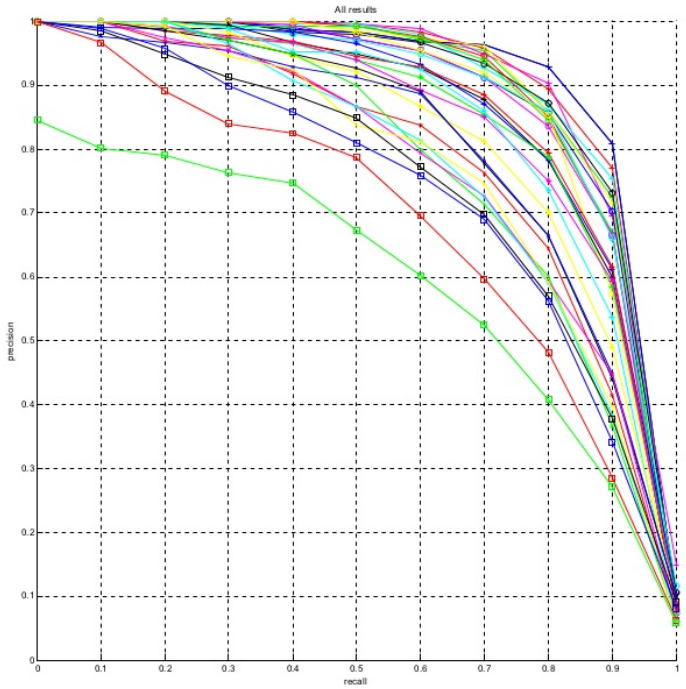
Girshick et al. 2014

Outline

1. Sliding Window Methods
2. Region-based Methods
3. Extra Topics

Evaluation

- Typically done with Average Precision (AP)
- When considering multiple classes, use mean (across classes) Average Precision (mAP)



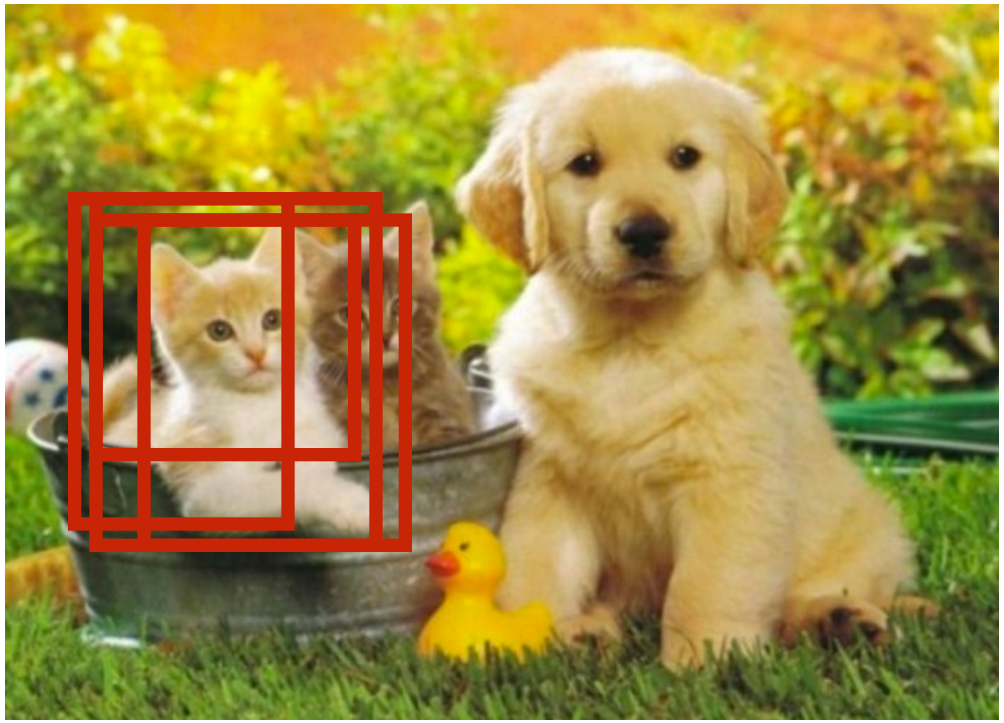
Context

- Surroundings can provide information
- Many methods use a weak version of this

What object is this?

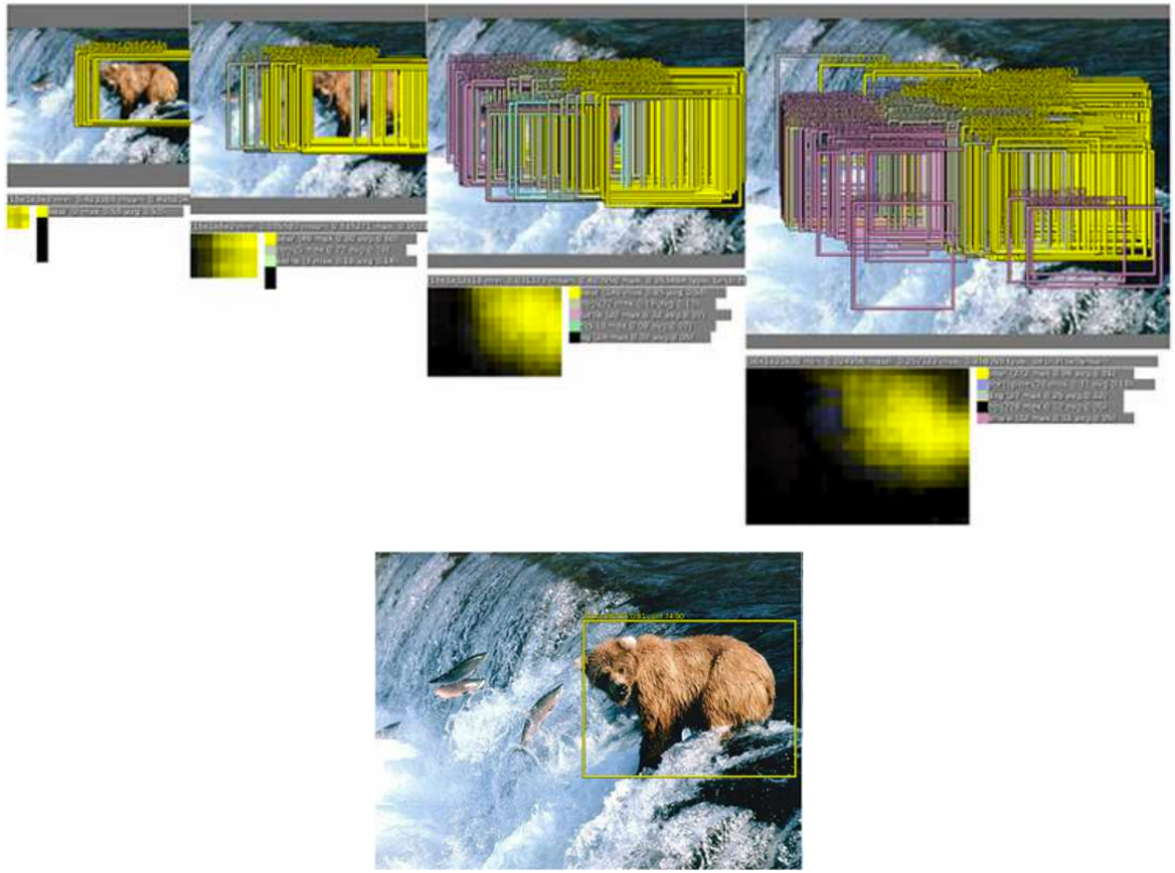
Non-maximal Suppression

- Turn multiple detections into one
- Common approach: merge bounding boxes with >= 0.5 (or some threshold) IoU, keep the higher scoring box.



OverFeat

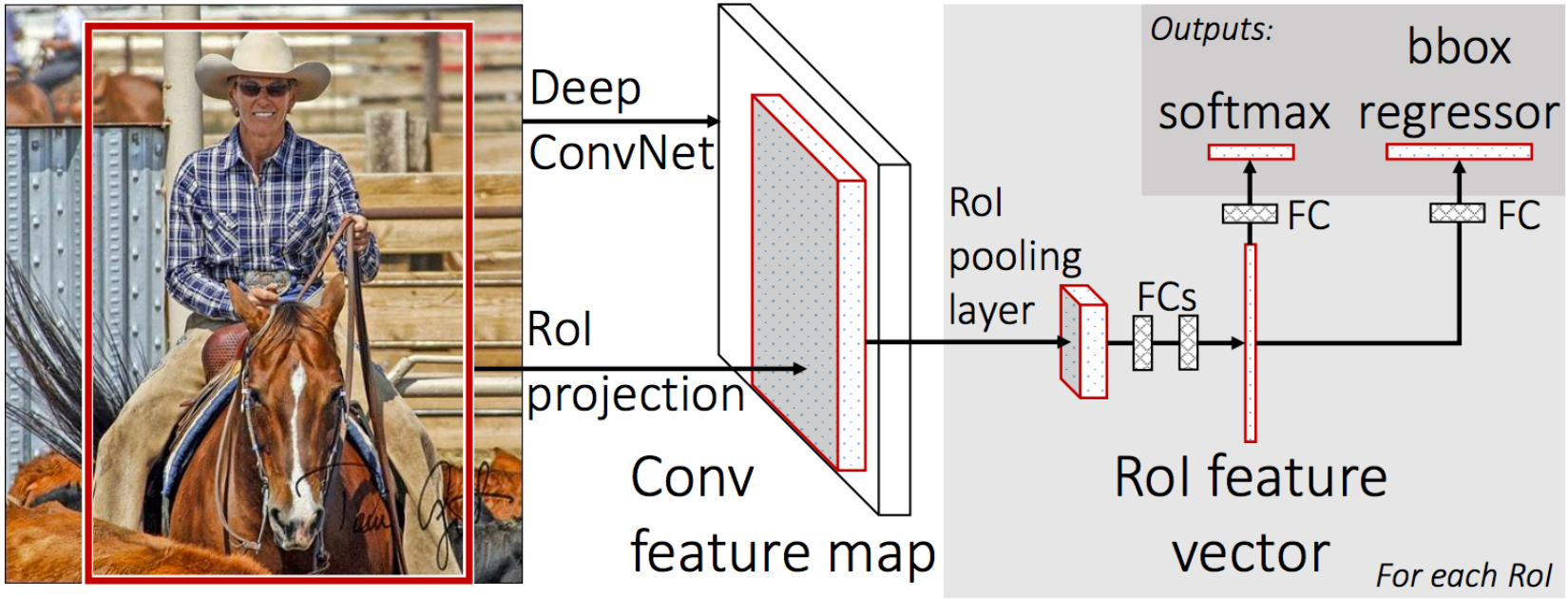
- Efficient sliding windows with CNNs



Sermanet et al. 2013

Fast R-CNN

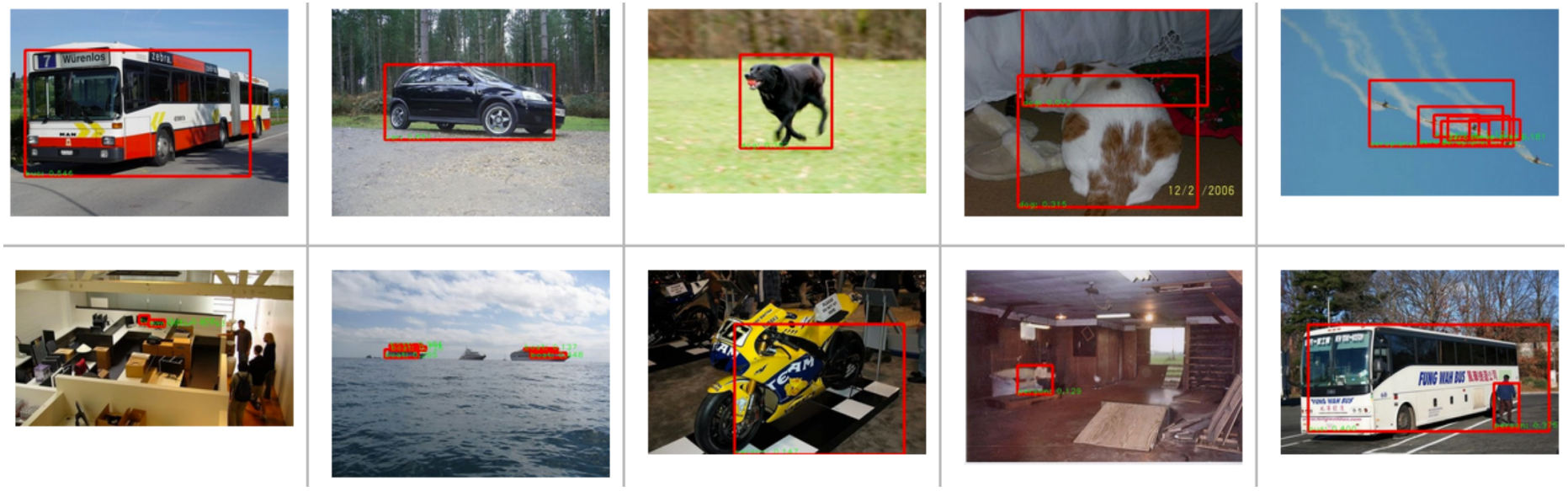
- Very new, reuses most CNN computation across regions



Girshick. 2015

Multibox

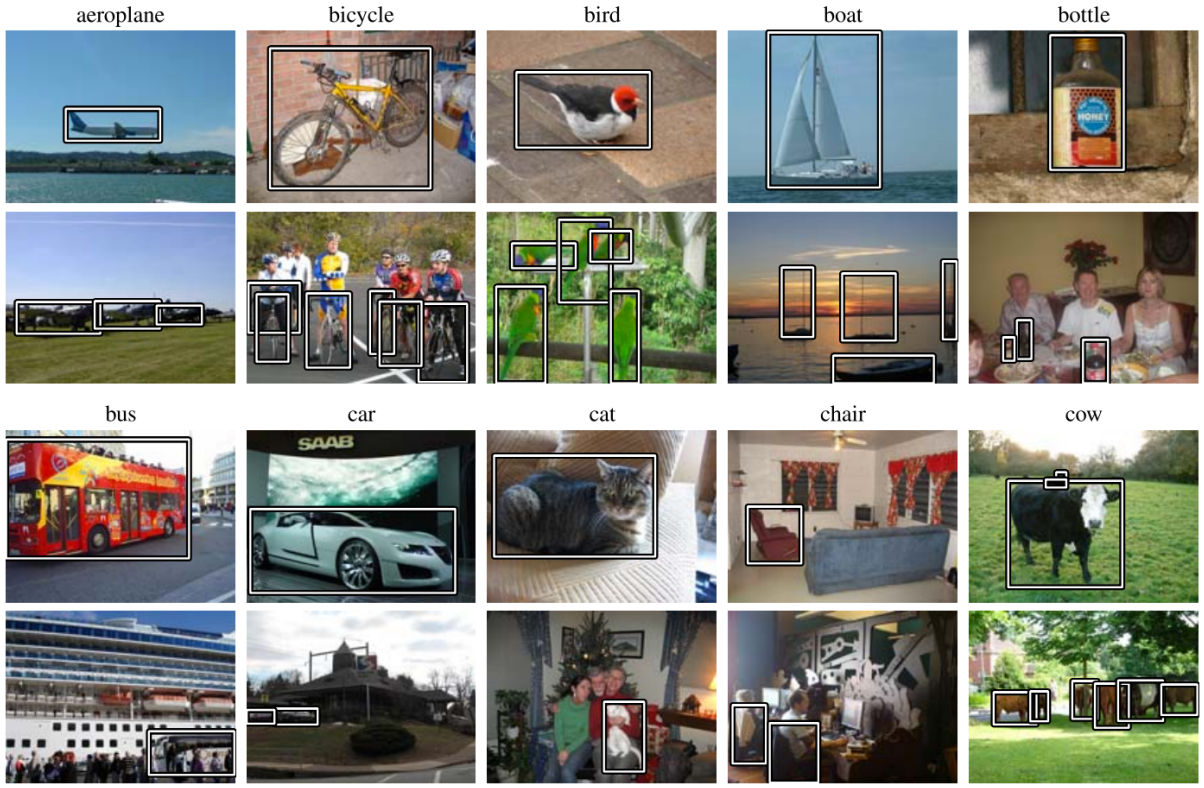
- Try to learn the region proposals



Erhan et al. 2014

Detection Challenges: PASCAL

- 20 Object Categories, thousands of images
- 2007-2012
- Was *the dataset* for a long time.



Detection Challenges: ILSVRC

- 200 Object Categories, 100,000s of images
- 2013-current
- Not all images fully annotated.