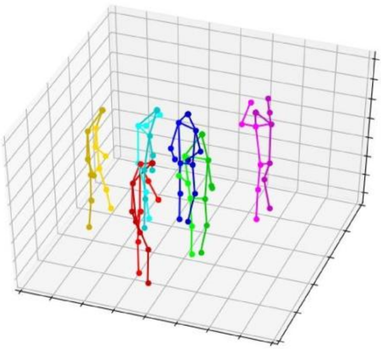
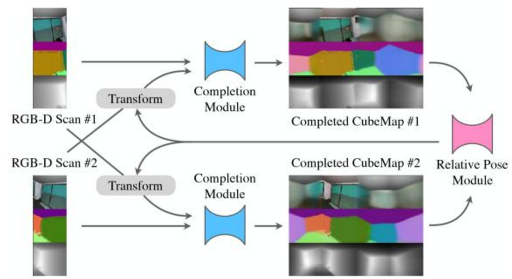
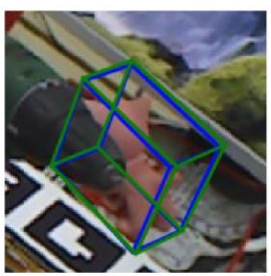
CS376 Computer Vision
Lecture 21: Object Detection

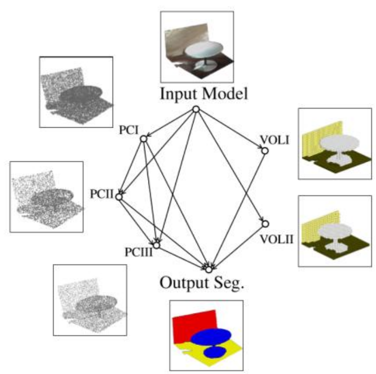
Qixing Huang

April 15th 2019









Window-based generic object detection

Generic category recognition: basic framework

- Build/train object model
  - Choose a representation
  - Learn or fit parameters of model / classifier
- Generate candidates in new image
- Score the candidates

Window-based models
Building an object model

Given the representation, train a binary classifier

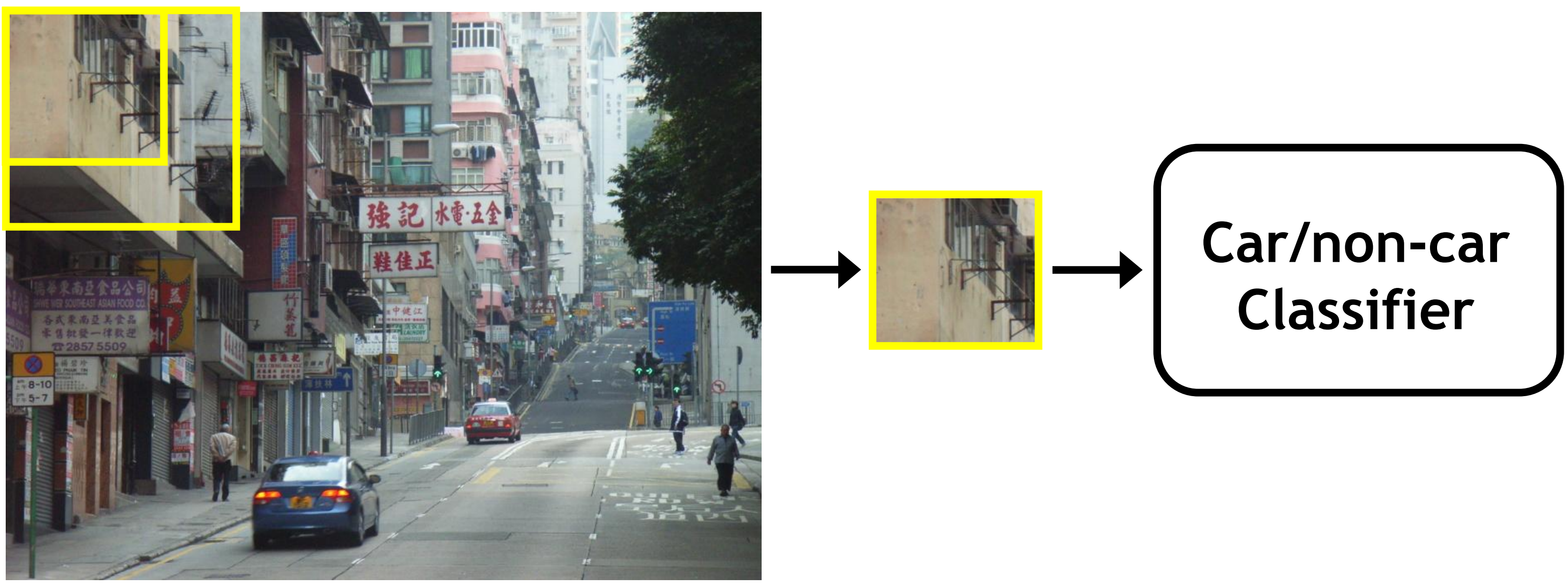
Car/non-car
Classifier

Yes, car.

No, not a car.

Slide: Kristen Grauman

Window-based models
Generating and scoring candidates

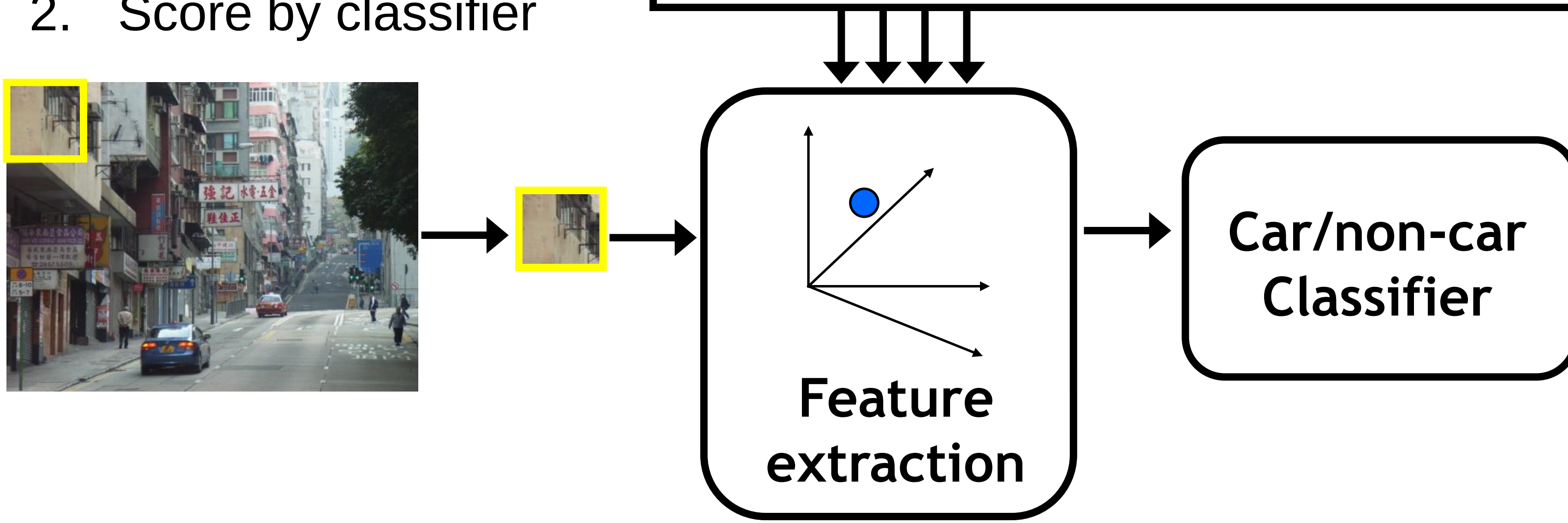


Slide: Kristen Grauman

Window-based object detection: recap

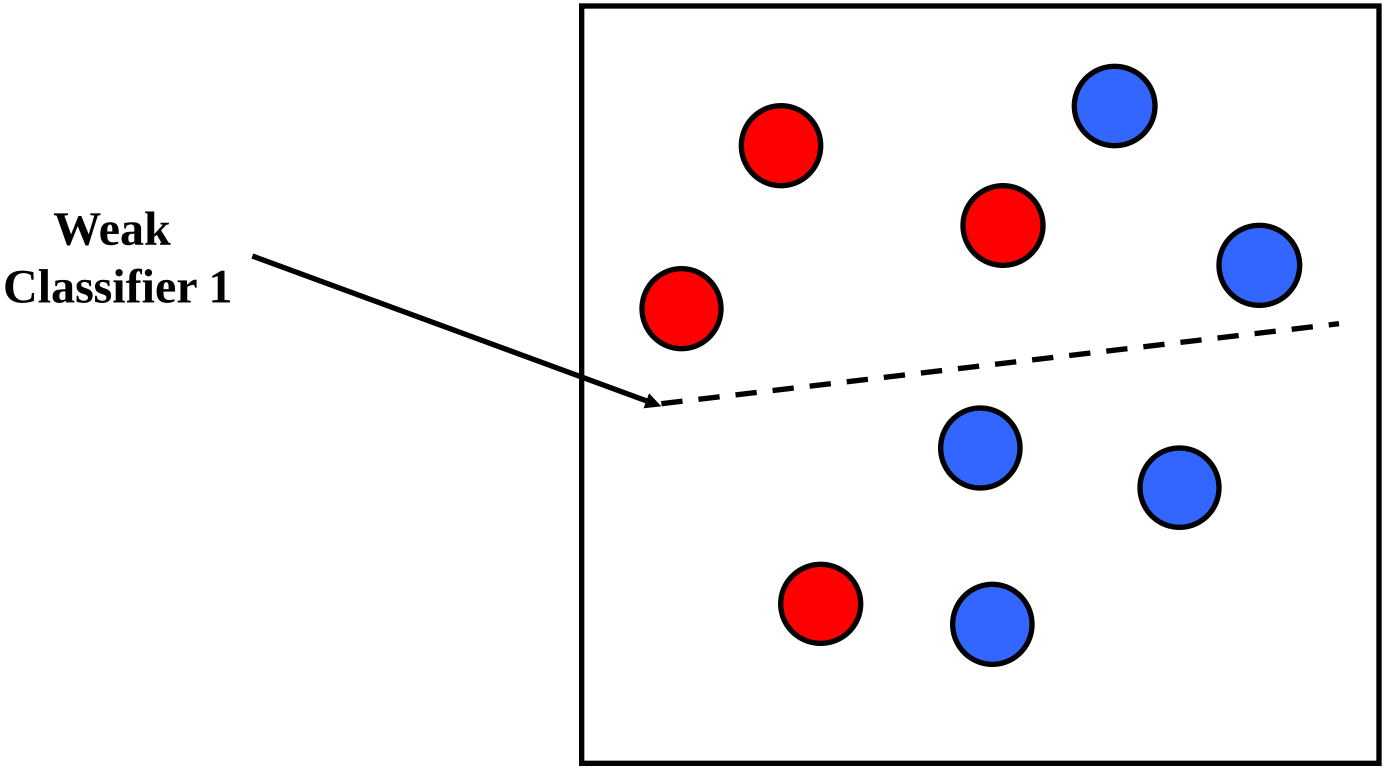
Training:
1. Obtain training data
2. Define features
3. Define classifier

Given new image:
1. Slide window
2. Score by classifier



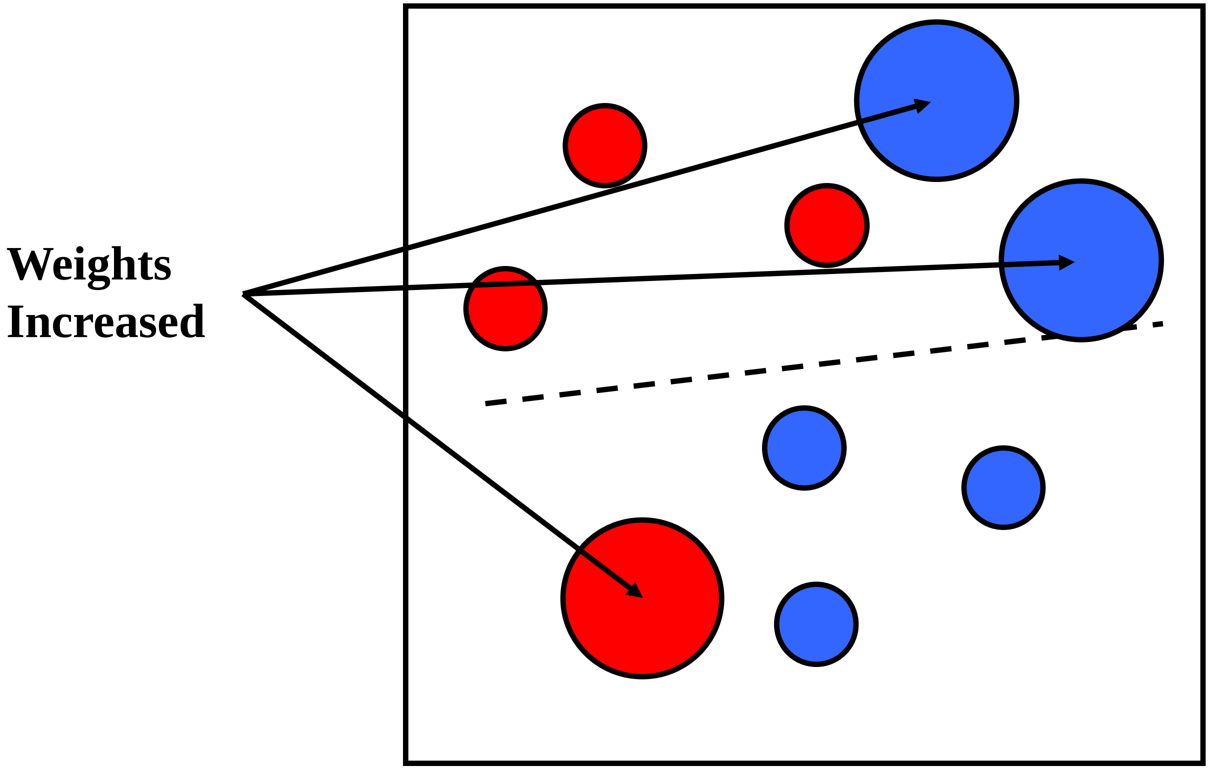
Boosting

Boosting intuition

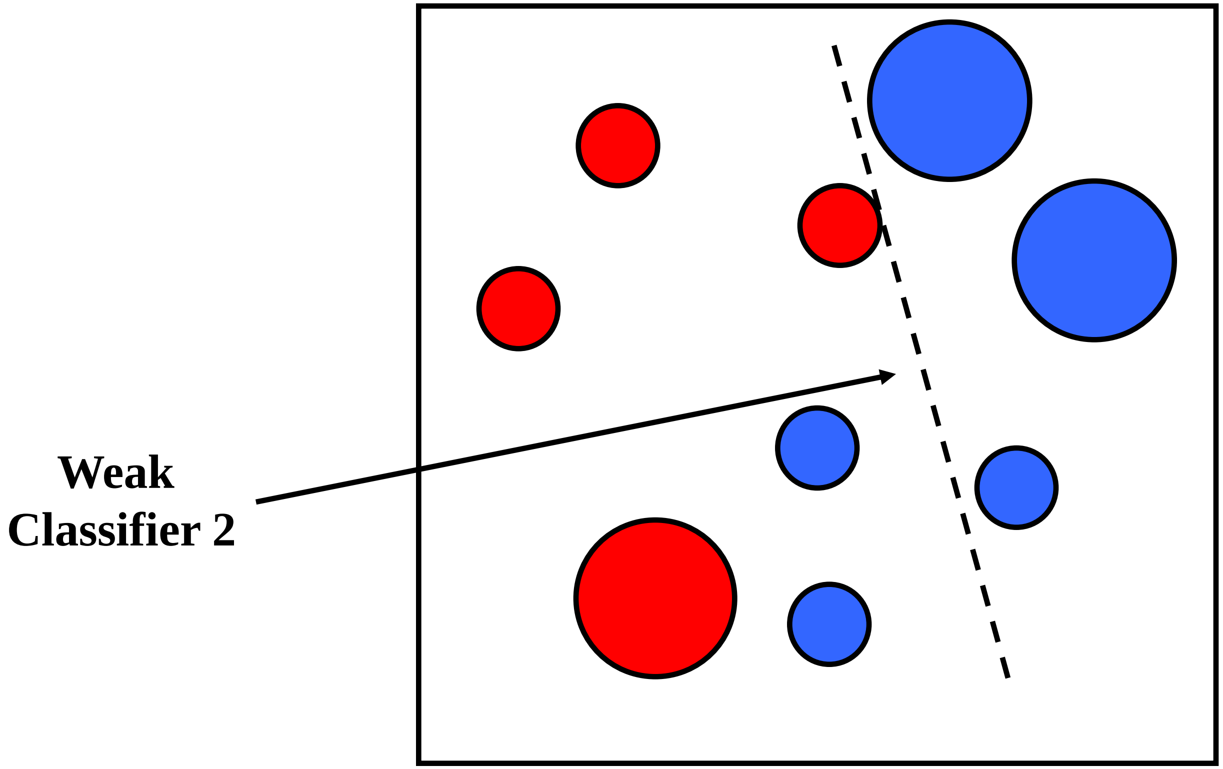


Slide credit: Paul Viola

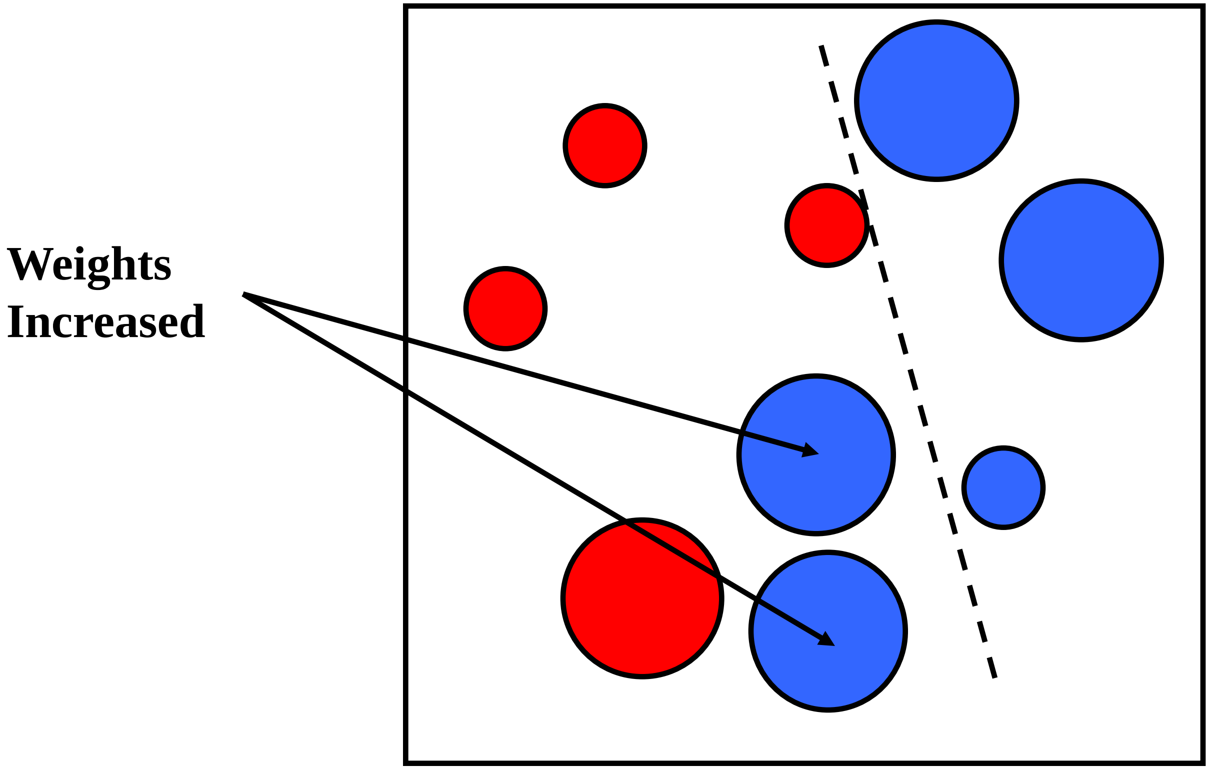
Boosting illustration



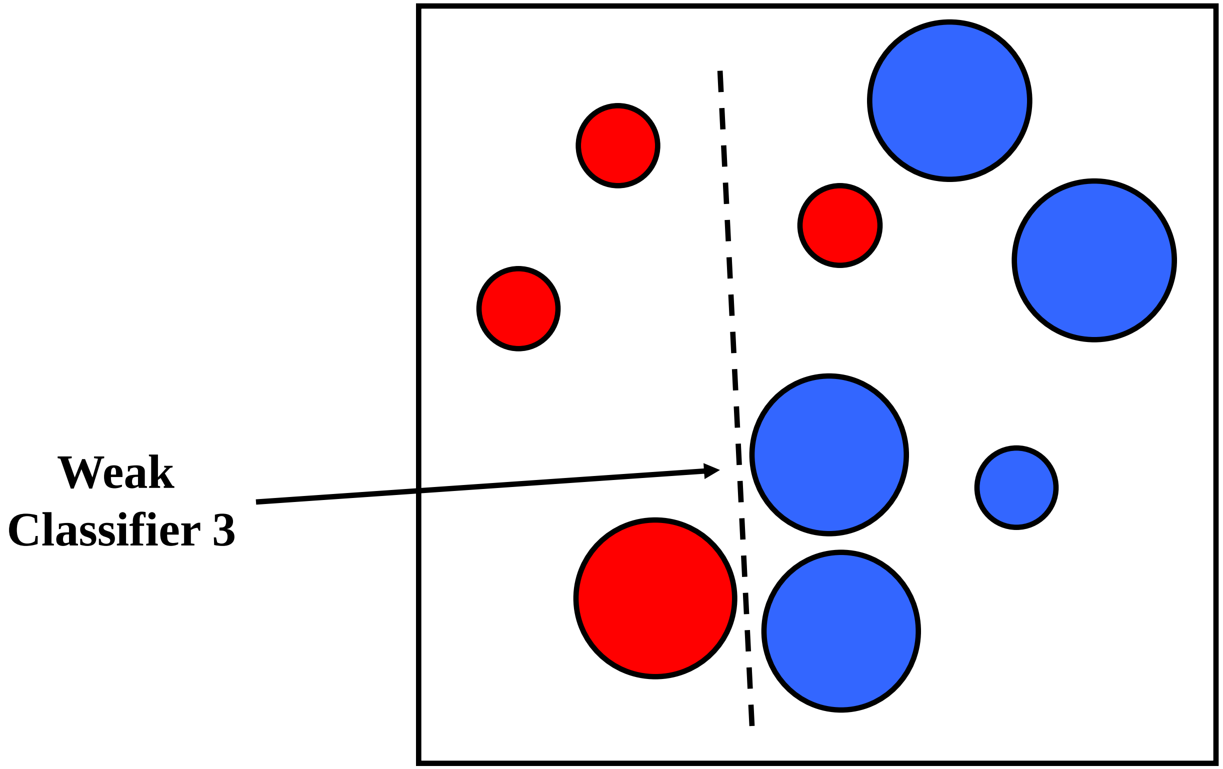
Boosting illustration



Boosting illustration

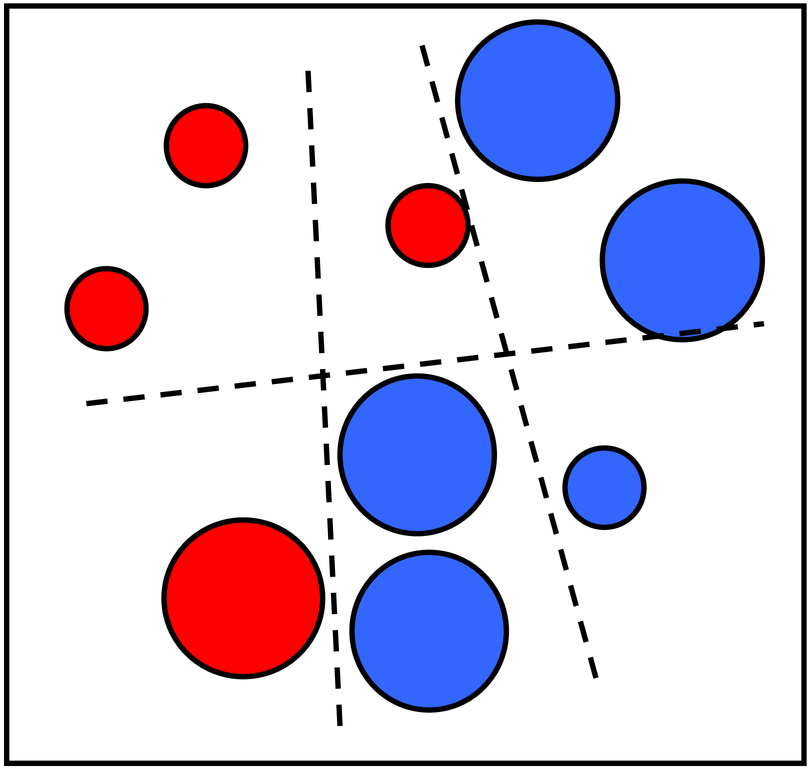


Boosting illustration



Boosting illustration

Final classifier is a combination of weak classifiers



Boosting: training

- Initially, weight each training example equally
- In each boosting round:
  - Find the weak learner that achieves the lowest *weighted* training error
  - Raise weights of training examples misclassified by current weak learner
- Compute final classifier as linear combination of all weak learners (weight of each learner is directly proportional to its accuracy)

- Exact formulas for re-weighting and combining weak learners depend on the particular boosting scheme (e.g., AdaBoost)

Slide credit: Lana Lazebnik

Viola-Jones face detector

ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 2001

Rapid Object Detection using a Boosted Cascade of Simple Features

Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

Michael Jones
mjones@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

Abstract

This paper describes a machine learning approach for vi-

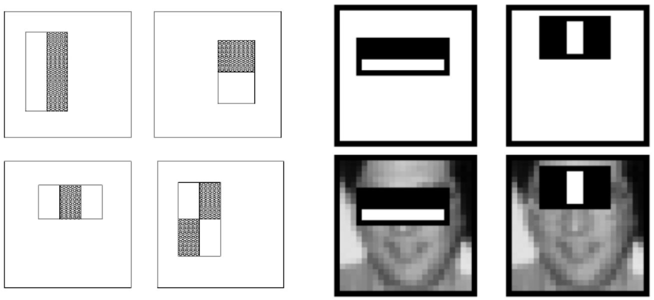
tected at 15 frames per second on a conventional 700 MHz Intel Pentium III. In other face detection systems, auxiliary information, such as image differences in video sequences,

Viola-Jones face detector

Main idea:

- Represent local texture with efficiently computable "rectangular" features within window of interest
- Select discriminative features to be weak classifiers
- Use boosted combination of them as final classifier
- Form a cascade of such classifiers, rejecting clear negatives quickly

Viola-Jones detector: features

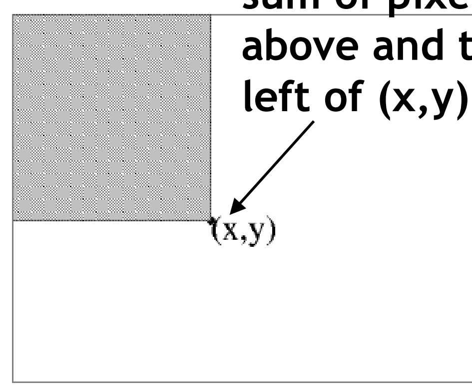


“Rectangular” filters

Feature output is difference between adjacent regions

Efficiently computable with integral image: any sum can be computed in constant time.

Value at (x,y) is sum of pixels above and to the left of (x,y)



Integral image

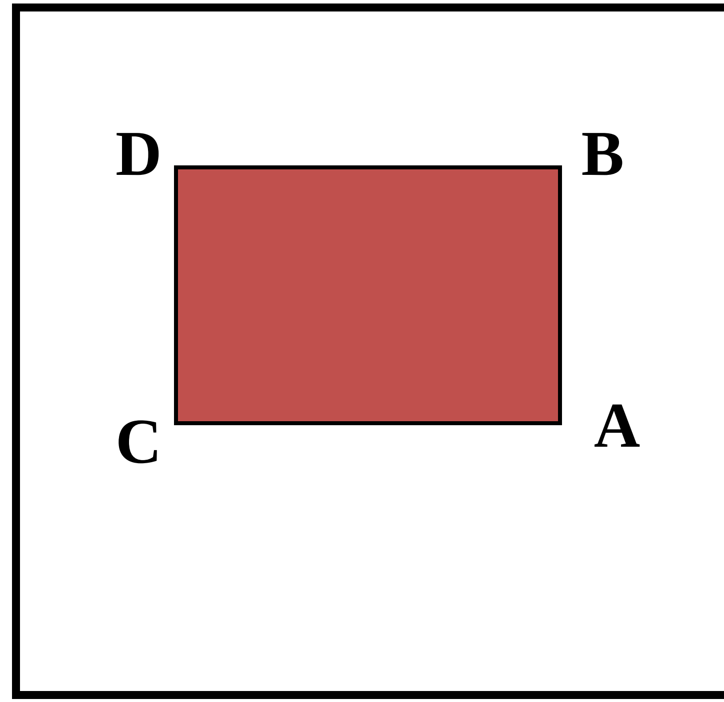
Slide: Kristen Grauman

Computing sum within a rectangle

- Let A,B,C,D be the values of the integral image at the corners of a rectangle
- Then the sum of original image values within the rectangle can be computed as:

  sum = A – B – C + D

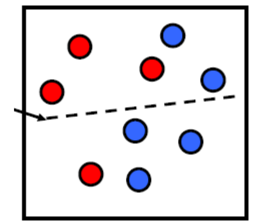
- Only 3 additions are required for any size of rectangle!



AdaBoost Algorithm

- Given example images $(x_1, y_1), \ldots, (x_n, y_n)$ 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 $t = 1, \ldots, T$:
  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, $\epsilon_j = \sum_i w_i \, |h_j(x_i) - y_i|$.
  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 $\alpha_t = \log \frac{1}{\beta_t}$

Start with uniform weights on training examples



$\{x_1, \ldots x_n\}$

For T rounds

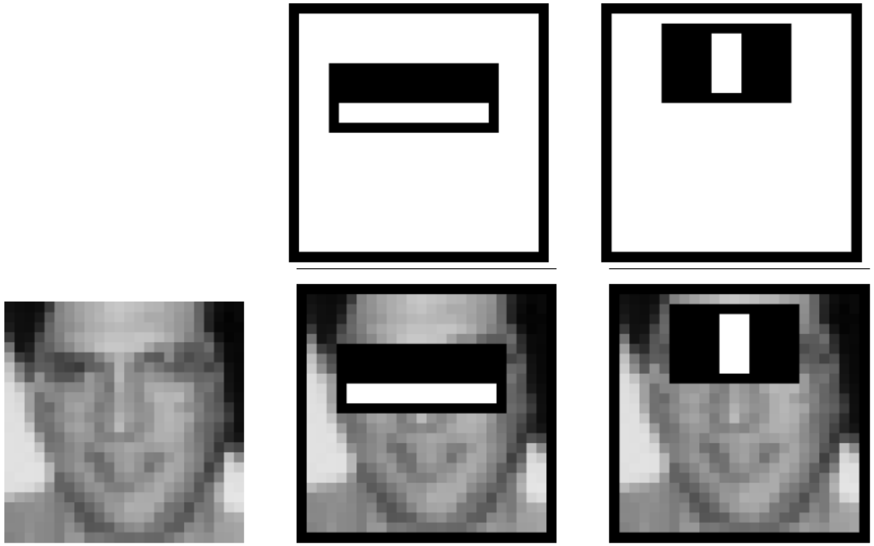
Evaluate *weighted* error for each feature, pick best.

Re-weight the examples:
Incorrectly classified -> more weight
Correctly classified -> less weight

Final classifier is combination of the weak ones, weighted according to error they had.

Freund & Schapire 1995

Viola-Jones Face Detector: Results



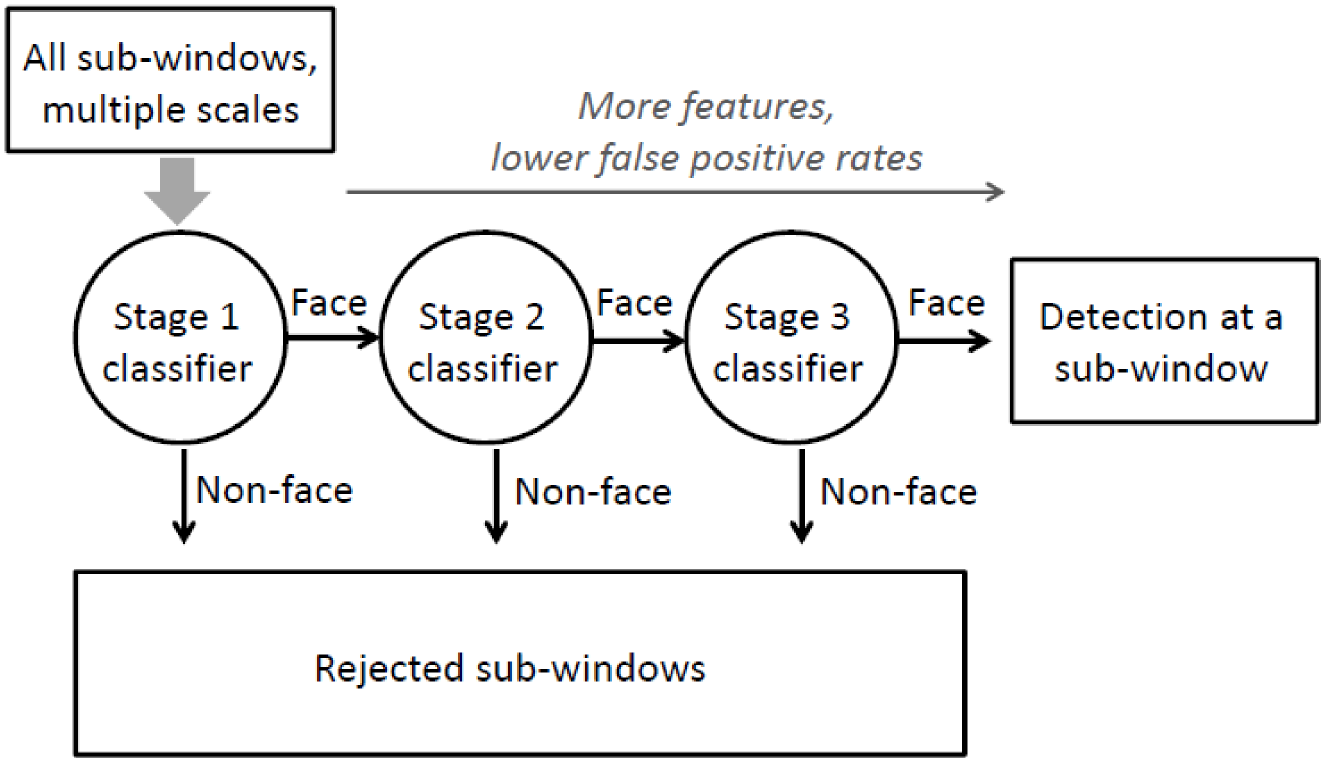
First two features selected

A practical issue

- Even if the filters are fast to compute, each new image has a lot of possible windows to search.

- How to make the detection more efficient?

Cascading classifiers for detection



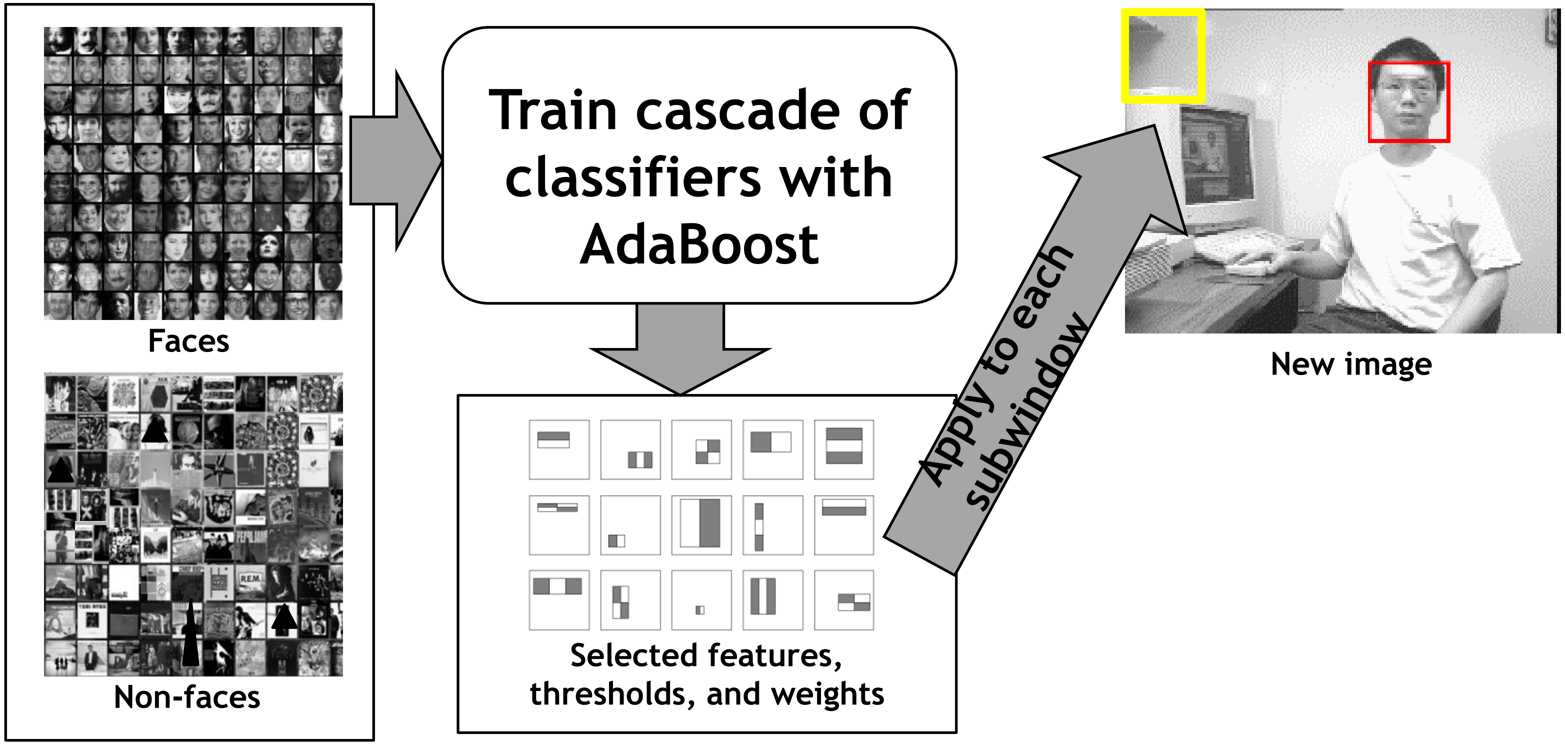
- Form a *cascade* with low false negative rates early on
- Apply less accurate but faster classifiers first to immediately discard windows that clearly appear to be negative

Slide: Kristen Grauman

Training the cascade

- Set target detection and false positive rates for each stage
- Keep adding features to the current stage until its target rates have been met
  - Need to lower AdaBoost threshold to maximize detection (as opposed to minimizing total classification error)
  - Test on a *validation set*
- If the overall false positive rate is not low enough, then add another stage
- Use false positives from current stage as the negative training examples for the next stage

Viola-Jones detector: summary



Train with 5K positives, 350M negatives

Real-time detector using 38 layer cascade

6061 features in all layers

[Implementation available in OpenCV]

Slide: Kristen Grauman

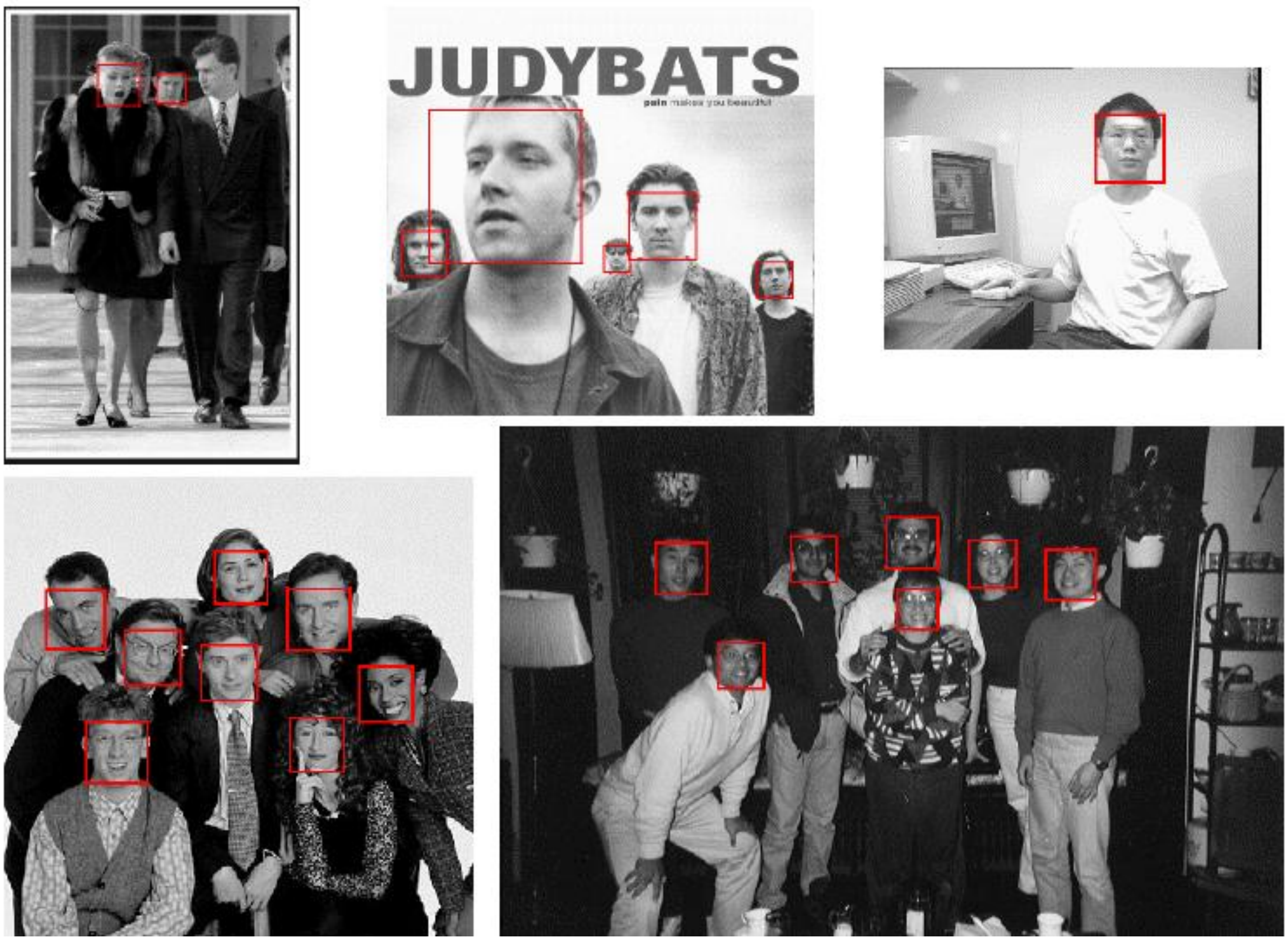
Viola-Jones detector: summary

- A seminal approach to real-time object detection
  - 15,700 citations and counting
- Training is slow, but detection is very fast
- Key ideas
  - *Integral images* for fast feature evaluation
  - *Boosting* for feature selection
  - *Attentional cascade* of classifiers for fast rejection of non-face windows

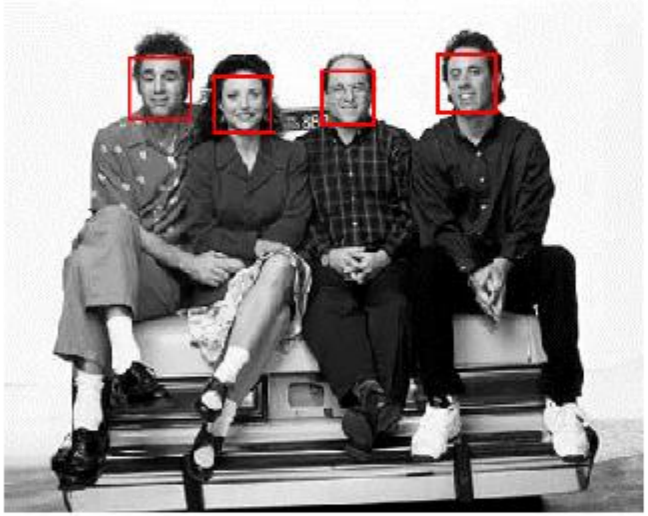
P. Viola and M. Jones. *Rapid object detection using a boosted cascade of simple features.* CVPR 2001.

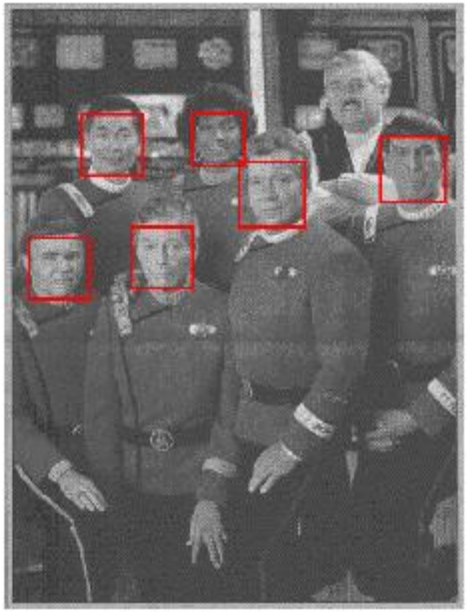
P. Viola and M. Jones. *Robust real-time face detection.* IJCV 57(2), 2004.

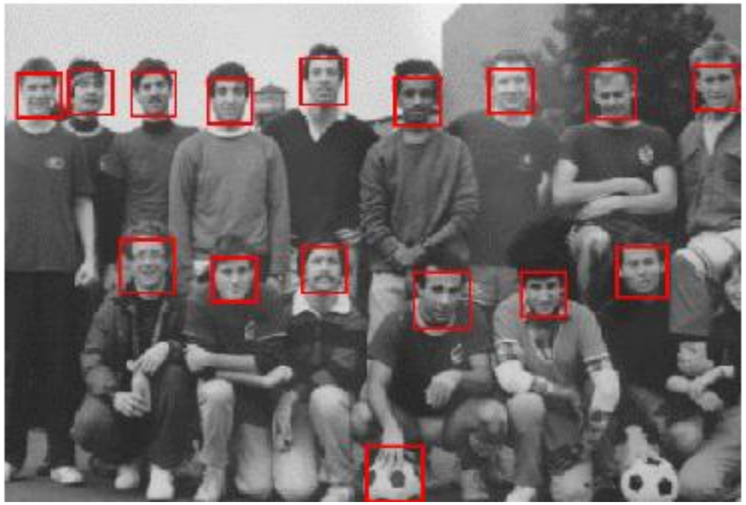
Viola-Jones Face Detector: Results

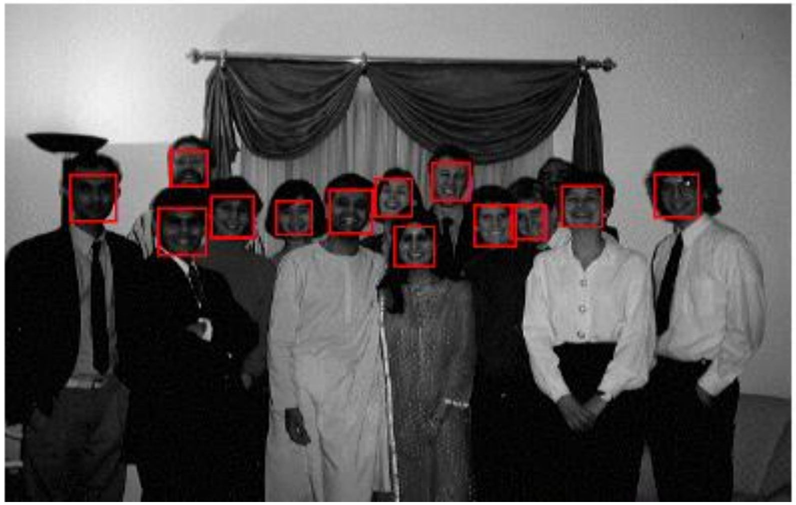


Viola-Jones Face Detector: Results





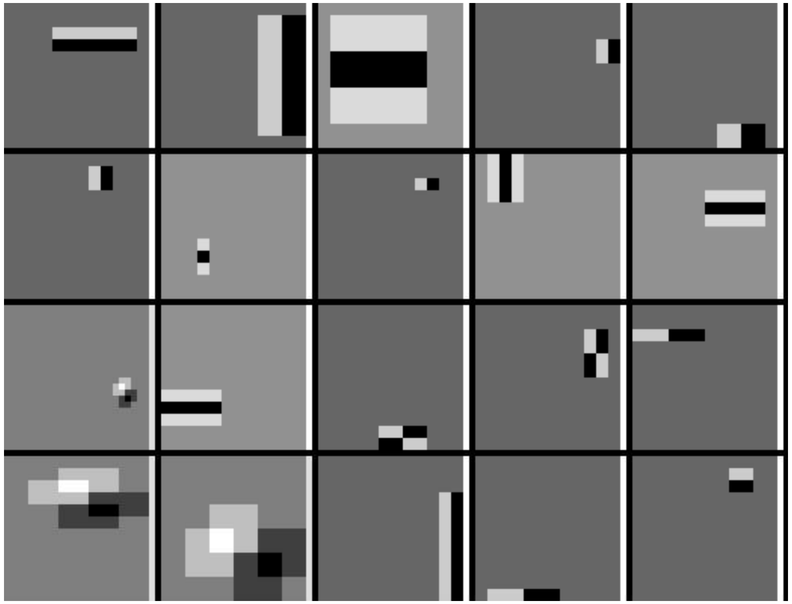




Viola-Jones Face Detector: Results

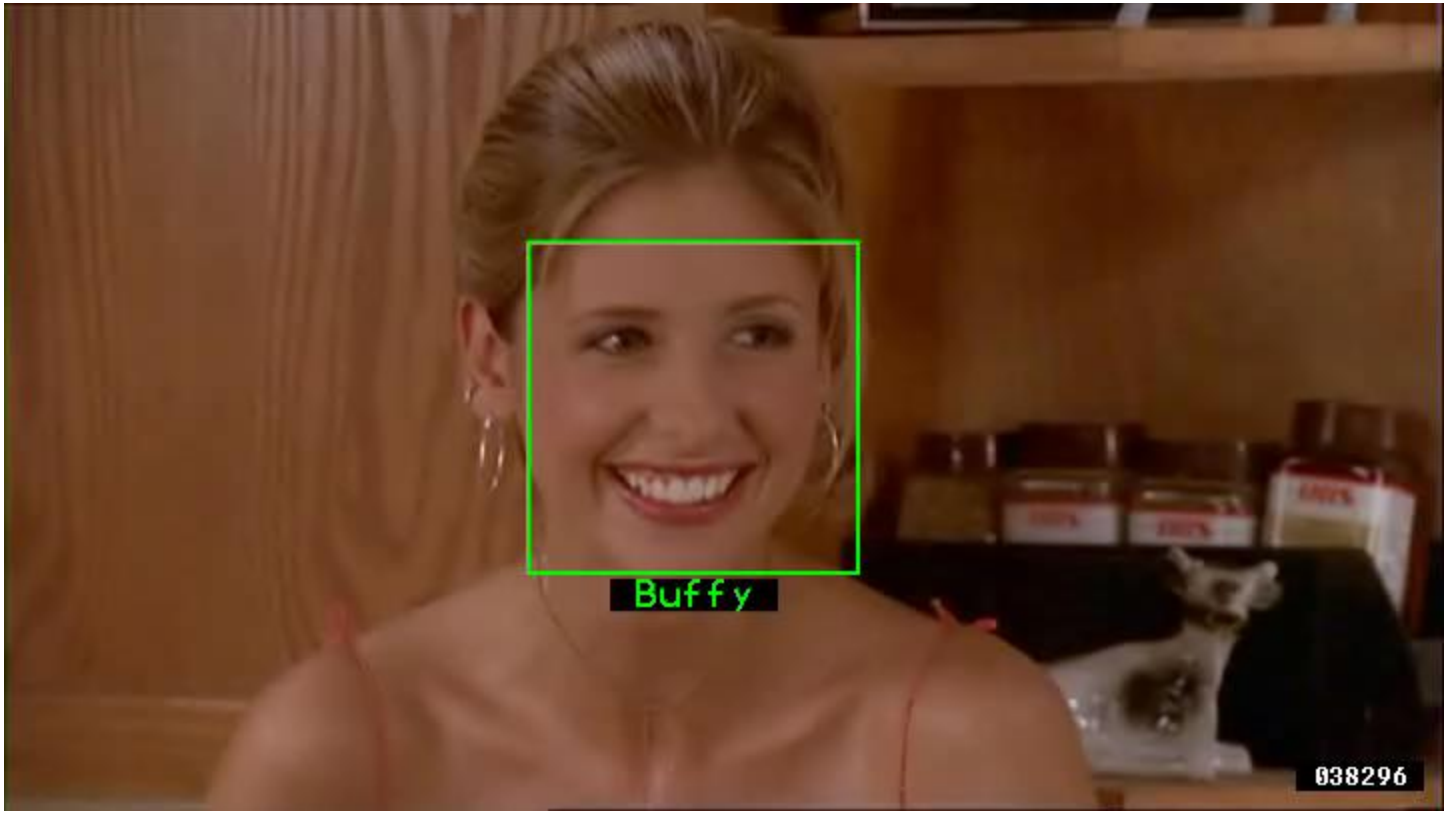
Detecting profile faces?

Can we use the same detector?



Viola-Jones Face Detector: Results

Example using Viola-Jones detector



Frontal faces detected and then tracked, character names inferred with alignment of script and subtitles.

Everingham, M., Sivic, J. and Zisserman, A.
"Hello! My name is... Buffy" - Automatic naming of characters in TV video, BMVC 2006. http://www.robots.ox.ac.uk/~vgg/research/nface/index.html

Window-based detection: strengths

- Sliding window detection and global appearance descriptors:
  - Simple detection protocol to implement
  - Good feature choices critical
  - Past successes for certain classes

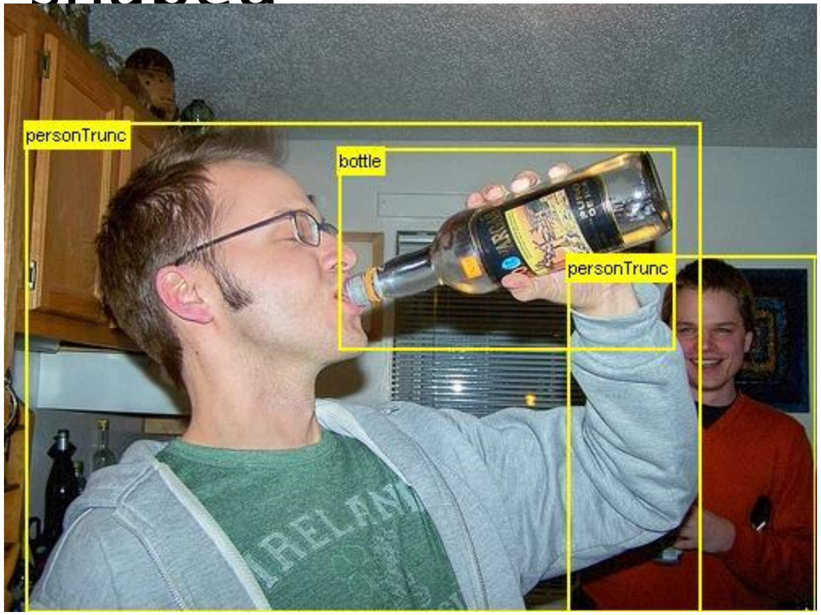
Slide: Kristen Grauman

Window-based detection: Limitations

- High computational complexity
  - For example: 250,000 locations x 30 orientations x 4 scales = 30,000,000 evaluations!
  - If training binary detectors independently, means cost increases linearly with number of classes
- With so many windows, false positive rate better be low

Limitations (continued)

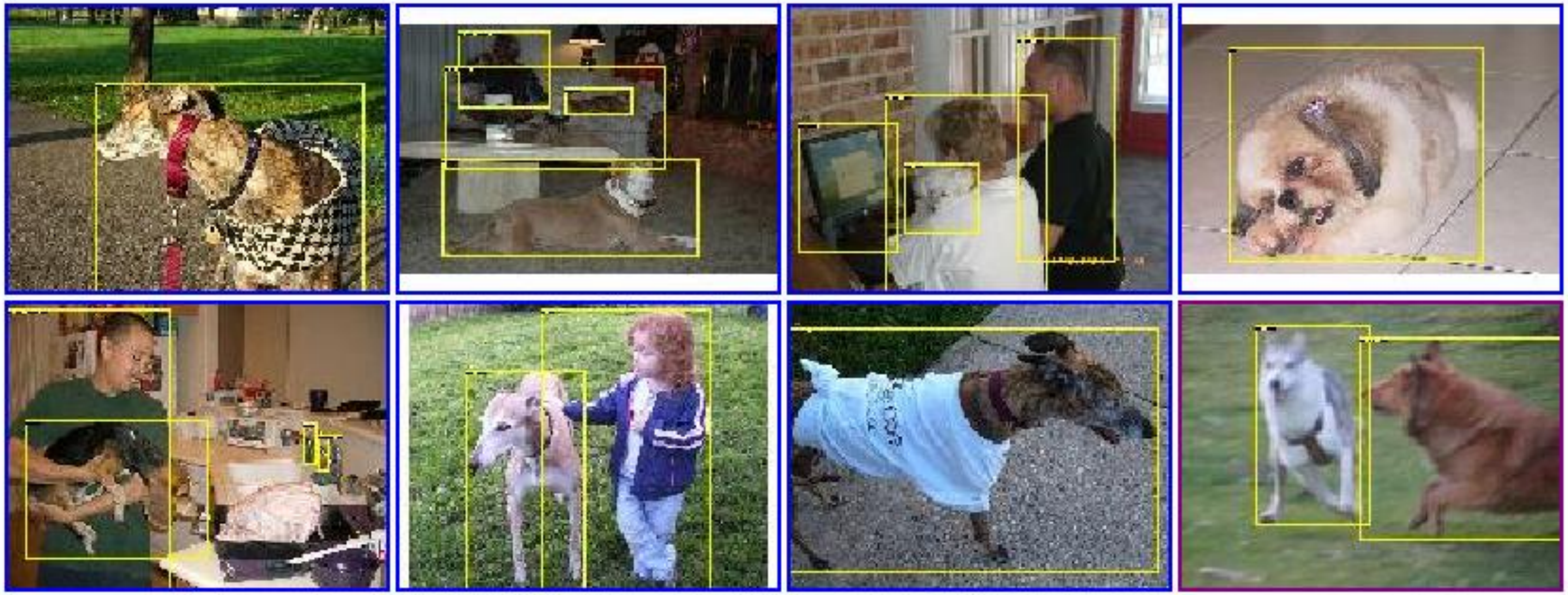
- Not all objects are “box” shaped



Slide: Kristen Grauman

Limitations (continued)

- Non-rigid, deformable objects not captured well with representations assuming a fixed 2d structure; or must assume fixed viewpoint



Slide: Kristen Grauman

Summary

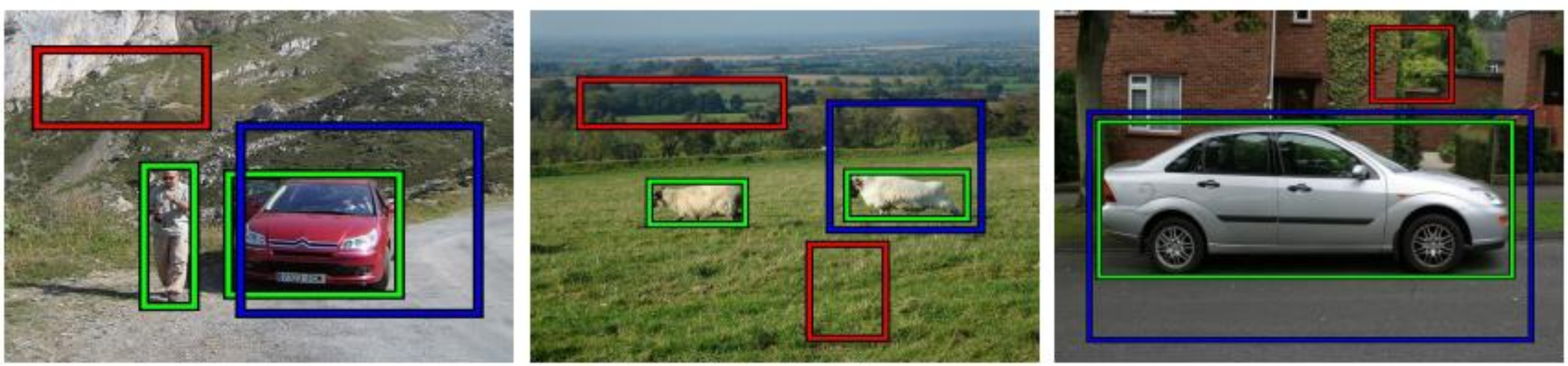
- Basic pipeline for window-based detection
  - Model/representation/classifier choice
  - Sliding window and classifier scoring
- Boosting classifiers: general idea
- Viola-Jones face detector
  - Exemplar of basic paradigm
  - Plus key ideas: rectangular features, Adaboost for feature selection, cascade
- Pros and cons of window-based detection

Object Proposals

Object proposals

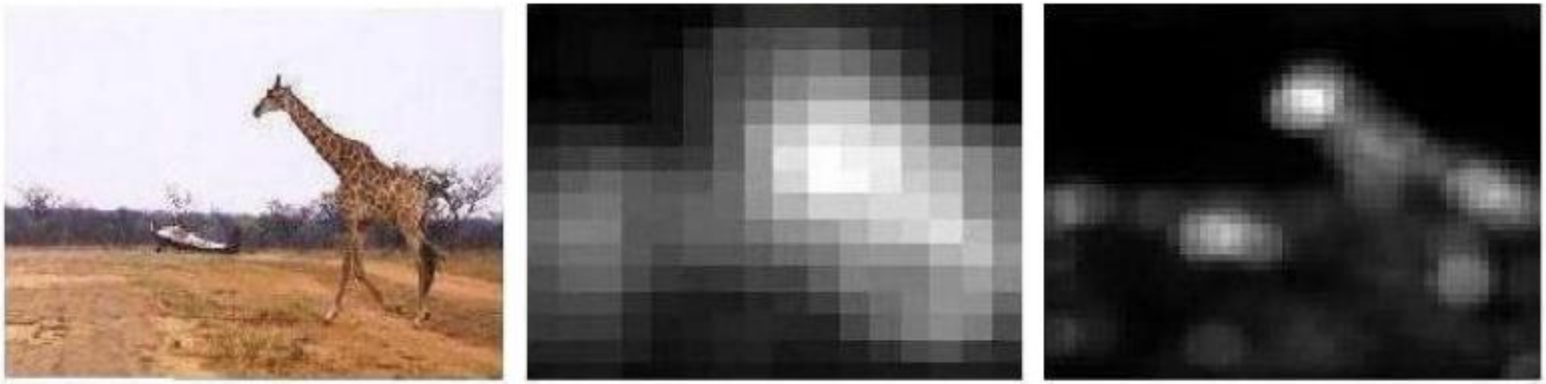
Main idea:

- Learn to generate category-independent regions/boxes that have object-like properties.
- Let object detector search over “proposals”, not exhaustive sliding windows



Alexe et al. Measuring the objectness of image windows, PAMI 2012

Object proposals



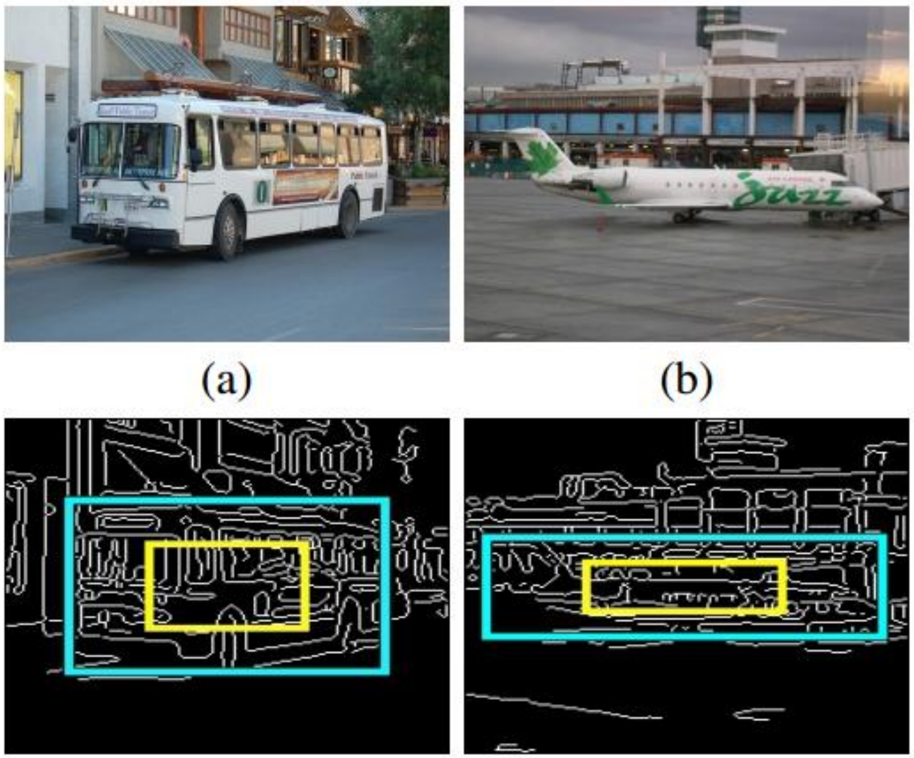
Multi-scale saliency

Color contrast

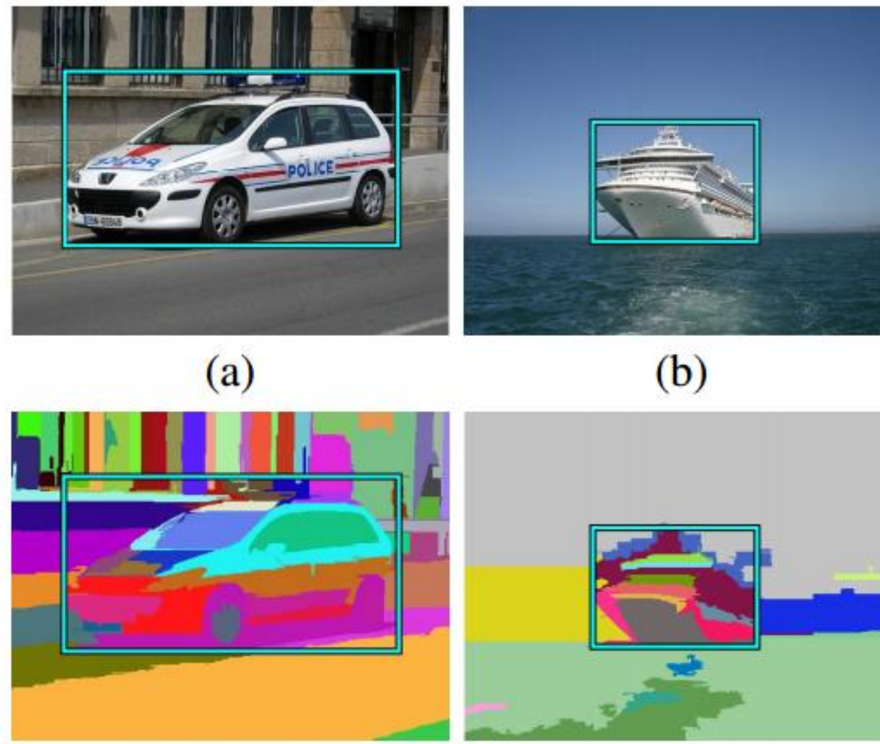
Alexe et al. Measuring the objectness of image windows, PAMI 2012

Object proposals

Edge density

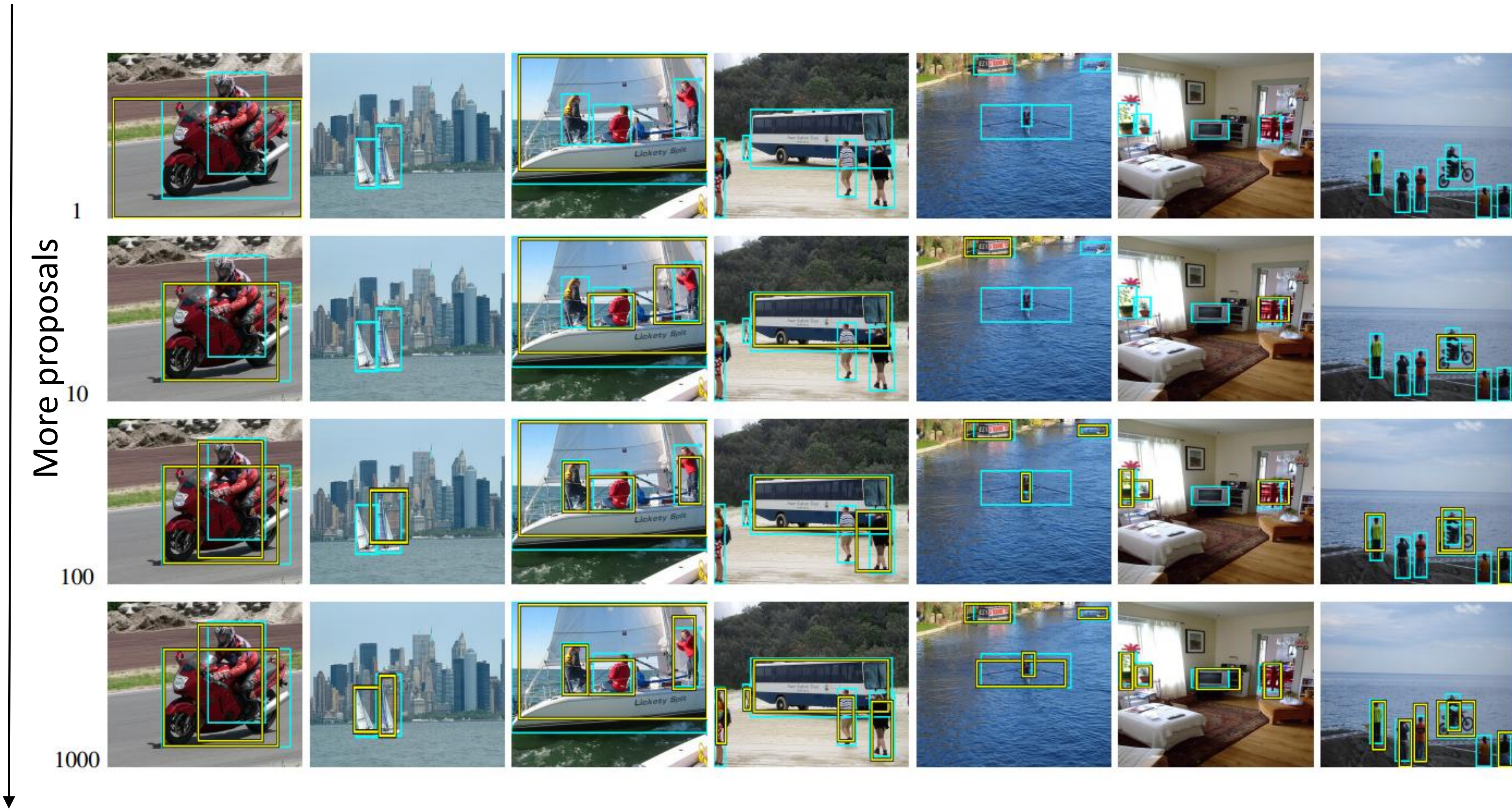


Superpipxel straddling



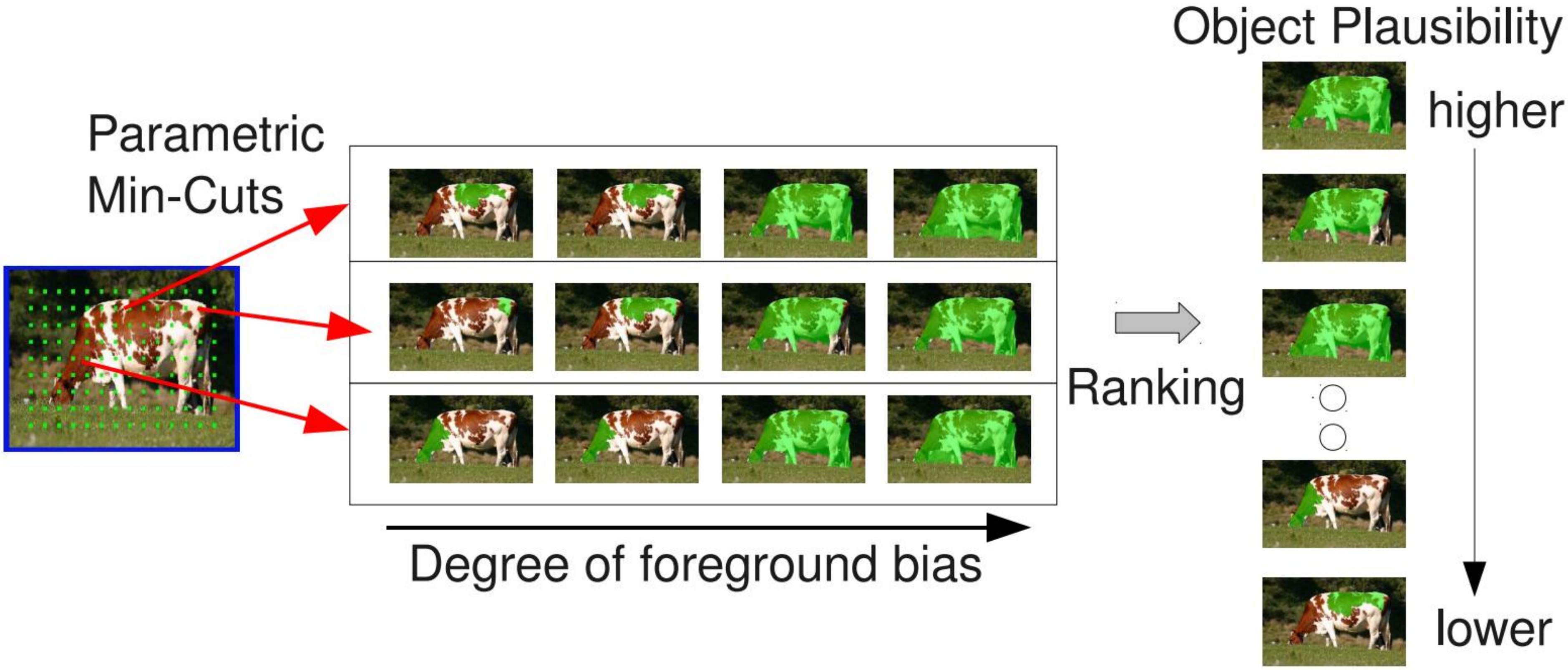
Alexe et al. Measuring the objectness of image windows, PAMI 2012

Object proposals



Alexe et al. Measuring the objectness of image windows, PAMI 2012

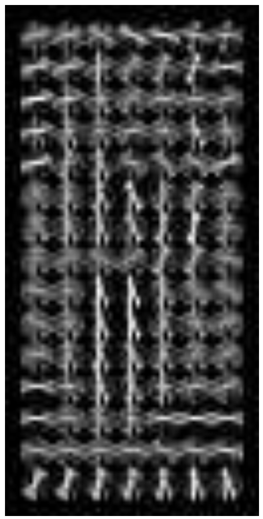
Region-based object proposals



- J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. PAMI, 2012.

Object Proposal Classification

Person detection with HoG's & linear SVM's



- Histogram of oriented gradients (HoG): Map each grid cell in the input window to a histogram counting the gradients per orientation.
- Train a linear SVM using training set of pedestrian vs. non-pedestrian windows.

Dalal & Triggs, CVPR 2005

Person detection with HoGs & linear SVMs

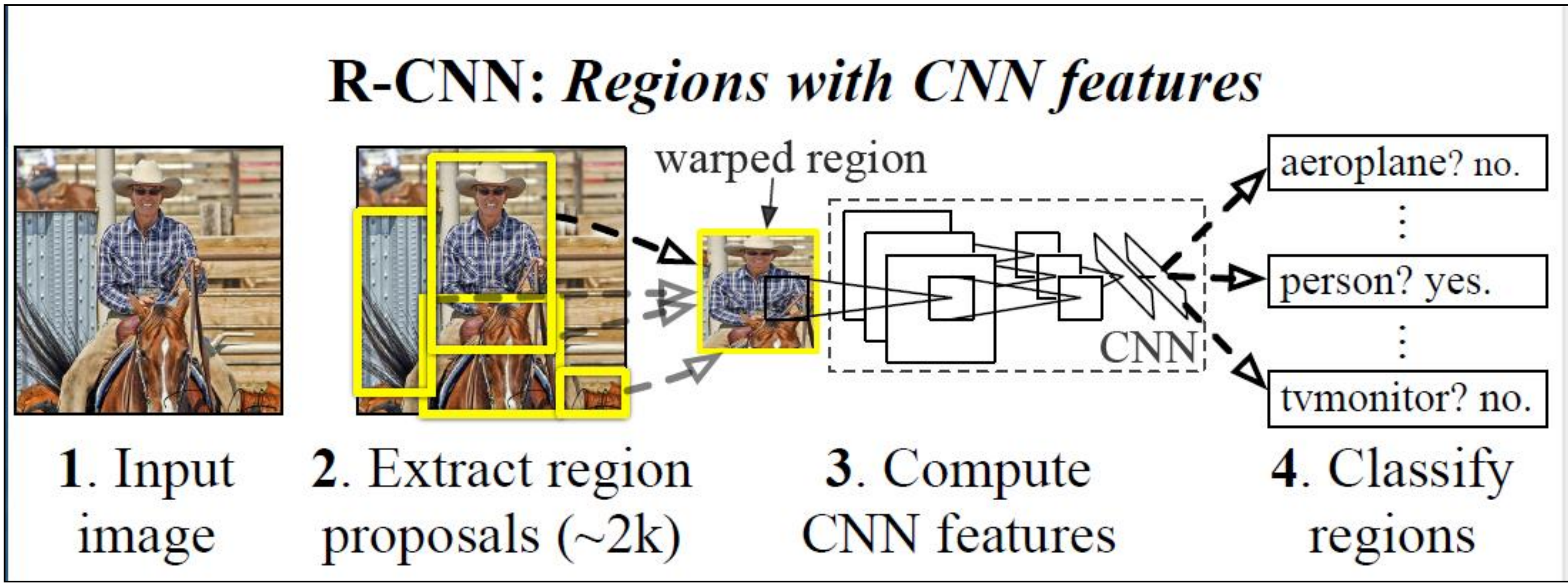
- Histograms of Oriented Gradients for Human Detection, Navneet Dalal, Bill Triggs, International Conference on Computer Vision & Pattern Recognition - June 2005
- http://lear.inrialpes.fr/pubs/2005/DT05/

Summary

- Object recognition as classification task
  - Boosting (face detection ex)
  - Support vector machines and HOG (person detection ex)

- Sliding window search paradigm
  - Pros and cons
  - Speed up with attentional cascade
  - Object proposals, proposal regions as alternative

Region CNNs



R-CNN: Regions with CNN features
warped region
aeroplane? no.
person? yes.
tvmonitor? no.
CNN
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Region CNN

- Pretraining
  - 1.2 Million images with class labels

- Fine-tuned on PASCAL VOC
  - 20K images with object labels

Deep Learning for Object Detection

R-CNN OverFeat DetectorNet DeepMultibox SPP-net Fast R-CNN MR-CNN SSD YOLO YOLOv2 G-CNN AttractioNet Mask R-CNN R-FCN RPN FPN Faster R-CNN …

Slide Credit: Ross Girshick

Common to all Methods

Start by modifying a classification network

Since R-CNN, this network is pre-trained, typically using ImageNet (cf. DetectorNet)

Slide Credit: Ross Girshick

Highest Information Gain Split: "Stage" Count

More than one stage

- DetectorNet (Szegedy et al.)
- R-CNN (Girshick et al.)
- SPP-net (He et al.)
- Fast R-CNN (Girshick)
- Faster R-CNN (Ren et al.)
- R-FCN (Dai et al.)
- Mask R-CNN (He et al.)

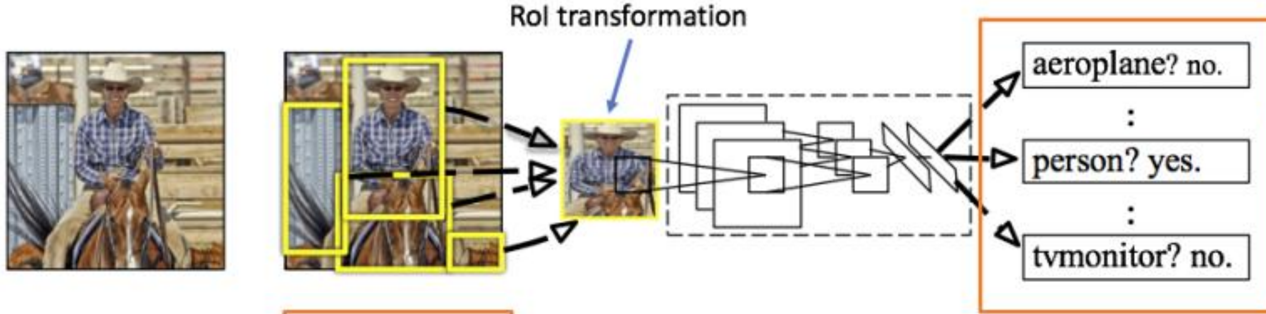
One stage

- OverFeat (Sermanet et al.)
- YOLO, YOLOv2 (Redmon et al.)
- SSD (Wei et al.)
- RetinaNet (Lin et al.) [Poster at WICV on Wed.]

Slide Credit: Ross Girshick

More than one “stage” (≈ proposal based; but doesn’t require proposals)

Classification of *reduced* output space elements



Cascade-like reduction in output space

Input image

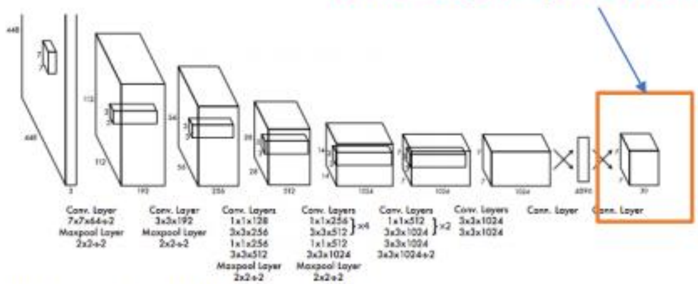
Object / region proposals

Deep Learning region classifier

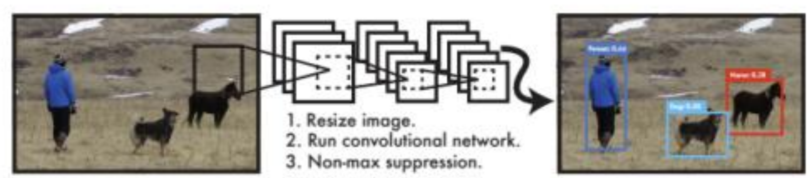
Region classification, box regression

One stage

Direct classification of *all* output space elements

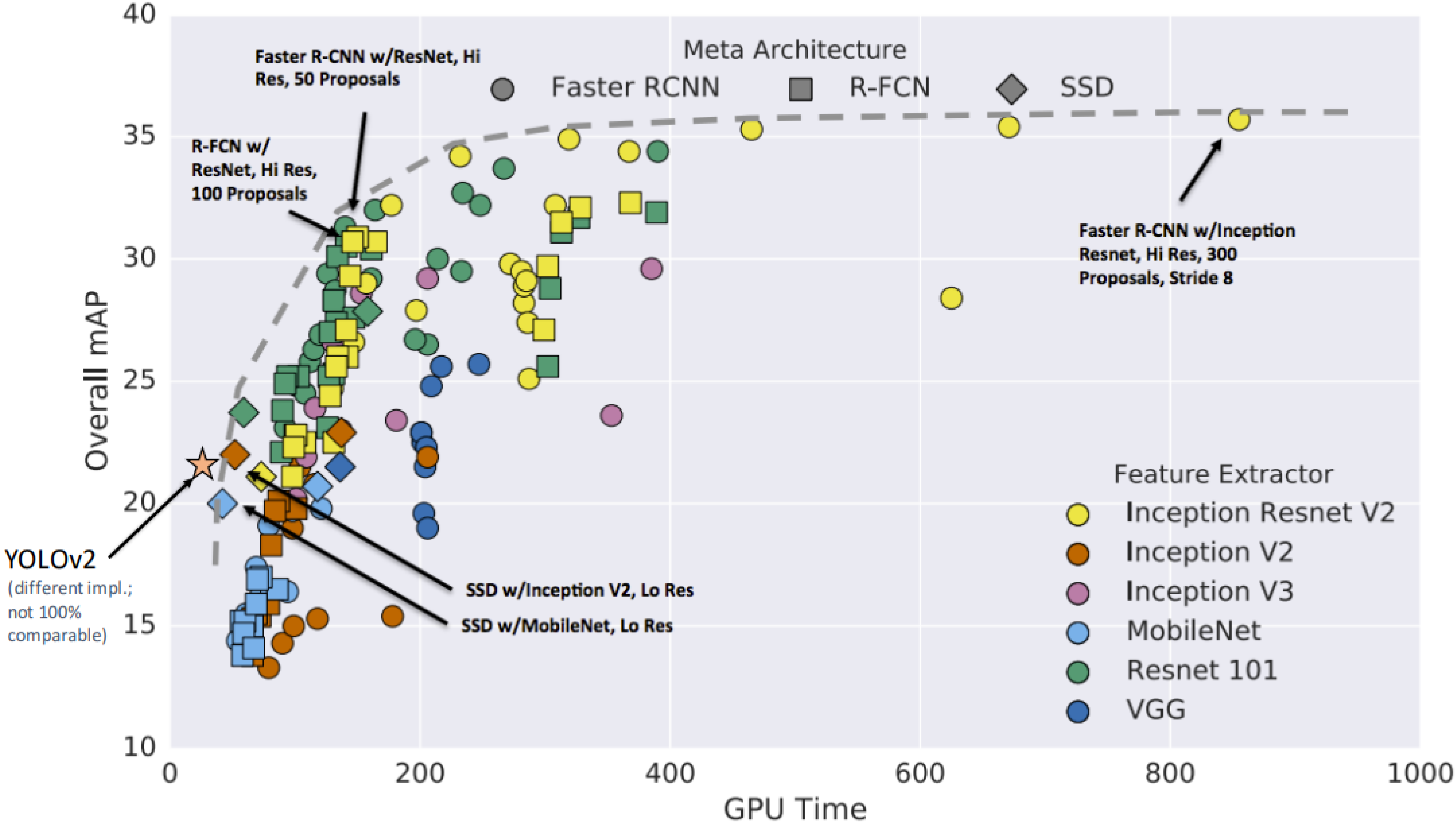


Redmond et al. You Only Look Once: Unified Real-time Object Detection. In CVPR 2016



“You only look once”
“Single shot”

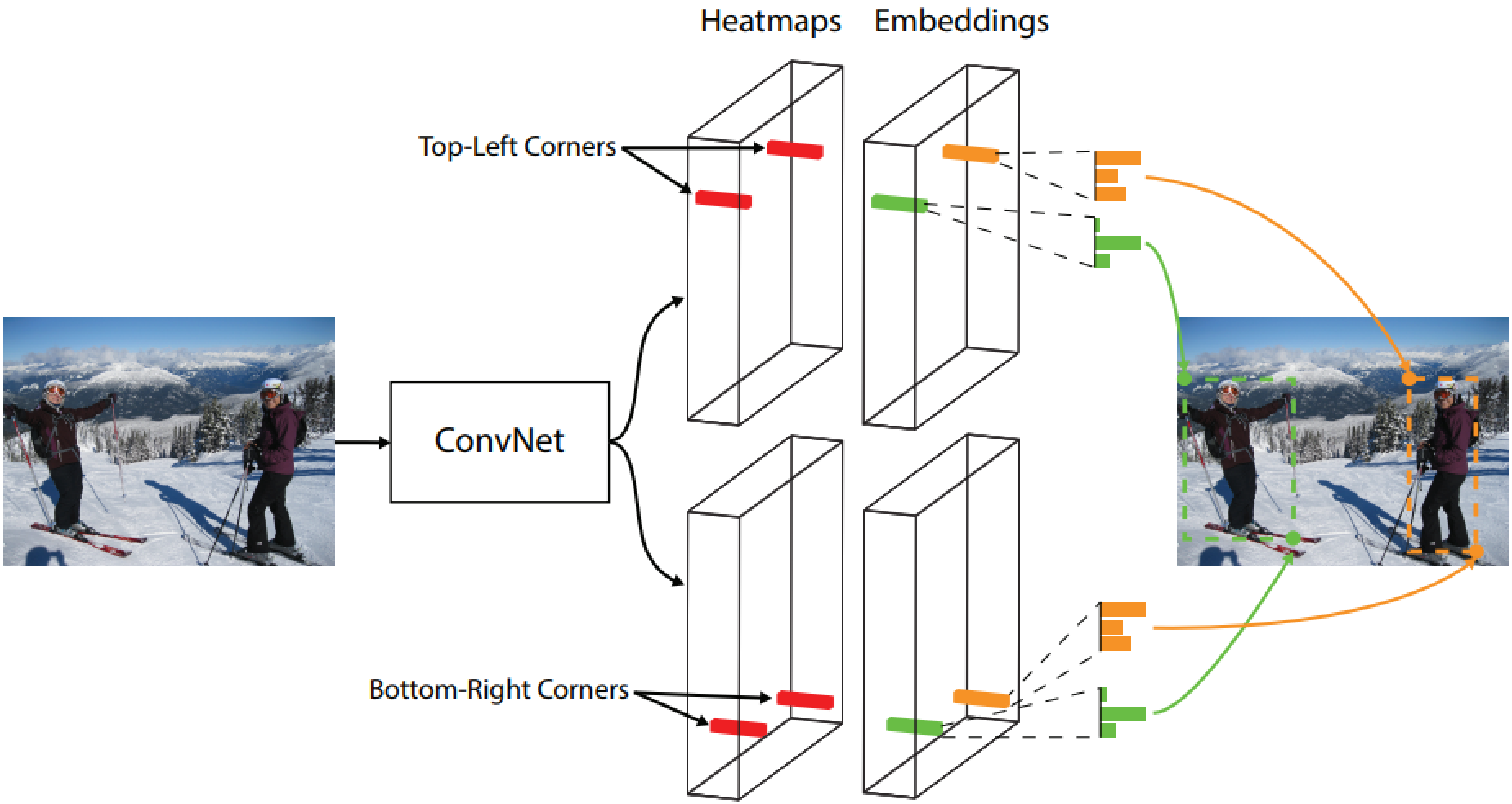
Slide Credit: Ross Girshick



Huang et al. Speed/Accuracy Tradeoffs for Modern Convolutional Object Detectors. CVPR 2017

Slide Credit: Ross Girshick

Regression-Based Techniques



Detect an object as a pair of bounding box corners grouped together

Slide Credit: Ross Girshick

State-of-the-art detectors

- Improve with more data
- Improve with increased model capacity
- Improve from transfer learning
- Immediately benefit from image classification research