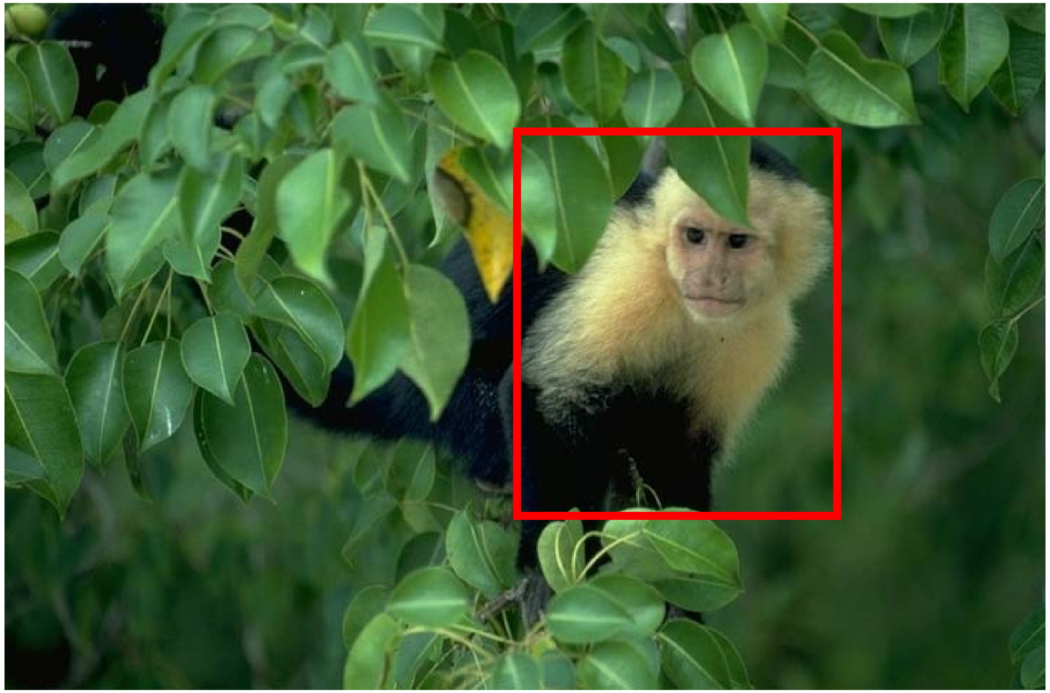
Category-level localization

Cordelia Schmid

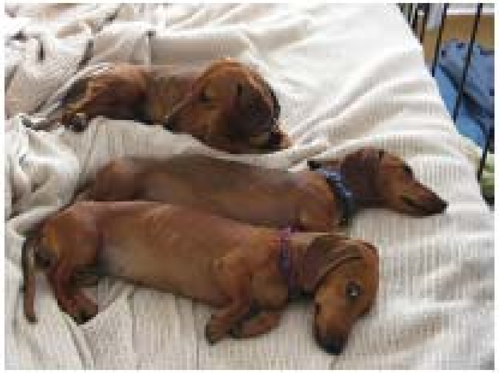
Recognition

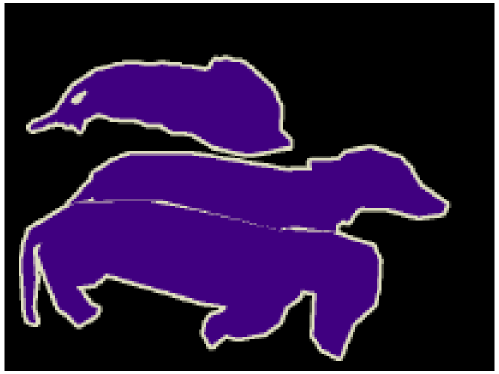
- Classification
  - Object present/absent in an image
  - Often presence of a significant amount of background clutter

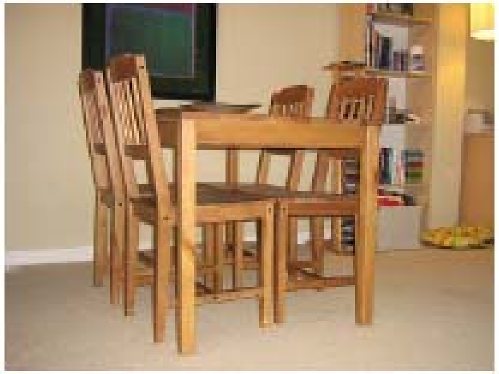
- Localization / Detection
  - Localize object within the frame
  - Bounding box or pixel-level segmentation

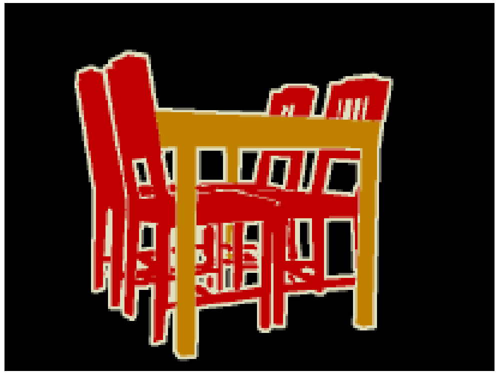


Pixel-level object classification



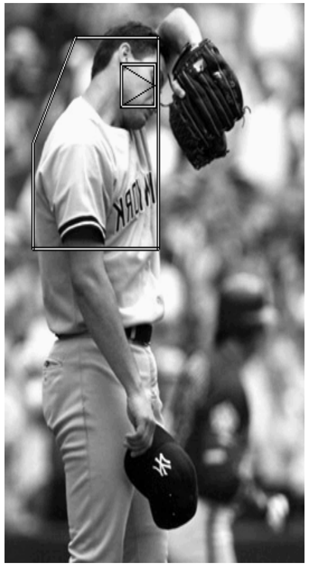


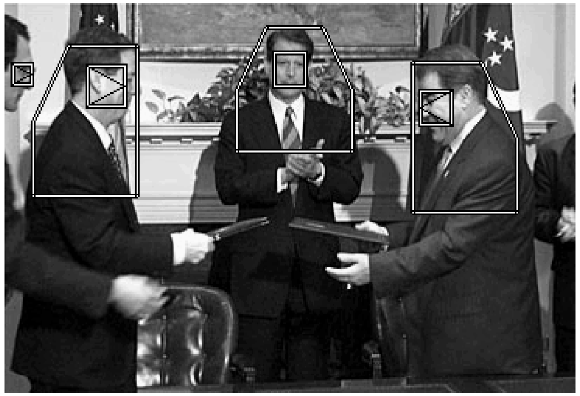


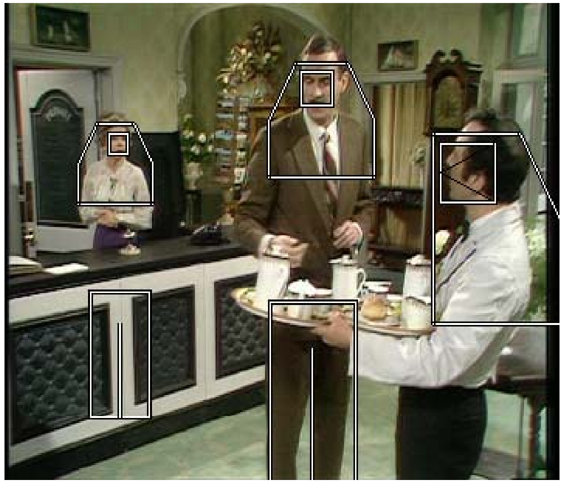


Difficulties

- Intra-class variations







- Scale and viewpoint change
- Multiple aspects of categories

Approaches

- Intra-class variation

  => Modeling of the variations, mainly by learning from a large dataset, for example by SVMs

- Scale + limited viewpoints changes

  => multi-scale approach or invariant local features

- Multiple aspects of categories

  => separate detectors for each aspect, front/profile face, build an approximate 3D “category” model

Approaches

- Localization (bounding box)
  - Hough transform
  - Sliding window approach

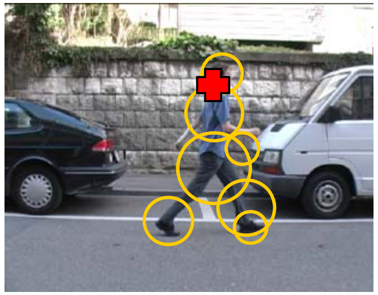
- Localization (segmentation)
  - Shape based
  - Pixel-based +MRF
  - Segmented regions + classification

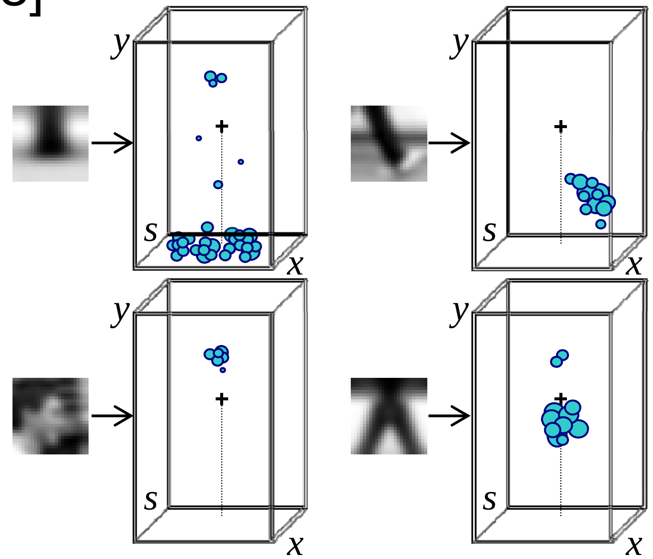
Hough voting

- Use Hough space voting to find objects of a class
- Implicit shape model [Leibe and Schiele ’03,’05]

Learning

- Learn appearance codebook
  - Cluster over interest points on training images
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
  - Centroid + scale is given





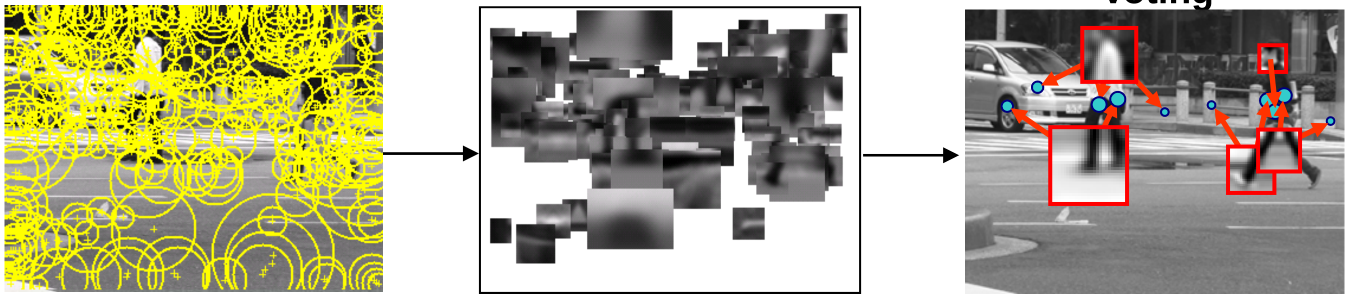
Spatial occurrence distributions

Recognition

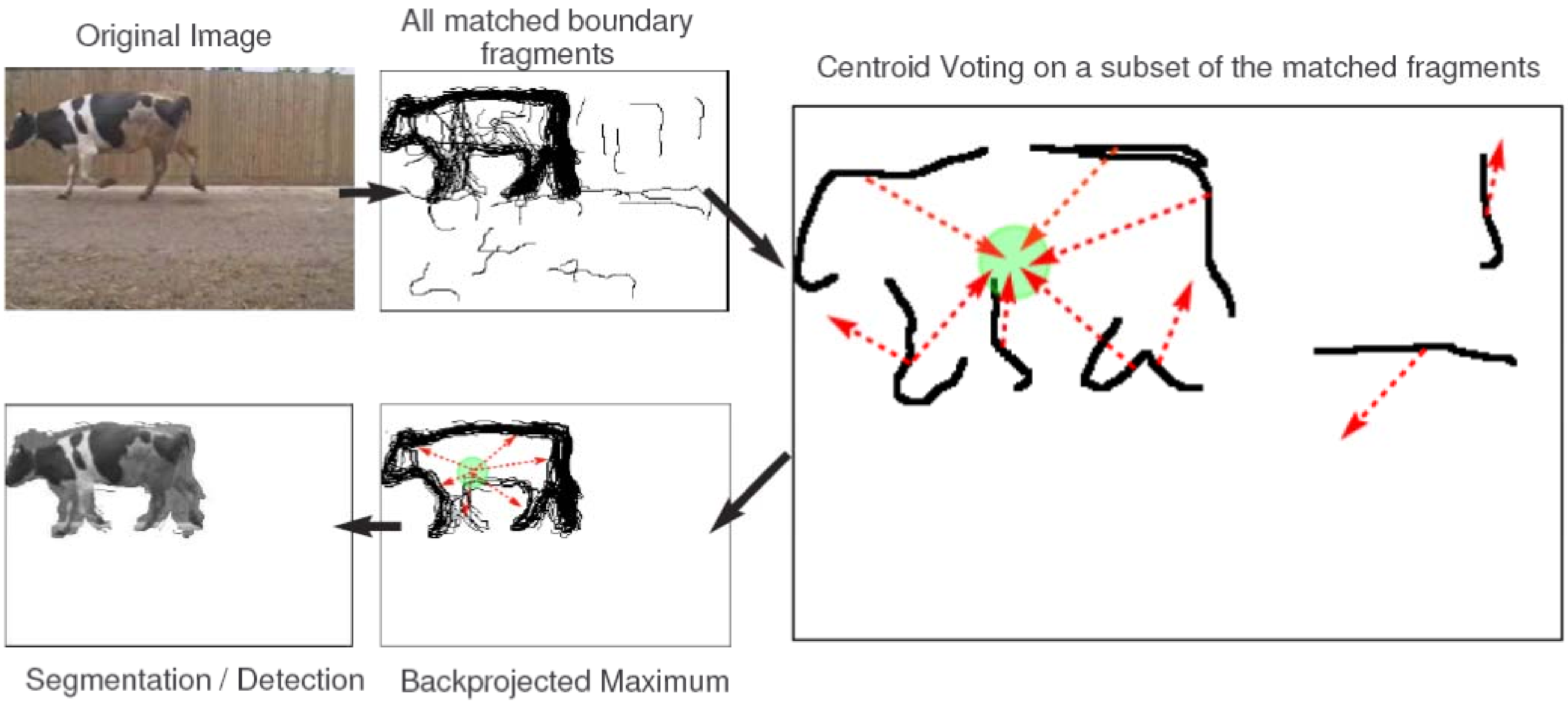
Interest Points

Matched Codebook Entries

Probabilistic Voting



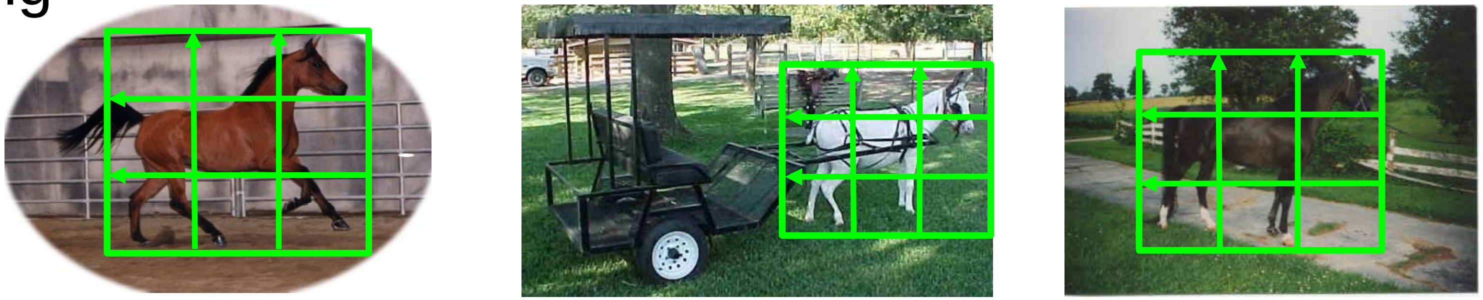
Hough voting


Original Image
All matched boundary fragments
Centroid Voting on a subset of the matched fragments
Segmentation / Detection
Backprojected Maximum

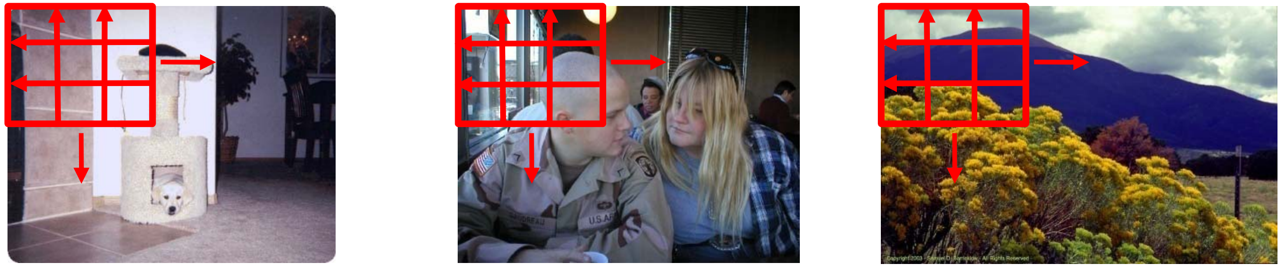
[Opelt, Pinz,Zisserman, ECCV 2006]

Localization with sliding window

Training



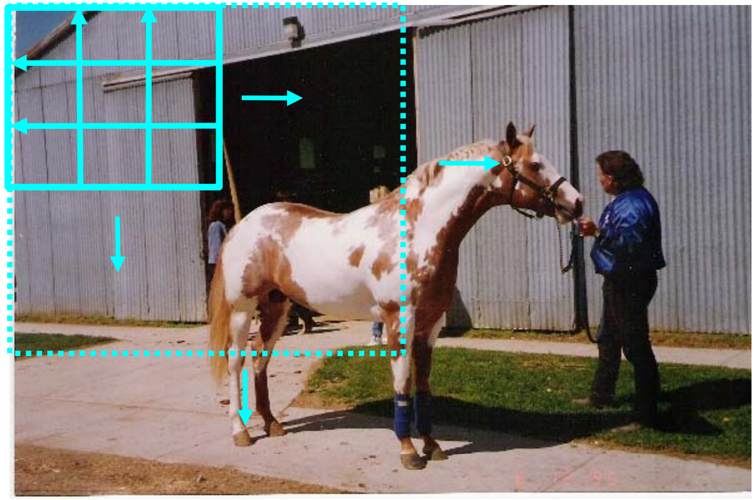
Positive examples



Negative examples

Description + Learn a classifier

Localization with sliding window

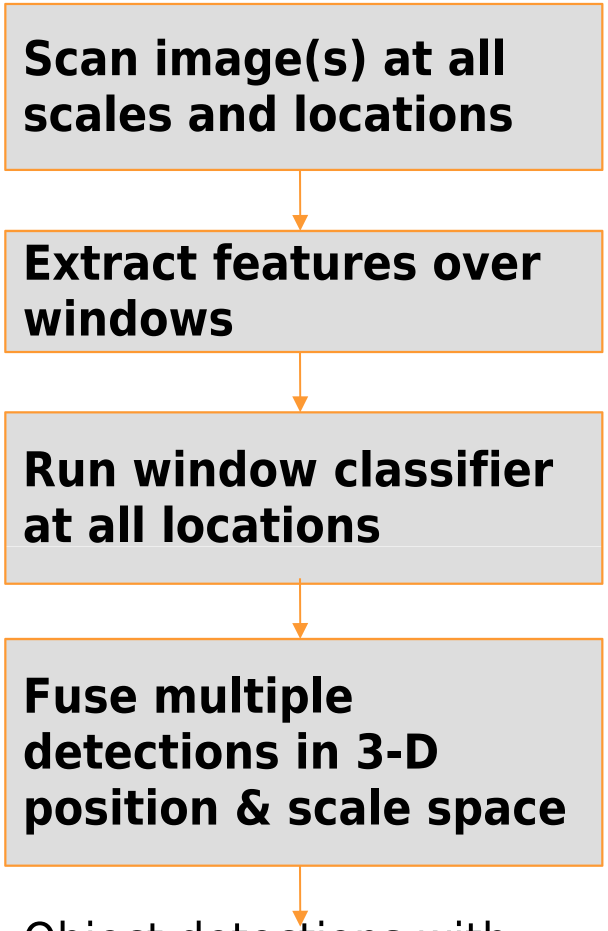


Testing at multiple locations and scales

Find local maxima, non-maxima suppression

Sliding Window Detectors

Detection Phase



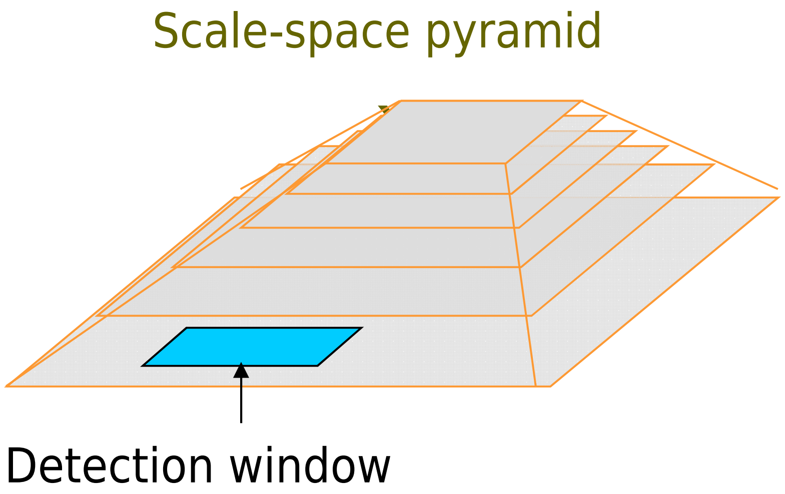
Scan image(s) at all scales and locations

Extract features over windows

Run window classifier at all locations

Fuse multiple detections in 3-D position & scale space

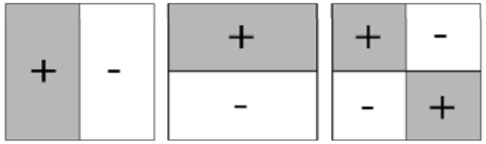
Object detections with bounding boxes



Haar Wavelet / SVM Human Detector

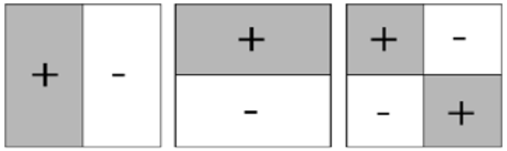
Training set (2k positive / 10k negative)

Haar wavelet descriptors



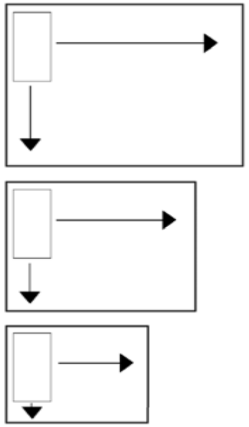
training

1326-D descriptor



descriptors

Test image

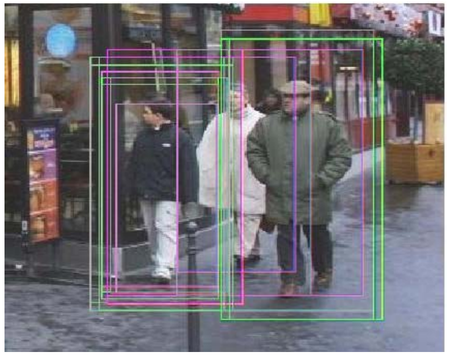


Multi-scale search

test

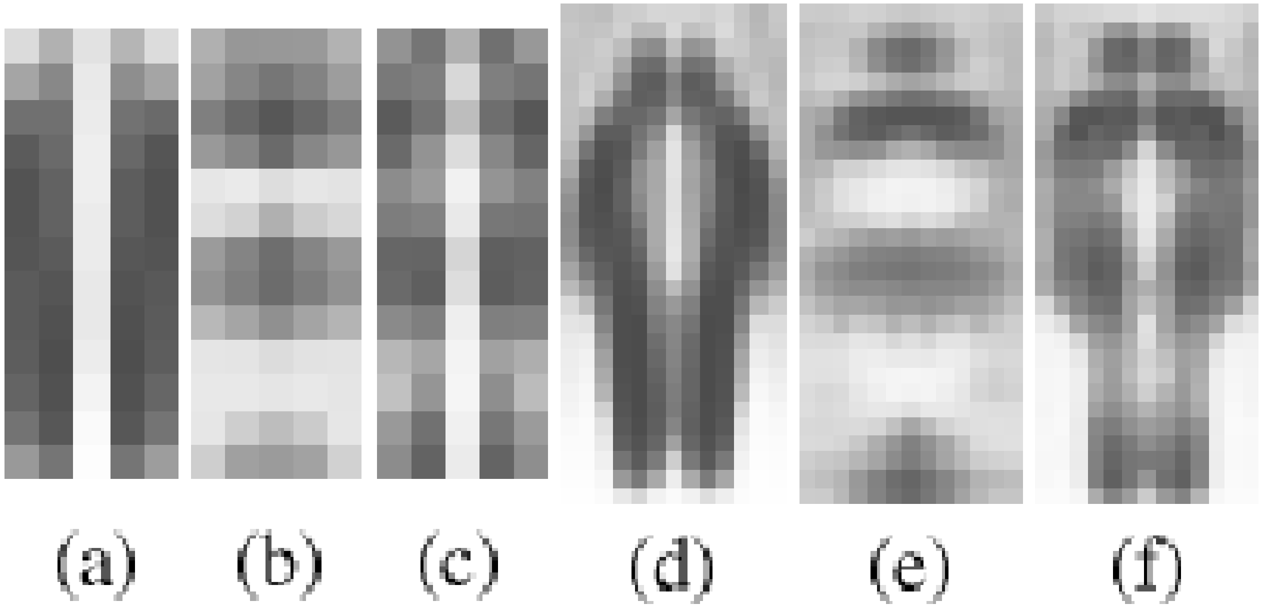
Support vector machine

results



[Papageorgiou & Poggio, 1998]

Which Descriptors are Important?

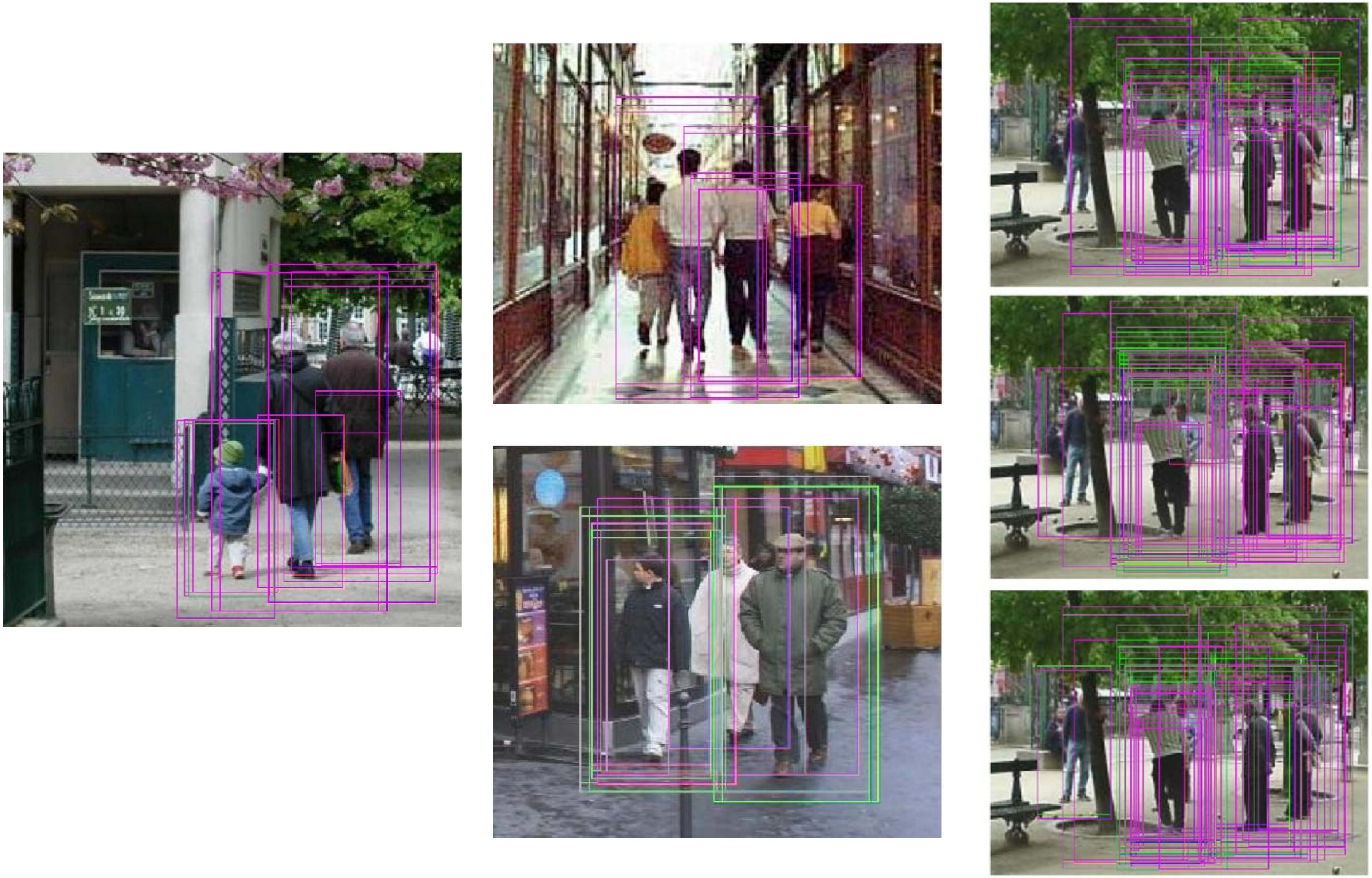


32x32 descriptors   *16x16 descriptors*

Mean response difference between positive & negative training examples

Essentially just a coarse-scale human silhouette template!

Some Detection Results



PASCAL VOC dataset - localization

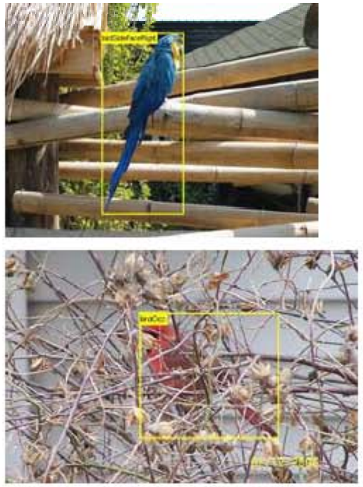
- 20 object classes (aeroplane, bicycle, bird, etc.)
- Bounding box annotations for training and evaluation
- Viewpoint information : front, rear, left, right, unspecified
- Other information : truncated, occluded, difficult

PASCAL dataset

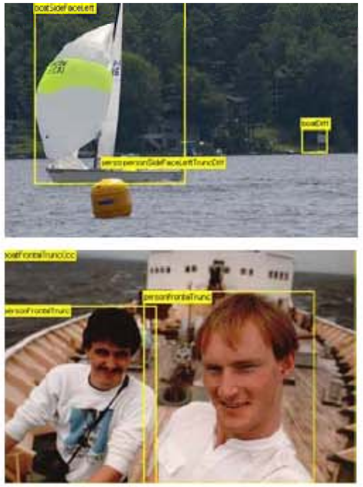
Aeroplane

Bicycle

Bird



Boat



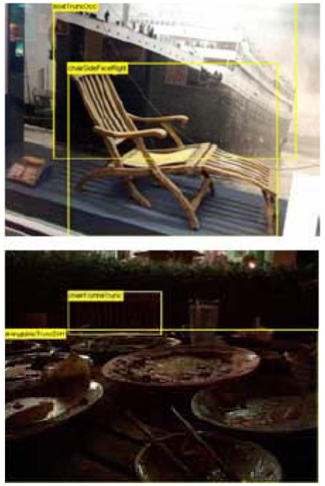
Bottle

Bus

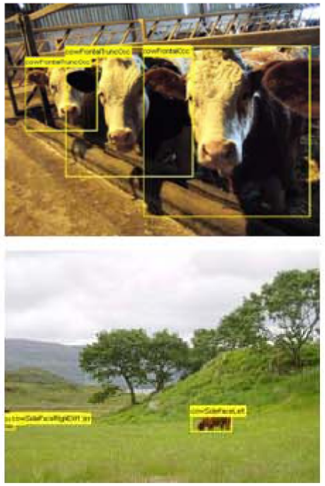
Car

Cat

Chair

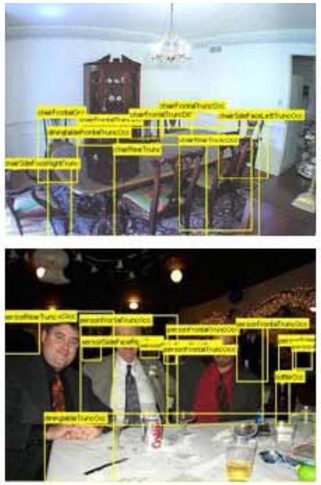


Cow



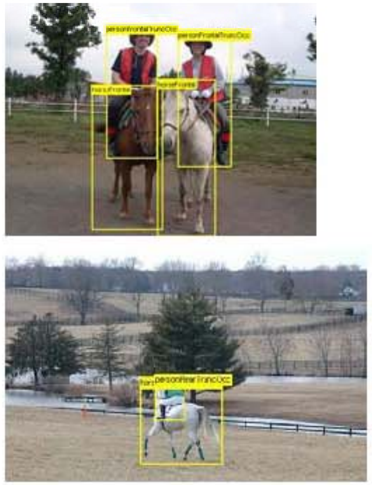
PASCAL dataset

Dining Table



Dog

Horse

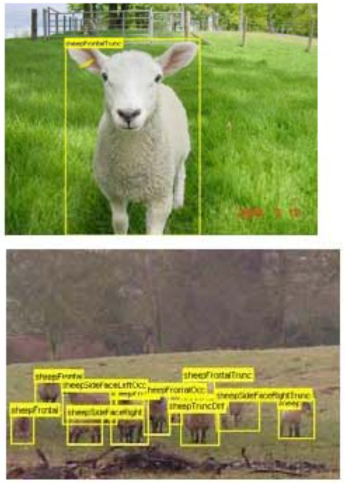


Motorbike

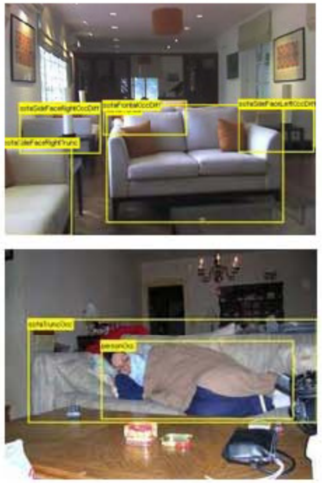
Person

Potted Plant

Sheep

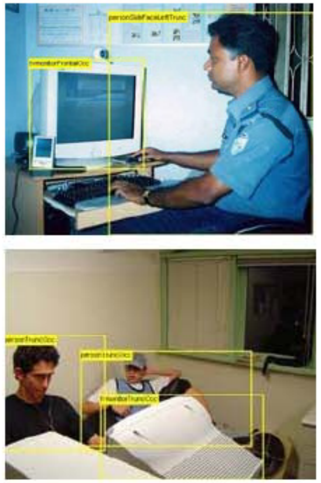


Sofa



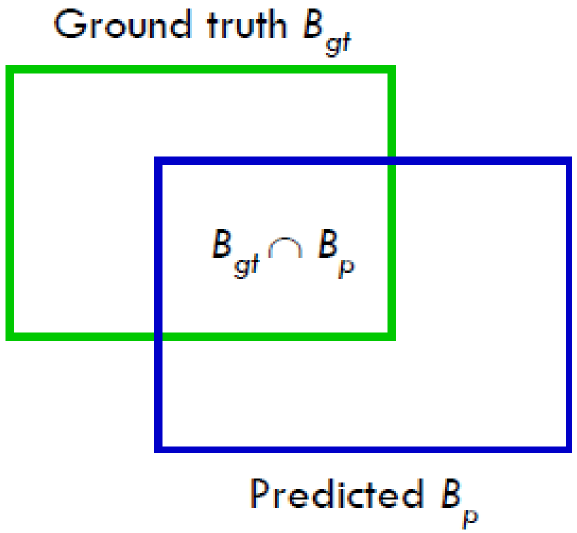
Train

TV/Monitor



Evaluating localization with bounding boxes

- Area of Overlap (AO) Measure

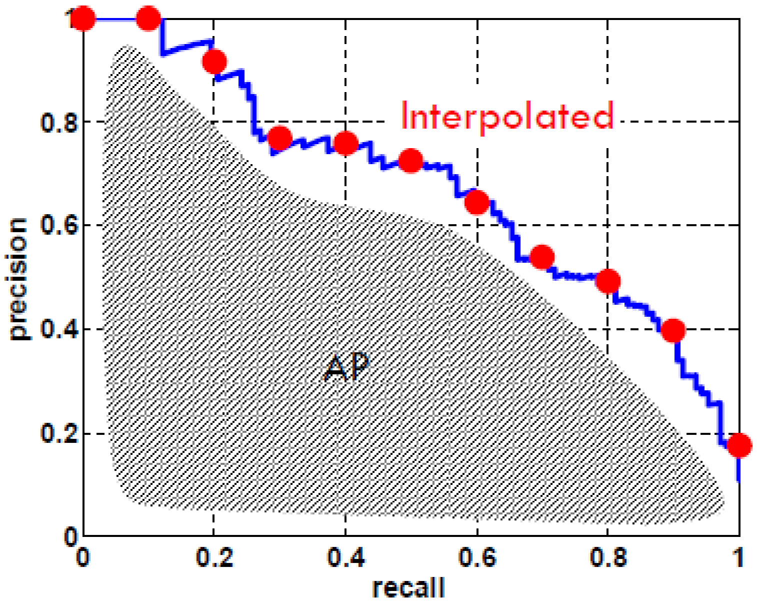


$$AO(B_{gt}, B_p) = \frac{|B_{gt} \bigcap B_p|}{|B_{gt} \bigcup B_p|}$$

- Need to define a threshold t such that $AO(B_{gt}, B_p)$ implies a correct detection: 50%

Evaluating localization with bounding boxes

- Average Precision [TREC] averages precision over the entire range of recall
  - Curve interpolated to reduce influence of "outliers"



- A good score requires both high recall and high precision
- Application-independent
- Penalizes methods giving high precision but low recall