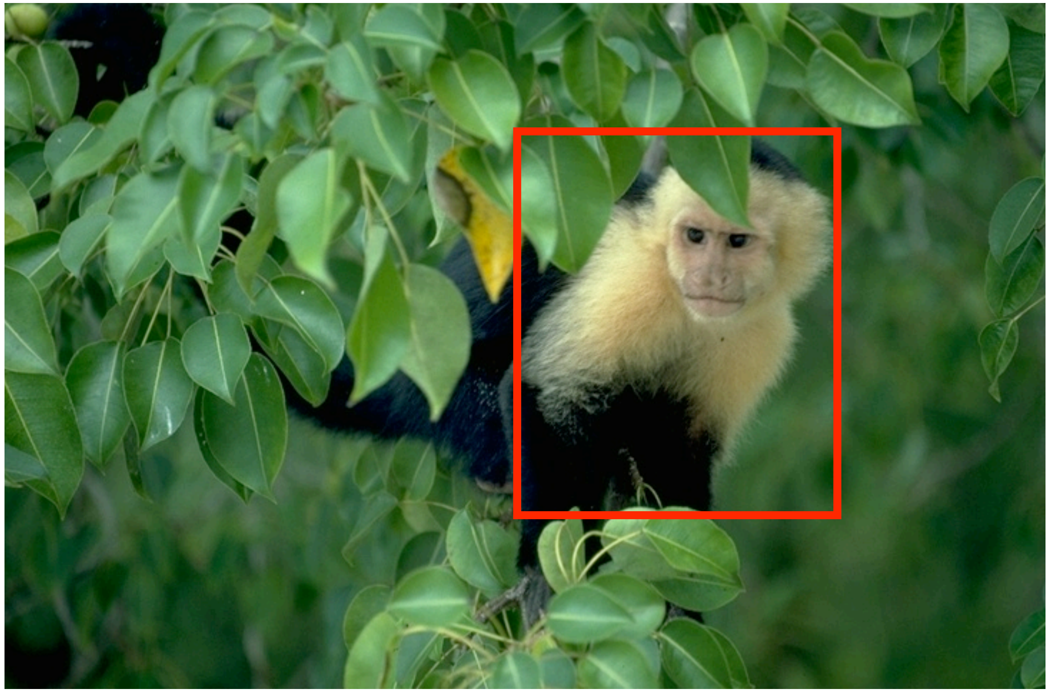
Category-level localization

Cordelia Schmid

Recognition

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

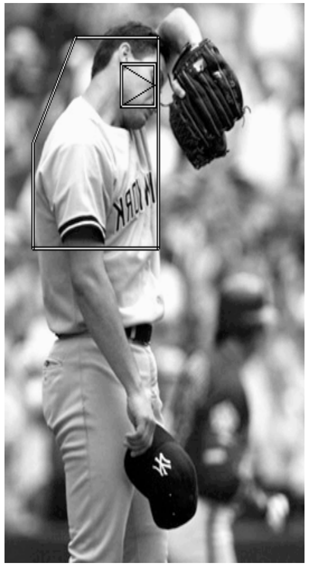
- Localization / Detection
  - Localize object within the frame
  - Bounding box or pixel-level segmentation

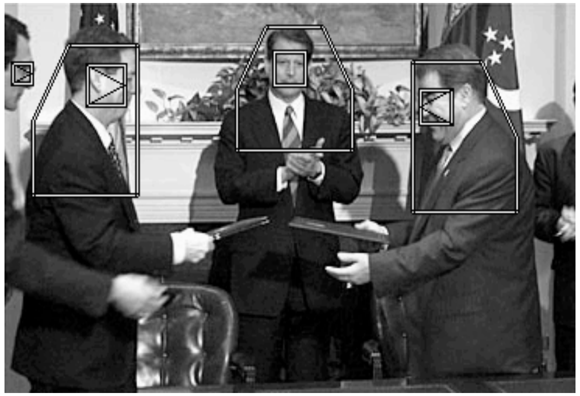


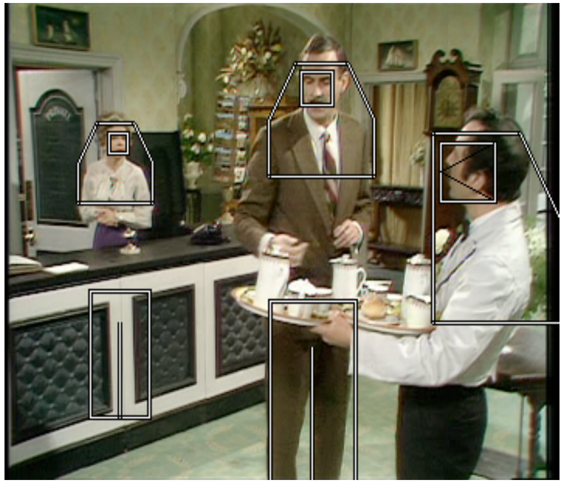
Pixel-level object classification

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

  => 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
  - Shape voting
  - Shape exemplars
  - Sliding window approach

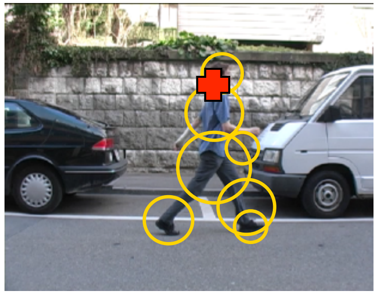
- Localization (segmentation)
  - Shape based
  - Pixel-based +MRF
  - Segmented regions + classification

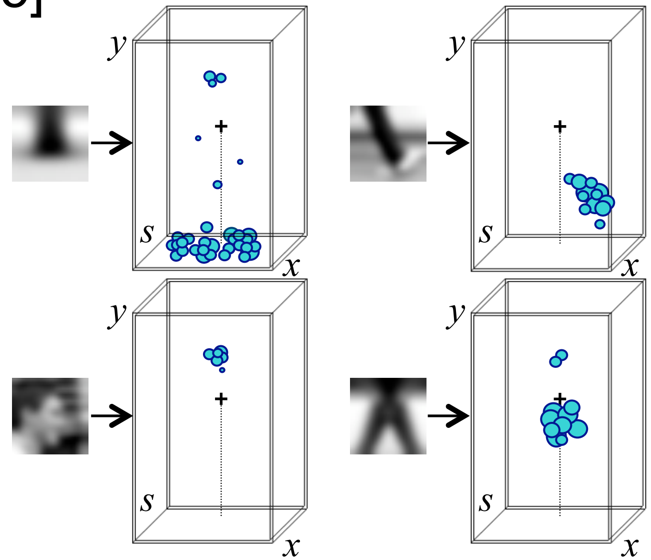
Hough voting

- 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

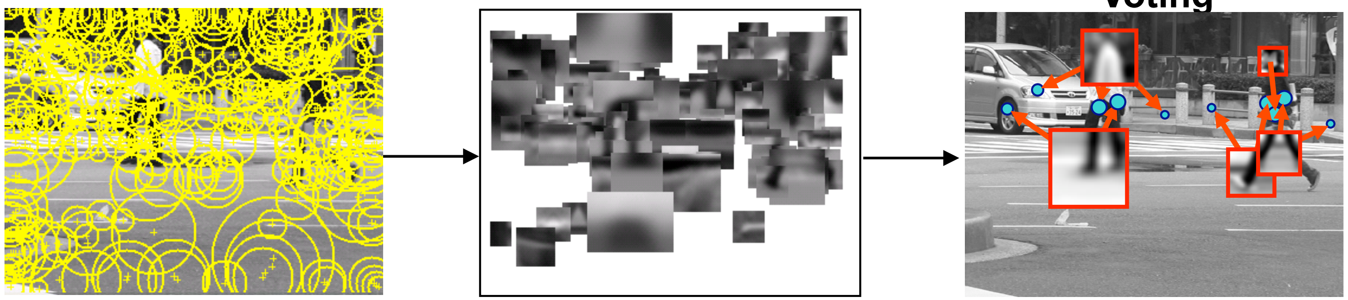




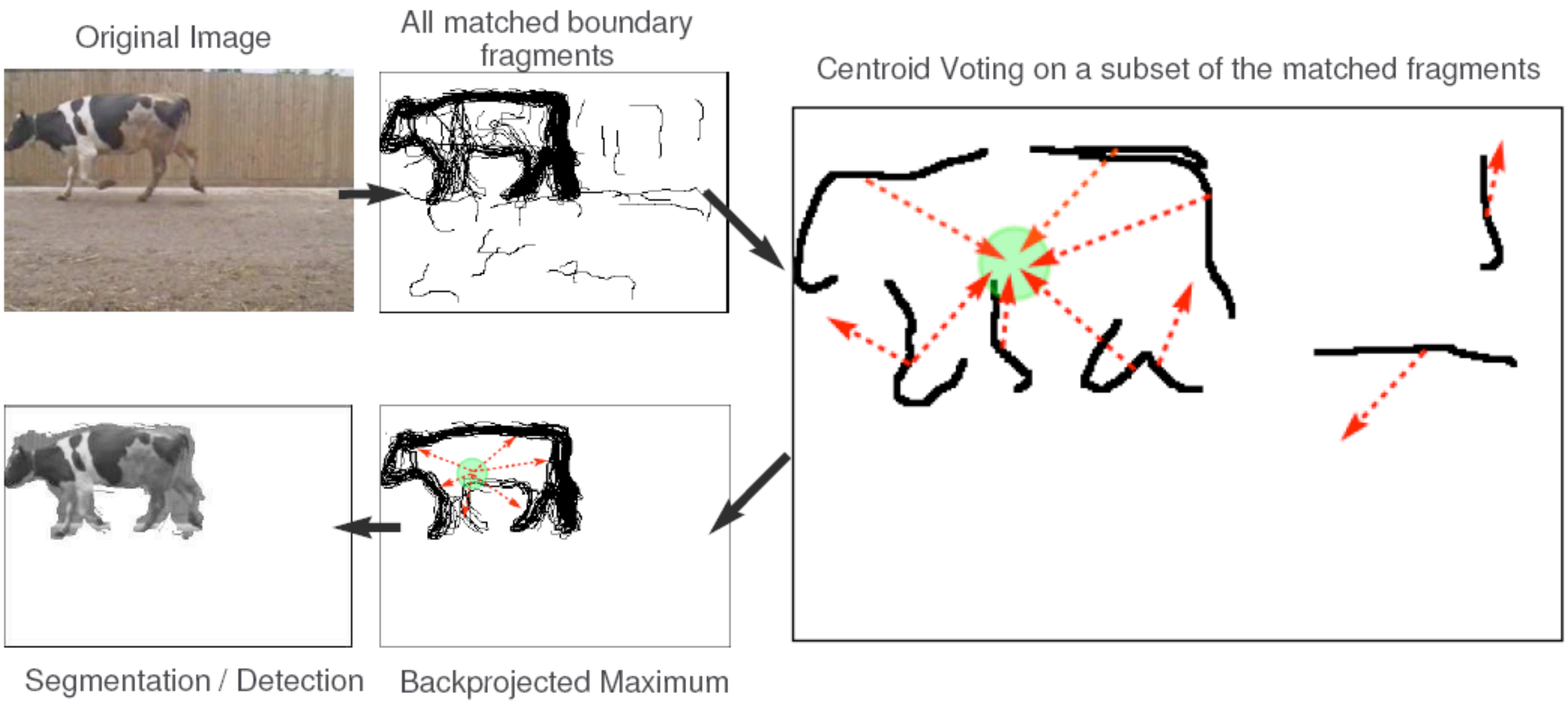
Spatial occurrence distributions

Recognition

Interest Points **Matched Codebook Entries** **Probabilistic Voting**



Hough voting



[Opelt, Pinz,Zisserman, ECCV 2006]

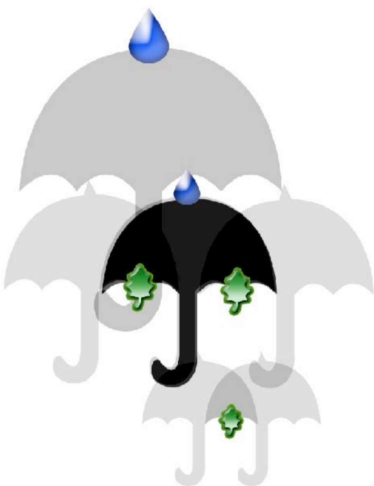
Masks for object localization

For each test feature:

- Select closest training features + corresponding masks
  (training requires images with shape outline)

- Align mask based on local co-ordinates system
  (transformation between training and test co-ordinate systems)

Sum masks weighted by matching distance



three features agree on object localization,
the object has higher weights

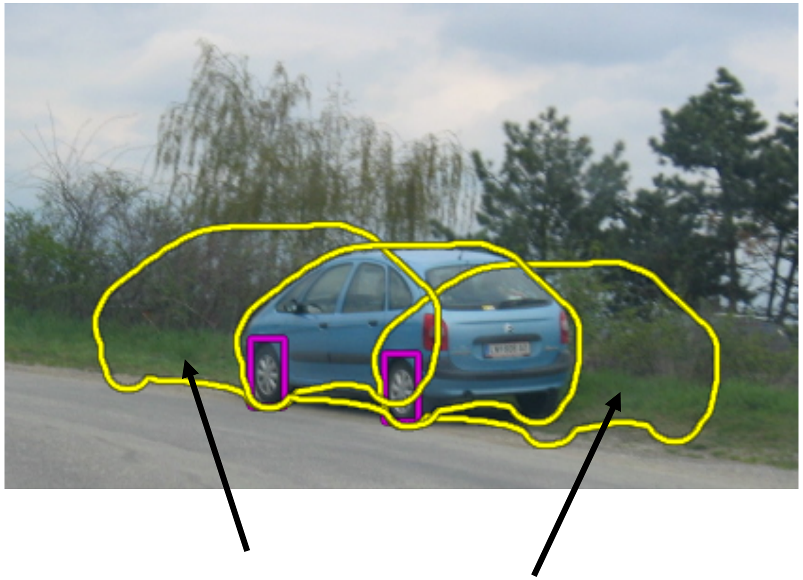
[Marszalek & Schmid, CVPR 2007]

Examples of “summed” masks

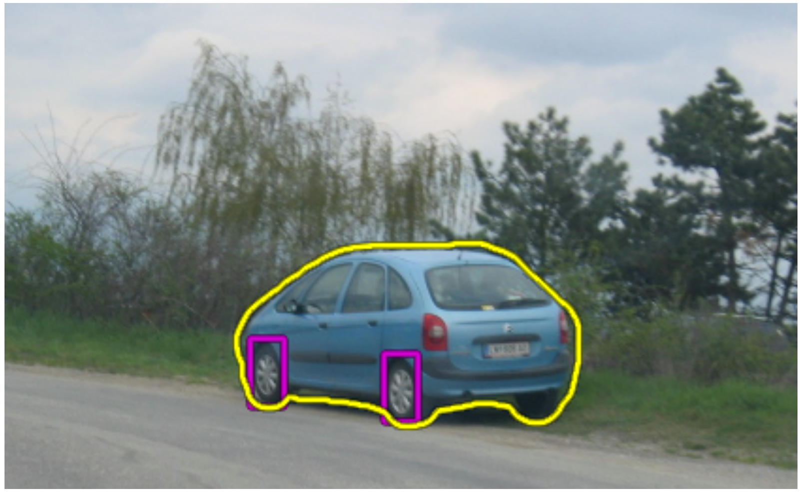
Object localization

- Cast hypothesis
  - Aligning the mask based on matching features
- Evaluate each hypothesis
  - SVM for local features
- Merge hypothesis to produce localization decisions
  - Online clustering of similar hypothesis, rejection of weak ones

Illustration of hypothesis evaluation



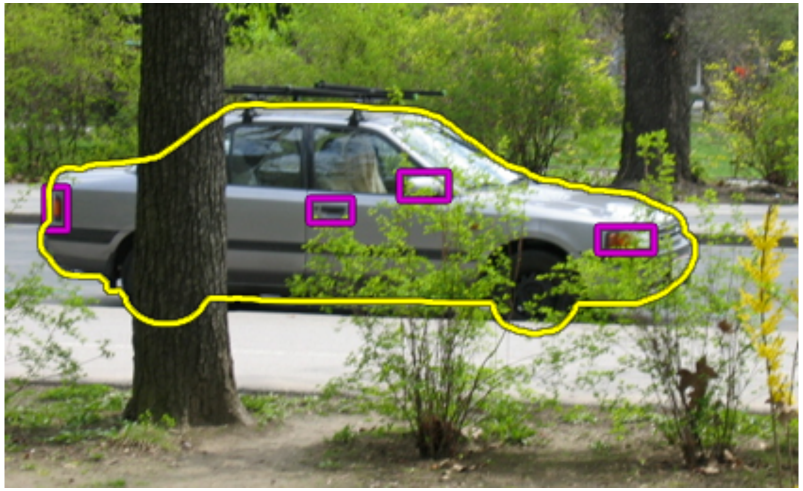
False hypotheses due to the
ambiguities of the wheels



Eliminated after the evaluation

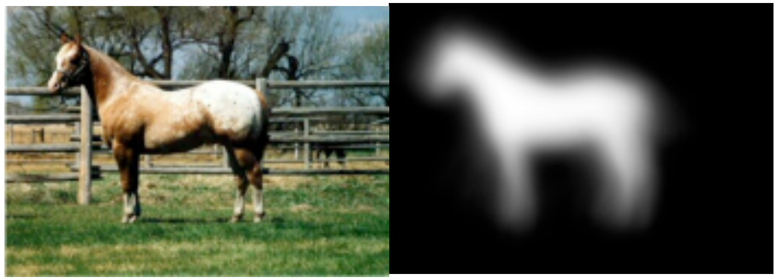
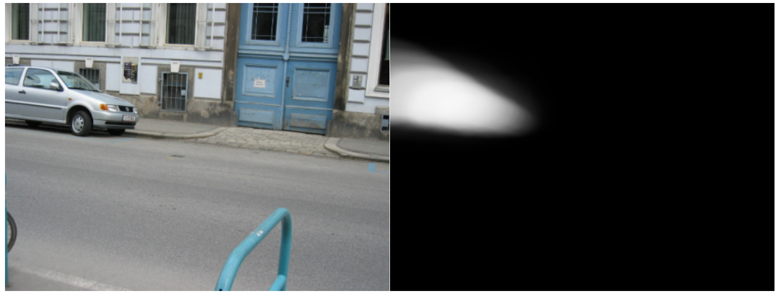
Illustration of hypotheses merging

Weak classifier response
due to occlusion



Merging of evidence based on
consistent object features

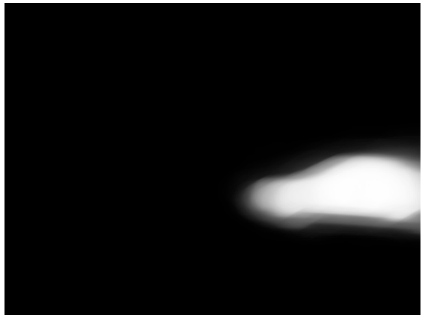
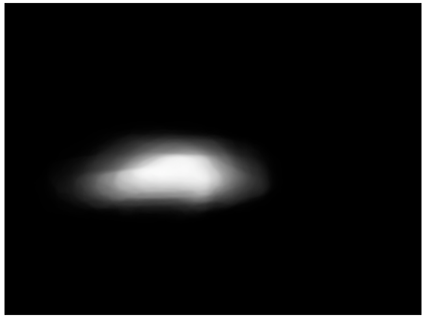
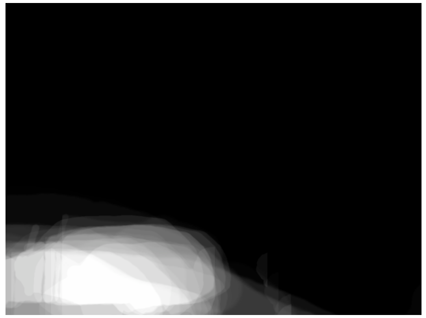
Localization results



Localization result

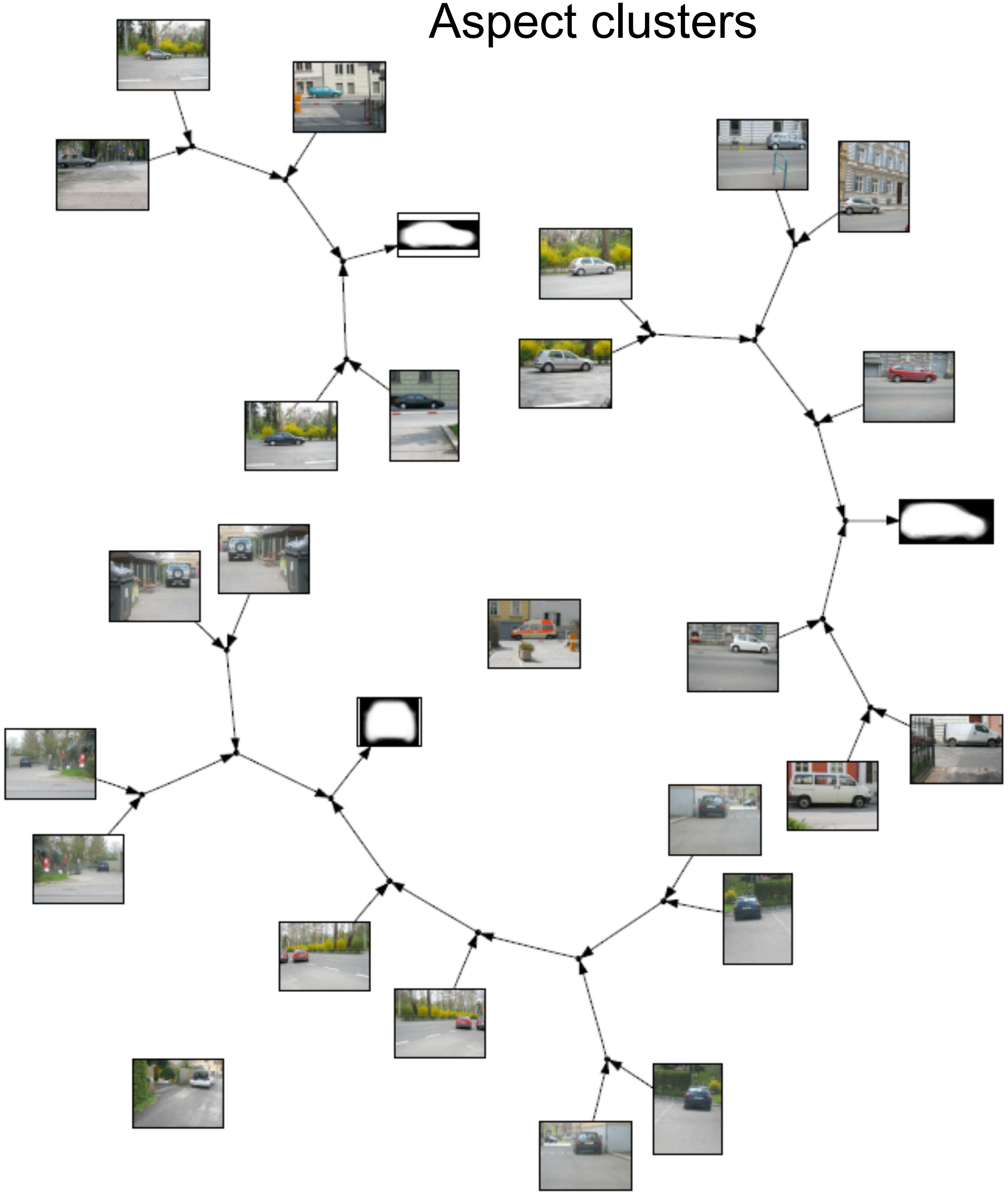
Illustration of subsequent hypotheses

Confidence value	1103.1	561.8	4.9

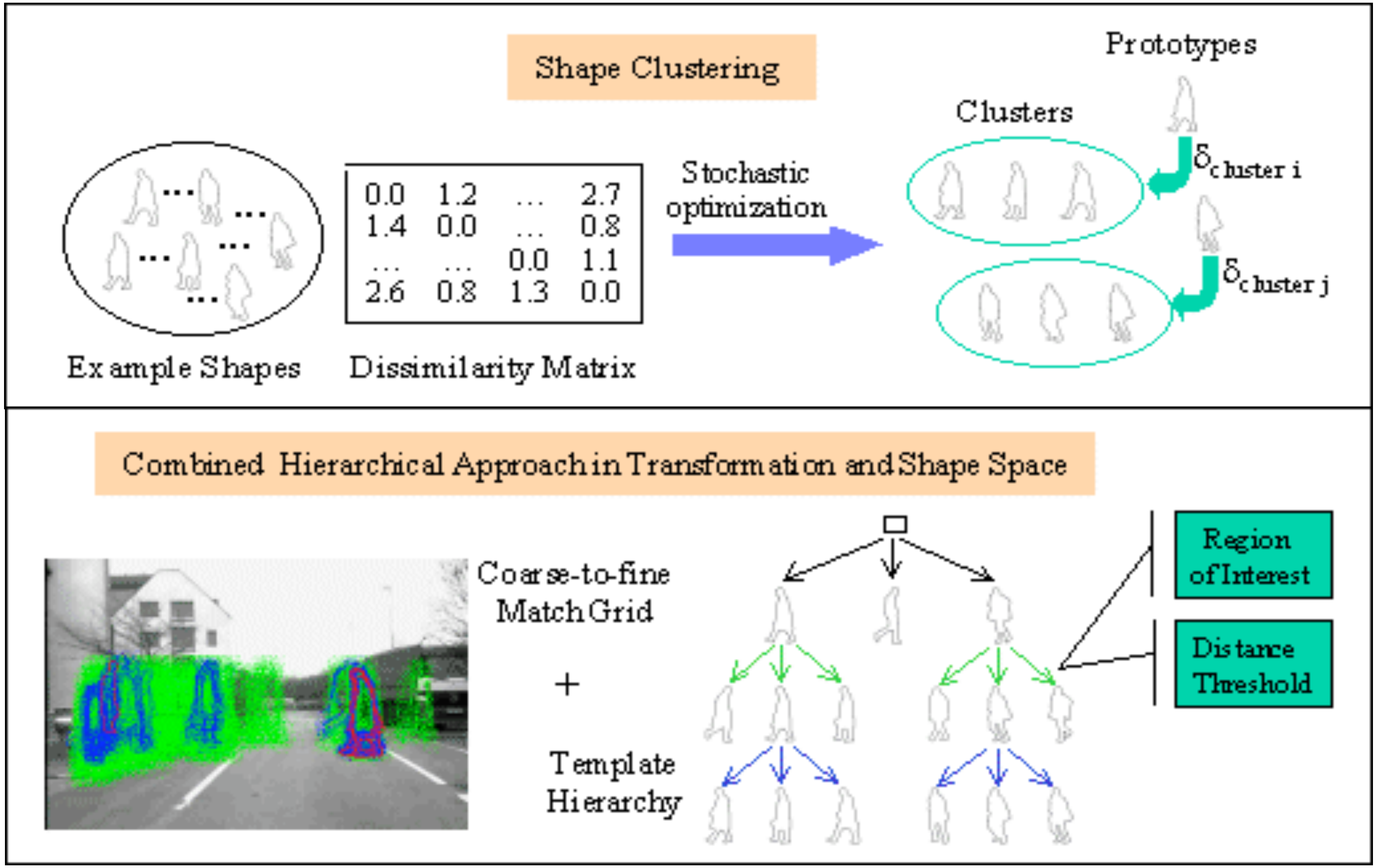



Aspect clusters



Exemplar based Pedestrian Detector

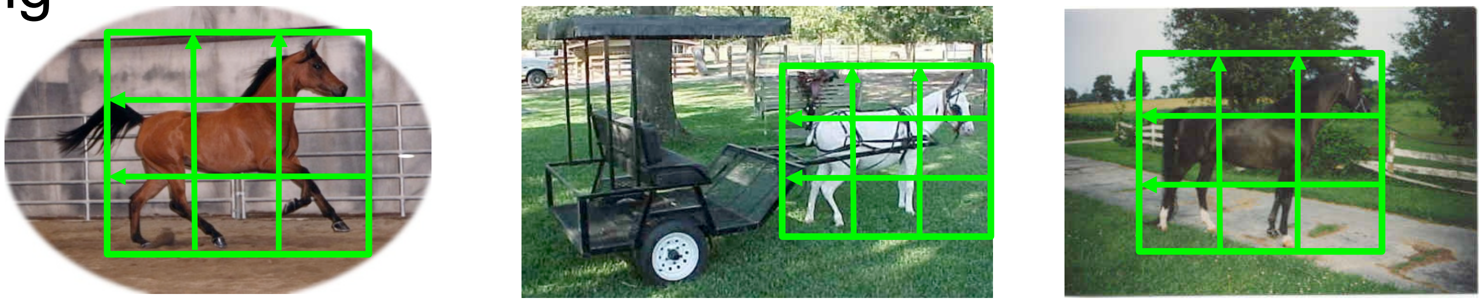
- ***Build model by clustering training examples hierarchically***
- ***At run-time, use similarity tree to find similar examples quickly***



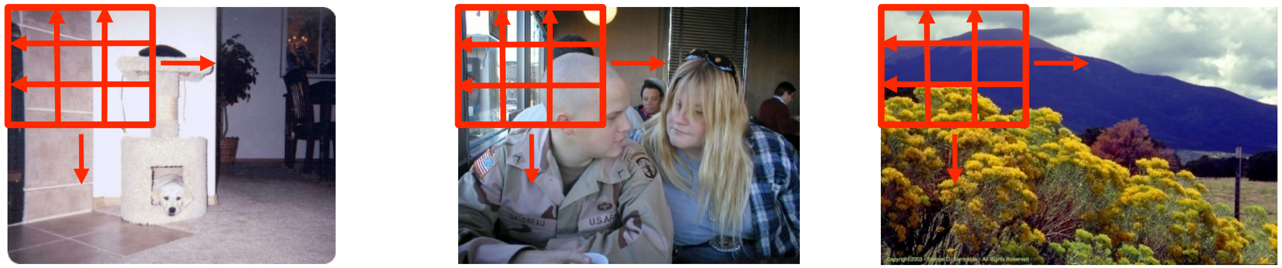
[D.Gavrila, ICPR'98]

Localization with sliding window

Training



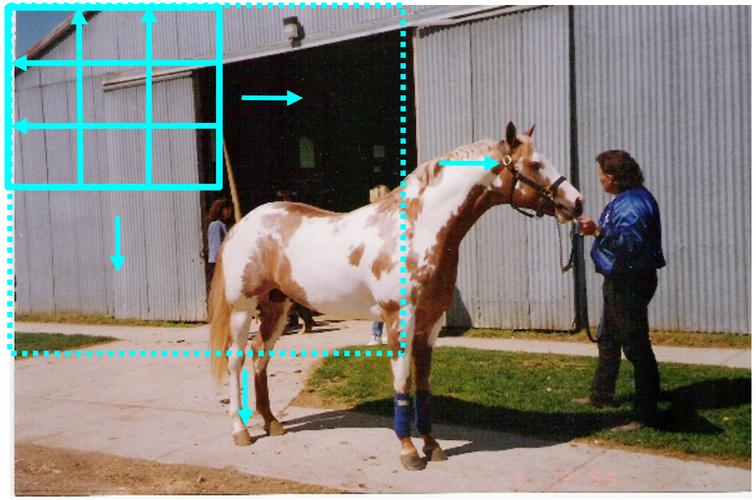
Positive examples



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

↓

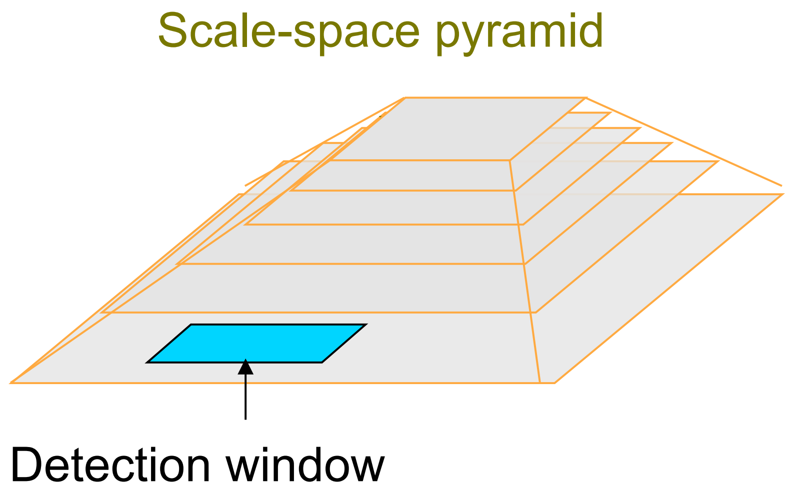
Run window classifier at all locations

↓

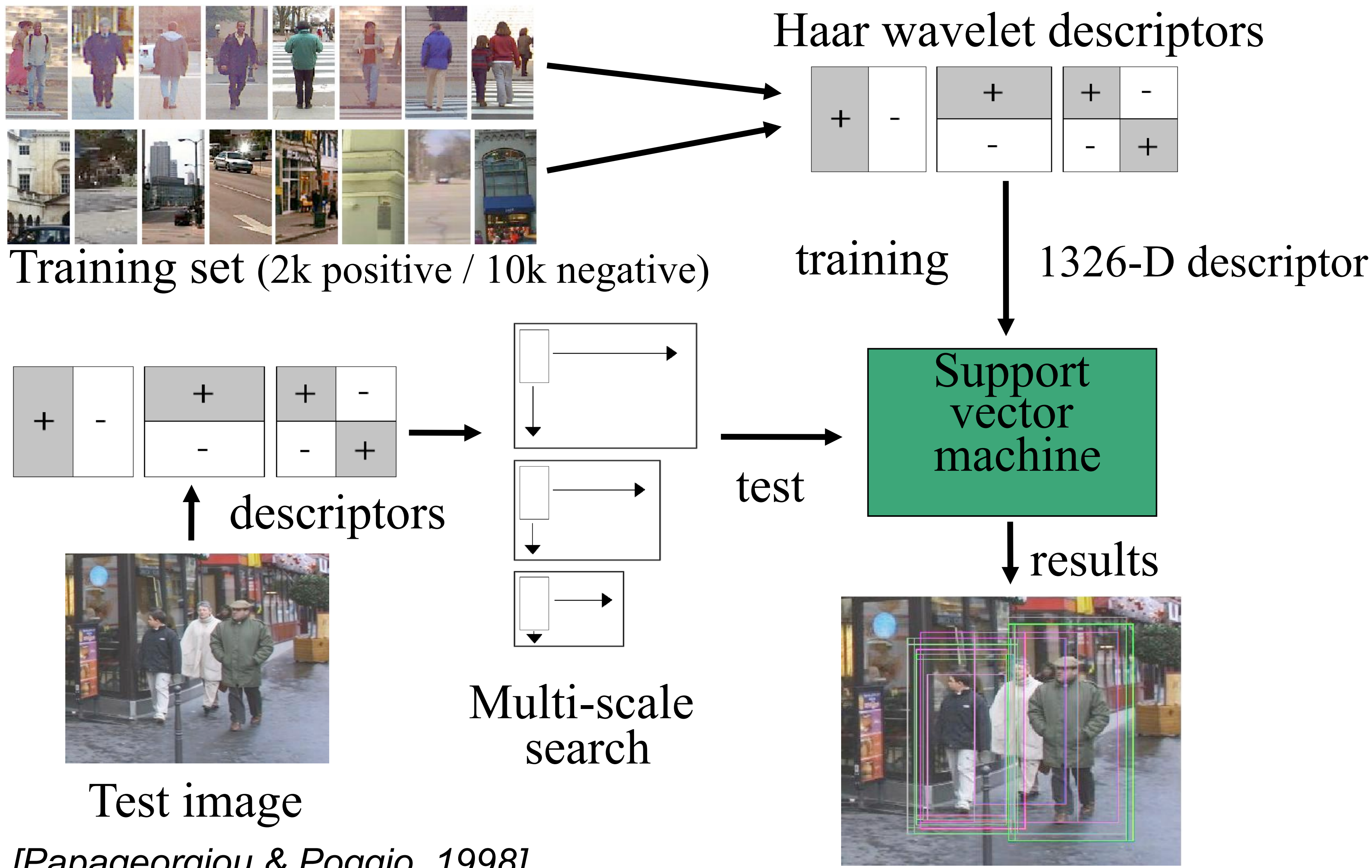
Fuse multiple detections in 3-D position & scale space

↓

Object detections with bounding boxes

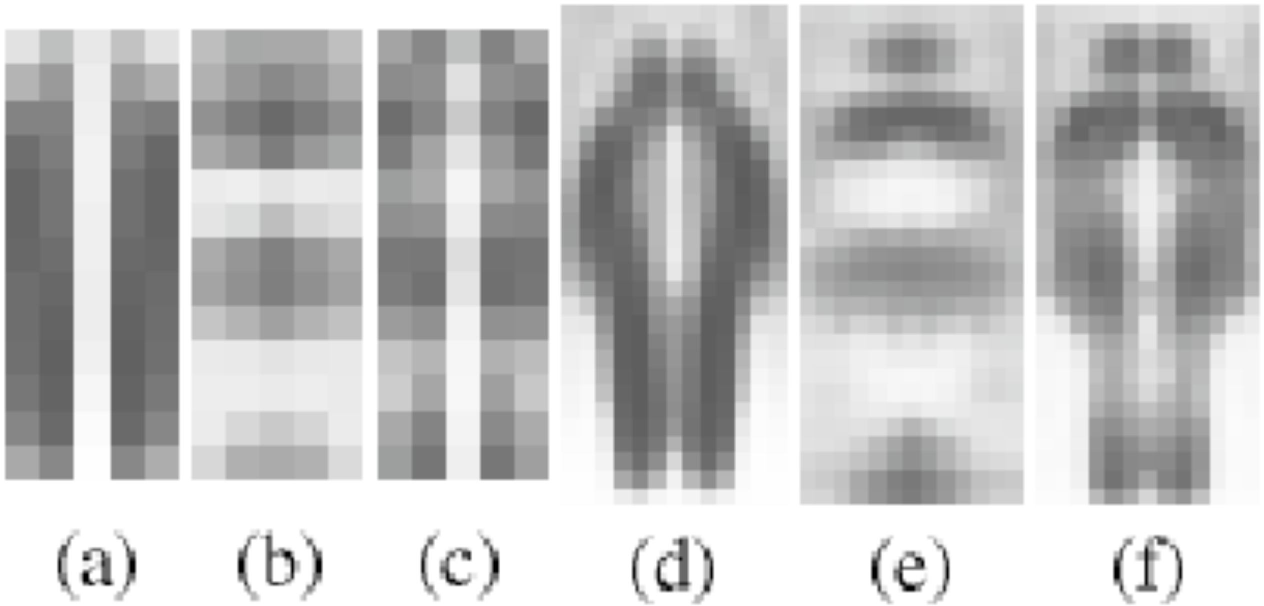


Haar Wavelet / SVM Human Detector



[Papageorgiou & Poggio, 1998]

Which Descriptors are Important?

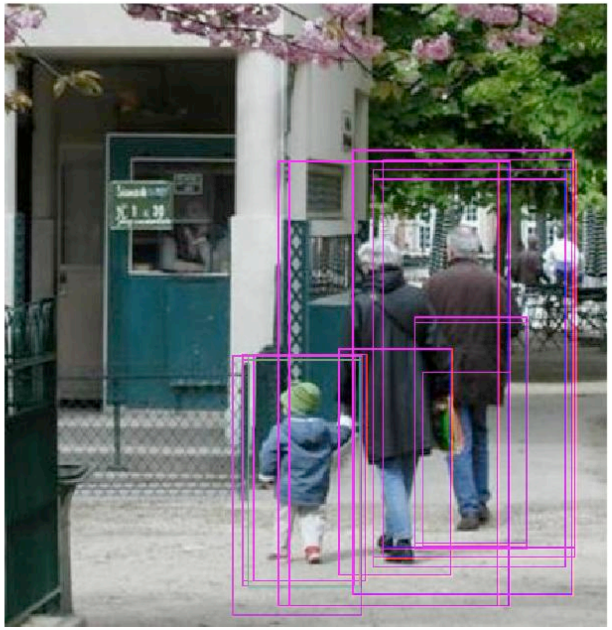


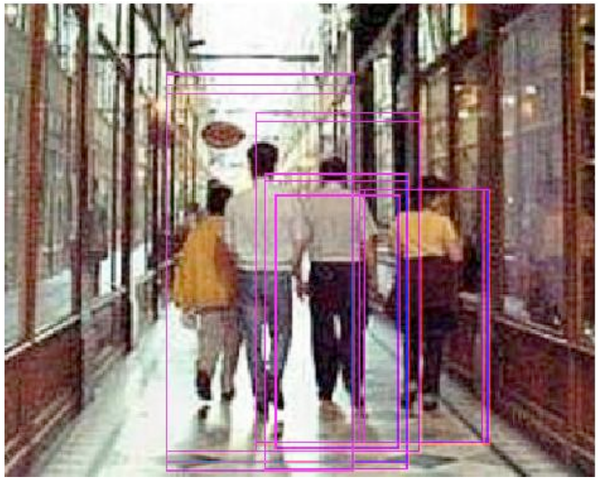
32x32 descriptors *16x16 descriptors*

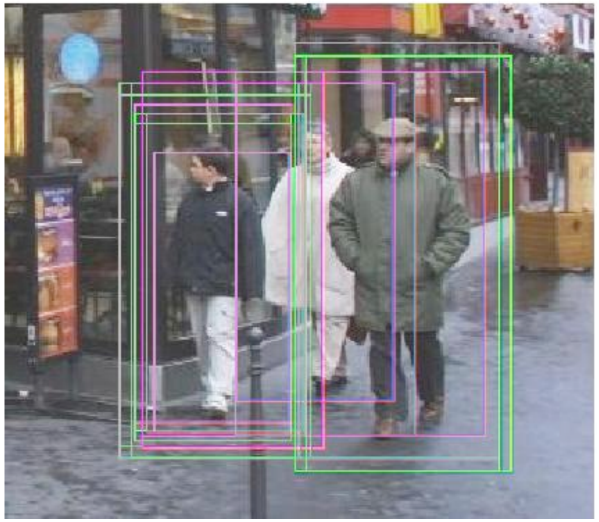
Mean response difference between positive & negative training examples

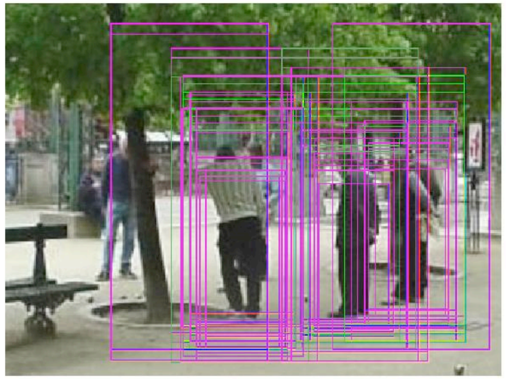
Essentially just a coarse-scale human silhouette template!

Some Detection Results





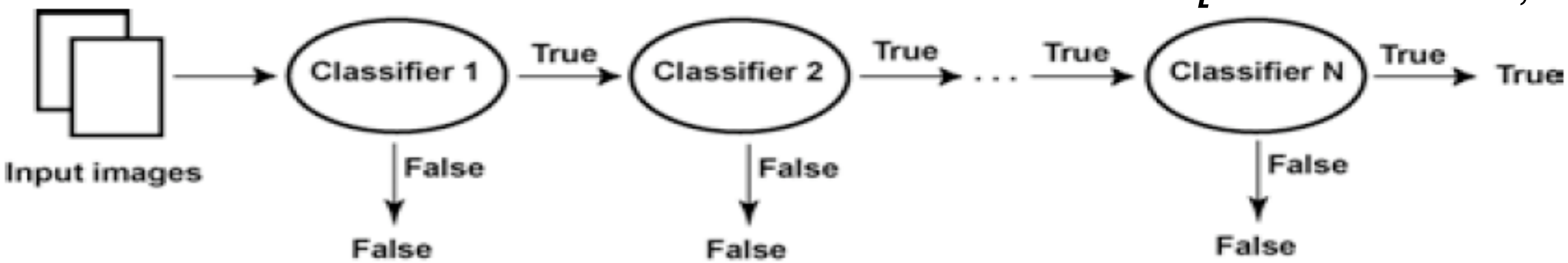


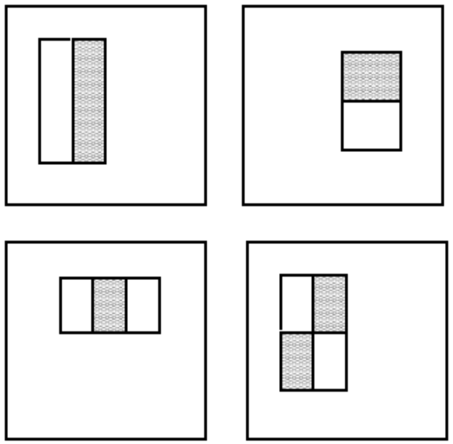


AdaBoost Cascade Face Detector

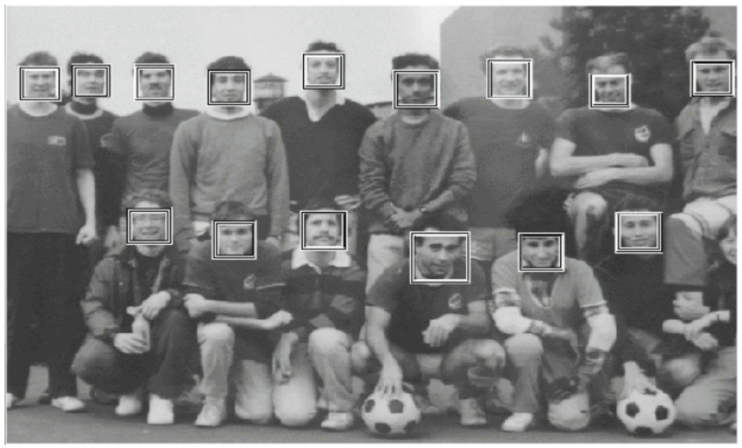
- A computationally efficient architecture that rapidly rejects unpromising windows
  - A chain of classifiers that each reject some fraction of the negative training samples while keeping almost all positive ones
- Each classifier is an AdaBoost ensemble of rectangular Haar-like features sampled from a large pool

[Viola & Jones, 2001]



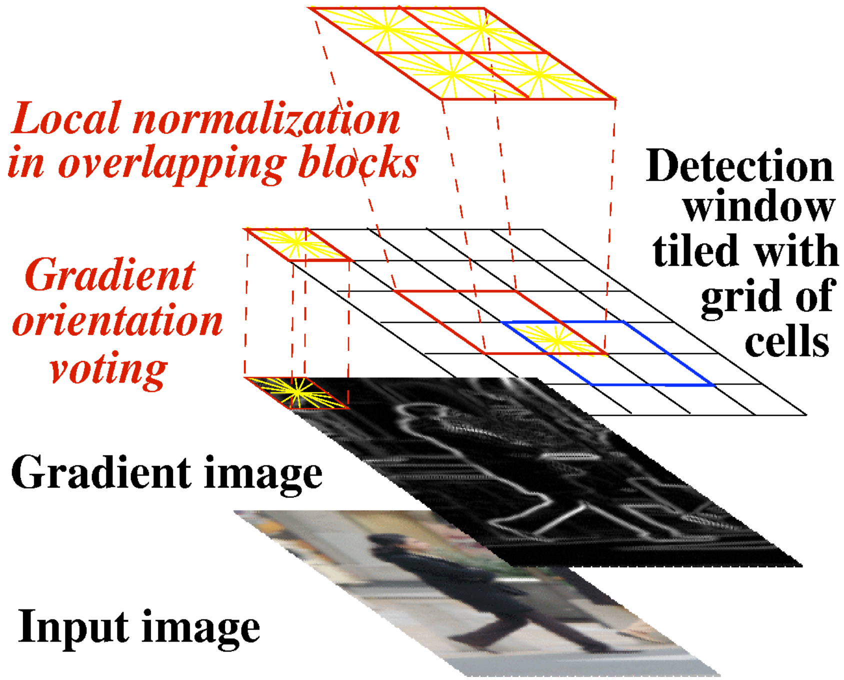


Rectangular Haar features and the first two features chosen by AdaBoost

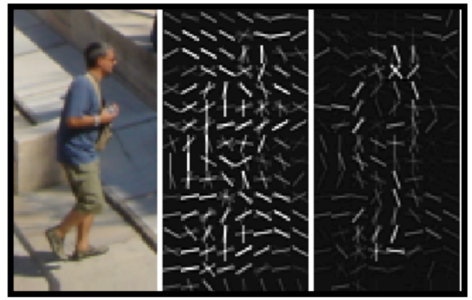


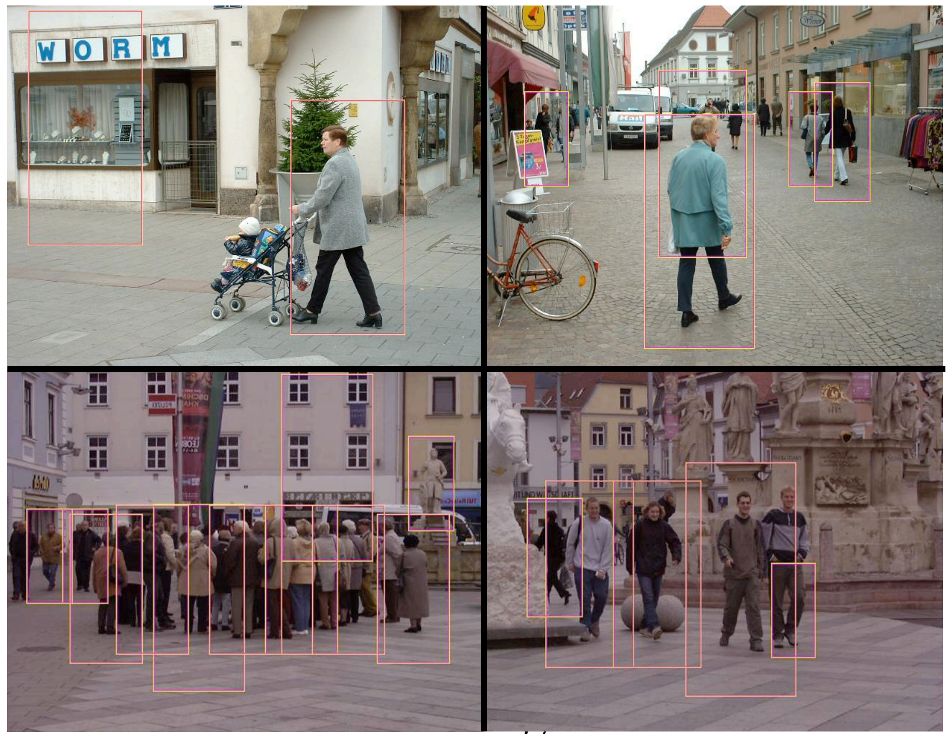
Histogram of Oriented Gradient Human Detector

- Descriptors are a grid of local Histograms of Oriented Gradients (HOG)
- Linear SVM for runtime efficiency
- Tolerates different poses, clothing, lighting and background
- Assumes upright fully visible people



Importance weighted responses



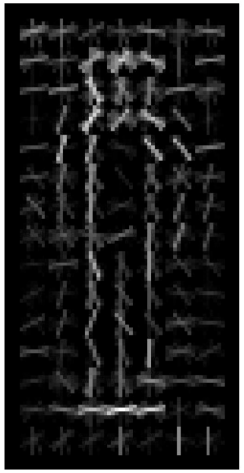


[Dalal & Triggs, CVPR 2005]

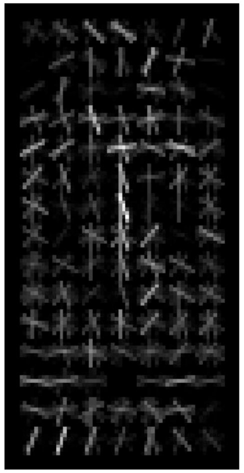
Descriptor Cues

Input example

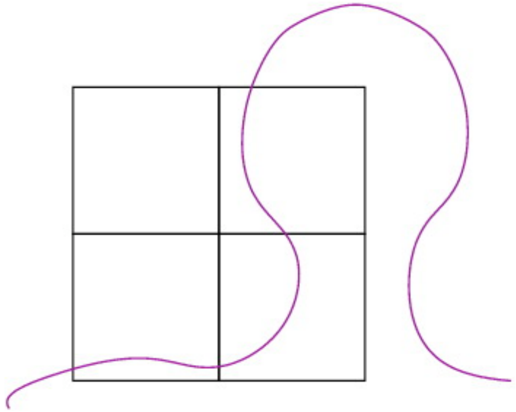
Average gradients



Weighted pos wts



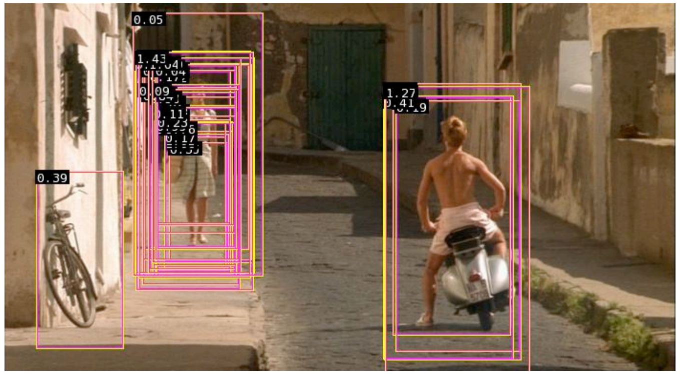
Weighted neg wts



Outside-in weights

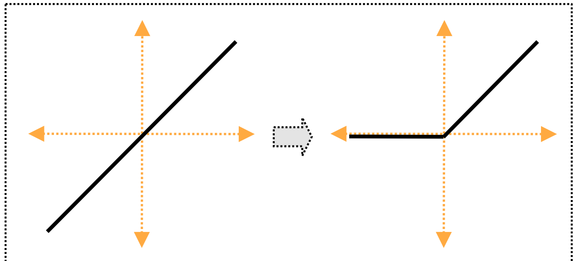
- Most important cues are head, shoulder, leg silhouettes
- Vertical gradients inside a person are counted as negative
- Overlapping blocks just outside the contour are most important

Multi-Scale Object Localisation

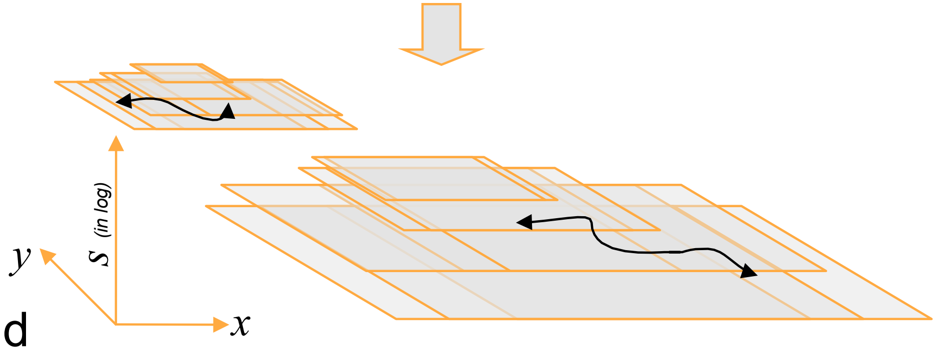


Multi-scale dense scan of detection window

Bias



Clip Detection Score



Threshold

Apply robust mode detection, like mean shift

Final detections

- Robust non-maximum suppression is important
- Fine scale transitions helps!

Human detection

Two layer detection [Harzallah et al. 2009]

- Combination of a linear with a non-linear SVM classifier
  - Linear classifier is used to preselection
  - Non-linear one for scoring
- Use of image classification for context information
- Winner of 11/20 classes in the PASCAL Visual Object Classes Challenge 2008 (VOC 2008)

PASCAL VOC 2008 dataset

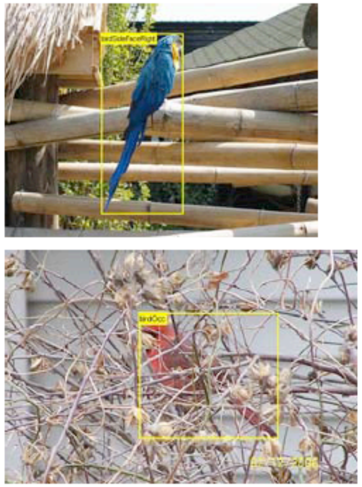
- 8465 image (4332 training and 4133 test) downloaded from Flickr, manually annotated
- 20 object classes (aeroplane, bicycle, bird, etc.)
- Between 130 and 832 images per class (except person 3828)
- On average 2-3 objects per image
- Viewpoint information : front, rear, left, right, unspecified
- Other information : truncated, occluded, difficult

PASCAL 2008 dataset

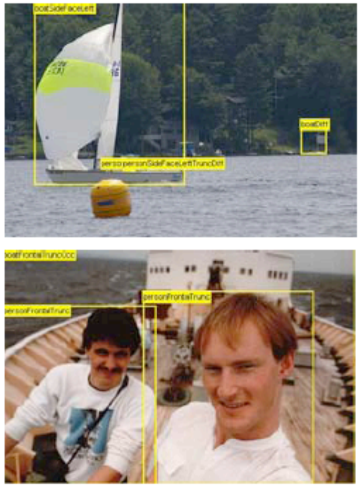
Aeroplane

Bicycle

Bird



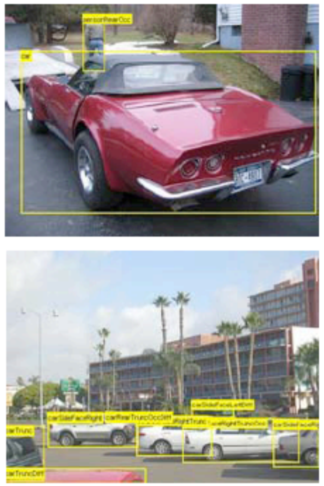
Boat



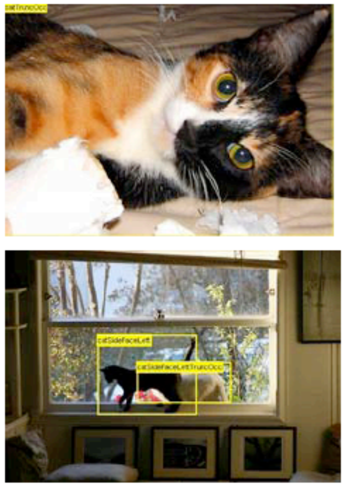
Bottle

Bus

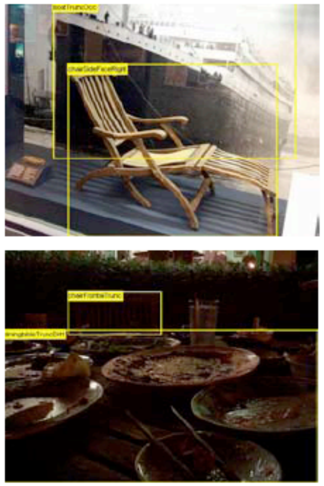
Car



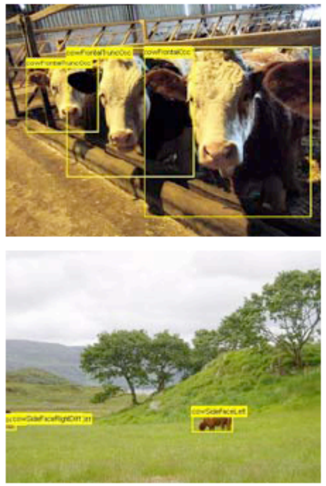
Cat



Chair



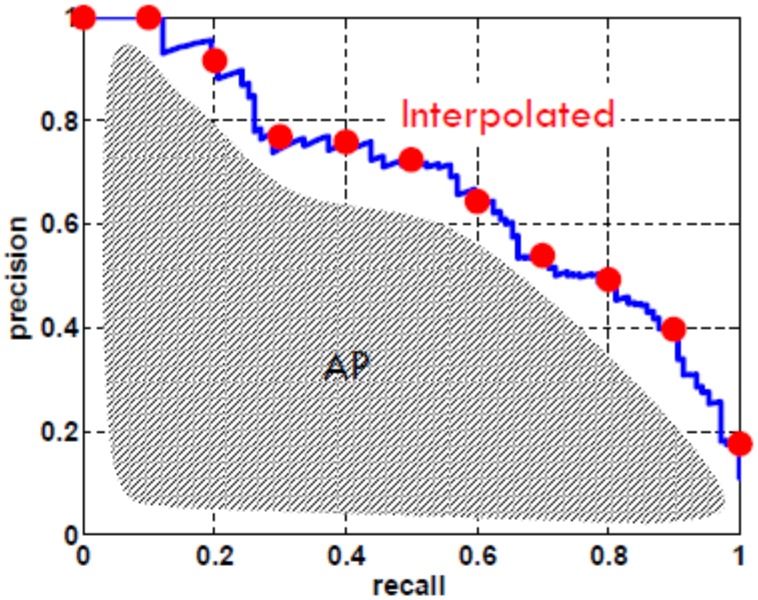
Cow



PASCAL 2008 dataset

Evaluation

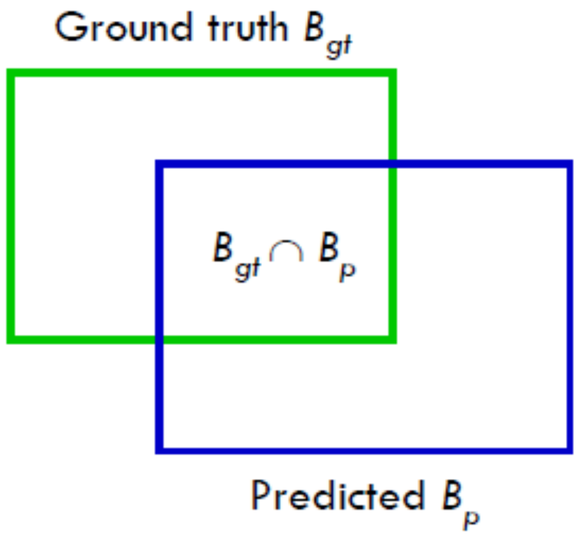
- Average Precision [TREC] averages precision over the entire range of recall
  - Curve interpolated to reduce influence of “outliers”



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

Evaluating bounding boxes

- Area of Overlap (AO) Measure

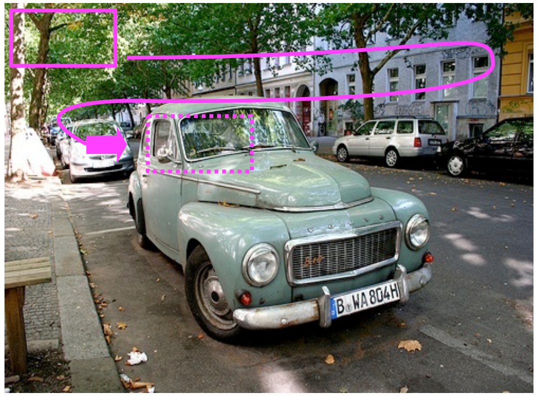


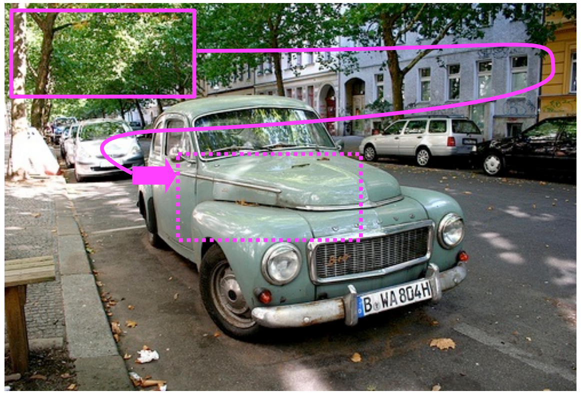
$$AO(B_{gt}, B_p) = \frac{|B_{gt} \bigcap B_p|}{|B_{gt} \bigcup B_p|}$$

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

Introduction [Harzallah et al. 2000]

- Method with sliding windows (Each window is classified as containing or not the targeted object)





- Learn a classifier by providing positive and negative examples

Generating training windows

- Adding positive training examples by shifting and scaling the original annotations [Laptev06]

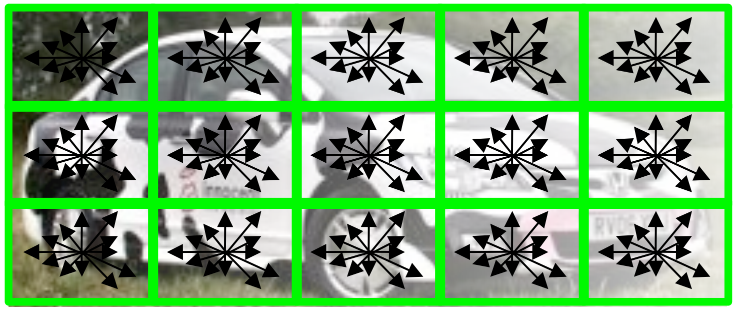
- Initial negative examples randomly extracted from background
- Training an initial classifier
- Retraining 4 times by adding false positives

Examples of false positives

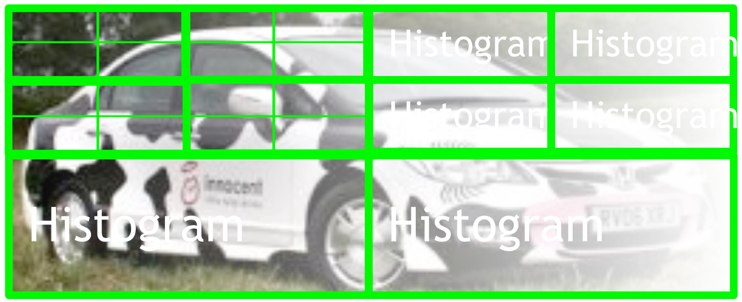
Image representation

- Combination of 2 image representations

- Histogram Oriented Gradient
  - Gradient based features
  - Integral Histograms



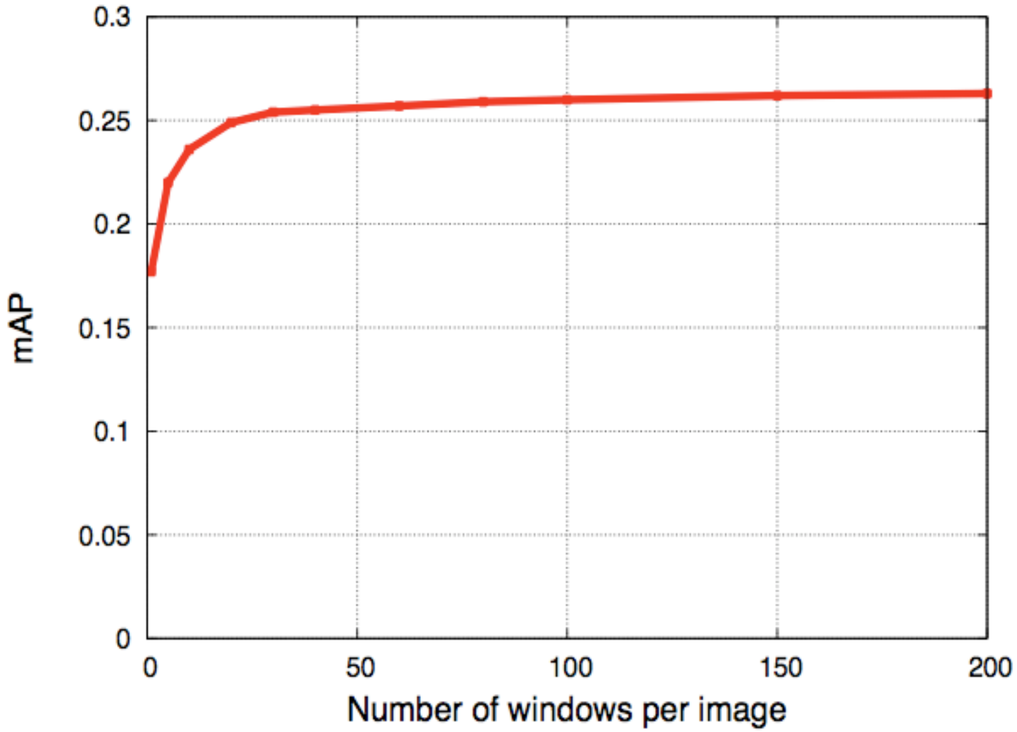
- Bag of Features
  - SIFT features extracted densely + k-means clustering
  - Pyramidal representation of the sliding windows
  - One histogram per tile



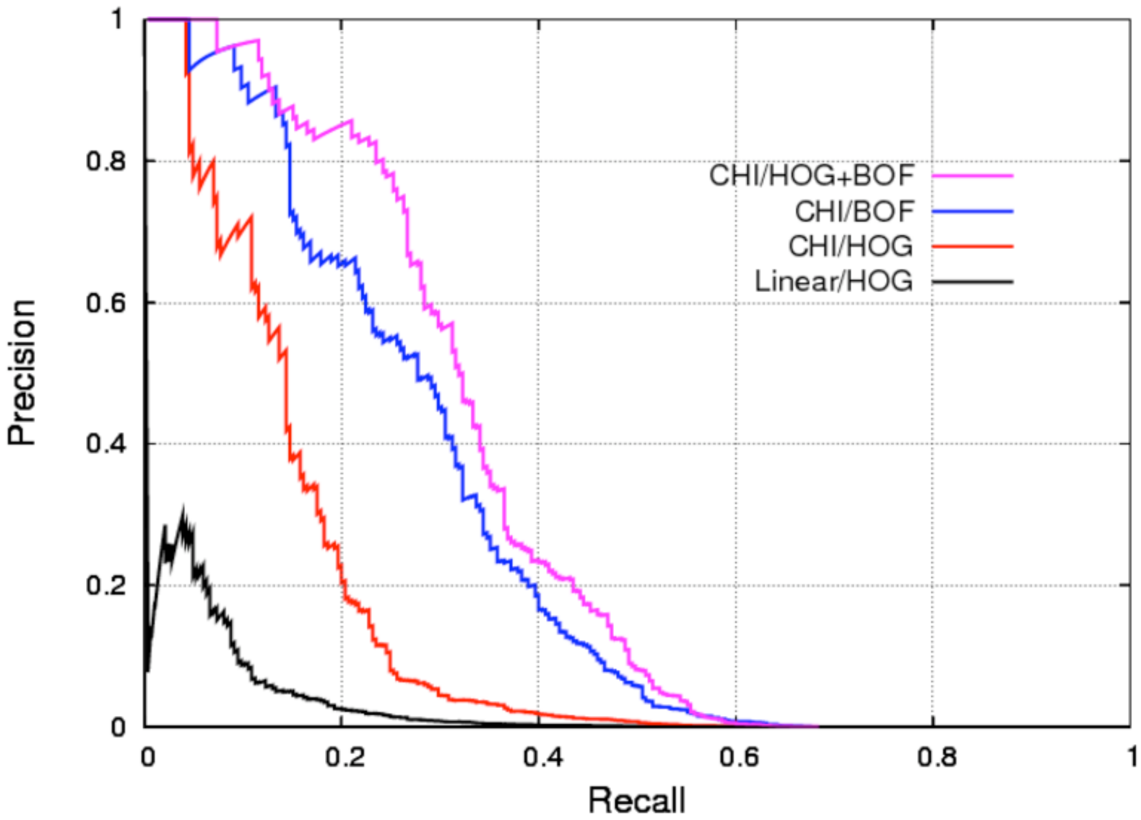
Efficient search strategy

- Reduce search complexity
  - Sliding windows: huge number of candidate windows
  - Cascades: pros/cons
- Two stage cascade:
  - Filtering classifier with a linear SVM
    - Low computational cost
    - Evaluation: capacity of rejecting negative windows
  - Scoring classifier with a non-linear SVM
    - X^2 kernel with a channel combination [Zhang07]
    - Significant increase of performance

Efficiency of the 2 stage localization

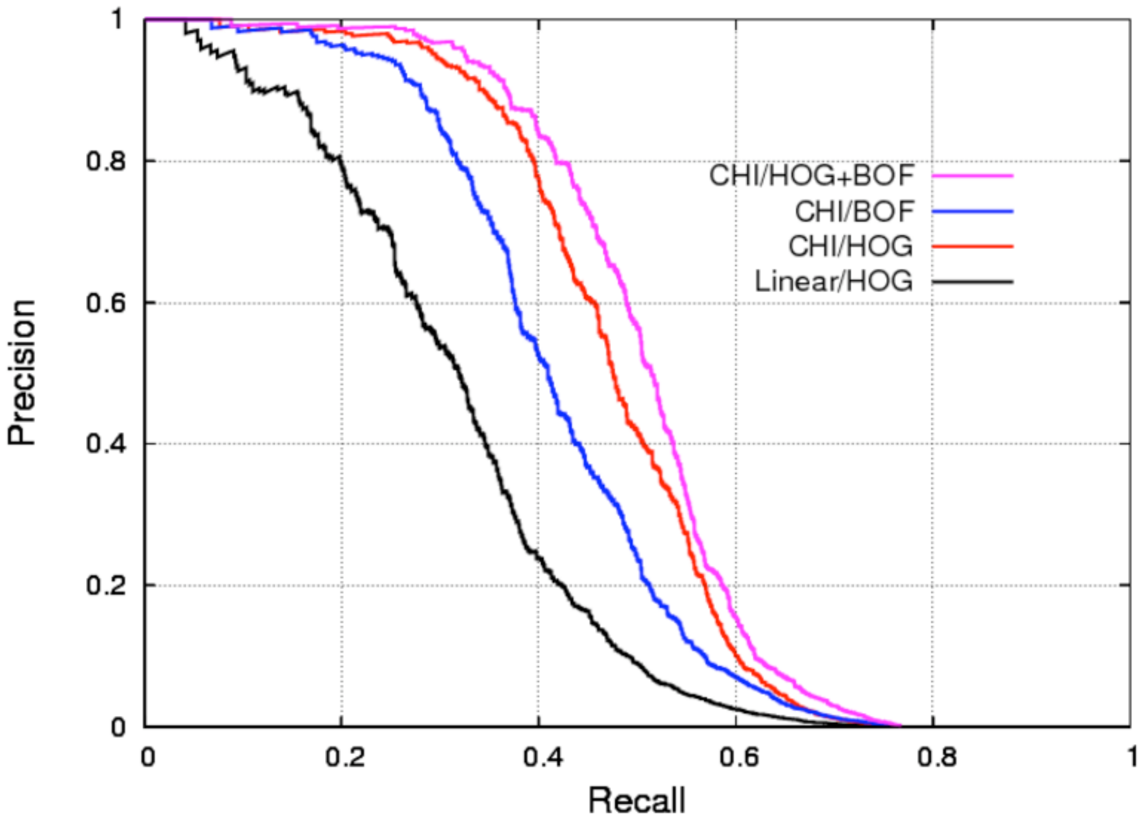


Localization performance: aeroplane



Method	AP
X^2, HOG+BOF	33.8
X^2, BOF	29.8
X^2, HOG	18.4
Linear, HOG	10.0

Localization performance: car



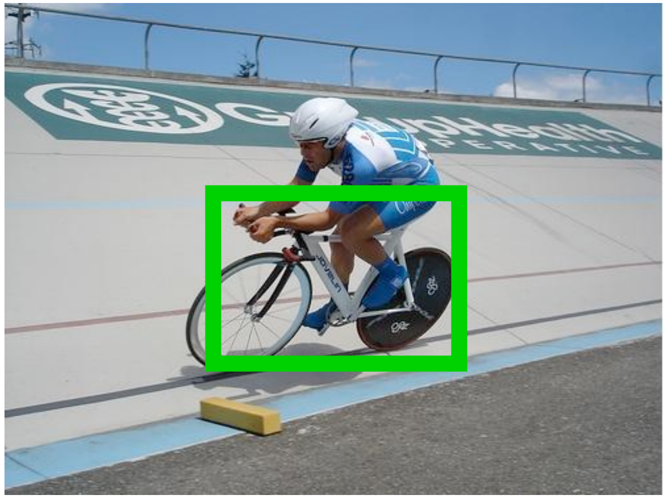
Method	AP
X^2, HOG+BOF	50.4
X^2, BOF	42.3
X^2, HOG	47.5
Linear, HOG	33.9

Localization performance

Mean Average Precision on all 20 classes, PASCAL 2007 dataset

Method	mAP
Linear, HOG	**14.6**
Linear, BOF	**15.0**
Linear, HOG+BOF	**17.6**
X^2, HOG	**21.9**
X^2, BOF	**23.1**
X^2, HOG+BOF	**26.3**

Localization examples: correct localizations



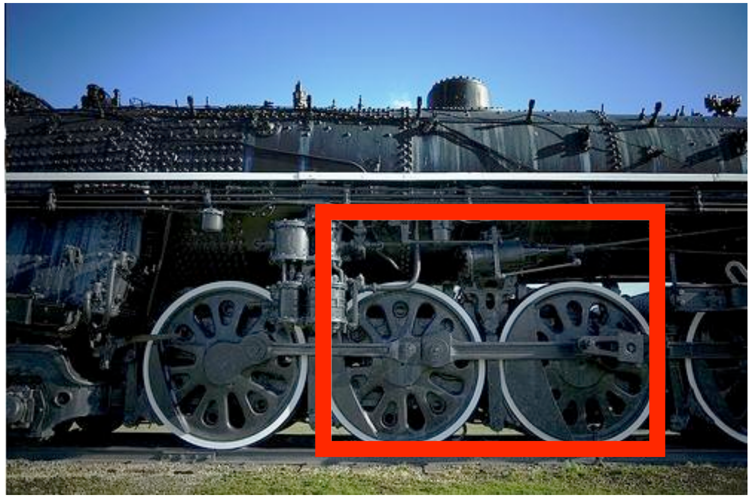
Bicycle

Car

Horse

Sofa

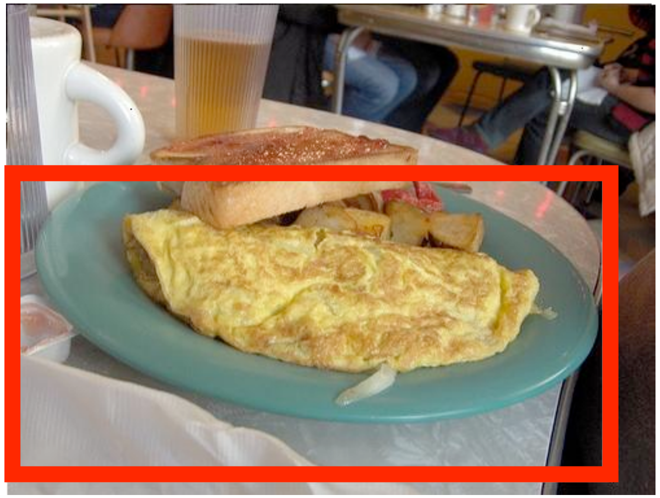
Localization examples: false positives



Bicycle

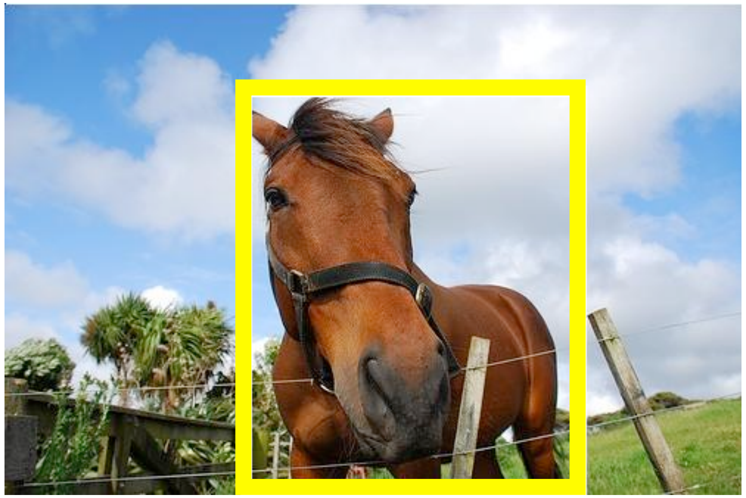
Car

Horse



Sofa

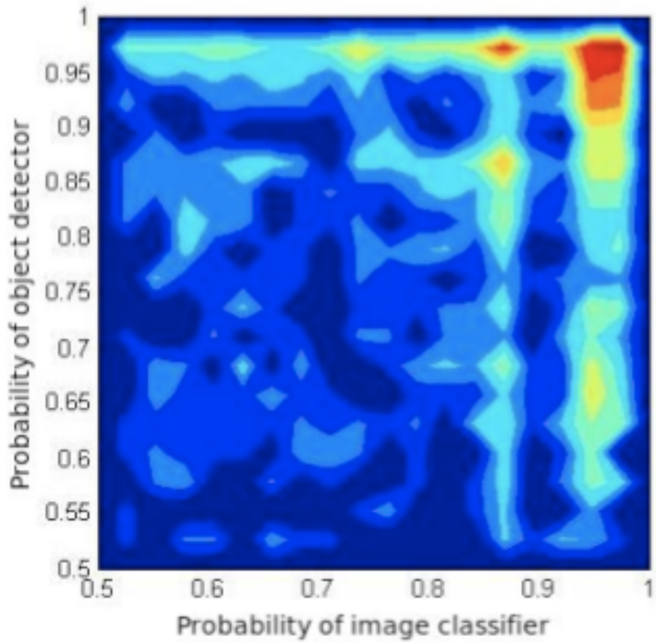
Localization examples: missed objects



Combining image classification and localization

- Image classification & localization use a different information

- For many TP only one has a high score
  - Truncated objects: hard for the detector
  - Small objects: ok for the detector but not for the classifier using global information



Combination model

- Input: classification (S_i) and localization (S_w) scores
- Output: probability that object is present
- Suppose that classification and localization outputs are independent:

$$P(O|S_w, S_i) \propto P(O|S_i) \times P(O|S_w)$$

Combination model

- For each modality (classification/detection): notion of *detectability* $P(D_i)$ for classifier and $P(D_w)$ for detector

- Encodes the ability to detect presence of the objects

- Assuming that the classifier/detector outputs conditional probabilities: $P(O|S_i, D_i)$ and $P(O|S_w, D_w)$

Combination model

- $P(O|S_i) = P(D_i)P(O|S_i, D_i) + P(\overline{D_i})P(O|S_i, \overline{D_i})$

- $P(O|S_w) = P(D_w)P(O|S_w, D_w) + P(\overline{D_w})P(O|S_w, \overline{D_w})$

- Final probability: $P(O|S_w, S_i) \propto P(O|S_i) \times P(O|S_w)$

- Handle both cases:
  - Object detectable by two modalities
  - Object detectable by only one modality

Combination model

- $P(O|S_i, \overline{D_i})$ and $P(O|S_w, \overline{D_w})$: constant value

- S_w = classification by localization: highest localization score

- Priors $P(D_i)$ and $P(D_w)$ class dependant

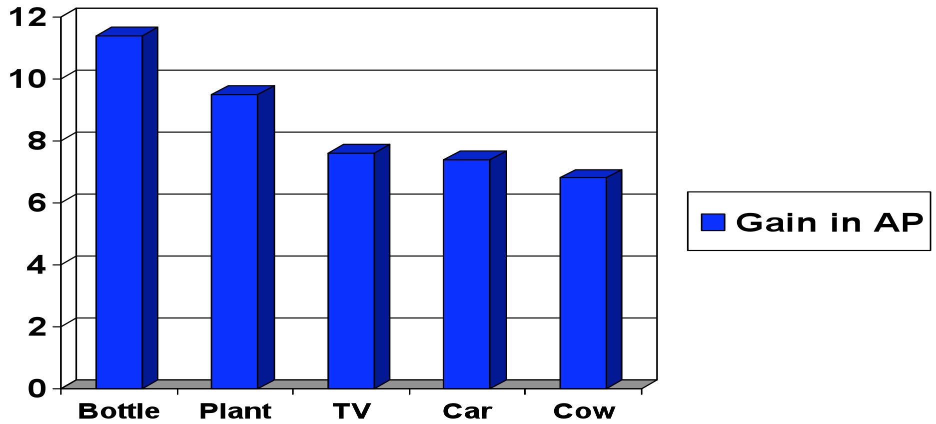
Combination experimental setup

- Image classifier : INRIA_flat classifier
  - SVM classifier X^2 kernel using multiple feature channels [Zhang07]
  - Excellent results in PASCAL 2008 challenge
- Detector : as described previously
- Experimental validation on PASCAL VOC 2007

Experimental results : gain obtained

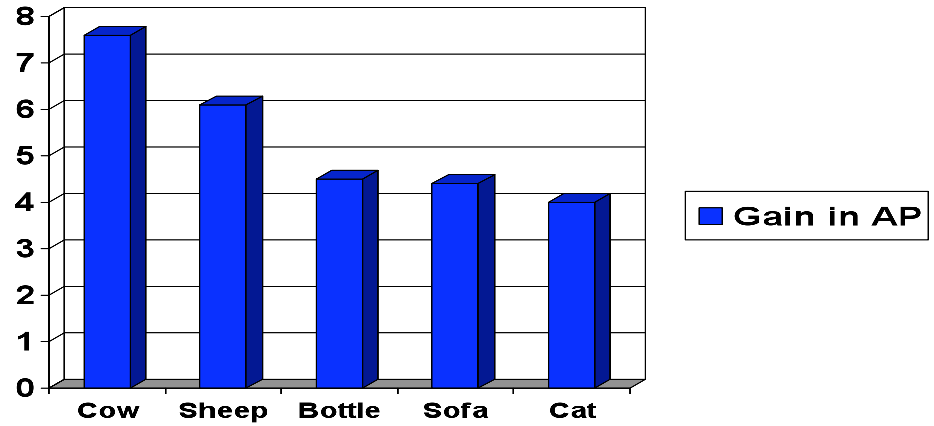
- Classification

Method	mAP
Base Classifier	**60.1**
Our Combination	**63.5**



- Localization

Method	mAP
Base Detector	**26.3**
Our Combination	**28.9**



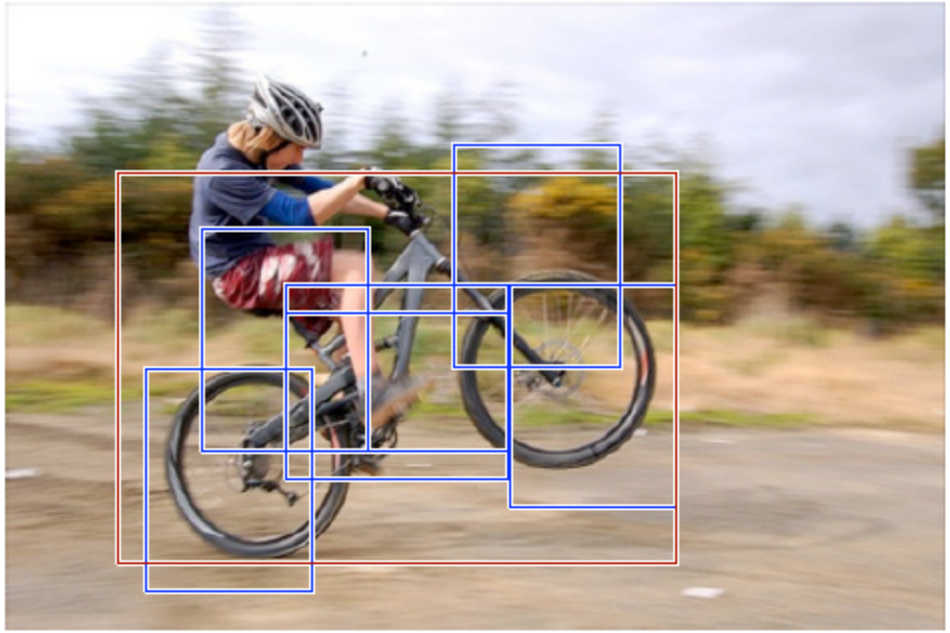
Experimental results

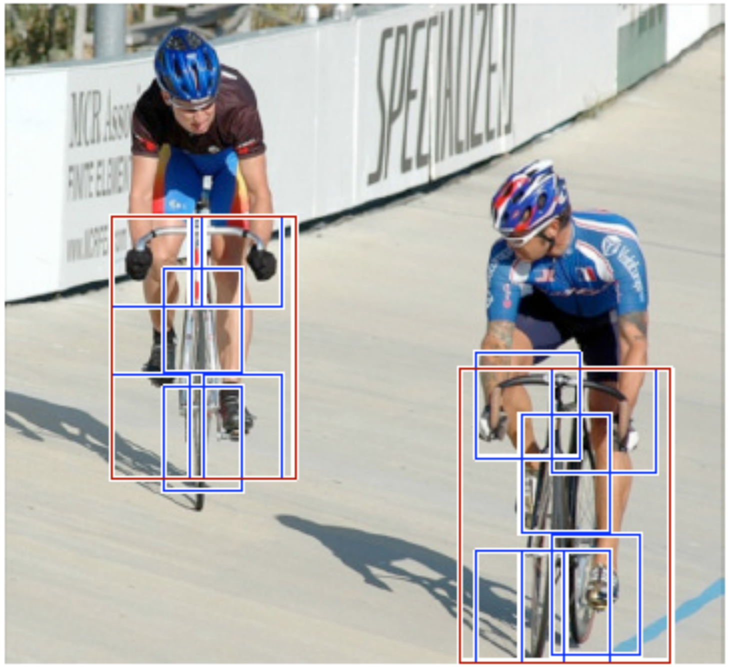
Correct but low score for car localization
High classification score for car
➡ score increased after combination

Experimental results

High classification score for car
No localization of car
➡ score decreased after combination

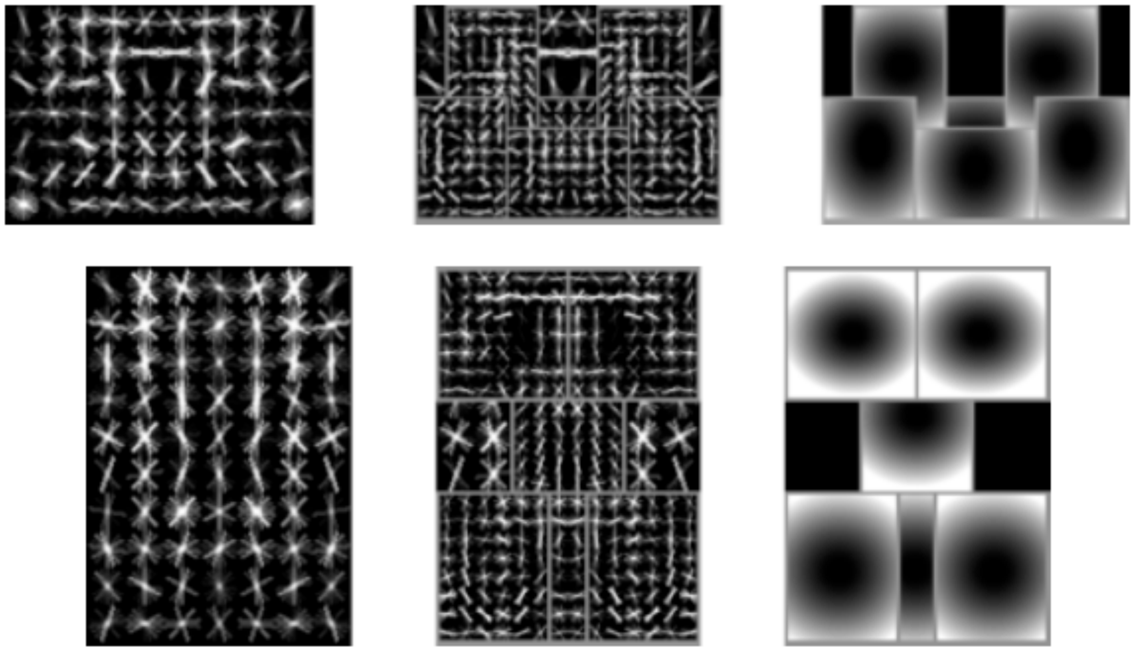
Flexible Model [Felsenszwalb et al. 2009]





- Mixture of deformable part models
- Each component has global template + deformable parts
- Fully trained from bounding boxes alone

Two component bike model



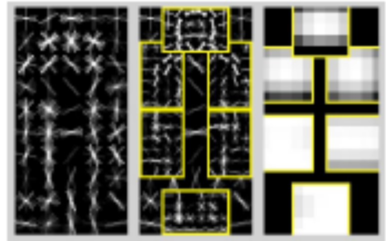
root filters
coarse resolution

part filters
finer resolution

deformation
models

Each component has a root filter F_0
and n part models (F_i, v_i, d_i)

Object hypothesis



$z = (p_0, \ldots, p_n)$

p_0 : location of root

$p_1, \ldots, p_n$: location of parts

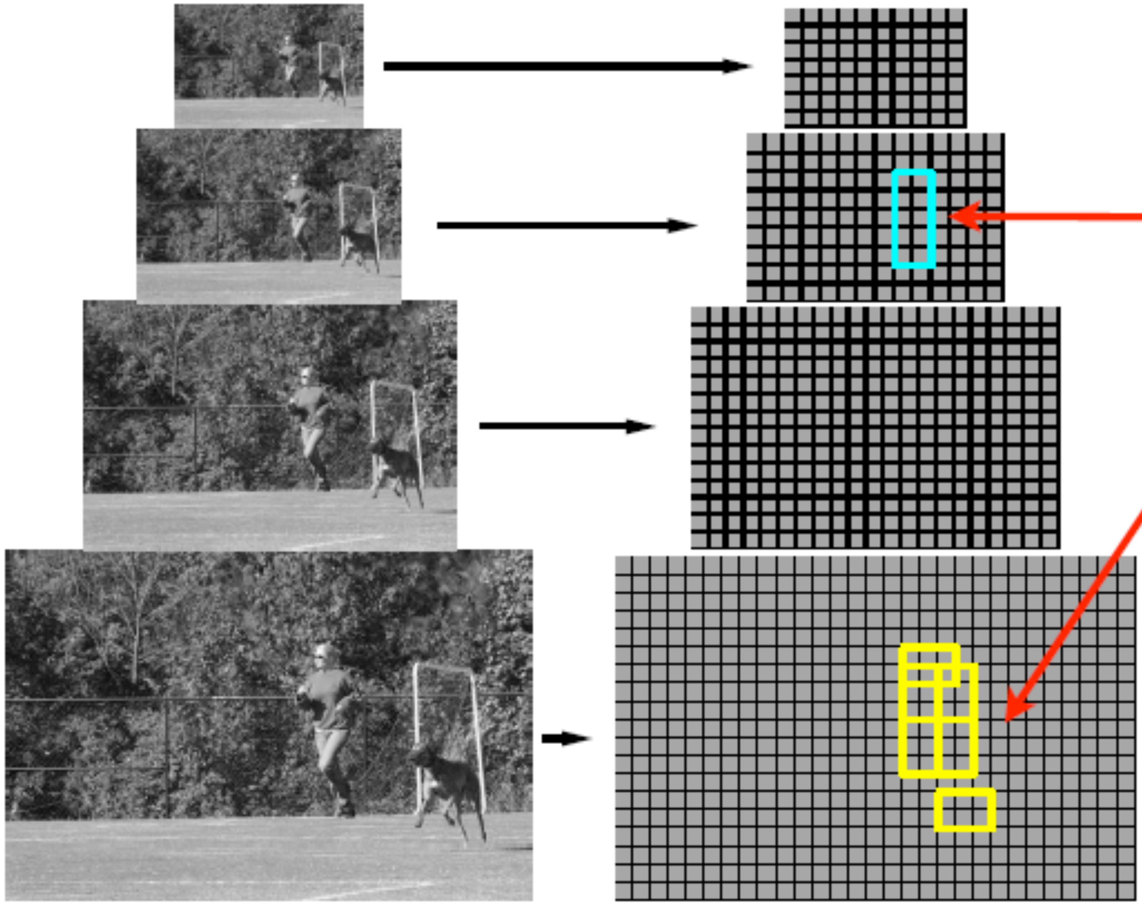


Image pyramid HOG feature pyramid

Score is sum of filter scores minus deformation costs

Multiscale model captures features at two-resolutions

Score of a hypothesis

"data term"

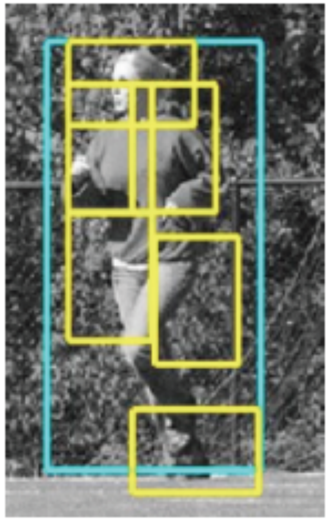
"spatial prior"

$$\text{score}(p_0, \ldots, p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2)$$

displacements

filters

deformation parameters



$$\text{score}(z) = \beta \cdot \Psi(H, z)$$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features

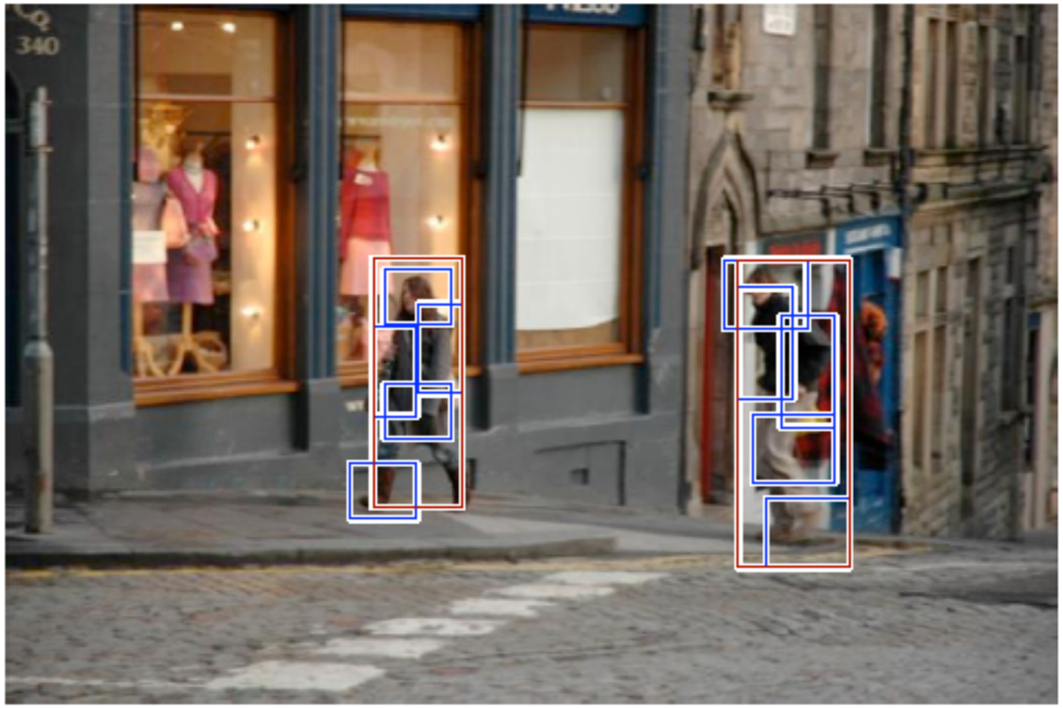
Matching

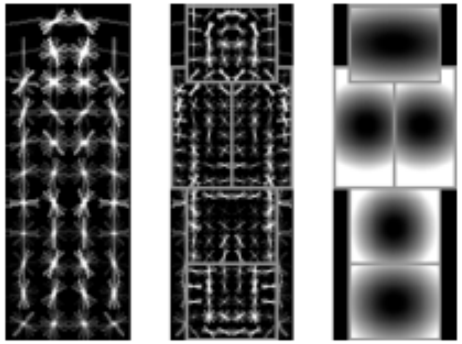
- Define an overall score for each root location
  - Based on best placement of parts

$$\text{score}(p_0) = \max_{p_1,\ldots,p_n} \text{score}(p_0,\ldots,p_n).$$

- High scoring root locations define detections
  - “sliding window approach”

Matching results

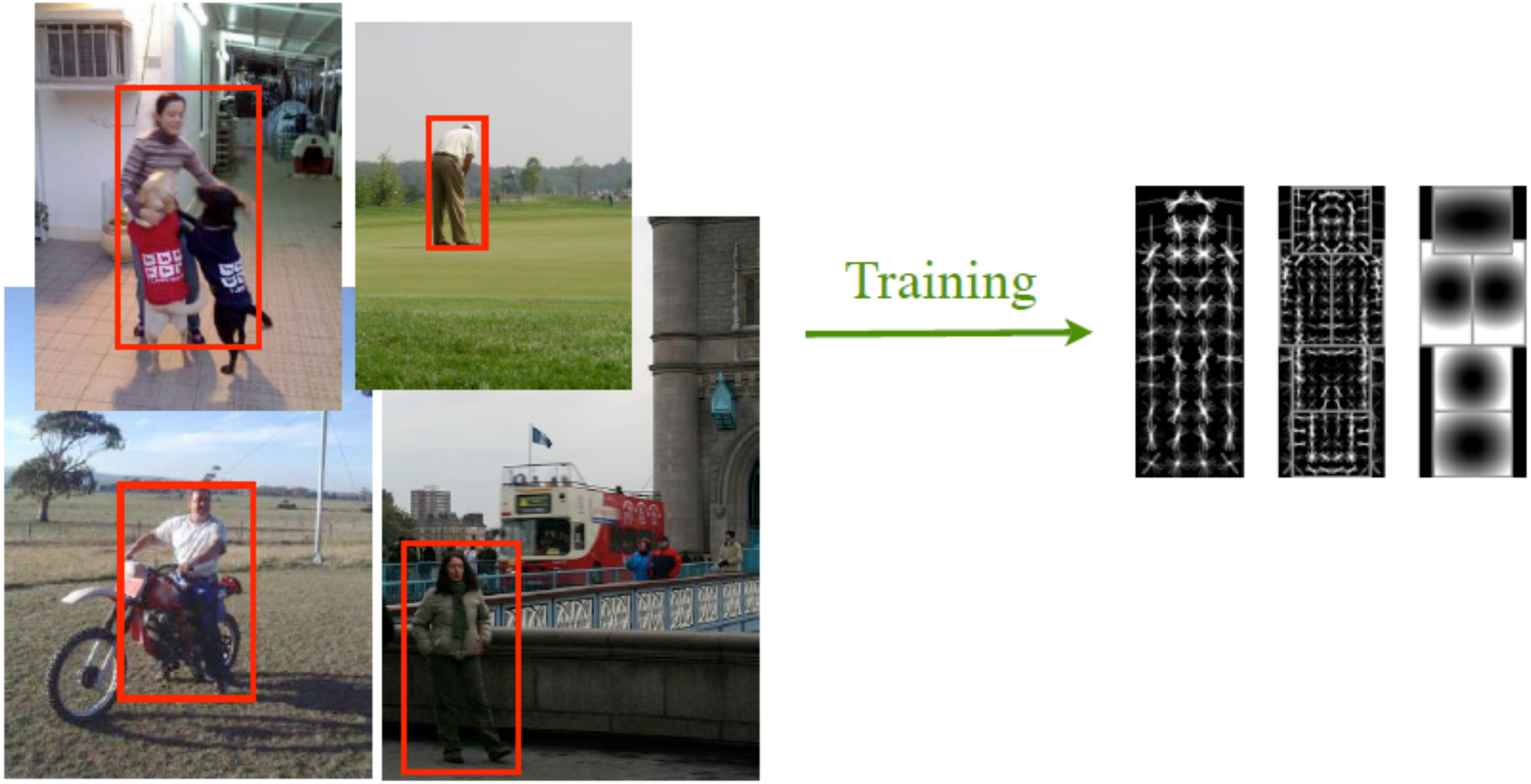




(after non-maximum suppression)

Training

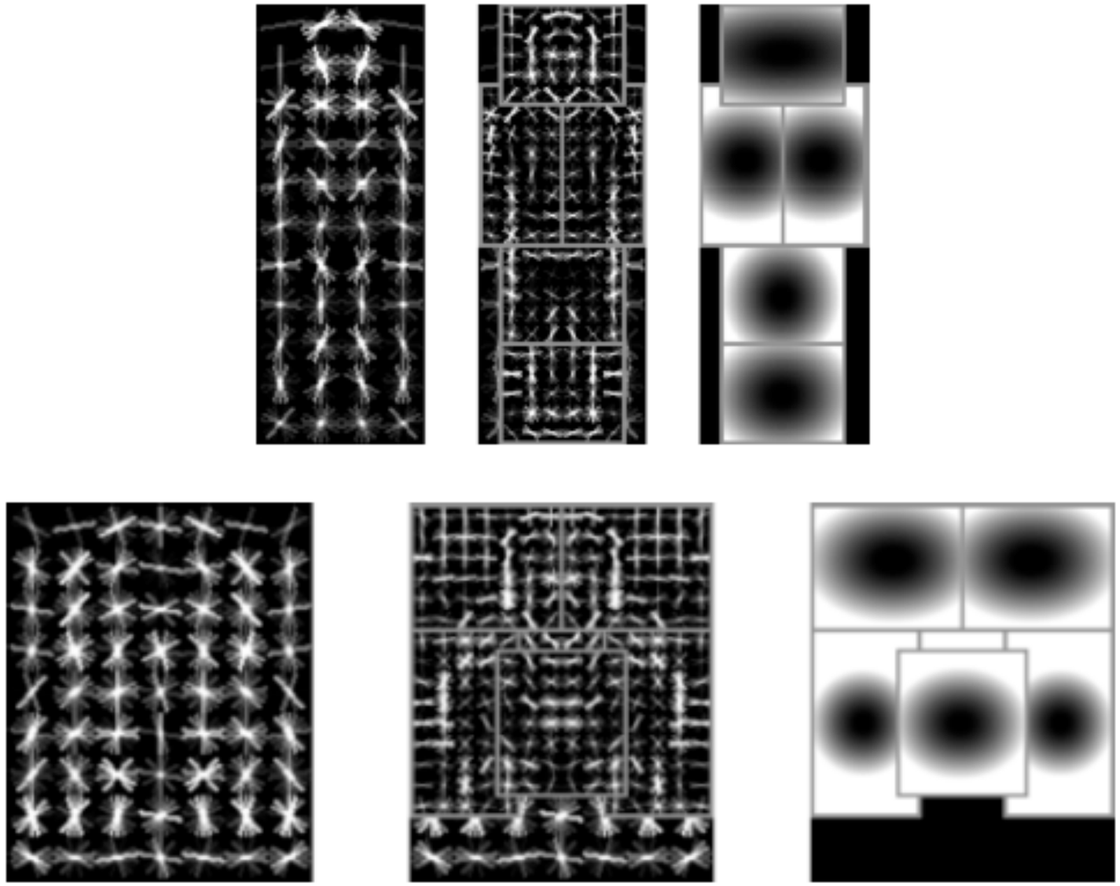
- Training data consists of images with labeled bounding boxes.
- Need to learn the model structure, filters and deformation costs.



Training Models

- Reduce to Latent SVM training problem
- Positive example specifies some z should have high score
- Bounding box defines range of root locations
  - Parts can be anywhere
  - This defines $Z(x)$ part locations

Person model



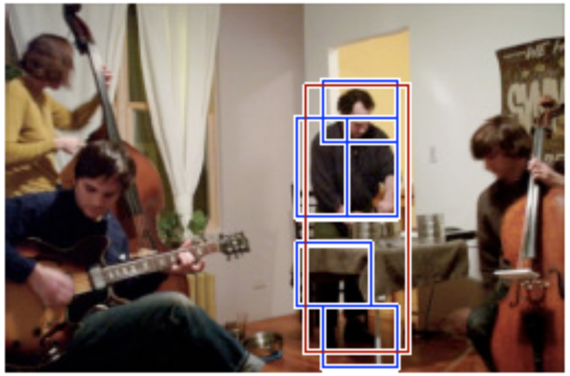
root filters
coarse resolution

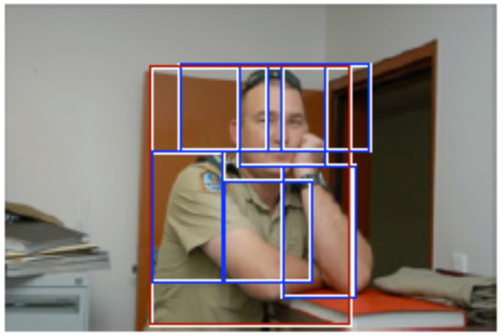
part filters
finer resolution

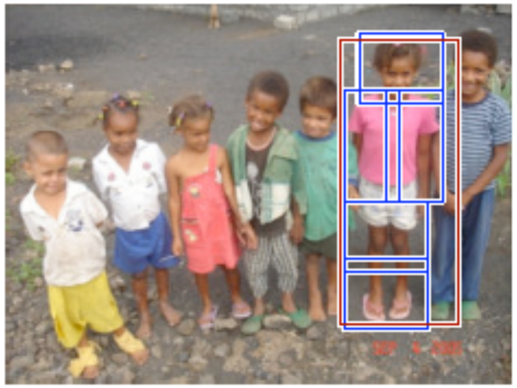
deformation
models

Person detections

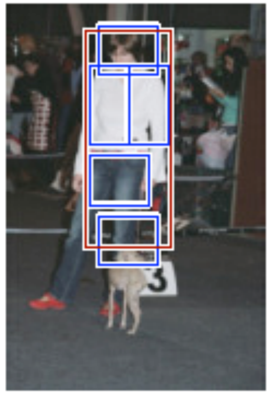
high scoring true positives







high scoring false positives
(not enough overlap)



Shape-based features for localization

- Classes with characteristic shape
  - Appearance, local patches are not adapted
  - shape-based descriptors are necessary

[Ferrari, Fevrier, Jurie & Schmid, PAMI'08]

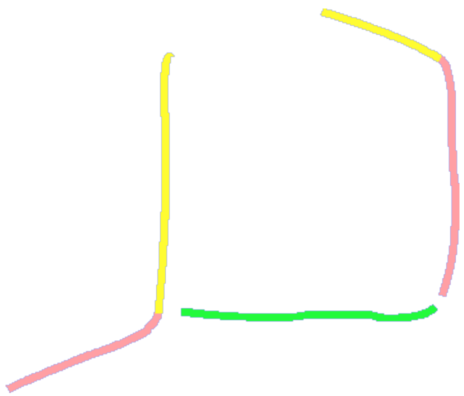
Pairs of adjacent segments (PAS)

Contour segment network

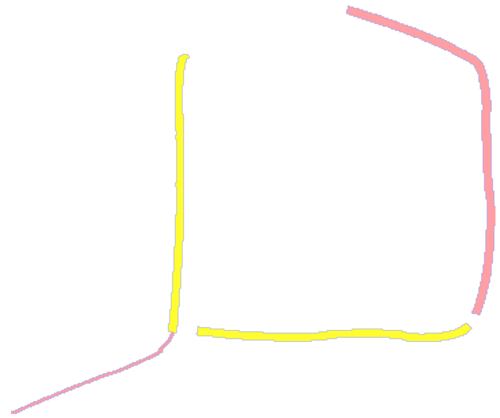
[Ferrari et al. ECCV'06]

1. Edgels extracted with Berkeley boundary detector
2. Edgel-chains partitioned into straight contour segments
3. Segments connected at edgel-chains' endpoints and junctions

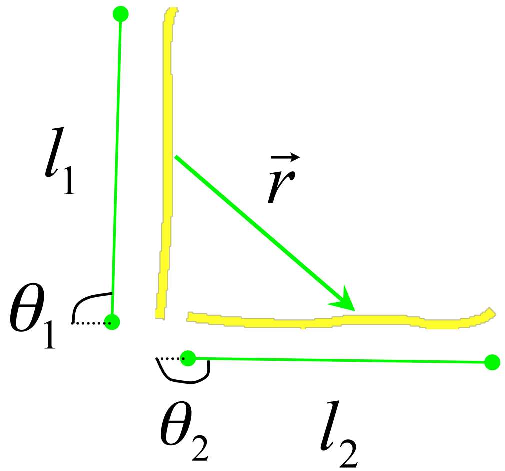
Pairs of adjacent segments (PAS)



Contour segment network



PAS = groups of two connected segments



PAS descriptor:

$$\left(\frac{r_x}{\|\vec{r}\|},\frac{r_y}{\|\vec{r}\|},\theta_1,\theta_2,\frac{l_1}{\|\vec{r}\|},\frac{l_2}{\|\vec{r}\|}\right)$$

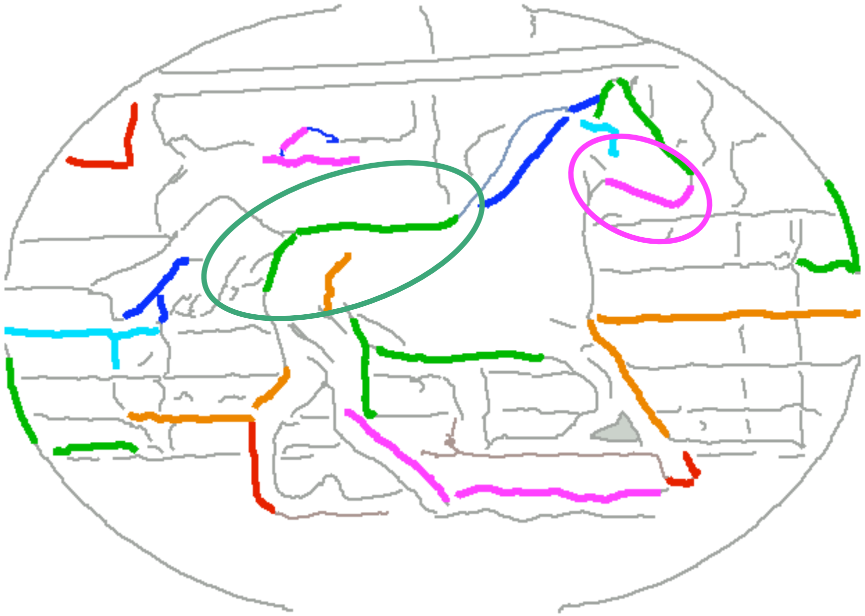
encodes *geometric* properties of the PAS

scale and translation invariant

compact, 5D

Features: pairs of adjacent segments (PAS)

Example PAS



Why PAS ?

+ can cover pure portions of the object boundary

+ intermediate complexity: good repeatability-informativeness trade-off

+ scale-translation invariant

+ connected: natural grouping criterion (need not choose a grouping neighborhood or scale)

PAS codebook

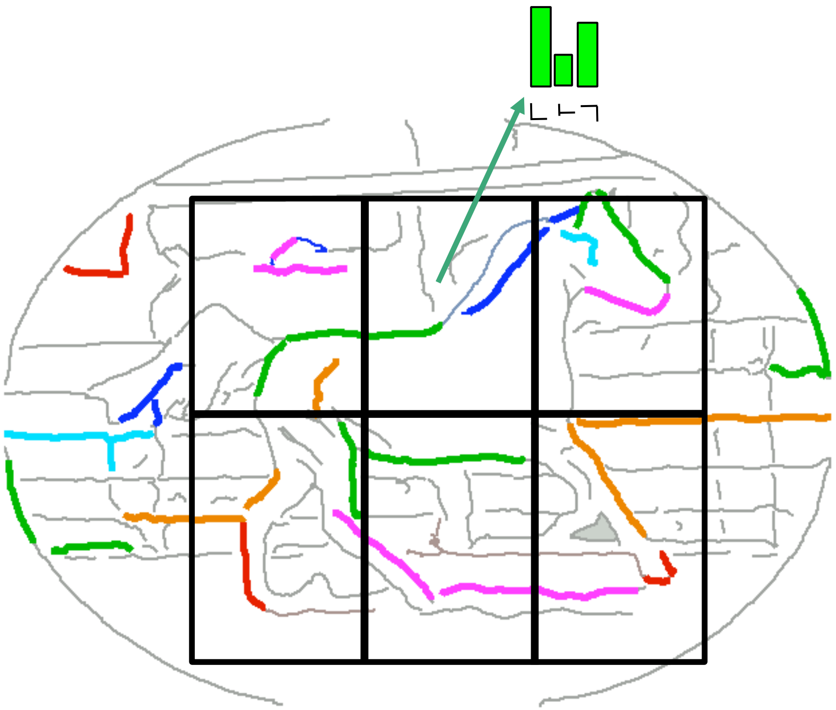
PAS descriptors are clustered into a vocabulary

a few types from 15 indoor images

- Frequently occurring PAS have intuitive, natural shapes
- As we add images, number of PAS types converges to just ~100
- Very similar codebooks come out, regardless of source images

→ general, simple features

Window descriptor



1. Subdivide window into tiles

2. Compute a separate bag of PAS per tile

3. Concatenate these semi-local bags

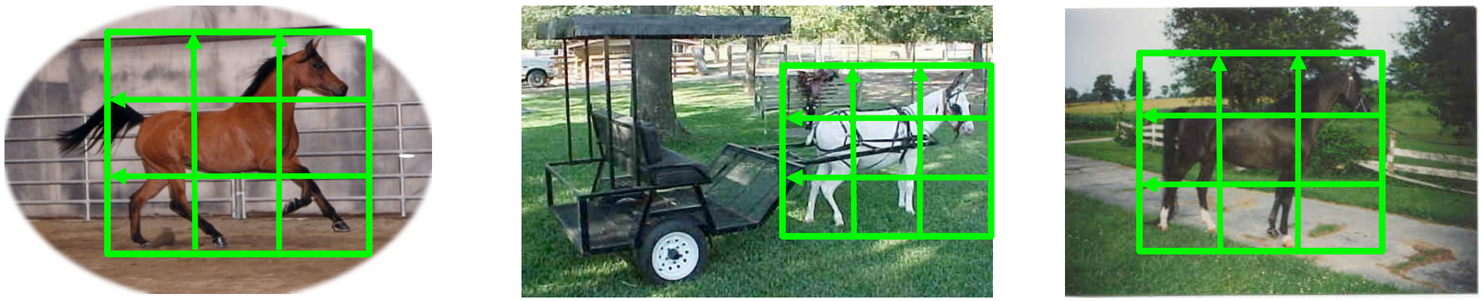
+ distinctive:
    records *which* PAS appear *where*
    weight PAS by average edge strength

+ flexible:
    soft-assign PAS to types, coarse tiling

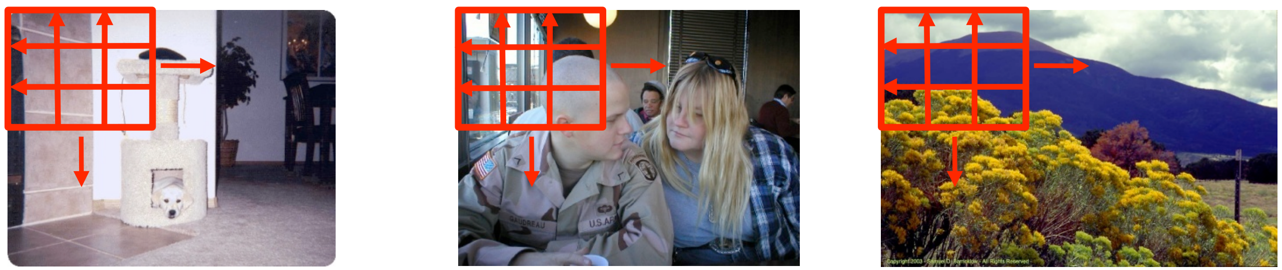
+ fast:
    computation with Integral Histograms

Training

1. Learn mean positive window dimensions $M_w \times M_h$
2. Determine number of tiles T
3. Collect positive example descriptors



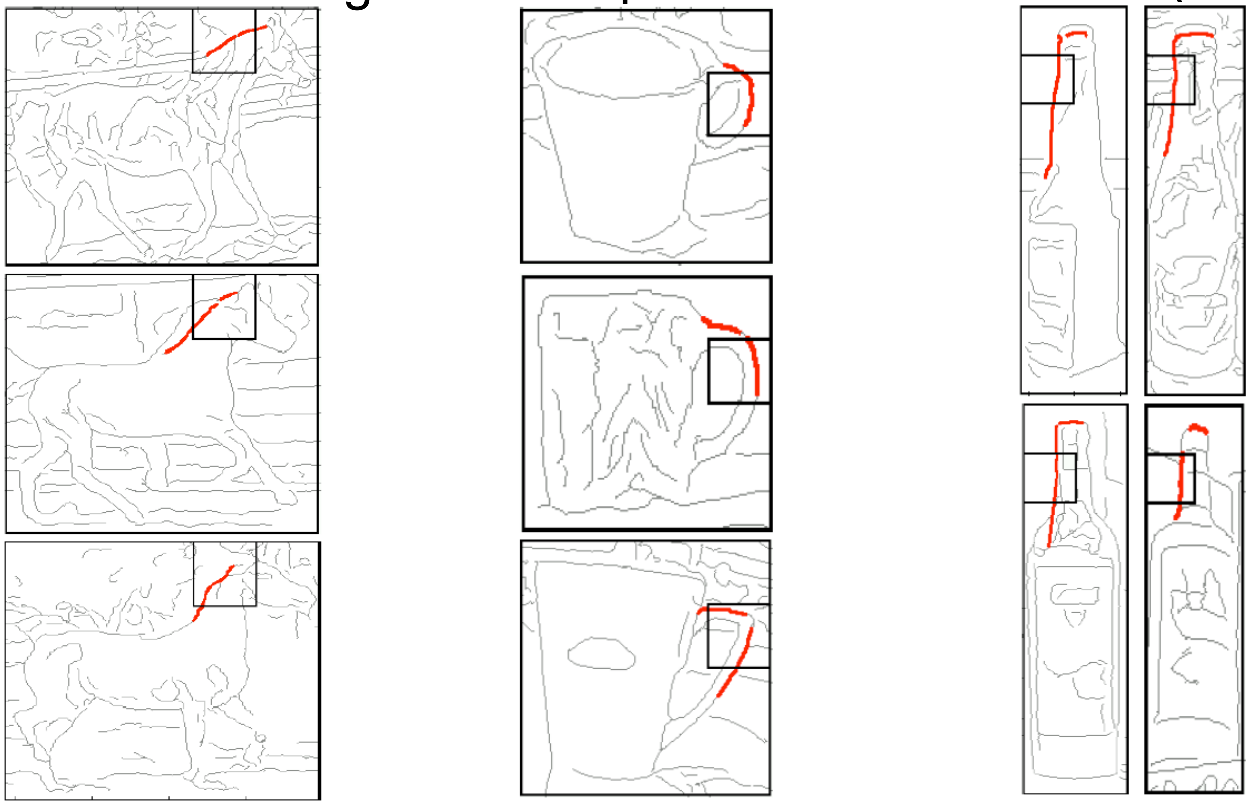
4. Collect negative example descriptors:
   slide $M_w \times M_h$ window over negative training images



Training

5. Train a linear SVM from positive and negative window descriptors

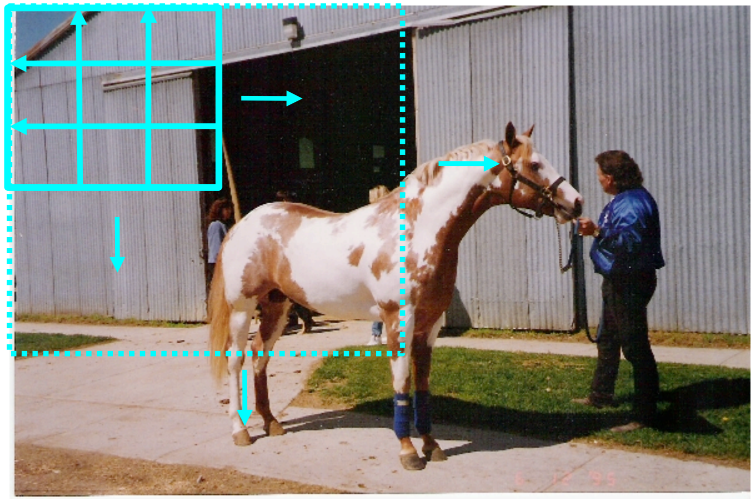
A few of the highest weighed descriptor vector dimensions (= 'PAS + tile')



\+ lie on object boundary (= local shape structures common to many training exemplars)

Testing

1. Slide window of aspect ratio M_w / M_h at multiple scales

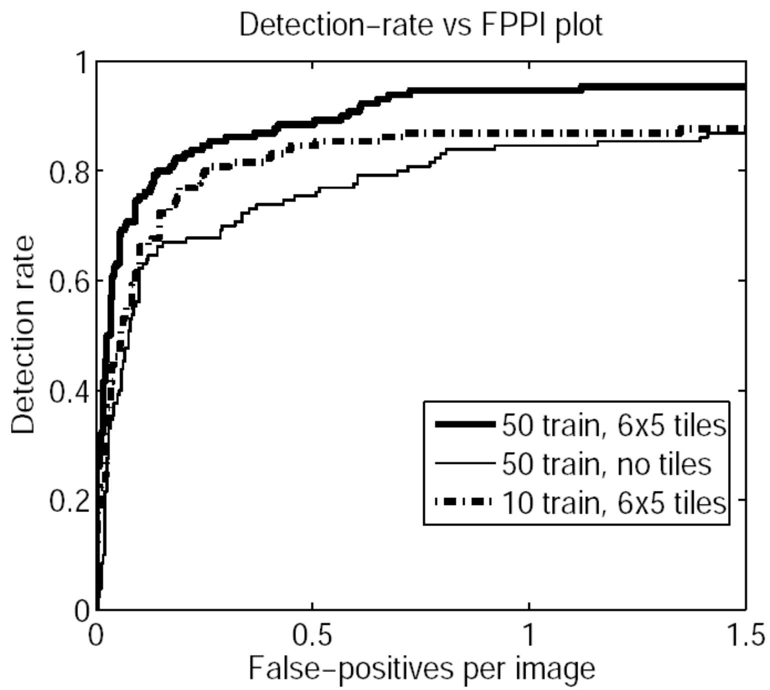


2. SVM classify each window + non-maxima suppression

$\longrightarrow$ detections

Experimental results – INRIA horses

Dataset: 170 positive + 170 negative images (training = 50 pos + 50 neg)
wide range of scales; clutter



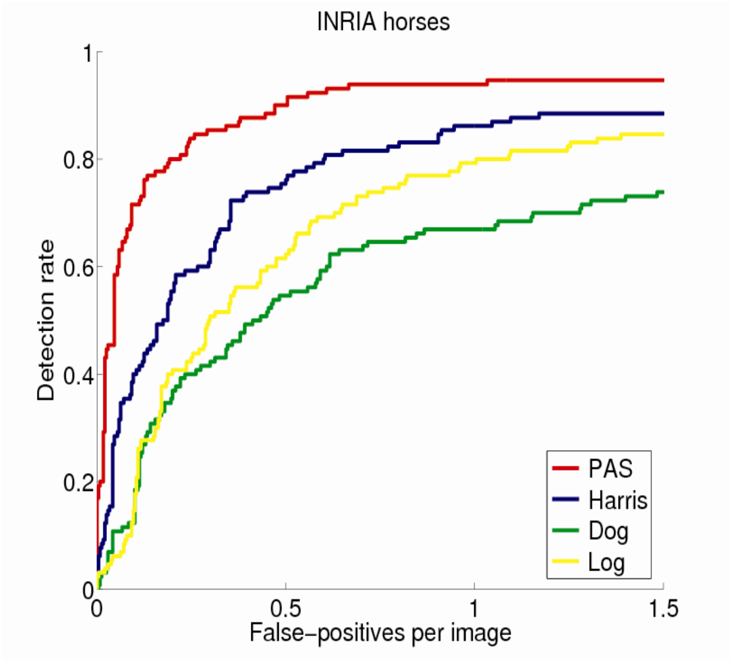
(missed and FP)

+ tiling brings a substantial improvement
optimum at T=30 → used for all other experiments

+ works well: 86% det-rate at 0.3 FPPI (50 pos + 50 neg training images)

Experimental results – INRIA horses

Dataset: 170 positive + 170 negative images (training = 50 pos + 50 neg)
wide range of scales; clutter



+ PAS better than any interest point detector

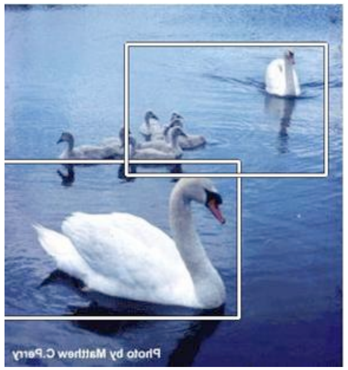
- all interest point (IP) comparisons with T=10, and 120 feature types (= optimum over INRIA horses, and ETHZ Shape Classes)
- IP codebooks are class-specific

Results – ETH shape classes

Dataset: 255 images, 5 classes; large scale changes, clutter
training = half of positive images for a class
+ same number from the other classes (1/4 from each)
testing = **all** other images

Results – ETH shape classes

Dataset: 255 images, 5 classes; large scale changes, clutter
training = half of positive images for a class
+ same number from the other classes (1/4 from each)
testing = **all** other images



Missed

Results – ETHZ Shape Classes

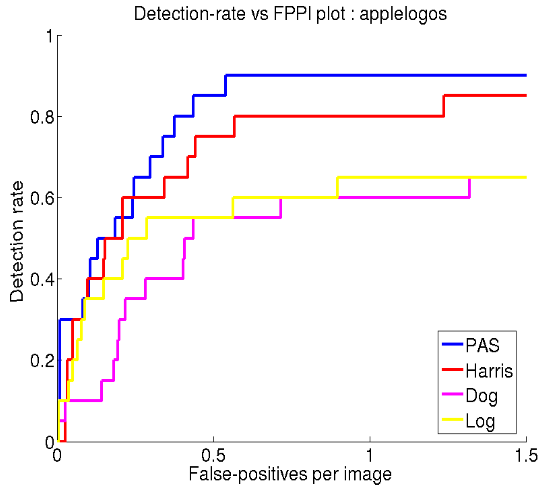
+ mean det-rate at 0.4 FPPI = 79%

+ class specific IP codebooks

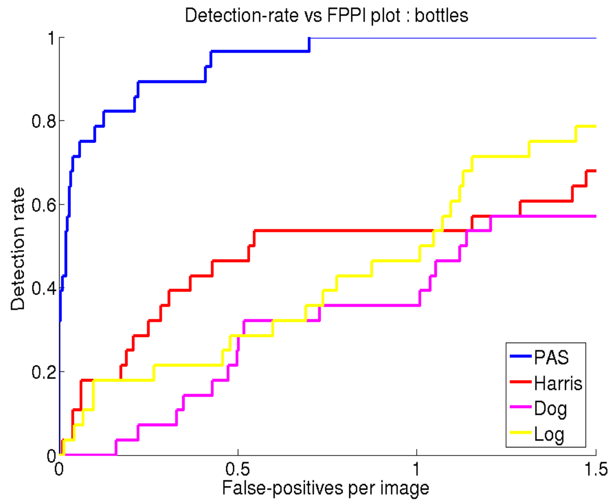
+ PAS >> I.P for
  apple logos, bottles, mugs
  PAS ~= IP for
  giraffes (texture!)
  PAS < IP for
  swan

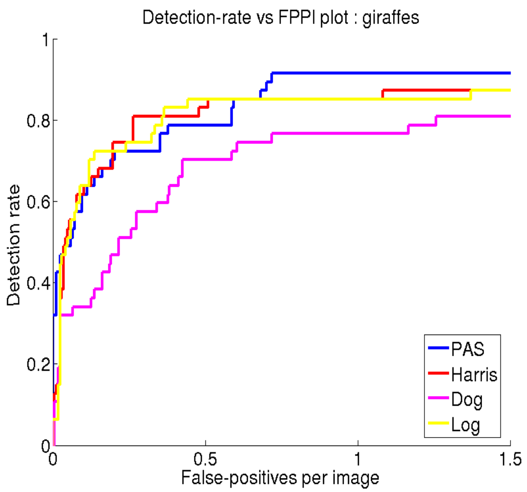
+ overall best IP: Harris-Laplace

Apple logos

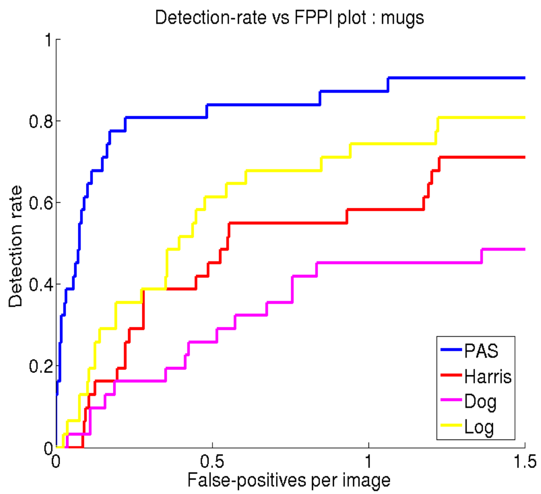


Bottles

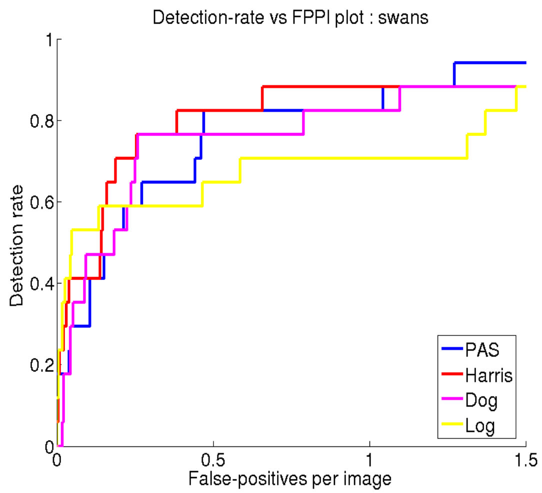




Giraffes



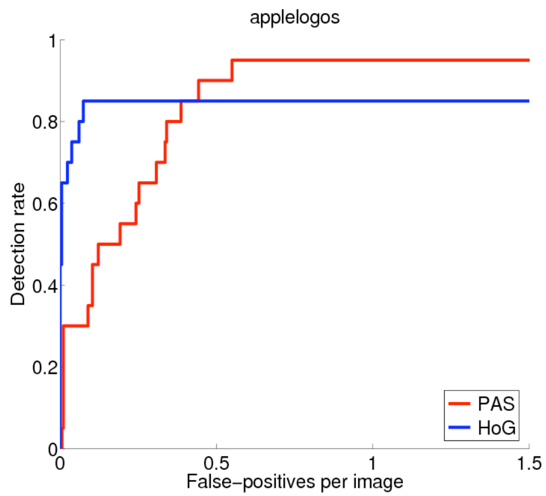
Mugs



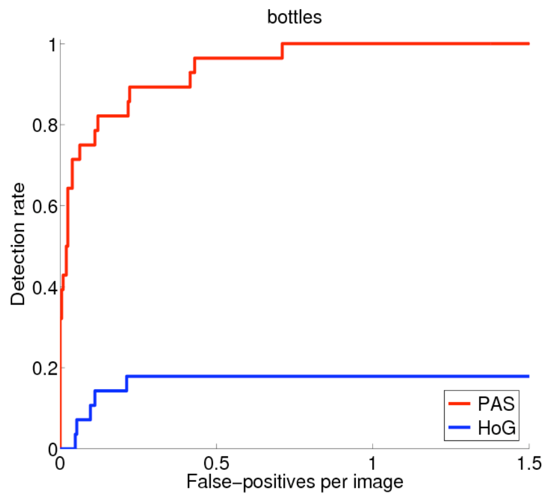
Swans

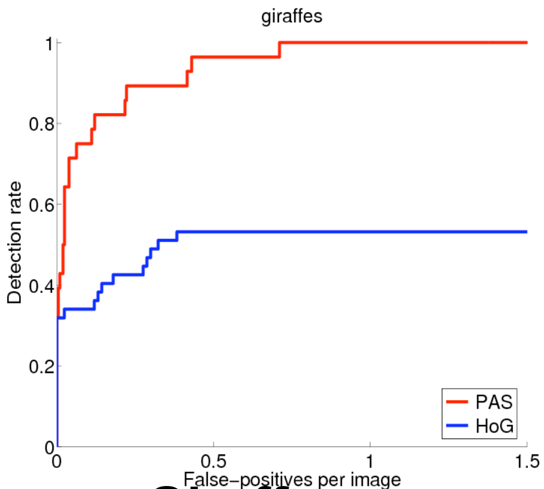
Comparison to HOG [Dalal & Triggs, CVPR'05]

Apple logos

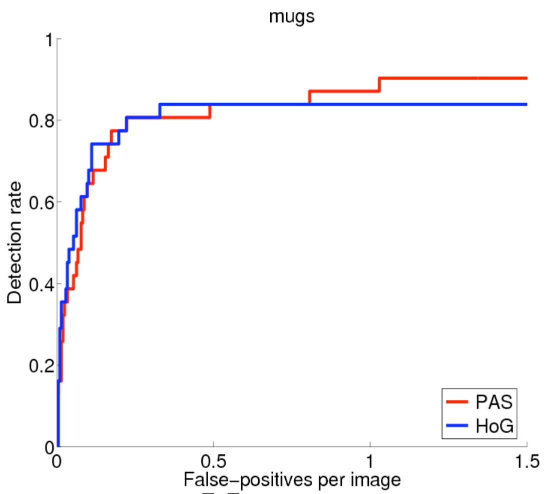


Bottles

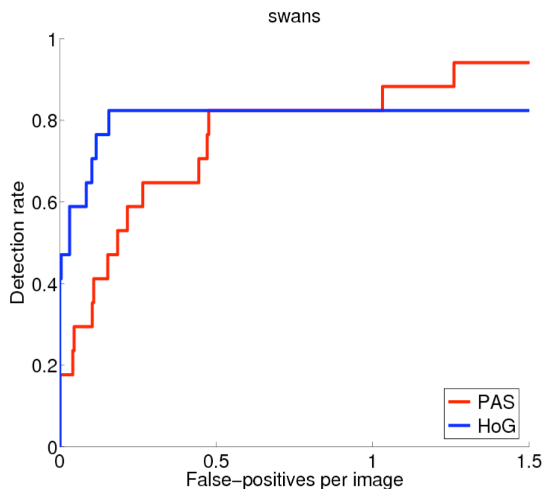




Giraffes



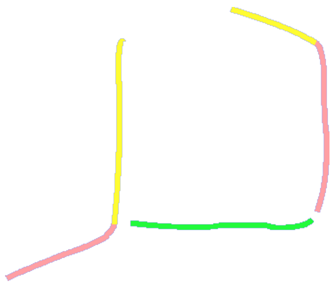
Mugs



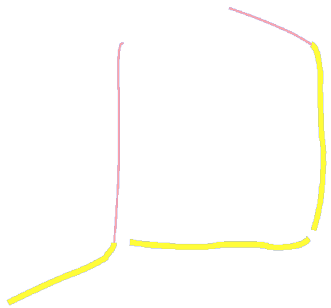
Swans

Generalizing PAS to *k*AS

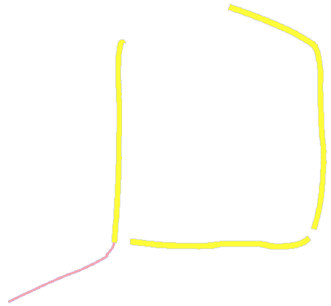
*k*AS: any path of length *k* through the contour segment network



segment network



3AS



4AS

scale+translation invariant descriptor with dimensionality 4*k*-2

k = feature complexity; higher *k* more informative, but less repeatable

overall mean det-rates (%)

	1AS	PAS	3AS	4AS
0.3 FPPI	69	77	64	57
0.4 FPPI	76	82	70	64

PAS do best !