

Category-level localization

Josef Sivic

<http://www.di.ens.fr/~josef>

Slides from Andrew Zisserman

Includes slides from: Ondra Chum, Alyosha Efros, Mark Everingham, Pedro Felzenszwalb, Rob Fergus, Kristen Grauman, Bastian Leibe, Ivan Laptev, Fei-Fei Li, Marcin Marszalek, Pietro Perona, Deva Ramanan, Bernt Schiele, Jamie Shotton, Andrea Vedaldi

Announcements

- Assignment 1 was due last week. **Have you sent it?**

- Assignment 2 is out. **Any questions?**

<http://www.di.ens.fr/willow/teaching/recvis11/assignment2/>

- Topic ideas for the final projects are out:

<http://www.di.ens.fr/willow/teaching/recvis11/finalproject/>

What we would like to be able to do...

- Visual scene understanding
- **What** is in the image and **where**



- Object categories, identities, properties, activities, relations, ...

Recognition Tasks

- **Image Classification**

- Does the image contain an aeroplane?



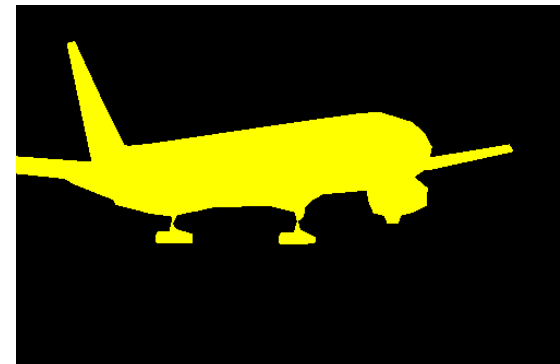
- **Object Class Detection/Localization**

- Where are the aeroplanes (if any)?



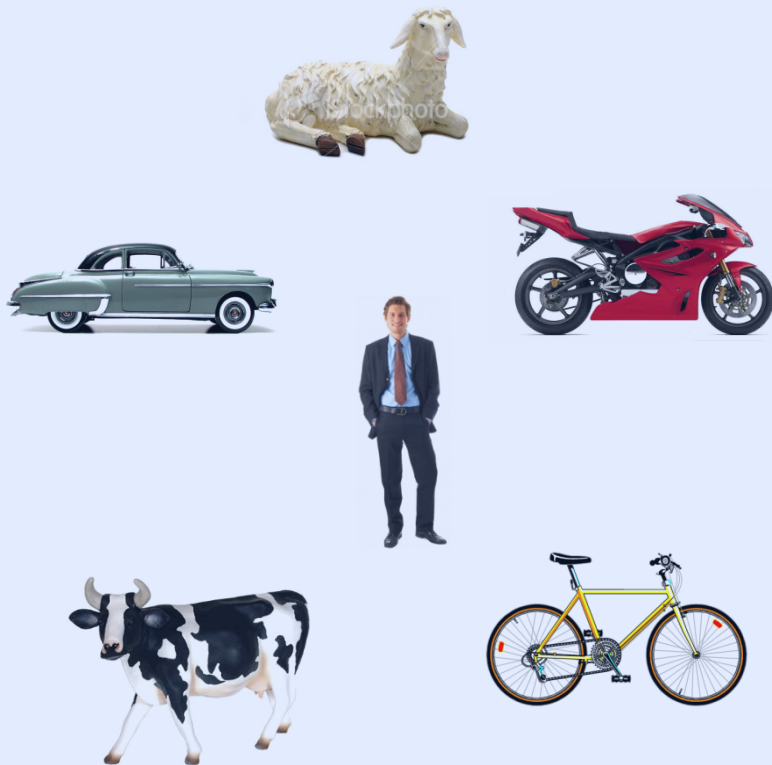
- **Object Class Segmentation**

- Which pixels are part of an aeroplane (if any)?



Things vs. Stuff

Thing (n): An object with a specific size and shape.



Ted Adelson, Forsyth et al. 1996.

Stuff (n): Material defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape.

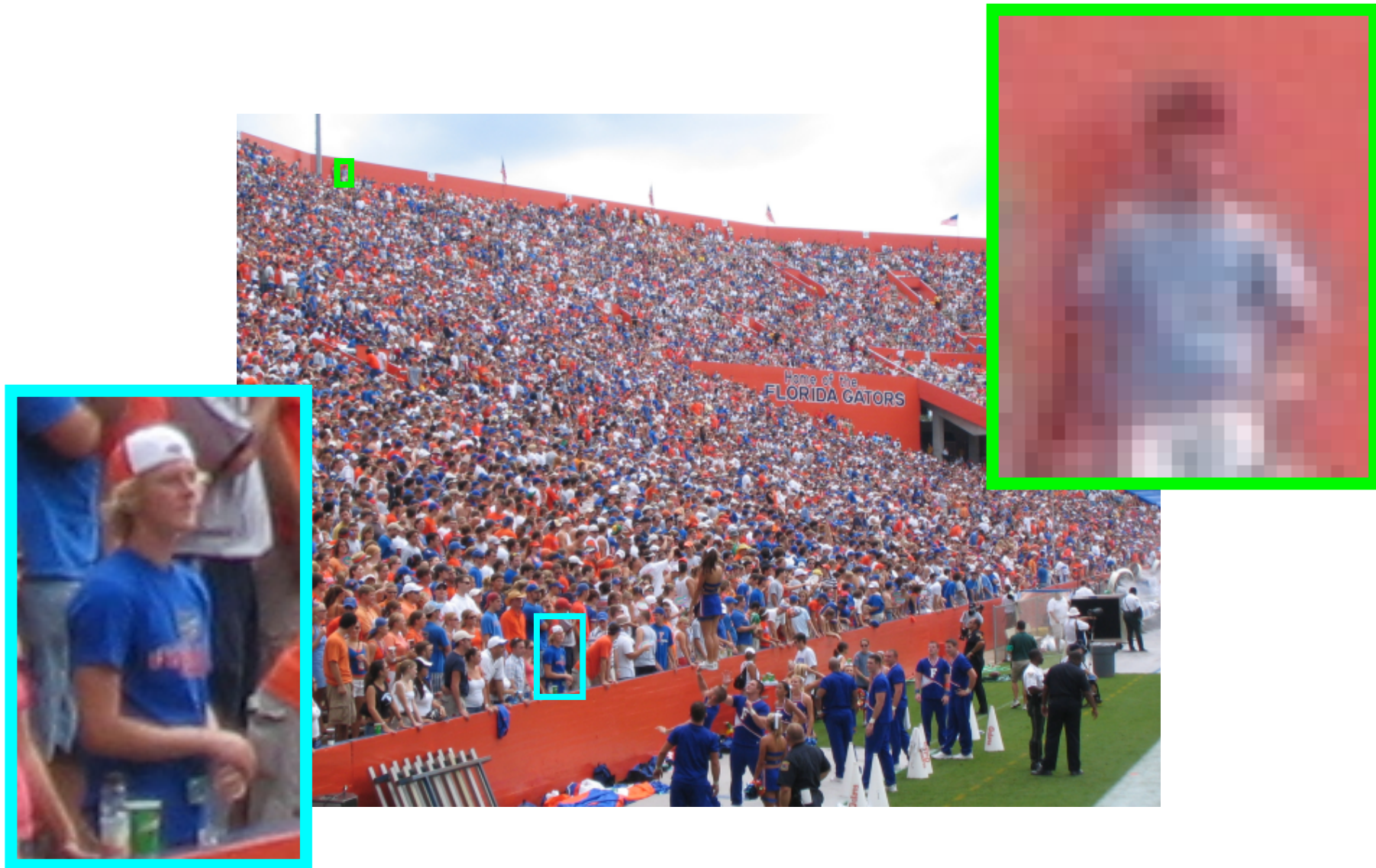


Recognition Task

- **Object Class Detection/Localization**
 - Where are the aeroplanes (if any)?
- **Challenges**
 - Imaging factors e.g. lighting, pose, occlusion, clutter
 - Intra-class variation
- **Compared to Classification**
 - Detailed prediction e.g. bounding box
 - Location usually provided for training



Challenges: Scale



Challenges: Background Clutter



Challenges: Occlusion and truncation



Challenges: Intra-class variation



Object Category Recognition by Learning

- Difficult to define model of a category. Instead, learn from example images



Level of Supervision for Learning

Image-level label



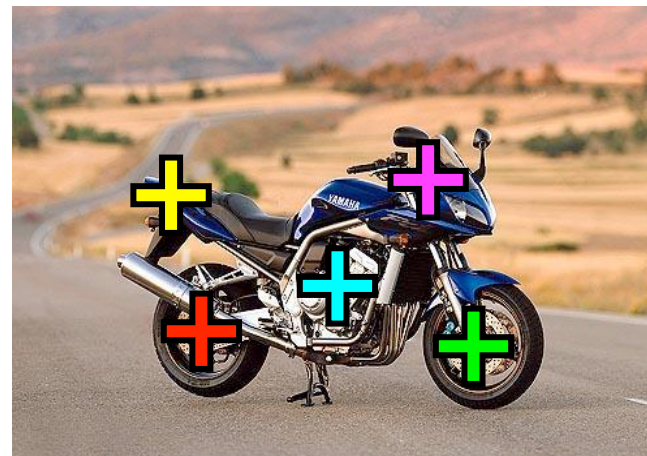
Bounding box



Pixel-level segmentation



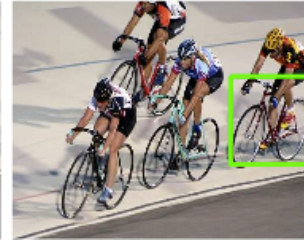
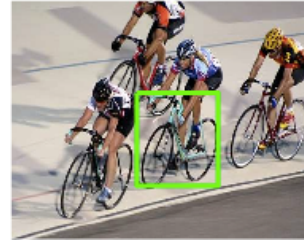
“Parts”



Preview of typical results



aeroplane



bicycle



car



cow



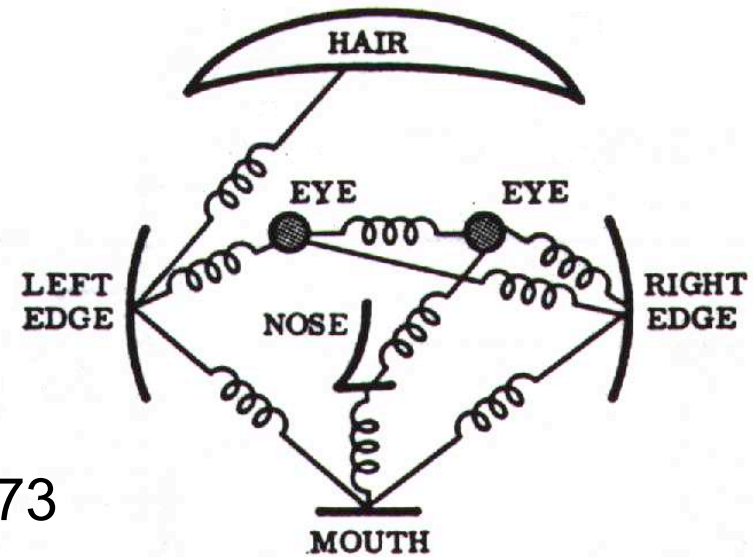
horse



motorbike

Class of model: Pictorial Structure

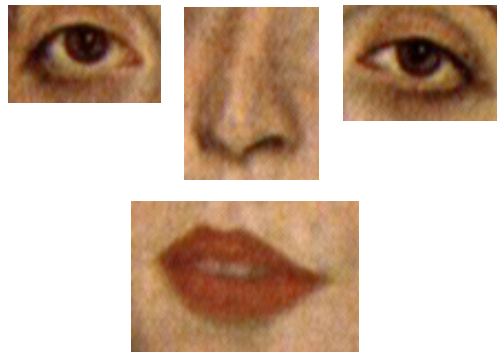
- Intuitive model of an object
- Model has two components
 1. parts (2D image fragments)
 2. structure (configuration of parts)
- Dates back to Fischler & Elschlager 1973



Is this complexity of representation necessary ?

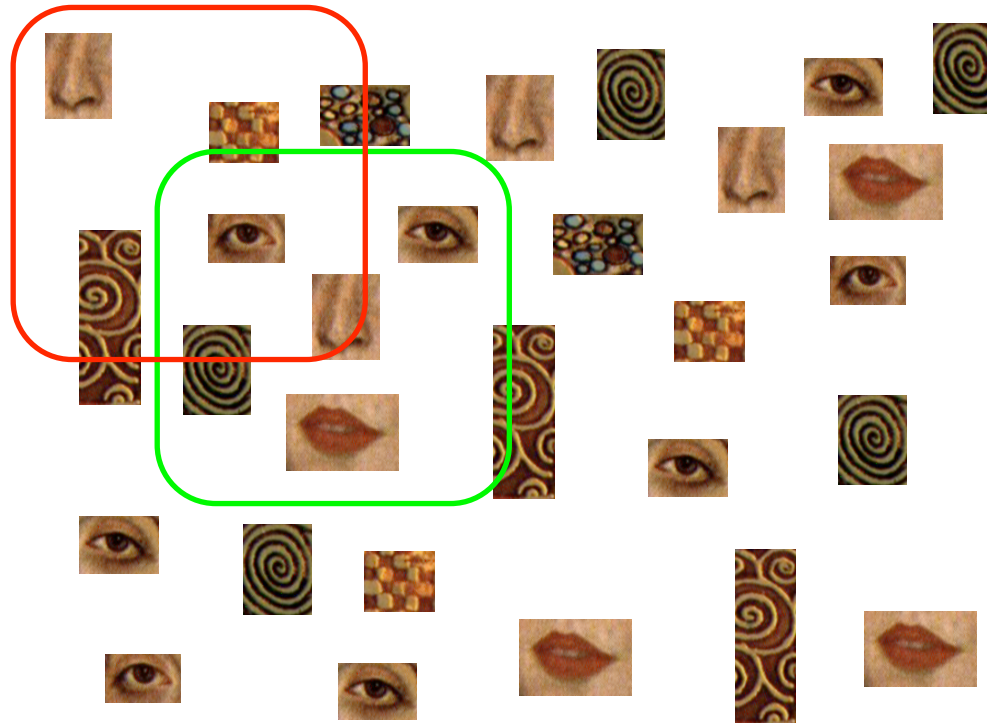
Which features?

Restrict deformations



Problem of background clutter

- Use a sub-window
 - At correct position, no clutter is present
 - Slide window to detect object
 - Change size of window to search over scale



Outline

1. Sliding window detectors
2. Features and adding spatial information
3. Histogram of Oriented Gradients (HOG)
4. Two state of the art algorithms and PASCAL VOC
5. The future and challenges

Outline

1. Sliding window detectors

- Start: feature/classifier agnostic
- Method
- Problems/limitations

2. Features and adding spatial information

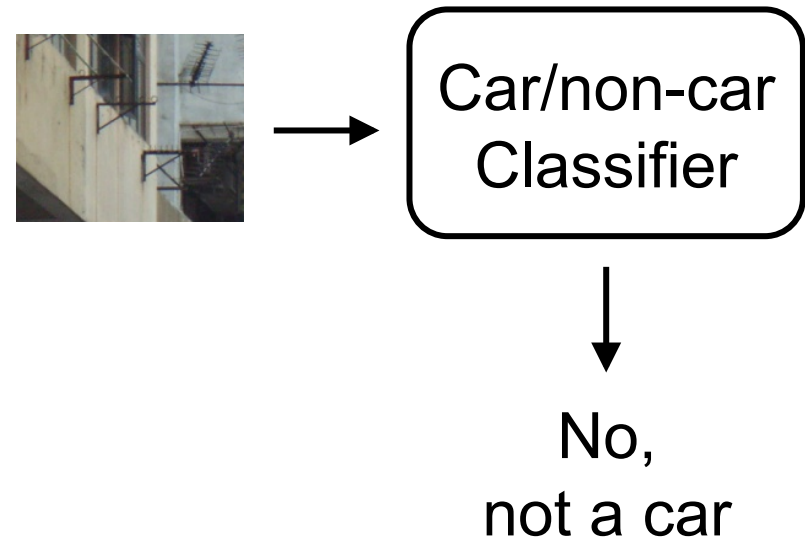
3. Histogram of Oriented Gradients (HOG)

4. Two state of the art algorithms and PASCAL VOC

5. The future and challenges

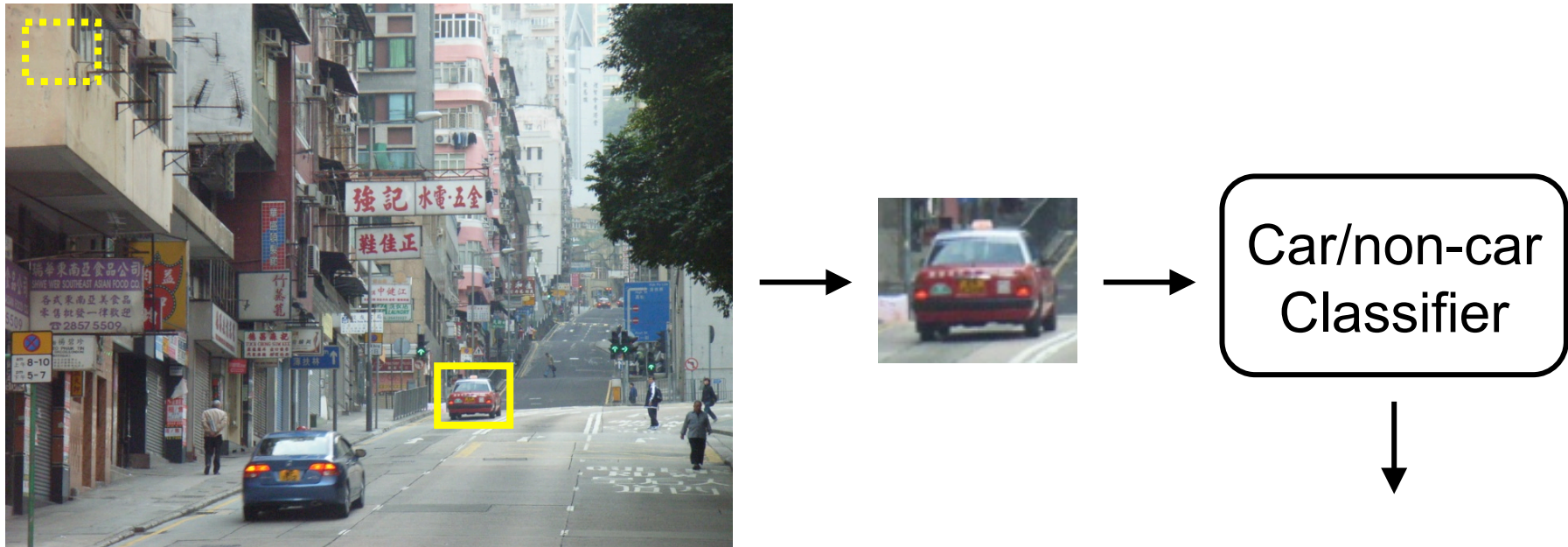
Detection by Classification

- Basic component: binary classifier



Detection by Classification

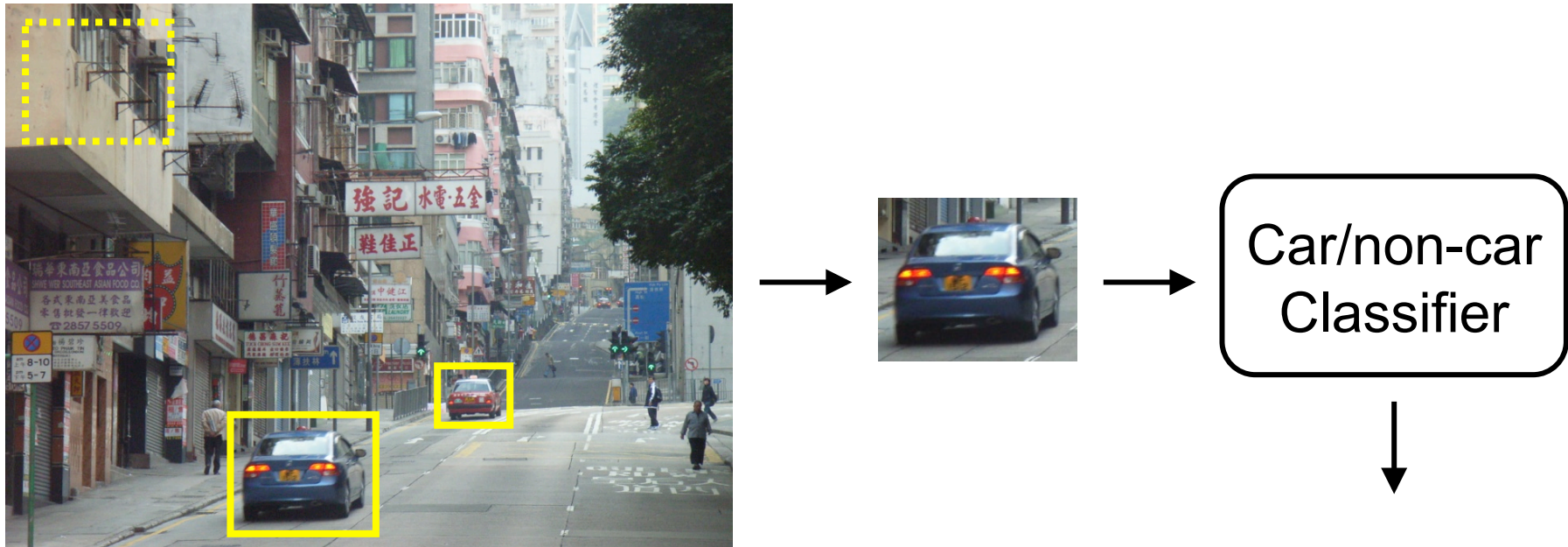
- Detect objects in clutter by search



- **Sliding window:** exhaustive search over position and scale

Detection by Classification

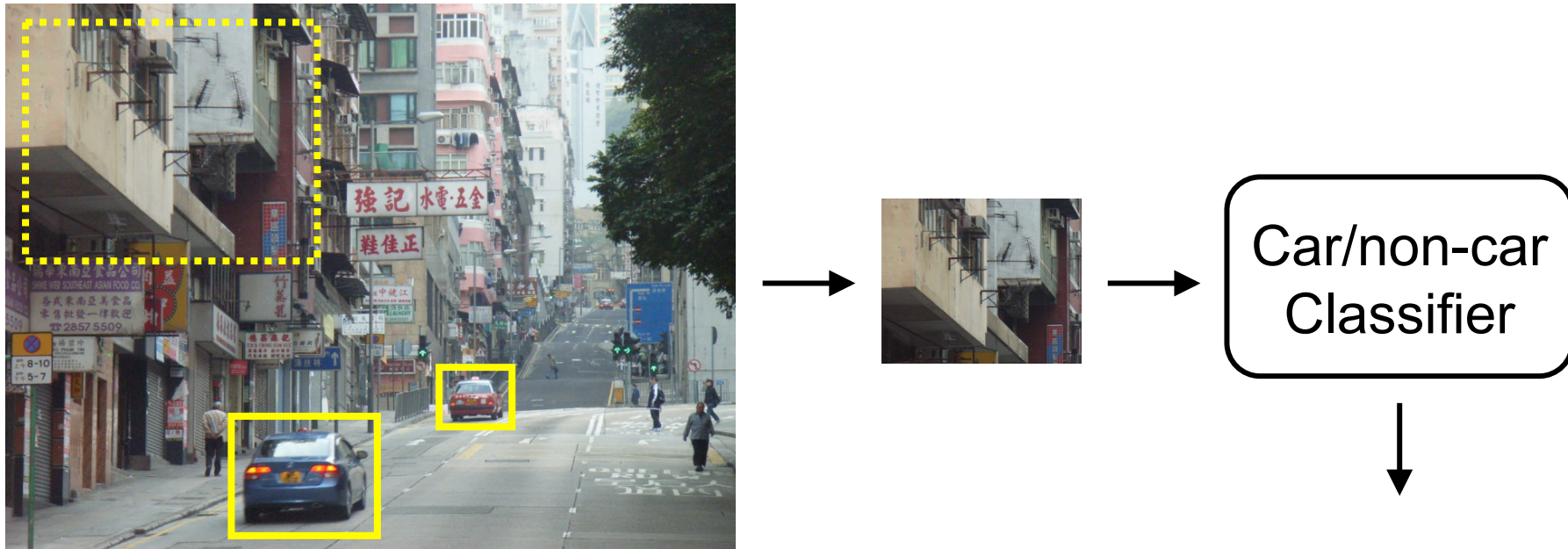
- Detect objects in clutter by search



- **Sliding window:** exhaustive search over position and scale

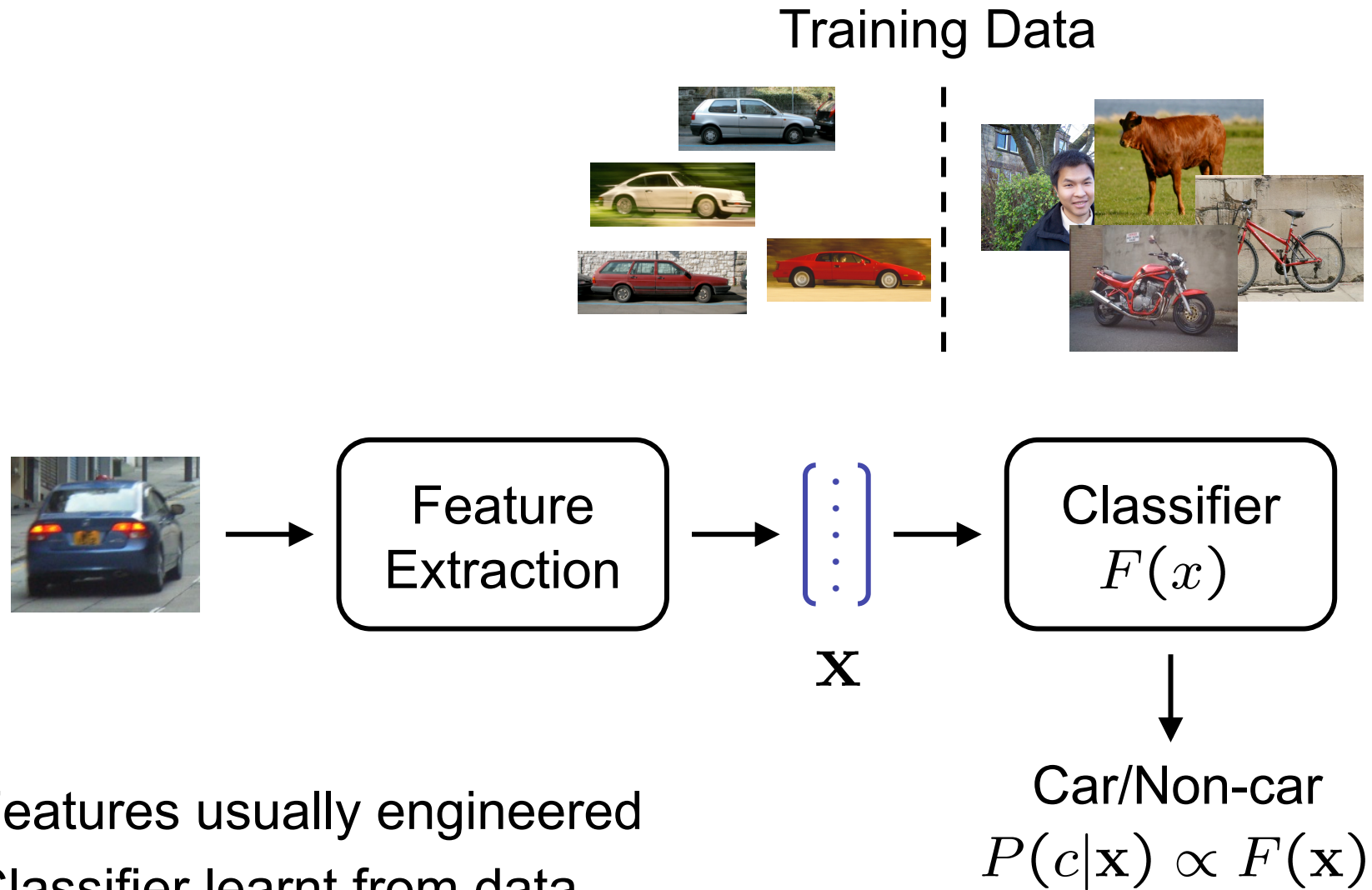
Detection by Classification

- Detect objects in clutter by search



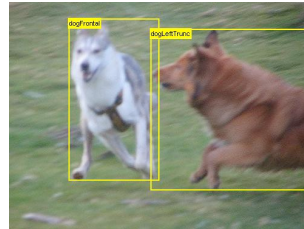
- **Sliding window:** exhaustive search over position and scale
(can use same size window over a spatial pyramid of images)

Window (Image) Classification



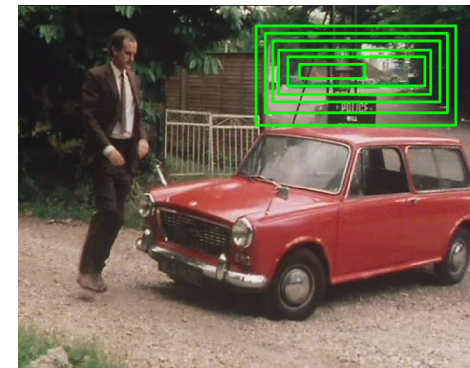
Problems with sliding windows ...

- aspect ratio
- granularity (finite grid)
- partial occlusion
- multiple responses



See recent work by

- Christoph Lampert et al CVPR 08, ECCV 08

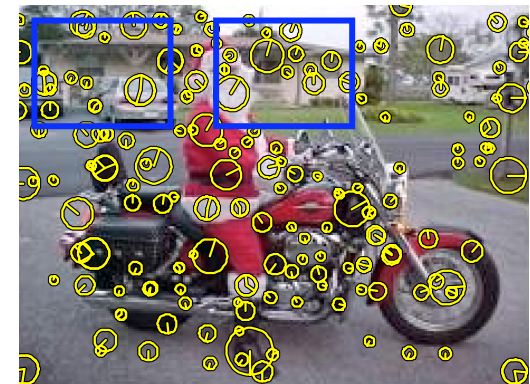
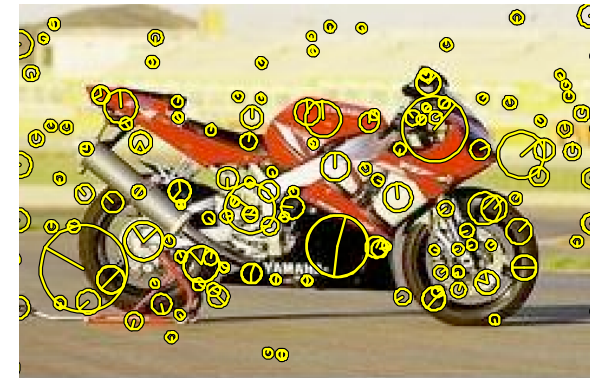
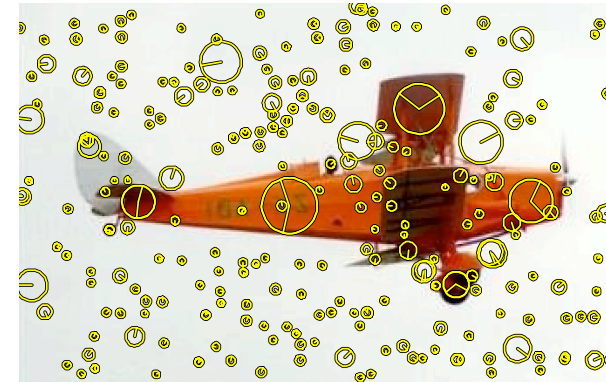


Outline

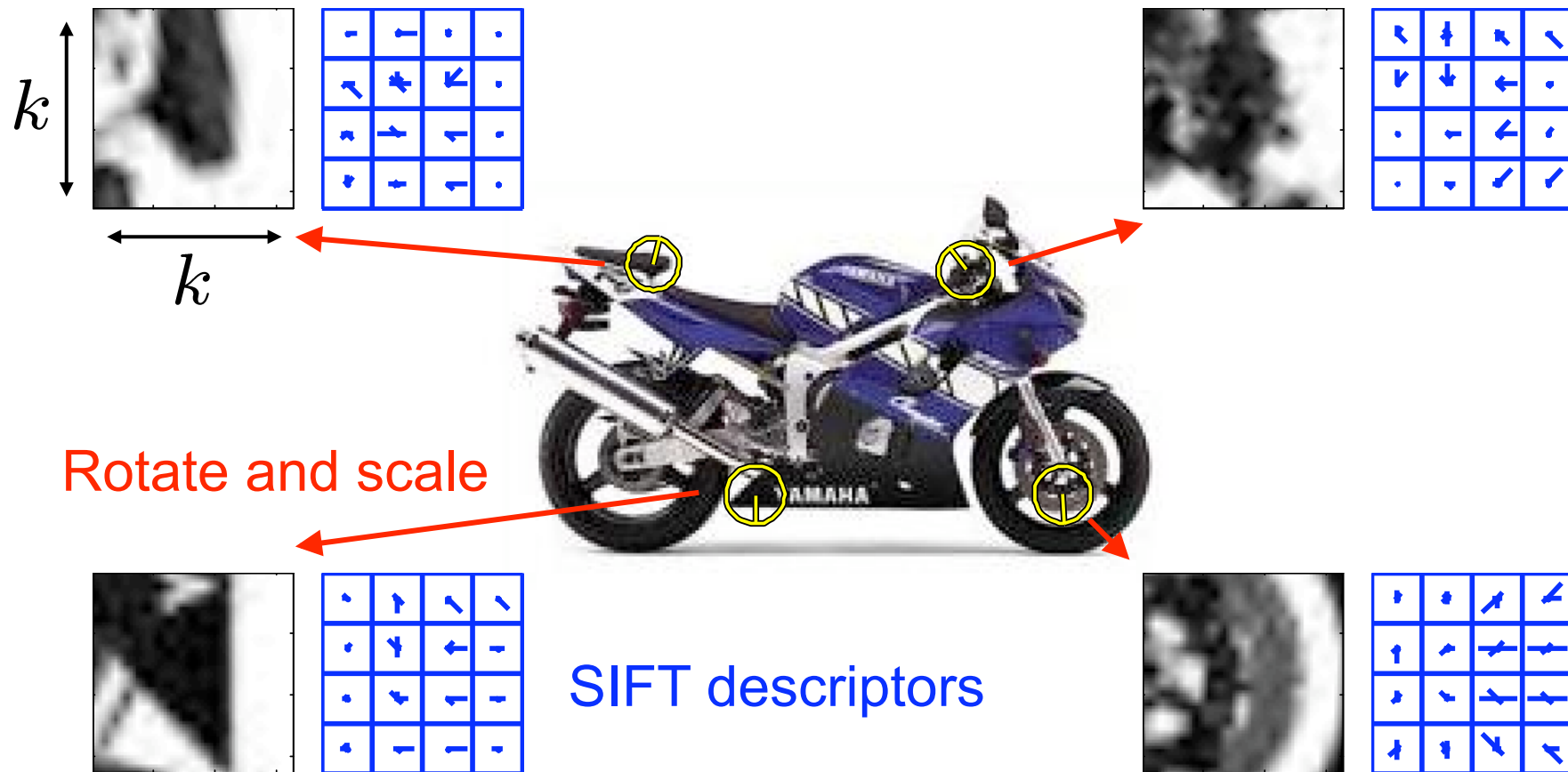
1. Sliding window detectors
2. Features and adding spatial information
 - Bag of visual word (BoW) models
 - Beyond BoW I: Constellation and ISM models
 - Beyond BoW II: Grids and spatial pyramids
3. Histogram of Oriented Gradients (HOG)
4. Two state of the art algorithms and PASCAL VOC
5. The future and challenges

Recap: Bag of (visual) Words representation

- Detect affine invariant local features (e.g. affine-Harris)
- Represent by high-dimensional descriptors, e.g. 128-D for SIFT
- How to summarize sliding window content in a fixed-length vector for classification?
 1. Map descriptors onto a common vocabulary of **visual words**
 2. Represent image as a histogram over visual words – a **bag of words**



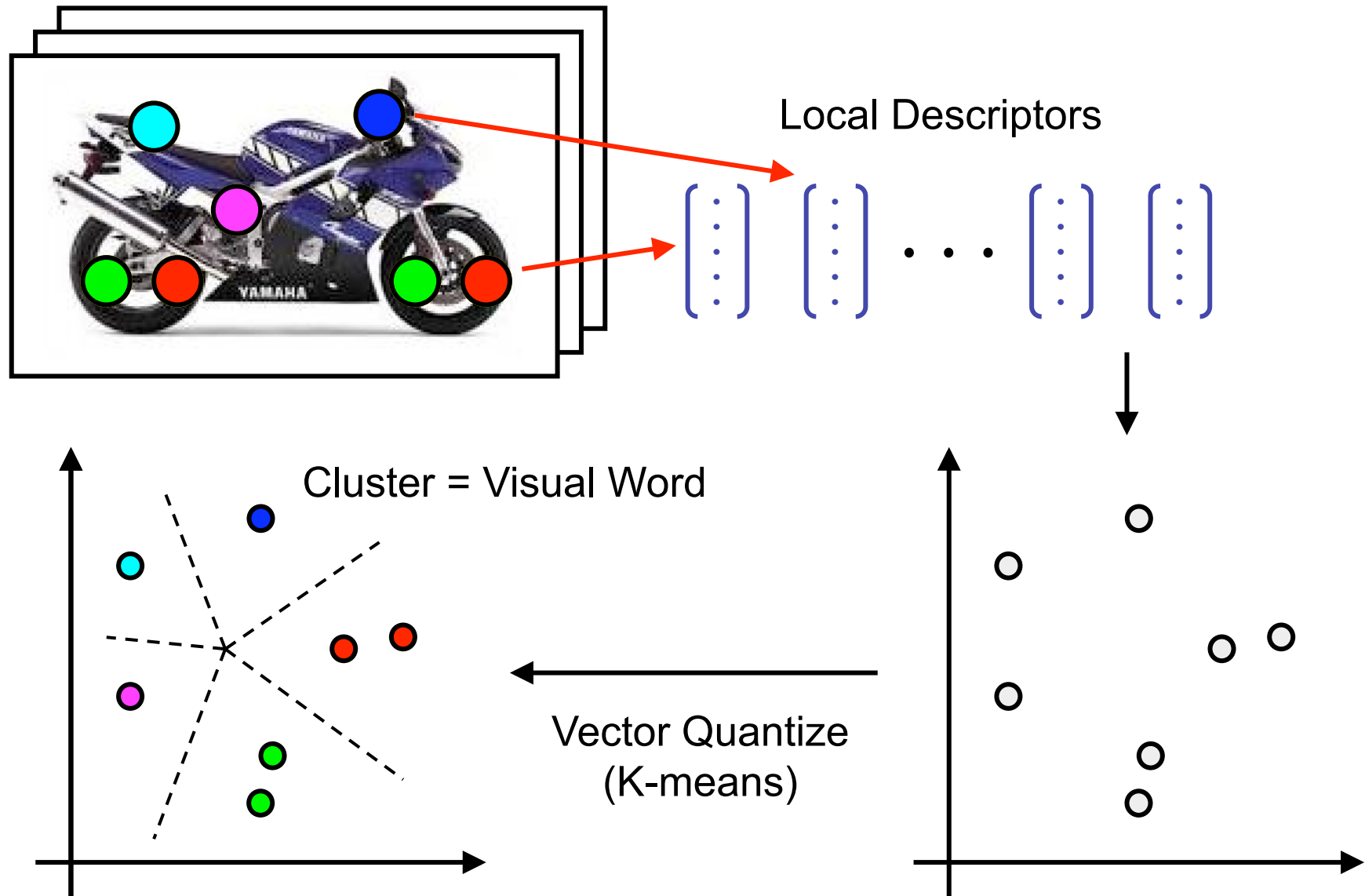
Local region descriptors and visual words



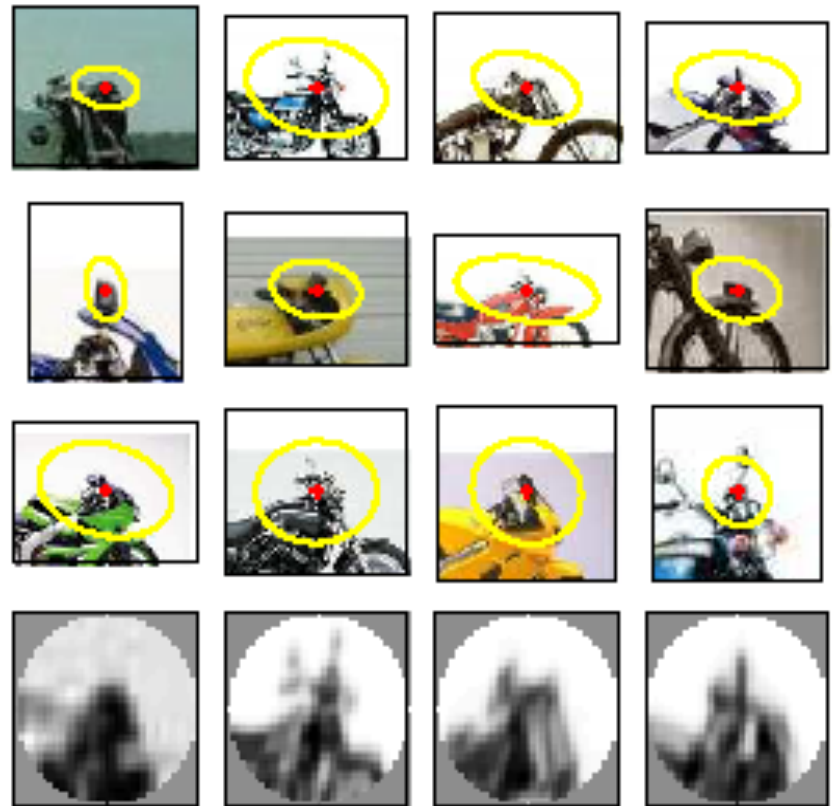
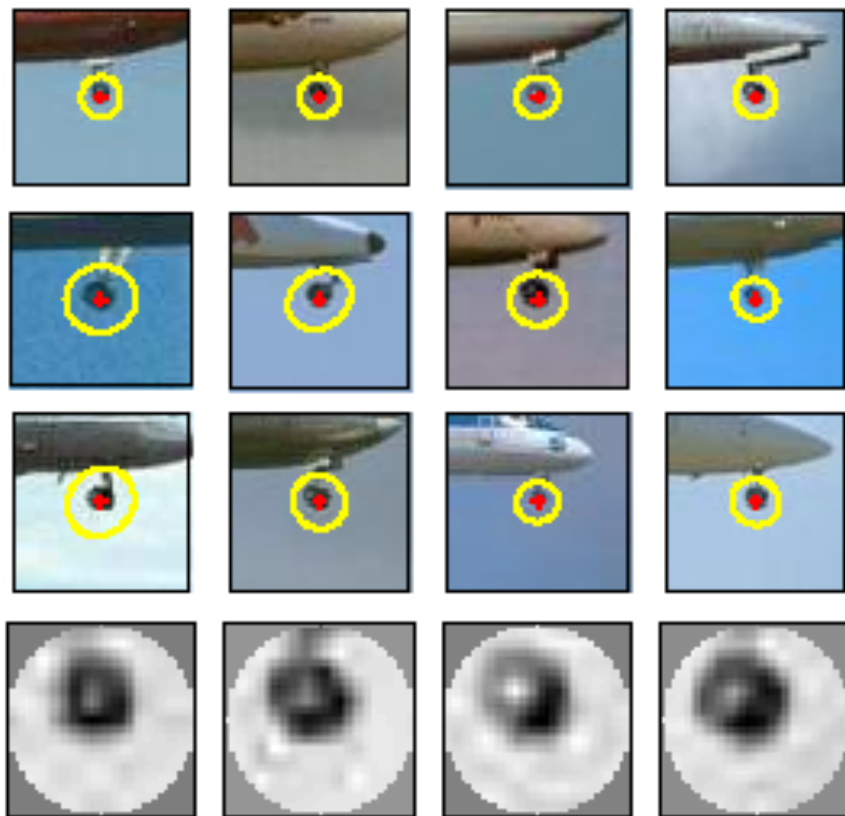
- Normalize regions to fixed size and shape
- Describe each region by a SIFT descriptor
- Vector quantize into visual words, e.g. using k-means

NB: aff. detectors/SIFT/visual words originally for view point invariant matching

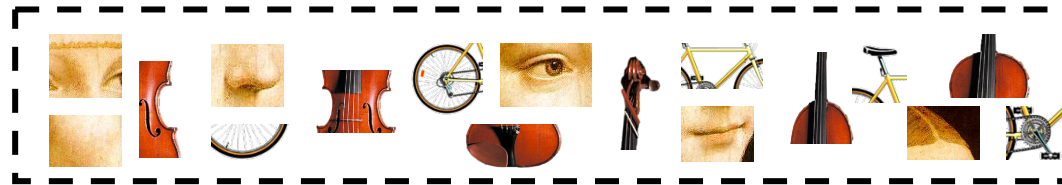
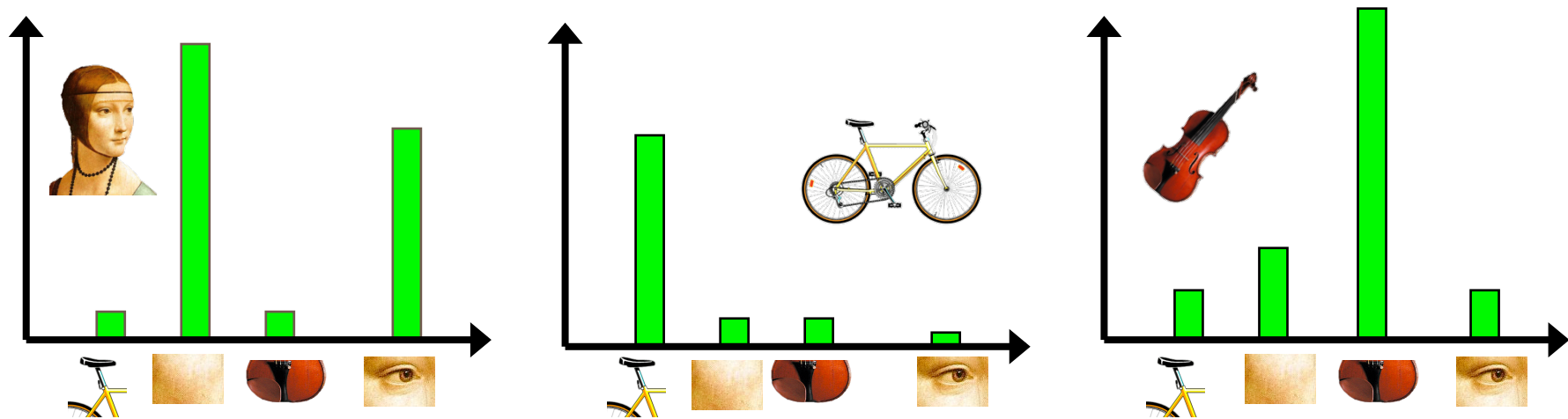
Visual Words



Example Visual Words

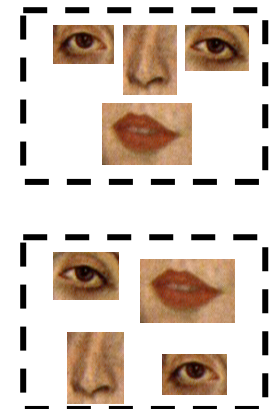


Intuition



Visual Vocabulary

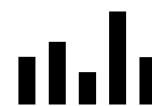
- Visual words represent “iconic” image fragments
- Feature detectors and SIFT give invariance to local rotation and scale
- Discarding spatial information gives configuration invariance



Learning from positive ROI examples



Bag of Words

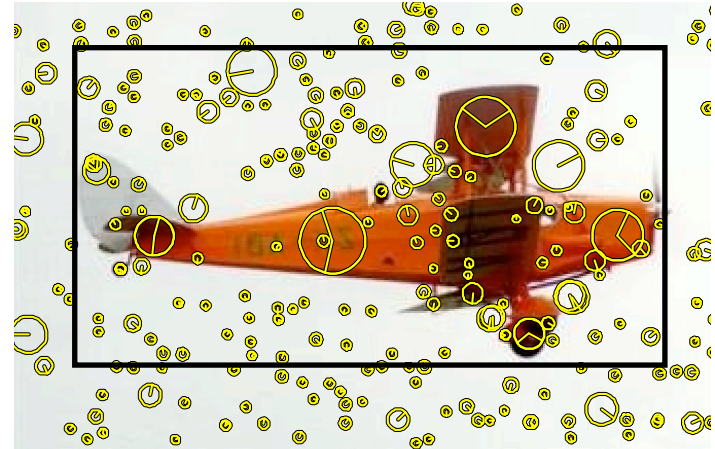


Feature Vector

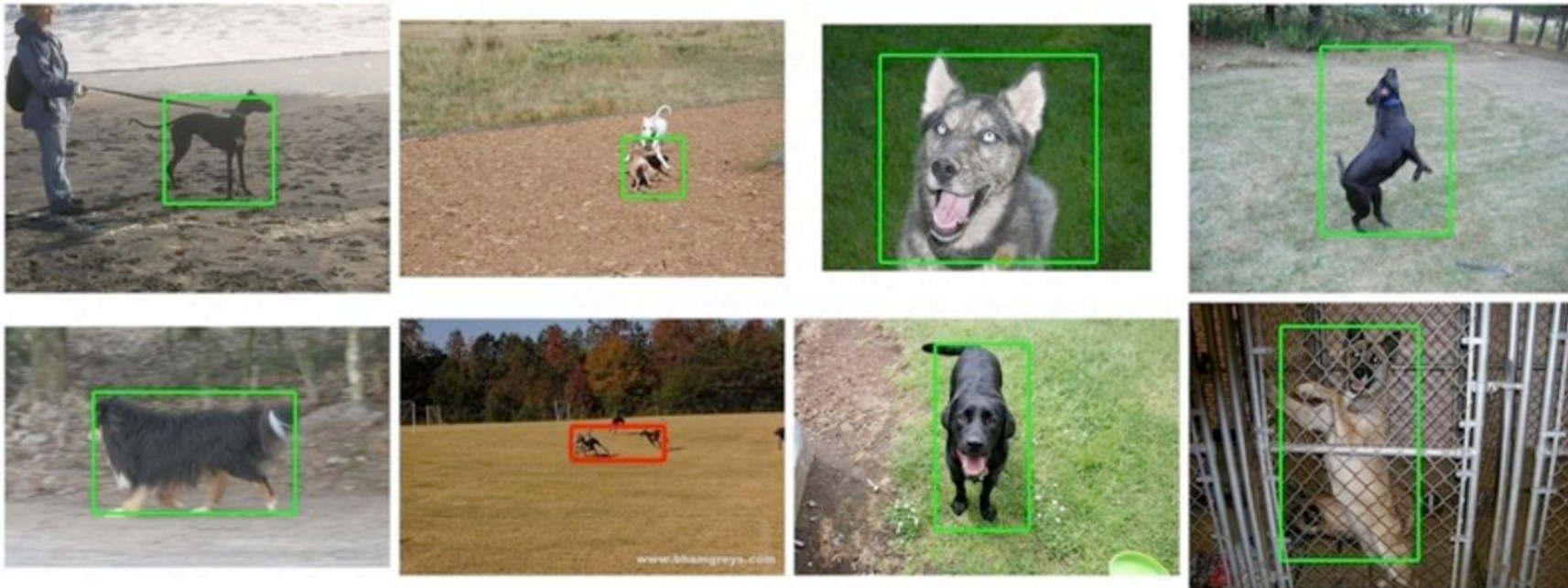


Sliding window detector

- Classifier: SVM with linear kernel
- BOW representation for ROI



Example detections for dog



Discussion: ROI as a Bag of Visual Words

- Advantages

- No explicit modelling of spatial information -> high level of invariance to position and orientation in image
- Fixed length vector -> standard machine learning methods applicable



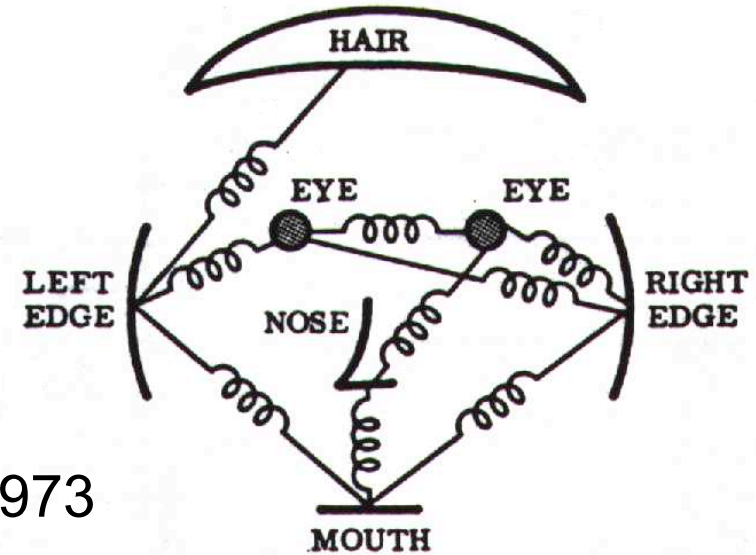
- Disadvantages

- No explicit modelling of spatial information -> less discriminative power
- Inferior to state of the art performance



Beyond BOW I: Pictorial Structure

- Intuitive model of an object
- Model has two components
 1. parts (2D image fragments)
 2. structure (configuration of parts)
- Dates back to Fischler & Elschlager 1973

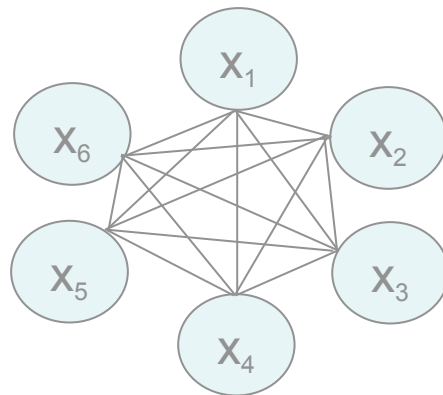


Two approaches that have investigated this spring like model:

- Constellation model
- Implicit shape model

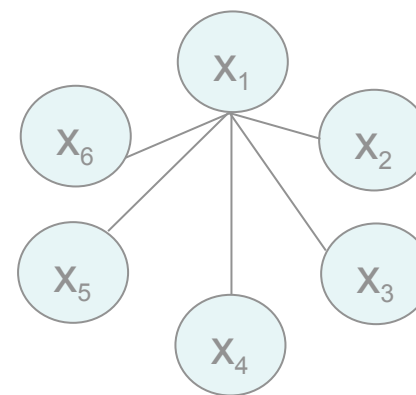
Spatial Models Considered

Fully connected shape model



e.g. Constellation Model
Parts fully connected
Recognition complexity: $O(N^P)$
Method: Exhaustive search

“Star” shape model

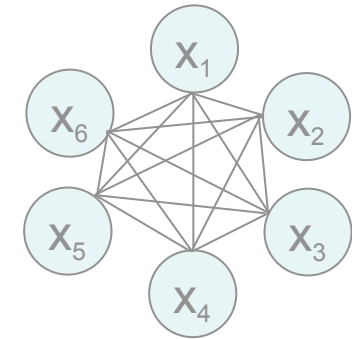


e.g. ISM
Parts mutually independent
Recognition complexity: $O(NP)$
Method: Gen. Hough Transform

Constellation model

Fergus, Perona & Zisserman, CVPR 03

- Explicit structure model – Joint Gaussian over all part positions
- Part detector determines position *and* scale
- Simultaneous learning of parts and structure
- Learn from images alone using EM algorithm

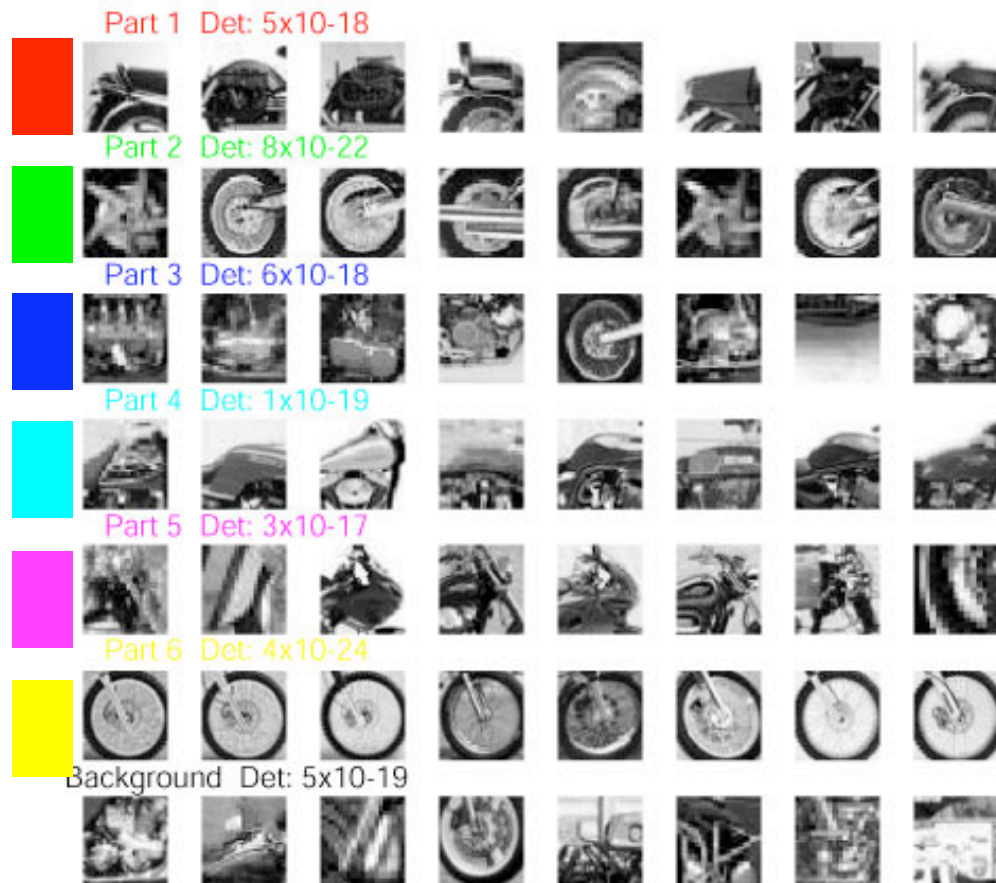


Given detections: learn a six part model by optimizing part and configuration similarity

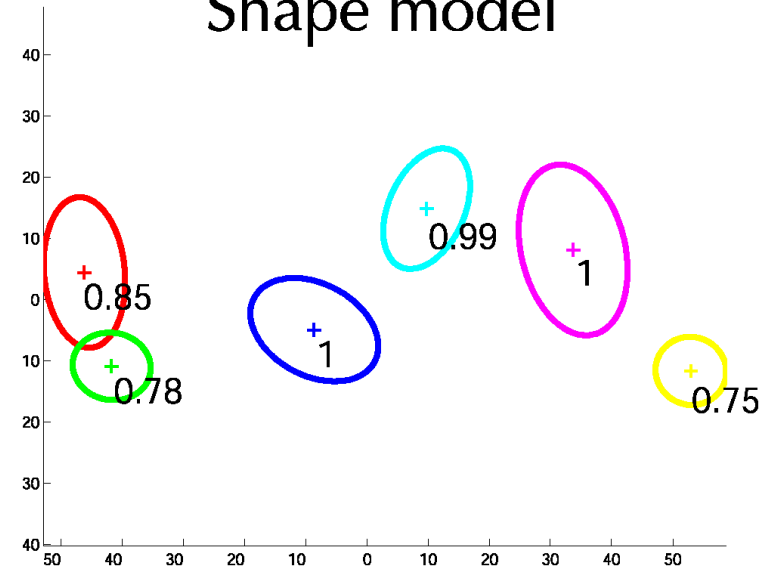


Example – Learnt Motorbike Model

Samples from appearance model

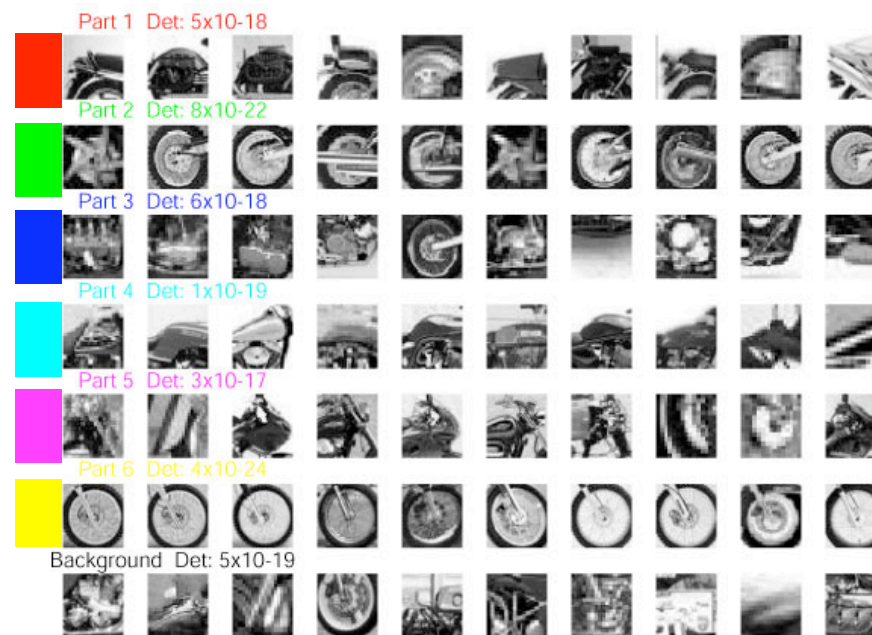
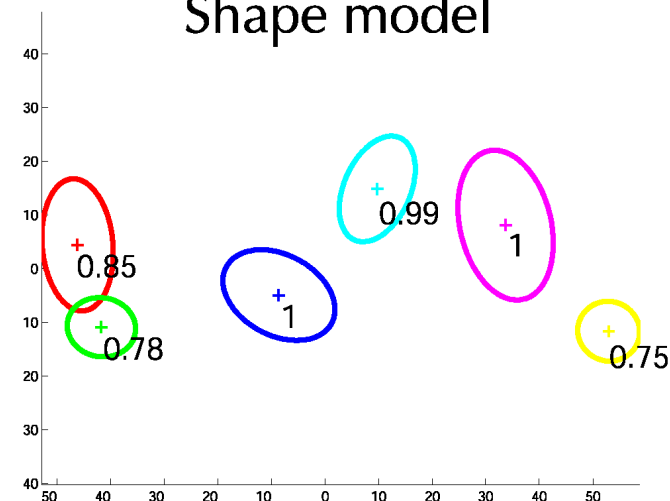
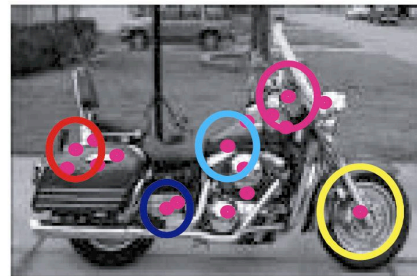
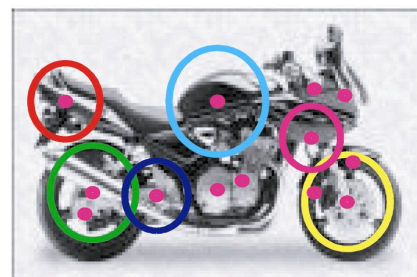
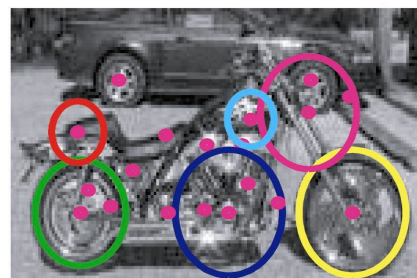
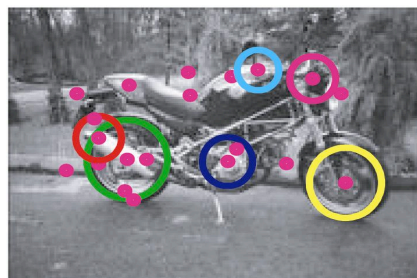
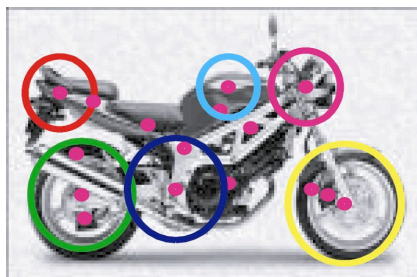


Shape model



Recognized Motorbikes

Shape model



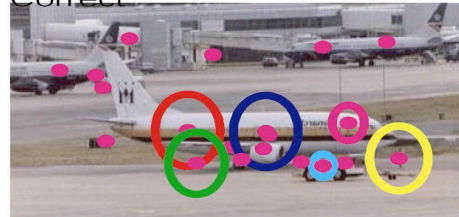
position of object determined

Airplanes

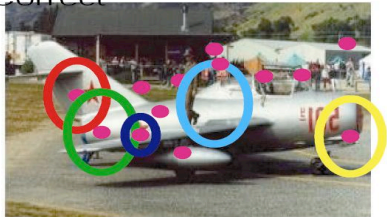
INCORRECT



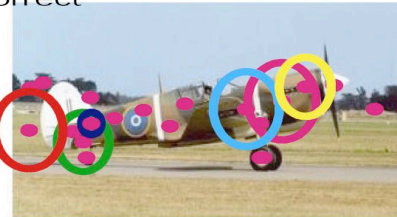
Correct



Correct



Correct



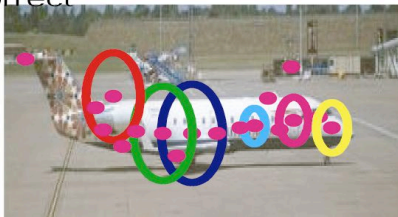
Correct



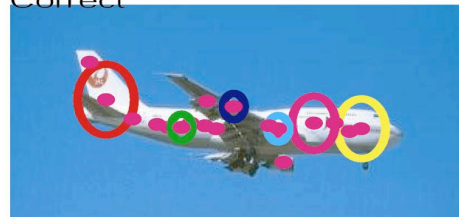
Correct



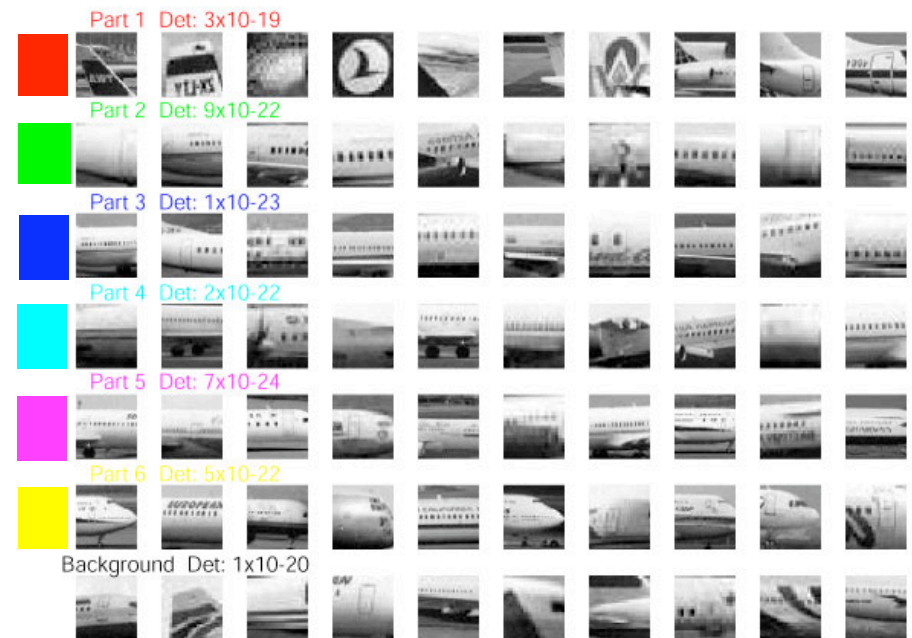
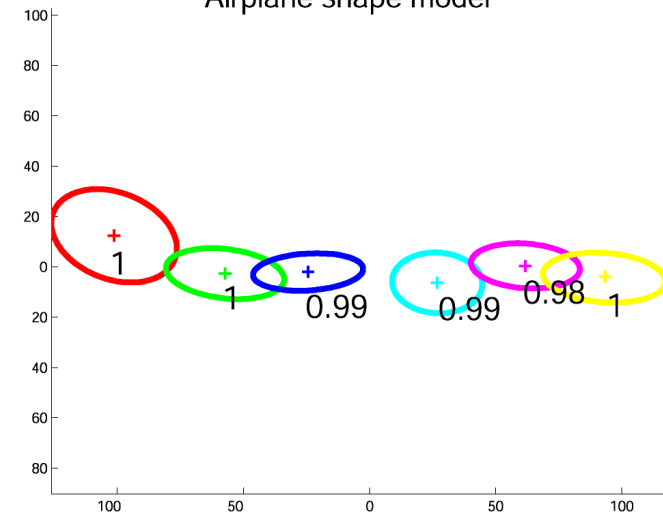
Correct



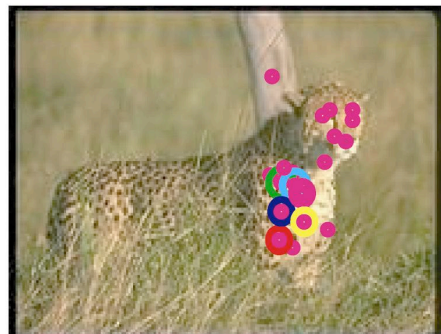
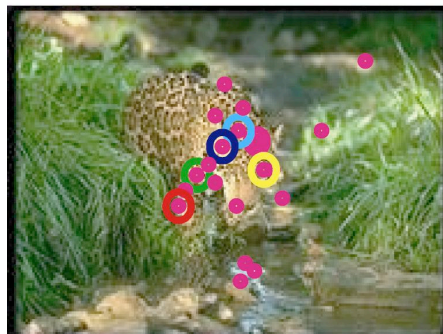
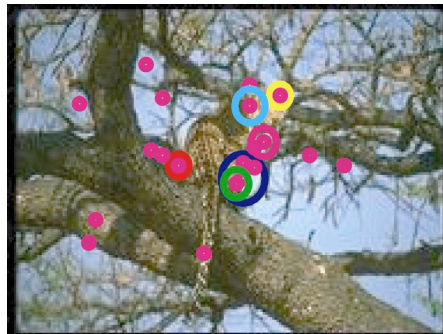
Correct



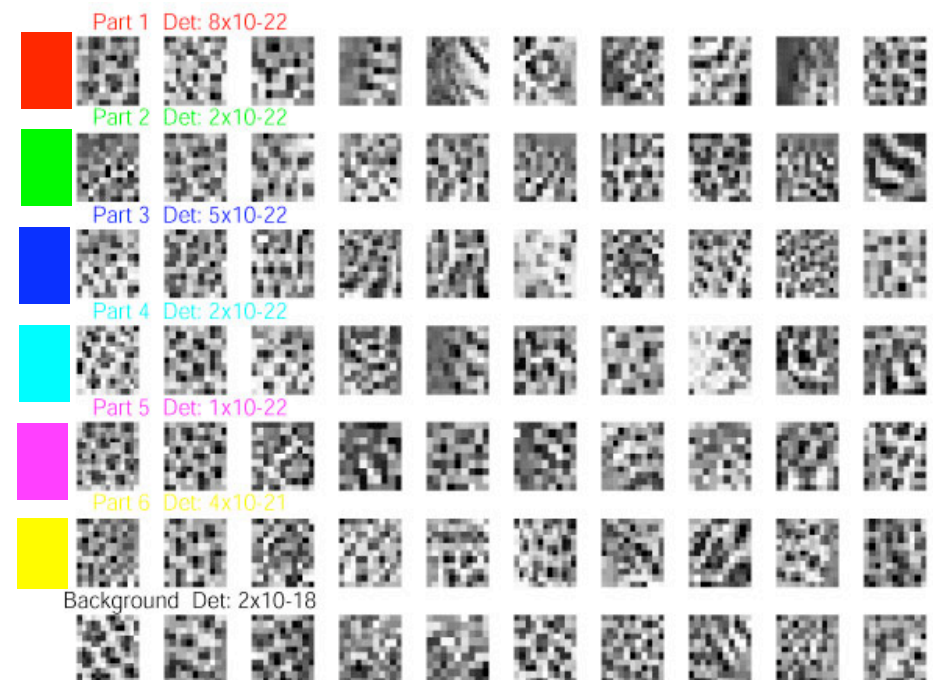
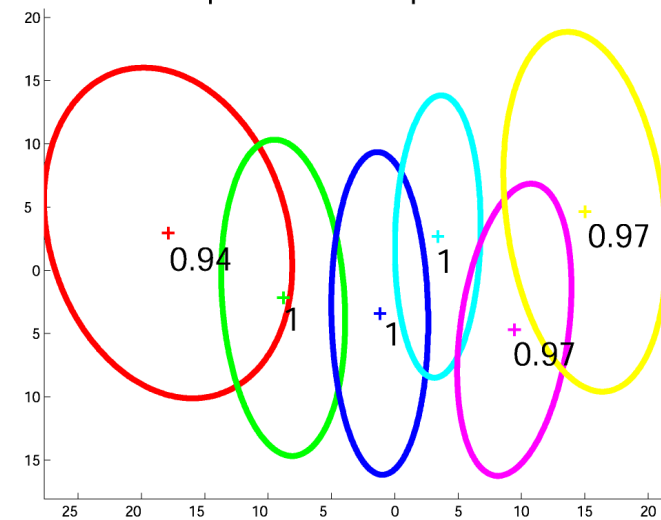
Airplane shape model



Spotted cats



Spotted cat shape model



Discussion: Constellation Model

- Advantages

- Works well for many different object categories
- Can adapt well to categories where
 - Shape is more important
 - Appearance is more important
- Everything is learned from training data
- Weakly-supervised training possible

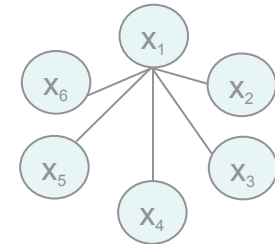
- Disadvantages

- Model contains many parameters that need to be estimated
- Cost increases exponentially with increasing number of parameters
- ⇒ Fully connected model restricted to small number of parts.

Implicit Shape Model (ISM)

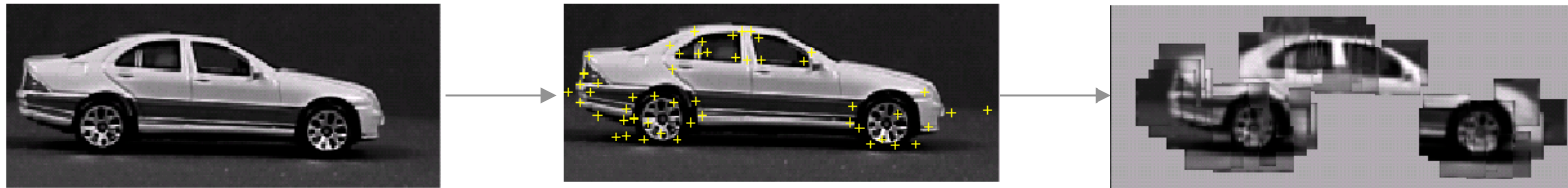
Leibe, Leonardis, Schiele, 03/04

- Basic ideas
 - Learn an appearance codebook
 - Learn a star-topology structural model
 - Features are considered independent given object centre
- Algorithm: probabilistic Generalized Hough Transform
 - Good engineering:
 - Soft assignment
 - Probabilistic voting
 - Continuous Hough space

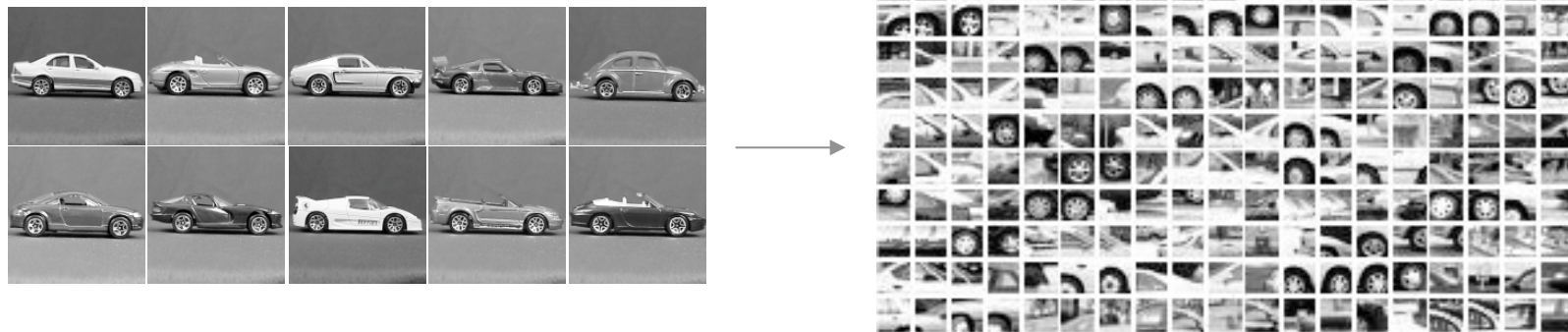


Codebook Representation

- Extraction of local object features
 - Interest Points (e.g. Harris detector)
 - Sparse representation of the object appearance



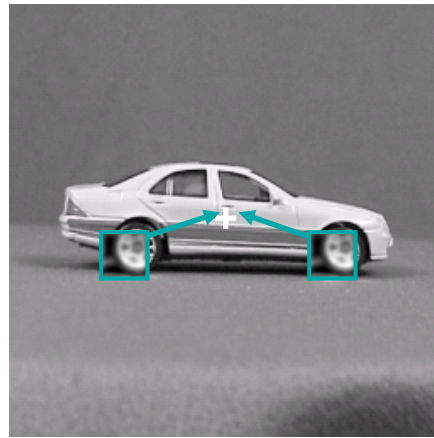
- Collect features from whole training set
- Example:



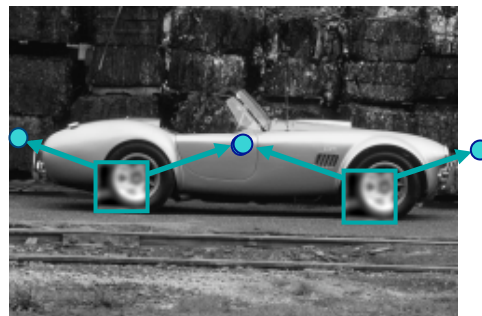
Class specific vocabulary

Leibe & Schiele 03/04: Generalized Hough Transform

- **Learning:** for every cluster, store possible “occurrences”

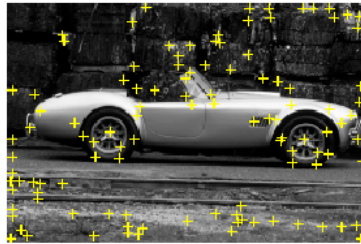


- **Recognition:** for new image, let the matched patches vote for possible object positions

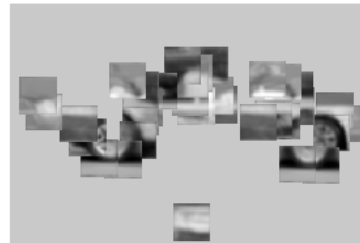


Leibe & Schiele 03/04: Generalized Hough Transform

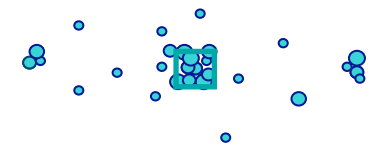
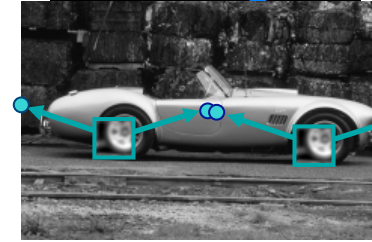
Interest Points



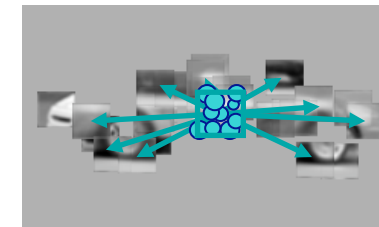
Matched Codebook Entries



Probabilistic Voting

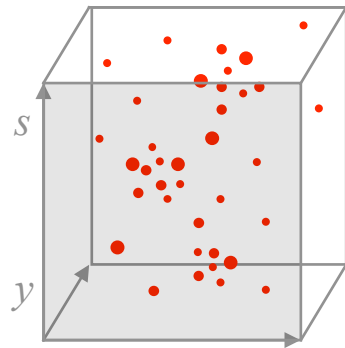


Voting Space
(continuous)

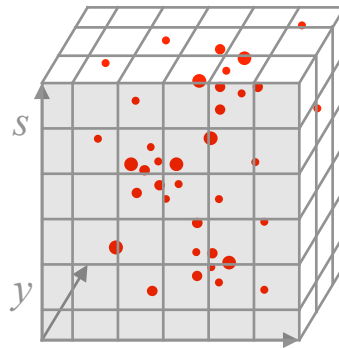


Backprojection
of Maximum

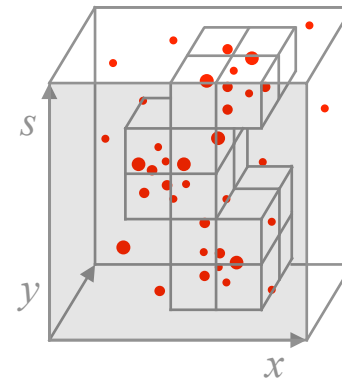
Scale Voting: Efficient Computation



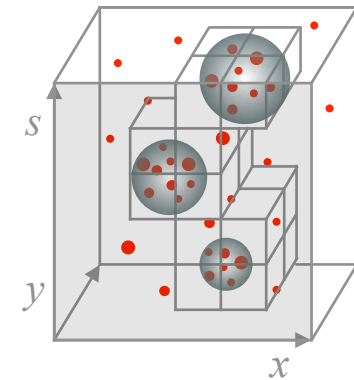
Scale votes



Binned
accum. array



Candidate
maxima



Refinement
(MSME)

- Mean-Shift formulation for refinement
 - Scale-adaptive *balloon density estimator*

$$\hat{p}(o_n, x) = \frac{1}{V_b} \sum_k \sum_j p(o_n, x_j | f_k, \ell_k) K\left(\frac{x - x_j}{b}\right)$$

Detection Results

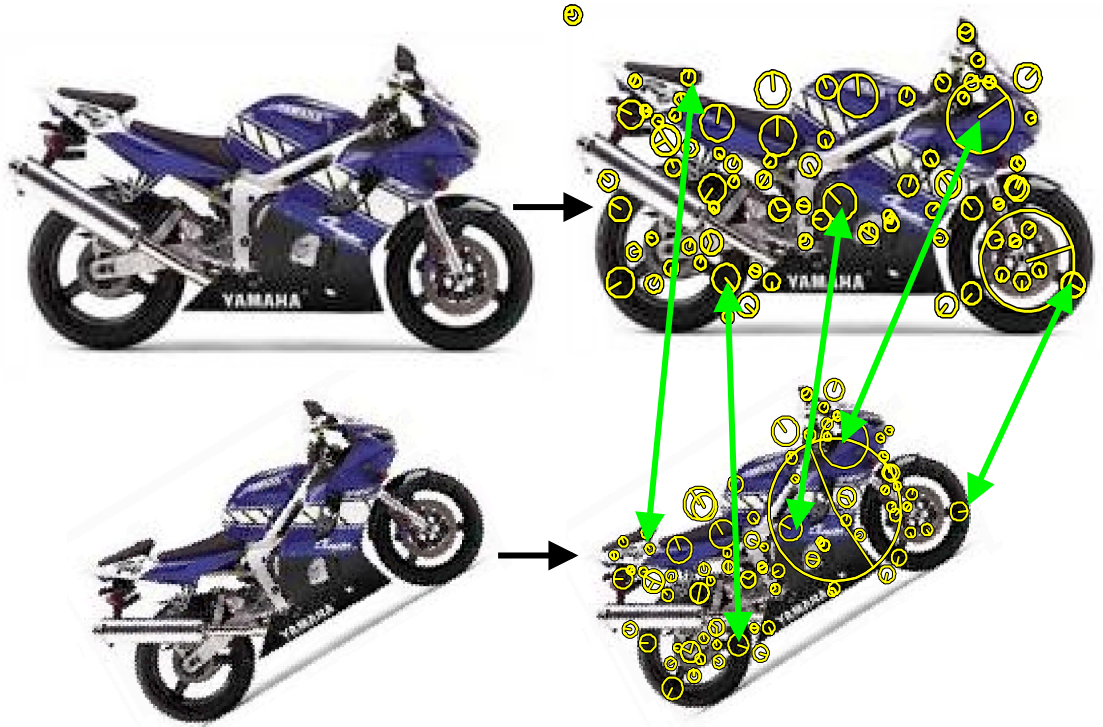
- Qualitative Performance
 - Recognizes different kinds of cars
 - Robust to clutter, occlusion, low contrast, noise



Discussion: ISM and related models

Advantages

- Scale and rotation invariance can be built into the representation from the start
- Relatively cheap to learn and test (inference)
- Works well for many different object categories
- Max-margin extensions possible, Maji & Malik, CVPR09



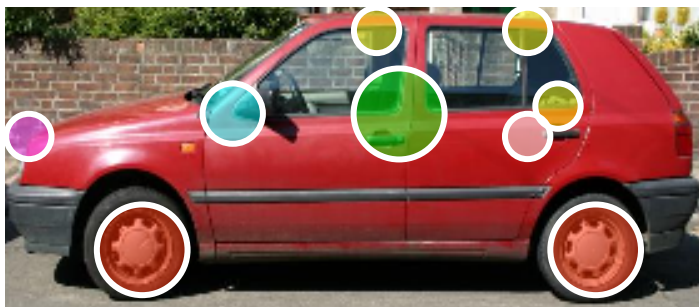
Disadvantages

- Requires searching for modes in the Hough space
- Similar to sliding window in this respect
- Is such a degree of invariance required? (many objects are horizontal)

Beyond BOW II: Grids and spatial pyramids

Start from BoW for ROI

- no spatial information recorded
- sliding window detector

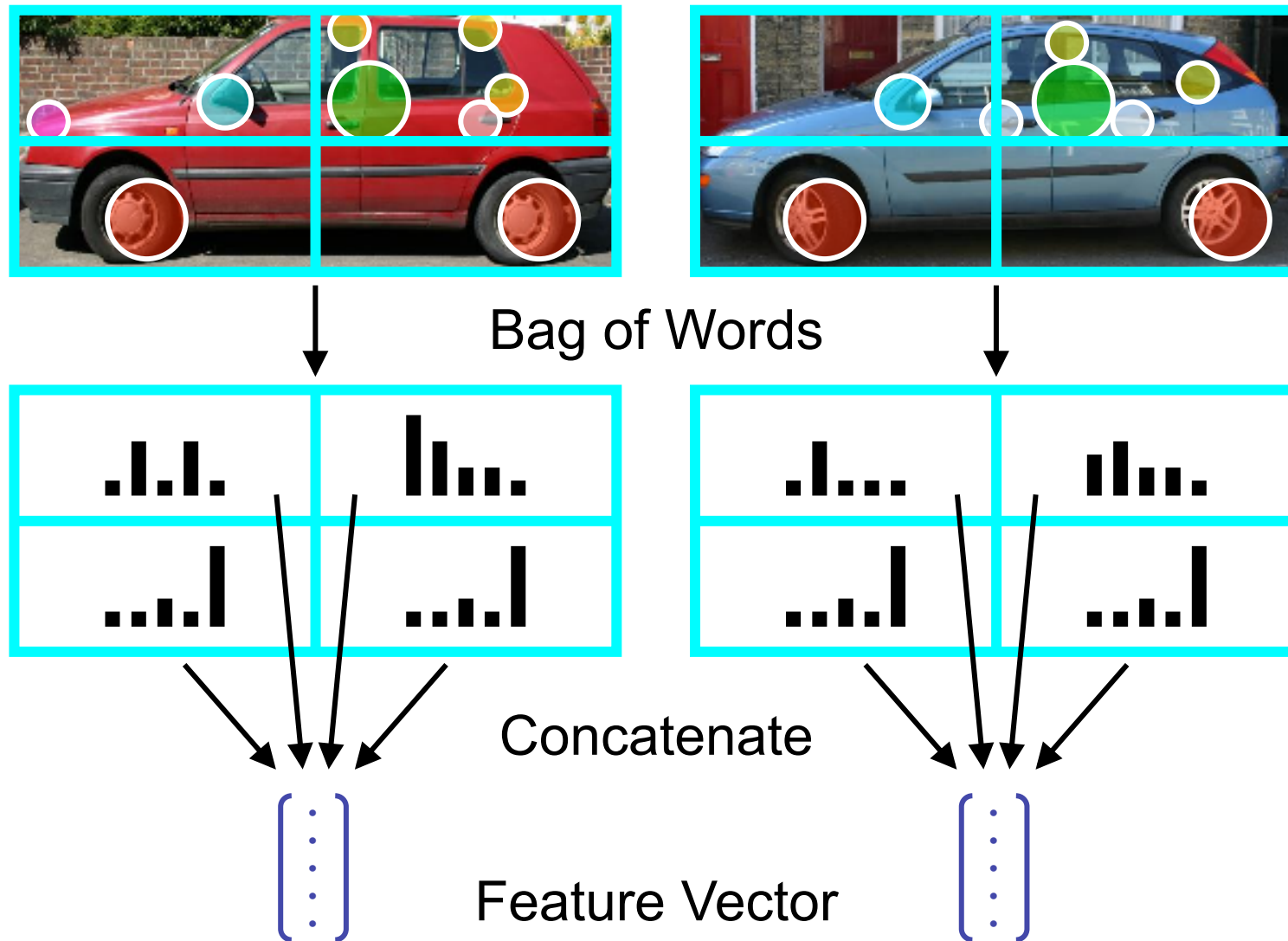


Bag of Words



Feature Vector

Adding Spatial Information to Bag of Words



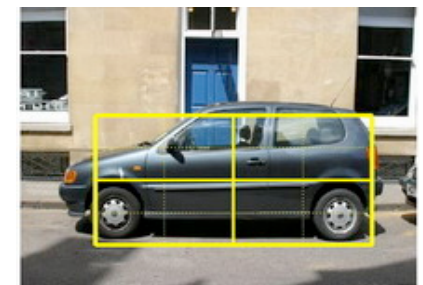
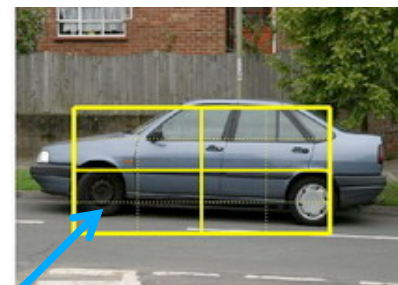
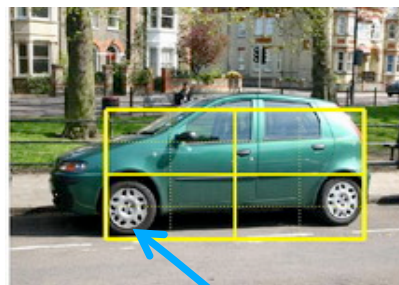
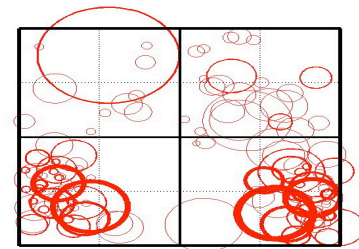
Keeps fixed length feature vector for a window

[Fergus et al, 2005]

Tiling defines (records) the spatial correspondence of the words



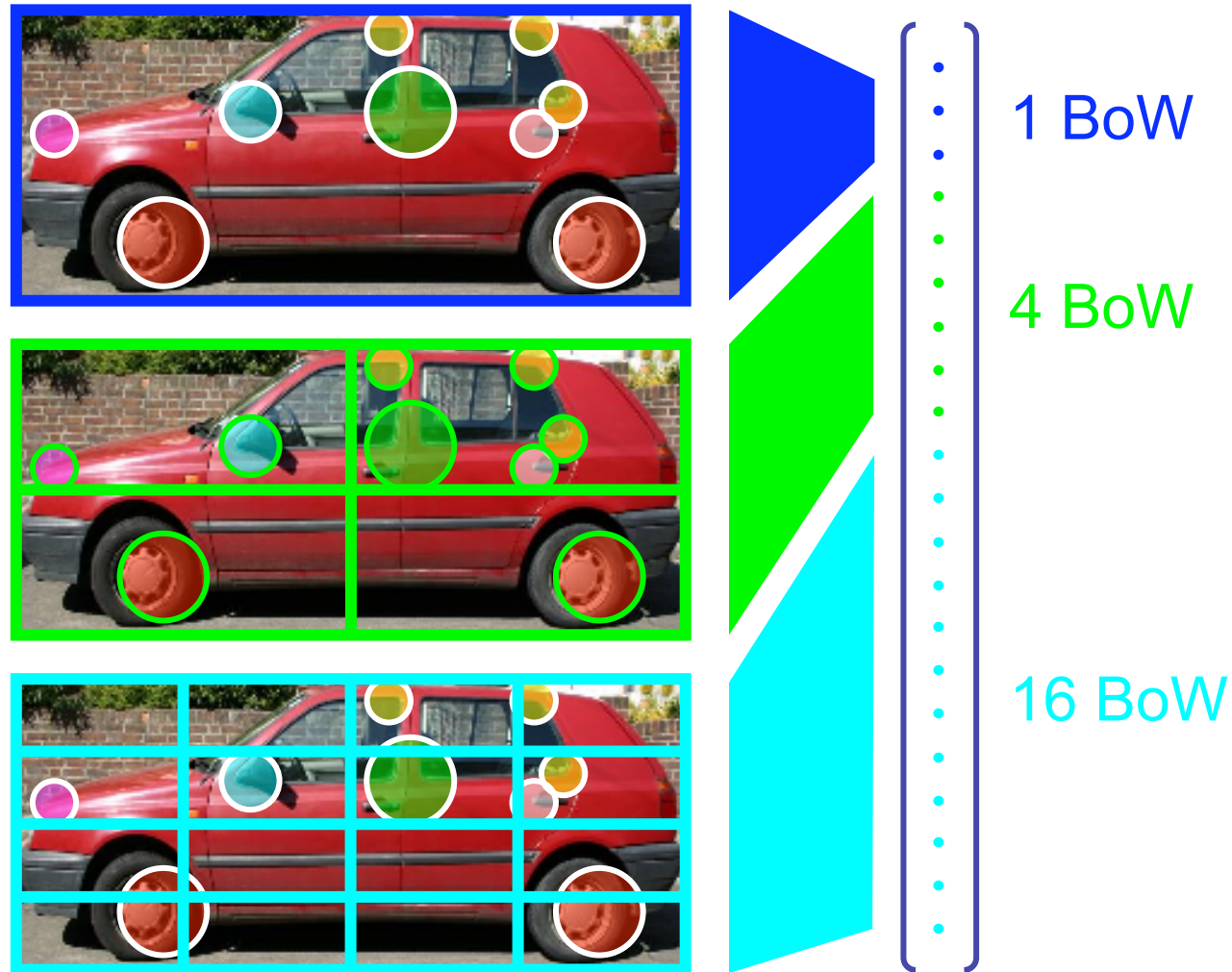
- parameter: number of tiles



If codebook has V visual words, then representation has dimension $4V$

Fergus et al ICCV 05

Spatial Pyramid – represent correspondence

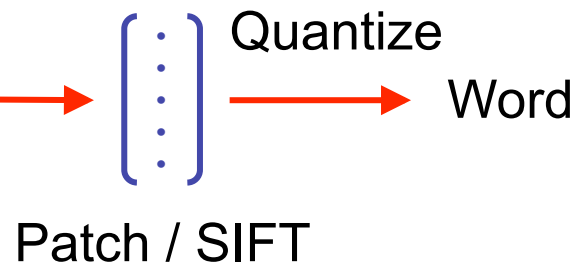
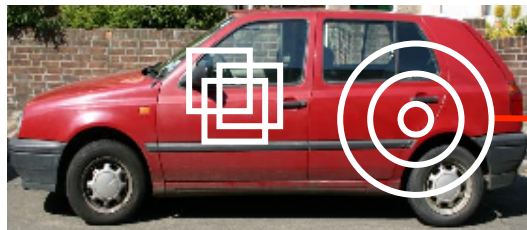
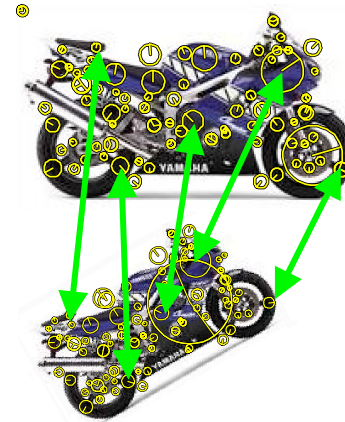


- As in scene/image classification can use pyramid kernel

[Grauman & Darrell, 2005] [Lazebnik et al, 2006]

Dense Visual Words

- Why extract only **sparse** image fragments?
- Good where lots of invariance is needed, but not relevant to sliding window detection?
- Extract **dense** visual words on an overlapping grid



[Luong & Malik, 1999]
[Varma & Zisserman, 2003]
[Vogel & Schiele, 2004]
[Jurie & Triggs, 2005]
[Fei-Fei & Perona, 2005]
[Bosch et al, 2006]

- More “detail” at the expense of invariance
- Pyramid histogram of visual words (PHOW)

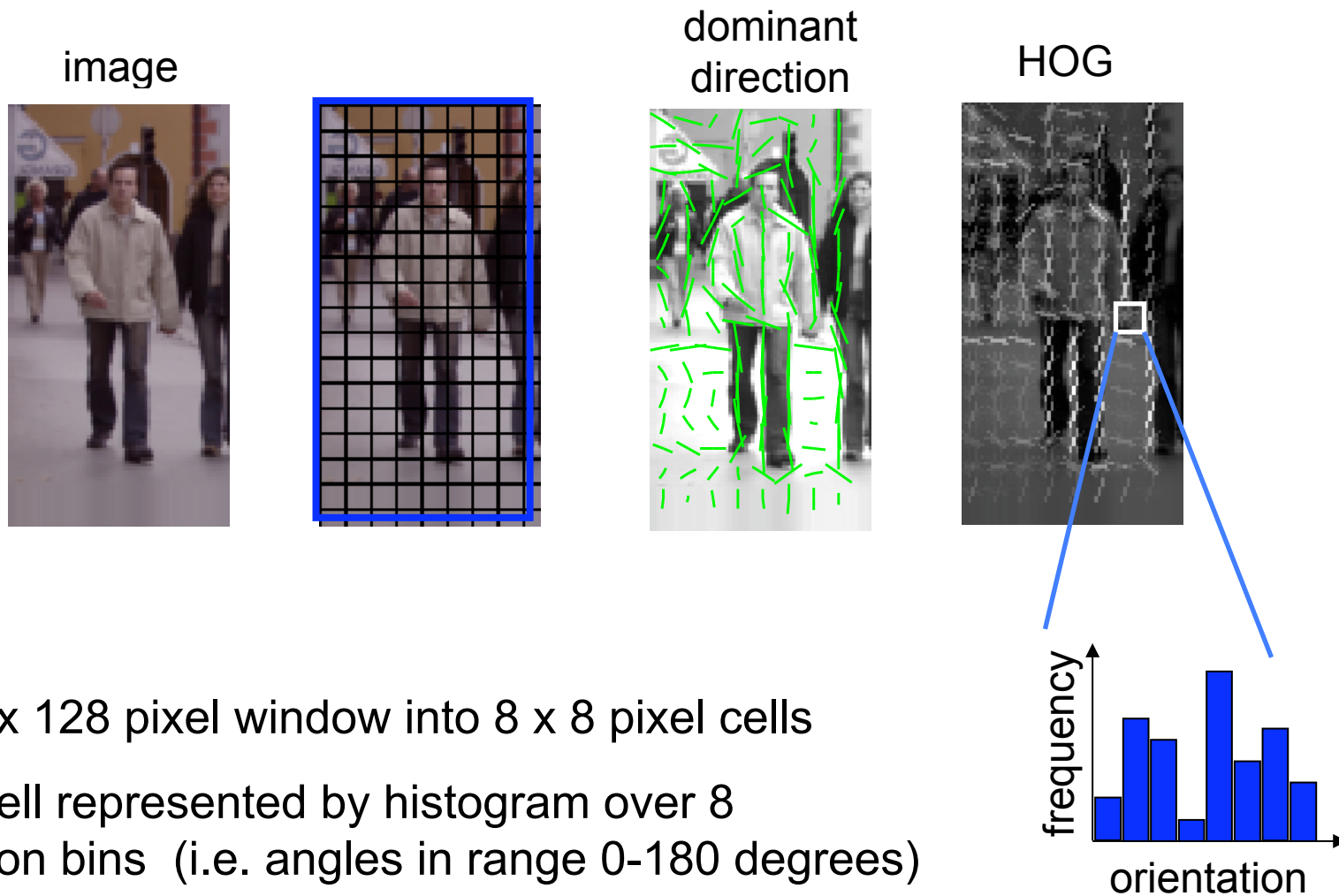
Outline

1. Sliding window detectors
2. Features and adding spatial information
3. Histogram of Oriented Gradients + linear SVM classifier
 - Dalal & Triggs pedestrian detector
 - HOG and history
 - Training an object detector
4. Two state of the art algorithms and PASCAL VOC
5. The future and challenges

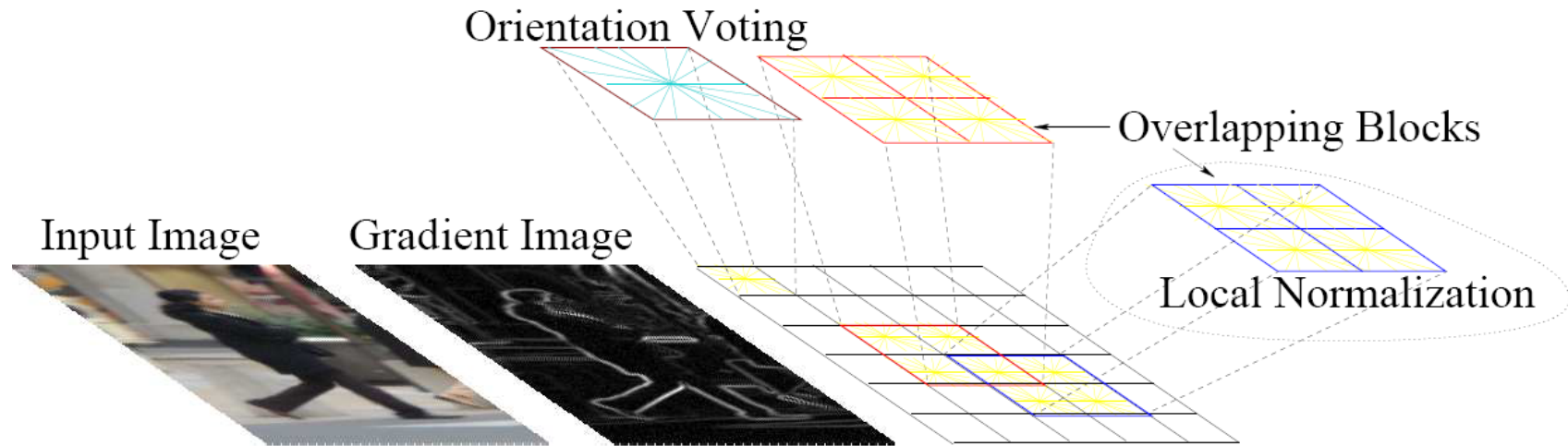
Dalal & Triggs CVPR 2005 Pedestrian detection

- Objective: detect (localize) standing humans in an image
- sliding window classifier
- train a binary classifier on whether a window contains a standing person or not
- Histogram of Oriented Gradients (HOG) feature
- although HOG + SVM originally introduced for pedestrians has been used very successfully for many object categories

Feature: Histogram of Oriented Gradients (HOG)

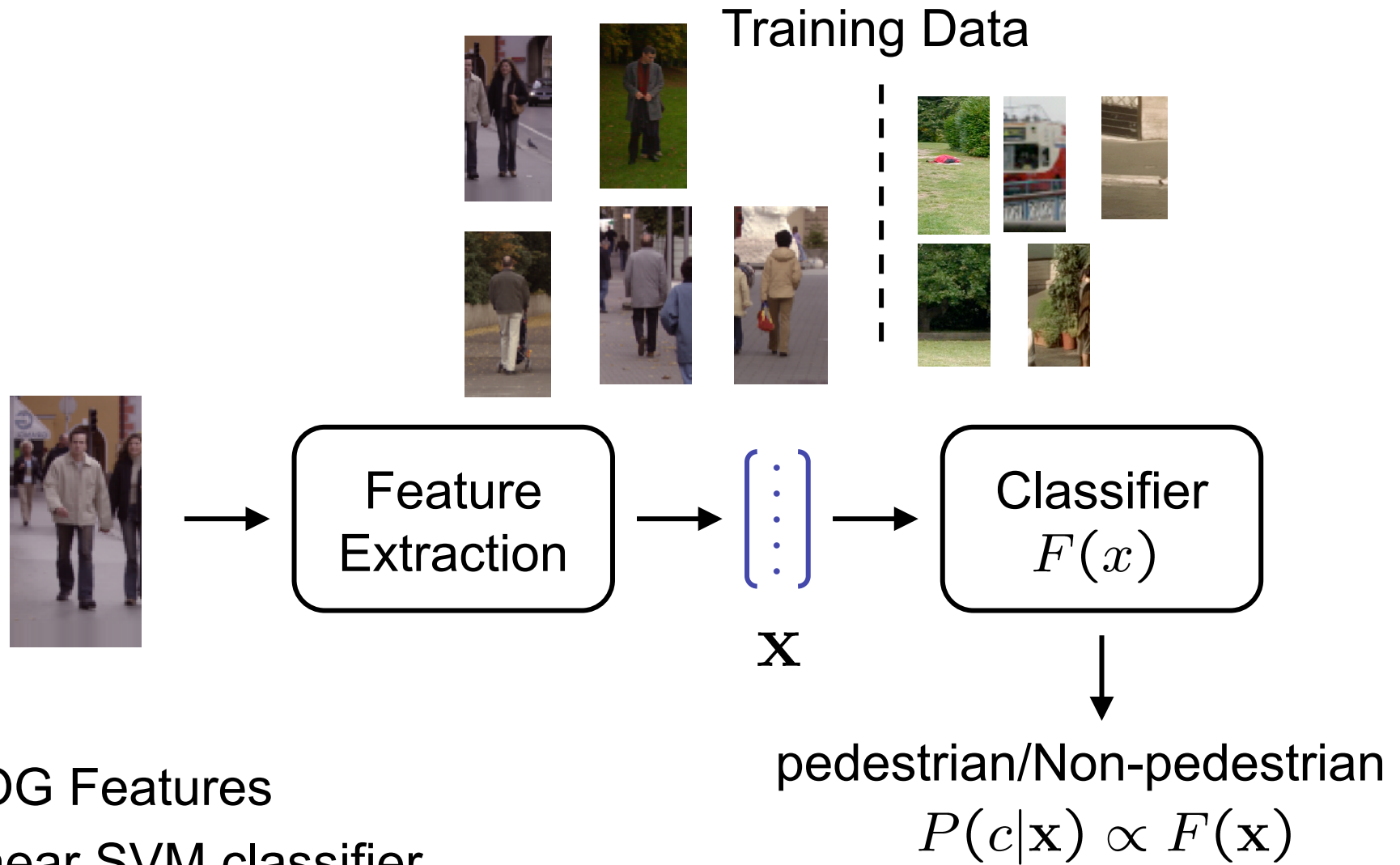


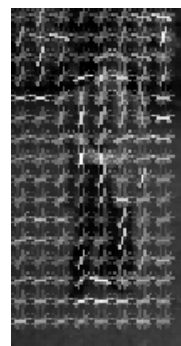
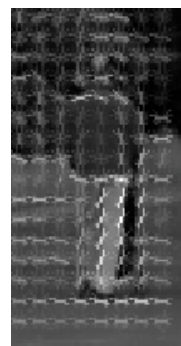
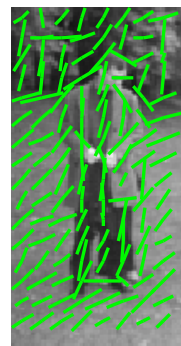
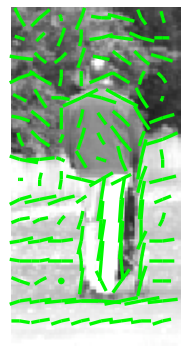
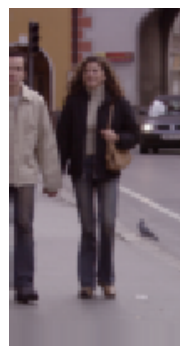
Histogram of Oriented Gradients (HOG) continued



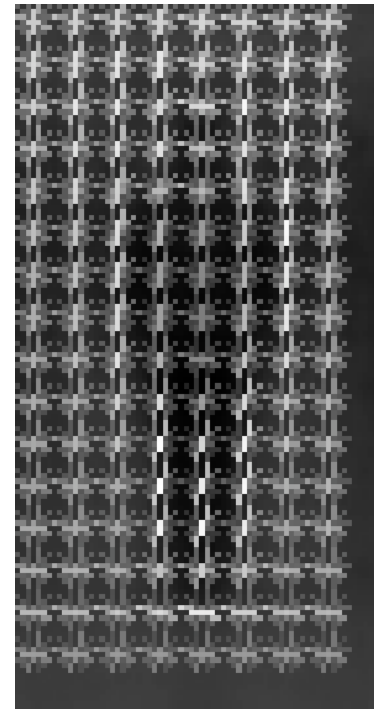
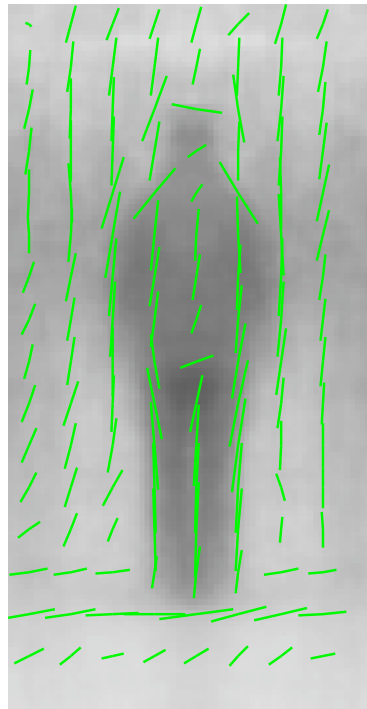
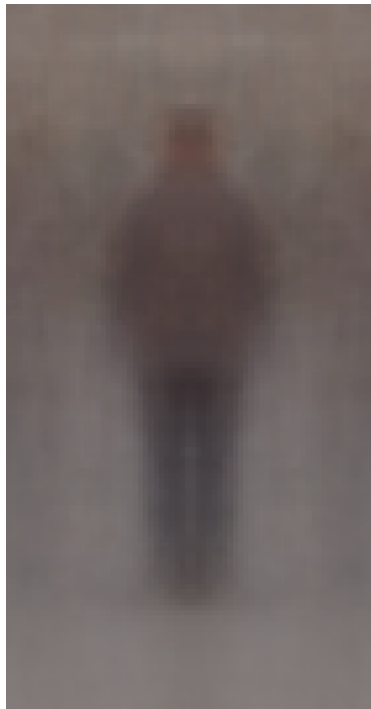
- Adds a second level of overlapping spatial bins re-normalizing orientation histograms over a larger spatial area
- Feature vector dimension (approx) = 16×8 (for tiling) $\times 8$ (orientations) $\times 4$ (for blocks) = 4096

Window (Image) Classification





Averaged examples



Classifier: linear support vector machine (linear SVM)

Advantages of linear SVM: $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$

- Training (Learning)
 - Very efficient packages for the linear case, e.g. LIBLINEAR for batch training and Pegasos for on-line training.
 - Complexity $O(N)$ for N training points (cf $O(N^3)$ for general SVM)
- Testing (Detection)

Non-linear $f(\mathbf{x}) = \sum_i^S \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b$

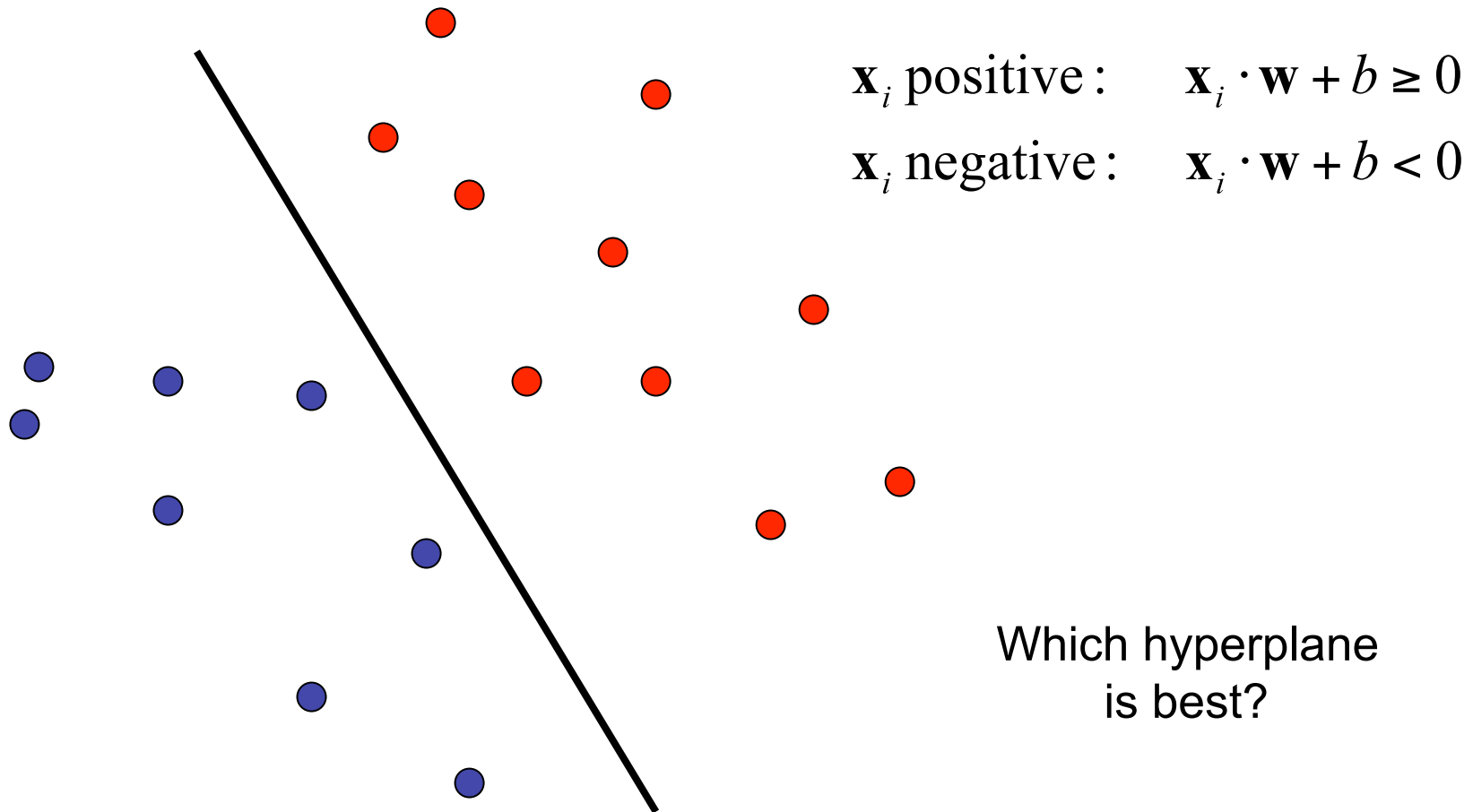
S = # of support vectors
= (worst case) N
size of training data

linear $f(\mathbf{x}) = \sum_i^S \alpha_i \mathbf{x}_i^T \mathbf{x} + b$
 $= \mathbf{w}^T \mathbf{x} + b$

Independent of size of training data

Linear classifiers

- Find linear function (*hyperplane*) to separate positive and negative examples





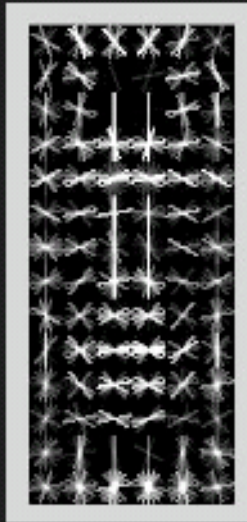
Dalal and Triggs, CVPR 2005

Learned model

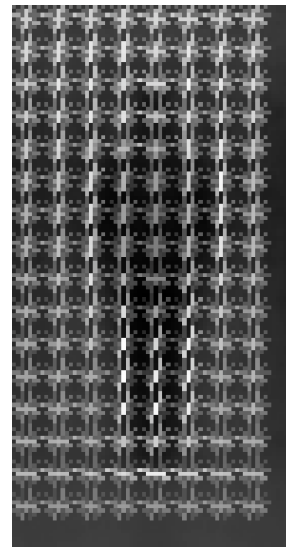
$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



positive
weights



negative
weights



average over
positive training data

What do negative weights mean?

$$wx > 0$$

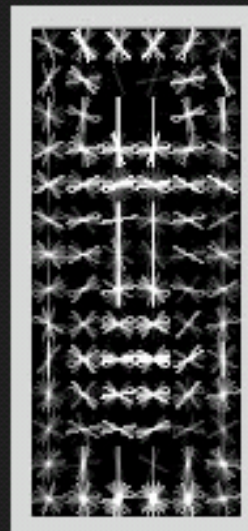
$$(w_+ - w_-)x > 0$$

$$w_+ > w_-$$

pedestrian
model



>



pedestrian
background
model

Complete system should compete pedestrian/pillar/doorway models

Discriminative models come equipped with own bg
(avoid firing on doorways by penalizing vertical edges)

Why does HOG + SVM work so well?

- Similar to SIFT, records spatial arrangement of **histogram** orientations
- Compare to learning only edges:
 - Complex junctions can be represented
 - Avoids problem of early thresholding
 - Represents also soft internal gradients
- Older methods based on edges have become largely obsolete



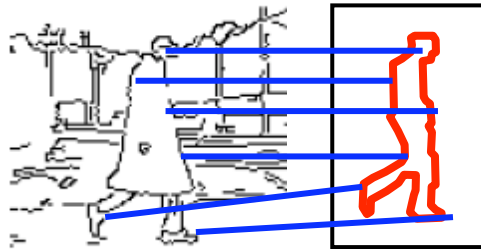
- HOG gives fixed length vector for window, suitable for feature vector for SVM

Chamfer Matching

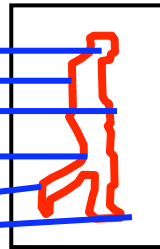
Input



Edges



Template



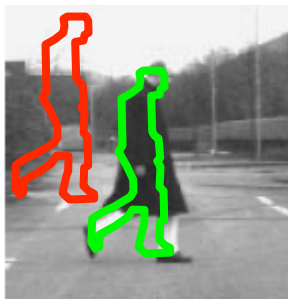
- Match points between template and image
- Measure mean distance
- Template edgel matches **nearest** image edgel

$$D(T, I) = \frac{1}{|T|} \sum_{p \in T} \min_{q \in I} d(p, q)$$

Distance Transform



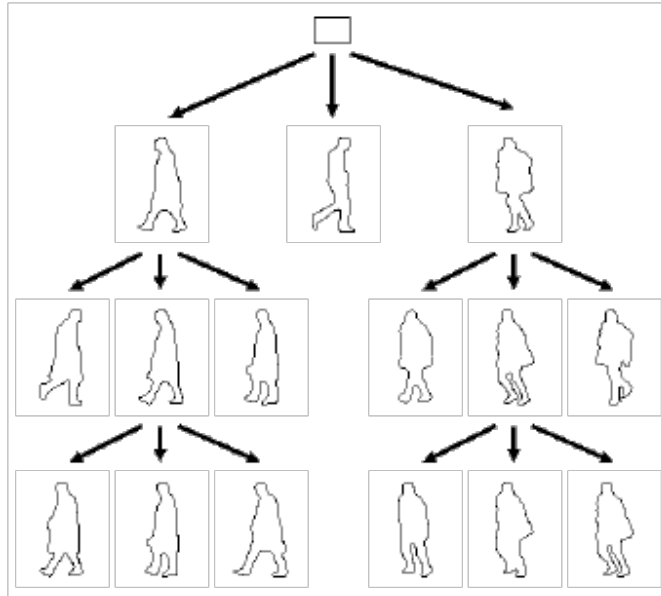
Best match



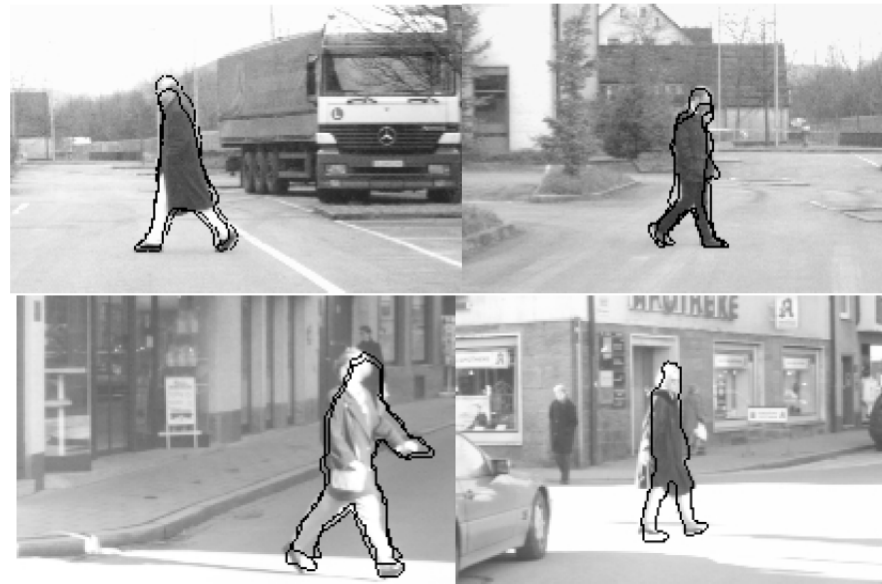
- Distance transform reduces min operation to array lookup
- Computable in linear time
- Localize by sliding window search

[Gavrila & Philomin, 1999]

Chamfer Matching



Hierarchy of Templates



Detections

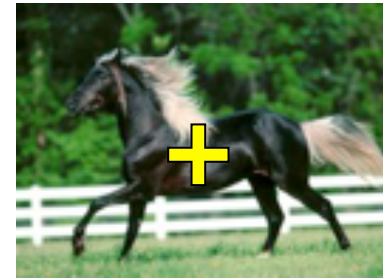
- In practice performs poorly in clutter
- Unoriented edges are not discriminative enough (too easy to find...)

[Gavrila & Philomin, 1999]

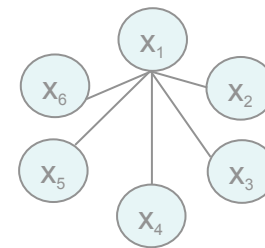
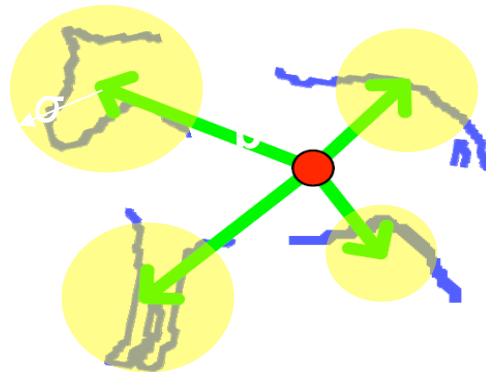
Contour-fragment models

Shotton et al ICCV 05, Opelt et al ECCV 06

- Generalized Hough like representation using contour fragments
- Contour fragments learnt from edges of training images

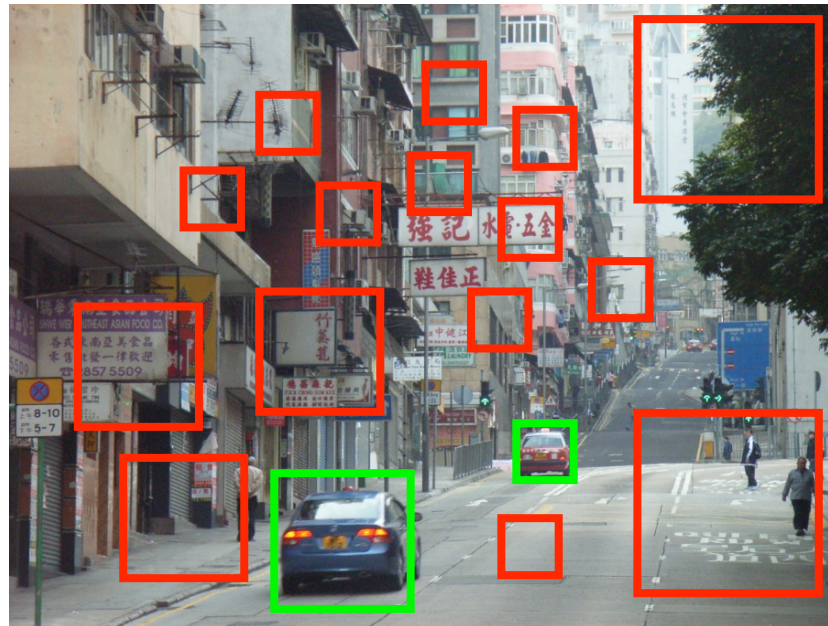


- Hough like voting for detection



Training a sliding window detector

- Object **detection** is inherently asymmetric: much more “non-object” than “object” data



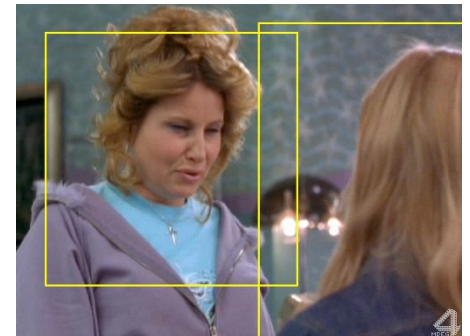
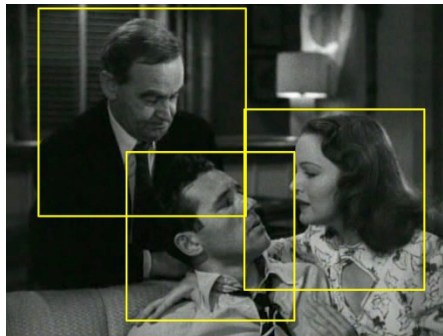
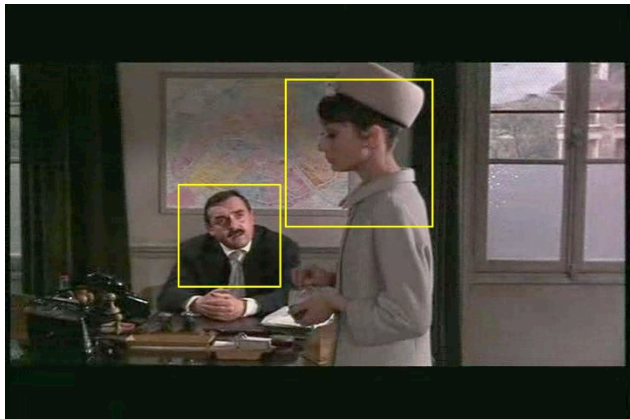
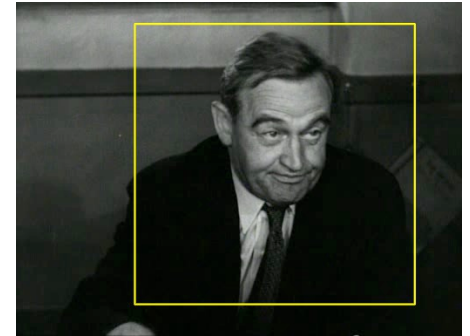
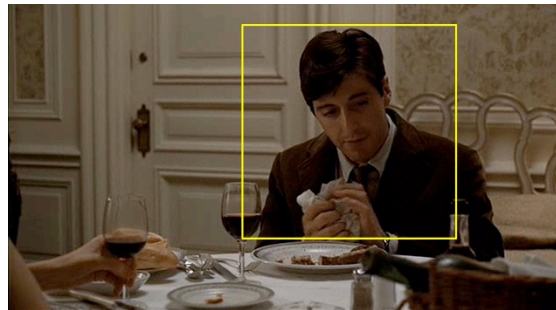
- Classifier needs to have very low false positive rate
- Non-object category is very complex – need lots of data

Example: train an upper body detector

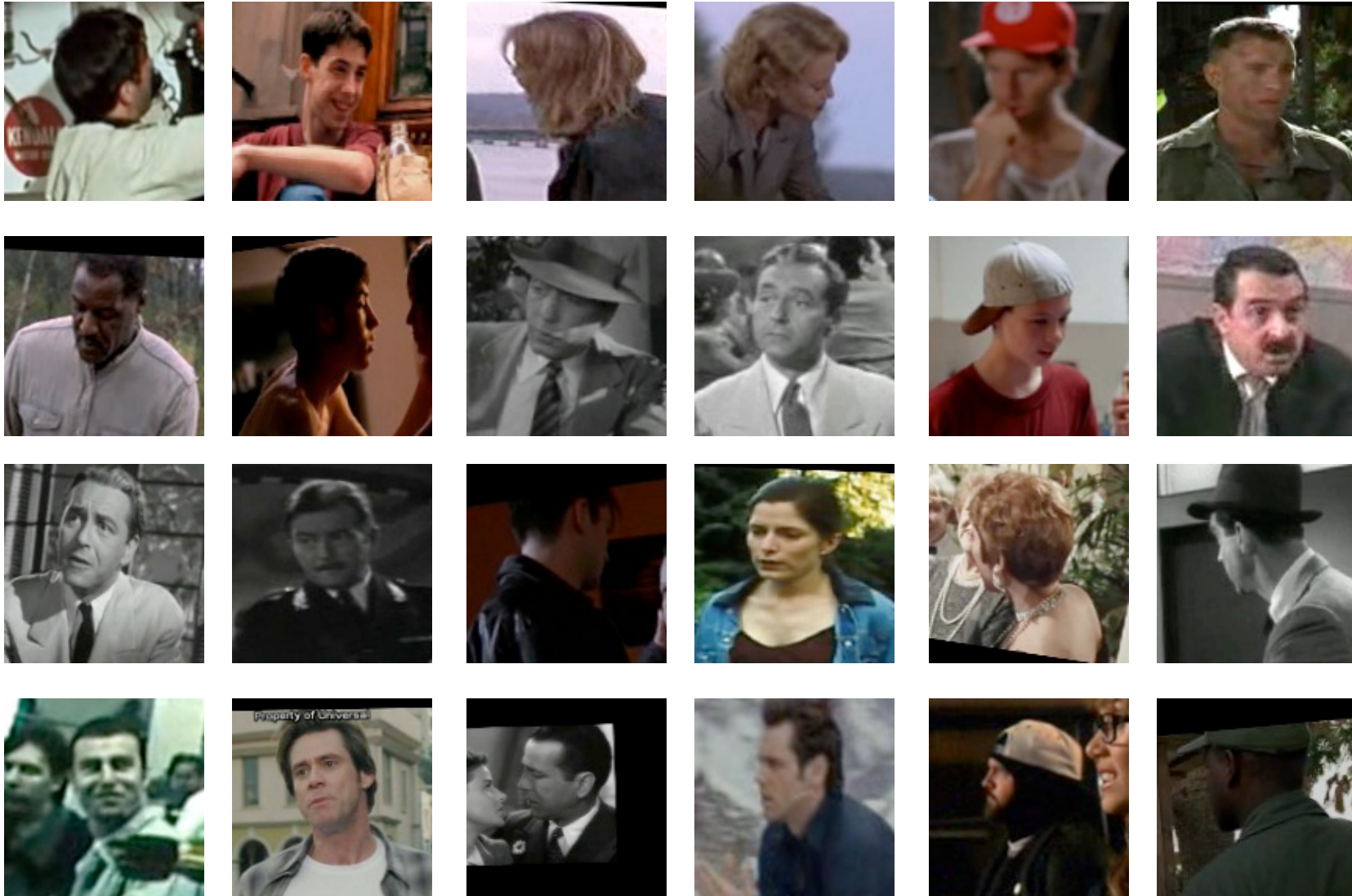
- Training data – used for training and validation sets
 - 33 Hollywood2 training movies
 - 1122 frames with upper bodies marked
- First stage training (bootstrapping)
 - 1607 upper body annotations jittered to 32k positive samples
 - 55k negatives sampled from the same set of frames
- Second stage training (retraining)
 - 150k hard negatives found in the training data



Training data – positive annotations

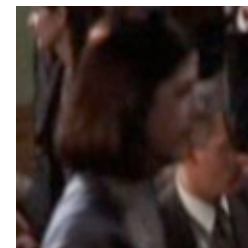
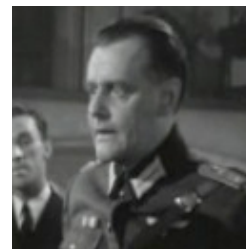
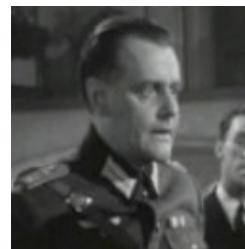
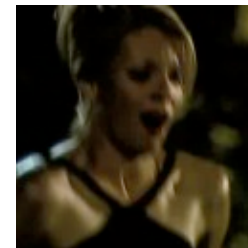
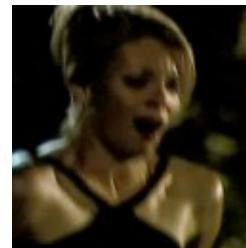
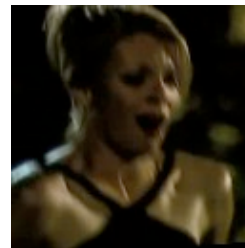
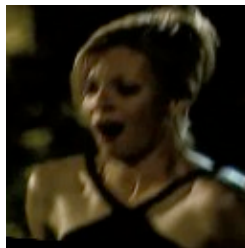
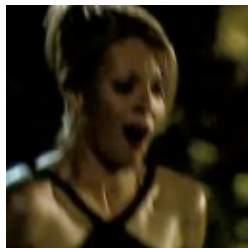
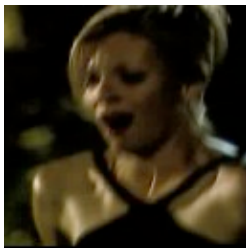


Positive windows

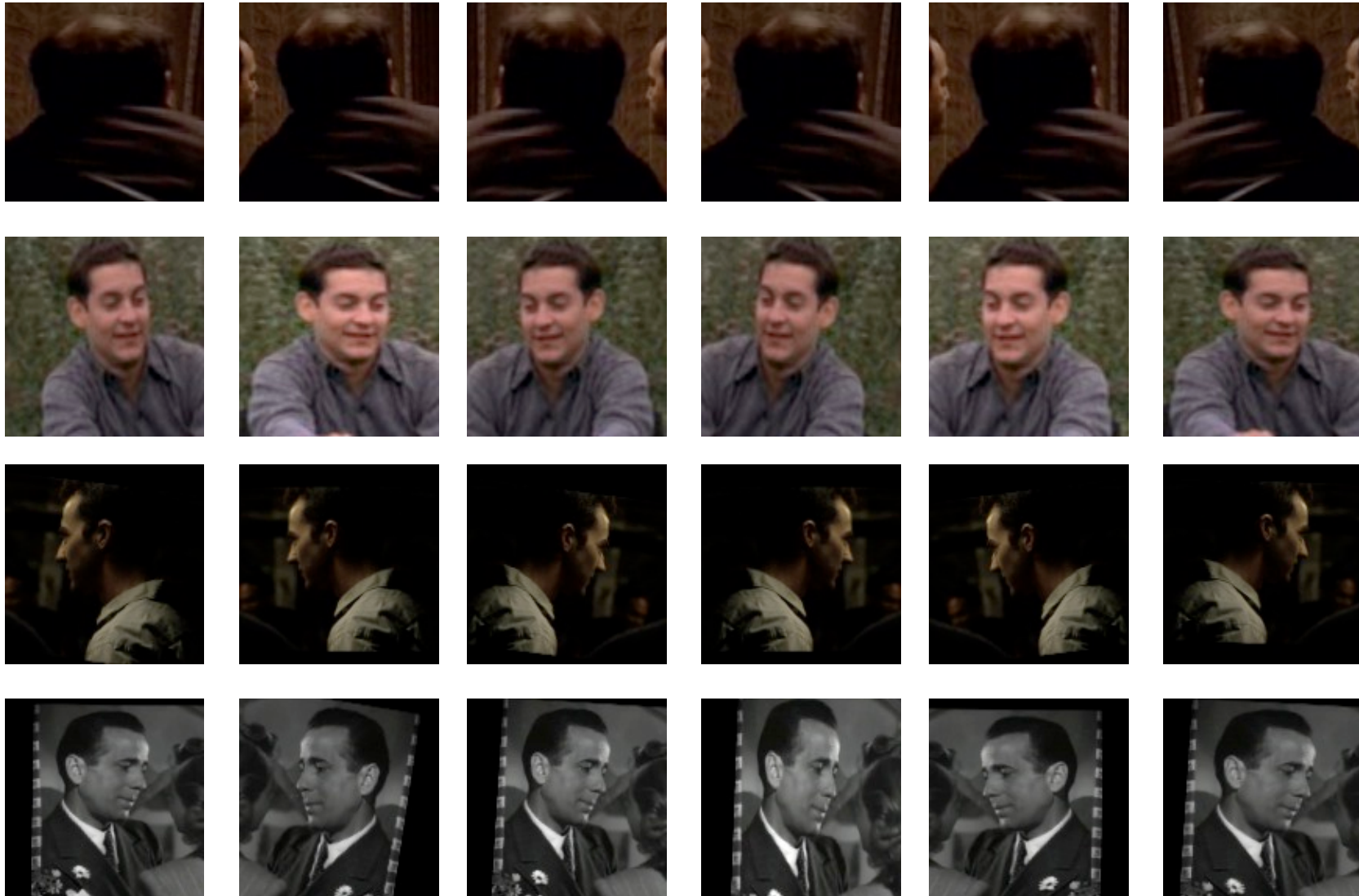


Note: common size and alignment

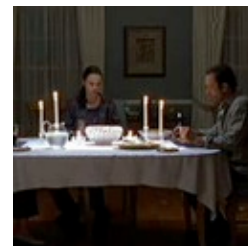
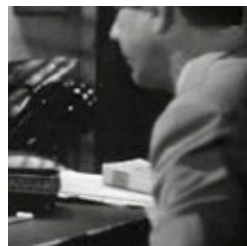
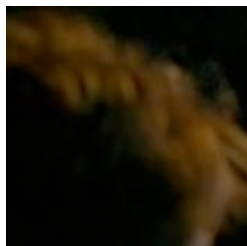
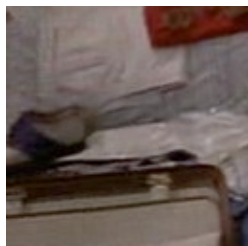
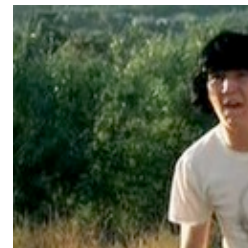
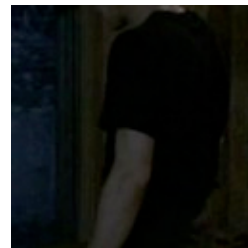
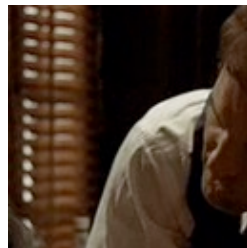
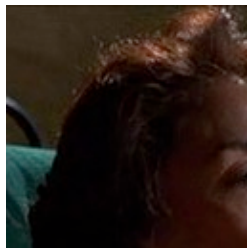
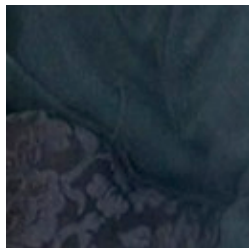
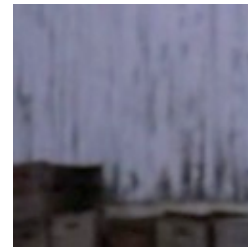
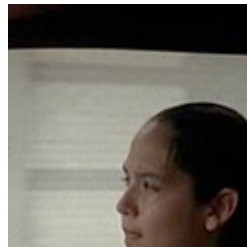
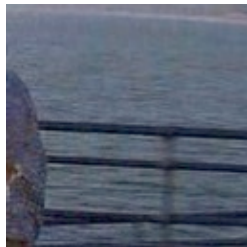
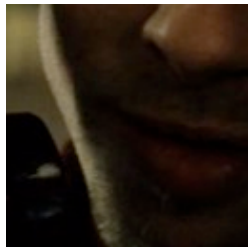
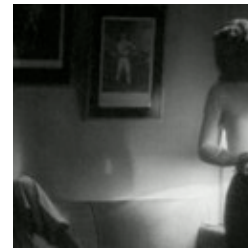
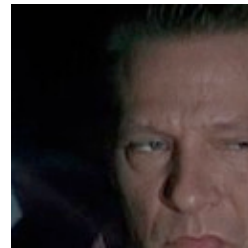
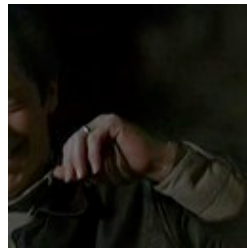
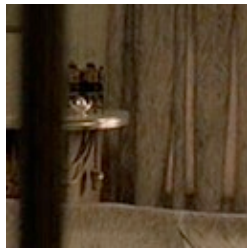
Jittered positives



Jittered positives



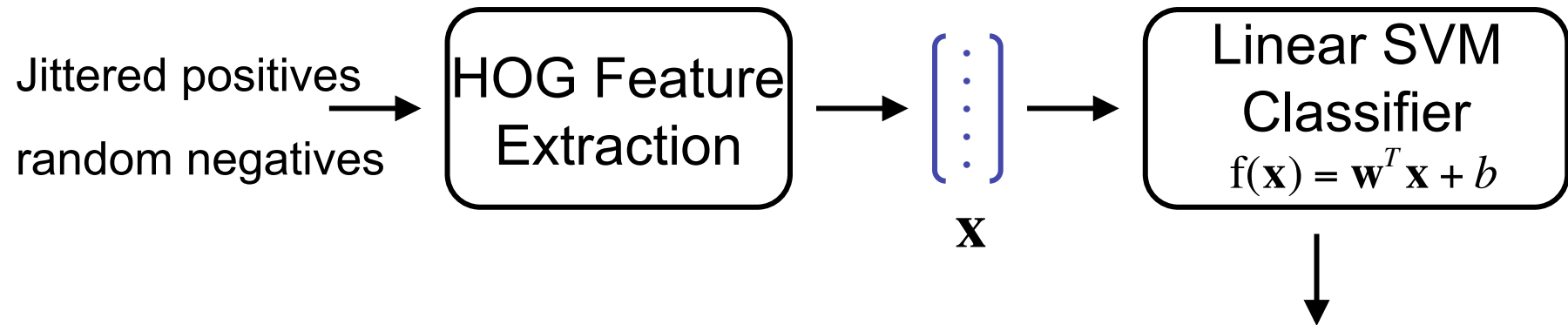
Random negatives



Random negatives



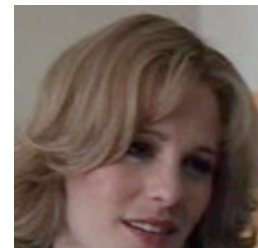
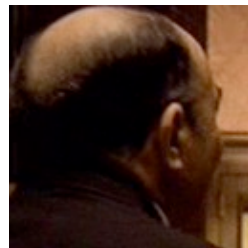
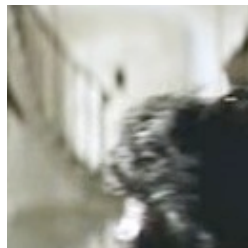
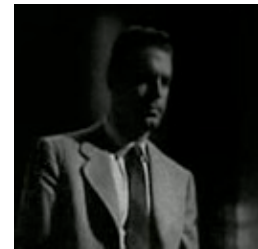
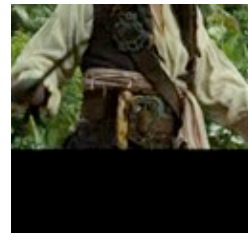
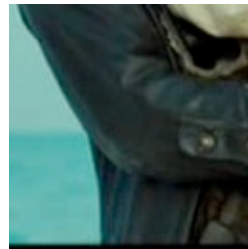
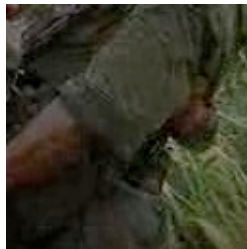
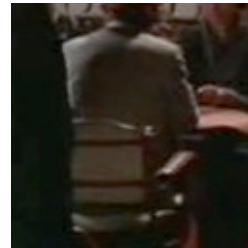
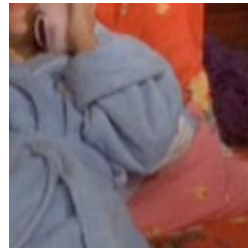
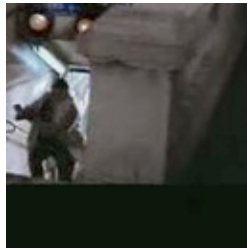
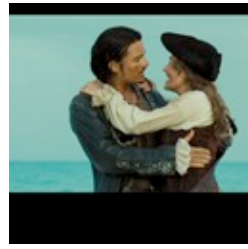
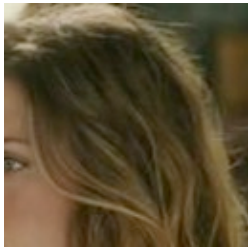
Window (Image) first stage classification



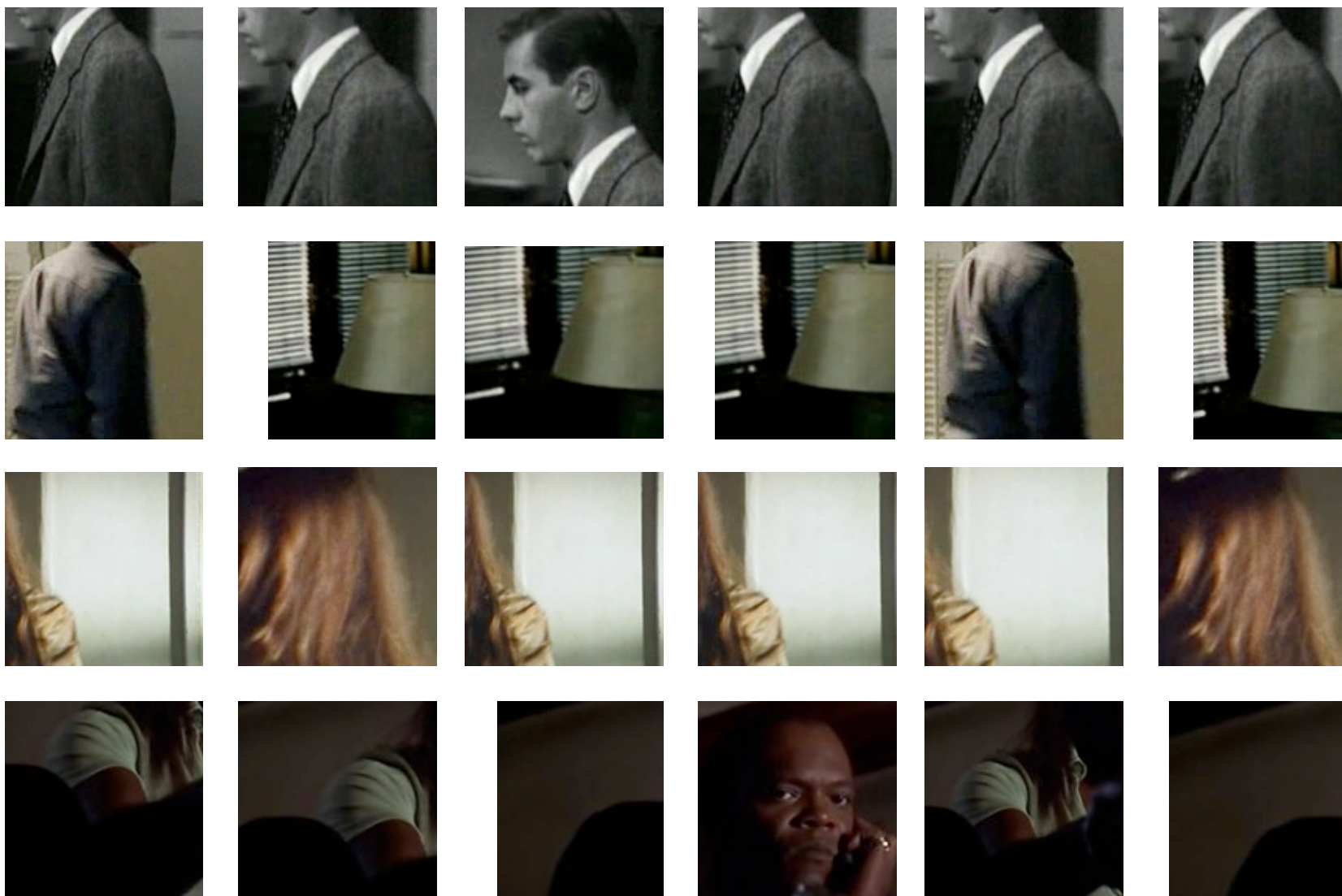
- find high scoring false positives detections
- these are the hard negatives for the next round of training
- cost = # training images x inference on each image



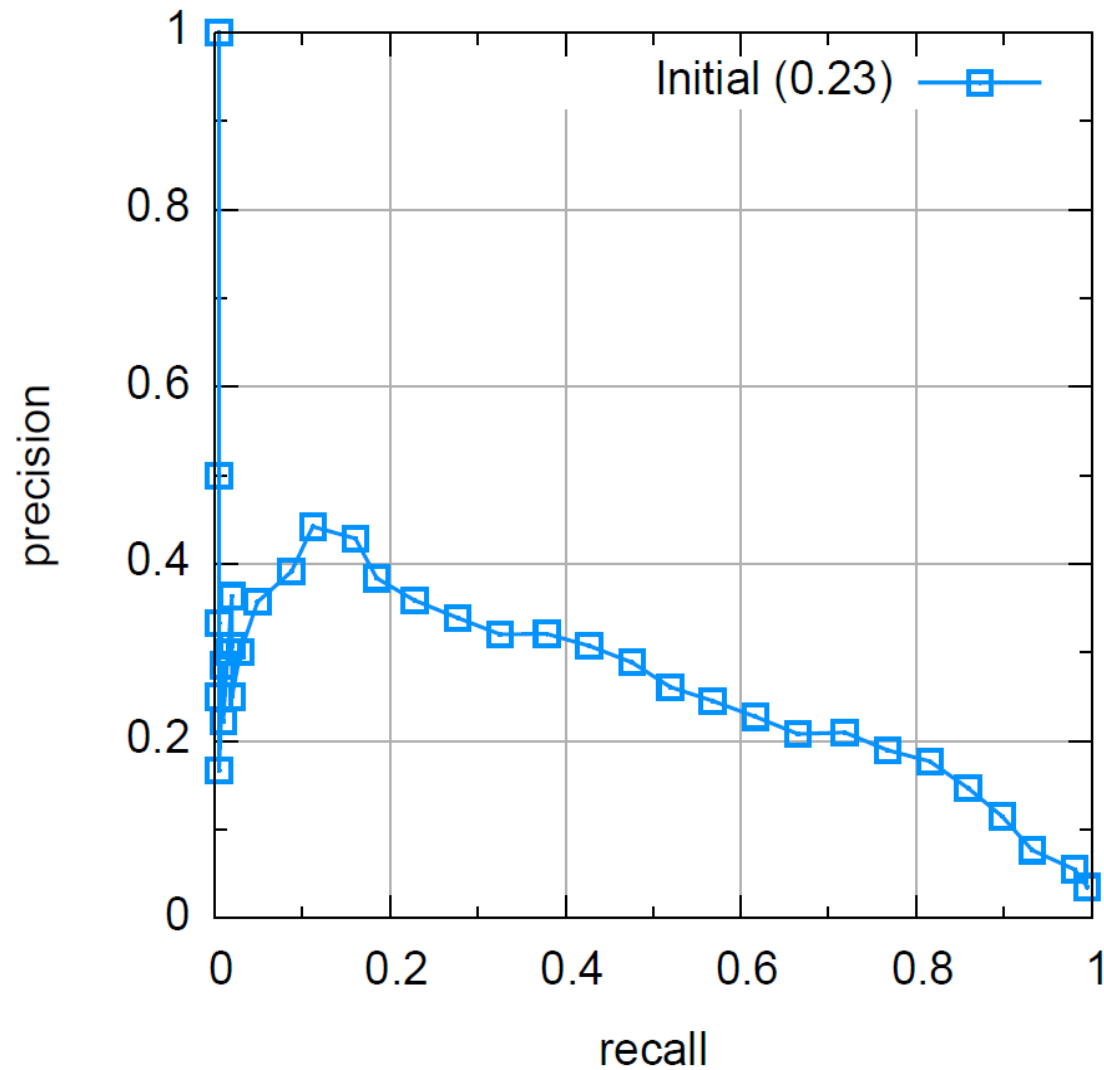
Hard negatives



Hard negatives

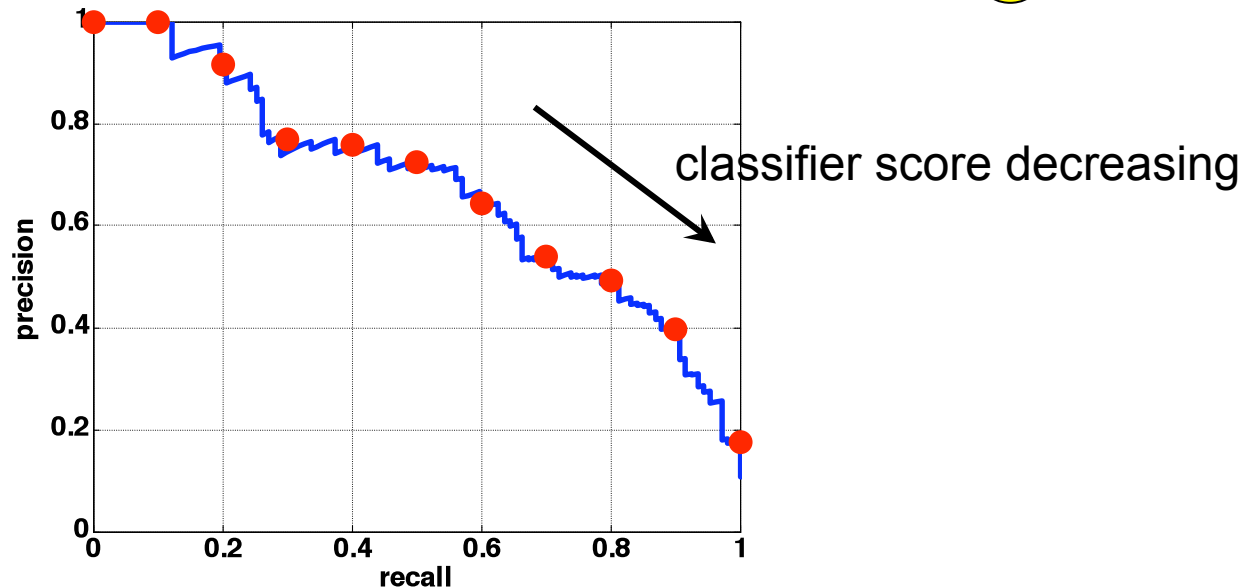
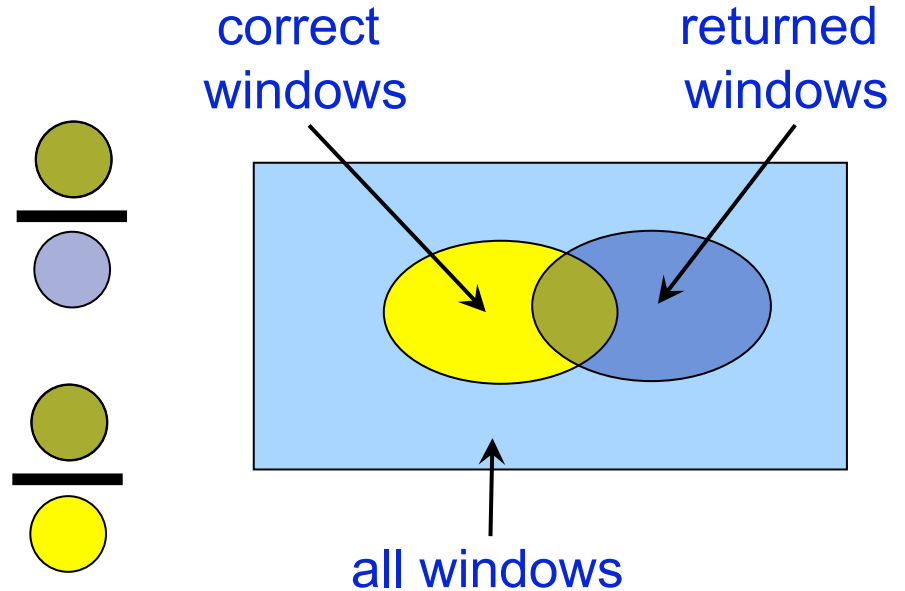


First stage performance on validation set

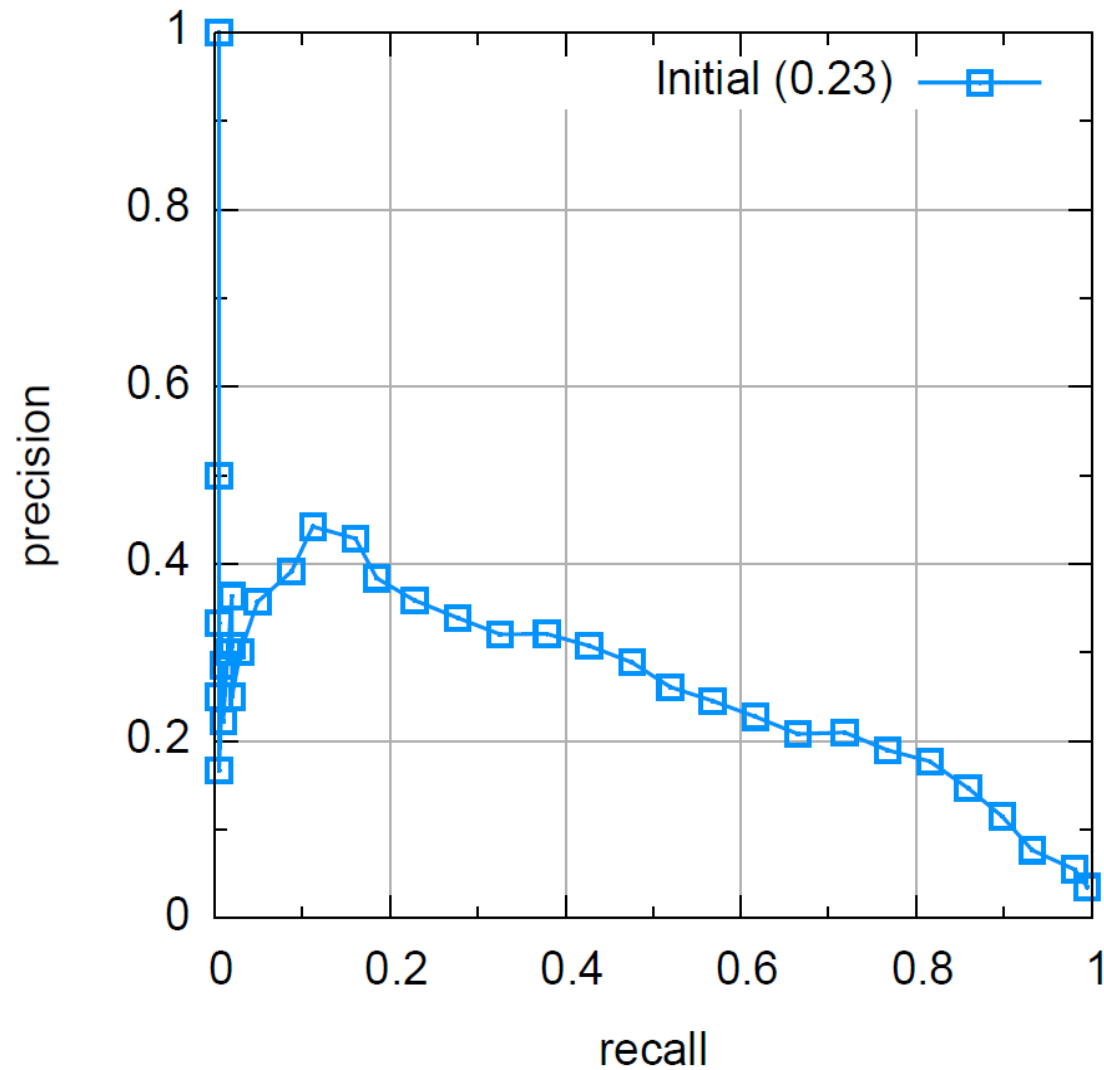


Precision – Recall curve

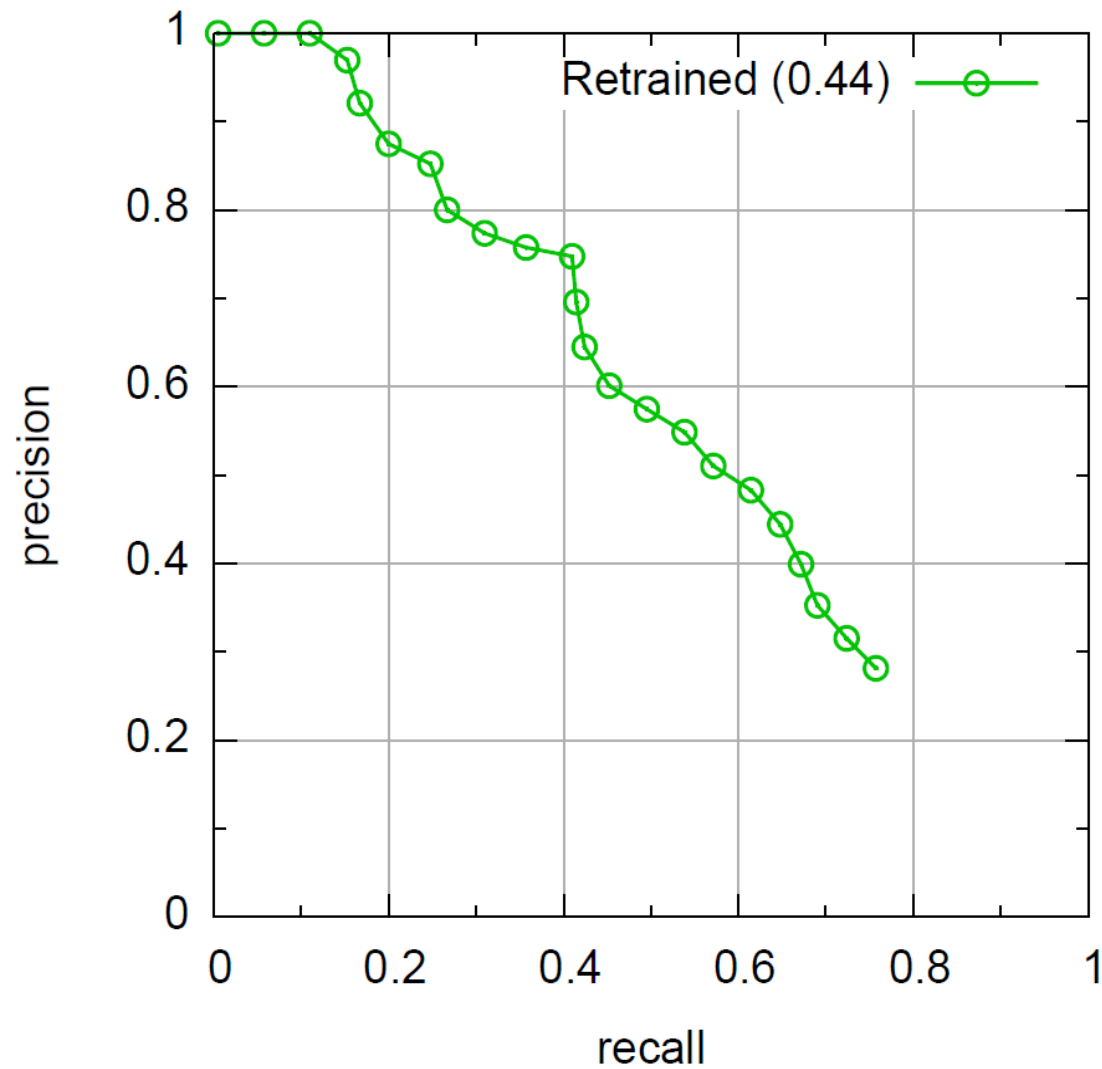
- **Precision:** % of returned windows that are correct
- **Recall:** % of correct windows that are returned



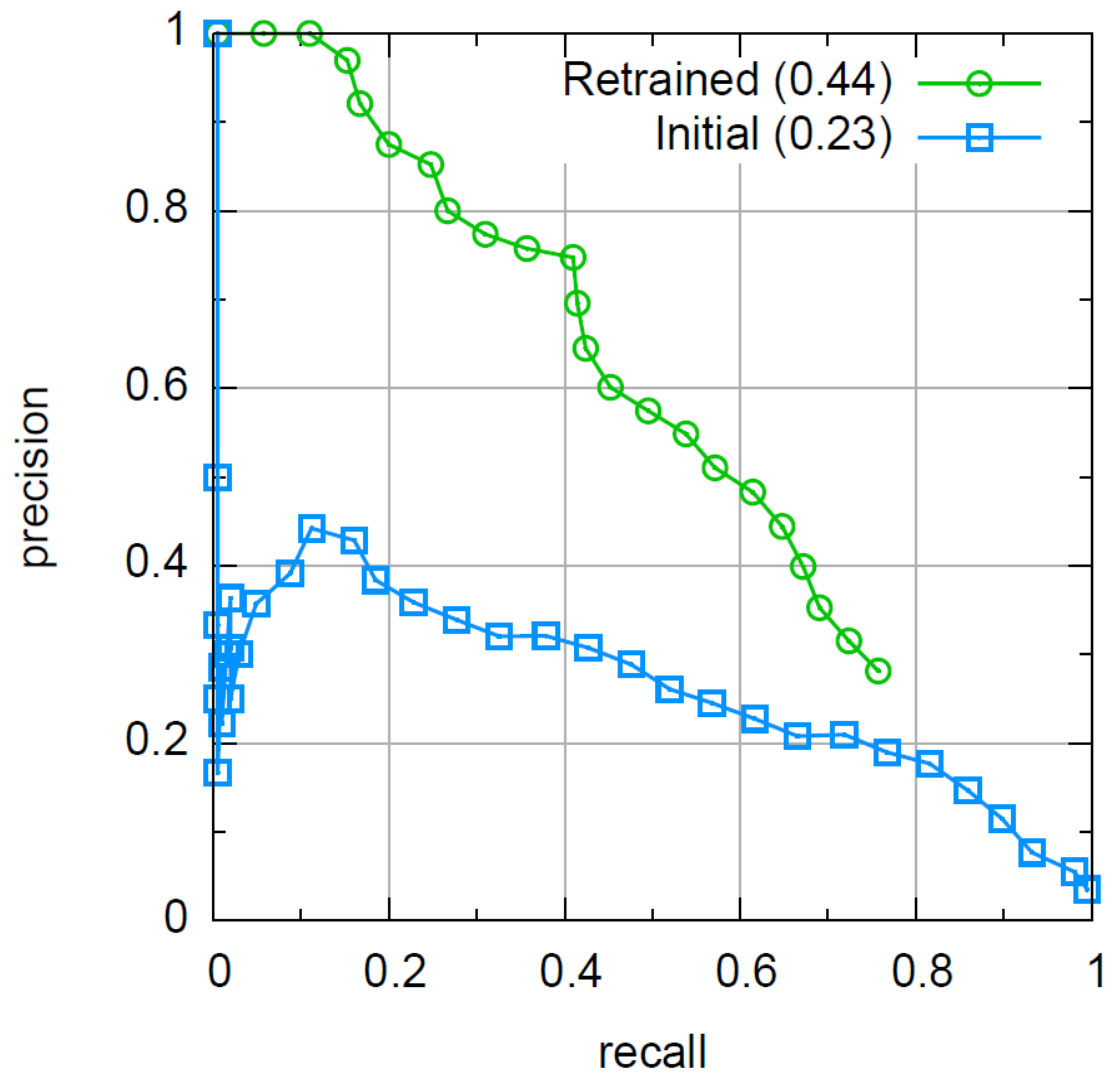
First stage performance on validation set



Performance after retraining

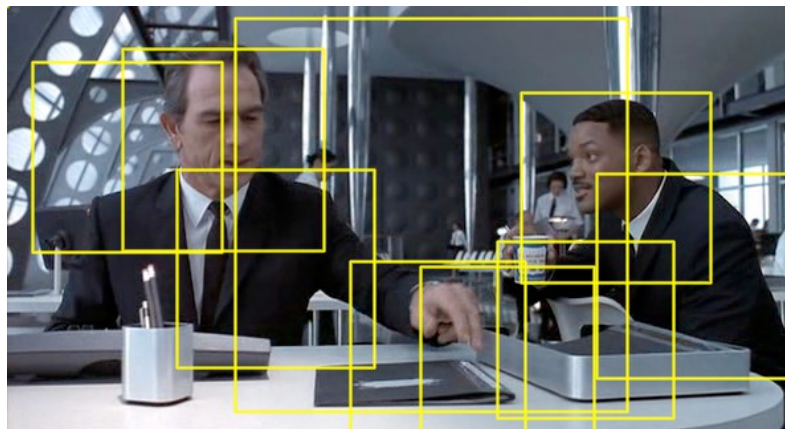


Effects of retraining

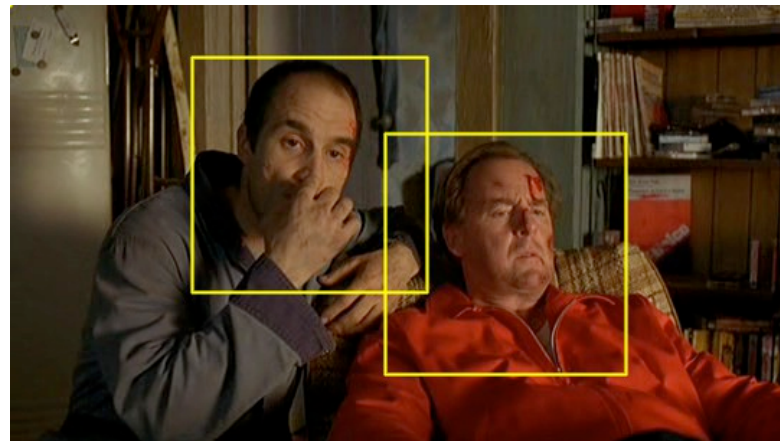
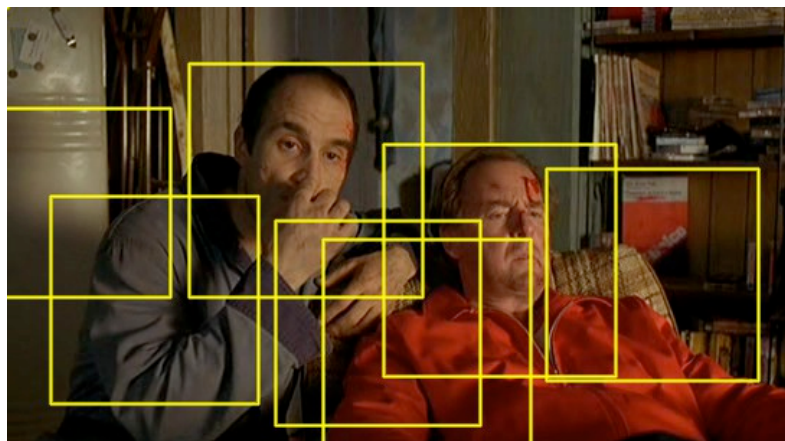
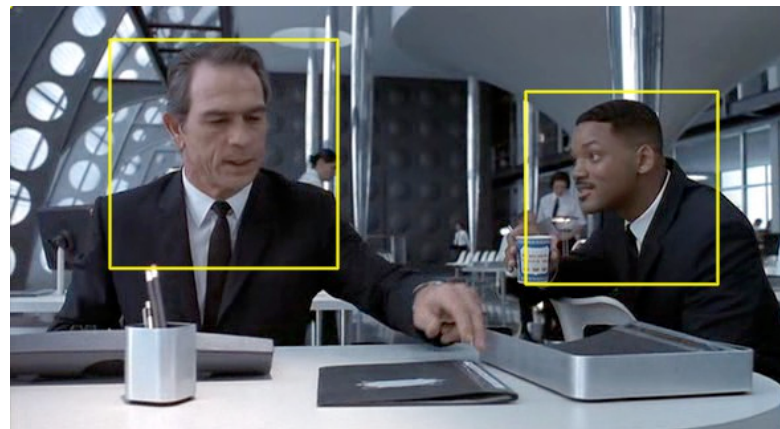


Side by side

before retraining

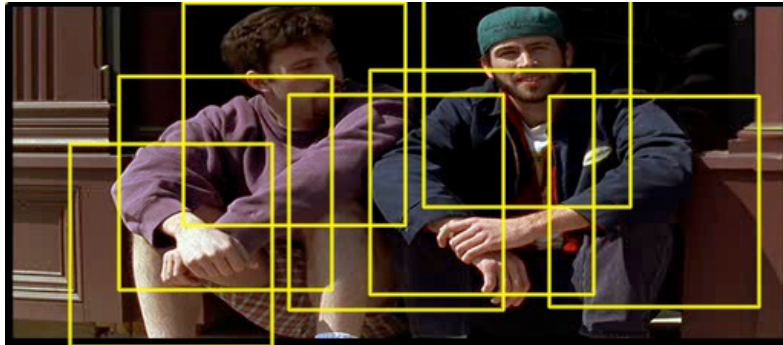


after retraining

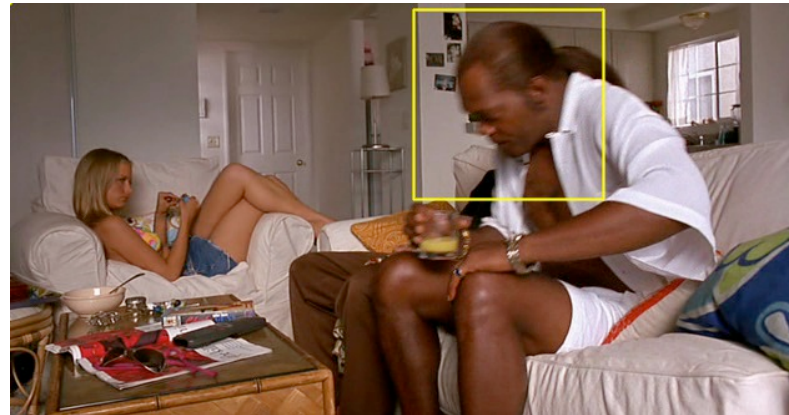
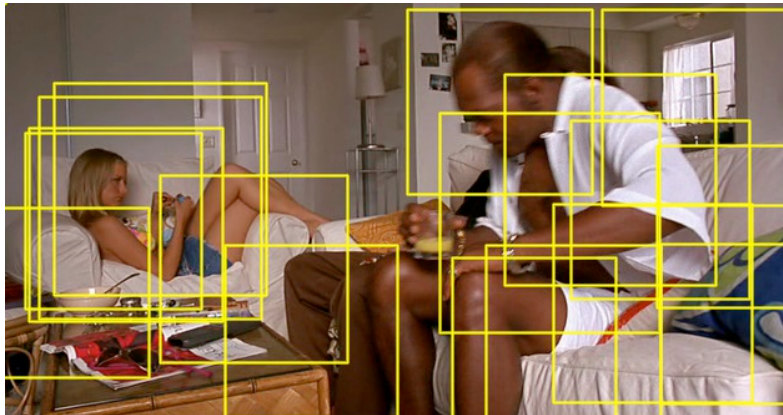


Side by side

before retraining

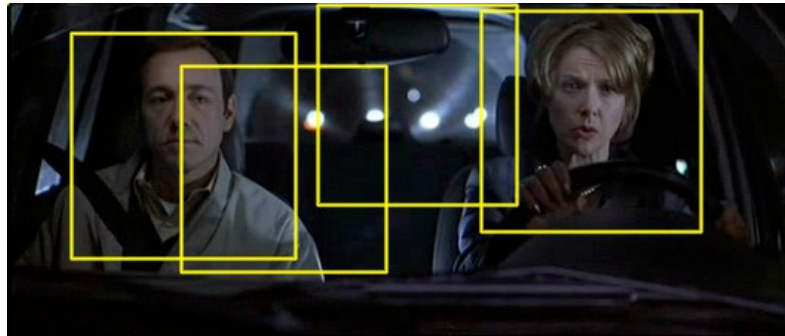


after retraining

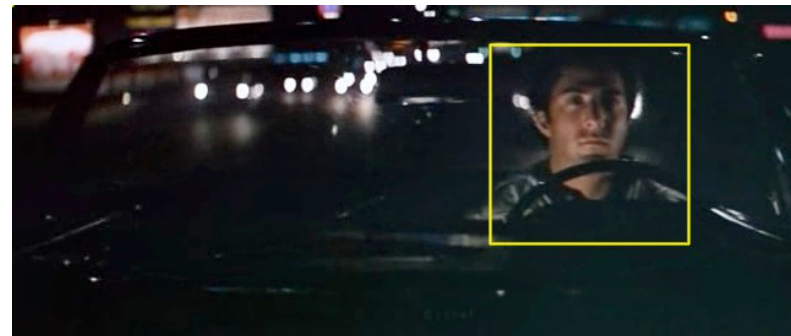
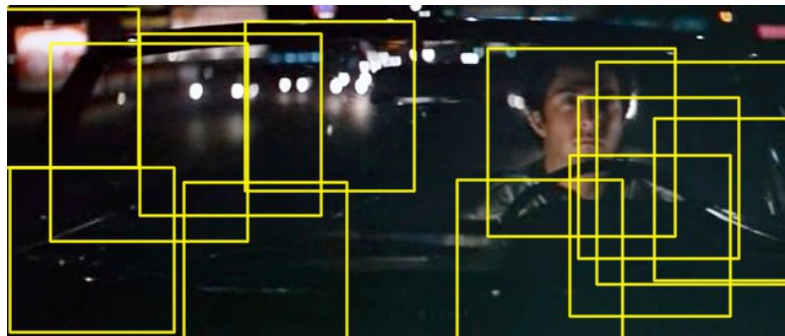


Side by side

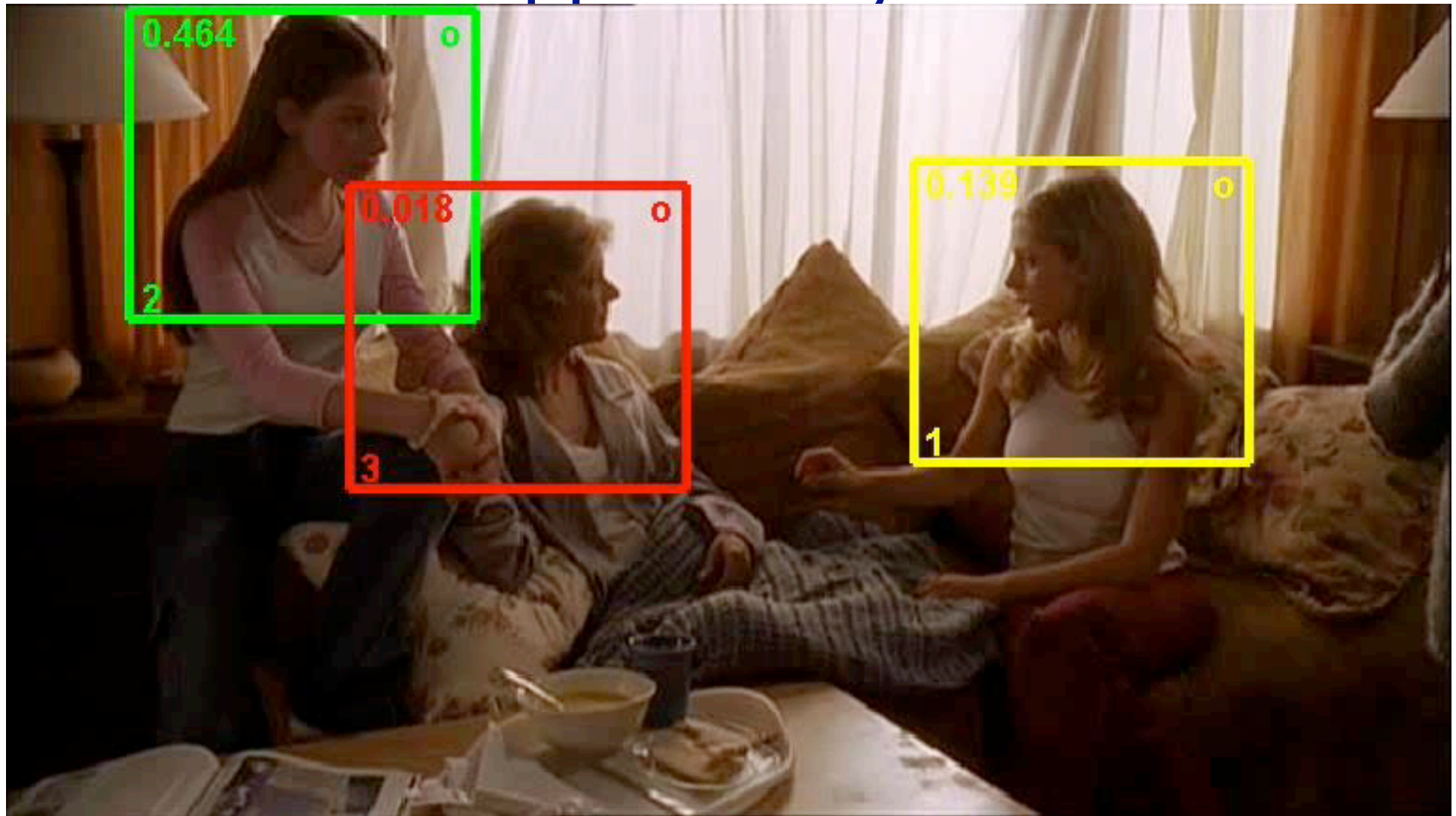
before retraining



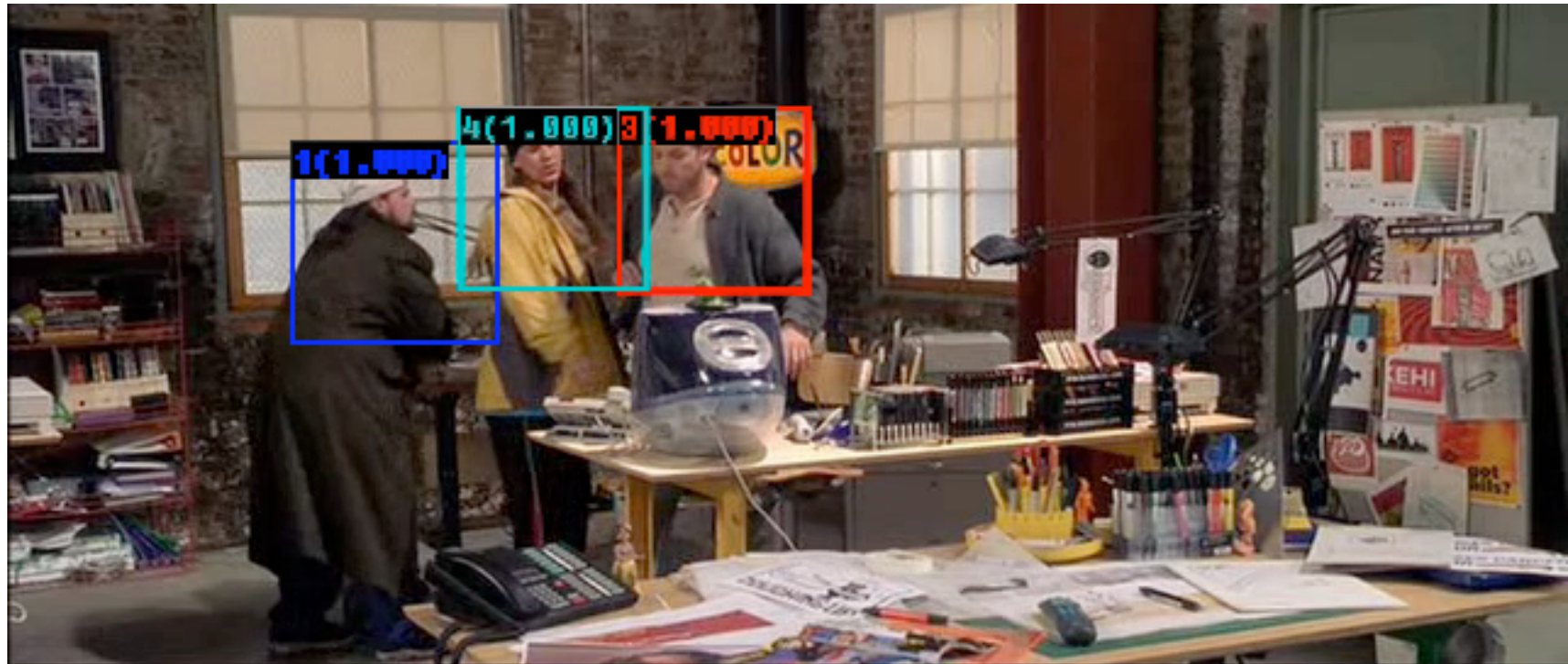
after retraining



Tracked upper body detections



Tracked upper body person detections



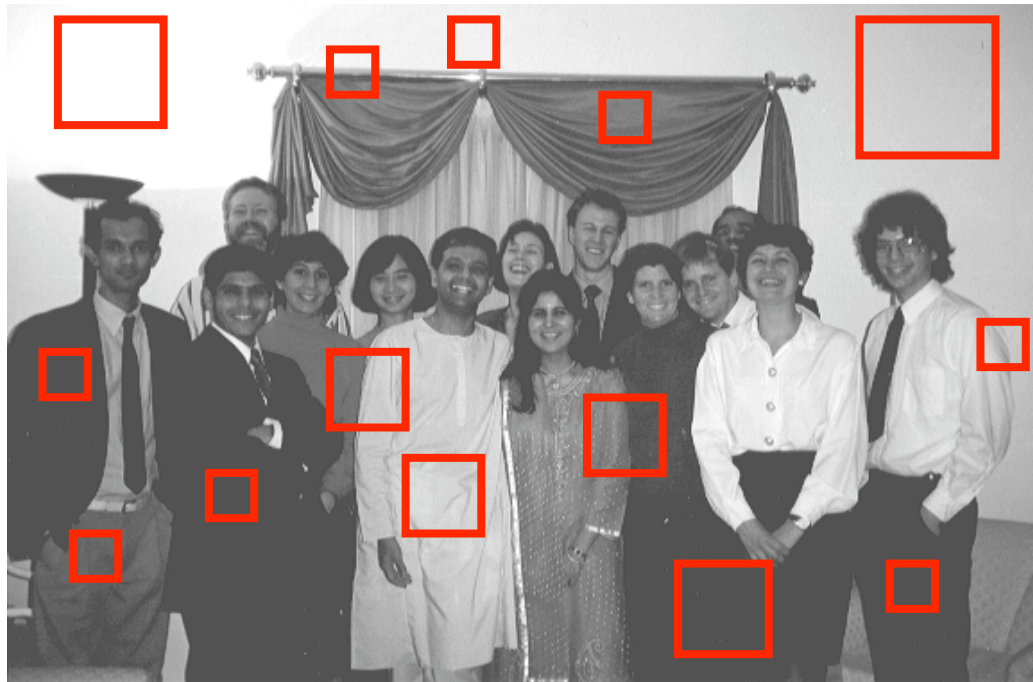
Combined face, upper body and full body detectors “vote” for upper body bounding boxes.

Detections are tracked and smoothed over video.

[Lezama, MVA thesis 2010]

Accelerating Sliding Window Search

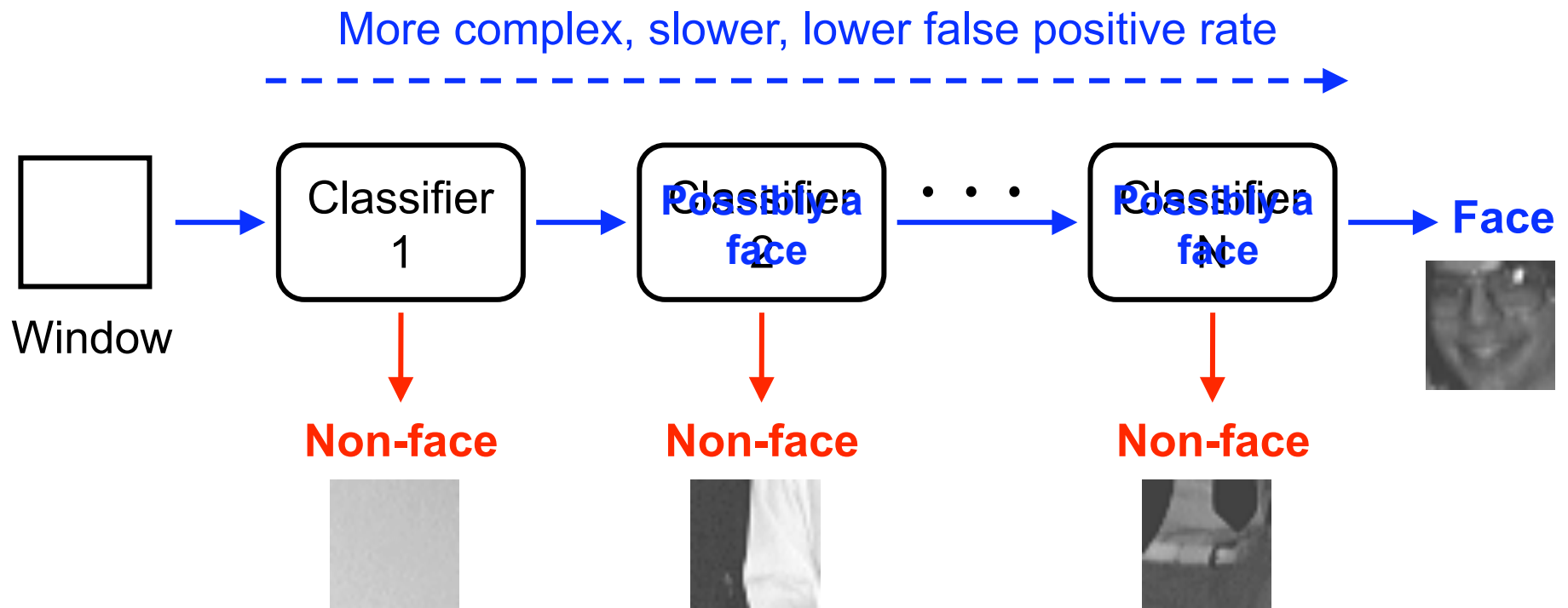
- Sliding window search is slow because so many windows are needed e.g. $x \times y \times \text{scale} \approx 100,000$ for a 320×240 image



- Most windows are clearly not the object class of interest
- Can we speed up the search?

Cascaded Classification

- Build a sequence of classifiers with increasing complexity



- Reject easy non-objects using simpler and faster classifiers

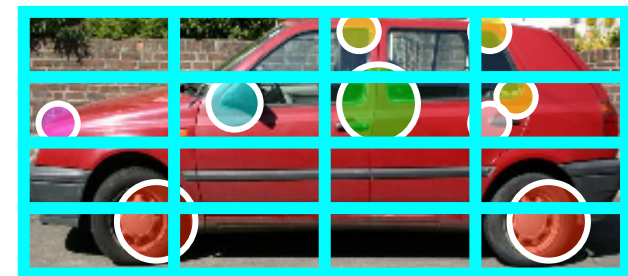
Cascaded Classification



- Slow expensive classifiers only applied to a few windows → significant speed-up
- Controlling classifier complexity/speed:
 - Number of support vectors [Romdhani et al, 2001]
 - Number of features [Viola & Jones, 2001]
 - Type of SVM kernel [Vedaldi et al, 2009]

Summary: Sliding Window Detection

- Can convert any image classifier into an object detector by sliding window. Efficient search methods available.
- Requirements for invariance are reduced by searching over e.g. translation and scale
- Spatial correspondence can be “engineered in” by spatial tiling



Outline

1. Sliding window detectors
2. Features and adding spatial information
3. HOG + linear SVM classifier
4. Two state of the art algorithms and PASCAL VOC
 - VOC challenge
 - Vedaldi et al – multiple kernels and features, cascade
 - Felzenswalb et al – multiple parts, latent SVM
5. The future and challenges

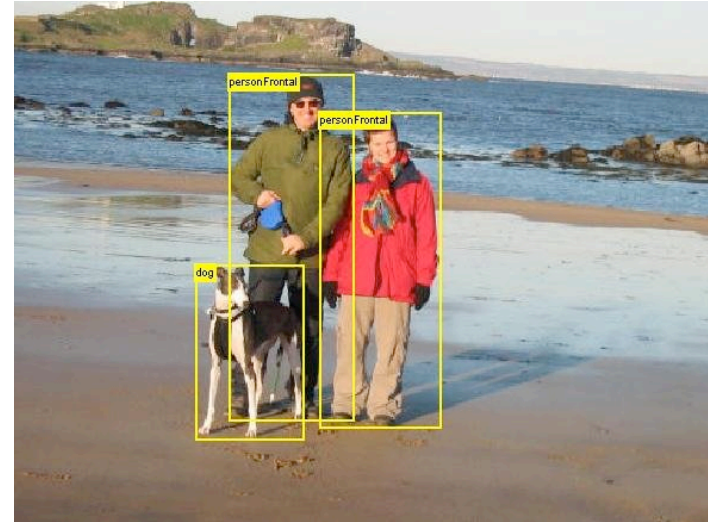
The PASCAL Visual Object Classes (VOC) Dataset and Challenge

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman



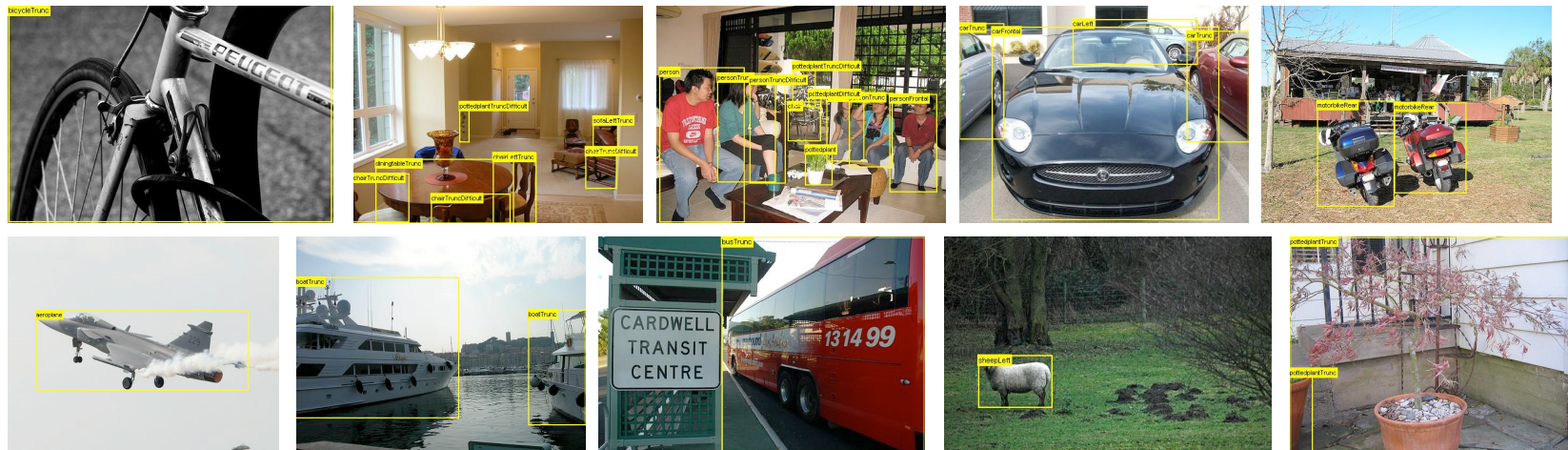
The PASCAL VOC Challenge

- Challenge in visual object recognition funded by PASCAL network of excellence
- Publicly available dataset of annotated images
- Main competitions in classification (is there an X in this image), detection (where are the X's), and segmentation (which pixels belong to X)
- “Taster competitions” in 2-D human “pose estimation” (2007-present) and static action classes
- Standard evaluation protocol (software supplied)



Dataset Content

- 20 classes: aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV
- Real images downloaded from flickr, not filtered for “quality”



- Complex scenes, scale, pose, lighting, occlusion, ...

Annotation

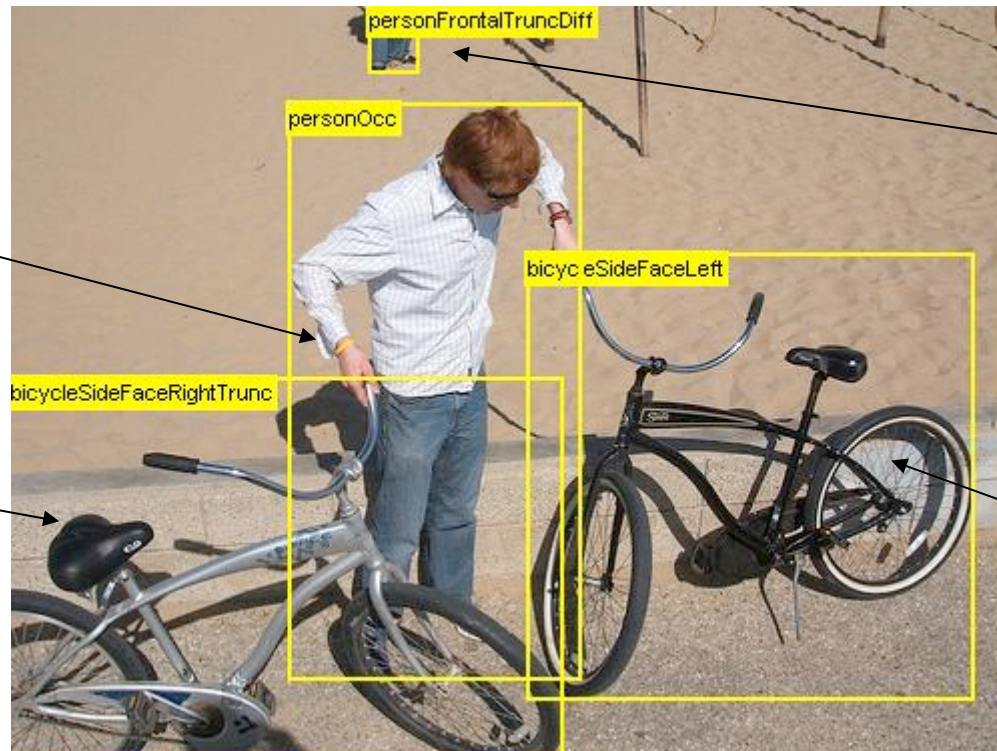
- Complete annotation of all objects
- Annotated in one session with written guidelines

Occluded

Object is significantly occluded within BB

Truncated

Object extends beyond BB



Difficult

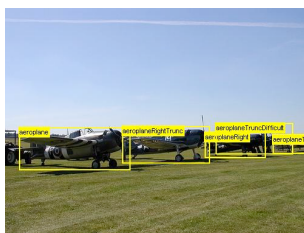
Not scored in evaluation

Pose

Facing left

Examples

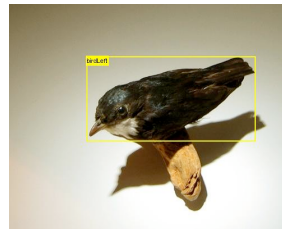
Aeroplane



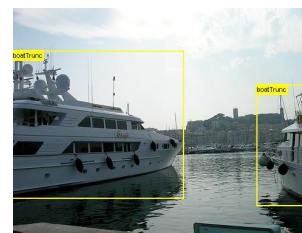
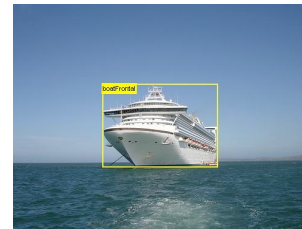
Bicycle



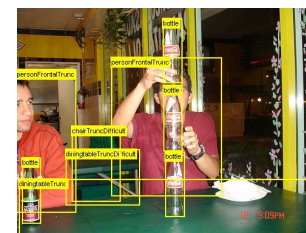
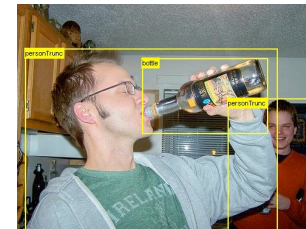
Bird



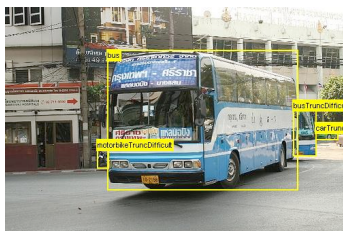
Boat



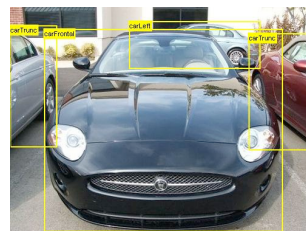
Bottle



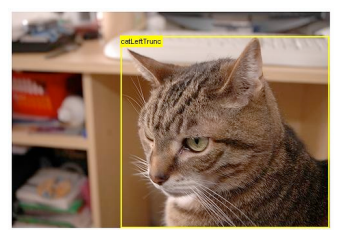
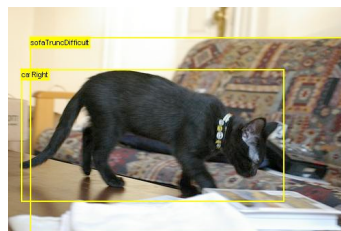
Bus



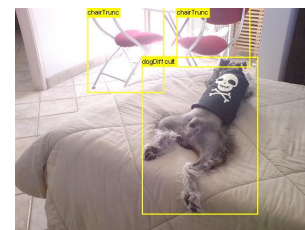
Car



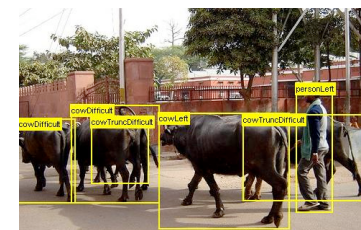
Cat



Chair

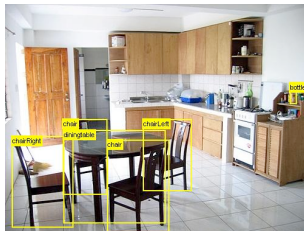


Cow

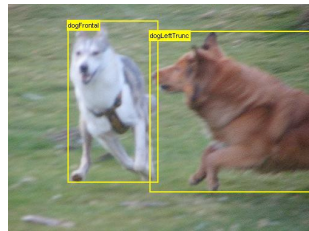


Examples

Dining Table



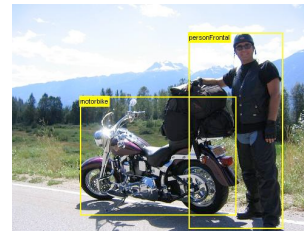
Dog



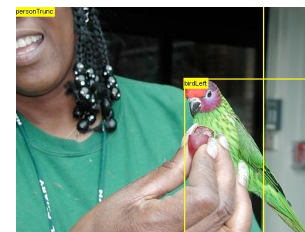
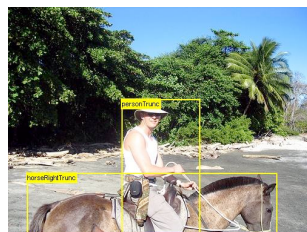
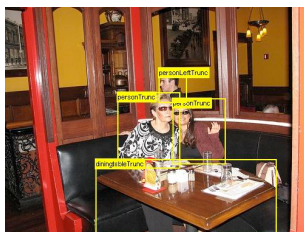
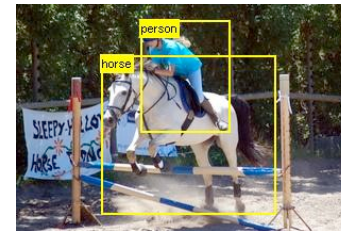
Horse



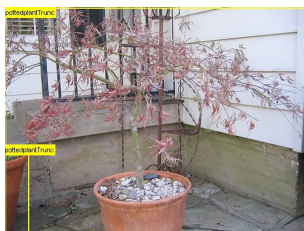
Motorbike



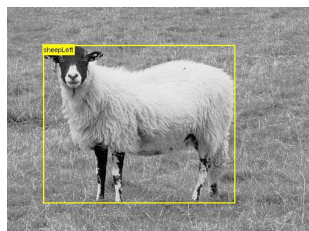
Person



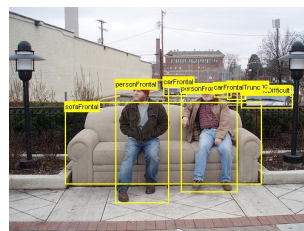
Potted Plant



Sheep



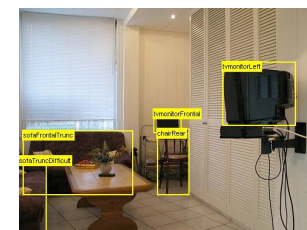
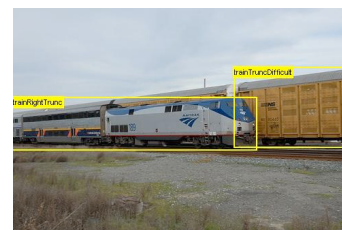
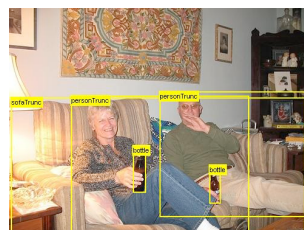
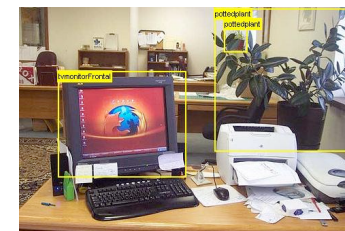
Sofa



Train



TV/Monitor



Main Challenge Tasks

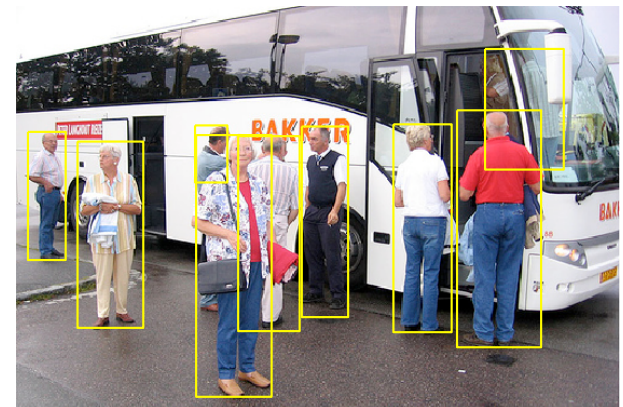
- **Classification**

- Is there a dog in this image?
- Evaluation by precision/recall



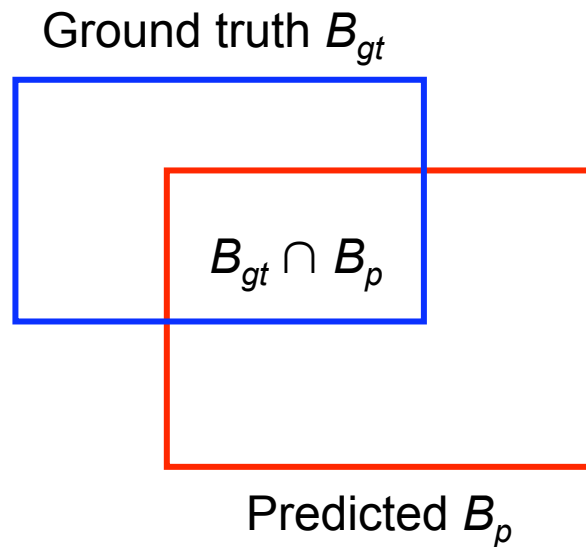
- **Detection**

- Localize all the people (if any) in this image
- Evaluation by precision/recall based on bounding box overlap



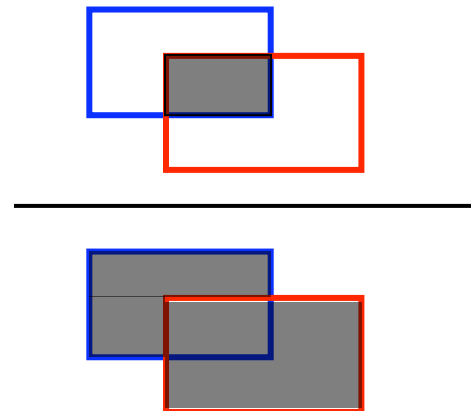
Detection: Evaluation of Bounding Boxes

- Area of Overlap (AO) Measure



$$AO(B_{gt}, B_p) = \frac{|B_{gt} \cap B_p|}{|B_{gt} \cup B_p|}$$

Detection if



> Threshold

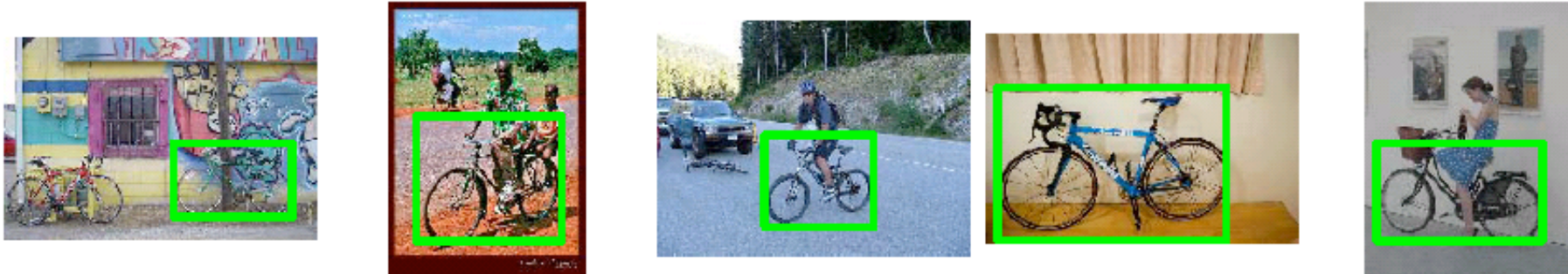
50%

Dataset Statistics

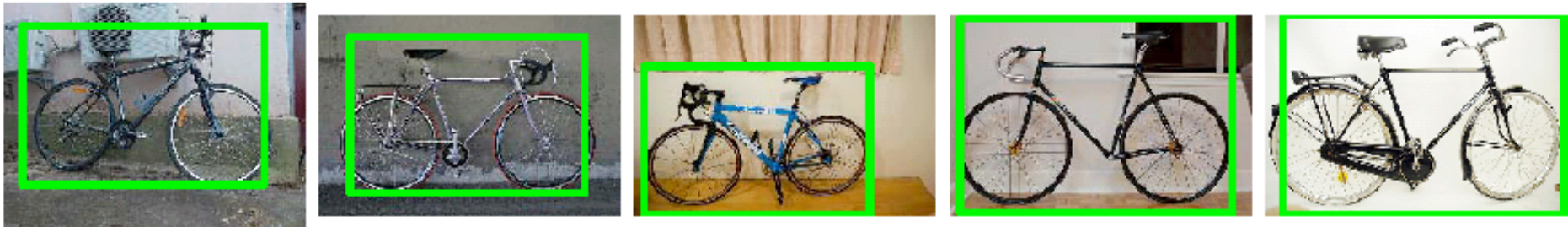
	train		val		trainval		test	
	Images	Objects	Images	Objects	Images	Objects	Images	Objects
Aeroplane	201	267	206	266	407	533		
Bicycle	167	232	181	236	348	468		
Bird	262	381	243	379	505	760		
Boat	170	270	155	267	325	537		
Bottle	220	394	200	393	420	787		
Bus	132	179	126	186	258	365		
Car	372	664	358	653	730	1,317		
Cat	266	308	277	314	543	622		
Chair	338	716	330	713	668	1,429		
Cow	86	164	86	172	172	336		
Diningtable	140	153	131	153	271	306		
Dog	316	391	333	392	649	783		
Horse	161	237	167	245	328	482		
Motorbike	171	235	167	234	338	469		
Person	1,333	2,819	1,446	2,996	2,779	5,815		
Pottedplant	166	311	166	316	332	627		
Sheep	67	163	64	175	131	338		
Sofa	155	172	153	175	308	347		
Train	164	190	160	191	324	381		
Tvmonitor	180	259	173	257	353	516		
Total	3,473	8,505	3,581	8,713	7,054	17,218	6,650	16,829

True Positives - Bicycle

UoCTTI_LSVM-MDPM



OXFORD_MKL

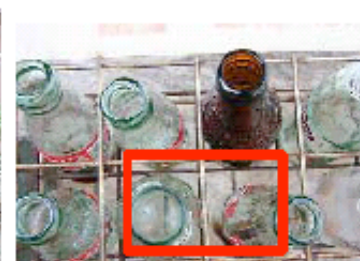


NECUIUC_CLS-DTCT



False Positives - Bicycle

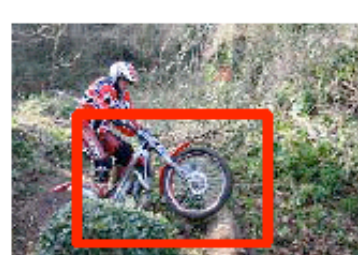
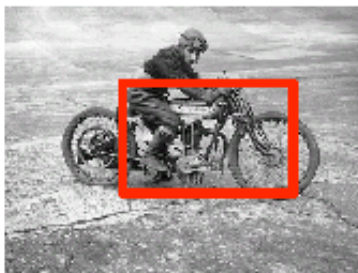
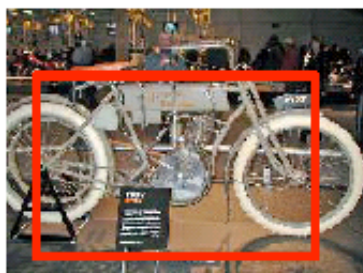
UoCTTI_LSVM-MDPM



OXFORD_MKL

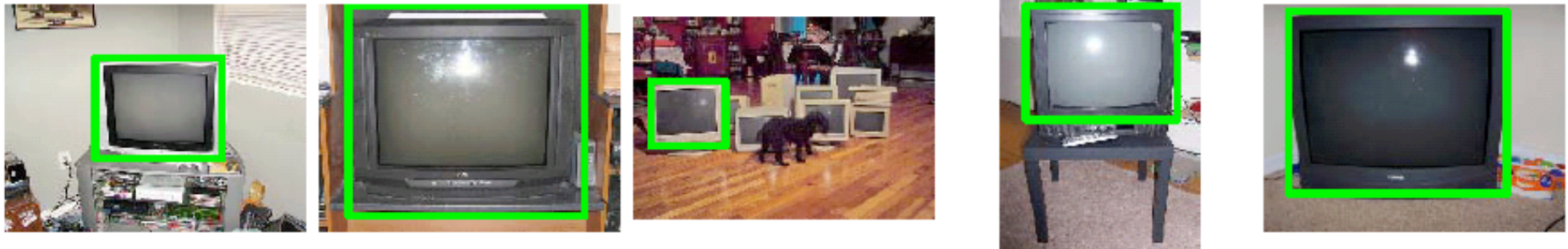


NECUIUC_CLS-DTCT



True Positives – TV/monitor

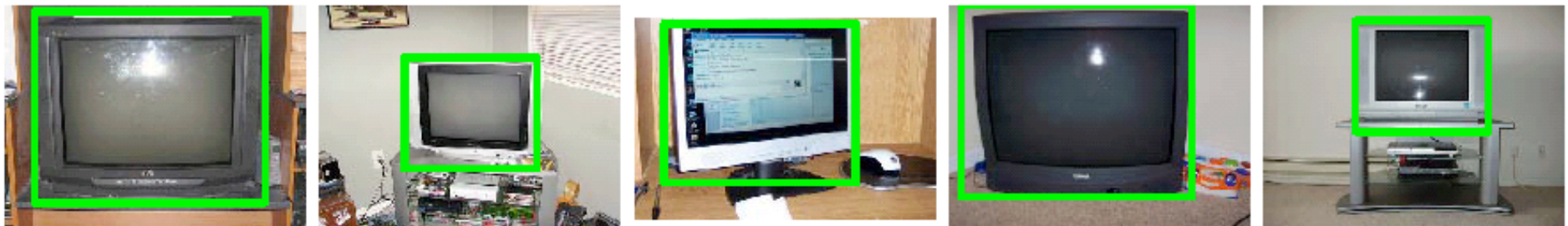
OXFORD_MKL



UoCTTI_LSVM-MDPM

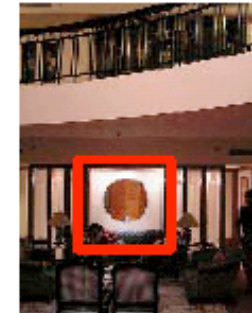
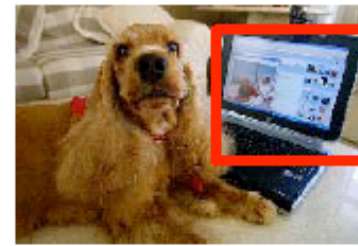
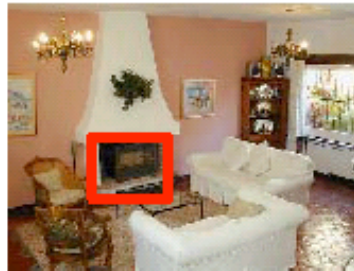
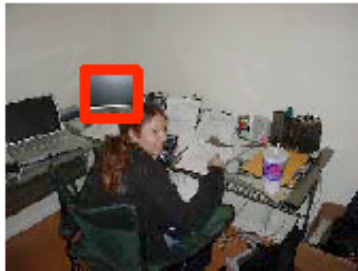


LEAR_CHI-SVM-SIFT-HOG-CLS

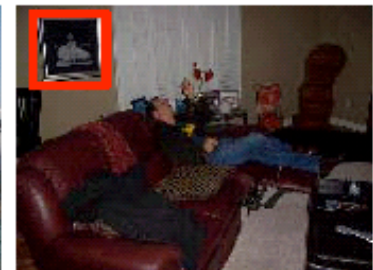
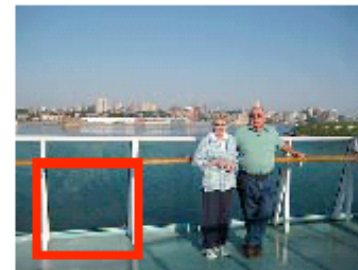
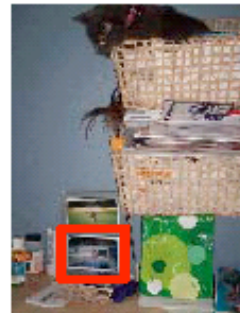


False Positives – TV/monitor

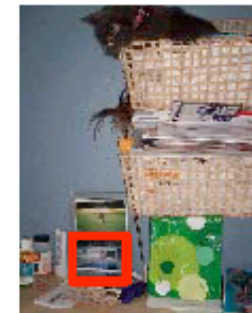
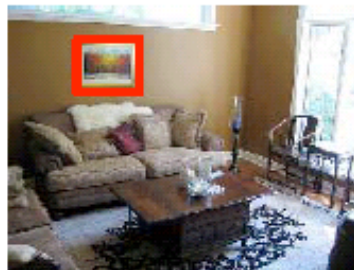
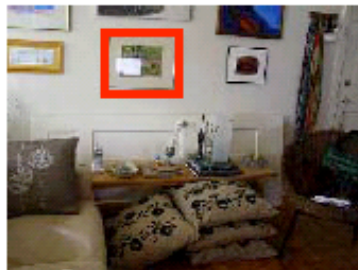
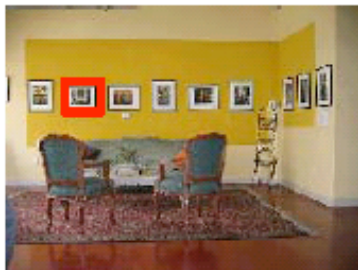
OXFORD_MKL



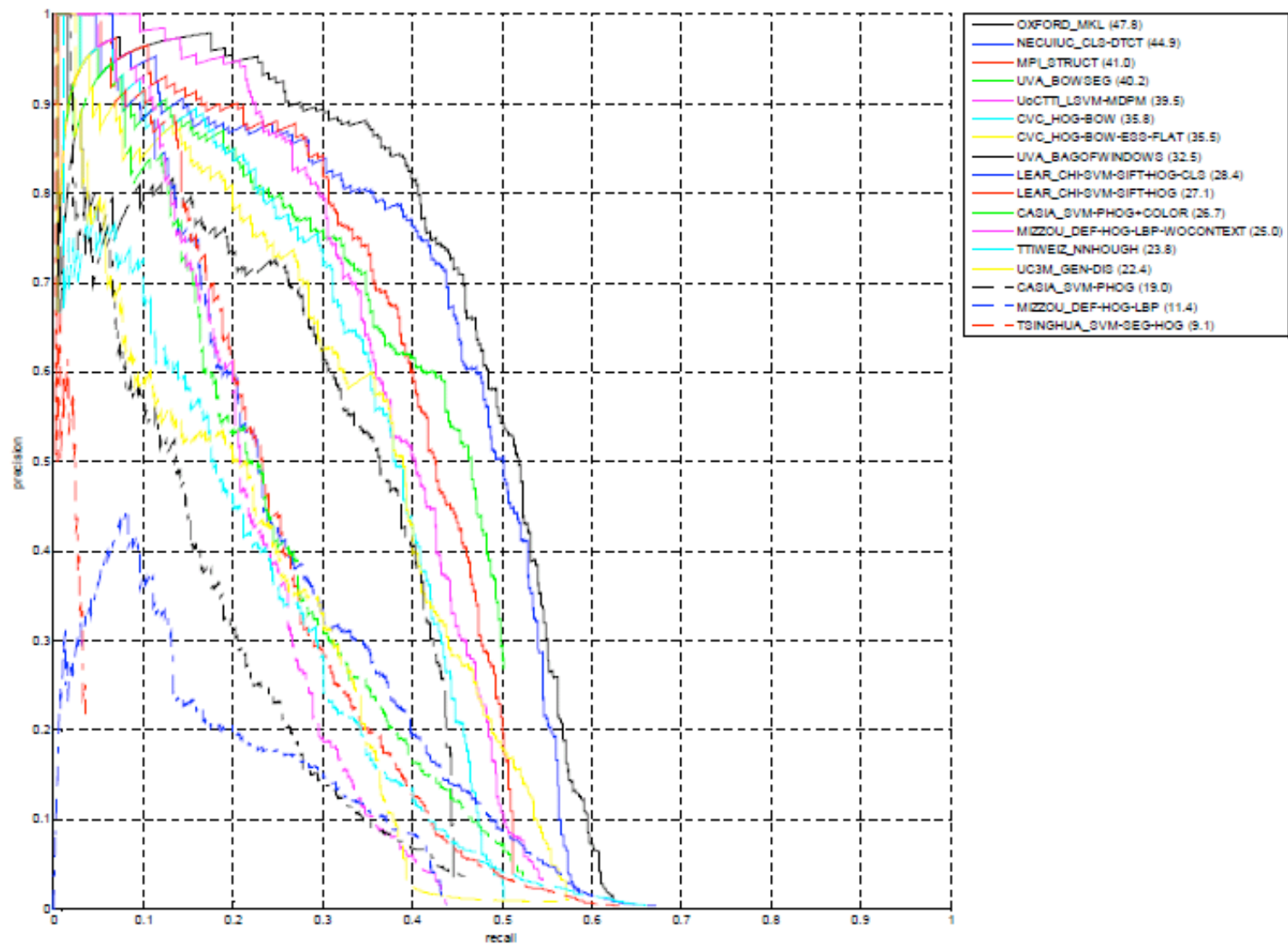
UoCTTI_LSVM-MDPM



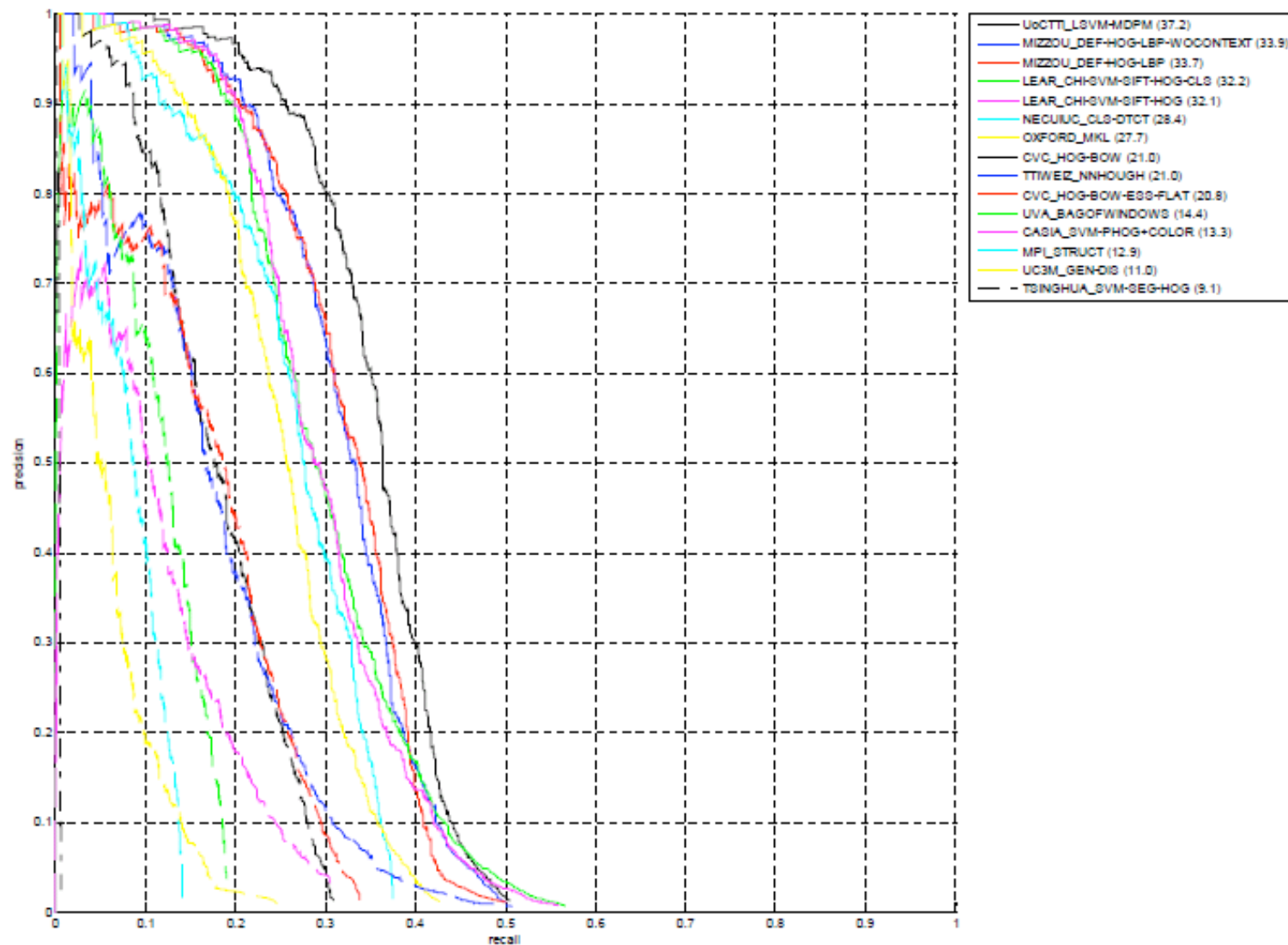
LEAR_CHI-SVM-SIFT-HOG-CLS



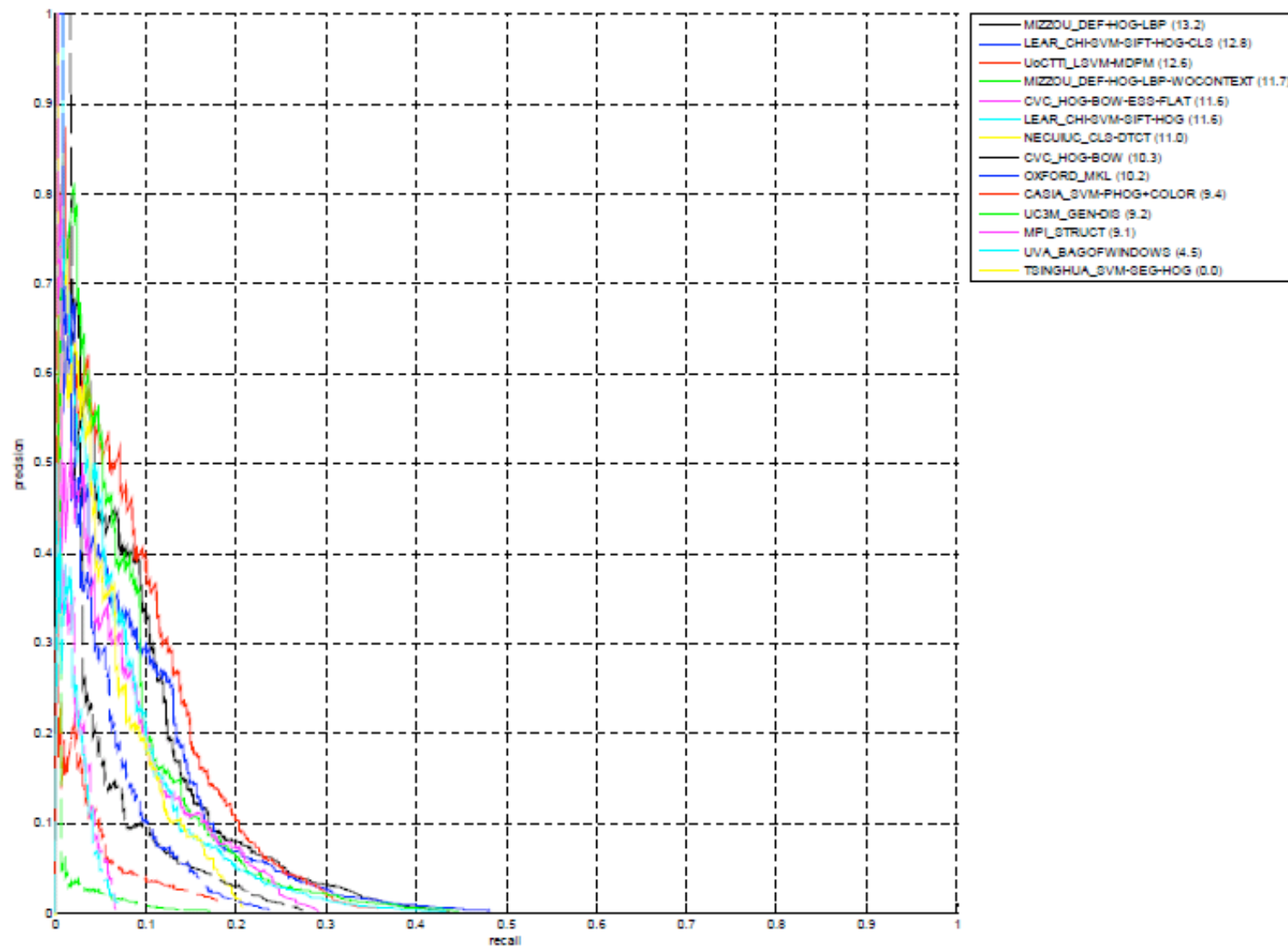
Precision/Recall - Aeroplane



Precision/Recall - Car

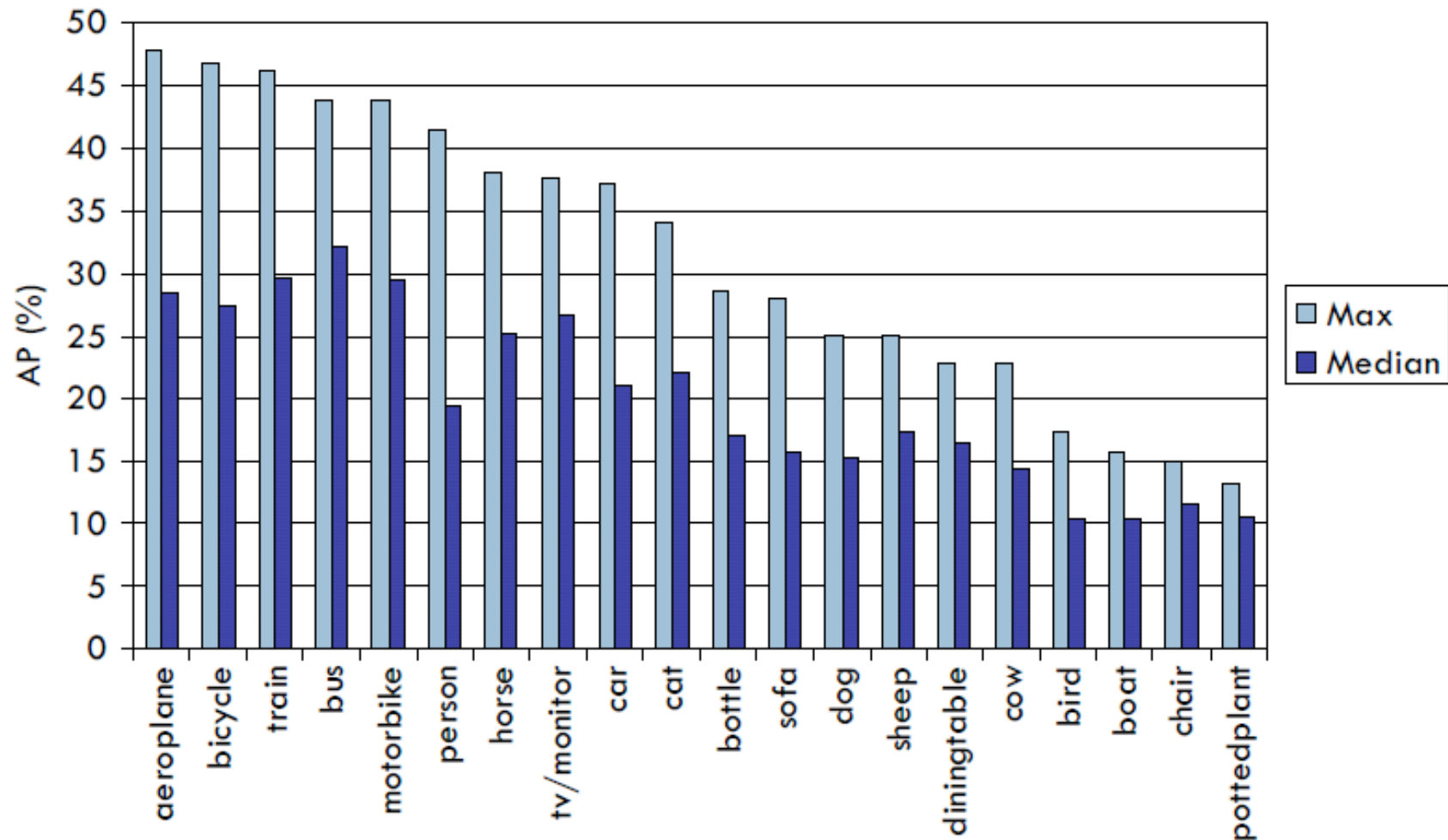


Precision/Recall – Potted plant



AP by Class

Detection



Wide variety of methods: sliding window, combination with whole image classifiers, segmentation based

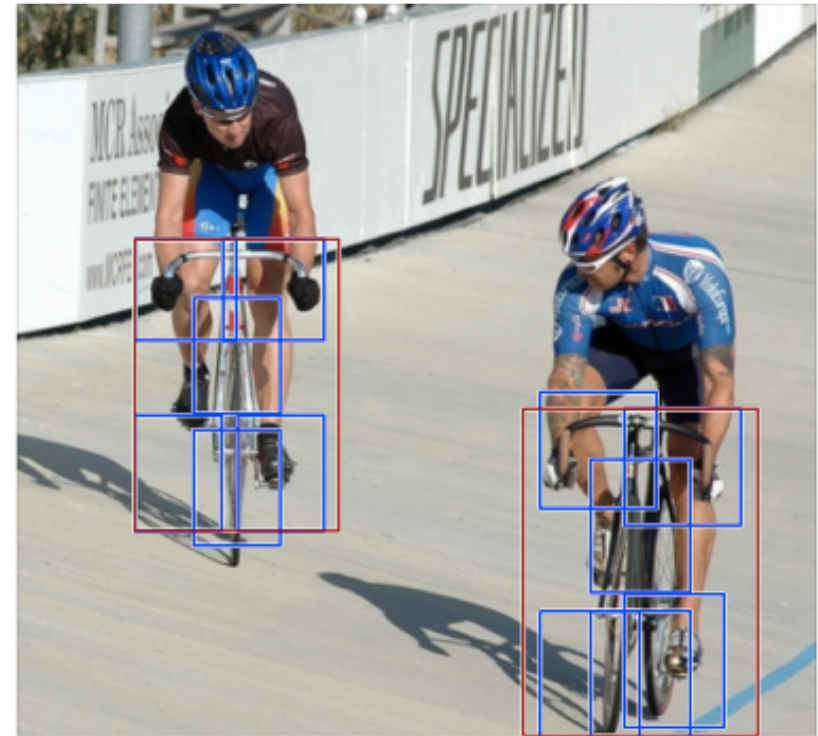
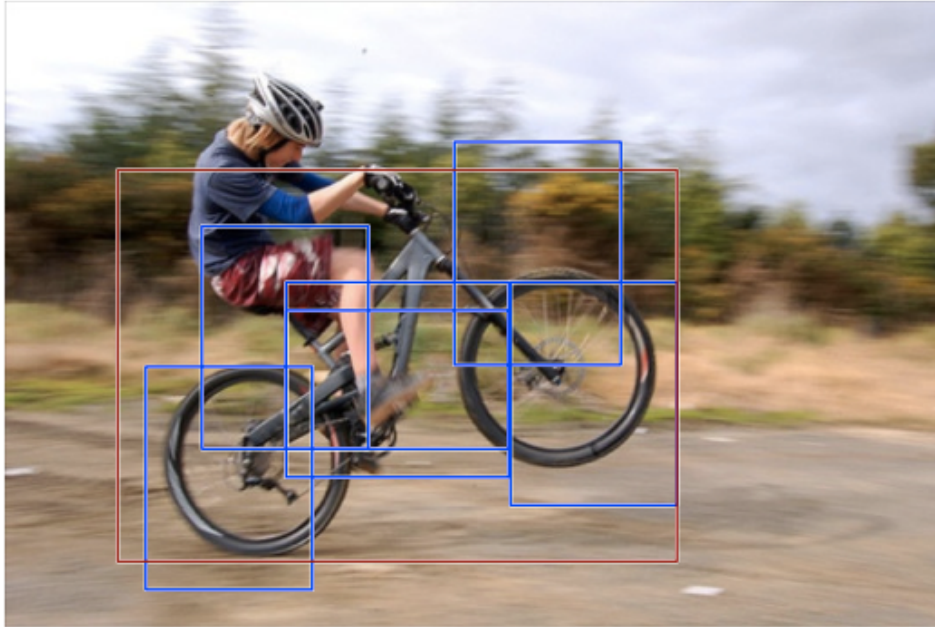
Object Detection with Discriminatively Trained Part Based Models

Pedro F. Felzenszwalb, David Mcallester,
Deva Ramanan, Ross Girshick

PAMI 2010

Matlab code available online:
<http://www.cs.brown.edu/~pff/latent/>

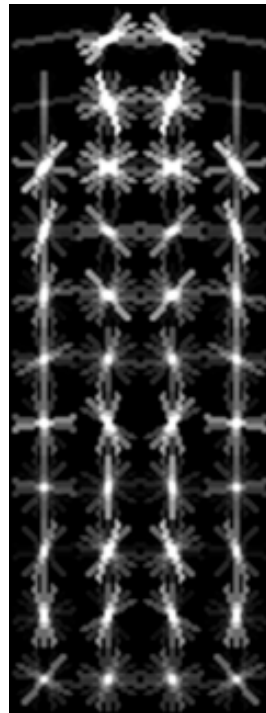
Approach



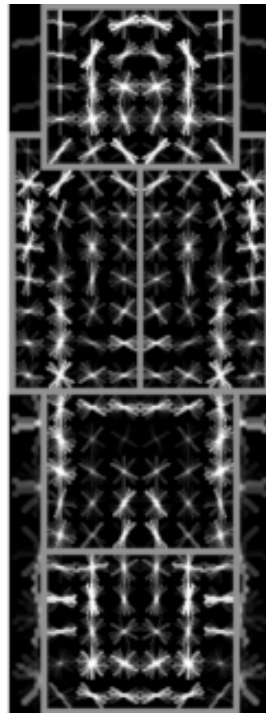
- Mixture of deformable part-based models
 - One component per “aspect” e.g. front/side view
- Each component has global template + deformable parts
- Discriminative training from bounding boxes alone

Example Model

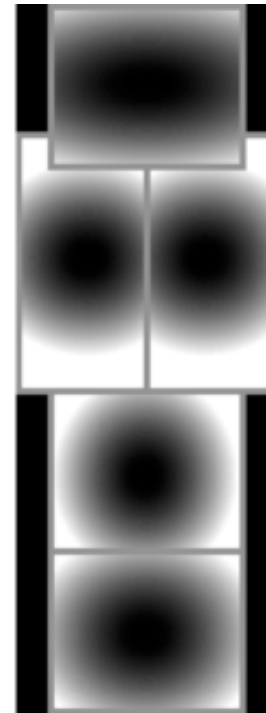
- One component of person model



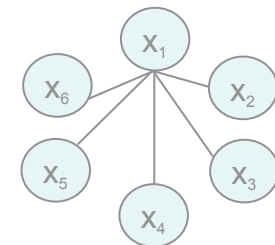
root filters
coarse resolution



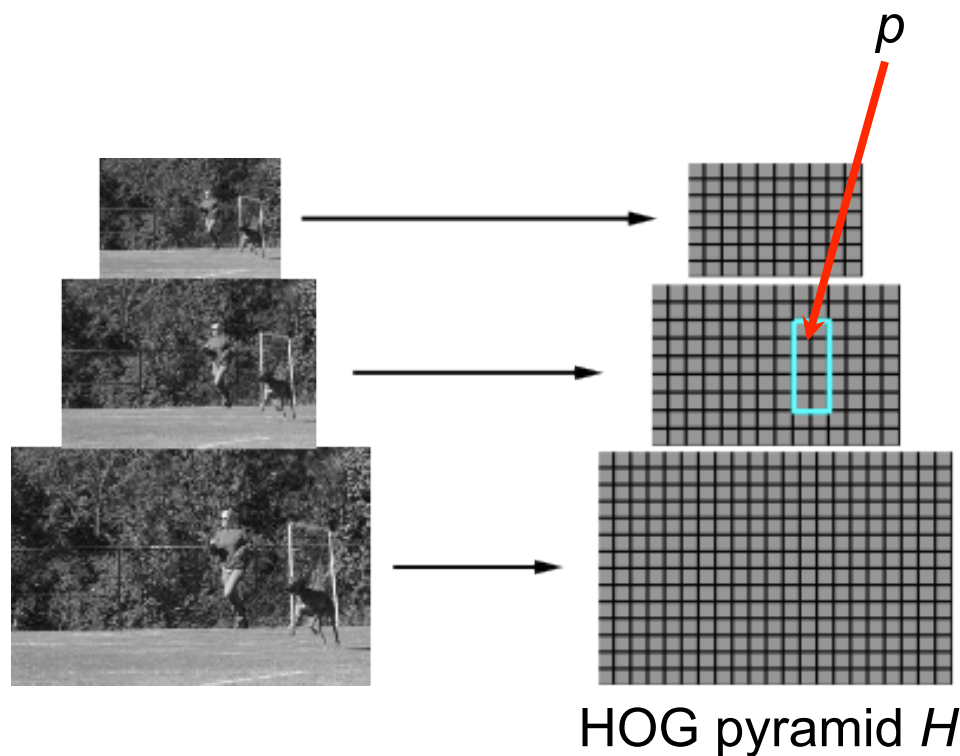
part filters
finer resolution



deformation
models



Starting Point: HOG Filter



Score of F at position p is
 $F \cdot \varphi(p, H)$

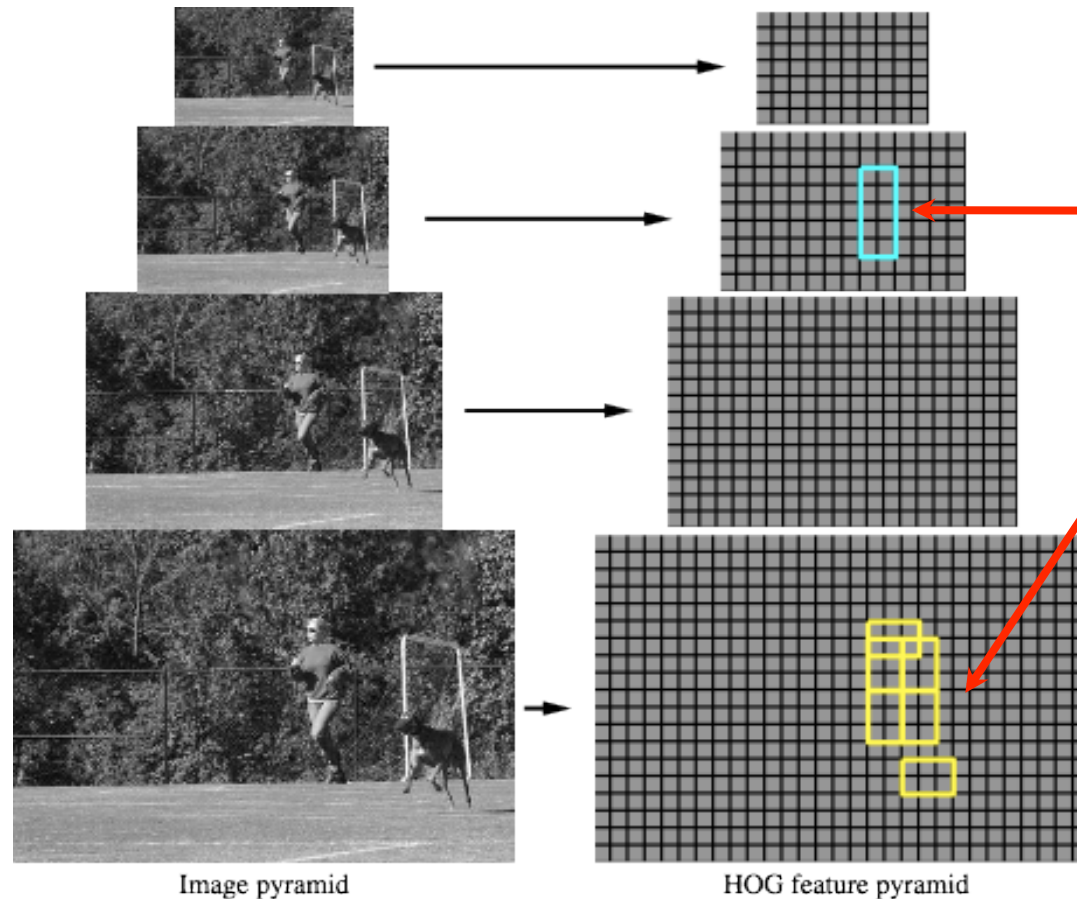
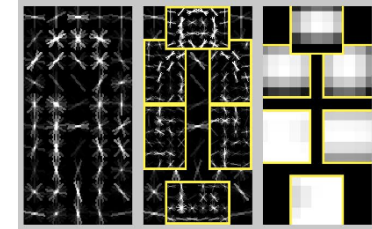
$\varphi(p, H)$ = concatenation of
HOG features from
subwindow specified by p

- Search: sliding window over position and scale
- Feature extraction: HOG Descriptor
- Classifier: Linear SVM

Dalal & Triggs [2005]

Object Hypothesis

- Position of root + each part
- Each part: HOG filter (at higher resolution)



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

p_0 : location of root

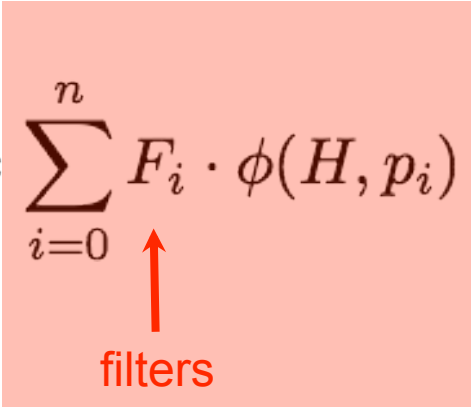
p_1, \dots, p_n : location of parts

Score is sum of filter
scores minus
deformation costs

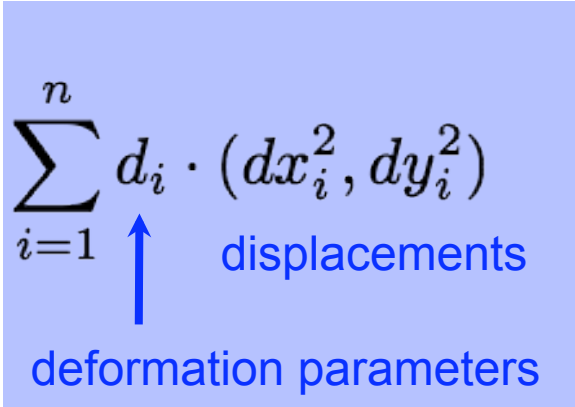
Score of a Hypothesis

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

Appearance term
Spatial prior

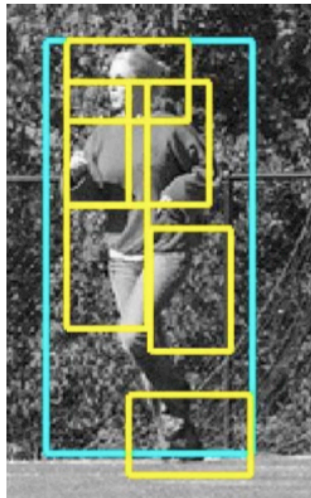


filters



displacements

deformation parameters



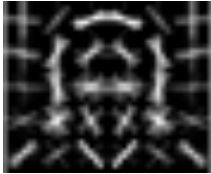
$$\text{score}(z) = \beta \cdot \Psi(H, z)$$

concatenation of filters
and deformation
parameters

concatenation of
HOG features and
part displacement
features

- Linear classifier applied to feature subset defined by hypothesis

Part Detection



head filter

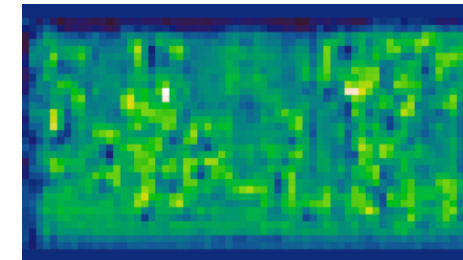
input image



Response of filter in l-th pyramid level

$$R_l(x, y) = F \cdot \phi(H, (x, y, l))$$

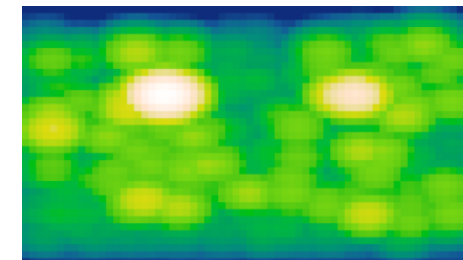
cross-correlation



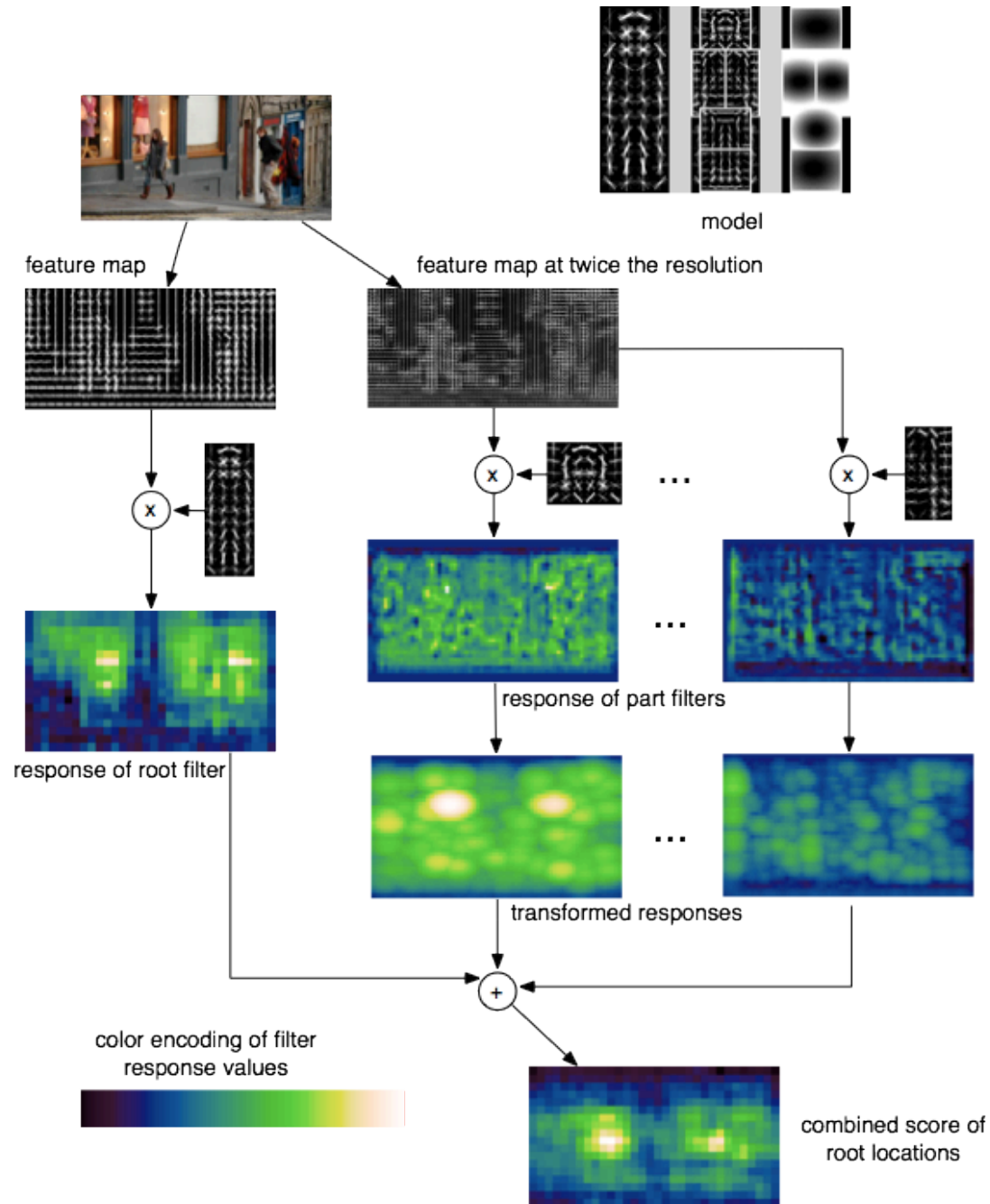
Transformed response

$$D_l(x, y) = \max_{dx, dy} (R_l(x + dx, y + dy) - d_i \cdot (dx^2, dy^2))$$

max-convolution, computed in linear time
(spreading, local max, etc)

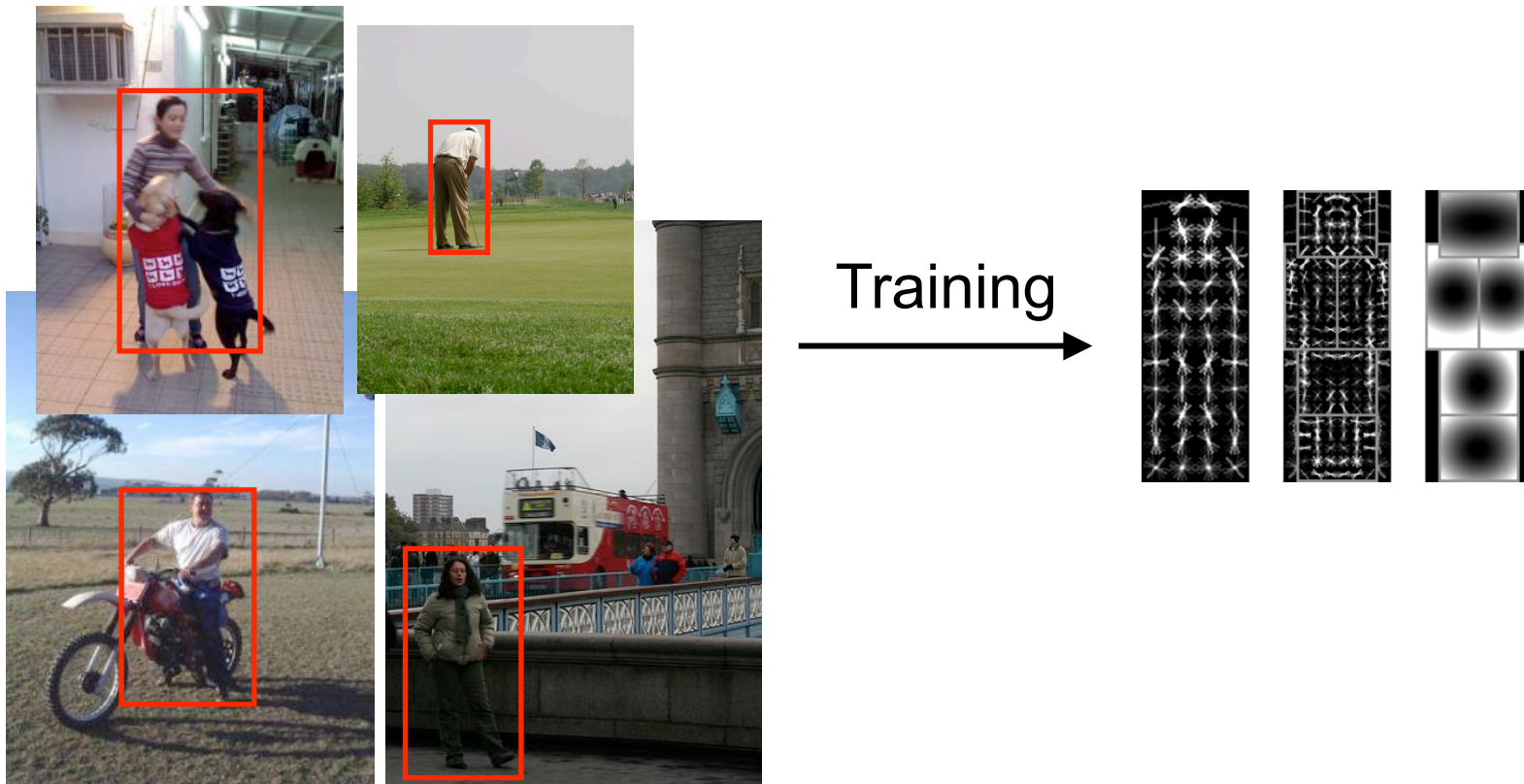


System



Training

- Training data = images + bounding boxes
- Need to learn: model structure, filters, deformation costs



Latent SVM (MI-SVM)

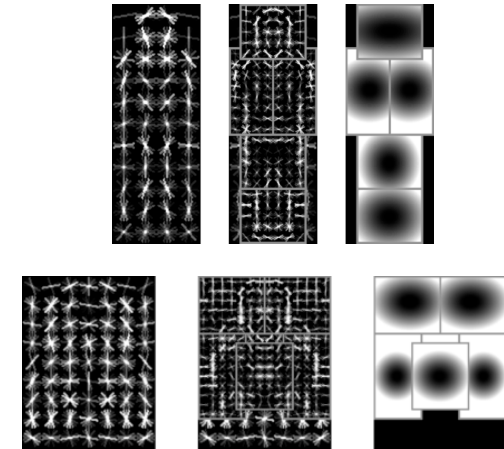
Classifiers that score an example x using

$$f_{\beta}(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$$

β are model parameters

z are latent values

- Which component?
- Where are the parts?



Training data $D = (\langle x_1, y_1 \rangle, \dots, \langle x_n, y_n \rangle)$ $y_i \in \{-1, 1\}$

We would like to find β such that: $y_i f_{\beta}(x_i) > 0$

Minimize Regularizer “Hinge loss” on one training example

$$L_D(\beta) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i f_{\beta}(x_i))$$

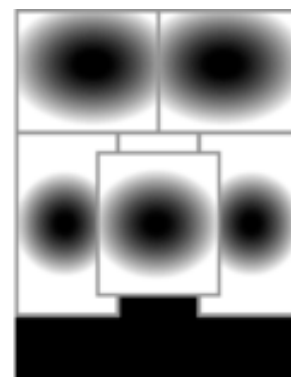
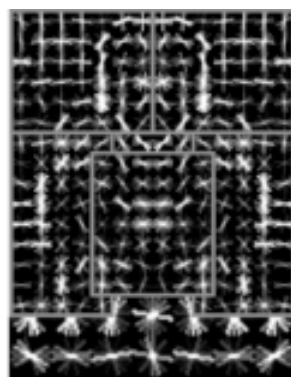
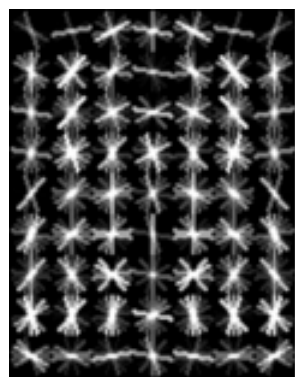
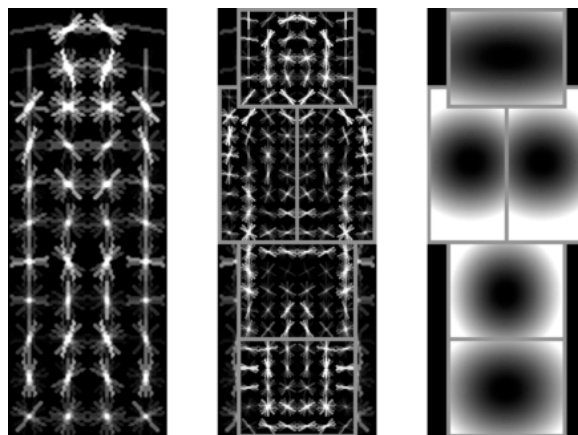
SVM objective

Latent SVM Training

$$L_D(\beta) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i f_{\beta}(x_i))$$

- Convex if we fix z for positive examples
 - Optimization:
 - Initialize β and iterate:
 - Pick best z for each positive example
 - Optimize β with z fixed
 - Local minimum: needs good initialization
 - Parts initialized heuristically from root
- } Alternation strategy

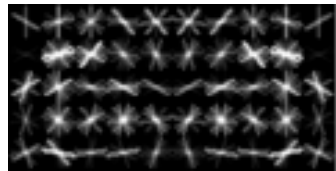
Person Model



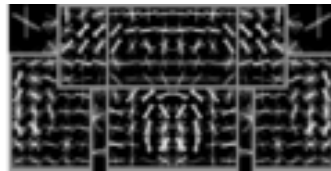
root filters part filters deformation
coarse resolution finer resolution models

Handles partial occlusion/truncation

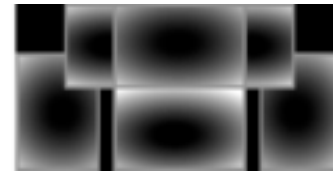
Car Model



root filters
coarse resolution



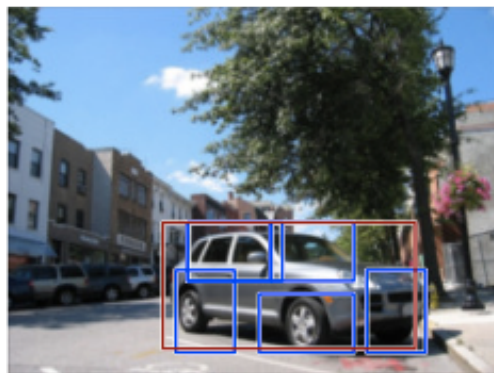
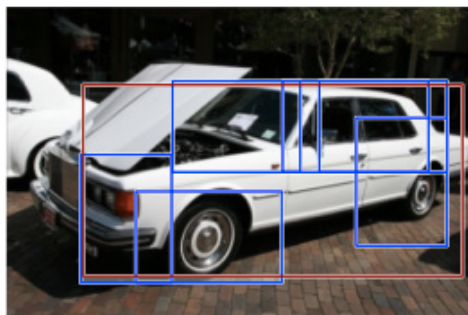
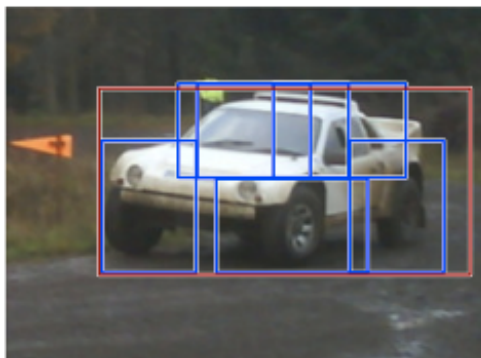
part filters
finer resolution



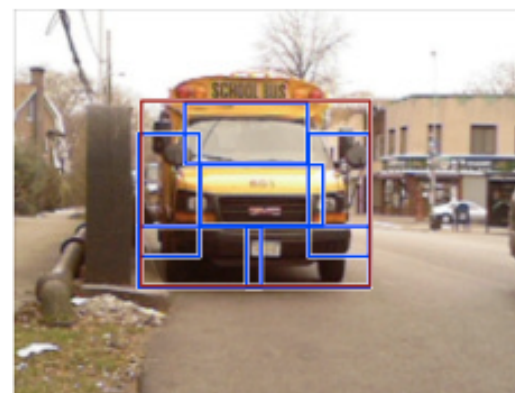
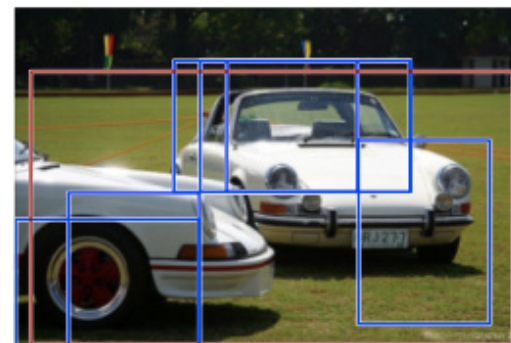
deformation
models

Car Detections

high scoring true positives

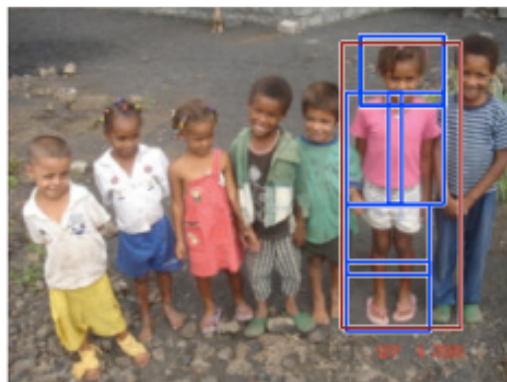
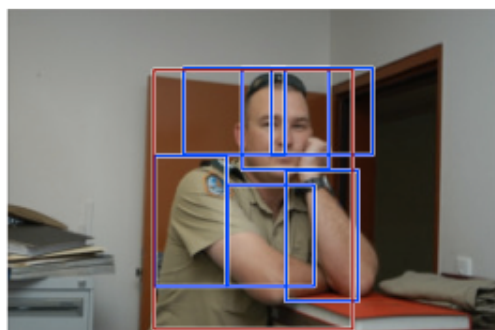
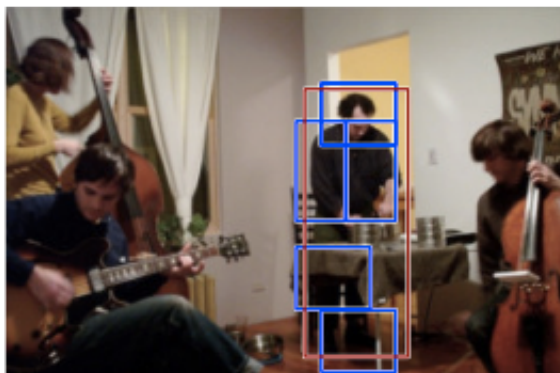


high scoring false positives

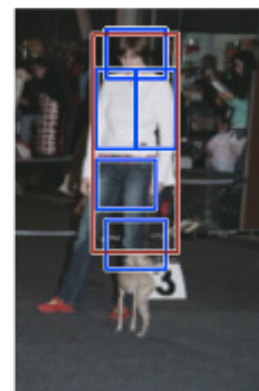


Person Detections

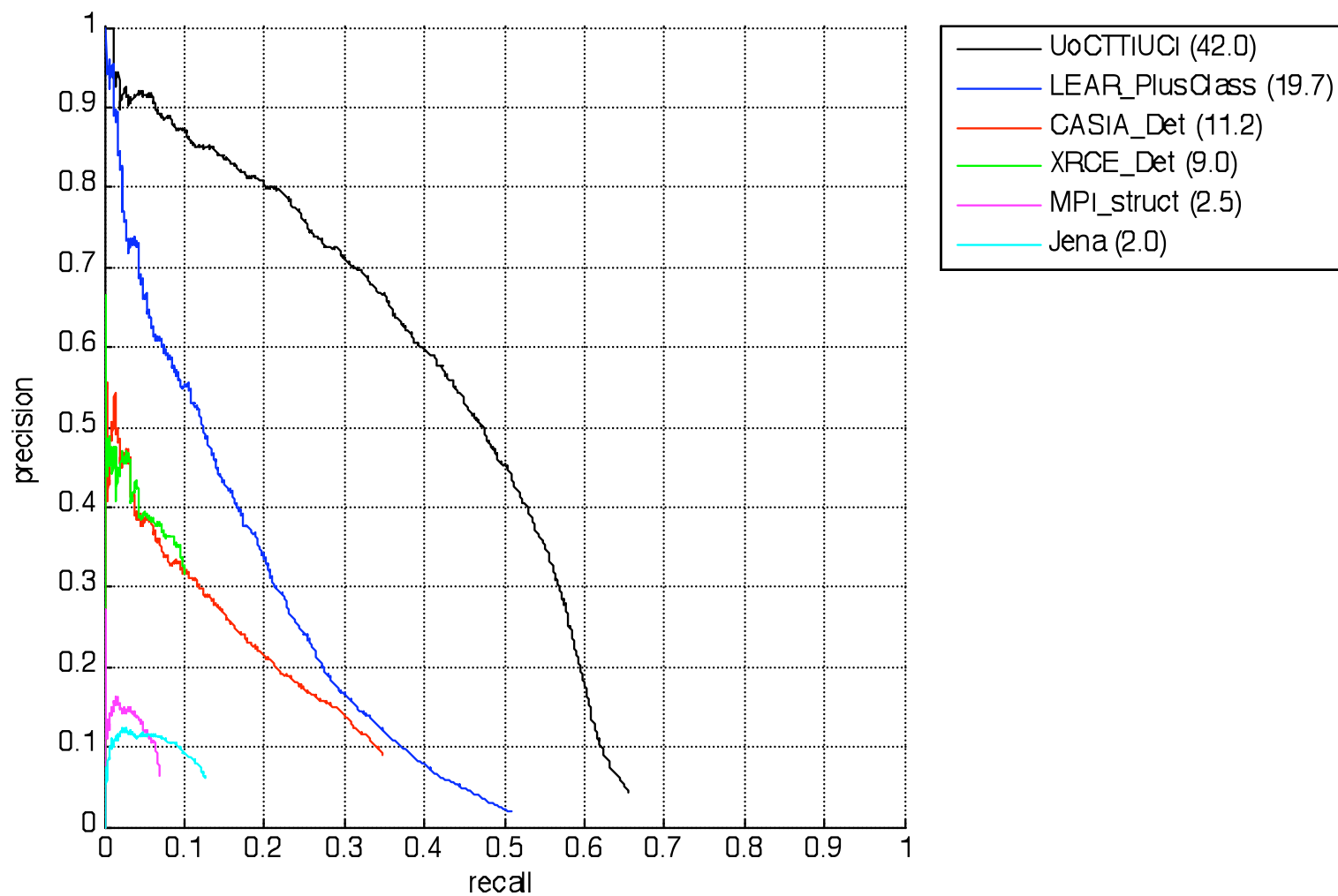
high scoring true positives



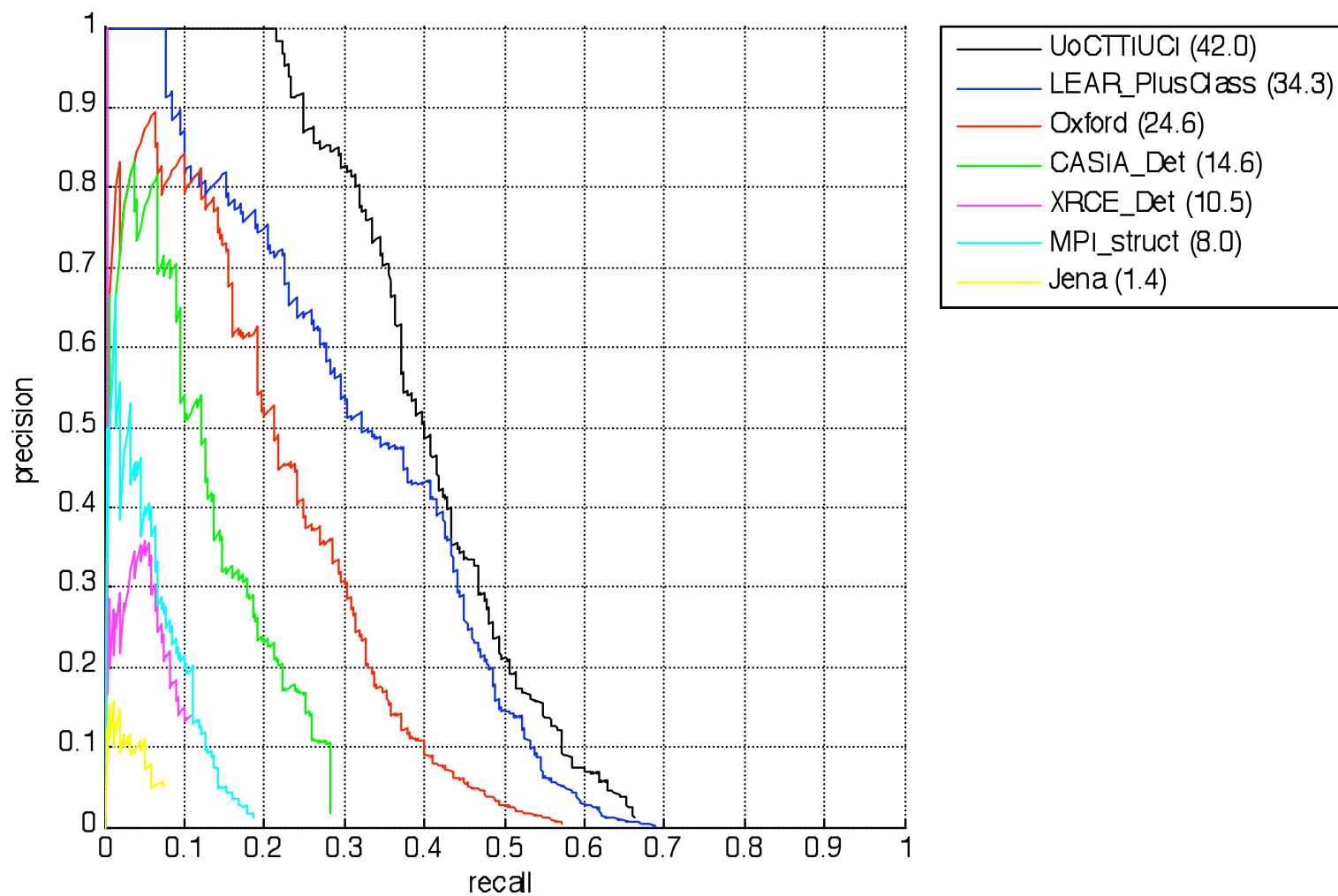
high scoring false positives
(not enough overlap)



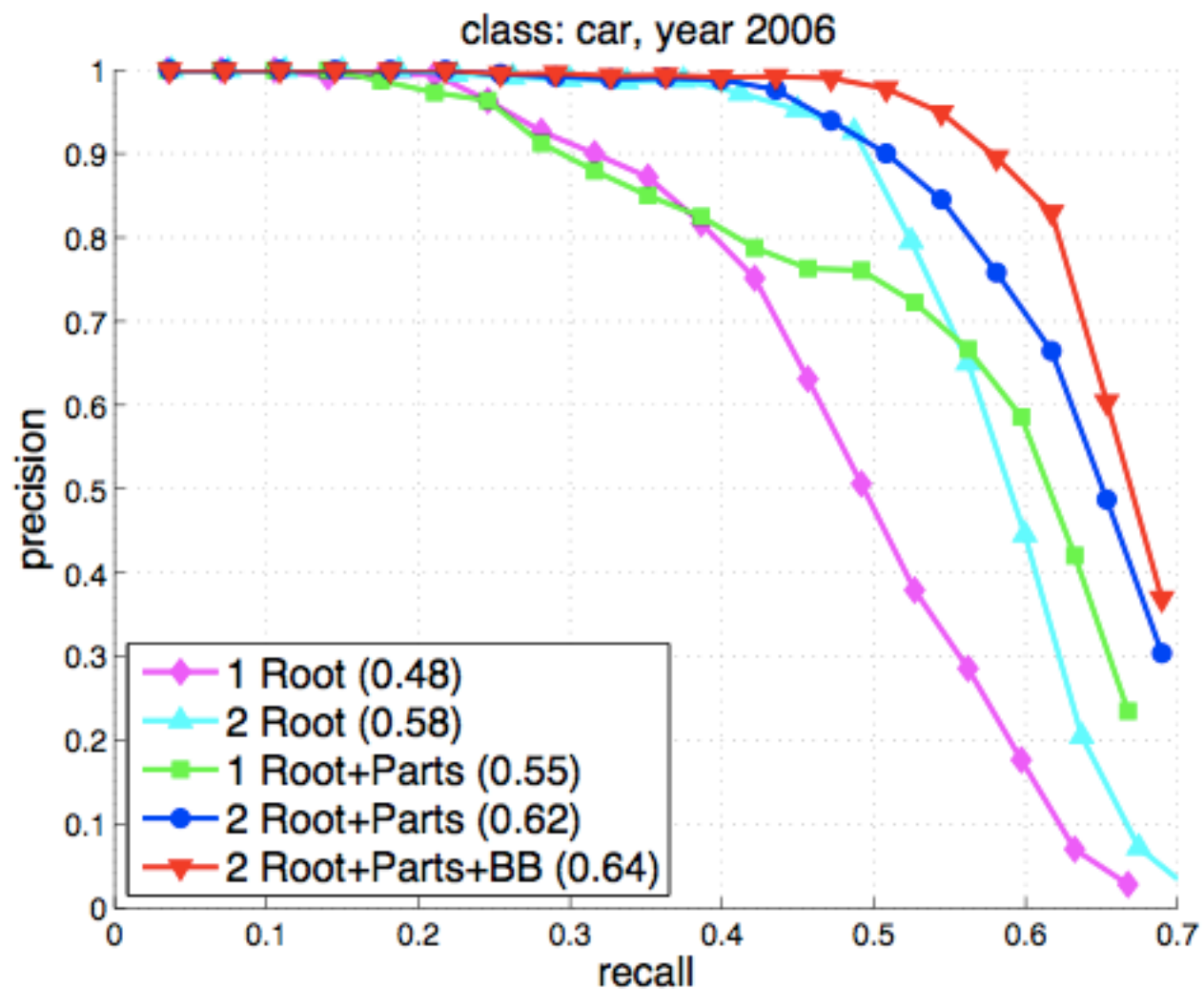
Precision/Recall: VOC2008 Person



Precision/Recall: VOC2008 Bicycle

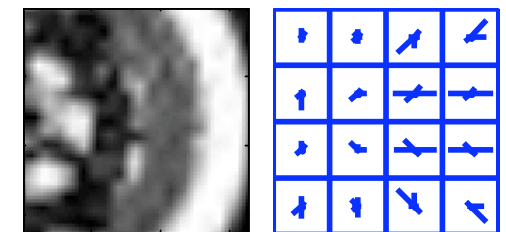
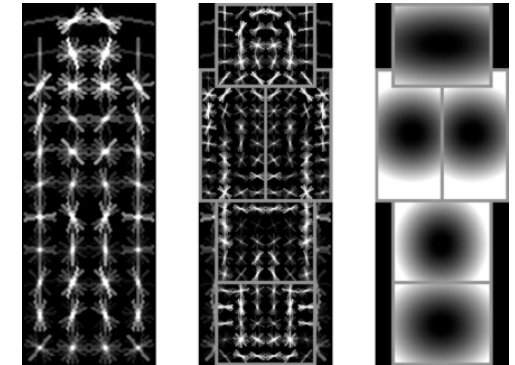
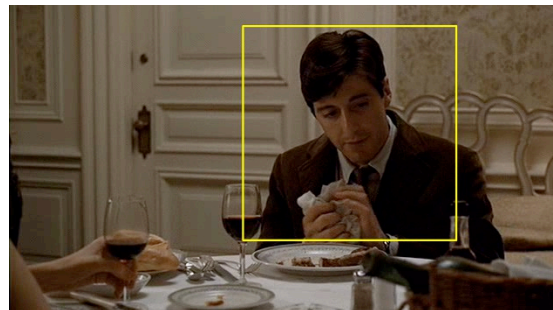
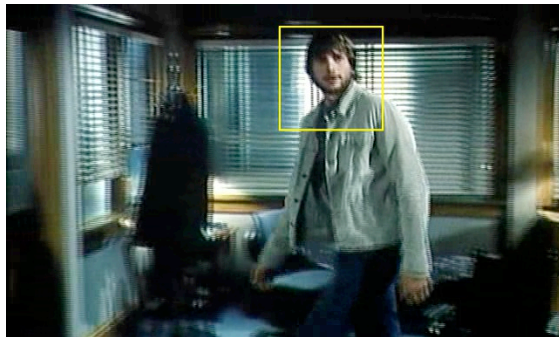


Comparison of Models



Summary

- Multiple features and multiple kernels boost performance
- Discriminative learning of model with latent variables for single feature (HOG):
 - Latent variables can learn best alignment in the ROI training annotation
 - Parts can be thought of as local SIFT vectors
 - Some similarities to Implicit Shape Model/ Constellation models but with discriminative/ careful training throughout



NB: Code available for latent model !

Outline

1. Sliding window detectors
2. Features and adding spatial information
3. HOG + linear SVM classifier
4. Two state of the art algorithms and PASCAL VOC
5. The future and challenges

Current Research Challenges

- Context
 - from scene properties: GIST, BoW, stuff
 - from other objects
 - from geometry of scene, e.g. Hoiem et al CVPR 06
- Occlusion/truncation
 - Winn & Shotton, Layout Consistent Random Field, CVPR 06
 - Vedaldi & Zisserman, NIPS 09
 - Yang et al, Layered Object Detection, CVPR 10
- 3D
- Scaling up – thousands of classes
 - Torralba et al, Feature sharing
 - ImageNet
- Weak and noisy supervision

