Computer Vision

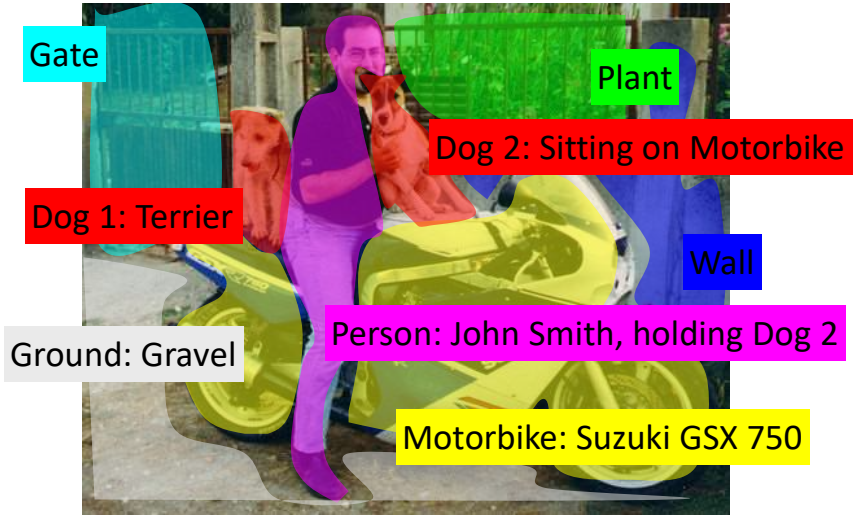
DATA.ML.300, 5 study credits

Esa Rahtu
Unit of Computing Sciences, Tampere University

Object category detection

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,…



Things vs. Stuff

Thing: Object with specific shape

Stuff: Material defined by homogenous or repetitive pattern. Has no specific spatial extent or shape.

Ted Adelson, Forsyth et al. 1996

Things vs. Stuff

Thing: Object with specific shape

Stuff: Material defined by homogenous or repetitive pattern. Has no specific spatial extent or shape.

Ted Adelson, Forsyth et al. 1996

Recognition tasks

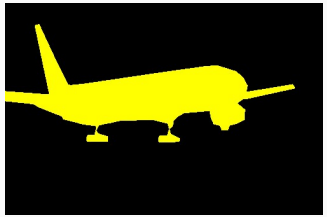
- Image classification
  - Does the image contain an aeroplane?

Recognition tasks

- Image classification
  - Does the image contain an aeroplane?
- Object class detection/localization
  - Where are the aeroplanes (if any)?

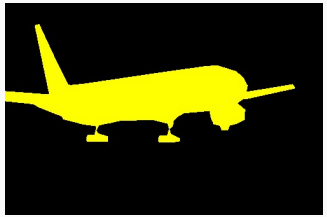
Recognition tasks

- 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?



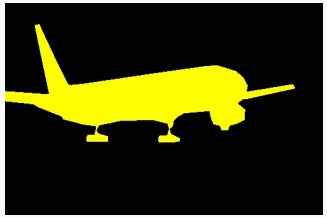
Recognition tasks

- 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?
- (Panoptic segmentation)



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 make aeroplane (if any)?
- (Panoptic segmentation)



Challenges?

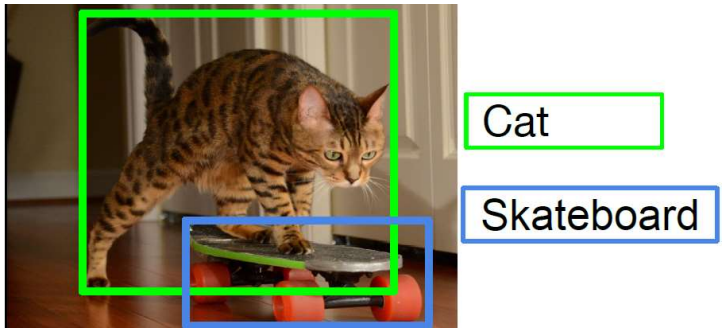
Background clutter

Occlusions and truncation

Intra class variation

So why bother?

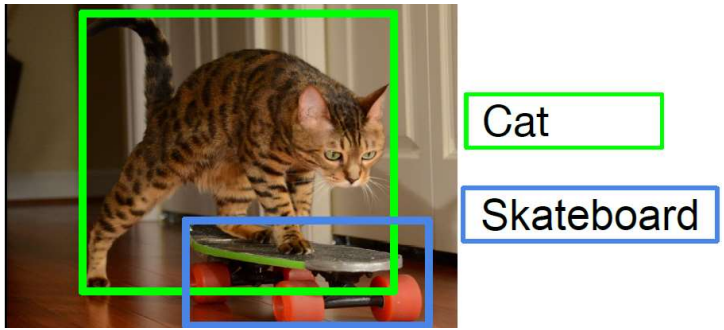
- Spatial relationships for image understanding and retrieval



Credits: A Zisserman

So why bother?

- Spatial relationships for image understanding and retrieval

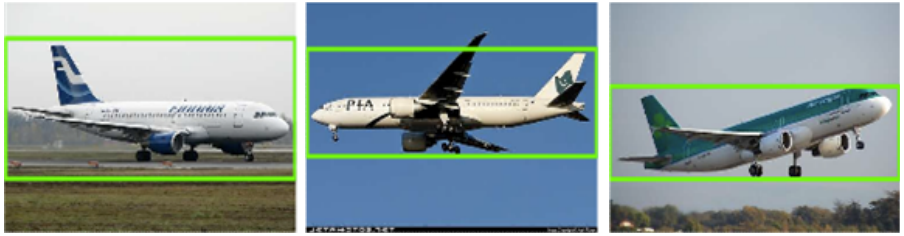


“a cat riding a skateboard”

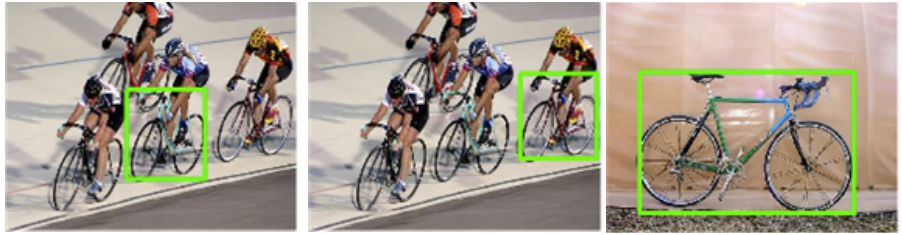
- Visual question and answering, object grasping/tracking

Credits: A Zisserman

Preview of typical results



Aeroplane



Bicycle

Car

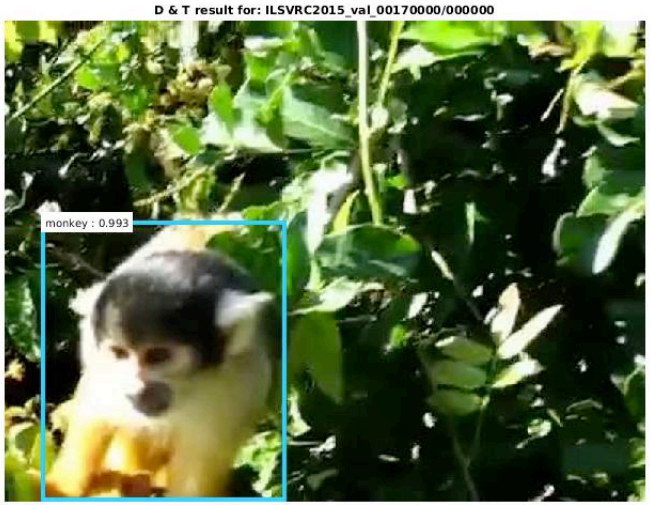
Cow

Horse

Motorbike

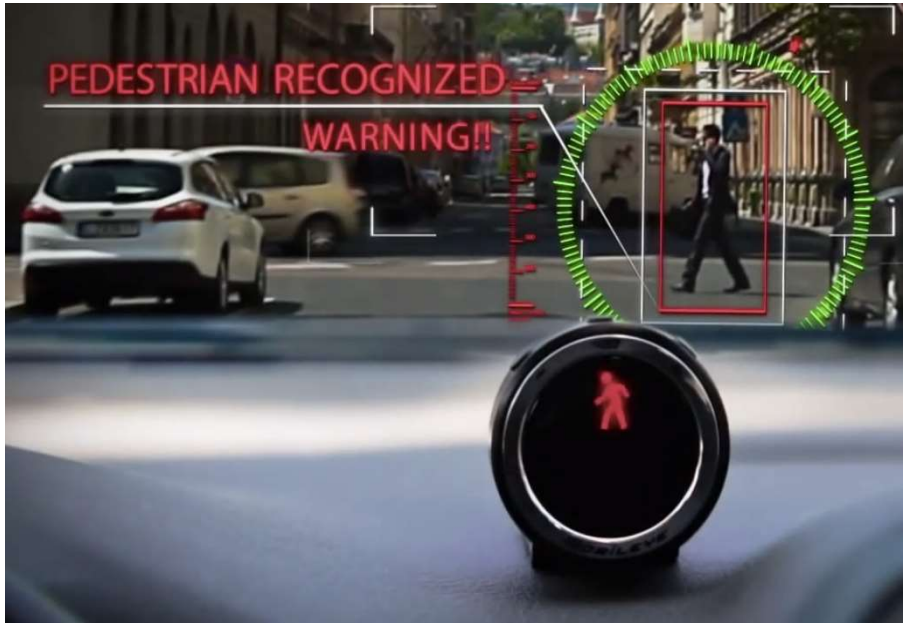
Credits: A Zisserman

Preview of tracking by detection



Detect to track and track to detect, Feichtenhofer, Pinz, Zisserman, ICCV 2017

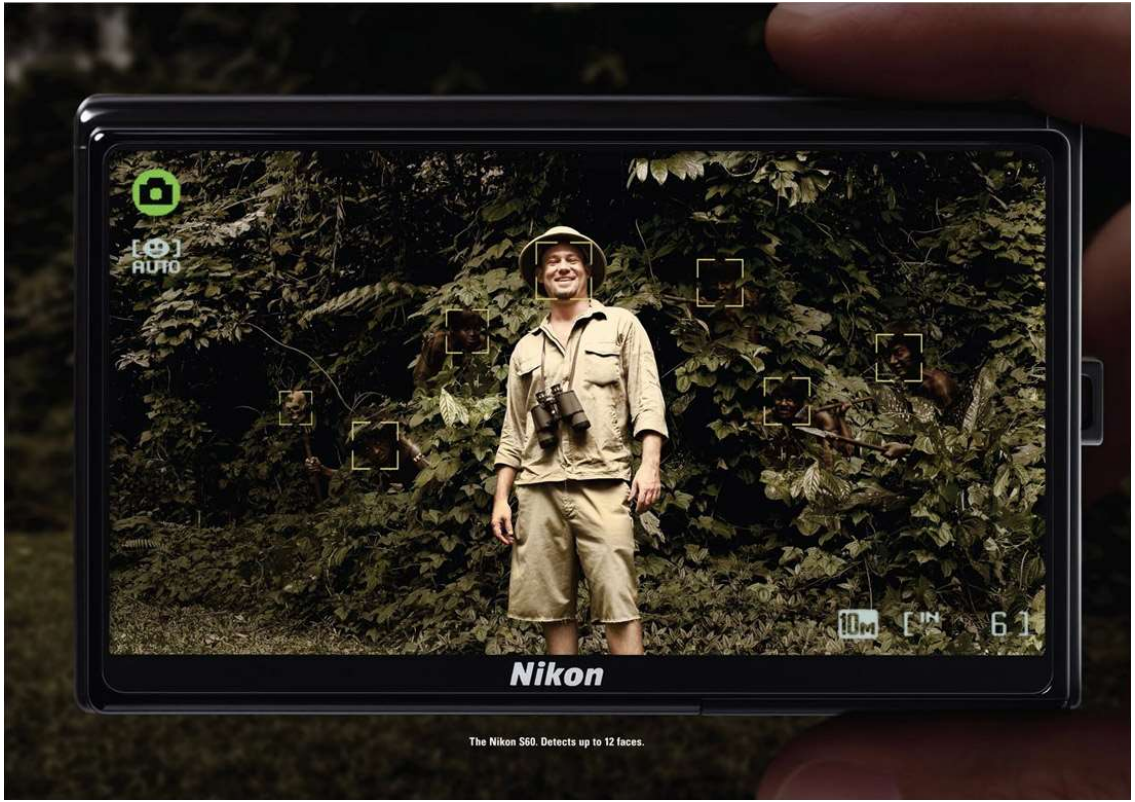
Application: collision prevention



www.mobileye.com

Application: Funny Nikon ads

"Nikon S60 detects up to 12 faces."



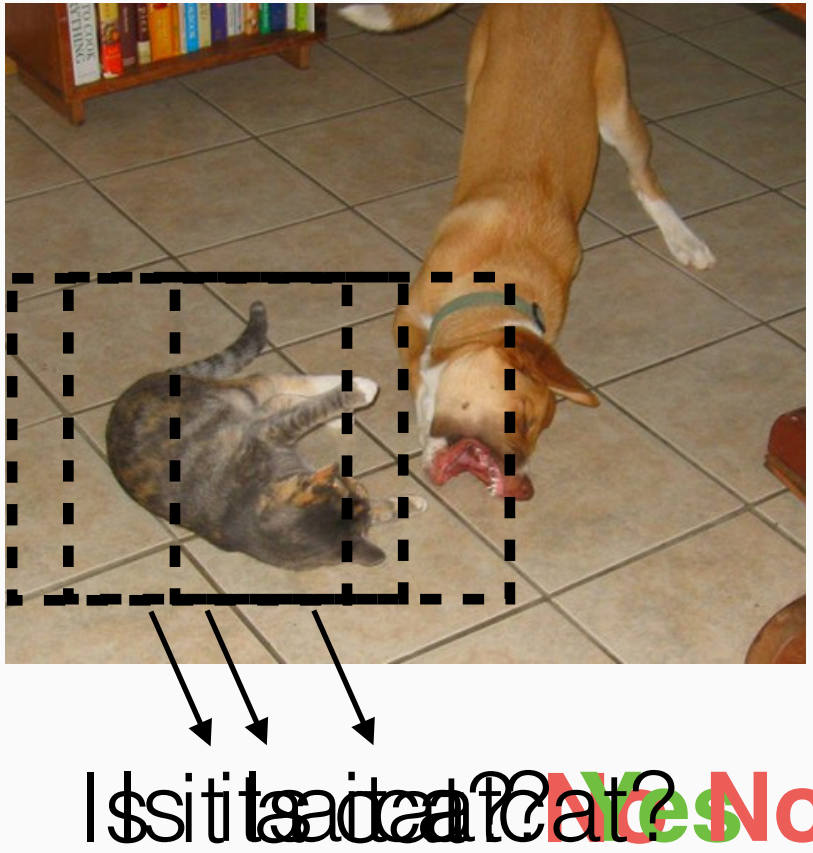
Outline for this week

- **Lecture 1: Principles of sliding window detectors**
  - Training a sliding window detector
  - Speeding up inferences
- Lecture 2: Deep networks for object category detection
  - Two-stage and one-stage networks
  - State-of-the-art

Sliding window detector

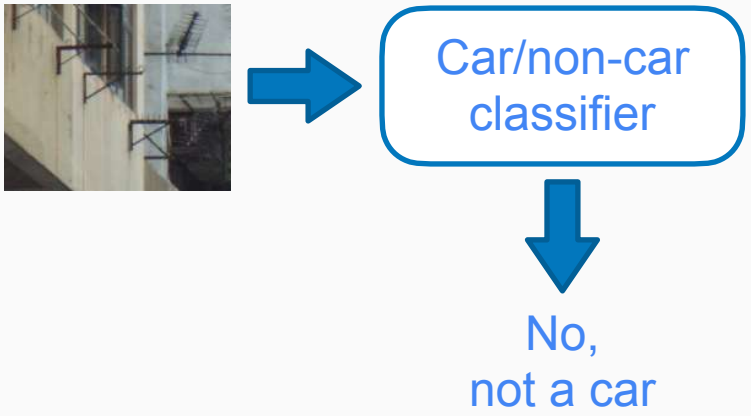
Problem of background clutter

- Use sub window:
  - At correct position, no clutter is present
  - Slide window to detect objects
  - Change size of the window to search over scales



Detection by classification

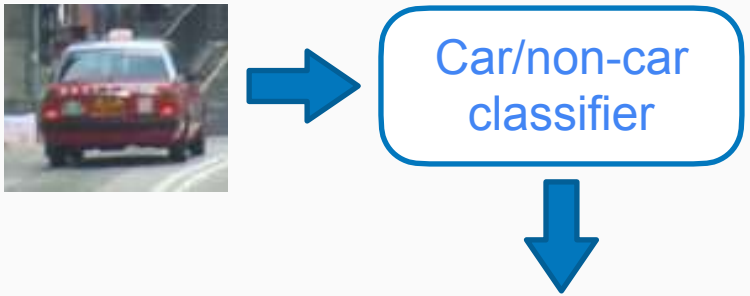
- Basic component: binary classifier



Detection by classification

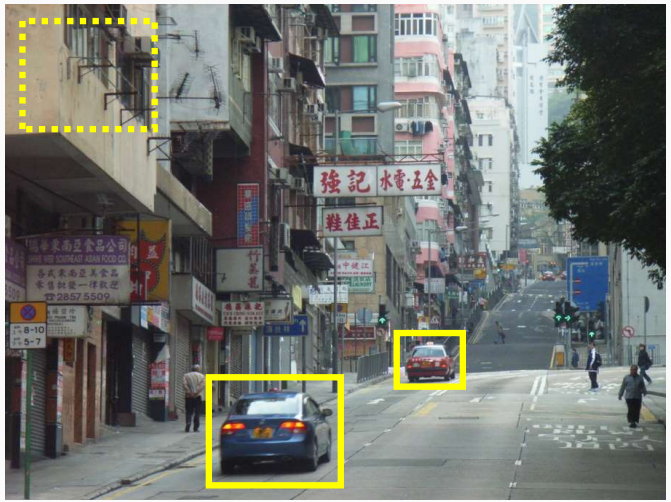
- Detect objects in clutter by **search**

Sliding window: exhaustive search over position and scale

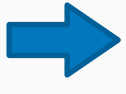


Detection by classification

- Detect objects in clutter by **search**



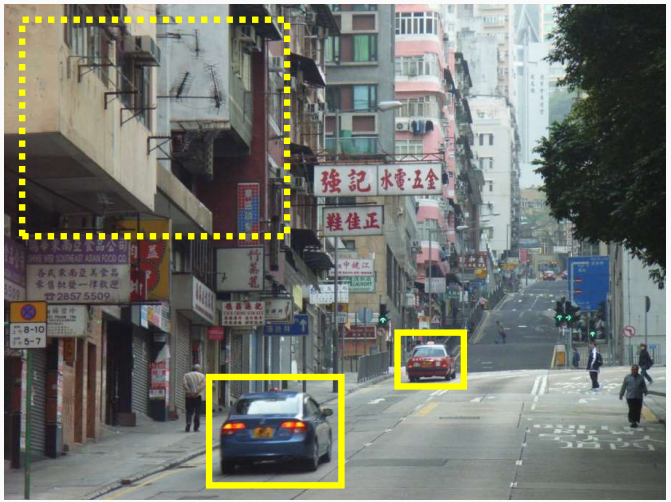
Sliding window: exhaustive search over position and scale



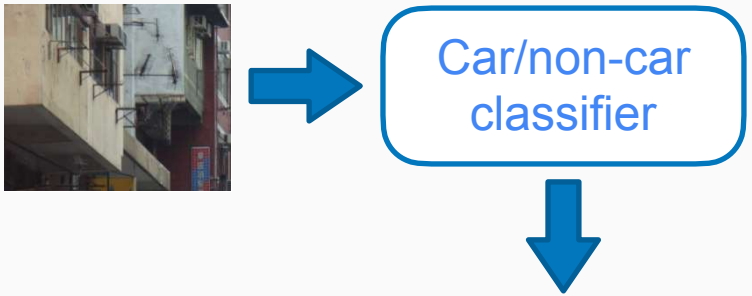
Car/non-car classifier

Detection by classification

- Detect objects in clutter by **search**



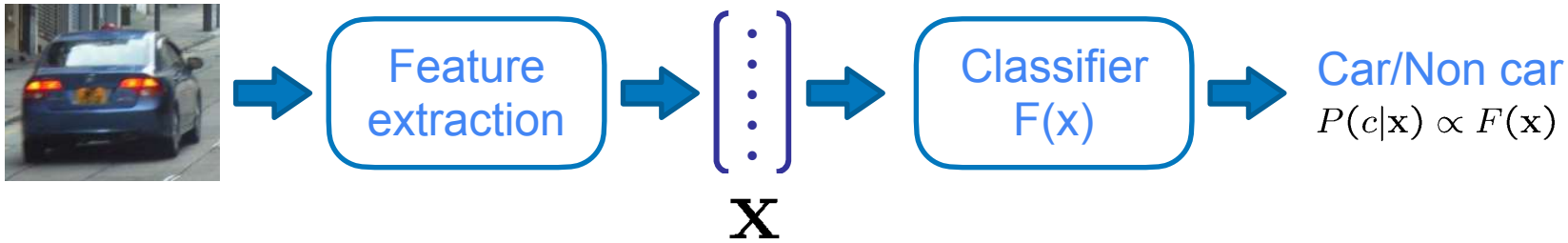
Sliding window: exhaustive search over position and scale



In practice one can use same window size over spatial pyramid

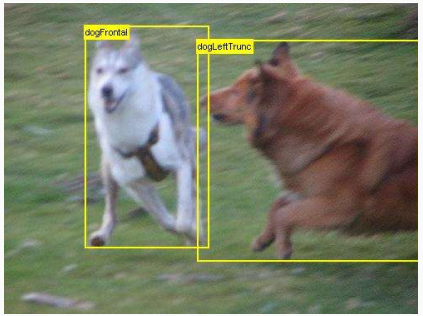
Window (image) classification

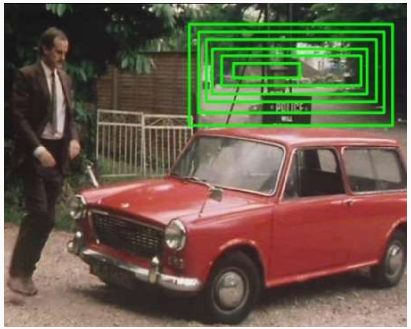
- Features usually engineered
- Classifier learned from data



Problems with sliding windows

- Aspect ratio
- Granularity (finite grid)
- Partial occlusions
- Multiple responses



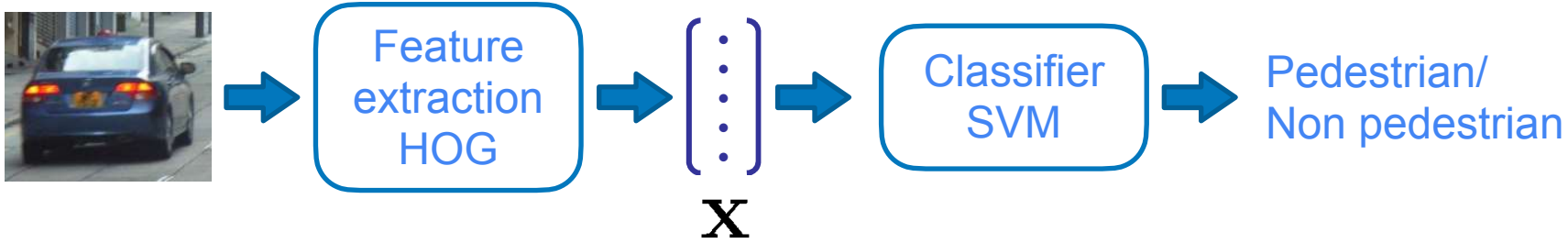


Example: pedestrian detection

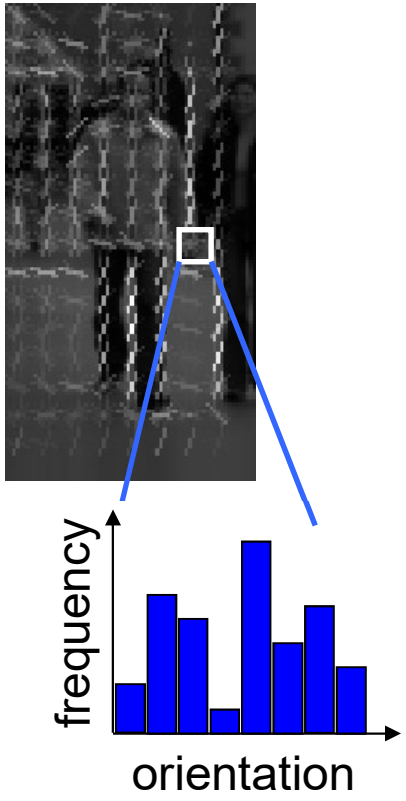
Dalal & Triggs CVPR 2005, pedestrian detection

- Objective: detect (localize) standing humans in an image
- Sliding window classifier
- Train a binary SVM classifier on whether a window contains a standing person or not
- **Histogram of Oriented Gradients (HOG) feature**
- Although introduced for pedestrian detection, it has been very successfully used with many other categories

Window (image) classification



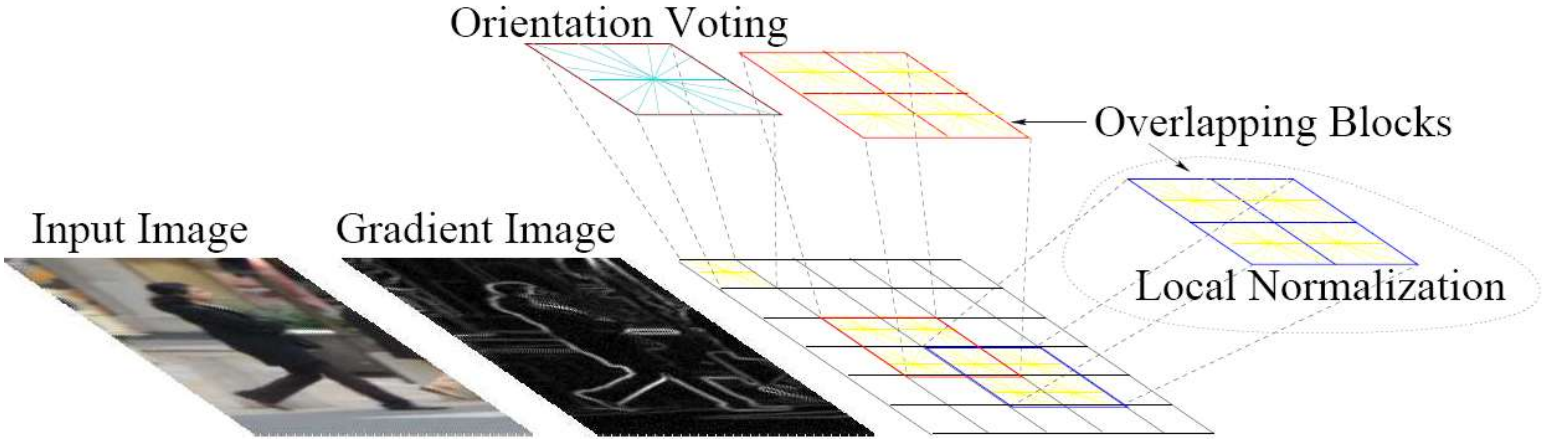
Feature: Histogram of Oriented Gradients (HOG)



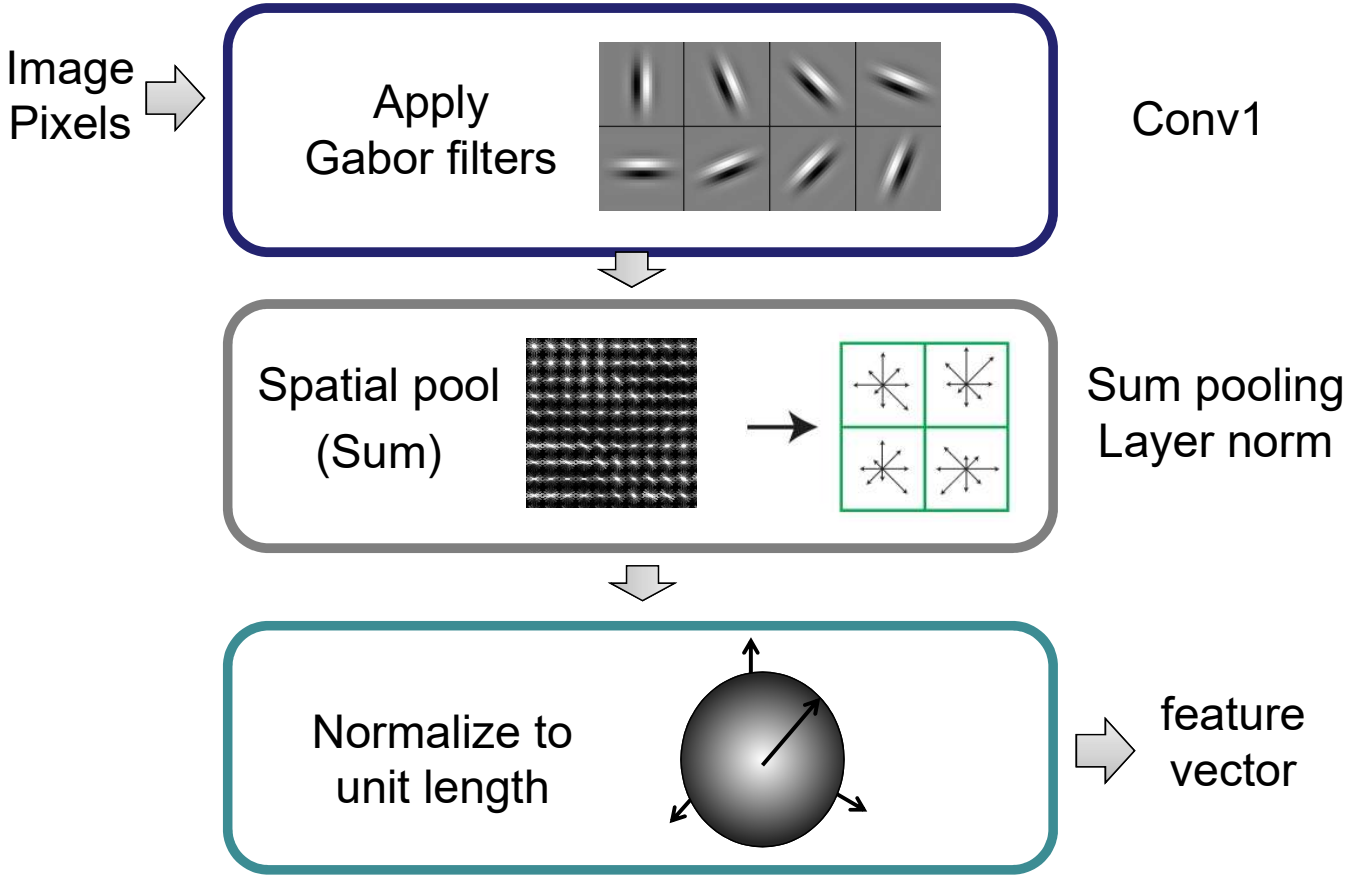
- Tile 64x128 pixel window into 8x8 pixel cells
- Calculate gradient image
- Each cell represented by histogram over 8 orientation bins (i.e. angles in range 0-180)

Feature: Histogram of Oriented Gradients (HOG)

- Adds a second level of overlapping spatial bins re-normalizing orientation histogram over larger spatial area
- Feature dimensions (approx) = 16 x 8 (for tiling) x 8 (orientations) x 4 (for blocks) = 4096

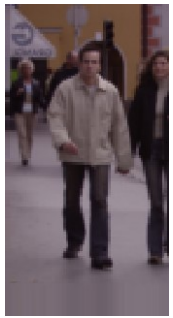


HOG descriptor similarity to CNN layers



Window (image) classification

- HOG features
- Linear SVM classifier



Feature extraction

$\mathbf{x}$

Classifier F(x)

Pedestrian/
Non pedestrian

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

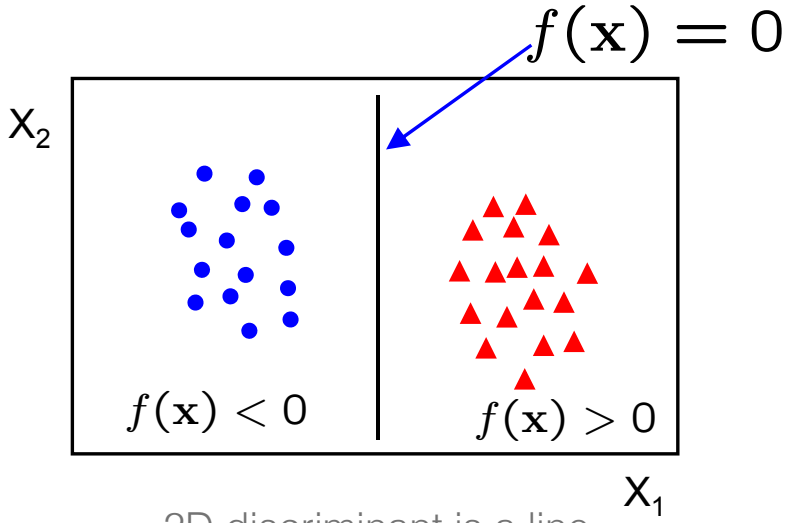
Reminder: Linear classifier

- A linear classifier has the form

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

- Learn such weights W and bias b that:

$$f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$



2D discriminant is a line

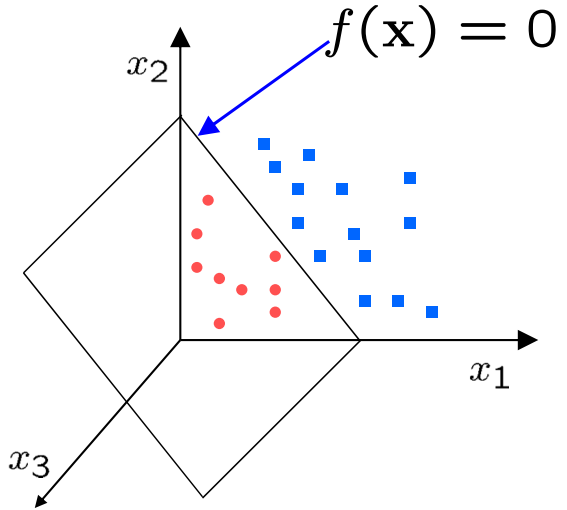
Reminder: Linear classifier

- A linear classifier has the form

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

- Learn such weights W and bias b that:

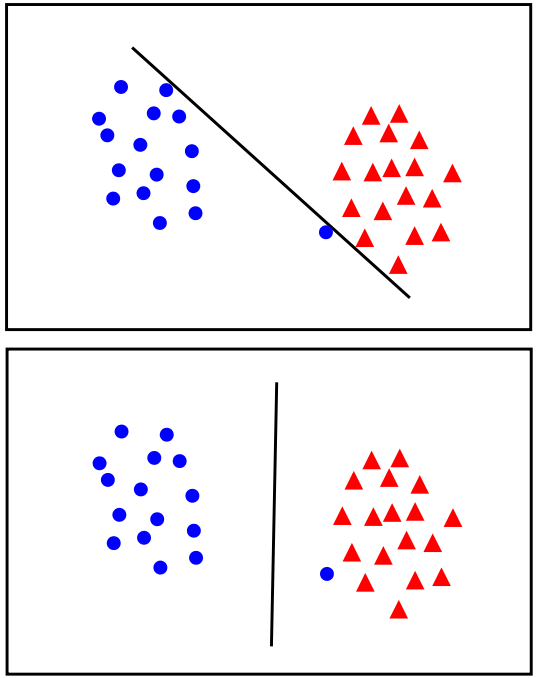
$$f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$



3D discriminant is a plane

Reminder: Linear separability again

- The points might be linearly separated but with very narrow margin
- The large margin solution might be better, even constraints are violated
- In general there is a trade off between the margin and the number of mistakes on the training data

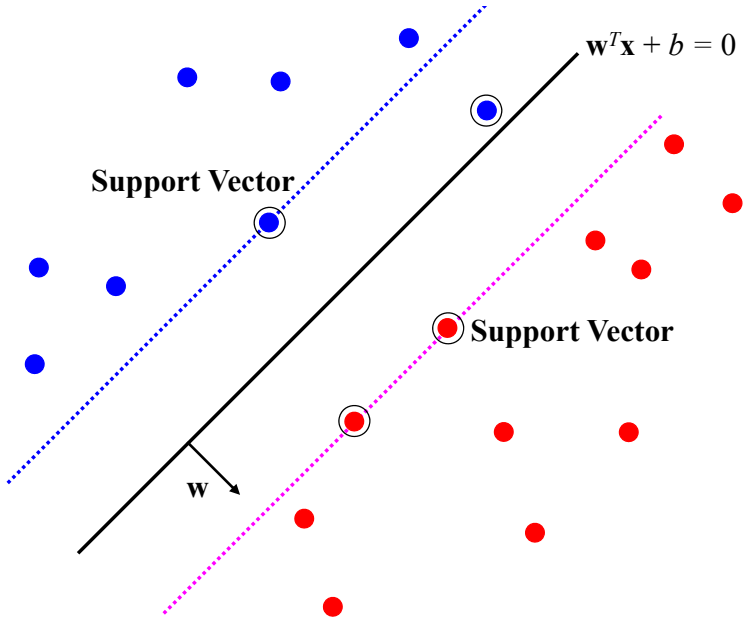


Support Vector Machine (SVM)

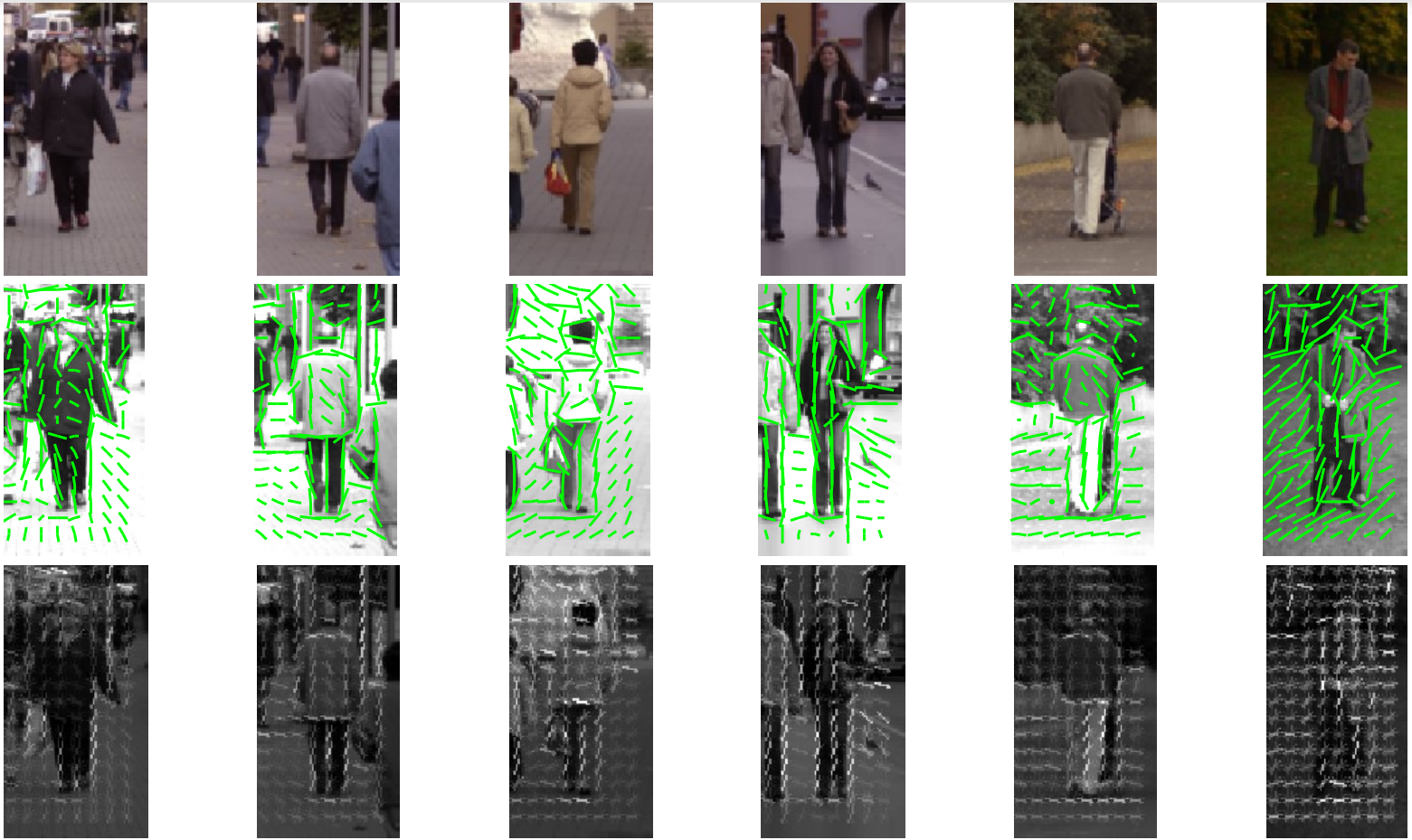
- Find good trade off between the classification margin and the misclassified training points.

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

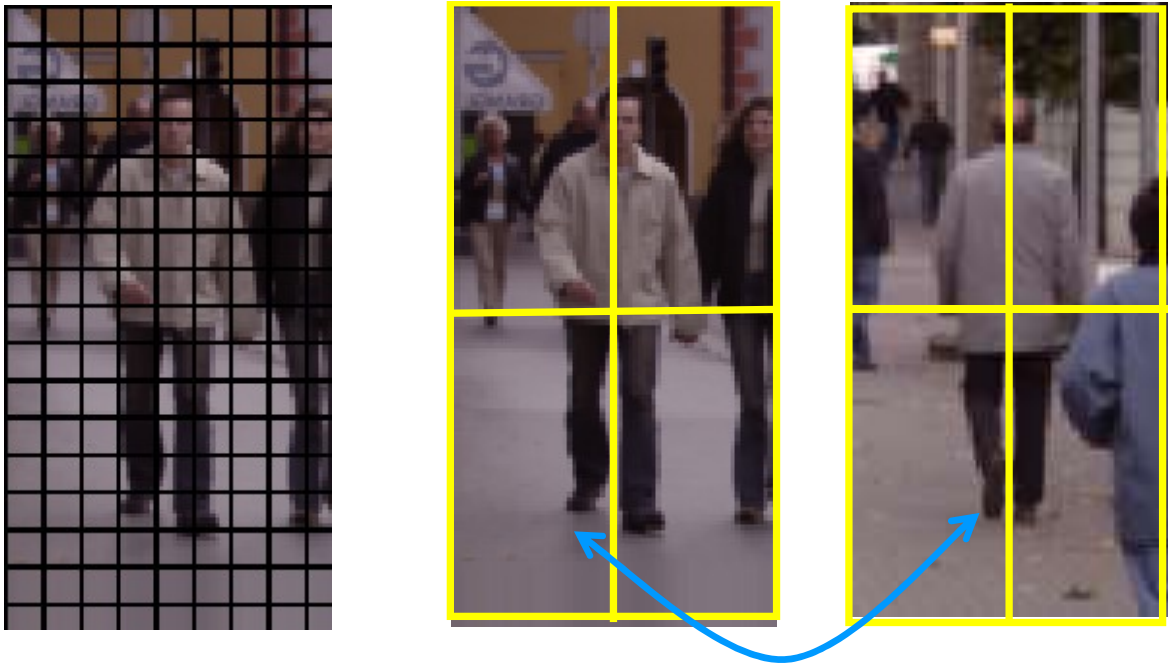
$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{w}||^2 + C\sum_i^N \max(0, 1 - y_i f(\mathbf{x}_i))$$



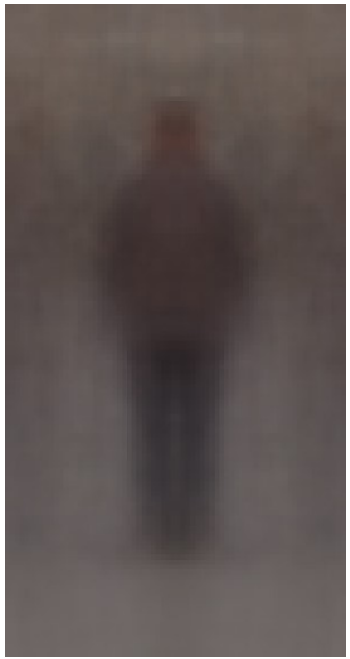
Examples

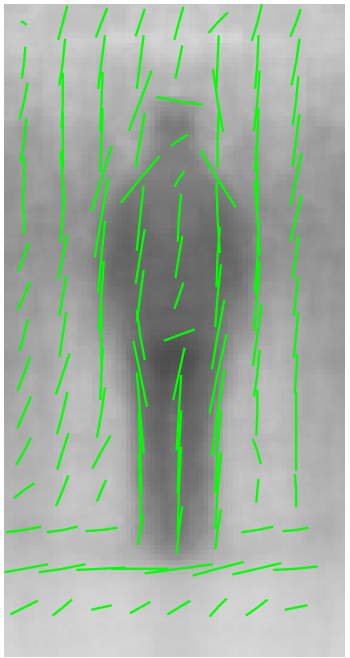


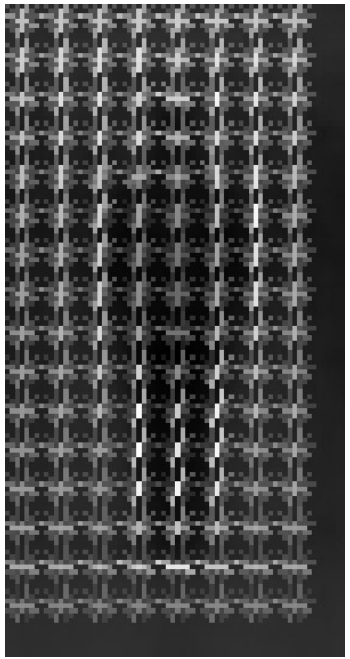
Tiling defines the spatial correspondence



Averaged examples



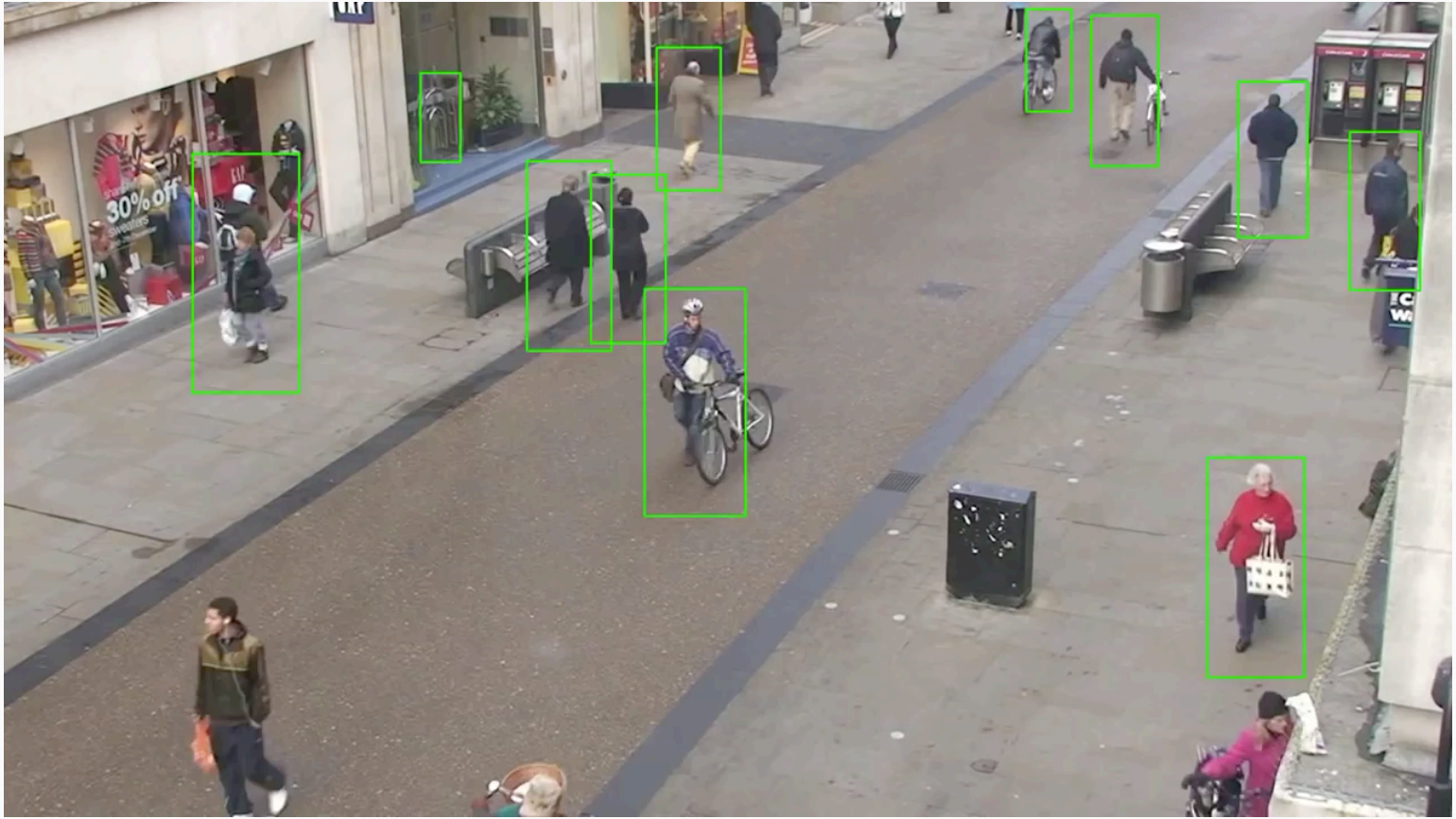




Example

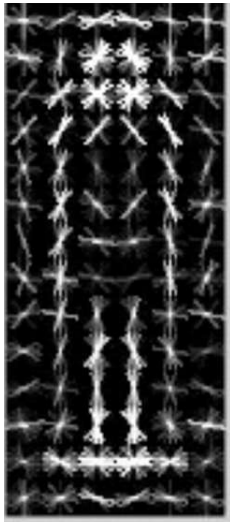
Dalal & Triggs 2005

Video example

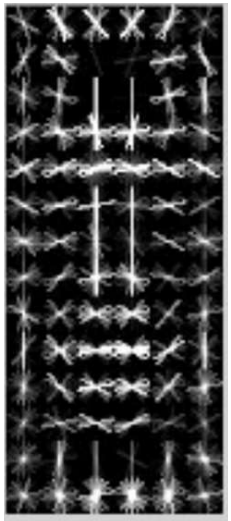


https://www.youtube.com/watch?v=b_DSTuxdGDw

Learned model

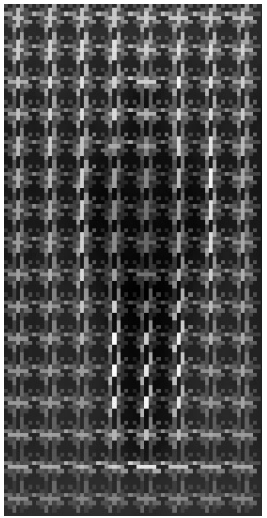


Positive weights



Negative weights

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



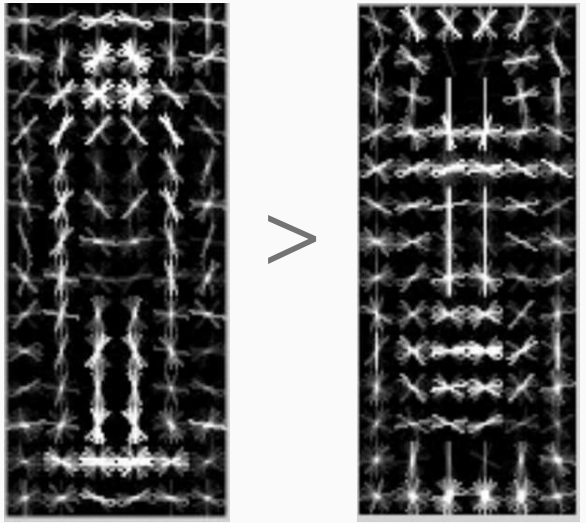
Average over positive training data

What do negative weights mean?

- Complete system would compete pedestrian/pillar/doorway models
- Discriminative models come with own background model

- Avoid detections on doorways by penalising vertical edges

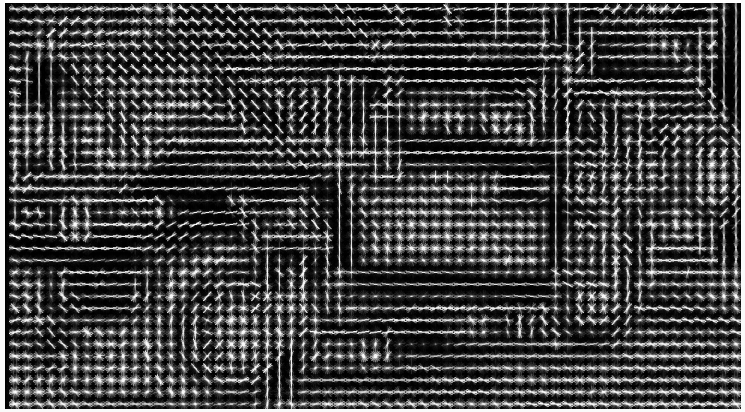
$$wx > 0$$
$$(w_+ - w_-)x > 0$$
$$w_+x > w_-x$$



Pedestrian model

Pedestrian **background** model

What is represented by HOG?



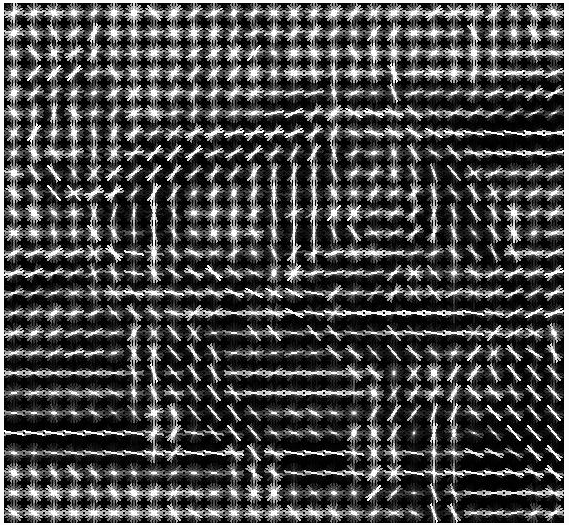
HOG

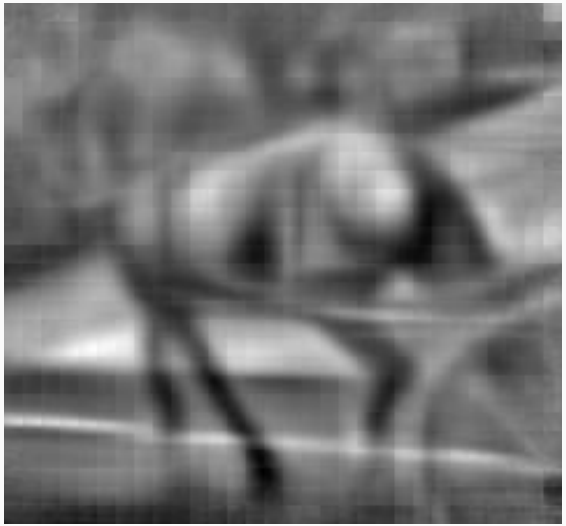
Inverse

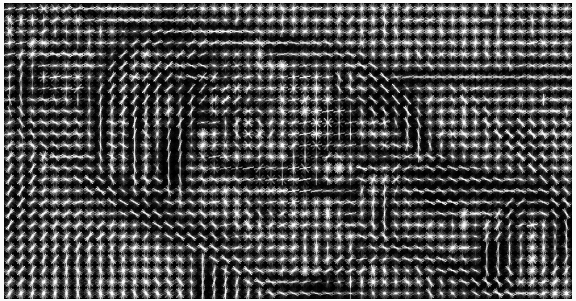
Original

Inverting and visualising feature for object detection
Carl Vondrick et al.
http://www.cs.columbia.edu/~vondrick/ihog/

What is represented by HOG?



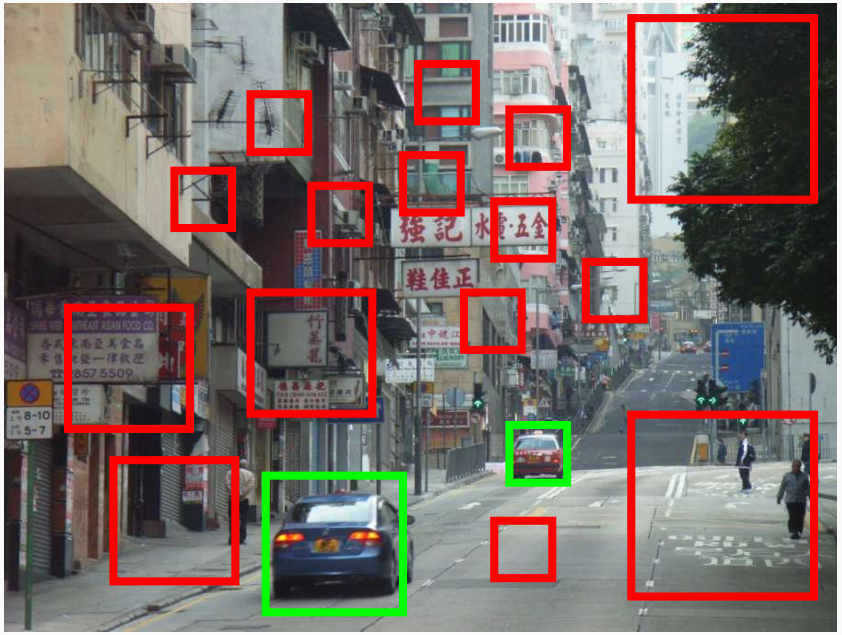




Training a sliding window detector

Training a sliding window detector

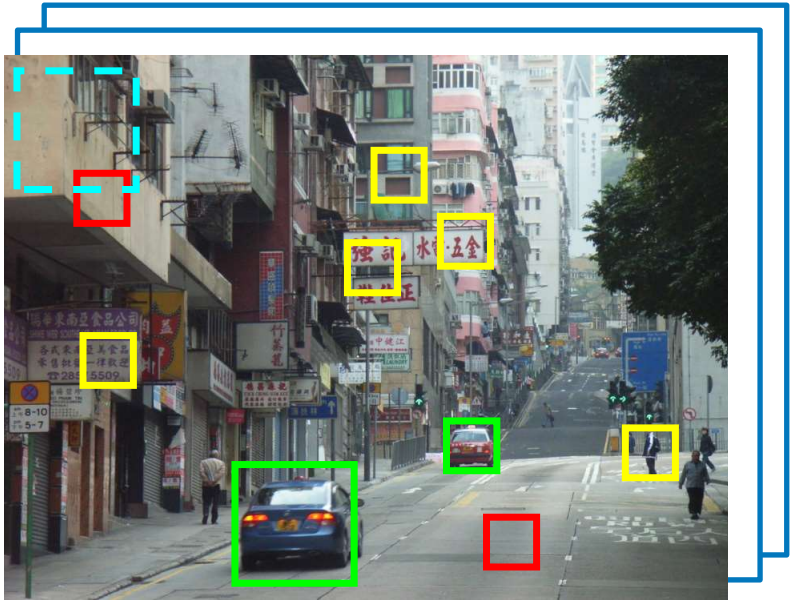
- Inherently asymmetric, i.e. much more “non-object” than “object
- Classifier needs to have very low false positive rate
- “Non-object” category is very complex and needs lots of data



Bootstrapping

1. Pick a negative training set at random
2. Train classifier
3. Run on training data
4. Add false positives to training set
5. Repeat from step 2

- Results in finite but diverse training set
- Forces classifier to use hard negatives



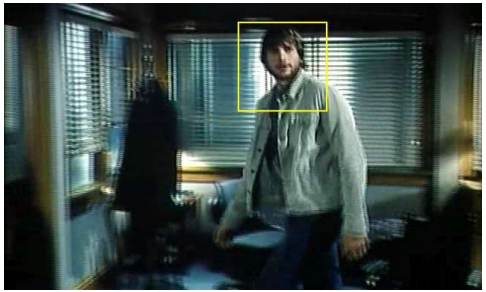
Example: upper body detector

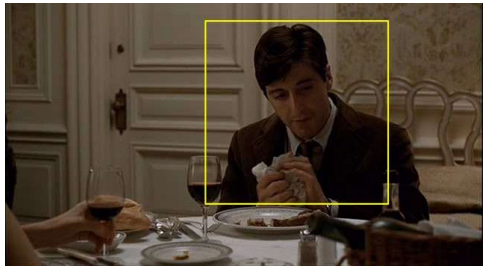
- Training data for classifier training and validation
  - Hollywood2 dataset (33 movies) - 1122 video frames with upper bodies marked
- First stage training (bootstrapping)
  - 1607 upper body annotations jittered to 32k positive examples
  - 55k negatives sampled from the same set of frames
- Second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

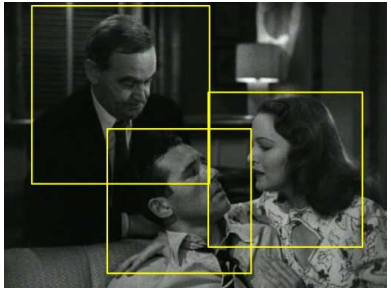
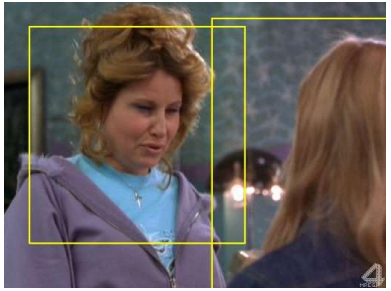


Credits: A Zisserman

Training data - positive annotations

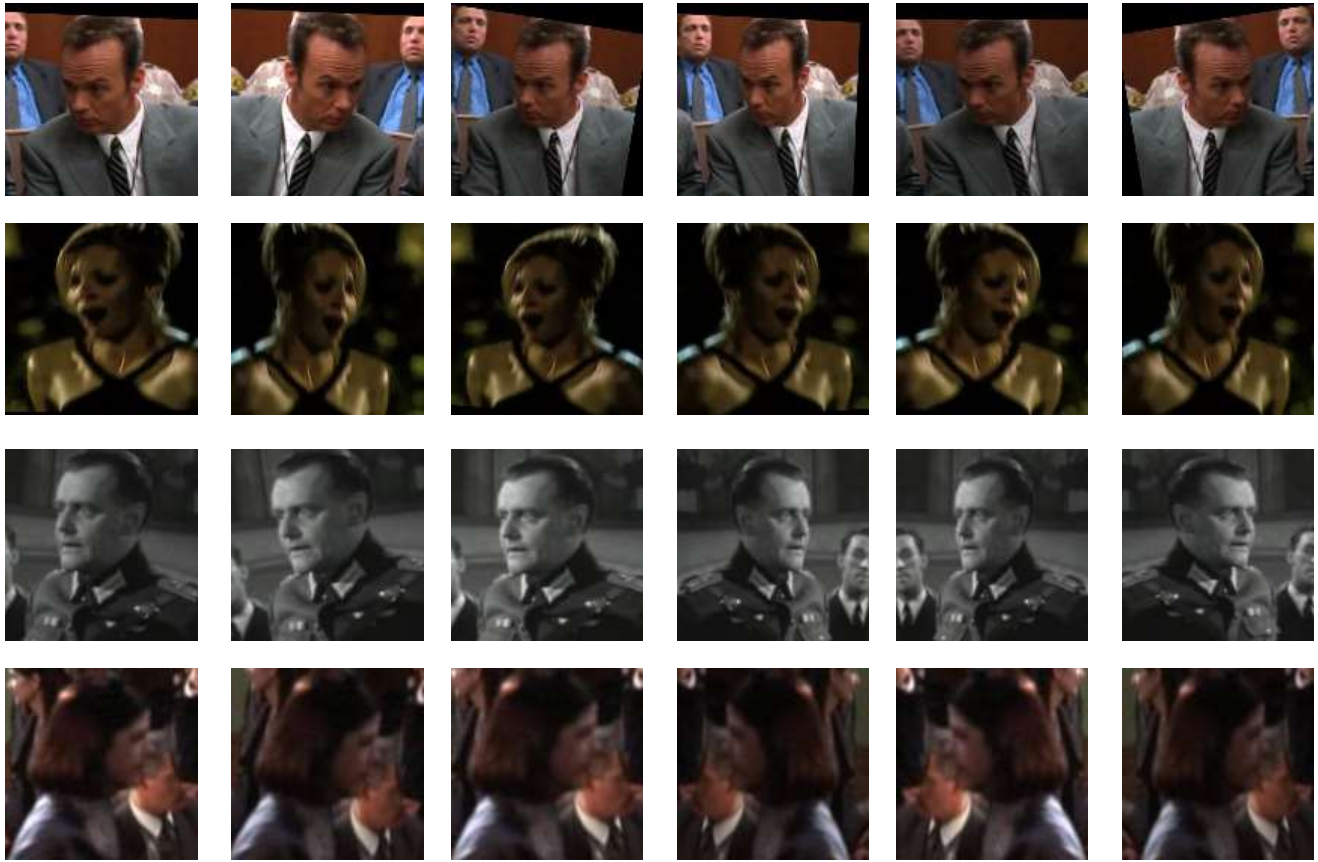



Positive windows

Note:
common size
and alignment

Jittered positive



Credits: A Zisserman

Jittered positive

Credits: A Zisserman

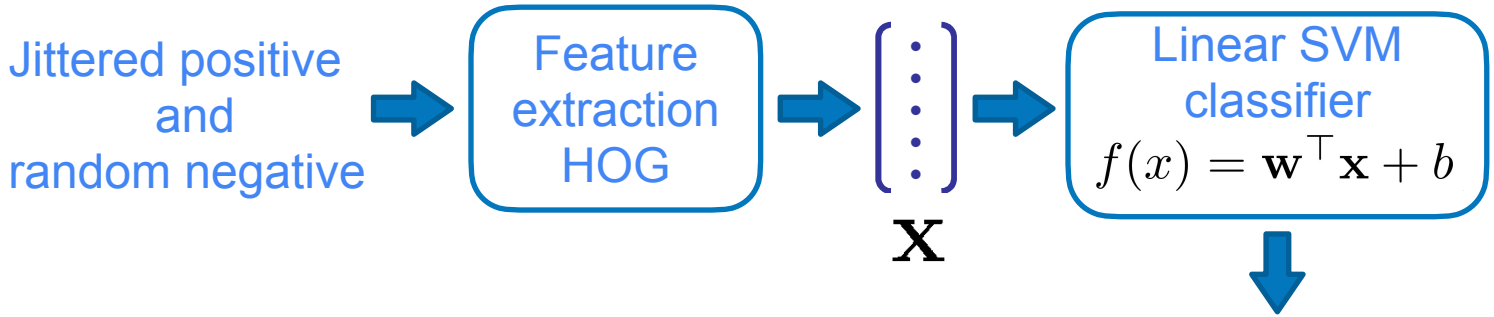
Random negatives

Credits: A Zisserman

Random negatives

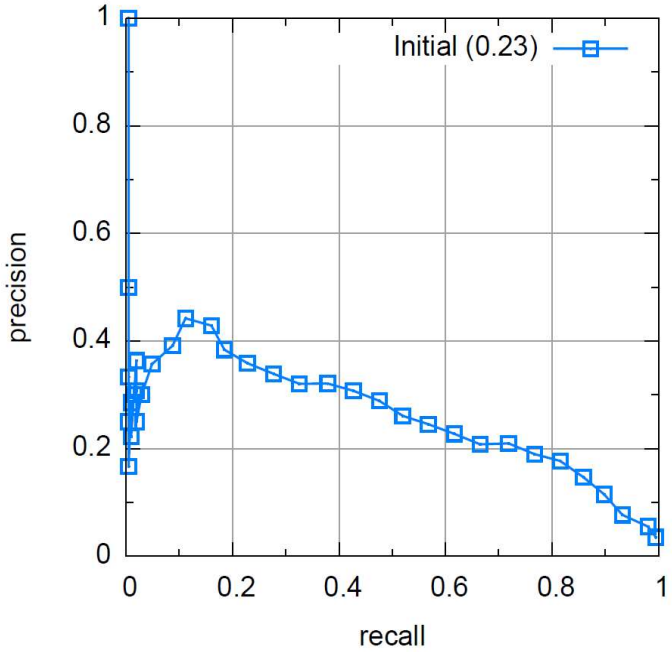
Credits: A Zisserman

Window (image) first stage classifier



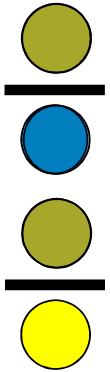
Credits: A Zisserman

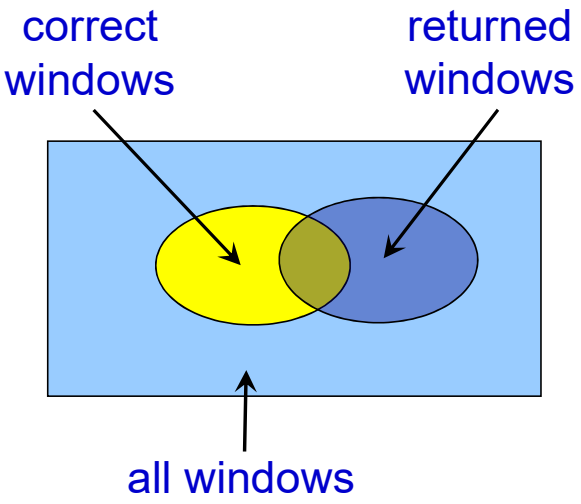
First stage performance on validation set

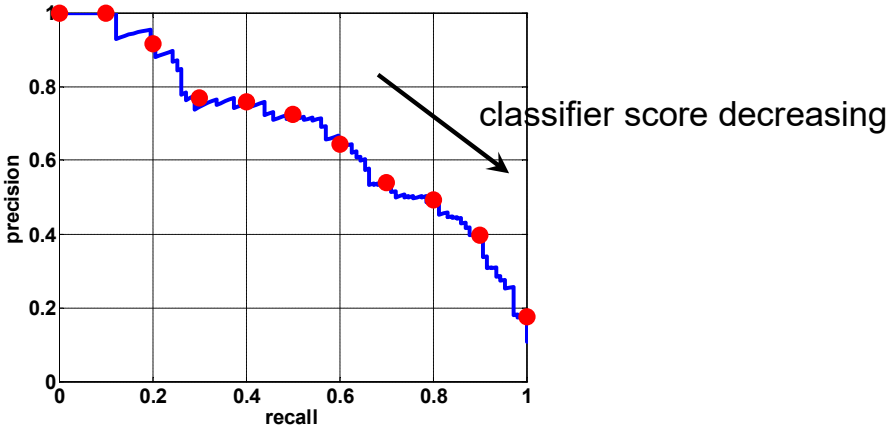


Reminder: Precision - Recall curve

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

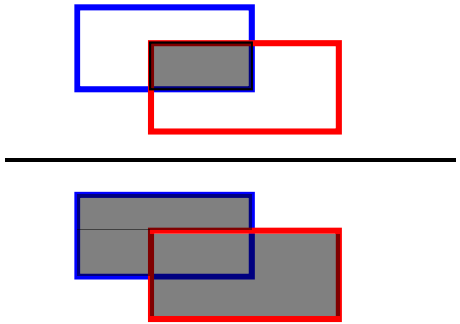




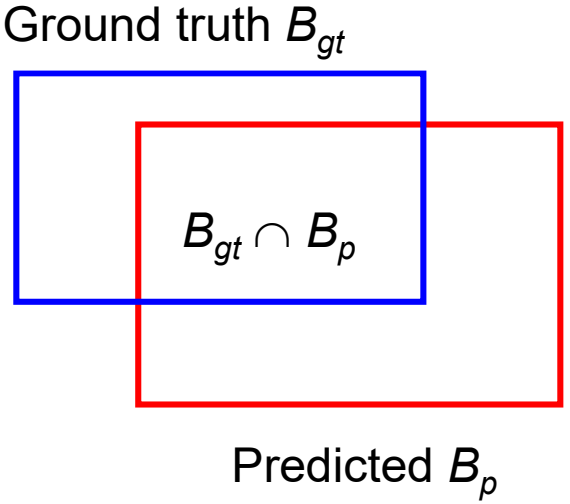


Evaluating the detected bounding boxes

- Area of overlap (AO) measure
- Correct detection if intersection over union larger than threshold

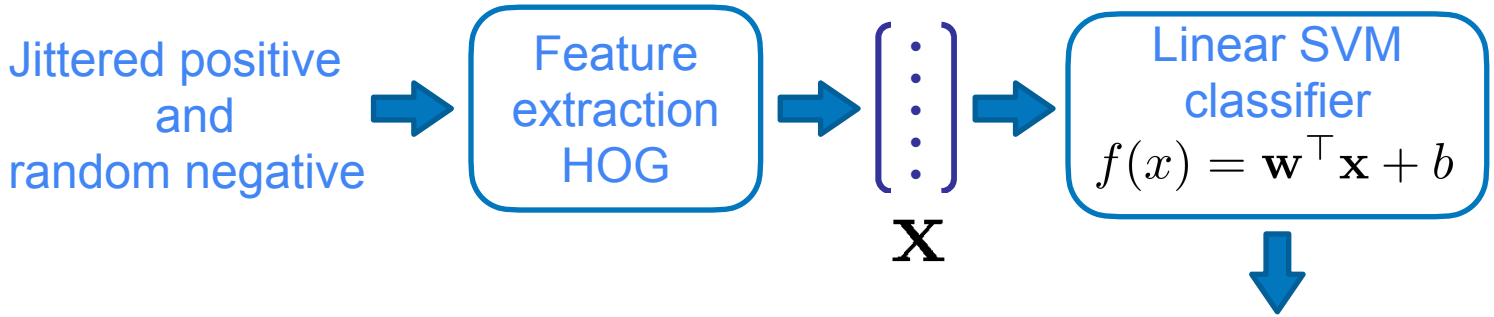


> threshold (50 %)



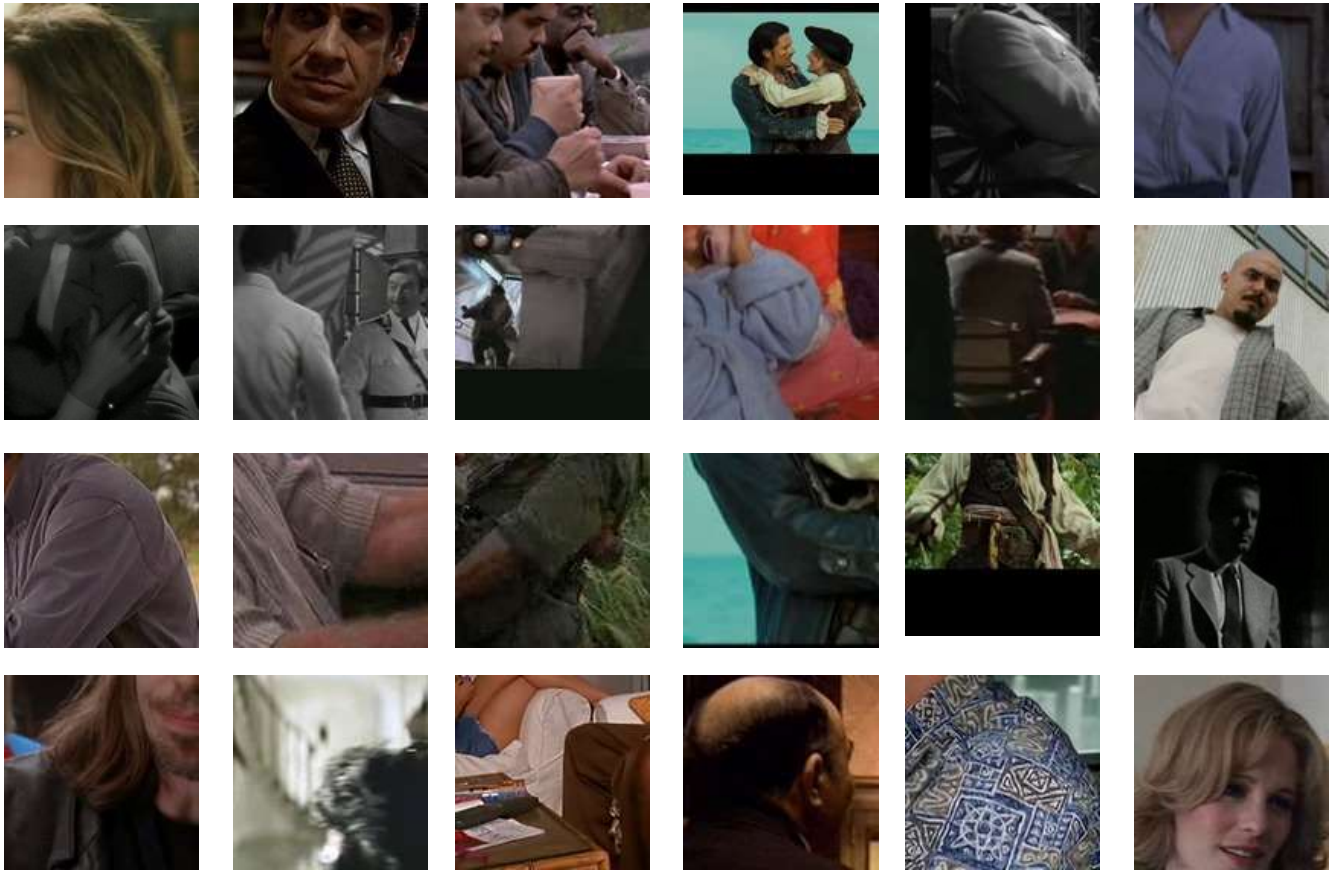
$$AO(B_{gt}, B_p) = \frac{|B_{gt} \bigcap B_p|}{|B_{gt} \bigcup B_p|}$$

Window (image) first stage classifier



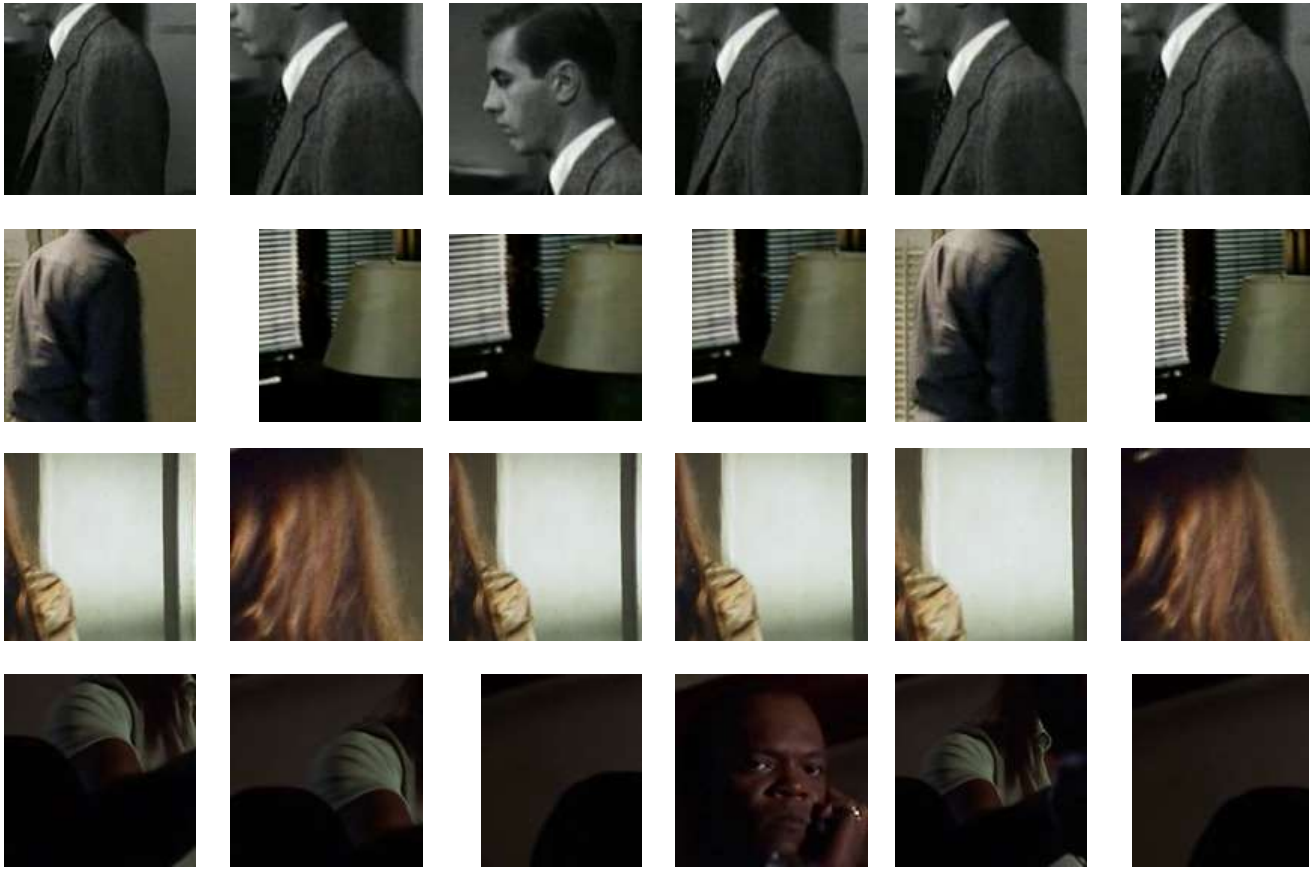
- Find high scoring false positive detections
- Use them as hard negatives for next training round
- Cost = # training image x inference time per image

Hard negatives



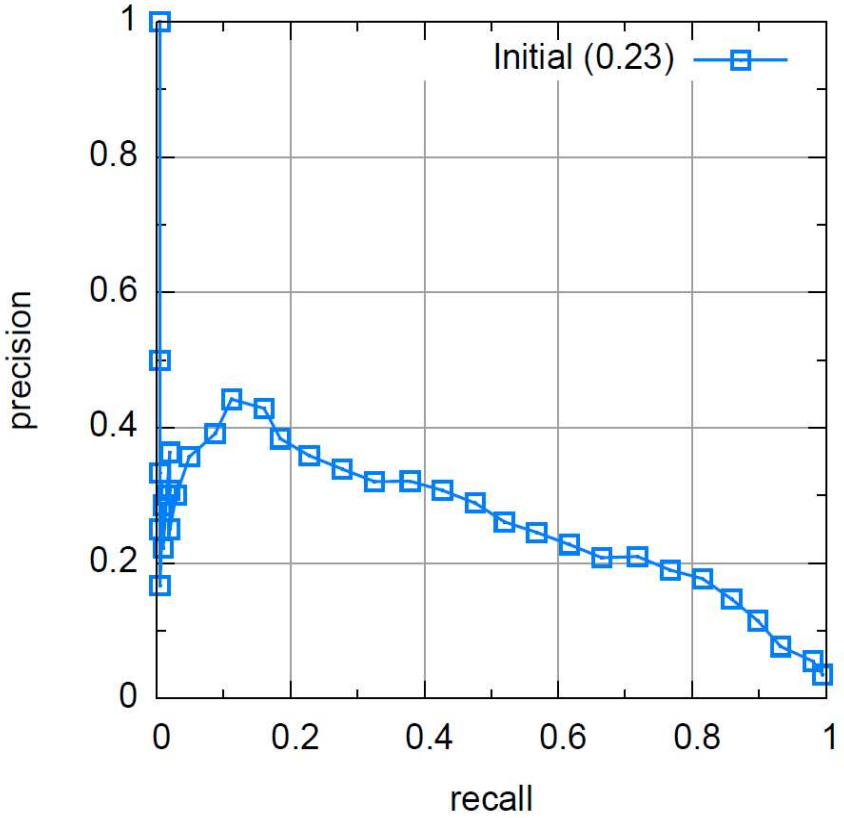
Credits: A Zisserman

Hard negatives

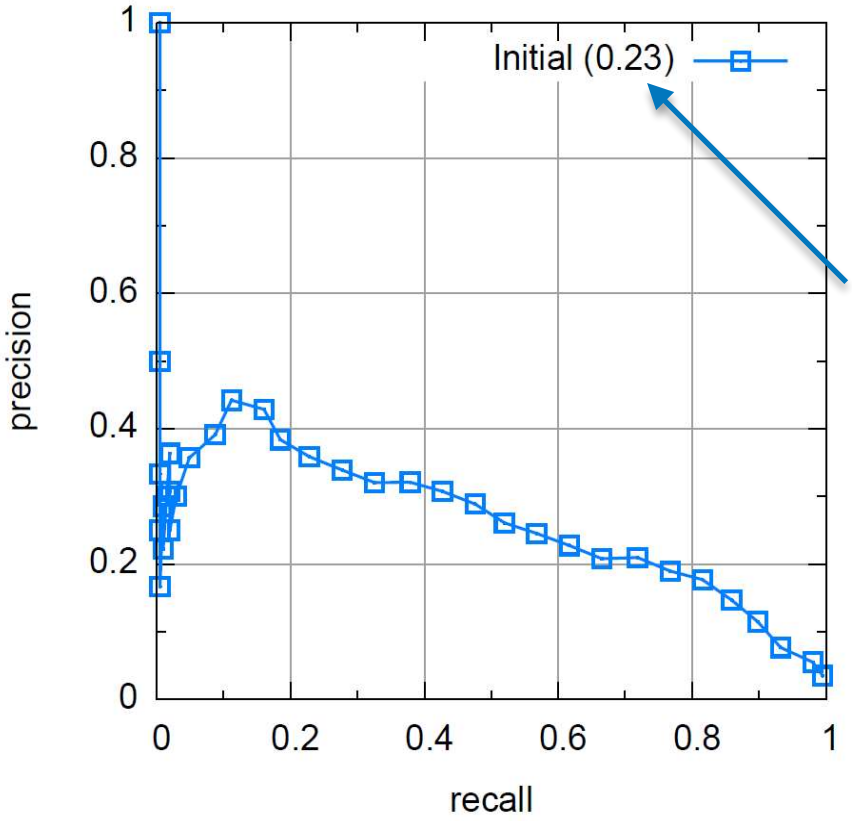


Credits: A Zisserman

First stage performance on validation set

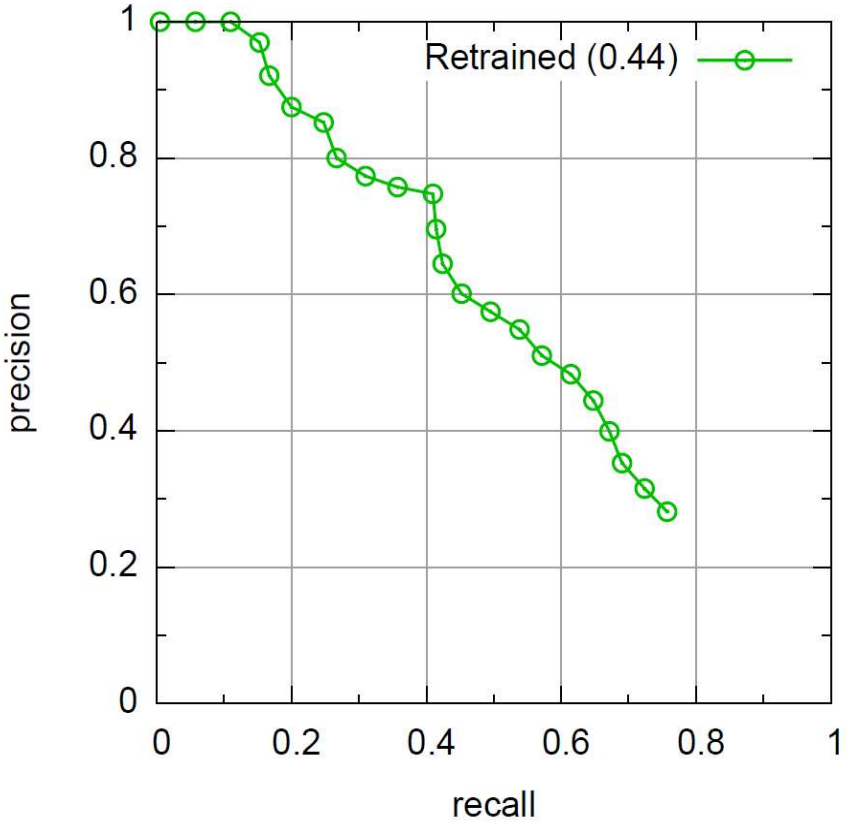


First stage performance on validation set

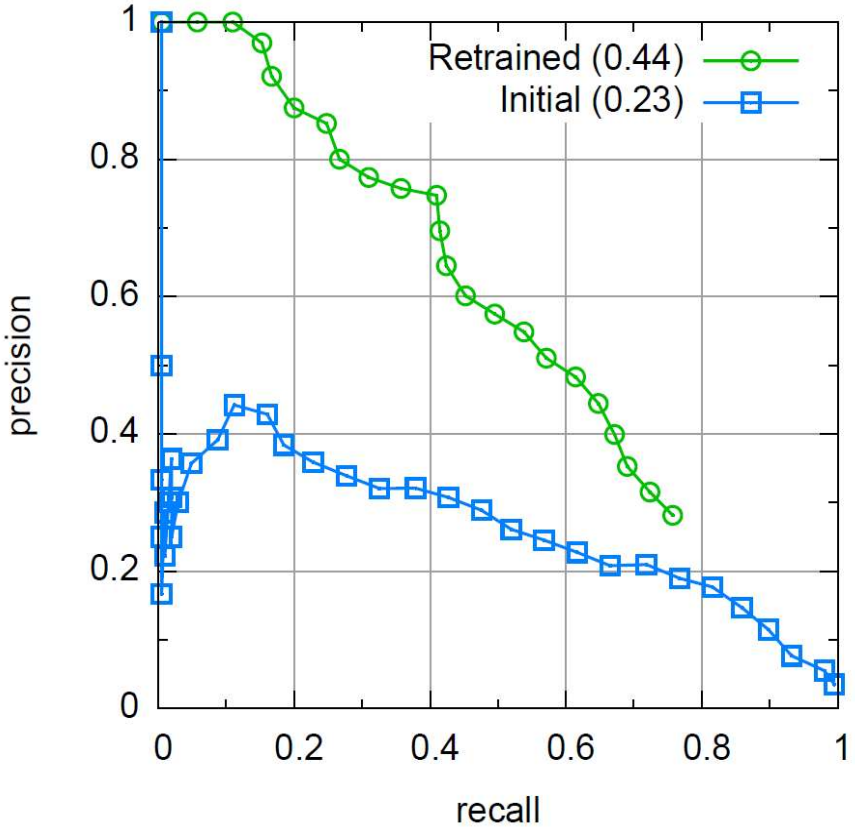


Average precision
(area under PR curve)

Performance after re-training

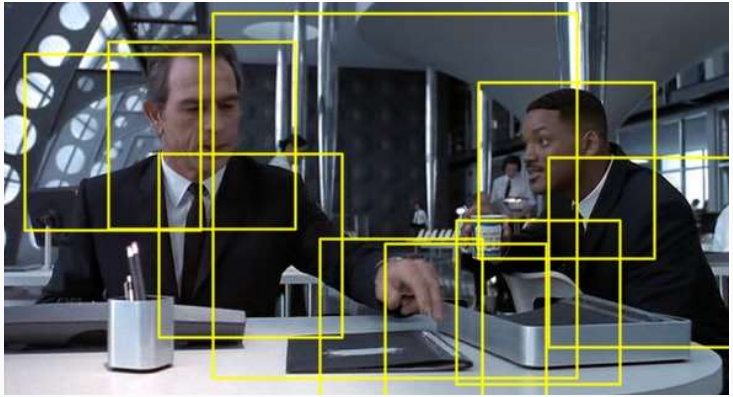


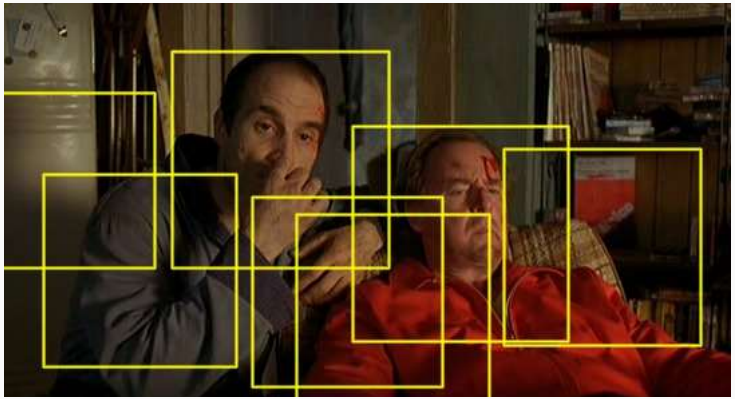
The effect of the re-training



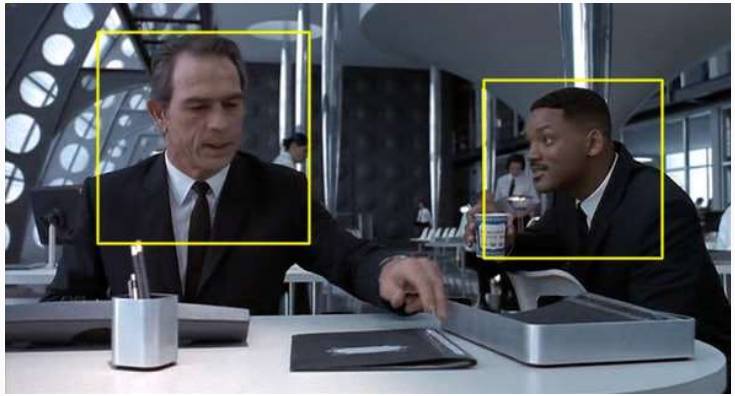
Side by side comparison

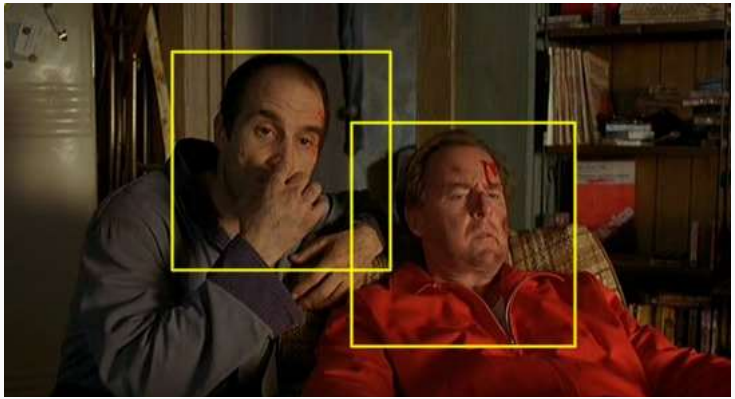
Before re-training





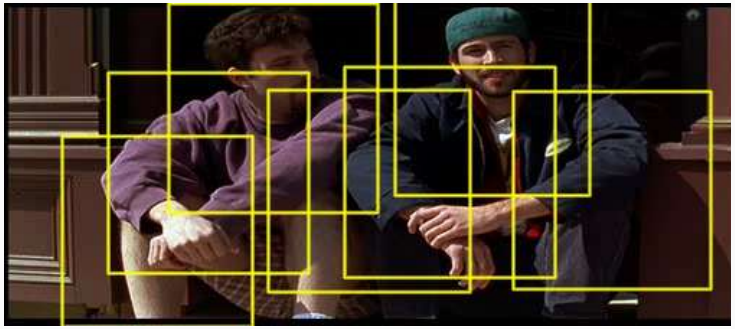
After re-training

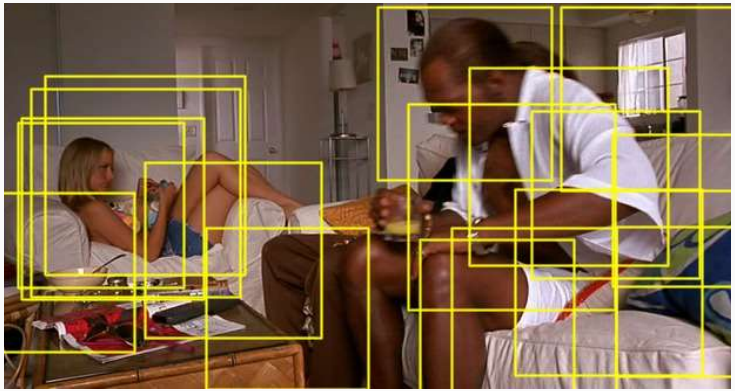




Side by side comparison

Before re-training

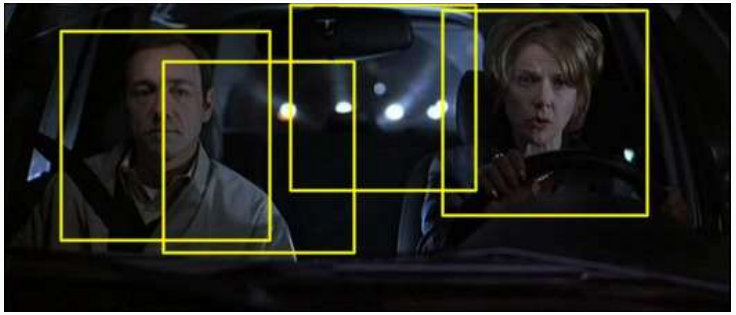


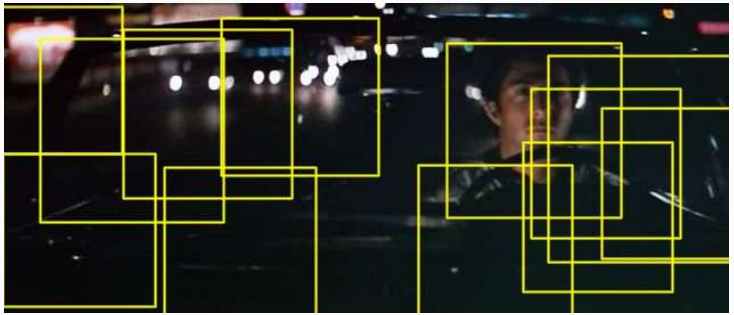


After re-training

Side by side comparison

Before re-training



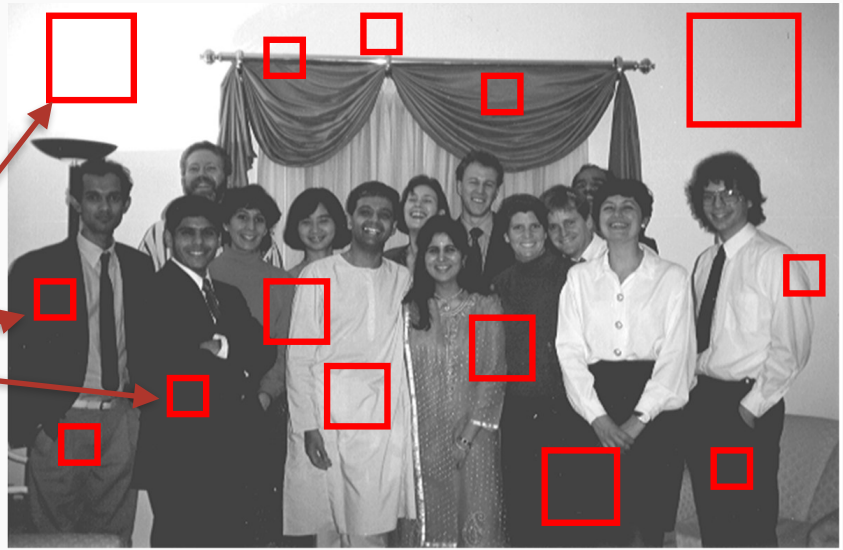


After re-training

Accelerating sliding window search

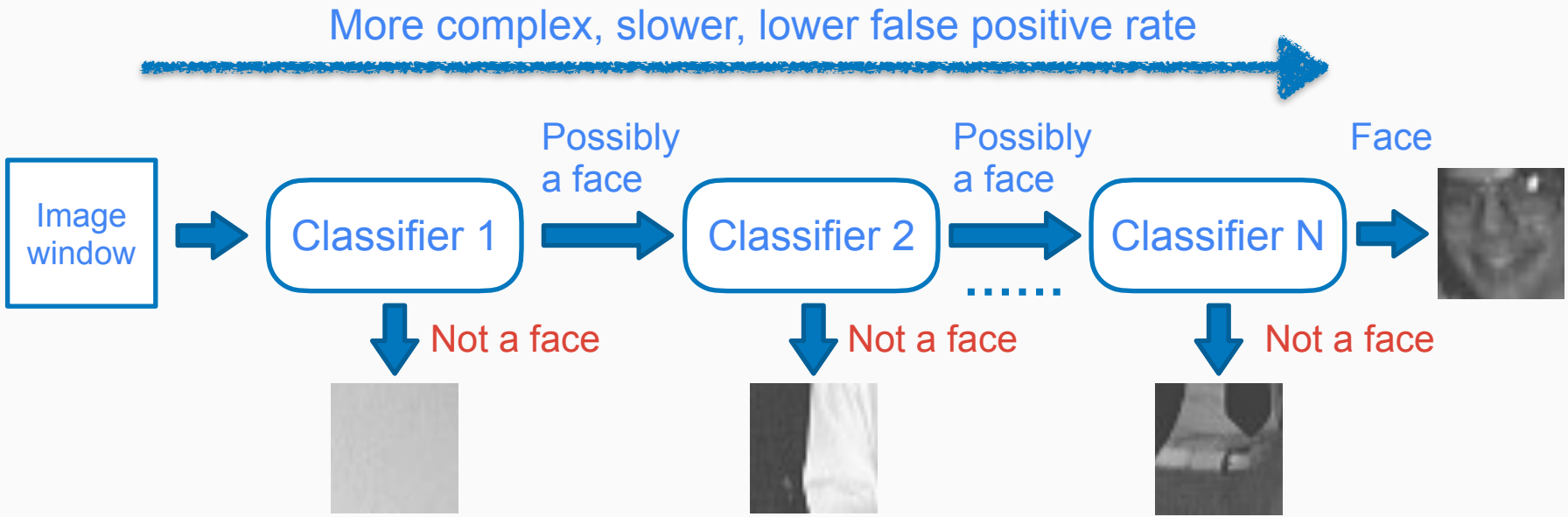
Accelerating sliding window search

- Sliding window search is slow since some many windows are needed
- m x n x scale = 100 000 windows for 320 x 240 image
- Most windows are clearly negative
- Is it possible to seed up the search?



Example: face detection

Cascaded classification



Reject easy non-objects using simpler and faster classifiers

Cascaded classification

- Slow and expensive classifiers only applied to a few windows
  -> significant speedup
- Controlling complexity vs. speed: number of features, number of parts..

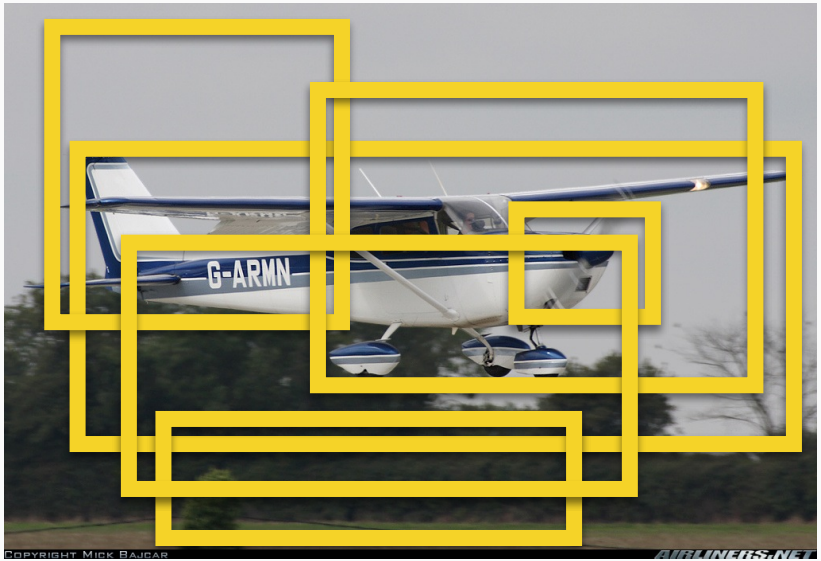
Detection proposals

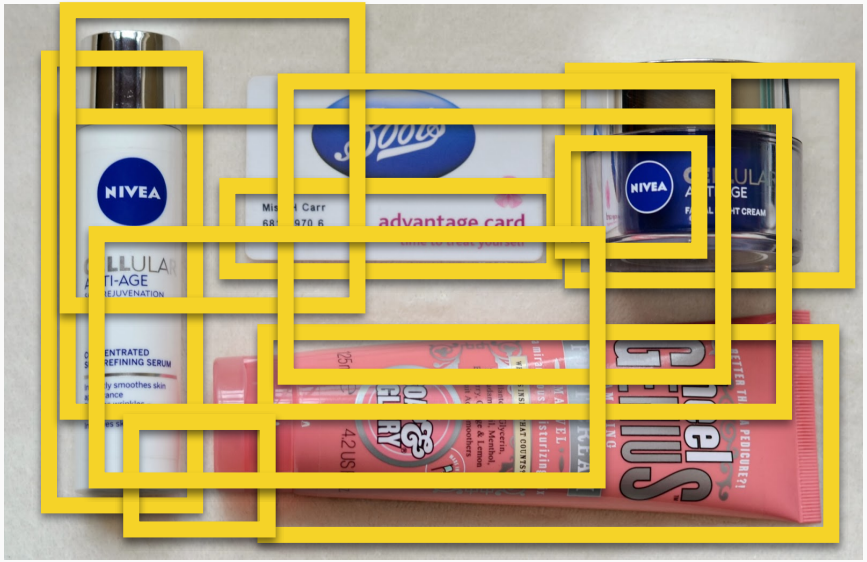
Propose image regions containing objects (rather than stuff)

Detection proposals

Propose image regions containing objects (rather than stuff)

Detection proposals





Aim to cover all objects with small amount of proposals e.g. 100-1000

Detection proposals

Can be boxes or segmentations, but must be class agnostic

Detection proposals: example method 1

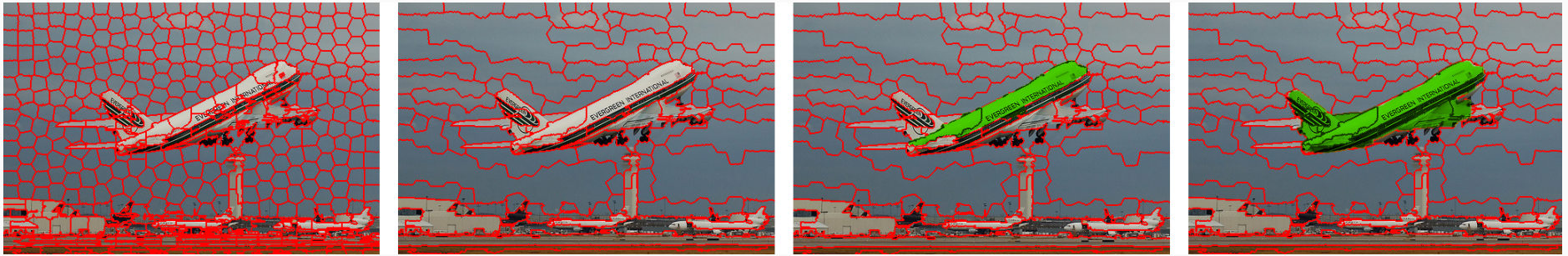
- Hierarchical segmentation: start with small and merge using cues
  Produces roughly 2000 regions per image with 95% of hitting relevant objects

Selective search for object recognition, Ujilings et al. IJCV 2013

Hierarchical merging

Video credits: Karen Simonyan

Detection proposals: example method 2



Detecting object segmentation proposals using global and local search, Rantalankila, Kannala, Rahtu, 2014

Detection proposals: example method 2

Detecting object segmentation proposals using global and local search, Rantalankila, Kannala, Rahtu, 2014

Detection proposals: example method 2

Detecting object segmentation proposals using global and local search, Rantalankila, Kannala, Rahtu, 2014

Things to remember

- Detection by sliding window classification
- Multiple scales (and aspect ratios) to detect objects of different size
- Importance of hard negative mining (due to class imbalance)
- Cascaded detectors
- Speed up training and inference by selecting sub-set of windows only

That’s it folks!