Finding People in Images and Videos

Navneet DALAL

GRAVIR, INRIA Rhône-Alpes

Thesis Advisors

Cordelia SCHMID et Bill TRIGGS

17 July, 2006

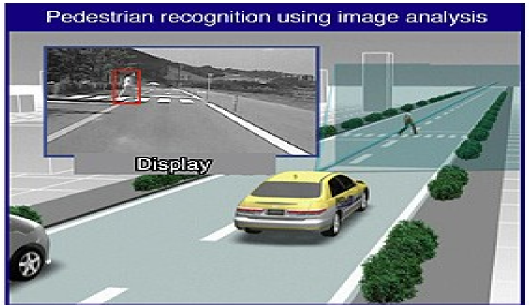
Institut National Polytechnique de Grenoble

Goals & Applications

Goal: Detect and localise people in images and videos

Applications:

- Images, films & multi-media analysis
- Pedestrian detection for smart cars
- Visual surveillance, behavior analysis



Difficulties

Wide variety of articulated poses

Variable appearance and clothing

Complex backgrounds

Unconstrained illumination

Occlusions, different scales

Videos sequences involves motion of the subject, the camera and the objects in the background

Main assumption: upright fully visible people

Talk Outline

- Overview of detection methodology
- Static images
  - Feature sets
  - Object localisation
  - Extension to other object classes
- Videos
  - Motion features
  - Optical flow estimation
- Part based person detection
- Conclusions and perspectives

Overview of Methodology

Detection Phase

Scan image(s) at all scales and locations

↓

Extract features over windows

↓

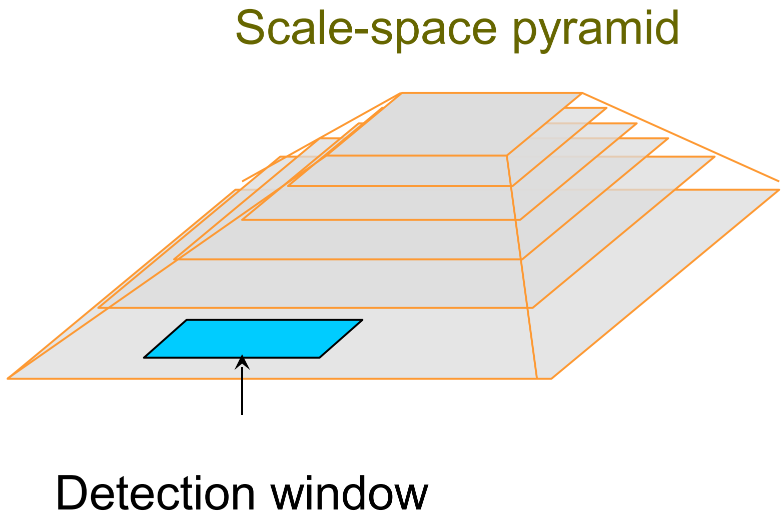
Run linear SVM classifier on all locations

↓

Fuse multiple detections in 3-D position & scale space

↓

Object detections with bounding boxes

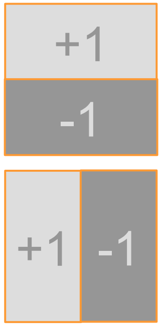


Focus on building robust feature sets (static & motion)

Finding People in Images

Existing Person Detectors/Feature Sets

Current Approaches

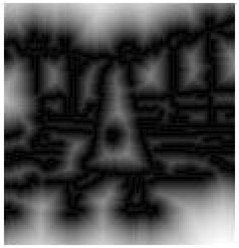


Haar wavelets + SVM:

- Papageorgiou & Poggio, 2000; Mohan et al 2000

Rectangular differential features + adaBoost:

- Viola & Jones, 2001



Edge templates + nearest neighbour:

- Gavrila & Philomen, 1999

Model based methods

- Felzenszwalb & Huttenlocher, 2000; Ioffe & Forsyth, 1999

Other works

- Leibe et al, 2005; Mikolajczyk et al, 2004

Orientation histograms

Freeman et al, 1996; Lowe, 1999 (SIFT); Belongie et al, 2002 (Shape contexts)

Static Feature Extraction

Input image

Detection window

Normalise gamma

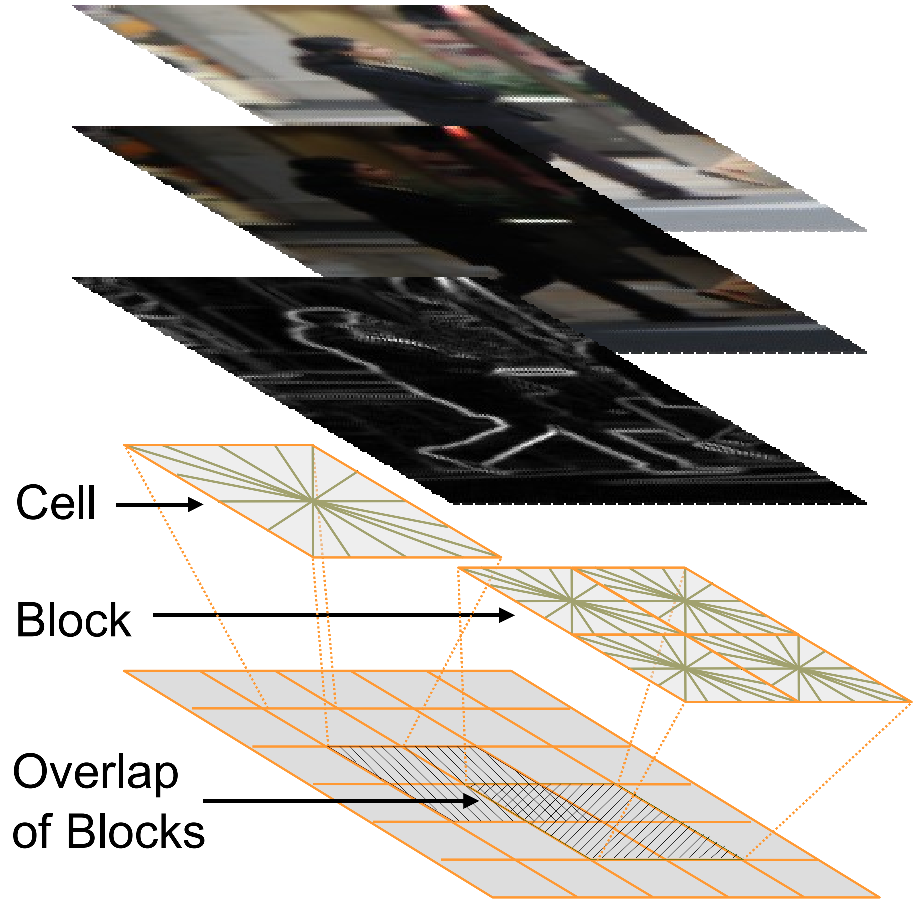
Compute gradients

Weighted vote in spatial & orientation cells

Contrast normalise over overlapping spatial cells

Collect HOGs over detection window

Linear SVM



Feature vector f = [..., ..., ...]

Overview of Learning Phase

Learning phase

Input: Annotations on training images

Create fixed-resolution normalised training image data set

Encode images into feature spaces

Learn binary classifier

Resample negative training images to create hard examples

Encode images into feature spaces

Learn binary classifier

Object/Non-object decision

Retraining reduces false positives by an order of magnitude!

HOG Descriptors

Parameters

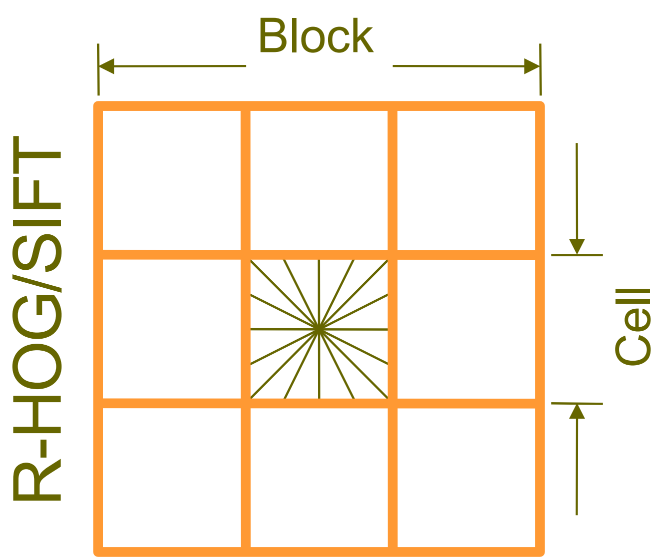
- Gradient scale
- Orientation bins
- Percentage of block overlap

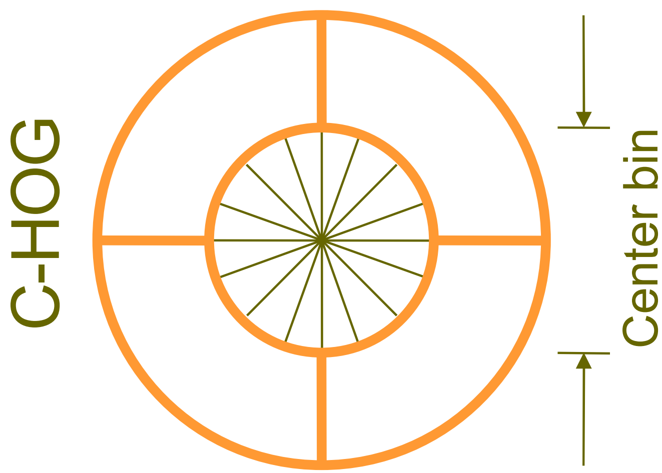
Schemes

- RGB or Lab, colour/gray-space
- Block normalisation
  - *L2*-norm, $v \leftarrow v / \sqrt{\|v\|_2^2 + \varepsilon}$

or

  - *L1*-norm, $v \leftarrow \sqrt{v / (\|v\|_1 + \varepsilon)}$



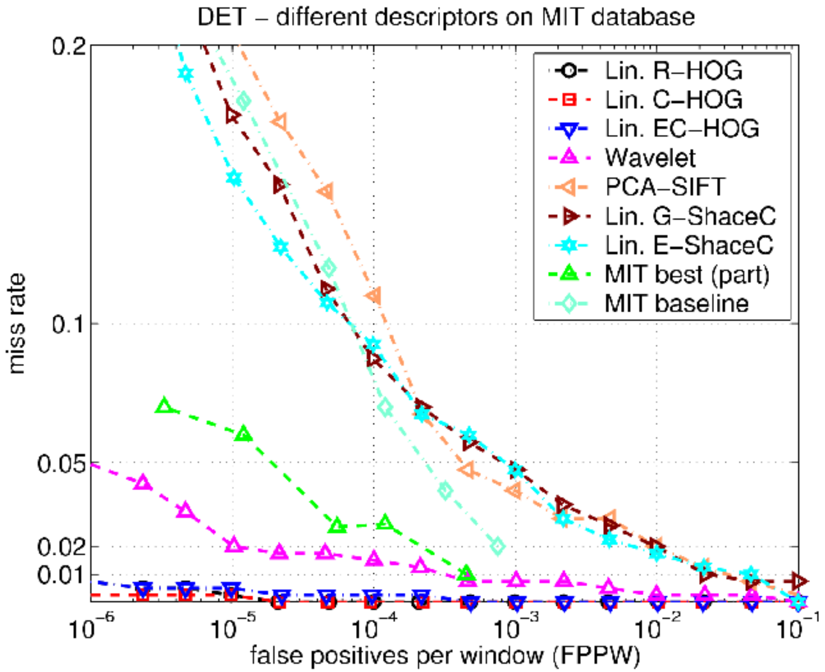


Evaluation Data Sets

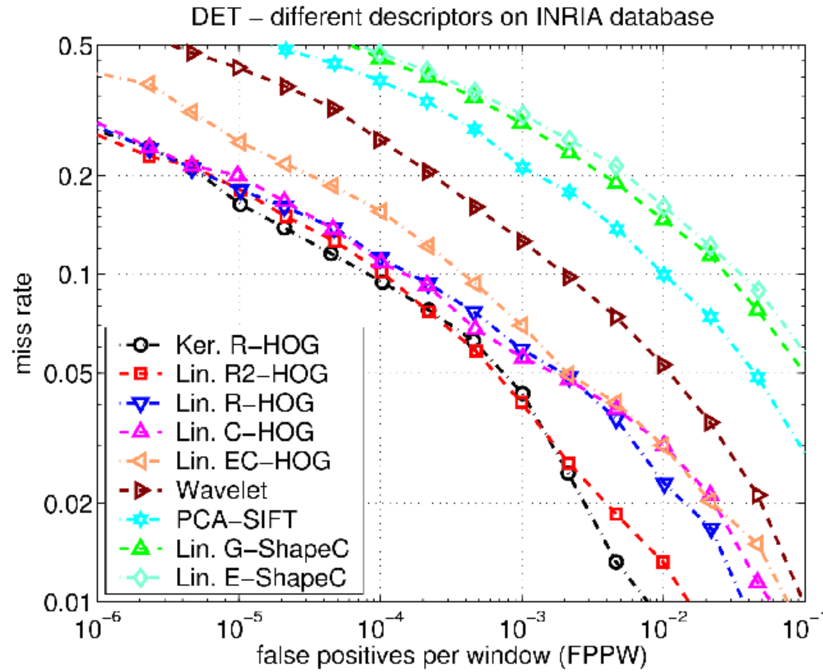
	MIT pedestrian database	INRIA person database
		
Train	507 positive windows Negative data unavailable	1208 positive windows 1218 negative images
Test	200 positive windows Negative data unavailable	566 positive windows 453 negative images
	Overall 709 annotations+ reflections	Overall 1774 annotations+ reflections

Overall Performance

MIT pedestrian database



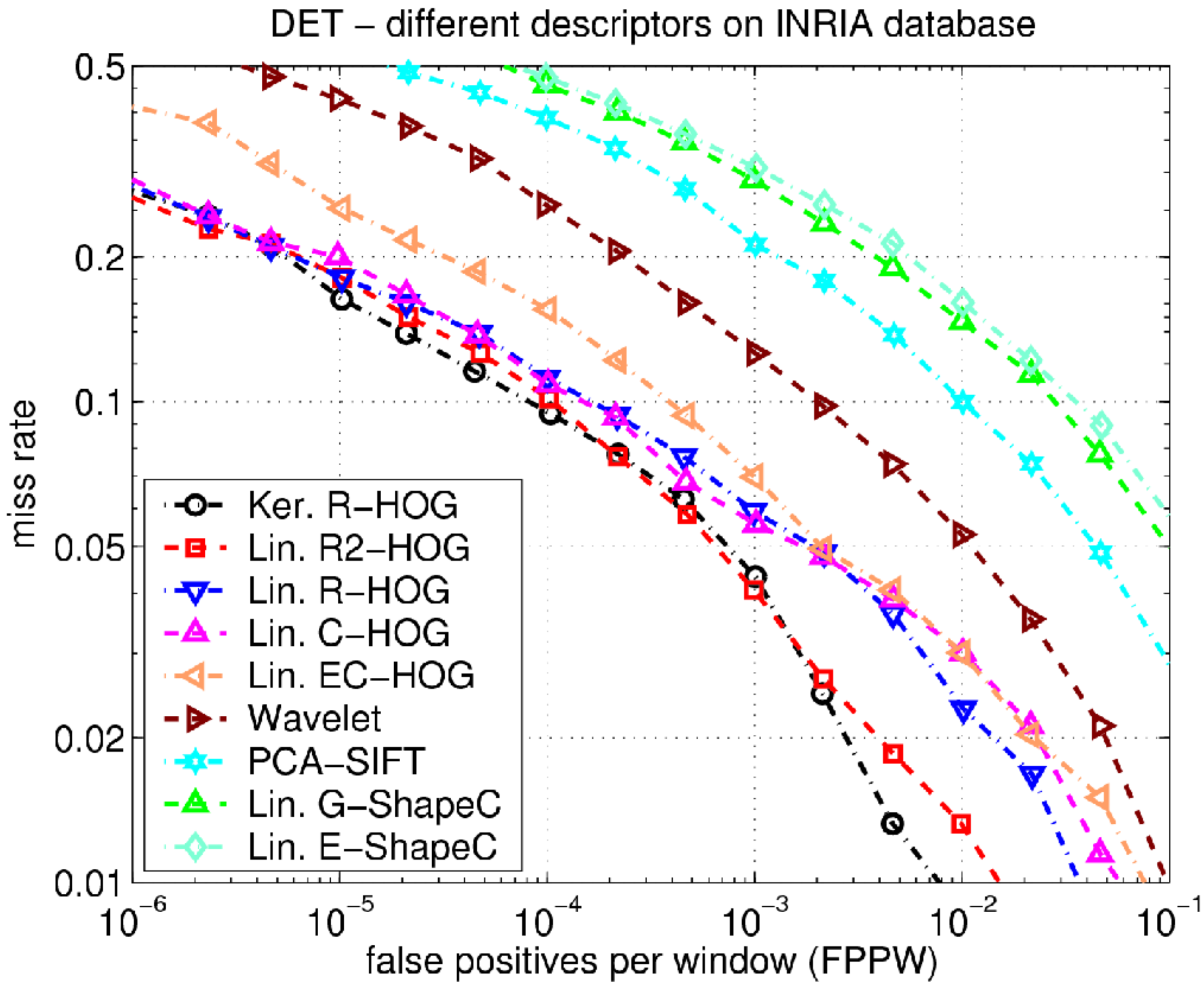
INRIA person database



R/C-HOG give near perfect separation on MIT database

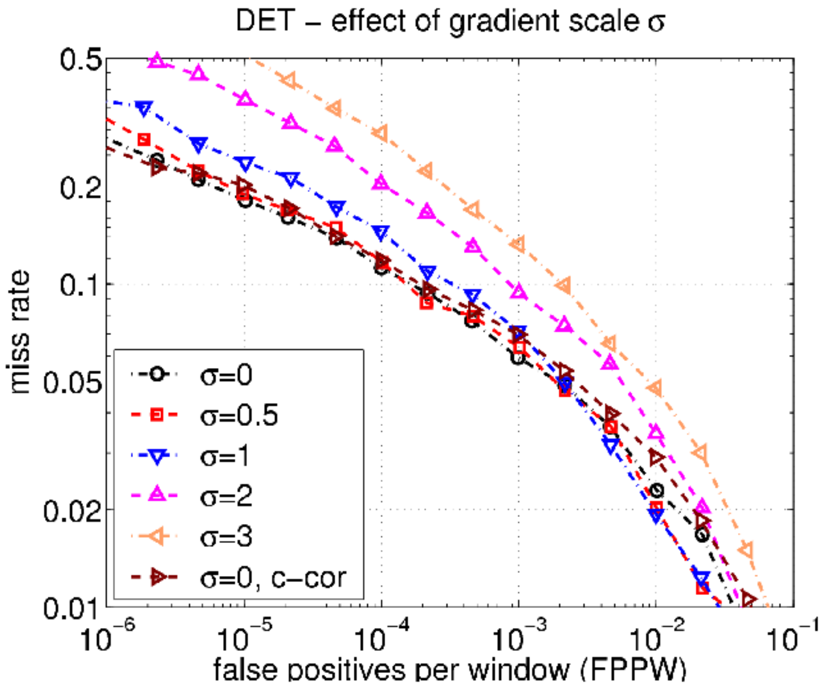
Have 1-2 order lower false positives than other descriptors

Performance on INRIA Database



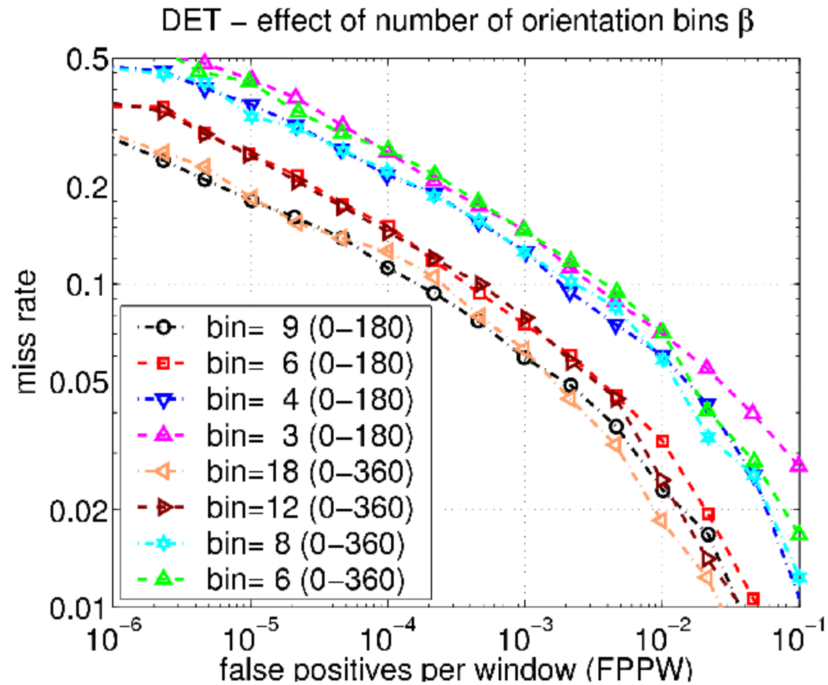
Effect of Parameters

Gradient smoothing, σ



Reducing gradient scale from 3 to 0 decreases false positives by 10 times

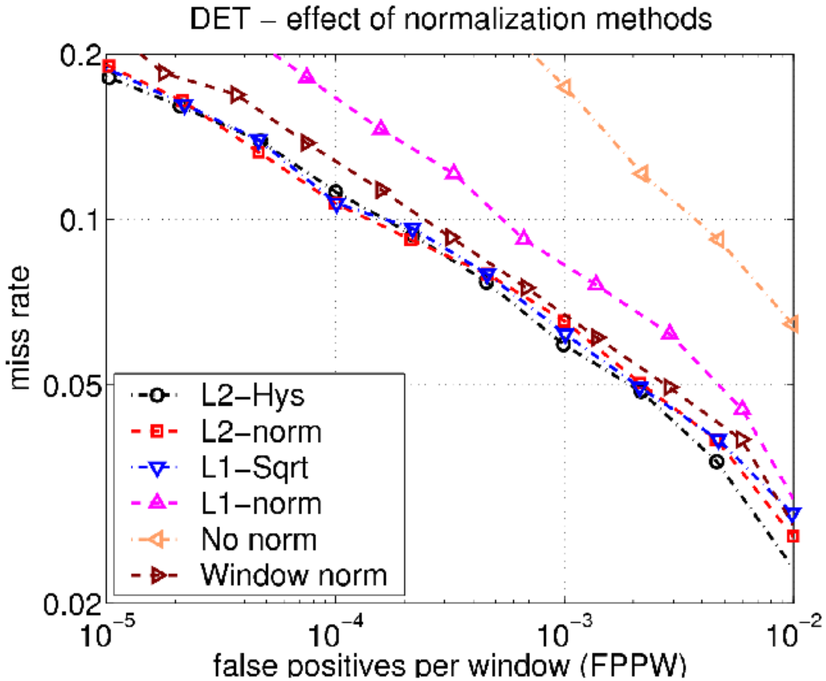
Orientation bins, β



Increasing orientation bins from 4 to 9 decreases false positives by 10 times

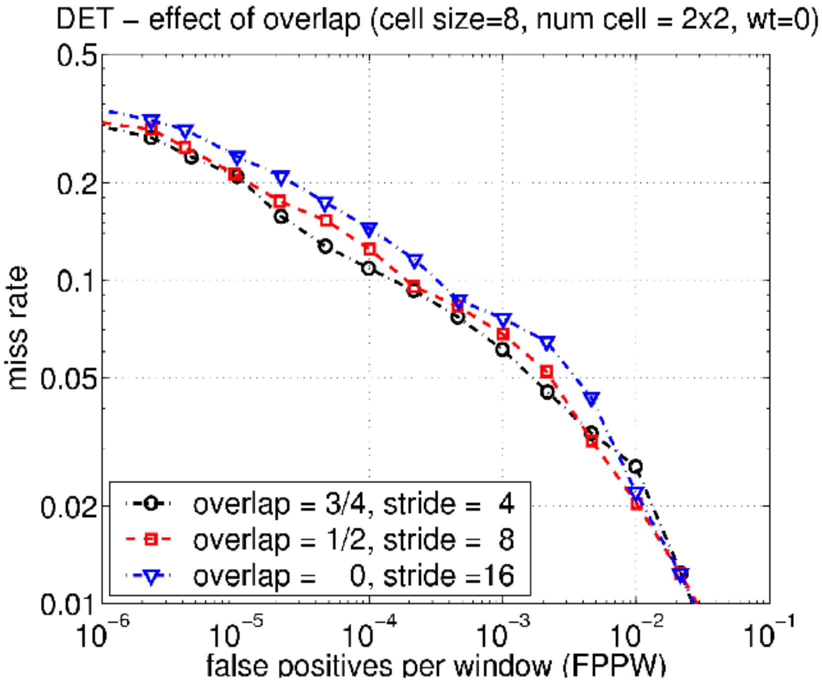
Normalisation Method & Block Overlap

Normalisation method



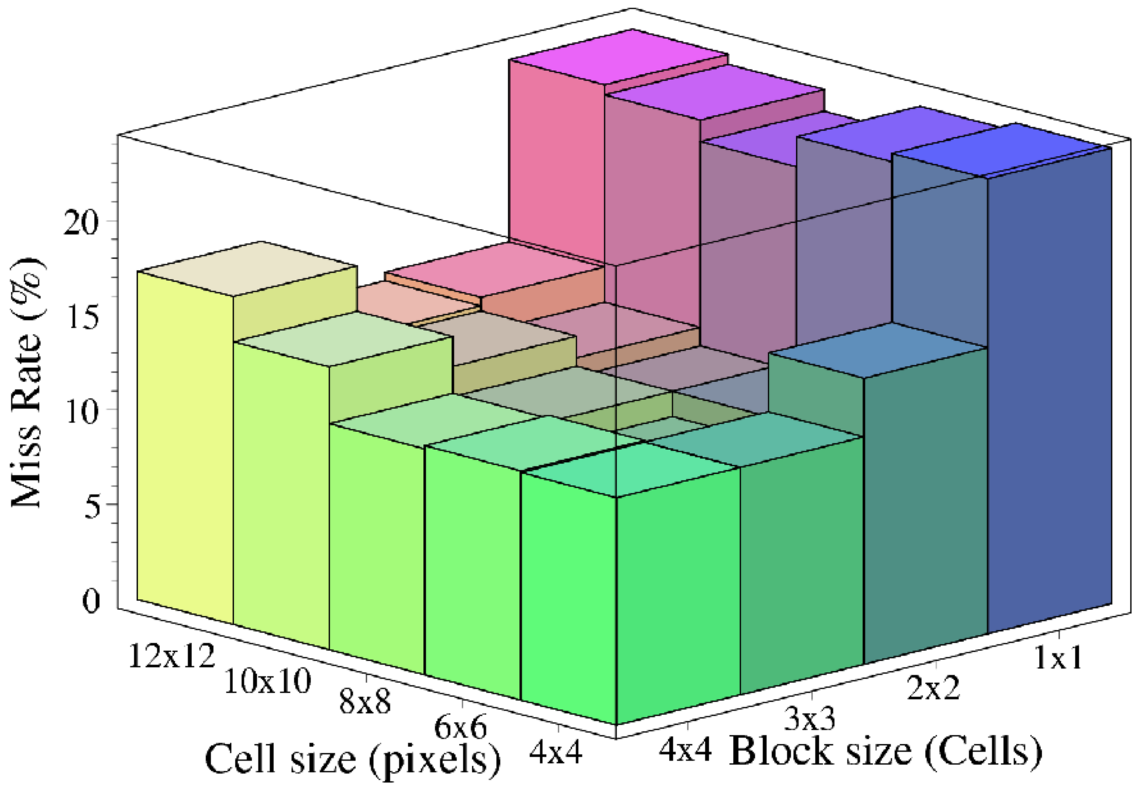
Strong local normalisation is essential

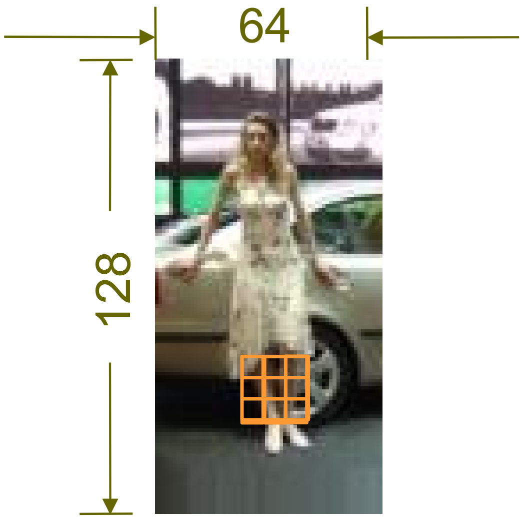
Block overlap



Overlapping blocks improve performance, but descriptor size increases

Effect of Block and Cell Size

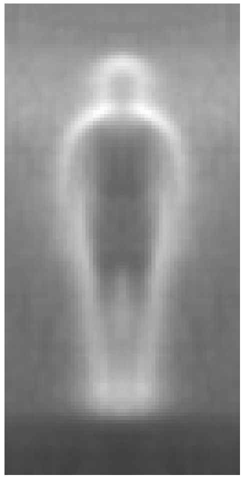




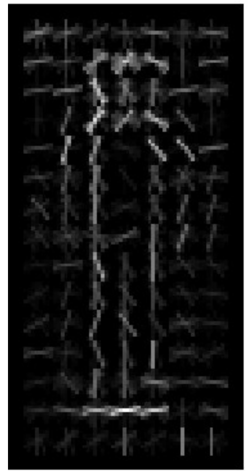
Trade off between need for local spatial invariance and need for finer spatial resolution

Descriptor Cues

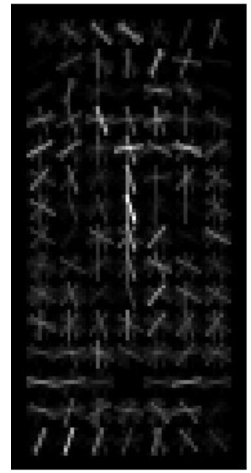
Input example



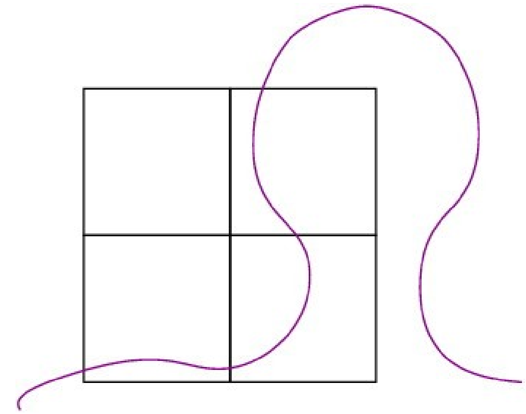
Average gradients



Weighted pos wts



Weighted neg wts



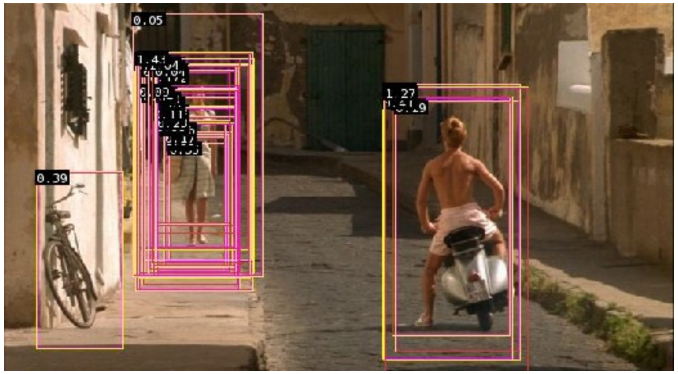
Outside-in weights

Most important cues are head, shoulder, leg silhouettes

Vertical gradients inside a person are counted as negative

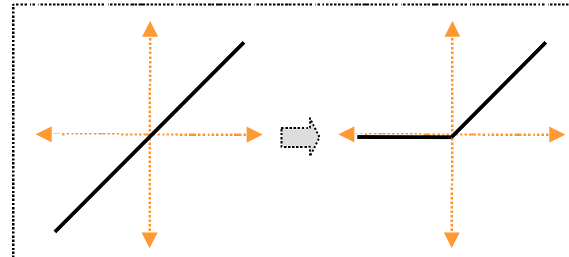
Overlapping blocks just outside the contour are most important

Multi-Scale Object Localisation

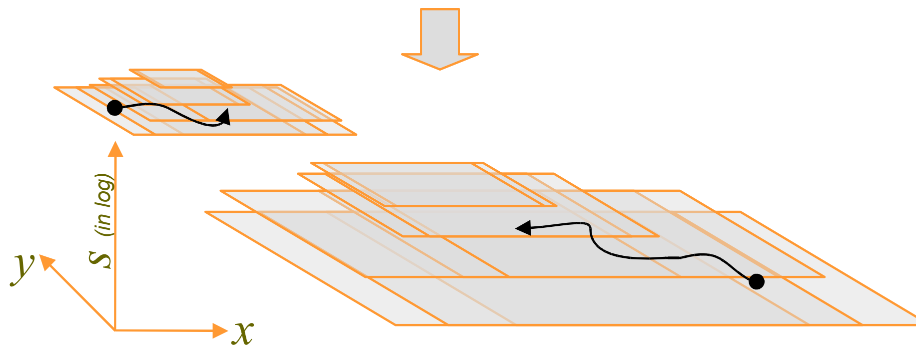


Multi-scale dense scan of detection window

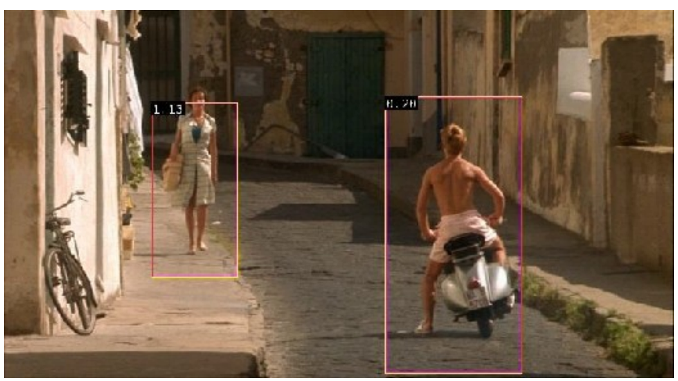
Bias



Clip Detection Score



Threshold



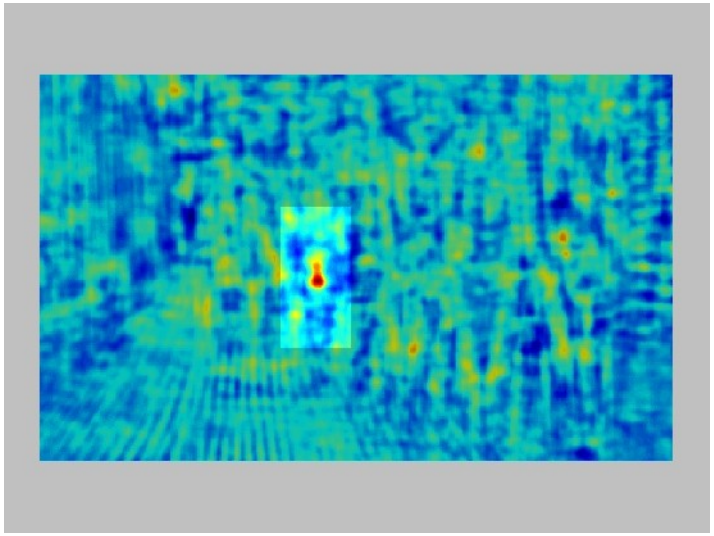
Final detections

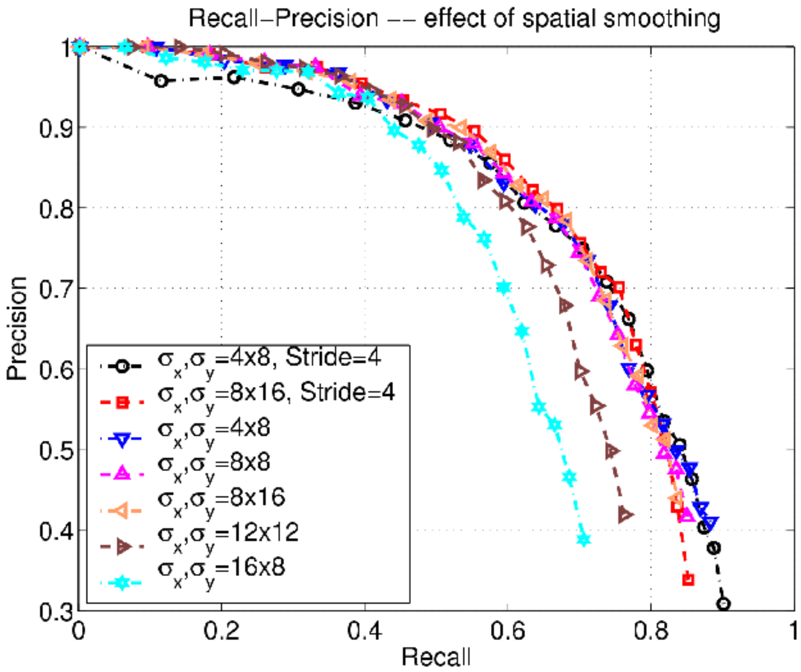
$$\mathrm{H}_i = [\exp(s_i)\sigma_x, \exp(s_i)\sigma_y, \sigma_s]$$

$$f(\mathbf{x}) = \sum_i^n w_i \exp\left(-\left\|(\mathbf{x}-\mathbf{x}_i)/\mathrm{H}_i^{-1}\right\|^2/2\right)$$

Apply robust mode detection, like mean shift

Effect of Spatial Smoothing



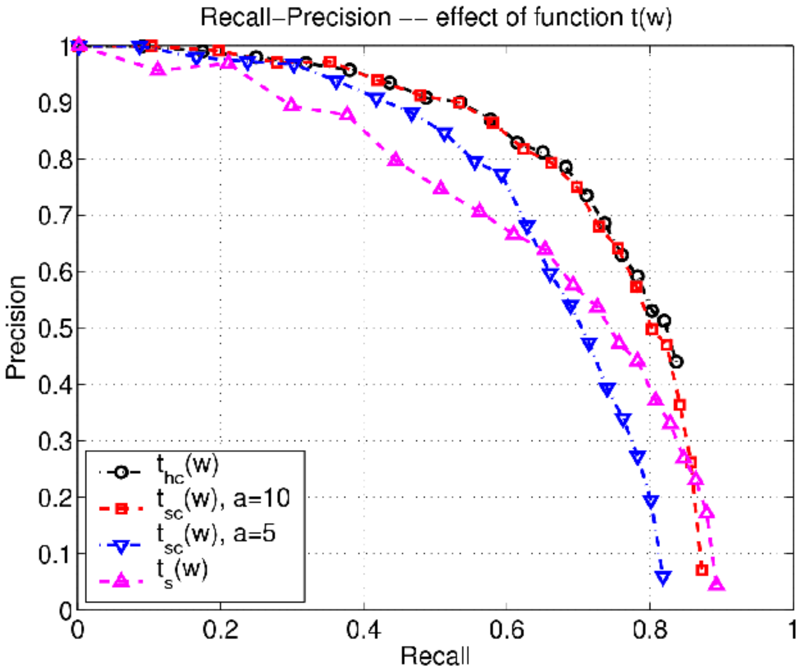


Spatial smoothing aspect ratio as per window shape, smallest sigma approx. equal to stride/cell size

Relatively independent of scale smoothing, sigma equal to 0.4 to 0.7 octaves gives good results

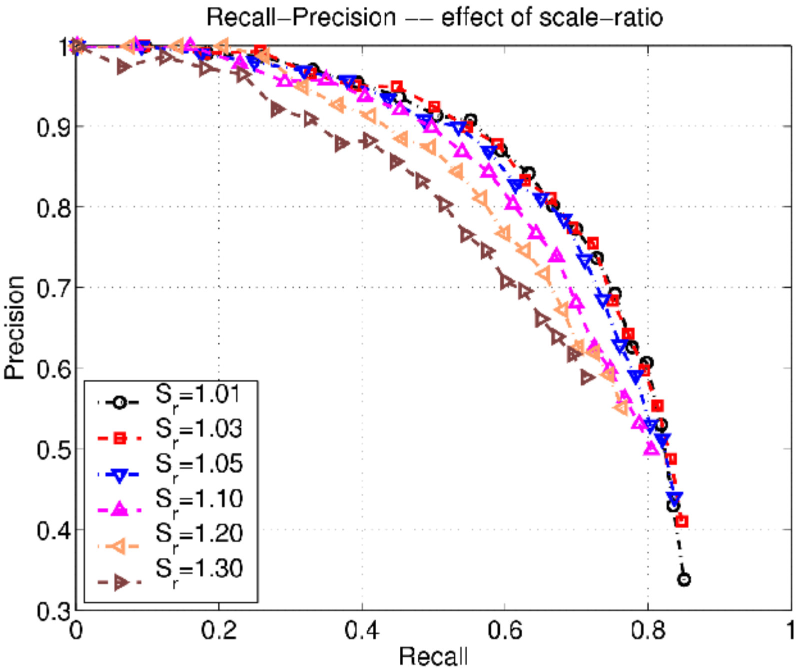
Effect of Other Parameters

Different mappings



Hard clipping of SVM scores gives the best results than simple probabilistic mapping of these scores

Effect of scale-ratio



Fine scale sampling helps improve recall

Results Using Static HOG

No temporal smoothing of detections

Conclusions for Static Case

Fine grained features improve performance

Rectify fine gradients then pool spatially

- No gradient smoothing, [1 0 -1] derivative mask
- Orientation voting into fine bins
- Spatial voting into coarser bins

Use gradient magnitude (no thresholding)

Strong local normalization

Use overlapping blocks

Robust non-maximum suppression

- Fine scale sampling, hard clipping & anisotropic kernel

Human detection rate of 90% at 10^{-4} false positives per window

Slower than integral images of Viola & Jones, 2001

Applications to Other Classes

M. Everingham et al. *The 2005 PASCAL Visual Object Classes Challenge*. Proceedings of the PASCAL Challenge

Parameter Settings

Most HOG parameters are stable across different classes

Parameters that change

- Gamma compression
- Normalisation methods
- Signed/un-signed gradients

Results from Pascal VOC 2006

	Person	Car	Motorbike	Bicycle	Bus	Sheep	Horse	Cow	Cat	Dog
Cam bridge	0.030	0.254	0.178	0.249	0.138	0.131	0.091	0.149	0.151	**0.118**
ENSMP	-	0.398	-	-	-	-	-	0.159	-	-
HOG	**0.164**	**0.444**	**0.390**	0.414	0.117	**0.251**	-	0.212	-	-
Laptev= HOG+ Ada-boost	0.114	-	0.318	**0.440**	-	-	**0.140**	0.224	-	-
TUD	0.074	-	0.153	-	-	-	-	-	-	-
TKK	0.039	0.222	0.265	0.303	**0.169**	0.227	0.137	**0.252**	**0.160**	0.113

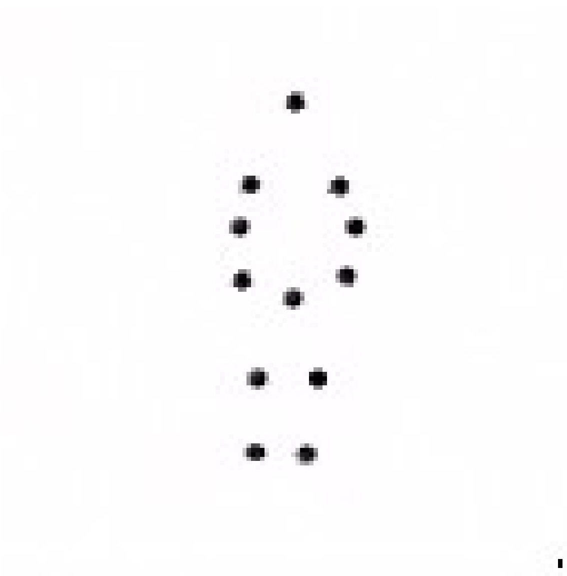
HOG outperformed other methods for 4 out of 10 classes

Its adaBoost variant outperformed other methods for 2 out of 10 classes

Finding People in Videos

Finding People in Videos

- Motivation
  - Human motion is *very* characteristic
- Requirements
  - Must work for moving camera and background
  - Robust coding of relative motion of human parts
- Previous works
  - Viola et al, 2003
  - Gavrila et al, 2004
  - Efros et al, 2003



Courtesy: R. Blake
Vanderbilt Univ

Handling Camera Motion

- Camera motion characterisation
  - Pan and tilt is locally translational
  - Rest is depth induced motion parallax
- Use local differential of flow
  - Cancels out effects of camera rotation
  - Highlights 3D depth boundaries
  - Highlights motion boundaries
- Robust encoding into oriented histograms
  - Some focus on capturing motion boundaries
  - Other focus on capturing internal motion or relative dynamics of different limbs

Motion HOG Processing Chain

Detection windows

↓

Normalise gamma & colour

↓

Compute optical flow

↓

Compute differential flow

↓

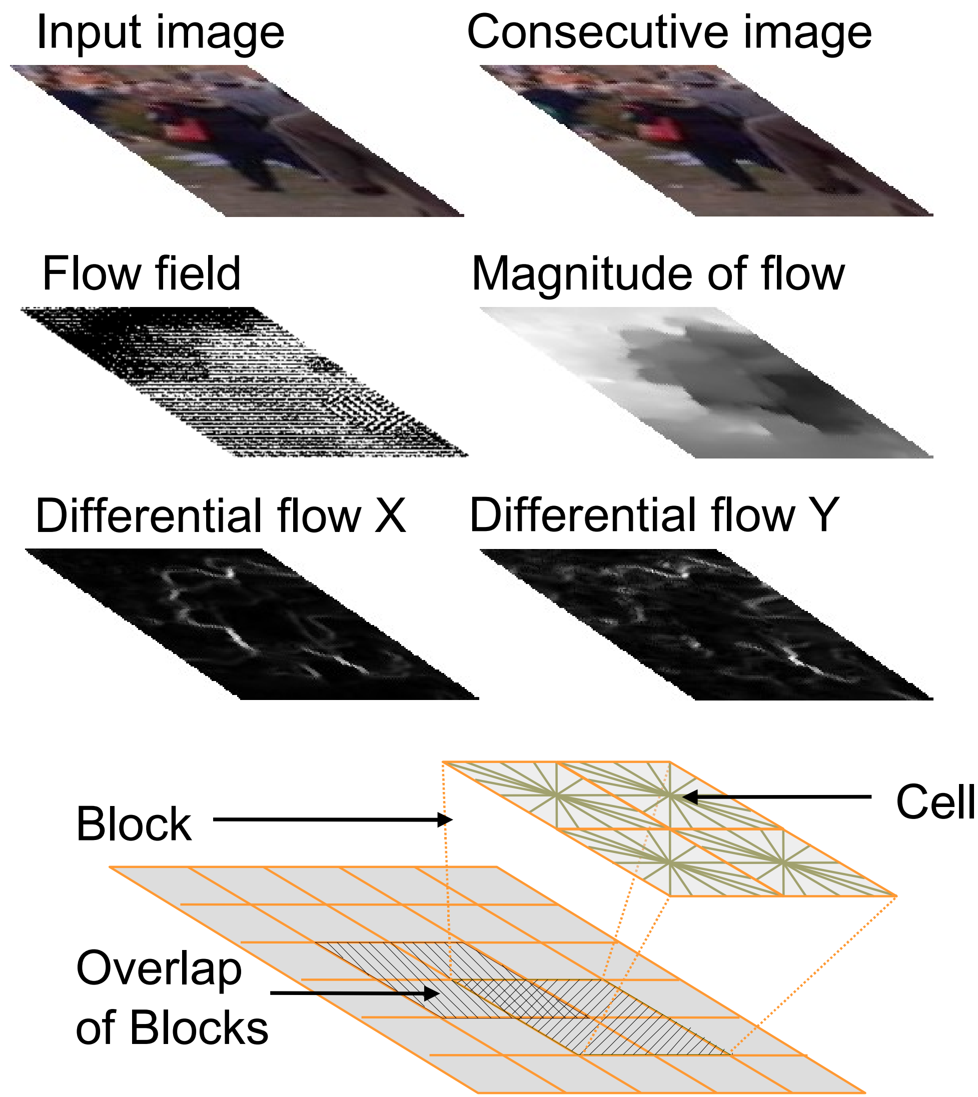
Accumulate votes for differential flow orientation over spatial cells

↓

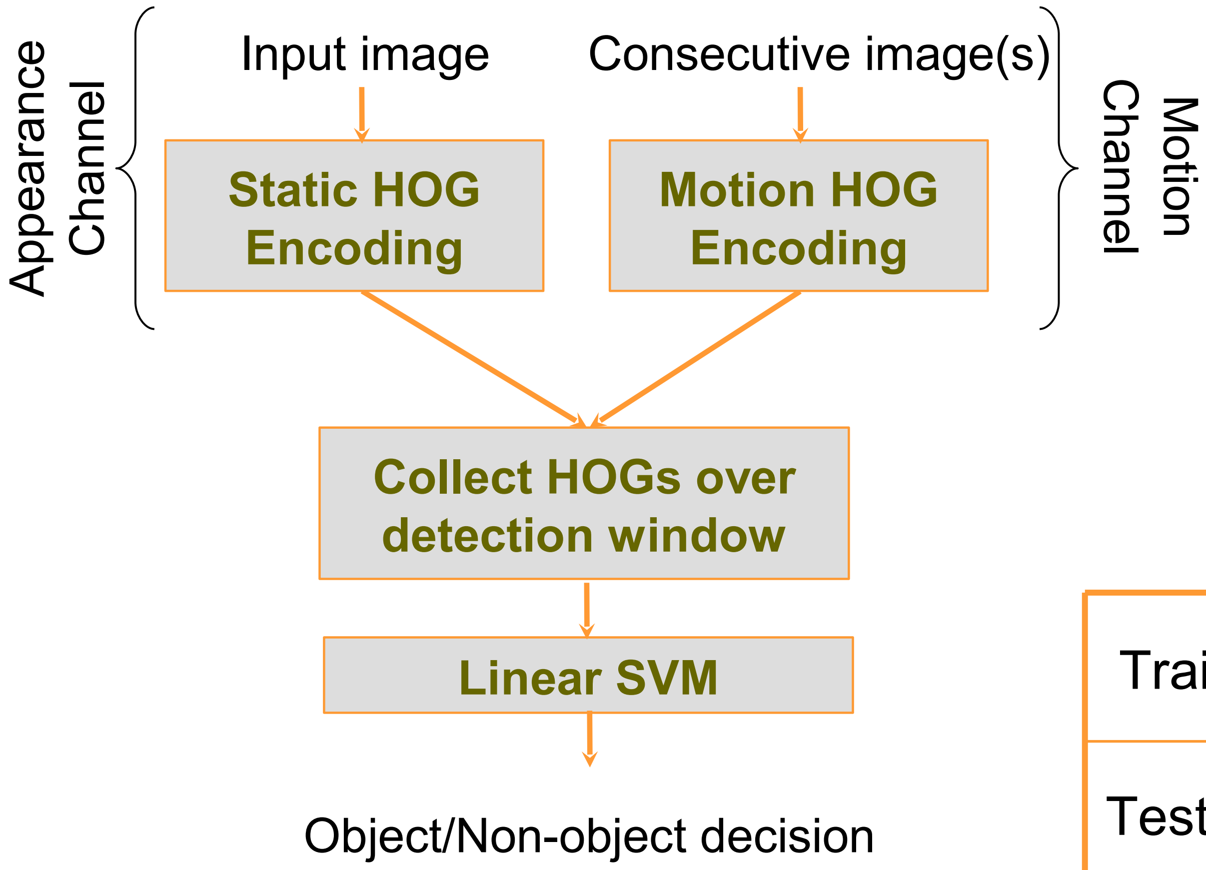
Normalise contrast within overlapping blocks of cells

↓

Collect HOGs for all blocks over detection window



Overview of Feature Extraction



Data Set

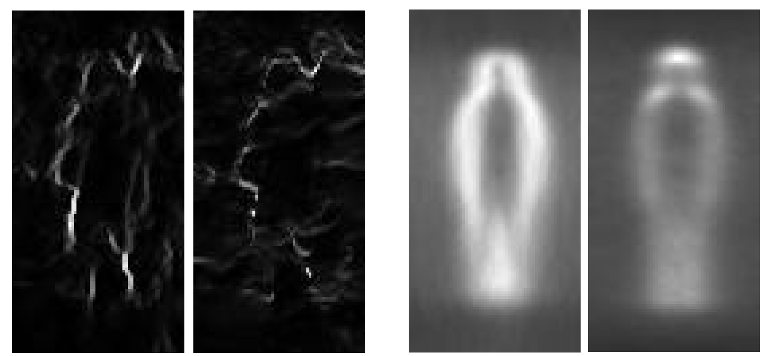
Train	5 DVDs, 182 shots 5562 positive windows
Test 1	Same 5 DVDs, 50 shots 1704 positive windows
Test 2	6 new DVDs, 128 shots 2700 positive windows

Coding Motion Boundaries

Treat x, y-flow components as independent images

Take their local gradients separately, and compute HOGs as in static images

First frame | Second frame | Estd. flow | Flow mag.



x-flow diff | y-flow diff | Avg. x-flow diff | Avg. y-flow diff

Motion Boundary Histograms (MBH) encode depth and motion boundaries

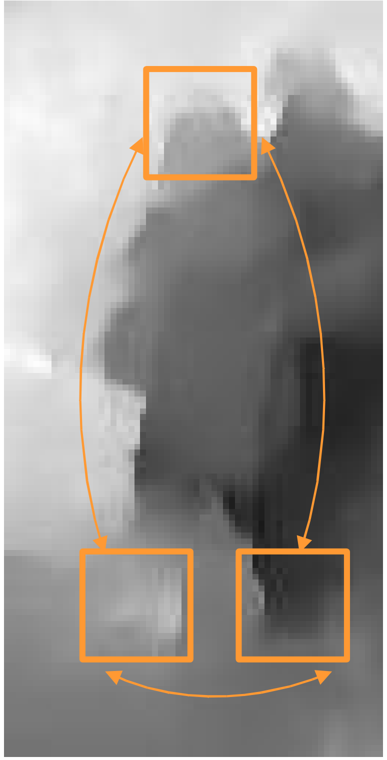
Coding Internal Dynamics

Ideally compute relative displacements of different limbs

- Requires reliable part detectors

Parts are relatively localised in our detection windows

Allows different coding schemes based on fixed spatial differences



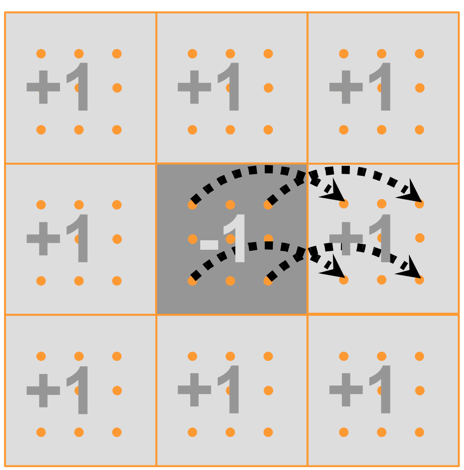
Internal Motion Histograms (IMH) encode relative dynamics of different regions

…IMH Continued

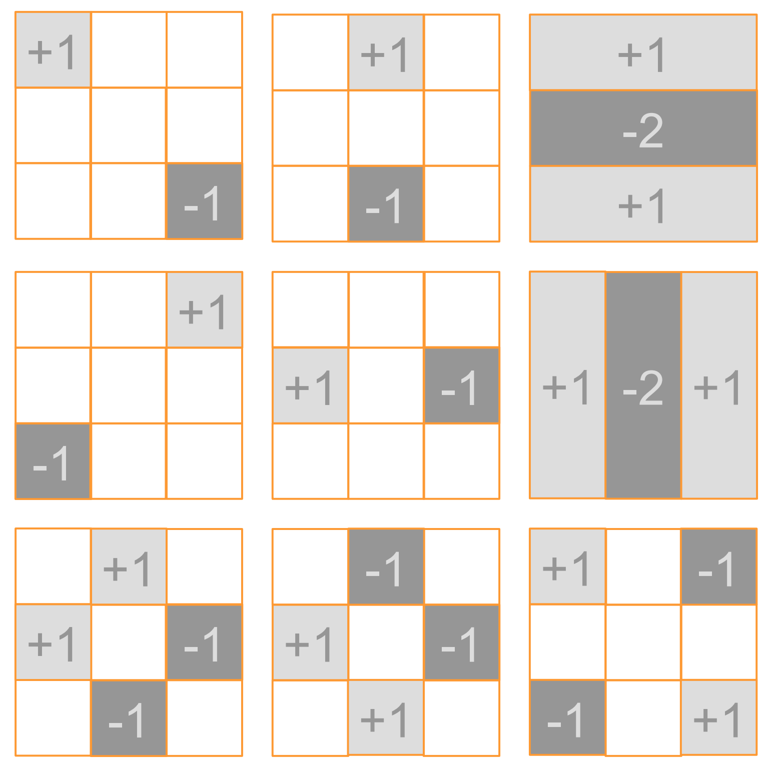
Simple difference

- Take x, y differentials of flow vector images $[I_x, I_y]$
- Variants may use larger spatial displacements while differencing, e.g. [1 0 0 0 -1]

Center cell difference



Wavelet-style cell differences



Flow Methods

Proesman's flow [Proesmans et al. ECCV 1994]

- 15 seconds per frame

Our flow method

- Multi-scale pyramid based method, no regularization
- Brightness constancy based damped least squares solution

$$[x, y]^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \beta\mathbf{I}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b}$$

  on 5X5 window

- 1 second per frame

MPEG-4 based block matching

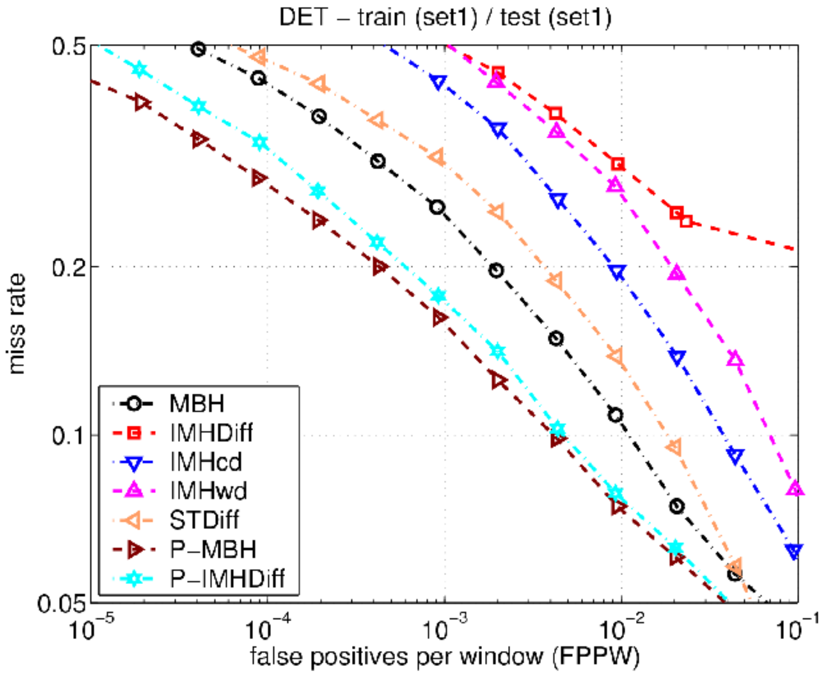
Input image

Proesman's flow

Our multi-scale flow

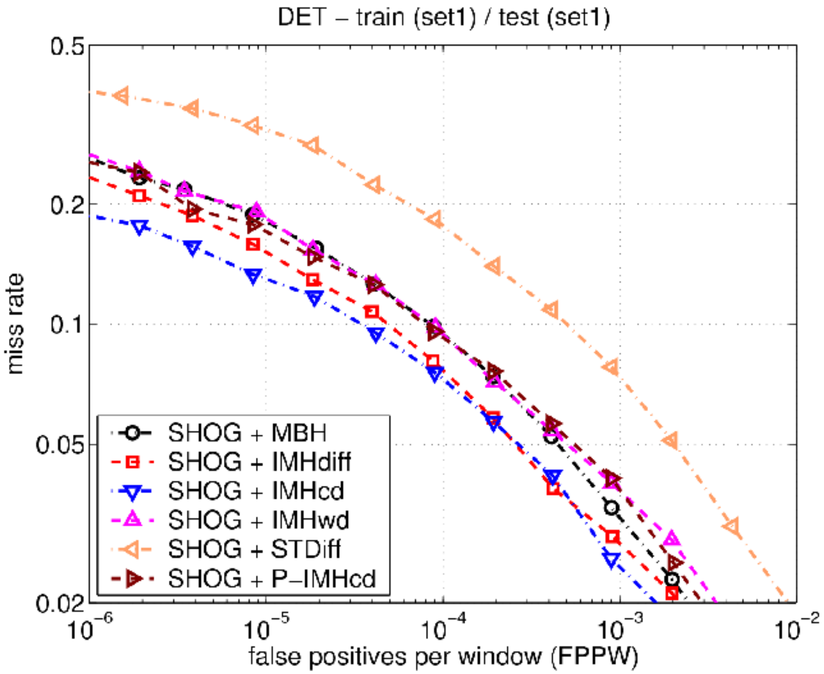
Performance Comparison

Only motion information



With motion only, MBH scheme on Proesmans' flow works best

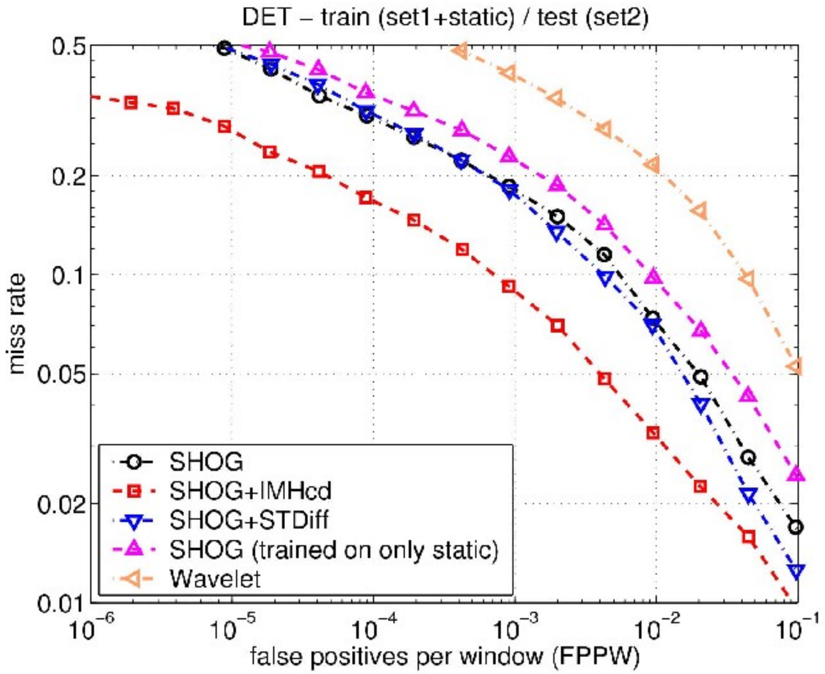
Appearance + motion



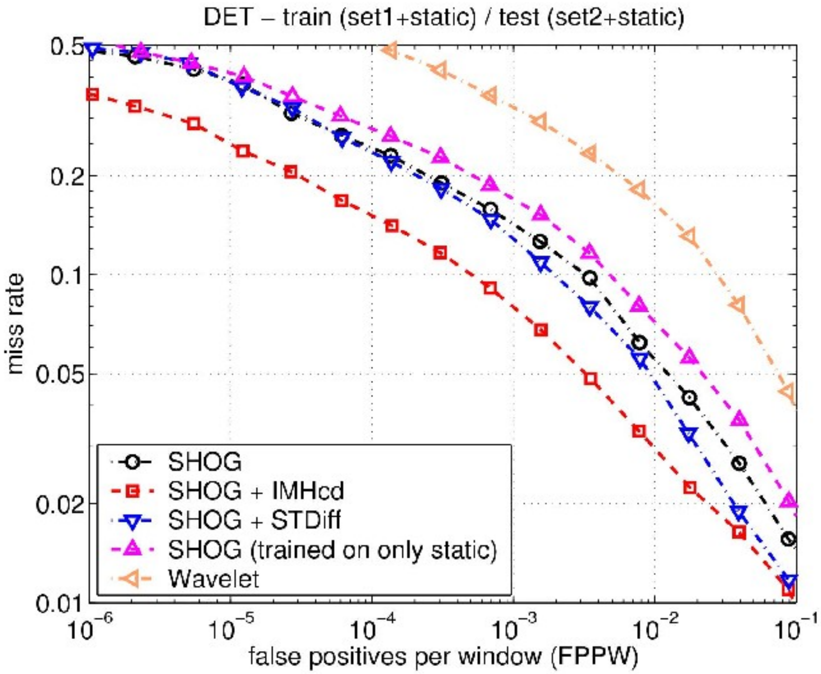
Combined with appearance, centre difference IMH performs best

Trained on Static & Flow

Tested on flow only



Tested on appearance + flow



Adding static images during test reduces performance margin

No deterioration in performance on static images

Motion HOG Video

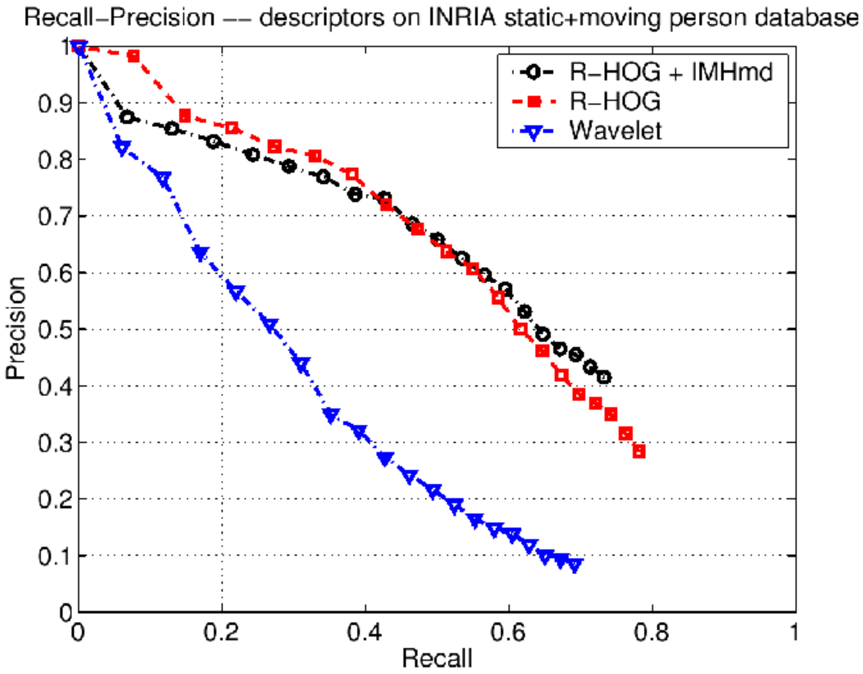
No temporal smoothing, each pair of frames treated independently

Recall-Precision for Motion HOG

Unresolved issue!

Recall-precision plots for the combined static + motion HOG shows there is no gain over the static HOG

Results are disappointing; probable reason is different internal biases during non-maximum suppression for static and motion HOG



Conclusions for Motion HOG

Summary

- When combined with appearance, IMH outperforms MBH
- Regularization in flow estimates reduces performance
- MPEG4 block matching looks good but motion estimates not good for detection
- Larger spatial difference masks help
- Strong local normalization is very important
- Relatively insensitive to number of orientation bins

Window classifier reduces false positives by 10 times

- Issue of unexpectedly low precision for full detector
- Slow compared to static HOG

Human Part Detectors

Current approaches:

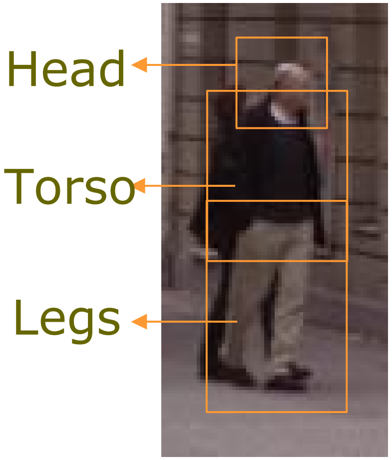
- Mohan et al, 2000;
- Mikolajczyk et al, 2004

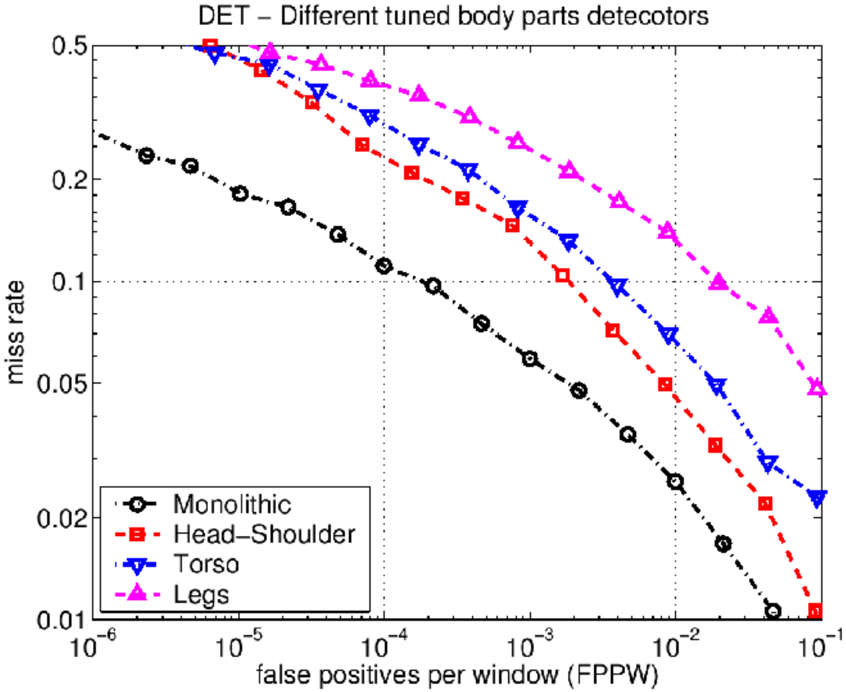
Current approach to part detectors

- Use manual part annotations to learn individual classifiers
- Parameters optimized for each detector

Other approaches

- Cluster block feature vectors to automatically learn different part representations





Part-based Human Detectors

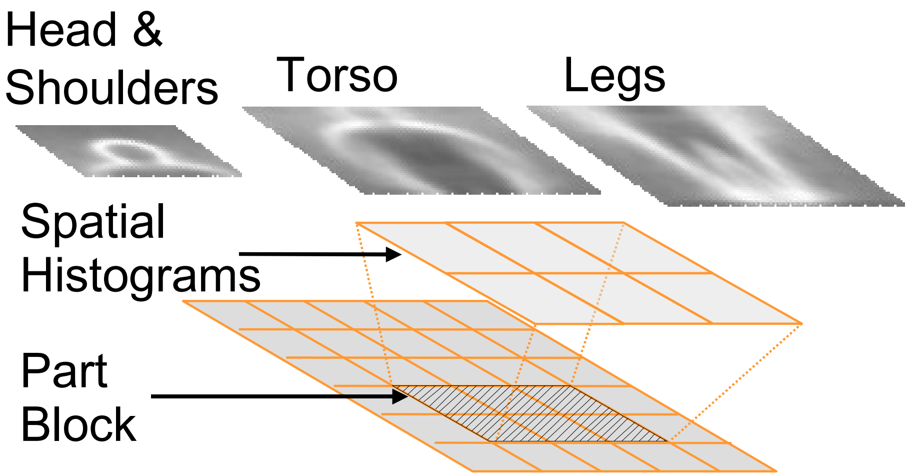
Learned part detectors

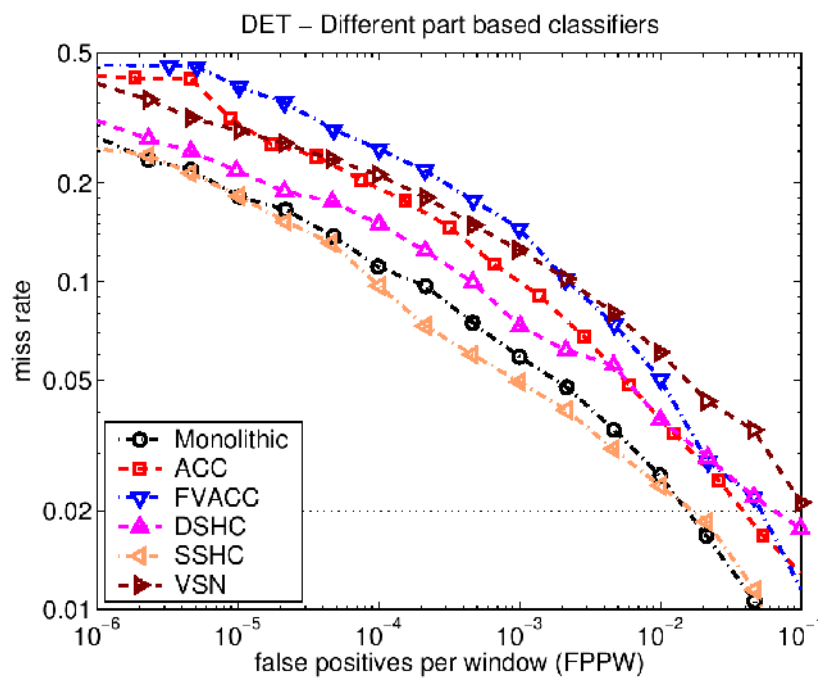
Scan part detectors over the detection window

Accumulate top 3 estimates for each part over spatial histograms

Collect spatial histograms for all parts over the detection window

Linear SVM





Contributions

Bottom-up approach to object detection

Robust feature encoding for person detection

Gives state-of-the-art results for person detection

Also works well for other object classes

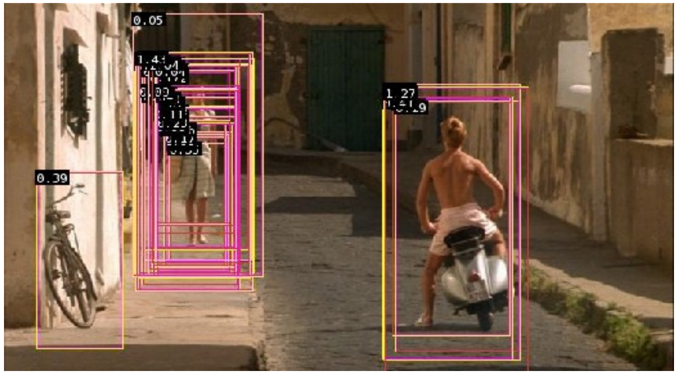
Proposed differential motion features vectors for feature extraction from videos

Future Work

- Fix the motion HOG integration
- Real time implementation is possible
- Use rejection cascade algorithms for selecting most relevant features
- Part based detector for handling partial occlusions
- Extend motion HOG to activity recognition
- Use higher level image analysis to improve performance

Thank You

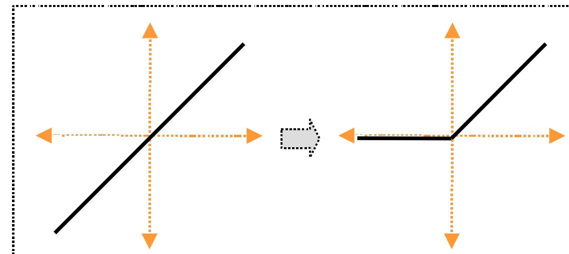
Multi-Scale Object Localisation



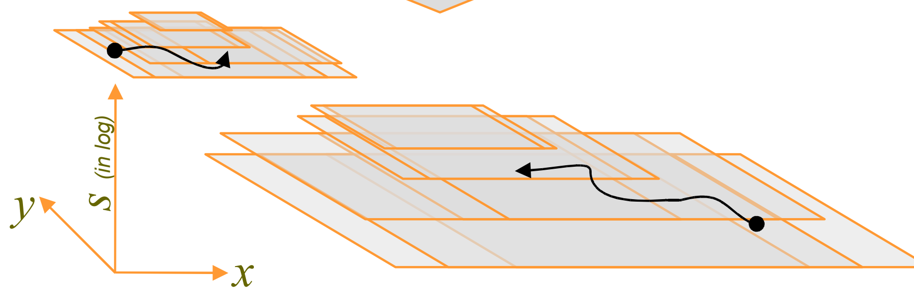
Multi-scale dense scan of detection window

Goal

Bias



Clip Detection Score

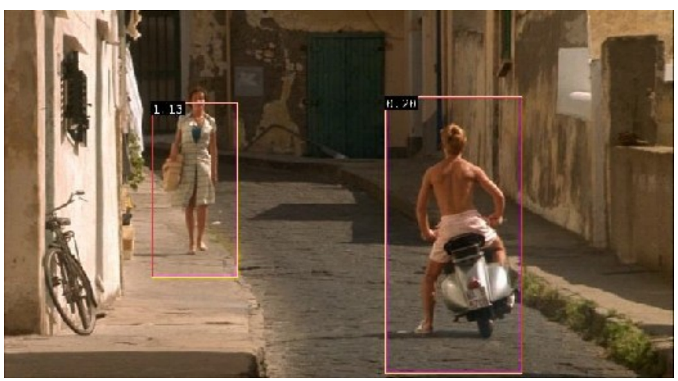


Threshold

$$\mathrm{H}_i = [\exp(s_i)\sigma_x, \exp(s_i)\sigma_y, \sigma_s]$$

$$f(\mathbf{x}) = \sum_i^n w_i \exp\left(- \left\| (\mathbf{x} - \mathbf{x}_i) / \mathrm{H}_i^{-1} \right\|^2 / 2 \right)$$

Apply robust mode detection, like mean shift



Final detections