Human detection using histogram of oriented gradients

Srikumar Ramalingam

School of Computing

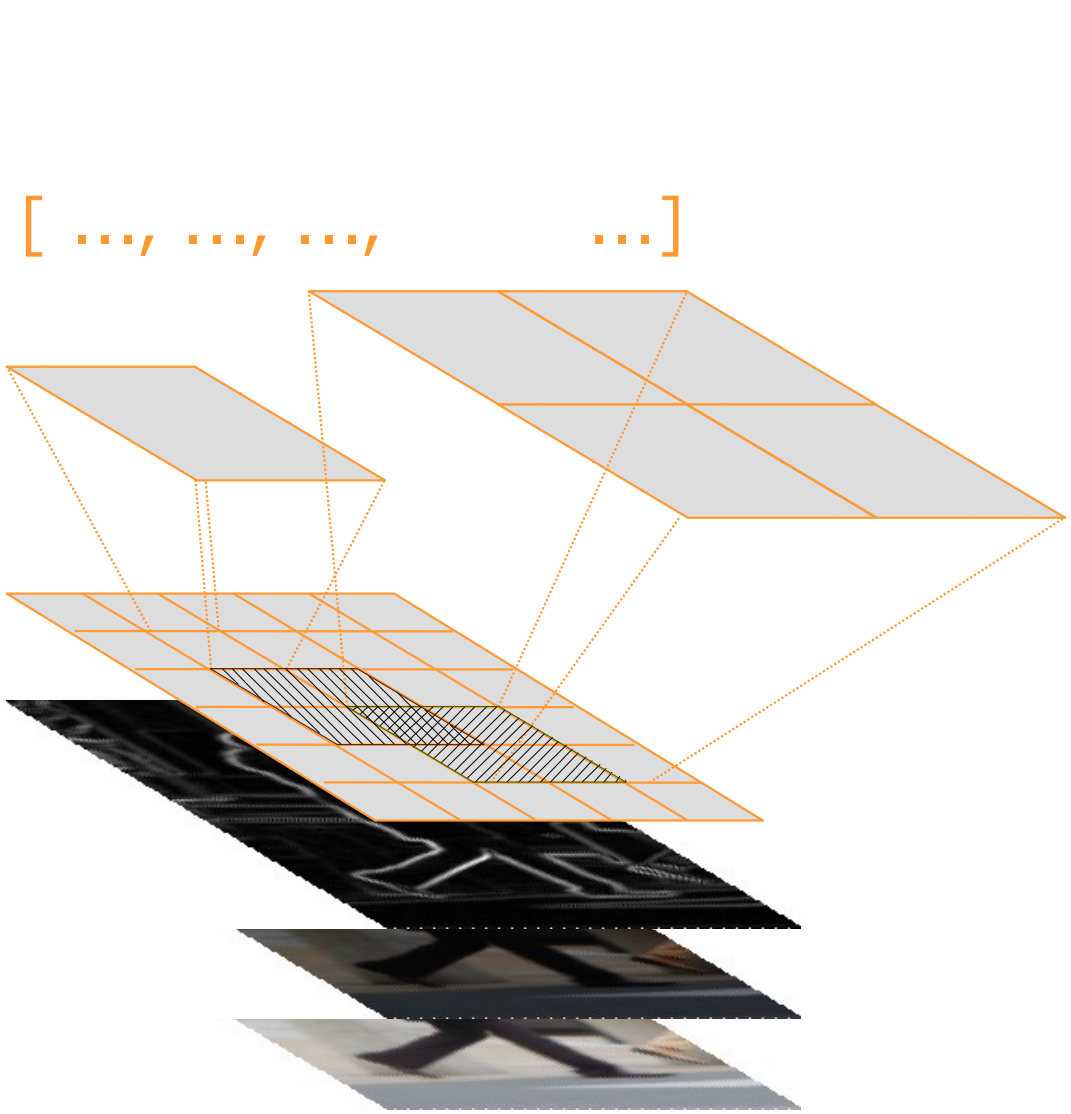
University of Utah

Reference

Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005.

https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

Descriptor Processing Chain



Object/Non-object

Linear SVM

Collect HOGs over detection window

Contrast normalize over overlapping spatial cells

Weighted vote in spatial & orientation cells

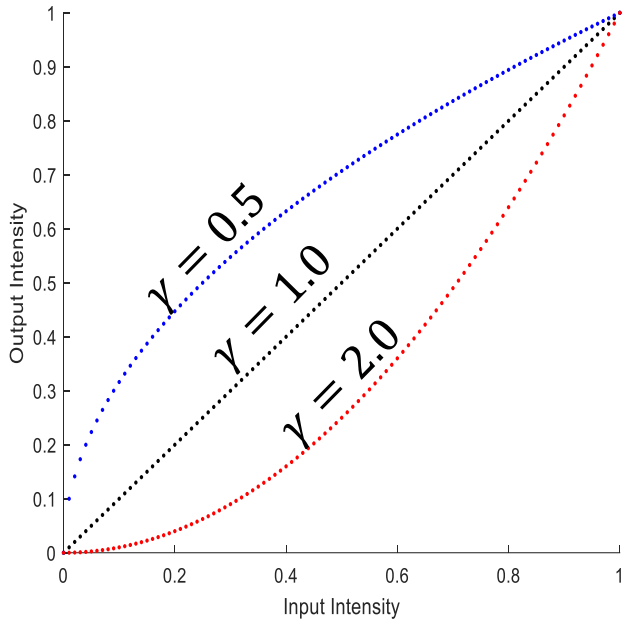
Compute gradients

Gamma compression

Image Window

Slide courtesy : Navneet Dalal

Gamma correction



$\gamma = 2.0$ $\gamma = 1.0$ $\gamma = 0.5$

- Each pixel has a brightness value or luminance that varies from 0 to 1.
- Different cameras do not capture the correct values for luminance, and there is usually a non-linear mapping.

$$I_{out} = I_{in}^{\gamma}$$

Gradient computation

$$G_x = (-1 \quad 0 \quad 1)$$

$$G_y = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

Important: No smoothing with a Gaussian filter is used prior to the computation of gradients.

Image I

$I_x = G_x \odot I$

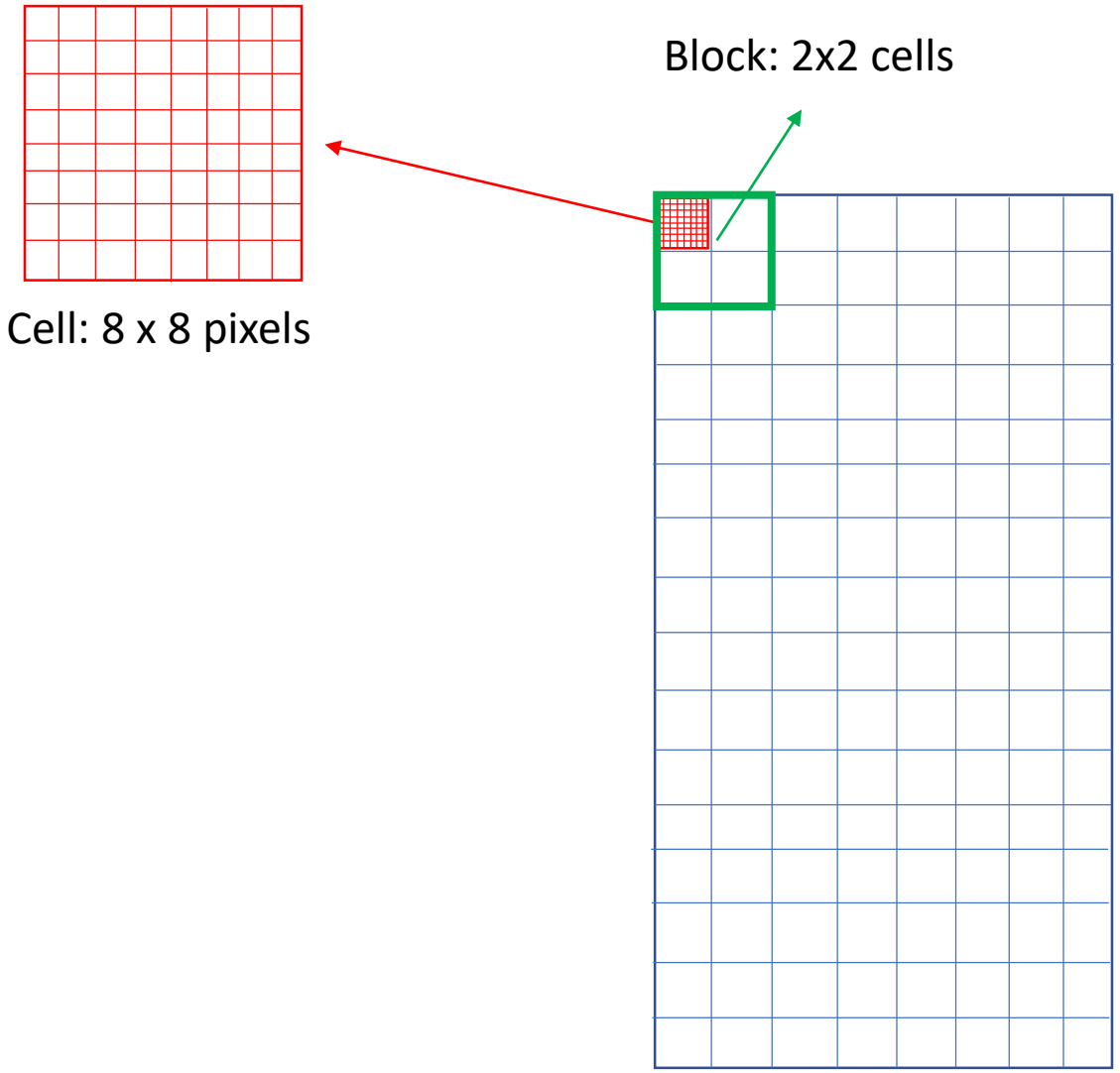
$I_y = G_y \odot I$

$$\theta = \tan^{-1} \frac{I_y}{I_x}$$

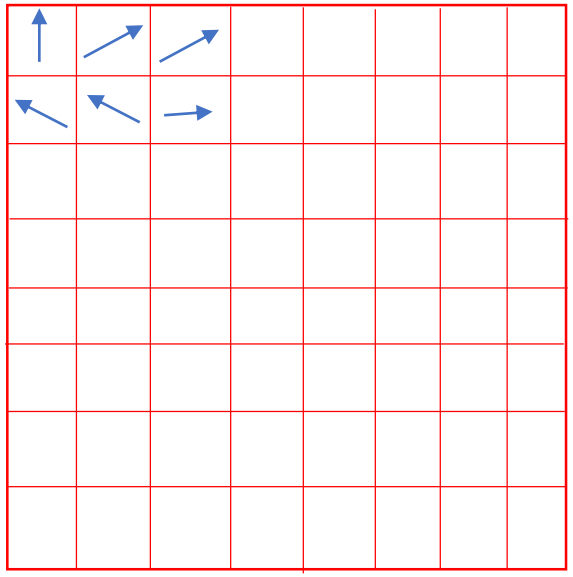
$$magnitute = \sqrt{I_x^2 + I_y^2}$$

Cells and Blocks in building the feature vector

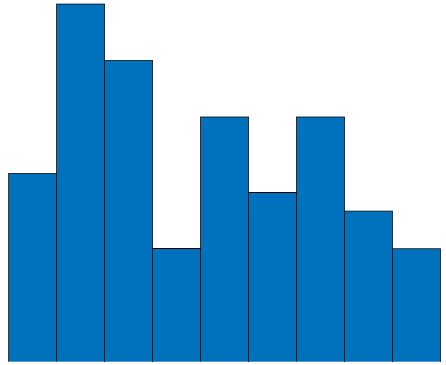
- The feature vector for human detection is built using cells and blocks.
- Each cell is a matrix of 8x8 pixels.
- Every block is a matrix of 2x2 cells, but the blocks are accumulated by overlapping with blocks from previous locations.
- We consider an image patch of size 64 x128 for detecting humans.



Histogram of orientations in every cell



Cell: 8 x 8 pixels



Histogram with 9 bins for orientations varying from 0 to 180 degrees.

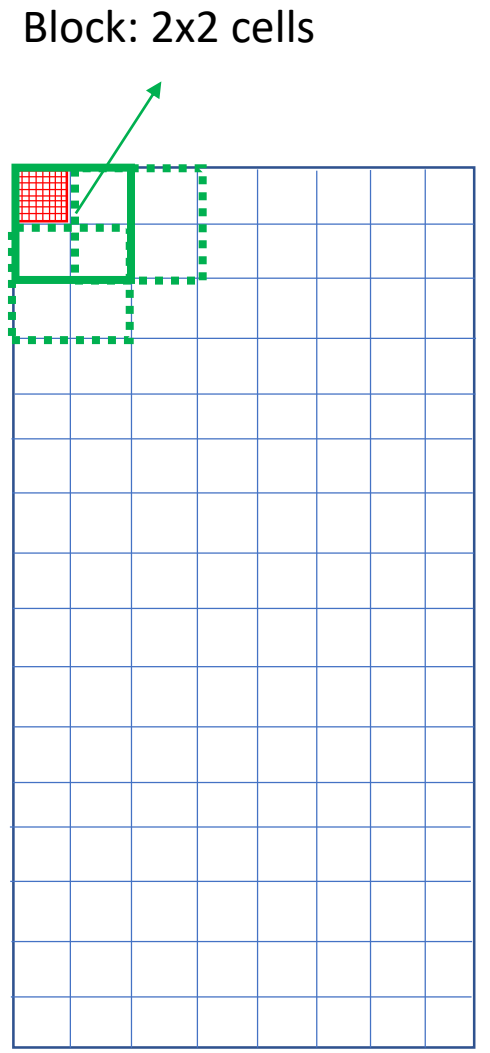
- We collect the magnitude and gradient angles for each pixel inside a cell to form the histogram with 9 bins (20 degree width for every bin for angles varying from 0 to 180 degrees).
- To avoid aliasing, votes are interpolated bilinearly from both sides based on the bin center.

Block Normalization

- Let v denote a 36×1 vector corresponding to a block aggregating the histograms from 4 cells.
- While using $L2$ normalization we have the following:

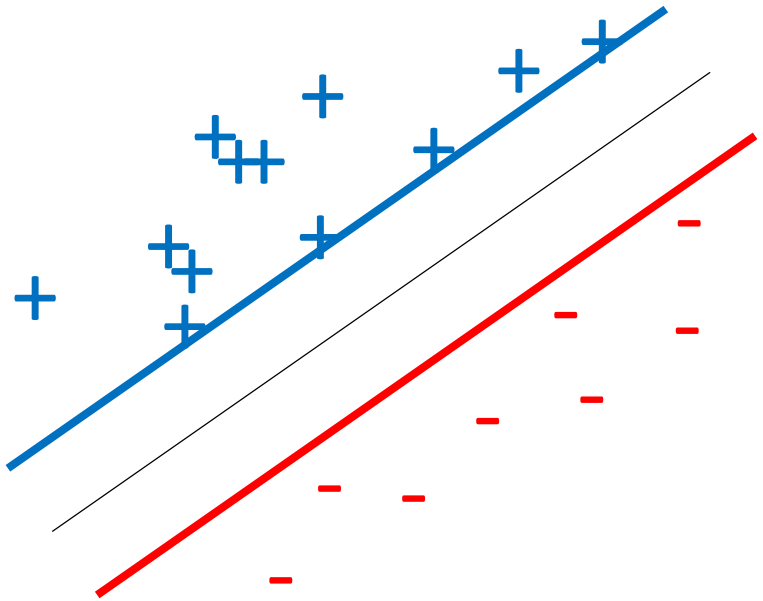
$$v = \frac{v}{\sqrt{|v|_2^2 + \epsilon}}$$

- Total length of the feature vector for an image patch of size 64 x 128 = 7 x 15 x 36 = 3780



Overlapping blocks : 7 x 15 for image patch of size 64 x 128

Support vector machines for pedestrian detection

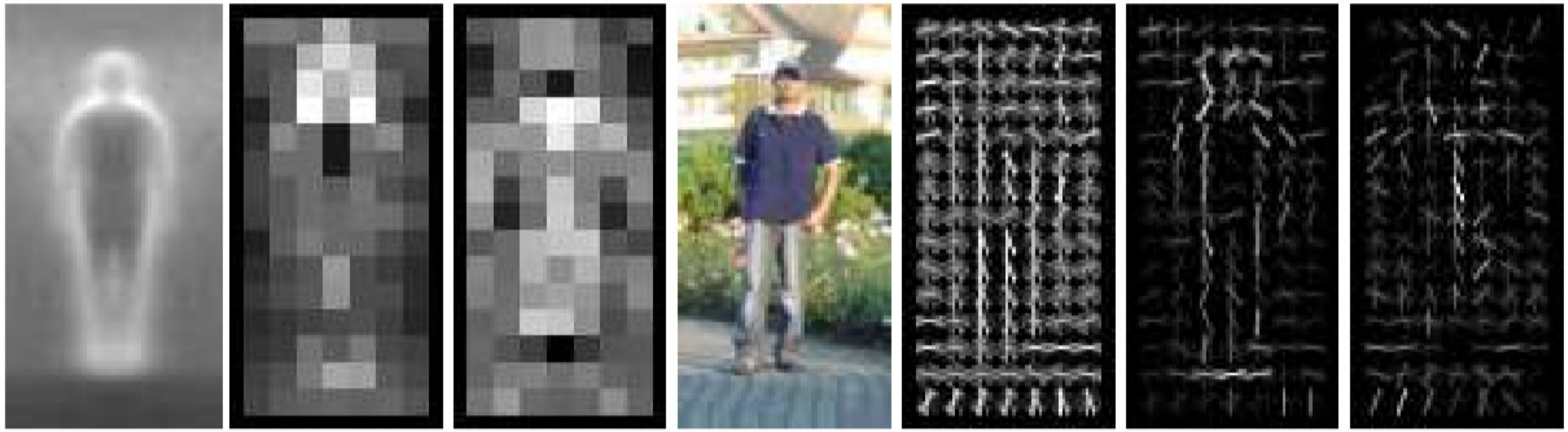


$$\min_{w}(|w| + C\sum \zeta_i$$

$$s.t.\left(w^T x_j + b\right)y_j \geq 1 - \zeta_i$$

- We are given samples as follows (x_i, y_i), where $x_i$is a 3760×1 dimensional feature vector, and y_i denotes the labels (+1 for positives, and -1 for negatives)
- For positives, we take image patches of dimensions 64x128 containing humans, and for negatives we take random image patches without any human in the images.
- We can use a linear SVM with $C = 0.01$.

SVM weight coefficients



average image from positive training samples | maximum SVM positive coefficient | maximum SVM negative coefficient | test image | Descriptor computed for the test image | Descriptor weighted by positive SVM weights | Descriptor weighted by negative SVM weights

- The most important cells are the ones that typically contain major human contours (especially the head and shoulders and the feet).

Threshold for Human detection

- For a given feature vector x_j , we check if

$$w^T x_j + b > T$$

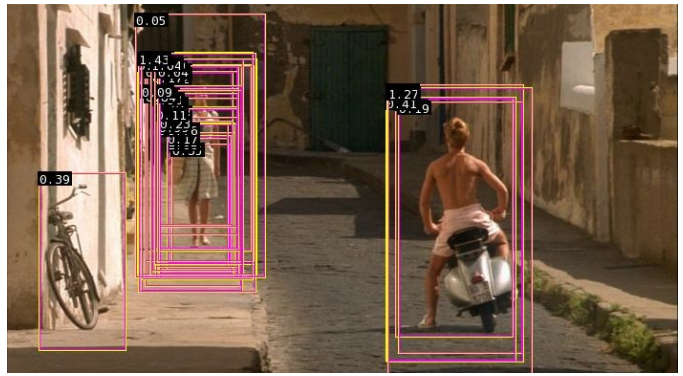
- T=0 is the natural threshold for SVM, but we use a slightly higher threshold, say 0.4, to avoid false positives.

Multi-scale detection

- The feature vectors are always computed from image windows of dimensions 64 x 128, but we scale the image to detect humans who are closer and further from the camera.
- We move the detection window using some fixed strides, say stride = 1 cell, toward the right and down directions.
- We decrease the scale of the image from scale, say s =1.0, to a small value, say s =0.1 or so, where the entire image becomes smaller than 64 x 128.
- At a smaller scale, say s = 0.2, we are looking for humans who appear large in the original image as they are close to the camera.

Multi-scale detections

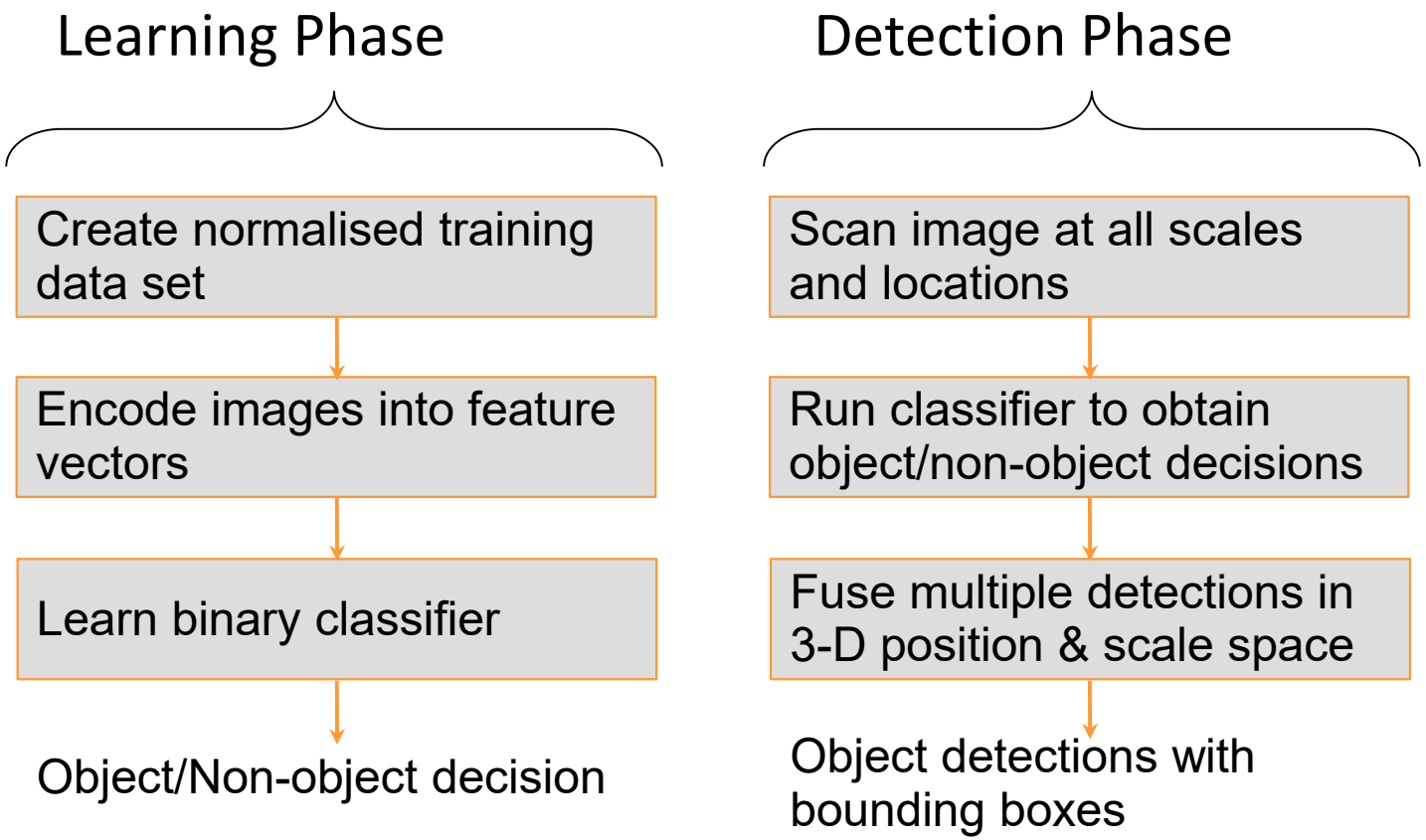
- Non-maxima suppression (NMS) – one possibility is to use greedy NMS. We select the best scoring window and discard other windows that have a significant overlap with this one. We repeat this.



After dense multi-scale scan of detection window

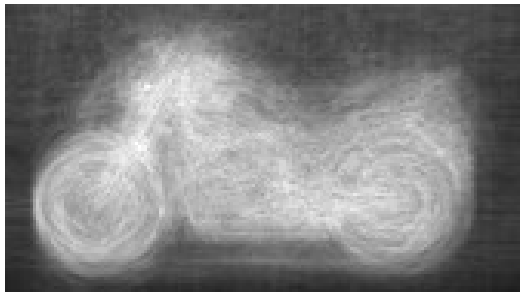
Final detections

Overall Architecture

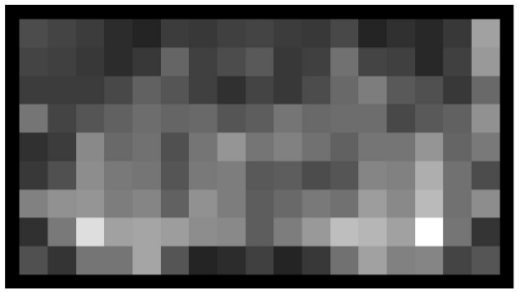


Slide courtesy : Navneet Dalal

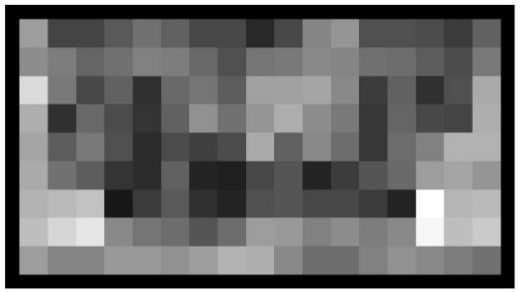
Descriptor Cues: Motorbikes



Average gradients

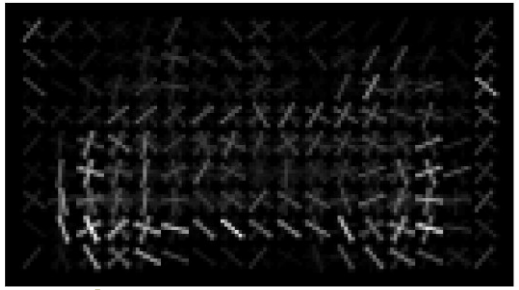


Weighted pos wts

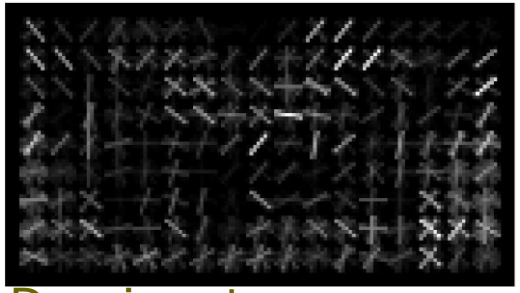


Weighted neg wts

Input window

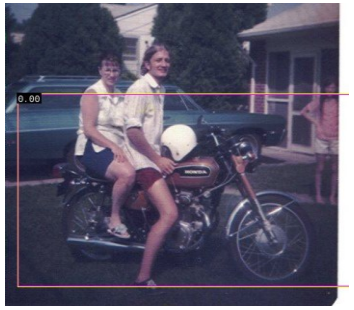


Dominant pos orientations



Dominant neg orientations

Detection Examples



Slide courtesy : Navneet Dalal

Key Descriptor Parameters

Class	Window Size	Avg. Size	# of Orient-ation Bins	Orientat-ion Range	Gamma Compre-ssion	Normal-isation Method
Person	64×128	Height 96	9	0˚–180˚	√RGB	L2-Hys
Car	104×56	Height 48	18	0˚–360˚	√RGB	L1-Sqrt
Bus	120×80	Height 64	18	0˚–360˚	√RGB	L1-Sqrt
Motorbike	120×80	Width 112	18	0˚–360˚	√RGB	L1-Sqrt
Bicycle	104×64	Width 96	18	0˚–360˚	√RGB	L2-Hys
Cow	128×80	Width 56	18	0˚–360˚	√RGB	L2-Hys
Sheep	104×60	Height 56	18	0˚–360˚	√RGB	L2-Hys
Horse	128×80	Width 96	9	0˚–180˚	RGB	L1-Sqrt
Cat	96×56	Height 56	9	0˚–180˚	RGB	L1-Sqrt
Dog	96×56	Height 56	9	0˚–180˚	RGB	L1-Sqrt

Slide courtesy : Navneet Dalal