Outline

1. Bag of visual words model for categorization
   - SVM classifier
2. Adding spatial information for localization
3. Databases and challenges
4. Spatial layout
5. Class based segmentation
   - Pixel level localization
6. Conclusions and the future

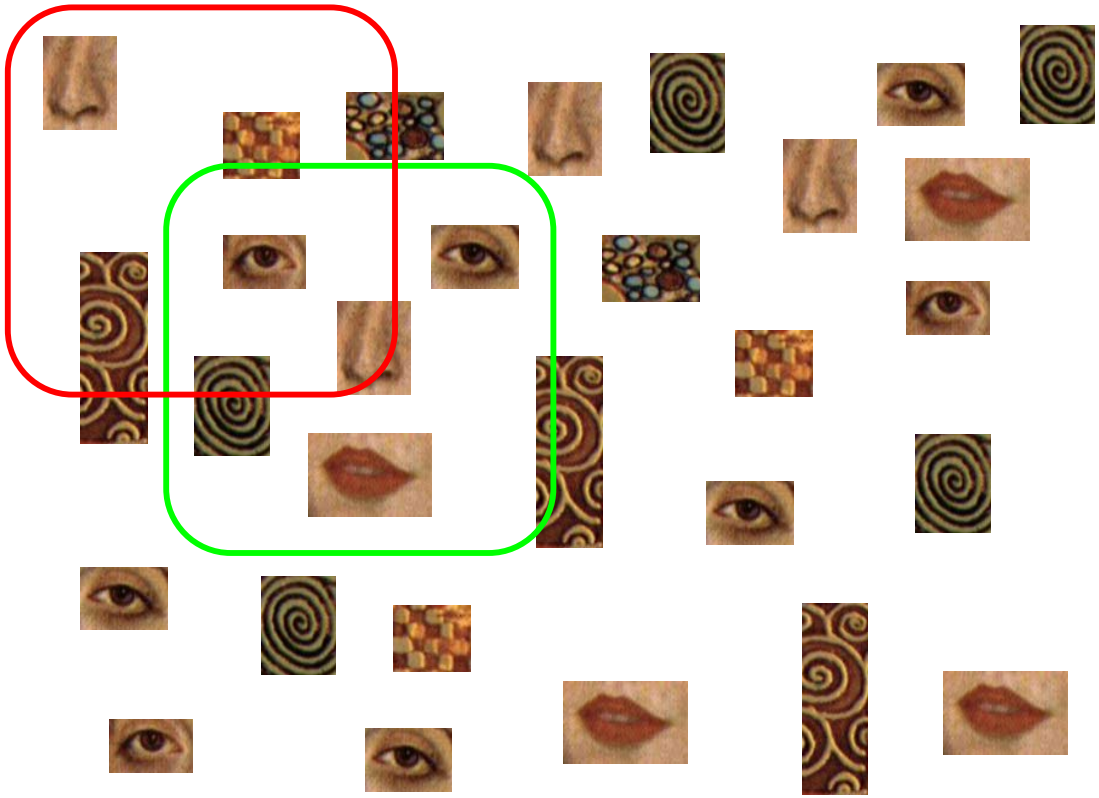
Beyond a bag of visual words

Outline

- Region of Interest (ROI)
  - jumping/sliding window for localization
- Spatial tiling
- Histogram of Gradients (HOG)
- Spatial pyramid
- Case study

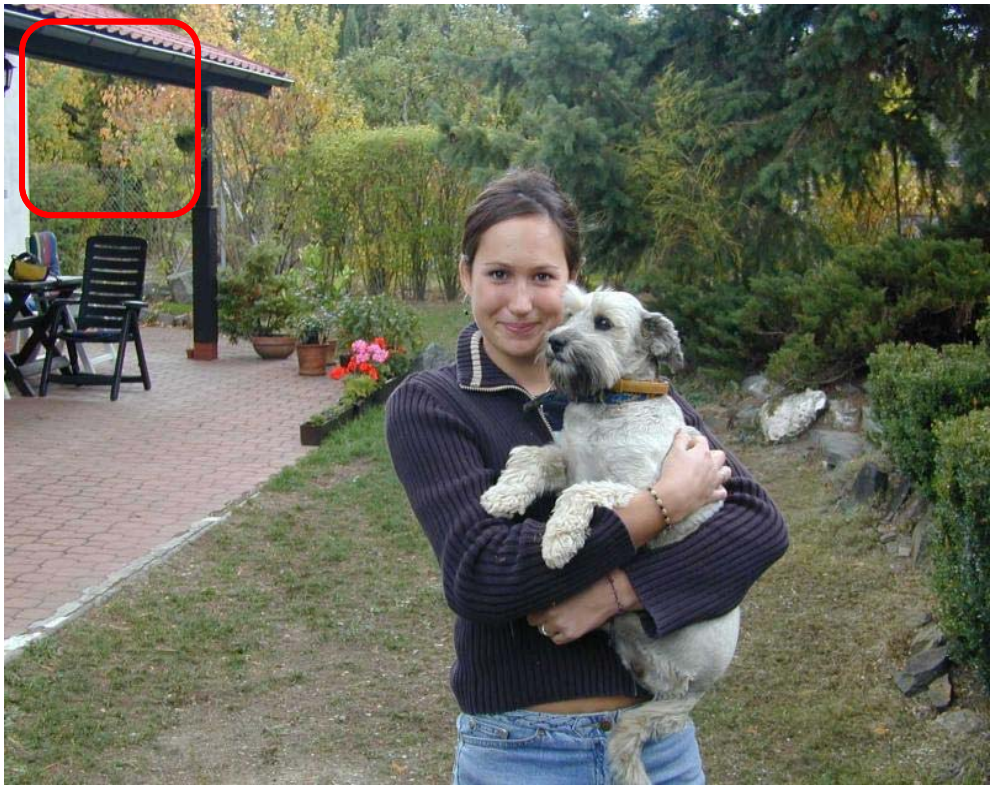
Region of Interest (ROI)

- Problem of background clutter
- Use a sub-window
  - At correct position, no clutter is present

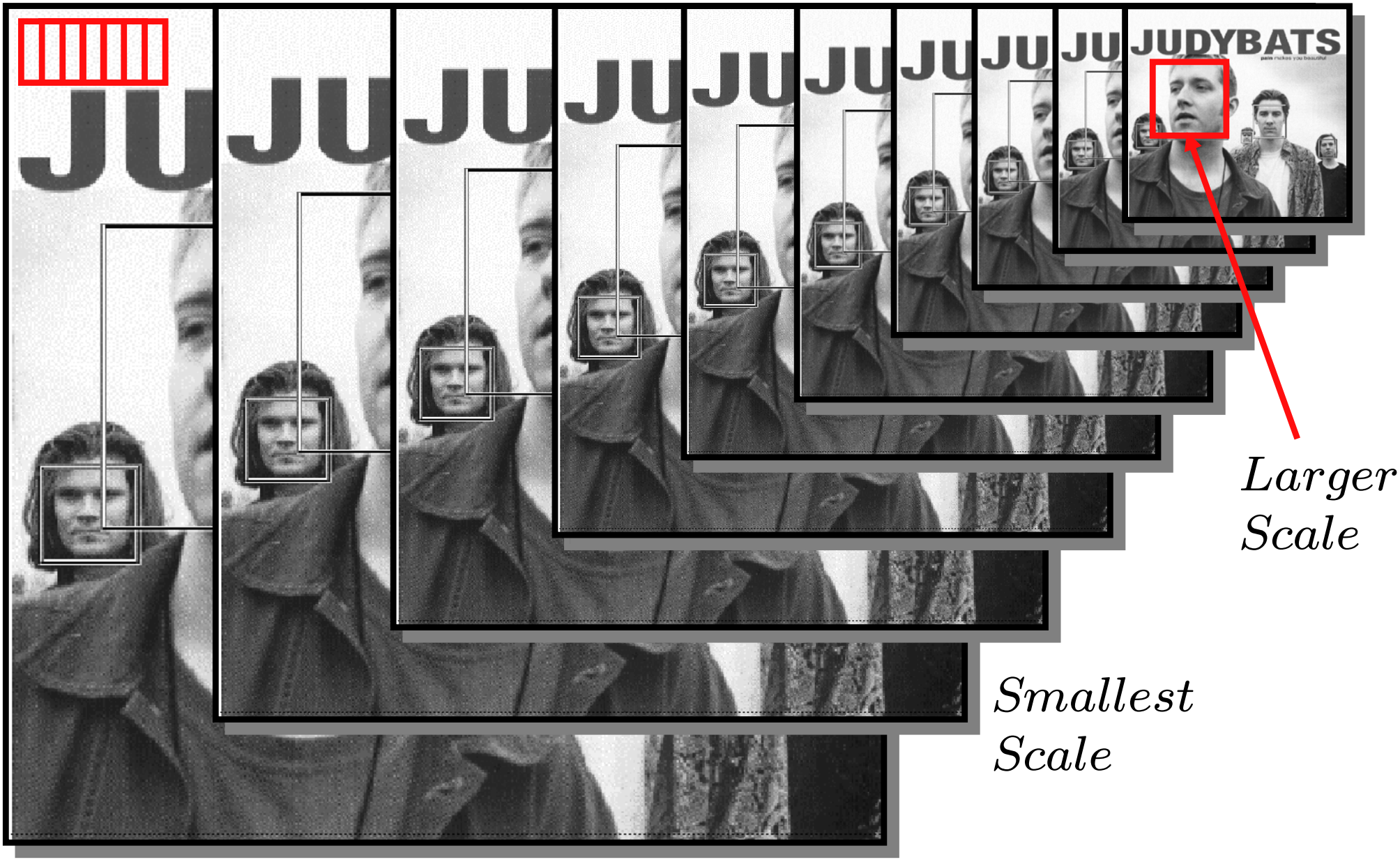


Sliding window detection

- Scale / orientation range to search over
- Speed
- Context

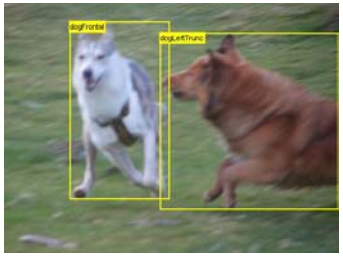


search over scale



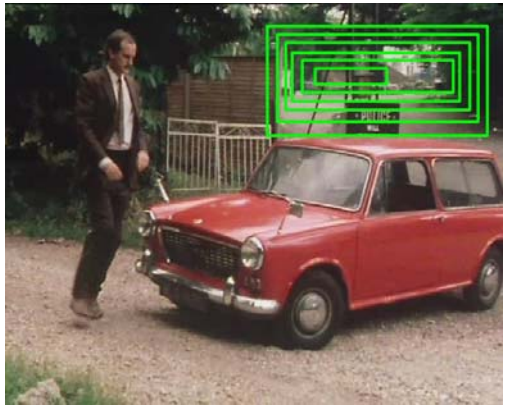
Problems with sliding windows …

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



See recent work by

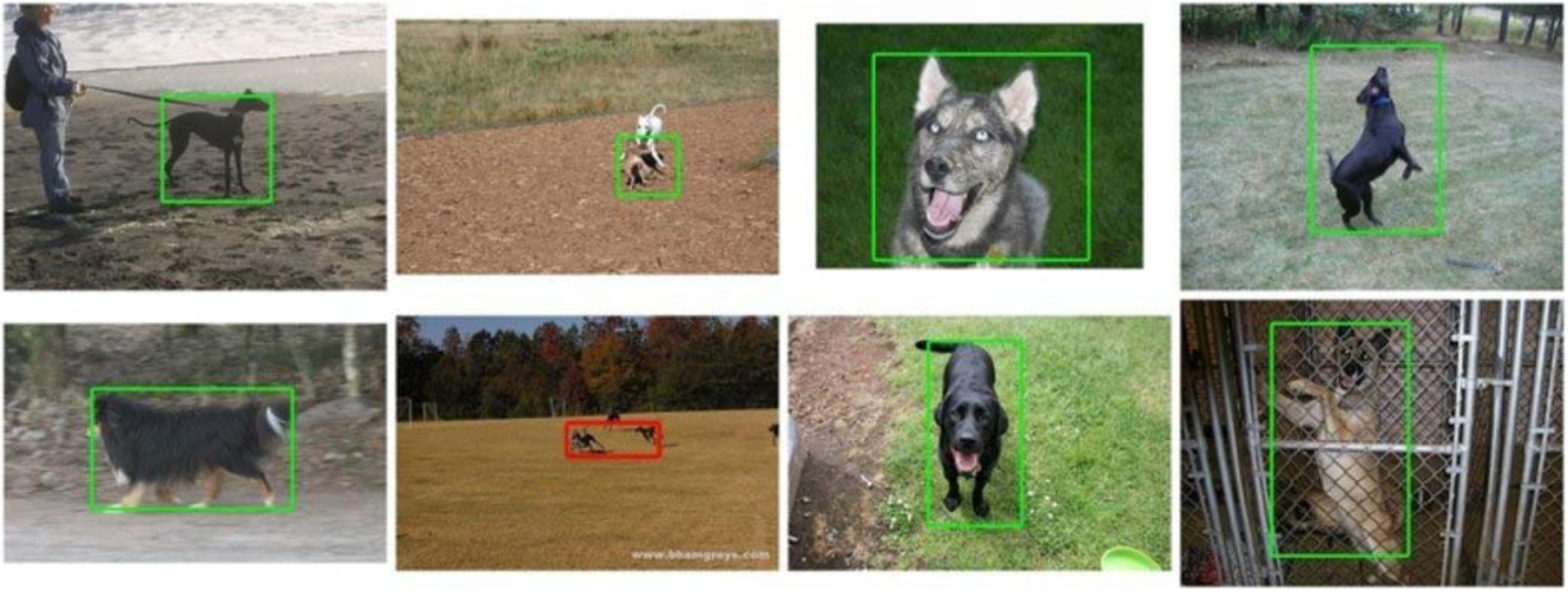
- Christoph Lampert et al CVPR 08, ECCV 08
- Bosch et al BMVC 08



Sliding window

- Classifier: SVM with linear kernel
- bag of visual word representation of ROI
- Stronger training: ROI on object instance

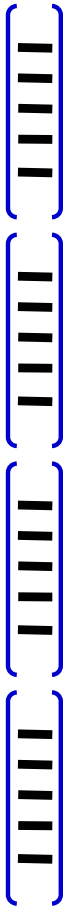
Example detections for dog



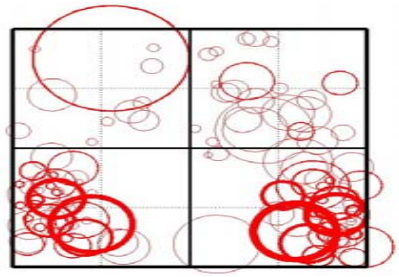
Lampert et al CVPR 08

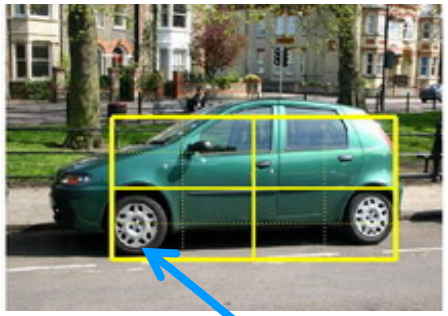
More spatial information - tiling

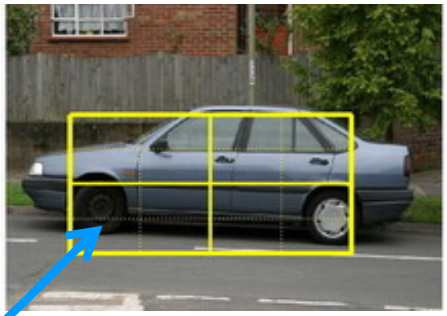
Use spatial grid to define correspondence

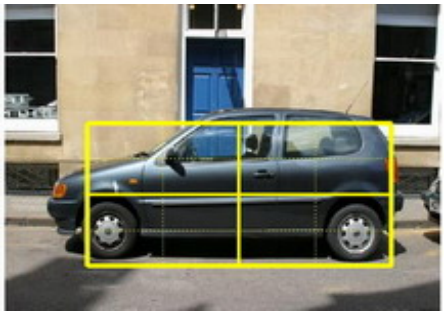


- parameter: number of tiles







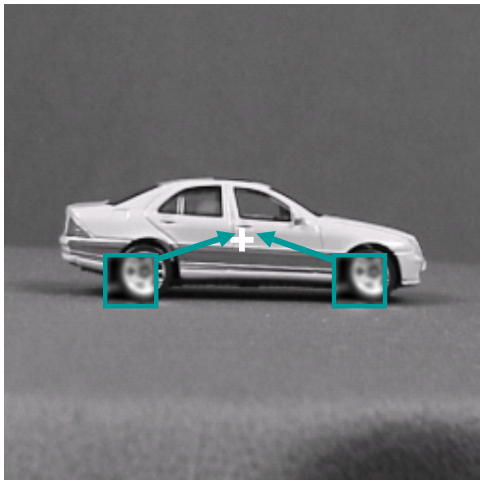


If codebook has V visual words, then representation has dimension 4V

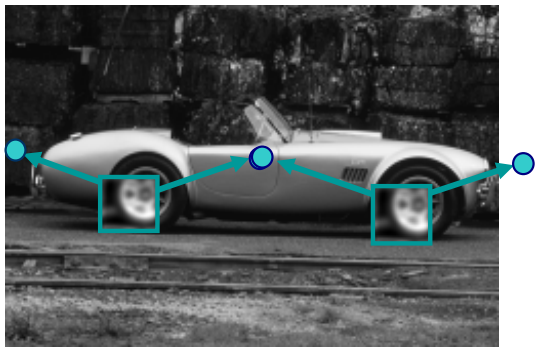
Fergus et al ICCV 05

Ex: Leibe & Schiele 03/04 : Generalized Hough Transform

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

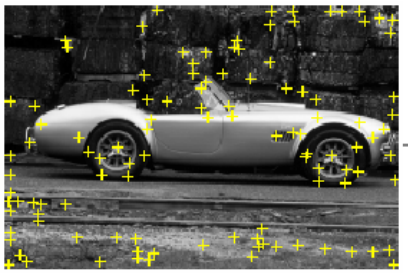


- Recognition: for new image, let the matched patches vote for possible object positions

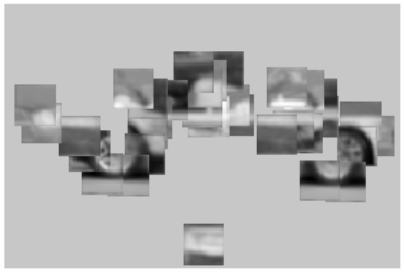


Ex: Leibe & Schiele 03/04 : Generalized Hough Transform

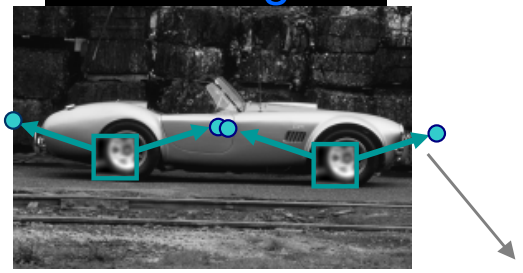
Interest Points

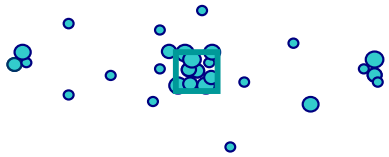


Matched Codebook Entries

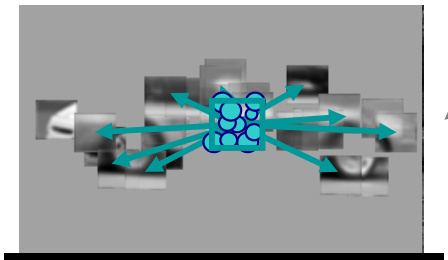


Probabilistic Voting



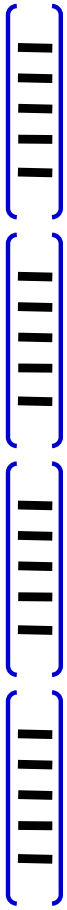


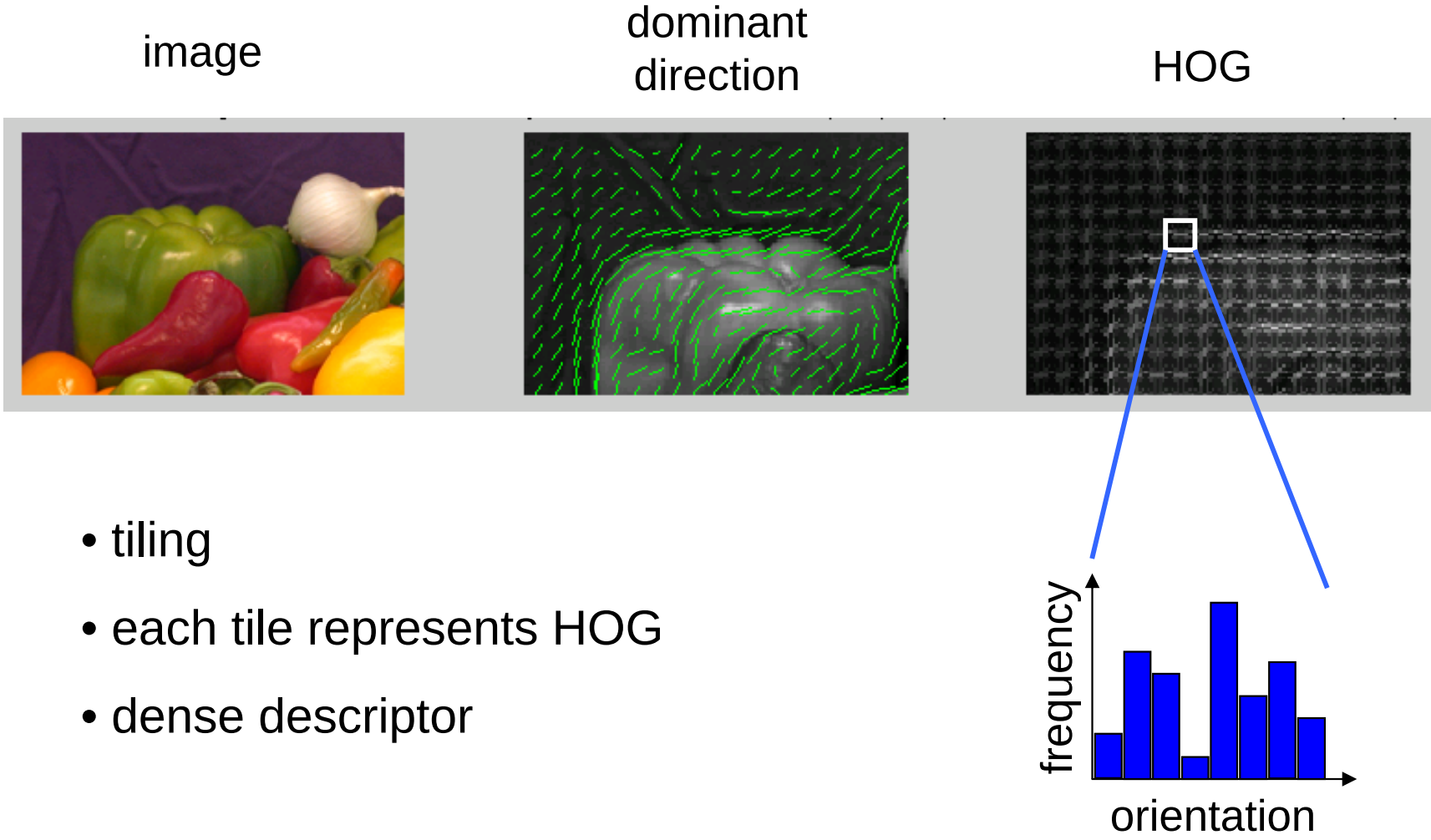
Voting Space (continuous)



Backprojection of Maximum

More features – histogram of orientations





- tiling
- each tile represents HOG
- dense descriptor

Counts in orientation bins can be thought of as visual words

Ex 1: Human (Pedestrian) Detection

Histograms of Oriented Gradients for Human Detection
Dalal & Triggs, CVPR 2005

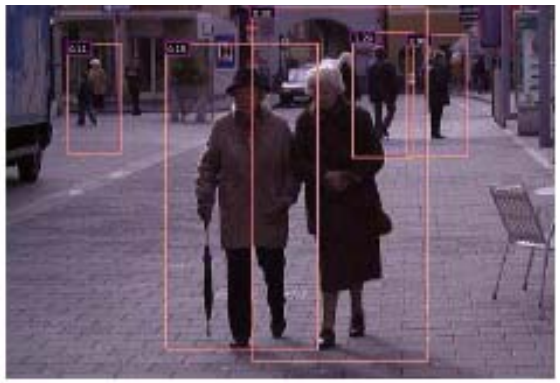
Detect & localize upright people in static images

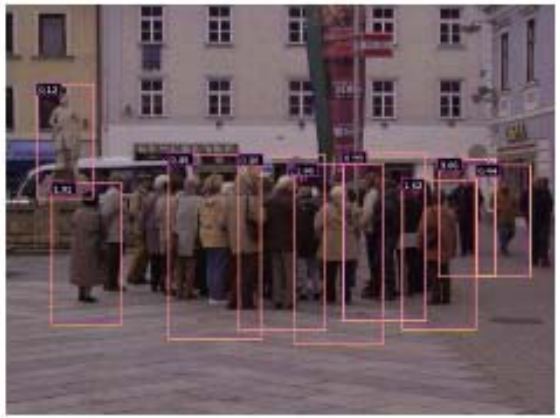
Challenges

- Wide variety of articulated poses
- Variable appearance/clothing
- Complex backgrounds
- Unconstrained illumination
- Occlusions, different scales

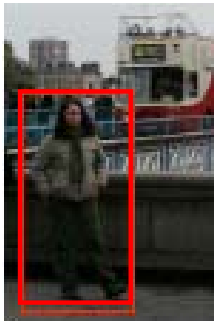
Applications

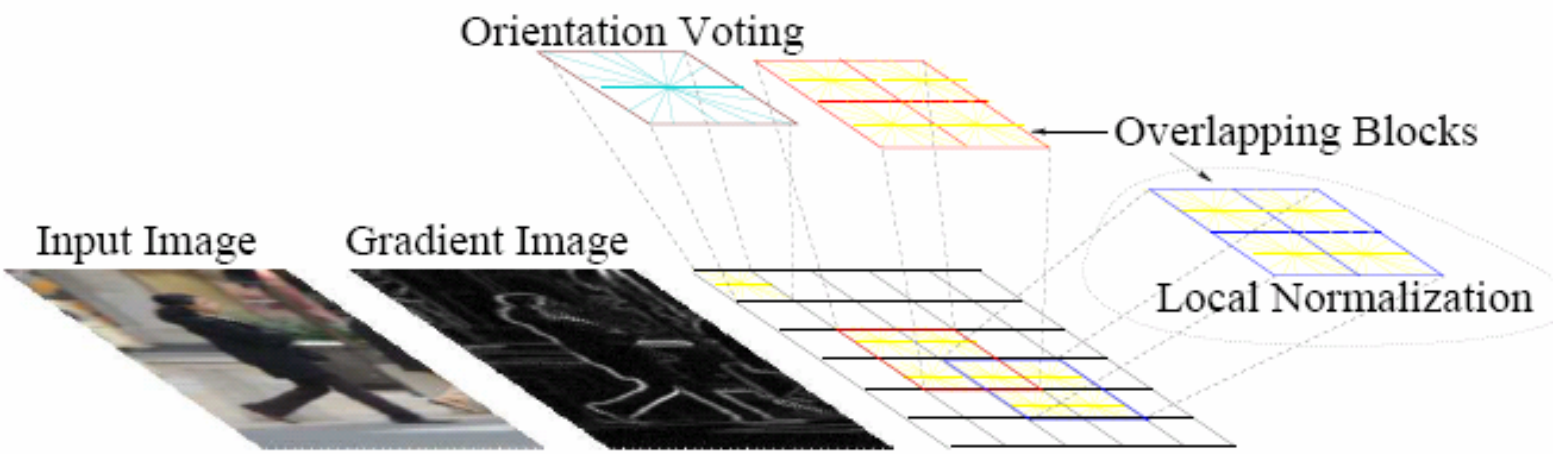
- Pedestrian detection for smart cars
- Film & media analysis
- Visual surveillance

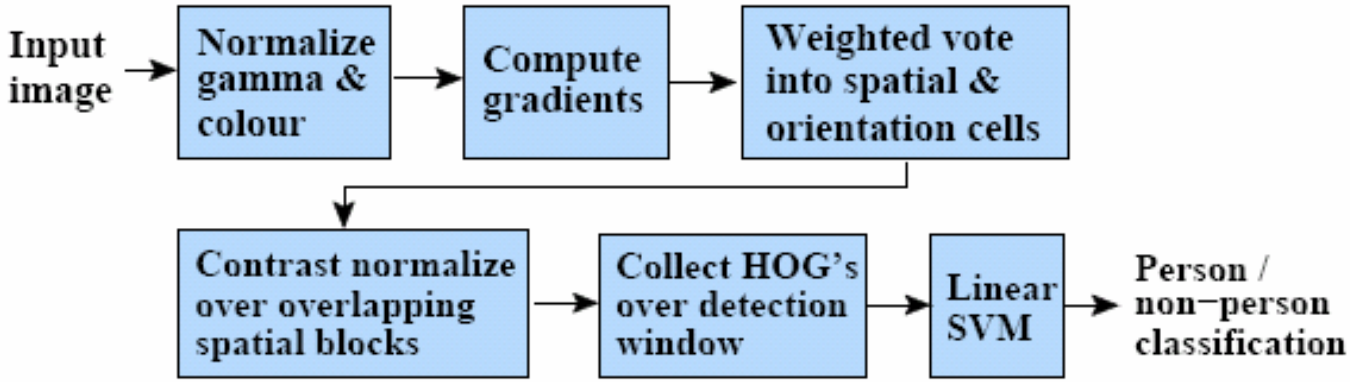




- training: ROI over pedestrian
- classification: linear SVM on HOG
- NB similarity to SIFT, GIST



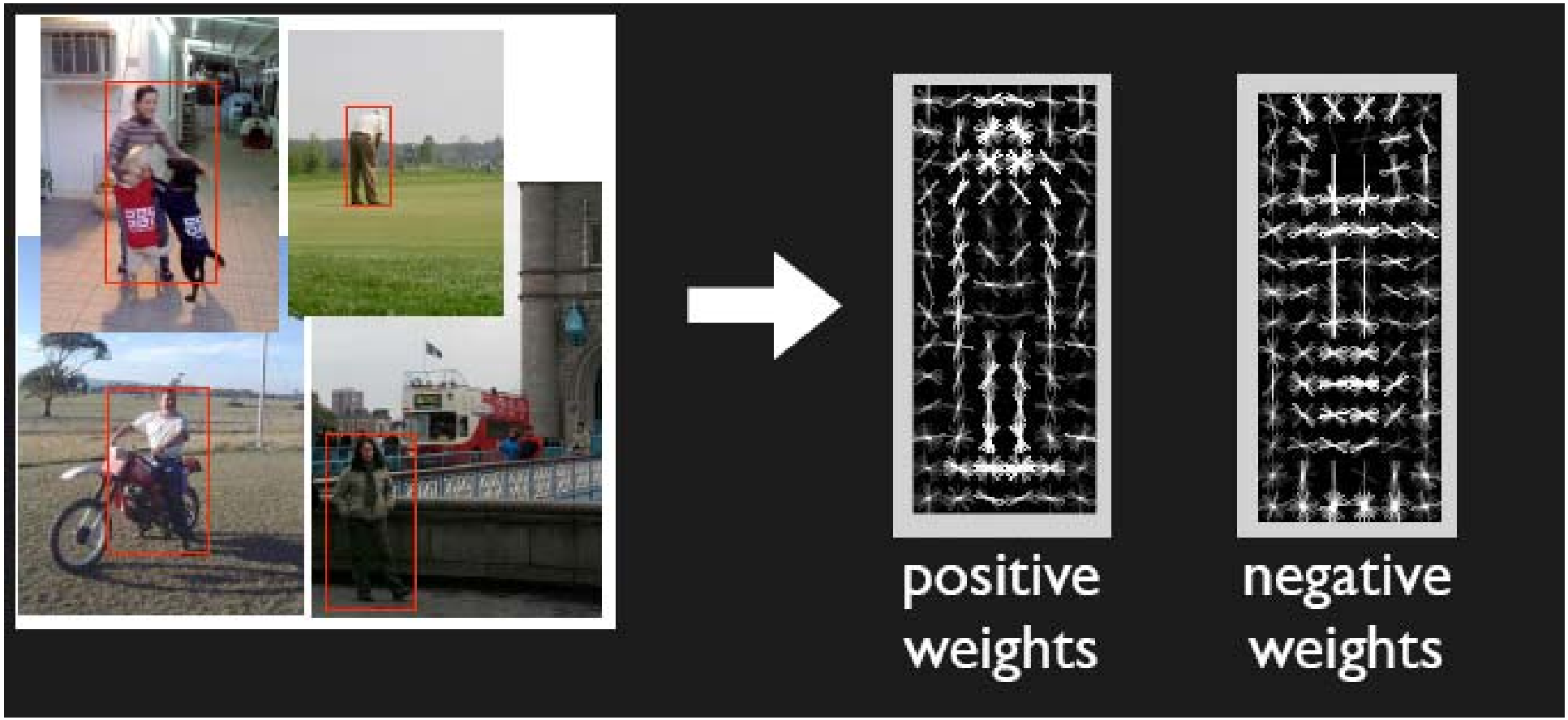




Dalal and Triggs, CVPR 2005

Learned model

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



Slide from Deva Ramanan

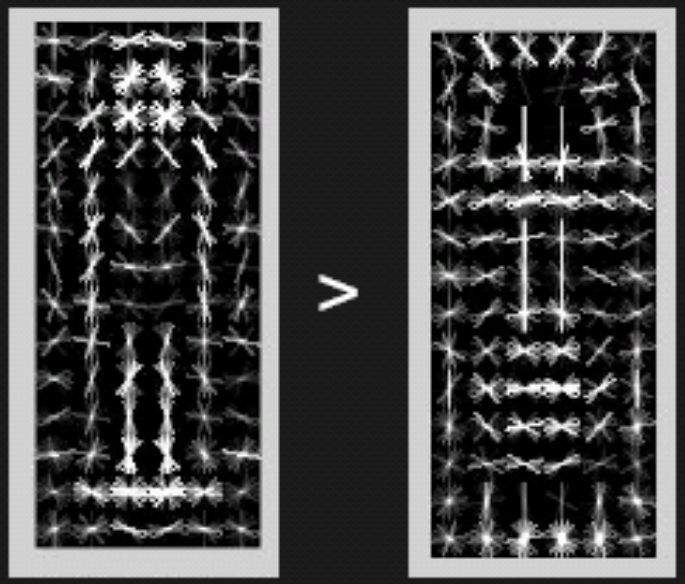
What do negative weights mean?

$$wx > 0$$

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

$$w_+ > w_-x$$

pedestrian model



>

pedestrian background model

Complete system should compete pedestrian/pillar/doorway models

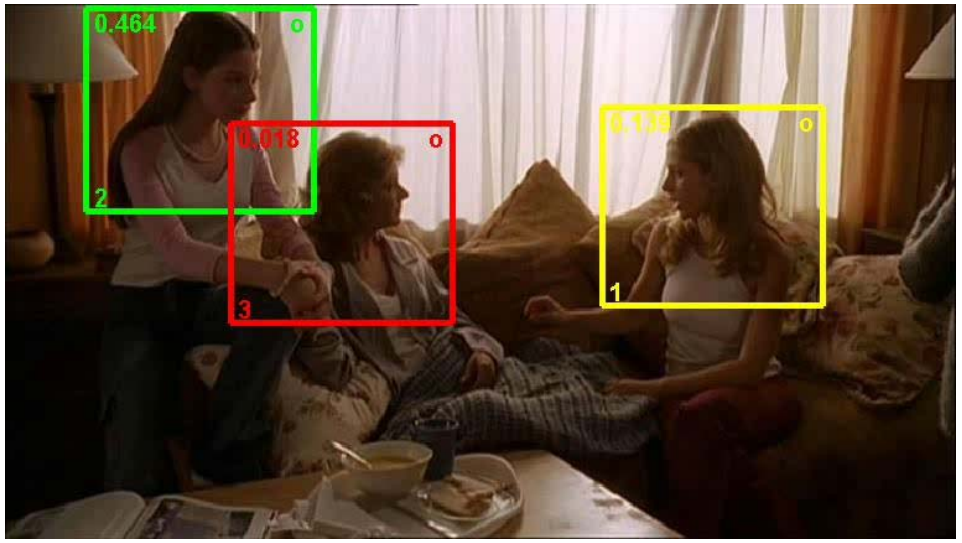
Discriminative models come equipped with own bg

(avoid firing on doorways by penalizing vertical edges)

Slide from Deva Ramanan

Ex 2: Upper body detector – using HOGs

average training data

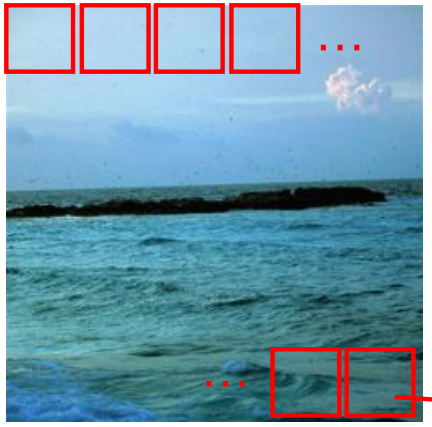


- Ferrari et al CVPR 08

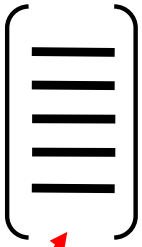
More features – dense visual words

DENSE PATCHES

Textons



Parameters: N – size of patch
M – distance between patches

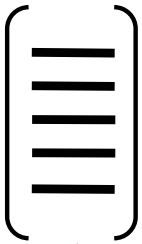


Row reorder gray values and form a vector of size N^2

Luong & Malik 1999,
Varma & Zisserman 2003

SIFT

Parameters: r – radi of patch
M – distance between patches



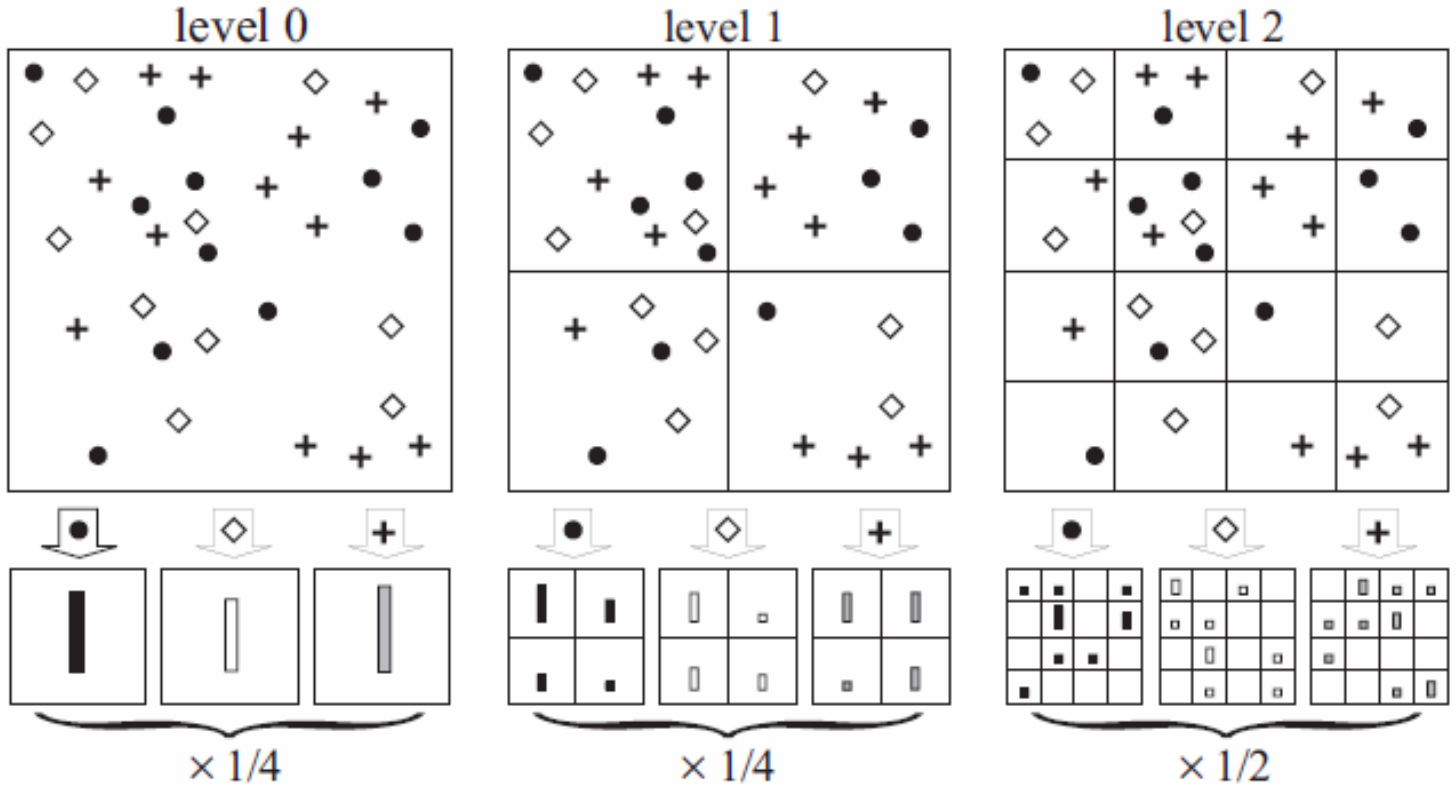
128- SIFT descriptor

Vogel & Schiele 2004,
Jurie & Triggs ICCV 05 ,
Fei-Fei & Perona CVPR 05,
Bosch et al ECCV 06

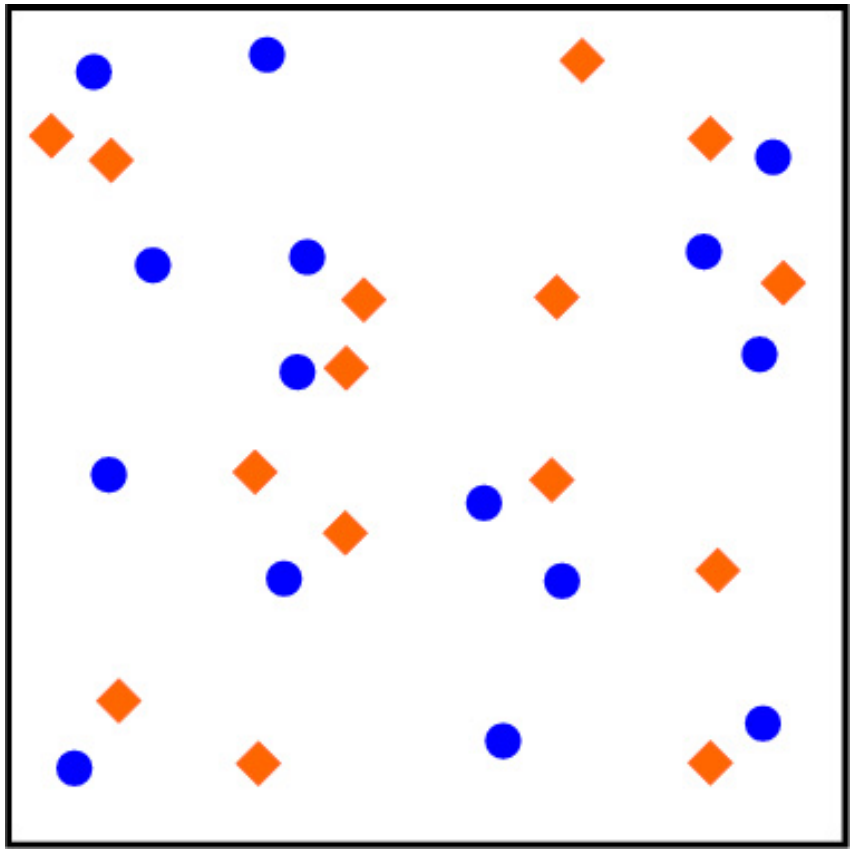
More Spatial information – Pyramid kernels

Lazebnik *et al.* [CVPR 2006]

- Divide image into grids of varying resolution, and give more weight to agreement in finer grids.
  - 2^l grids at level l
- Intersect histograms, multiply by weight.

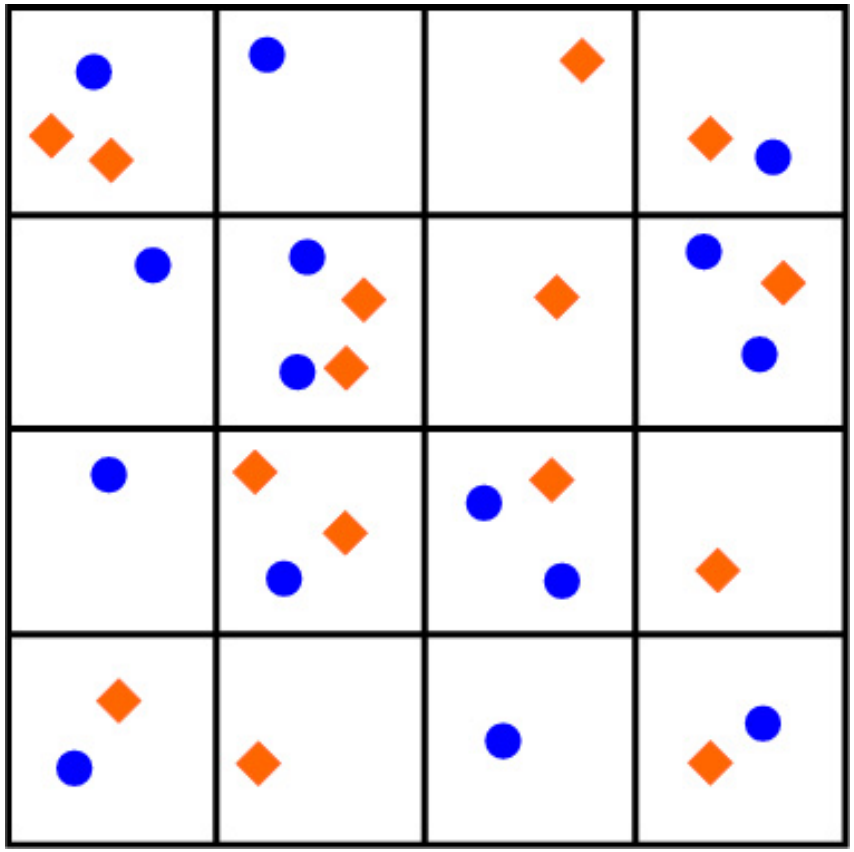


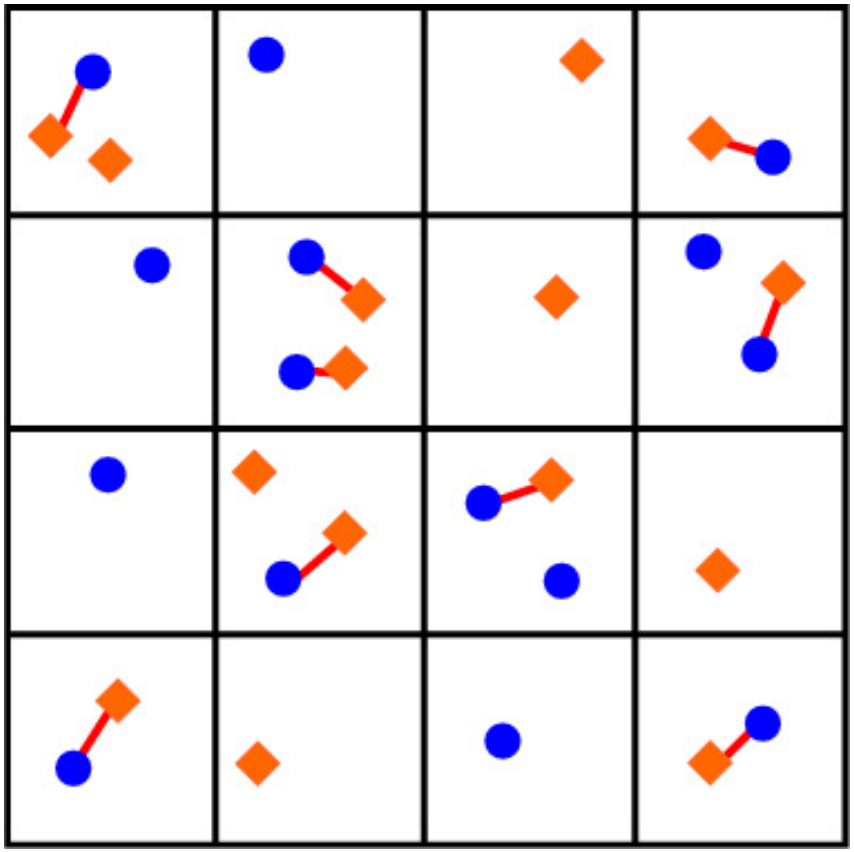
More Spatial information – Pyramid kernels



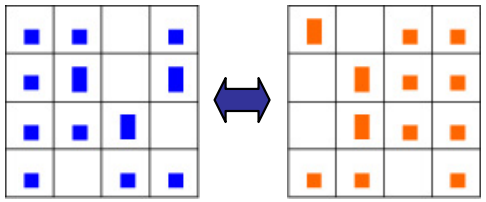
Spatial Pyramid Kernels for Geometry/Appearance Matching
(Lazebnik et al CVPR'06)

Based on Grauman & Darrell ICCV 05

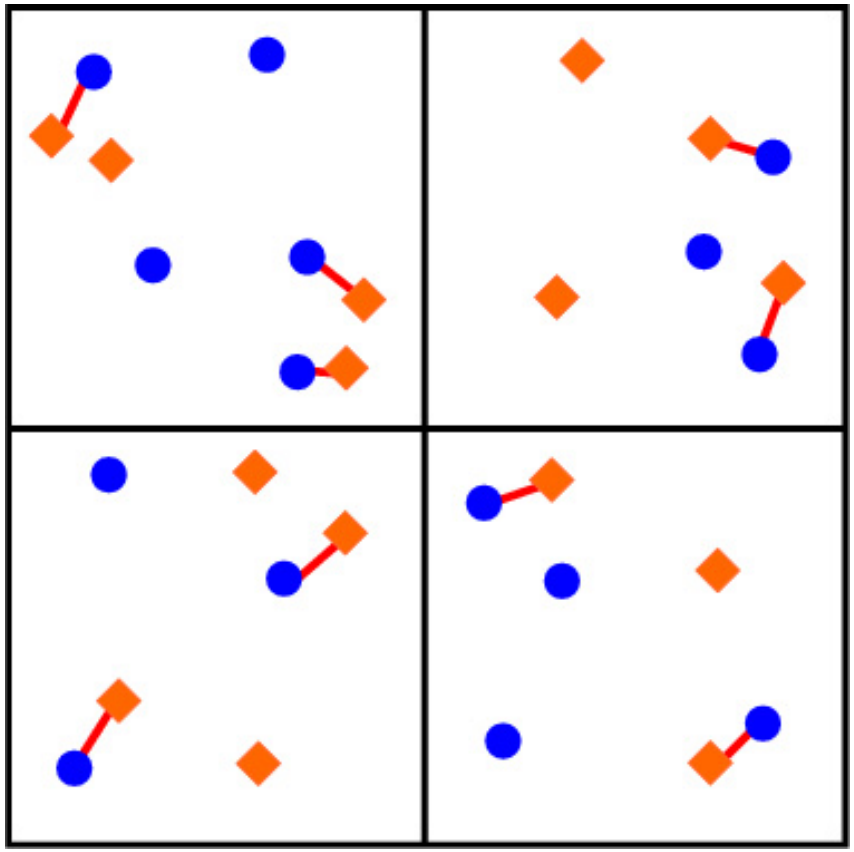




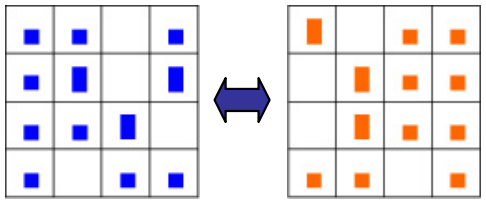
9 matches x 1



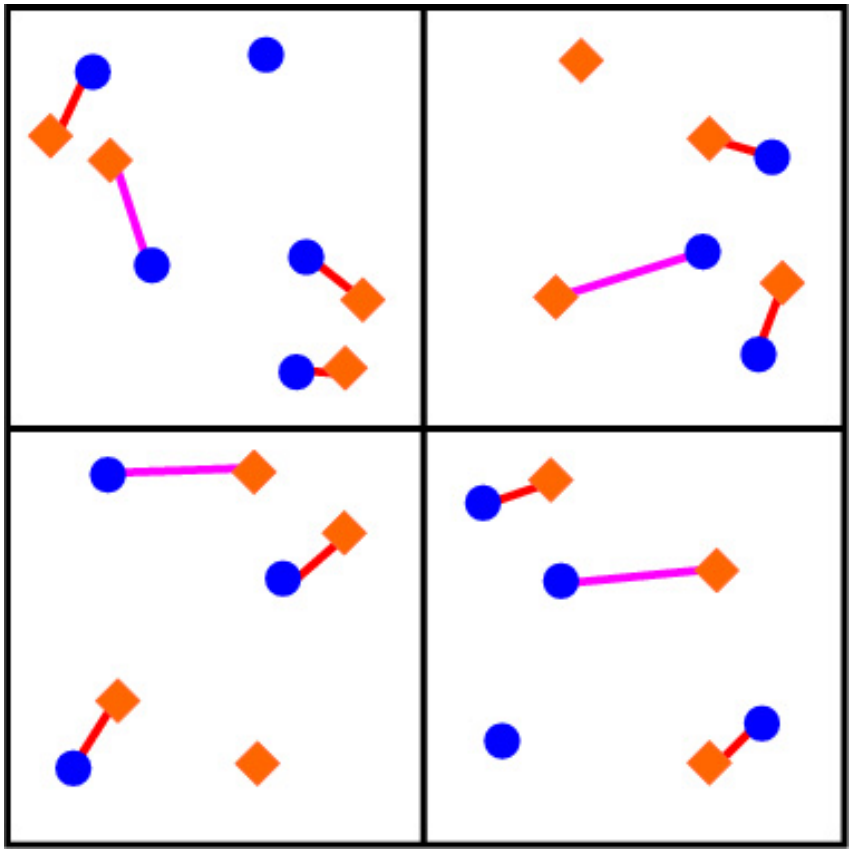
= 9



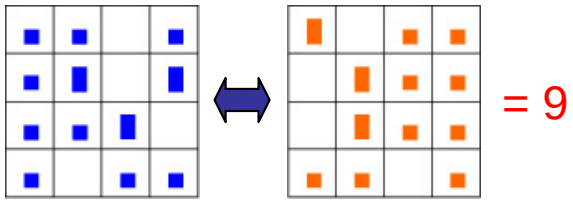
9 matches x 1



= 9

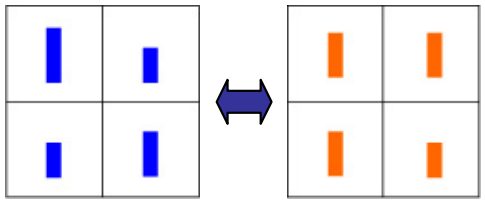


9 matches x 1

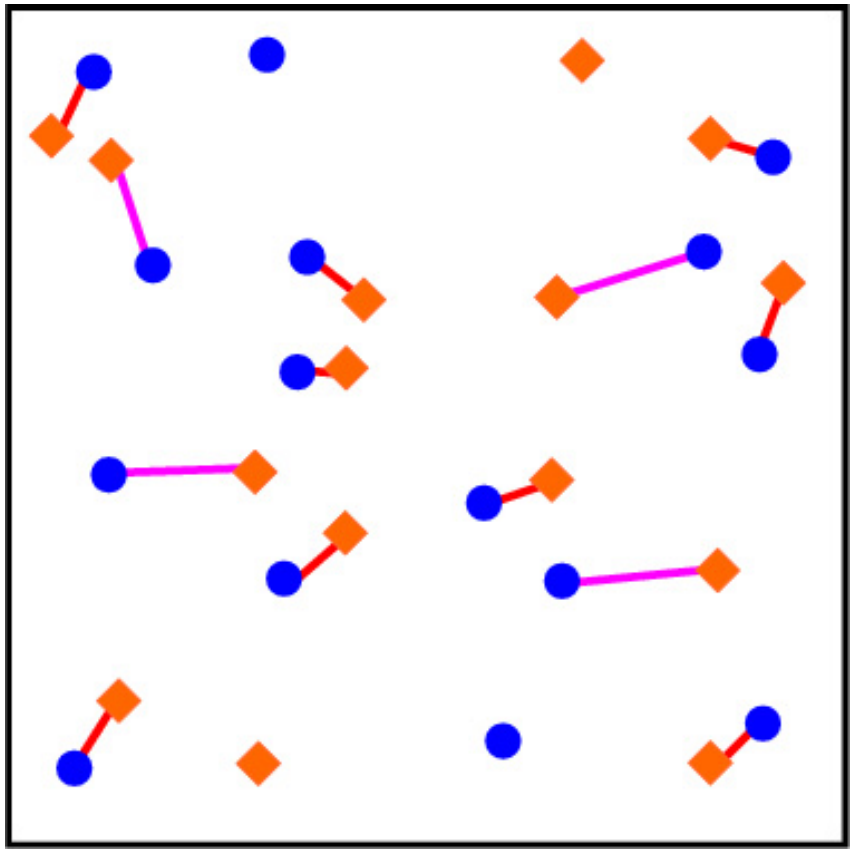


= 9

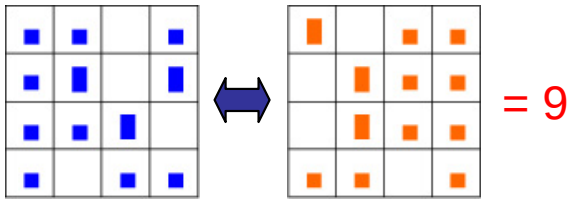
4 matches x ½



= 2

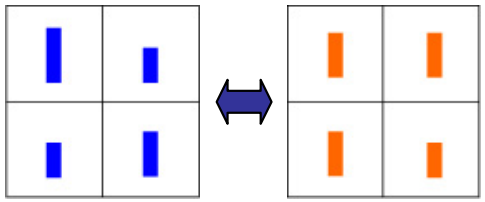


9 matches x 1

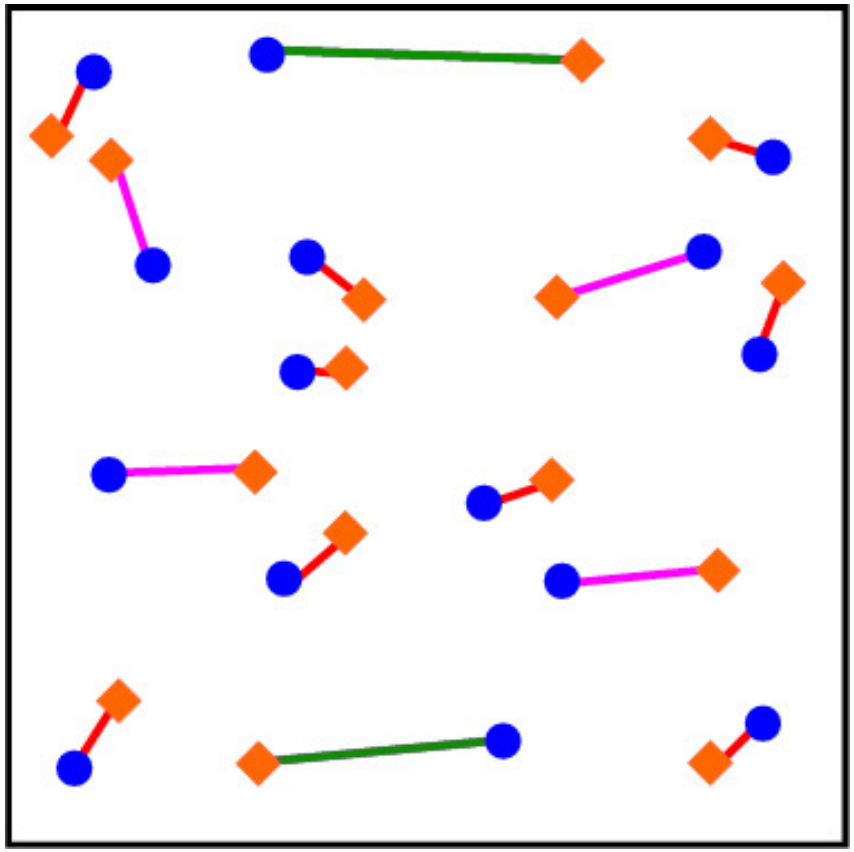


= 9

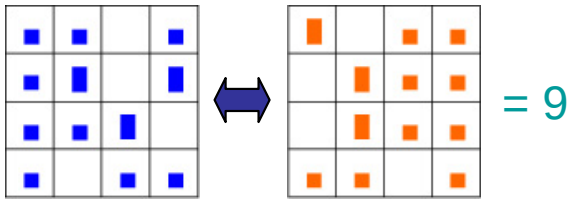
4 matches x ½



= 2

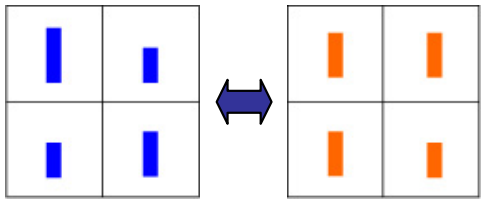


9 matches x 1



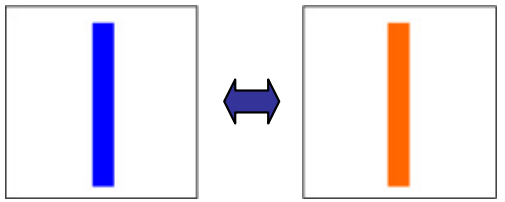
= 9

4 matches x ½



= 2

2 matches x ¼

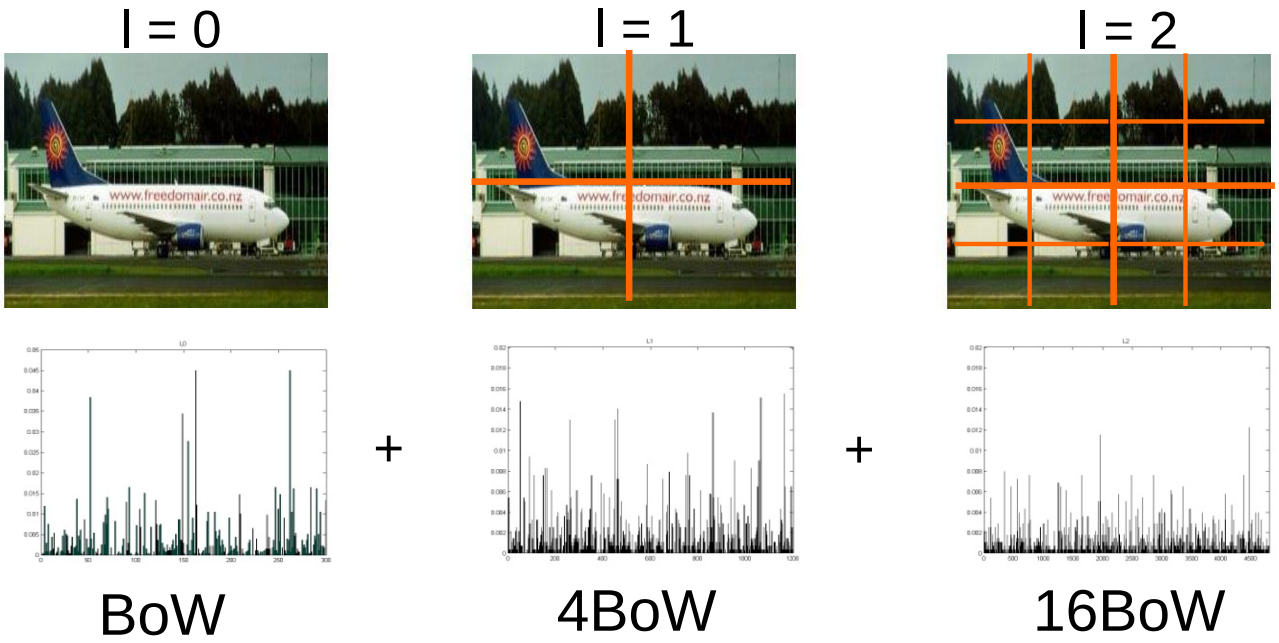


= 1/2

Total matching weight (value of *spatial pyramid kernel*): 9 + 2 + 0.5 = 11.5

Pyramid spatial layout for appearance patches – for images

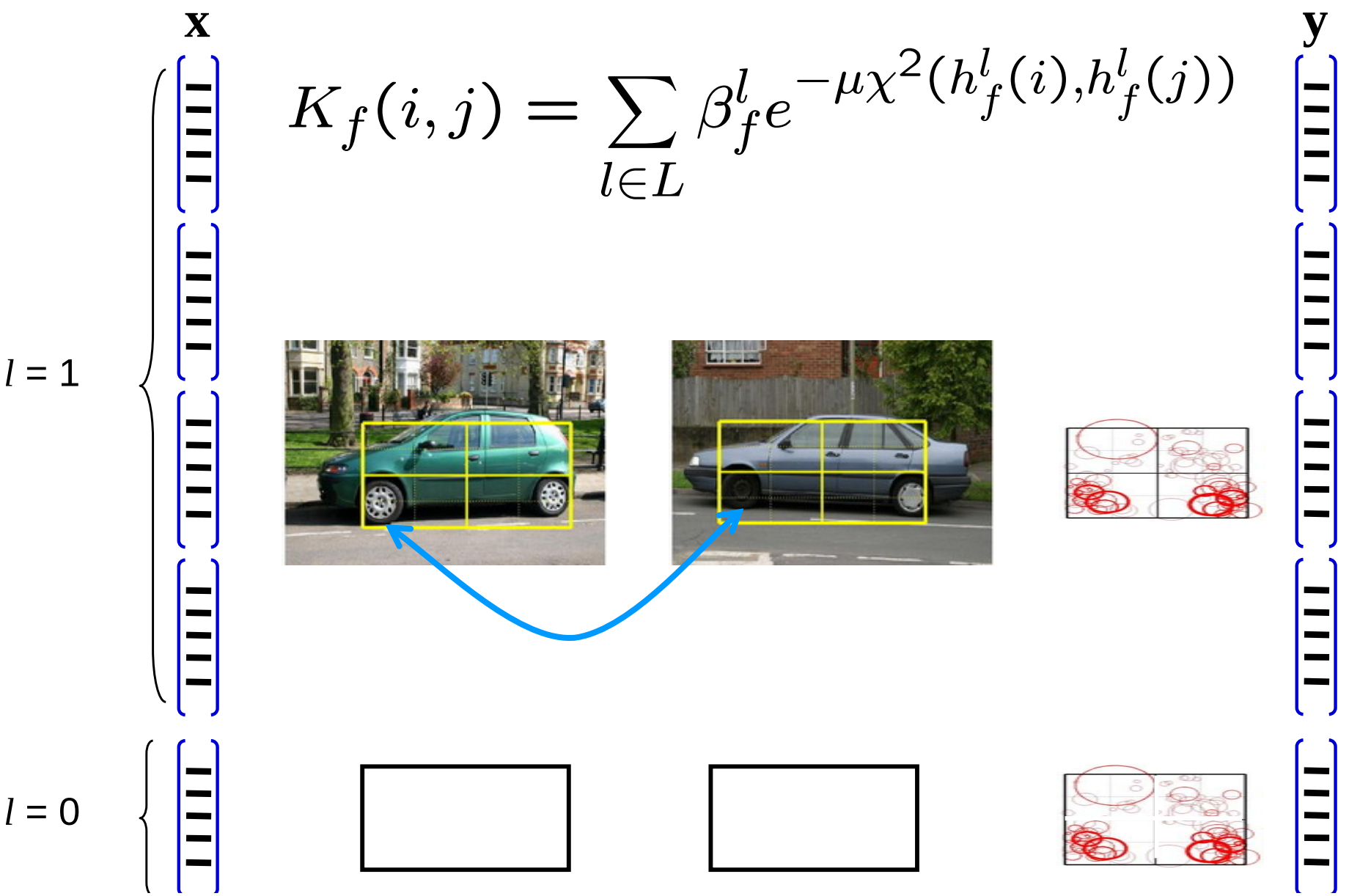
Represent appearance as dense grid of visual words



$$\kappa^L(X,Y) = \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}}\mathcal{I}^\ell$$

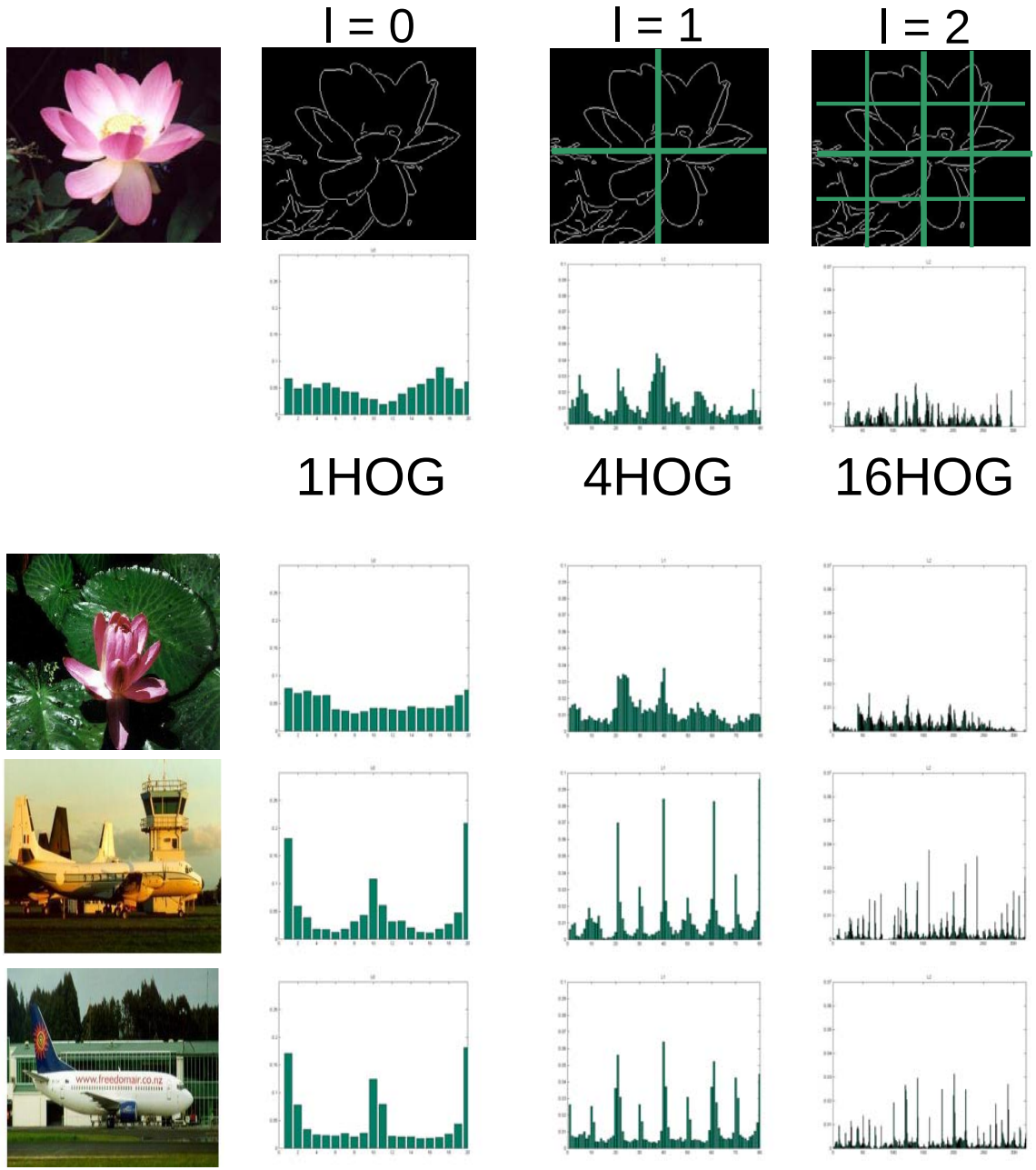
Generalizations

- Use chi-squared kernel instead of histogram intersection

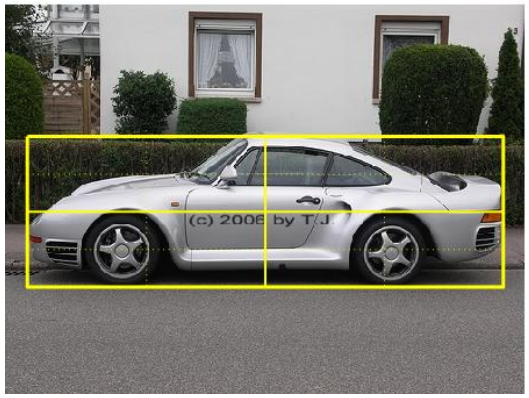


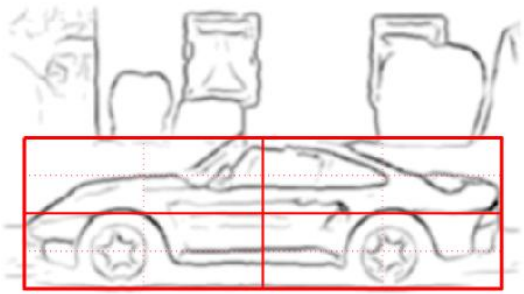
Pyramid HOG – for images

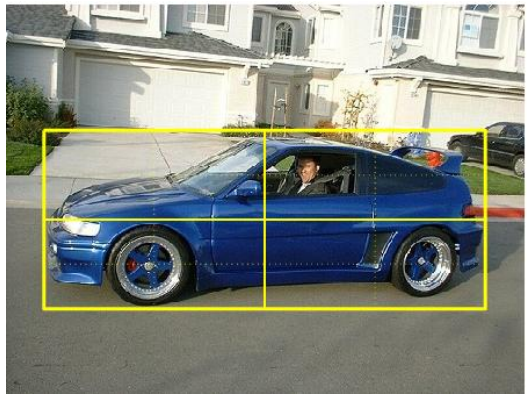
Represent
local orientated
gradients

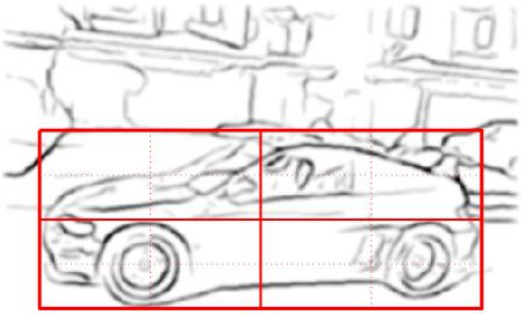


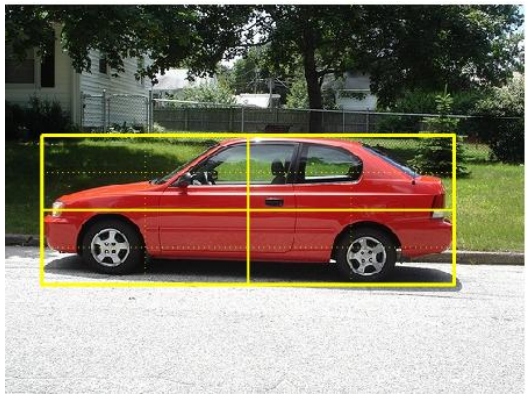
Pyramid HOG for image regions

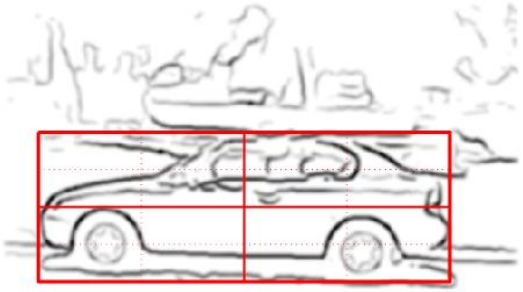


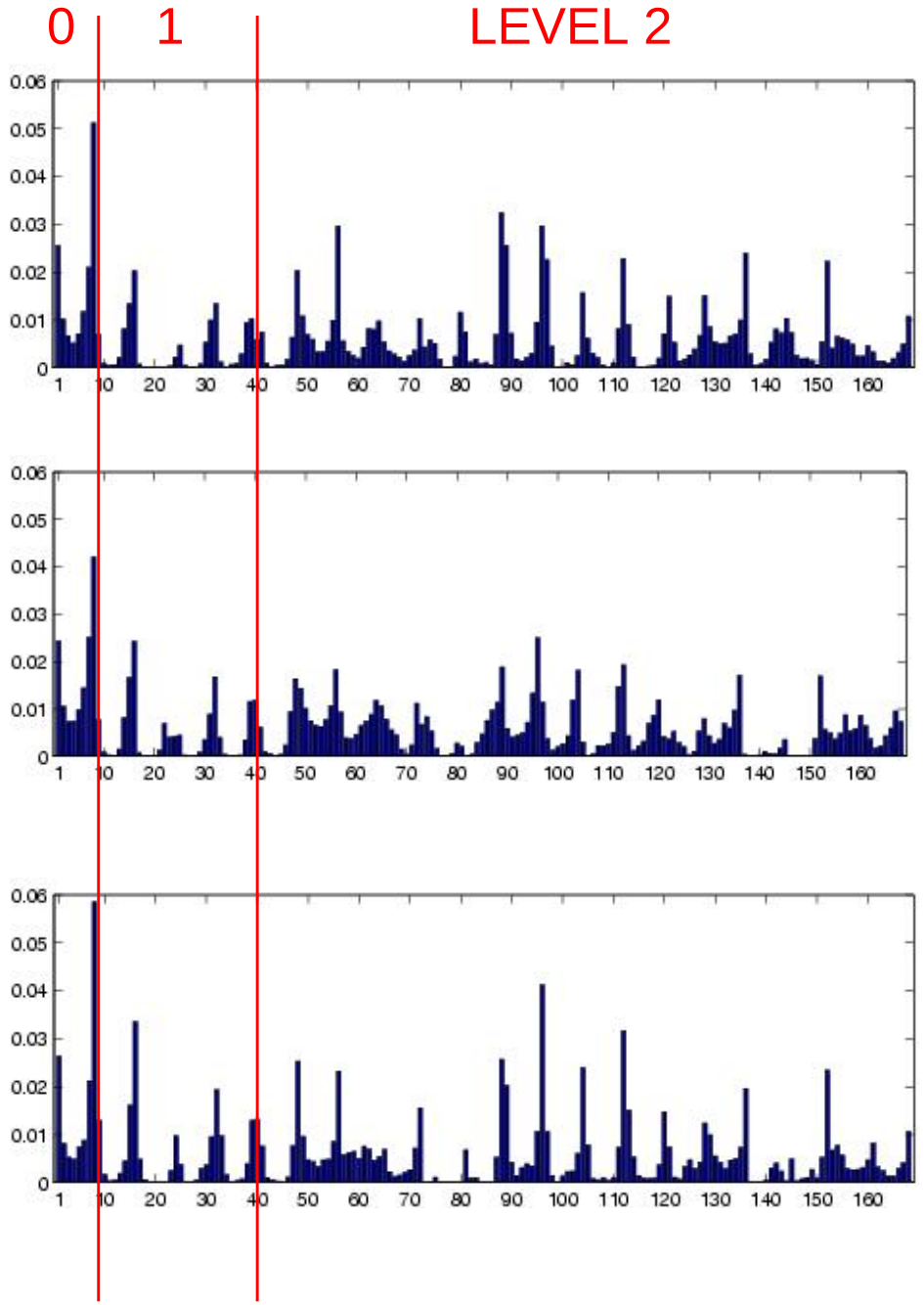




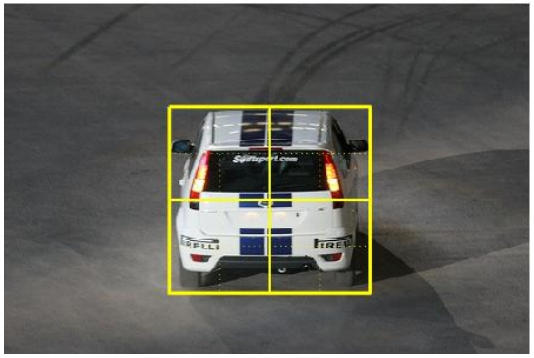


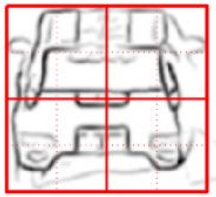


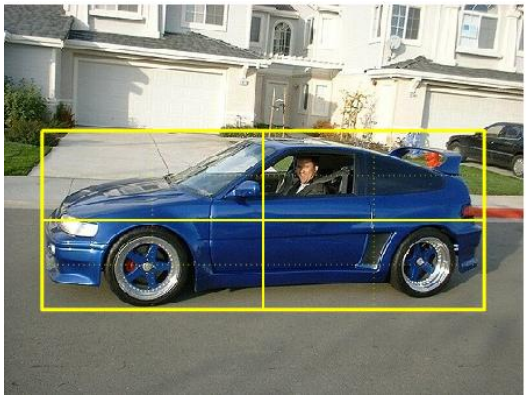


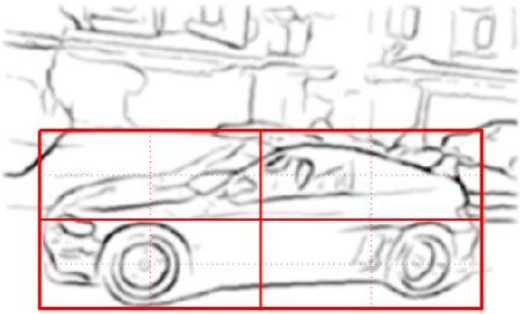


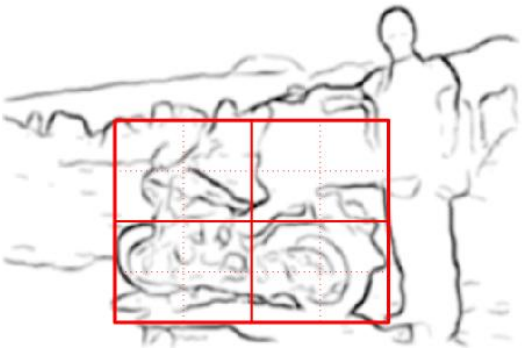
Pyramid HOG for image regions

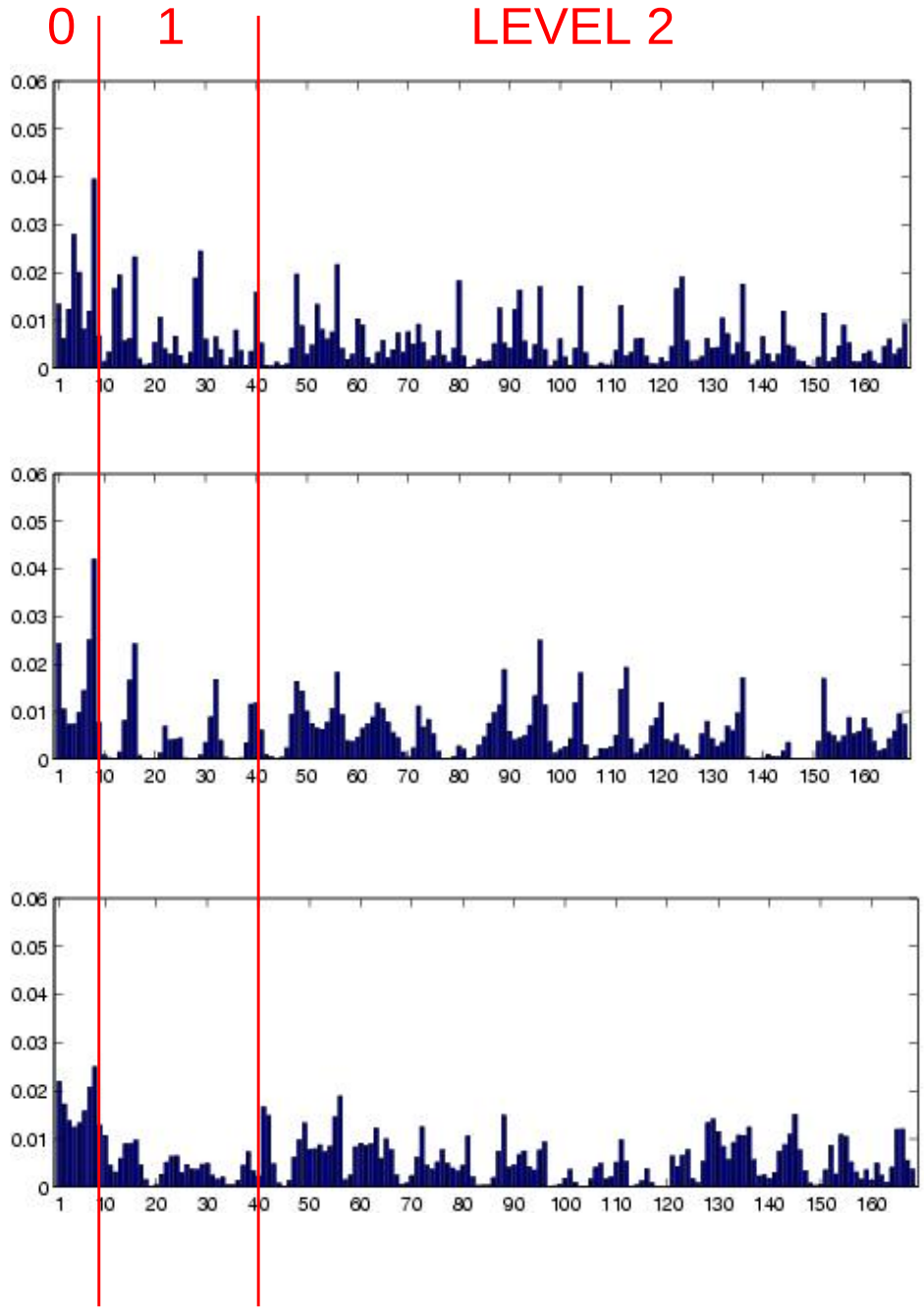






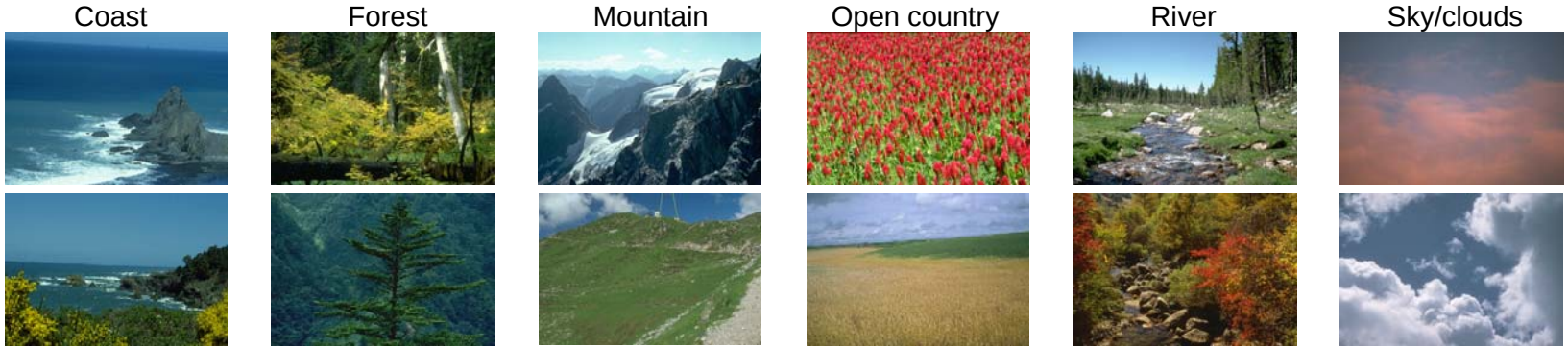




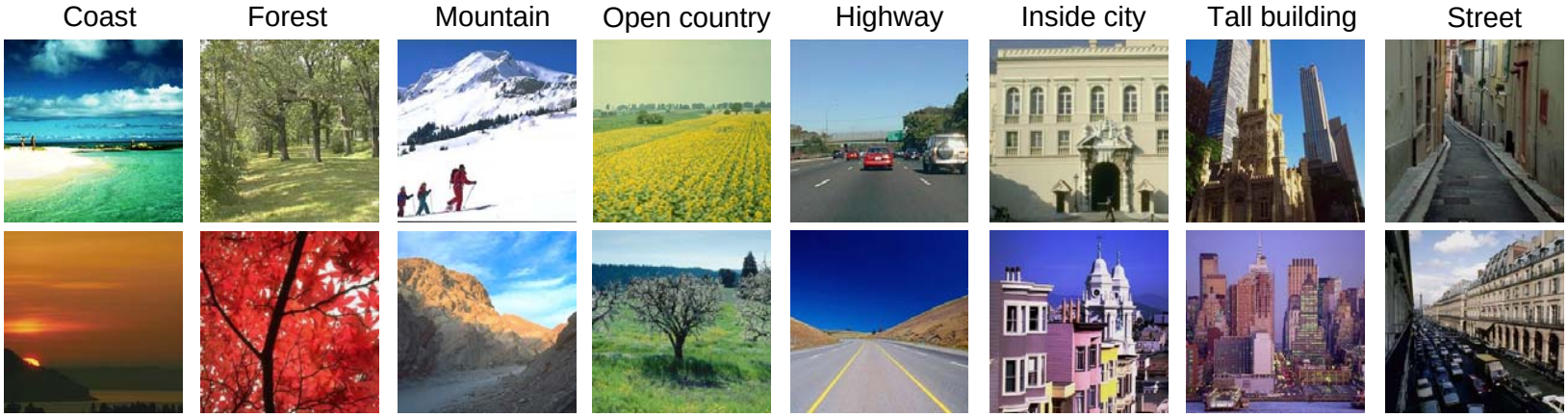


Case study: scene classification

Vogel & Schiele - **VS**



Oliva & Torralba - **OT**



Fei Fei & Perona - **FP**

Lazebnik et al. - **LSP**

Vogel & Schiele DATASET

702 images
6 categories

VS dataset

Coast

Forest

Mountain

Open country

River

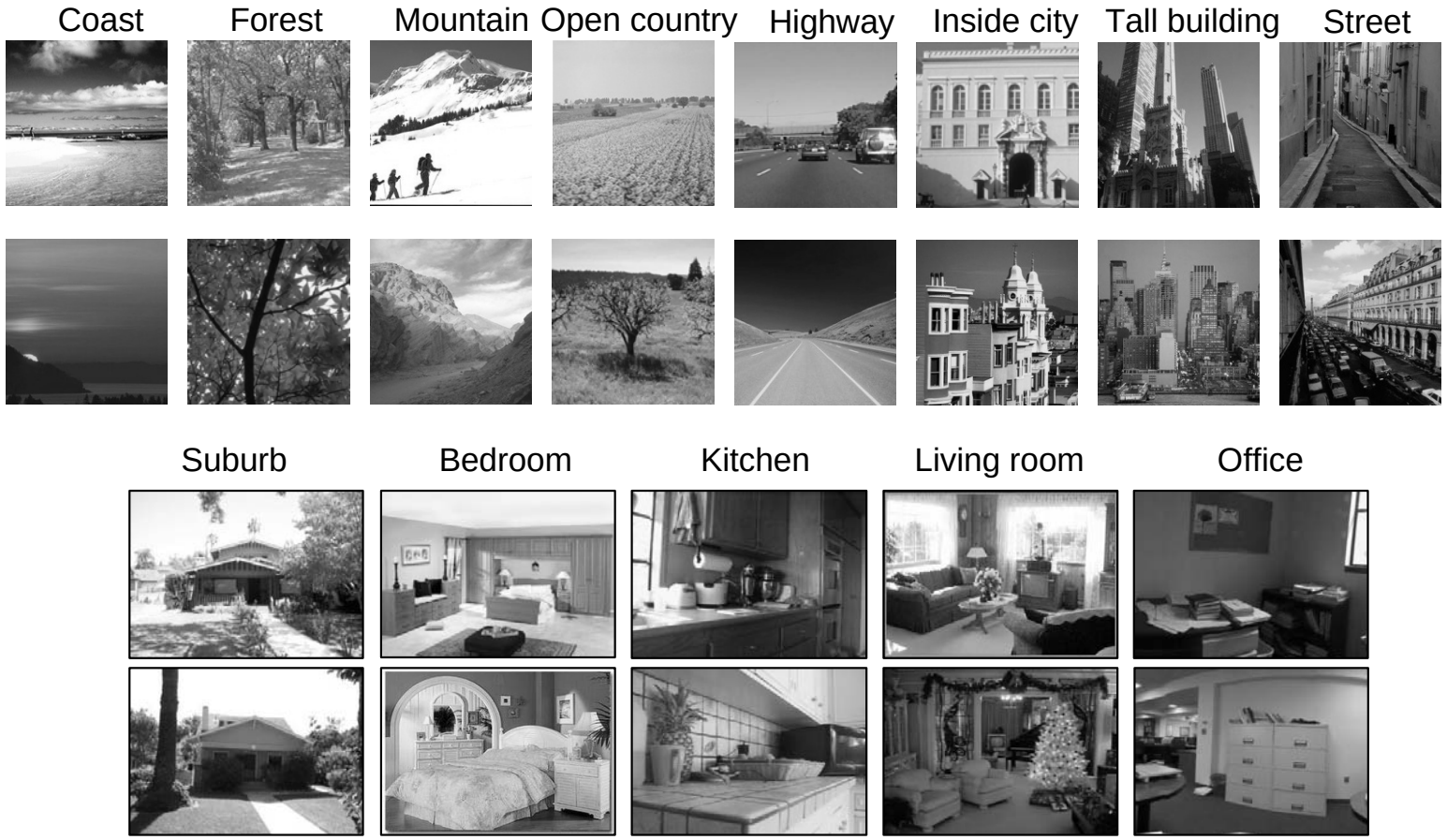
Sky/clouds

Oliva & Torralba DATASET

2688 images
8 categories

OT dataset

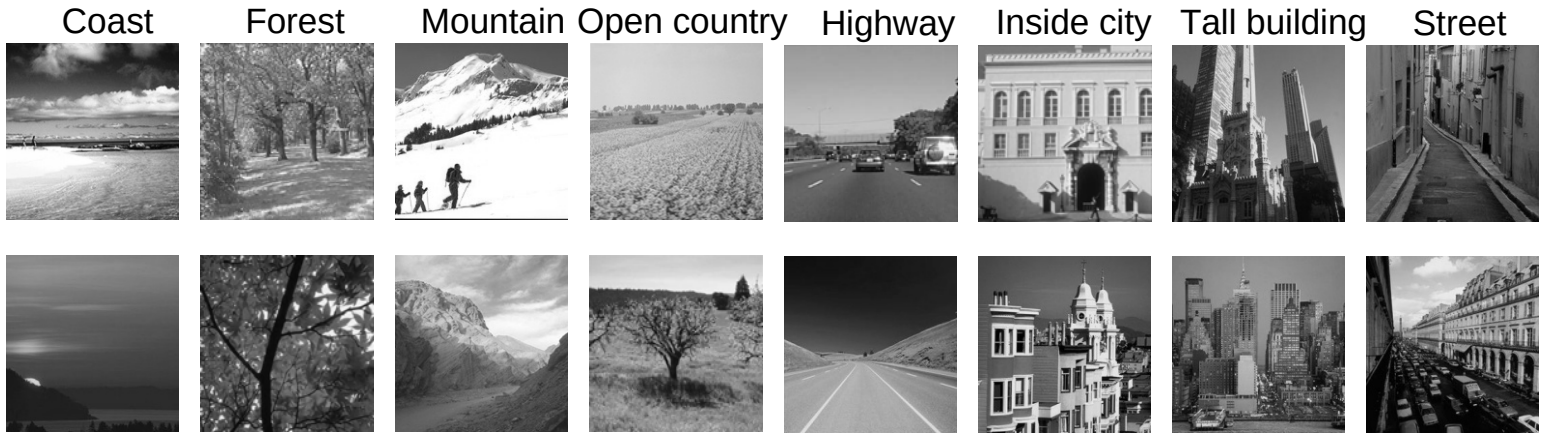
Fei-Fei & Perona DATASET

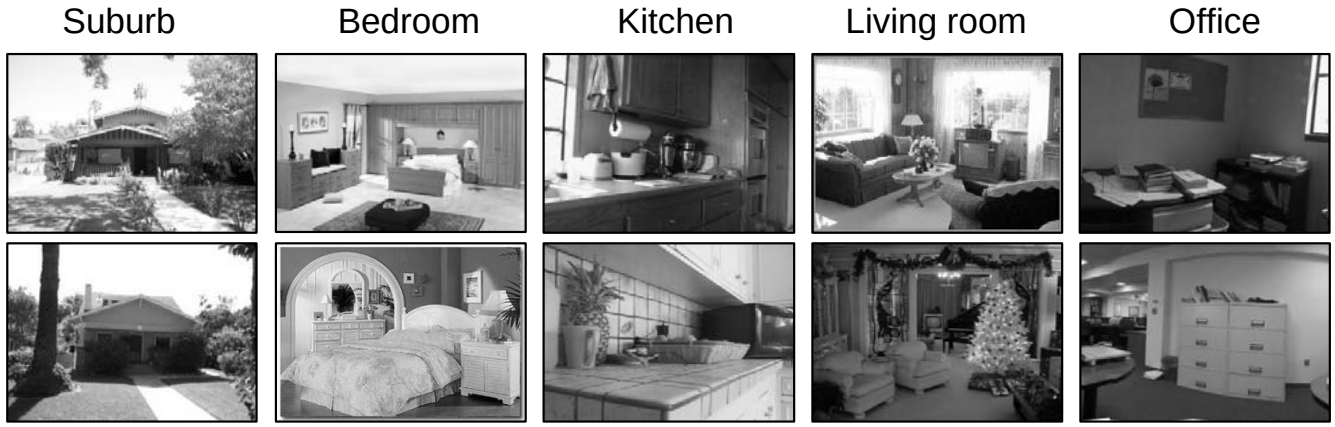


3759 images
13 categories

FP dataset

Lazebnik, Schmid & Ponce DATASET





4385 images
15 categories

LSP dataset

Mulit-way Classification

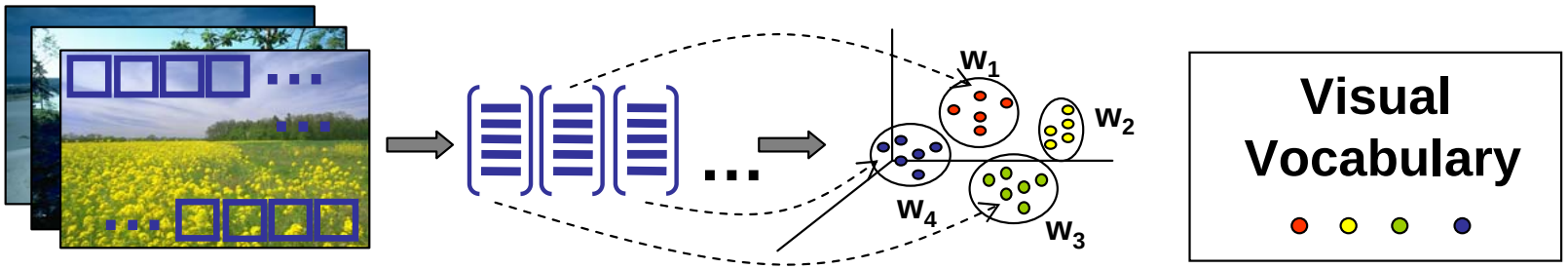
- For each class '*c*' learn a 1-vs-rest SVM classifier
- Classification of test image I according to:

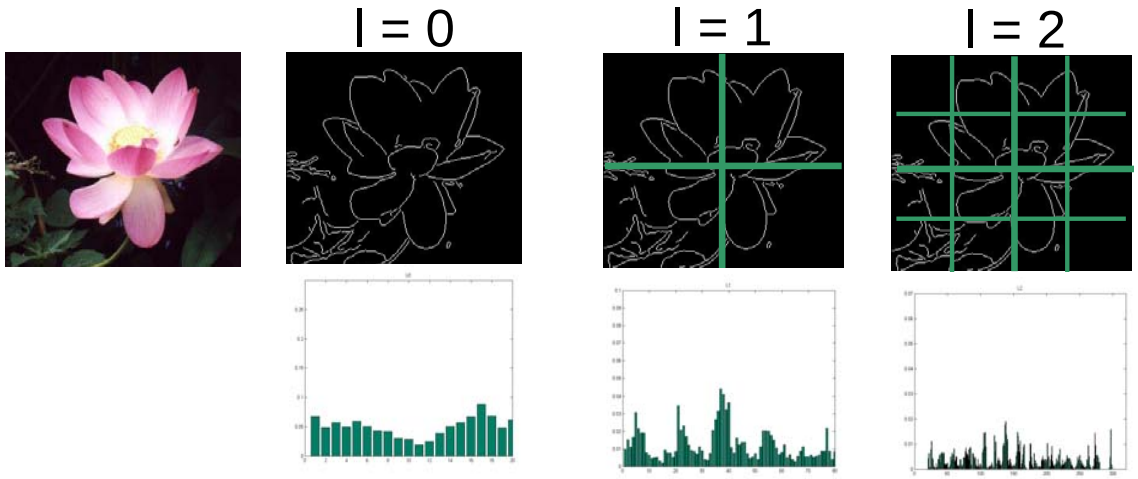
$$c^* = \arg\max_c D_c(I)$$

- where $D_c(I)$ is the distance for the SVM for class *c*

Features

- bag of visual words
- HOG
- spatial pyramid of visual words
- spatial pyramid HOG



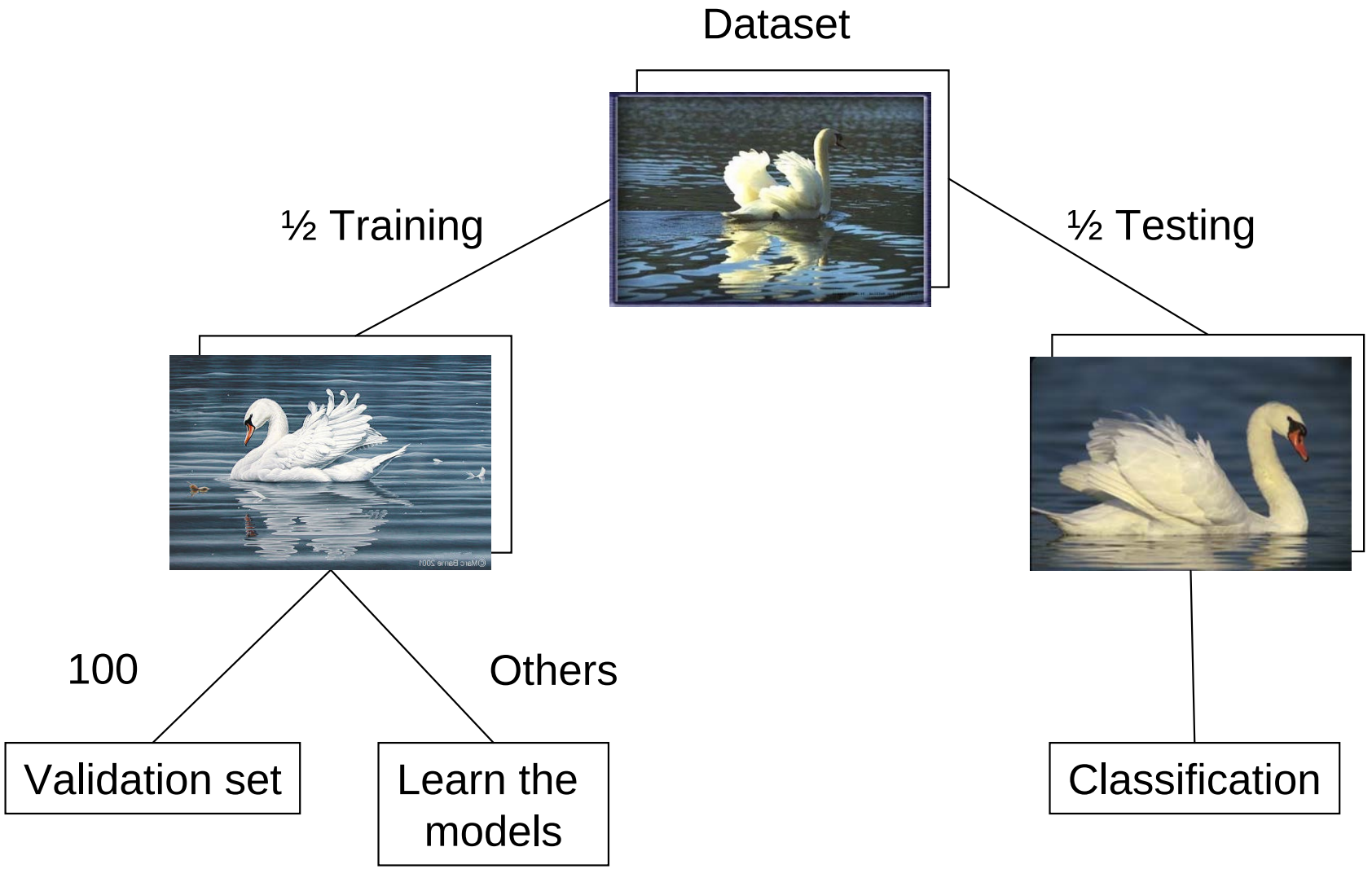


Parameters

- vocabulary size V
- level weightings
- feature combination weights

Methodology for learning parameter values

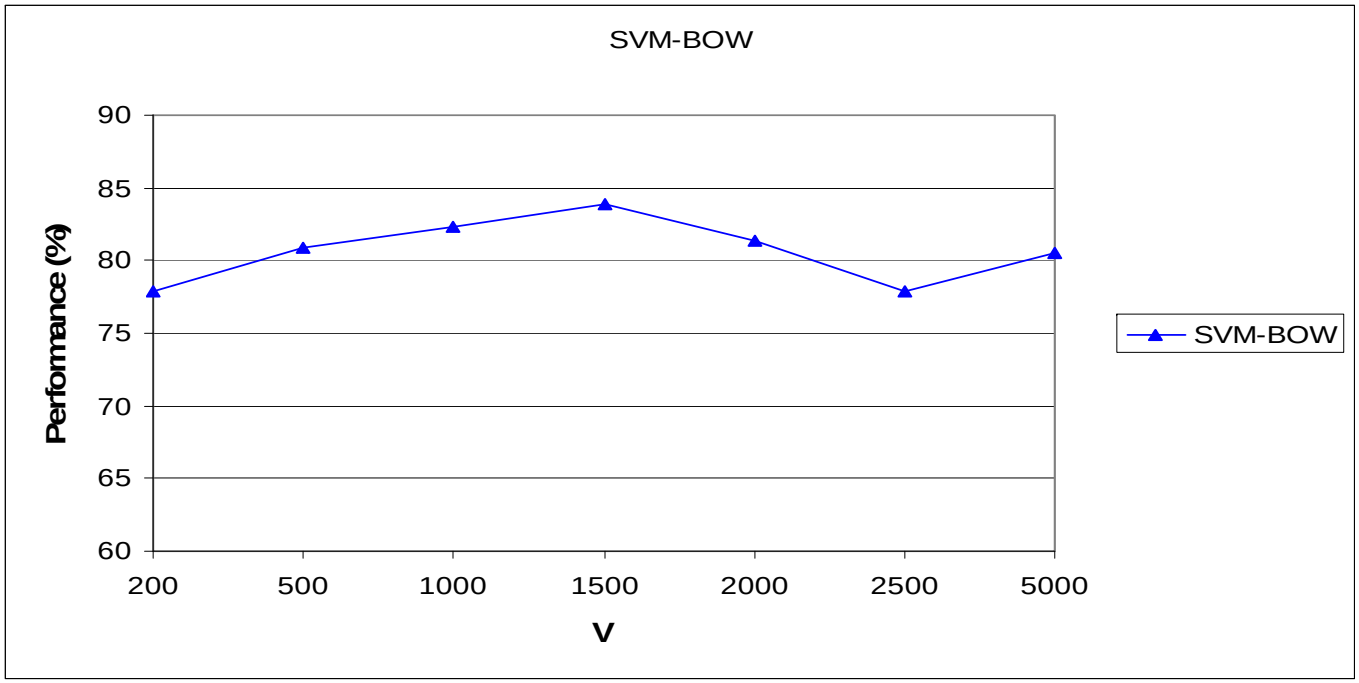
- optimize classification performance on a validation set
- 1 vs rest SVM classifier



Optimize vocabulary size V on validation set

2688 images
8 categories

OT dataset



Spatial pyramid with learnt weights

- Learn level weights – linear combination of kernels

$$K_f(i,j) = \sum_{l \in L} \beta_f^l e^{-\mu \chi^2(h_f^l(i), h_f^l(j))}$$

dense visual words

level weight

base kernel

- if weights common to all classes:

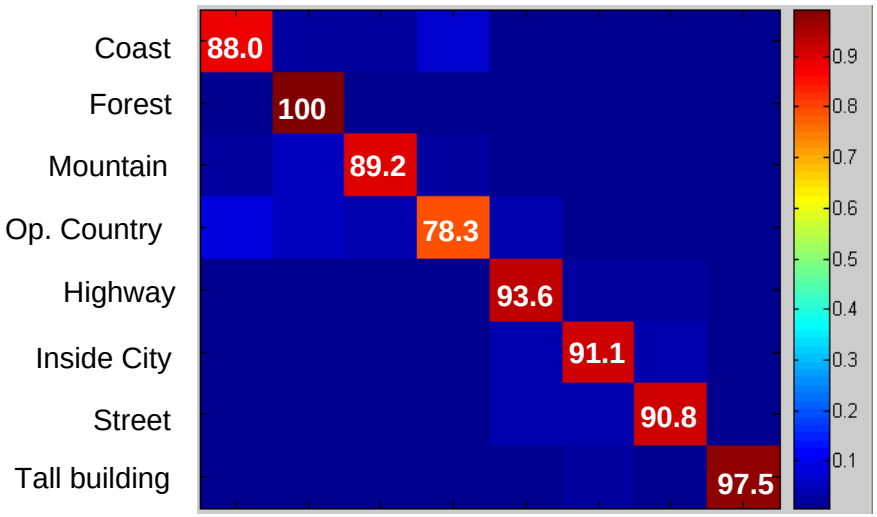
without optimization: $\beta_0 = 0.25, \ \beta_1 = 0.25, \ \beta_2 = 0.5$

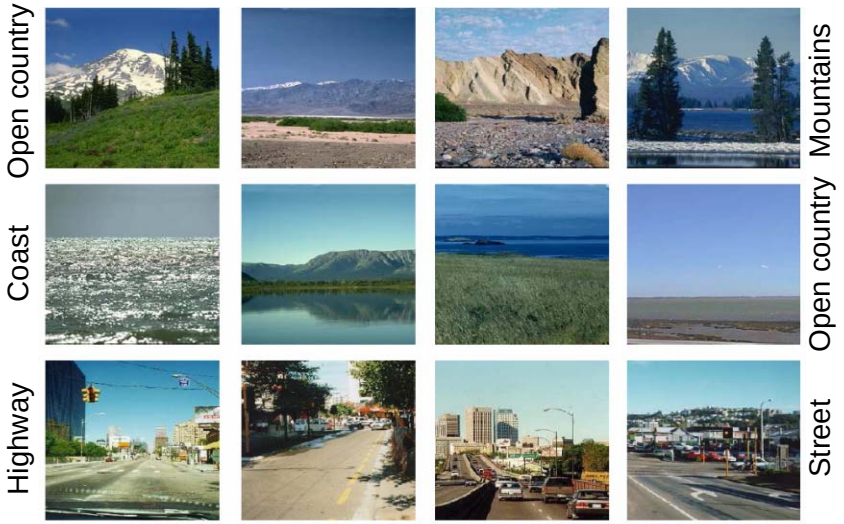
with optimization: $\beta_0 : \beta_2 = 1.0, \ \beta_1 : \beta_2 = 0.8$

Optimized values for each dataset
Same number of Images as authors

SVM one against all – $\mathbf{X}^2$ kernel for visual words
Up to L = 2 for spatial pyramid

Dataset	# of categ.	# train	# test	SP	Authors
OT	8	800	1888	87.1	83.7 Oliva & Torralba
OT	4 Natural	1000	472	93.3	89.0 Oliva & Torralba
OT	4 Man-Made	1000	216	94.2	89.0 Oliva & Torralba
VS	6	600	100	**88.6**	74.1 Vogel & Schiele
FP	13	1300	2459	85.5	65.2 Fei-Fei & Perona
LSP	15	1500	2986	83.5	81.4 Lazebnik et al.





Spatial pyramid with feature combination

- feature weights – linear combination of kernels

$$K_{\text{opt}}(i,j) = \sum_{f \in F} d_f K_f(i,j)$$

feature weight

feature kernel

- dense visual words
- HOG

Lazebnik et al. dense visual words	dense visual words optimized	dense visual words & HOG optimized
81.1	83.5	90.2

Take home messages

- Lite use of spatial information
  - tiling, spatial pyramid
- Combination of features
  - visual words, sparse, dense, HOG
- Learn parameters on validation set

More classifiers …

- SVM Classifier
  - good performance
  - convex optimization
- Logistic regression
- Adaboost
  - e.g. used by Viola & Jones face detector
  - slow to learn, fast to test
- Random forests
  - fast to learn, fast to test
  - Jamie Shotton tutorial