max planck institut informatik

UNIVERSITÄT DES SAARLANDES

High Level Computer Vision

Bag of Words Model and Part-Based Models for Object Class Recognition

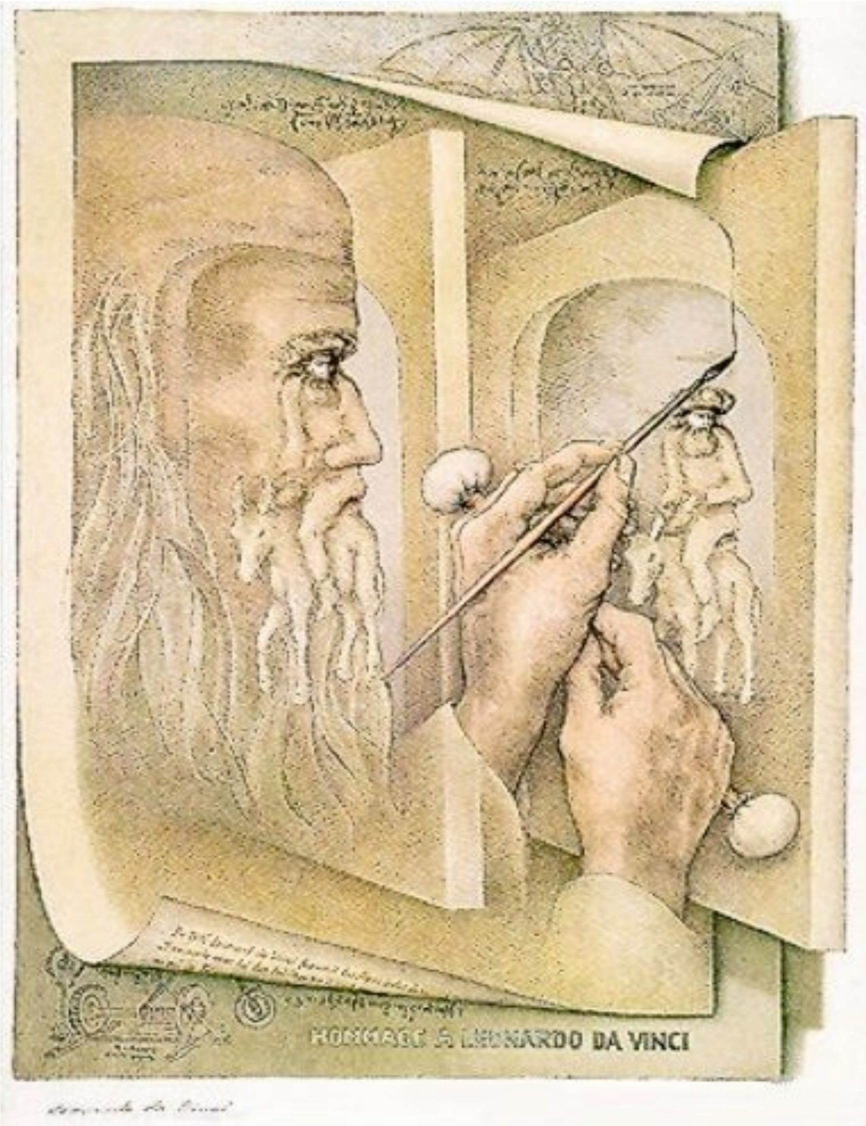
Bernt Schiele - schiele@mpi-inf.mpg.de
Mario Fritz - mfritz@mpi-inf.mpg.de

https://www.mpi-inf.mpg.de/hlcv

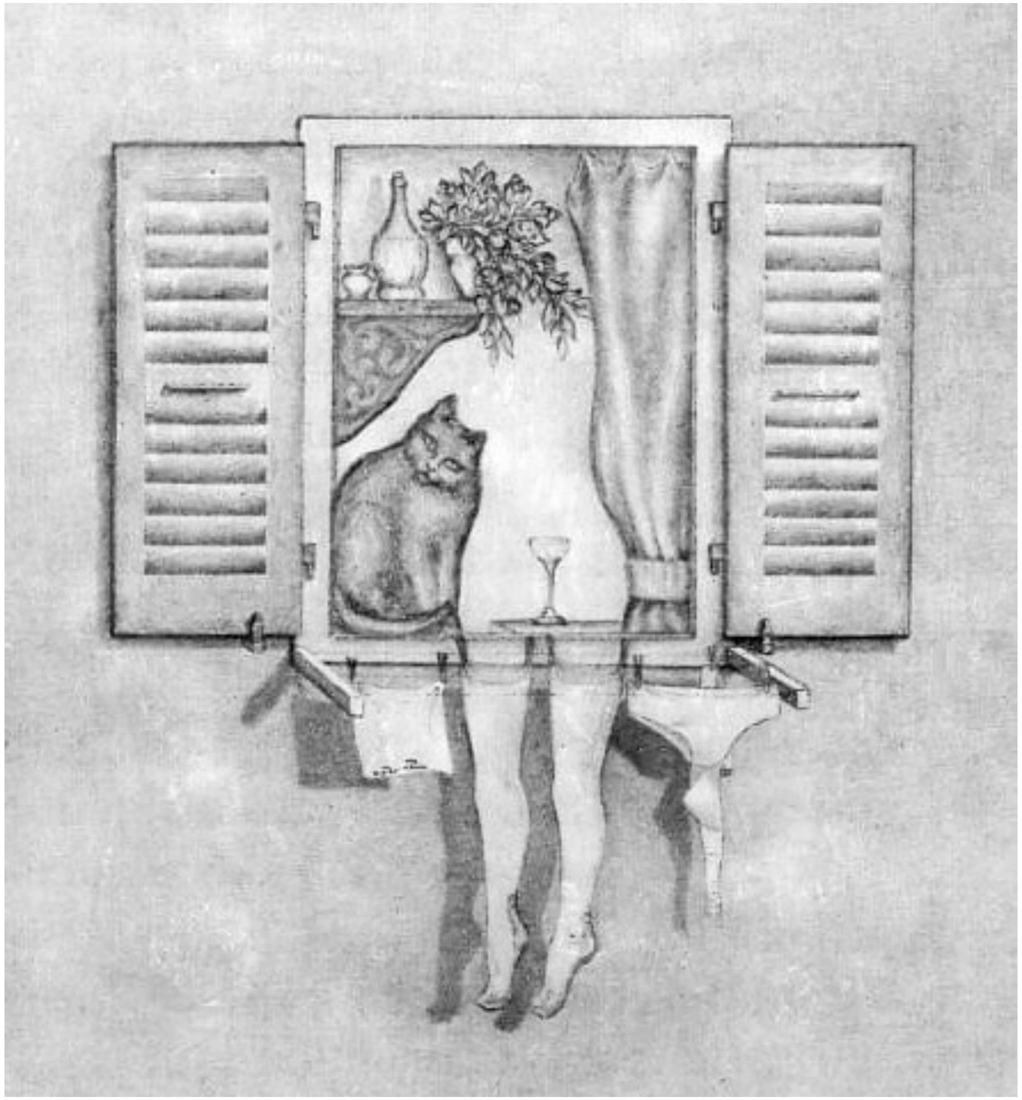
Object Recognition (reminder)

- Different Types of Recognition Problems:
  - Object **Identification**
    - recognize your apple, your cup, your dog
    - sometimes called: "instance recognition"
  - Object **Classification**
    - recognize any apple, any cup, any dog
    - also called: generic object recognition, object categorization, …
    - typical definition: 'basic level category'

Complexity of Recognition

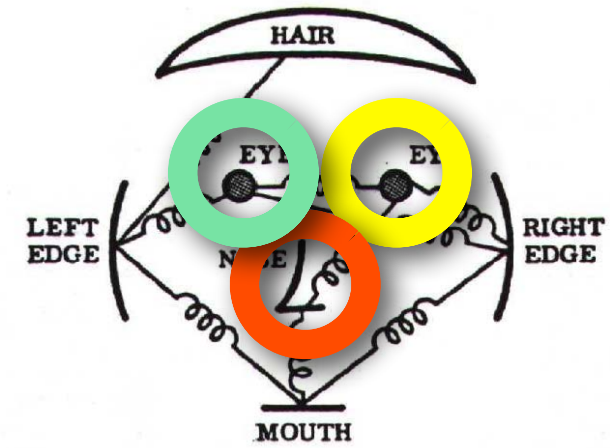


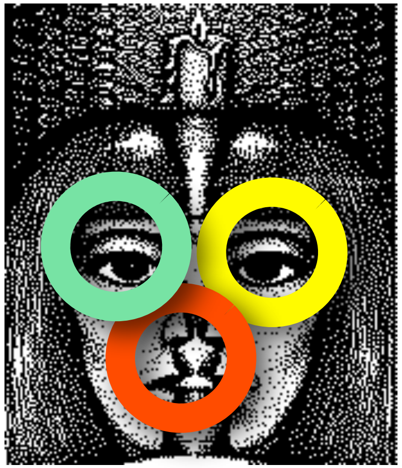
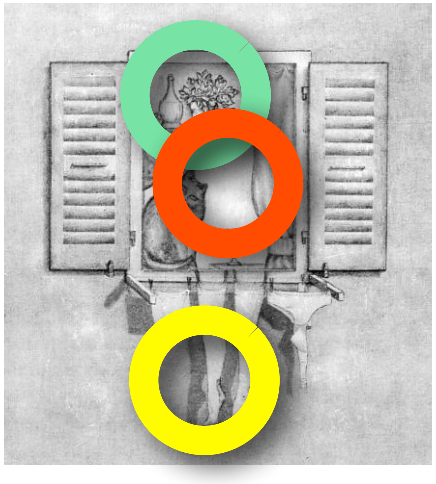
Complexity of Recognition



Class of Object Models: Part-Based Models / Pictorial Structures

- Pictorial Structures [Fischler & Elschlager 1973]
  - Model has two components
    - **parts** (2D image fragments)
    - **structure** (configuration of parts)

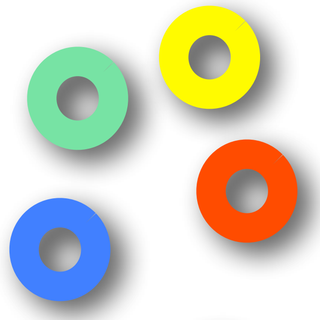



“State-of-the-Art” in Object Class Representations

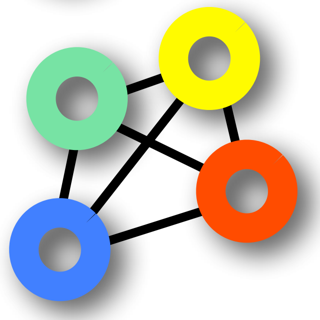
- Bag of Words Models (BoW)
  - object model = histogram of local features
  - e.g. local feature around interest points

BoW: no spatial relationships



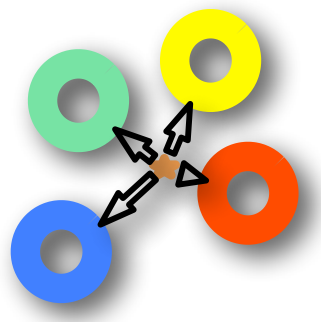
- Global Object Models
  - object model = global feature object feature
  - e.g. HOG (Histogram of Oriented Gradients)

e.g. HOG: fixed spatial relationships



- Part-Based Object Models
  - object model = models of parts & spatial topology model
  - e.g. constellation model or ISM (Implicit Shape Model)

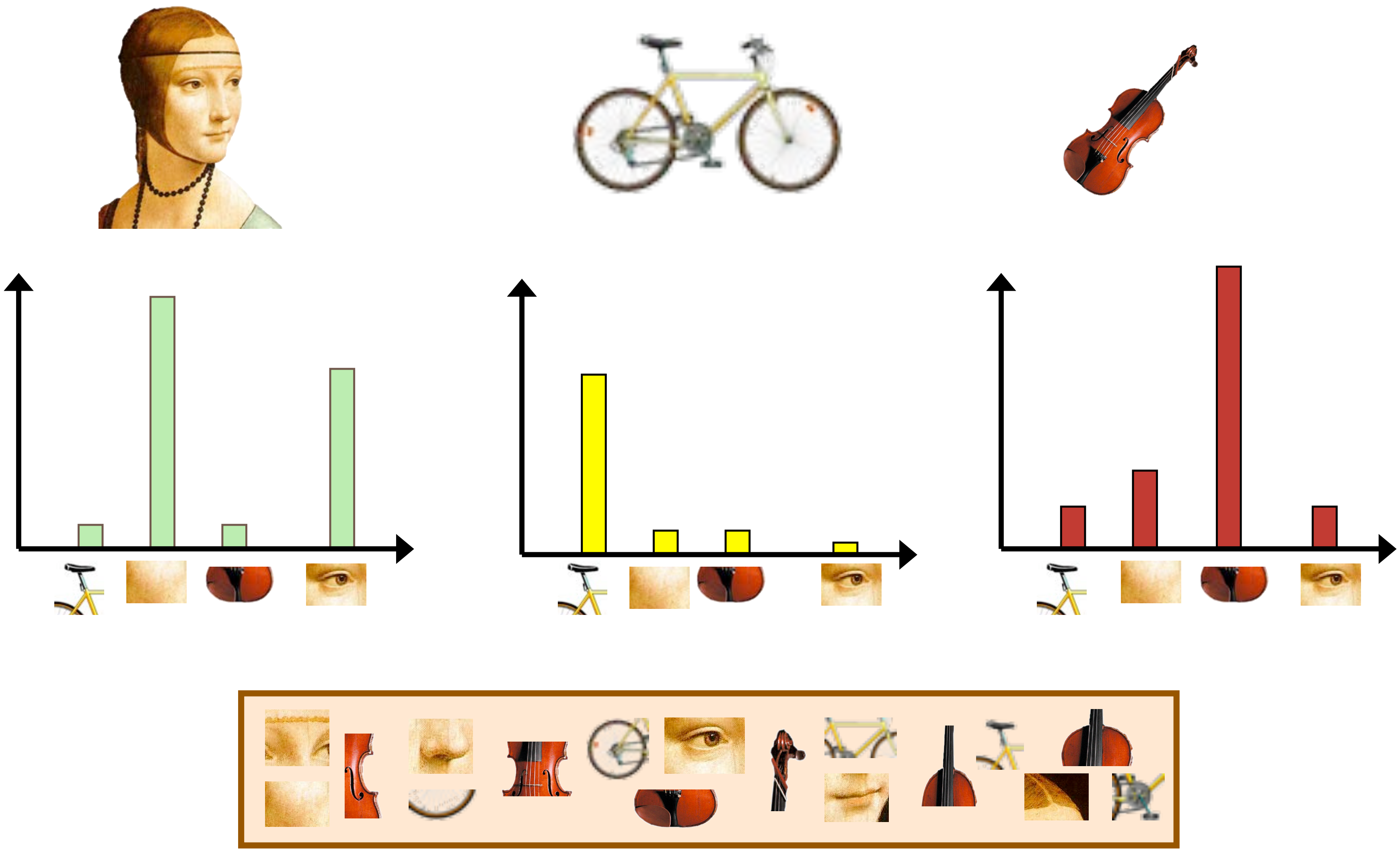
e.g. ISM: flexible spatial relationships



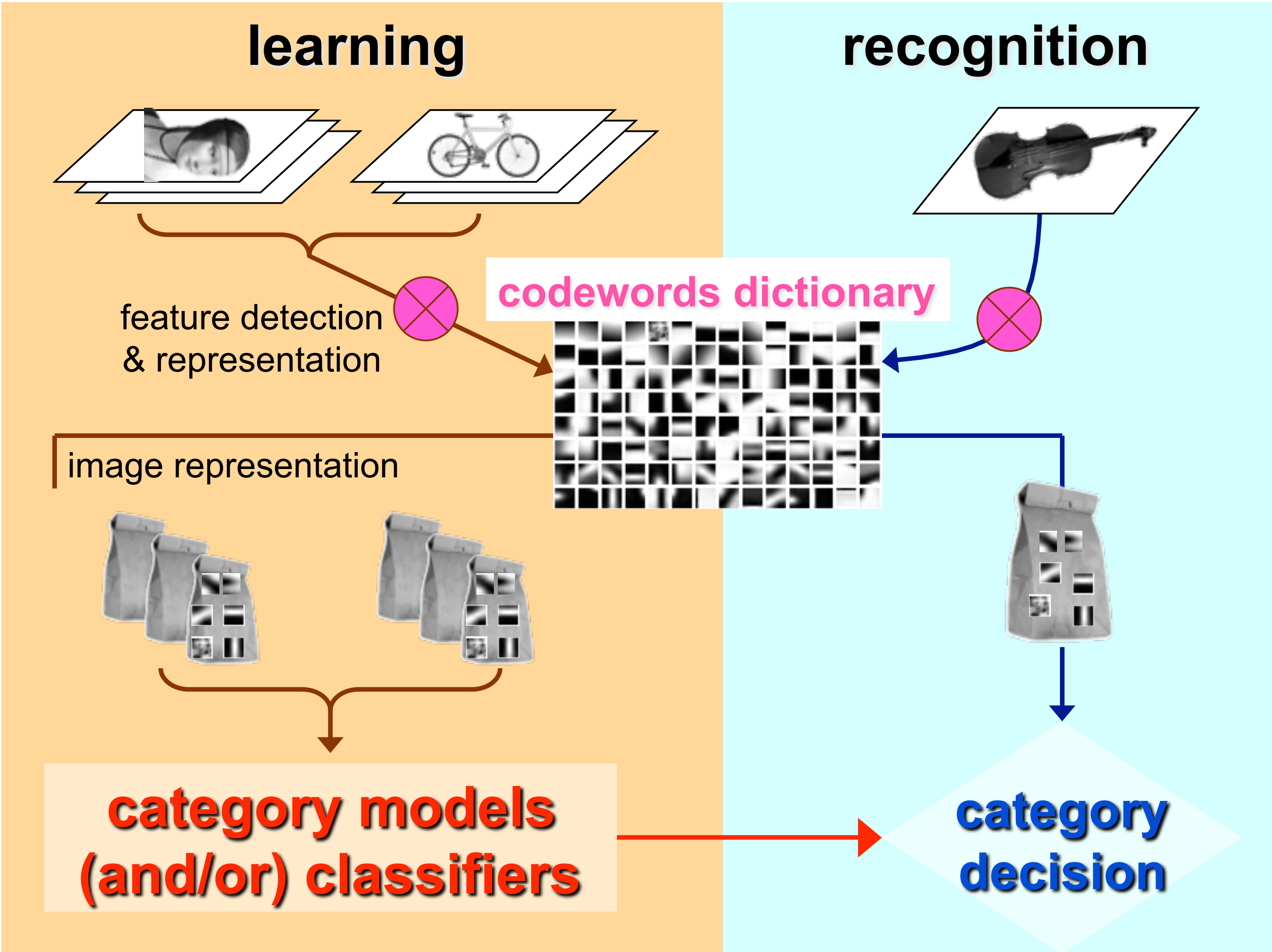
Bag-of-Words Model (BoW) for Object Categorization

Object → Bag of 'words'

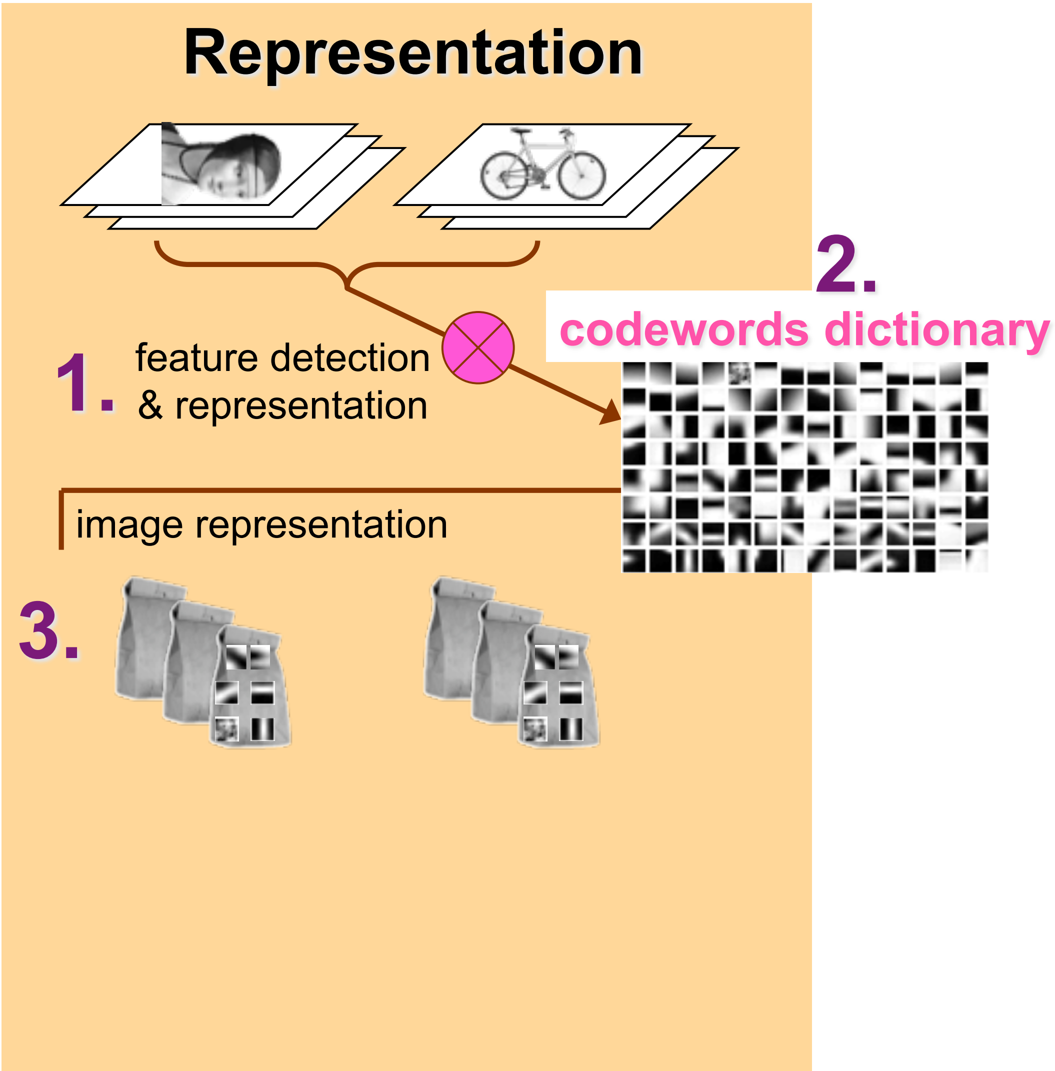
Visual words distributions



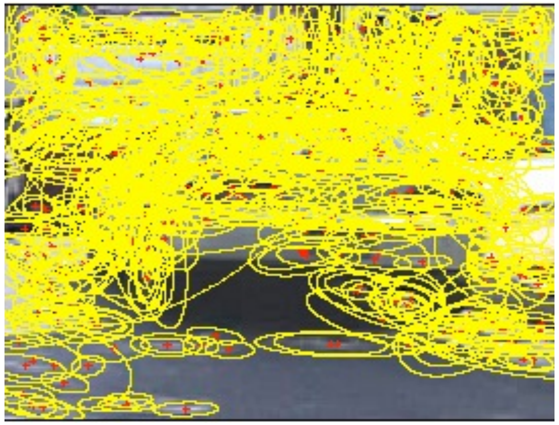
Bag-of-Words Model: Overview



Bag-of-Words Model: Object Representation & Learning

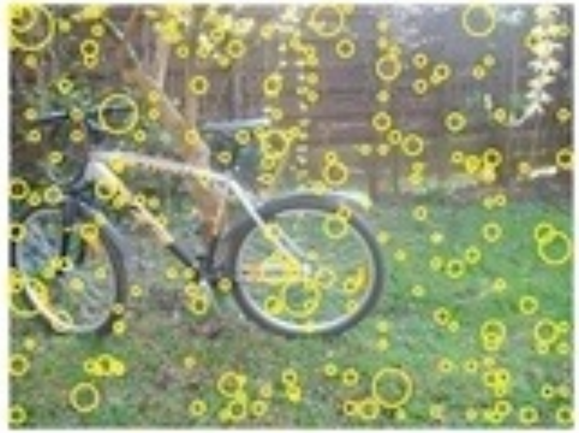


Sampling Strategies

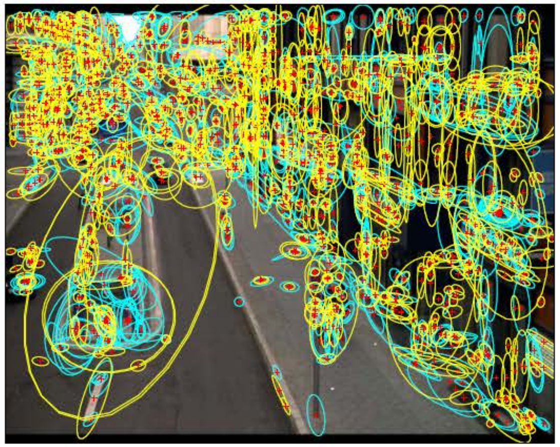


Sparse, at interest points

Dense, uniformly



Randomly

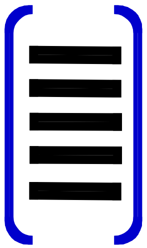


Multiple interest operators

- To find specific, textured objects, sparse sampling from interest points often more reliable.
- Multiple complementary interest operators offer more image coverage.
- For object categorization, dense sampling offers better coverage.

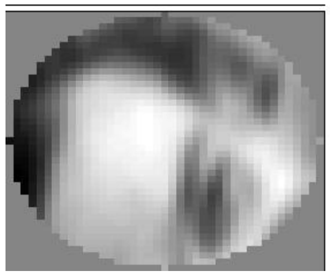
[See Nowak, Jurie & Triggs, ECCV 2006]

BoW-1. Feature detection and representation

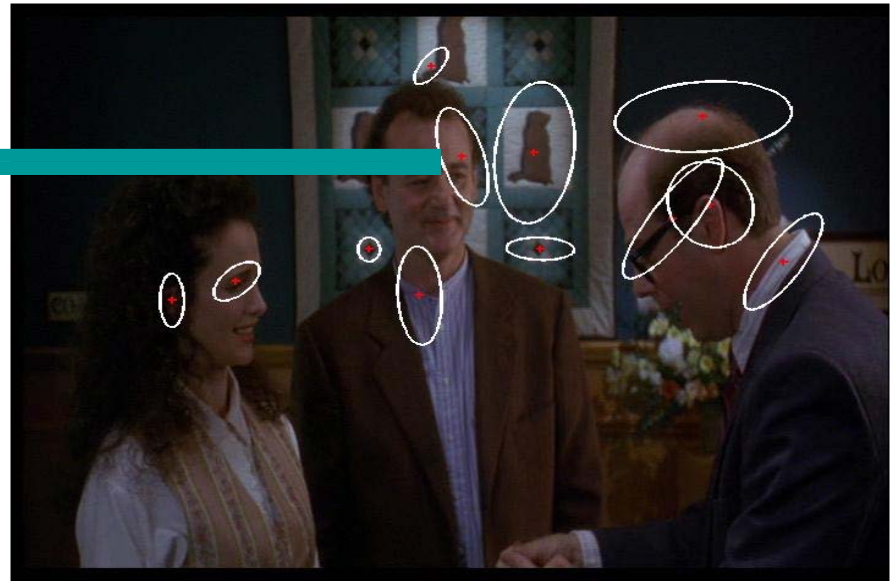


Compute SIFT descriptor

[Lowe'99]



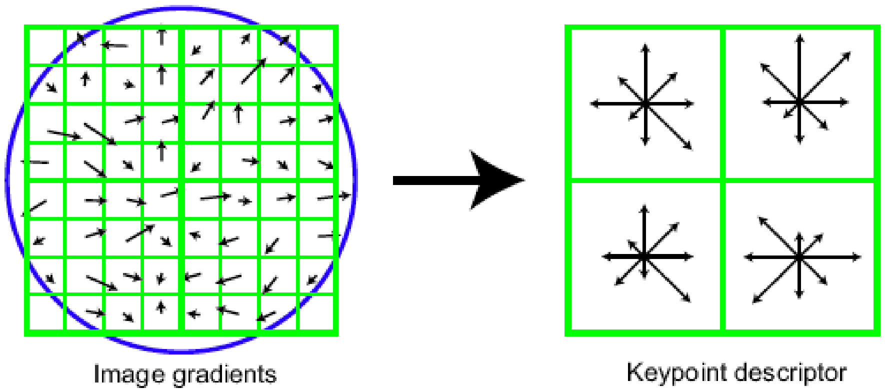
Normalize patch



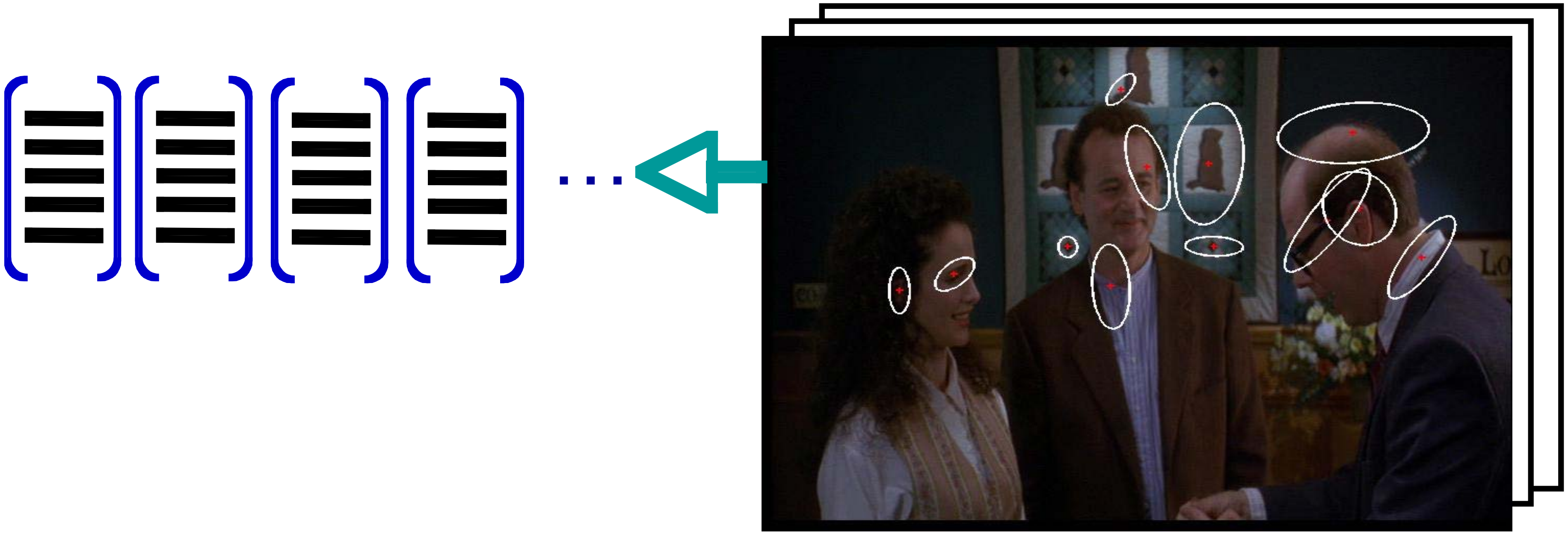
SIFT - Scale Invariant Feature Transform [Lowe]

- Interest Points:
  - Difference of Gaussians

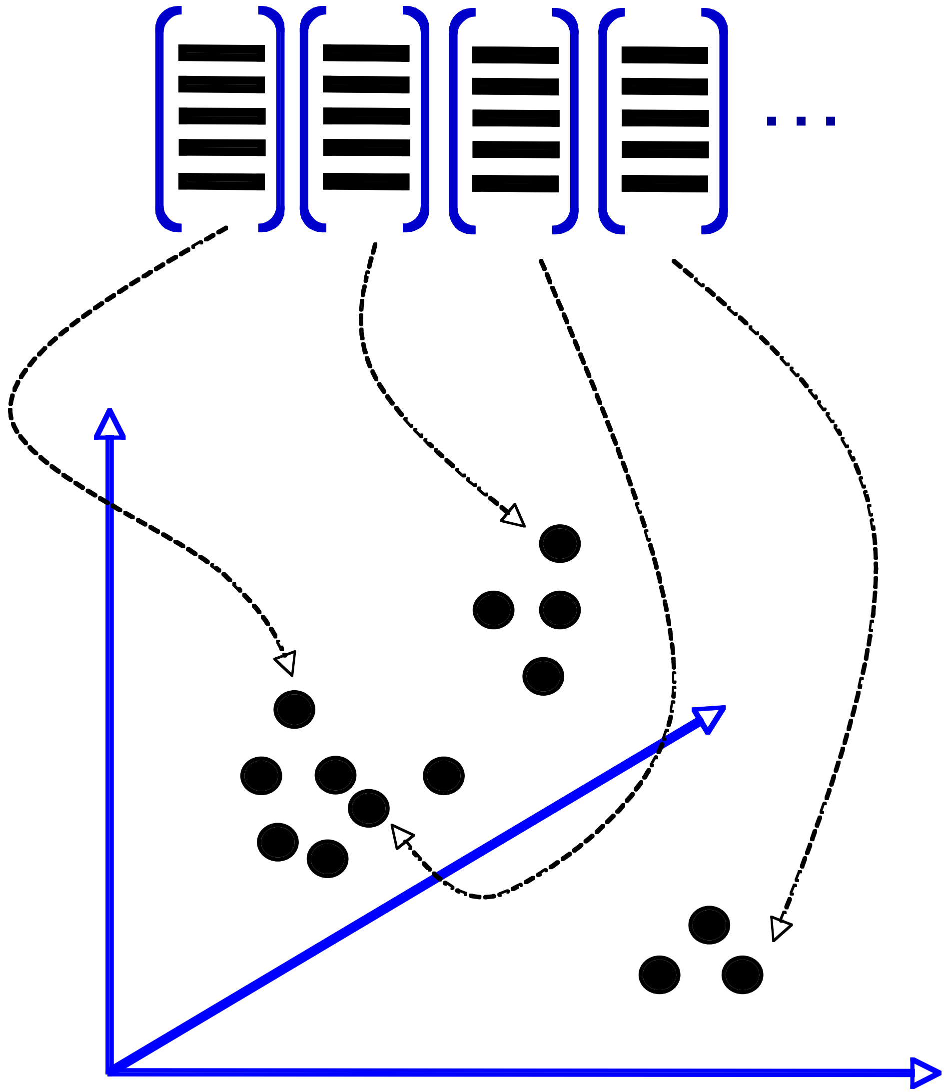
- Feature Descriptor:
  - local histogram of 4x4 local orientation histograms (each over 16x16 pixels),
    - 8 orientations x 4 x 4 = 128 dimensions
  - example: 2x2 local orientation histogram (each of 4x4 pixels):



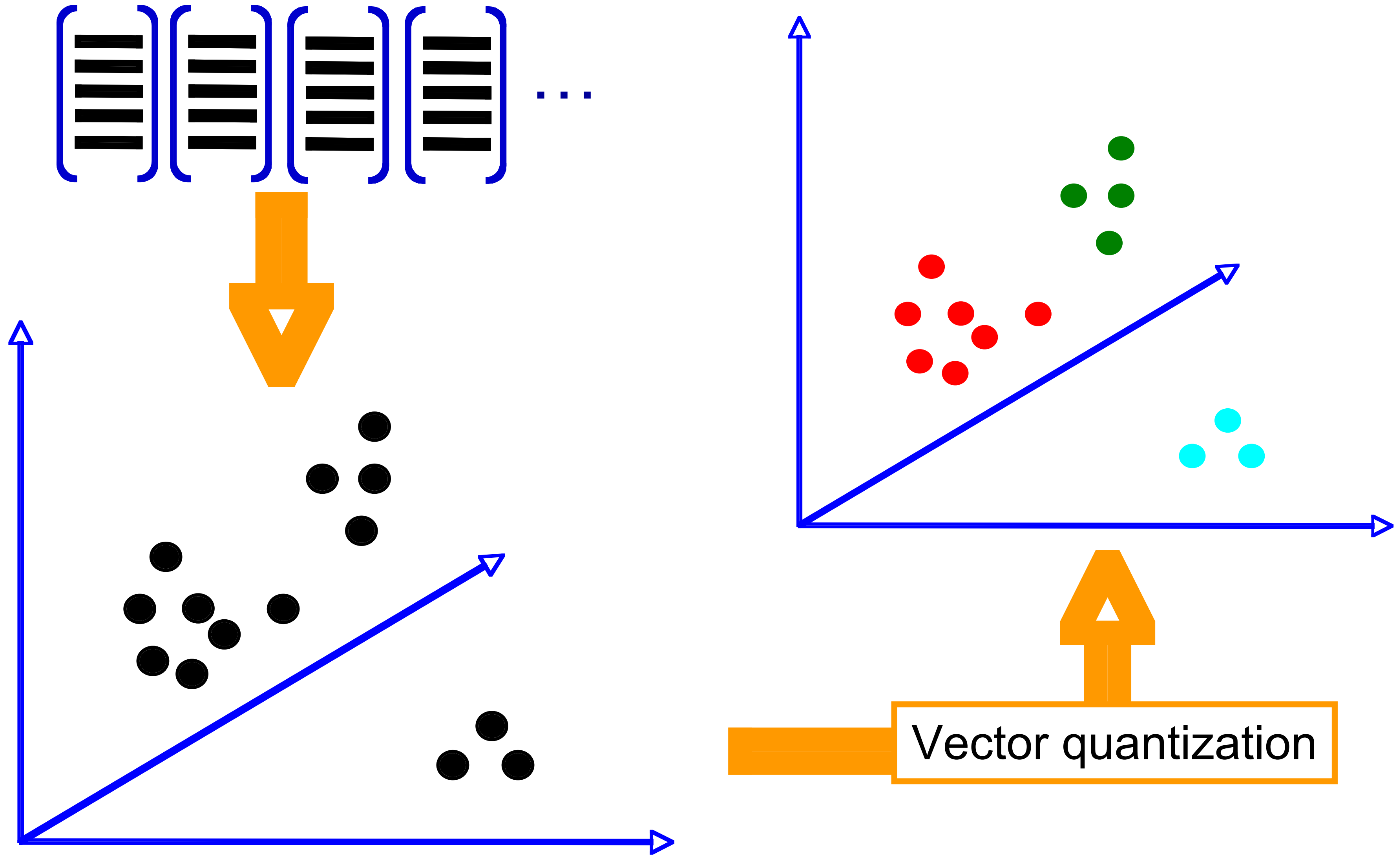
BoW-1. Feature detection and representation



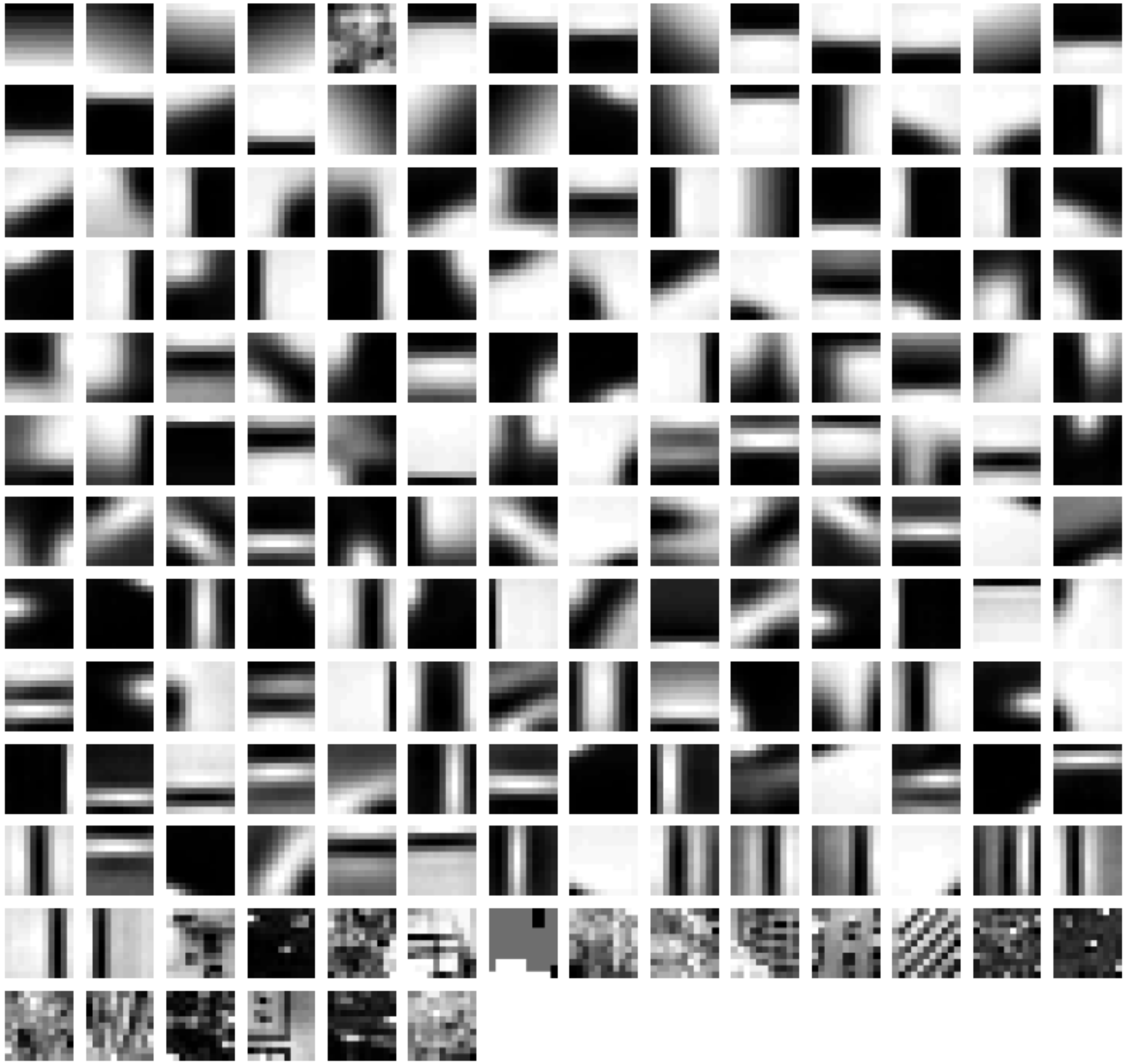
BoW-2. Codewords (= “visual words”) dictionary formation



BoW-2. Codewords dictionary formation

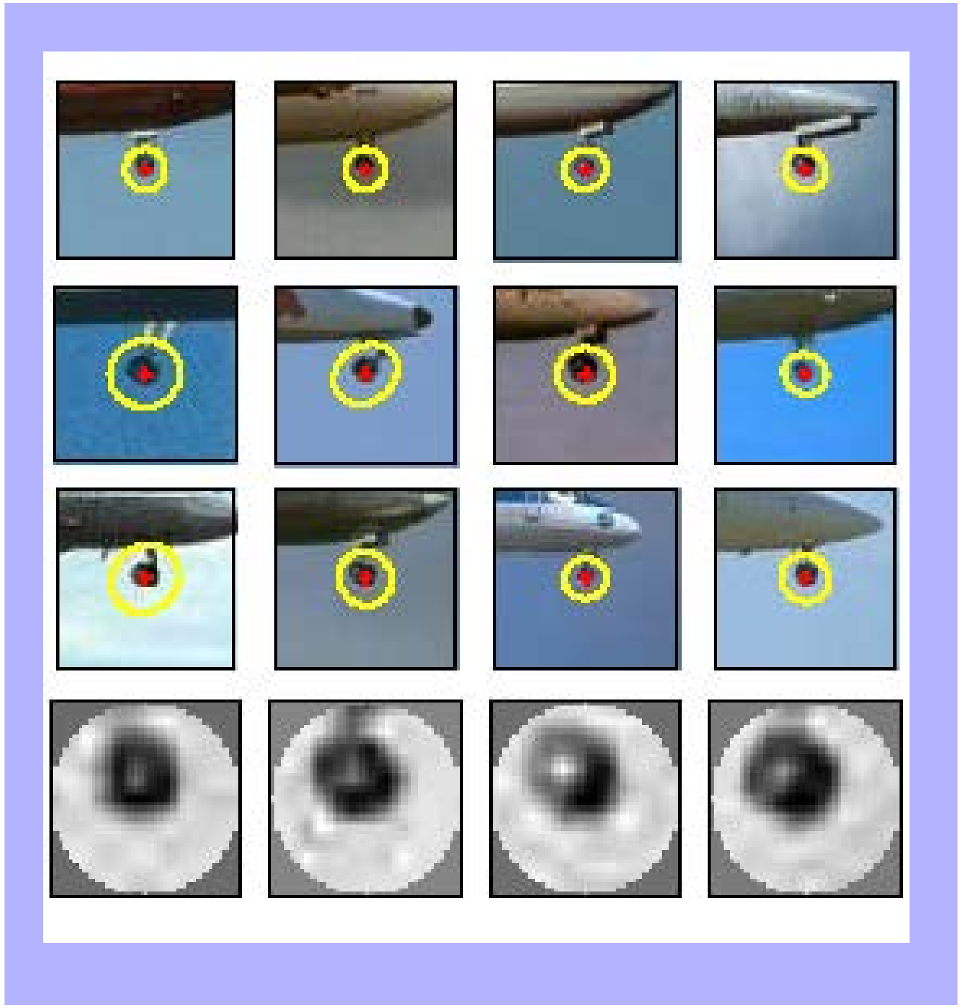


BoW-2. Codewords dictionary formation



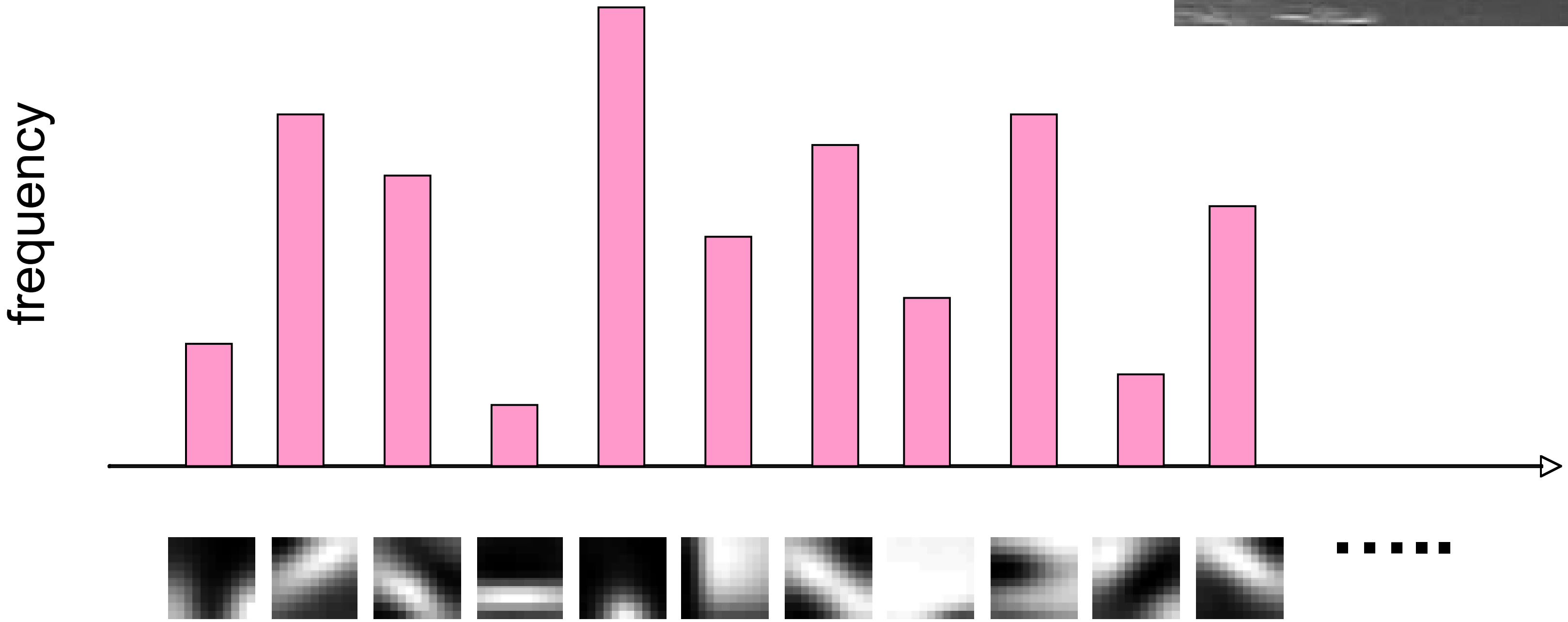
Fei-Fei et al. 2005

Image patch examples of codewords / "visual words"

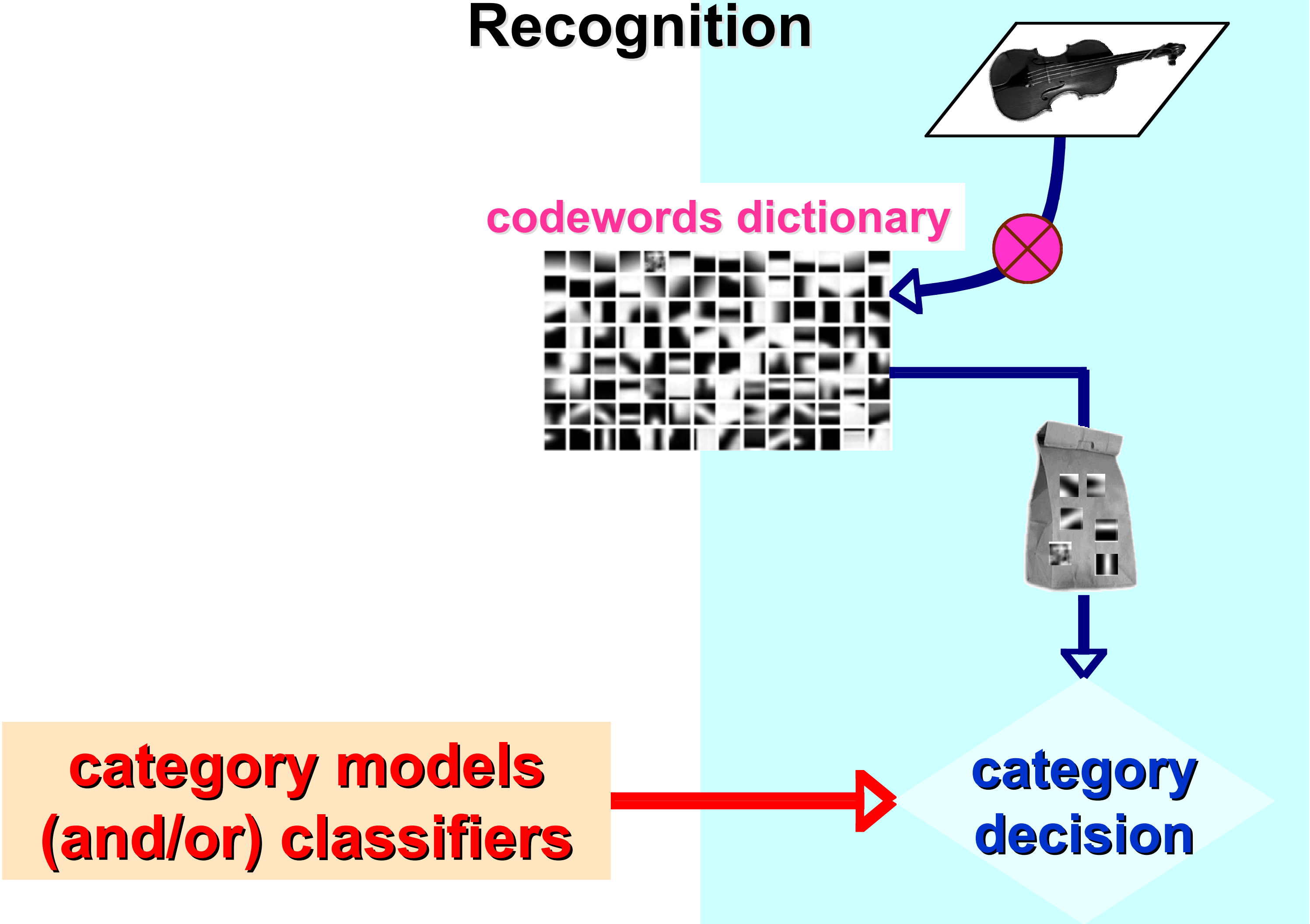


Sivic et al. 2005

BoW-3. Object / Image representation: Histogram over Codewords / Visual Words

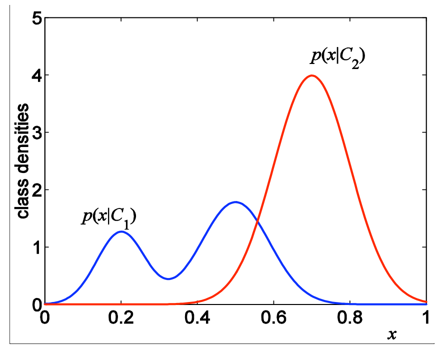


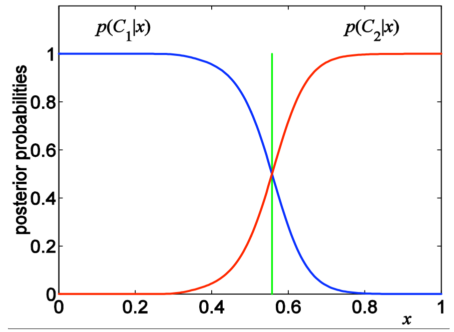
Learning and Recognition



Learning and Recognition

- Generative method:
  - graphical models
- Discriminative method:
  - Support Vector Machine (SVM)





category models (and/or) classifiers

Generative Models explored

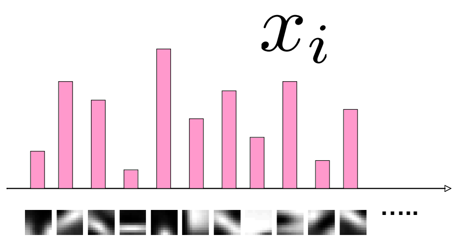
- Naïve Bayes classifier
  - Csurka Bray, Dance & Fan, 2004

- Hierarchical Bayesian text models (pLSA and LDA)
  - Background: Hoffman 2001, Blei, Ng & Jordan, 2004
  - Object categorization: Sivic et al. 2005, Sudderth et al. 2005
  - Natural scene categorization: Fei-Fei et al. 2005

Naïve Bayes Classifier

- Classify image using histograms of occurrences on visual words:

$$\mathbf{x} =$$



  if only present/absence of a word is taken into account:

$$x_i \in \{0, 1\}$$

- Naïve Bayes classifier assumes that visual words are conditionally independent given object class

$$P(\mathbf{x}|c) = \prod_{i=1}^{m} P(x_i|c)$$

Based on lecture by Prof.
T. Hofmann

Naive Bayes Classifier

- Multinomial model for each object class:

$$P(\mathbf{x}|c) = \prod_{i=1}^{m} P(x_i|c)$$

- Class priors: $P(c)$, with $\sum_c P(c) = 1$

- Posterior probabilities: $$\boxed{P(c|\mathbf{x}) = \frac{P(c)\prod_{t=1}^{n} P(x_t|c)}{\sum_{c'} P(c')\prod_{t=1}^{n} P(x_t|c')}}$$

Naive Bayes Classifier: Decision

- Bayes optimal decision:

$$c^* = \text{argmax}_c \; P(c|\mathbf{x})$$

$$\boxed{= \text{argmax}_c \left[\log P(c) + \sum_{t=1}^{n} \log P(x_t|c) \right]}$$

Image Classification with Naive Bayes

- Image dataset: 7 object categories, arbitrary views, partial occlusions

Csurka et al. 2004

Example of feature extraction

All features detected in the image

Features corresponding to two different visual words

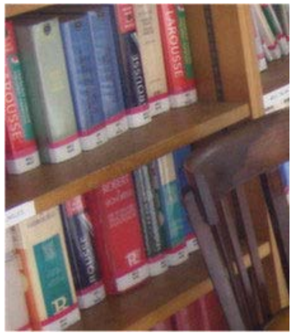
Csurka et al. 2004

Recognition results:

Table 1. Confusion matrix and the mean rank for the best vocabulary (*k=1000*).

True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**76**	4	2	3	4	4	13
buildings	2	**44**	5	0	5	1	3
trees	3	2	**80**	0	0	5	0
cars	4	1	0	**75**	3	1	4
phones	9	15	1	16	**70**	14	11
bikes	2	15	12	0	8	**73**	0
books	4	19	0	6	7	2	**69**
Mean ranks	1.49	1.88	1.33	1.33	1.63	1.57	1.57

Examples of correctly classified images:

Summary & Discussion: BoW for Object Categorization

- Bag-of-words representation:
  - Sparse representation of object category
  - Many machine learning methods are directly applicable.
  - Robust to occlusions
  - Allows sharing of representation between multiple classes
- Problems:
  - Localization of objects in images is problematic
  - Spatial distribution of visual words is not modeled, all these images have equal probability for bag-of-words methods:

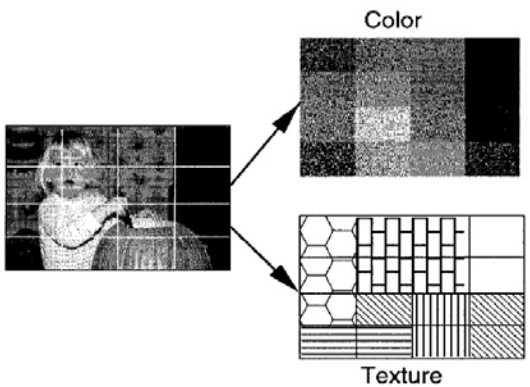
Beyond Bag-of-Words: Spatial Pyramid Matching

- Address the problem of preserving “some” spatial information
- Still applicable to local feature representations
- Idea:
    - compute local bag of words representations
    - concatenate the representations
- following slides form Svetlana Lazebnik

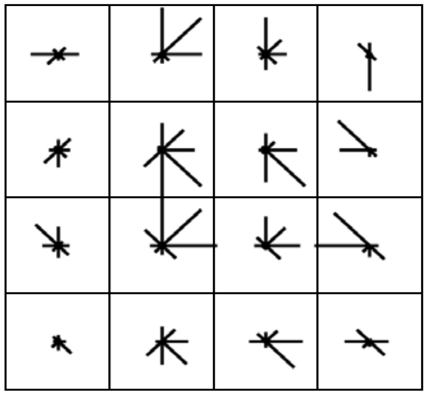
Overview

- A "pre-attentive" approach: recognize the scene as a whole without examining its constituent objects   Biederman (1988), Thorpe et al. (1996), Fei-Fei et al. (2002), Renninger & Malik (2004)
- Inspiration: *locally orderless images*   Koenderink & Van Doorn (1999)


- Previous work: "subdivide-and-disorder" strategy



Szummer & Picard (1997)



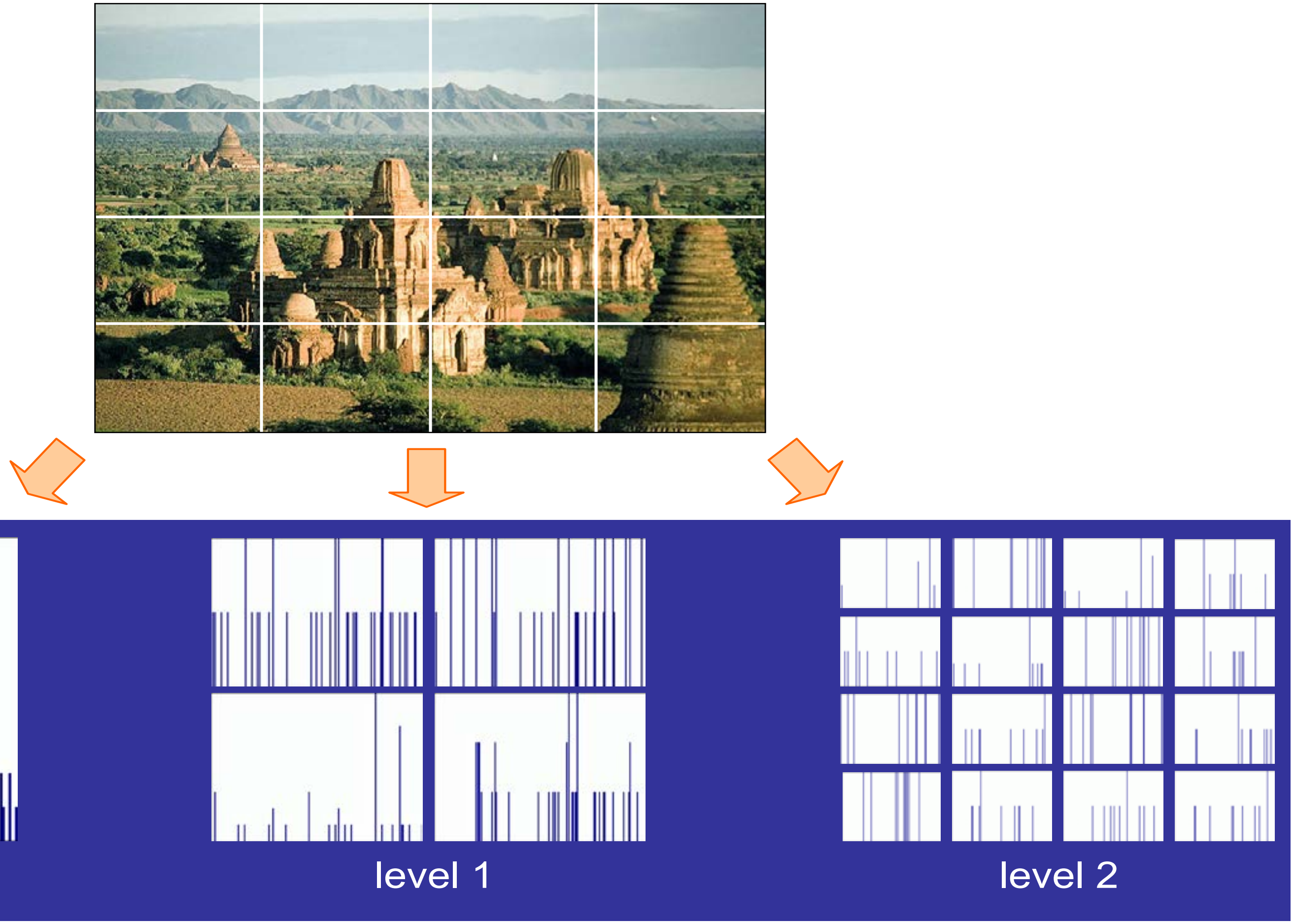
SIFT: Lowe (1999, 2004)

Gist: Torralba et al. (2003)

2

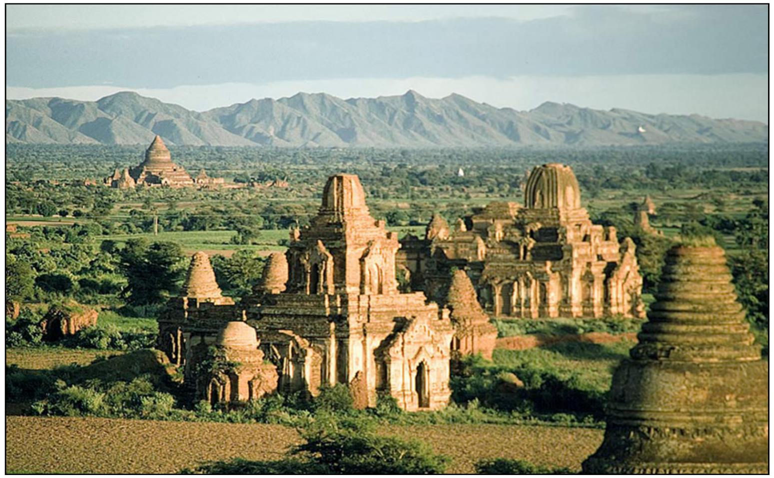
Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution
- Based on *pyramid match kernels* Grauman & Darrell (2005)
  - Grauman & Darrell: build pyramid in feature space, discard spatial information
  - Our approach: build pyramid in image space, quantize feature space

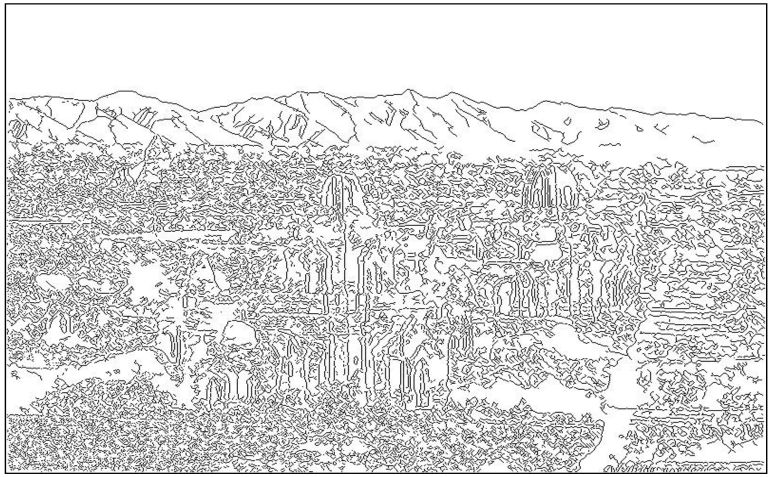


3

Feature extraction

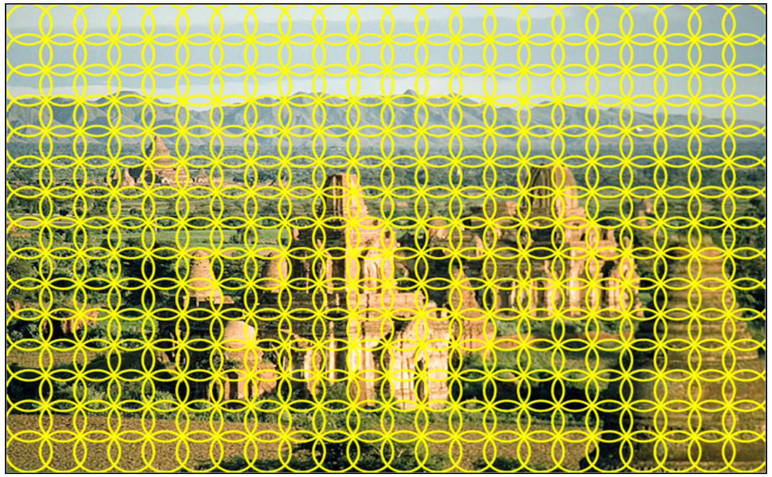


Weak features



Edge points at 2 scales and 8 orientations (vocabulary size 16)

Strong features



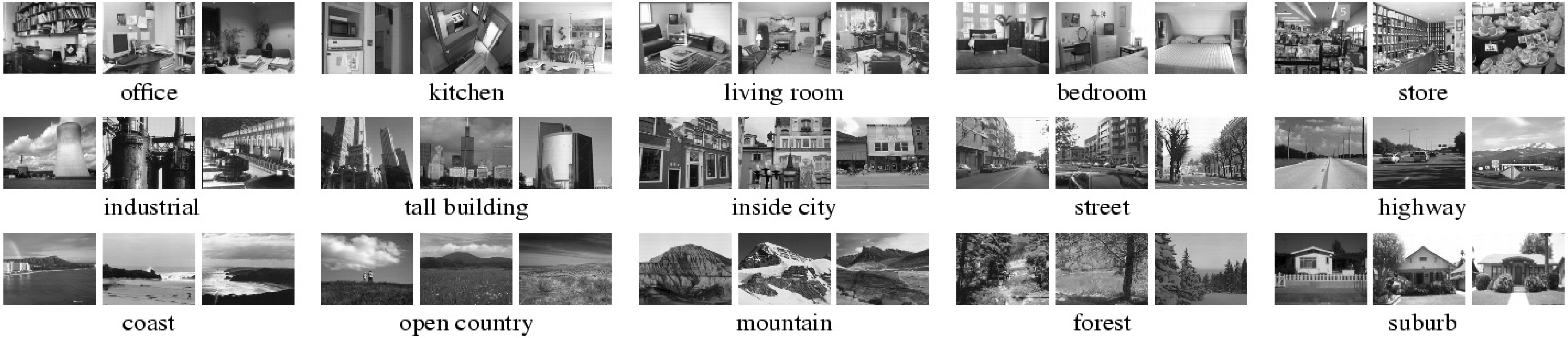
SIFT descriptors of 16x16 patches sampled on a regular grid, quantized to form visual vocabulary (size 200, 400)

5

Scene category dataset

Fei-Fei & Perona (2005), Oliva & Torralba (2001)

`http://www-cvr.ai.uiuc.edu/ponce_grp/data`



Multi-class classification results (100 training images per class)

	Weak features (vocabulary size: 16)		Strong features (vocabulary size: 200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3

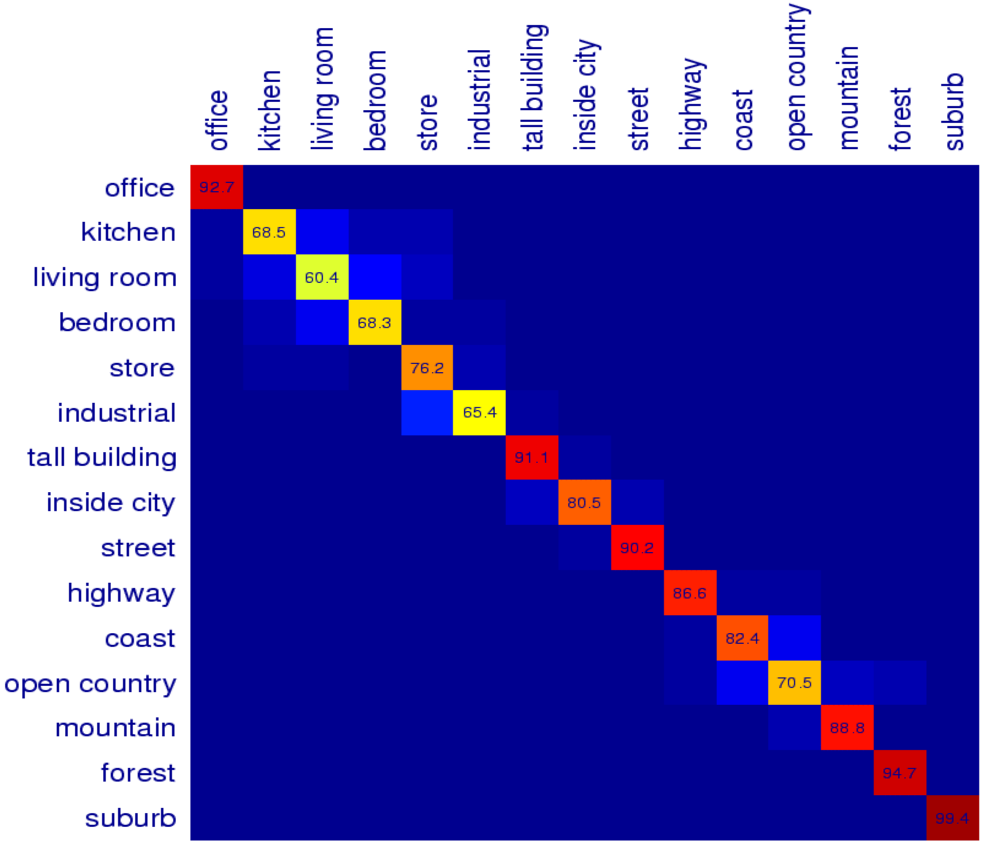
Fei-Fei & Perona: 65.2%

6

Scene category retrieval

7

Scene category confusions



Difficult indoor images


kitchen


living room


bedroom

8

Caltech101 dataset

Fei-Fei et al. (2004)

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

Multi-class classification results (30 training images per class)

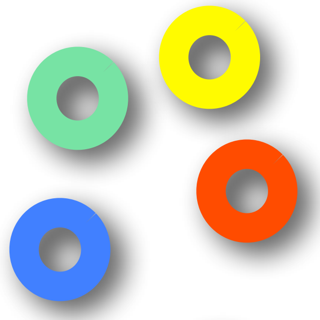
	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

9

“State-of-the-Art” in Object Class Representations

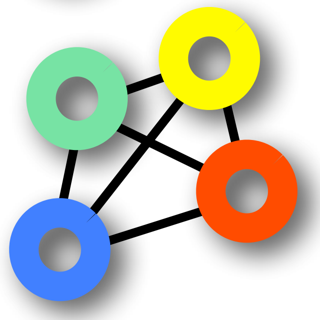
- Bag of Words Models (BoW)
  - object model = histogram of local features
  - e.g. local feature around interest points

BoW: no spatial relationships



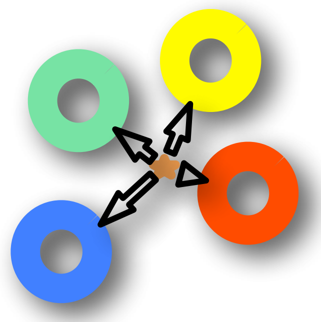
- Global Object Models
  - object model = global feature object feature
  - e.g. HOG (Histogram of Oriented Gradients)

e.g. HOG: fixed spatial relationships



- Part-Based Object Models
  - object model = models of parts & spatial topology model
  - e.g. constellation model or ISM (Implicit Shape Model)

e.g. ISM: flexible spatial relationships



Part-Based Models - Overview Today (more next week)

- Part-Based using Manual Labeling of Parts
  - Detection by Components
  - Multi-Scale Parts

- The Constellation Model
  - automatic discovery of parts and part-structure

- The Implicit Shape Model (ISM)
  - parts obtained by clustering interest-points
  - star-model to model configuration of parts

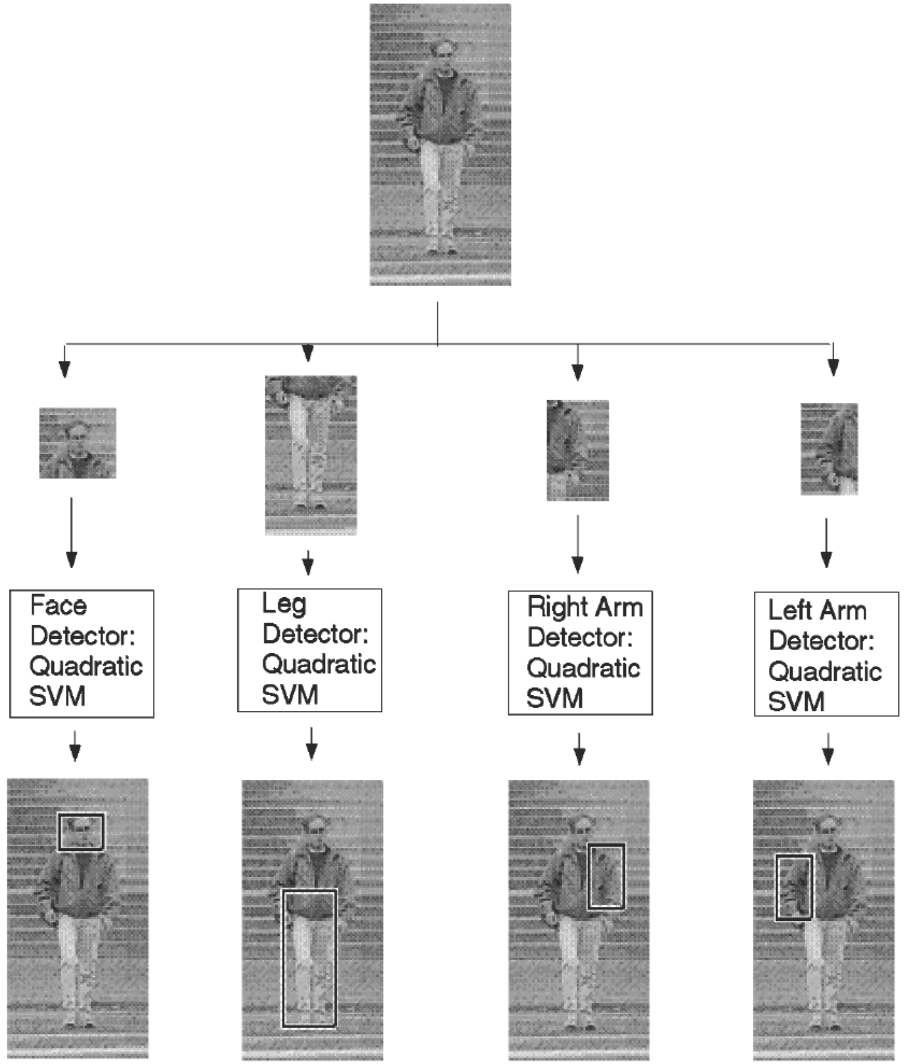
Manually Selected Parts

- Simplest solution
  - Let a human expert select a set of parts
  - (If it doesn't work, take a different human expert)

Mohan, Papageorgiou, Poggio, '01

Example 1: Detection by Components

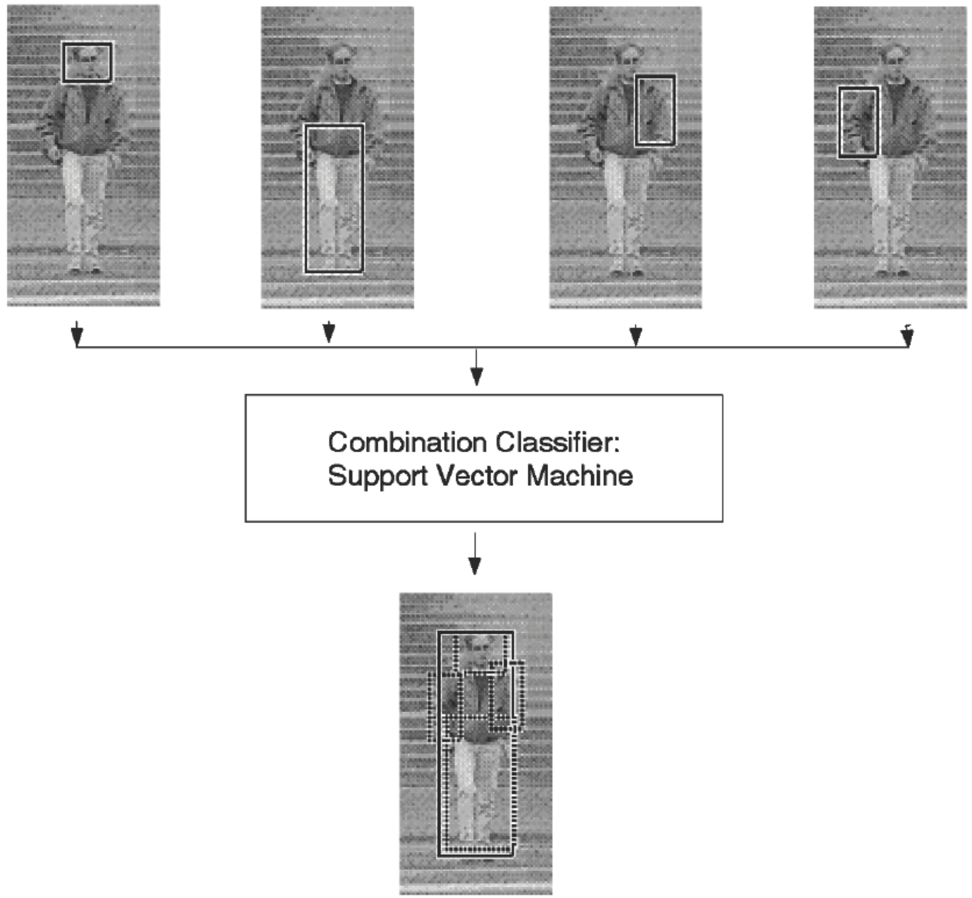
- Application
  - Pedestrian detection
- Representation by 4 parts
  - Part candidates are selected by a human expert
  - Part detectors are learned and applied independently
  - The "most suitable" head, leg, and arms are identified by the part detectors



Mohan, Papageorgiou, Poggio, '01

Example 1: Detection by Components

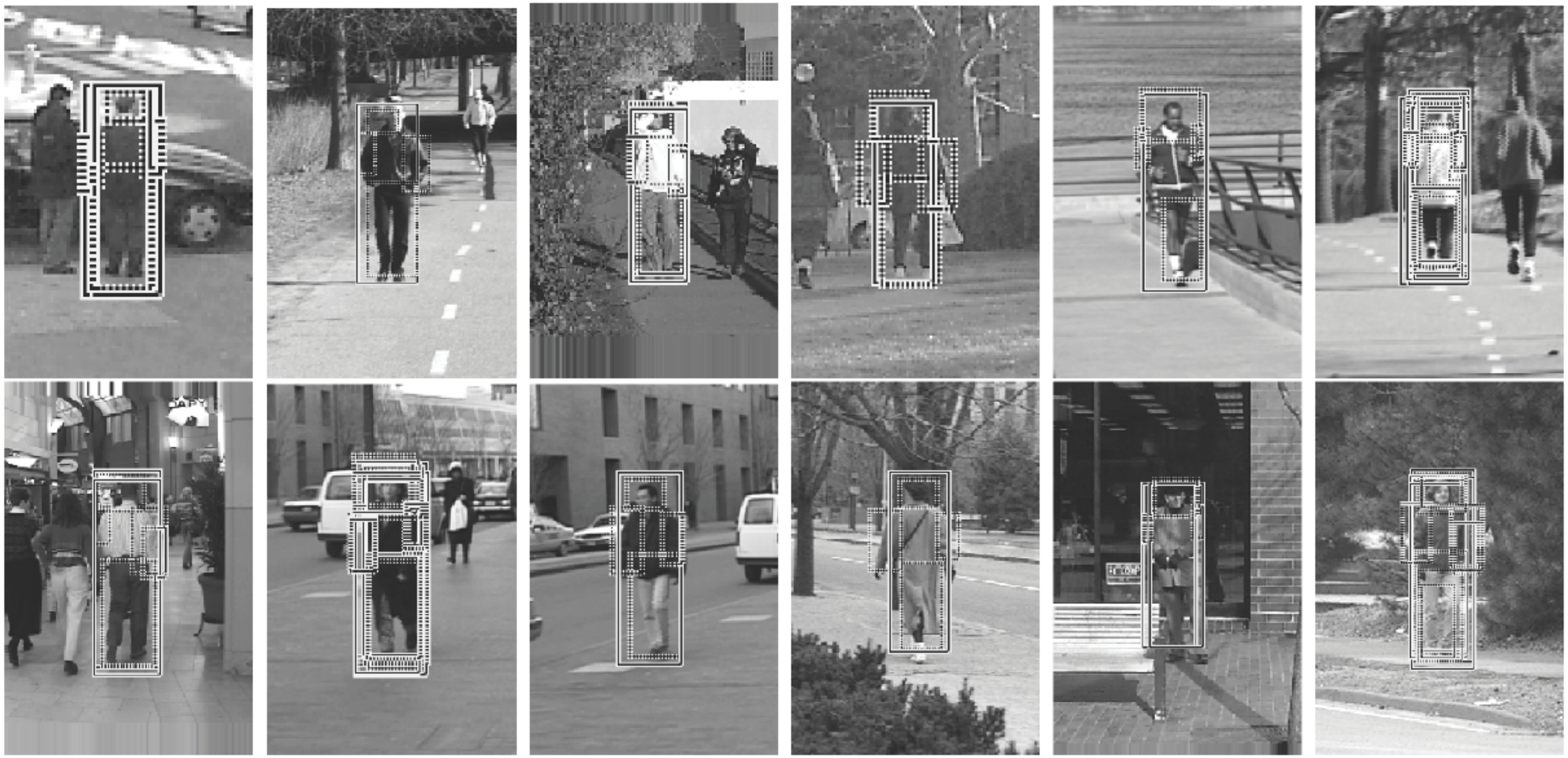
- "Structural model" via a Combination Classifier (stacking)
  - Part scores are fed into the combination classifier
  - Combination classifier classifies the pattern as "person" or "non-person"
  - The person is detected as an ensemble of its parts



Mohan, Papageorgiou, Poggio, '01

Example 1: Detection by Components

- Detection results



Mohan, Papageorgiou, Poggio, '01

Example 1: Detection by Components

- Robustness to occlusion
  - System still detects pedestrians if a part is not visible

Mohan, Papageorgiou, Poggio, ‘01

Discussion

- Approach
  - Manually selected set of parts - Specific detector trained for each part
  - Spatial model trained on part activations
  - Evaluate joint likelihood of part activations
- Advantages
  - Parts have intuitive meaning.
  - Standard detection approaches can be used for each part (e.g. SVMs or AdaBoost).
  - Works well for specific categories.
- Disadvantages
  - Parts need to be selected manually
    - Semantically motivated parts sometimes don't have a simple appearance distribution
    - No guarantee that some important part hasn't been missed
  - When switching to another category, the model has to be rebuilt from scratch.

$\Rightarrow$ Goal: Model that can be automatically learned for many categories

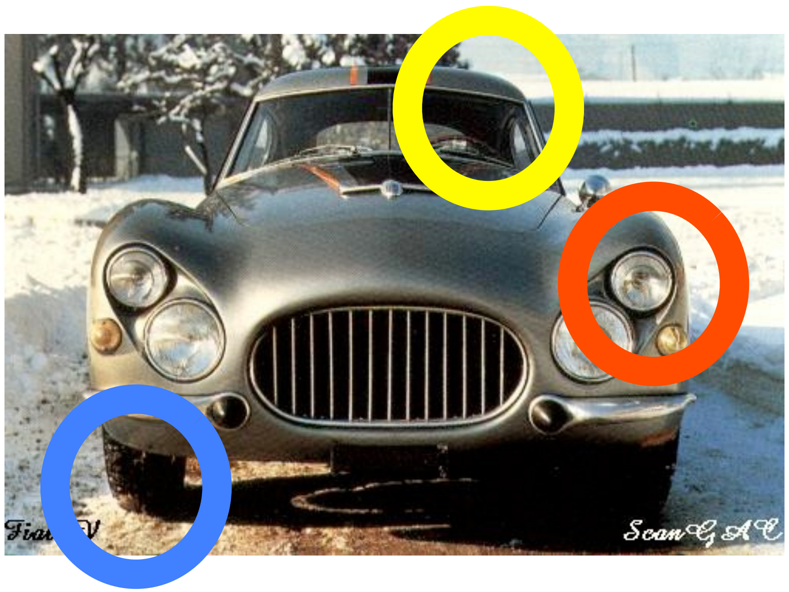
Part-Based Models - Overview Today (more next week)

- Part-Based using Manual Labeling of Parts
  - Detection by Components
  - Multi-Scale Parts

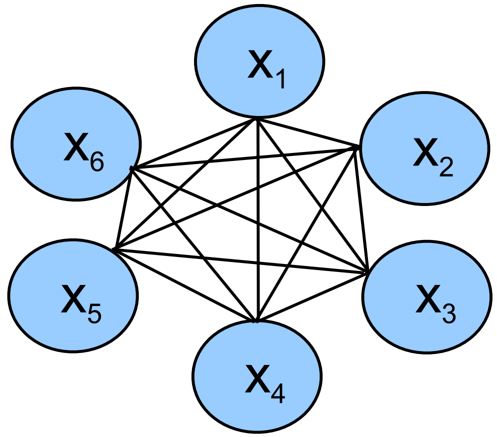
- The Constellation Model
  - automatic discovery of parts and part-structure

- The Implicit Shape Model (ISM)
  - parts obtained by clustering interest-points
  - star-model to model configuration of parts

Constellation of Parts



Fully connected shape model



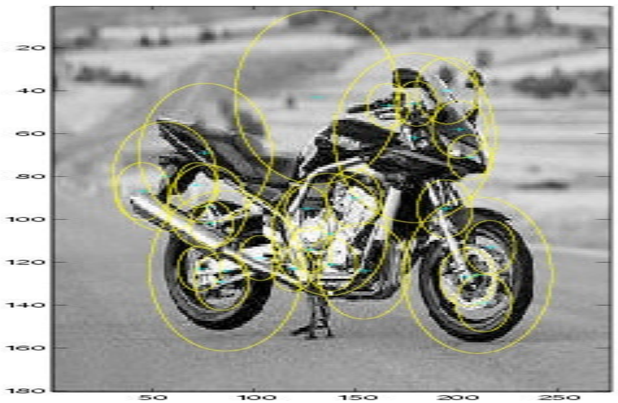
Weber, Welling, Perona, ’00;
Fergus, Zisserman, Perona, 03

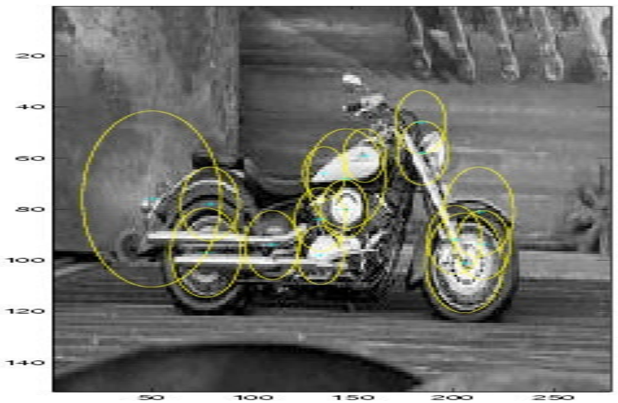
Automatic Part Learning

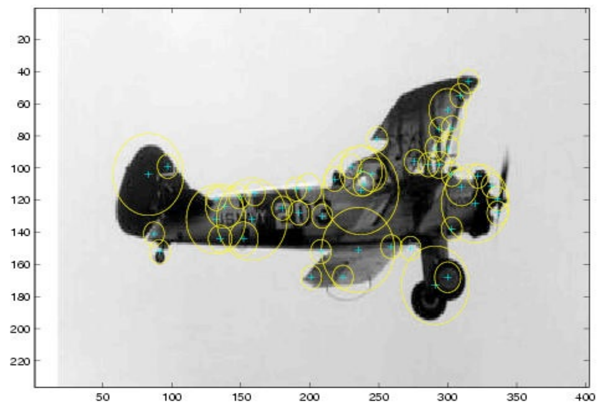
- Basic idea consists of two steps

Fergus, Zisserman, Perona, ‘03

  - “Part” candidates in each image
    - take the output regions of an interest point detector as part candidates (use scale-invariant interest point detector for that).
    - interest point detector “guarantees” (sort of ;-) that similar structures will be detected in all images (keyword: repeatability)
  - “Part learning”
    - find those regions, that occur repeatedly on different instances of the same object:
    - for this: group (=cluster) the extracted regions to find those that are characteristic for the object category.

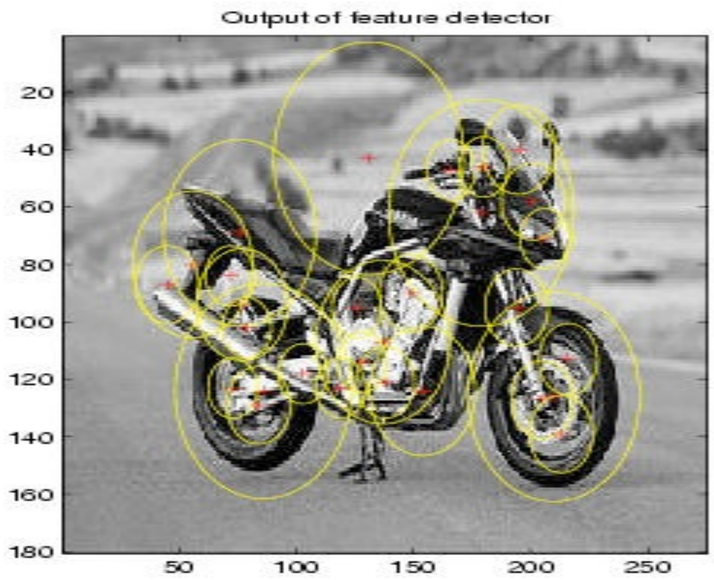


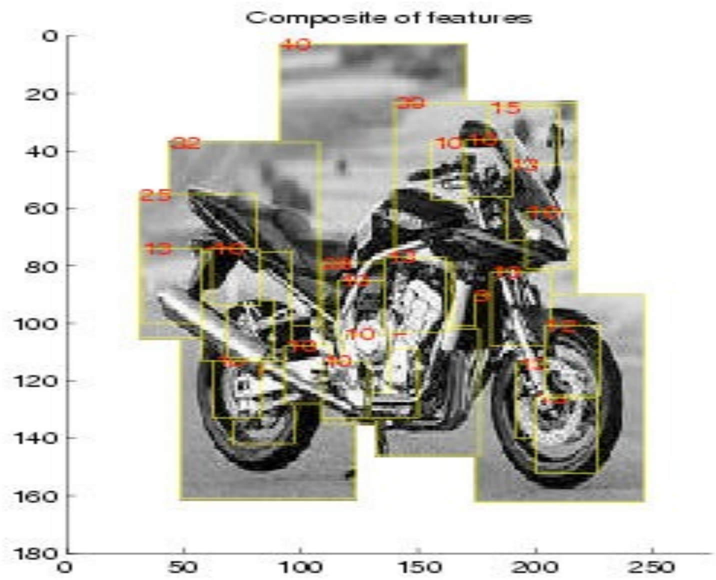




Representation of Appearance

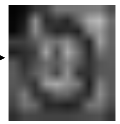
Fergus, Zisserman, Perona, '03





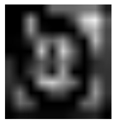
interest point detection

11x11 patch →



size normalized

Normalize →

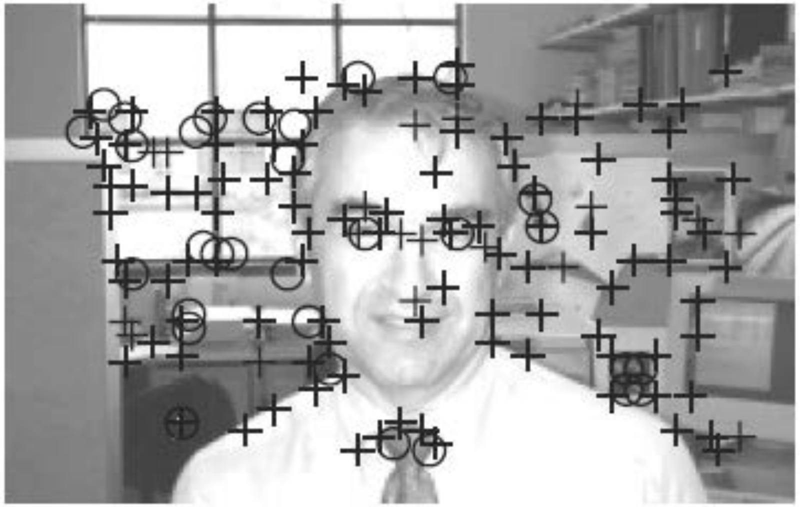


luminance normalized

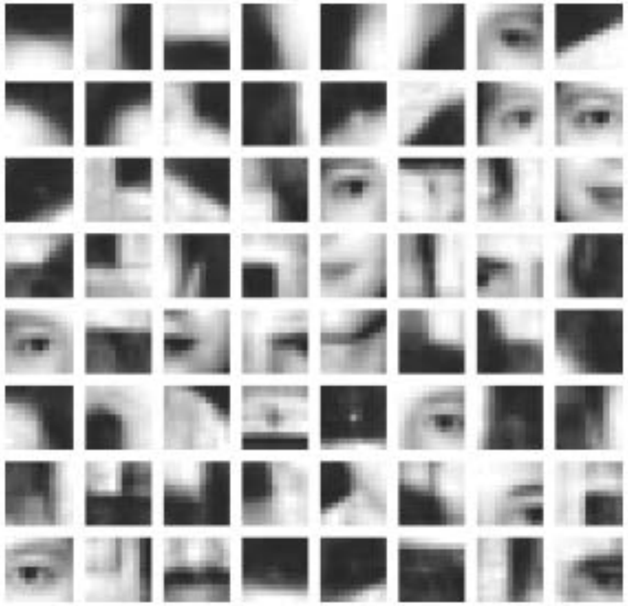
Projection onto PCA basis →

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{15} \end{pmatrix}$$

Selected Features & “Parts” (=feature clusters)



interest points



100 clusters

Weber, Welling, Perona, ‘00

Weakly Supervised Training

200 images containing faces

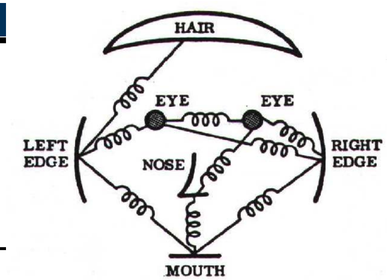
200 background images

- Repeating structures (clusters in appearance space and in location space) are more likely to belong to the object category than to the background.
  ⇒ Clusters should mainly represent objects.

Weber, Welling, Perona, '00

Constellation Model

Weber, Welling, Perona, '00;
Fergus, Zisserman, Perona, 03

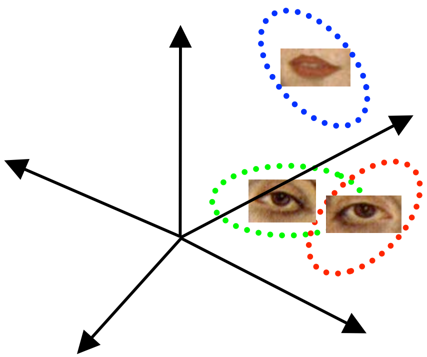


- **Joint model** for **appearance** and **structure** (=shape)
  - X: positions, A: part appearance, S: scale
  - h: Hypothesis = assignment of features (in the image) to parts (of the model)

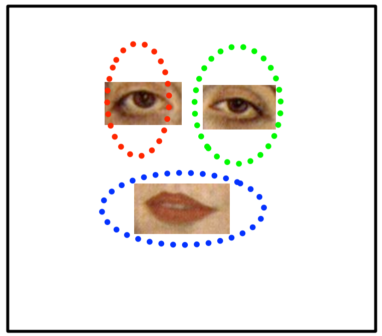
$$p(\mathbf{X},\mathbf{S},\mathbf{A}|\,\theta) = \sum_{\mathbf{h}\in H} p(\mathbf{X},\mathbf{S},\mathbf{A},\mathbf{h}|\,\theta)$$

$$= \sum_{\mathbf{h}\in H} \underbrace{p(\mathbf{A}|\mathbf{X},\mathbf{S},\mathbf{h},\theta)}_{Appearance}\,\underbrace{p(\mathbf{X}|\mathbf{S},\mathbf{h},\theta)}_{Shape}\,\underbrace{p(\mathbf{S}|\mathbf{h},\theta)}_{Rel.\ Scale}\,\underbrace{p(\mathbf{h}|\theta)}_{Other}$$

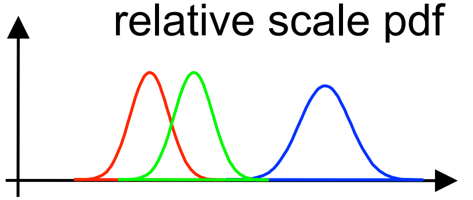
Gaussian part appearance pdf



Gaussian shape pdf

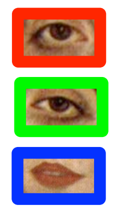


Gaussian relative scale pdf



Log(scale)

Prob. of detection



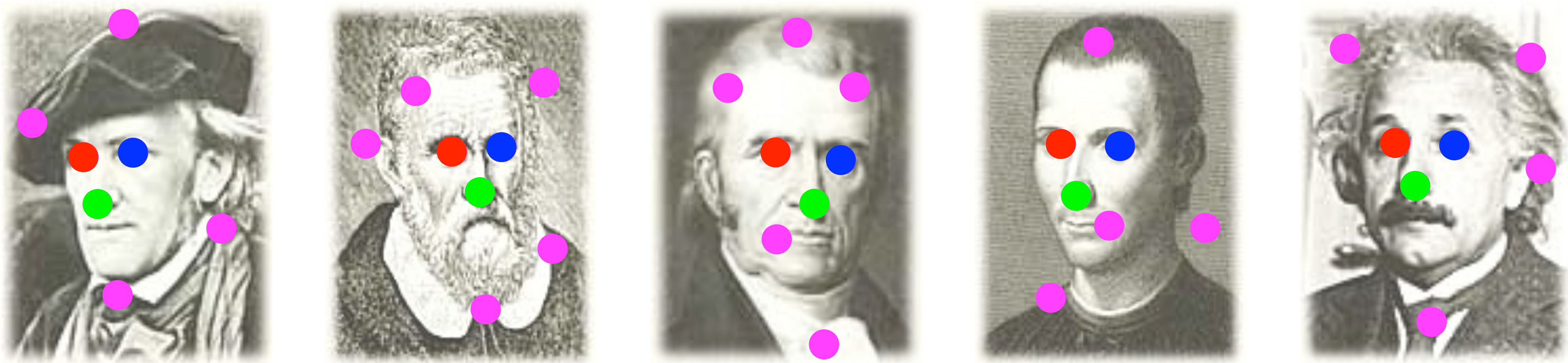
Training Procedure

- Need to solve two problems
  - Select a subset of appearance clusters as **part candidates**
    - Greedy strategy
    - Start with 3-part model, then test if additional part improves the results
  - **Learn** the parameters of their **joint probability** density over **appearance & structure**
    - Expectation Maximization (EM) algorithm

Weber, Welling, Perona, ‘00

Learning

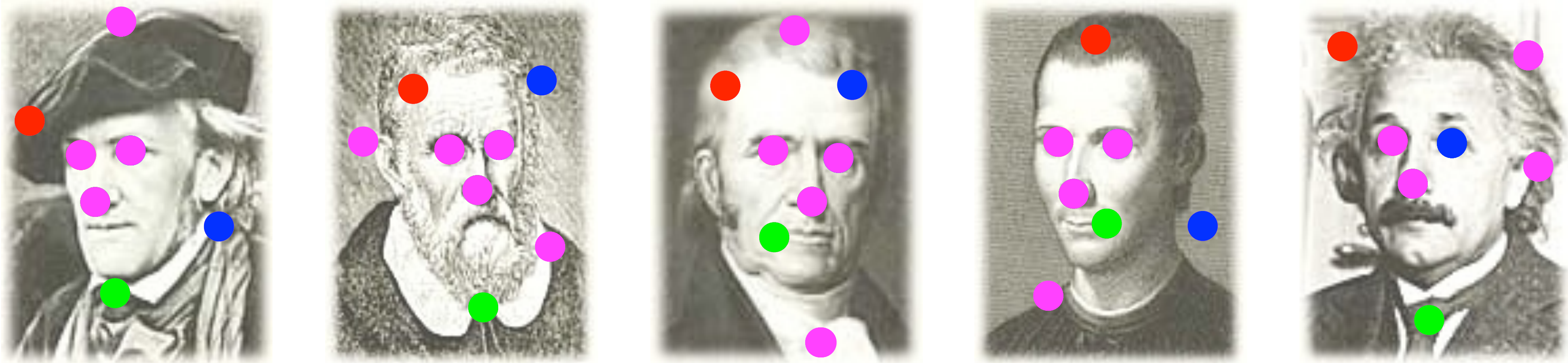
- Task: Estimation of model parameters
- Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to foreground/background
- Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters



Learning Procedure

- Find regions: their location, scale & appearance
- Initialize model parameters
- Use EM and iterate to convergence
  - E-step: Compute assignments for which regions are foreground/background
  - M-step: Update model parameters

- Trying to maximize likelihood – consistency in shape & appearance



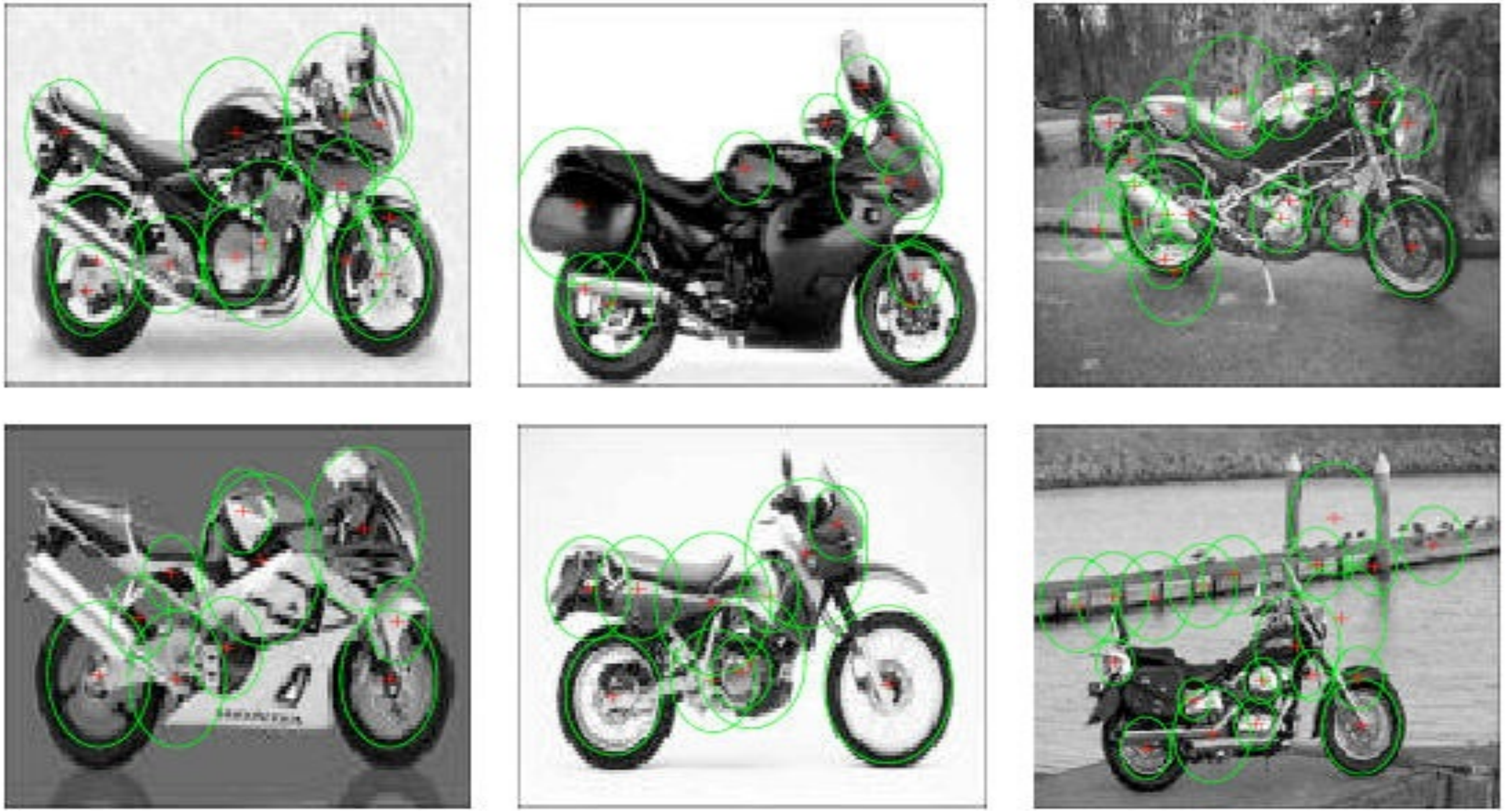
Experiments

- Data sets
  - Motorbikes, Airplanes, Faces, Cars from side and behind, Spotted cats
  - and background images
  - Between 200 and 800 images per category

- Training
  - 50% of images
  - position of object unknown within image (called weakly supervised)
- Testing
  - 50% of images
  - Simple object present/absent test
  - ROC equal error rate computed, using background set of images

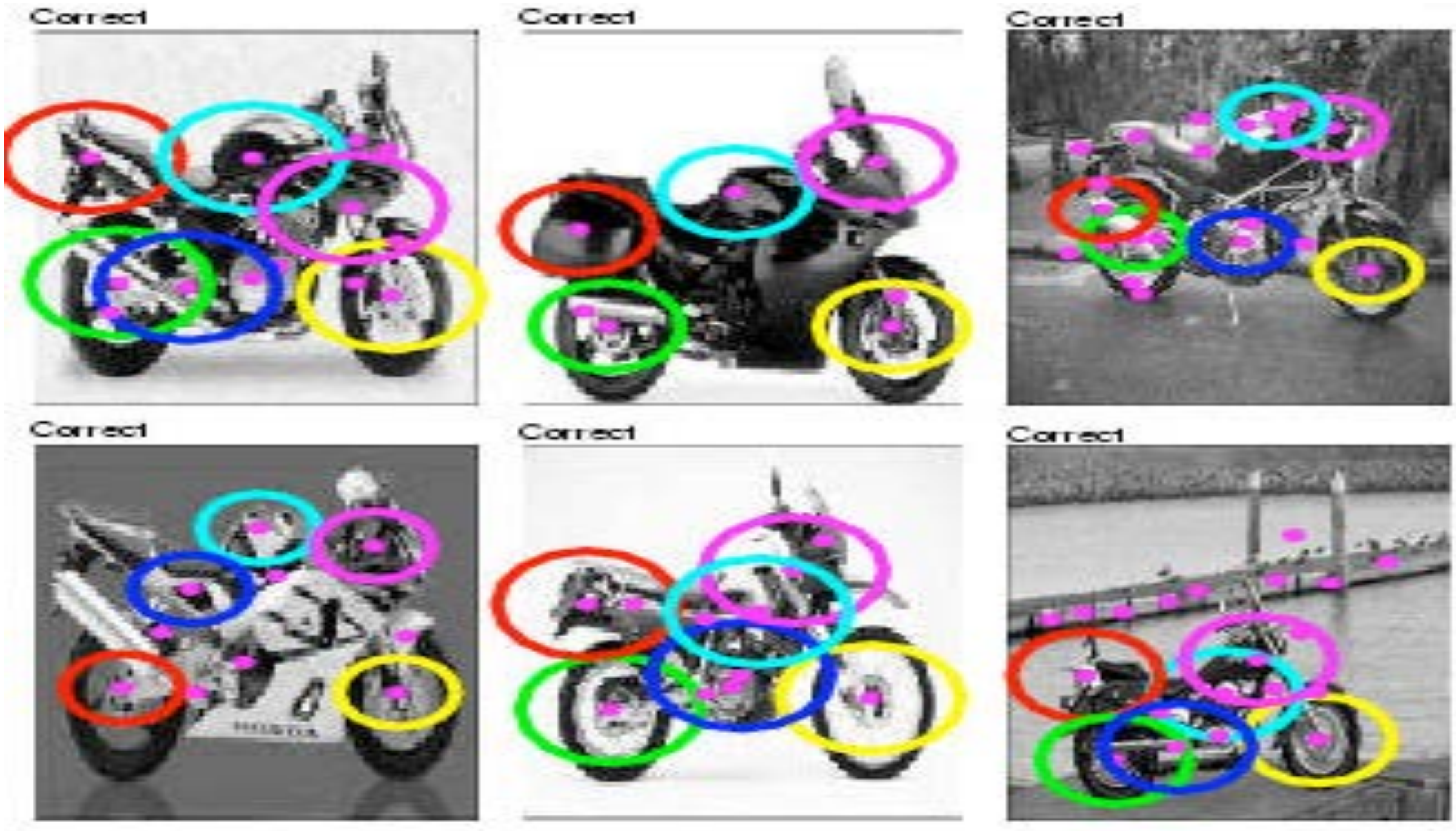
Fergus, Zisserman, Perona, '03

Example: Motorbikes - Part Hypotheses



Fergus, Zisserman, Perona, ‘03

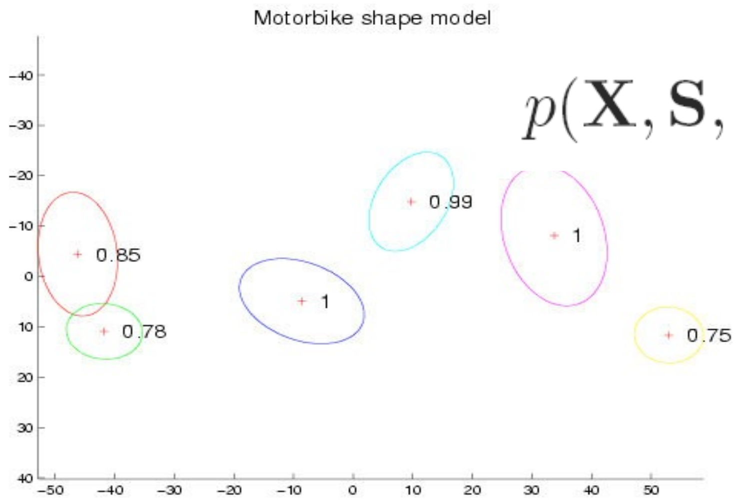
Example: Motorbikes - Learned Parts



Fergus, Zisserman, Perona, ‘03

Equal error rate: 7.5%

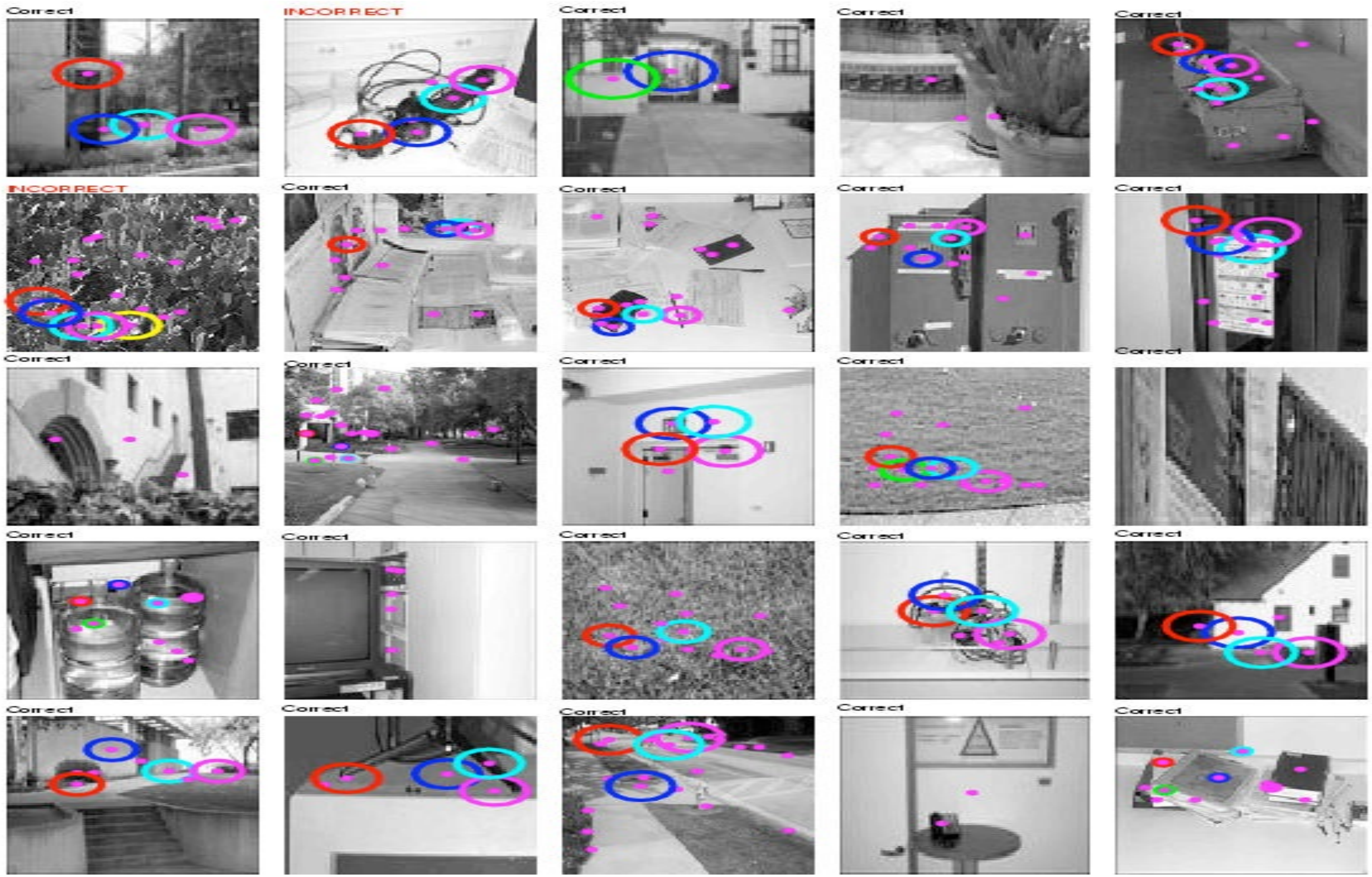
Motorbikes - Constellation Model



$$p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \mathbf{h} | \theta)$$

Fergus, Zisserman, Perona, '03

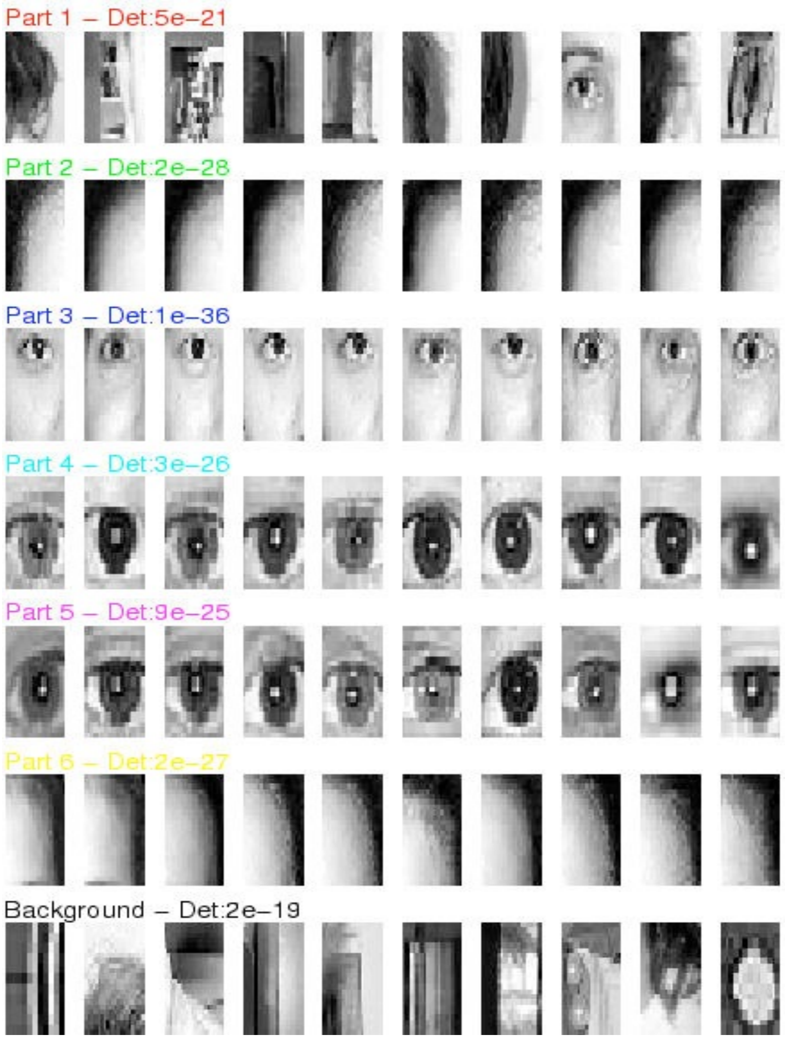
Background Images

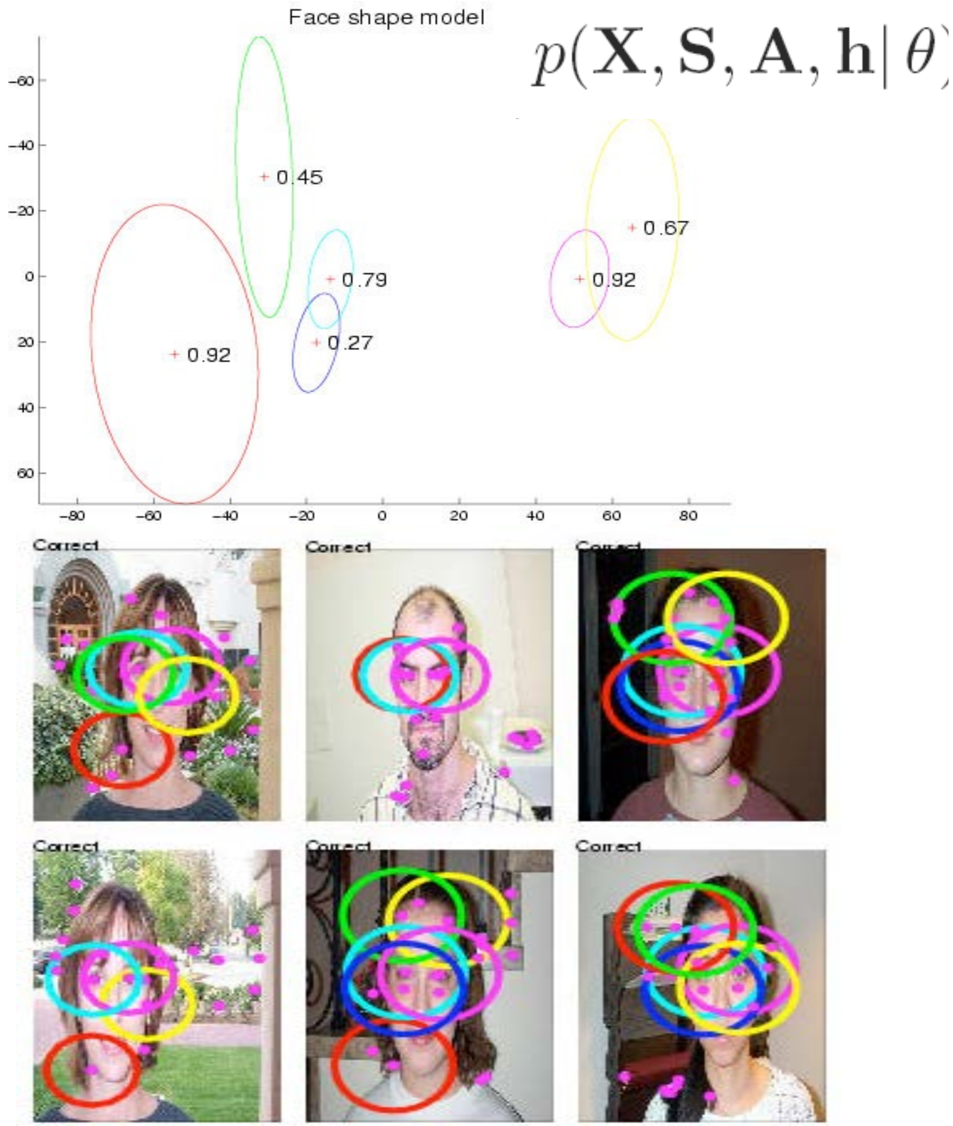


Fergus, Zisserman, Perona, ‘03

Equal error rate: 4.6%

Frontal Faces - Constellation Model

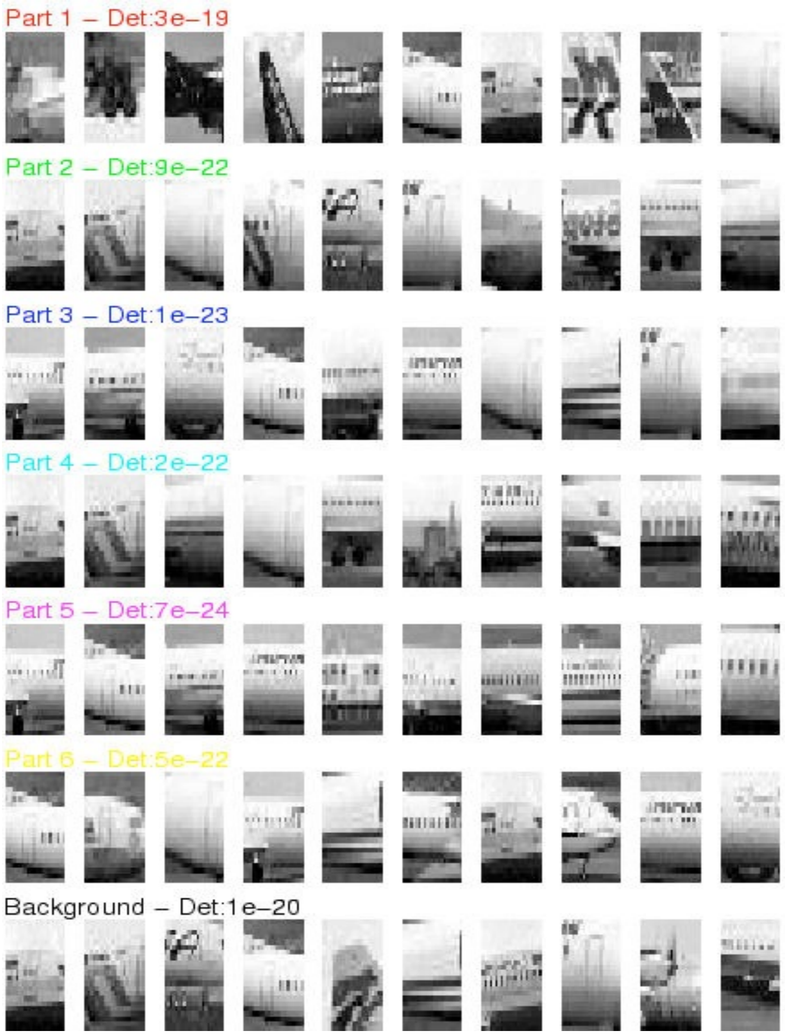


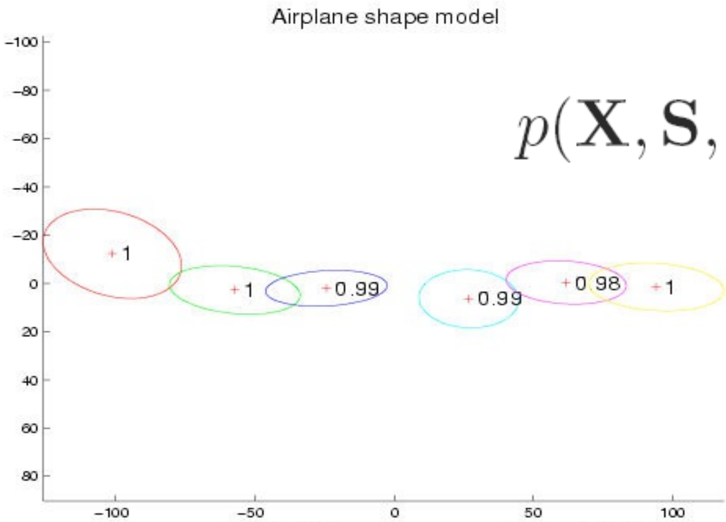


Fergus, Zisserman, Perona, ‘03

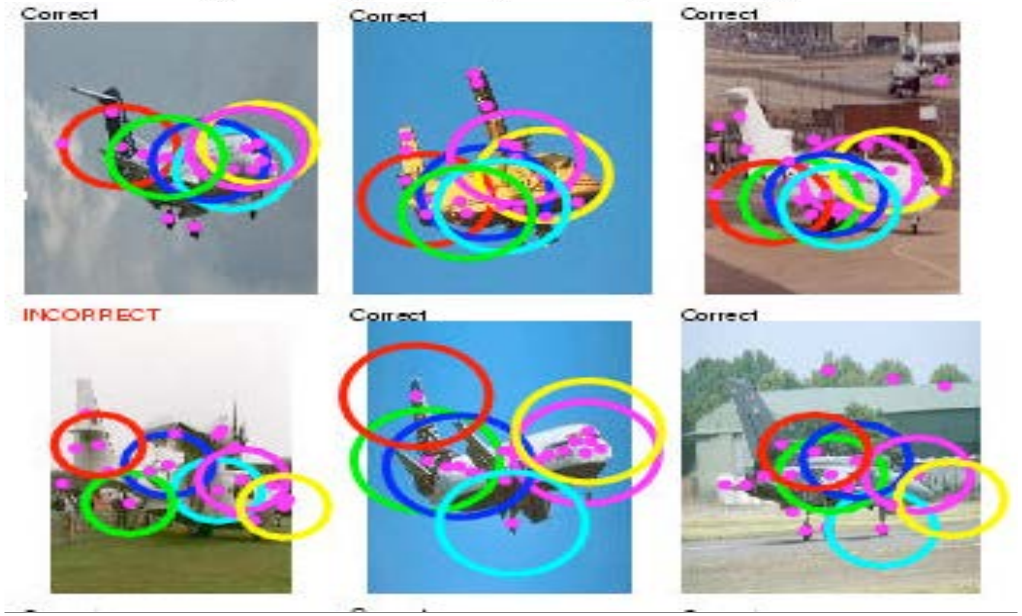
Equal error rate: 9.8%

Airplanes - Constellation Model





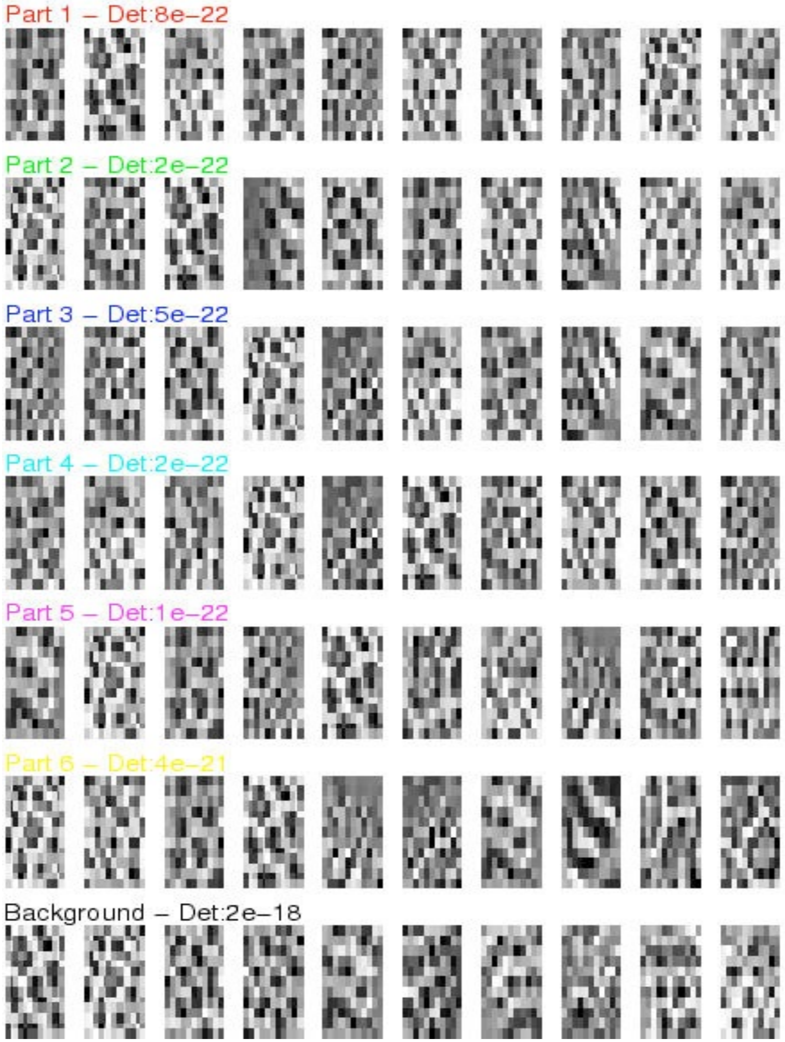
$$p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \mathbf{h} | \theta)$$

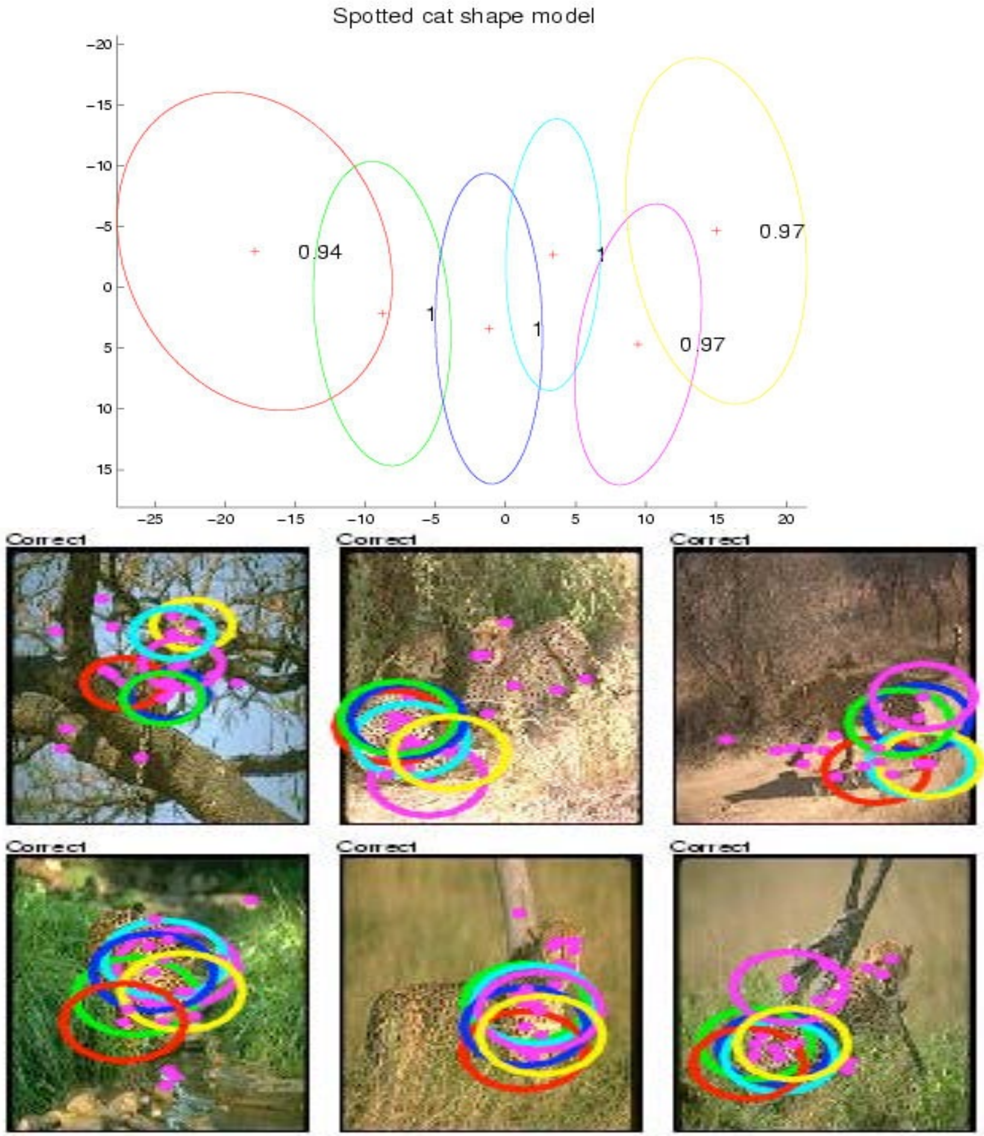


Fergus, Zisserman, Perona, '03

Equal error rate: 10.0%

Spotted Cats - Constellation Model

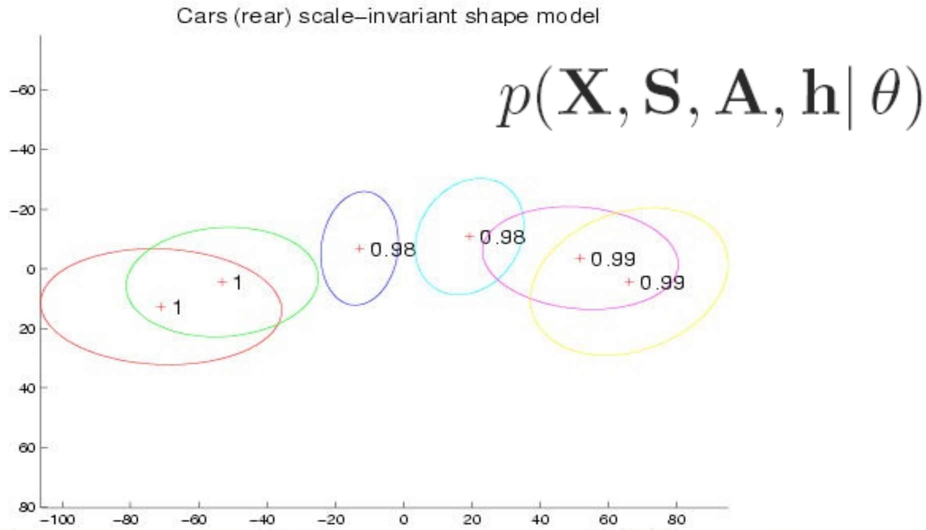


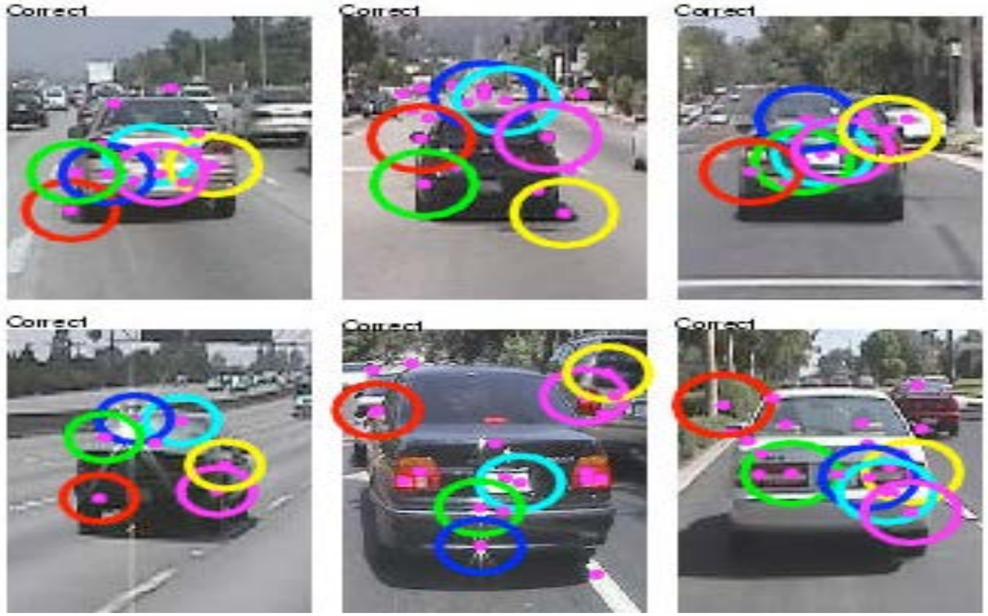


Fergus, Zisserman, Perona, '03

Equal error rate: 9.7%

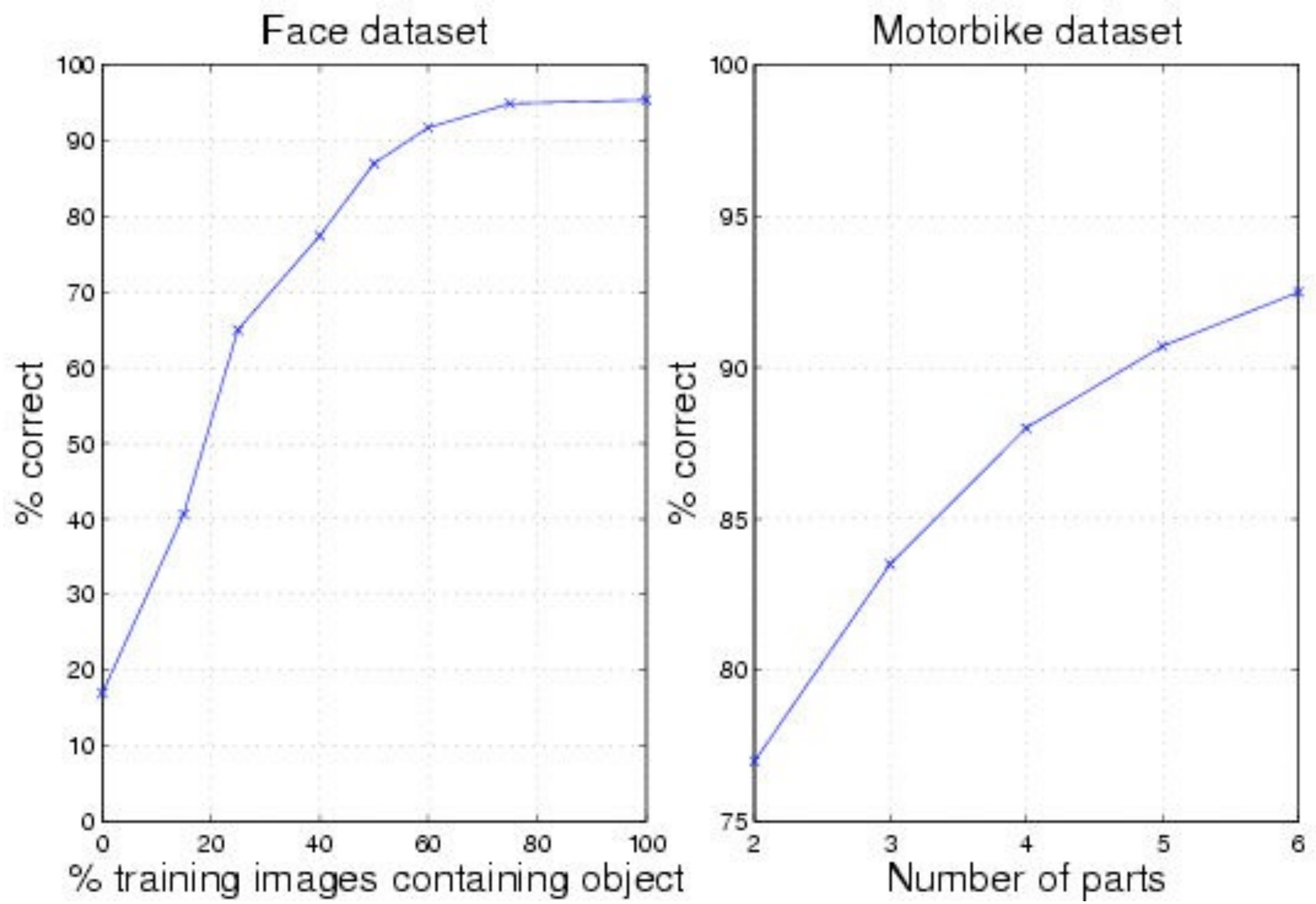
Cars (Rear Views) - Constellation Model





Fergus, Zisserman, Perona, '03

Robustness of the Algorithm



Fergus, Zisserman, Perona, '03

Discussion

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

- Disadvantages
  - Model contains many parameters that need to be estimated
  - Cost increases exponentially with increasing number of parameters (that is in particular with the # of parts !)

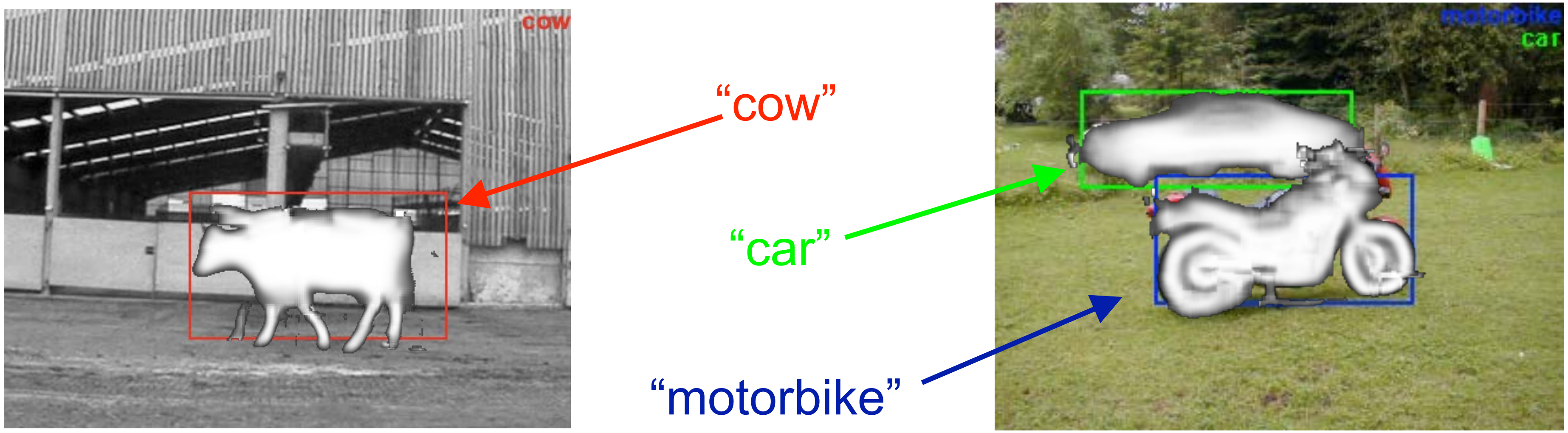
Part-Based Models - Today

- Part-Based using Manual Labeling of Parts
  - Detection by Components
  - Multi-Scale Parts

- The Constellation Model
  - automatic discovery of parts and part-structure

- The Implicit Shape Model (ISM)
  - parts obtained by clustering interest-points
  - star-model to model configuration of parts

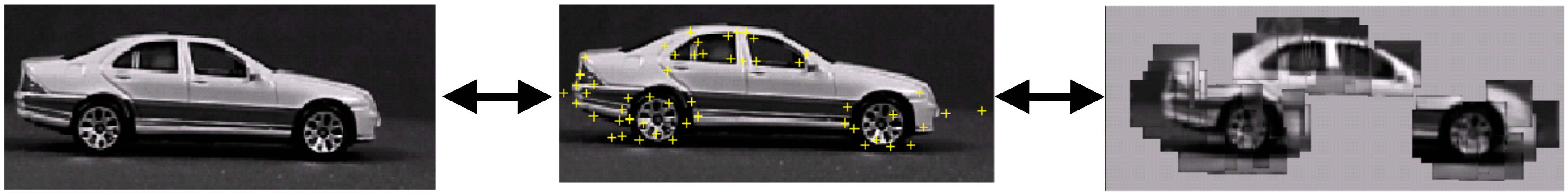
Implicit Shape Model: Object Categorization



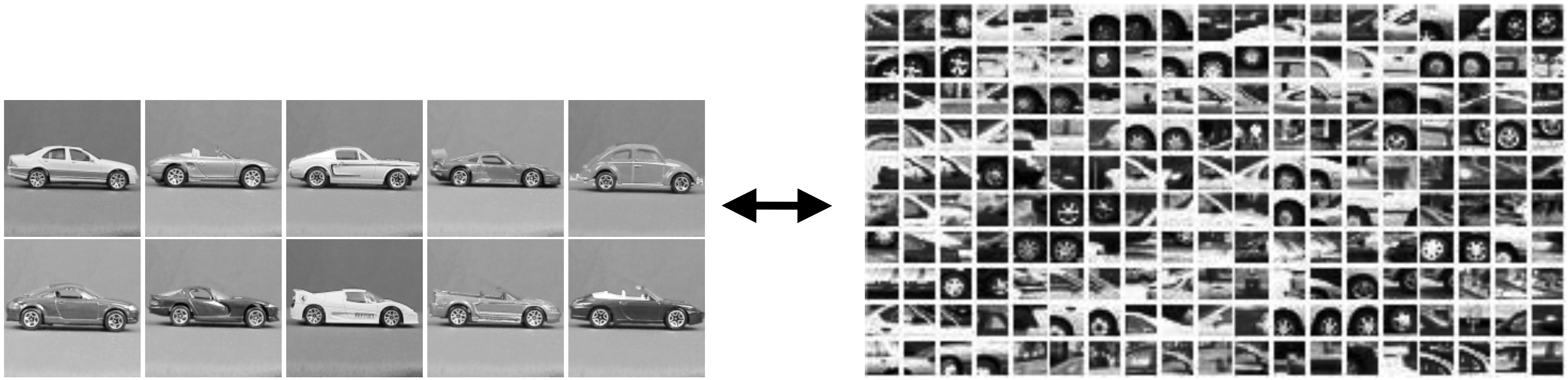
- Goals
  - Learn to recognize object categories
  - Detect and localize them in real-world scenes
  - Segment objects from background
- Combination with top-down segmentation
  - Initial hypothesis generation
  - Category-specific figure-ground segmentation - used to verify object hypothesis

Codebook Representation

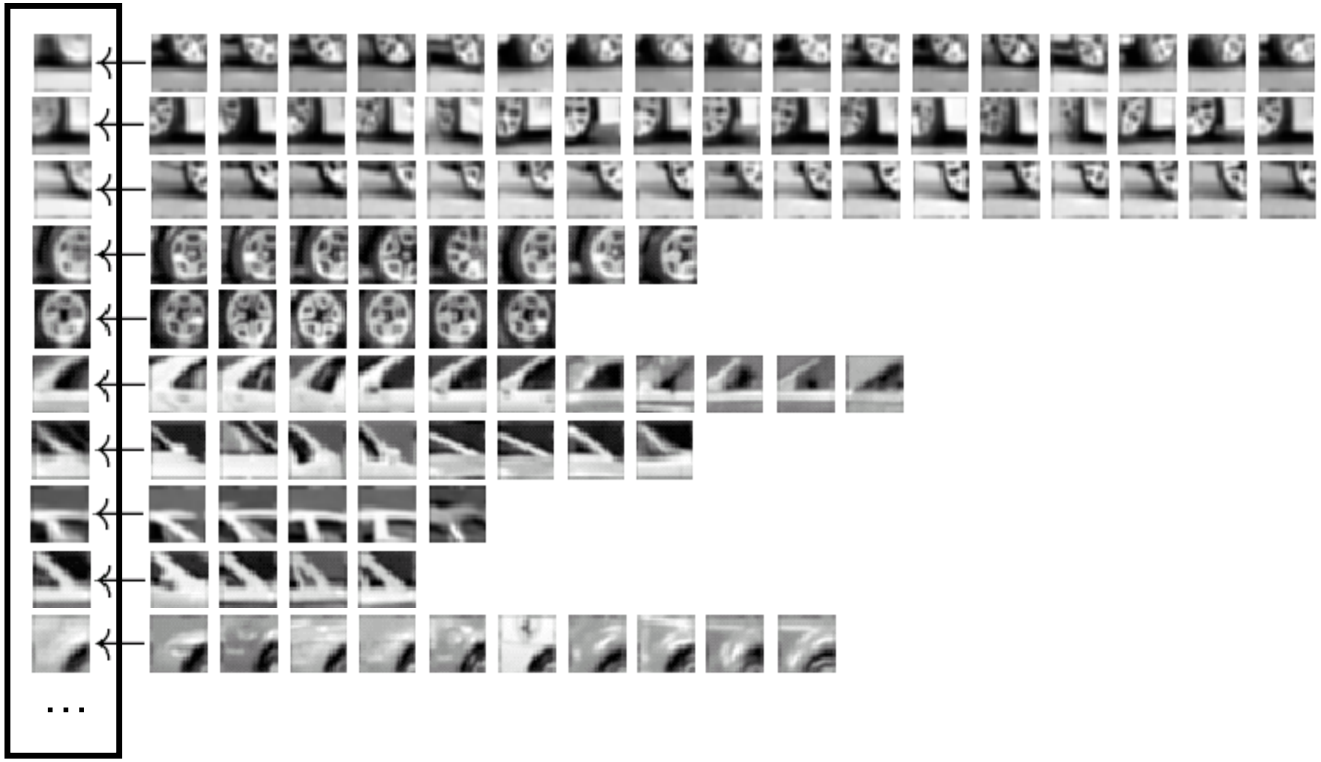
- Extraction of local object patches
  - Interest Points (e.g. Harris detector, Hes-Lap, DoG, ...)
  - inspired by [Agarwal & Roth, 02]



- Collect patches from whole training set
  - Example:



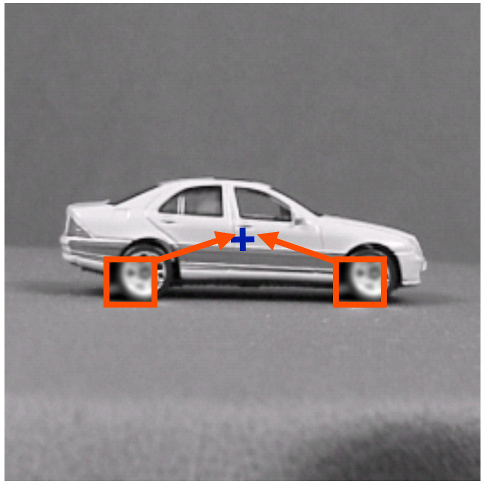
Appearance Codebook



- Clustering Results
  - Visual similarity preserved
  - Wheel parts, window corners, fenders, ...
  - Store cluster centers as Appearance Codebook

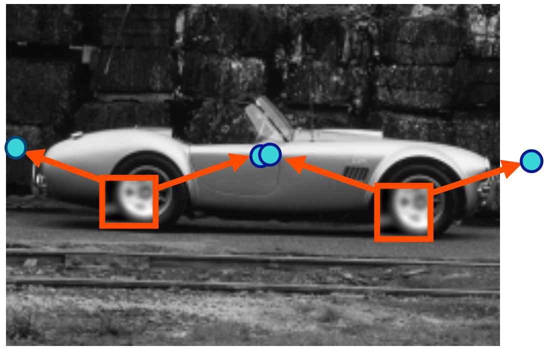
Learning the Spatial Layout

- For every codebook entry, store possible “occurrences”



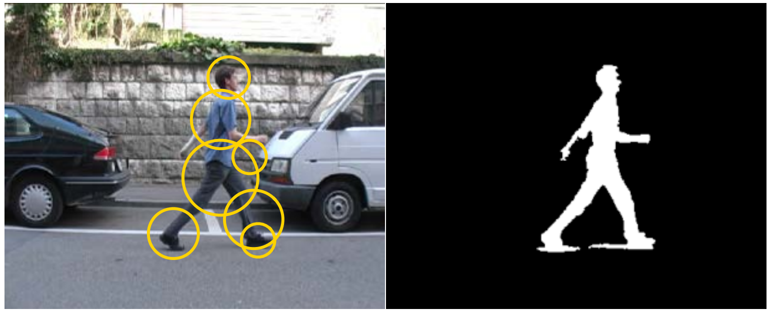
- Object identity
- Pose
- Relative position

For new image, let the matched patches vote for possible object positions

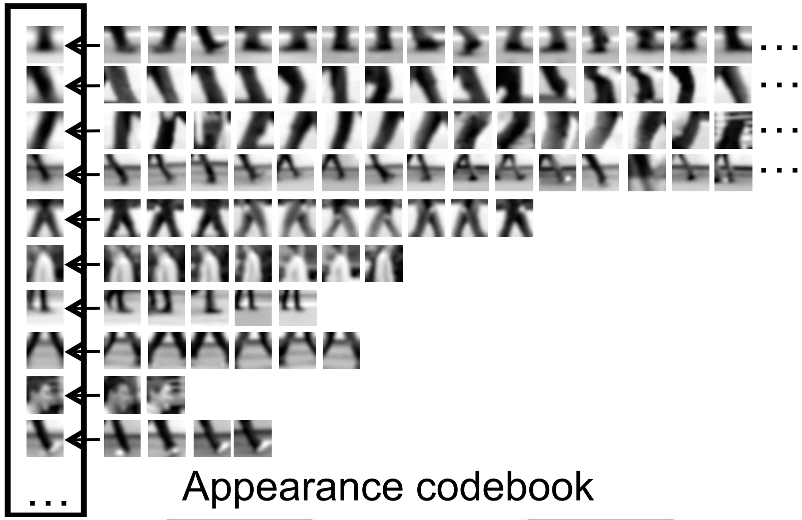


- Object identity
- Pose
- Relative position

Implicit Shape Model - Representation



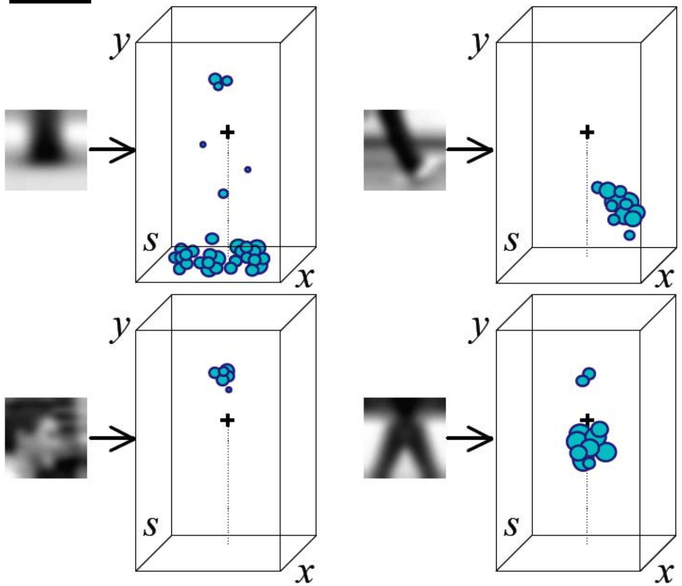
105 training images
(+motion segmentation)



Appearance codebook

- Learn appearance codebook
  - Extract patches at DoG interest points
  - Agglomerative clustering ⇒ codebook

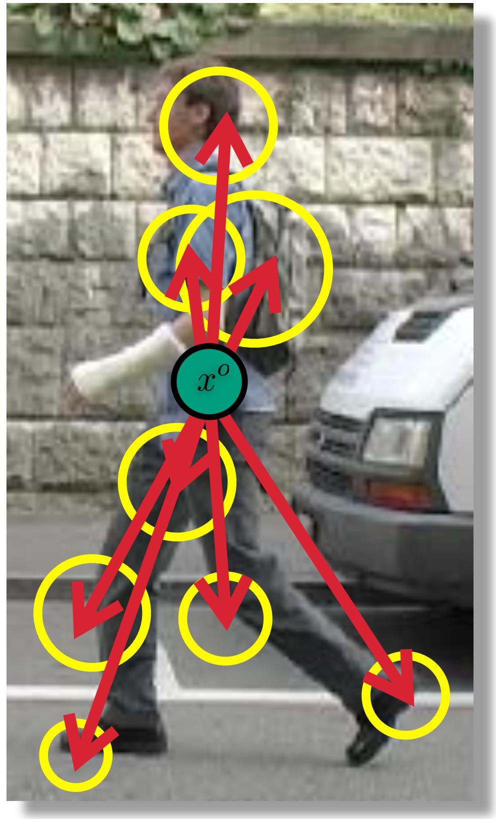
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object



Spatial occurrence distributions

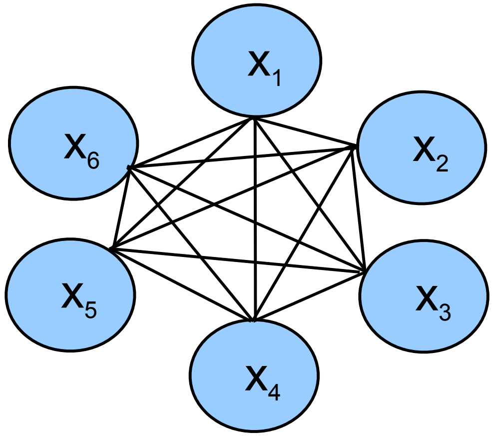
Object Detection: ISM (Implicit Shape Model)

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]



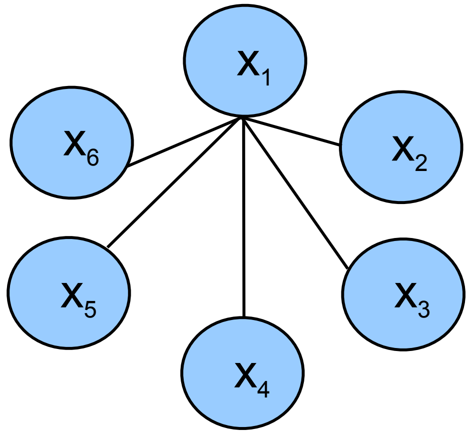
Spatial Models for Categorization

Fully connected shape model



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

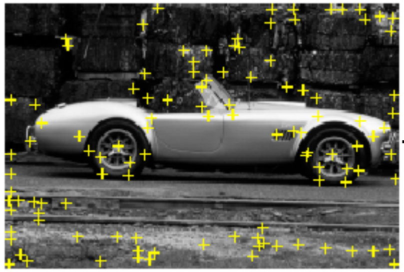
"Star" shape model



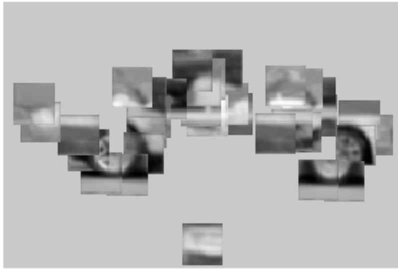
- e.g. ISM (Implicit Shape Model)
- Parts mutually independent
- Recognition complexity: O(NP)
- Method: Generalized Hough Transform

Object Categorization Procedure

Interest Points



Matched Codebook Entries



Probabilistic Voting

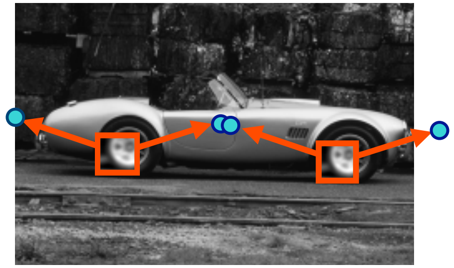
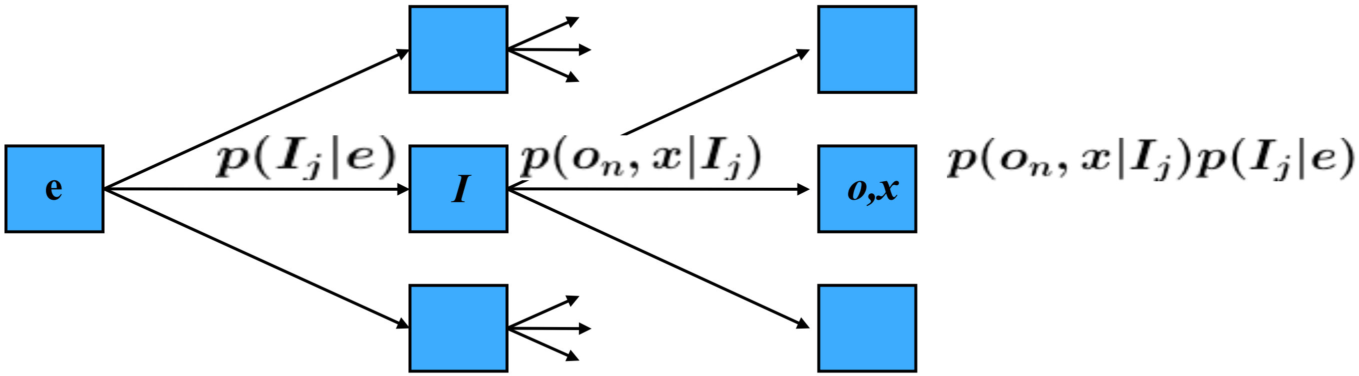


Image Patch

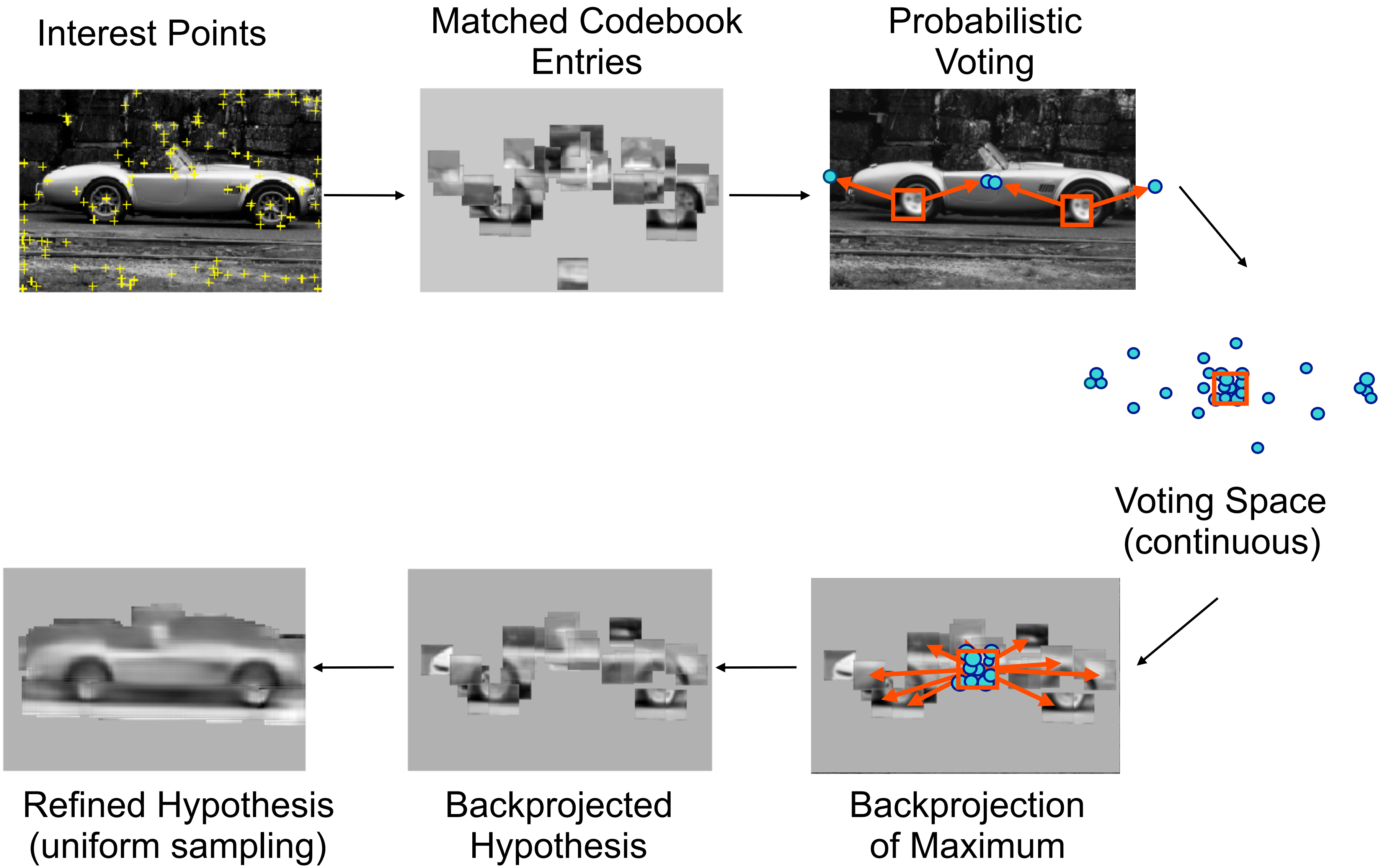
Interpretation (Codebook match)

Object Position



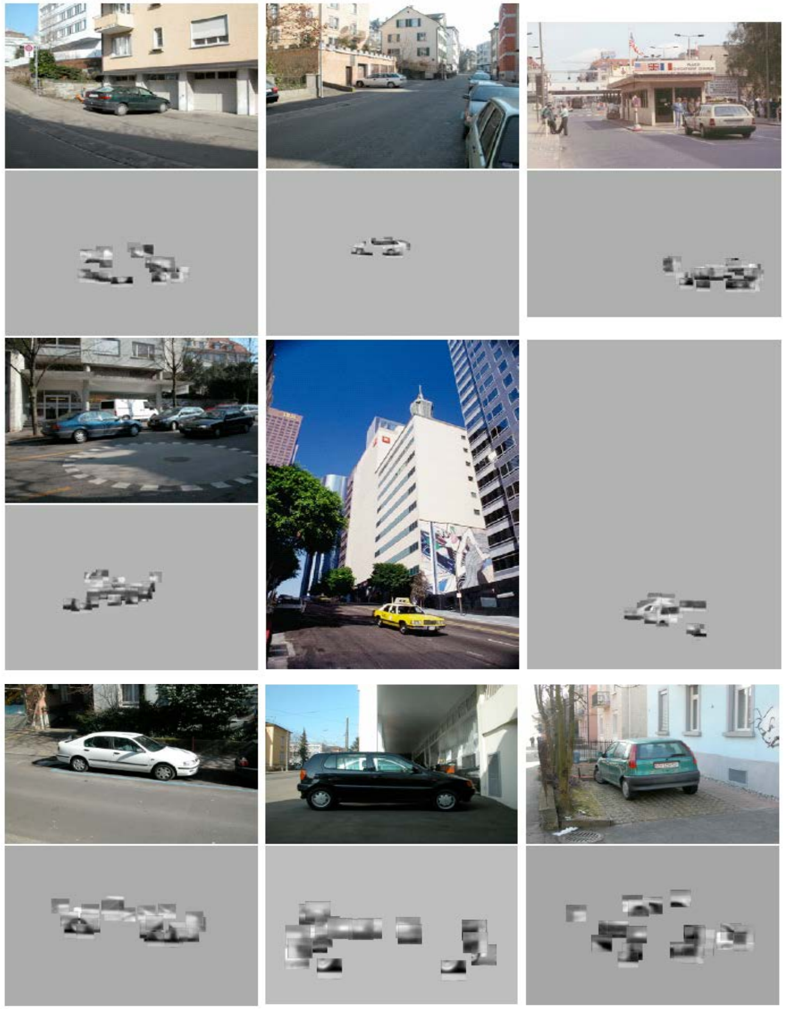
$$p(o_n, x|e) = \sum_j p(o_n, x|I_j)p(I_j|e)$$

Object Categorization Procedure

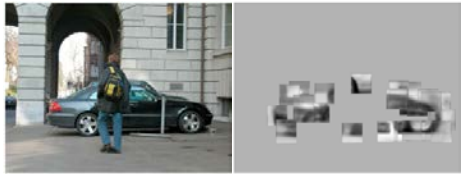


Car Categorization - Qualitative Results

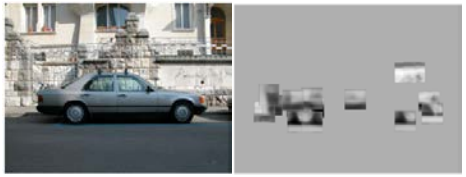
- 1st hypothesis



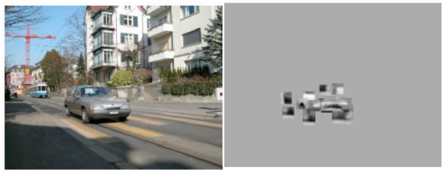
2nd hypothesis



4th hypothesis

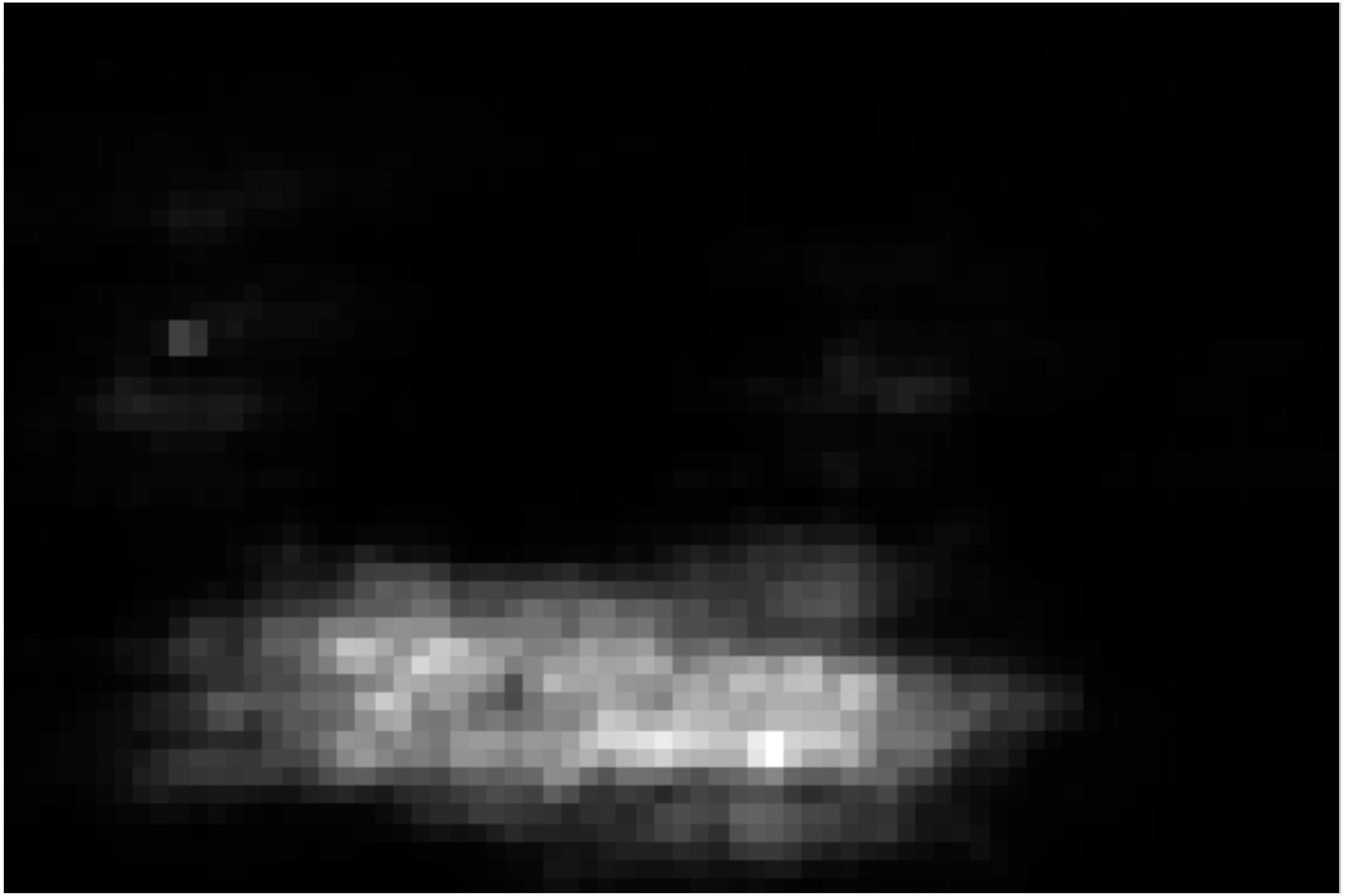


7th hypothesis



8th hypothesis

Results on Cows



Prob. Votes

Results on Cows

1’st hypothesis

Results on Cows

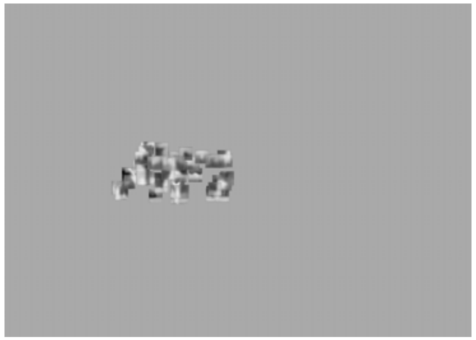
2'nd hypothesis

Results on Cows

3'rd hypothesis

More Results on Cows…

16’th hypothesis



8’th hypothesis

2’nd hypothesis

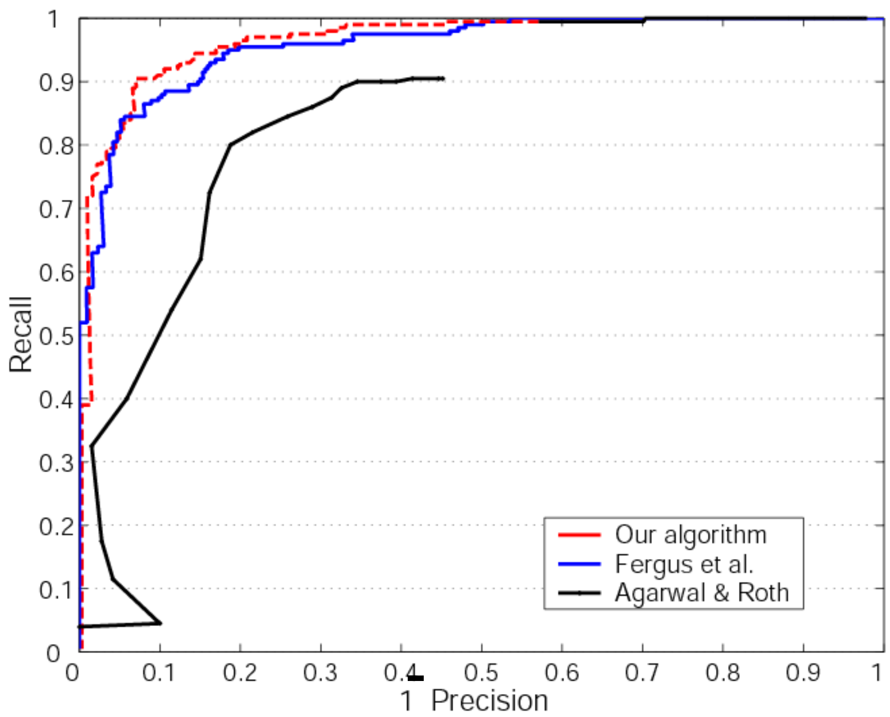
14’th hypothesis

Detection Results

- Qualitative Performance (UIUC database - 200 cars)
  - Recognizes different kinds of cars
  - Robust to clutter, occlusion, low contrast, noise

Leibe, Leonardis, Schiele, '04

Quantitative Evaluation



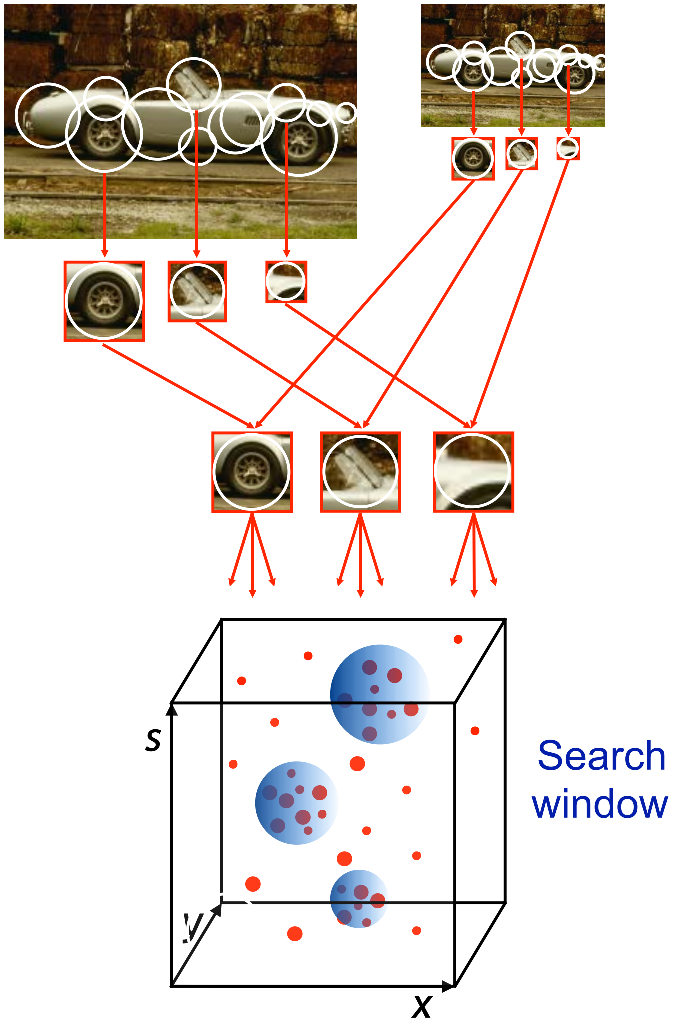
- Results on UIUC car database
  - (170 images containing 200 cars)
  - Good performance, similar to Constellation Model
  - Still some false positives

Scale Invariance

- Scale-invariant feature selection
  - Scale-invariant interest points
  - Rescale extracted patches
  - Match to constant-size codebook
- Generate scale votes
  - Scale as 3rd dimension in voting space

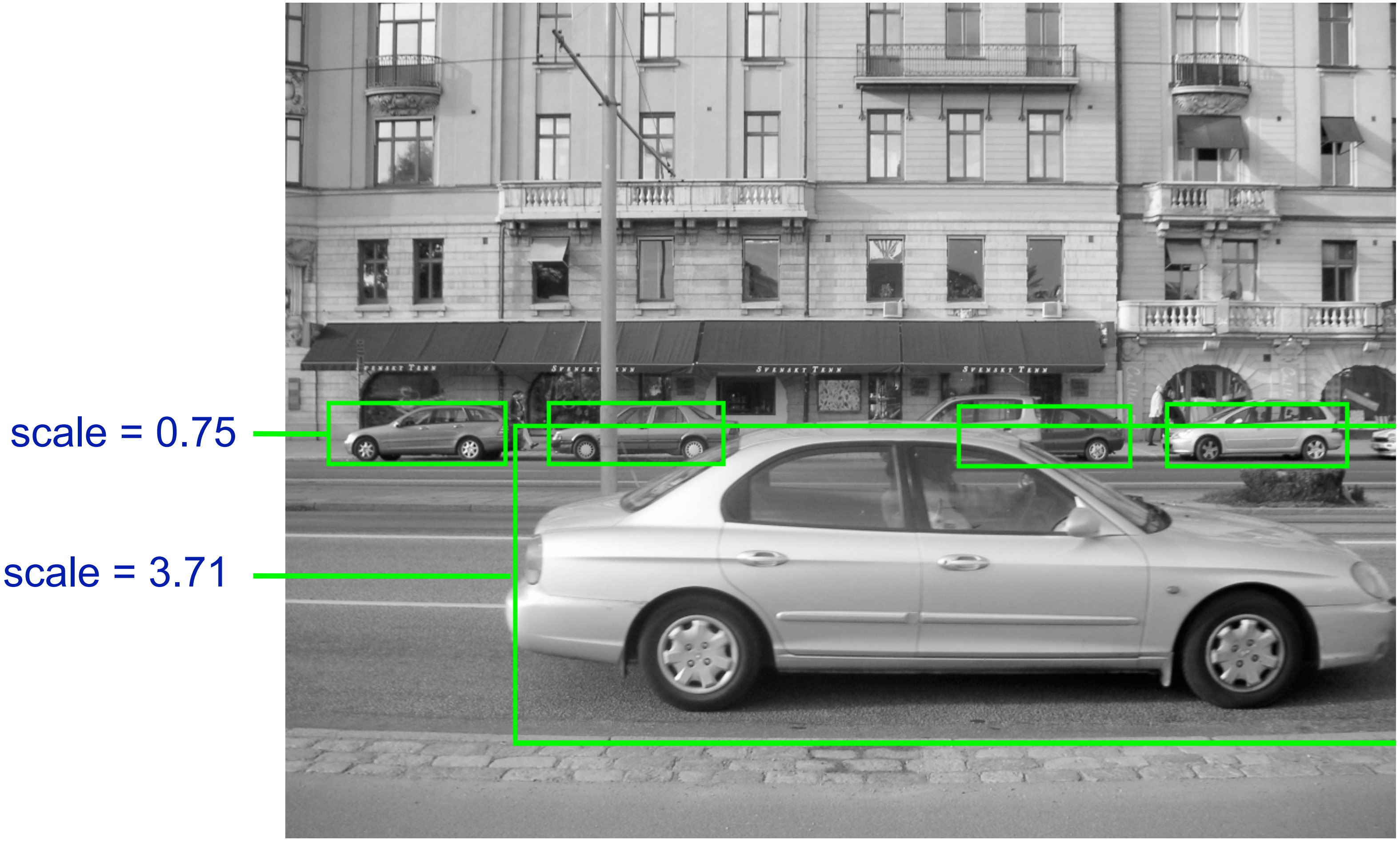
$$x_{vote} = x_{img} - x_{occ}(s_{img}/s_{occ})$$
$$y_{vote} = y_{img} - y_{occ}(s_{img}/s_{occ})$$
$$s_{vote} = (s_{img}/s_{occ})$$

  - Search for maxima in 3D voting space



Leibe, Schiele ‘04

Qualitative Detection Results



Altogether, objects detected with factor 5.0 scale differences!

Discussion

- Approach: Implicit Shape Model
  - Generate appearance codebook
  - Learn spatial occurrence distribution for each codebook entry
  - Recognition using a probabilistic extension of the Generalized Hough Transform

- Advantages
  - Highly flexible shape model
  - Each image feature acts independently
  - Possible to learn good object models already from very few (50-100) training examples
  - Recognition is fast!

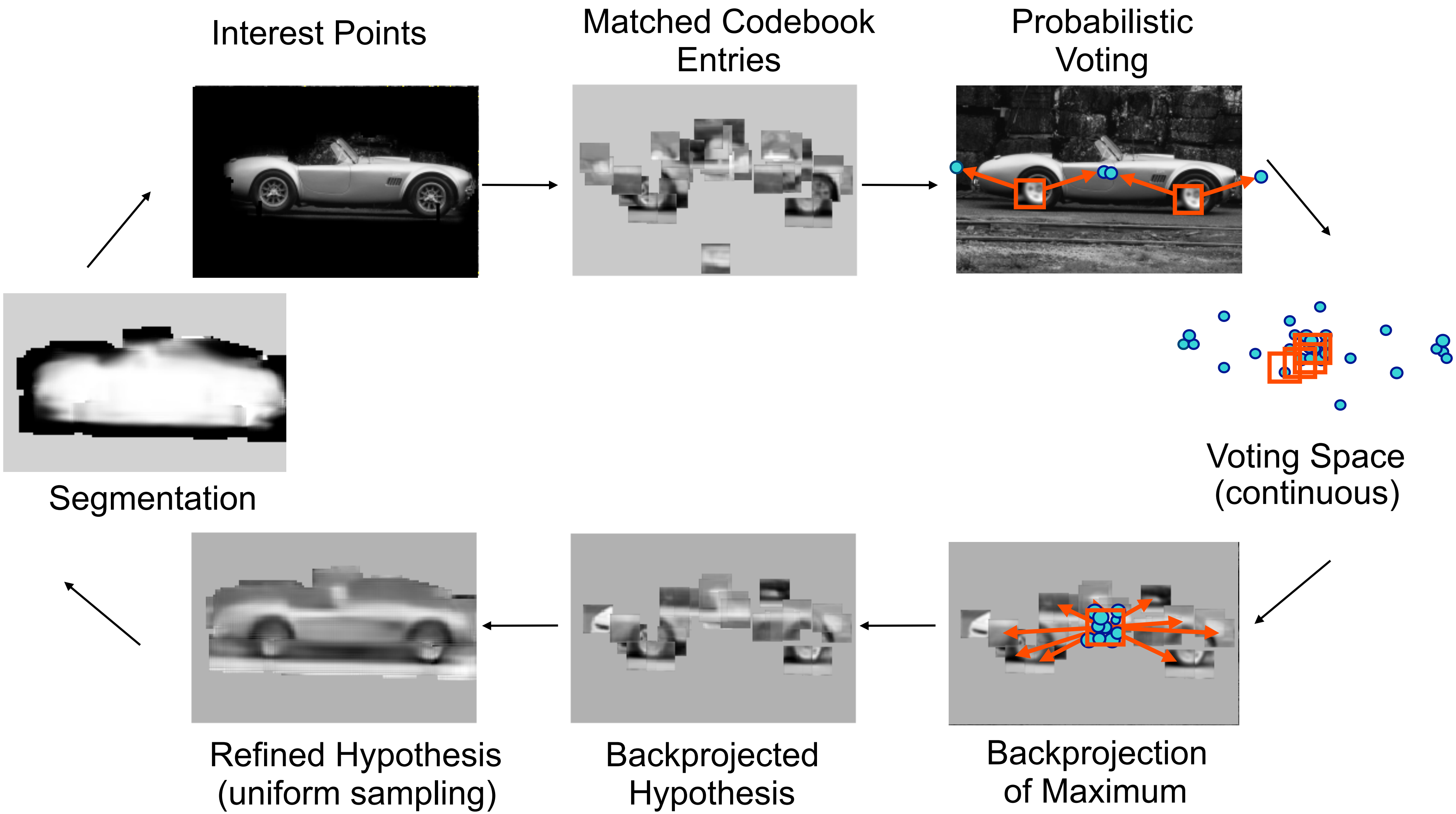
Discussion (2)

- Disadvantages
  - Each feature acts independently
    ⇒ Assumption violated if sampled patches overlap
  - Only loose constraints on object shape
  - False positives on structured regions of the background

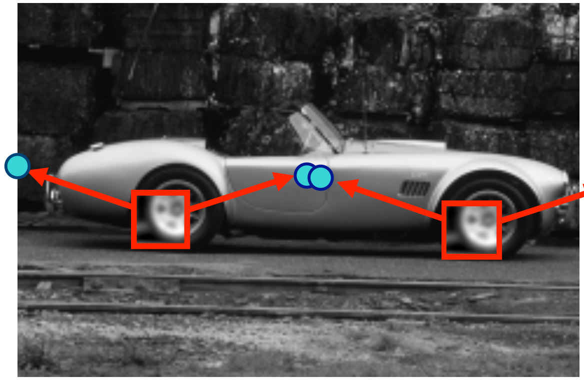
⇒ Hypothesis verification needed

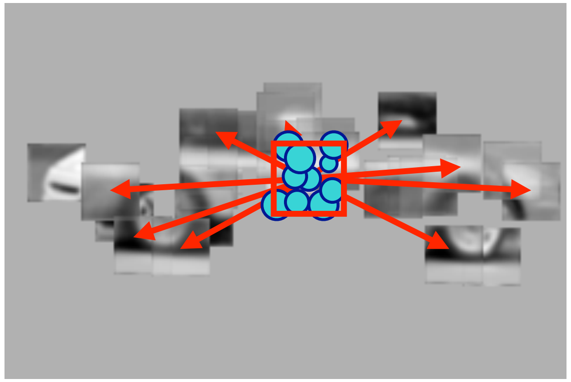
- Idea: Combination with top-down segmentation
  - Initial hypothesis generation
  - Category-specific figure-ground segmentation
  - Hypothesis verification using segmentation

"Closing the Loop"



Segmentation: Probabilistic Formulation





- Influence of patch e on object hypothesis

$$\boxed{p(\mathbf{e} \mid o_n, x)} = \frac{p(o_n, x \mid \mathbf{e})\, p(\mathbf{e})}{p(o_n, x)} = \frac{\sum_I p(o_n, x \mid I)\, p(I \mid \mathbf{e})\, p(\mathbf{e})}{p(o_n, x)}$$

- Backprojection to patches e and pixels p:

$$p(\mathbf{p} = \mathit{figure} \mid o_n, x) = \sum_{\mathbf{p} \in \mathbf{e}} p(\mathbf{p} = \mathit{figure} \mid \mathbf{e}, o_n, x)\, \boxed{p(\mathbf{e} \mid o_n, x)}$$

Leibe, Schiele, '03

Segmentation: Probabilistic Formulation

- Resolve patches by interpretations (codebook entries) I

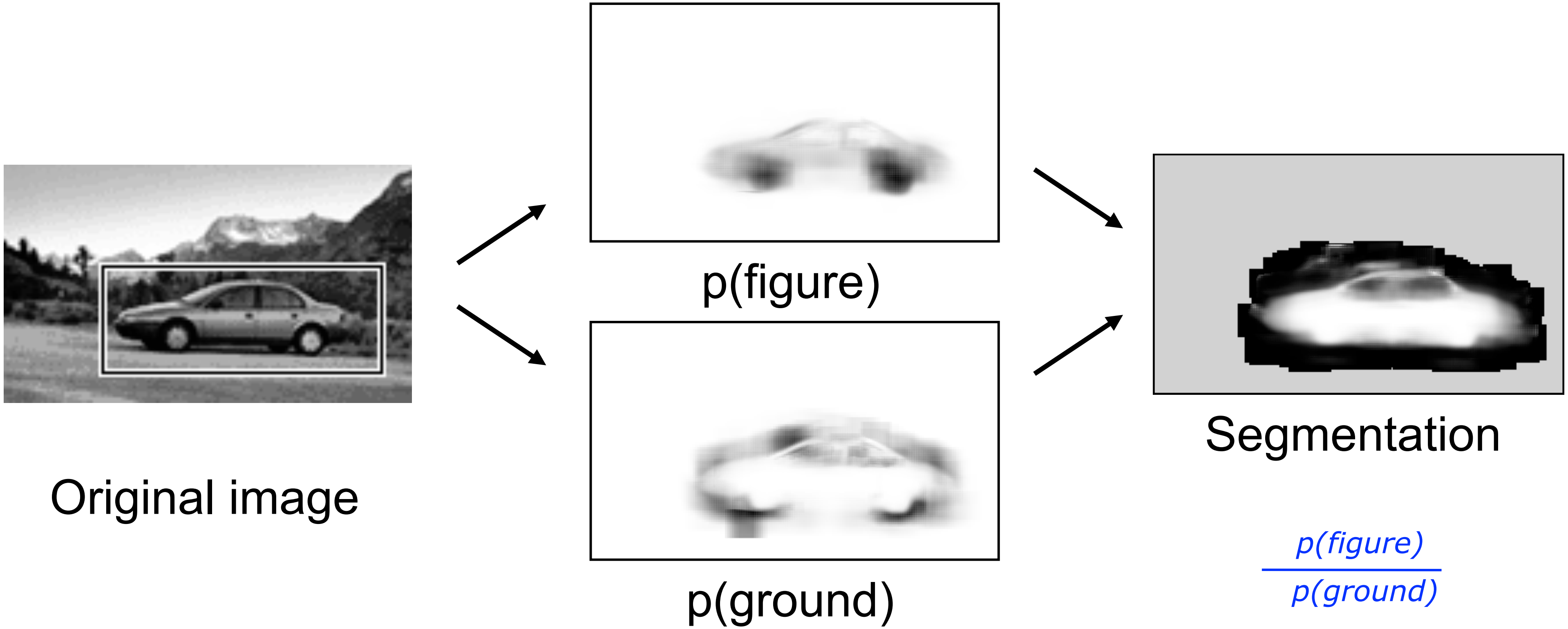
$$p(\mathbf{p} = figure \mid o_n, x) = \sum_{\mathbf{p} \in \mathbf{e}} \sum_I p(\mathbf{p} = figure \mid \mathbf{e}, I, o_n, x) \boxed{p(\mathbf{e}, I \mid o_n, x)}$$

$$= \sum_{\mathbf{p} \in \mathbf{e}} \sum_I \underbrace{p(\mathbf{p} = figure \mid I, o_n, x)}_{\text{Segmentation information}} \underbrace{\boxed{\frac{p(o_n, x \mid I) p(I \mid \mathbf{e}) p(\mathbf{e})}{p(o_n, x)}}}_{\text{Influence on object hypothesis}}$$

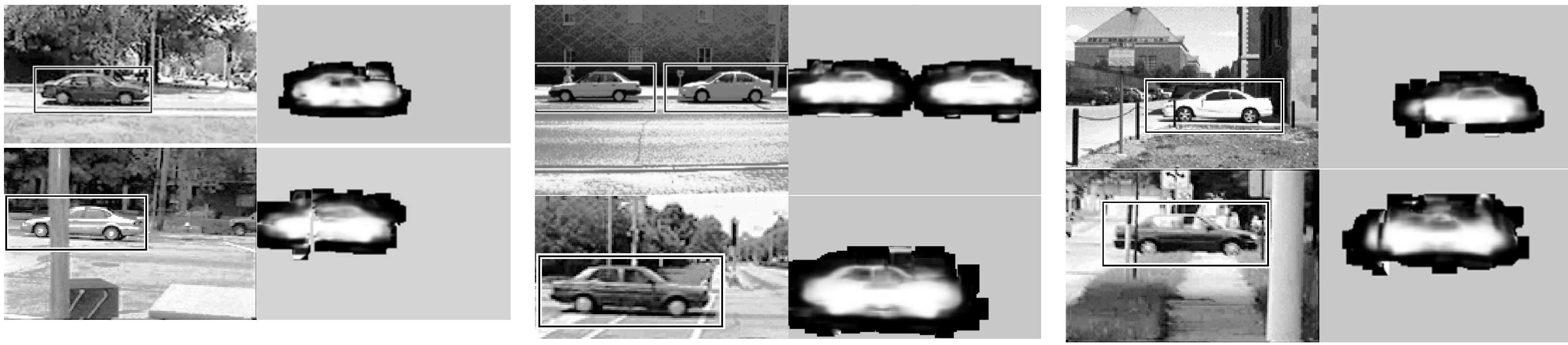
⇒ Store patch segmentation mask for every occurrence position!

Leibe, Schiele, '03

Segmentation

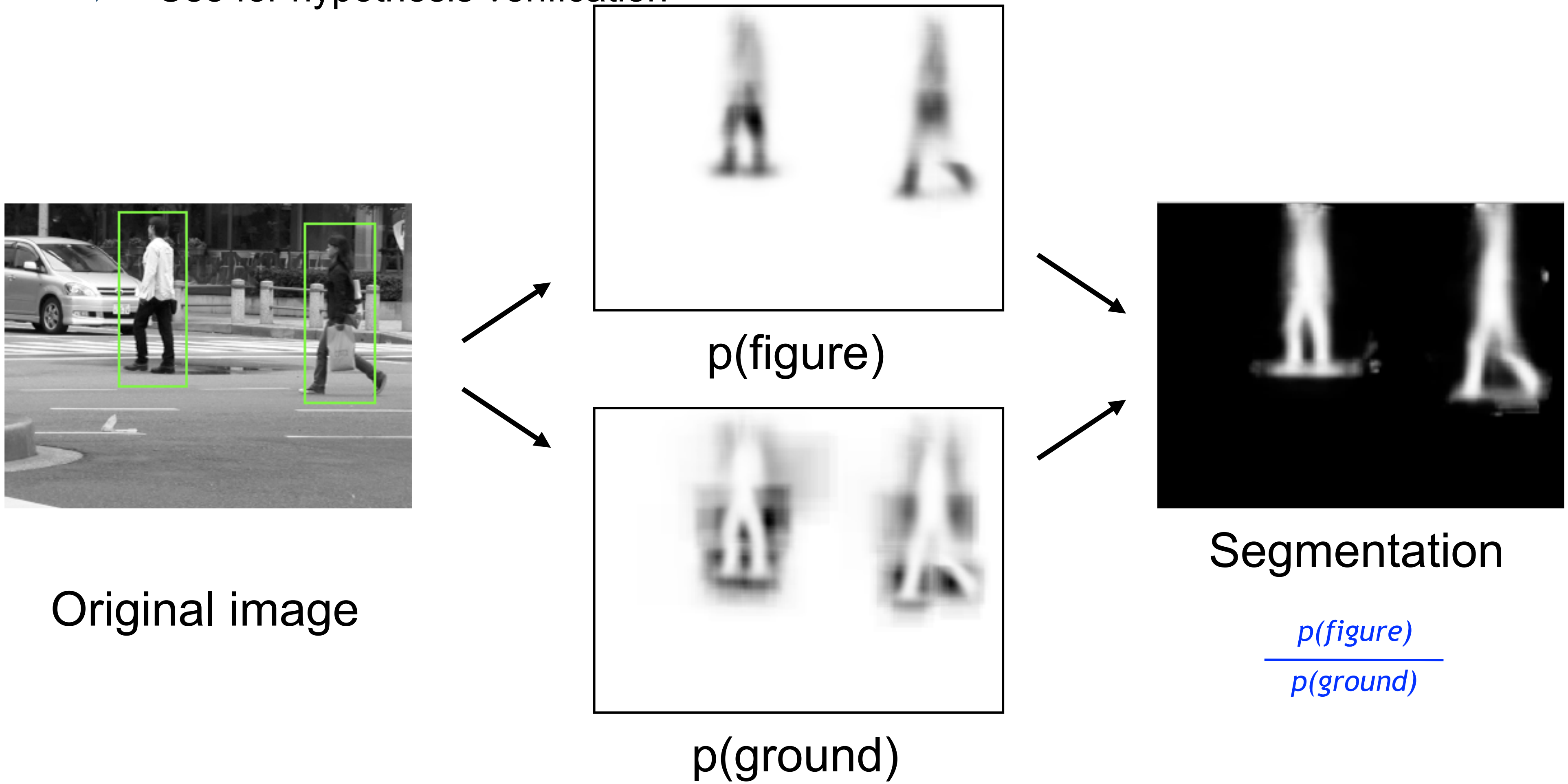


$$\frac{p(figure)}{p(ground)}$$



Segmentation

- Interpretation of p(figure) map
  - per-pixel confidence in object hypothesis
  - Use for hypothesis verification



Top-Down Driven Segmentation

- Example 1:

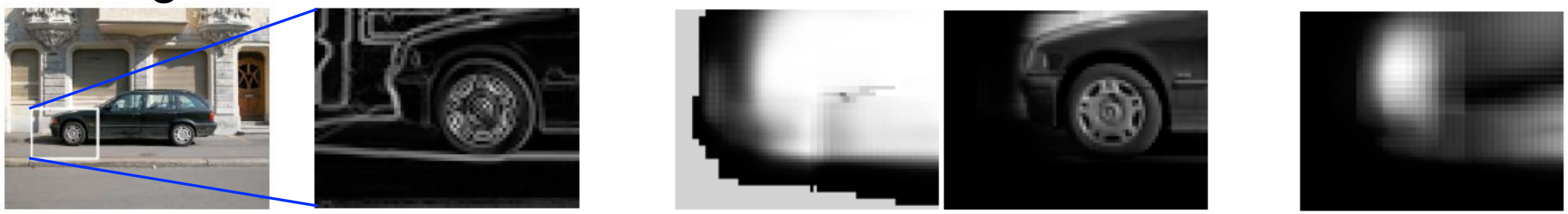
Leibe, Schiele, '03

image hypothesis segmentation p(figure)

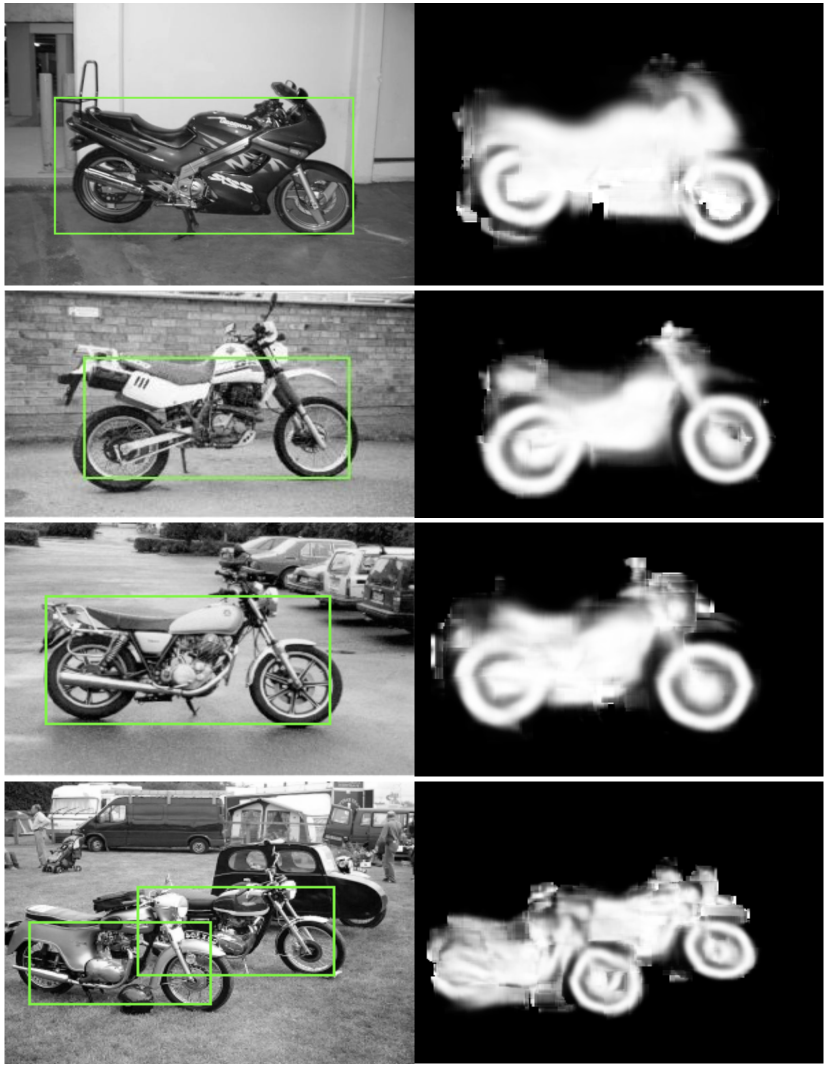
  - Pedestrian is segmented out since it does not contribute to the car hypothesis

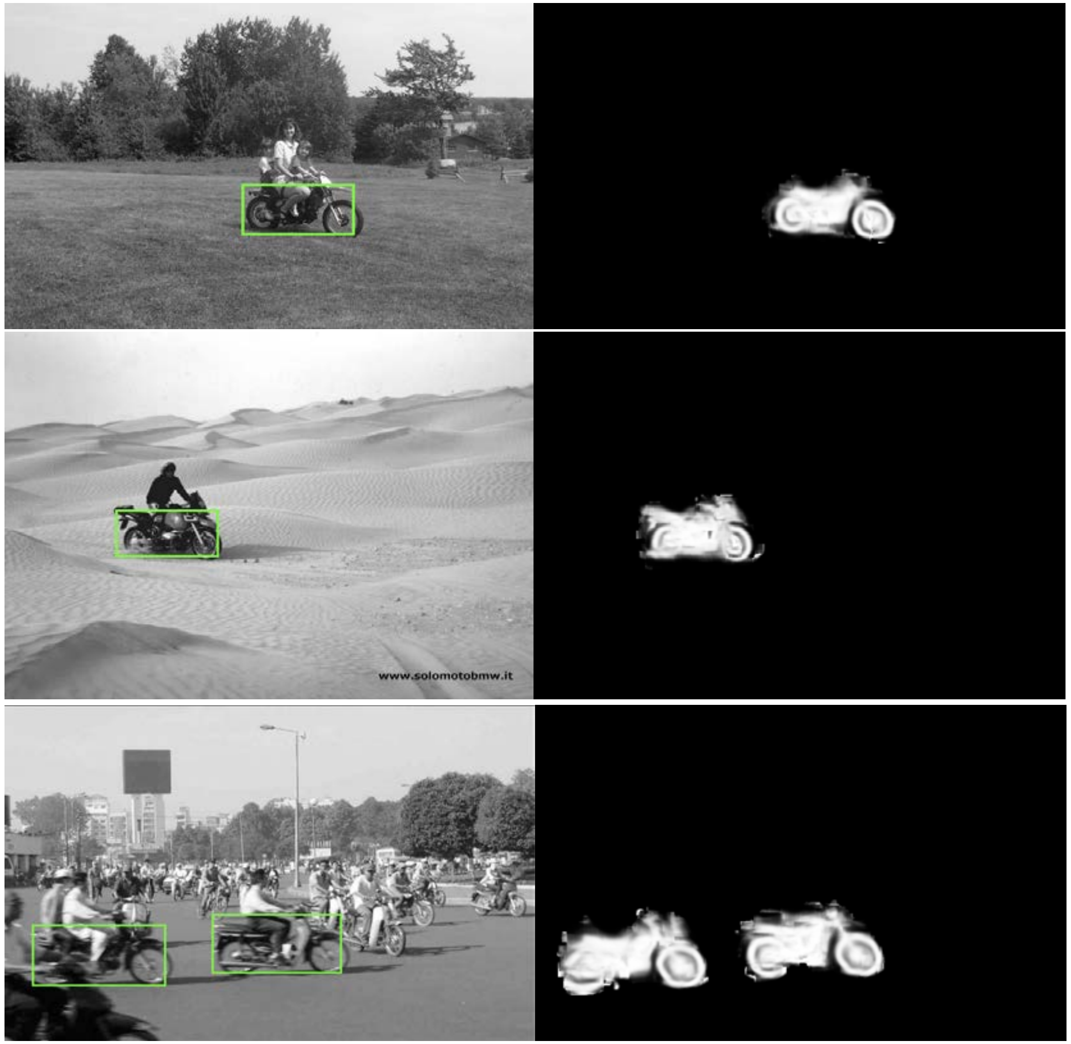
- Example 2:

image sub-image contours segmentation p(figure)



Motorbikes: Segmentation Results





Leibe, Schiele, '04

Hypothesis Verification: Motivation

- Secondary hypotheses
  - Desired property of algorithm! $\Rightarrow$ robustness to partial occlusion
  - Standard solution: reject based on bounding box overlap

$\Rightarrow$ Problematic - may lead to missing detections!
$\Rightarrow$ Use segmentations to resolve ambiguities instead

Leibe, Leonardis, Schiele, ‘04

Formalization in MDL Framework

- Savings of a hypothesis [Leonardis, IJCV'95]

$$S_h = K_0 S_{area} - K_1 S_{model} - K_2 S_{error}$$

- with
  - S_{area} : #pixels N in segmentation
  - S_{model} : model cost, assumed constant
  - S_{error} : estimate of error, according to

$$S_{error} = \sum_{\mathbf{p} \in Seg(h)} (1 - p(\mathbf{p} = figure|h))$$

- Final form of equation

$$S_h = -\frac{K_1}{K_0} + \left(1 - \frac{K_2}{K_0}\right) N + \frac{K_2}{K_0} \sum_{\mathbf{p} \in Seg(h)} p(\mathbf{p} = figure|h)$$

Formalization in MDL Framework (2)

- Savings of combined hypothesis

$$S_{h_1 \cup h_2} = S_{h_1} + S_{h_2} - S_{area}(h_1 \cap h_2) + S_{error}(h_1 \cap h_2)$$

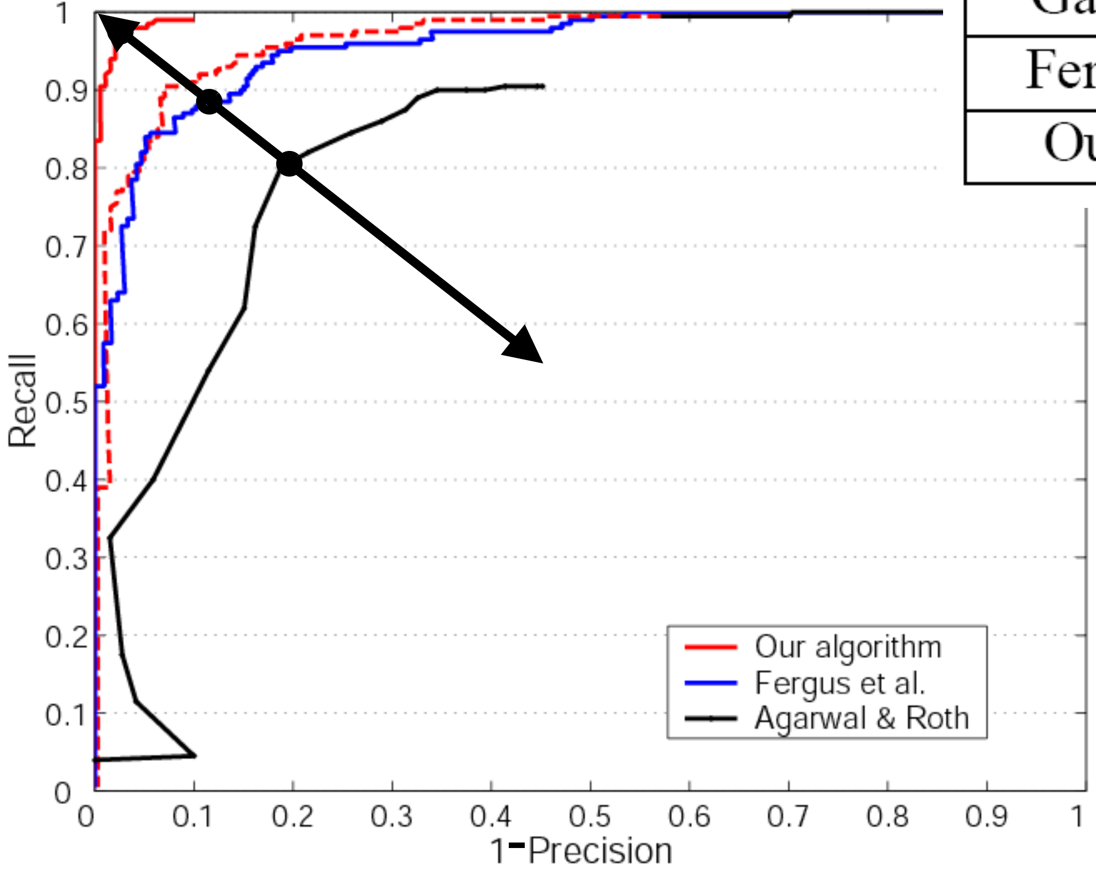
- Goal: Find combination (vector m) that best explains the image
  - Quadratic Boolean Optimization problem [Leonardis et al, 95]

$$S(\hat{m}) = \max_m m^T Q m = \max_m m^T \begin{bmatrix} S_{h_1} & \cdots & \frac{1}{2} S_{h_1 \cap h_N} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} S_{h_1 \cap h_2} & \cdots & S_{h_N} \end{bmatrix} m$$

  - In practice often sufficient to compute greedy approximation

Performance after Verification Stage

- Direct Comparison



Method	Equal Error Rate
Agarwal & Roth	~79%
Garg et al.	~88%
Fergus et al.	88.5%
Our algorithm	97.5%

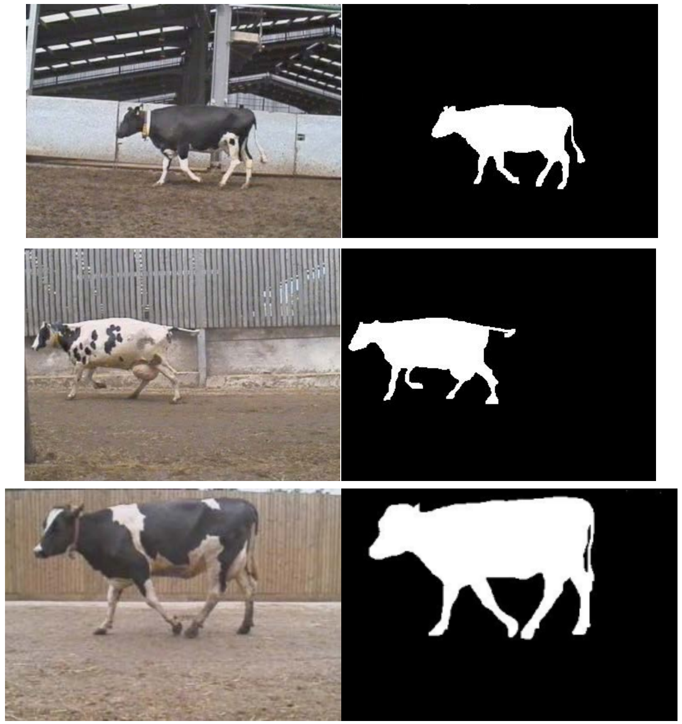
195/200 correct detections
5 false positives

Other Categories: Cows

- Articulated Object Recognition
  - Use set of cow sequences (from Derek Magee@Leeds)

- Extract frames from subset of sequences

Train on 113 images (+ segmentation)



Leibe, Leonardis, Schiele, '04

Cows: Results on Novel Sequences

- Object Detections
  - Single-frame recognition - No temporal continuity used!

Leibe, Leonardis, Schiele, '04

Cows: Results on Novel Sequences (2)

Leibe, Leonardis, Schiele, '04

- Segmentations from interest points
  - Single-frame recognition - No temporal continuity used!

Cows: Results on Novel Sequences (3)

Leibe, Leonardis, Schiele, '04

- Segmentations from refined hypotheses
  - Single-frame recognition - No temporal continuity used!

Another Example

- Object Detections

Leibe, Leonardis, Schiele, '04

Another Example (2)

- Segmentations from interest points

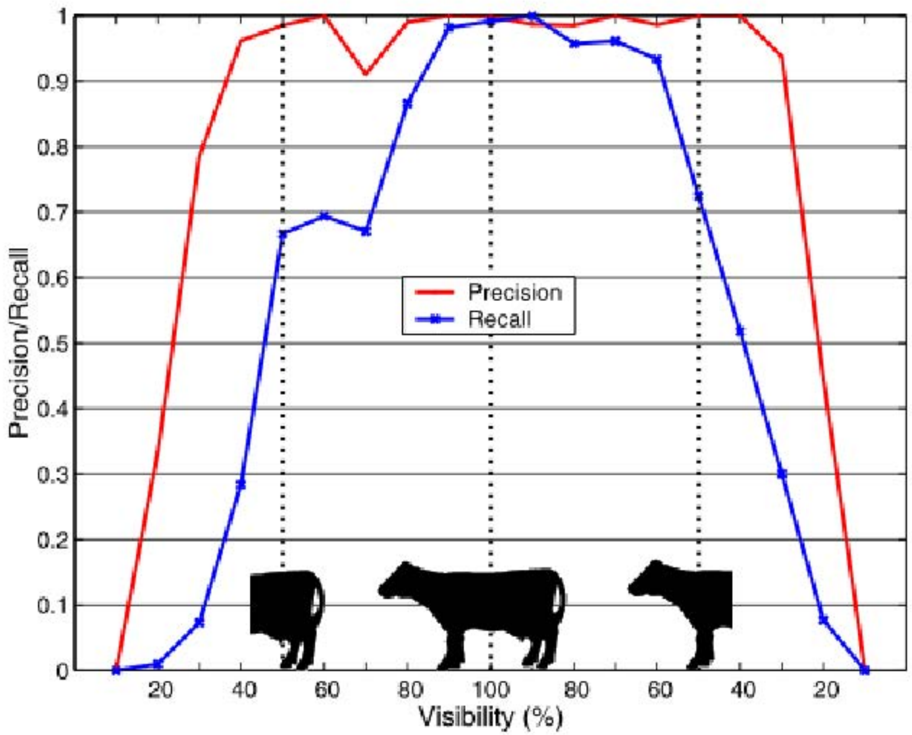
Leibe, Leonardis, Schiele, ‘04

Another Example (3)

- Segmentations from refined hypotheses

Leibe, Leonardis, Schiele, '04

Robustness to Occlusion



- Quantitative results (14 sequences, 2217 frames total)
  - No difficulties recognizing fully visible cows (99.1% recall)
  - Robust to significant partial occlusion!
  - Some detections even with 20-30% visibility

Example Detections