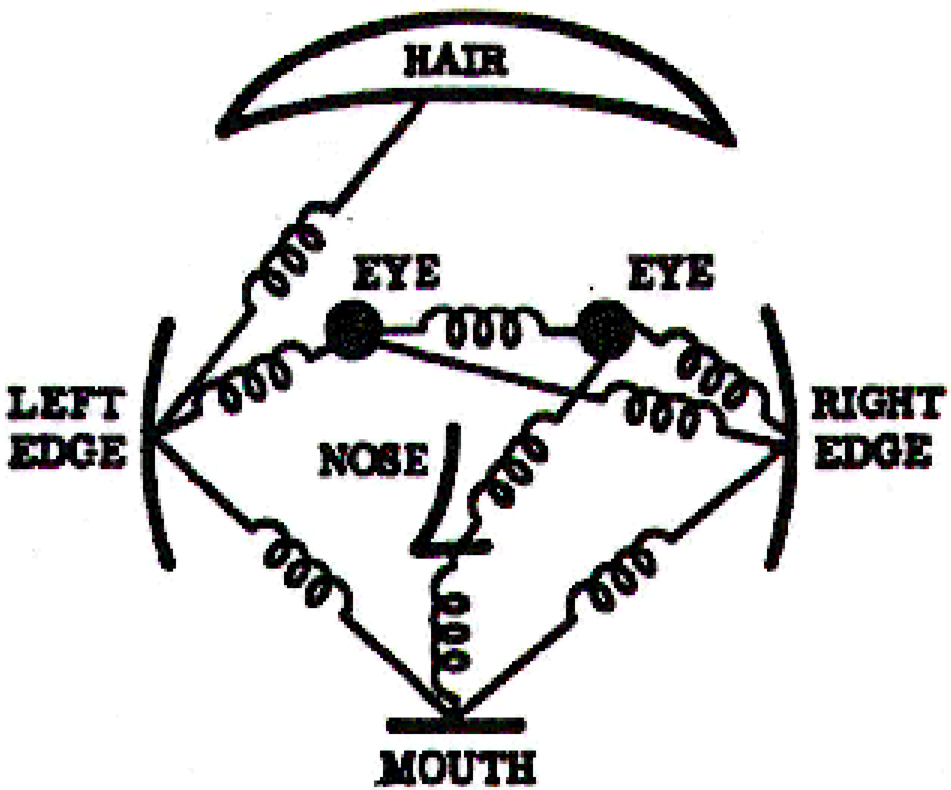
Beyond bags of features: Adding spatial information



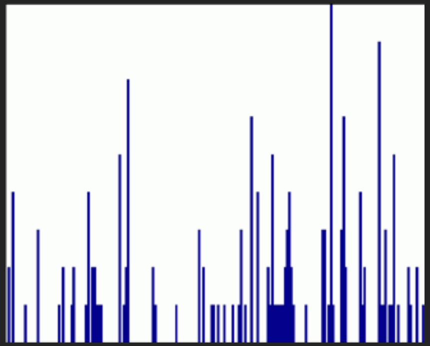
Slides by Lana Lazebnik, some adapted from Fei-Fei Li, Rob Fergus, and Antonio Torralba

Adding spatial information

- Forming vocabularies from pairs of nearby features – “doublets” or “bigrams”
- Computing bags of features on sub-windows of the whole image
- Using codebooks to vote for object position
- Generative part-based models

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution

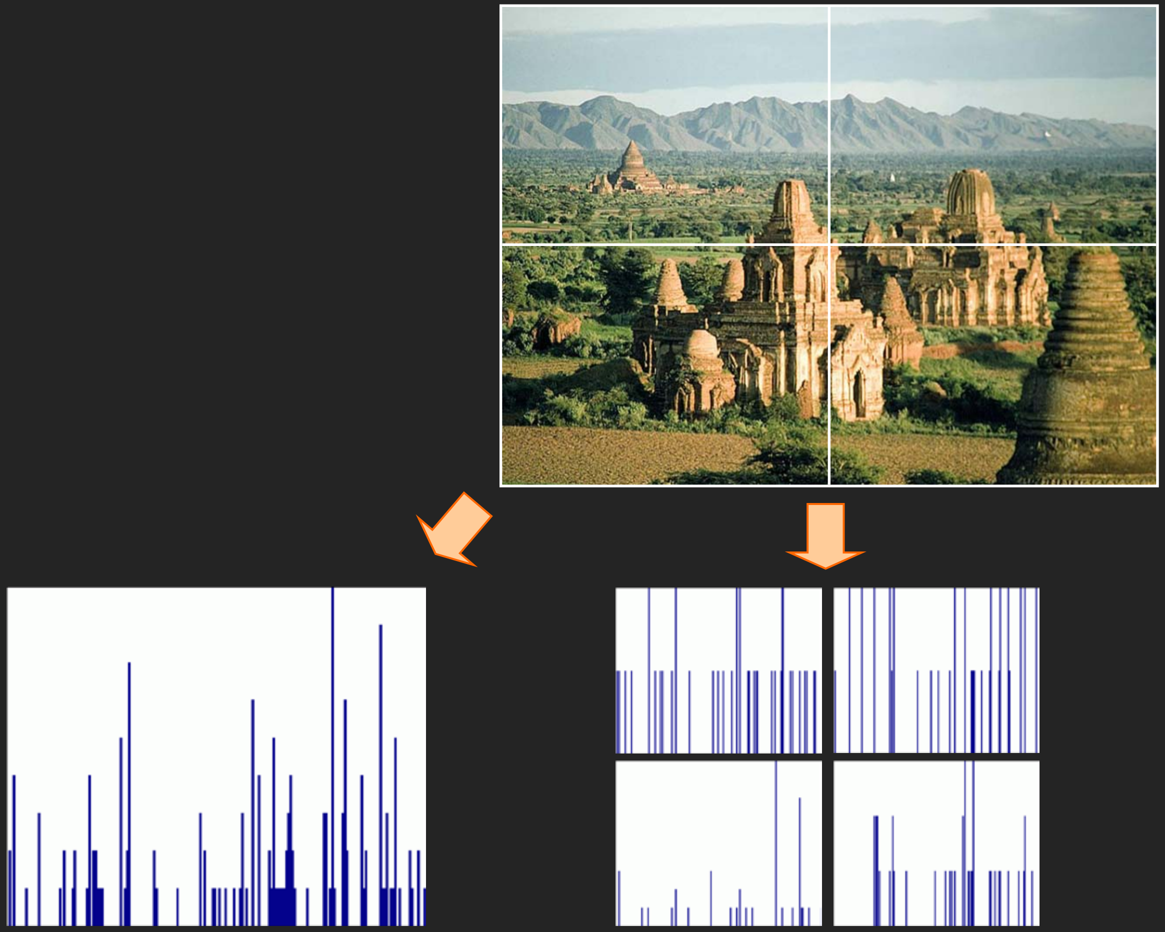


level 0

Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



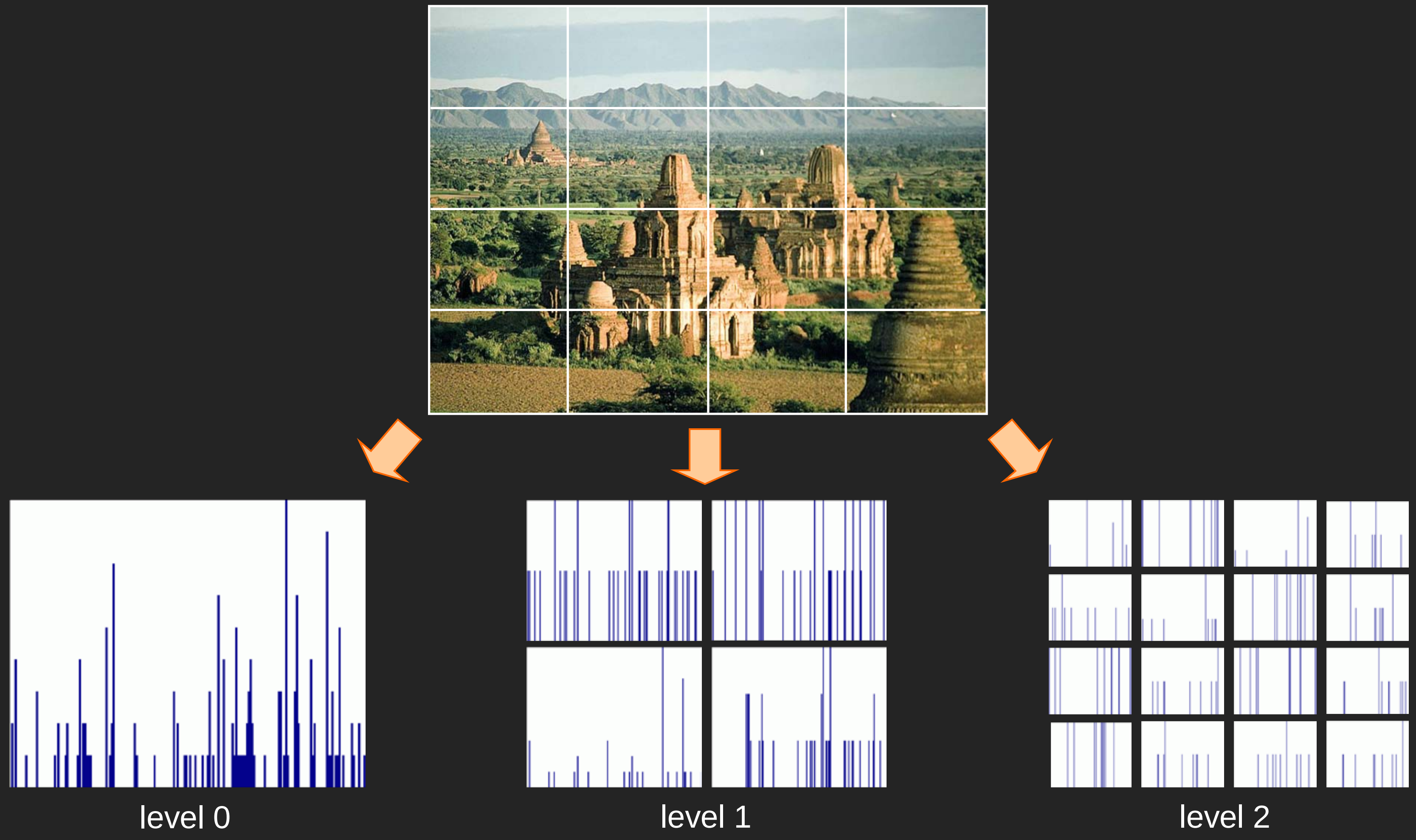
level 0

level 1

Lazebnik, Schmid & Ponce (CVPR 2006)

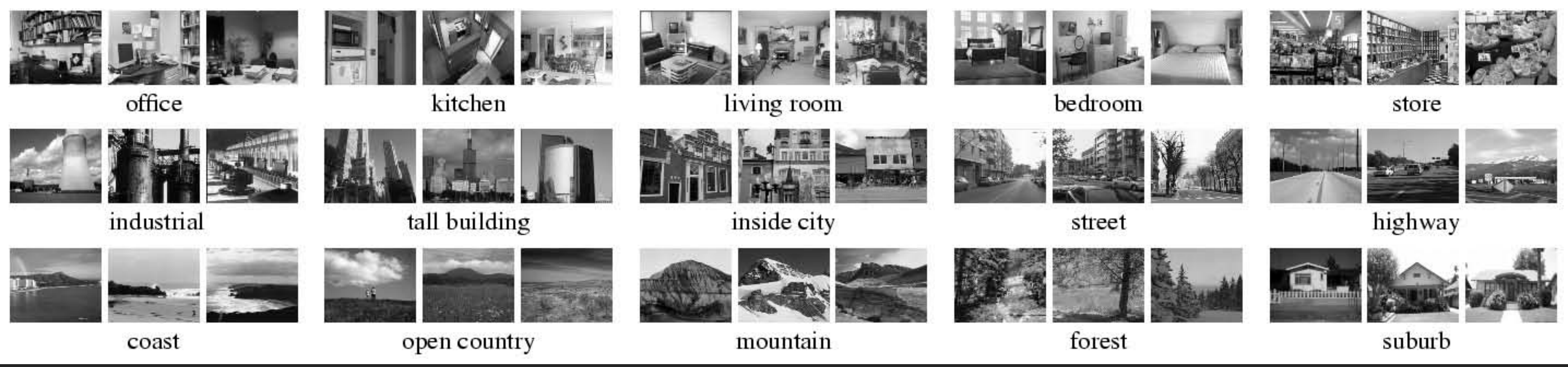
Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution



Lazebnik, Schmid & Ponce (CVPR 2006)

Scene category dataset

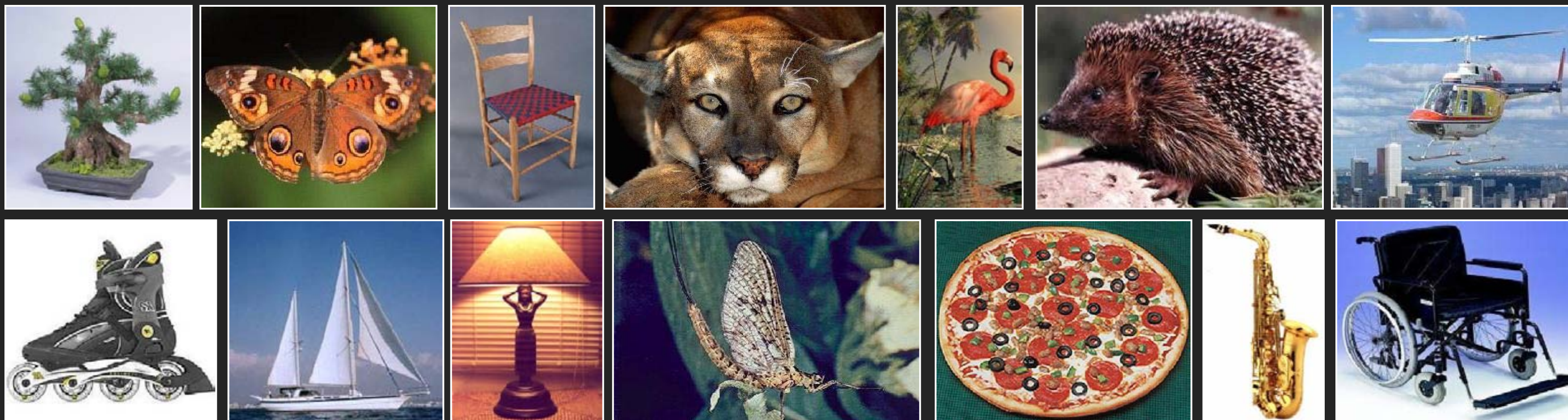


Multi-class classification results
(100 training images per class)

	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

Caltech101 dataset

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

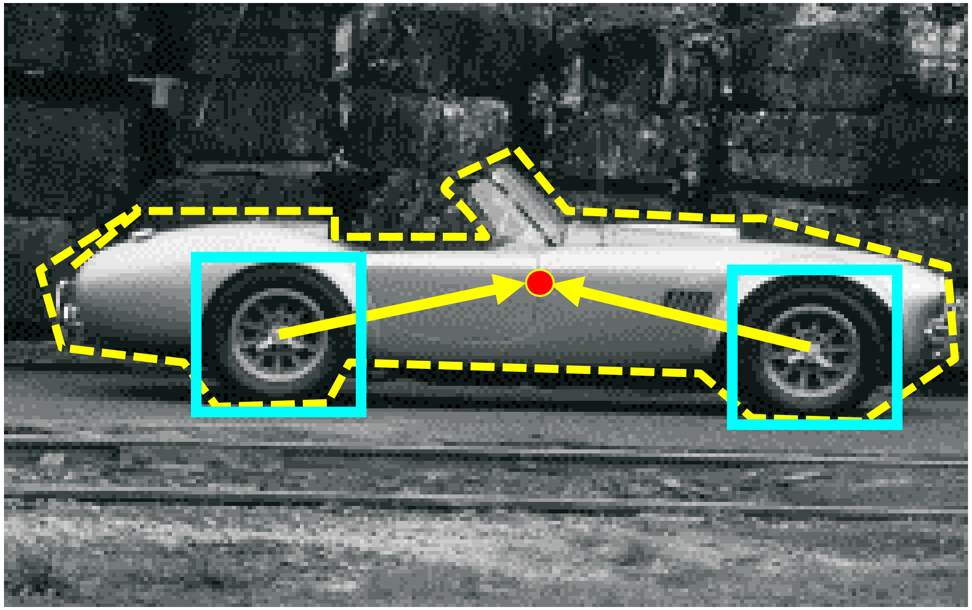


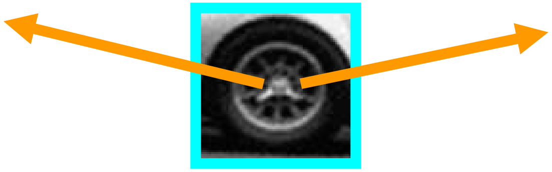
Multi-class classification results (30 training images per class)

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

Implicit shape models

- Visual codebook is used to index votes for object position





visual codeword with displacement vectors

training image annotated with object localization info

B. Leibe, A. Leonardis, and B. Schiele, [Combined Object Categorization and Segmentation with an Implicit Shape Model](), ECCV Workshop on Statistical Learning in Computer Vision 2004

Implicit shape models

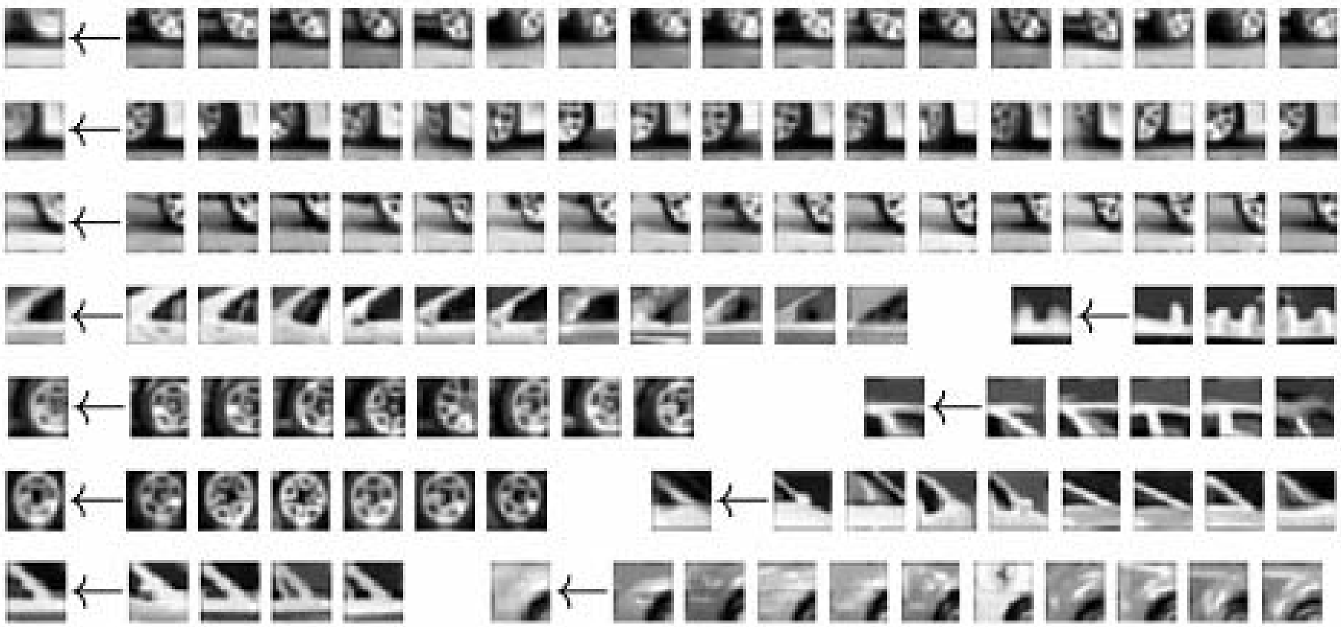
- Visual codebook is used to index votes for object position

test image

B. Leibe, A. Leonardis, and B. Schiele, [Combined Object Categorization and Segmentation with an Implicit Shape Model](), ECCV Workshop on Statistical Learning in Computer Vision 2004

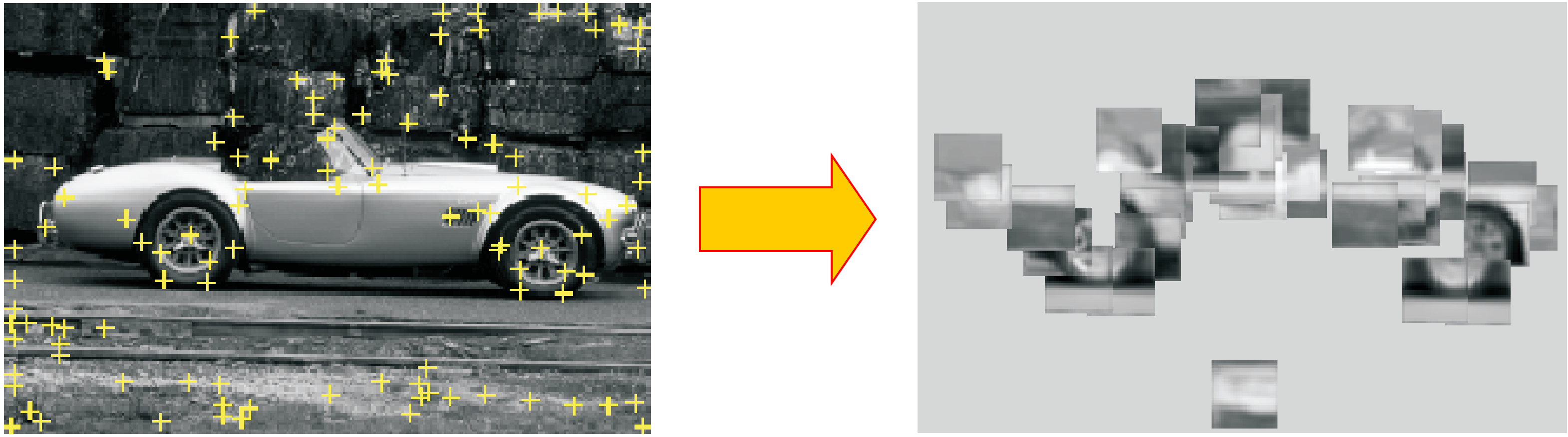
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering



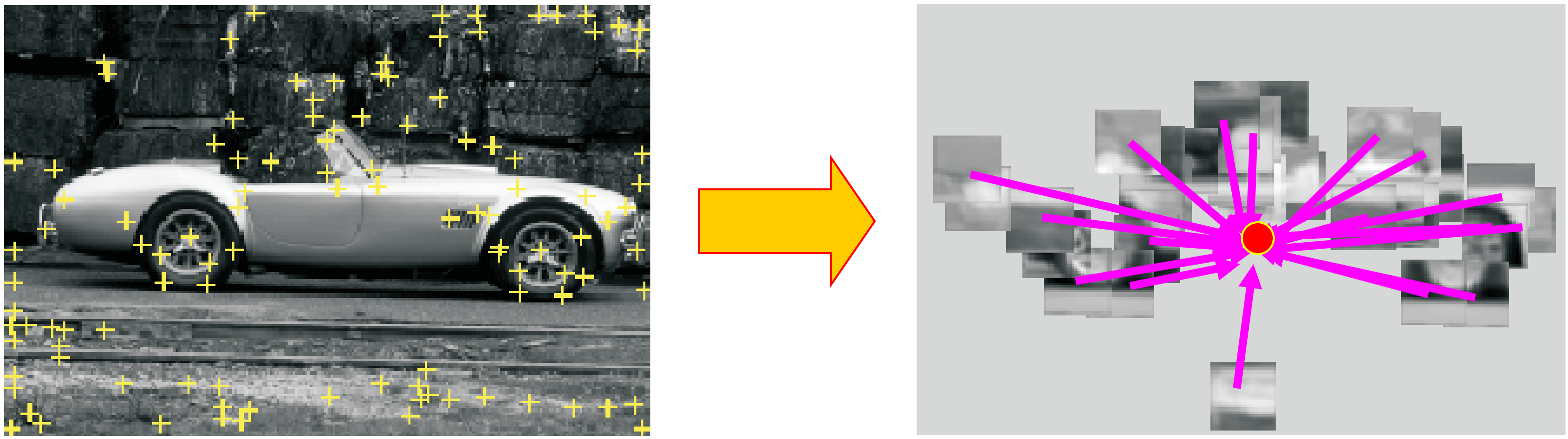
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest codebook entry



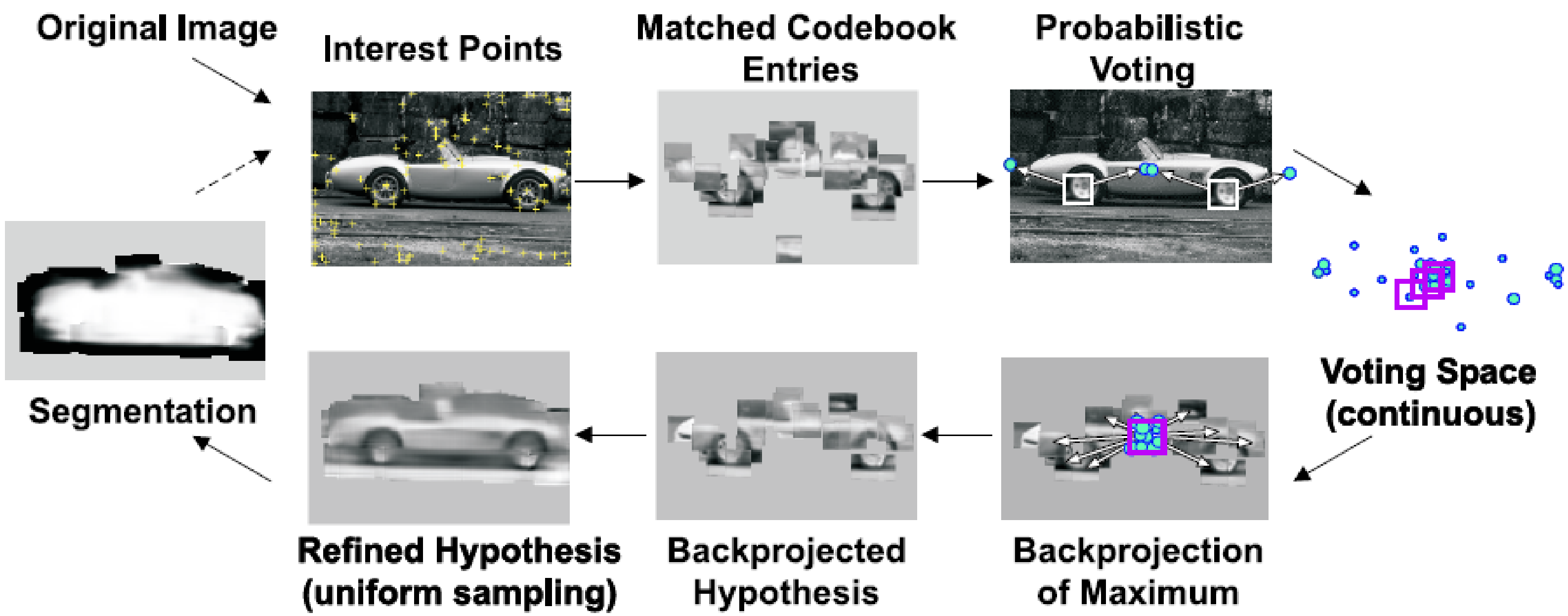
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest codebook entry
3. For each codebook entry, store all positions it was found, relative to object center



Implicit shape models: Testing

1. Given test image, extract patches, match to codebook entry
2. Cast votes for possible positions of object center
3. Search for maxima in voting space
4. Extract weighted segmentation mask based on stored masks for the codebook occurrences



Example: Results on Cows

Original image

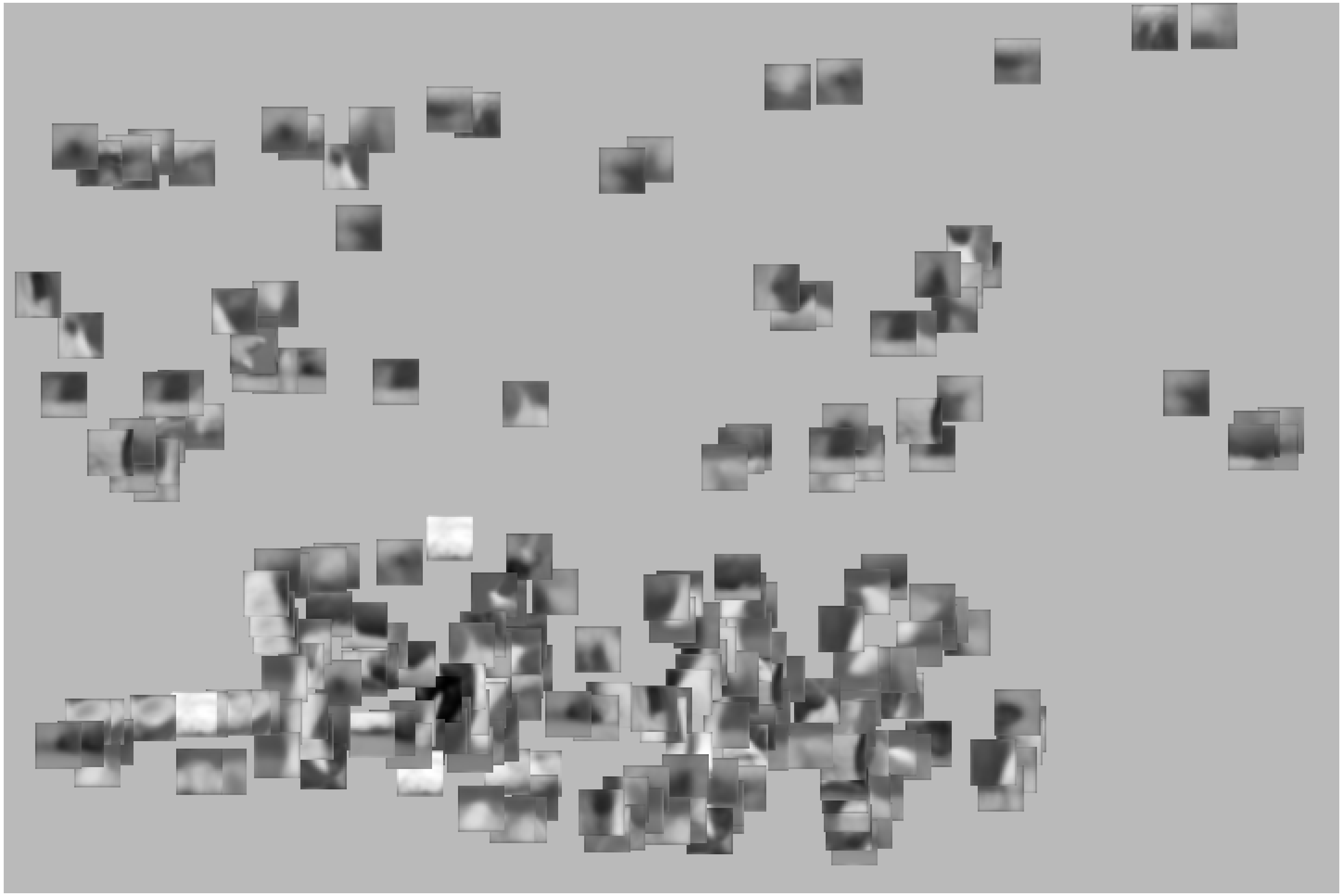
Source: B. Leibe

Example: Results on Cows

Interest points

Source: B. Leibe

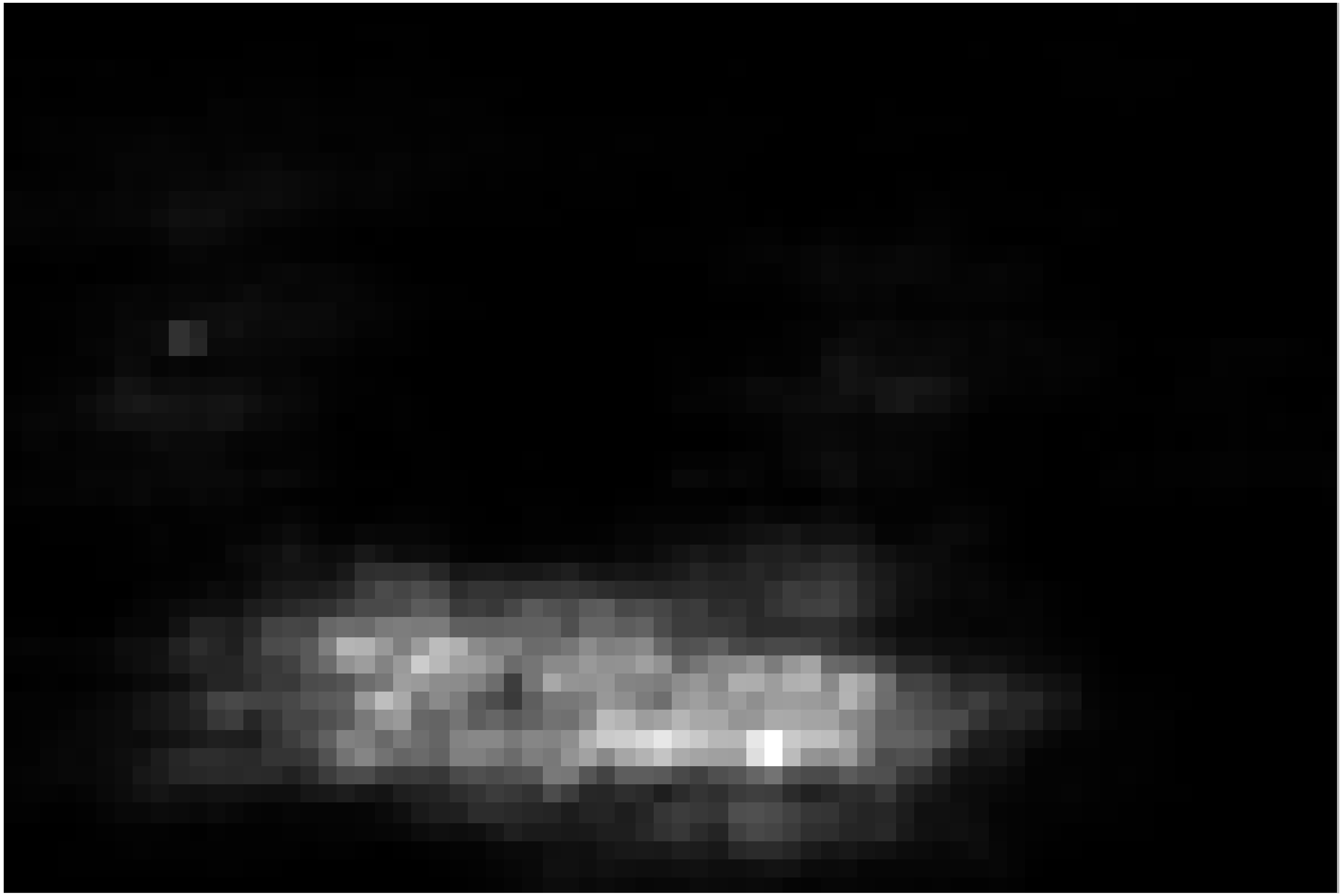
Example: Results on Cows



Matched patches

Source: B. Leibe

Example: Results on Cows



Probabilistic votes

Source: B. Leibe

Example: Results on Cows

Hypothesis 1

Source: B. Leibe

Example: Results on Cows

Hypothesis 2

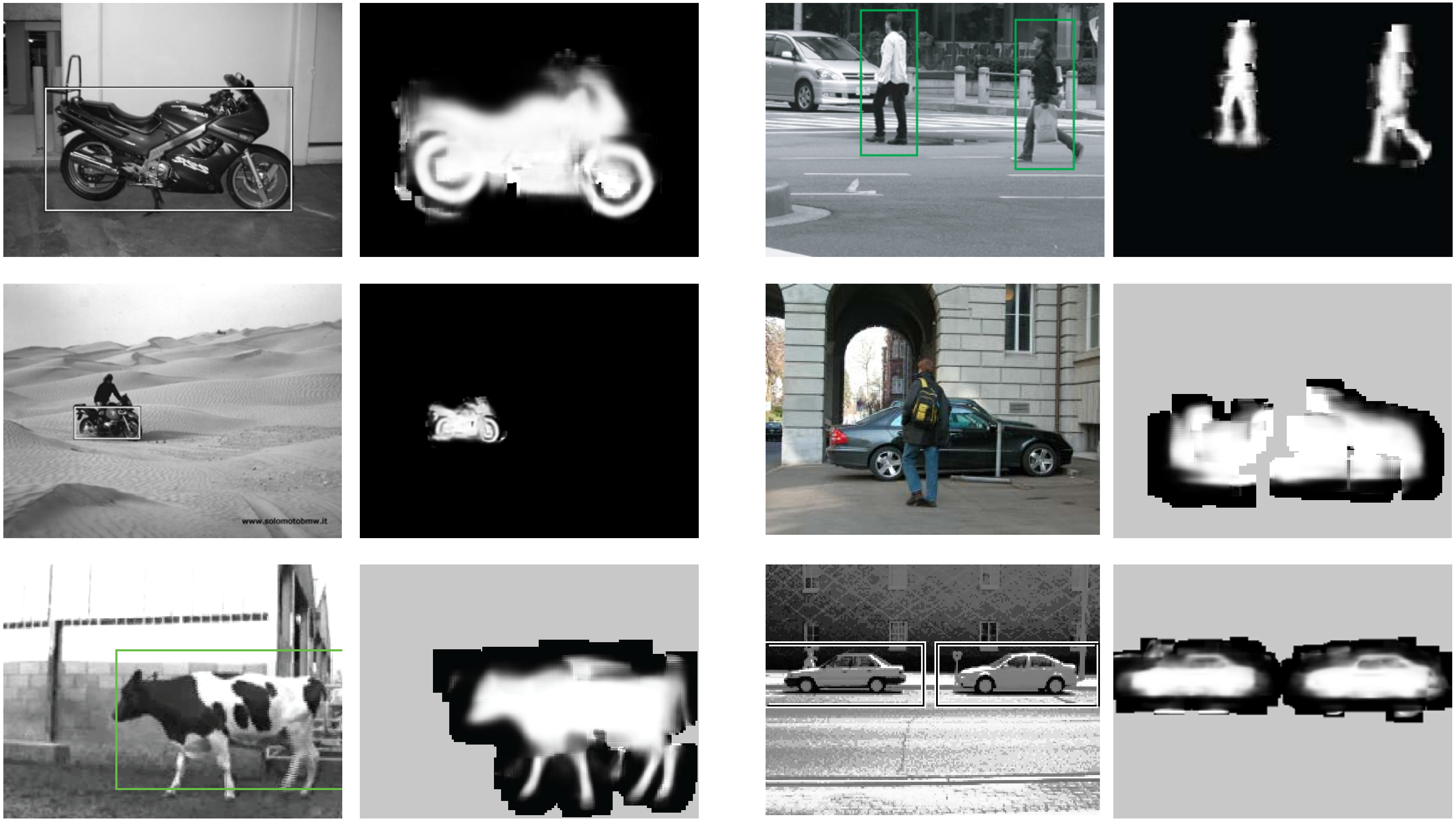
Source: B. Leibe

Example: Results on Cows

Hypothesis 3

Source: B. Leibe

Additional examples



B. Leibe, A. Leonardis, and B. Schiele, [Robust Object Detection with Interleaved Categorization and Segmentation](), IJCV 77 (1-3), pp. 259-289, 2008.

Generative part-based models

R. Fergus, P. Perona and A. Zisserman, **[Object Class Recognition by Unsupervised Scale-Invariant Learning](#)**, CVPR 2003

Probabilistic model

$$P(image \mid object) = P(appearance, shape \mid object)$$

Part descriptors

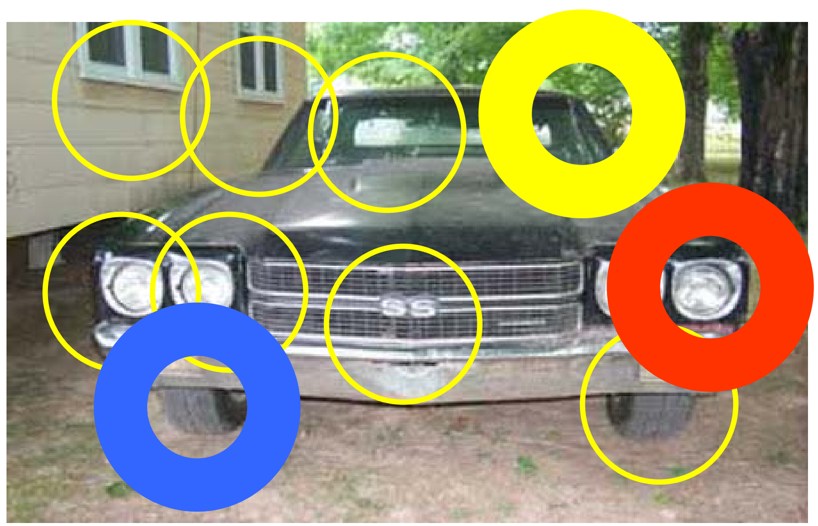
Part locations

Candidate parts

Probabilistic model

$$P(image \mid object) = P(appearance, shape \mid object)$$

Part 1



Part 3

Part 2

Probabilistic model

$$P(image \mid object) = P(appearance, shape \mid object)$$

$$= \max_h P(appearance \mid h, object)\, p(shape \mid h, object)\, p(h \mid object)$$

h: assignment of features to parts

Part 1

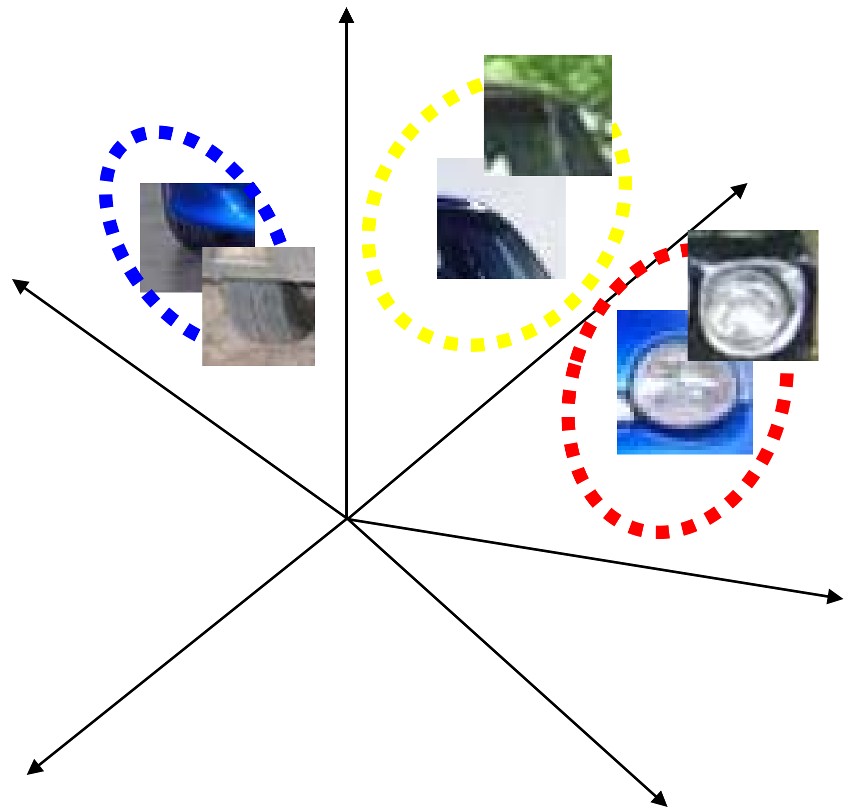
Part 3

Part 2

Probabilistic model

$$P(image \mid object) = P(appearance, shape \mid object)$$

$$= \max_h \boxed{P(appearance \mid h, object)}\, p(shape \mid h, object)\, p(h \mid object)$$



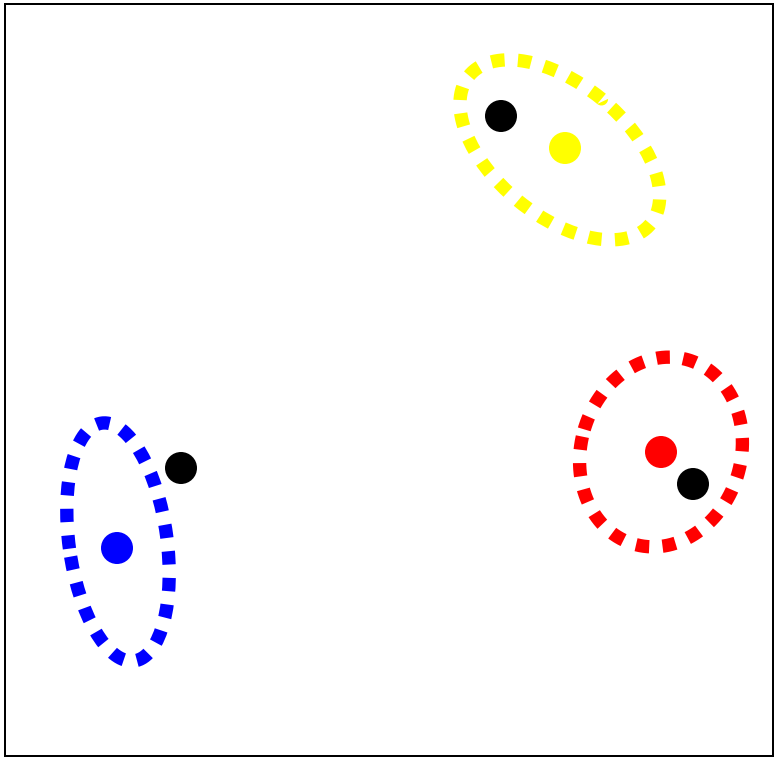
Distribution over patch descriptors

High-dimensional appearance space

Probabilistic model

$$P(image \mid object) = P(appearance, shape \mid object)$$

$$= \max_h P(appearance \mid h, object) \boxed{p(shape \mid h, object)} p(h \mid object)$$

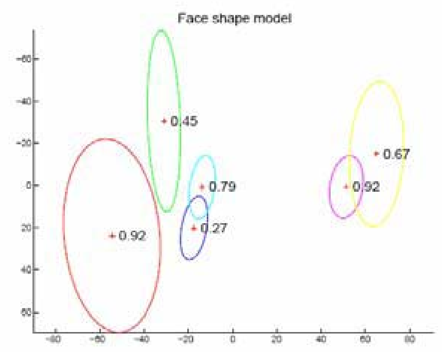


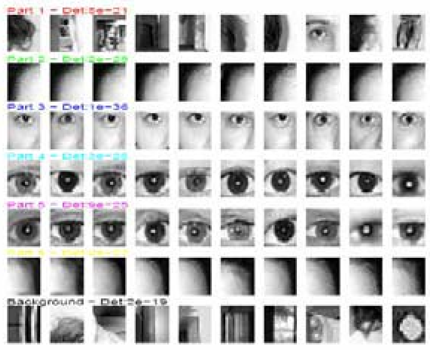
2D image space

Distribution over joint part positions

Results: Faces

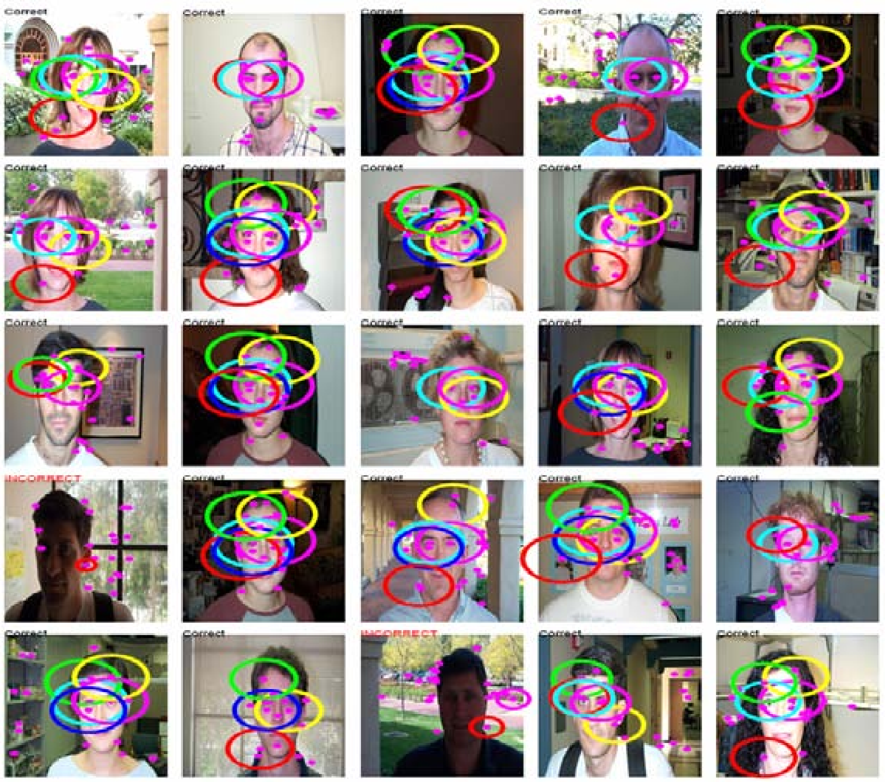
Face shape model



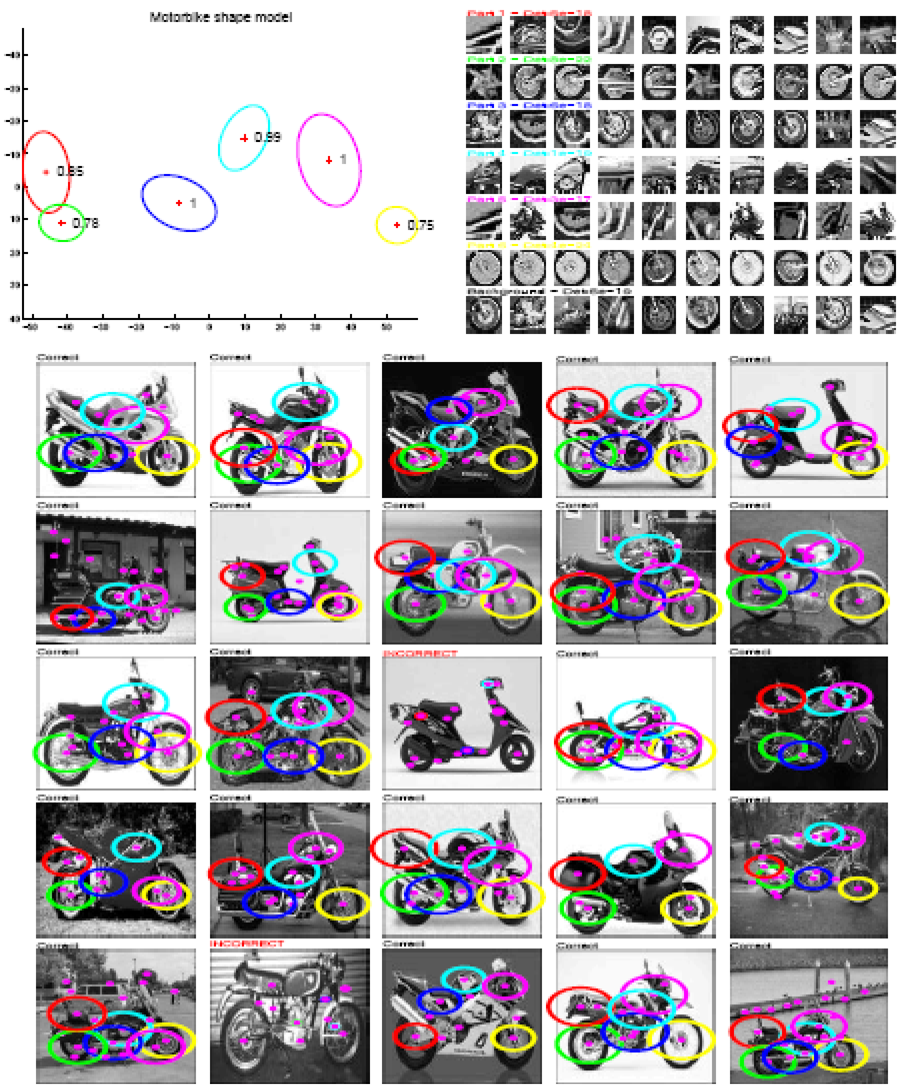


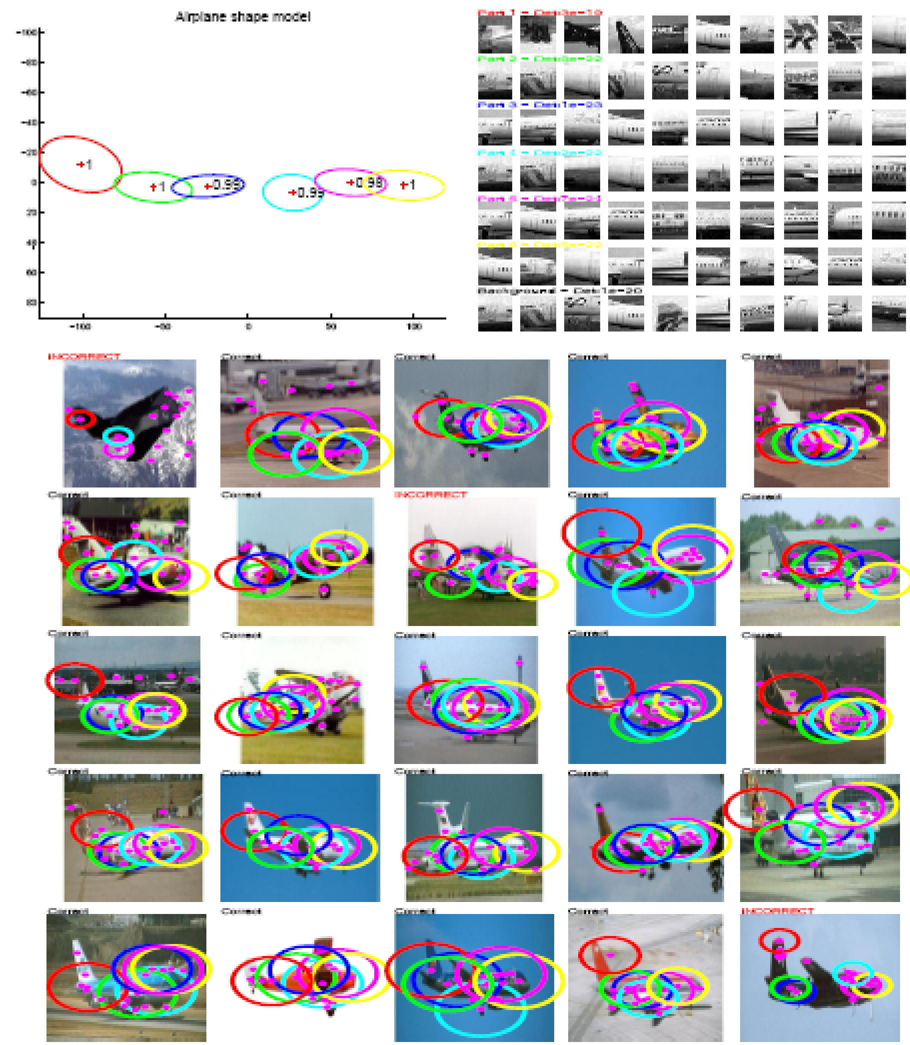
Patch appearance model

Recognition results

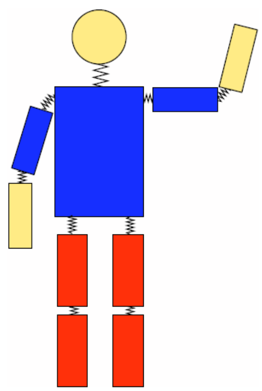


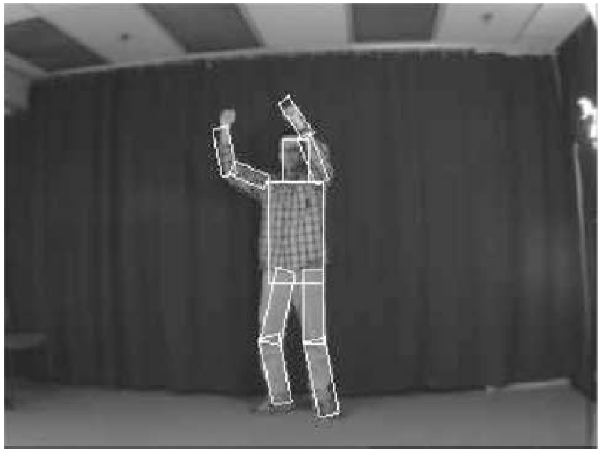
Results: Motorbikes and airplanes





Pictorial structures

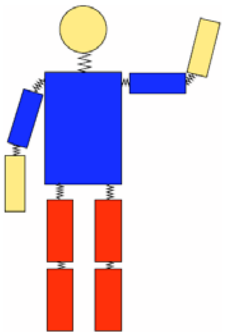


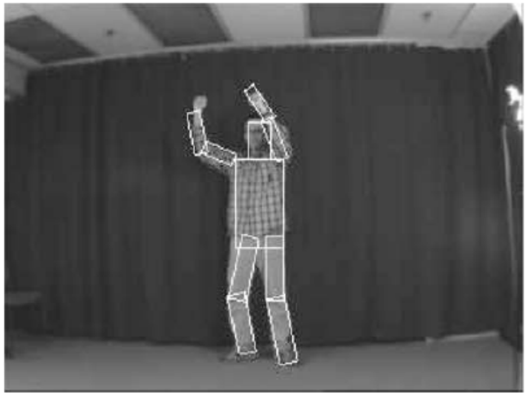


- Set of parts (oriented rectangles) connected by edges
- Recognition problem: find the most probable part layout $l_1, \ldots, l_n$ in the image

P. Felzenszwalb and D. Huttenlocher, **Pictorial Structures for Object Recognition**, IJCV 61(1), 2005

Pictorial structures





- MAP formulation: maximize posterior

$$P(l_1,\ldots,l_n \mid \mathrm{Im}) \propto P(\mathrm{Im} \mid l_1,\ldots,l_n) P(l_1,\ldots,l_n) = \underbrace{\prod_i P(\mathrm{Im}(l_i))}_{\text{Appearance}} \underbrace{\prod_{i,j \in E} P(l_i \mid l_j)}_{\text{Geometry}}$$

- Energy-based formulation: minimize minus the log of probability:

$$E(l_1,\ldots,l_n) = \underbrace{\sum_i m_i(l_i)}_{\text{Matching cost}} + \underbrace{\sum_{i,j} d_{ij}(l_i,l_j)}_{\text{Deformation cost}}$$

Summary: Adding spatial information

- Doublet vocabularies
  - Pro: takes co-occurrences into account, some geometric invariance is preserved
  - Con: too many doublet probabilities to estimate
- Spatial pyramids
  - Pro: simple extension of a bag of features, works very well
  - Con: no geometric invariance, no object localization
- Implicit shape models
  - Pro: can localize object, maintain translation and possibly scale invariance
  - Con: need supervised training data (known object positions and possibly segmentation masks)
- Generative part-based models
  - Pro: very nice conceptually, can be learned from unsegmented images
  - Con: combinatorial hypothesis search problem