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

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

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

[Diagram: An object is linked to a bag labeled 'Bag of 'words''.
Visual words distributions
Bag-of-Words Model: Overview

learning

- feature detection & representation
- image representation

recognition

- codewords dictionary

category models (and/or) classifiers

category decision
Bag-of-Words Model: Object Representation & Learning

1. feature detection & representation
2. codewords dictionary
3. image representation
Sampling Strategies

Sparse, at interest points

Dense, uniformly

Randomly

• 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):

![Image gradients](image1.png) → ![Keypoint descriptor](image2.png)
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

codewords dictionary

category models (and/or) classifiers

category decision
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
  - Natural scene categorization: Fei-Fei et al. 2005
Generative Models

- Basis is typically the Bayes’ Theorem:
  - $c$ = object class and $x$ = object representation
  
  Likelihood = probability of $x$ given $c$  
  A priori probability of class $c$

  \[
P(c|x) = \frac{P(x|c)P(c)}{P(x)}\]

  Posterior = probability of class $c$ given $x$  
  Probability of $x$

- Classification with

  \[
c^* = \arg \max_c P(c|x) = \arg \max_c \frac{P(x|c)P(c)}{P(x)} \\
  \propto \arg \max_c P(x|c)P(c)\]
Naïve Bayes Classifier

- Classify image using histograms of occurrences on visual words:

\[ x = \begin{array}{c}
\begin{array}{c}
\vdots
\end{array}
\end{array} \]

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

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

Based on lecture by Prof. T. Hofmann
Naive Bayes Classifier

- Multinomial model for each object class:

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

- Class priors: \( P(c) \), with \( \sum_{c} P(c) = 1 \)

- Posterior probabilities:

\[ P(c|x) = \frac{P(c) \prod_{t=1}^{n} P(x_t|c)}{\sum_{c'} P(c') \prod_{t=1}^{n} P(x_t|c')} \]
Naive Bayes Classifier: Decision

- Bayes optimal decision:

\[
c^* = \text{argmax}_c \ P(c \mid x)
\]

= \text{argmax}_c \left[ \log P(c) + \sum_{t=1}^{n} \log P(x_t \mid 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 \((k=1000)\).

<table>
<thead>
<tr>
<th>True classes</th>
<th>faces</th>
<th>buildings</th>
<th>trees</th>
<th>cars</th>
<th>phones</th>
<th>bikes</th>
<th>books</th>
</tr>
</thead>
<tbody>
<tr>
<td>faces</td>
<td>76</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>buildings</td>
<td>2</td>
<td>44</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>trees</td>
<td>3</td>
<td>2</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>cars</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>75</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>phones</td>
<td>9</td>
<td>15</td>
<td>1</td>
<td>16</td>
<td>70</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>bikes</td>
<td>2</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>8</td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>books</td>
<td>4</td>
<td>19</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>69</td>
</tr>
<tr>
<td>Mean ranks</td>
<td>1.49</td>
<td>1.88</td>
<td>1.33</td>
<td>1.33</td>
<td>1.63</td>
<td>1.57</td>
<td>1.57</td>
</tr>
</tbody>
</table>

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
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
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)
### 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)

<table>
<thead>
<tr>
<th>Level</th>
<th>Weak features (vocabulary size: 16)</th>
<th>Strong features (vocabulary size: 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-level</td>
<td>Pyramid</td>
</tr>
<tr>
<td>0 (1 × 1)</td>
<td>45.3 ±0.5</td>
<td></td>
</tr>
<tr>
<td>1 (2 × 2)</td>
<td>53.6 ±0.3</td>
<td>56.2 ±0.6</td>
</tr>
<tr>
<td>2 (4 × 4)</td>
<td>61.7 ±0.6</td>
<td>64.7 ±0.7</td>
</tr>
<tr>
<td>3 (8 × 8)</td>
<td>63.3 ±0.8</td>
<td><strong>66.8 ±0.6</strong></td>
</tr>
</tbody>
</table>

Fei-Fei & Perona: 65.2%
Scene category retrieval

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved images</th>
</tr>
</thead>
<tbody>
<tr>
<td>kitchen</td>
<td>living room</td>
</tr>
<tr>
<td>store</td>
<td></td>
</tr>
<tr>
<td>tall bldg</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>inside city</td>
<td></td>
</tr>
<tr>
<td>tall bldg</td>
<td></td>
</tr>
<tr>
<td>mountain</td>
<td></td>
</tr>
<tr>
<td>forest</td>
<td></td>
</tr>
</tbody>
</table>
## Caltech101 dataset

Fei-Fei et al. (2004)


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

<table>
<thead>
<tr>
<th>Level</th>
<th>Weak features (16)</th>
<th>Strong features (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-level</td>
<td>Pyramid</td>
</tr>
<tr>
<td>0</td>
<td>15.5 ±0.9</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>31.4 ±1.2</td>
<td>32.8 ±1.3</td>
</tr>
<tr>
<td>2</td>
<td>47.2 ±1.1</td>
<td>49.3 ±1.4</td>
</tr>
<tr>
<td>3</td>
<td>52.2 ±0.8</td>
<td><strong>54.0 ±1.1</strong></td>
</tr>
</tbody>
</table>
"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

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

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

⇒ 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
  - “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.

Fergus, Zisserman, Perona, ‘03
Representation of Appearance

Fergus, Zisserman, Perona, ‘03

11x11 patch
interest point
detection

size
normalized

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

- 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.

200 images containing faces
200 background images

Weber, Welling, Perona, ‘00
Constellation Model

- **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)

\[
p(X, S, A \mid \theta) = \sum_{h \in H} p(X, S, A, h \mid \theta) = \sum_{h \in H} p(A \mid X, S, h, \theta) p(X \mid S, h, \theta) p(S \mid h, \theta) p(h \mid \theta)
\]

- Gaussian part appearance pdf
- Gaussian shape pdf
- Gaussian relative scale pdf
- 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 using scale-invariant interest point detector
  - 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

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**

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**

- Part 1 – Det:2e–19
- Part 2 – Det:3e–18
- Part 3 – Det:2e–20
- Part 4 – Det:2e–22
- Part 5 – Det:3e–18
- Part 6 – Det:2e–18
- Background – Det:4e–20

$P(X, S, A, h | \theta)$

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

"cow" "motorbike" "car"
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”

For new image, let the matched patches vote for possible object positions
Implicit Shape Model - Representation

- 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

105 training images (+motion segmentation)
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

\[
p(i_j | e) \quad p(o_n, x | i_j) \quad o, x \quad p(o_n, x | i_j)p(i_j | e)
\]

\[
p(o_n, x | e) = \sum_j p(o_n, x | i_j)p(i_j | e)
\]
Object Categorization Procedure

Interest Points → Matched Codebook Entries → Probabilistic Voting

- Voting Space (continuous)
- Backprojection of Maximum
- Backprojected Hypothesis
- Refined Hypothesis (uniform sampling)
Car Categorization - Qualitative Results

- 1st hypothesis

- 2nd hypothesis

- 4th hypothesis

- 7th hypothesis

- 8th hypothesis
Results on Cows

Original image

Interest points

Matched patches

Prob. Votes
Results on Cows

1’st hypothesis
Results on Cows

2’nd hypothesis
Results on Cows

3’rd hypothesis
More Results on Cows…
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
    \[ x_{\text{vote}} = x_{\text{img}} - x_{\text{occ}} \left( \frac{S_{\text{img}}}{S_{\text{occ}}} \right) \]
    \[ y_{\text{vote}} = y_{\text{img}} - y_{\text{occ}} \left( \frac{S_{\text{img}}}{S_{\text{occ}}} \right) \]
    \[ s_{\text{vote}} = \left( \frac{S_{\text{img}}}{S_{\text{occ}}} \right) \]
  - 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"

Interest Points → Matched Codebook Entries → Probabilistic Voting

Segmentation → Refined Hypothesis (uniform sampling) → Backprojected Hypothesis → Backprojection of Maximum

Voting Space (continuous)
Segmentation: Probabilistic Formulation

- Influence of patch $e$ on object hypothesis

\[
p(e \mid o_n, x) = \frac{p(o_n, x \mid e)p(e)}{p(o_n, x)} = \sum_I p(o_n, x \mid I)p(I \mid e)p(e)
\]

- Backprojection to patches $e$ and pixels $p$:

\[
p(p = \text{figure} \mid o_n, x) = \sum_{p \in e} p(p = \text{figure} \mid e, o_n, x)p(e \mid o_n, x)
\]

Leibe, Schiele, '03
Segmentation: Probabilistic Formulation

- Resolve patches by interpretations (codebook entries) \( I \)

\[
p(p = \text{figure} \mid o_n, x) = \sum_{p \in e} \sum_{I} p(p = \text{figure} \mid e, I, o_n, x) p(e, I \mid o_n, x)
\]

\[
= \sum_{p \in e} \sum_{I} p(p = \text{figure} \mid I, o_n, x) \frac{p(o_n, x \mid I) p(I \mid e) p(e)}{p(o_n, x)}
\]

- Segmentation information
- Influence on object hypothesis

\( \Rightarrow \) Store patch segmentation mask for every occurrence position!

Leibe, Schiele, ‘03
Segmentation

Original image

\[ p(\text{figure}) \]

\[ p(\text{ground}) \]

Segmentation

\[ \frac{p(\text{figure})}{p(\text{ground})} \]
Segmentation

- Interpretation of $p(\text{figure})$ map
  - per-pixel confidence in object hypothesis
  - Use for hypothesis verification

Original image

$\overset{p(\text{figure})}{\rightarrow}$

$\overset{p(\text{ground})}{\leftarrow}$

Segmentation
Top-Down Driven Segmentation

- Example 1:
  - Pedestrian is segmented out since it does not contribute to the car hypothesis

- Example 2:
  - Image with sub-image contours
Motorbikes: Segmentation Results

Leibe, Schiele, ‘04
Hypothesis Verification: Motivation

- Secondary hypotheses
  - Desired property of algorithm! ⇒ robustness to partial occlusion
  - Standard solution: reject based on bounding box overlap
    ⇒ Problematic - may lead to missing detections!
    ⇒ Use segmentations to resolve ambiguities instead

Leibe, Leonardis, Schiele, ’04
Formalization in MDL Framework
(MDL = Minimum Description Length)

- Savings of one hypothesis [Leonardis, IJCV’95]

\[ S_h = K_0 S_{\text{area}} - K_1 S_{\text{model}} - K_2 S_{\text{error}} \]

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

\[ S_{\text{error}} = \sum_{p \in \text{Seg}(h)} (1 - p(p = \text{figure}|h)) \]

- Final form of equation

\[ S_h = -\frac{K_1}{K_0} + \left(1 - \frac{K_2}{K_0}\right) N + \frac{K_2}{K_0} \sum_{p \in \text{Seg}(h)} p(p = \text{figure}|h) \]
Formalization in MDL Framework (2)

- Savings of combined hypotheses

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

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

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

- In practice often sufficient to compute greedy approximation
Performance after Verification Stage

- Direct Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Equal Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarwal &amp; Roth</td>
<td>~79%</td>
</tr>
<tr>
<td>Garg et al.</td>
<td>~88%</td>
</tr>
<tr>
<td>Fergus et al.</td>
<td>88.5%</td>
</tr>
<tr>
<td>Our algorithm</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

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)

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

Leibe, Leonardis, Schiele, '04
Cows: Results on Novel Sequences (3)

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

Leibe, Leonardis, Schiele, ‘04
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