Object Recognition

• Intro
• Recognition of 3D objects
  • Recognition of object categories:
    • Bag of world models
    • Part based models
    • 3D object categorization
Challenges: intra-class variation
Usual Challenges:

Variability due to:

• View point
• Illumination
• Occlusions
Basic properties

• Representation
  – 2D Bag of Words (BoW) models;
  – 2D Star/ISM models;
  – Multi-view models;

• Learning
  – Generative & Discriminative BoW models
  – Probabilistic Hough voting
  – Generative multi-view models

• Recognition
  – Classification with BoW
  – Detection via Hough voting
Basic properties

• **Representation**
  – 2D Bag of Words (BoW) models;
  – 2D Star/ISM models;
  – Multi-view models;

• **Learning**
  – Generative & Discriminative BoW models
  – Probabilistic Hough voting
  – Generative multi-view models

• **Recognition**
  – Classification with BoW
  – Detection via Hough voting
definition of “BoW”

– Independent features
– histogram representation

codewords dictionary
representation

feature detection & representation

image representation

category models (and/or) classifiers

codewords dictionary

category decision

representation

recognition
Learning and Recognition

1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models

category models (and/or) classifiers
Discriminative classifiers

category models

Class 1

Class N

Model space
Discriminative models

**Nearest neighbor**

Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

**Support Vector Machines**

Guyon, Vapnik, Heisele, Serre, Poggio...

**Neural networks**

LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998...

**Latent SVM Structural SVM**

Felzenszwalb 00
Ramanan 03...

**Boosting**

Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,...
Support vector machines

- Find hyperplane that maximizes the margin between the positive and negative examples

![Diagram of support vectors and margin]

Support vectors: \( x_i \cdot w + b = \pm 1 \)

Distance between point and hyperplane:
\[
| x_i \cdot w + b | \over \| w \| 
\]

Margin = 2 / \( \| w \| \)

Solution: \[ w = \sum_i \alpha_i y_i x_i \]

Classification function (decision boundary):
\[ w \cdot x + b = \sum_i \alpha_i y_i x_i \cdot x + b \]
Support vector machines

- Classification

\[ \mathbf{w} \cdot \mathbf{x} + b = \sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b \]

Test point

\[ \text{if } \mathbf{x} \cdot \mathbf{w} + b \geq 0 \rightarrow \text{ class 1 } \]
\[ \text{if } \mathbf{x} \cdot \mathbf{w} + b < 0 \rightarrow \text{ class 2 } \]

Nonlinear SVMs

- Datasets that are linearly separable work out great:

- But what if the dataset is just too hard?

- We can map it to a higher-dimensional space:
Nonlinear SVMs

• General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:

\[ \Phi: x \rightarrow \varphi(x) \]

lifting transformation

Slide credit: Andrew Moore
Overfitting

- A simple dataset.
- Two models
Overfitting

- Let’s get more data.
- Simple model has better generalization.
Overfitting

• As complexity increases, the model overfits the data
• Training loss decreases
• Real loss increases
• We need to penalize model complexity = to regularize
1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models

category models (and/or) classifiers
Generative models

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

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

- Discriminative methods model posterior
- Generative methods model likelihood and prior
- Bayes rule:

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} = \frac{p(\text{image} \mid \text{zebra})}{p(\text{image} \mid \text{no zebra})} \cdot \frac{p(\text{zebra})}{p(\text{no zebra})}
\]

posterior ratio likelihood ratio prior ratio
Some notations

• $w$: a collection of all $N$ codewords in the image

$w = [w_1, w_2, \ldots, w_N]$

• $c$: category of the image
the Naïve Bayes model

\[ p(c \mid w) \sim p(c)p(w \mid c) = p(c) p(w_1, \ldots, w_N \mid c) \]

Prior prob. of the object classes

Image likelihood given the class
### the Naïve Bayes model

\[
p(c \mid w) \sim p(c)p(w \mid c) = p(c) p(w_1, \ldots, w_N \mid c)
\]

\[
= p(c) \prod_{n=1}^{N} p(w_n \mid c)
\]

- **Prior prob. of the object classes**
- **Image likelihood given the class**
- **Likelihood of n\textsuperscript{th} visual word given the class**

**Assume that each feature (codewords) is conditionally independent** 
*given the class*

\[
p(w_1, \ldots, w_N \mid c) = \prod_{i=1}^{N} p(w_i \mid c)
\]
the Naïve Bayes model

\[ p(c \mid w) \sim p(c)p(w \mid c) = p(c) \ p(w_1, \ldots, w_N \mid c) \]

\[ = p(c) \prod_{n=1}^{N} p(w_n \mid c) \]

Prior prob. of the object classes

Image likelihood given the class

Likelihood of n\textsuperscript{th} visual word given the class
**Classification/Recognition**

\[ c^* = \arg \max_c p(c \mid w) \propto p(c) \prod_{n=1}^{N} p(w_n \mid c) \]

Object class decision

- How do we learn \( P(w_i \mid c_j) \)?
- From empirical frequencies of code words in images from a given class

![Graph showing class densities](image)
Our in-house database contains 1776 images in seven classes\(^1\): faces, buildings, trees, cars, phones, bikes and books. Fig. 2 shows some examples from this dataset.
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>6</td>
<td>7</td>
<td>2</td>
<td>69</td>
<td></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>

E = 28%

Table 2. Confusion matrix and mean rank for SVM (k=1000, linear kernel).

<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>98</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>34</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>buildings</td>
<td>1</td>
<td>63</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>trees</td>
<td>1</td>
<td>10</td>
<td>81</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>cars</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>85</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>phones</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>55</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>bikes</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>91</td>
<td>0</td>
</tr>
<tr>
<td>books</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>Mean ranks</td>
<td>1.04</td>
<td>1.77</td>
<td>1.28</td>
<td>1.30</td>
<td>1.83</td>
<td>1.09</td>
<td>1.39</td>
</tr>
</tbody>
</table>

E = 15%
Summary: Generative models

- Naïve Bayes
  - *Unigram models* in document analysis
  - Assumes conditional independence of words given class
  - Parameter estimation: frequency counting
Other generative BoW models

- Hierarchical Bayesian topic models (e.g. pLSA and LDA)
  - Natural scene categorization: Fei-Fei et al. 2005
Generative vs discriminative

• Discriminative methods
  – Computationally efficient & fast

• Generative models
  – Convenient for weakly- or un-supervised, incremental training
  – Prior information
  – Flexibility in modeling parameters
Weakness of BoW the models

- No rigorous geometric information of the object components
- It’s intuitive to most of us that objects are made of parts – no such information
- Not extensively tested yet for
  - View point invariance
  - Scale invariance
- Segmentation and localization unclear
EECS 442 – Computer vision

Object Recognition

- Intro
- Recognition of 3D objects
  - Recognition of object categories:
    - Bag of world models
    - Part based models
    - 3D object categorization
Basic properties

• **Representation**
  - 2D Bag of Words (BoW) models;
  - 2D Star/ISM models;
  - Multi-view models;

• **Learning**
  - Generative & Discriminative BoW models
  - Probabilistic Hough voting
  - Generative multi-view models

• **Recognition**
  - Classification with BoW
  - Detection via Hough voting
Problem with bag-of-words

- All have equal probability for bag-of-words methods
- Location information is important
Part Based Representation

- Object as set of parts

- Model:
  - Relative locations between parts
  - Appearance of part

Figure from [Fischler & Elschlager 73]
History of Parts and Structure approaches

- Fischler & Elschlager 1973
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Perona et al. ‘95, ’96, ’98, ’00, ’03, ’04, ’05
- Ullman et al. 02
- Felzenszwalb & Huttenlocher ’00, ’04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04
- Many papers since 2000
Deformations

A

B

C

D
Presence / Absence of Features
Background clutter
Sparse representation

Computationally tractable ($10^5$ pixels $\rightarrow 10^1 -- 10^2$ parts)
But throw away potentially useful image information
Discriminative

Parts need to be distinctive to separate from other classes
Hierarchical representations

- Pixels → Pixel groupings → Parts → Object

Images from [Amit98, Bouchard05]
Different connectivity structures

- a) Constellation \([O(N^6)]\)
- b) Star shape \([O(N^2)]\)
- c) \(k\)-fan \((k = 2) [O(N^3)]\)
- d) Tree \([O(N^2)]\)
- e) Bag of features \([O(N^6)]\)
- f) Hierarchy \([O(N^3)]\)
- g) Sparse flexible model

- Fergus et al. ’03
- Fei-Fei et al. ’03
- Crandall et al. ’05
- Leibe ’05: Felzenszwalb ’09
- Crandall et al. ’05
- Felzenszwalb & Huttenlocher ’00
- Csurka ’04
- Vasconcelos ’00
- Bouchard & Triggs ’05
- Carneiro & Lowe ’06

from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006
Stochastic Grammar of Images

S.C. Zhu et al. and D. Mumford
Context and Hierarchy in a Probabilistic Image Model
Jin & Geman (2006)

animal head instantiated by tiger head

animal head instantiated by bear head

e.g. discontinuities, gradient

e.g. linelets, curvelets, T-junctions

e.g. animals, trees, rocks

e.g. contours, intermediate objects

e.g. discontinuities, gradient
Different connectivity structures

- a) Constellation [13]
  Fergus et al. ’03
  Fei-Fei et al. ’03

- b) Star shape [9, 14]
  Crandall et al. ’05
  Leibe 05; Felzenszwalb 09

- c) $k$-fan ($k = 2$) [9]
  Crandall et al. ’05
  Felzenszwalb & Huttenlocher ’00

- d) Tree [12]

- e) Bag of features [10, 21]
  Csurka ’04
  Vasconcelos ’00

- f) Hierarchy [4]
  Bouchard & Triggs ’05

- g) Sparse flexible model
  Carneiro & Lowe ’06

from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006
Star models by Latent SVM

Felzenszwalb, McAllester, Ramanan, 08
• Source code:
Star models by Latent SVM

– Search strategy: Sliding Windows
  • Simple but large computational complexity \((x,y, S, \theta, N \text{ of classes})\)
Implicit shape models by generalized Hough voting

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
Basic properties

• Representation
  – How to represent an object category; which classification scheme?

• Learning
  – How to learn the classifier, given training data

• Recognition
  – How the classifier is to be used on novel data
Object representation:
Constellation of parts w.r.t object centroid

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
Object representation:
How to capture constellation of parts?
Using Hough Voting

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
Basic properties

- Representation
  - How to represent an object category; which classification scheme?

- Learning
  - How to learn the classifier, given training data

- Recognition
  - How the classifier is to be used on novel data
Hough Transform Formulation

• Parts in query image vote for a learnt model
• Significant aggregations of votes correspond to models
• Complexity: \# parts * \# votes
  – Significantly lower than brute force search (e.g., sliding window detectors)
• Popular for detecting parameterized shapes
  – Hough’59, Duda&Hart’72, Ballard’81,…
Hough transform

Given a set of points, find the curve or line that explains the data points best

Hough transform


Given a set of points, find the curve or line that explains the data points best.

\[ y = mx + n \]

Hough space

\[ y_1 = m x_1 + n \]
Hough transform


Issue: parameter space \([m,n]\) is unbounded...

*Use a polar representation for the parameter space*

\[
x \cos \theta + y \sin \theta = \rho
\]
Hough transform - experiments

features

votes
IDEA: introduce a grid a count intersection points in each cell
Issue: Grid size needs to be adjusted…
Hough transform - experiments

Issue: spurious peaks due to uniform noise
Hough transform - conclusions

Good:

• All points are processed independently, so can cope with occlusion/outliers
• Some robustness to noise: noise points unlikely to contribute consistently to any single bin

Bad:

• Spurious peaks due to uniform noise
• Trade-off noise-grid size (hard to find sweet point)
Hough transform - experiments

Courtesy of TKK Automation Technology Laboratory
Generalized Hough transform

- What if we want to detect arbitrary shapes defined by boundary points and a reference point?

At each boundary point, compute displacement vector: \( \mathbf{r} = a - p_i \).

For a given model shape: store these vectors in a table indexed by gradient orientation \( \theta \).

[Dana H. Ballard, Generalizing the Hough Transform to Detect Arbitrary Shapes, 1980]
Generalized Hough transform

To detect the model shape in a new image:

• For each edge point
  – Index into table with its gradient orientation $\theta$
  – Use retrieved $r$ vectors to vote for position of reference point
• Peak in this Hough space is reference point with most supporting edges

Assuming translation is the only transformation here, i.e., orientation and scale are fixed.
Circle model

Example

Query

\[ P_1 \rightarrow \theta = 0 \Rightarrow R = [rx, ry] = [1, 0] \Rightarrow C_1 = P_1 + R \]
\[ P_2 \rightarrow \theta = 45 \Rightarrow R = [rx, ry] = [0.7, 0.7] \Rightarrow C_2 = P_2 + R \]
\[ P_k \rightarrow \theta = -180 \Rightarrow R = [rx, ry] = [-1, 0] \Rightarrow C_k = P_k + R \]

\[ \theta \quad rx \quad ry \]
\[ 0 \quad 1 \quad 0 \]
\[ 45 \quad 0.7 \quad 0.7 \]
\[ 90 \quad 0 \quad 1 \]
\[ 135 \quad -0.7 \quad 0.7 \]
\[ \ldots \]
\[ 270 \quad 0.7 \quad -0.7 \]
Implicit shape models

- Instead of indexing displacements by gradient orientation, index by “visual codeword”

→ Visual codebook is used to index votes for object position [center] and scale

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

- Instead of indexing displacements by gradient orientation, index by “visual codeword”

→ Visual codebook is used to index votes for object position [center] and scale

<table>
<thead>
<tr>
<th>CW</th>
<th>rx</th>
<th>ry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9</td>
<td>.1</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

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

• Instead of indexing displacements by gradient orientation, index by “visual codeword”

→ Visual codebook is used to index votes for object position [center] and scale

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
3. For each codebook entry, store all positions relative to object center [center is given] and scale [bounding box is given]
Implicit Shape Model - Recognition

Interest Points → Matched Codebook Entries → Probabilistic Voting

3D Voting Space (continuous)

[Leibe, Leonardis, Schiele, SLCV'04; IJCV'08]
Implicit Shape Model - Recognition

Interest Points → Matched Codebook Entries → Probabilistic Voting

Segmentation → Backprojected Hypotheses → Backprojection of Maxima

3D Voting Space (continuous)
Probabilistic Hough Transform

Detection Score

Position Posterior
distribution of the centroid
given the Codeword $C_i$
observed at location $l_j$.

Codeword Match

confidence (or
weight) of the
codeword $C_i$.

Learnt using a
max margin
formulation
Maji et al, CVPR 2009

$f = \text{features}$
$l = \text{location of the features}$.
$C = \text{codebook entry}$
$O = \text{object class}$
$x = \text{object center}$
Example: Results on Cows
Example: Results on Cows

Interest points
Example: Results on Cows

Matched patches
Example: Results on Cows

Prob. Votes
Example: Results on Cows

1\textsuperscript{st} hypothesis
Example: Results on Cows

2\textsuperscript{nd} hypothesis
Example: Results on Cows

3rd hypothesis
Example Results: Chairs

Office chairs

Dining room chairs
You Can Try It At Home...

- Linux binaries available
  - Including datasets & several pre-trained detectors
  - [http://www.vision.ee.ethz.ch/bleibe/code](http://www.vision.ee.ethz.ch/bleibe/code)
Extension: Learning Feature Weights

Subhransu Maji and Jitendra Malik, Object Detection Using a Max-Margin Hough Transform, CVPR 2009

Weights can be learned optimally using a max-margin framework.
Learned Weights (ETHZ shape)

Naïve Bayes

Max-Margin

Influenced by clutter
(rare structures)

Important Parts

blue (low), dark red (high)
Learned Weights (UIUC cars)

Naïve Bayes

Important Parts

Max-Margin

blue (low), dark red (high)
## Detection Results (ETHZ dataset)

Recall @ 1.0 False Positives Per Window

<table>
<thead>
<tr>
<th>Category</th>
<th>Uniform</th>
<th>Naive Bayes</th>
<th>Max-margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applelogos</td>
<td>70.0</td>
<td>70.0</td>
<td>85.0</td>
</tr>
<tr>
<td>Bottles</td>
<td>62.5</td>
<td>71.4</td>
<td>67.0</td>
</tr>
<tr>
<td>Giraffes</td>
<td>47.1</td>
<td>47.1</td>
<td>55.0</td>
</tr>
<tr>
<td>Mugs</td>
<td>35.5</td>
<td>35.5</td>
<td>55.0</td>
</tr>
<tr>
<td>Swans</td>
<td>47.1</td>
<td>47.1</td>
<td>42.5</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>52.4</strong></td>
<td><strong>54.2</strong></td>
<td><strong>60.9</strong></td>
</tr>
</tbody>
</table>
Conclusions

• **Pros:**
  – Works well for many different object categories
    • Both rigid and articulated objects
  – Flexible geometric model
    • Can recombine parts seen on different training examples
  – Learning from relatively few (50-100) training examples
  – Optimized for detection, good localization properties

• **Cons:**
  – Needs supervised training data
    • Object bounding boxes for detection
    • Segmentations for top-down segmentation
  – No discriminative learning
Object detection and 3d shape recovery from a single image

- M. Sun, B. Xu, G. Bradski, S. Savarese, ECCV 2010
- M. Sun, S. Kumar, G. Bradski, S. Savarese, 3DIM-PVT 2011

- Object categories
- Arbitrary Texture, Topology
- Uncontrolled Environment
- Un-calibrated Camera
Depth Encoded Hough Voting

: Object location \( x \)

: Part with center location \( l \) and scale \( s \)
Depth Encoded Hough Voting

\[ S(O, x) \propto \sum_{i,j} p(x|O, C_i, l_j)p(C_i|f_j)p(O|C_i, l_j) \]

\[ \sum_{i,j} \int p(x|O, C_i, l_j, s_j, d_j)p(C_i|f_j)p(O|C_i, l_j, s_j, d_j)p(s_j|l_j, d_j) \, ds_j \]

C = Codebook
f = Part appearance(feature), l = 2D location
s = Scale of the part, d = Depth information,
Object detection and 3d shape recovery from a single image

3d reconsctruction
Object detection and 3d shape recovery from a single image

3d pose estimation
Object detection and 3d shape recovery from a single image

3d modeling
Object detection and 3d shape recovery from a single image

CAD Alignment

Texture Completion
Summary

• Part based models enable more descriptive characterization of objects
  – Semantically useful regions
  – Depth
  – Occlusions

• ISM models
  – Generative
  – Enable segmentation
  – Require large supervision