Lecture 15:
Object recognition:
Bag of Words models &
Part-based generative models

Professor Fei-Fei Li
Stanford Vision Lab
Basic issues

• 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
What we will learn today?

• Bag of Words model (Problem Set 4 (Q2))
  – Basic representation
  – Different learning and recognition algorithms

• Constellation model
  – Weakly supervised training
  – One-shot learning (supplementary materials)

• (Problem Set 4 (Q1))
What we will learn today?

• Bag of Words model *(Problem Set 4 (Q2))*
  – Basic representation
  – Different learning and recognition algorithms

• Constellation model
  – Weakly supervised training
  – One-shot learning (supplementary materials)

• *(Problem Set 4 (Q1))*
Part 1: Bag-of-words models

This segment is based on the tutorial “Recognizing and Learning Object Categories: Year 2007”, by Prof L. Fei-Fei, A. Torralba, and R. Fergus
Related works

• Early “bag of words” models: mostly texture recognition

• Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  – Hoffman 1999; Blei, Ng & Jordan, 2004; Teh, Jordan, Beal & Blei, 2004

• Object categorization
  – Csurka, Bray, Dance & Fan, 2004; Sivic, Russell, Efros, Freeman & Zisserman, 2005; Sudderth, Torralba, Freeman & Willsky, 2005;

• Natural scene categorization
  – Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006
Object → Bag of ‘words’
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach us through our eyes. For a long time it was thought that retinal image was seen by the brain as a movie slide, but subsequent discoveries have shown that we know the visual world is more complex. Following the discovery by Hubel and Wiesel, they have demonstrated that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004’s $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% increase in 2005 to $750bn, compared with $660bn. The US has been annoyed that China’s leaders have repeatedly agreed that the yuan is not yet ready for foreign exchange. China’s government also needs the yuan to fall to meet domestic demand so that it can keep its exports competitive. China’s leaders have not permitted the yuan against the dollar to move freely, but the US wants the yuan to be allowed to rise freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
definition of “BoW”

– Independent features

face

bike

violin
definition of “BoW”

– Independent features
– histogram representation

codewords dictionary
Lecture 15 - Fei-Fei Li

Representation

- feature detection & representation
- image representation

recognition

- codewords dictionary

learning

- category models (and/or) classifiers
1. Feature detection and representation
1. Feature detection and representation

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
1. Feature detection and representation

- **Regular grid**
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005

- **Interest point detector**
  - Csurka, et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic, et al. 2005
1. Feature detection and representation

- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005

- Interest point detector
  - Csurka, Bray, Dance & Fan, 2004
  - Fei-Fei & Perona, 2005
  - Sivic, Russell, Efros, Freeman & Zisserman, 2005

- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)
1. Feature detection and representation

- Detect patches
  - [Mikojaczyk and Schmid '02]
  - [Mata, Chum, Urban & Pajdla, '02]
  - [Sivic & Zisserman, '03]

- Normalize patch
- Compute SIFT descriptor
  - [Lowe'99]
1. Feature detection and representation
2. Codewords dictionary formation
2. Codewords dictionary formation

Cluster center = code word

Clustering/ vector quantization
2. Codewords dictionary formation

Fei-Fei et al. 2005
Image patch examples of codewords

Sivic et al. 2005
Visual vocabularies: Issues

• How to choose vocabulary size?
  – Too small: visual words not representative of all patches
  – Too large: quantization artifacts, overfitting

• Computational efficiency
  – Vocabulary trees
    (Nister & Stewenius, 2006)
3. Bag of word representation

- Nearest neighbors assignment
- K-D tree search strategy

Codewords dictionary
3. Bag of word representation

- Codewords dictionary
- Codewords frequency

Fei-Fei Li
Lecture 15 - 24 16-Nov-11
Representation

1. feature detection & representation

2. codewords dictionary

3. image representation
Learning and Recognition

codewords dictionary

category models (and/or) classifiers

category decision
1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models

category models (and/or) classifiers
Discriminative classifiers

category models

Model space

Class 1

Class N
Discriminative classifiers

Query image

Winning class: pink

Model space
Nearest Neighbors classifier

Query image

Winning class: pink

• Assign label of nearest training data point to each test data point
K-Nearest Neighbors classifier

Query image

Winning class: pink

• For a new point, find the k closest points from training data
• Labels of the k points “vote” to classify
• Works well provided there is lots of data and the distance function is good
Lecture 15 - Fei-Fei Li

• For $k$ dimensions: $k$-D tree = space-partitioning data structure for organizing points in a $k$-dimensional space
• Enable efficient search

K-Nearest Neighbors classifier

• Voronoi partitioning of feature space for 2-category 2-D and 3-D data

16-Nov-11 32
Functions for comparing histograms

- **L1 distance**
  \[ D(h_1, h_2) = \sum_{i=1}^{N} |h_1(i) - h_2(i)| \]

- **\(\chi^2\) distance** (popular)
  \[ D(h_1, h_2) = \sum_{i=1}^{N} \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)} \]

- **Quadratic distance (cross-bin)**
  \[ D(h_1, h_2) = \sum_{i,j} A_{ij} (h_1(i) - h_2(j))^2 \]

Learning and Recognition

1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models
Discriminative classifiers (linear classifier)

category models

Class 1

Class N

Model space
Support vector machines

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

Distance between point and hyperplane:

\[ \frac{|x_i \cdot w + b|}{\|w\|} \]

Margin:

\[ \frac{2}{\|w\|} \]

Support vectors:

\[ x_i \cdot w + b = \pm 1 \]

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

\[ w \cdot x + b = \sum_i \alpha_i y_i x_i \cdot x + b \]

Test point

if \( x \cdot w + b \geq 0 \) \( \rightarrow \) class 1

if \( x \cdot w + b < 0 \) \( \rightarrow \) class 2

C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998
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:

\[ \Phi: \mathbf{x} \rightarrow \varphi(\mathbf{x}) \]
Nonlinear SVMs

• Nonlinear decision boundary in the original feature space:

\[ \sum_i \alpha_i y_i K(x_i, x) + b \]

• The kernel \( K = \) product of the lifting transformation \( \varphi(x) \):

\[ K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \]

NOTE:
• It is not required to compute \( \varphi(x) \) explicitly:
• The kernel must satisfy the “Mercer inequality”
Kernels for bags of features

• Histogram intersection kernel:

\[ I(h_1, h_2) = \sum_{i=1}^{N} \min(h_1(i), h_2(i)) \]

• Generalized Gaussian kernel:

\[ K(h_1, h_2) = \exp \left( -\frac{1}{A} D(h_1, h_2)^2 \right) \]

• \( D \) can be Euclidean distance, \( \chi^2 \) distance etc...

Pyramid match kernel

- Fast approximation of Earth Mover’s Distance
- Weighted sum of histogram intersections at multiple resolutions (linear in the number of features instead of cubic)

Spatial Pyramid Matching

What about multi-class SVMs?

• No “definitive” multi-class SVM formulation
• In practice, we have to obtain a multi-class SVM by combining multiple two-class SVMs
• One vs. others
  – Training: learn an SVM for each class vs. the others
  – Testing: apply each SVM to test example and assign to it the class of the SVM that returns the highest decision value
• One vs. one
  – Training: learn an SVM for each pair of classes
  – Testing: each learned SVM “votes” for a class to assign to the test example

Credit slide: S. Lazebnik
SVMs: Pros and cons

• Pros
  – Many publicly available SVM packages: http://www.kernel-machines.org/software
  – Kernel-based framework is very powerful, flexible
  – SVMs work very well in practice, even with very small training sample sizes

• Cons
  – No “direct” multi-class SVM, must combine two-class SVMs
  – Computation, memory
    • During training time, must compute matrix of kernel values for every pair of examples
    • Learning can take a very long time for large-scale problems
Object recognition results

- ETH-80 database of 8 object classes  
  \[ (Eichhorn \text{ and Chapelle 2004}) \]

- Features:
  - Harris detector
  - PCA-SIFT descriptor, \( d=10 \)

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Complexity</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match [Wallraven et al.]</td>
<td>( O(dm^2) )</td>
<td>84%</td>
</tr>
<tr>
<td>Bhattacharyya affinity</td>
<td>( O(dm^3) )</td>
<td>85%</td>
</tr>
<tr>
<td>Pyramid match</td>
<td>( O(dmL) )</td>
<td>84%</td>
</tr>
</tbody>
</table>

Slide credit: Kristen Grauman
Discriminative models

**Nearest neighbor**

- 10^8 examples
- 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,...

Source: Vittorio Ferrari, Kristen Grauman, Antonio Torralba

Fei-Fei Li  Lecture 15 - 47  16-Nov-11
1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models

→ Model the probability distribution that produces a given bag of features
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
Some notations

• \textbf{w}: a collection of all N codewords in the image
  \[
  \mathbf{w} = [w_1, w_2, \ldots, w_N]
  \]

• \textbf{c}: category of the image
the Naïve Bayes model

\[ \text{Posterior} = p(c \mid w) \propto p(c) p(w \mid c) \]

\( p(c) \) = Prior prob. of the object classes
\( p(w \mid c) \) = Image likelihood given the class

\( p(c \mid w) \) = Probability that image I is of category c
the Naïve Bayes model

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

Object class decision

Likelihood of ith visual word given the class

Estimated by empirical frequencies of code words in images from a given class
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>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>
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
What we will learn today?

• Bag of Words model (Problem Set 4 (Q2))
  – Basic representation
  – Different learning and recognition algorithms

• Constellation model
  – Weakly supervised training
  – One-shot learning (supplementary materials)

• (Problem Set 4 (Q1))
Model: Parts and Structure
Parts and Structure Literature

- 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
- Huttenlocher et al. ’00
- Agarwal & Roth ’02
- etc…
The Constellation Model

- T. Leung: Representation
  - Shape statistics – F&G ’95
  - Affine invariant shape – CVPR ’98

- M. Burl: Detection
  - CVPR ’96
  - ECCV ’98

  - ECCV ’00
  - Multiple views - F&G ’00
  - Discovering categories - CVPR ’00

- R. Fergus: Joint shape & appearance learning
  - Generic feature detectors
  - CVPR ’03
  - Polluted datasets - ECCV ’04

- L. Fei-Fei: One-Shot Learning
  - Incremental learning
  - ICCV ’03
  - CVPR ’04
Deformations
Presence / Absence of Features
Background clutter
Generative probabilistic model

Foreground model
- Gaussian shape pdf

Clutter model
- Uniform shape pdf

Prob. of detection
- 0.8
- 0.75
- 0.9

# detections
- $P_{\text{Poisson}}(N_1|\lambda_1)$
- $P_{\text{Poisson}}(N_2|\lambda_2)$
- $P_{\text{Poisson}}(N_3|\lambda_3)$

Assumptions:
(a) Clutter independent of foreground detections
(b) Clutter detections independent of each other

Example
1. Object Part Positions
2. Part Absence
3a. N false detect
3b. Position f. detect

Fei-Fei Li
Lecture 15 - 65
16-Nov-11
Learning Models `Manually’

- Obtain set of training images
- Choose parts
- Label parts by hand, train detectors
- Learn model from labeled parts
Recognition

1. Run part detectors exhaustively over image

2. Try different combinations of detections in model
   - Allow detections to be missing (occlusion)

3. Pick hypothesis which maximizes:
   \[
   \frac{p(Data | Object, Hyp)}{p(Data | Clutter, Hyp)}
   \]

4. If ratio is above threshold then, instance detected
So far.....

• Representation
  – Joint model of part locations
  – Ability to deal with background clutter and occlusions

• Learning
  – Manual construction of part detectors
  – Estimate parameters of shape density

• Recognition
  – Run part detectors over image
  – Try combinations of features in model
  – Use efficient search techniques to make fast
Unsupervised Learning
Weber & Welling et. al.
(Semi) Unsupervised learning

- Know if image contains object or not
- But no segmentation of object or manual selection of features
Highly textured neighborhoods are selected automatically
produces 100-1000 patterns per image
Unsupervised detector training - 2

“Pattern Space” (100+ dimensions)
Unsupervised detector training - 3

100-1000 images ~100 detectors
Learning

- Take training images. Pick set of detectors. Apply detectors.
- 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
ML using EM

1. Current estimate

2. Assign probabilities to constellations

3. Use probabilities as weights to re-estimate parameters. Example: \( \mu \)

\[
\text{Large } P \times \quad \text{Small } P \times \quad + \quad \ldots =
\]

new estimate of \( \mu \)
Detector Selection

• Try out different combinations of detectors (Greedy search)

Detectors (≈100)

Choice 1

Choice 2

Parameter Estimation

Model 1

Model 2

Predict / measure model performance (validation set or directly from model)
Frontal Views of Faces

- 200 Images (100 training, 100 testing)
- 30 people, different for training and testing
Learned face model

Pre-selected Parts

Test Error: 6% (4 Parts)

Parts in Model

Sample Detection

Model Foreground pdf
Face images
Background images
Car from Rear

Preselected Parts

Test Error: 13% (5 Parts)

Parts in Model

Model Foreground pdf

Sample Detection

Fei-Fei Li
Detections of Cars
3D Object recognition – Multiple mixture components
3D Orientation Tuning

Orientation Tuning

% Correct vs. angle in degrees

Frontal vs. Profile

Fei-Fei Li

Lecture 15 - 85 16-Nov-11
So far (2).....

- **Representation**
  - Multiple mixture components for different viewpoints

- **Learning**
  - Now semi-unsupervised
  - Automatic construction and selection of part detectors
  - Estimation of parameters using EM

- **Recognition**
  - As before

- **Issues:**
  - Learning is slow (many combinations of detectors)
  - Appearance learnt first, then shape
Issues

• Speed of learning
  – Slow (many combinations of detectors)

• Appearance learnt first, then shape
  – Difficult to learn part that has stable location but variable appearance
  – Each detector is used as a cross-correlation filter, giving a hard definition of the part’s appearance

• Would like a fully probabilistic representation of the object
Object categorization

Fergus et. al.

CVPR ’03, IJCV ‘06
Detection & Representation of regions

• Find regions within image
• Use salient region operator
  (Kadir & Brady 01)

Location

(x, y) coords. of region centre

Scale

Radius of region (pixels)

Appearance

Normalize

11x11 patch

Projection onto PCA basis

Gives representation of appearance in low-dimensional vector space
Motorbikes example

• Kadir & Brady saliency region detector
Generative probabilistic model (2)

Foreground model

Gaussian shape pdf

Gaussian part appearance pdf

Gaussian relative scale pdf

Prob. of detection

Clutter model

Uniform shape pdf

Gaussian background appearance pdf

Uniform relative scale pdf

Poission pdf on # detections

Based on Burl, Weber et al. [ECCV ’98, ’00]

Fei-Fei Li

Lecture 15 - 91
16-Nov-11
Motorbikes

Samples from appearance model

Shape model

Part 1: Det: 5x10-18
Part 2: Det: 6x10-22
Part 3: Det: 6x10-18
Part 4: Det: 1x10-19
Part 5: Det: 3x10-17
Part 6: Det: 4x10-24
Background: Det: 5x10-19

Fei-Fei Li

92 16-Nov-11
Recognized Motorbikes

Shape model

Fei-Fei Li

Lecture 15 - 93 16-Nov-11
Background images evaluated with motorbike model
Frontal faces
Airplanes
Spotted cats
## Summary of results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fixed scale experiment</th>
<th>Scale invariant experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>7.5</td>
<td>6.7</td>
</tr>
<tr>
<td>Faces</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>Airplanes</td>
<td>9.8</td>
<td>7.0</td>
</tr>
<tr>
<td>Cars (Rear)</td>
<td>15.2</td>
<td>9.7</td>
</tr>
<tr>
<td>Spotted cats</td>
<td>10.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

% equal error rate

Note: Within each series, same settings used for all datasets
Comparison to other methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ours</th>
<th>Others</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorbikes</td>
<td>7.5</td>
<td>16.0</td>
<td>Weber et al. [ECCV ‘00]</td>
</tr>
<tr>
<td>Faces</td>
<td>4.6</td>
<td>6.0</td>
<td>Weber</td>
</tr>
<tr>
<td>Airplanes</td>
<td>9.8</td>
<td>32.0</td>
<td>Weber</td>
</tr>
<tr>
<td>Cars (Side)</td>
<td>11.5</td>
<td>21.0</td>
<td>Agarwal–Roth [ECCV ‘02]</td>
</tr>
</tbody>
</table>

% equal error rate
Why this design?

- Generic features seem to well in finding consistent parts of the object
- Some categories perform badly – different feature types needed
- Why PCA representation?
  - Tried ICA, FLD, Oriented filter responses etc.
  - But PCA worked best
- Fully probabilistic representation lets us use tools from machine learning community
What we have learned today?

• Bag of Words model \((\text{Problem Set 4 (Q2)})\)
  – Basic representation
  – Different learning and recognition algorithms

• Constellation model
  – Weakly supervised training
  – One-shot learning (supplementary materials)

• \((\text{Problem Set 4 (Q1)})\)
Supplementary materials

• One-Shot learning using Constellation Model
One-Shot learning
Fei-Fei et. al.

ICCV ’03, PAMI ‘06
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Examples</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viola et al.</td>
<td>~10,000</td>
<td>Faces</td>
</tr>
<tr>
<td>Schneiderman, et al.</td>
<td>~2,000</td>
<td>Faces, Cars</td>
</tr>
<tr>
<td>Rowley et al.</td>
<td>~500</td>
<td>Faces</td>
</tr>
</tbody>
</table>
Number of training examples

Generalisation performance

6 part Motorbike model

Previously

Classification error (%) vs. \( \log_2 \) (Training images)

Test
Train
How do we do better than what statisticians have told us?

- Intuition 1: use Prior information
- Intuition 2: make best use of training information
Prior knowledge: means

Appearance

Shape

likely

unlikely
Bayesian framework

\[ P(\text{object} \mid \text{test}, \text{train}) \quad \text{vs.} \quad P(\text{clutter} \mid \text{test}, \text{train}) \]

Bayes Rule

\[ p(\text{test} \mid \text{object}, \text{train}) \quad p(\text{object}) \]

Expansion by parametrization

\[ \int p(\text{test} \mid \theta, \text{object}) \quad p(\theta \mid \text{object}, \text{train}) \quad d\theta \]
Bayesian framework

\[ P(\text{object} \mid \text{test, train}) \quad \text{vs.} \quad P(\text{clutter} \mid \text{test, train}) \]

\[
p(\text{test} \mid \text{object, train}) \quad p(\text{object})
\]

Expansion by parametrization

\[
\int p(\text{test} \mid \theta, \text{object}) \quad p(\theta \mid \text{object, train}) \quad d\theta
\]

Previous Work: \( \delta(\theta^{\text{ML}}) \)
Bayesian framework

\[ P(\text{object} \mid \text{test, train}) \quad \text{vs.} \quad P(\text{clutter} \mid \text{test, train}) \]

Bayes Rule

\[ P(\text{test} \mid \text{object, train}) \cdot p(\text{object}) \]

Expansion by parametrization

\[ \int p(\text{test} \mid \theta, \text{object}) \cdot p(\theta \mid \text{object, train}) \, d\theta \]

One-Shot learning:

\[ p(\text{train} \mid \theta, \text{object}) \cdot p(\theta) \]
Each object model $\theta$

Gaussian shape pdf

Gaussian part appearance pdf
Lecture 15 - Fei-Fei Li

Model Structure

model distribution: $p(\theta)$
- conjugate distribution of $p(train|\theta,\text{object})$

Each object model $\theta$

Gaussian shape pdf

Gaussian part appearance pdf

model (\theta) space

$\theta_n$, $\theta_1$, $\theta_2$
Learning Model Distribution

\[ p(\theta | \text{object, train}) \propto p(\text{train} | \theta, \text{object}) p(\theta) \]

- use **Prior** information
- Bayesian learning
  - marginalize over \( \theta \)
  - **Variational EM** *(Attias, Hinton, Minka, etc.)*
Variational EM

E-Step

M-Step

new $\theta$'s

new estimate of $p(\theta|\text{train})$

prior knowledge of $p(\theta)$

Random initialization

Fei-Fei Li
Experiments

Training:
1- 6 randomly drawn images

Testing:
50 fg/ 50 bg images object present/absent

Datasets
- faces
- airplanes
- spotted cats
- motorbikes

117 16-Nov-11
Experiments: obtaining priors

- airplanes
- spotted cats
- motorbikes
- faces

model ($\theta$) space
Experiments: obtaining priors

- airplanes
- faces
- motorbikes

model (θ) space

spotted cats
Lecture 15 - Fei-Fei Li

Performance comparison

Maximum-Likelihood
Bayesian OneShot

Number of training examples

Correct
Correct
Correct

Correct
INCORRECT
Correct

Shape Model (Training # = 1)
Lecture 15 - Fei-Fei Li

Performance comparison

Number of training examples

Performance (equal error rates)

Maximum-Likelihood
Bayesian OneShot

Correct
Correct
Correct

INcorrect
INcorrect
Correct

Shape Model (Training # = 1)
Performance comparison

Number of training examples

- Maximum-Likelihood
- Bayesian OneShot

Correct

Shape Model (Training # = 1)
Performance comparison

Number of training examples

Performance (equal error rates)

Maximum-Likelihood
Bayesian OneShot

INCORRECT
Correct
Correct
Correct
Correct
Correct
Correct

Shape Model (Training # = 1)
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Examples</th>
<th>Categories</th>
<th>Results (error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viola et al.</td>
<td>~10,000</td>
<td>Faces</td>
<td>7-21%</td>
</tr>
<tr>
<td>Schneiderman, et al.</td>
<td>~2,000</td>
<td>Faces, Cars</td>
<td>5.6 – 17%</td>
</tr>
<tr>
<td>Rowley et al.</td>
<td>~500</td>
<td>Faces</td>
<td>7.5 – 24.1%</td>
</tr>
<tr>
<td>Bayesian One-Shot</td>
<td>1 ~ 5</td>
<td>Faces, Motorbikes, Spotted cats, Airplanes</td>
<td>8 – 15%</td>
</tr>
</tbody>
</table>