Lecture 16: Object recognition: Part-based generative models

Professor Fei-Fei Li
Stanford Vision Lab
What we will learn today?

- Introduction
- Constellation model
  - Weakly supervised training
  - One-shot learning
- (Problem Set 4 (Q1))
Challenges: intra-class variation
Usual Challenges:

Variability due to:

• View point
• Illumination
• Occlusions
Basic issues

• **Representation**
  - 2D Bag of Words (BoW) models;
  - Part-based models;
  - Multi-view models;

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

• **Recognition**
  - Classification with BoW
  - Classification with Part-based models
Basic issues

• **Representation**
  – 2D Bag of Words (BoW) models;
  – Part-based models;
  – Multi-view models;

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

• **Recognition**
  – Classification with BoW
  – Classification with Part-based models
Basic issues

• Representation
  – 2D Bag of Words (BoW) models;
  – Part-based models;
  – Multi-view models (Lecture #19);

• Learning
  – Generative & Discriminative BoW models
  – Generative models
  – Probabilistic Hough voting

• Recognition
  – Classification with BoW
  – Classification with Part-based models
Problem with bag-of-words

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

<table>
<thead>
<tr>
<th>T. Leung</th>
<th>Representation</th>
<th>Shape statistics – F&amp;G '95</th>
<th>CVPR '96</th>
<th>ECCV '98</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Burl</td>
<td>Detection</td>
<td>Affine invariant shape – CVPR '98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| M. Weber | M. Welling | Unsupervised Learning | ECCV '00 | Multiple views - F&G '00 | Discovering categories - CVPR '00 |

<table>
<thead>
<tr>
<th>R. Fergus</th>
<th>Joint shape &amp; appearance learning</th>
<th>CVPR '03</th>
<th>Polluted datasets - ECCV '04</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Generic feature detectors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L. Fei-Fei</th>
<th>One-Shot Learning</th>
<th>ICCV '03</th>
<th>CVPR '04</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incremental learning</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Deformations

A

B

C

D

Fei-Fei Li
Presence / Absence of Features

Fei-Fei Li

Lecture 16 - 13  28-Nov-11
Background clutter
Generative probabilistic model

Foreground model (3-particle)

Gaussian shape pdf

Prob. of detection

Uniform shape pdf

# 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 16 - 15

28-Nov-11
Learning Models `Manually’

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

Goal: $\mu, \Sigma$
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:

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.

Image: label is given (i.e. a "face" image)

Object: no label is given (i.e. don't know where each part is)
(Semi) Unsupervised learning

- Know if image contains object or not
- But no segmentation of object or manual selection of features
Unsupervised detector training - 1

- 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

Fei-Fei Li  Lecture 16 - 23  28-Nov-11
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: $\mu$

$P = P \left( \left( \begin{array}{c} x_1 \\ y_2 \\ y_3 \end{array} \right) \right) \left( \begin{array} {c} \mu^* \\ \Sigma^* \end{array} \right)$

New estimate of $\mu$
Detector Selection

• Try out different combinations of detectors (Greedy search)

- Choice 1
  - Parameter Estimation
  - Model 1

- Choice 2
  - Parameter Estimation
  - 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

Model Foreground pdf

Sample Detection
Face images
Background images
Car from Rear

Test Error: 13% (5 Parts)

Preselected Parts

Sample Detection

Parts in Model

Model Foreground pdf
Detections of Cars
Background Images
3D Object recognition – Multiple mixture components
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

\[
\begin{pmatrix}
  x_1 \\
  y_1 \\
  x_2 \\
  y_2 \\
  x_3 \\
  y_3
\end{pmatrix}
\]

\[\mu_{\text{shape}} \]

\[\Sigma_{\text{shape}} \]

Gaussian part appearance pdf

\[
\begin{pmatrix}
  a_{i1} \\
  a_{i2} \\
  \vdots \\
  a_{i15}
\end{pmatrix}
\]

\[
\mu_{\text{part}} \]

\[
\Sigma_{\text{part}} \]

Gaussian relative scale pdf

\[\log(\text{scale})\]

Prob. of detection

0.8
0.75
0.9

Clutter model

Uniform shape pdf

Gaussian background appearance pdf

Uniform relative scale pdf

Poisson pdf on # detections

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

Samples from appearance model

Shape model

Fei-Fei Li
Recognized Motorbikes

Shape model

Part 1  Det: 5x10^-18
Part 2  Det: 6x10^-22
Part 3  Det: 6x10^-18
Part 4  Det: 5x10^-12
Part 5  Det: 5x10^-12
Background  Det: 5x10^-19
Background images evaluated with motorbike model
Frontal faces
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 work 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
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

Test
Train

Classification error (%) vs. log₂ (Training images)

Number of training examples

Previously

6 part Motorbike model

Classification error (%) vs. log₂ (Training images)
How do we do better than what statisticians have told us?

• Intuition 1: use Prior information

• Intuition 2: make best use of training information
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

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

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

Bayes Rule

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

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

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

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

Bayes Rule

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
model (θ) space

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

Each object model θ

Gaussian shape pdf

Gaussian part appearance pdf
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

Lecture 16 - 64 28-Nov-11
Experiments

Training:
1-6 randomly drawn images

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

Datasets

faces
airplanes
spotted cats
motorbikes
Experiments: obtaining priors

- airplanes
- spotted cats
- motorbikes
- faces

model (θ) space
Experiments: obtaining priors

- airplanes
- faces
- motorbikes
- spotted cats

Model (θ) space
Lecture 16 - Fei-Fei Li

Performance comparison

- Maximum-Likelihood
- Bayesian OneShot

Number of training examples

Correct

INCORRECT

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

- Maximum-Likelihood
- Bayesian OneShot

INCORRECT

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>
What we have learned today?

• Introduction
• Constellation model
  – Weakly supervised training
  – One-shot learning
• (Problem Set 4 (Q1))