Object Categorization: an overview & two models

Li Fei-Fei
Agenda

• Introduction to
  “Object Categorization”

• “Bag of Words” models

• Part-based models
object  

Perceptible  

Object  

Vision  

Material thing

1. Something visible or perceivable through one or more of the senses, especially sight or touch; an object of the senses.
2. A focus of attention, thinking, thought, or action: an object of continued policy.
3. The purpose or goal of a specific action or effort: the object of an investigation.
   a. A noun, pronoun, or noun phrase that receives or is affected by the action of a verb within a sentence.
   b. A noun or substantive governed by a preposition.
5. Philosophy. Something intangible or perceptible by the mind.
6. Computer Science. A discrete item that can be selected and maneuvered, such as an onscreen graphic. In object-oriented programming, objects include data and the procedures necessary to operate on that data.
Plato said...

• Ordinary objects are classified together if they `participate' in the same abstract Form, such as the Form of a Human or the Form of Quartz.

• Forms are proper subjects of philosophical investigation, for they have the highest degree of reality.

• Ordinary objects, such as humans, trees, and stones, have a lower degree of reality than the Forms.

• Fictions, shadows, and the like have a still lower degree of reality than ordinary objects and so are not proper subjects of philosophical enquiry.
How many object categories are there?

~10,000 to 30,000

Biederman 1987
So what does object recognition involve?
Identification: is that Potala Palace?
Verification: is that a lamp?
Detection: are there people?
Object categorization

- mountain
- tree
- building
- banner
- street lamp
- vendor
- people
Scene and context categorization

- outdoor
- city
- ...

[Image of a cityscape with a large building in the background and people walking on the streets]
Challenges 1: view point variation

Michelangelo 1475-1564
Challenges 2: illumination
Challenges 3: occlusion

Magritte, 1957
Challenges 4: scale
Challenges 5: deformation

Xu, Beihong 1943
Challenges 6: background clutter

Klimt, 1913
History: single object recognition
History: single object recognition

- Mahamud and Herbert, 2000
- Ferrari, Tuytelaars, and Van Gool, 2004
- Rothganger, Lazebnik, and Ponce, 2004
- Moreels and Perona, 2005
- ...
Challenges 7: intra-class variation
History: early object categorization
• Turk and Pentland, 1991
• Belhumeur, Hespanha, & Kriegman, 1997
• Schneiderman & Kanade 2004
• Viola and Jones, 2000

• Amit and Geman, 1999
• LeCun et al. 1998
• Belongie and Malik, 2002

• Schneiderman & Kanade, 2004
• Argawal and Roth, 2002
• Poggio et al. 1993
Object categorization: the statistical viewpoint

\[ p(\text{zebra} \mid \text{image}) \]

vs.

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

• 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
Object categorization: the statistical viewpoint

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

- Discriminative methods model posterior
- Generative methods model likelihood and prior
Discriminative

- Direct modeling of $\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})}$
Generative

- Model $p(\text{image} \mid \text{zebra})$ and $p(\text{image} \mid \text{no zebra})$

<table>
<thead>
<tr>
<th>$p(\text{image} \mid \text{zebra})$</th>
<th>$p(\text{image} \mid \text{no zebra})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Middle</td>
</tr>
<tr>
<td>High</td>
<td>Middle $\rightarrow$ Low</td>
</tr>
</tbody>
</table>
Three main issues

• Representation
  – How to represent an object category

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

• Recognition
  – How the classifier is to be used on novel data
Representation

- Generative / discriminative / hybrid
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
Representation

– Generative / discriminative / hybrid
– Appearance only or location and appearance
– Invariances
  • View point
  • Illumination
  • Occlusion
  • Scale
  • Deformation
  • Clutter
  • etc.
Representation

– Generative / discriminative / hybrid
– Appearance only or location and appearance
– Invariances
– Part-based or global w/sub-window
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
- Parts or global w/sub-window
- Use set of features or each pixel in image
Learning

Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning.
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- Methods of training: generative vs. discriminative

\[ p(x|C_1), p(x|C_2) \]

\[ p(C_1|x), p(C_2|x) \]
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

– Level of supervision
  • Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
– Level of supervision
  • Manual segmentation; bounding box; image labels; noisy labels
– Batch/incremental (on category and image level; user-feedback )
– Training images:
  • Issue of overfitting
  • Negative images for discriminative methods
  Priors
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning
– What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
– Level of supervision
  • Manual segmentation; bounding box; image labels; noisy labels
– Batch/incremental (on category and image level; user-feedback )
– Training images:
  • Issue of overfitting
  • Negative images for discriminative methods
– Priors
Recognition

– Scale / orientation range to search over
– Speed
– Context
(b) $P(\text{person}) = \text{uniform}$

(d) $P(\text{person} | \text{geometry})$

(f) $P(\text{person} | \text{viewpoint})$

(g) $P(\text{person} | \text{viewpoint, geometry})$
“Bag-of-words” models
• Definition of BoW
• System outline
• Various BoW models
  – Generative
  – Discriminative
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 the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the cerebral cortex, as a movie screen, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that the perception of the image is a more complex process. By following the visual impulses along their path to the various compartments of the visual cortex, Hubel and Wiesel 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 exports to $750bn, compared with a 18% rise in imports to $660bn. This may further annoy the US, which has long argued that China’s exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade 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.
A clarification: definition of “BoW”

- Looser definition
  - Independent features
A clarification: definition of “BoW”

- Looser definition
  - Independent features

- Stricter definition
  - Independent features
  - histogram representation
learning

- feature detection & representation
- image representation

recognition

- category models (and/or) classifiers
- codewords dictionary

decision

recognition
Representation

1. feature detection & representation

2. codewords dictionary

3. image representation
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

Vector quantization

Slide credit: Josef Sivic
2. Codewords dictionary formation

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

Sivic et al. 2005
3. Image representation
Representation

1. feature detection & representation

2. codewords dictionary

3. image representation
Learning and Recognition

codewords dictionary

category models (and/or) classifiers

category decision
Learning and Recognition

1. Generative method:
   - graphical models

2. Discriminative method:
   - SVM
Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)
Probabilistic Latent Semantic Analysis (pLSA)

d → z → w

D → N

“face”

Sivic et al. ICCV 2005
the pLSA model

\[ p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j) \]

Observed codeword distributions
Codeword distributions per theme (topic)
Theme distributions per image

Slide credit: Josef Sivic
Recognition using pLSA

\[ z^* = \arg \max_z p(z \mid d) \]
Learning the pLSA parameters

Maximize likelihood of data using EM

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i \mid d_j) n(w_i, d_j)$$

$$\sum_{k=1}^{K} P(z_k \mid d_j) P(w_i \mid z_k)$$

M … number of codewords

N … number of images

Observed counts of word $i$ in document $j$
the pLSA model
pLSA for human action classification

Niebles, Wang & Fei-Fei, BMVC 2006
pLSA for human action classification

Figure skating actions

Camel spin  Sit spin  Stand spin

Niebles, Wang & Fei-Fei, BMVC 2006
pLSA for human action classification

Niebles, Wang & Fei-Fei, BMVC 2006
the LDA model: natural scene categ.

Latent Dirichlet Allocation (LDA)

"beach"
the LDA model: natural scene categ.
the LDA model: natural scene categ.
z is fixed

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)

Hoffman, 2001

Blei et al., 2001
OPTIMOL: automatic Online Picture collection via Incremental Model Learning, by Li, Wang & Fei-Fei CVPR 2007
OPTIMOL: automatic Online Picture collection via Incremental Model Learning, by Li, Wang & Fei-Fei CVPR 2007

Dataset

Incremental learning

Category model

Classification

Downloaded Web images

Keyword: panda
OPTIMOL: automatic Online Picture collection via Incremental Model Learning, by Li, Wang & Fei-Fei CVPR 2007
1. Generative method:
   - graphical models

2. Discriminative method:
   - SVM
Discriminative methods based on ‘bag of words’ representation

Decision boundary

Zebra

Non-zebra
Discriminative methods based on ‘bag of words’ representation

- Grauman & Darrell, 2005, 2006:
  - SVM w/ Pyramid Match kernels
- Others
  - Csurka, Bray, Dance & Fan, 2004
  - Serre & Poggio, 2005
Summary: Pyramid match kernel

\[ K_\Delta (\Psi(X), \Psi(Y)) \]

optimal partial matching between sets of features

Grauman & Darrell, 2005, Slide credit: Kristen Grauman
Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

\[
\mathcal{I}(H(X), H(Y)) = \sum_{j=1}^{r} \min(H(X)_j, H(Y)_j)
\]

Slide credit: Kristen Grauman
Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

$$\mathcal{I} (H(X), H(Y)) = \sum_{j=1}^{r} \min (H(X)_j, H(Y)_j)$$

matches at this level

$$N_i = \mathcal{I} (H_i(X), H_i(Y)) - \mathcal{I} (H_{i-1}(X), H_{i-1}(Y))$$

matches at previous level

Difference in histogram intersections across levels counts *number of new pairs* matched

Slide credit: Kristen Grauman
Pyramid match kernel

\[
K_\Delta \left( \Psi(X), \Psi(Y) \right) = \sum_{i=0}^{L} \frac{1}{2^i} \left( I(H_i(X), H_i(Y)) - I(H_{i-1}(X), H_{i-1}(Y)) \right)
\]

- Weights inversely proportional to bin size
- Normalize kernel values to avoid favoring large sets

Slide credit: Kristen Grauman
Example pyramid match

Level 0

$X$

$Y$

$N_0 = 2$

$w_0 = 1$

$H_0(X)$

$H_0(Y)$

$I_0 = 2$

Slide credit: Kristen Grauman
Example pyramid match

Level 1

$X$

$Y$

$H_1(X)$

$H_1(Y)$

$I_1 = 4$

$N_1 = 4 - 2 = 2$

$w_1 = \frac{1}{2}$

Slide credit: Kristen Grauman
Example pyramid match

Level 2

$N_2 = 5 - 4 = 1$

$w_2 = \frac{1}{4}$

$H_2(X)$

$H_2(Y)$

$I_2 = 5$

Slide credit: Kristen Grauman
Example pyramid match

\[ K_\Delta = \sum_{i=0}^{L} w_i \, N_i \]

\[ = 1(2) + \frac{1}{2}(2) + \frac{1}{4}(1) = 3.25 \]

\[ K = \max_{\pi: X \to Y} \sum_{x_i \in X} S(x_i, \pi(x_i)) \]

\[ = 1(2) + \frac{1}{2}(3) = 3.5 \]

Slide credit: Kristen Grauman
Summary: Pyramid match kernel

$$K_{\Delta} (\Psi(X), \Psi(Y)) = \sum_{i=0}^{L} \frac{1}{2^i} \left( \mathcal{I}(H_i(X), H_i(Y)) - \mathcal{I}(H_{i-1}(X), H_{i-1}(Y)) \right)$$

difficulty of a match at level $i$  
number of new matches at level $i$

optimal partial matching between sets of features

Slide credit: Kristen Grauman
Learning

- Feature detection & representation
- Image representation
- Codewords dictionary
- Category models (and/or) classifiers

Recognition

- Category decision
Problem with bag-of-words

- All have equal probability for bag-of-words methods
- Location information is important
“Part-based models”
• Introduction
• An example: the constellation model
• Other models
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…
Model: Parts and Structure
The Constellation Model

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

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

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

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

L. Fei-Fei
- One-Shot Learning
- ICCV ’03
- Incremental learning
- CVPR ’04
Generative probabilistic model

**Foreground model**  
Gaussian shape pdf

**Prob. of detection**

Uniform shape pdf

**Clutter model**

# 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
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

\[
\begin{pmatrix}
0 & \ldots & N_1 \\
0 & \ldots & N_2 \\
0 & \ldots & N_3 \\
0 & \ldots & N_4
\end{pmatrix}
\]

\[h = \begin{pmatrix}
2 \\
3 \\
0 \\
2
\end{pmatrix}\]

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

3. Pick hypothesis which maximizes:

\[
\frac{p(\text{Data} | \text{Object}, \text{Hyp})}{p(\text{Data} | \text{Clutter}, \text{Hyp})}
\]

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, Perona, 1998, 2000a, b
(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
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
ML using EM

1. Current estimate

2. Assign probabilities to constellations

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

$\text{Large P} \times \text{new estimate of } \mu$ $\text{Small P} \times + \ldots = \text{new estimate of } \mu$
Detector Selection

• Try out different combinations of detectors (Greedy search)

Detectors ($\approx 100$)

Choice 1

$\square$ $\square$ $\square$ $\square$ $\square$

$\rightarrow$ Parameter Estimation

Model 1

$\square$

Choice 2

$\square$ $\square$ $\square$ $\square$ $\square$

$\rightarrow$ Parameter Estimation

Model 2

$\square$

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
Car from Rear

Preselected Parts

Test Error: 13% (5 Parts)

Parts in Model

Model Foreground pdf

Sample Detection
Detections of Cars
Background Images
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
Object categorization

Fergus, Perona, Zisserman, CVPR 2003
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 → \( \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{15} \end{pmatrix} \)

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
- Poisson pdf on # detections

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

Samples from appearance model

Shape model
Recognized Motorbikes
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
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
How many object categories are there?

~10,000 to 30,000

Biederman 1987
Savarese, 2003
One-Shot learning
Fei-Fei et. al.

ICCV ‘03
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Examples</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burl et al. Weber et al.</td>
<td>200 ~ 400</td>
<td>Faces, Motorbikes, Spotted cats, Airplanes, Cars</td>
</tr>
<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

Classification error (%)

\[
\log_2 (\text{Training images})
\]

Previously
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

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

Bayes Rule

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

Expansion by parametrization

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

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

**Bayes Rule**

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

**Expansion by parametrization**

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

**Previous Work:**

\[ \delta(\theta^{\text{ML}}) \]
Bayesian framework

P(object | test, train) vs. P(clutter | test, train)

Bayes Rule

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

Expansion by parametrization

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

One-Shot learning:

\[ p \left( \text{train} \mid \theta, \text{object} \right) p \left( \theta \right) \]
Model Structure

Each object model $\theta$

- Gaussian shape pdf
- Gaussian part appearance pdf

model ($\theta$) space

$\theta_n$ $\theta_2$ $\theta_1$
Each object model $\theta$

- Gaussian shape pdf
- Gaussian part appearance pdf

model distribution: $p(\theta)$

- conjugate distribution of $p(train|\theta, object)$
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
Experiments

Training:
1-6 randomly drawn images

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

Datasets

faces
airplanes
spotted cats
motorbikes

[www.vision.caltech.edu]
Experiments: obtaining priors

airplanes

spotted cats

motorbikes

model ($\theta$) space

faces

Miller, et al. ‘00
Experiments: obtaining priors

- Airplanes
- Faces
- Motorbikes
- Spotted cats
Performance comparison

Number of training examples

Performance (equal error rates)

Maximum-Likelihood
Bayesian OneShot

Correct
Correct
Correct
Correct

Shape Model (Training # = 1)
Performance comparison

- Maximum-Likelihood
- Bayesian OneShot

Number of training examples

Correct
INCORRECT
Correct
Correct

Shape Model (Training # = 1)
Performance comparison

Number of training examples

- Maximum-Likelihood
- Bayesian OneShot

Correct

Correct

INCORRECT

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>Burl, et al. Weber, et al.</td>
<td>200 ~ 400</td>
<td>Faces, Motorbikes, Spotted cats, Airplanes, Cars</td>
<td>5.6 - 10 %</td>
</tr>
<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>
Caltech 101 Dataset

Fei-Fei et al. 2004, 2006a, 2006b
Caltech 101 Dataset

Performance comparison for 101 categories

- Maximum Likelihood
- Maximum a posteriori
- Bayesian

Fei-Fei et al. 2004, 2006a, 2006b
Some class-specific graphs

• Articulated motion
  – People
  – Animals

• Special parameterisations
  – Limb angles

Images from [Kumar05, Felzenszwalb05]
Hierarchical representations

- Pixels → Pixel groupings → Parts → Object

- Multi-scale approach increases number of low-level features

- [Amit98]
- [Bouchard05]
Implicit shape model

- Use Hough space voting to find object
- Leibe and Schiele ’03,’05

Learning

- Learn appearance codebook
  - Cluster over interest points on training images

- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
  - Centroid is given

Recognition

**Interest Points**

**Matched Codebook Entries**

**Probabilistic Voting**

Spatial occurrence distributions
Deformable Template Matching

Berg et al. CVPR 2005

• Formulate problem as Integer Quadratic Programming
• $O(N^P)$ in general
• Use approximations that allow $P=50$ and $N=2550$ in <2 secs
Dense layout of parts

Layout CRF: Winn & Shotton, CVPR ‘06

Part labels (color-coded)
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. contours, intermediate objects
e.g. animals, trees, rocks
e.g. discontinuities, gradient
The correspondence problem

• Model with P parts
• Image with N possible locations for each part

• \(N^P\) combinations!!!
Different graph structures

- Fully connected
  - $O(N^6)$

- Star structure
  - $O(N^2)$

- Tree structure
  - $O(N^2)$

- Sparser graphs cannot capture all interactions between parts
Reconciliate BoW w/ Part-based Models

- Feature level
  - Spatial influence through correlogram features:
    Savarese, Winn and Criminisi, CVPR 2006
    (Wednesday, 3pm)
Reconciliate BoW w/ Part-based Models

- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Wang, Zhang & Fei-Fei, CVPR 2006
  - Niebles & Fei-Fei, CVPR 2007
Reconciliate BoW w/ Part-based Models

• Feature level

• Generative models
  – Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  – Wang, Zhang & Fei-Fei, CVPR 2006
  – Niebles & Fei-Fei, CVPR 2007
Reconciliate BoW w/ Part-based Models

- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Wang, Zhang & Fei-Fei, CVPR 2006
  - Niebles & Fei-Fei, CVPR 2007
Reconciliate BoW w/ Part-based Models

- Feature level
-Generative models
- Discriminative methods
  - Lazebnik, Schmid & Ponce, 2006
Reconciliate BoW w/ Part-based Models

- Feature level
- Generative models
- Discriminative methods
- 3D object categorization:
  - Savarese & Fei-Fei, (Wednesday, 3pm)
Reconciliate BoW w/ Part-based Models

- Feature level
- Generative models
- Discriminative methods
- 3D object categorization

- Still a wide open field of research
We have...

• Introduction to
  “Object Categorization”
  – Invariance issues
  – Representation, learning and recognition

• “Bag of Words” models
  – Generative models
  – Discriminative models

• Part-based models
  – Constellation model
  – Others
Online resources

Recognizing and Learning Object Categories

ICCV 2005 short courses

Li Fei-Fei (UIUC), Rob Fergus (Oxford-MIT), Antonio Torralba (MIT)

Slides

Matlab code

This set of three demos illustrates the concepts behind several approaches for object recognition. The code consists of Matlab scripts (which should run under both Windows and Linux). The code is for teaching/research purposes only.

Datasets

These are pointers to the datasets used in the demos:

- Caltech datasets
- LabelMe dataset and annotation tool

Other datasets:

http://vision.cs.princeton.edu/
Assisted technology

Pedestrian and car detection

Lane detection

Alan Yuille’s stuff
Security
Improving online search

Query: STREET

Organizing photo collections
Computational photography

[Face priority AE] When a bright part of the face is too bright
Toys and robots
Digit recognition, AT&T labs
http://www.research.att.com/~yann/

LeCun, et al. 1998
Medical Imaging
Thank you!

Postdoc opening: Fall 2007