Generative Models for Visual Objects and Object Recognition via Bayesian Inference

Li Fei-Fei
Beckman Institute
outline

• The question of object categorization
• Brief overview
  – Generative
  – Discriminative
• Generative models
  – Bag of words
  – Constellation
  – Others
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.
Bruegel, 1564
How many object categories are there?

10,000 to 30,000

Biederman 1987
So what does object recognition involve?
Verification: is that a bus?
Detection: are there cars?
Identification: is that a picture of Mao?
Object categorization

- sky
- building
- flag
- banner
- face
- street lamp
- wall
- bus
- cars
Scene and context categorization

- outdoor
- city
- traffic
- ...

[Image of a cityscape with traditional Chinese architecture and traffic]
Challenges 1: view point variation

Michelangelo 1475-1564
Challenges 2: illumination

slide credit: S. Ullman
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 et al. 1997
• Schneiderman et al. 2004
• Viola and Jones, 2000

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

• Schneiderman et al. 2004
• Argawal and Roth, 2002
• Poggio et al. 1993
~10,000 to 30,000
Scenes, Objects, and Parts

Object categorization: the statistical viewpoint

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

- 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

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

- Posterior ratio
- Likelihood ratio
- Prior ratio

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

• Direct modeling of \( \frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} \)

Decision boundary

Zebra

Non-zebra

![Image of Okapi](image1.png)

![Image of Motorcycle](image2.png)

![Image of Zebra](image3.png)
**Generative**

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

<table>
<thead>
<tr>
<th>$p(image \mid zebra)$</th>
<th>$p(image \mid no \ zebra)$</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
Bag-of-words models
Related works

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

• Hierarchical Bayesian models for documents (pLSA, LDA, etc.)
  – Hoffman 1999; Blei et al, 2004; Teh et al. 2004

• Object categorization
  – Dorko et al. 2004; Csurka et al. 2003; Sivic et al. 2005; Sudderth et al. 2005;

• Natural scene categorization
  – Fei-Fei et al. 2005
Object → Bag of ‘words’
Analogy to documents

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 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, but now we know 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. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes 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% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. These figures are likely to 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.
learning

feature detection & representation

image representation

recognition

codewords dictionary

category models (and/or) classifiers

category decision
Representation

1. feature detection & representation

2. codewords dictionary

3. image representation
1. Feature detection and representation
1. Feature detection and representation

- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005
1. Feature detection and representation

- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005

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

- Regular grid
  - Vogel et al. 2003
  - Fei-Fei et al. 2005

- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei et al. 2005
  - Sivic et al. 2005

- Other methods
  - Random sampling (Ullman et al. 2002)
  - Segmentation based patches (Barnard et al. 2003)
1. Feature detection and representation

- Compute SIFT descriptor [Lowe'99]
- Normalize patch
- Detect patches [Mikojaczyk and Schmid '02]
  [Matas et al. '02]
  [Sivic et al. '03]

Slide credit: Josef Sivic
1. Feature detection and representation
2. Codewords dictionary formation
2. Codewords dictionary formation
2. Codewords dictionary formation

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

Sivic et al. 2005
3. Image representation

![Car image]

**Bar Chart:**
- **X-axis:** Codewords
- **Y-axis:** Frequency

The bar chart illustrates the frequency distribution of codewords in an image. The rightmost bar shows a significantly higher frequency compared to the others.
1. feature detection & representation

2. codewords dictionary

3. image representation
Learning and Recognition

codewords dictionary

category models (and/or) classifiers

category decision
2 case studies

1. Naïve Bayes classifier
   - Csurka et al. 2004

2. Hierarchical Bayesian text models (pLSA and LDA)
   - Background: Hoffman 2001, Blei et al. 2004
   - Natural scene categorization: Fei-Fei et al. 2005
First, some notations

- \( w_n \): each patch in an image
  - \( w_n = [0,0,...1,...,0,0]^T \)
- \( w \): a collection of all \( N \) patches in an image
  - \( w = [w_1,w_2,...,w_N] \)
- \( d_j \): the \( j^{th} \) image in an image collection
- \( c \): category of the image
- \( z \): theme or topic of the patch
Case #1: 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

Prior prob. of the object classes

Image likelihood given the class

Csurka et al. 2004
Our in-house database contains 1776 images in seven classes\textsuperscript{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>

Csurka et al. 2004
Case #2: Hierarchical Bayesian
text models

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)

Hoffman, 2001

Blei et al., 2001
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

"face"

Sivic et al. ICCV 2005
Case #2: Hierarchical Bayesian text models

Latent Dirichlet Allocation (LDA)

Fei-Fei et al. ICCV 2005
Case #2: the pLSA model
Case #2: 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
Case #2: Recognition using pLSA

\[ z^* = \arg \max_z p(z | d) \]
Case #2: Learning the pLSA parameters

Maximize likelihood of data using EM

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

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

Observed counts of word $i$ in document $j$

M … number of codewords
N … number of images

Slide credit: Josef Sivic
Learning and Recognition

codewords dictionary

category models (and/or) classifiers

category decision
Invariance issues

• Scale and rotation
  – Implicit
  – Detectors and descriptors

Kadir and Brady. 2003
Invariance issues

- Scale and rotation
- Occlusion
  - Implicit in the models
  - Codeword distribution: small variations
  - (In theory) Theme (z) distribution: different occlusion patterns
Invariance issues

• Scale and rotation
• Occlusion
• Translation
  – Encode (relative) location information
Invariance issues

• Scale and rotation
• Occlusion
• Translation

• View point (in theory)
  – Codewords: detector and descriptor
  – Theme distributions: different view points

Fergus et al. 2005
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 from our eyes. For a long time it was believed that the retinal image was transmitted point by point to visual centers in the brain. This view is like a movie screen on which the image is projected. Through the discoveries of Hubel and Wiesel we now know that the perception of visual images is a more complex process. By following the path of images to the various cell layers of the optical 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.
• Intuitive
• (Could use) generative models
  – Convenient for weakly- or un-supervised training
  – Prior information
  – Hierarchical Bayesian framework

Sivic et al., 2005, Sudderth et al., 2005
Model properties

- Intuitive
- (Could use) generative models
- Learning and recognition relatively fast
  - Compare to other methods
Weakness of the model

- 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
part-based models

Slides courtesy to Rob Fergus for “part-based models”
Problem with bag-of-words

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

- Representation
  - Computational complexity
  - Design choices

- Recognition
  - Demos

- Learning
  - Automated methods
Representation
Model: Parts and Structure
Representation

• Object as set of parts
  – Generative representation

• Model:
  – Relative locations between parts
  – Appearance of part

• Issues:
  – How to model location
  – How to represent appearance
  – Sparse or dense (pixels or regions)
  – How to handle occlusion/clutter

Figure from [Fischler73]
Example scheme

• Model shape using Gaussian distribution on location between parts
• Model appearance as pixel templates
• Represent image as collection of regions
  – Extracted by template matching: normalized-cross correlation

• Manually trained model
  – Click on training images
Sparse representation

+ Computationally tractable (\(10^5\) pixels \(\rightarrow\) \(10^1\) -- \(10^2\) parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition

- Throw away most image information
- Parts need to be distinctive to separate from other classes
History of Idea

- 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
- Felzenszwalb & Huttenlocher ’00
- Many papers since 2000
The correspondence problem

- Model with $P$ parts
- Image with $N$ possible locations for each part

- $N^P$ combinations!!!
Connectivity of parts

- Complexity is given by size of maximal clique in graph
- Consider a 3 part model
  - Each part has set of N possible locations in image
  - Location of parts 2 & 3 is independent, given location of L
  - Each part has an appearance term, independent between parts.
Connectivity of parts

• To find best match in image, we want most probable state of $L$, $L^*$

• Run max-product message passing

$$m_a(L) = S(L)$$
$$m_b(L) = \max_2 S(L,2)A(2)$$
$$m_c(L) = \max_3 S(L,3)A(3)$$
$$m_d(L) = A(L)$$

Take $O(N^2)$ to compute:
For each of the $N$ values of $L$, need to find max over $N$ states

$L^* = \max_L (m_a(L)m_b(L)m_c(L)m_d(L))$
Different graph structures

- Fully connected
- Star structure
- Tree structure

- $O(N^6)$
- $O(N^2)$
- $O(N^2)$

- Sparser graphs cannot capture all interactions between parts
Regions or pixels

- # Regions << # Pixels
- Regions increase tractability but lose information
- Generally use regions:
  - Local maxima of interest operators
  - Can give scale/orientation invariance

Figures from [Kadir04]
How to model location?

• Explicit: Probability density functions
• Implicit: Voting scheme

• Invariance
  – Translation
  – Scaling
  – Similarity/affine
  – Viewpoint
Explicit shape model

• Probability densities
  – Continuous (Gaussians)
  – Analogy with springs

• Parameters of model, $\mu$ and $\Sigma$
  – Independence corresponds to zeros in $\Sigma$

$$\mu = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix} \quad \Sigma = \begin{pmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1y_1 & x_1y_2 & x_1y_3 \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2y_1 & x_2y_2 & x_2y_3 \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3y_1 & x_3y_2 & x_3y_3 \\ y_1x_1 & y_1x_2 & y_1x_3 & y_1y_1 & y_1y_2 & y_1y_3 \\ y_2x_1 & y_2x_2 & y_2x_3 & y_2y_1 & y_2y_2 & y_2y_3 \\ y_3x_1 & y_3x_2 & y_3x_3 & y_3y_1 & y_3y_2 & y_3y_3 \end{pmatrix}$$
Shape

• Shape is “what remains after differences due to translation, rotation, and scale have been factored out”. [Kendall84]

\[ p(x_1, \ldots, x_P, y_1, \ldots, y_P) = p_{\text{Pose}}(t_x, t_y, s, \theta) \cdot p_{\text{Shape}}(u_3, \ldots, u_P, v_1, \ldots, v_P) \]

\[ X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_N \end{bmatrix} \]

\[ U = \begin{bmatrix} u_3 \\ \vdots \\ x_N \\ v_3 \\ \vdots \\ v_N \end{bmatrix} \]

• Statistical theory of shape [Kendall, Bookstein, Mardia & Dryden]

Figures from [Leung98]
Euclidean & Affine Shape

- Translation, rotation and scaling $\Rightarrow$ Euclidean Shape
- Removal of camera foreshortenings $\Rightarrow$ Affine Shape

Assume Gaussian density in figure space.

What is the probability density for the shape variables in each of the different spaces?

Figures from [Leung98]
Translation-invariant shape

- Figure space density:
  \[ x = [x_1, \ldots, x_P, y_1, \ldots, y_P]^T \quad p(x) = \mathcal{N}(x|\mu, \Sigma) \]

- Translation-invariant form
  \[ x^* = Lx \quad \text{e.g. } P=3, \text{ move } 1^{\text{st}} \text{ part to origin} \]
  \[ L = \begin{pmatrix}
    0 & 0 & 0 & 0 & 0 & 0 \\
    -1 & 1 & 0 & 0 & 0 & 0 \\
    -1 & 0 & 1 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & -1 & 1 & 0 \\
    0 & 0 & 0 & -1 & 0 & 1 
  \end{pmatrix} \]
  \[ x^* = [0, x_2-x_1, \ldots, x_P-x_1, 0, y_2-y_1, \ldots, y_P-y_1] \]

- Shape space density is still Gaussian
  \[ p(x)^* = \mathcal{N}(x^*|\mu^*, \Sigma^*) \]
  \[ \mu^* = L\mu \quad \Sigma^* = L\Sigma L^T \]
Affine Shape Density

- Affine Shape density (Dryden-Mardia):

\[ p_u(U) = \frac{(N-3)!e^{-g/2}}{\pi^{(N-3)}} \sqrt{\prod_{k=1}^{4} \lambda_{ki}^k L_{ki}} \sum_{i=1}^{4} \lambda_{ki}^k \left\{-\frac{\rho_i^2}{2}\right\} \]

- Euclidean Shape density is of similar form

- Can learnt parameters of DM density with EM!

[Leung98],[Welling05]
Other invariance methods

- Search over transformations
  - Large space (# pixels x # scales ....)
  - Closed form solution for translation and scale (Helmer and Lowe '04)

- Features give information
  - Characteristic scale
  - Characteristic orientation (noisy)

Figures from Mikolajczyk & Schmid
Representation of appearance

• Dependencies between parts
  – Common to assume independence
  – Need not be
  – Symmetry

• Needs to handle intra-class variation
  – Task is no longer matching of descriptors
  – Implicit variation (VQ appearance)
  – Explicit probabilistic model of appearance (e.g. Gaussians in SIFT space or PCA space)
Representation of appearance

• Invariance needs to match that of shape model

• Insensitive to small shifts in translation/scale
  – Compensate for jitter of features
  – e.g. SIFT

• Illumination invariance
  – Normalize out
  – Condition on illumination of landmark part
Representation of occlusion

- Explicit
  - Additional match of each part to missing state

- Implicit
  - Truncated minimum probability of appearance

![Graph showing log probability against Appearance space with a peak at \( \mu_{\text{part}} \)]
Representation of background clutter

- **Explicit model**
  - Generative model for clutter as well as foreground object

- **Use a sub-window**
  - At correct position, no clutter is present
Learning
Learning situations

• Varying levels of supervision
  – Unsupervised
  – Image labels
  – Object centroid/bounding box
  – Segmented object
  – Manual correspondence
    (typically sub-optimal)

• Generative models naturally incorporate labelling information (or lack of it)

• Discriminative schemes require labels for all data points
Learning using EM

• Task: Estimation of model parameters
• Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to parts
• Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters
Example scheme, using EM for maximum likelihood learning

1. Current estimate of $\theta$

2. Assign probabilities to constellations

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

$$\text{Large P} \quad x \quad \text{Small P} \quad x \quad + \quad \ldots \quad = \quad \text{new estimate of } \mu$$
Priors

• Implicit
  – Structure of dependencies in model
  – Parameterisation of model
  – Feature detectors

• Explicit
  – $p(\theta)$
  – MAP / Bayesian learning
  – Fei-Fei ‘03
Learning Shape & Appearance simultaneously

Fergus et al. '03
Learn appearance then shape

Weber et al. ‘00

Preselected Parts (≈100)

Choice 1

Parameter Estimation

Model 1

Choice 2

Parameter Estimation

Model 2

Predict / measure model performance
(validation set or directly from model)
Discriminative training

- Sparse so parts need to be distinctive of class

- Boosted parts and structure models
  - Amores et al. CVPR 2005
  - Bar Hillel et al. CVPR 2005

- Discriminative features
  - Weber et al. 2000
  - Ullman et al.

- Train discriminatively on parameters of generative model
  - Holub, Welling, Perona ICCV 2005
Number of training images

• More supervision, fewer images needed
  – Few unknown parameters

• Less supervision, more images.
  – Lots of unknown parameters

• Over-fitting problems
Number of training examples

Generalisation performance

6 part Motorbike model

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

Priors
Parts and Structure models

Summary

• Correspondence problem
• Efficient methods for large # parts and # positions in image
• Challenge to get representation with desired invariance
• Minimal supervision

Future directions:
• Multiple views
• Approaches to learning
• Multiple category training
Recognition
What task?

• Classification
  – Object present/absent
  – Sum over all matches (Bayesian)
  – Take best

• Detection
  – Localize object within the frame
  – Slide sub-window across image
  – Use features to define a basis
Efficient search methods

• Interpretation tree (Grimson ’87)
  – Condition on assigned parts to give search regions for remaining ones
  – Branch & bound, A*
Parts and Structure demo

- Gaussian location model – star configuration
- Translation invariant only
  - Use 1st part as landmark
- Appearance model is template matching
- Manual training
  - User identifies correspondence on training images
- Recognition
  - Run template for each part over image
  - Get local maxima → set of possible locations for each part
  - Impose shape model - O(N^2P) cost
  - Score of each match is combination of shape model and template responses.
Demo images

- Sub-set of Caltech face dataset
- Caltech background images
A simple parts and structure object detector

ICCV 2005 short courses on
Recognizing and Learning Object Categories

An intuitive way to represent objects is as a collection of distinctive parts. Such a release model both the relative positions of the parts as well as their appearance, giving a sparse representation that captures the essence of the object.

This simple demo illustrates the concepts behind many such "parts and structure" approaches. For simplicity, training is manually guided with the user hand clicking on the distinctive parts of a few training images. A simple model is then built for use in recognition. Two different recognition approaches are provided: one relying on feature points [1]; the other using the efficient methods of Fehnmeier and Fattal [2].

The code consists of Matlab scripts (which should run under both Windows and Linux). The Image Processing Toolbox is required. The code is for teaching/research purposes only. If you find a bug, please email me at [email protected] point net point edu.

Download

Download the code and dataset (24 MBytes).

Operation of code

To run the demo:
1. Throsh the zip file into a new directory (e.g., mywork/myproject/demos).
Demo (2)
Demo (3)
Demo (4)
How much does shape help?

- Crandall, Felzenszwalb, Huttenlocher CVPR’05
- Shape variance increases with increasing model complexity

(a) Airplane, 1-fan

(b) Airplane, 2-fan
combined segmentation and recognition
Aim

- Given an image and object category, to segment the object

Segmentation should (ideally) be
- shaped like the object e.g. cow-like
- obtained efficiently in an unsupervised manner
- able to handle self-occlusion
In this section: brief paper reviews

• Jigsaw approach: Borenstein & Ullman, 2001, 2002
• Concurrent recognition and segmentation: Yu and Shi, 2002
• Image parsing: Tu et al. 2003
• Interleaved segmentation: Liebe & Schiele, 2004, 2005
• OBJCUT: Kumar et al. 2005
• LOCUS: Winn and Jojic, 2005
Image parsing: Tu, Zhu and Yuille 2003
Image parsing: Tu, Zhu and Yuille 2003

a. Input image  
b. Segmentation  
c. Object recognition  
d. Synthesized image
OBJCUT:
shape prior -- Layered Pictorial Structures (LPS)

• Generative model
• Composition of parts + spatial layout

Parts in Layer 2 can occlude parts in Layer 1

Kumar, et al. 2004, 2005
Probability of labelling in addition has
• Unary potential which depend on distance from $\Theta$ (shape parameter)

Kumar, et al. 2004, 2005
OBJCUT: Results

Using LPS Model for Cow

In the absence of a clear boundary between object and background
LOCUS model

Class shape \( \pi \)

Class edge sprite \( \mu^o, \sigma^o \)

Shared between images

Deformation field \( D \)

Position & size \( T \)

Different for each image

Mask \( m \)

Background appearance \( \lambda^0 \)

Image

Object appearance \( \lambda^1 \)

Edge image \( e \)

Winn and Jojic, 2005
remarks

• Strength
  – Explains every pixel of the image
  – Useful for image editing, layering, etc.

• Issues
  – Invariance issues
    • (especially) scale, view-point variations
Summary

• Methods reviewed here
  – Bag of words
  – Parts and structure
  – Discriminative methods
  – Combined Segmentation and recognition

• Resources online
  – Slides
  – Code
  – Links to datasets

http://people.csail.mit.edu/torralba/iccv2005/
List properties of ideal recognition system

• Representation
  – 1000’s categories,
  – Handle all invariances (occlusions, view point, …)
  – Explain as many pixels as possible (or answer as many questions as you can about the object)
  – fast, robust

• Learning
  – Handle all degrees of supervision
  – Incremental learning
  – Few training images

• …
Recognizing and Learning Object Categories

ICCV 2005 short courses

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

Slides

- Talks

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://people.csail.mit.edu/torralba/iccv2005/