EECS 598 – Special Topics in Computer vision

Intro to Object Categorization

- Introduction
- Bag of world models

This segment is adapted from the tutorial “Recognizing and Learning Object Categories: Year 2007”, by Prof A. Torralba, R. Fergus and F. Li
How many object categories are there?

~10,000 to 30,000

Biederman 1987
~10,000 to 30,000
Challenges: viewpoint variation
Challenges: illumination

image credit: J. Koenderink
Challenges: scale
Challenges: deformation
Challenges:
occlusion

Magritte, 1957
Challenges: background clutter

Kilmeny Niland. 1995
Challenges: intra-class variation
No intra-class variation: single object recognition
So what does object recognition involve?
Identification: is that Potala Palace?
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 people walking in the foreground and a mountain in the background]
Other recognition problems

• Action recognition
• Event recognition
Some early works on 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
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

- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.
Representation

- Use set of features or each pixel in image
Representation

Bag of features, part based, global

Global representation

Bag of features/parts representations

Constellation of features/parts
Representation

A statistical view point
- generative models
- discriminative models
Object categorization: the statistical viewpoint

\[ p(zebra \mid image) \]

vs.

\[ p(no\ zebras \mid image) \]

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

- Bayes rule:

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}
\]
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**: Discriminative methods model posterior
- **Likelihood ratio**: Generative methods model likelihood and prior
- **Prior ratio**
Discriminative

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

Decision boundary

- Zebra
- Non-zebra

Images:
- Left: Okapi (non-zebra)
- Middle: Motorcycle (non-zebra)
- Right: Zebra
Discriminative
Generative

\( p(\text{image} \mid \text{zebra}) \quad p(\text{image} \mid \text{no zebra}) \)

\begin{tabular}{|c|c|}
\hline
\( p(\text{image} \mid \text{zebra}) \) & \( p(\text{image} \mid \text{no zebra}) \) \\
\hline
Low & Middle \\
\hline
High & Middle \\
\hline
\end{tabular}
Generative

$p(x|C_1)$

$p(x|C_2)$
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
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
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

- machine learning useful to model intraclass variability
Learning

- Machine learning useful to model intra-class variability
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
Learning

- Machine learning useful to model intra-class variability
  - 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

- Machine learning useful to model intra-class variability
  - 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

Contains a motorbike
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
Recognition

– Scale / orientation range to search over
– Speed
– Context
Context

(b) $P(\text{person}) = \text{uniform}$

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

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

(g) $P(\text{person} \mid \text{viewpoint, geometry})$
Applications: Assisted driving

Pedestrian and car detection

Lane detection

- Collision warning systems with adaptive cruise control,
- Lane departure warning systems,
- Rear object detection systems,
Computational photography

[Face priority AE] When a bright part of the face is too bright
Improving online search

Query: STREET

Organizing photo collections
Applications of computer vision

- Factory inspection
- Assistive technologies
- Surveillance
- Autonomous driving, robot navigation
- Security
Intro to Object Recognition

• Introduction
• Bag of world models

This segment is based on the tutorial “Recognizing and Learning Object Categories: Year 2007”, by Prof A. Torralba, R. Fergus and F. Li
Bag-of-words models

This segment is adapted from the tutorial “Recognizing and Learning Object Categories: Year 2007”, by Prof A. Torralba, R. Fergus and F. Li
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; Grauman & Darrell 2005

• Natural scene categorization
  – Vogel & Schiele, 2004; Fei-Fei & Perona, 2005; Bosch, Zisserman & Munoz, 2006; Lazebnik et al, 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 brain as a movie screen. Through the discoveries of Hubel and Wiesel we now know that the perception of an image is a more complicated process. 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 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% jump in exports to $750bn, compared with an 18% rise in imports to $660bn. This is likely to annoy the US, which has long argued that China’s exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees that 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

codewords dictionary

category models (and/or) classifiers

category decision
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]

Slide credit: Josef Sivic
1. Feature detection and representation
Example: color feature

Source: K. Grauman
Example: color feature
Example: color feature
2. Codewords dictionary formation
2. Codewords dictionary formation

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. Image representation

![Diagram showing frequency of codewords with examples of codewords represented as bar graphs and images.](image.png)
Representation

1. feature detection & representation

2. codewords dictionary

3. category models
Learning & Recognition

1. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models

category models (and/or) classifiers
Discriminative classifiers

category models

Class 1

Class N
Discriminative classifiers

Query image

Winning class: pink
Nearest Neighbors classifier

Query image

Winning class: pink

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

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

Source: D. Lowe
K-Nearest Neighbors

- 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

\[ k = 5 \]

Source: D. Lowe
Linear classifiers

• Find linear function (hyperplane) to separate positive and negative examples

\[ \mathbf{x}_i \text{ positive: } \mathbf{x}_i \cdot \mathbf{w} + b \geq 0 \]
\[ \mathbf{x}_i \text{ negative: } \mathbf{x}_i \cdot \mathbf{w} + b < 0 \]
Support vector machines

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

Support vector machines

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

\[ x_i \text{ positive } (y_i = 1): \quad x_i \cdot w + b \geq 1 \]
\[ x_i \text{ negative } (y_i = -1): \quad x_i \cdot w + b \leq -1 \]

For support vectors, \( x_i \cdot w + b = \pm 1 \)

Distance between point and hyperplane:

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

Therefore, the margin is \( 2 / \|w\| \)

Credit slide: S. Lazebnik
Functions for comparing histograms

• **L1 distance**

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

• **\( \chi^2 \) distance**

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

• **Quadratic distance (cross-bin)**

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

Jan Puzicha, Yossi Rubner, Carlo Tomasi, Joachim M. Buhmann: Empirical Evaluation of Dissimilarity Measures for Color and Texture. ICCV 1999
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

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

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

Graphical model

$$p(c \mid w) \sim p(c)p(w \mid c)$$

Prior prob. of the object classes
Image likelihood given the class

Csurka et al. 2004
The Naïve Bayes model

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

- Assume that each feature is conditionally independent *given the class*

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

Csurka et al. 2004
the Naïve Bayes model

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

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\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>75</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></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
Summary: Generative models

• Naïve Bayes
  – *Unigram models* in document analysis
  – Assumes conditional independence of words given class
  – Parameter estimation: frequency counting
EECS 598 – Special Topics in Computer vision

Next lecture

- SVM classification using pyramid matching kernels

This segment is adapted from the tutorial “Recognizing and Learning Object Categories: Year 2007”, by Prof A. Torralba, R. Fergus and F. Li