Lecture 14: Introduction to Object Recognition & Bag-of-Words (BoW) Models

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

• Introduction to object recognition
  – Representation
  – Learning
  – Recognition

• Bag of Words models *(Problem Set 4 (Q2))*
  – Basic representation
  – Different learning and recognition algorithms
What are the different visual recognition tasks?
Classification:
Does this image contain a building? [yes/no]

Yes!
Classification:
Is this an beach?
Image Search

Organizing photo collections
Detection:
Does this image contain a car? [where?]
Detection:
Which object does this image contain? [where?]
Detection:
Accurate localization (segmentation)
Detection: Estimating object semantic & geometric attributes

Object: Person, back; 1-2 meters away
Object: Police car, side view, 4-5 m away
Object: Building, 45º pose, 8-10 meters away
It has bricks
Applications of computer vision

- Computational photography
- Assistive technologies
- Surveillance
- Security
- Assistive driving
Categorization vs Single instance recognition

Does this image contain the Chicago Macy building’s?
Categorization vs Single instance recognition

Where is the crunchy nut?
Applications of computer vision

• Recognizing landmarks in mobile platforms
Activity or Event recognition
What are these people doing?
Visual Recognition

• Design algorithms that are capable to
  – Classify images or videos
  – Detect and localize objects
  – Estimate semantic and geometrical attributes
  – Classify human activities and events

Why is this challenging?
How many object categories are there?

~10,000 to 30,000
Challenges: viewpoint variation

Michelangelo 1475-1564
Challenges: illumination

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

Magritte, 1957
Challenges: background clutter

Kilmeny Niland. 1995
Challenges: intra-class variation
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
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
Representation
- Building blocks: Sampling strategies

Interest operators
Dense, uniformly
Multiple interest operators
Randomly

Image credits: Fei-Fei, E. Nowak, J. Sivic
Representation
– Appearance only or location and appearance
Representation

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

– To handle intra-class variability, it is convenient to describe an object categories using probabilistic models
– Object models: Generative vs Discriminative vs hybrid
Object categorization: the statistical viewpoint

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

vs.

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

- Bayes rule:

\[
P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}
\]

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})}
\]
Object categorization: the statistical viewpoint

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

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

• Bayes rule:

\[ P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)} \]

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

posterior ratio

likelihood ratio

prior ratio
Object categorization: 
the statistical viewpoint

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

• 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
Discriminative models

- Modeling the posterior ratio:
  \[
  \frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})}
  \]
Discriminative models

**Nearest neighbor**
- 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,...
Generative models

- Modeling the likelihood ratio:

\[
p(image \mid zebra) \over p(image \mid no \ zebra)
\]
Generative models

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

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(image \mid C_1)</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>p(image \mid C_2)</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

\[ p(x \mid C_1) \quad p(x \mid C_2) \]

Lecture 14  8-Nov-11
Generative models

- Naïve Bayes classifier
  - Csurka Bray, Dance & Fan, 2004

- Hierarchical Bayesian topic models (e.g. pLSA and LDA)
  - Natural scene categorization: Fei-Fei et al. 2005

- 2D Part based models
  - Constellation models: Weber et al 2000; Fergus et al 2000
  - Star models: ISM (Leibe et al 05)

- 3D part based models:
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
Learning

• Learning parameters: What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
Learning

• Learning parameters: 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

• Priors
Learning

• Learning parameters: 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

• Priors

• Training images:
  • Issue of overfitting
  • Negative images for discriminative methods
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
Recognition

– Recognition task: classification, detection, etc..
Recognition

– Recognition task

– Search strategy: Sliding Windows
  
  • Simple
  
  • Computational complexity \((x, y, S, \theta, N\) of classes)

  - BSW by Lampert et al 08
  - Also, Alexe, et al 10

Viola, Jones 2001,
Recognition

– Recognition task

– Search strategy: Sliding Windows

  • Simple
  • Computational complexity \((x, y, S, \theta, N \text{ of classes})\)

- BSW by Lampert et al 08
- Also, Alexe, et al 10

• Localization
  • Objects are not boxes

Viola, Jones 2001,
Recognition

– Recognition task

– Search strategy: Sliding Windows
  • Simple
  • Computational complexity \((x, y, S, \theta, N\) of classes)
    - BSW by Lampert et al 08
    - Also, Alexe, et al 10
  • Localization
    • Objects are not boxes
    • Prone to false positive

Non max suppression:
Canny ’86
....
Desai et al , 2009
Recognition

- Recognition task
- Search strategy
- Attributes

- It has metal
- It is glossy
- Has wheels

- Savarese, 2007
- Sun et al. 2009
- Liebelt et al., ’08, 10
- Farhadi et al 09

Category: car
Azimuth = 225º
Zenith = 30º

- Farhadi et al 09
- Lampert et al 09
- Wang & Forsyth 09
Recognition

– Recognition task
– Search strategy
– Attributes
– Context

Semantic:
• Torralba et al 03
• Rabinovich et al 07
• Gupta & Davis 08
• Heitz & Koller 08
• L-J Li et al 08
• Yao & Fei-Fei 10

Geometric
• Hoiem, et al 06
• Gould et al 09
• Bao, Sun, Savarese 10
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
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 the cortex through our eyes. For a long time, scientists have been searching for a retinal image which is analogous to a movie screen, where the visual centers in the brain resemble the movie screen. However, it has been discovered that our perception is actually a lot more complex than this. Following 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% rise in exports to $750bn, compared with $560bn in 2004, and a 21% rise in imports to $660bn. This move 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 it also needs the yuan to be able to compete with foreign currency. China has already raised 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.
definition of “BoW”

– Independent features

face

bike

violin
definition of “BoW”

– Independent features
– histogram representation

codewords dictionary
Representation

- feature detection & representation
- image representation

Recognition

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

- **Compute SIFT descriptor**
  - [Lowe'99]

- **Normalize patch**

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

Slide credit: Josef Sivic
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

- Frequency distribution of codewords

- Visual representation of codewords
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. Discriminative method:
   - NN
   - SVM

2. Generative method:
   - graphical models
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

- 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-Nearest Neighbors classifier

• Voronoi partitioning of feature space for 2-category 2-D and 3-D data
• For k dimensions: k-D tree = space-partitioning data structure for organizing points in a k-dimensional space
• Enable efficient search
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)**

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

Learning and Recognition

1. Discriminative method:
   - NN
   - SVM

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

category models

Model space

Class 1

Class N
Support vector machines

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

Support vectors: \[ x_i \cdot w + b = \pm 1 \]

Distance between point and hyperplane:

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

Margin = \[ 2 / ||w|| \]

Solution:

\[ w = \sum_i \alpha_i y_i x_i \]

Classification function (decision boundary):

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

Credit slide: S. Lazebnik
Support vector machines

• Classification

$$\mathbf{w} \cdot \mathbf{x} + b = \sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b$$

Test point

- If $\mathbf{x} \cdot \mathbf{w} + b \geq 0 \rightarrow \text{class 1}$
- If $\mathbf{x} \cdot \mathbf{w} + b < 0 \rightarrow \text{class 2}$

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:
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>]</td>
<td>( O(dm^2) )</td>
</tr>
<tr>
<td>Bhattacharyya affinity [Kondor &amp; Jebara]</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**
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005...

**Support Vector Machines**
Guyon, Vapnik, Heisele, Serre, Poggio...

**Latent SVM**
Structural SVM
Felzenszwalb 00
Ramanan 03...

**Neural networks**
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
...

**Boosting**
Viola, Jones 2001,
Torralba et al. 2004,
Opelt et al. 2006,...
Learning and Recognition

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
  \[ w = [w_1, w_2, ..., w_N] \]

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

\[
p(c | w) \propto p(c) p(w | c)
\]

**Posterior** = probability that image \( I \) is of category \( c \)

- **Prior prob. of the object classes**
- **Image likelihood given the class**
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\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>
</tbody>
</table>

| Mean ranks   | 1.49  | 1.88      | 1.33  | 1.33 | 1.63   | 1.57  | 1.57  |

Csurka et al. 2004
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 have learned today?

• Introduction to object recognition
  – Representation
  – Learning
  – Recognition

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