Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)
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
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; the cerebral cortex was a movie screen so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path 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.

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. The figures are likely to further annoy the US, which has long argued that China’s exports are unfairly stimulated 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 also needs to do more to boost domestic demand so more goods stay 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 rise 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
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

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

- **Normalize patch**

- **Detect patches**
  [Mikojaczyk 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

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

![Image representation diagram]

- **frequency**
- **codewords**
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

*category models (and/or) classifiers*
2 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
First, some notations

- \( w_n \): each patch in an image
  - \( w_n = [0,0,\ldots,1,\ldots,0,0]^T \)

- \( w \): a collection of all \( N \) patches in an image
  - \( w = [w_1,w_2,\ldots,w_N] \)

- \( d_j \): 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: 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>
Case #2: Hierarchical Bayesian text models

**Probabilistic Latent Semantic Analysis (pLSA)**

Hoffman, 2001

**Latent Dirichlet Allocation (LDA)**

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

Probabilistic Latent Semantic Analysis (pLSA)

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

Latent Dirichlet Allocation (LDA)

Fei-Fei et al. ICCV 2005
Demo

- Course website

Two bag-of-words classifiers

ICCV 2005 short courses on
Recognizing and Learning Object Categories

A simple approach to classifying images is to treat them as a collection of regions, describing only their appearance and ignoring their
have been successfully used in the text community for analyzing documents and are known as "bag-of-words" models, since each doc
distribution over fixed vocabulary. Using such a representation, methods such as probabilistic latent semantic analysis (pLSA) [1]
(LDA) [2] are able to extract coherent topics within document collections in an unsupervised manner.

Recently, Fei-Fei et al. [3] and Sivic et al. [4] have applied such methods to the visual domain. The demo code implements pLSA, inc.
For comparison, a Naive Bayes classifier is also provided which requires labelled training data, unlike pLSA.

The code consists of Matlab scripts (which should run under both Windows and Linux) and a couple of 32-bit Linux binaries for doing
representation. Hence the whole system will need to be run on Linux. The code is for teaching/research purposes only. If you find a b
where csail point mit point edu.

Download

Download the code and datasets (32 Mbabytes)

Operation of code

To run the demos
task: face detection – no labeling
Demo: feature detection

- Output of crude feature detector
  - Find edges
  - Draw points randomly from edge set
  - Draw from uniform distribution to get scale
Demo: recognition examples
Learning and Recognition

1. Generative method:
   - graphical models

2. Discriminative method:
   - SVM

category models
(and/or) classifiers
Discriminative methods based on ‘bag of words’ representation
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

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

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

Histogram intersection

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

Slide credit: Kristen Grauman
Object recognition results

- ETH-80 database
  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>(O(dm^2))</td>
<td>84%</td>
</tr>
<tr>
<td>Bhattacharyya affinity</td>
<td>(O(dm^3))</td>
<td>85%</td>
</tr>
<tr>
<td>[Kondor &amp; Jebara]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyramid match</td>
<td>(O(dmL))</td>
<td>84%</td>
</tr>
</tbody>
</table>

Slide credit: Kristen Grauman
Object recognition results

- Caltech objects database
  101 object classes
- Features:
  - SIFT detector
  - PCA-SIFT descriptor, $d=10$
- 30 training images / class
- 43% recognition rate
  (1% chance performance)
- 0.002 seconds per match

Slide credit: Kristen Grauman
learning

feature detection & representation

image representation

codewords dictionary

category models (and/or) classifiers

recognition

category decision

(and/or) classifiers
What about spatial info?
What about spatial info?

- Feature level
  - Spatial influence through correlogram features: Savarese, Winn and Criminisi, CVPR 2006
What about spatial info?

- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007
What about spatial info?

- Feature level
- Generative models
  - Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  - Niebles & Fei-Fei, CVPR 2007
What about spatial info?

- Feature level
- Generative models
- Discriminative methods
  - Lazebnik, Schmid & Ponce, 2006
Invariance issues

• Scale and rotation
  – Implicit
  – Detectors and descriptors
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
    • Sudderth, Torralba, Freeman & Willsky, 2005, 2006
    • Niebles & Fei-Fei, 2007
Invariance issues

- Scale and rotation
- Occlusion
- Translation
- View point (in theory)
  - Codewords: detector and descriptor
  - Theme distributions: different view points

Fergus, Fei-Fei, Perona & Zisserman, 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 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; the cerebral cortex was a movie screen, so to speak, upon which the image was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception there is a considerably more complex process. By following the visual impulses along their path to the various columns of the retinal image to the various columns of the cerebral 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.

- Intuitive
  - Analogy to documents
Model properties

• Intuitive
  – Analogy to documents
  – Analogy to human vision

Olshausen and Field, 2004, Fei-Fei and Perona, 2005
Model properties

- Intuitive
- generative models
  - Convenient for weakly- or un-supervised, incremental training
  - Prior information
  - Flexibility (e.g. HDP)

Sivic, Russell, Efros, Freeman, Zisserman, 2005

Li, Wang & Fei-Fei, CVPR 2007
Model properties

- Intuitive
- Generative models
- Discriminative method
  - Computationally efficient

Grauman et al. CVPR 2005
Model properties

- Intuitive
- Generative models
- Discriminative method
- 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