Problem

The Myth of Human Vision

Can you understand the following?

People may have difficulties to understand different texts, but NOT images.

Photo courtesy to luster
Problem

Can the computer vision system recognize objects or scenes like human?
Why This Problem Important?

Traditional companies
[Matrox Imaging, Cognex Vision]

Searching engines
[Google Image Search, Picasa]

Mobile Apps
[snaptell, leafsnap]
Examples of Object Recognition Dataset

http://www.vision.caltech.edu
Problem

Indoor
- airlock
- anechoic chamber
- bookbindery
- bowling
- dais
- boat deck
- hatchway
- hunting lodge
- parlor
- pilothouse

Armoury
- brewery
- departure lounge
- jewellery shop
- police office
- piazza

Urban
- access road
- campus
- carport
- fire escape
- floating bridge
- launchpad
- loading dock
- plantation

Aqueduct
- alleyway
- cathedral
- flying bridge
- lookout station
- plantation

Nature
- apple orchard
- arbor
- archipelago
- crag
- cromlech
- ditch
- gorge
- grassland
- mountain
- marsh
- mineshaft
- river
- rock outcrop

http://groups.csail.mit.edu/vision/SUN/
## Overview of Classification Model

<table>
<thead>
<tr>
<th>Coding</th>
<th>Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram of SIFT</td>
<td>Vector quantization</td>
</tr>
<tr>
<td>Uncertainty-Based Quantization</td>
<td>Soft quantization</td>
</tr>
<tr>
<td>Sparse Coding</td>
<td>Soft quantization</td>
</tr>
<tr>
<td>Fisher vector/Supervector</td>
<td>GMM probability estimation</td>
</tr>
</tbody>
</table>

**Coding**: to map local features into a compact representation

**Pooling**: to aggregate these compact representation together
Outlines

• Histogram of local features
• Bag of words model
• Soft quantization and sparse coding
• Fisher vector and supervector
Histogram of Local Features
Bag of Words Models

Recall of Last Class

- Powerful local features
  - DoG
  - Hessian, Harris
  - Dense-sampling

Non-fixed number of local regions per image!
Bag of Words Models

Recall of Last Class (2)

- Histograms can provide a fixed size representation of images
- Spatial pyramid/gridding can enhance histogram presentation with spatial information
Bag of Words Models

Histogram of Local Features

frequency

condewords

\[ \text{dim} = \# \text{codewords} \]
Bag of Words Models

Histogram of Local Features (2)

\[ \text{dim} = \#\text{codewords} \times \#\text{grids} \]
Bag of Words Models

Local Feature Quantization

Slide courtesy to Fei-Fei Li
Local Feature Quantization
Local Feature Quantization

- Vector quantization
- Dictionary learning
Dictionary for Codewords

Visual Recognition And Search

Columbia University, Spring 2013
Bag of Words Models

Most slides in this section are courtesy to Fei-Fei Li
Object \rightarrow \text{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 us through our eyes. For a long time, the visual images were thought to be processed in sensory centers in the brain. However, with the introduction of a movie still camera, it became possible to analyze visual images in the human eye. Hubel and Wiesel discovered that the nervous system of the human eye has a more complex structure than previously thought. For the various stages of image transmission 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% increase in exports to $750bn, compared with $660bn. This could annoy the US, which is determined to keep the yuan stable against the dollar. China's central bank has already agreed to appreciate the yuan, 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.
Bag of Words Models

Underlining Assumptions - Image
learning

- feature detection & representation
- image representation

recognition

- category models (and/or) classifiers
- decision
Bag of Words Models

Borrowing Techniques from Text Classification

- PLSA
- Naïve Bayesian Model

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
Bag of Words Models

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)

Hoffman, 2001

Blei et al., 2001
Bag of Words Models

Probabilistic Latent Semantic Analysis (pLSA)

\[ d \rightarrow z \rightarrow w \]

"face"

Sivic et al. ICCV 2005
Bag of Words Models

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

Parameter estimated by EM or Gibbs sampling
Bag of Words Models

Recognition using pLSA

\[ z^* = \arg \max_z p(z \mid d) \]

\[
\begin{align*}
\text{Bag of Words} & \quad = \quad \text{Bag of Words} \\
\text{P(w|d)} & \quad = \quad \text{P(w|z)} \\
\text{P(z|d)} & \quad = \quad \text{P(z|d)}
\end{align*}
\]
Bag of Words Models

Scene Recognition using LDA

Latent Dirichlet Allocation (LDA)

Fei-Fei et al. ICCV 2005
Bag of Words Models

Spatial-Coherent Latent Topic Model

Cao and Fei-Fei, ICCV 2007
Bag of Words Models

Simultaneous Segmentation and Recognition
Bag of Words Models

Pros and Cons

*Images differ from texts!*

Bag of Words Models are good in
- Modeling prior knowledge
- Providing intuitive interpretation

But these models suffer from
- Loss of information in quantization of “visual words”
- Loss of spatial information

Better coding
Better pooling
Soft Quantization and Sparse Coding
Soft Quantization

Hard Quantization

\[ CB(w) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1 & \text{if } w = \arg \min_{v \in V} (D(v, r_i)) \\ 0 & \text{otherwise,} \end{cases} \]
Soft Quantization

Uncertainty-Based Quantization

Model the uncertainty across multiple codewords

\[ K_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{x^2}{\sigma^2}\right) \]

\[ \text{UNC}(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_\sigma(D(w, r_i))}{\sum_{j=1}^{|V|} K_\sigma(D(v_j, r_i))} \]

Gemert et al, Visual Word Ambiguity, PAMI 2009
Soft Quantization

Intuition of UNC

Gemert et al, Visual Word Ambiguity, PAMI 2009
Soft Quantization

Improvement of UNC

Caltech-101 absolute difference per category

Classification Rate (%) vs. Category Label

- Hard Assignment
- Codeword Uncertainty
Soft Quantization

Sparse Coding-Based Quantization

Hard quantization can be viewed as an “extremely sparse representation”

\[
\min ||x - Dz||_2^2 \quad s.t. \quad ||z||_0 = 1
\]

A more general but hard to solve representation

\[
\min \limits_{\tilde{z}} ||x - D\tilde{z}||_2^2 + \lambda ||z||_0
\]

In practice we consider

\[
\min \limits_{\tilde{z}} ||x - D\tilde{z}||_2^2 + \lambda ||z||_1
\]  

Sparse coding
Soft Quantization

Sparse Coding-Based Quantization

Hard quantization can be viewed as an “extremely sparse representation”

\[
\min_{\mathbf{z}} \| \mathbf{x} - \mathbf{Dz} \|_2^2 \quad \text{s.t.} \quad \| \mathbf{z} \|_0 = 1
\]

A more general but hard to solve representation

\[
\min_{\mathbf{z}} \| \mathbf{x} - \mathbf{Dz} \|_2^2 + \lambda \| \mathbf{z} \|_0
\]

In practice we consider

\[
\min_{\mathbf{z}} \| \mathbf{x} - \mathbf{Dz} \|_2^2 + \lambda \| \mathbf{z} \|_1 \quad \text{Sparse coding}
\]
Yang et al obtain good recognition accuracy by combining sparse coding with spatial pyramid and dictionary training.

*More details will be in group presentation.*

*Yang et al, Linear Spatial Pyramid Matching using Sparse Coding for Image Classification, CVPR 2009*
Fisher Vector and Supervector

One of the most powerful image/video classification techniques

Thanks to Zhen Li and Qiang Chen constructive suggestions to this section
Fisher Vector and Supervector

Winning Systems

Classification task

PASCAL

Large Scale Visual Recognition Challenge

IMAGENET

2009 2010 2011 2012

NEC

NUS

2010 2011

Xerox
### Fisher Vector and Supervector

#### Literature


*These papers are not very easy to read.*

*Let me take a simplified perspective via coding&pooling framework*
Fisher Vector and Supervector

Coding without Information Loss

• Coding with hard assignment \( x \rightarrow D_x \)

• Coding with soft assignment \( x \rightarrow \sum_i \rho_i D_i \)

• How to keep all the information?
Fisher Vector and Supervector

Coding without Information Loss

- Coding with hard assignment \( x \rightarrow D_x \)

- Coding with soft assignment \( x \rightarrow \sum_i \rho_i^x D_i \)

- How to keep all the information?

\[
x \rightarrow [\rho_1^x x, \rho_2^x x, \rho_3^x x, \cdots]
\]
Fisher Vector and Supervector

An Intuitive Illustration

**Coding**

\[ x_1 \iff \rho_1 x_1 \quad \rho_2 x_1 \quad \rho_3 x_1 \quad \cdots \]

\[ x_2 \iff \rho_1 x_2 \quad \rho_2 x_2 \quad \rho_3 x_2 \quad \cdots \]

\[ \cdots \]
Fisher Vector and Supervector

An Intuitive Illustration

**Coding**

\[ \begin{align*}
\mathbf{x}_1 & \leftrightarrow \rho_1^1 \mathbf{x}_1 \\
\mathbf{x}_2 & \leftrightarrow \rho_2^2 \mathbf{x}_2 \\
\ddots & \ddots
\end{align*} \]

Component 1

\[ \begin{align*}
\rho_1^1 \mathbf{x}_1 \\
\rho_2^2 \mathbf{x}_2
\end{align*} \]

Component 2

\[ \begin{align*}
\rho_1^1 \mathbf{x}_1 \\
\rho_2^2 \mathbf{x}_2
\end{align*} \]

Component 3

\[ \begin{align*}
\rho_3^1 \mathbf{x}_1 \\
\rho_2^2 \mathbf{x}_2
\end{align*} \]
Fisher Vector and Supervector

An Intuitive Illustration

Component 1
\[ \rho_1^1 x_1 + \rho_1^2 x_2 \]

Component 2
\[ \rho_2^1 x_1 + \rho_2^2 x_2 \]

Component 3
\[ \rho_3^1 x_1 + \rho_3^2 x_2 \]

Pooling
\[ \sum_t \rho_1^t x_t, \sum_t \rho_2^t x_t, \sum_t \rho_3^t x_t \]
Fisher Vector and Supervector

Implementation of Supervector

In speech (speaker identification), supervector refer to stacked means of adaptive GMMs.

\[ p(k \mid x_t) = \frac{w_k \mathcal{N}(x_t; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(x_t; \mu_j, \Sigma_j)} \]

\[ m_k = \frac{1}{\sum_t p(k \mid x_t)} \sum_t p(k \mid x_t) x_t \]

Supervector = \[ [m_1, m_2, \cdots, m_K] \]
Fisher Vector and Supervector

Interpretation with Supervector

Origin distribution

New distribution

*Picture from Reynolds, Quatieri, and Dunn, DSP, 2001*
In practice, a normalization process using the covariance matrix often improves the performance

$$\left[ \sqrt{w_1 \Sigma_1^{-1/2}} \mathbf{m}_1; \cdots; \sqrt{w_K \Sigma_K^{-1/2}} \mathbf{m}_K \right]$$

Moreover, we can subtract the original mean vector for the ease of normalization

$$\left[ \sqrt{w_1 \Sigma_1^{-1/2}} (\mathbf{m}_1 - \mu_1); \cdots; \sqrt{w_K \Sigma_K^{-1/2}} (\mathbf{m}_K - \mu_K) \right]$$

The representation is also called Hierarchical Gaussianization (HG).
Fisher Vector

Let $u_\lambda$ be the probability density function with parameter $\lambda$. 

[Jaakkola and Haussler, NIPS 98] suggested that $X$ can be described by the derivative subject to $\lambda$:

$$G_X^\lambda = \frac{1}{T} \nabla_\lambda \log u_\lambda(X)$$

Now we can define the Fisher Kernel:

$$K(X, Y) = G_X^\lambda \top F_{-1}^\lambda G_Y^\lambda$$

where $F_{-1}^\lambda$ is called the Fisher information matrix:

$$F_{-1}^\lambda = L_\lambda \top L_\lambda$$
Fisher Vector with GMM

Consider the Gaussian Mixture Model

\[ \lambda = \{ w_i, \mu_i, \sigma_i, i = 1 \ldots K \} \]

We consider

\[ g^X_\lambda = L_\lambda G^X_\lambda \]

With GMM, Fisher vectors can be obtained:

Let

\[ w^X_i = \frac{1}{T} \sum_{t=1}^{T} \gamma_t(i) \]

\[ g^X_\lambda = \frac{w^X_i}{\sqrt{w_i}} \sigma^{-1}_i (m_i - \mu_i) \]

The Fisher vector
Fisher Vector and Supervector

Comparison

- Supervector

\[ \sqrt{w_1} \Sigma_1^{-1/2} (m_1 - \mu_1); \cdots; \sqrt{w_K} \Sigma_K^{-1/2} (m_K - \mu_K) \]

- Fisher vector

\[ \frac{w_1^X}{\sqrt{w_1}} \sigma_1^{-1} (m_1 - \mu_1); \cdots; \frac{w_K^X}{\sqrt{w_K}} \sigma_K^{-1} (m_K - \mu_K) \]
Fisher Vector and Supervector

Comparison

Diagonal covariance matrix

\[ \sqrt{w_1} \mathbf{\Sigma}_1^{-1/2} (\mathbf{m}_1 - \mu_1); \cdots; \sqrt{w_K} \mathbf{\Sigma}_K^{-1/2} (\mathbf{m}_K - \mu_K) \]

Posterior estimation of \( w_1 \)

\[ \frac{w_1^X}{\sqrt{w_1}} \mathbf{\Sigma}_1^{-1} (\mathbf{m}_1 - \mu_1); \cdots; \frac{w_K^X}{\sqrt{w_K}} \mathbf{\Sigma}_K^{-1} (\mathbf{m}_K - \mu_K) \]

Diagonal covariance with same derivation

The two representations are almost the same even with different motivations.
Fisher Vector and Supervector

How to Code Your Own

• Learn from existing code
  http://lear.inrialpes.fr/src/inria_fisher/ (Linux or Mac)
• Learn from public GMM code
• Be careful with pitfalls
  – Probability is comparable to machine’s rounding error: compute logP instead of P
  – Try different normalization strategy
  – Try to make the code efficient
Summary

**Coding**
- Histogram of SIFT
- Vector quantization
- Uncertainty-Based Quantization
- Soft quantization
- Sparse Coding
- Soft quantization
- Fisher vector/Supervector
- GMM probability estimation
- GMM adaptation

**Pooling**
- Histogram aggregation
- Histogram aggregation
- Max pooling
• Read the deformable part model
• Enjoy the Rogerio’s talk next week
☺
• Project proposal deadline Feb 19
• Project presentation Feb 21

Todo Before Next Class
Summary