Machine Learning in Computer Vision

Fei-Fei Li
What is (computer) vision?

- When we “see” something, what does it involve?
- Take a picture with a camera, it is just a bunch of colored dots (pixels)
- Want to make computers understand images
- Looks easy, but not really…
What is it related to?

- Biology
- Psychology
- Neuroscience
- Cognitive sciences
- Computer Vision
- Information retrieval
- Machine Learning
- Speech
- Robotics
- Computer Science
- Information Engineering
- Maths
- Physics
Quiz?
What about this?
A picture is worth a thousand words.
--- Confucius
or *Printers’ Ink* Ad (1921)
A picture is worth a thousand words.
--- Confucius
or *Printers’ Ink* Ad (1921)
A picture is worth a thousand words.

--- Confucius

or *Printers’ Ink* Ad (1921)
Today: machine learning methods for object recognition
outline

• Intro to object categorization
• Brief overview
  – Generative
  – Discriminative
• Generative models
• Discriminative models
How many object categories are there?

~10,000 to 30,000

Biederman 1987
Challenges 1: view point variation

Michelangelo 1475-1564
Challenges 2: illumination
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

• Lowe, et al. 1999, 2003
• Mahamud and Herbert, 2000
• Ferrari, Tuytelaars, and Van Gool, 2004
• Rothganger, Lazebnik, and Ponce, 2004
• Moreels and Perona, 2005
• ...
Challenges 7: intra-class variation
Object categorization: the statistical viewpoint

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

\[ \text{vs.} \]

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

- 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

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

- **Posterior ratio**
- **Likelihood ratio**
- **Prior ratio**

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

- Direct modeling of

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})}
\]
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
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
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
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 the brain via our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain, just as a movie screen displays an image and through the discoveries of Hubel and Wiesel we now know that the visual perception in the brain is a more complex process. By following the visual impulses along their path to the various cell layers of the visual 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 a 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 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

- 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
3. Image representation

![Image Representation Diagram]

- Frequency
- Codewords

---

- Car Image Insert
1. Feature detection & representation
2. Codewords dictionary
3. Image representation
category models (and/or) classifiers

codewords dictionary

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
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)
Case #2: Hierarchical Bayesian text models

Probabilistic Latent Semantic Analysis (pLSA)

“face”

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

"beach"

Latent Dirichlet Allocation (LDA)

Fei-Fei et al. ICCV 2005
Another application

• Human action classification
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

Sudderth et al. 2005
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 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; it was seen as a movie screen into which the retinal image 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 cerebral cortex, Hubel and Wiesel have been able to demonstrate that the message about an 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.
Model properties

• 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”
One-shot learning of object categories

Fei-Fei et al. ‘03, ‘04, ‘06
One-shot learning of object categories

Fei-Fei et al. ‘03, ‘04, ‘06

P. Bruegel, 1562
model representation

One-shot learning of object categories

Fei-Fei et al. ‘03, ‘04, ‘06
X (location)

(x,y) coords. of region center

A (appearance)

(normalize)

11x11 patch

Projection onto PCA basis

\[
\begin{pmatrix}
c_1 \\
c_2 \\
\vdots \\
c_{10}
\end{pmatrix}
\]
The Generative Model

\[ \begin{align*} 
X & \xrightarrow{\mu^X, \Gamma^X} \ h \\
A & \xrightarrow{\mu^A, \Gamma^A} 
\end{align*} \]

\[ \text{normalize 11x11 patch} \]

\[ \text{Projection onto PCA basis} \]

\[ \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{10} \end{pmatrix} \]

\( X \) (location)

(x,y) coords. of region center

\( \mu^X, \Gamma^X \)

\( \mu^A, \Gamma^A \)

\( h \)

\( X \)

\( A \) (appearance)
The Generative Model

- Observed variables
- Hidden variable
- Parameters
- Observed variables
The Generative Model

\[ \theta = \{ \mu^X, \Gamma^X, \mu^A, \Gamma^A \} \]

ML/MAP

Weber et al. ’98 ’00, Fergus et al. ’03
The Generative Model

where \( \theta = \{ \mu^X, \Gamma^X, \mu^A, \Gamma^A \} \)
The Generative Model

\[ \theta_1 \]

\[ \theta_2 \]

\[ \theta_n \]

ML/MAP
The Generative Model

Parameters to estimate: \( \{m^X, \beta^X, a^X, B^X, m^A, \beta^A, a^A, B^A\} \)
i.e. parameters of Normal-Wishart distribution
The Generative Model

Fei-Fei et al. ‘03, ‘04, ‘06
The Generative Model

Fei-Fei et al. ‘03, ‘04, ‘06
1. human vision

2. model representation

3. learning & inferences

4. evaluation & dataset & application

One-shot learning of object categories
One-shot learning of object categories

Fei-Fei et al. 2003, 2004, 2006
learning & inferences

Bayesian

One-shot learning of object categories

Fei-Fei et al. 2003, 2004, 2006
Variational EM

M-Step

new estimate of $p(\theta|\text{train})$

prior knowledge of $p(\theta)$

E-Step

Random initialization

Attias, Jordan, Hinton etc.
evaluation & dataset

One-shot learning of object categories

Fei-Fei et al. 2004, 2006a, 2006b
One-shot learning of object categories

Fei-Fei et al. 2004, 2006a, 2006b

evaluation & dataset -- Caltech 101 Dataset
One-shot learning of object categories

Fei-Fei et al. 2004, 2006a, 2006b
Part 3: discriminative methods
Discriminative methods

Object detection and recognition is formulated as a classification problem. The image is partitioned into a set of overlapping windows … and a decision is taken at each window about if it contains a target object or not.
Discriminative vs. generative

- Generative model
  *(The artist)*

- Discriminative model
  *(The lousy painter)*

- Classification function

\[
\text{label} = F_{\text{Zebra}}(\text{Data})
\]
Discriminative methods

Nearest neighbor
10^8 examples
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
...

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

Support Vector Machines and Kernels
Guyon, Vapnik
Heisele, Serre, Poggio, 2001
...

Conditional Random Fields
McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
...
Formulation

• Formulation: binary classification

\[
\text{Features } x = X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_N \quad X_{N+1} \quad X_{N+2} \quad \ldots \quad X_{N+M}
\]

\[
\text{Labels } y = -1 \quad +1 \quad -1 \quad -1 \quad ? \quad ? \quad ? \quad ?
\]

Training data: each image patch is labeled as containing the object or background

Test data

• Classification function

\[\hat{y} = F(x)\] Where \(F(x)\) belongs to some family of functions

• Minimize misclassification error

(Not that simple: we need some guarantees that there will be generalization)
Overview of section

- Object detection with classifiers

- **Boosting**
  - Gentle boosting
  - Weak detectors
  - Object model
  - Object detection

- Multiclass object detection
Why boosting?

- A simple algorithm for learning robust classifiers
  - Freund & Shapire, 1995
  - Friedman, Hastie, Tibshhirani, 1998

- Provides efficient algorithm for sparse visual feature selection
  - Tieu & Viola, 2000
  - Viola & Jones, 2003

- Easy to implement, not requires external optimization tools.
Boosting

• Defines a classifier using an additive model:

\[ F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots \]
Boosting

• Defines a classifier using an additive model:

\[ F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots \]

• We need to define a family of weak classifiers

\[ f_k(x) \] from a family of weak classifiers
Boosting

- It is a sequential procedure:

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bigcirc) \\
-1 & (\bigotimes) 
\end{cases} \]

and a weight:

\[ w_t = 1 \]
Toy example

Weak learners from the family of lines

Each data point has a class label:

\[ y_t = \begin{cases} +1 & (\circ) \\ -1 & (\bullet) \end{cases} \]

and a weight:

\[ w_t = 1 \]

\[ h \Rightarrow p(\text{error}) = 0.5 \quad \text{it is at chance} \]
This one seems to be the best

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bullet) \\
-1 & (\bigcirc) 
\end{cases} \]

and a weight:

\[ w_t = 1 \]

This is a ‘\textbf{weak classifier}’: It performs slightly better than chance.
Toy example

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bullet) \\
-1 & (\circ) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]

We set a new problem for which the previous weak classifier performs at chance again.
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\circ) \\
-1 & (\bigcirc) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
Toy example

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bullet) \\
-1 & (\circ) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]

We set a new problem for which the previous weak classifier performs at chance again.
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} 
  +1 & (\bullet) \\
  -1 & (\circ) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
The strong (non-linear) classifier is built as the combination of all the weak (linear) classifiers.
From images to features: Weak detectors

We will now define a family of visual features that can be used as weak classifiers ("weak detectors")

\[ h_i(I, x, y) \]

Takes image as input and the output is binary response. The output is a weak detector.
Weak detectors

Textures of textures

Tieu and Viola, CVPR 2000

\[ g_{i,j,k} = \sum_{\text{pixels}} | |I * f_i| \downarrow_2 * f_j| \downarrow_2 * f_k \]

Every combination of three filters generates a different feature

This gives thousands of features. Boosting selects a sparse subset, so computations on test time are very efficient. Boosting also avoids overfitting to some extend.
Weak detectors

Haar filters and integral image
Viola and Jones, ICCV 2001

The average intensity in the block is computed with four sums independently of the block size.
Weak detectors

Other weak detectors:
• Carmichael, Hebert 2004
• Yuille, Snow, Nitzbert, 1998
• Amit, Geman 1998
• Papageorgiou, Poggio, 2000
• Heisele, Serre, Poggio, 2001
• Agarwal, Awan, Roth, 2004
• Schneiderman, Kanade 2004
• …
Weak detectors

**Part based**: similar to part-based generative models. We create weak detectors by using parts and voting for the object center location.

These features are used for the detector on the course web site.
Weak detectors

First we collect a set of part templates from a set of training objects.

Vidal-Naquet, Ullman (2003)
Weak detectors

We now define a family of “weak detectors” as:

\[ h_i(I, x, y) = [I \otimes P_i] * g_i \]

Better than chance
Weak detectors

We can do a better job using filtered images

\[ h_i(I, x, y) = [I \ast f_i \otimes P_i] \ast g_i \]

Still a weak detector but better than before
Training

First we evaluate all the N features on all the training images.

Feature 1

\[
\begin{bmatrix}
\star & \star \\
\end{bmatrix} \otimes
\begin{bmatrix}
\star & \star \\
\end{bmatrix} =
\begin{bmatrix}
\star & \star \\
\end{bmatrix}
\]

Feature N

\[
\begin{bmatrix}
\star & \star \\
\end{bmatrix} \otimes
\begin{bmatrix}
\star & \star \\
\end{bmatrix} =
\begin{bmatrix}
\star & \star \\
\end{bmatrix}
\]

Then, we sample the feature outputs on the object center and at random locations in the background:

Positive Training Vector:

\[
\begin{bmatrix}
1 \\
2 \\
3 \\
N-1 \\
N \\
\end{bmatrix}
\]

Negative Training Vectors:

\[
\begin{bmatrix}
1 \\
2 \\
3 \\
N-1 \\
N \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 \\
2 \\
3 \\
N-1 \\
N \\
\end{bmatrix}
\]
Representation and object model

Selected features for the screen detector

Lousy painter
Representation and object model

Selected features for the car detector
Overview of section

- Object detection with classifiers

- Boosting
  - Gentle boosting
  - Weak detectors
  - Object model
  - **Object detection**

- Multiclass object detection
Example: screen detection
Example: screen detection

Feature output

Thresholded output

Weak ‘detector’
Produces many false alarms.
Example: screen detection

- Feature output
- Thresholded output
- Strong classifier at iteration 1
Example: screen detection

Second weak ‘detector’
Produces a different set of false alarms.
Example: screen detection

Feature output → Thresholded output → Strong classifier

Strong classifier at iteration 2
Example: screen detection

Feature output → Thresholded output → Strong classifier

Strong classifier at iteration 10
Example: screen detection

Feature output → Thresholded output → Strong classifier

Adding features

Strong classifier at iteration 200
applications
Face-hunting cameras boost Nikon

Japanese camera maker Nikon has tripled its profits on the back of strong sales of digital cameras that automatically focus on human faces.

Face recognition cameras like the Coolpix L1 are popular.
**Sample image:** Subject as seen on the COOLPIX 5900 camera’s color LCD and when using Nikon’s Face-priority AF function.
Document Analysis

Digit recognition, AT&T labs
http://www.research.att.com/~yann/
Medical Imaging
Toys and robots
Finger prints
Surveillance
Searching the web