Object Recognition: Conceptual Issues

Slides adapted from Fei-Fei Li, Rob Fergus, Antonio Torralba, and K. Grauman
Issues in recognition

The statistical viewpoint
Generative vs. discriminative methods
Model representation
Generalization, bias vs. variance
Supervision
Datasets
Discrete Random Variables

$X$ denotes a random variable.

$X$ can take on a countable number of values in \{x_1, x_2, \ldots, x_n\}.

$P(X=x_i)$, or $P(x_i)$, is the probability that the random variable $X$ takes on value $x_i$.

$P(.)$ is called probability mass function.
Continuous Random Variables

$X$ takes on values in the continuum.

$p(X=x)$, or $p(x)$, is a probability density function.

$\Pr(x \in (a, b)) = \int_{a}^{b} p(x) \, dx$

E.g.

![Graph of probability density function](image-url)
Joint and Conditional Probability

\[ P(X=x \text{ and } Y=y) = P(x,y) \]

If \( X \) and \( Y \) are independent then

\[ P(x,y) = P(x) \cdot P(y) \]

\( P(x \mid y) \) is the probability of \( x \) given \( y \)

\[ P(x \mid y) = \frac{P(x,y)}{P(y)} \]

\[ P(x,y) = P(x \mid y) \cdot P(y) \]

If \( X \) and \( Y \) are independent then

\[ P(x \mid y) = P(x) \]
Law of Total Probability, Marginals

Discrete case

\[ \sum_x P(x) = 1 \]

\[ P(x) = \sum_y P(x, y) \]

\[ P(x) = \sum_y P(x \mid y) P(y) \]

Continuous case

\[ \int p(x) \, dx = 1 \]

\[ p(x) = \int p(x, y) \, dy \]

\[ p(x) = \int p(x \mid y) p(y) \, dy \]
Bayes Formula

\[ P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x) \]

\[ \Rightarrow \]

\[ P(x \mid y) = \frac{P(y \mid x) P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \]
Normalization

\[ P(x \mid y) = \frac{P(y \mid x) P(x)}{P(y)} = \eta P(y \mid x) P(x) \]

\[ \eta = P(y)^{-1} = \frac{1}{\sum_{x} P(y \mid x)P(x)} \]

Algorithm:

\[ \forall x : aux_{x \mid y} = P(y \mid x) P(x) \]

\[ \eta = \frac{1}{\sum_{x} aux_{x \mid y}} \]

\[ \forall x : P(x \mid y) = \eta aux_{x \mid y} \]
Bayes' Rule

Bayes Rule for point probabilities

\[ P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)} \]

or in distribution form

\[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \]

Useful for assessing diagnostic probability from causal probability:

\[ P(\text{Cause} \mid \text{Effect}) = \frac{P(\text{Effect} \mid \text{Cause})P(\text{Cause})}{P(\text{Effect})} \]

E.g., let \( M \) be meningitis, \( S \) be stiff neck:

\[ P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008 \]
Bayes' Rule

Bayes Rule for point probabilities

\[ P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008 \]

\[ P(s) = P(s \mid m)P(m) + P(s \mid \neg m)P(\neg m) \]

Here we assumed that prior probability is known quantity – exactly 0.001 (point estimate)
Bayesian approach capture the uncertainty about prior as a distribution
Then the posterior will also be distribution
Simple Example of Recognition

Suppose we obtain measurement $z$
What is $P(zebra|z)$?
Causal vs. Diagnostic Reasoning

$P(\text{zebra}|z)$ is diagnostic.

$P(z|\text{zebra})$ is causal.

Often causal knowledge is easier to obtain.

Bayes rule allows us to use causal knowledge:

$$P(\text{zebra} | z) = \frac{P(z|\text{zebra})P(\text{zebra})}{P(z)}$$

*count frequencies!*
Combining Evidence

Suppose we obtain another observation \( z_2 \).

How can we integrate this new information?

More generally, how can we estimate 
\[
P(x \mid z_1 \ldots z_n)
\]?
Example: Second Measurement

\[ P(z_2|\text{zebra}) = 0.5 \quad \quad P(z_2|\neg\text{zebra}) = 0.6 \]
\[ P(\text{zebra}|z_1) = \frac{2}{3} \]

\[
P(\text{zebra} | z_2, z_1) = \frac{P(z_2 | \text{zebra}) \cdot P(\text{zebra} | z_1)}{P(z_2 | \text{zebra}) \cdot P(\text{zebra} | z_1) + P(z_2 | \neg\text{zebra}) \cdot P(\neg\text{zebra} | z_1)}
\]

\[
= \frac{1 \cdot \frac{2}{3}}{2 \cdot \frac{2}{3} + \frac{3}{5} \cdot \frac{1}{3}} = \frac{5}{8} = 0.625
\]

• \(z_2\) lowers the probability that the picture is zebra.
Object categorization: the statistical viewpoint

- MAP decision: \( p(\text{zebra} \mid \text{image}) \) vs. \( p(\text{no zebra} \mid \text{image}) \)

\[
P(x \mid y) = \frac{P(y \mid x) P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}
\]
**Object categorization: the statistical viewpoint**

- **MAP decision:**
  
  \[ p(zebra \mid image) \]
  
  vs.
  
  \[ p(no \ zebra \mid image) \]

- **Bayes rule:**

  \[ p(zebra \mid image) \propto p(image \mid zebra) p(zebra) \]

  \[ \text{posterior} \quad \text{likelihood} \quad \text{prior} \]
Object categorization: the statistical viewpoint

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

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

- Direct modeling of $p(\text{zebra} \mid \text{image})$
Generative methods

- Model $p(\text{image} \mid \text{zebra})$ and $p(\text{image} \mid \text{no zebra})$

<table>
<thead>
<tr>
<th>$p(\text{image} \mid \text{zebra})$</th>
<th>$p(\text{image} \mid \text{no zebra})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Middle</td>
</tr>
<tr>
<td>High</td>
<td>Middle $\rightarrow$ Low</td>
</tr>
</tbody>
</table>
Generative vs. discriminative methods

• Generative methods
  + Interpretable
  + Can be learned using images from just a single category
    – Sometimes we don’t need to model the likelihood when all we want is to make a decision

• Discriminative methods
  + Efficient
  + Often produce better classification rates
    – Can be hard to interpret
    – Require positive and negative training data
Learning approaches proceed in supervised way: need some labeled data

Unsupervised

“Weakly” supervised

Supervised

Definition depends on task
What task?

- **Classification**
  - Object present/absent in image
  - Background may be correlated with object

- **Localization / Detection**
  - Localize object within the frame
  - Bounding box or pixel-level segmentation
Caltech 101 & 256

http://www.vision.caltech.edu/Image_Datasets/Caltech101/
http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

2008 Challenge classes:

Person: person

Animal: bird, cat, cow, dog, horse, sheep

Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train

Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

• Main competitions
  – **Classification**: For each of the twenty classes, predicting presence/absence of an example of that class in the test image
  – **Detection**: Predicting the bounding box and label of each object from the twenty target classes in the test image

http://pascallin.ecs.soton.ac.uk/challenges/VOC/

• “Taster” challenges
  – **Segmentation:** Generating pixel-wise segmentations giving the class of the object visible at each pixel, or "background" otherwise
  
  – **Person layout:** Predicting the bounding box and label of each part of a person (head, hands, feet)
Labeling with games

http://www.gwap.com/gwap/

Figure 1. Partners agreeing on an image in the ESP Game. Neither player can see the other’s guesses.

Figure 2. Peekaboom. “Peek” tries to guess the word associated with an image slowly revealed by “Boom.”

LabelMe

http://labelme.csail.mit.edu/
Summary

• Recognition is the “grand challenge” of computer vision

• History
  – Geometric methods
  – Appearance-based methods
  – Sliding window approaches
  – Local features
  – Parts-and-shape approaches
  – Bag-of-features approaches

• Issues
  – Generative vs. discriminative models
  – Supervised vs. unsupervised methods
  – Tasks, datasets
Discriminative Methods

- Object/scene category recognition
- Multi-class classification problem
- Supervised setting
- Given examples of images with category labels
- Learn classifier to predict labels of new images
Discriminative methods

- Brief overview of machine learning
- Linear regression
- Logistic regression
- Decision trees
- Boosting idea
- SVM

*Slides from Russell and Norvig, AI book*
Inductive learning method

• Construct/adjust $h$ to agree with $f$ on training set
• ($h$ is consistent if it agrees with $f$ on all examples)

• E.g., curve fitting:

![Diagram of curve fitting](image)
Inductive learning method

- Construct/adjust $h$ to agree with $f$ on training set
- ($h$ is consistent if it agrees with $f$ on all examples)

E.g., curve fitting:
Inductive learning method

• Construct/adjust $h$ to agree with $f$ on training set
• ($h$ is consistent if it agrees with $f$ on all examples)

• E.g., curve fitting:
Inductive learning method

• Construct/adjust $h$ to agree with $f$ on training set
• ($h$ is consistent if it agrees with $f$ on all examples)

• E.g., curve fitting:
Linear regression

• Fit function to the data so you can predict future values
• Given data \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)
• Find such hypothesis \(f\), such that \(y = f(x)\) has small error of future data
• Choose the form of function and estimate the data using linear least squares techniques (in closed form, or gradient descent)
Logistic Regression

- Similar as linear regression, but now $y$'s are only +/- or 0/1 denoting positive or negative examples of a class
- We know that our function should have values 0 or 1
- Transform to continuous setting – construct such as $h$ where function would have values only between 0-1

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

- $g(.)$ is a sigmoid, logistic function

$$g(z) = \frac{1}{1 + e^{-z}}$$

- Property of the derivative

$$g'(z) = g(z)(1 - g(z))$$
Logistic Regression

• How do we find parameter $\theta$
• Find such parameters so as to maximize likelihood of the data assume that

$$P(y = 1|x; \theta) = h_\theta(x)$$
$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

• More compactly

$$p(y|x; \theta) = (h_\theta)^y(1 - h_\theta)^{(1-y)}$$

• Resulting gradient ascent rule

$$\theta_j = \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$$

• Closely related to perceptron learning algorithm where values are forced to be 0-1 - no clear probabilistic interpretation
Ensemble learning

- Idea: instead of keeping simple hypothesis – keep multiple of them – ensemble of hypotheses
- Increased expressive power
- Enables us to combine several simpler hypotheses to get better prediction
Ensemble learning

- **Boosting**
- Idea: weighted training set – each training example has an associated weight \( w \geq 0 \)
- The higher the weight, higher importance of the example in the learning stage
- Start with all examples with weight 1.
- Generate hypothesis \( h_1 \)
- \( h_1 \) classifies some examples correctly some incorrectly
- Next hypothesis should do better on the misclassified examples – i.e. increase their weights
- Given new weighted training set generate \( h_2 \) etc
- We will see examples of these later
Linear classifiers

• Find linear function (*hyperplane*) to separate positive and negative examples – many such hyperplanes

\[
x_i \text{ positive: } x_i \cdot w + b \geq 0
\]
\[
x_i \text{ negative: } x_i \cdot w + b < 0
\]

Which hyperplane is best?
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 \( 2 / ||w|| \)

Summary: Discriminative methods

• Nearest-neighbor and k-nearest-neighbor classifiers
  – L1 distance, $\chi^2$ distance, quadratic distance, Earth Mover’s Distance

• Support vector machines
  – Linear classifiers
  – Margin maximization
  – The kernel trick
  – Kernel functions: histogram intersection, generalized Gaussian, pyramid match
  – Multi-class

• Of course, there are many other classifiers out there
  – Neural networks, boosting, decision trees, …