Recognizing and Learning Object Categories: Year 2007

Li Fei-Fei, Princeton
Rob Fergus, MIT
Antonio Torralba, MIT

Agenda

- Introduction
- Bag-of-words models
- Part-based models
- Discriminative methods
- Segmentation and recognition
- Datasets & Conclusions

How many object categories are there?

~10,000 to 30,000

Challenges 1: view point variation

Biederman 1987

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

• 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
History: early 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

Object categorization: the statistical viewpoint

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

Bayes rule:

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

- Discriminative methods model posterior
- Generative methods model likelihood and prior
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
– Invariances
  • View point
  • Illumination
  • Occlusion
  • Scale
  • Deformation
  • Clutter
  • etc.

– Part-based 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

– Methods of training: generative vs. discriminative

– 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)
Recognition

- Scale / orientation range to search over
- Speed
- Context

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

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 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, 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.

A clarification: definition of “BoW”

- Looser definition
  - Independent features
A clarification: definition of “BoW”

- Looser definition
  - Independent features
- Stricter definition
  - Independent features
  - Histogram representation
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

- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation based patches (Barnard, Duygulu, Forsyth, de Freitas, Blei, Jordan, 2003)

Slide credit: Josef Sivic
1. Feature detection and representation

2. Codewords dictionary formation

---

Slide credit: Josef Sivic

Fei-Fei et al. 2005
3. Image representation

- Image patch examples of codewords

- Frequency 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. Generative method:
   - graphical models

2. Discriminative method:
   - SVM

**category models (and/or) classifiers**

Discriminative methods based on ‘bag of words’ representation

- Zebra
- Non-zebra

Decision boundary

- Grauman & Darrell, 2005, 2006:
  - SVM w/ Pyramid Match kernels
- Others
  - Csurka, Bray, Dance & Fan, 2004
  - Serre & Poggio, 2005
Pyramid Match (Grauman & Darrell 2005)

Histogram intersection

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

matches at this level

matches at previous level

\[ N_i = I(H_i(X), H_i(Y)) - I(H_{i-1}(X), H_{i-1}(Y)) \]

Difference in histogram intersections across levels counts number of new pairs matched

Pyramid match kernel

\[ K_\Delta(\Psi(X), \Psi(Y)) = \sum_{i=0}^{L} \frac{1}{2^i} \left( I(H_i(X), H_i(Y)) - I(H_{i-1}(X), H_{i-1}(Y)) \right) \]

- Weights inversely proportional to bin size
- Normalize kernel values to avoid favoring large sets

Slide credit: Kristen Grauman
Example pyramid match

Level 0

$X$  
\[ N_0 = 2 \]
\[ w_0 = 1 \]

$Y$

$H_0(X)$  
$H_0(Y)$

$\mathcal{I}_0 = 2$

Slide credit: Kristen Grauman

Level 1

$X$

$H_1(X)$

$Y$

$H_1(Y)$

$\mathcal{I}_1 = 4$

$N_1 = 4 - 2 = 2$

$w_1 = \frac{1}{2}$

Slide credit: Kristen Grauman

Example pyramid match

Level 2

$X$

$N_2 = 5 - 4 = 1$

$w_2 = \frac{1}{4}$

$Y$

$H_2(X)$  
$H_2(Y)$

$\mathcal{I}_2 = 5$

Slide credit: Kristen Grauman

Example pyramid match

Pyramid match

$K_{\Delta} = \sum_{i=0}^{L} w_i N_i$

$= 1(2) + \frac{1}{2}(2) + \frac{1}{4}(1) = 3.25$

Optimal match

$K = \max_{\pi : X \to Y} \sum_{x_i \in X} S(x_i, \pi(x_i))$

$= 1(2) + \frac{1}{2}(3) = 3.5$

Slide credit: Kristen Grauman
Summary: Pyramid match kernel

optimal partial matching between sets of features

\[
K_{\Delta} (\Psi(X), \Psi(Y)) = \frac{1}{2^i} \left( I(H_i(X), H_i(Y)) - I(H_{i-1}(X), H_{i-1}(Y)) \right)
\]

difficulty of a match at level i  
number of new matches at level i

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

What about spatial info?
What about spatial info?

• Feature level
  – Spatial influence through correlogram features: Savarese, Winn and Criminisi, CVPR 2006

• Generative models
  – Sudderth, Torralba, Freeman & Willsky, 2005, 2006
  – Niebles & Fei-Fei, CVPR 2007

• Discriminative methods
  – Lazebnik, Schmid & Ponce, 2006
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 2: part-based models
by Rob Fergus (MIT)

Problem with bag-of-words

- All have equal probability for bag-of-words methods
- Location information is important

Overview of section

- Representation
  - Computational complexity
  - Location
  - Appearance
  - Occlusion, Background clutter
- Recognition
Model: Parts and Structure

Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

Figure from [Fischler & Elschlager 73]

History of Parts and Structure approaches

- Fischler & Elschlager 1973
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Felzenszwalb & Huttenlocher ‘00, ‘04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ‘03, ’04
- Many papers since 2000

Sparse representation

+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition

- Throw away most image information
- Parts need to be distinctive to separate from other classes
Region operators

- Local maxima of interest operator function
- Can give scale/orientation invariance

Figures from [Kadir, Zisserman and Brady 04]

The correspondence problem

- Model with P parts
- Image with N possible assignments for each part
- Consider mapping to be 1-1

- $N^P$ combinations!!!

Connectivity of parts

- Complexity is given by size of maximal clique in graph
- Consider a 3 part model
  - Each part has set of N possible locations in image
  - Location of parts 2 & 3 is independent, given location of L
  - Each part has an appearance term, independent between parts.

Variable graph

Shape Model

- L

Factor graph

- L
- S(L)
- S(L,2)
- S(L,3)
- A(L)
- A(2)
- A(3)

Factors

Shape

Appearance
Different connectivity structures

- O(N^6)
- O(N^2)
- O(N^3)

Fergus et al. '03
Fei-Fei et al. '03
Crandall et al. '05
Fergus et al. '05
Felzenszwalb & Huttenlocher '00
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandall et al. '05
Fergus et al. '05
Crandal
Dense layout of parts
Layout CRF: Winn & Shotton, CVPR ‘06

How to model location?
- Explicit: Probability density functions
- Implicit: Voting scheme

- Invariance
  - Translation
  - Scaling
  - Similarity/affine
  - Viewpoint

Explicit shape model
- Cartesian
  - E.g. Gaussian distribution
  - Parameters of model, $\mu$ and $\Sigma$
  - Independence corresponds to zeros in $\Sigma$
  - Burl et al. ’96, Weber et al. ‘00, Fergus et al. ’03

- Polar
  - Convenient for invariance to rotation

Mikolajczyk et al., CVPR ’06

Implicit shape model
- Use Hough space voting to find object
- Leibe and Schiele ‘03,’05

Learning
- Learn appearance codebook
  - Cluster over interest points on training images

- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
  - Centroid is given

Recognition
- Interest Points
- Matched Codebook Entries
- Probabilistic Voting
Multiple view points

Hoiem, Rother, Winn, 3D LayoutCRF for Multi-View Object Class Recognition and Segmentation, CVPR ’07

Thomas, Ferrari, Leibe, Tuytelaars, Schiele, and L. Van Gool. Towards Multi-View Object Class Detection, CVPR 06

Representation of appearance

• Needs to handle intra-class variation
  – Task is no longer matching of descriptors
  – Implicit variation (VQ to get discrete appearance)
  – Explicit model of appearance (e.g. Gaussians in SIFT space)

• Dependency structure
  – Often assume each part’s appearance is independent
  – Common to assume independence with location

Representation of appearance

• Invariance needs to match that of shape model

• Insensitive to small shifts in translation/scale
  – Compensate for jitter of features
  – e.g. SIFT

• Illumination invariance
  – Normalize out

Appearance representation

• SIFT

• PCA

• Decision trees
  [Lepetit and Fua CVPR 2005]
Background clutter

- Explicit model
  - Generative model for clutter as well as foreground object
- Use a sub-window
  - At correct position, no clutter is present

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

Demo Web Page

Learning situations

- Varying levels of supervision
  - Unsupervised
  - Image labels
  - Object centroid/bounding box
  - Segmented object
  - Manual correspondence (typically sub-optimal)
- Generative models naturally incorporate labelling information (or lack of it)
- Discriminative schemes require labels for all data points
Learning using EM

- Task: Estimation of model parameters
- Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to parts
- Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters

Example scheme, using EM for maximum likelihood learning

1. Current estimate of $\theta$
2. Assign probabilities to constellations
3. Use probabilities as weights to re-estimate parameters. Example: $\mu$
   
Learning Shape & Appearance simultaneously

Fergus et al. ’03

Last part: datasets and object collections
## Collecting datasets (towards $10^{6-7}$ examples)

- **ESP game (CMU)**
  Luis Von Ahn and Laura Dabbish 2004

- **LabelMe (MIT)**
  Russell, Torralba, Freeman, 2005

- **StreetScenes (CBCL-MIT)**
  Bileschi, Poggio, 2006

- **WhatWhere (Caltech)**
  Perona et al, 2007

- **PASCAL challenge**
  2006, 2007

- **Lotus Hill Institute**
  Song-Chun Zhu et al 2007

## Labeling with games

- **ESP game** (CMU)
  Luis Von Ahn, Laura Dabbish 2004

- **LabelMe (MIT)**
  Russell, Torralba, Freeman, 2005

- **StreetScenes (CBCL-MIT)**
  Bileschi, Poggio, 2006

- **WhatWhere (Caltech)**
  Perona et al, 2007

- **PASCAL challenge**
  2006, 2007

- **Lotus Hill Institute**
  Song-Chun Zhu et al 2007
The PASCAL Visual Object Classes Challenge 2007

The twenty object classes that have been selected are:

**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

M. Everingham, Luc van Gool, C. Williams, J. Winn, A. Zisserman 2007

---

LabelMe

M. Everingham, Luc van Gool, C. Williams, J. Winn, A. Zisserman 2007

---

Caltech 101 & 256

G. Griffin, A. Holub, P. Perona, 2007

---

How to evaluate datasets?

Fei-Fei, S. Fergus, P. Perona, 2004

Russell, S. Torralba, S. Freman, 2005

How many labeled examples? How many classes? Segments or bounding boxes? How many instances per image? How small are the targets? Variability across instances of the same classes (viewpoint, style, illumination). How different are the images?

How representative of the visual world is? What happens if you nail it?
Summary

- Methods reviewed here
  - Bag of words
  - Parts and structure
  - Discriminative methods
  - Combined Segmentation and recognition

- Resources online
  - Slides
  - Code
  - Links to datasets

List properties of ideal recognition system

- Representation
  - 1000’s categories,
  - Handle all invariances (occlusions, view point, …)
  - Explain as many pixels as possible (or answer as many questions as you can about the object)
  - fast, robust

- Learning
  - Handle all degrees of supervision
  - Incremental learning
  - Few training images

- …