Lecture 4

Representing objects for 2D & 3D object recognition
What’s visual recognition?
Classification:
Does this image contain a building? [yes/no]

Yes!
Classification:
Is this an beach?

No!
Image Search or Indexing

Organizing photo collections
Detection:
Does this image contain a car? [where?]
Detection:
Which object does this image contain? [where?]
Detection:
Accurate localization (segmentation)
Object detection is useful...

Computational photography

Autonomous driving

Security

Surveillance
Categorization vs Single instance recognition

Which building is this? Marshall Field building in Chicago
Visual search and landmarks recognition

Google Goggles
Visual search and landmarks recognition
Detection: Estimating object geometric properties
(3D pose, shape, occlusion patterns, etc...)

Object: Building, 45º pose, 8-10 meters away
It has bricks

Object: Person, back; 1-2 meters away

Object: Police car, side view, 4-5 m away
Estimating object pose and 3D shape is useful

Autonomous driving

Robotic manipulation
Recognition tasks

– Classify/detect object categories
– Classify/Detect specific object instances
– Estimate object geometrical properties
Representation, representation, representation!
Different tasks lead to different representations
Recognition tasks

– Classify/detect object categories
– Classify/Detect specific object instances
– Estimate semantic and geometrical attributes
Recognition tasks

– Classify/detect object categories
– Classify/Detect specific object instances
– Estimate semantic and geometrical attributes
Challenges: viewpoint variation
Challenges: illumination

image credit: J. Koenderink
Challenges: scale
Challenges: deformation
Challenges: occlusion

Magritte, 1957
Challenges: background clutter

Kilmeny Niland. 1995
Challenges: intra-class variation
Scalable to millions of categories
Representation

- which features or descriptors? (SIFT, HOG, …)

Interest operators

Dense, uniformly

Multiple interest operators

Randomly

Image credits: F.-F. Li, E. Nowak, J. Sivic
Representation

- Appearance only
- 2D location and appearance
- 3D location and appearance
Representation

Generative – vs – discriminative
Generative – vs – discriminative

- Generative: Infer a function that can generate (explain) your observations
Generative – vs – discriminative

• Discriminative: Infer a function that can separate (discriminate) your observations
Generative models

• **Naïve Bayes classifier**
  - Csurka Bray, Dance & Fan (2004)

• **Hierarchical Bayesian topic models (e.g. pLSA and LDA)**
  - Object categorization: Sivic et al. (2005), Sudderth et al. (2005)
  - Natural scene categorization: Fei-Fei et al. (2005)

• **Auto-encoders & decoders**

• **Restricted Boltzmann machines (RBM)**
  - P. Smolensky (1986)
Discriminative models

**Nearest neighbor**

- Shakhnarovich, Viola, Darrell 2003
- Berg, Berg, Malik 2005...

**Support Vector Machines**

- Guyon, Vapnik, Heisele, Serre, Poggio...

**Neural networks**

- LeCun, Bottou, Bengio, Haffner 1998
- Rowley, Baluja, Kanade 1998
- ... 

**Latent SVM**

- Structural SVM
- Felzenszwalb 00
- Ramanan 03...

**Boosting**

- Viola, Jones 2001,
  Torralba et al. 2004,
  Opelt et al. 2006,...
Learning (not representation)

- How to learn parameters of generative functions or separating functions?
Learning (not representation)

• How to learn parameters of generative functions or separating functions?

• Level of supervision
  • Noisy labels; image labels; bounding box; manual segmentation; part annotation

• Batch/incremental

• Priors
Recognition (not representation)

– Search strategy:
  – Sliding Windows
    • Viola, Jones 2001
    • Dalal and Bill Triggs, 2005
Recognition (not representation)

- Search strategy:
  - Sliding Windows
  - Bottom-up cues (segmentation)

Felzenszwalb and Huttenlocher, 2004
Recognition (not representation)

- Search strategy:
  - Sliding Windows
  - Bottom-up cues (segmentation)
  - Saliency
Recognition tasks

– Classify/detect object categories
  • Bow models
  • Template-based models
  • Part-based models
  • Hierarchical models
Bag of words models

• Used for image and object classification

• Designed to handle variability due to:
  • View point
  • Illumination
  • Occlusions
  • Intra-class
Inspired by works on document analysis!

• 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
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 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 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 rise of 18% 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.

sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image

Hubel, Wiesel

China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

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 rise of 18% 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.
definition of “BoW”

– Independent features

face

bike

violin
definition of “BoW”

– Independent features
– histogram representation
What’s the main limitation of such a representation?
**Representation**

1. feature detection & description

2. codewords dictionary

3. BOW representation

**recognition**

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

**category decision**
Category models

Class 1

Class N
Discriminative classifiers

category models

model space

Class 1

Class N
Spatial Pyramid Matching

Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. S. Lazebnik, C. Schmid, and J. Ponce.. 2006

\[ H = \begin{bmatrix} H_0^1 & H_1^2 & \ldots & H_4^2 & H_1^3 & \ldots & H_{16}^3 \end{bmatrix} \]

or, \[ H = \text{combination of } H_i^j \text{ with appropriate weights} \]
Discriminative classifiers

category models

model space

Class 1

Class N
Caltech 101

Pyramid matching

![Graph showing mean recognition rate per class vs. number of training examples per class for different methods. The graph includes a legend with various methods and their corresponding lines on the graph.]
Represent an object as a collection of HoG templates

1. Extract fixed-sized window at each position and scale
2. Compute HOG (histogram of gradient) features within each window
3. Score the window with a linear SVM classifier
4. Perform non-maxima suppression to remove overlapping detections with lower scores
Issues with a pyramid-based or template-based representation

- Occlusions
- Truncations
- Deformations
Part Based Representation

- Object as set of parts
- Model:
  - Relative locations between parts
  - Appearance of each part
Deformations
Presence / Absence of Features
Sparse representation

Computationally tractable (10^5 pixels $\rightarrow$ 10^1 -- 10^2 parts)

... but throw away potentially useful image information
Larger discrimination power

Parts can help fine-grained discrimination
Different connectivity structures

- a) Constellation
  - Fergus et al. '03
  - Fei-Fei et al. '03

- b) Star shape
  - Crandall et al. '05
  - Leibe 05; Felzenszwalb 09

- c) $k$-fan ($k = 2$)
  - Crandall et al. '05

- d) Tree
  - Felzenszwalb & Huttenlocher '00

- e) Bag of features
  - Csurka '04
  - Vasconcelos '00

- f) Hierarchy
  - Bouchard & Triggs '05

- g) Sparse flexible model
  - Carneiro & Lowe '06

from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006
Deformable Part Models (DPM)


- root filters (coarse)
- part filters (fine)
- deformation models
Mixture of components

SVM with mixtures of components
Different connectivity structures

a) Constellation [13]

b) Star shape [9, 14]

c) $k$-fan ($k = 2$) [9]

d) Tree [12]

e) Bag of features [10, 21]

Csurka ’04

Vasconcelos ’00

f) Hierarchy [4]

Bouchard & Triggs ’05

g) Sparse flexible model

Carneiro & Lowe ’06

from Sparse Flexible Models of Local Features
Gustavo Carneiro and David Lowe, ECCV 2006
Hierarchical representations

- Pixels $\rightarrow$ Pixel groupings $\rightarrow$ Parts $\rightarrow$ Object
Hierarchical representations

- Deep learning architectures and ConvNets

Fukushima, 1980
LeCun, 1987
Hierarchical representations

• Deep learning architectures and ConvNets

(Lee et al., 2009)
Summary

• Handcrafted representations (features, parts) have been surpassed by learned-based representations via CNNs

• Caveat:
  – a lot of training data are required
  – Images are not actionable
Recognition tasks

- Classify/detect object categories
- Classify/Detect specific object instances
- Estimate semantic and geometrical attributes
Visual search and landmarks recognition

Google Goggles
Key Challenges

Variability due to: View point, illumination, occlusions
Key Challenges

Intra-class variability doesn’t need to be modeled!
1963: Block world

(a) Original picture.

(b) Computer display of picture (reflected by mistake).

(c) Differentiated picture.

(d) Feature points selected.

Larry Roberts
80s: First 3D object detectors

- Marr '78, '82
- Ballard, '81
- Grimson & L.-Perez, '87
- Lowe, '87
- Forsyth et al. '91
- Edelman et al. '91
- Ullman & Barsi, '91
- Rothwell '92
- Linderberg, '94
- Murase & Nayar '94
Feature-based detectors

- Handle severe occlusions
- Fast!

Lowe. ’99, ’04
Detecting food in your kitchen!

Edward Hsiao, Alvaro Collet and Martial Hebert. **Making specific features less discriminative to improve point-based 3D object recognition.** *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2010.

Hsiao, Alvaro Collet and Martial Hebert, **Occlusion Reasoning for Object Detection under Arbitrary Viewpoint**, PAMI 2014
Detecting IKEA furniture!

Common property:
Object representation is simple: features + locations

How do these methods handle extreme viewpoint changes for specific objects?
Basic idea

**Goal:** given a query image $I$, find object model that matches with $I$.

**Model:** collection of points on planar surface.
Recognition

• Find matches between “model” points and “query” points
• Using N matches to fit a transformation (e.g. homography)
• If matches and selected model are correct, the fitting error is small
Recognition

- Find matches between “model” points and “query” points
- Using N matches to fit a transformation (e.g. homography)
- If matches and selected model are correct, the fitting error is small

Verification: The hypothesis generates **high** fitting error
Verification: The hypothesis generates low fitting error.

Recognition

- Find matches between “model” points and “query” points
- Using N matches to fit a transformation (e.g. homography)
- If matches and selected model are correct, the fitting error is small

\[
\begin{align*}
\text{query} & \quad \text{Model 2} \\
\text{Verification: The hypothesis generates low fitting error} & \\
\text{• Generate hypothesis} & \\
\text{• Verify hypothesis} & \\
\text{• Select hypothesis with lowest fitting error} & \\
\text{• Generate recognition results} & 
\end{align*}
\]
Recognition

- Find matches between “model” points and “query” points
- Using N matches to fit a transformation (e.g., homography)
- If matches and selected model are correct, the fitting error is small

Verification: The hypothesis generates low fitting error
• Representation is very simple
  • constellation of features description + locations

• Recognition stage (classification) is much more complex!
  • hypothesis generation and verification

• End-to-end deep learning methods are still on par with traditional methods
Recognition tasks

- Classify/detect object categories
- Classify/Detect specific object instances
- Estimate geometrical properties
Properties of a 3D object detector

- Detect objects under generic view points
- Estimate object pose
- Infer parts and functional properties
Properties of a 3D object detector

- Detect objects under generic view points
- Estimate object pose
- Infer parts and functional properties
- Work at different levels of specificity

Azimuth $\theta$, Zenith $\varphi$
Why is a 3D representation important?
Tea Kettle!

Lid

Spout

Base

Handle
Object parts, they physical properties, and their interrelations in the 3D space are key for object understanding.
Why is a 3D representation important?

- It is easier to update once it is built.
Why is a 3D representation important?

- It is easier to update once it is built.
- It allows predicting the perceptual future.
Why is a 3D representation important?

• It is easier to update once it is built.
• It allows predicting the perceptual future
• It enable robustness to view point changes
Why is a 3D representation important?

- It is easier to update once it is built.
- It allows predicting the perceptual future
- It enable robustness to view point changes
- It makes images of objects actionable
Mixture of 2D models

- Weber et al. ’00
- Schneiderman et al. ’01
- Ullman et al. 02
- Fergus et al. ’03
- Torralba et al. ’03
- Felzenszwalb & Huttenlocher ’03
- Leibe et al. ’04
- Shotton et al. ’05
- Grauman et al. ’05
- Savarese et al., ′06
- Todorovic et al. ’06
- Vedaldi & Soatto ’08
- Zhu et al 08
- Gu & Ren, ′10

CONS: Non scalable to large number of categories/view-points • Just b. boxes • Cannot estimate 3D pose or 3D layout
2D $\frac{1}{2}$ implicit models

- Su, Sun, Fei-Fei, Savarese, CVPR 2009
- Sun, Su, Fei-Fei, Savarese, ICCV 2009
- Thomas et al. ‘06-09
- Kushal, et al., ‘07
- Farhadi ‘09
- Zhu et al. ‘09
- Ozuysal et al. ‘10
- Stark et al.’10
- Payet & Todorovic, 11
- Glasner et al., ‘11

Parts relationship can be probabilistic and learnt automatically
Implicit 3D models – graph-based representations

- Canonical parts captures view invariant diagnostic appearance information
- 2d ½ structure linking parts via weak geometry
- Parts and relationship are modeled in a probabilistic fashion
- Semi-supervised: only class labels, not view point or part annotations
Examples of learnt part-based models

Car
Examples of learnt part-based models

Travel iron
3D explicit models

- Sun, Xu, Bradski, Savarese, ECCV 2010
- Sun, Kumar, Bradski, Savarese, 3DIM-PVT 2011
- Kumar, Sun, Savarese, CVPR 12
- Xiang & Savarese, CVPR 12
- Hoiem, et al., '07
- Chiu et al., '07
- Liebelt et al., '08, 10
- Xiao et al., '08
- Yi et al., 09
- Arie-Nachimson & Barsi ’09
- Sandhu et al., ‘09
- Hu & Zhu ’10

- Enable 6DOF object pose estimation
- 3D layout estimation of object parts
3D explicit models

3D Aspect Layout, Xiang & Savarese, 2012
VDPM, Pepik et al. 2013
3D Voxel patterns, Xiang et al., 2015-2016
Choy et al. 2015

- Part are learning from data and possibly processed through CNNs
Objects become actionable!

<table>
<thead>
<tr>
<th>Object</th>
<th>a</th>
<th>e</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAR</td>
<td>330</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>MOUSE</td>
<td>300</td>
<td>45</td>
<td>23</td>
</tr>
<tr>
<td>SHOE</td>
<td>240</td>
<td>45</td>
<td>11</td>
</tr>
</tbody>
</table>
Objects become actionable!

CHAIR  a=0  e=30  d=7

TABLE  a=60  e=15  d=2

BED    a=30  e=15  d=2.5

SOFA   a=345 e=15  d=3.5

ImageNet dataset [Deng et al. 2010]
Objects become actionable!

Part localization results
Ability to reason about occlusions

[Xiang, et al 2015-2016]
Ability to predict the perceptual future

Xiang, et al., ECCV 2014

Tracking by detection + online tracking of individual parts
Ability to predict the perceptual future

Xiang, et al., ECCV 2014

PROS: Very good understanding the object 3D layout;
CONS: not as accurate as recent CNN-based approaches
Joint detection and pose via CNNs

Results on 3D-ObjectNet

Xiang et al, ECCV 2016

• 100 categories
• 90,127 images
• 201,888 objects

• All annotated with 3D pose and shapes
Joint detection and pose via CNNs


PROS: more accurate, enable very fast detection results
CONS: still large margins for improvements (especially pose); limited 3D
Next lecture

E Simo-Serra, E Trulls, L Ferraz, I Kokkinos, P Fua, F Moreno-Noguer,
Discriminative learning of deep convolutional feature point descriptors, ICCV 2015

Y Xiang, W Choi, Y Lin, S Savarese,
Data-Driven 3D Voxel Patterns for Object Category Recognition, CVPR 2015

R Socher, B Huval, B Bhat, CD Manning, AY Ng,
Convolutional-recursive deep learning for 3d object classification, NIPS 2012