Lecture 14

Announcements

• Midterm review on Friday
• Midterm next Monday!
  • 1:30-3:50pm in this class room
  • previous years midterms are available!
  • Open notes/books but no electronics!
Lecture 14

Visual recognition

• Datasets
• 3D object detection
Caltech 101

Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. L. Fei-Fei, R. Fergus, and P. Perona. CVPR 2004, Workshop on Generative-Model Based Vision. 2004

- Pictures of objects belonging to 101 categories.
- About 40 to 800 images per category. Most categories have about 50 images.
- The size of each image is roughly 300 x 200 pixels.
Caltech 101 images
Caltech-101: Drawbacks

- Smallest category size is 31 images: \( N_{\text{train}} \leq 30 \)

- Too easy?
  - left-right aligned
  - Rotation artifacts
  - Saturated performance
Caltech-256


- Smallest category size now 80 images
- About 30K images

- Harder
  - Not left-right aligned
  - No artifacts
  - More categories

- New and larger clutter category

![Images of various objects]
Largest dataset for object categories up to date

- ~20K categories;
- 14 million images;
- ~700im/categ;
- free to public at www.image-net.org

J. Deng, et al. (2009-2016)
From prof. Li’s group @ Stanford!
IMAGENET is a knowledge ontology

- Taxonomy

- S: (n) Eskimo dog, husky (breed of heavy-coated Arctic sled dog)
  - direct hypernym | inherited hypernym | sister term
  - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
  - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) “the dog barked all night.”
  - S: (n) canine, canid (any of various fissiped mammals with non-retractile claws and typically long muzzles)
  - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) “terrestrial carnivores have four or five clawed digits on each limb”
  - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
  - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
  - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
  - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
  - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
  - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
  - S: (n) living thing, animate thing (a living (or once living) entity)
  - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) “how big is that part compared to the whole?” “the team is a unit”
  - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) “it was full of rackets, balls and other objects”
  - S: (n) physical entity (an entity that has physical existence)
  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman
Dataset Content

- 20 classes: aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV

- Real images downloaded from flickr, not filtered for “quality”

- Complex scenes, scale, pose, lighting, occlusion, ...
Annotations

- Complete annotation of all objects
- Annotated in one session with written guidelines

**Occluded**
Object is significantly occluded within BB

**Truncated**
Object extends beyond BB

**Difficult**
Not scored in evaluation

**Pose**
Facing left
## History

<table>
<thead>
<tr>
<th>Year</th>
<th>Images</th>
<th>Objects</th>
<th>Classes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>2,232</td>
<td>2,871</td>
<td>4</td>
<td>Collection of existing and some new data.</td>
</tr>
<tr>
<td>2006</td>
<td>5,304</td>
<td>9,507</td>
<td>10</td>
<td>Completely new dataset from flickr (+MSRC)</td>
</tr>
<tr>
<td>2007</td>
<td>9,963</td>
<td>24,640</td>
<td>20</td>
<td>Increased classes to 20. Introduced tasters.</td>
</tr>
<tr>
<td>2008</td>
<td>8,776</td>
<td>20,739</td>
<td>20</td>
<td>Added “occlusion” flag.”</td>
</tr>
<tr>
<td>2012</td>
<td>11,530</td>
<td>27,450</td>
<td>20</td>
<td>Added segmentation masks</td>
</tr>
</tbody>
</table>

- **Challenge**: annotation of test set is withheld until after challenge
PASCAL 3D+
Xiang, Mottaghi, Savarese (2014)

- 12 rigid categories from PASCAL VOC are annotated with 3D pose and aligned with 3D cad models
- Benchmark for continuous 3D pose estimation and shape recovery of object categories
ShapeNet3D


• Covers 55 common object categories with about 51,300 unique 3D models.

• 12 object categories of PASCAL 3D+
ObjectNet3D

Xiang, et al., (2016)

- ~100 rigid categories from ImageNet and PASCAL
- Annotated with 3D pose and aligned with 3D cad models from ShapeNet
OpenSurfaces  Bell et al., 2014-2015

<table>
<thead>
<tr>
<th>Materials</th>
<th>Reflectances</th>
<th>Textures</th>
</tr>
</thead>
</table>

+80K annotated images of materials
MSR COCO dataset

• Dataset for image recognition, segmentation, and captioning

• Features:
  – Object segmentation
  – Recognition in Context
  – Multiple objects per image
  – More than 300,000 images
  – More than 2 Million instances
  – 80 object categories
  – 5 captions per image
Knowledge-based dataset that connects structured image concepts to language

108,249 Images
4.2 Million Region Descriptions
1.7 Million Visual Question Answers
2.1 Million Object Instances
1.8 Million Attributes
1.8 Million Relationships
Scene understanding

LabelMe, Russell et al., 2005
3D scene understanding

Stanford 2D-3D-Semantics Dataset, Armeni et al., 2016

6 buildings ~500 rooms ~6000m²
area ~6000 Building Elements

http://buildingparser.stanford.edu/dataset.html
3D scene understanding

Stanford, Princeton, MatterPort 2017

- 10,800 aligned 3D panoramic views (RGB + depth per pixel)
- 194,400 RGB + depth images
- 90-scale scenes.
- Scenes were captured with Matterport’s Pro 3D Camera
More Datasets....

**UIUC Cars (2004)**
S. Agarwal, A. Awan, D. Roth

**CMU/VASC Faces (1998)**
H. Rowley, S. Baluja, T. Kanade

**FERET Faces (1998)**
P. Phillips, H. Wechsler, J. Huang, P. Raus

**COIL Objects (1996)**
S. Nene, S. Nayar, H. Murase

**MNIST digits (1998-10)**
Y LeCun & C. Cortes

**KTH human action (2004)**
I. Leptev & B. Caputo

**Sign Language (2008)**
P. Buehler, M. Everingham, A. Zisserman

**Segmentation (2001)**

**3D Textures (2005)**
S. Lazebnik, C. Schmid, J. Ponce

**CuRRET Textures (1999)**
K. Dana B. Van Ginneken S. Nayar J. Koenderink

**CAVIAR Tracking (2005)**
R. Fisher, J. Santos-Victor J. Crowley

**Middlebury Stereo (2002)**
D. Scharstein R. Szeliski
Lecture 14
Visual recognition

• 3D object detection
  • Introduction
  • Single instance 3D object detectors
  • Generic 3D object detectors
3D object detection
Properties of a 3D object detector

- Detect objects under generic view points
- Estimate object pose & 3D shape
Properties of a 3D object detector

- Detect objects under generic view points
- Estimate object pose & 3D shape
- Work at different levels of specificity
Properties of a 3D object detector

- Detect objects under generic view points
- Estimate object pose & 3D shape
- Work at different levels of specificity
- Limited amount of supervision
Models for 3d Object detection
Models for 3D Object detection

Supervision

Degree of 3D

Intra-class variability

Single instance models

Mixture of 2D single view models
Models for 3d Object detection

Supervision

Degree of 3D

Intra-class variability

Single instance models

2D ½ implicit models

Mixture of 2D single view models
Models for 3d Object detection

- Single instance models
- 3D explicit models
- 2D ½ implicit models
- Mixture of 2D single view models
Models for 3d Object detection

- Single instance models
- 3D explicit models
- 2D 1/2 implicit models
- Mixture of 2D single view models

Variables:
- Supervision
- Degree of 3D
- Intra-class variability
Models for 3d Object detection

Supervision vs. Degree of 3D vs. Intra-class variability

- Single instance models
- 3D explicit models
- 2D ½ implicit models
- Mixture of 2D single view models
Single 3D object recognition
Single 3D object recognition
1963: Block world

(a) Original picture.
(b) Computer display of picture (reflected by mistake).
(c) Differentiated picture.
(d) Feature points selected.
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
Key Challenges

Variability due to:

- View point
- Illumination
- Occlusions
- Arbitrary texture

NOTE: intra-class variability doesn’t need to be modeled
Modern 3D object recognition

Recognition paradigm:
Hypothesis generation & validation

- Rothganger et al. ’04, ’06
- Brown et al, ’05
- Lowe ‘99, ‘04
- Ferrari et al. ’04, ‘06
- Lazebnick et al ’04
- Hsiao et al., ’11-14
- Lim et al., ‘13-16
Object representation: 2D or 3D location of key points

Affine Harris-Laplace detector

- x, y
- Scale
- Orientation

Key idea: use scale and orientation to normalize descriptors
View invariant descriptors

View 1

View 2

Rectification

SIFT
Basic scheme

- Representation
  - Features
  - 2D/3D Geometrical constraints

- Model learning

- Recognition
  - Hypothesis generation
  - Validation
Model learning

Build a 3D model:

- N images of object from N different views
- Extract key points from each view
- Match key points between 2 views
- Use affine structure from motion to compute:
  - Keypoints 3D location and orientation
  - Camera locations from 2 views
- Find connected components
- Use bundle adjustment to refine the model
- Upgrade model to Euclidean assuming zero skew and square pixels
Learnt models

\[ x, y, z + h, v + \text{SIFT descriptor} \]
Basic scheme

- Representation
  - Features
  - 2D/3D Geometrical constraints

- Model learning

- Recognition [object instance from object model]
  - Hypothesis generation
  - Model verification
Recognition

Goal: given a query image I, detect object instance and estimate its pose

Equivalent to: from a collection of learnt object models, find object model that fits object in image

Equivalent to a fitting problem!

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

RANSAC!
Sample set = set of points in 2D

Algorithm:
1. Select random sample of minimum required size to fit model \[?\] =[2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set
Repeat 1-3 until model with the most inliers over all samples is found
Algorithm:
1. Select random sample of minimum required size to fit model [?] 
2. Compute a putative model from sample set 
3. Compute the set of inliers to this model from whole data set 
Repeat 1-3 until model with the most inliers over all samples is found 

\[ |O| = 6 \]
**Algorithm:**

1. Select random sample of minimum required size to fit model $[?] = [2]$
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

$|O| = 14$

Sample set = set of points in 2D
2D model detection

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

**Model:** collection of points on planar surface
2D model detection

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small
2D model detection

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

Verification: The hypothesis generates *high* fitting error
2D model detection

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

Verification: The hypothesis generates low fitting error
2D model detection

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

Verification: The hypothesis generates low fitting error

Model 2

query
2D model detection

- Find matches between “model” points and “query” points
- Using N matches to fit homographic transformation
- If matches and selected model are correct, the fitting error is small

Verification: if the model is wrong, each hypothesis generates **high** fitting errors!
Recognition

Class: apple

1. Find matches between model and test image features
2. Generate hypothesis:
   • Compute transformation $p = MP$, from N matches
Recognition

Class: apple

1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation $p = MP$, from $N$ matches

\[N=2, \text{ if affine camera \& affine key points}\]
Recognition

Class: apple

1. Find matches between model and test image features
2. Generate hypothesis:
   • Compute transformation $p = MP$, from $N$ matches
     $\rightarrow$ Generate hypothesis of object location and pose w.r.t. camera

$p = MP$
Recognition

Class: apple

1. Find matches between model and test image features
2. Generate hypothesis:
   • Compute transformation $p = M \cdot P$, from $N$ matches
   \(N=2\), if affine camera & affine key points
   → Generate hypothesis of object location and pose w.r.t. camera
3. Model verification
   • Use $M$ to project other 3D model features into test image
   • Compute residual = $D$(projections, measurements)
4. Repeat steps 2 and 3 until residual doesn’t decrease anymore
5. Repeat steps 1-4 for different object instance C (apple, teddy bear, etc…)
6. M and C corresponding to min residual return the estimated object pose and object instance
Object to recognize

Matches verified with geometrical constraints

Initial matches based on appearance

Recovered pose

Courtesy of Rothganger et al
Detection and pose estimations results

• Handle severe clutter
3D object detectors

- Handle severe occlusions
- Fast!
Detecting food in your kitchen!


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

Visual search and landmarks recognition

Google Goggles
Visual search and landmarks recognition
Limitations of single instance 3D object detectors

- Cannot handle intra-class variability.

Why?

- Models capture fine-grained details of the object instance which are not shared across instances in the same class
- Hypothesis-generation and verification scheme is not designed to maximize discrimination power
Models for 3d Object detection

Supervision

Degree of 3D

Intra-class variability

- Single instance models
- 3D explicit models
- 2D ½ implicit models
- Mixture of 2D single view models
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

DPM and ISM are examples!

CONS: Single view models are independent • Non scalable to large number of categories/view-points • Just b. boxes • Cannot estimate 3D pose or 3D layout
Models for 3d Object detection

- Single instance models
- 3D explicit models
- 2D \( \frac{1}{2} \) implicit models
- Mixture of 2D single view models

Diagram:
- Supervision axis
- Degree of 3D axis
- Intra-class variability axis
2D ½ implicit models

- Savarese & Fei-Fei, ICCV 07
- Savarese & Fei-Fei, ICCV 07
- 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
Linking features or parts across views:
Perspective or affine transformation constraints

\[ x' = H x \]
Linking features or parts across views:
Epipolar Transformation Constraints

\[ l' = F^T x \]
\[ x' \in l' \]
Implicit 3D models – built upon the ISM

- Sparse set of interest points or parts of the objects are linked across views.
- These links are used to transfer votes across views.
- Each detected codeword votes for the object centroid within nearby views.

Thomas et al. ‘06
Leibe et al. ‘04

Multi-view model
Implicit 3D models – graph-based representations

Savarese, Fei-Fei, ICCV 07
Sun, et al, CVPR 2009, ICCV 09

• 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
Examples of learnt part-based models

Travel iron
Experimental results

• Object detection from any viewing angles
• Accurate estimation of the object pose
• Synthesis of object appearance from unseen view points
Object detection and pose estimation

3D object dataset, 2007
Object detection and pose estimation
Synthesizing novel views

Predicting object appearance from novel views

**PROS:** Flexible and easy to learn • Enable unsupervised discovery of parts

**CONS:** Limited accuracy • Unable to model part configurations in 3D
Models for 3d Object detection

- Single instance models
- 3D explicit models
- 2D ½ implicit models
- Mixture of 2D single view models

Axes:
- Supervision
- Degree of 3D
- Intra-class variability
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

Yan, et al. ’07
• Part configuration is modeled as a 3D conditional random fields with maximal margin parameter estimation
3D object detectors

3D object dataset [Savarese & Fei-Fei, ICCV 07]
3D object detectors

ImageNet dataset [Deng et al. 2010]
3D object detectors

- Part localization on the 3DObject dataset

**PROS:** Large discrimination power; Able to capture part configurations in 3D

**CONS:** Require more supervision; slow...
Next lecture

• 3D scene understanding
Agenda on recognition

Classification (images; areas)
- bag of words
- Pyramid matching

Detection (use slides with 3 axis: category; supervision; 3D info. Use intro job talk)
- Template based (holistic; part based)
- Multi-view (single instance; categories; 3D pose)

Scene understanding
- Segmentation (bottom up; semantic)
- 3D scene understanding

Activity understanding
• Model the object as collection of parts for any T and S on the viewing sphere
A multi-view generative part-based model

\[ \pi \sim \text{Dir}(\alpha) \]
\[ R \sim \text{Mult}(\pi) \]
\[ Y_n \sim \text{Mult}(\eta) \]
\[ \eta = \text{Part Appearance} \]
\[ X_n \sim N(\theta) \]
\[ \theta = \text{Part Location/shape} \]

\( Y_n = \text{Codeword} \)
\( X_n = \text{Location} \)

\[ X_n \leftarrow A \cdot X \]
• Learning: estimate the latent variables and relevant parameters, given the observations
• Variational EM can be used

$\alpha = \text{Part Prop. Prior}$
$\pi \sim Dir(\alpha)$
$R \sim Mult(\pi)$
$Y_n \sim Mult(\eta)$
$\eta = \text{Part Appearance}$
$X_n \sim N(\theta)$
$\theta = \text{Part Location/shape}$

$X_n \leftarrow A \cdot X$

Within triangle constraints:

\[ M_{i \rightarrow j} \cdot m^i \approx m^j \]

Encoded as a penalty term in variational EM.
Incorporating geometrical constraints

View morphing constraints:

\[ m(S) = \sum_{g=1}^{3} m^g_T \cdot s_g \]

\[ W(S) = \sum_{g=1}^{3} W^g_T \cdot s_g \]

\[ m(S) = \text{Center} \]

\[ W(S) = \text{Shape} \]

\[ \begin{align*}
\Sigma &= WW^T \\
\theta &= (m, \Sigma)
\end{align*} \]

Encoded as a penalty term in variational EM

Seitz & Dyer SIGGRAPH 96
Xiao & Shah CVIU '04
Incorporating geometrical constraints

Semi-supervised

- Class label
- Object bounding box
- No part labels
- No pose labels
  [unlike Sun CVPR 09]

- No need to observe same object instance from multiple views
  [unlike Savarese & Fei-Fei, 07, 08]
Incremental learning

- Enable unorganized and on-line collection training images
- Increase efficiency in learning (no need large storage space)
Incremental learning

- Assign new training image to a triangle of the view sphere
- Evidence of training image is used to update model parameters
- Re-estimate sufficient statistics in an iterative fashion
Evolution of learnt parts

Part Evolution
# 3D object detectors

- **Best results up-to-date in pose estimation and 3D part estimation**

## Cars from 3D Object dataset [Savarese 07]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint (cars)</td>
<td><strong>93.4%</strong></td>
<td>85.4</td>
<td>85.3</td>
<td>81</td>
<td>70</td>
<td>67</td>
<td>48.5</td>
</tr>
</tbody>
</table>

## Cars from EPFL dataset [Ozuysal 09]

<table>
<thead>
<tr>
<th>Method</th>
<th>ours</th>
<th>Ours - baseline</th>
<th>DPM [7]</th>
<th>[8]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint (cars)</td>
<td><strong>64.8%</strong></td>
<td>58.1</td>
<td>56.6</td>
<td>41.6</td>
</tr>
</tbody>
</table>

## Chairs, tables and beds from IMAGE NET [Deng et al. CVPR09]

<table>
<thead>
<tr>
<th>Method</th>
<th>ours</th>
<th>Ours - baseline</th>
<th>DPM [7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint (cars)</td>
<td><strong>63.4%</strong></td>
<td>34.0</td>
<td>49.5</td>
</tr>
</tbody>
</table>

---

3D object detectors

- Part localization on the 3DObject dataset