Introduction to object recognition
Course overview

<table>
<thead>
<tr>
<th>Dates</th>
<th>Wednesdays (13:30pm - 14:50pm)</th>
<th>Fridays (13:30pm - 14:50pm)</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/5 &amp; 9/7</td>
<td>Introduction slides</td>
<td>Image filtering slides</td>
<td></td>
</tr>
<tr>
<td>9/12 &amp; 9/14</td>
<td>Image pyramids and Fourier transform slides slides</td>
<td>Hough transform slides</td>
<td></td>
</tr>
<tr>
<td>9/19 &amp; 9/21</td>
<td>Feature and corner detection slides</td>
<td>Feature descriptor and matching slides</td>
<td>Assig. 1 due</td>
</tr>
<tr>
<td>9/26 &amp; 9/28</td>
<td>2D transformations slides</td>
<td>Image homographies slides</td>
<td></td>
</tr>
<tr>
<td>10/3 &amp; 10/5</td>
<td>Camera models slides</td>
<td>Two-view geometry slides</td>
<td>Assig. 2 due</td>
</tr>
<tr>
<td>10/10 &amp; 10/12</td>
<td>Stereo slides</td>
<td>Structure from motion</td>
<td></td>
</tr>
<tr>
<td>10/17 &amp; 10/19</td>
<td>Radiometry and reflectance</td>
<td>Photometric stereo and shape from shading</td>
<td>Assig. 3 due</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/24 &amp; 10/26</td>
<td>Introduction to recognition</td>
<td>Bag of words</td>
<td></td>
</tr>
<tr>
<td>10/31 &amp; 11/2</td>
<td>Tracking</td>
<td>Segmentation</td>
<td>Assig. 4 due</td>
</tr>
<tr>
<td>11/7 &amp; 11/9</td>
<td>Neural networks</td>
<td>Convolutional neural networks 1</td>
<td></td>
</tr>
<tr>
<td>11/14 &amp; 11/16</td>
<td>Convolutional neural networks 2</td>
<td>Color</td>
<td>Assig. 5 due</td>
</tr>
<tr>
<td>11/21 &amp; 11/23</td>
<td>Image processing pipeline</td>
<td>Faces</td>
<td></td>
</tr>
<tr>
<td>11/28 &amp; 11/30</td>
<td>Optical flow</td>
<td>Wrap-up</td>
<td>Assig. 6 due</td>
</tr>
</tbody>
</table>
## Course overview

<table>
<thead>
<tr>
<th>Dates</th>
<th>Wednesdays (13:30pm - 14:50pm)</th>
<th>Fridays (13:30pm - 14:50pm)</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/5 &amp; 9/7</td>
<td>Introduction slides</td>
<td>Image filtering slides</td>
<td></td>
</tr>
<tr>
<td>9/12 &amp; 9/14</td>
<td>Image pyramids and Fourier transforms slides</td>
<td>Hough transform slides</td>
<td></td>
</tr>
<tr>
<td>9/19 &amp; 9/21</td>
<td>Feature and slides</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/26 &amp; 9/28</td>
<td>2D transformations slides</td>
<td>Image homographies slides</td>
<td></td>
</tr>
<tr>
<td>10/3 &amp; 10/5</td>
<td>Camera models slides</td>
<td>Two-view geometry slides</td>
<td></td>
</tr>
<tr>
<td>10/10 &amp; 10/12</td>
<td>Stereo slides</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/17 &amp; 10/19</td>
<td>Radiometry and shading</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Image processing

- Semantic Computer Vision
  - The biggest innovation in Computer Vision
  - Physic-based vision, motion, others…

### Geometry-based Computer Vision
What do we mean by ‘object recognition’?
Is this a street light?
(Verification / classification)
Where are the people? (Detection)
What is this? (Potala palace) (Identification)
What's in the scene?
(semantic segmentation)

Sky

Mountain

Trees

Building

Vendors

People

Ground
Object categorization

- mountain
- tree
- building
- banner
- street lamp
- vendor
- people
What type of scene is it? (Scene categorization)

Outdoor
Marketplace
City
Activity / Event Recognition

what are these people doing?
Object recognition
Is it really so hard?

This is a chair

Find the chair in this image

Output of normalized correlation
Object recognition
Is it really so hard?

Find the chair in this image

Pretty much garbage
Simple template matching is not going to make it

A “popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts.” Nivatia & Binford, 1977.
And it can get a lot harder

How do humans do recognition?

• We don’t completely know yet
• But we have some experimental observations.
Observation 1

- We can recognize familiar faces even in low-resolution images
Observation 2:

Jim Carrey  Kevin Costner

• High frequency information is not enough
What is the single most important facial features for recognition?
What is the single most important facial features for recognition?
Observation 4:

- Image Warping is OK
Spatial configuration matters too
Spatial configuration matters too
The list goes on

Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About

• http://web.mit.edu/bcs/sinha/papers/19results_sinha_etal.pdf
Why is this hard?

Variability: Camera position
Illumination
Shape parameters

Set of Images
How many object categories are there?

~10,000 to 30,000
Challenge: variable viewpoint

Michelangelo 1475-1564
Challenge: variable illumination

image credit: J. Koenderink
Challenge: scale
Challenge: deformation
Deformation
Challenge: Occlusion

Magritte, 1957
Challenge: background clutter

Kilmeny Niland. 1995
Challenge: Background clutter
Challenge: intra-class variations
Common approaches
Common approaches: object recognition

Feature Matching

Spatial reasoning

Window classification

Bag of words
Feature matching
What object do these parts belong to?
Some local features are very informative.

An object as a collection of local features (bag-of-features):

- deals well with occlusion
- scale invariant
- rotation invariant

Are the positions of the parts important?
Why not use SIFT matching for everything?

- Works well for object *instances*
- Not great for generic object *categories*
Pros

• Simple
• Efficient algorithms
• Robust to deformations

Cons

• No spatial reasoning
Common approaches: object recognition

- Feature Matching
- Spatial reasoning
- Window classification
Spatial reasoning
The position of every part depends on the positions of all the other parts

Many parts, many dependencies!
1. Extract features
2. Match features
3. Spatial verification
1. Extract features
2. Match features
3. Spatial verification
1. Extract features
2. Match features
3. Spatial verification

an old idea...
The Representation and Matching of Pictorial Structures

MARTIN A. FISCHLER AND ROBERT A. ELSCIAGER

Abstract—The primary problem dealt with in this paper is the following. Given some description of a visual object, find that object in an actual photograph. Part of the solution to this problem is the specification of a descriptive scheme, and a metric on which to base the decision of "goodness" of matching or detection.

We offer a combined descriptive scheme and decision metric which is general, intuitively satisfying, and which has led to promising experimental results. We also present an algorithm which takes the above descriptions, together with a matrix representing the intensities of the actual photograph, and then finds the described object in the matrix. The algorithm uses a procedure similar to dynamic programming in order to cut down on the vast amount of computation otherwise necessary.

One desirable feature of the approach is its generality. A new programming system does not need to be written for every new description; instead, one just specifies descriptions in terms of a certain set of primitives and parameters.

Description for left edge of face

\[
\text{VALUE}(X) = (E + F + G + H) - (A + B + C + D)
\]

Note: \text{VALUE}(X) is the value assigned to the \text{L(EV)A} corresponding to the location \text{X} as a function of the intensities of locations \text{A through H} in the sensed scene.
Pros

• Retains spatial constraints
• Robust to deformations

Cons

• Computationally expensive
• Generalization to large inter-class variation (e.g., modeling chairs)
Window-based
Template Matching

1. get image window
2. extract features
3. classify

When does this work and when does it fail?
How many templates do you need?
Per-exemplar

exemplar template top hits from test data

find the ‘nearest’ exemplar, inherit its label
Template Matching

1. get image window (or region proposals)
2. extract features
3. compare to template

Do this part with one big classifier ‘end to end learning’
Bag-of-features models
Bag-of-features models

Object → Bag of ‘words’
And one approach rules them all
Convolutional Neural Networks

**Convolution**

Image patch (raw pixels values)

response of one ‘filter’

A 96 x 96 image convolved with 400 filters (features) of size 8 x 8 generates about 3 million values ($89^2 \times 400$)

**Pooling**

Image patch (raw pixels values)

response of one ‘filter’

max/min response over a region

Pooling aggregates statistics and lowers the dimension of convolution
Deep Learning or CNNs

- Since 2012, huge impact..., best results
- Can soak up all the data for better prediction
IMAGENET Large Scale Visual Recognition Challenge

**Year 2010**
- NEC-UIUC
  - Dense grid descriptor: HOG, LBP
  - Coding: local coordinate, super-vector
  - Pooling, SPM
  - Linear SVM
  - [Lin CVPR 2011]

**Year 2012**
- SuperVision
  - [Krizhevsky NIPS 2012]

**Year 2014**
- GoogleLeNet
  - VGG
  - MSRA
  - Convolution
  - Pooling
  - Softmax
  - Other
  - FC
  - softmax
  - conv
  - maxpool
  - [Szegedy arxiv 2014] [Simonyan arxiv 2014] [He arxiv 2014]
History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• Mid-1990s: sliding window approaches
• Late 1990s: local features
• Early 2000s: parts-and-shape models
• Mid-2000s: bags of features
• Present trends: data-driven methods, **deep learning**
PASCAL VOC 2005-2012

20 object classes
Classification: person, motorcycle

Detection

Person

Motorcycle

Action: riding bicycle

Segmentation

22,591 images

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- 20 object categories (aeroplane to TV/monitor)

- Three (+2) challenges:
  - Classification challenge (is there an X in this image?)
  - Detection challenge (draw a box around every X)
  - Segmentation challenge (which class is each pixel?)
Classification Challenge

- Predict whether at least one object of a given class is present in an image

is there a cat?
Pascal VOC 2007 Average Precision
Pascal VOC 2012 Average Precision
Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)
False Positives - Person

UoCTTI_LSVM-MDPM

MIZZOU_DEF-HOG-LBP

NECUUIUC_CLS-DTCT
“Near Misses” - Person

UoCTTI_L SVM-MDPM

MIZZOU_DEF-HOG-LBP

NECUIUC_CLS-DTCT
True Positives - Bicycle

UoCTTI_LSVM-MDPM

OXFORD_MKL

NECUIUC_CLS-DTCT
False Positives - Bicycle

UoCTTI_LSVM-MDPM

OXFORD_MKL

NECUIUC_CLS-DTCT
Where to from here?

• Scene Understanding
  • Big data – lots of images
  • Crowd-sourcing – lots of people
  • Deep Learning – lots of compute
24 Hrs in Photos

http://www.kesselskramer.com/exhibitions/24-hrs-of-photos

installation by Erik Kessels
Data Sets

• ImageNet
  – Huge, Crowdsourced, Hierarchical, Iconic objects

• PASCAL VOC
  – Not Crowdsourced, bounding boxes, 20 categories

• SUN Scene Database, Places
  – Not Crowdsourced, 397 (or 720) scene categories

• LabelMe (Overlaps with SUN)
  – Sort of Crowdsourced, Segmentations, Open ended

• SUN Attribute database (Overlaps with SUN)
  – Crowdsourced, 102 attributes for every scene

• OpenSurfaces
  – Crowdsourced, materials

• Microsoft COCO
  – Crowdsourced, large-scale objects
Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

20 object classes 22,591 images
1000 object classes 1,431,167 images

Variety of object classes in ILSVRC
Variety of object classes in ILSVRC

Amount of Texture
- Screwdriver
- Hatchet
- Ladybug
- Honeycomb

Color Distinctiveness
- Coffee mug
- Cleaver
- Bagel
- Red Wine

Shape Distinctiveness
- Jigsaw Puzzle
- Foreland
- Lipstick
- Bell

Real-world Size
- Orange
- Mask
- Parachute
- Airliner
References

Basic reading:
• Szeliski, Chapter 14.