Visual object recognition and localization

Part 1: Introduction to visual recognition

Ivan Laptev
ivan.laptev@inria.fr
INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d’Informatique, Ecole Normale Supérieure, Paris, France

Includes slides from Steven M. Seitz and A. Torralba
What is Computer Vision?
What is Computer Vision?
Computer vision grand challenge: Dynamic scene understanding

Objects:
cars, glasses, people, etc...

Actions:
drinking, running, door exit, car enter, etc...

Scene categories:
indoors, outdoors, street scene, etc...

Geometry:
Street, wall, field, stair, etc...

Constraints
Why visual recognition?
• Image and video data is exploding (cheap cameras, easy storage and sharing)

~2.5 Billion new images / month

~5K image uploads every min.

~30M surveillance cameras in US

=> ~700K video hours/day

And even more with future wearable devices
Applications

Very large-scale image search

Person identification

Driver-less cars

Robotics
Applications

First appearance of N. Sarkozy on TV

Sociology research: Influence of character smoking in movies

Education: How do I make a pizza?

Motion capture and animation

Predicting crowd behavior

Counting people

Where is my cat?
Current solution: train a *person-throws-cat-into-trash-bin* classifier
What are the challenges?
Object recognition
Is it really so hard?

Find the chair in this image

This is a chair

Output of normalized correlation

Slide credit: A. Torralba
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

Slide credit: A. Torralba
And it can get a lot harder


Slide credit: A. Torralba
Visual variability
Visual variability

View variation

Light variation

Partial visibility

Within-class variation
Many *object* categories

ImageNet: >22,000 object categories
Many scene categories

SUN dataset: ~900 scene categories
Many action categories

Interacting w. computer

Taking a photograph

Playing music

Riding bike

Riding horse

Running

Walking
Subtle differences

Same or different person?

Same

Different
What is the recent progress?
Recent progress

1990s:
Automated quality inspection (controlled lighting, scale, …)

Now:
Facial expression recognition
Gesture recognition
Object and action recognition in natural images and video

Industry

Research

Recognition at the level of a few toy objects (COIL 20 dataset)
Example: Google Goggles
What is the recent progress?
Example: Photo Tourism

[Snavely, Seitz and Szeliski ’06]

Input photographs

Scene reconstruction

Photo Explorer

Relative camera positions and orientations
Point cloud
Sparse correspondence
Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

• 150,000 images from Flickr.com associated with the tags "Rome" or "Roma"
• Matching and reconstruction: 21 hours on a cluster with 496 compute cores
Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

- 150,000 images from Flickr.com associated with the tags "Rome" or "Roma"
- Matching and reconstruction: 21 hours on a cluster with 496 compute cores
Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

- 250,000 Venice images from Flickr.com
- Matching and reconstruction: 27 hours on a cluster with 496 compute cores
Reconstruction and photo explorer
[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

- 250,000 Venice images from Flickr.com
- Matching and reconstruction: 27 hours on a cluster with 496 compute cores
Convolutional Neural Networks

- Success in character recognition [LeCun’88].
- Limited performance on natural images until 2012.

ILSVRC’12: 1.2M images, 1K classes

<table>
<thead>
<tr>
<th>Method</th>
<th>Top 5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT + FVs [7]</td>
<td>26.2%</td>
</tr>
<tr>
<td>1 CNN</td>
<td>—</td>
</tr>
<tr>
<td>5 CNNs</td>
<td>16.4%</td>
</tr>
<tr>
<td>1 CNN*</td>
<td>—</td>
</tr>
<tr>
<td>7 CNNs*</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

2014-2015
- GoogLeNet: 6.6%
- VGG: 6.8%
- Baidu: 5.3%
- Human: 5.1%
Other recent successes of ConvNets: Human pose estimation

DeepPose: Human Pose Estimation via Deep Neural Networks
A. Toshev and C. Szegedy, CVPR 2014
Other recent successes of ConvNets: Face recognition

DeepFace: Closing the Gap to Human-Level Performance in Face Verification
Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf, CVPR 2014

Figure 3. The ROC curves on the LFW dataset. Best viewed in color.
Summary of the course

Problems:

1. Instance-level recognition
   
   Today

2. Category-level recognition
   Object classes
   as well as scene and action classes
   
   Tomorrow

Methods:

- Local features,
  Bag-of-Features (BOF) representations

- Geometric verification,
  Indexing

- Machine learning:
  Support Vector Machines (SVM)
  Convolutional Neural Nets (CNN)
Practical sessions:

If you are not fluent in Matlab, go through this tutorial: