Visual Recognition & Search

January 22, 2009

Introductions

• Class: Thursday 3:30-6:30 PM
• Instructor: Kristen Grauman
  grauman at cs.utexas.edu
  CSA 114
• Office hours: by appointment
• TA: Harshdeep Singh
• Class page: link from
  http://www.cs.utexas.edu/~grauman/
  Check for updates to schedule.
Plan for today

- Topic overview: What is visual recognition and search? Why are these hard problems? What sorta works?

- Course overview: Requirements, syllabus tour
Computer Vision

• Automatic understanding of images and video
  – Computing properties of the 3D world from visual data (*measurement*)
  – Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. (*perception and interpretation*)
  – Algorithms to mine, search, and interact with visual data (*search and organization*)

Vision for measurement

[Images of real-time stereo, structure from motion, tracking]
Vision for perception, interpretation

Objects
Activities
Scenes
Locations
Text / writing
Faces
Gestures
Motions
Emotions…

Visual search, organization

Query
Image or video archives
Relevant content
Why recognition and search?

– Recognition a fundamental part of perception
  • e.g., robots, autonomous agents

– Organize and give access to visual content
  • Connect to information
  • Detect trends and themes

• Why now?

Vision in 1963

Today: visual data in the wild

- Personal photo albums
- Movies, news, sports
- Surveillance and security
- Medical and scientific images

Slide credit: L. Lazebnik

Today: visual data in the wild

- 350 mil. photos, 1 mil. added daily
- 1.6 bil. images indexed as of summer 2005
- 916,271 titles
- 10 mil. videos, 65,000 added daily

Slide by Lana Lazebnik
Autonomous agents able to detect objects


Linking to info with a mobile device

Situated search
Yeh et al., MIT

MSR Lincoln

kooaba
Finding visually similar objects

Exploring community photo collections

- Snively et al.

Simon & Seitz
Discovering visual patterns

Plan for today

- Topic overview: **What is visual recognition and search? Why are these hard problems?** What sorta works?

- Course overview: Requirements, syllabus tour
The Instance-Level Recognition Problem

John’s car

The Categorization Problem

- How to recognize ANY car
Levels of Object Categorization

- Different levels of recognition
  - *Which* object class is in the image? ⇒ Obj/Img classification
  - *Where* is it in the image? ⇒ Detection/Localization
  - *Where exactly* — which pixels? ⇒ Figure/Ground segmentation

Object Categorization

- Task Description
  - “Given a small number of training images of a category, recognize a-priori unknown instances of that category and assign the correct category label.”

- Which categories are feasible visually?
Visual Object Categories

- Basic Level Categories in human categorization
  [Rosch 76, Lakoff 87]
  - The highest level at which category members have similar perceived shape
  - The highest level at which a single mental image reflects the entire category
  - The level at which human subjects are usually fastest at identifying category members
  - The first level named and understood by children
  - The highest level at which a person uses similar motor actions for interaction with category members

Visual Object Categories

- Basic-level categories in humans seem to be defined predominantly visually.
- There is evidence that humans (usually) start with basic-level categorization before doing identification.

⇒ Basic-level categorization is easier and faster for humans than object identification!
⇒ How does this transfer to automatic classification algorithms?
How many object categories are there?

~10,000 to 30,000

Source: Fei-Fei Li, Rob Fergus, Antonio Torralba.

Biederman 1987
Other Types of Categories

- Functional Categories
  - e.g. chairs = “something you can sit on”

- Ad-hoc categories
  - e.g. “something you can find in an office environment”
Challenges: robustness

Illumination
Object pose
Clutter
Occlusions
Intra-class appearance
Viewpoint

Realistic scenes are crowded, cluttered, have overlapping objects.
Challenges: importance of context

slide credit: Fei-Fei, Fergus & Torralba

Challenges: importance of context
Challenges: complexity

- Thousands to millions of pixels in an image
- 3,000-30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images indexed by Google Image Search
- 18 billion+ prints produced from digital camera images in 2004
- 295.5 million camera phones sold in 2005
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

Challenges: learning with minimal supervision
What “works” today

• Reading license plates, zip codes, checks

Source: Lana Lazebnik

What “works” today

• Reading license plates, zip codes, checks
• Fingerprint recognition

Source: Lana Lazebnik
What “works” today

• Reading license plates, zip codes, checks
• Fingerprint recognition
• Face detection

Source: Lana Lazebnik
• Active research area with exciting progress!

Today’s challenge
This course

• Focus on current research in
  – visual category and object recognition
  – image/video retrieval
  – organization, exploration, interaction with visual content

• High-level vision and learning problems, innovative applications.

Goals

• Understand current approaches
• Analyze
• Identify interesting research questions
Expectations

- **Discussions** will center on recent papers in the field
  - Paper reviews
- **Student presentations**
  - Papers and background reading
  - Demos
- **Projects**
  - Research-oriented
- **Workload = reasonably high**

Prerequisites

- Courses in:
  - Computer vision
  - Machine learning
  - Basic probability
  - Linear algebra

- Ability to analyze high-level conference papers
Paper reviews

• For each class, review two of the assigned papers.
• Post by Wed night 10 PM on Google docs (instructions are on Blackboard)
• Don’t review papers the week(s) you are presenting.

Paper review guidelines

• Brief (2-3 sentences) summary
• Main contribution
• Strengths? Weaknesses?
• How convincing are the experiments? Suggestions to improve them?
• Extensions?
• Additional comments, unclear points
• Relationships observed between the papers we are reading
• ½ page to 1 page.
Presentation guidelines

• Read 3-4 selected papers in topic area
• Well-organized talk, about 30 minutes

• What to cover?
  – Problem overview, motivation
  – Algorithm explanation, technical details
  – Any commonalities, important differences between techniques covered in the papers.

• See class webpage for more details.

Demo guidelines

• Implement/download code for a main idea in the paper and show us toy examples:
  – Experiment with different types of (mini) training/testing data sets
  – Evaluate sensitivity to important parameter settings
  – Show (on a small scale) an example in practice that highlights a strength/weakness of the approach

• Present in class – about 20-30 minutes.
• Post webpage with links to any tools or data.
Timetable for presenters

• By the Thursday the week before your presentation is scheduled:
  – Email draft slides to me, and schedule a time to meet and discuss.

• The week of your presentation:
  – Refine slides, practice presentation, know about how long each part requires.

• The day of your presentation:
  – Send final slides (and, for demos, pointer to webpage) to me.

Presenter feedback

• Preparedness
• Coverage of topic
• Organization and clarity of presentation
• Enthusiasm, use of engaging examples
• Serves to start discussion, quality of discussion points raised
Demo feedback

• Preparedness
• Clarity of message and organization
• Technical detail and relevance to reading
• Enthusiasm, use of engaging examples

Projects

Possibilities:
  – Extend a technique studied in class
  – Analysis and empirical evaluation of a technique
  – Comparison between two approaches
  – Design and evaluate a novel approach

• Work in pairs
Grading policy

• 20% participation
  – includes attendance and paper reviews
• 20% demo
• 20% paper presentation
• 40% project

Important dates

• March 26 : project proposals due (tentative)
• April 16 : project progress report / draft (tentative)
• May 7 : Final project papers due
• May 7 and May 8 : Final presentations
  – May 8 is Friday after last class.
Syllabus tour

I. Categorizing and matching objects
II. Surrounding cues
III. Data-driven visual learning
IV. Searching and browsing visual content

Sliding windows and global representations

- Sliding window protocol for detection
- Good features for “patch” appearance, global descriptors
- Building detectors with discriminative classifiers
- Faces, pedestrians as case studies

• (Next week)
Distances and kernels, bags of words representations

Local features: interest operators and descriptors

How to summarize local content?
How to *match* or compare images with local descriptors?

Distances and kernels, bags of words representations

- Constructing a visual “vocabulary”
Distances and kernels, bags of words representations

Local features: interest operators and descriptors

Correspondence kernels
• how to compute matches efficiently?

Learning feature significance
• which features are most discriminative?
Part-based models

- Representing part appearance plus structure
- Summarizing repeated parts
- Efficient matching

Image annotation process

Classifiers can be trained from labeled data…
Image annotation process

- What data should be labeled?
- How can the task be streamlined with semi-automatic tools?
- How can it be more enticing?
- What makes an image dataset useful/not so useful?

Image annotation process

Contribute to LabelMe

LabelMe is to collect contributions from many people so that we can build a large high-quality database for research on object recognition. The following are some basic guidelines for labeling the images.

1. Label as many objects and regions as you can within the same image.
2. Follow the object boundary, ignoring occlusions.
3. Label regions, objects, and parts.

We are interested in objects such as cars, polygons, and tables. But we are also interested in regions such as sky, buildings, sidewalk, walk, etc.

You can also label parts (e.g., the legs of a table, the wheel of a car).

We can use that information later to reason about what objects are part of others by studying how many times they overlap.
Syllabus tour

I. Categorizing and matching objects
II. Surrounding cues
III. Data-driven visual learning
IV. Searching and browsing visual content

Inferring 3D cues from single images

Geometric context is important to scene understanding.

- What are the primary surfaces and their orientations?
- How can this be inferred with a single snapshot?
Scene recognition

Many objects occur only in certain scenes, and scene types are a useful summary of a shot.

- What kind of scene is it? Indoor/outdoor, city/mountain?
- Holistic representations for scenes

Oliva & Torralba

FeiFei & Perona
Context

• The context of the scene, the other objects, and the spatial layout could tell us a lot about what is reasonable to detect.
Context

[Images of various scenes as context]

Context

[Diagram illustrating segment recognizer, semantic context, and spatial context]

Torralba et al.

Galleguillos et al.
Syllabus tour

I. Categorizing and matching objects
II. Surrounding cues
III. Data-driven visual learning
IV. Searching and browsing visual content

Leveraging internet data

- The internet offers unprecedented access to lots of data: both images and surrounding cues.

Mining for themes, connecting to geo-tags

Quack et al.
Leveraging internet data

The value of volume

Dealing with noisy sources

Text, language, and imagery

President George W. Bush makes a state-
ment in the Rose Garden, while Secretary of
Defence Donald Rumsfeld looks on, July 23,
2003. Rumsfeld said the United States would
release graphic photographs of the dead wom-
ans of Saddam Hussein to prove they were killed
by American troops. Photo by Lucy Dano-
ing/Rex

British director Sam Mendes and his part-
tner actress Kate Winslet arrive at the London
premiere of 'The Road to Perdition' on Sep-
tember 18, 2002. The film stars Tom Hanks as
a Chicago hit man who has a sonorous fam-
ily life and co-stars Paul Newman and Ade
Lind. REUTERS/Dan Chung

World number one Lleyton Hewitt of Aus-
tralia hits a return to Novak Maks of Cugit
in Japan Open tennis championships in
Tokyo October 3, 2002. REUTERS/Erkio
Sakag

German supermodel Claudia Schiffer gave
birth to a baby boy by Cesarean section
January 30, 2003, her spokesman said. The
baby is the first child for both Schif-
fer, 32, and her husband, British film pro-
ducer Matthew Vaughn, who was at her side
for the birth. Schiffer is seen on the Ger-
man television show 'Fei Dem!' ('Wetten
Das...?') in Bonn, Germany, on January 26,
2002. (Alexandra Winkler/Zuma)

US President George W. Bush (L) makes his
mark while Secretary of State Colin Pow-
ell (R) looks on before signing the US Leader-
ship Against HIV/AIDS, Tuberculosis and
Malaria Act of 2003 at the Department of
State in Washington, DC, April 2, 2002. The
five-year plan is designed to help prevent and
meet AIDS, especially in more than 40 Ama-
rican and Caribbean nations. (AFP/Luke
Frazza)
Text, language, and imagery

Everingham et al.

Text, language, and imagery

Cour et al.
Unsupervised learning and discovery

- What are common visual patterns?
- What is unusual, or salient?

Syllabus tour

I. Categorizing and matching objects
II. Surrounding cues
III. Data-driven visual learning
IV. Searching and browsing visual content
Fast indexing and search

- With large archives, how to access the relevant content rapidly with good image metrics?

Browsing: query refinement and summarization

- How will a user peruse resulting content efficiently?
- How can a user intervene in the search process?
- Visualizing the aggregation of multiple users’ photos
Browsing: query refinement and summarization

Snavely et al.
Social networks and image tagging

- What information (helpful for recognition) does a community of users provide?
- Why and how do people contribute tags?
- When do they agree? What is objective?
Not covered in this course

• Low-level processing
• Basic machine learning methods

• I will assume you already know these, or are willing to pick them up on your own.

Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>22-Jan</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>29-Jan</td>
<td>Categorising and matching objects</td>
<td>Global appearance, window-based recognition</td>
</tr>
<tr>
<td>5-Feb</td>
<td></td>
<td>Distances and kernels</td>
</tr>
<tr>
<td>12-Feb</td>
<td></td>
<td>Part-based models</td>
</tr>
<tr>
<td>19-Feb</td>
<td></td>
<td>Image annotation process</td>
</tr>
<tr>
<td>26-Feb</td>
<td>Surrounding cues</td>
<td>Inferring 3d cues from a single image</td>
</tr>
<tr>
<td>5-Mar</td>
<td></td>
<td>Scene recognition</td>
</tr>
<tr>
<td>12-Mar</td>
<td></td>
<td>Context</td>
</tr>
<tr>
<td>19-Mar</td>
<td>Spring break - no class</td>
<td></td>
</tr>
<tr>
<td>26-Mar</td>
<td>Data-driven visual learning</td>
<td>Leveraging internet data</td>
</tr>
<tr>
<td>2-Apr</td>
<td></td>
<td>Text, language, and imagery</td>
</tr>
<tr>
<td>9-Apr</td>
<td></td>
<td>Unsupervised learning and discovery</td>
</tr>
<tr>
<td>16-Apr</td>
<td>Searching and browsing visual content</td>
<td>Fast indexing and search</td>
</tr>
<tr>
<td>23-Apr</td>
<td></td>
<td>Browsing: query refinement and summarization</td>
</tr>
<tr>
<td>30-Apr</td>
<td></td>
<td>Social networks and image tagging</td>
</tr>
<tr>
<td>7-May</td>
<td>Final project presentations</td>
<td></td>
</tr>
<tr>
<td>8-May</td>
<td>Final project presentations</td>
<td></td>
</tr>
</tbody>
</table>
For next week

• Read and review:
  – Viola & Jones, CVPR 2001
  – Dalal & Triggs, CVPR 2005
  – **Review syllabus, select topic preferences**
    (3 for demo, 3 for paper topics)
    • Email me by Monday.

• First student presenters will be on Feb 5.