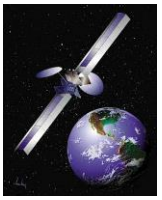




Visual Recognition
Spring 2016






Introductions

- **Instructor**: Prof. Kristen Grauman
- **TA**: Kai-Yang Chiang

Today

- Course overview
- Requirements, logistics

What is computer vision?

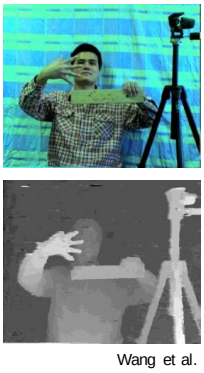
Done?

Computer Vision

- Automatic understanding of images and video
  1. Computing properties of the 3D world from visual data *(measurement)*

1. Vision for measurement

Real-time stereo



Wang et al.

Structure from motion

Snavely et al.

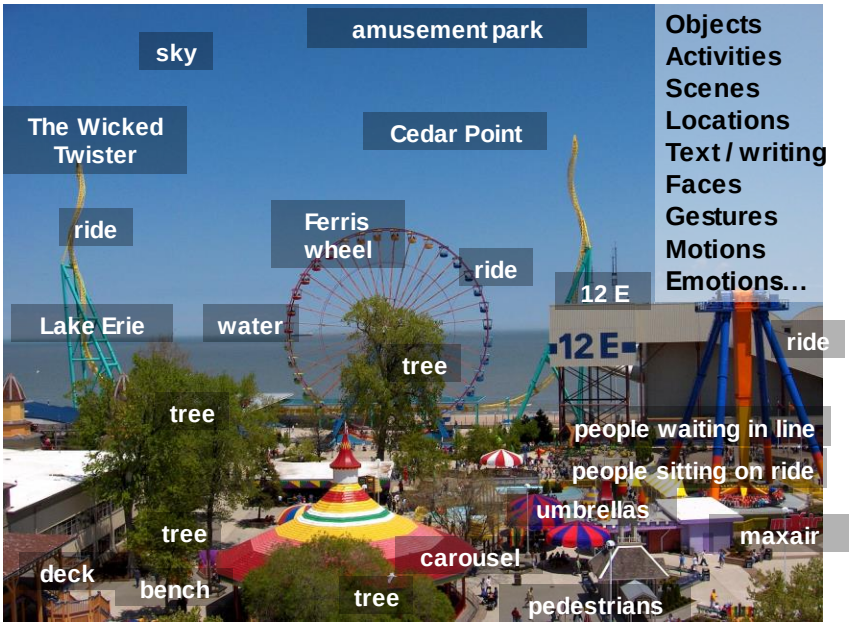
Tracking

Demirdjian et al.

Computer Vision

- Automatic understanding of images and video
  1. Computing properties of the 3D world from visual data *(measurement)*
  2. Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. *(perception and interpretation)*

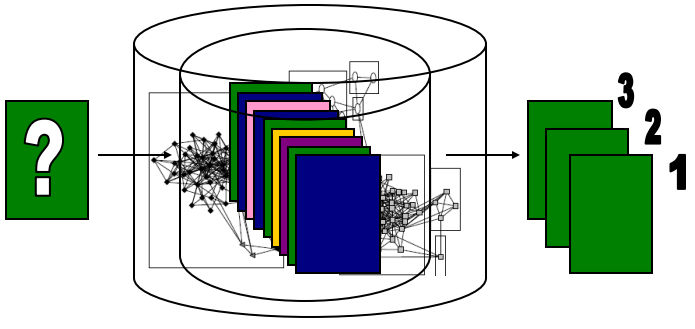
2. Vision for perception, interpretation



Computer Vision

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

3. Visual search, organization

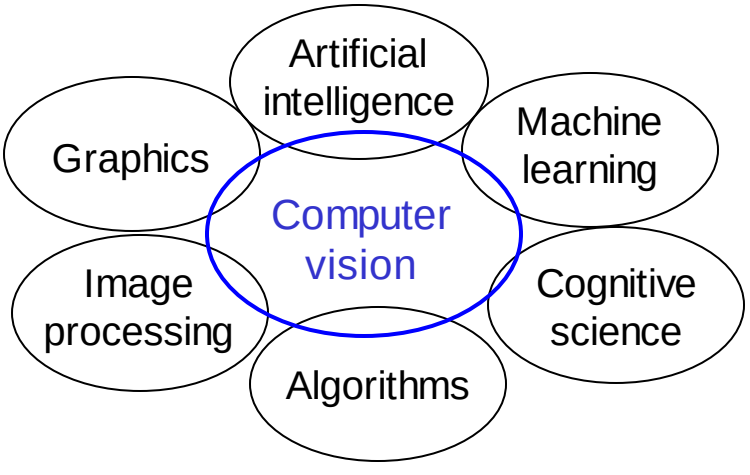


Computer Vision

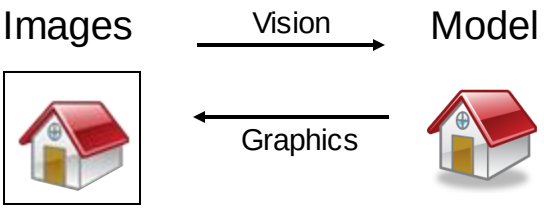
- Automatic understanding of images and video
  1. Computing properties of the 3D world from visual data *(measurement)*
  2. Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities. *(perception and interpretation)*
  3. Algorithms to mine, search, and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

Course focus

Related disciplines

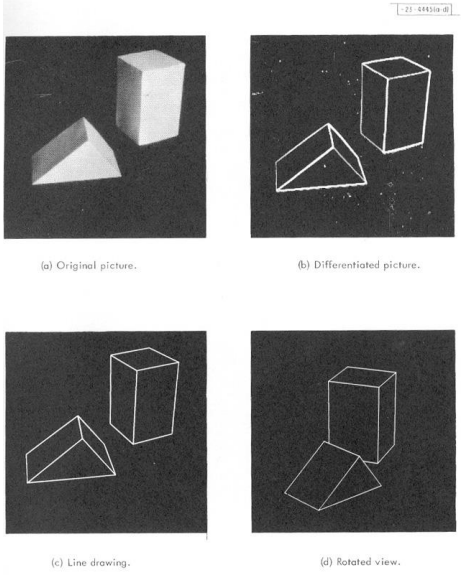


Vision and graphics



Inverse problems: analysis and synthesis.

Visual data in 1963



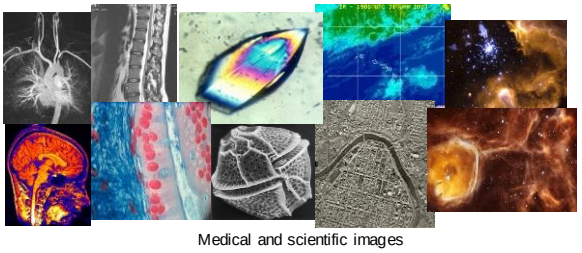
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

Visual data in 2016

Personal photo albums

Movies, news, sports

Surveillance and security



Medical and scientific images

Slide credit; L. Lazebnik

Why recognition?

- 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

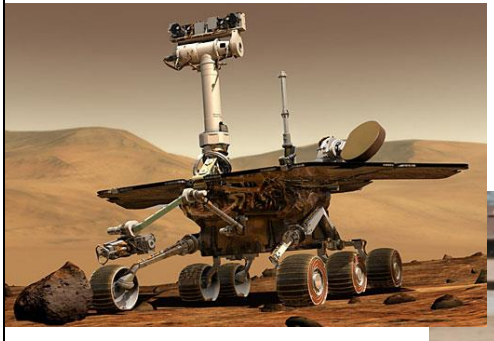
- Why now?

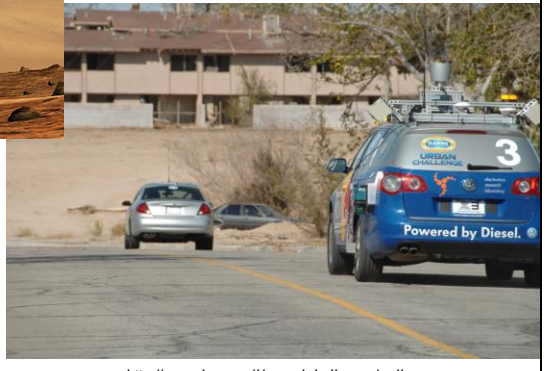
Faces

Camera waits for everyone to smile to take a photo [Canon]

Setting camera focus via face detection

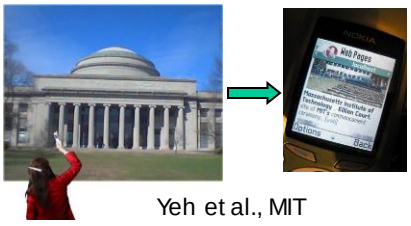
Autonomous agents able to detect objects





http://www.darpa.mil/grandchallenge/gallery.asp

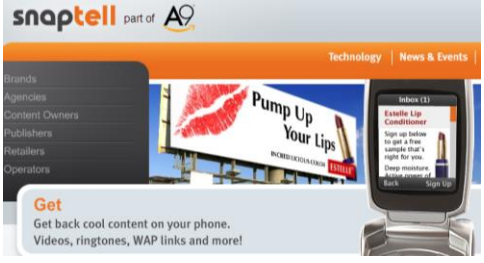
Posing visual queries



Yeh et al., MIT

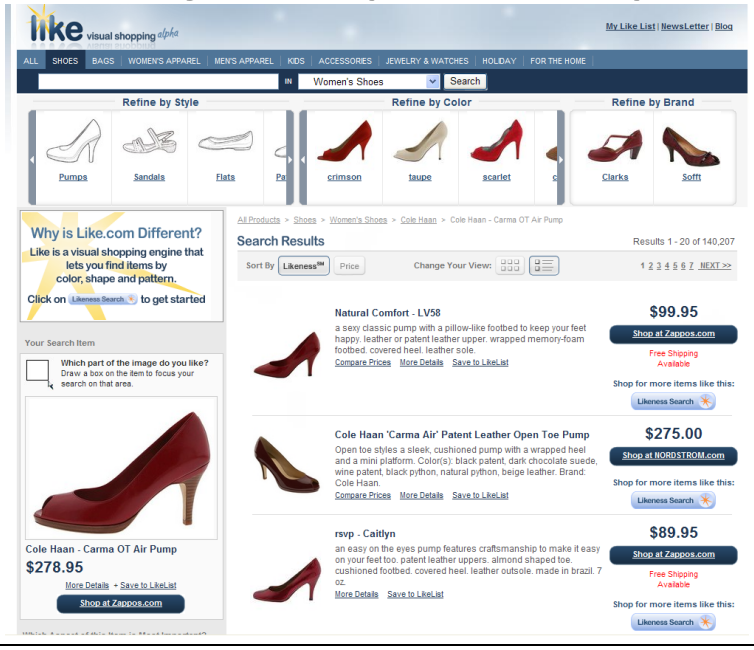
Digital Field Guides Eliminate the Guesswork

Belhumeur et al.

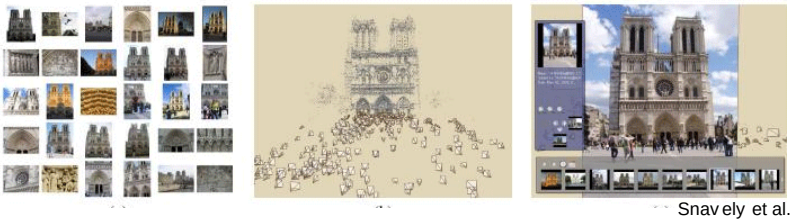


Kooaba, Bay & Quack et al.

Finding visually similar objects



Exploring community photo collections

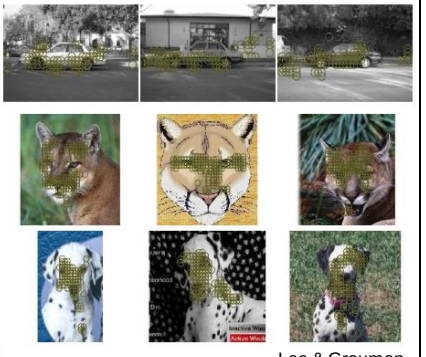


Snavely et al.

Simon & Seitz

Discovering visual patterns

Objects Sivic & Zisserman



Lee & Grauman

Categories

Wang et al.

Actions

Auto-annotation

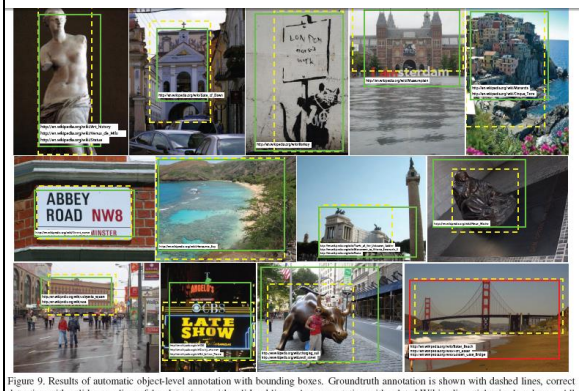
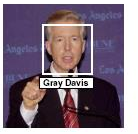


Figure 9. Results of automatic object-level annotation with bounding boxes. Groundtruth annotation is shown with dashed lines, correct detection with solid green lines, false detections with solid red lines. Auto-annotation with related Wikipedia articles is also shown. All results are also labeled with their GPS position and estimated tags (not shown here).

Gammeter et al.

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung



Incumbent California Gov. Gray Davis (news - web site) leads Republican challenger Bill Simon by 10 percentage points – although 17 percent of voters are still undecided, according to a poll released October 22, 2002 by the Public Policy Institute of California. Davis is shown speaking to reporters after his debate with Simon in Los Angeles, on Oct. 7. (Jim Ruymen/Reuters)

T. Berg et al.

Video-based interfaces

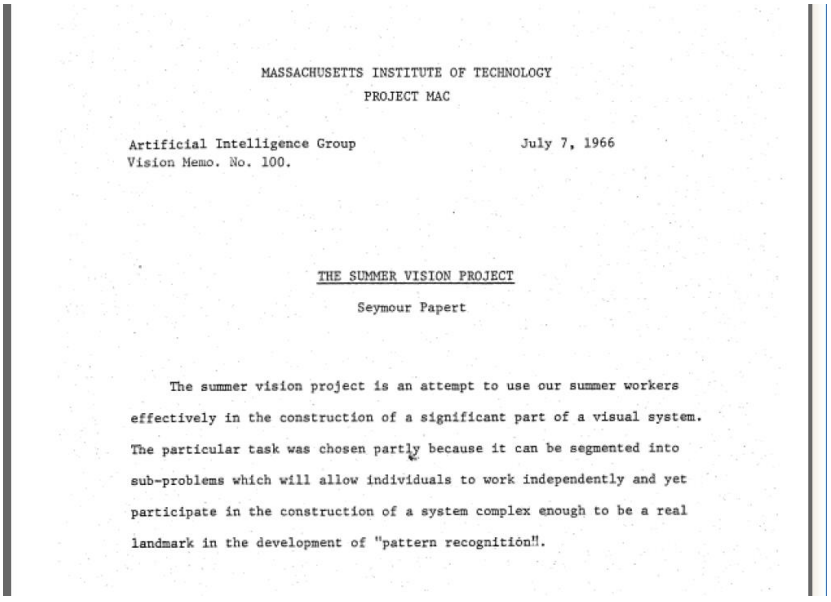
Human joystick, NewsBreaker Live

Assistive technology systems
Camera Mouse, Boston College

Microsoft Kinect

What else?

Obstacles?



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

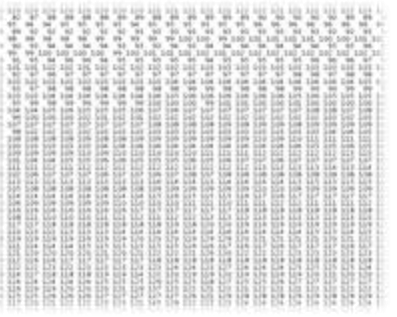
July 7, 1966

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

What the computer gets



Why is vision difficult?

- Ill-posed problem: real world much more complex than what we can measure in images
  - 3D → 2D
- Impossible to literally "invert" image formation process

Challenges: many nuisance parameters

Illumination

Object pose

Clutter

Occlusions

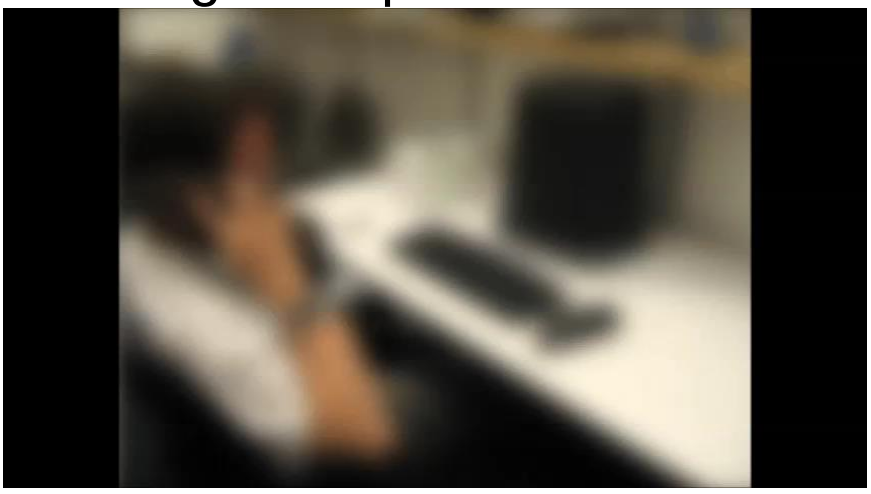
Intra-class appearance

Viewpoint

Challenges: intra-class variation

slide credit: Fei-Fei, Fergus & Torralba

Challenges: importance of context



Video credit: Rob Fergus and Antonio Torralba

Challenges: importance of context

Video credit: Rob Fergus and Antonio Torralba

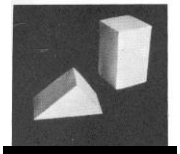
Challenges: importance of context

slide credit: Fei-Fei, Fergus & Torralba

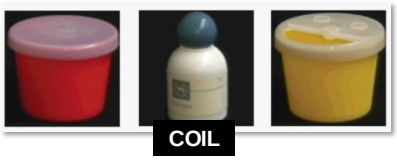
Challenges: complexity

- Millions of pixels in an image
- 30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images online
- 144K hours of new video on YouTube daily
- …
- About half of the cerebral cortex in primates is devoted to processing visual information [Felleman and van Essen 1991]

Progress charted by datasets



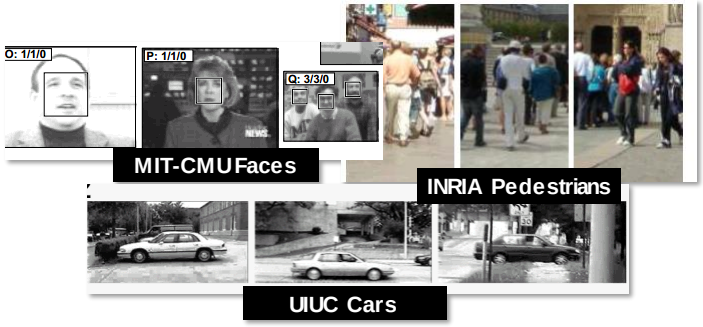
Roberts 1963



COIL

1963 … 1996

Progress charted by datasets



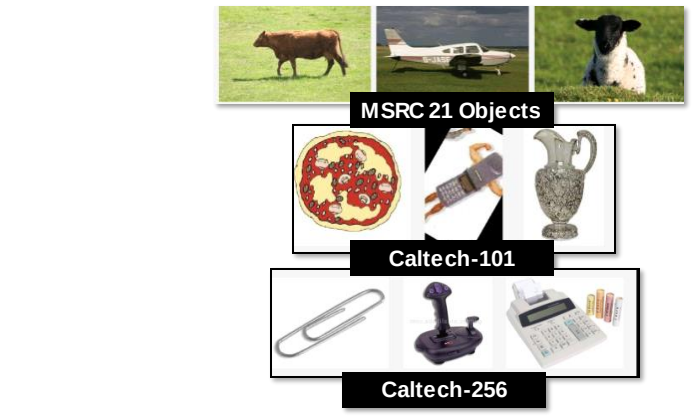
MIT-CMUFaces

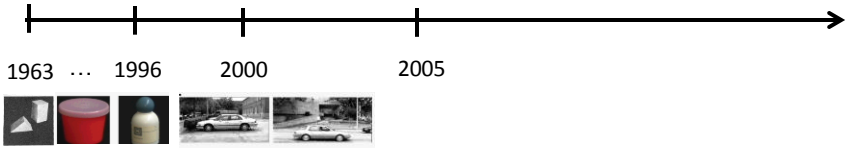
INRIA Pedestrians

UIUC Cars

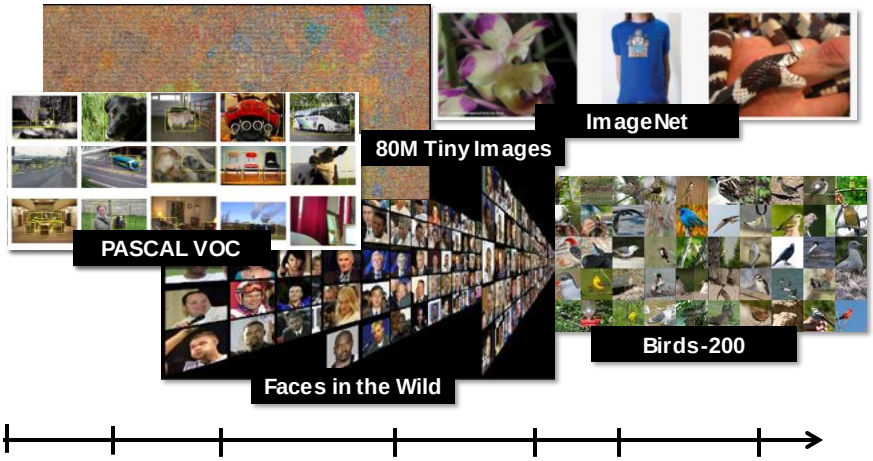
1963 … 1996 2000

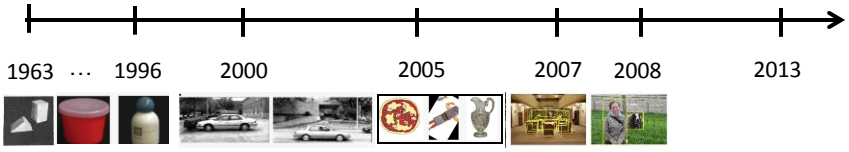
Progress charted by datasets



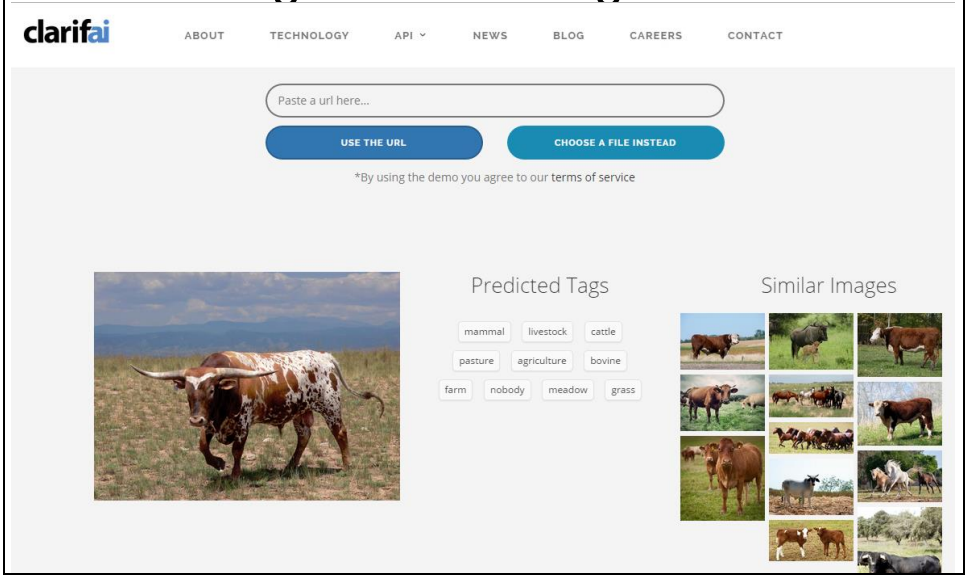


Progress charted by datasets





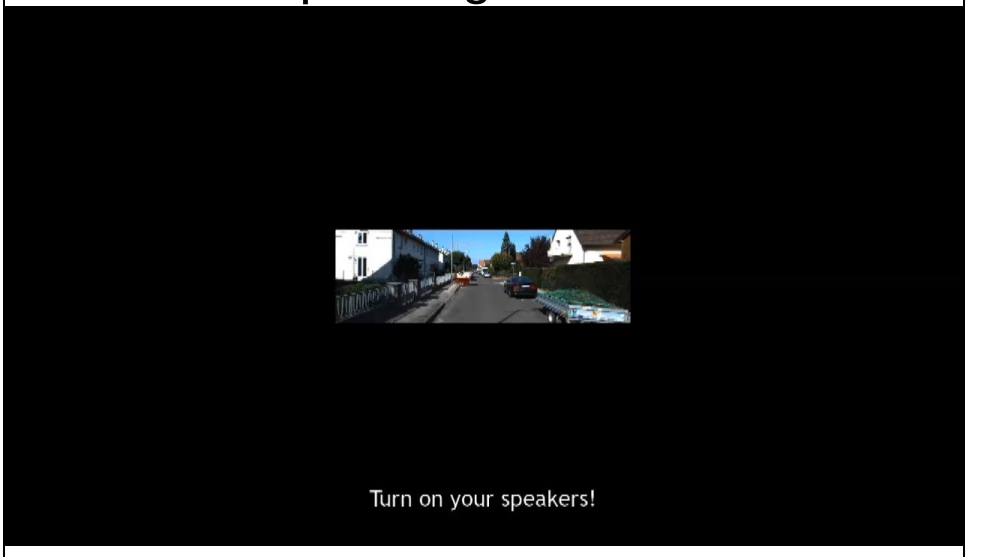
Expanding horizons: large-scale recognition



Expanding horizons: captioning

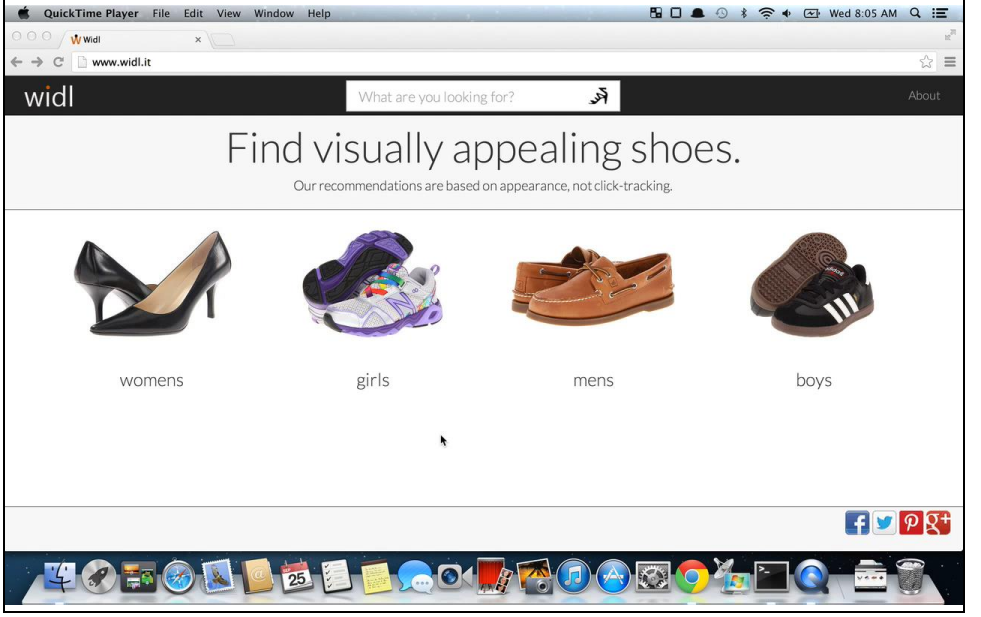
https://pdollar.wordpress.com/2015/01/21/image-captioning/

Expanding horizons:

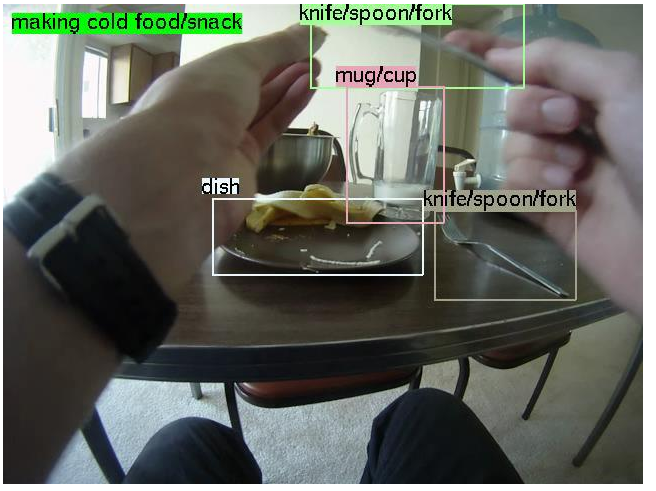


KITTI dataset – Andreas Geiger et al.

Expanding horizons: interactive visual search



Expanding horizons: first-person vision



Activities of Daily Living – Hamed Pirsiavash et al.

This course

- Focus on current research in
  - Object recognition and categorization
  - Image/video retrieval, annotation
  - Some activity recognition
- 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 each week
- **Student presentations**
  - Papers
  - Experiment
- **2 implementation assignments**
- **Project**

Workload is fairly high

Prerequisites

- Courses in:
  - Computer vision
  - Machine learning
- Ability to analyze high-level conference papers

Paper reviews & discussion pts

- Each week, review two of the assigned papers.
- Separately, summarize some "discussion points"
- Post each separately to Piazza following instructions on course "requirements" page.
- Skip reviews 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? What's inspiring?
- Additional comments, unclear points
- Relationships observed between the papers we are reading

Discussion point guidelines

- ~2-3 sentences per reviewed paper
- Recap of salient parts of your reviews
  - Key observations, lingering questions, interesting connections, etc.
- Will be shared to our class via Piazza
- Discussion points required for each class session (due 8 pm Tues)
- All encouraged to browse and post before and after class

Paper presentation guidelines

- Read the selected paper
- Well-organized talk, about 15 minutes

- What to cover?
  - Problem overview, motivation
  - Algorithm explanation, technical details
  - Any commonalities, important differences between techniques covered in the papers.
  - Demos, videos, other visuals etc. from authors

- See handout and class webpage for more details.

Experiment 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 to analyze a strength/weakness of the approach
- Present in class – about 20 minutes.
- Share links to any tools or data.

Timetable for presenters

- For papers or experiments, by the Wednesday **the week before** your presentation is scheduled:
  - Email draft slides to me, and schedule a time to meet, do dry run, discuss.
  - Hard deadline: 5 points per day late
- See course webpage for examples of good reviews, presentations.

Projects

Possibilities:

- Extend a technique studied in class
- Analysis and empirical evaluation of an existing technique
- Comparison between two approaches
- Design and evaluate a novel approach
- Thorough survey / review paper

- Work in pairs, except for survey.

Grades

- Grades will be determined as follows:
  - 25% **participation** (includes attendance, in-class discussions, paper reviews)
  - 15% **coding** assignments
  - 35% **presentations** (includes drafts submitted one week prior, and in-class presentation)
  - 25% **final project** (includes proposal, draft, presentation, final paper)

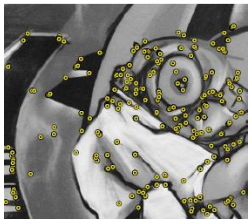
Miscellaneous

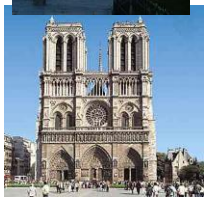
- Feedback welcome and useful!
- Slides, announcements via class website
- Discussion including assignment questions on Piazza
- No laptops, phones, etc. open in class please.

Syllabus tour

- **The core**
  - Instance recognition
  - Category recognition
  - Mid-level representations
  - Object detection
- **Advanced topics**
  - Great outdoors
  - Social signals
  - Noticing and remembering
  - Low-supervision learning
  - 3d scenes and objects
  - Recognition in action
  - Attributes and parts
  - Language and vision

Instance recognition





Local invariant features, detection and description

Matching models to images

Indexing specific objects with bag-of-words descriptors

Category recognition

Recognition as an image classification problem

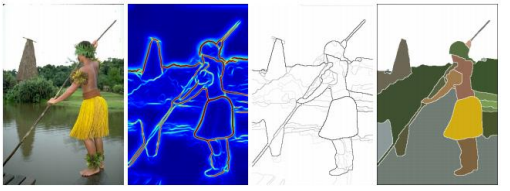
Discriminative methods

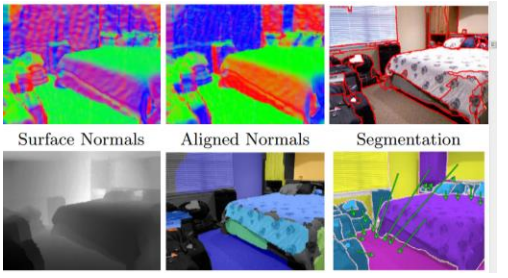
Image descriptors

Convolutional neural networks

Large-scale image collections

Mid-level representations

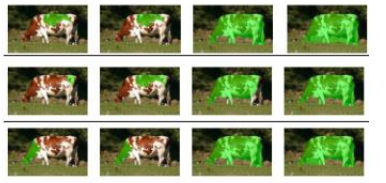




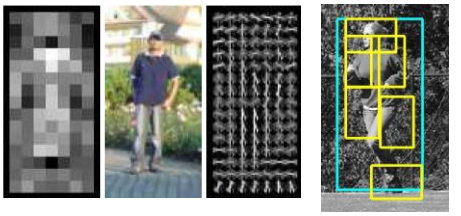
Segmentation

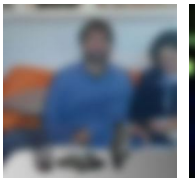
Category-independent region ranking

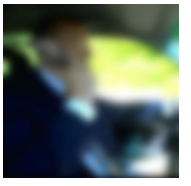
Surface estimation



Object detection







Localizing objects within an image

Efficient search

Part-based models

Semantic segmentation

Voting

Context

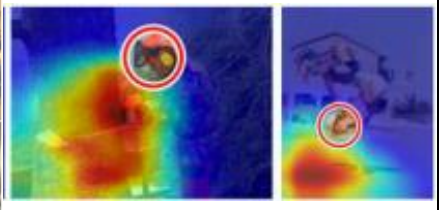
Syllabus tour

- **The core**
  - Instance recognition
  - Category recognition
  - Mid-level representations
  - Object detection
- **Advanced topics**
  - Great outdoors
  - Social signals
  - Noticing and remembering
  - Low-supervision learning
  - 3d scenes and objects
  - Recognition in action
  - Attributes and parts
  - Language and vision

Great outdoors

- Linking and visualizing multi-view data from tourist photos
- Image-based geolocalization
- Natural scene text detection
- Discovering correlated non-visual properties in street-side imagery

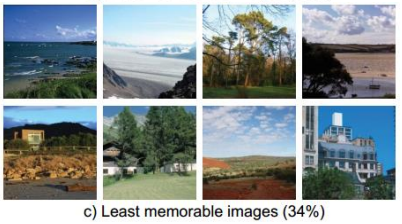
Social signals



- Cues from people in images: body pose, social groups and roles, attention, gaze following, scene structure

Noticing and remembering

a) Most memorable images (86%)



c) Least memorable images (34%)

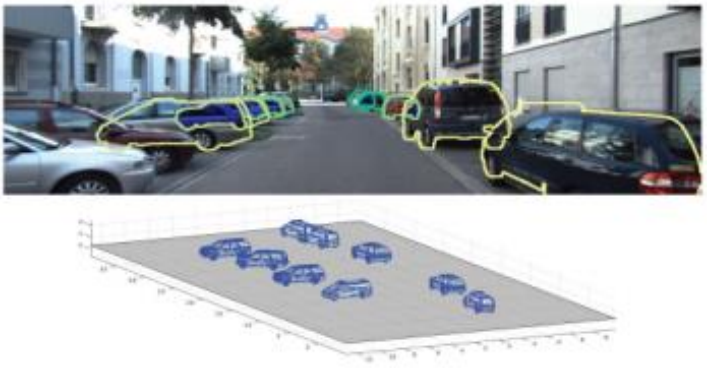
- Predicting what gets noticed or remembered in images and video. Saliency, importance, memorability, photography biases.

Low-supervision learning

- Feature learning, semantics learning. Leveraging free or nearly free cues for supervision. Internet data, video, egomotion, context...

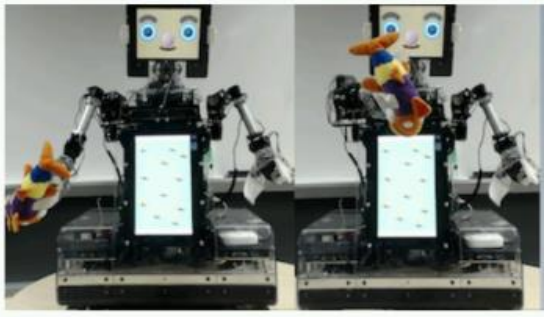
3d scenes and objects

- 3d structure (single views, panoramas, RGBD) and scene layout for visual recognition



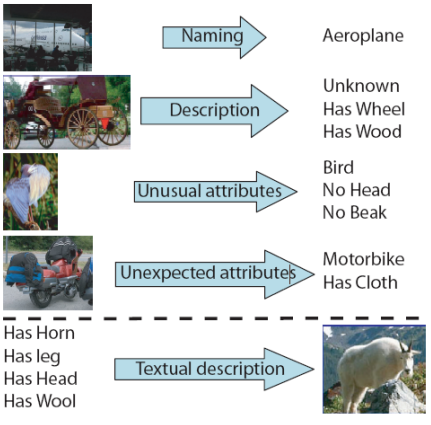
Recognition in action

- Learning how to move for recognition, manipulation. 3D objects and the next best view.



Attributes and parts

Beyond naming object by category, we should be able to describe their properties, or use descriptions to understand novel objects.



Language and vision

- Image-word embeddings, question answering, captioning.

Not covered

- Low-level image processing
- Basic machine learning methods

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

Coming up

- Do reading and paper reviews/discussion point posts for weeks 1 and 2
  - Instance recognition
  - Category recognition
- No class next week