CS 1674: Intro to Computer Vision

Introduction

Prof. Adriana Kovashka
University of Pittsburgh
August 29, 2016

About the Instructor



Born 1985 in
Sofia, Bulgaria



Got BA in 2008 at
Pomona College, CA
(Computer Science &
Media Studies)



Got PhD in 2014
at University of
Texas at Austin
(Computer Vision)

Course Info

- **Course website:** http://people.cs.pitt.edu/~kovashka/cs1674
- **Instructor**: Adriana Kovashka (kovashka@cs.pitt.edu)
  - Please use "CS1674" at the beginning of your Subject
- **Office**: Sennott Square 5129
- **Office hours**: MW, 3:30pm - 4:25pm
- **TA**: ChangSheng Liu (to be confirmed)
- **TA's office hours:** to be determined

Textbooks

- Computer Vision: Algorithms and Applications by Richard Szeliski
- Visual Object Recognition by Kristen Grauman and Bastian Leibe
- More resources available on course webpage

- Your notes from class are your best study material, slides are not complete with notes

Course Goals

- To learn about the basic computer vision tasks and approaches
- To get experience with some computer vision techniques
- To learn absolute basics of machine learning
- To think critically about vision approaches, and to see connections between works and potential for improvement

Plan for Today

- Blitz introductions
- What is computer vision?
  - Why do we care?
  - What are the challenges?
  - What is the current research like?
- Course structure and policies
- Overview of topics (if time)

Blitz Introductions

Blitz Introductions (5-10 sec)

- What is your name?
- Tell us one fun thing about yourself!

(I'll ask you more questions in HW1.)

Computer Vision

What is computer vision?

Done?

"We see with our brains, not with our eyes“ (Oliver Sacks and others)

What is computer vision?

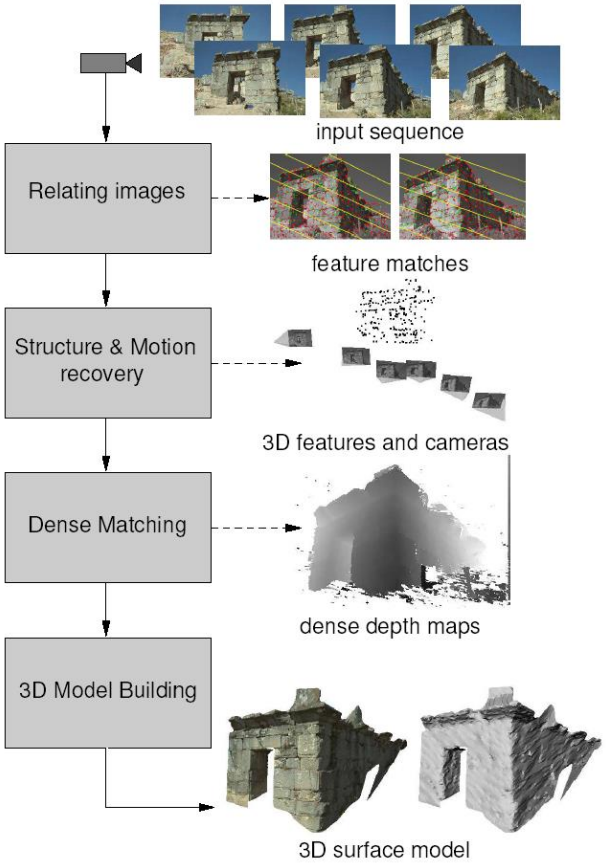
- Automatic understanding of images and video
  - Computing properties of the 3D world from visual data *(measurement)*
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities *(perception and interpretation)*
  - Algorithms to mine, search, and interact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

Vision for measurement

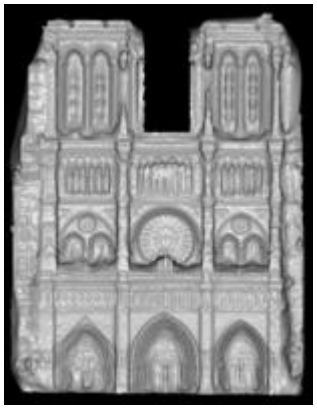
Real-time stereo

Pollefeys et al.

Structure from motion



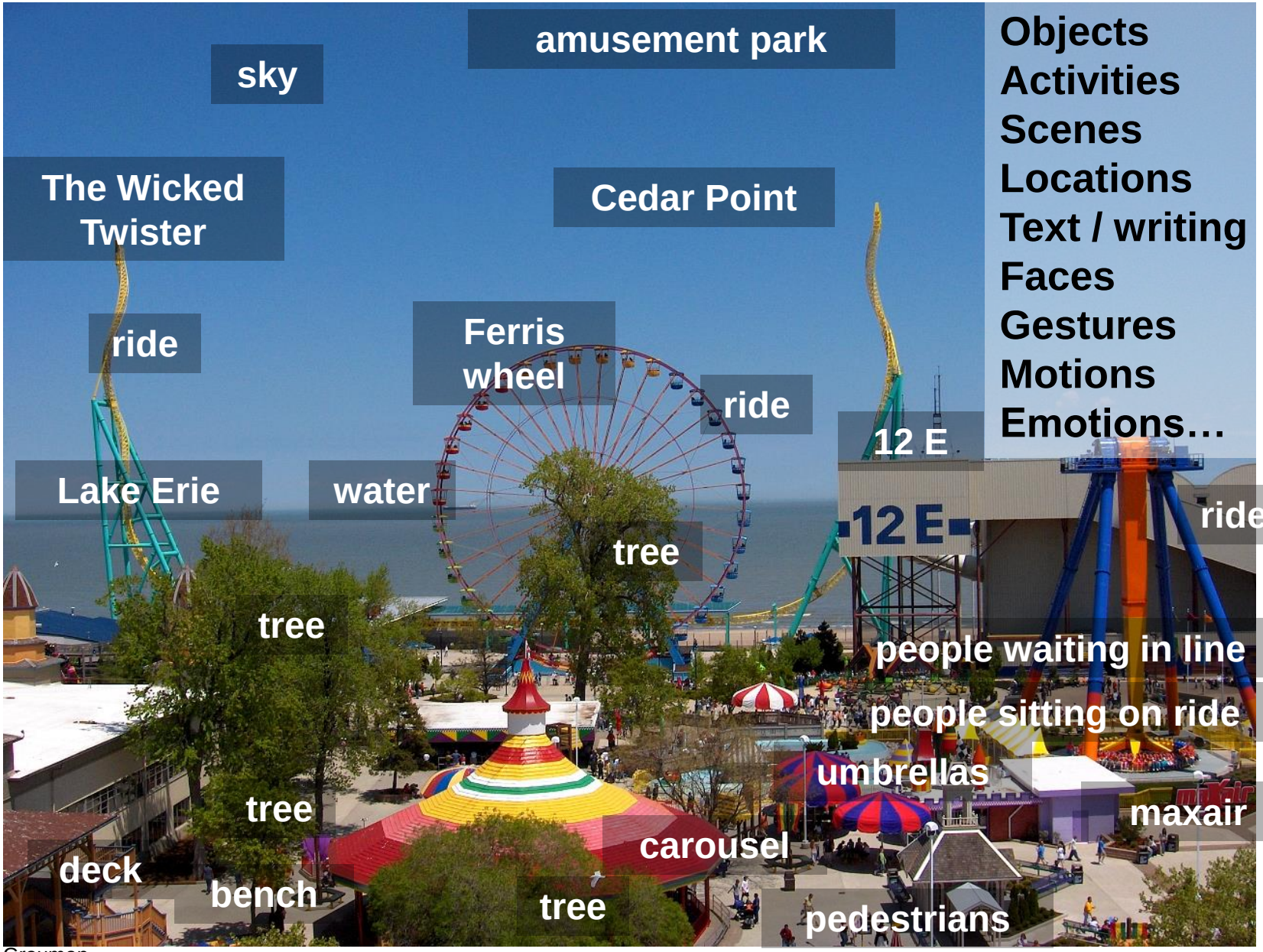
Multi-view stereo for community photo collections



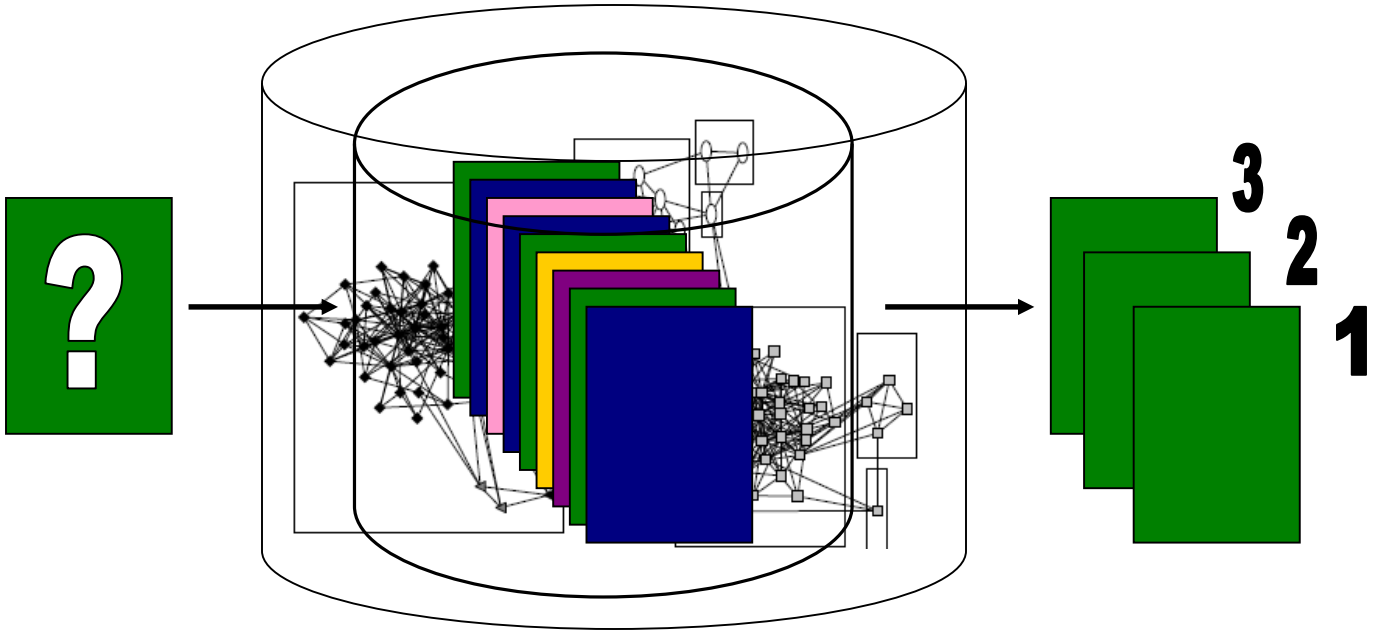
Goesele et al.

Slide credit: L. Lazebnik

Vision for perception, interpretation

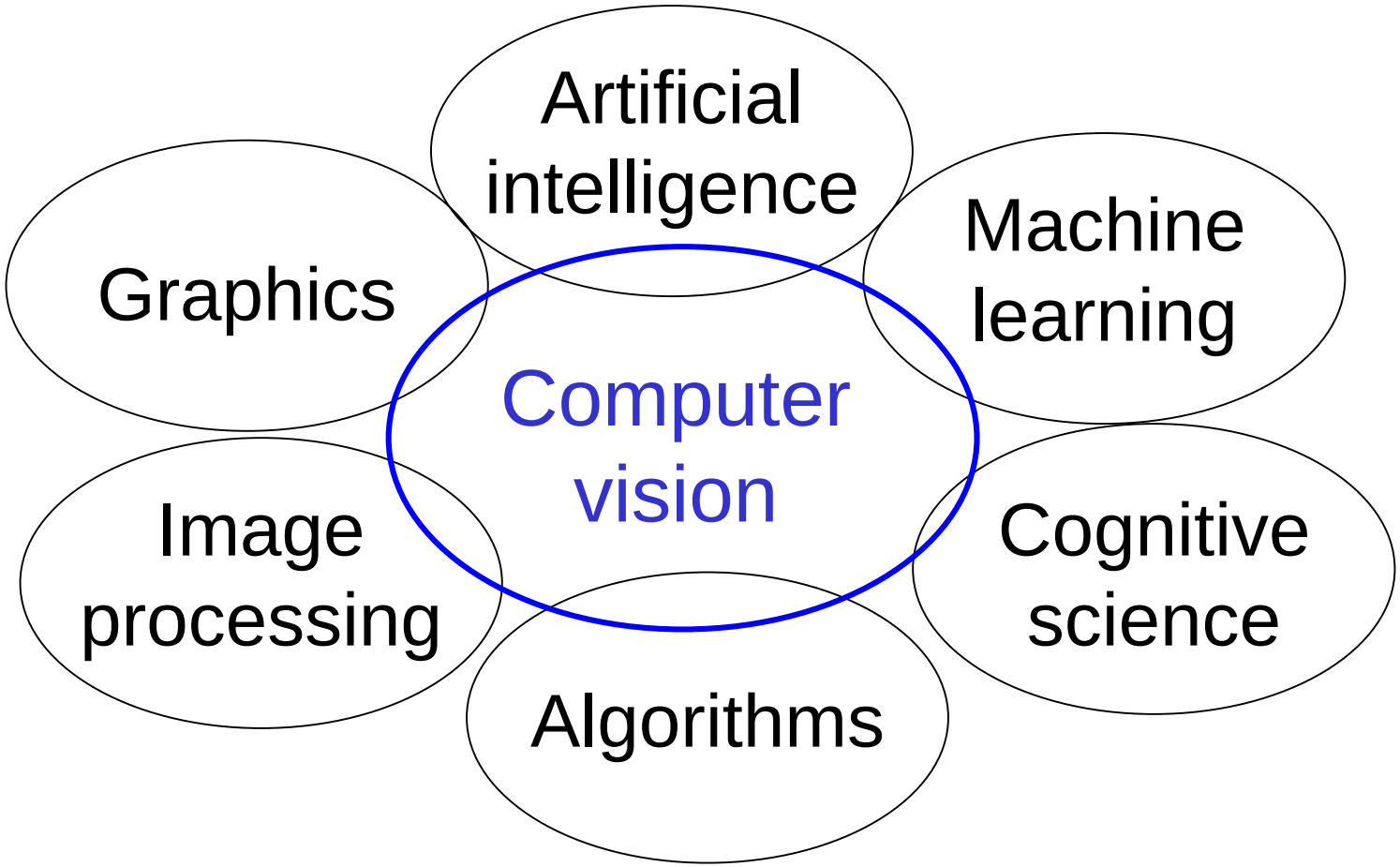


Visual search, organization

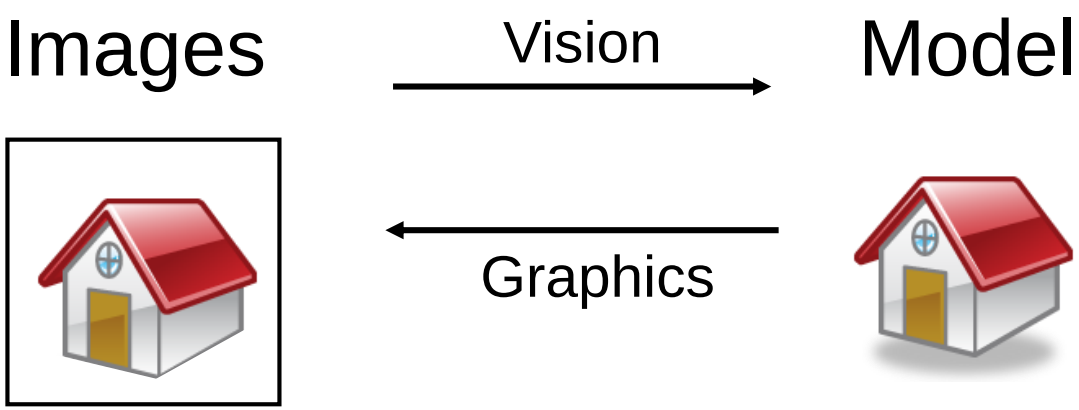


Query **Image or video archives** **Relevant content**

Related disciplines



Vision and graphics



Inverse problems: analysis and synthesis.

Why vision?

- As image sources multiply, so do applications
  - Relieve humans of boring, easy tasks
  - Human-computer interaction
  - Perception for robotics / autonomous agents
  - Organize and give access to visual content
  - Description of image content for the visually impaired
  - Fun applications (e.g. transfer art styles to my photos)

Why vision?

144k hours uploaded to YouTube daily
4.5 mil photos uploaded to Flickr daily
10 bil images indexed by Google

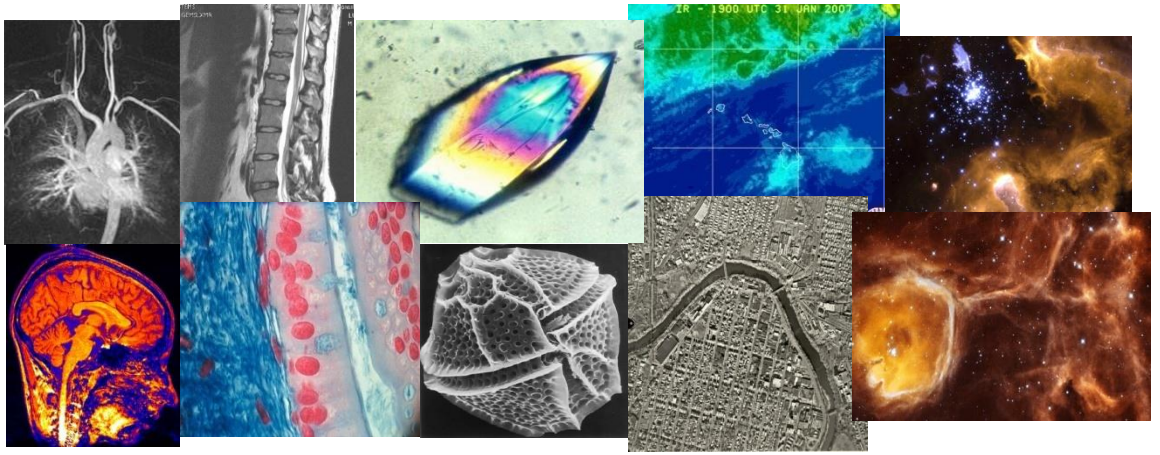
- Images and video are everywhere!

Personal photo albums

Movies, news, sports

Instagram flickr shutterstock YouTube gettyimages

Surveillance and security



Medical and scientific images

Adapted from Lana Lazebnik

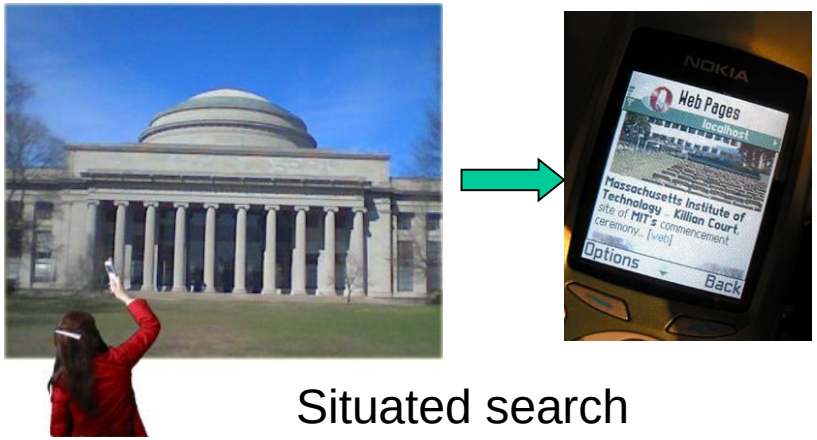
Faces and digital cameras

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

Setting camera focus via face detection

Face recognition

Linking to info with a mobile device



Situated search
Yeh et al., MIT

kooaba

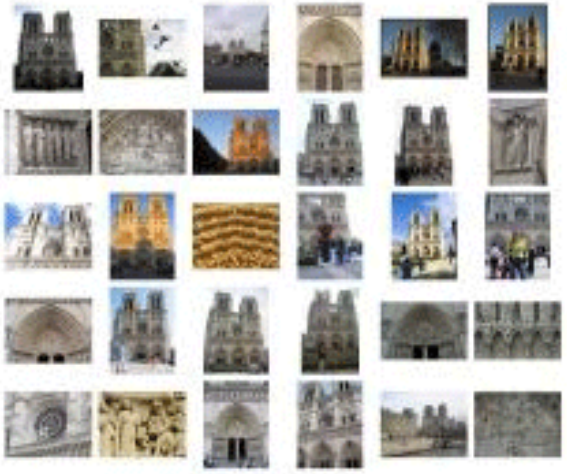
MSR Lincoln

Exploring photo collections

Photo Tourism

Exploring photo collections in 3D

Microsoft



(a)

(b)

(c)

Snavely et al.

Special visual effects

The Matrix

What Dreams May Come

Mocap for *Pirates of the Carribean*, Industrial Light and Magic
Source: S. Seitz

Interactive systems

Shotton et al.

Video-based interfaces

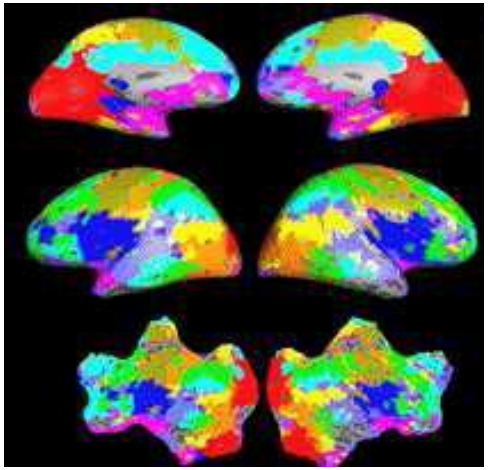
YouTube Link

Human joystick
NewsBreaker Live

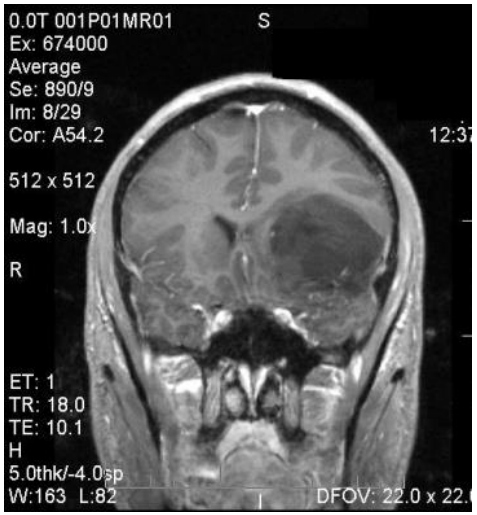
Assistive technology systems
Camera Mouse
Boston College

Vision for medical & neuroimages

Image guided surgery
MIT AI Vision Group

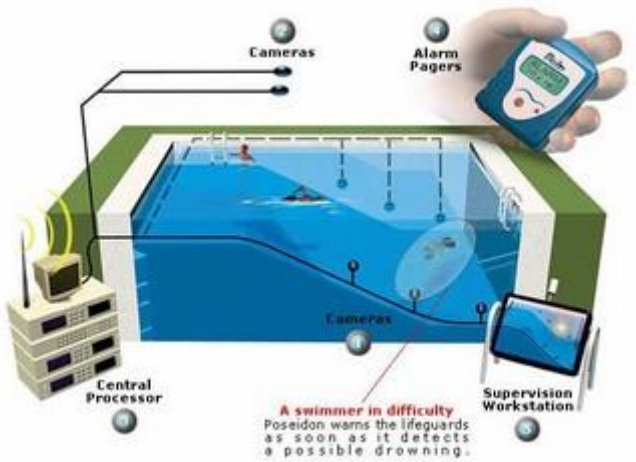


fMRI data
Golland et al.



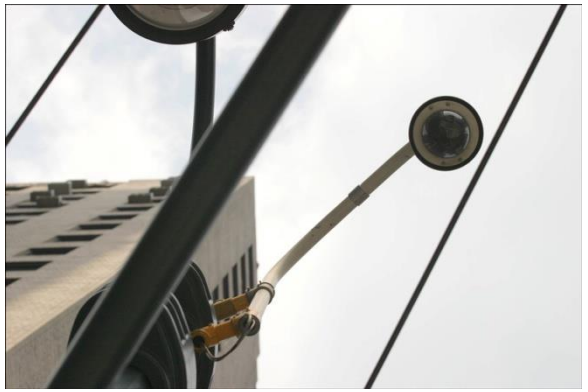
Safety & security

Navigation,
driver safety



Monitoring pool
(Poseidon)

Pedestrian detection
MERL, Viola et al.



Surveillance

Healthy eating

Im2calories by Myers et al., ICCV 2015
figure source

FarmBot.io
YouTube Link

Self-training for sports?

Pirsiavash et al., “Assessing the Quality of Actions”, ECCV 2014

Image generation

Figure 3: Generated bedrooms after five epochs of training. There appears to be evidence of visual under-fitting via repeated noise textures across multiple samples such as the base boards of some of the beds.

Radford et al., ICLR 2016

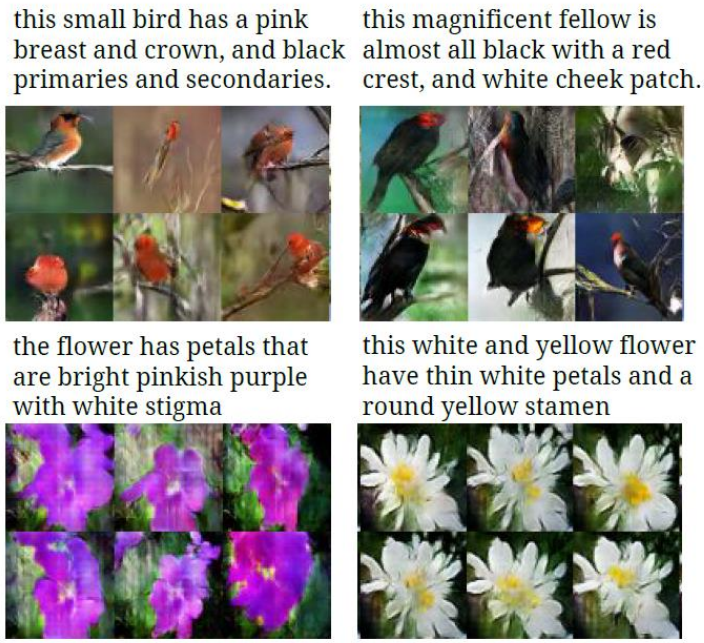


Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories. Right: captions are from training set categories.

Reed et al., ICML 2016

Seeing AI

YouTube link

Obstacles?

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

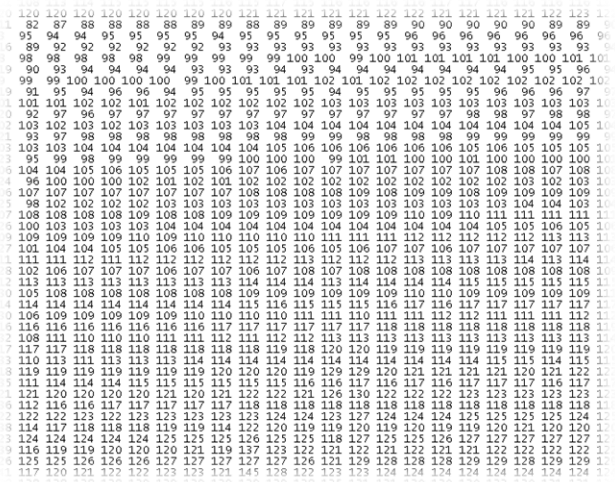
THE SUMMER VISION PROJECT

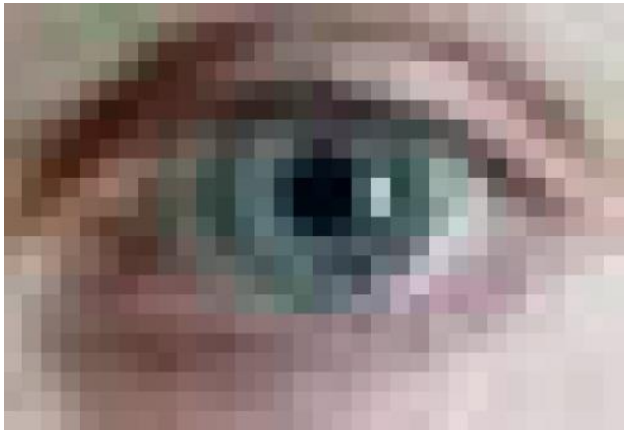
Seymour Papert

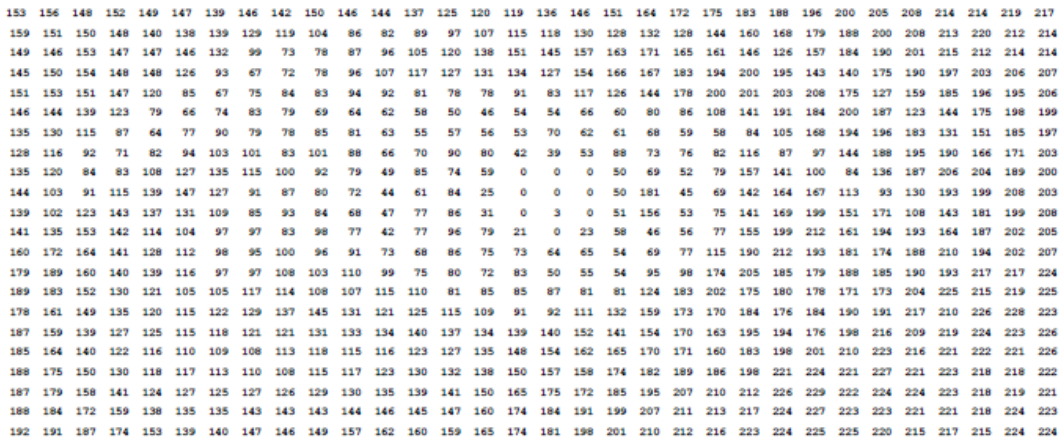
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the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

Read more about the history: Szeliski Sec. 1.2

What the computer gets







Why is this problematic?

Adapted from Kristen Grauman and Lana Lazebnik

Why is vision difficult?

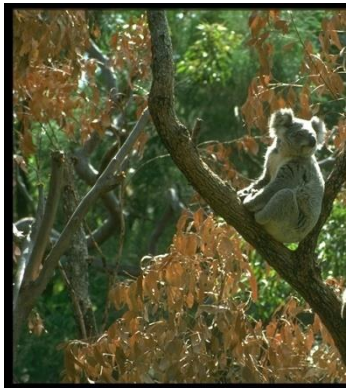
- Ill-posed problem: real world much more complex than what we can measure in images
  - 3D → 2D
- Impossible to literally “invert” image formation process with limited information
- Need information outside of this particular image to generalize what image portrays (e.g. to resolve occlusion)

Adapted from Kristen Grauman

Challenges: many nuisance parameters

Illumination

Object pose



Clutter

Occlusions

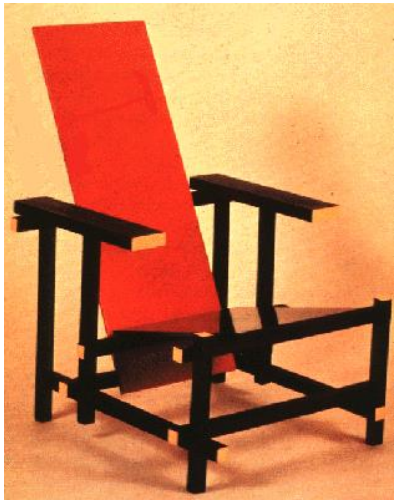
Intra-class appearance

Viewpoint

Think again about the pixels…

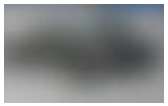
Challenges: intra-class variation

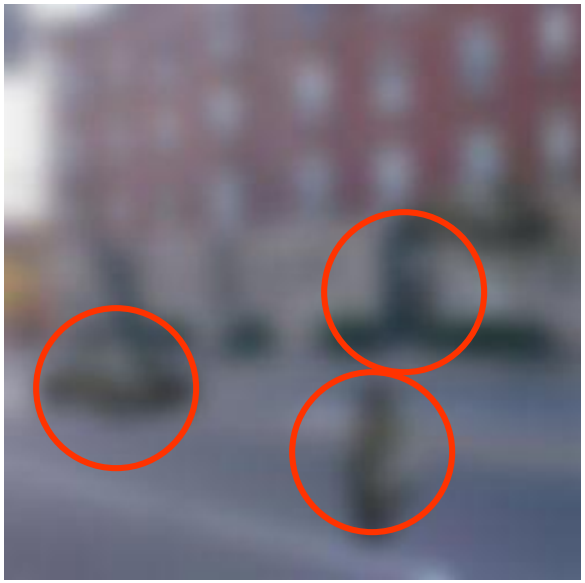
CMOA Pittsburgh



slide credit: Fei-Fei, Fergus & Torralba

Challenges: importance of context





slide credit: Fei-Fei, Fergus & Torralba

Challenges: Complexity

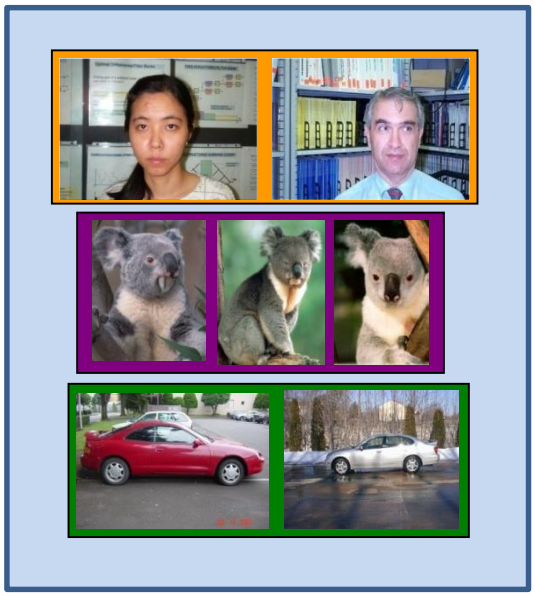
- Thousands to millions of pixels in an image
- 3,000-30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images indexed by Google Image Search
- 18 billion+ prints produced from digital camera images in 2004
- 295.5 million camera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

Challenges: Limited supervision

Less

More

Unlabeled, multiple objects



Classes labeled, some clutter

Cropped to object, parts and classes labeled

Challenges: Vision requires reasoning

What color are her eyes?
What is the mustache made of?

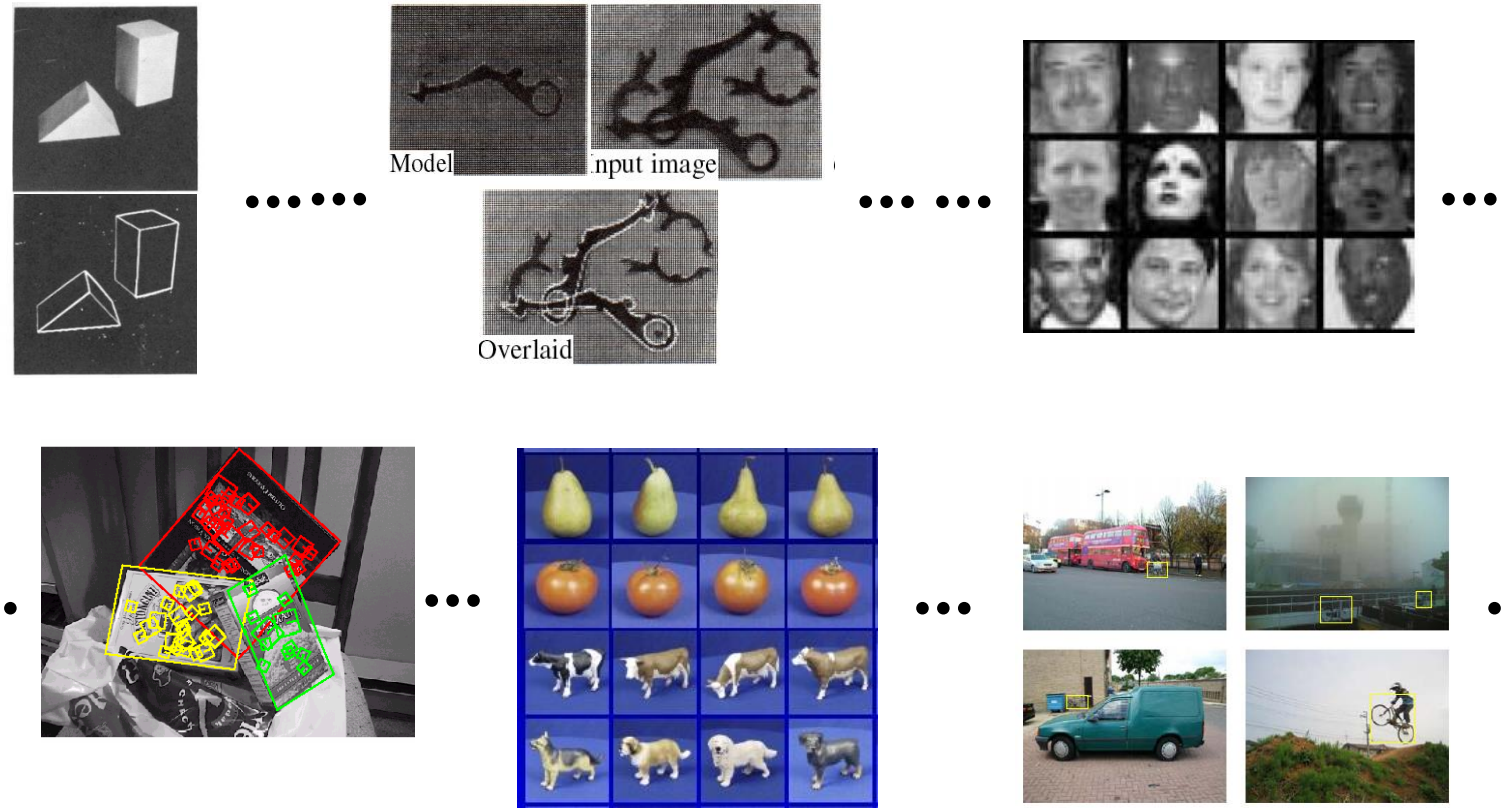
How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Antol et al., “VQA: Visual Question Answering”, ICCV 2015

- Ok, clearly the vision problem is deep and challenging…time to give up?
- Active research area with exciting progress!

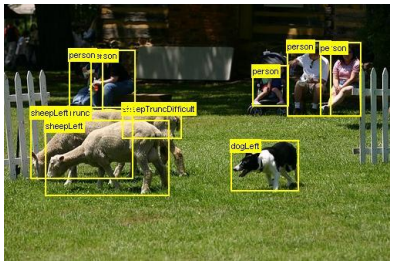


Datasets today



ImageNet:
22k categories, 14mil images


Microsoft COCO:
80 categories, 300k images



PASCAL:
20 categories, 12k images

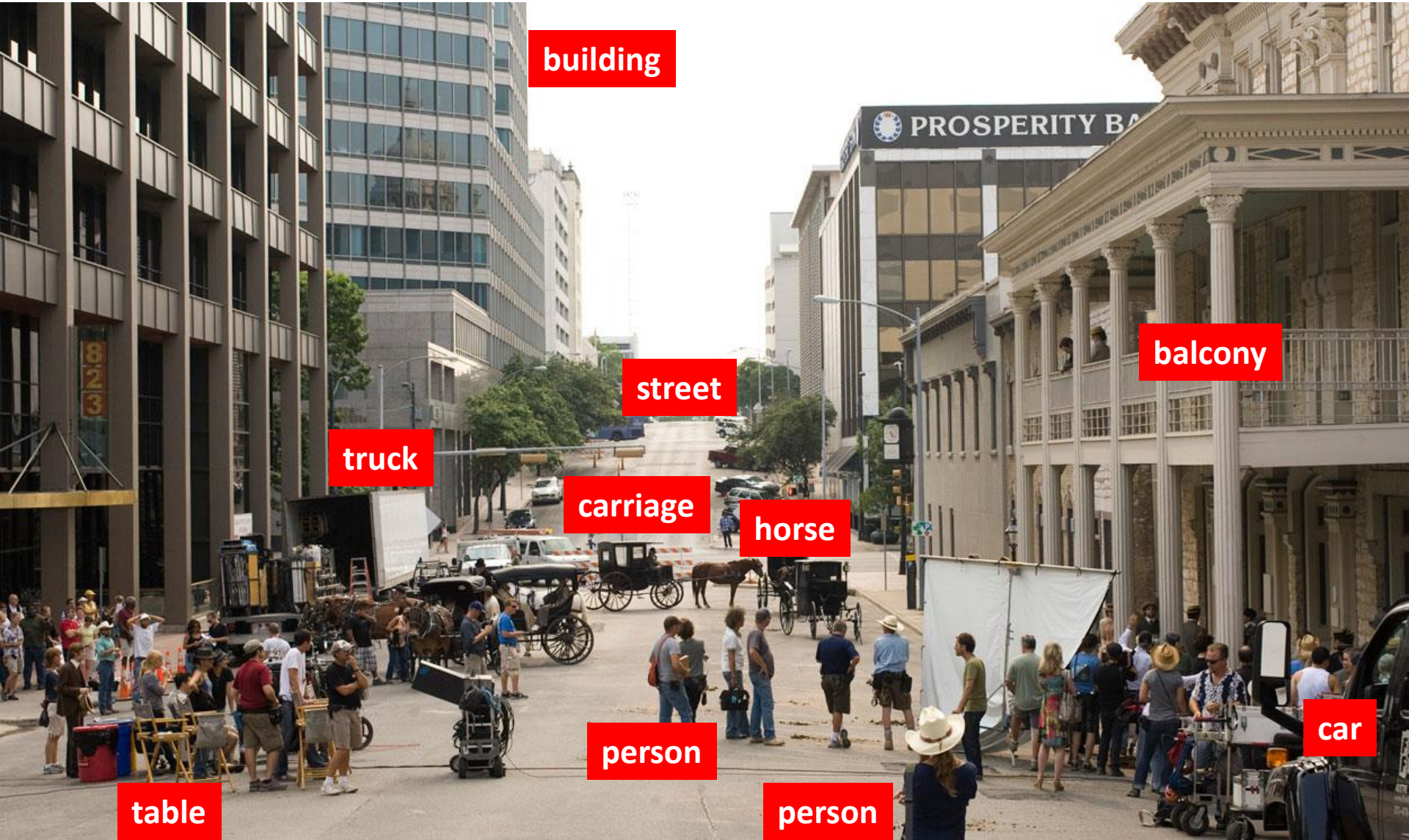


SUN:
5k categories, 130k images

Some Visual Recognition Problems

Recognition: What is this?

Recognition: What objects do you see?



Detection: Where are the cars?

Activity: What is this person doing?

Scene: Is this an indoor scene?

Instance: Which city? Which building?

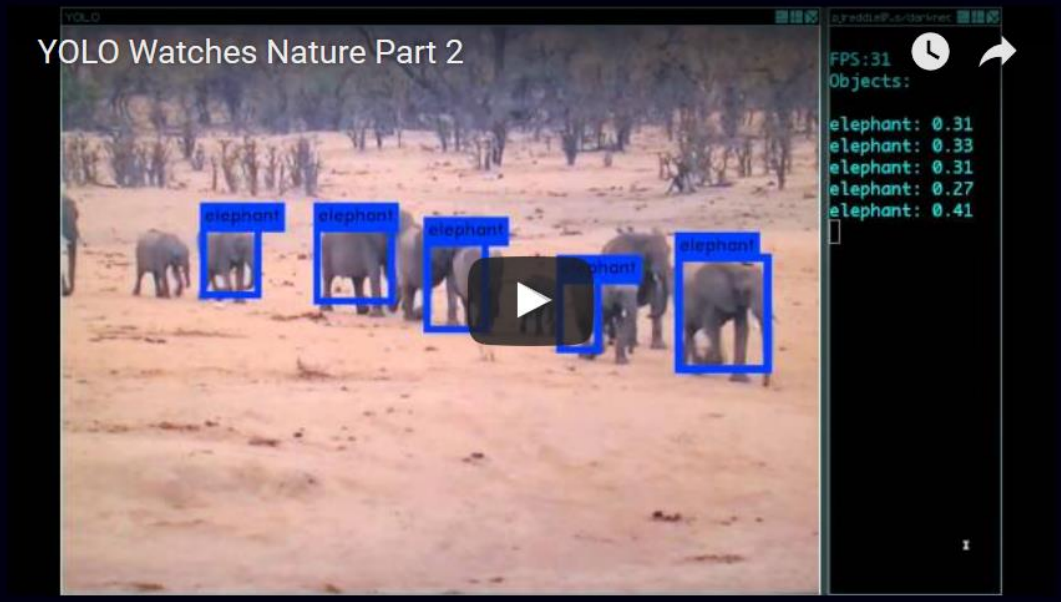
Visual question answering:
What are all these people participating in?

The Latest at CVPR 2016

* CVPR = IEEE Conference on Computer Vision and Pattern Recognition

Our ability to detect objects has gone
from 34 mAP in 2008
to 73 mAP at 7 FPS (frames per second)
or 63 mAP at 45 FPS
in 2016

YOLO: Real-Time Object Detection



You only look once (YOLO) is a system for detecting objects on the Pascal VOC 2012 dataset. It can detect the 20 Pascal object classes:

- person
- bird, cat, cow, dog, horse, sheep
- aeroplane, bicycle, boat, bus, car, motorbike, train
- bottle, chair, dining table, potted plant, sofa, tv/monitor

You Only Look Once: Unified, Real-Time Object Detection

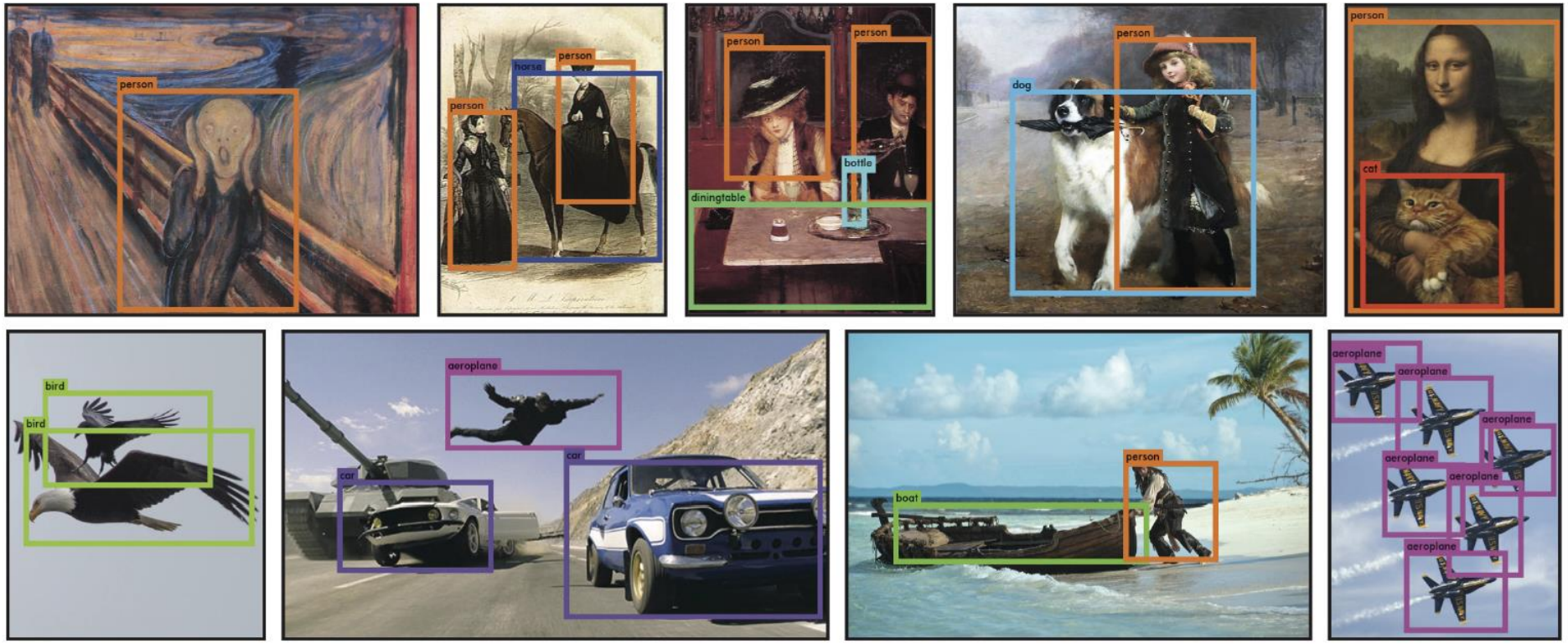
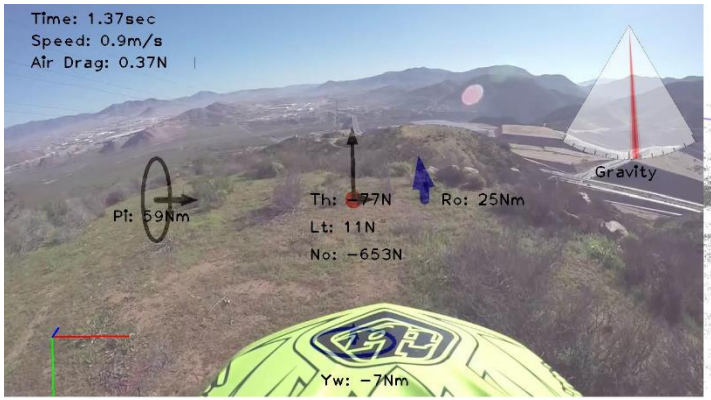
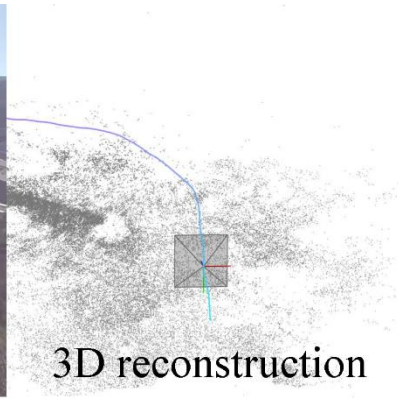


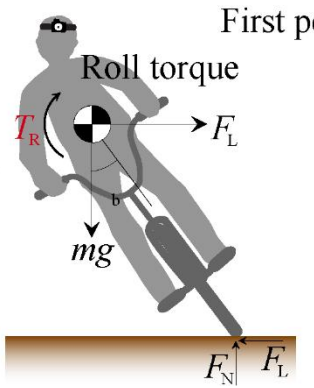
Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

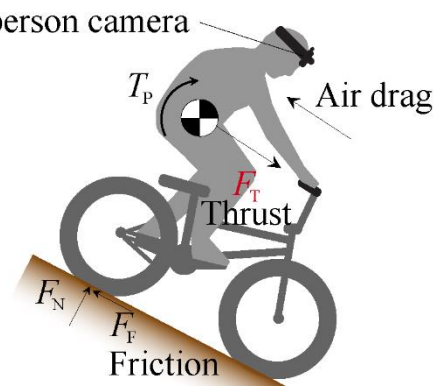
Redmon et al., CVPR 2016

Force from Motion:
Decoding Physical Sensation from a First Person Video



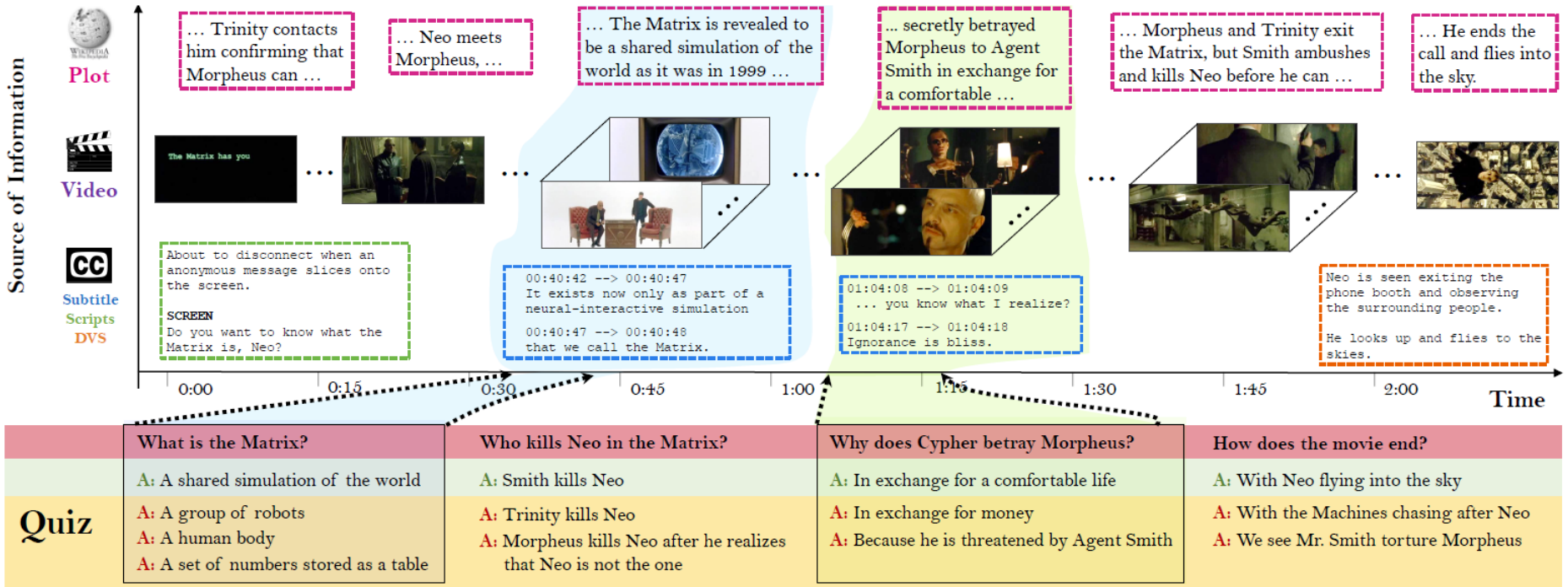






Park et al., CVPR 2016

MovieQA: Understanding Stories in Movies through Question-Answering

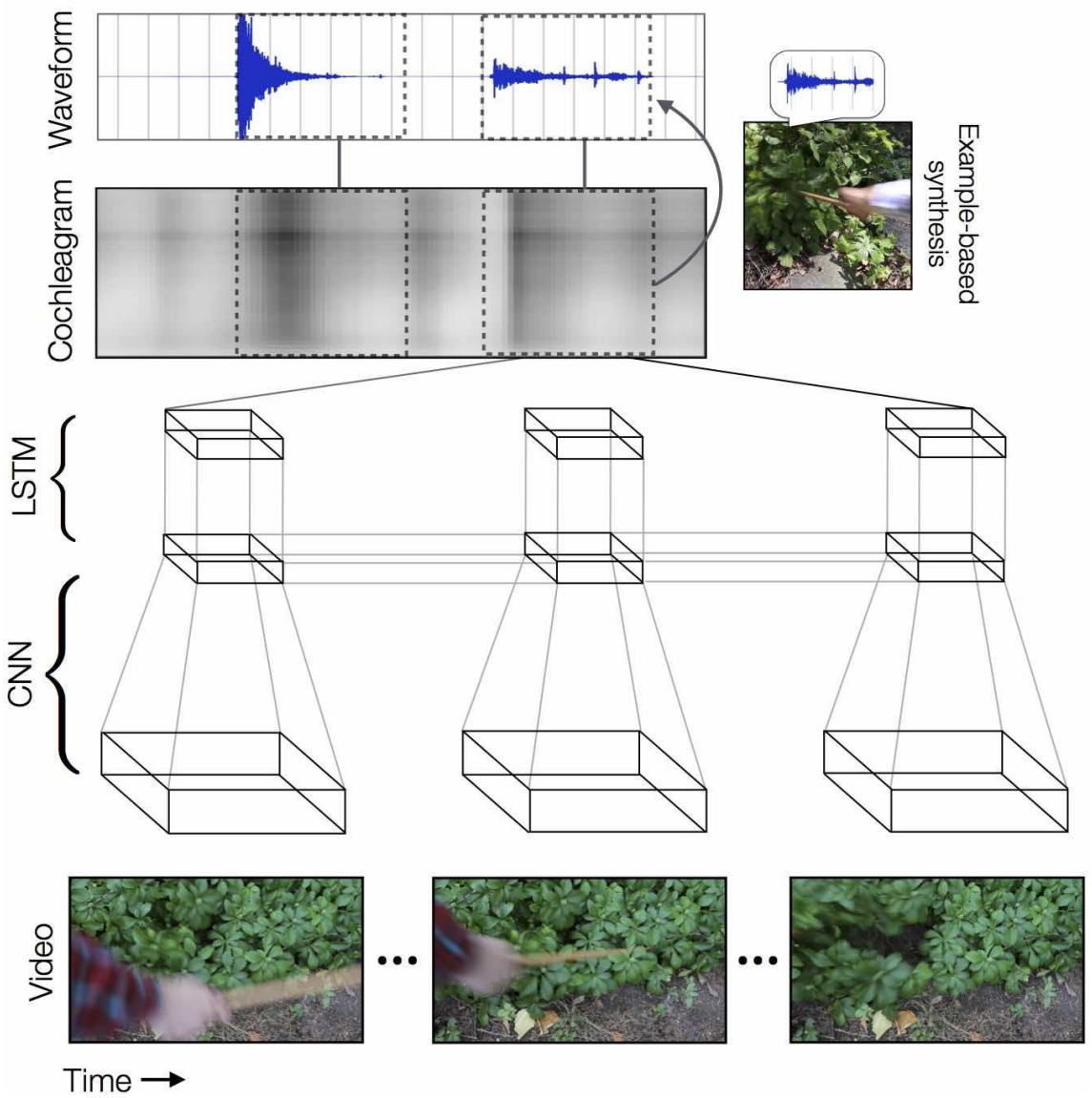


	What is the Matrix?	Who kills Neo in the Matrix?	Why does Cypher betray Morpheus?	How does the movie end?
	A: A shared simulation of the world	A: Smith kills Neo	A: In exchange for a comfortable life	A: With Neo flying into the sky
Quiz	A: A group of robots A: A human body A: A set of numbers stored as a table	A: Trinity kills Neo A: Morpheus kills Neo after he realizes that Neo is not the one	A: In exchange for money A: Because he is threatened by Agent Smith	A: With the Machines chasing after Neo A: We see Mr. Smith torture Morpheus

Figure 1: Our MovieQA dataset contains 14,944 questions about 408 movies. It contains multiple sources of information: plots, subtitles, video clips, scripts, and DVS transcriptions. In this figure we show example QAs from *The Matrix* and localize them in the timeline.

Tapaswi et al., CVPR 2016

Visually Indicated Sounds



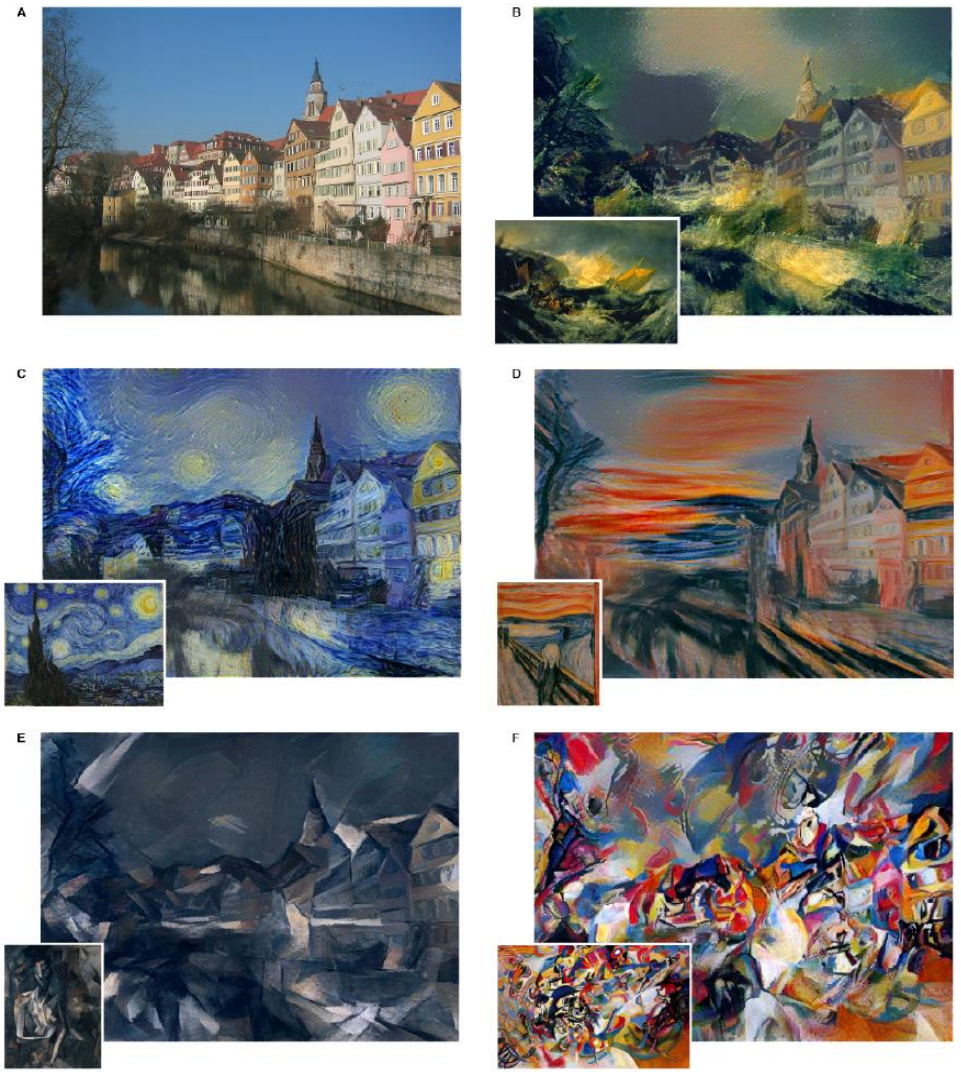
Owens et al., CVPR 2016

Anticipating Visual Representations from Unlabeled Video

Figure 5: **Example Action Forecasts:** We show some examples of our forecasts of actions one second before they begin. The left most column shows the frame before the action begins, and our forecast is below it. The right columns show the ground truth action. Note that our model does not observe the action frames during inference.

Vondrick et al., CVPR 2016

Image Style Transfer Using Convolutional Neural Networks

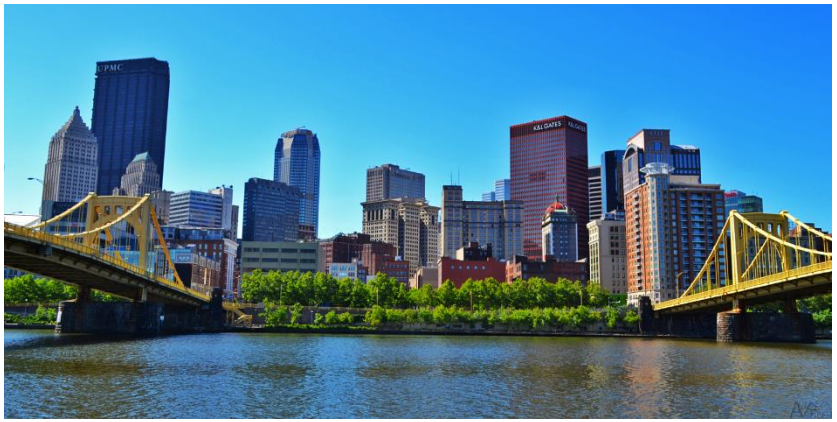


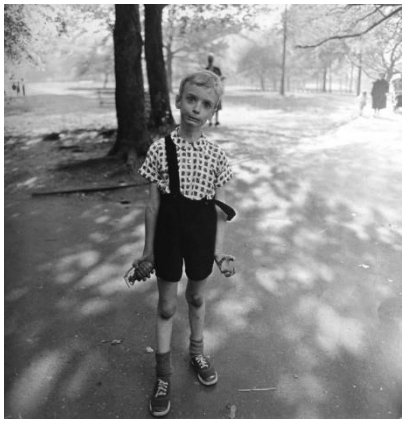
Gatys et al., CVPR 2016

DeepArt.io – try it for yourself!

(Image Style Transfer Using Convolutional Neural Networks)

Images:





Styles:

DeepArt.io – try it for yourself!

(Image Style Transfer Using Convolutional Neural Networks)

Results:

Seeing Behind the Camera: Identifying the Authorship of a Photograph

(a) (b) (c)

Figure 1: Three sample photographs from our dataset taken by Hine, Lange, and Wolcott, respectively. Our top-performing feature is able to correctly determine the author of all three photographs, despite the very similar content and appearance of the photos.

Thomas and Kovashka, CVPR 2016

Is computer vision solved?

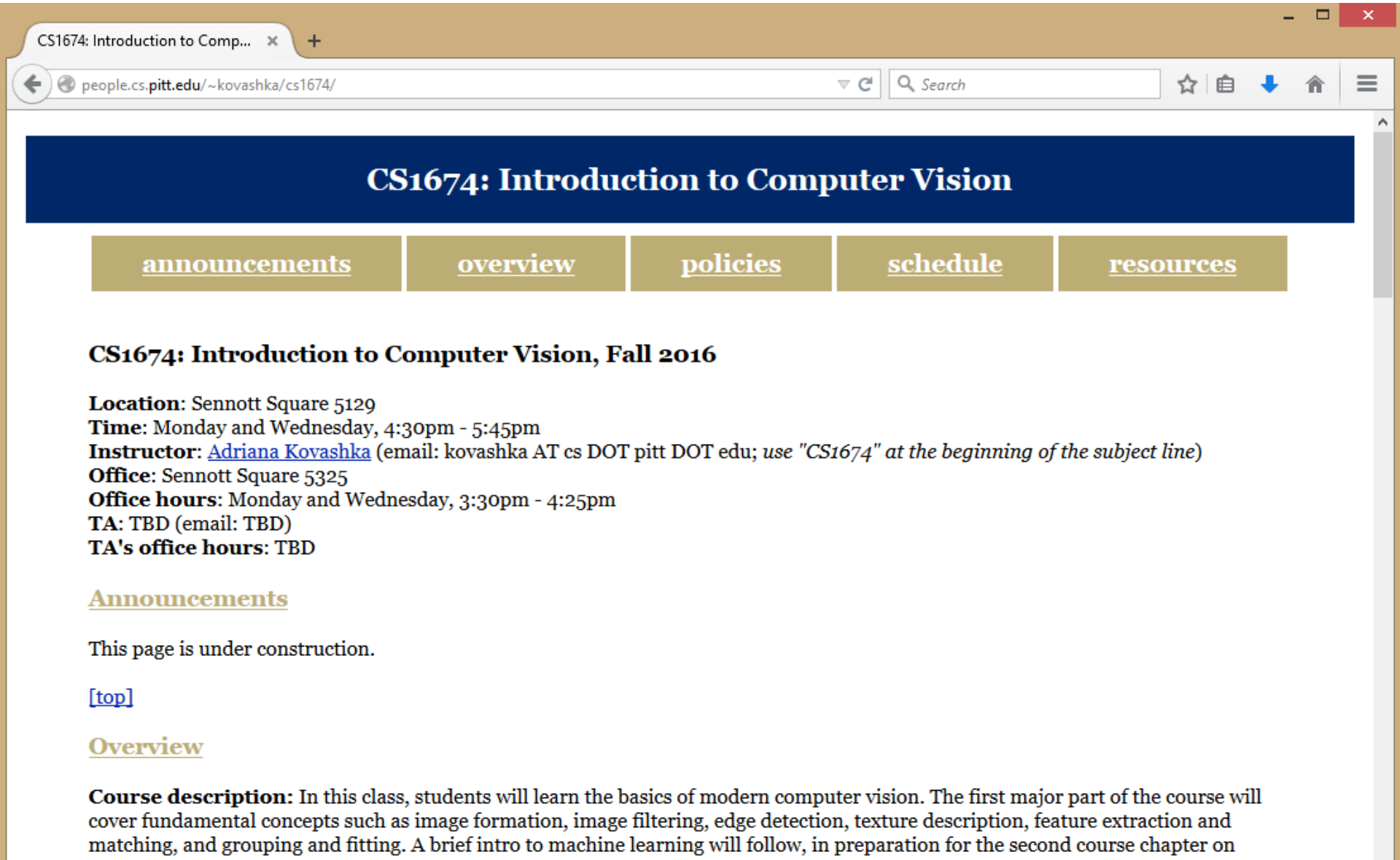
- Given an image, we can guess with 81% accuracy what object categories are shown (ResNet)
- … but we only answer “why” questions about images with 14% accuracy!

Why does it seem that it's solved?

- Deep learning makes excellent use of massive data (labeled for the task of interest?)
  - But it's hard to understand *how* it does so
  - It doesn't work well when massive data is not available and your task is different than tasks for which data is available
- Sometimes the manner in which deep methods work is not intellectually appealing, but our "smarter" / more complex methods perform worse

Course Structure and Policies

Course Website + Schedule



Readings

- Readings for class *n+1* will be posted by 11:59pm on the day of class *n*
- They provide background and more detailed explanations of topics we'll cover
- Generally they provide more info than what you need for the exams
- Sometimes I will post research papers as reading

Course Components

- Written HW (11 assignments x 1% each = 11%)
- Programming HW (11 assignments x 4% each = 44%)
- Midterm exam (15%)
- Final exam (25%)
- Participation (5%)

Written Homework

- Short-answer or multiple-choice answers
- Will help ensure that you understand the most recent topics covered, so you can do the programming homework
- Due two days before the programming assignments
- **Free late days do not apply**
- I grade these assignments

Programming Homework

- Implement a technique or practice concepts we discussed
- One assignment roughly every week
- **Please comment your code!**
- Free late days apply (see next slide)
- TA grades these

Late Policy

- On programming assignments **only**:
- You get 3 "free" late days, i.e., you can submit homework a total of 3 days late.
- For example, you can submit one problem set 12 hours late, and another 60 hours late.
- Once you've used up your free late days, you will incur a penalty of 25% from the total project credit possible for each late day.
- A late day is anything from 1 minute to 24 hours.

Homework Submission

- Navigate to the CourseWeb page for CS1674, click on "Assignments" and the corresponding HW ID
- Your written answers should be a single .pdf/.doc/.docx file
- Your code should be a single zip file with .m files (and images/results if requested)
- Name the file YourFirstName_YourLastName.<extension>
- Homework is due at 11:59pm on the due
- Grades will appear on CourseWeb

Exams

- One mid-term and one final exam (15% / 25%)
- Midterm counts for less because based on only 9/23 “meat” lectures (excluding intro, Matlab, reviews)
- The final exam will focus on the latter half of the course
- Exams will be preceded by review sessions (if our schedule allows it)
- **There will be no make-up exams unless you or a close relative is seriously ill (<u>excludes</u> cold/flu)**

Participation

- 5% of grade will be based on attendance and participation
- Answer questions asked by instructor and others
- Ask meaningful questions
- Ask or answer questions on Piazza
- Bring in relevant articles about recent developments in computer vision

- Feedback is welcome!

Collaboration Policy

- You will work individually. The work you turn in must be your own work.
- You can discuss the assignments with your classmates, but do not look at their code or answers.
- You **cannot** use posted solutions, search for code on the internet or use Matlab's implementations of something you are asked to write.
- When in doubt, ask the instructor or TA!
- **Plagiarism will cause you to fail the class and receive disciplinary penalty.**

Disabilities

- If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will verify your disability and determine reasonable accommodations for this course.

Medical Conditions

- If you have a medical condition which will prevent you from doing a certain assignment, you must inform the instructor of this **before** the deadline.
- You must then submit documentation of your condition within a week of the assignment deadline.
- **There will be no make-up exams.**

No Classroom Recording

- To ensure the free and open discussion of ideas, students may not record classroom lectures, discussion and/or activities without the advance written permission of the instructor, and any such recording properly approved in advance can be used solely for the student's own private use.

Warnings

Warning #1

- This class is **a lot of work**
- This time I've opted for shorter, more manageable HW assignments, but there is more of them
- I expect you'd be spending **6-8 hours** on homework each week

- … But you get to understand algorithms and concepts in detail!

Warning #2

- Some parts will be **hard** and require that you pay close attention!
- … I will use the written HW to gauge how you're doing
- … I will also pick on students randomly to answer questions
- **Use instructor's and TA's office hours!!!**

- … You will learn a lot!

Warning #3

- Programming assignments will be in Matlab since that's very common in computer vision, and is optimized for work with matrices
- Matlab also has great documentation
- HW1 is just Matlab practice
- Some people **won't like Matlab** (I like it!)

- … You will learn a new programming language!

Evidence that you should take my warnings seriously

- HW1W is due on Monday (Labor Day)
- HW1P is due on Wednesday (Sept. 7)

If this doesn’t sound like your cup of coffee…

- … drop deadline is September 9 (next Friday).

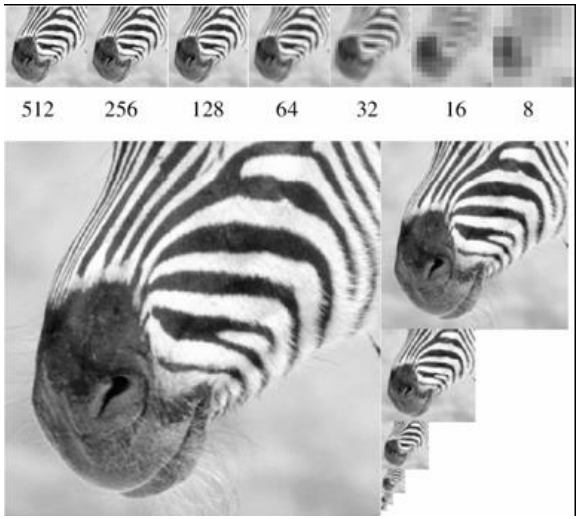
Note to Waitlisted Students

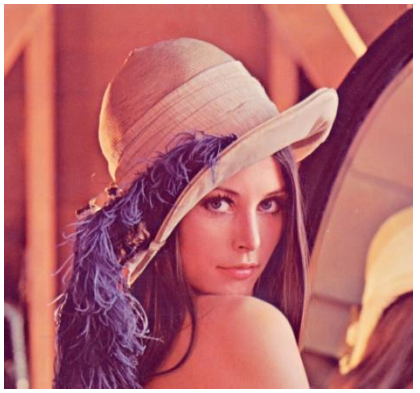
- Keep coming to class if it sounds interesting!

Questions?

Overview of Topics

Features and filters

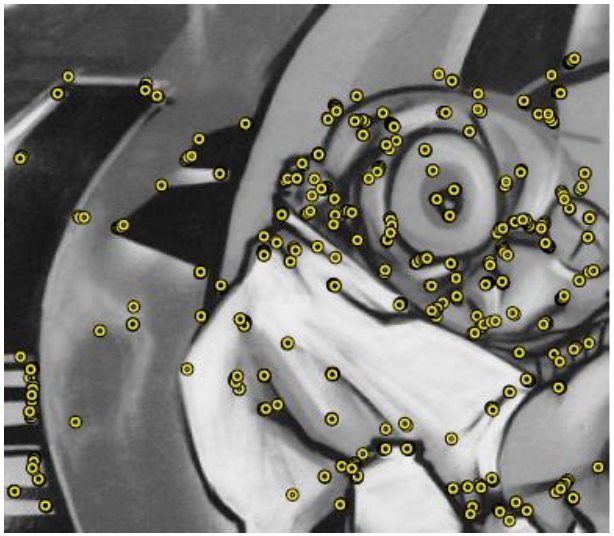


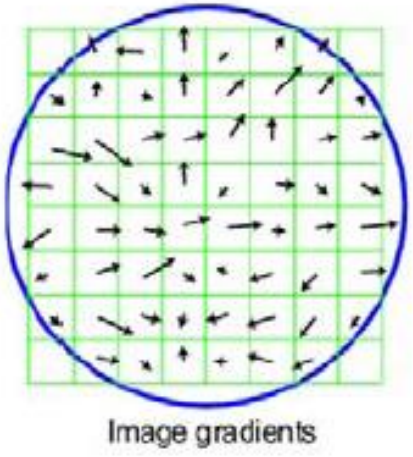


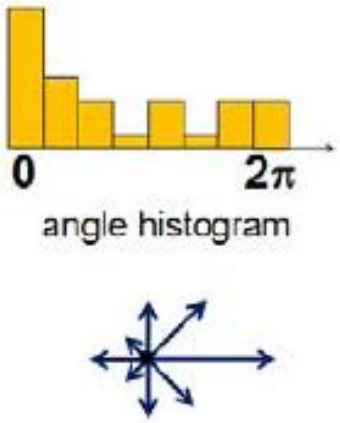
- Transforming and describing images; textures, colors, edges

Features and filters

- Detecting distinctive + repeatable features
- Describing images with local statistics





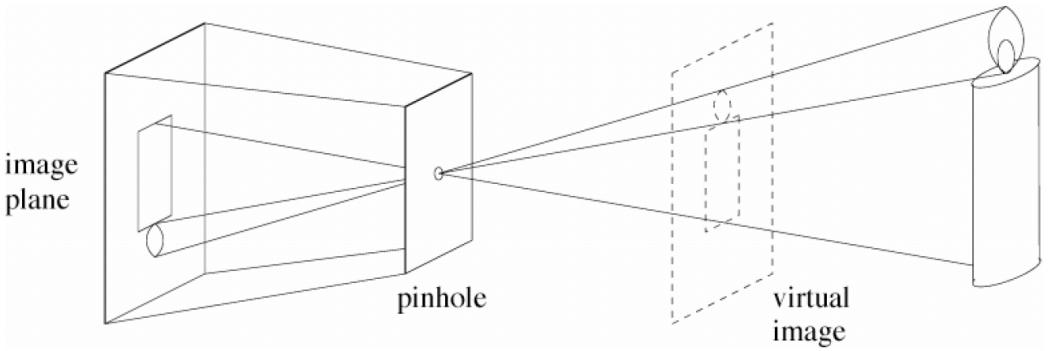


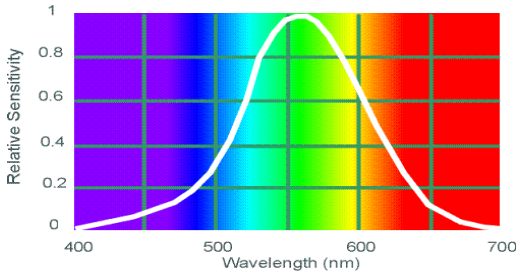
Indexing and search

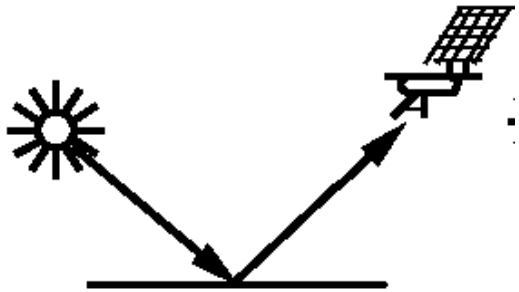
- Matching features and regions across images

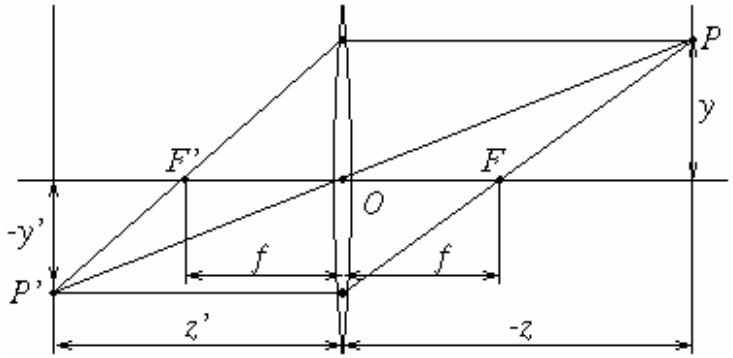
Image formation

- How does light in 3d world project to form 2d images?

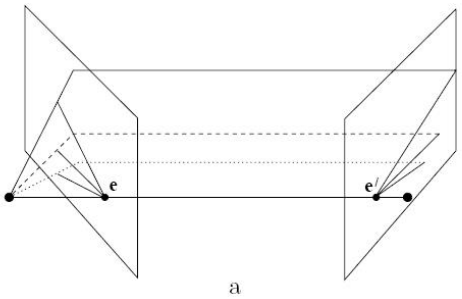




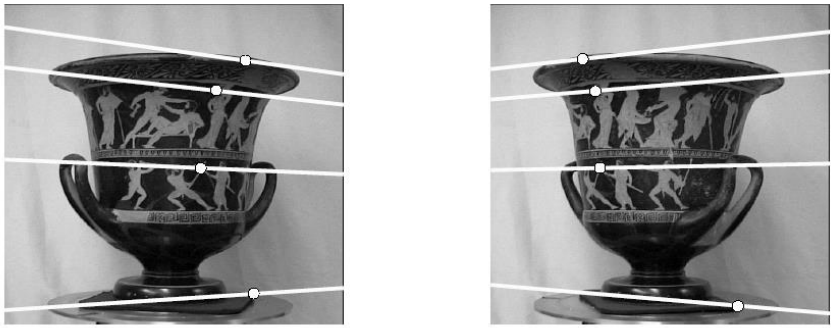




Multiple views

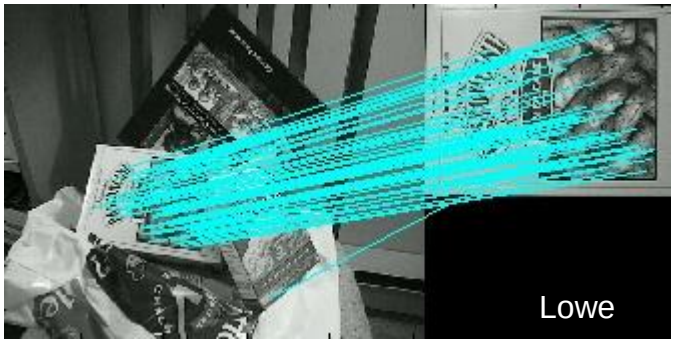


a

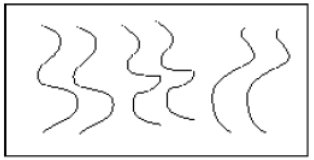


Hartley and Zisserman

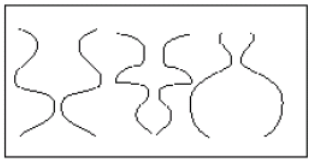
- Multi-view geometry, matching, invariant features, stereo vision



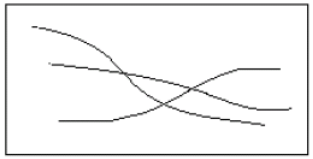
Grouping and fitting



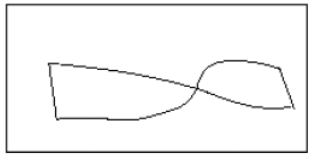
Parallelism



Symmetry



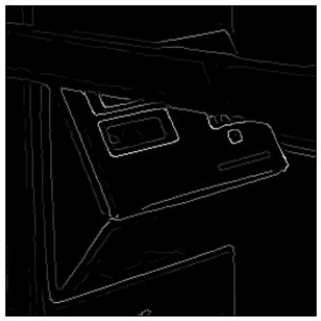
Continuity

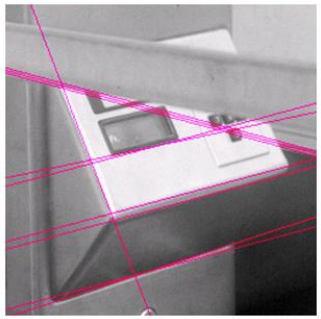


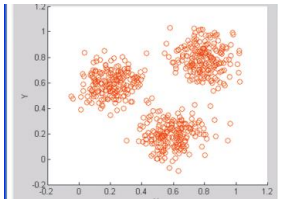
Closure

[fig from Shi et al]

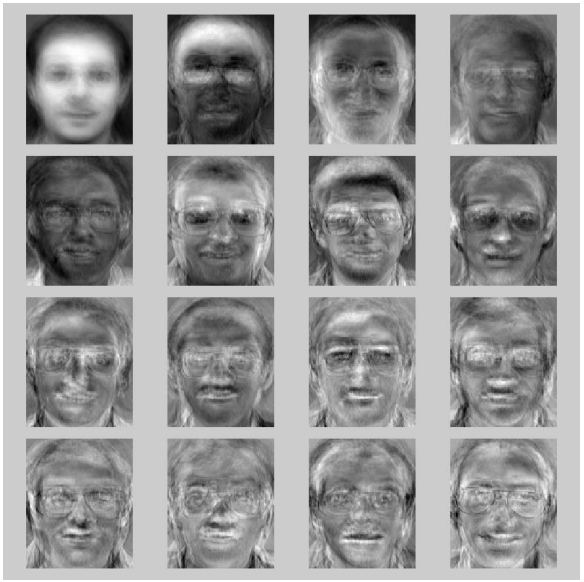
- Clustering, segmentation, fitting; what parts belong together?

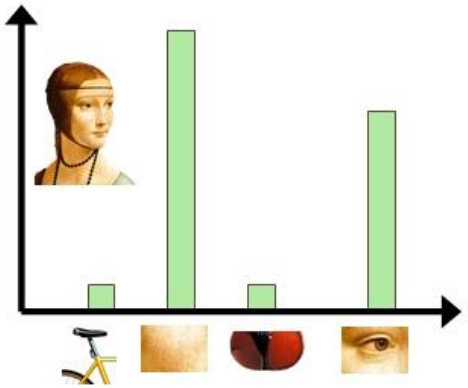


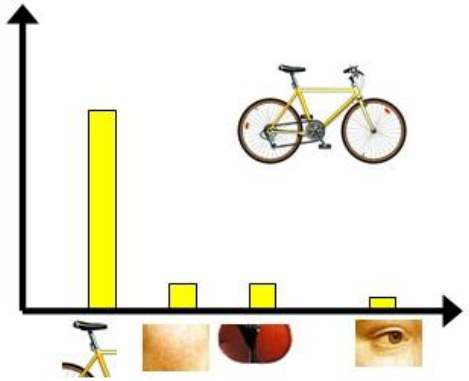


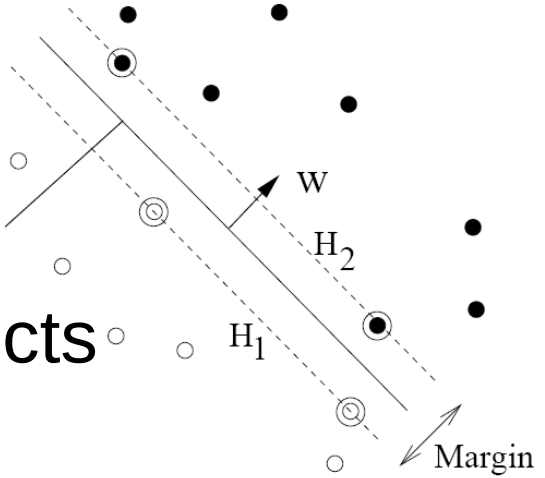


Visual recognition





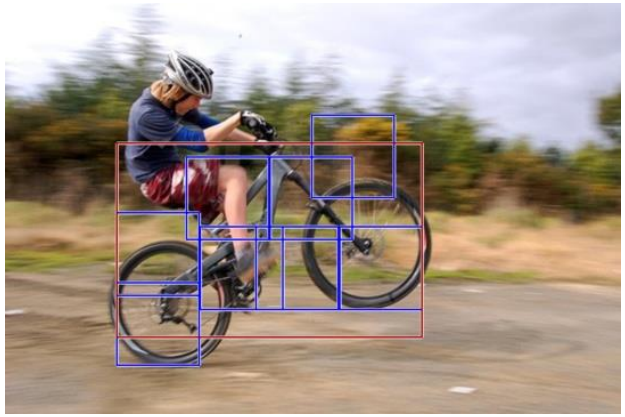


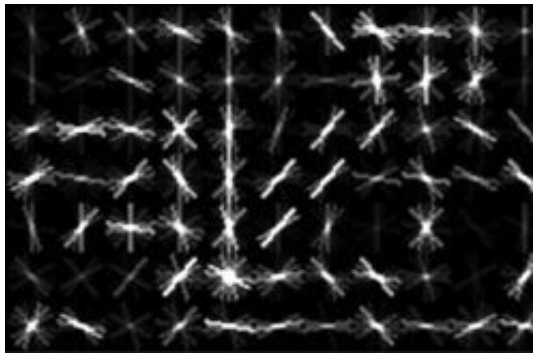


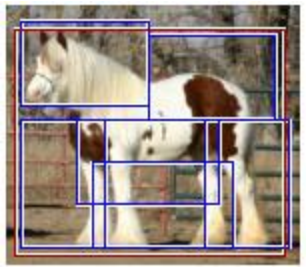
- Recognizing objects and categories, learning techniques

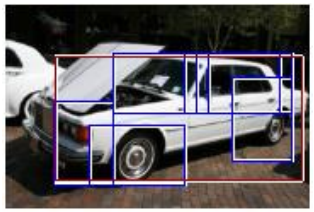
Object detection

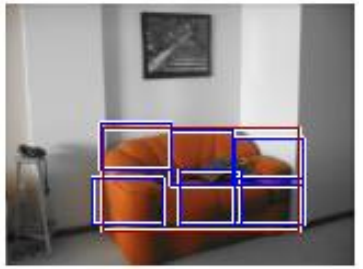
- Detecting novel instances of objects
- Classifying regions as one of several categories











Attribute-based description

- Describing the high-level properties of objects
- Allows recognition of unseen objects

Naming → Aeroplane

Description → Unknown, Has Wheel, Has Wood

Unusual attributes → Bird, No Head, No Beak

Unexpected attributes → Motorbike, Has Cloth

otter	
black:	yes
white:	no
brown:	yes
stripes:	no
water:	yes
eats fish:	yes

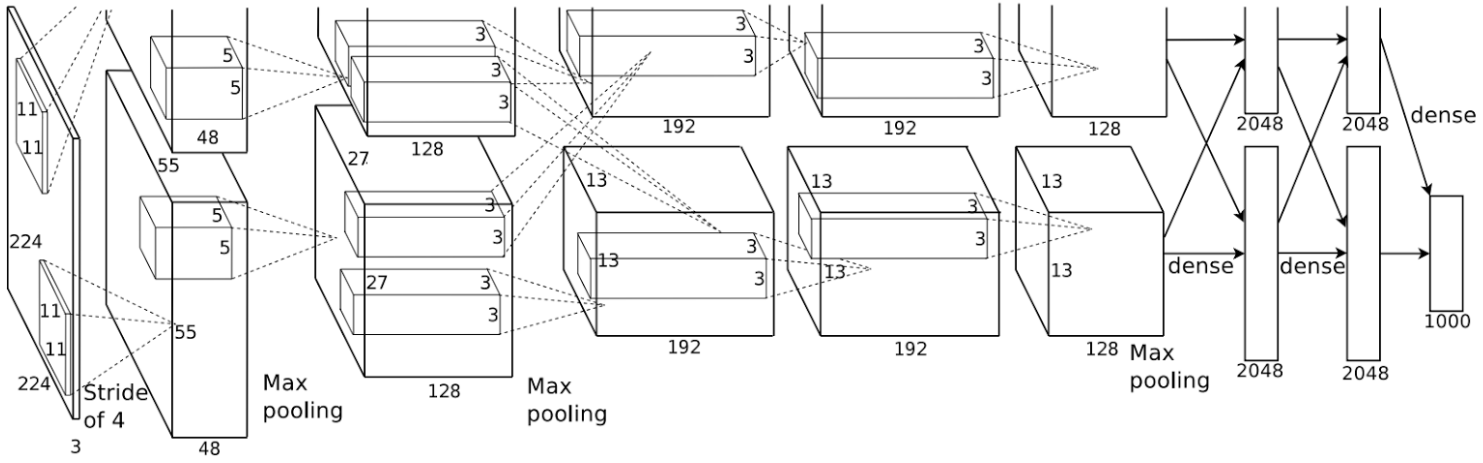
polar bear	
black:	no
white:	yes
brown:	no
stripes:	no
water:	yes
eats fish:	yes

zebra	
black:	yes
white:	yes
brown:	no
stripes:	yes
water:	no
eats fish:	no

Convolutional neural networks

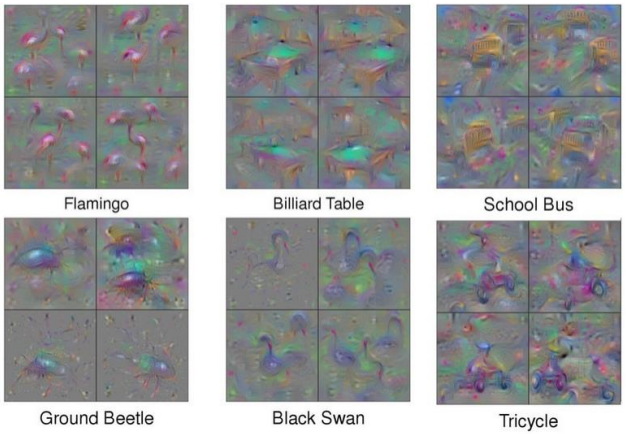
- State-of-the-art on many recognition tasks

Image



Prediction

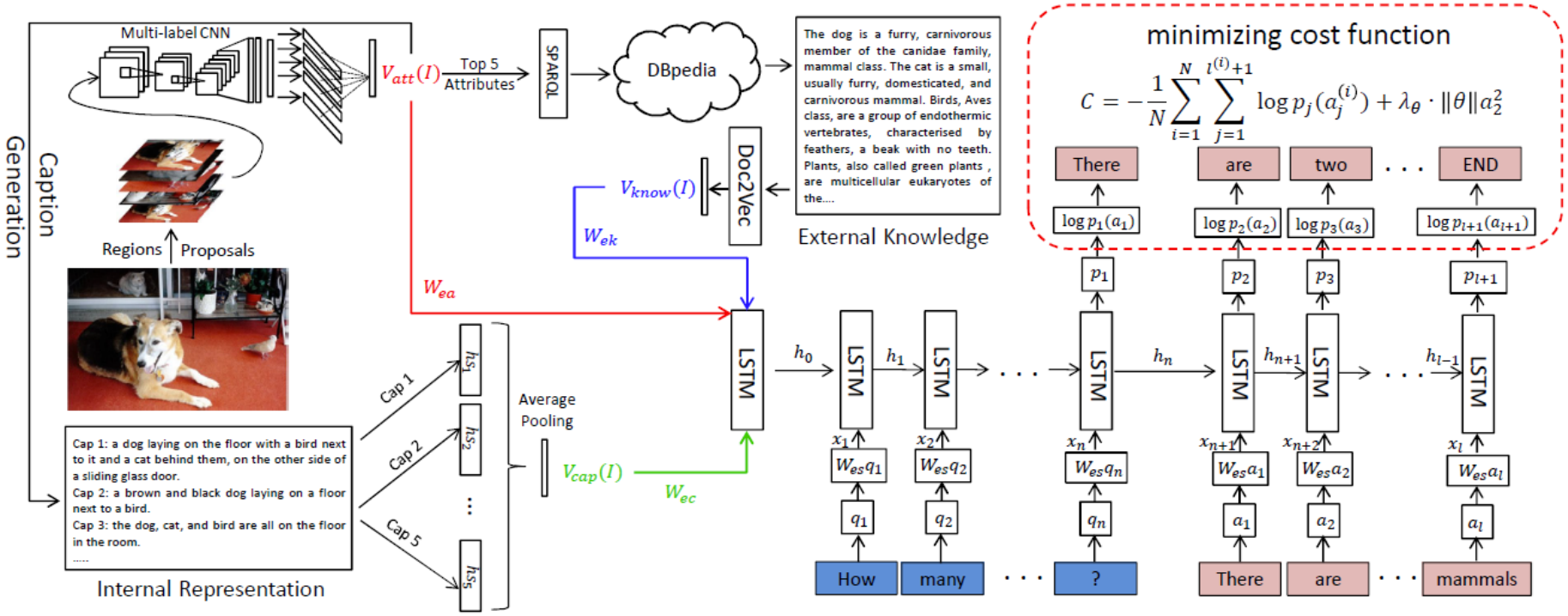
Krizhevsky et al.



Yosinski et al., ICML DL workshop 2015

Recurrent neural networks

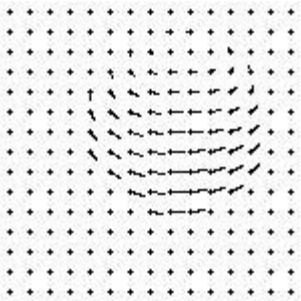
- Sequence processing, e.g. question answering

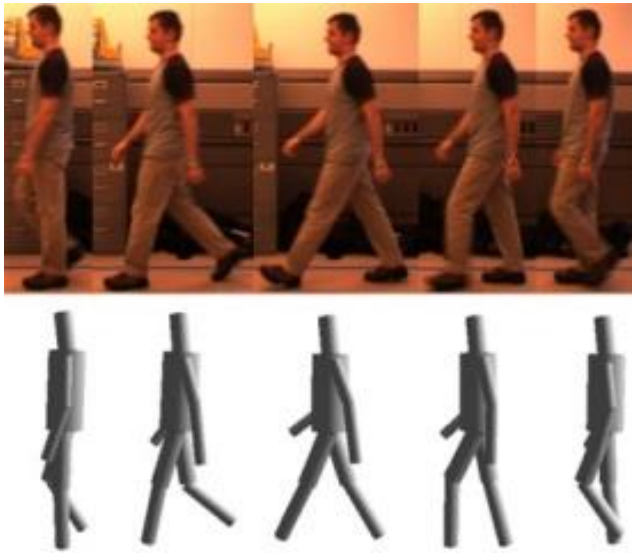


Wu et al., CVPR 2016

Motion and tracking

- Tracking objects, video analysis

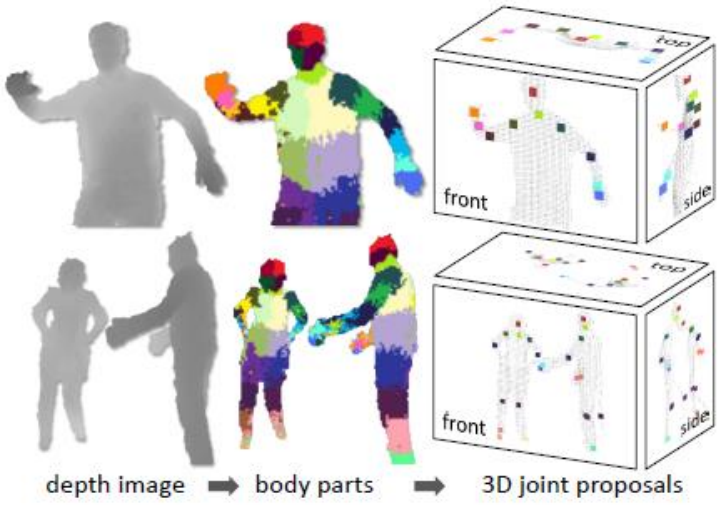


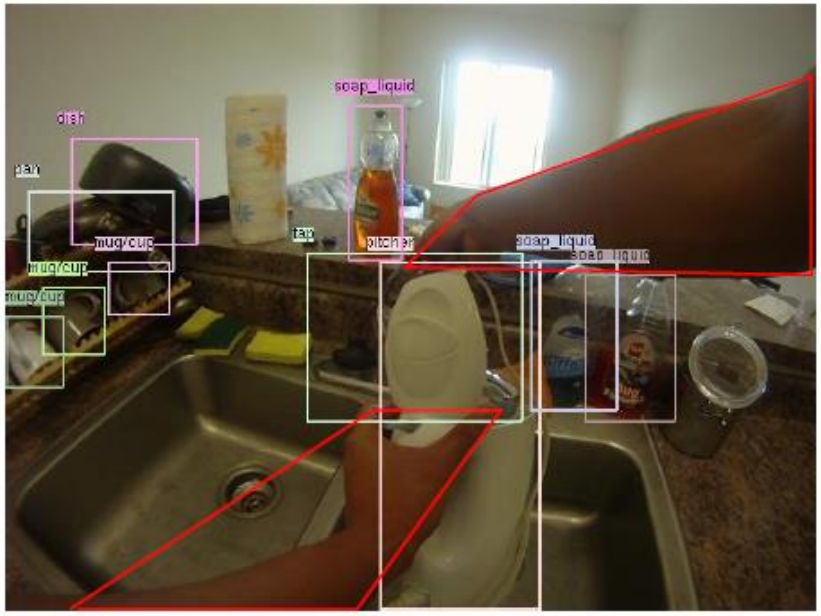


Tomas Izo

Pose and actions

- Automatically annotating human pose (joints)
- Recognizing actions in first-person video





Next Time

- Matlab tutorial
- HW1P out – will read in class

Homework

- HW1W due on Monday
- Read Szeliski Sec. 1.1-1.2
- Enroll for Piazza