CS 1674: Intro to Computer Vision

Introduction

Prof. Adriana Kovashka
University of Pittsburgh
January 9, 2018

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_sp18
- **Instructor**: Adriana Kovashka (kovashka@cs.pitt.edu)
- **Office:** Sennott Square 5325
- **Class:** Tue/Thu, 2:30pm-3:45pm
- **Office hours**: Tue/Thu, 9:30am-11am, 1pm-2pm

TA

- Nils Murrugarra-Llerena (nineil@cs.pitt.edu)
- **Office:** Sennott Square 5404
- **Office hours:** TBD
  - **Do this Doodle by the end of Friday:**

    https://doodle.com/poll/htaw6kudn5paspmc

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

Matlab Tutorials and Exercises

http://www.cs.pitt.edu/~kovashka/cs2770_sp18/tutorial.m
http://www.cs.pitt.edu/~kovashka/cs2770_sp18/myfunction.m
http://www.cs.pitt.edu/~kovashka/cs2770_sp18/myotherfunction.m

https://people.cs.pitt.edu/~milos/courses/cs2750/Tutorial/
http://www.math.udel.edu/~braun/M349/Matlab_probs2.pdf
http://www.facstaff.bucknell.edu/maneval/help211/basicexercises.html

Ask the TA or instructor if you have any problems.

Types of computer vision

- Lower-level vision
  - Analyzing textures, edges and gradients in images, without concern for the semantics (e.g. objects) of the image
- Higher-level vision
  - Making predictions about the semantics or higher-level functions of content in images (e.g. objects, attributes, styles, motion, etc.)
  - Involves machine learning

Course Goals

- To learn the basics of low-level image analysis
- To learn about some classic and modern approaches to high-level computer vision tasks
- To get experience with some computer vision techniques
- To learn/apply basic machine learning (a key component of modern computer vision)
- To think critically about vision approaches, and to see connections between works

Policies and Schedule

http://people.cs.pitt.edu/~kovashka/cs1674_sp18

- Grading and course components
- Homework assignments
- Schedule

Warnings

Warning #1

- This class is **a lot of work**
- 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!

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

- … please drop the class!
- Drop deadline is January 19

Note to Waitlisted Students

- Keep coming to class if it sounds interesting!

Questions?

Plan for Today

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

Blitz Introductions (5-10 sec)

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

(I'll ask you more questions in HW1W.)

Computer Vision

What is computer vision?

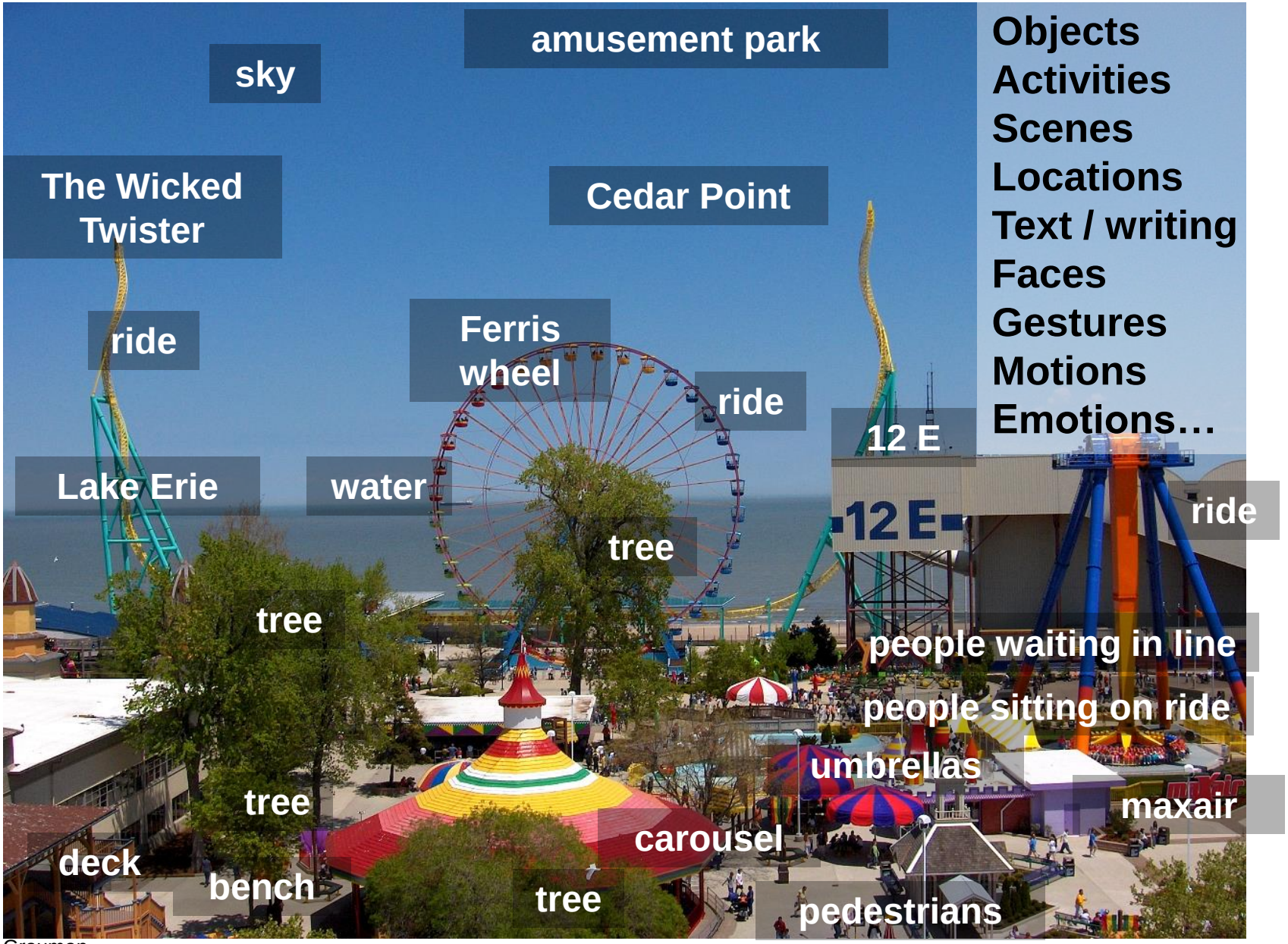
Done?

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

What is computer vision?

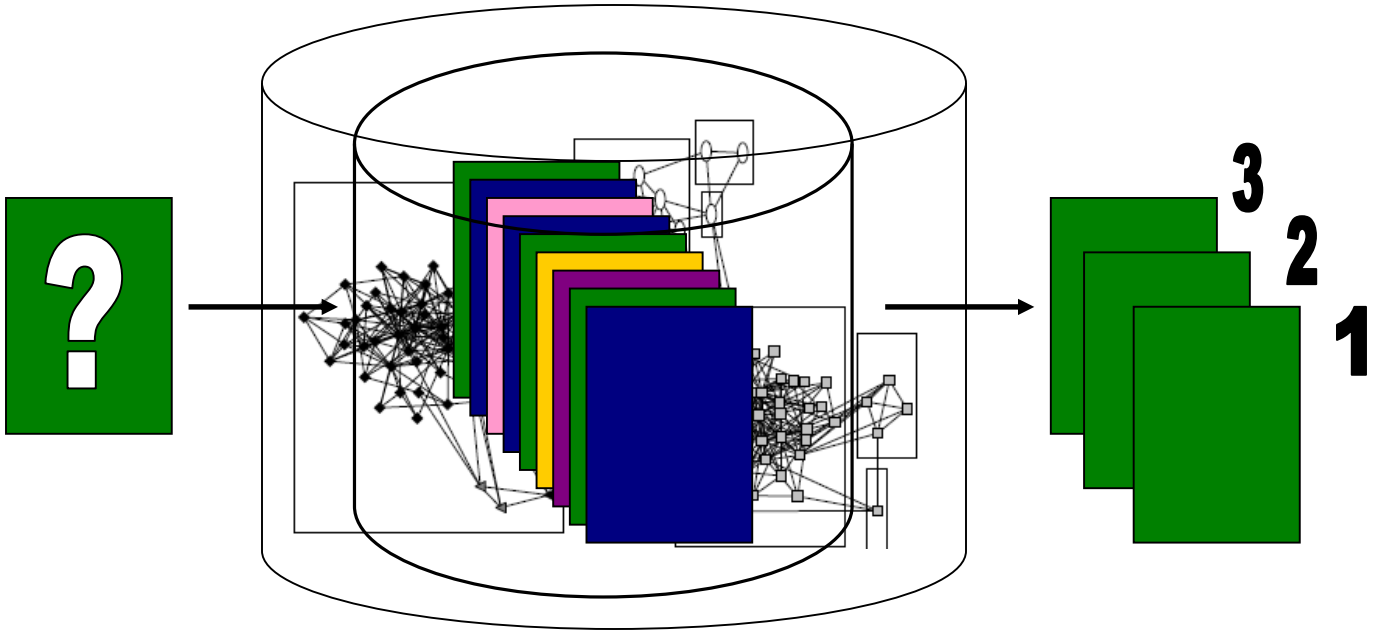
- Automatic understanding of images and video
  - Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities *(perception and interpretation)*
  - Algorithms to mine, search, and interact with visual data (*search and organization*)
  - Computing properties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

Vision for perception, interpretation



Kristen Grauman

Visual search, organization



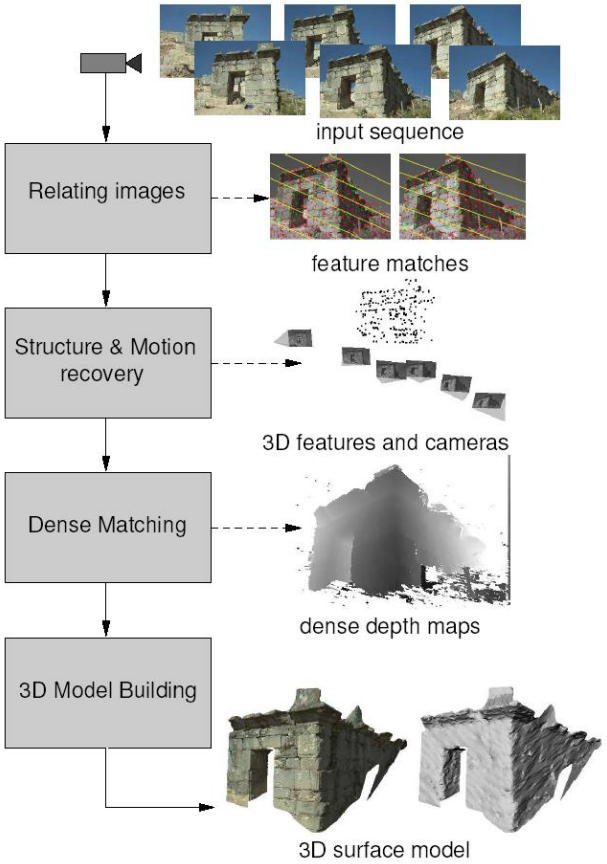
Query **Image or video archives** **Relevant content**

Vision for measurement

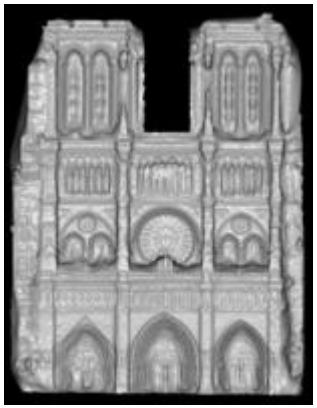
Real-time stereo

Pollefeys et al.

Structure from motion



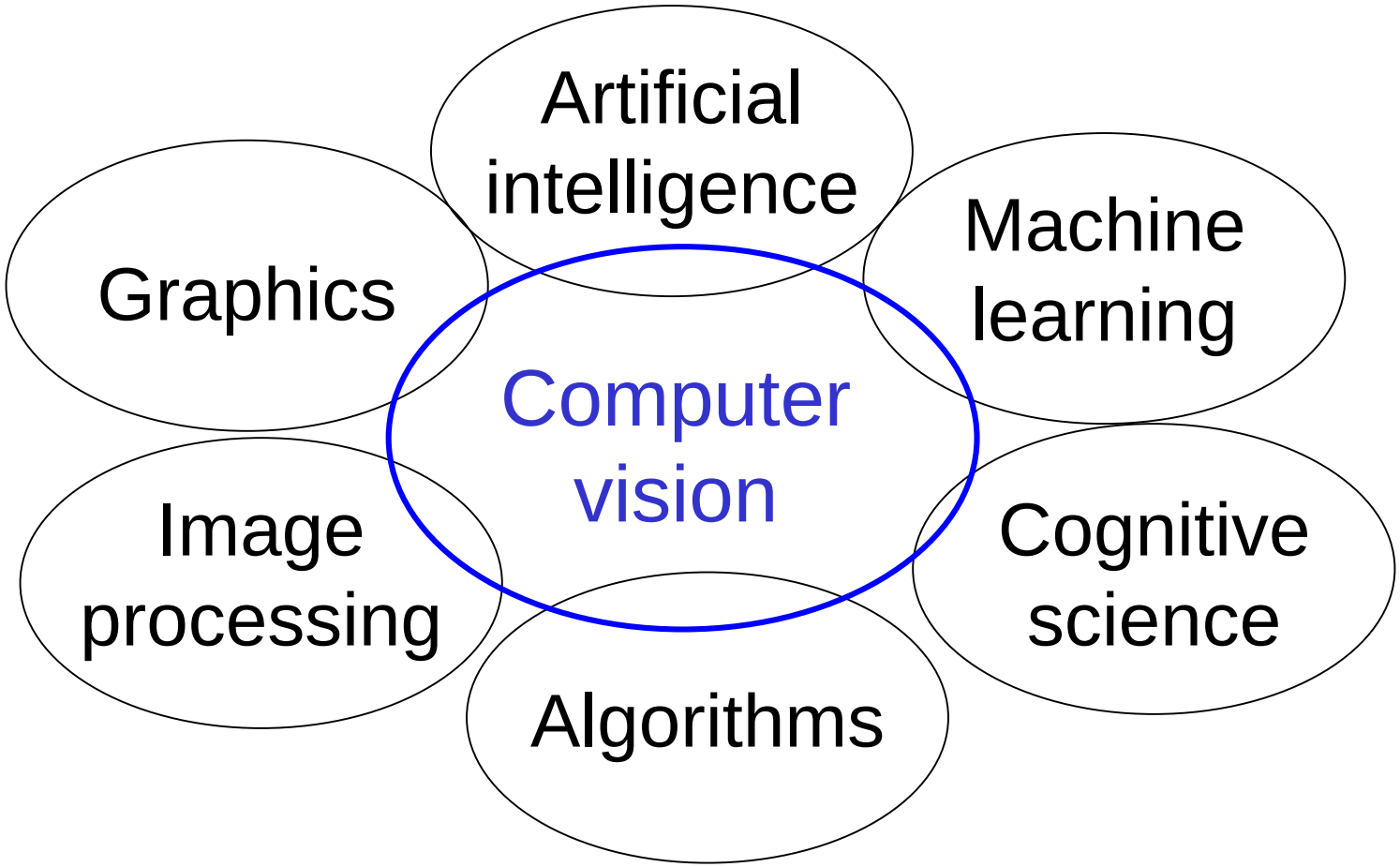
Multi-view stereo for community photo collections



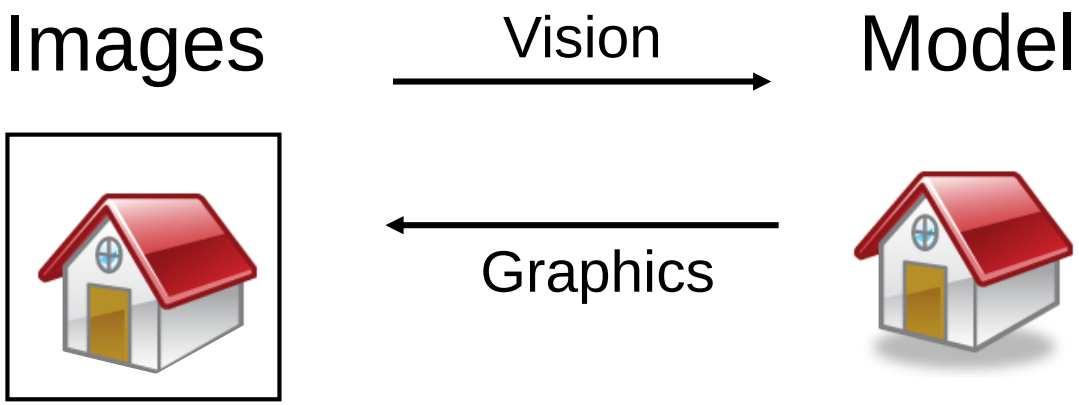
Goesele et al.

Slide credit: L. Lazebnik

Related disciplines



Vision and graphics



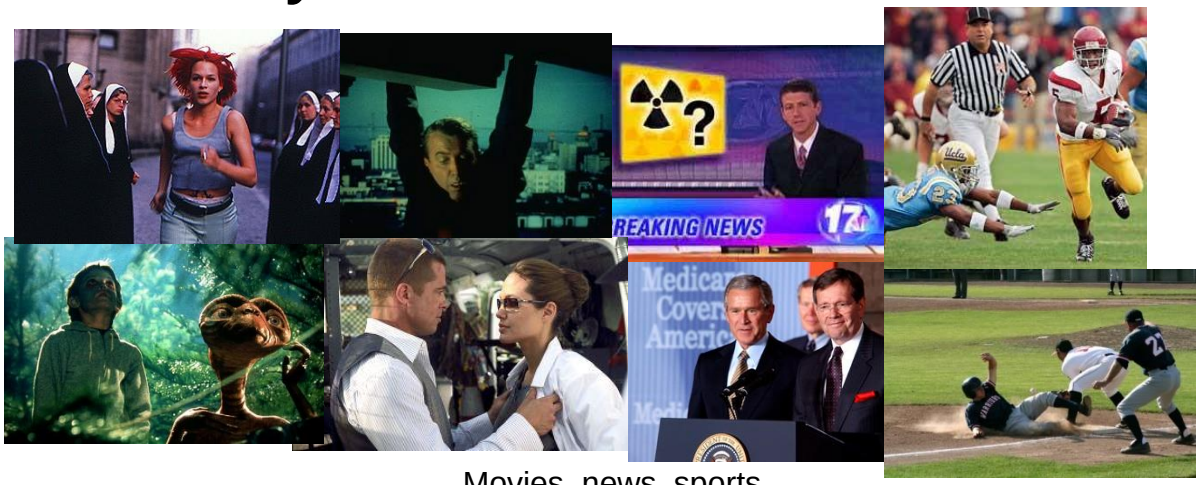
Inverse problems: analysis and synthesis.

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!

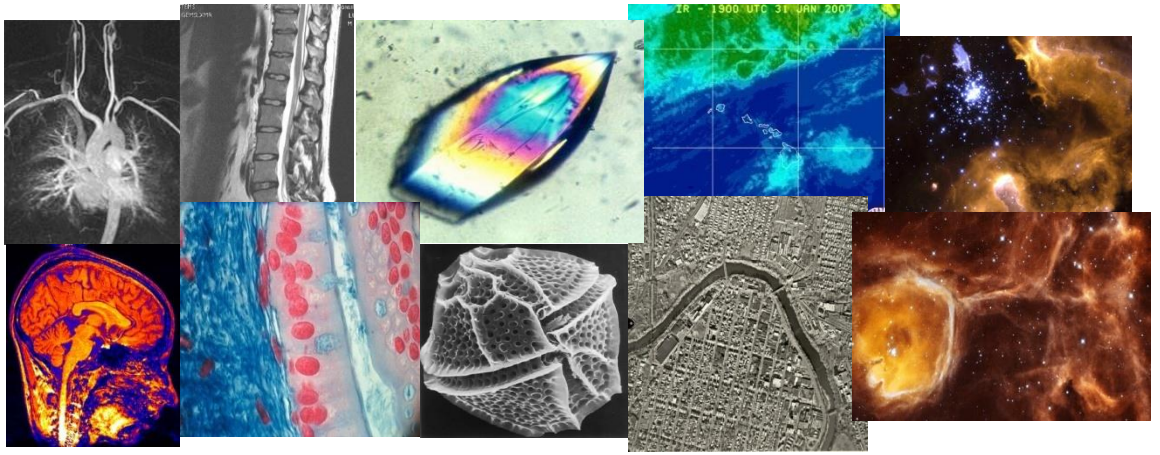
Personal photo albums



Movies, news, sports

Instagram flickr shutterstock YouTube gettyimages

Surveillance and security



Medical and scientific images

Adapted from Lana Lazebnik

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)

Things that work well

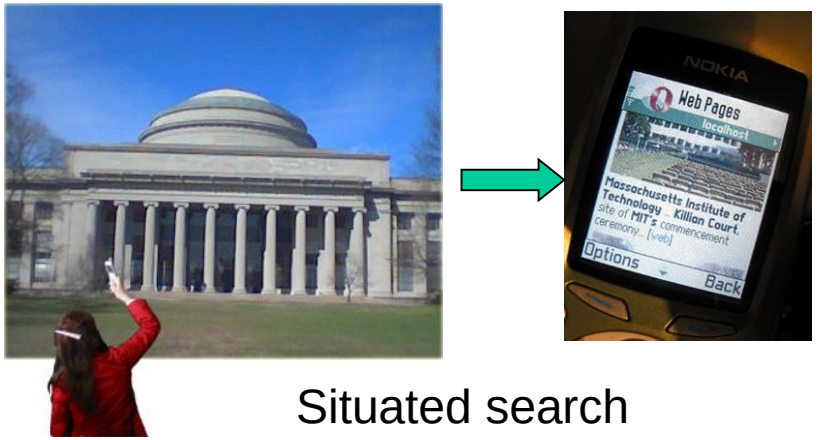
Faces and digital cameras

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.

Interactive systems

Shotton et al.

Video-based interfaces

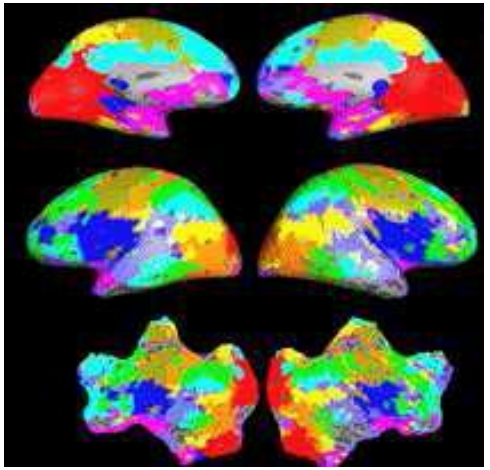
YouTube Link

Human joystick
NewsBreaker Live

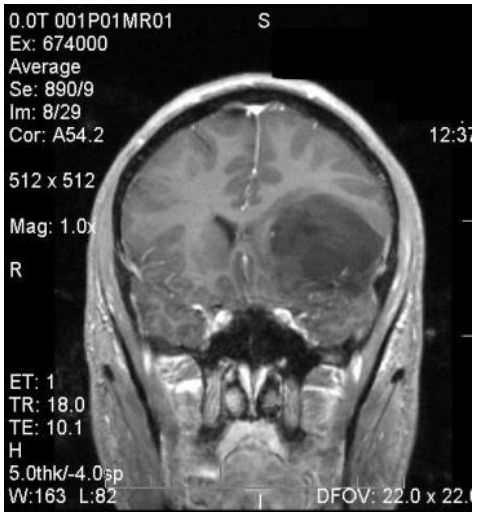
Assistive technology systems
Camera Mouse
Boston College

Vision for medical & neuroimages

Image guided surgery
MIT AI Vision Group

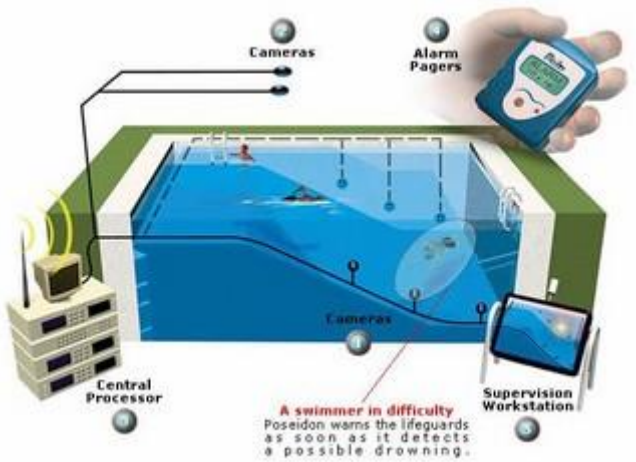


fMRI data
Golland et al.



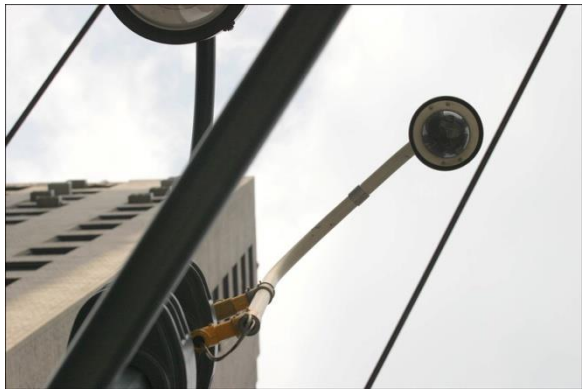
Safety & security

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

Things that need more work

The latest at CVPR* and ICCV**

* IEEE/CVF Conference on Computer Vision and Pattern Recognition
** IEEE/CVF International Conference on Computer Vision

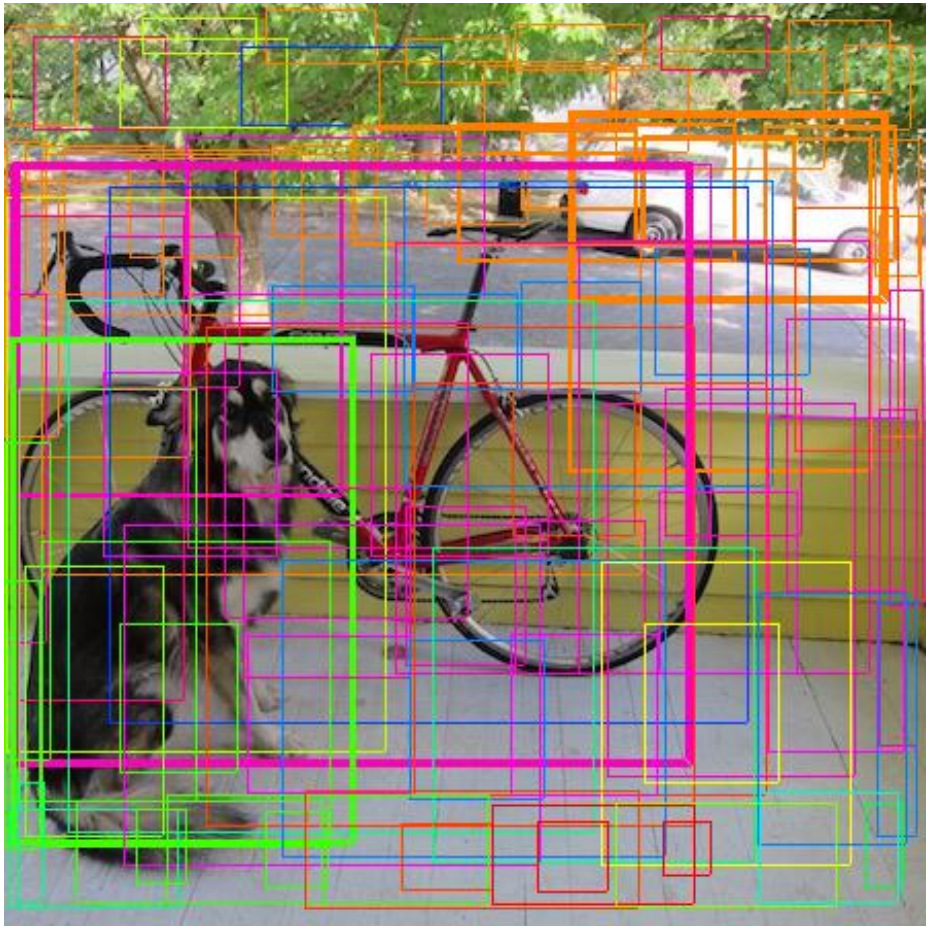
Accurate object detection in real time

	Pascal 2007 mAP	**Speed**	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img
YOLO	69.0	45 FPS	22 ms/img

2 feet

Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016

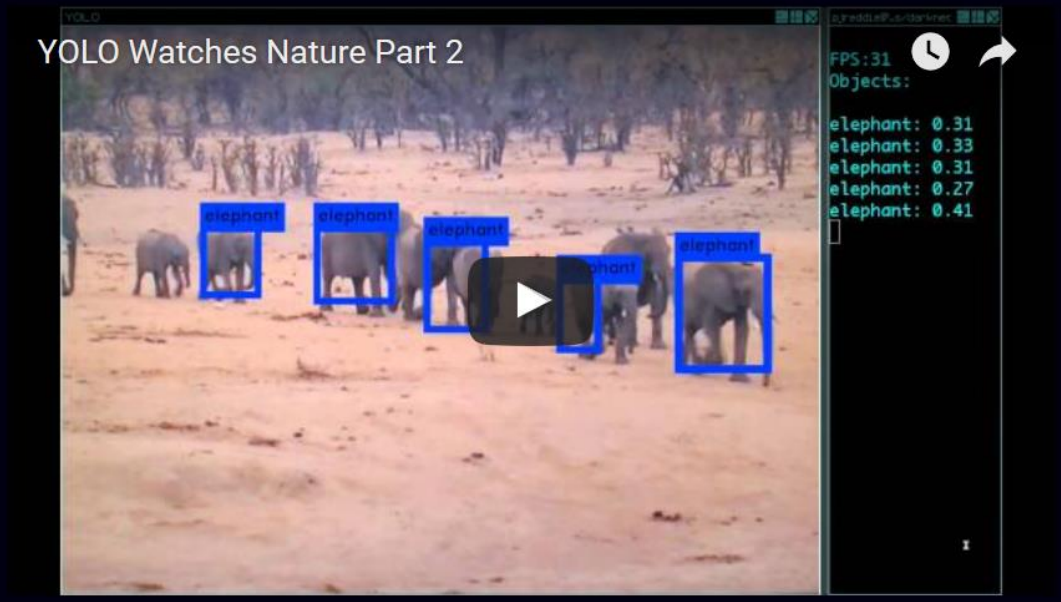
Accurate object detection in real time



Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016

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

Recognition in novel modalities

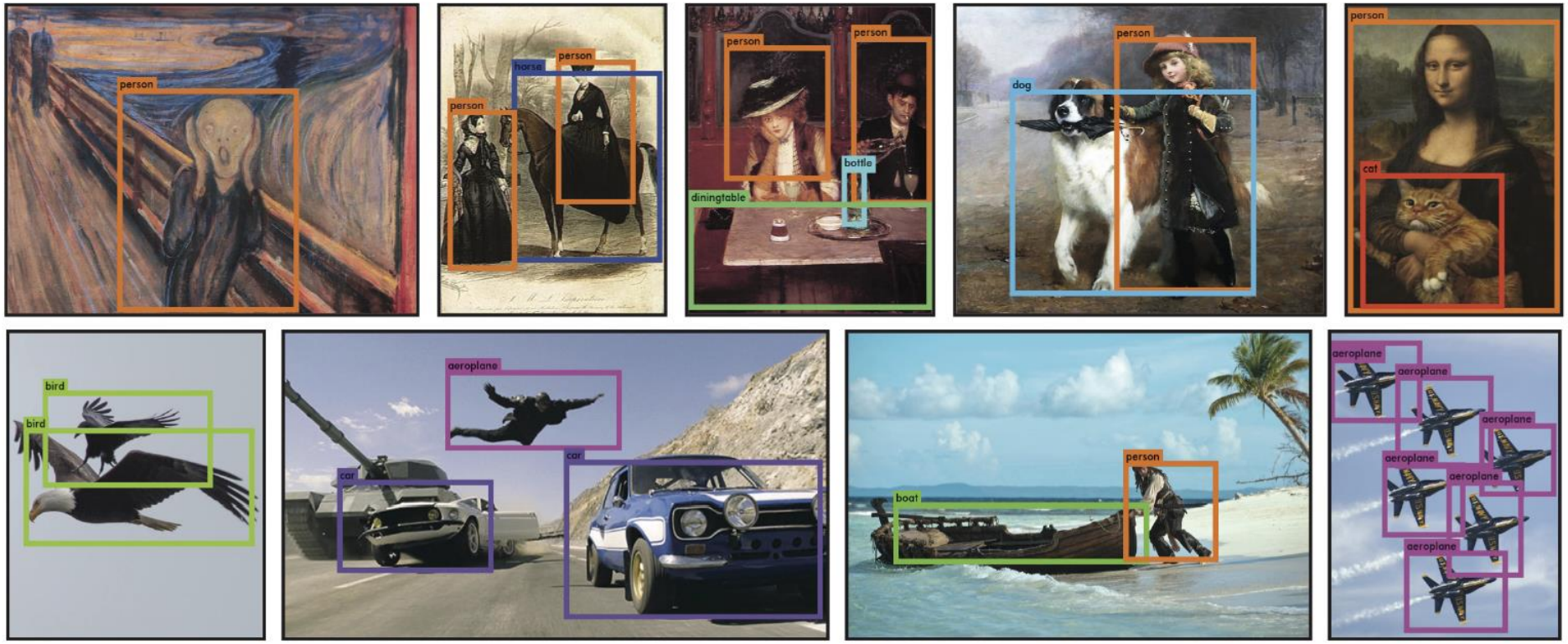
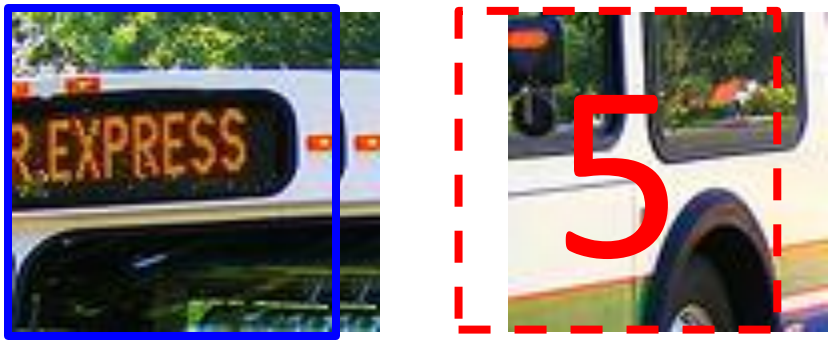


Figure 6: Qualitative Results. YOLO running on sample artwork and natural images from the internet. It is mostly accurate although it does think one person is an airplane.

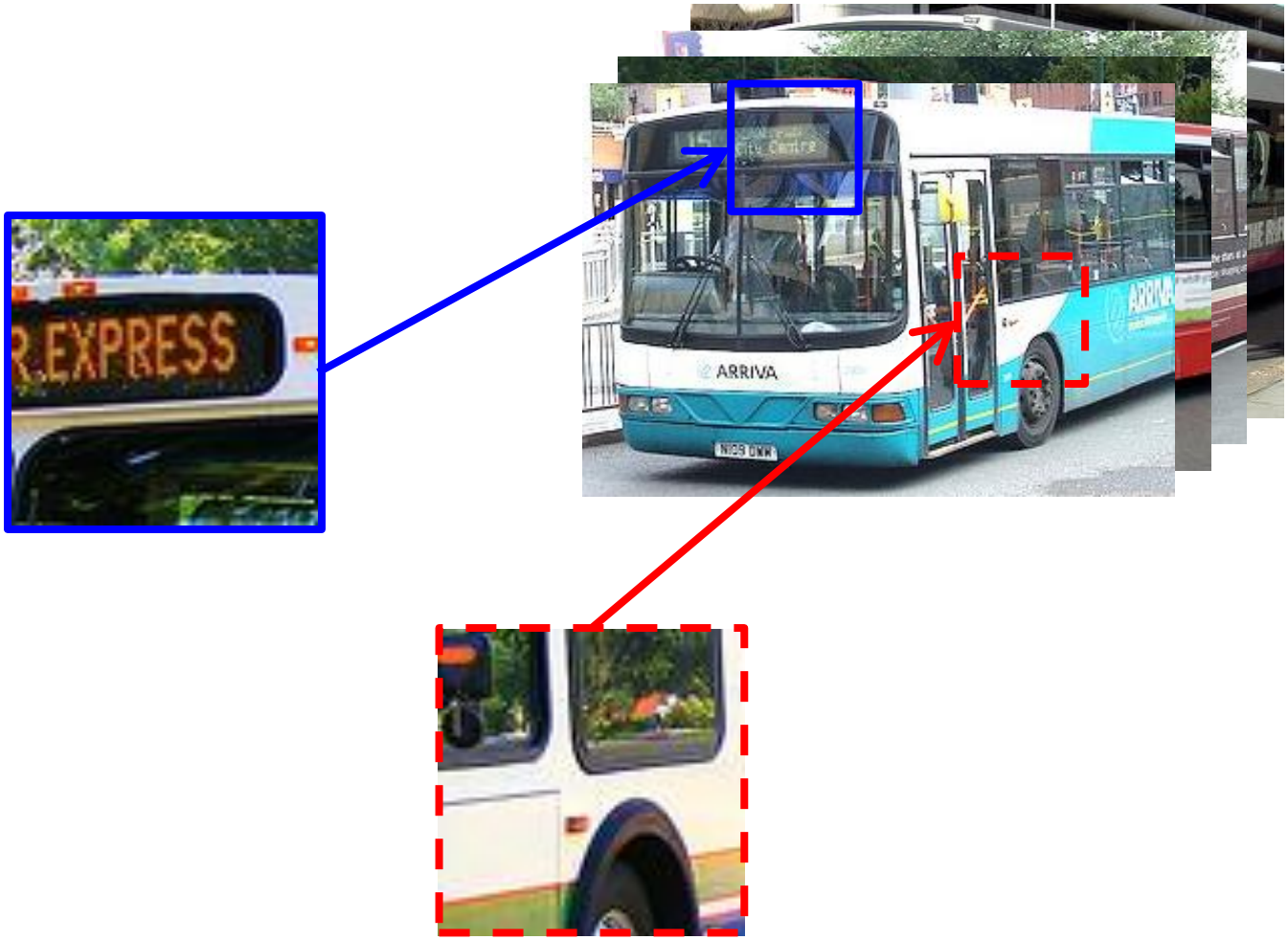
Redmon et al., CVPR 2016

Context Prediction for Images



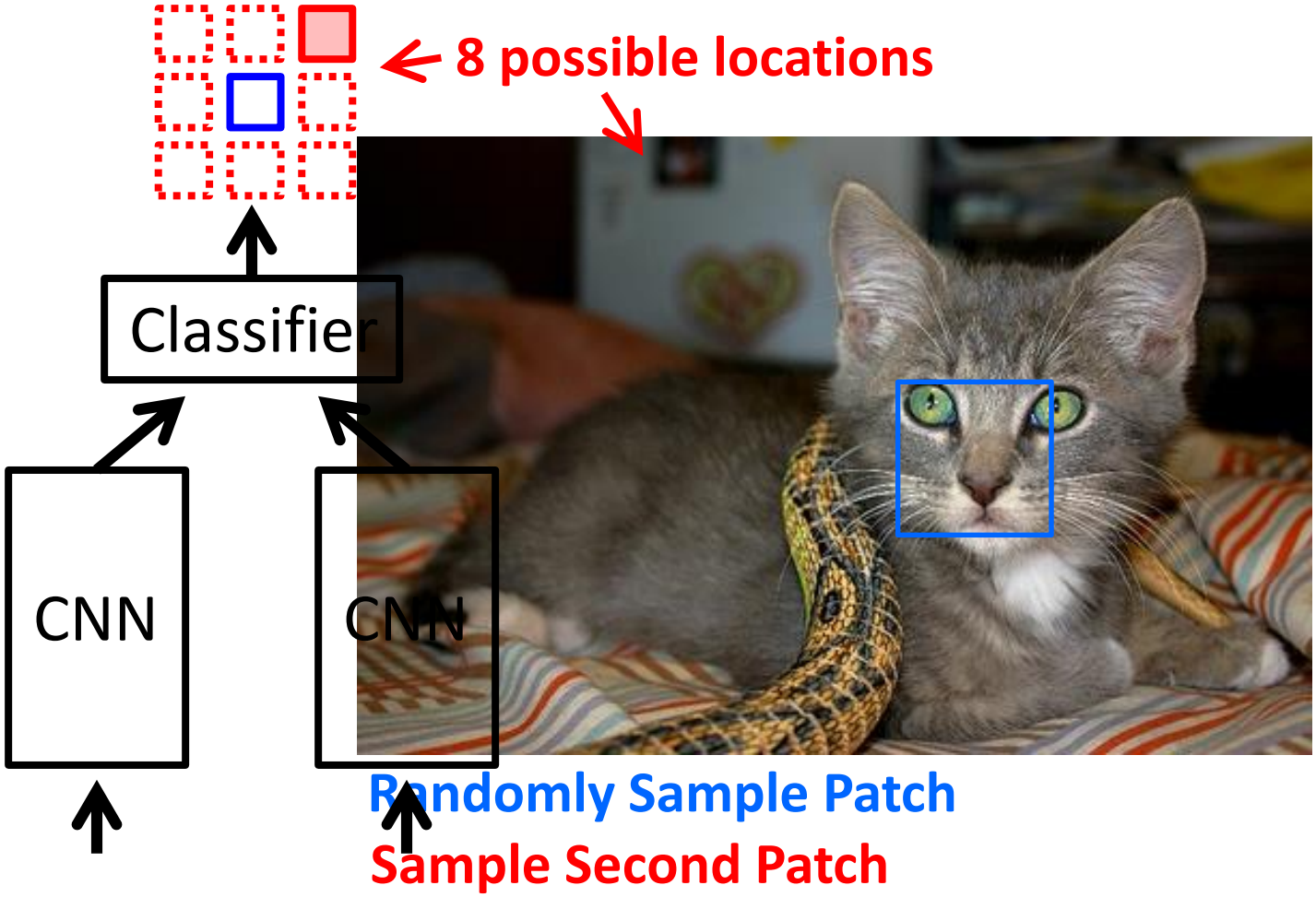
Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Semantics from a non-semantic task



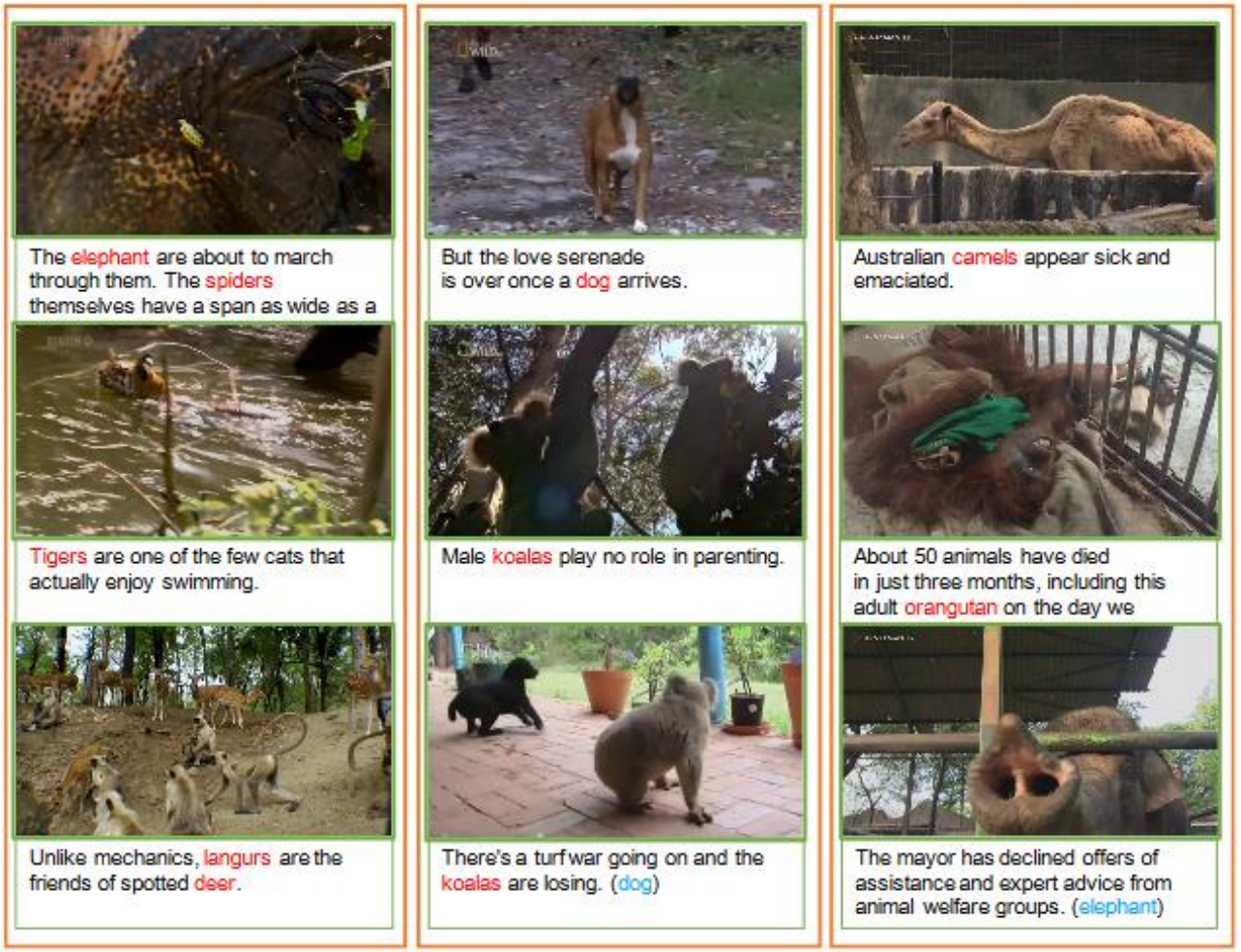
Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Relative Position Task



Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Discover and Learn New Objects from Documentaries



Chen et al., CVPR 2017

MovieQA: Understanding Stories in Movies through Question-Answering

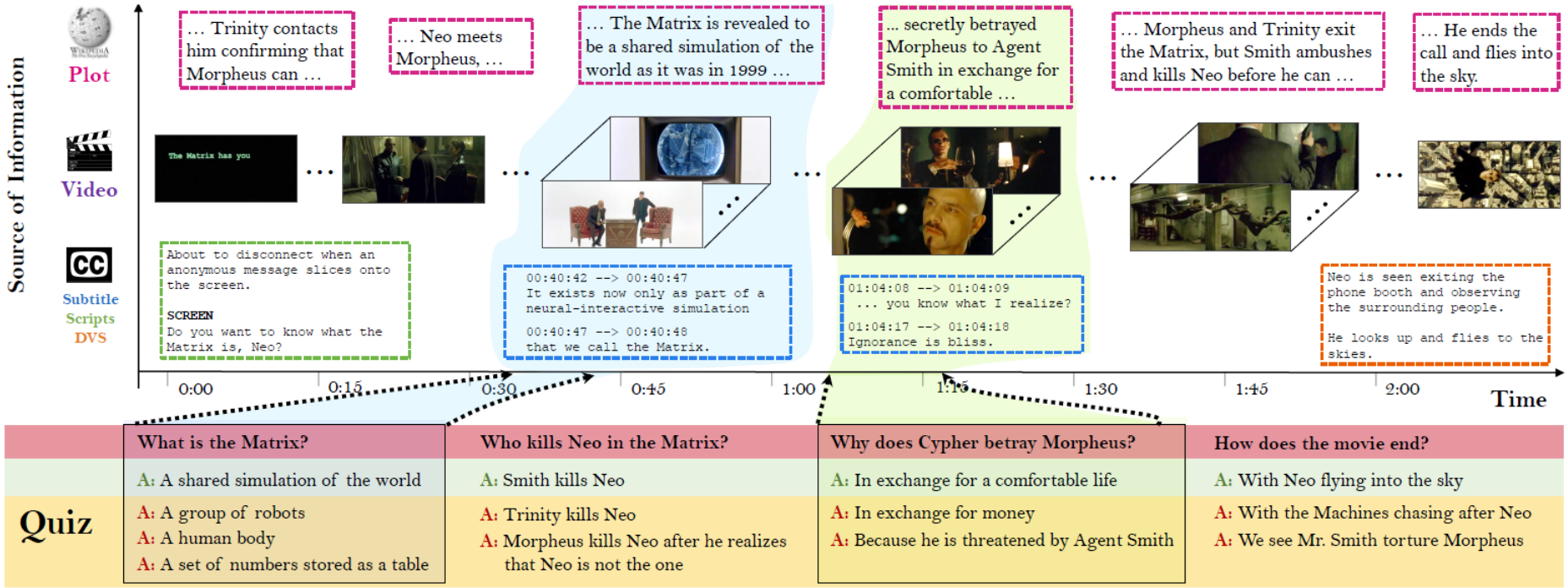


Figure 1: Our MovieQA dataset contains 14,944 questions about 408 movies. It contains multiple sources of information: plots, subtitles, video clips, scripts, and DVS transcriptions. In this figure we show example QAs from *The Matrix* and localize them in the timeline.

Tapaswi et al., CVPR 2016

Social Scene Understanding: End-To-End Multi-Person Action Localization and Collective Activity Recognition

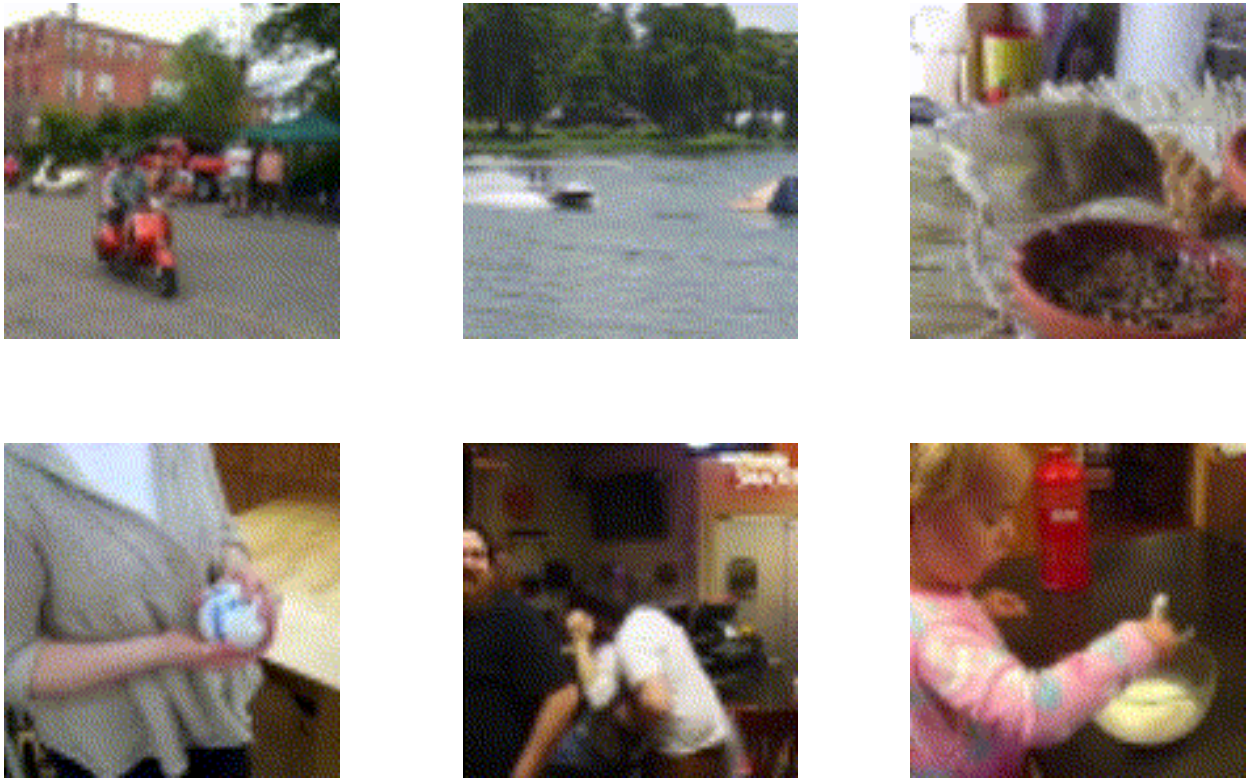
Bagautdinov et al., CVPR 2017

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.

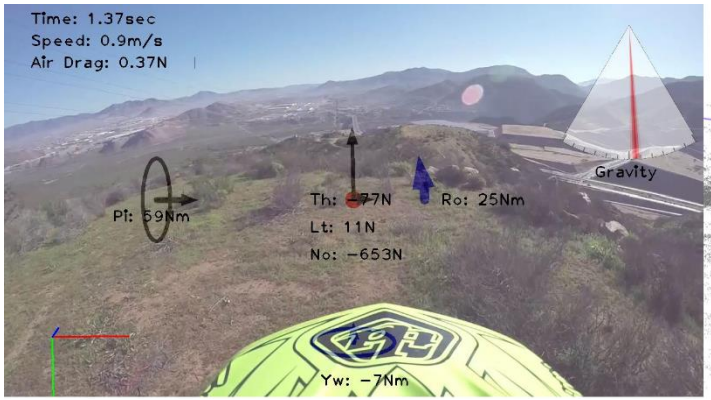
Vondrick et al., CVPR 2016

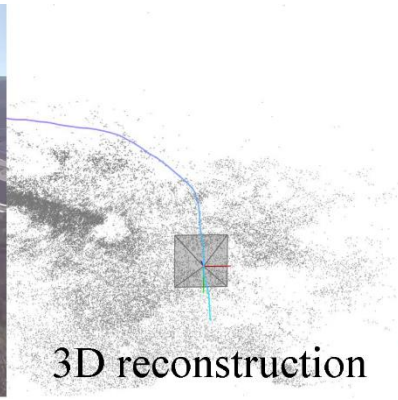
Generating the Future with Adversarial Transformers

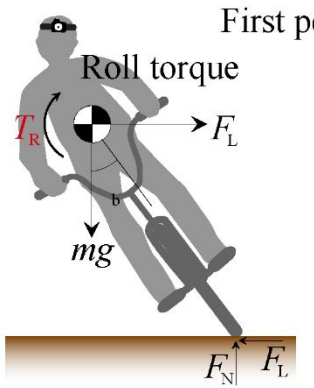


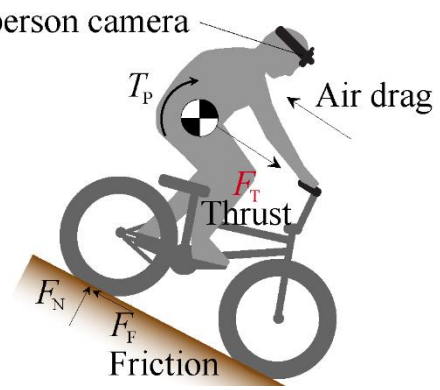
Vondrick and Torralba, CVPR 2017

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









Park et al., CVPR 2016

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

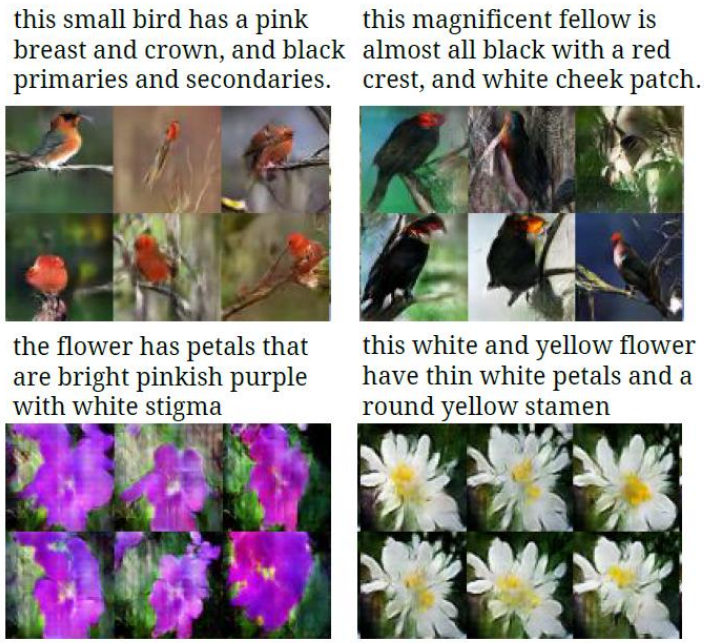
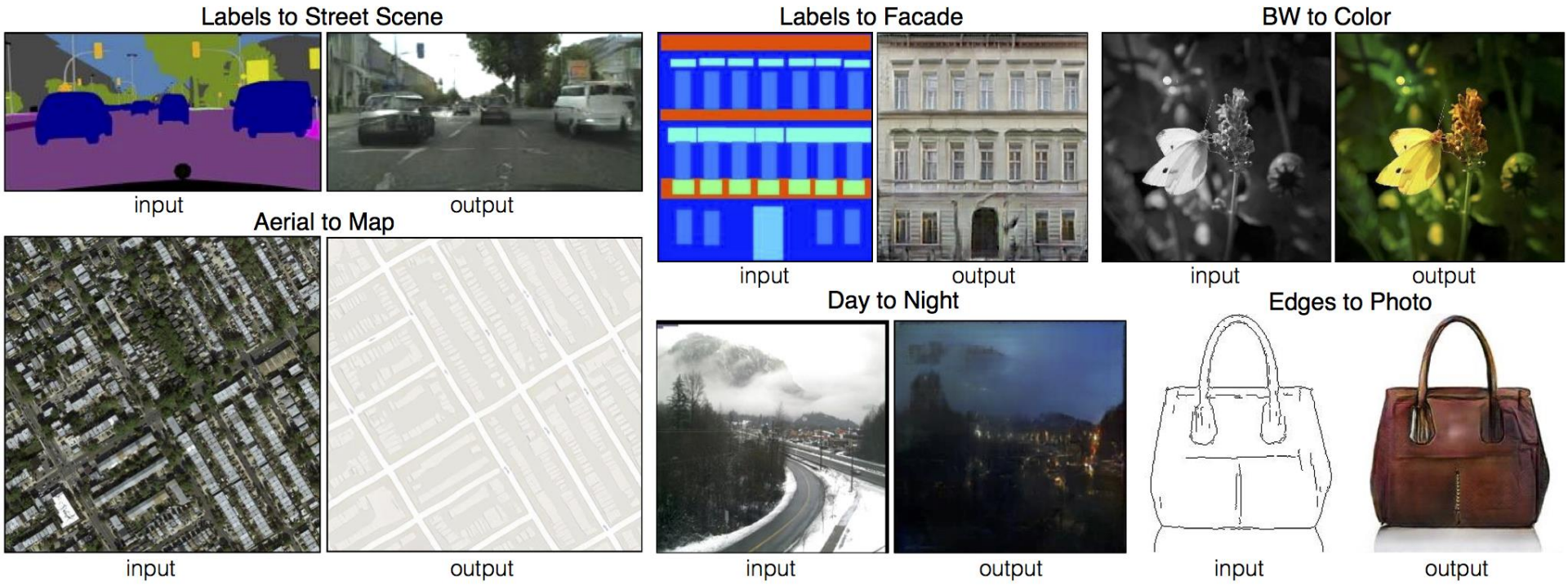


Figure 1. Examples of generated images from text descriptions. Left: captions are from zero-shot (held out) categories. Right: captions are from training set categories.

Reed et al., ICML 2016

Image-to-Image Translation with Conditional Adversarial Nets



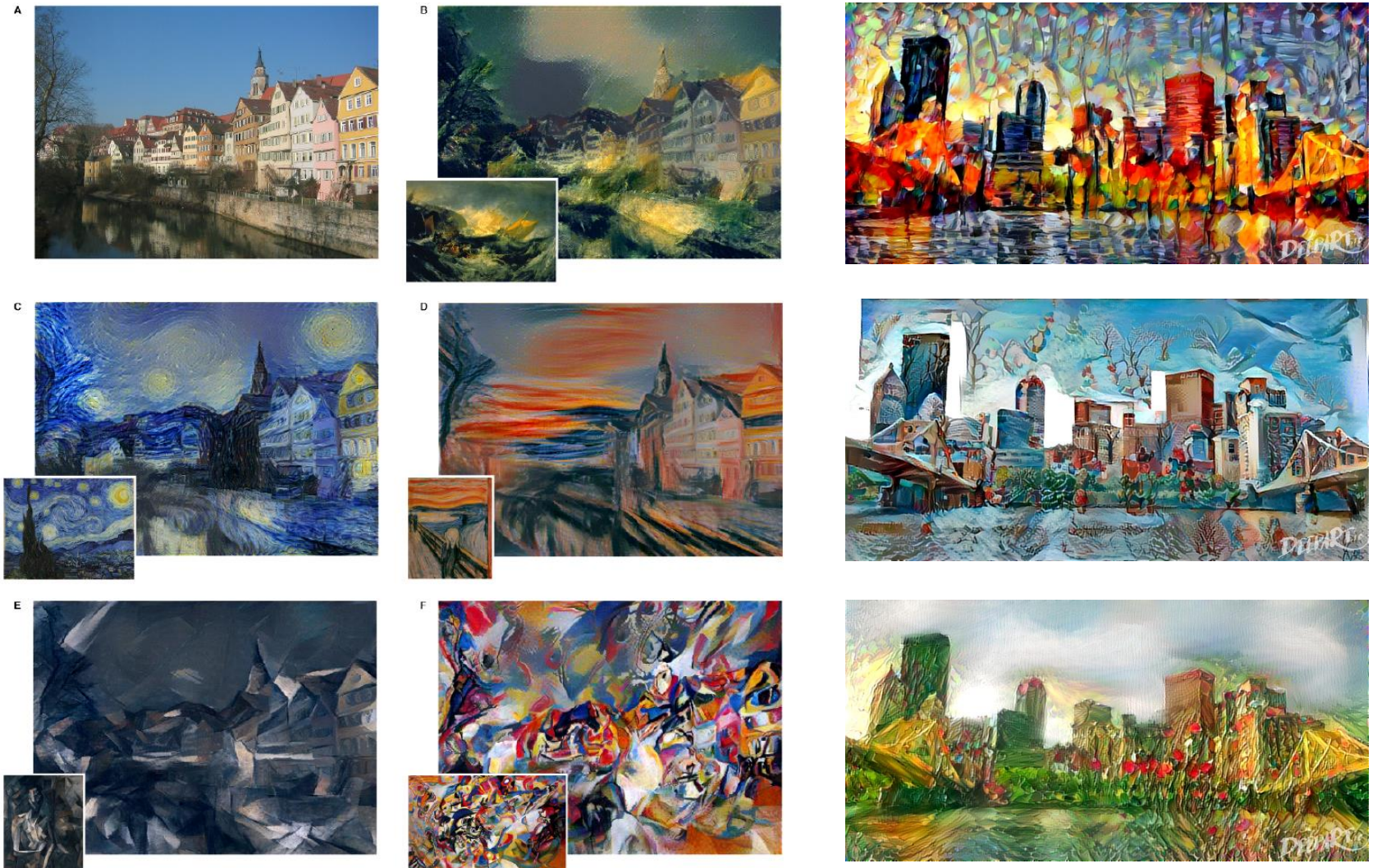
Isola et al., CVPR 2017

Scribbler: Controlling Deep Image Synthesis with Sketch and Color

Figure 1. A user can sketch and scribble colors to control deep image synthesis. On the left is an image generated from a hand drawn sketch. On the right several objects have been deleted from the sketch, a vase has been added, and the color of various scene elements has been constrained by sparse color strokes. For best resolution and additional results, see scribbler.eye.gatech.edu

Sangkloy et al., CVPR 2017

Image Style Transfer Using Convolutional Neural Networks



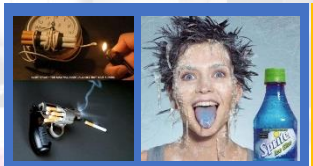
DeepArt.io – try it for yourself!

Gatys et al., CVPR 2016

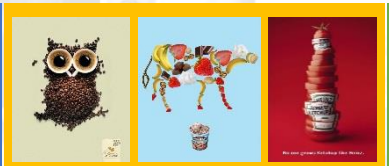
Automatic Understanding of Image and Video Advertisements

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas,
Zuha Agha, Nathan Ong, Adriana Kovashka
University of Pittsburgh

Understanding advertisements is more challenging than simply recognizing physical content from images, as ads employ a variety of strategies to persuade viewers.



Symbolism



Atypical Objects

Culture/Memes

We collect an advertisement dataset containing 64,832 images and 3,477 videos, each annotated by 3-5 human workers from Amazon Mechanical Turk.

Image	Topic	204,340	Strategy	20,000
	Sentiment	102,340	Symbol	64,131
	Q+A Pair	202,090	Slogan	11,130
Video	Topic	17,345	Fun/Exciting	15,380
	Sentiment	17,345	English?	17,374
	Q+A Pair	17,345	Effective	16,721

Here are some sample annotations in our dataset.

What's being advertised in this image?

Cars, automobiles

What sentiments are provoked in the viewer?

Amused, Creative, Impressed, Youthful, Conscious

What strategies are used to persuade viewer?

Symbolism, Contrast, Straightforward, Transferred qualities

What should the viewer do, and why should they do this?

- I should buy Volkswagen because it can hold a big bear.
- I should buy VW SUV because it can fit anything and everything in it.
- I should buy this car because it can hold everything I need.

More information available at http://cs.pitt.edu/~kovashka/ads

Hussein et al., CVPR 2017

Is computer vision solved?

- Given an image, we can guess with 96% accuracy what object categories are shown (ResNet)
- … but we only answer “why” questions about images with 14% accuracy!

Why does it seem like 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

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".

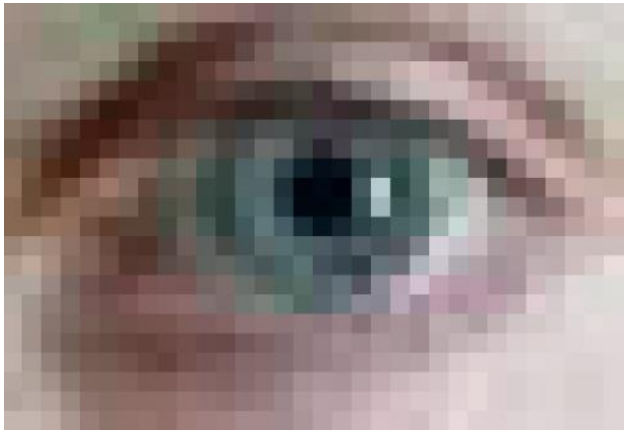
Read more about the history: Szeliski Sec. 1.2

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

What the computer gets



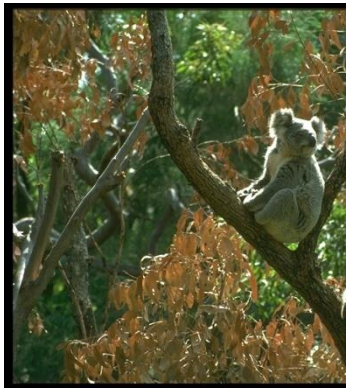
Why is this problematic?

Adapted from Kristen Grauman and Lana Lazebnik

Challenges: many nuisance parameters

Illumination

Object pose



Clutter

Occlusions

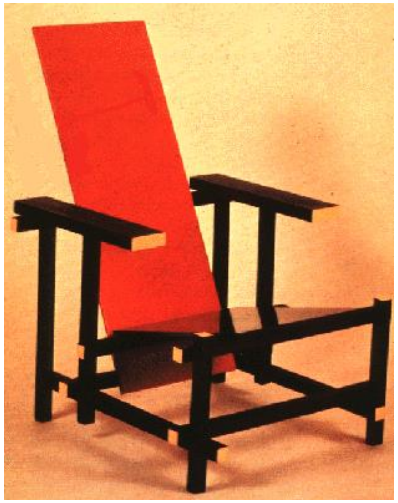
Intra-class appearance

Viewpoint

Think again about the pixels…

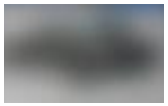
Challenges: intra-class variation

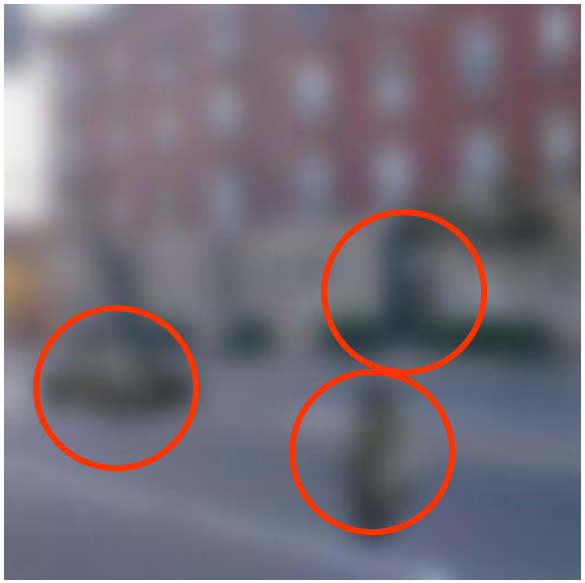
CMOA Pittsburgh



slide credit: Fei-Fei, Fergus & Torralba

Challenges: importance of context





slide credit: Fei-Fei, Fergus & Torralba

Challenges: Complexity

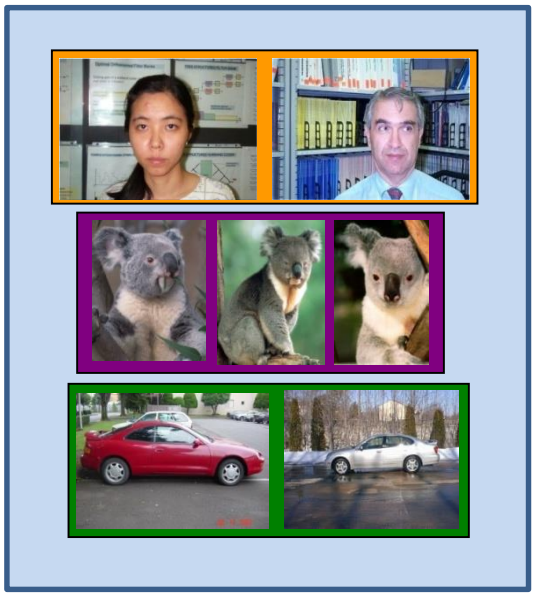
- Thousands to millions of pixels in an image
- 3,000-30,000 human recognizable object categories
- 30+ degrees of freedom in the pose of articulated objects (humans)
- Billions of images indexed by Google Image Search
- 1.424 billion smart camera phones sold in 2015
- About half of the cerebral cortex 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?

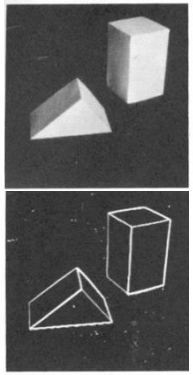
Is this person expecting company?
What is just under the tree?

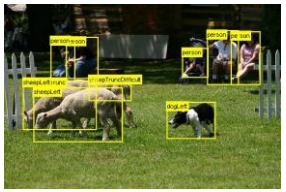
Does it appear to be rainy?
Does this person have 20/20 vision?

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

Evolution of datasets

- Challenging problem → active research area





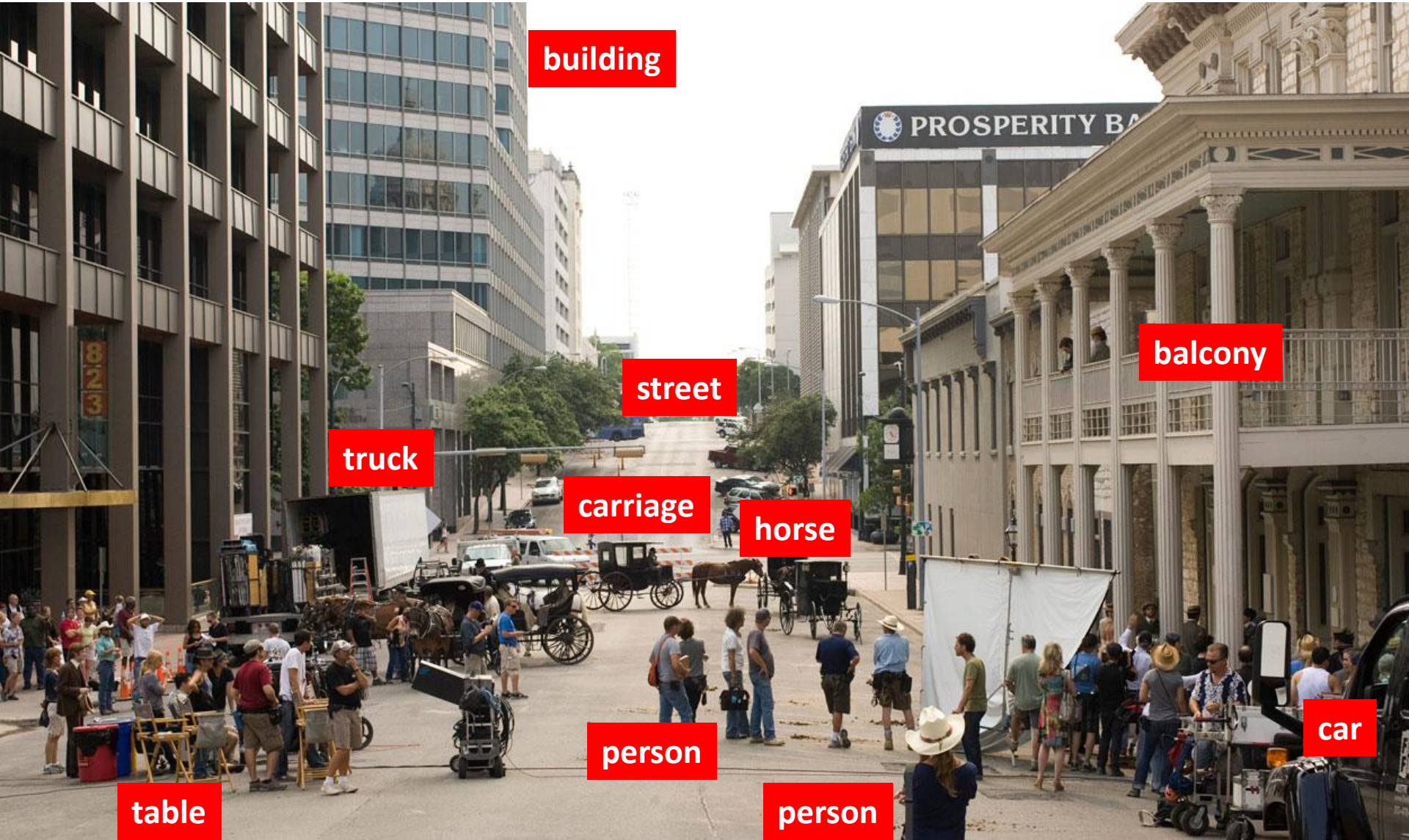
PASCAL:
20 categories, 12k images

ImageNet:
22k categories, 14mil images

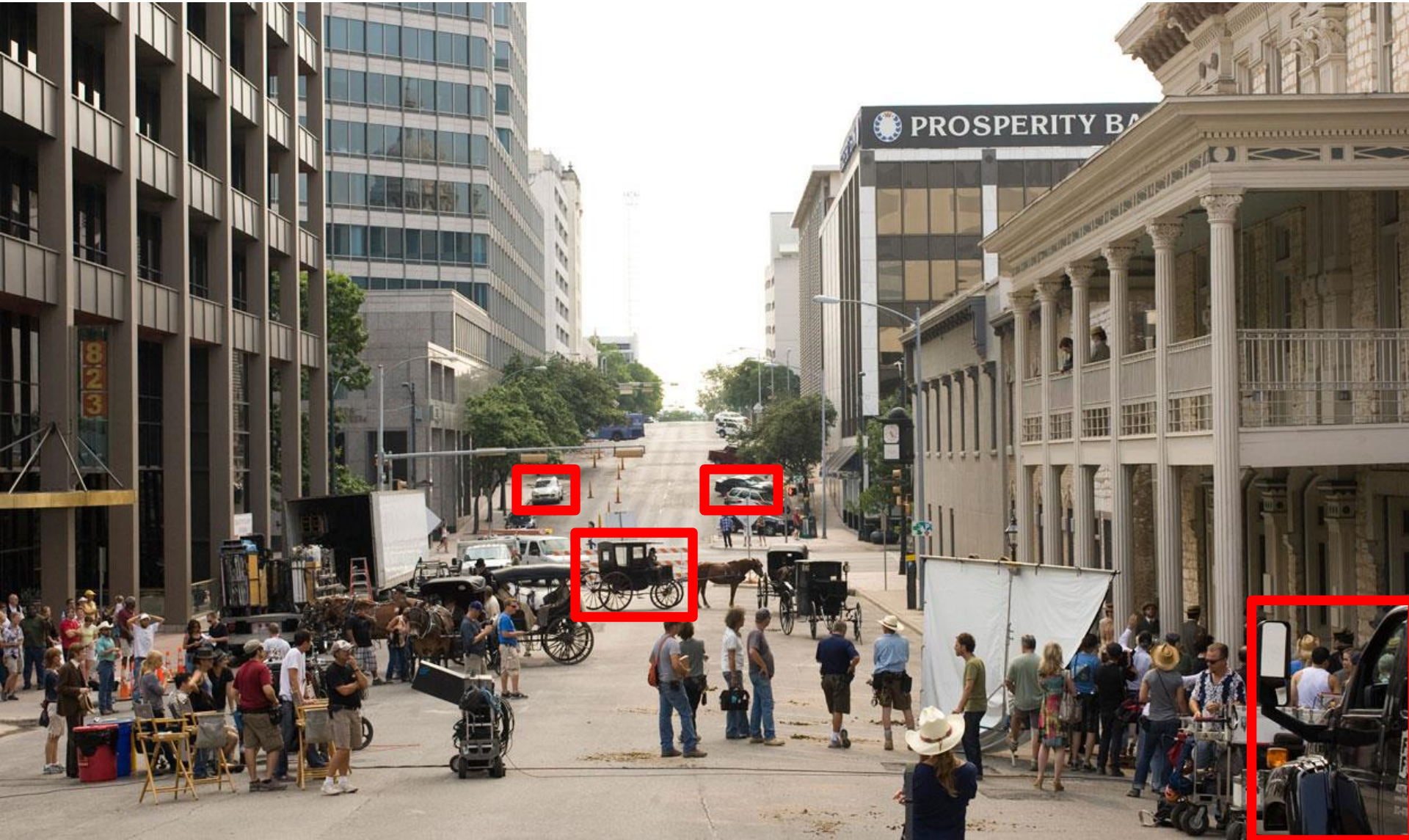
Microsoft COCO:
80 categories, 300k images

Some Visual Recognition Problems: Why are they challenging?

Recognition: What objects do you see?



Detection: Where are the cars?



Activity: What is this person doing?

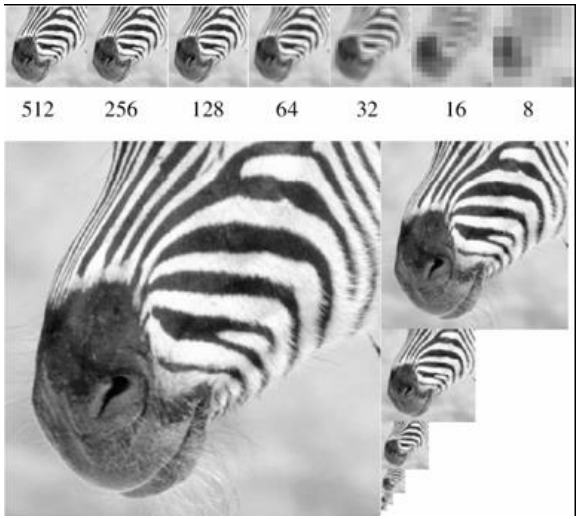
Scene: Is this an indoor scene?

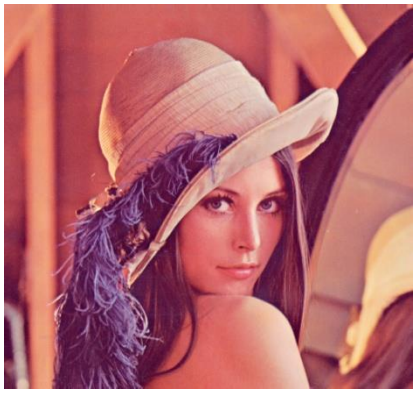
Instance: Which city? Which building?

Visual question answering:
Why is there a carriage in the street?

Overview of topics

Features and filters

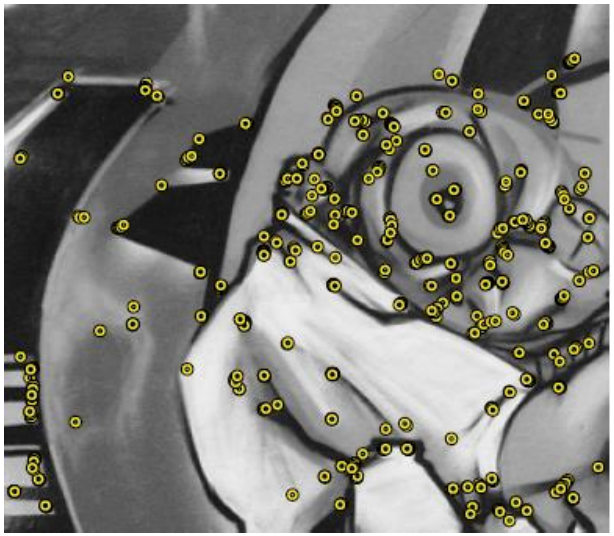


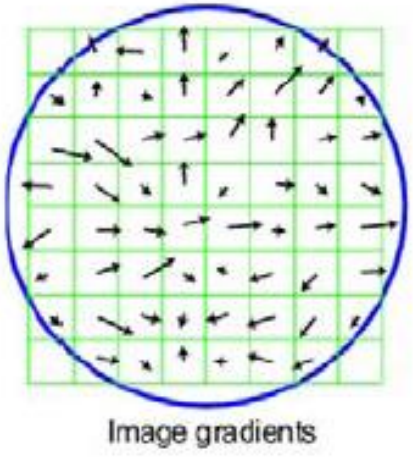


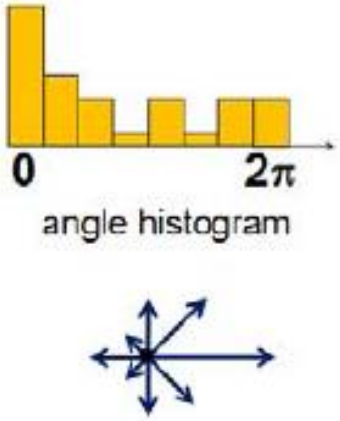
- Transforming and describing images; textures, colors, edges

Features and filters

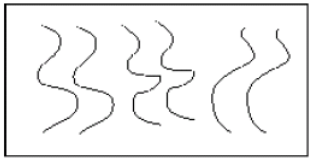
- Detecting distinctive + repeatable features
- Describing images with local statistics



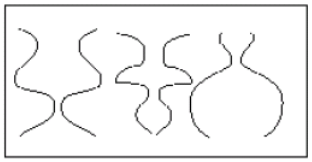




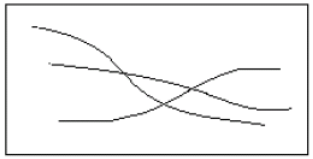
Grouping and fitting



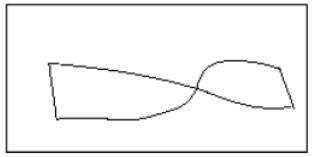
Parallelism



Symmetry



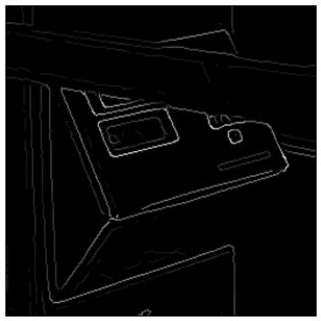
Continuity

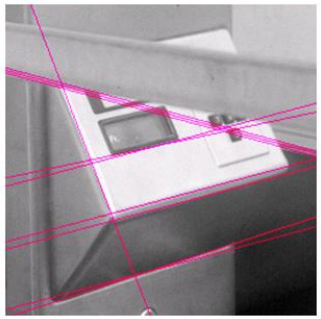


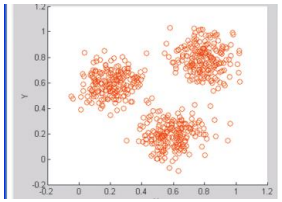
Closure

[fig from Shi et al]

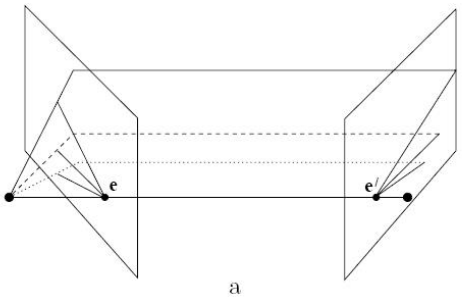
- Clustering, segmentation, fitting; what parts belong together?



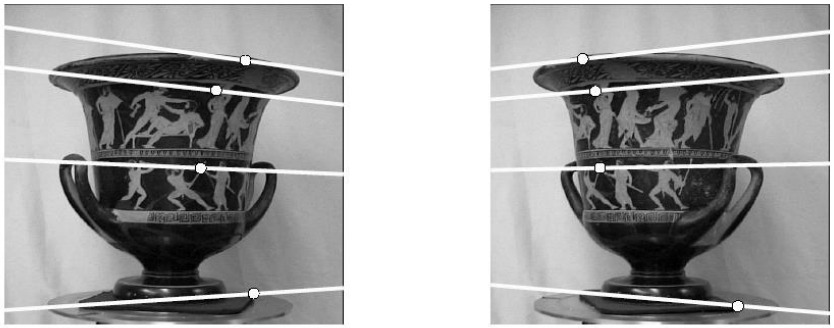




Multiple views



a



Hartley and Zisserman

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

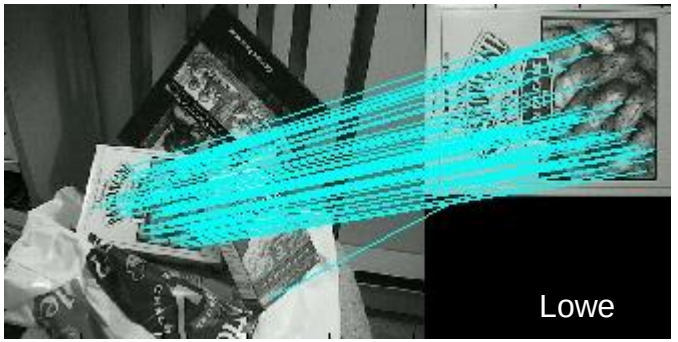
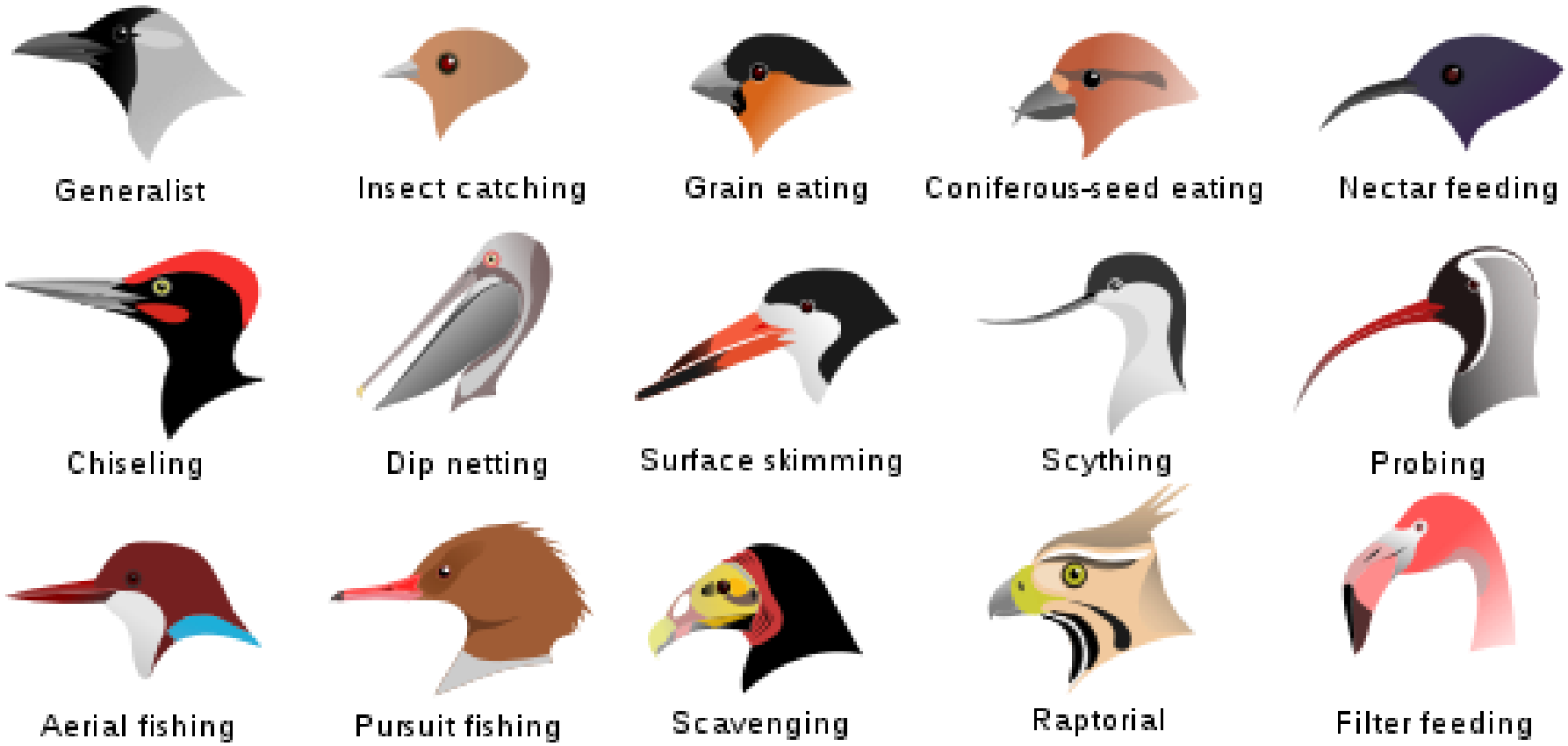


Image categorization

- Fine-grained recognition



Visipedia Project

Slide credit: D. Hoiem

Region categorization

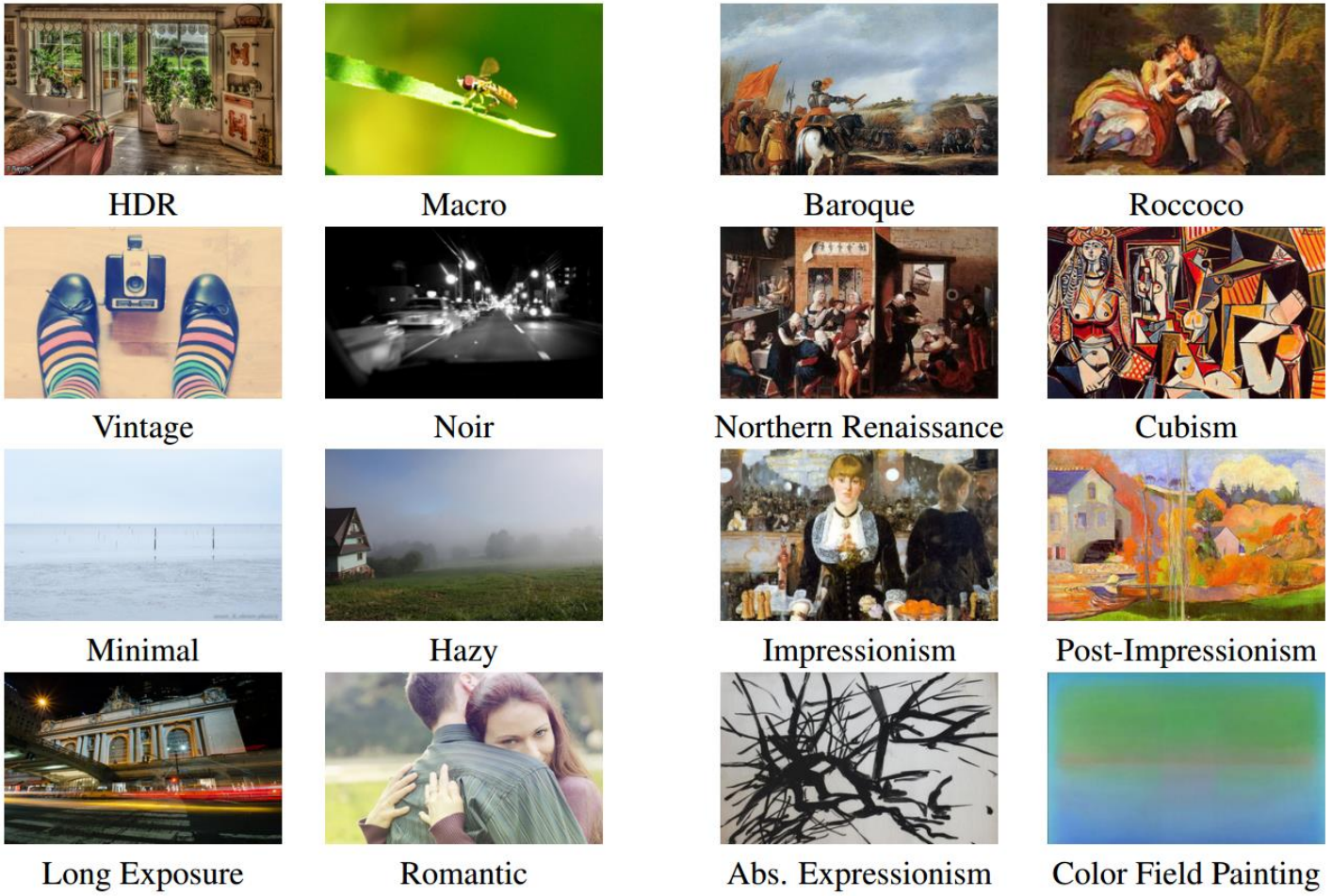
- Material recognition

[Bell et al. CVPR 2015]

Slide credit: D. Hoiem

Image categorization

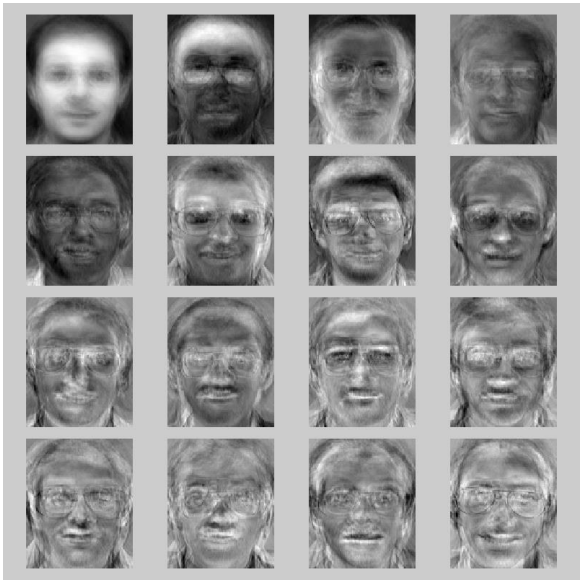
- Image style recognition

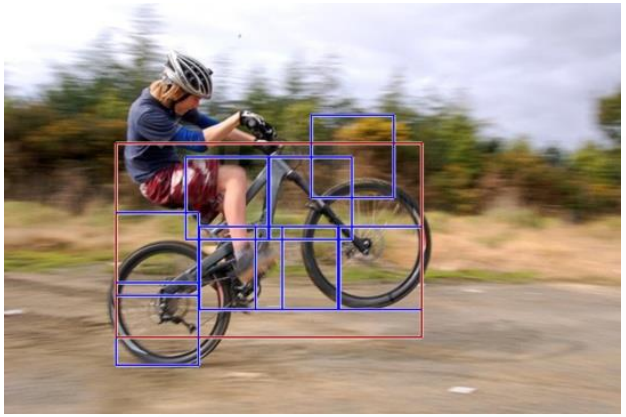


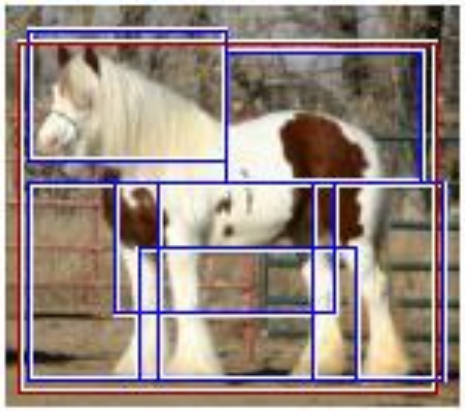
[Karayev et al. BMVC 2014]

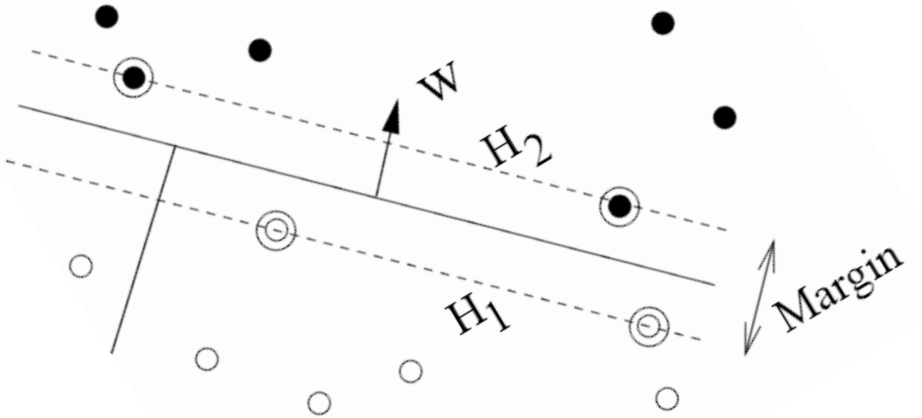
Slide credit: D. Hoiem

Visual recognition and SVMs









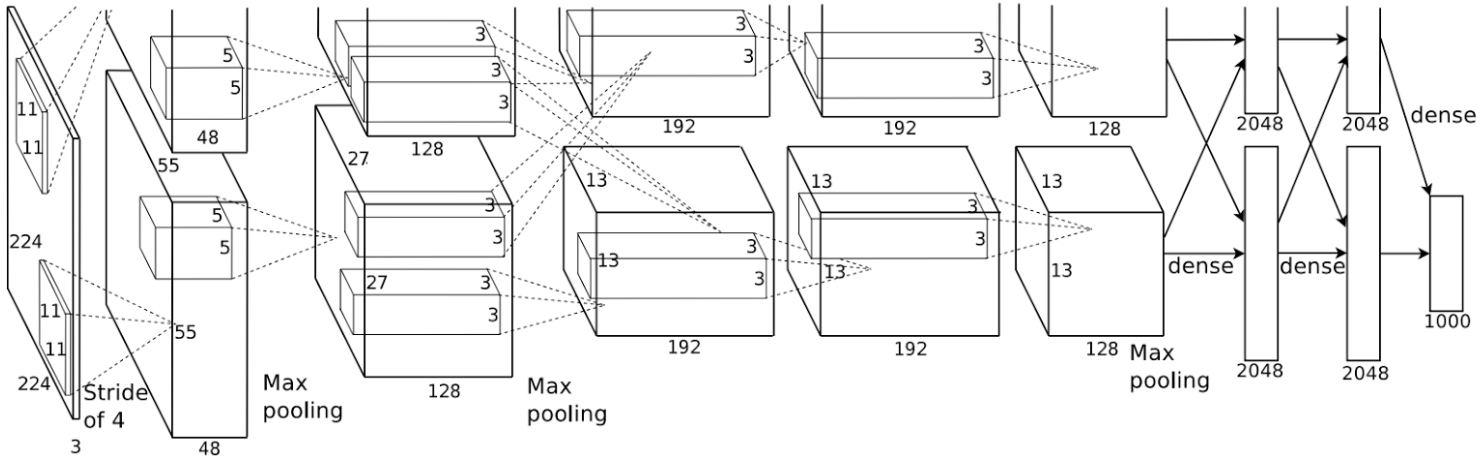
- Recognizing objects and categories, learning techniques

Adapted from Kristen Grauman

Convolutional neural networks (CNNs)

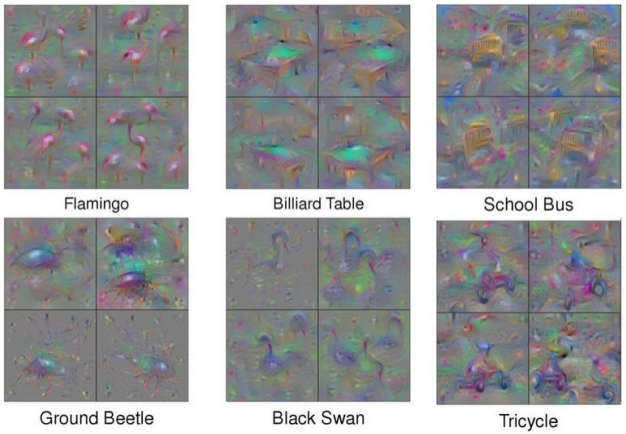
- State-of-the-art on many recognition tasks

Image



Prediction

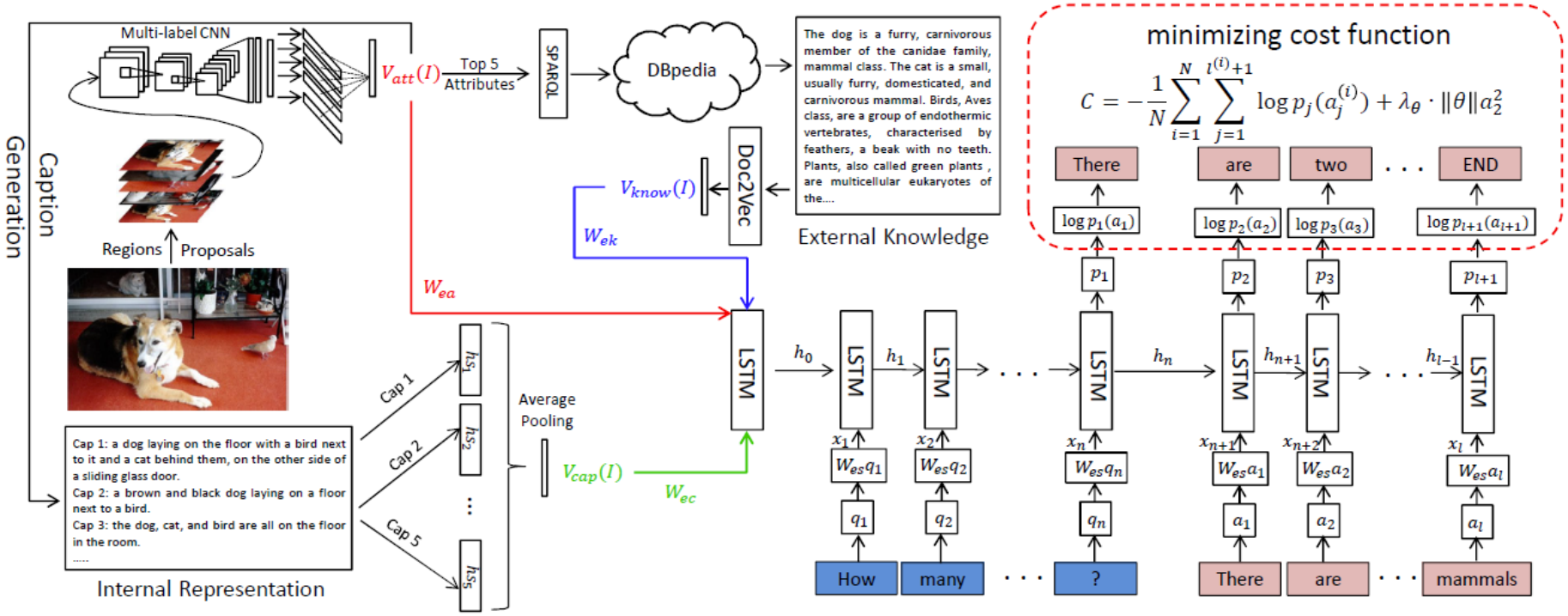
Krizhevsky et al., NIPS 2012



Yosinski et al., ICML DL workshop 2015

Recurrent neural networks

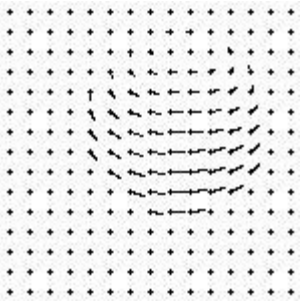
- Sequence processing, e.g. question answering

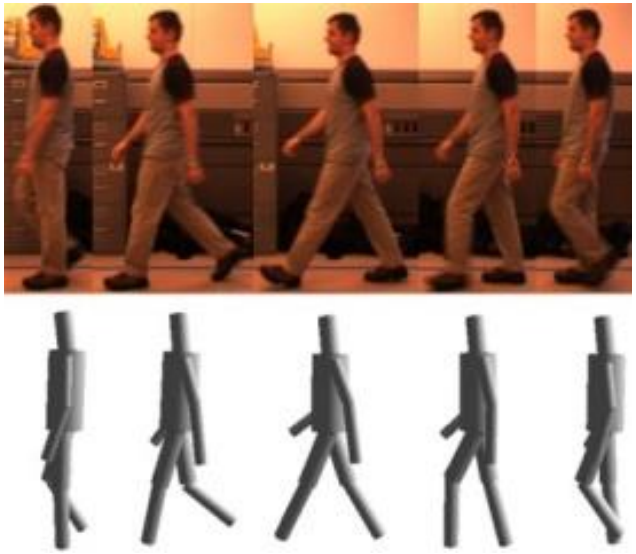


Wu et al., CVPR 2016

Motion and tracking

- Tracking objects, video analysis

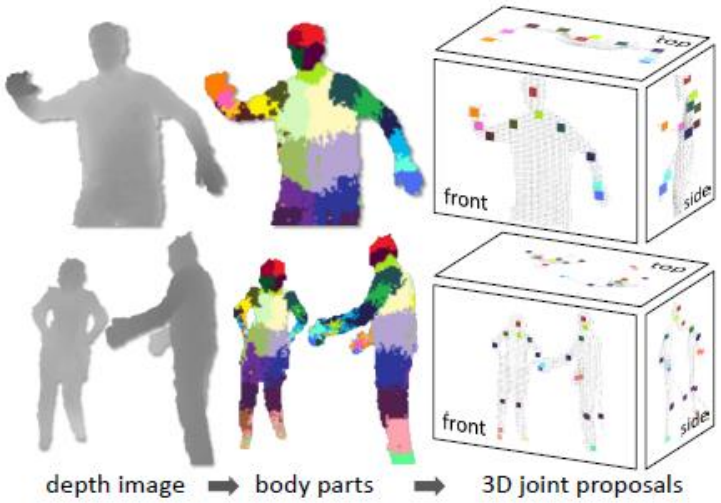


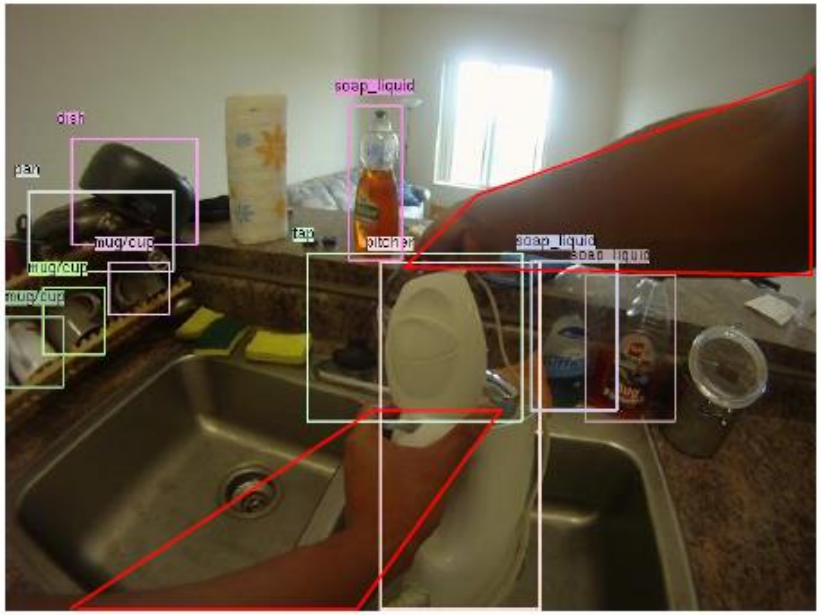


Tomas Izo

Pose and actions

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





Your Homework

- Fill out Doodle
- Read entire course website
- Do first reading