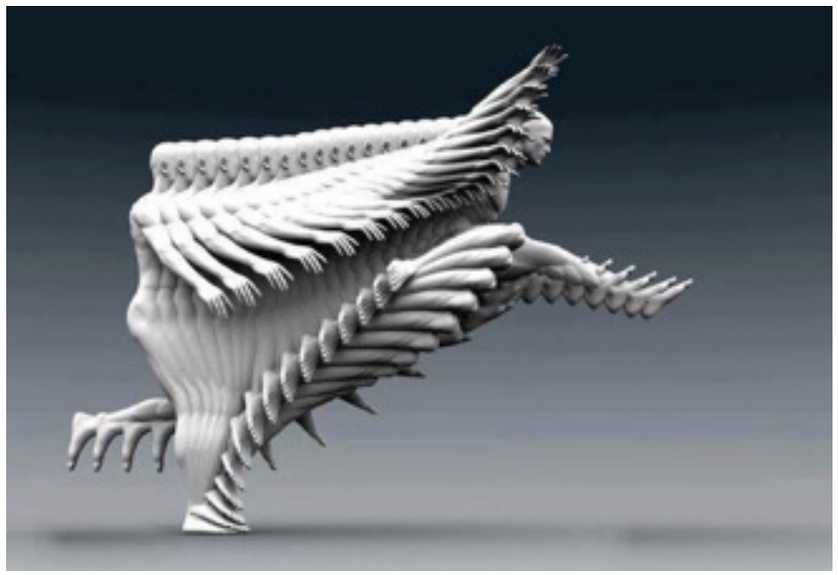
16-899H: Human Activity Analysis



"The single, frozen feild of view provides only impoverished information about the world. The visual system did not evolve for this."

J. J. Gibson *Ecological Approach to Visual Perception*

Me

Deva Ramanan
Associate Professor
Robotics Institute
Carnegie Mellon University
Elliot Dunlap Smith Hall (EDSH), Rm 221
deva@cs.cmu.edu
412-268-6966
Mailing address

Bio

I recently moved to CMU after spending 8 wonderful years as a faculty at UC-Irvine. A more formal bio is here.

Research

My research focuses on computer vision, often motivated by the task of understanding people from visual data. My work tends to make heavy use of machine learning techniques, often using the human visual system as inspiration. For example, temporal processing is a key component of human perception, but is still relatively unexploited in current visual recognition systems. Machine learning from big (visual) data allows systems to learn subtle statistical regularities of the visual world. But humans have the ability to learn from very few examples. Here's a recent talk (from 2015) that discusses some thoughts on these issues.

And you?

Format of course

http://16899.courses.cs.cmu.edu/

Seminar: everyone will present 1-3 times
(depending on enrollment)

Course project: you are free to use your research, but it should some have element of video analysis
[take this opportunity to try something new!]

Goals for the course

Mixture of:

Vision classics
Frontiers of research

Gain experience with:

Presenting a topic in-depth
Research discussions

The goal

"understanding" this video

Some thoughts...

Video analysis has traditionally taken
a back seat to image analysis

Why?
Storage and processing costs.
If we can't process images, how can we process frames?

Some thoughts...

Video analysis has traditionally taken
a back seat to image analysis

Why?

Storage and processing costs.

Much less of an issue, but still a nuisance

If we can't process images, how can we process frames?

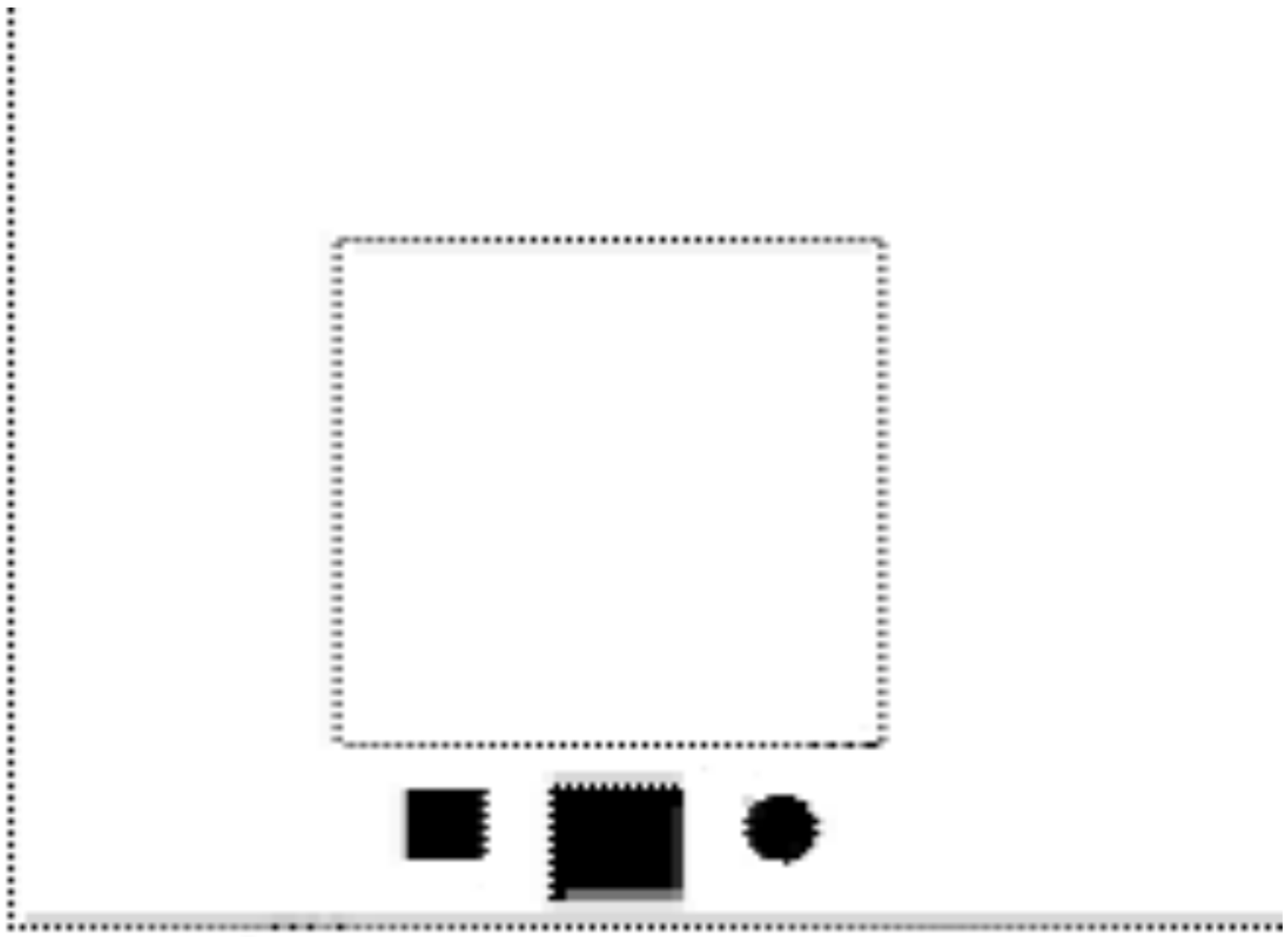
Static image processing has rapidly improved

Video repositories...

8 years worth of video is uploaded to YouTube... each day

A YouTube-size repository (# of vids) is uploaded to social media sites (Vine, Instagram).. every 3 months

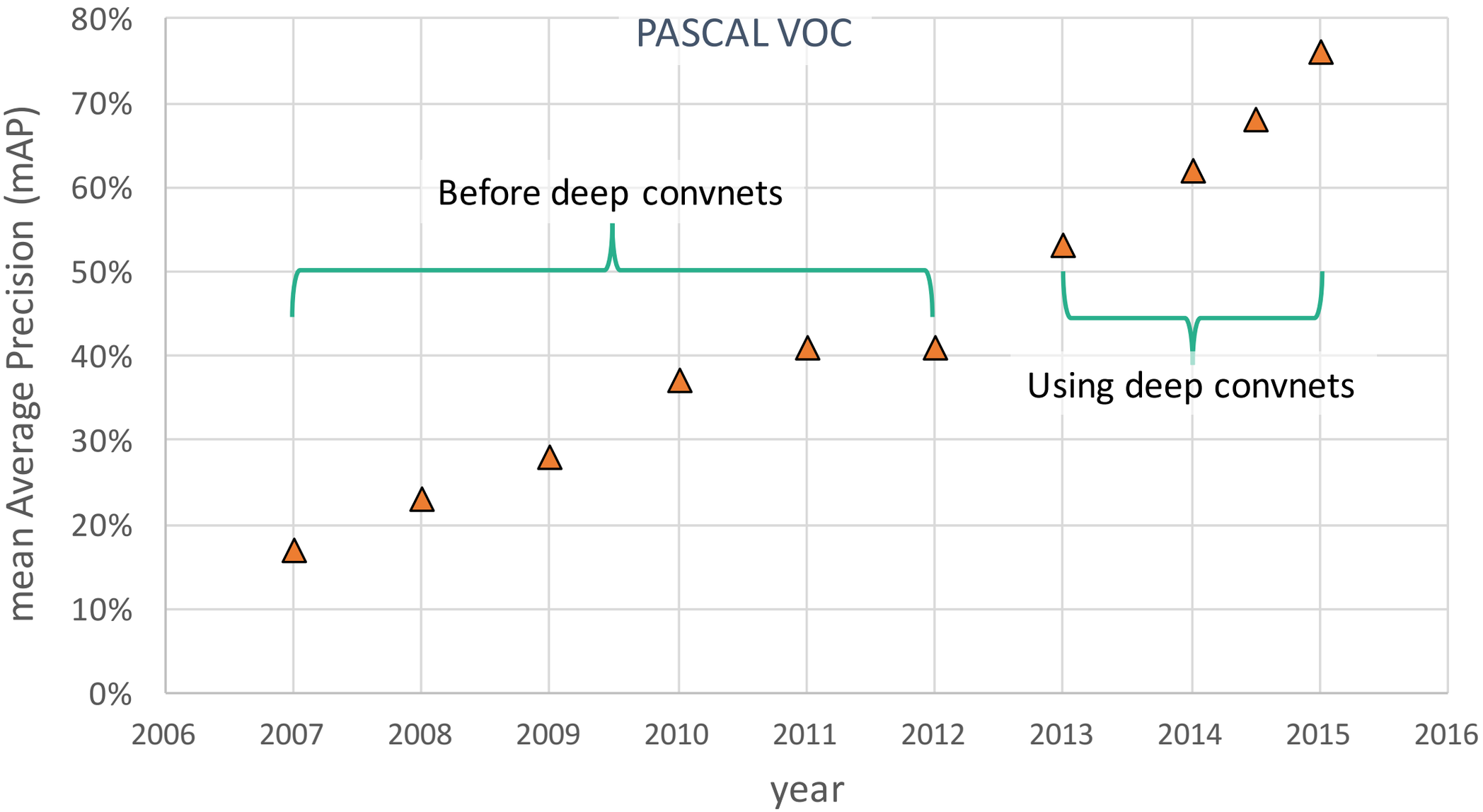
Activity recognition with weak image-based features



Recognition without almost no image-based features

(in that a static image in uninterpretable)

Today's world of deep learning



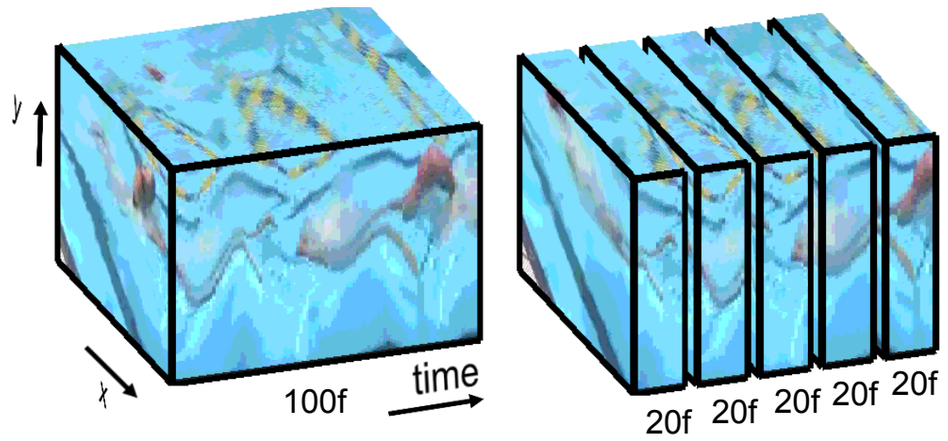
Biological motivation

Hubel and Weisel's iconic experiments on simple vs complex "pooling"cells

Complex cells are tuned to movement

"Clicks" are action potentials generated by instrumented cortical neuron

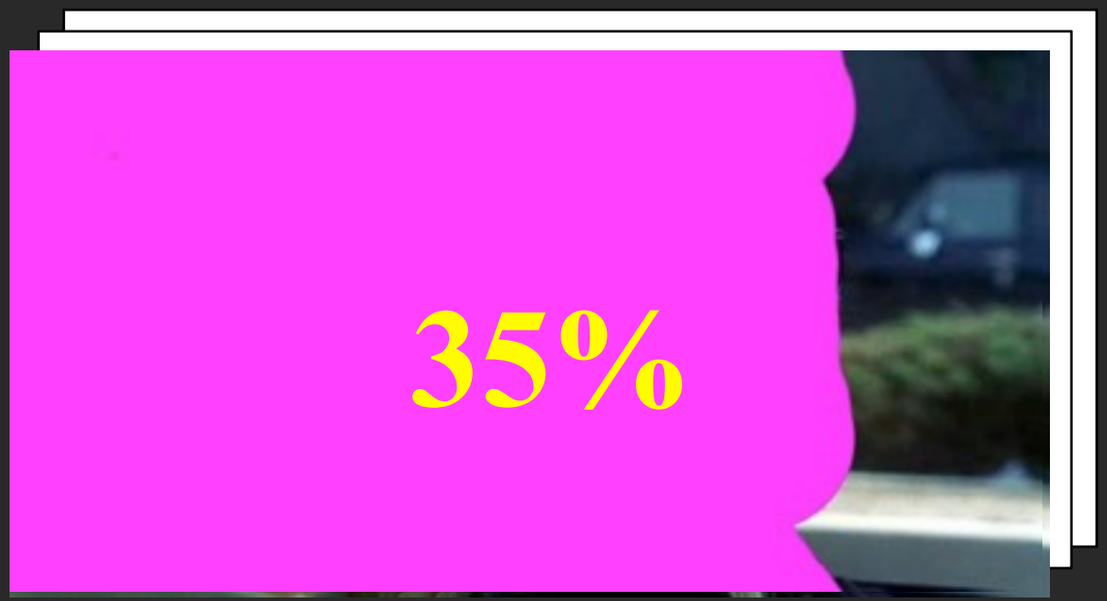
Deep video features



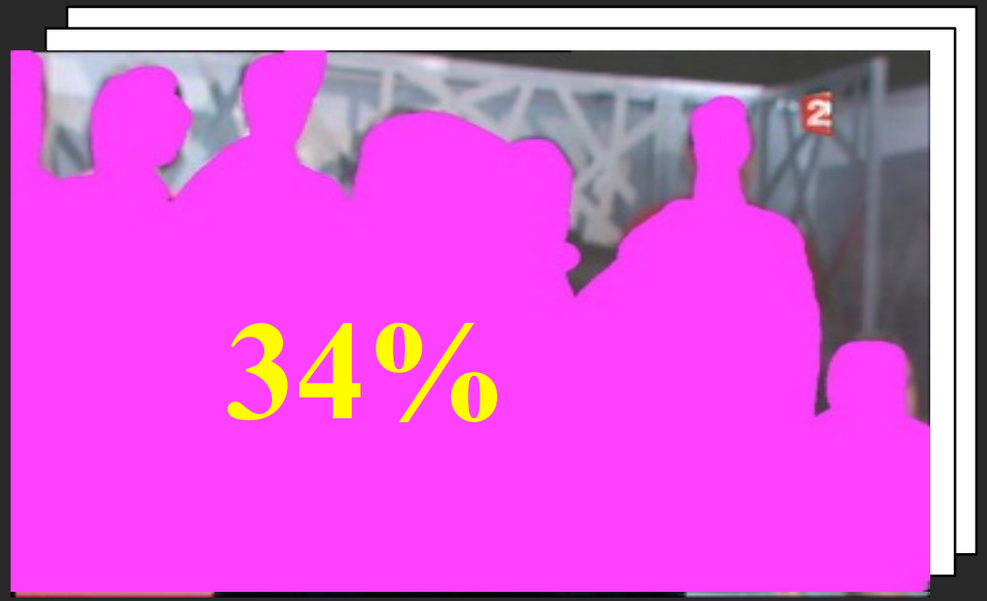
	Method	UCF101	HMDB51
Hand-crafted	[5] IDT+FV	85.9	57.2
	[29] IDT+HSV	87.9	61.1
	[30] IDT+MIFS	89.1	65.1
	[31] IDT+SFV	-	66.8
CNN (RGB)	[12] Slow fusion (from scratch)	41.3	-
	[13] C3D (from scratch)	44[1]	-
	[12] Slow fusion	65.4	-
	[6] Spatial stream	73.0	40.5
	[13] C3D (1 net)	82.3	-
	LTC$_{RGB}$	**82.4**	-
	[13] C3D (3 nets)	85.2	-
CNN (Flow)	[6] Temporal str.	83.7[2]	54.6[2]
	LTC$_{Flow}$	**85.2**	**59.0**
Fusion	[6] Two-stream (avg. fusion)	86.9	58.0
	[6] Two-stream (SVM fusion)	88.0	59.4
	[32] Convolutional pooling	88.2	-
	[32] LSTM	88.6	-
	[22] TDD	90.3	63.2
	[13] C3D+IDT	90.4	-
	[22] TDD+IDT	91.5	65.9
	[33] Transformations	92.4	62.0
	LTC$_{Flow+RGB}$	**91.7**	**64.8**
	LTC$_{Flow+RGB}$**+IDT**	**92.7**	**67.2**

Why focus on people?

How many person-pixels are in a video?



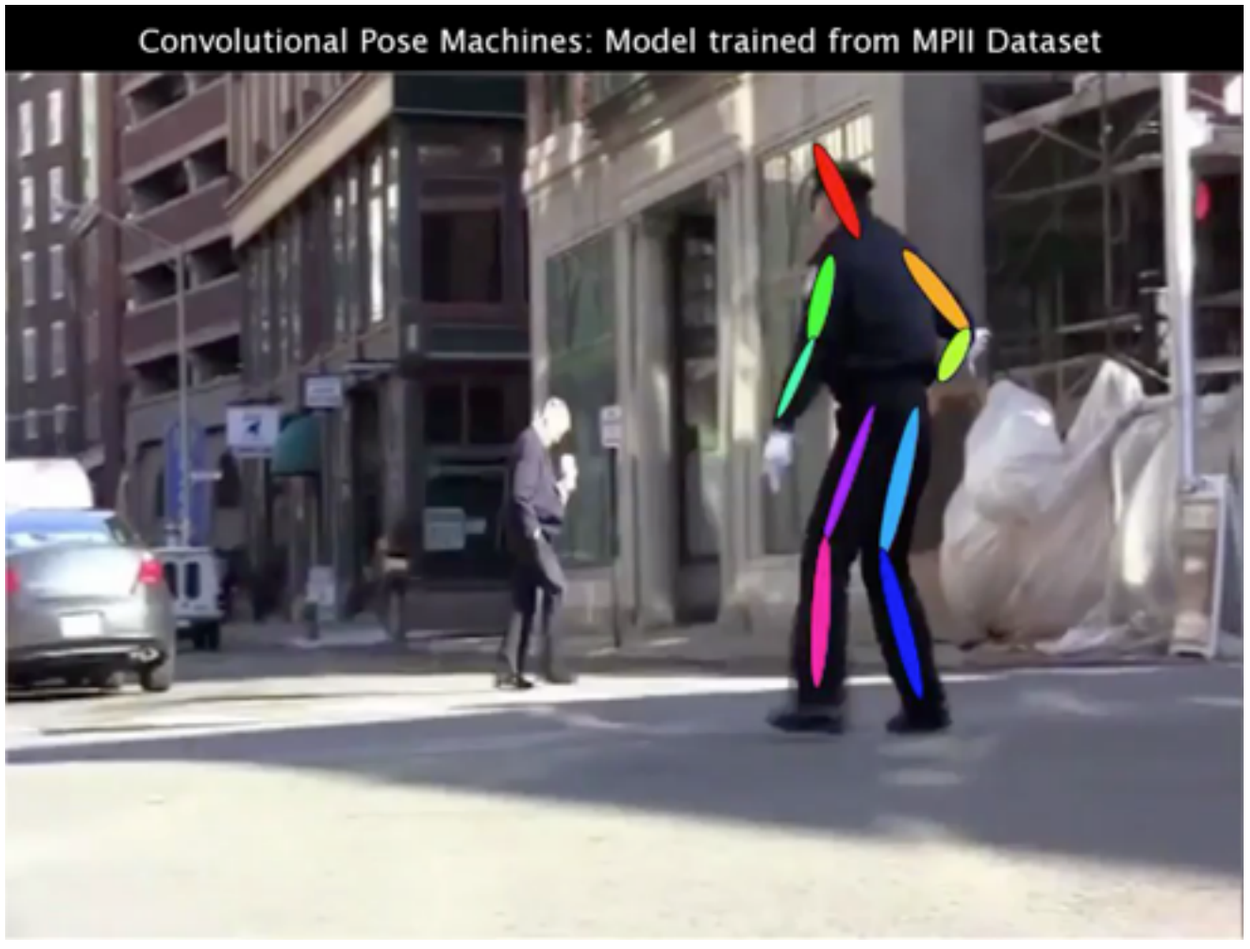
Movies



TV

YouTube

Visual front-end seems to work!



"Convolutional Pose Machines" CMU

Battle-plan for solving vision

Representations: How do we represent visual phenomena (objects/scenes/actions)?

Inference: How to compute with a given representation?

Learning: How do we learn representations from data?

Applications: How we ensure that we build useful pieces along the way?

Battle-plan for solving vision

Representations: How do we represent visual phenomena (objects/scenes/actions)?

Inference: How to compute with a given representation?

Learning: How do we learn representations from data?

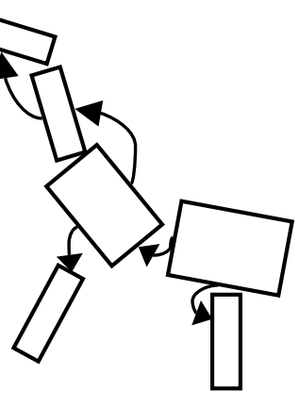
Applications: How we ensure that we build useful pieces along the way?

- Videos can help use address these tasks
- Deep learning appears to blur line between representations & inference

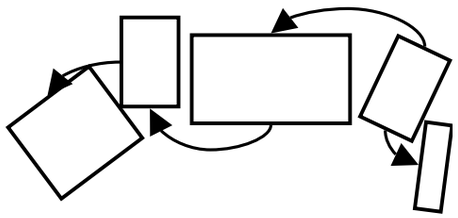
Learning structured models of objects from videos

Model objects as collection of "part," where parts are groups of pixels that move coherently

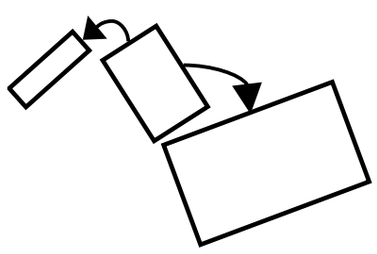
Find animals in other images



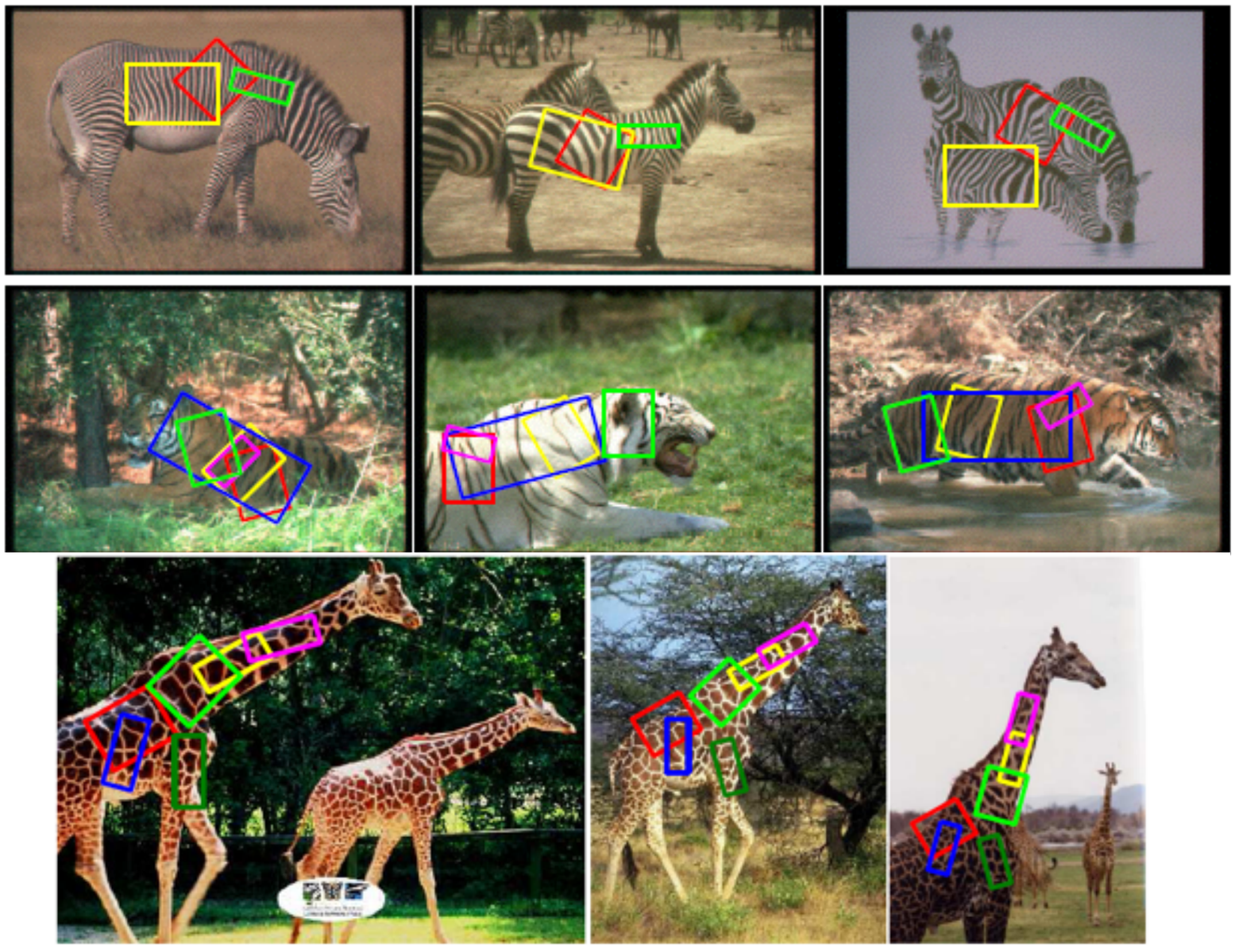
giraffe structure



tiger structure

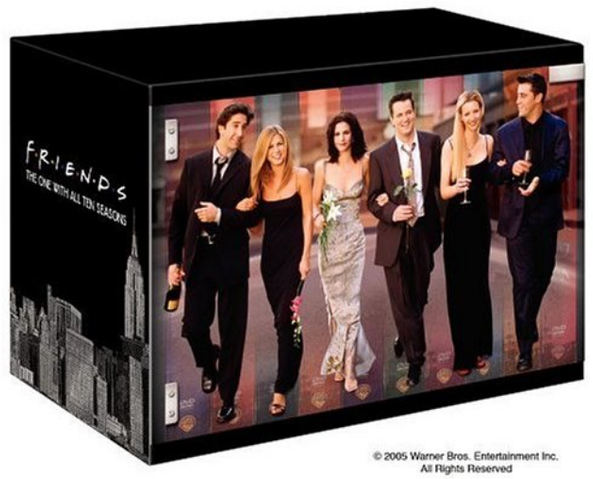


zebra structure



Can learn models for object detection from video

Semi-supervised learning from video



© 2005 Warner Bros. Entertainment Inc.
All Rights Reserved

Rich data source:
Spans 10s of years
Variation due to pose, age, weight gain, hairstyles,...

More realistic datasets to learn/evaluate face models

Learn about 3D facial structure

Global “scene” analysis

Requires understanding at both high-level (actions/itentions/goals) and low-level (pose tracking, optical flow)

Applications: assistive technology

Kinect camera hacked to help blind users nagivate

Applications: personal video logging

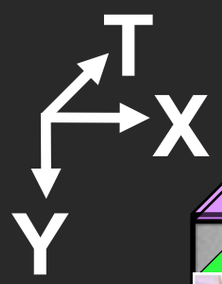
Outline

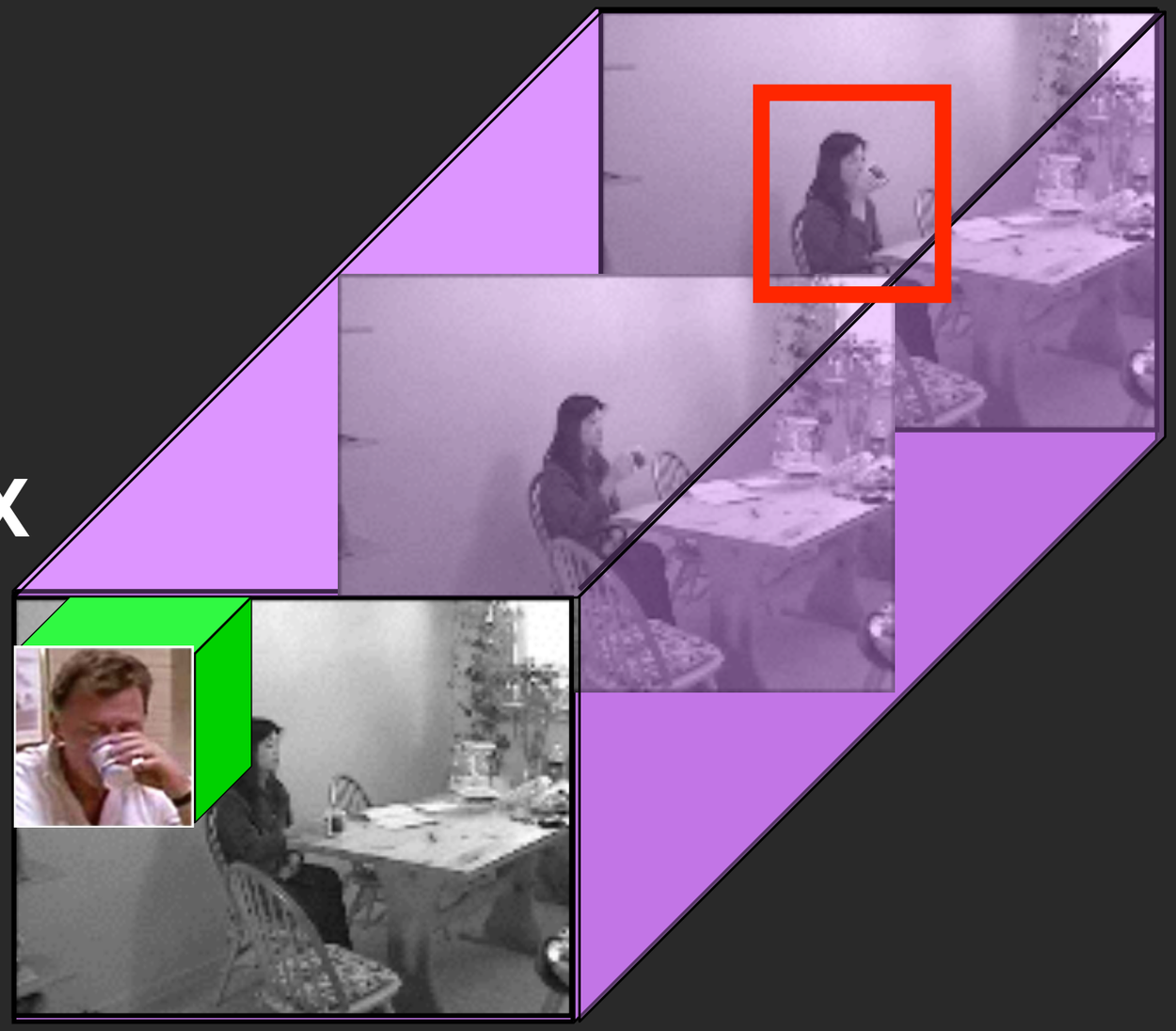
- Motivation
- Why is it hard?
- Open questions
- Logistics

One approach

Generalize ‘object detection’ techniques to spacetime volumes

Spacetime (XYT) template



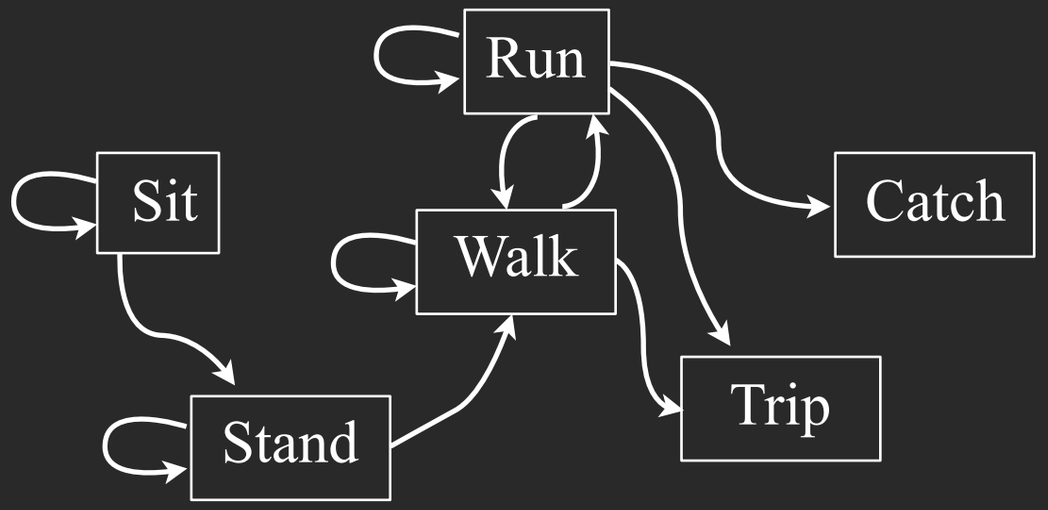


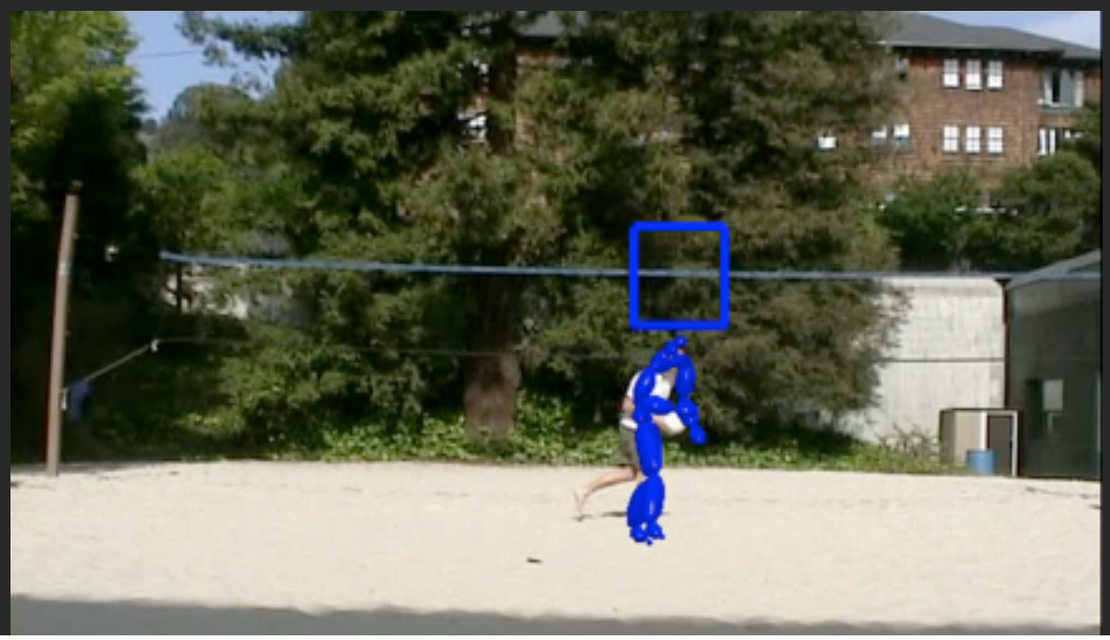
Grab-Cup Event

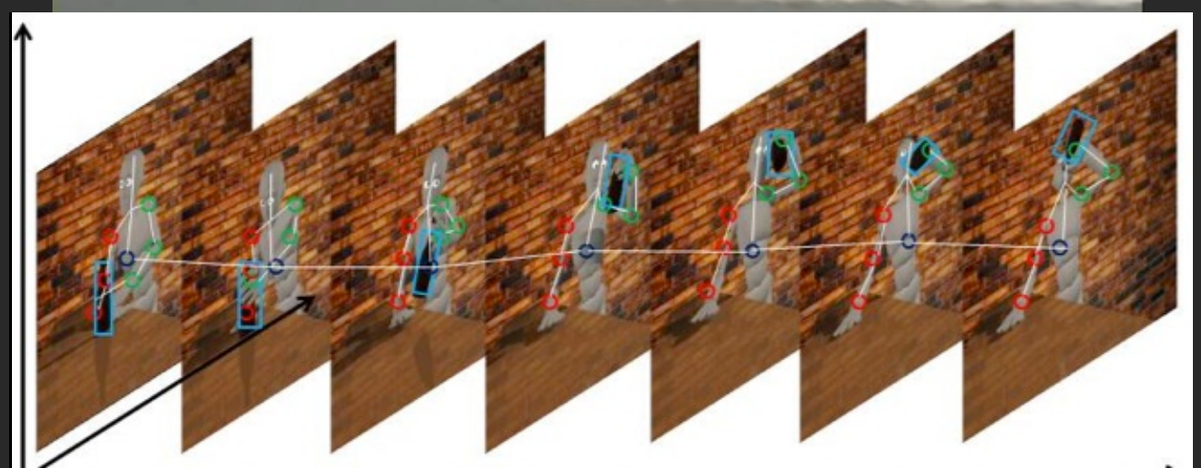
Spacetime correlation

Shechtman & Irani, CVPR05

Doesn’t seem likely to scale for higher-level activities

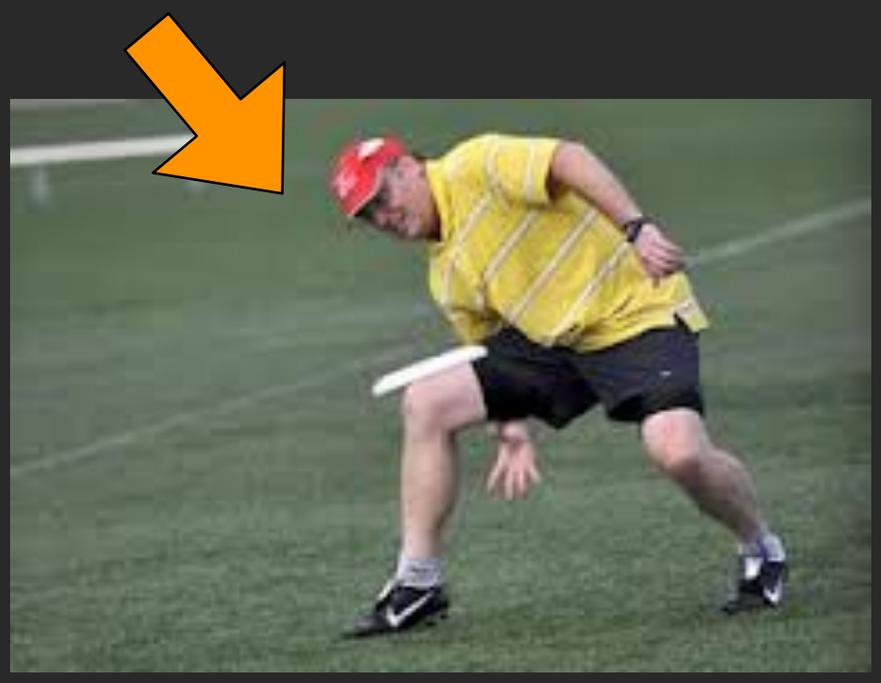






But what’s the desired output here?

Long-tail distributions



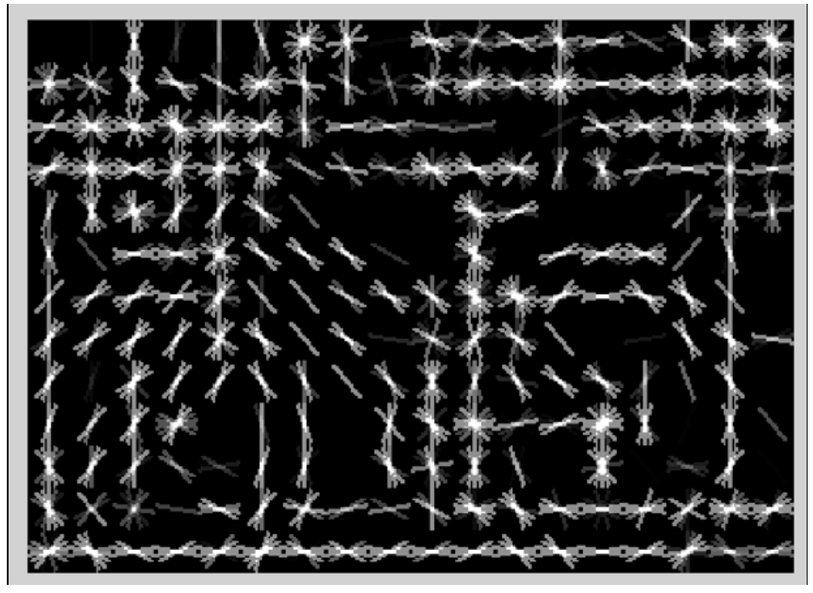
Challenge: actions seem to follow an extremely heavy tail distribution

Complicates dataset collection and annotation

Open questions

Can we do better than hand-engineered temporal features?

This would clearly have significant impact



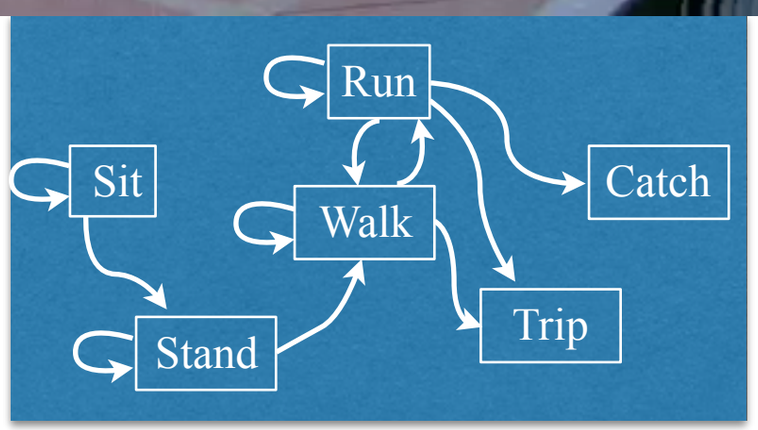
Open questions

Recognizing objects in videos: can we do better than processing each image independently?

Open questions

What are representations that capture temporal relations?

Markov models, grammars, temporal interval logics, etc.

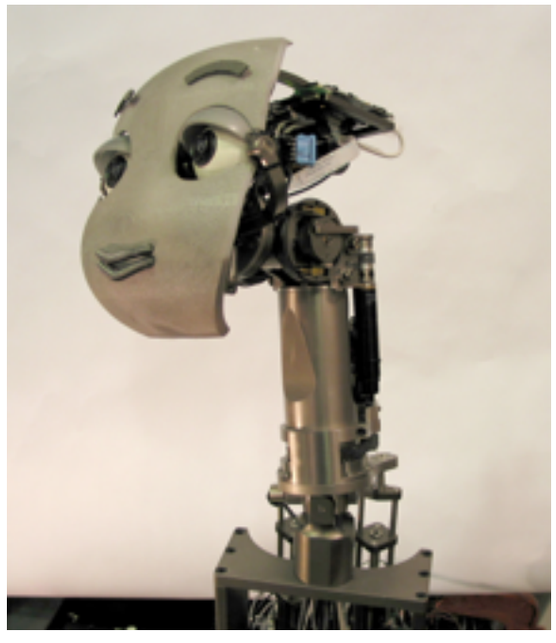


Goal: Produce output such as

"woman may have left item in building"

Seeing as an action

Perceptual inference as a decision-making process



Integrate deep models with attentional cascades (or policies)?

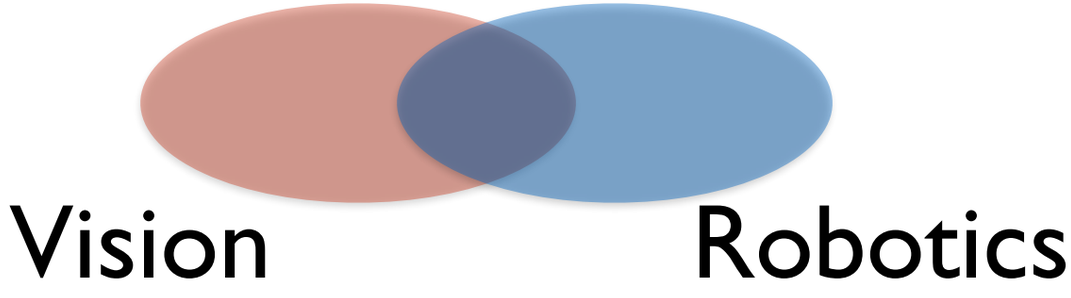
Functional view of recognition

J. J. Gibson

The Ecological Approach to Visual Perception

"If you know what can be done with a graspable detached object, what it can be used for, you can call it whatever you please"

Active vision



Real-time active perception is finally within reach!

Structure of papers

Low-level (xyt features)

Mid-level (tracking / pose / detection)

High-level (actions / activities)

Topics we'll address along the way:

Latest & greatest in the deep world
Reinforcement learning
Prediction
Datasets

Project

- Would like it to be more than just research "as usual"
- Take the opportunity to explore something new
- I will assume you have resources (hardware, machines, etc.) but let me know if not

Homework

- Start thinking about papers and projects
  (initial list will be up by end of day)

- I've already reached out to folks for some suggestions (thanks! - happy to include more)

http://16899.courses.cs.cmu.edu/lec.html

Presentations/ audience participation

- I think communication is a hugely undervalued skill for researchers

Tips for Giving Clear Talks

Kayvon Fatahalian
Sept 2015

https://www.cs.cmu.edu/~kayvonf/misc/cleartalktips.pdf

I'll also require folks to begin presentation days with a one-sentence description of what they liked about the presented paper (suggestion from LP)