Action recognition

Deva Ramanan

UC Irvine

(Many slides from Ivan Laptev, David Forsyth)

What is the problem we're trying to solve?

Surprisingly hard to be precise about....

Laptev et al

Laptev et al

Computer vision grand challenge: Dynamic scene understanding

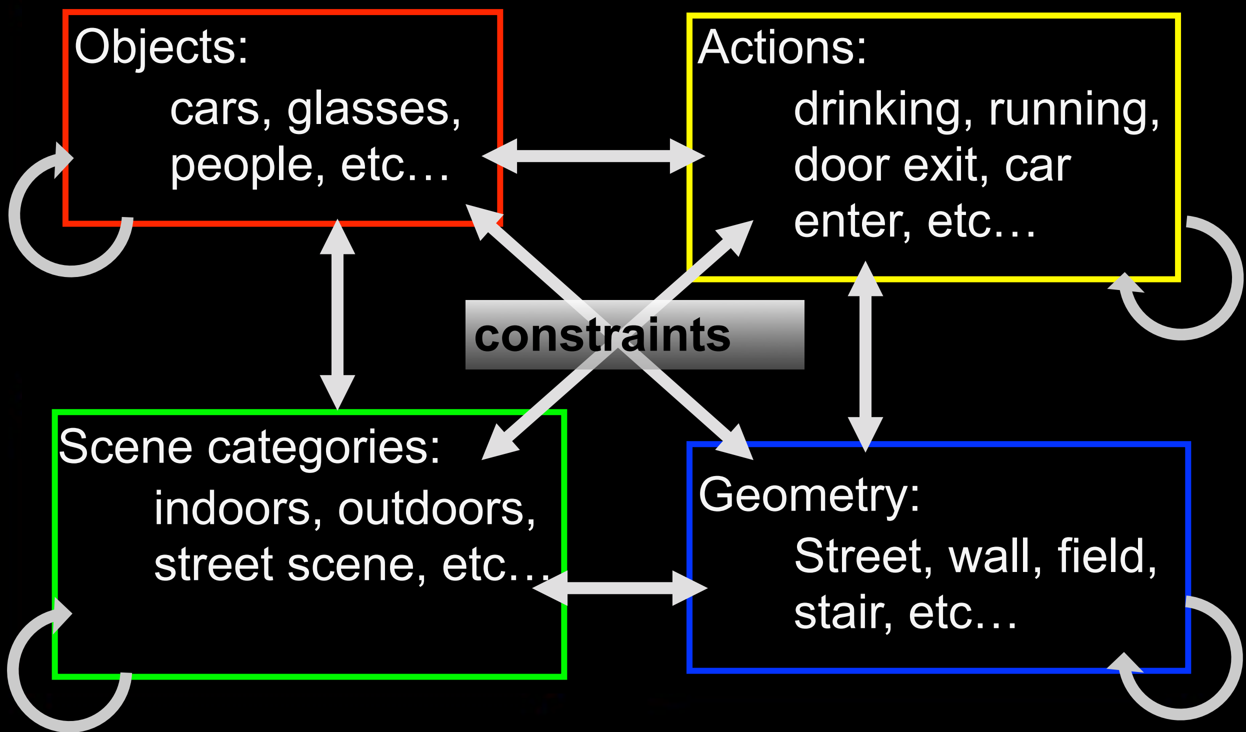
Objects:
cars, glasses, people, etc…

Actions:
drinking, running, door exit, car enter, etc…

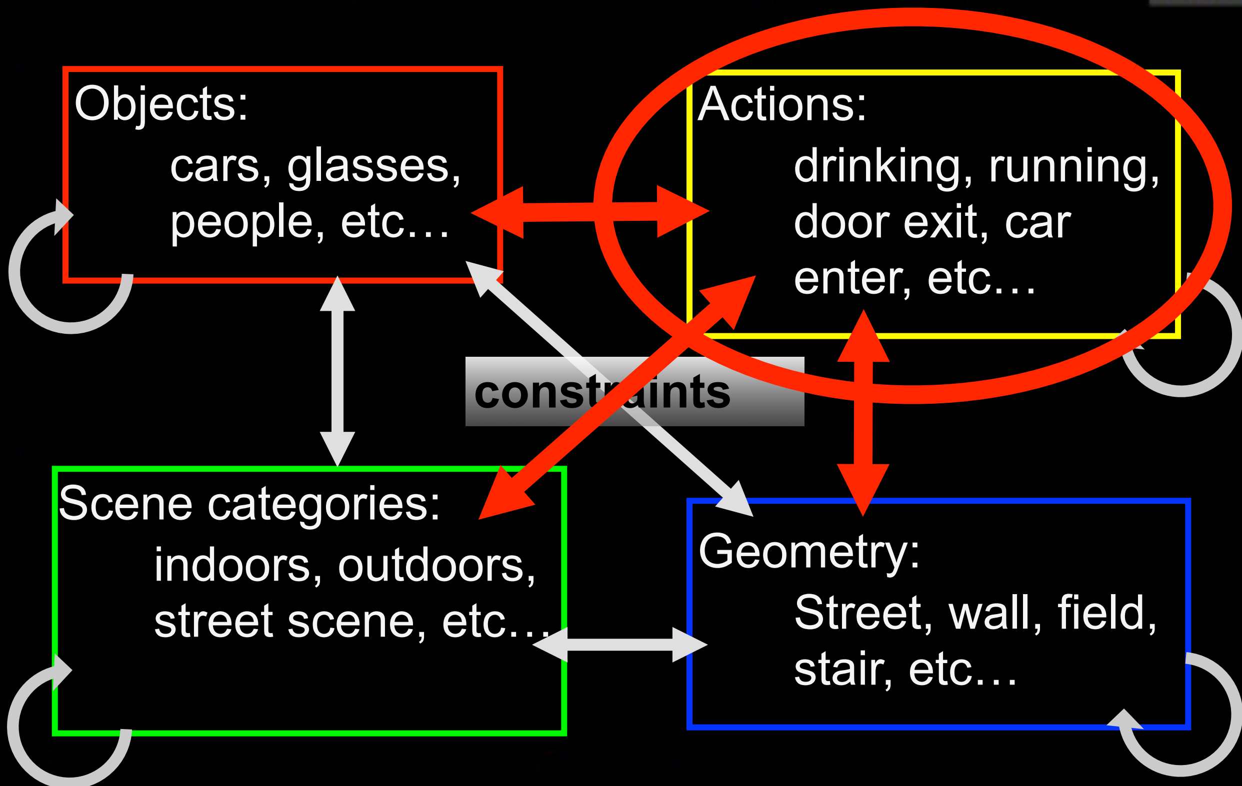
Scene categories:
indoors, outdoors, street scene, etc…

Geometry:
Street, wall, field, stair, etc…

Computer vision grand challenge: Dynamic scene understanding



Computer vision grand challenge: Dynamic scene understanding



Why focus on people/actions?

Lots (most?) of training data comes in video form

Data:

TV-channels recorded since 60's

>34K hours of video uploads every day

~30M surveillance cameras in US
=> ~700K video hours/day

Lots (most?) of test data comes in video form

Wearable: where did I leave my keys?

Education: How do I make a pizza?

Predicting crowd behavior
Counting people

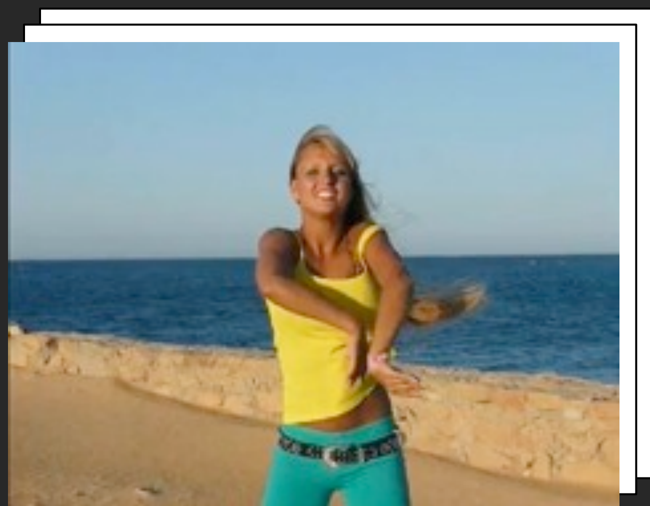
Motion capture and animation

Why focus on people/actions?

How many person-pixels are in the video?

Movies

TV



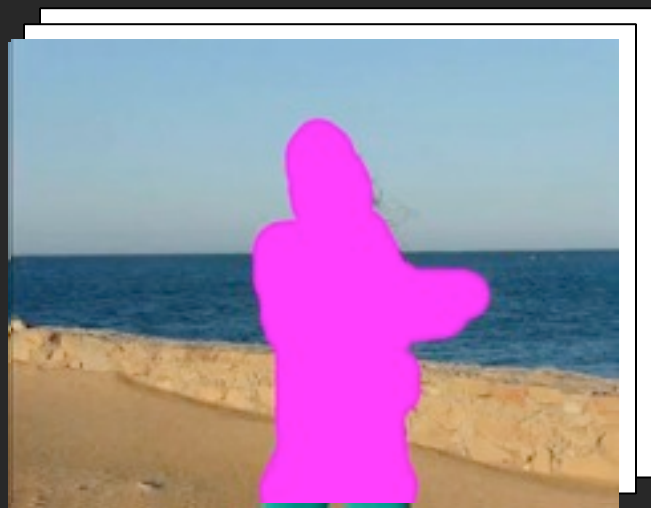
YouTube

Why focus on people/actions?

How many person-pixels are in the video?

Movies

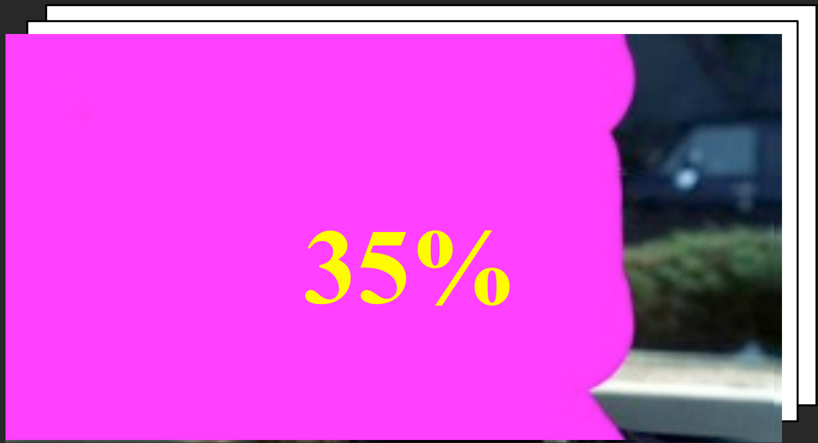
TV



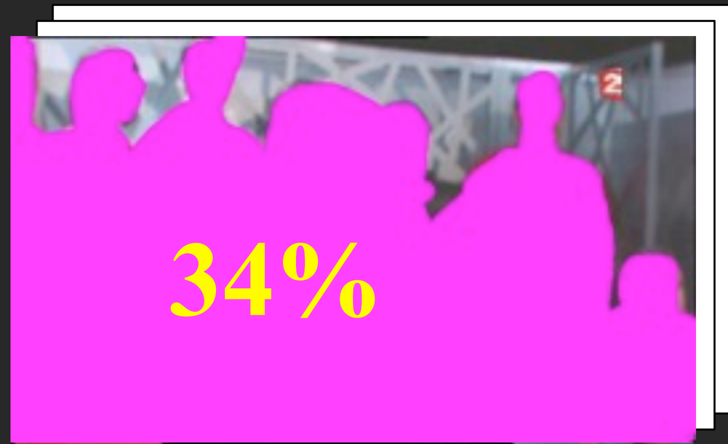
YouTube

Why focus on people/actions?

How many person-pixels are in the video?



Movies



TV

YouTube

One approach

Generalize ‘object detection’ techniques to spacetime windows

Ke et al, ICCV05

One approach

Generalize 'object detection' techniques to spacetime windows

Ke et al, ICCV05

One approach

Generalize 'object detection' techniques to spacetime windows

Spacetime (XYT) template

Grab-Cup Event

Ke et al, ICCV05

Spacetime correlation

Shechtman & Irani, CVPR05

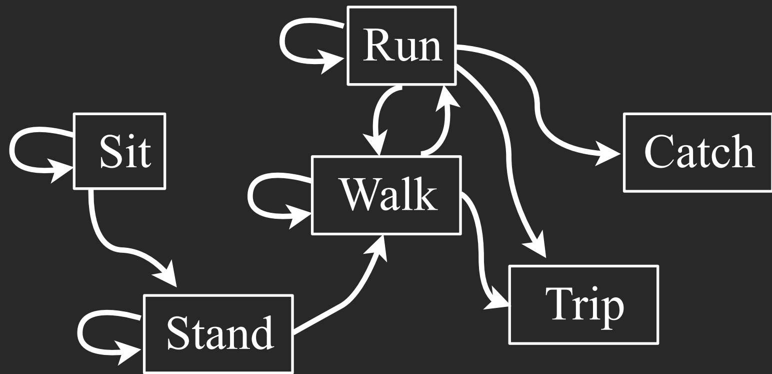
Spacetime correlation

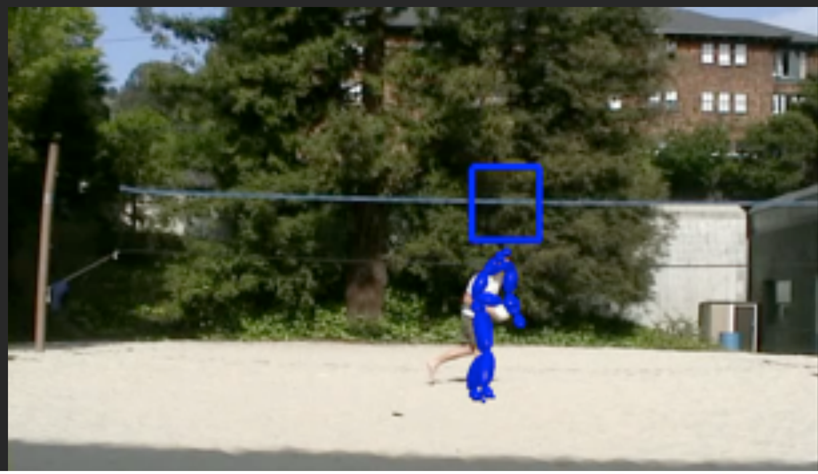
Shechtman & Irani, CVPR05

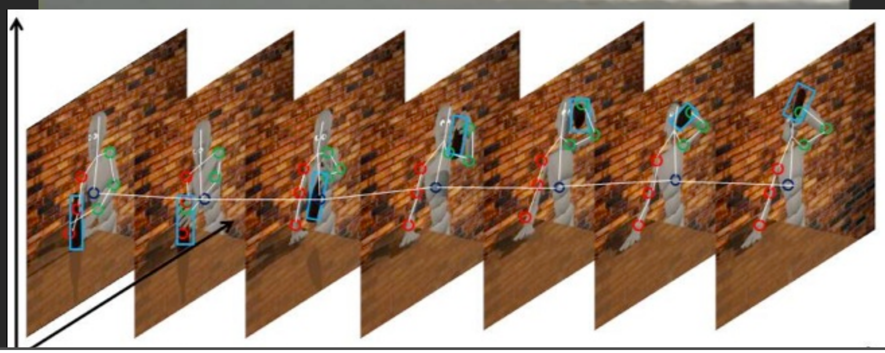
Spacetime correlation

Shechtman & Irani, CVPR05

Flexible spacetime part templates







But what's the desired output here?

But what’s the desired output here?

Will we have a “throw cat in the trash bin” template?

But what’s the desired output here?

Will we have a “throw cat in the trash bin” template?

Long-tail distributions

Challenge: actions seem to follow an extremely heavy tail distribution

Complicates dataset collection and annotation

Roadmap

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

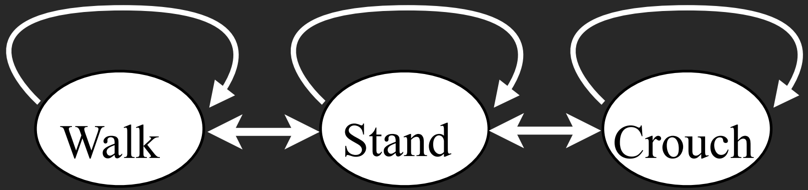


Image benchmarks

Russell, et al. CVPR 2008.

Deng, et al. CVPR 2009.

Everingham, et al. IJCV 2010.

Ahn and Dabbish. CHI 2004.

Torralba, et al. PAMI 2008.

Xiao, et al. CVPR 2010.

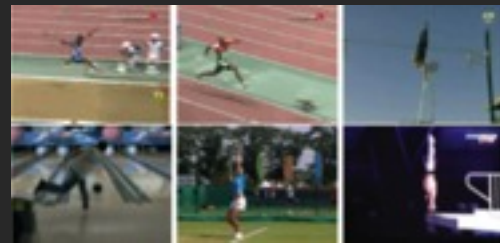
Like it or not, crucial for advances in the field

Large-scale annotated video datasets are more rare - why?

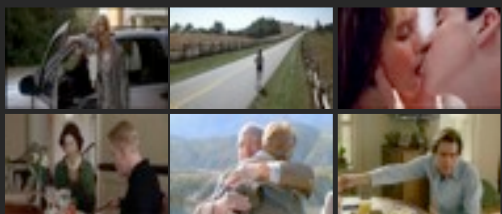
Action recognition benchmarks

UCF 101 Sports, 2013

KTH, ICPR’04



Olympics sport, BMVC’10



Hollywood, CVPR’09

UCF Youtube, CVPR’08

VIRAT, CVPR’11

1) Video is cumbersome to label (difficult to define natural categories outside sports)
2) Collecting interesting but natural video is surprisingly hard
3) Most current work focuses on K-way classification (similar to image recognition 10 years ago)

KTH

Classification performance around 100%
"Outdated"

TRECVID

“Woodworking” action

board-trick, feeding animal, fishing, wedding, woodworking, birthday, changing vehicle tire, flash mob, vehicle unstuck, grooming an animal, sandwich making, parade, parkour, repairing appliance and sewing,...

State-of-the-art is around 5-10% accuracy

Challenge 1: how we do know what to label?

Look for cues in language (how do people describe images/videos?)

...

RICK

Why weren't you honest with me? **Why did** you keep your marriage a secret?

Rick sits down with Lisa.

LISA

Oh, it wasn't my secret, Richard. Victor wanted it that way. Not even our closest friends knew about our marriage.

...

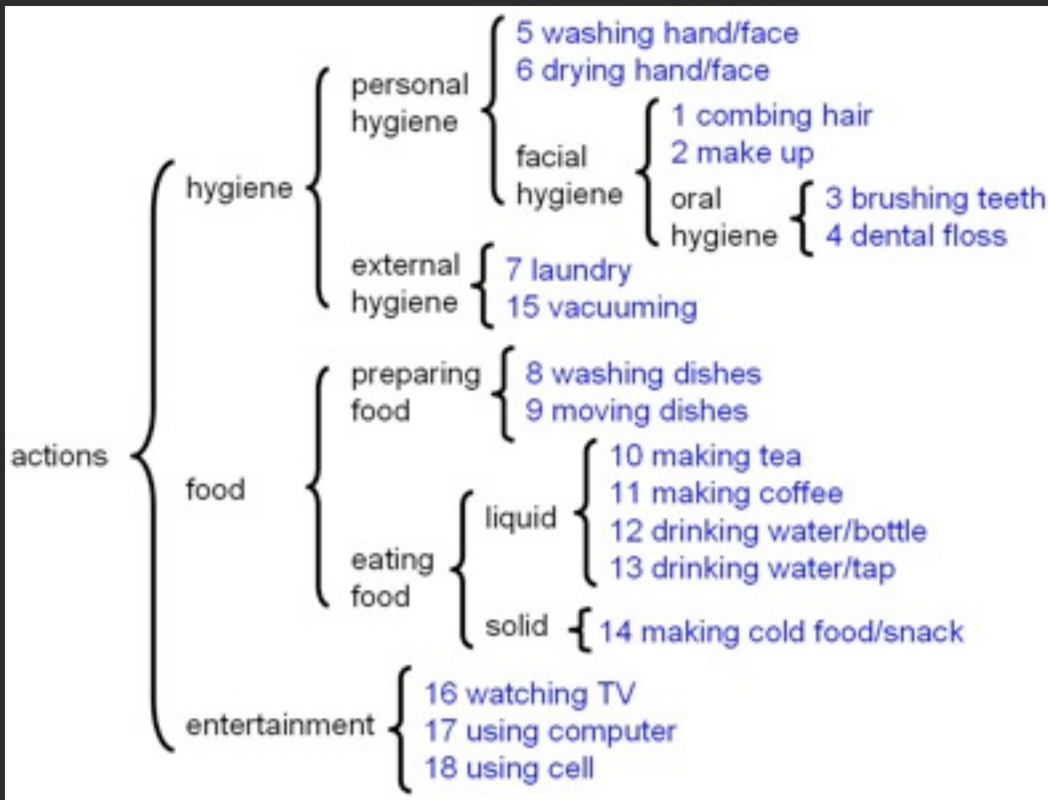
Mining movie scripts
Everingham et al. BMVC
Laptev et al 08.

This is a lot of technology.
Somebody's screensaver of a pumpkin.
Black laptop is connected to black Dell monitor.
Old school Computer monitor with many stickers on it.
A refrigerator full of food.

Ask people on turk for descriptions
Farhadi et al ECCV10

Challenge 1: how we do know what to label?

Look for cues in medical literature on “activities of daily living” (ADLs)



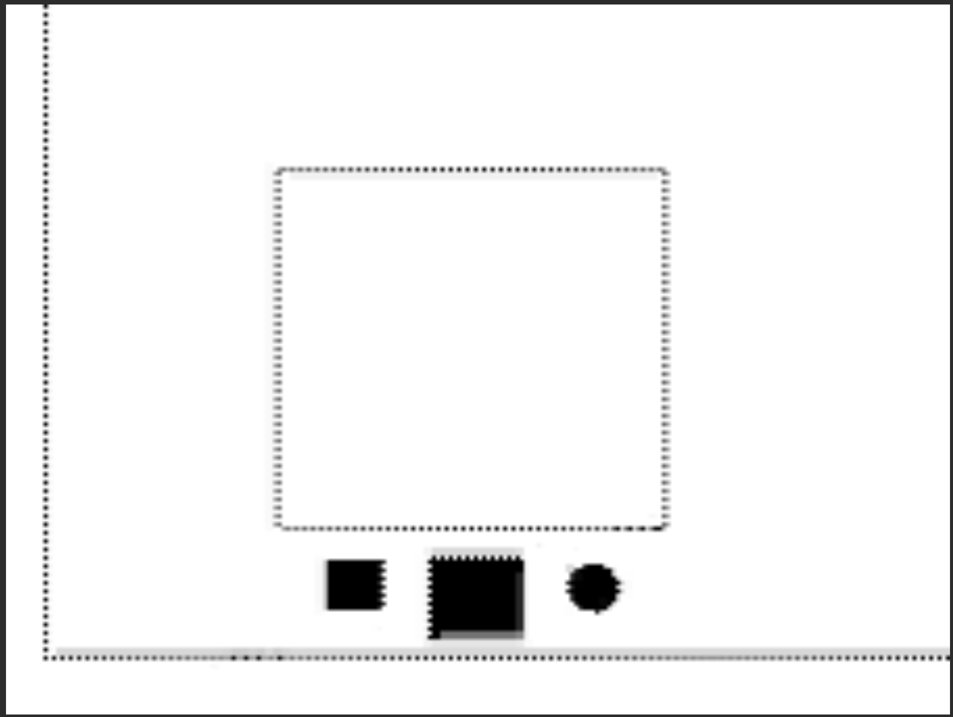
Pirsivash & Ramanan CVPR12

Table 1: AMAT Compound Tasks and Task Components

- I. Cut “Meat”
  1. Pick up knife and fork*
  2. Cut “meat” (Play-Doh)*†
  3. Fork to mouth
- II. Foam “Sandwich”
  4. Pick up foam “sandwich”
  5. “Sandwich” to mouth
- III. Eat With Spoon
  6. Pick up spoon
  7. Pick up dried kidney bean with spoon
  8. Spoon to mouth
- IV. Drink From Mug
  9. Grasp mug handle
  10. Mug to mouth
- V. Comb Hair
  11. Pick up comb
  12. Comb hair†
- VI. Open Jar
  13. Grasp jar top*
  14. Screw jar top open*
- VII. Tie Shoelace
  15. Tie shoelace*†
- VIII. Use Telephone
  16. Phone received to ear
  17. Press phone number

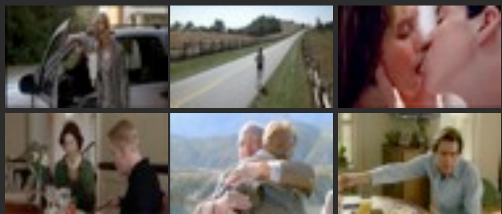
Challenge 1: how we do know what to label?

Actions vs goal-directed behaviors

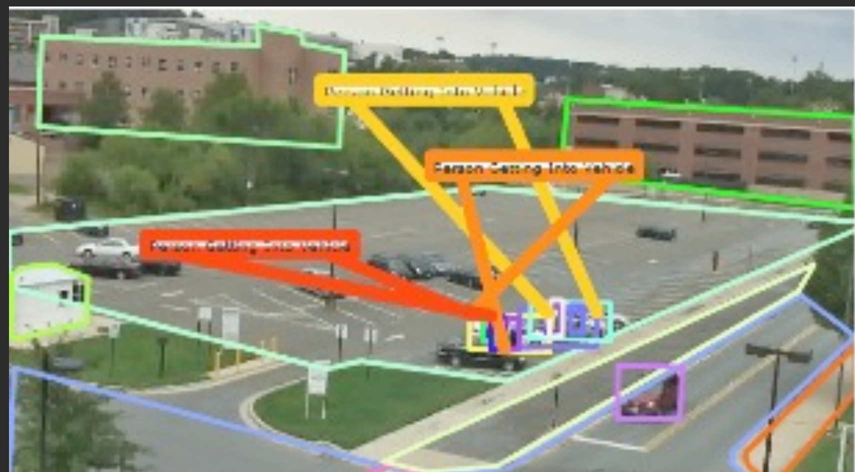


Chase vs follow

Challenge 2: how we do obtain interesting data?



Script it, using actors



Use real but “boring” data

Challenge 2: how we do obtain interesting data?

Egocentric/wearable cameras

Pirsivash & Ramanan CVPR12

"Functional" ADLs

Easy to capture variety-rich data

Challenge 2: how we do obtain interesting data?

Egocentric/wearable cameras

Pirsivash & Ramanan CVPR12

"Functional" ADLs
Easy to capture variety-rich data

Challenge 2: how we do obtain interesting data?

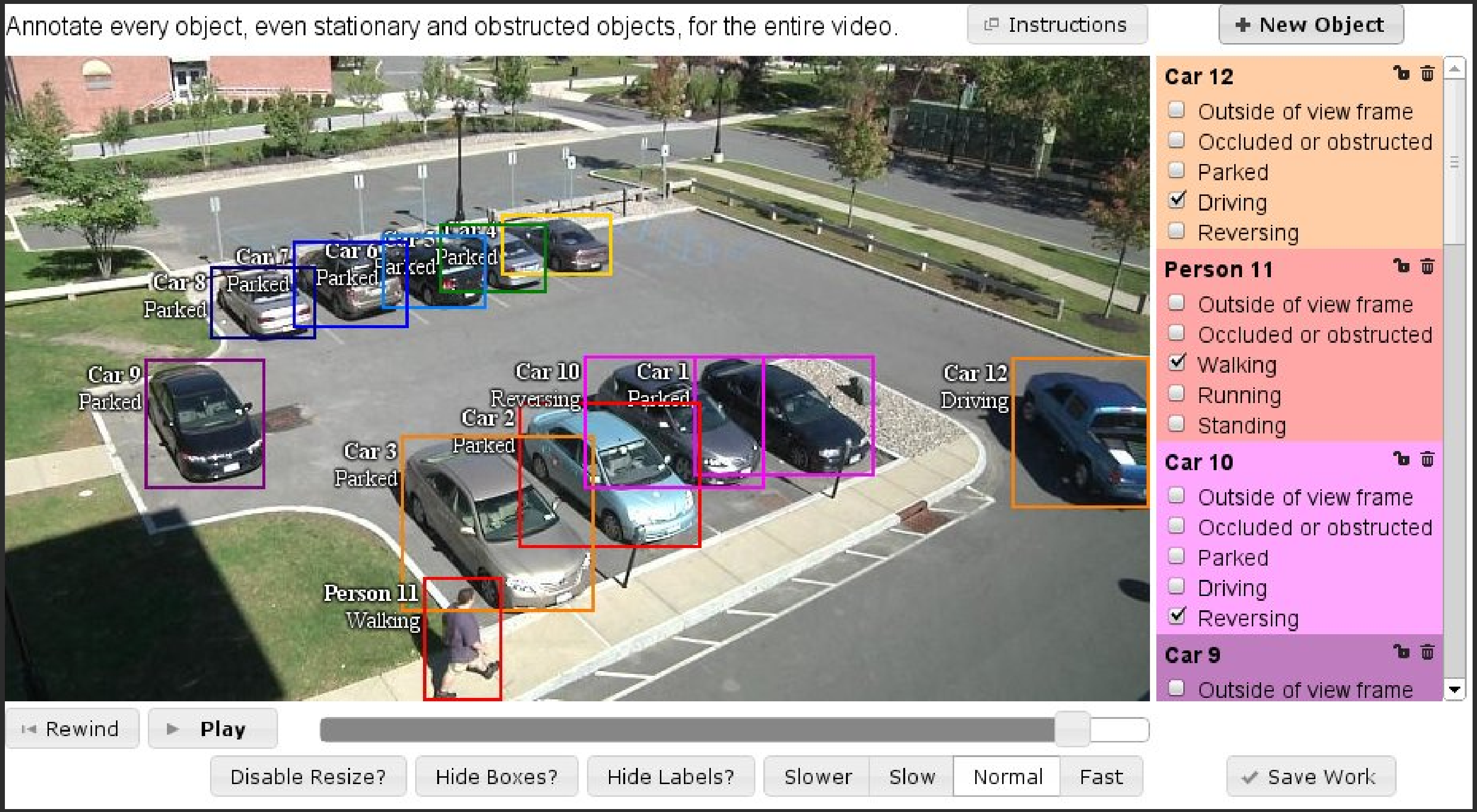
Egocentric/wearable cameras

Pirsivash & Ramanan CVPR12

“Functional” ADLs
Easy to capture variety-rich data

Challenge 3: how we do produce detailed annotations?

Crowdsource labeling



Vondrick et al "VATIC" ECCV10, NIPS11, IJCV 13

Lessons: Interface design matters

Use experts, not the crowd

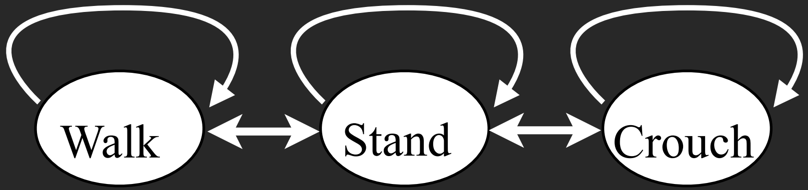
Active annotation helps

Roadmap

Data/benchmark analysis

Spatiotemporal features

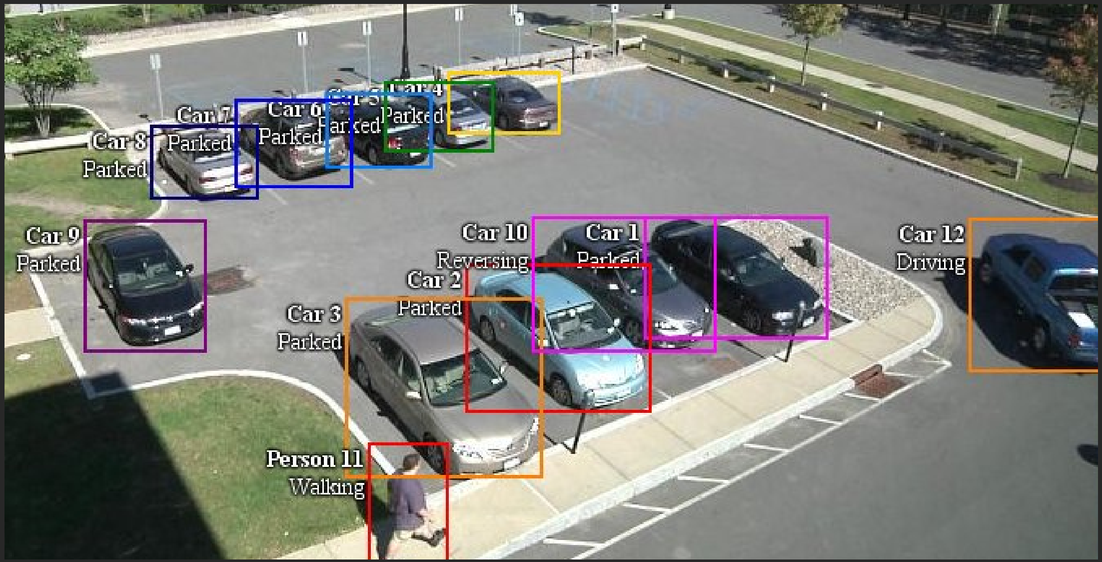
Spatiotemporal models



Spacetime features

Simple approach: just use spatial features

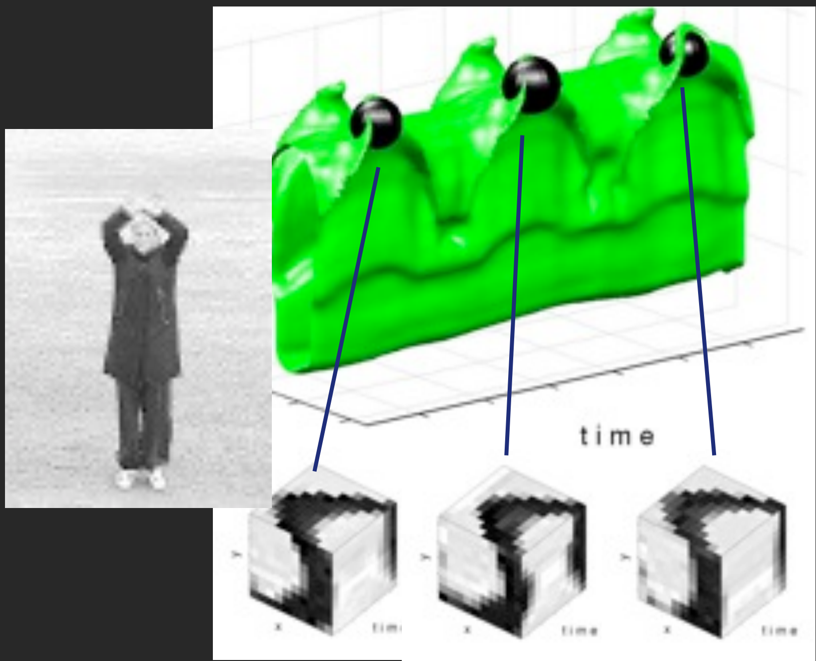
Surprisingly (and annoyingly) effective

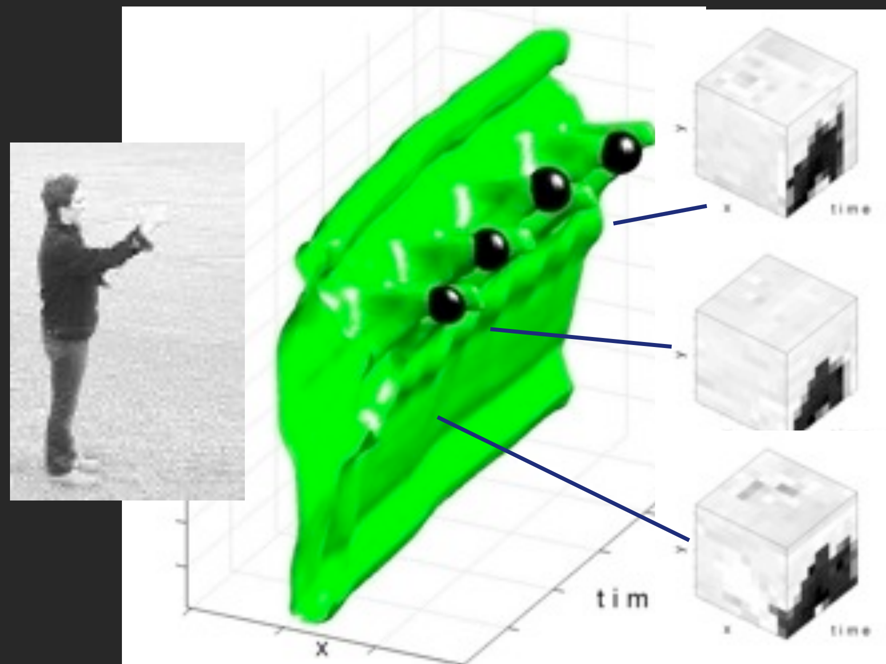


Build a bank of static-image detectors (of poses, objects, ...)

Exploiting motion

Spatiotemporal interest points (STIPS)





[Laptev 2005]

XYT descriptor evaluation

Detection (AP)

Descriptors

	Harris3D	Cuboids	Hessian	Dense
HOG3D	43.7%	45.7%	41.3%	45.3%
HOG/HOF	45.2%	46.2%	46.0%	**47.4%**
HOG	32.8%	39.4%	36.2%	39.4%
HOF	43.3%	42.9%	43.0%	45.5%
Cuboids	-	45.0%	-	-
E-SURF	-	-	38.2%	-

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]

Capturing the “right” temporal motion

Image motion confounds camera translation, object translation, and nonrigid deformations

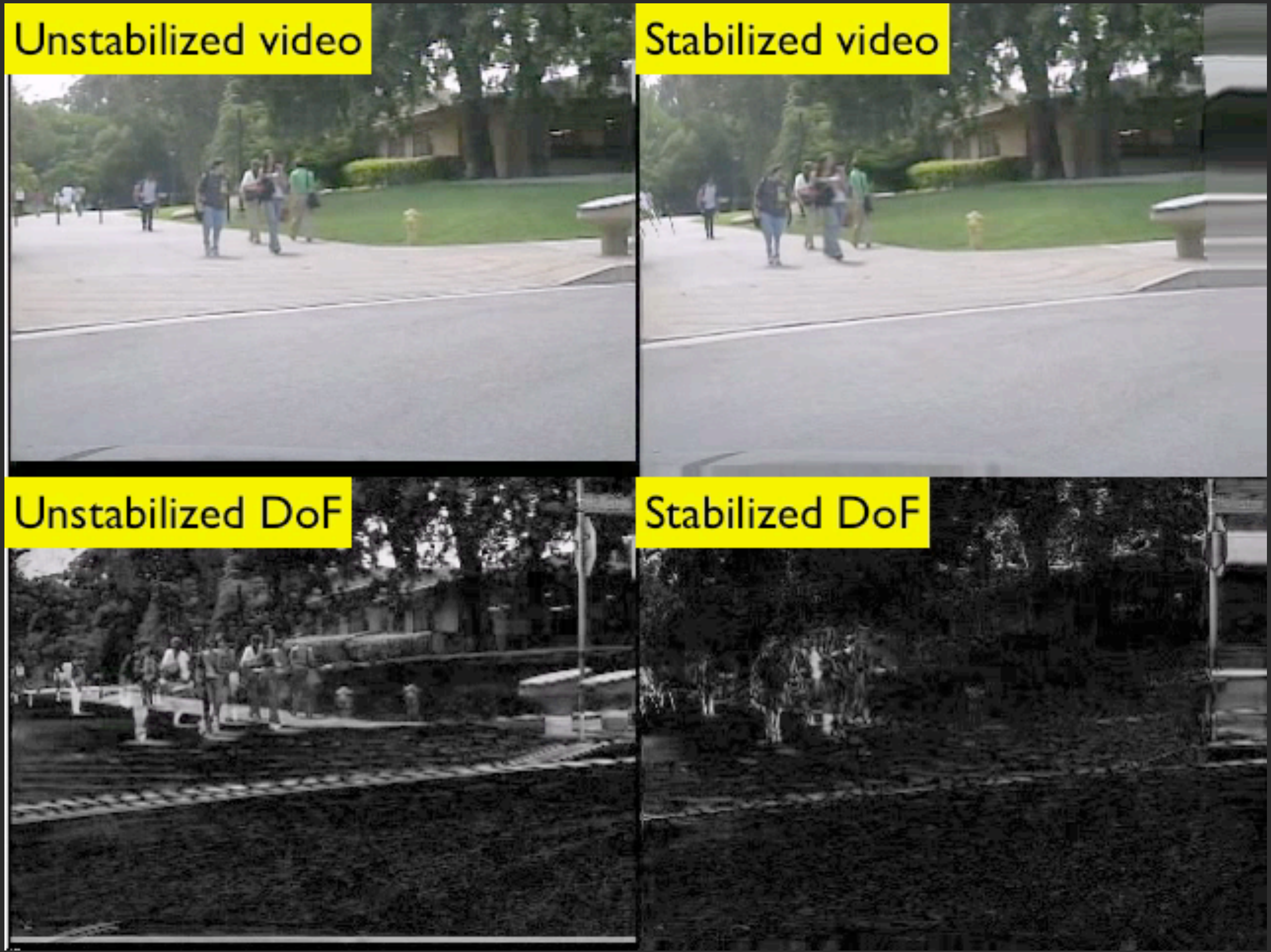
Capturing the “right” temporal motion

Image motion confounds camera translation, object translation, and nonrigid deformations

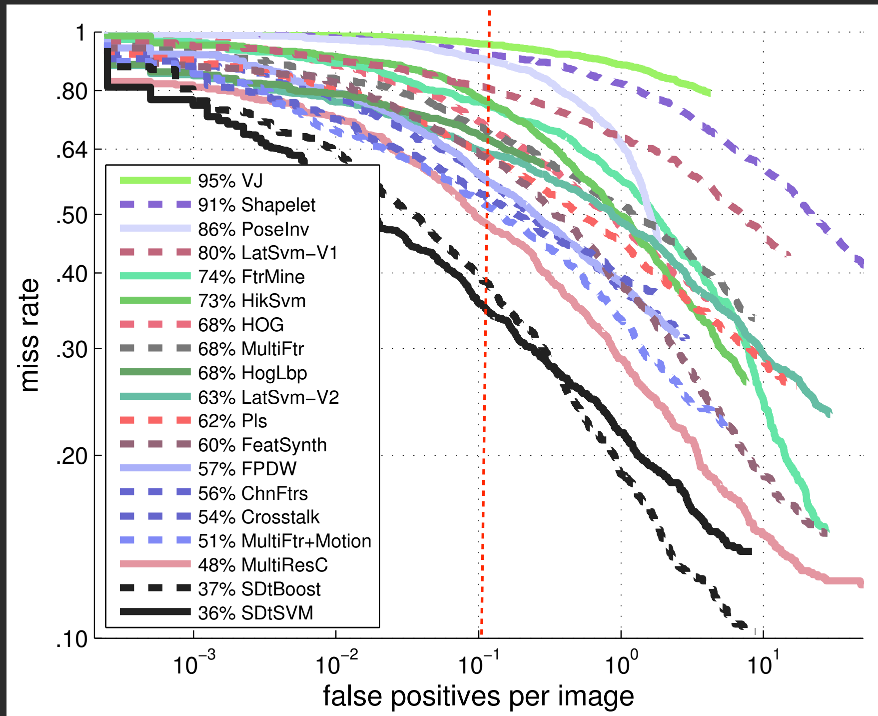
Stabilized camera

Stabilized object

Stabilized camera + object


Unstabilized video
Stabilized video
Unstabilized DoF
Stabilized DoF

Motion features for detection in videos



Caltech Pedestrian Benchmark; reduce miss rate from 48% to 36%

Park et al CVPR13

Roadmap

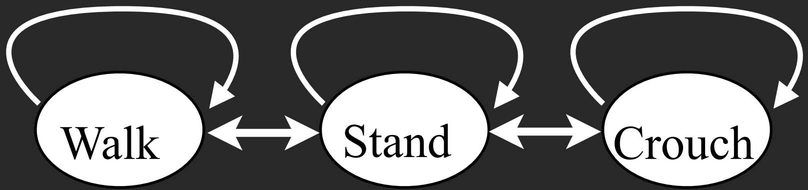
Data/benchmark analysis

Spatiotemporal features

Motion Features

Tracking

Spatiotemporal models



Why do we need to track?

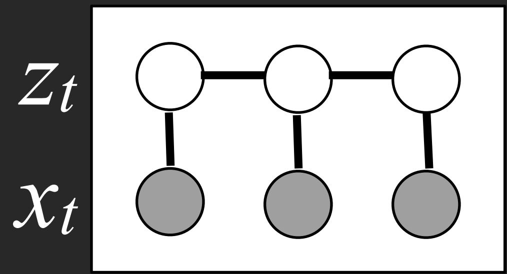
Spacetime window maybe “shearing”

Why do we need to track?

Spacetime window maybe “shearing”

Tracking

Immense literature



S(x,z) = local templates + spatial relations + temporal relations

Historically, last term has been focus of tracking community

Given z_t, predict z_{t+1} with $P(z_{t+1}|z_t)$

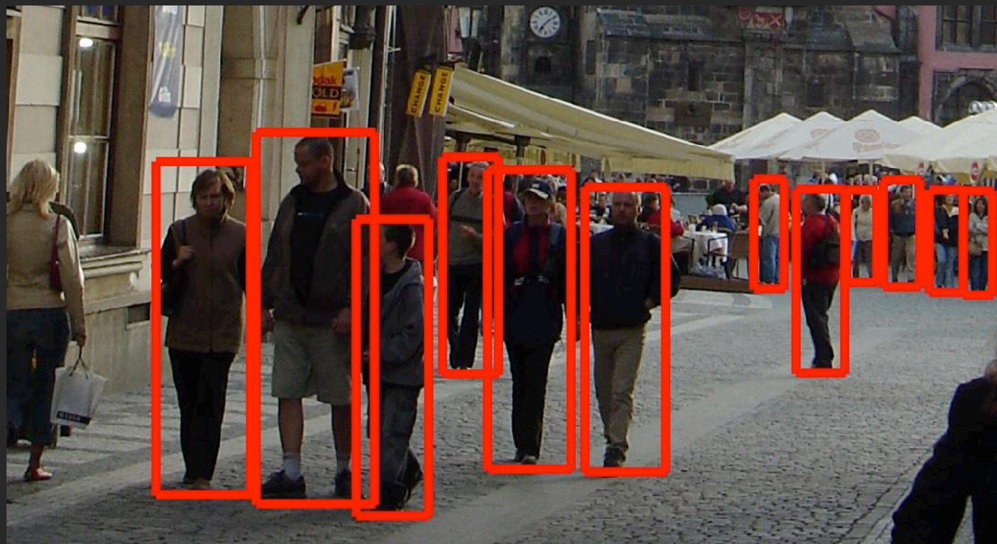
e.g., particle filtering, Isard & Blake

Extreme form of problem: multi-object tracking

Estimate number of tracks and their extent

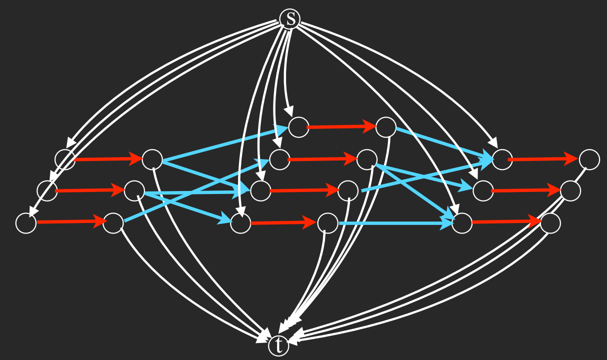
Do not assume manual initialization
Estimate birth and death of each track

Tracking by detection



Detect candidates
Link detections over time into tracks

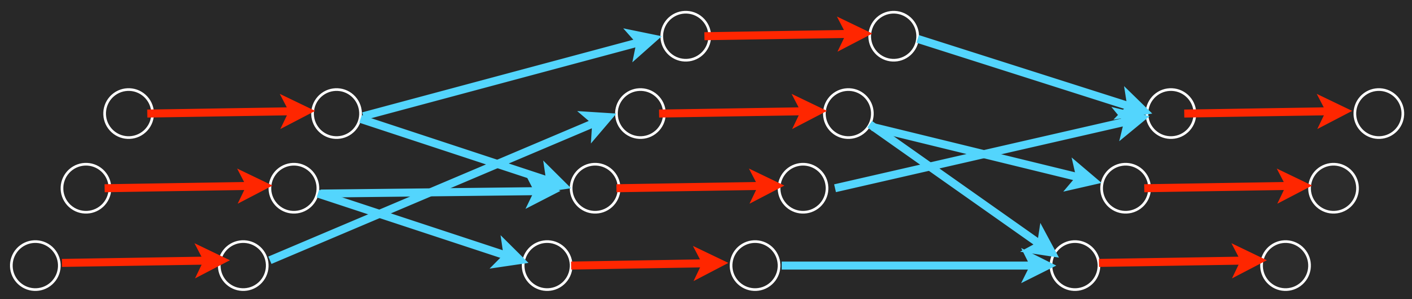
Multi-object tracking as integer/linear programming



View as combinatorial problem of what detections to turn on/off

Jiang et al CVPR07
Zhang et al CVPR08
Berclaz et al PAMI2011
Andriyenko and Schindler ECCV10
Pirsiavash, Ramanan, Fowlkes CVPR11
Butt and Collins CVPR13

Trellis Graph

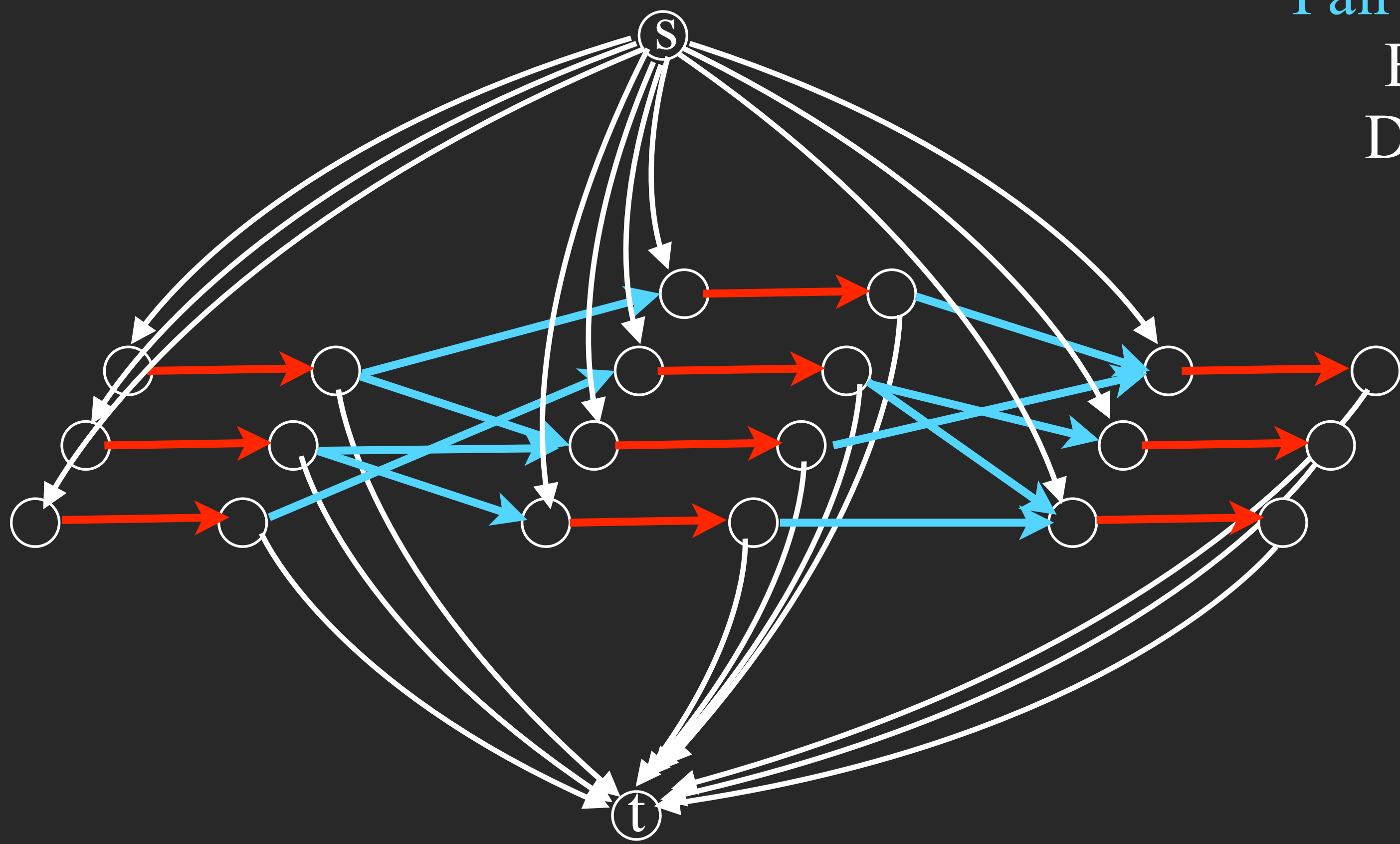


Local cost of window
Pairwise cost of transition

Use dynamic programming (DP) to find single track
(e.g., Vitterbi algorithm)

Trellis Graph

Local cost
Pairwise cost
Birth cost
Death cost

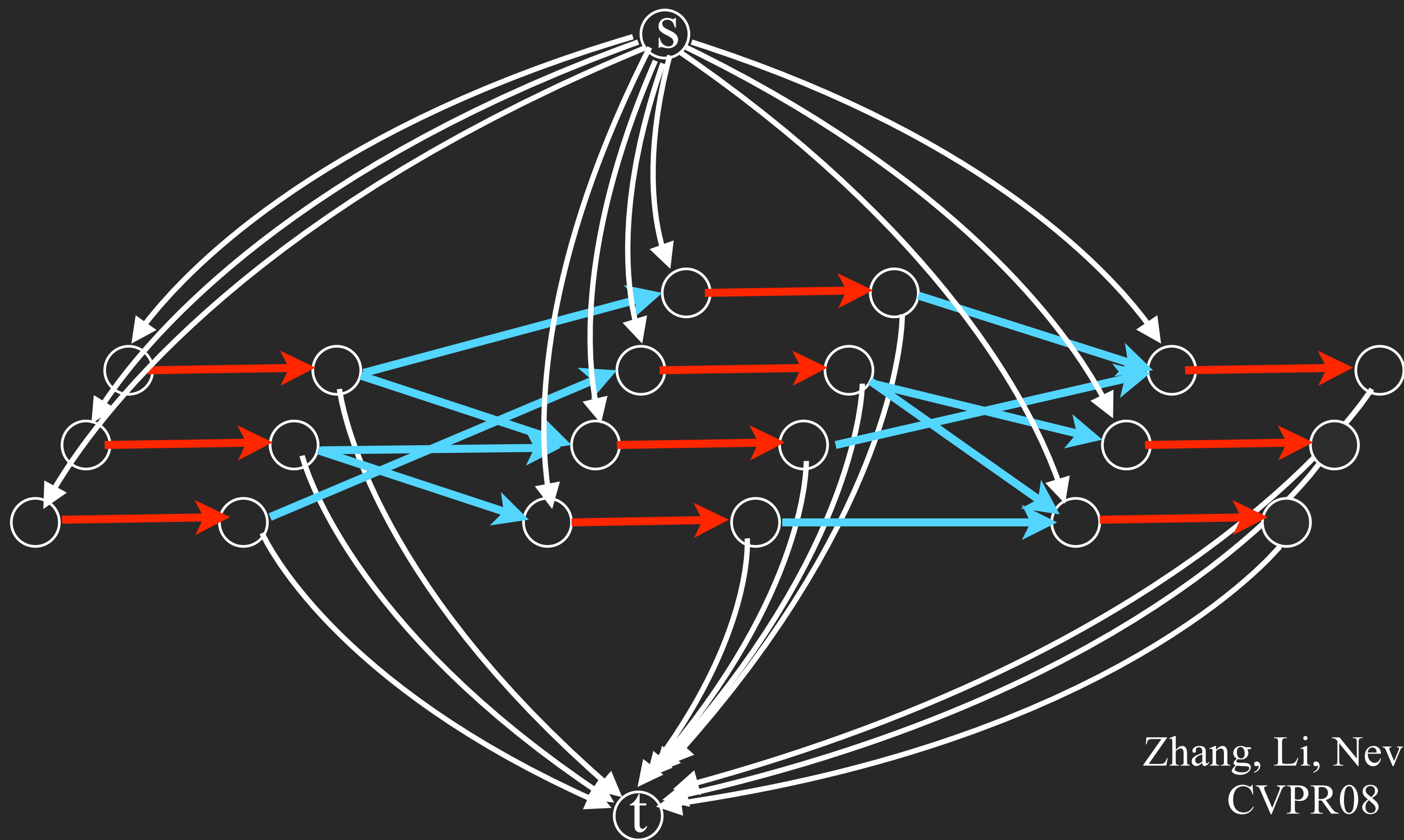


Shortest path from S to T = best variable-length track

Still can use DP

Min-cost flow problem

(generalization of min-cut / max-flow)



Zhang, Li, Nevatia
CVPR08

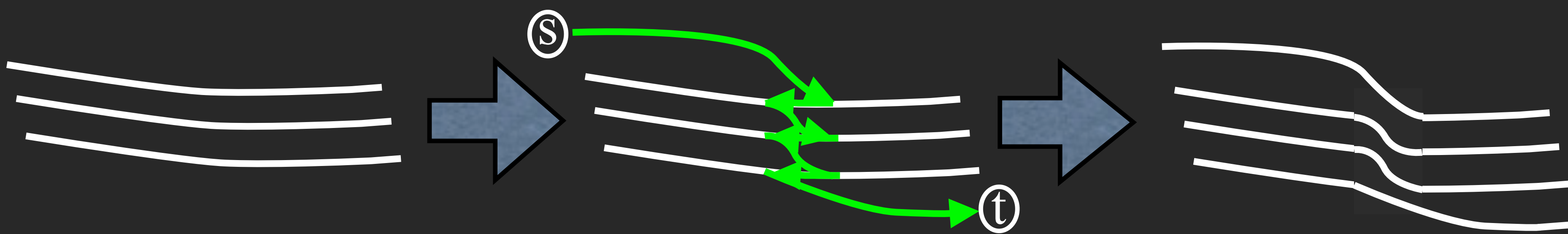
Cost of a K-unit flow = sum of flow along each edge * cost

1) Capacity along each edge is 1
2) Sum of flow into a node = sum of flow out
(ensures non-overlapping tracks)

Exact solution for K>1

Problem: once we instantiate a track, we cannot edit it

Solution: compute shortest path on residual graph
augmented with reserve edges



New tracks can "suck flow" out of existing tracks

Keep repeating until next instantiated track increases cost

Okay... so what about tracking articulations?

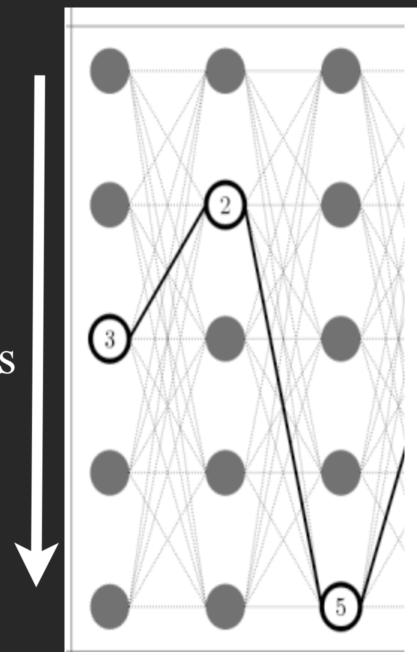
Which one is correct?

What should a single-image pose estimation alg. output?

N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex model

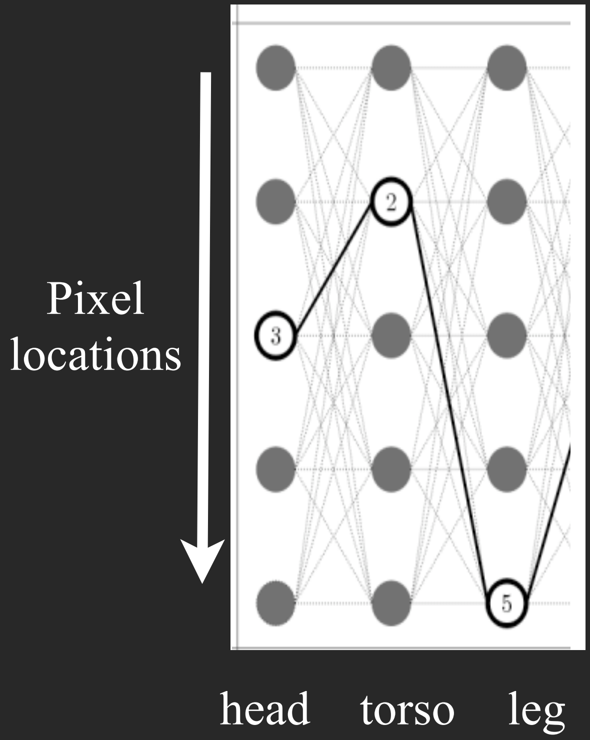
Popular in speech, but why not vision?



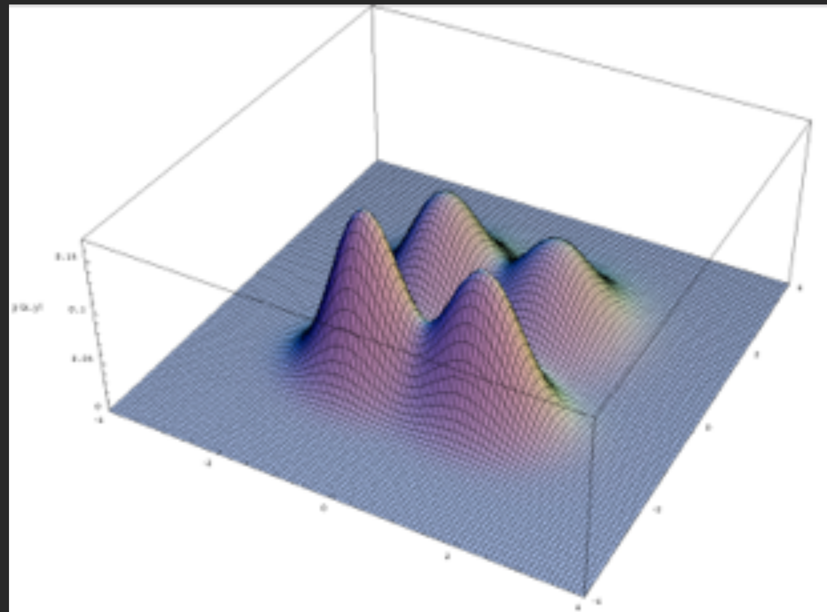
N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex model

Popular in speech, but why not vision?



N-best maximal decoding

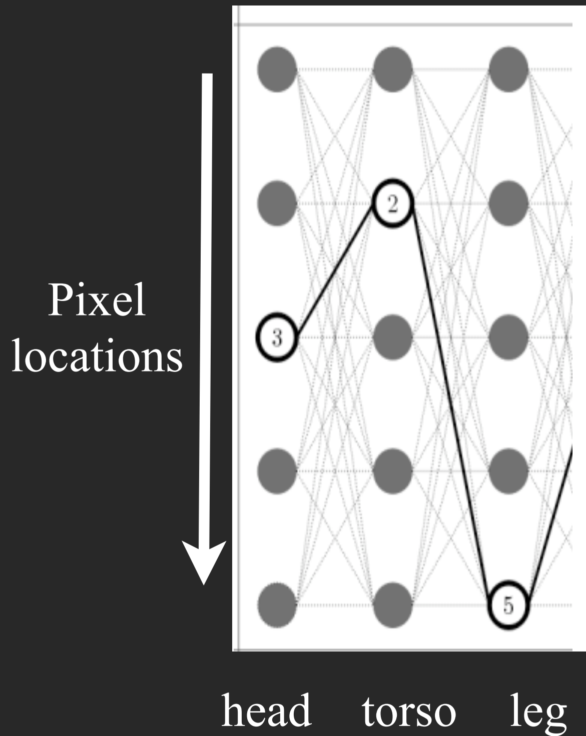


N-best with “NMS” or “mode-finding”

Park and Ramanan, ICCV11

Yadollahpour et al. ECCV12

N-best maximal decoding



Intuition: backtrack from all part “max-marginals”, not just root

(can we done without any noticeable increase in computation)

Park and Ramanan, ICCV 2011

N-best maximal decoding

Park & Ramanan, "N-best decoders for part models" ICCV 2011

Find N-best "modes" rather than N-best poses

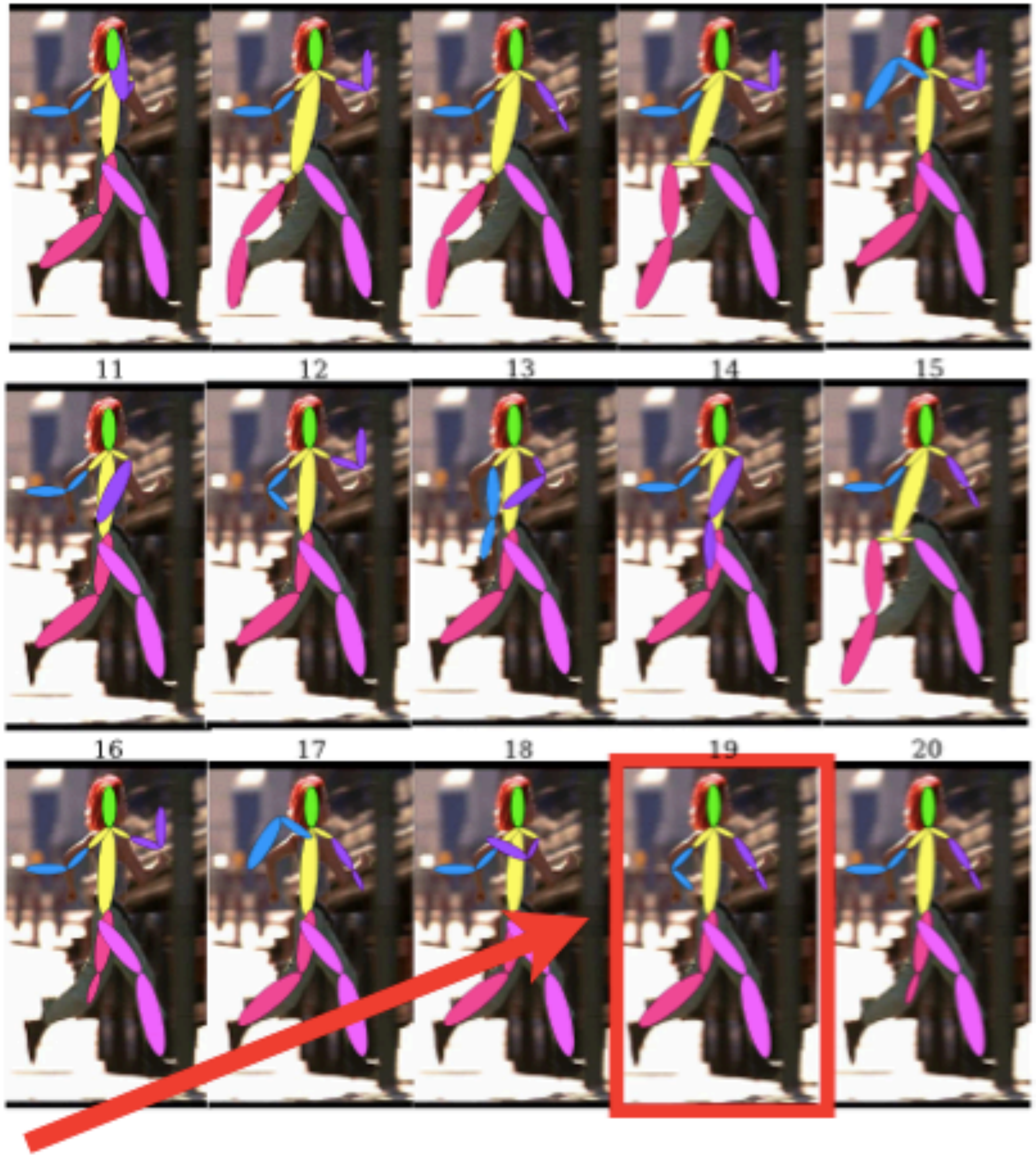
Philosophy: Delay hard decisions as much as possible

Candidate interest ~~points~~

Candidate ~~parts~~

Candidate poses

Maximal poses from a single frame



Correct one picked out by temporal context (tracker)

Aside: other ways of representing uncertainty

$$P(z|x) \propto e^{S(x,z)}$$

Log-linear conditional models

Ramanan NIPS 06

Tracking by articulated detection

Problem: linking up these detections won't work

Recall: Why is finding people difficult?

variation in appearance

variation in pose and viewpoint

occlusion & clutter

Classic “nuisance factors” in image recognition

Recall: Why is finding people difficult?

~~variation in appearance~~

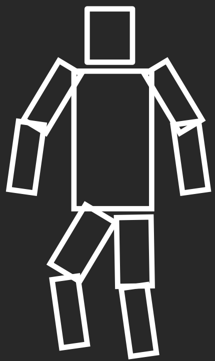
variation in pose and viewpoint

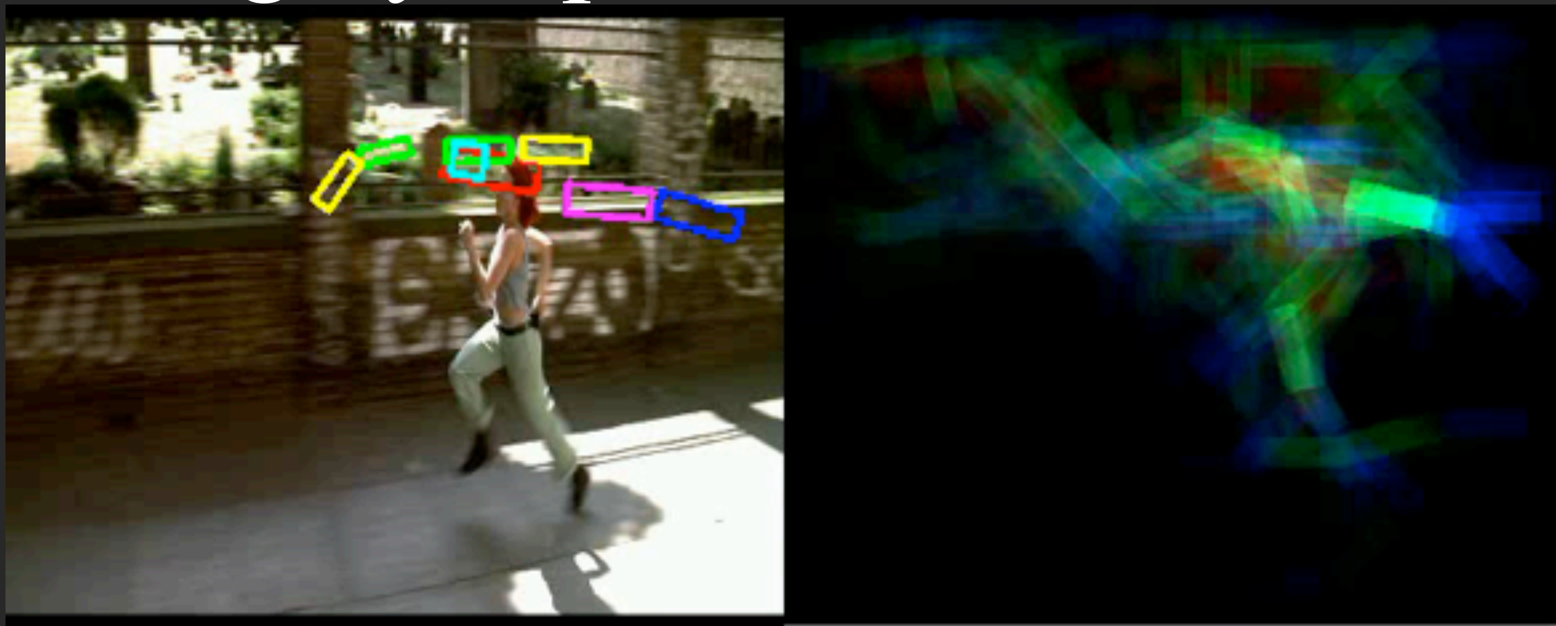
occlusion & clutter

Classic “nuisance factors” in image recognition

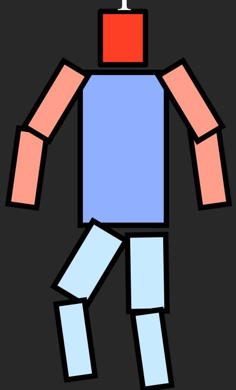
Tracking by repeated detection

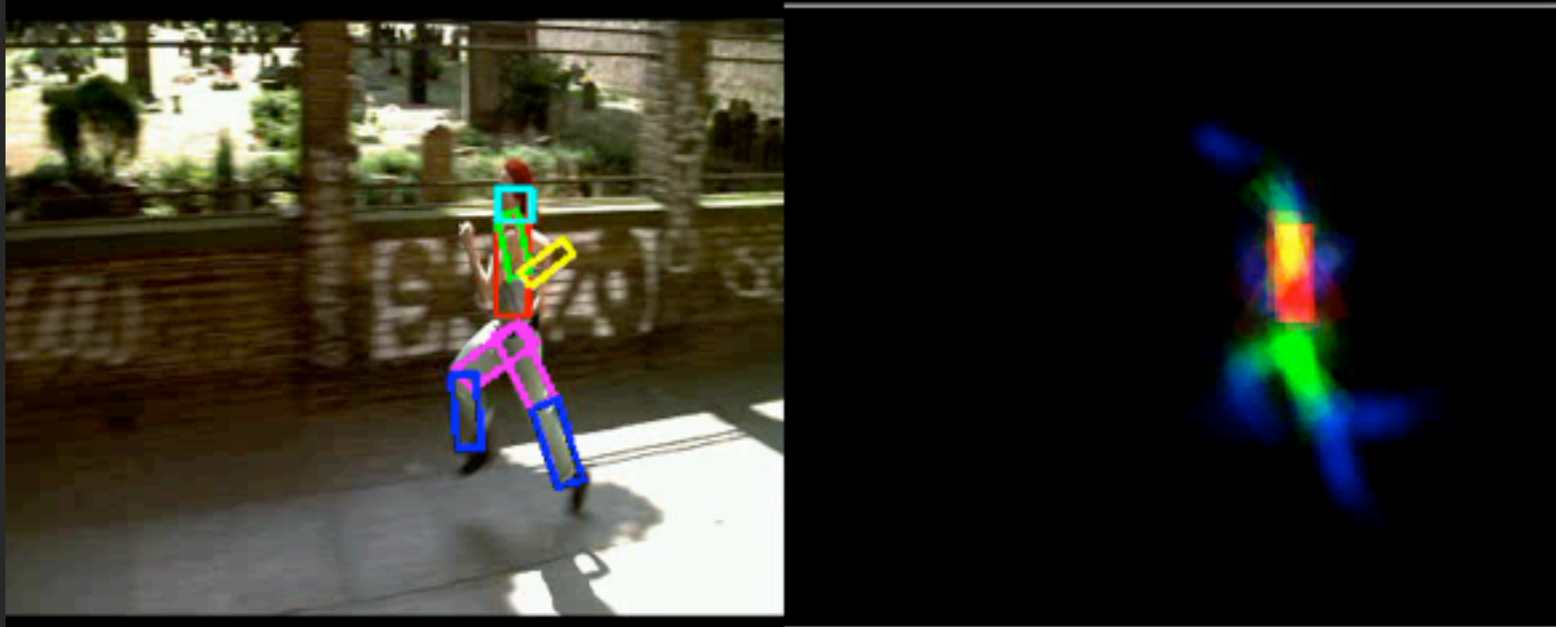
Generic Person Template



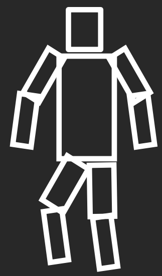


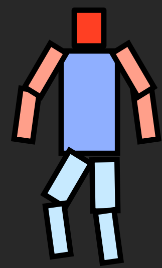
‘Lola’ Template





Tracking as model-building

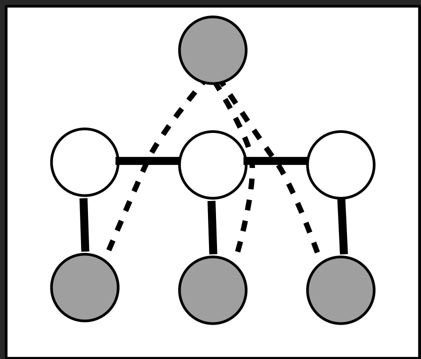


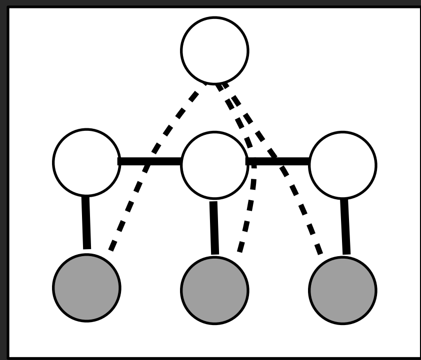


Model-based tracking

Latent-variable tracking

ORourke & Badler 80
Hogg 83
Rehg & Kanade 95
Ioffe & Forsyth 01
Toyama & Blake 01
Sigal et al. 04



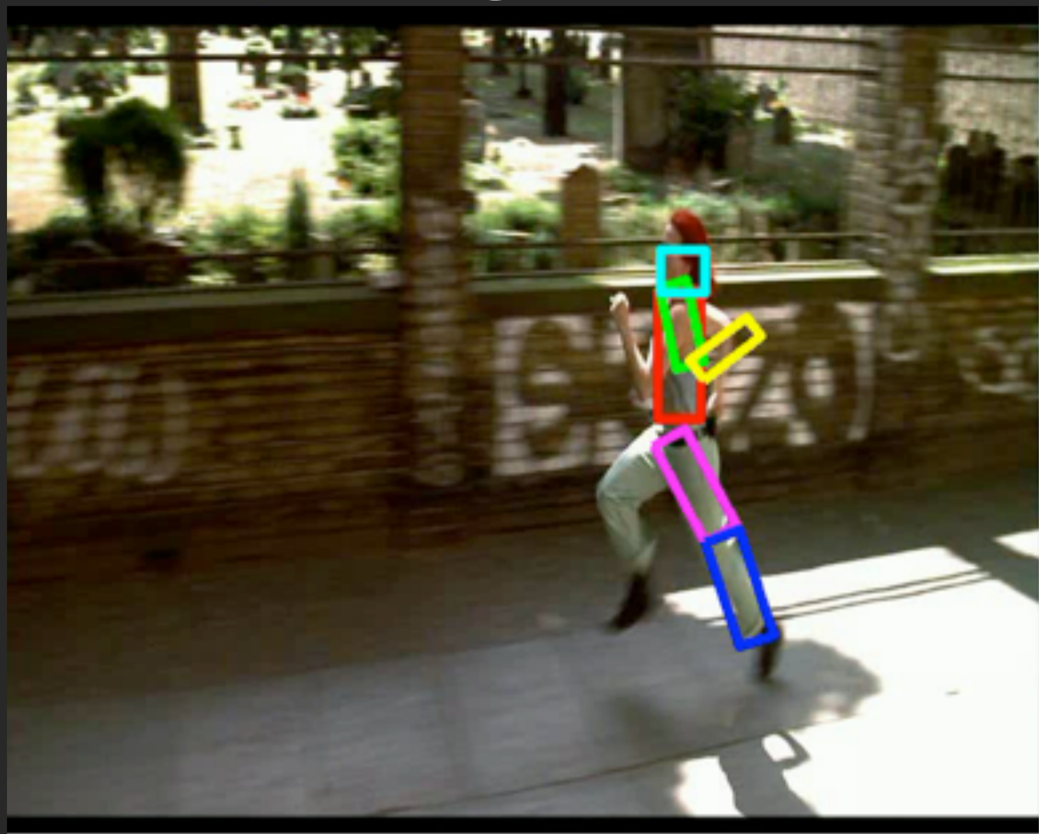


Ramanan et al.
PAMI 07

A generic object template must be invariant

We want to build a model of the object as we track it

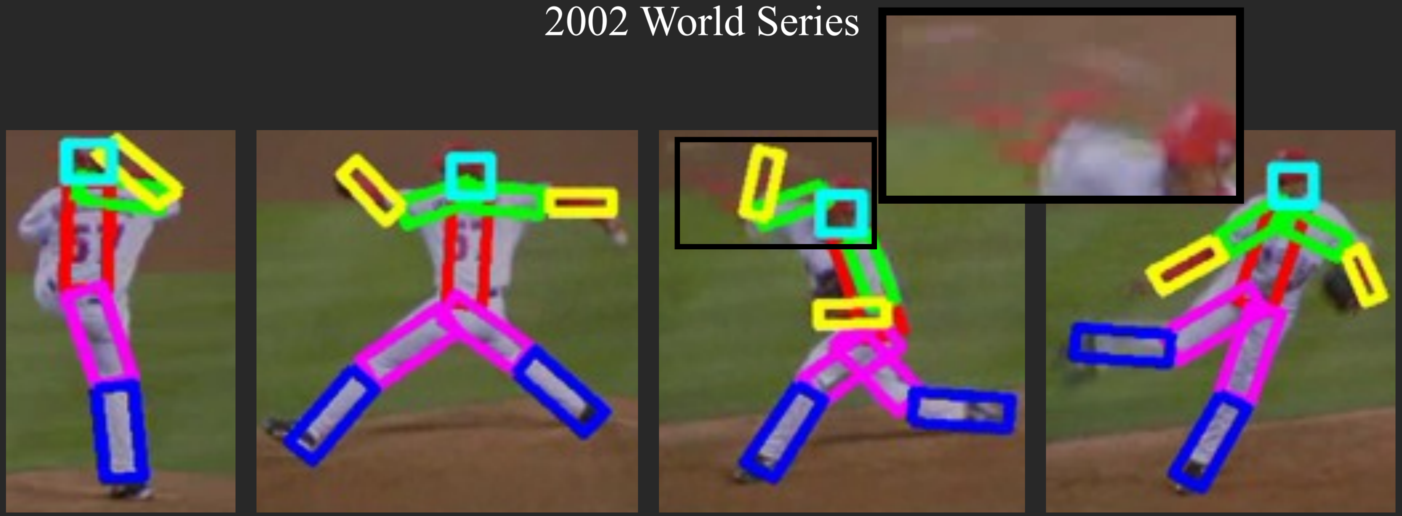
Track through occlusions



Ramanan, Forsyth, and Zisserman PAMI 07

Discriminative clothing models

2002 World Series



motion blur &interlacing

Track long footage (10,000 frames)

Michelle Kwan,1998 Olympics

extreme pose　　motion blur　　fast movement

Olympic woes

silver, not gold →

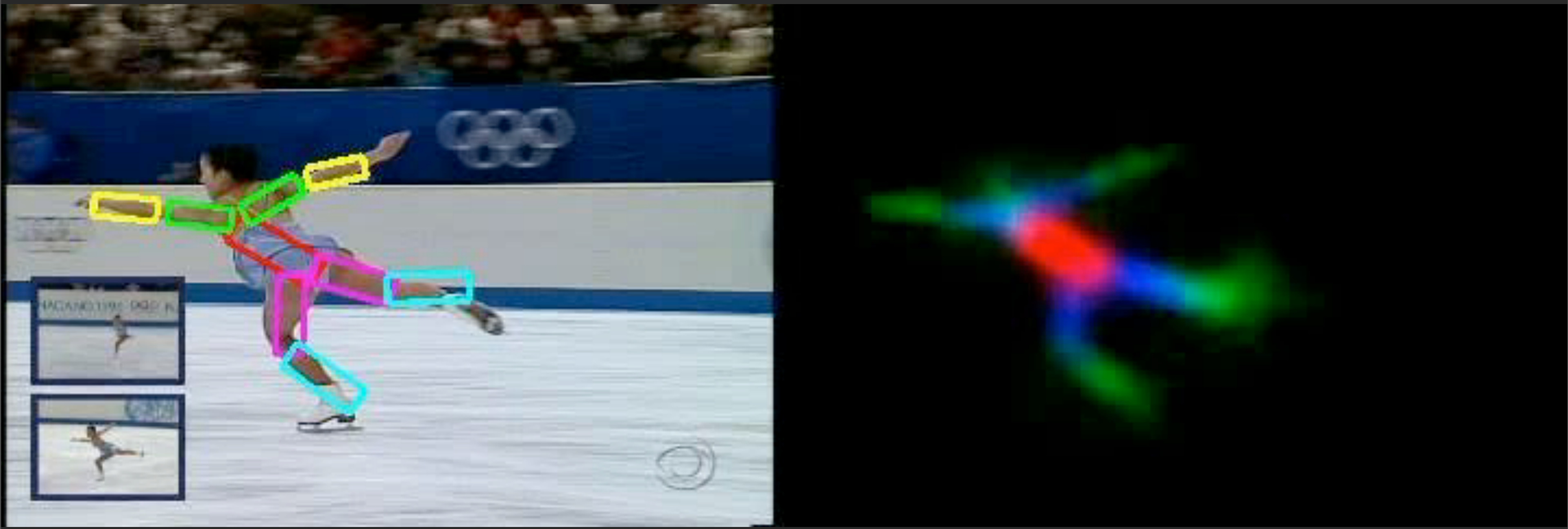
Olympic woes

silver, not gold →

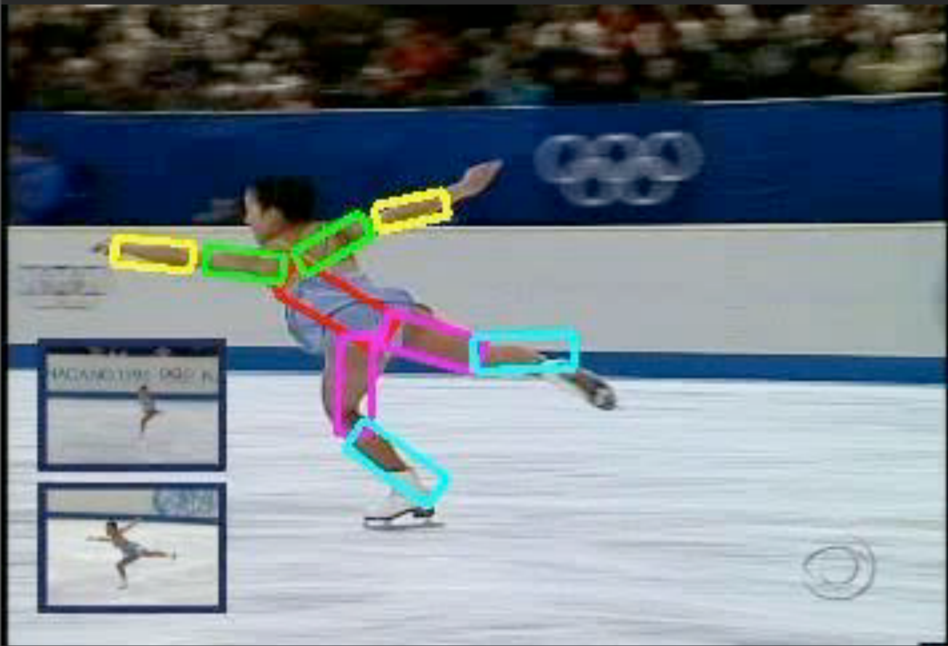
Kwan led after the short program. In the long program, skating to Lyra Angelica by the British composer William Awyn, the 17-year-old turned in a clean, if cautious, effort. Kwan didn't make a major error -- with only one slight wobble on a triple jump -- earning her a solid row of 5.9s on presentation from the judges. As flowers rained upon the ice from her fans, the gold medal, it seemed, was hers. Still, her conservative routine earned five 5.7s for technical merit, and the door was opened, however slight, for Lipinski.

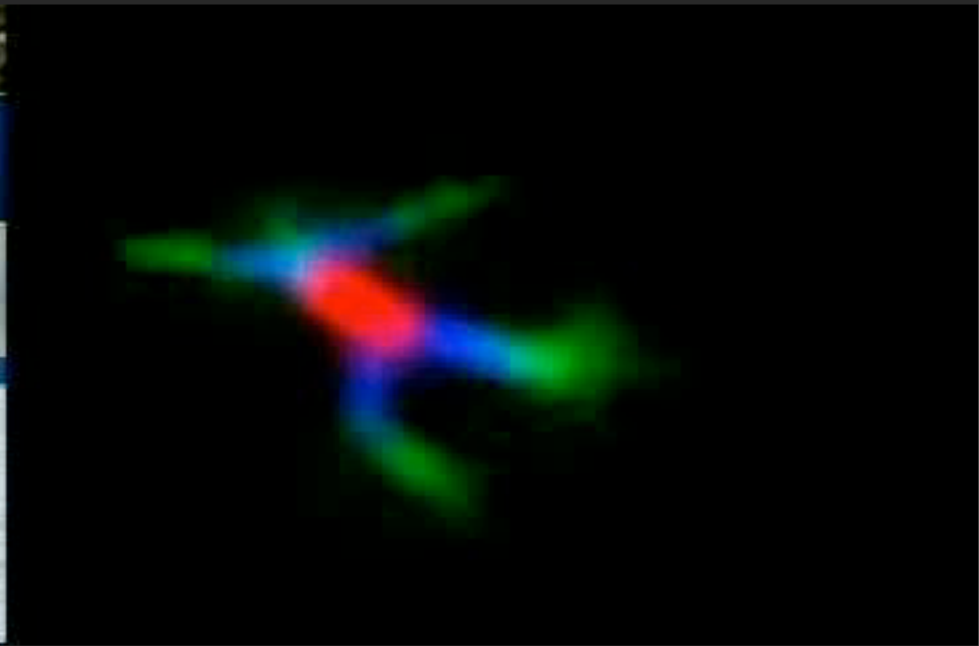
http://espn.go.com/classic/biography/s/Kwan_Michelle.html

The culprit



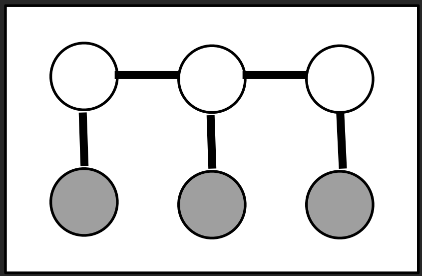
The culprit





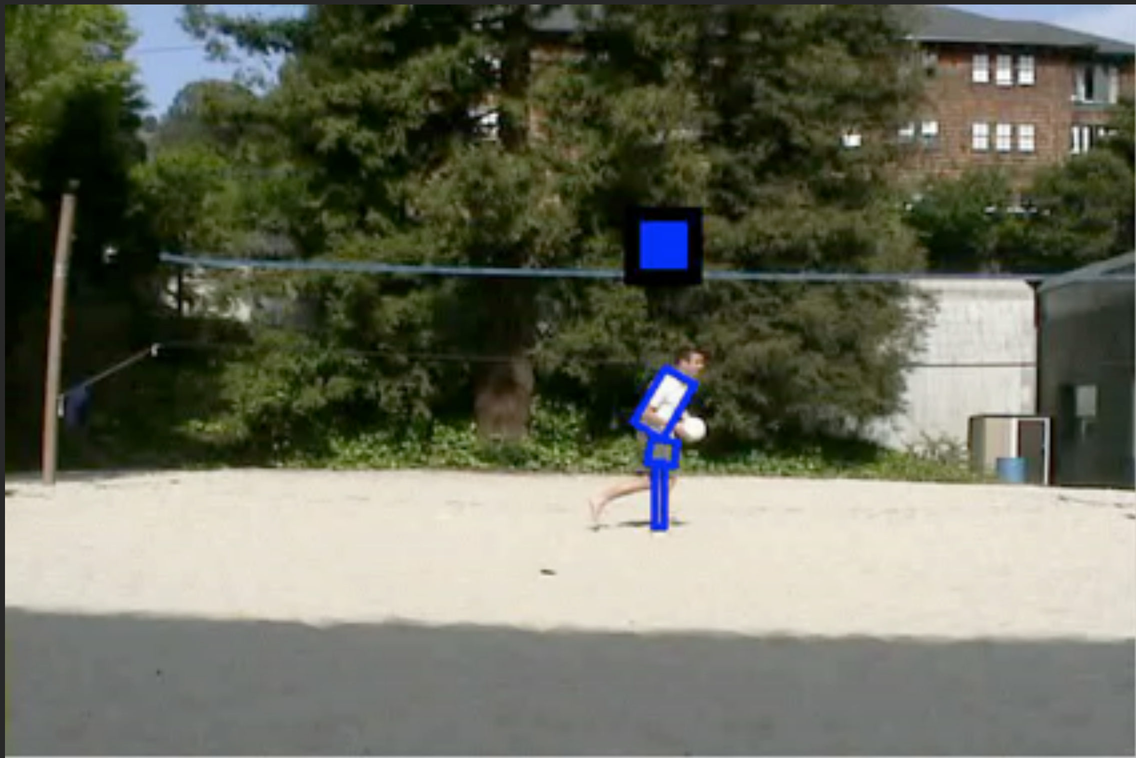
Unexpected/unlikely motions often very important

The motion prior $P(z_{t+1}|z_t)$ may smooth out such subtleties



Tracking multiple people

Independently track each figure



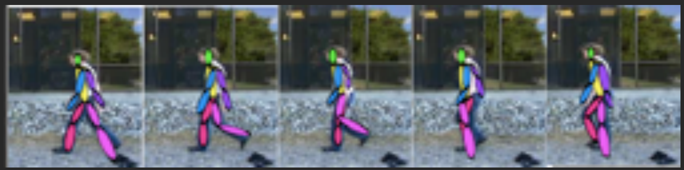
Ramanan & Forsyth
CVPR 03

Clothing appearance is no longer a nuisance

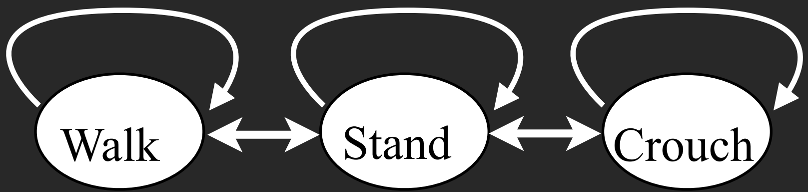
Roadmap

Data/benchmark analysis

Spatiotemporal features



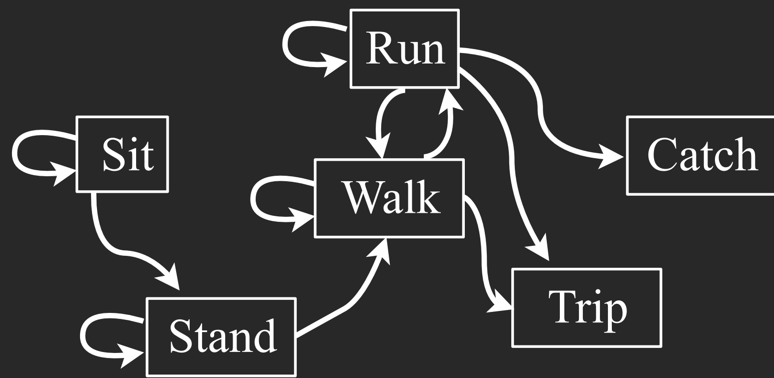
Spatiotemporal models



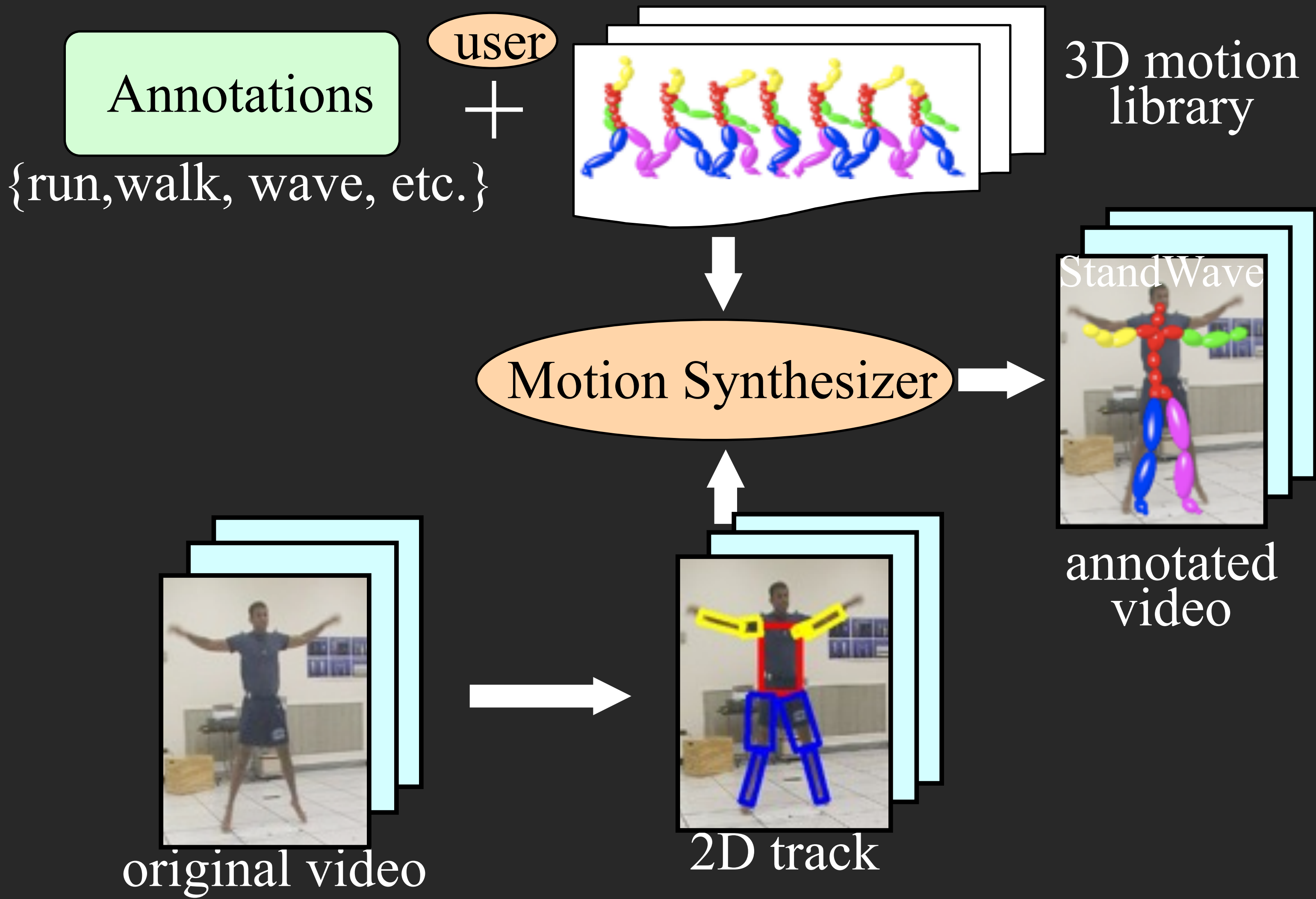
Spatiotemporal models

Data-driven

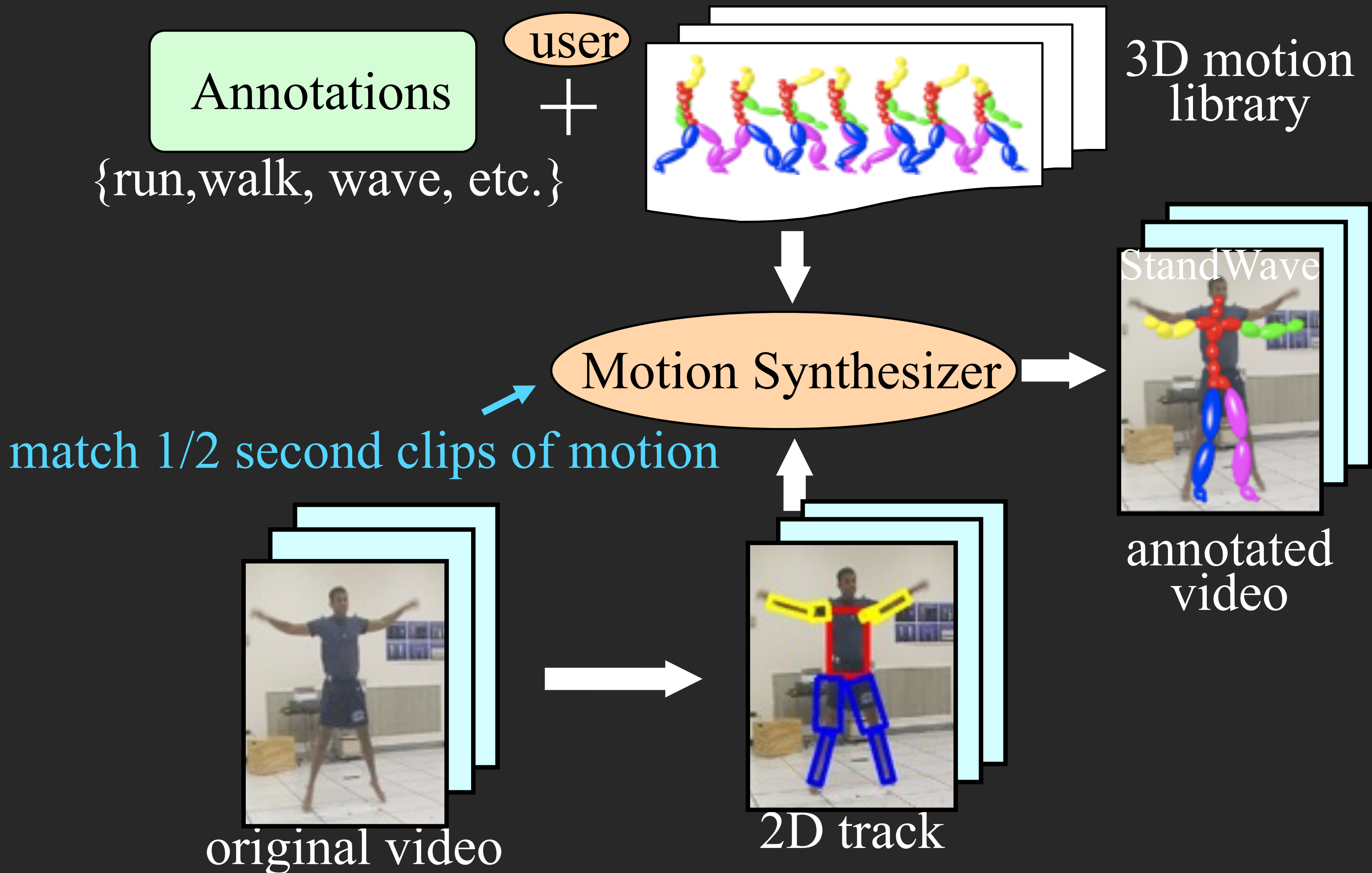
Model-driven



Data-driven action recognition



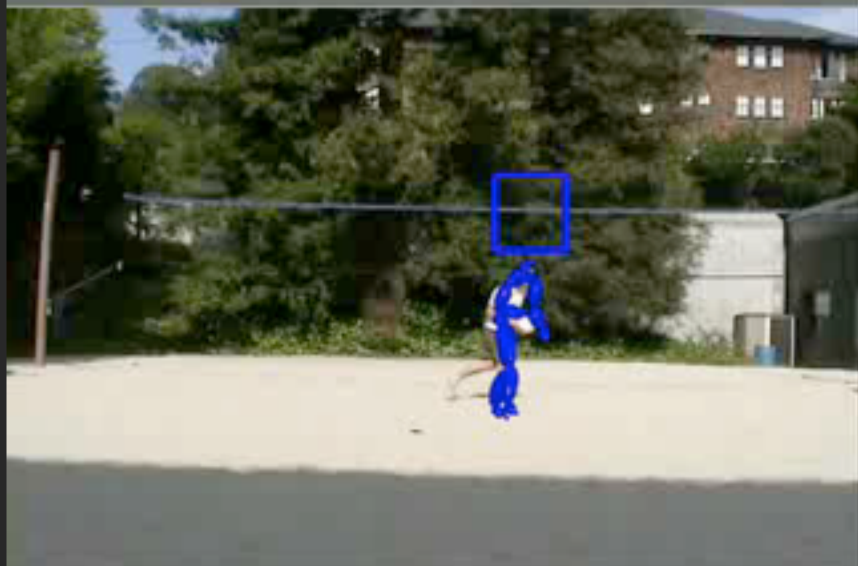
Data-driven action recognition



Results

(Low-level) tracking

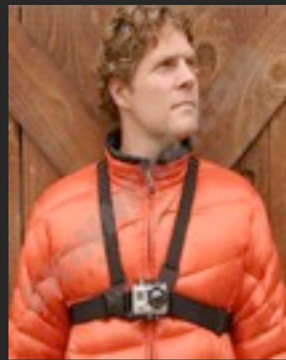
(High-level) spatiotemporal models



Pipeline surprisingly rare (e.g., doesn't work on TrecVid)

Recognizing structured actions

Making tea from a wearable camera



Start boiling water

Do other things (while waiting)

Pour in cup

Drink tea

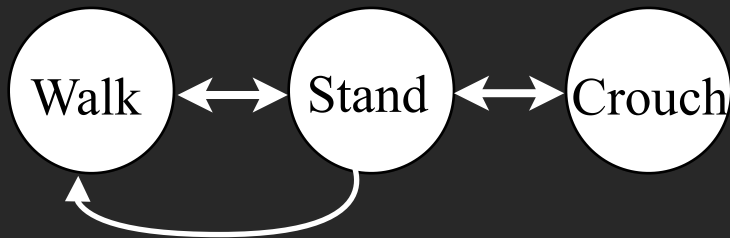
start boiling water

wait

steep tea leaves

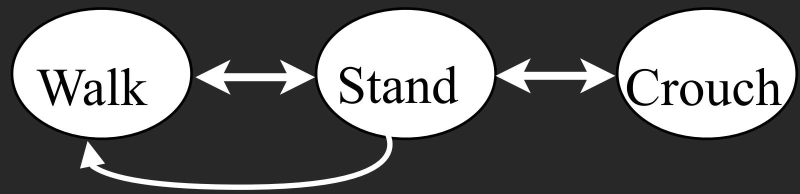
time

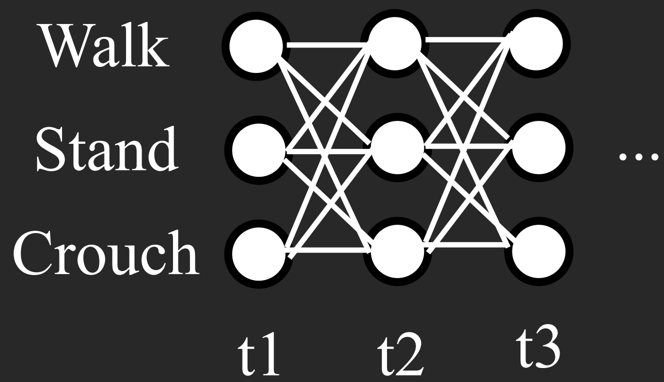
How do we capture long-term structure?



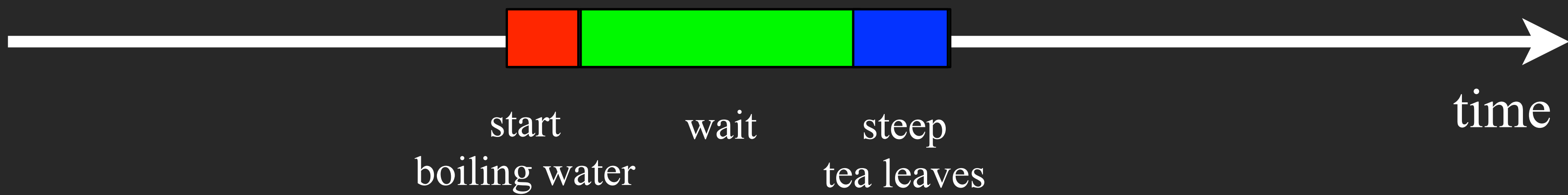
Markov models

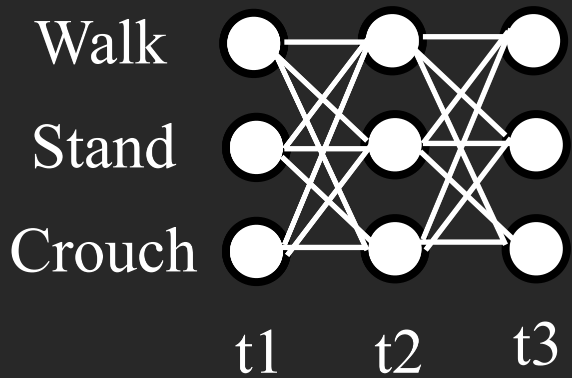
What’s magic behind semi-markov models?

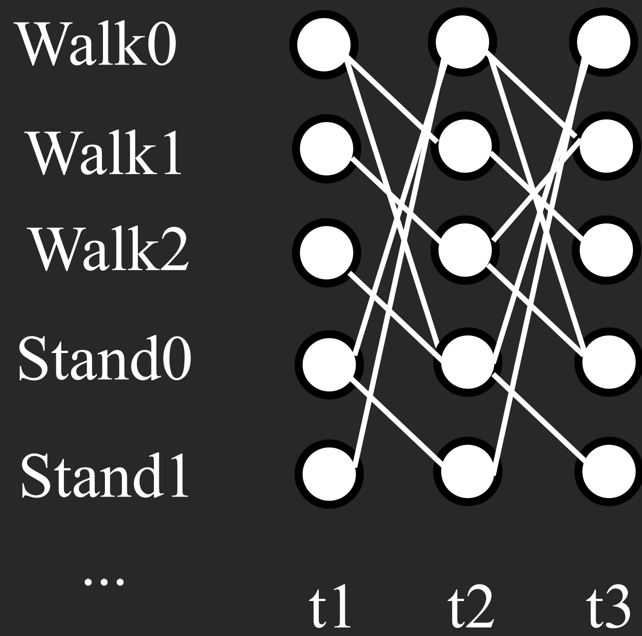




Semi-markov models



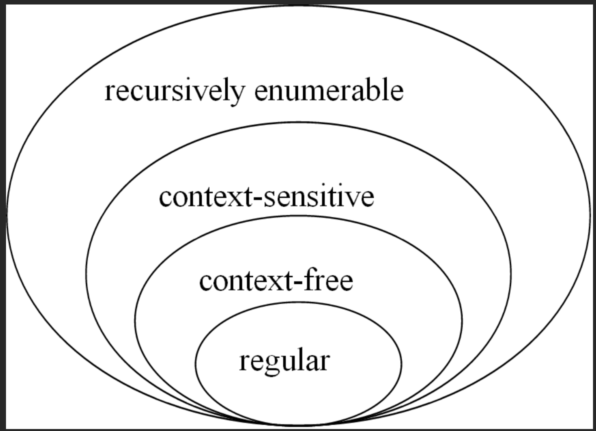




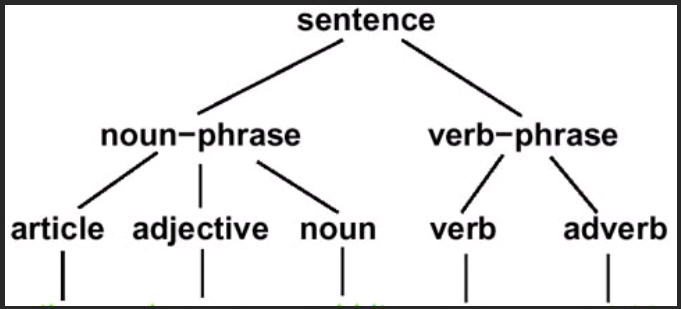
Add counting states and force sparse transitions (Walk0 to Walk1)

Counting state costs can model arbitrary priors over segment lengths

How do we capture long-term structure?



Exploit models for language



“The hungry rabbit eats quickly”

Context-free grammar

Example grammar

"yank"

"pause"

"press"

"background"

Clean&Jerk action =

Snatch action =

$$S \rightarrow$$

$$S \rightarrow S$$

$$S \rightarrow S$$

Example parse

time

Zhu et al
Bobick et al

A look back

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models