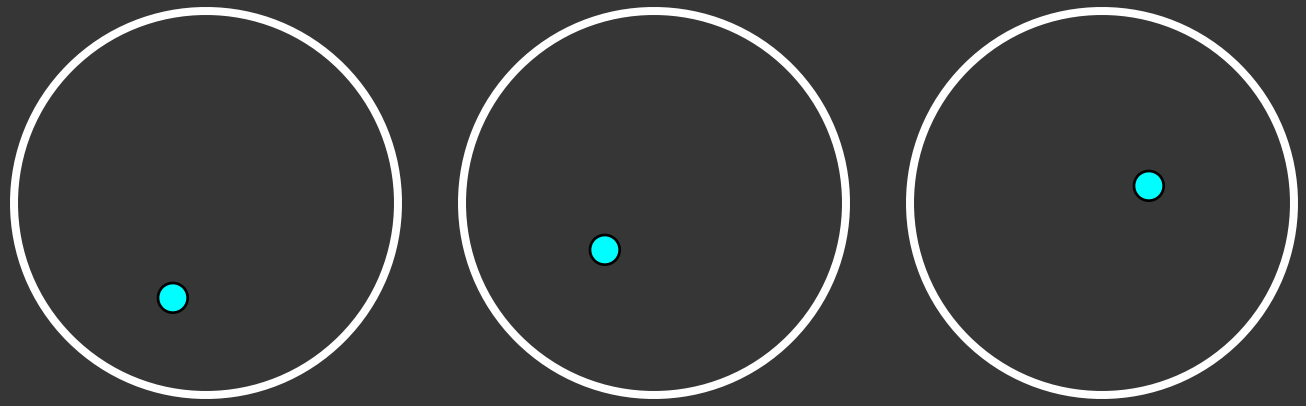
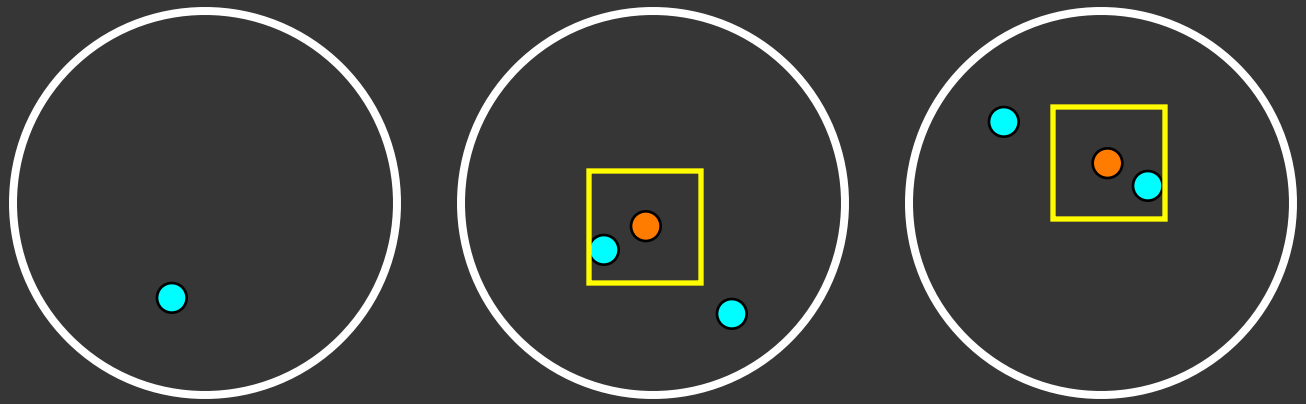
People Tracking: Data Association

Deva Ramanan
Toyota Technological Institute at Chicago

What's data association?

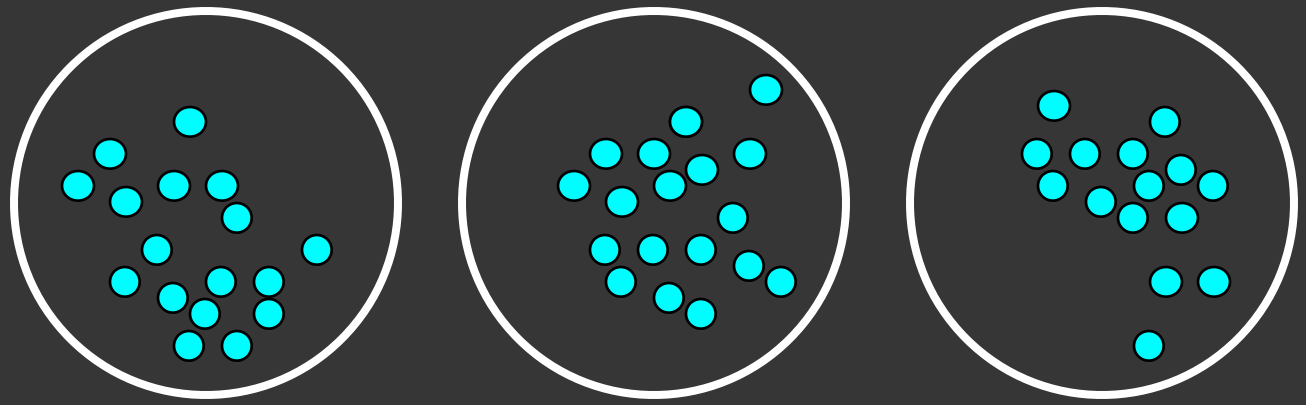


"Bare-bones" Kalman filter: acts like a smoother



Simple strategy: gate window around prediction

But what about….



Millions of alternative measurements
(Better analogy with images)

Can't just use motion prediction

Inference involves not just smoothing,
but identifying which measurement to smooth

Is it that bad?

Probably not

Observation: 95% of the time, people + backgrounds are boring
Can track using motion priors and/or background models
Is data association solved?

What about other 5%?

Chicago White Sox
World Series

Andy Serkis's performance
Lord of the Rings

Berkeley campus

What about other 5%?

Chicago White Sox
World Series

Andy Serkis's performance
Lord of the Rings

Berkeley campus

(Perhaps) its more interesting to track

Why is finding the "people-pixels" hard?

variation in appearance

variation in pose & aspect

occlusion & clutter

Roadmap of what lies ahead

- Caveat: Not as cool as Cristian/David's talks
- Big issues:
  - Invariances (pose, illumination, intra-class)
  - Image features
  - Search (top-down vs bottom-up)
- Very similar to object detection
  - see ICCV05 tutorial `Recognizing and Learning Object Classes' by Fei-Fei, Fergus, and Torralba

Strategy 1: Pixel-based approaches
bg subtraction

- Subtract im from bg estimate
- bg estimate = known image or statistical average of history

Haritaoglu et al. PAMI00, Stauffer & Grimson, PAMI00

Strategy 1: Pixel-based approaches
fg enhancement

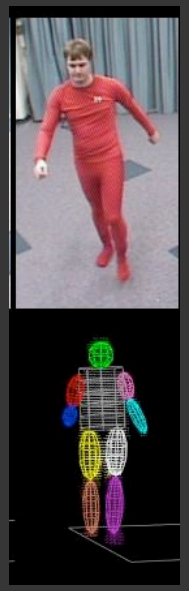
- skin detection

-Compute P(rgb|skin)
vs P(rgb|~skin)

-Tuned for Caucasians

Jones and Rehg, IJCV02
Fleck et al ECCV96

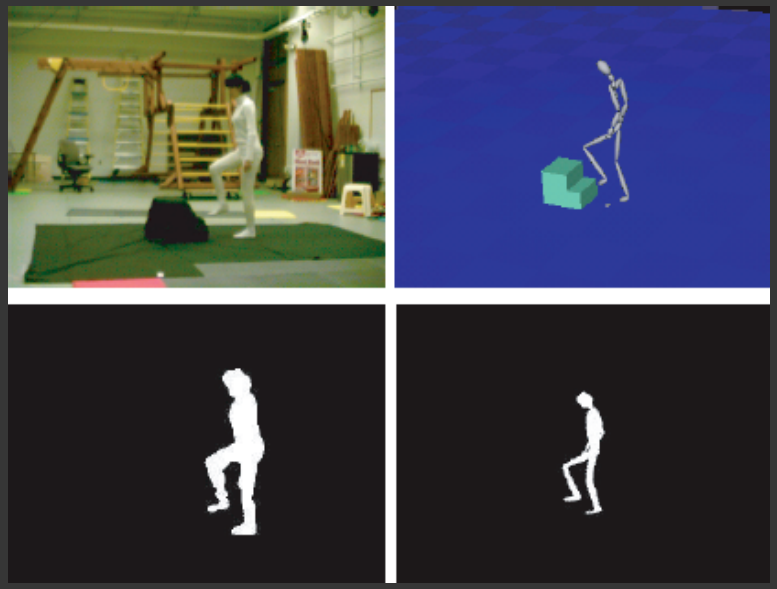
- color detection



Mikic et al CVPR01

Strategy 1: Pixel-based approaches
fg enhancement

If it can be used, it generally should be!

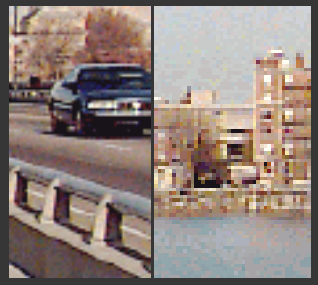


Lee et al, SIGGRAPH02

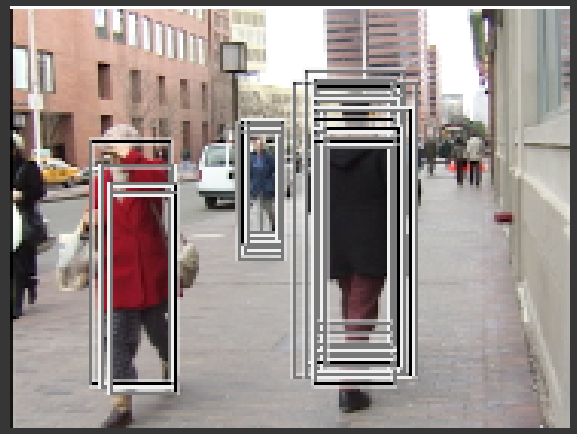
Easy to implement & reliable in controlled situations (ie, markerless motion capture)

Strategy 2: Scanning window

(+)



(-)



Papageorgiou and Poggio, ICIP99

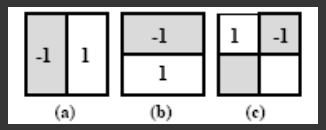
Dalal and Triggs, CVPR05

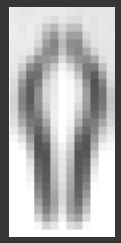
Learn pedestrian vs background
classifier from training data

Strategy 2: Scanning window features

Need invariance to appearance; focus on contours

- Haar wavelet features



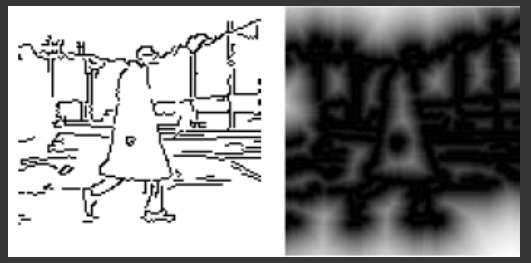


Papageorgiou and Poggio, ICIP99

- Histogram of Gradients (HOG)/SIFT descriptors

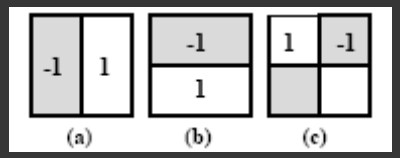
Dalal &Triggs CVPR05

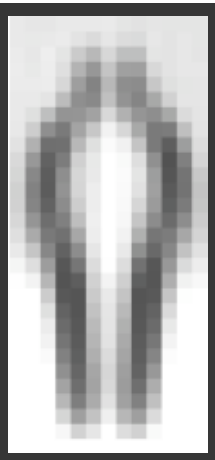
- Edges
(evaluated with chamfer score)



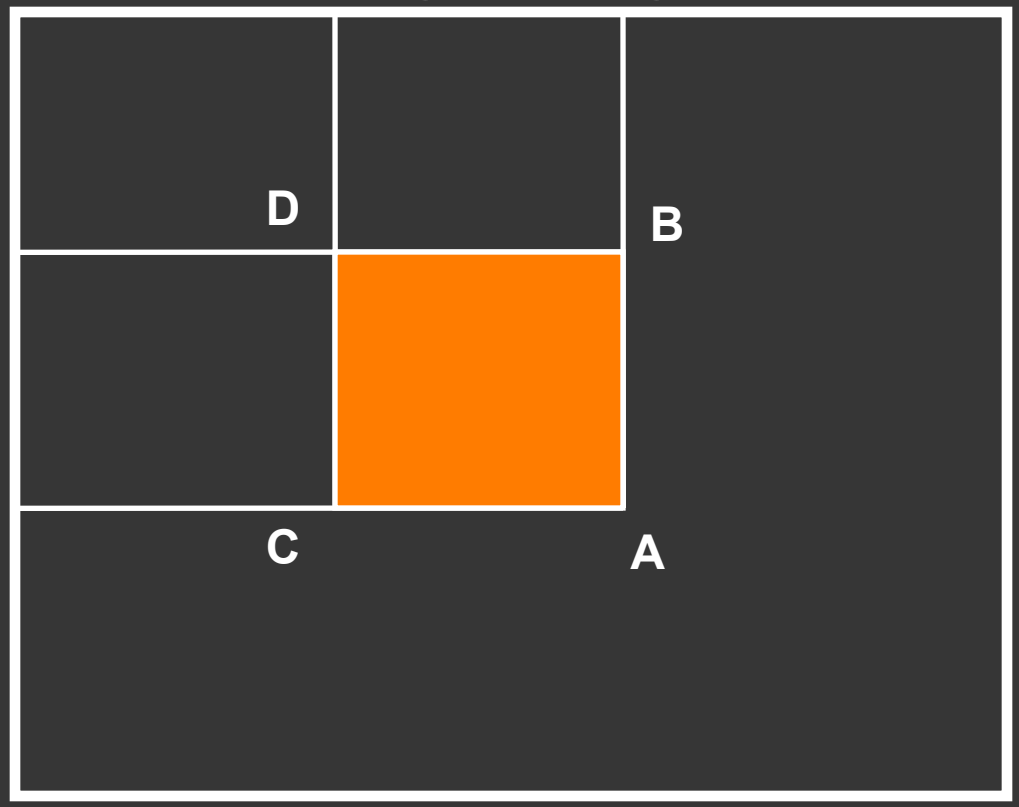
Gavrila and Philomin ICCV99

Features: Haar wavelets





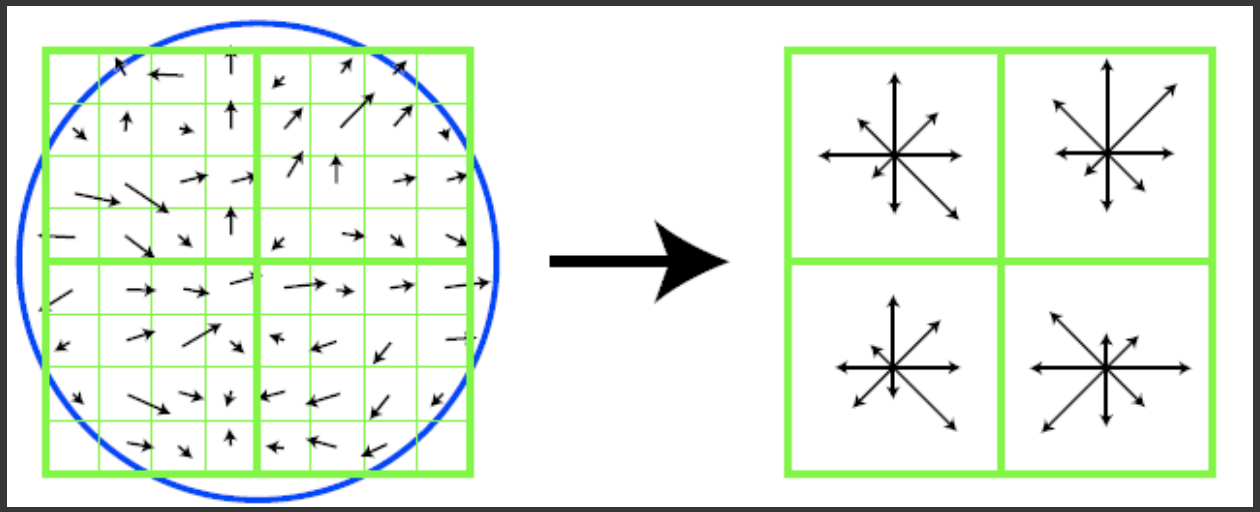
Integral Image



Sum = A - B - C +D

Viola Jones CVPR01

Features: histograms of gradients



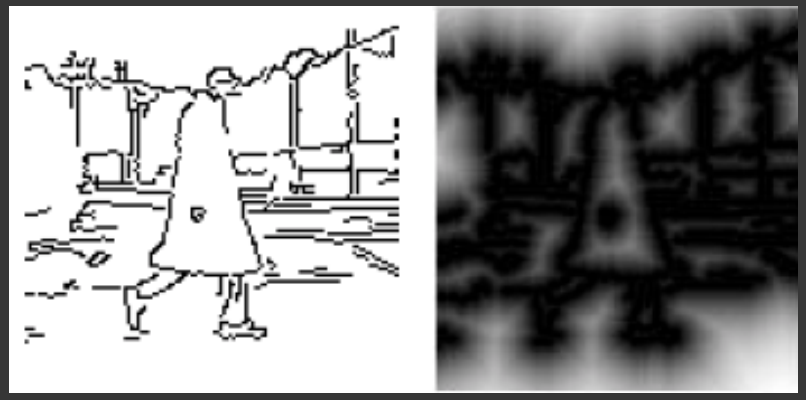
Gradients within 8X8 patch

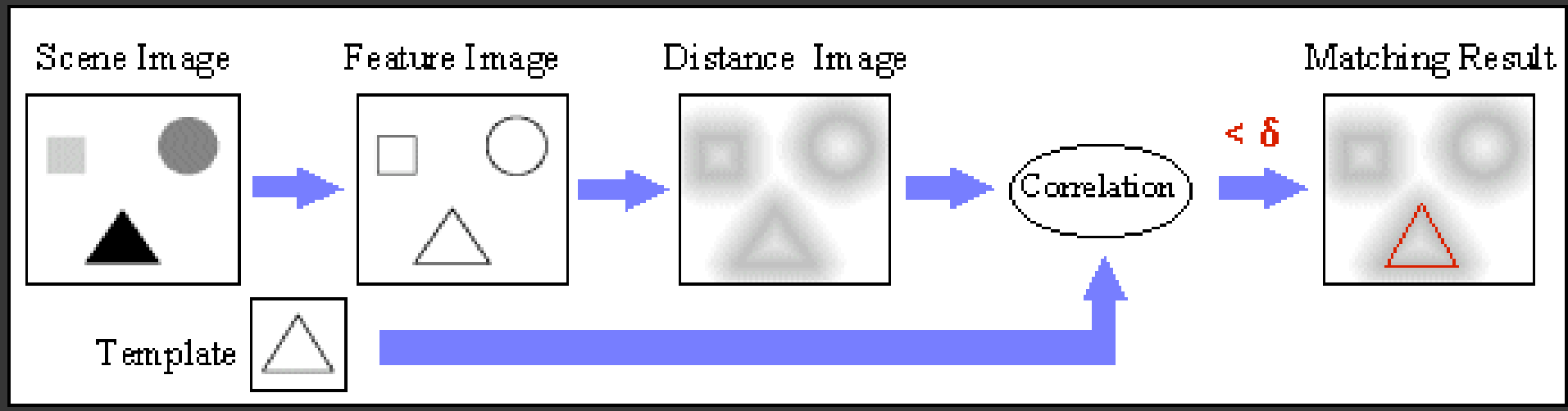
Bin into local (4X4) neighborhoods
& 8 orientations

Lowe IJCV2004
Dalal &Triggs CVPR05
Freeman and Roth IAFGR 1995

Binning achieves invariance to small patch offsets

Features: oriented chamfer edges





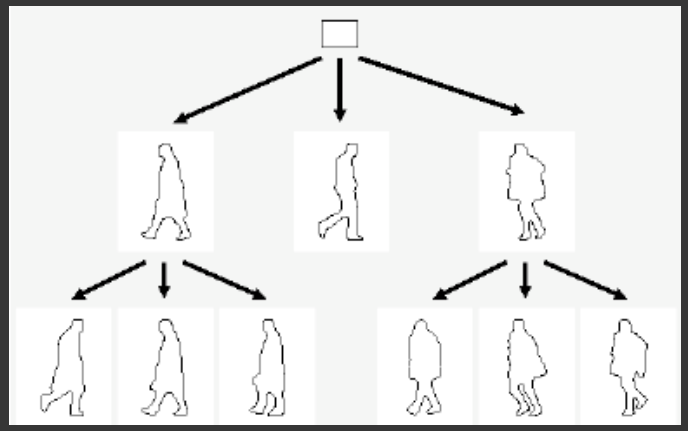
Gavrila and Philomin ICCV99

Matching can handle small deformations in the template/scene

Strategy 2: Scanning window
efficient scanning

•Coarse-to-fine search

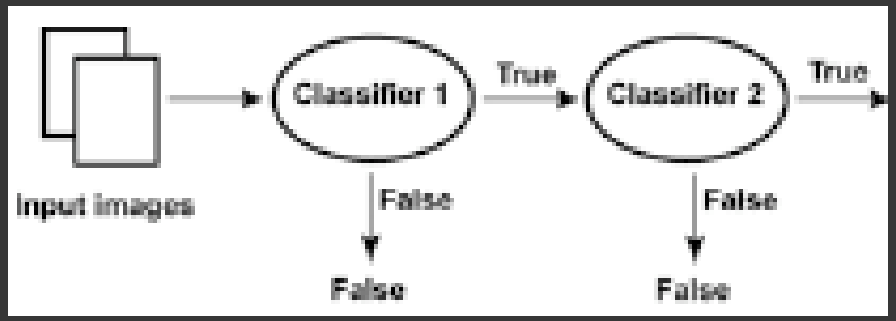
-coarse-to-fine in both template and image domain



Gavrila and Philomin, ICCV99
Stenger et al, ICCV03

•Cascade

-prune away most windows with initial classifier



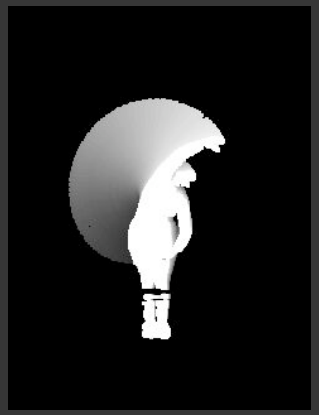
Viola and Jones CVPR01
Viola et al, ICCV03

Strategy 3: XYT window

Single frame might not be enough to find person

Motion History Image

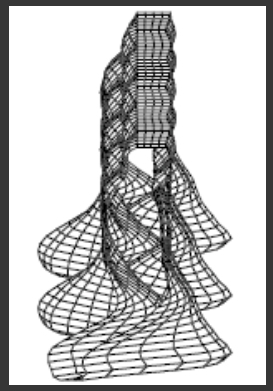
1)Do bg subtraction
2)MHI = pixel is brighter the more recently it was fg

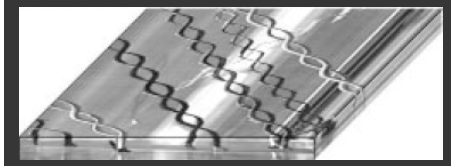


Bobick and Davis, PAMI01

Strategy 3: XYT window (cont'd)

- Look for symmetry in XYT slices

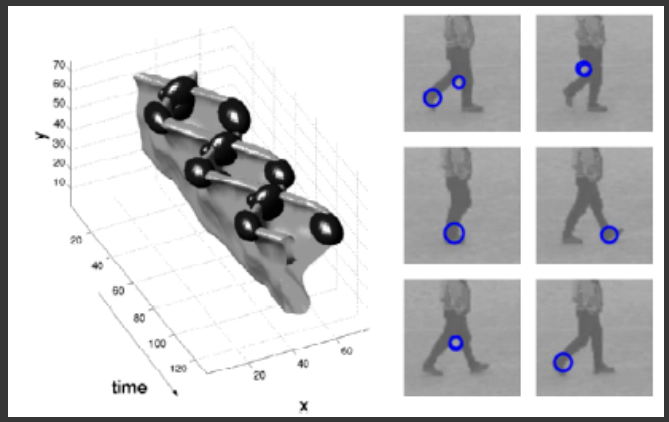




- Look for XYT interest points

Niyogi and Adelson CVPR94
Polana and Nelson, ICPR94

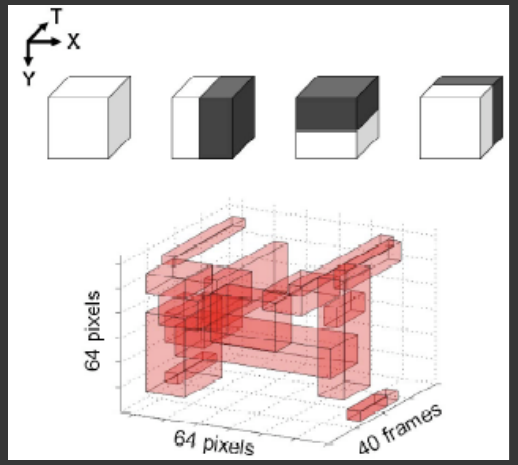
-define harris detector for XYT



Laptev and Lindeberg ICCV03

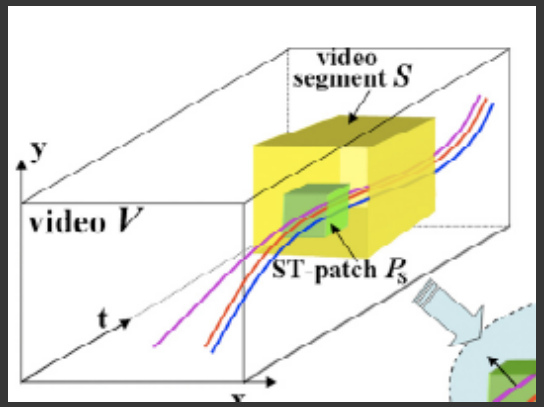
Strategy 3: XYT window (cont'd)

- Define XYT feature for classifier
  - -applied to flow
  - (invariant to appearance)



Viola et al ICCV03, Ke et al ICCV05

- Define XYT template & correlate
  - -use local flow
  - as feature



Shechtman and Irani CVPR05

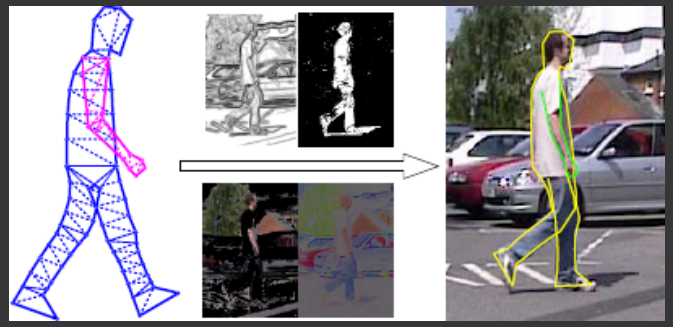
Strategy 3: XYT window (cont'd)

Shechtman and Irani CVPR05

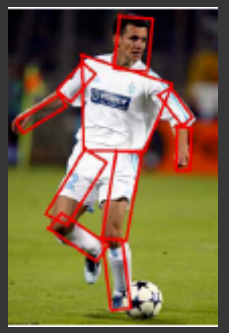
Strategy 4:
Top-down pose estimation

- Compute $P(\Theta|I) \alpha P(I|\Theta)P(\Theta)$ by sampling methods
  - Iteratively search space of body poses Θ
    - sample from prior or data-driven proposal
  - Works well with informative likelihood (skin) and/or prior (walking)

Lee & Cohen CVPR04



Zhang et al, CVPR04

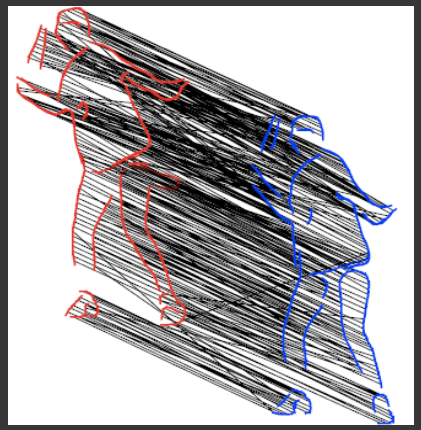


Hua et al CVPR05

Strategy 4: Top-down pose estimation (cont'd)

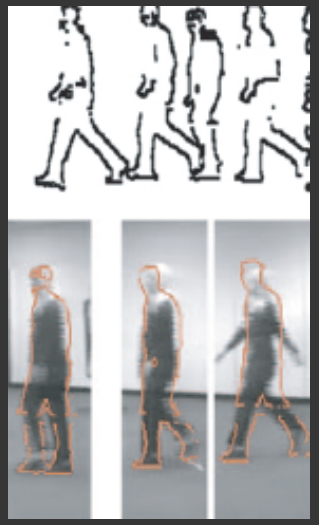
- Match exemplars
  - Encode articulations by templates or on-the-fly deformations
  - Seems to be limited to standard poses (useful for tracker initialization)

Model template



Query edge map

Sullivan & Carlsson, ECCV02
Loy et al ECCV04
Mori & Malik ECCV02



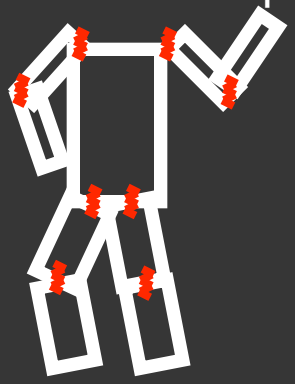
Gavrila & Philomin, ICCV99
Toyama & Blake ICCV01

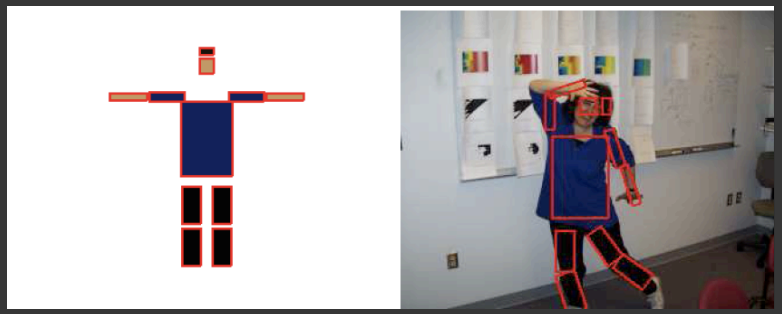
- Efficient search
  - Coarse to fine
  - Approx. Nearest Neighbors (Shakhnarovich et al ICCV03)

Strategy (5)
Bottom-up parts: assembly

Detect parts & then assemble

- Dynamic programming (tree model)
  - For N candidate parts $O(N^2)$, but can speed up to O(N) with distance transform

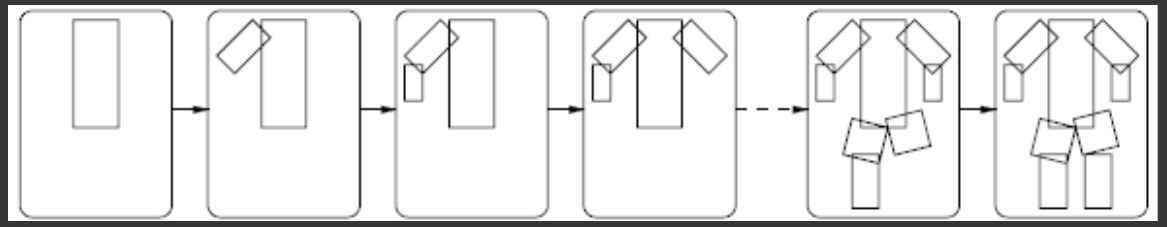




Felzenszwalb & Huttenlocher, CVPR00

Ioffe & Forsyth, ICCV01

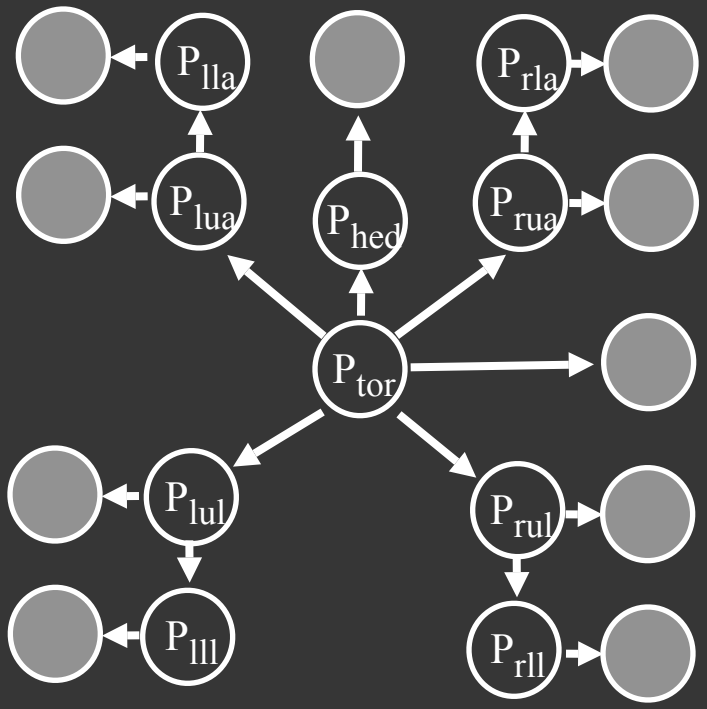
- Iteratively sample good assemblies

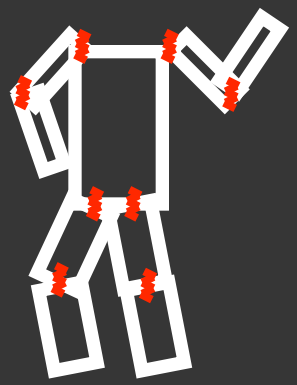


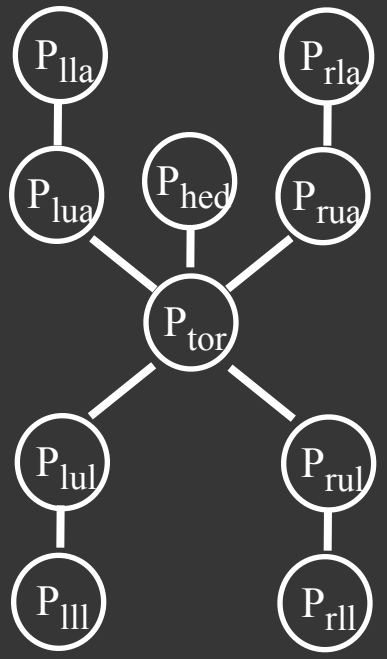
Ioffe & Forsyth, ICCV99

Pictorial structure model

Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)







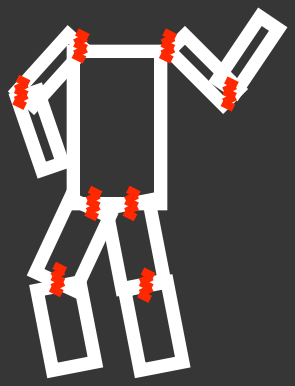
$$\Pr(P_{tor}, P_{arm}, \ldots | \mathrm{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\mathrm{Im}(P_i))$$

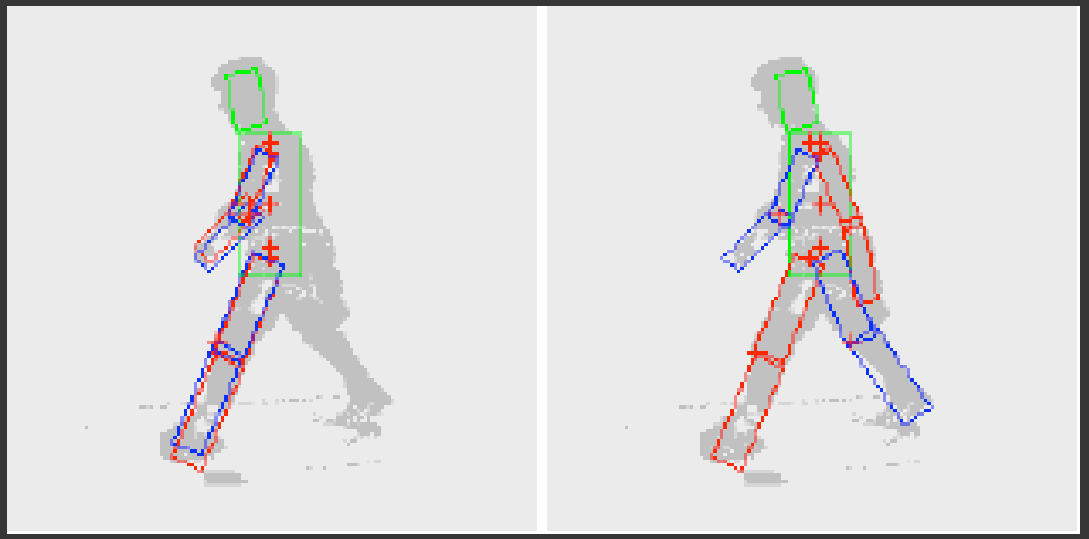
part geometry

part appearance

Trouble with trees

- Limbs attracted to regions of high likelihood
  (local image evidence is double-counted)

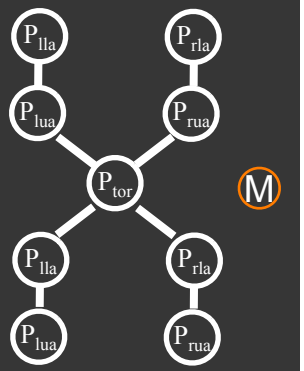


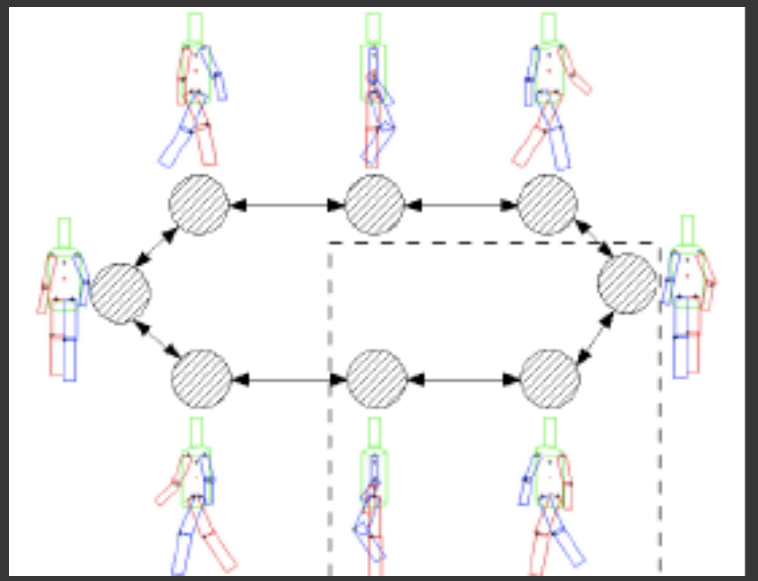


Lan & Huttenlocher, ICCV05

Tree extensions: augment model

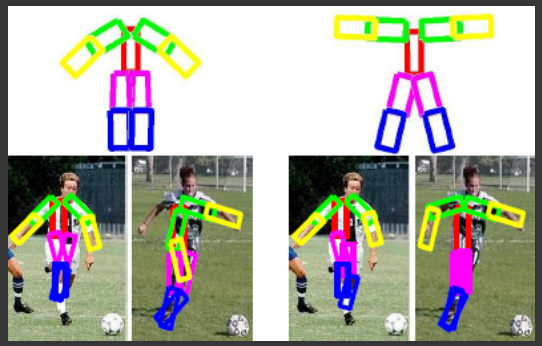
- Latent variable models (mixture component)





Lan & Huttenlocher, ICCV05
Lan & Huttenlocher, CVPR04

- Train tree model discriminatively (CRF)

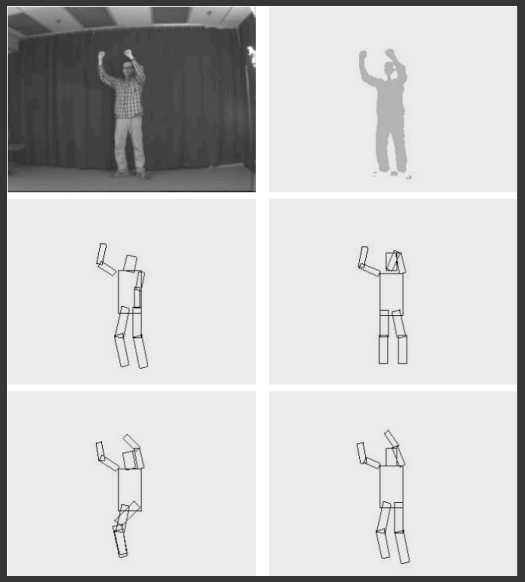


Ramanan & Sminchisescu CVPR06

Tree extensions: sample the posterior

Because its a tree, sample from true posteror; `burn in' time = 0

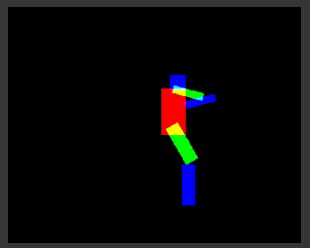
Evaluate samples with global model

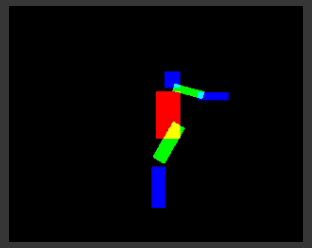


Felzenszwalb and Huttenlocher, IJCV05

Find modes in posterior

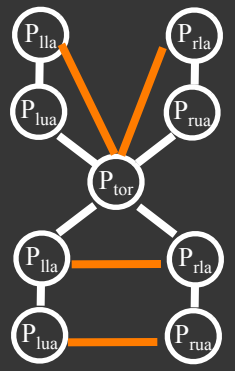
Ramanan et al, PAMI06



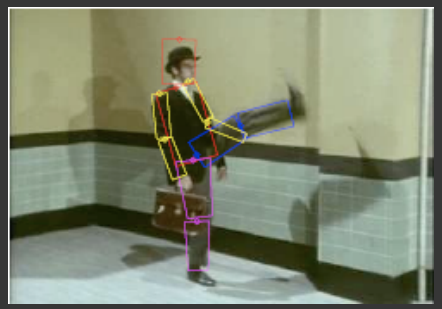


Tree extensions: Don't use a tree!

Add loops enforcing non-occlusion

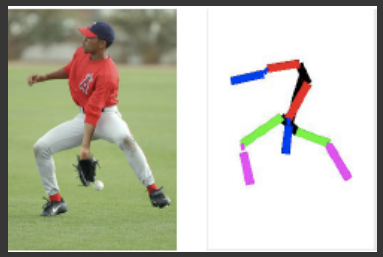


- Use loopy belief propogation (nonparametric msgs)



Sigal and Black CVPR06

- Combinatorial search (integer quadratic program)



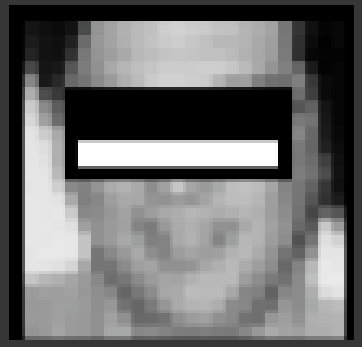
Mori et al, CVPR04

Ren et al, ICCV05

Part-based: What are good parts?

Build a part detector

- Face detector (adaboost, SVMs, NN)



Viola & Jones, IJCV01

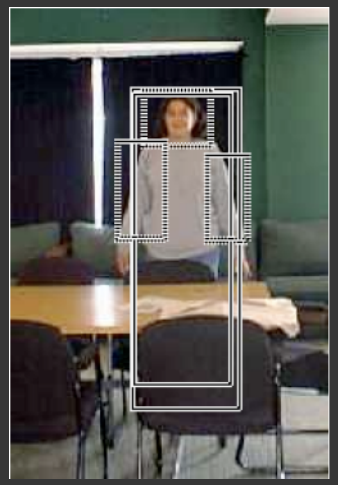
Schneiderman and Kanade, CVPR98

….

- Rich body of literature

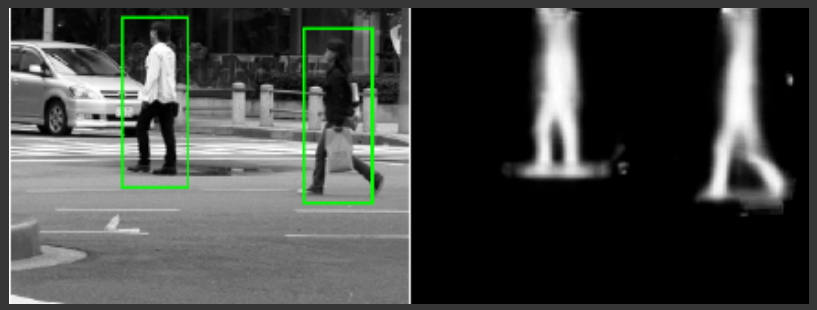
Part-based: What are good parts?

- Train a body part detector with SVM, adaboost, etc.



Mohan et al, PAMI01
Ronfard et al, ECCV02
Mikolajczyk et al ECCV04

- Learn a body part model
  -grayscale patches, filter responses



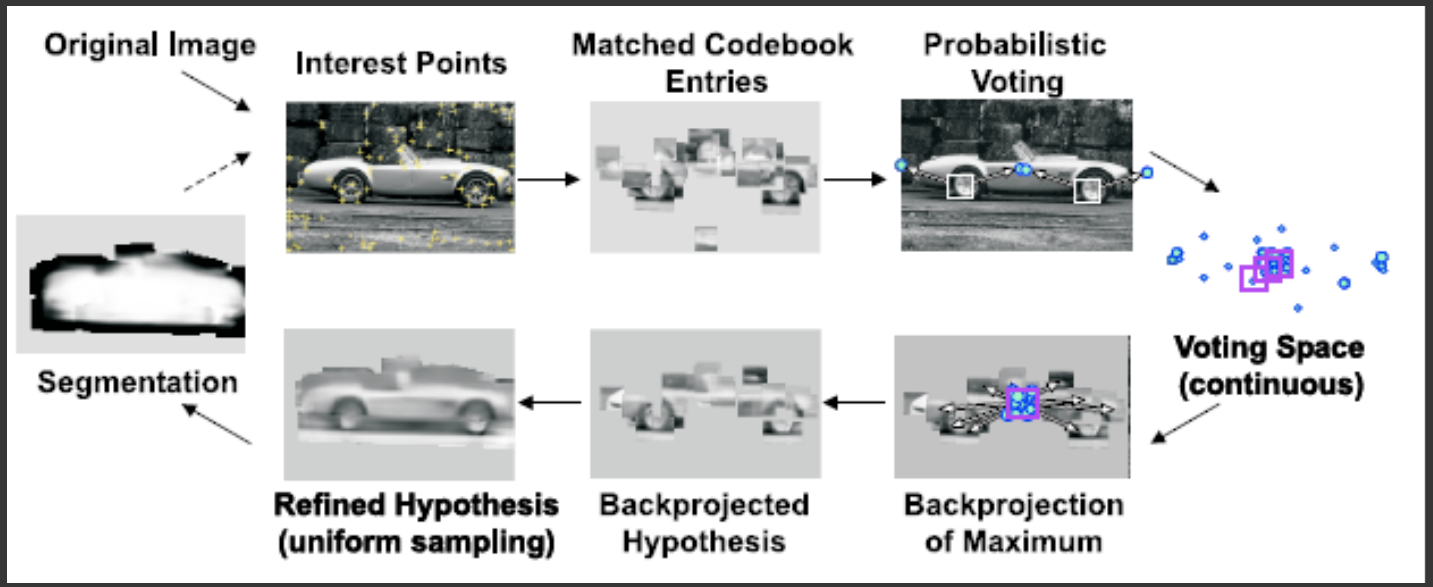
Liebe et al CVPR05
Roth et al CVPR04

- Look for limb-like segments

Mori et al, CVPR04
Ren et al, ICCV05
Mori ICCV05

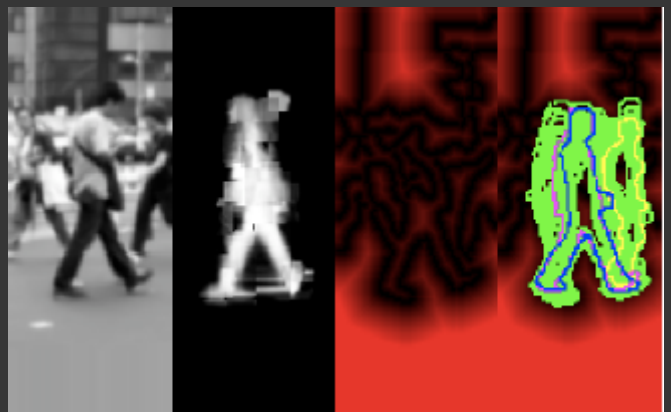
Hypothesize parts and test

Implicit shape model



Liebe et al BMV03
Liebe et al CVPR05

1)Use parts to hypothesize person location
2)Test segmentation with a chamfer score



A quick look back: how do we find person-pixels?

(1) Pixel-based bg/fg labelling

-bg subtraction or fg enhancement

(2) Scanning window pedestrian classifiers

(3) XYT-based people detectors

(4) Top-down models

-exemplars, or sampling-based pose estimation

(5) Bottom-up parts-based models

-stitch parts by dynamic programming

Recall: why is data association is hard?

variation in appearance

variation in pose & aspect

occlusion & clutter

bottom-up
part detection

Recall: why is data association is hard?

variation in appearance

variation in pose & aspect

still hard

occlusion & clutter

bottom-up
part detection

Recall: why is data association is hard?

variation in appearance

variation in pose & aspect

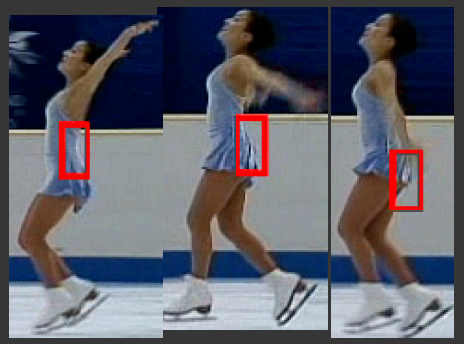
but we don't need intra-class invariance

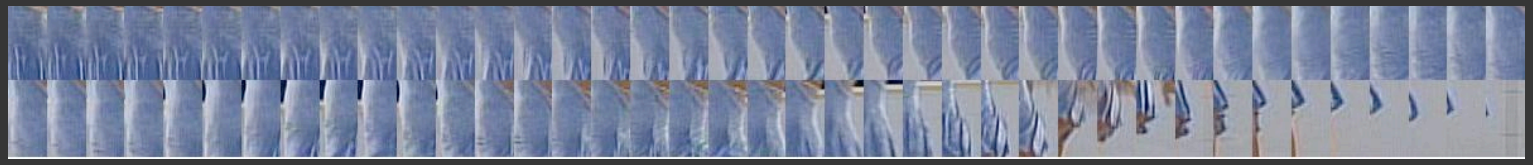
bottom-up part detection

occlusion & clutter

Try updating person-specific appearance

After hand-initializing…

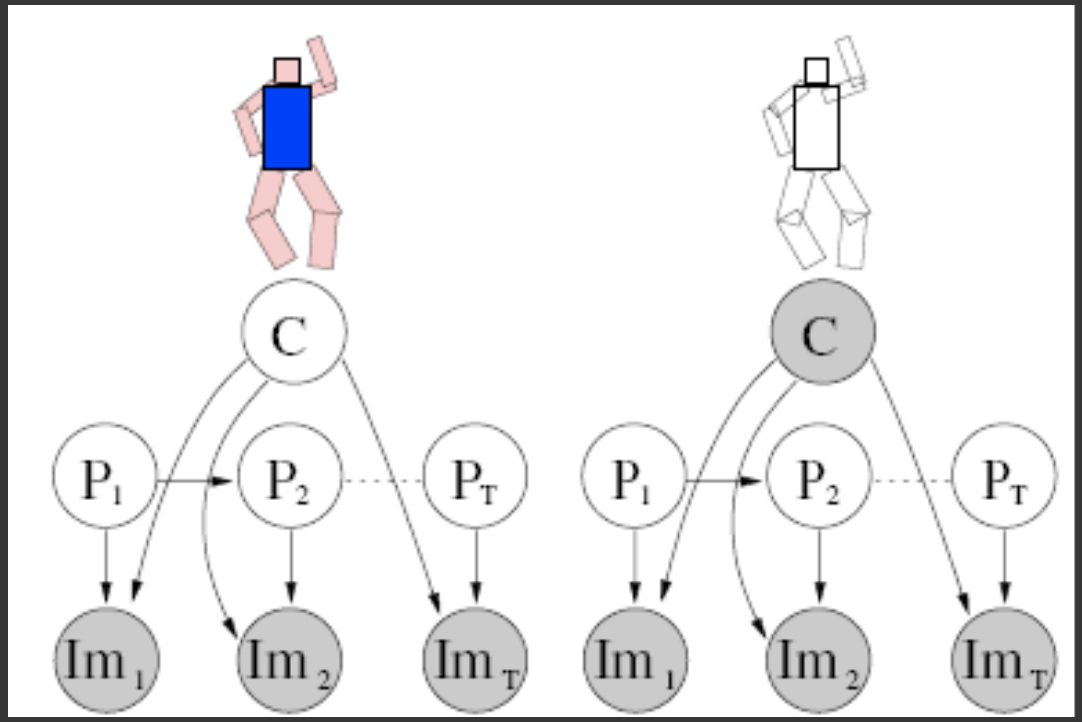




"Telephone Game" can causes drift

Can fix with a accurate data association (i.e. bg subtraction)

Model-based Tracking



If we know model *a priori* => regular Markov model

But model must necessarily be detuned

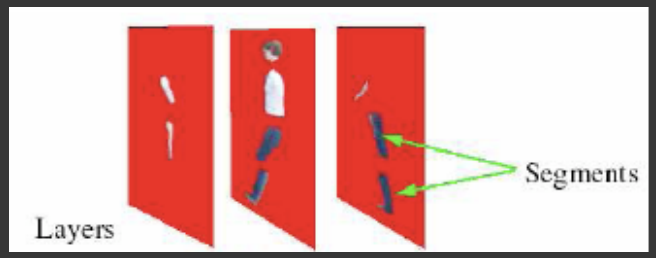
We want to learn template on-the-fly

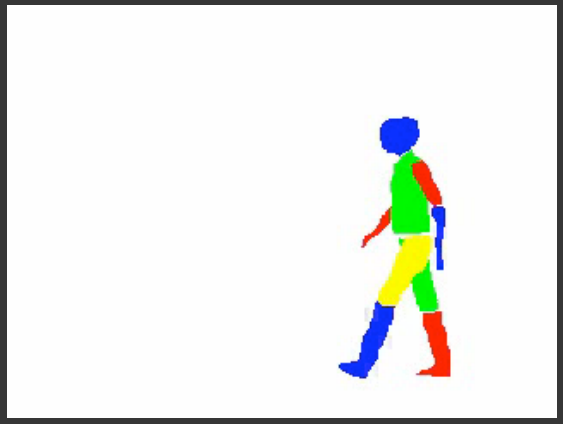
Building models on-the-fly by EM

Jojic & Frey, CVPR01

Learn templates, alpha masks, and depth ordering

Building models by EM (cont'd)



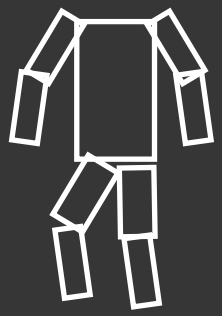


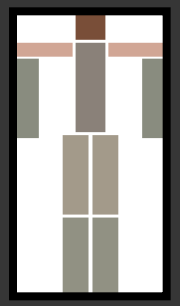
Kumar et al, ICCV05

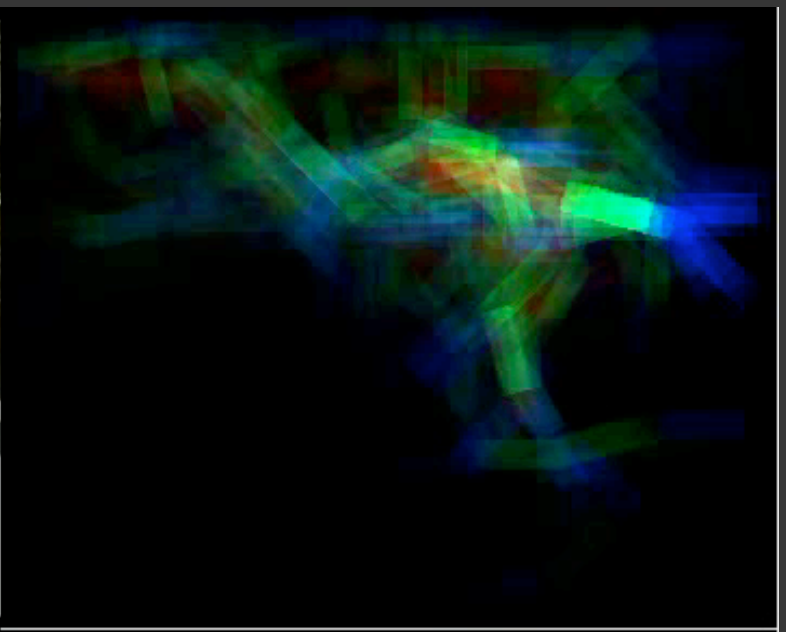
Add low-level segmentation cue to the model

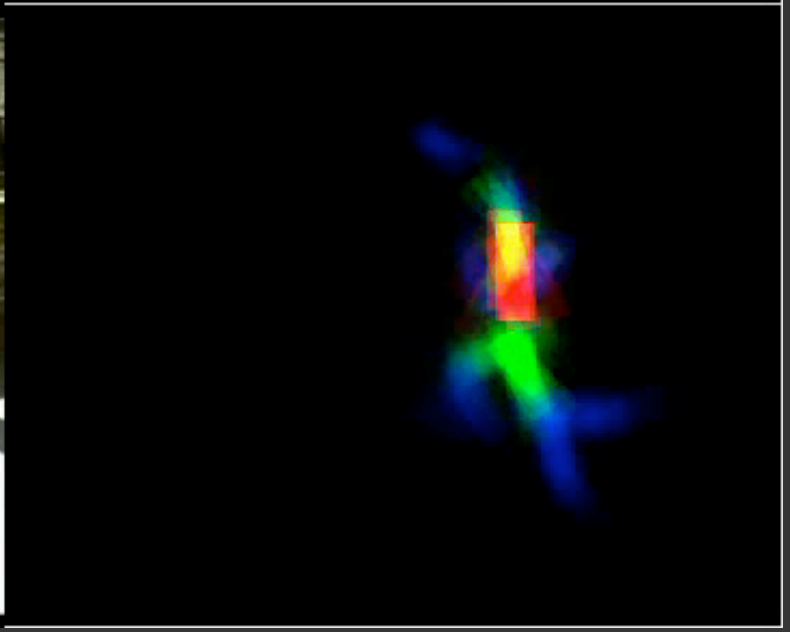
Add temporal illumination variable

Are learned models better?

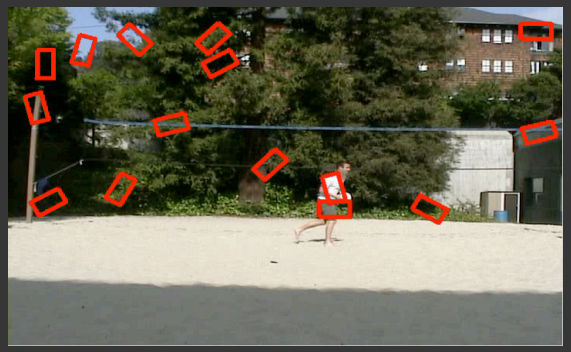




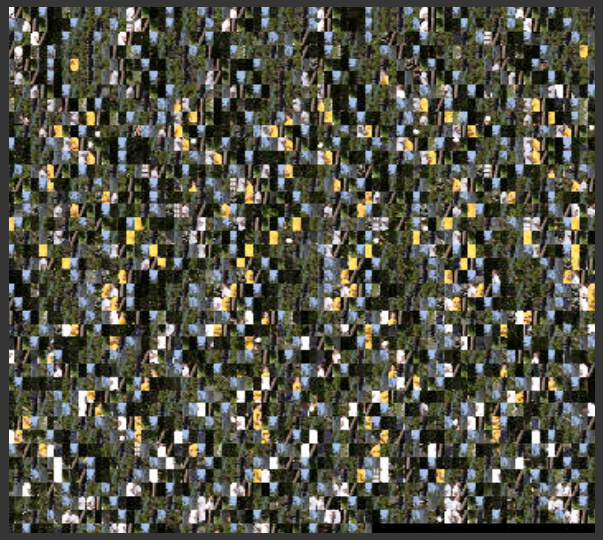




Build models by clustering candidate parts



detected torsos

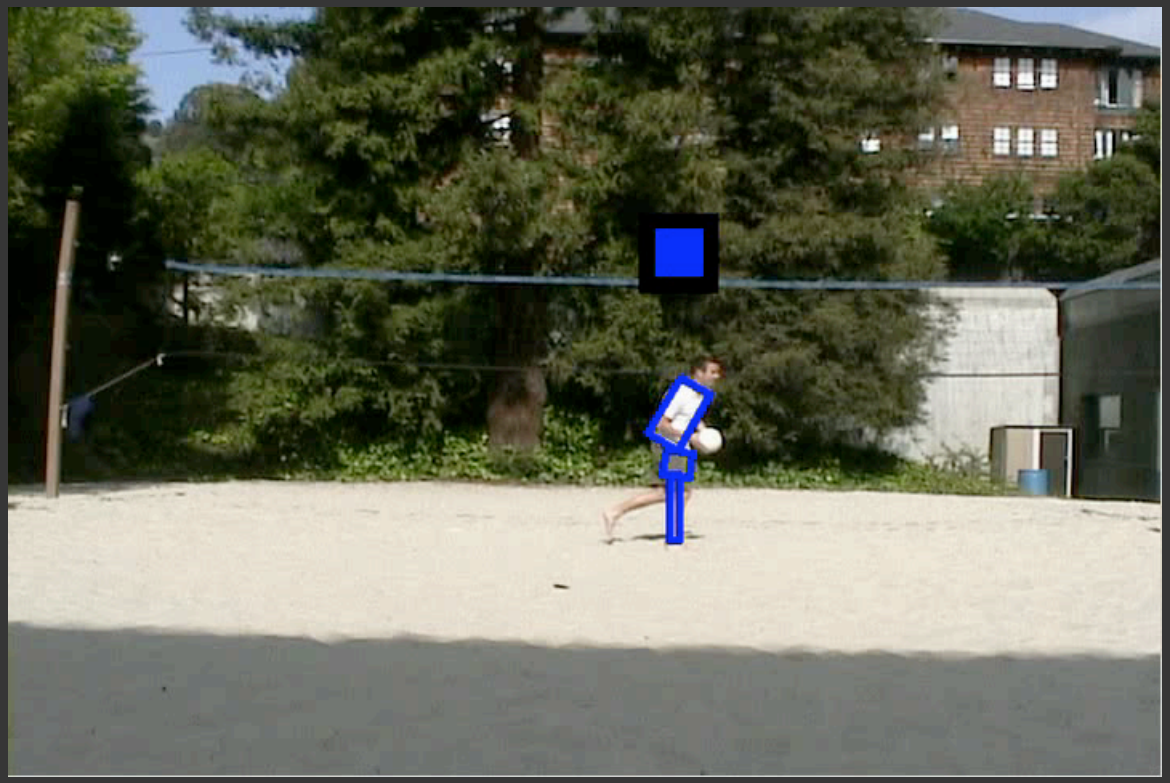


bag of detected torso patches

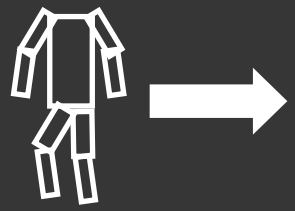
clustered detections

keep ones that don't move

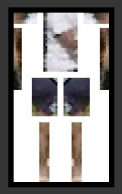
Track multiple people by model-building + detection



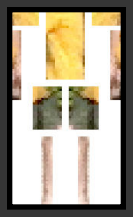
Ramanan & Forsyth CVPR03



person detector



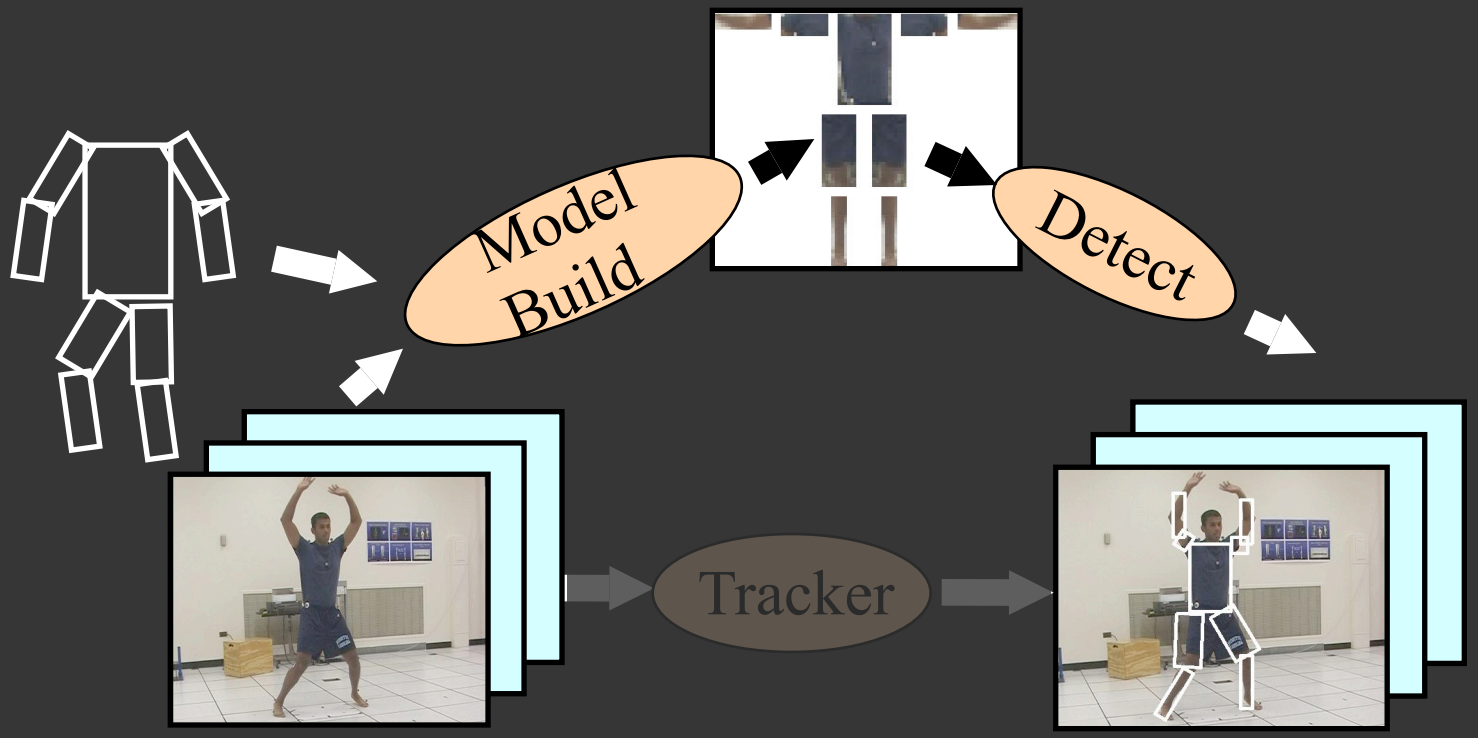
Deva detector



Bryan detector

John detector

How can we build models reliably?



Look for easy frames!

Which frames are easy?

People take on a variety of poses, aspects, scales

self-occlusion

rare pose

motion blur

non-distinctive pose

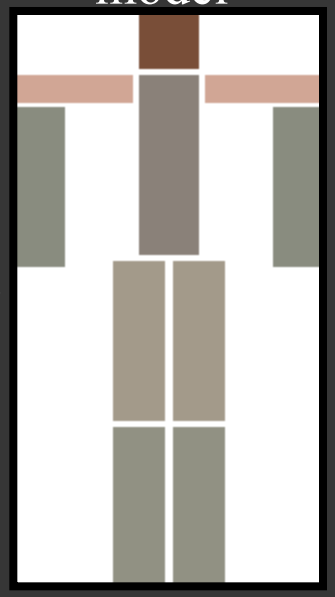
too small

just right
detect this

(Pick you're favorite)
window-based pedestrian classifier
XYT template
top-down exemplars

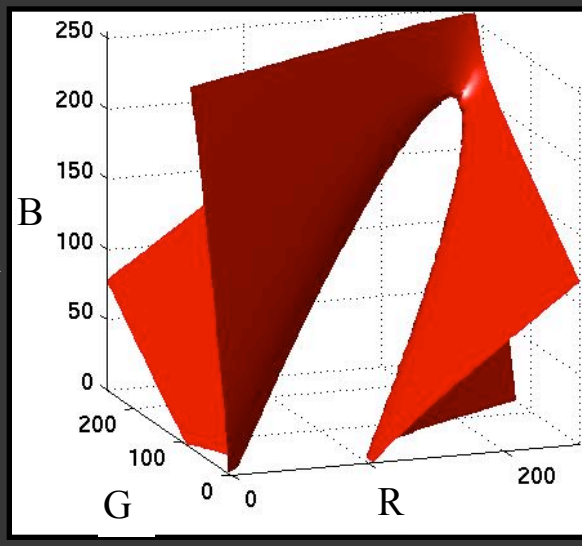
Build model

build model



limb pixel masks

learn limb classifier



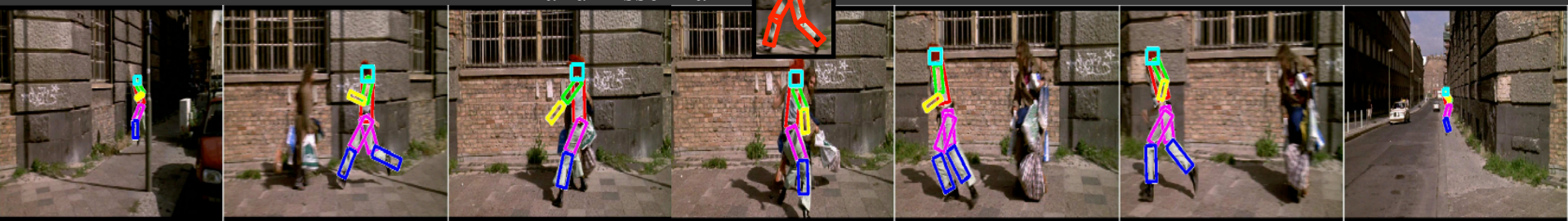
Sequence-specific
discriminative features for tracking

Collins et al. CVPR03
Avidan CVPR05
Ramanan et al CVPR05

`Run Lola Run'

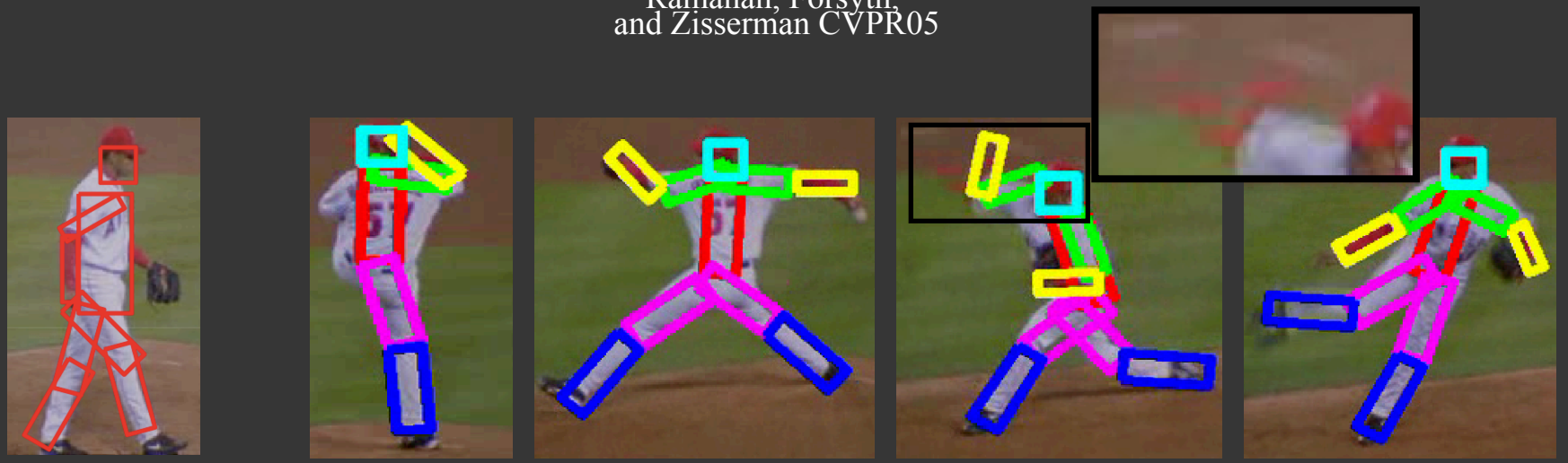
Ramanan, Forsyth, and Zisserman

CVPR05



How likely is a ‘typical’ pose?

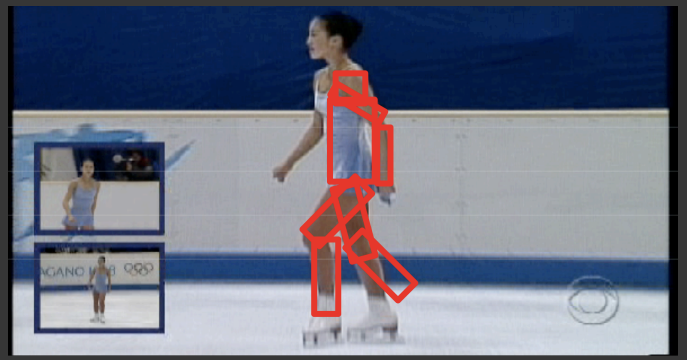
Ramanan, Forsyth, and Zisserman CVPR05



motion blur &interlacing

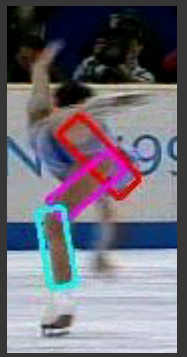
Track long footage (7600 frames)

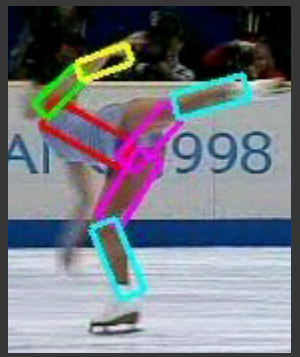
Ramanan, Forsyth, and Zisserman CVPR05



extreme pose

motion blur





fast movement

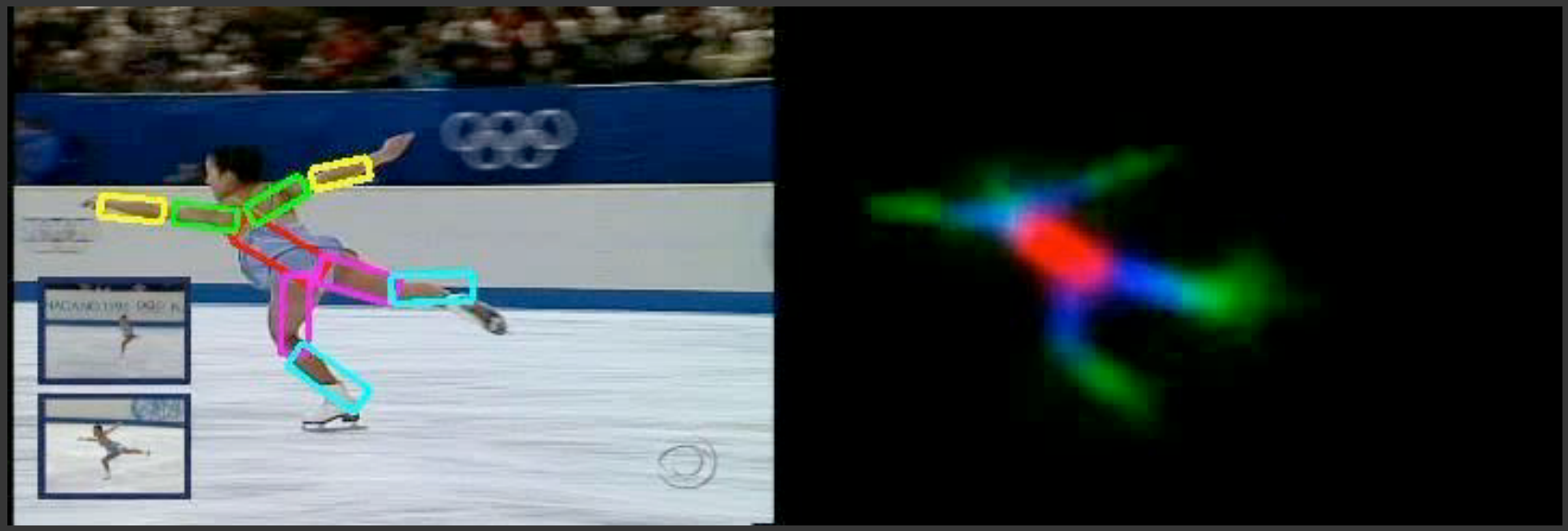
Olympic woes

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



Unexpected/unlikely motions often very important

Motion and pose priors maybe misleading

Leaves us with data association(?)

Evaluation-data

- Synthetic (rendered 3D mocap data)
  - likely too unrealistic
- Coarse body labels (PETS data)
- Video + Mocap ground truth
  - Brown dataset

    (registered video + mocap)

http://www.cs.brown.edu/~ls/Software/index.html

  - CMU dataset

    (not registered)

http://mocap.cs.cmu.edu/

- Hard to get lots of variety-rich sequences
  - Probably requires manual labeling

Evaluation-metrics

- Detection rates of body parts
- Localization error of joints (in image or 3D)
- Error in joint angles

Not clear how these translate to
a specific application

Data association: A look back

How do we pull the person out from bg?

Probably need to use image data (as opposed to dynamics)

Low level image features seem important

Learn for each sequence?

Discriminative features seem to address bg clutter

How do we detect people in images?

Hard problem, but worth solving!

Bottom-up seems to address pose variation

Better use of XYT volume?