Human Detection and Tracking

CS 543 - D.A. Forsyth

Why is human motion important?

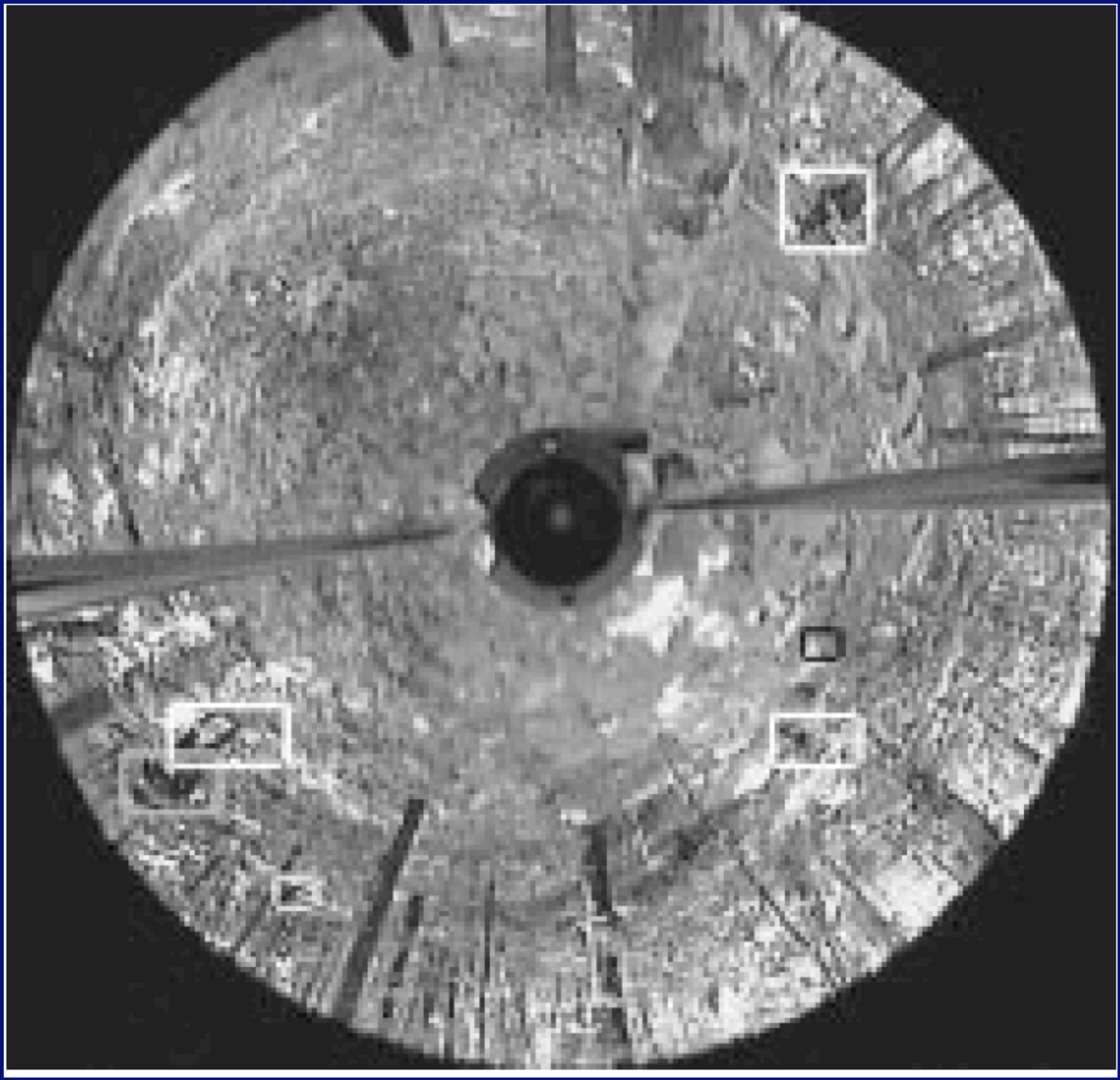
- Surveillance
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- Synthesis
  - games; movies;
- Biomechanics
  - spot diseases; learn new facts
- People are interesting
  - movies; news

Core Problems

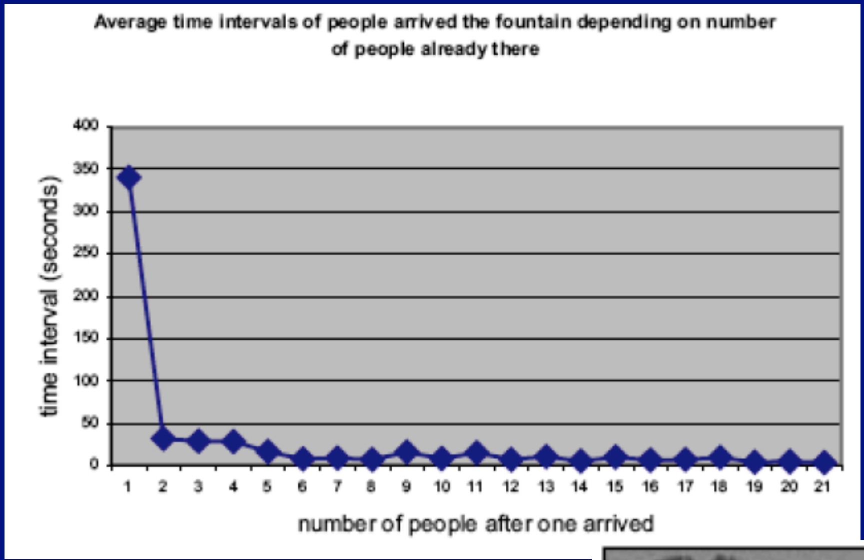
- It is not known what needs to be known
  - or, what should we extract from video to do what task?
- It is hard to find people
  - Appearance
  - Aspect
- It is hard to track people in detail
  - Small parts that move fast and unpredictably
- It is hard to describe what they are doing
  - Behaviour composes
    - sometimes in complex ways
  - A canonical vocabulary is not known

Where you are can tell what you are doing

Intille et al 95, 97

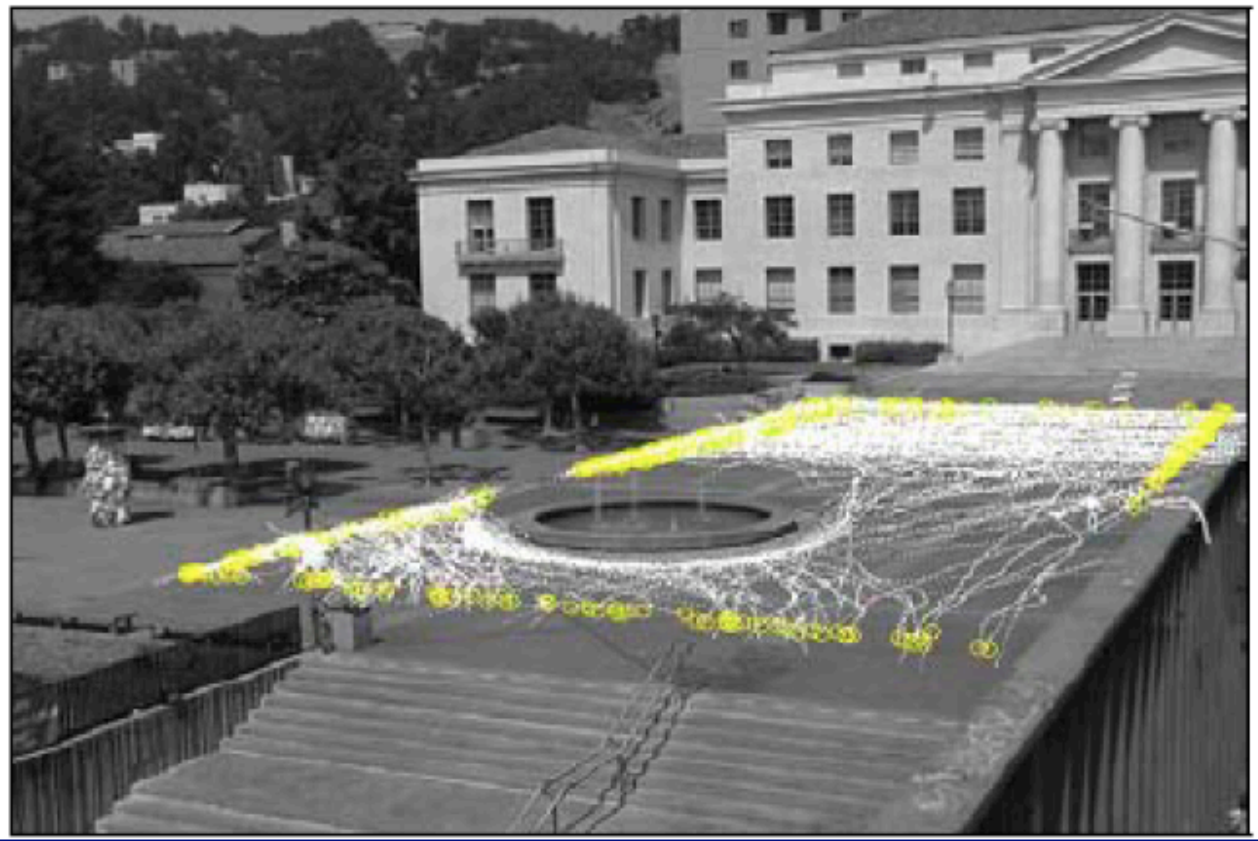


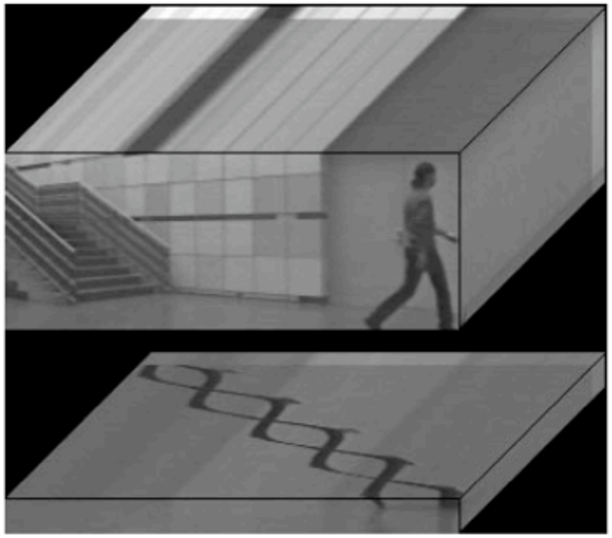
And can suggest you are doing something you shouldn't be
Boult 2001



Curious phenomena in public spaces

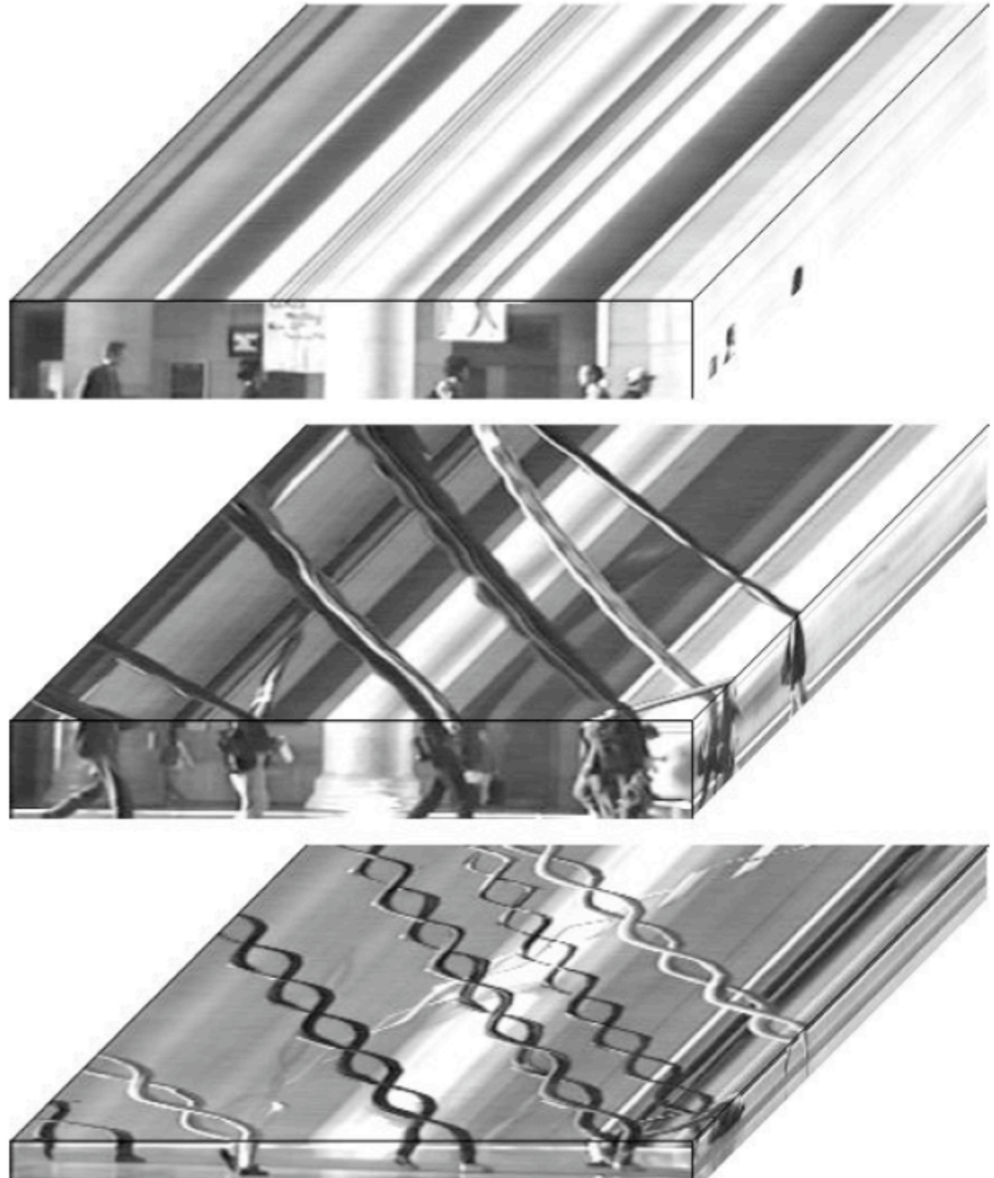
Yan+Forsyth, 04

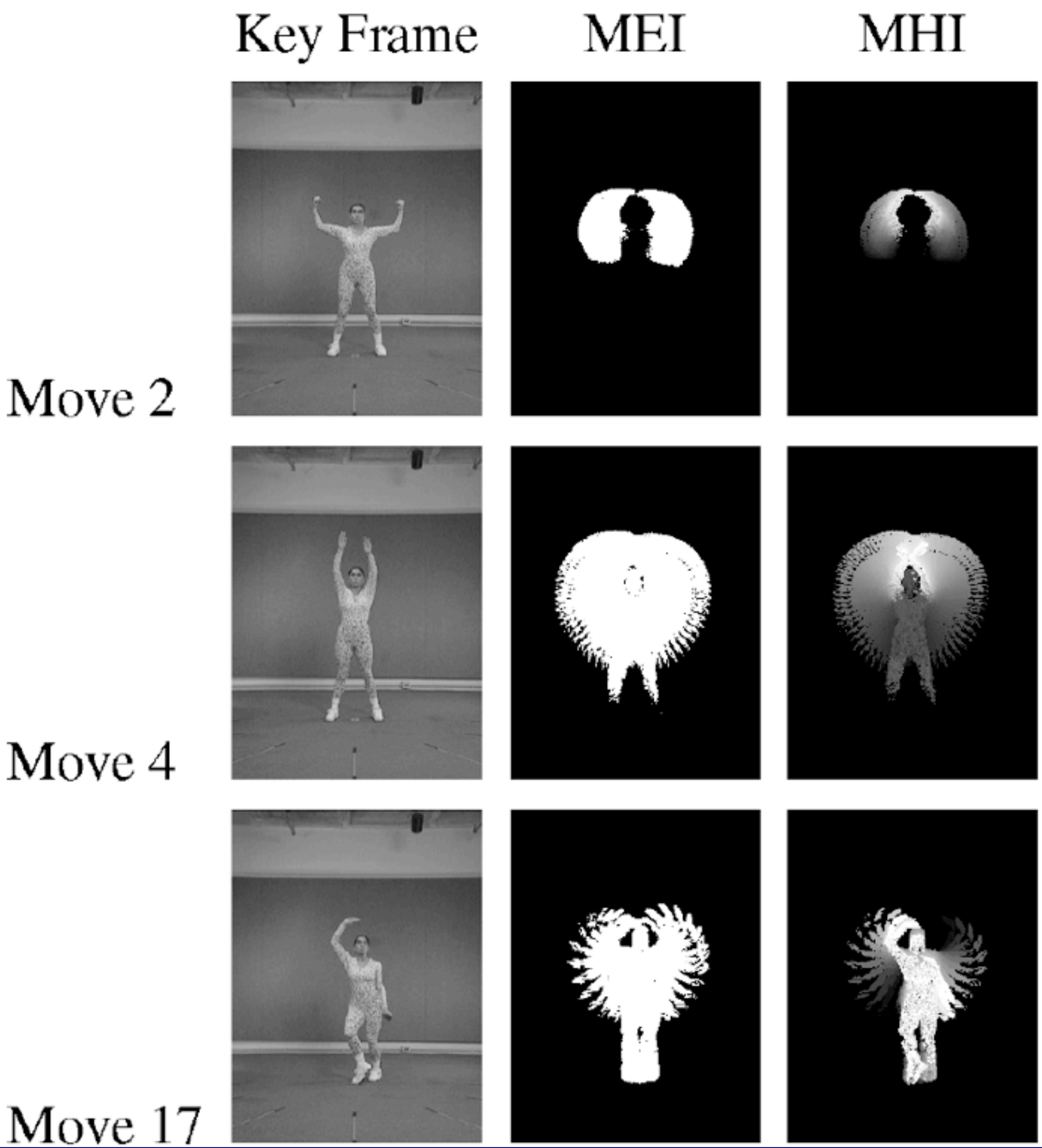




Niyogi Adelson 94

Particular activities often have characteristic appearance patterns. Braids appear at the legs of a walker.

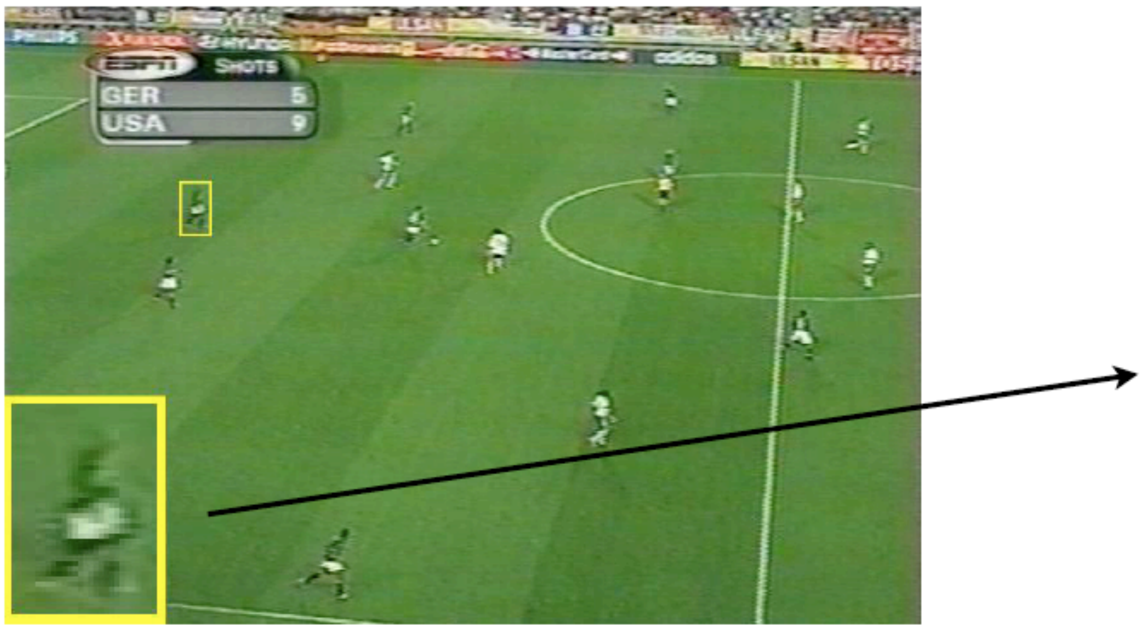


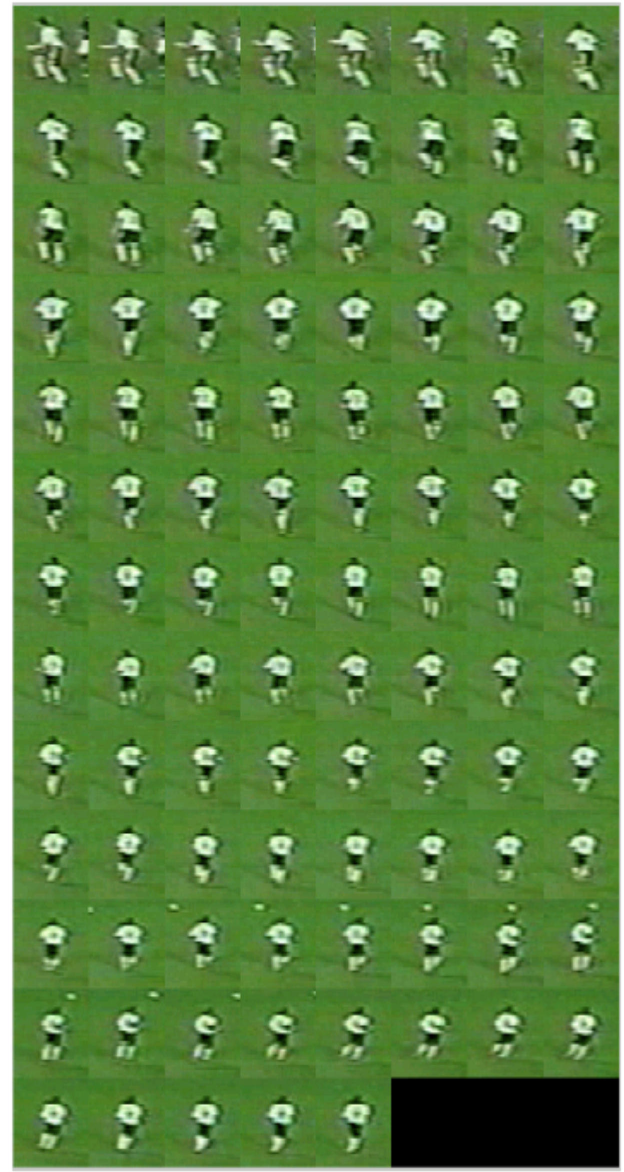


Bobick + Davis, 97

Efros et al, 03

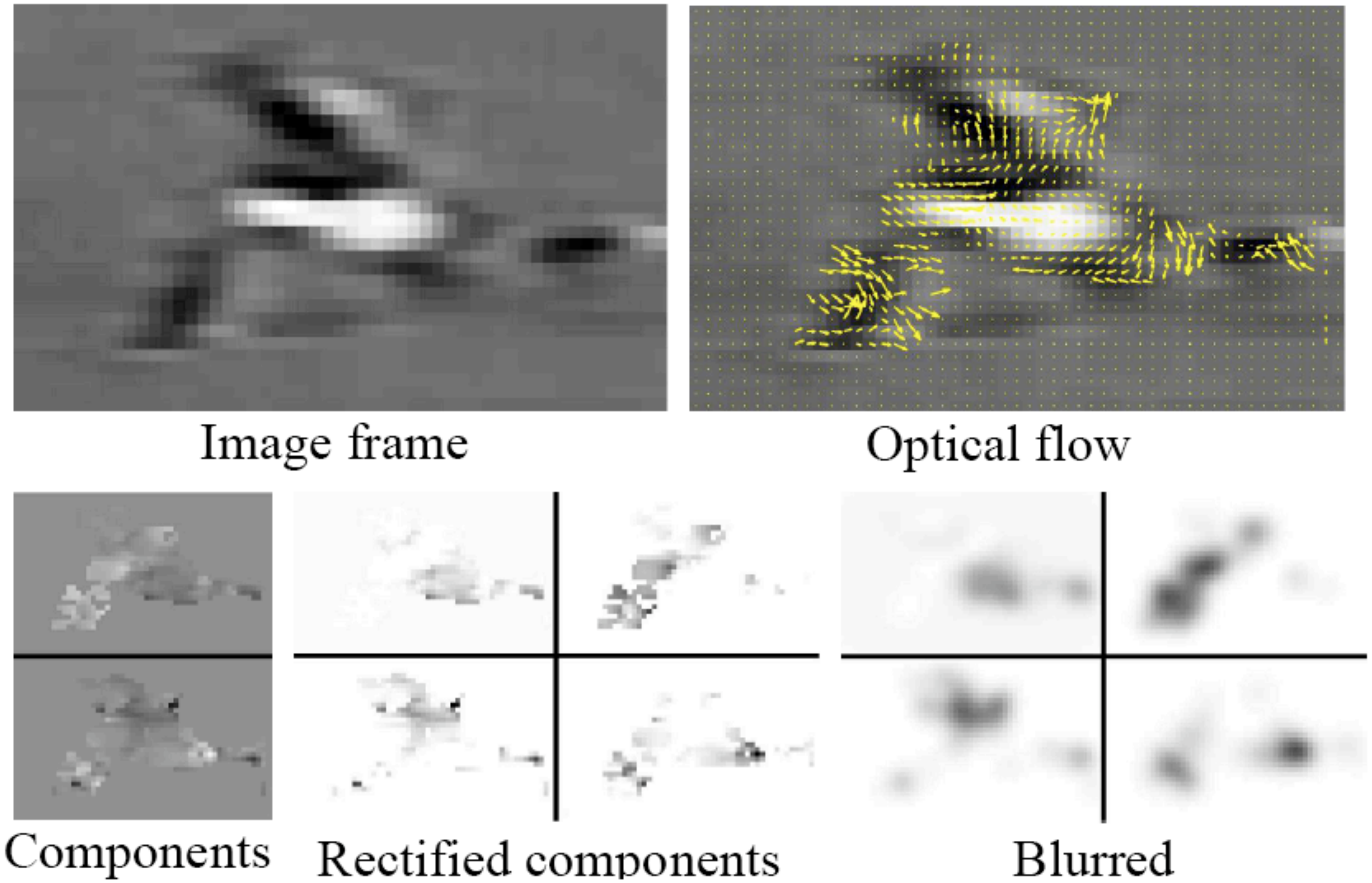
Motion is a powerful cue at low resolution





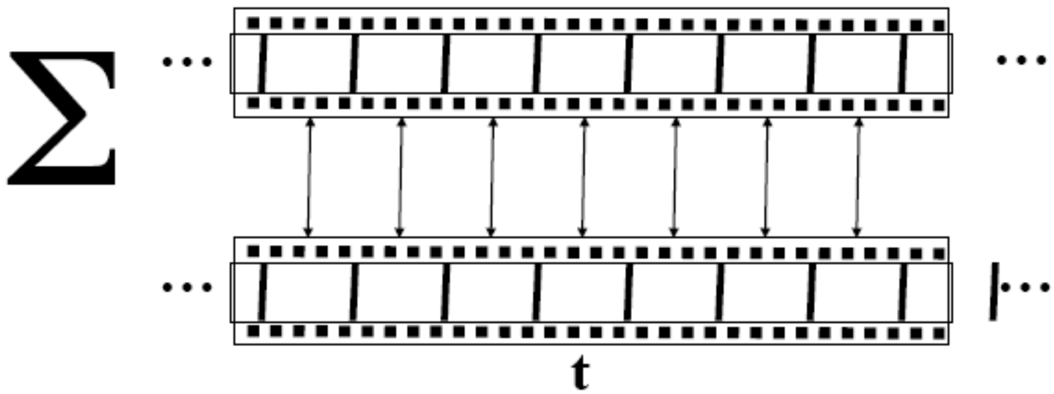
Efros et al 03

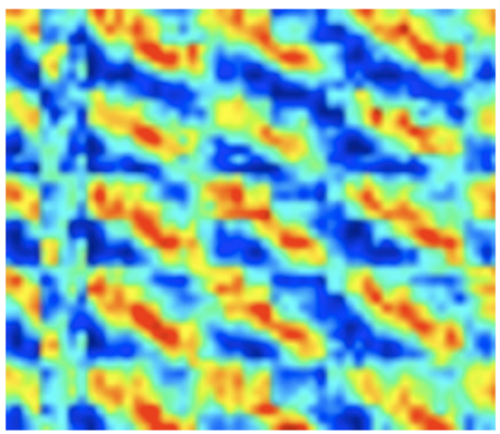
Motion Descriptor

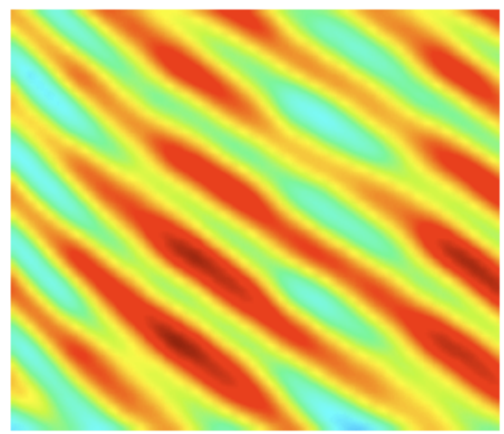


Efros et al 03

Comparing motion descriptors

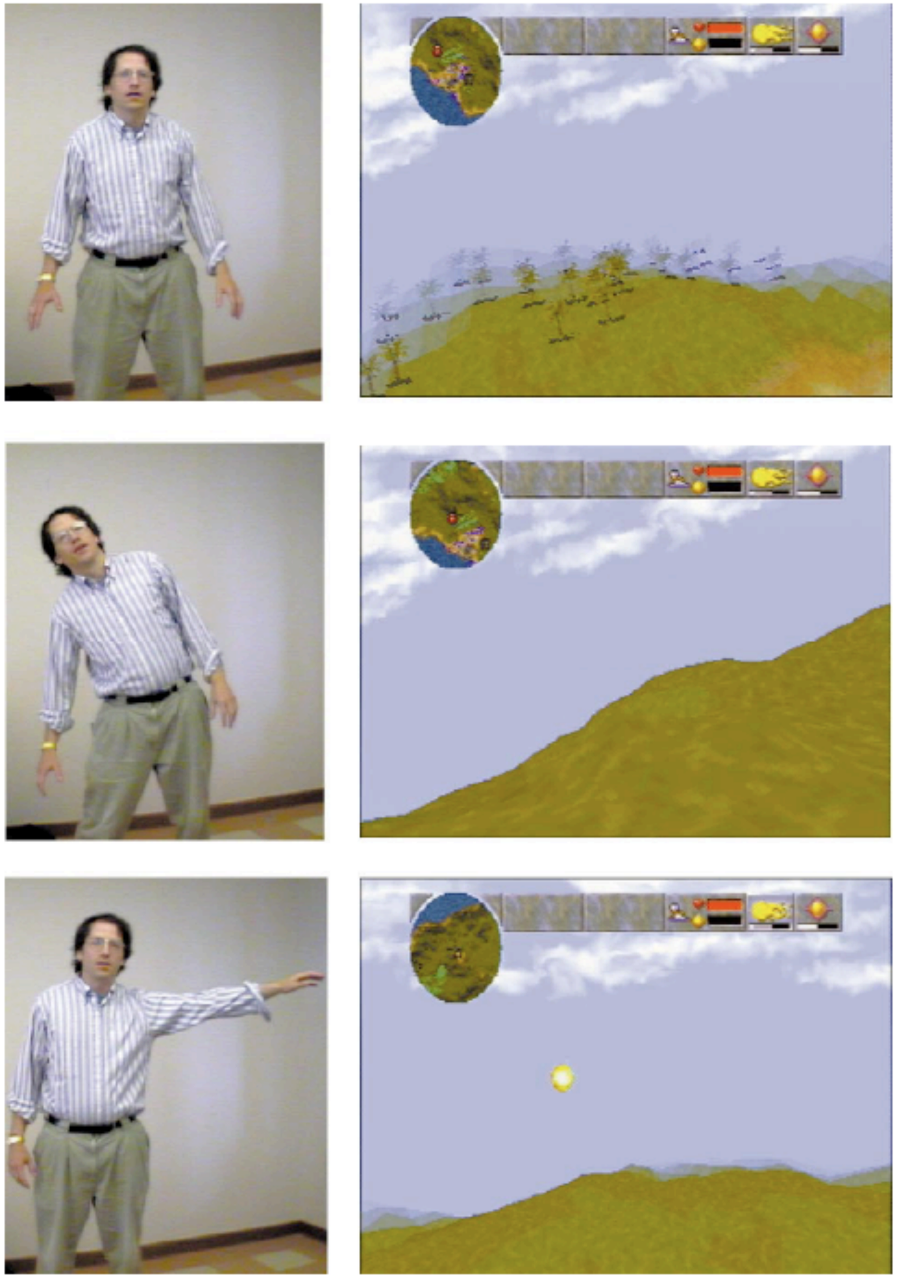



frame-to-frame similarity matrix


motion-to-motion similarity matrix

Efros et al 03

[redacted]



Bill Freeman flies a magic carpet.

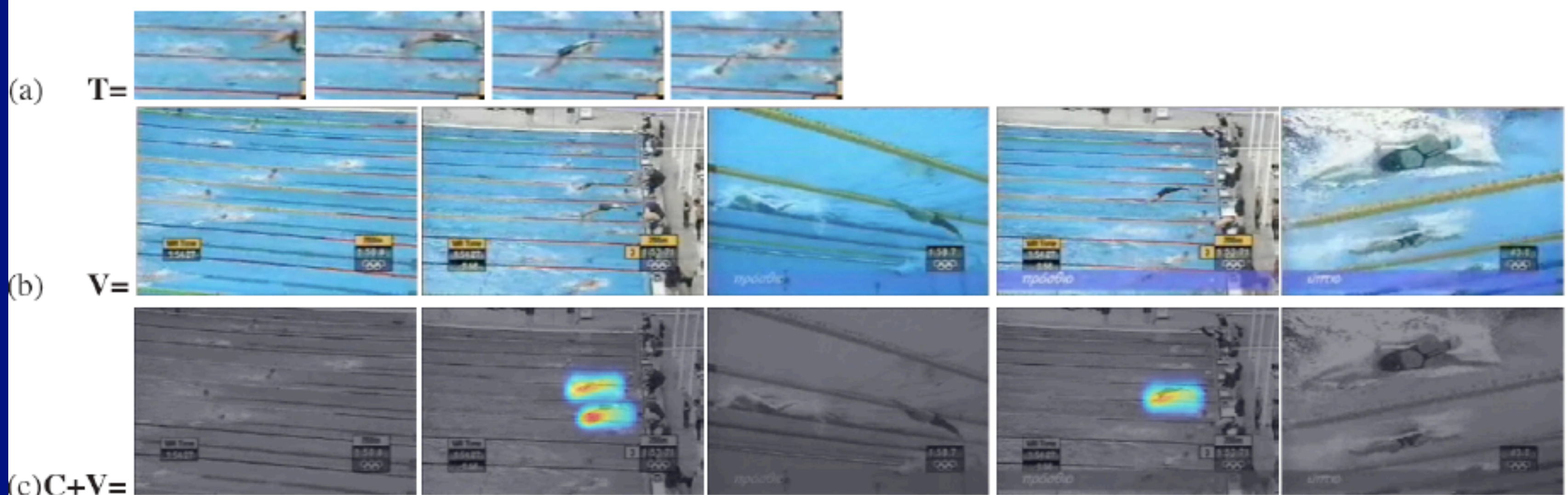
Orientation histograms detect body configuration to control bank, raised arm to fire magic spell.

Freeman et al, 98.

9 An example of a user playing a Decathlon event, the javelin throw. The computer's timing of the set and release for the javelin is based on when the integrated downward and upward motion exceeds predetermined thresholds.

Motion fields set javelin timing
Freeman et al 98

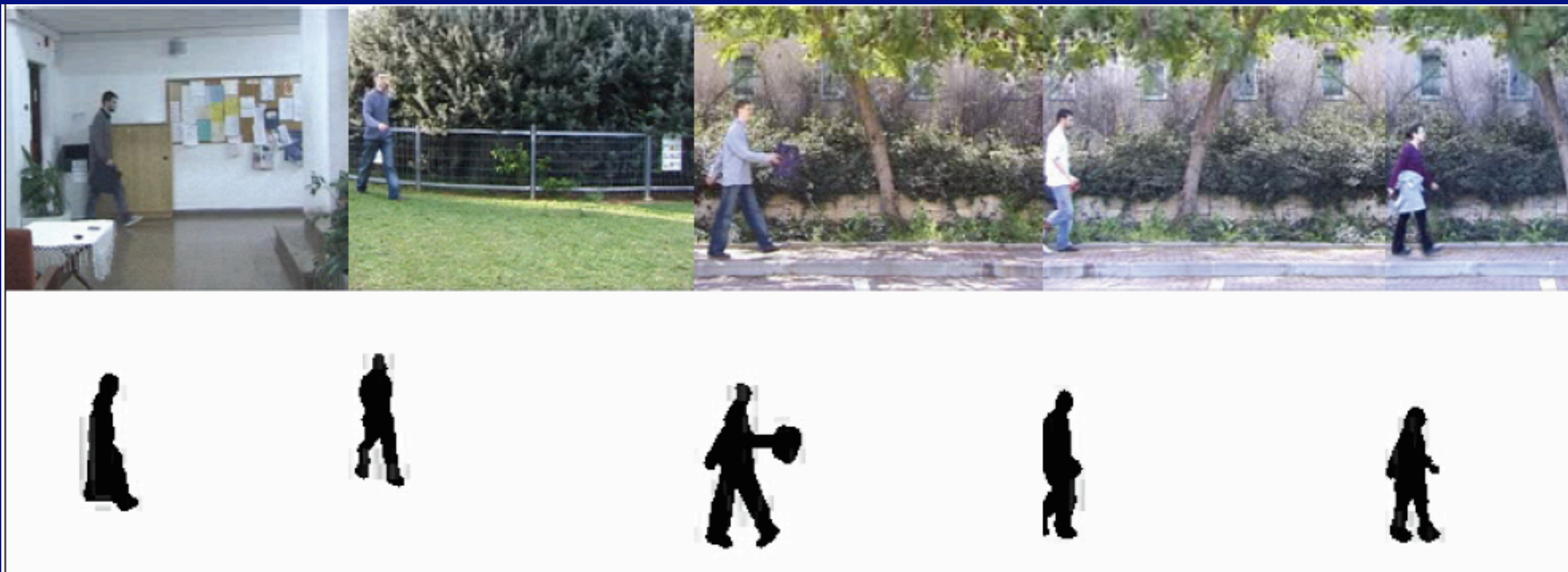
Sony's eyetoy estimates motion fields,
links these to game inputs.
Huge hit in EU, well received in US



Correlation-like matching can reveal motion matches to queries
Schechtman Irani 05

Spatio-temporal volume is important

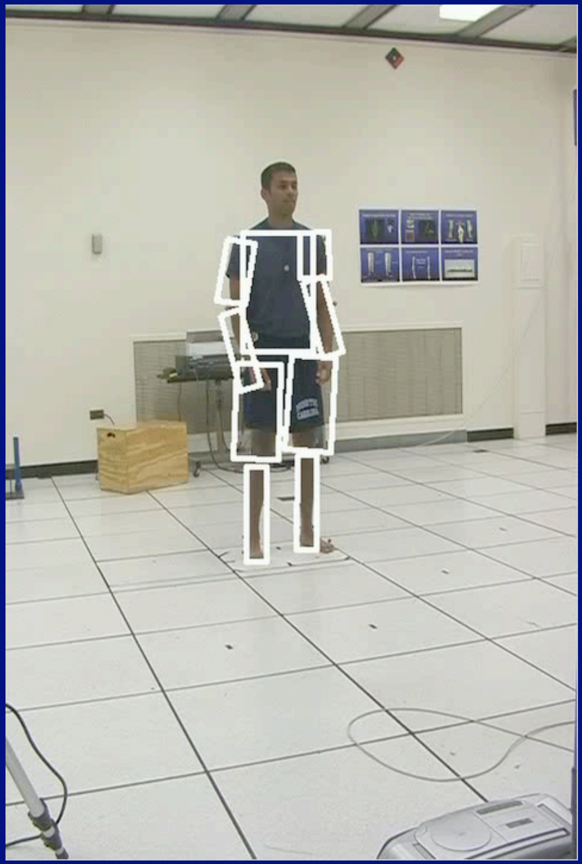
Blank et al 05



Extract silhouettes
Smooth to get volume
Compute moment representation on s-t volume referred to body
Match

Blank et al 05

Motion transduction



Pictorial structures

- For models with the right form, one can test "everything"
  - model is a set of cylindrical segments linked into a tree structure
    - model should be thought of as a 2D template
      - segments are cylinders, so no aspect issue there
      - 3D segment kinematics implicitly encoded in 2D relations
      - easy to build in occlusion
  - putative image segments are quantized
  - => dynamic programming to search all matches
  - What to add next? (DP deals with this)
  - Pruning? (Irrelevant)
  - Can one stop?
    - (Use a mixture of tree models, with missing segments marginalized out)
  - Known segment colour - Felzenszwalb-Huttenlocher 00
  - Learned models of colour, layout, texture - Ramanan Forsyth 03, 04

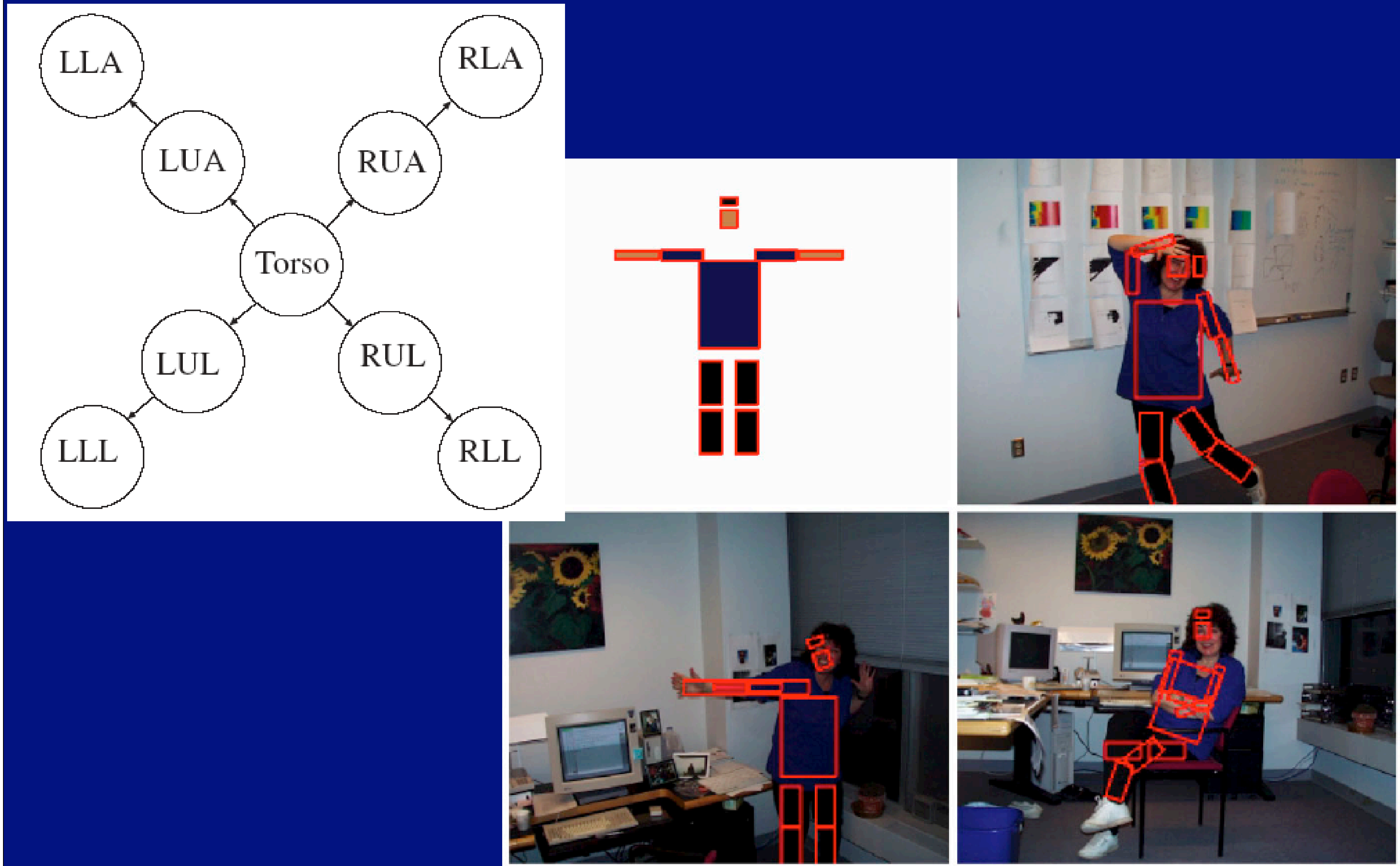


Figure from “Efficient Matching of Pictorial Structures,”
P. Felzenszwalb and D.P. Huttenlocher, Proc. Computer Vision and Pattern Recognition
2000, c 2000, IEEE as used in Forsyth+Ponce, pp 636, 640

Human tracking options

- People as blobs (+appearance)
  - Grimson et al 98; Stauffer et al 00; Haritaoglu et al 98, 00; Okuma et al 04
- People as motion fields
  - Bregler 97;Boyd+Little 98
- People as blobs+motion fields
  - Efros et al 03
- Kinematics
  - Hogg 83; Rohr 93; Deutscher et al 00; Toyama+Blake 02; SidenbladhBlackFleet 00; JuBlackYacoob 96; Song Perona 00; etc

Why is kinematic tracking hard?

- It's hard to detect people
  - until recently, all human trackers were manually started
- People move fast, and can move unpredictably
  - dynamics gives limited constraint on future configuration
  - appearance changes over time (shading, aspect, etc)
- Some body parts are small and tend to have poor contrast
  - particularly difficult to track
    - lower arms (small, fast, look like other things);
    - upper arms (poor contrast)

variation in pose & aspect

self-occlusion & clutter

variation in appearance

Strategies

- Markov model of (appearance, configuration)
  - 3D Models
    - compare to image
      - variations in dynamical constraints, complexity of inference
        - Hogg 83; Rohr 93; Bregler+Malik 98; Sidenbladh Black Fleet 00; Deutscher Blake Reid 00
  - 2D model
    - Ju Black Yacoob 96; Cham + Rehg 99
- Not quite Markov, but
  - templates encode appearance, then assume markovian dynamics
    - Toyama+Blake 02
- Track by detection
  - Song+Perona (motion) 00; Ioffe+Forsyth (appearance) 01; Mori+Malik (appearance) 02

Opportunistic detection

People take on a variety of poses, aspects, scales

self-occlusion


rare pose


motion blur

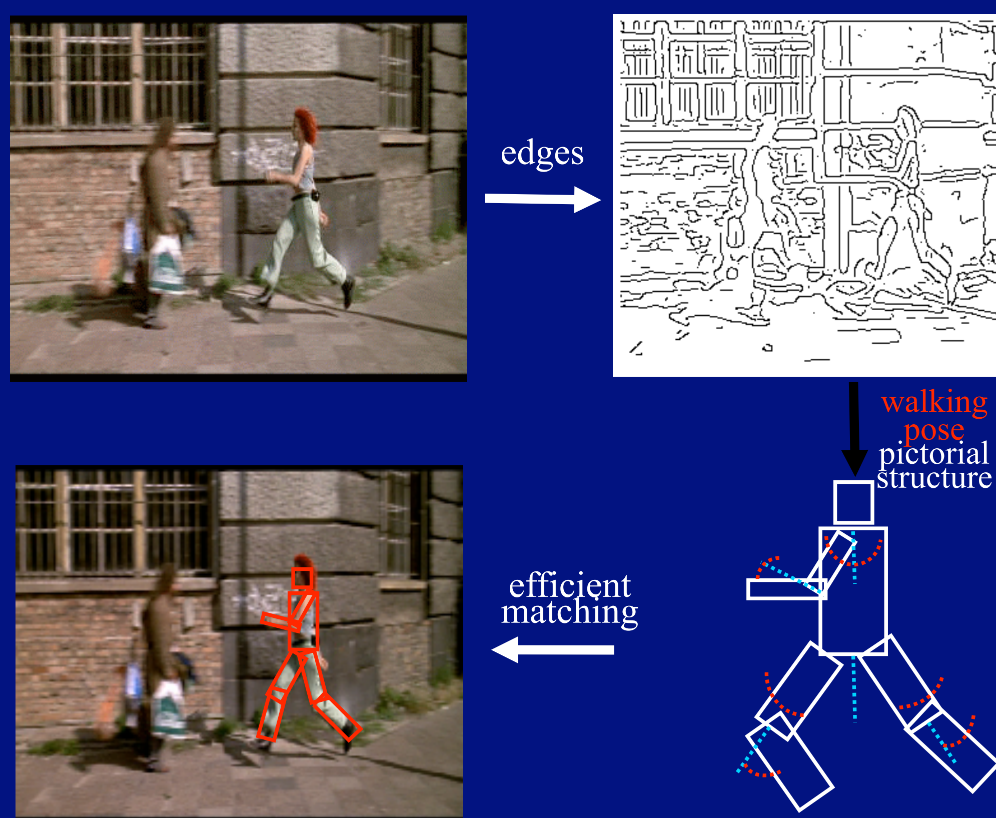

non-distinctive pose


too small


just right
detect this

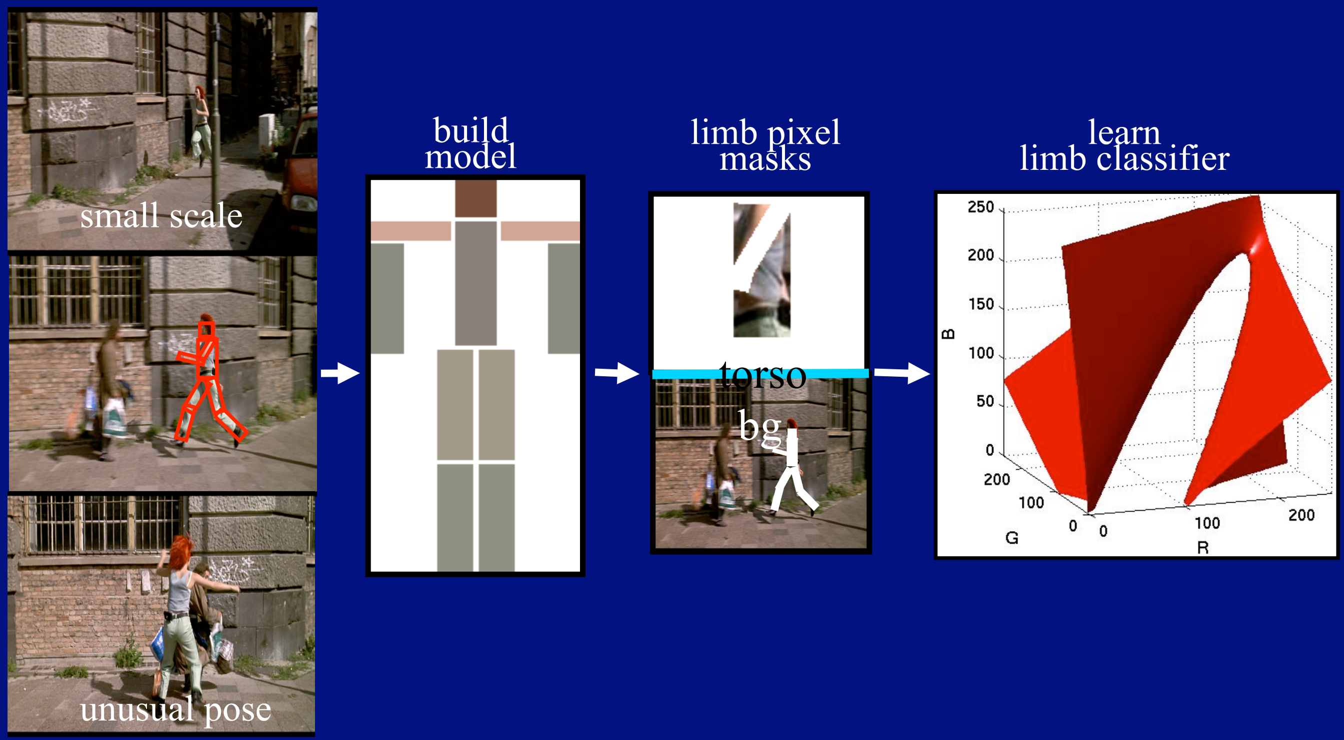
Ramanan, Forsyth and Zisserman CVPR05

Stylized pose detector



Ramanan, Forsyth and Zisserman CVPR05

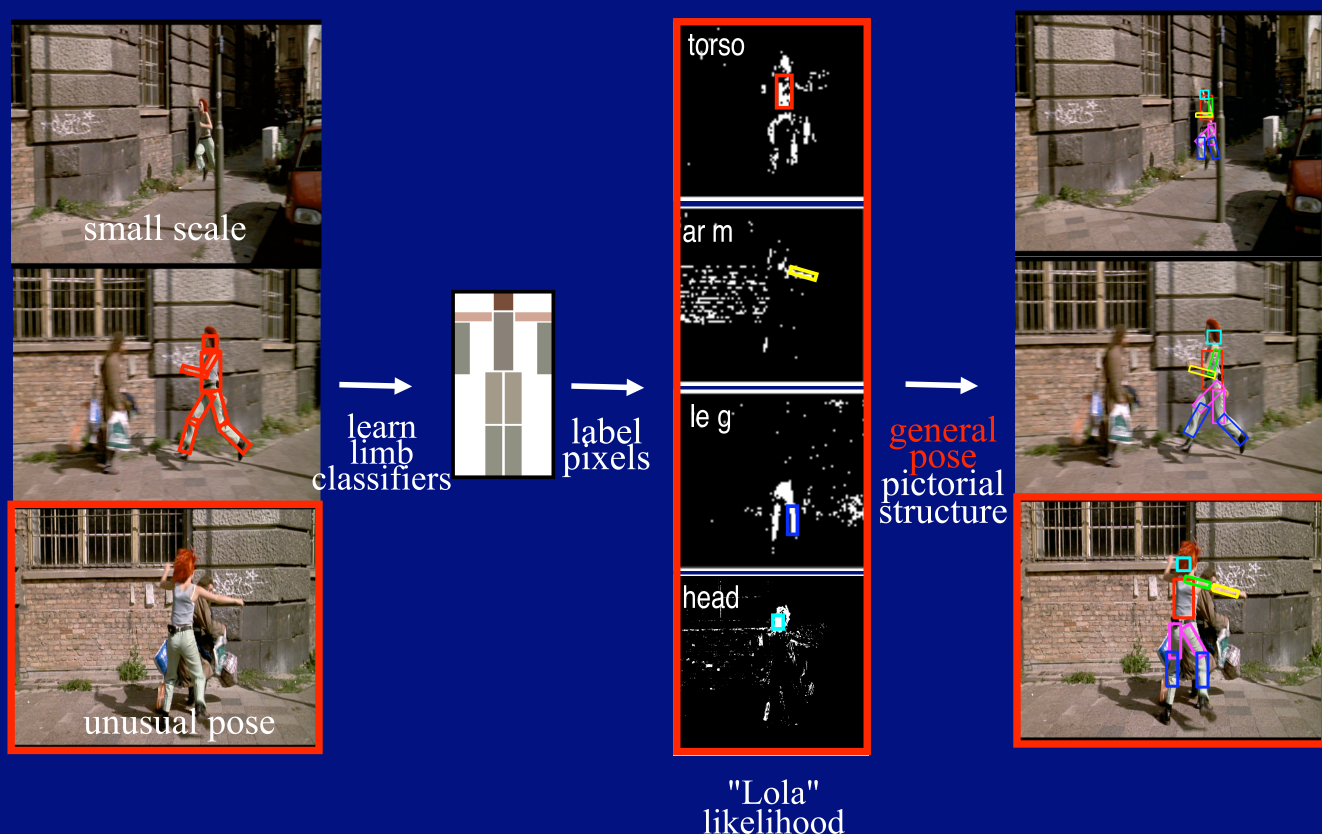
Model building



Ramanan, Forsyth and Zisserman CVPR05

Ramanan, Forsyth and Zisserman CVPR05

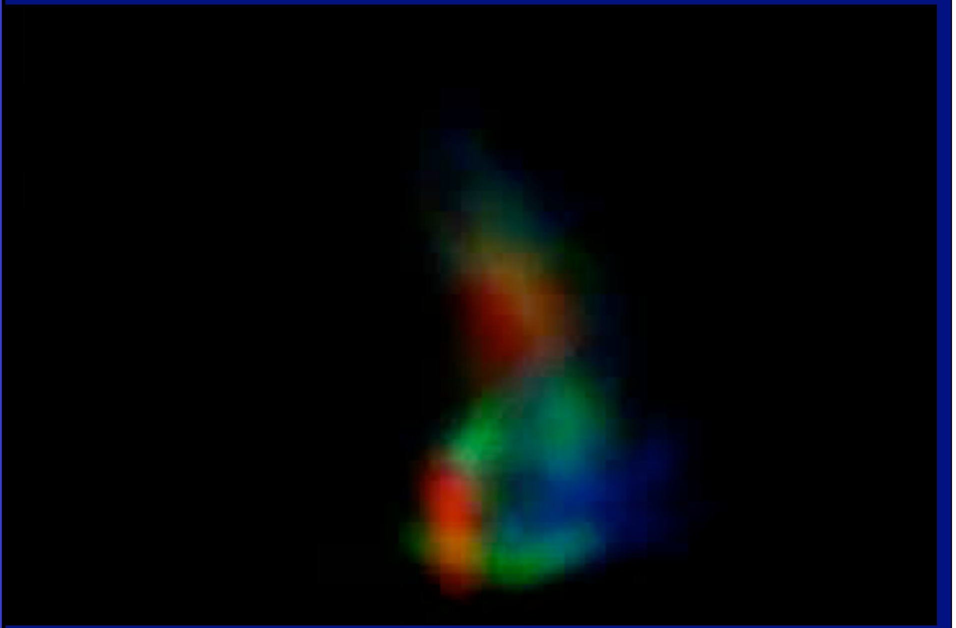
Build and detect models



Ramanan, Forsyth and Zisserman CVPR05

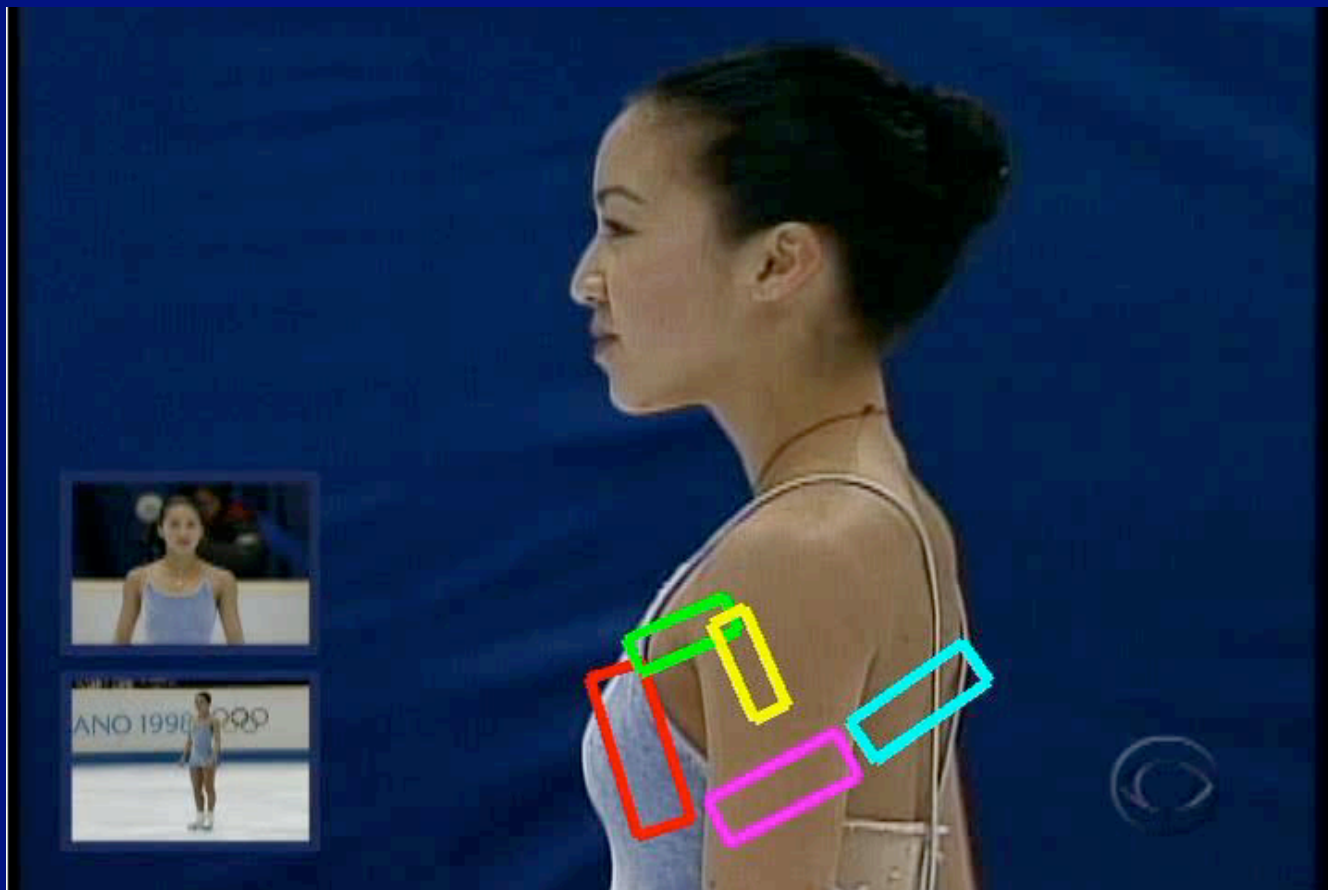
Ramanan, Forsyth and Zisserman CVPR05

Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05

Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05

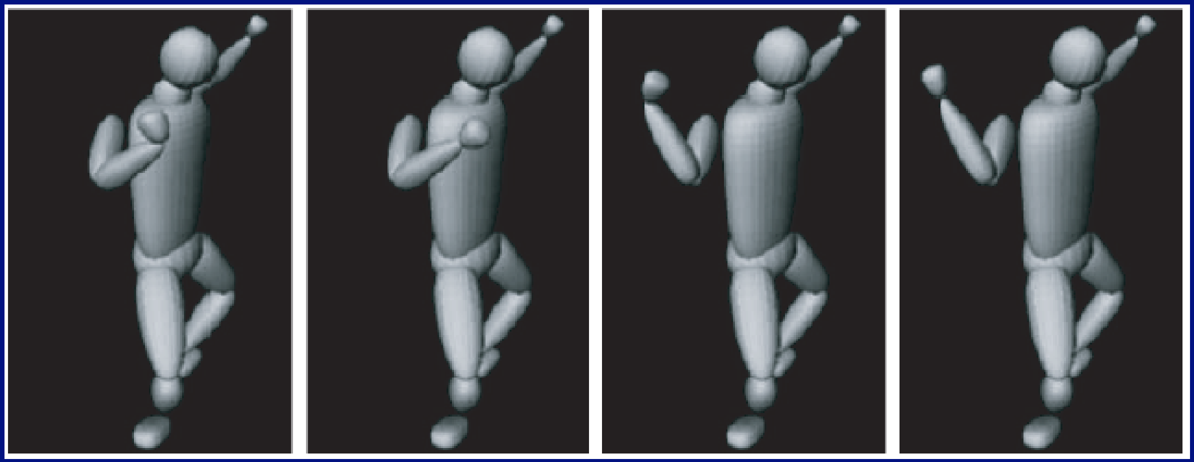
Lifting

- Infer 3D configuration from image configuration
- Useful for
  - view independent activity recognition
  - user interfaces
  - video motion capture

Taylor, 00

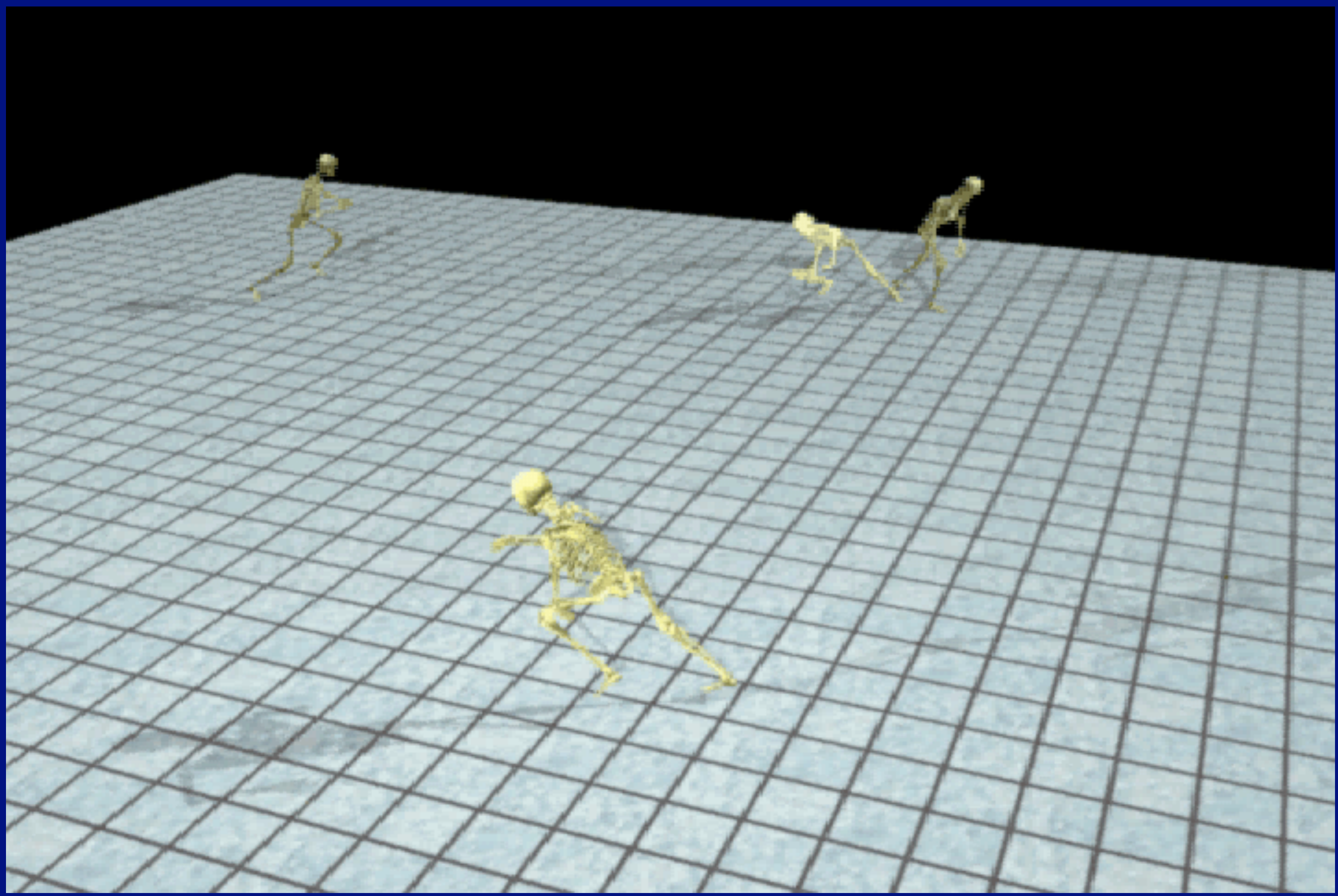
Ambiguity

- Troubled question
  - lifts are ambiguous (Orthography; Sminchicescu+Triggs 03; etc)
  - but ambiguities
    - can be ignored
      - Taylor 00; Barron+Kakadiaris 00
    - can be dodged
      - Ramanan+Forsyth 03; Howe et al 00
- Summary+musings in Forsyth etal 06



Sminchisescu+Triggs, 03

Animating people



Points

- Some properties of motion, illustrated by animation
  - motion composes
    - across time
    - across the body
  - motion can be easy to annotate
    - but good from bad is hard
  - motion clusters well

Motion synthesis

- Methods
  - By animator
  - By combining observations
    - old tradition of move trees; also (Kovar et al 02, Lee et al 02, Arikan +Forsyth 02, Arikan et al 03,Gleicher et al 03)
  - By physical models
    - old tradition; (Witkin+Kass, 88; Witkin+Popovic 99; Funge et al 88; Fang+Pollard 03, 04)
  - By biomechanical models
    - old tradition; (Liu+Popovic 02; Abe et al 04; Wu+Popovic 03; Liu +Popovic 02)
  - By statistical models
    - old tradition (e.g. Ramsey+Silverman 97); Li et al 02; Safanova et al 04; Mataric et al 99; Mataric 00; Jenkins+Mataric 04;

Motion graph

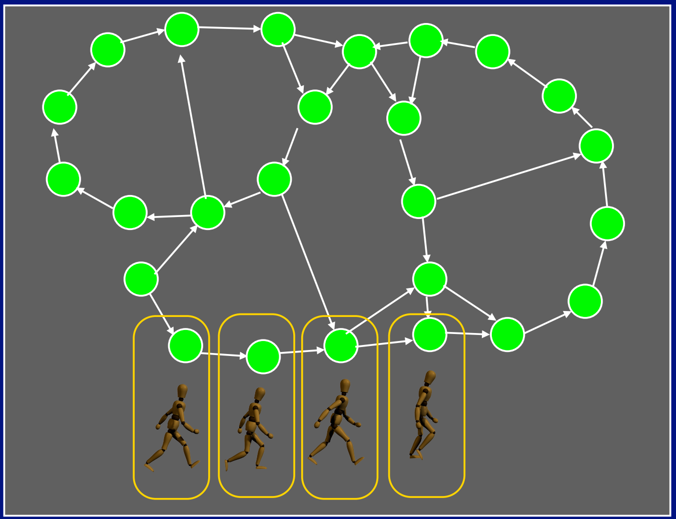
- Take measured frames of motion as nodes
  - from motion capture, given us by our friends
- Edge from frame to any that could succeed it
  - decide by dynamical similarity criterion
  - see also (Kovar et al 02; Lee et al 02)
- A path is a motion
- Search with constraints
  - like root position+orientation, etc.
  - Local (Kovar et al 02)
  - With some horizon
    - Lee et al 02; Ikemoto, Arikan+Forsyth 05
  - Whole path
    - Arikan+Forsyth 02; Arikan et al 03

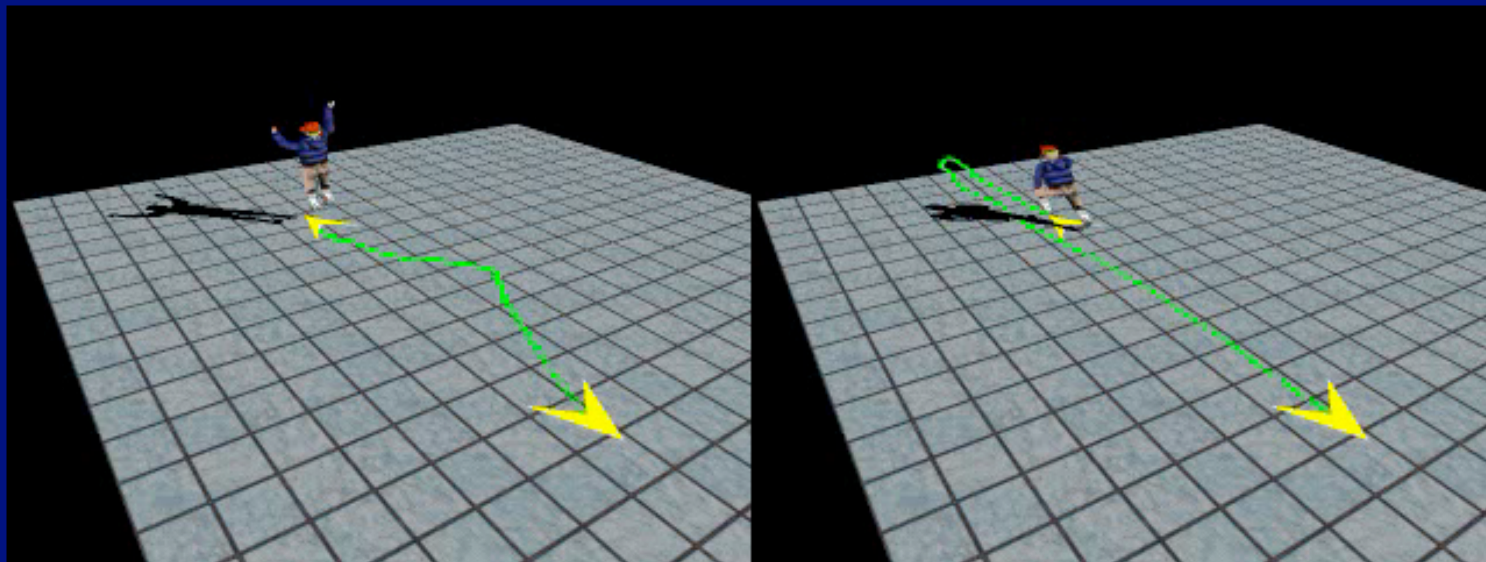
Motion Graph:

Nodes = Frames

Edges = Transition

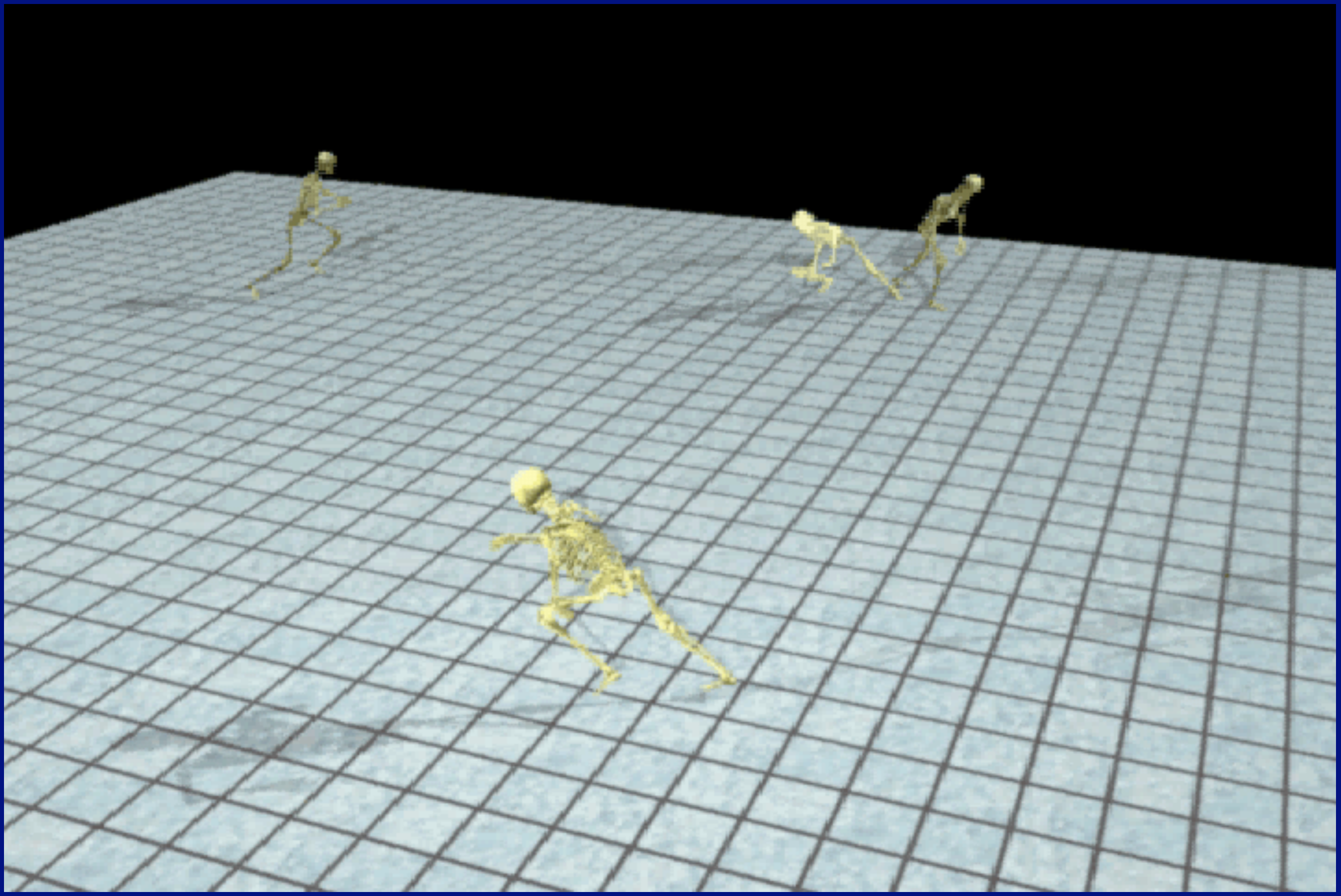
A path = A motion





- Characteristic features
  - most demands are radically underconstrained
  - motion is simultaneously
    - hugely ambiguous
    - “low entropy”
- Suggests using “summaries”

Arikan+Forsyth 02;
Lee et al 02;

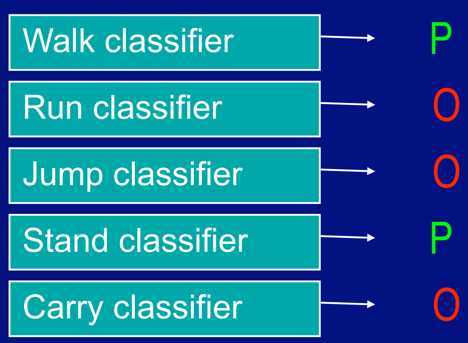


Arikan+Forsyth 02

Annotation - desirable features

- Composability
  - run and wave;
- Comprehensive but not canonical vocabulary
  - because we don't know a canonical vocabulary
- Speed and efficiency
  - because we don't know a canonical vocab.

- Can do this with one classifier per vocabulary item
  - use an SVM applied to joint angles
  - form of on-line learning with human in the loop
  - works startlingly well (in practice 13 bits)



Arikan+Forsyth+O'Brien 03

	?	?	?	...	?	*n* - frames
Walk	P	P	P		P	
Run	●	●	●	Motion demand	●	
Jump	●	●	●		●	
Wave	P	P	O		O	
Carry	●	●	●		●	

Synthesis by dynamic programming

Arikan+Forsyth+O'Brien 03

Dynamic programming practicalities

- Scale
  - Too many frames to synthesize
  - Too many frames in motion graph
- Obtain good summary path, refine
  - Form long blocks of motion, cluster
  - DP on stratified sample
    - split blocks on "best" path
    - find similar subblocks
      - DP on this lot
        - etc. to 1-frame blocks

Arikan+Forsyth+O'Brien 03

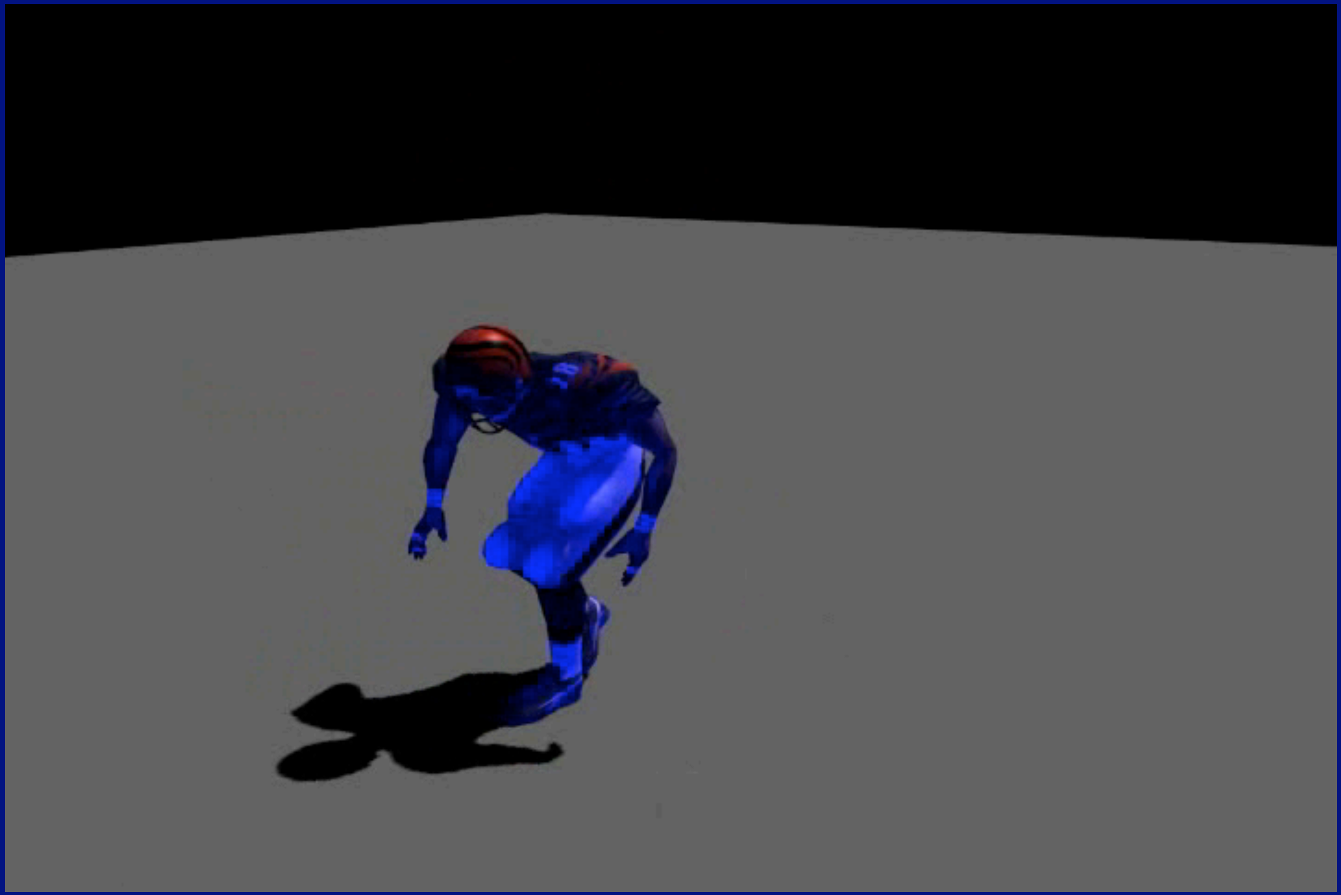
Arikan+Forsyth+O'Brien 03

Transplantation

- Motions clearly have a compositional character
  - Why not cut limbs off some motions and attach to others?
    - we get some bad motions
  - build a classifier to tell good from bad
    - avoid foot slide by leaving lower body alone

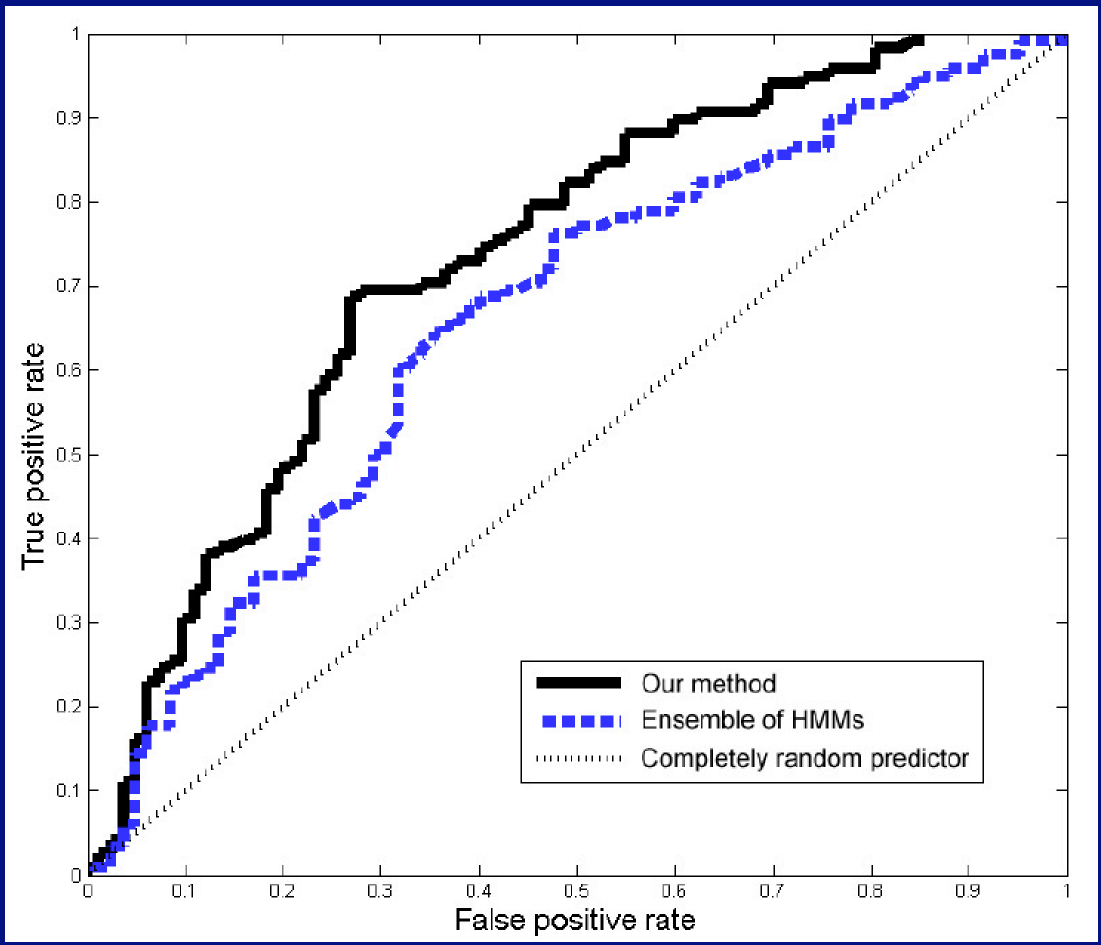
Ikemoto+Forsyth 04

Ikemoto+Forsyth 04



Ikemoto+Forsyth 04

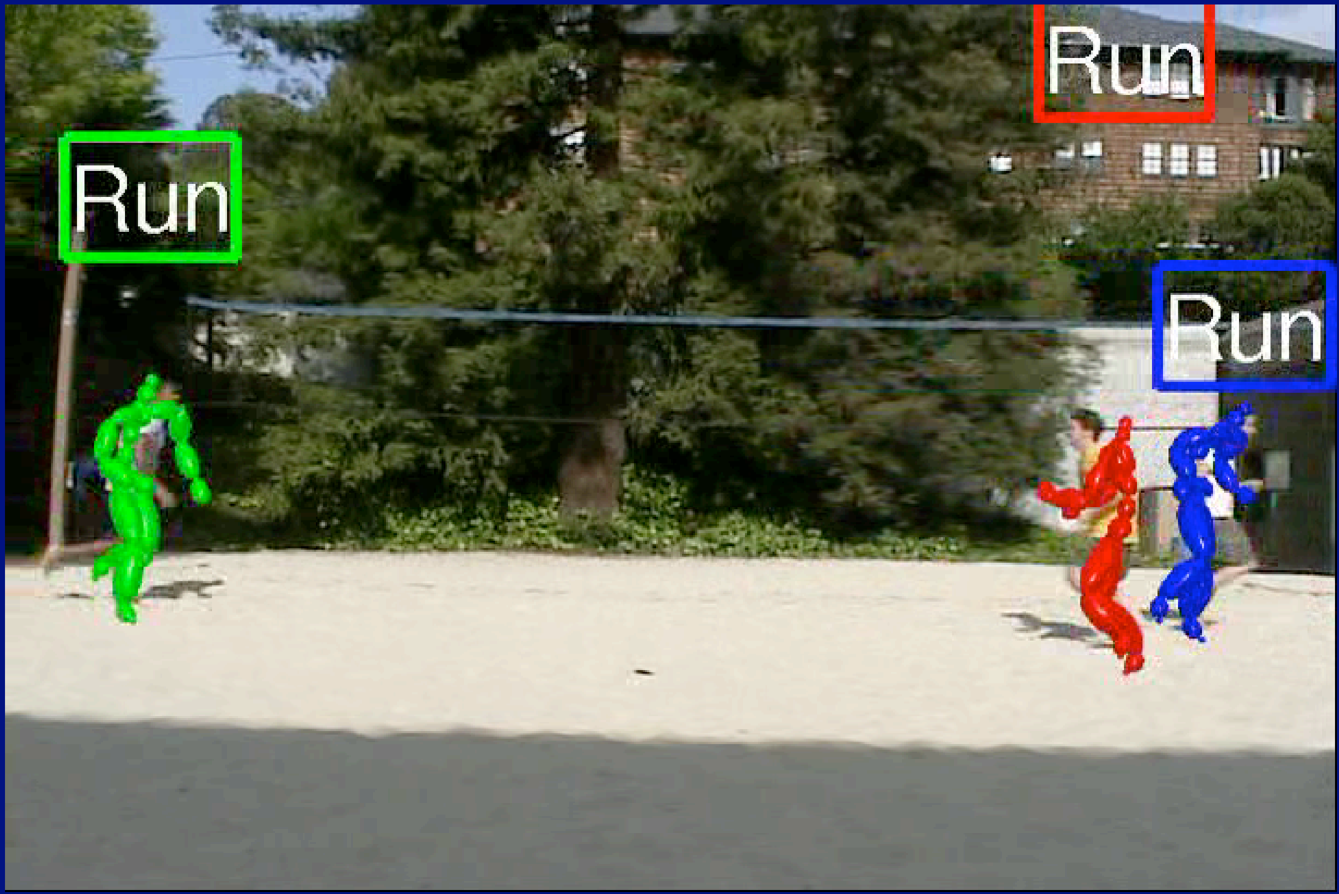
It is hard to tell good from bad automatically



cf Ren et al 05 for HMM’s

Ikemoto Arikan Forsyth 07

Activity recognition



Naming activities

- Absence of a canonical vocabulary is a serious problem
  - strategies
    - adopt specialized domains (Bobick+Davis 01, Efros et al 03)
    - guess a vocabulary (Efros et al 03)
    - match motion to motion and avoid the issue (Efros et al 03)
    - use vocab useful for synthesis (Ramanan et al 03)

Bobick & Davis. PAMI01

Activity recognition

- By comparison to labelled data
  - benefit from temporal smoothing
    - aka motion synthesis
- By inference on a generative model
  - so we can search for activities without having ever seen them
    - composition over body and space
- By discriminative method
  - transfer learning by feature construction deals with
    - aspect
    - shortage of training data

Annotating observations by synthesis