Activity and Kinematics

D.A. Forsyth, UIUC (was U.C. Berkeley; was U.Iowa)
Leslie Ikemoto, Okan Arikan, of Animeeple
Deva Ramanan of TTI/UC Irvine
Ali Farhadi of UIUC Nazli Ikizler of Bilkent U (now Boston U; soon Hacettepe U)
Alex Sorokin, UIUC Du Tran, UIUC Duan Tran, UIUC, Wei Yan, Texas A+M

Thanks to: Electronic Arts, Sony SCEA, ONR MURI, NSF, DHS

Core questions

- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - style and composition
- How should we transduce the signal?
  - infer body segments or not
- Bias and generalization
  - inevitable problems with complex high dimensional signals

What should activity recognition say?

- Report names of activity of all actors (?!?)
    - but we might not have names
    - and some might not be important
- Make useful reports about what’s going on
    - what is going to happen?
    - how will it affect me?
    - who’s important?
- Do activity categories exist?
    - allow generalization
        - future behavior; non-visual properties of activities

Unfamiliar activities present no real problem

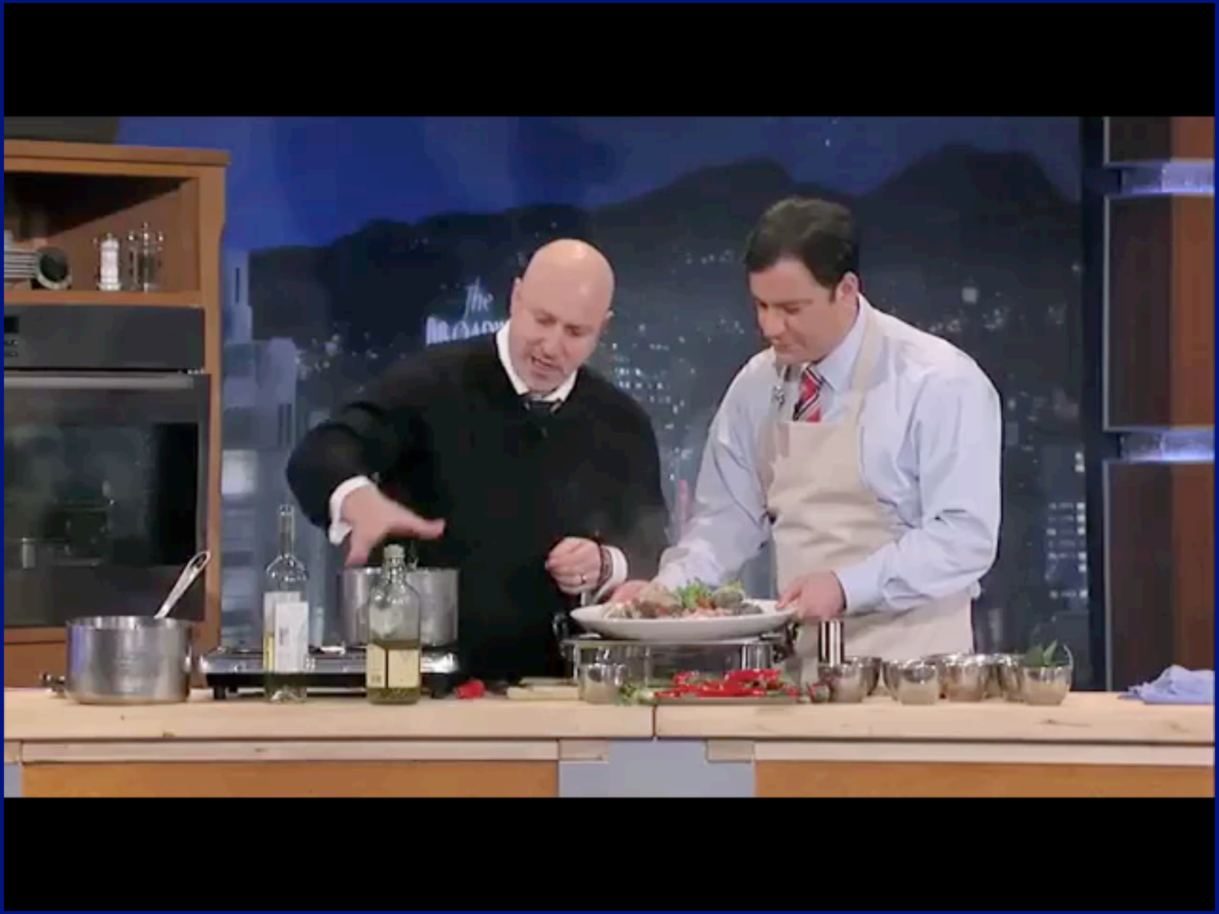
Unfamiliar activities present no real problem

Unfamiliar activities present no real problem

Kinematic detail can be informative

Weird actors present no real problem

Interactions often tell story



How is it going to affect me?

What outcome do we expect?

How are other people feeling?

What will they do?

What outcome do we expect?

How are other people feeling?

What will they do?

What outcome do we expect?

How are other people feeling?

What will they do?

How many adults were on the platform and what were they doing?

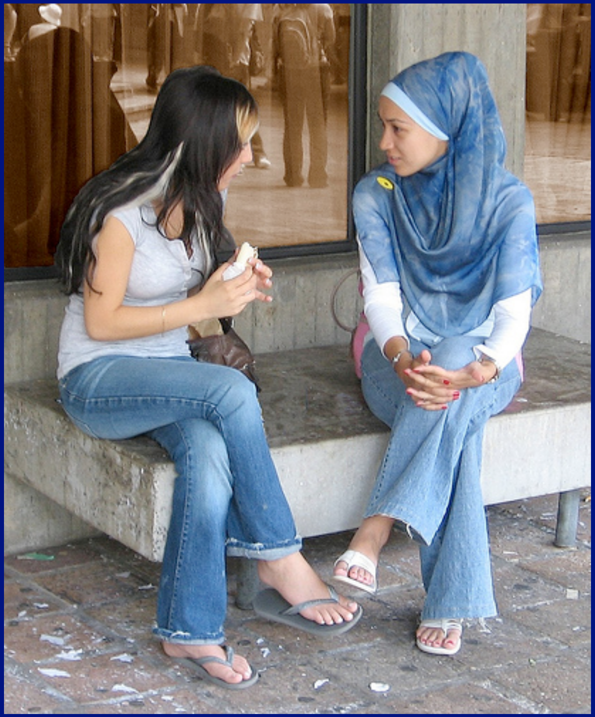
What’s going to happen to the baby?

What outcome do we expect?

How are other people feeling?

What will they do?

Choosing what to report



Two girls take a break to sit and talk .

Two women are sitting , and one of them is holding something .

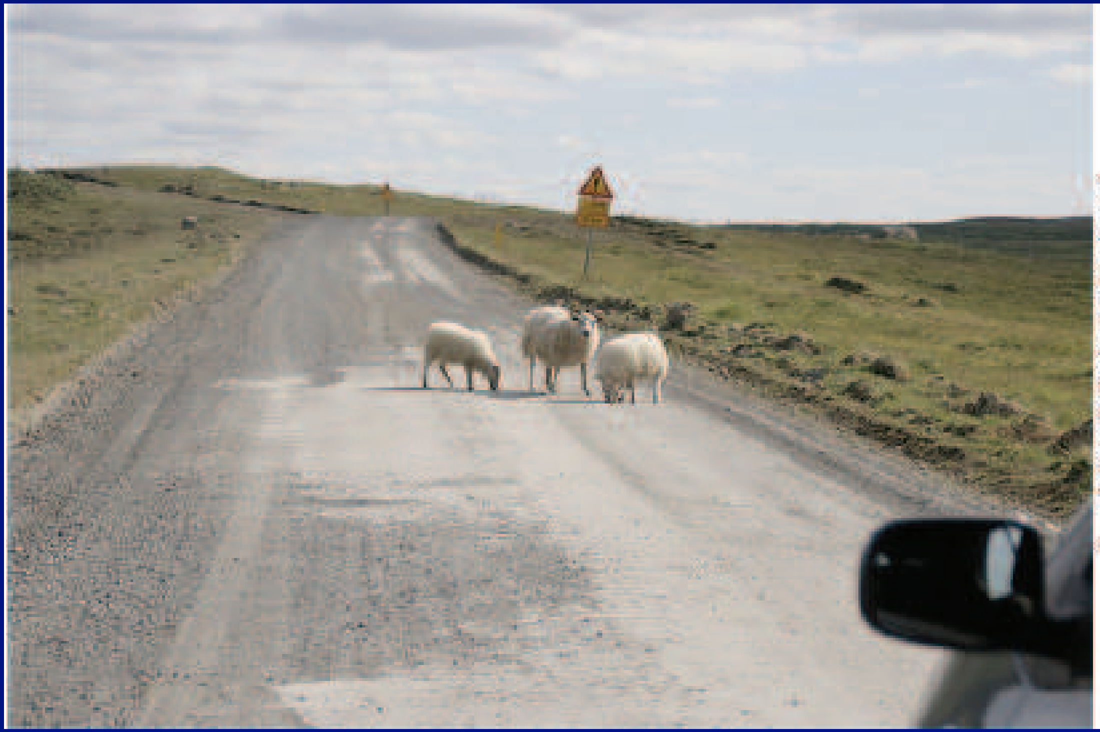
Two women chatting while sitting outside

Two women sitting on a bench talking .

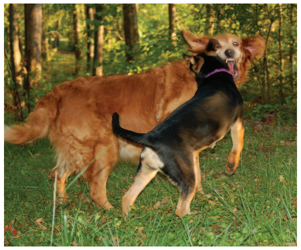
Two women wearing jeans , one with a blue scarf around her head , sit and talk .

Sentences from Julia Hockenmaier’s work

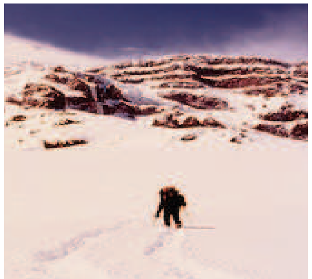
Rashtchian ea 10



The goats on the way
A car on a rural dirt and gravel road approaches a group of three sheep grazing.
A small group of sheep in a dirt road.
Three sheep on a rural road, about to block traffic.
Three sheeps on the road out of nowhere.

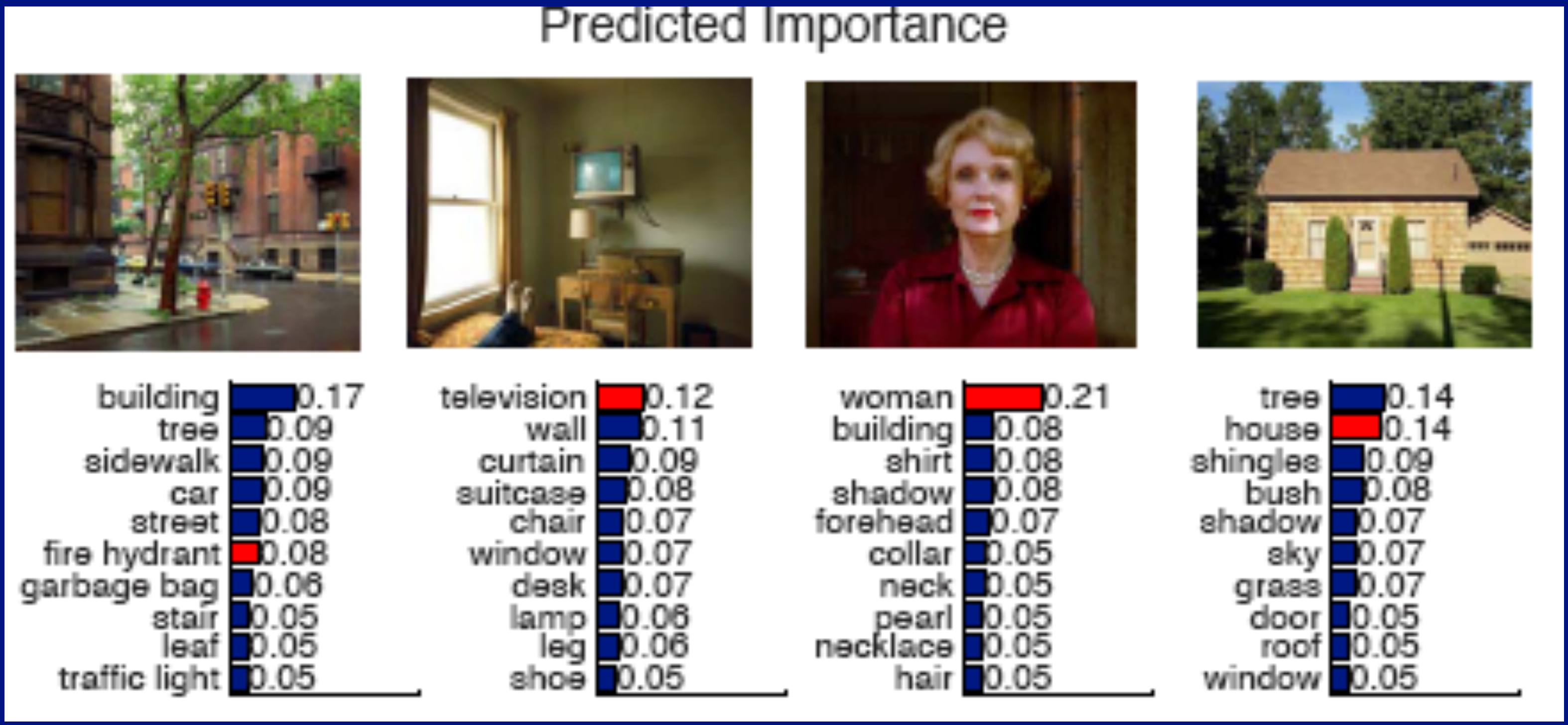


A golden retriever (ANIMAL) is playing with **a smaller black and brown dog** (ANIMAL) in a **pink collar** (CLOTHING).
A smaller black dog (ANIMAL) is fighting with **a larger brown dog** (ANIMAL) in **a forest** (NAT_BACKGROUND).
A smaller black and brown dog (ANIMAL) is jumping on a **large orange dog** (ANIMAL).
Brown dog (ANIMAL) with **mouth** (BODY_PART) open near **head**(BODY_PART) of **black and tan dog** (ANIMAL).
Two dogs (ANIMAL) playing near **the woods** (NAT_BACKGROUND).



A lone hiker (PERSON) treks through **deep snow** (NAT_BACKGROUND) near **rocky peaks**(NAT_BACKGROUND).
A mountain climber (PERSON) on **a snowy plain**(NAT_BACKGROUND) near **a mountain top**(NAT_BACKGROUND).
A person (PERSON) travels down **a snowy path**(NAT_BACKGROUND) into **the mountains** (NAT_BACKGROUND).
Someone (PERSON) is walking through **the snow**(NAT_BACKGROUND) with **snow-covered mountains** (NAT_BACKGROUND) behind **then**(PERSON).
On a **mountain top**(NAT_BACKGROUND) **a climber** (person) is seen in the distance(orientation), **black figure**(PERSON) against **white snow**(BACKGROUND_NATURAL).

Hodosh ea 2010



Spain ea 08; red is human importance, blue is urn model

Good properties of recognition

- Bias robust
  - biases, sparsity in training data don’t affect test behaviour (much)
- Unfamiliarity
  - Make useful statements about objects whose name isn’t yet known
- Manage deviant objects
  - Say how a detected object is different from the usual
- Learn by X
  - Single picture
  - Reading
    - Description (0 pictures; zero shot learning)
- Accuracy
  - be good at recognizing known objects

Core questions

- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - composition and style
- How should we transduce the signal?
  - infer body segments or not
- Bias and generalization
  - inevitable problems with complex high dimensional signals

Motion Capture

The motion signal

- There is no reliable method for generating novel motions
  - some special cases work OK

  - Keys for special cases
    - data driven methods work well for temporal composition
    - Some motions can be blended successfully
    - Contacts create special problems
    - There are complex, cross-body correlations

- There must be some set of motion primitives

Data driven methods and composition

- Composition is an important source of complexity
  - (flexibility for planning, control)
- We can join motions up in time to make new motions
  - The process is now quite well understood
  - Good quality can be obtained
  - Useful in animation
- We can join up parts of motion across the body
  - But it doesn’t always work (and we don’t know why, really)

Cut and Paste works well over time

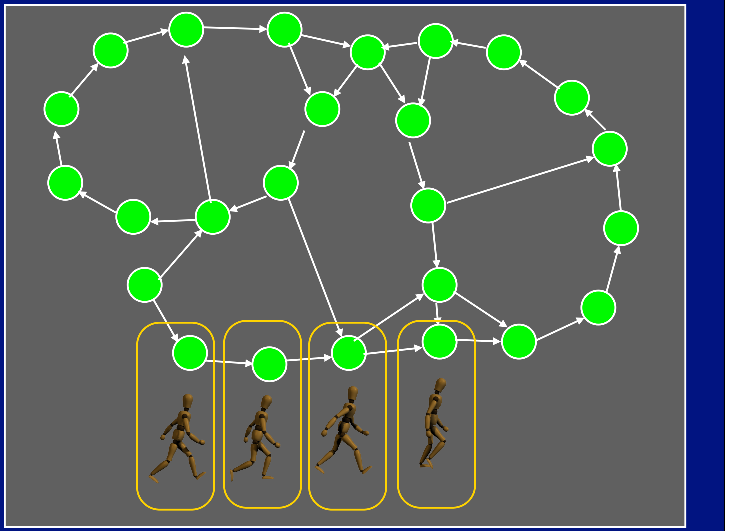
- Motion graph: by analogy with
    - text synthesis, texture synthesis, video textures
- Take measured frames of motion as nodes
    - from motion capture, given us by our friends
- Edge from frame to any that could succeed it
    - decide by dynamical similarity criterion
    - see also (Kovar et al 02; Lee et al 02)
- A path is a motion
- Search with constraints
    - like root position+orientation, etc.
    - In various ways
        - Local (Kovar et al 02)
        - Lee et al 02; Ikemoto, Arikan+Forsyth 05
        - Arikan+Forsyth 02; Arikan et al 03

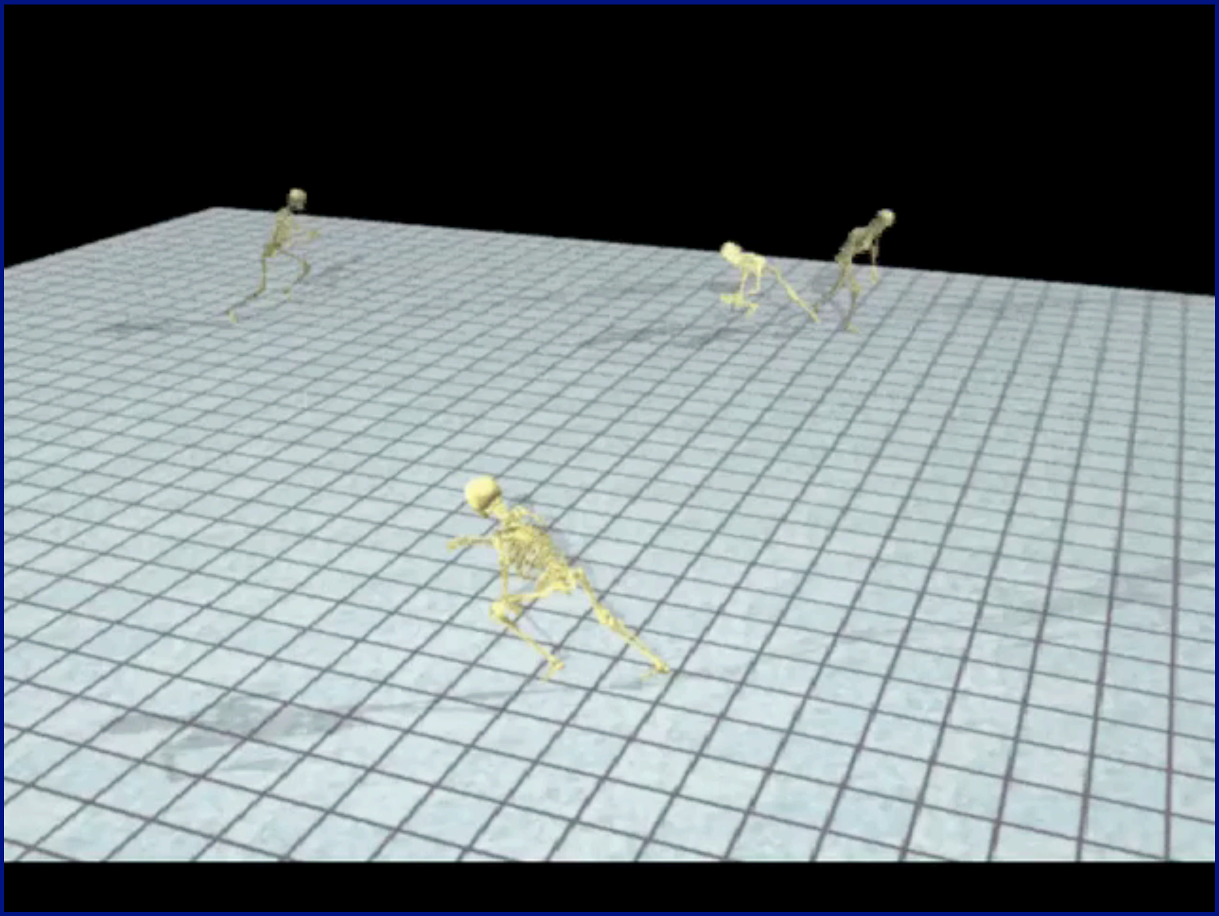
Motion Graph:

Nodes = Frames

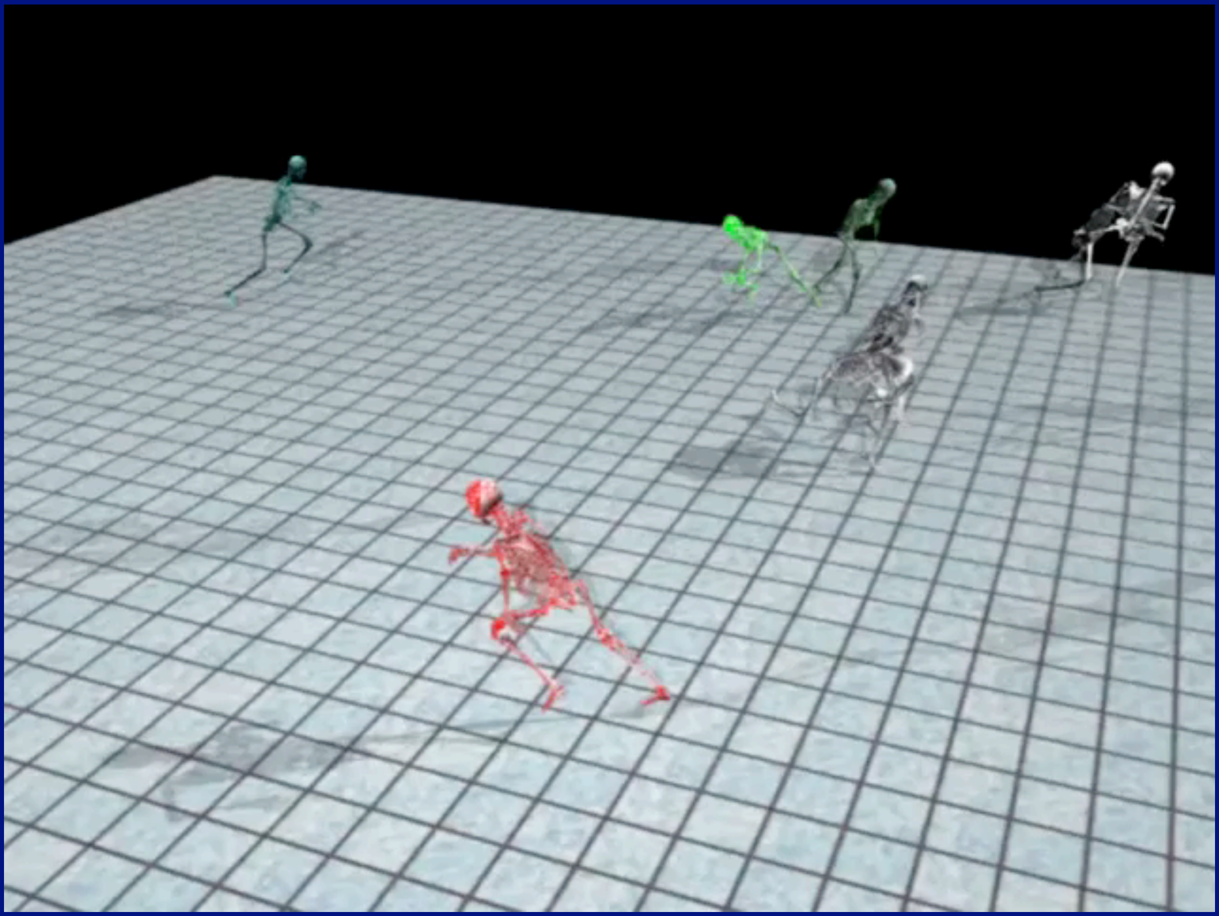
Edges = Transition

A path = A motion

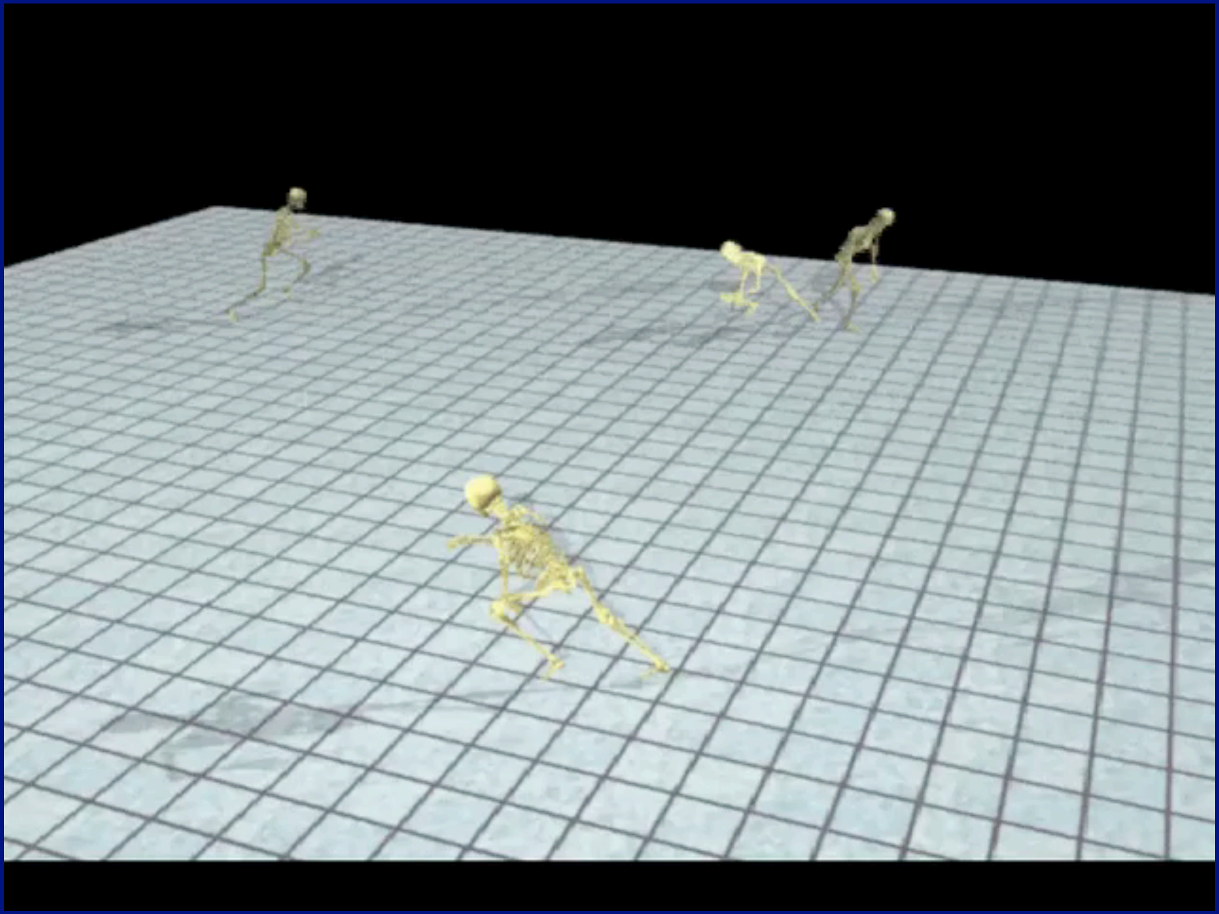




Arikan+Forsyth 02



Arikan+Forsyth 02



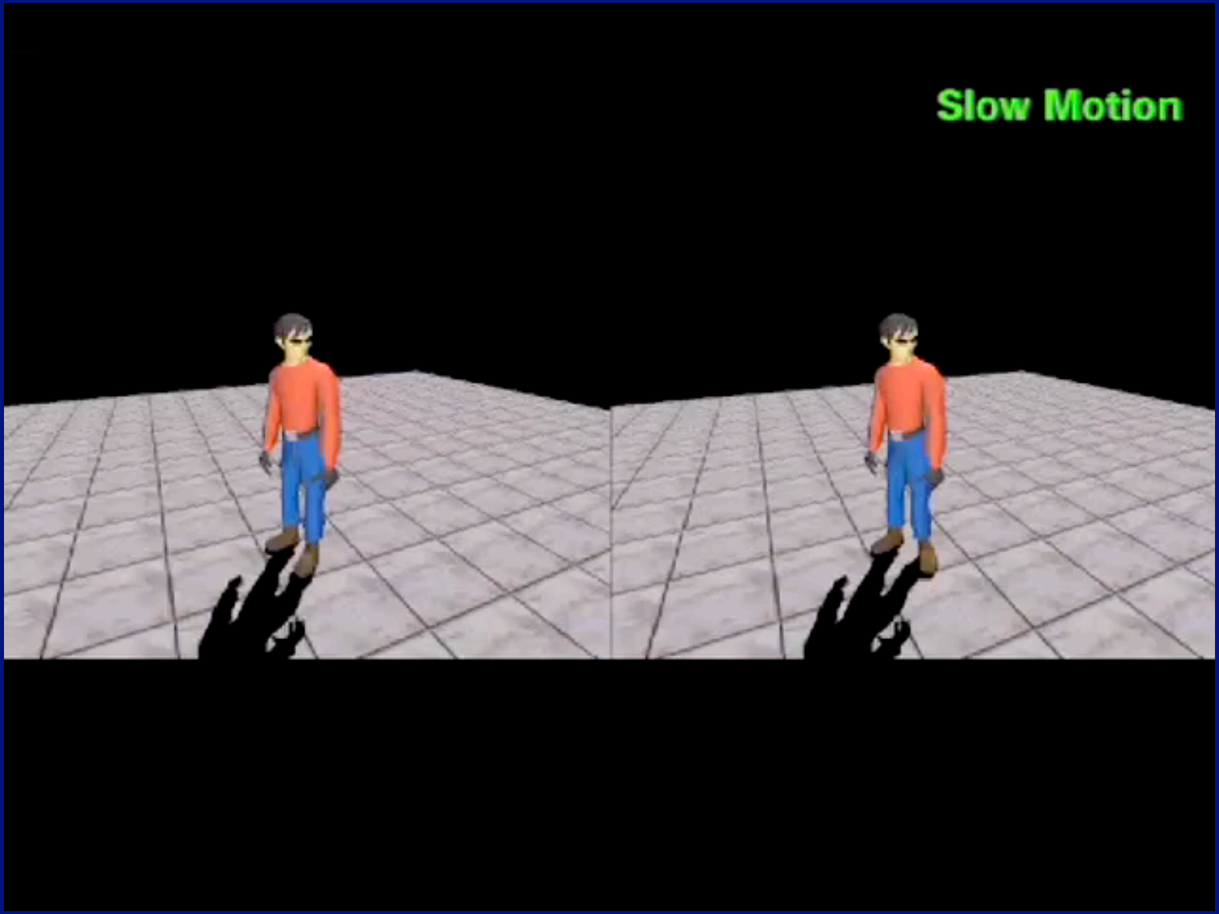
Arikan+Forsyth 02

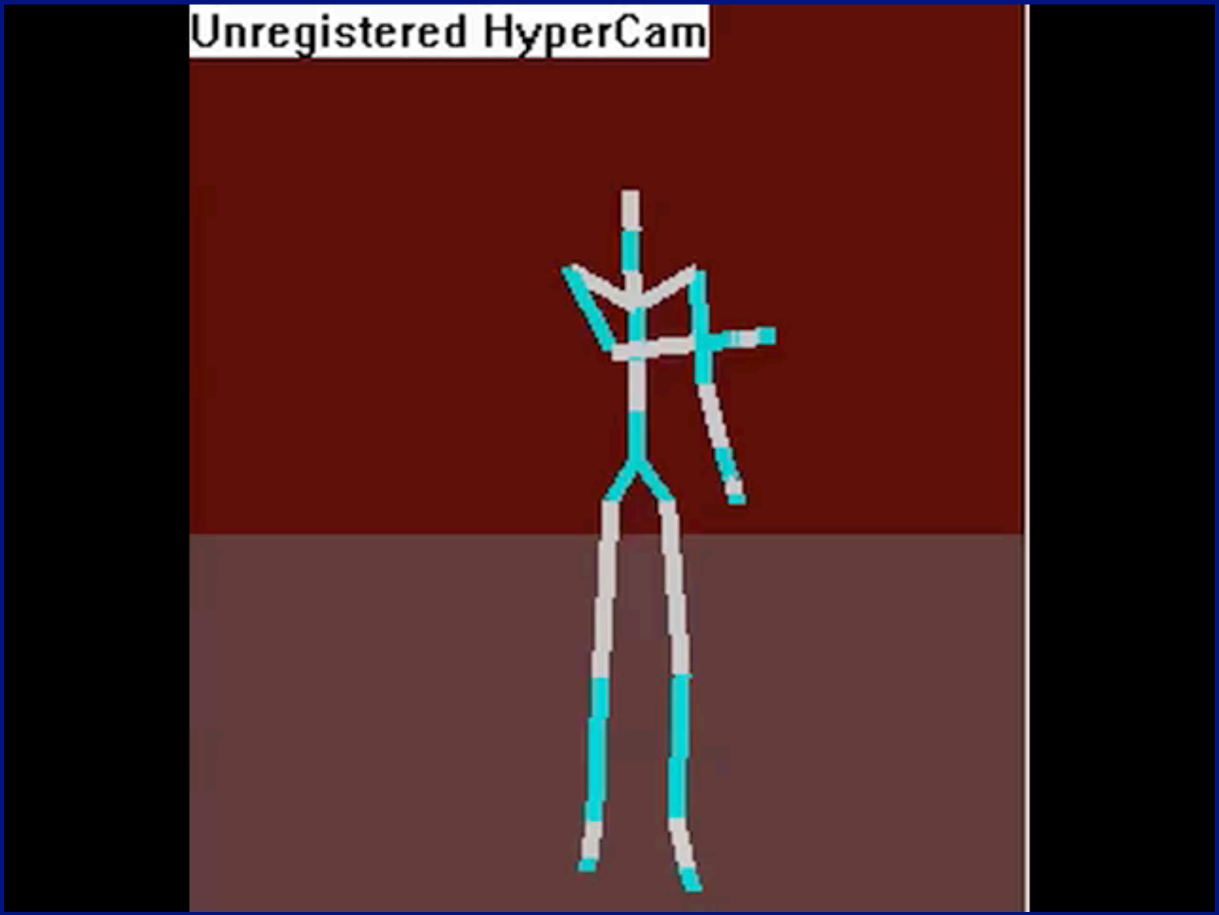
Non data-driven methods don’t work yet

- Temporally fast phenomena are important to perception
  - means obvious methods work poorly
    - Blending works ok sometimes
    - Compression works ok sometimes
    - Tracking works ok sometimes

  - All mess up contacts

Footskate

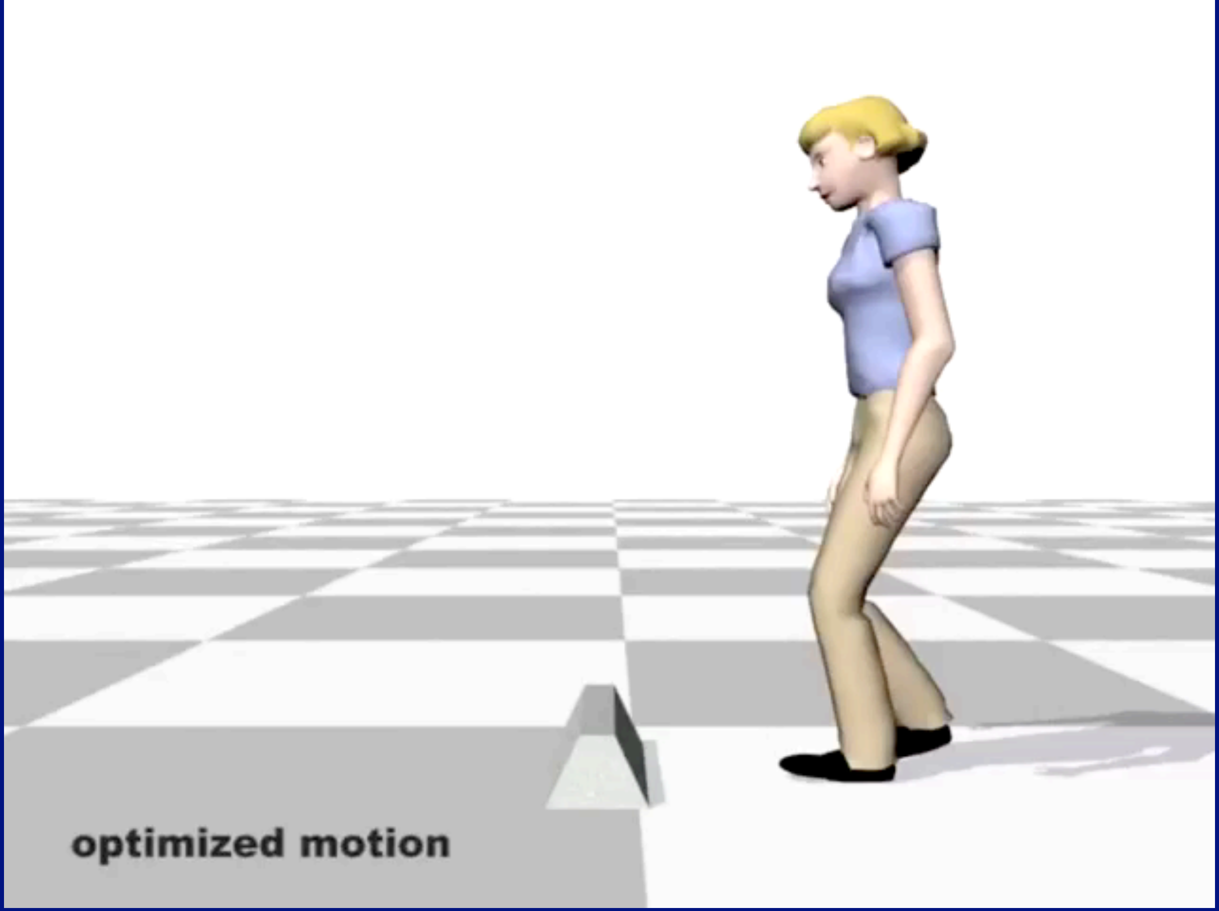




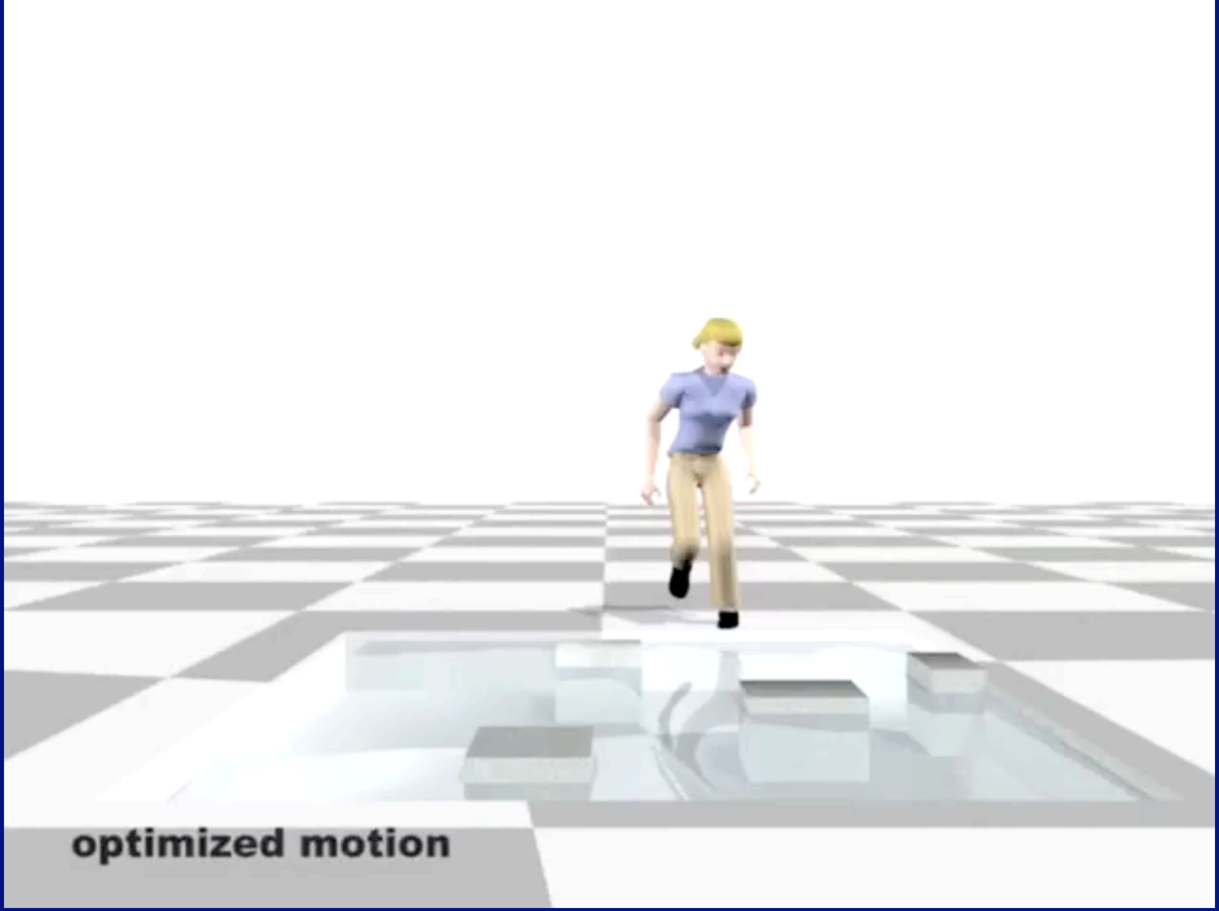
Mataric et al,

Unregistered HyperCam

Mataric et al,



Safonova ea 04



Safonova ea 04

The Benefit of Interpolation

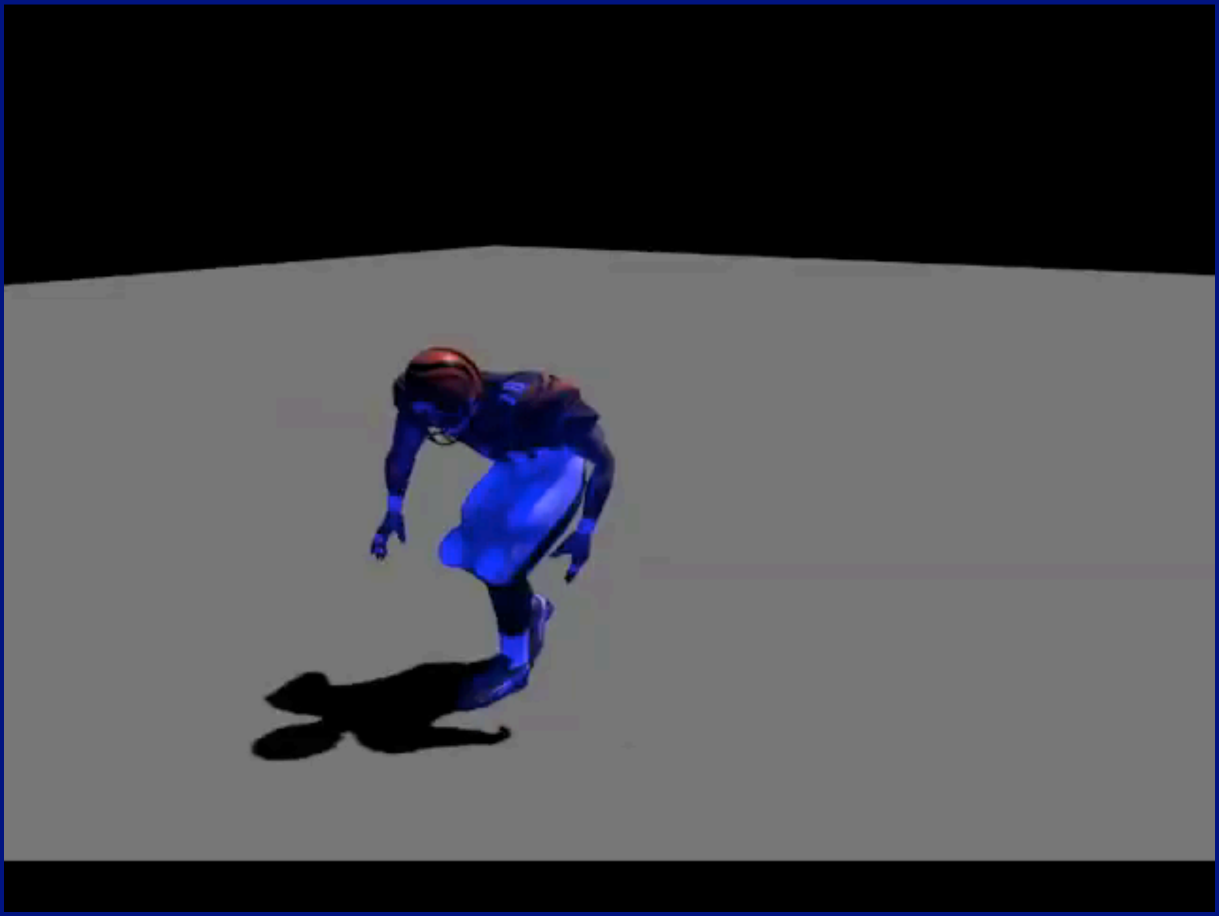
Safonova ea 07

Transplantation

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

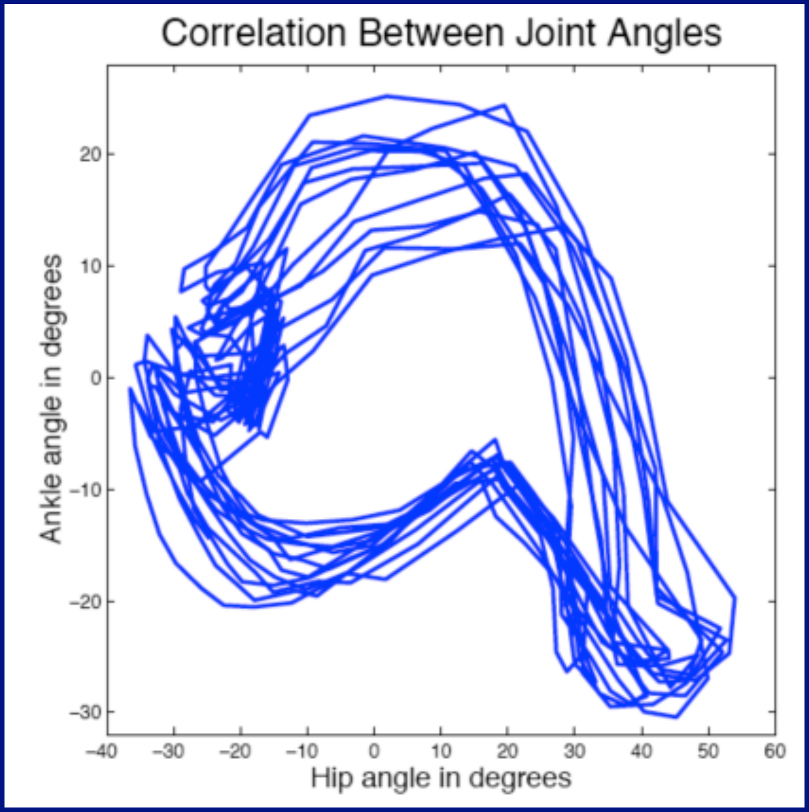
Ikemoto+Forsyth 04

Ikemoto+Forsyth 04



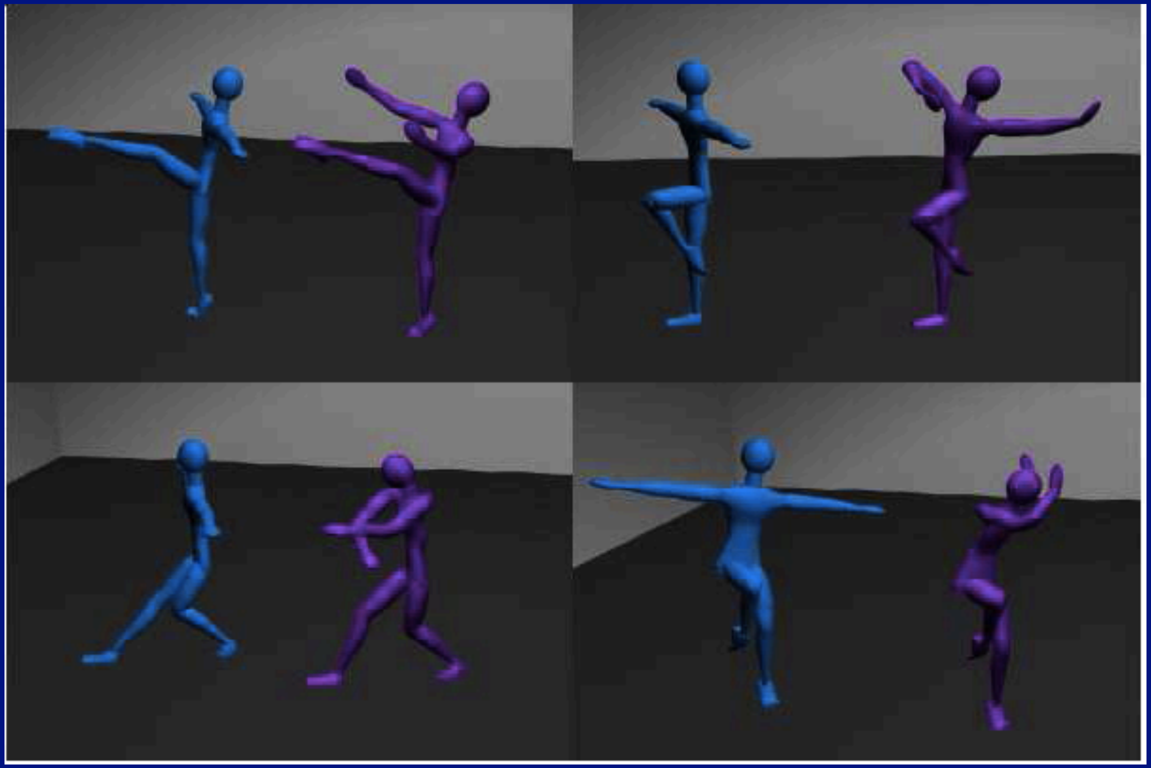
Ikemoto+Forsyth 04

Joint angles are heavily correlated



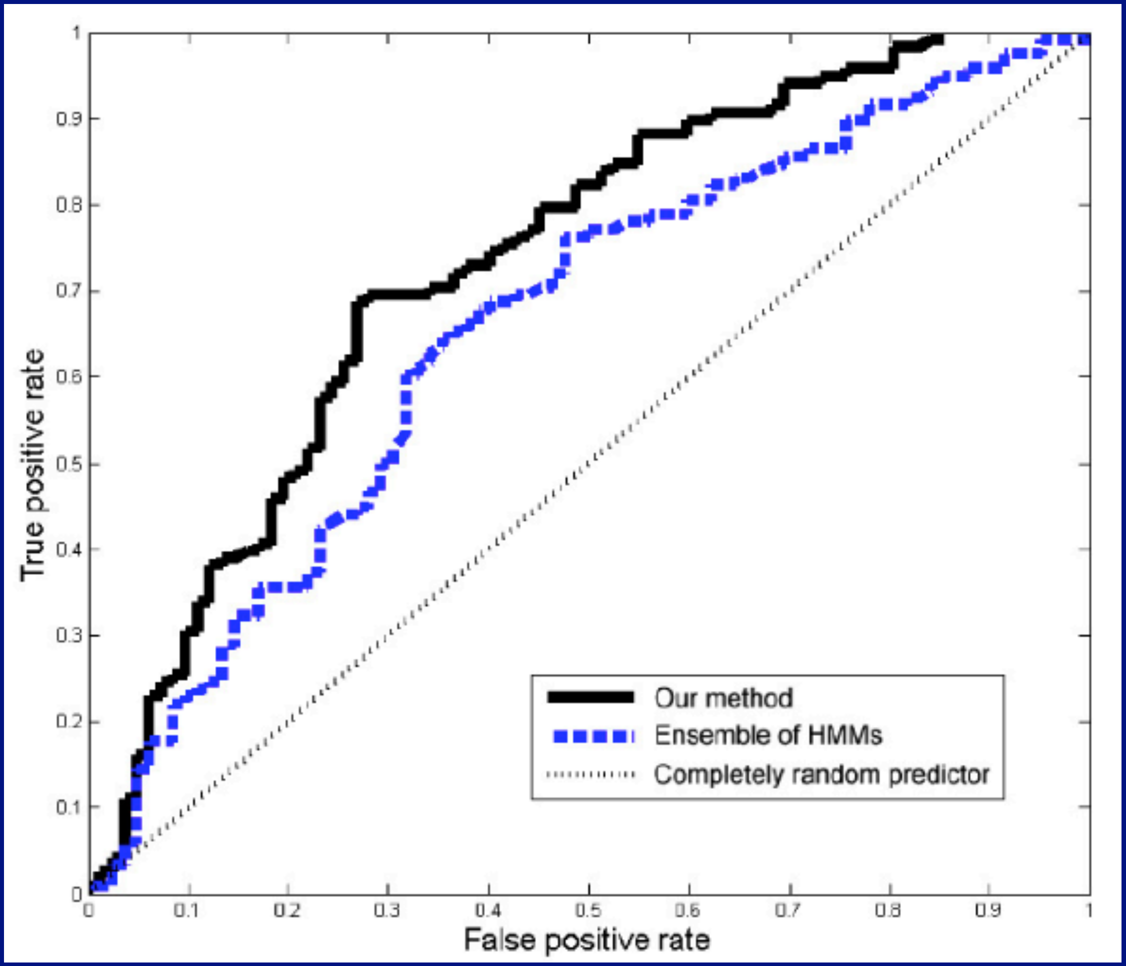
Pullen + Bregler 02

Joint angles are heavily correlated



Pullen + Bregler 02

Hard to tell good from bad



Ikemoto Arikan Forsyth 07 cf Ren et al 05 for HMM’s

Why should we care?

- People seem very aware of detail in other peoples motion
  - footplants, contacts, etc.
  - maybe cues to what motion comes next?
- Temporal composition rules!
  - because nothing else looks natural
  - very hard to escape at present
  - consequence: major shortage of motion capture data
- Body composition seems like the right direction
  - but details are hard to get right
  - covariance across body might help us?

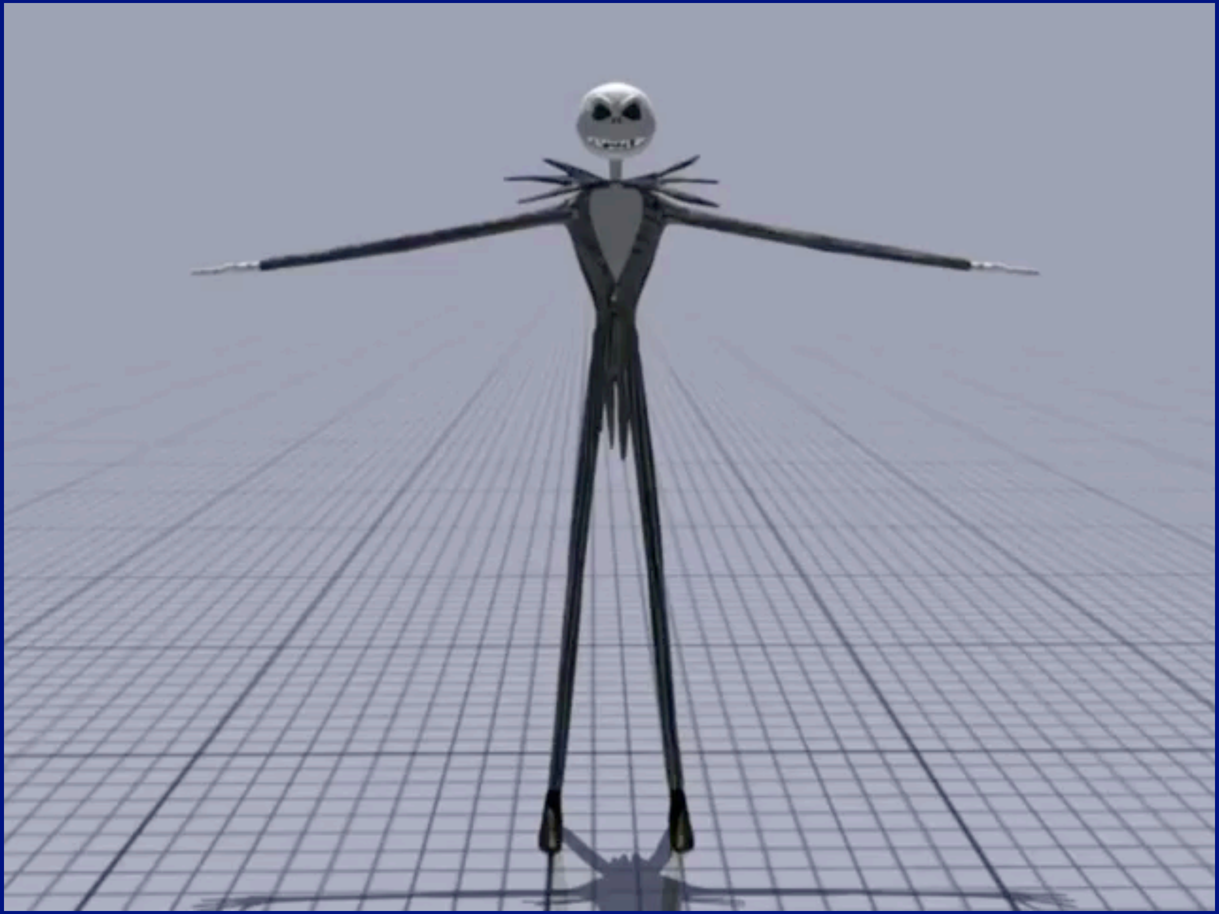
Style

- Qualitative properties of motion, including
  - individual characteristics
  - modifiers, eg: clumsy, fast, heavy, forceful, graceful

- Animation problem:
  - Control new character with old motion, preserving new character’s style

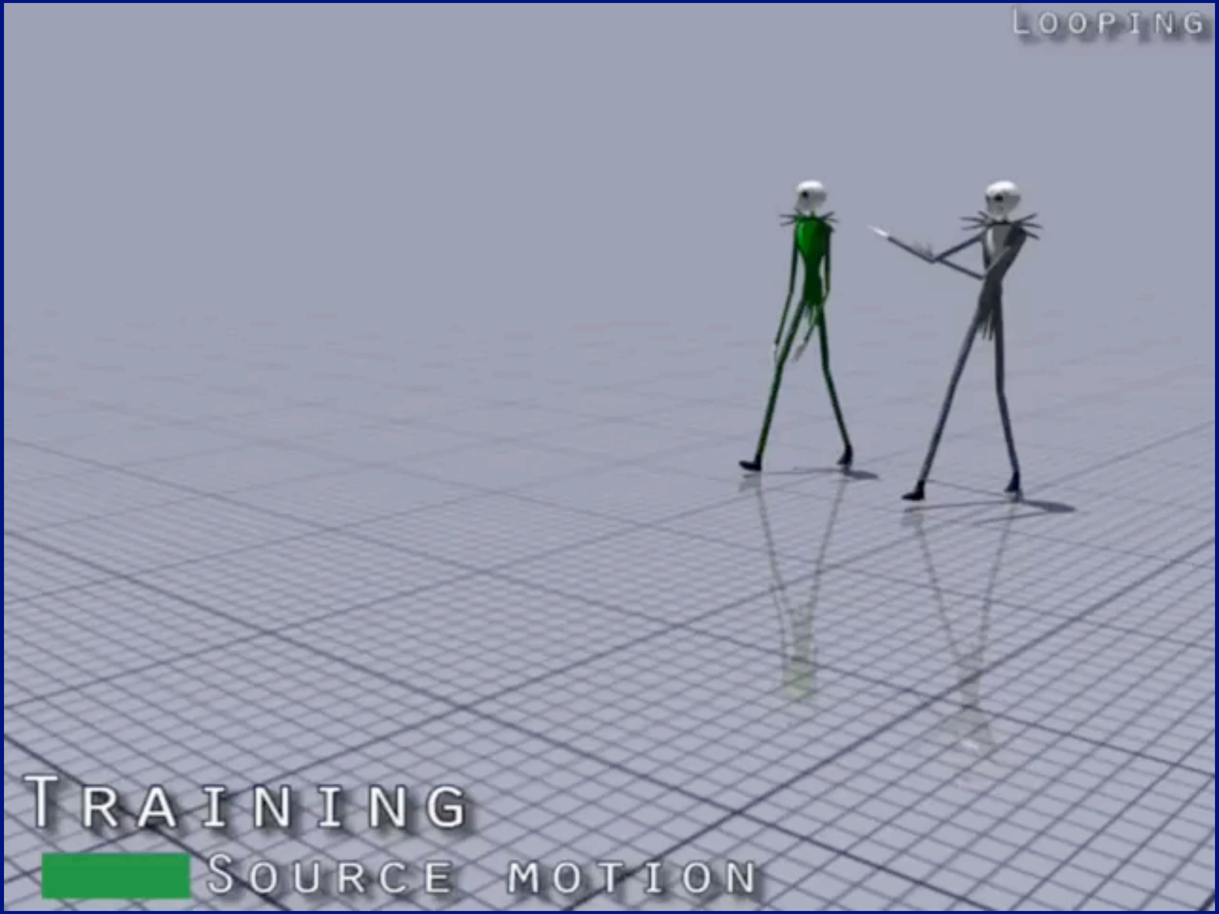
- Vision problem:
  - infer style descriptors, identity from observed motion

Kinematic style transfer



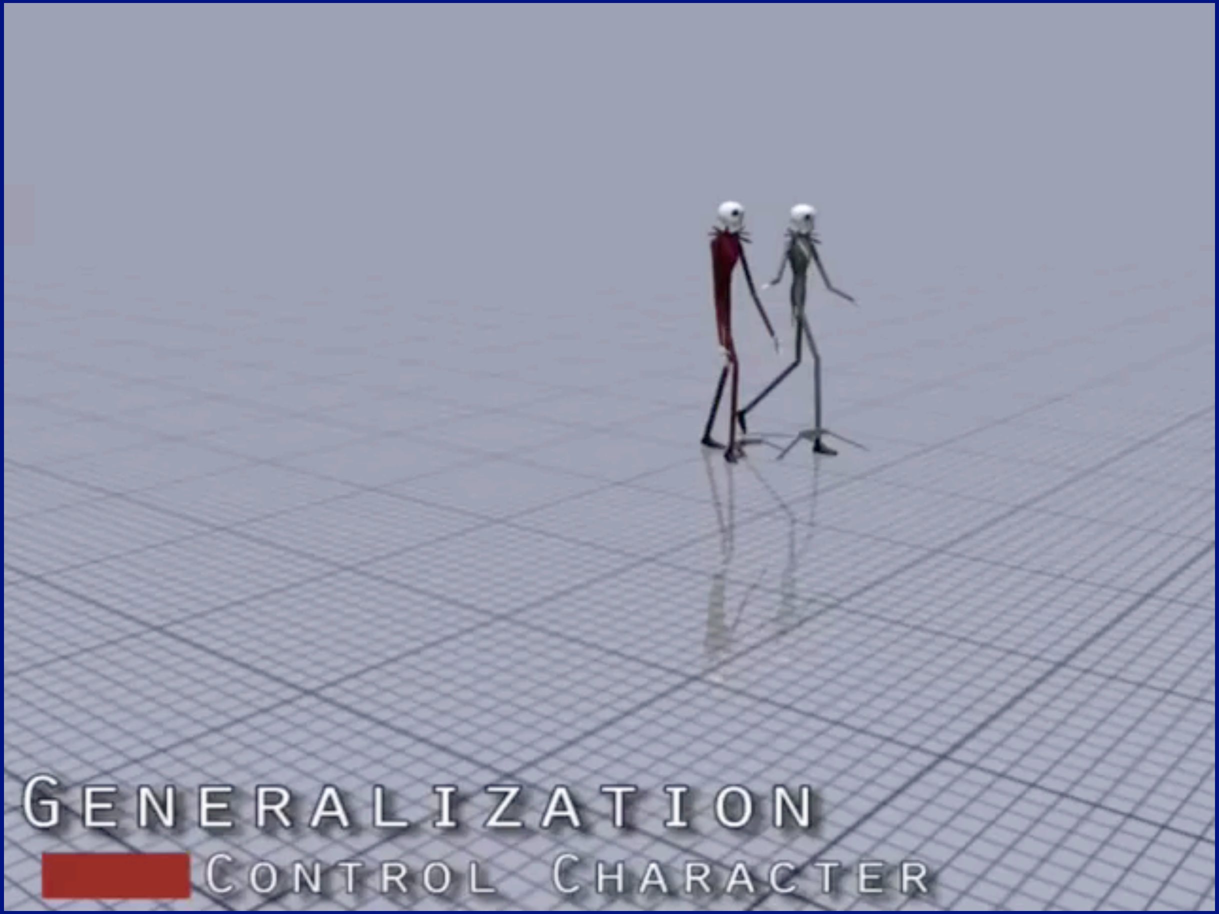
Ikemoto ea 09

Kinematic style transfer

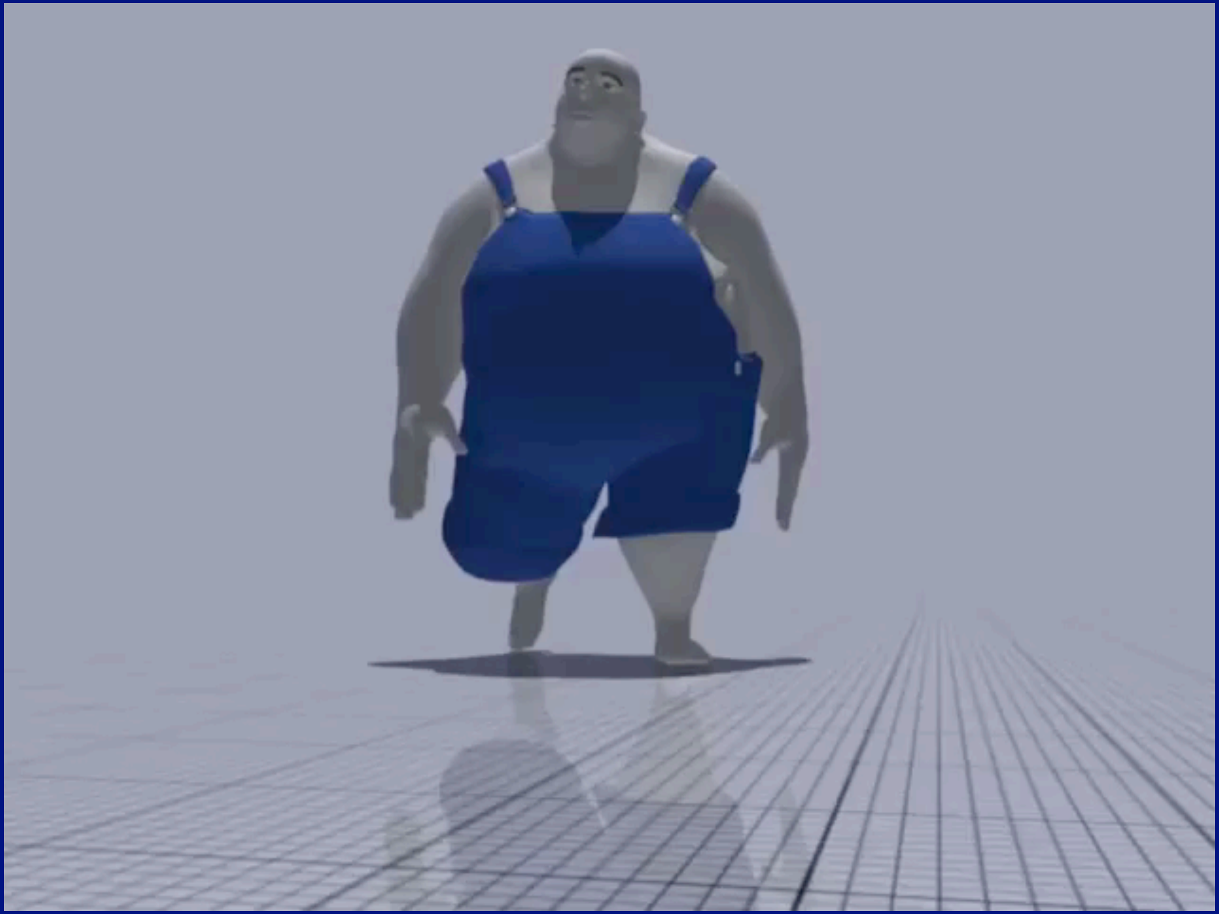


Ikemoto ea 09

Kinematic style transfer



Kinematic style transfer



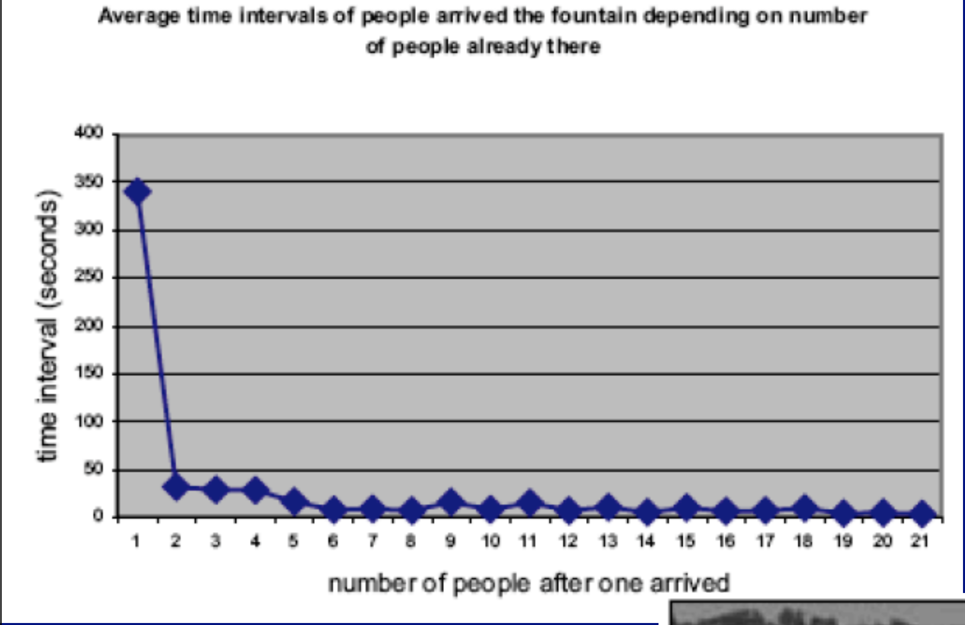
Ikemoto ea 09

Why should we care?

- How is the person moving?
  - rather than what are they doing
- May identify individuals

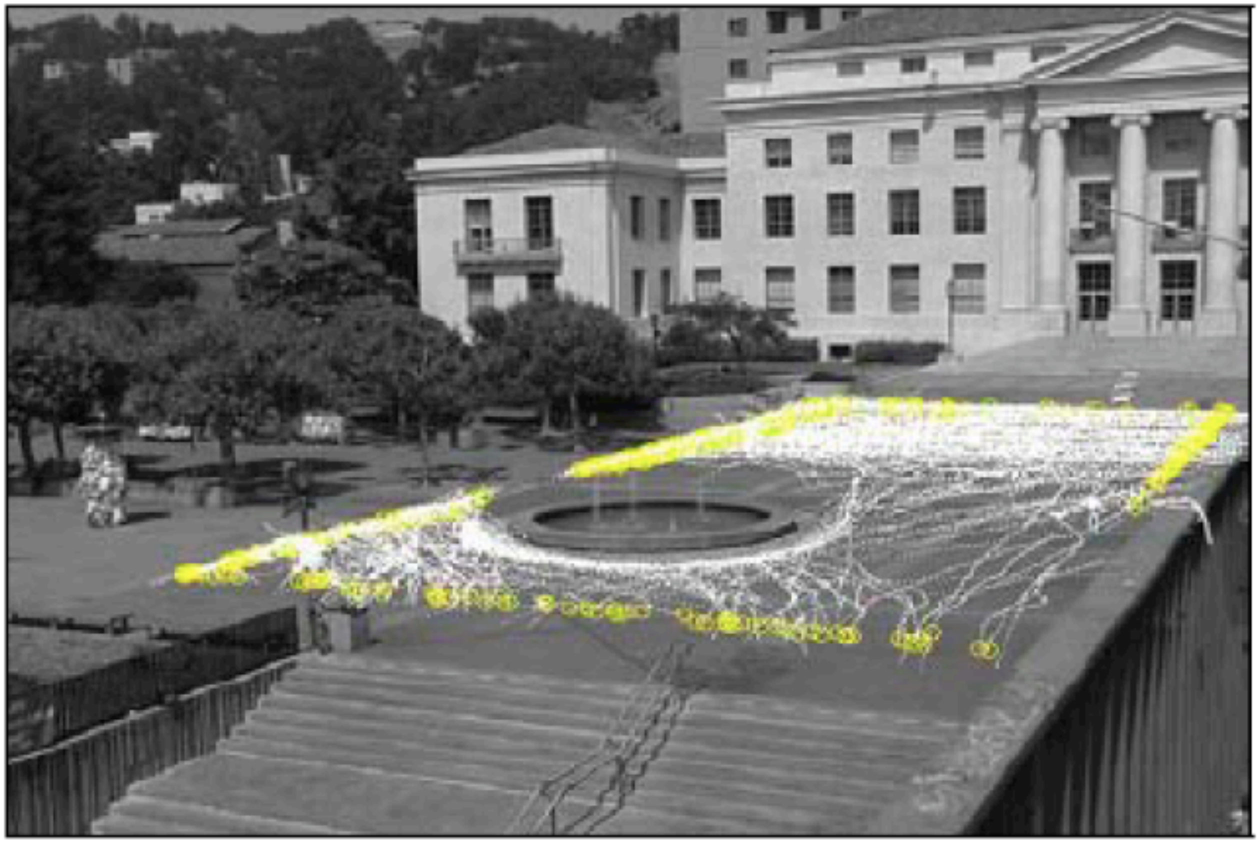
Core questions

- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - style and composition
- How should we transduce the signal?
  - infer body segments or not
- Bias and generalization
  - inevitable problems with complex high dimensional signals



Point tracks reveal curious phenomena in public spaces

Yan+Forsyth, 04



Transduction

- Frames can be distinctive
- Multiple views seem to help
- Key questions:
  - segment body parts or not
  - how to represent timing
  - how to represent style

Why is kinematic tracking hard?

- It’s hard to detect people
  - until recently, 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

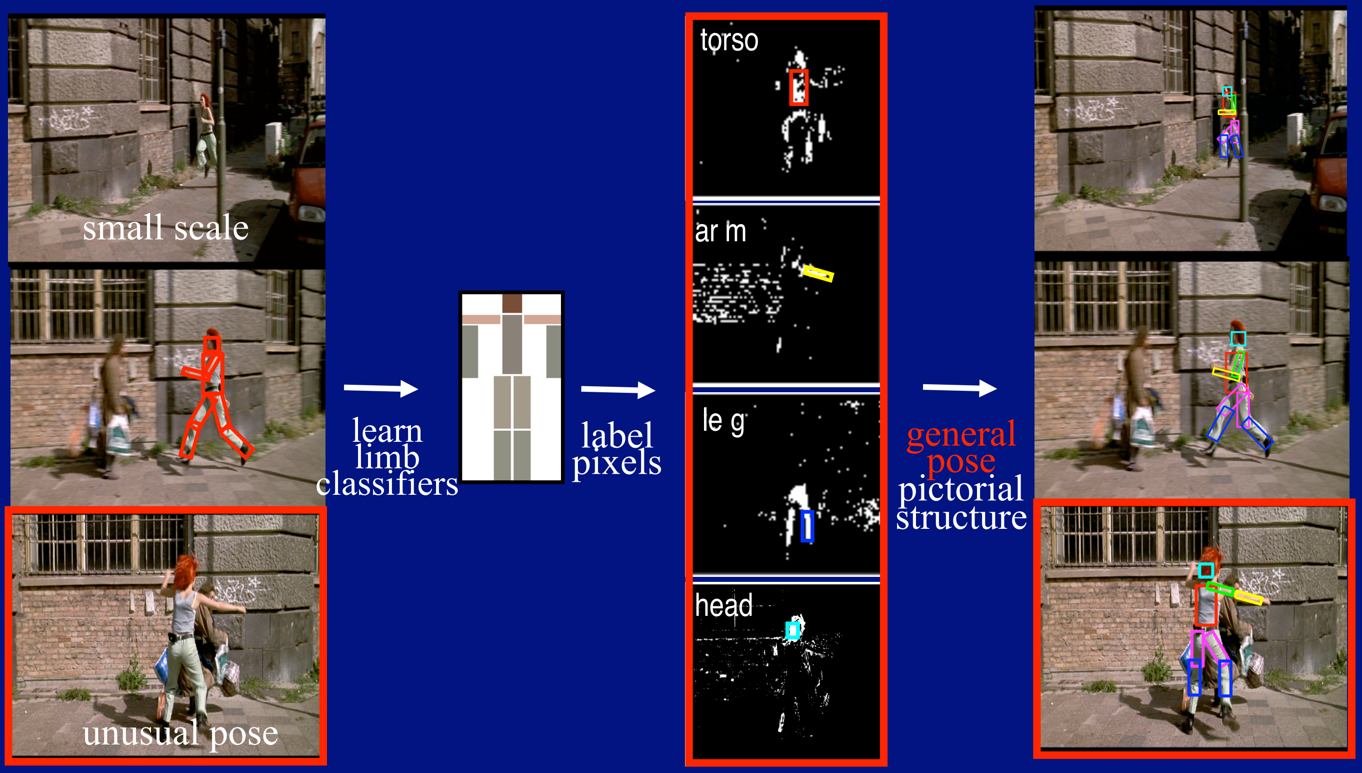
self-occlusion & clutter

variation in appearance

Kinematic tracking background

- Desirable for:
  - Video motion capture
  - HCI
  - Activity recognition
- Main threads:
  - 3D representation vs. 2D representation
  - Mechanics of inference
    - multiple modes in posterior
    - speed

Build and detect models



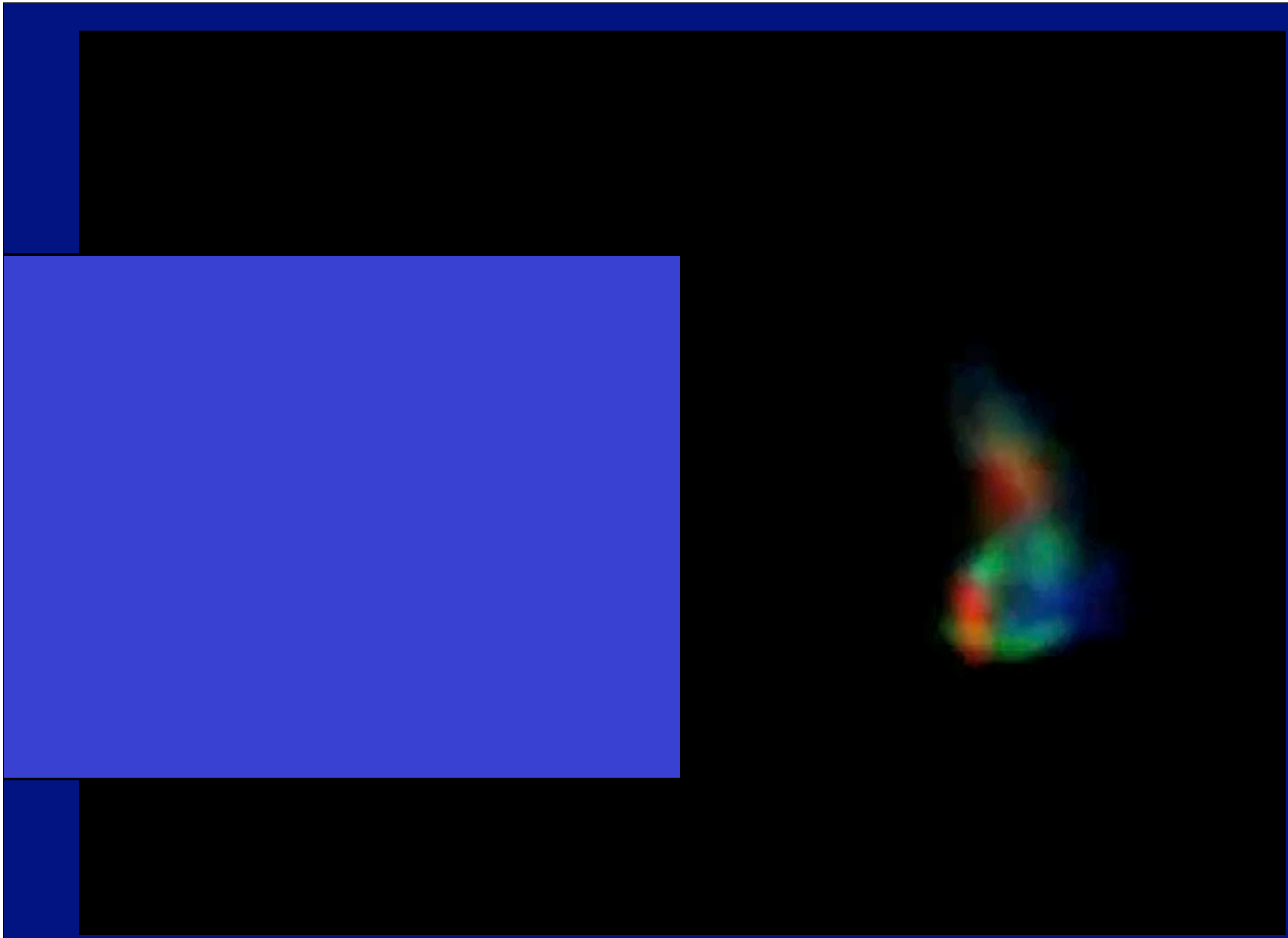
learn limb classifiers

label pixels

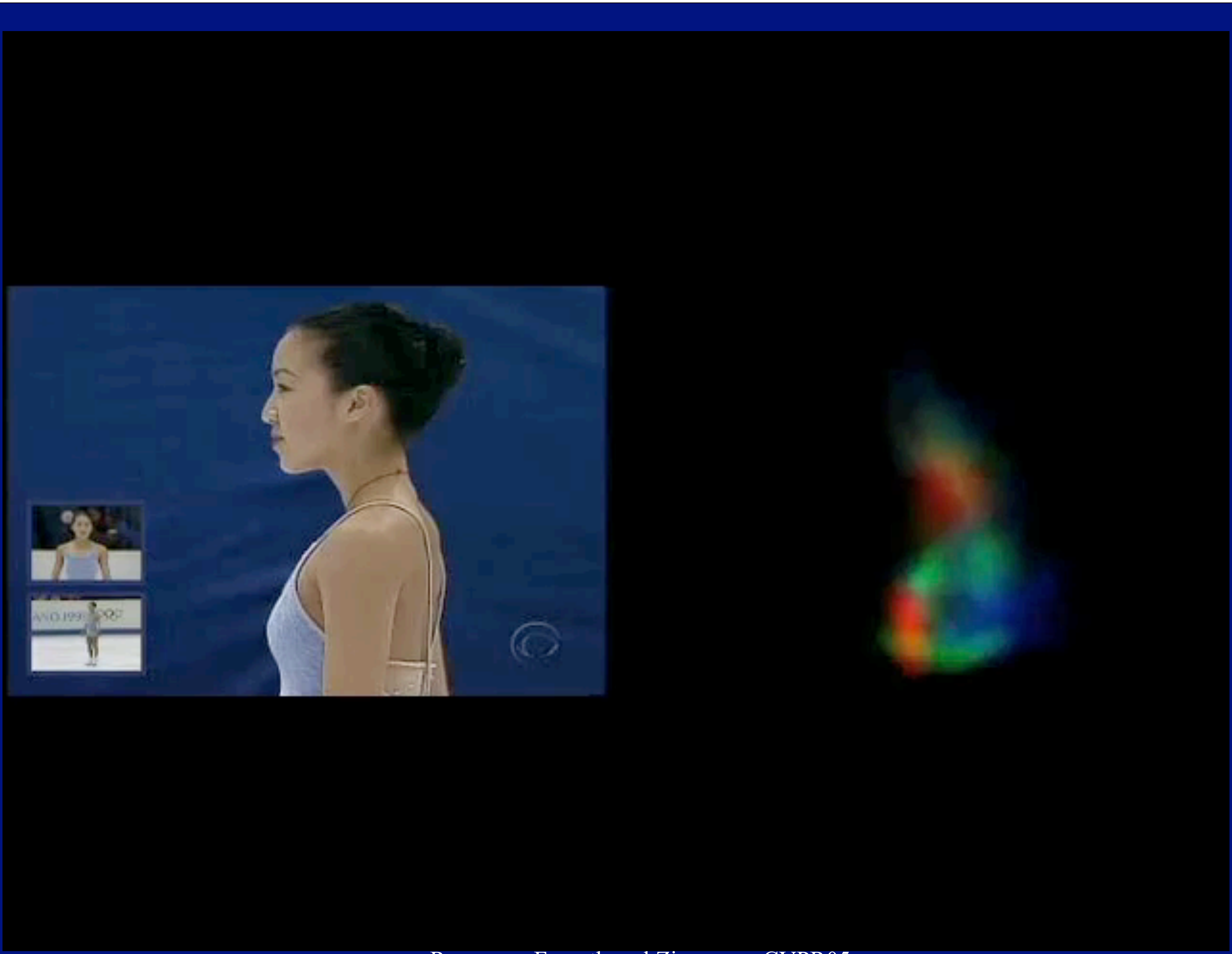
general pose pictorial structure

"Lola" likelihood

Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05



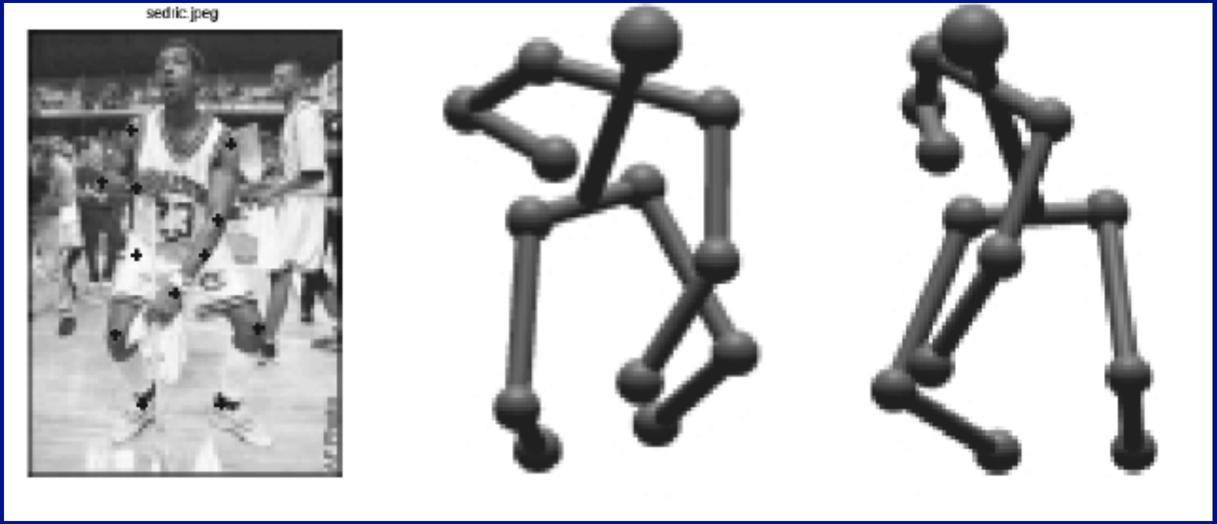
Ramanan, Forsyth and Zisserman CVPR05

Coming to tracking

- Advances in human parsing
  - Appearance/layout interaction (Ramanan 06)
  - Improved appearance models (Ferrari et al 08; Eichner Ferrari 10)
  - Branch+bound (Tian Sclaroff 10)
  - Interactions with objects (Yao Fei-Fei 10; Desai et al 10)
  - Coverage and background (Buehler ea 08; Jiang 09)
  - Full relational models (Tran Forsyth 10)

Lifting

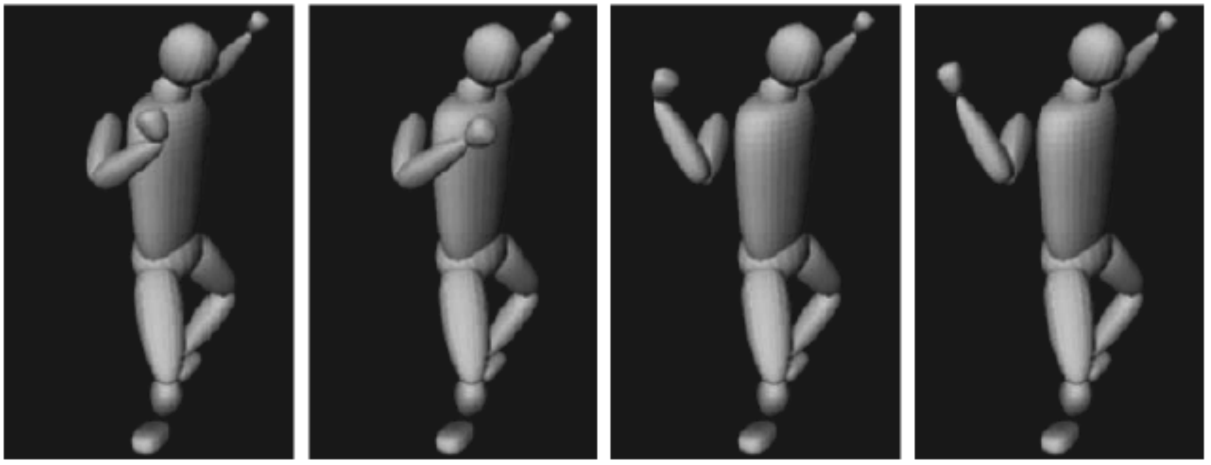
- Infer 3D configuration from image configuration
- Useful for
  - view independent activity recognition
  - user interfaces
  - video motion capture



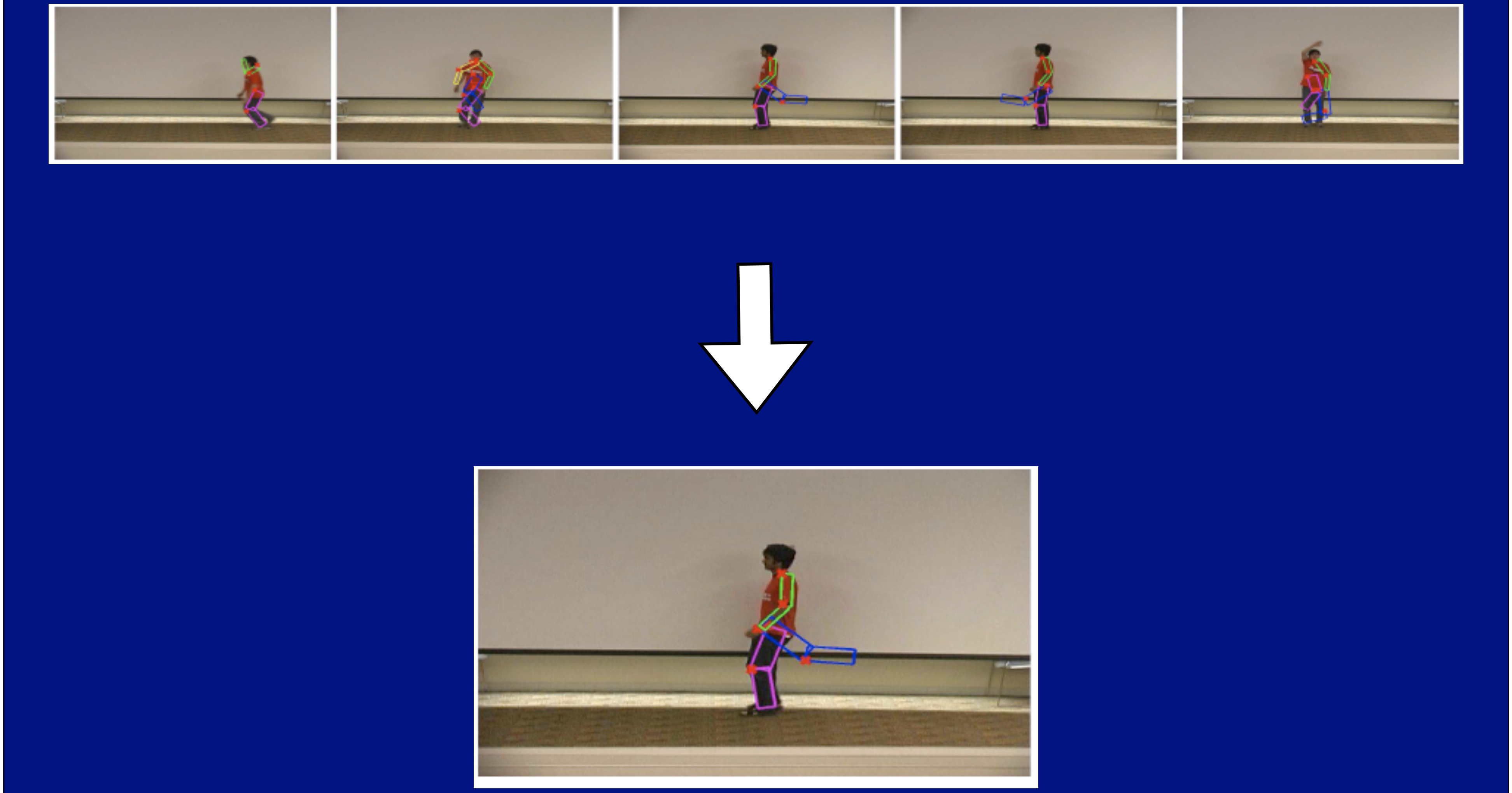
Taylor, 00

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



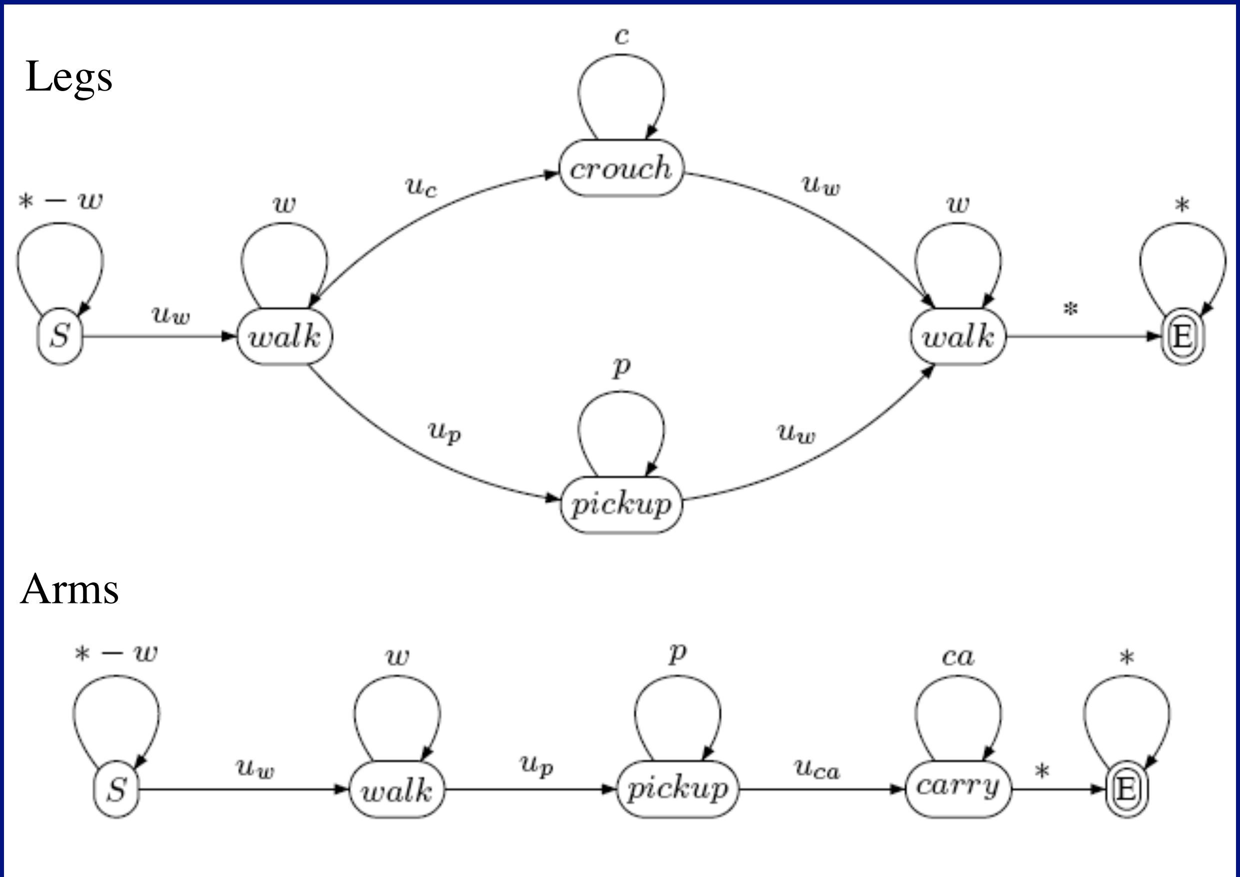
Core questions

- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - style and composition
- How should we transduce the signal?
  - infer body segments or not
- Bias and generalization
  - inevitable problems with complex high dimensional signals

Naming activities

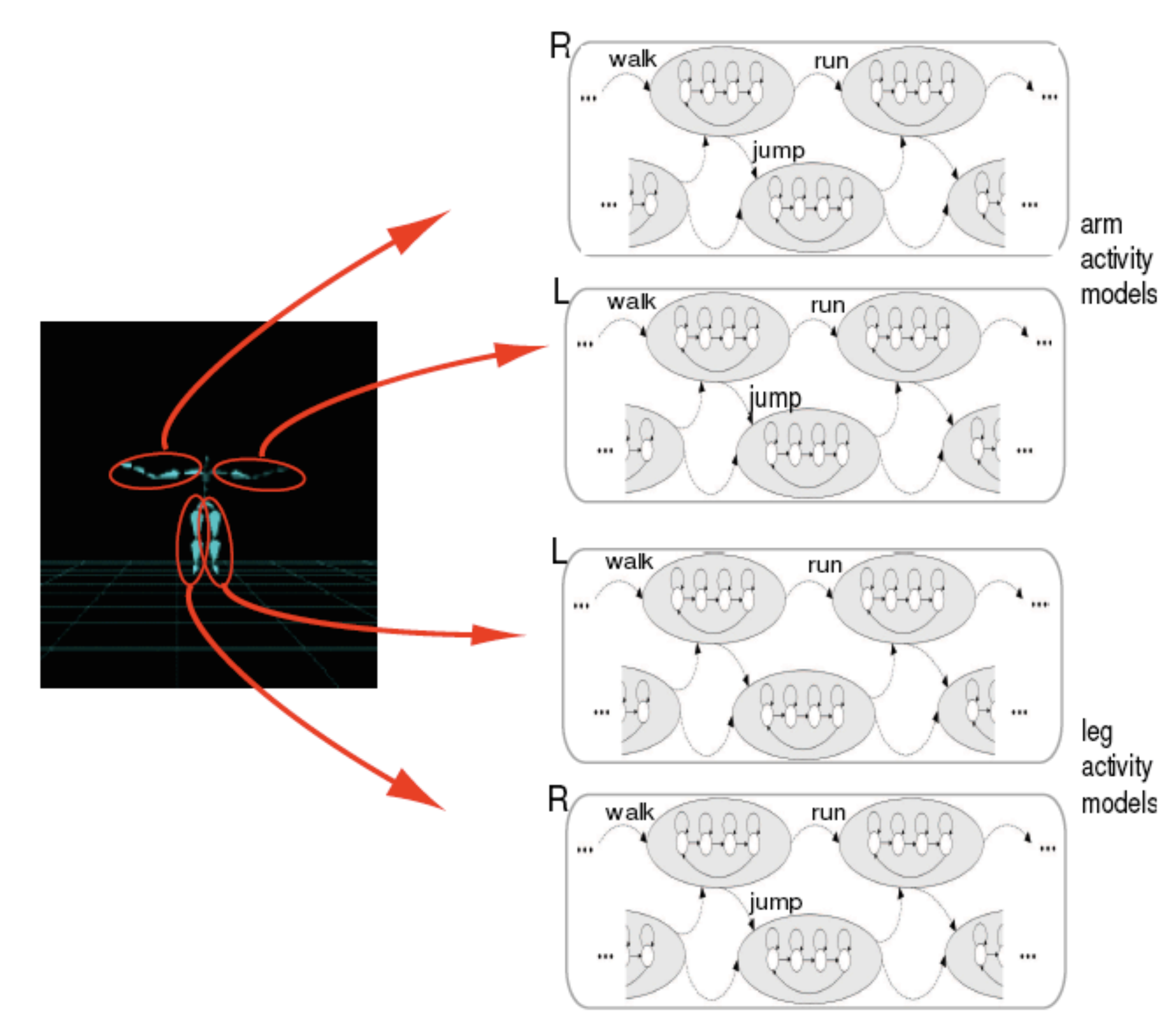
- Build a set of basic labels
  - guess them: walk, run, stand, reach, crouch, etc.
- Composite Activity model:
  - Product of finite state automata for arms, legs built from MoCap
  - Arms, legs each have local short timescale activity models for basic labels
  - Link these models into a large model, using animation-legal transitions

Naming activities

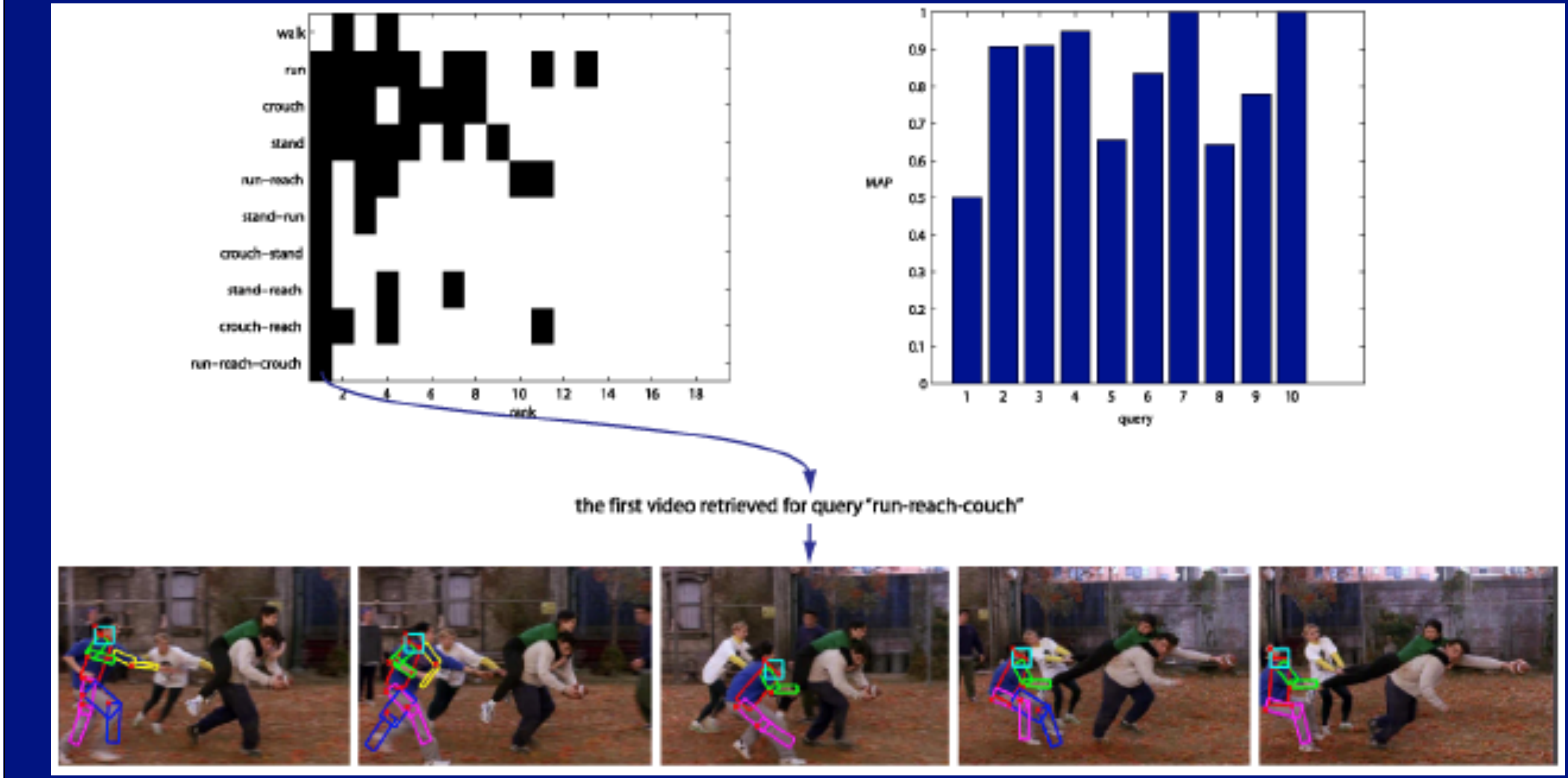


Ikizler Forsyth 07,08

Composition



Ikizler Forsyth 07,08



the first video retrieved for query "run-reach-couch"

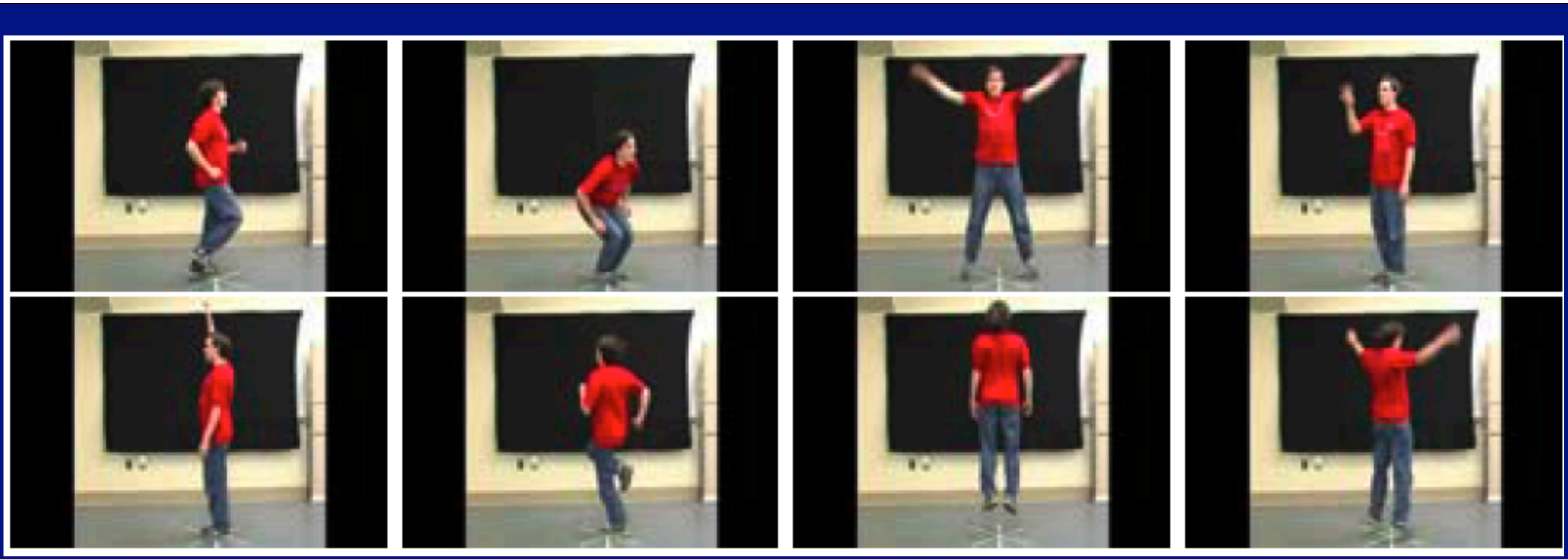
Searching for complex human activities with no visual examples N İkizler, DA Forsyth - IJCV, 2008

Emission

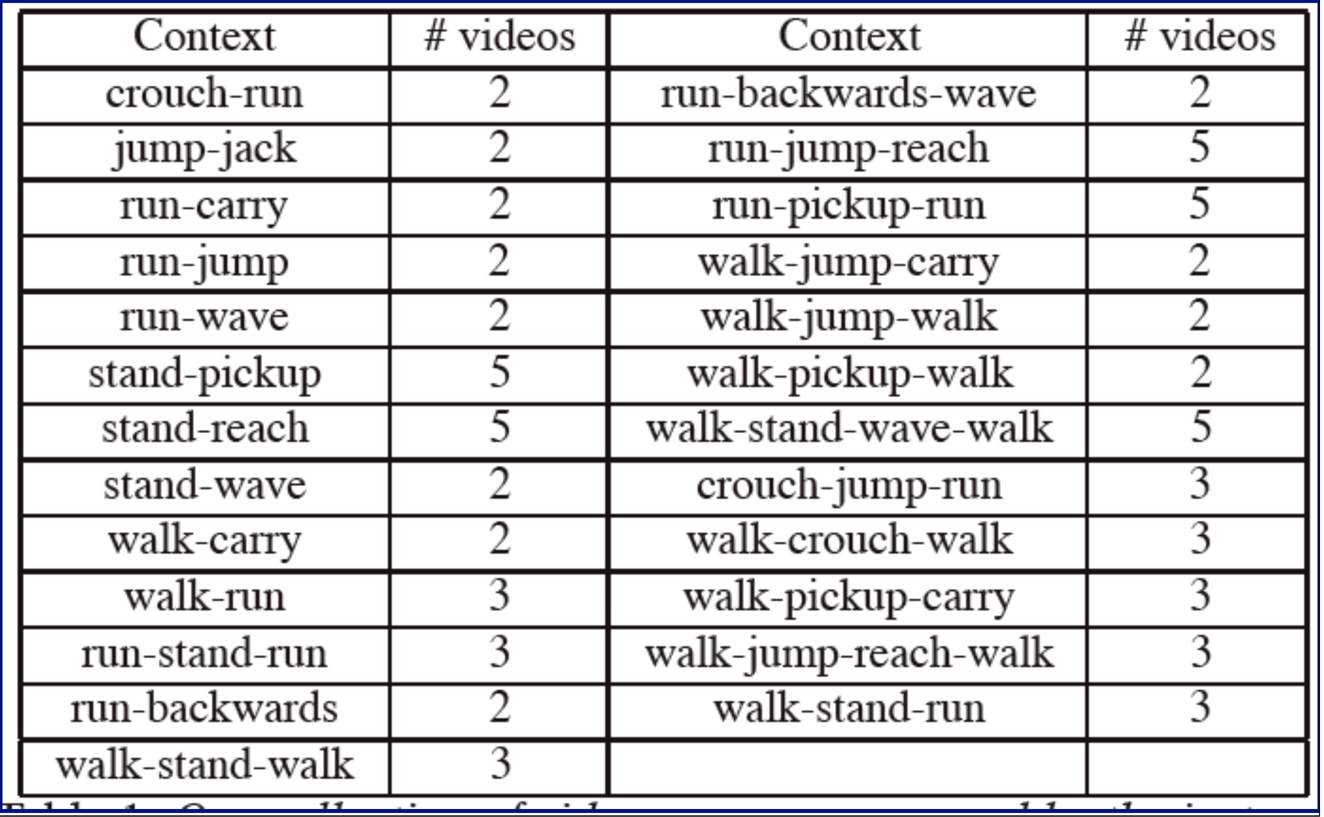
- Transduction
  - Track the body, as above
  - Lift “snippets” of each quarter
    - vector quantized
  - impose root consistency
- Emission
  - emit cluster center from state according to table
  - table learned by EM, known dynamical model

Query for motions with no examples

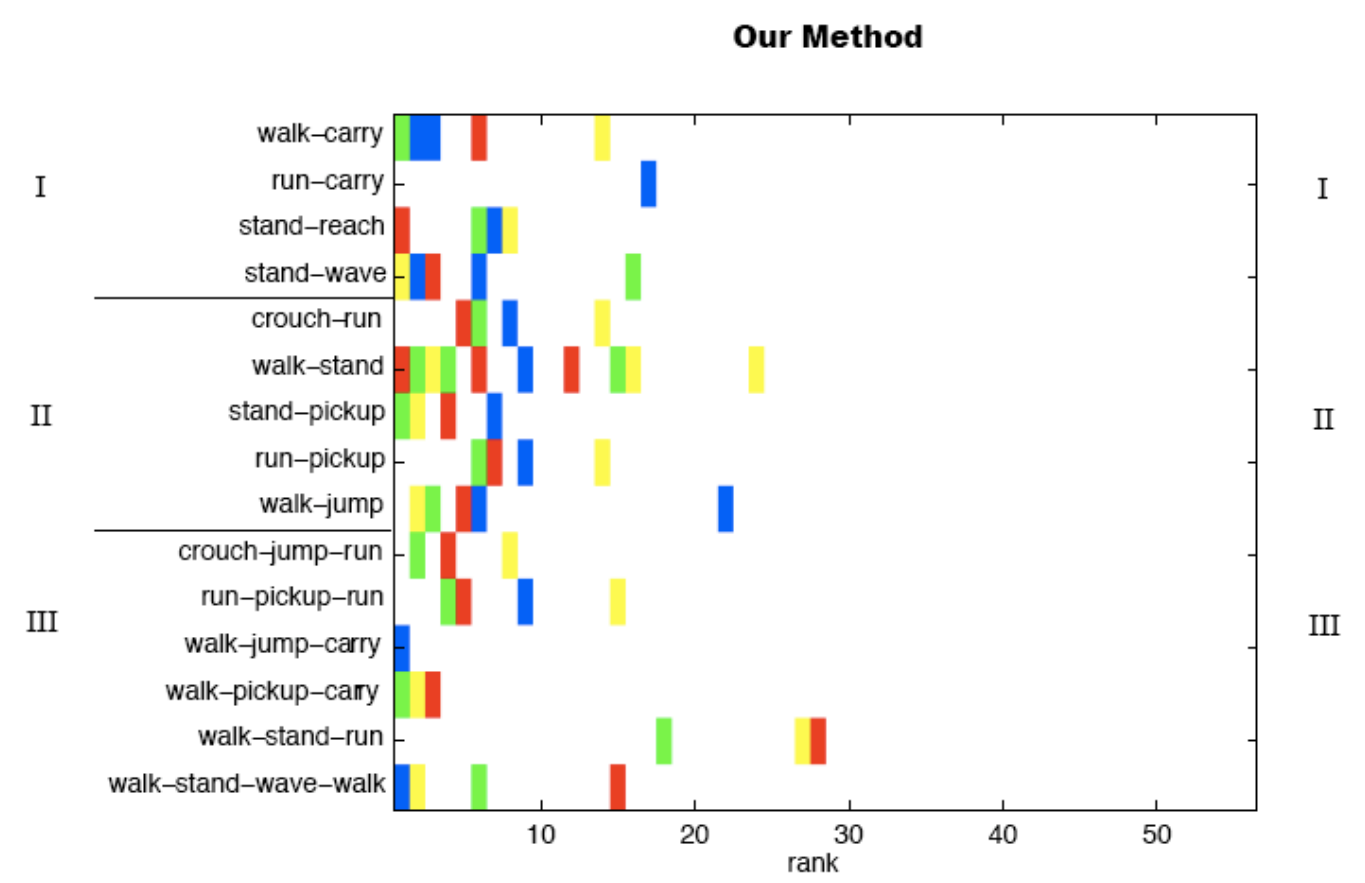
- Primary attraction
  - “natural” query language
- Rank sequences by P(FSA|data, model)
  - e.g. P(leg-walk-arm-walk-then-leg-walk-arm-reach| data, model)
  - DP variant will do this easily



Ikizler Forsyth 07,08

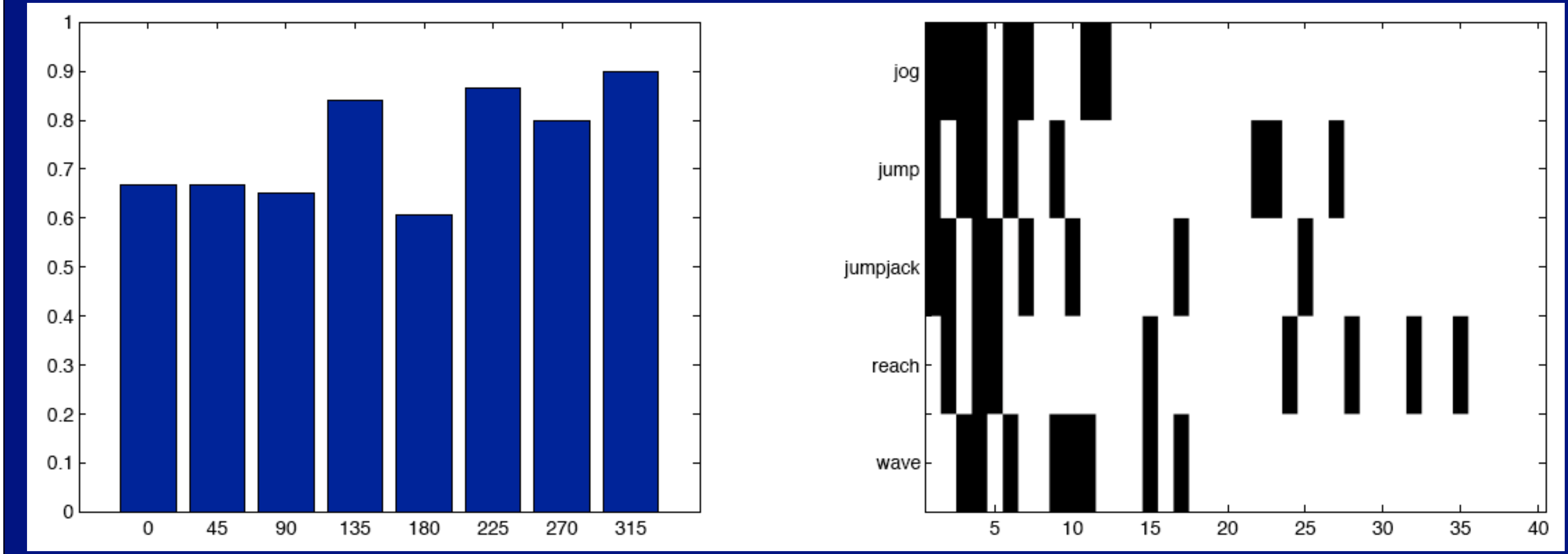


Context	# videos	Context	# videos
crouch-run	2	run-backwards-wave	2
jump-jack	2	run-jump-reach	5
run-carry	2	run-pickup-run	5
run-jump	2	walk-jump-carry	2
run-wave	2	walk-jump-walk	2
stand-pickup	5	walk-pickup-walk	2
stand-reach	5	walk-stand-wave-walk	5
stand-wave	2	crouch-jump-run	3
walk-carry	2	walk-crouch-walk	3
walk-run	3	walk-pickup-carry	3
run-stand-run	3	walk-jump-reach-walk	3
run-backwards	2	walk-stand-run	3
walk-stand-walk	3		



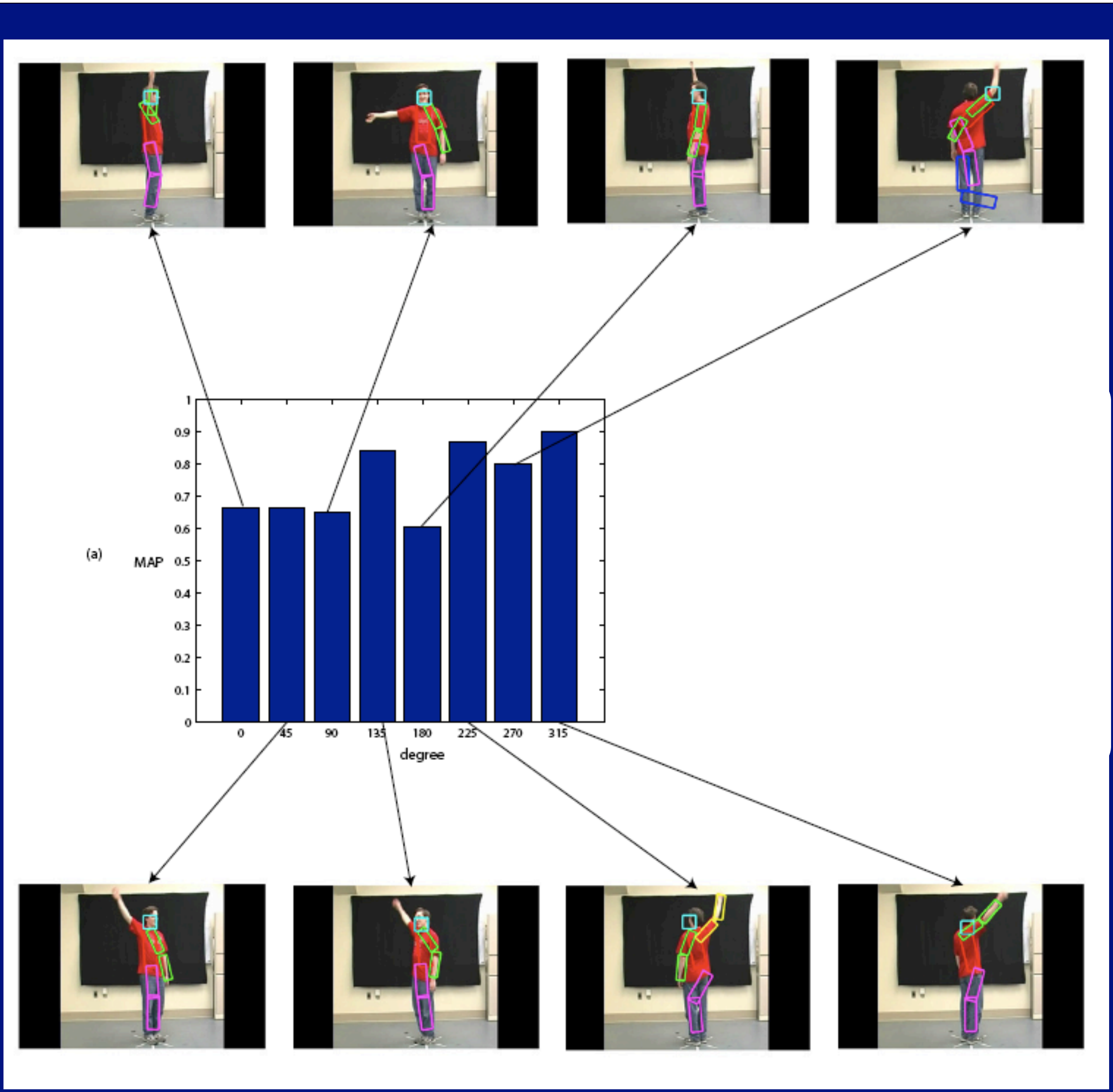
Ikizler Forsyth 07,08

The effect of aspect

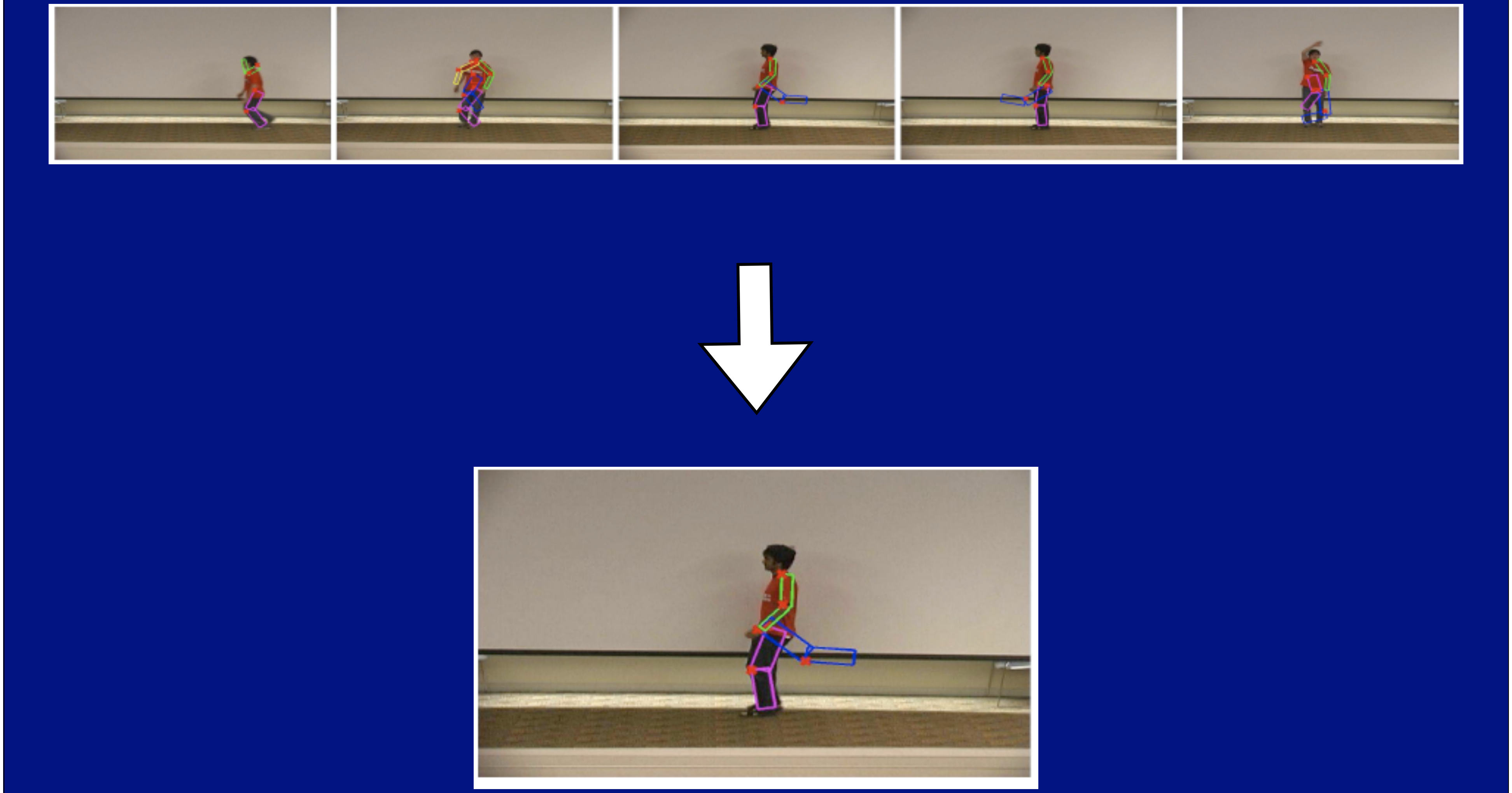


Jog; Jump; Jumpjack; Reach; Wave

Ikizler Forsyth 07, 08



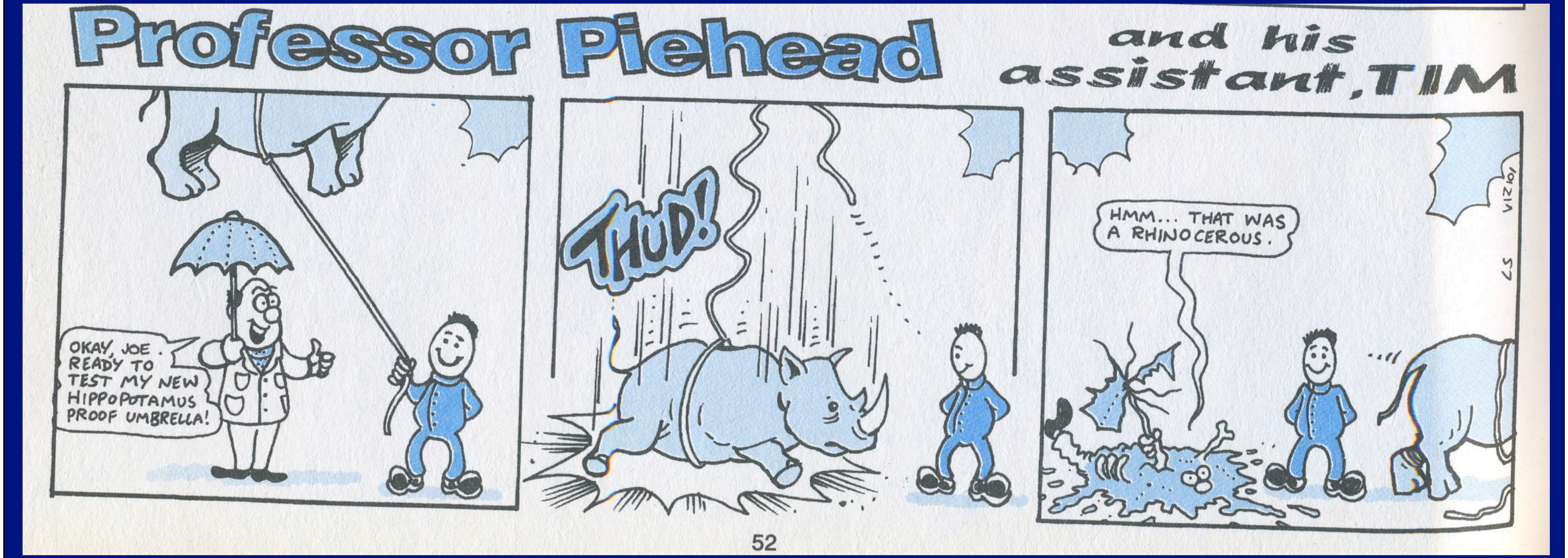
Ikizler Forsyth 07, 08



Core questions

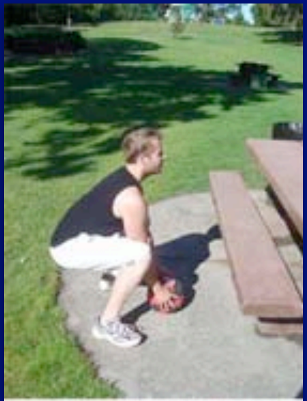
- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - style and composition
- How should we transduce the signal?
  - infer body segments or not
- Bias and generalization
  - inevitable problems with complex high dimensional signals

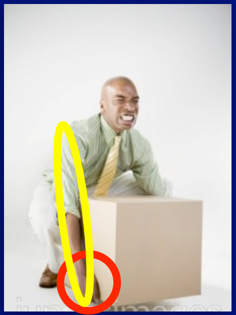
What is an object like?

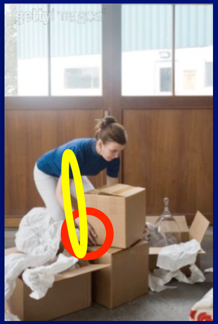

Professor Piehead and his assistant, TIM
OKAY, JOE. READY TO TEST MY NEW HIPPOPOTAMUS PROOF UMBRELLA!
THUD!
HMM... THAT WAS A RHINOCEROUS.
VIZ 101
LS
52

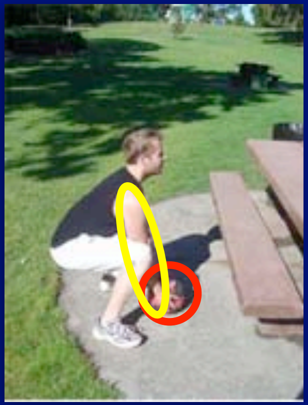
Viz comic, issue 101

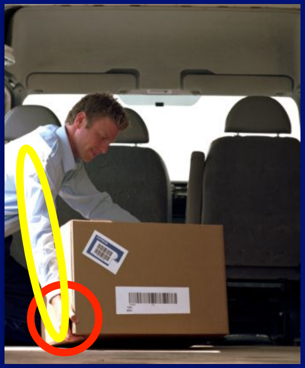
ROCCOS
GREEN BEANS
Josi











Activity attributes

- Fast/Gentle
- Clumsy/adroit

- Having hand contact

- Arms sticking out

Bias affects representation

- Other kinds of semantics
  - Ramanan’s activity example
    - where you are often reveals what you are doing
    - but how do we encode where you are
      - x-y coords?
      - near the stove?

Thanks

UIUC Vision & Graphics groups
UC Berkeley Vision & Graphics Groups
Oxford Visual Geometry Group, particularly Andrew Zisserman
Dept. Homeland Security
ONR MURI
NSF
Electronic Arts
Sony Computer Entertainment