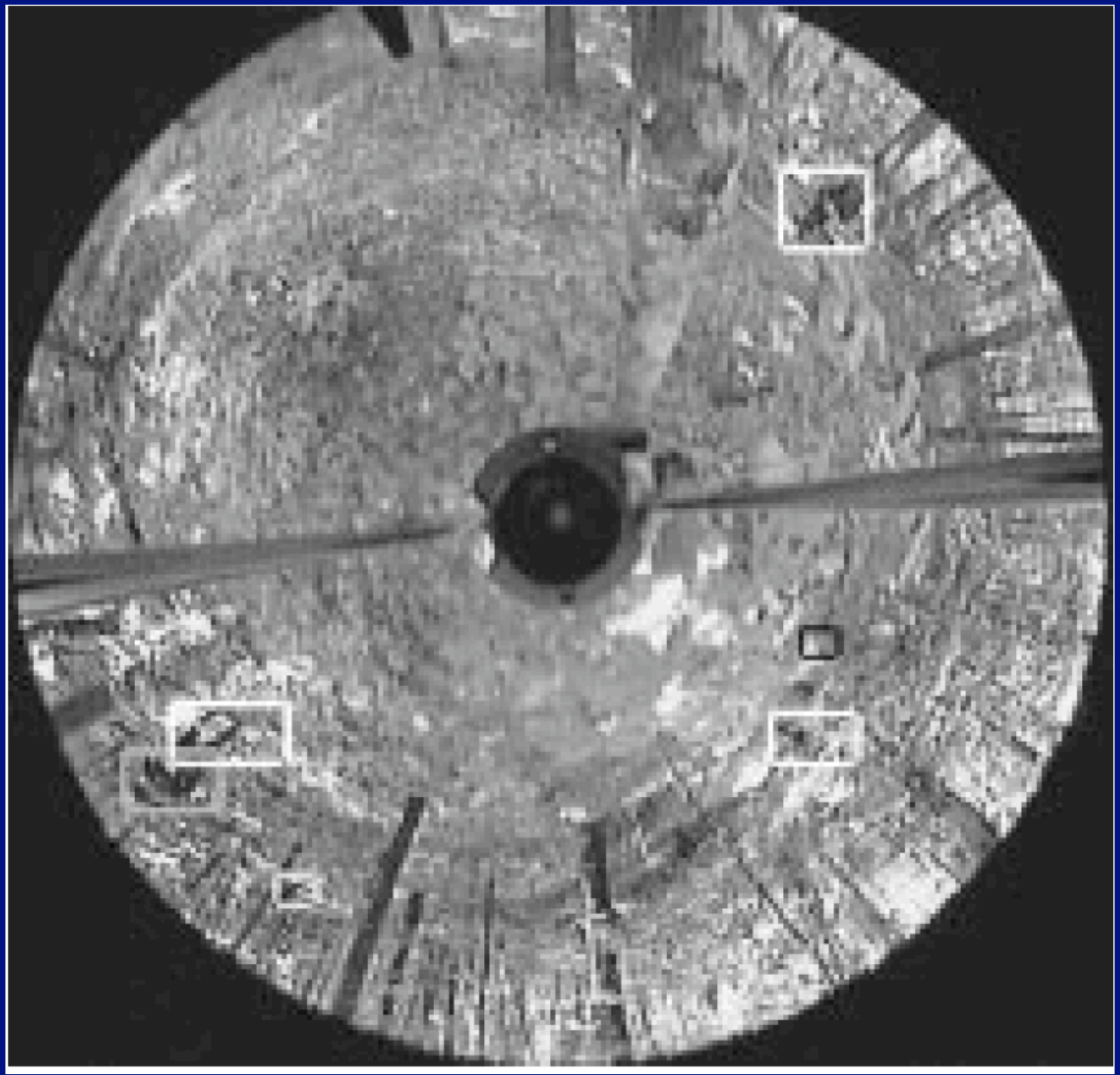
Activity and Kinematics

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 U. Washington (was UIUC)
Nazli Ikizler of Bilkent U (now Boston U; soon Hacettepe U)
Alex Sorokin, Crowdflower, Du Tran, NTU Duan Tran, UIUC, Wei Yan, Texas A+M
Yang Wang (was UIUC, now Edmonton)
Daphne Tsatsoulis, UIUC

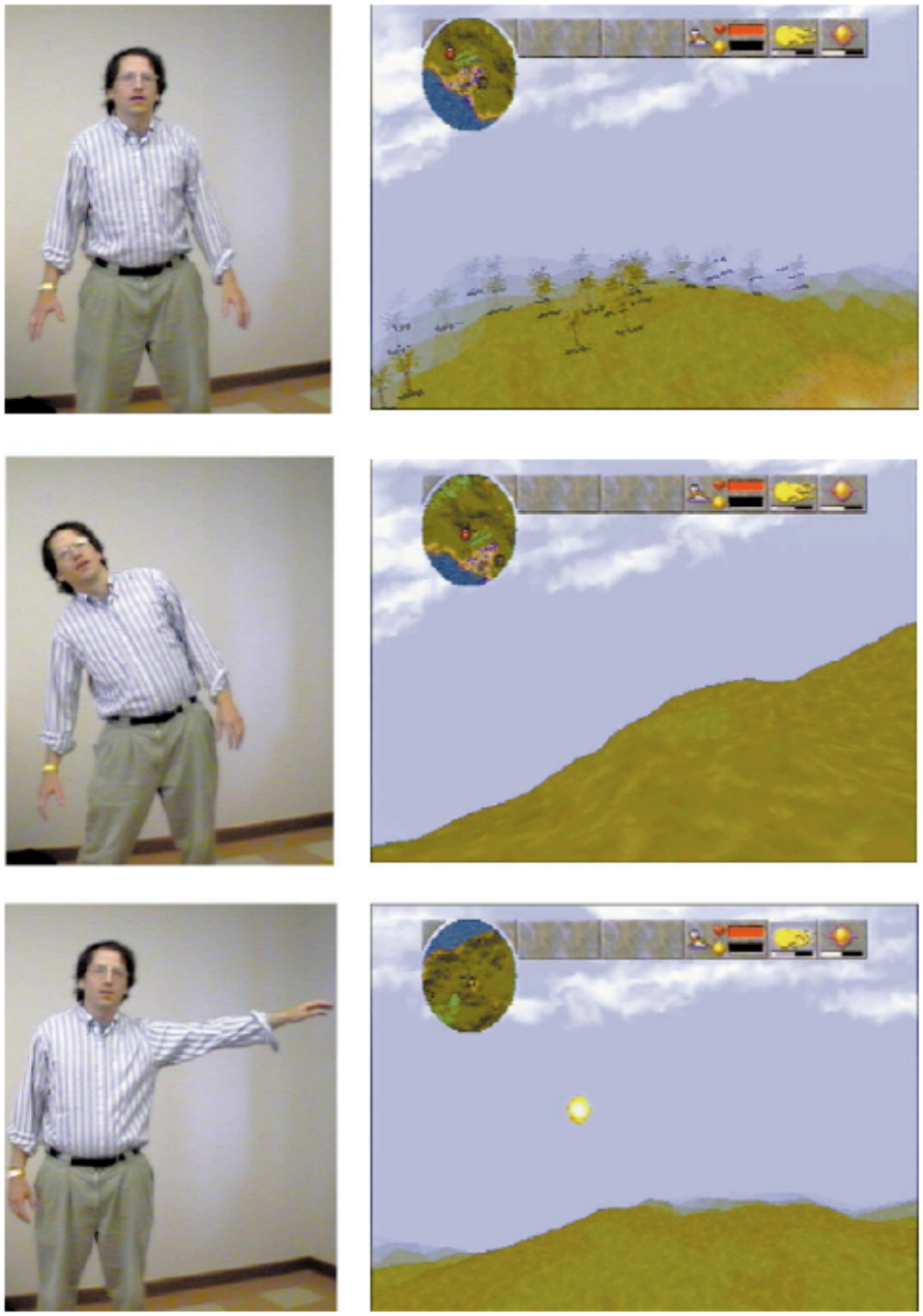
Thanks to: Electronic Arts, Sony SCEA, ONR MURI, NSF, DHS

Why are humans important?

- Surveillance
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- Synthesis
  - games; movies;
- Safety applications
  - pedestrian detection
- People are interesting
  - movies; news



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



Bill Freeman flies a magic carpet.

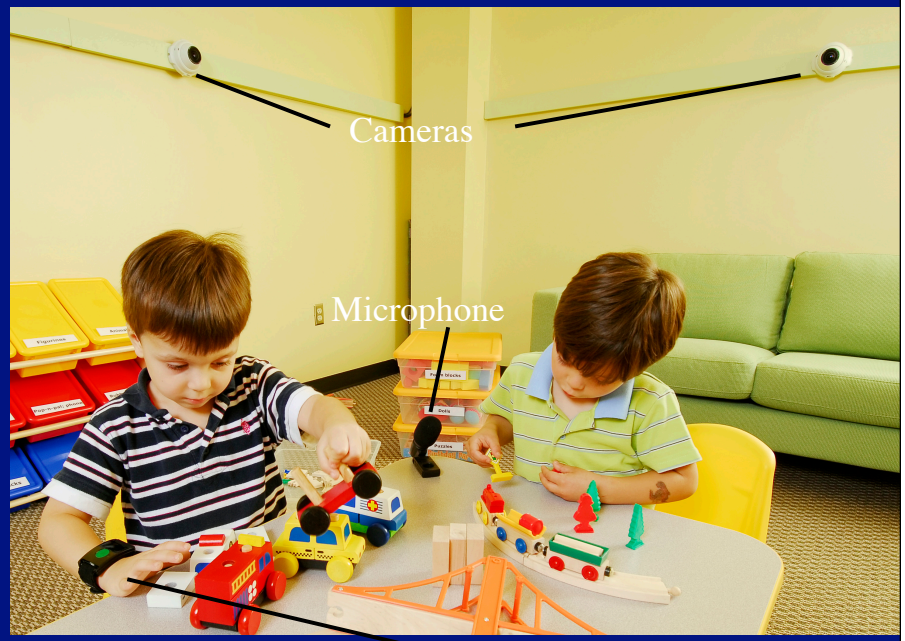
Orientation histograms detect body configuration to control bank, raised arm to fire magic spell.

Freeman et al, 98.

Idea becomes:

Eyetoy
WII
Eyetoy 2
Kinect

Computational Behavioural Science

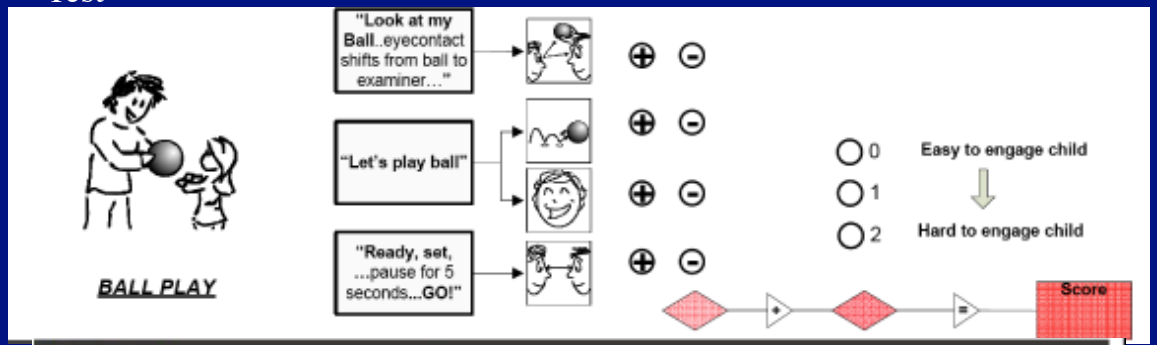


- Observe people
  - Using vision, physiological markers
  - Interacting, behaving naturally
  - In the wild
- drive feedback for therapy
  - Eg reward speech
- Applications
  - Model: screen for ASD
  - Other:
    - Any w here large scale observations help
      - Support in home care
      - Support care for demented patients
      - Support stroke recovery
      - Support design of efficient buildings
- 10M$, 5yr NSF award under Expeditions program
  - GaTech, UIUC(DAF, Karahalios), MIT, CMU, Pittsburgh, USC, Boston U

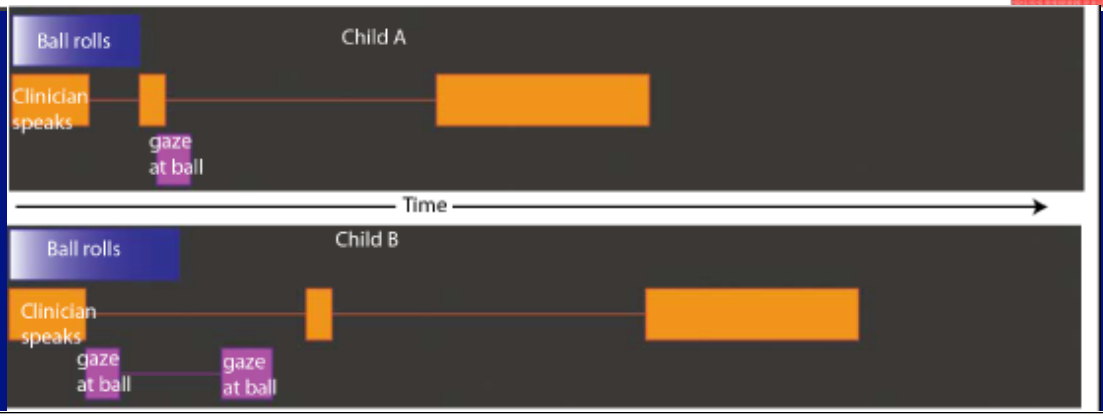
Rapid ABC

- Easily administered screening test
  - Challenge:
    - Automatic evaluation
    - To use unskilled screeners

Test

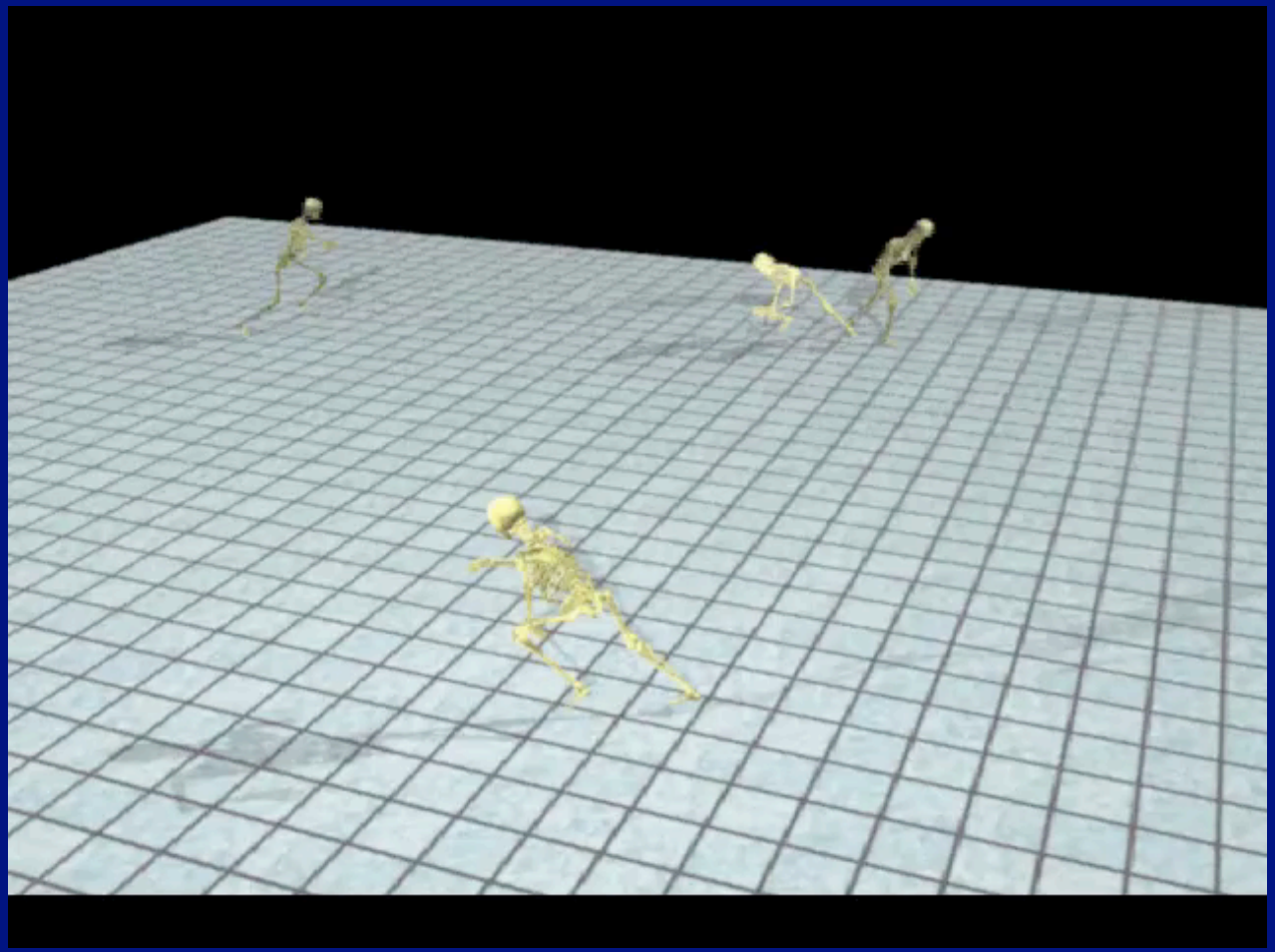


Outcome s



Why are humans important?

- Surveillance
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- Synthesis
  - games; movies;
- Safety applications
  - pedestrian detection
- People are interesting
  - movies; news

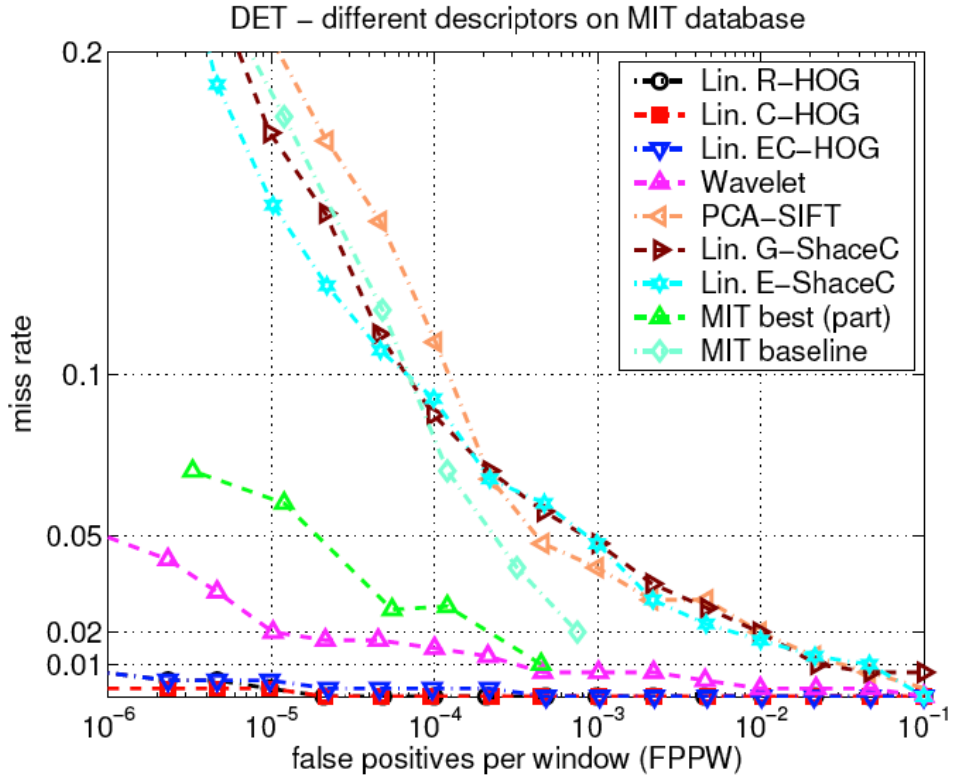


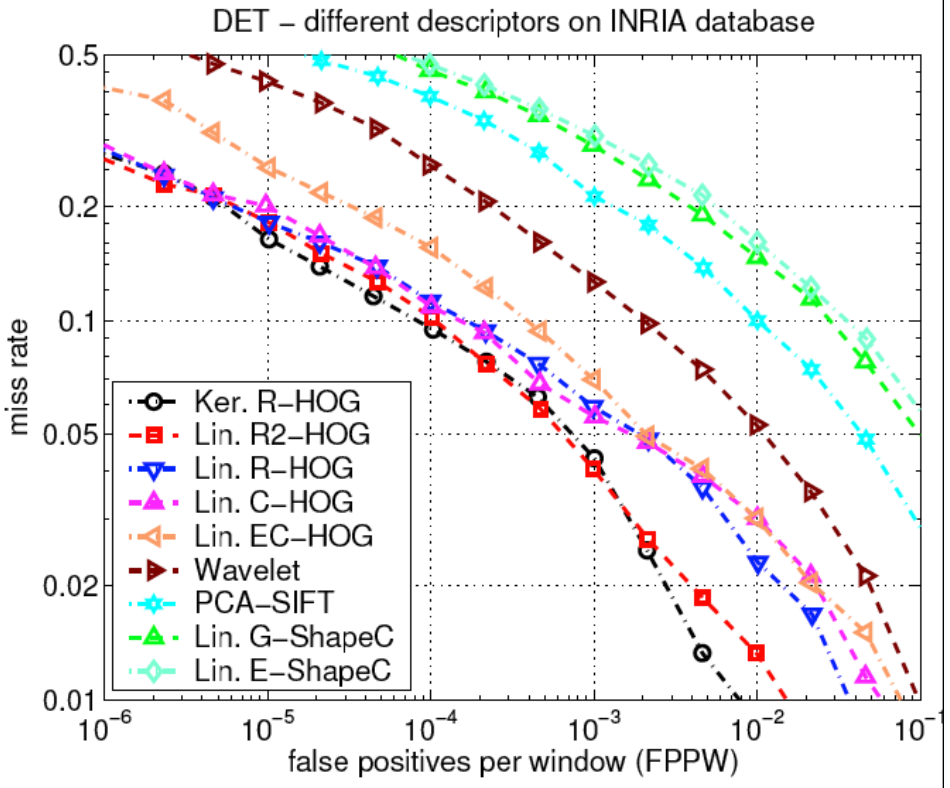
Arikan+Forsyth 02

Why are humans important?

- Surveillance
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- Synthesis
  - games; movies;
- Safety applications
  - pedestrian detection
- People are interesting
  - movies; news

From Dalal+Triggs, 05





Why are humans important?

- Surveillance
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- Synthesis
  - games; movies;
- Safety applications
  - pedestrian detection
- People are interesting
  - movies; news

News Faces

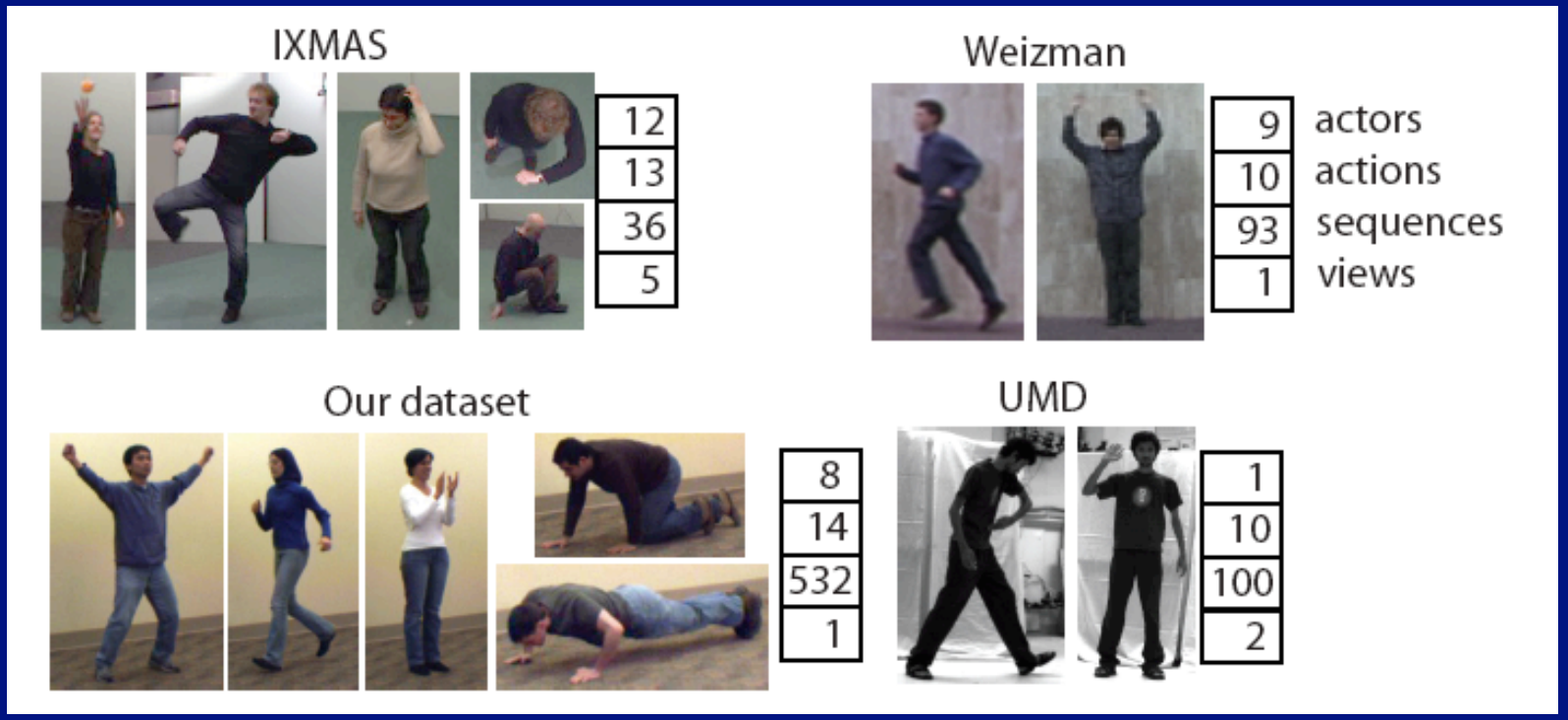
- 5e5 captioned news images
- Mainly people "in the wild"
- Correspondence problem
  - some images have many (resp. few) faces, few (resp. many) names (cf. Srihari 95)
- Process
  - Extract proper names
  - Detect faces (Vogelhuber Schmid 00) 44773 big face responses
  - Rectify faces 34623 properly rectified
  - Kernel PCA rectified faces
  - Estimate linear discriminants
  - Now have (face vector; name_1,...., name_k) 27742 for k<=4
- Apply a form of modified k-means

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/ Reuters

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

Datasets


IXMAS
12
13
36
5
Weizman
9 actors
10 actions
93 sequences
1 views
Our dataset
8
14
532
1
UMD
1
10
100
2

Discriminative results

Dataset	Algorithm	Chance	Protocols								
			Discriminative task				Reject	Few examples			
			L1SO	L1AAO	L1AO	L1VO	UNa	FE-1	FE-2	FE-4	FE-8
Weizman	NB(k=300)	10.00	91.40	93.50	95.70	N/A	0.00	N/A	N/A	N/A	N/A
	1NN	10.00	95.70	95.70	96.77	N/A	0.00	53.00	73.00	89.00	96.00
	1NN-M	10.00	100.00	100.00	100.00	N/A	0.00	72.31	81.77	92.97	100.00
	1NN-R	9.09	83.87	84.95	84.95	N/A	84.95	17.96	42.04	68.92	84.95
	1NN-MR	9.09	**89.66**	**89.66**	**89.66**	N/A	**90.78**	N/A	N/A	N/A	N/A
Our	NB(k=600)	7.14	98.70	98.70	98.70	N/A	0.00	N/A	N/A	N/A	N/A
	1NN	7.14	98.87	97.74	98.12	N/A	0.00	58.70	76.20	90.10	95.00
	1NN-M	7.14	**99.06**	**97.74**	**98.31**	N/A	0.00	88.80	94.84	95.63	98.86
	1NN-R	6.67	95.86	81.40	82.10	N/A	81.20	27.40	37.90	51.00	65.00
	1NN-MR	6.67	**98.68**	**91.73**	**91.92**	N/A	**91.11**	N/A	N/A	N/A	N/A
IXMAS	NB(k=600)	7.69	80.00	78.00	79.90	N/A	0.00	N/A			
	1NN	7.69	81.00	75.80	80.22	N/A	0.00				
	1NN-R	7.14	**65.41**	**57.44**	**57.82**	N/A	**57.48**				
UMD	NB(k=300)	10.00	100.00	N/A	N/A	97.50	0.00	N/A			
	1NN	10.00	100.00	N/A	N/A	97.00	0.00				
	1NN-R	9.09	**100.00**	N/A	N/A	**88.00**	**88.00**				

Works well, depending on task; not rejecting improves things
metric learning improves things

Tran + Sorokin 08

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

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?

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

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

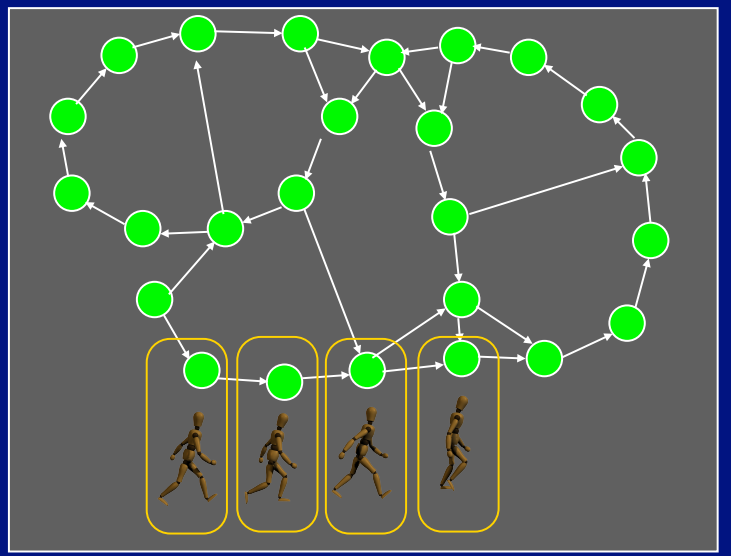
- Motion graph: by analogy with
  - text synthesis, texture synthesis, video textures
- Take measured frames of motion as nodes
  - from motion capture, given us by our friends
- Edge from frame to any that could succeed it
  - decide by dynamical similarity criterion
  - see also (Kovar et al 02; Lee 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

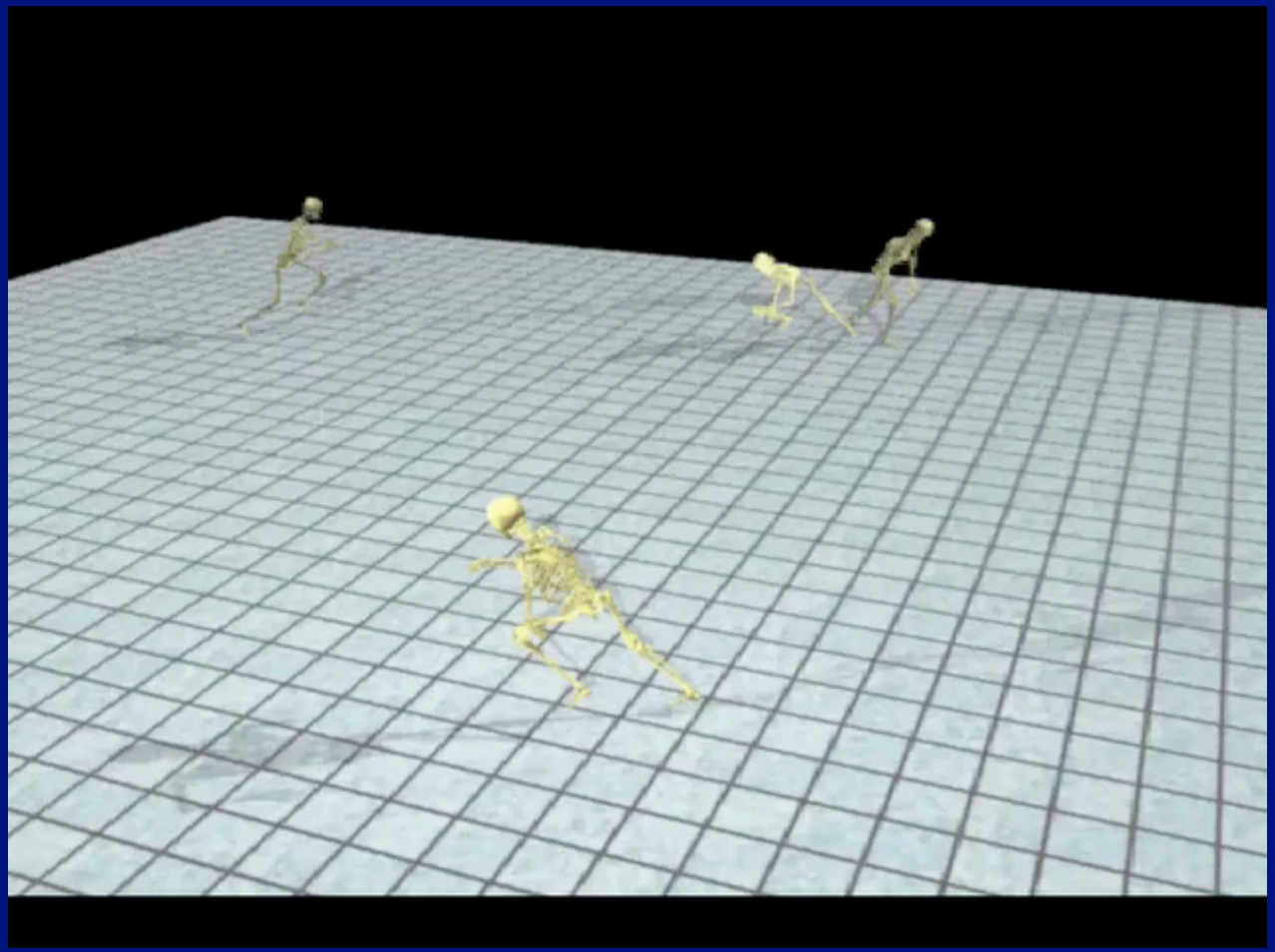
Motion Graph:

Nodes = Frames

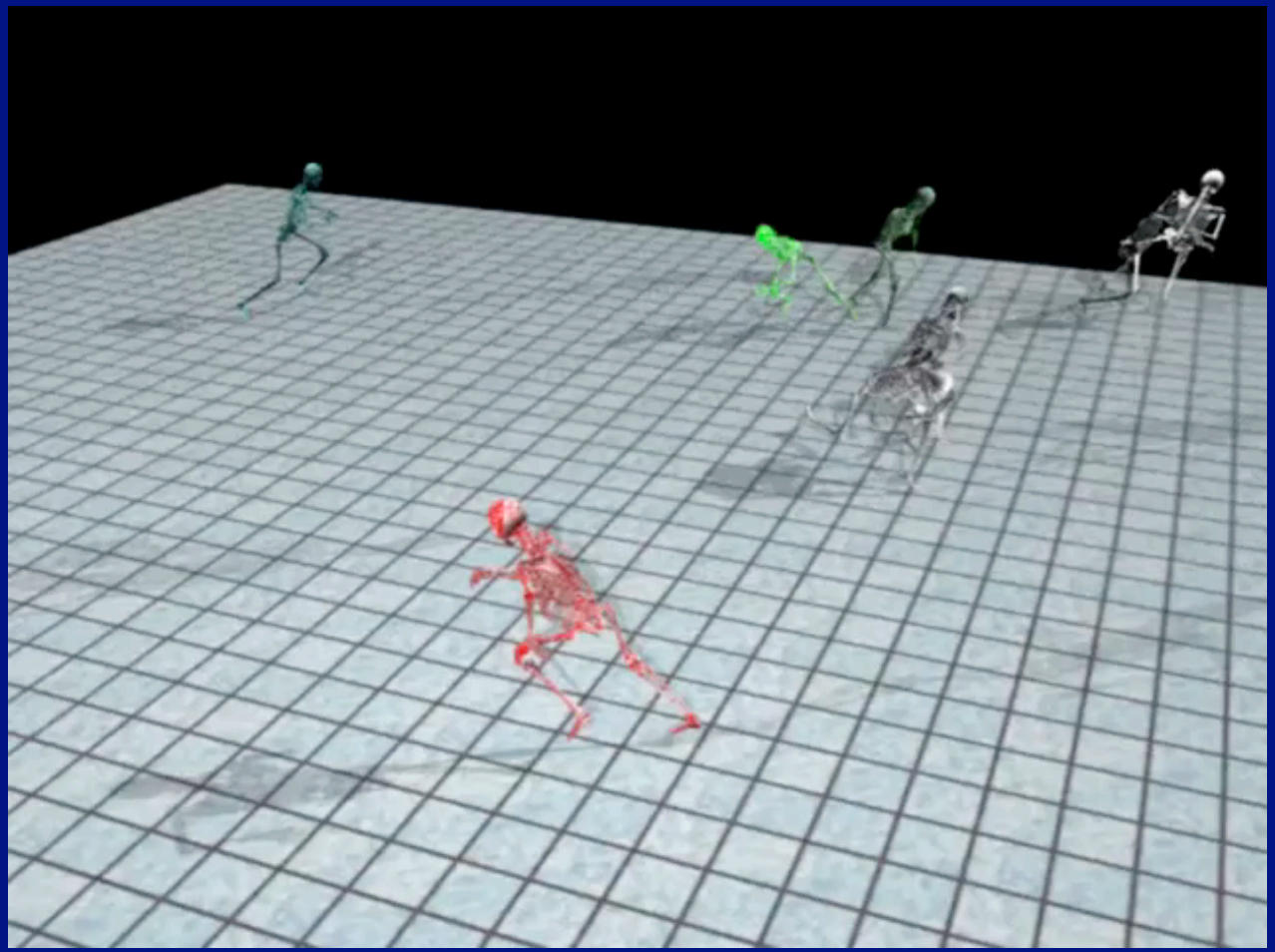
Edges = Transition

A path = A motion

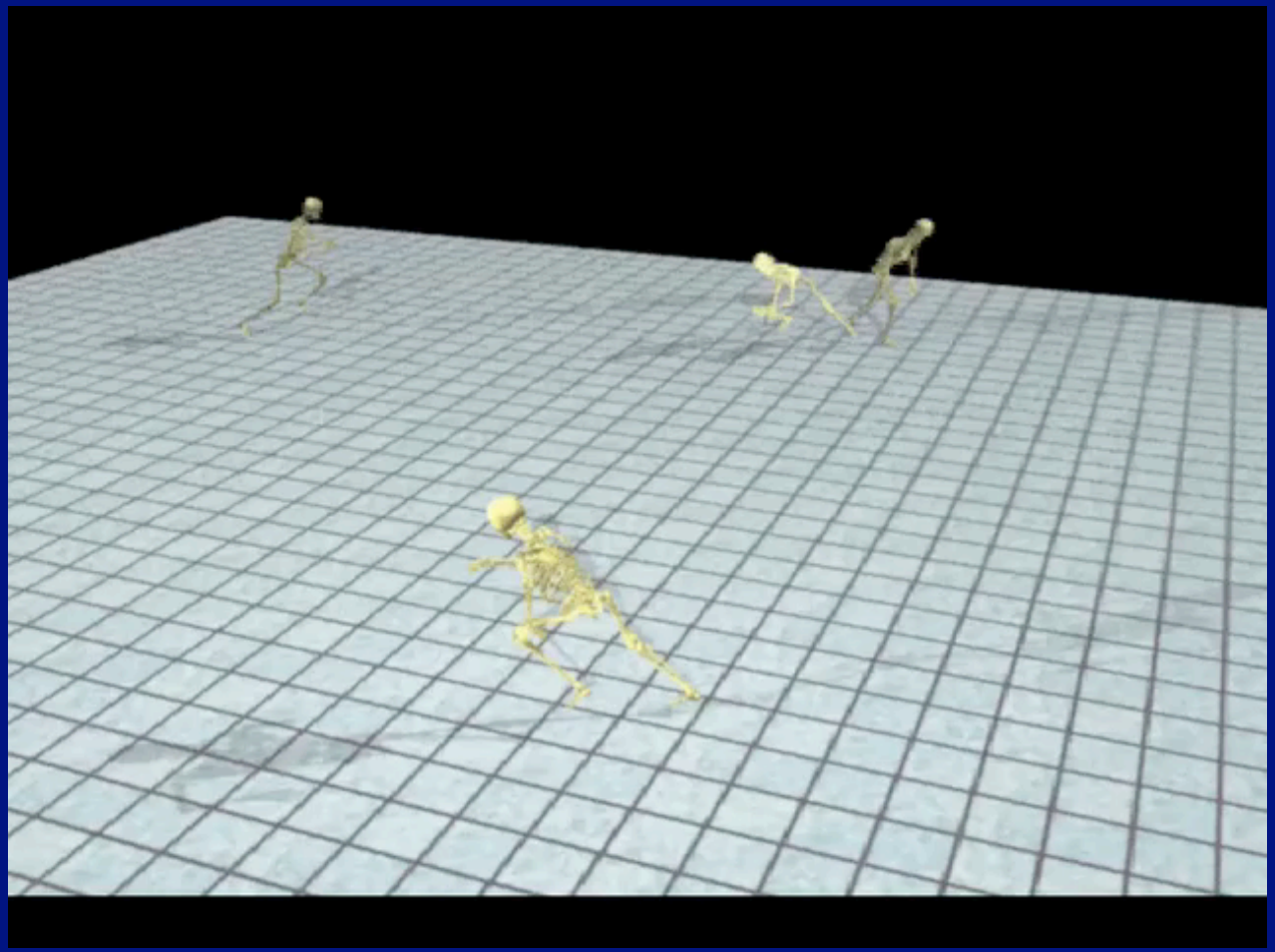




Arikan+Forsyth 02



Arikan+Forsyth 02

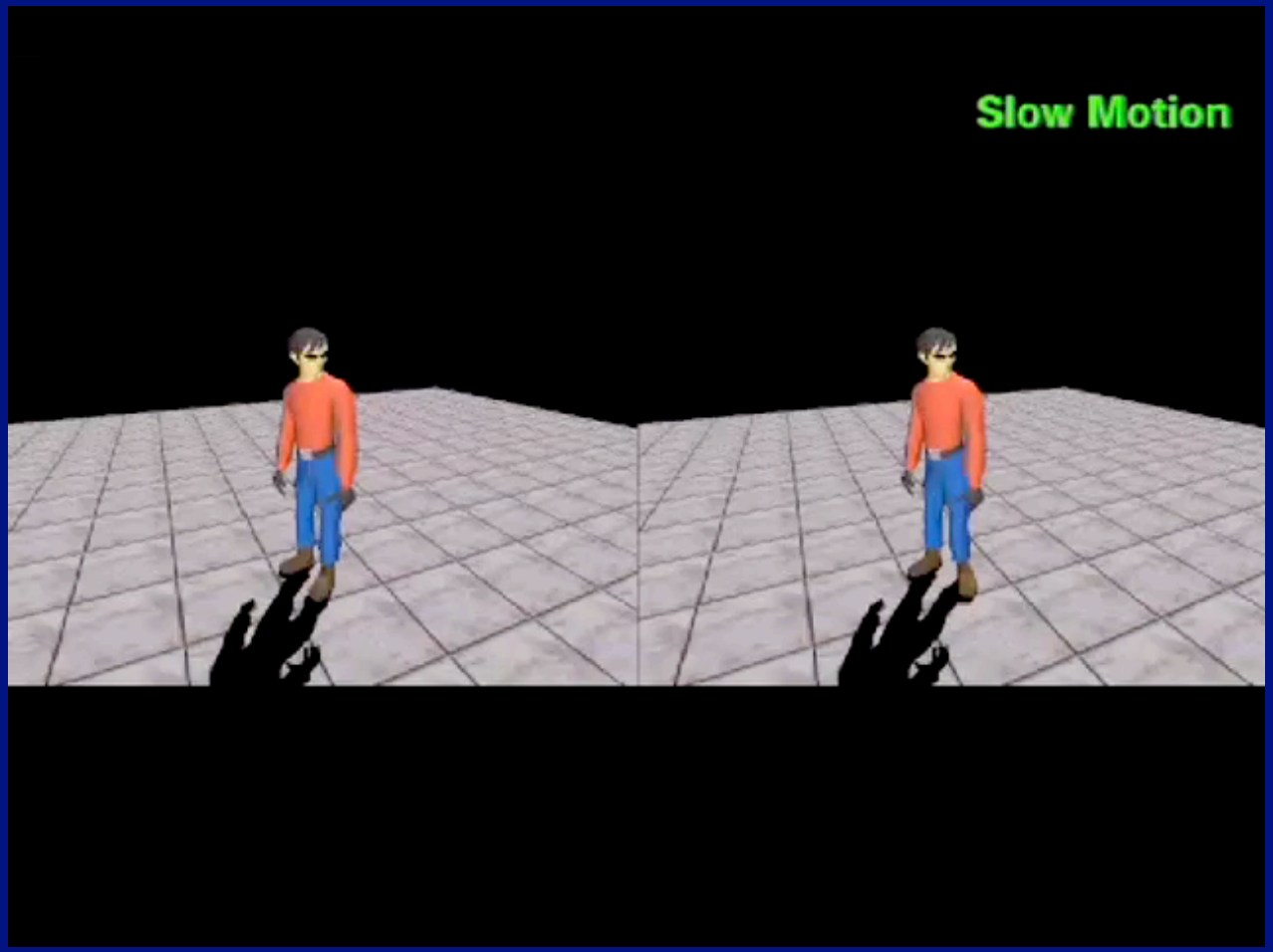


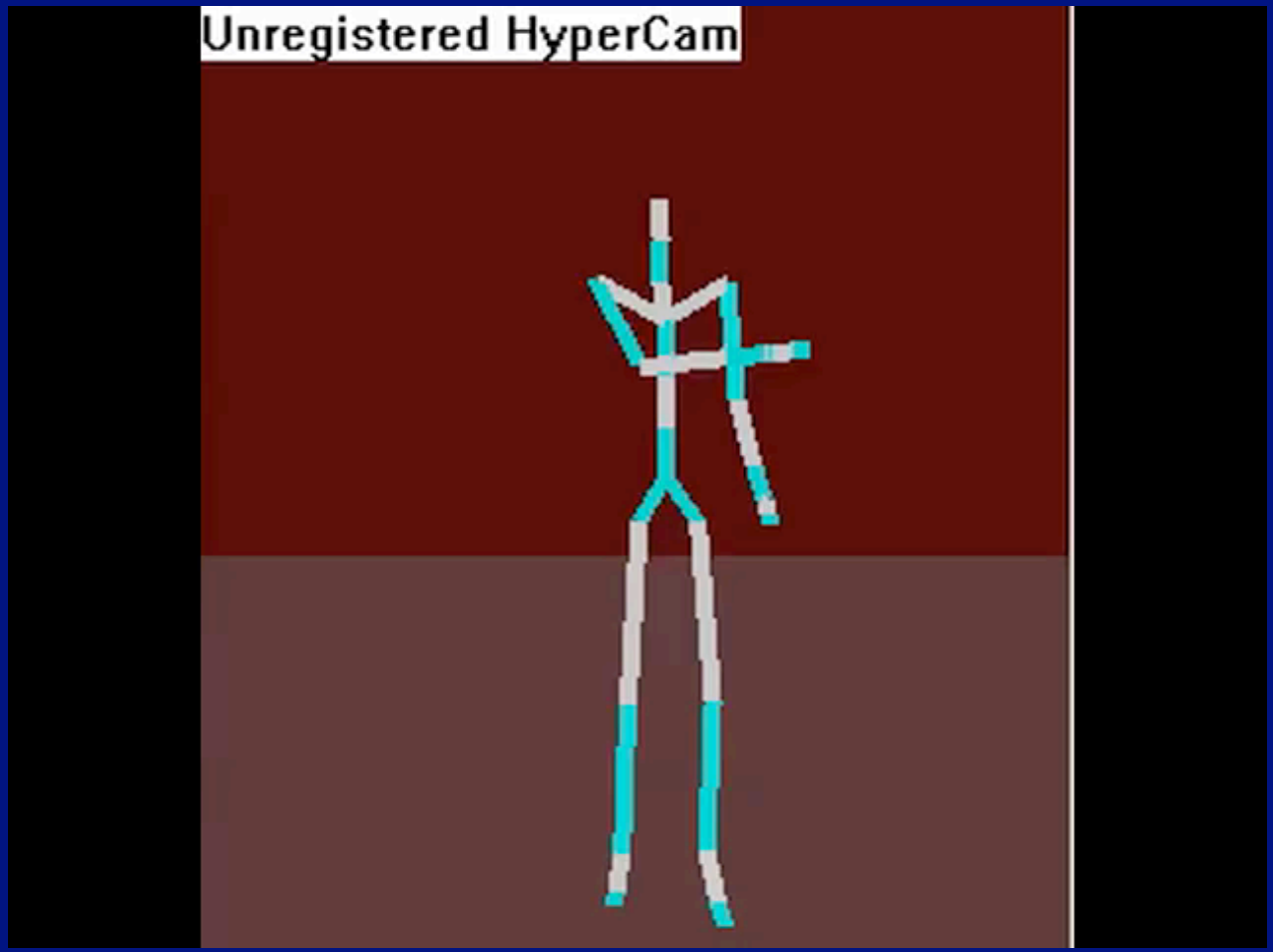
Arikan+Forsyth 02

Non data-driven methods don’t work yet

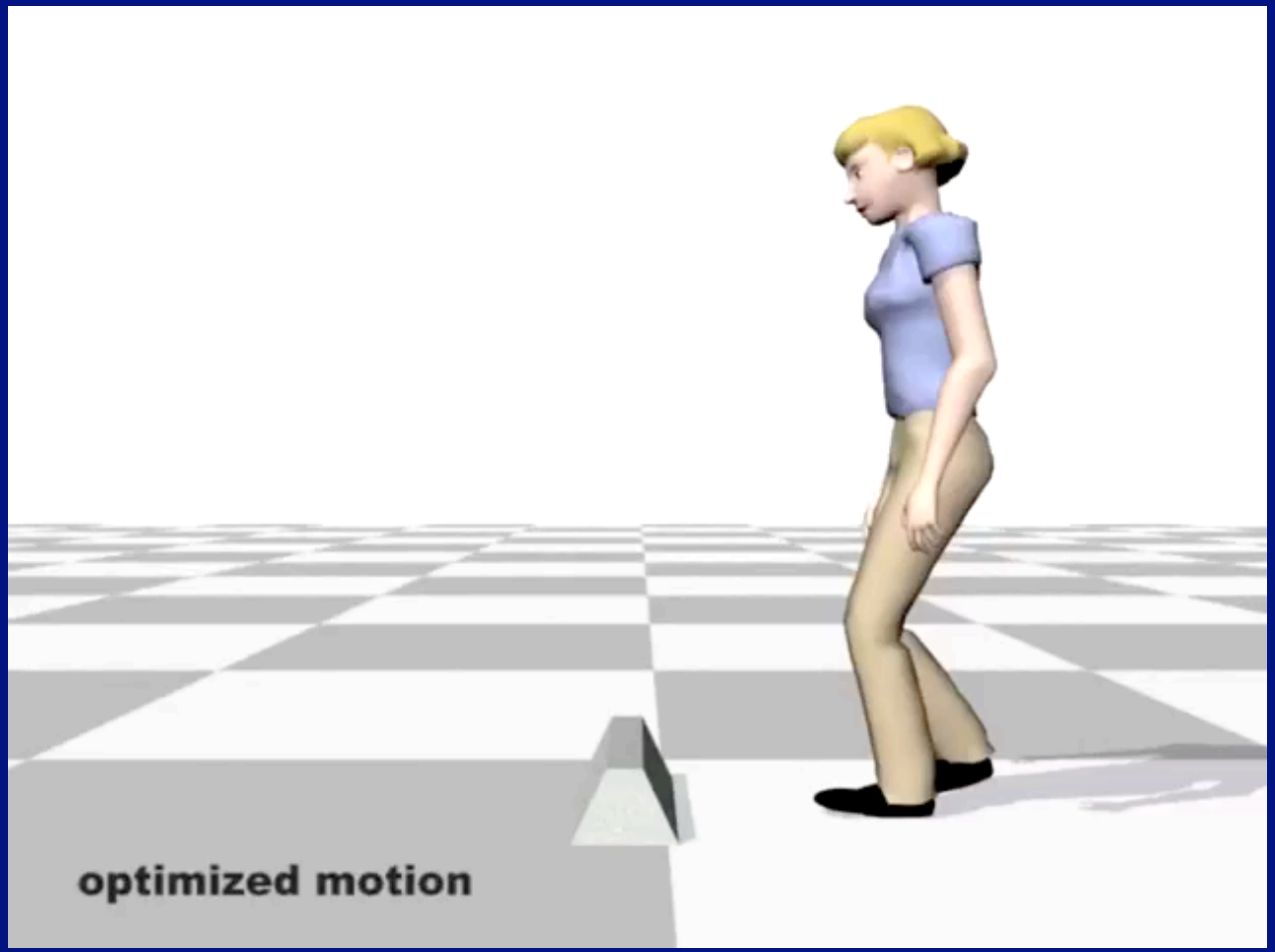
- Temporally fast phenomena are important to perception
  - means obvious methods work poorly
    - Blending works ok sometimes
    - Compression works ok sometimes
    - Tracking works ok sometimes
  - All mess up contacts

Footskate





Mataric et al,



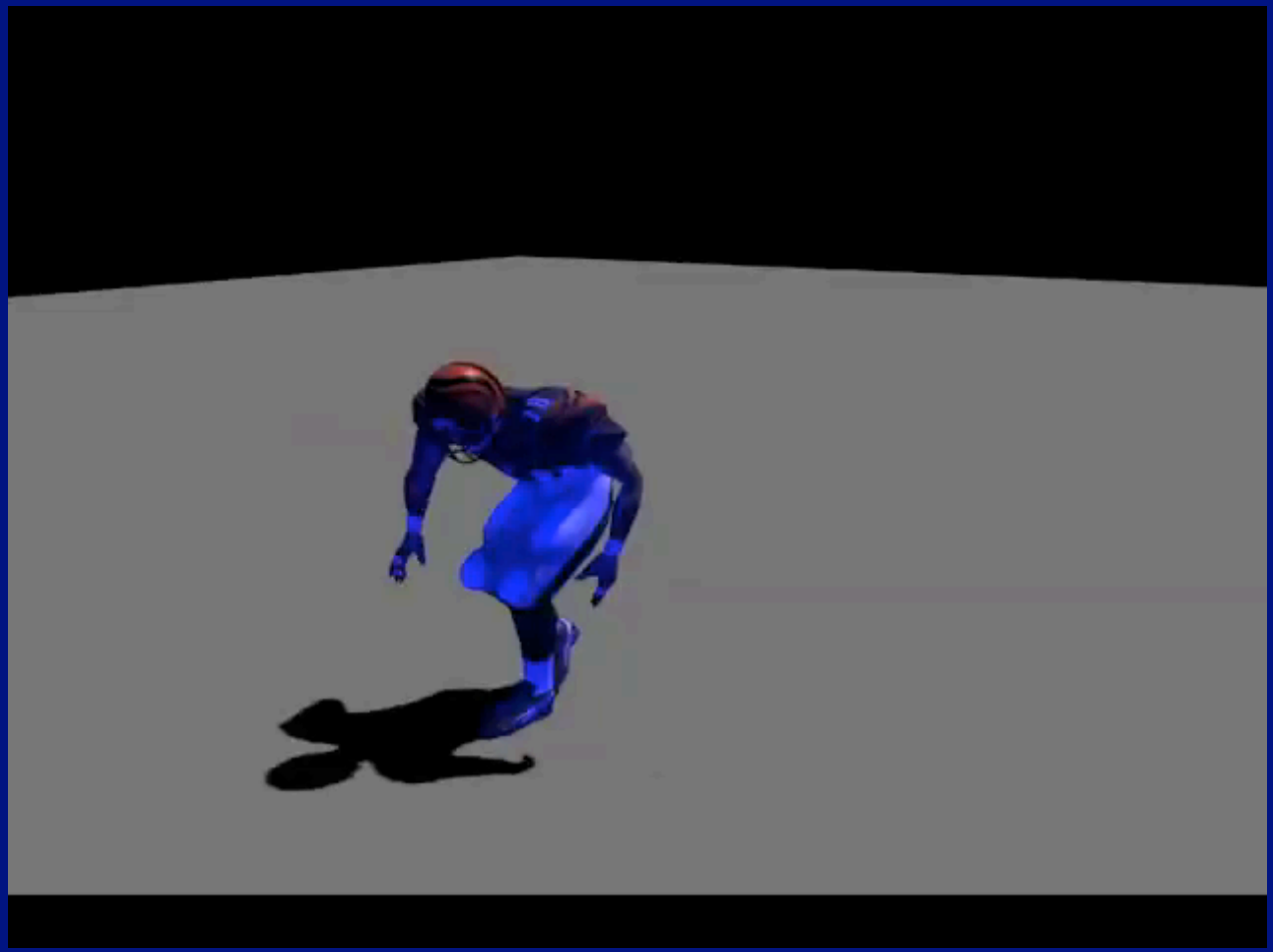
Safonova ea 04

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

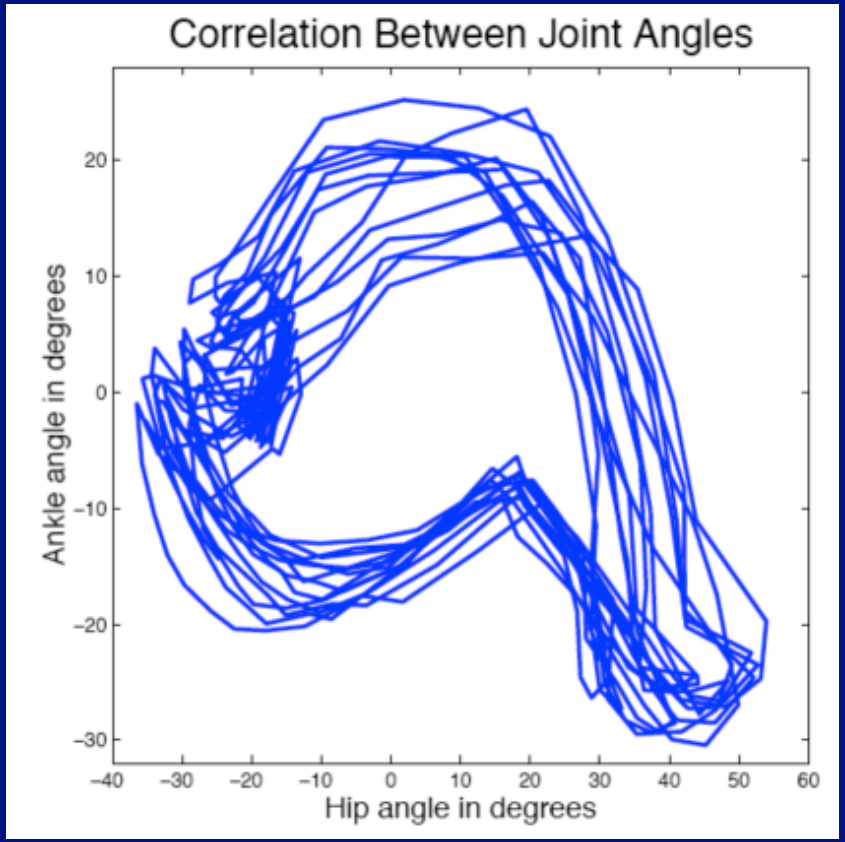
Ikemoto+Forsyth 04

Ikemoto+Forsyth 04



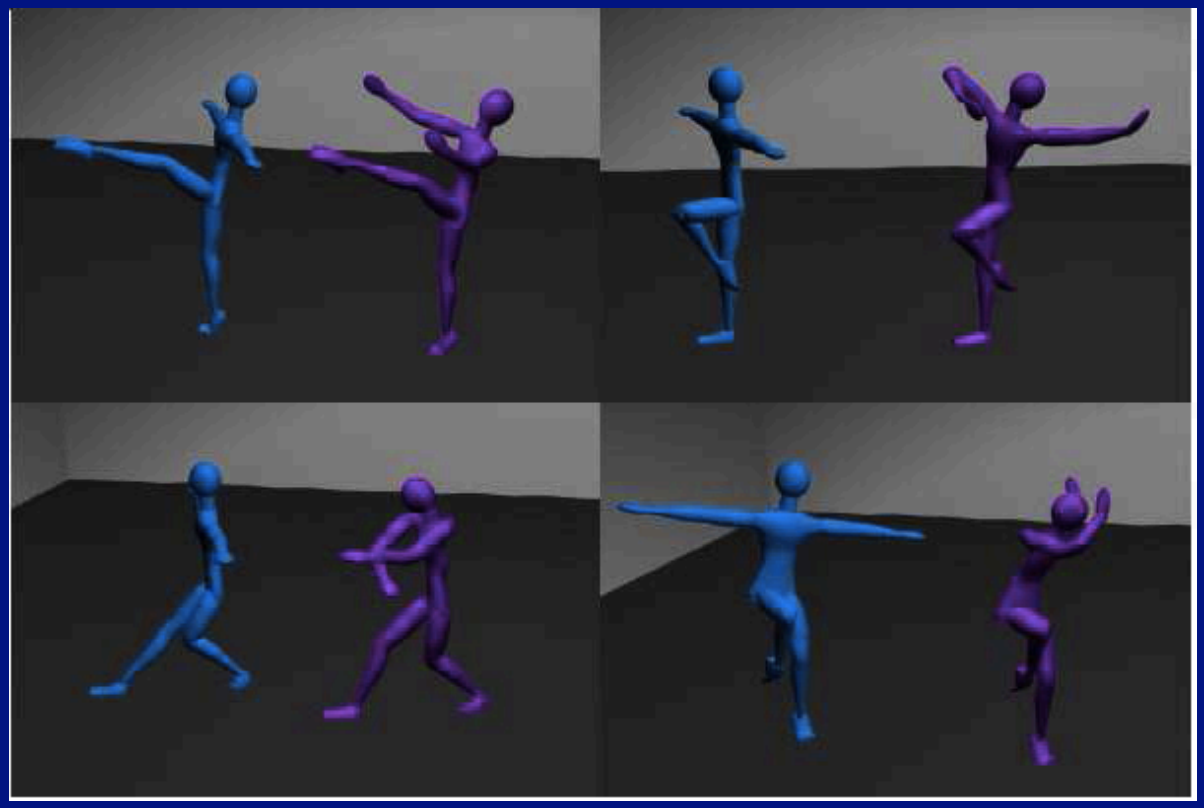
Ikemoto+Forsyth 04

Joint angles are heavily correlated



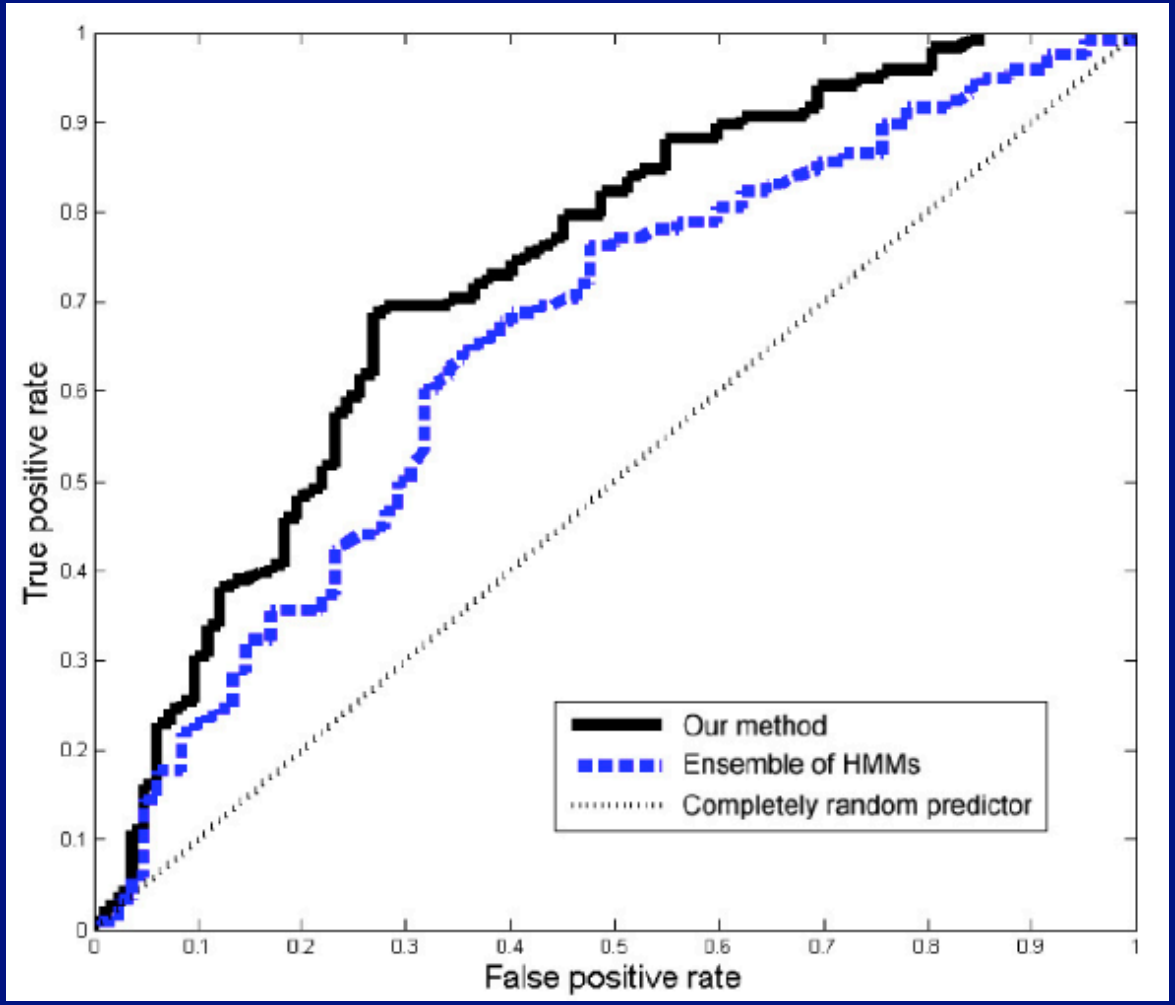
Pullen + Bregler 02

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

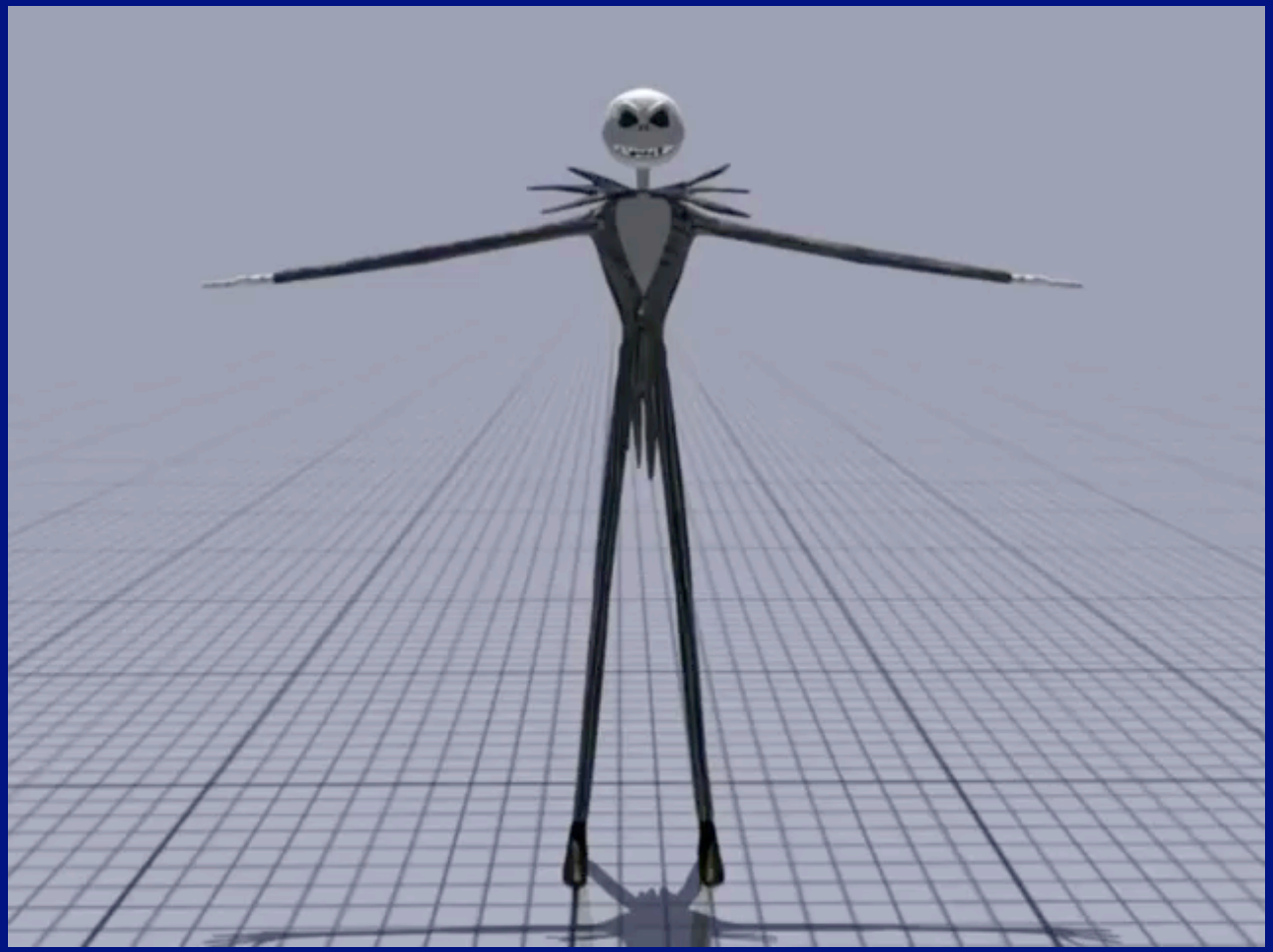
Why should we care?

- 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

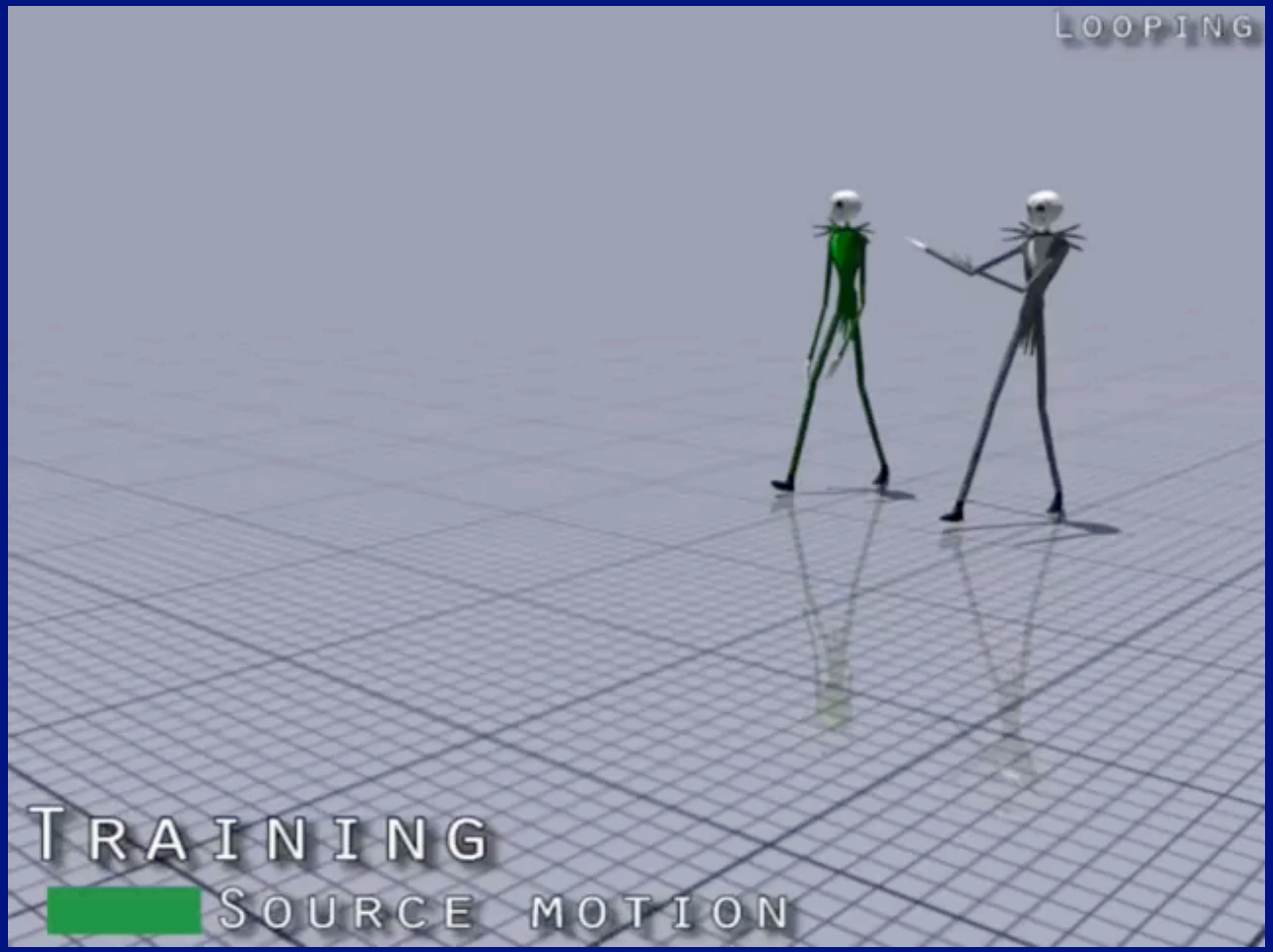
- Qualitative properties of motion, including
  - individual characteristics
  - modifiers, eg: clumsy, fast, heavy, forceful, graceful
- Animation problem:
  - Control new character with old motion, preserving new character's style
- Vision problem:
  - infer style descriptors, identity from observed motion

Kinematic style transfer



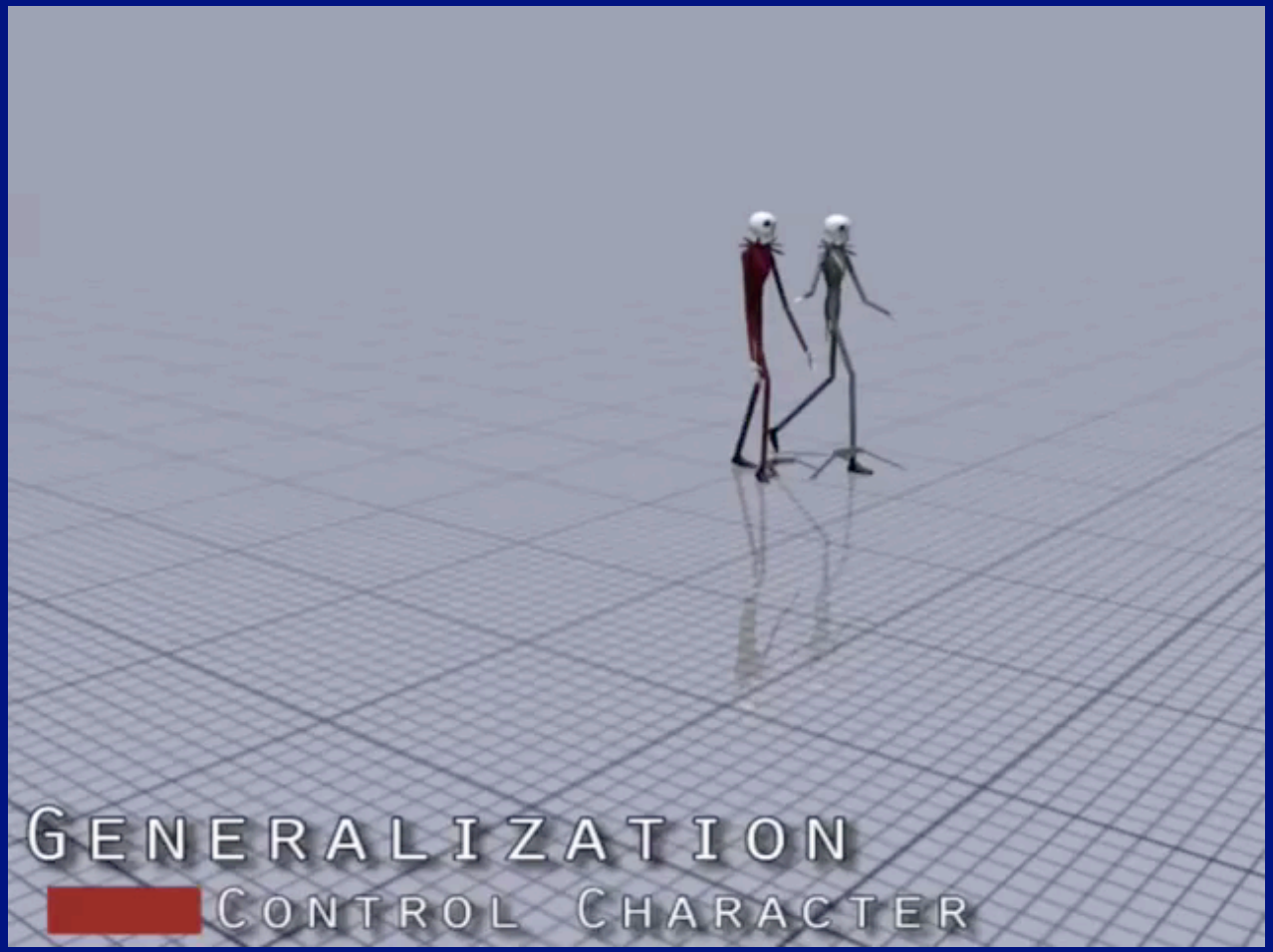
Ikemoto ea 09

Kinematic style transfer

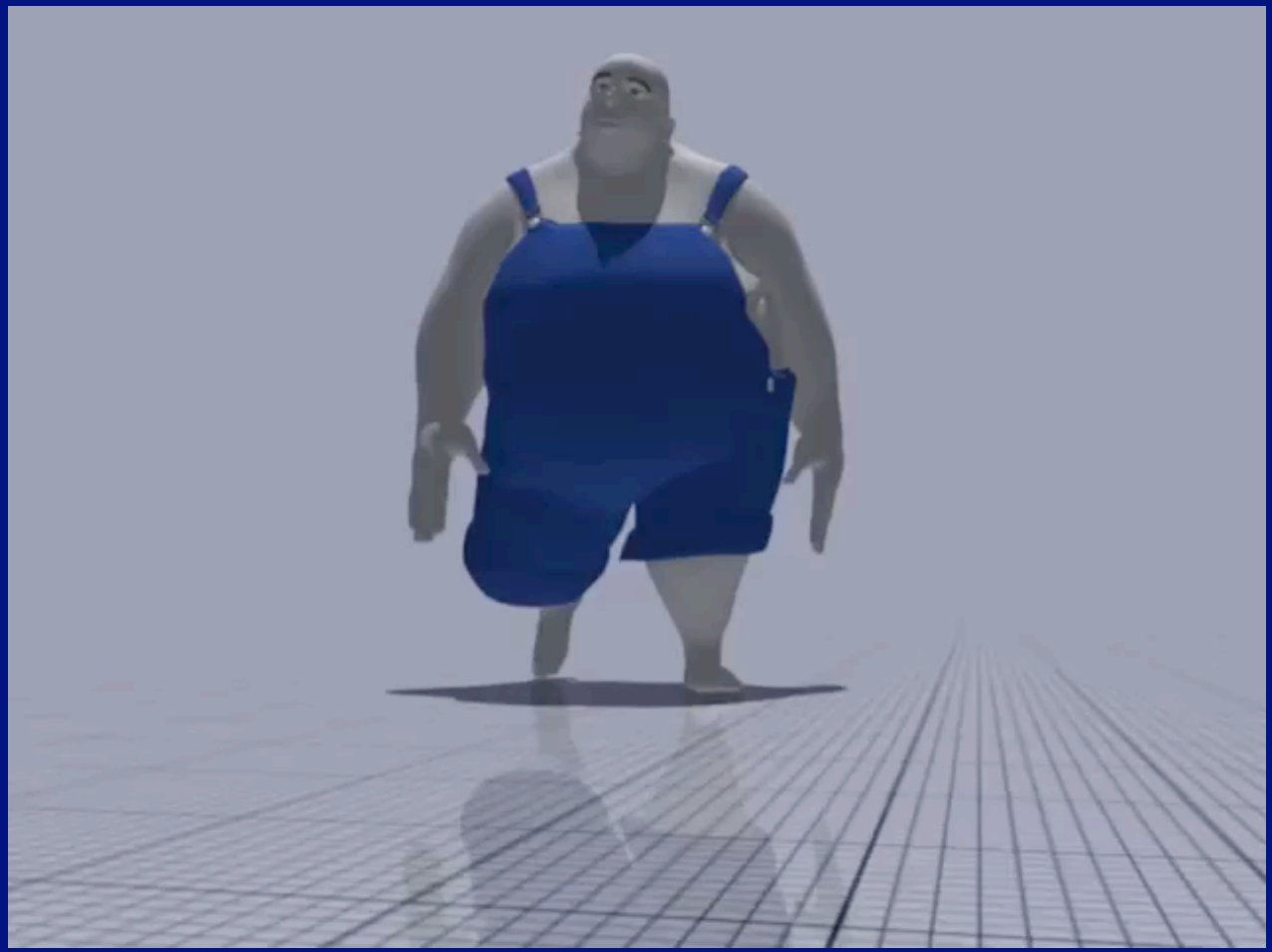


Ikemoto ea 09

Kinematic style transfer



Kinematic style transfer



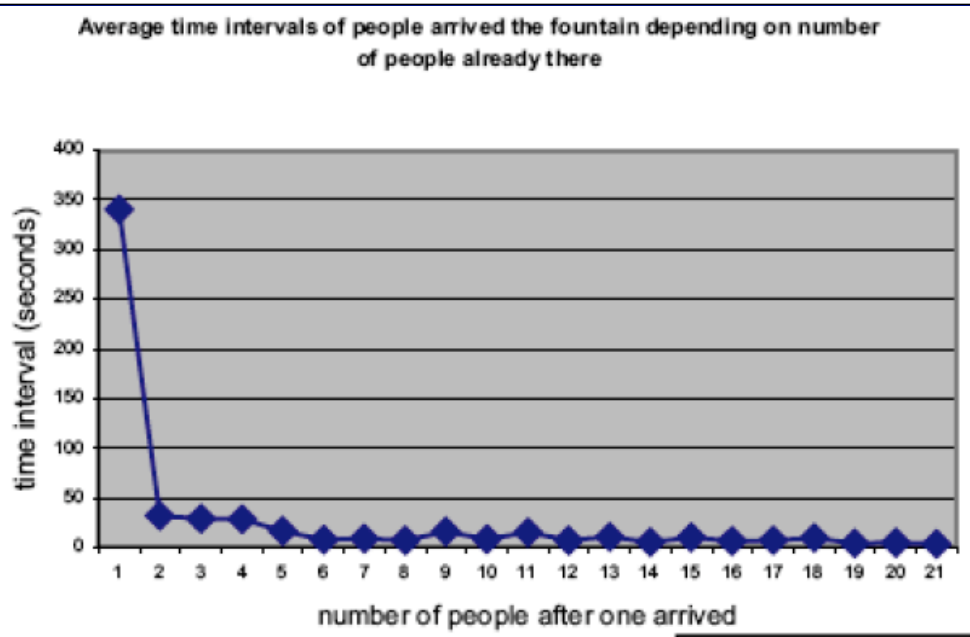
Ikemoto ea 09

Why should we care?

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

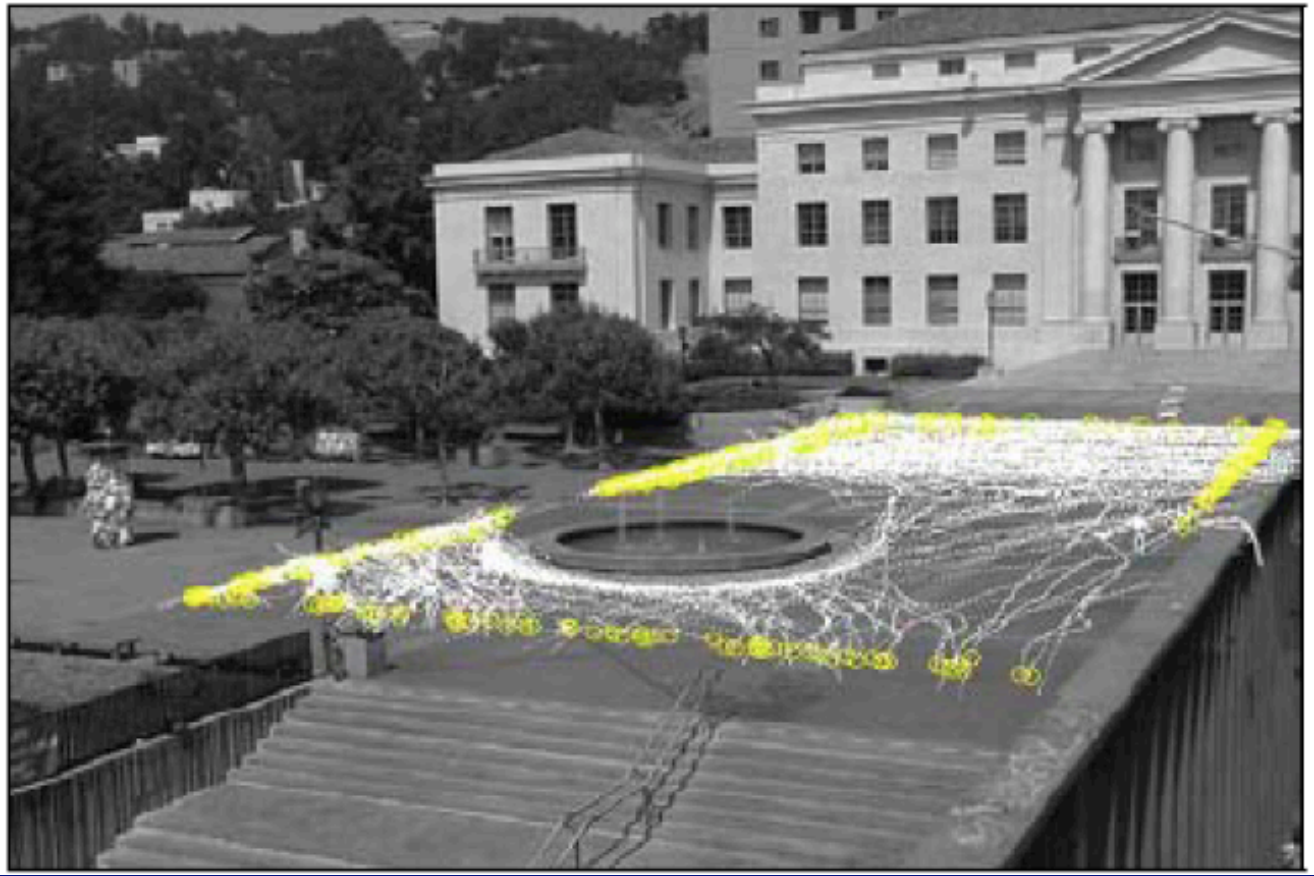
Core questions

- 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

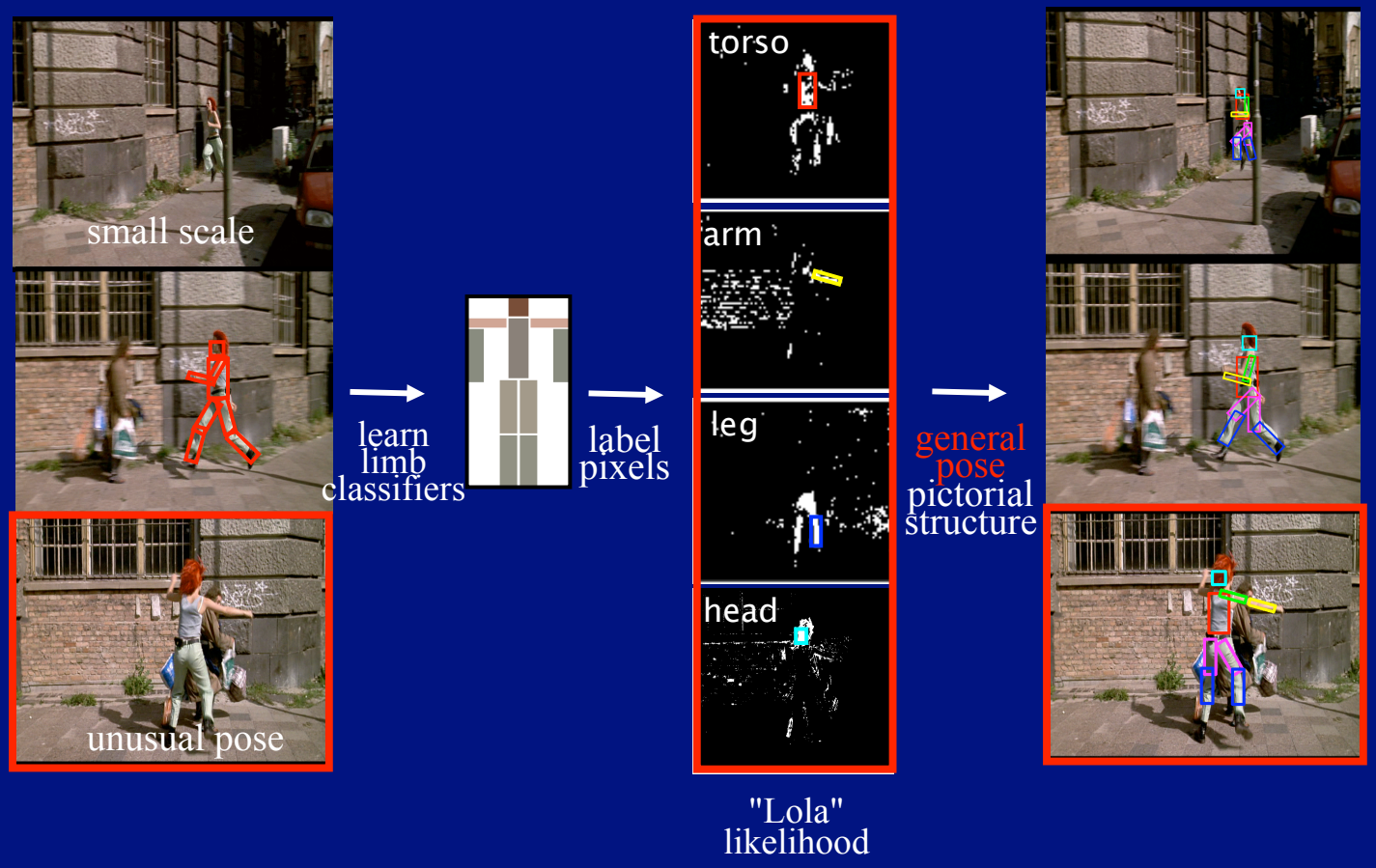
self-occlusion & clutter

variation in appearance

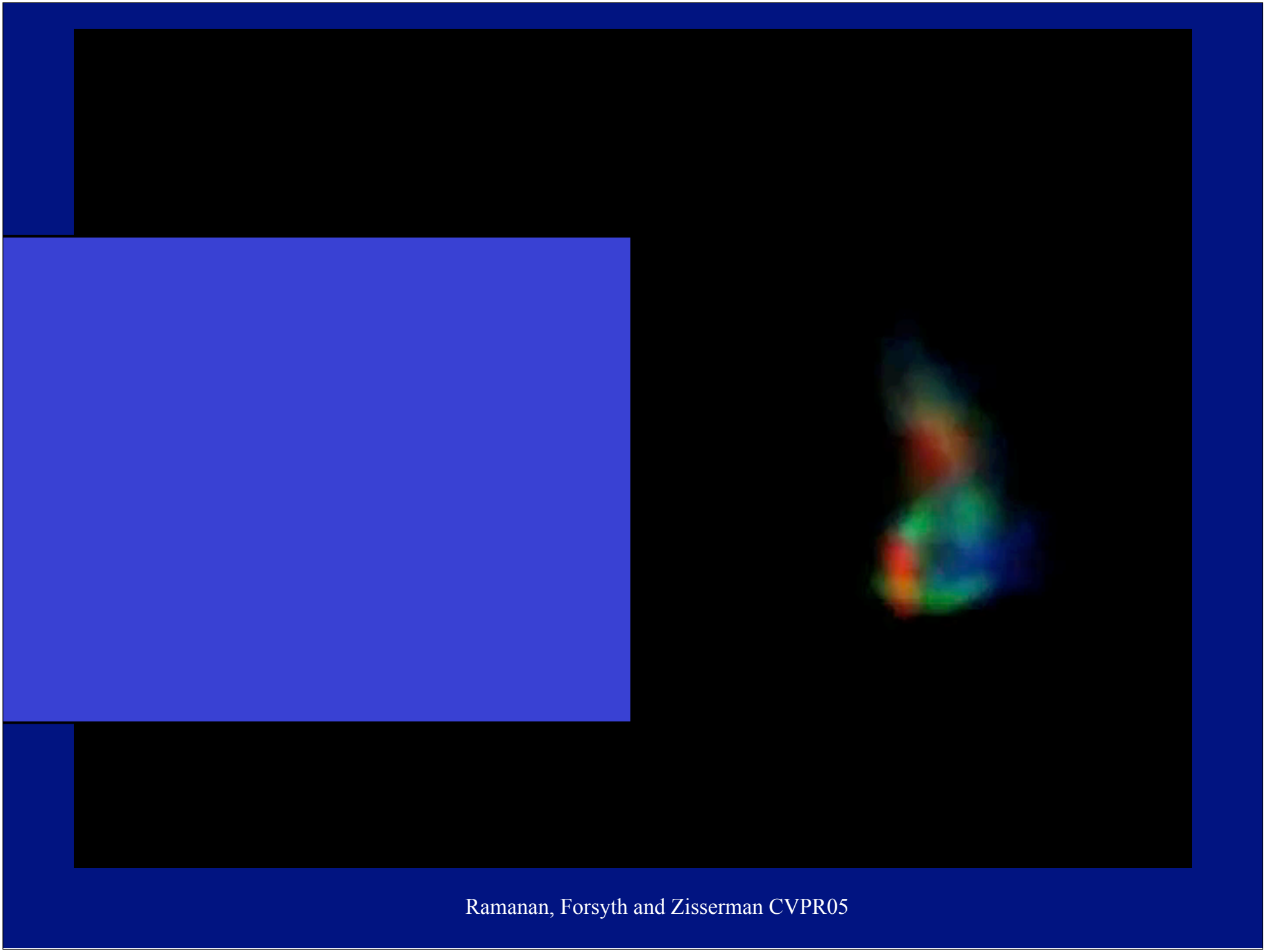
Kinematic tracking background

- 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



Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05

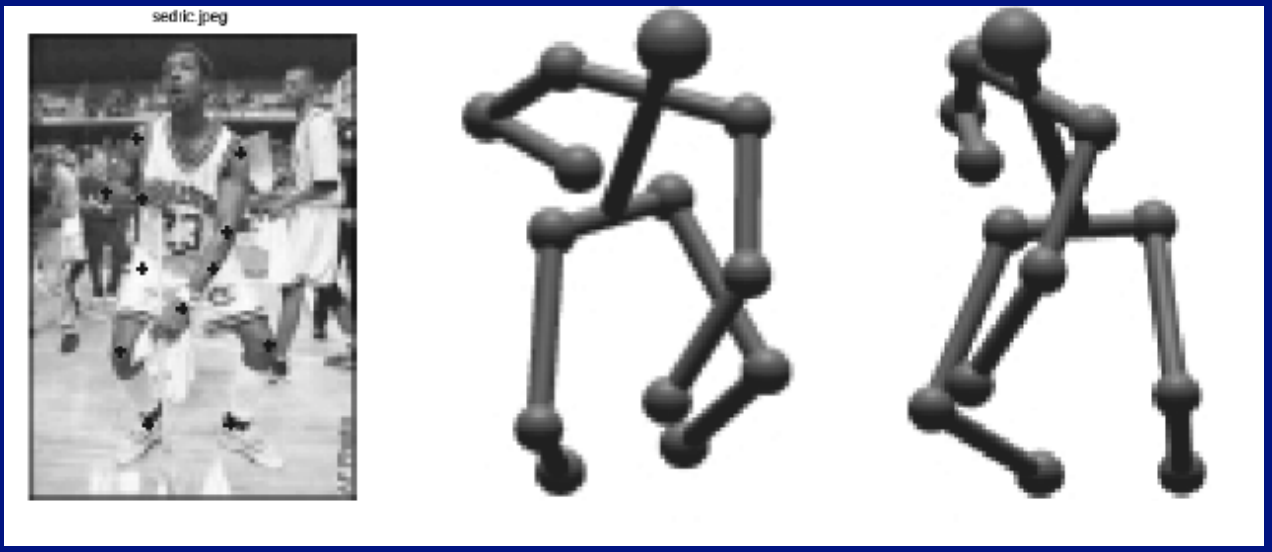
Ramanan, Forsyth and Zisserman CVPR05

Coming to tracking

- Amazing 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)
  - Complex spatial models (Sapp ea 10a)
  - Cascade models (Sapp ea 10b)
  - Full relational models (Tran Forsyth 10)
  - Search over "clumps" of body parts (Bourdaev Malik 09; Wang ea 11)
  - Tiny parts (Yang+Ramanan 11)

Lifting

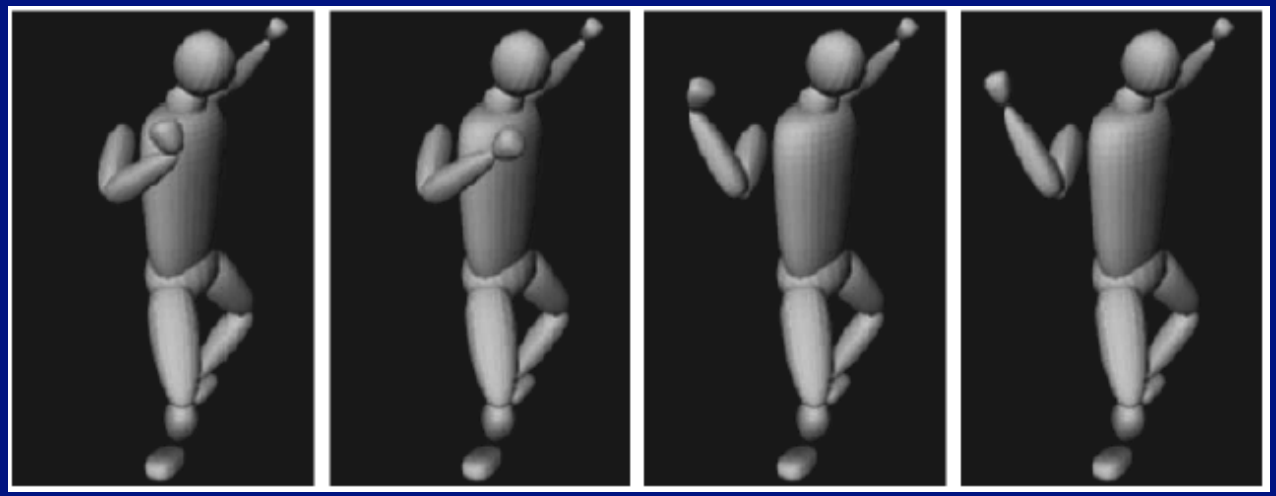
- Infer 3D configuration from image configuration
- Useful for
  - view independent activity recognition
  - user interfaces
  - video motion capture



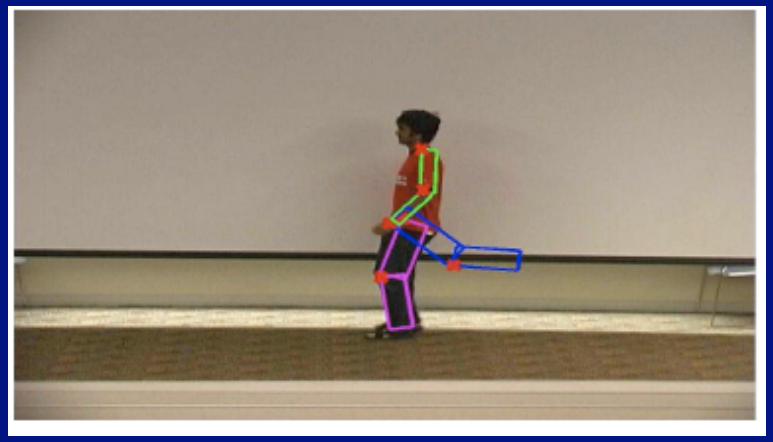
Taylor, 00

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



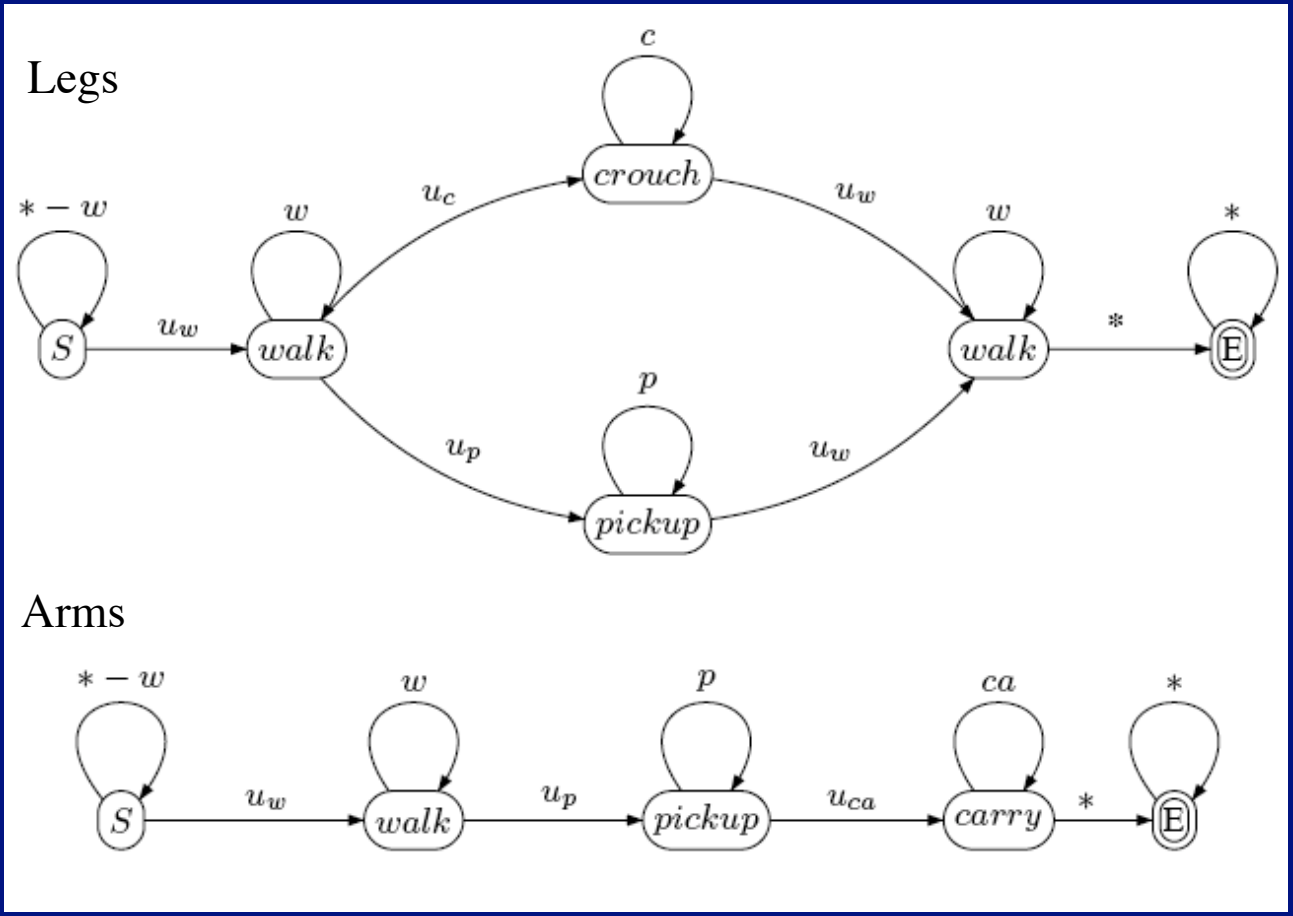
Core questions

- 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

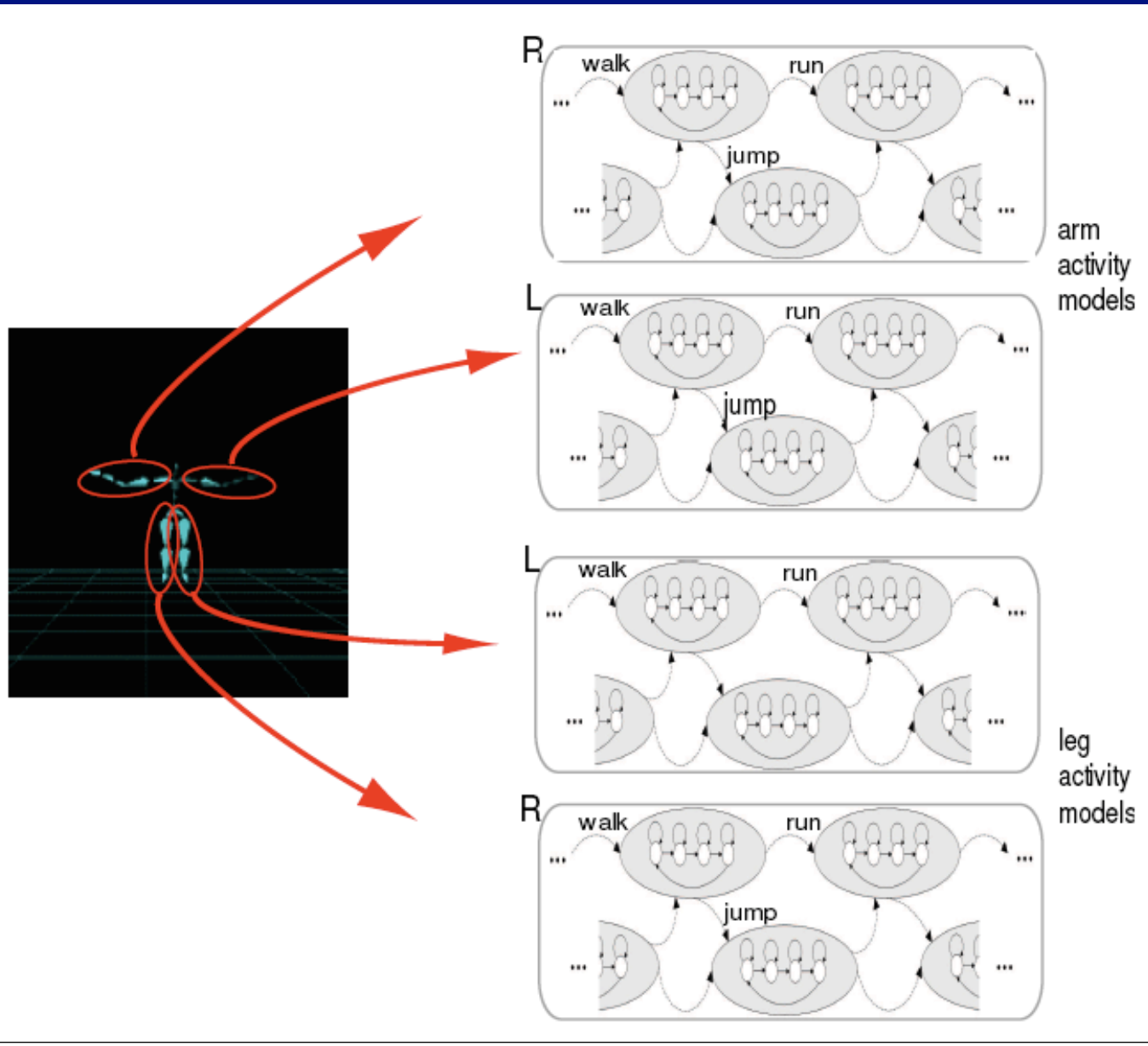
Naming activities

- 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



Composition



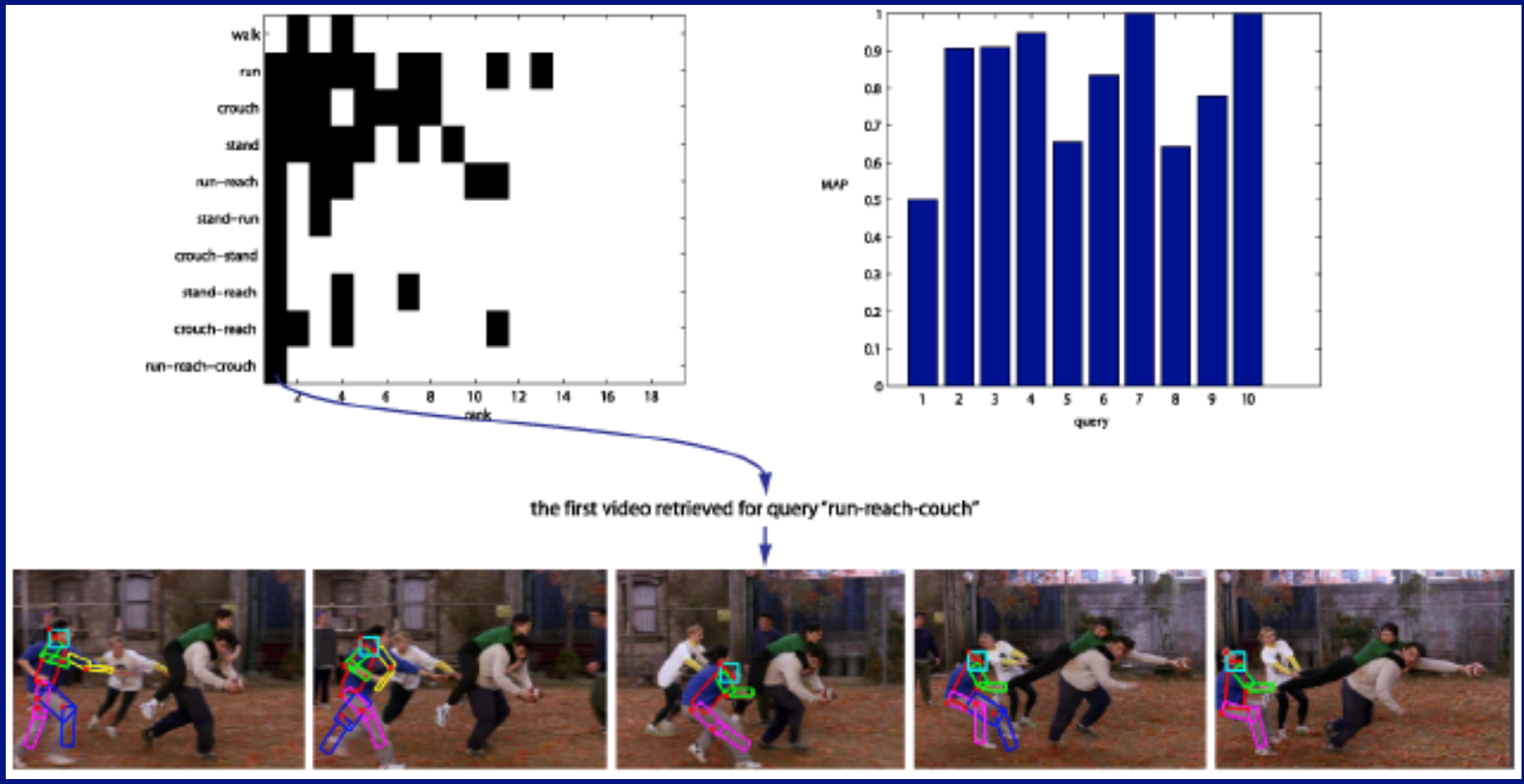
Ikizler Forsyth 07,08

Emission

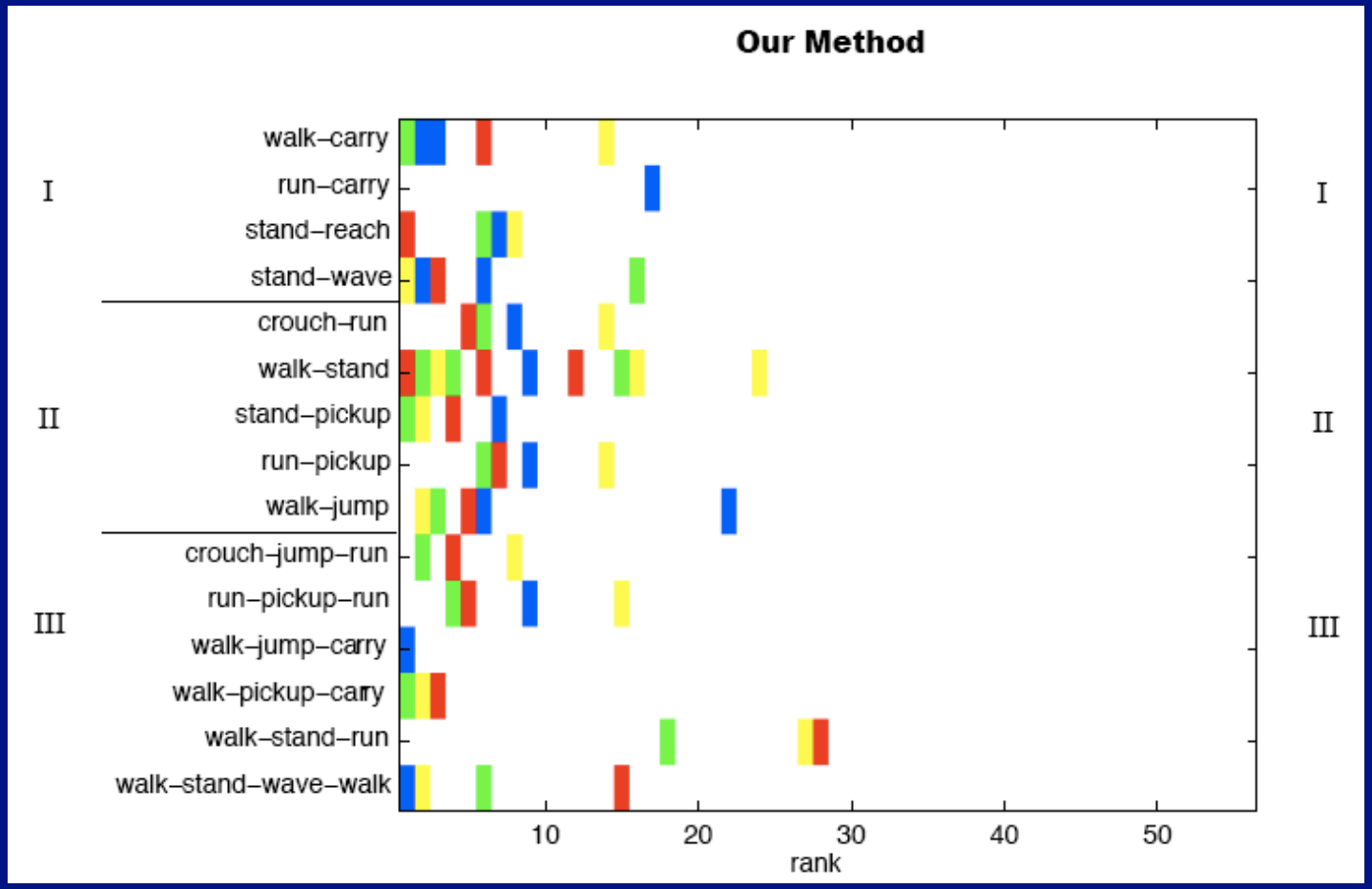
- Transduction
  - Track the body, as above
  - Lift “snippets” of each quarter
    - vector quantized
  - impose root consistency
- Emission
  - emit cluster center from state according to table
  - table learned by EM, known dynamical model

Query for motions with no examples

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

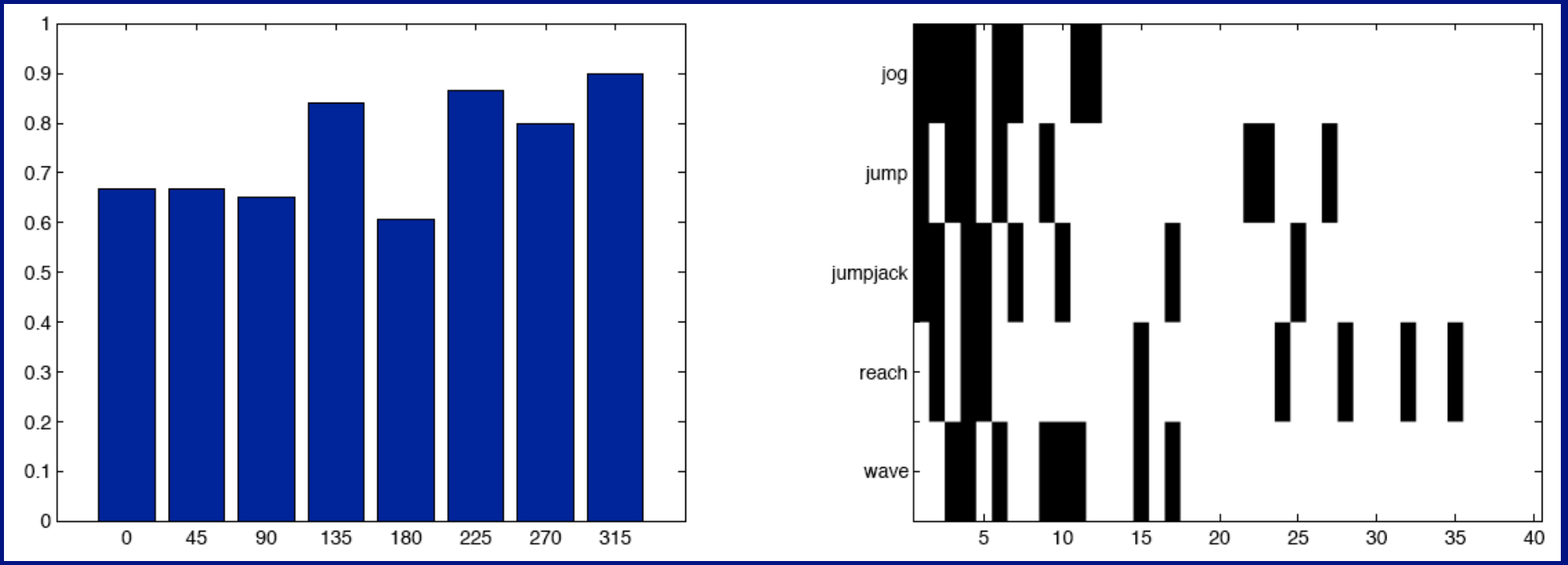


Searching for complex human activities with no visual examples N İkizler, DA Forsyth - IJCV, 2008



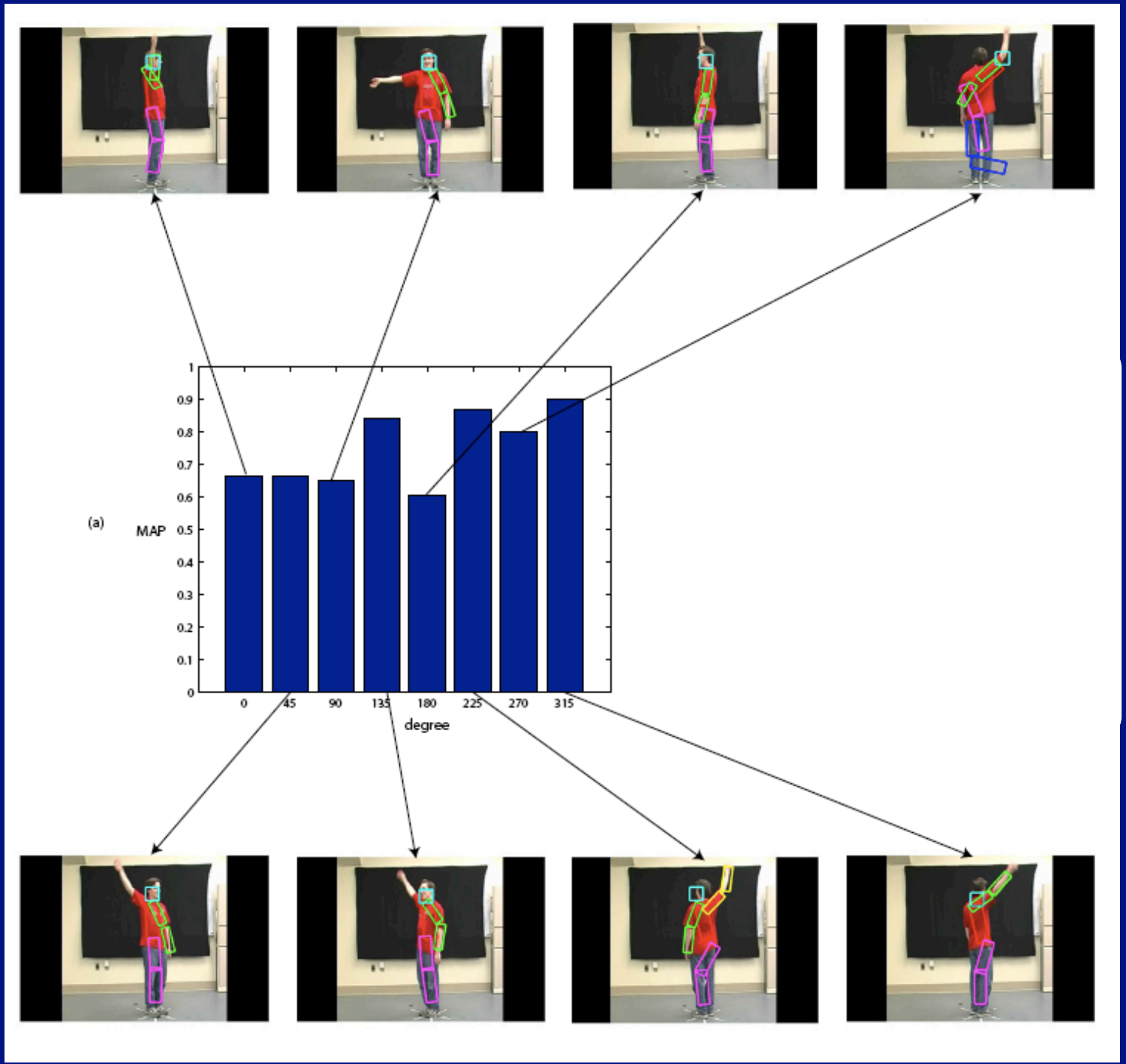
Ikizler Forsyth 07,08

The effect of aspect



Jog; Jump; Jumpjack; Reach; Wave

Ikizler Forsyth 07, 08



Ikizler Forsyth 07, 08












cooking - Google Search

http://images.google.com/search?hl=en&site=&tbm=isch&source=hp&q=cooking&oq=cooking&gs_l=img.3..0l1

Reader PRIVATE Google

Apple Yahoo! Google Maps YouTube Wikipedia News (314) Popular

Related searches: cooking **pot** cooking **clip art** cooking **food** cooking **cartoon** **kids** cooking cooking **class**











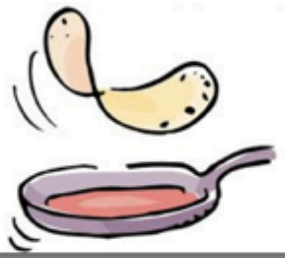






Page 2



white

011-4318-1111 / 1110

Netherlands

landlines. Mobile, charge cards and other

p://images.google.com/search?hl=en&site=&tbm=isch&source=hp&q=cooking&oq=cooking&gs_l=img.3..0l10 Reader PRIVATE Google

Yahoo! Google Maps YouTube Wikipedia News (314) Popular

About 186,000,000 results (0.10 seconds) SafeSearch off

Related searches: food preparation **cartoons** **safe** food preparation food preparation **hygiene**

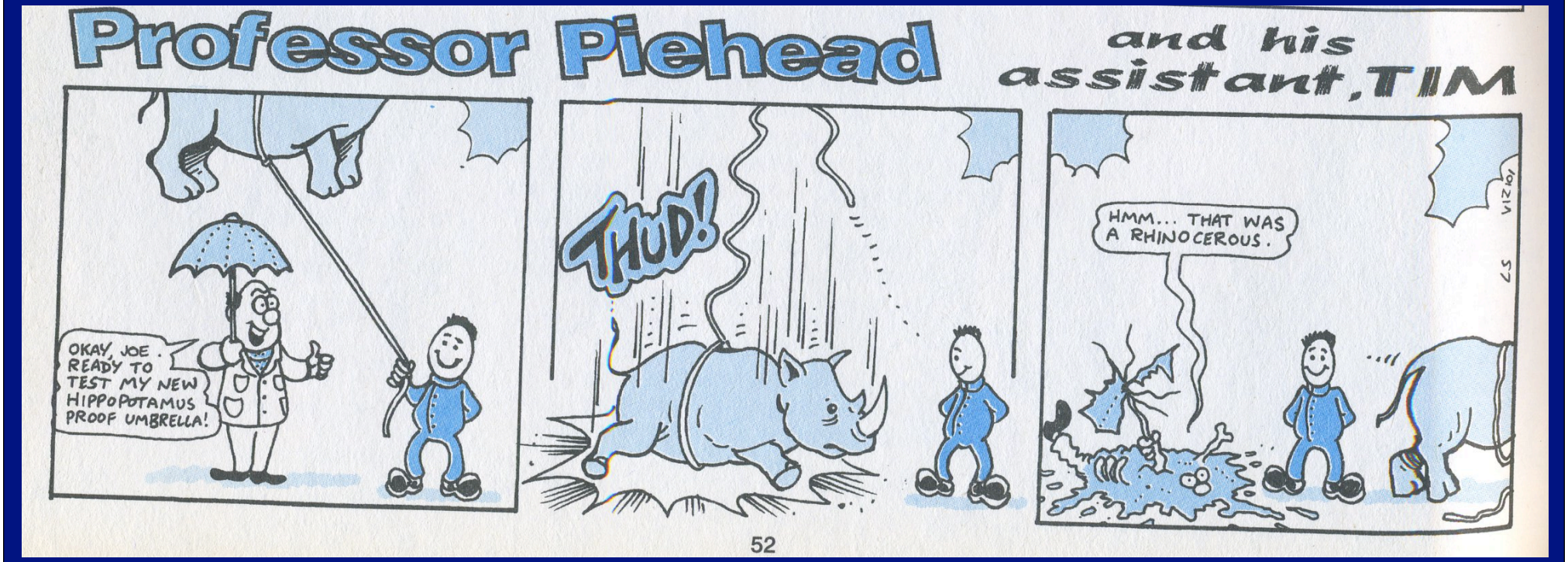
Page 2

011-4316-11111 1110 Netherlands landlines. Mobile, charge cards and other

Core questions

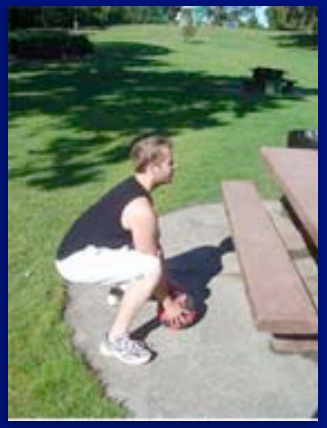
- What should we say about motion?
  - and what is worth mentioning?
- What properties does the signal have?
  - style and composition
- How should we transduce the signal?
  - infer body segments or not
- Representations of the body
  - what should we say about people and what's around them?

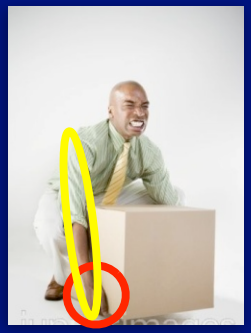
What is an object like?

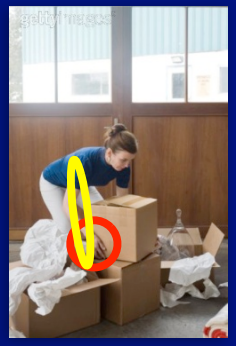


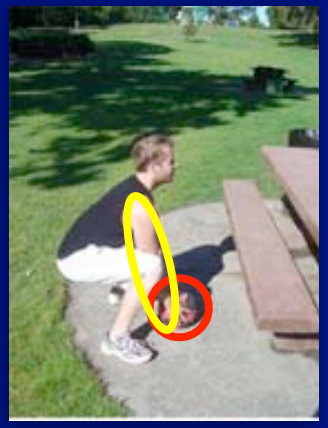
Viz comic, issue 101


ROCKIES



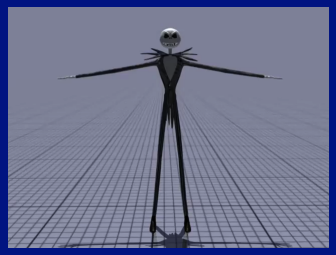






Activity attributes

- Gaze and focus
- Style
  - Fast/Gentle



- Timing
  - arms in phase with legs

- Contact
  - Having hand contact

- Kinematic
  - Arms sticking out

Nearby objects and free space

Contact? but where are the hands?

- Possible method
  - rack up hierarchical poselet features after Wang et al 11
    - these are small and large poselets
  - build a decision tree using Bourdev's dataset for 3D body configuration
  - make decisions based on image features
  - Result
    - which box in 3D wrt the body is the hand in for this person?
- "Evaluation"
  - render hand pose on image using inferred camera

Tsatsoulis nd

Where is it happening?

- where you are often reveals what you are doing
  - but how do we encode where you are
    - x-y coords? (Big bias problems)
    - near the stove? (perhaps smaller bias problems)



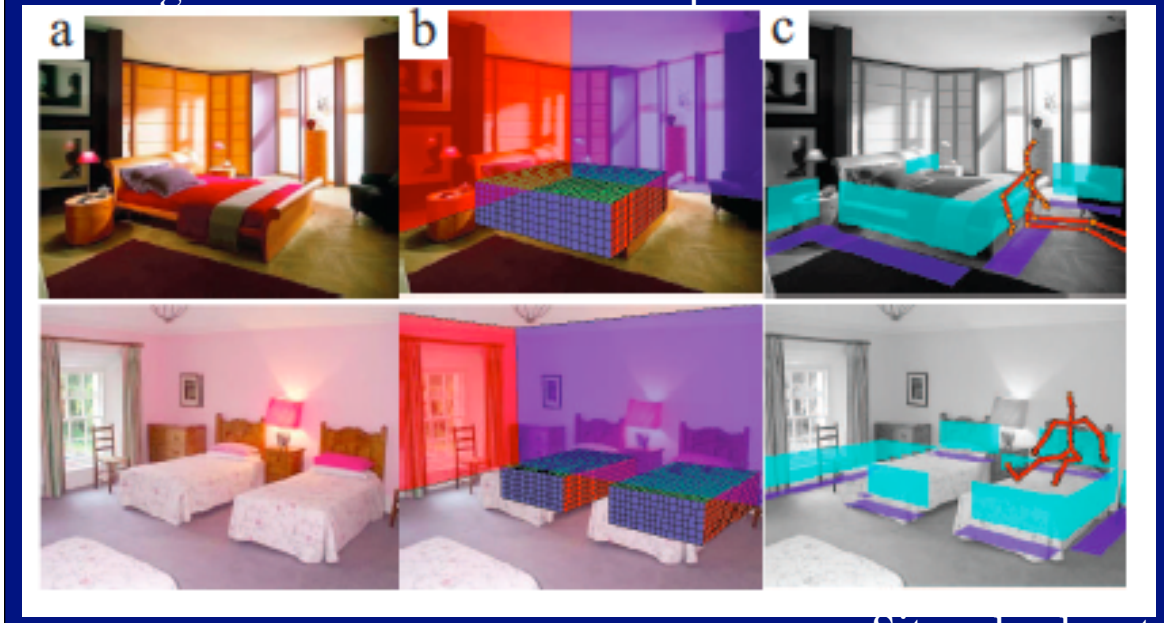



Thinking about Free Space

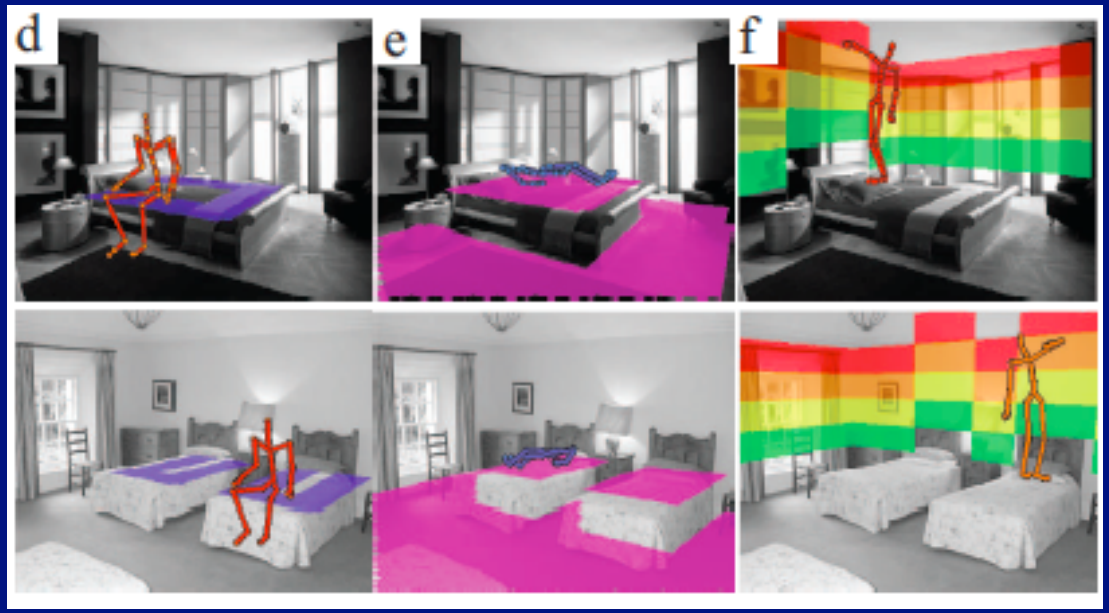
- Rooms are important
- Rough 3D representations help recognition

- Free space has motion "potential"
  - it tells you where you can move
  - and we can recover descriptions of free space that are good for motion

Image Geometric repn Sit with backrest

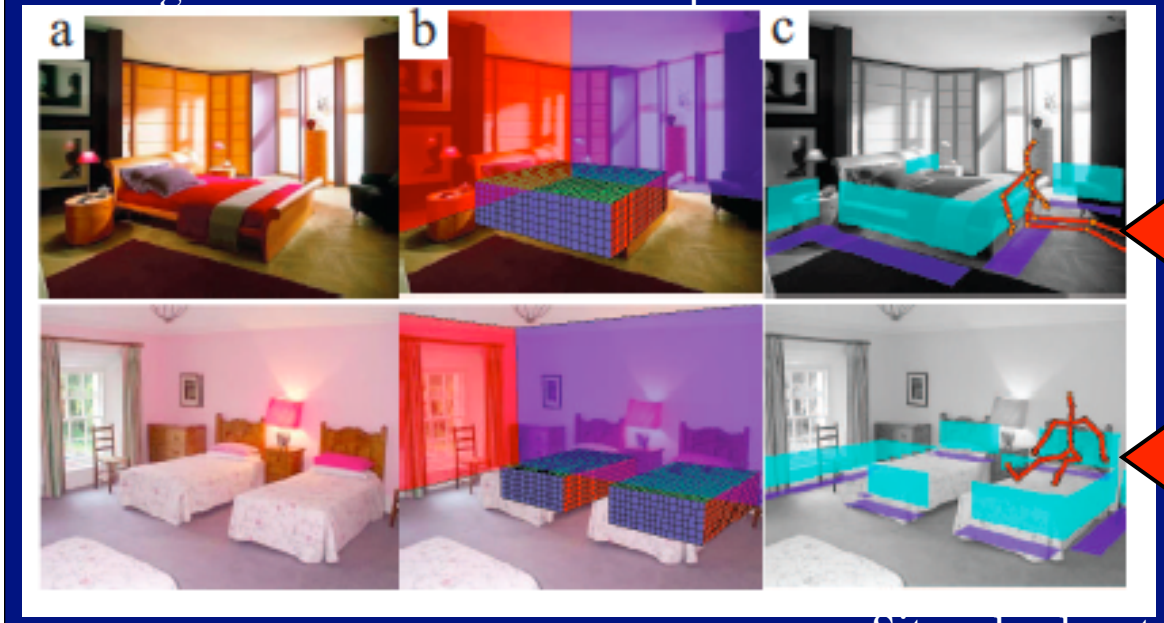


Sit no backrest Lie down Reach+touch



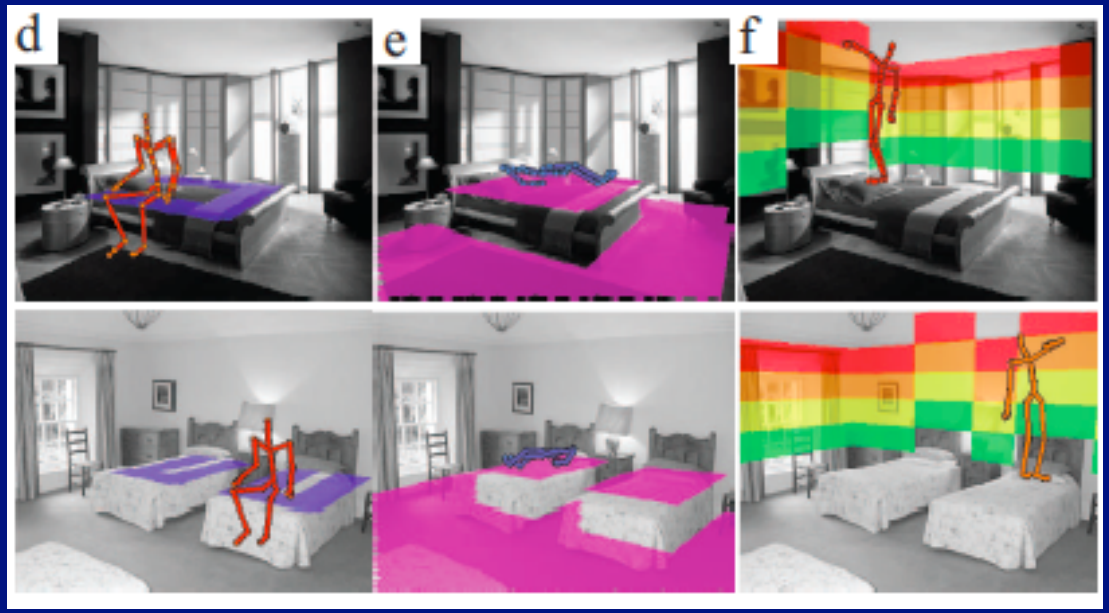
Gupta ea ’11

Image Geometric repn Sit with backrest



3D details are HARD to get right

Sit no backrest Lie down Reach+touch



Gupta ea ’11

The indoor scene dataset

Finding boxy objects - issues

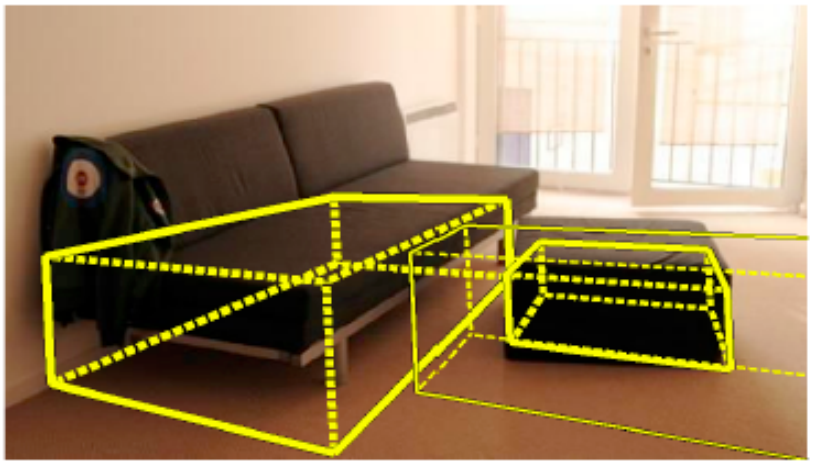
- Hard to distinguish beds, sofas, some tables, chairs
  - Block out free space first
  - (Perhaps) then use layout, etc. to figure out details
- Free space is useful
  - you can move there
  - objects might appear there
- Estimating free space correctly is HARD
  - small 2D errors give big 3D errors
    - floor is inclined; perspective

Finding boxy objects - process

- Assume
  - they're axis aligned
    - search angles later
  - quantizing to a small set of aspect ratios is OK
- Group local cues with a face finder to get proposal boxes
  - as in beds
- Check these boxes with context cues
  - what boxes are nearby? what are their scores?
  - discriminative, as in Farhadi+Sadeghi, '11
- Refine these boxes
  - to improve 3D estimate

Contrast refinement

- Many boxy objects aren't really boxes
  - they have legs
- This matters
  - small 2D errors make big 3D errors
  - the inclination of the floor
- Check the base of the box, adjust



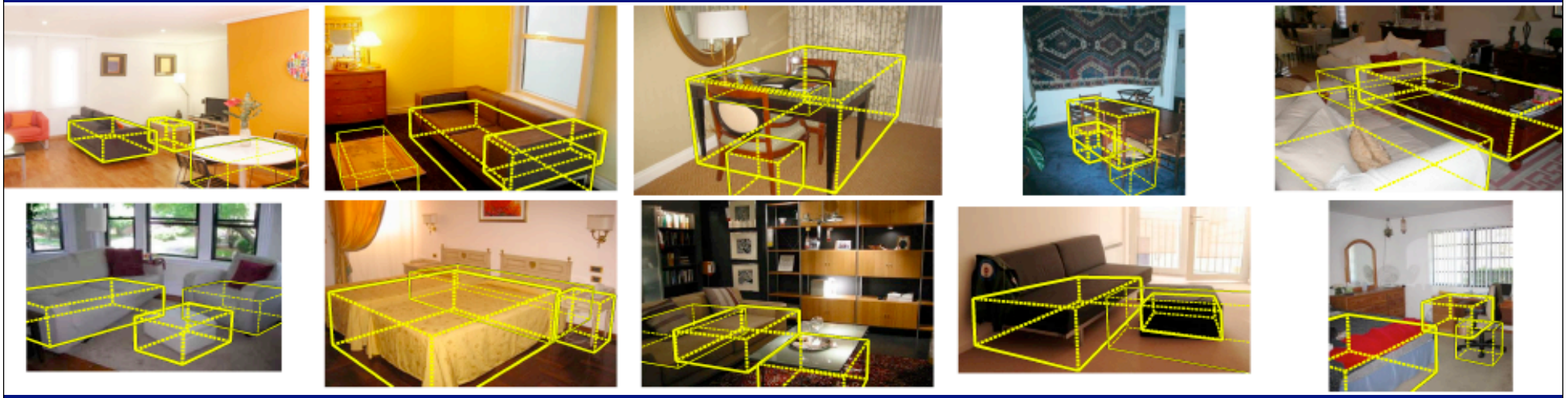
Contrast refinement helps

Initial

Refined

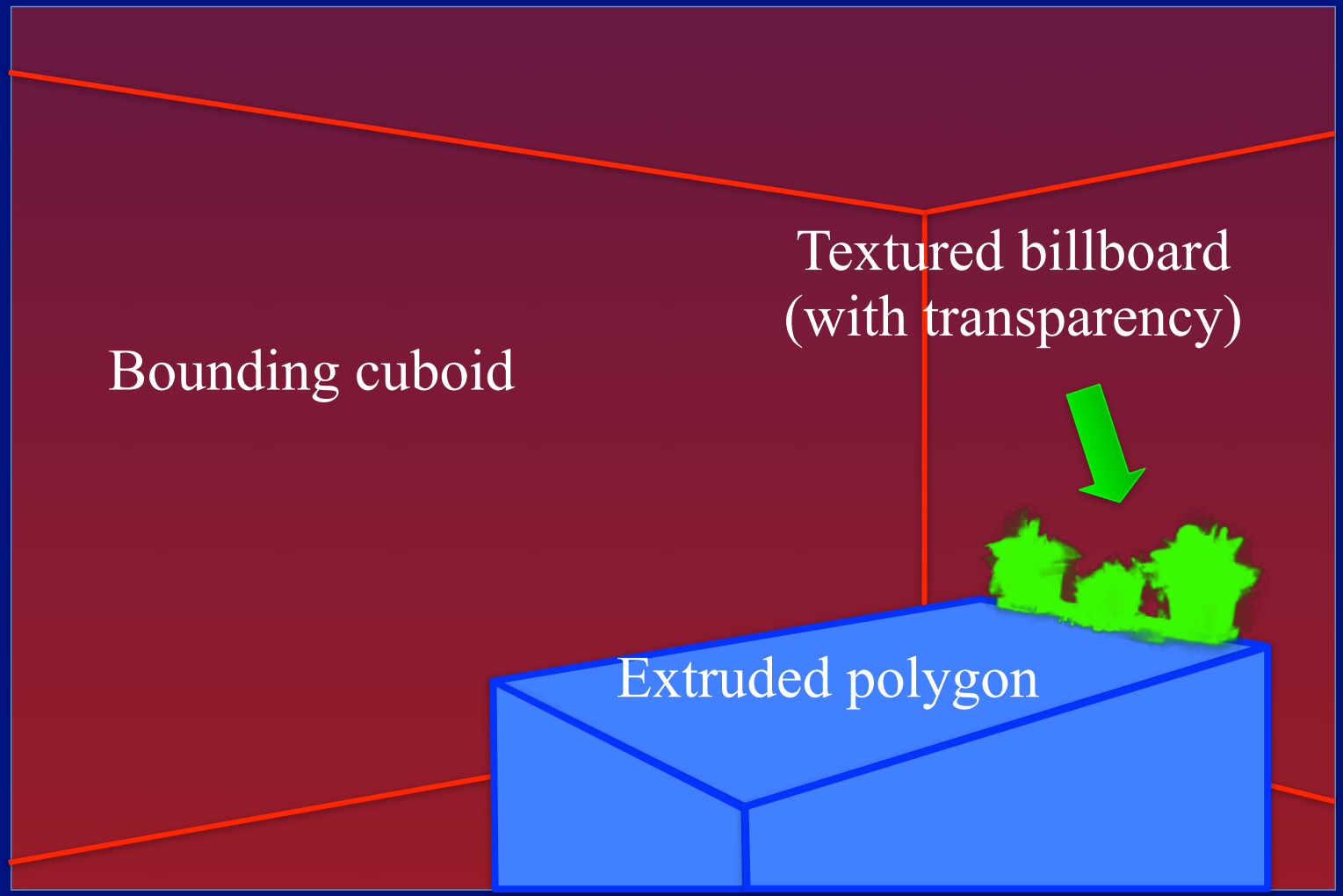
Floorplan in green

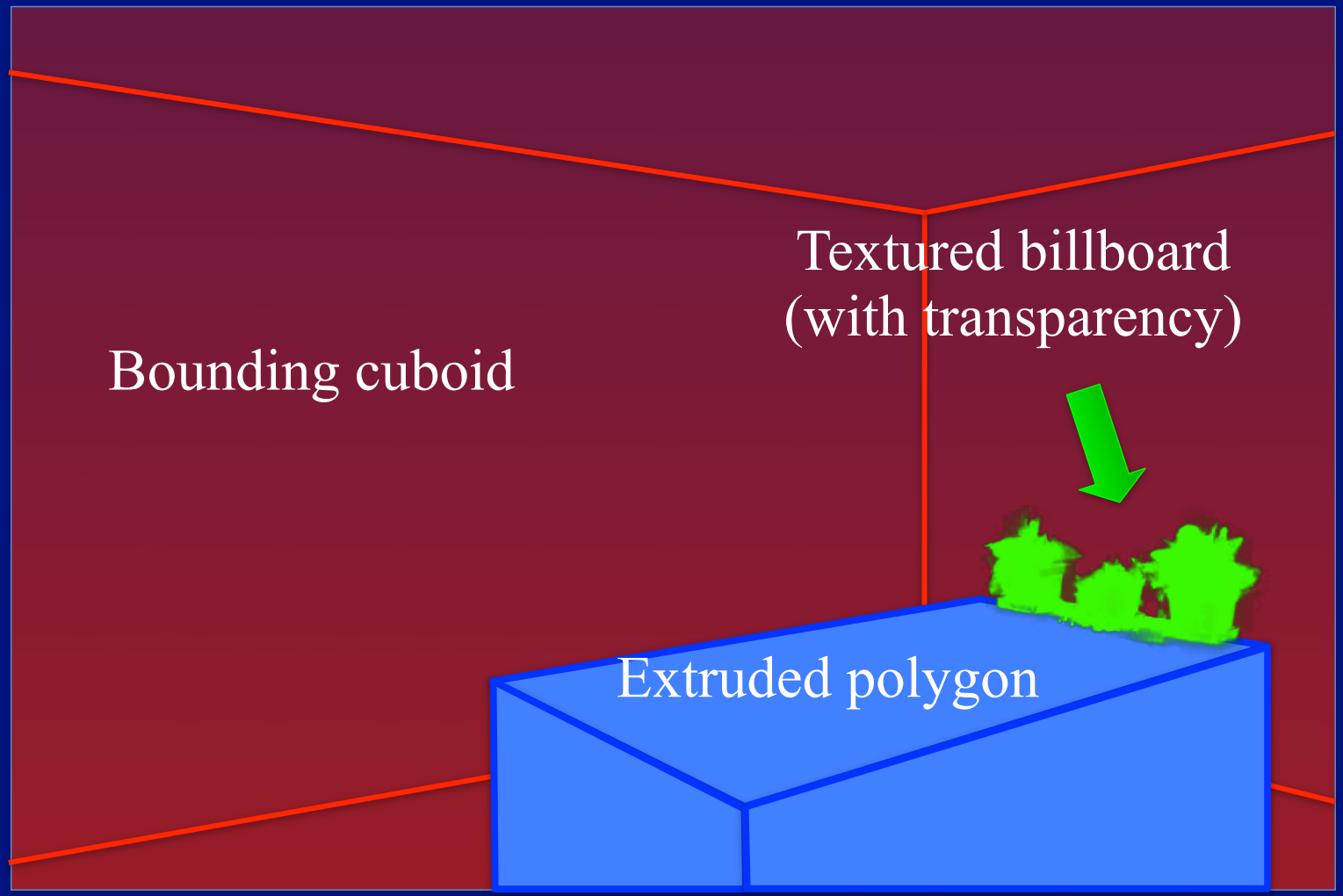
Detected boxy objects



Stage lighting

From Koenderink slides on image texture and the flow of light


Textured billboard
(with transparency)
Bounding cuboid
Extruded polygon


Textured billboard
(with transparency)
Bounding cuboid
Extruded polygon

System Overview

Input image

Estimate geometry

Estimate materials

Estimate lighting

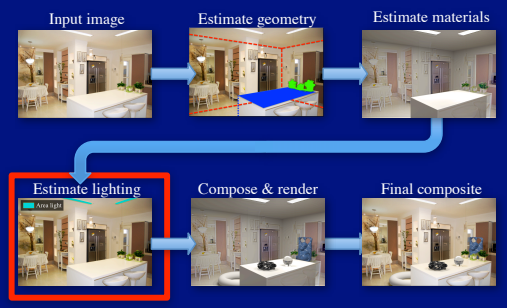
Compose & render

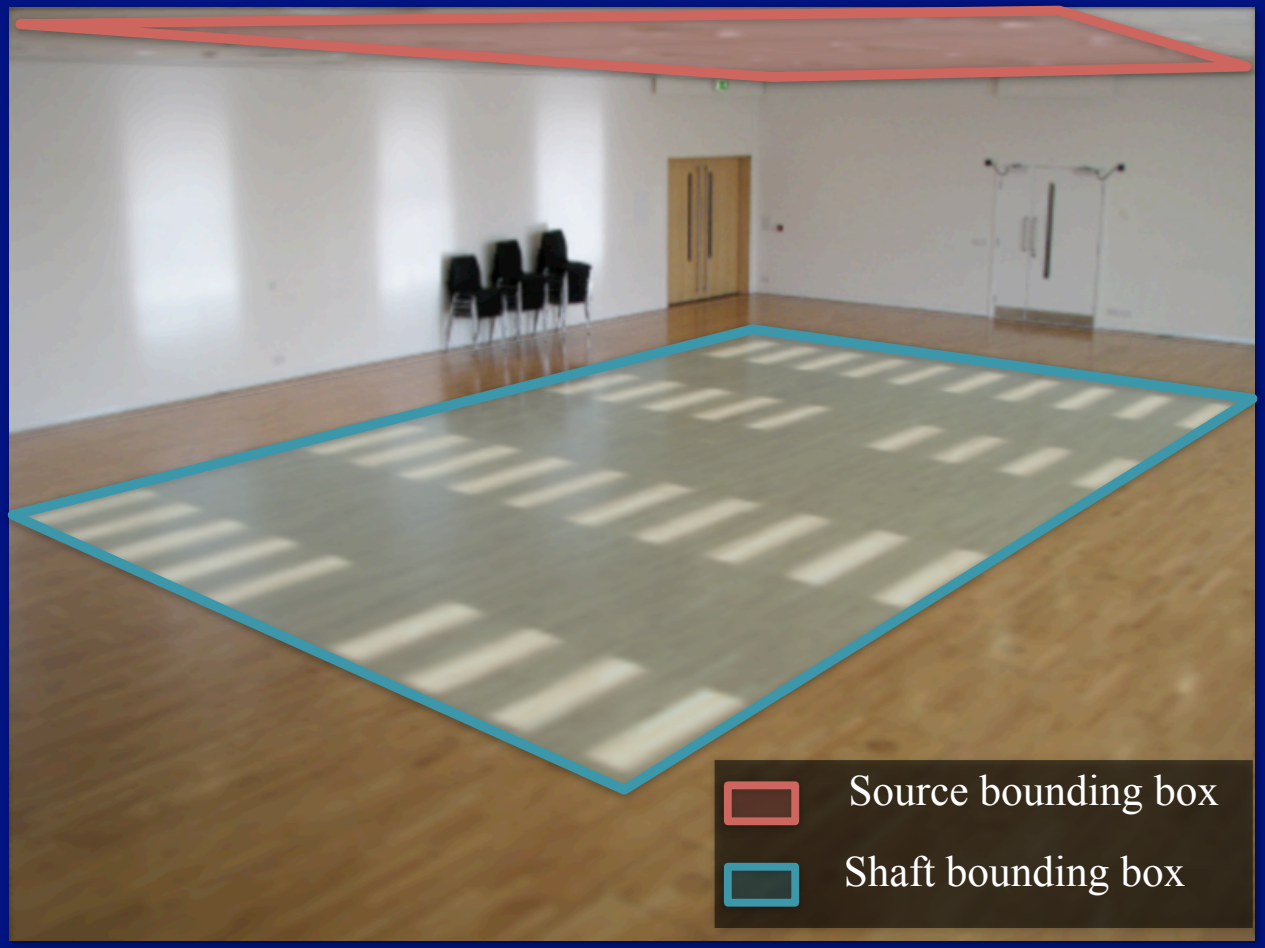
Final composite

Secret sauce: Consistency

Secret sauce: Physical renderer

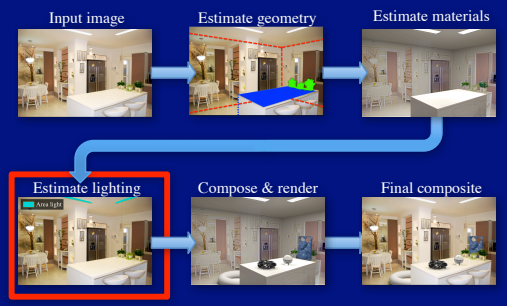
External light shafts

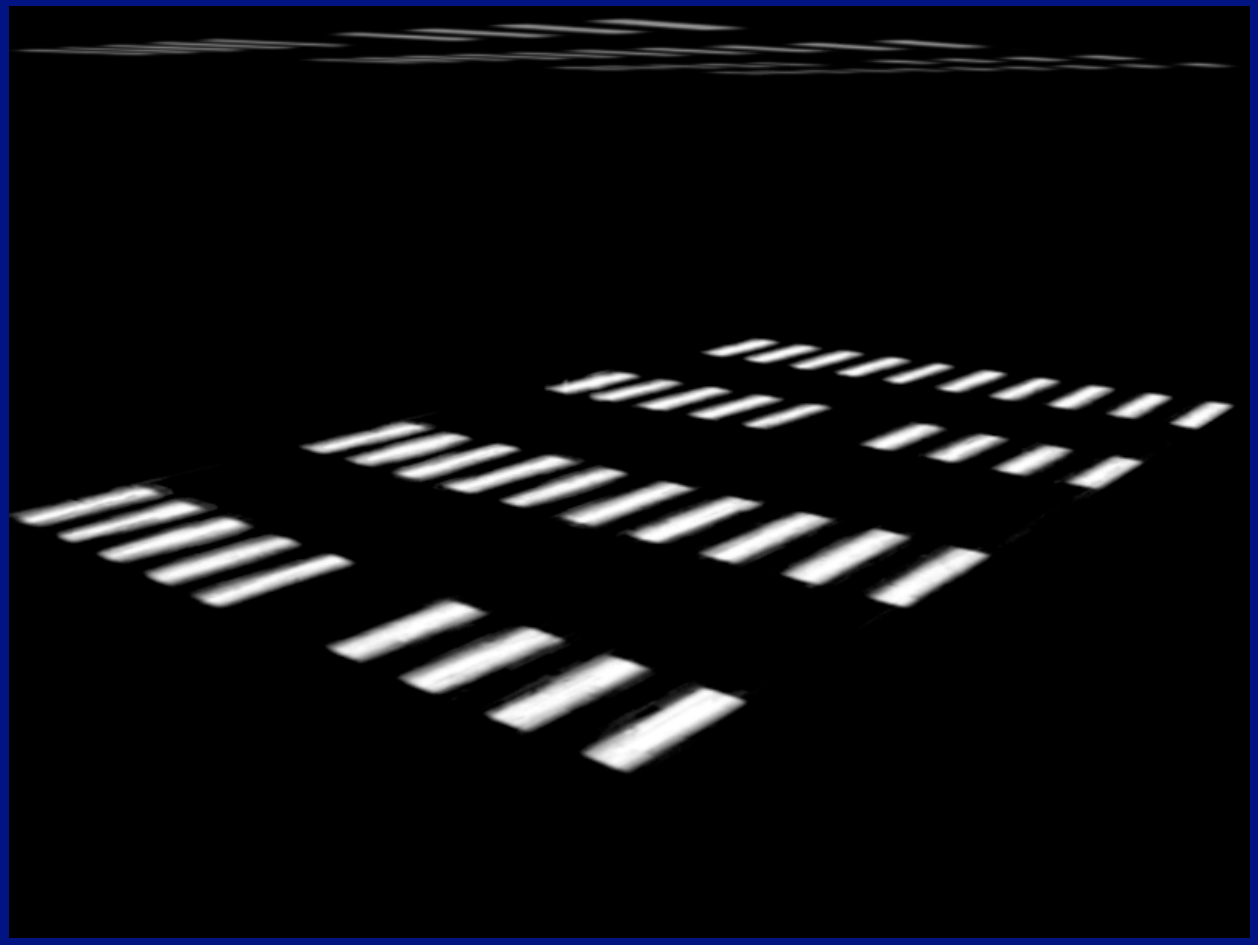




Shafts are “inverse shadows”

External light shafts





Shadow matting via Guo et al. [2011]
Many other shadow detectors, eg Panagoploulos ea 09, 10

Results

Results

Results

Results

Results

Results

Results

Results

Materials

- We know the light coming in
  - we can make pictures using it!
- We know the light coming out
  - in the picture
- With approximate shape, fit parametric material model
  - albedo, specular albedo, phong parameter, roughness
    - (stuff used by modern renderers)

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