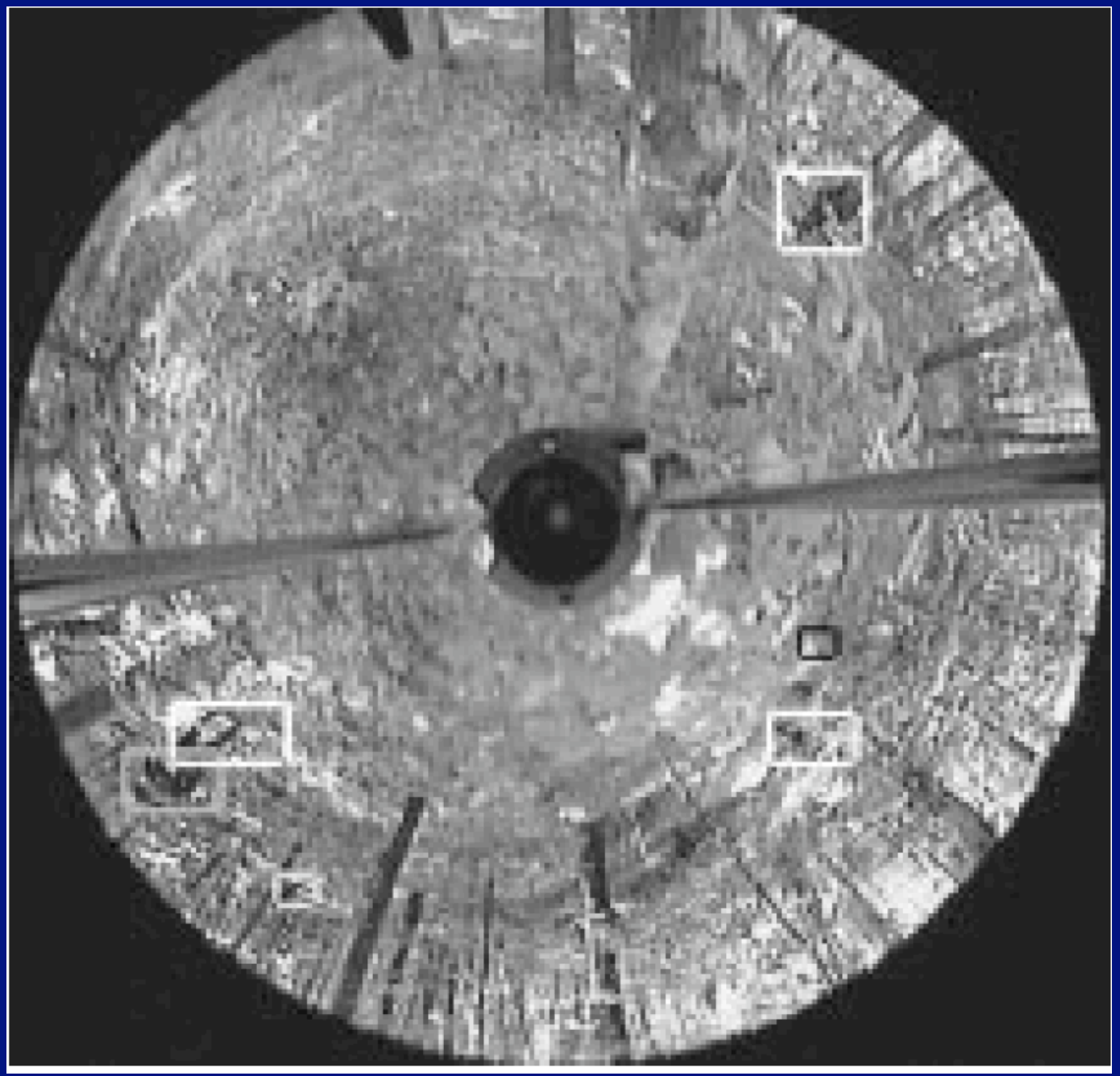
Looking at people (again!)

D.A. Forsyth, UIUC (was U.C. Berkeley; was U.Iowa)
contributions from:
Derek Hoiem (UIUC), 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

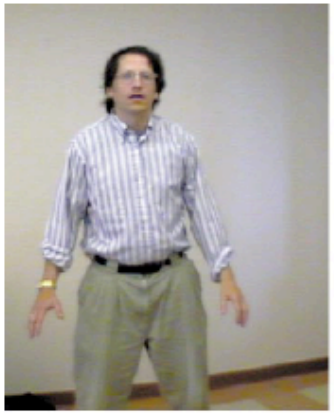
Thanks to: Electronic Arts, Sony SCEA, ONR MURI, NSF, DHS

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.

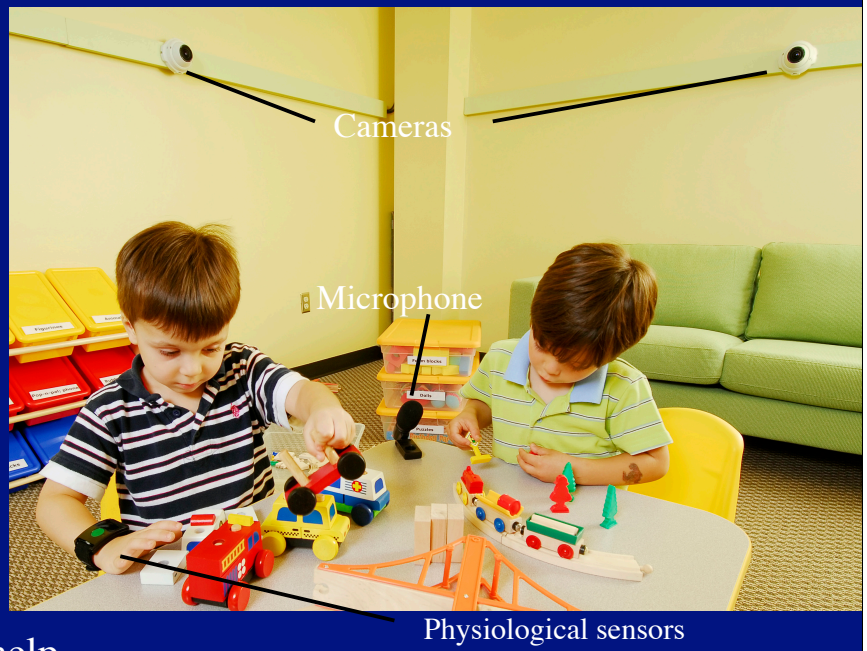
Freeman et al, 98.

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

Computational Behavioural Science



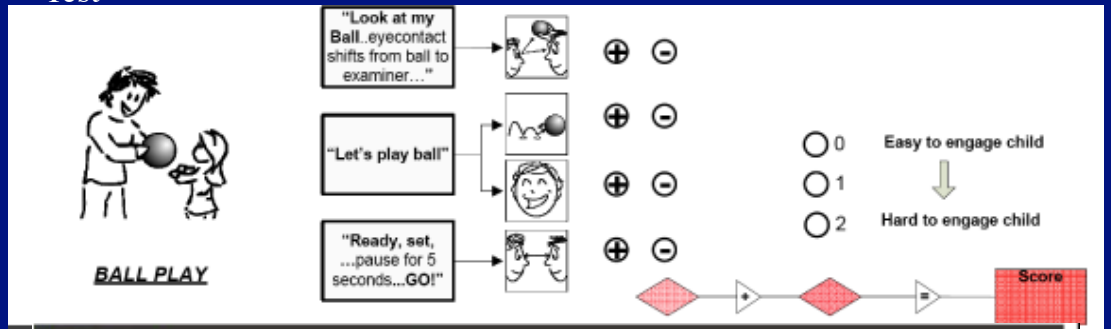
- Observe people
  - Using vision, physiological markers
  - Interacting, behaving naturally
  - In the wild
- drive feedback for therapy
  - Eg reward speech
- Applications
  - Model: screen for ASD
  - Other:
    - Any w here large scale observations help
      - Support in home care
      - Support 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

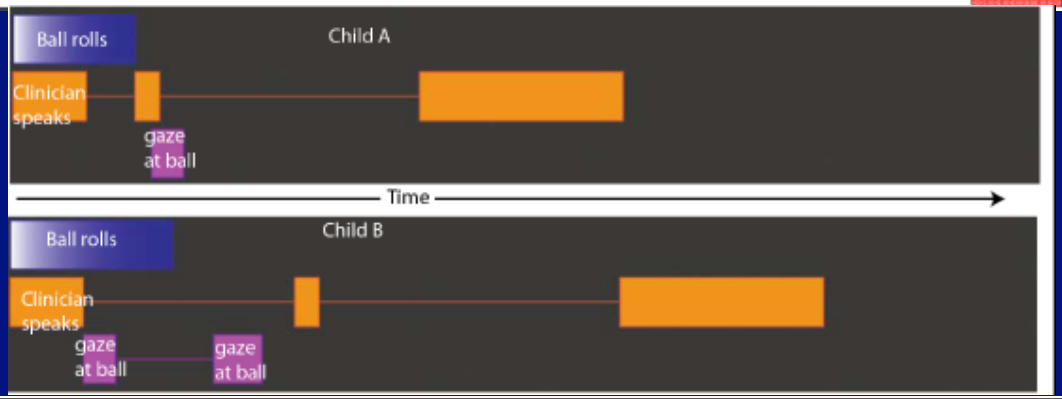
- Easily administered screening test
  - Challenge:
    - Automatic evaluation
    - To use unskilled screeners

View

Test

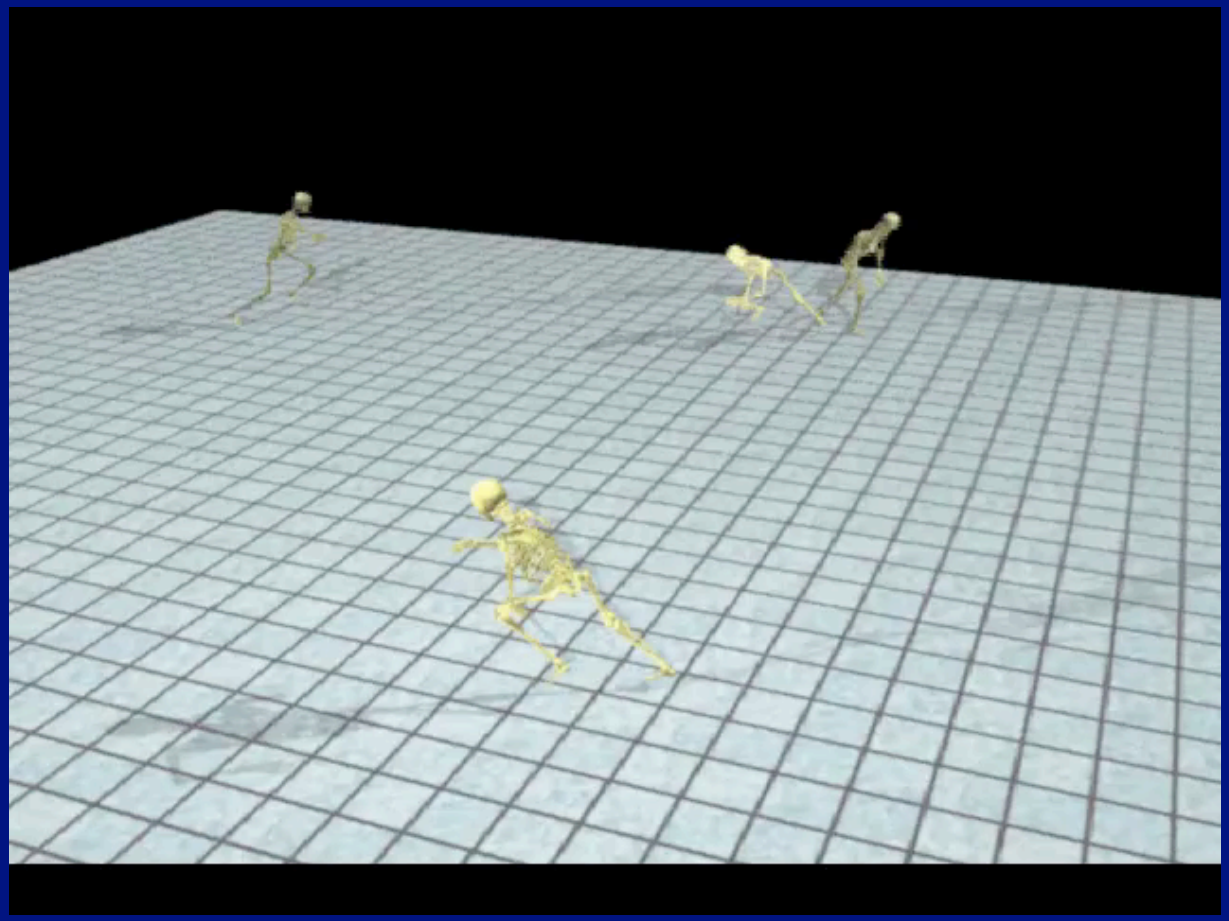


Outcome s



Why are humans important?

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

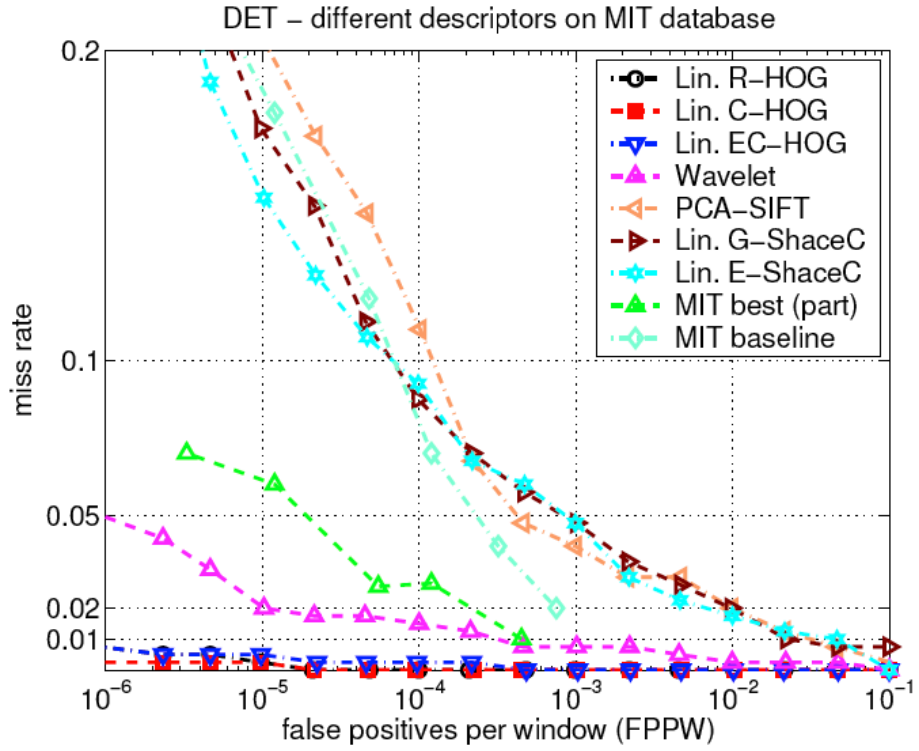


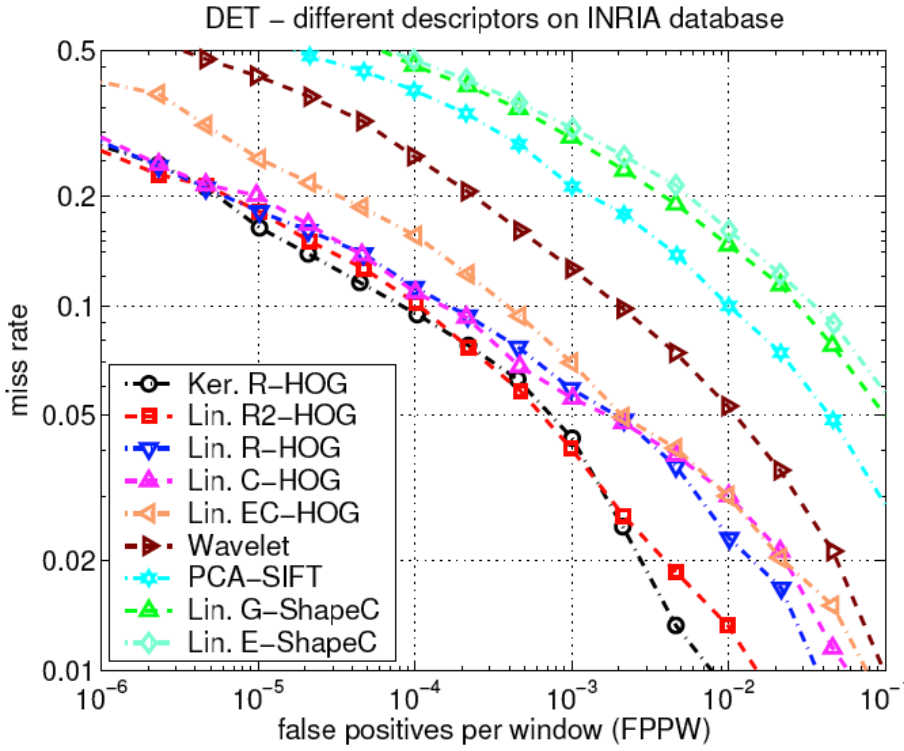
Arikan+Forsyth 02

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

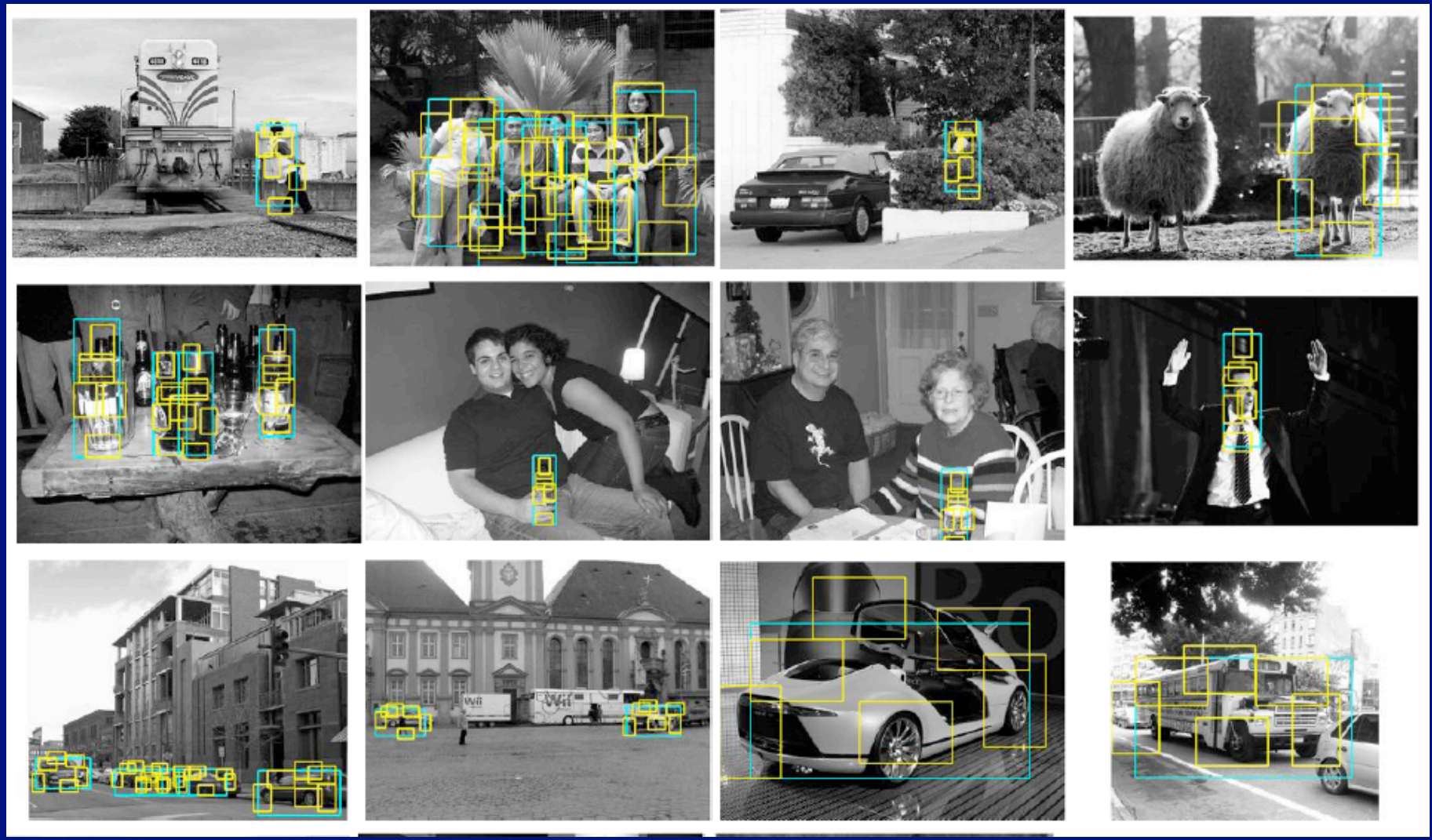
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/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

Structure

- What can we do?
  - mainly, tag some known activities with classifiers
- What should we be doing?
  - building representations to describe the unfamiliar
- How do we get information from the image signal?
  - tracking/parsing the body to get arms and legs
- What's the form of the representation?
  - contact, timing, style attributes

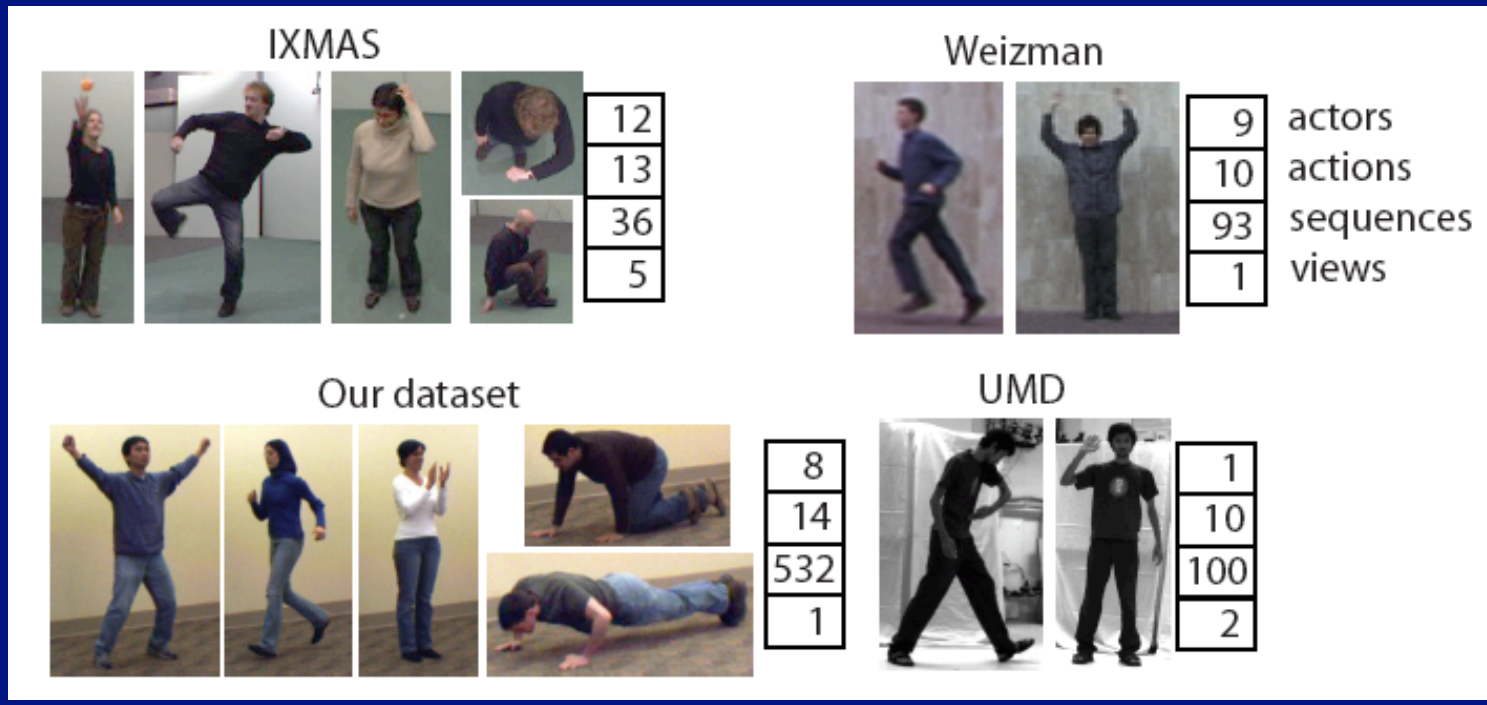
What we can do

- Primary machine is the classifier
  - features in, decision out
  - train with examples
- Decision is typically motion label
  - “run”, “walk”, “fight”, etc.
  - drawn from vocabularies of 5-50 (or so, depending on paper)



P. Felzenszwalb, D. McAllester, D. Ramanan. “A Discriminatively Trained, Multiscale, Deformable Par
Model” CVPR 2008.

Datasets

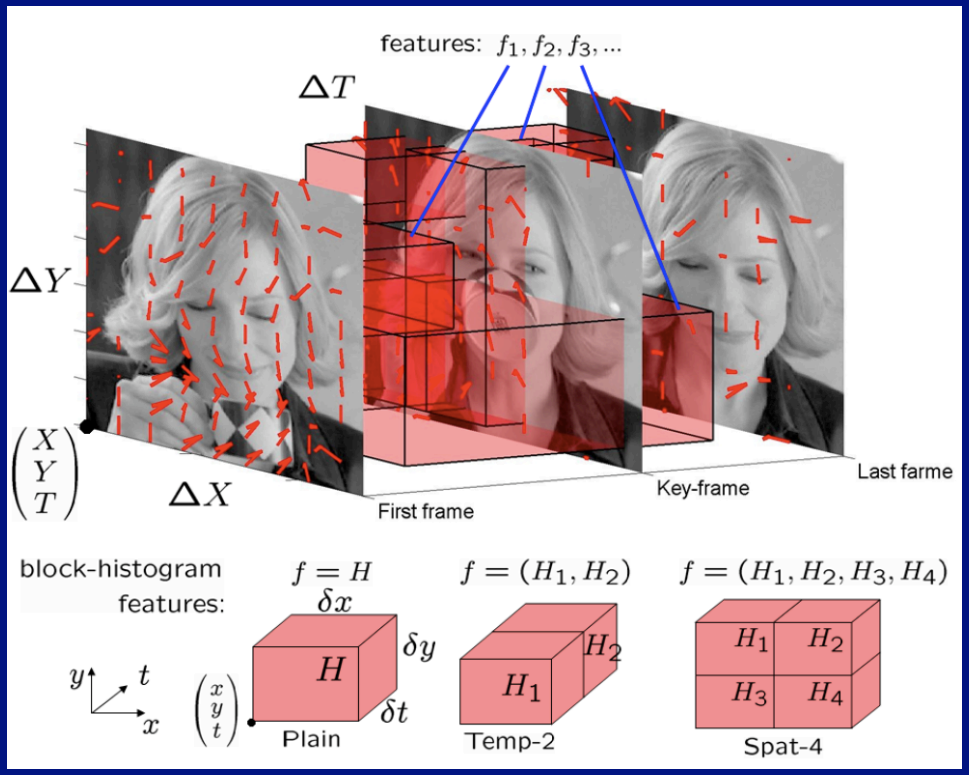

IXMAS
12
13
36
5
Weizman
9 actors
10 actions
93 sequences
1 views
Our dataset
8
14
532
1
UMD
1
10
100
2

Discriminative results

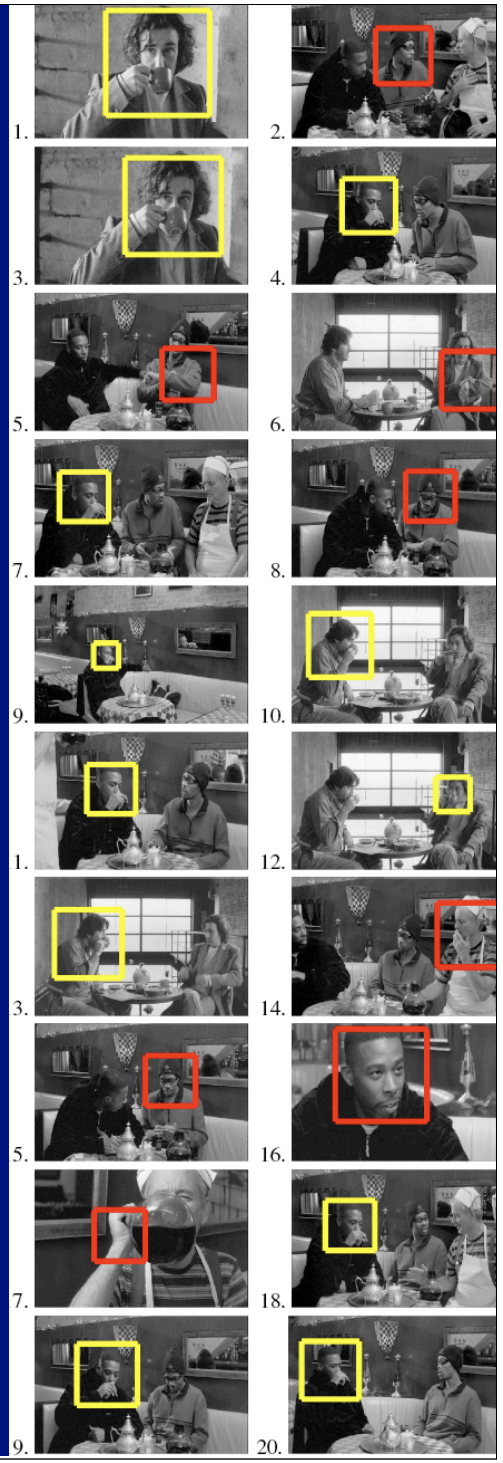
Dataset	Algorithm	Chance	Protocols								
			Discriminative task				Reject	Few examples			
			L1SO	L1AAO	L1AO	L1VO	UNa	FE-1	FE-2	FE-4	FE-8
Weizman	NB(k=300)	10.00	91.40	93.50	95.70	N/A	0.00	N/A	N/A	N/A	N/A
	1NN	10.00	95.70	95.70	96.77	N/A	0.00	53.00	73.00	89.00	96.00
	1NN-M	10.00	100.00	100.00	100.00	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

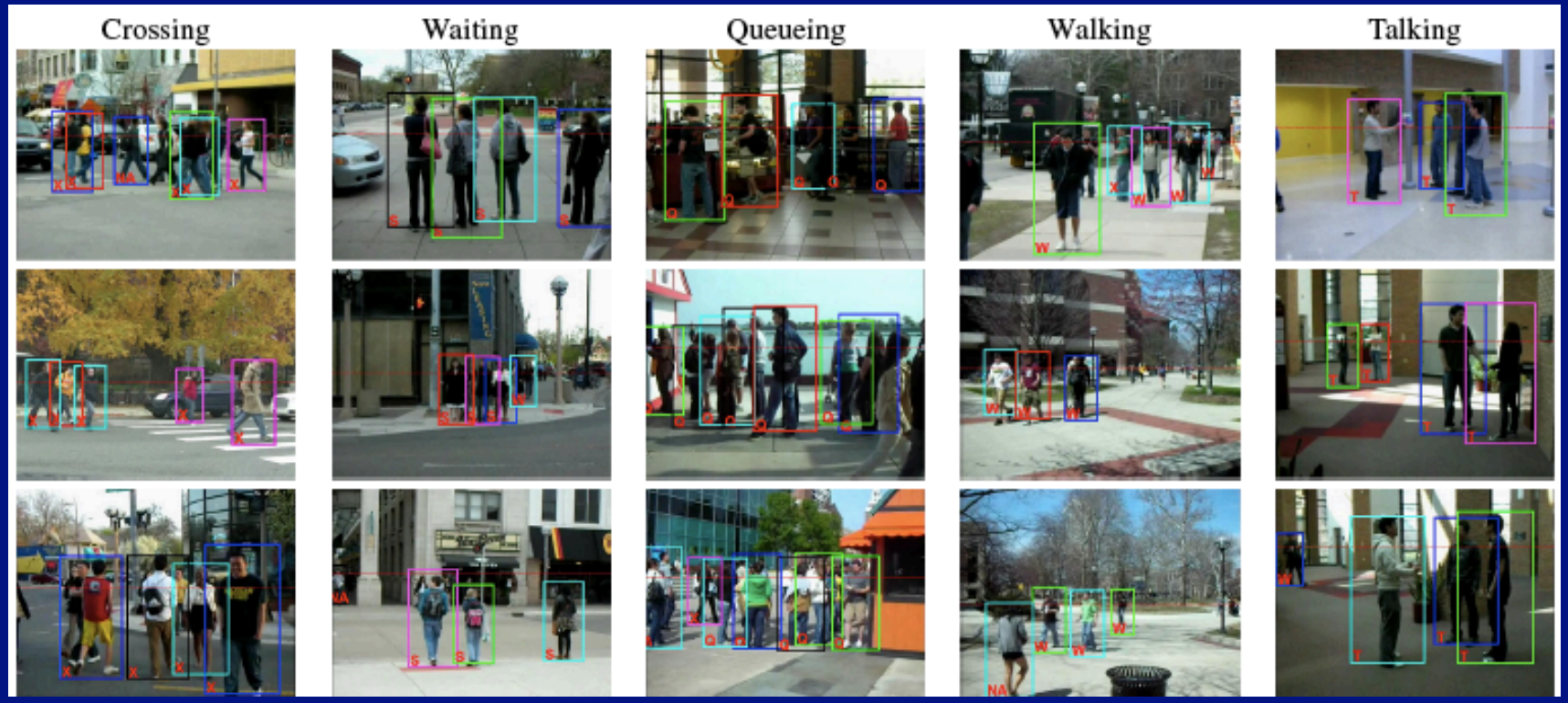
Tran + Sorokin 08



Laptev Perez 2007
see also Laptev et al 08

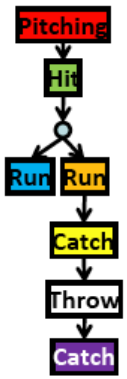


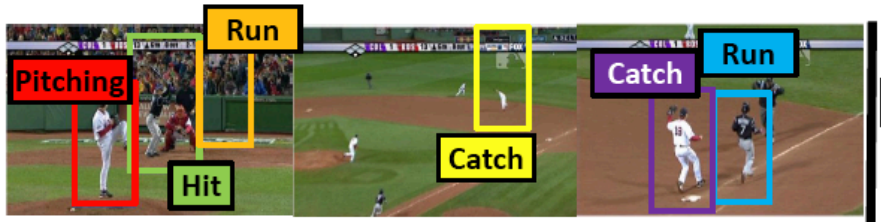
Movies and captions: Laptev et al 08



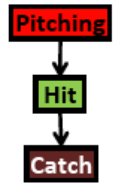
Choi Shahid Savarese 09

Predicting stylized narrations

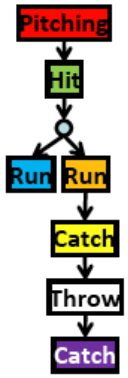


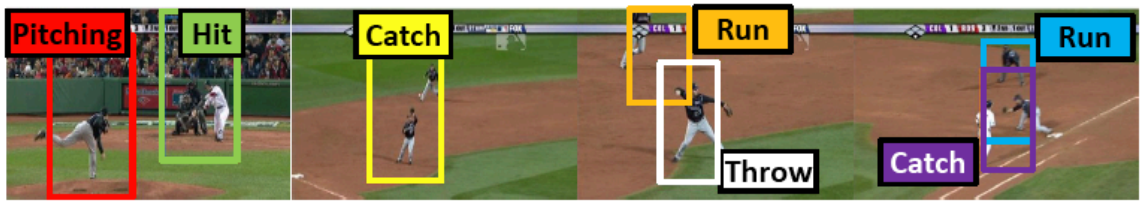


Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder catches the ball after Fielder runs towards the ball. Fielder catches the ball before Fielder throws to the base. Fielder throws to the base and then Fielder at Base catches the ball at base .

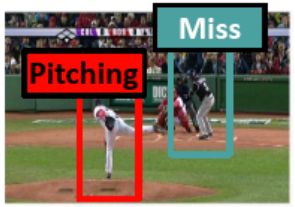


Pitcher pitches the ball and then Batter hits. Fielder catches the ball after Batter hits.





Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder runs towards the ball and then Fielder catches the ball. Fielder throws to the base after Fielder catches the ball. Fielder throws to the base and then Fielder at Base catches the ball at base .



Pitcher pitches the ball and then Batter does not swing.

Gupta ea 09

Structure

- What can we do?
  - mainly, tag some known activities with classifiers
- What should we be doing?
  - building representations to describe the unfamiliar
- How do we get information from the image signal?
  - tracking/parsing the body to get arms and legs
- What's the form of the representation?
  - contact, timing, style attributes

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?


PM

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 .

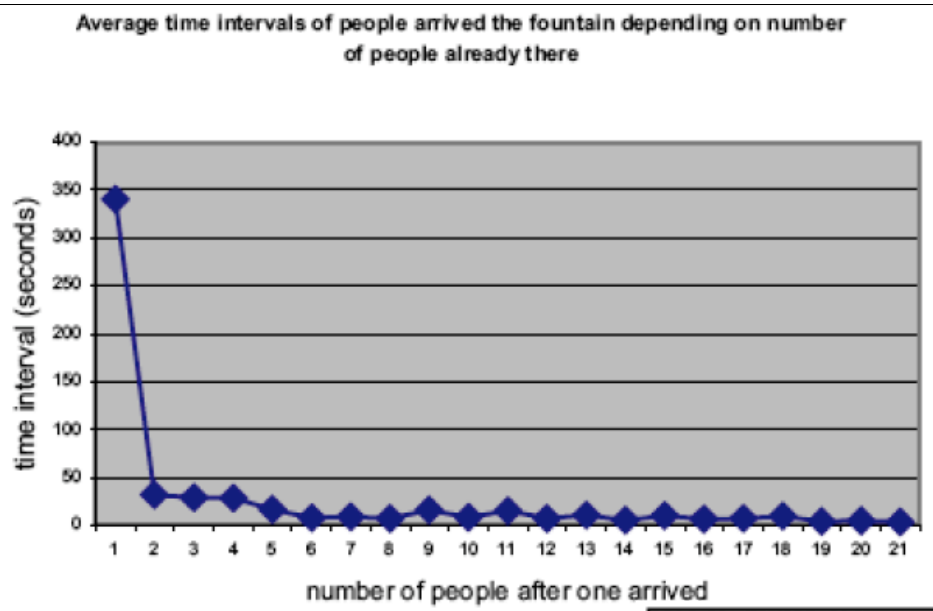
Sentences from Julia Hockenmaier's work

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.

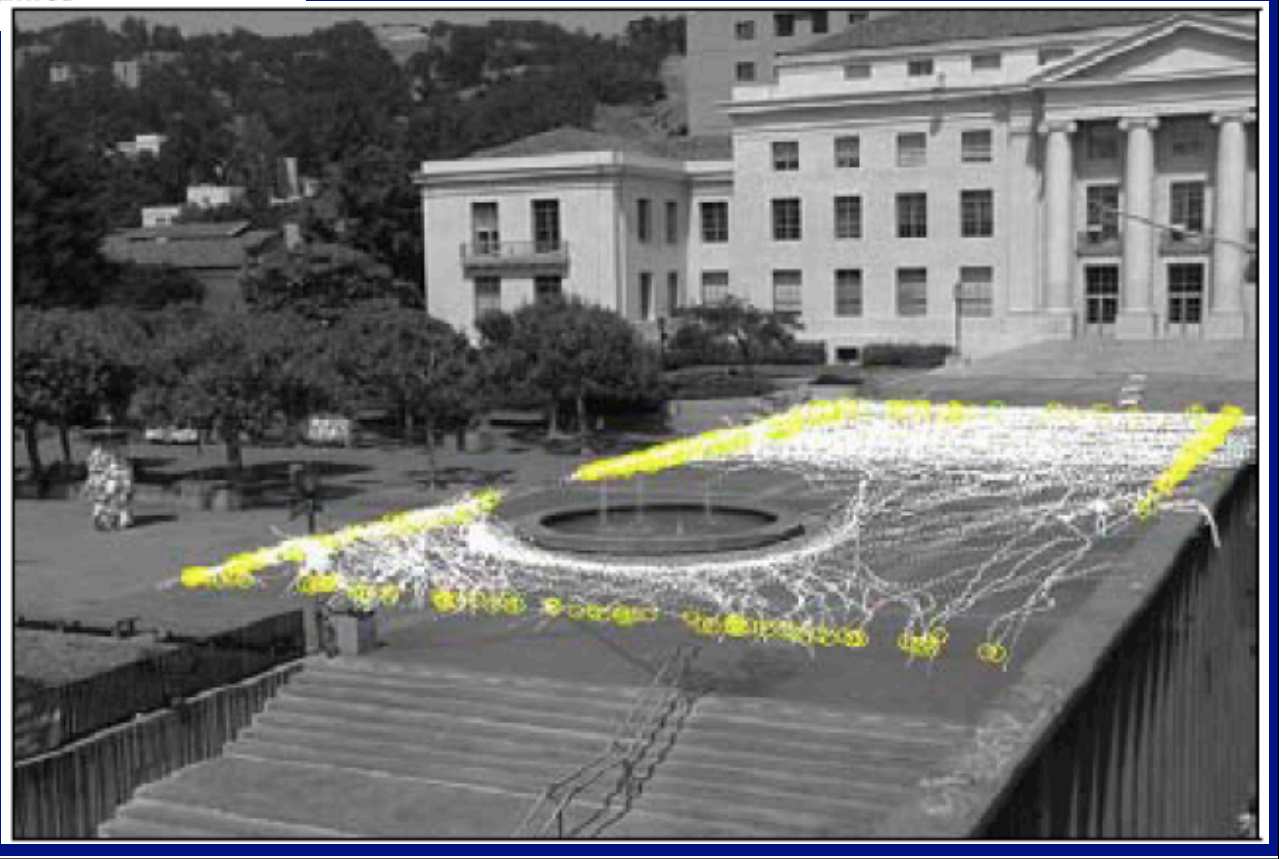
Structure

- What can we do?
  - mainly, tag some known activities with classifiers
- What should we be doing?
  - building representations to describe the unfamiliar
- How do we get information from the image signal?
  - tracking/parsing the body to get arms and legs
- What's the form of the representation?
  - contact, timing, style attributes



Point tracks reveal curious phenomena in public spaces

Yan+Forsyth, 04



Goals, intentions, outcomes

- Probably need to know some of body configuration
  - to reason about current contacts
    - man is on bicycle
    - woman is on platform
  - to reason about future contacts, eg
    - man is flying off bicycle and will hit water
    - woman is reaching for baby carriage
  - to reason about unfamiliar movements
    - what is he doing with his arm?

Why is kinematic tracking hard?

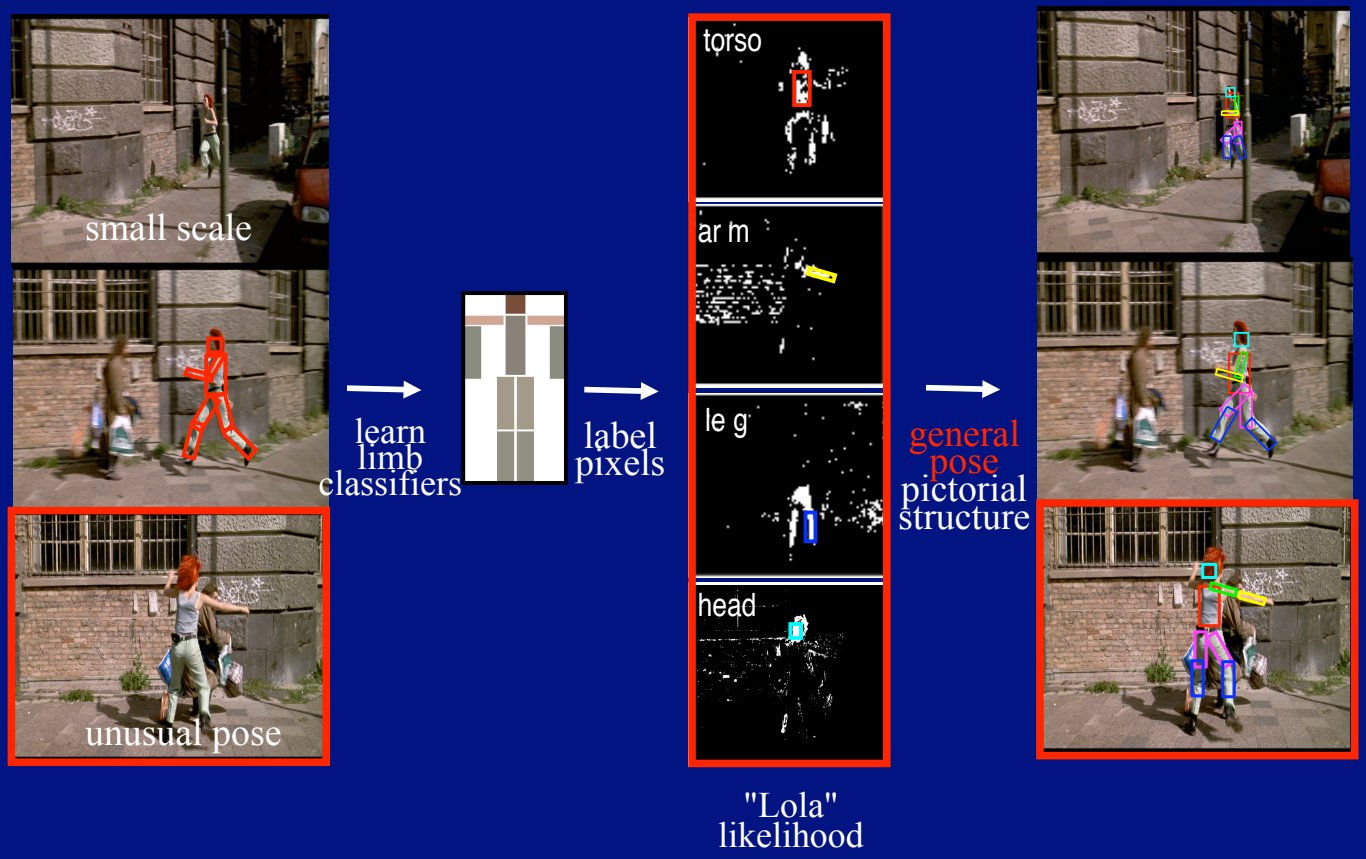
- It's hard to detect people
  - until recently, human trackers were manually started
- People move fast, and can move unpredictably
  - dynamics gives limited constraint on future configuration
  - appearance changes over time (shading, aspect, etc)
- Some body parts are small and tend to have poor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

variation in pose & aspect

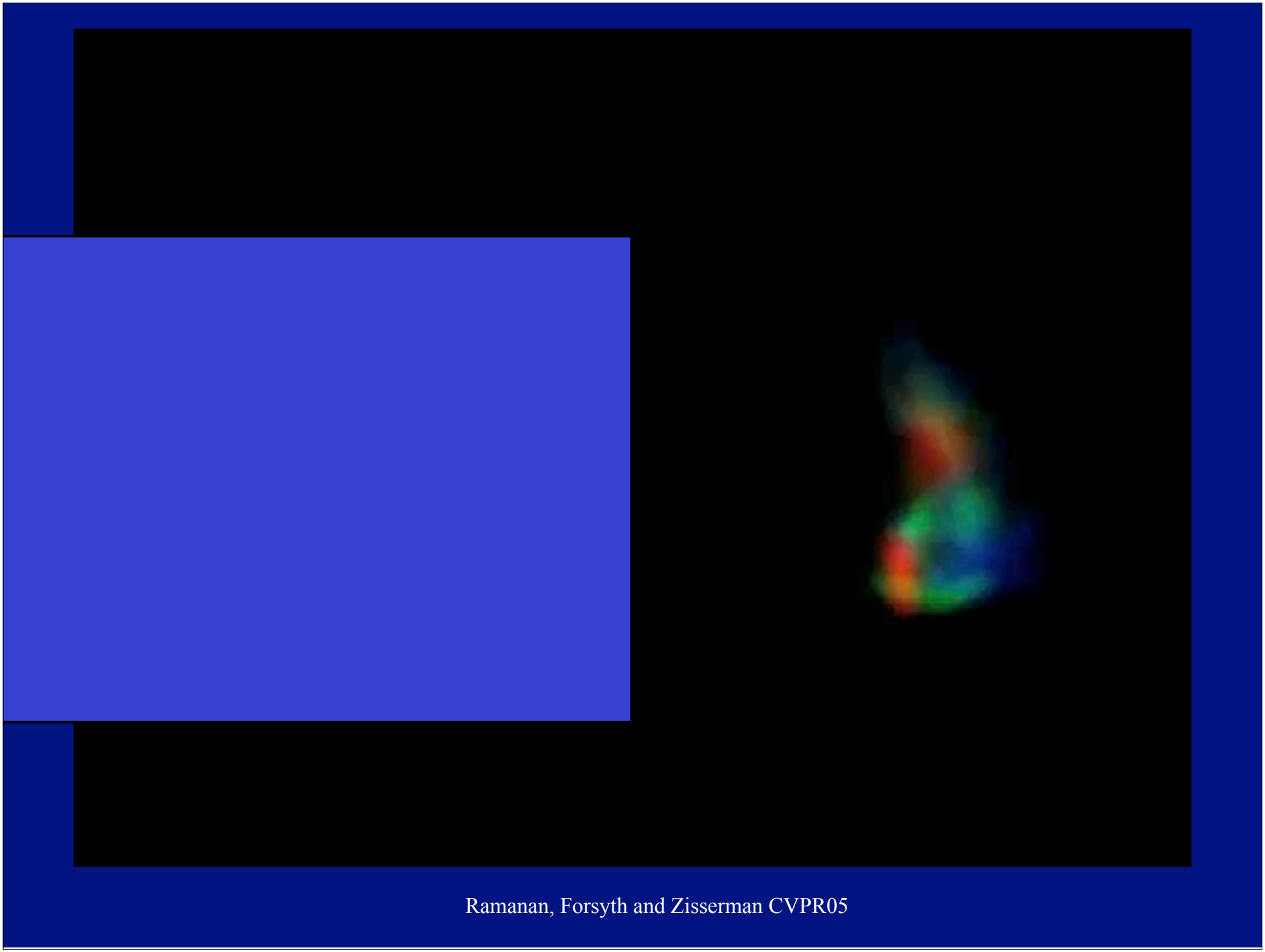
self-occlusion & clutter

variation in appearance

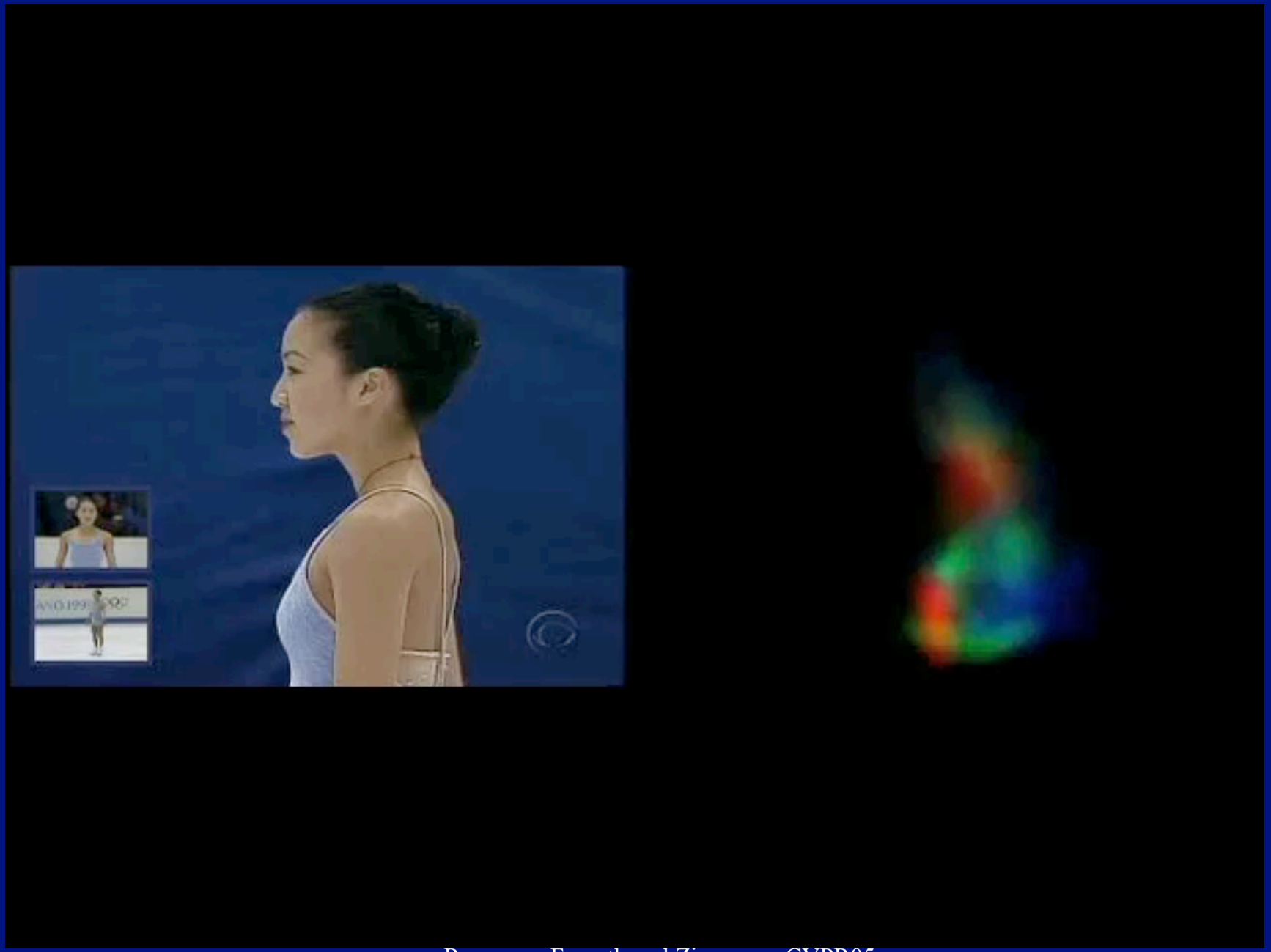
Build and detect models



Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05



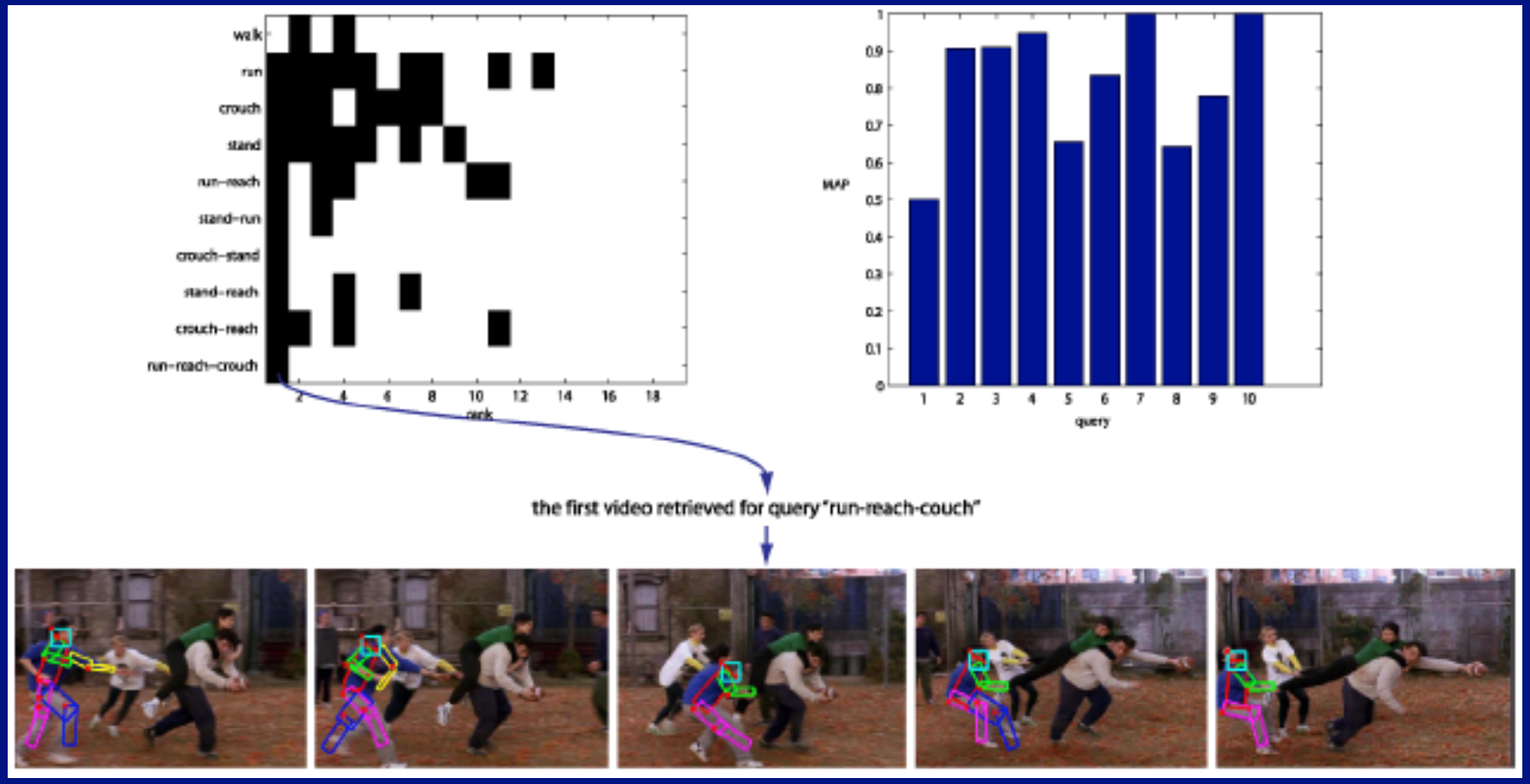
Ramanan, Forsyth and Zisserman CVPR05

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)
  - Complex spatial models (Sapp ea 10a)
  - Cascade models (Sapp ea 10b)
  - Full relational models (Tran Forsyth 10)

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



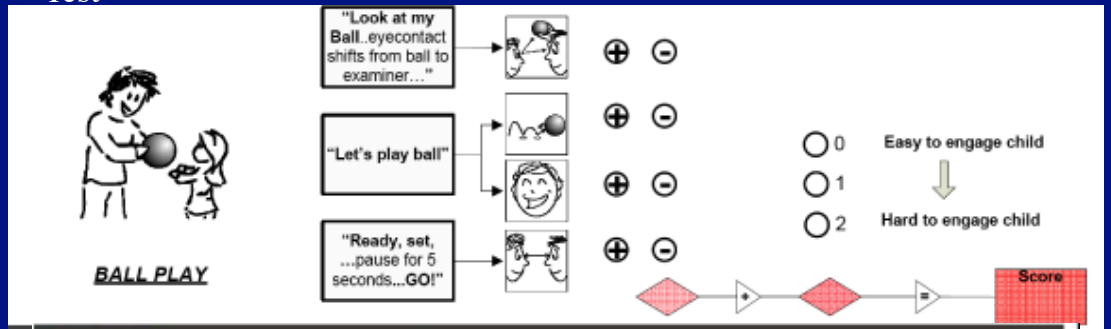
Searching for complex human activities with no visual examples N İkizler, DA Forsyth - IJCV, 2008

Rapid ABC

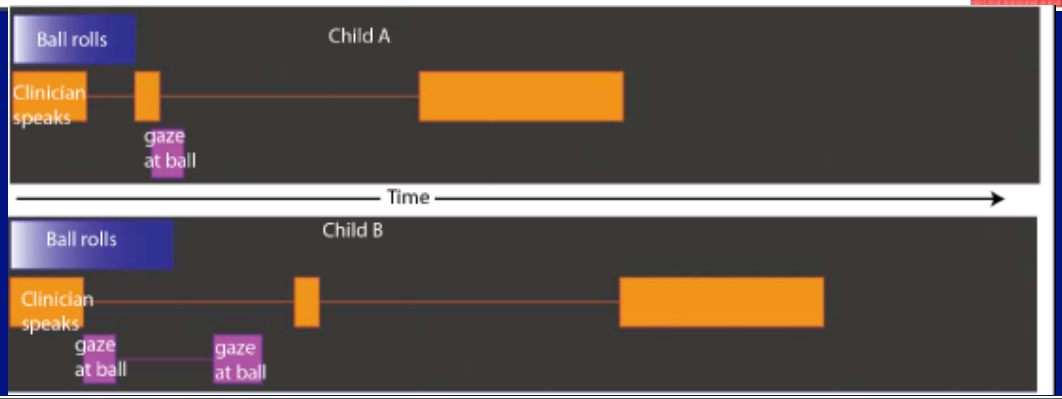
- Easily administered screening test
  - Challenge:
    - Automatic evaluation
    - To use unskilled screeners

View

Test



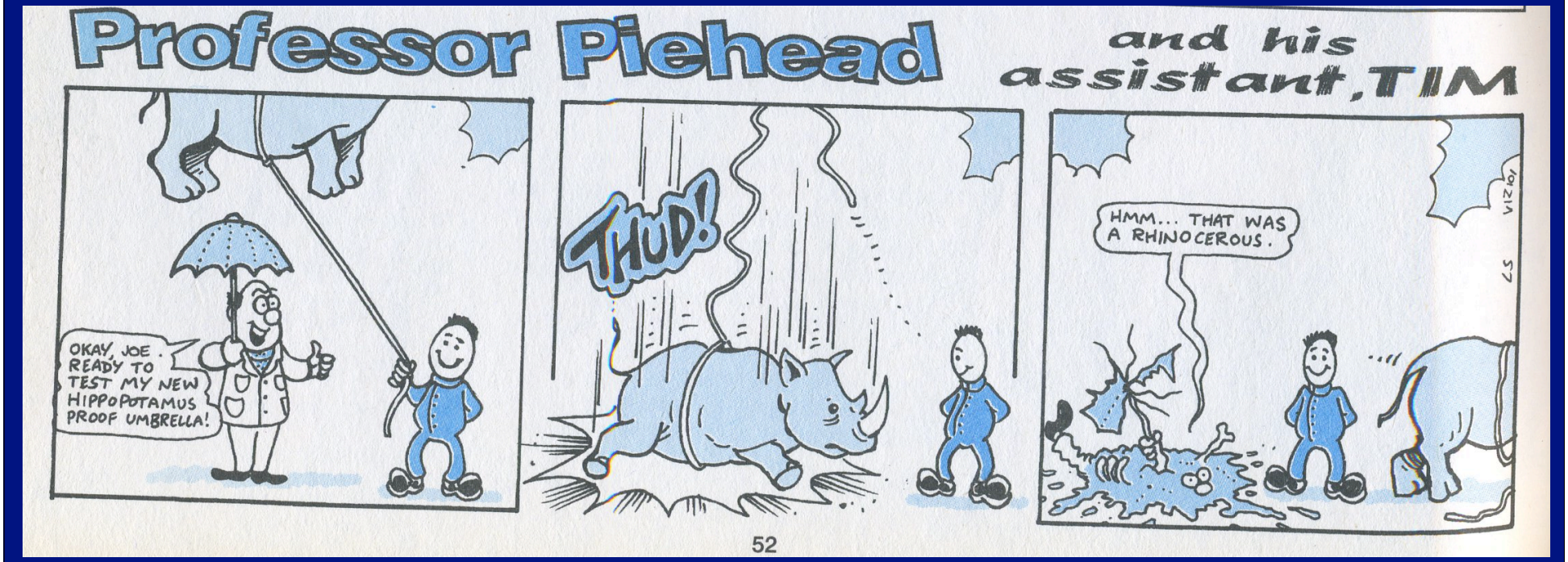
Outcome s



Structure

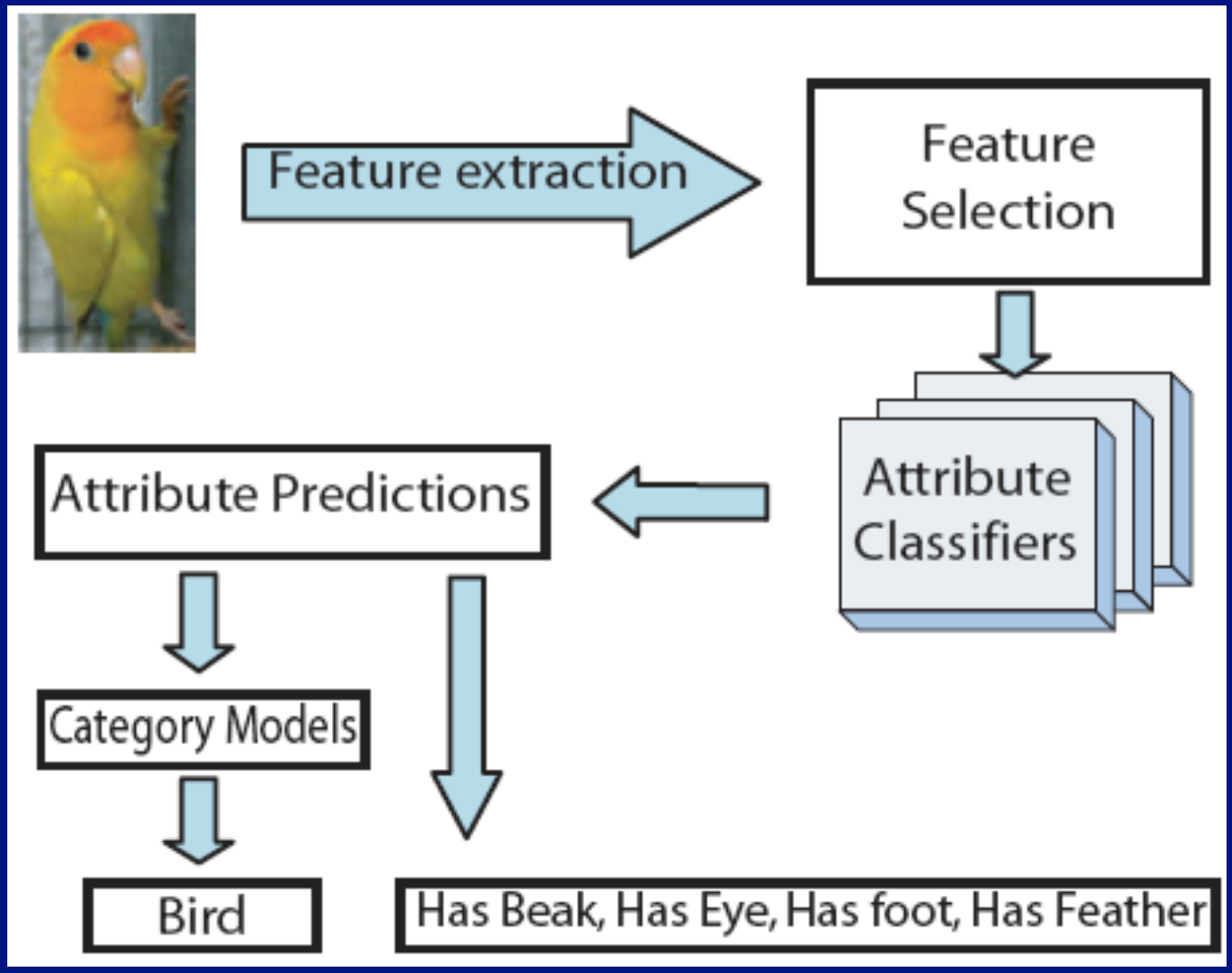
- What can we do?
  - mainly, tag some known activities with classifiers
- What should we be doing?
  - building representations to describe the unfamiliar
- How do we get information from the image signal?
  - tracking/parsing the body to get arms and legs
- What's the form of the representation?
  - contact, timing, style attributes

What is an object like?



Viz comic, issue 101

Possible architecture



Attribute phenomena

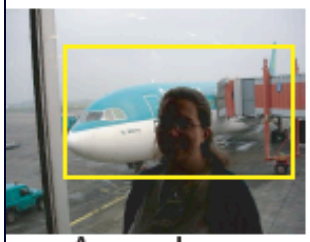
- Some are easily predicted from pictures
  - eg “red”, “wooden”
- Some are properly inherited from category
  - eg “mammal”
- They are heavily correlated
  - easy binary variable argument
- Some are “stuff”-like
  - eg “red”, “wooden”
- Others “thing”-like
  - eg “wheel”, “leg”
- Within class variation
  - Different instances of the same category could have different attributes

“Stuff” -- shape doesn’t matter (sky, grass, bush)
cf mass noun

“Thing” -- shape matters (cow, cat, car)
cf count noun

Farhadi ea 09

Missing attributes



Aeroplane
No "wing"

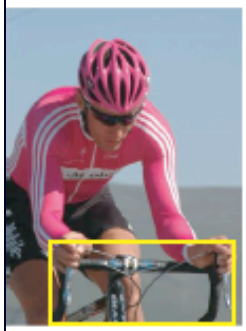
Car
No "window"

Boat
No "sail"

Aeroplane
No "jet engine"

Motorbike
No "side mirror"

Car
No "door"

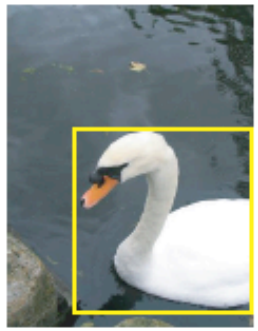


Bicycle
No "wheel"

Sheep
No "wool"

Train
No "window"

Sofa
No "wood"



Bird
No "tail"

Bird
No "leg"

Bus
No "door"

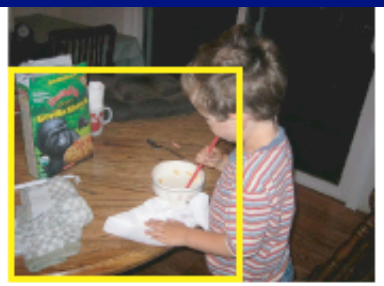
Farhadi ea 09

Extra attributes

Bird
"Leaf"

Bus
"face"

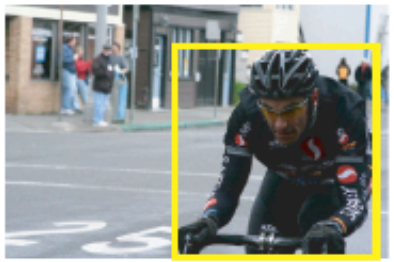
Motorbike
"cloth"



DiningTable
"skin"

People
"Furn. back"

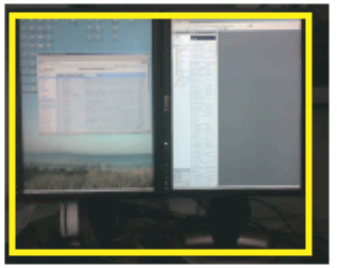
Aeroplane
"beak"



People
"label"

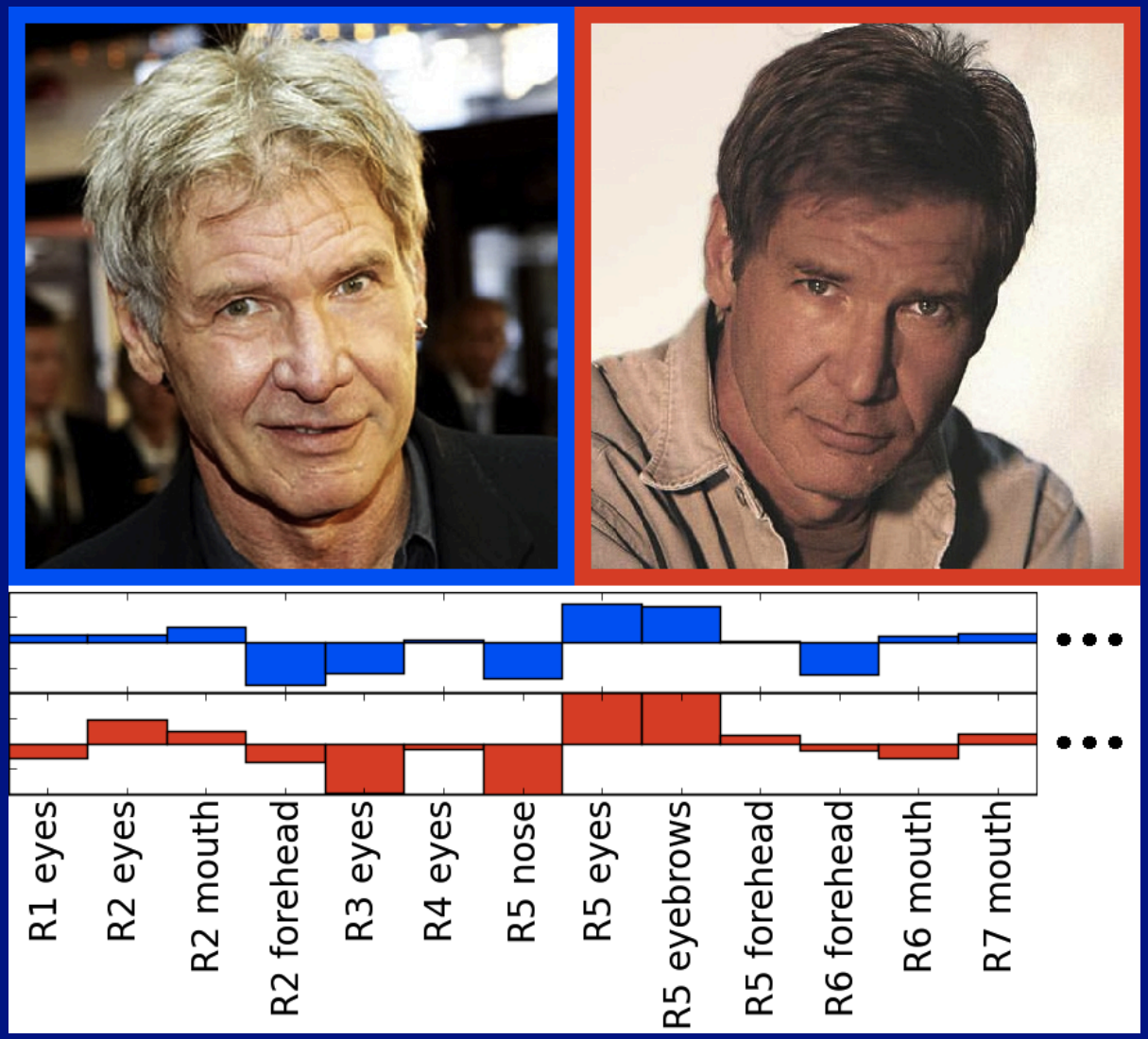
Sofa
"wheel"

Bike
"Horn"



Monitor
window"

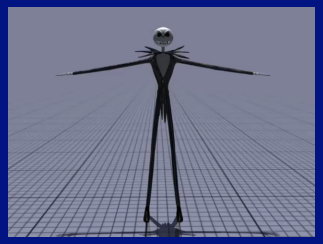
Farhadi ea 09



"Attribute and Simile Classifiers for Face Verification," ICCV 2009. (N. Kumar, A. Berg, P. Belhumeur, S. K. Nayar)

Activity attributes

- Gaze and focus
- Style
  - Fast/Gentle



- Timing
  - arms in phase with legs

- Contact
  - Having hand contact

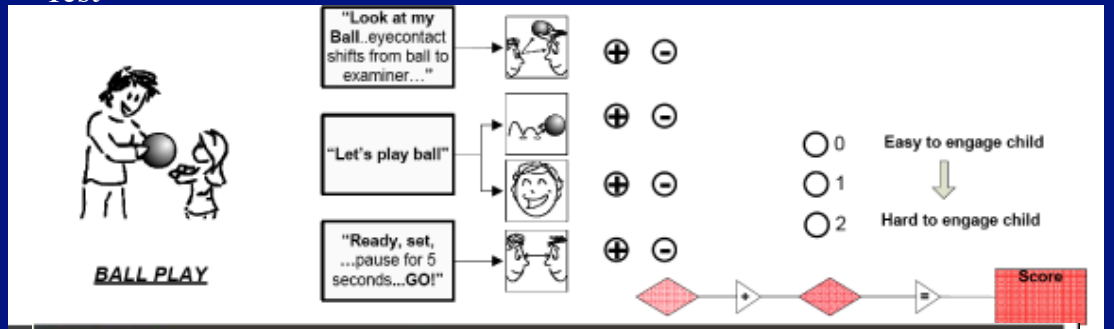
- Kinematic
  - Arms sticking out

Nearby objects and free space

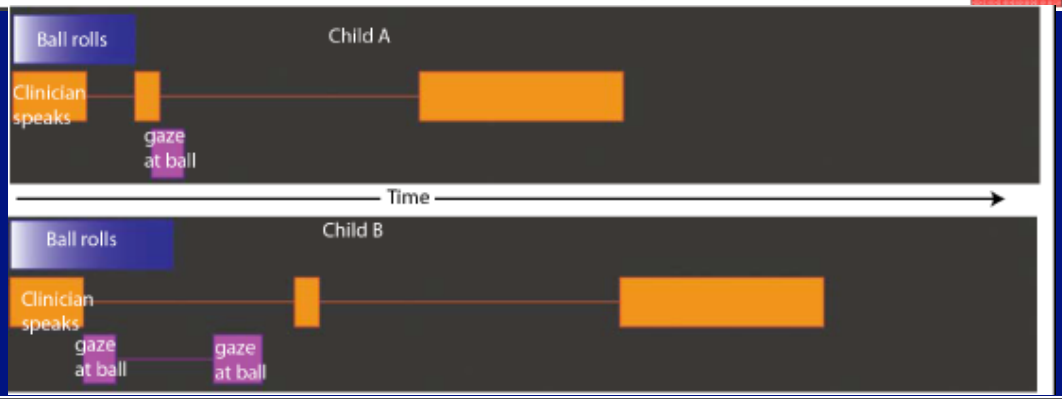
Gaze and focus: Rapid ABC

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

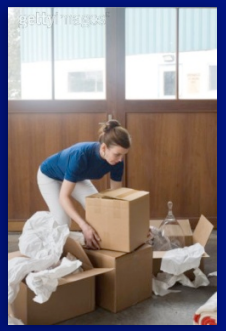
Test

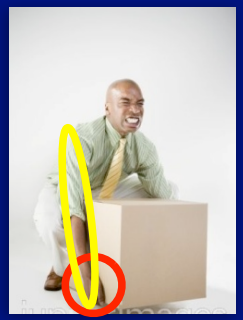


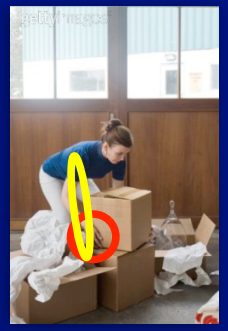
Outcome s

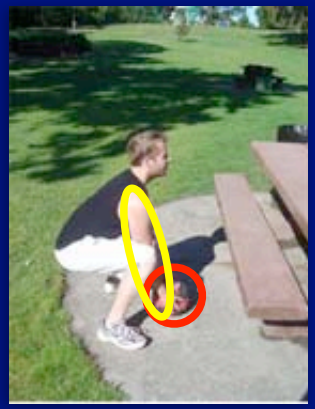


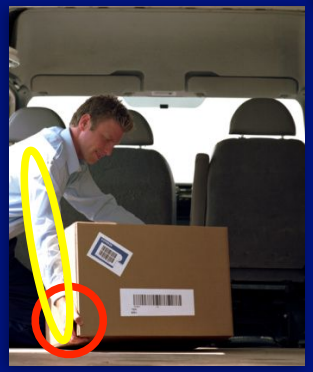
Contact and kinematics: Picking things up











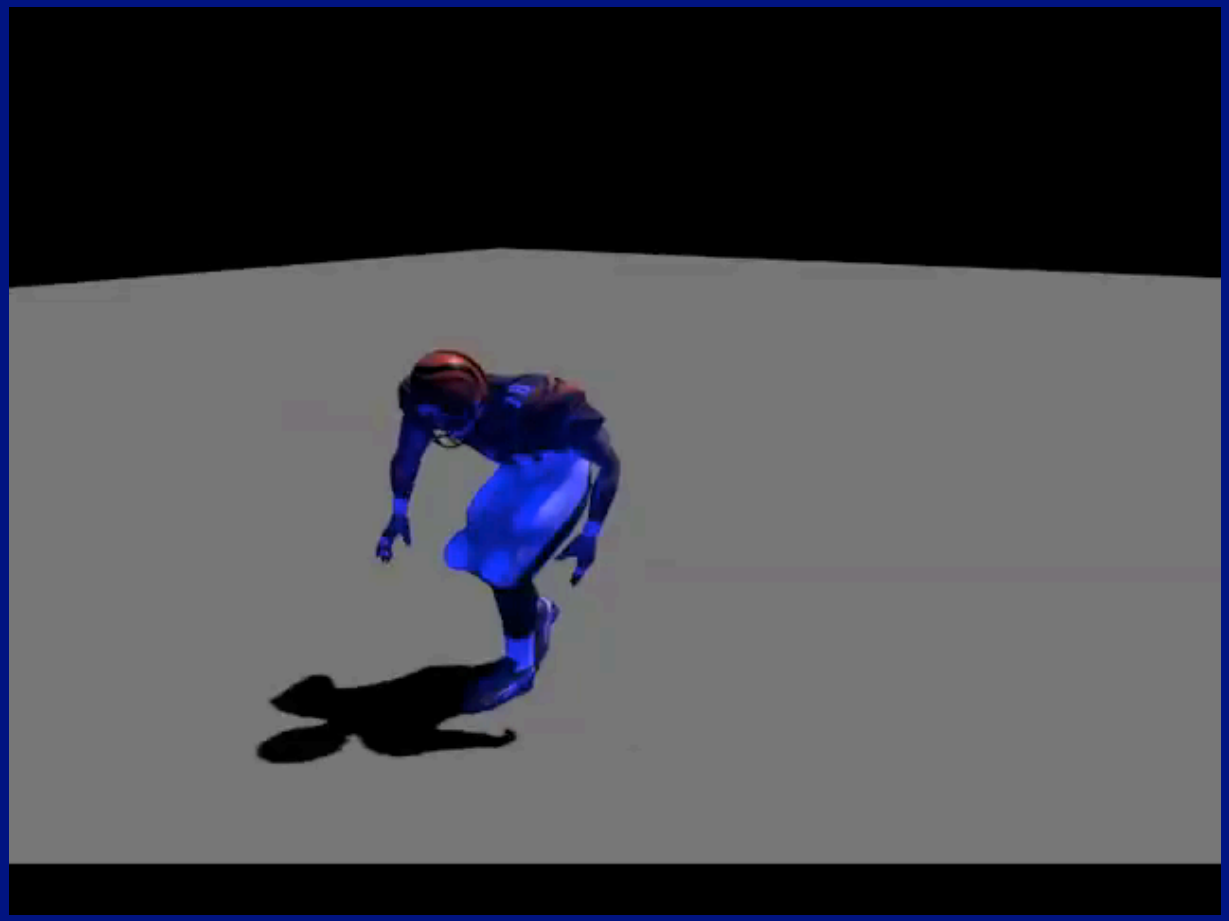
Animation tells us about attributes

- Relative timing of movements across the body matters
  - however, no real models here
- Contact matters
  - people are highly sensitive to incoherent contacts
- Style matters
  - people are good at consistency between motion style and body shape

Relative timing matters

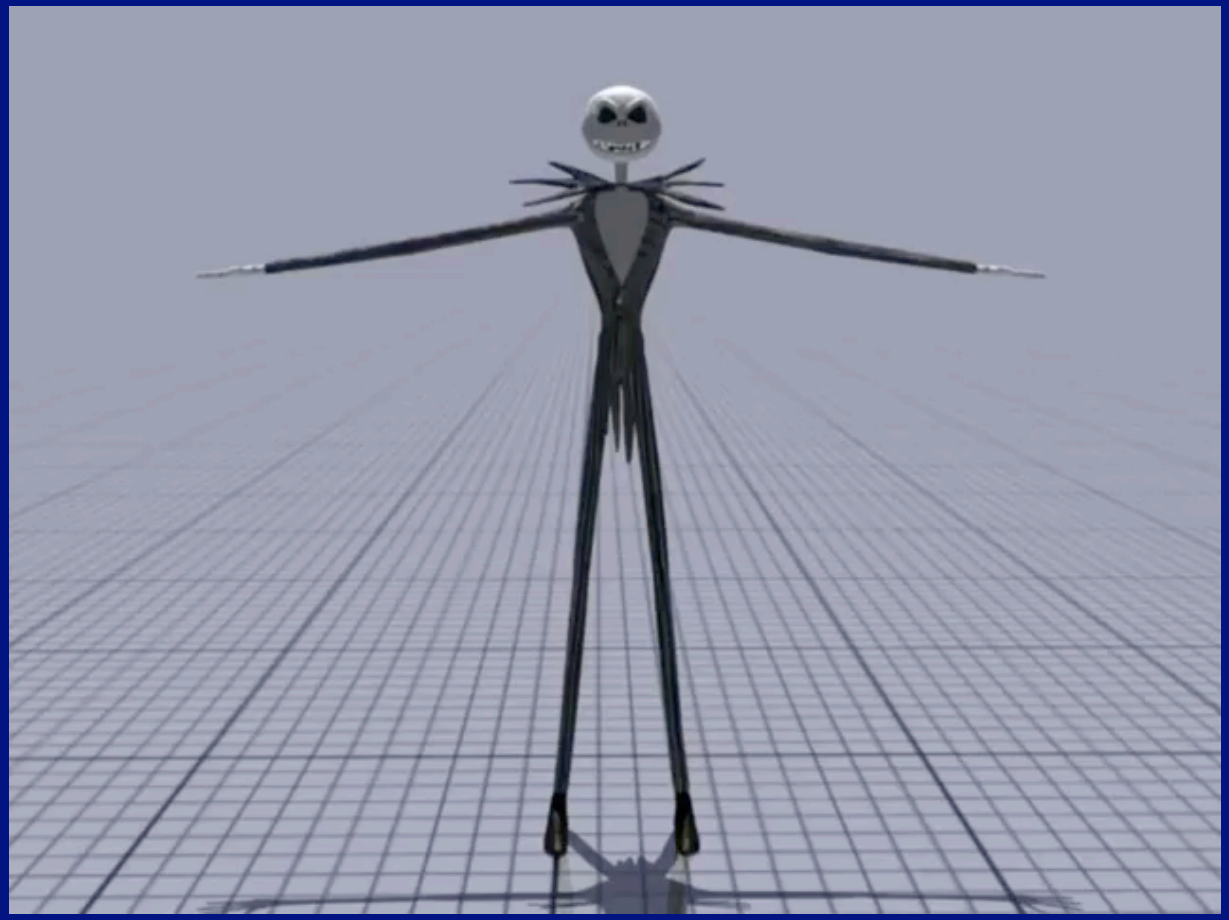
Ikemoto+Forsyth 04

Relative timing matters



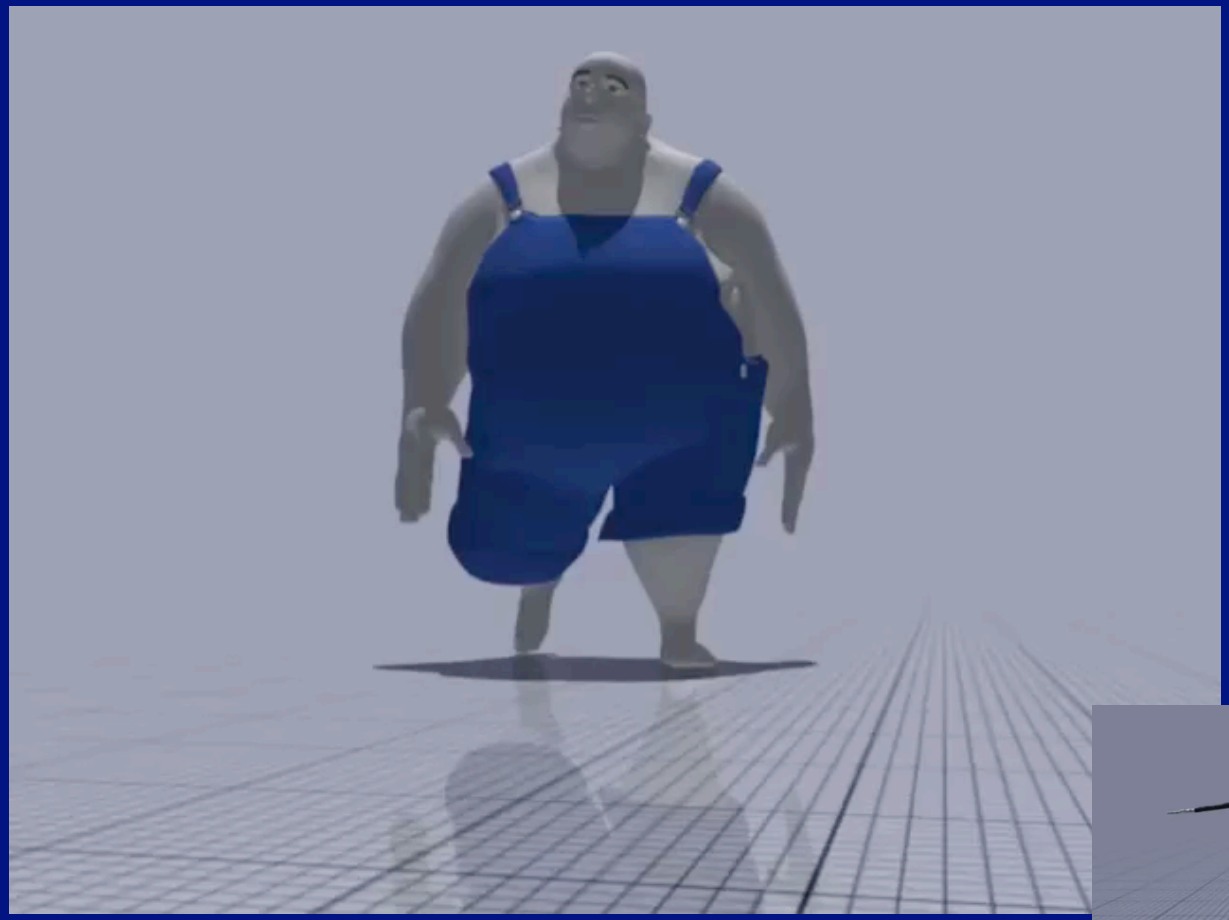
Ikemoto+Forsyth 04

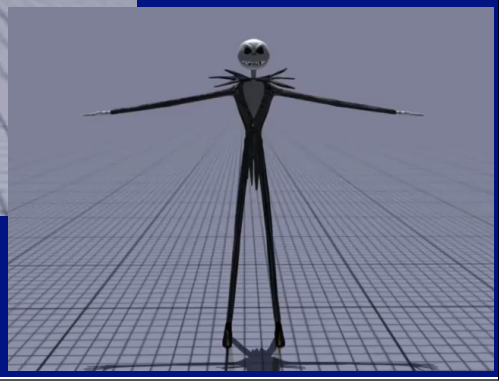
Different bodies have different styles



Ikemoto ea 09

Different bodies have different styles





Ikemoto ea 09

Open question: similarity

- This activity is like that one
  - therefore, the outcome might be similar
- In what way like? how do we score this?
- Advantages
  - strong improvements in training with few examples
    - (Wang, 10; some cases)
  - perhaps allows recognition/prediction with no examples

Summary

- Extend attribute based representations to describe activity
  - starting at least with
    - Gaze/focus
    - Style
    - Timing
    - Contact
    - Kinematics
    - Nearby objects or free space
- Select what is important from sequences
  - perhaps for predictive purposes
- Build procedures to use similarity of motion/outcome
  - to train models with little data

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