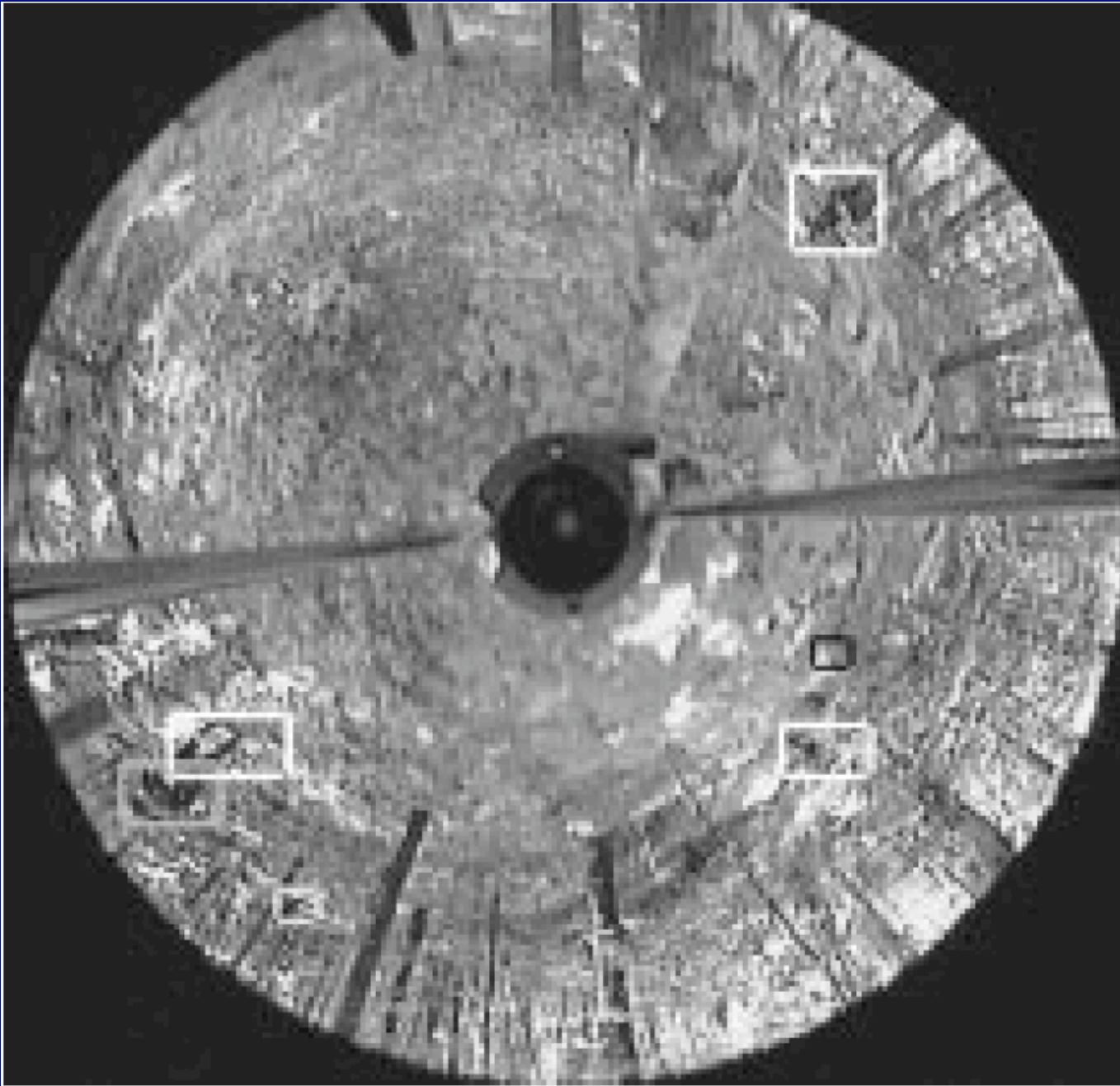


Human Detection and Tracking

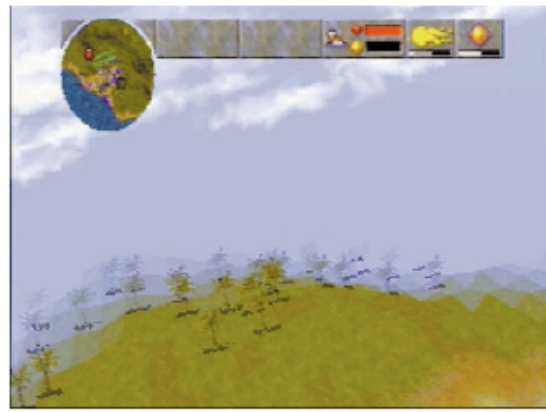
CS 543 - D.A. Forsyth

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

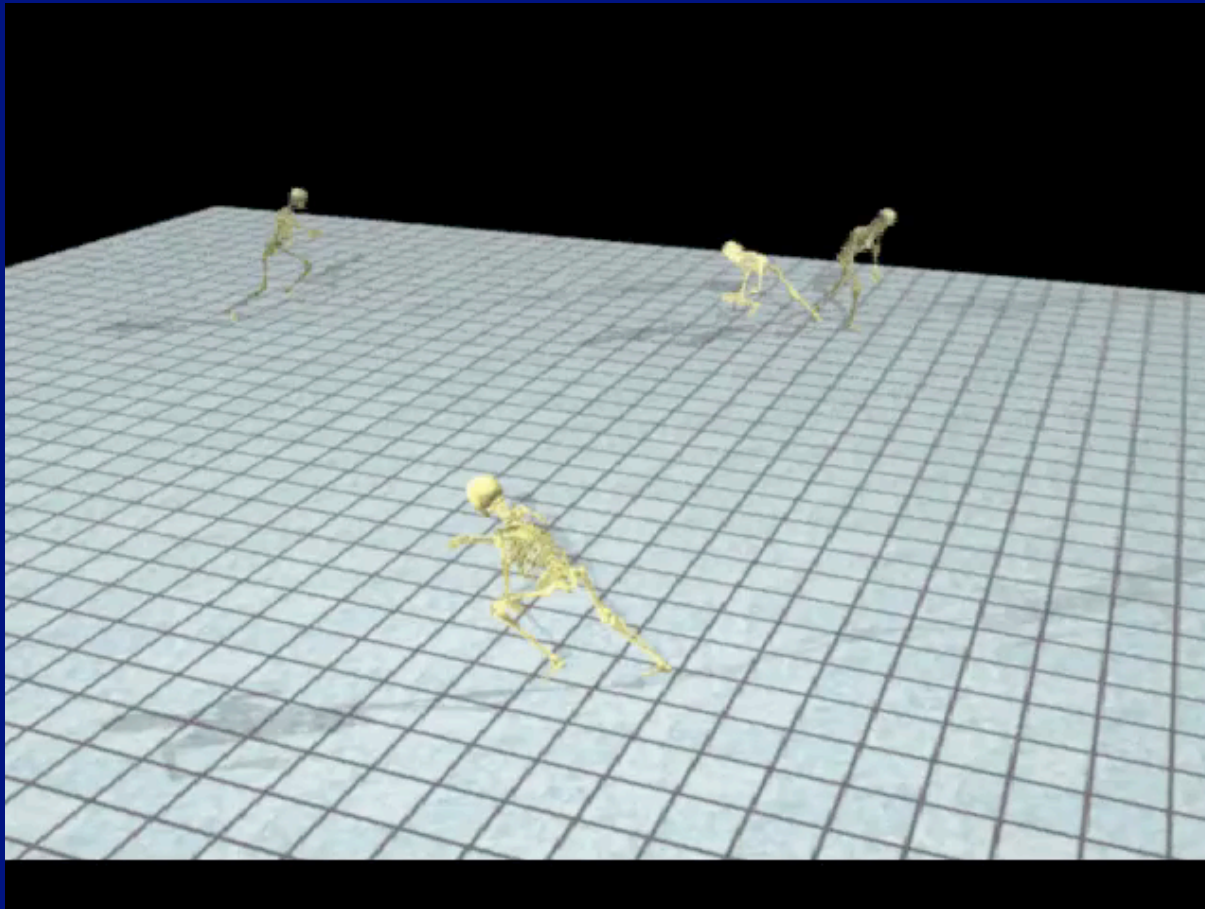


Sony's eyetoy estimates motion fields,
links these to game inputs.
Huge hit in EU, well received in US



Why are humans important?

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



Why are humans important?

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

Why are humans important?

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

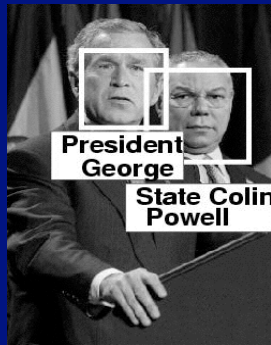
News Faces

- 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 $k \leq 4$
- 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





US President George W. Bush (L) makes remarks while Secretary of State Colin Powell (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/ Luke Frazza)



German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer Matthew Vaughn, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

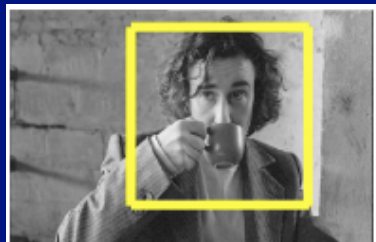


British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

Topics

- Transducing the body
 - getting some representation out of picture/video
- Inferring activity
 - going from that representation to a description of activity

Activity by Appearance



Points

- Quite simple features can reveal what people are doing
 - location
 - patterns of motion
 - “appearance”

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)

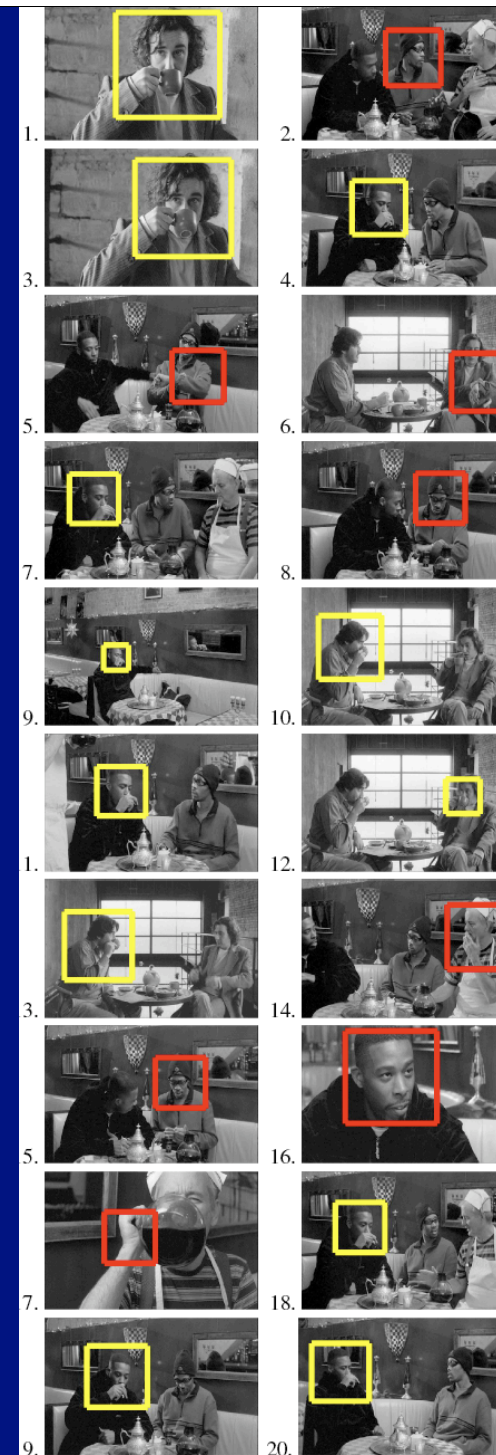
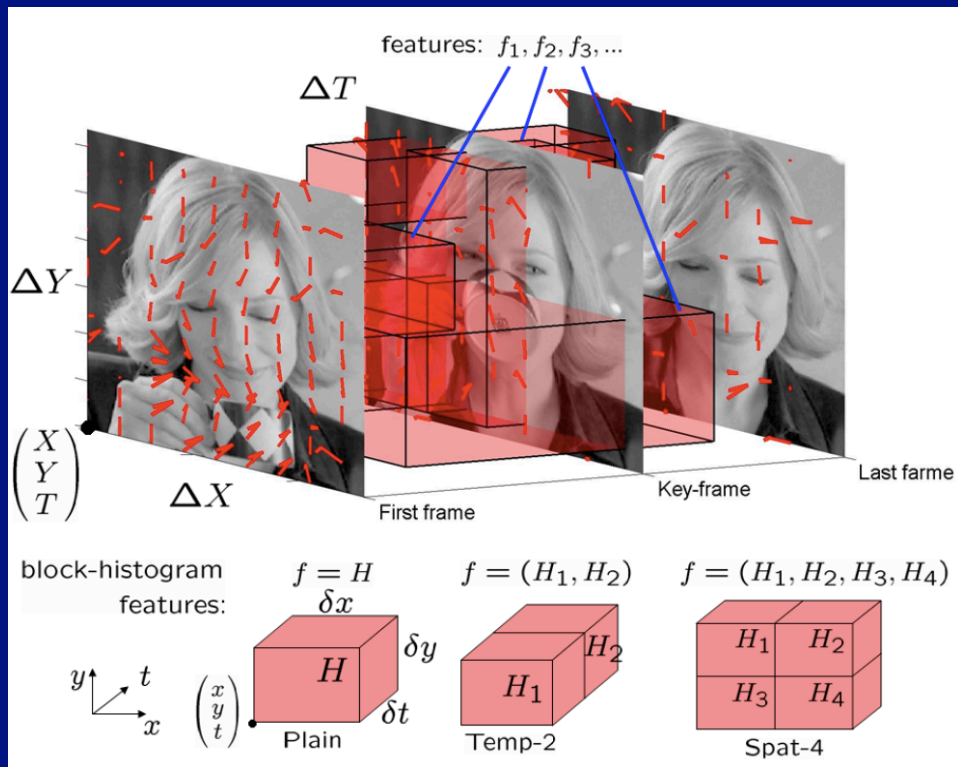


Bobick & Davis. PAMI01















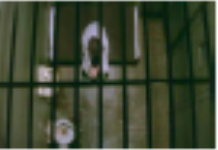
















Surveillance: Where you are can tell what you are doing



Intille et al 95, 97



Laptev Perez 2007
see also Laptev et al 08

	AnswerPhone	GetOutCar	HandShake	HugPerson	Kiss	SitDown	SitUp	StandUp
TP								
TN								
FP								
FN								

Movies and captions: Laptev et al 08

Datasets

IXMAS



Weizman



Our dataset



UMD

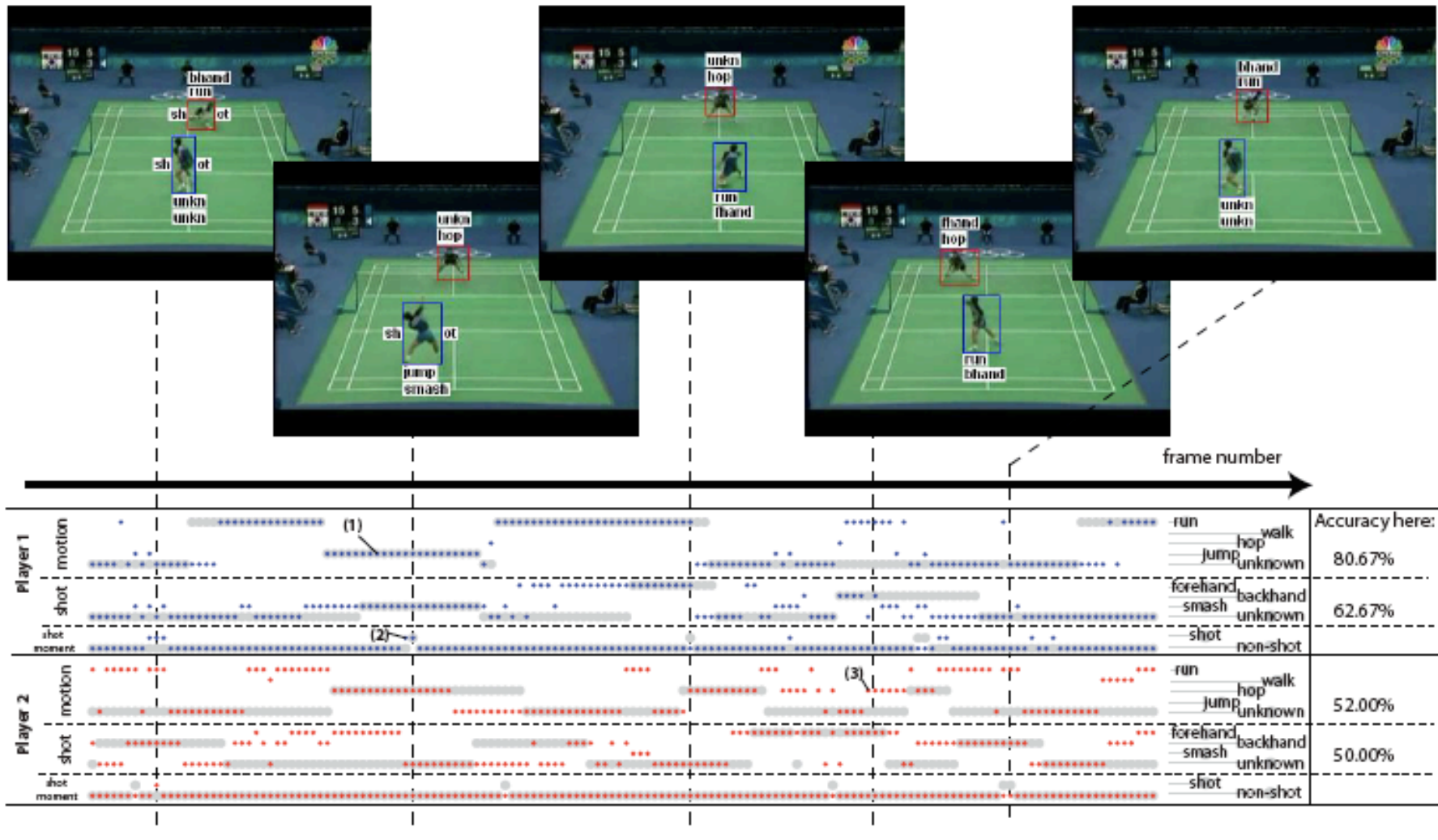


Discriminative results

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				

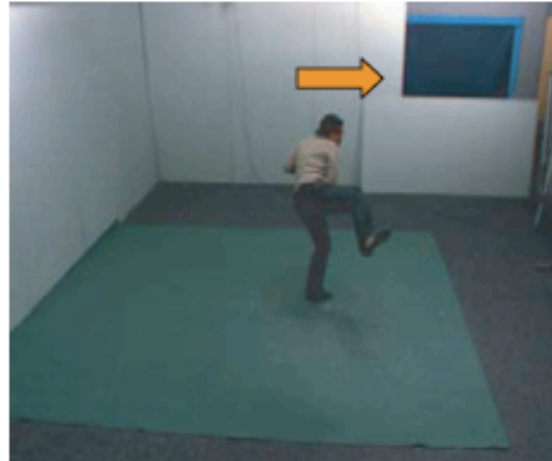
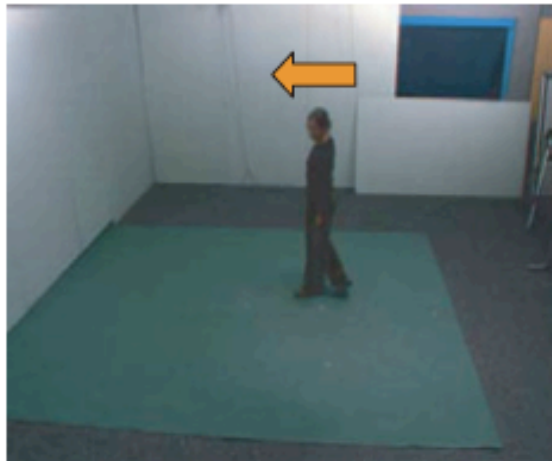
Works well, depending on task; not rejecting improves things
metric learning improves things

Youtube video



IXMAS and Aspect

Camera 0



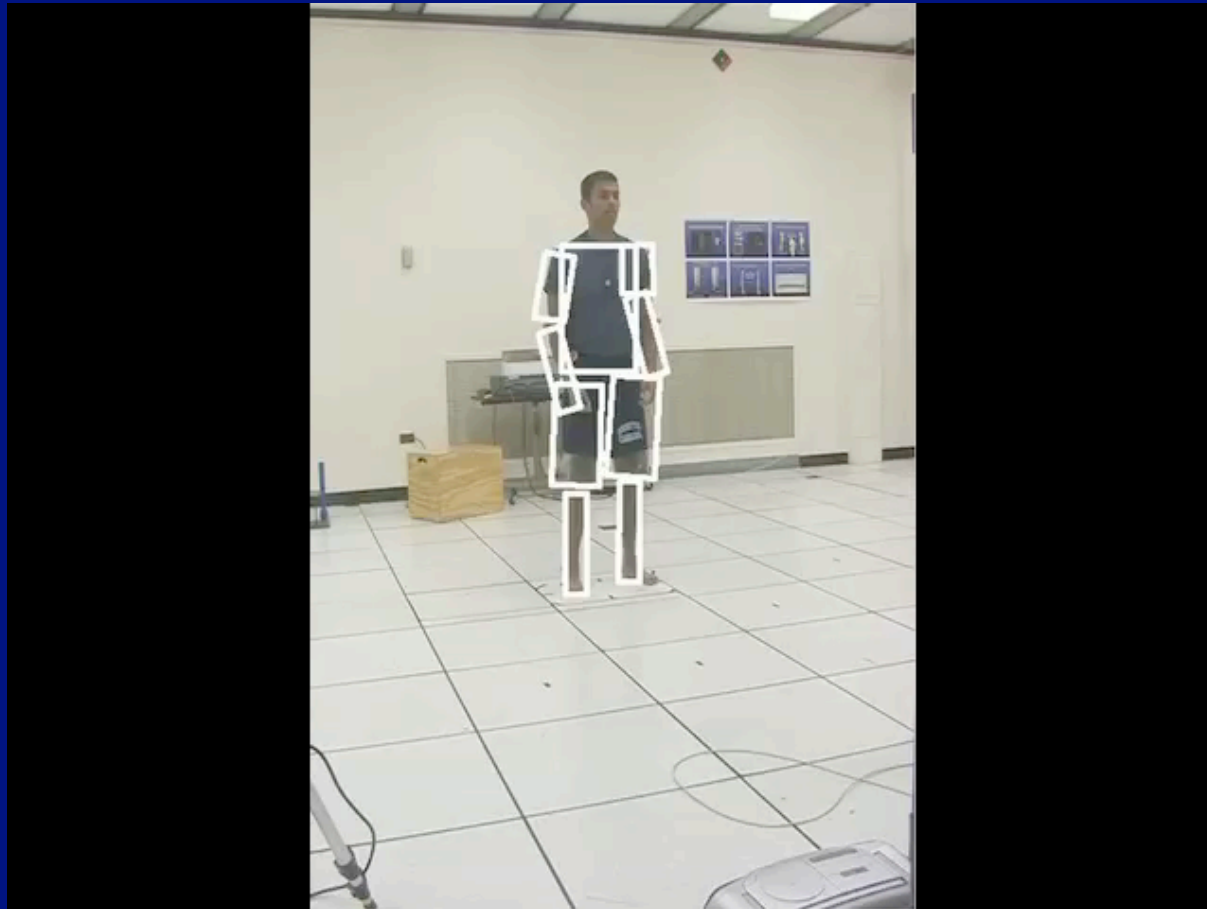
Camera 4



The Effects of Aspect

	Camera 0		Camera 1		Camera 2		Camera 3		Camera 4	
FO	76		76		68		73		51	
	WT		WT		WT		WT		WT	
Camera 0	NA		35		16		8		10	
Camera 1	38		NA		15		8		11	
Camera 2	16		16		NA		6		11	
Camera 3	8		8		8		NA		8	
Camera 4	12		11		15		9		NA	

Motion transduction



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, 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

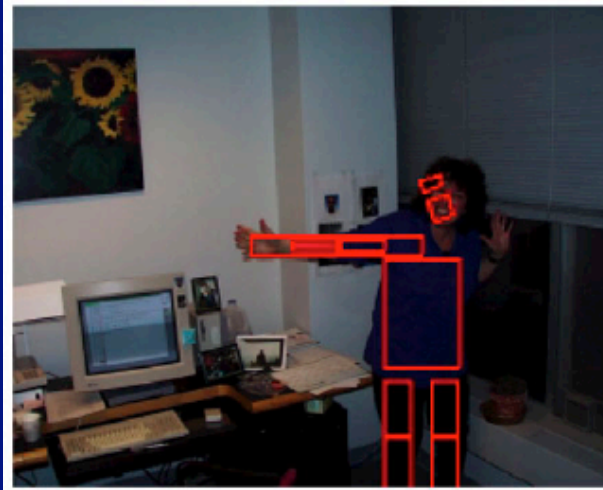
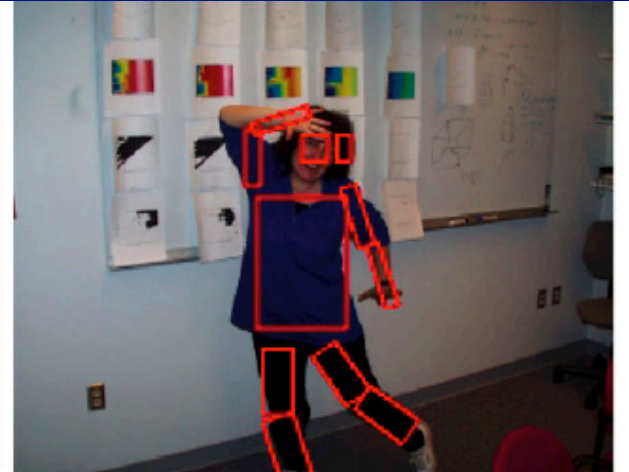
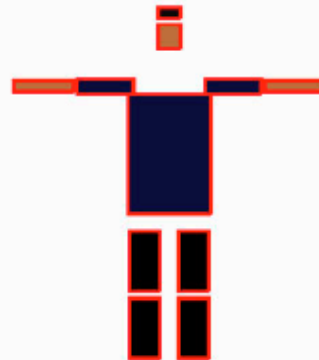
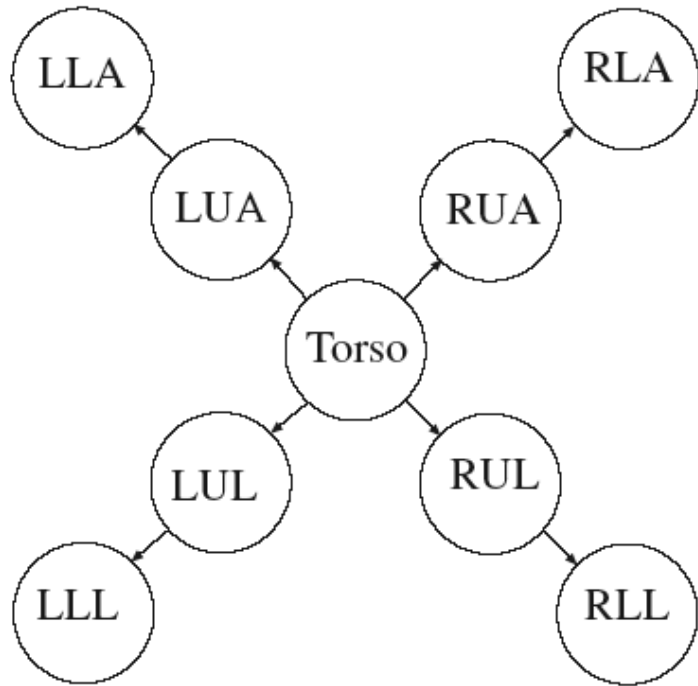
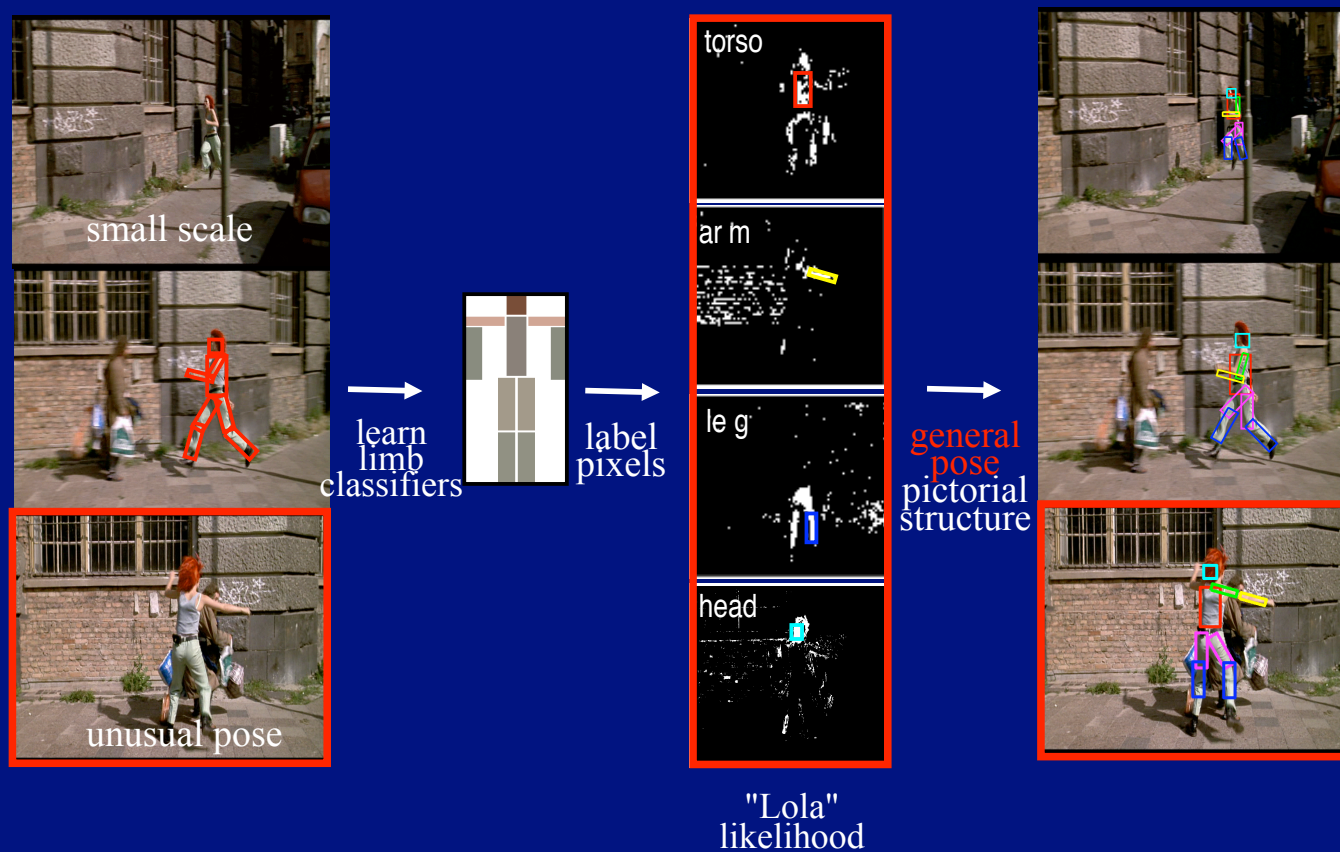


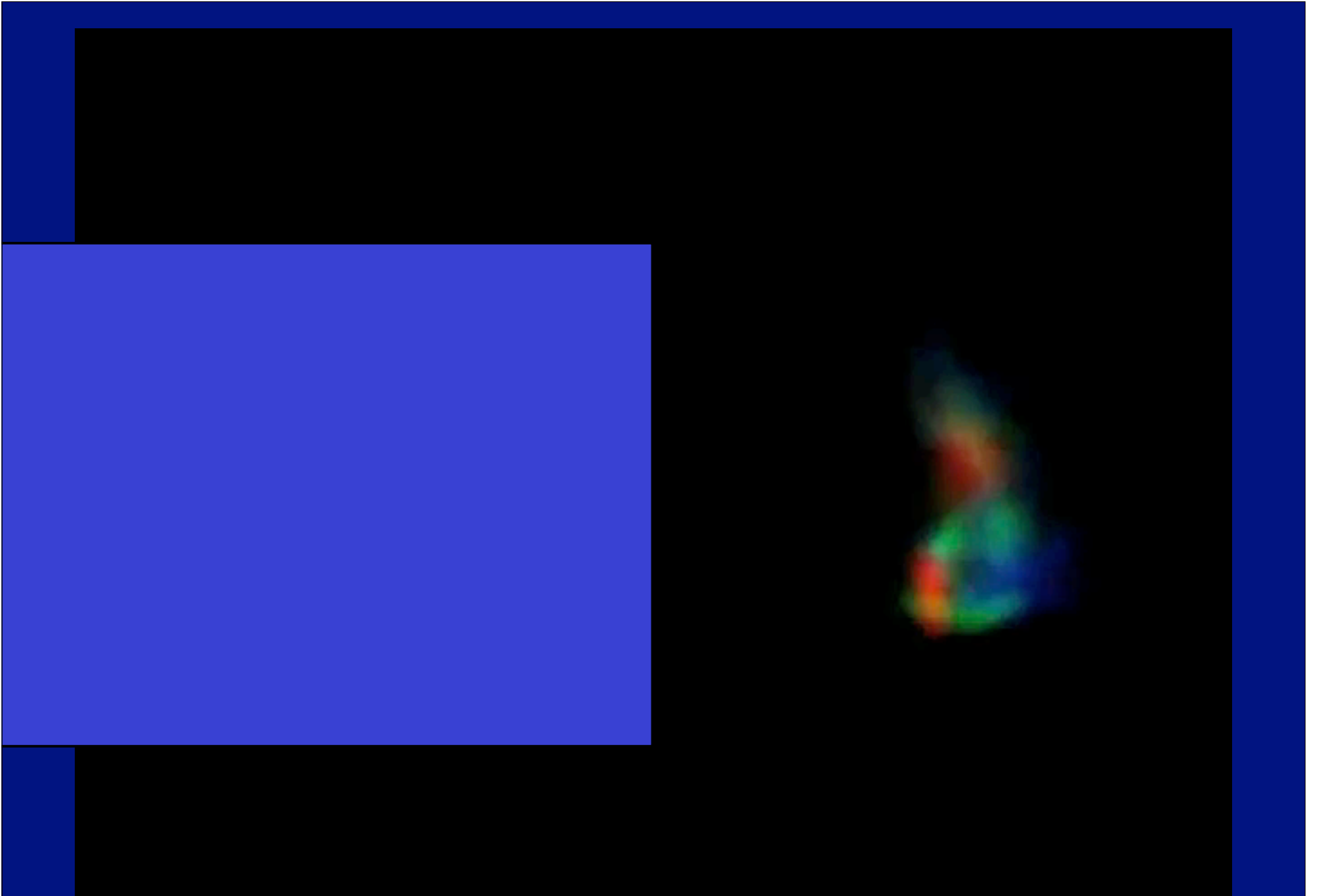
Figure from "Efficient Matching of Pictorial Structures,"
P. Felzenszwalb and D.P. Huttenlocher, Proc. Computer Vision and Pattern Recognition
2000, c 2000, IEEE as used in Forsyth+Ponce, pp 636, 640

Build and detect models





Ramanan, Forsyth and Zisserman CVPR05

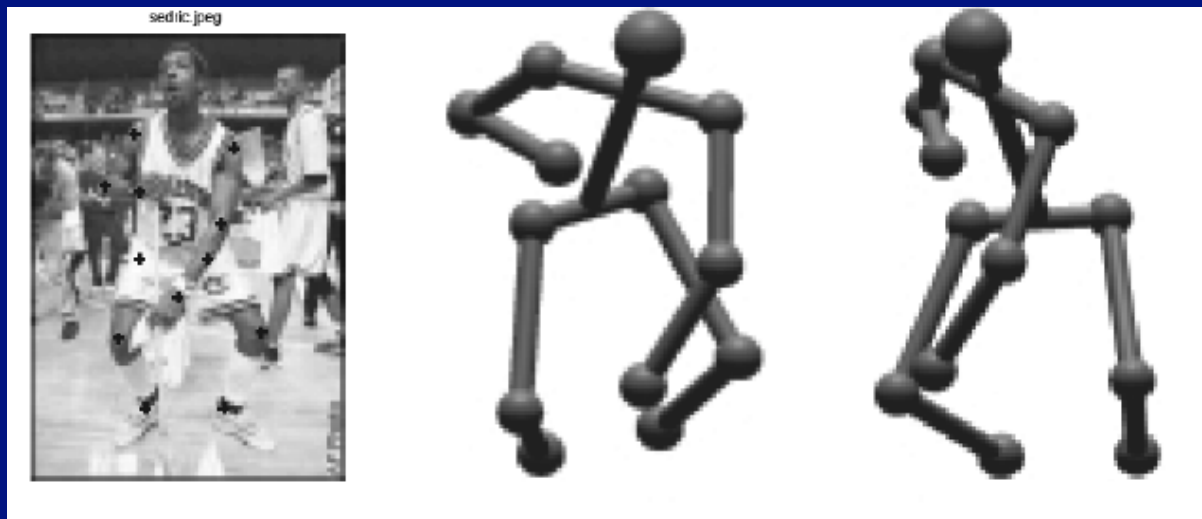




Ramanan, Forsyth and Zisserman CVPR05

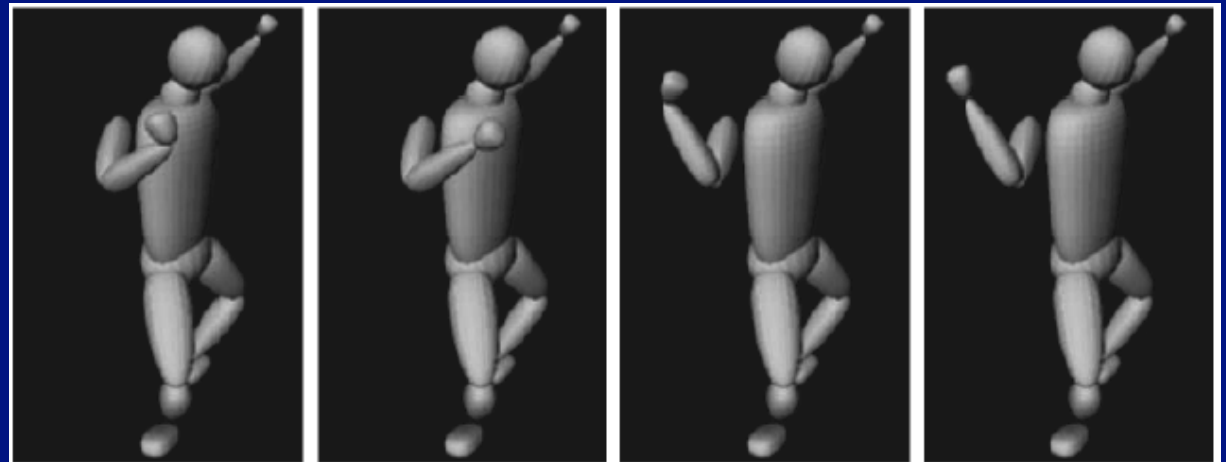
Lifting

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



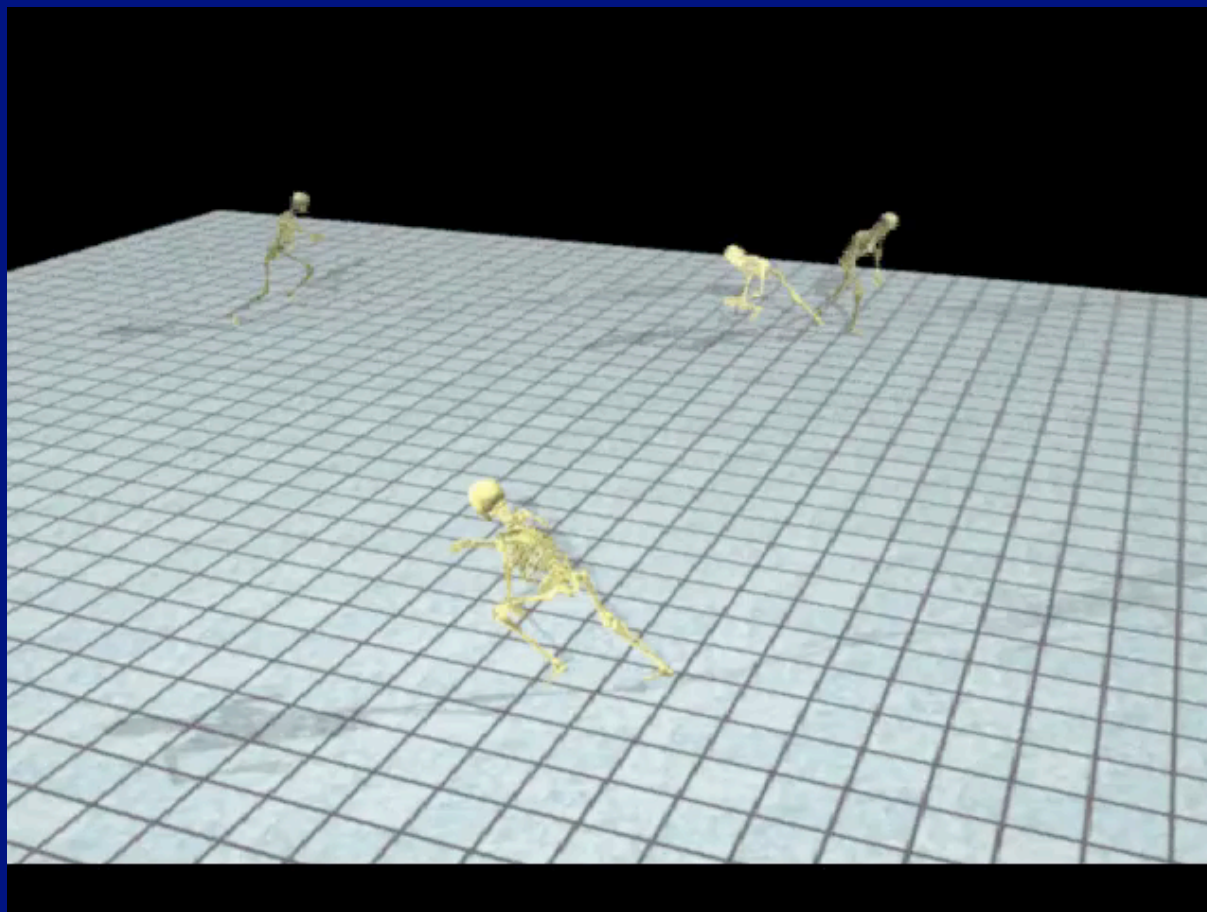
Ambiguity

- Troubled question
 - lifts are ambiguous (Orthography; Sminchiescu+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 et al 06



Sminchiescu+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

Cut and Paste works well over time

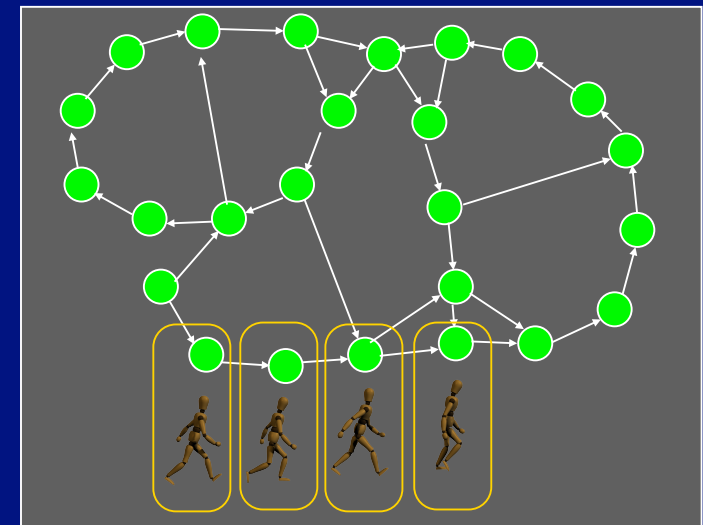
- Motion graph: by analogy with
 - text synthesis, texture synthesis, video textures
- Take measured frames of motion as nodes
 - from motion capture, given us by our friends
- Edge from frame to any that could succeed it
 - decide by dynamical similarity criterion
 - see also (Kovar et al 02; Lee et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

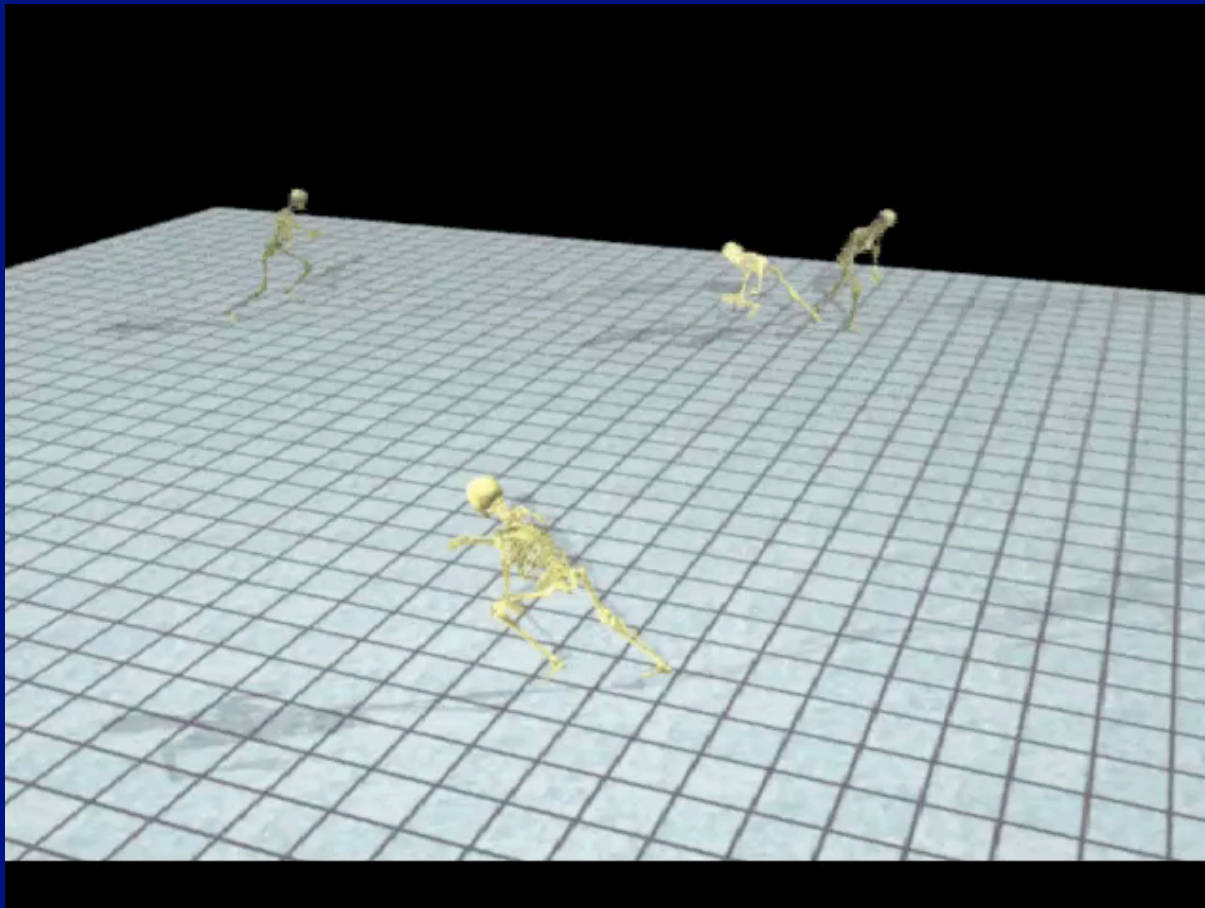
Motion Graph:

Nodes = Frames

Edges = Transition

A path = A motion

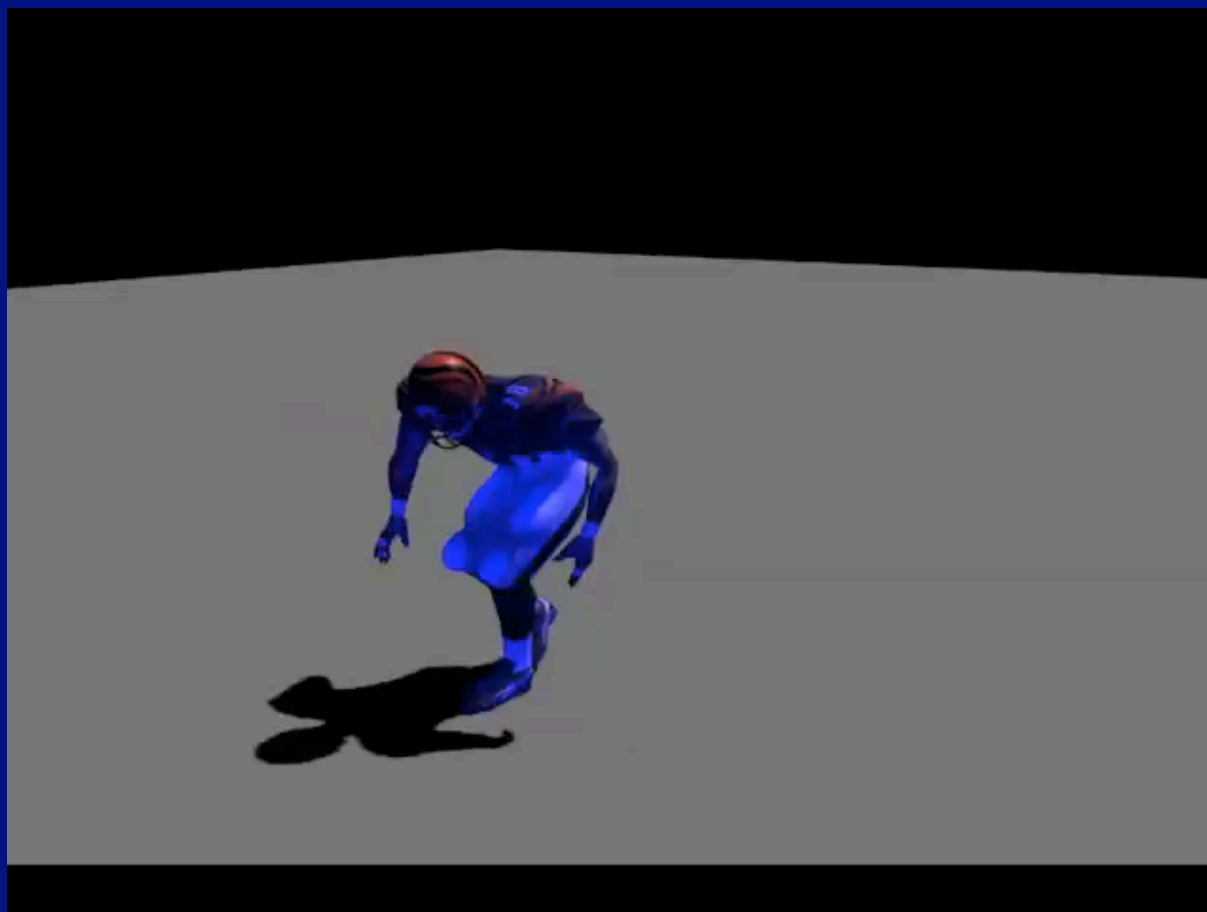




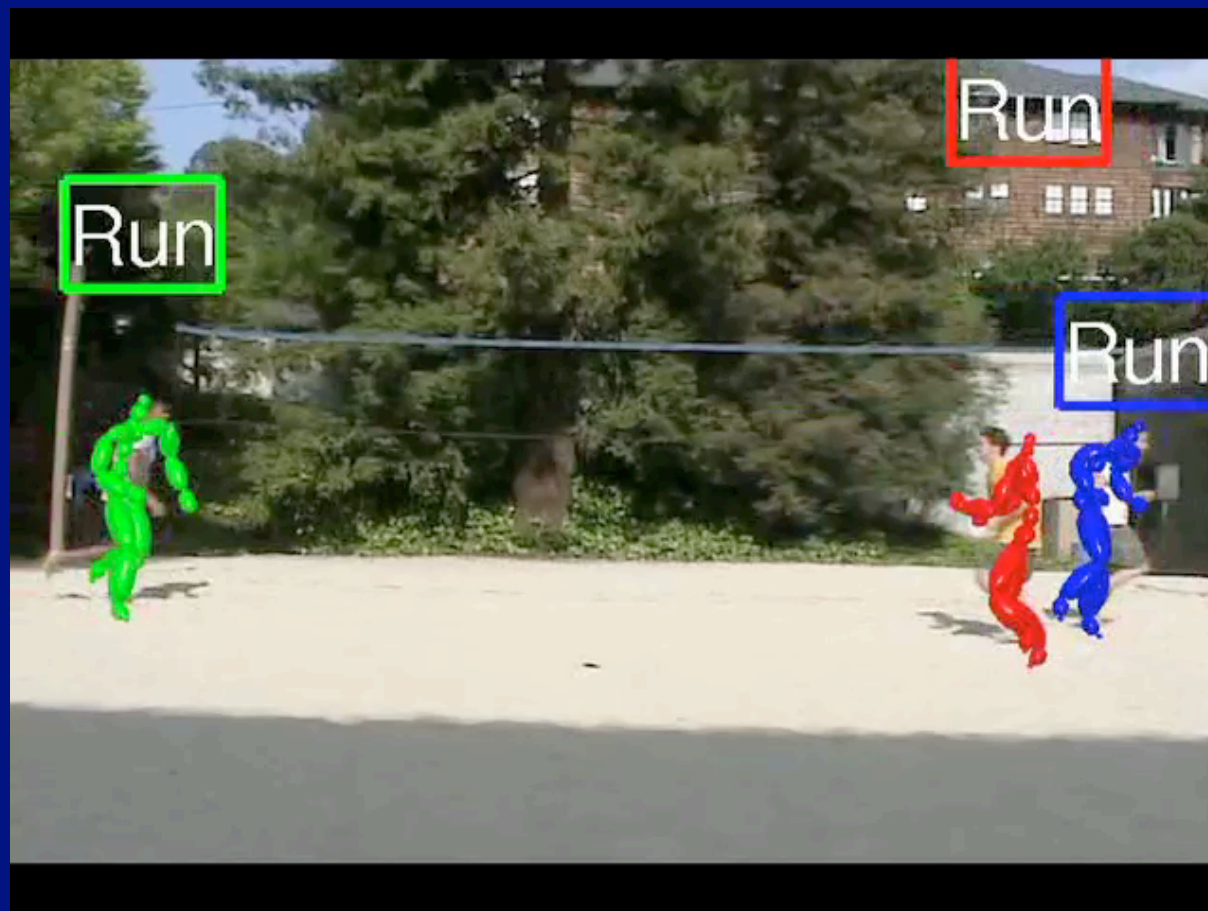
Transplantation

- 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





Activity from tracks



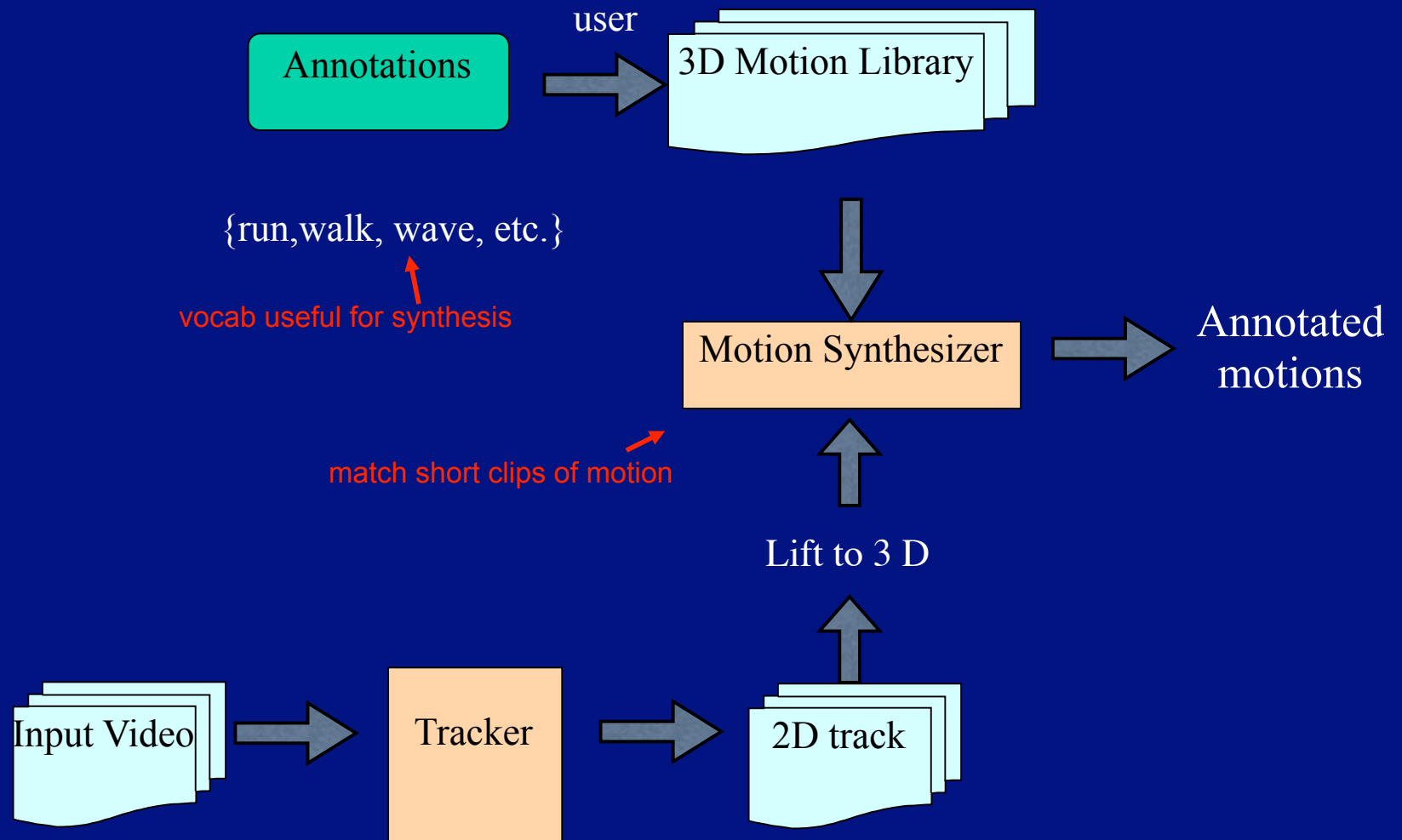
Themes

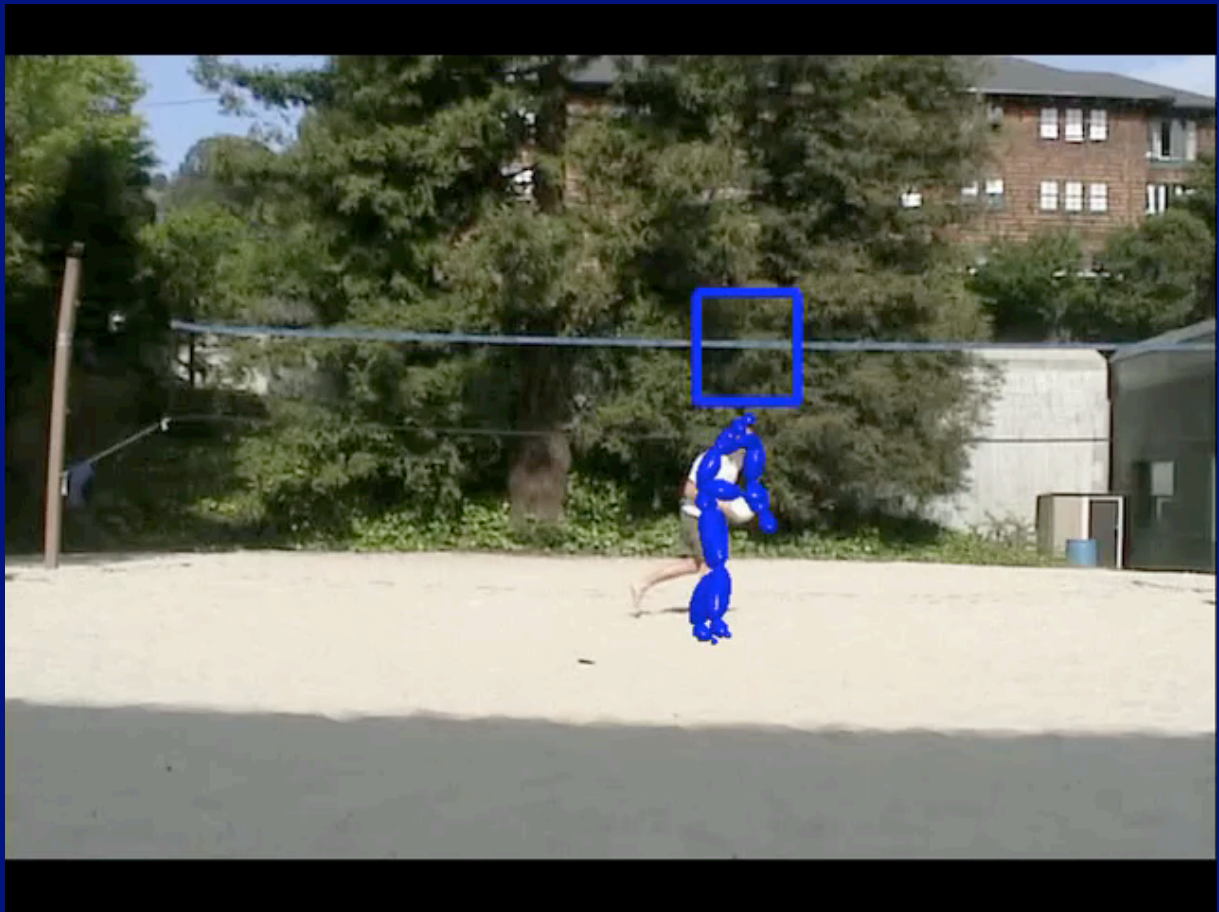
- Activity recognition has important special properties
 - No taxonomy - the structure of categories is hard, not well understood
 - Activity composes in complex ways
- Hence, most activities have no name
 - we're not really tagging videos "run" vs. "walk"
- Signal representations should
 - respect composition
 - be comparative

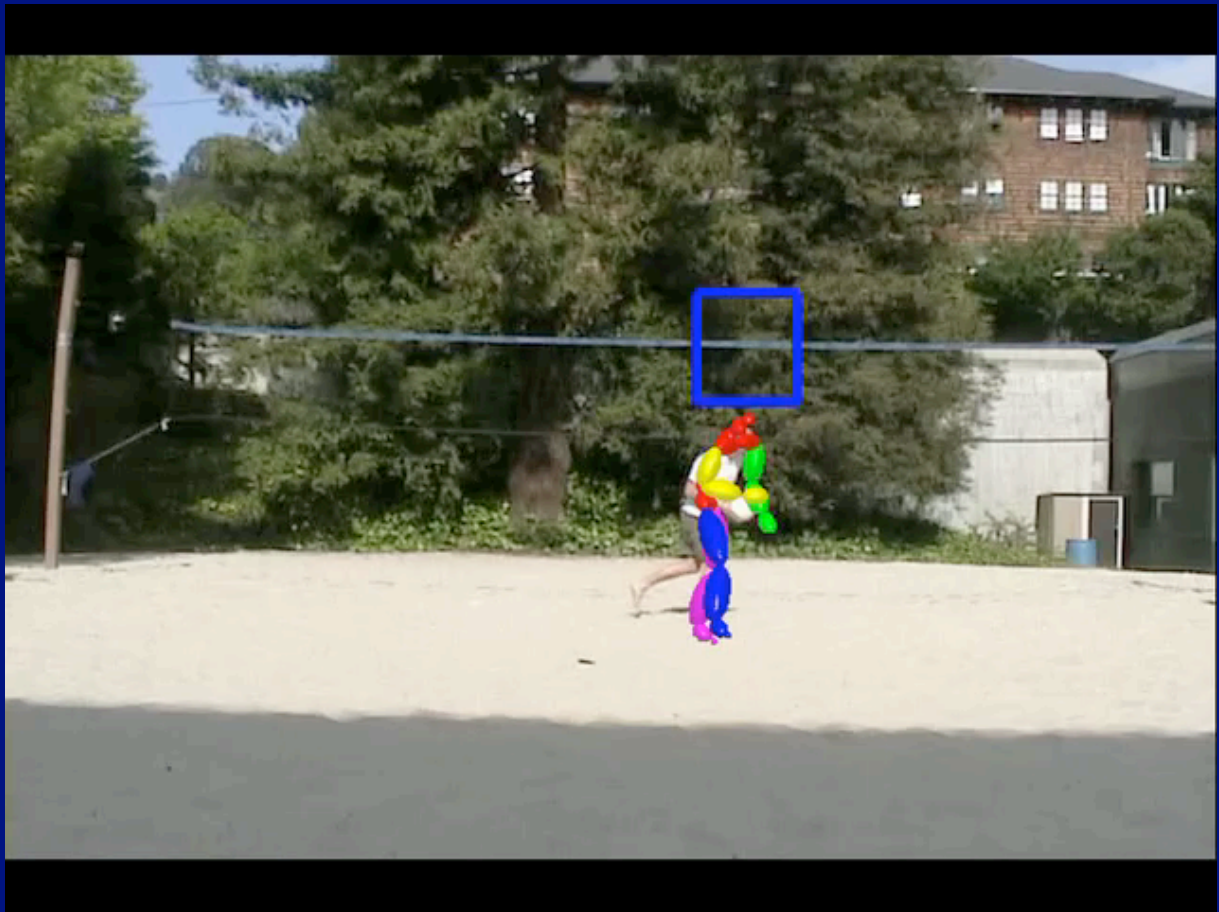
Composition and Activity

- 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)

Annotating observations by synthesis

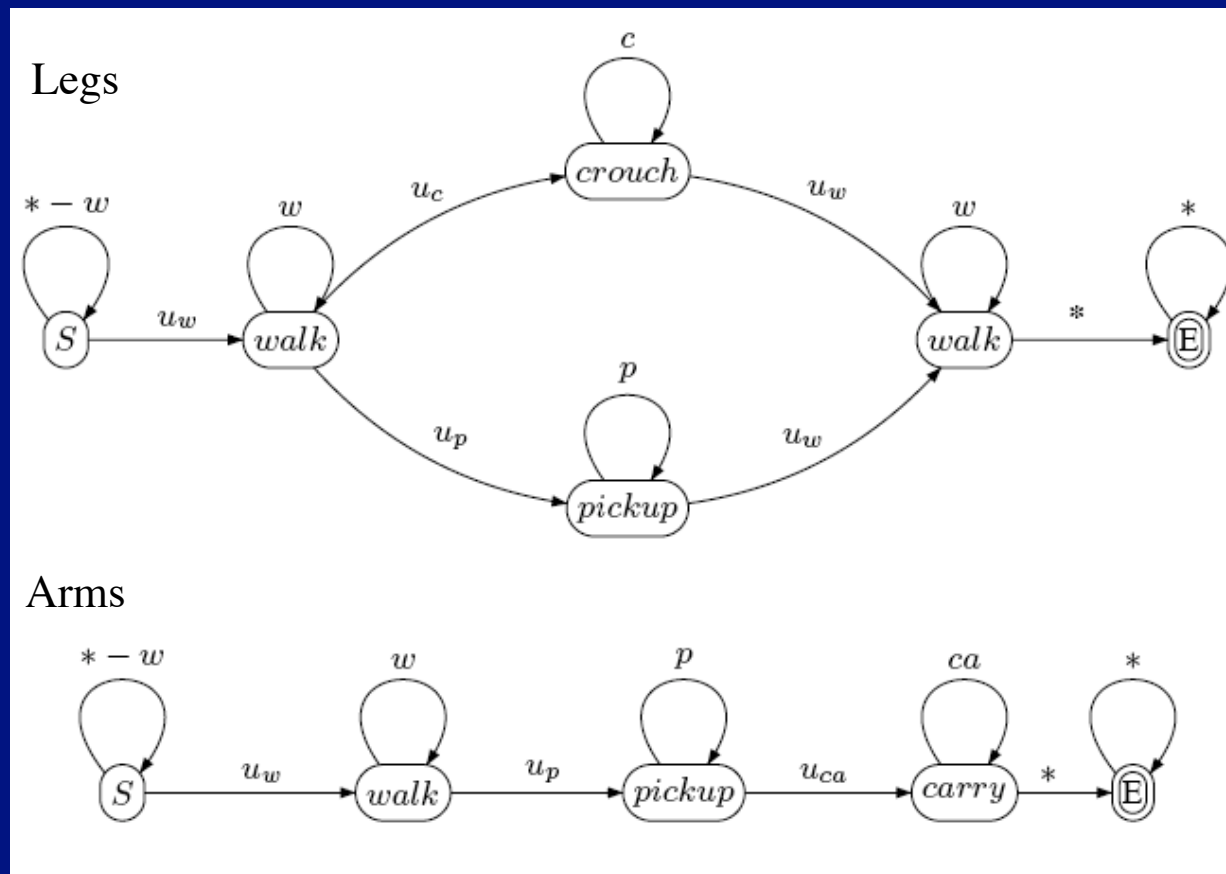








Composite activity names



Main problems

- What should the output be?
 - fixed vocabulary
 - easier, but people aren't like that
 - composite
 - how, in general
- What should the signal representation be?
 - fixed vocabulary, appearance wins hands down
 - but you'll make mistakes if someone does something funny
 - composite, tracks seem to be essential
 - but they're really noisy, so you'll make mistakes