Looking at People

D.A. Forsyth, UIUC (was U.C. Berkeley; was U.Iowa)
Leslie Ikemoto, James O'Brien, Ryan White, Anthony Lobay
all of UC Berkeley
Okan Arikan of UT Austin Deva Ramanan of TTI/UC Irvine
Ali Farhadi of UIUC Nazli Ikizler of Bilkent 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 is human motion important?

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

Themes

- Activity recognition has important special properties
  - No taxonomy - the structure of categories is hard, not well understood
  - Activity composes in complex ways
- Current signal representations are unsatisfactory
  - track and lift, work in 3D
    - good for aspect, composition
    - accuracy in localizing limbs is very difficult
  - spatio-temporal volumes
    - aspect is tough but manageable
    - composition across time easy, across the body mysterious.
    - attribute reasoning may be useful.

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)

Motion synthesis

- Problem
  - Produce a human motion that meets some constraints and looks good
- Methods
  - By animator
  - By combining observations
    - old tradition of move trees; also (Kovar et al 02, Lee et al 02, Arikan +Forsyth 02, Arikan et al 03,Gleicher et al 03)
  - By physical models, biomechanical models, statistical models (see review)
- Why do we care?
  - Exposes important practical properties of human motion.

Cut and Paste works well over time

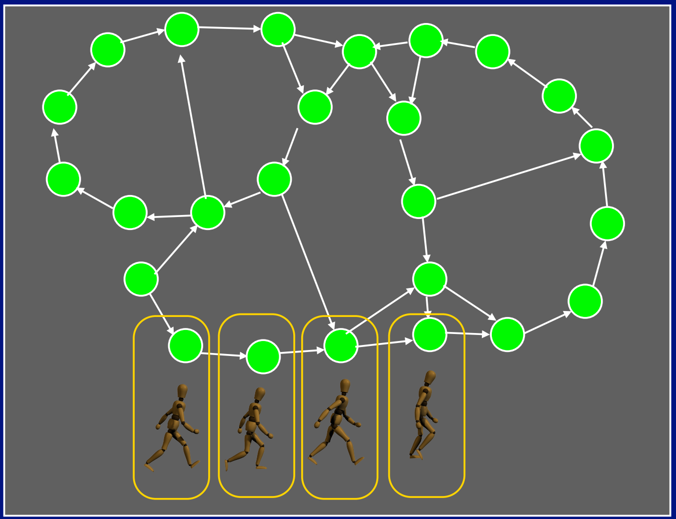
- Motion graph: by analogy with
  - text synthesis, texture synthesis, video textures
- Take measured frames of motion as nodes
  - from motion capture, given us by our friends
- Edge from frame to any that could succeed it
  - decide by dynamical similarity criterion
  - see also (Kovar et al 02; Lee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

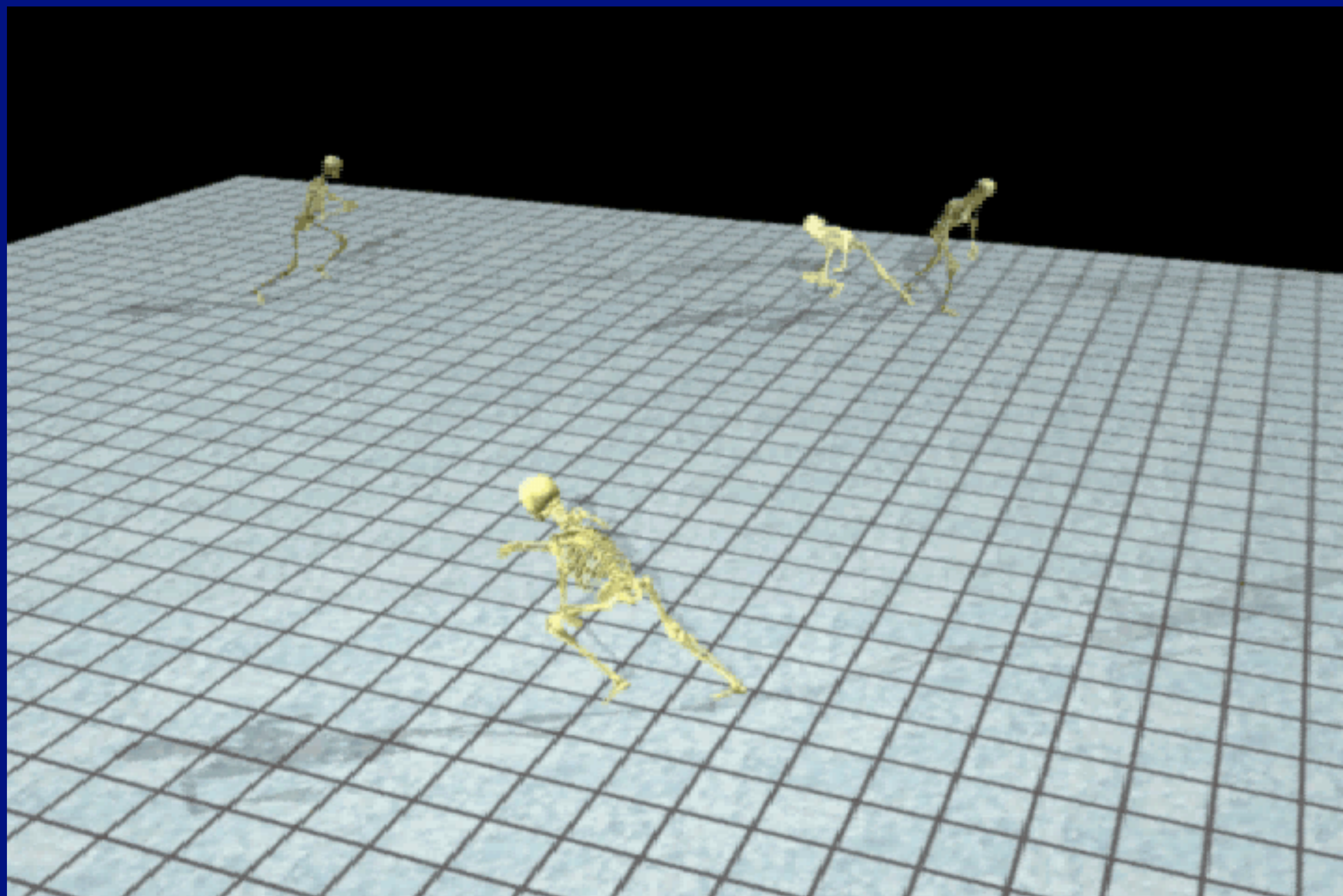
Motion Graph:

Nodes = Frames

Edges = Transition

A path = A motion





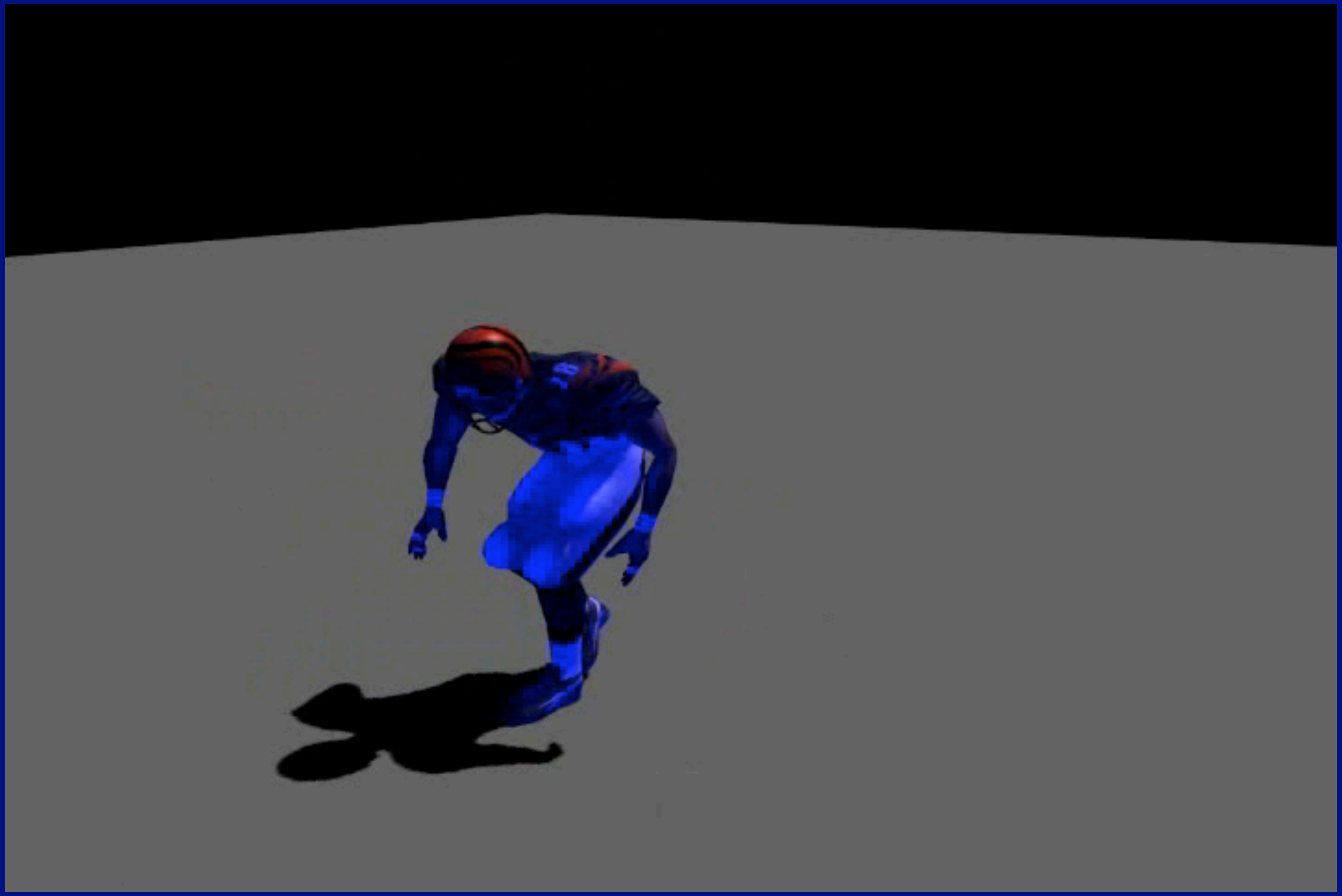
Arikan+Forsyth 02

Transplantation

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

Ikemoto+Forsyth 04

Ikemoto+Forsyth 04



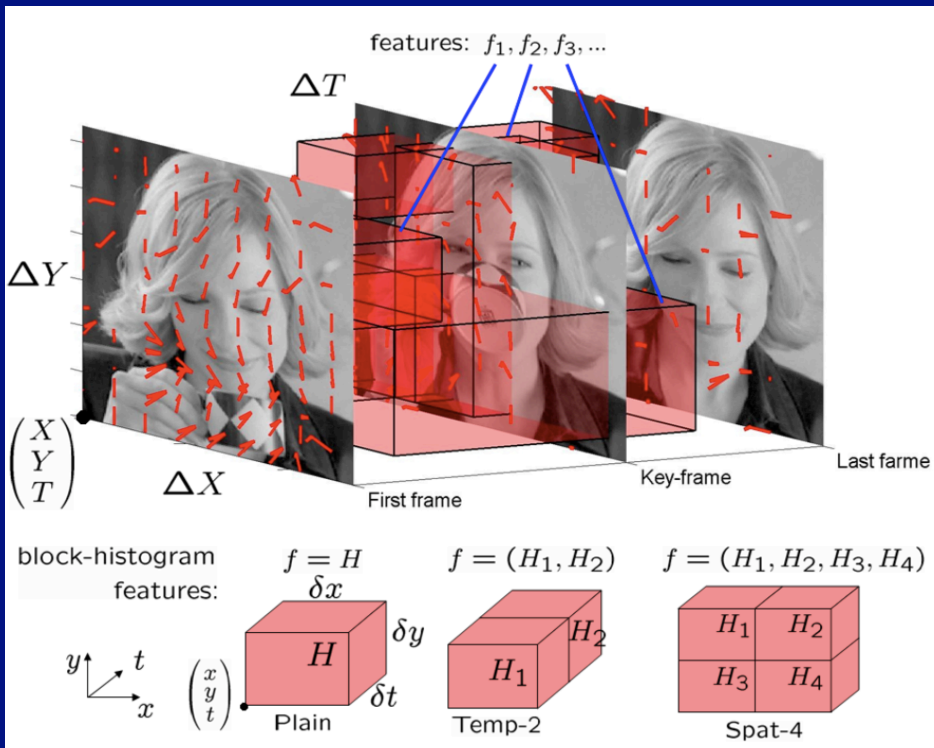
Ikemoto+Forsyth 04

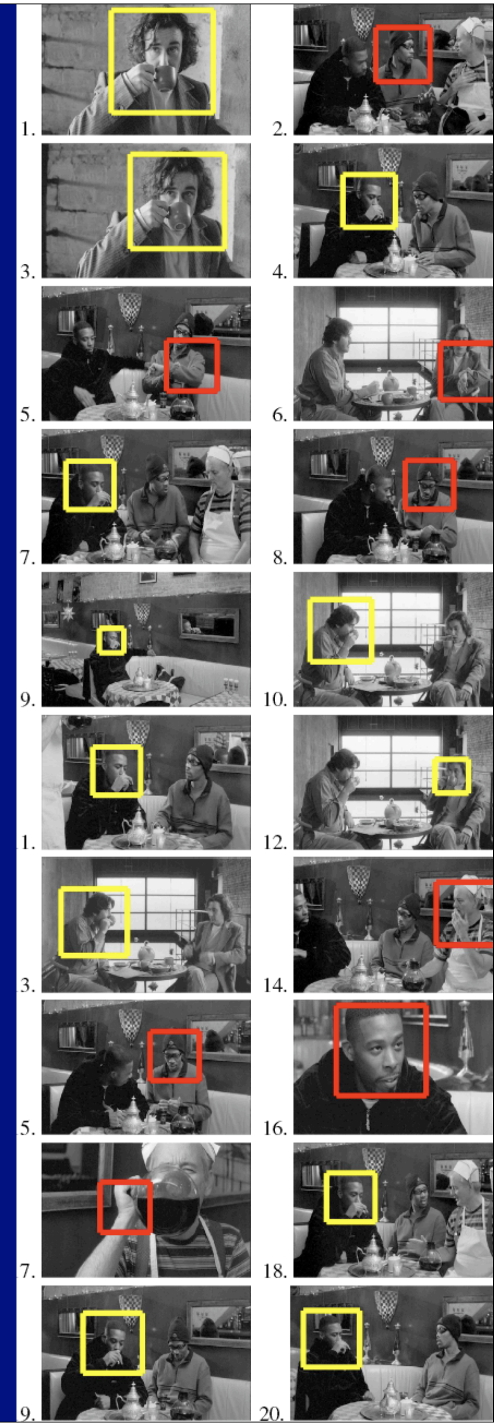
What are people doing?

- Core problem
  - It is not known what needs to be known
    - or, what should we measure?

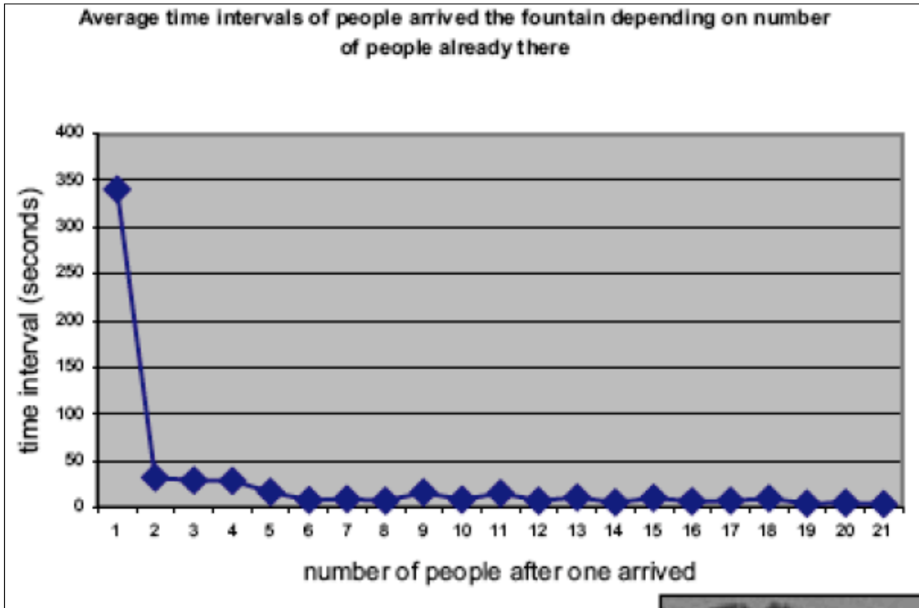
What is the right signal representation?

- Spatio-temporal features
  - Laptev Perez 07
- 3D Kinematic track
  - with some work, a 3D representation of arms, legs, torso, etc.
- Appearance
  - Spatio-temporal features localized to the body



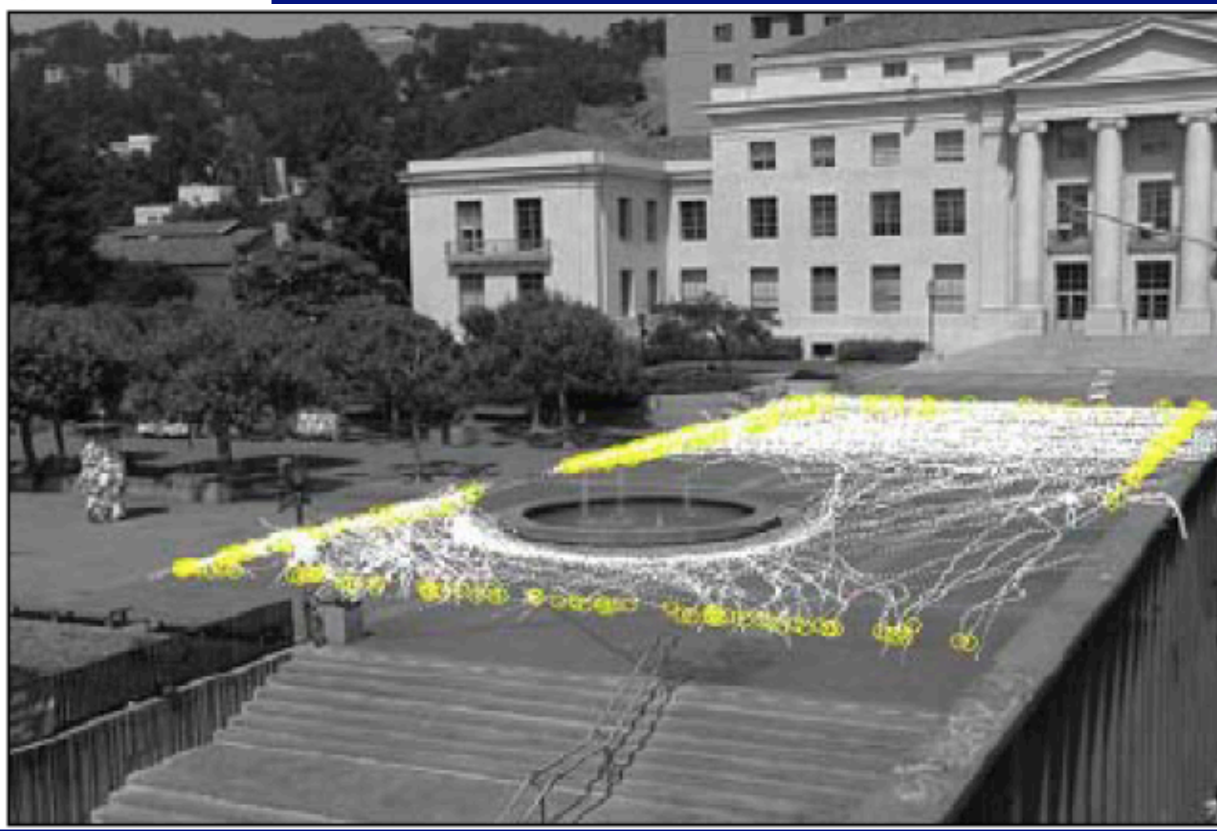


Laptev Perez 2007
see also Laptev et al 08



Point tracks reveal curious phenomena in public spaces

Yan+Forsyth, 04



Tracking

- Hard, but
  - you can do it
  - great advantages for aspect, composition
- Major problems with accuracy, seem likely to be ongoing
  - but ferrari zisserman, etc.

Why is kinematic tracking hard?

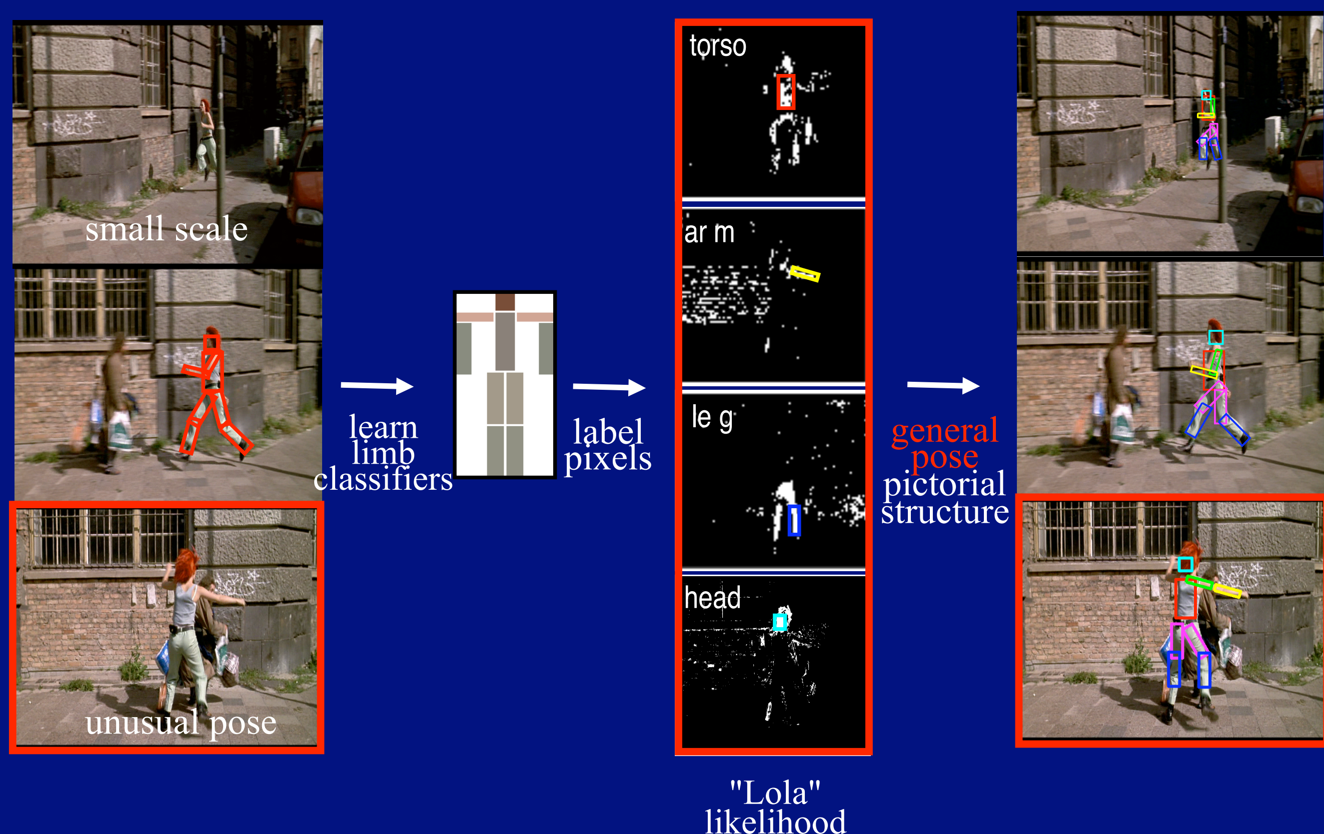
- It's hard to detect people
  - until recently, human trackers were manually started
- People move fast, and can move unpredictably
  - dynamics gives limited constraint on future configuration
  - appearance changes over time (shading, aspect, etc)
- Some body parts are small and tend to have poor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


variation in pose & aspect


self-occlusion & clutter

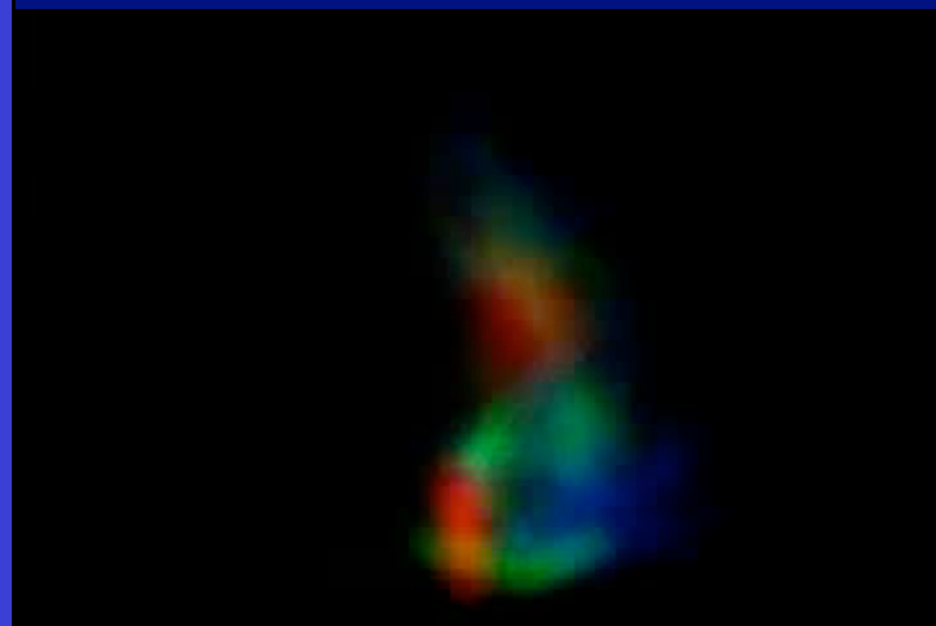

variation in appearance

Build and detect models



Ramanan, Forsyth and Zisserman CVPR05

Ramanan, Forsyth and Zisserman CVPR05



Ramanan, Forsyth and Zisserman CVPR05

Ramanan, Forsyth and Zisserman CVPR05

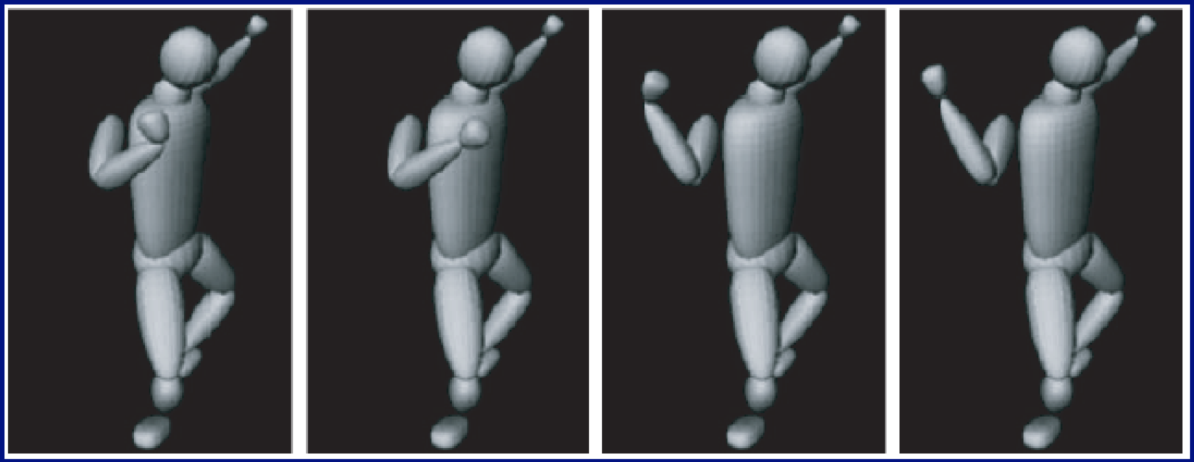
Lifting

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

Taylor, 00

Ambiguity

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



Sminchisescu+Triggs, 03

Naming activities

- With what? (no canonical vocabulary)
  - Choose actions with names
    - (e.g. gymnastics Bobick+Davis 01, ballet Efros et al 03)
  - Match motion to motion, avoid the issue (e.g. Efros 03)
  - Vocabulary of tags (eg Ramanan+Forsyth 03)
- Never enough data
  - “Noise” in transduction
    - aspect, appearance
    - tracking, lifting, silhouettes
    - intraclass variation in activity
  - Complex taxonomy
    - composition

Fiercely hard to learn models from video

- Generative dynamical models
  - dynamical parameters hard to learn
    - too many parameters
    - or insufficiently expressive

- Discriminative models
  - not enough training data
    - of the right aspect, clothing, etc.

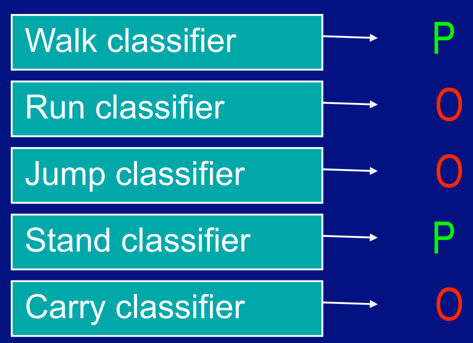
Label motion capture data

- Data
  - released to the research community by Electronic Arts, 2002
  - Or one could use Georgia Tech data, etc.

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

Annotation

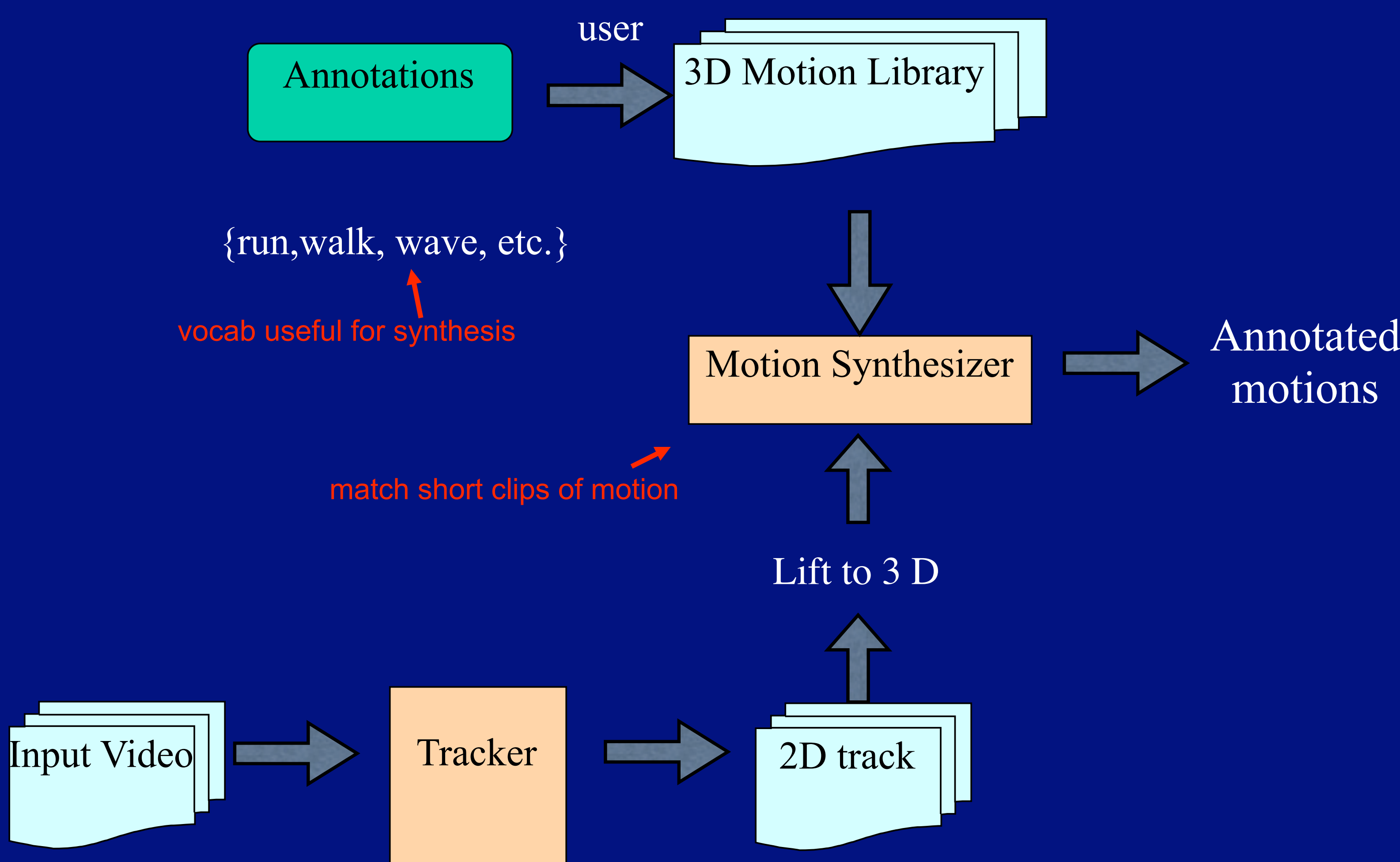
- Can do this with one classifier per vocabulary item
  - use an SVM applied to joint angles
  - form of on-line learning with human in the loop
  - works startlingly well (in practice 13 bits)

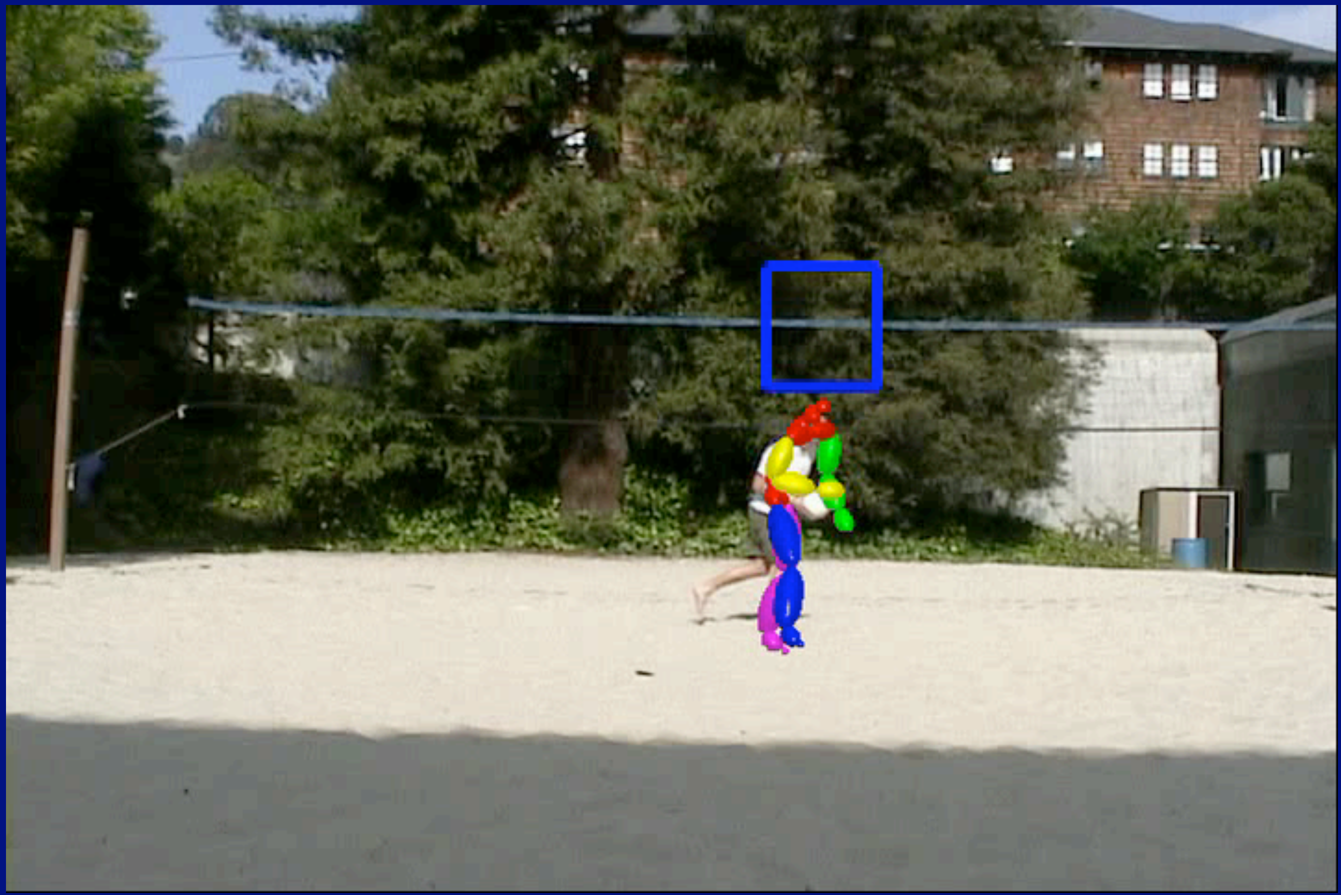


Arikan+Forsyth+O'Brien 03

Arikan+Forsyth+O'Brien 03

Annotating observations by synthesis

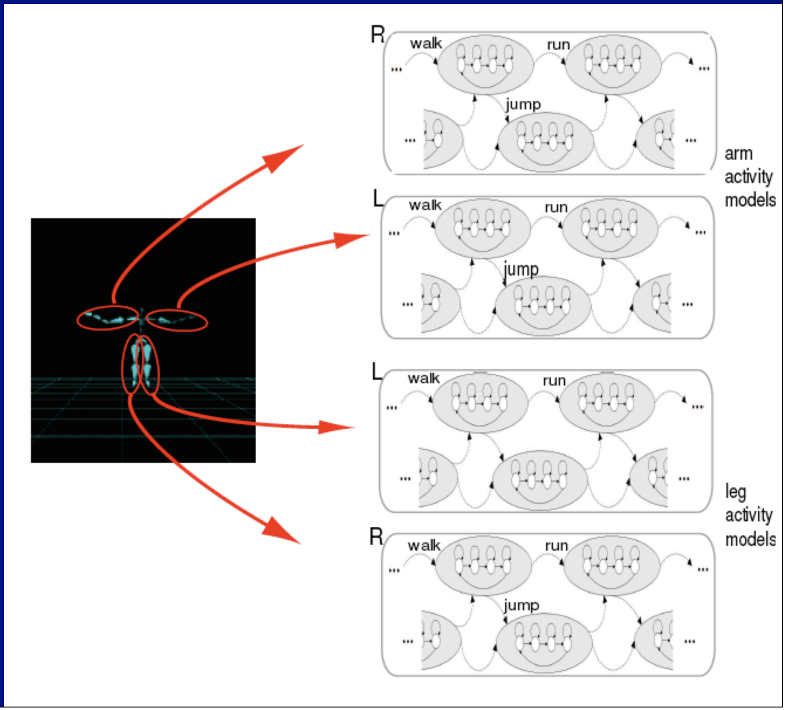




Ramanan Forsyth 04

Composition, authoring and transfer

- Activity composes across time and across the body
  - and we may have no examples of a particular activity
  - but we should like to query
    - generative model learned on annotated motion capture data
    - string together short timescale models
      - across time
      - across the body
    - Author longer timescale models
      - by kinematic consistency
      - by query



Generative model

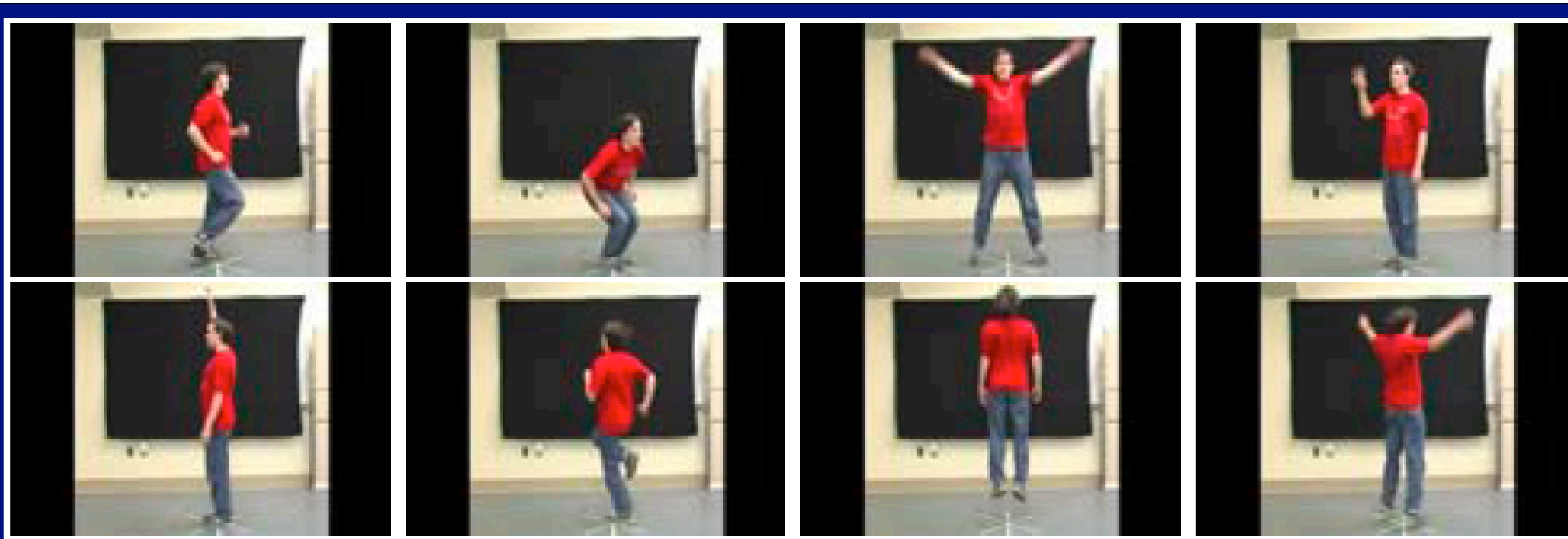
- Many states
  - but few parameters to learn
- Annotation vocabulary
  - original 13 annotations
    - Less: 3 direction labels, 1 ambiguous term
    - each limb can have at most one annotation

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

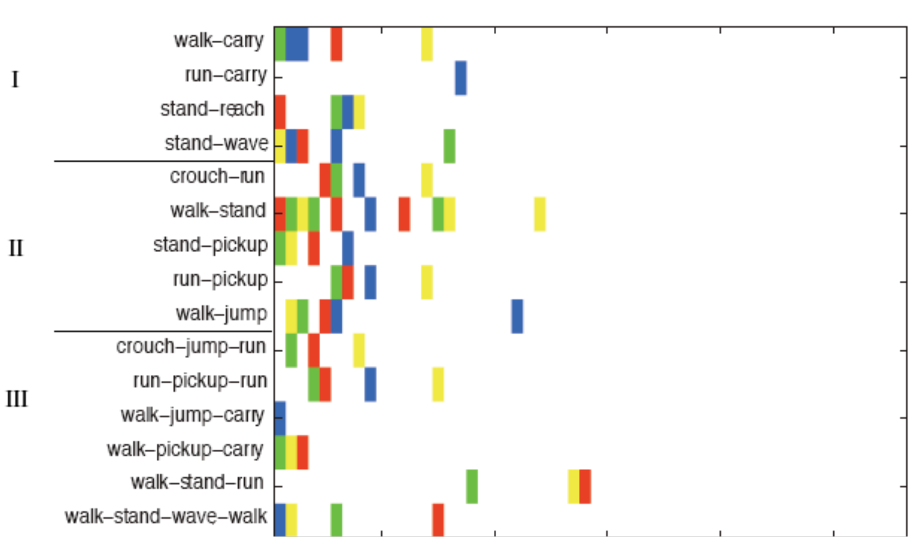
- Primary attraction
  - “natural” query language
- Rank sequences by
  - e.g. P(leg-walk-arm-walk-then-leg-walk-arm-reachl data, model)



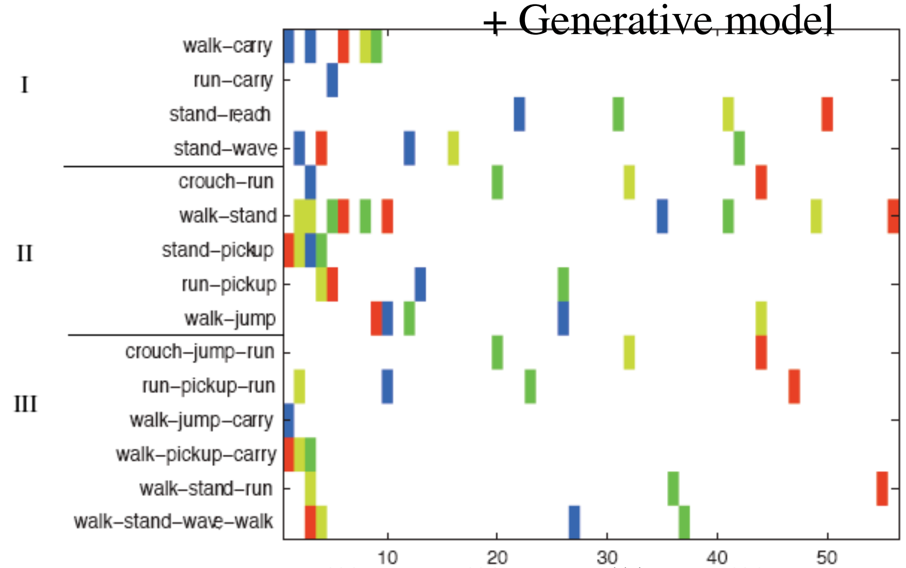
Ikizler Forsyth 07,08

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		

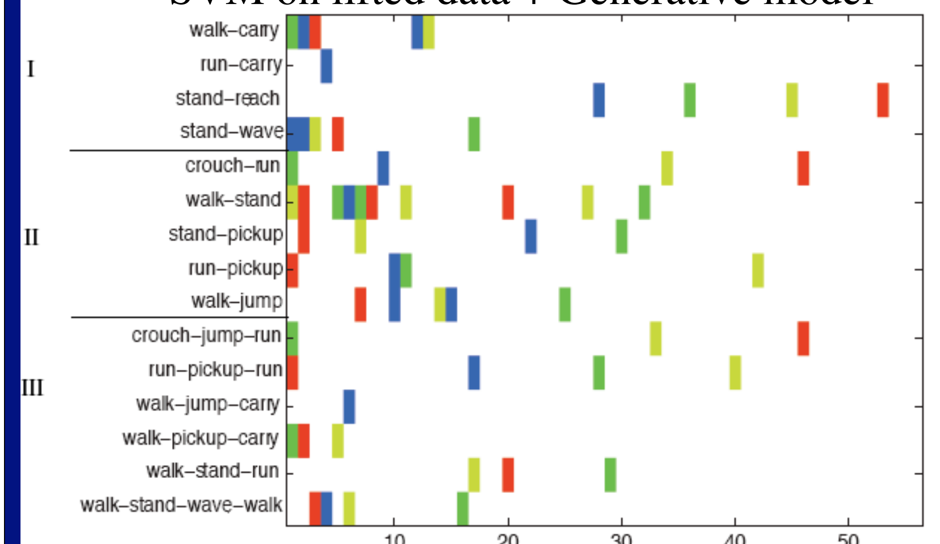
Our method

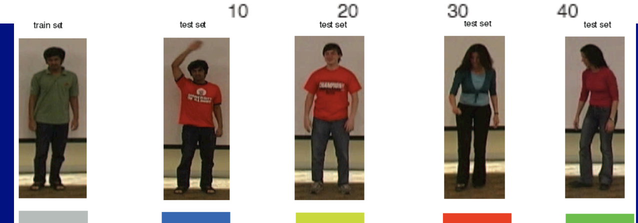


SVM on image appearance + Generative model



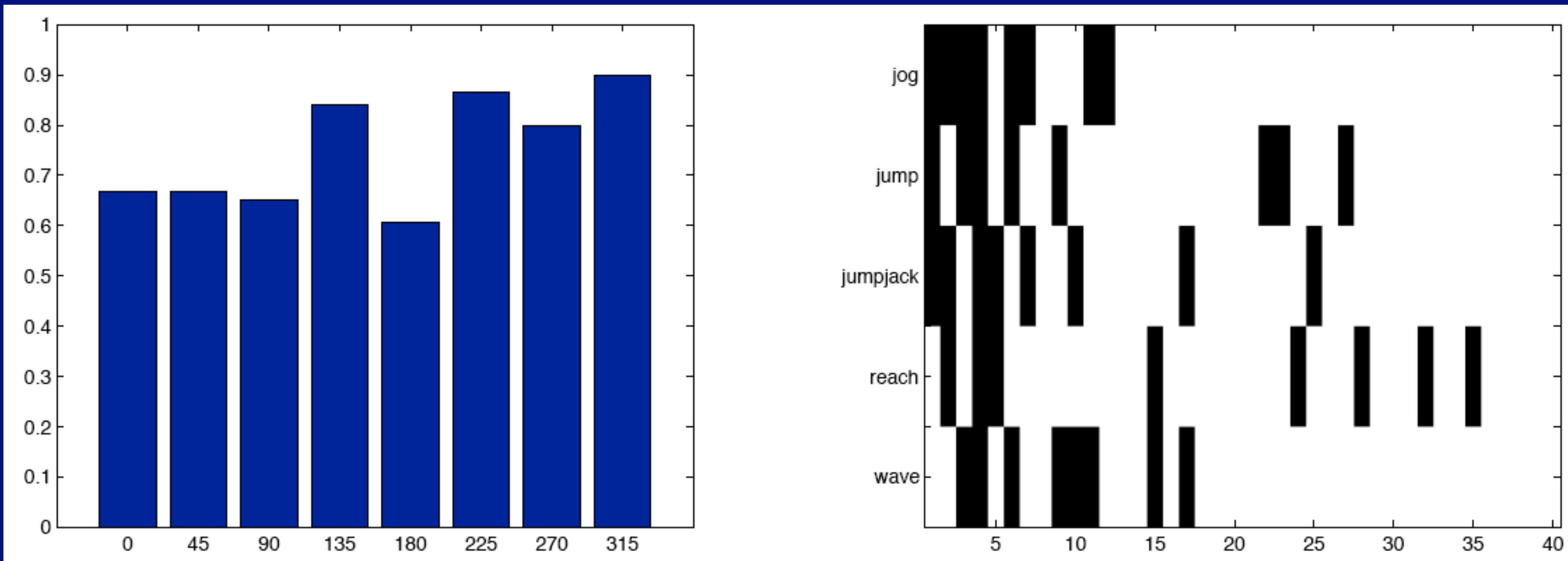
SVM on lifted data + Generative model





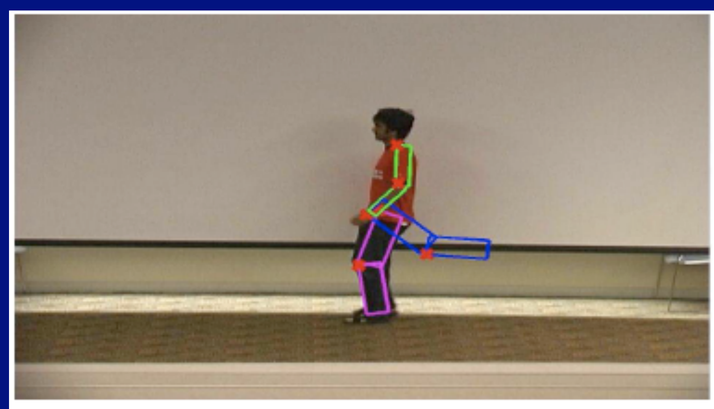
Ikizler Forsyth 07,08

The effect of aspect



Jog; Jump; Jumpjack; Reach; Wave

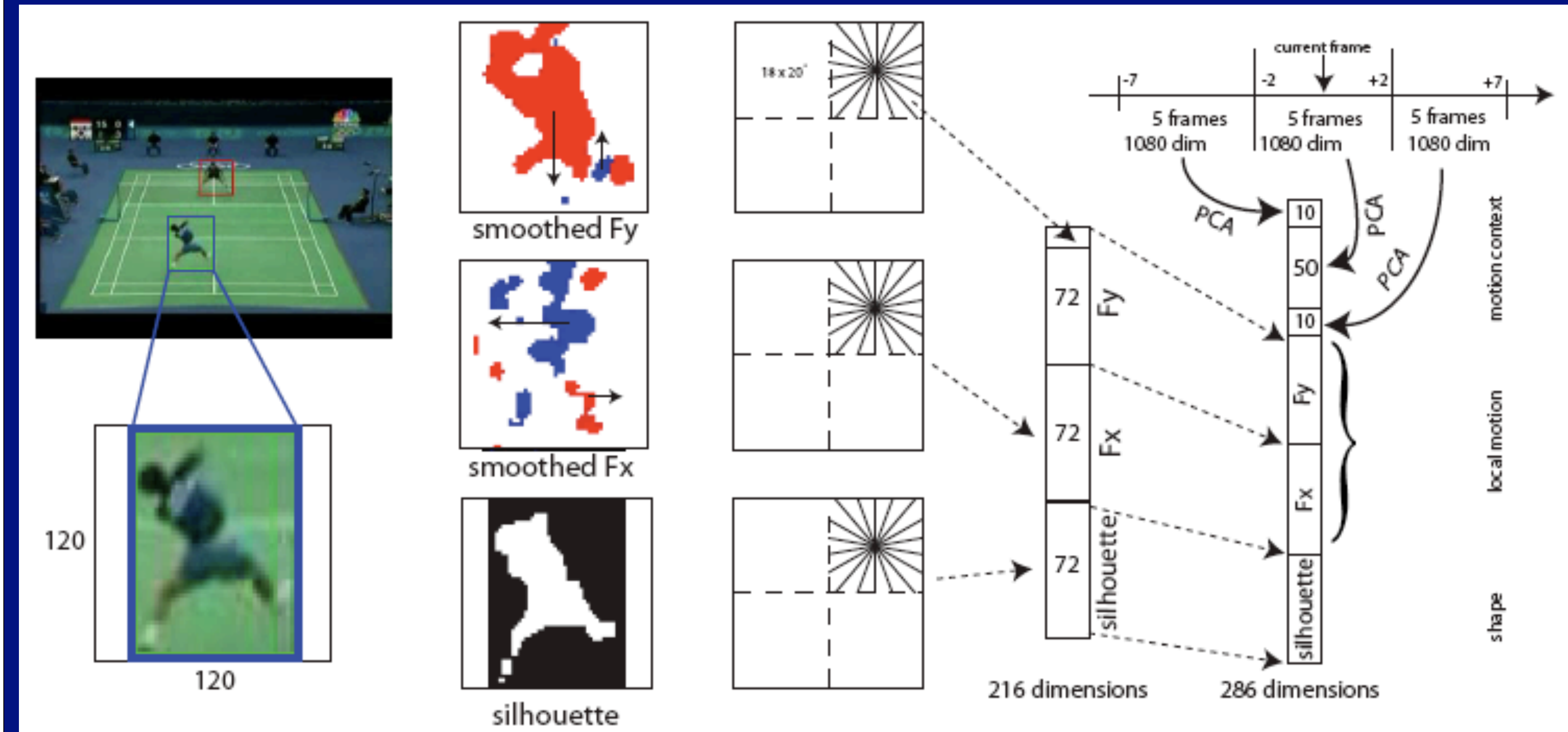
Ikizler Forsyth 07, 08



Appearance and activity

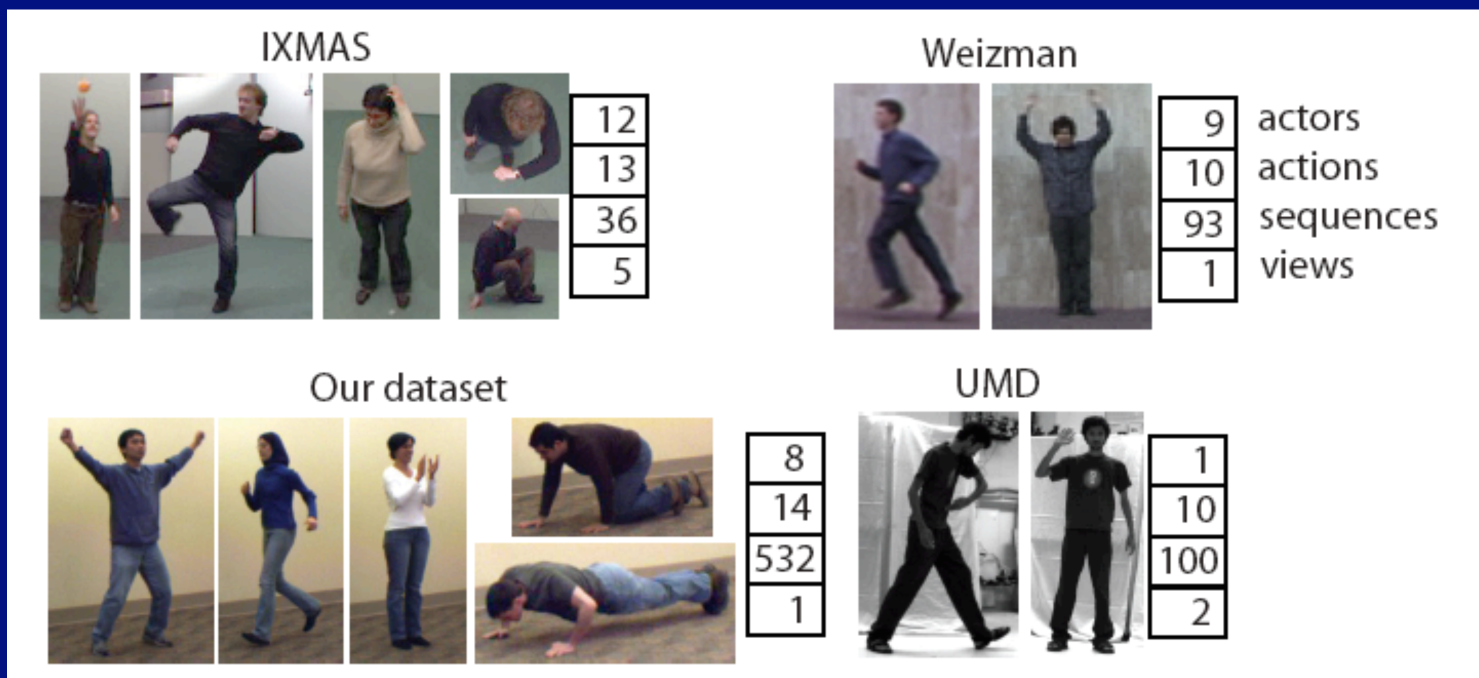
- Location can be a powerful guide to activity
  - Intille et al 95, 97
- Configuration, motion are distinctive
  - Polana Nelson 93; Niyogi Adelson 94; Bobick+Davis 97; Efros et al 03; Blank et al 05
  - spatiotemporal volumes are good (Blank et al 05)

An Appearance feature



Tran and Sorokin 08, after Duygulu and Ikizler 07

Datasets


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

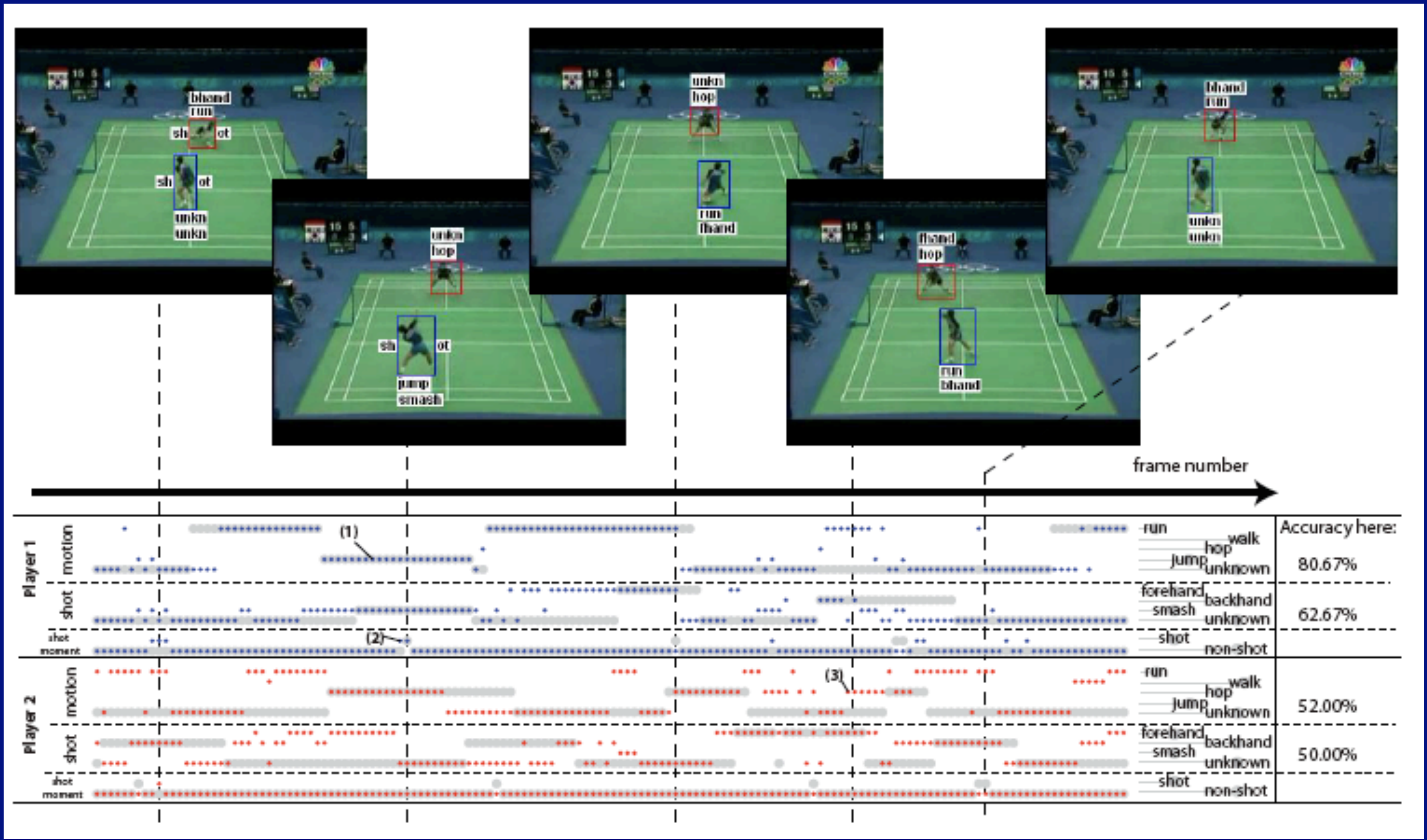
Discriminative results

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

Youtube video



Tran + Sorokin 08

IXMAS and Aspect

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	

Learning to recognize from the wrong view

- Idea:
  - Build features that are robust to aspect changes
  - AND
    - encode aspect explicitly in discriminative procedures

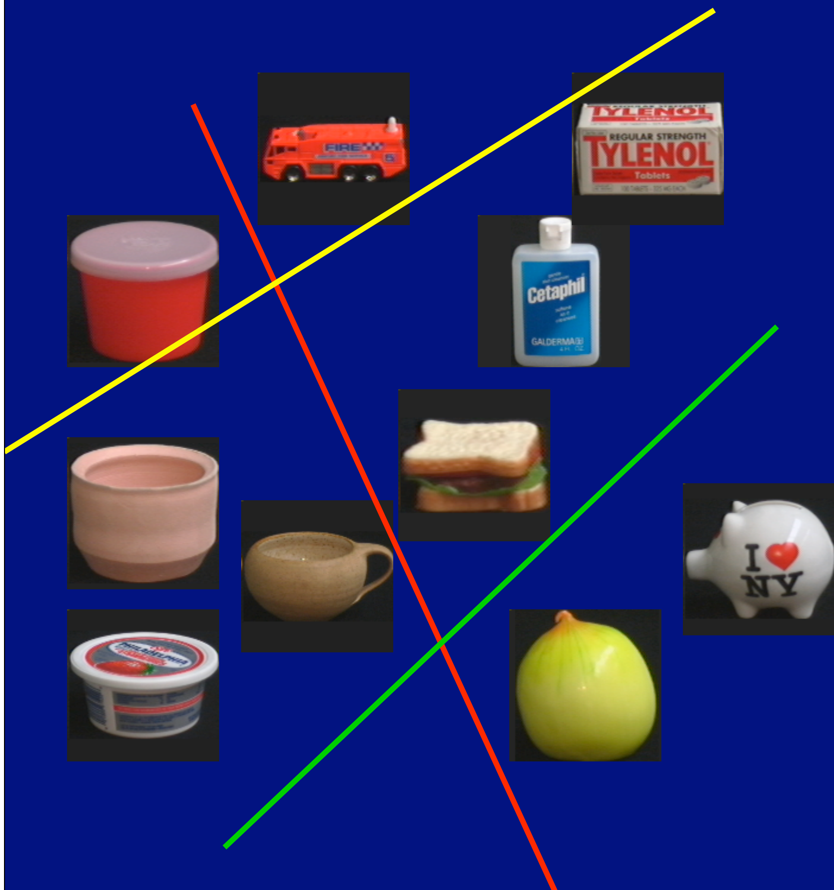
Farhadi Kamali 08

Comparative Features

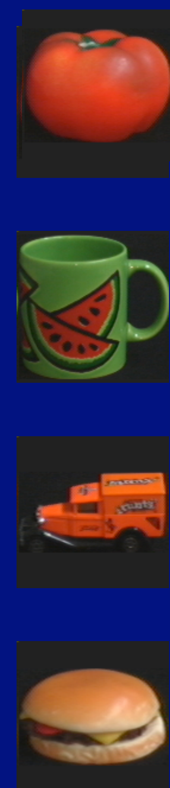
- Comparisons seem to behave well under change of aspect

[Images from COIL-100 Dataset]

Best splits & comparative features



Known objects



Unknown objects

0	0	1
1	1	0
0	1	1
1	0	0

Comparative features

[Images from COIL-100 Dataset]

Learning to recognize from the wrong view

- Idea:
  - tag training examples with an aspect variable
  - this is unknown, but we have an estimate
  - estimate classifier, correct aspect variable, at the same time
  - Recognition:
    - use non-parametric estimate of aspect var

Farhadi Kamali 08

Results

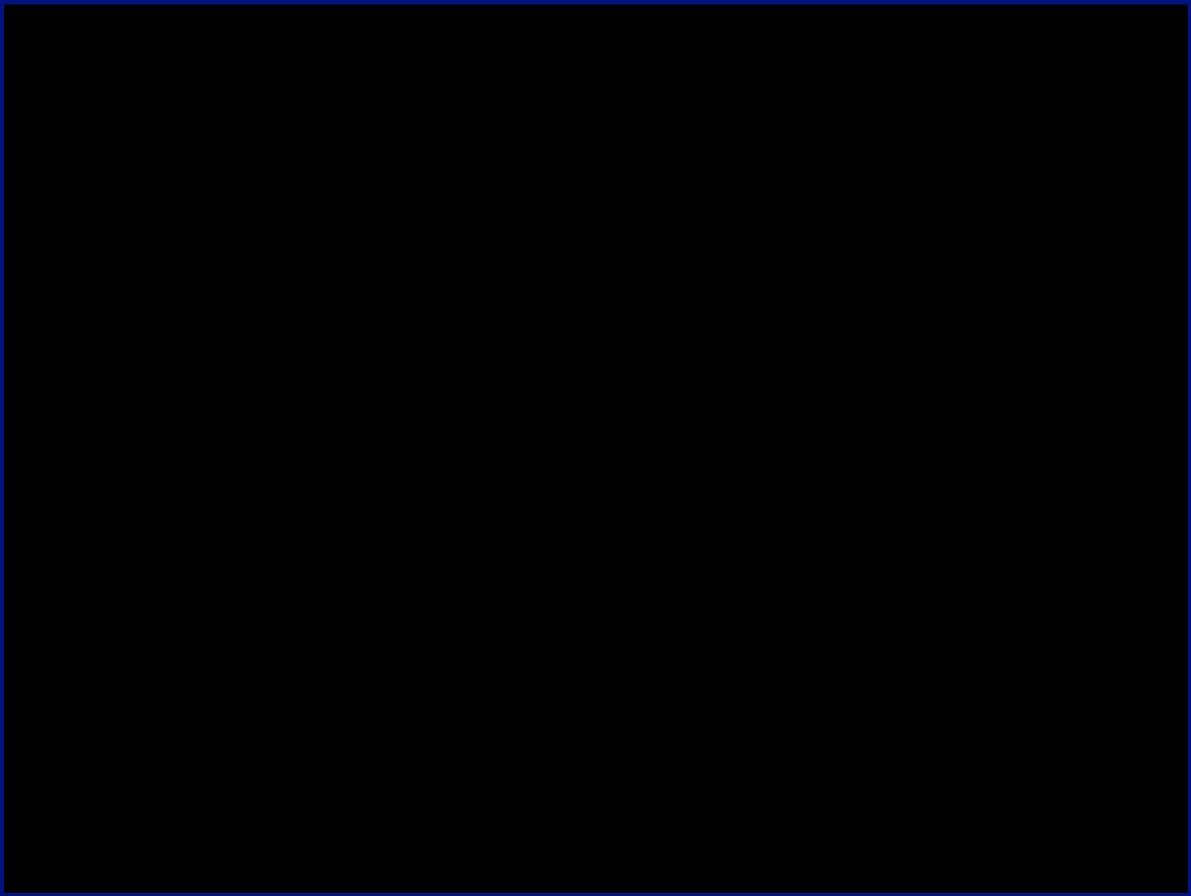
	Camera 0			Camera 1			Camera 2			Camera 3			Camera 4		
	QV	SS	CV	QV	SS	CV	QV	SS	CV	QV	SS	CV	QV	SS	CV
Camera 0	76	76	**84**	72	78	**79**	61	69	**79**	62	**70**	68	30	45	**76**
Camera 1	69	**77**	72	76	78	**85**	64	**74**	**74**	68	67	**70**	41	44	**66**
Camera 2	62	66	**71**	67	71	**82**	68	74	**87**	67	64	**76**	43	54	**72**
Camera 3	63	69	**75**	72	70	**75**	68	63	**79**	73	68	**87**	44	44	**76**
Camera 4	51	39	**80**	55	39	**73**	51	52	**73**	53	34	**79**	51	66	**80**

But what about composition?

Conclusions

- Absent taxonomy/composition is a major nuisance
  - if it were not for this question, appearance methods would win hands down
- What do we need to say about activity?
  - should we name activity, or reason about goals, intentions?
  - what about the objects nearby?
- Object recognition is in a fool’s paradise
  - unknown names, etc.

Bonus question



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

Composition

- Very little is known
- Idea
  - Activity recognition is more like clustering than like recognition
- Features
  - describe activities by comparison to other activities
    - rather than with absolute discriminative repn

American Sign Language (ASL)

- Generative models popular in the literature
  - Using HMM’s
    - [Grobel, Assan 97], [Bauer, Hienz 00], [Vogler, Metaxas 98,99,03], [Gao, et. al. 00], [Bowden et. al. 04], [Kadous 96], [Matsuo 97], [Zieren, Kraiss 04], [Starner+Pentland 95] etc - long literature
- Few discriminative models
  - Discriminative word spotter for small vocabulary
    - [Farhadi, Forsyth 06]

Easy to get dubious data

Dictionary

Generated by SigningAvatar

Comparisons are good features

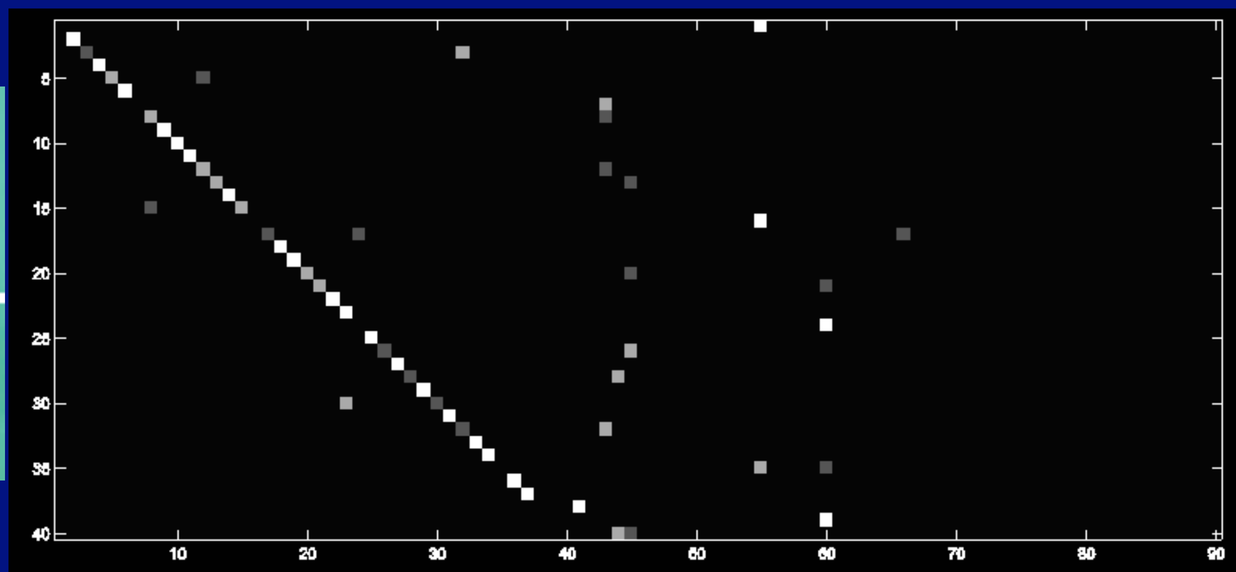
- Evidence
  - By adroit use of comparisons in sign language domain, we can
    - Build a set of comparative features
    - Learn to recognize a new word from one, dictionary example

Avatar   Human

- Learn on avatar
- Test on human signer
- Target words: 40 words
- Shared words: 50 words
  - Vocabulary size: 90 words
    - 40->90 classification problem
- 99.1% error rate using SVM

Results

90-Class Classification results on words that have never been seen in this rendering



Class confusion matrix for transfer from frontal avatar to frontal human signer.
64.17% of classification attempts are successful. (error rate of 35.83%)
Classified words have never been seen in frontal human signer.
Controls:
Without comparative features 98.2% error rate (c.f. our error rate of 35.8%)
PCA instead of random projections 64.3% error rate (c.f. our error rate of 35.8%)