Activity, Appearance, Aspect and Attributes

D.A. Forsyth, UIUC
with
Okan Arikan (UT Austin), Nazli Ikizler (Boston U), Leslie Ikemoto (animate-me), Derek Hoiem (UIUC), Ali Farhadi (UIUC), Ian Endres (UIUC), Ryan White (Euclid media)

Obtain dataset

Build features

Mess around with classifiers, probability, etc

Produce representation

Computer vision

Obtain dataset

Build features

Light entertainment
(the way we do it)

Mess around with classifiers, probability, etc

Computer vision

Produce representation

Big questions

Computer vision

- What signal representation should we use for activity recognition?
  - Compare
    - Appearance (do not segment bits and pieces explicitly)
    - Kinematic (segment bits and pieces explicitly)

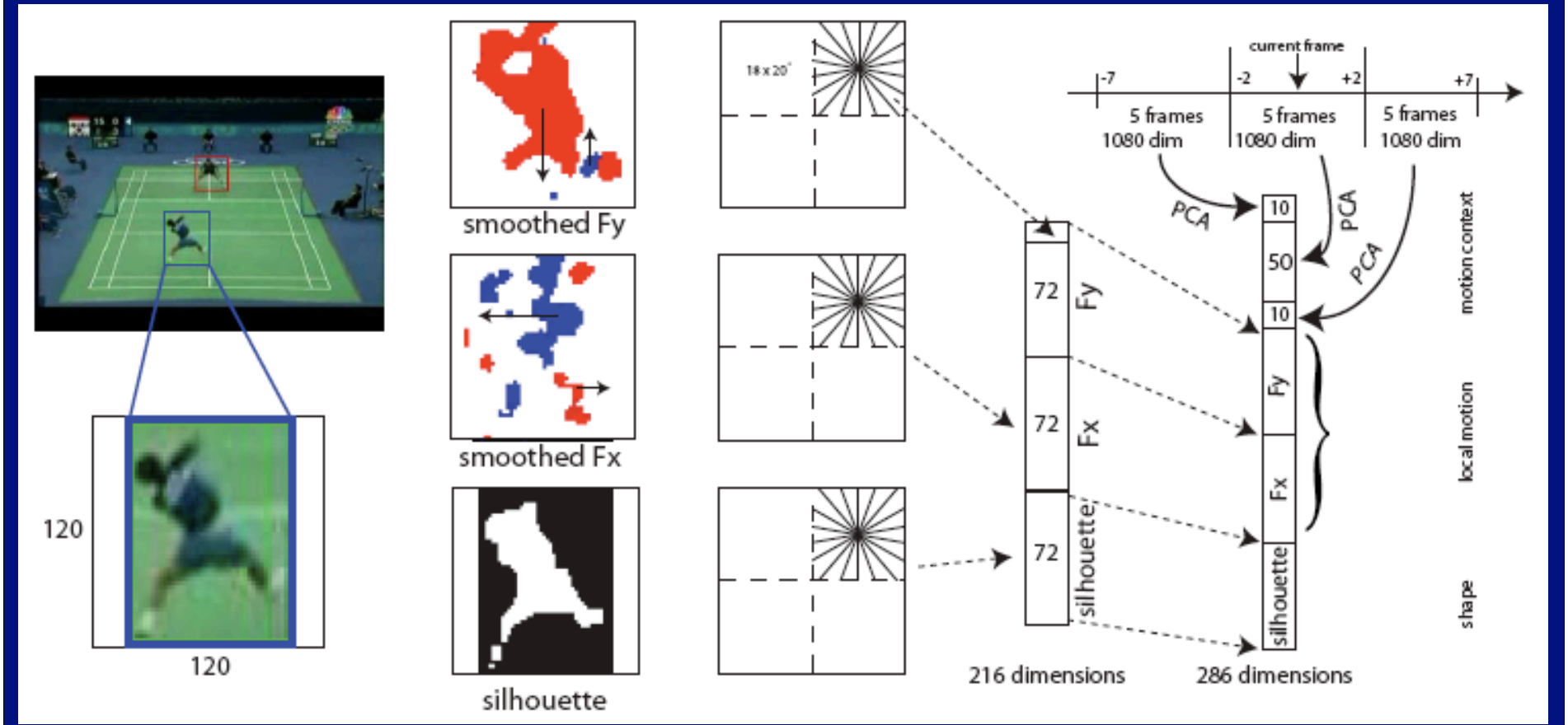
Computer vision

- Does a discriminative framework make any sense? For activity
  - Compare
    - Walk; run; etc
    - (rather vague)

Appearance features

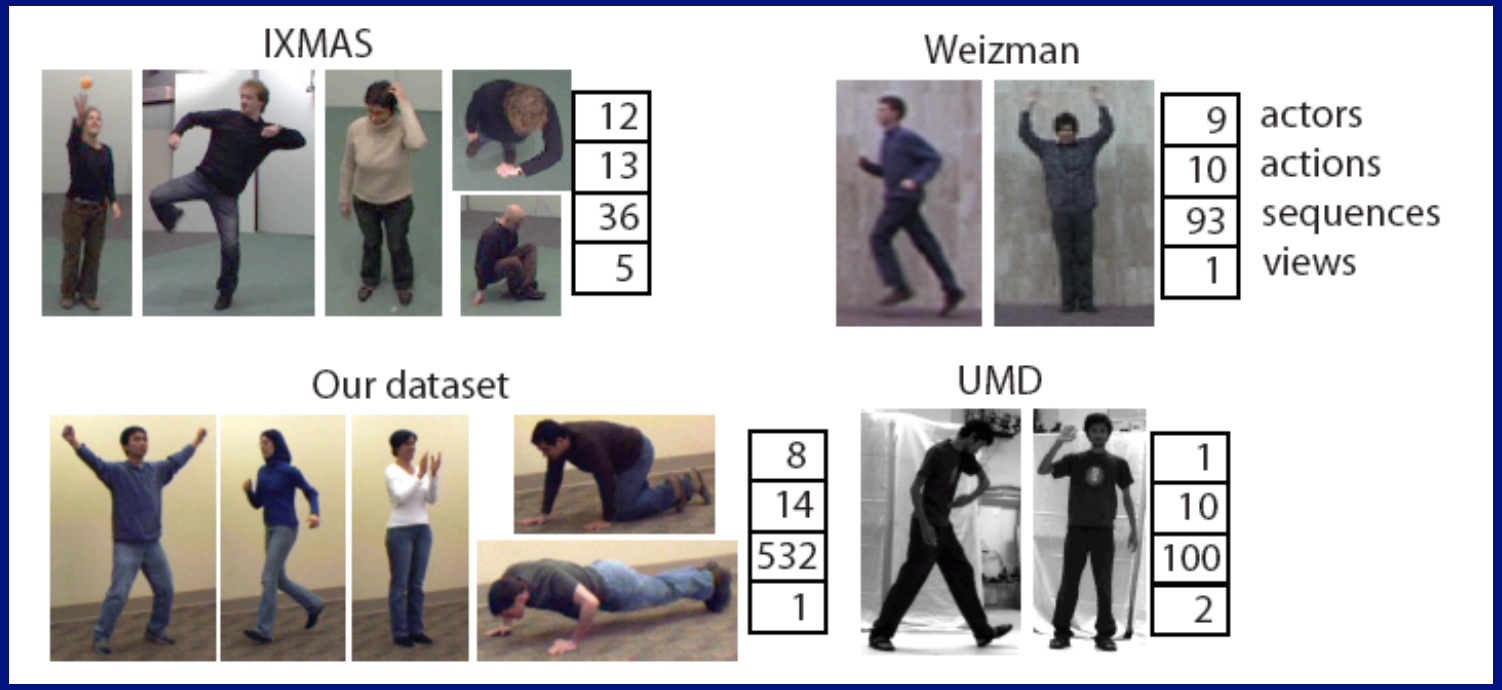
- Less nasty segmentation
  - (body from background, perhaps not even that)
    - Spatio-temporal volumes
      - (e.g. Davis+Bobick 97; Blank et al 05)
    - Motion trends/flow fields
      - (e.g. BobickDavis 96; Davis 01; Efros et al 03; Laptev+Perez 07; Laptev et al 08)
    - Spatio-temporal interest points
      - (e.g.Niebles et al 06; 08; Scovanner et al 07)
    - Various mixtures of these

An Appearance feature



Tran and Sorokin 08, after Duygulu and Ikizler 07

Datasets



Tran + Sorokin 08

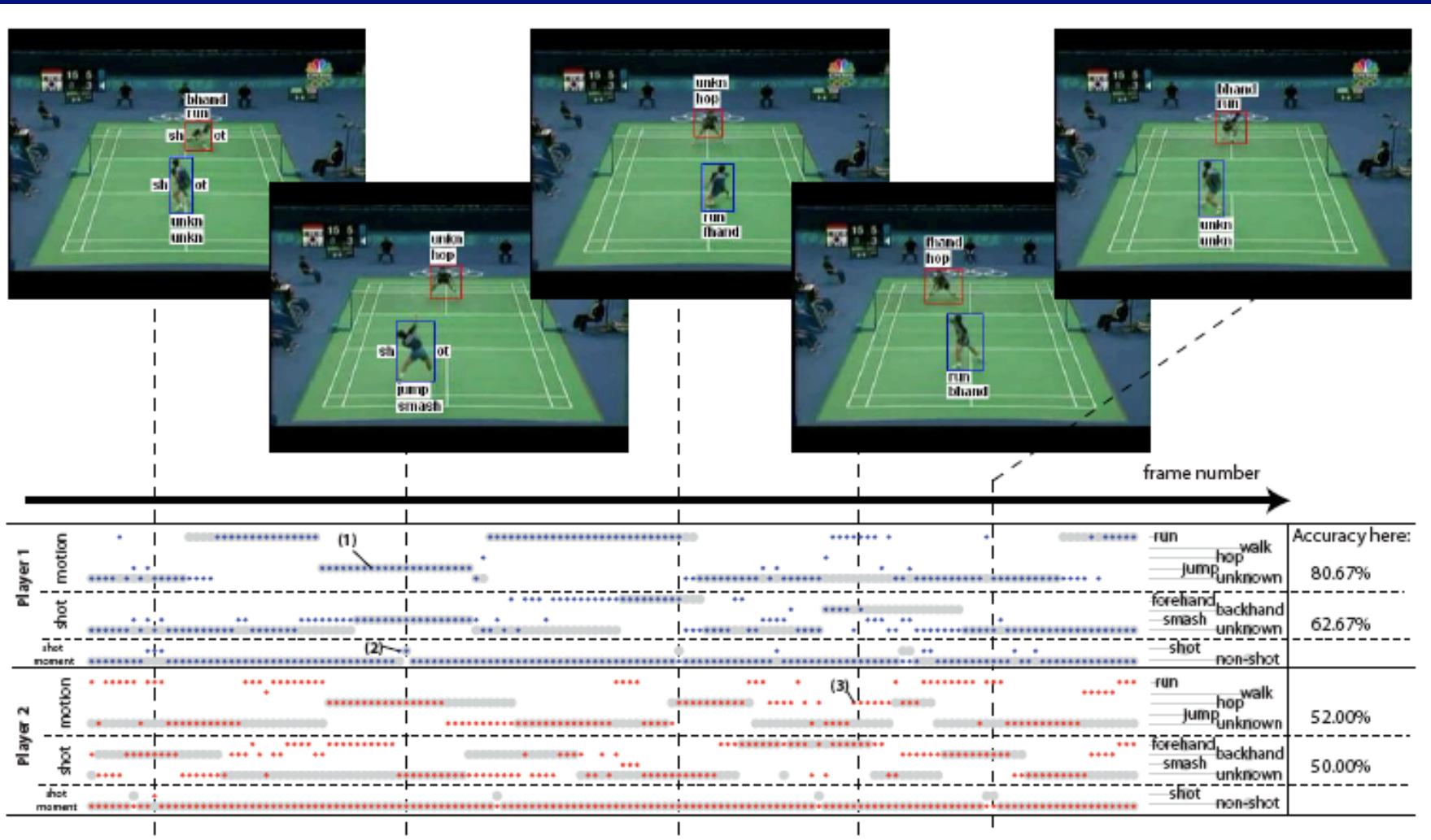
Discriminative results

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**				

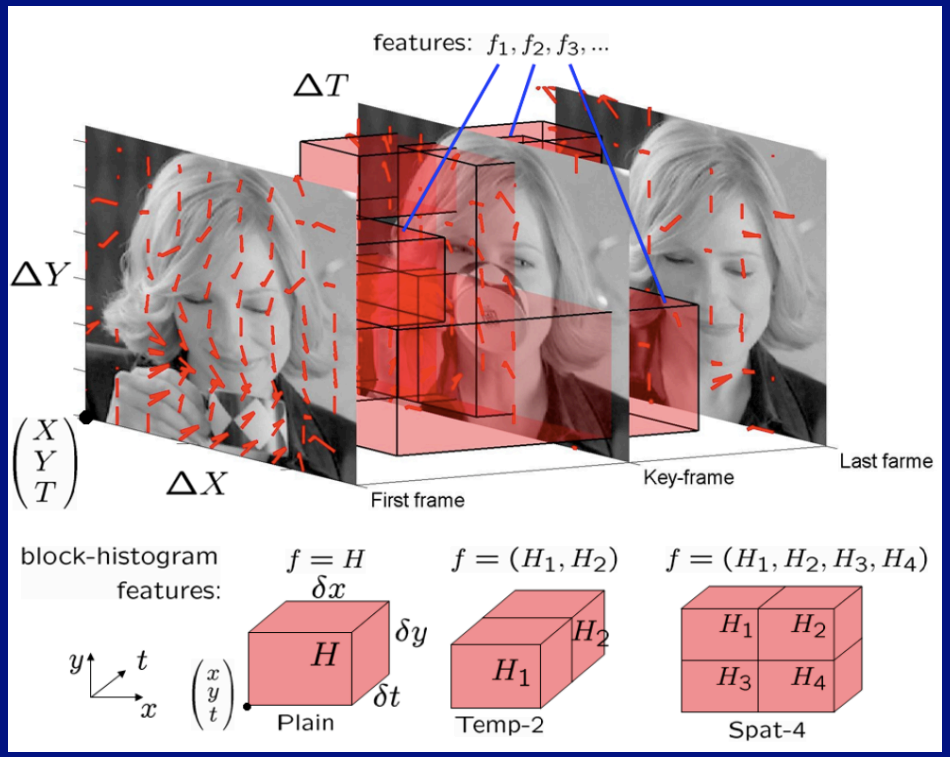
Works well, depending on task; not rejecting improves things
metric learning improves things

Tran + Sorokin 08

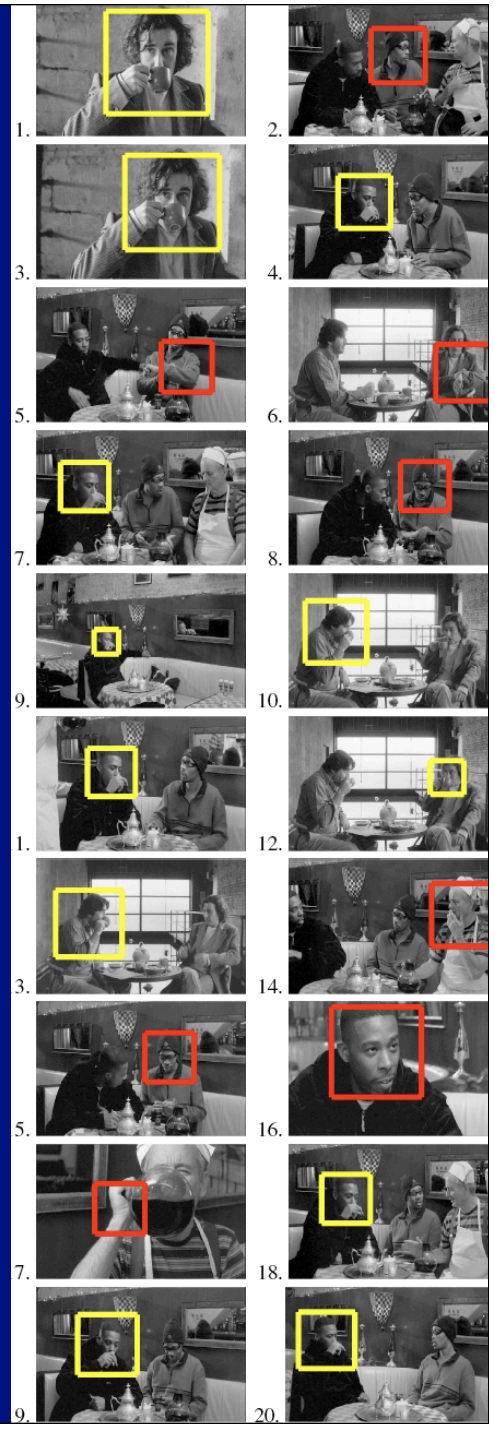
Youtube video



Tran + Sorokin 08



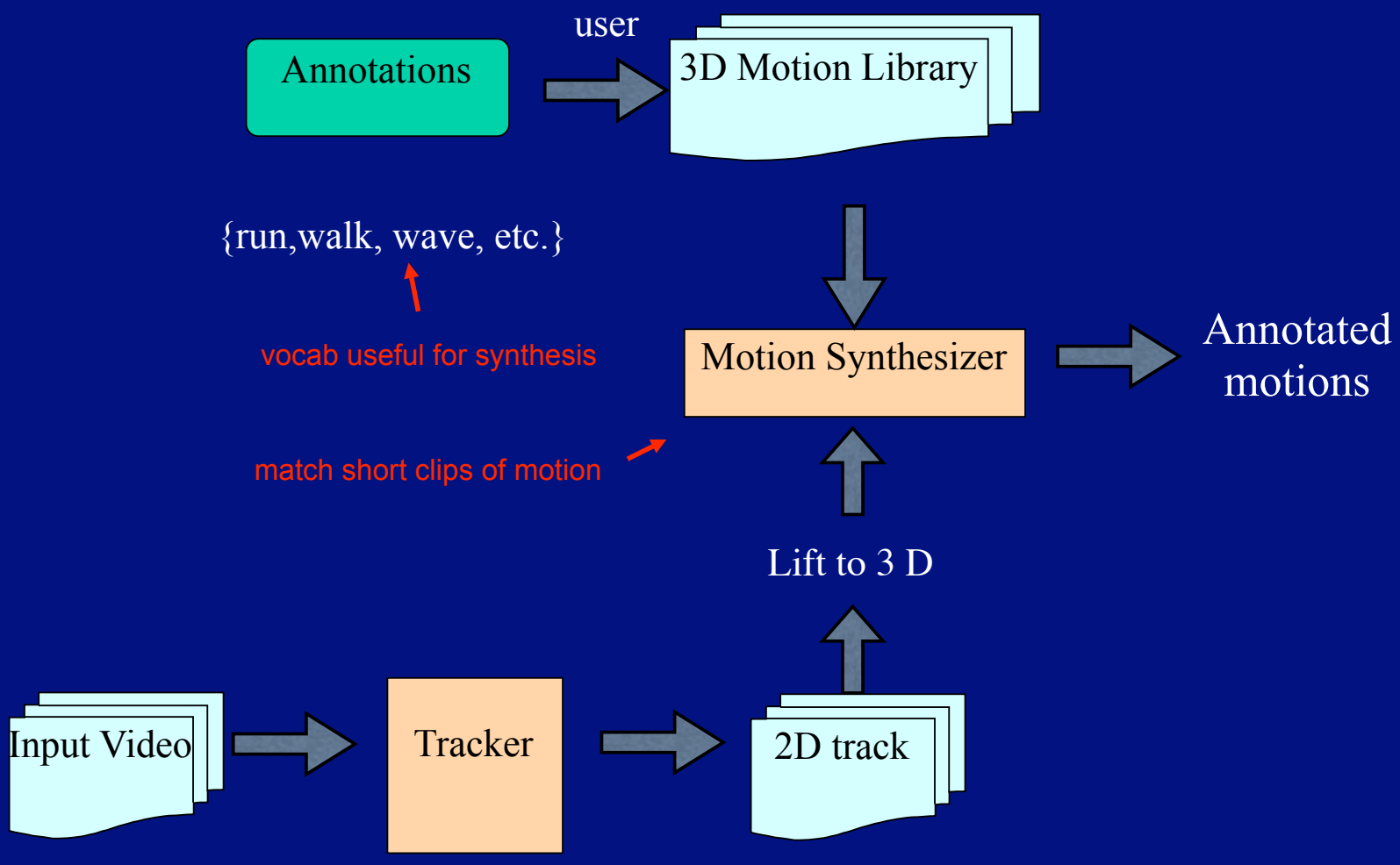
Laptev Perez 2007
see also Laptev et al 08

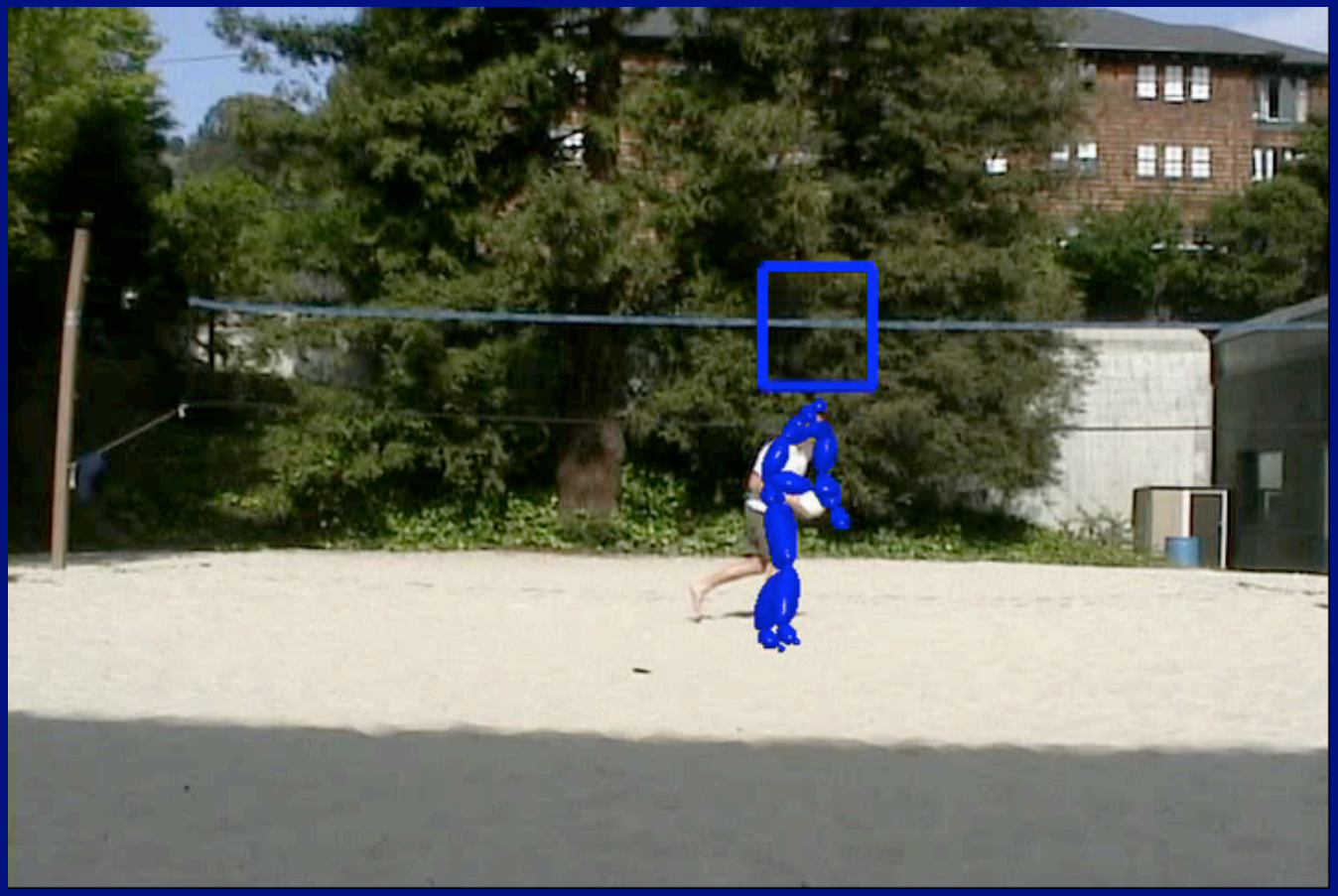


Kinematic features

- Find body parts
  - with geometric/appearance model (deformable template)
  - cardboard people
    - (eg Ju et al 96; Sidenbladh et al 2000)
  - pictorial structures
    - (eg Felzenszwalb Huttenlocher 05)
  - kinematic tracks
    - (eg Ramanan et al 05)
  - repeated model-based segmentation
    - (eg Ferrari et al 08)

Annotating observations by synthesis





Ramanan+Forsyth 03

Criteria

- Base accuracy?
  - appearance wins hands down on current datasets
- Aspect
  - appearance can be fixed
- Do they solve the right problem?
  - advantage: kinematic

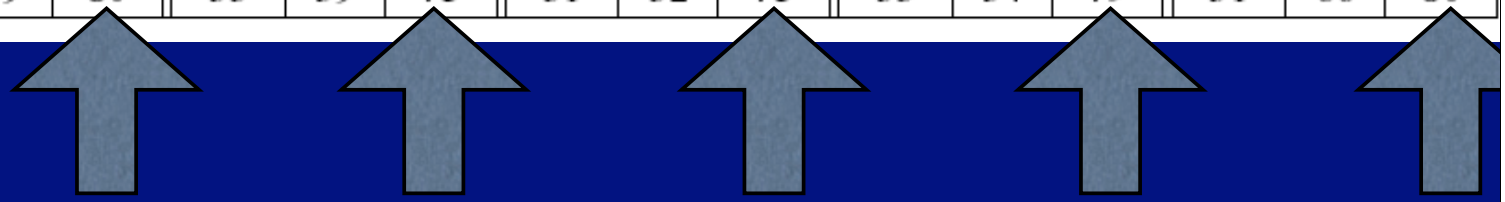
IXMAS and Aspect

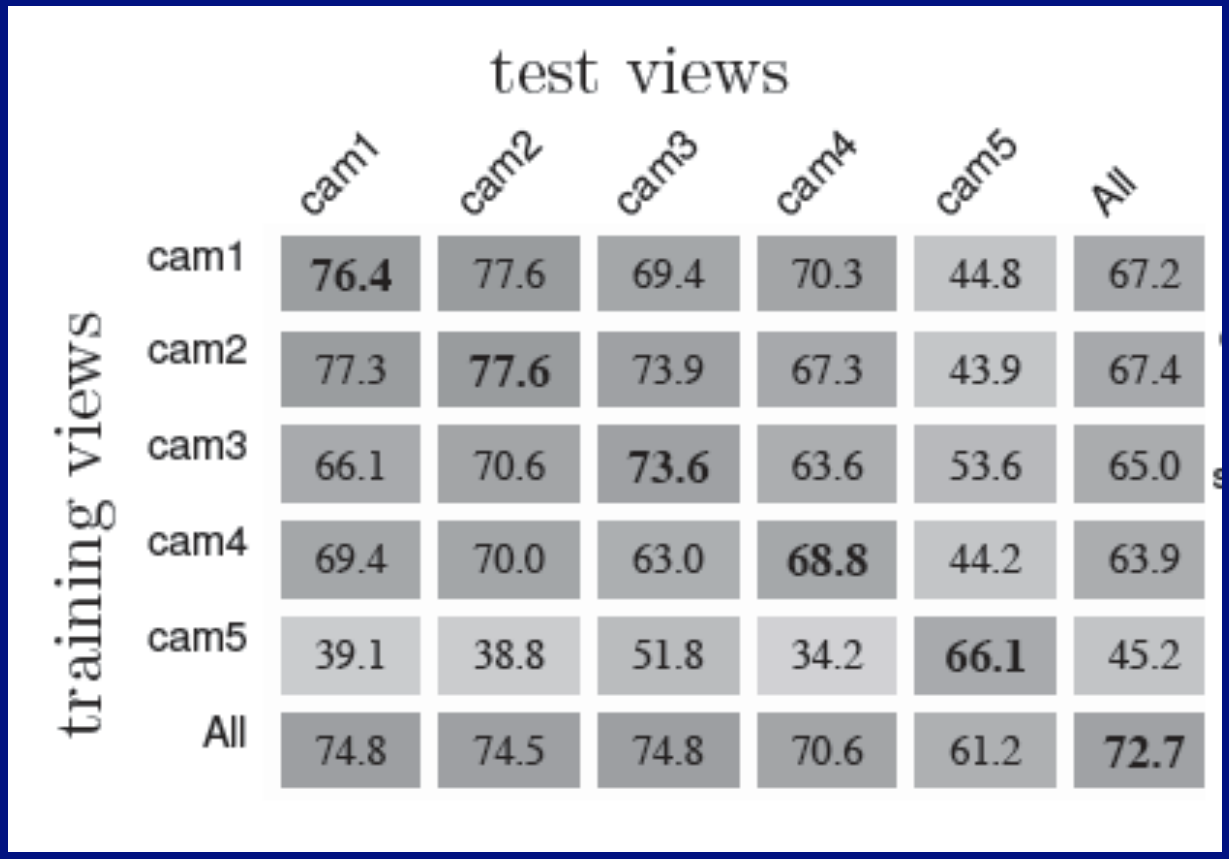
The Effects of Aspect

	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

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**





Junejo et al 08, different feature construction, same dataset

The problem we have been solving

- Rack up a bunch of activity categories, and discriminate
  - how many categories are enough?
  - can one movement have two categories?
  - what are the categories?
    - the verb argument (probably) fails
      - if there are few movement, many goal verbs
    - introspection suggests too few words

AnswerPhone, GetOutCar, Handshake, Kiss, Hugperson, SitDown, SitUp, StandUp

Goal achieved by body movement

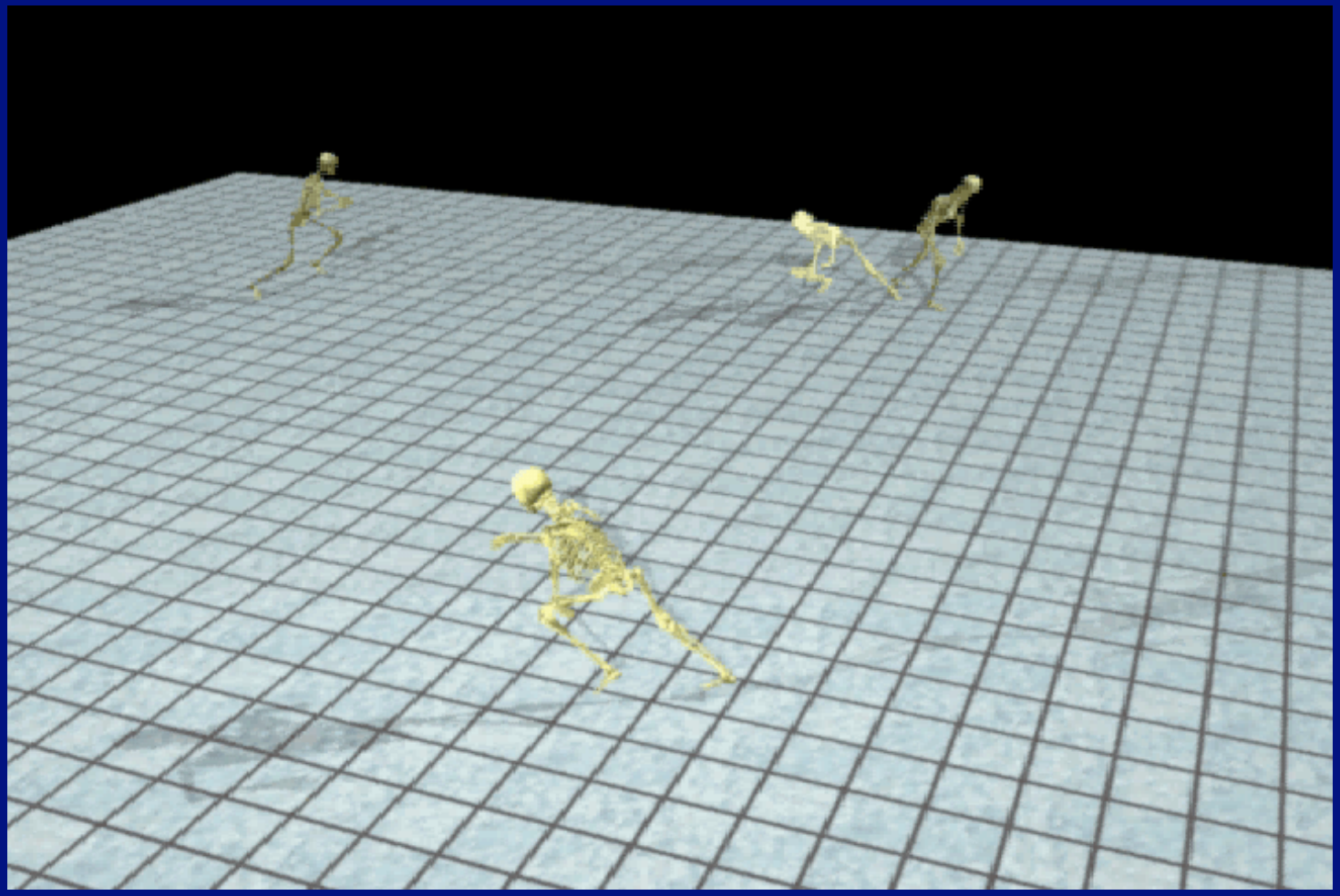
Body movement

Components of the problem we should be trying to map

- Activity composes freely into complex structures
- Most human activities cause changes of state, meet goals
  - similar movements will meet different goals
  - different movements can meet the same goal
- We should probably be trying to “recognize” things
  - whose names we do not know
    - fluid, changing categories, affected by
      - nearby objects
      - observer, observation context
  - for which we have seen no examples

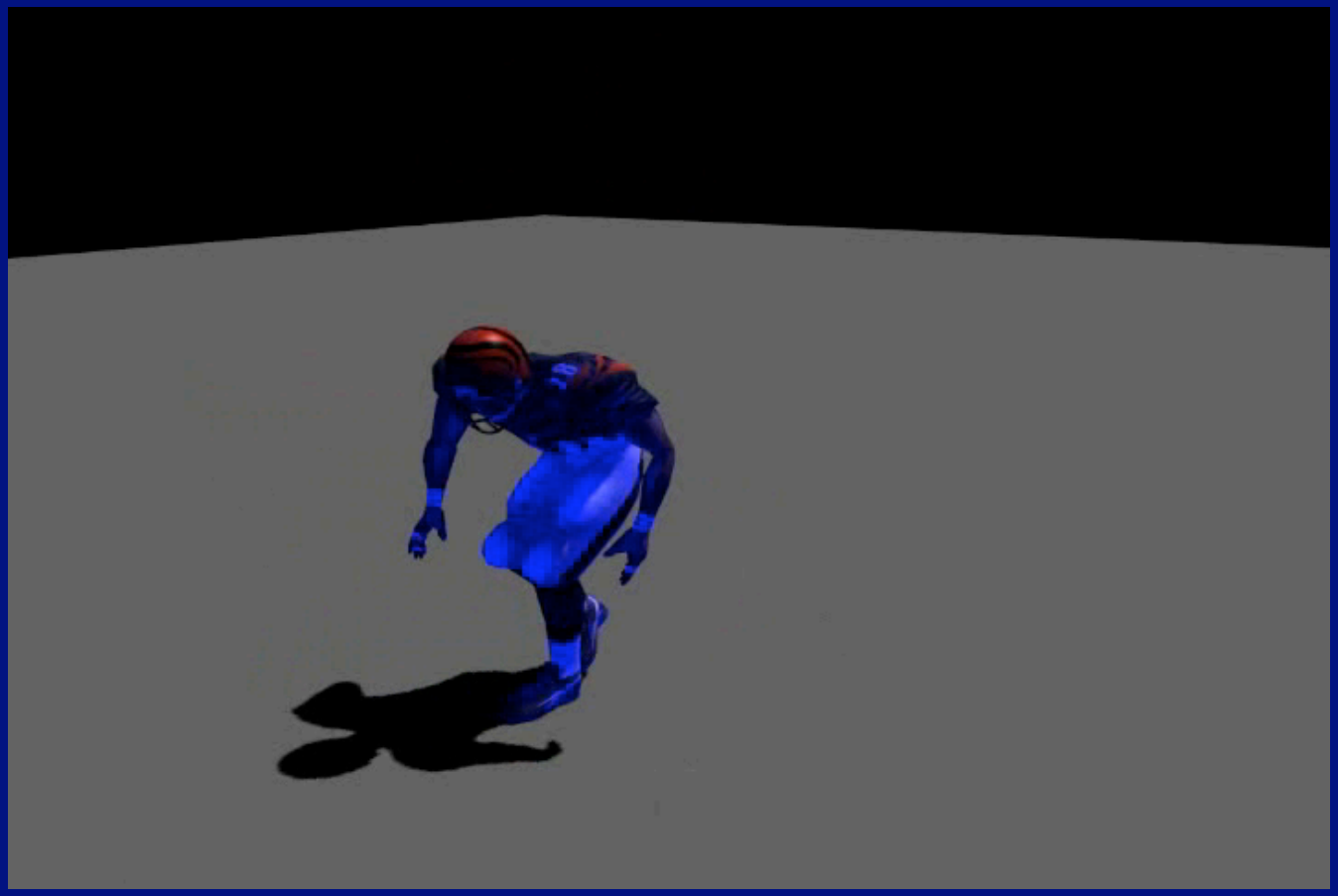
Composition and Activity

- 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)



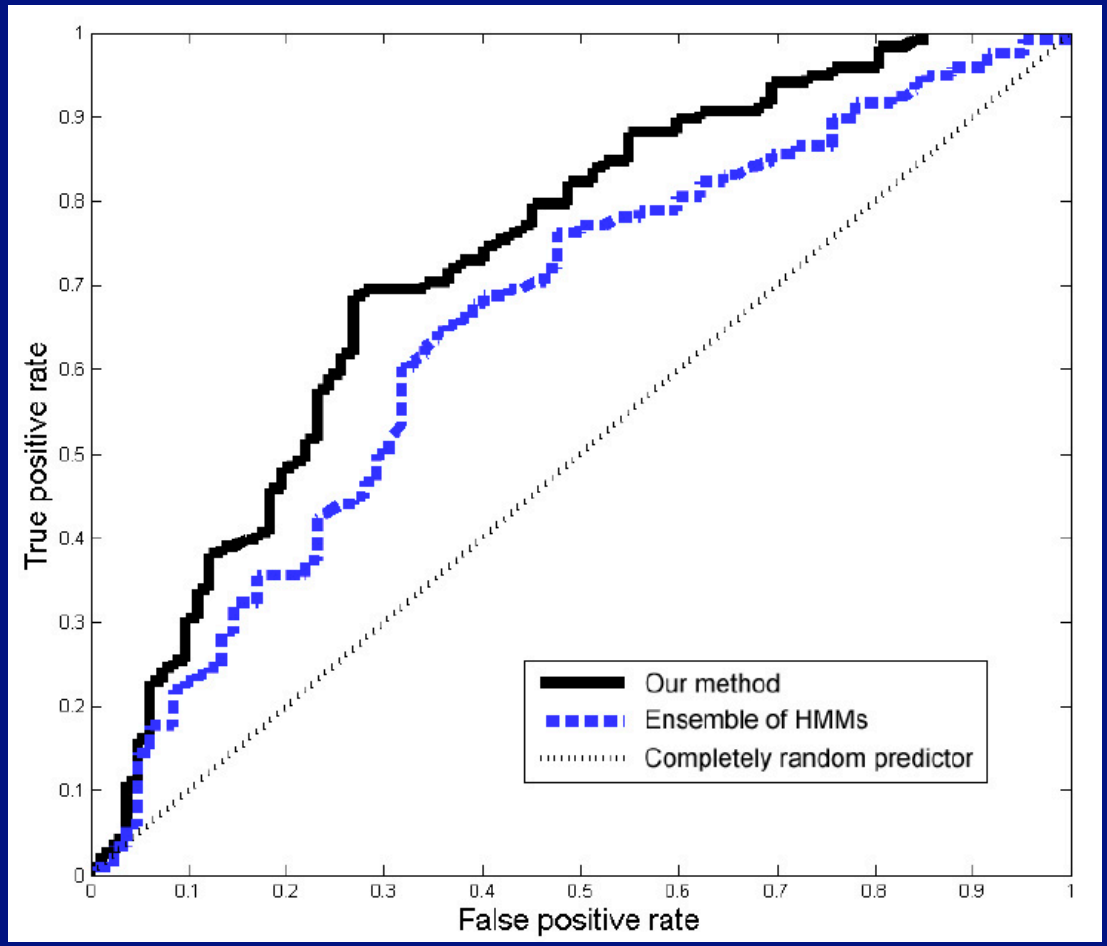
Arikan+Forsyth 02

Ikemoto+Forsyth 04



Ikemoto+Forsyth 04

Hard to tell good from bad

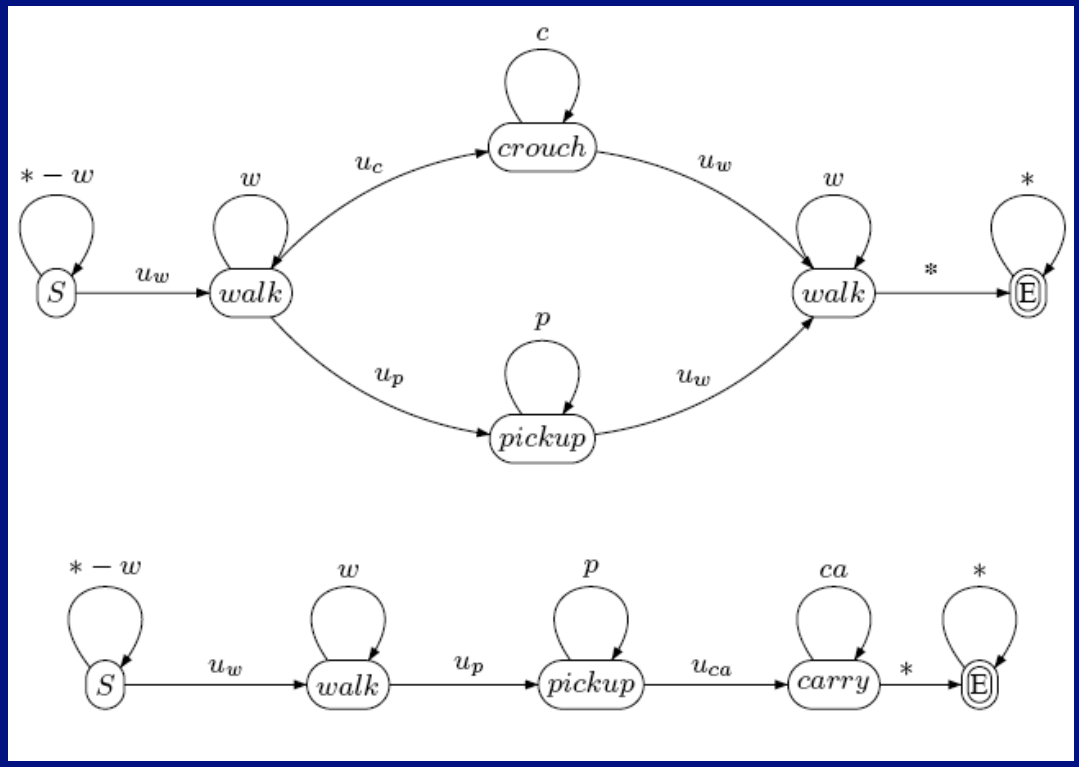


Ikemoto Arikan Forsyth 07

cf Ren et al 05 for HMM's

“Recognizing” composites

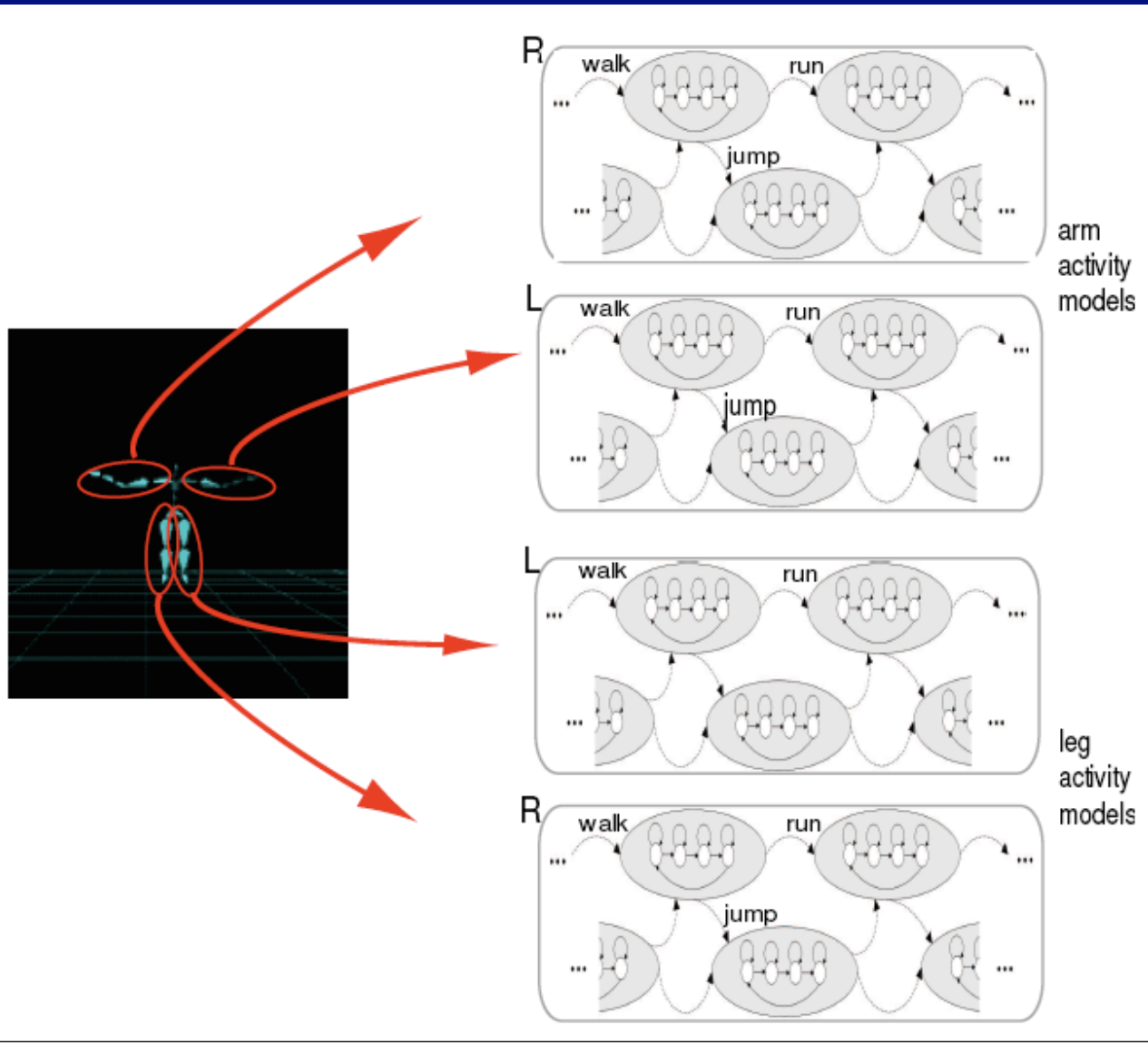
- Rank sequences by P(FSAldata, model)
  - e.g. P(leg-walk-arm-walk-then-leg-walk-arm-reachl data, model)
  - DP variant will do this easily



Building a composite model

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

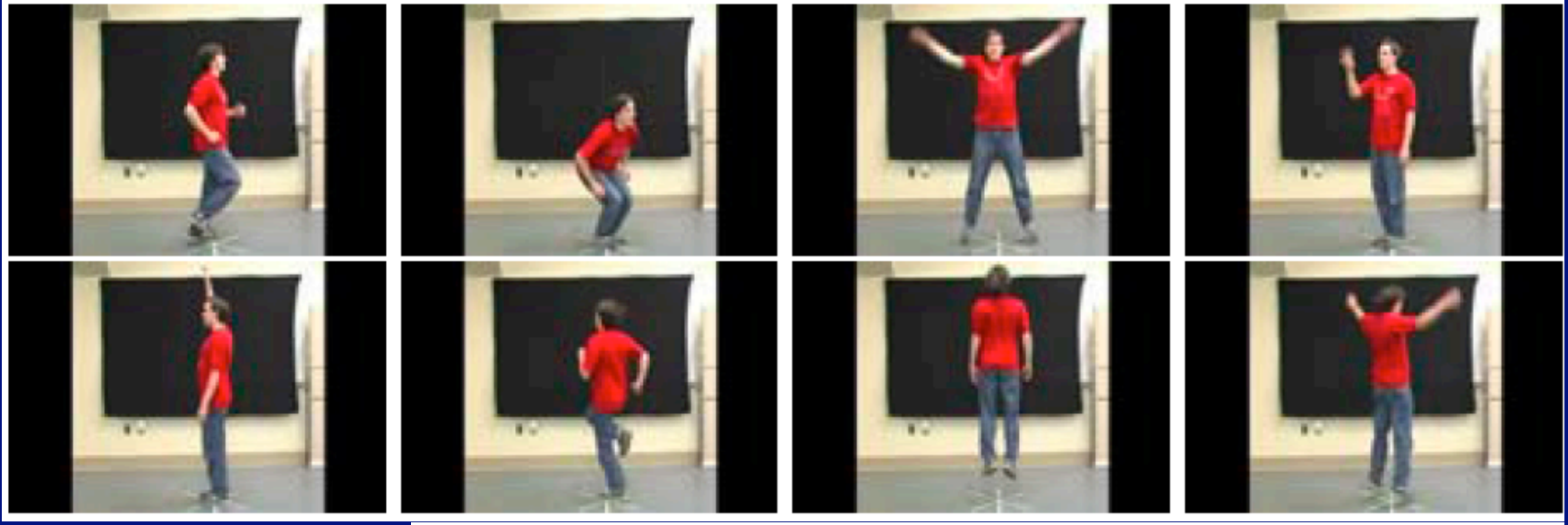
Composition



Ikizler Forsyth 07,08

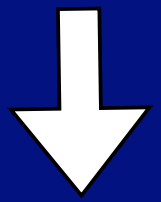
Emission

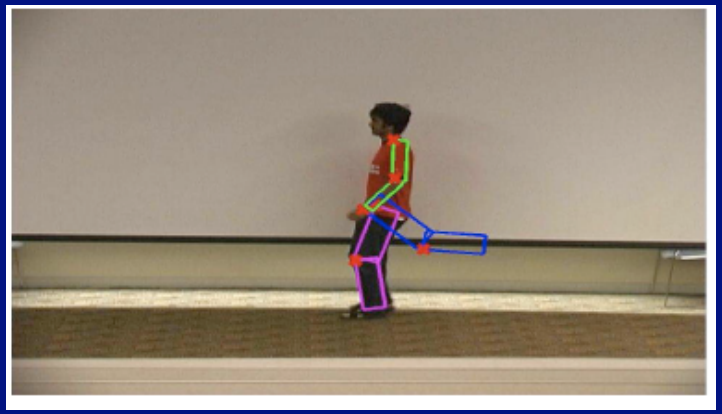
- Transduction
  - Track the body, as above
  - Lift “snippets” of each quarter
    - vector quantized
  - impose root consistency
- Emission
  - emit cluster center from state according to table
  - table learned by EM, known dynamical model



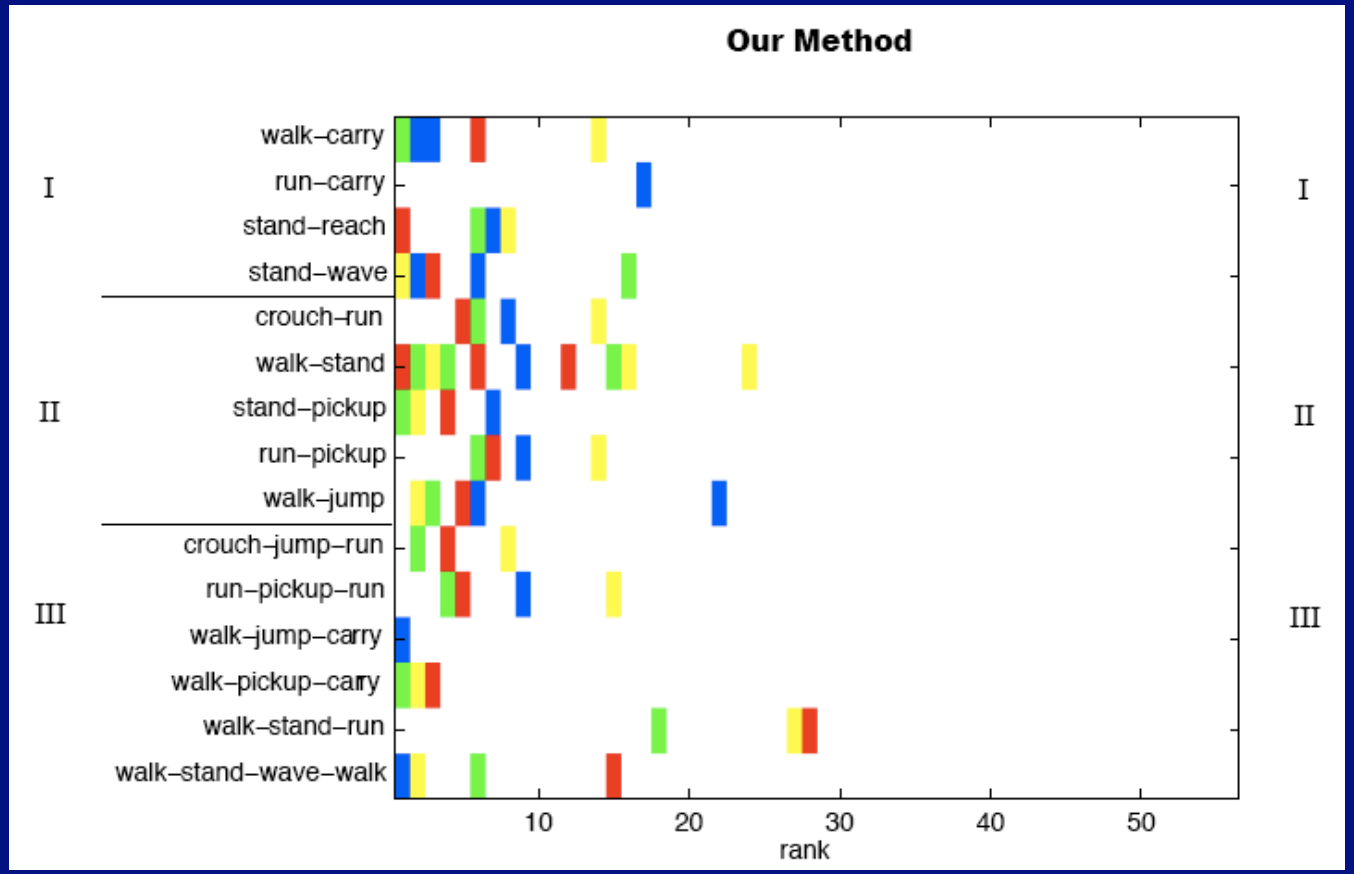
Ikizler Forsyth 07,08

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





Ikizler Forsyth 07,08



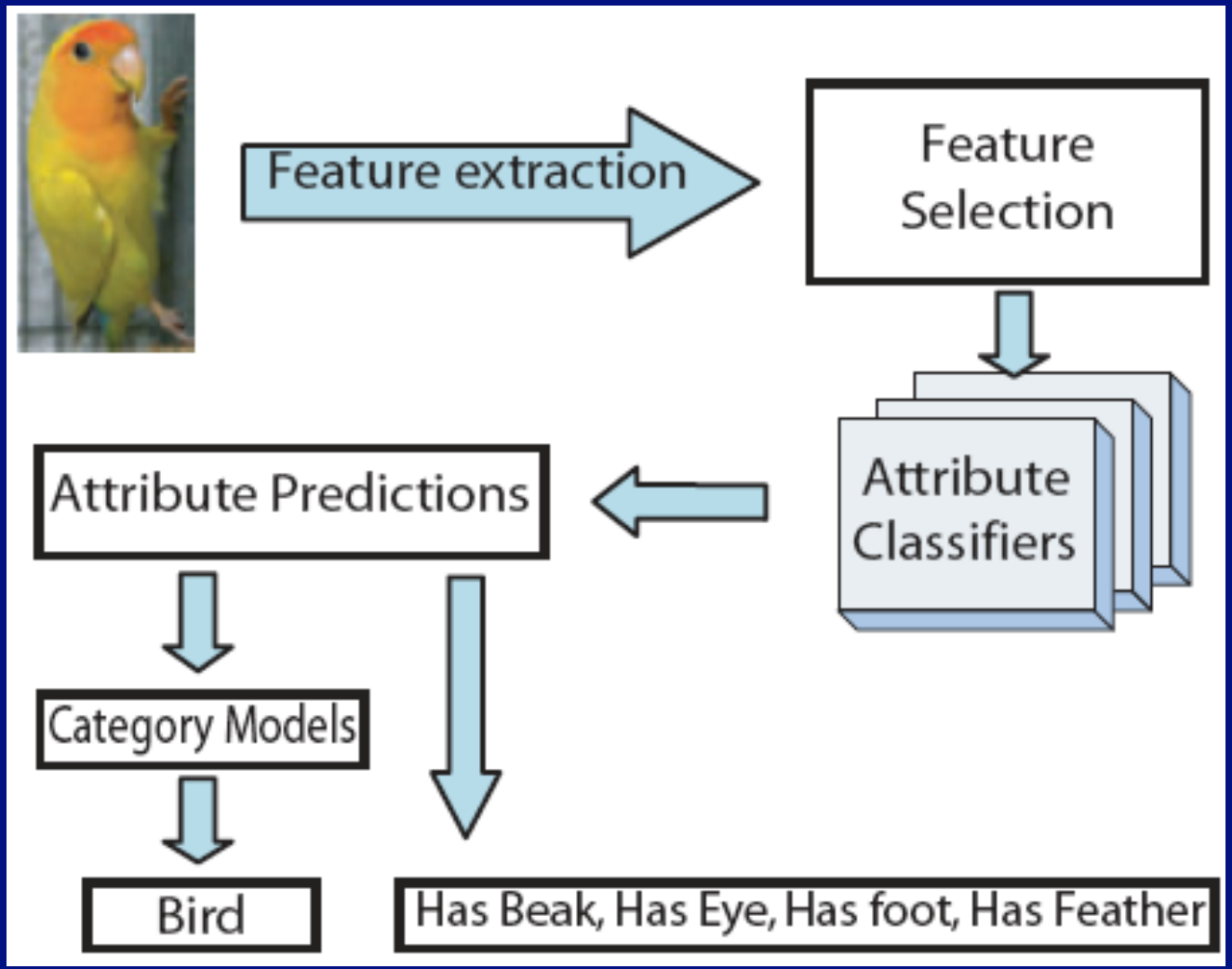
Ikizler Forsyth 07,08

How do you describe something whose name is unknown?

- Attributes
  - Properties shared by many object categories
  - Material (like)
    - glass, wood, furry, red, etc.
  - Part (like)
    - has wheel, has head, has tail, etc.
  - Shape (like)
    - is 2D Boxy, is cylindrical, etc
- What do we need to say about activity?
  - should we name activity, or reason about goals, intentions?
  - what about the objects nearby?

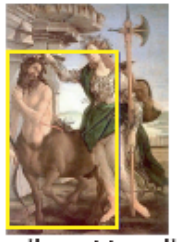
Farhadi et al 09;
cf Blaschke ?09;
Ferrari Zisserman 07;

General architecture



'is 3D Boxy'
'is Vert Cylinder'
'has Window'
'has Row Wind'
'has Headlight'

'has Hand'
'has Arm'
'has Screen'
'has Plastic'
'is Shiny'



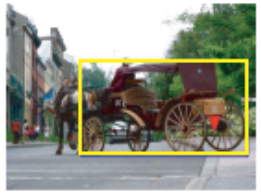
'has Head'
'has Hair'
'has Face'
'hasSaddle'
'has Skin'

'has Head'
'has Torso'
'has Arm'
'has Leg'
'has Wood'

'has Head'
'has Ear'
'has Snout'
'has Nose'
'has Mouth'

'has Head'
'has Ear'
'has Snout'
'has Mouth'
'has Leg'

has Furniture Back'
as Horn'
s Screen'
'has Plastic'
'is Shiny'



' is 3D Boxy'
'has Wheel'
'has Window
'is Round'
' 'has Torso'

'has Tail'
'has Snout'
'has Leg'
'has Text'
'has Plastic'

'has Head'
'has Ear'
'has Snout'
'has Leg'
'has Cloth'

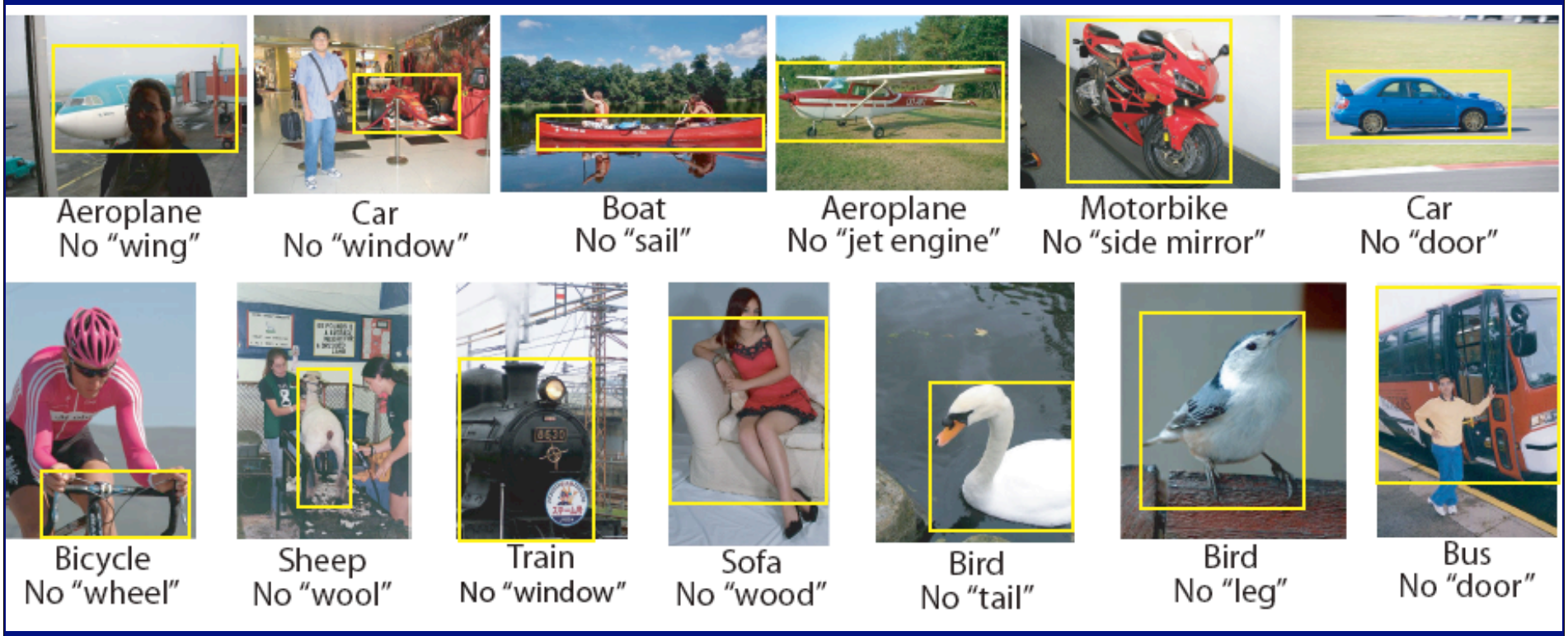
'is Horizontal Cylinder'
'has Beak'
'has Wing'
'has Side mirror'
'has Metal'

'has Head'
'has Snout'
'has Horn'
'has Torso'
'has Arm'

How is an object different from typical?

- Pragmatics suggests this is how adjectives are chosen
  - If we are sure it's a cat, and we know that
    - an attribute is different from normal
    - the detector is usually reliable
  - we should report the missing/extra attribute

Missing attributes



Extra attributes

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 has to deal with similar issues
  - unknown names, etc.