Activity representation and recognition

Take home points

- There is very seldom a taxonomy
- Generative models based around FSA/HMM are popular
- Discriminative models are well worth using
- Very little clear information about best ways to proceed.

Core difficulties

- The configuration of the body remains difficult to transduce
  - and may not be essential to understand what's going on
  - whence appearance, location based methods
- There is no natural taxonomy of activity
  - but we're beginning to get beyond walk, run, jump
  - introspection suggests taxonomy may be wrong approach?
- Composition and nulls create fearsome complexity
  - few representational methods can really deal with this
- The role of dynamics is uncertain
- What needs to be transduced?

Classes of method

- Appearance based
- Logical representations
- Finite state representations
  - fitted HMM
  - switching linear dynamical systems
- Discriminative methods
- Authored models

Temporal scale and activity

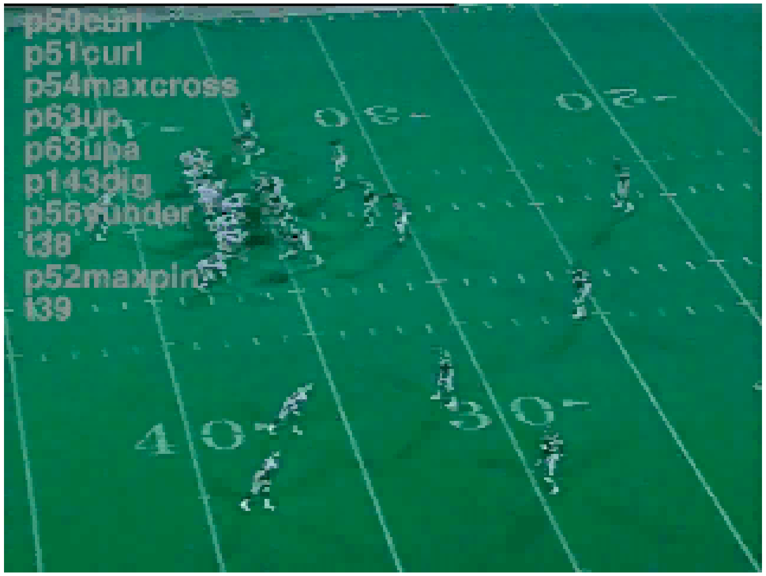
- Very short timescales
  - not much happens
    - low dimensional models seem to work in animation
    - motion compresses well
  - but body configuration is diagnostic
- Medium timescales
  - Motions can be (at least):
    - sustained (running, walking, jogging, etc. --- typically periodic)
    - punctate (jump, punch, kick)
    - parametric (reach, etc.)
- Long timescales
  - Motions are complex composites
    - visiting an ATM
    - reading a book
    - cooking a meal

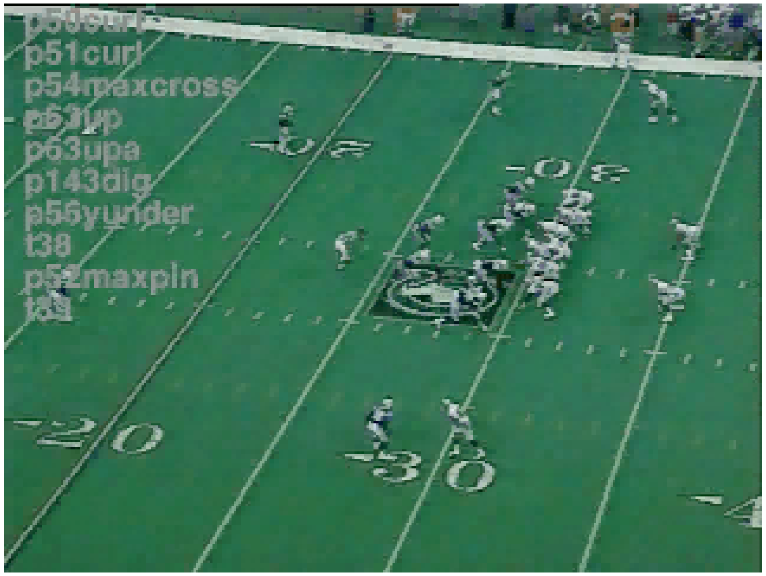
Appearance

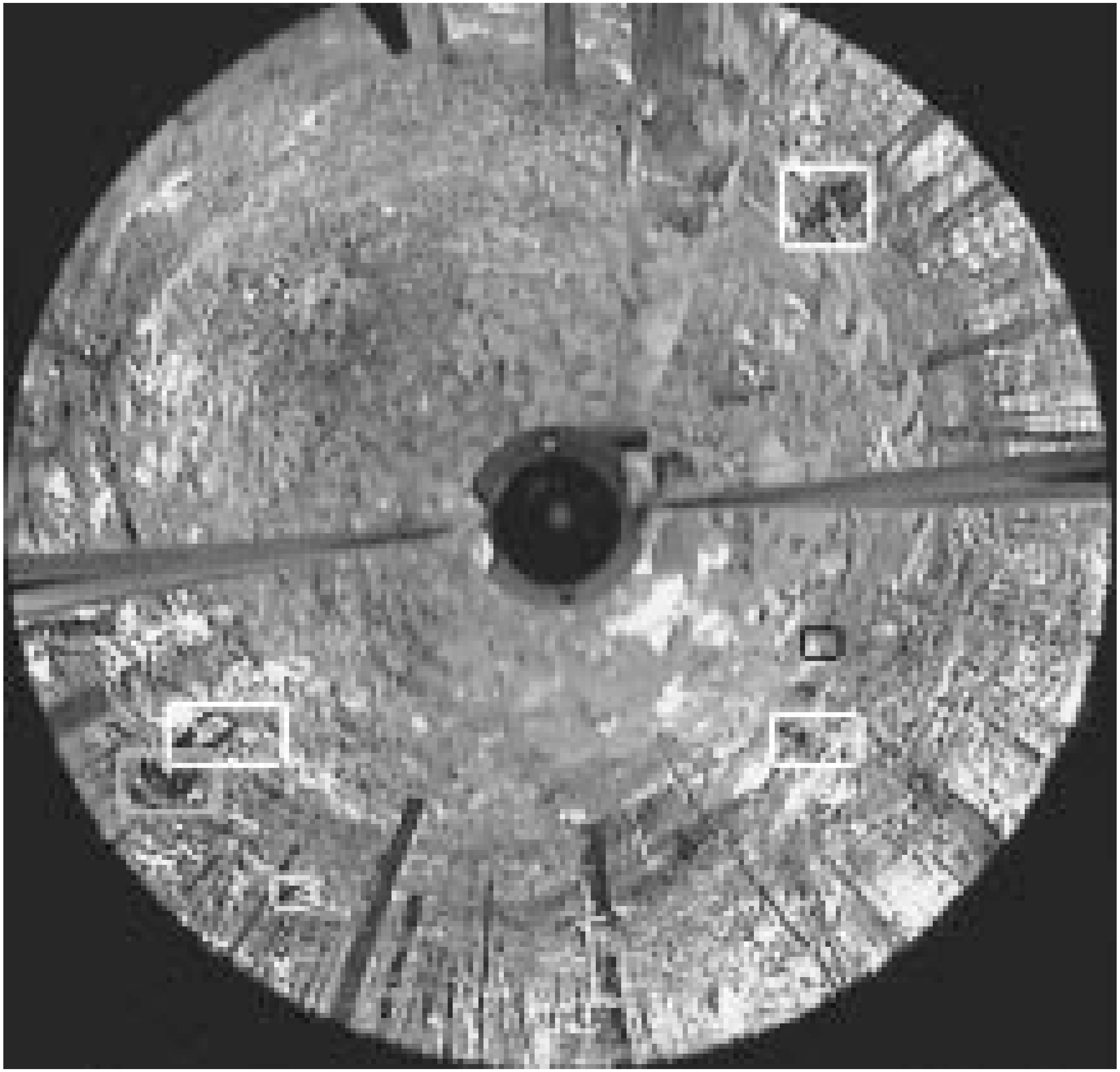
- Activities lead to characteristic patterns of image appearance
  - in grey level
  - in optic flow

Where you are is often a very powerful guide to what you are doing

Intille et al 95, 97



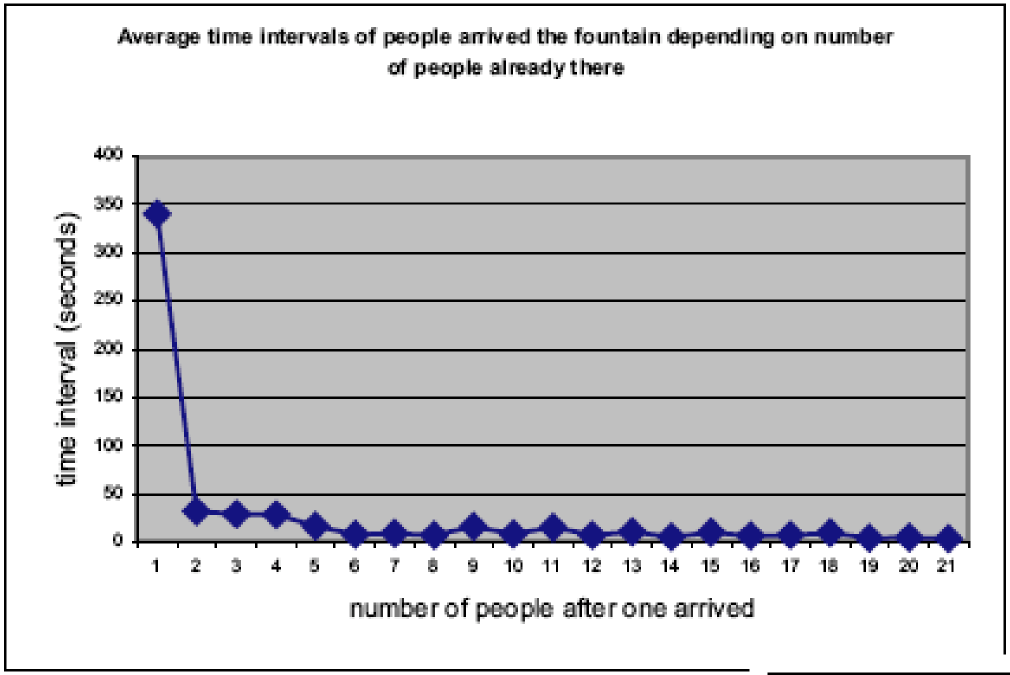




And can suggest
you are doing what
you should not be

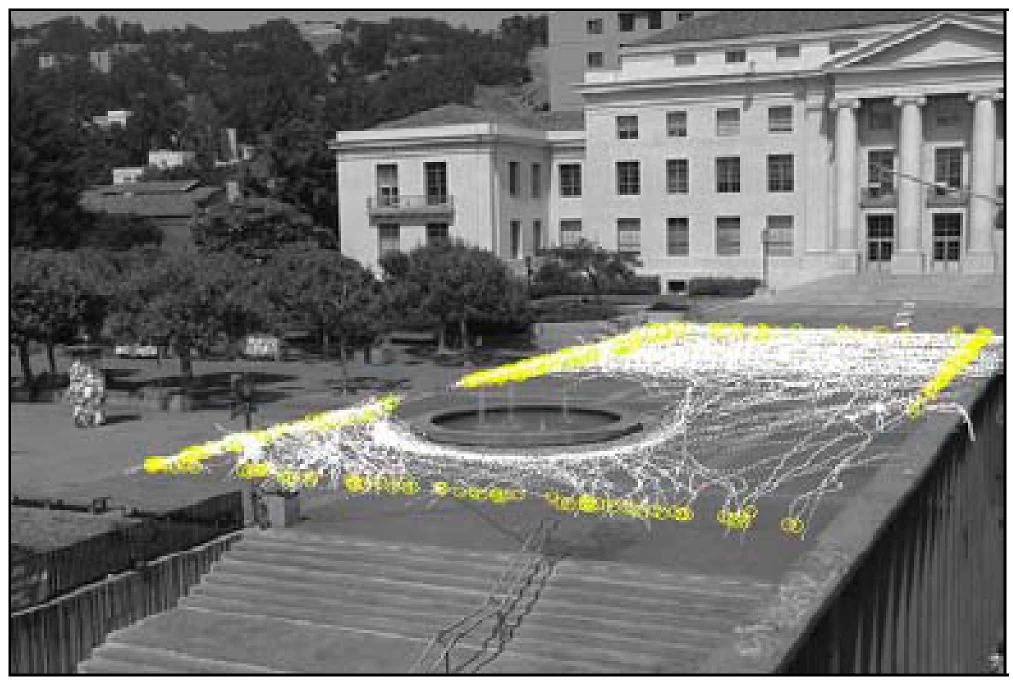
Boult et al 2001

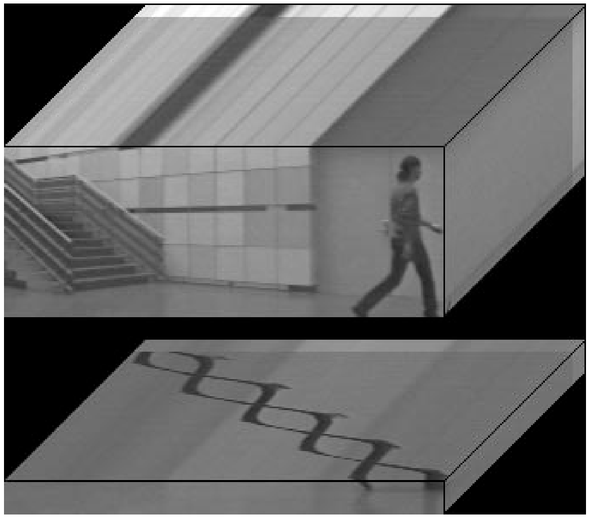
Surveillance by omnidirectional cameras,
detection of anomalous pixel groups



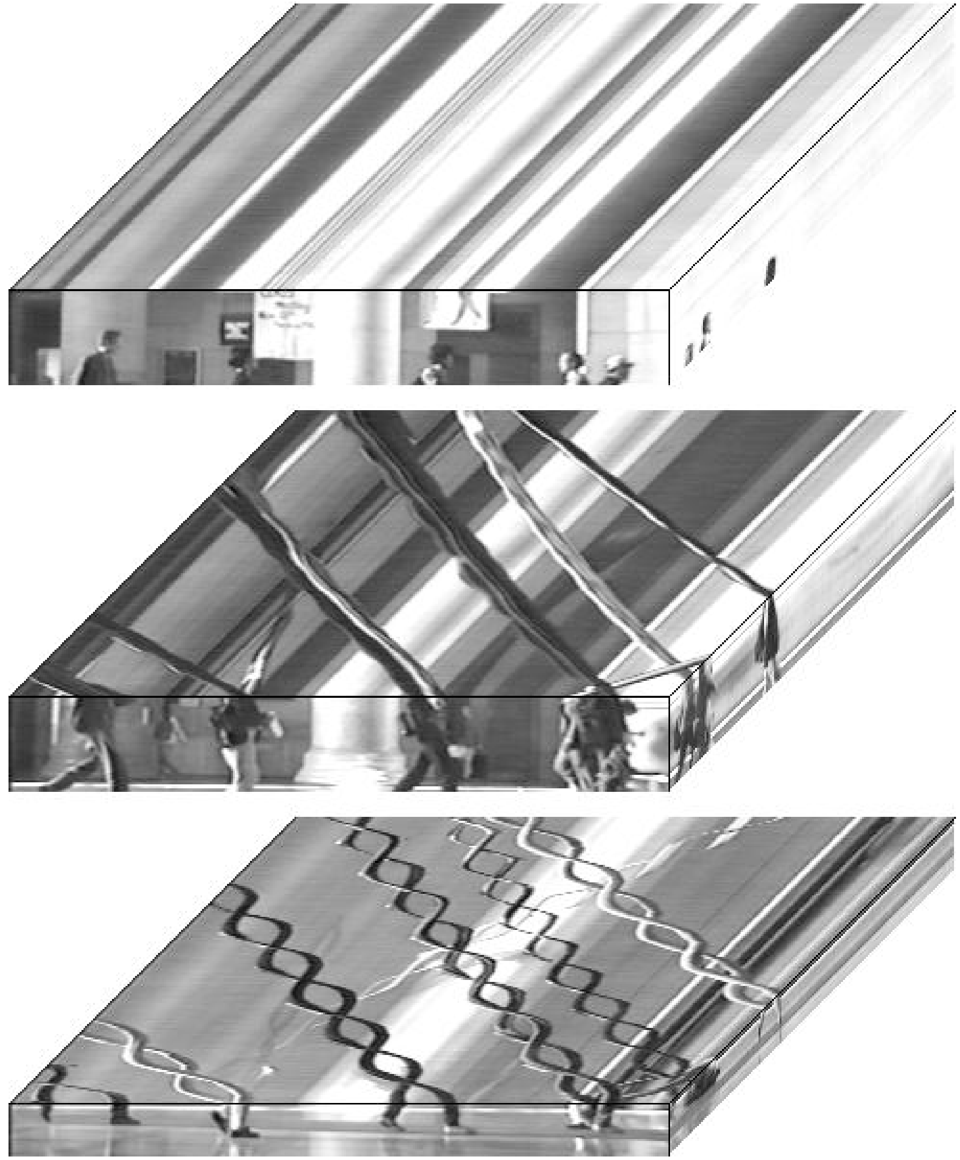
Numerous curious phenomena related to location

Yan + Forsyth 04

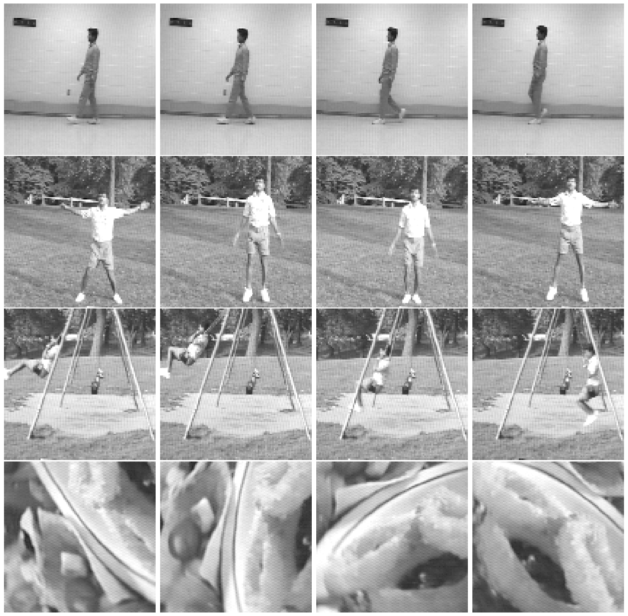




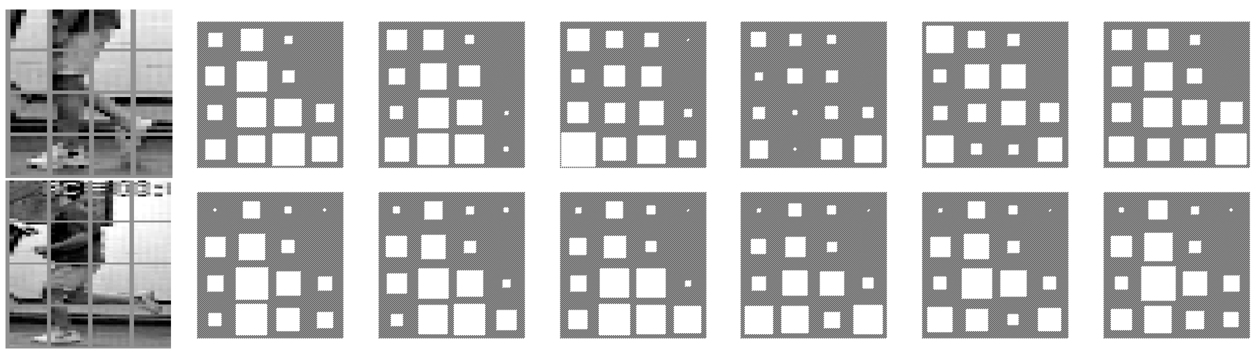
Niyogi Adelson 94

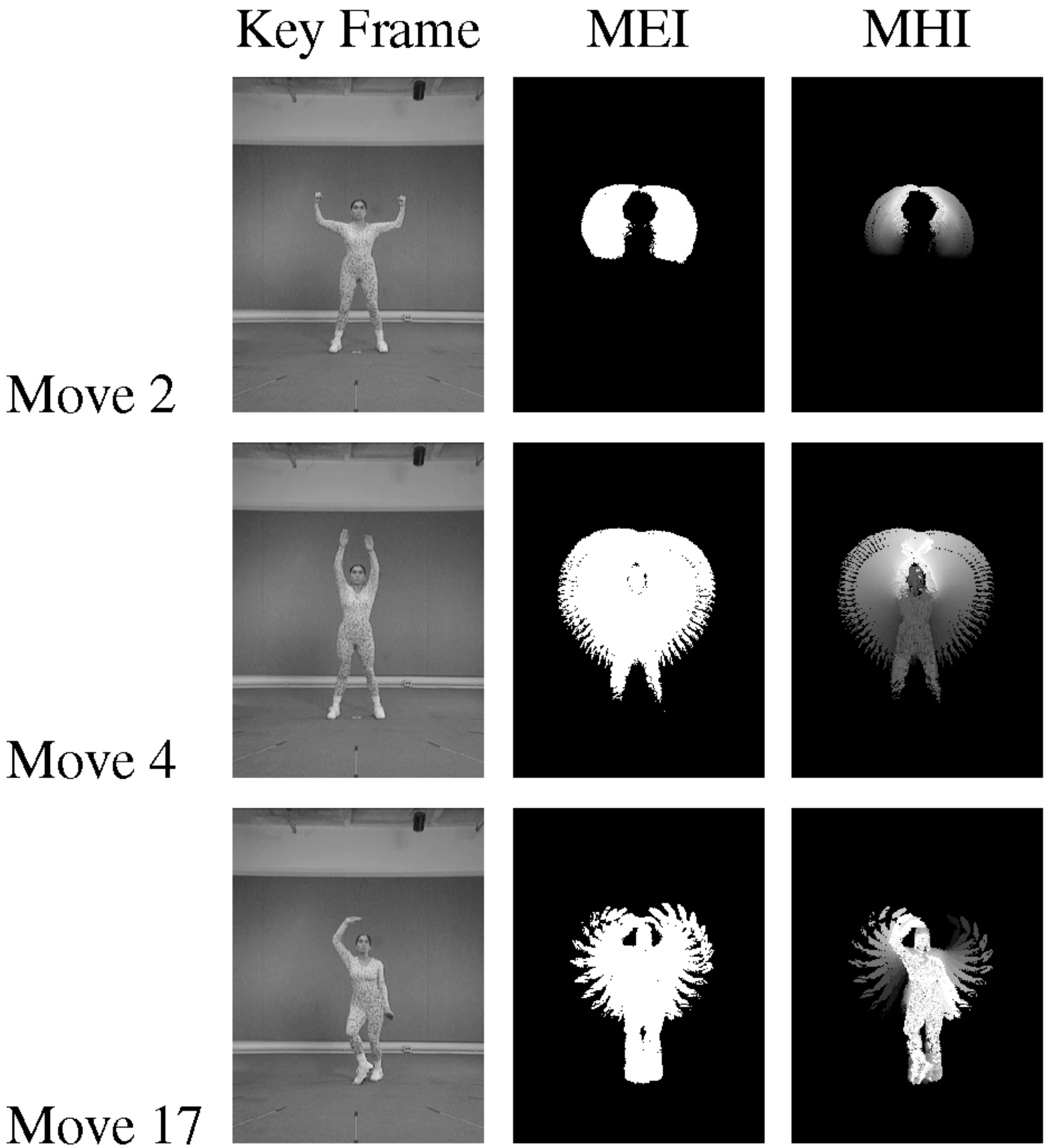


Particular activities often have characteristic appearance patterns. Braids appear at the legs of a walker.



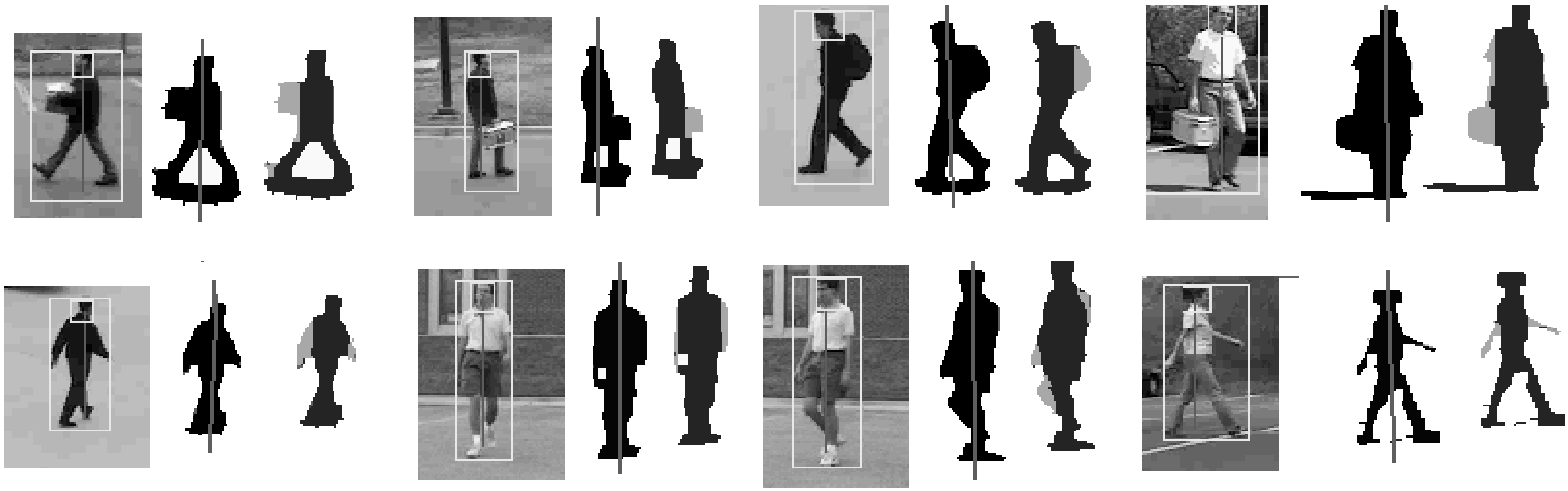
Polana Nelson 93, 94





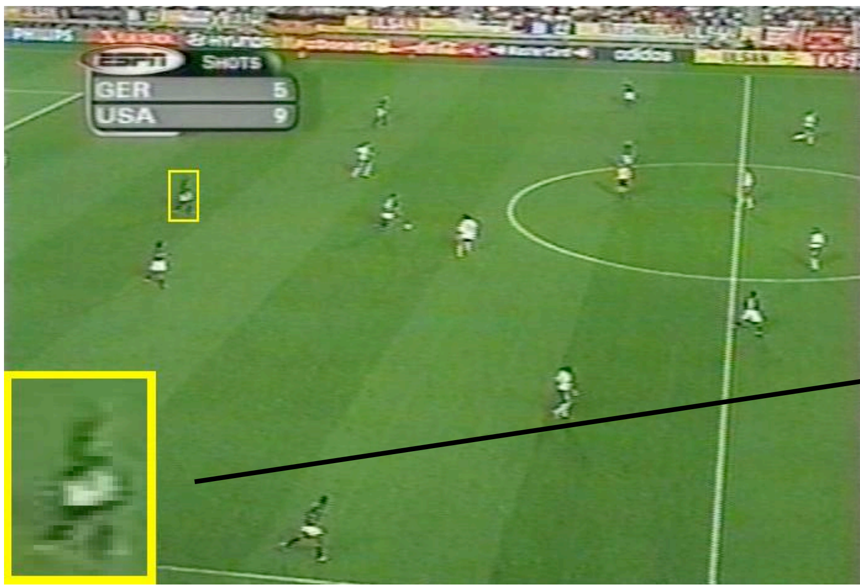
Bobick + Davis, 97

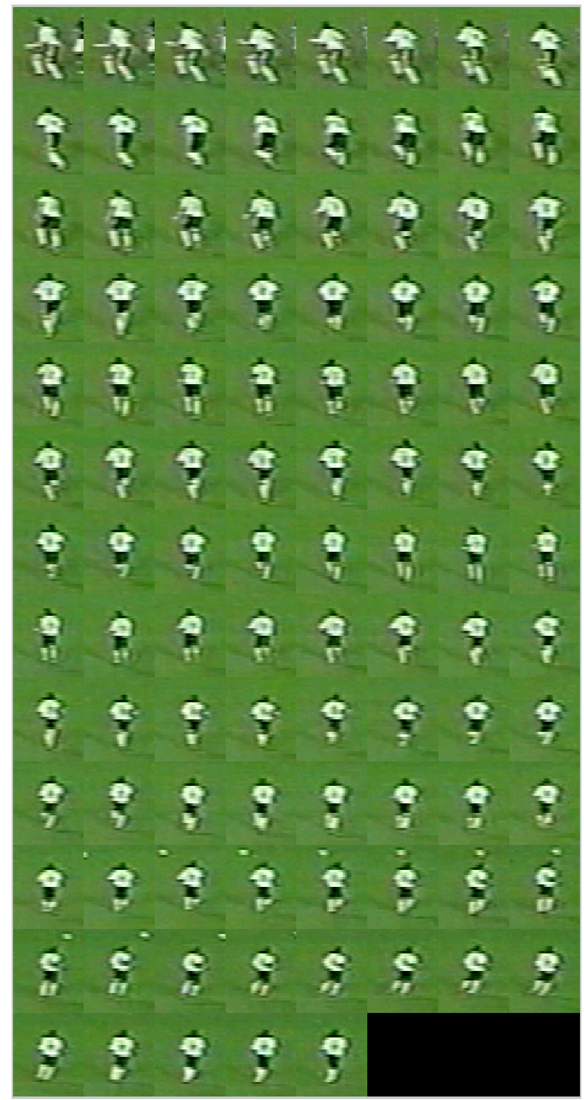
The appearance of a silhouette can show whether a person is carrying something



Haritaoglu, Cutler, Harwood, Davis

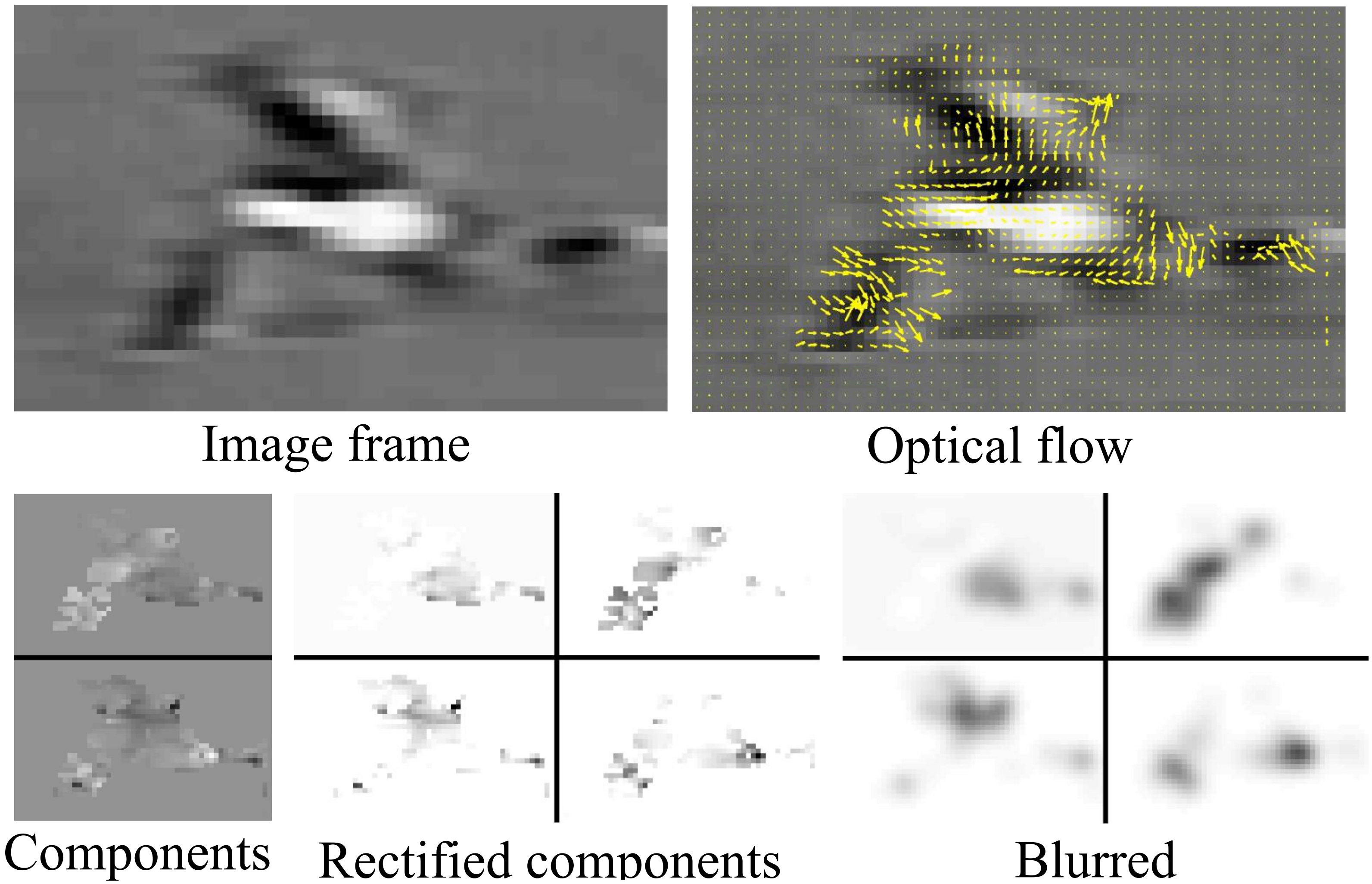
Motion is a powerful cue at low resolution





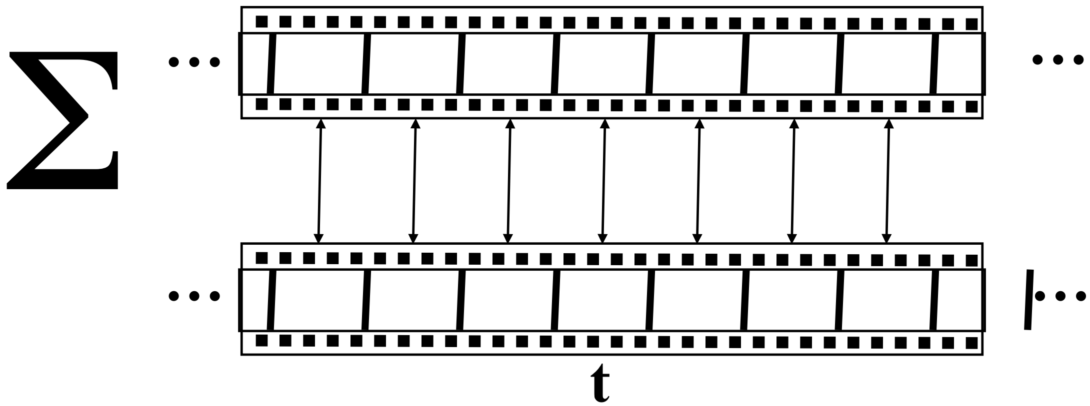
Efros et al 03

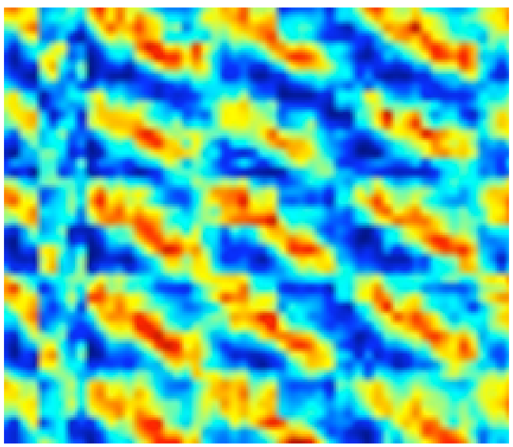
Motion Descriptor



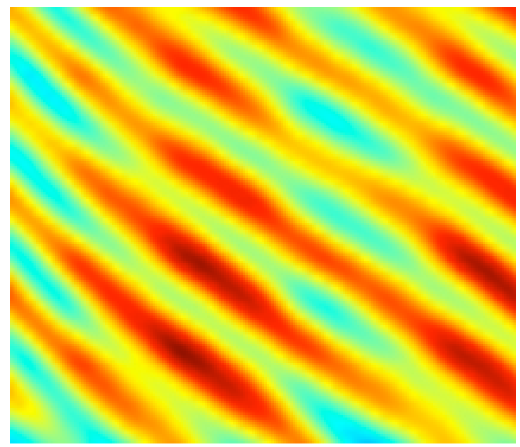
Efros et al 03

Comparing motion descriptors





frame-to-frame
similarity matrix

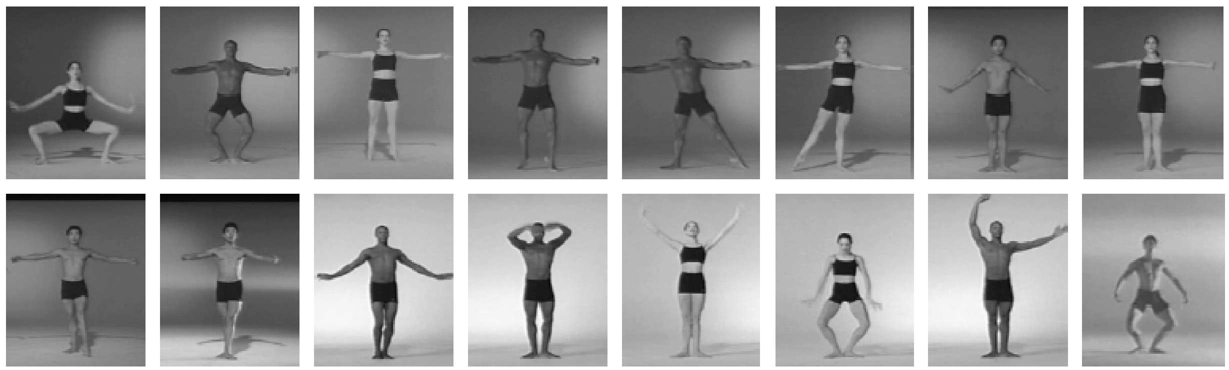


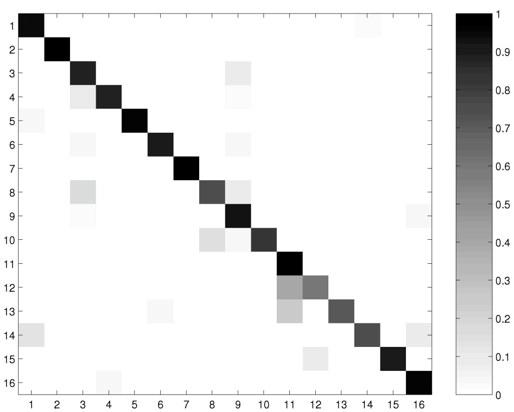
motion-to-motion
similarity matrix

Efros et al 03

Classifying Ballet Actions

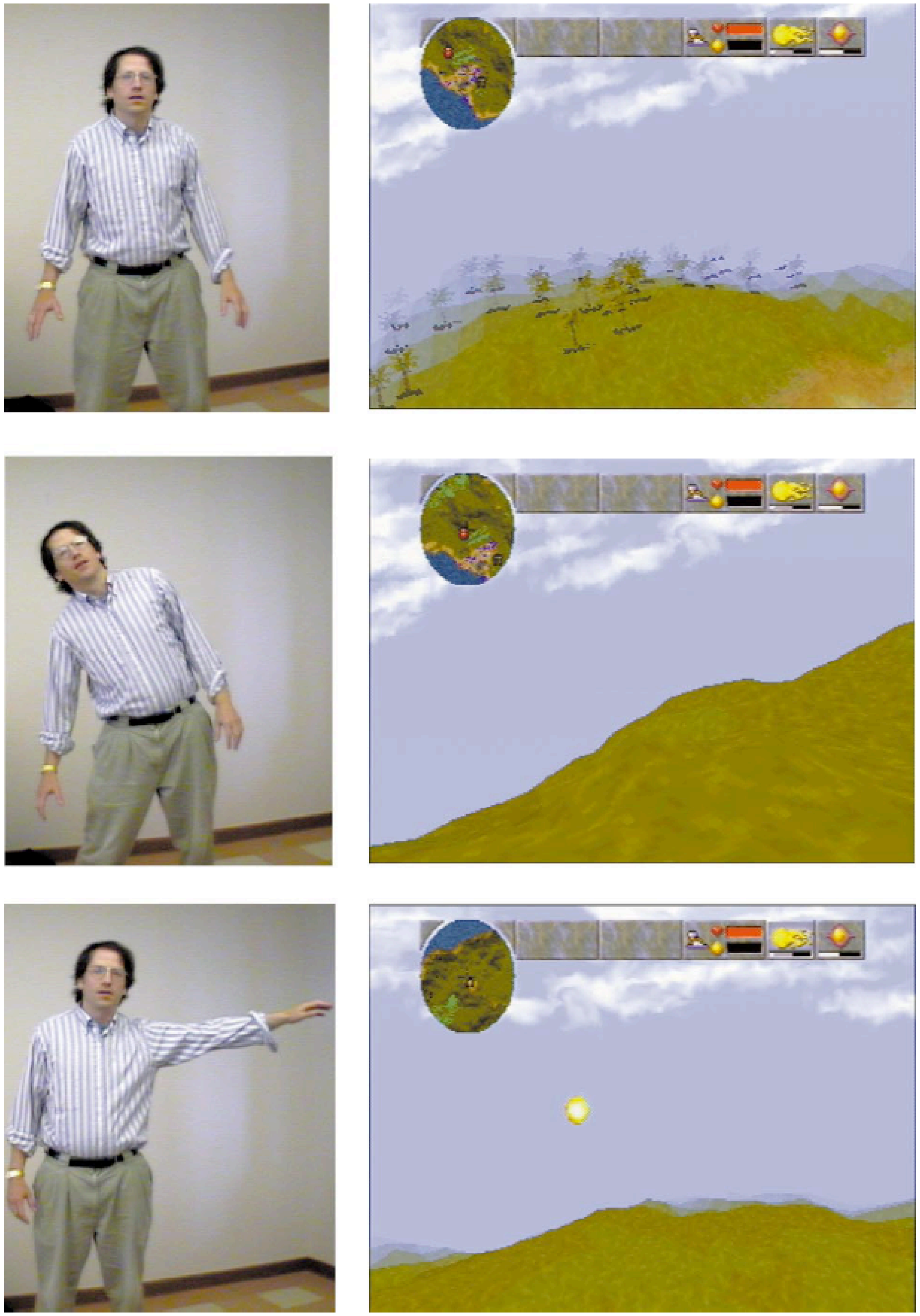
16 Actions. Men used to classify women and vice versa.





Efros et al 03

Applications in Computer Games



Bill Freeman flies a magic carpet.

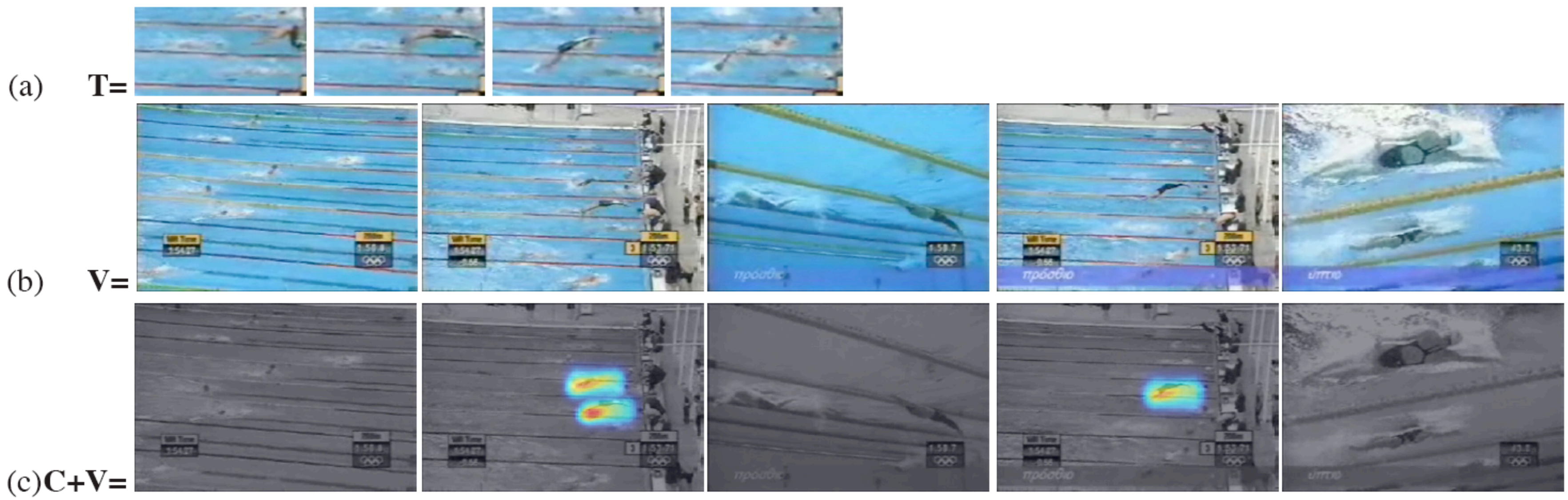
Orientation histograms detect body configuration to control bank, raised arm to fire magic spell.

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



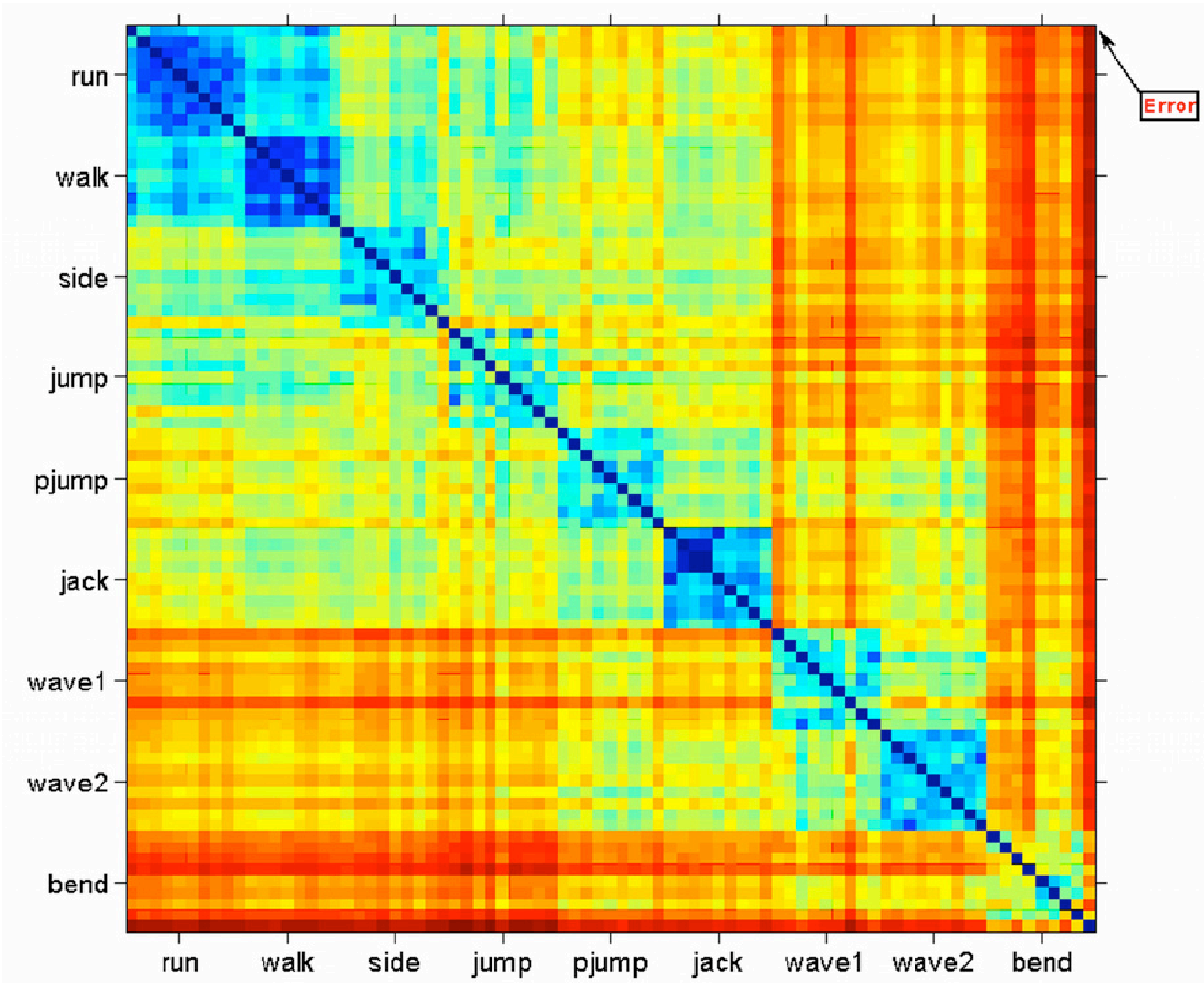
Correlation-like matching can reveal motion matches to queries
Schechtman Irani 05

Spatio-temporal volume is important

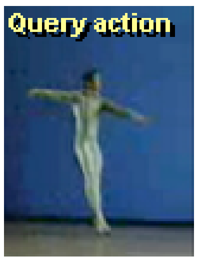
Blank et al 05

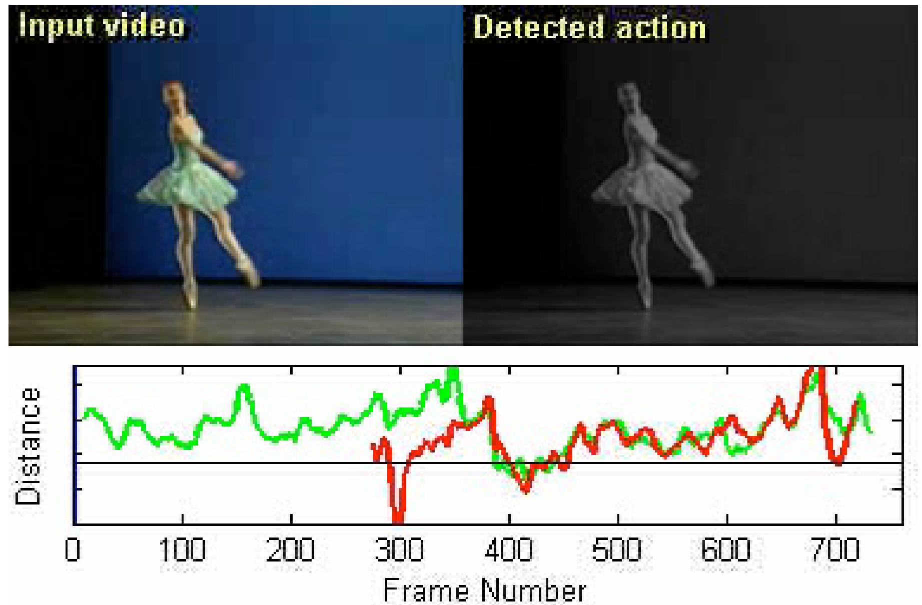
Extract silhouettes
Smooth to get volume
Compute moment representation on s-t volume referred to body
Match

Blank et al 05



Distance matrix between sequences of named motions, obtained by computing distances as above, applying spectral clustering, then reordering. Blue is small, red is large. Generally, similar names have small distances.
Blank et al 05



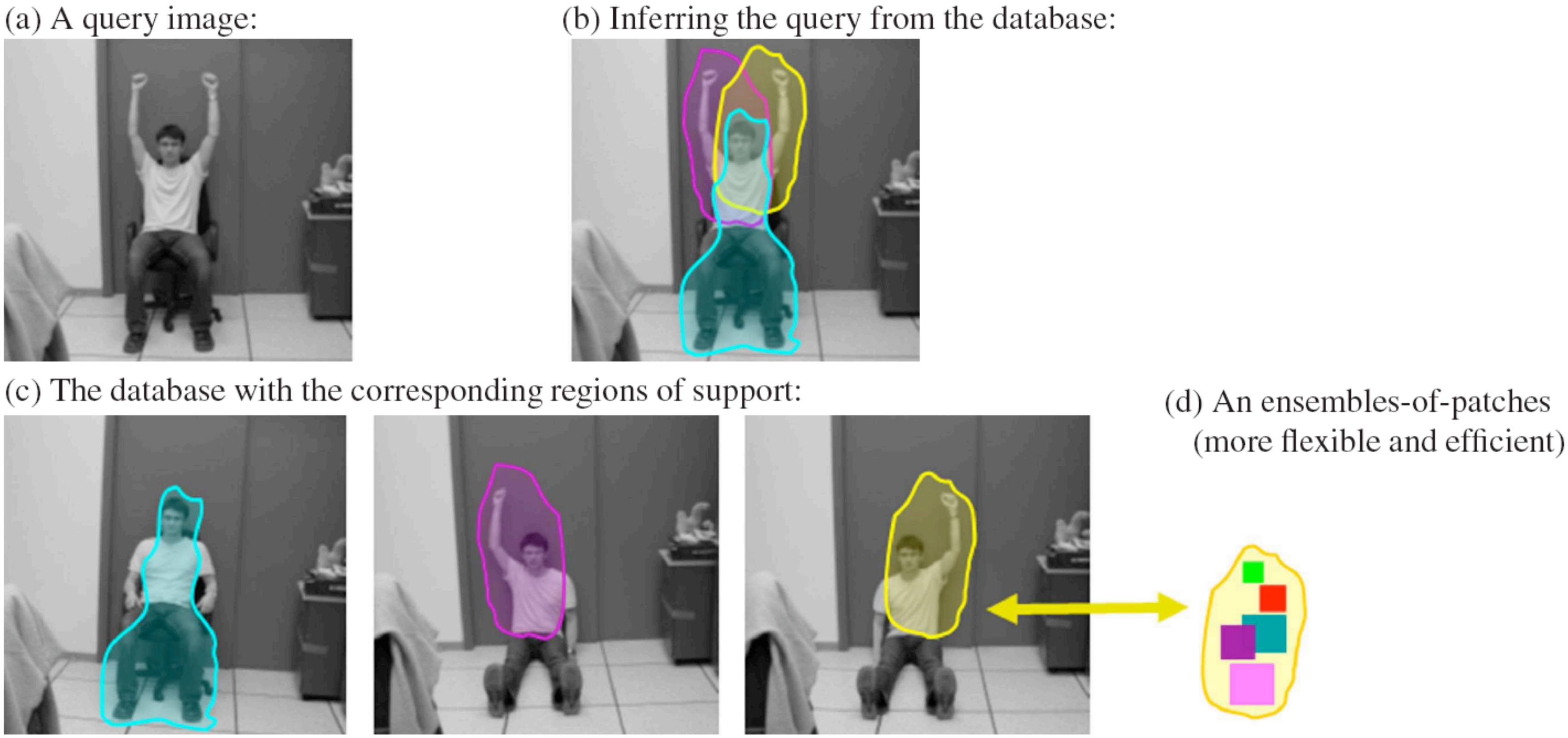


Working in a motion query framework relieves the need for a motion taxonomy. Features computed as before, we now seek sequences with small distances.

Blank et al 05

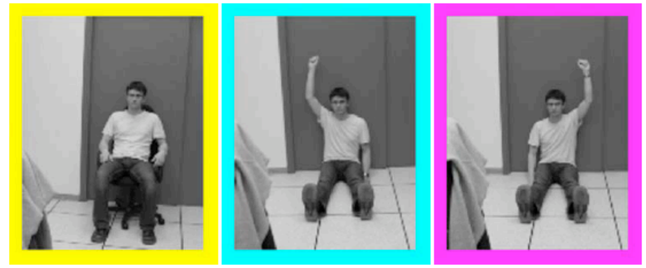
Detecting anomalous activities

- We may have no examples
- Taxonomy is unhelpful, because it won't be complete
  - and may not cover the cases we care about



Anomaly as a failure to be easily encodable
“Normal” motions have been seen before, at least in part.
Boiman+Irani, 05

(a) The database images (3 poses):



(b) Query images:

(c) Red highlights the detected "unfamiliar" image configurations (unexpected poses):

Anomaly as a failure to be easily encodable
Anomalous motions are poorly encoded by example frames
Boiman+Irani, 05

Irani et al 05

Irani et al 05

Strengths

- Can be accurate at discrimination
- Query/Match paradigm can avoid taxonomy issue
    - but requires examples for query
- Strong at low resolutions
- Location may be a very strong cue to activity in some cases

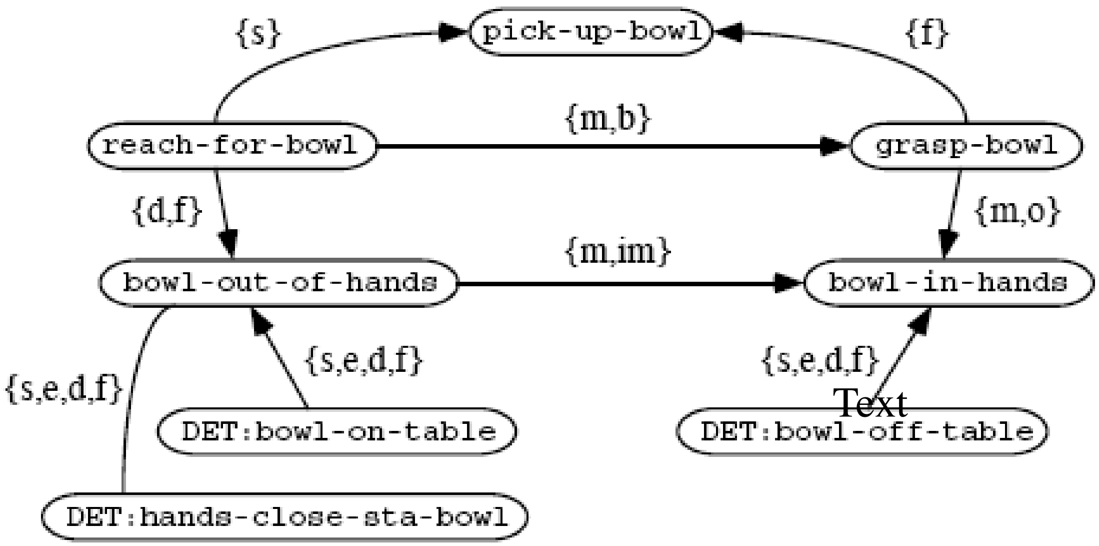
Critiques

- Segmentation is crucial, and harder than it is made to seem
- View variation may present a problem
- Composition presents problems
- Nulls present problems

Logical models of activity

- Logical formulas in primitives
  - spatial relations, motion, support, contact, attachment
  - with noise free transduction (Siskind, 92, 95)
  - analogous with HMM's (Siskind+Morris, 96)
  - Attractions
    - may be quite a broad class of representation
    - very general activities (visit to the ATM) might be of this type
  - Unproven

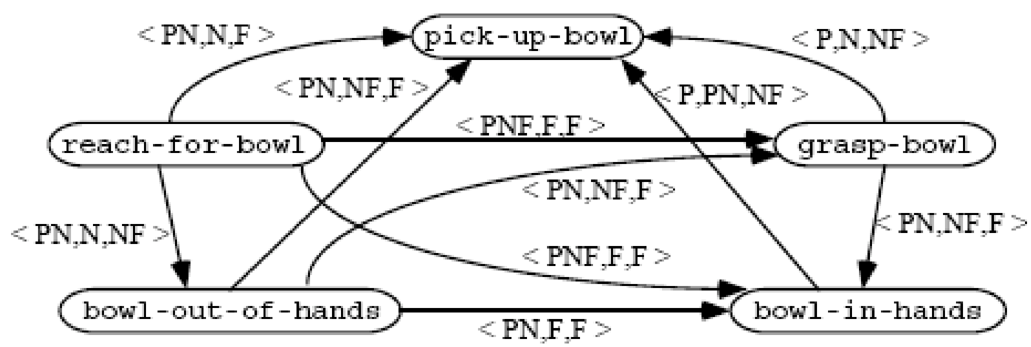
Temporal Calculus



Start with an interval algebra structure for an activity with detectors, relations between events such as start, finish, etc.

Allow relations to take form Past, Now, Future

Infer relations from detector responses
Note dynamic representation does not represent "speed"
Pinhanez Bobick 98



Sign Language as a Problem Domain

- Advantages
  - large data sets can be found
  - in principle, right answer can be known
  - cooperative subjects? and rich problem
  - socially useful, perhaps
- State of the art quite advanced for small vocab, controlled views
  - otherwise rather open

ASL Rough SOA

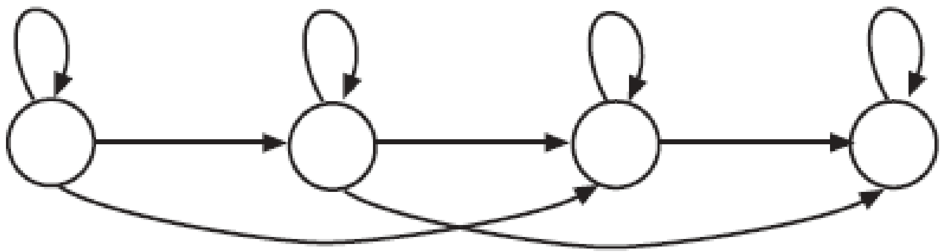
- Recognition rates
  - 90% on 40 signs (Starner+Pentland 95)_
  - 262 isolated signs (Grobel+Assan)
  - continuous German 97 signs (Bauer+Heinz)
  - 90’s on 53 words (Vogler+Metaxas)
  - 90s on 131 Korean using datagloves (Kim et al)
  - etc. see printed text
- But there is no continuous transcription system for large vocab
  - nothing resembling modern speech systems
  - nothing resembling modern MT systems

HMM'S - core ideas

- Finite state machine maintains hidden state; there are stochastic state transitions at known time steps
- At each time step, a measurement is emitted with probability conditioned on the hidden state
- Inference
  - Dynamic programming
  - beam search
- Learning
  - EM

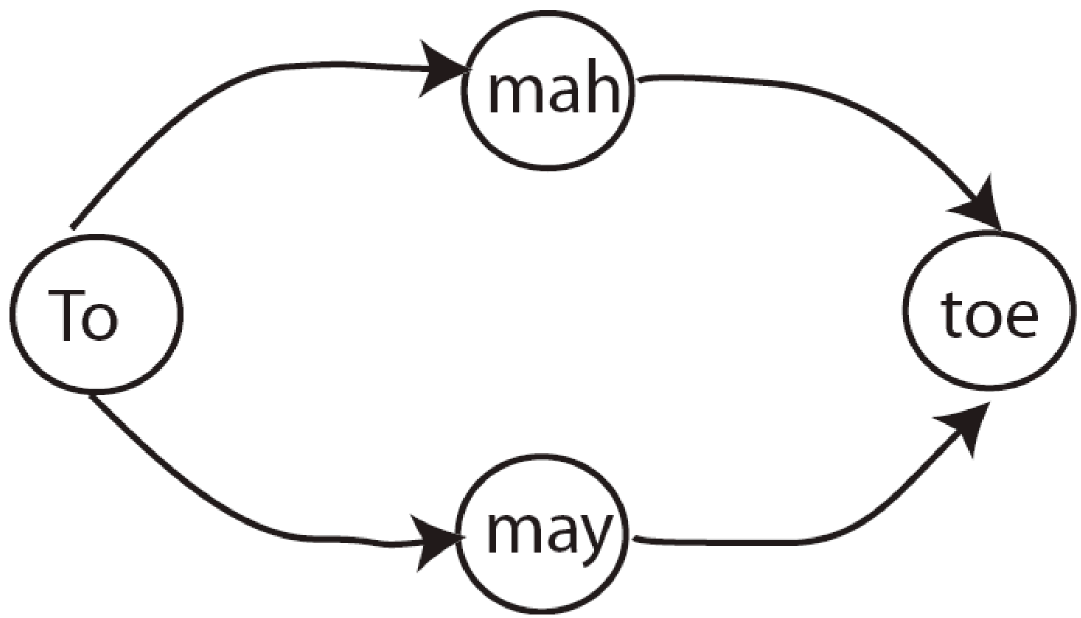
HMM's in speech understanding

- A string of words is modelled at several levels, e.g.
  - trigram word models
  - pronunciation dictionary per word
  - context dependence of phonemes
  - acoustic model of context dependent phones
- Each is an FSM
  - these are composed
  - missing parameters can be supplied in a variety of ways
    - count in text (trigrams)
    - pronunciation dictionary
    - learned from data (acoustics)
- Result: enormous state space model with relatively few pars to learn

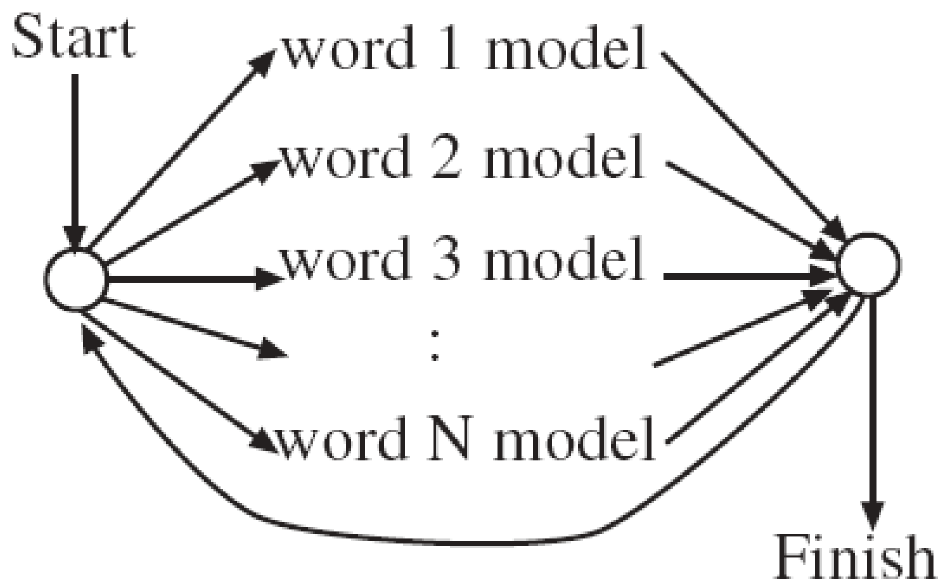


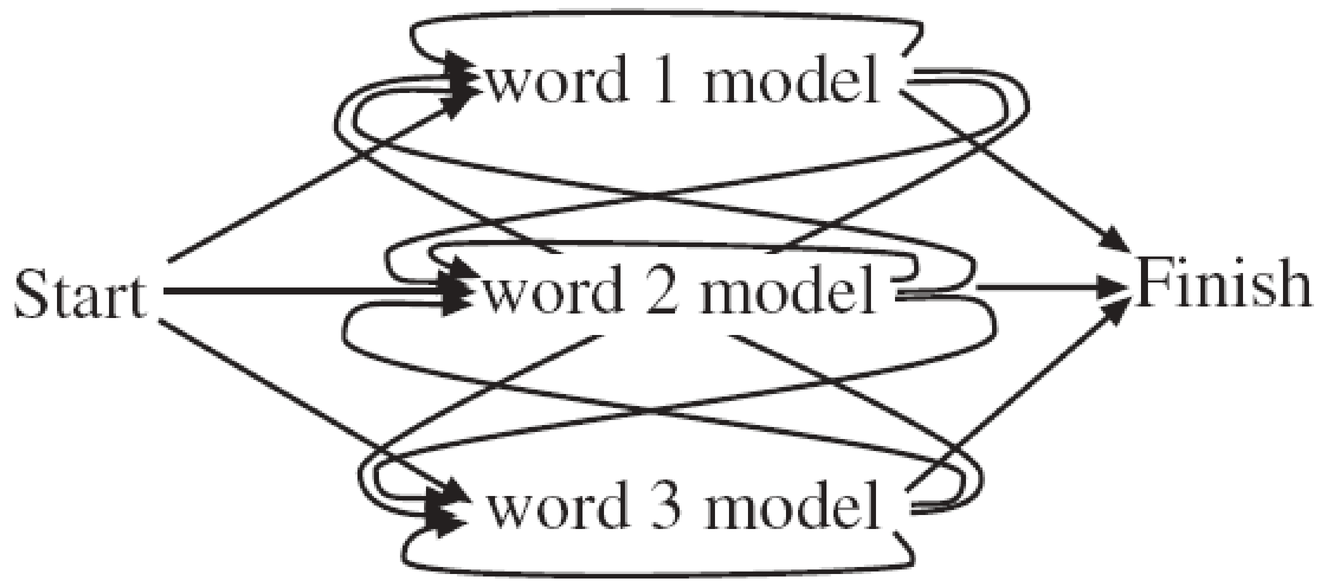
Phoneme model

Pronounciation model



Language models





HMM's in activity recognition

- Gesture
  - No pronunciation dictionaries, trigram models, etc. available
    - very difficult to learn with large state spaces
      - various hacks
- Sign language
  - No pronunciation dictionaries, trigram models, etc. available
    - but (perhaps) lots of data
      - no pooling phone data over examples
      - data essentially discriminative
- Surveillance
  - same story

Figure 5: Human area extraction
a)original, b)background, c)extracted

Table 3: Recognition rate (%) (experiment 2)

	Training data player			
Test data player	A	B	A+B	C
C	61.2	66.8	70.8	100.0

Symbol sequence	60 61 61 62 62 62 63 63 64 64 65 66 66 66 67 68 68 69 69 70 70 70 71 71

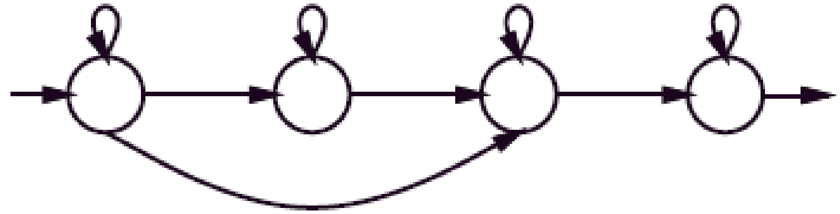
Figure 6: Example of extracted tennis action and symbol sequence(forehand volley).
(Underlined symbol is assigned to frame of above figure.)

Activity recognition by HMM's used discriminatively (choose the HMM with the highest likelihood), silhouettes for tennis activities.

Yamato et al 1992

Table 1: ASL Vocabulary Used

part of speech	*vocabulary*
pronoun	I you he we you(pl) they
verb	want like lose dontwant dont
	love pack hit loan
noun	box car book table paper pa
	bicycle bottle can wr
	umbrella coat pencil
	magazine fish mouse
adjective	red brown black gray

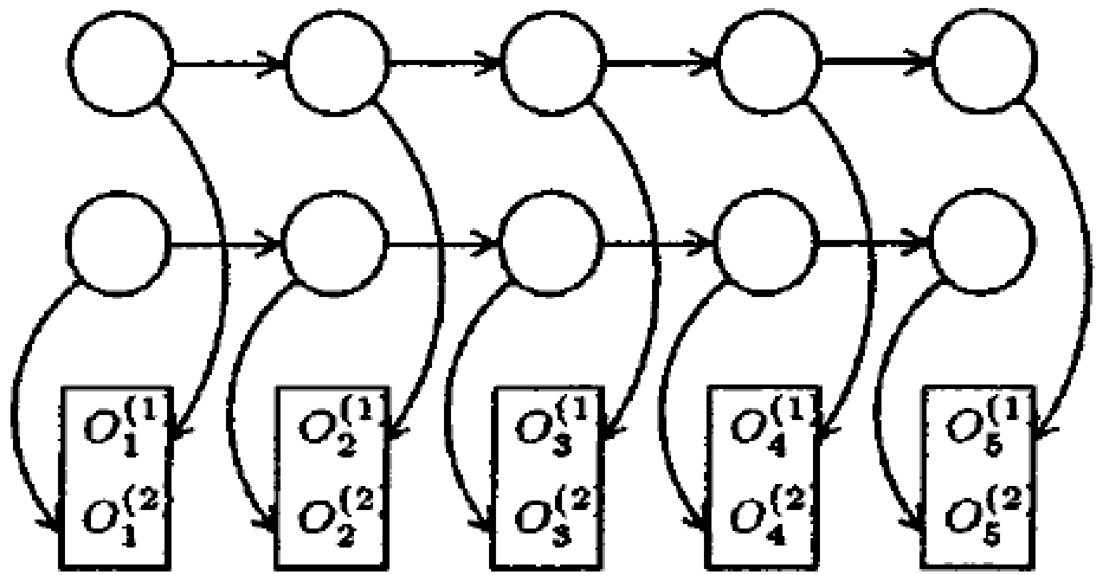


Starner Pentland 95

	on training	*on indep. test set*
grammar	99.5%	99.2%
no gram.	92.0% (97% corr.) (D=9, S=67, I=121, N=2470)	91.3% (97% corr.) (D=1, S=16, I=26, N=495)

Variant HMM's

- Goal:
  - reduce learning complexity of transition probability matrix
- Methods:
  - variant architectures
  - variant training algorithms

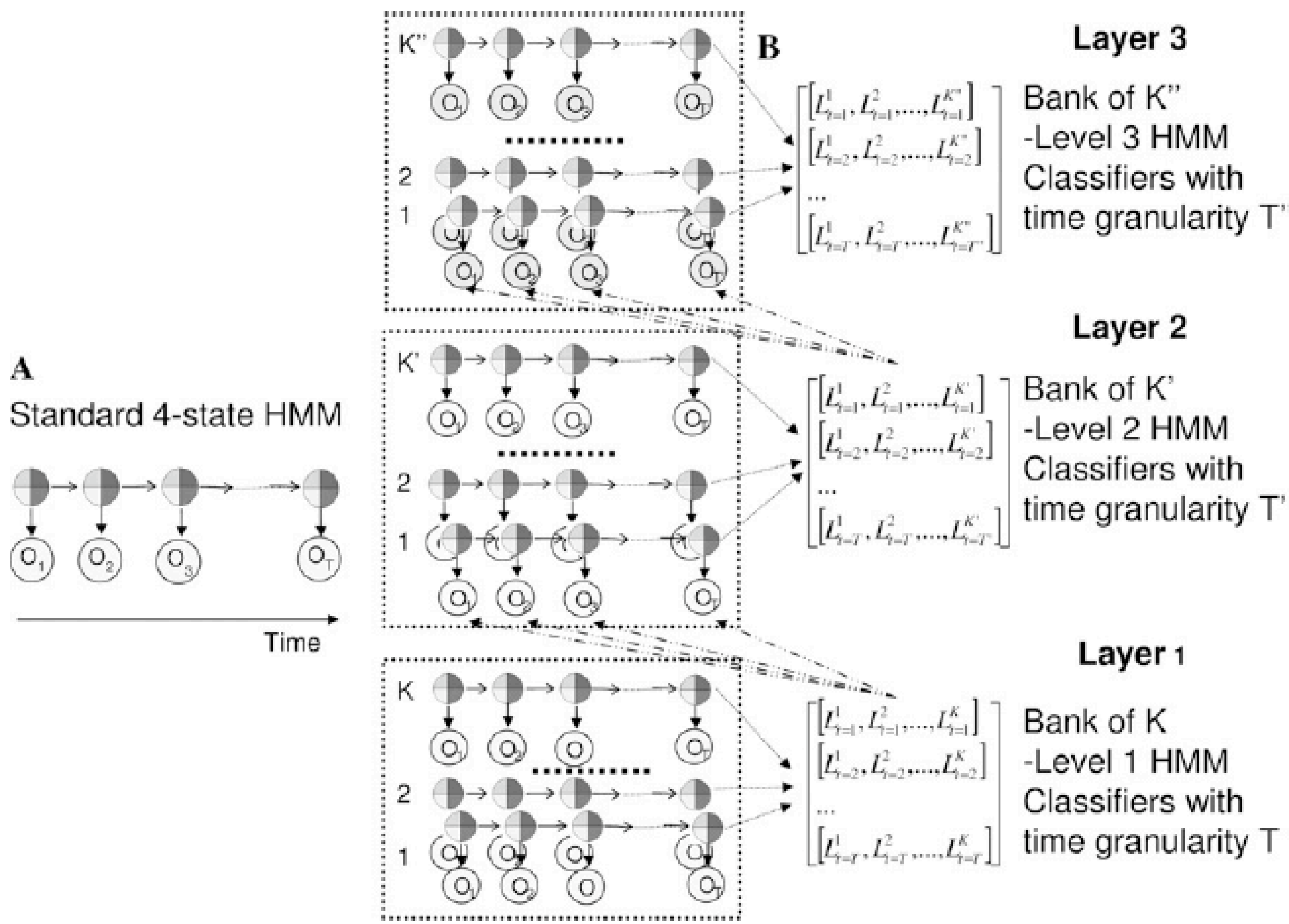


Factorial HMM's Ghahramani+Jordan 97

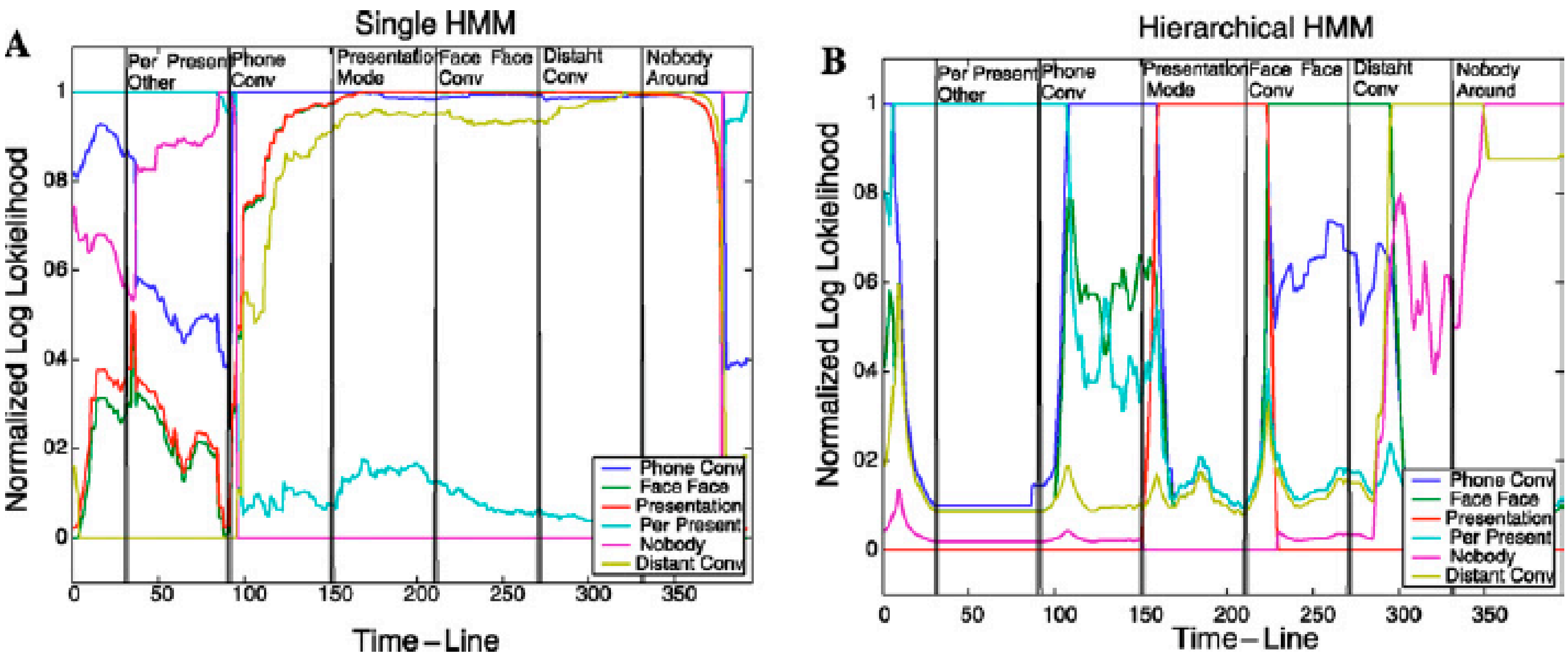
Layered HMM

Note that, in principle, one could build a single HMM,

but the quantization process reduces number of parameters to learn



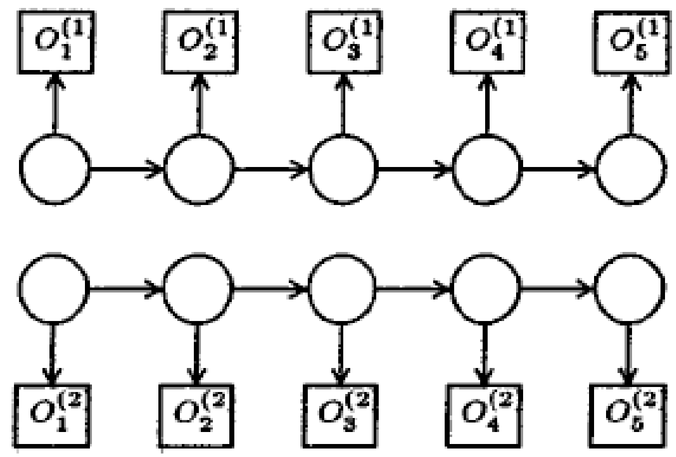
Oliver et al 04



Oliver et al 04

Parallel HMM's

- ASL words
  - strong hand produces one sequence, weak hand another (or nothing)
- Possible squaring of the space of phoneme models
- Use
  - phoneme transcription of words
  - one HMM for each hand
  - require inferred path to be consistent
-



Parallel HMM's Vogler and Metaxas 01

Parallel HMM's

TABLE 2
Regular HMMs: Results of the Recognition Experiments

Level	Accuracy	Details
sentence	80.81%	$H = 80^a$, $S = 19^b$, $N = 99^c$
sign	93.27%	$H = 294$, $D = 3^d$, $S = 15$, $I = 3^e$, $N = 312$

Note. 80.81% of the sentences were recognized correctly, and 93.27% of the signs were recognized correctly.

[a] H denotes the number of correctly recognized sentences or signs.

[b] S denotes the number of substitution errors.

[c] N denotes the total number of signs or sentences in the test set.

[d] D denotes the number of deletion errors.

[e] I denotes the number of insertion errors.

TABLE 3
PaHMMs: Results of the Recognition Experiments, with Merging of the Token Probabilities at the Phoneme Level

Level	Accuracy	Details
sentence	84.85%	$H = 84$, $S = 15$, $N = 99$
sign	94.23%	$H = 297$, $D = 3$, $S = 12$, $I = 3$, $N = 312$

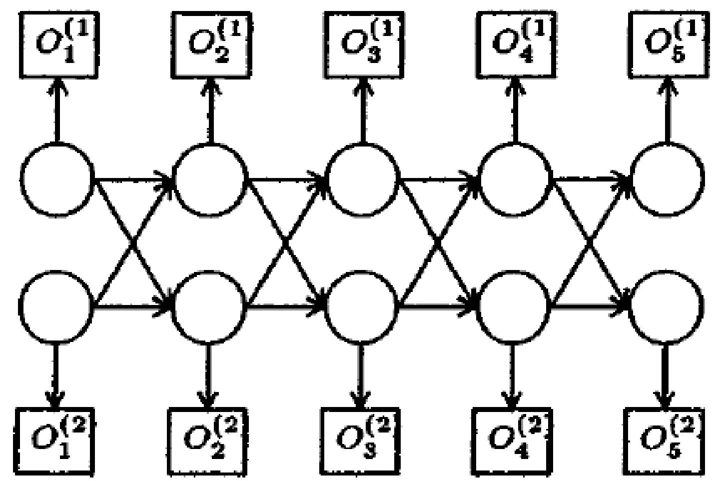
Note. See Table 2 for an explanation of the terminology.

- Small improvement on HMM's using
  - 3D arm configuration data, 3D tracked visual data

Coupled HMM's

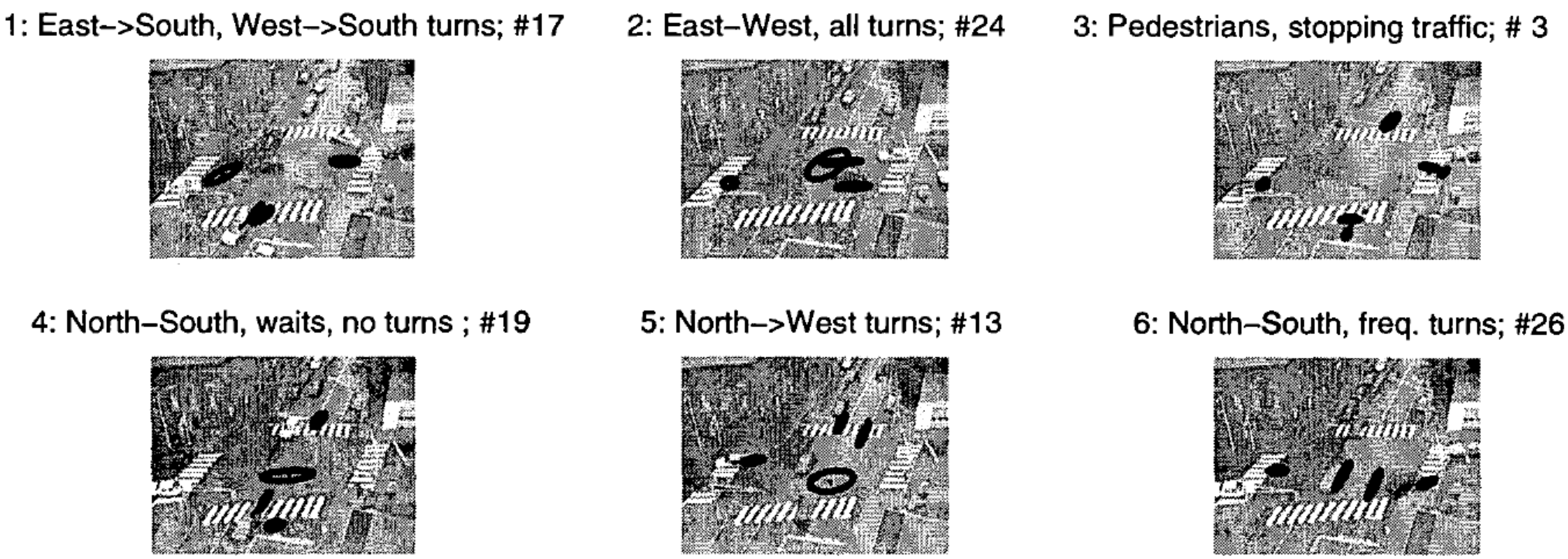
- Observations in two classes, states split, state transition matrix coupled, variant estimation algorithm
- Improvement in discriminative results for very small state models, three gestures

	Single HMMs	Linked HMMs	Coupled HMMs
accuracy	69.2%	36.5%*	**94.2%**
# params	25+30+180	27+18+54	36+18+54



Coupled HMM - Brand et al. 97

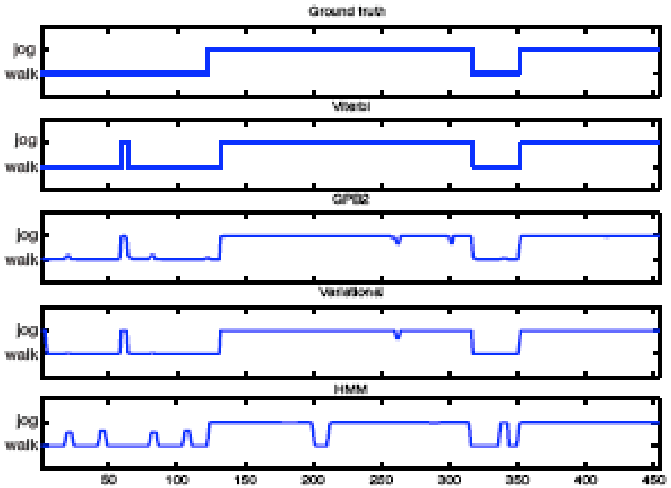
Finite state models of activity



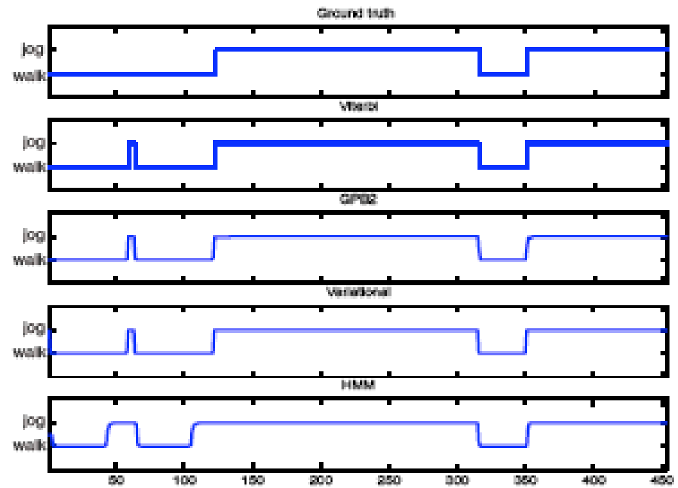
Variant generalized HMM with variant learning method, 6 states
Kettnaker Brand 99

Switching Linear Dynamical Systems

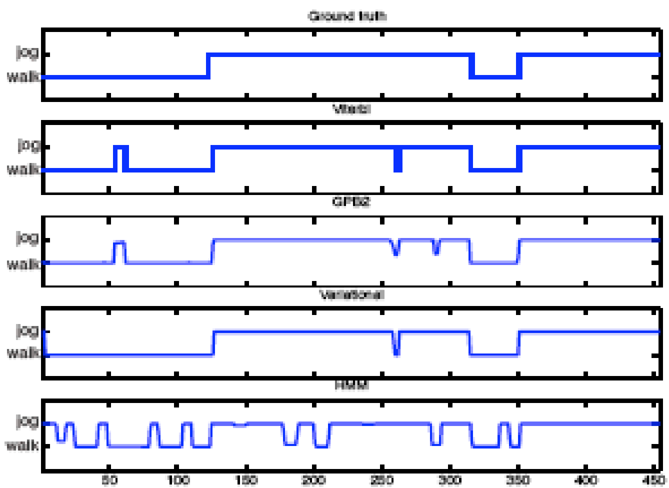
- Linear dynamical system
  - consists of state vector, linear state transition process, linear emission process
  - fair model for some forms of activity, at least at short timescales
    - handwriting
    - dance (Li et al 02)
- Switching
  - discrete state transition process chooses LDS
- In vision
  - Bregler, 97; Pavlovic Rehg 2000



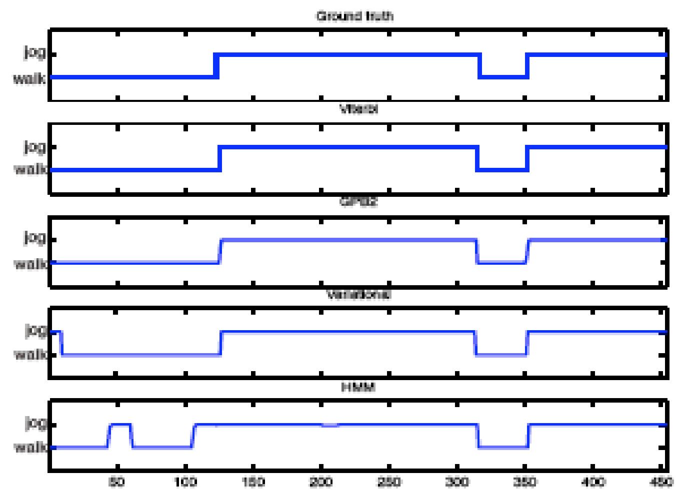
(a) One switching state, second order SLDS.



(b) Four switching state models, first order SLDS.



(c) Two switching state models, second order SLDS.



(d) Four switching state models, second order SLDS.

- Pavlovic Rehg 2000

Discriminative models of activity

- Matching inferred body to labelled 3D configuration data

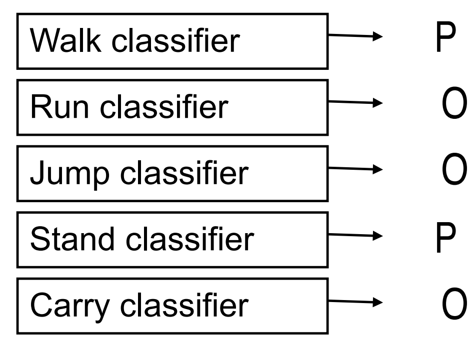
Synthesis with off-line control

- Annotate motions
  - using a classifier and on-line learning
  - efficient human-in-the loop training
- Produce a sequence that meets annotation demands
  - a form of dynamic programming

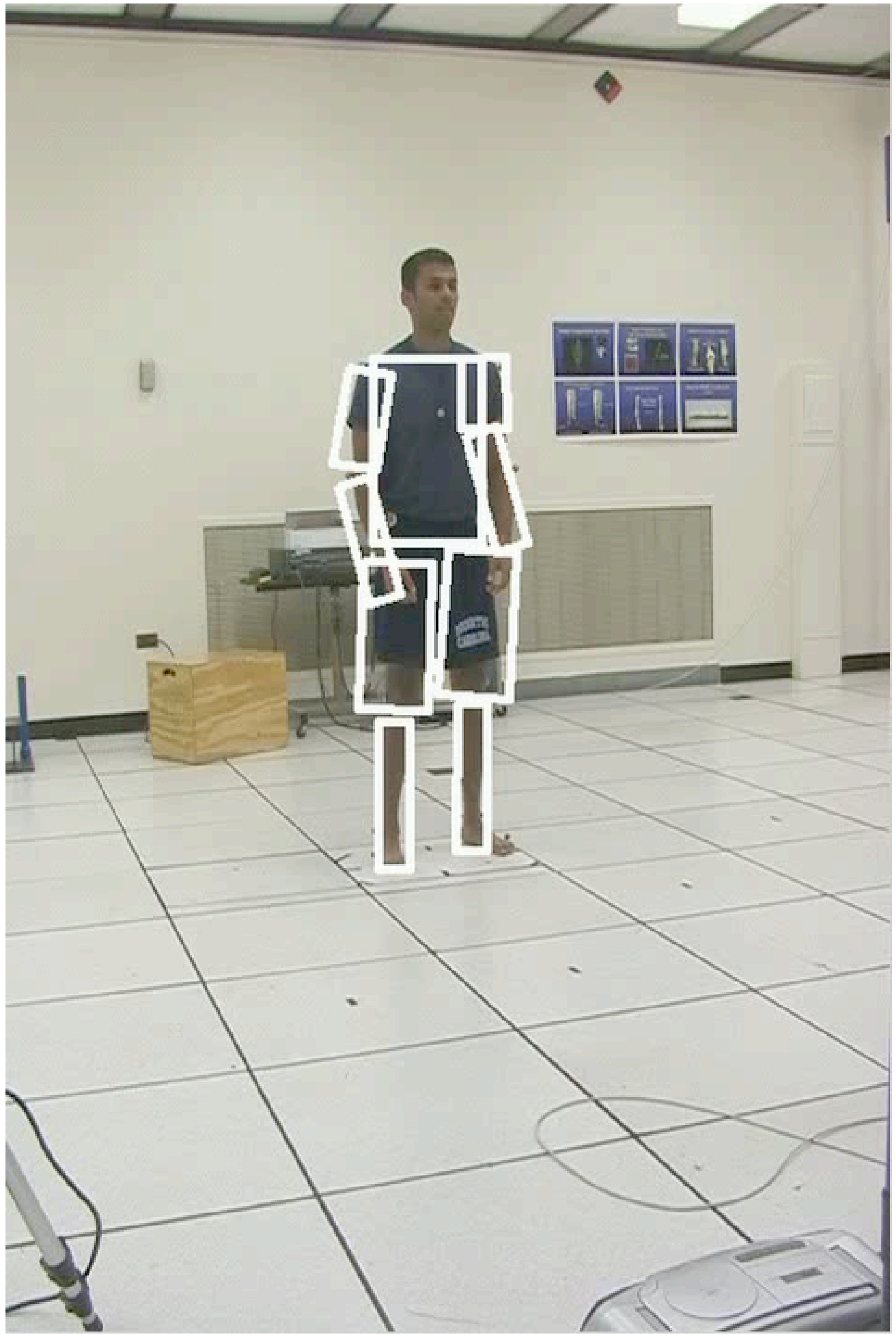
Annotation - desirable features

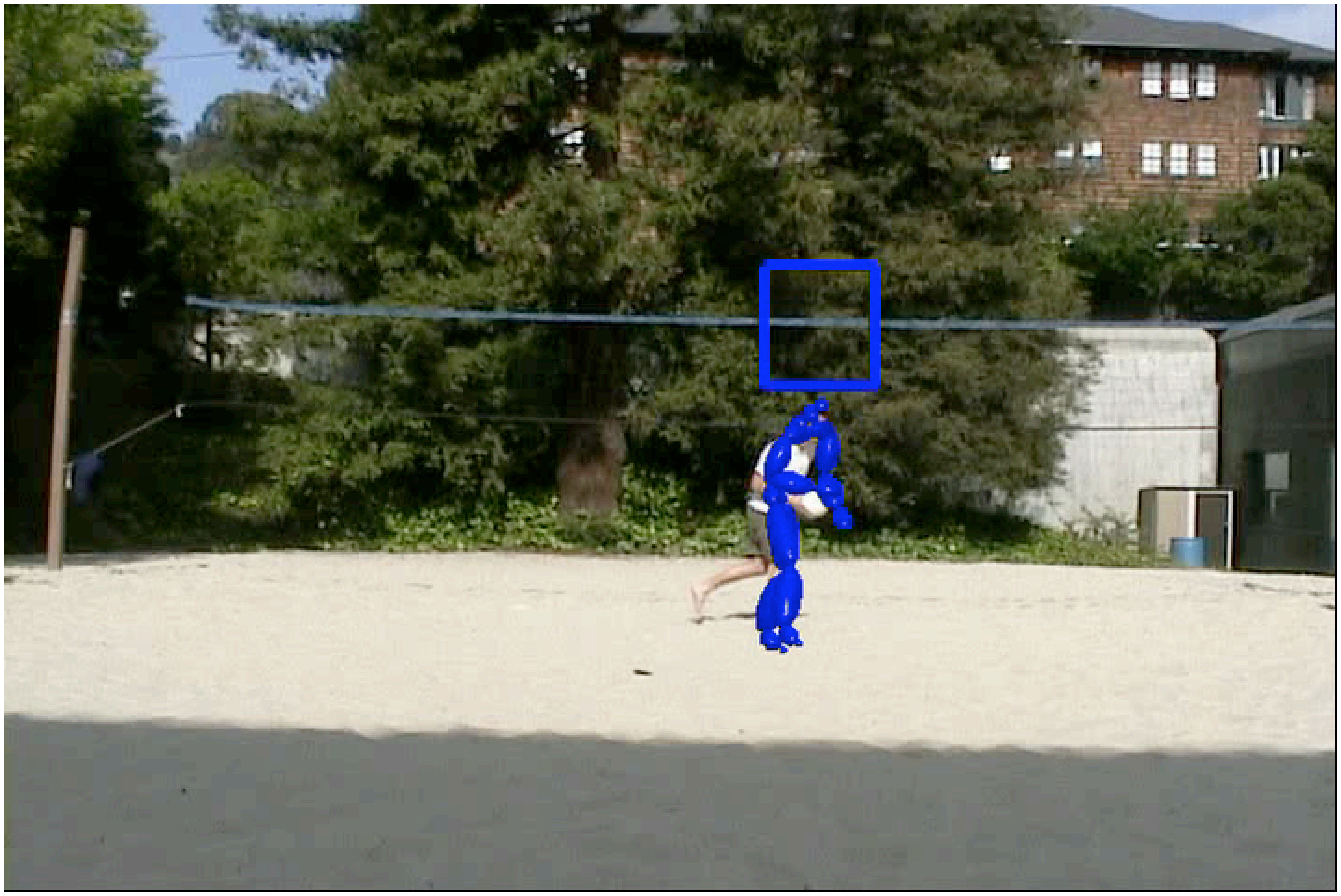
- Composability
  - run and wave;
- Comprehensive but not canonical vocabulary
  - because we don’t know a canonical vocabulary
- Speed and efficiency
  - because we don’t know a canonical vocab.

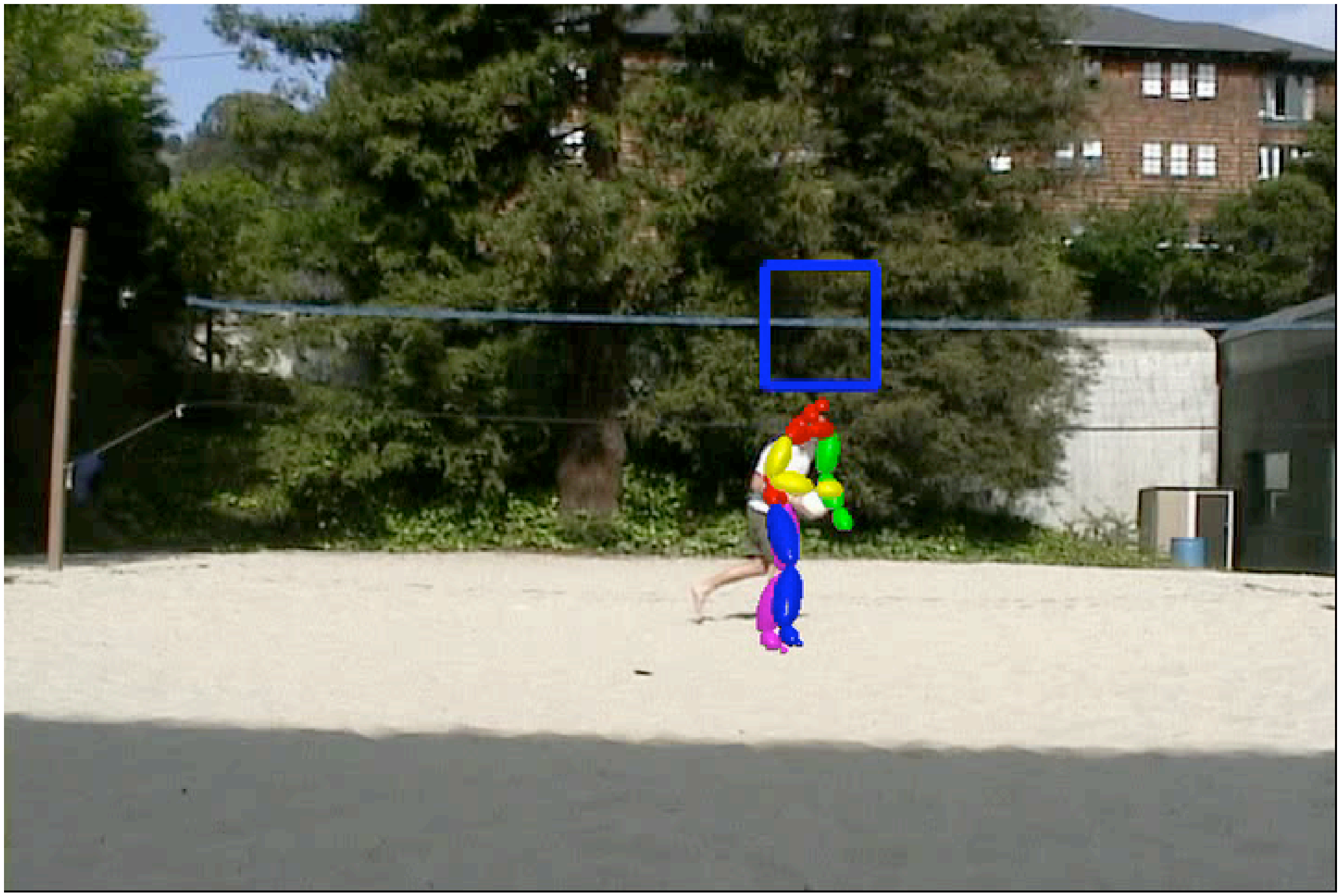
- Can do this with one classifier per vocabulary item
  - use an SVM applied to joint angles
  - form of on-line learning with human in the loop
  - works startlingly well (in practice 13 bits)



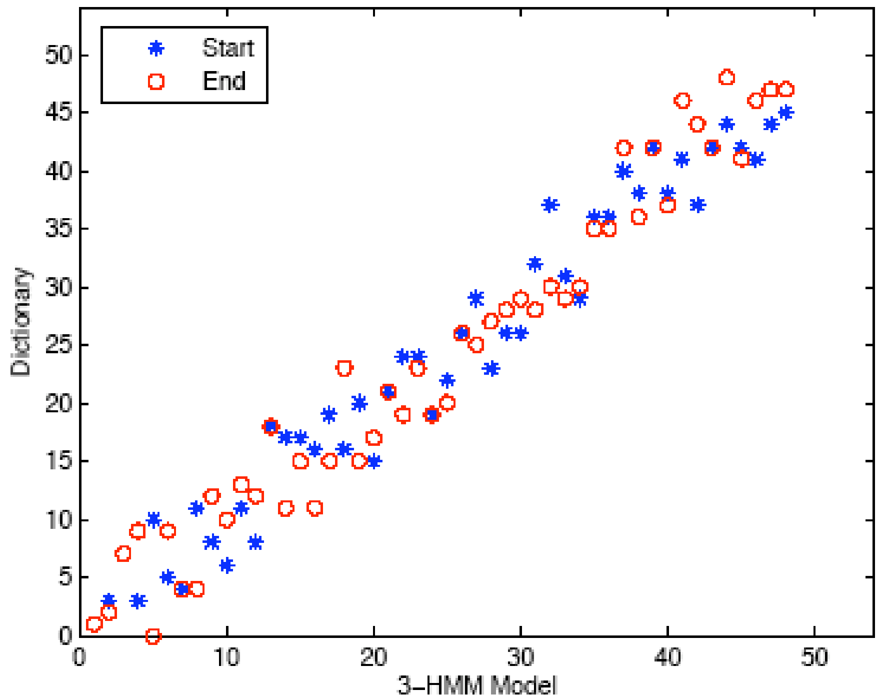
Arikan+Forsyth+O’Brien 03





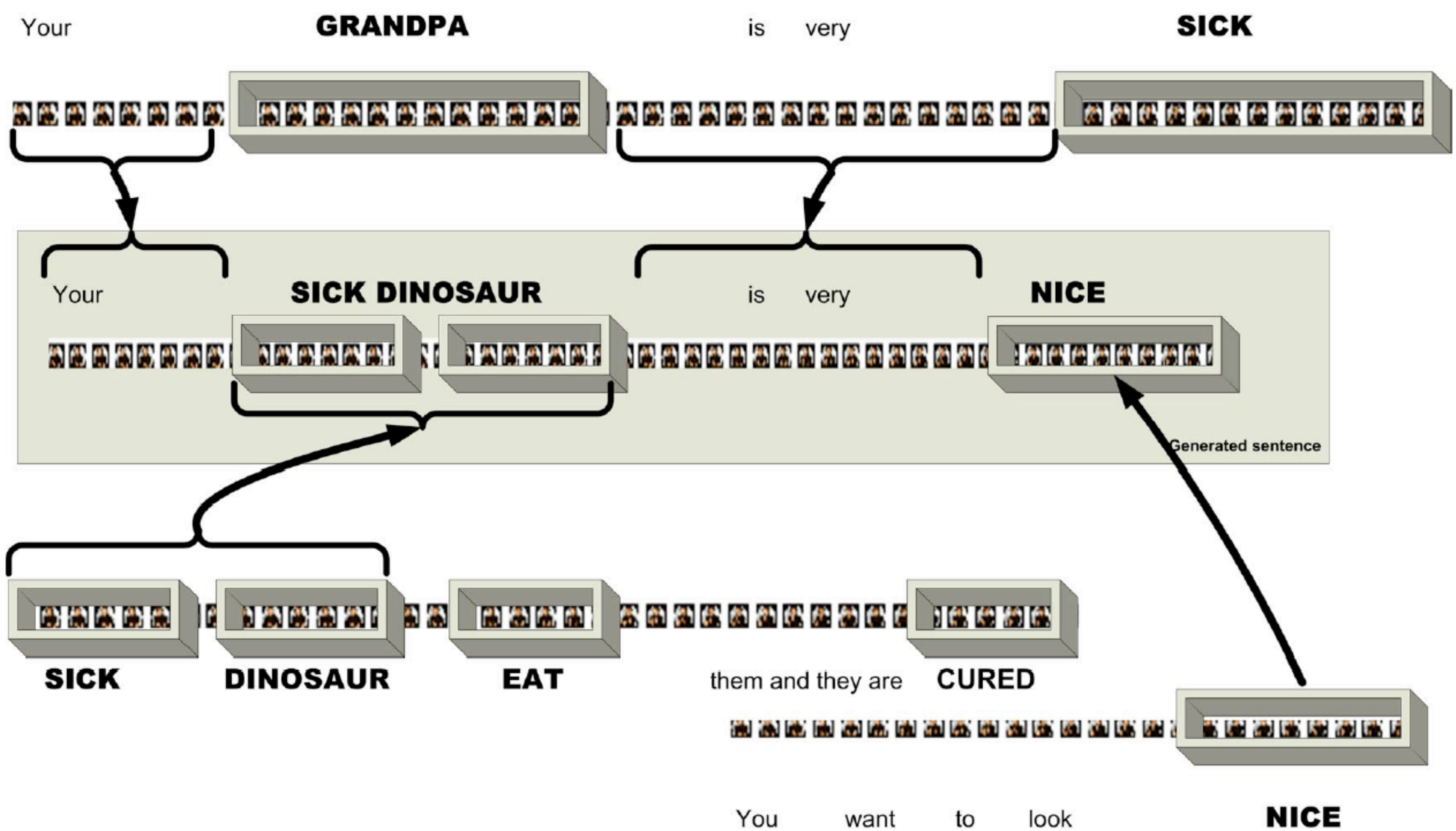


Examples of words (subtitles for land before time III: journey to the mists) at signal res



Find word boundaries by voting using
3 distinct generative models

Farhadi+Forsyth 06

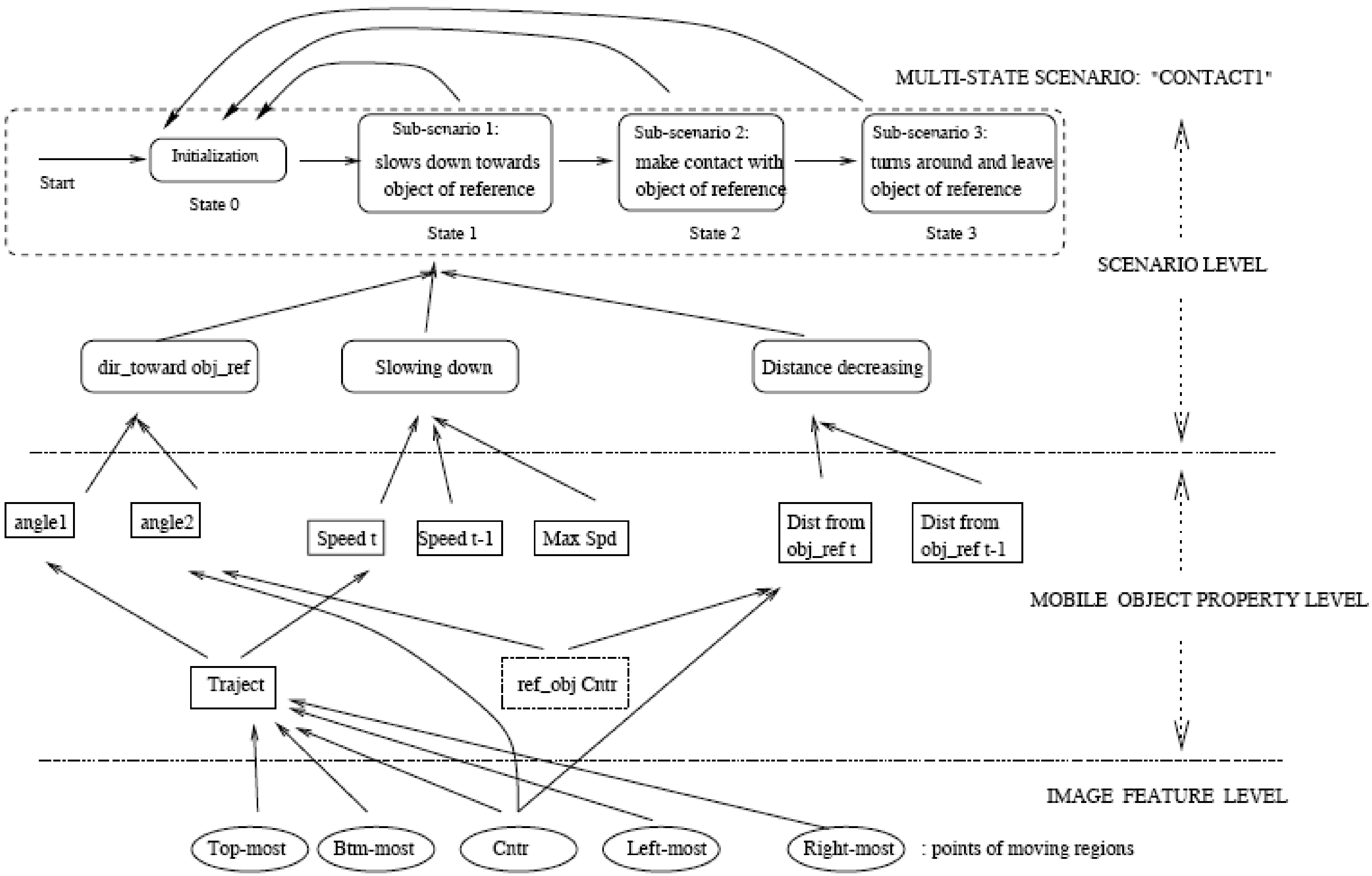


Spot words using multiclass logistic regression trained on small blocks of frames; regime involves base and derived forms of words to control dimension problems

Farhadi+Forsyth 06

Authored representations

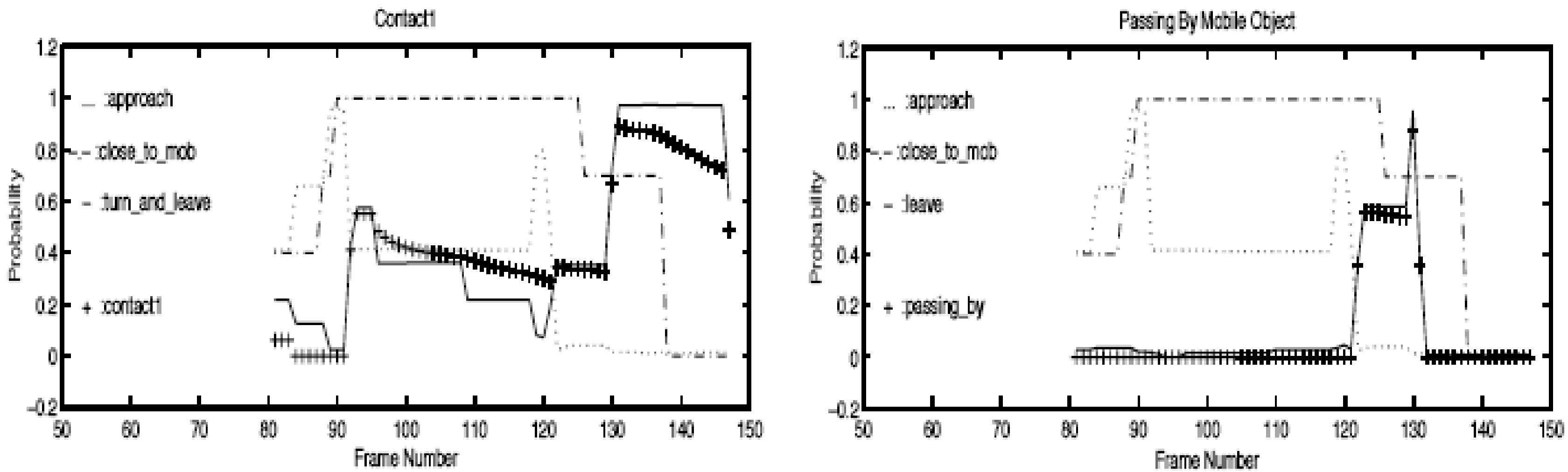
- Build a system of representation that allows authoring a query
  - typically, an FSA or RE
  - but could be a query video as above?



Explicit models of activity - Hongeng et al 00

frame A.63 frame A.95 frame A.111 frame A.157

(a) Detection and tracking of moving regions for scenario "CONTACT1".



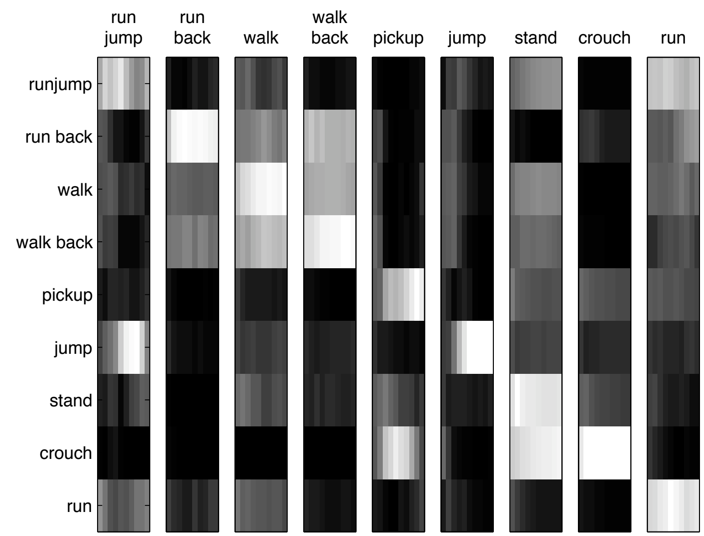
I) CONTACT1 II) PASSING_BY

(b) Recognition results of two competing activities.

Explicit models of activity - Hongeng et al 00

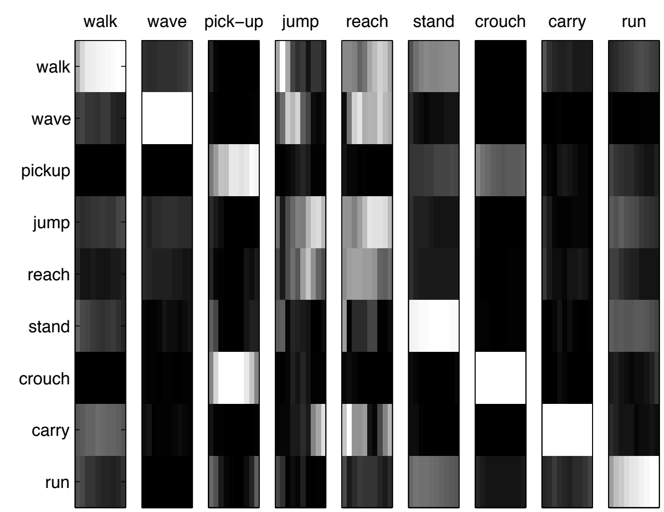
Composite representations

Legs - CCM

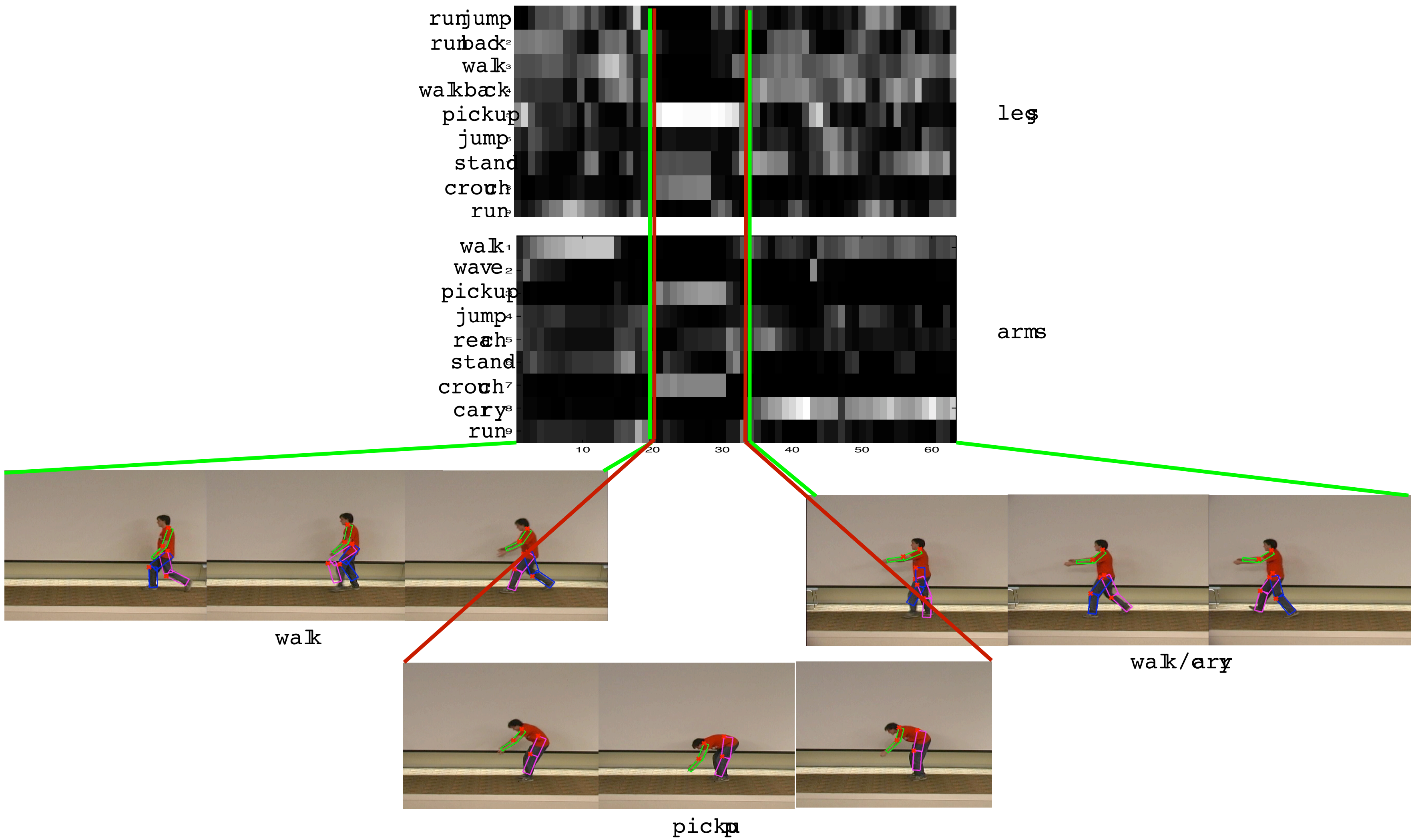


- Build HMM for each of arm, leg
  - for each of a set of labels
- Link states with similar emissions
  - Large composite model
  - Blocks of states csp to activities
- Now search with FSA
  - alphabet
    - composites
      - leg-run-arm-wave
  - P(endstate| measurements)

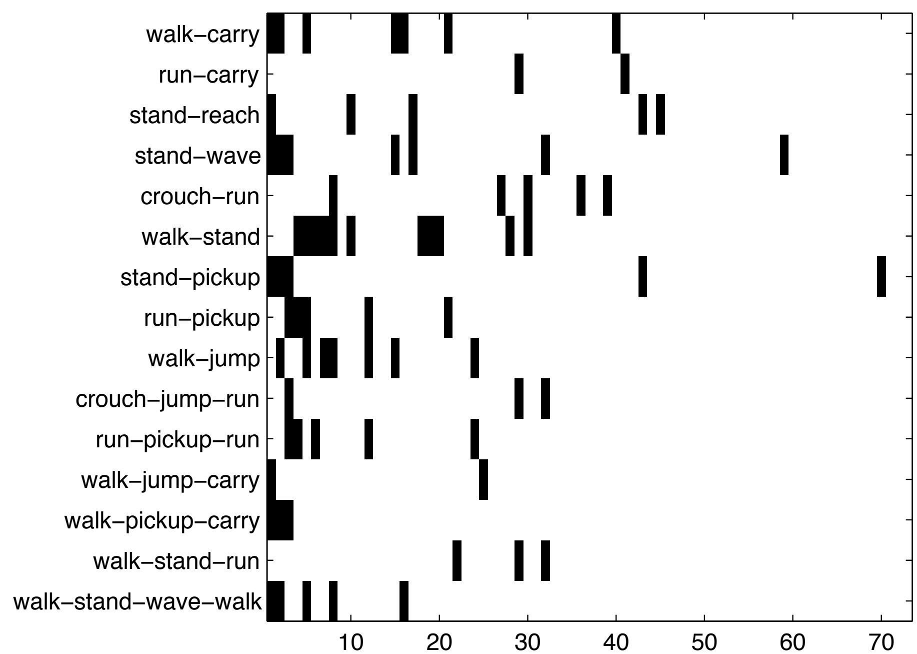
Arms - CCM



Ikizler+Forsyth, 06?



runjump
runback
walk
walkback
pickup
jump
stand
crouch
run
walk
wave
pickup
jump
reach
stand
crouch
carry
run
10
20
30
40
50
60
legs
arms
walk
walk/carry
pickup



Ranked query results for composite queries for 73 videos, black is relevant
Ikizler+Forsyth 06?

Take home points

- There is very seldom a taxonomy
- It is not clear what is important
  - expressive models of what the body is doing?
  - location information?
  - other sensors?
- Generative models based around FSA/HMM are popular
- Discriminative models are well worth using
- Very little clear information about best ways to proceed.