Representing activities

- Requirements
  - dynamical structure
    - cope with sequencing, etc.
  - logical structure
    - cope with different orderings, etc.
  - View
    - probably tolerant to view changes
- Applications
  - Sign language understanding
  - Gesture based interfaces
  - Surveillance

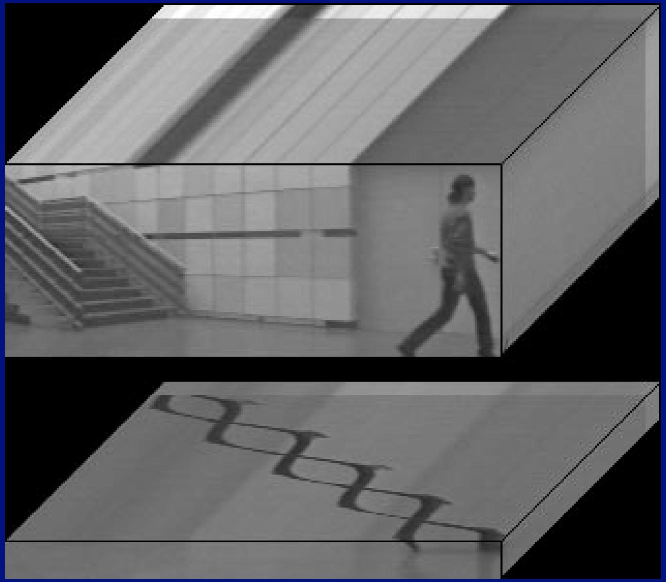
Absence of taxonomy

- Work with activities that have a taxonomy
- Detect “unusualness”
- Match (this is like that)
- Invent taxonomies

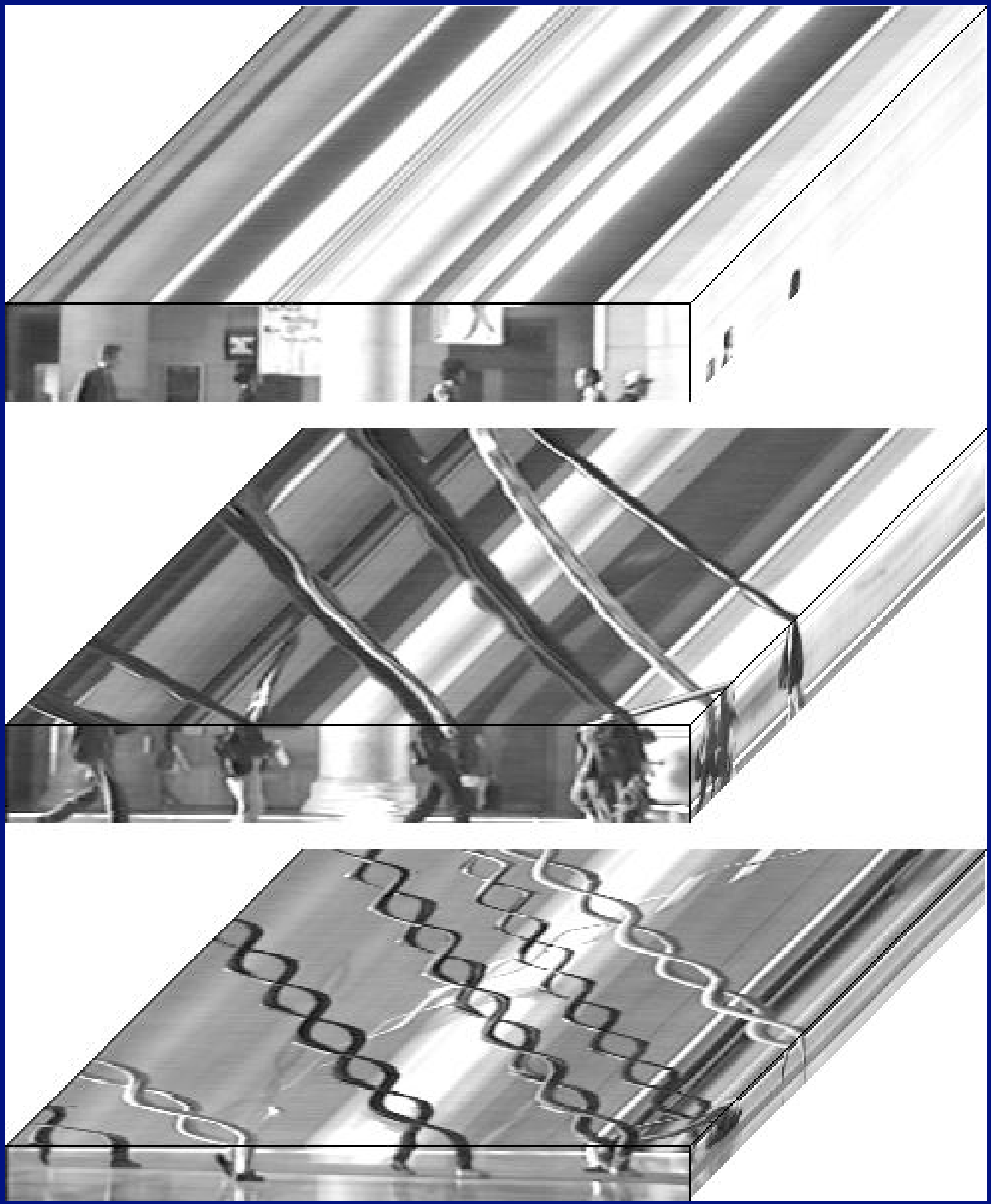
- Should there be intermediate levels of representation?

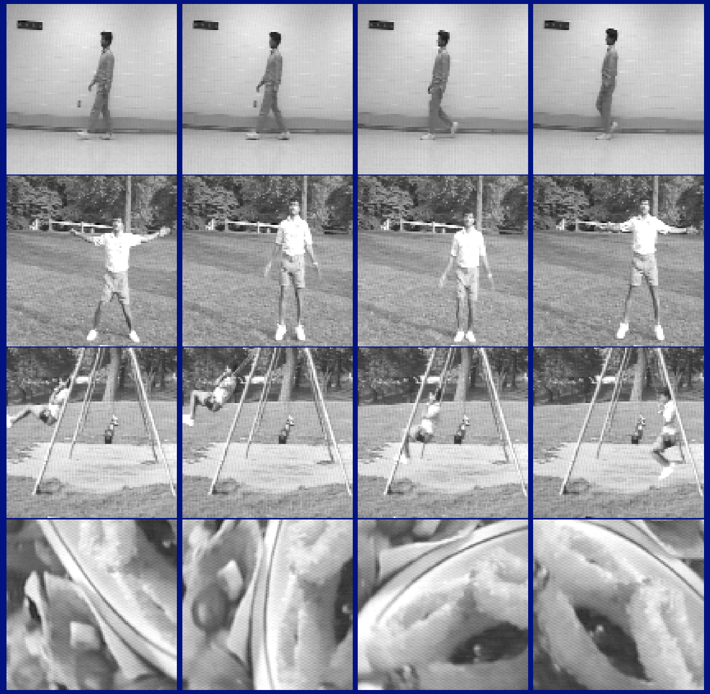
Appearance as a cue

- Many movements have quite characteristic appearances

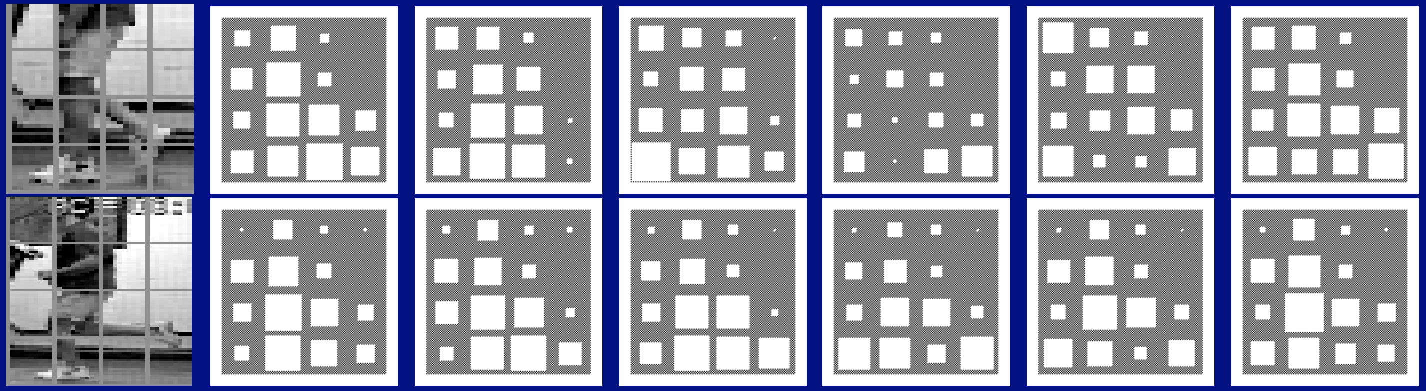


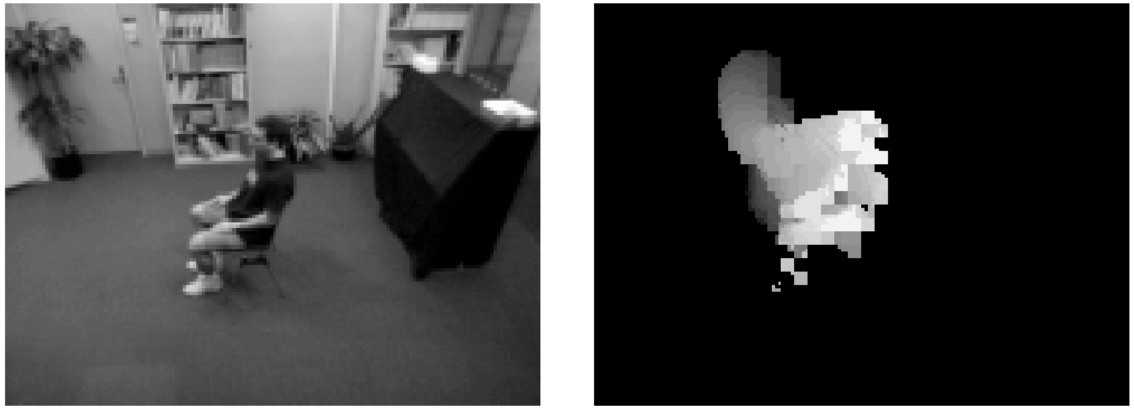
Niyogi Adelson 94



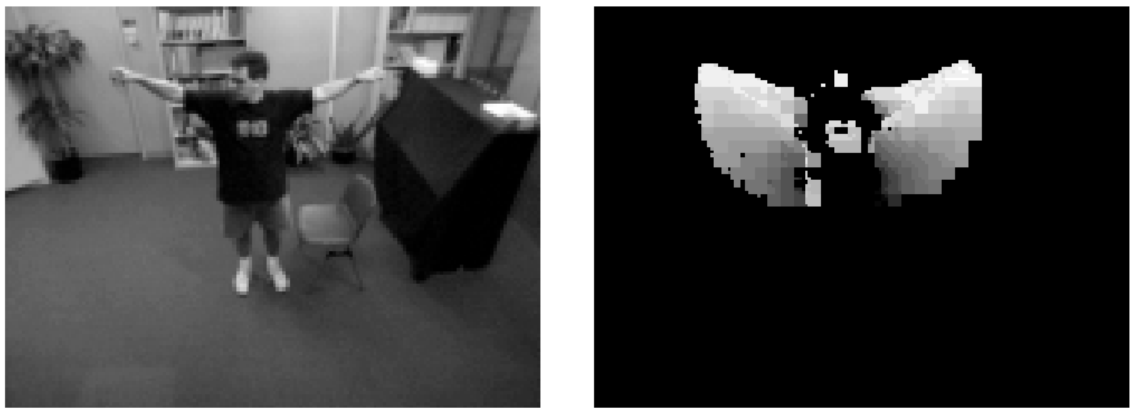


Polana Nelson 93, 94

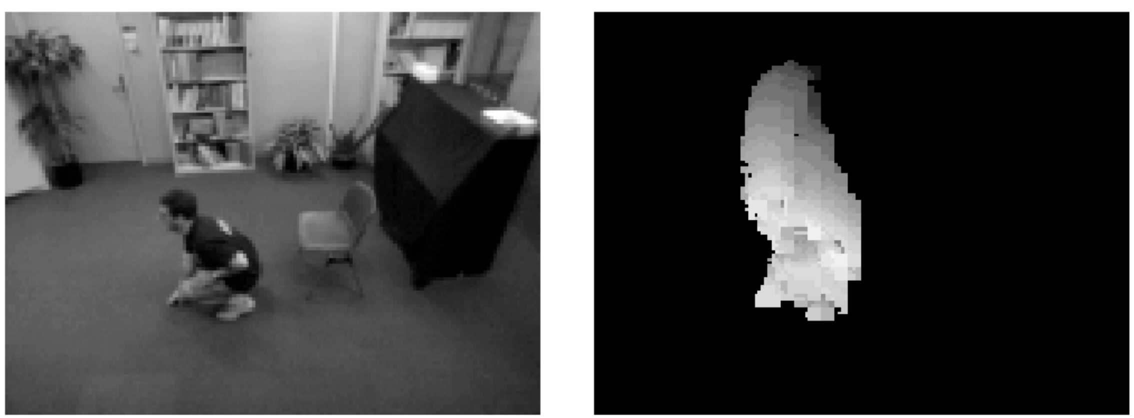




sit-down | sit-down MHI

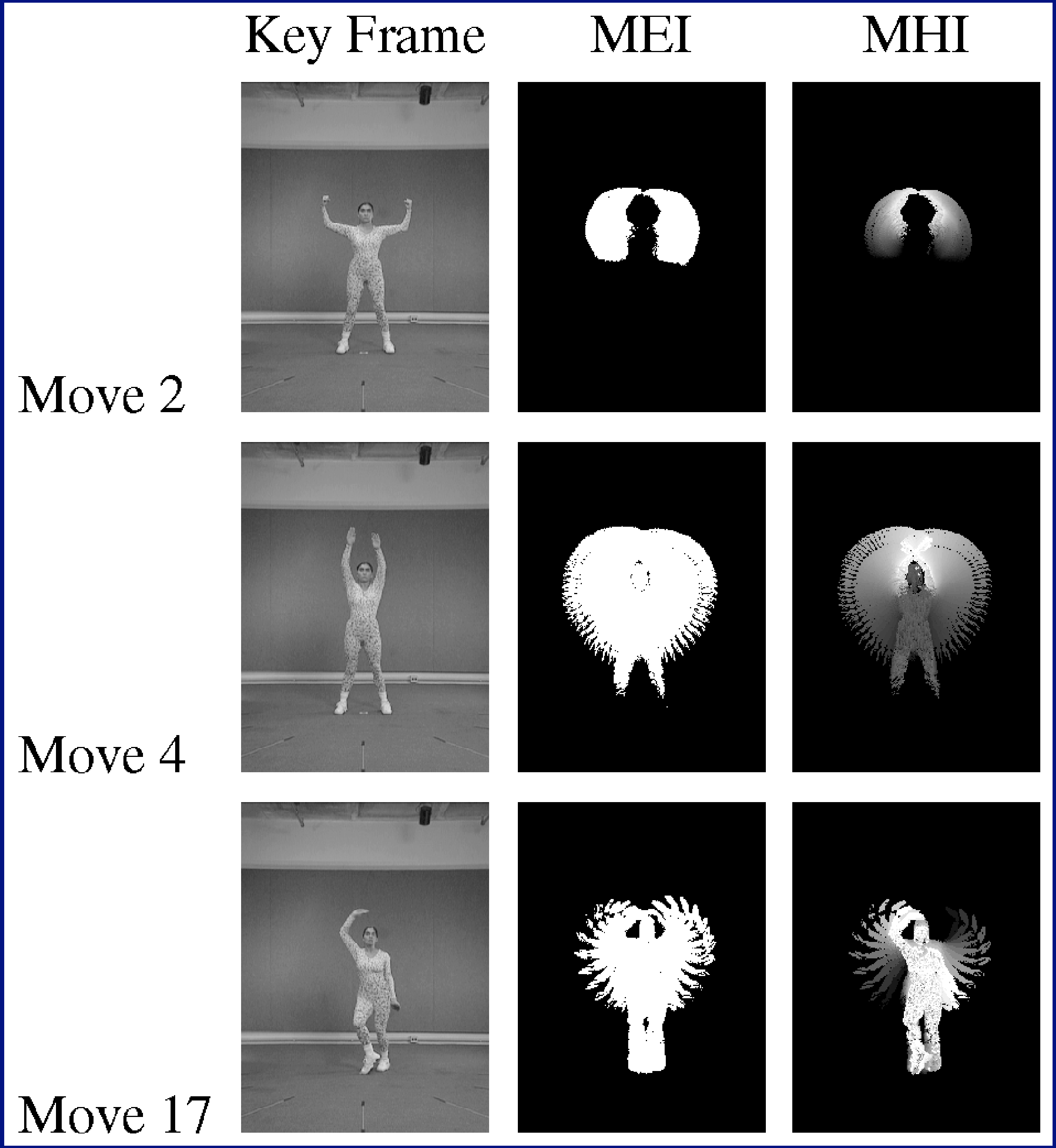


arms-wave | arms-wave MHI

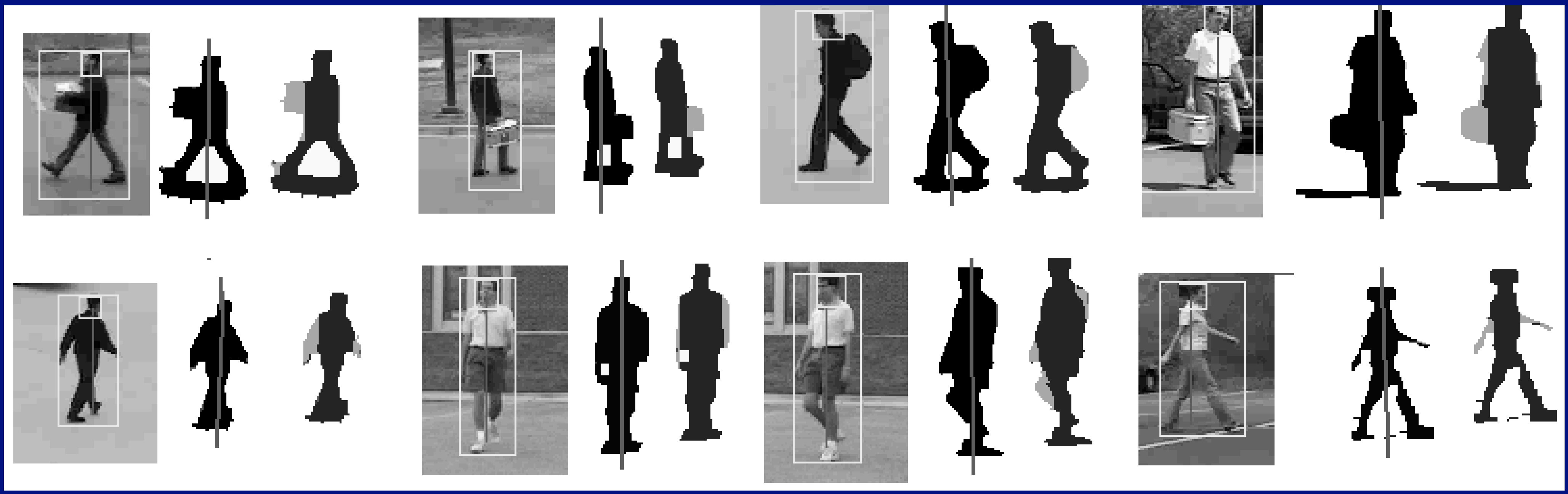


crouch-down | crouch-down MHI

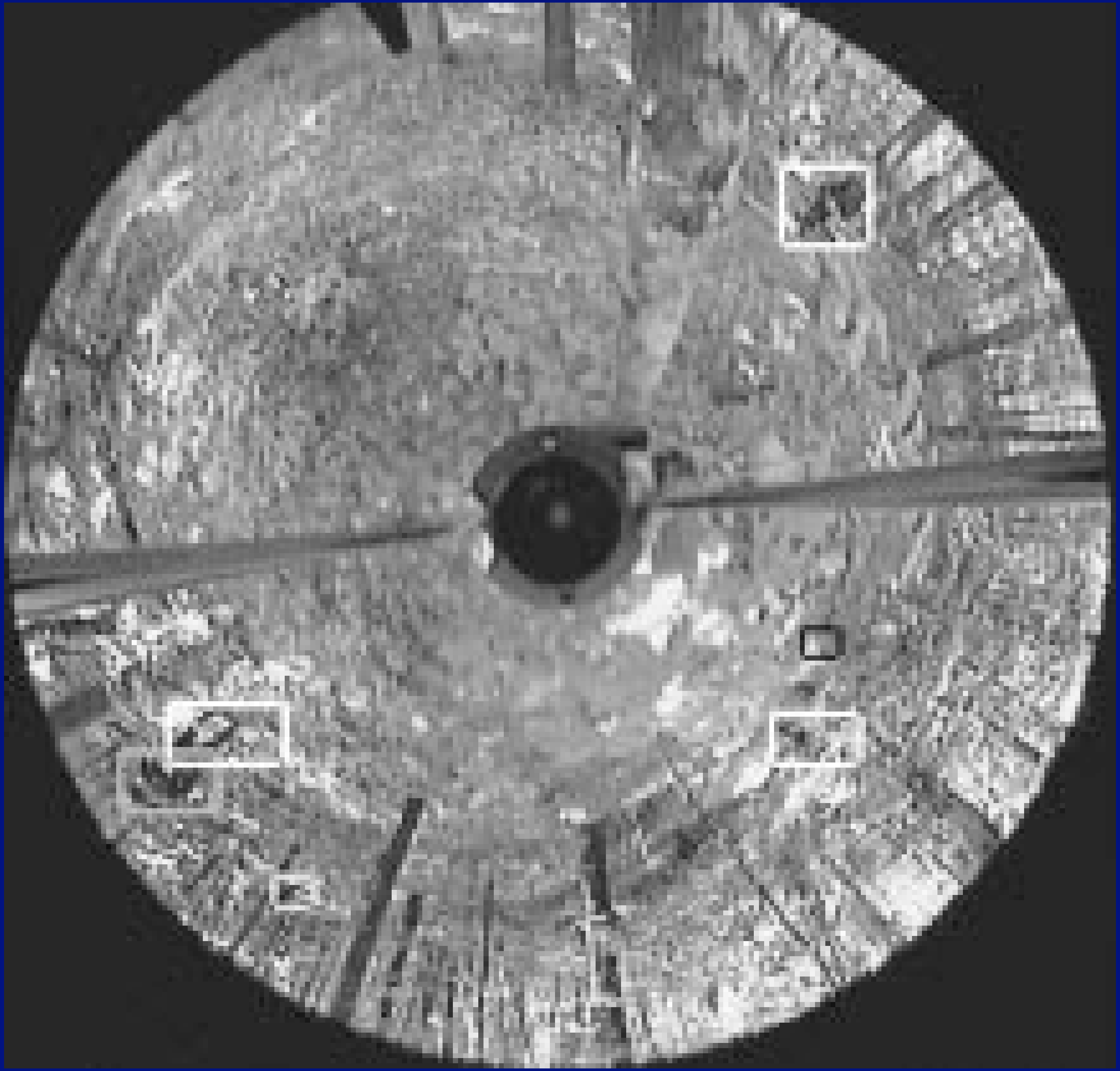
Bobick + Davis, 97



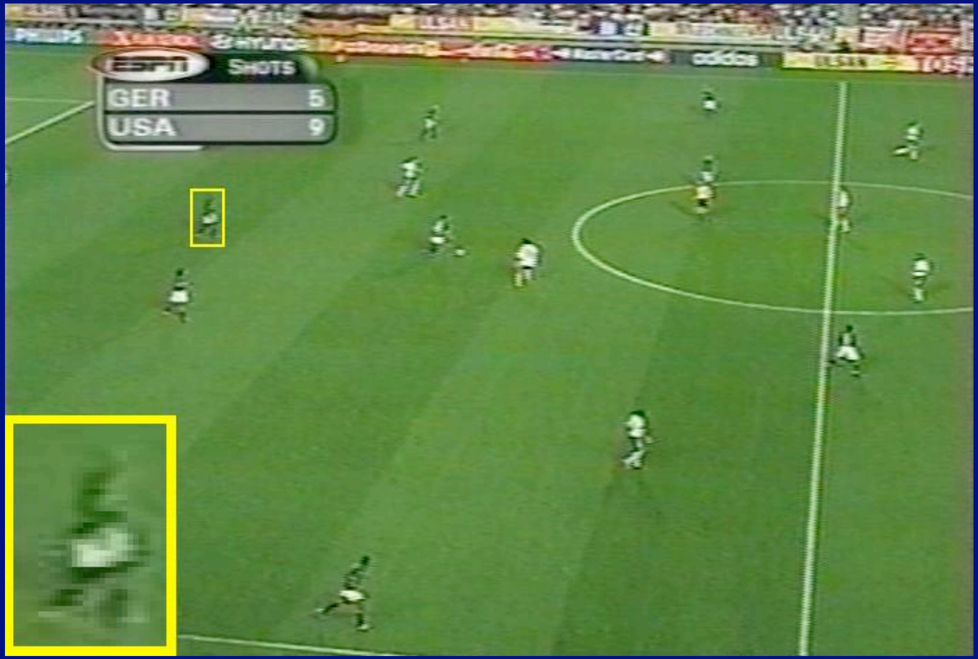
Bobick + Davis, 97



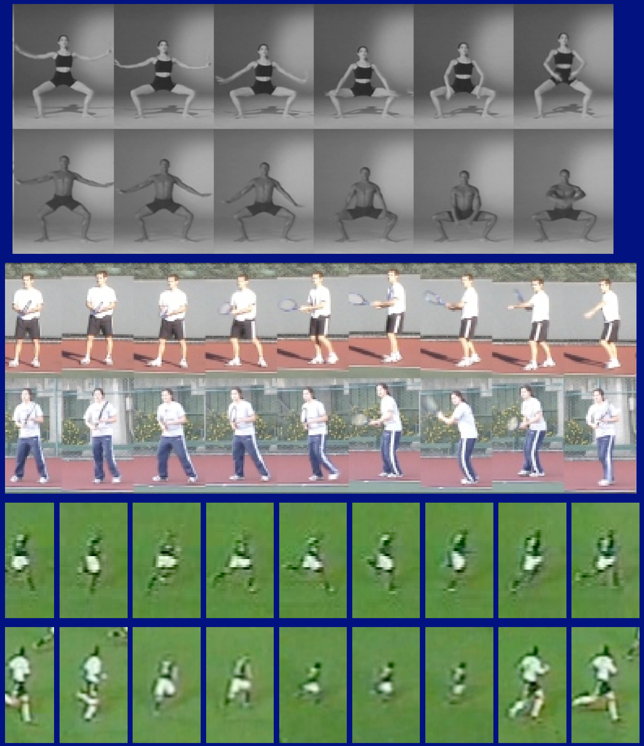
Haritaoglu, Cutler, Harwood, Davis



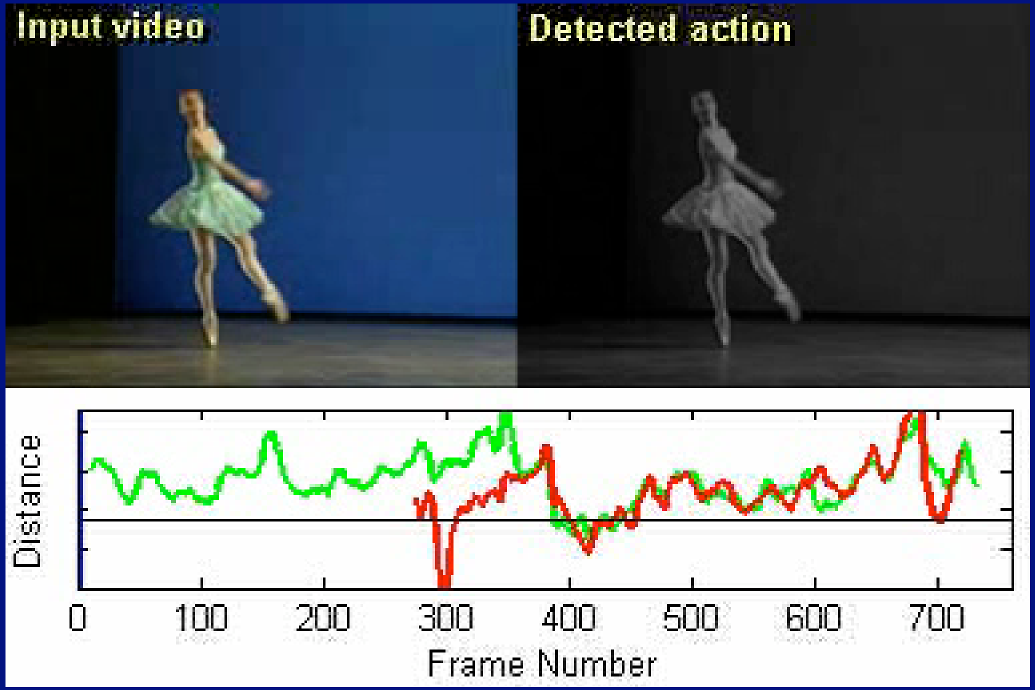
Boult et al 2001



Efros et al 03



Irani et al 05



Irani et al 05

Irani et al 05

Irani et al 05

HMM'S - core ideas

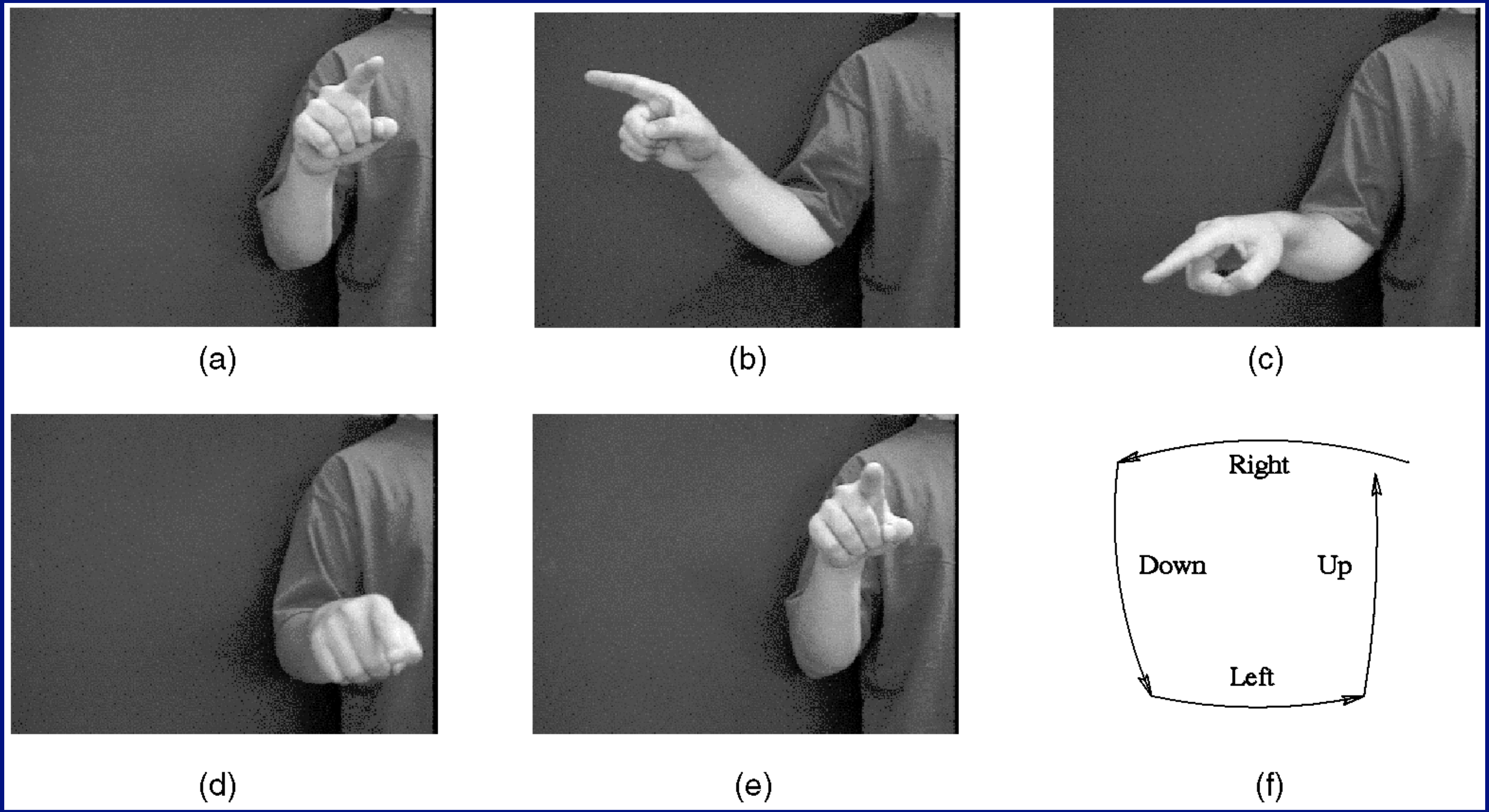
- Finite state machine maintains hidden state; there are stochastic state transitions at known time steps
- At each time step, a measurement is emitted with probability conditioned on the hidden state
- Inference
  - Dynamic programming
  - beam search
- Learning
  - EM

HMM’s in speech understanding

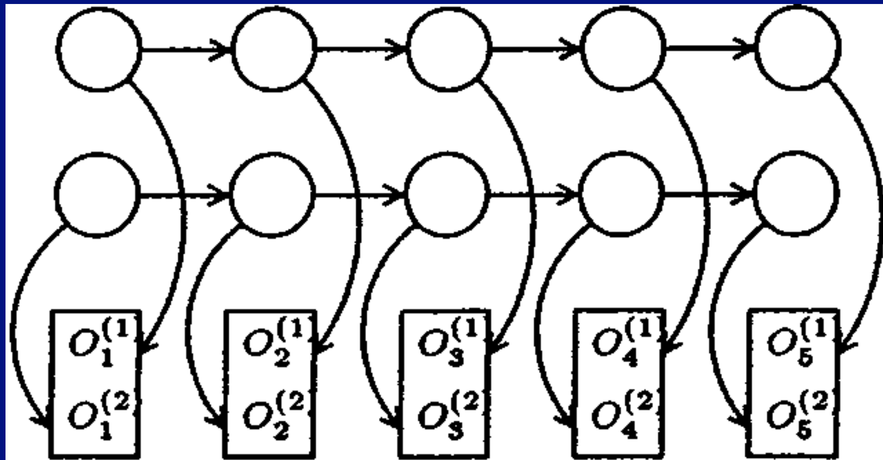
- A string of words is modelled at several levels, e.g.
  - trigram word models
  - pronunciation dictionary per word
  - context dependence of phonemes
  - acoustic model of context dependent phones
- Each is an FSM
  - these are composed
  - missing parameters can be supplied in a variety of ways
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

HMM's in activity recognition

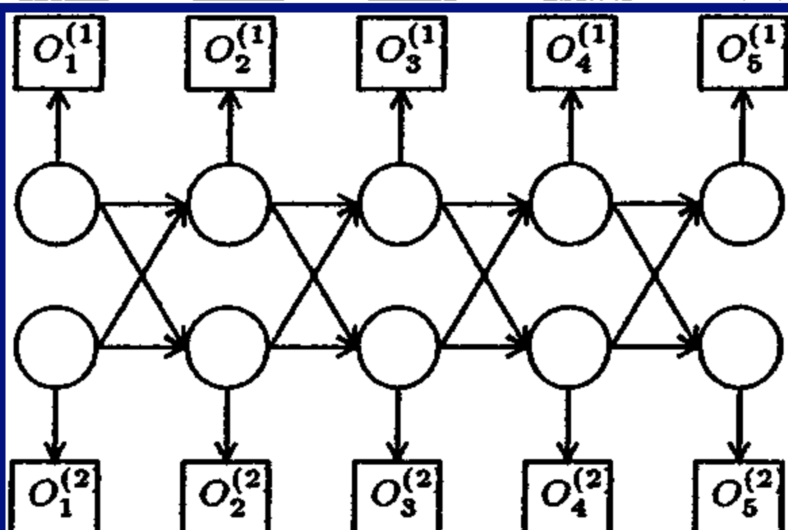
- Gesture
  - No pronunciation dictionaries, trigram models, etc. available
    - very difficult to learn with large state spaces
      - various hacks
- Sign language
  - No pronunciation dictionaries, trigram models, etc. available
    - but (perhaps) lots of data
      - no pooling phone data over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



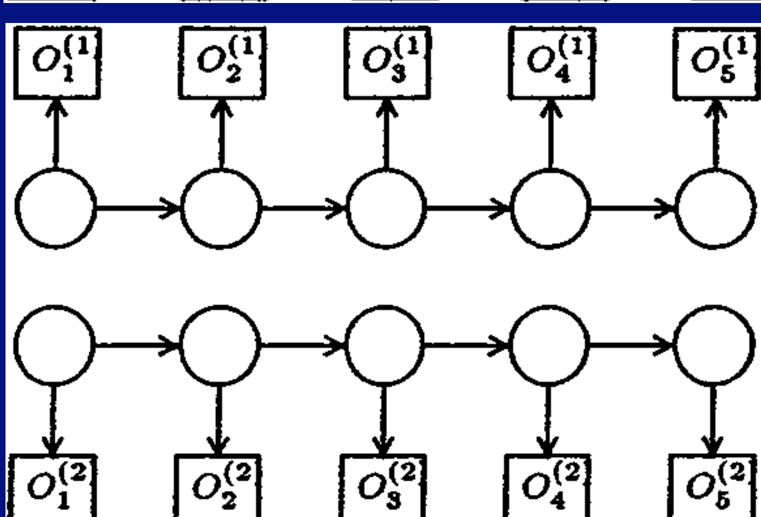
Ivanov and Bobick 2000



Factorial HMM’s



Coupled HMM’s

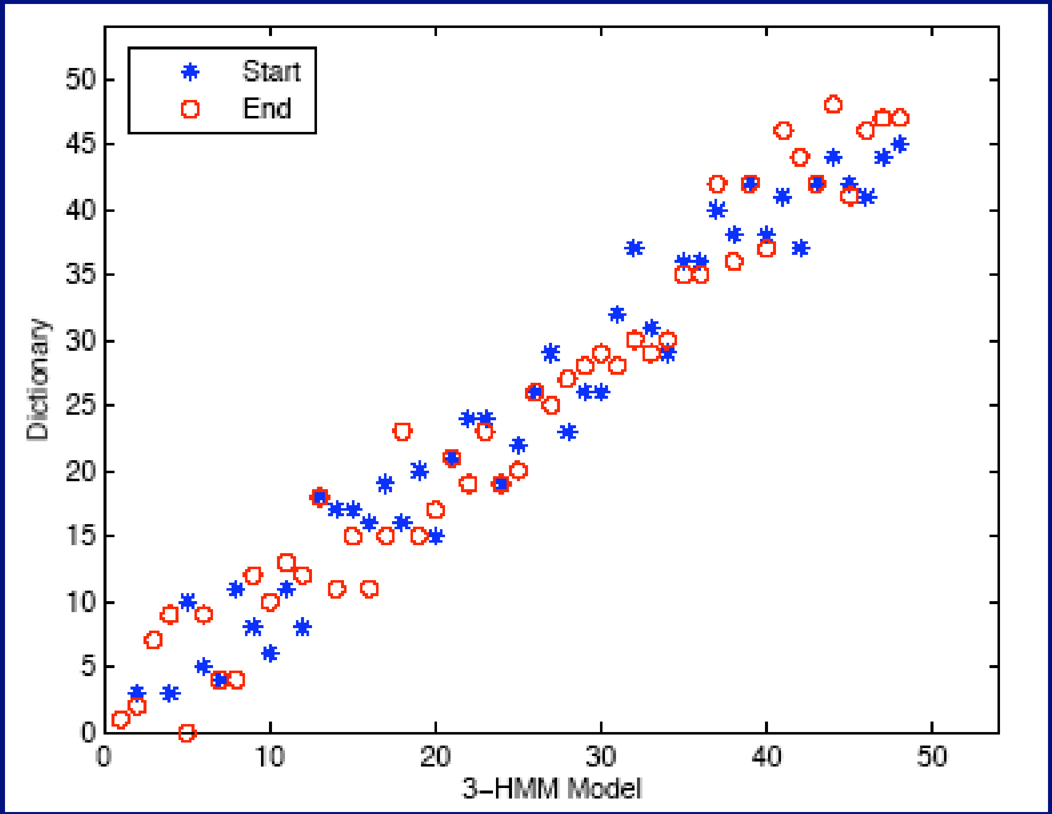


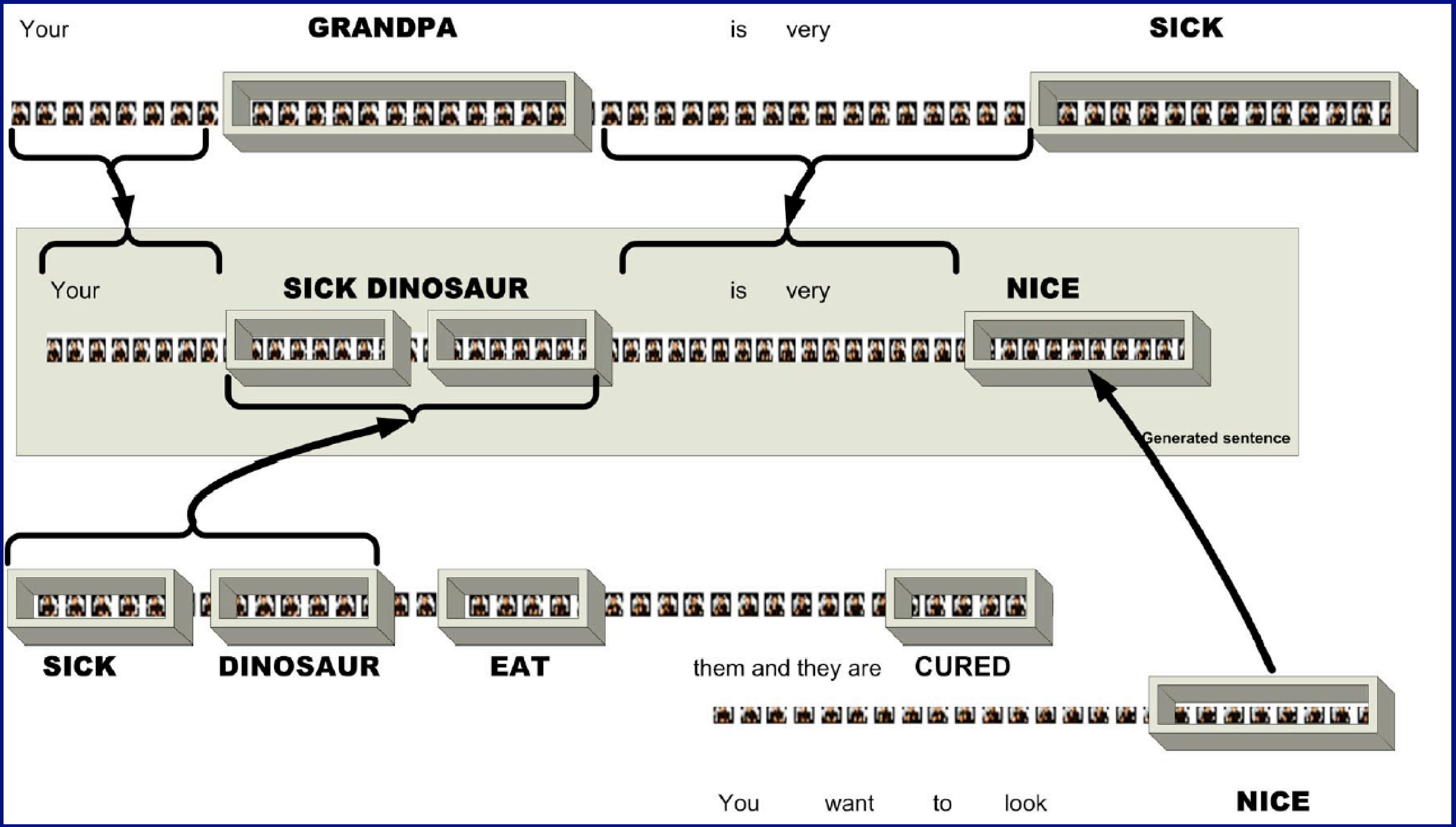
Parallel HMM’s

on rates

The best-known system for sign matching is due to Starner and Pentland [419, 420]. Features are image moments of the hand region; signers either wear coloured gloves, or hands are identified using a skin filter. A Hidden Markov Model (HMM) is used to model individual signs; signs are strung together with a rigid language model (`pronoun verbnoun adjective pronoun`). Authors report a recognition rate of 90% with a vocabulary of 40 signs. Grobel and Assan recognize isolated signs under similar conditions for a 262-word vocabulary using HMM's [227]. This work was extended to recognize continuous German sign language with a vocabulary of 97 signs by Bauer and Hienz [34]. Vogler and Metaxas have built a system that uses estimates of arm position, recovered either from a physical sensor mounted on the body or from a system of three cameras that measures arm position fairly accurately [455, 456, 459]. For a vocabulary of 53 words, and an independent word language model, they report a word recognition accuracy of the order of 90%. A more recent system attempted to recognize phonemes with HMM's; Vogler and Metaxas were able to recognize signs from a 22 word vocabulary with similar recognition rates for phoneme and word models (without handshapes in [457], with handshapes in [458]).

Kadous transduced isolated Australian sign language signs with a powerglove, reporting a recognition rate of 80% using decision trees [305]. Matsuo *et al* transduced Japanese sign language with stereo cameras, using decision tree methods to recognize a vocabulary of 38 signs [278]. Kim *et al.* transduce Korean sign language using datagloves, reporting 94% accuracy in recognition for 131 Korean signs [228]. Al-Jarrah and Halawani report high recognition accuracy for 30 Arabic manual alphabet signs recognized from monocular views of a signer using a fuzzy inference system [12]. Gao *et al.* describe recognizing isolated signs drawn from a vocabulary of 5177 using datagloves and an HMM model [141, 465]. Their system is not speaker-independent: they describe relatively high accuracy for the original signer, and a significant reduction in performance for other signers. Similarly, Zieren and Kraiss report high, but not speaker independent, accuracy for monocular recognition of German sign language drawn from a vocabulary of 152 signs [487]. Akyol and Canzler describe an information terminal which can recognize 16 signs with a high, user-independent, recognition rate; their system uses HMM's to infer signs from monocular views of users wearing coloured gloves [11]. Bowden *et al.* use independent component analysis to obtain state estimates from a set of discriminative visual features; each sign is encoded as a Markov chain, learned from a single example [52]. They report high accuracy recognition from a lexicon of 49 signs using a very small training set.


Start
End
Dictionary
3–HMM Model
0
10
20
30
40
50
5
15
25
35
45


Your
GRANDPA
is very
SICK
Your
SICK DINOSAUR
is very
NICE
Generated sentence
SICK
DINOSAUR
EAT
them and they are
CURED
You want to look
NICE