Frontiers of
Human Activity Analysis

J. K. Aggarwal
Michael S. Ryoo
Kris M. Kitani

THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

ETRI

Carnegie Mellon
THE ROBOTICS INSTITUTE

Single layered approaches

Two different views

- Activities as human movements
  - Semantic-oriented
  - 3-D body-part estimation
  - Tracking

- Sequence

- Activities as video observations
  - Data-oriented
  - Spatio-temporal features
  - Bag-of-words

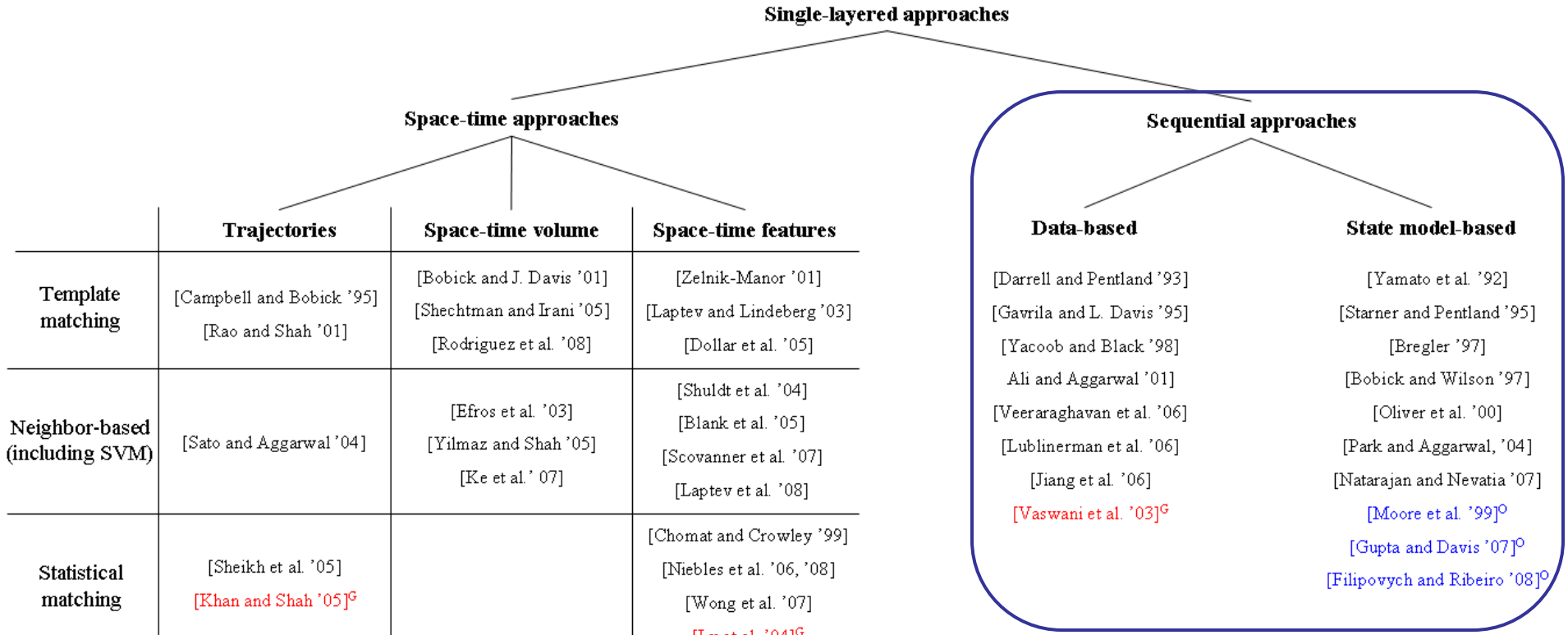
- Space-time distribution

Single layered approaches

Sequential approaches

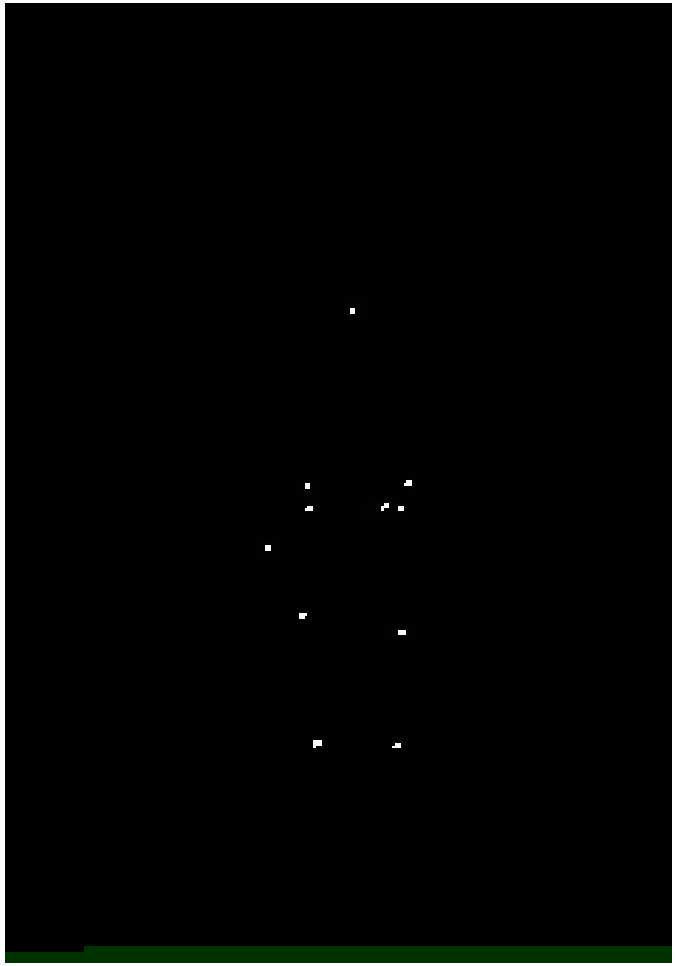
Sequential approaches

- Actions as a set of videos
  - Videos as feature sequences



Sequential approaches

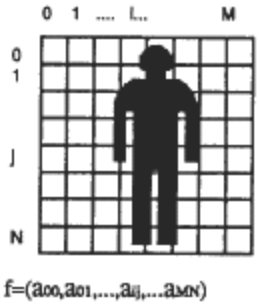
- Motivation
  - An action is a ***sequence*** of body-part states
  - Each frame in an action video describes a particular body-part configuration
    - Example:
      11 points
      body configuration of ‘kicking’



Action recognition using HMMs

- Recognition using hidden Markov models
  - Each HMM generates a particular sequence of features.
  - Matching observed features with the model.
    - An action -> a set of sequences of features

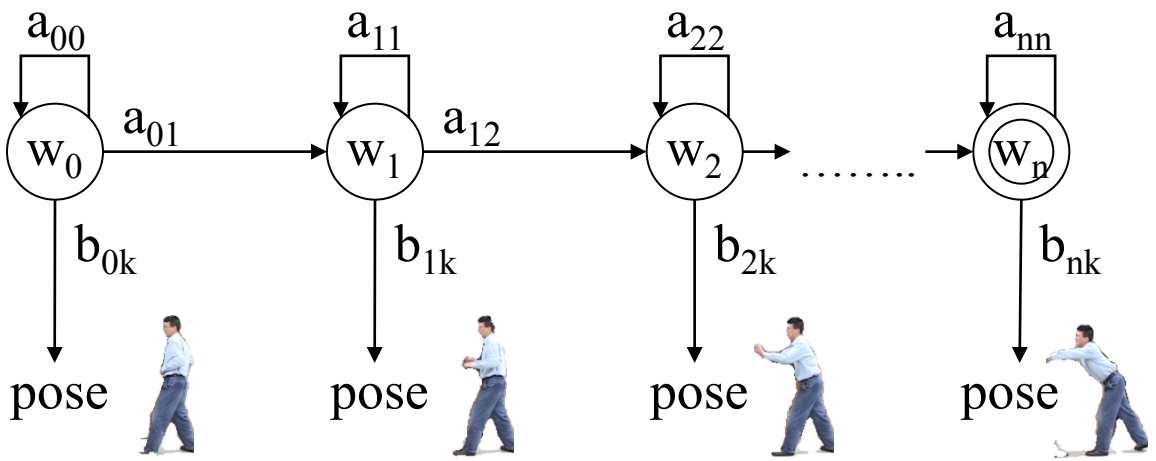
- [Yamato et al. CVPR 1992]: Tennis plays



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

HMMs for actions

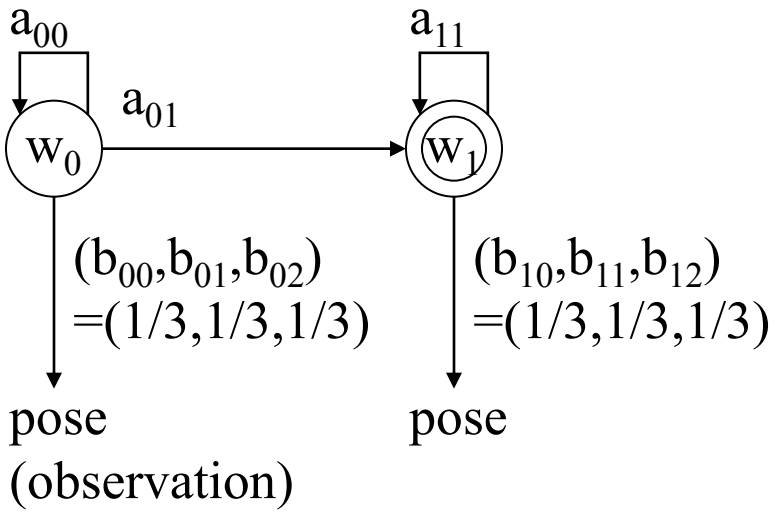
- Human action as a pose sequence
- Each hidden state is trained to generate a particular body posture.
  - Each HMM produces a pose sequence: action



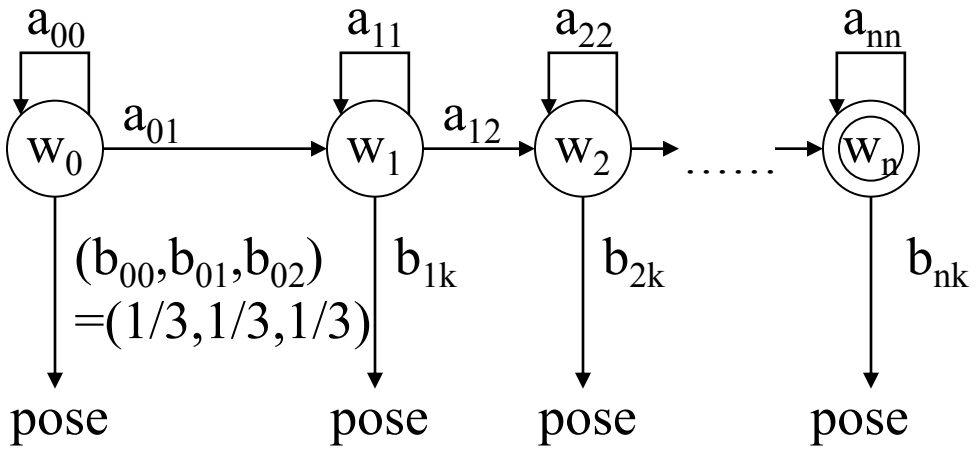
Hidden Markov models

- This is a classic evaluation problem of HMMs.
  - Given observations V^T (a sequence of poses), find the HMM M_i that maximizes $P(V^T|M_i)$: forward algorithm.
  - Transition probabilities a_{ij} and observations probabilities b_{ik} are pre-trained using training data.

Noise HMM

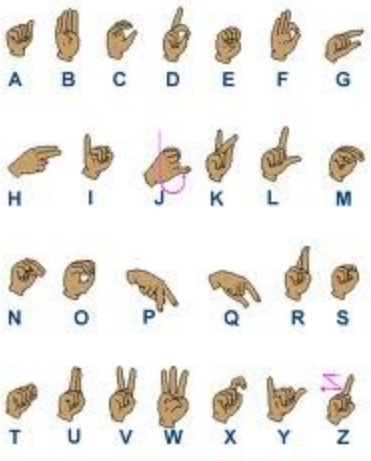


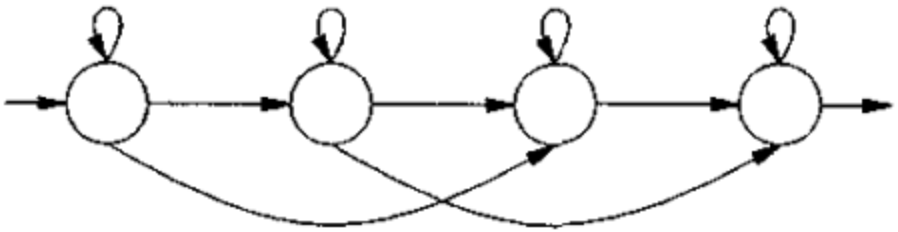
Structure of other basic HMMs



HMMs for hand gestures

- HMMs for gesture recognition
  - American Sign Language (ASL)
  - Sequential HMMs
    - Features from colored globes

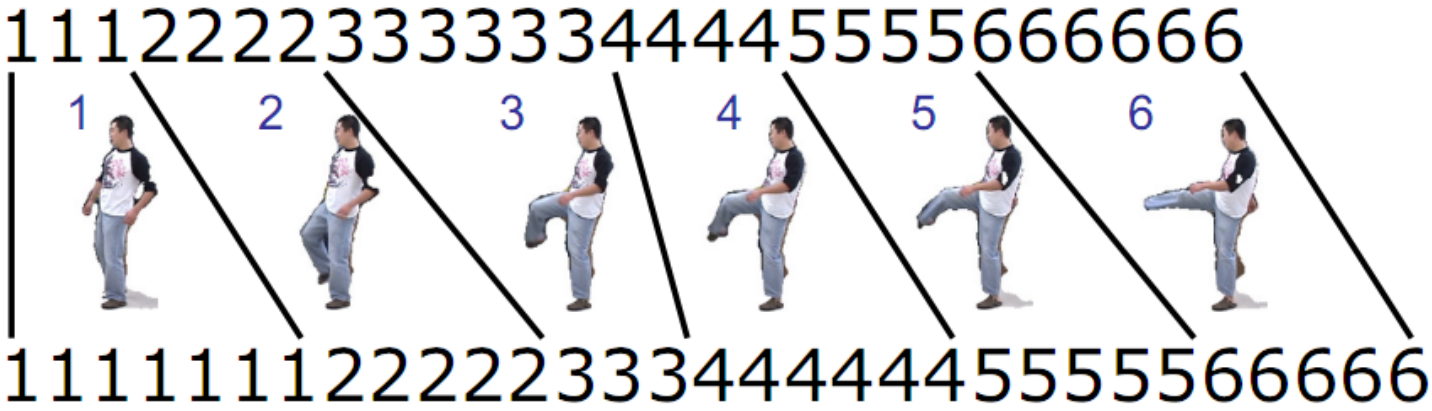




[Starner, T. and Pentland, A., Real-time American Sign Language recognition from video using hidden Markov models. International Symposium on Computer Vision, 1995.]

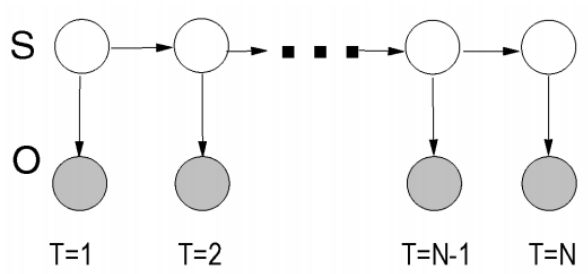
Dynamic time warping

- Dynamic programming algorithm to match two strings (e.g. sequences).
  - [Gavrila and L. Davis, 1995]
  - Each frame generates a symbol (of a feature vector)

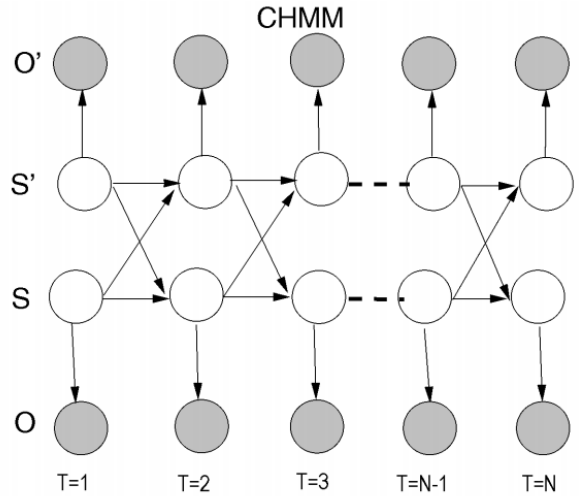


Coupled HMMs

- Pentland CHMMs
  - Human-human interactions
    - Two types of states for two agents
  - Synthetic agents for training HMMs



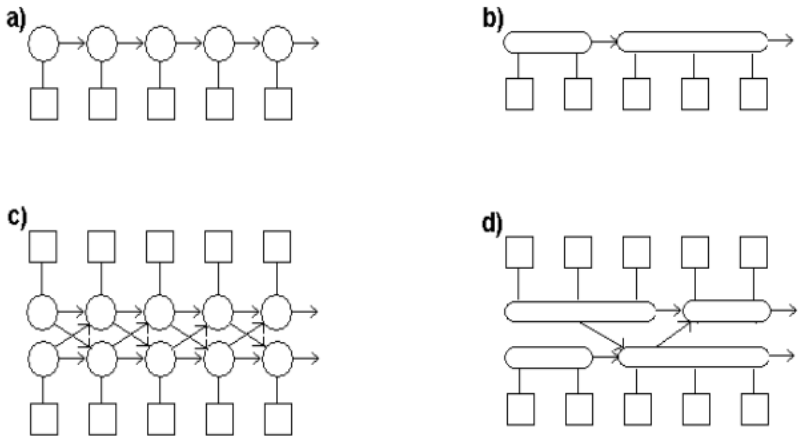
Vs.



[Oliver, N. M., Rosario, B., and Pentland, A. P., A Bayesian computer vision system for modeling human interactions. IEEE T PAMI, 2000.]

HMM Variations

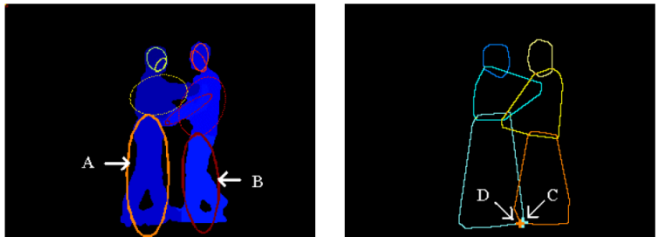
- Coupled hidden semi-Markov models
  - Natarajan and Nevatia 2007
  - Human-human interactions
  - Activities with varying durations
    - Models probabilistic distributions of state durations.

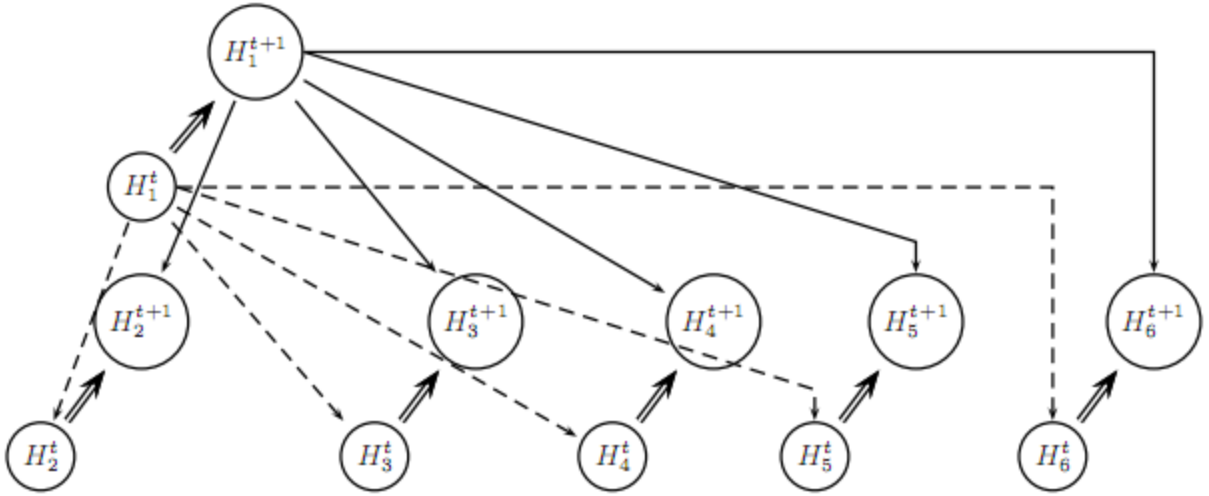


[Natarajan, P. and Nevatia, R., Coupled hidden semi Markov models for activity recognition. WMVC 2007]

Dynamic Bayesian networks

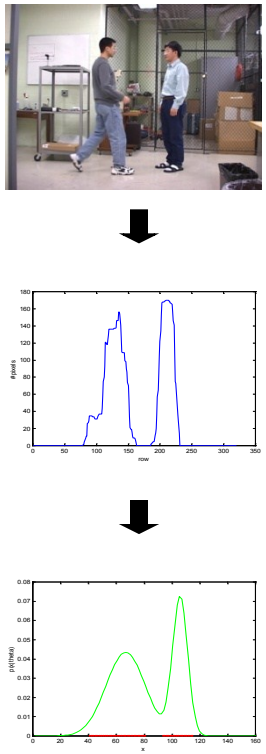
- Diverse variations
  - Dynamic Bayesian networks
    - Body-part analysis



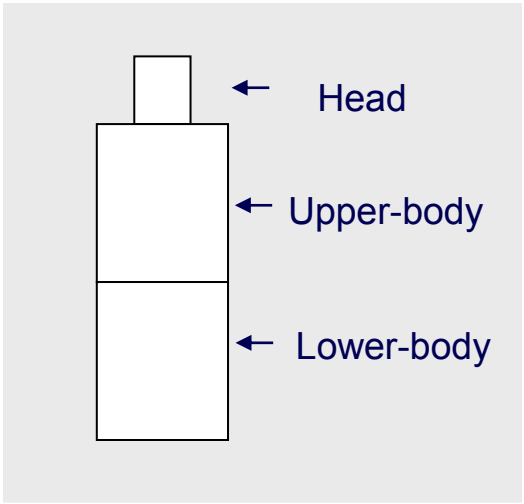


[Park, S. and Aggarwal, J. K., A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, 2004]

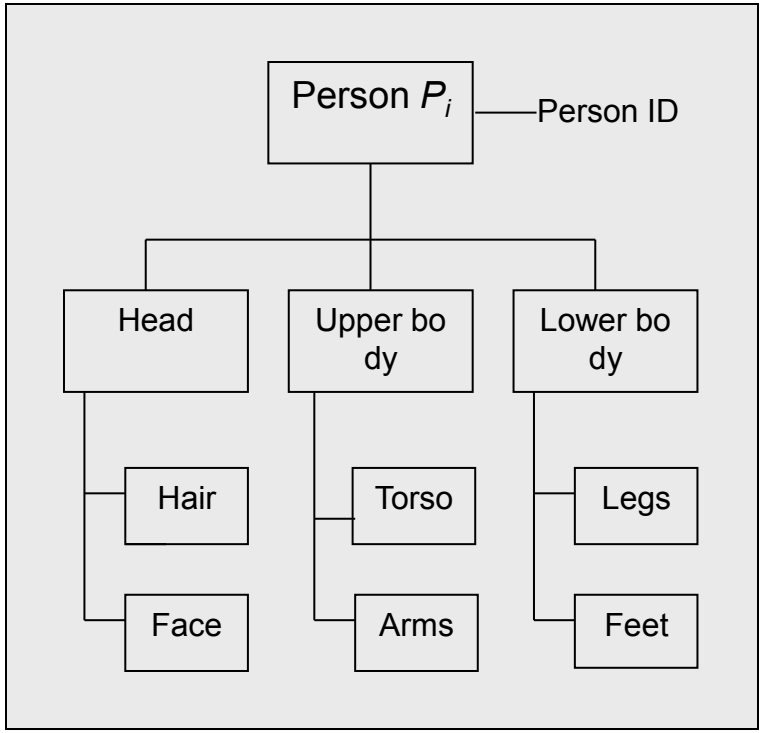
Hierarchical human-body modeling



Vertical projection



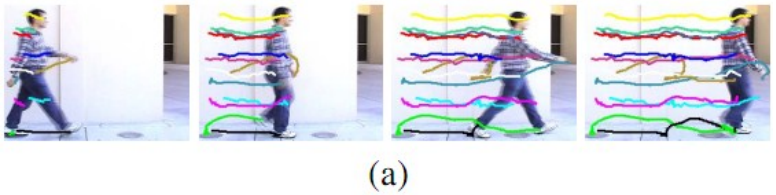
A simple human body model



Data structure

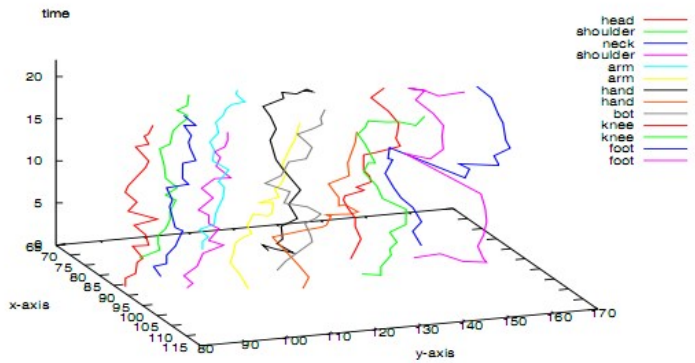
Space-time trajectories

- Trajectory patterns
  - Yilmaz and Shah, 2005 – UCF
  - Joint trajectories in 3-D XYT space.
  - Compared trajectory shapes to classify human actions.



(a)

(b)



(c)

[Yilmaz, A. and Shah, M., Recognizing human actions in videos acquired by uncalibrated moving cameras, ICCV 2005]

Sequential approaches - summary

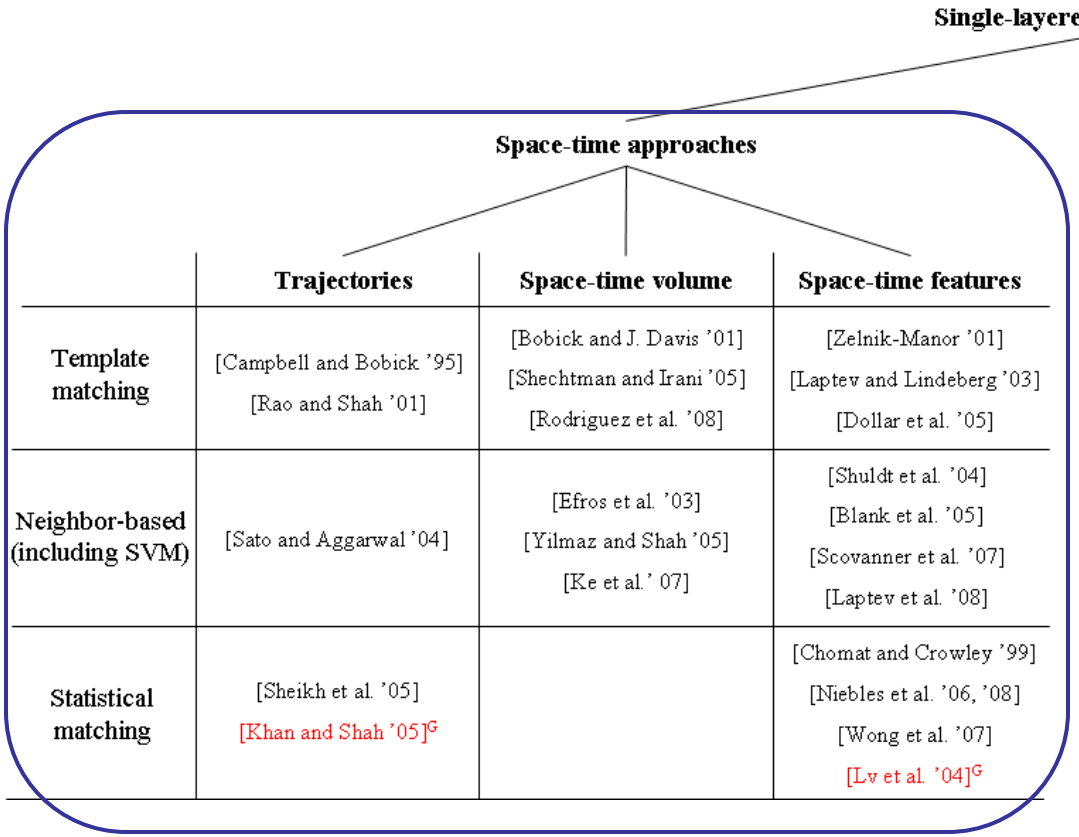
- Designed for modeling sequential dynamics
  - Markov process
  - Motion features are extracted *per frame*
- Limitations
  - Feature extraction
    - Assumes good observation models
  - Complex human activities?
    - Large amount of training data

Single layered approaches

Space-time approaches

Space-time approaches

- Actions as a set of videos
  - Videos as space-time volumes

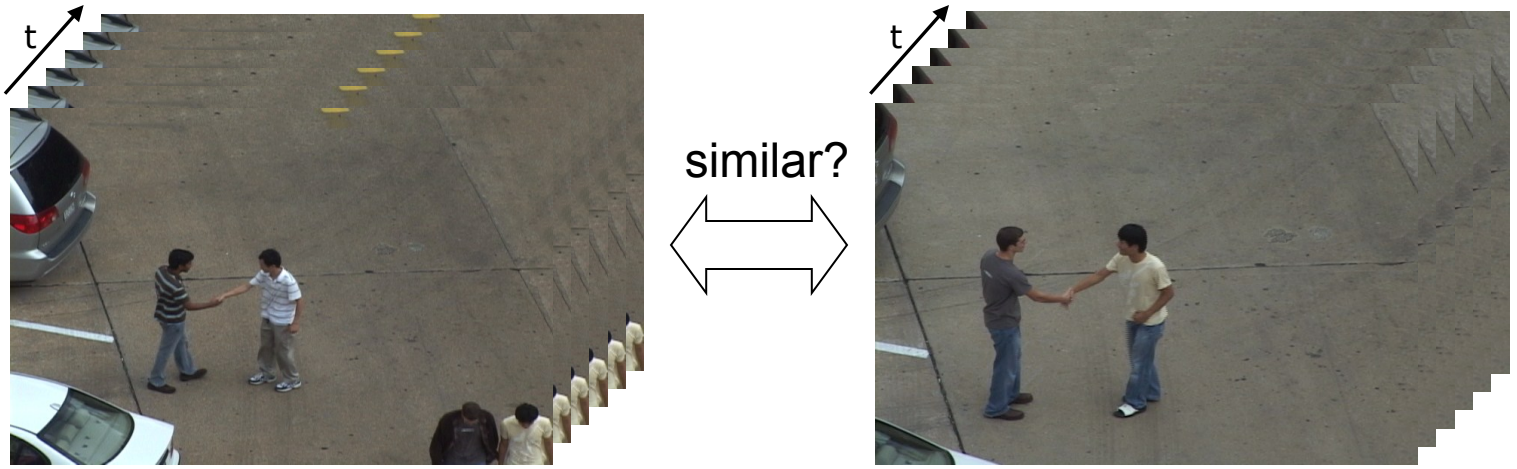


	Trajectories	Space-time volume	Space-time features
Template matching	[Campbell and Bobick '95] [Rao and Shah '01]	[Bobick and J. Davis '01] [Shechtman and Irani '05] [Rodriguez et al. '08]	[Zelnik-Manor '01] [Laptev and Lindeberg '03] [Dollar et al. '05]
Neighbor-based (including SVM)	[Sato and Aggarwal '04]	[Efros et al. '03] [Yilmaz and Shah '05] [Ke et al.' 07]	[Shuldt et al. '04] [Blank et al. '05] [Scovanner et al. '07] [Laptev et al. '08]
Statistical matching	[Sheikh et al. '05] [Khan and Shah '05]^{G}		[Chomat and Crowley '99] [Niebles et al. '06, '08] [Wong et al. '07] [Lv et al. '04]^{G}

Data-based	State model-based
[Darrell and Pentland '93]	[Yamato et al. '92]
[Gavrila and L. Davis '95]	[Starner and Pentland '95]
[Yacoob and Black '98]	[Bregler '97]
Ali and Aggarwal '01]	[Bobick and Wilson '97]
[Veeraraghavan et al. '06]	[Oliver et al. '00]
[Lublinerman et al. '06]	[Park and Aggarwal, '04]
[Jiang et al. '06]	[Natarajan and Nevatia '07]
[Vaswani et al. '03]^{G}	[Moore et al. '99]^{O}
	[Gupta and Davis '07]^{O}
	[Filipovych and Ribeiro '08]^{O}

Space-time approaches

- Videos as 3-D XYT volumes
- Problem: matching between two volumes
  - Match volumes directly
    - Compare volumes from testing videos with those from training videos.

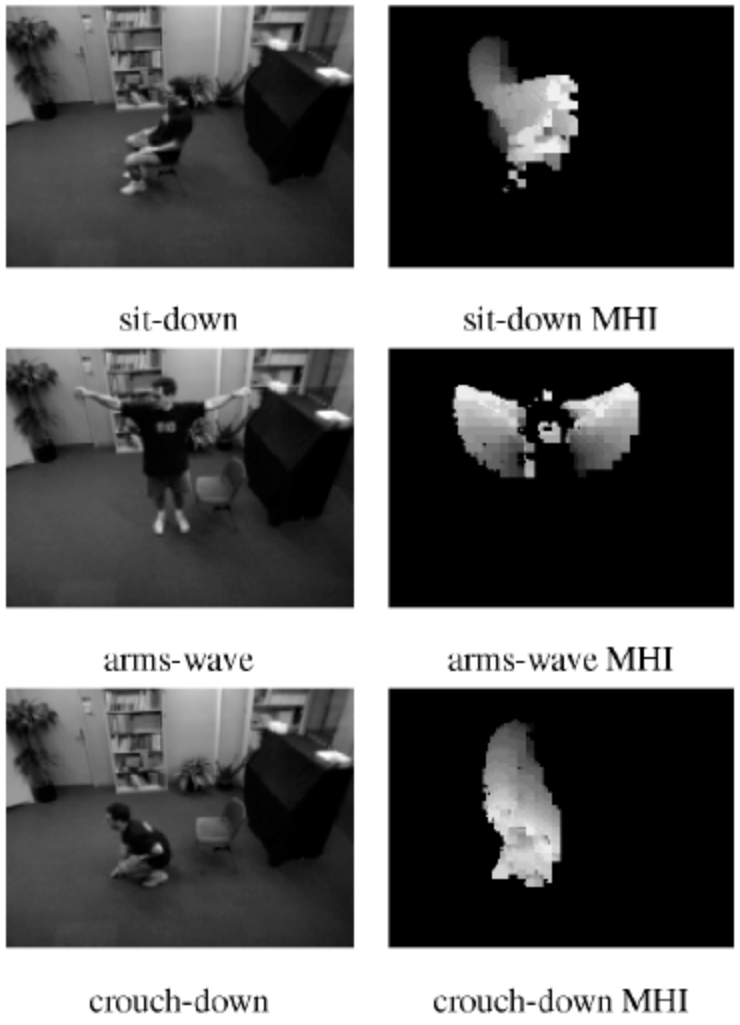


Training video

Testing video

Motion history images

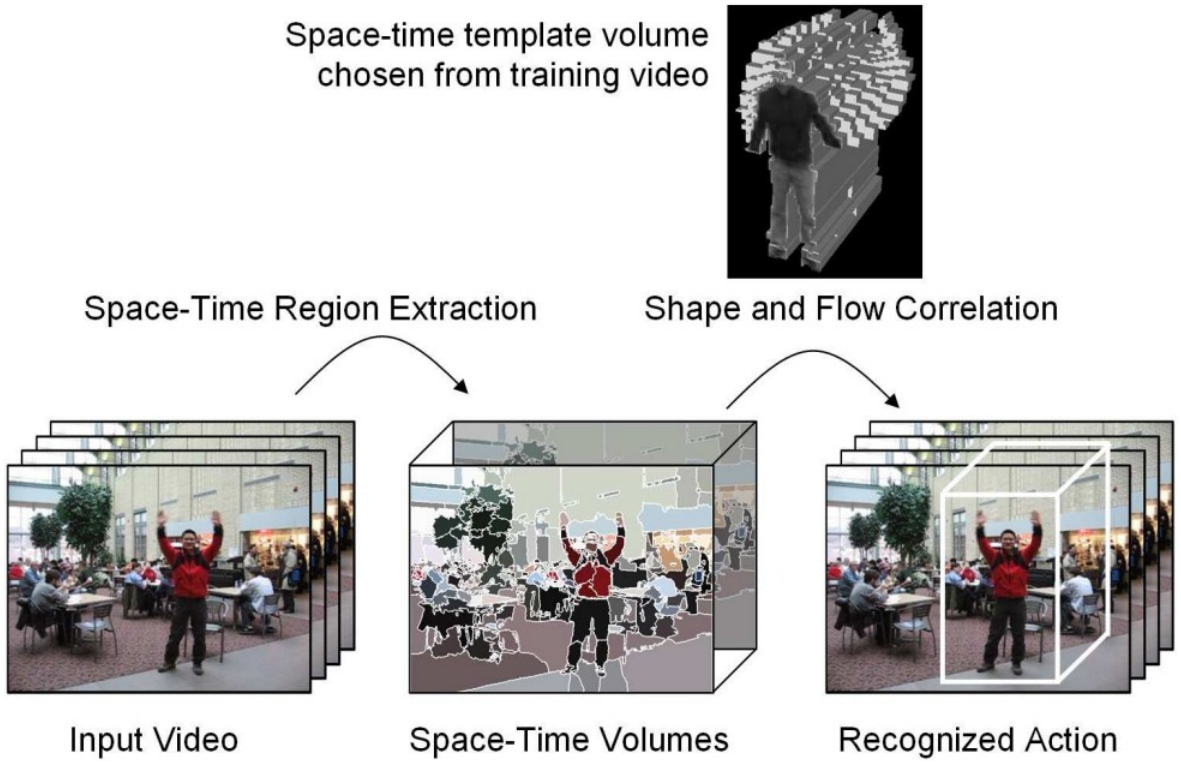
- Matching two volumes
- Bobick and J. Davis, 2001
  - Motion history images (MHIs)
  - Weighted projection of a XYT foreground volume
  - Template matching



[Bobick, A. and Davis, J., The recognition of human movement using temporal templates. IEEE T PAMI 23(3), 2001]

3-D volume matching

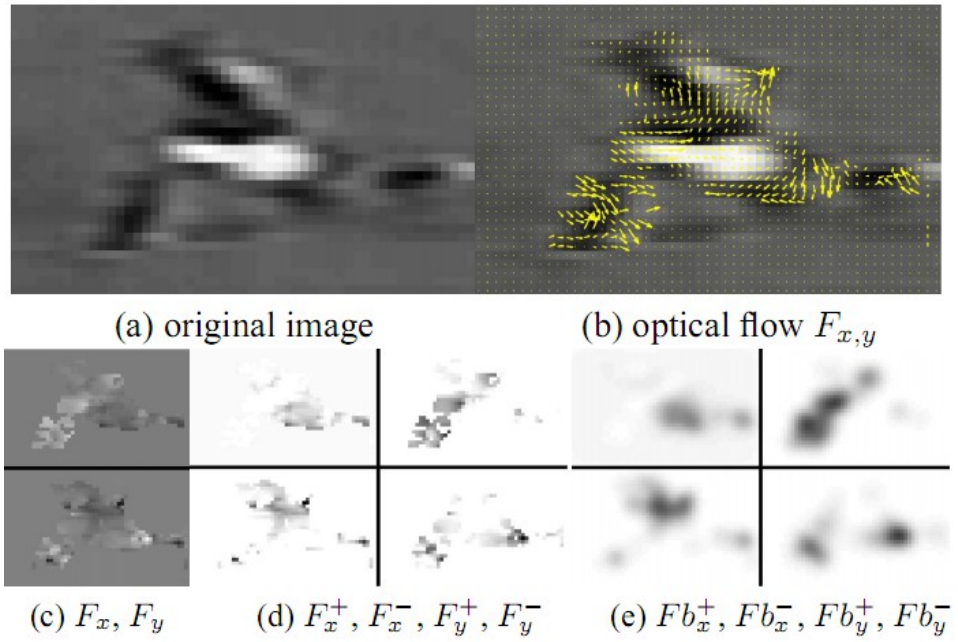
- Ke, Suktankar, Herbert 2007
  - Volume matching based on its segments.
  - Segment matching scores are combined.



[Ke, Y., Sukthankar, R., and Hebert, M., Spatio-temporal shape and flow correlation for action recognition. CVPR 2007]

Global features from volumes

- Efros et al. 2003
  - Concatenated optical flow features from 3-D XYT volumes
  - Analyzed soccer plays from low-resolution videos.

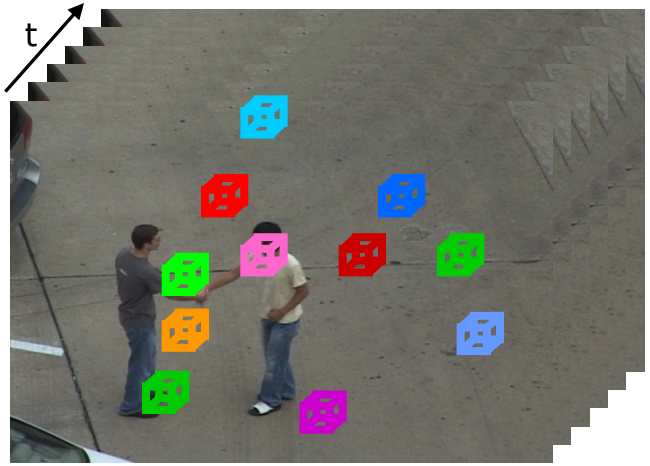


[Efros, A., Berg, A., Mori, G., and Malik, J., Recognizing action at a distance, ICCV 2003]

Sparse features from videos

- Problem: matching between two videos
  - Match volumes directly?
  - Extracts sparse features -
    - Video version of SIFT features

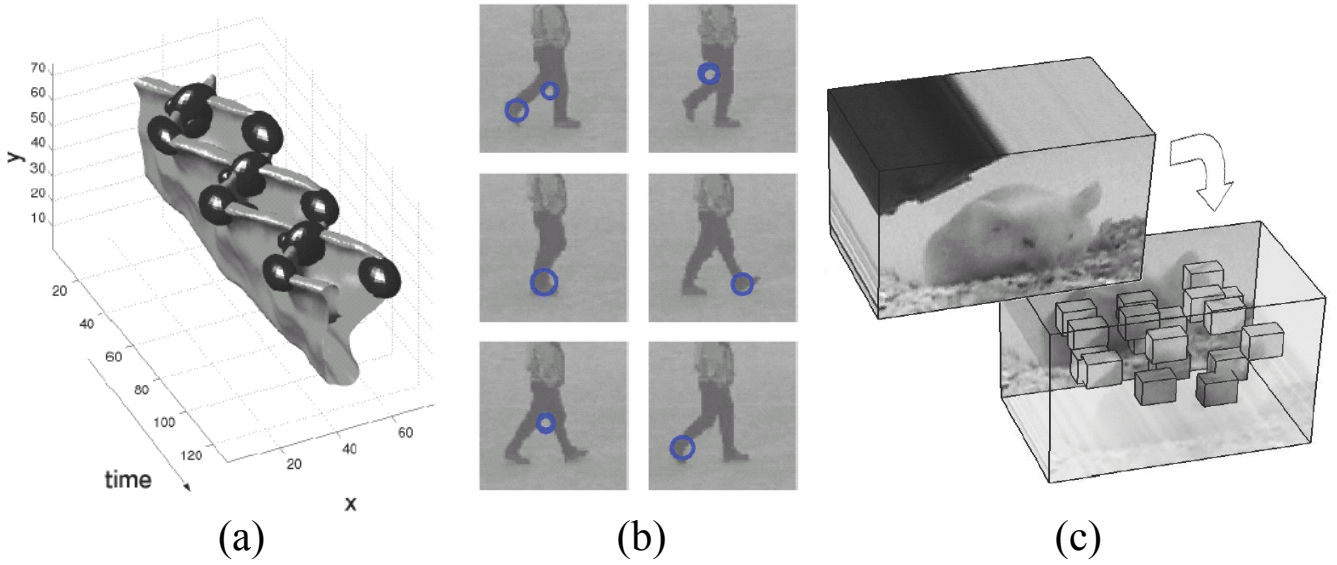
SIFT for images



Features for videos

Sparse features from videos

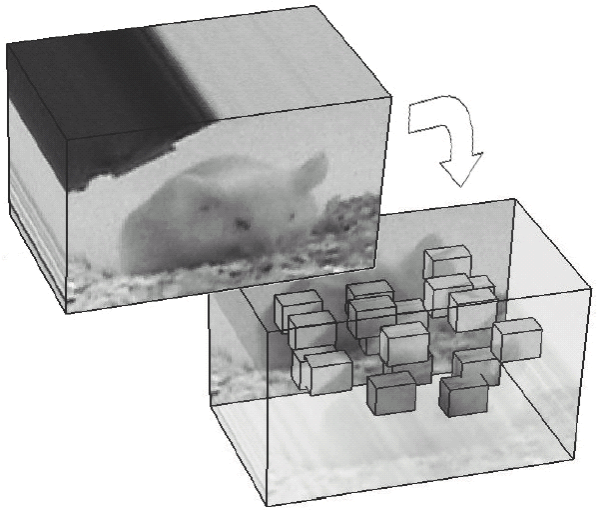
- Spatio-temporal features
  - Reliable under noise, background changes, lighting condition changes, …
  - Laptev 2003, Dollar et al. 2005



(a) (b) (c)

Cuboid features

- Cuboid descriptors
  - Dollar *et al.*, Cuboid, VS-PETS 2005
  - Appearances of local 3-D XYT volumes
    - Raw appearance
    - Gradients
    - Optical flows
  - Captures salient periodic motion.



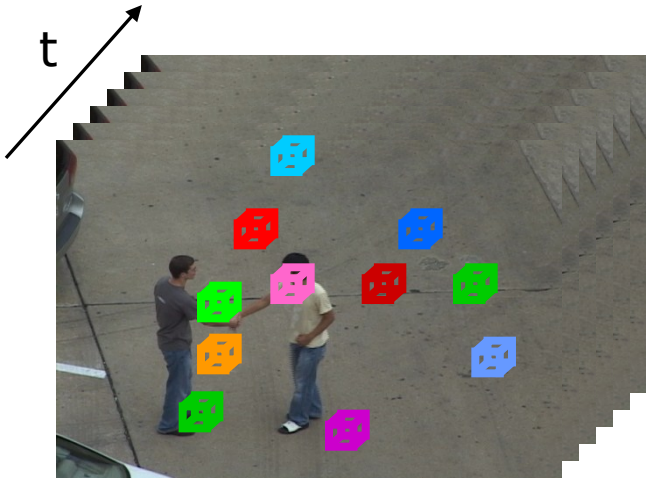
[Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S., Behavior recognition via sparse spatio-temporal features, VS-PETS 2005]

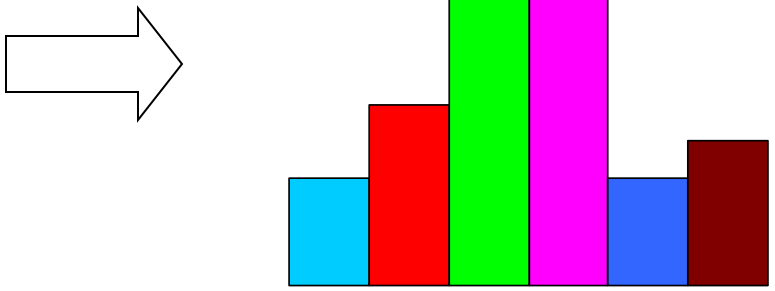
STIP interest point detector

- Laptev and Linderberg 2003
  - Simple periodic actions
    - Spatio-temporal local features + SVMs
  - Introduced the ***KTH dataset***
  - Local descriptor based on Harris corner detector

[Schuldt, C., Laptev, I., and Caputo, B., Recognizing human actions: A local SVM approach, ICPR 2004]

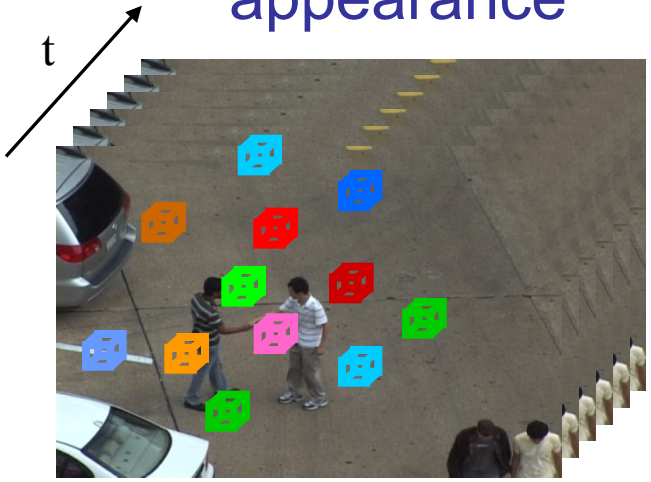
Bag-of-words representation

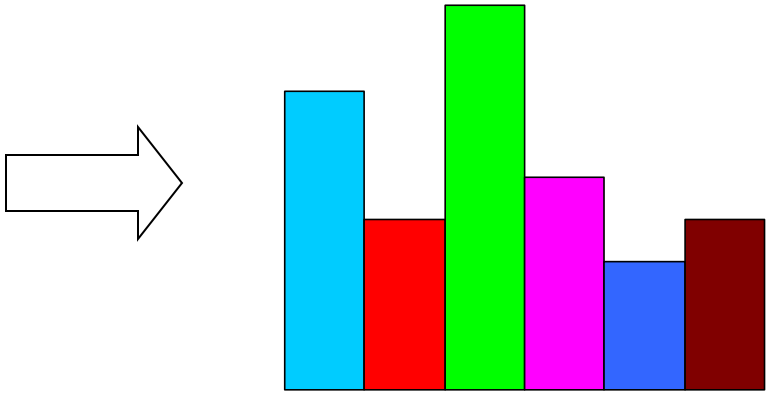




Classify features based on their appearance

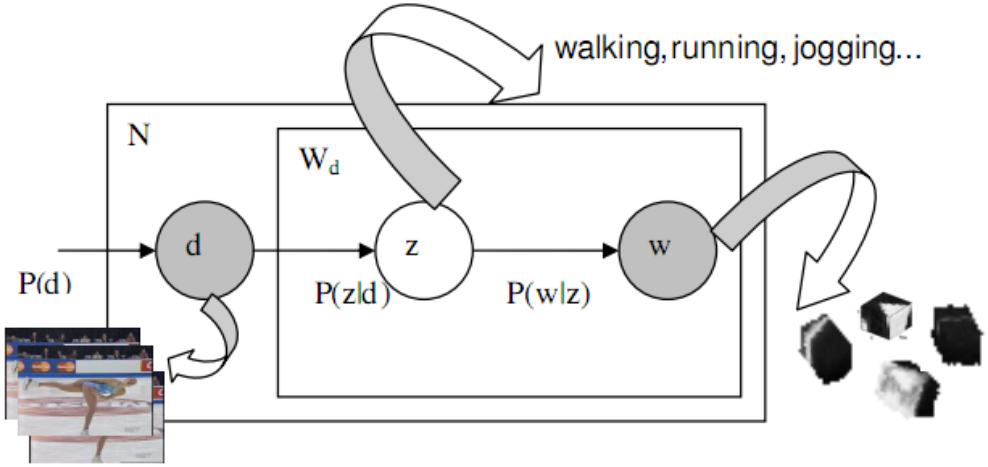
Histogram (bag-of-words) similarity





pLSA models for actions

- pLSA from text recognition
  - Probabilistic latent semantic analysis
  - Reasoning the probability of features originated from a particular action video.



[Niebles, J. C., Wang, H., and Fei-Fei, L., Unsupervised learning of human action categories using spatial-temporal words, BMVC 2006]

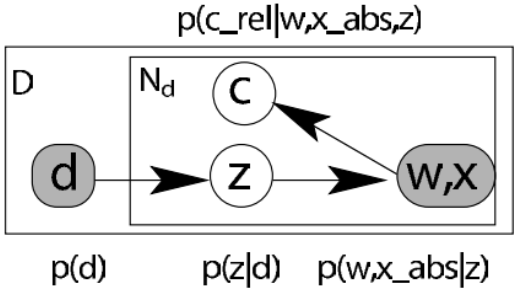
Approach overview

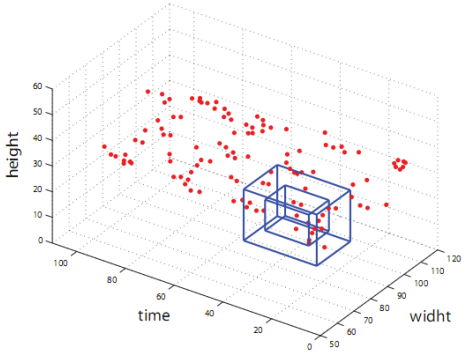
- Recognition using local spatio-temporal features
  - Bag-of-words
  - Classifiers
    - e.g. SVMs, pLSA, …

- Extensions
  - Structural considerations
  - Hybrid features
  - Grouping features

Structural considerations

- Bag-of-features ignores structure.
- Structures?
  - Wong et al. 2007
    - pLSA-ISM: encodes relative locations of features
  - Savarese et al. 2008
    - Feature correlation: pairwise proximity



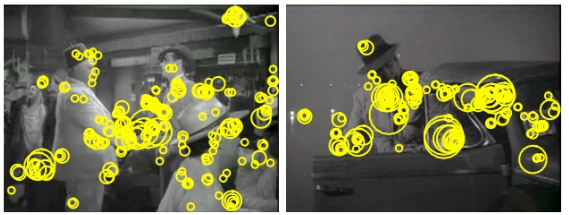


[Wong, S.-F., Kim, T.-K., and Cipolla, R., Learning motion categories using both semantic and structural information, CVPR 2007]
[Savarese, S., DelPozo, A., Niebles, J., and Fei-Fei, L., Spatial-temporal correlatons for unsupervised action classification, WMVC 2008]

Local features for movie scenes

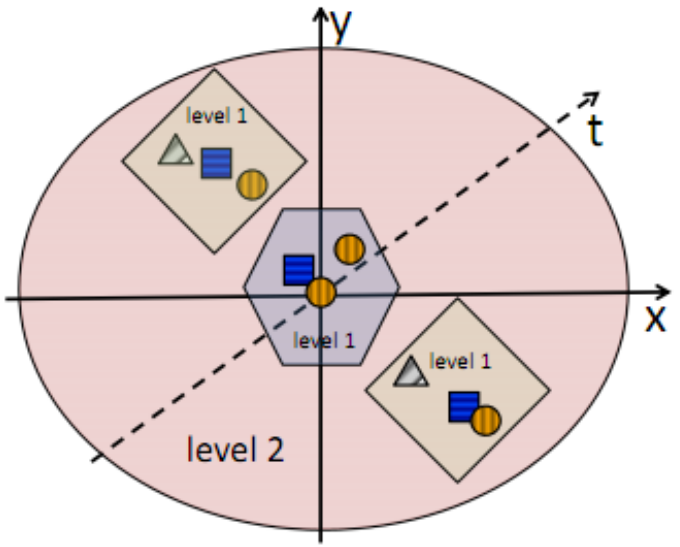
- Laptev et al. 2009 – movies
  - Movie scenes with camera movements
    - Instantaneous actions

  - Improved their local descriptor [Laptev 03] for analyzing movie videos.
    - Gradients + optical flows



Grouping features

- Groups a small number of features
  - 2~3 features which appear jointly
  - Spatially/temporally adjacent features
    - grouping
- Multiple levels
  - Hierarchical?



[Kovashka A. and Grauman K., Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, CVPR 2010]

XYT approaches: pros and cons

- Advantages
  - Robust under noise
    - Background changes, camera movements, …
    - YouTube-type videos
- Limitations
  - Bag-of-words
    - Spatio-temporal relations among features are ignored.
  - Not hierarchical
    - Difficult to model complex activities

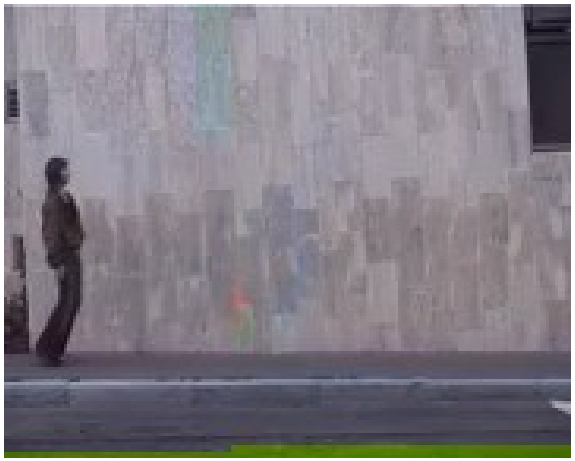
Summary: single layered

- In general, suitable for ***action*** recognition
  - Single actor
  - Structural variations?
- Handle uncertainties reliably
  - Strong to noise, background, illuminations, …
- Stochastic decision
  - Can be served as building blocks.
- A large number of training videos required.

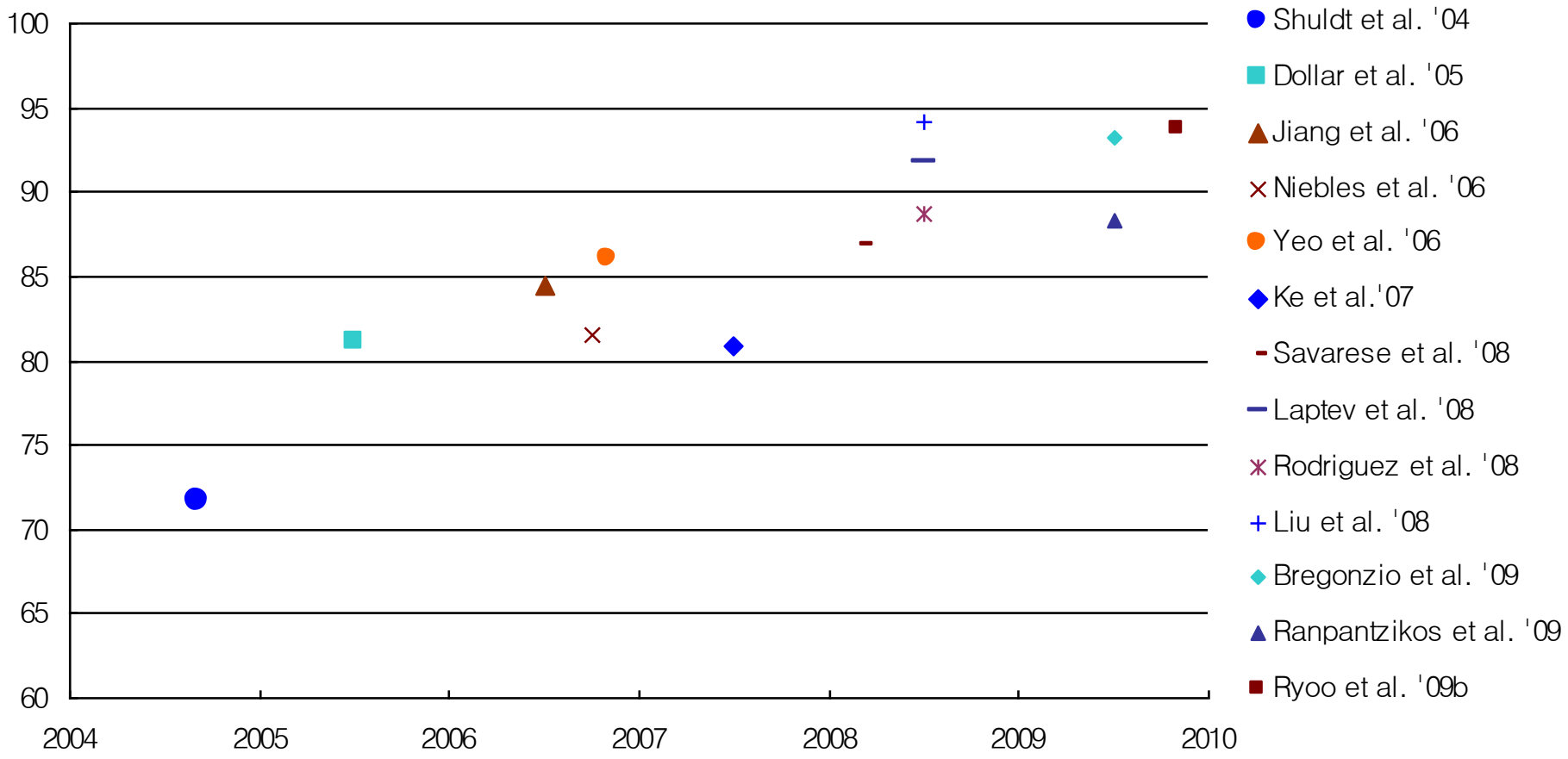
Datasets

- KTH dataset
  - Single action video classification
    - Single actor
    - One action per video

- Weizmann dataset
  - Similar to the KTH dataset (single action)



KTH results


100
95
90
85
80
75
70
65
60
2004
2005
2006
2007
2008
2009
2010
Shuldt et al. '04
Dollar et al. '05
Jiang et al. '06
Niebles et al. '06
Yeo et al. '06
Ke et al.'07
Savarese et al. '08
Laptev et al. '08
Rodriguez et al. '08
Liu et al. '08
Bregonzio et al. '09
Ranpantzikos et al. '09
Ryoo et al. '09b

New datasets

- Hollywood dataset [Laptev 07,08]
  - Movie scenes
  - Goal: recognition in complex environments
    - Moving cameras
    - Background changes
  - Action classification
    - Segmented videos
    - Atomic movements (e.g. kissing)

New datasets

- UT-Interaction dataset
  - Multiple actors
    - Human interactions
    - Pedestrians
  - Continuous videos

- UT-Tower dataset
  - Low-resolution
  - Simple actions

Single layered: References

Space-Time approaches

- Bobick, A. and Davis, J., The recognition of human movement using temporal templates. IEEE T PAMI 23(3), 2001.
- Efros, A., Berg, A., Mori, G., and Malik, J., Recognizing action at a distance, ICCV 2003.
- Schuldt, C., Laptev, I., and Caputo, B., Recognizing human actions: A local SVM approach, ICPR 2004.
- Yilmaz, A. and Shah, M., Recognizing human actions in videos acquired by uncalibrated moving cameras, ICCV 2005.
- Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S., Behavior recognition via sparse spatio-temporal features, VS-PETS 2005.
- Niebles, J. C., Wang, H., and Fei-Fei, L., Unsupervised learning of human action categories using spatial-temporal words, BMVC 2006.
- Ke, Y., Sukthankar, R., and Hebert, M., Spatio-temporal shape and flow correlation for action recognition. CVPR 2007.
- Wong, S.-F., Kim, T.-K., and Cipolla, R., Learning motion categories using both semantic and structural information, CVPR 2007
- Savarese, S., DelPozo, A., Niebles, J., and Fei-Fei, L., Spatial-temporal correlatons for unsupervised action classification, WMVC 2008.
- Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B., Learning realistic human actions from movies, CVPR 2008.

Singled-layered: References (2)

Sequential approaches

- Yamato, J., Ohya, J., and Ishii, K., Recognizing human action in time-sequential images using hidden Markov model. CVPR 1992.
- Gavrila, D. and Davis, L., Towards 3-D model-based tracking and recognition of human movement. In International Workshop on Face and Gesture Recognition 1995.
- Starner, T. and Pentland, A., Real-time American Sign Language recognition from video using hidden Markov models. International Symposium on Computer Vision, 1995.
- Oliver, N. M., Rosario, B., and Pentland, A. P., A Bayesian computer vision system for modeling human interactions. IEEE T PAMI, 2000.
- Park, S. and Aggarwal, J. K., A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, 2004.
- Natarajan, P. and Nevatia, R., Coupled hidden semi Markov models for activity recognition. WMVC 2007.