Understanding Human Motion: Recognition and Retrieval of Human Activities

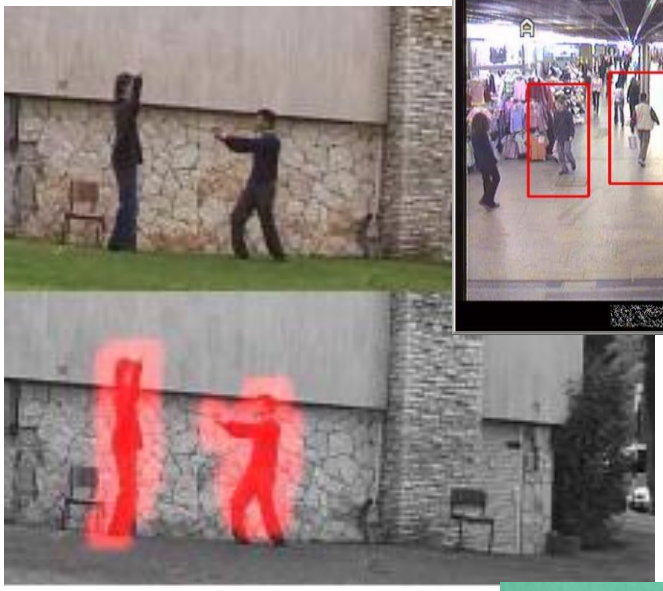
Nazlı İkizler
Bilkent University

PhD dissertation

May 2008

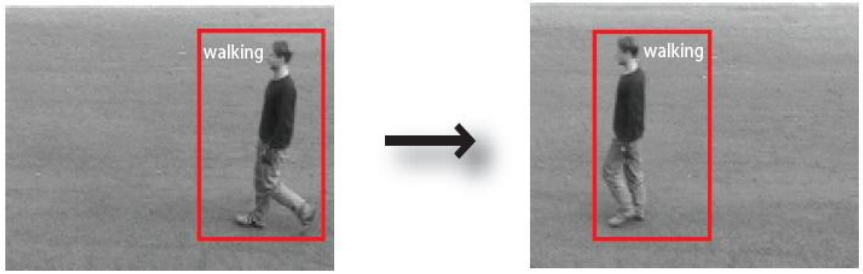
Why understand human motion?

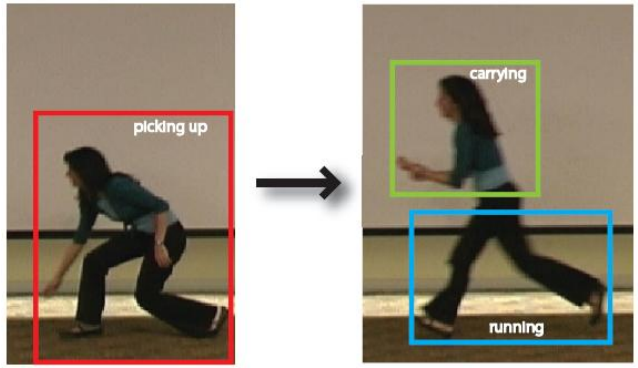
- Better security systems
  - search unusual events
- Surveillance
  - evaluate patient actions
  - evaluate criminal actions
- Gesture based interfaces
  - Games, interactive PCs, house devices
- Sign language
- and more…



Outline of this talk

- Related work on human motion understanding
- Single human actions
  - Silhouette presence
    - Histogram of Oriented Rectangles
  - Silhouette absence
    - Line and flow histograms
  - Still images
  - Results
- Complex human actions
  - Separate limbs models on HMMs
  - 2D-3D lifting
  - View-invariance
  - Results
- Conclusions and Future Work





Act, Action and Activity

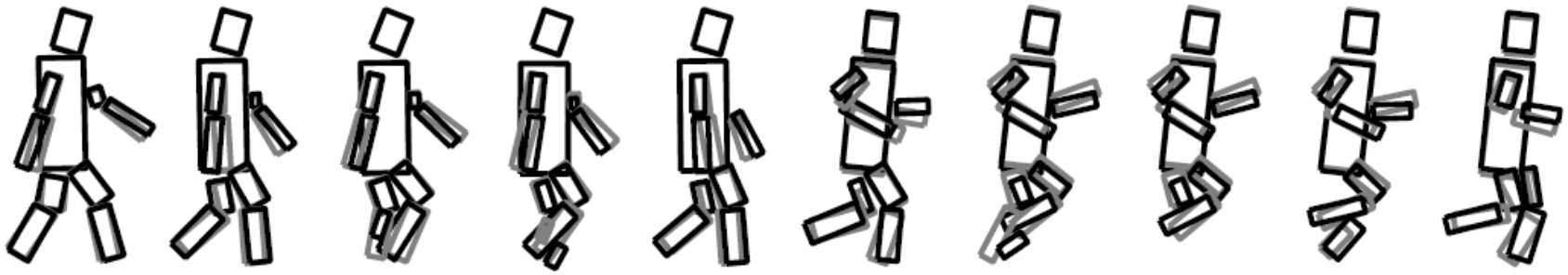
- Act: Short-timescale representations like a *forward-step* or a *hand-raise*
- Action: Medium timescale movements like *walking*, *running*, *jumping*
  - Typically composites of multiple acts
- Activity: Long timescale movements
  - Complex composites of actions
  - Composition can be
    - across time
    - across body

Related Work on Human Motion Understanding

- Three basic threads
  - Motion Primitives
  - Methods with Dynamical Models
    - Explicit dynamical models
    - Partial dynamical models
  - Discriminative Methods
    - Templates
    - Bag-of-features

Motion Primitives

- Motion primitives: Clusters of motion of the same type
- Form motion primitives on short timescales and analyze how they're strung together within action
- Feng and Perona (Feng 2002) use movelets for this purpose, and construct *two-frame movelets*. They model the ordering of the movelets by HMMs.



- Advantage: Fewer parameter estimation, actions can share primitives

Explicit Dynamical Models

- Hidden Markov Models
  - Linear, Coupled, Layered, Factorial, Parallel HMMs
- Finite State Methods
  - Hongeng et al, coarse scale person activities
  - Zhao and Nevatia, finite state model for walking, running and standing
  - Hong et al, model gesture via finite state models
- Conditional Random Fields

Partial Dynamical Models

- Pinhanez and Bobick: temporal relations between primitivies (using linear interval algebra).
- PNF network: Past-Now-Future
- Describe actions using PNF propagation.
- Infer relations from detector response (they used responses from simulated detectors)

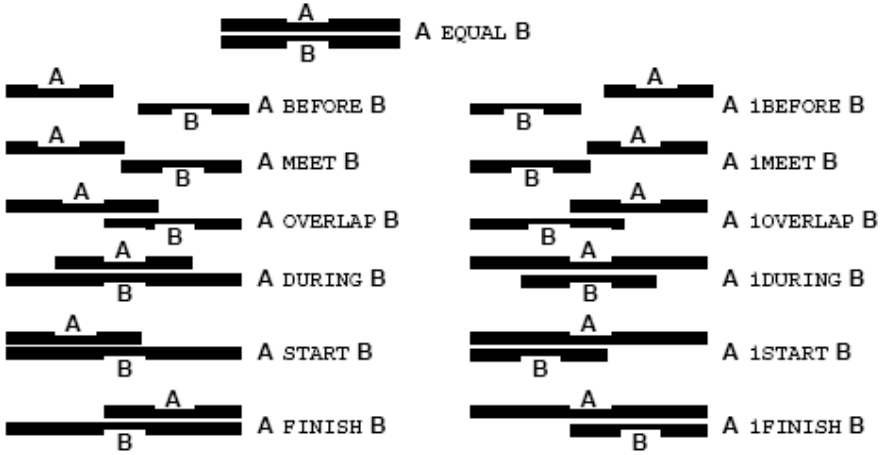


Figure 1: The possible 13 primitive time relationships

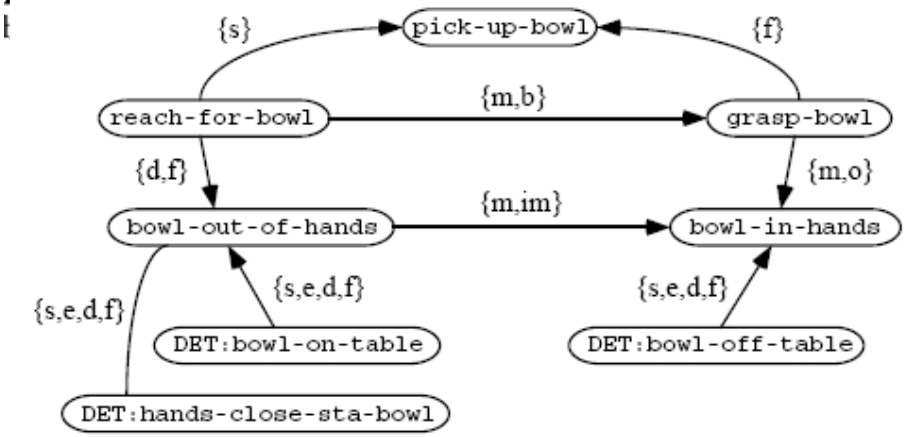
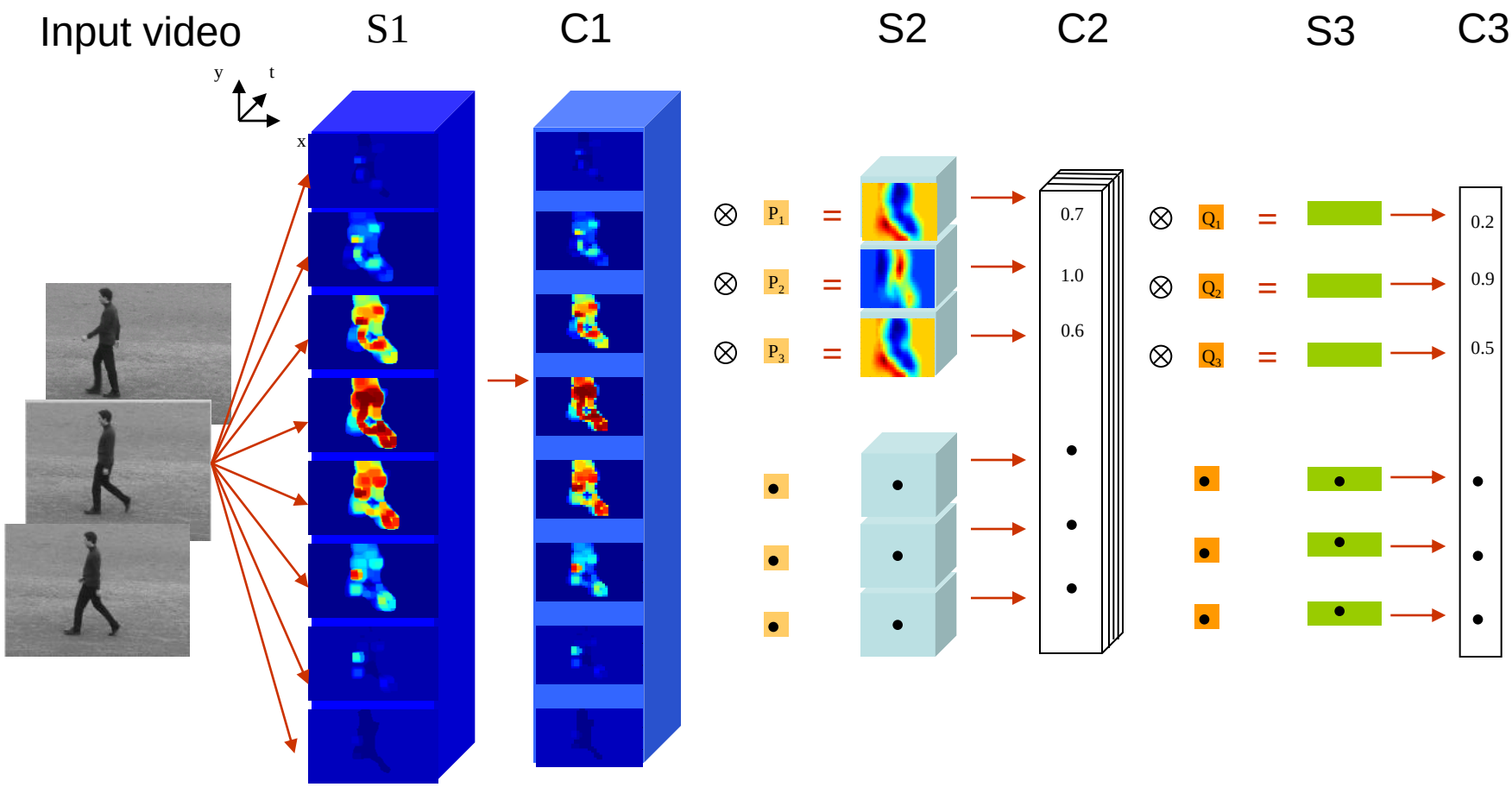


Figure 2: IA-network corresponding to the temporal structure of a "pick-up bowl" action .

(Spatio-temporal) Templates

S ------selectivity

C ------invariance



Jhuang et al ICCV 2007

Template matching with spatio-temporal filters

Local max over each S1 map (tolerance to position)

Template matching with stored prototypes -->S2 maps

Global max over each S2 map ---(invariant to position)

Template matching with stored prototypes ->S3 maps

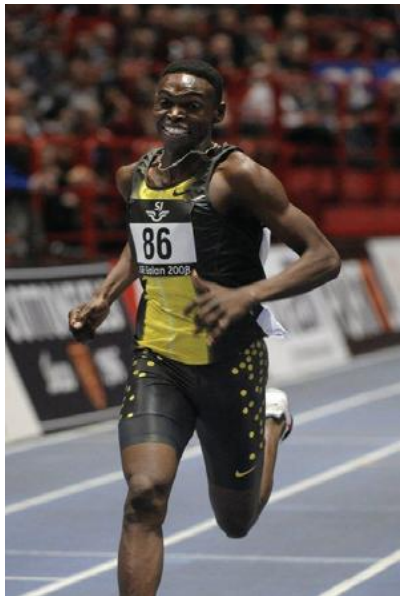
Global max over each S3 map (invariant to shift in time)

Single Action

“One actor, one action, simple background” paradigm

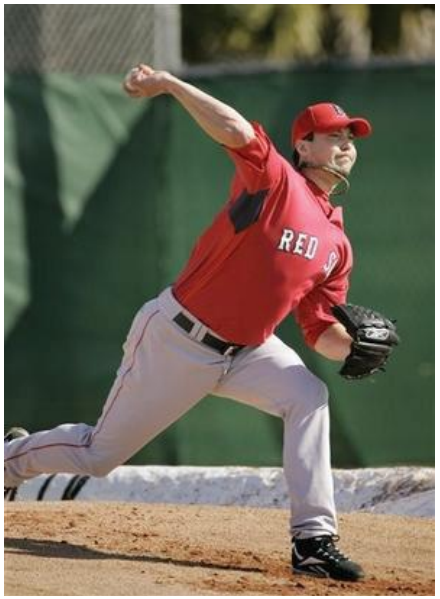
- H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV 2007*.
- Saad Ali, Arslan Basharat, and Mubarak Shah, Chaotic Invariants for Human Action Recognition. In *ICCV 2007*.
- S.-F. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR 2007*.
- J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR 2007*.
- L. Wang and D. Suter. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *CVPR* 2007
- J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatio-temporal words. In Proc *BMVC, 2006*.
- P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS 2005*.
- C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR 2004*.

What do these people do?



running

walking



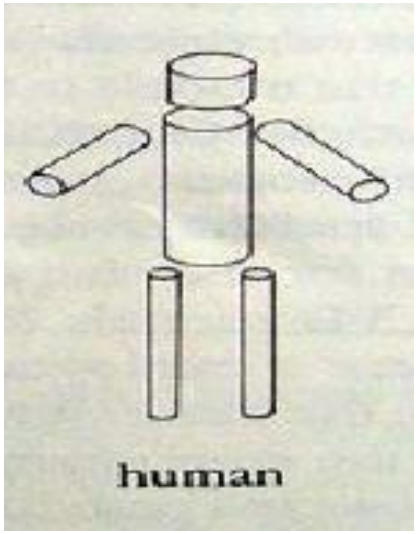
throwing

crouching

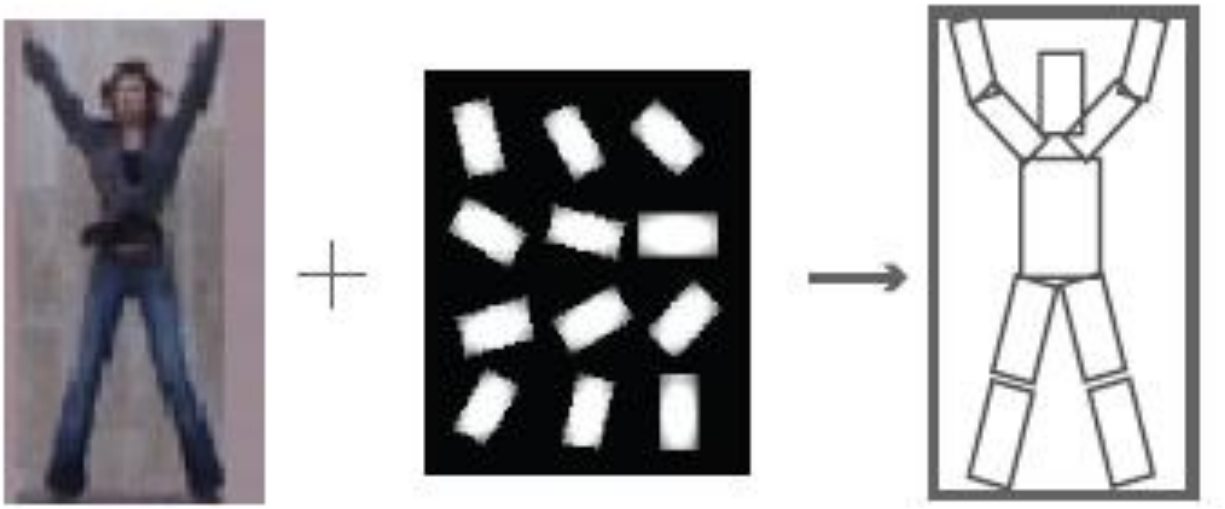
- Pose tells a lot about the actions.
- How can we describe the pose?

Pose as a Collection of Rectangles

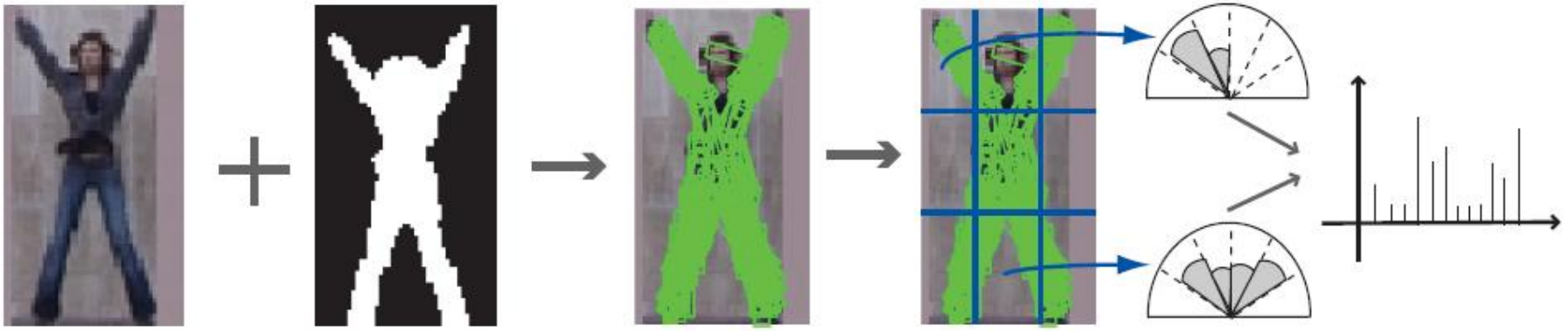
- Human body is composed of cylindrical parts.
- The projection of a cylinder on 2D is a rectangle.
- Body can be thought as a collection of rectangular regions
- We can represent the pose based on the orientation of these rectangles



David Marr's Theory of Vision (1982)



Histogram of Oriented Rectangles (HOR)

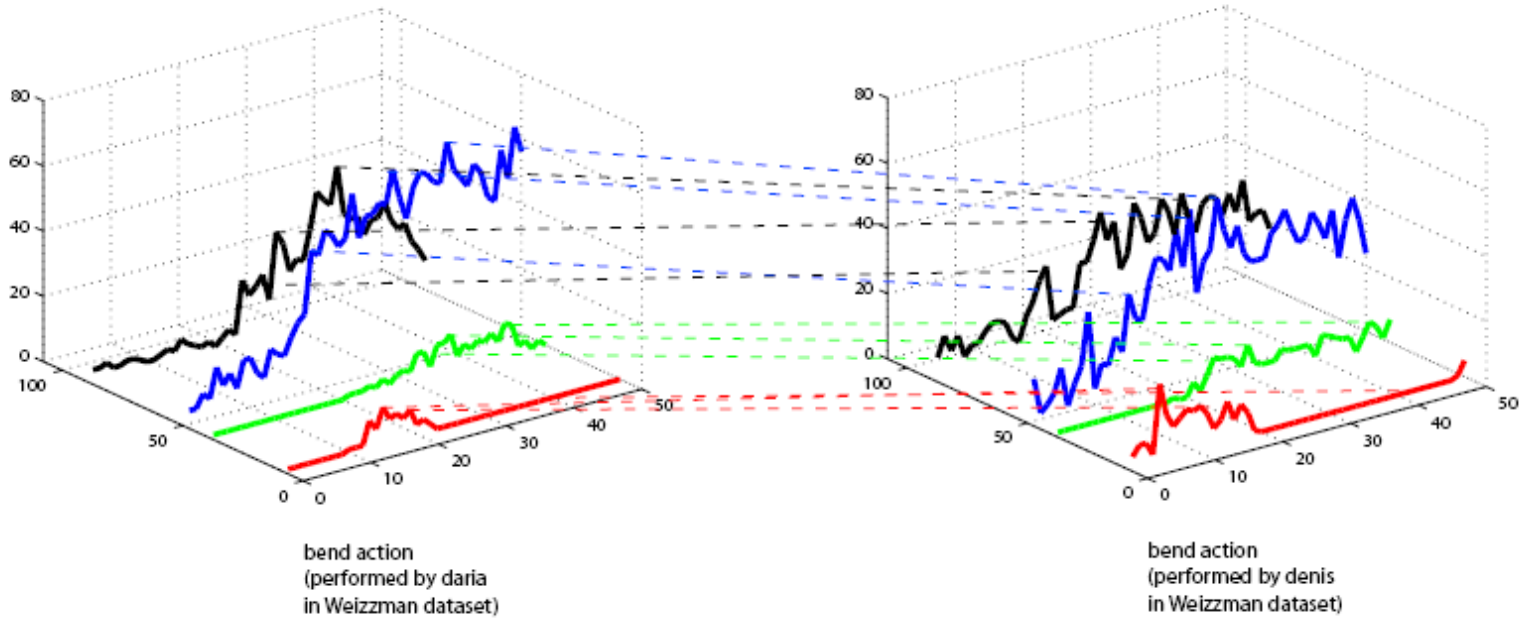


- Tracker finds the human subject
  - Extract the silhouettes
- Rectangular regions are extracted using convolution of a zero-padded rectangular 2D Gaussian on different orientations and scales
  - 12 angles 15 apart

Recognition

- We have utilized various methods for recognition
  - Template Matching
    - Nearest Neighbor
    - Global Histogramming
  - Sequence Matching
    - Dynamic Time Warping (DTW)
  - Discriminative Classification
    - Support Vector Machines (SVM)

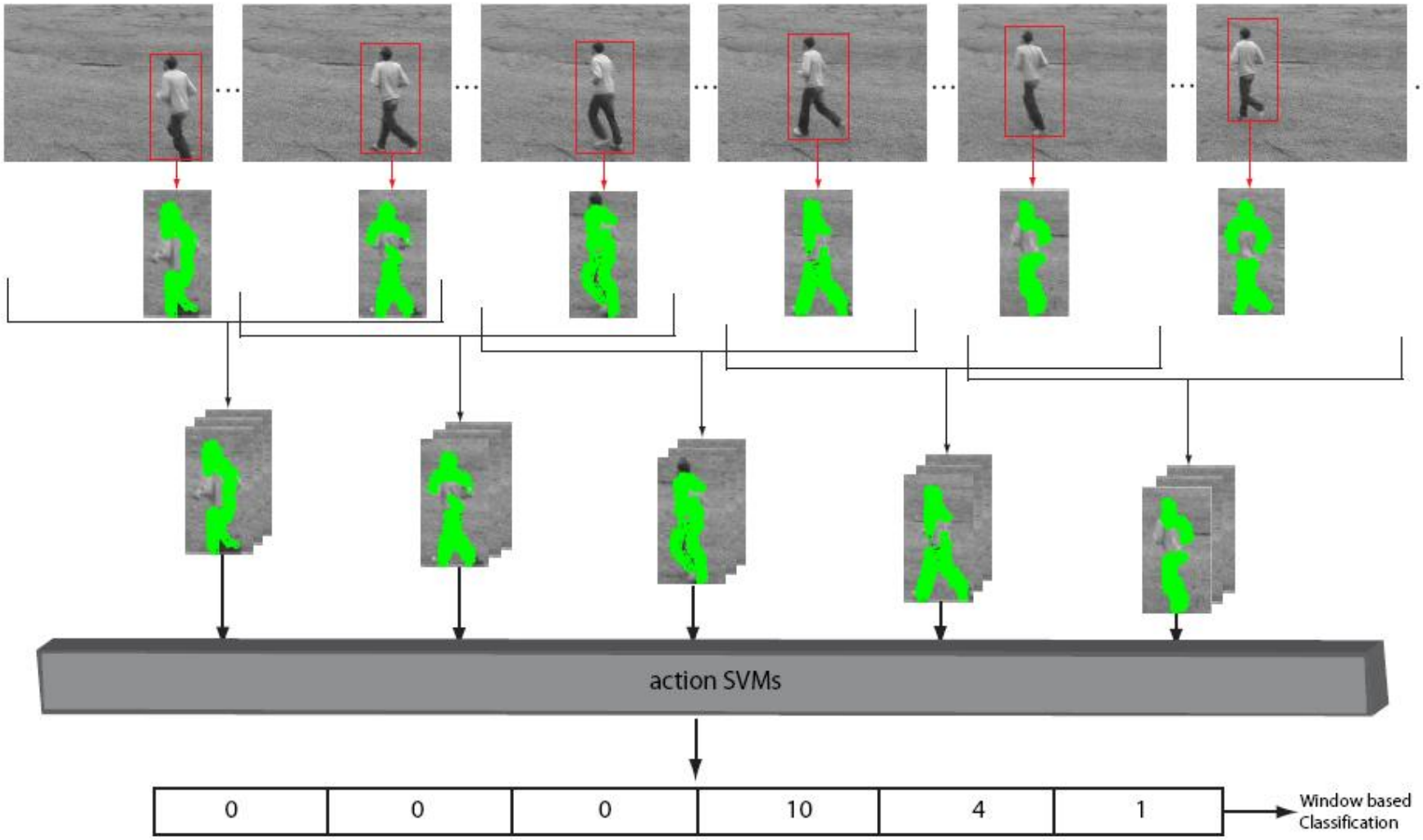
Dynamic Time Warping



- Align each of the features using DTW
- Compute the global cost

$$\chi^2(H_i, H_j) = \frac{1}{2}\sum_n \frac{(H_i(n) - H_j(n))^2}{H_i(n) + H_j(n)}$$

Support Vector Machines



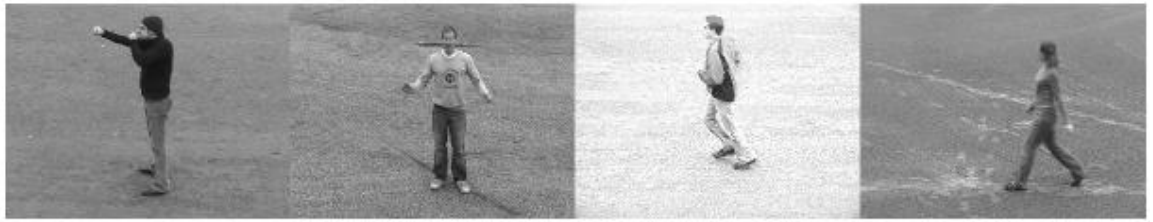
- Use snippets of frames and form histogram of oriented rectangles over a window (HORW)

Experimental Setup

- Weizzman Dataset
  - 9 actions(*walk, run, jump, jump in place, jump jack, gallop sideways, bend, one-hand wave, two-hands wave*)
  - 9 subjects

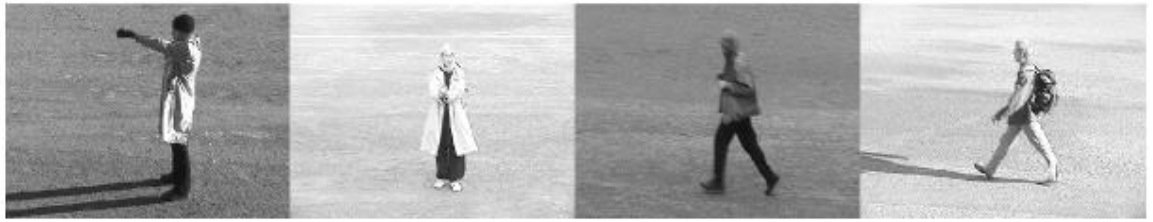
Experimental Setup

- KTH dataset
  - 6 actions (boxing, handclapping, handwaving, jogging, running, walking)
  - 25 subjects
  - 4 recording conditions (outdoor, indoor, zoom and viewpoint change, carry items and different outfits)



(a) s1 condition: outdoor(standard recording)

(b) s2 condition: zoom effect and different viewpoints



(c) s3 condition: different outfits and carry items

Experimental Evaluation of HOR

Classification Method	Feature	Weizzman	KTH
NearestNeighbor	HOR	96.30%	75.46%
	HORW	97.53%	72.22%
GlobalHist	HOR	96.30%	71.76%
	HORW	69.14%	57.41%
SVM	HOR	97.53%	77.31%
	HORW	95.06%	85.65%
DTW	HOR	**100%**	74.54%
	HORW	96.30%	78.24%
v+SVM	HOR	98.77%	81.48%
	HORW	95.06%	**89.35%**
v+DTW	HOR	**100%**	81.02%
	HORW	98.77%	83.8%

- For KTH, v+SVM performs better, because DTW surpasses time difference and makes running and jogging closer.
- Choosing window usage (HOR vs HORW) depends on the nature of the actions one wishes to discriminate.

Experimental Evaluation of HOR

Method	Accuracy
HOR	**100%**
Blank et al. [12]	99.64%
Jhuang et al. [48]	98.8%
Wang et al. [96]	97.78%
Niebles et al. [63]	72.8%

Comparison to other methods on the Weizzman dataset

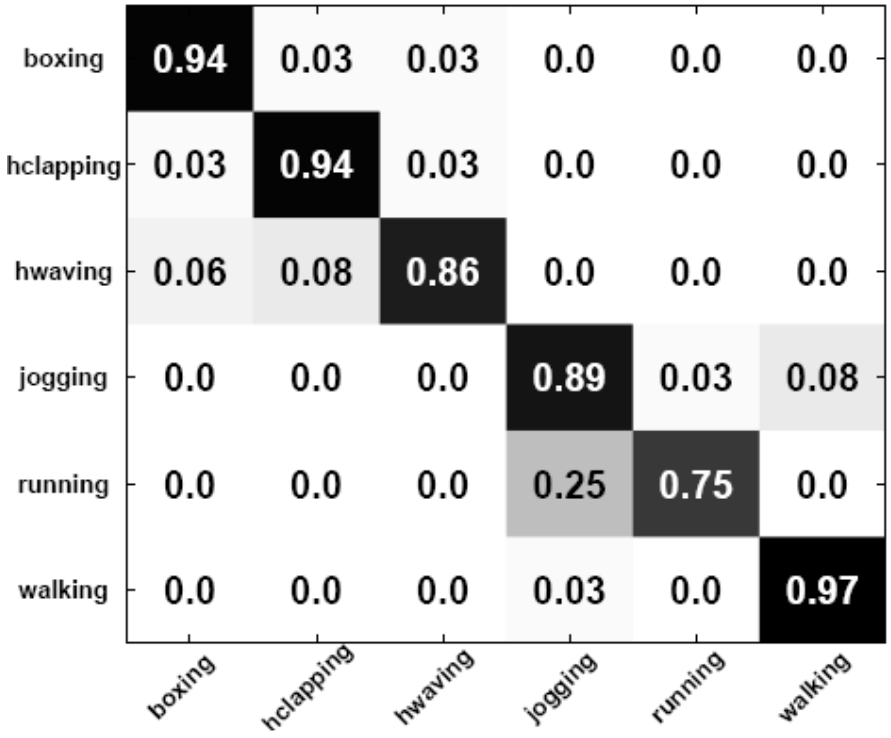
Method	Accuracy
Jhuang et al. [48]	91.7%
Wong et al. [100]	91.6%
HORW	**89.4%**
Niebles et al. [64]	81.5%
Dollár et al. [24]	81.2%
Ke et al. [50]	80.9%
Schuldt et al. [84]	71.7%

Comparison to other methods on the KTH dataset

Comparison to HOGs on the KTH

	HOG	HOR	HORW
SVM	76.85%	77.31%	85.65%
DTW	67.59%	74.54%	78.24%
v+SVM	82.41%	81.48%	89.35%

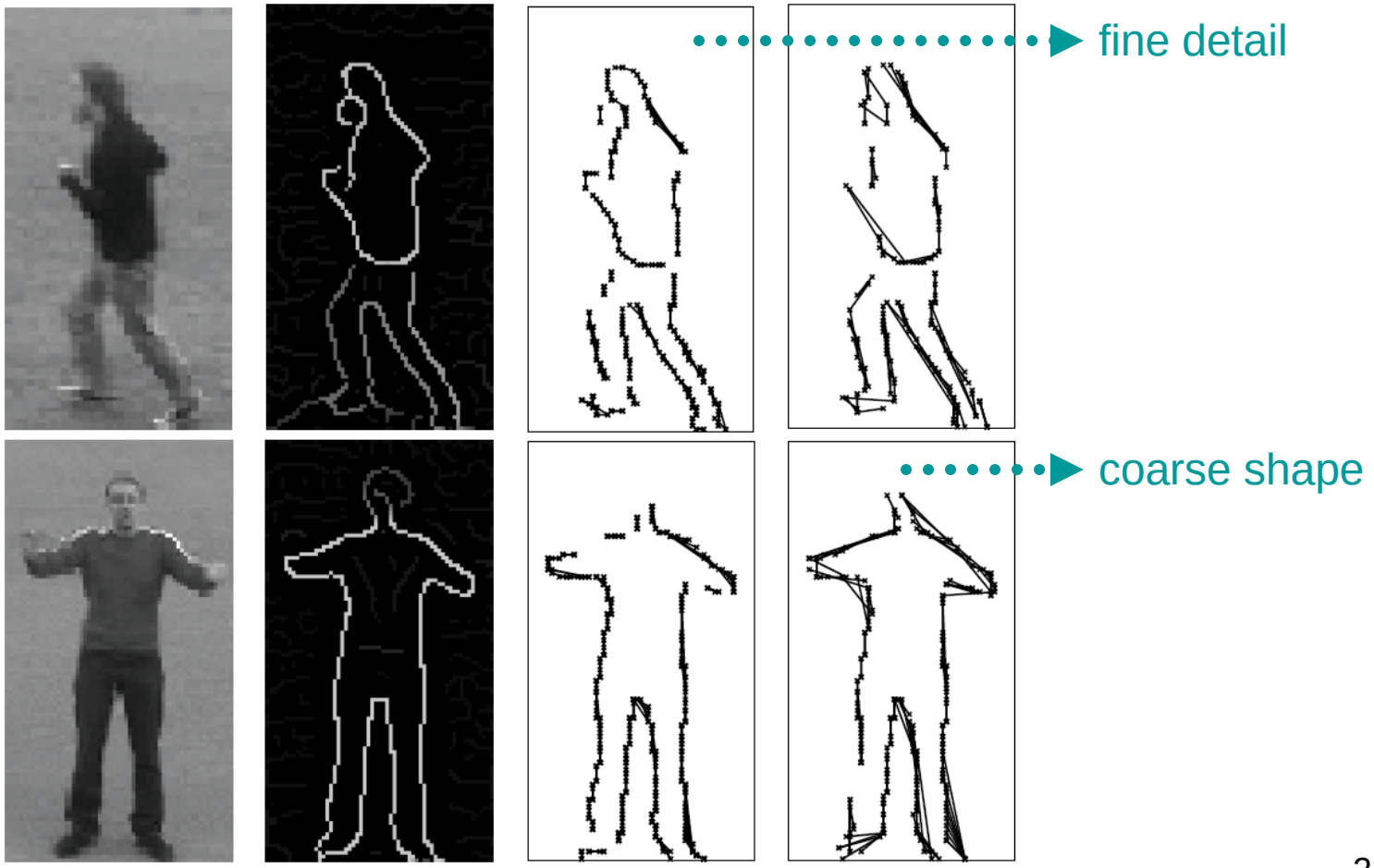
Experimental Evaluation of HOR



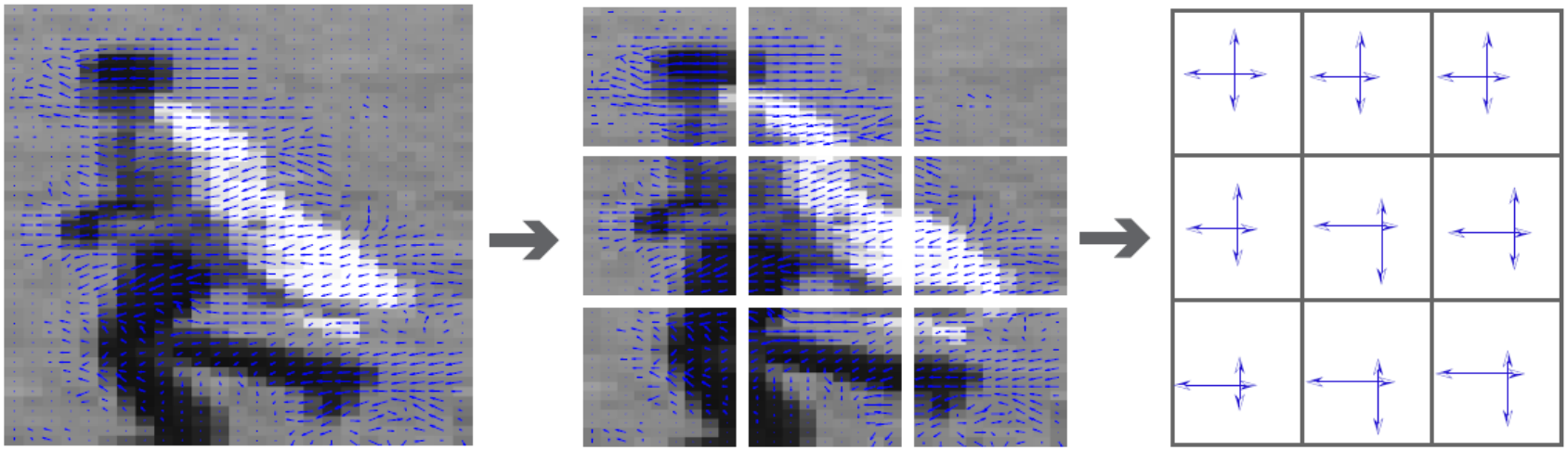
Confusion matrix for the KTH dataset

Boundary-fitted Lines

- In the absence of silhouettes, we can use lines fitted to the boundaries (Pb) (Martin PAMI2004) of human figures

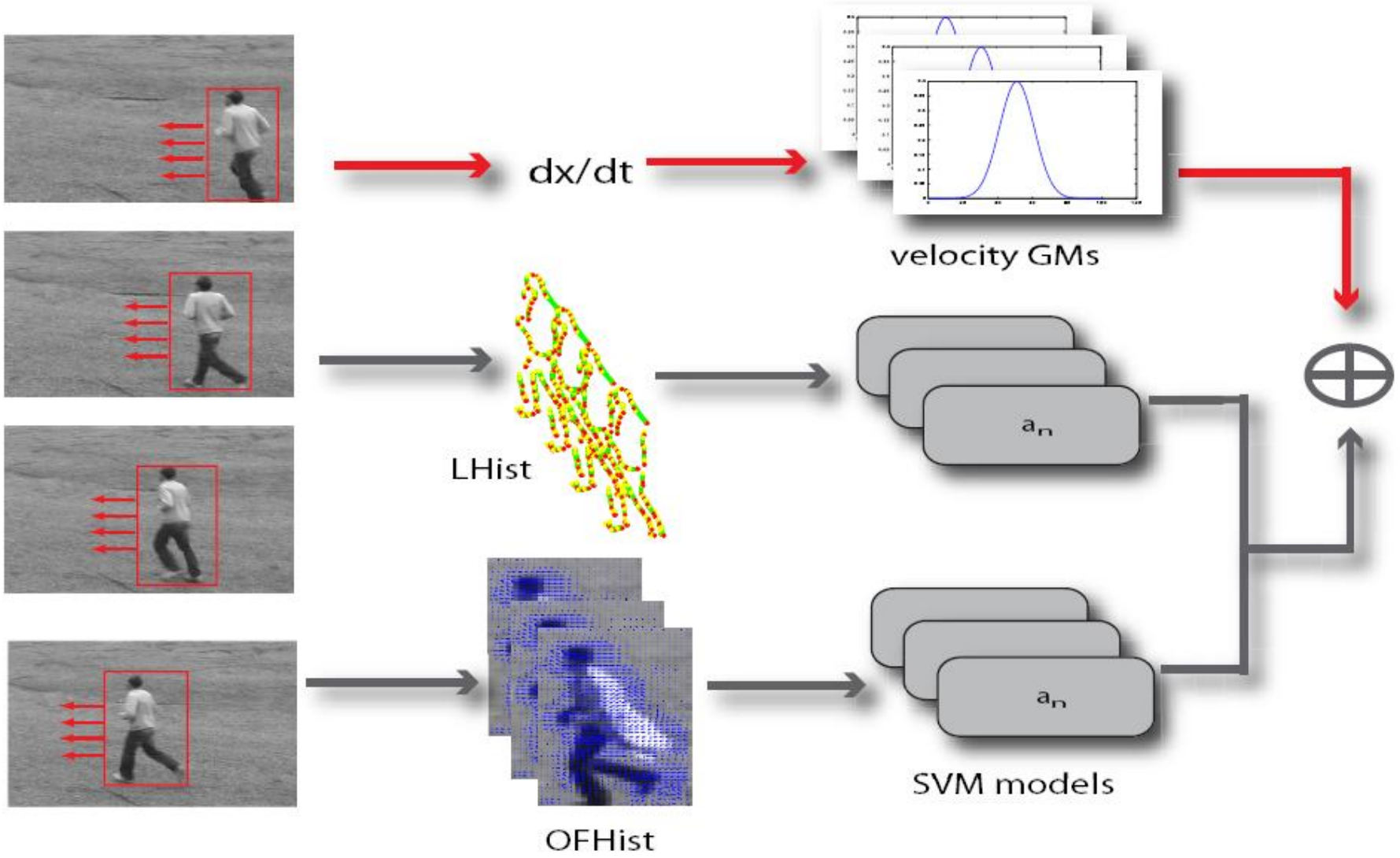


..and Optical Flow

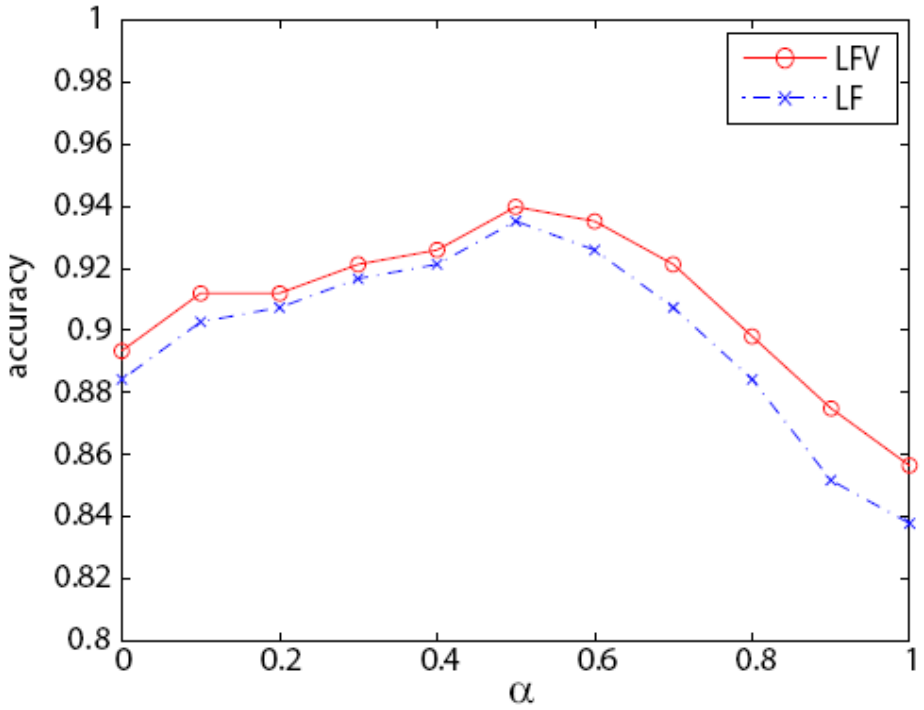


- Dense block-based optical flow calculation
  - L_1 block distance
  - 5x5 template size with a window size of 3

Recognition with LHist and OFHist



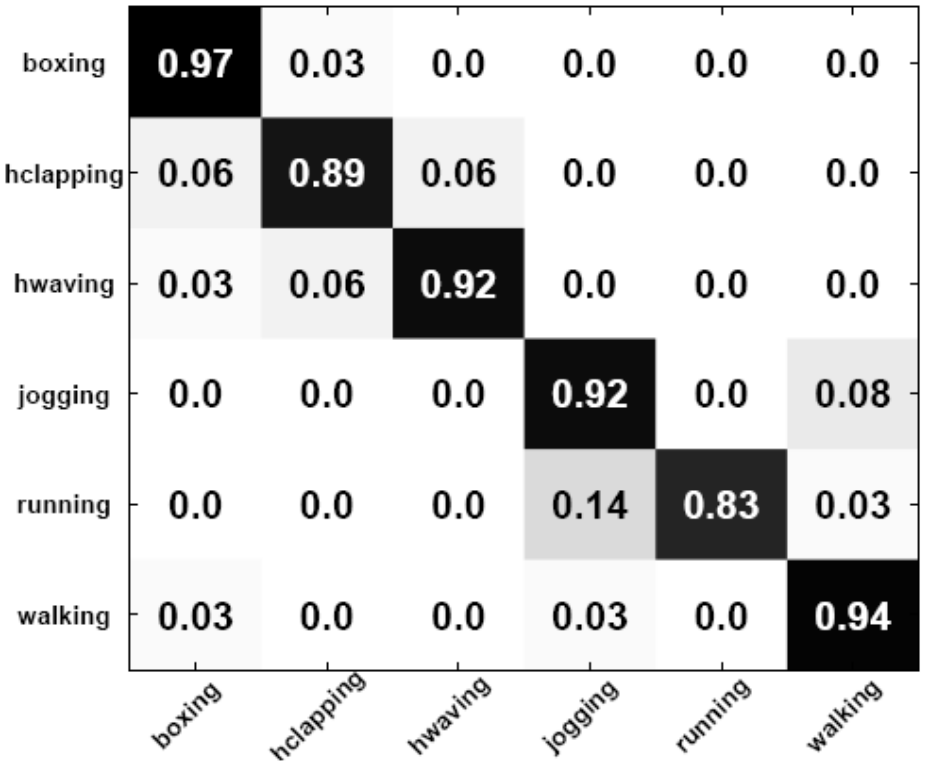
Line and Flow Results



- Choice of α in combining line and flow classification vectors

$$\mathbf{c}_f = \alpha\, \mathbf{c}_s + (1-\alpha)\mathbf{c}_m$$

Line and Flow Results



Method	Accuracy
LFV	**94.0%**
Jhuang [48]	91.7%
Wong [100]	91.6%
Niebles [64]	81.5%
Dollár [24]	81.2%
Ke [50]	80.9%
Schuldt [84]	71.7%

Condition	LFV	Jhuang [48]
s1	**98.2%**	96.0%
s2	**90.7%**	86.1%
s3	88.9%	**89.8%**
s4	**98.2%**	94.8%

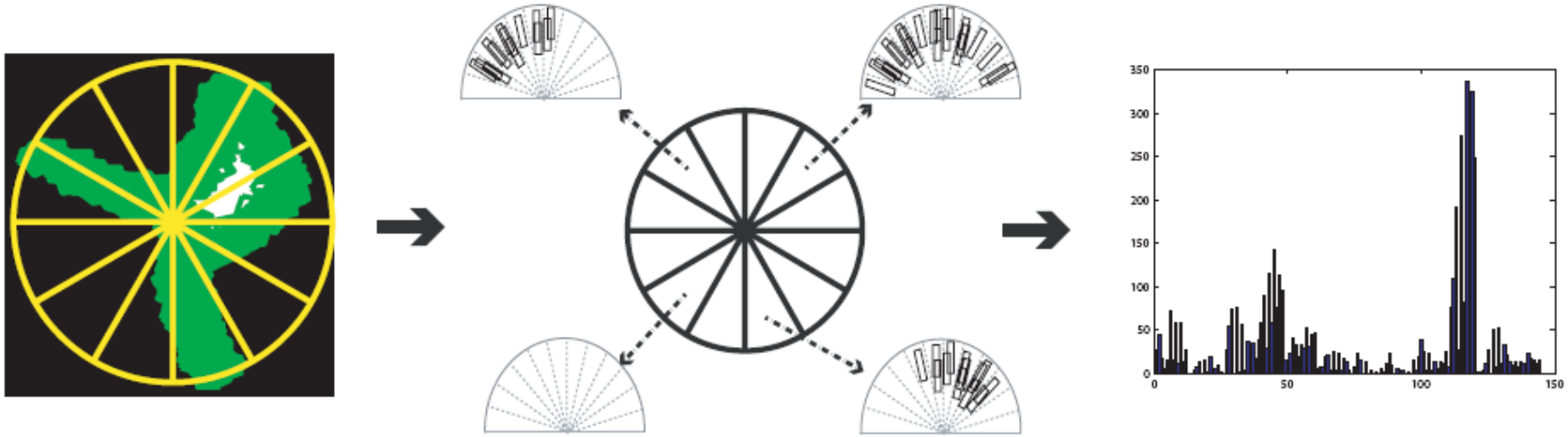
- Shape and flow are complimentary to each other.
- Again, this depends on the nature of the actions in mention.

Action Recognition in Still Images

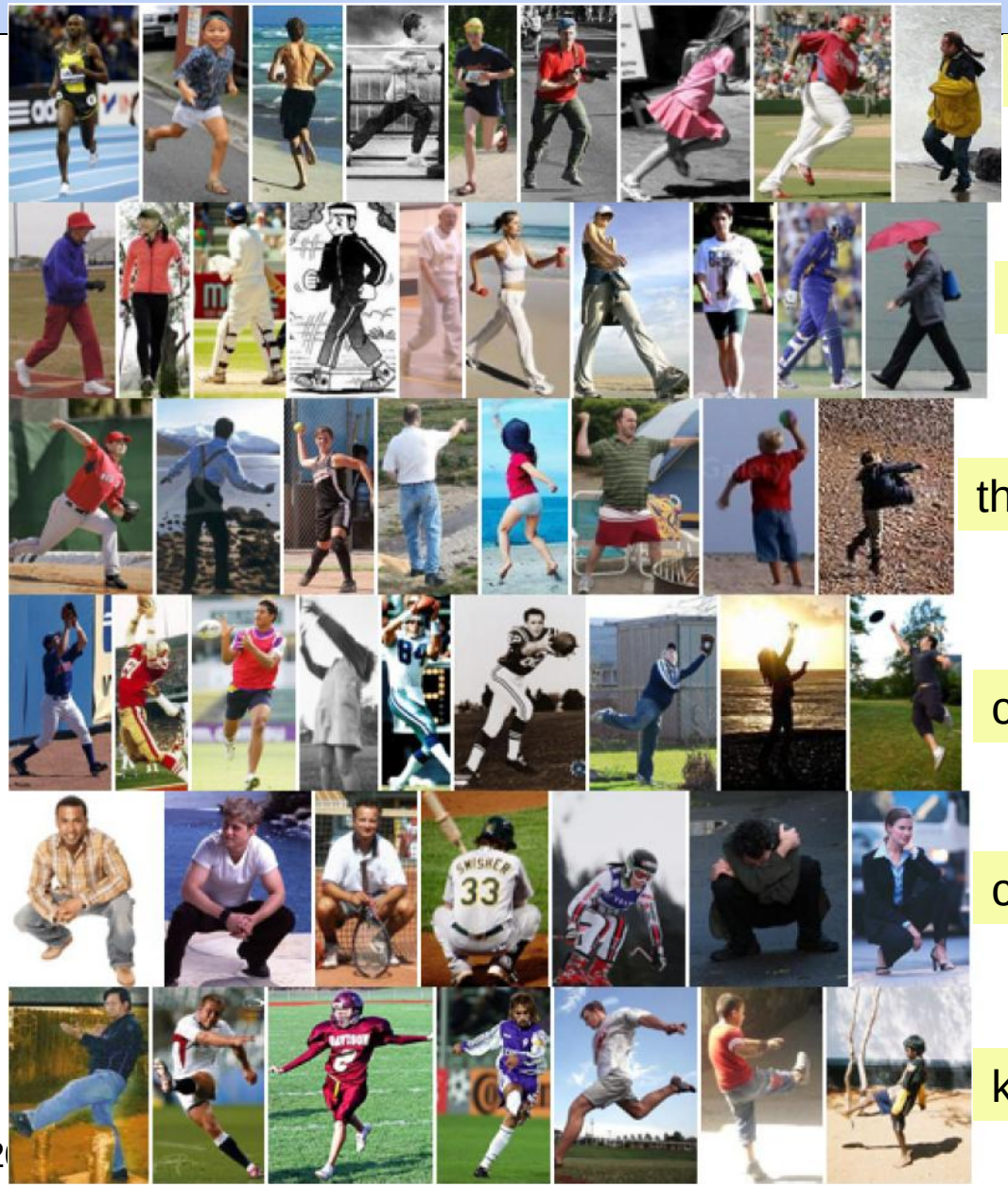
- Never done before
- First identify the person
  - Pose estimation by Ramanan based on CRFs.
- Extract the silhouette
- Use Histograms of Oriented Rectangles

Recognition by CHORs

- Form Circular HORs (CHORs)
  - Don’t have the bounding box, therefore use the maximum probability of parse as the center and use a polar grid.
- Classification based on LDA+SVM
  - We have lesser examples, so a discrimination on feature set is useful : Use Linear Discriminant Analysis (LDA)
  - SVM with Radial Basis Functions



Still Image Results



running

walking

throwing

catching

crouching

kicking

ActionWeb dataset - 467 images collected from the web

Correctly classified action images

Still Image Results

(a) catch, walk, catch, throw

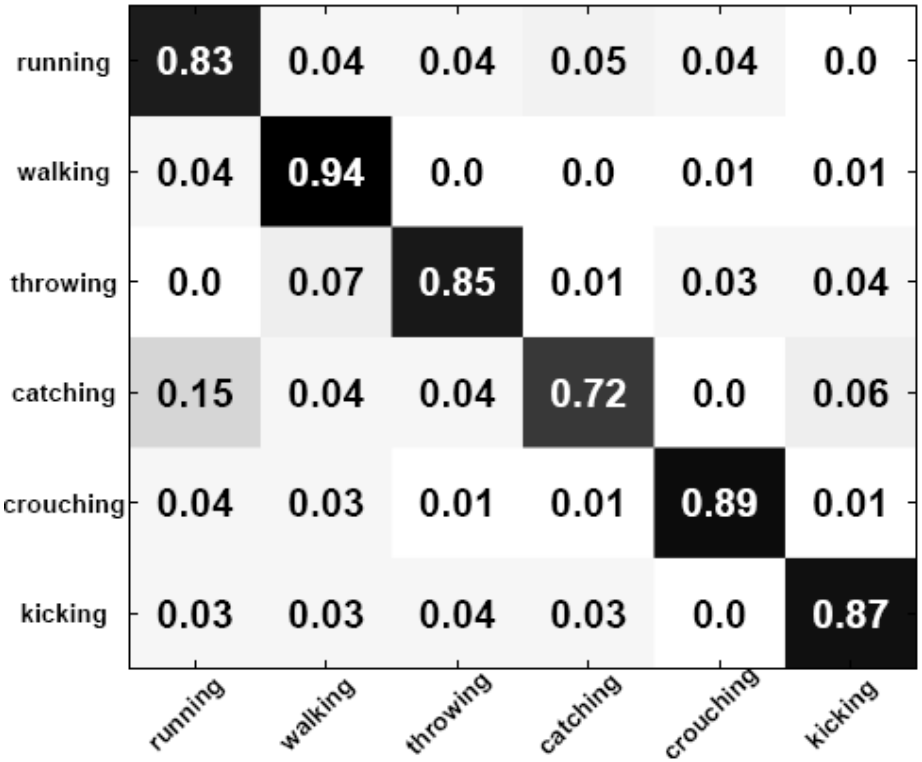
(b) run, run, run, kick

(c) catch, kick, walk, crouch

(d) run, throw, run, run

(e) kick, walk, walk, catch

(f) throw, walk, run, throw



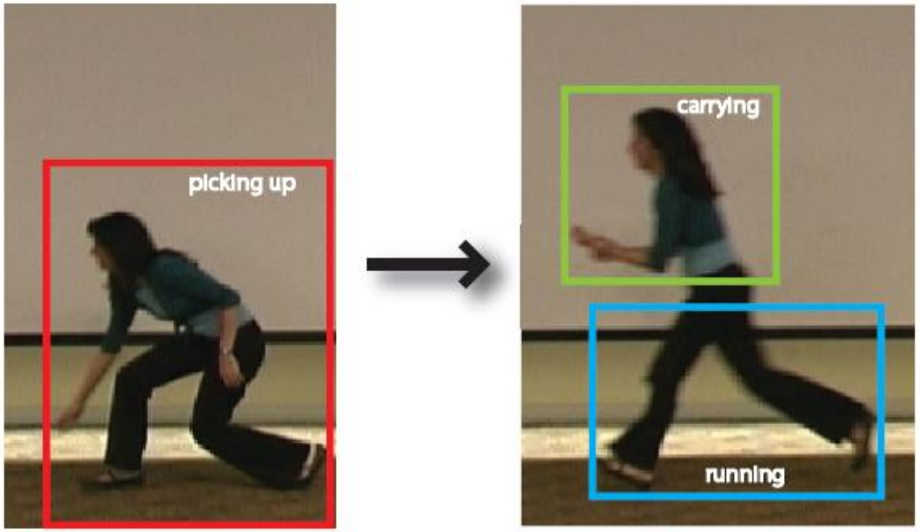
Total accuracy 85.1%

Misclassified action images

Still Image Clustering

Kmeans results with k = 100

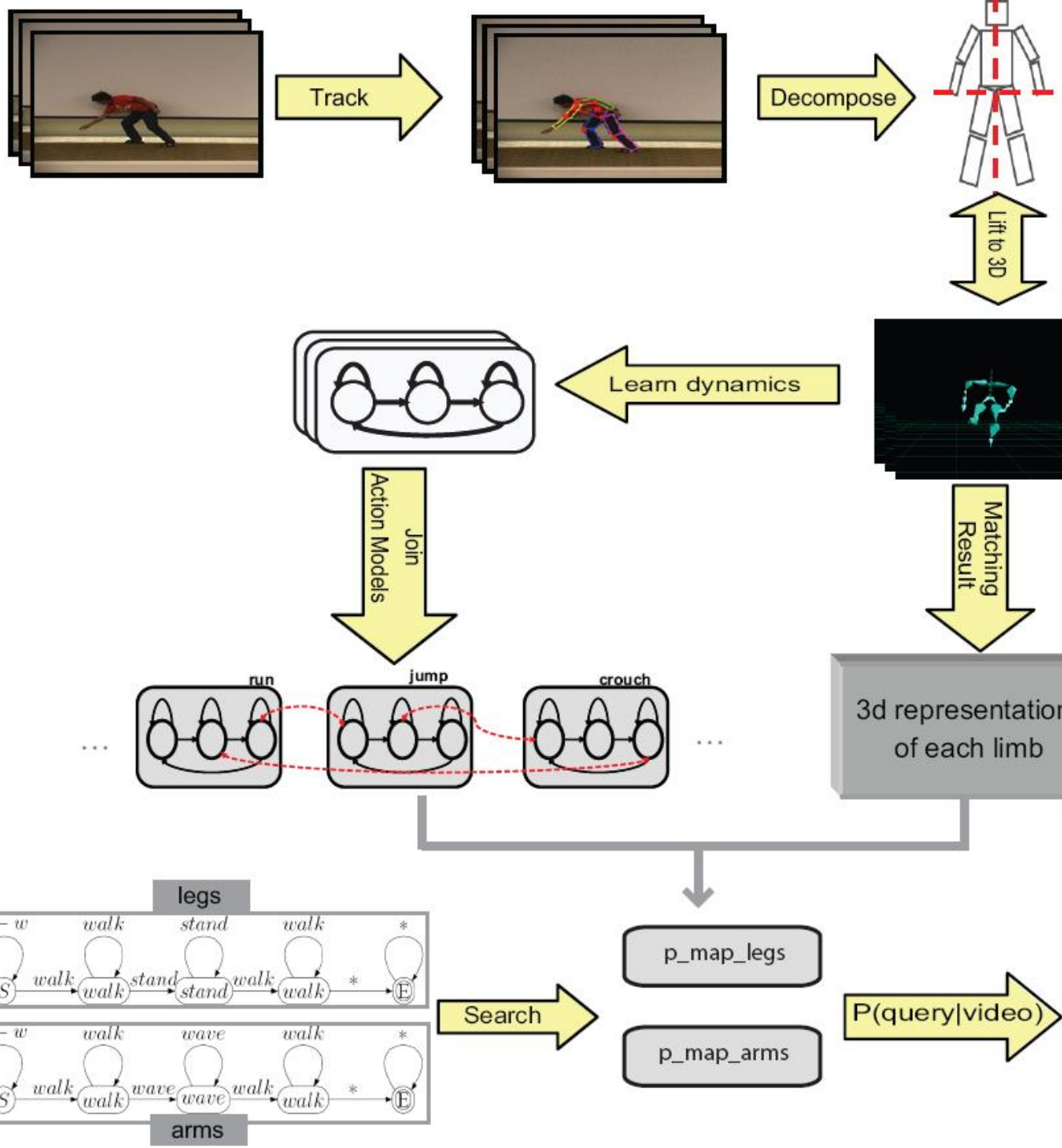
Complex Composite Activities



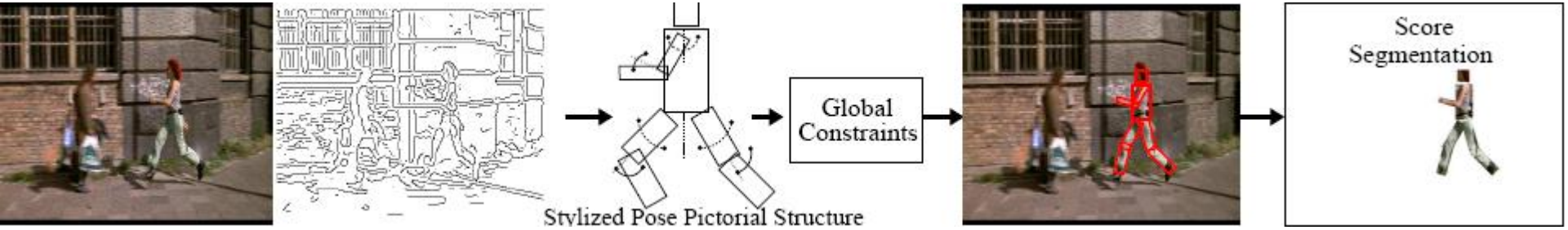
Complex Composite Activities

- Template models will not scale well to complex composite activities.
  - Will need lots of training data for every combination
- We need a generic method to represent composite human activities
  - Also handle view-invariance, which is an inherit property of everyday activity.
- We want to make videos searchable, even for complex composite actions
  - Capability to search without examples
  - Simple and effective query language
- Composition across time and across space is possible by using activity segments of the body

Overall System Architecture



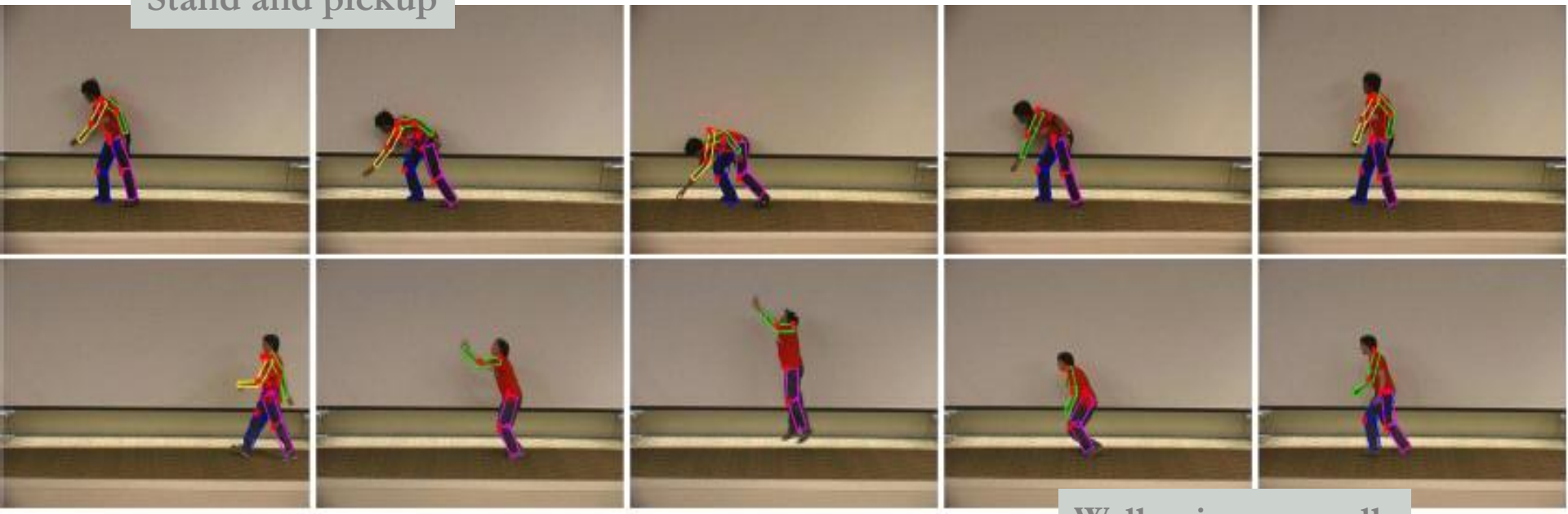
Tracking



- Tracker of Ramanan et al.
- Appearance based tracker,
  - initiates itself by finding a lateral walking pose,
  - then builds a quadratic linear regressor over the limb segments.
- Dynamic programming over the MAP estimates of candidate limbs gives the best body configuration

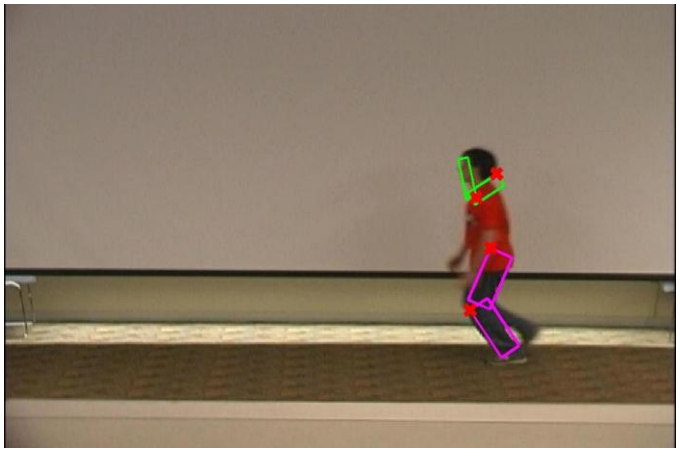
Tracking (good examples)

Stand and pickup

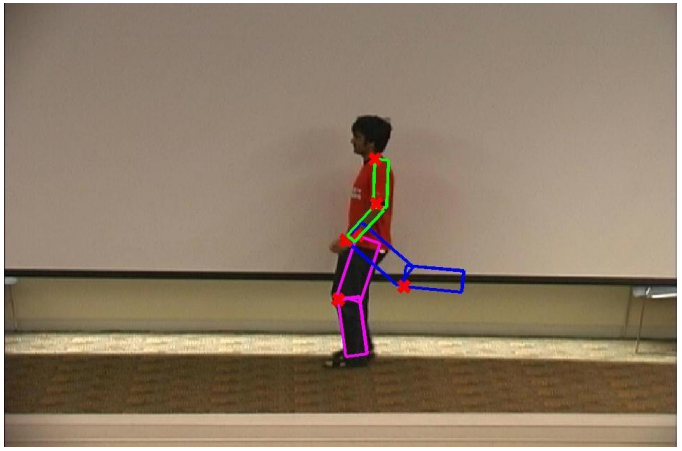


Walk – jump - walk

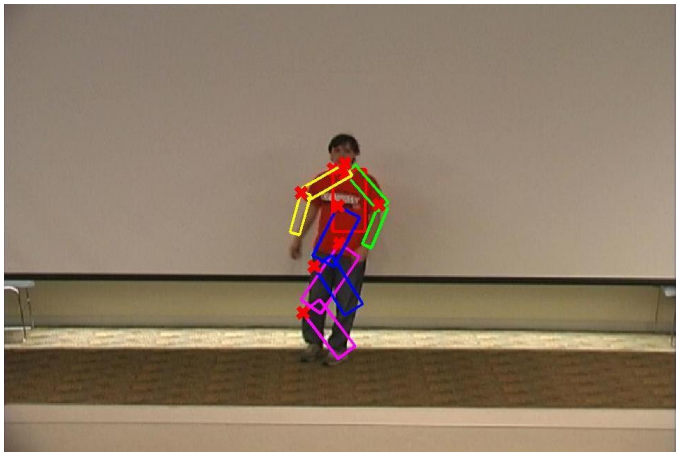
Not a perfect world



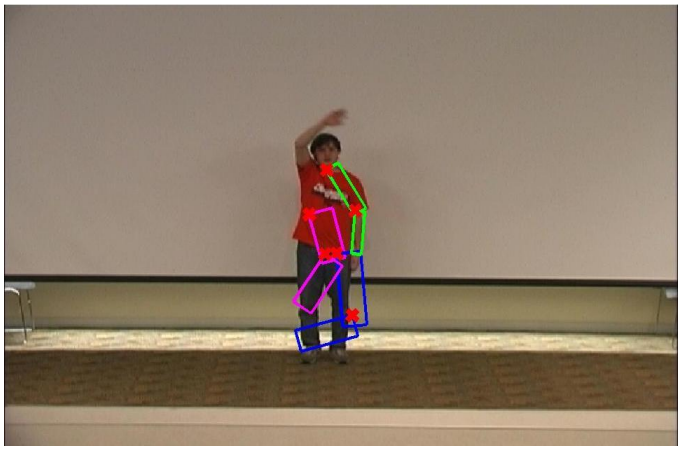
Appearance and motion blur



Occlusion of limbs

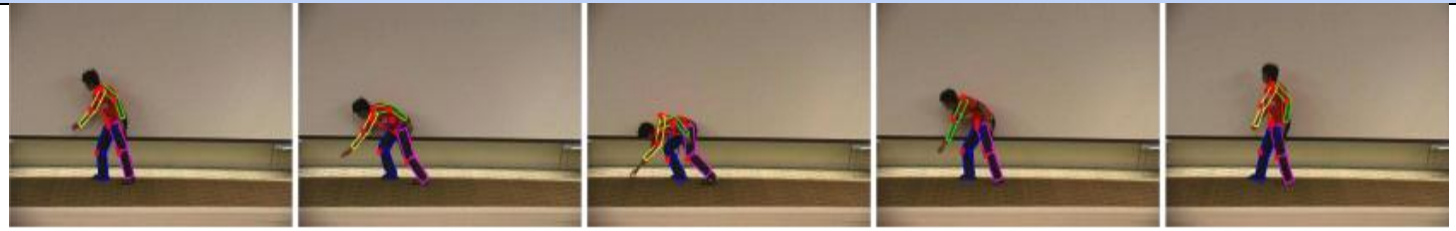


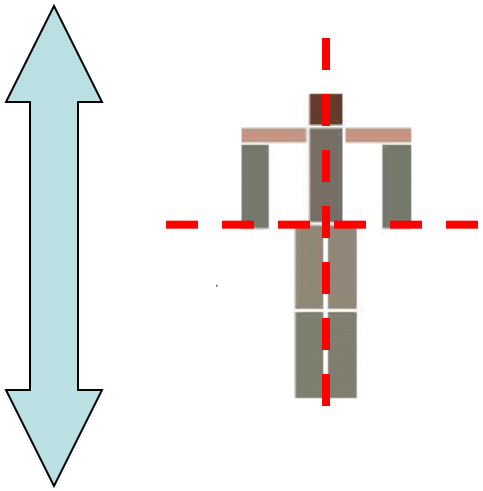
Rectangle search failure



Motion blur, bad classification

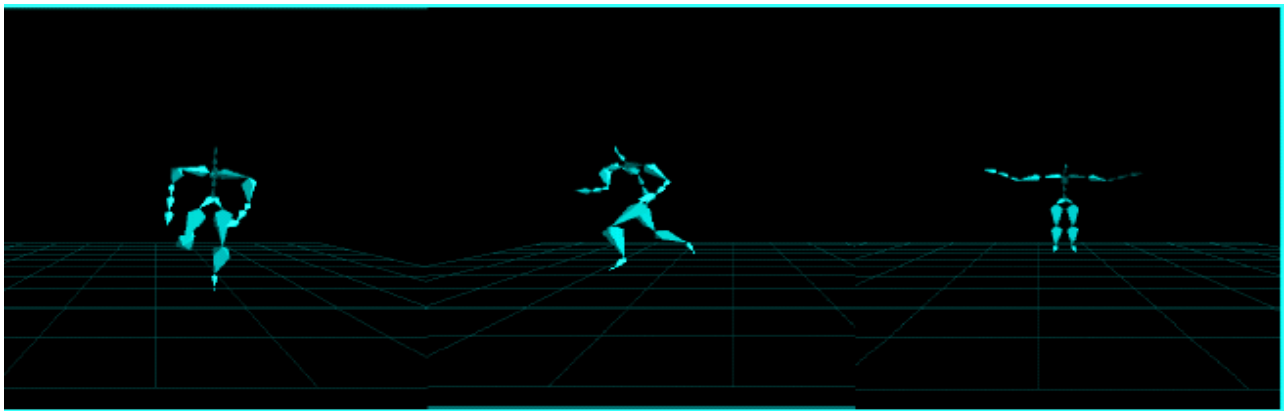
Lifting 2D tracks to 3D





- Divide 2d tracks to body parts → left arm, right arm, left leg, right leg
- Match each body part 2d track with motion capture dataset frames
  - Use snippets of video

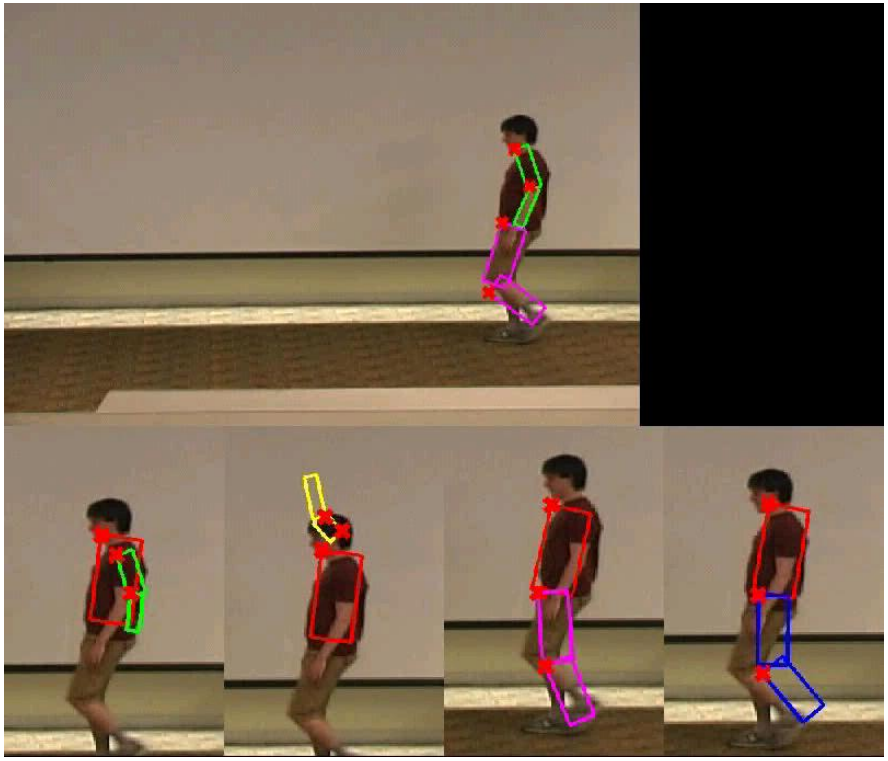
3D Motion Capture Dataset synthesized by Electronic Arts composed of American Football Movements

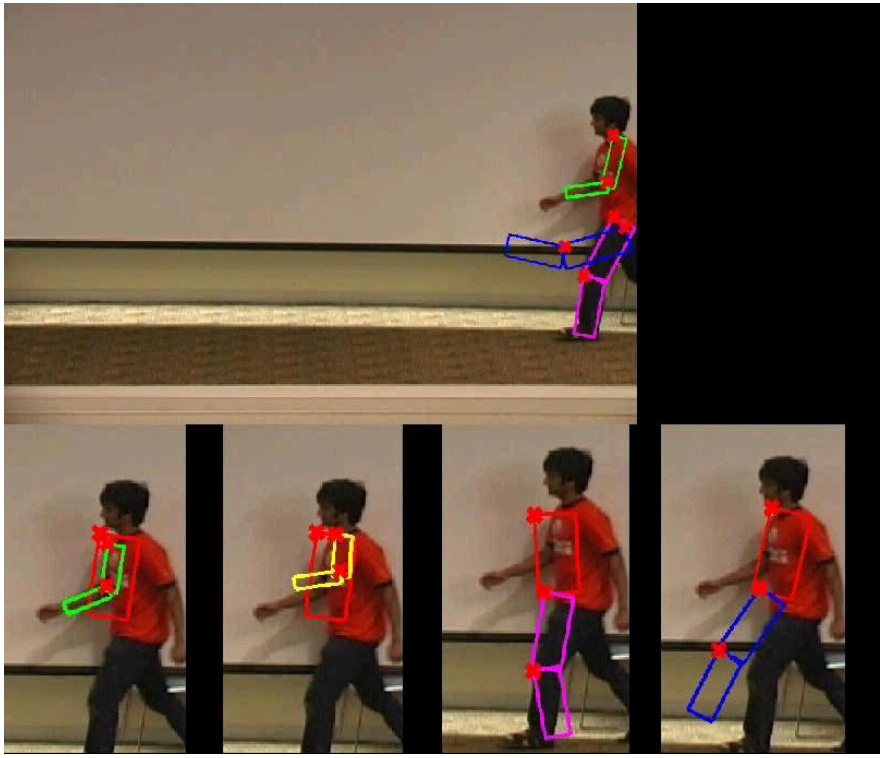


Lifting Procedure

- Assuming an orthographic camera in the root coordinate system of the motion capture dataset,
  - First match legs with dynamic programming using snippet method,
    - allow left and right legs to come from different mocap snippets.
  - Based on best leg camera
    - match the arms
    - force arm camera to be close to leg camera
  - Matching is done via dynamic programming
    - over 20 cameras and top 10 best matches

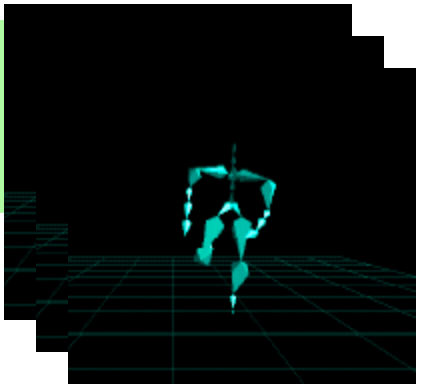
Example Videos



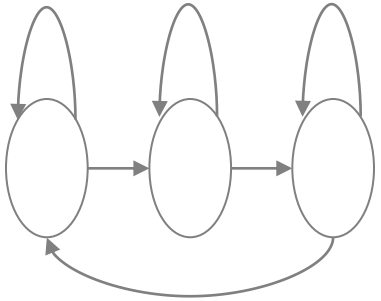


Representing Activities

Motion capture snippets



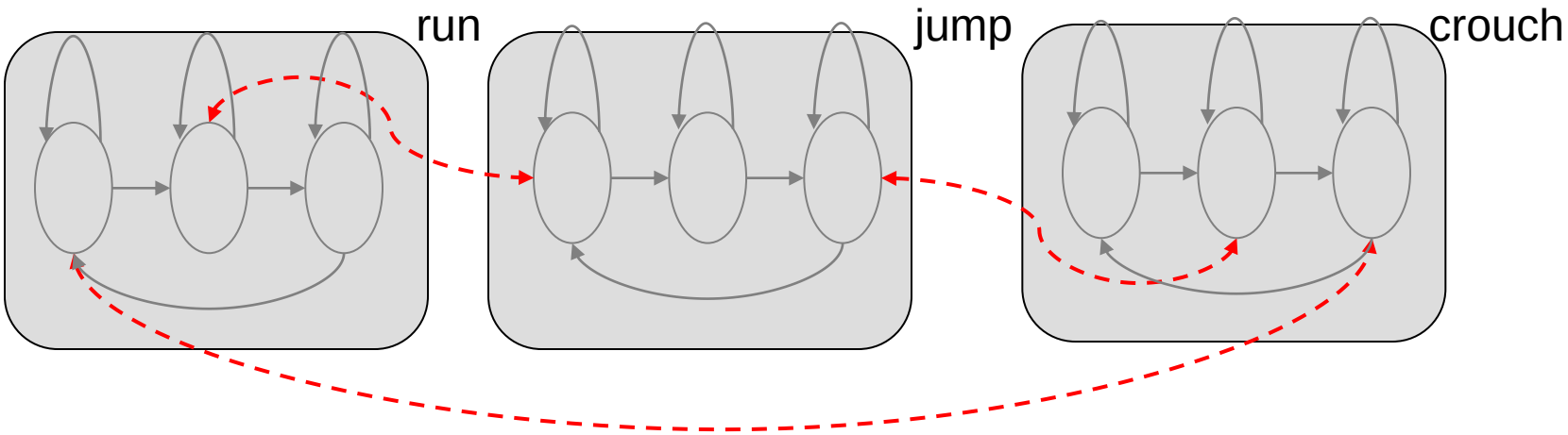
vector quantize



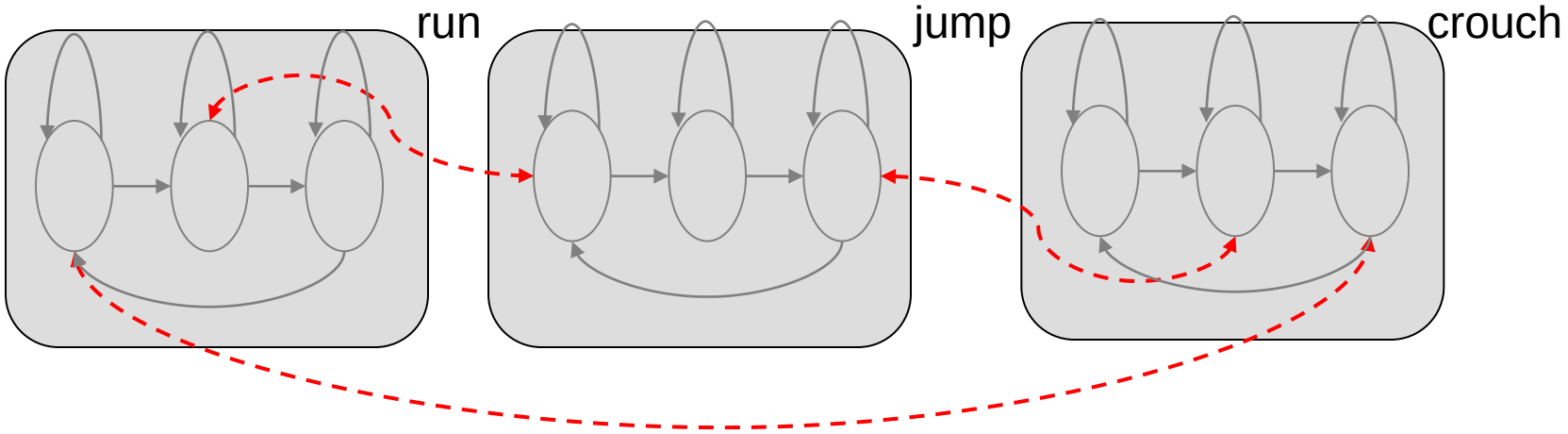
Limb activity model

Form 3 state Hidden Markov Models

Transfer Learning



Limb Activity Models

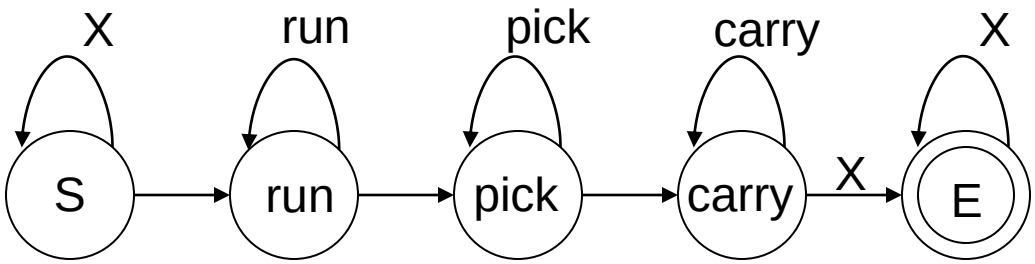


- Join HMMs that have the closest emission states
  - State similarity is computed by the Euclidean distance in 3d space

$$dist(A_m, B_n) = \sum_{o_m=1}^{N} \sum_{o_n=1}^{N} p(o_m)p(o_n)C(o_m, o_n)$$

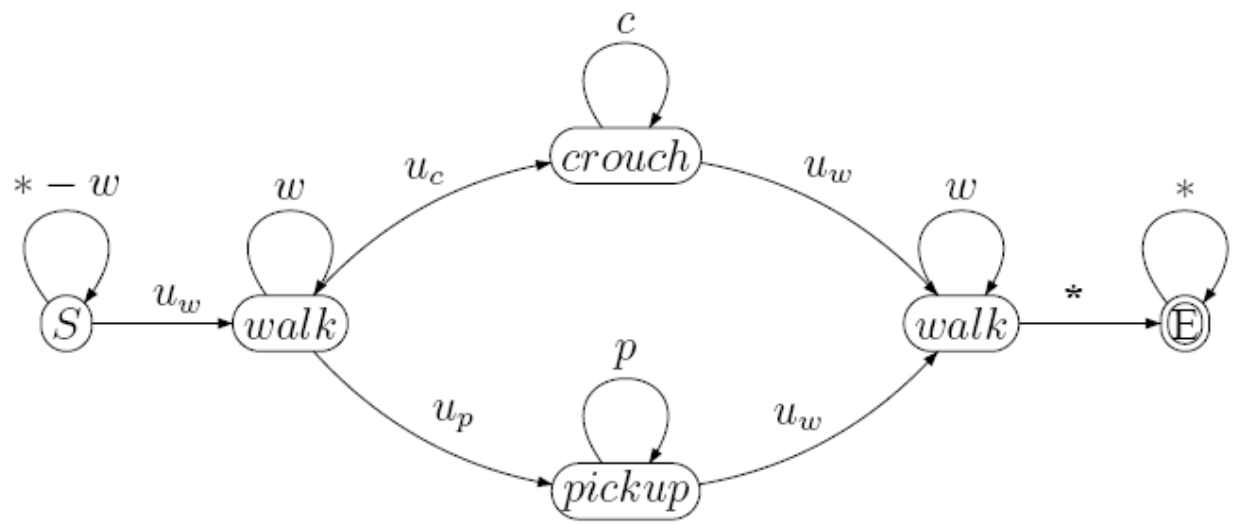
- For each limb, form a big activity model composed of smaller actions
- By this way, we achieve automatic motion segmentation

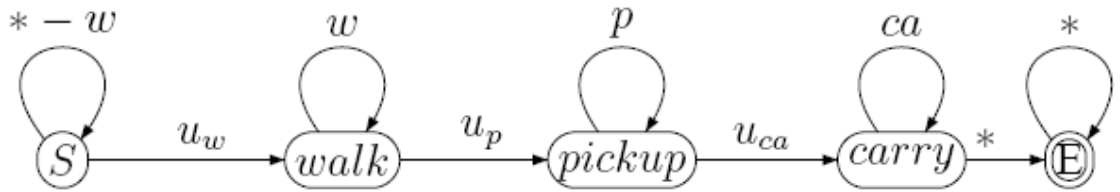
Searching Activities with Finite State Models

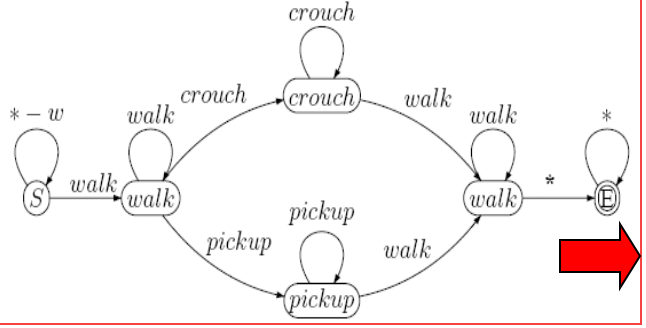


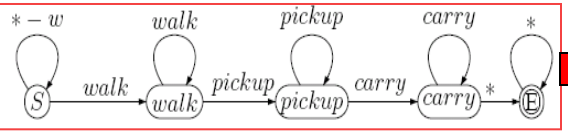
- Author different queries for different body parts
  - Legs walk, arms wave
- No need to specify location
- No need for example motion segment

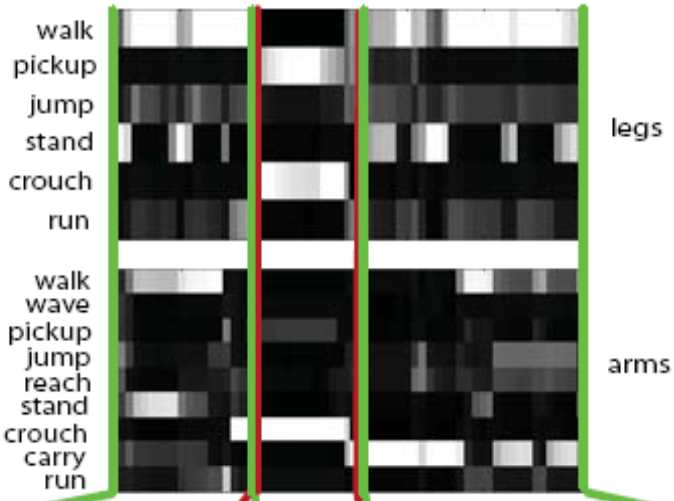
Example FSA for writing complex queries

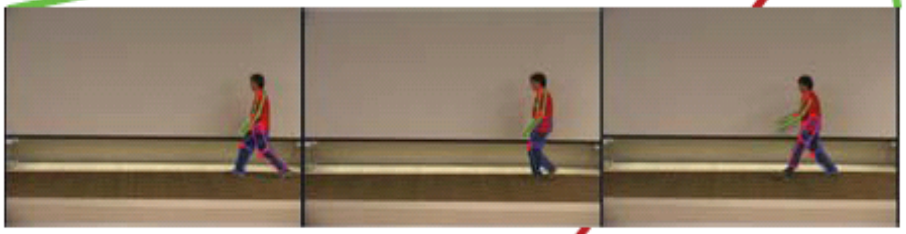




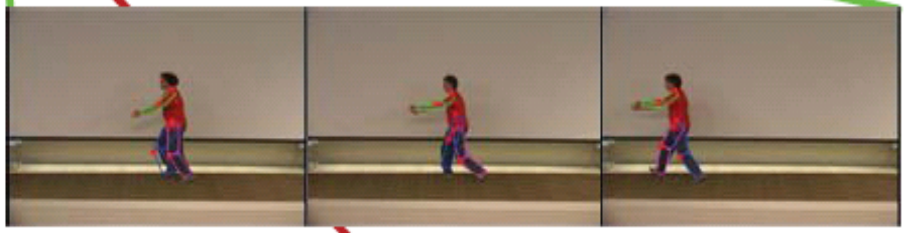




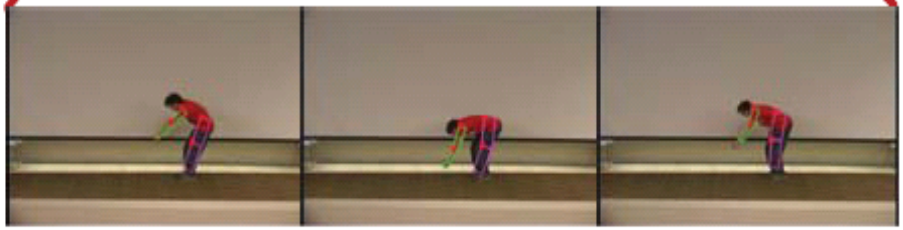




legs: walk
arms: walk



legs: walk
arms: carry

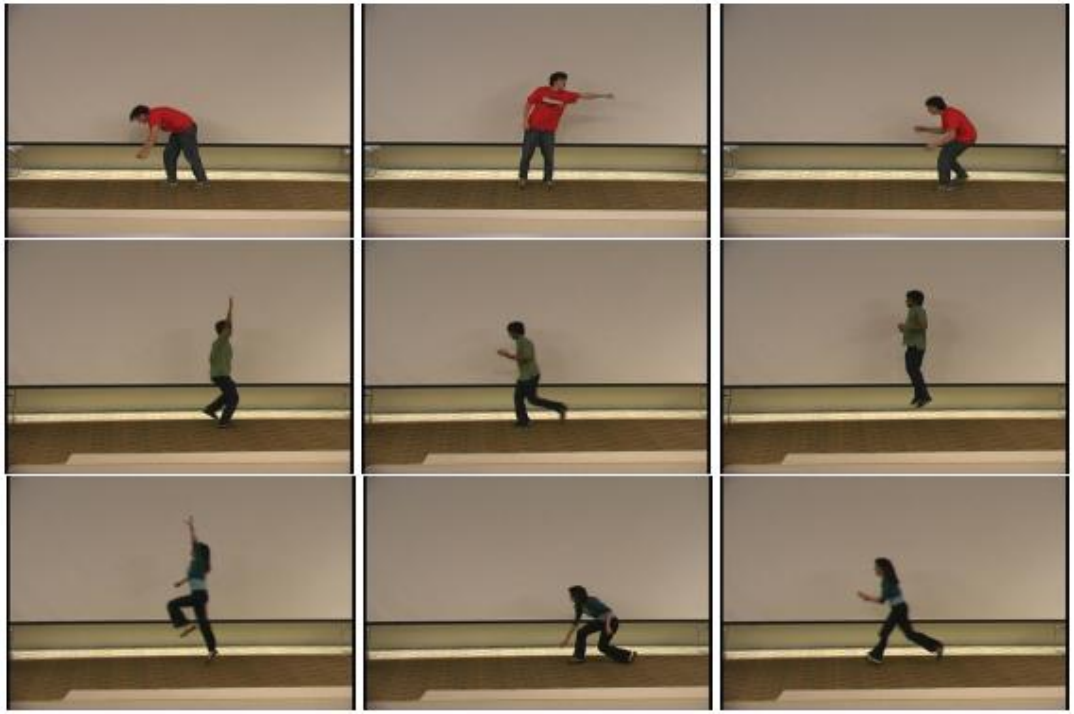


legs: pickup | crouch
arms: pickup

Complex-composite queries possible:
"Find me videos where legs are doing X followed by Y, arms are doing U followed by Z"

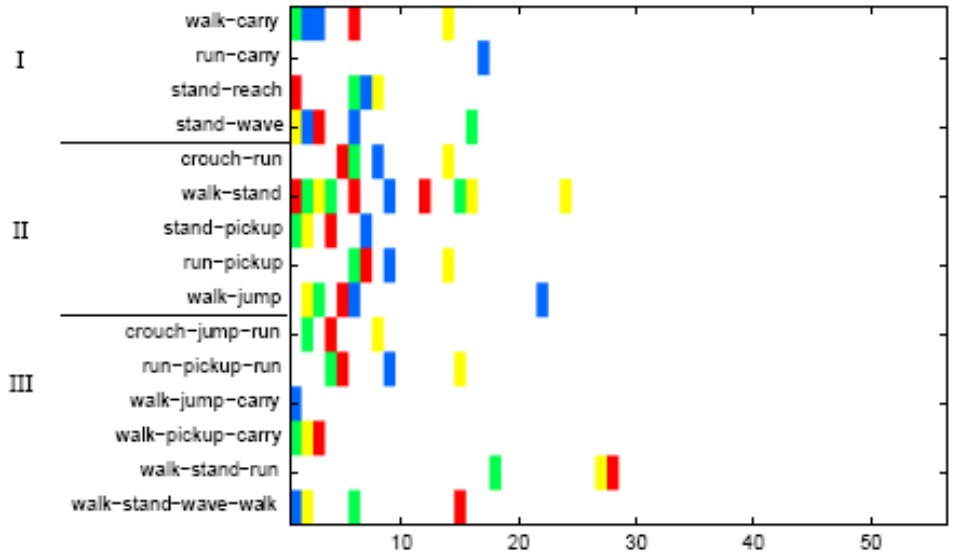
Experiments for complex activities

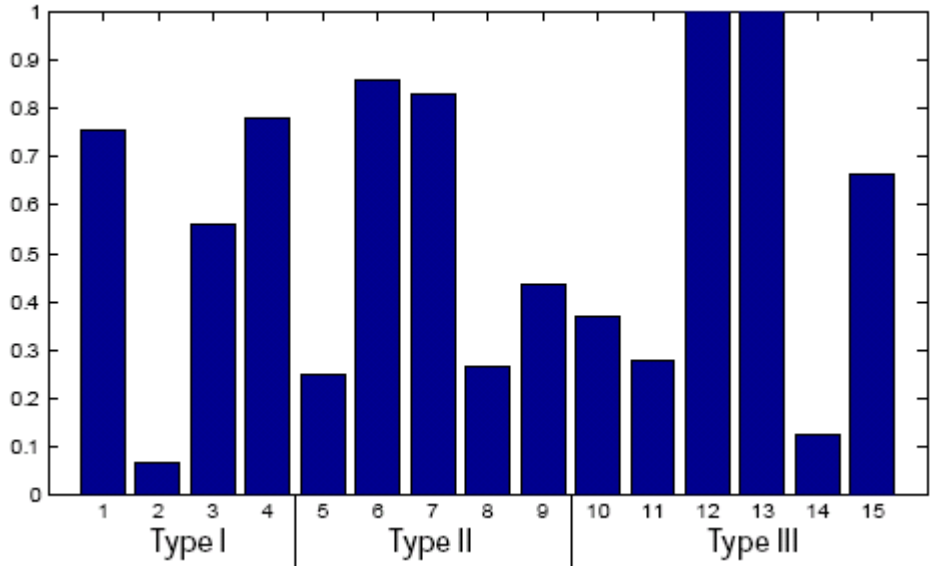
UIUC complex activity dataset



- 3 actors, 5 outfits, 13 different combination of activities
- 73 movies in total

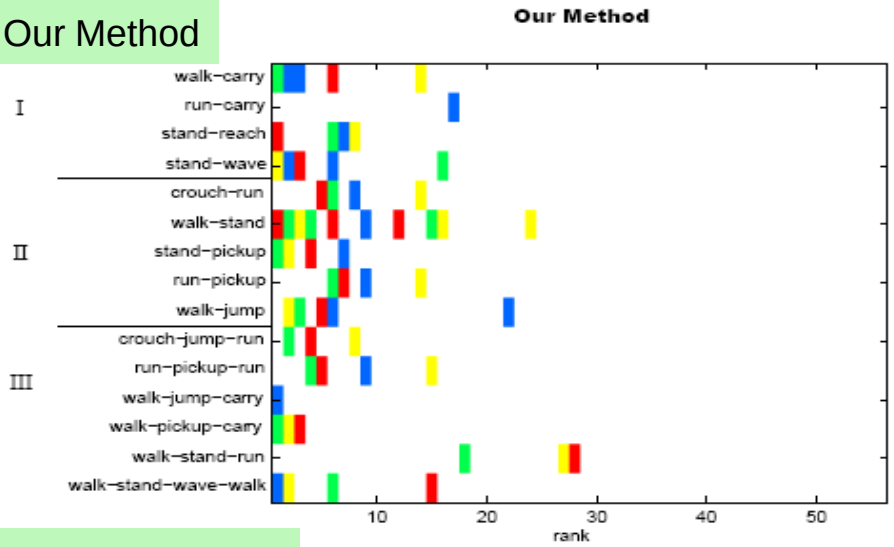
Limb HMM Results (k=40)



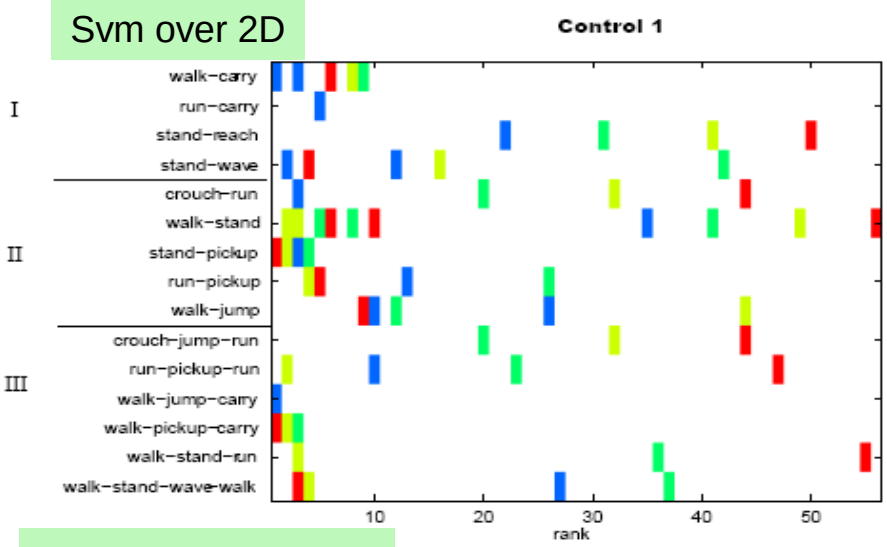


- Queries are insensitive to change in clothing
- For some queries MAP is quite high.
- For 2nd query, a short sequence is missed, that's why the MAP is low.
- Mean average precision is 0.5636

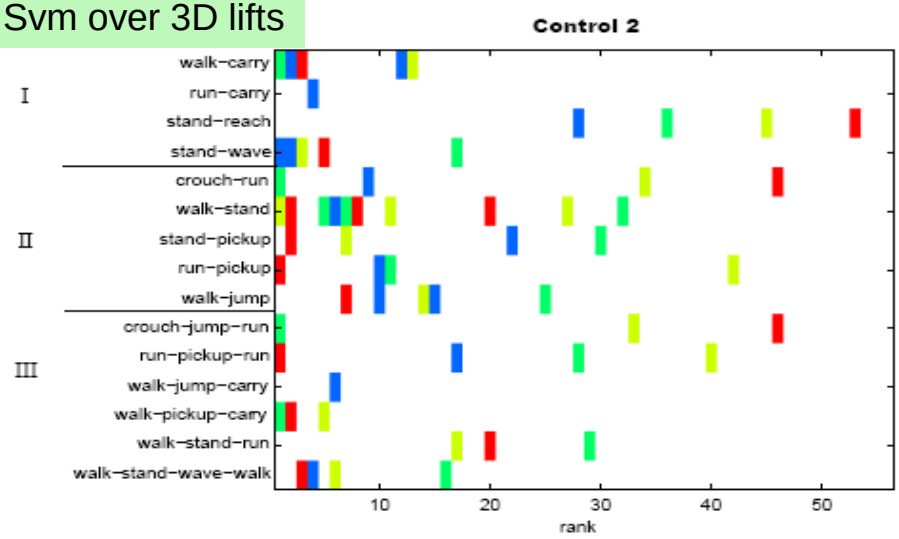
Our Method



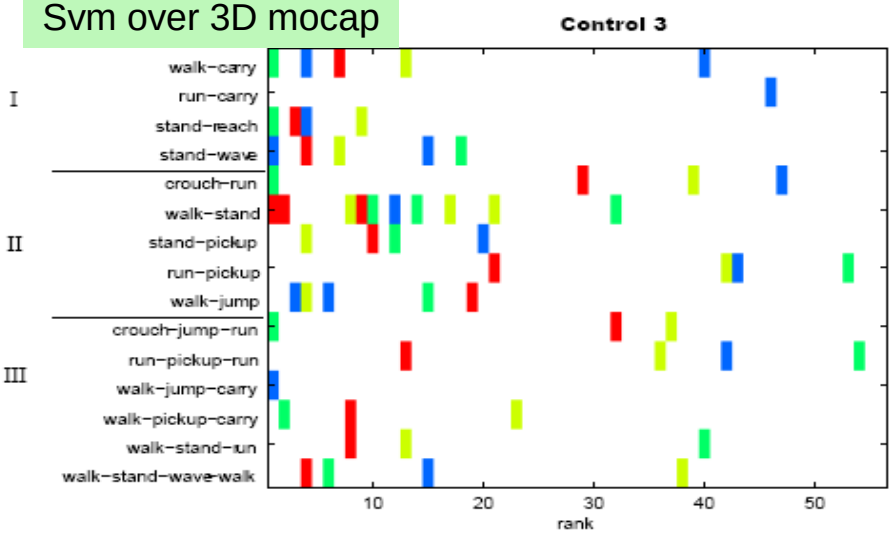
Svm over 2D

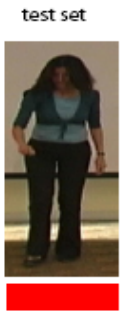


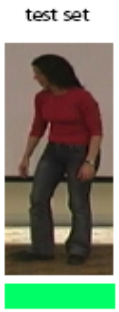
Svm over 3D lifts



Svm over 3D mocap

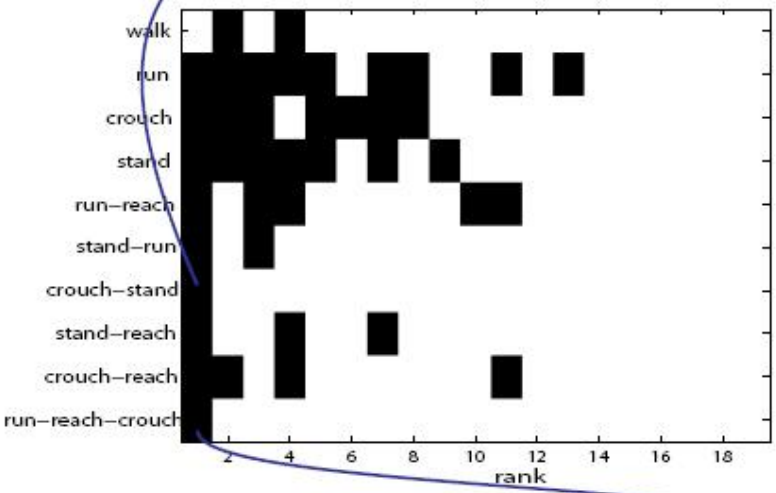


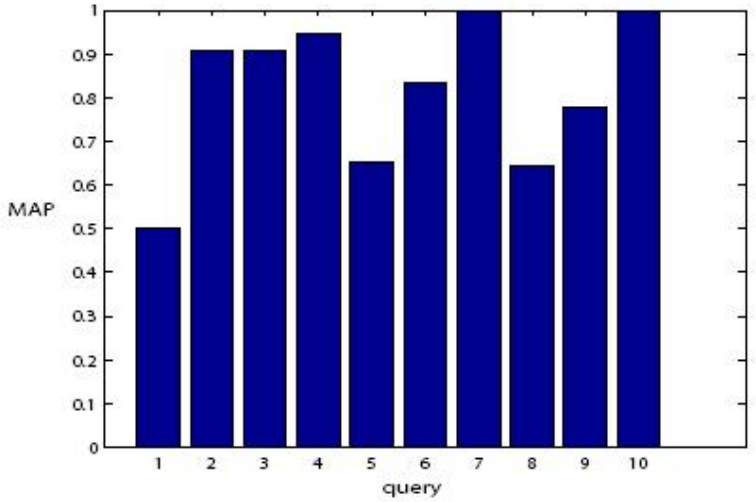




More complicated case…

the first video retrieved for the query "crouch-stand"





the first video retrieved for the query "run-reach-couch"

Indifference to viewpoints

Figure 4. *Example frames from our dataset of single activities with different views.* **Top row:** *Jogging 0 degrees, Jump 45 degrees, jumpjack 90 degrees, reach 135 degrees.* **Bottom row:** *wave 180 degrees, jog 225 degrees, jump 270 degrees, jumpjack 315 degrees.*

- 5 actions, 8 viewpoints = 40 videos
- jog, jump, jumpjack, reach, wave

Conclusions – Single Actions

- Pose, indeed, tells a lot.
- Shape and motion cues are complimentary to each other, depending on the nature of the actions at hand
- Dense templates, both for shape and for optical flow is not needed, a coarse directional and spatial binning gives enough information
- Using compact representations reduces the classification time substantially and also requires less training examples.

Conclusions – Complex Activities

- Model composition is needed for defining composite activities.
- Transfer learning helps a lot in activity recognition, if we didn't use motion capture data, we had to have a great deal of training videos
  - We can easily broaden our set of activities this way
- By joining models of atomic actions, we do not need train examples for activity sequences and we perform minimum parameter estimation
  - We can query for activities that we've never seen before
- Using 3D models is a crucial part of the effective activity recognition, since it is very difficult to train the model in every viewing direction.
- The generative nature of the HMMs helps to compensate different levels of sustainability and makes composition across time easier.
  - By following the transitions between states of the limb activity models, we achieve automatic segmentation of the motion.

Future Work

- Rectangle-based pose description can be augmented to handle the view-invariance case
- Improved 2d-3d lifting
- Use more discriminative features as front-end
- Enrich the set of motions by extending motion capture dataset
- A canonical action vocabulary and an activity ontology is needed
- Modeling interactions between multiple-people

Relevant Publications

- Nazlı İkizler and David A. Forsyth, “Searching for Complex Human Activities with No Visual Examples” *Accepted for publication in International Journal of Computer Vision (IJCV), 2008.*
- Nazlı İkizler and Pınar Duygulu, “Histogram of Oriented Rectangles: A New Pose Descriptor for Human Action Recognition”, *submitted to Journal of Image and Vision Computing (IMAVIS).*
- Nazlı İkizler and Pınar Duygulu, “Human Action Recognition Using Distribution of Oriented Rectangular Patches”, *2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation held in conjunction with Eleventh IEEE International Conference on Computer Vision (ICCV 2007), October 2007.*
- Nazlı Ikizler and David A. Forsyth, “Searching Video for Complex Activities with Finite State Models” *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), June 2007.*
- Nazlı İkizler, R. Gökberk Cinbiş and Pınar Duygulu, “Action Recognition with Line and Flow Histograms”, *submitted*.
- Nazlı İkizler, R. Gökberk Cinbiş, Selen Pehlivan and Pınar Duygulu, “Recognizing Actions From Still Images”, *submitted*.