Recognizing Actions by Shape-Motion Prototype Trees

ICCV 2009

Zhe Lin, Zhuolin Jiang, Larry S. Davis

Computer Vision Laboratory
University of Maryland at College Park

March 10th, 2009

UMIACS
University of Maryland Institute for Advanced Computer Studies

Action Recognition

- Action Recognition
  - Static action/ gesture
    - Signal from a single image
    - Visual cues: shape
  - Dynamic action/ gesture
    - Signal from a sequence
    - Visual cues: shape and motion
- Applications
  - Video surveillance
  - Multimedia analysis
  - Human-robot interaction (HRI)

Dynamic actions

Goal & Challenges

- Problem
  - Develop an efficient and robust system to recognize human gestures and actions from a moving platform and under cluttered, dynamic background.

- Challenges
  - Cluttered, dynamic background
  - Moving platforms (e.g. robots)
  - Moving gestures/ actions
  - Appearance variation
  - Occlusions

Motivation

- Explicit human pose estimation is very difficult and time consuming due to high dimensionality, occlusions and color ambiguity.

Instead, our approach is based on implicit pose estimation by learning and matching action prototypes.

Contributions

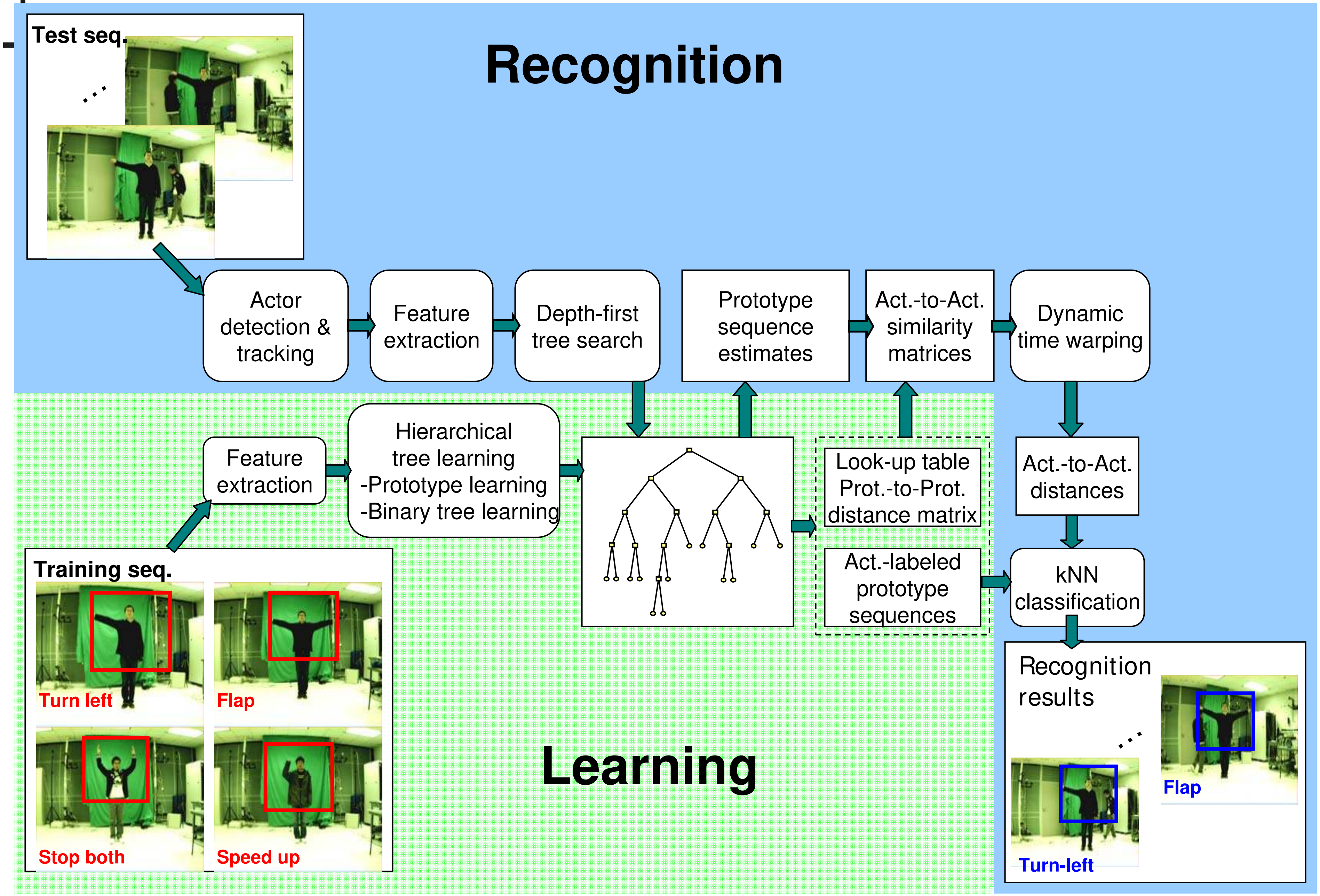
- A prototype-based approach is introduced for robustly detecting and matching prototypes, and recognizing actions against dynamic backgrounds.

- Actions are modeled by learning a prototype tree in a joint shape-motion space via hierarchical k-means clustering.

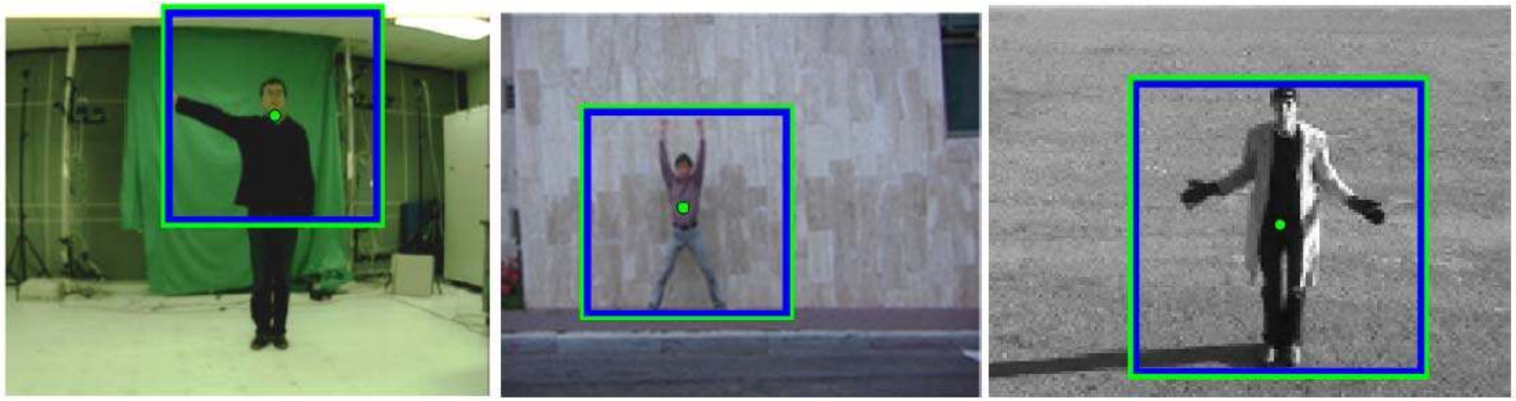
- Frame-to-frame distances are rapidly estimated via fast prototype tree search and look-up table indexing.

- A new challenging dataset consisting of 14 gestures is introduced for public use.

Overview

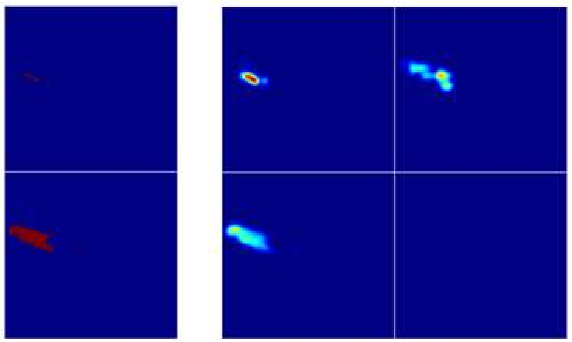

Test seq.
Recognition
Actor detection & tracking
Feature extraction
Depth-first tree search
Prototype sequence estimates
Act.-to-Act. similarity matrices
Dynamic time warping
Feature extraction
Hierarchical tree learning
-Prototype learning
-Binary tree learning
Look-up table
Prot.-to-Prot. distance matrix
Act.-labeled prototype sequences
Act.-to-Act. distances
kNN classification
Training seq.
Turn left
Flap
Stop both
Speed up
Learning
Recognition results
Flap
Turn-left

Action Representation by Joint Shape-Motion Descriptors



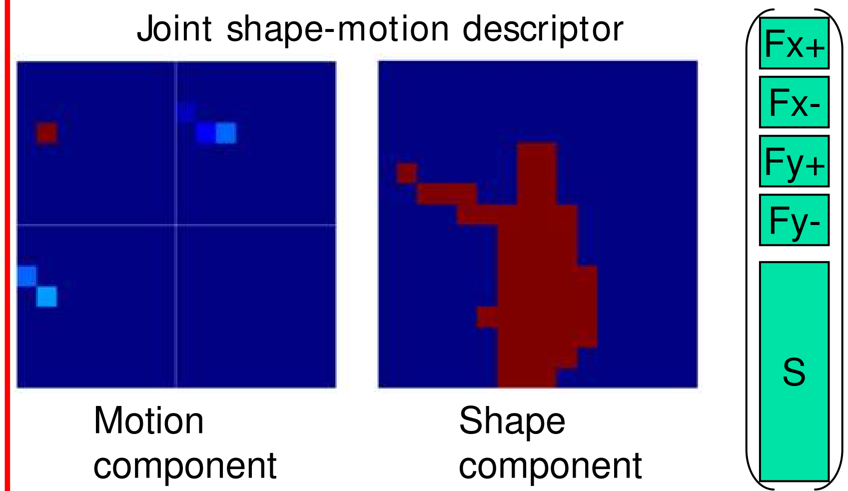
Action Interest Regions

A training image with optical flow



Motion observations [Efros et al. ICCV'03]

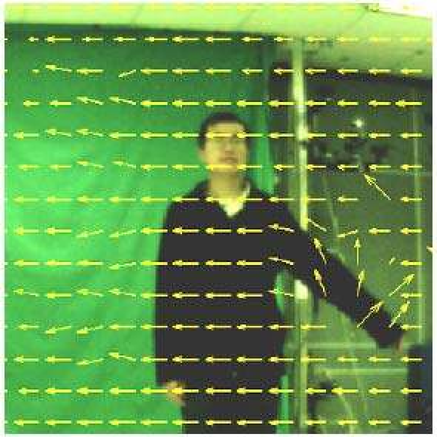
Joint shape-motion descriptor



Motion component

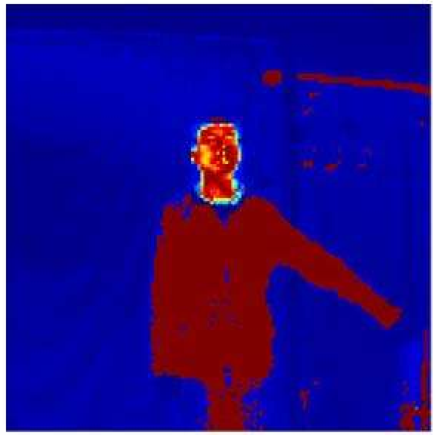
Shape component

Motion Compensation

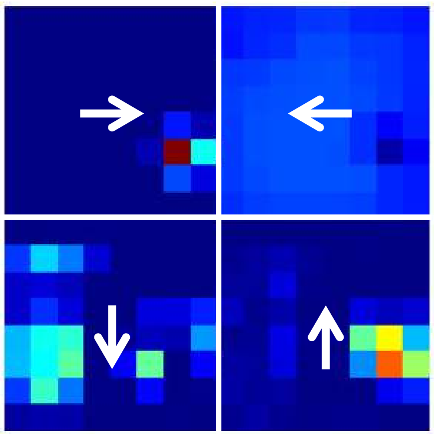


Raw optical flow field

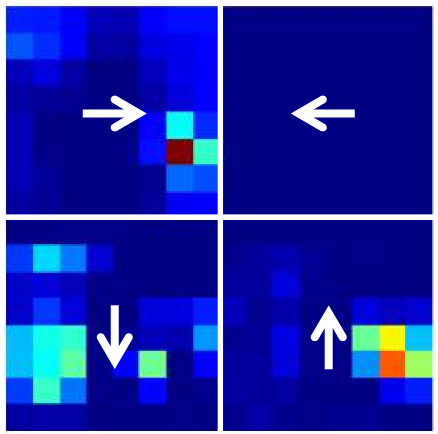
Compensated flow field



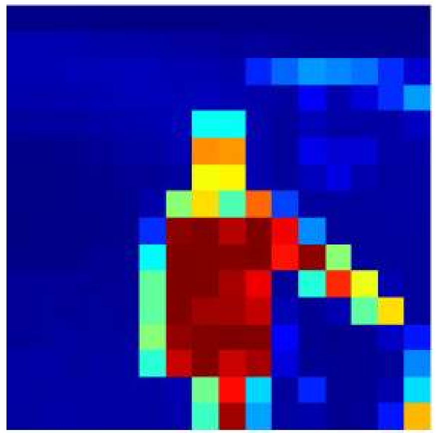
Combined appearance
–based likelihood map



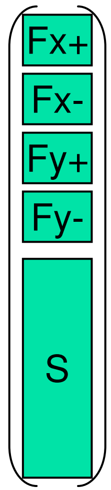
Raw motion descriptor



Compensated motion
descriptor

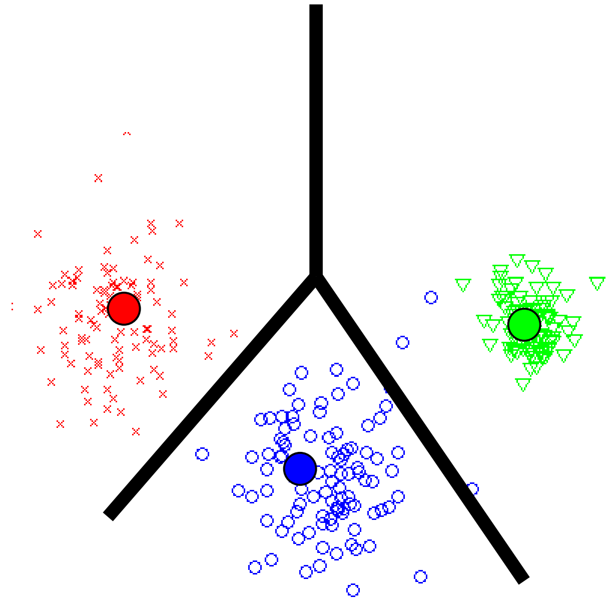


Shape descriptor



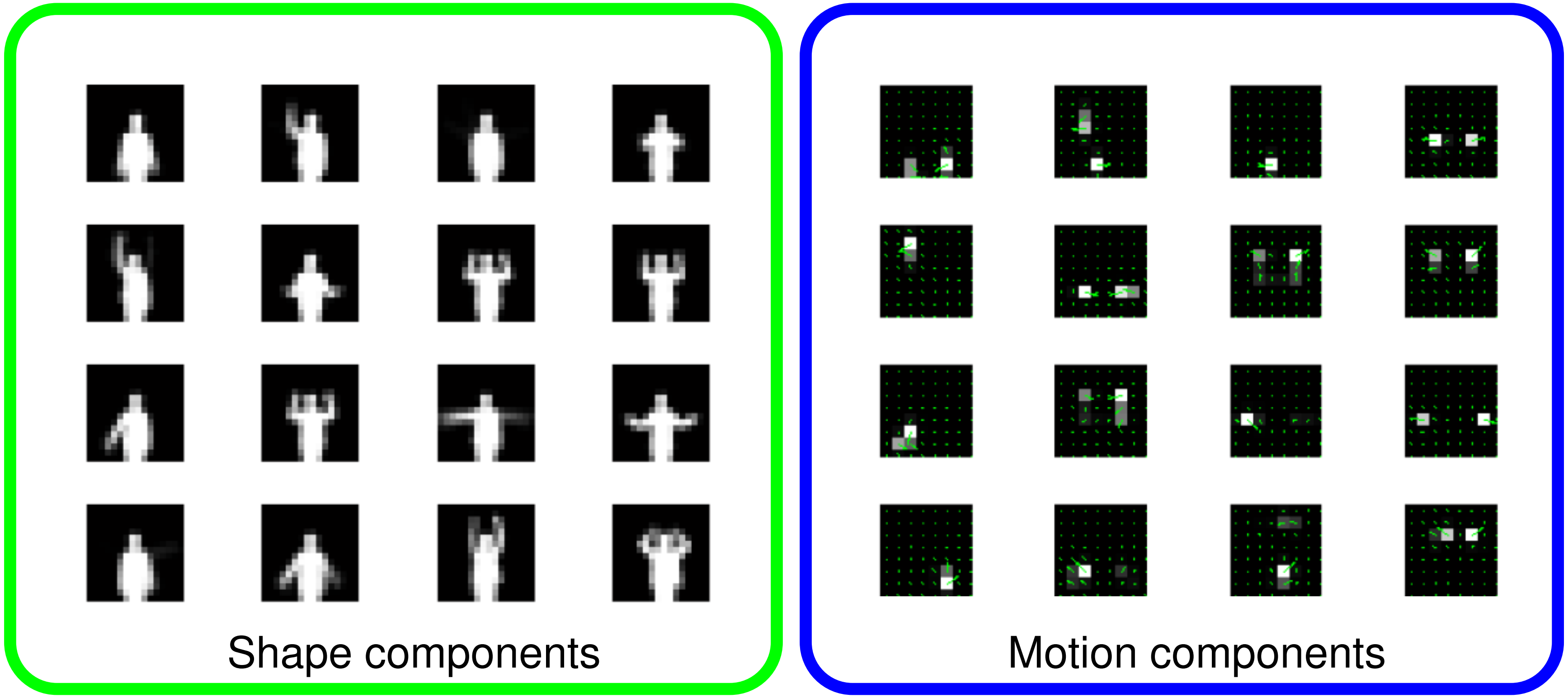
Learning Action Prototypes

- For handling large training database of action videos, we represent actions as a set of basic action units called prototypes.
- Hierarchical *k*-means clustering in a joint shape and motion space using the Euclidean distances
  - Use *k*-means centers as the shape-motion prototypes.
  - Construct a prototype-to-prototype distance matrix that is used as a look-up table to speed-up the recognition process.
  - Build a prototype tree to rapidly search for matching prototypes



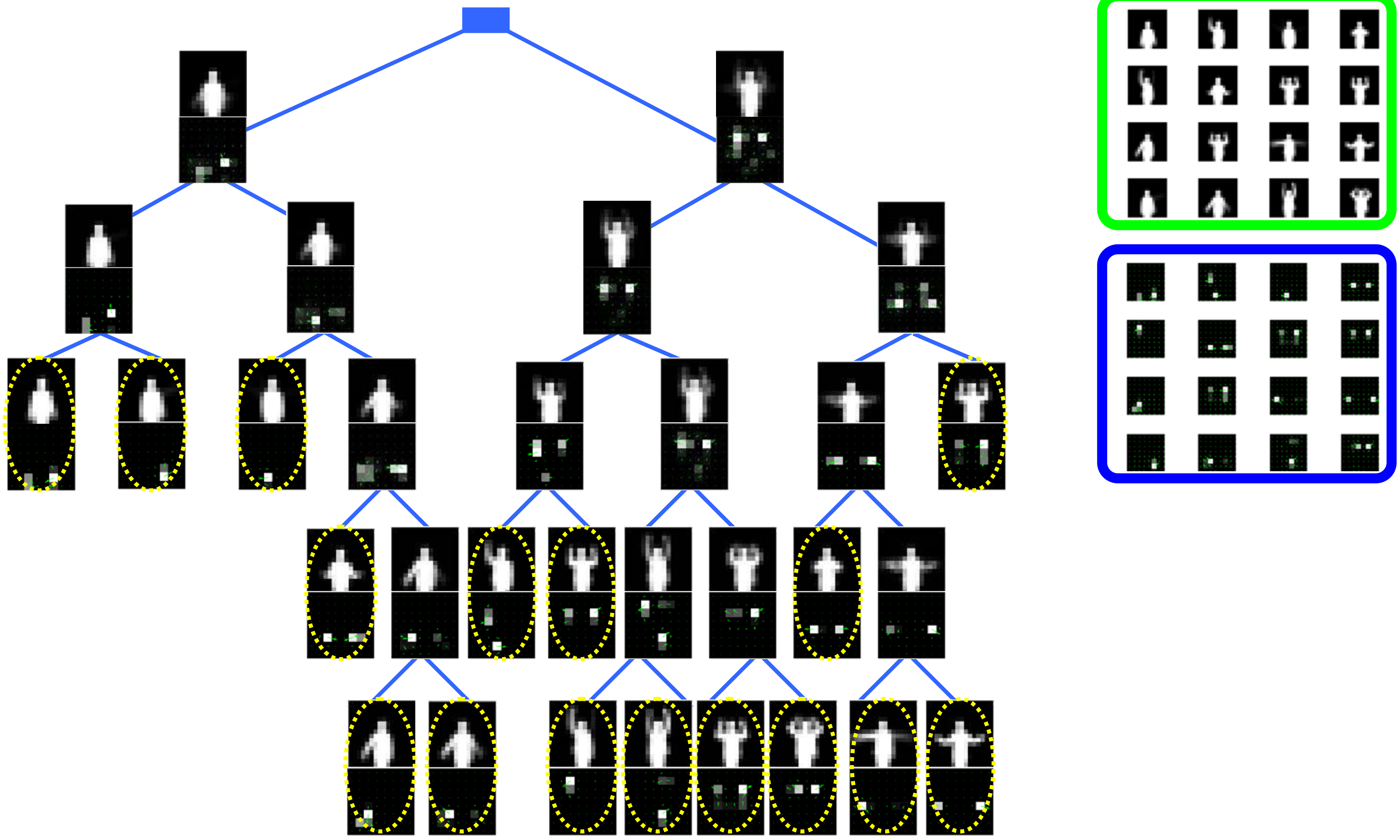
Action Prototypes

- Shape-Motion Prototypes
  - Implicit pose models

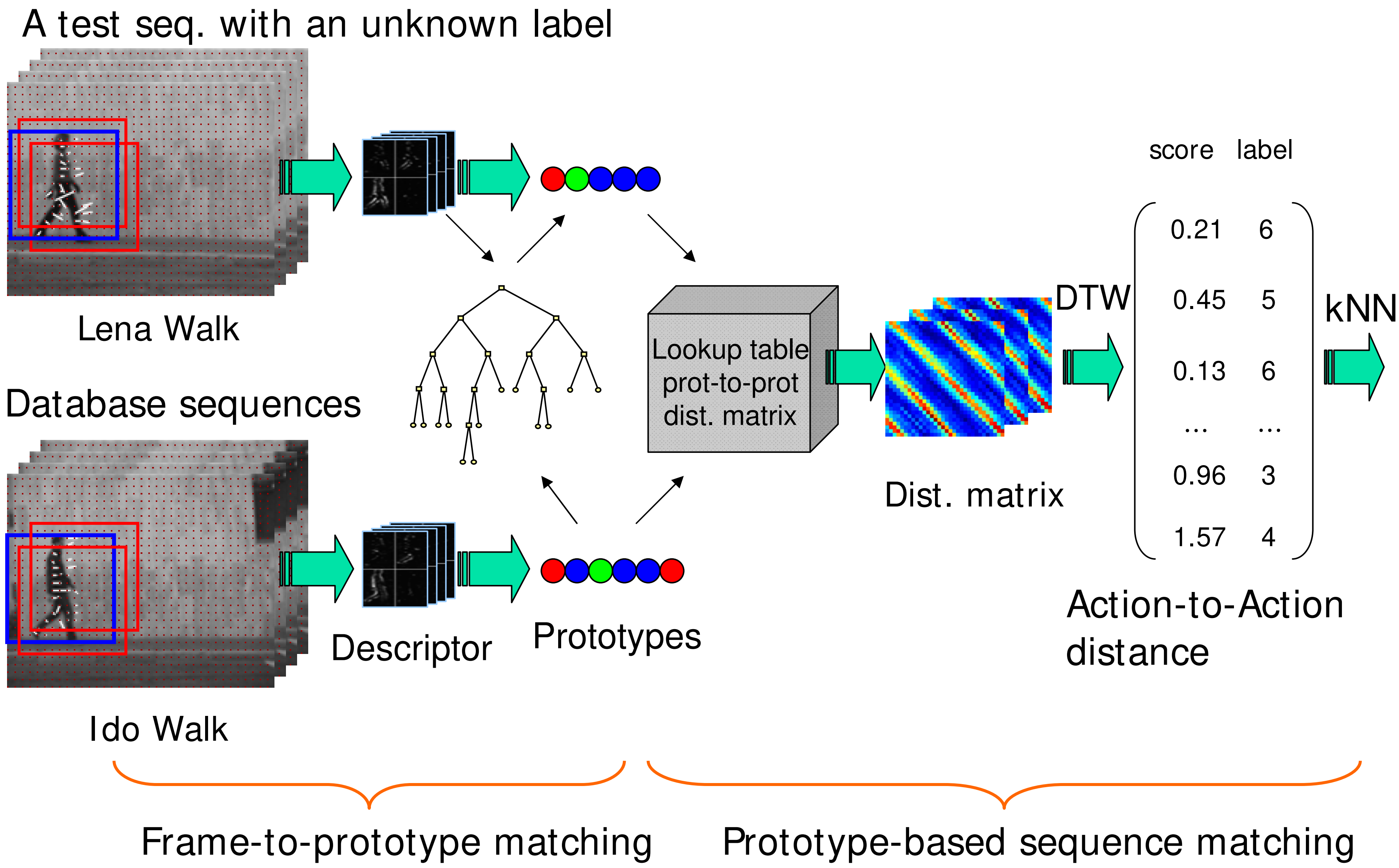


Number of clusters: K = 16

Binary Prototype Tree



Action Recognition Process



Frame-to-prototype matching

- Joint likelihood model

$$p(V,\theta,\alpha) \propto p(\theta,\alpha|V)$$
$$= \underbrace{p(\theta|V,\alpha)}_{\text{Prototype matching term}}\underbrace{p(\alpha|V)}_{\text{Action localization term}}$$

V - observation r.v.

θ - prototype r.v.

α - Localization r.v.

$$\boxed{p(\theta|V,\alpha) = exp(-d(D(V,\alpha),D(\theta)))}$$

$$\boxed{p(\alpha|V) = \frac{L(\alpha|V) - L_{min}}{L_{max} - L_{min}}}$$

- Optimization problem

$$(\theta^*,\alpha^*) = \arg\max_{\theta,\alpha} p(V,\theta,\alpha)$$

Frame-to-prototype matching

- Joint likelihood optimization
  - Only search using the space (set) of learned action prototypes instead of the entire high-dim. pose space, making the method computationally efficient.

For a set of samples $\alpha_1, \alpha_2 \ldots \alpha_P$

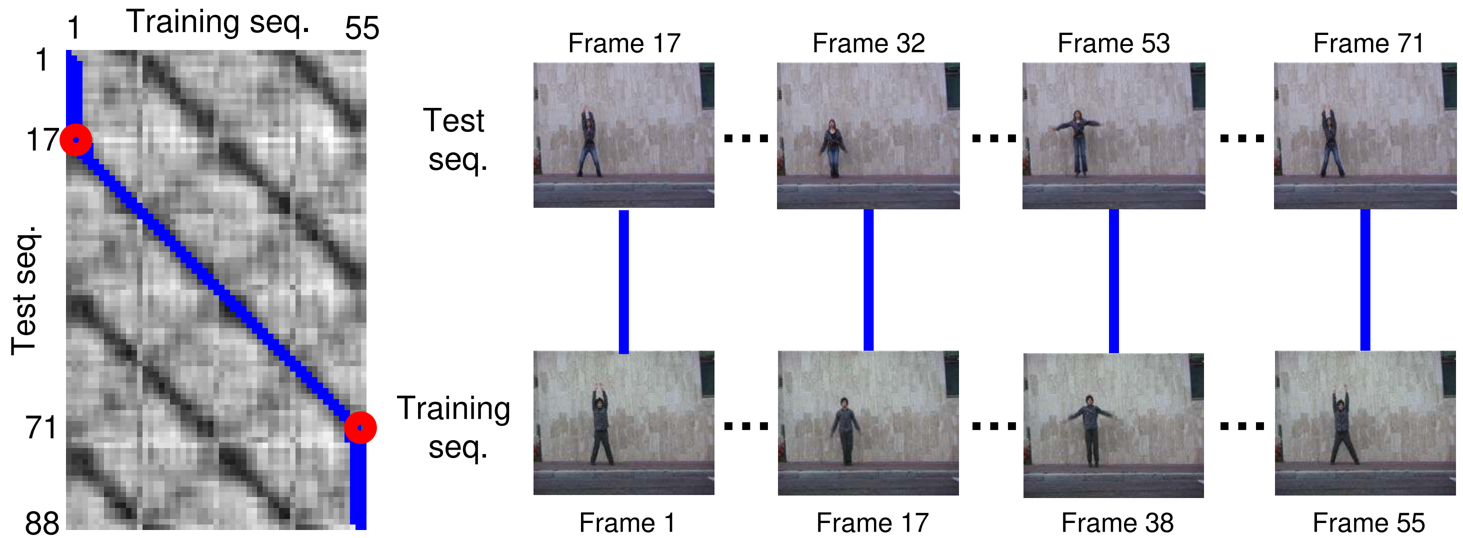
$$\theta^*(\alpha_p) = \arg\max_{\theta \in \Theta} p(V, \theta, \alpha_p)$$

$$J(\alpha_p) = \exp(-d(D(V_t, \alpha_p), D(\theta^*(\alpha_p))))L(\alpha|V_t)$$

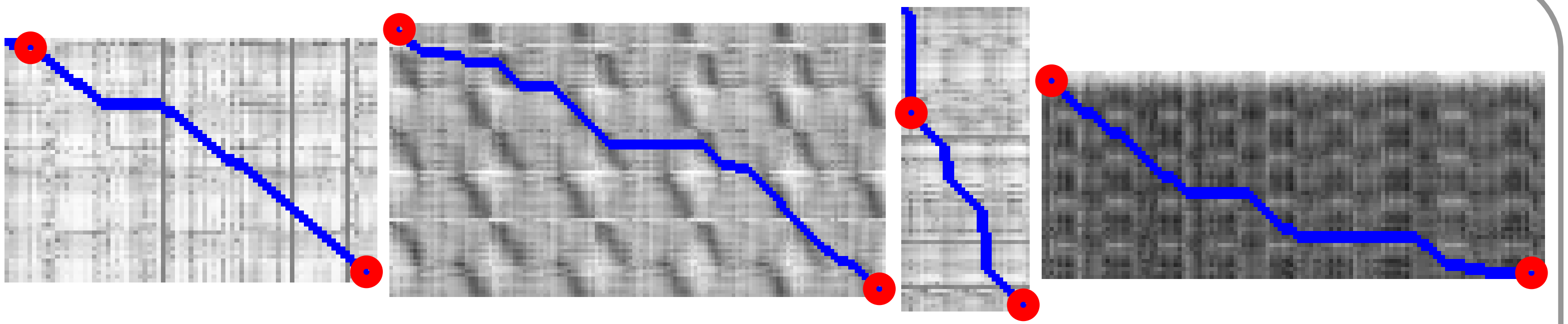
$$\alpha_p^* = \arg\max_{\alpha_p, p=1,2\ldots P} J(\alpha_p).$$

Prototype-based Sequence Matching

Dynamic sequence alignment



Same actions performed by different people



Different actions performed by different people

Alignment-based Recognition

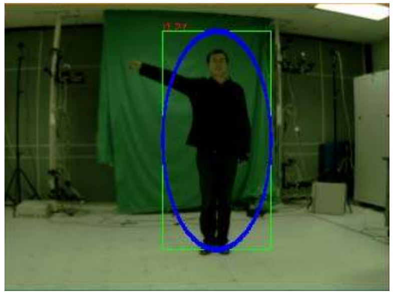
- Action-to-Action distance

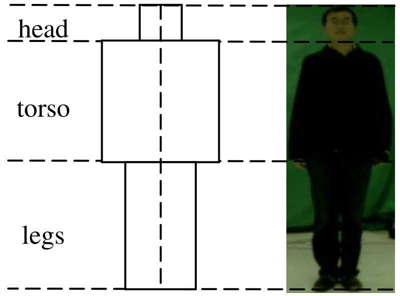
$$Dist(G_x, G_y) = \frac{\sum_{l=l_{start}}^{l_{end}} dist(x_{l,i}^*, y_{l,j}^*)}{(l_{end} - l_{start} + 1)}$$

- Recognition
  - K-nearest neighbor (k-NN) classification
  - Non-modeled actions are rejected by thresholding the action-to-action distances

Action Localization

- Initialize by background subtraction or a generic human detector, and track the person by a local mode seeking-based tracker, such as meanshift [Comaniciu03]





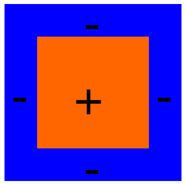
- Build a part-based appearance model using nonparametric kernel density estimation

$$p(y) = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{d} K_{\sigma_j}(y_j - x_{ij})$$

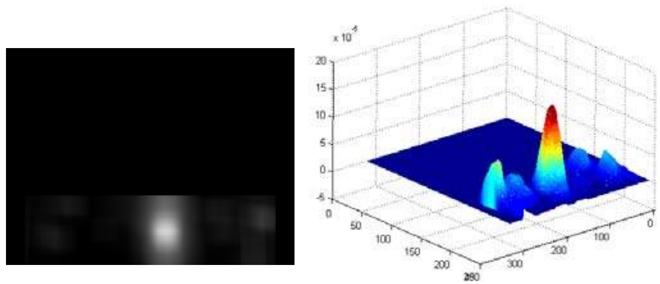
- Form appearance-based likelihood maps by linearly combing part likelihood (probability) maps

head likelihood Torso likelihood Leg likelihood

- Location likelihood: the difference of average appearance-based likelihood between inside and outside the rectangular regions - Like a generalized Laplacian operator

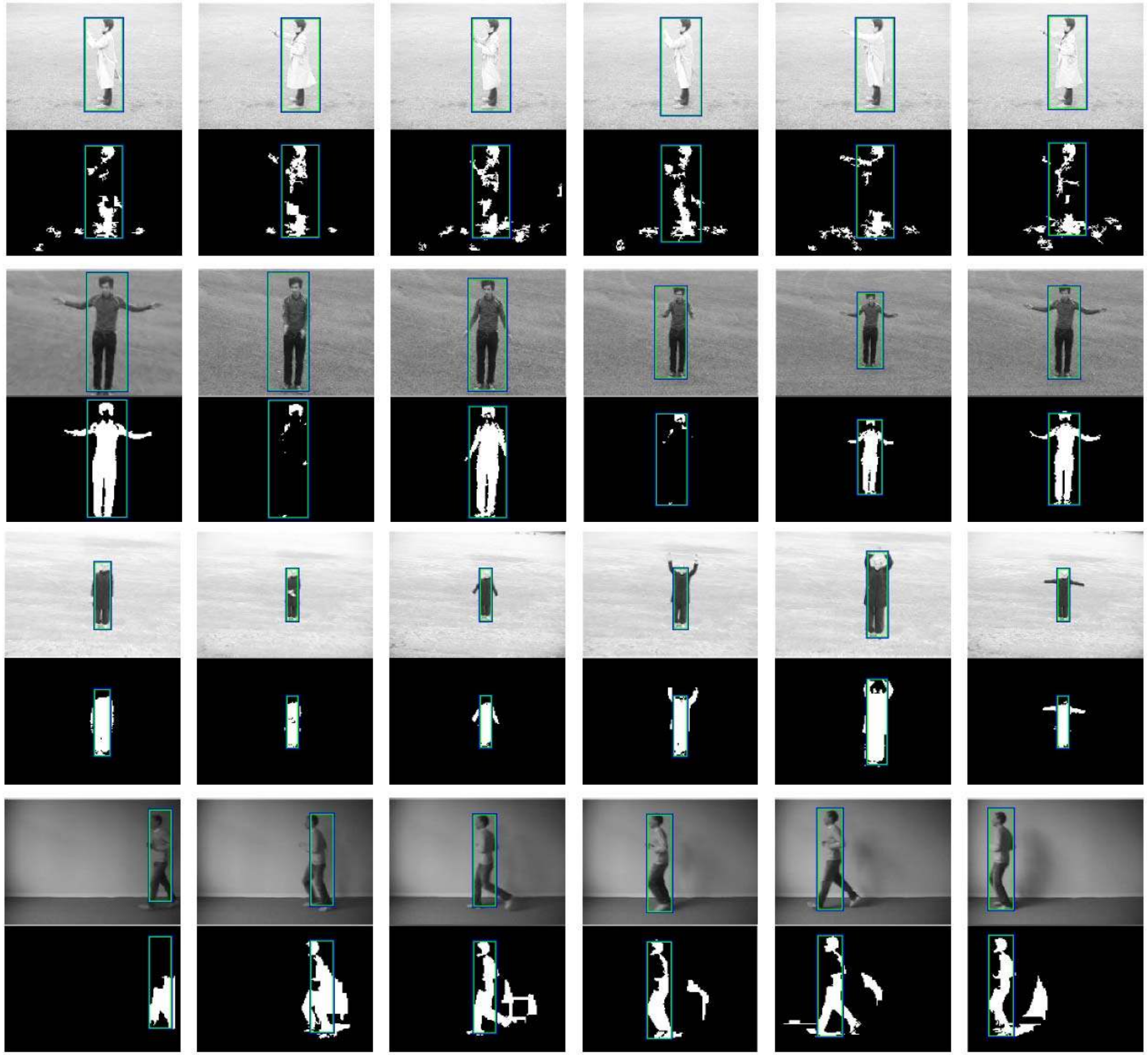


Filter



Location likelihood map

Example: Action Localization



Experiments

- Application
  - Gesture Recognition
  - Action Recognition

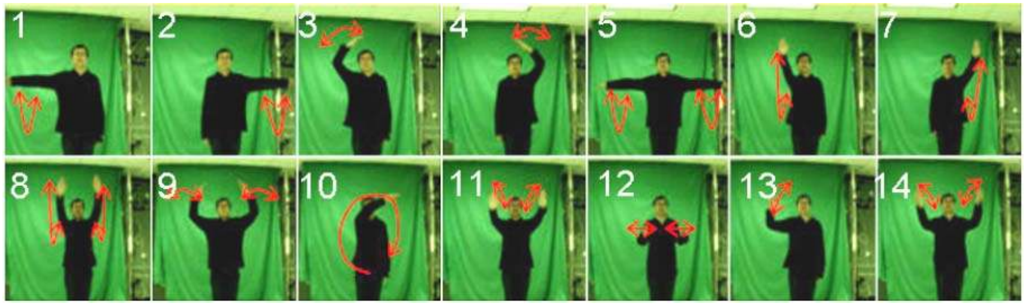
- Joint Shape-Motion Descriptor
  - 16*16 shape descriptor
  - Four channels of 8*8 motion descriptor
  - Total dimension: 512

Datasets


Stereo sensor


“Robot”

- Kecklab Gesture Dataset



The Keck gesture dataset contains 252 videos of 14 gestures performed by 3 individuals.

- Training sequences (fixed camera)
- Testing sequences (moving camera)
- Noisy, occluded testing sequences (moving camera, moving objects, occlusions)

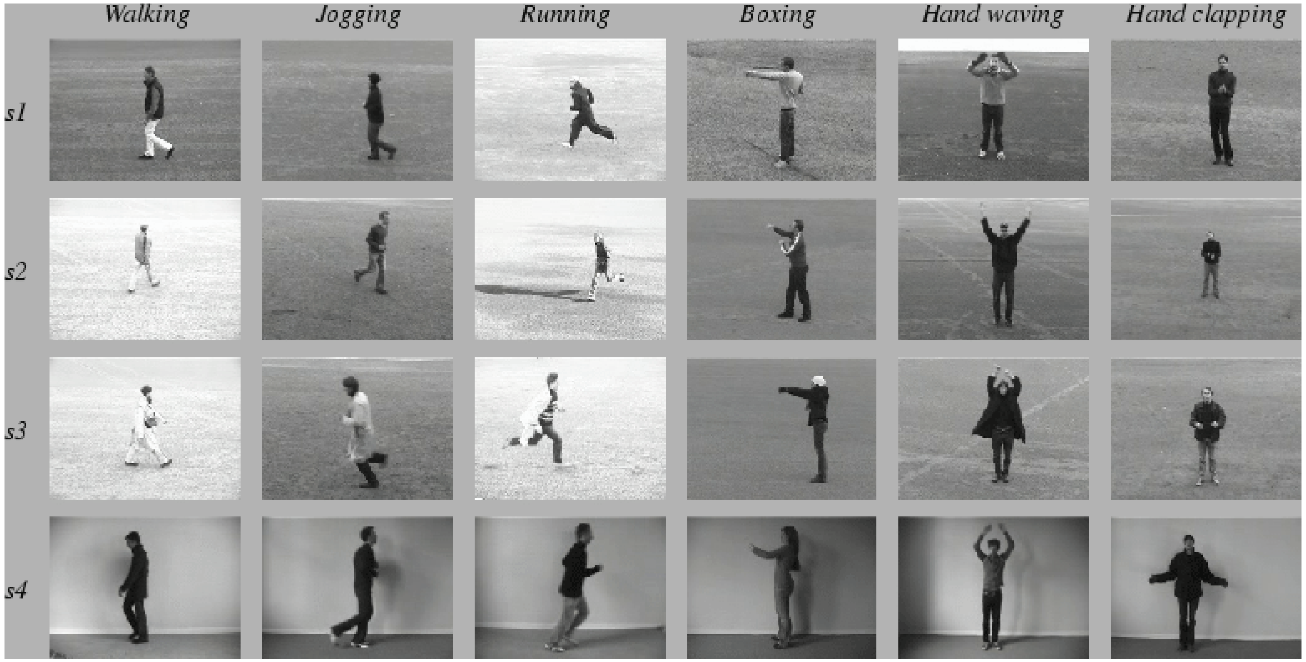
- Weizmann Action Dataset

The Weizmann action dataset contains 90 videos of 10 actions performed by 9 individuals.

L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In IEEE Trans. PAMI, 29(12):2247–2253, 2007.
http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

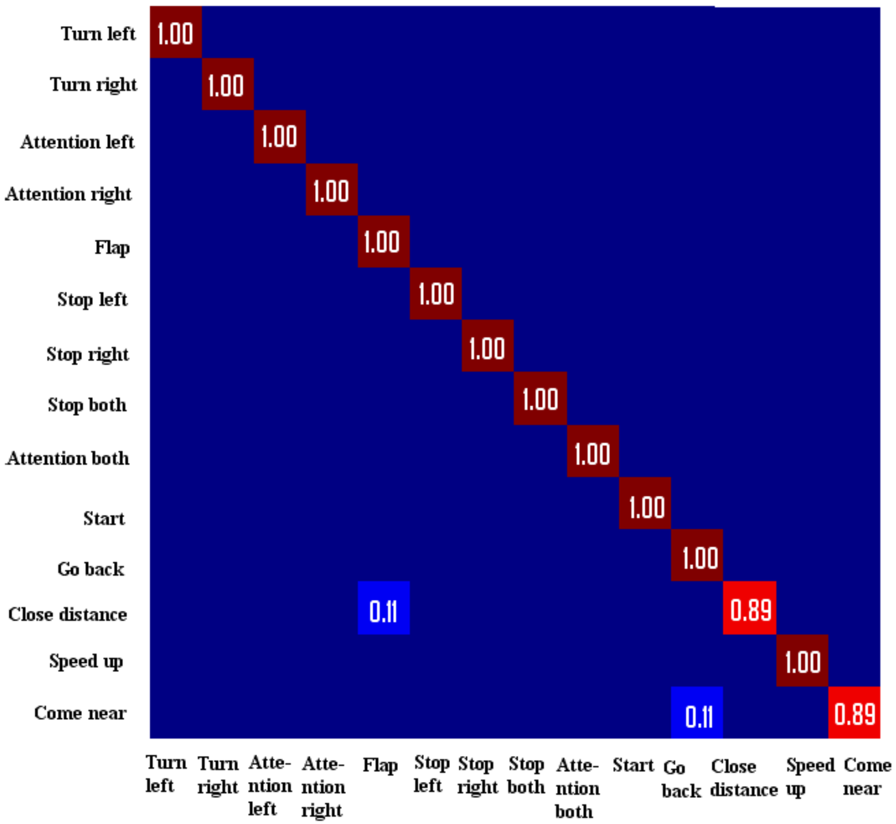
Datasets

- KTH Dataset http://www.nada.kth.se/cvap/actions/
  - 2391 action sequences (50-250 frames)
  - 25 people, 4 scenarios, 6 actions
    - 6 actions (Boxing, walking, running, hand-crapping, hand-waving, jogging)
  - 4 scenarios (**S1**: outdoor, **S2**: outdoor with scale variation, **S3**: outdoor with different cloths, **S4**: indoor)



Results on the Gesture Dataset

- Static camera, static background
  - 'leave-one-person-out' experiment on training data



method	recog. rate (%)
motion only	92.86
shape only	92.86
joint shape and motion	95.24

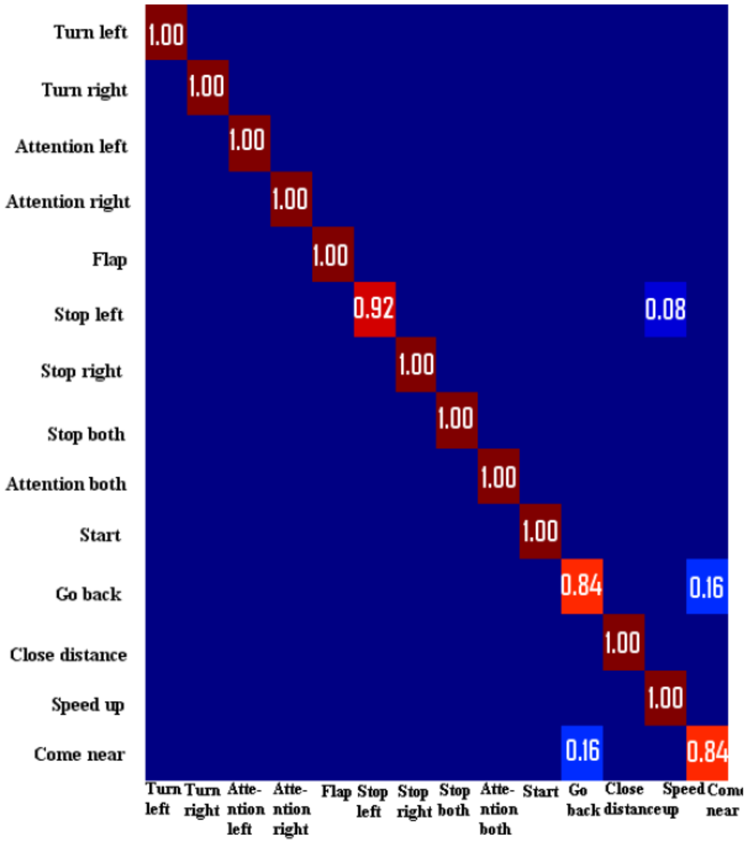
method	recog. rate(%)	avg. time(s)
descriptor dist.	95.24	0.1545
look-up(80 prot.)	90.48	0.0232
look-up(100 prot.)	92.86	0.0256
look-up(120 prot.)	90.48	0.0223
look-up(140 prot.)	92.86	0.0227
look-up(160 prot.)	95.24	0.0232
look-up(180 prot.)	95.24	0.0256

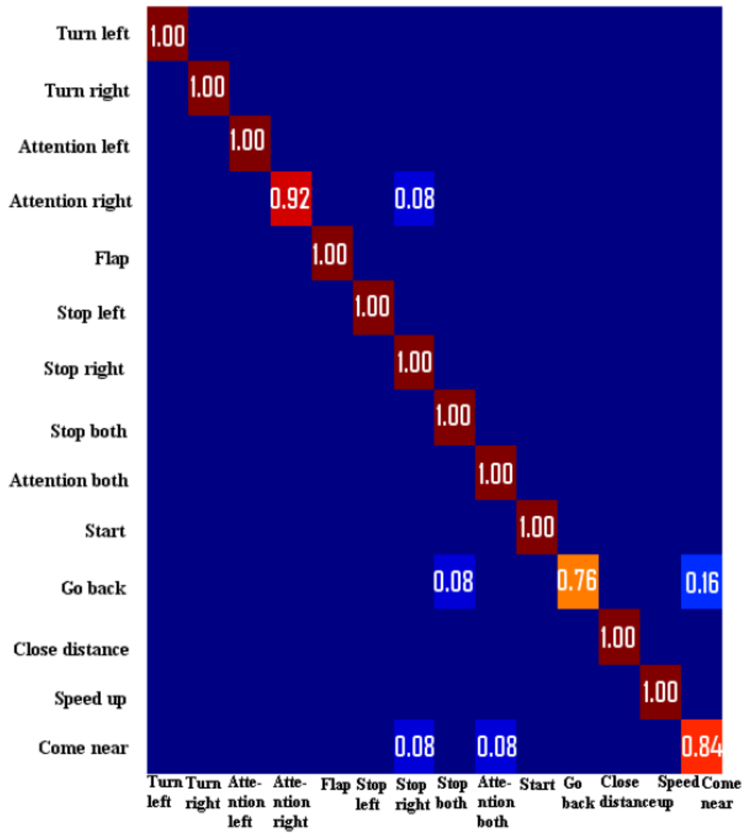
Results on Keck Gesture Dataset

- Moving camera, dynamic background

method	recog. rate (%)
motion only	87.5
shape only	53.57
joint shape and motion	91.07

method	recog. rate (%)	avg. time(s)
descriptor dist.	91.07	0.0965
look-up(160 prot.)	82.14	0.0077
look-up(180 prot.)	89.29	0.0078



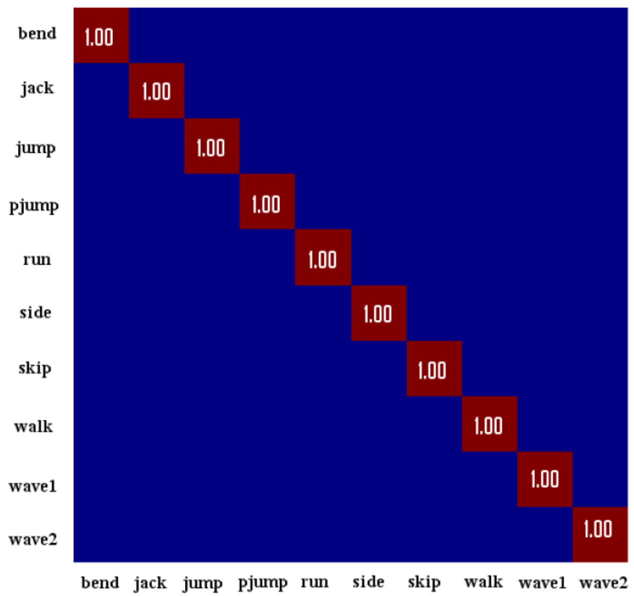


Results on Weizmann Dataset

http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

method	recog. rate (%)
motion only	88.89
shape only	81.11
joint shape and motion	100

method	recog. rate (%)	avg. time(s)
descriptor dist.	100	0.0134
look-up(80 prot.)	96.67	0.0005
look-up(100 prot.)	97.78	0.0005
look-up(120 prot.)	97.78	0.0006
look-up(140 prot.)	100	0.0005
look-up(160 prot.)	98.89	0.0006
look-up(180 prot.)	100	0.0005
Fathi&Mori [7]	100	N/A
Jhuang *et al.* [10]	98.8	N/A
Niebles&Fei-Fei [16]	72.8	N/A



Leave-one-person-out experiments

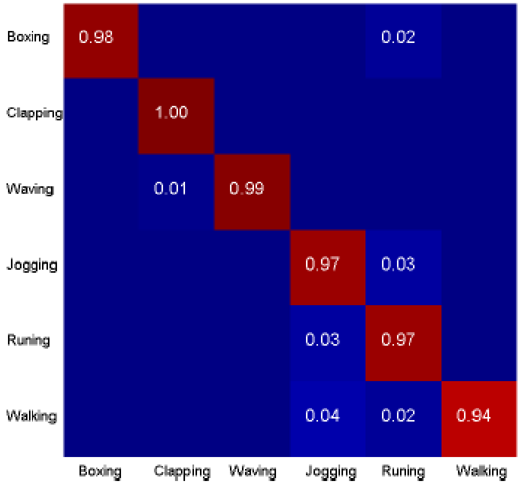
Results on the KTH Dataset

method	recognition rate (%) / time (ms) s1	s2	s3	s4
descriptor dist.	98.83 / 15.2	94 / 19.3	94.78 / 14.5	95.48 / 16.7
look-up(200 pr.)	96.83 / 0.9	85.17 / 1.2	92.26 / 0.8	85.79 / 1.1
look-up(240 pr.)	97.50 / 0.9	83.50 / 1.3	91.08 / 0.8	90.30 / 1.1
look-up(300 pr.)	96.66 / 0.9	86.17 / 1.2	90.07 / 0.8	89.97 / 1.1
Schindler [22]	93.0 / N/A	81.1 / N/A	92.1 / N/A	96.7 / N/A
Jhuang [9]	96.0 / N/A	86.1 / N/A	89.8 / N/A	94.8 / N/A
Ahmad [1]	90.17 / N/A	84.83 / N/A	89.83 / N/A	85.67 / N/A

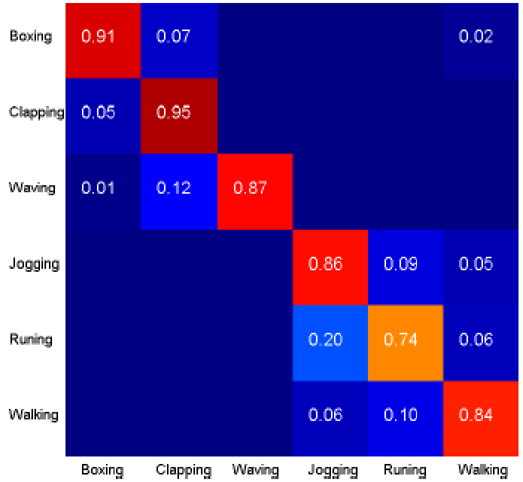
method	recognition rate (%) average of all scenarios	all scenarios in one
Our approach	95.77	93.43
Schindler [22]	90.73	92.7
Ahmad [1]	87.63	88.83
Jhuang [9]	91.68	N/A
Liu [15]	94.15	N/A
Niebles [17]	N/A	81.5
Dollar [5]	N/A	81.17
Schuldt [23]	N/A	71.72
Fathi [8]	N/A	90.50
Nowozin [20]	N/A	87.04
Wang [28]	N/A	92.43

Leave-one-person-out experiments

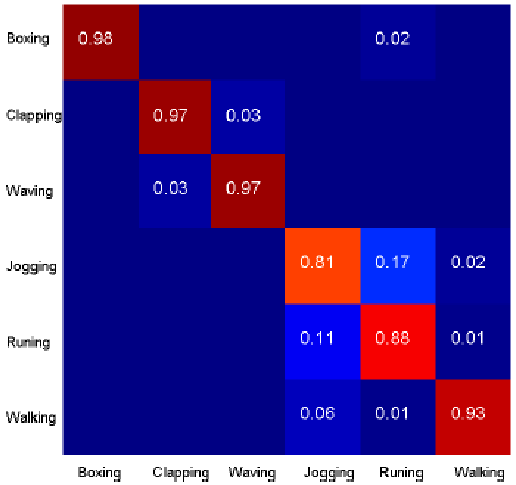
Results on the KTH Dataset



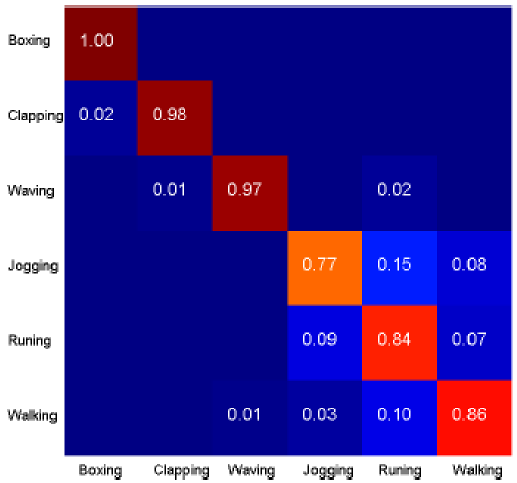
(a) s1 scenario using prototypes



(b) s2 scenario using prototypes



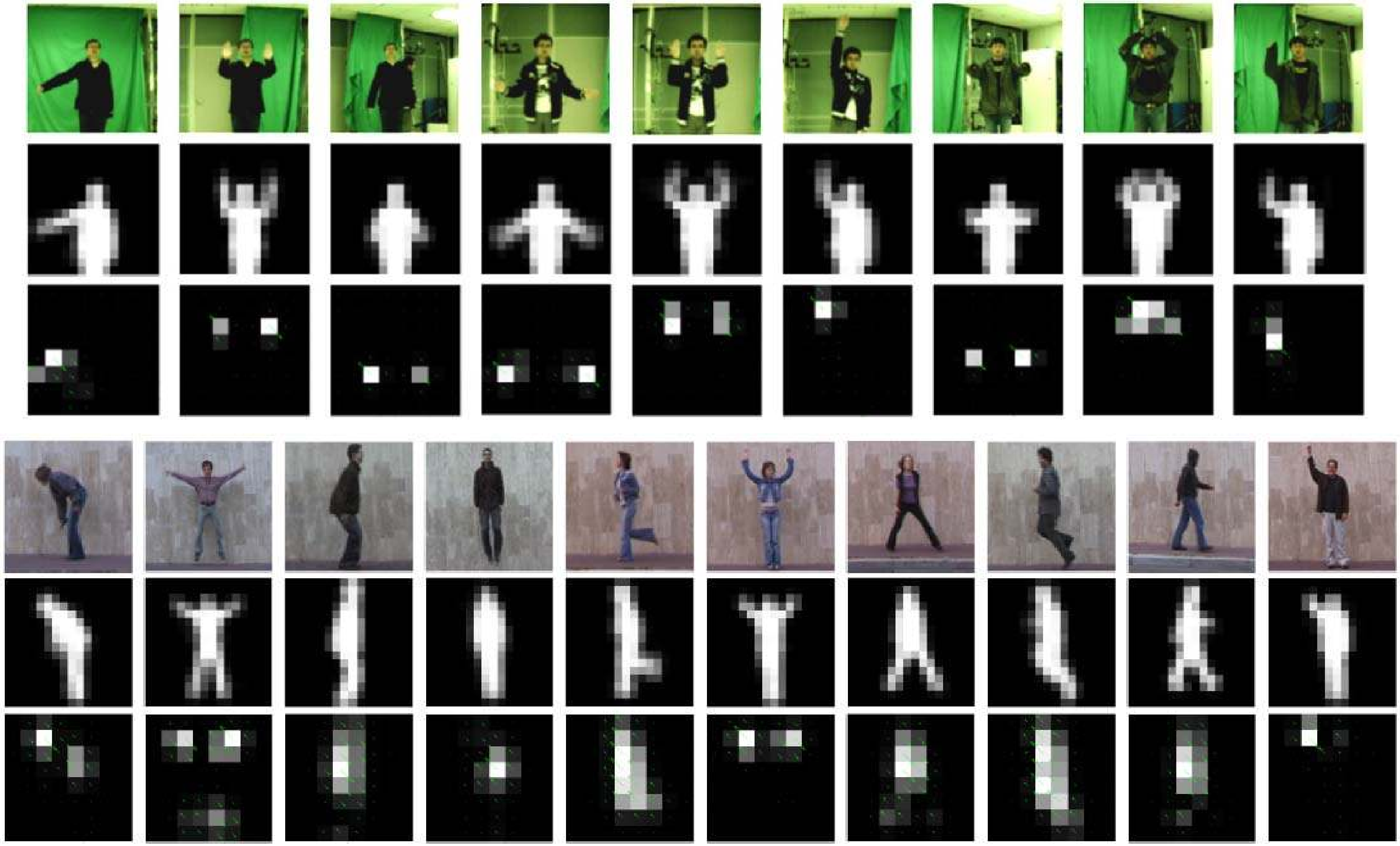
(c) s3 scenario using prototypes



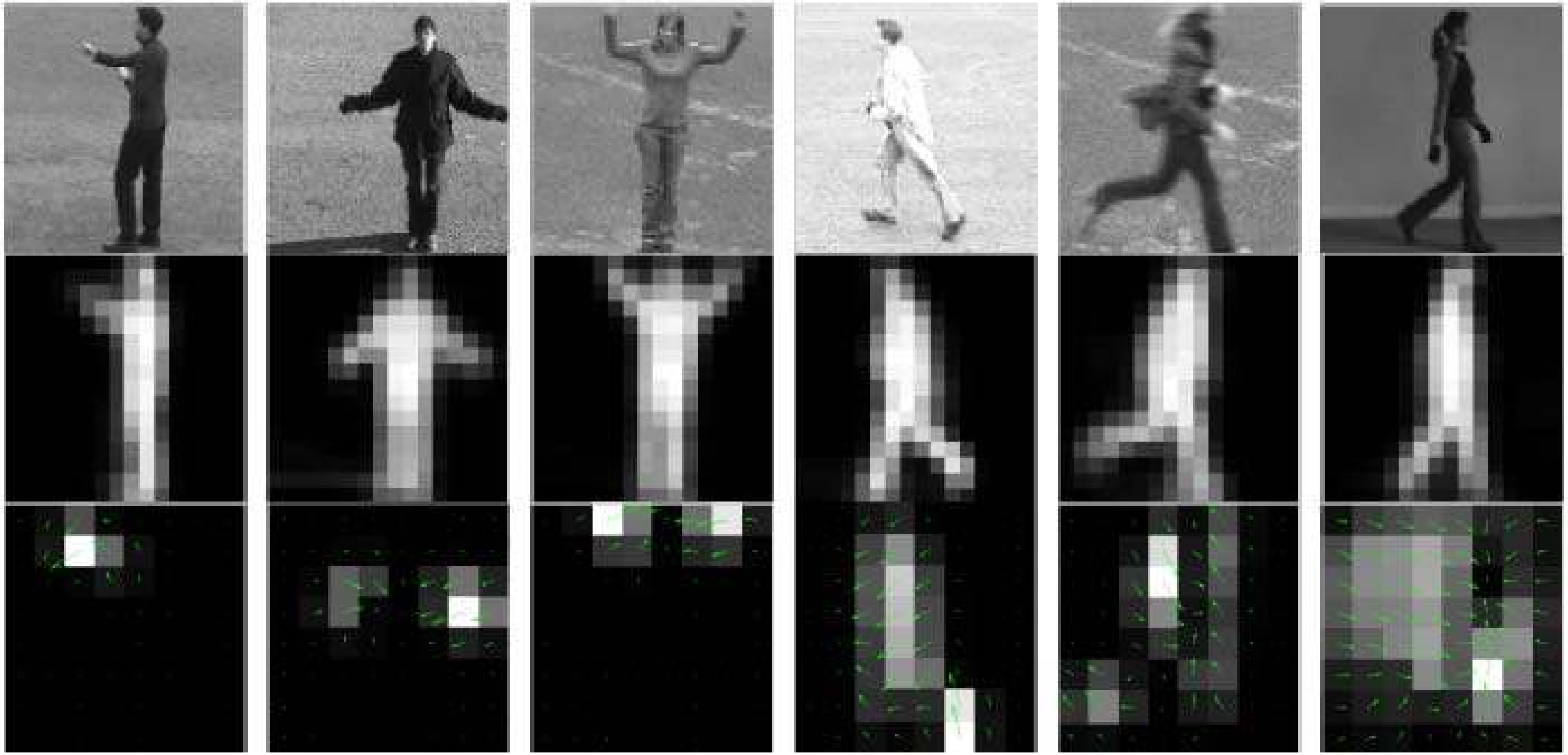
(d) s4 scenario using prototypes

Leave-one-person-out experiments

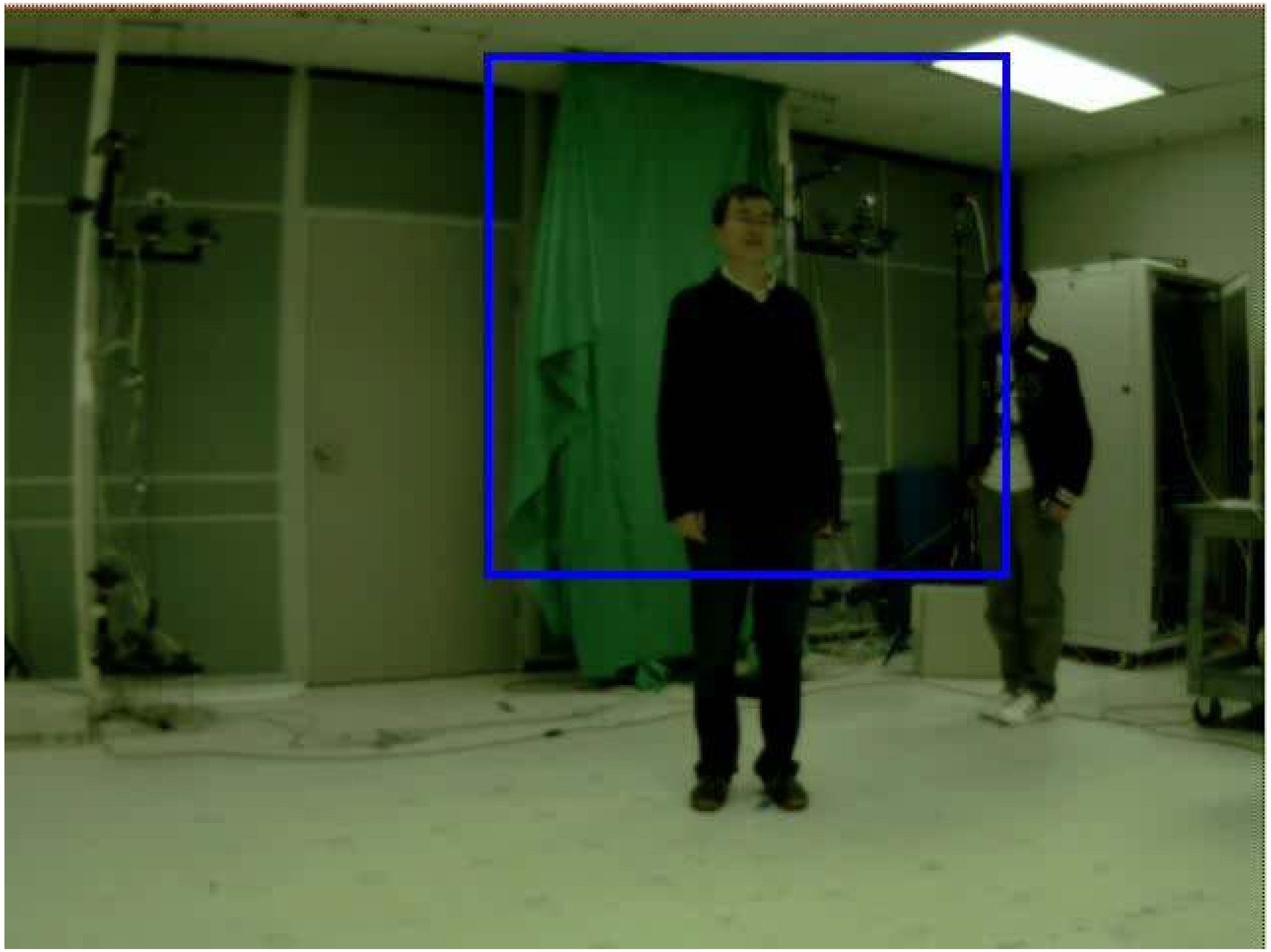
Results on Frame-to-Prototype Matching (Gesture & Weizmann)



Results on Frame-to-Prototype Matching (KTH)



Action Recognition Demo



Summary on Action Recognition

- Conclusions
  - The approach learns action prototypes in a joint shape and motion space to perform accurate and efficient action recognition.
  - The approach can handle challenging cases, such as moving camera and dynamic background.

- Future work
  - Discriminative feature and prototype learning algorithms for improving recognition performance (SVM, Adaboost).
  - Simultaneous action detection and recognition based on hierarchical shape-motion models.

Conclusion and Future Work

- Even though robust performance can be obtained in these fundamental components, there are still many unsolved problems.
  - Incorporation of scene-specific cues or high-level spatial or temporal contexts would make human movement analysis more reliable and accurate.
  - Integrating the fundamental components into a robust surveillance system working in challenging real-world scenarios.