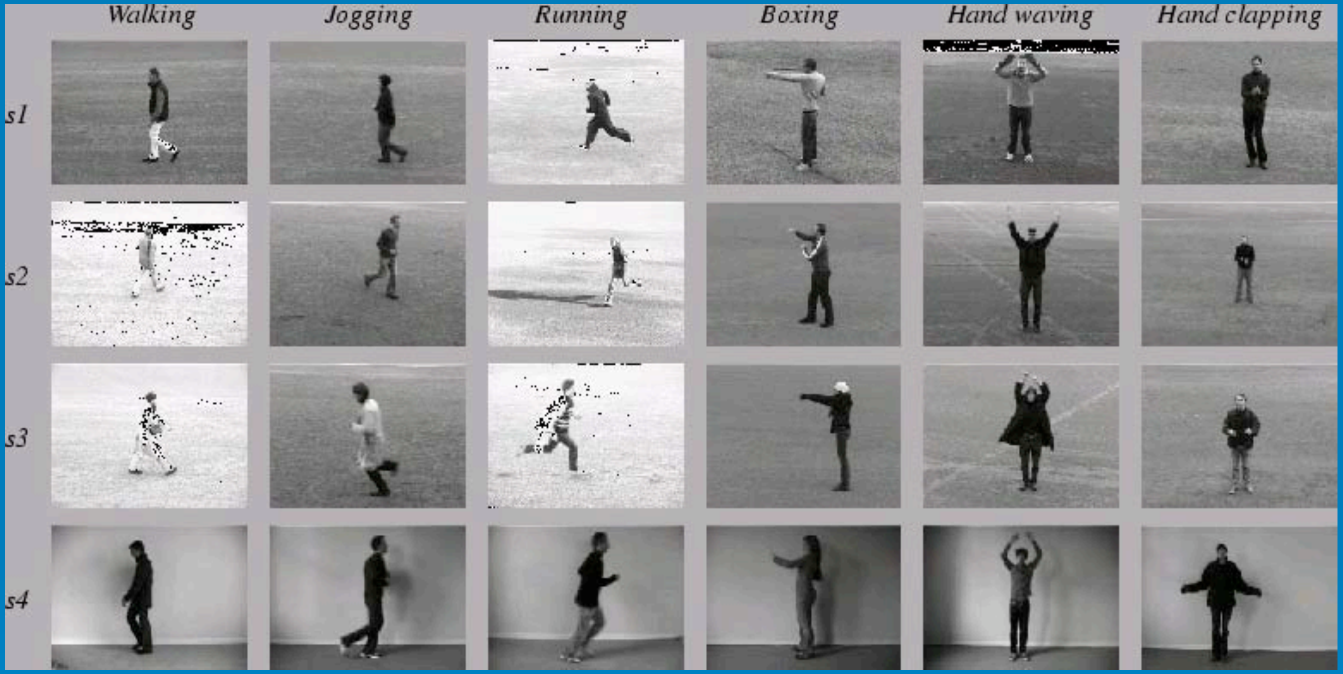
View-Independent Action Recognition

Daniel Weinland, Edmond Boyer, Remi Ronfard

INRIA, Rhone-Alpes, France.

Problem

- Action Recognition



[Schuldt04]

- Common assumptions on view:
  - Views are fronto-parallel.
  - Actors face camera / parallel to viewing plane.

Problem

- Application to Realistic Scenarios?
  - Arbitrary viewpoints.
  - No constraints on the actor’s orientation.

Overview

- Background and Related Work
  - Applications
  - Model based vs. template based approaches
  - View independence
- Motion History Volumes (MHVs)
  - 3D action representation → multiple cameras
  - Invariant representation → Fourier descriptors
- Single view recognition using 3D exemplar model
  - Learning: 3D model / multiple cameras
  - Recognition: 2D / single view
  - Probabilistic modelling of view changes

Applications

Entertainment

HCI

Surveillance

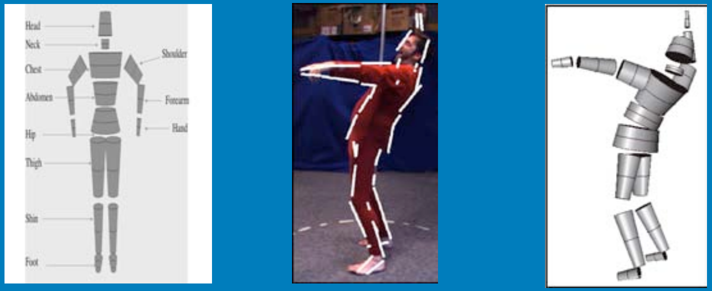
Ambient Intelligence

Sports

Group Actions

Related work:
Two Main Directions

Model based:



[Knossow06]

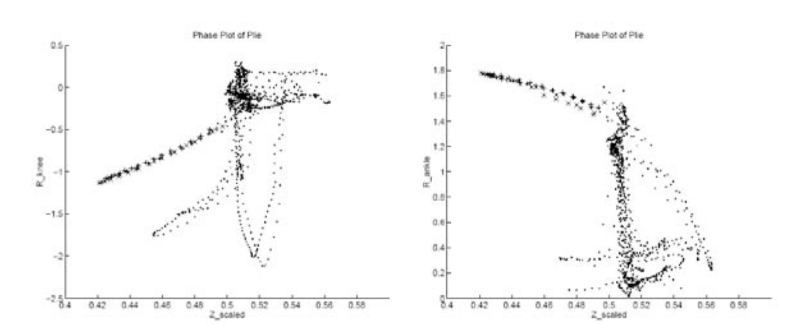
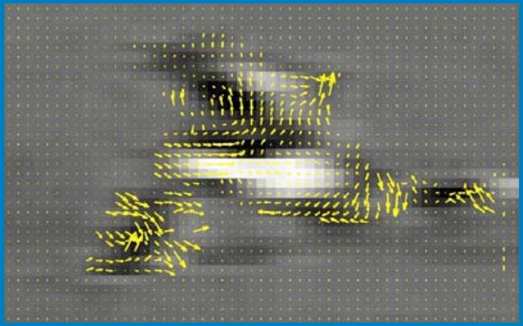


Figure 2: × and + marks time steps during pliés for two dancers; "." marks time steps during other movements. Angles are in radians. These two views of the phase space show how the plié lies in a region separated from the other movements (data subsampled for clarity).

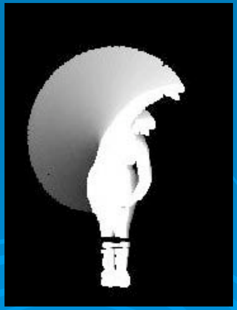
[Campbell95]

Template based:



Optical Flow [Efros03]

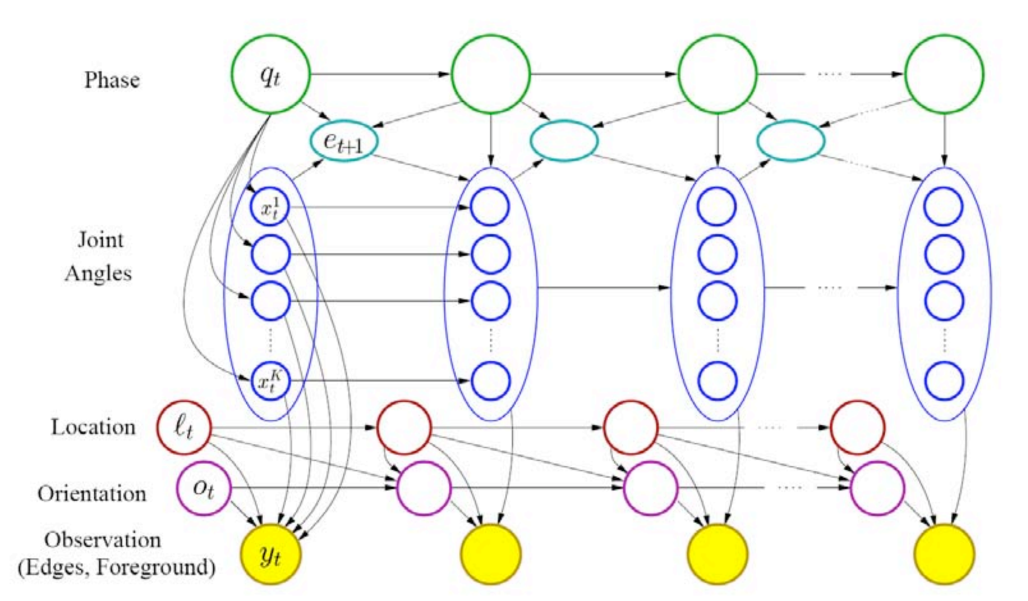
Space-Time Volumes [Blank05]

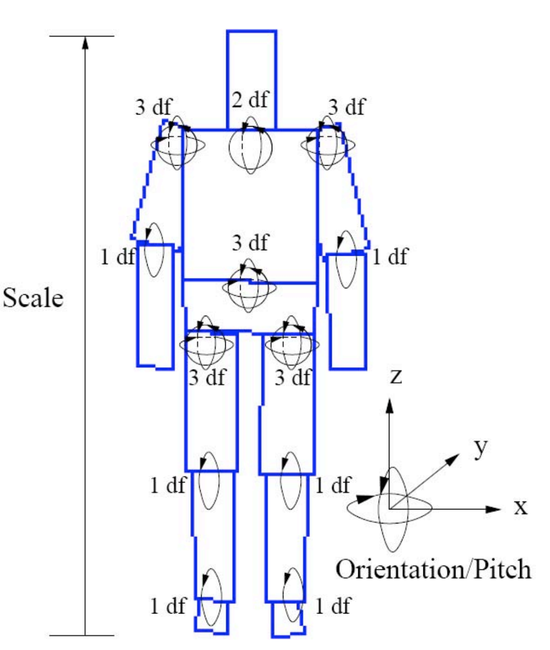


Motion History Images [Bobick96]

View independence: Model based

- ↑ Can model orientation as independent variable.
- ↓ Joint model is difficult to extract.

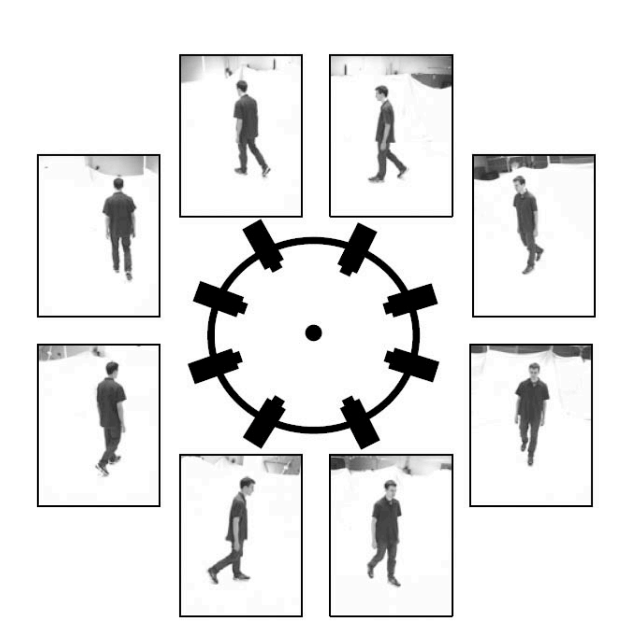




[Peursum07]

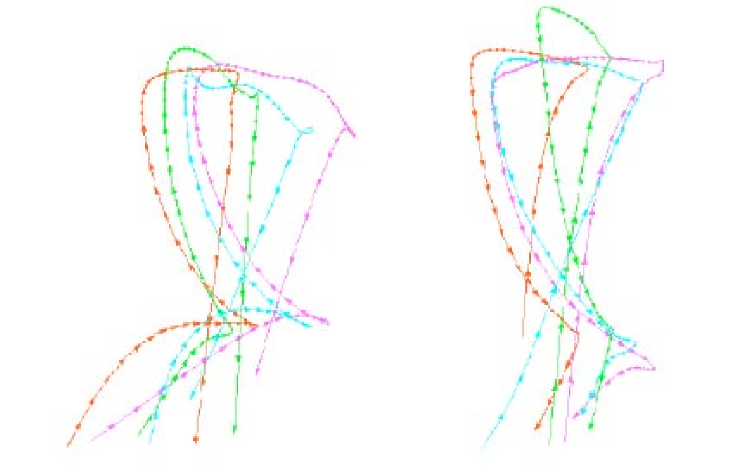
View Independence: Template Based

One model per view



- ↓ Views are independent → no modelling of temporal change in view.
- ↓ Limited number of views and fixed camera setup.

Invariance

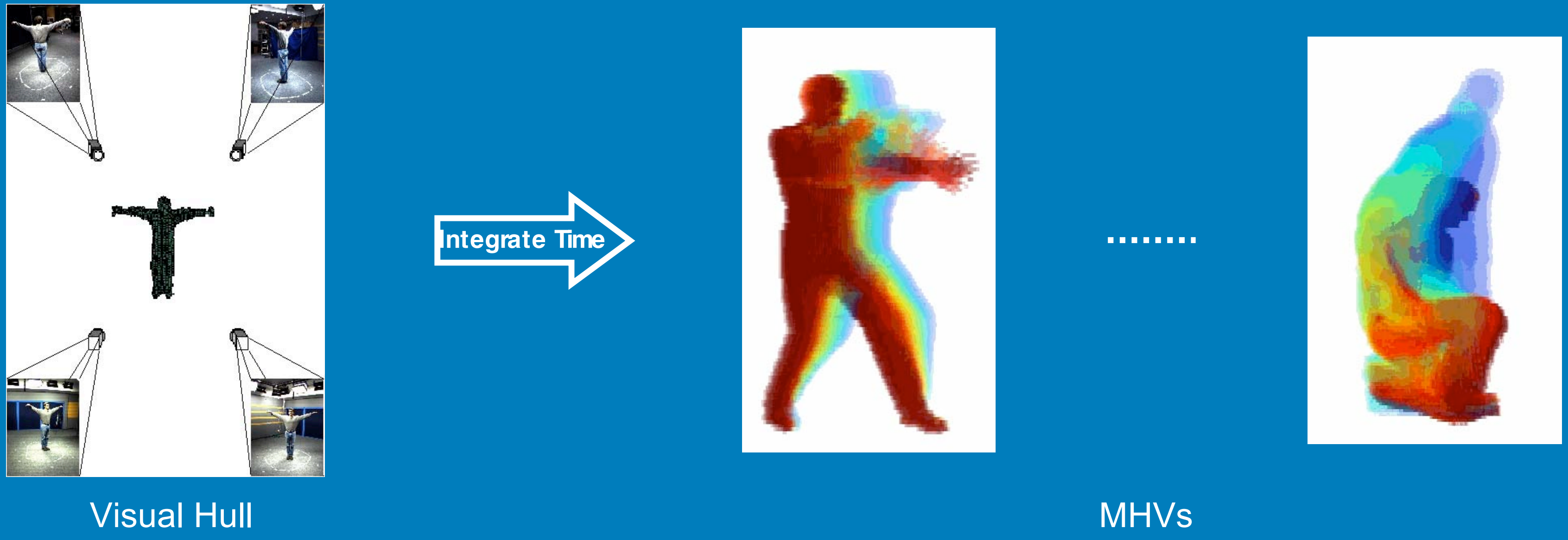


[Rao&Shah01]

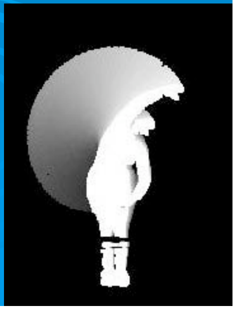
- ➢ In 2D using linear dependence in 3D (rank constraints / factorization)
- ↓ Many Ambiguities

↑ Effective and robust to obtain

Motion History Volumes



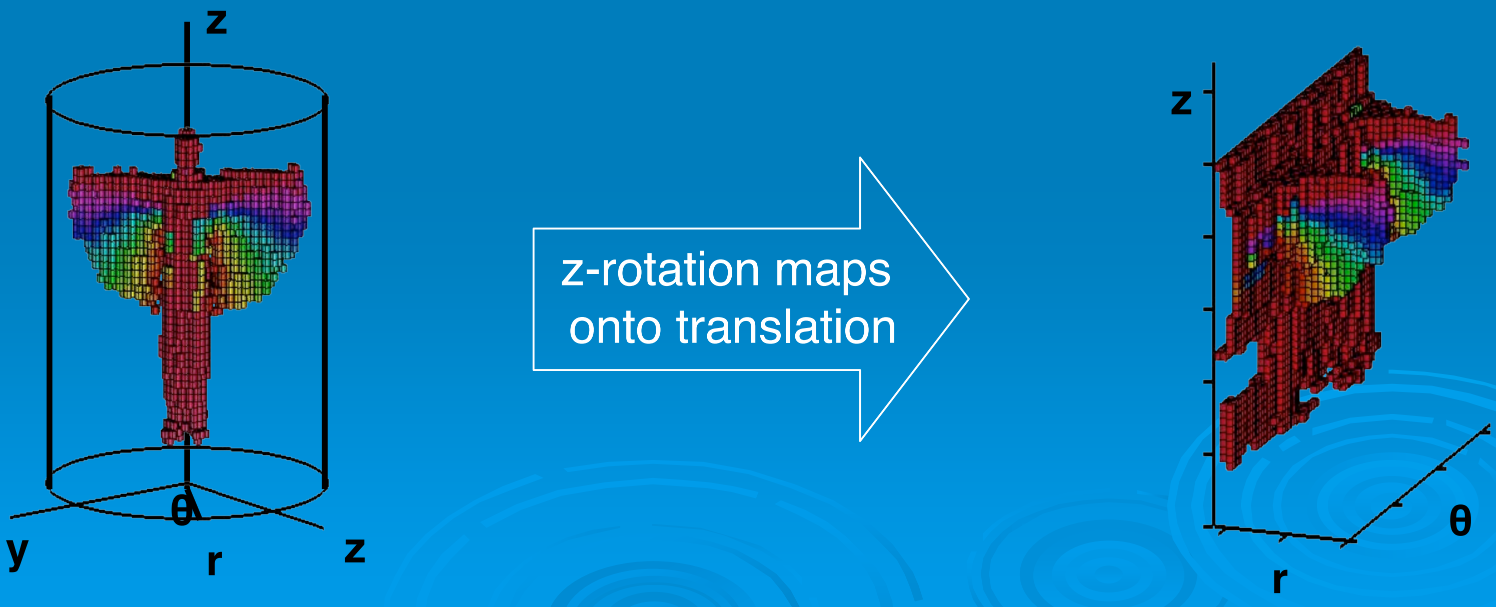
- Extends Motion History Images to volumetric representation.
- Template representation → Joint model free.
- Needs multiple calibrated cameras.



View Invariance: Cylindrical Representation

- Assumption: For similar actions main difference in scale, translation, and rotation around vertical axis.

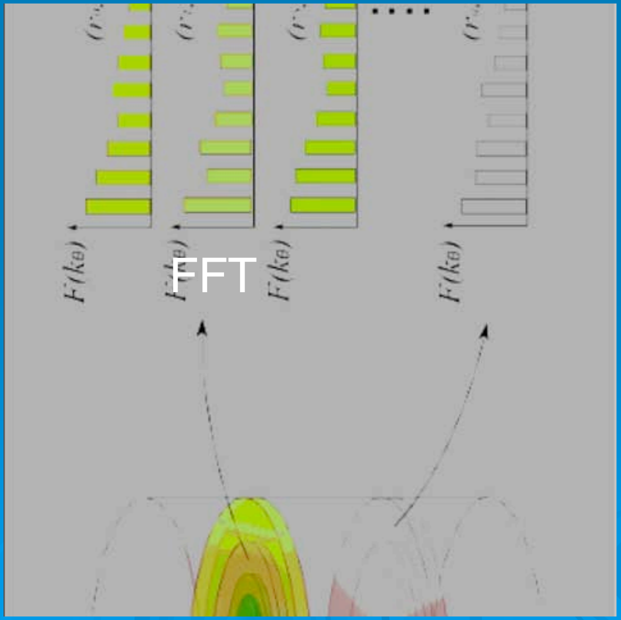
- Map into normalized and object-centred cylindrical coordinate system.



z-rotation maps onto translation

View Invariance:
Fourier Descriptors

- Magnitudes of Fourier-transform invariant to translation shifts *(Fourier-shift theorem).*
- Features based on Fourier-magnitudes over θ for each value r, z.



Feature Vector

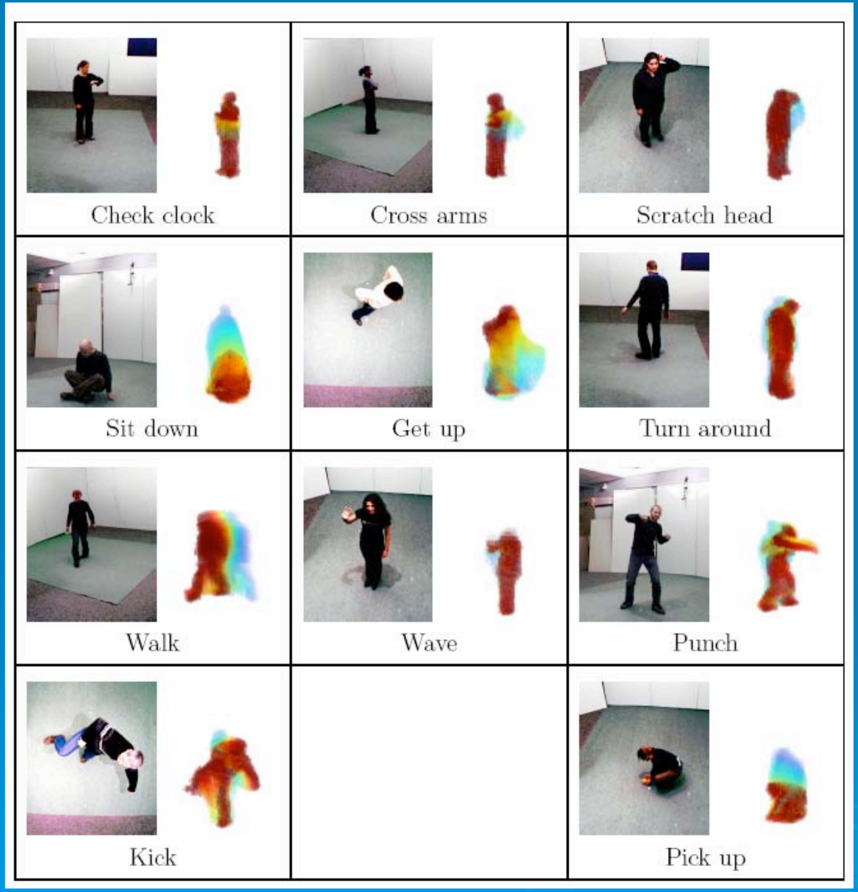
Learning Actions

- Variations in body and style:
  - Learn models of motion.

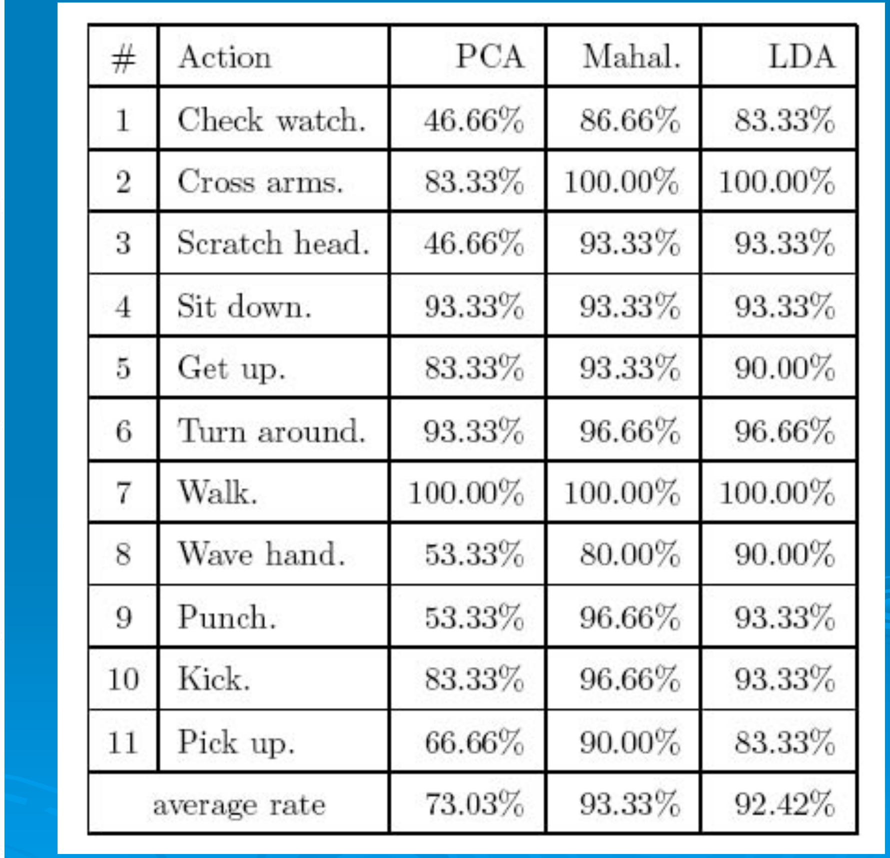
- Simple approach:
  - Each class is represented by it’s mean value.
  - Closest mean assignment.
  - Dimensional reduction using normalized principal component analysis (PCA) or linear discriminant analysis (LDA).
  - Leave one out validation.

Results

ixmas dataset (*https://charibdis.inrialpes.fr/*)
5 cameras, 330 samples (11 actions, 10 actors, 3 executions)



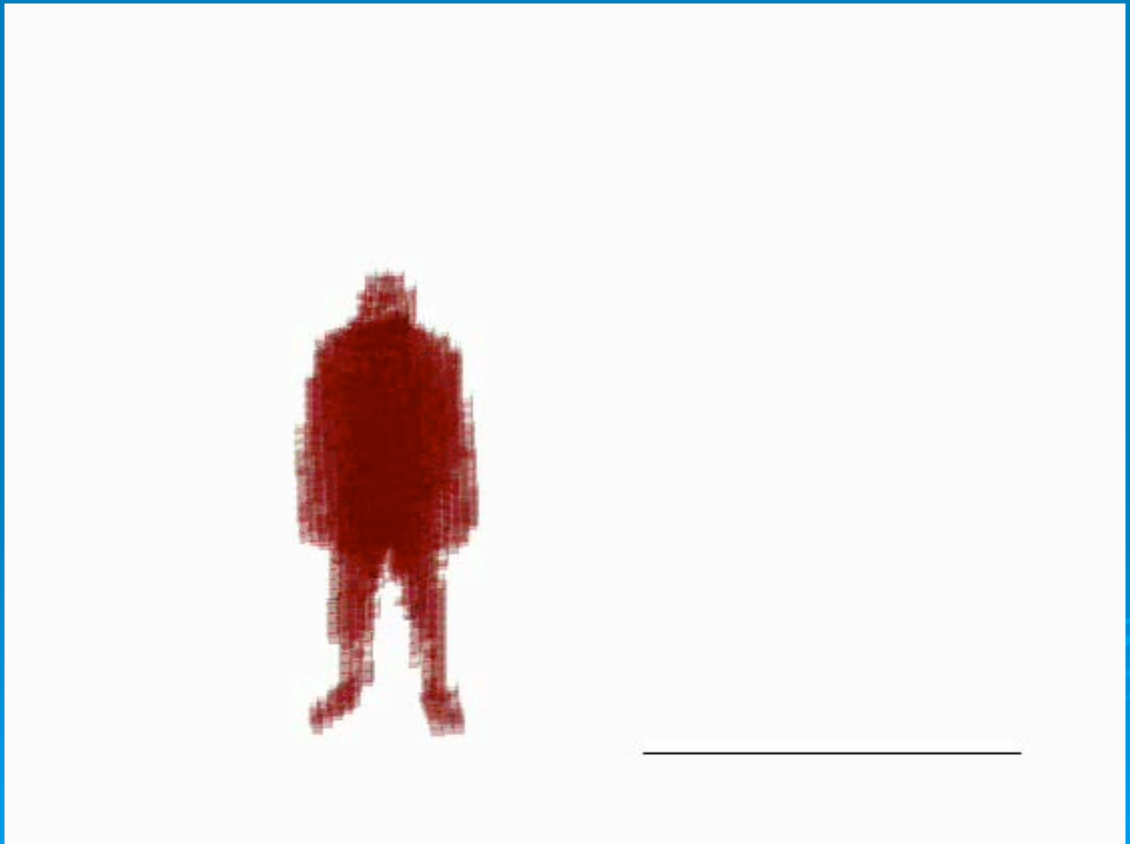
Check clock
Cross arms
Scratch head
Sit down
Get up
Turn around
Walk
Wave
Punch
Kick
Pick up



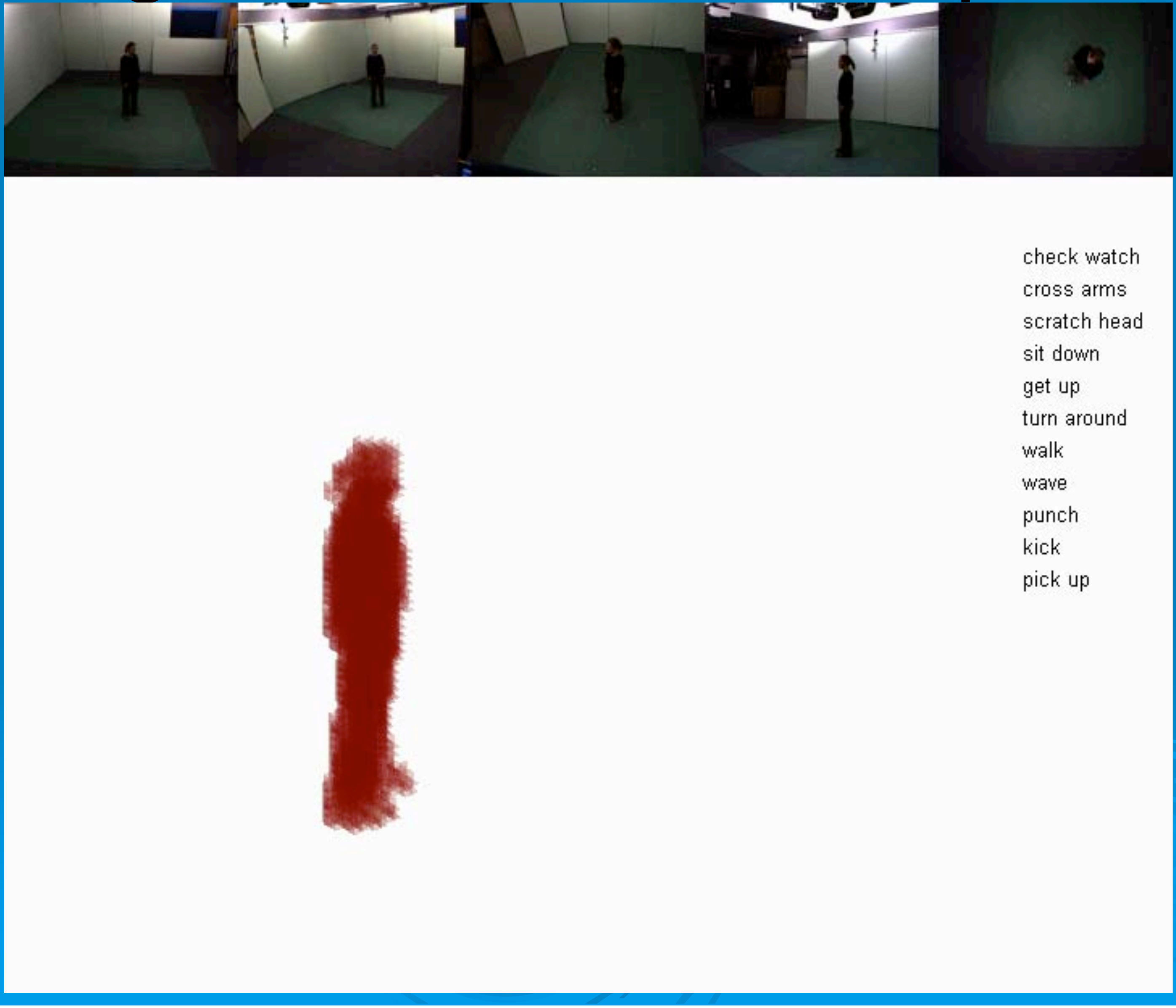
#	Action	PCA	Mahal.	LDA
1	Check watch.	46.66%	86.66%	83.33%
2	Cross arms.	83.33%	100.00%	100.00%
3	Scratch head.	46.66%	93.33%	93.33%
4	Sit down.	93.33%	93.33%	93.33%
5	Get up.	83.33%	93.33%	90.00%
6	Turn around.	93.33%	96.66%	96.66%
7	Walk.	100.00%	100.00%	100.00%
8	Wave hand.	53.33%	80.00%	90.00%
9	Punch.	53.33%	96.66%	93.33%
10	Kick.	83.33%	96.66%	93.33%
11	Pick up.	66.66%	90.00%	83.33%
average rate		73.03%	93.33%	92.42%

Temporal Segmentation based on Motion Energy

- “Global velocity” based on MHVs in constant time window.
- Segmentation criteria = local energy Minima.



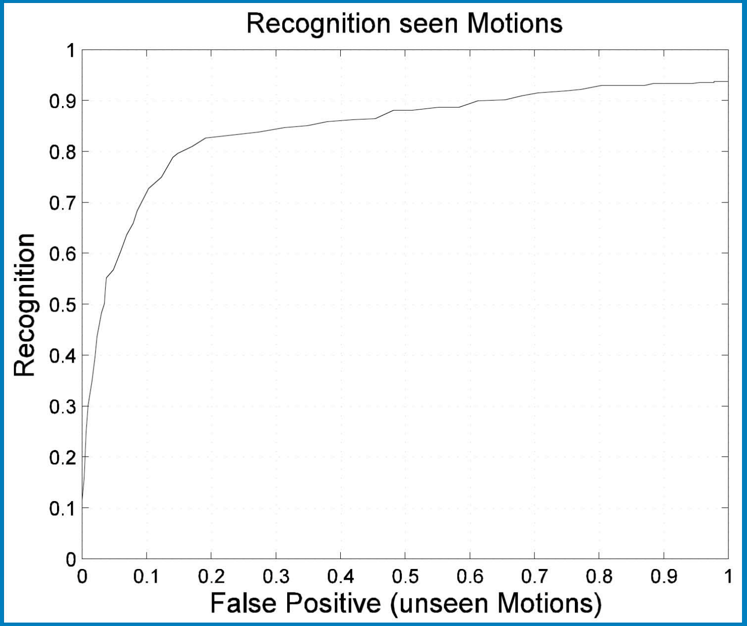
Recognition on Raw Sequences



Recognition on Videos

- Use threshold to detect actions.
- Everything lower → "garbage"-class.

- 23 minutes of video, 1188 templates:
  - 82.79% overall rate.
  - 78.79% recognized.
  - 14.08% false positive.



☝ Modelling garbage-class difficult / few existing solutions!

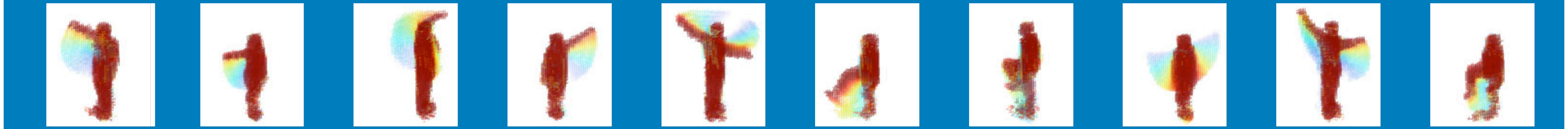
Automatic Discovery of Action Taxonomies

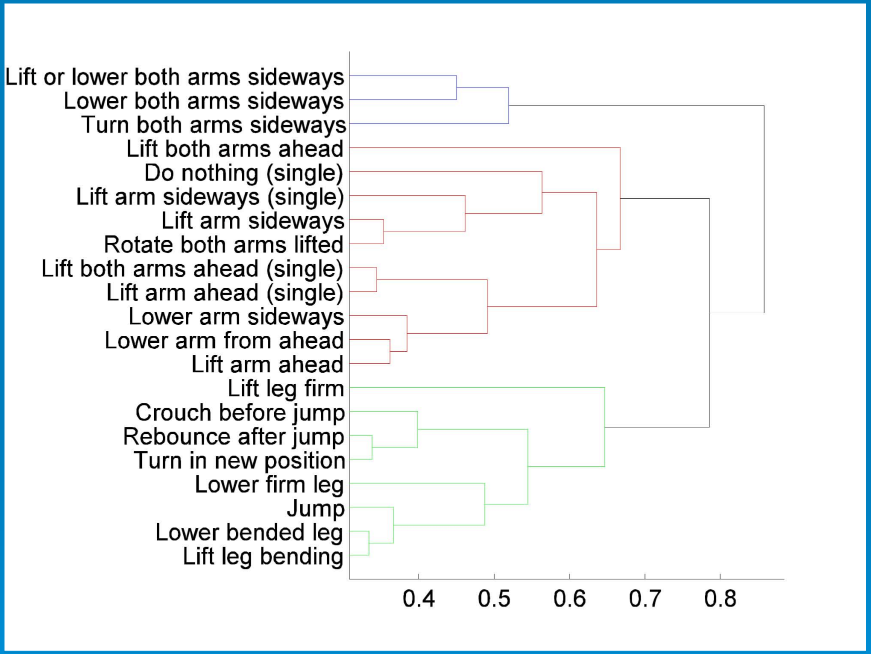
- We have:
  - Discriminative motion descriptor (MHVs)
  - Automatic motion segmentation

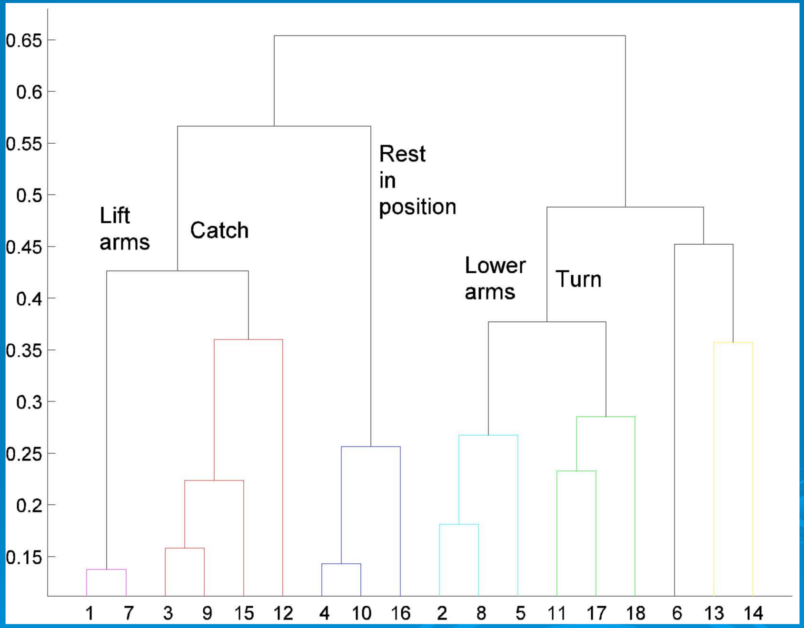
➔ Semi-supervised action recognition:

- Segment and cluster complex actions → discover motion primitives

Experiments

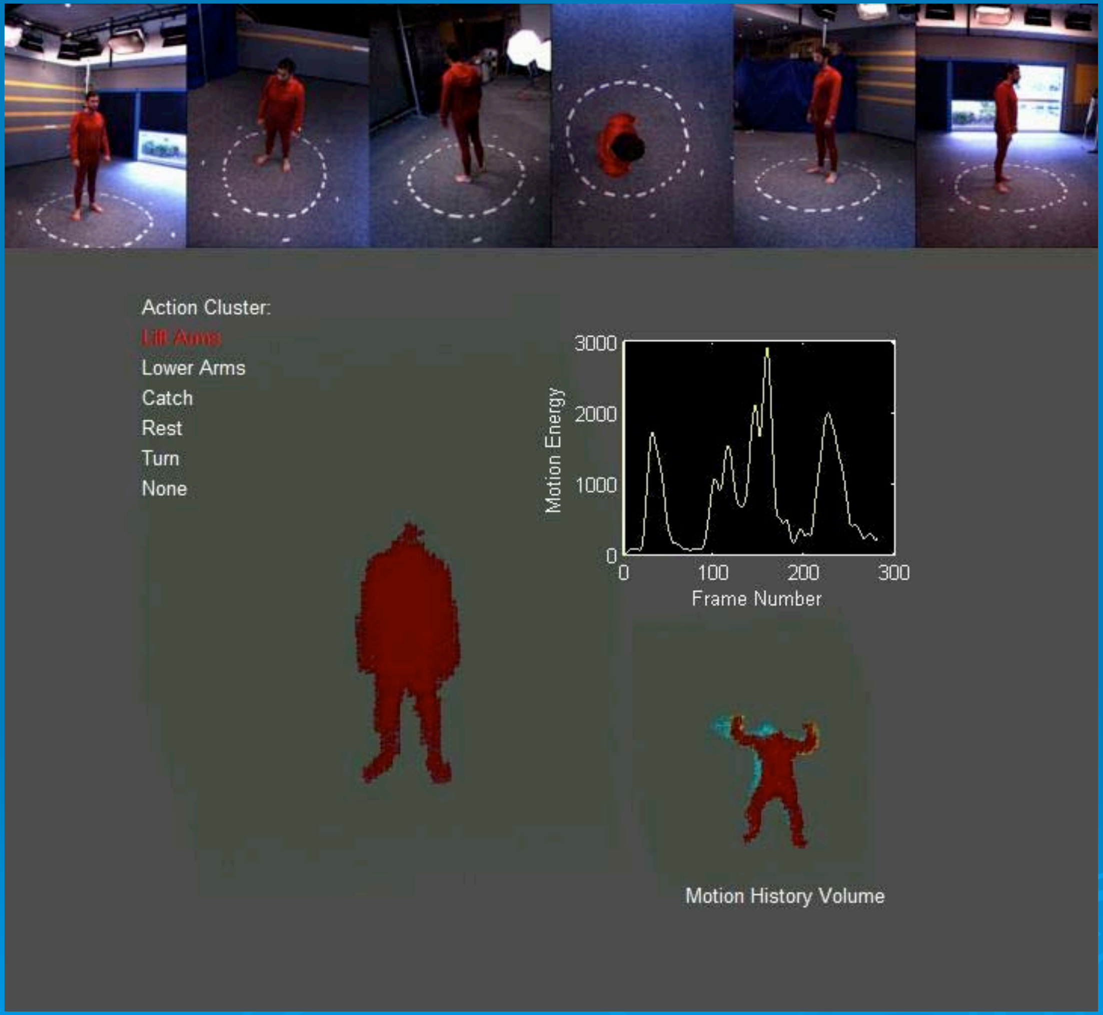




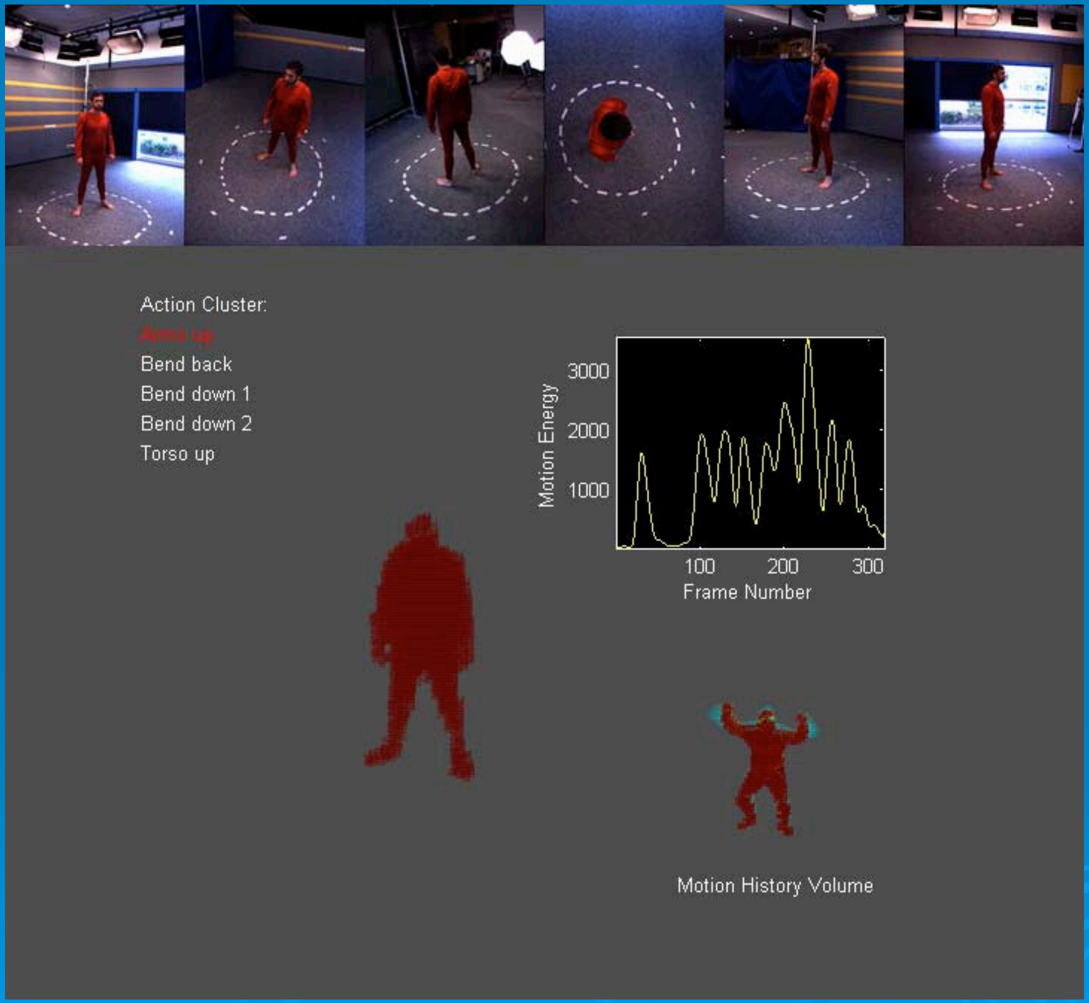


- Clustering on Primitive motions (103 sequences).
- Clustering on composite actions.

Clustering on Composite Actions



Clustering (2)



Action Cluster:

Arms up

Bend back

Bend down 1

Bend down 2

Torso up

Part 2

Action Recognition from Arbitrary Views using 3D Exemplars

Daniel Weinland, Edmond Boyer, Remi Ronfard
INRIA, Rhone-Alpes, France.

Problem

- Application to Realistic Scenarios?
  - Arbitrary viewpoints.
  - No constraints on the actor’s orientation.

Problem

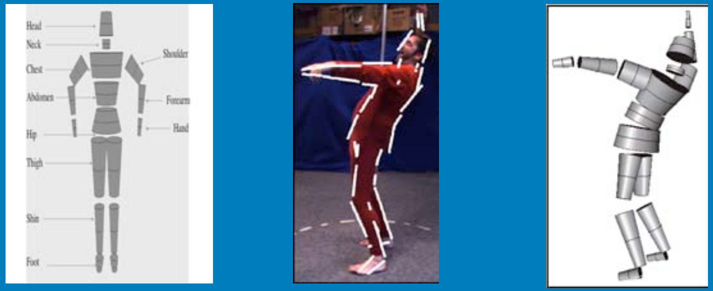
- Application to Realistic Scenarios?
  - Arbitrary viewpoints.
  - No constraints on the actor’s orientation.

- MHVs:
  - ↑ View-invariant.
  - ↓ Need 3D reconstruction from multiple calibrated cameras.

- Idea:
  - A method that can recognize actions from arbitrary number and configuration of cameras (even a single!)
  - Still use 3D during learning.

Related work:
Two Main Directions

Model based:



[Knossow06]

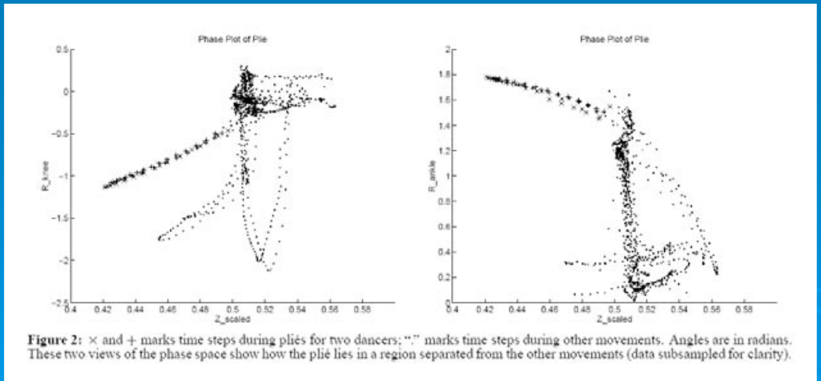
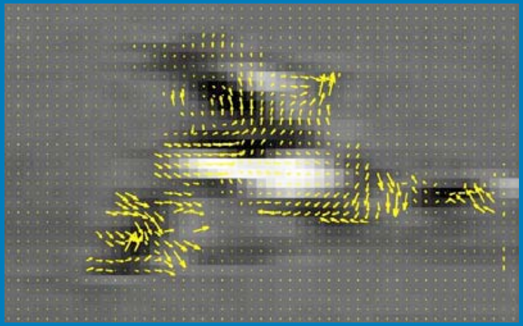


Figure 2: × and + marks time steps during pliés for two dancers; "." marks time steps during other movements. Angles are in radians. These two views of the phase space show how the plié lies in a region separated from the other movements (data subsampled for clarity).

[Campbell95]

Template based:



Optical Flow [Efros03]

Space-Time Volumes [Blank05]

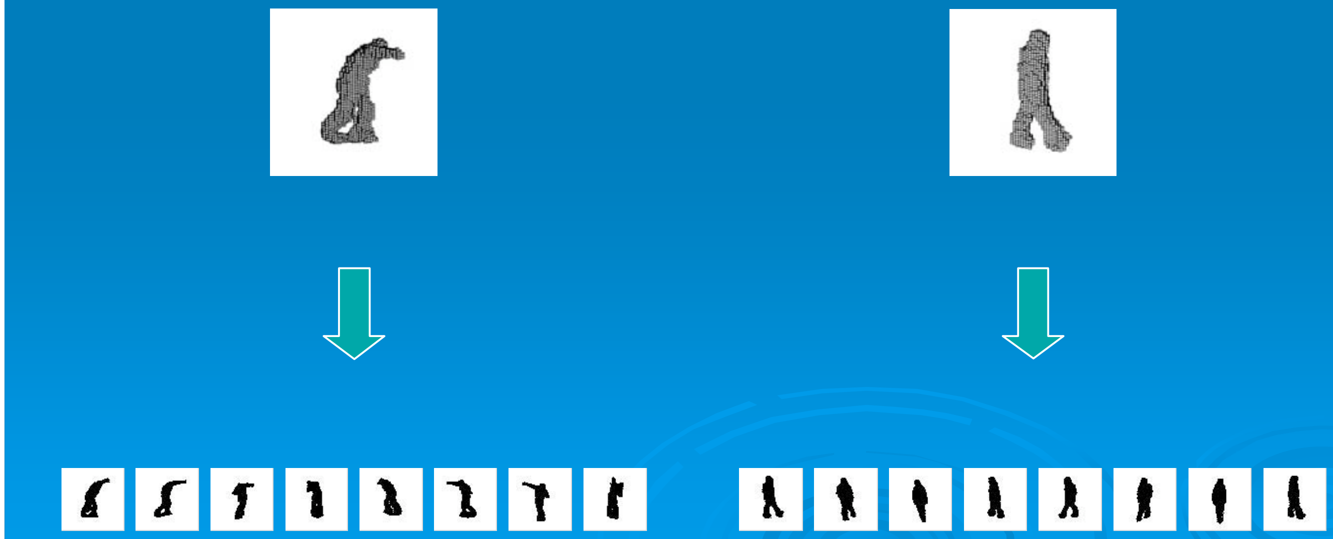
Motion History Volumes [weinland05]

Idea

- Take advantages of both directions:
  - Template based: Effective features, no joint space modelling
  - Model based: Simple generative modelling of view and orientation (by projecting a 3D model into 2D).

- Explicit modelling of view transformation as latent variable [Frey & Jojic 2000, Toyama & Blake 2001]

→ A 3D template based model that generates arbitrary 2D views.



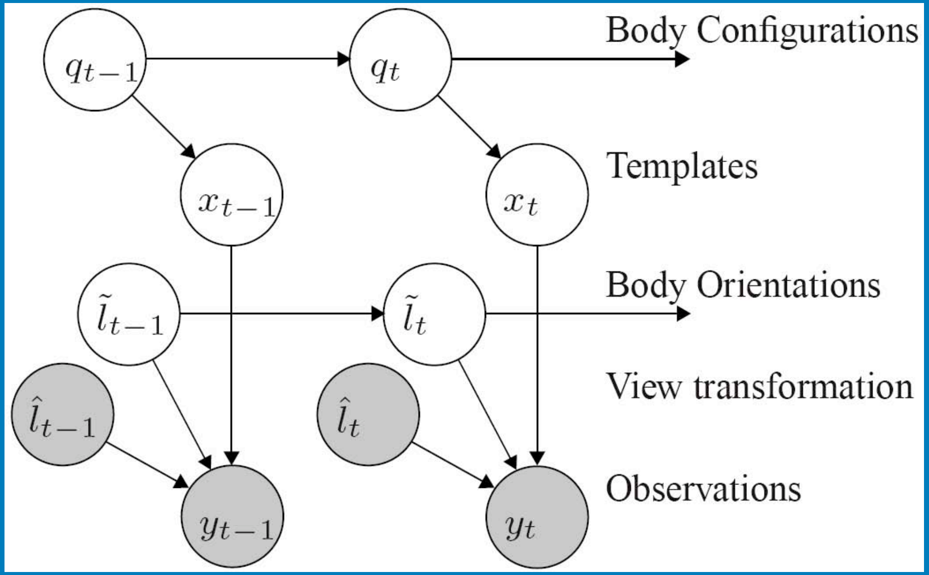
Exemplars (Visual Hulls)

$X = f\, x_{i\,2\,[1:::M\,]}\, g$

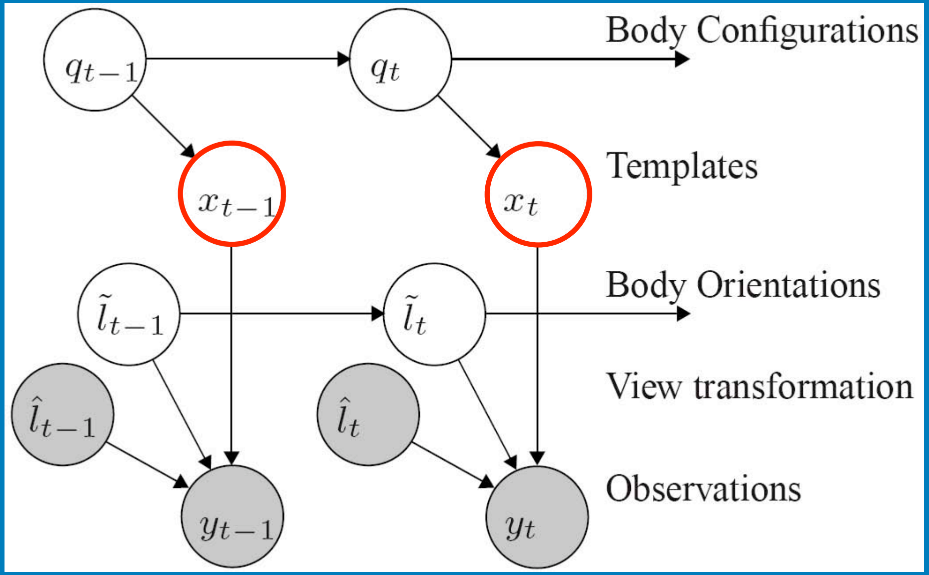
$P_{\hat{\mu}}(x) = \hat{P}[R_{\mu}; u]x$

Projections (Silhouettes)

Probabilistic Model of Actions



Probabilistic Model of Actions

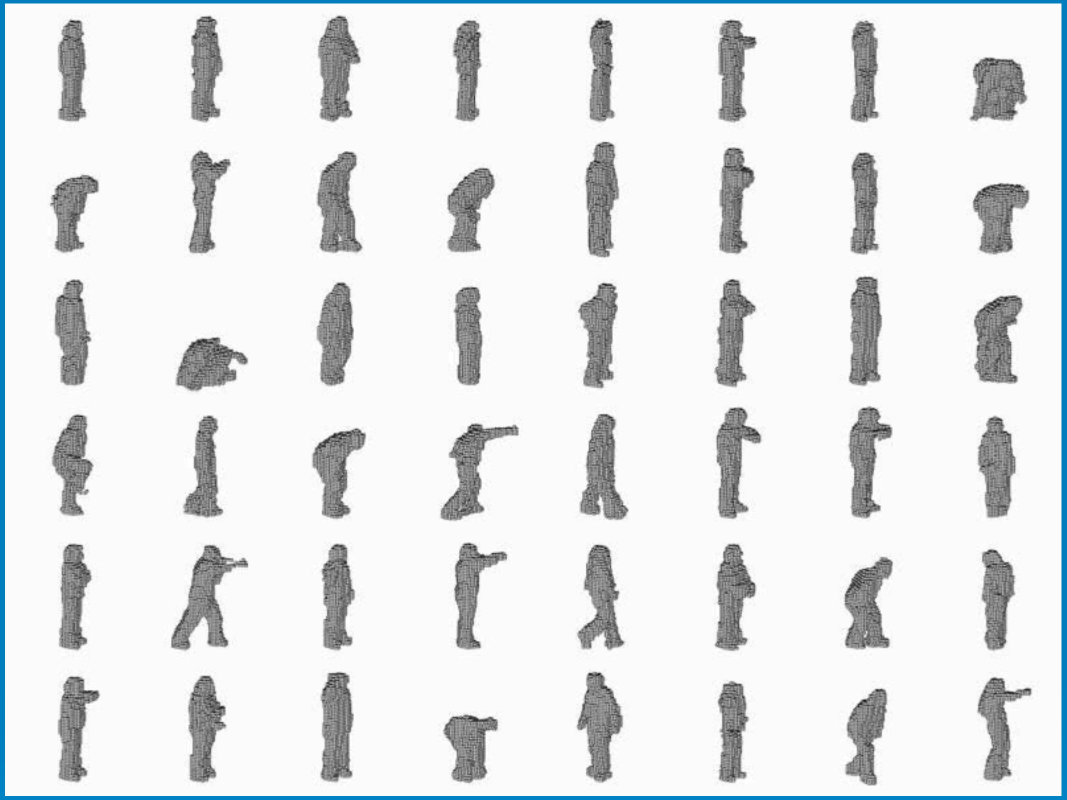


➢ Templates

- Not result from body models and joint configurations.
- Represented by a set of M exemplary templates: $X = f x_{i\,2\,[1:::M\,]} g,$

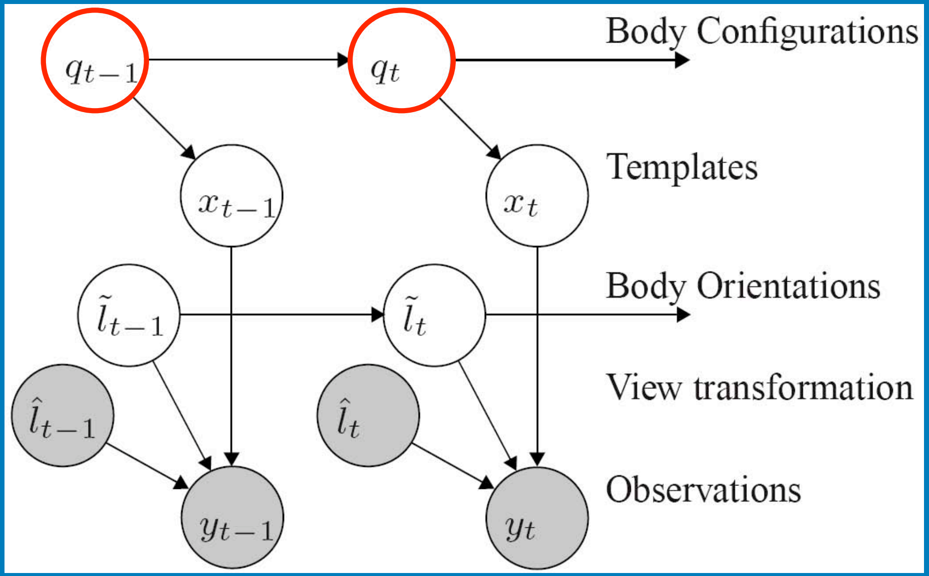
  learned from three dimensional training sequences.

Visual Hull Exemplars



- 3D and 2D features are *geometrically consistent* → 2D templates are obtained simply by projecting 3D templates.
- Silhouettes sequences are discriminative with respect to actions.
- Powerful distance functions exist, e.g. chamfer distance.

Probabilistic Model of Actions

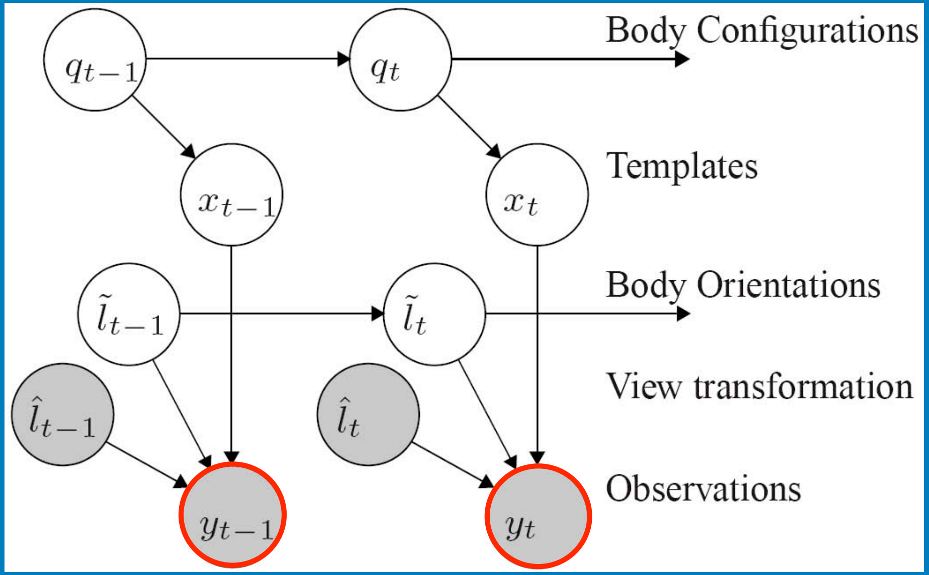


- Hidden State Sequence (Action Dynamics)
  - Discrete N-state latent variable q
  - Follows a first order Markov chain

    $$p(q_t | q_{t-1}, \ldots, q_1) = p(q_t | q_{t-1})$$

  - Intuitively: a quantization of the joint motion space into action-characteristic configurations

Probabilistic Model of Actions

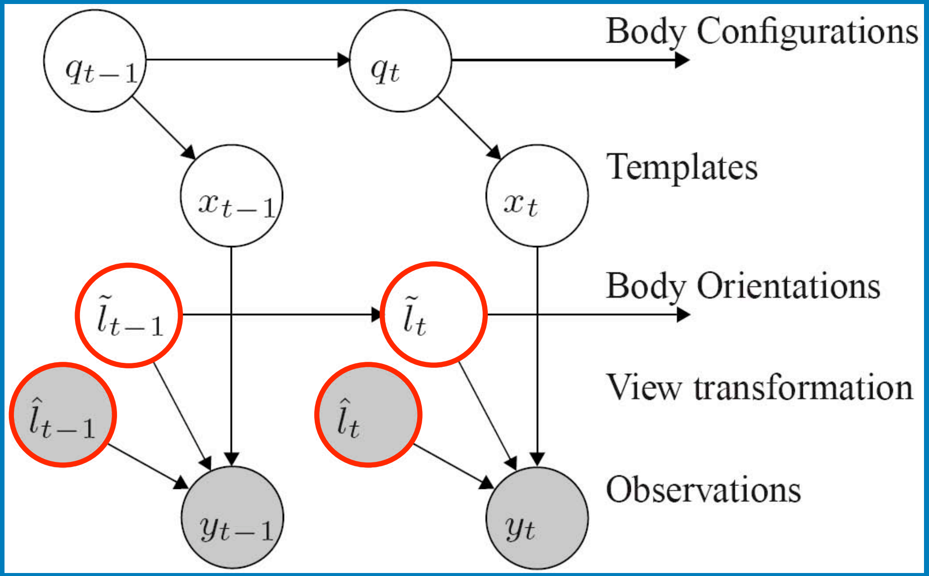


- Observations
  - 2D observations y_t result from a geometric transformation P of the 3D exemplars X:

$$p(y_t | x_t = x_i; \hat{l}_t; \tilde{l}_t) \propto \frac{1}{Z} \exp\left(- d(y_t; P_{\hat{l}\tilde{l}}(x_i))/\sigma^2 \right)$$

  - d is distance function (e.g. Euclidean or Chamfer distance),
  - σ is scale → non-parametric modeling (Parzen)

Probabilistic Model of Actions



- Geometric Transformation

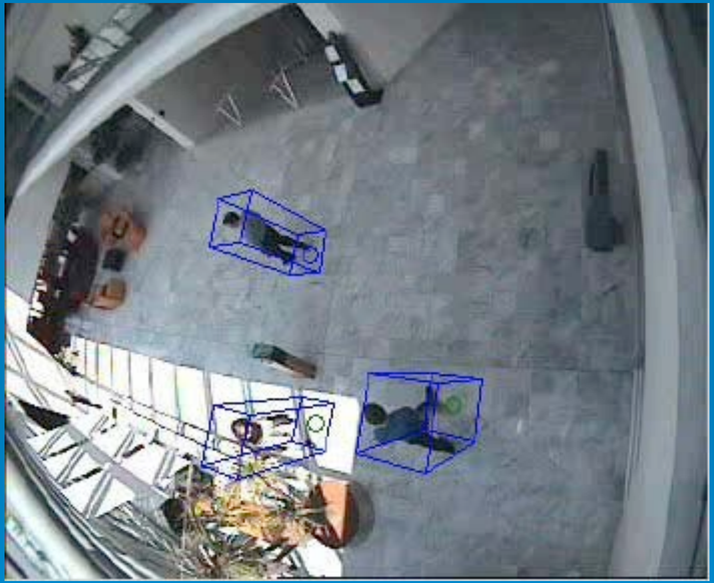
$$P_{\hat{l}\tilde{l}}(x) = \hat{P}[R_{\mu}; u]x$$

- $\hat{l}$: observed parameters: camera calibration P, position u
- $\tilde{l}$: latent parameters: body orientation = rotation R around vertical axis, follow a first order Markov process: $p(\tilde{l}_t | \tilde{l}_{t-1})$

Observed Parameters

- Camera calibration
- Position on Ground

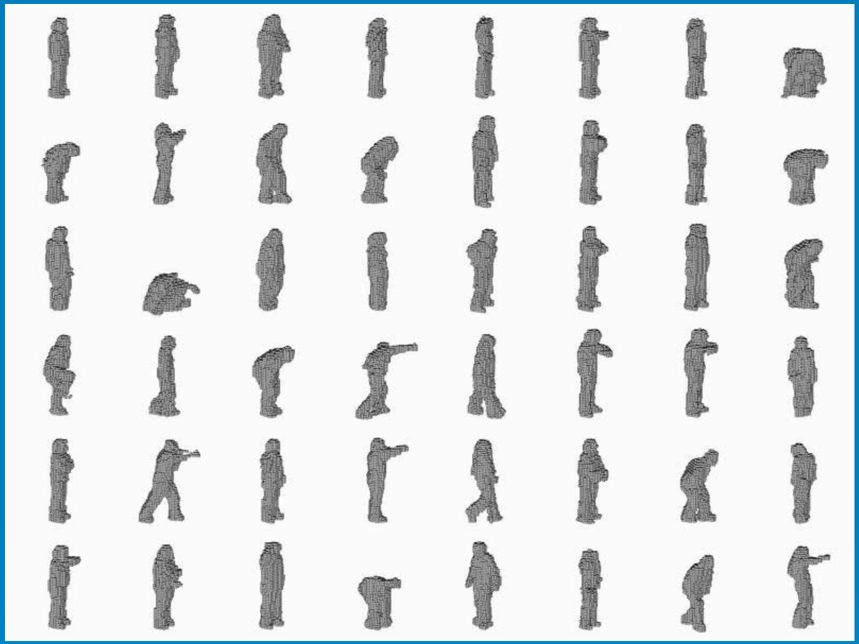
→ Pose



→ Reduce parameters during Marginalization

Hidden Parameters

- Body orientation is hidden



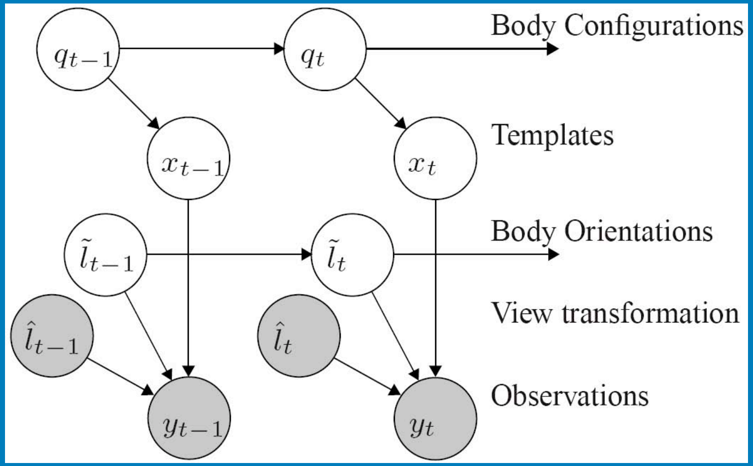
→ We have to marginalize over all possible values!

Action Recognition

- C action classes → C parallel “HMMs”
- each class is represented by

¸ c2 f 1;::::;C g

$$= f p(q_t j q_{t\text{¡}\ 1}; c); p(q_1 j c); p(x_t j q_t; c)g$$



- all actions {c = 1..C} share a common set of exemplars and kernel parameters (tied mixture HMM).
- a sequence of observations Y={y_1,…,y_T} is classified with respect to the maximum a posteriori (MAP) estimate:

$$g(Y) = \text{ar g max}_c p(Y j ¸_c) p(¸_c)$$

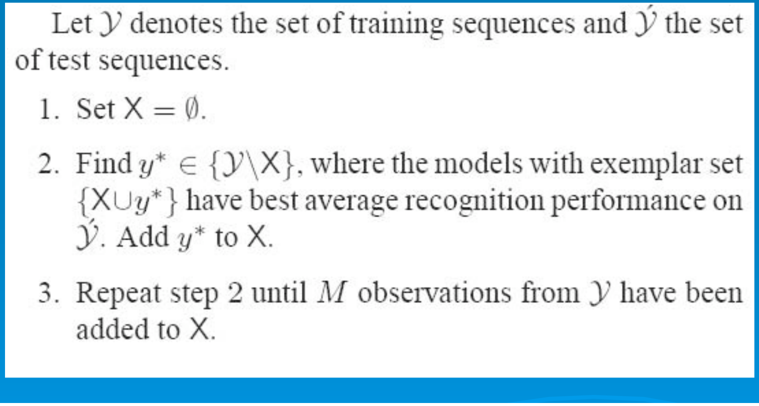
Model Learning

- Models Learning consists 2 main operations:
  - selecting or identifying exemplars
  - learning probabilities
- Both steps are coupled.

Exemplar Selection

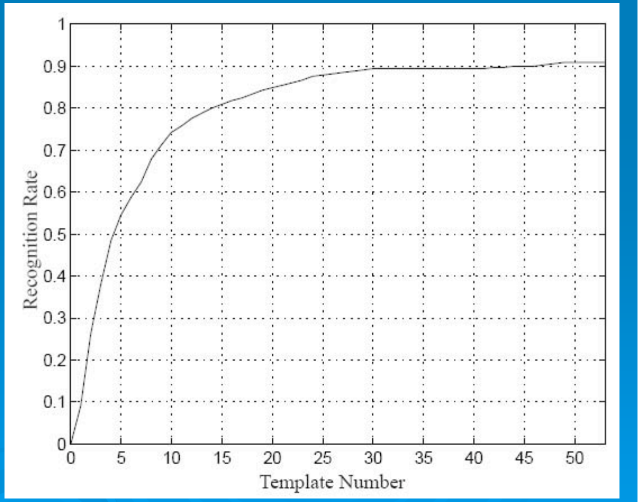
- How to identify discriminative exemplars?
- Clustering (e.g. k-means) tend to cluster different poses performed by similar actors rather than similar poses performed by different actors.

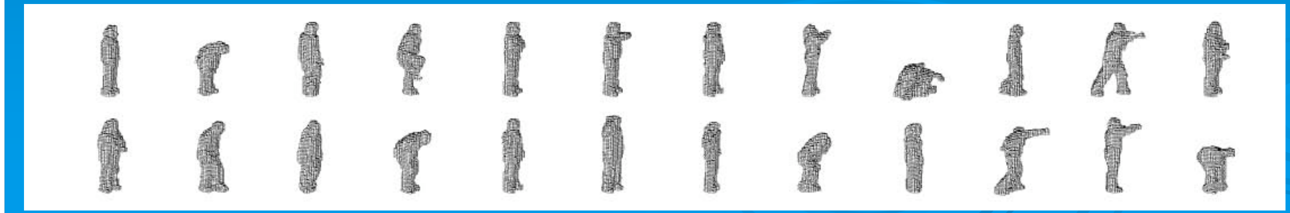
→ discriminative feature selection approach trough combinatorial subset selection: Wrapper - forward selection



Let $\mathcal{Y}$ denotes the set of training sequences and $\acute{\mathcal{Y}}$ the set of test sequences.

1. Set $\mathsf{X} = \emptyset$.
2. Find $y^* \in \{\mathcal{Y} \backslash \mathsf{X}\}$, where the models with exemplar set $\{\mathsf{X} \cup y^*\}$ have best average recognition performance on $\acute{\mathcal{Y}}$. Add y^* to X.
3. Repeat step 2 until M observations from $\mathcal{Y}$ have been added to X.





Learning Probabilities

- In 3D independent from viewing process and under ideal conditions (aligned data).
- In 2D with conditions similar when learning or recognizing.

- Learning trough forward-backward algorithm (HMM).

Action Recognition

- Y is classified using the MAP estimate.
- Computed via forward variable (HMM):

$$\alpha(q_t | \lambda_c) = p(y_1, \ldots, y_t, q_t | \lambda_c)$$

$$p(Y | \lambda_c) = \sum_{q_T} \alpha(q_T | \lambda_c)$$

- Observations from Multiple cameras:

$$p(y_t^1, \ldots, y_t^K | x_t, \hat{\theta}_t, \tilde{\rho}_t) \propto \prod_{y_t^k}^{K} p(y_t^k | x_t, \hat{\theta}_t, \tilde{\rho}_t)$$

Experiments

Xmas dataset

5 cameras, 330 samples (11 actions, 10 actors, 3 executions)

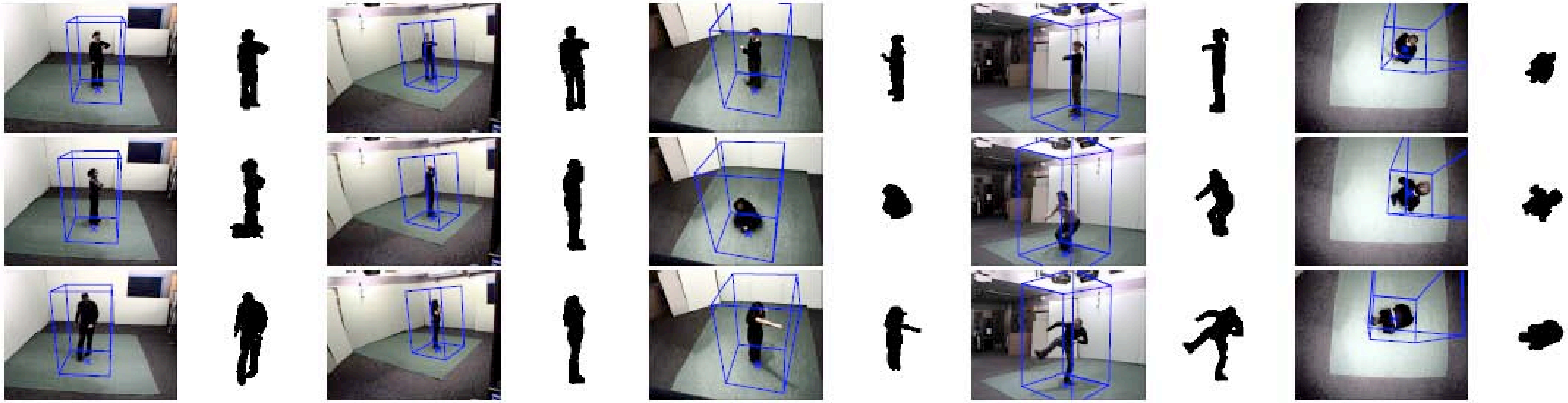


Figure 4. Camera setup and extracted silhouettes: (Top) the action "watch clock" from the 5 different camera views. (Middle and bottom) sample actions: "cross arms", "scratch head", "sit down", "get up", "turn", "walk", "wave", "punch", "kick", and "pick up". Volumetric templates are mapped onto the estimated interest region indicated by blue box.

Learning in 3D

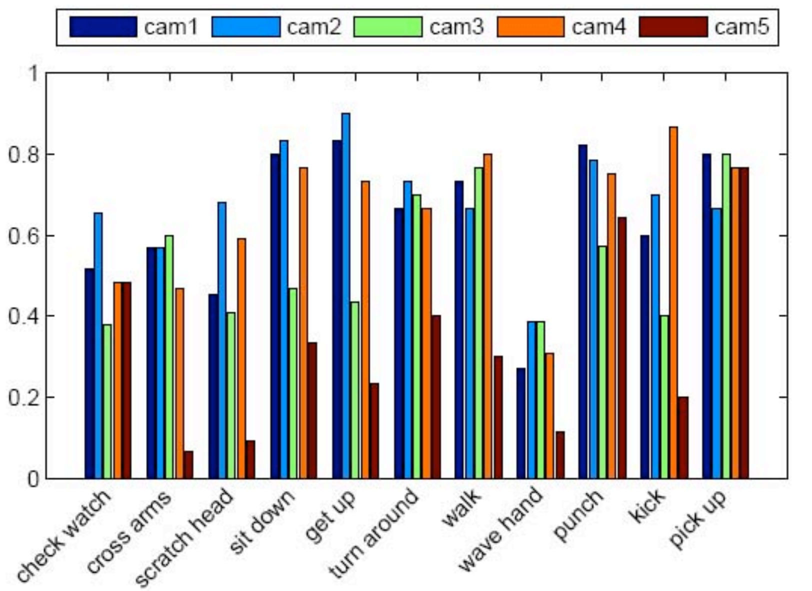
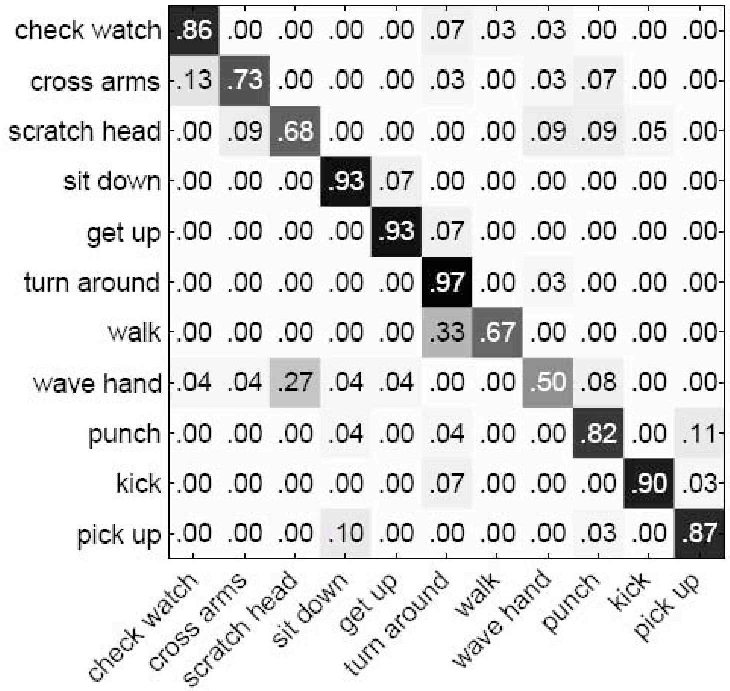
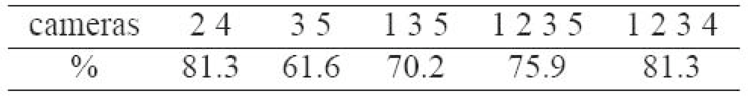


Figure 5. Recognition rates when learning in 3D and recognizing in 2D. The average rates per camera are {65.4, 70.0, 54.3, 66.0, 33.6}.



	check watch	cross arms	scratch head	sit down	get up	turn around	walk	wave hand	punch	kick	pick up
check watch	.86	.00	.00	.00	.00	.07	.03	.03	.00	.00	.00
cross arms	.13	.73	.00	.00	.00	.03	.00	.03	.07	.00	.00
scratch head	.00	.09	.68	.00	.00	.00	.00	.09	.09	.05	.00
sit down	.00	.00	.00	.93	.07	.00	.00	.00	.00	.00	.00
get up	.00	.00	.00	.00	.93	.07	.00	.00	.00	.00	.00
turn around	.00	.00	.00	.00	.00	.97	.00	.03	.00	.00	.00
walk	.00	.00	.00	.00	.00	.33	.67	.00	.00	.00	.00
wave hand	.04	.04	.27	.04	.04	.00	.00	.50	.08	.00	.00
punch	.00	.00	.00	.04	.00	.04	.00	.00	.82	.00	.11
kick	.00	.00	.00	.00	.00	.07	.00	.00	.00	.90	.03
pick up	.00	.00	.00	.10	.00	.00	.00	.00	.03	.00	.87

Figure 6. Confusion matrix for recognition using cameras 2 and 4. Note that actions performed with the hand are confused, *e.g.* "wave" and "scratch head" as well as "walk" and "turn".

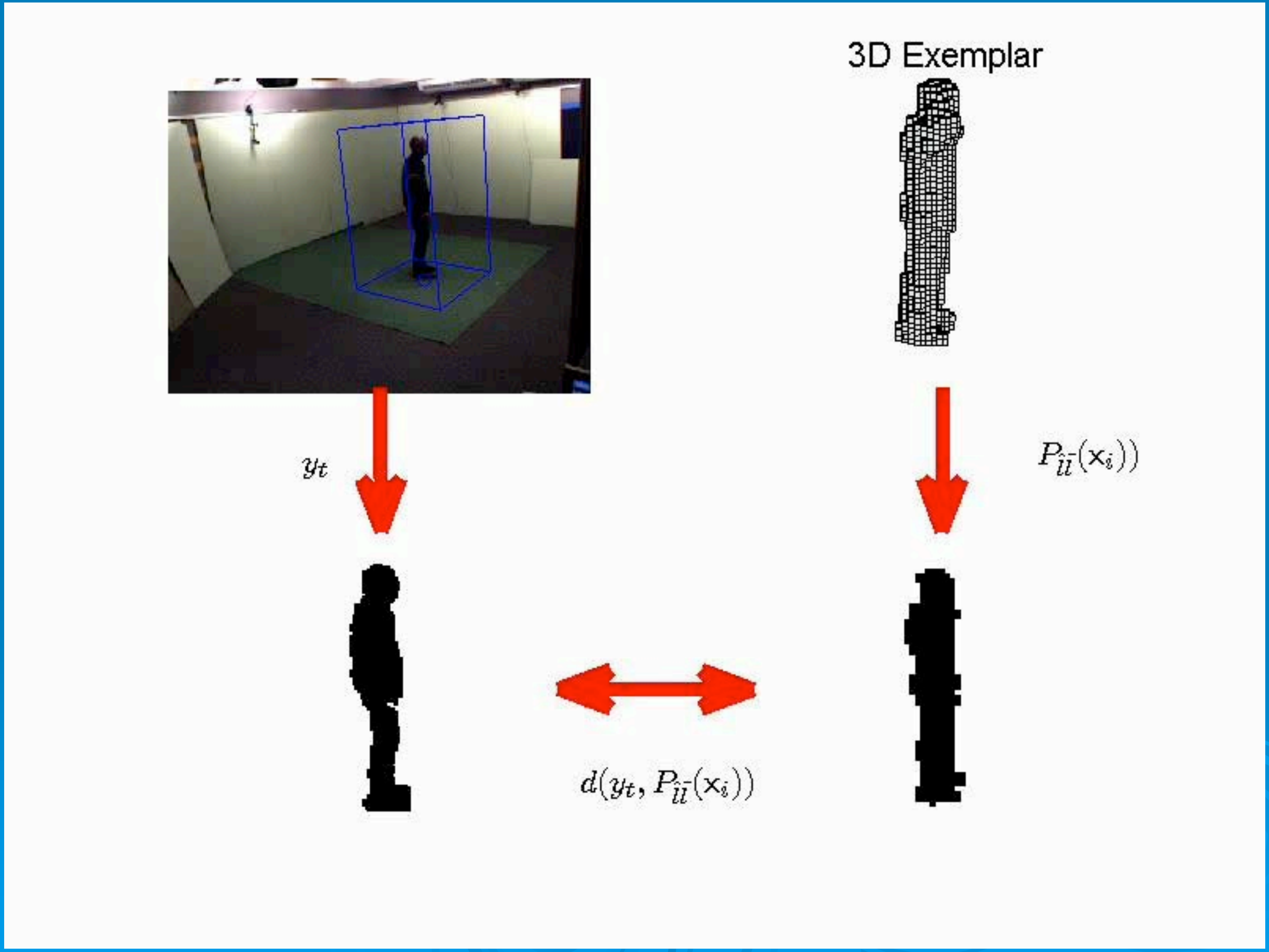


cameras	2 4	3 5	1 3 5	1 2 3 5	1 2 3 4
%	81.3	61.6	70.2	75.9	81.3

Table 1. Recognition rates with camera combinations. For comparisons, a full 3D recognition considering 3D manually aligned models as observations, instead of 2D silhouettes, yields 91.11%.

Rate with MHVs: 93.33%

Video


3D Exemplar
y_t
$P_{\hat{l}\bar{l}}(\mathsf{x}_i))$
$d(y_t, P_{\hat{l}\bar{l}}(\mathsf{x}_i))$

Conclusion

- MHVs:
  - Discriminative action descriptor for recognition and segmentation of motion streams.
  - Multiple calibrated cameras.
- Exemplar based model:
  - Probabilistic model of action and view transform.
  - Arbitrary number of cameras.
- Future Work:
  - Extend exemplar based model, e.g. other distance functions and template representations.
  - External cues, e.g. 3D scene information and objects.

Thank You!

Results

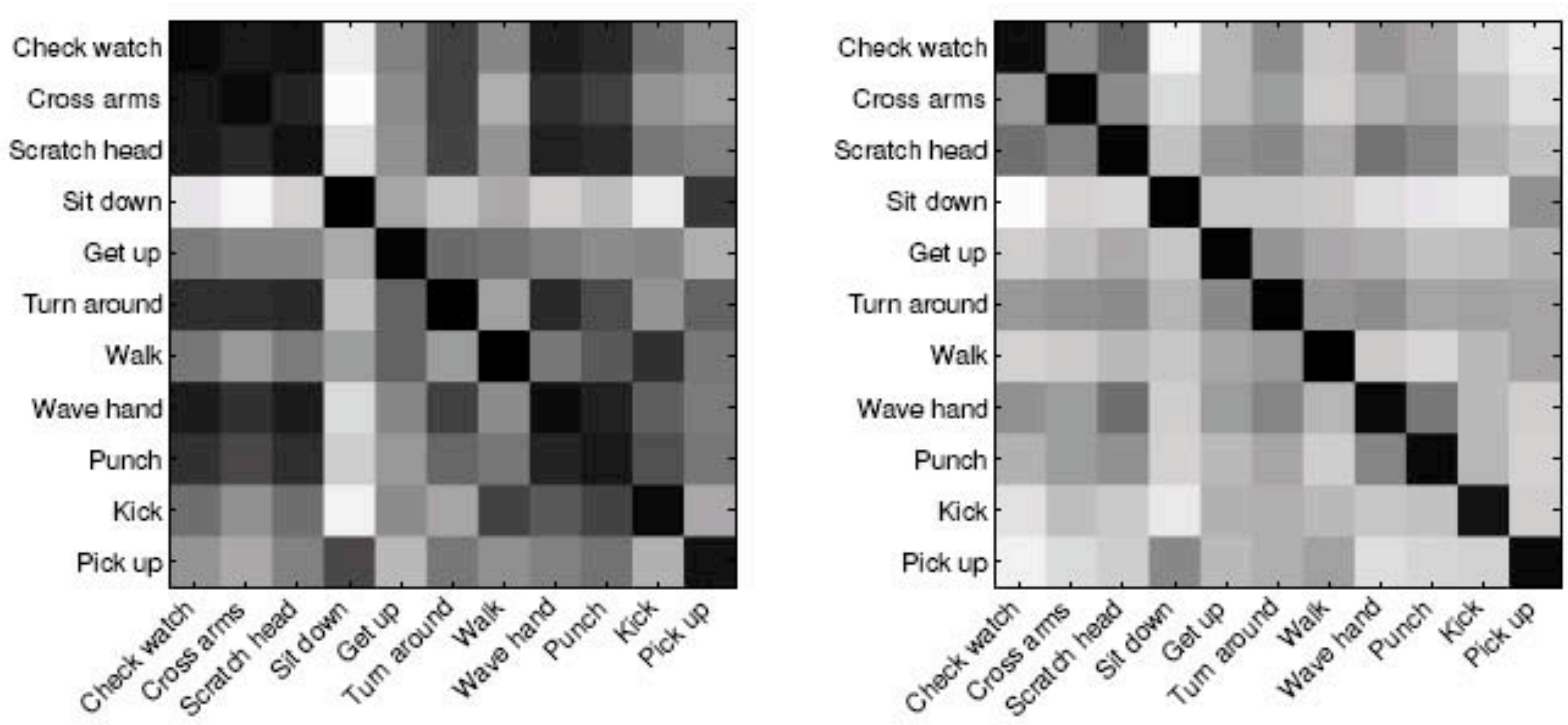


Fig. 8. Average class distance: (Left) before discriminant analysis. (Right) after discriminant analysis.