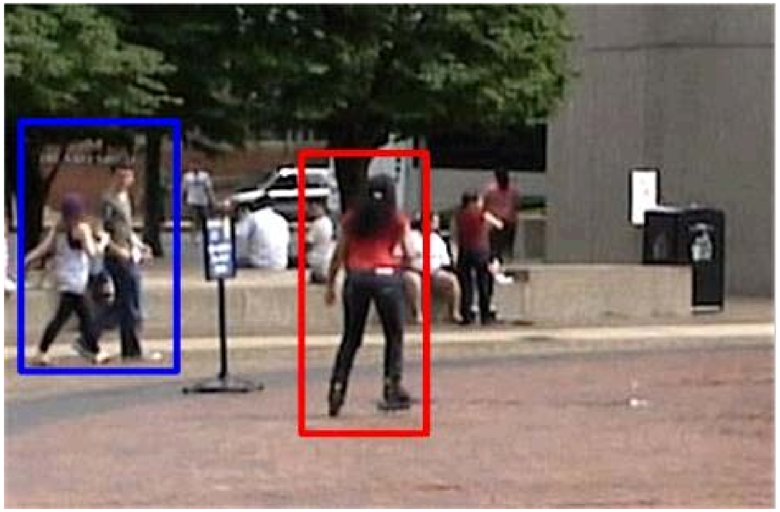
Action Recognition in Video by Sparse Representation on Covariance Manifolds of Silhouette Tunnels

Kai Guo, Prakash Ishwar, and Janusz Konrad

Department of Electrical & Computer Engineering

BOSTON UNIVERSITY

Motivation



- Recognize actions in video
  - walk
  - skate
- Applications

Surveillance: 

Sports & entertainment: 

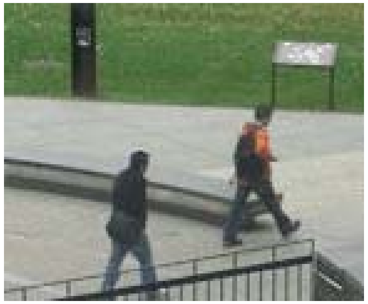
Tools for hearing impaired: 

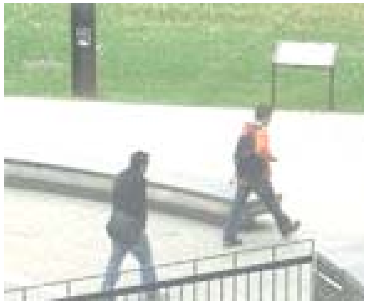
Wildlife habitat monitoring: 

Challenges and assumptions

Scene

- Multiple objects
- Clutter and occlusion
- Illumination variability





Challenges and assumptions

Scene

- ~~Multiple objects~~
- ~~Clutter and occlusion~~
- ~~Illumination variability~~

Scene

- Single object
- No significant clutter and occlusion
- No significant illumination change

Challenges and assumptions

Scene

- Multip... ic...
- Clutter a... sion
- Illu... tion varia...

Acquisition

- Camera motion
- Camera viewpoint and zoom
- Camera imperfections

Scene

- Single object
- No significant clutter and occlusion
- No significant illumination change

Challenges and assumptions

~~**Scene**~~

- ~~Multip[illegible]ic[illegible]~~
- ~~Clutter a[illegible]sion~~
- ~~Illu[illegible]tion varia[illegible]~~

~~**Acquisiti[illegible]**~~

- ~~Camera [illegible]~~
- ~~Camer[illegible]t and zoom~~
- ~~Ca[illegible]a imperfect[illegible]~~

Scene

- Single object
- No significant clutter and occlusion
- No significant illumination change

Acquisition

- Single camera, fixed viewpoint
- No significant camera motion and distortion

Challenges and assumptions

~~Scene~~

- ~~Multip ... ic~~
- ~~Clutter a ... sion~~
- ~~Illu ... tion varia ... y~~

~~Acquisiti~~

- ~~Camer ... i~~
- ~~Camer ... t and zoom~~
- ~~Ca ... a imperfect~~

Action

- Non-rigid objects
- Complex motion (ballet video)
- Intra and inter object motion variability

Scene

- Single object
- No significant clutter and occlusion
- No significant illumination change

Acquisition

- Single camera, fixed viewpoint
- No significant camera motion and distortion

Challenges and assumptions

Scene
- Multip ic
- Clutter a sion
- Illu tion varia

Scene
- Single object
- No significant clutter and occlusion
- No significant illumination change

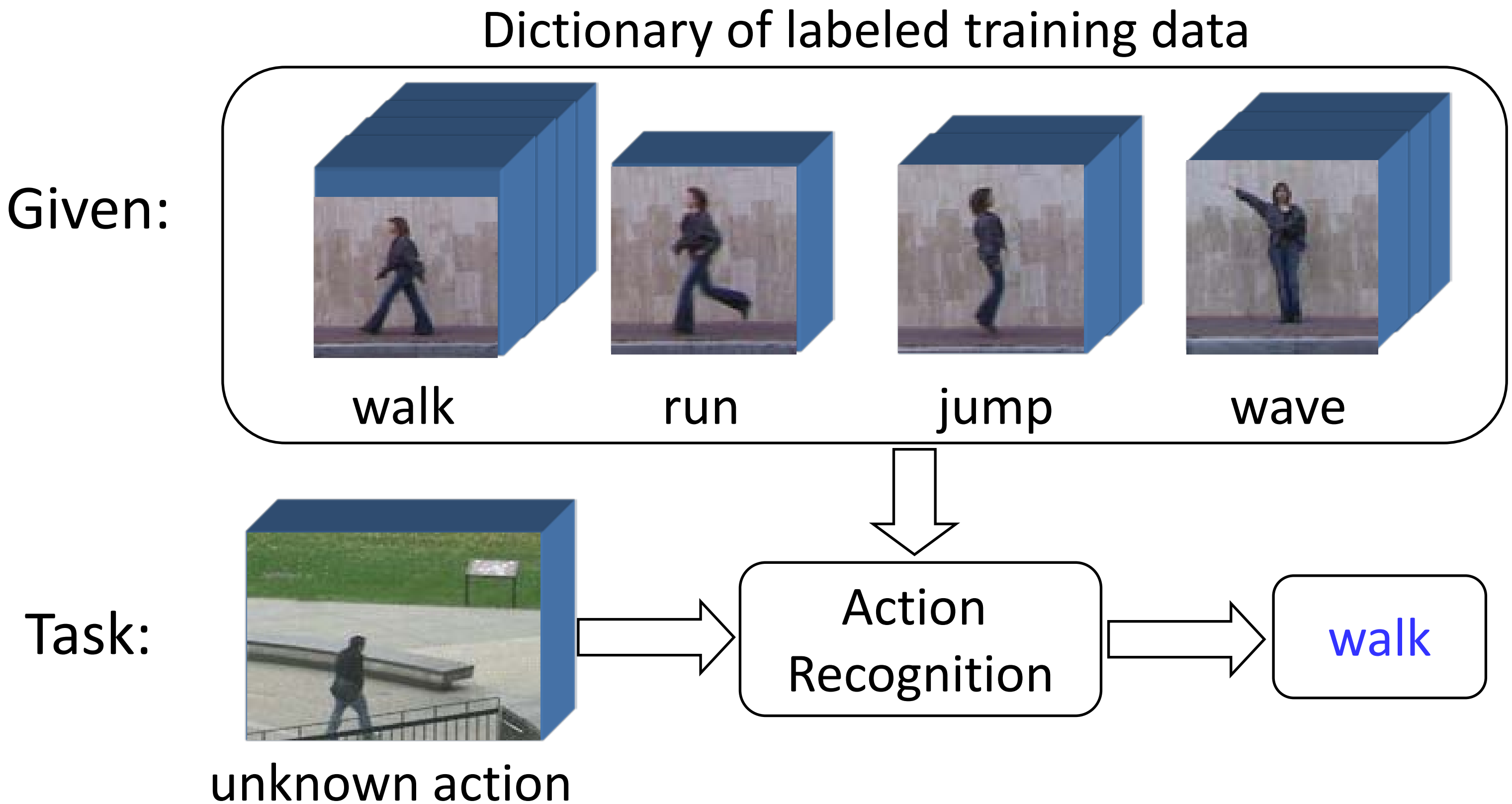
Acquisiti
- Camera
- Camera t and zoom
- Ca a imperfect

Acquisition
- Single camera, fixed viewpoint
- No significant camera motion and distortion

Action
- No rigid objects
- Co ex motio et video)
- Intra l i object motion variabi

Action
- Non-rigid objects
- Complex motion
- Intra and inter object motion variability

Problem statement



Recall: challenges

- non-rigid object
- complex motion
- intra and inter object motion variability

Related work

Action features →

Action classifiers ↓

	Shape-based features	Interest -point based features	Geometric human body features	Motion-based features
Nearest Neighbor	[Gorelick-Irani-PAMI'07] [Bobick-Davis-PAMI'01] [Collins-Gross-ICAFGR'02]	[Dollar-Rabaud-VS PETS'05]	[Cunado-Nixon-CVIU'03] [Wang-Ning-ICSVT'04]	[Seo-Milanfar-PAMI (submitted)], [Liu-Ali-CVPR'08] [Lowe-IJCV'04]
SVM	[Ikizler-Duygulu-LNCS'07] [Ahmad-Lee-Journal of Multimedia'10]	[Shuldt-Laptev-ICPR'04] [Laptev-CVPR'08]	[Goncalves-Bernardo-CVPR'95]	[Danafar-Gheissari-ACCV07], [Scovanner-Ali-ACM Multimedia'07]
Boosting	[Zhang-Liu-ICCV'09]	[Smith-Shah-ICCV'05]	-	[Alireza-Mori-CVPR'08], [Ke-Sukthankar-ICCV'05]
Graphical (Probabilistic) model	[Chen-Wu-ICDMW'08]	[Niebles-Lei-IJCV'08] [Wong-Cipolla-ICCV'07]	[Rohr-CVGIP'94]	[Ali-Shah-PAMI'10]

Action recognition framework

- Action recognition = Supervised learning problem,

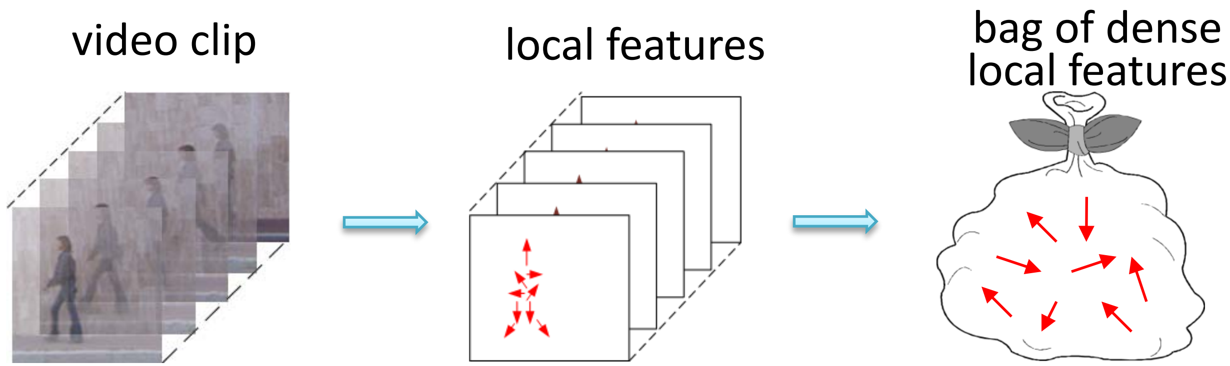
  where data samples are video clips

- Two main ingredients:

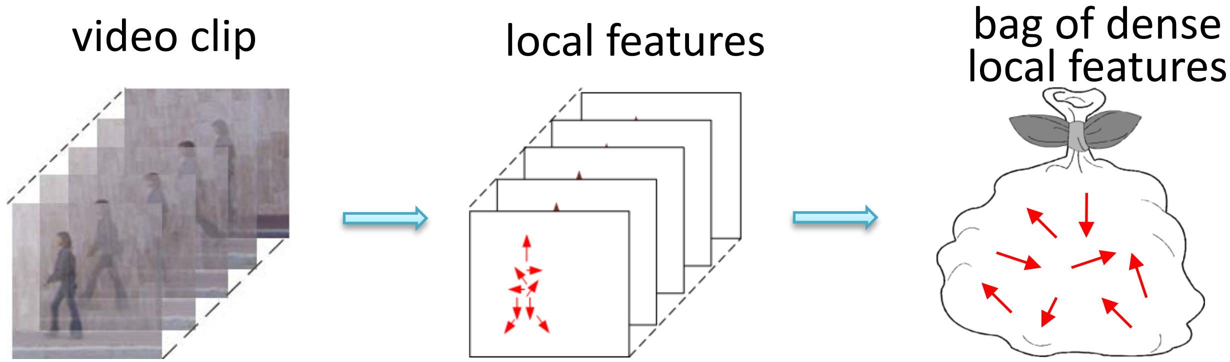
— **Representation** of samples

— **Classification**

Action representation

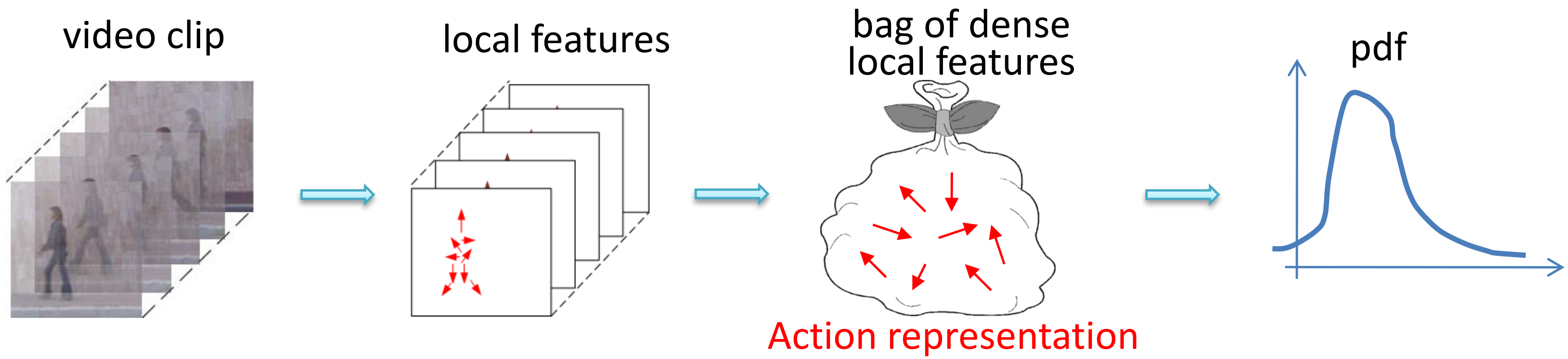


Action representation



- How to reduce the dimension of bag of local features?

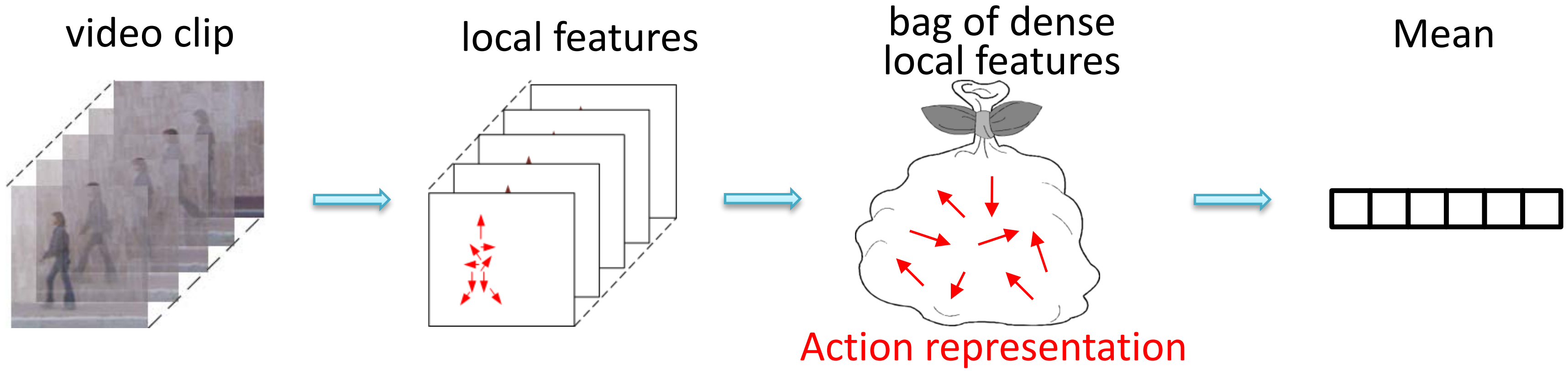
Action representation



- How to reduce the dimension of bag of local features?

  – Ideally, one should learn and compare pdfs of features
    – Problem: it may not reduce the dimensionality

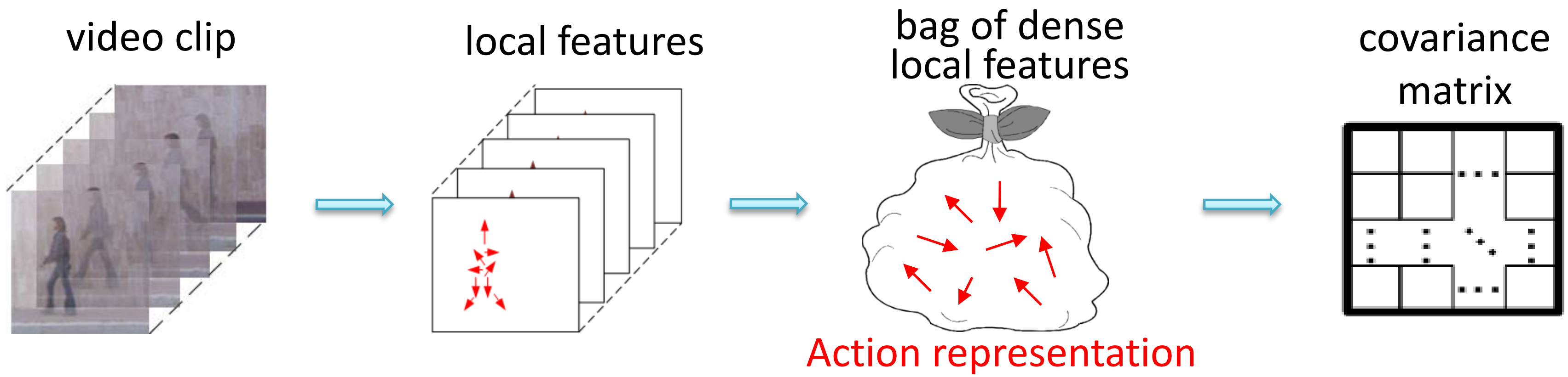
Action representation



- How to reduce the dimension of bag of local features?

— Idea-1: Learn and compare 1st order statistics (mean)
  — Problem: not sufficiently discriminative

Action representation

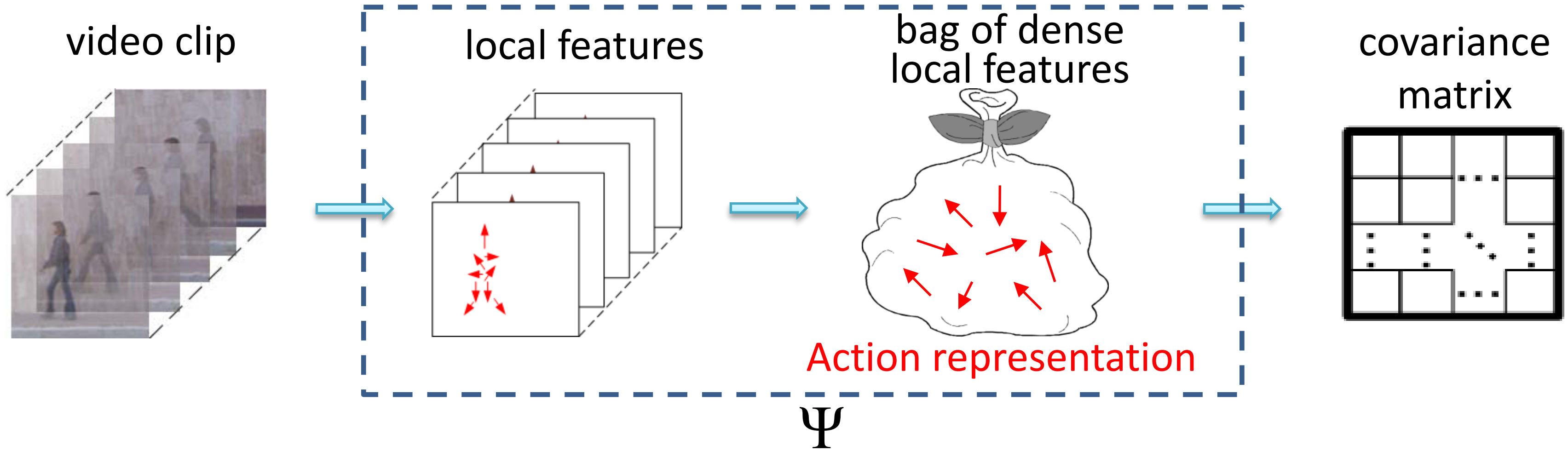


- How to reduce the dimension of bag of local features?

— Idea-2: Learn and compare 2nd order statistics (covariance)

[Tuzel-Porikli-Meer PAMI'08]

Action representation



- How to reduce the dimension of bag of local features?

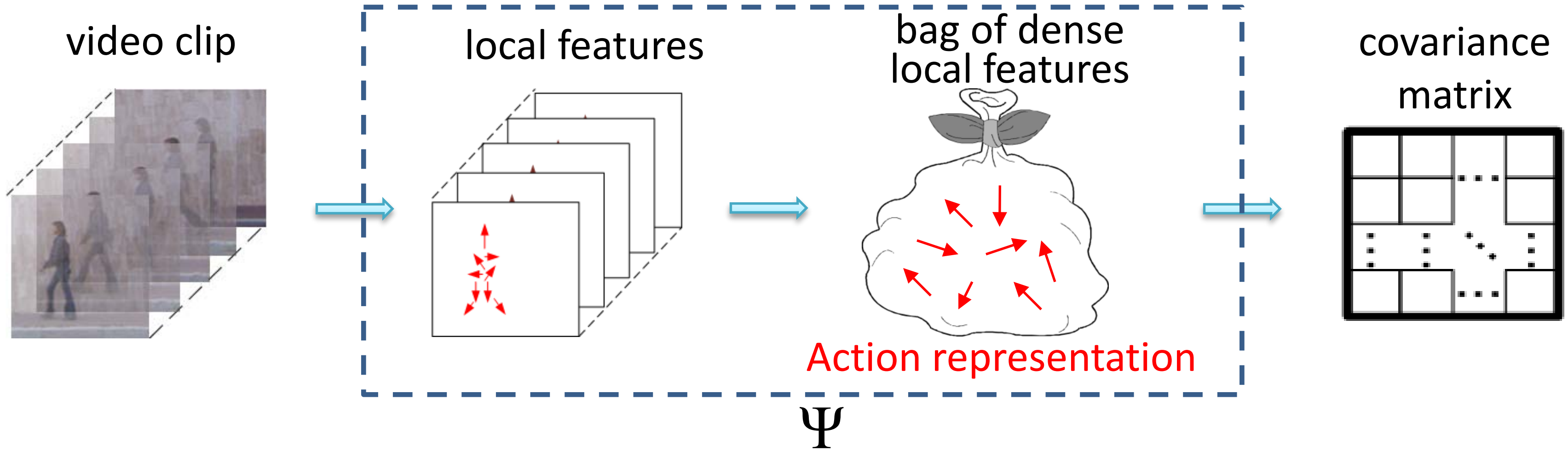
— Idea-2: Learn and compare 2nd order statistics (covariance)

[Tuzel-Porikli-Meer PAMI'08]

— Output: feature covariance matrix

(e.g., 13-dim vector -> 91-dim covariance matrix)

Action representation



- How to reduce the dimension of bag of local features?

— Idea-2: Learn and compare 2nd order statistics (covariance)

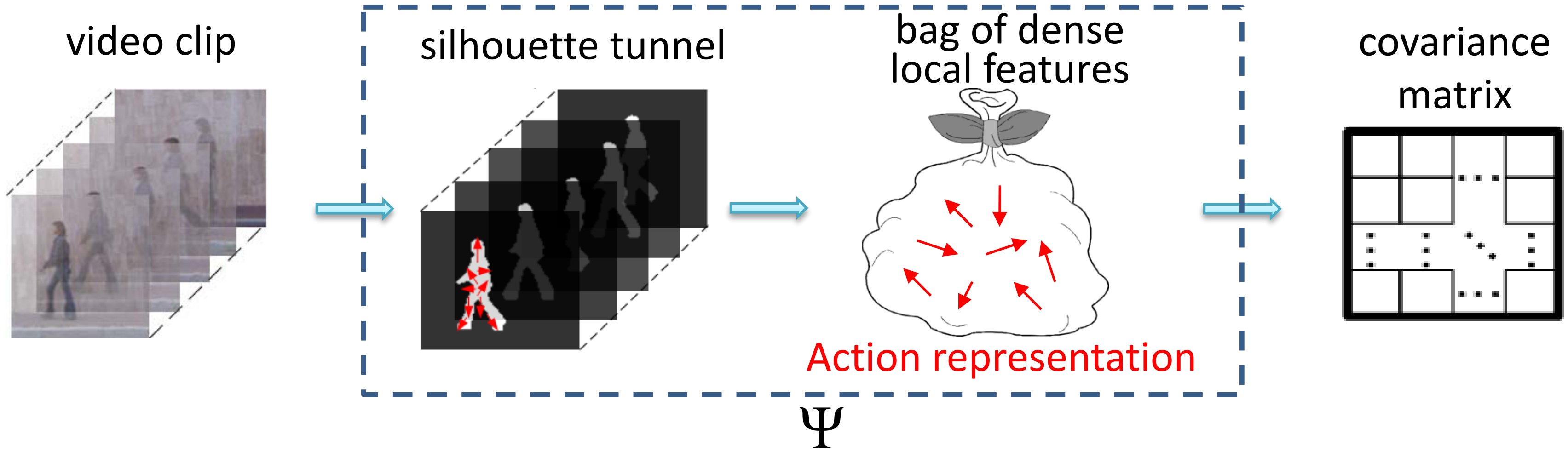
[Tuzel-Porikli-Meer PAMI'08]

— Output: feature covariance matrix

(e.g., 13-dim vector -> 91-dim covariance matrix)

Main thesis: covariance matrix is "sufficient" for action recognition

Action representation



- How to reduce the dimension of bag of local features?

— Idea-2: Learn and compare 2nd order statistics (covariance)

[Tuzel-Porikli-Meer PAMI'08]

— Output: feature covariance matrix

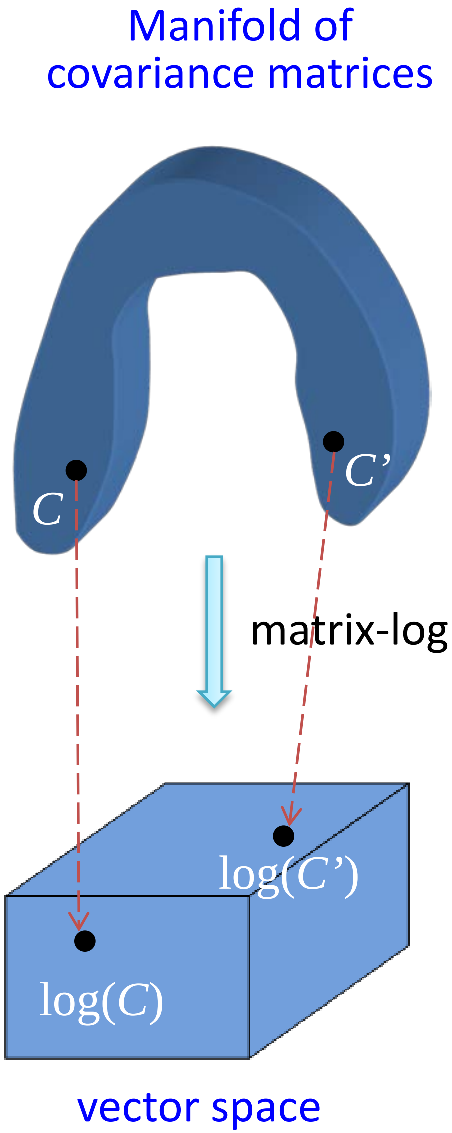
(e.g., 13-dim vector -> 91-dim covariance matrix)

Main thesis: covariance matrix is "sufficient" for action recognition

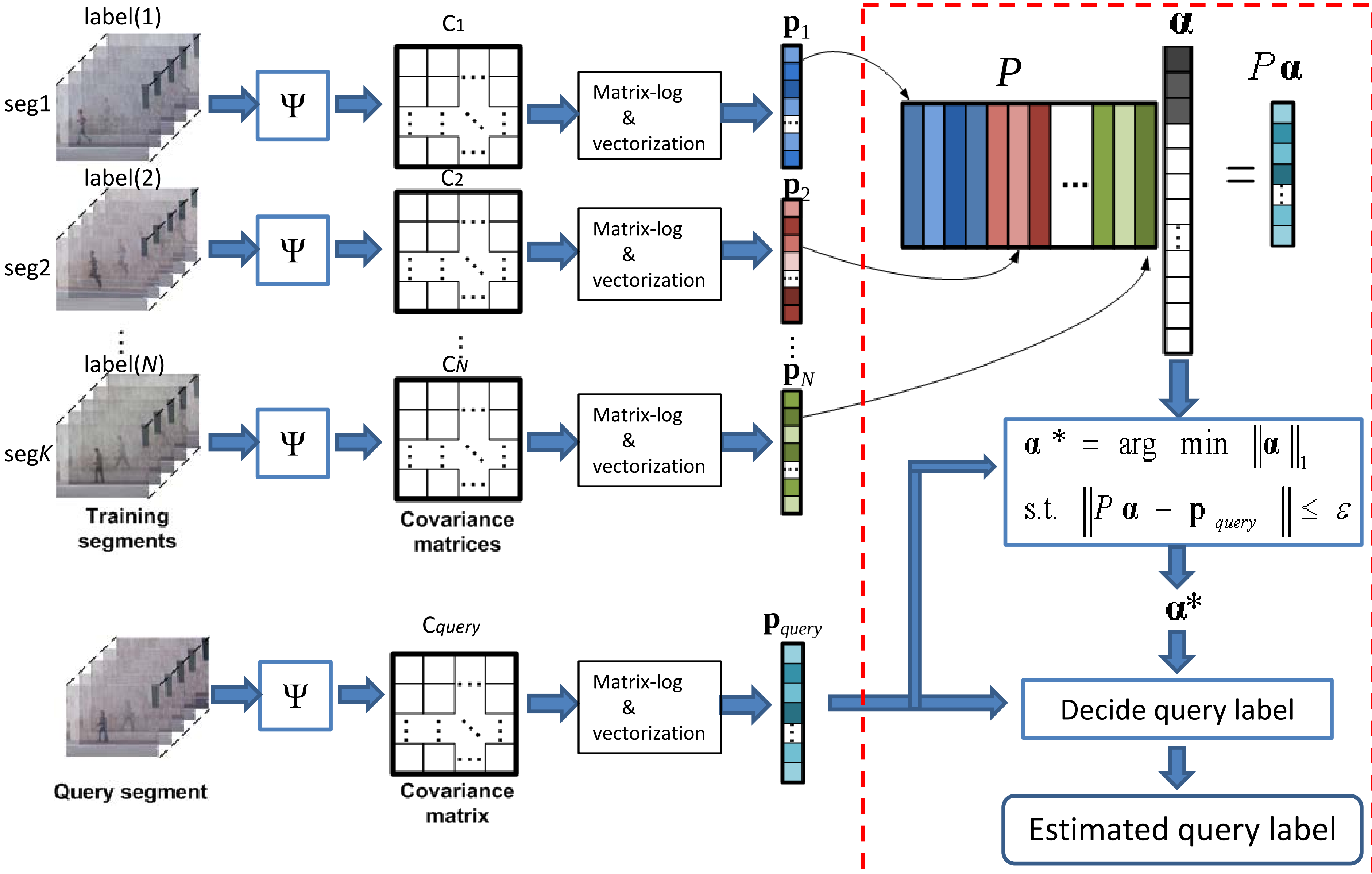
Covariance manifold

- Covariance matrices form:
— a Riemannian manifold
— not a vector space

- Matrix-log maps a Riemannian manifold to a vector space [Arsigny-Pennec-Ayache'06]
— $C = UDU^T$
— $\log(C) := U \log(D) U^T$



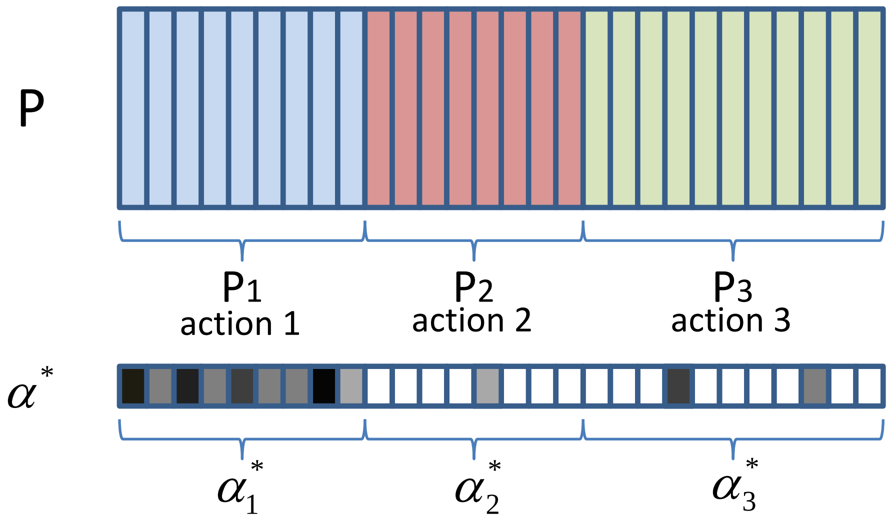
Classification based on sparse linear representation



[Wright et al., PAMI'09]

Classification algorithm

Each coefficient of α^* weights the contribution of training segments to query segment



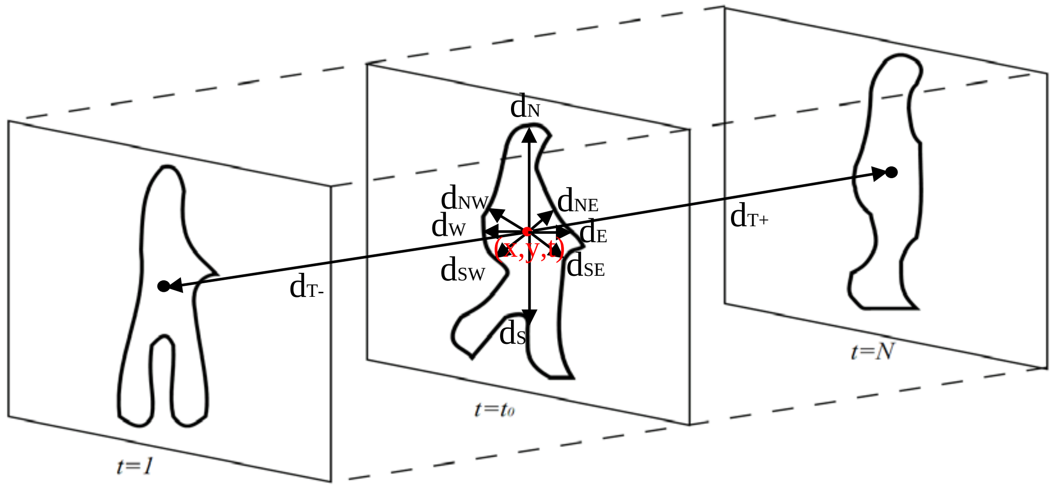
Step 1: Compute residual error:

$$R_i(\mathbf{p}_{query}) = \| \mathbf{p}_{query} - P_i \alpha_i^* \|_2$$

Step 2: Determine query label

$$label(\mathbf{p}_{query}) = \arg\min_i R_i(\mathbf{p}_{query})$$

Silhouette-based local features



$$\mathbf{s} = \begin{pmatrix} x \\ y \\ t \end{pmatrix} \Longrightarrow \mathbf{f(s)} = \begin{pmatrix} x \\ y \\ t \\ d_N \\ d_E \\ \cdots \\ d_{T+} \\ d_{T-} \end{pmatrix}_{13x1}$$

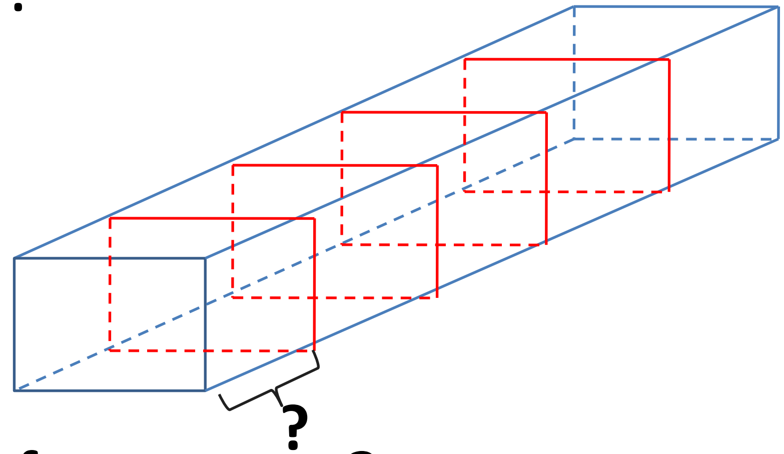
- Dimensionality reduction

$$C_S := \frac{1}{|S|} \sum_{\mathbf{s} \in S} (\mathbf{f}(\mathbf{s}) - \boldsymbol{\mu})(\mathbf{f}(\mathbf{s}) - \boldsymbol{\mu})^T$$

Implementation issues

- How to process a long video?

— Break it into segments

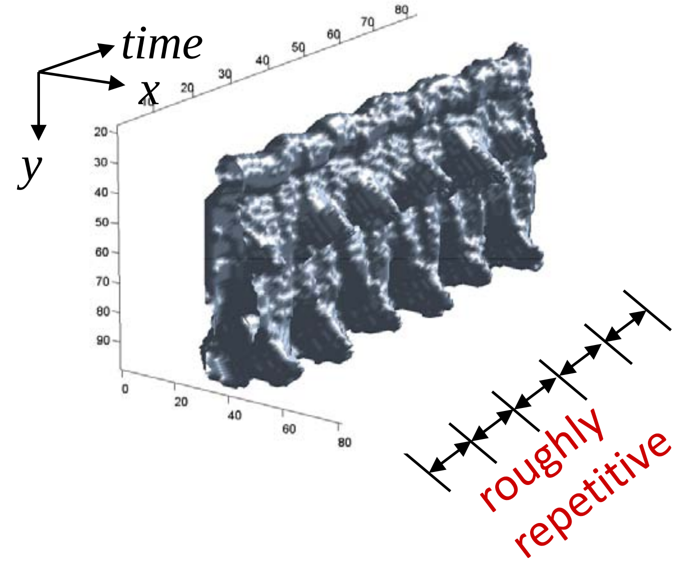


- What should be the length of segments?

—Period for human action ≈ 0.4 — 0.8 s

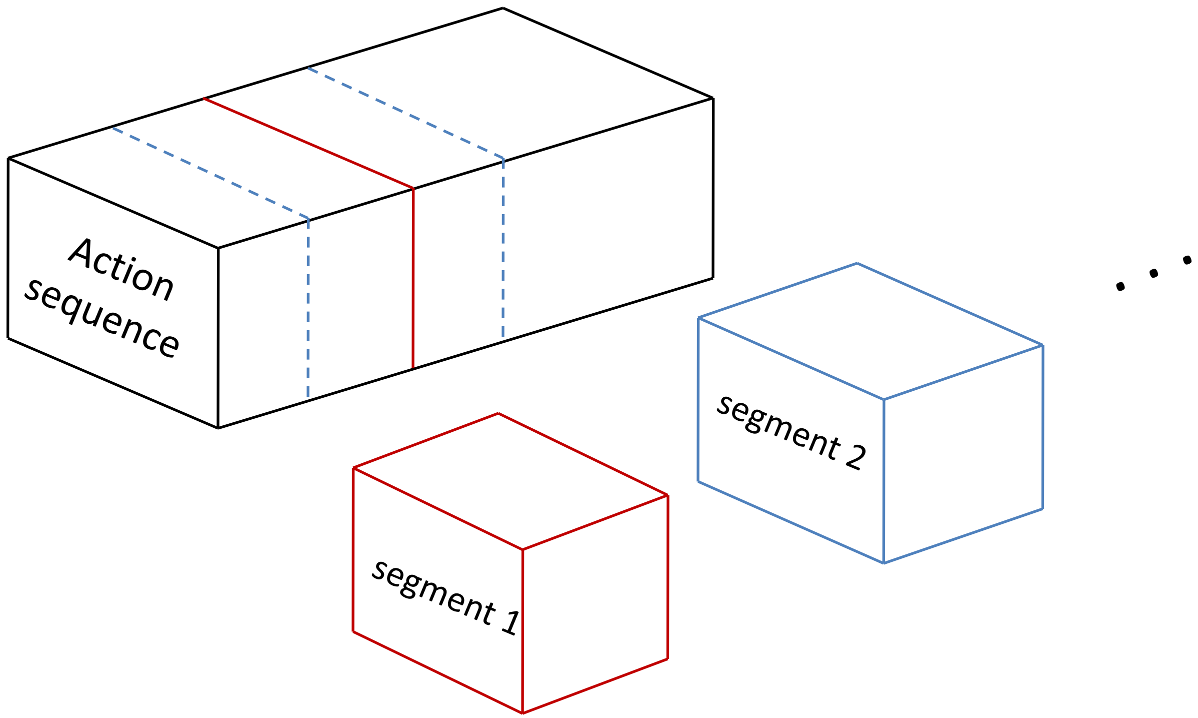
—Segment length ≈ 10 — 20 frames

(@25 fps video)



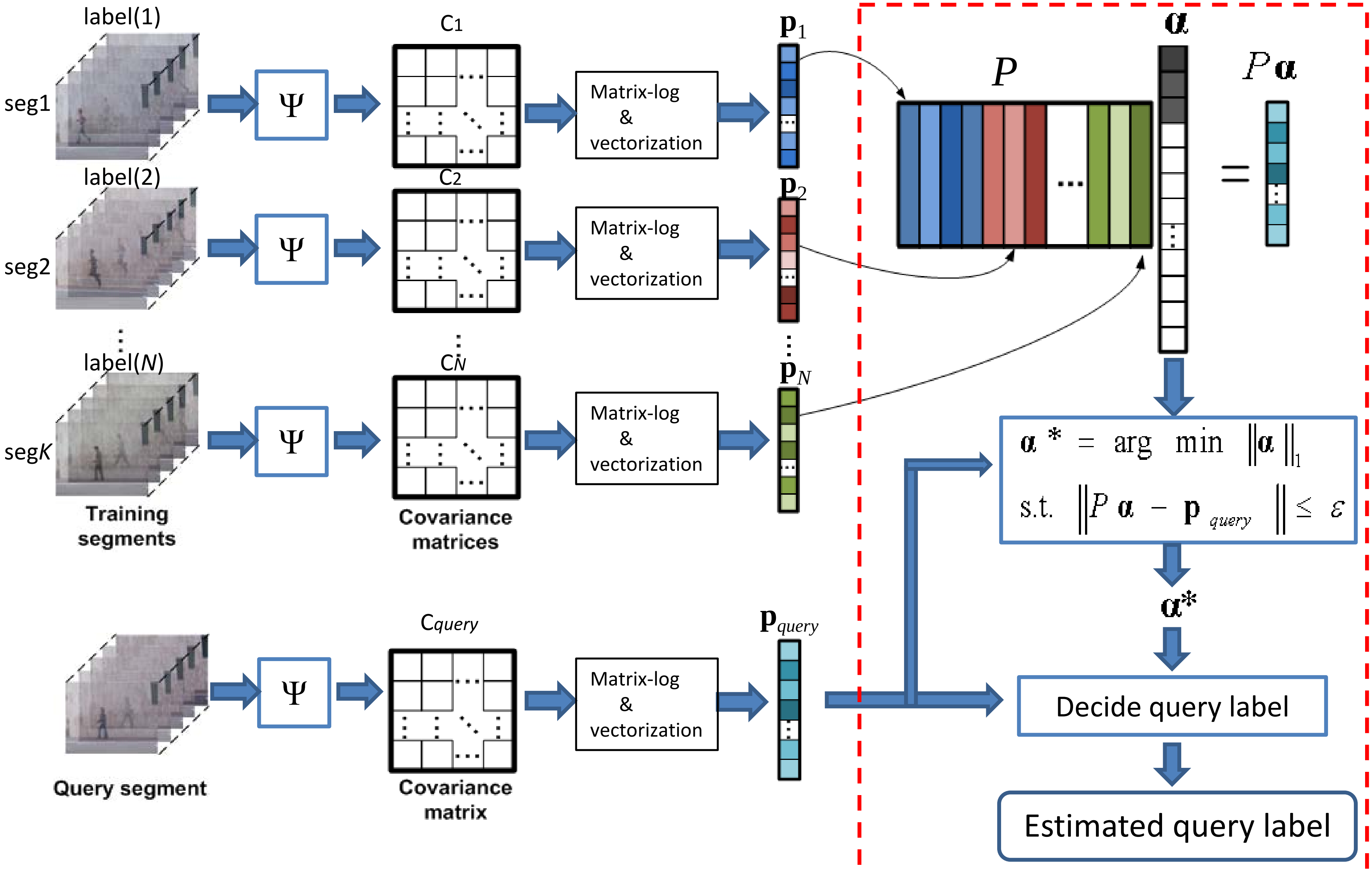
Action segments

- No knowledge about the beginning and end of periods
  - Use overlapping segments



- Additional benefits of overlapping segments
  - Reduced sensitivity to temporal action misalignment
  - Richer dictionary

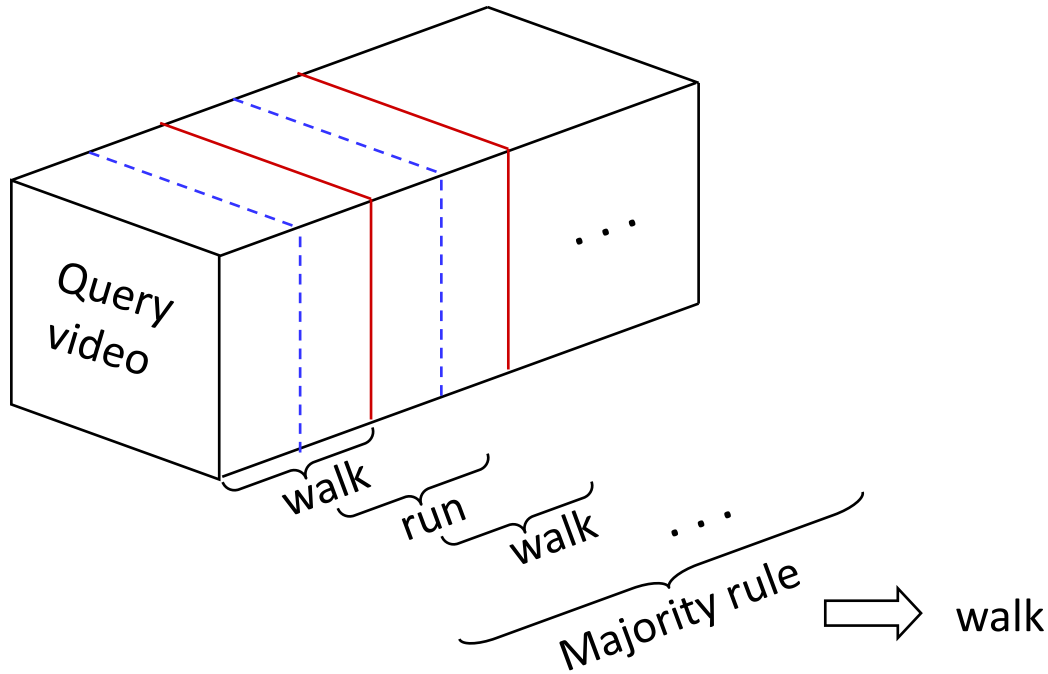
Segment-level classification



[Wright et al., PAMI'09]

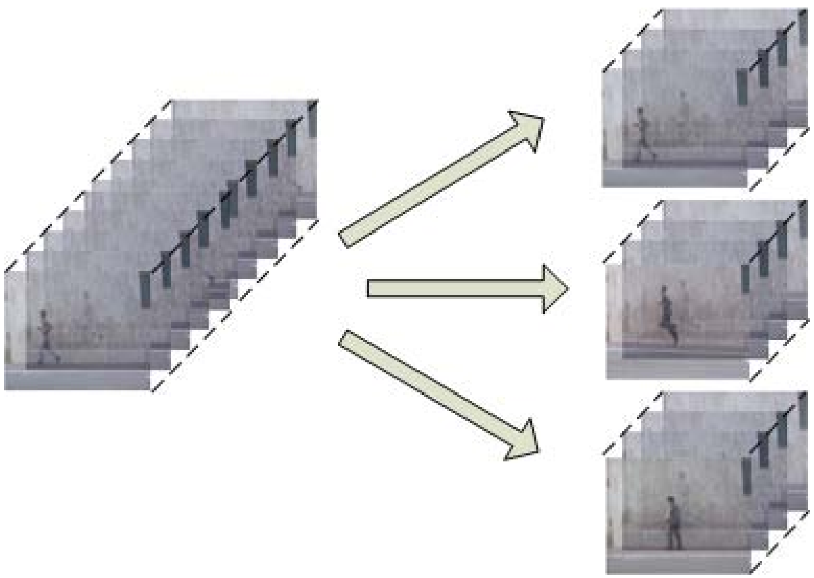
From segment decisions to a sequence decision

- How to get the labels of query video ?
  - Majority rule



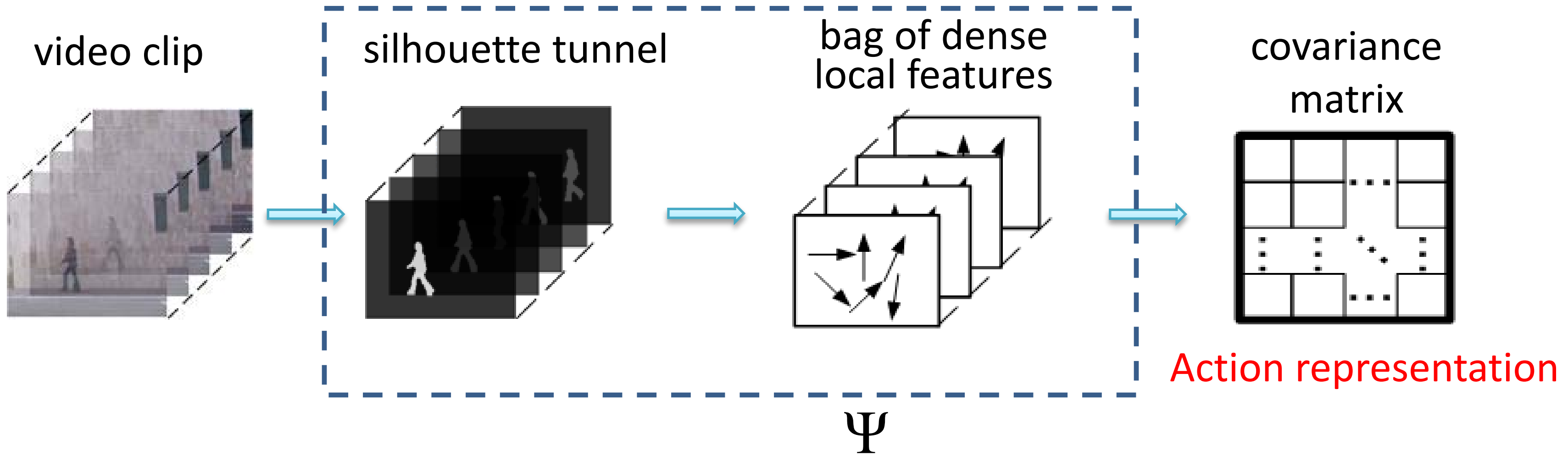
Summary of the approach

- Partitioning into segments



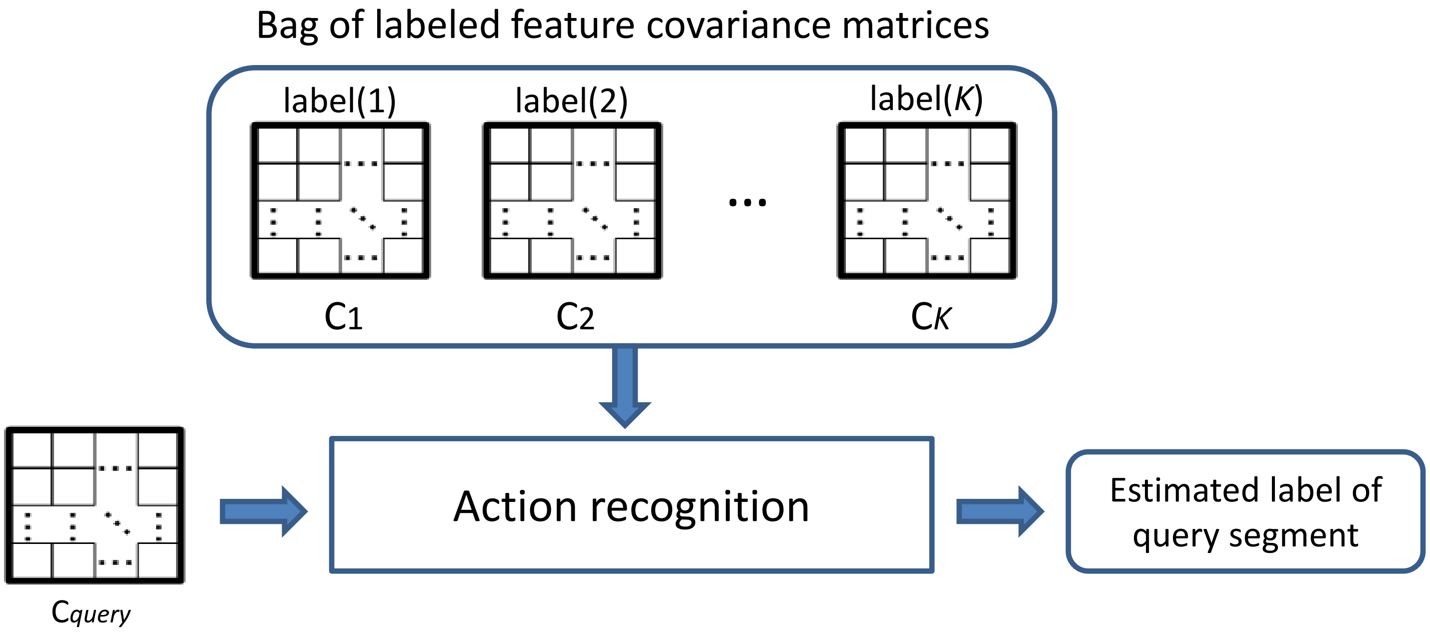
Summary of the approach

- Partitioning into segments
- Action representation for each segment



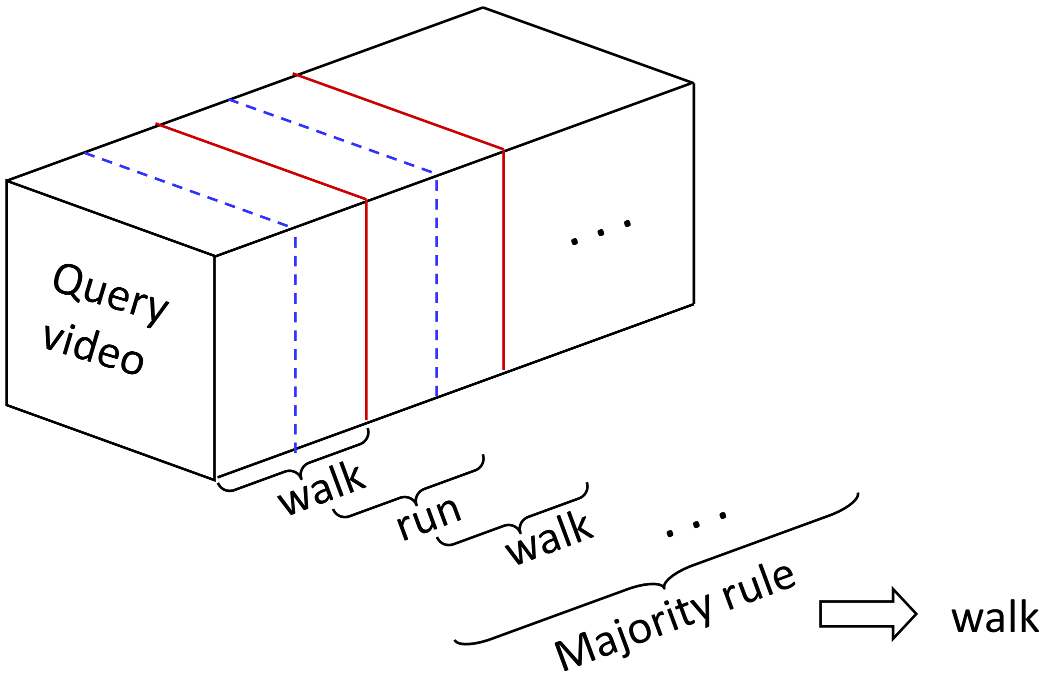
Summary of the approach

- Partitioning into segments
- Action representation for each segment
- Segment-wise action recognition



Summary of the approach

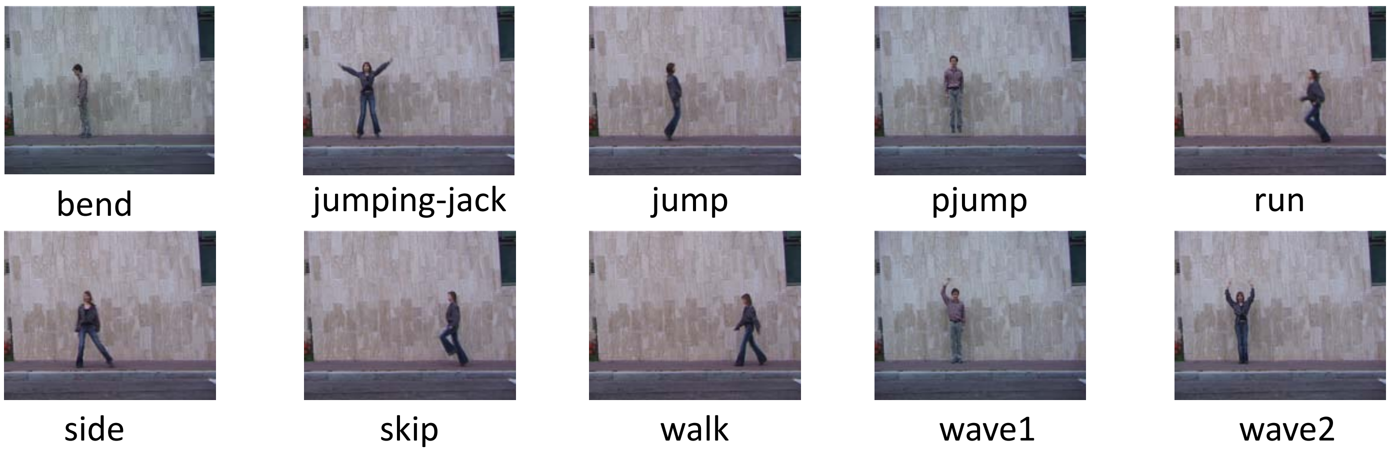
- Partitioning into segments
- Action representation for each segment
- Segment-wise action recognition
- Decision fusion



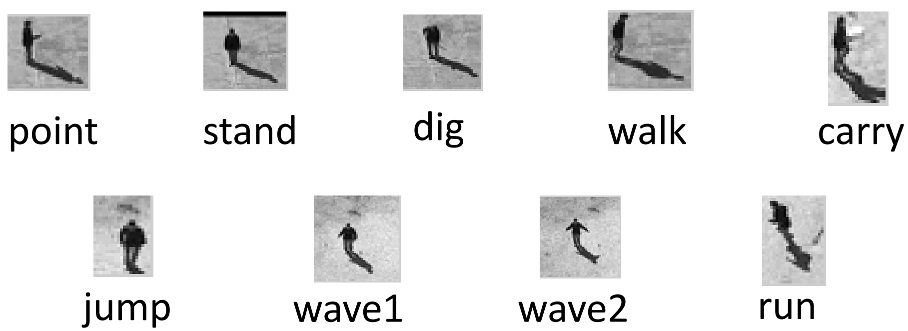
Experimental results

- **Datasets**

— Weizmann: 9 persons x 10 actions (180x144)



— UT-Tower: 6 persons x 9 actions x 2 times (about 90x70)



Performances

- Correct classification rate (CCR)
  - SEG-CCR: % of correctly classified query segments
  - SEQ-CCR: % of correctly classified query sequences

- Weizmann dataset: (LOOCV, N=8)

Method	Proposed	NN-based	Gorelick	Niebles	Ali	Seo
SEG-CCR	96.74%	97.05%	97.83%	—	95.75%	—
SEQ-CCR	100%	100%	—	90%	—	96%

- UT-Tower dataset: (LOOCV, N=8)

Method	Proposed	NN-based
SEG-CCR	96.15%	93.53%
SEQ-CCR	97.22%	96.30%

Computational complexity

- **Platform**: Dual Core 2.2 GHz + 2GB Memory + Matlab 7.6

- **Action representation**

  video: 180 x 144 x 84 — 0.12 sec/frame (8.3fps)

- **Action classification**

  0.07 sec/segment (14.3fps)

Conclusions

- We proposed a novel approach to action recognition:
  - —action representation = covariance matrix of local features
  - —action classification = sparse-representation-based classifier
- The proposed approach has
  - — state-of-the-art performance on Weizmann dataset
  - — 100% performance on non-static actions in low-resolution UT-Tower dataset
  - — low memory requirements with close to real-time performance