Human Motion Categorization & Detection

Juan Carlos Niebles[1,2]

Advisor : Prof. Fei-Fei Li[1]

1 Princeton University

2 UNIVERSIDAD DEL NORTE
BARRANQUILLA - COLOMBIA

Juan Carlos Niebles[1,2]

Princeton University

UNIVERSIDAD DEL NORTE
BARRANQUILLA - COLOMBIA

surveillance

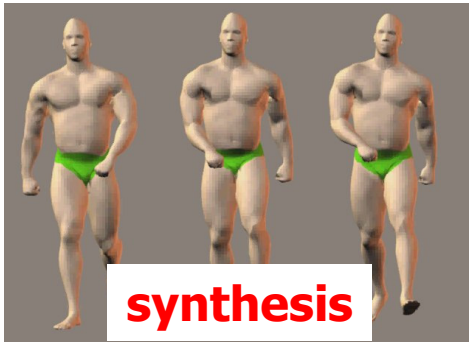
YouTube

YAHOO!

msn

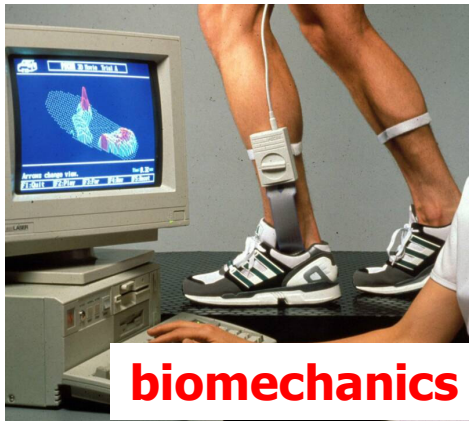
search

sport analysis

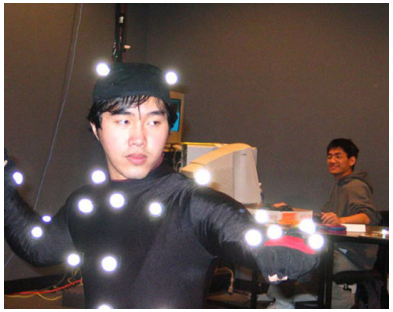


synthesis

Human Motion Analysis



biomechanics

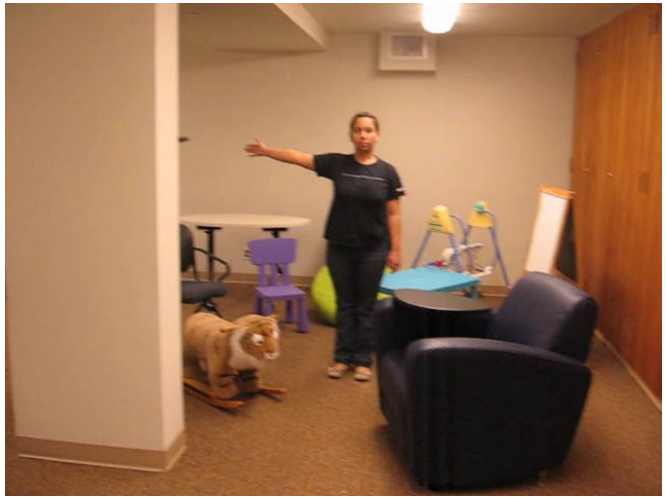


motion capture

game interfaces

sign language

In this talk



- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences

In this talk

- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences

Identification of human actions in video

Challenges:

- **Camera Motion**
- **Complex Background**
- **Viewpoint Change**

Localizing human actions in long sequences

Multiple actions by different people in one single sequence

Camera Motion

Dynamic Background

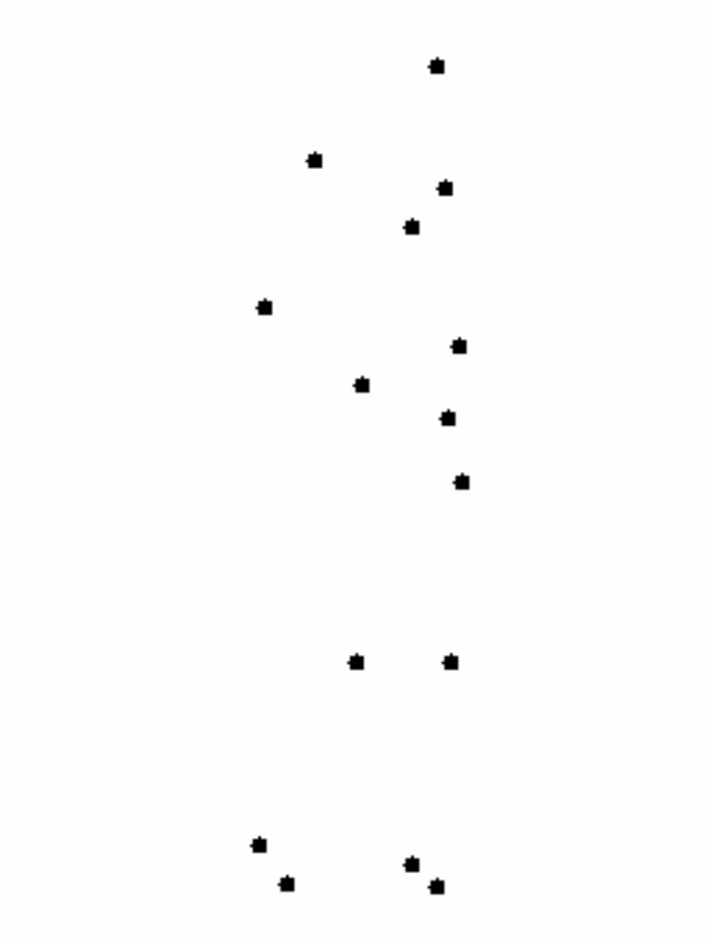
Previous Work

1. Recognition by template correlation/matching
    - Efros et al, ICCV '03
    - Shechtman et al, CVPR '05
    - Ramanan et al, NIPS '03

2. Action analysis using graphical models
    - Song et al, PAMI '03
    - Fanti et al, ICCV '05
    - Boiman et al, ICCV '05
    - Bergler, CVPR '97

3. Spatial-temporal Interesting Points
    - Laptev et al, ICCV '03; Schuldt et al, ICPR '04
    - Dollar et al, ICCV VS-PETS '05
    - Ke et al, ICCV '05

Motivation



Sparse and *Local* representation

Spatio-temporal information

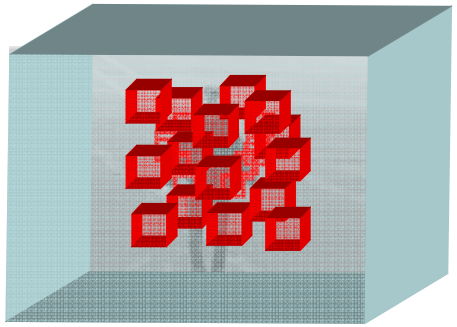
Johansson, 1973

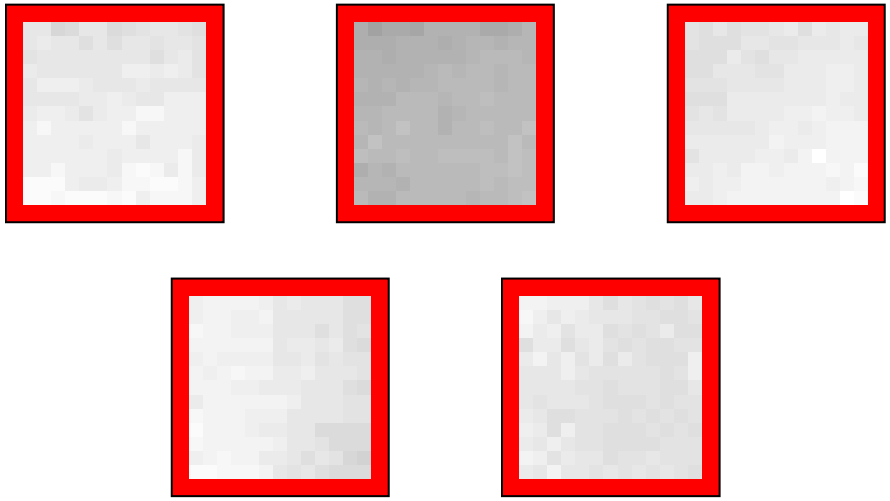
Spatio-Temporal Interest Points

Detection

$$\mathbf{R} = (\mathbf{I} * g_x * g_y * h_{ev})^2 + (\mathbf{I} * g_x * g_y * h_{od})^2$$

Local optimum of R define the position of features





[Dollar et al '05]

Spatial-Temporal Interest Points

- walking
- running
- jogging
- boxing
- handclapping
- handwaving

Spatial-Temporal Interest Points

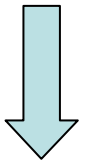
Detection

$$\mathbf{R} = (\mathbf{I} * g_x * g_y * h_{ev})^2 + (\mathbf{I} * g_x * g_y * h_{od})^2$$

Local optimum of R define the position of features

Description

Spatial-temporal cube

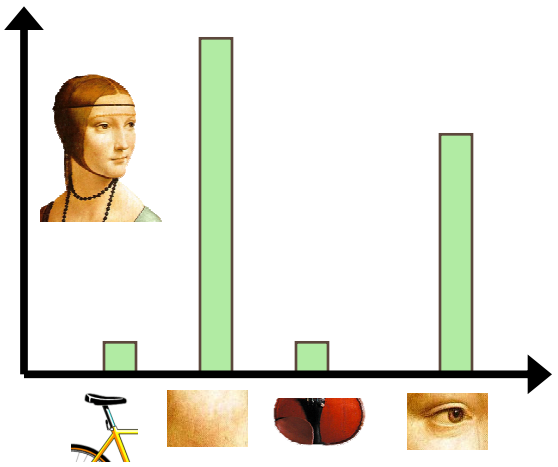


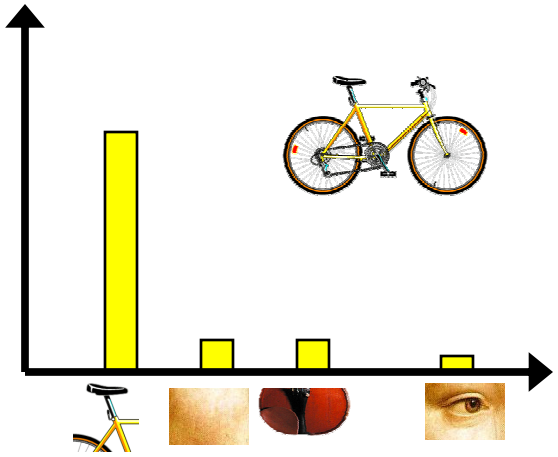
Concatenated brightness gradient

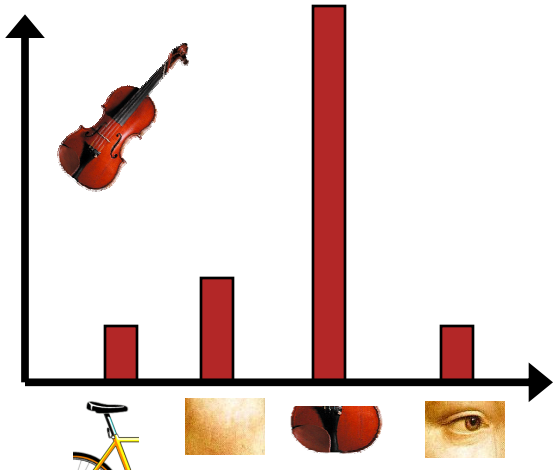
Experiments showed that optical flow descriptor is equally effective

[Dollar et al '05]

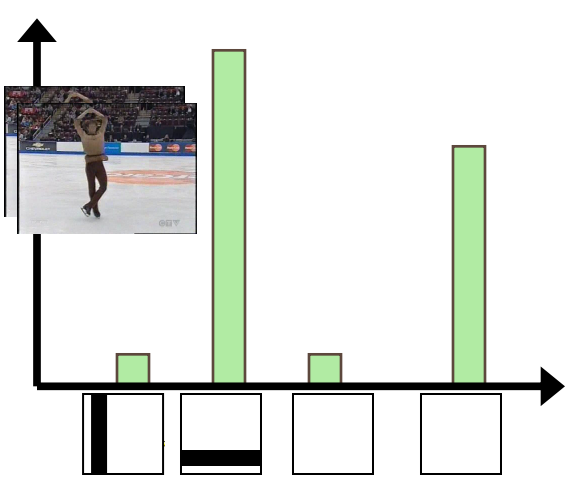
Object → Bag of 'words'

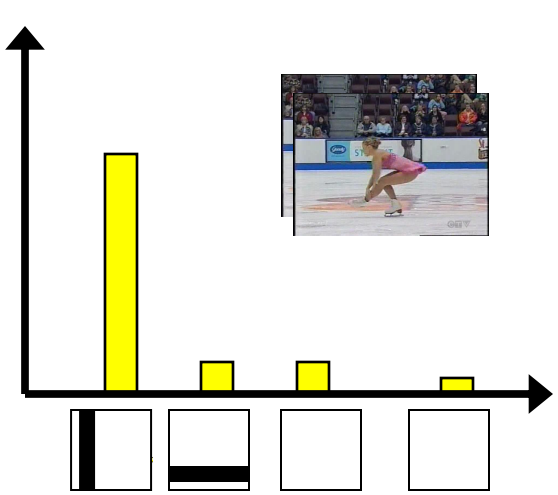


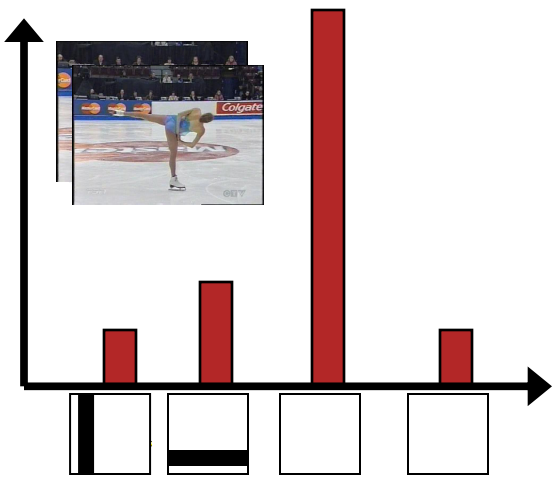


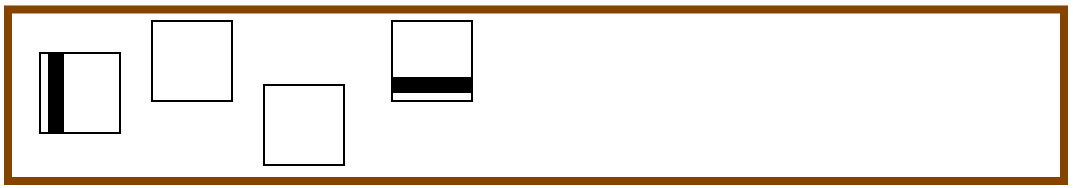


Action → Bag of motion ‘words’

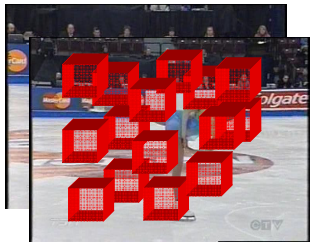




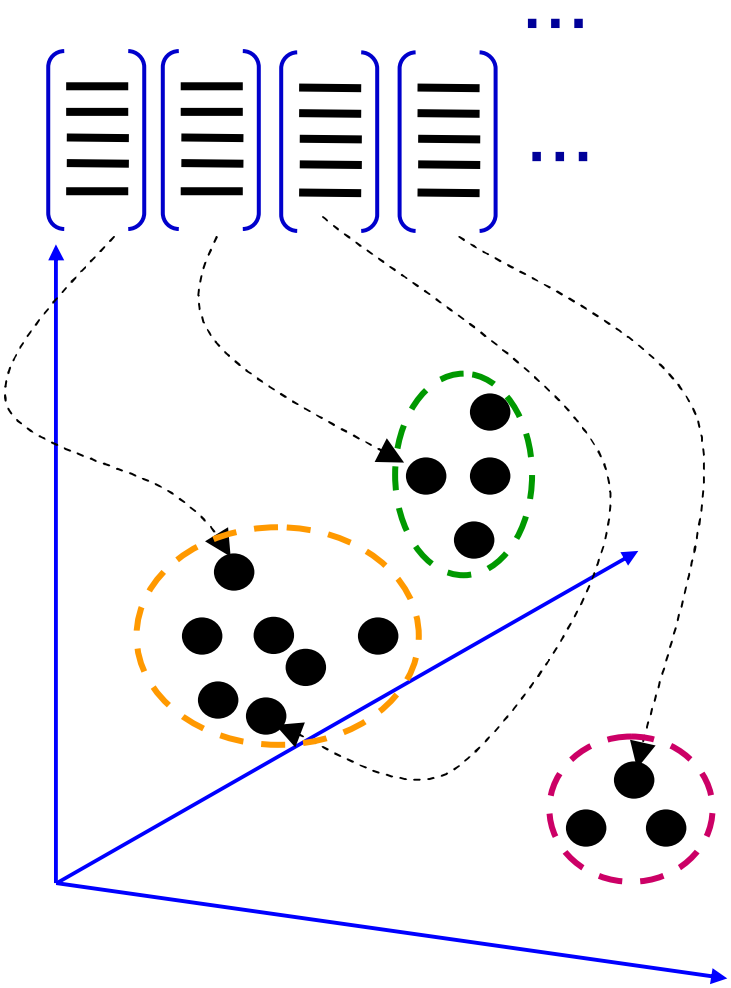




Codebook and Representation



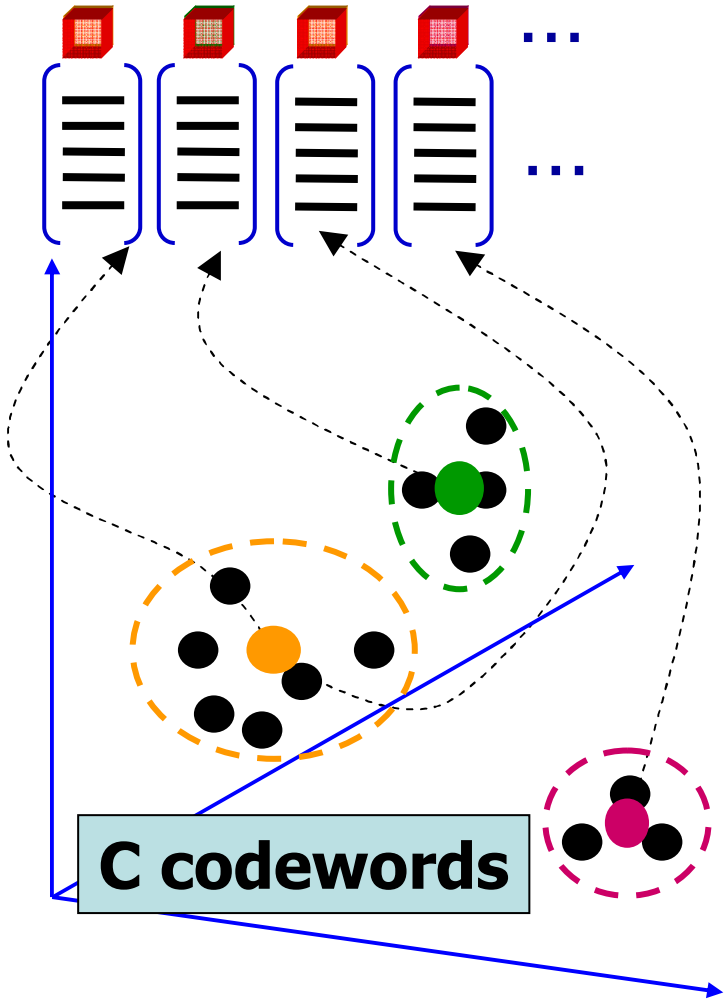
Training Videos

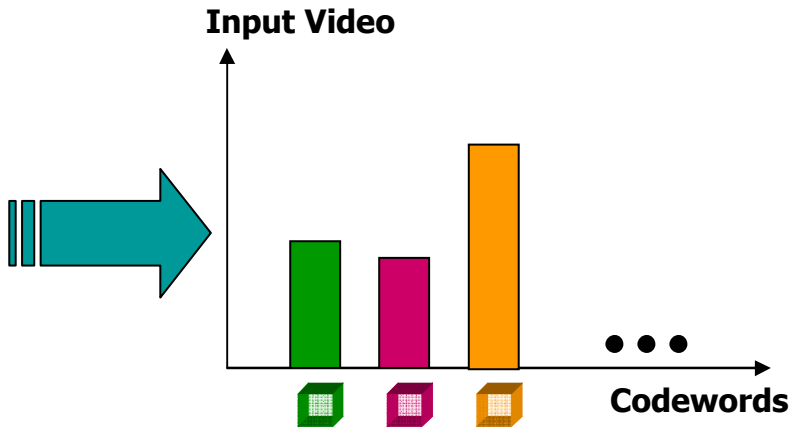


Codebook and Representation

Training Videos

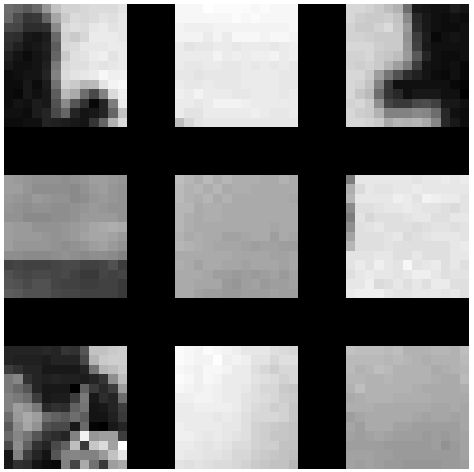
Feature Space



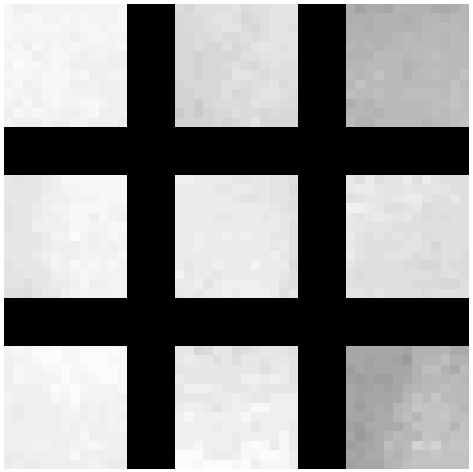


Experiment I: Codebook

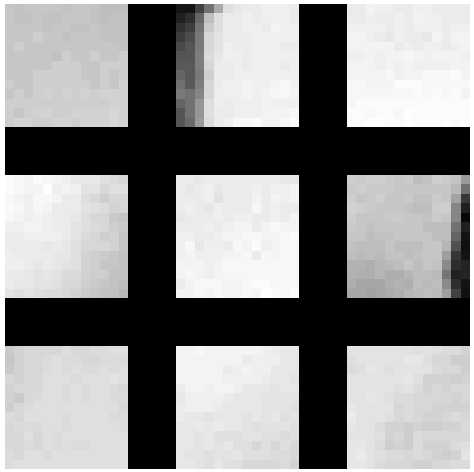
walking

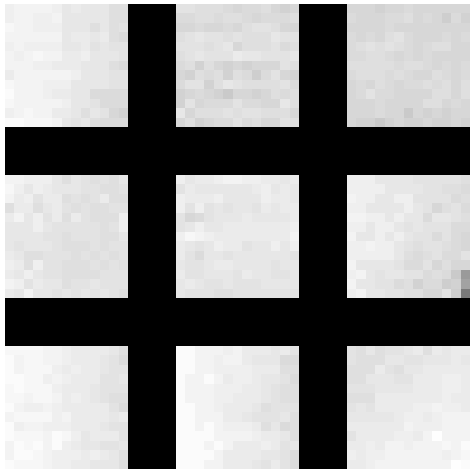


running



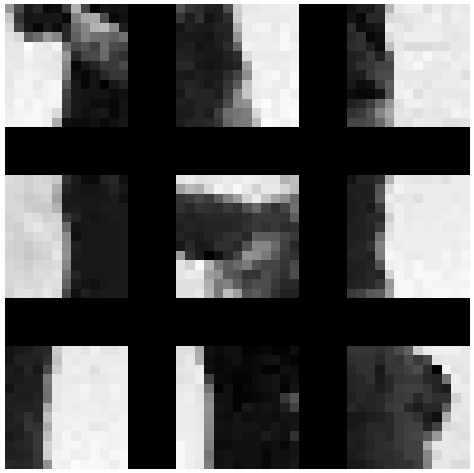
jogging



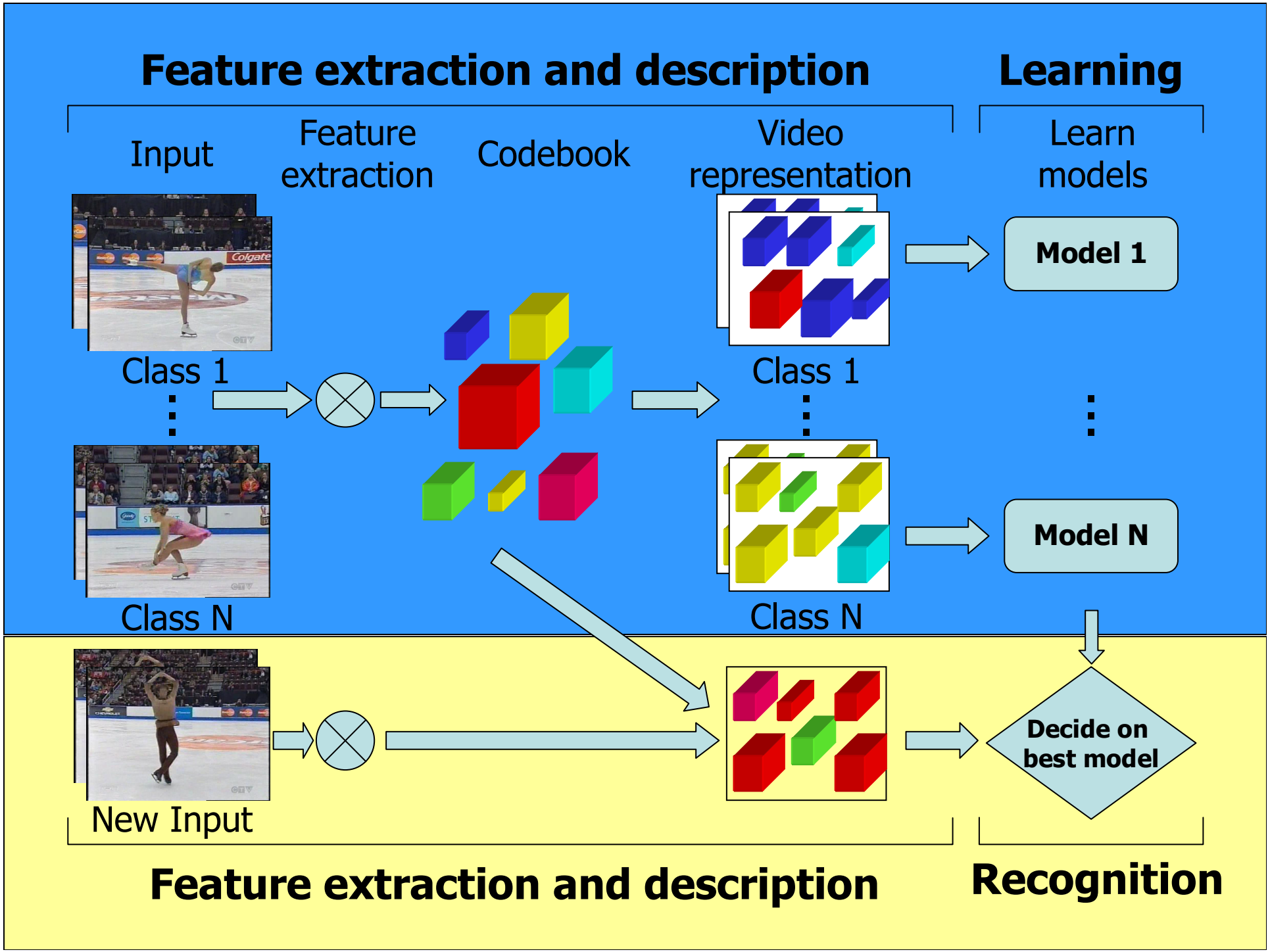


handwaving

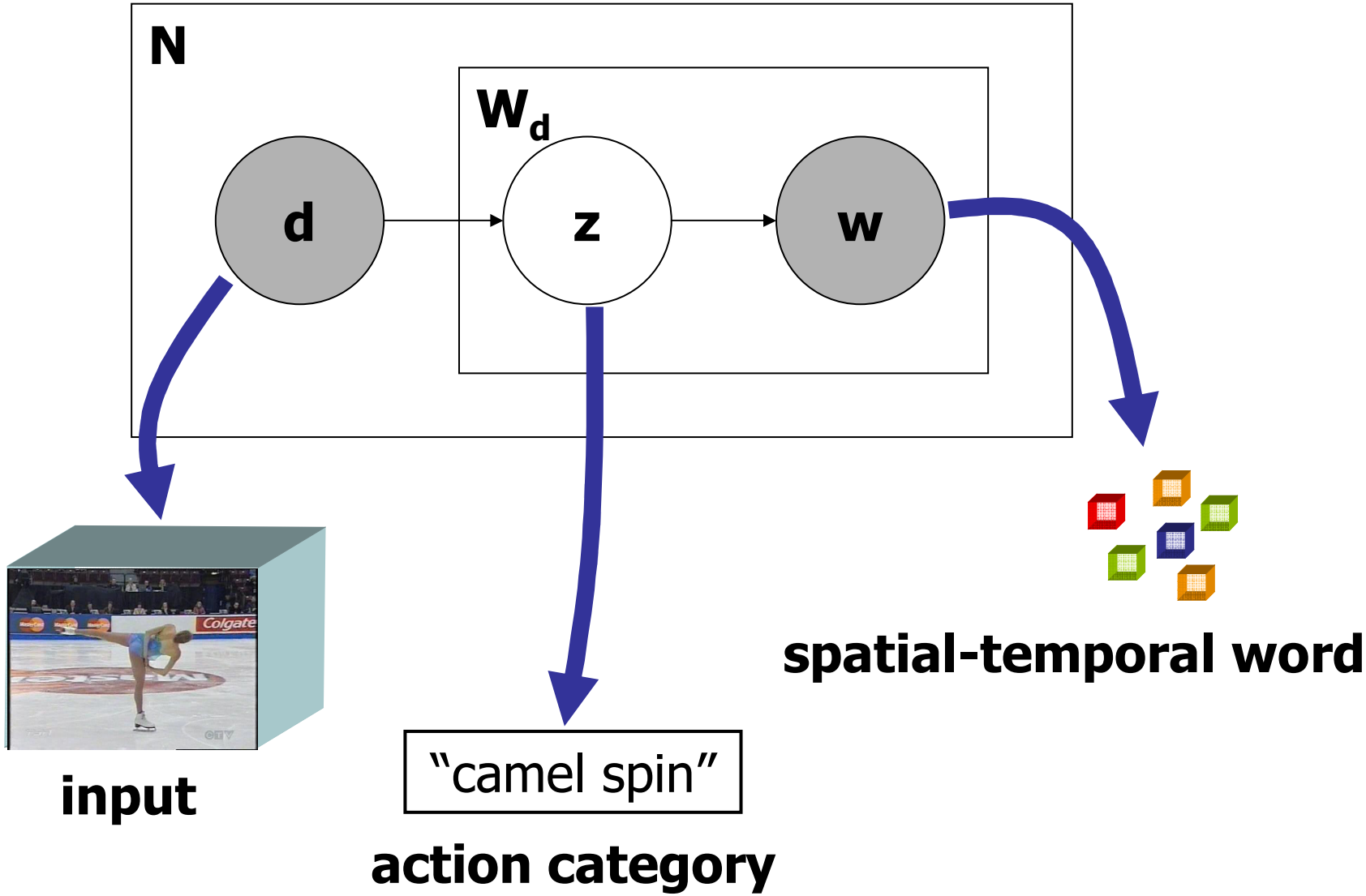
handclapping



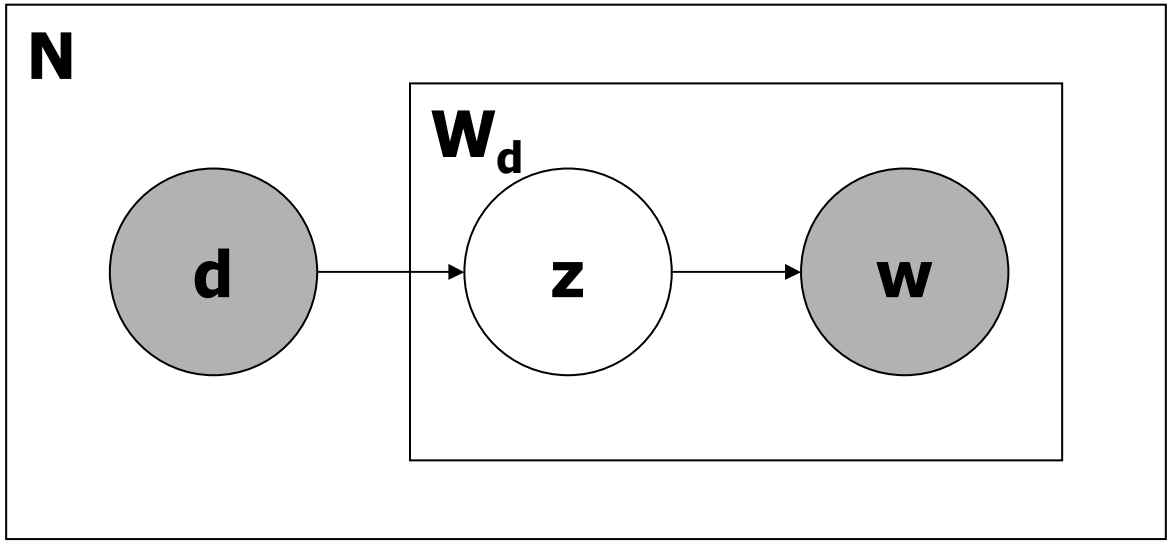
boxing


Feature extraction and description
Learning
Input
Feature extraction
Codebook
Video representation
Learn models
Class 1
Class N
Class 1
Class N
Model 1
Model N
New Input
Decide on best model
Feature extraction and description
Recognition

Unsupervised learning using pLSA



pLSA Model



$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$

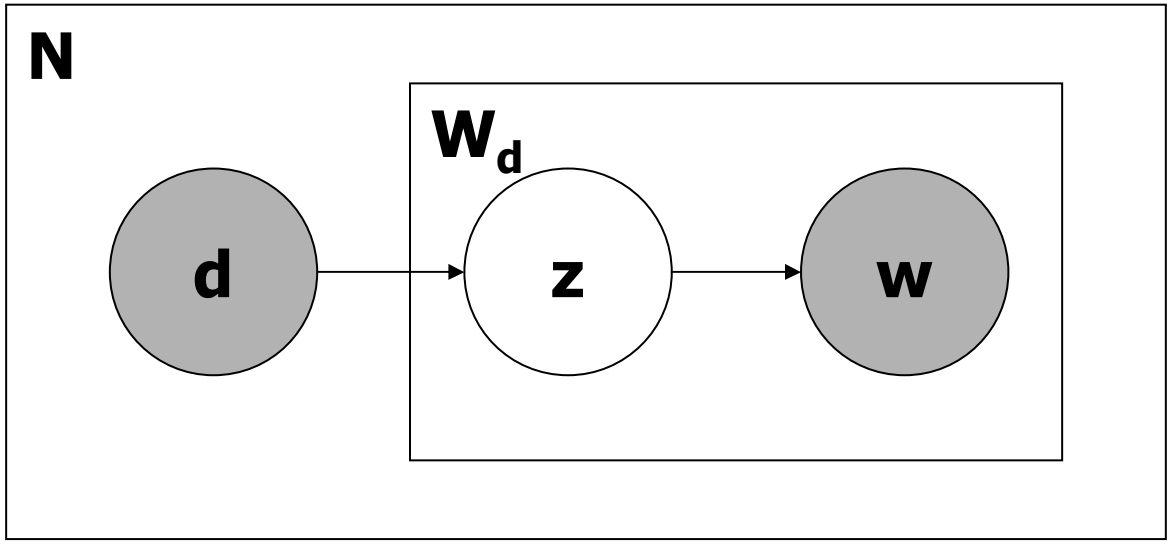
action category vectors

Word distribution per action category

action category weights

Action category distribution per video

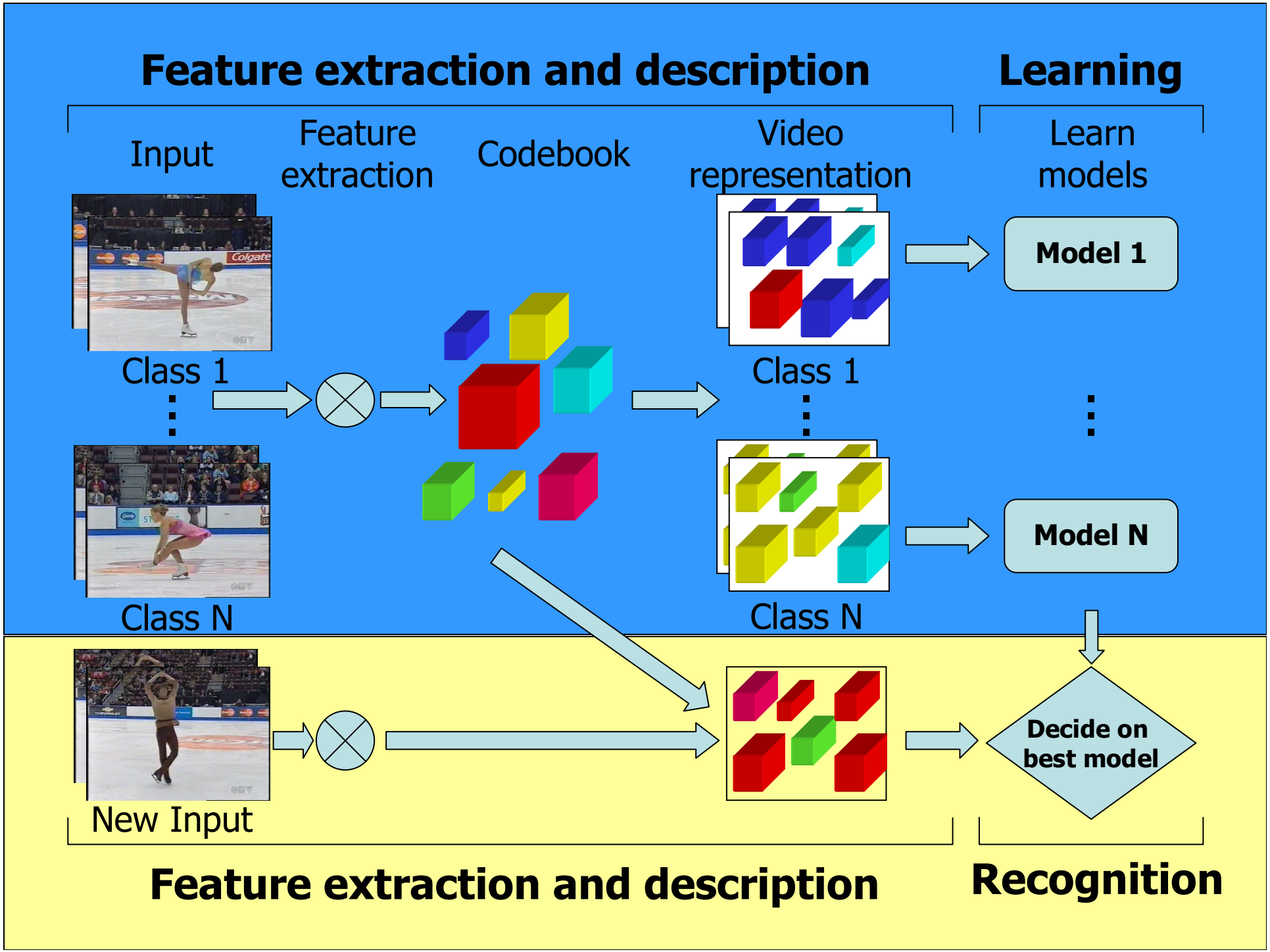
pLSA Model



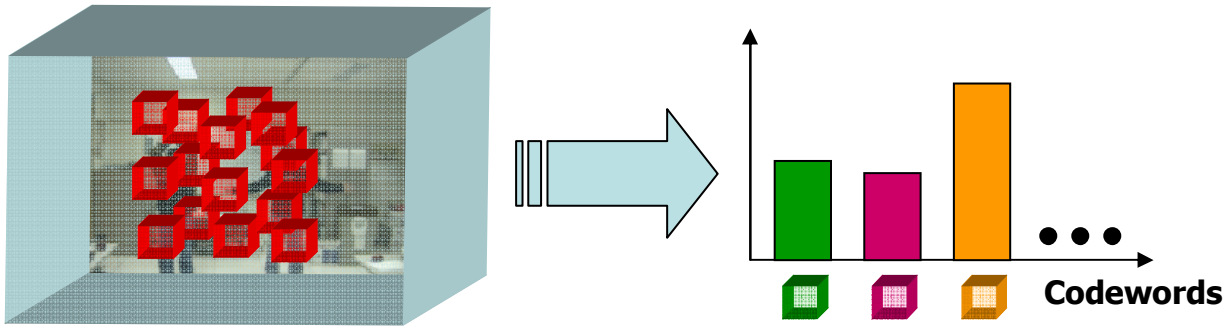
$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$

Unsupervised Learning

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} p(w_i \mid d_j)^{n(w_i, d_j)}$$


Feature extraction and description
Learning
Input
Feature extraction
Codebook
Video representation
Learn models
Class 1
Class N
Class 1
Class N
Model 1
Model N
New Input
Decide on best model
Feature extraction and description
Recognition

Recognition



New video

Action category distribution given test video

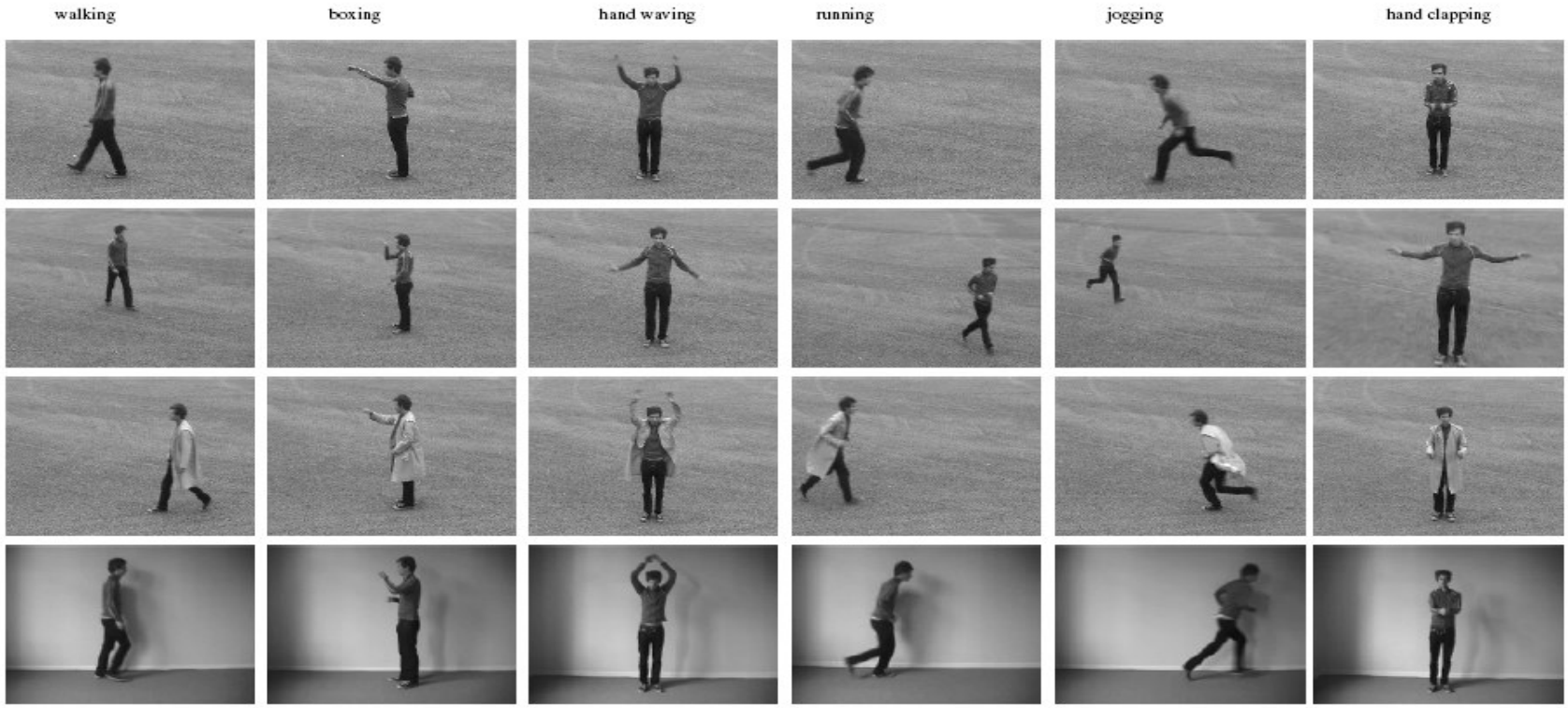
Word distribution given action category

$$P(w \mid d_{test}) = \sum_{k=1}^{K} P(z_k \mid d_{test}) P(w \mid z_k)$$

$$\text{action category} = \arg\max_{k} P(z_k \mid d_{test})$$

Experiment I:

KTH dataset
[Schuldt et al., 2004]:



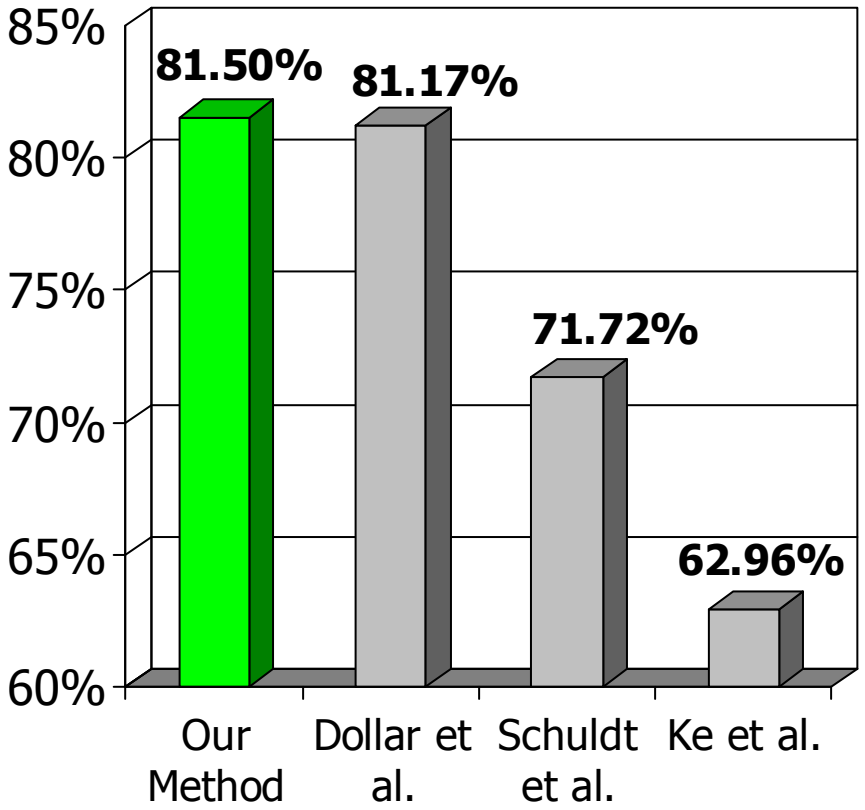
25 persons, indoors and outdoors, 4 long sequences per person

Experiment I: Performance

- Leave-one person out cross validation
- Average performance: 81.50%

walking	.79	.01	.14	.00	.06	.00
running	.01	.88	.11	.00	.00	.00
jogging	.11	.36	.52	.00	.01	.00
handwaving	.00	.00	.00	.93	.01	.06
handclapping	.00	.00	.00	.00	.77	.23
boxing	.00	.00	.00	.00	.00	1.00

- Unsupervised training
- Handle multiple motions



Experiment I: Caltech dataset




walking

running

Trained with the KTH data

Tested with the Caltech dataset

Only words from the corresponding action are shown

Experiment I: A longer sequence

walking

running

Trained with the KTH data

Tested with our own data

Experiment I: Multiple motions

handclapping

handwaving

Trained with the KTH data

Tested with our own data

Experiment II:

Figure Skating data set:
[Y.Wang, G.Mori et al, CVPR 2006]

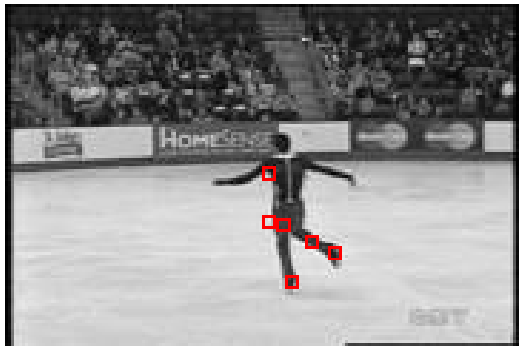
7 persons, 3 action classes: camel spin, stand spin, sit spin

Experiment II: Examples

Figure skating actions

Camel spin

Sit spin



Stand spin

Experiment II: Long Sequences

references

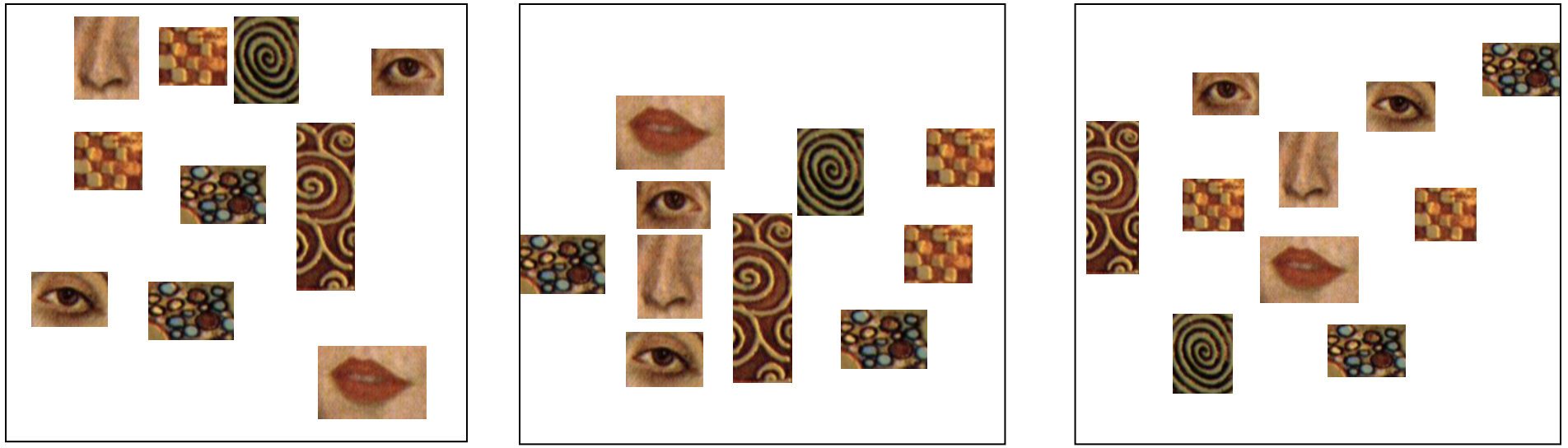
- Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. **IJCV**. September. 2008.

- with
  - Hongcheng Wang

but…

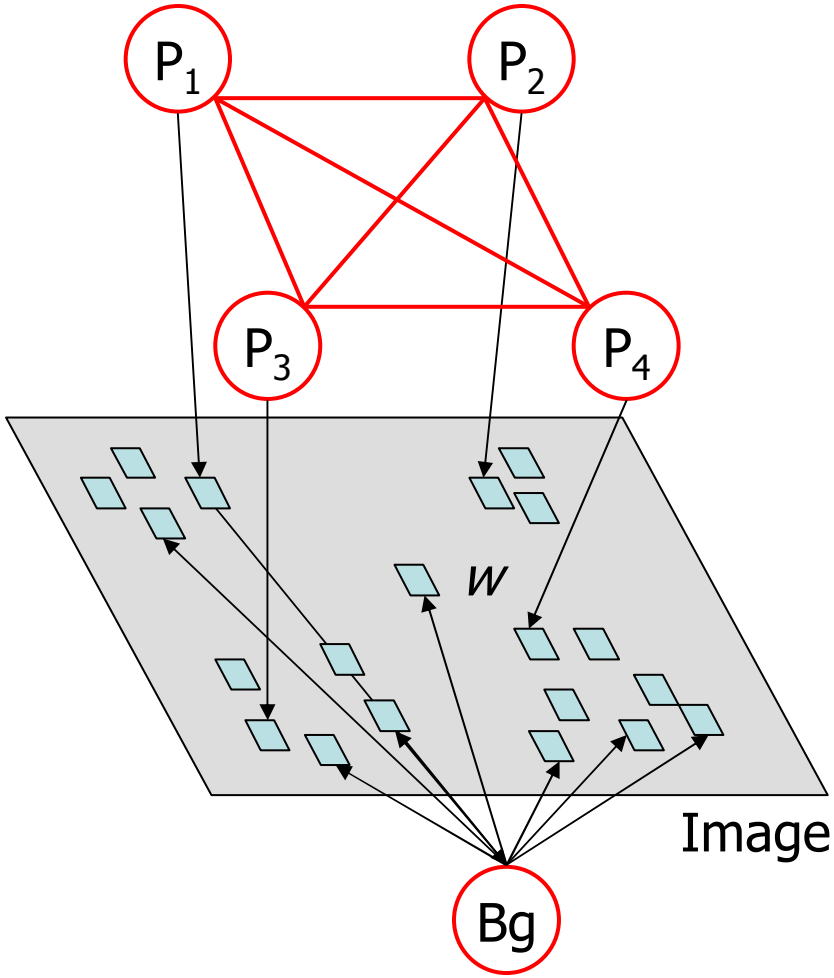
- We need models that exploit geometrical arrangements of features

in the object recognition world



- all have same P under bag-of-words model.
- part-based models that capture geometrical information
  - Constellation model
  - Pictorial structures
  - etc

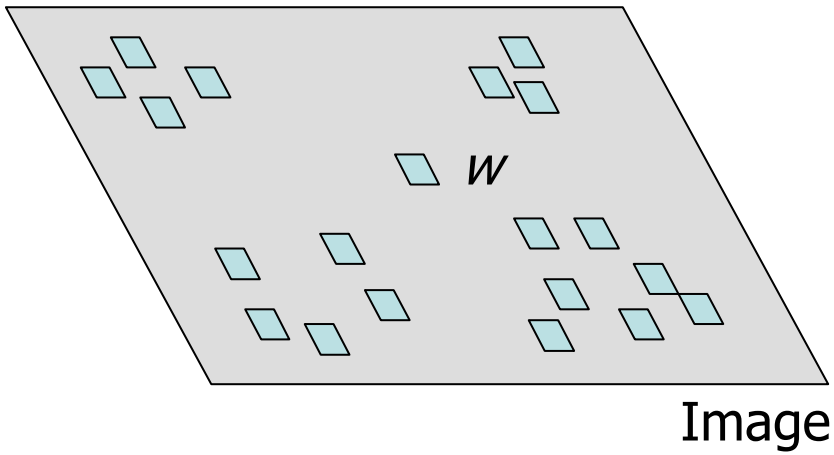
Constellation Model



☹ Small number of features

☺ Strong shape representation

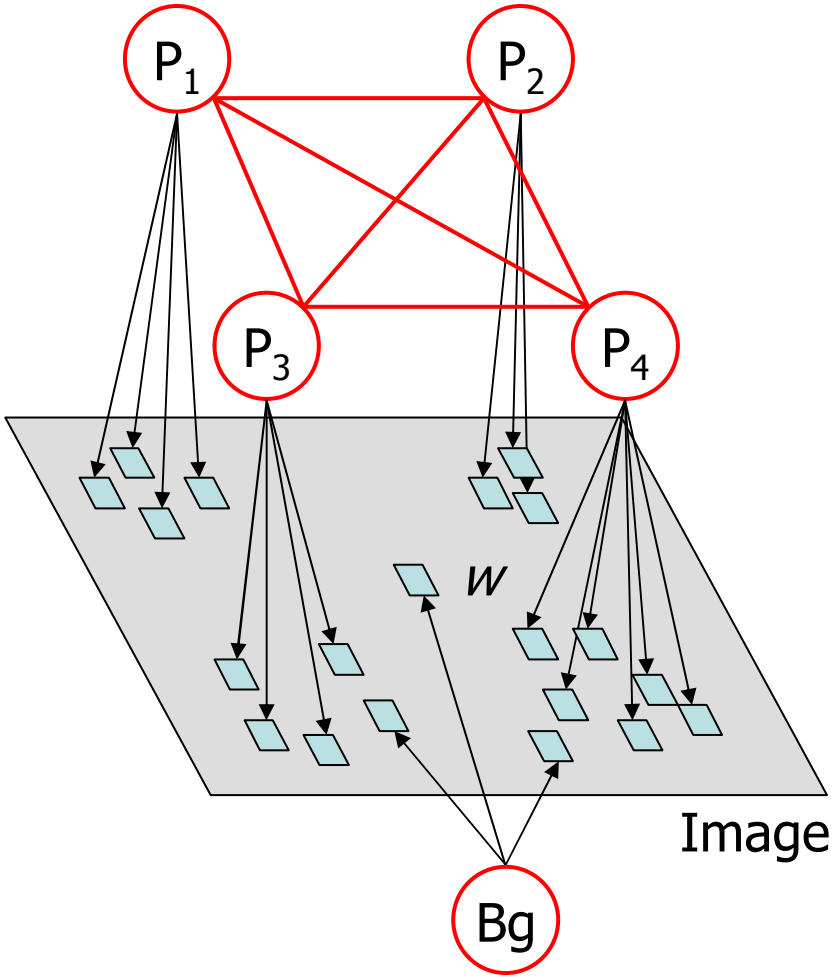
bags of features



☺ Large number of features

☹ No geometrical or shape information

Constellation of bags of features



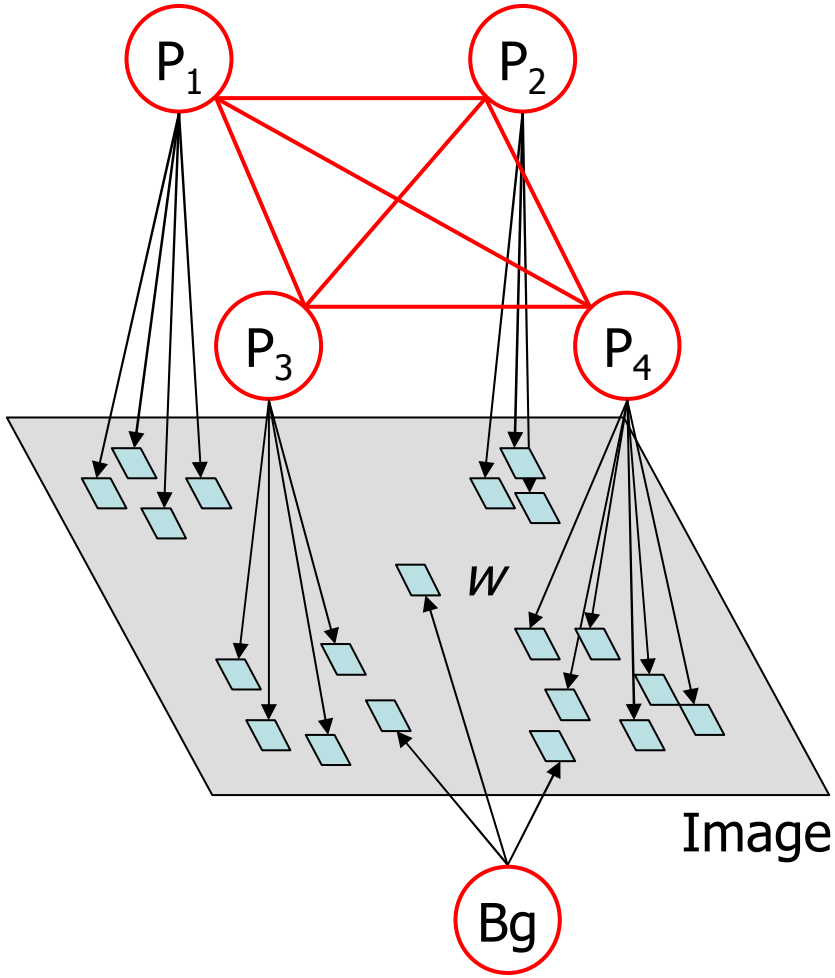
☺ Large number of features

☺ Strong shape representation

Constellation of bags of features

- P_p : Parts
- w : Observed patches (shape, motion and position)

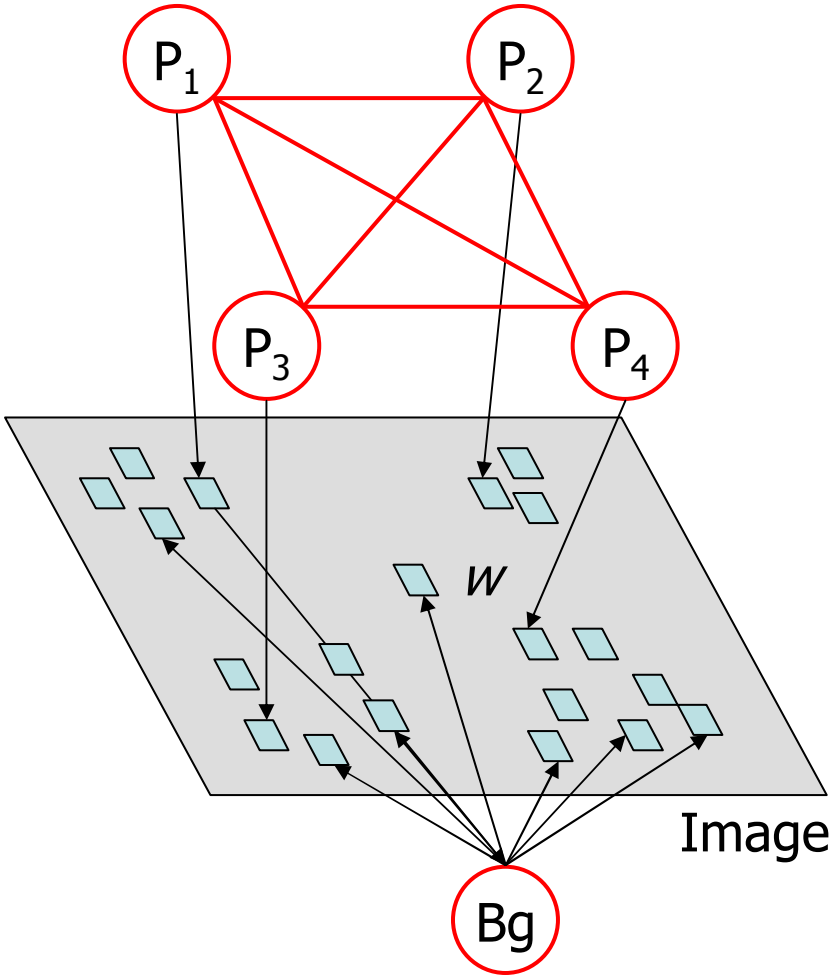
- P's are *similar* to the parts in Constellation model
- Each Part is the parent of a "bag-of-features"



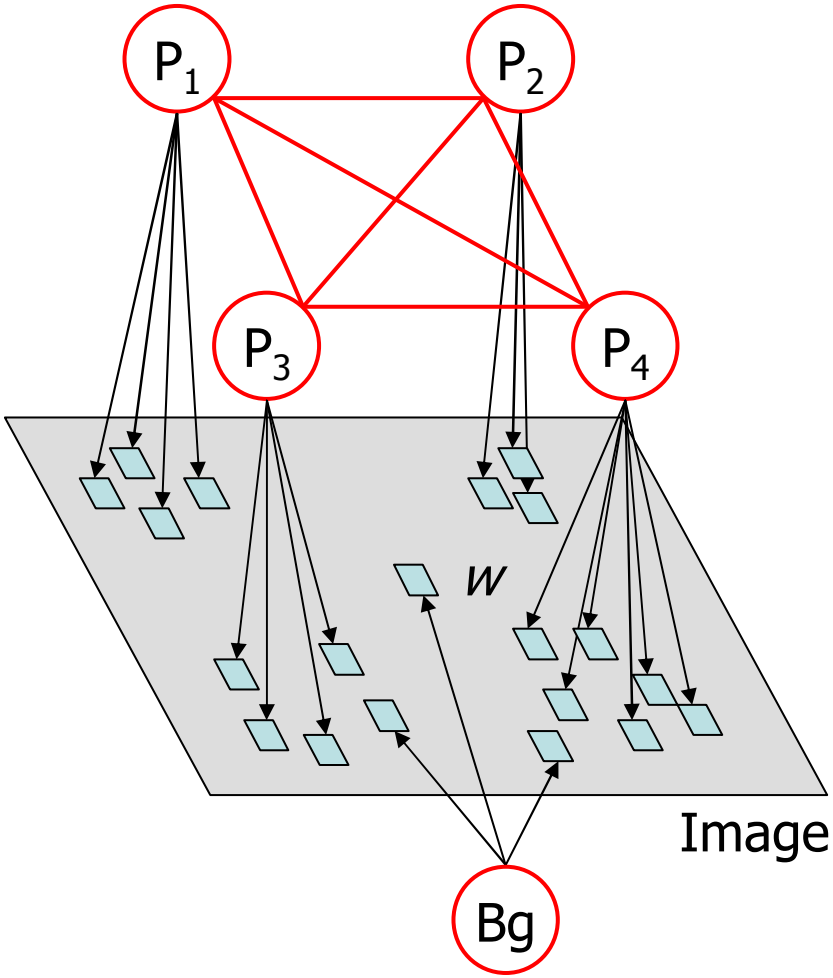
☺ Large number of features

☺ Strong shape representation

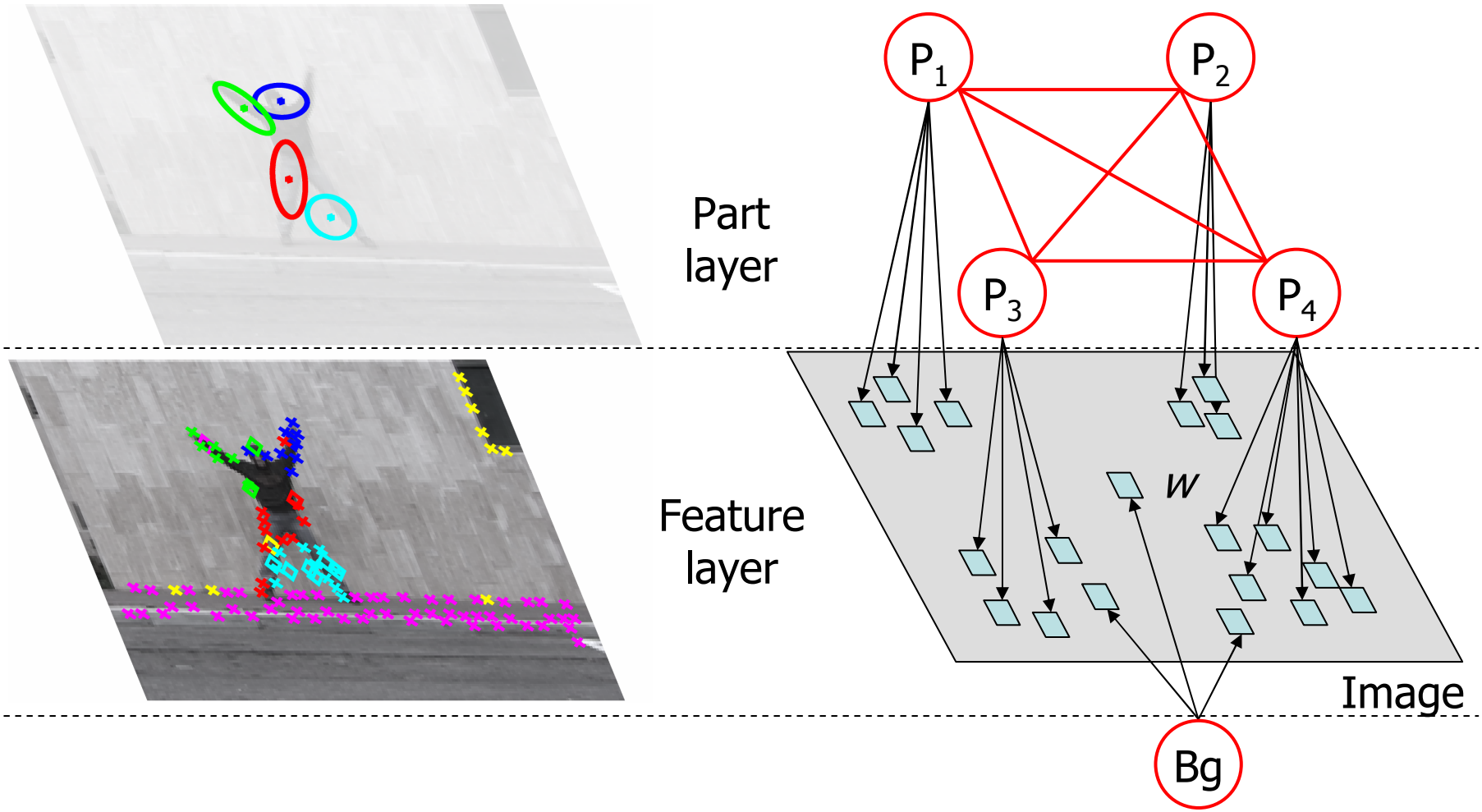
Constellation Model



Constellation of bags of features



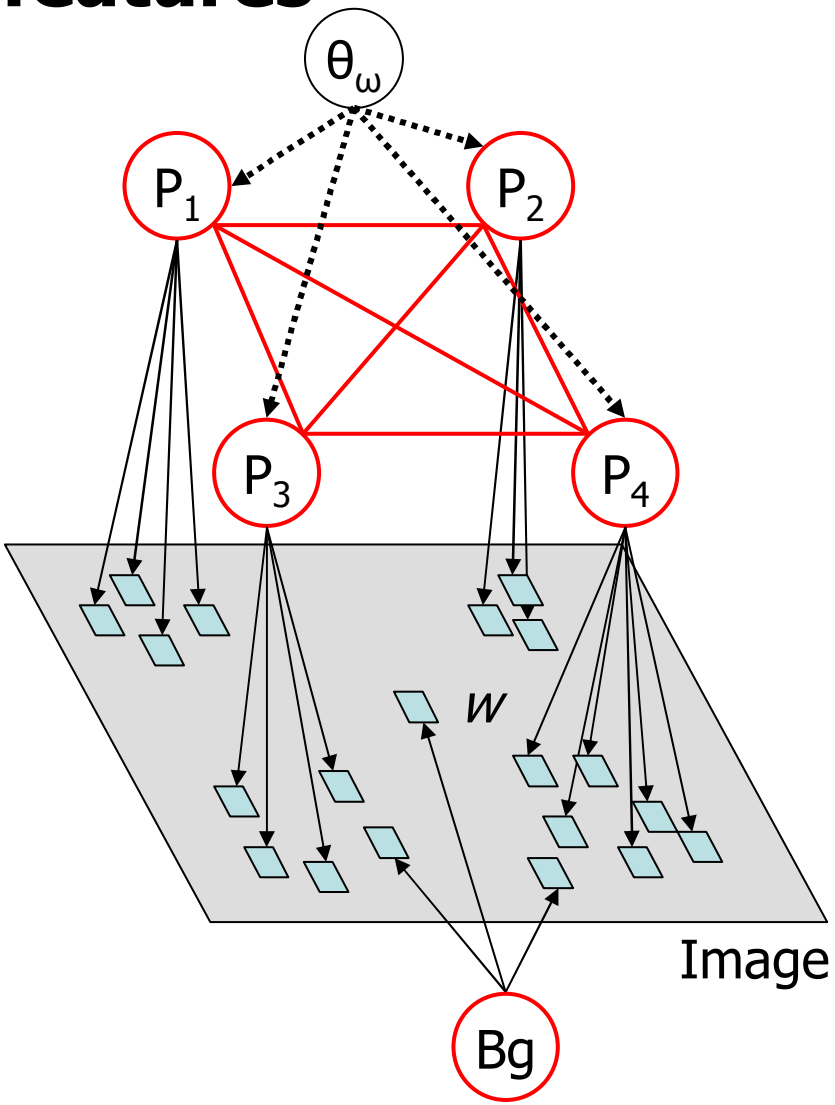
Constellation of bags of features



Constellation of bags of features

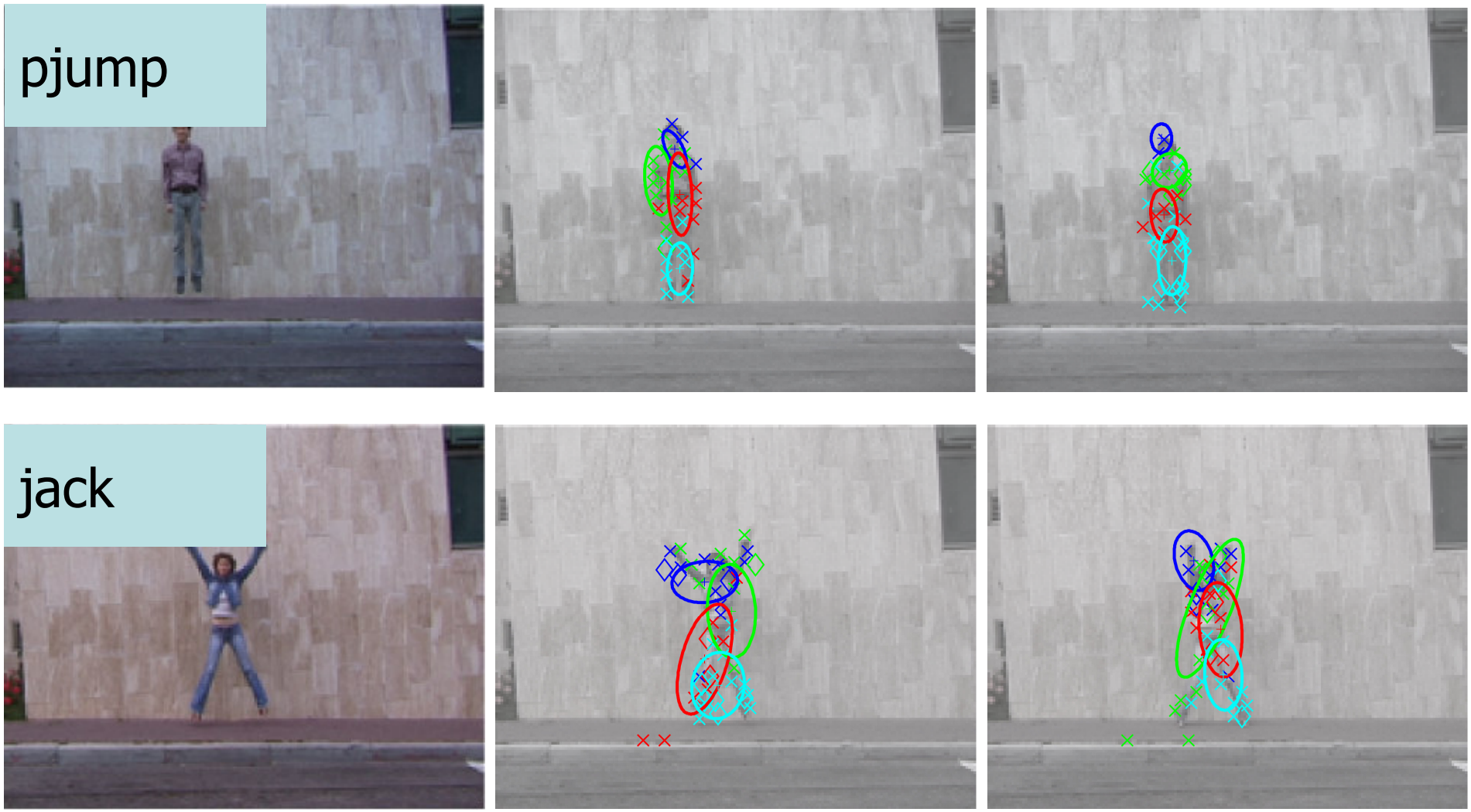
bend

jack

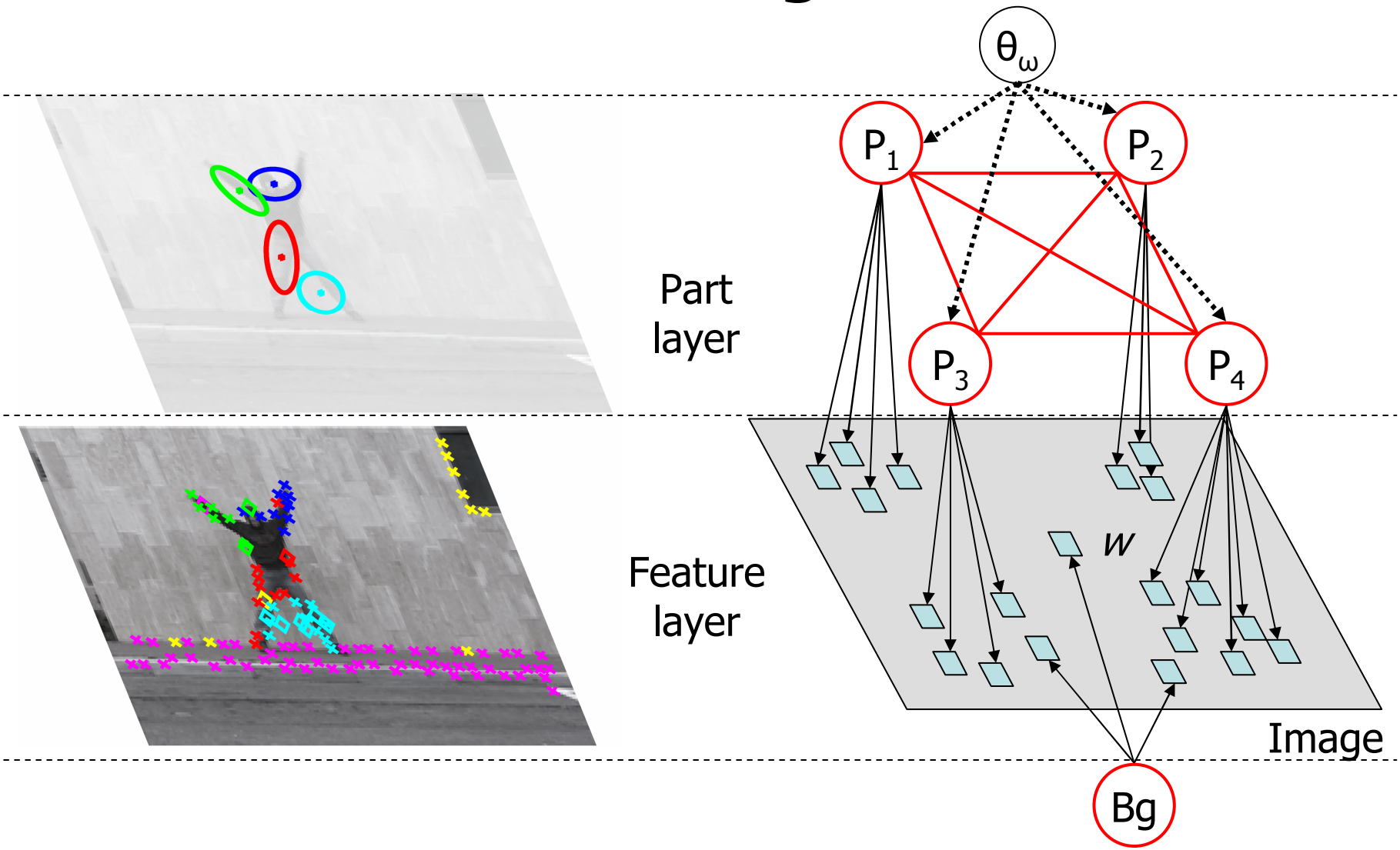


- Use a mixture to account for data multimodality.

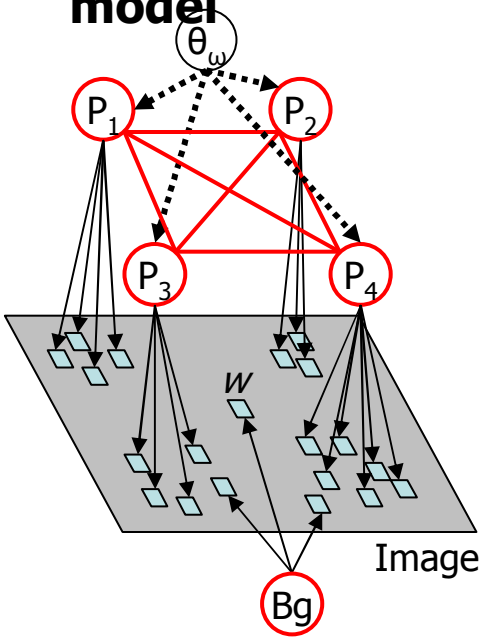
learned models



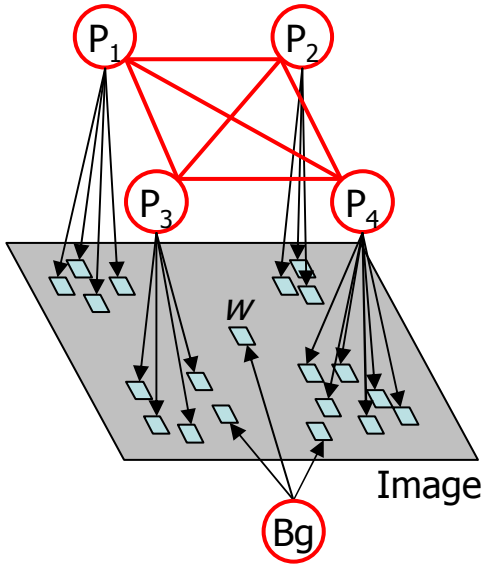
Constellation of bags of features



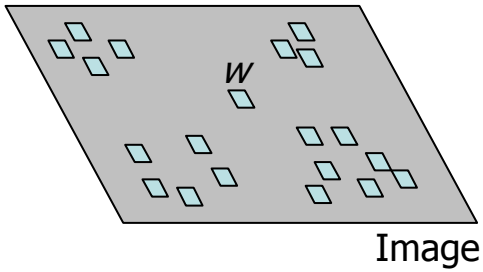
Mixture model



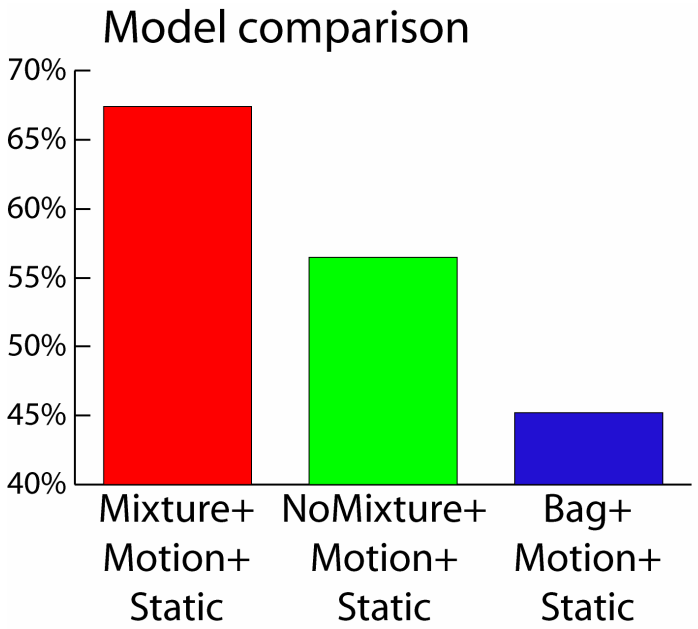
Single component



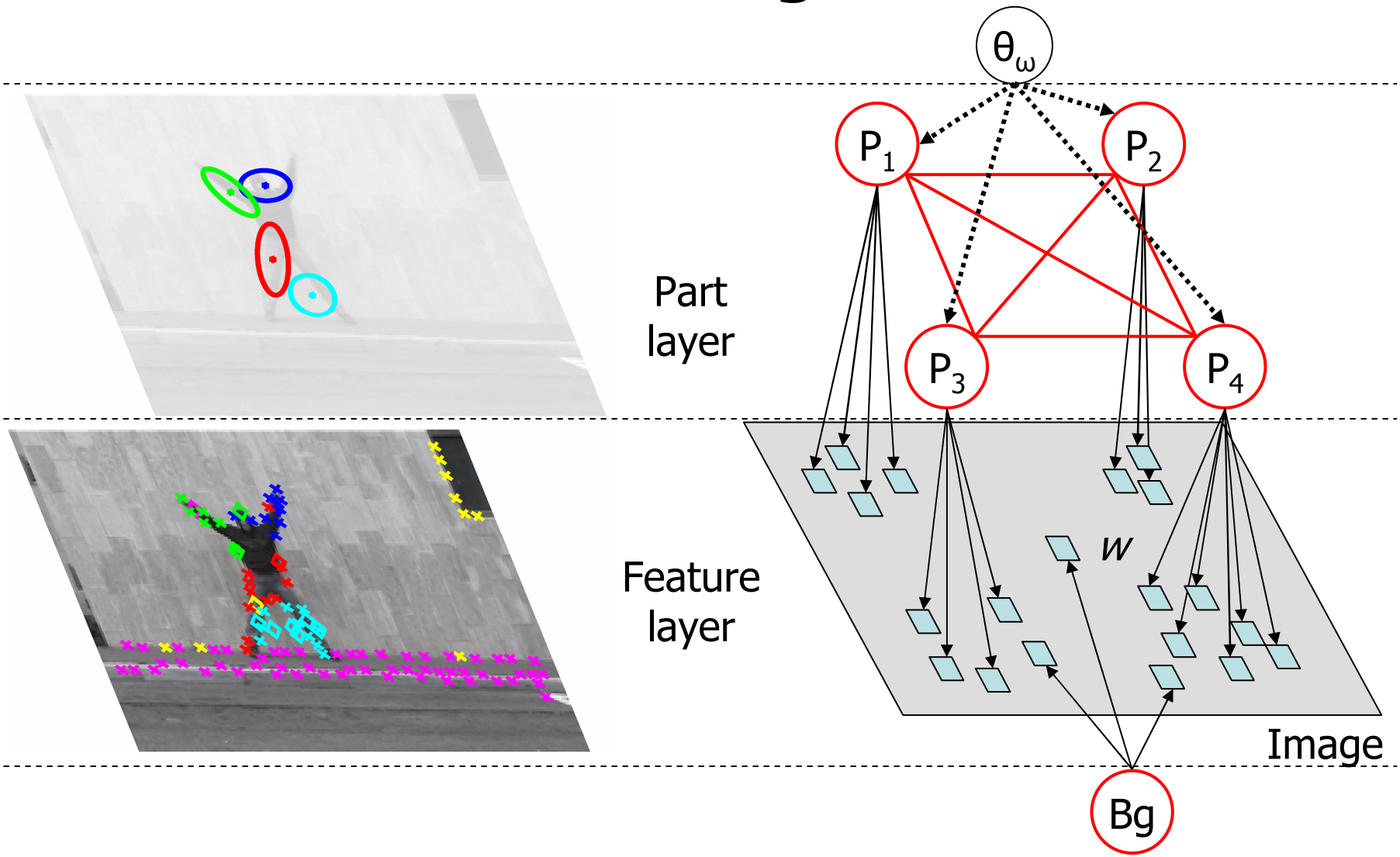
bag of words



Classification performance



Constellation of bags of features



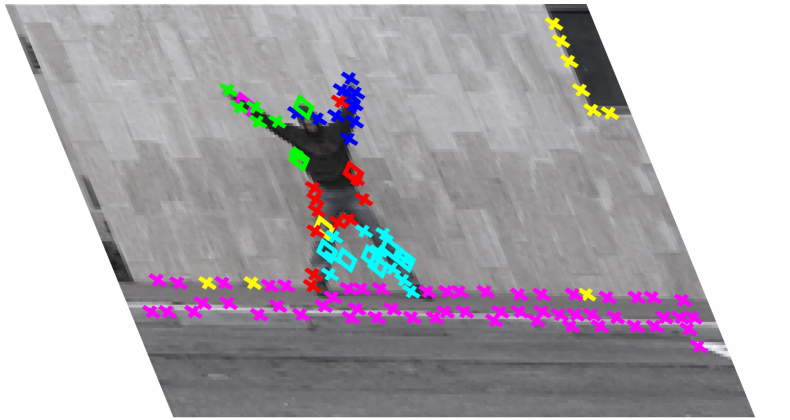
Constellation of bags of features

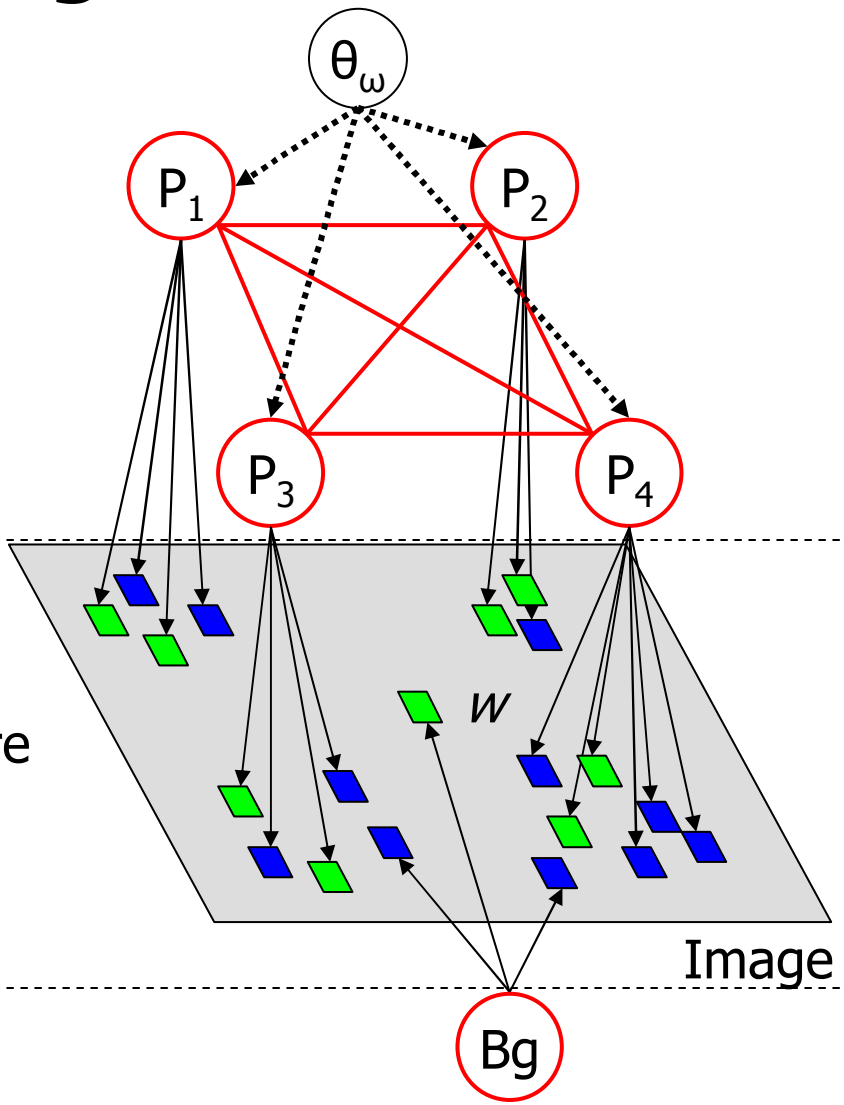
- Static features

  Edge map + Shape Context
  [Belongie '03]

- Motion features

  Interest Points + ST-gradients
  [Dollar '05]





Representation

video frame: **w** = {**x**,**a**}
$w_i = \{x_i, a_i\}$
x_i: i-th feature position
a_i: i-th feature appearance

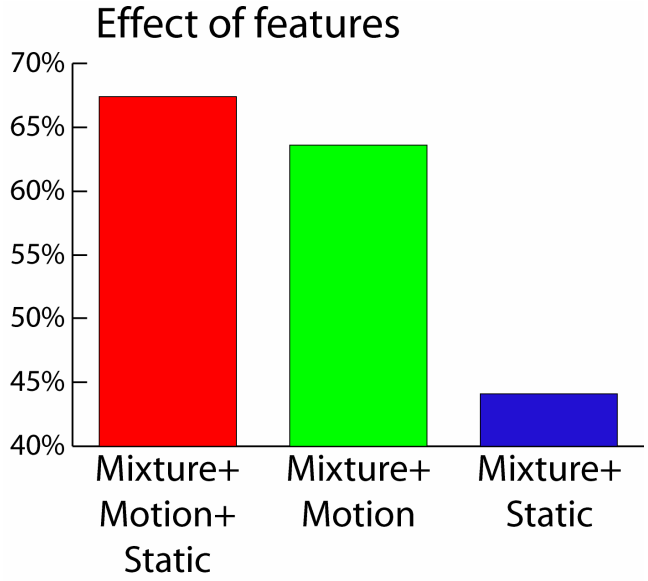
Constellation of bags of features

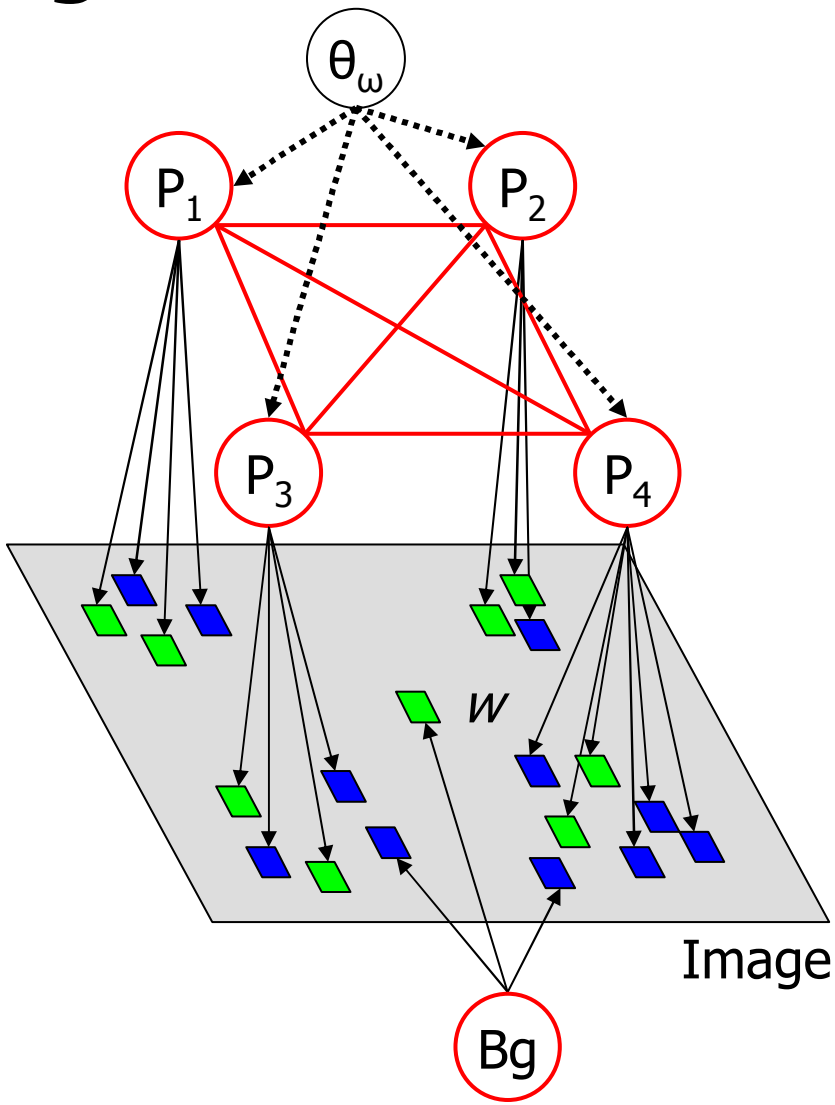
- Static features

  Edge map + Shape Context
  [Belongie '03]

- Motion features

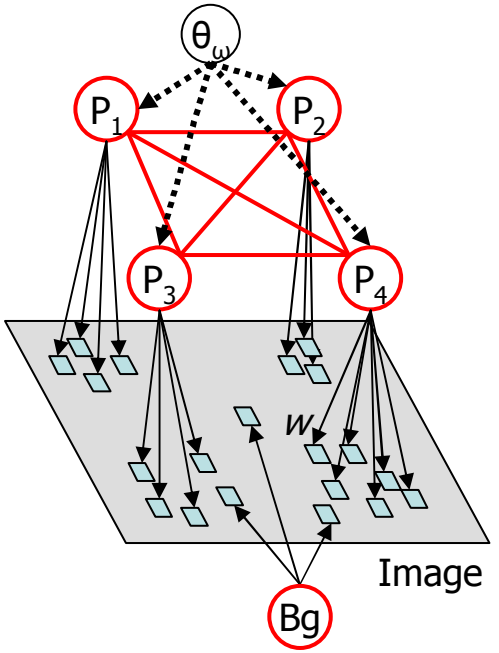
  Interest Points + ST-gradients
  [Dollar '05]





Constellation of bags of features

w	shape, motion and location features
Y	set of possible locations for the P parts (a discretized grid over the image)
h P	index variable to select a specific position of the parts among all the possibilities (in the grid)
m	chooses an specific assignment of features w to parts P

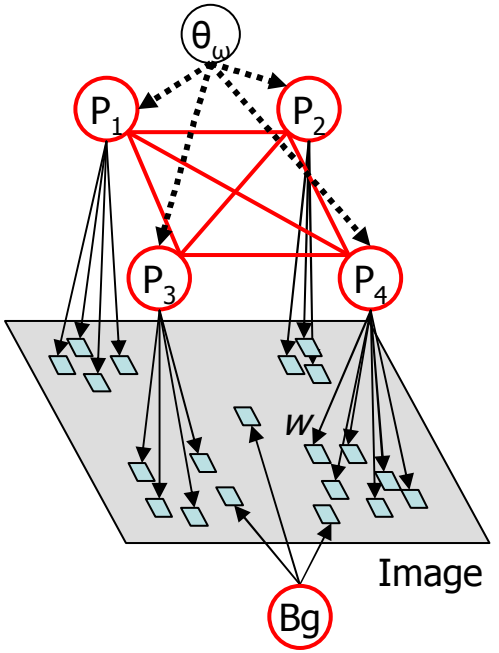


$$p\left(\mathbf{w},\mathbf{Y}\middle|\boldsymbol{\theta}\right) \approx$$

$$\sum_{\omega=1}^{\Omega}\left[\pi_{\omega}\sum_{\mathbf{h}\in H} p\left(\mathbf{h}\middle|\theta_{\omega}\right)\underbrace{p\left(\mathbf{Y}\middle|\mathbf{h},\theta_{\omega}\right)}_{\textit{Part layer}}\underbrace{p\left(\mathbf{w}\middle|\mathbf{Y},\mathbf{m}^{*},\mathbf{h},\theta_{\omega}\right)}_{\textit{Local feature layer}}\right]$$

Constellation of bags of features

w	shape, motion and location features
Y	set of possible locations for the P parts (a discretized grid over the image)
h P	index variable to select a specific position of the parts among all the possibilities (in the grid)
m	chooses an specific assignment of features w to parts P

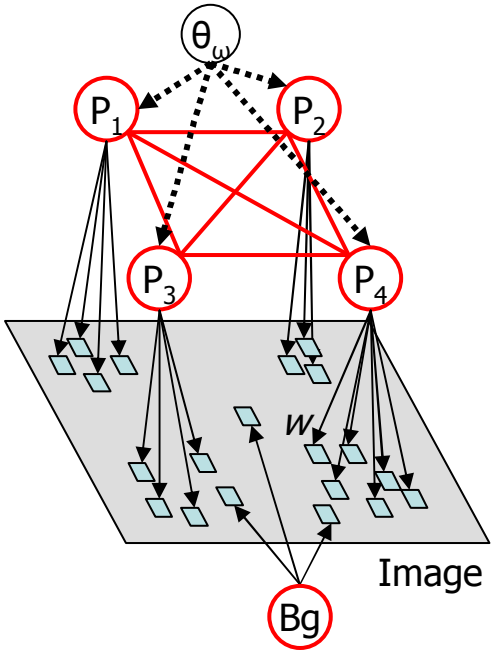


Part layer term: (constellation)

$$p(\mathbf{Y} \mid \mathbf{h}, \theta_\omega) = N(\mathbf{Y}_\mathbf{T}(\mathbf{h}) \mid \boldsymbol{\mu}_{L,\omega}, \boldsymbol{\Sigma}_{L,\omega})$$

Constellation of bags of features

w	shape, motion and location features
Y	set of possible locations for the P parts (a discretized grid over the image)
h P	index variable to select a specific position of the parts among all the possibilities (in the grid)
m	chooses an specific assignment of features w to parts P



Local feature layer term:

$$p\left(\mathbf{w}\middle|\mathbf{Y},\mathbf{m}^{*},\mathbf{h},\boldsymbol{\theta}_{\omega}\right)=$$

$$\prod_{\mathbf{w}_j\in Bg}\underbrace{p\left(x_j^r\middle|\theta_0^X\right)}_{Bg\ Shape}\underbrace{p\left(a_j\middle|\theta_0^A\right)}_{Bg\ Appearance}\prod_{p=1}^{P}\prod_{\mathbf{w}_i\in P_p}\underbrace{p\left(x_i^r\middle|\mathbf{Y},h_p,\theta_p^X\right)}_{Part\ Shape}\underbrace{p\left(a_i\middle|\theta_p^A\right)}_{Part\ Appearance}$$

learning

Constellation shape:
Gaussian

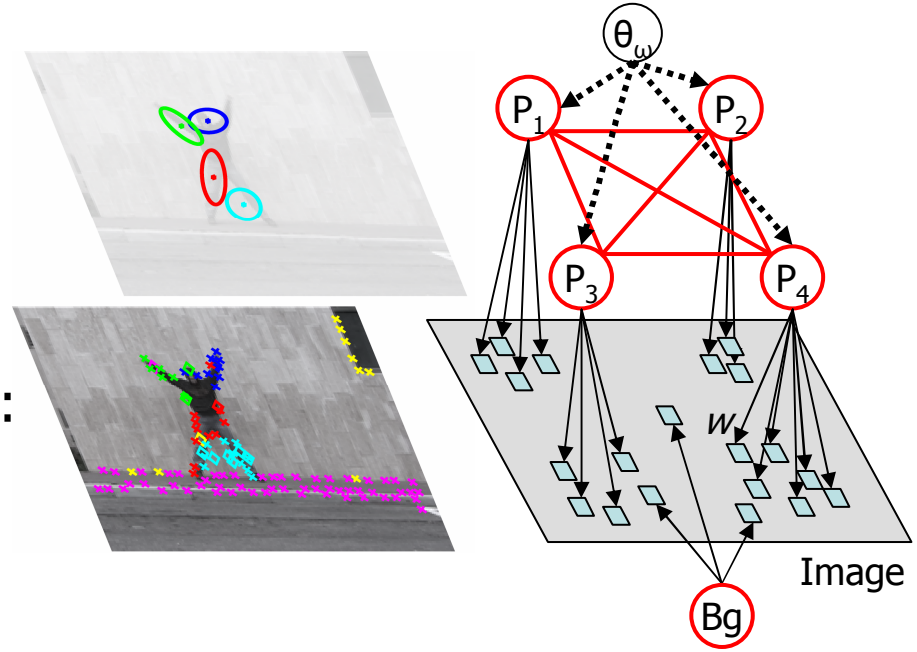
$$\boldsymbol{\mu}_{L,\omega}, \boldsymbol{\Sigma}_{L,\omega}$$

Part shape:
Zero mean Gaussian

$$\boldsymbol{\Sigma}^{X}_{p,\omega}$$

Part appearance:
Multinomial

$$\boldsymbol{\theta}^{A}_{p,\omega}$$

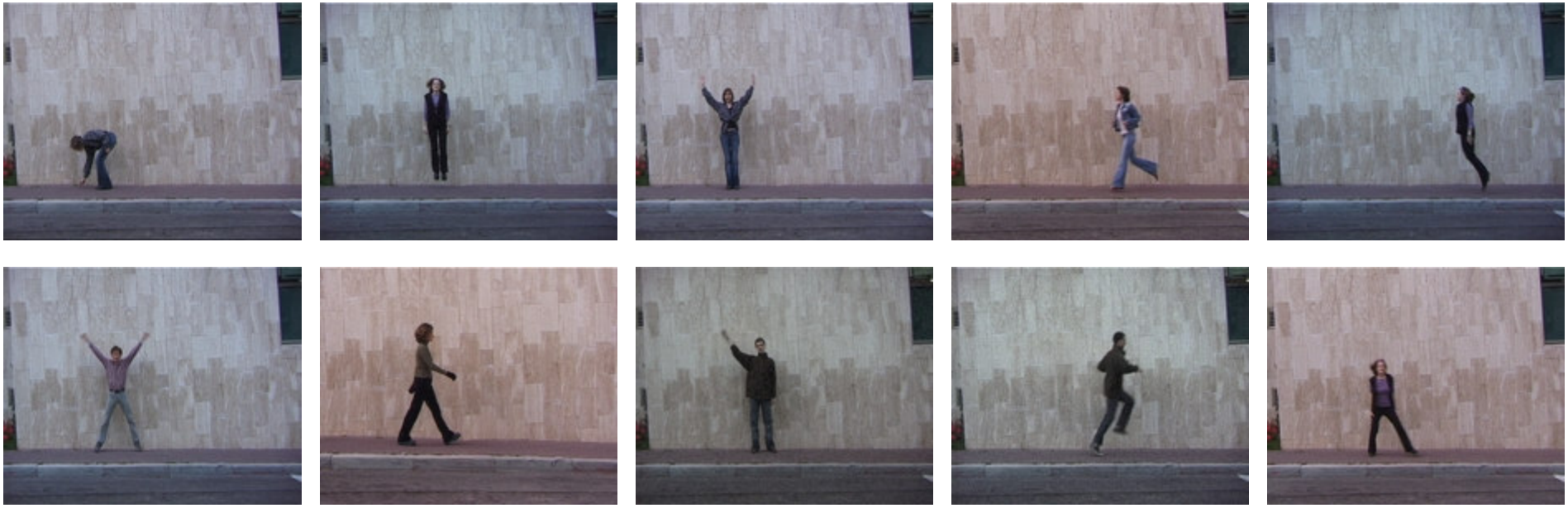


Parameter learning with EM

$$\theta_{\omega} = \left\{ \boldsymbol{\mu}_{L,\omega}, \boldsymbol{\Sigma}_{L,\omega}, \boldsymbol{\Sigma}^{X}_{p,\omega}, \boldsymbol{\theta}^{A}_{p,\omega}, \boldsymbol{\theta}^{X}_{0}, \boldsymbol{\theta}^{A}_{0} \right\} \qquad p = 1 \ldots P$$

$$\omega = 1 \ldots \Omega$$

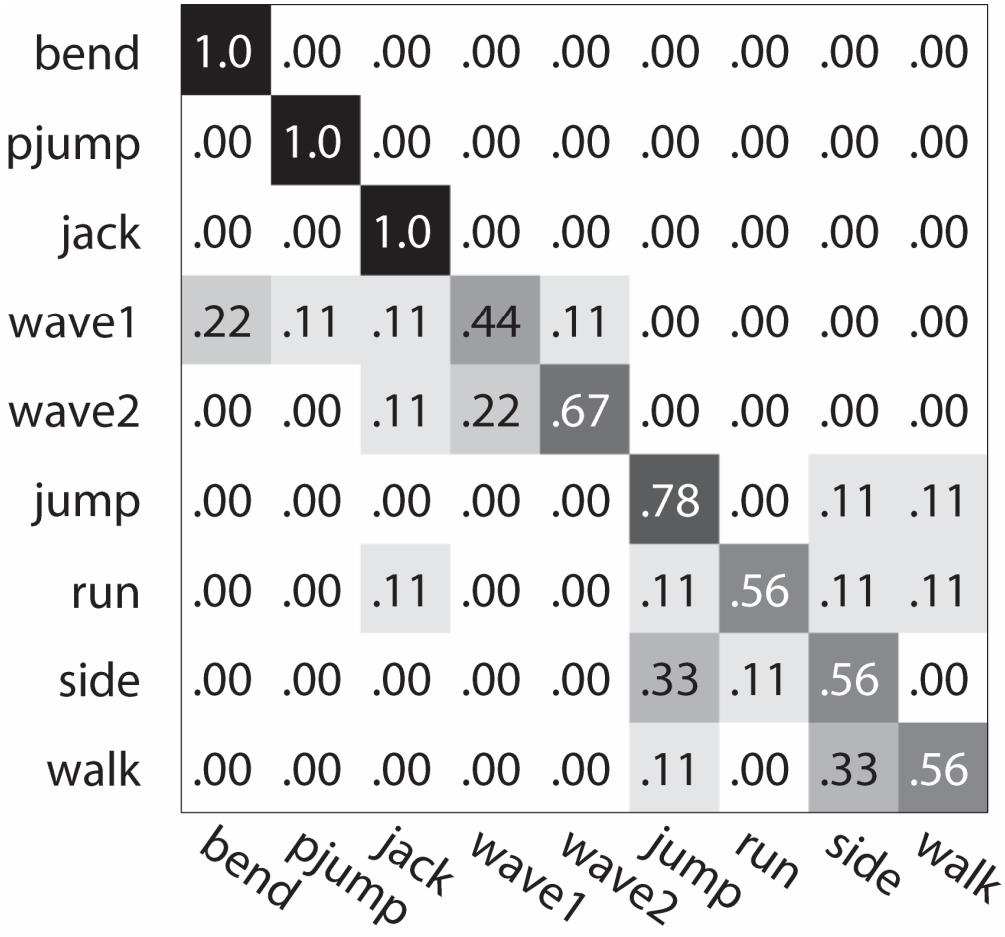
Dataset



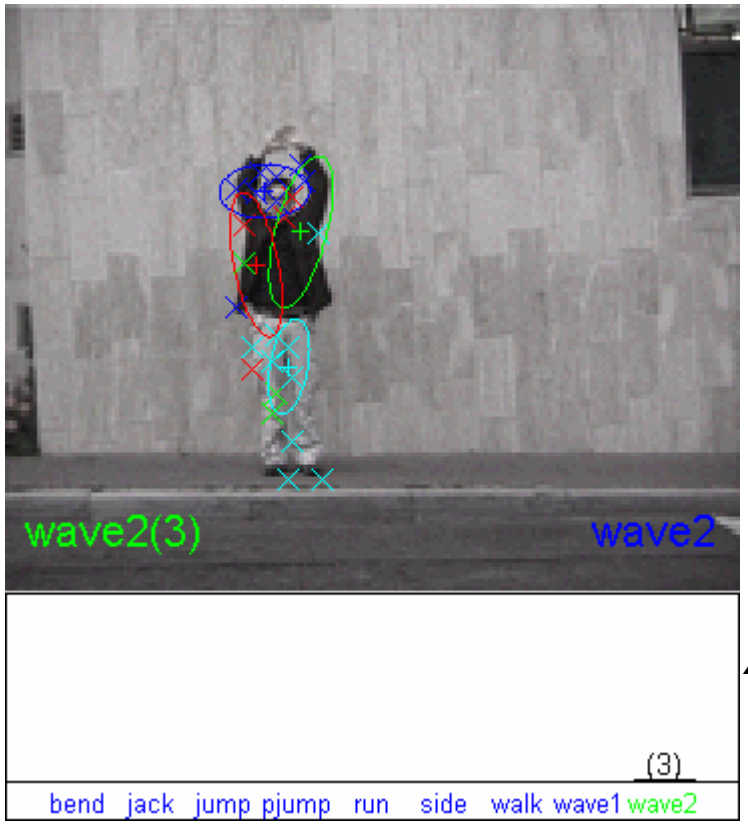
"Actions as Space-Time Shapes"
M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri.
IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

experimental results

- 9 action classes, performed by 9 subjects [Blank et al 2005]
- Leave one out cross-validation
- Video Classification performance: 72.8%



Results



Predicted frame label

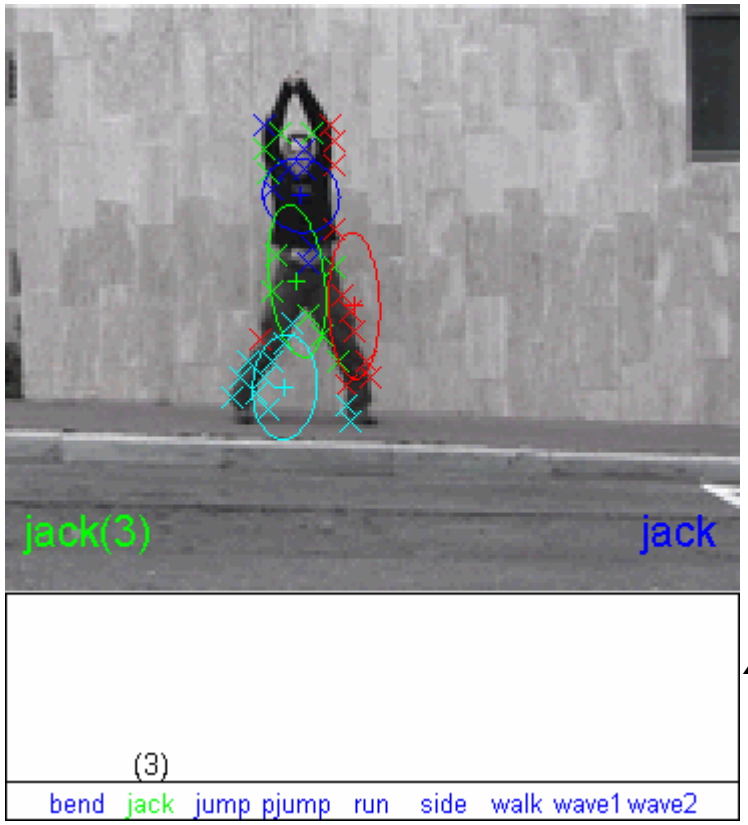
Ground truth frame label

Predicted sequence label

Accumulated votes

Correct **Incorrect**

Results



Predicted frame label

Ground truth frame label

Accumulated votes

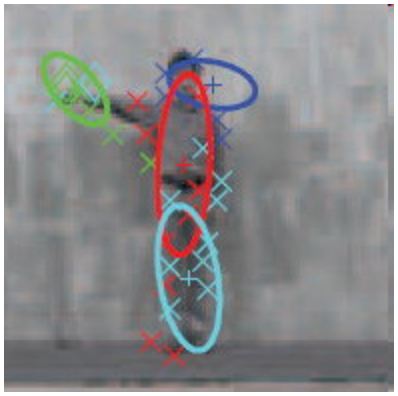
Predicted sequence label

Correct

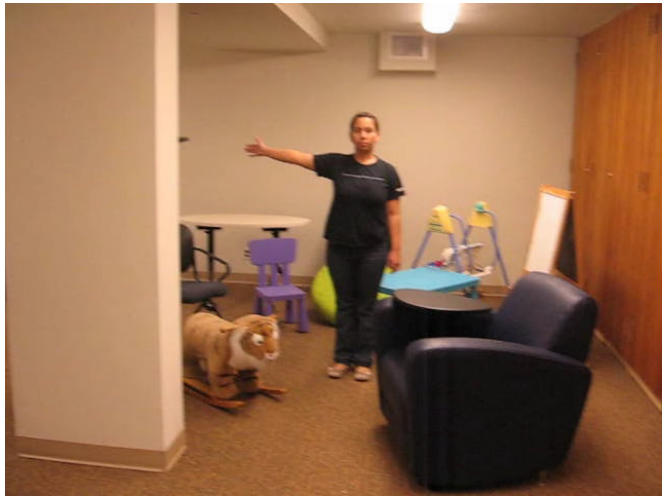
Incorrect

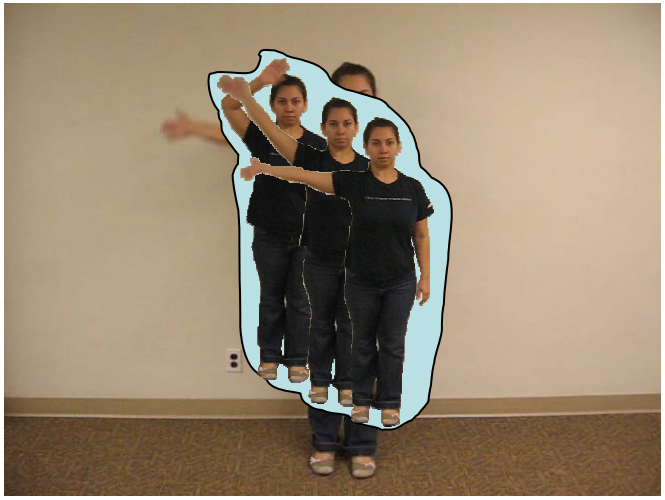
references

- Juan Carlos Niebles and Li Fei-Fei. A Hierarchical Model of Shape and Appearance for Human Action Classification. **CVPR** , 2007



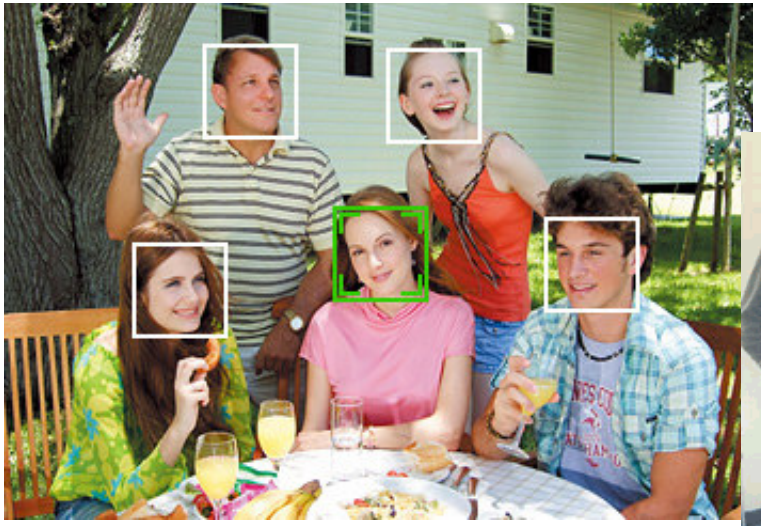
In this talk

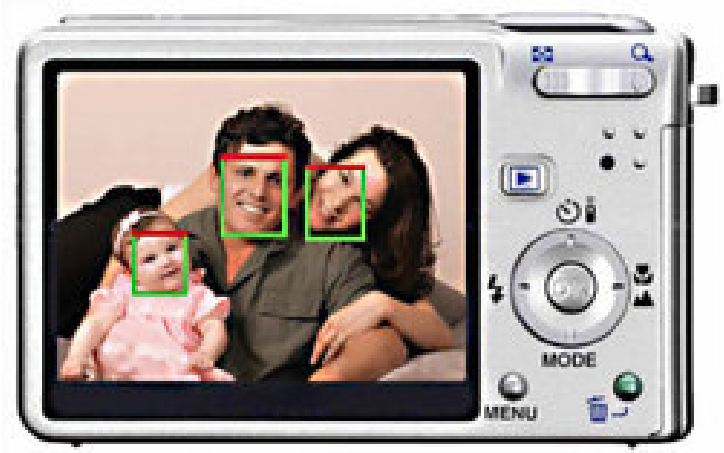




- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences

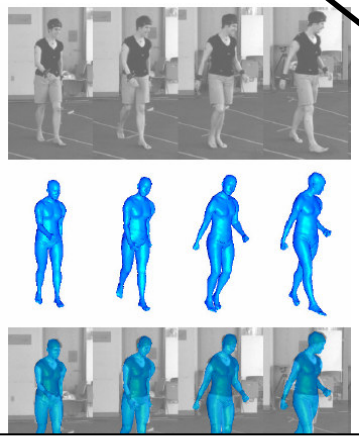
- detect moving people in YouTube videos
- extract the spatio-temporal volume that contains each person





Fast and robust algorithms for face detection

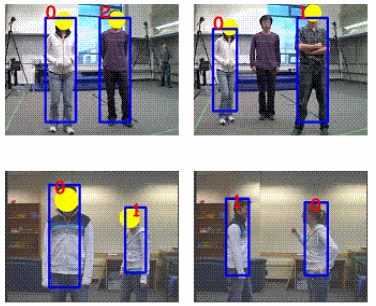
complexity of video sequences



Balan et al '07

real-world sequences

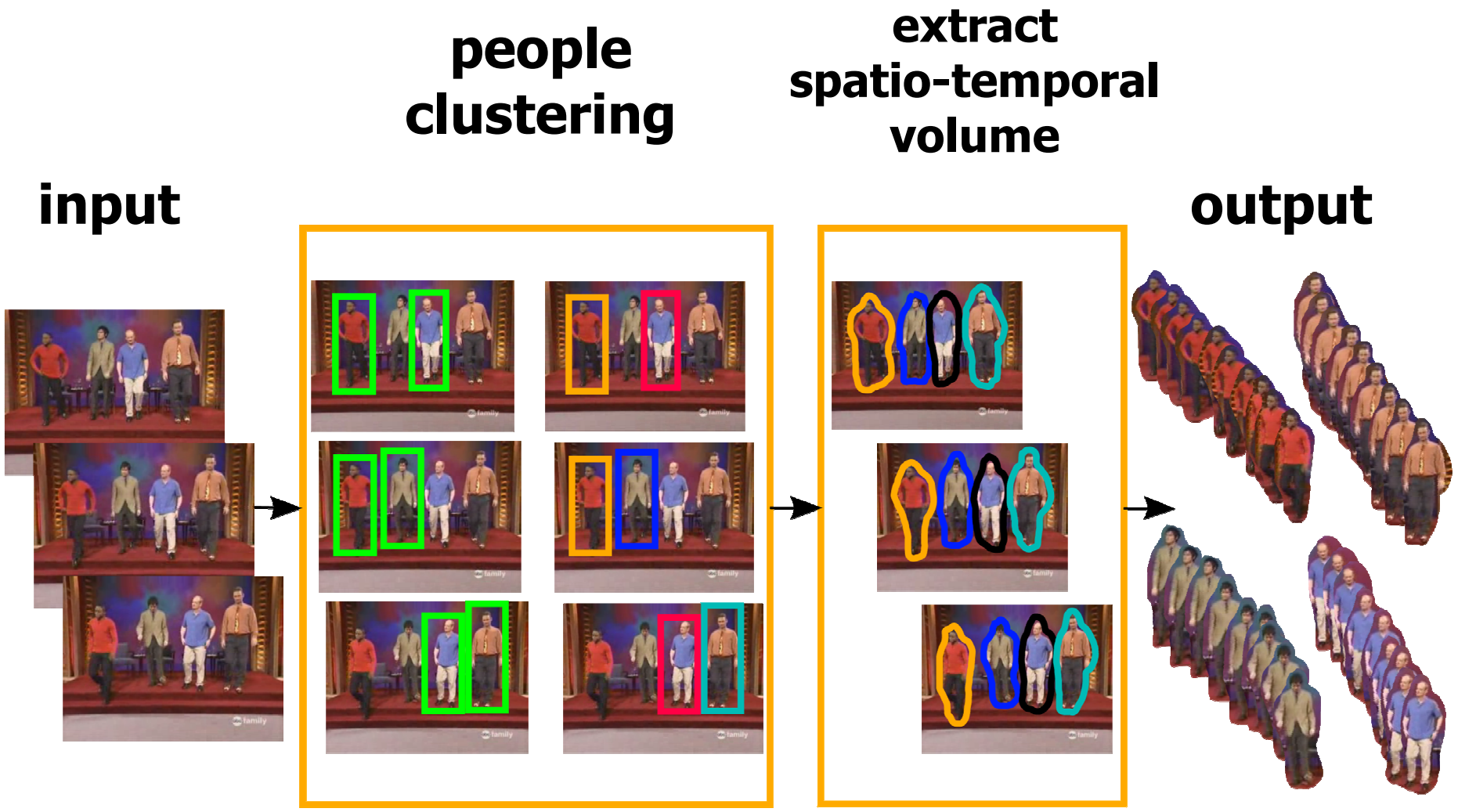
- Compression artifacts & low quality
- Videos contain multiple shots
- Unknown number of humans
- Arbitrary human motion and poses
- Unknown camera parameters and motion
- Background clutter, motion and occlusions



Wu&Nevatia '07
...

extraction detail

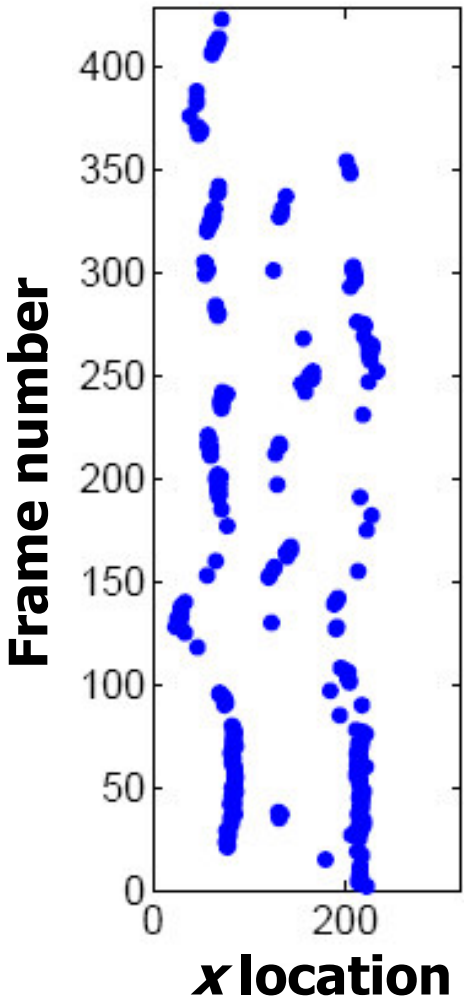
System Overview



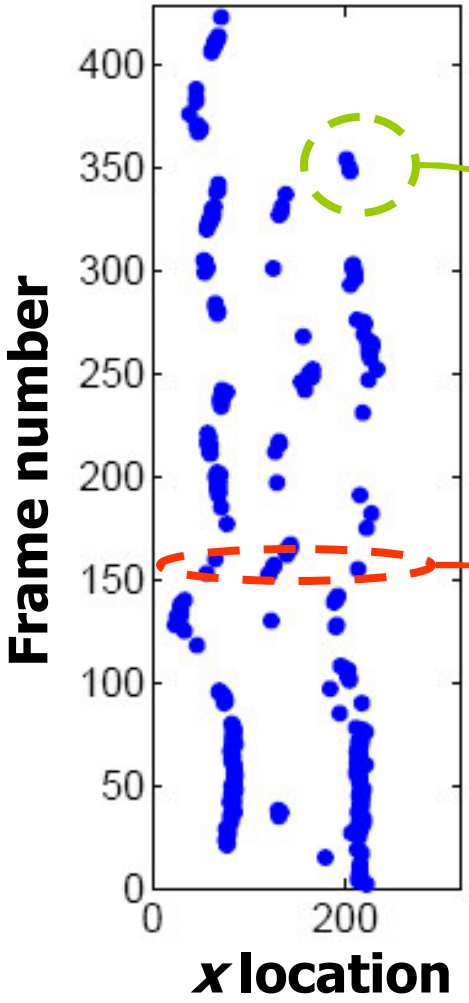
pedestrian detection

- Upright human detection is somewhat successful
  - Dalal & Triggs 05, Sabzmeydani & Mori 07, Laptev 06, and more

clustering ped detections



clustering ped detections

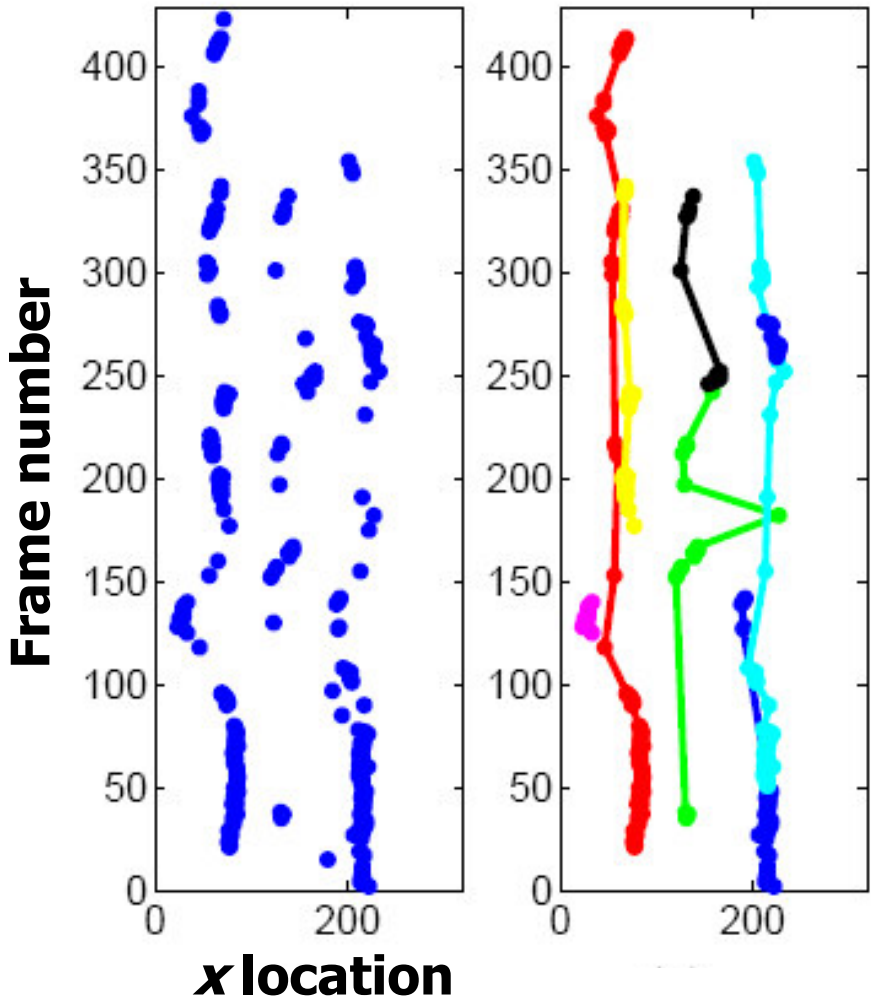


- Must-link: detections with similar appearance and consistent spatial locations

- Cannot-link: detections that occur within the same frame

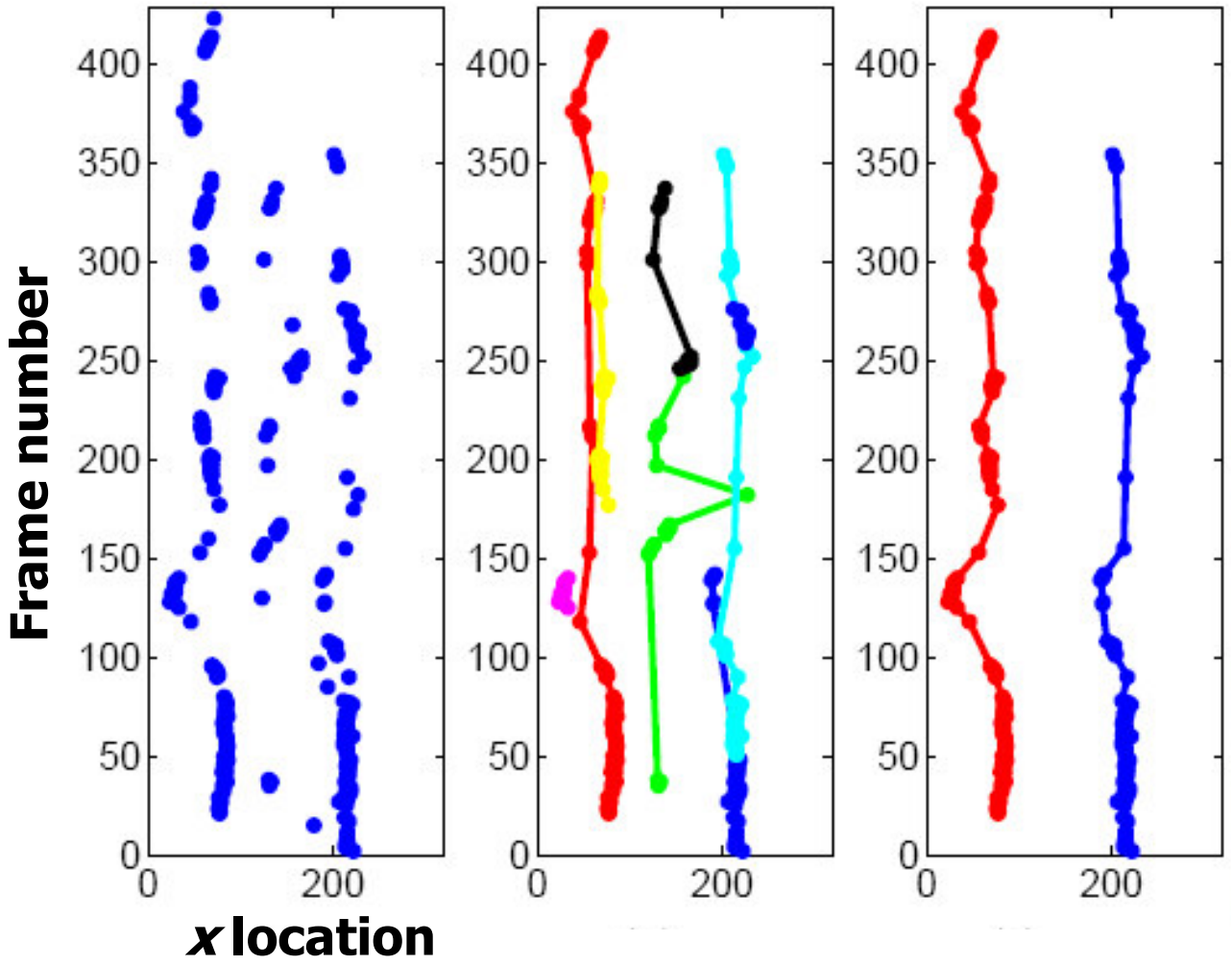
Ø Clustering with constraints, Klein et al '02

First clustering step



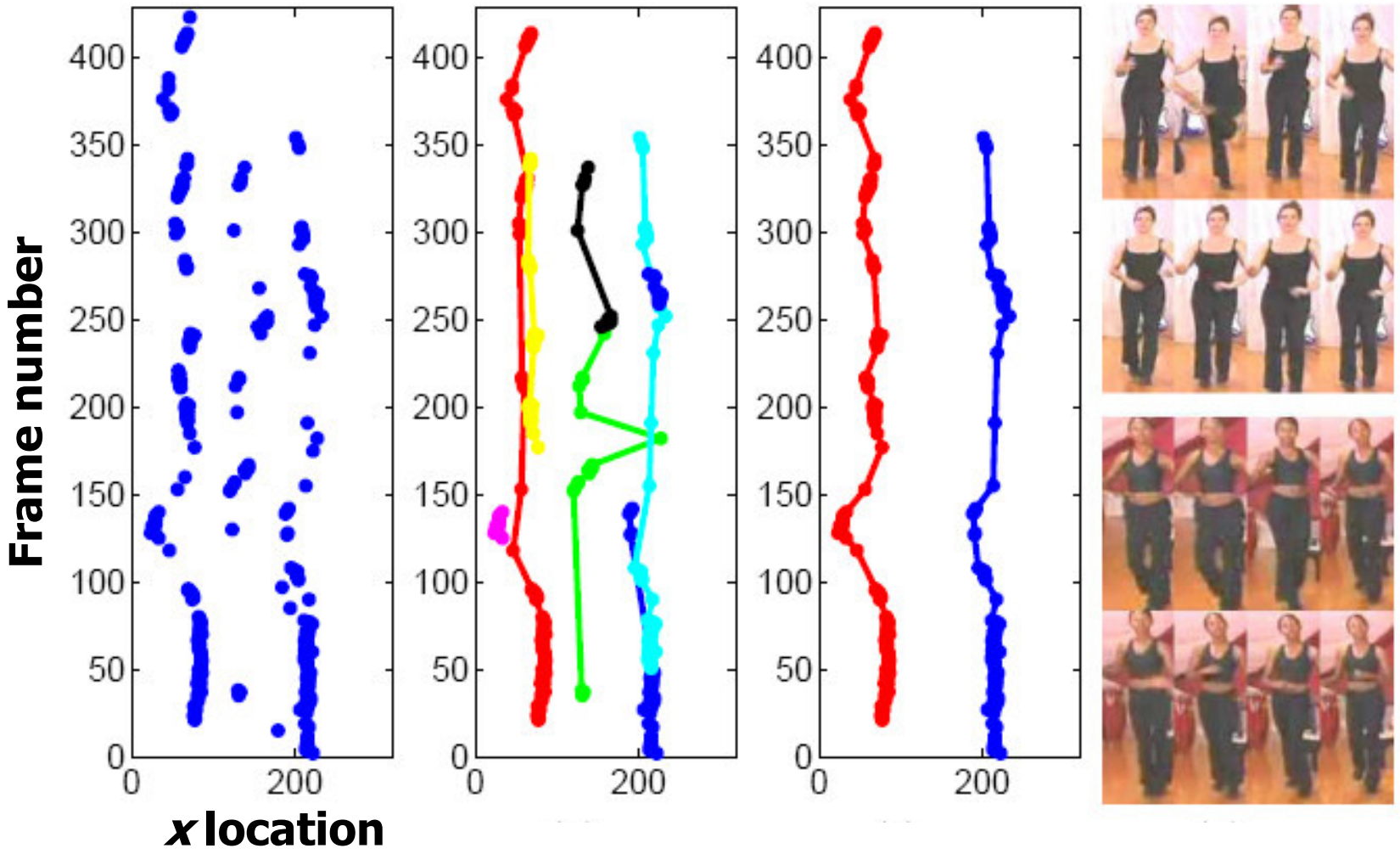
- Clustering with constraints [Klein et al ’02]
- Use simple & cheap descriptors:
  global color histogram
- Gives ‘over-segmented’ clusters
  Conservative clustering threshold

Second clustering step



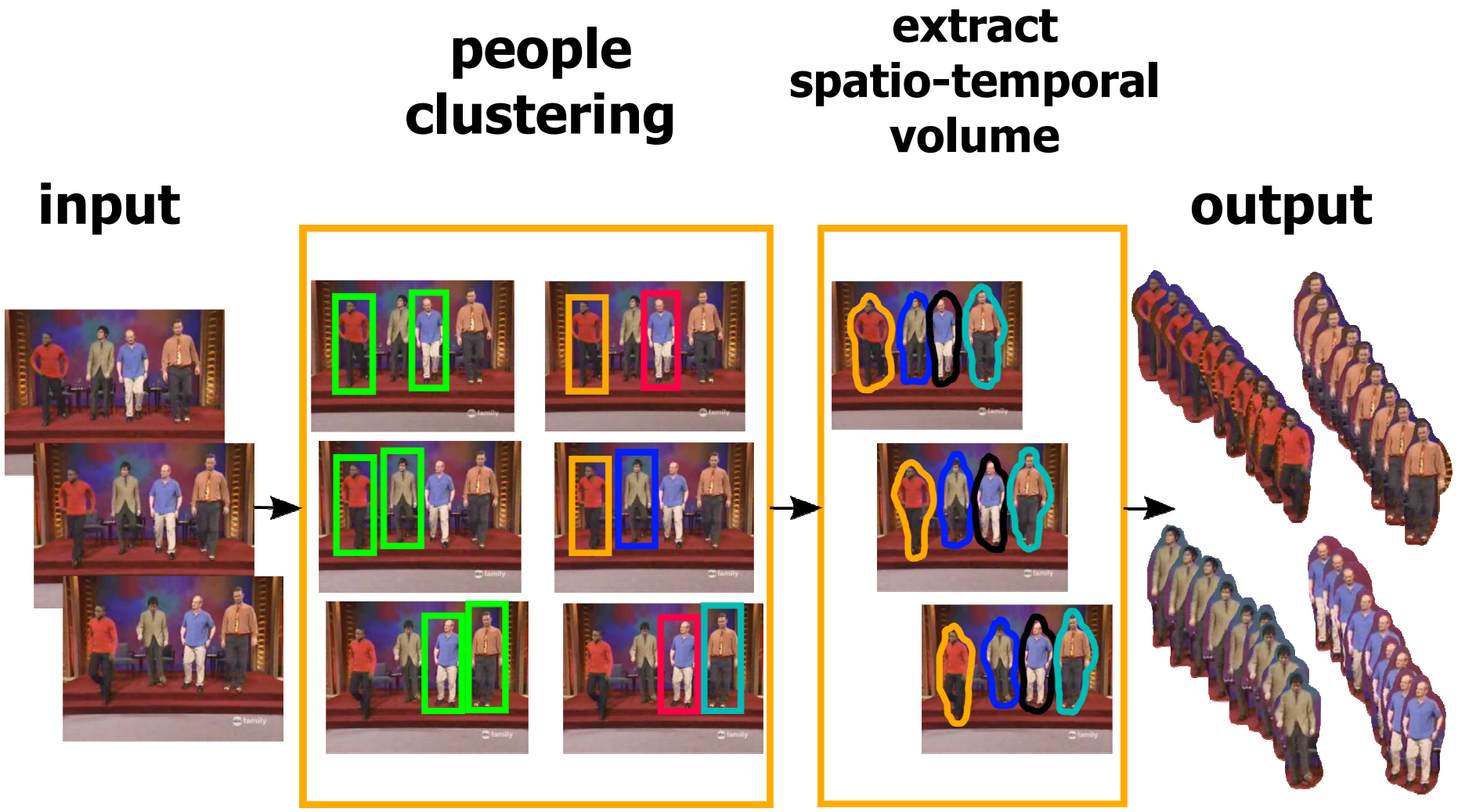
- Clustering with constraints [Klein et al ’02]
- More expensive descriptor:
  head & torso estimates
- Reject tracks
  Poor head torso estimates

Detection & clustering result



Output = One cluster per person in the sequence

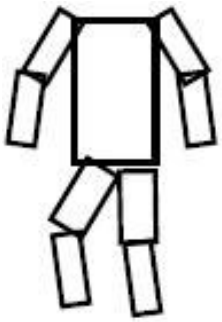
System Overview



extracting the human region in a single frame [Ramanan 06]

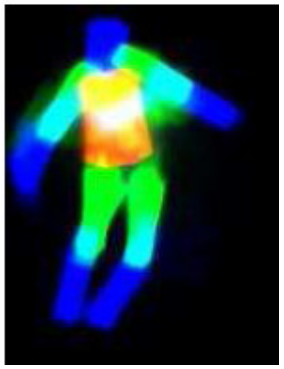
Original image

+



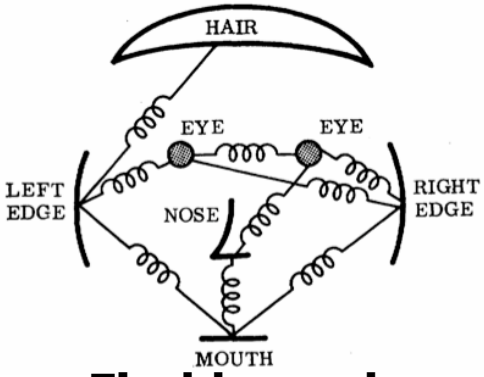
Top-down person model

→



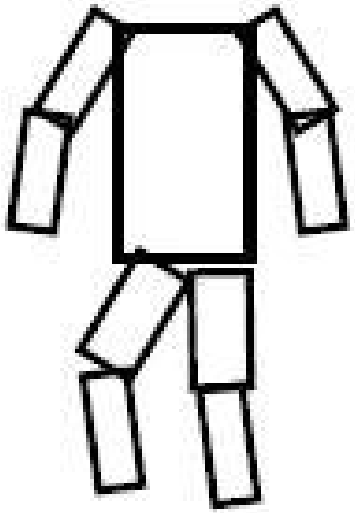
Estimated person region

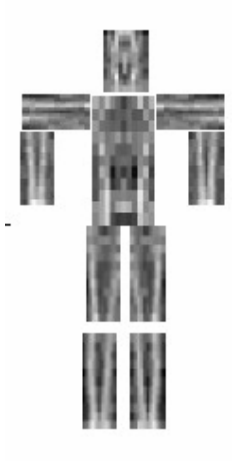
- Pictorial structure representation
  - each part represents local visual properties
  - 'springs' capture spatial relationships
  - stretch & fit to find right configuration



Fischler and Elschlager, 1973

images in this slide from: Deva Ramanan





Body part templates

- N-part model
- A configuration is given by $L = \{l_i\}; i = 1 \ldots N$, with $l_i = \{x_i, y_i, \theta_i\}$
- We are interested in

$$p(L|I,\theta) \propto p(I|L,\theta)p(L|\theta)$$

$$P(L|I,\theta) \propto \left(\prod_{i=1} p(I|l_i,u_i) \prod_{(v_i,v_j)\in E} p(l_i,l_j|c_{ij}) \right)$$

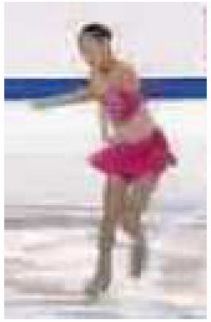
Image evidence (measurement) **shape prior**

- efficient inference
- measurement is the bottle-neck

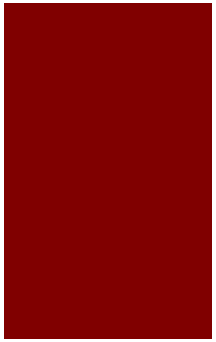
expensive convolutions with non-separable filters

reducing measurement computation

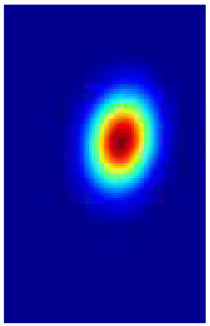
- Benefits
  - can be limited to a smaller space, reducing computation time
  - avoid distracting background observations
  - final estimation accuracy can be improved
- Use temporal consistency
  - Compute full model at first frame
  - Propagate part estimations to next frame



input image

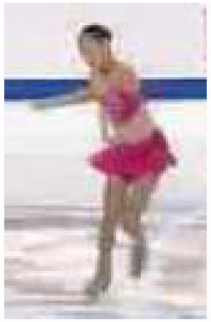


original search space for torso (full frame)



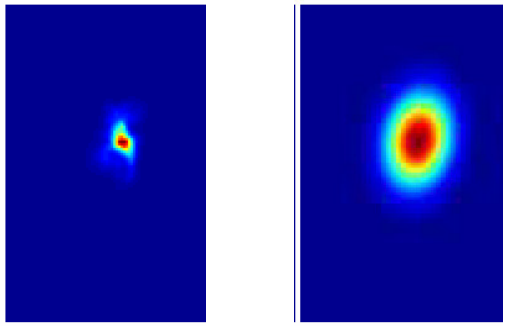
reduced search space for torso

reducing measurement computation



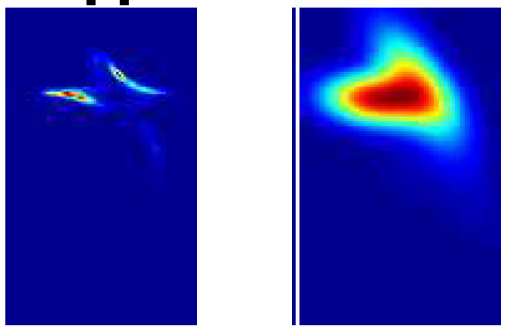
input image

torso



marginal **Diffused GMM approximation**

upper-left arm



marginal **Diffused GMM approximation**

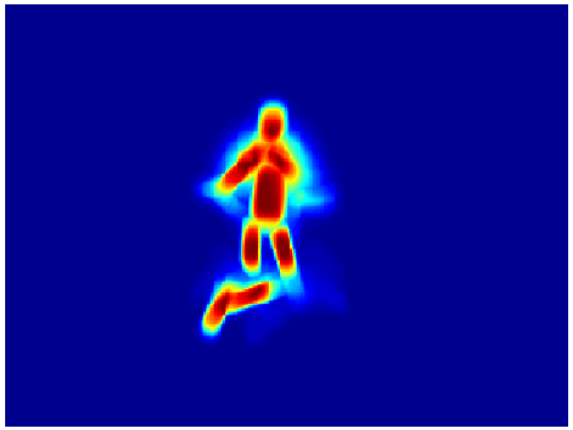
- GMM approximation is done via Kernel Density Approximation (KDA) [Han et al. 'PAMI08]
- Densities are propagated in a Bayesian filtering framework

$$p(\mathbf{X}_t \mid \mathbf{Z}_{1:t}) \propto p(\mathbf{Z}_t \mid \mathbf{X}_t) p(\mathbf{X}_t \mid \mathbf{Z}_{1:t-1})$$

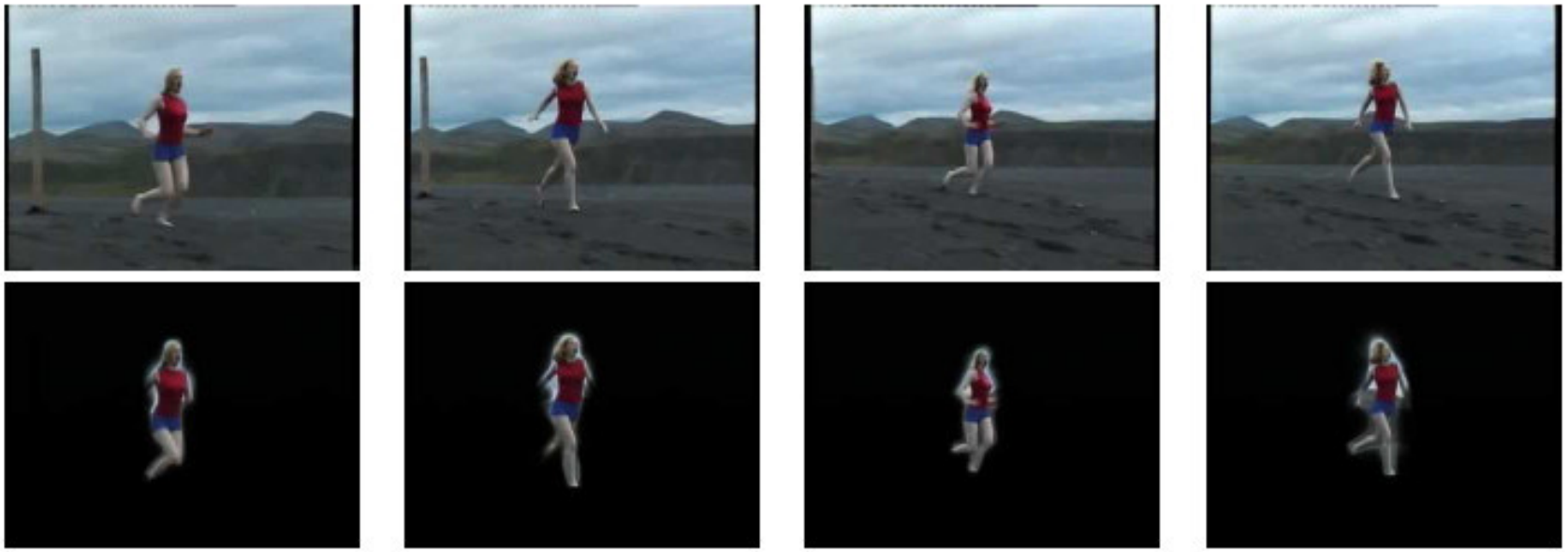
$$= \left(\sum_{i=1}^{N_1} \mathcal{N}(\kappa_i, \mathbf{x}_i, \mathbf{P}_i) \right) \left(\sum_{j=1}^{N_2} \mathcal{N}(\tau_j, \mathbf{y}_j, \mathbf{Q}_j) \right)$$

extracting the human region

Projecting the part posteriors into the image gives a rough segmentation of the body region

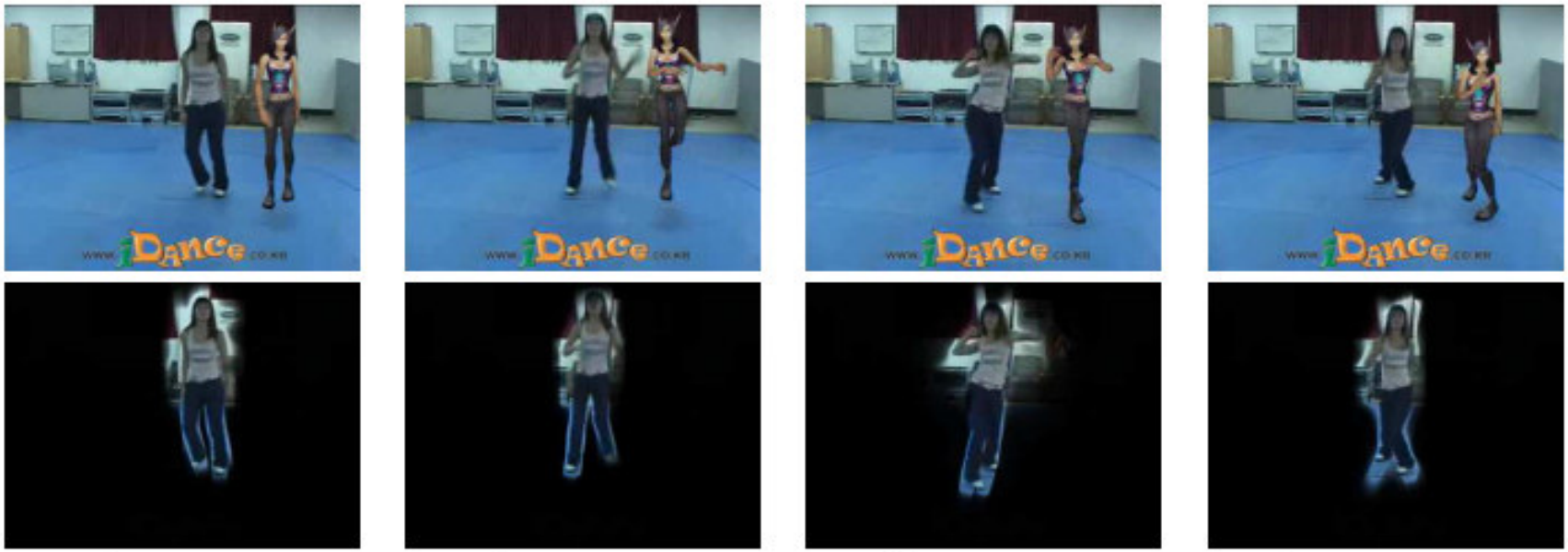




results



frame 1 frame 7 frame 18 frame 24

results



frame 1 frame 10 frame 60 frame 136

results

frame 11 frame 74 frame 86 frame 300

experiments

Precision Recall Comparison

	Detection only			Detection & Clustering			Full model		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
Rate	0.89	0.31	0.46	0.89	0.30	0.45	0.83	0.73	**0.78**
	0.90	0.25	0.39	0.91	0.24	0.38	0.87	0.62	**0.72**
	0.92	0.19	0.32	0.92	0.19	0.32	0.86	0.51	**0.64**
	0.93	0.16	0.27	0.94	0.15	0.27	0.92	0.43	**0.58**
	0.94	0.13	0.24	0.94	0.13	0.23	0.88	0.32	**0.46**

Computation Time

~ 1 order of magnitude speed up

more results

references

- Juan Carlos Niebles, Bohyung Han, Andras Ferencz and Li Fei-Fei. Extracting Moving People from Internet Videos. **ECCV** 2008

- With:

Bohyung Han

Andras Ferencz

Conclusions

- Spatio-temporal & spatial features + statistical classifier
  - Bag of words (unsupervised)
  - Constellation of bags of features

- Human Motion extraction
  - Real world sequences from YouTube

Interesting Research Issues

- Motion taxonomy/vocabulary
- higher level activities
- …

In this talk

- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences

Thank you