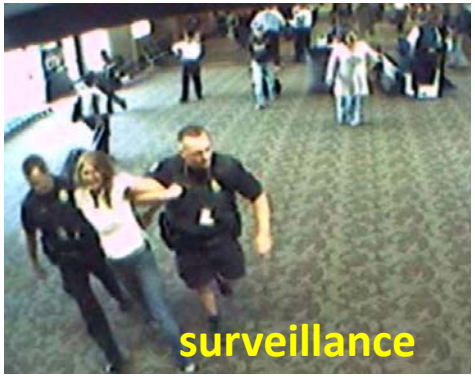
Lecture 18: Human Motion Recognition

Professor Fei-Fei Li

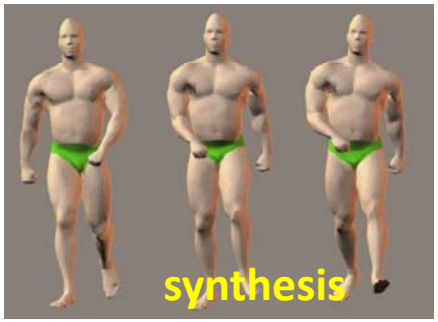
Stanford Vision Lab

What we will learn today?

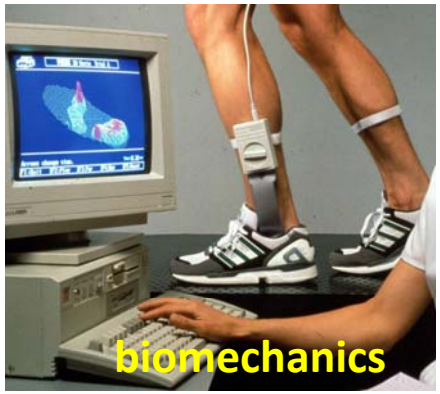
- Introduction
- Motion classification using template matching
- Motion classification using spatio-temporal features
- Motion classification using pose estimation



search



Human Motion Analysis



It's challenging…

Variation in Appearance

Variation in Pose

Variation in view-point

Occlusion & clutter

Adapted from http://luthuli.cs.uiuc.edu/~daf/tutorial.html

It's challenging…

- No clear action/movement classes
  - Unless in specific domain
- At different temporal spans, motion may acquire different meanings
- Interactions influence the meaning
  - With objects
  - With people
- Goals & intentions

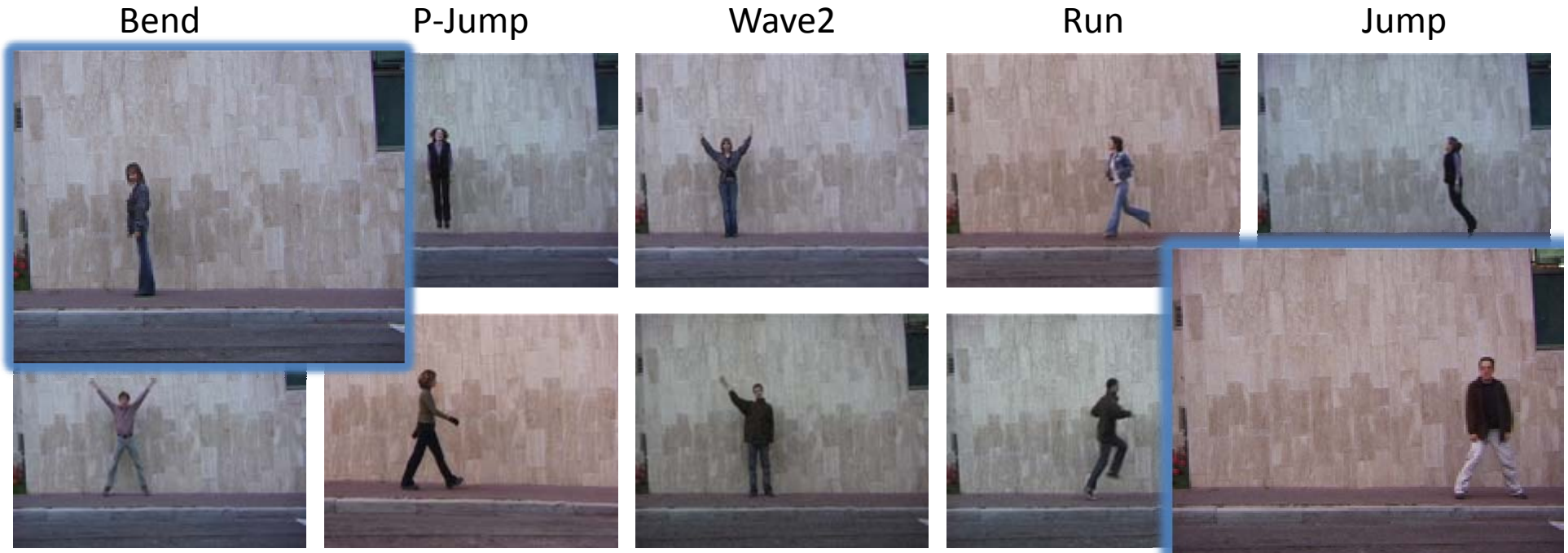
Current motion technology

- Mostly action discrimination
  - Walking or jumping?
- Mostly in simple scenarios
- We don't have a winner feature/representation yet

KTH Dataset

[Schuldt et al ICPR 04]

Weizmann Dataset



[Blank et al ICCV 2005]

UCF Sports

Lifting

Golf-swing

13 action classes

- Diving
- Golf-swing
- Kicking
- Walk
- Lifting
- Riding-Horse
- Run
- …

[Rodriguez et al CVPR 2008]

Hollywood Human Actions

12 Action Classes

- AnswerPhone
- Eat
- FightPerson
- GetOutCar
- HandShake
- Kiss
- Run
- SitDown
- SitUp
- StandUp

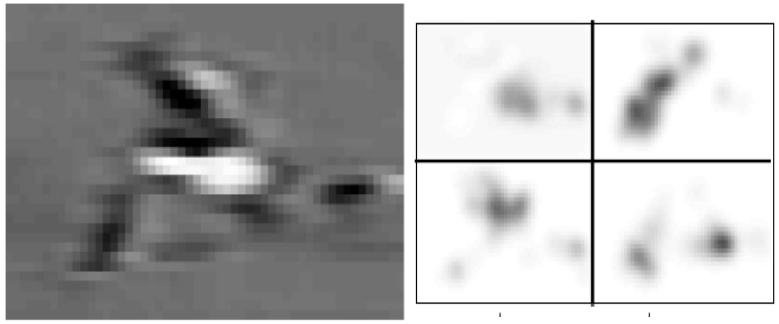





[Laptev et al CVPR08/CVPR09]

Three representations

- **Global template**



- Local video patches

- Kinematic/body pose

Fields of view

Far field

- 3-pixel man
- blob tracking

Near field

- 300-pixel man
- limb tracking

Medium Field

- 30-pixel man
- motions

The Big Picture

- Goal: Action classification
  - E.g. walk left, walk right, run left, run right, etc
- Steps:
  1. Tracking person in video
  2. Feature extraction
  3. Classification
     - Nearest Neighbor
     - AdaBoost

How To Gather Action Data

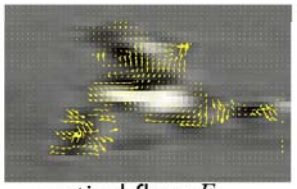
A person in a particular body configuration should always map to approximately the same stabilized image

-> We lose all translational information but optical flow features allow us to distinguish between running and walking.

- Tracking
  - Simple correlation-based tracker
  - User-initialized
- Forms a spatio-temporal volume for each person

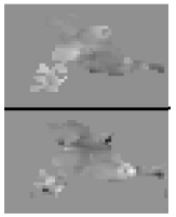
Low-level Motion Features

1. Compute Optical Flow (Lucas and Kanade algorithm).



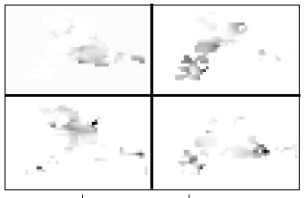
optical flow $F_{x,y}$

2. Optical flow vector field F is split into horizontal and vertical components (F_x, F_y).



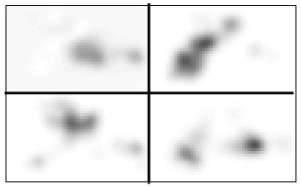
F_x, F_y

3. F_x and F_y are split into four non-negative channels F_{x+}, F_{x-} , F_{y+}, F_{y-}.



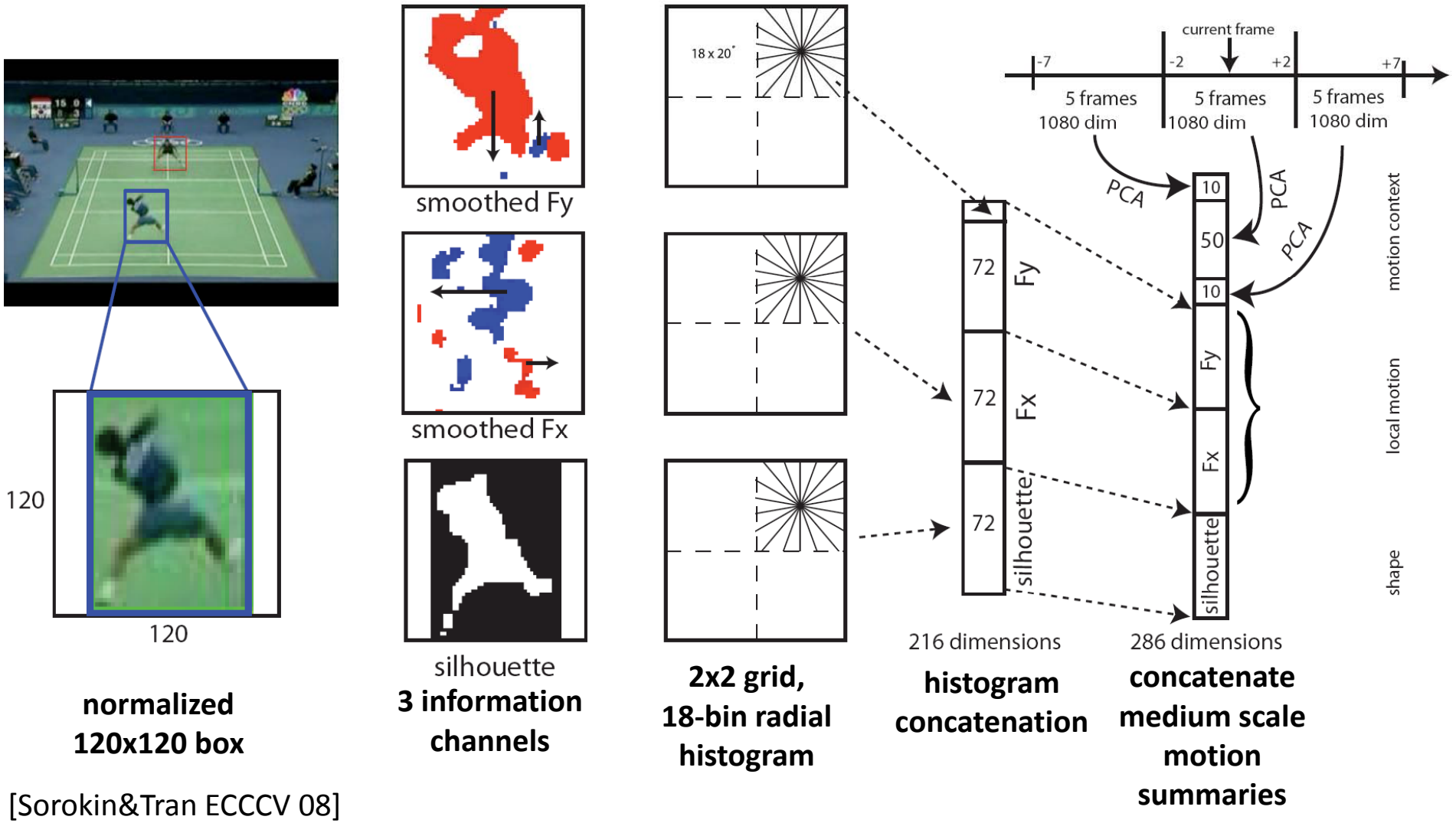
$F_x^+, F_x^-, F_y^+, F_y^-$

4. Each channel is blurred to reduce noise.



$Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

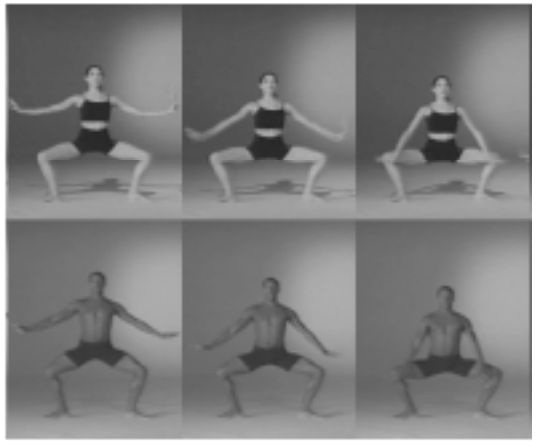
Motion Context Descriptor: A Feature that Combines Shape and Motion



[Sorokin&Tran ECCCV 08]

Classifying Actions

- Nearest Neighbors [Efros et al ICCV 2003]
- Boosting [Fathi & Mori CVPR 2008]
- Metric Learning [Tran & Sorokin ECCV 2008]



Nearest Neighbors

[Efros et al, ICCV 2003]

- Assume we have template videos containing examples of actions we want to classify
  - We form motion descriptors (i.e. feature vectors) for both the train and test data
- Find the k-nearest neighbors and use majority voting to identify the action class
- The motion descriptors are constructed to be discriminative so a simple classification algorithm is sufficient.

Nearest Neighbor Results

[Efros et al, ICCV 2003]

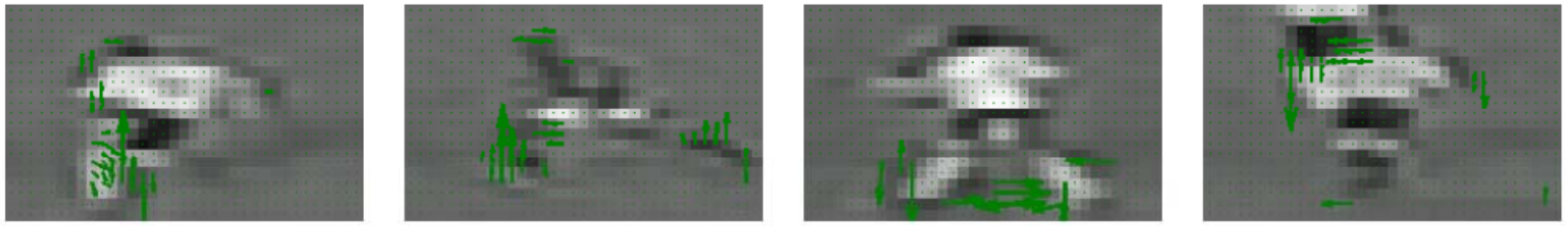
Input video: **Nearest neighbor match:**

Action Classification Result

Boosting

[Fathi and Mori CVPR 2008]

- Low-level features alone are not capable of discriminating between positive and negative classes.
  - They are weak-classifiers
  - Combine them via AdaBoost
- Mid-level features are obtained via AdaBoost from low-level features.



(Examples of low-level features selected for mid-level features.)

- Final classifier obtained via Adaboost from mid-level features.

Boosting

[Fathi and Mori CVPR 2008]

- Mid-level motion features

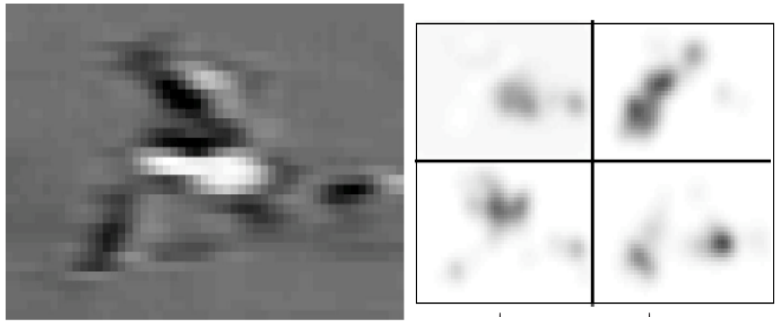
$$H_i(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t^i h_t^i(x) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- Final classifier

$$C(s) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t g_t(s) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Three representations

- Global template



- **Local video patches**

- Kinematic/body pose

Big Picture: Local Video Patches

- Similar to SIFT, we would like to identify key points in videos
  - Space-time Interest Points
- As done in bag-of-words approach, we can combine the 'Space-time Interest Points' to form a feature vector
  - Bag-of-space-time features
- Allows the use of a discriminative classifier (e.g. SVM)

Motivation

Sparse and *Local* representation

Spatio-temporal information

Johansson, 1973

Local Video Patches

No global assumptions $\Rightarrow$

Consider local spatio-temporal neighborhoods

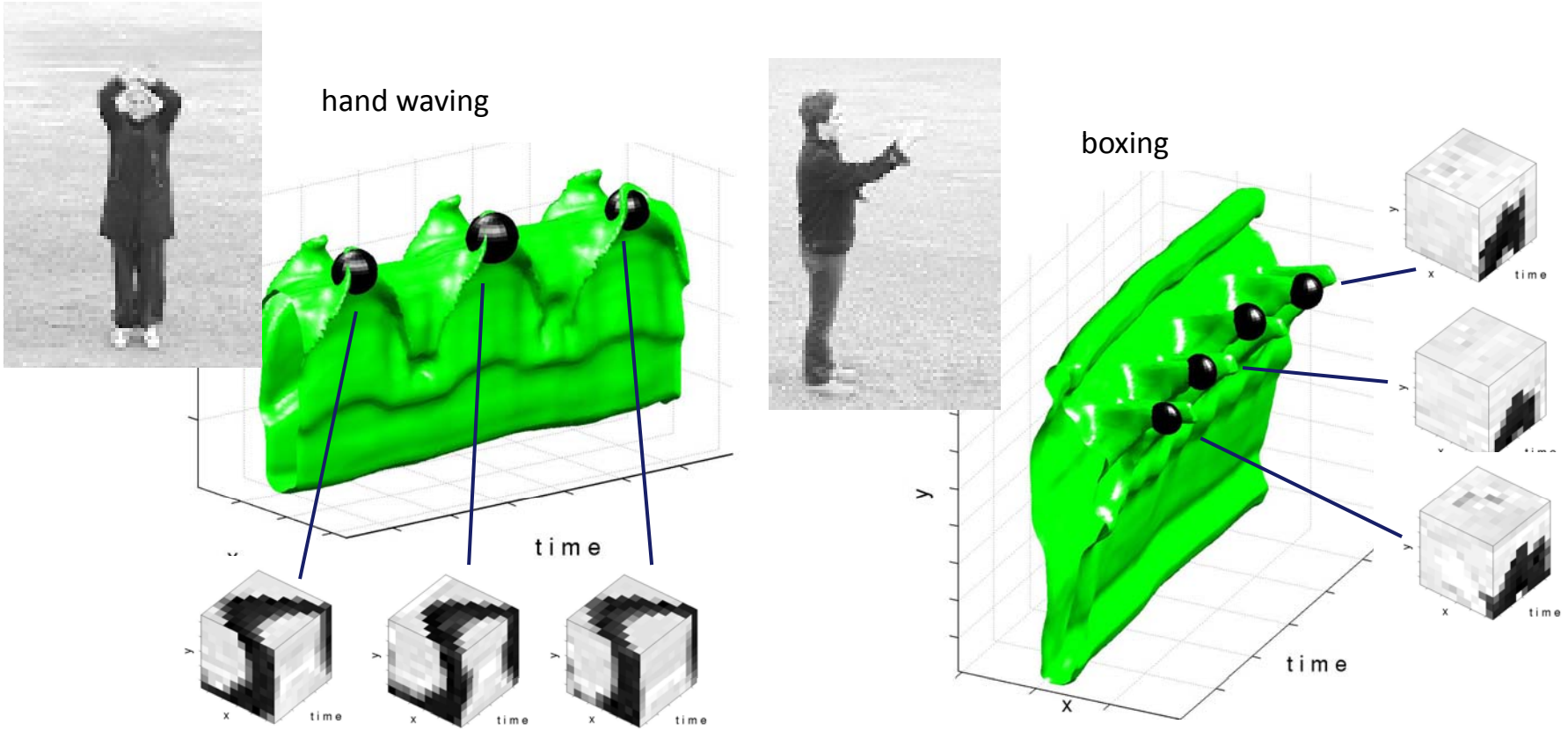
hand waving

boxing

Local Video Patches

No global assumptions $\Rightarrow$

Consider local spatio-temporal neighborhoods



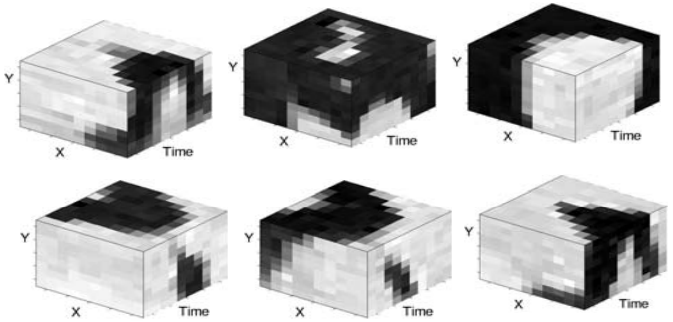
Local Video Patches

Bag of space-time features + Statistical classifier -> Action Classification

[Schuldt'04, Niebles'06, Zhang'07, Laptev'08,'09]

Space-time interest point detection

Collection of space-time patches

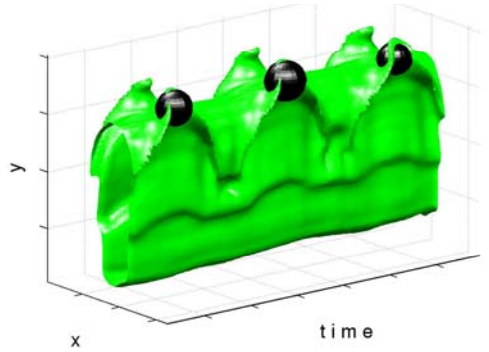


Video Feature Representation

Classifier

Space-Time Features: Detector

- Extend the idea of 2D interest points to 3D interest points in space and time.
- Build on the idea of Harris and Förstner interest point operators
- We want to find regions with large variation in both space and time domains



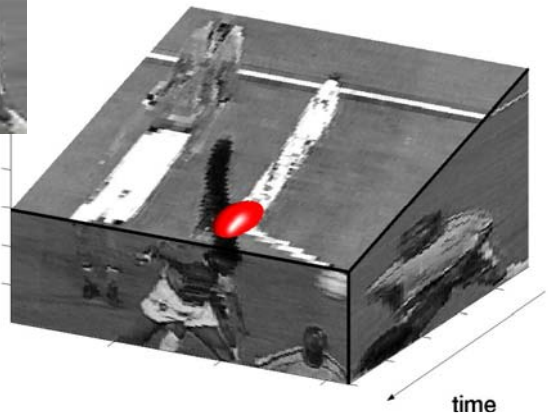
Space-Time Features: Detector

- Space-time corner detector

  [Laptev, IJCV 2005]

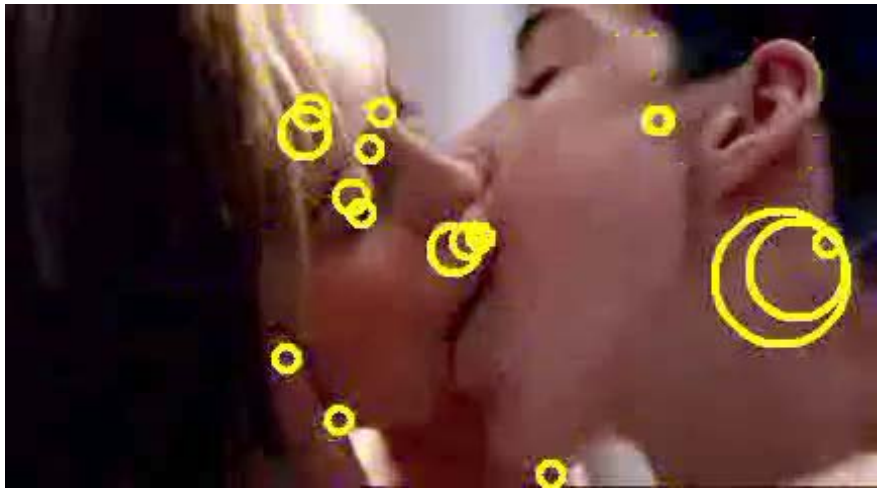
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$



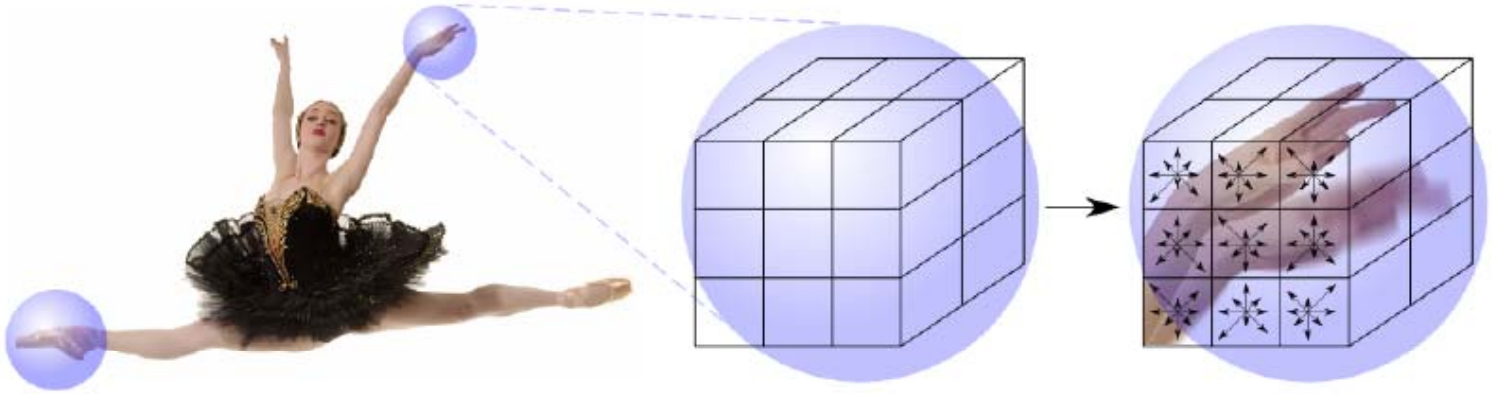
- Dense scale sampling (no explicit scale selection)

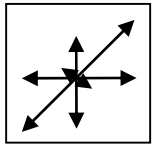
$$(\sigma^2, \tau^2) = \mathcal{S} \times \mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}}, \mathcal{T} = 2^{\{1,2\}}$$



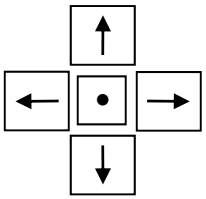
Space-Time Features: Detector

Feature extraction given interest point locations:





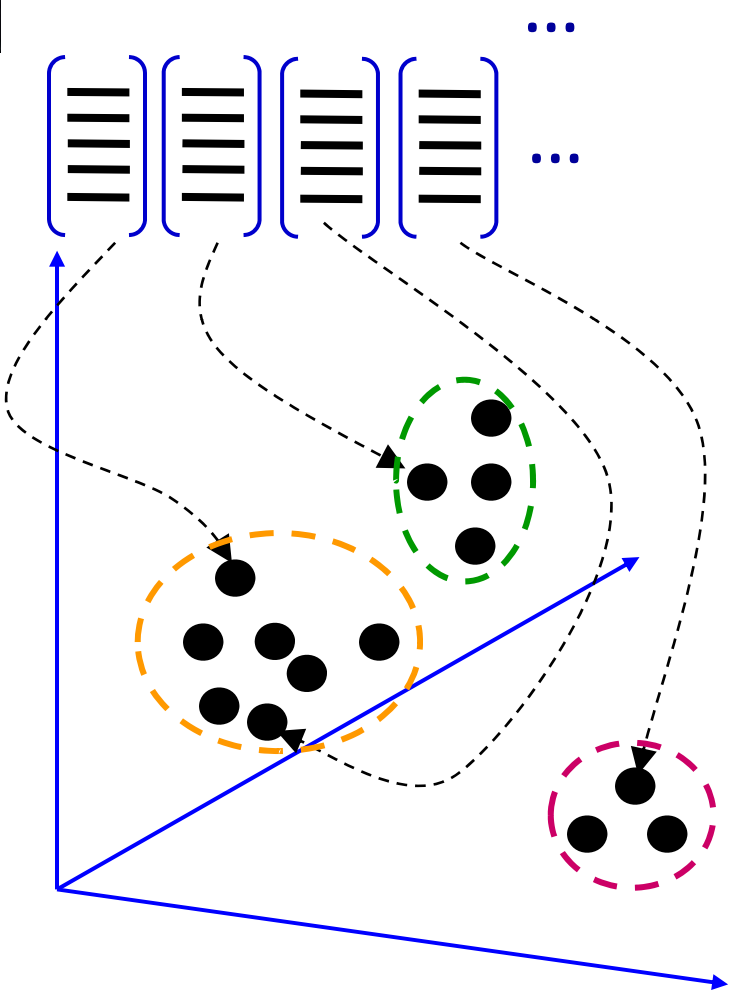
Histogram of oriented gradient (HOG)



Histogram of oriented flow (HOF)

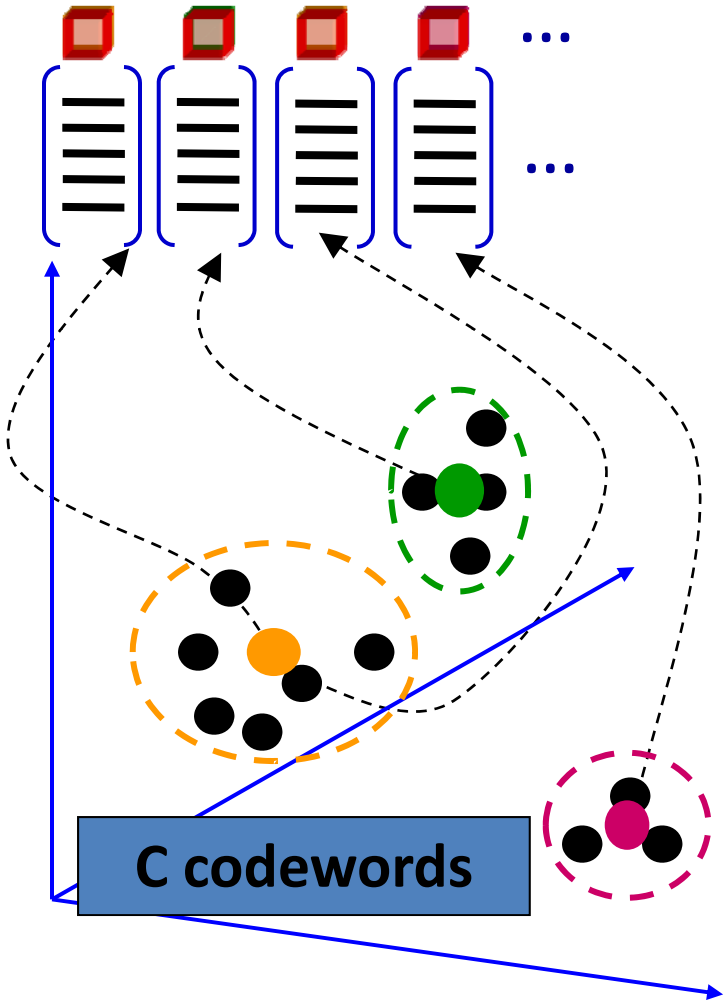
Codebook and Representation

Training Videos



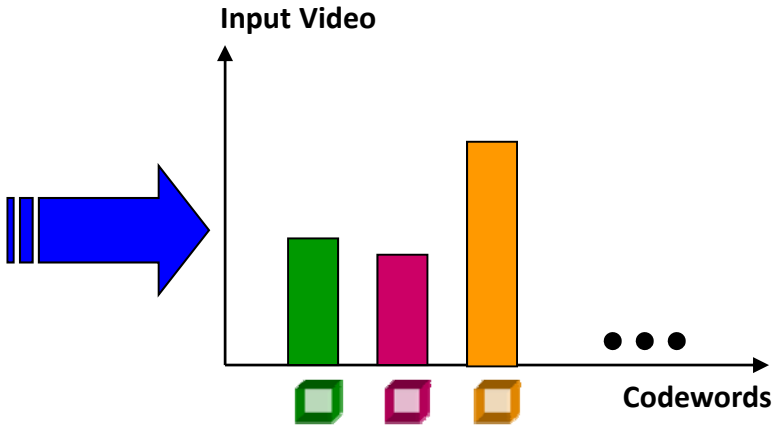
Codebook and Representation

Training Videos



Feature Space

Bag of features!

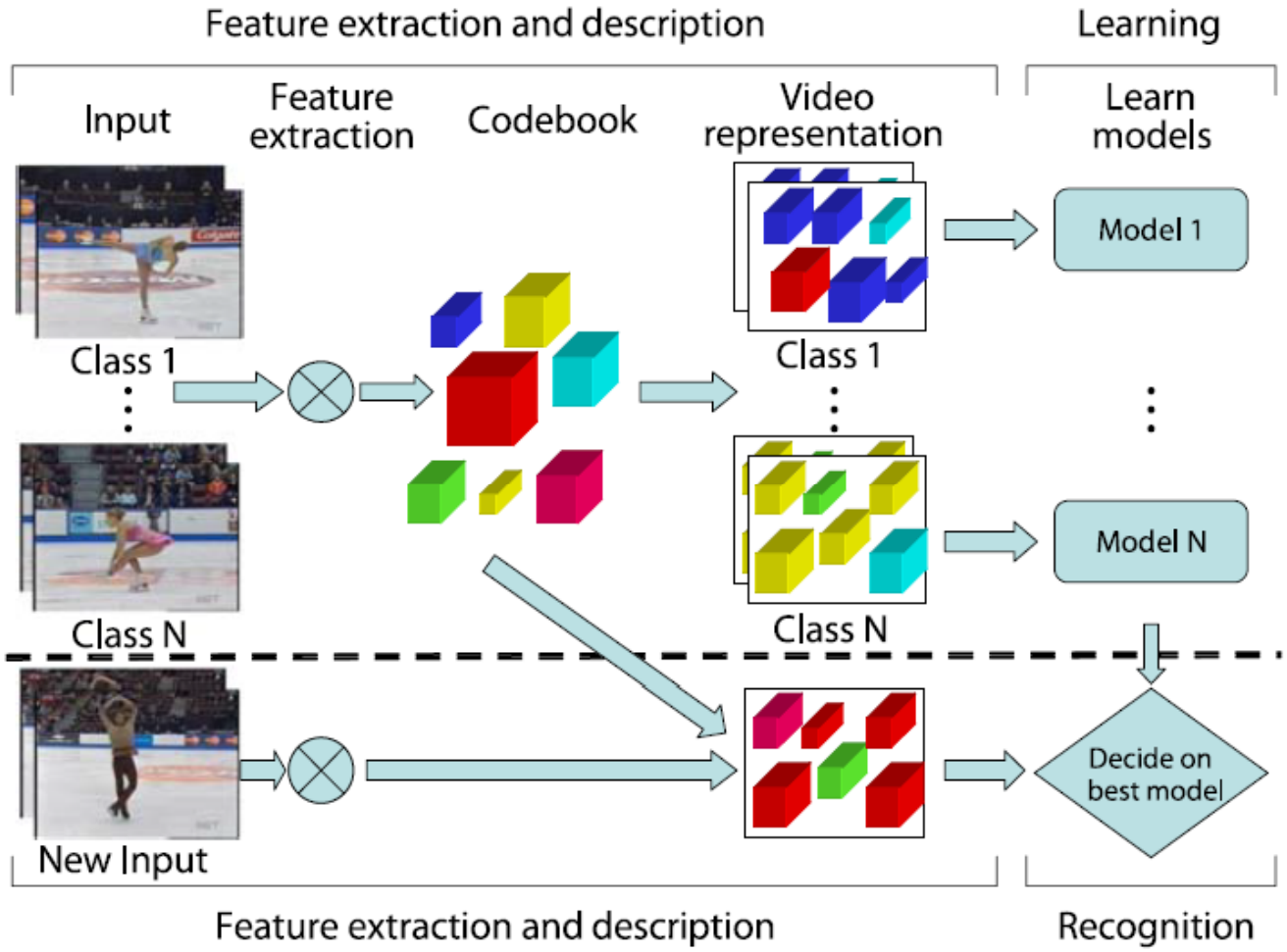


Classifying Actions

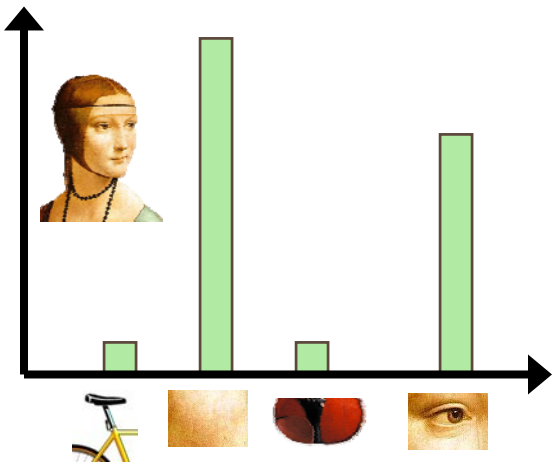
- SVM [Schuldt et al ICPR04, Laptev CVPR09]
- Topic Models (pLSA, LDA) [Niebles et al IJCV 08]
- …

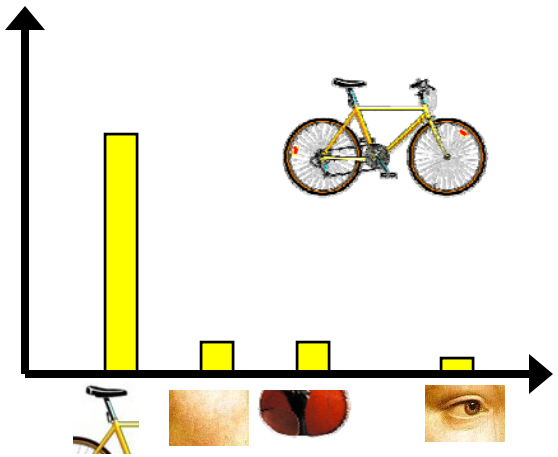
Topic Models (pLSA, LDA)

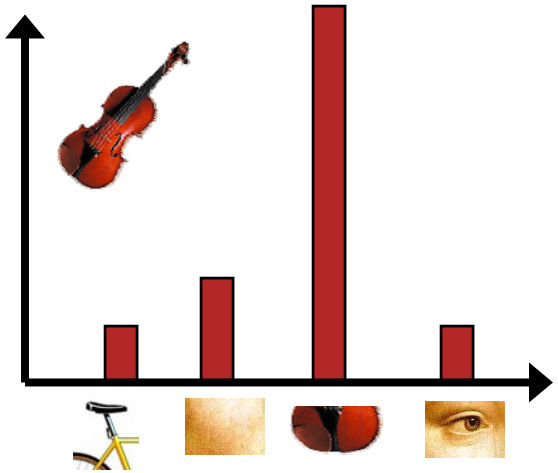
[Niebles et al IJCV 08]



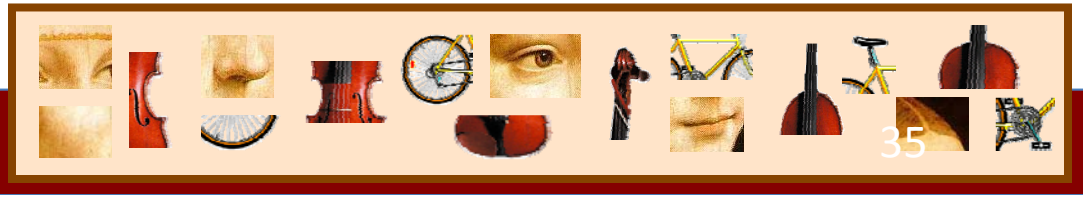
Object → **Bag of 'words'**



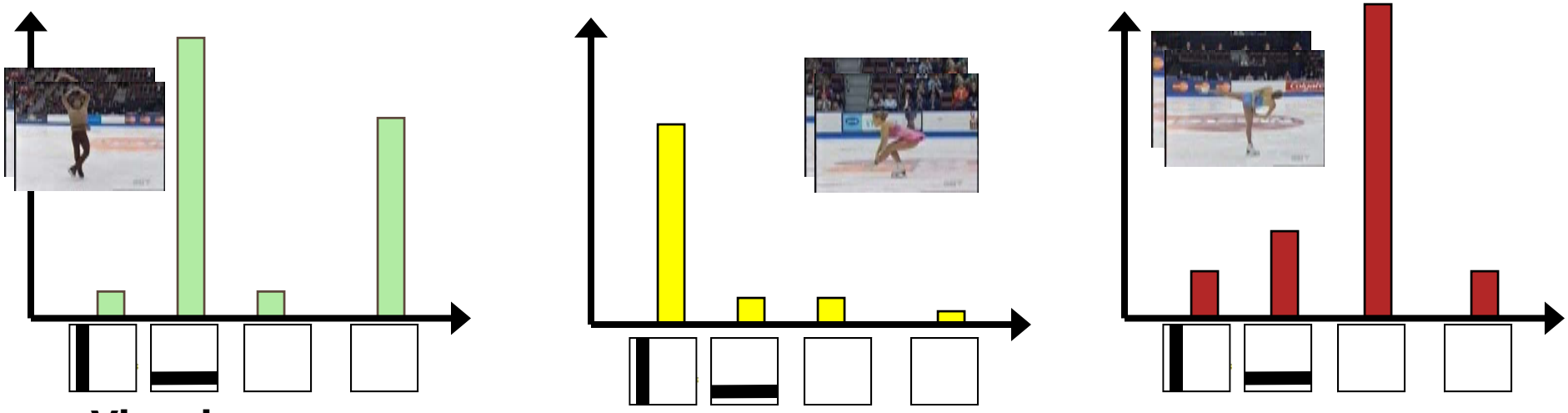




Visual dictionary

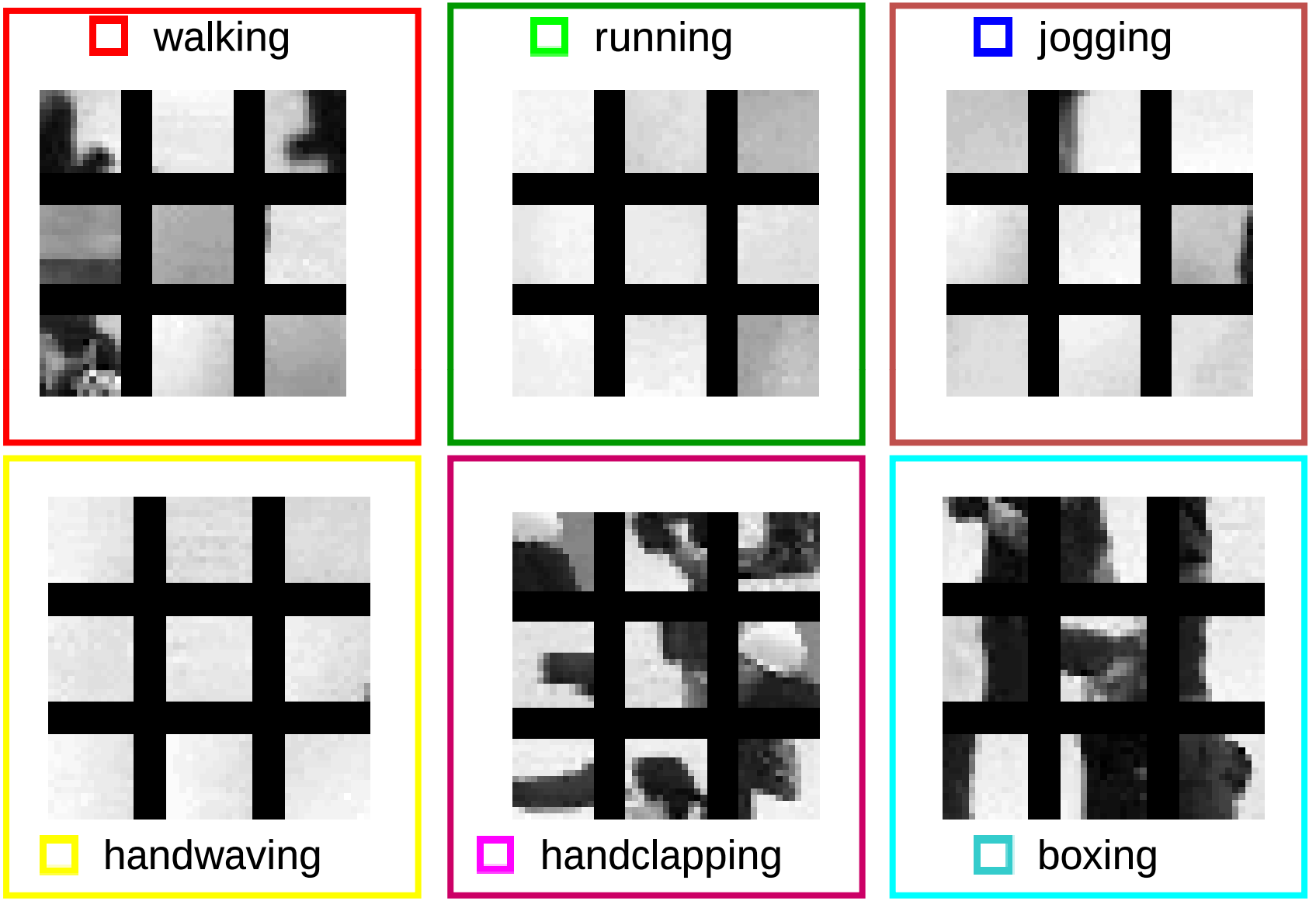


Action → **Bag of motion 'words'**

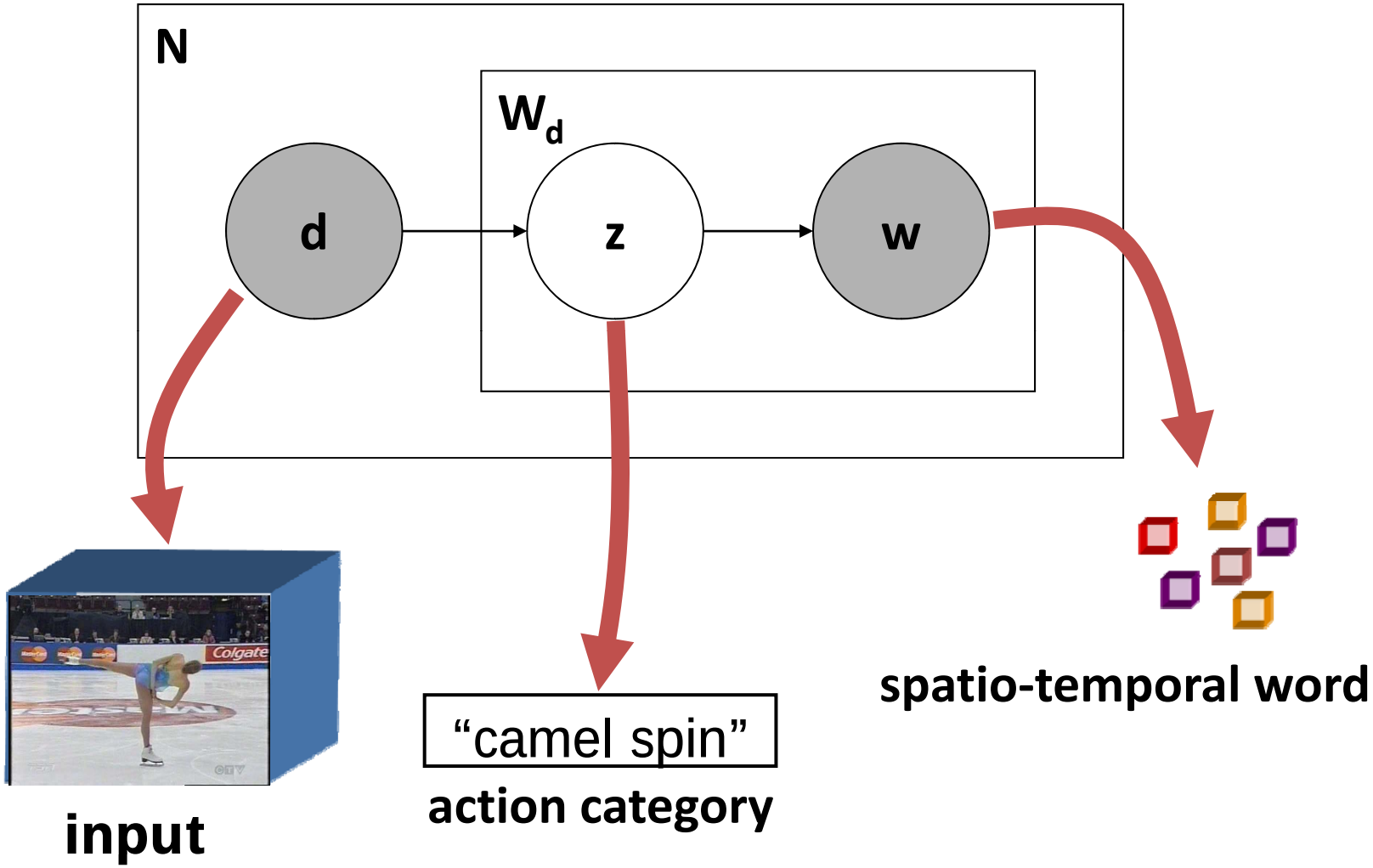


Visual dictionary

Codebook

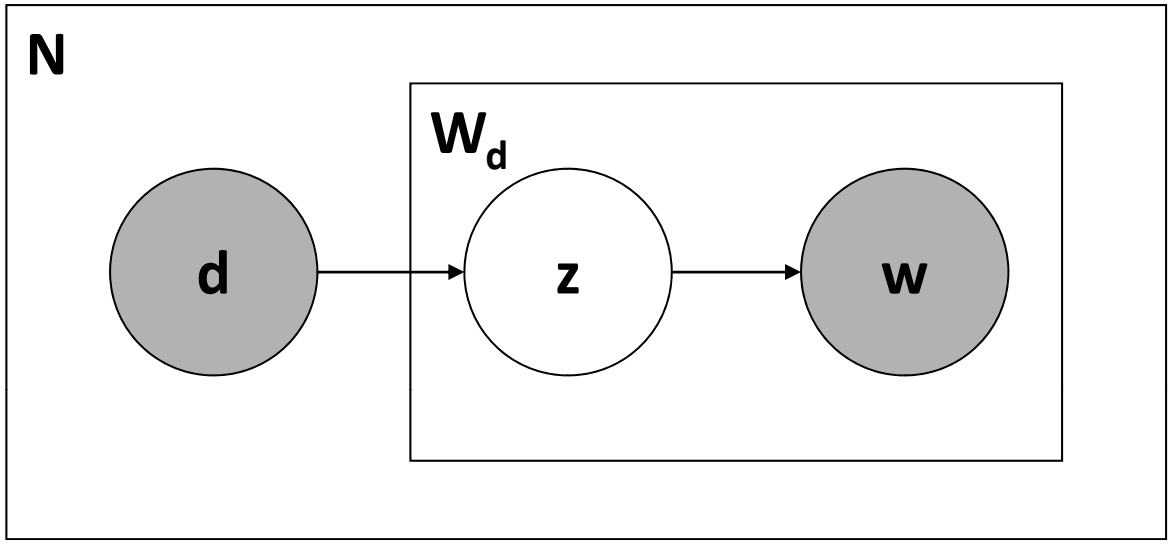


Unsupervised learning using pLSA



J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

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

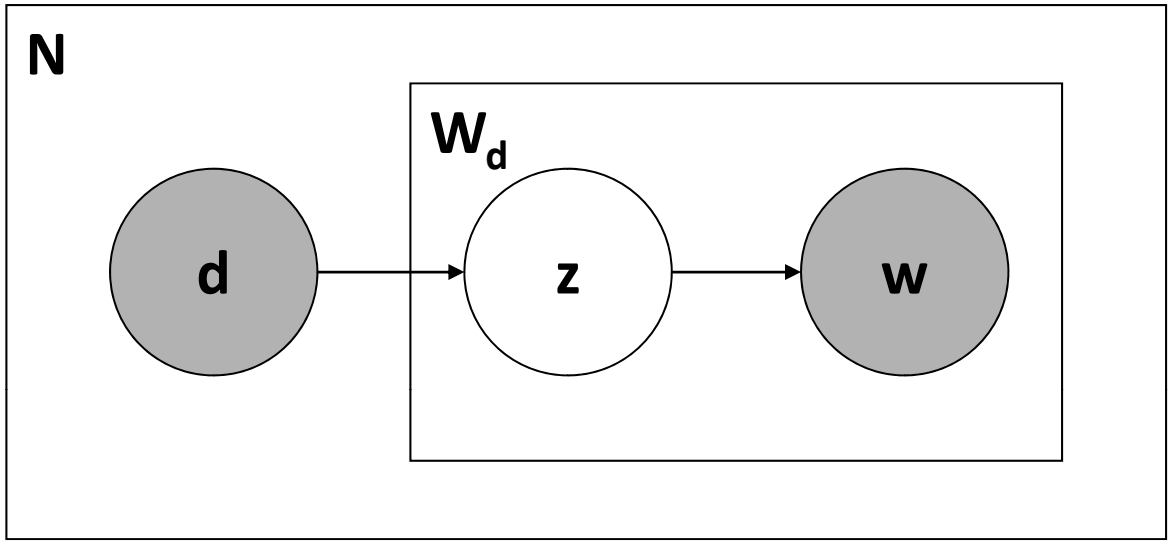
Word distribution per action category

action category weights

Action category distribution per video

J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

pLSA Model

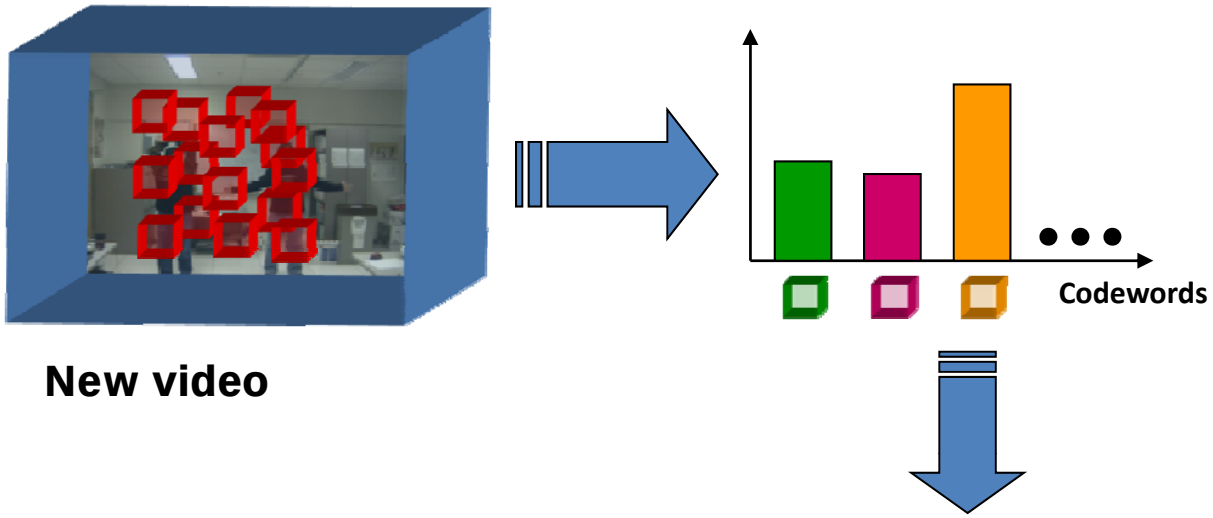


$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k)\, p(z_k \mid d_j)$$

Unsupervised Learning

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

Recognition



Action category distribution given test video

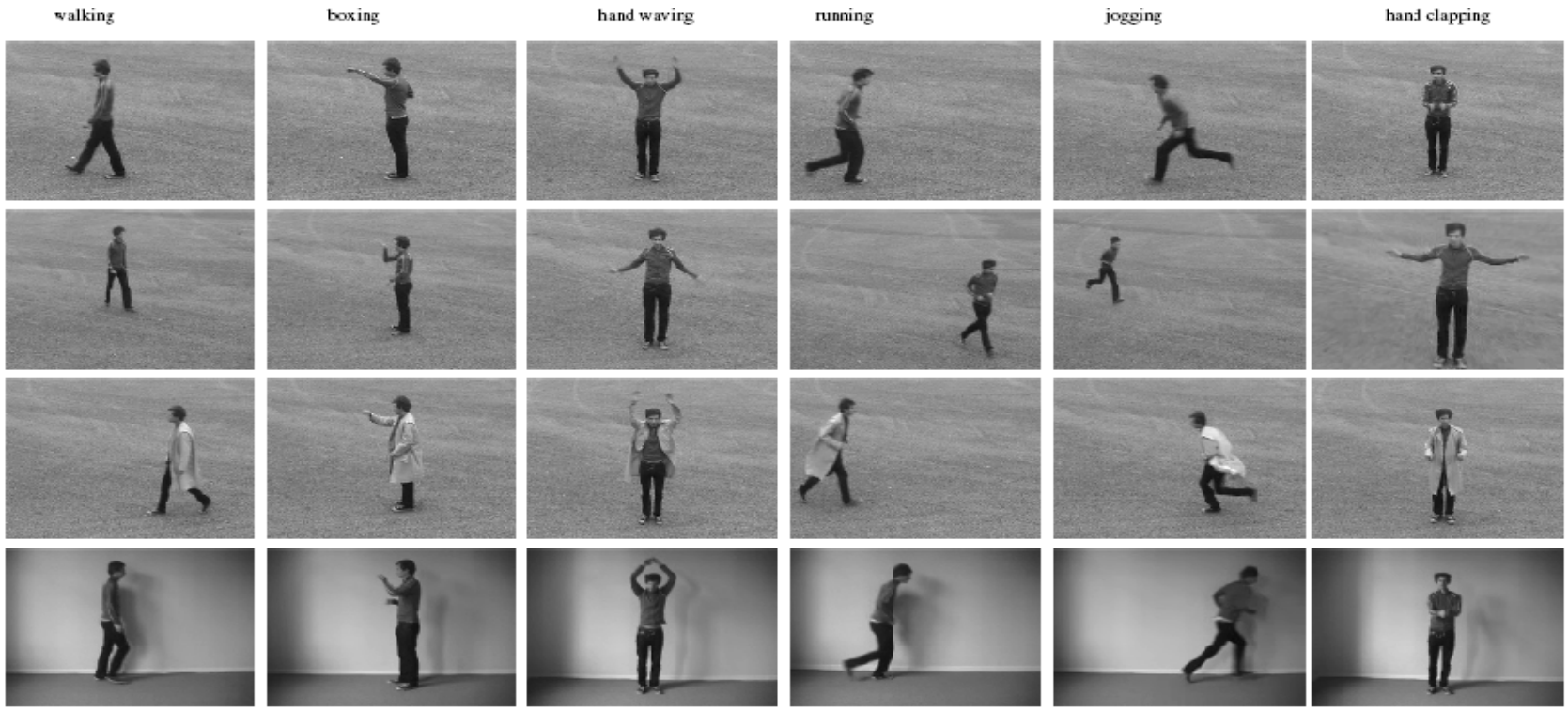
Word distribution given action category

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

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

Experiment I:

KTH dataset
[Schuldt et al., 2004]:



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

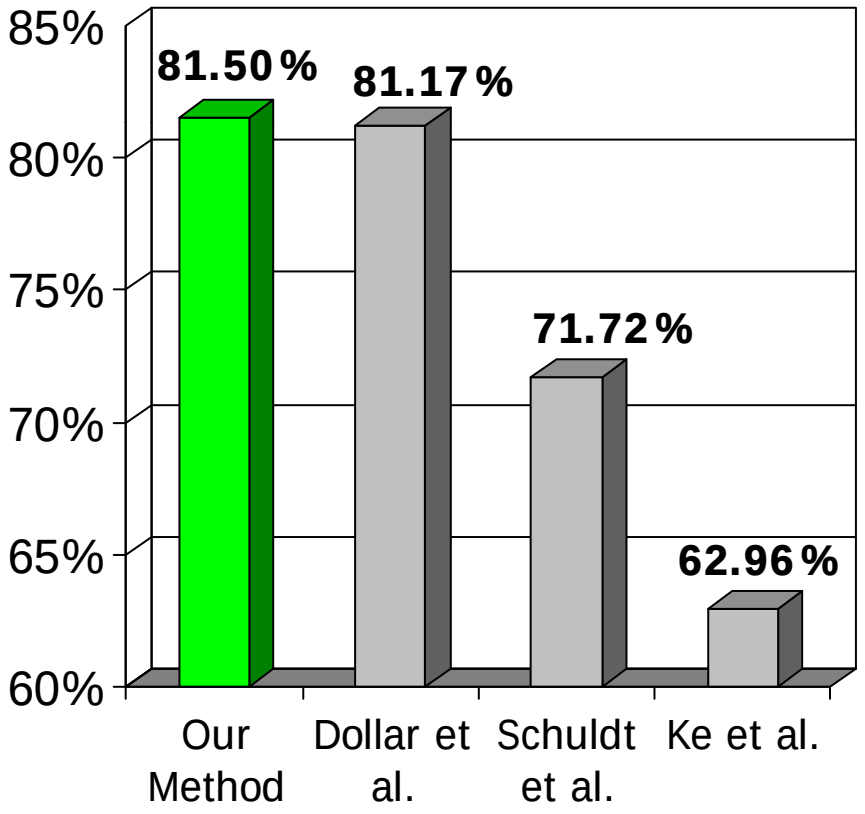
J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

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

J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

Experiment I: A longer sequence

walking

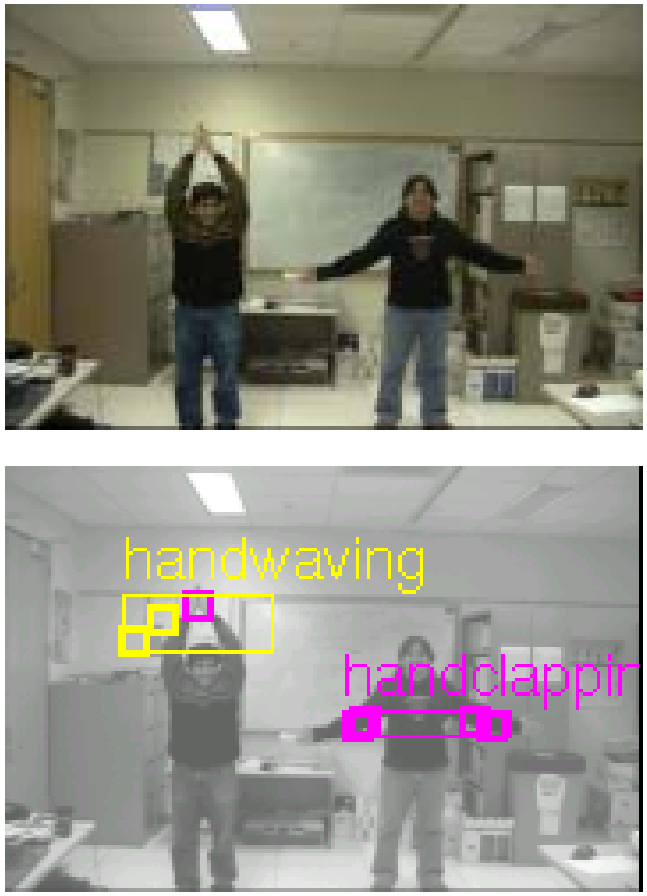
running

Trained with the KTH data

Tested with our own data

J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

Experiment I: Multiple motions



handclapping

handwaving

Trained with the KTH data

Tested with our own data

J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

Experiment II:

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

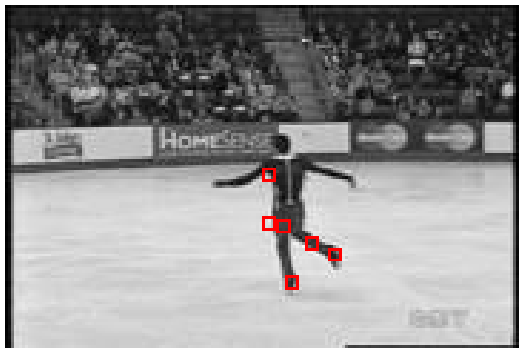
7 persons, 3 action classes: camel spin, stand spin, sit spin
J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

Experiment II: Examples

Figure skating actions

Camel spin

Sit spin



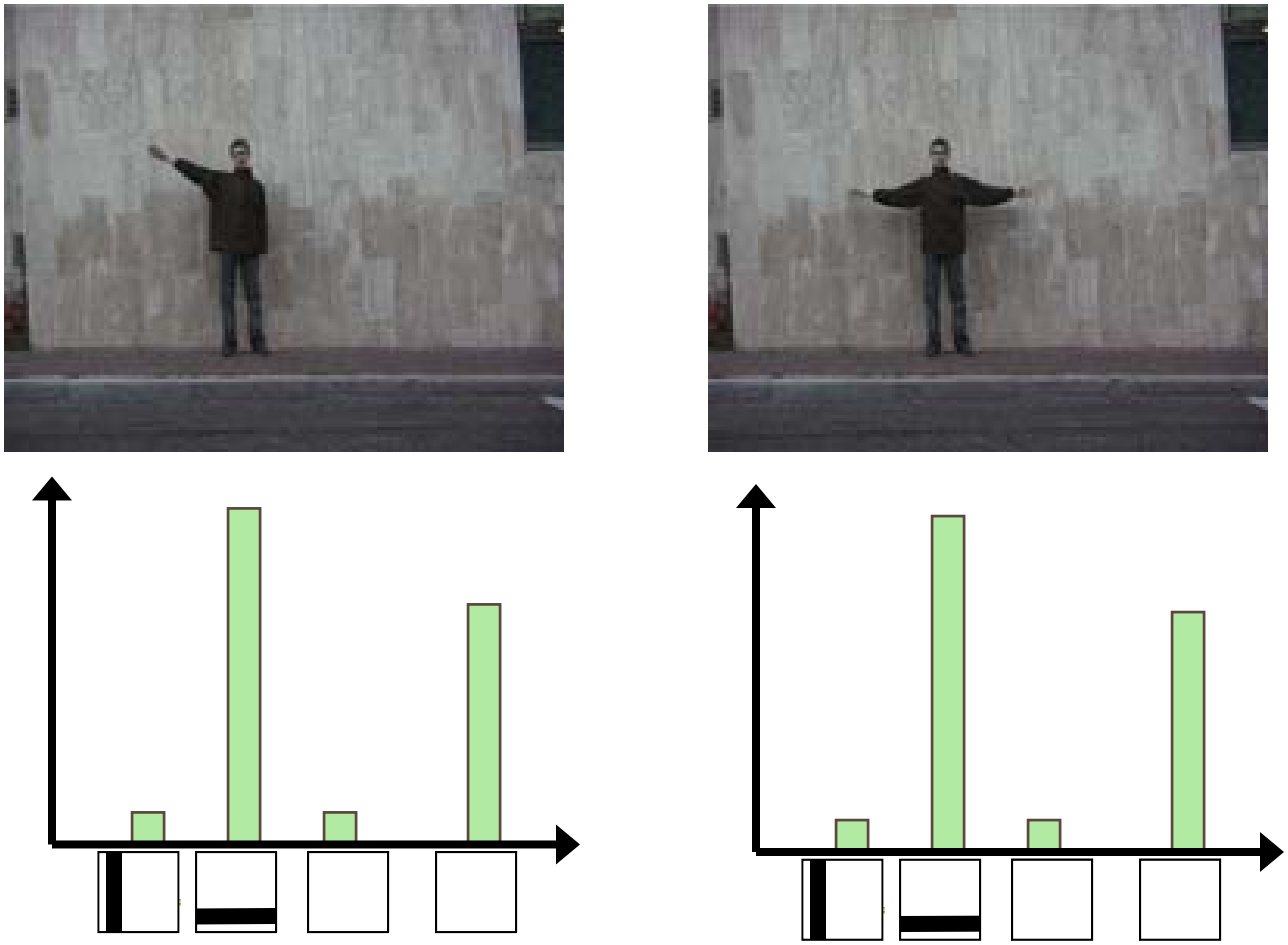
Stand spin

J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

Experiment II: Long Sequences

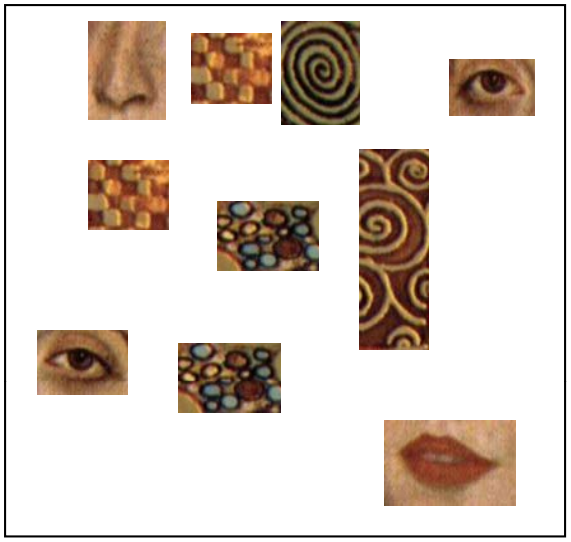
J.C. Niebles, H. Wang, L. Fei-Fei, BMVC, 2006

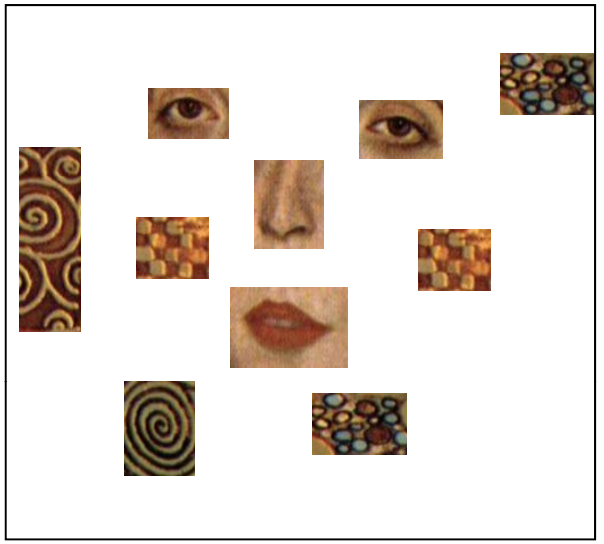
But…



- We need models that exploit geometrical arrangements of features

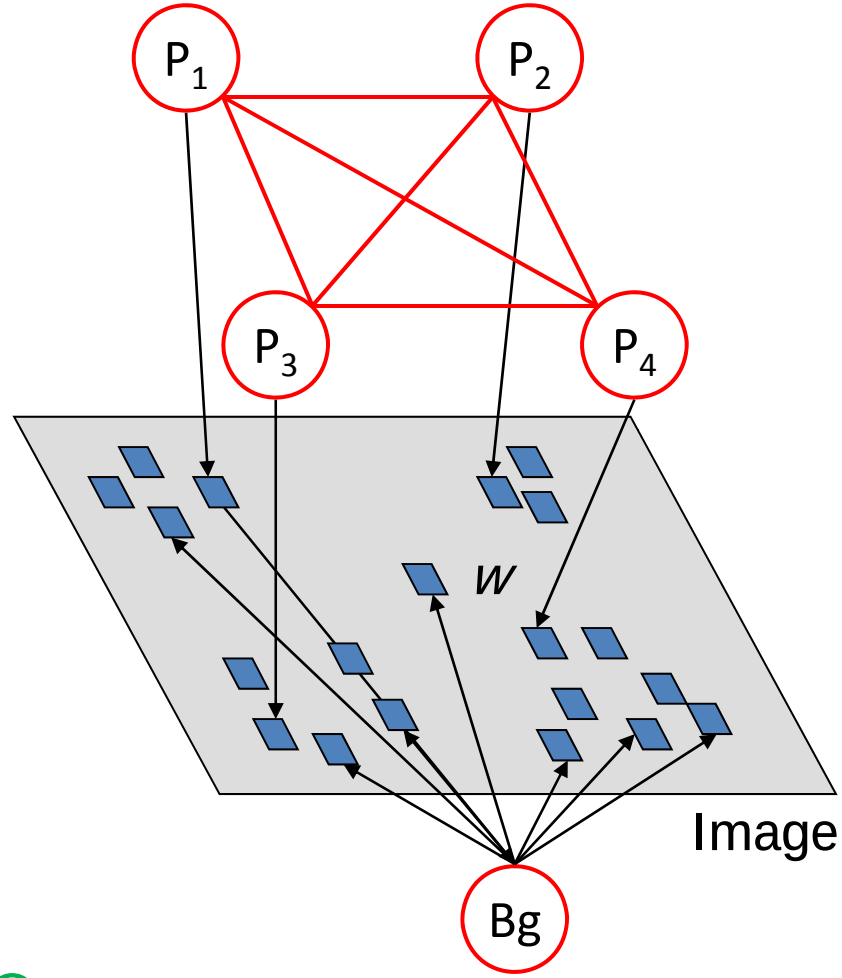
In the object recognition world





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

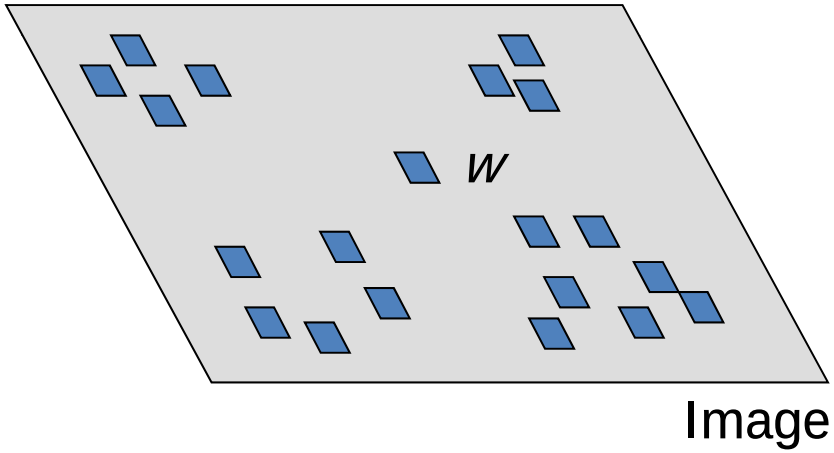
Constellation Model



☺ Strong shape representation

☹ Small number of features

bags of features



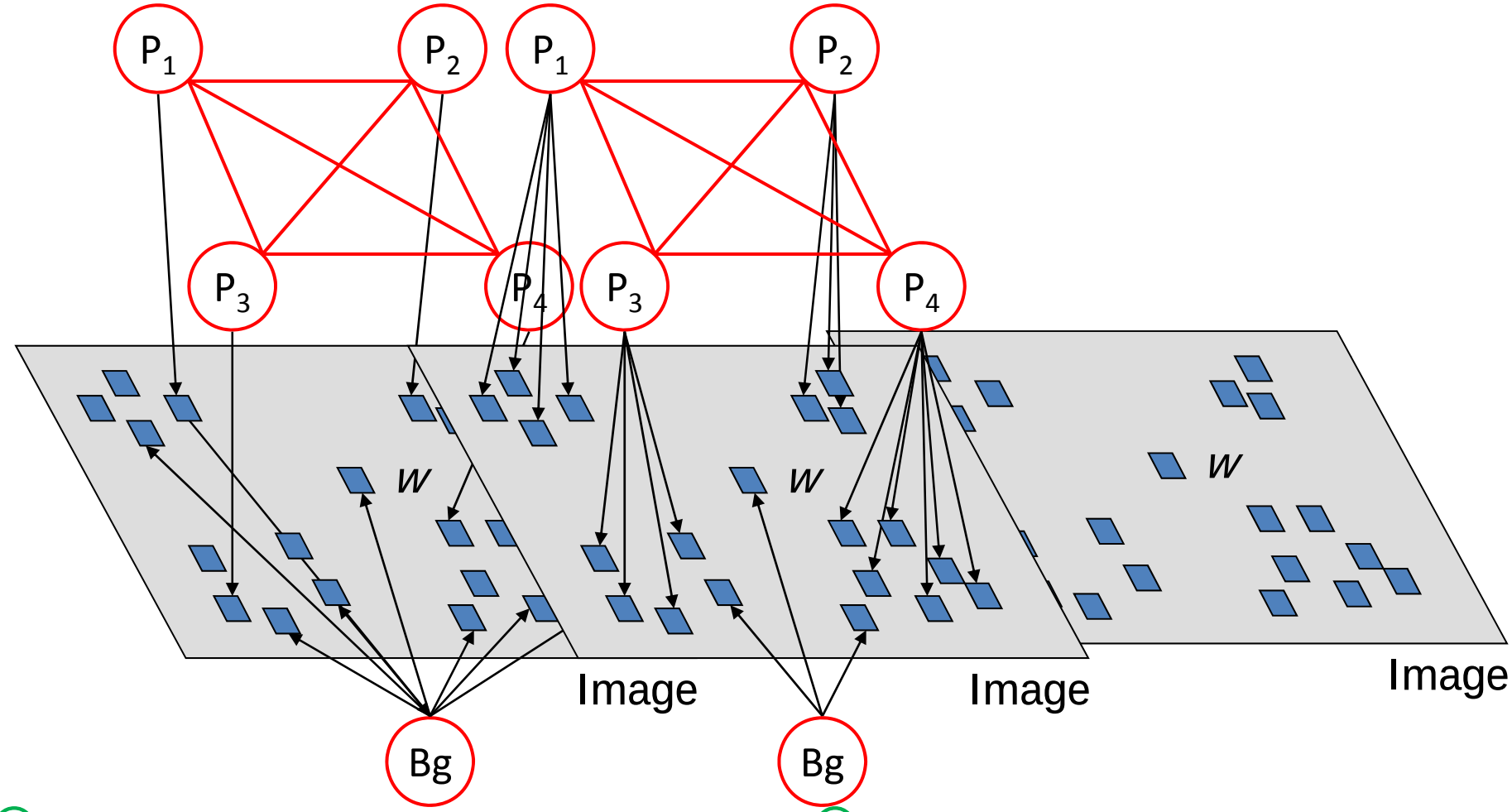
☺ Large number of features

☹ No shape information

Constellation model

Constellation of bags of features

bags of features

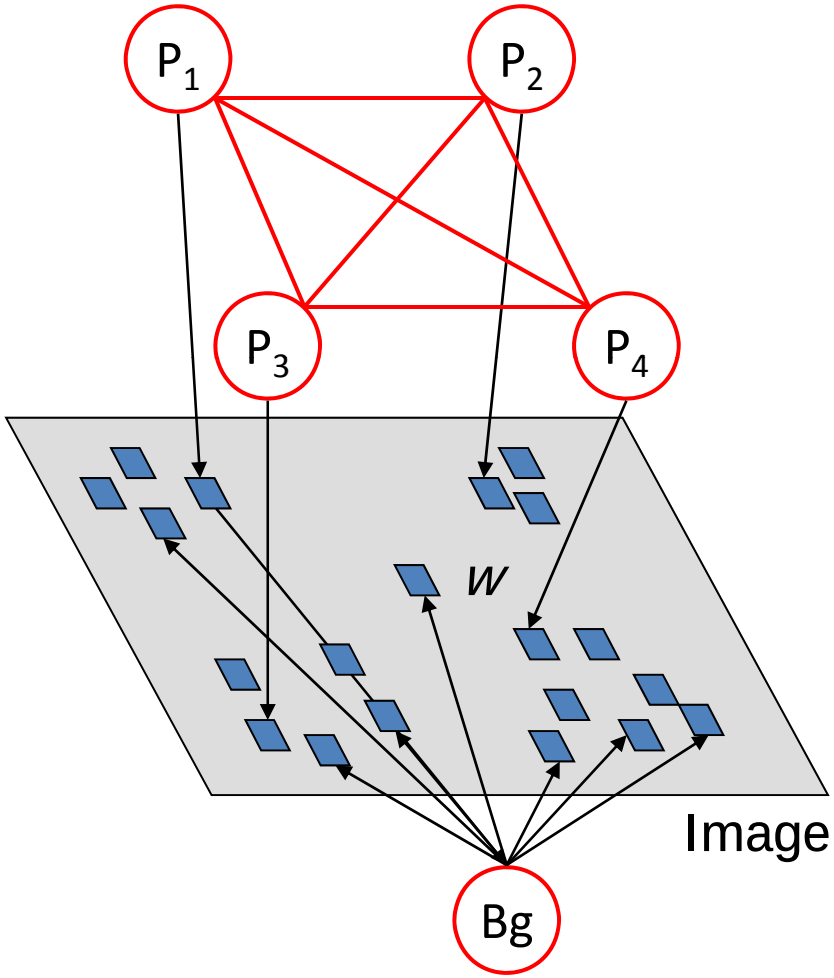


☺ Strong shape representation

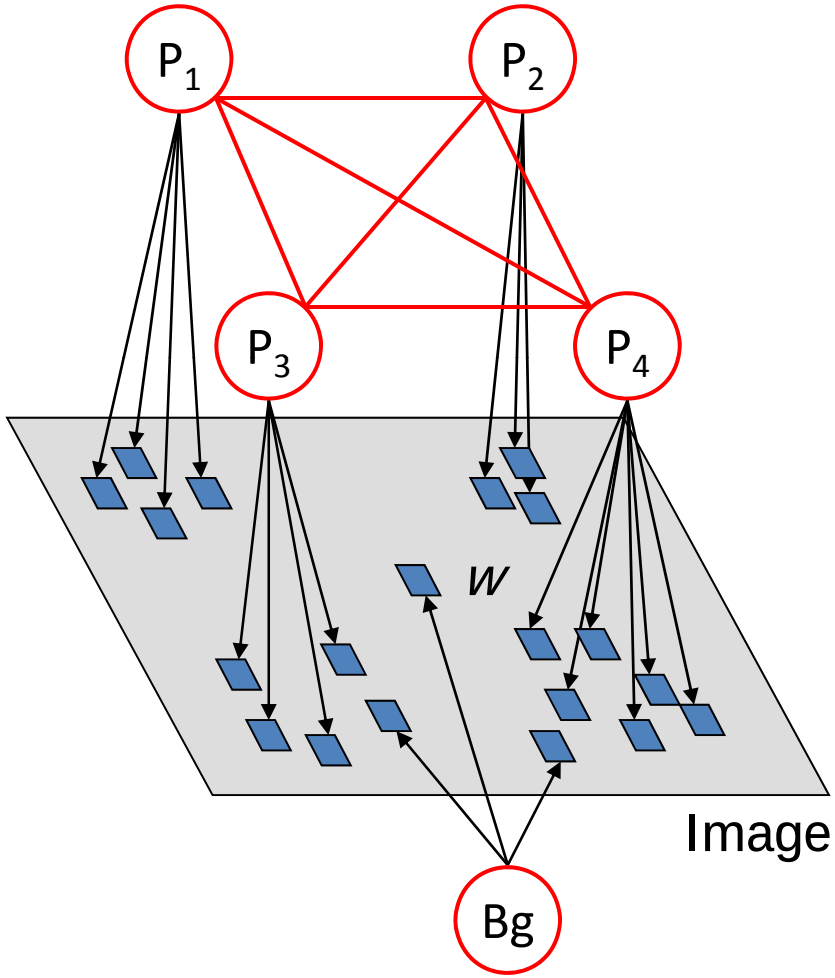
☺ Large number of features

J.C. Niebles, & L. Fei-Fei, CVPR, 2007

Constellation Model

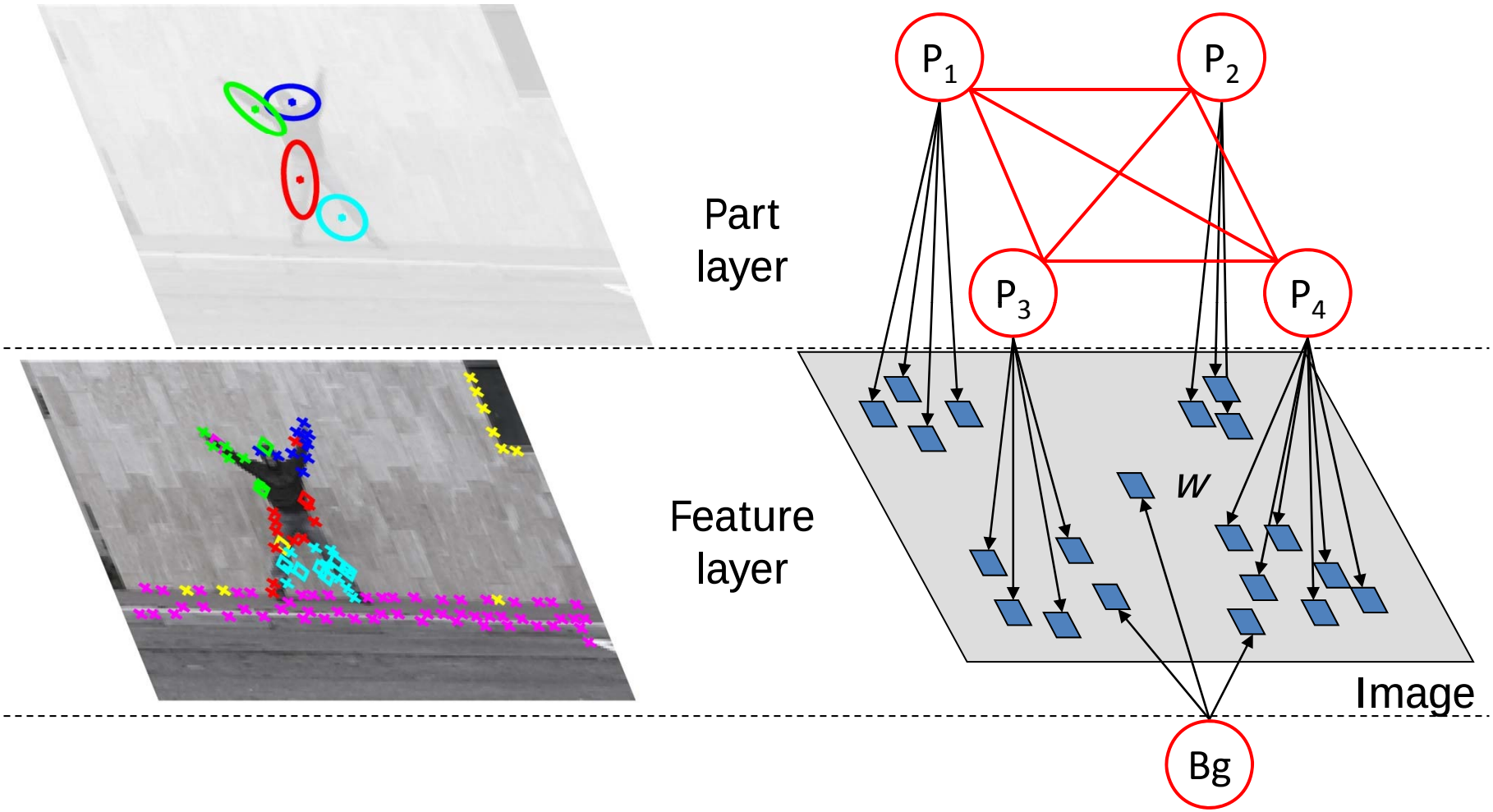


Constellation of bags of features



J.C. Niebles, & L. Fei-Fei, CVPR, 2007

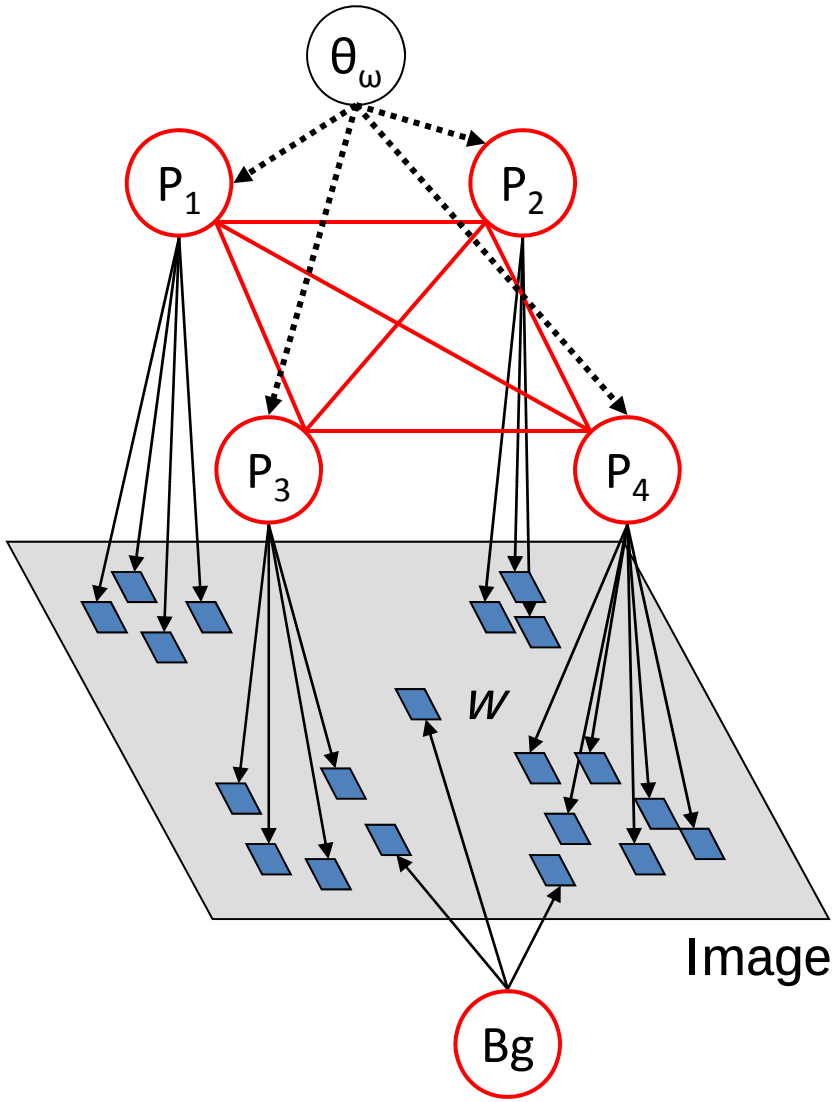
Constellation of bags of features



J.C. Niebles, & L. Fei-Fei, CVPR, 2007

Constellation of bags of features

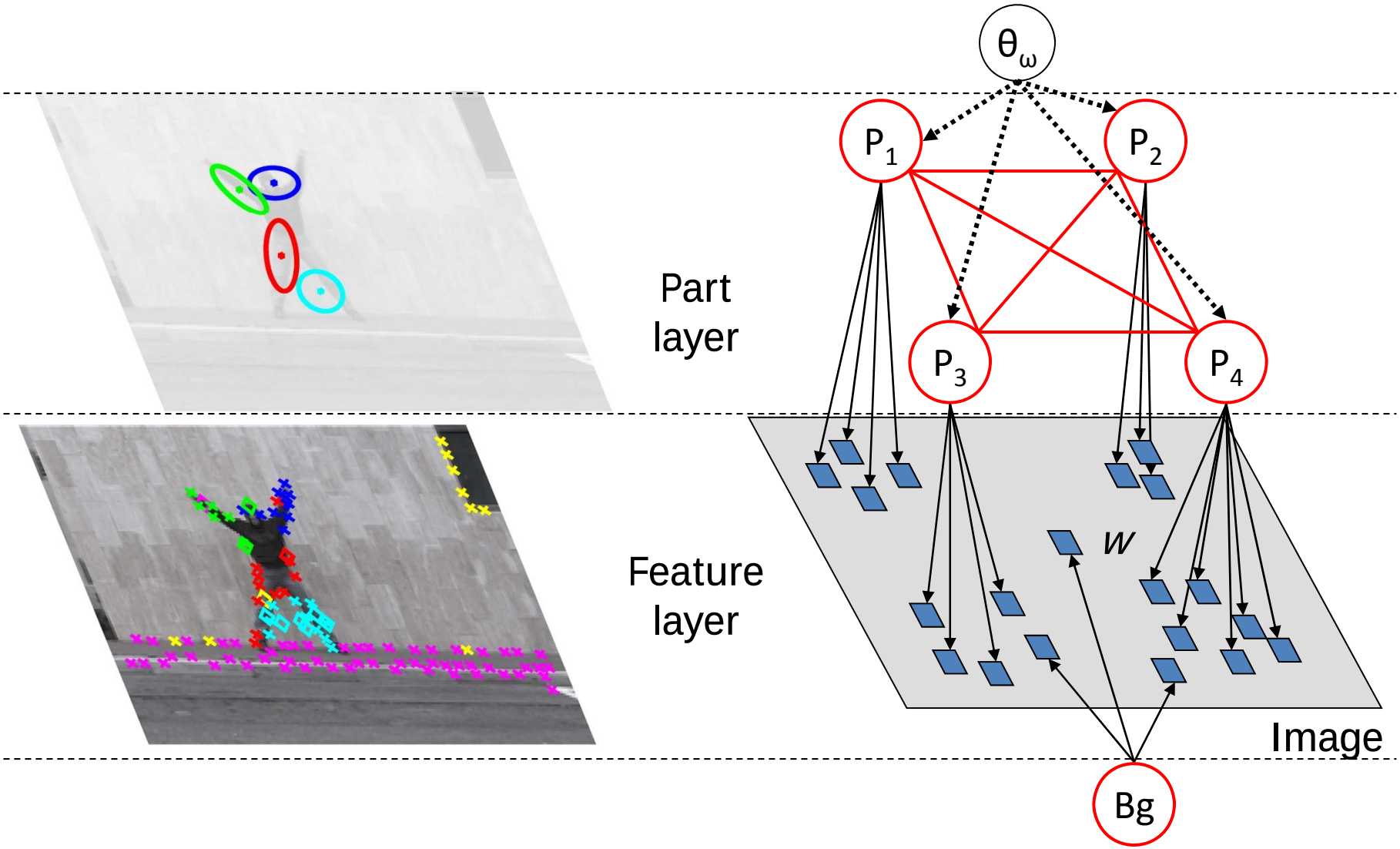
bend



- Use a mixture to account for data multimodality.

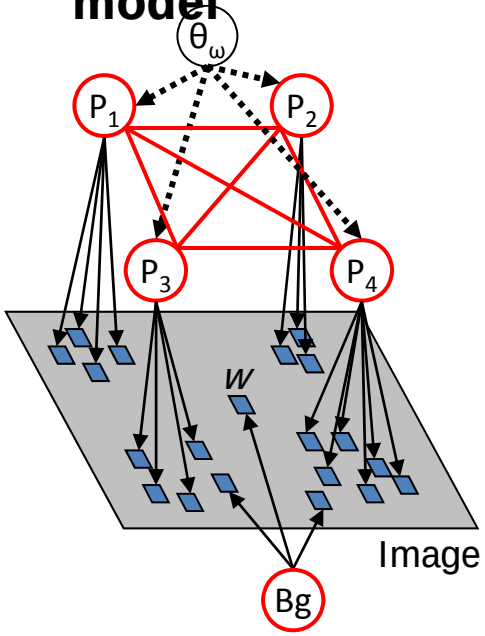
J.C. Niebles, & L. Fei-Fei, CVPR, 2007

Constellation of bags of features

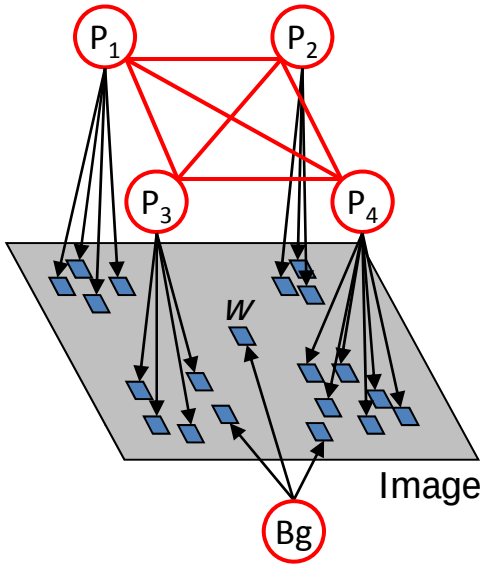


J.C. Niebles, & L. Fei-Fei, CVPR, 2007

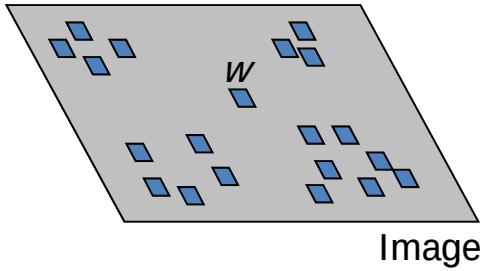
Mixture model



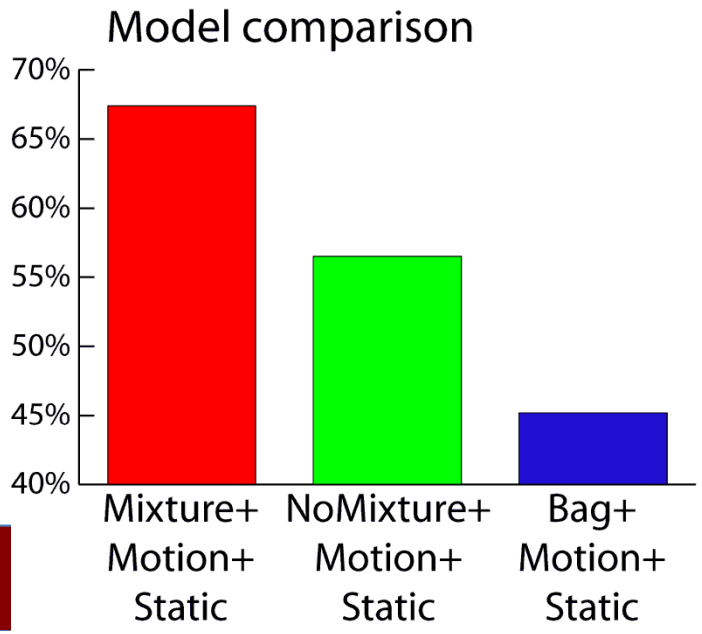
Single component



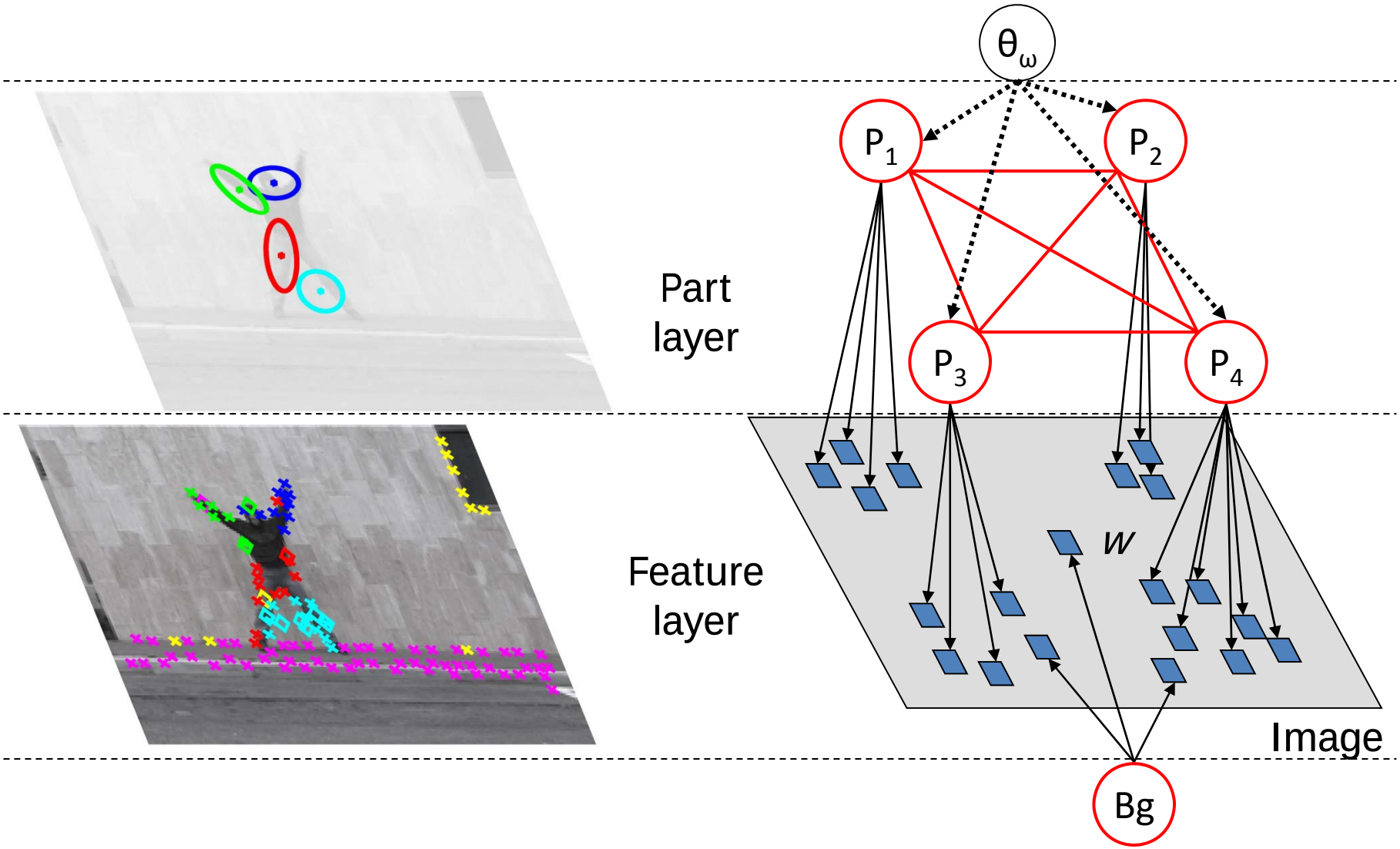
bag of words



Classification performance



Constellation of bags of features



J.C. Niebles, & L. Fei-Fei, CVPR, 2007

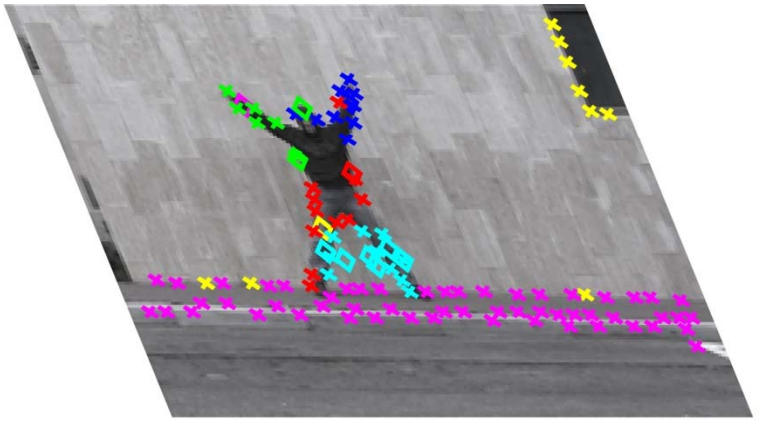
Constellation of bags of features

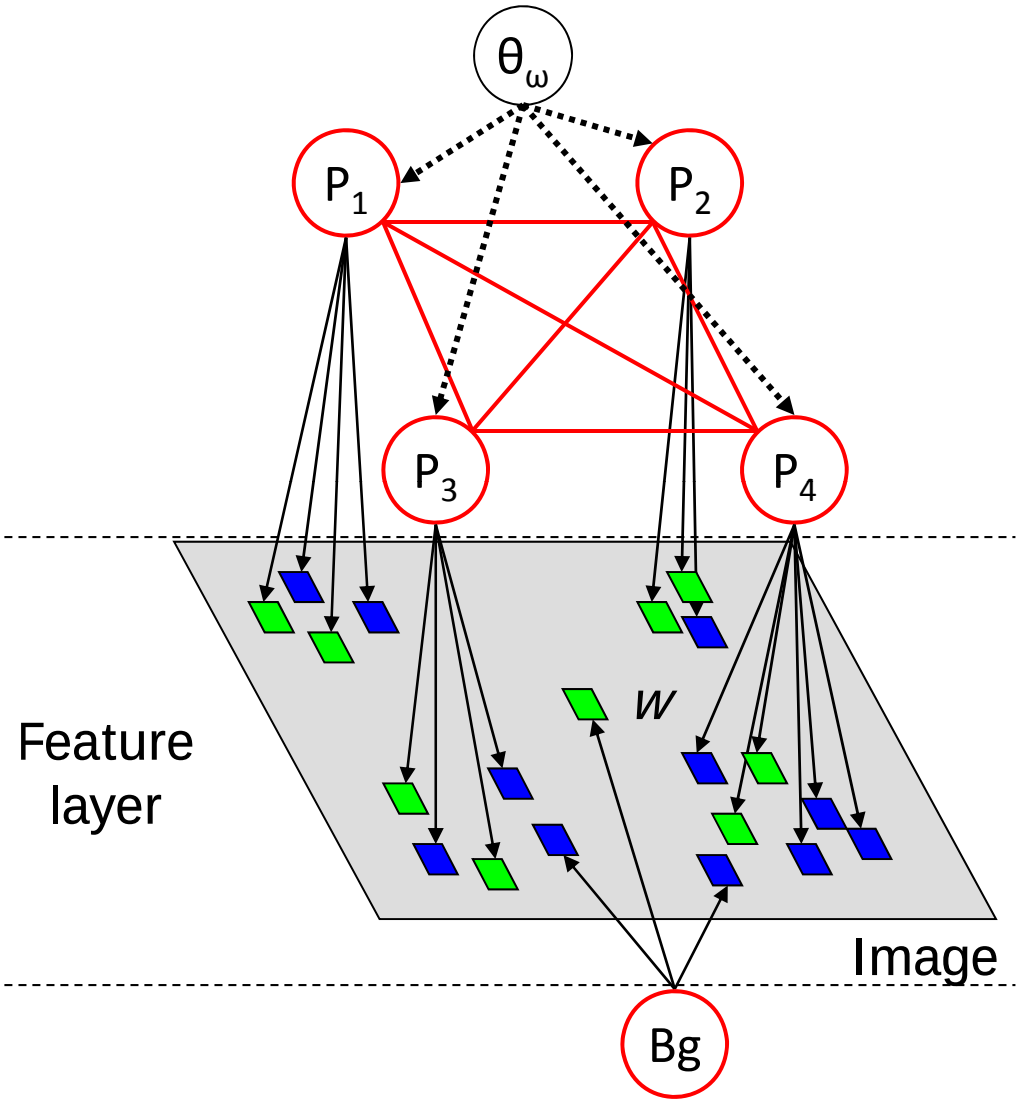
- Static features

  Edge map + Shape Context
  [Belongie '03]

- Motion features

  Interest Points + ST-gradients
  [Dollar '05]





J.C. Niebles, & L. Fei-Fei, CVPR, 2007

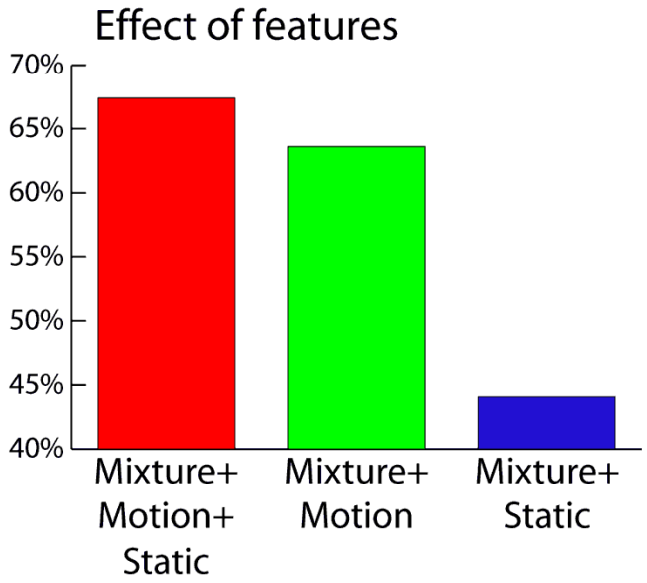
Constellation of bags of features

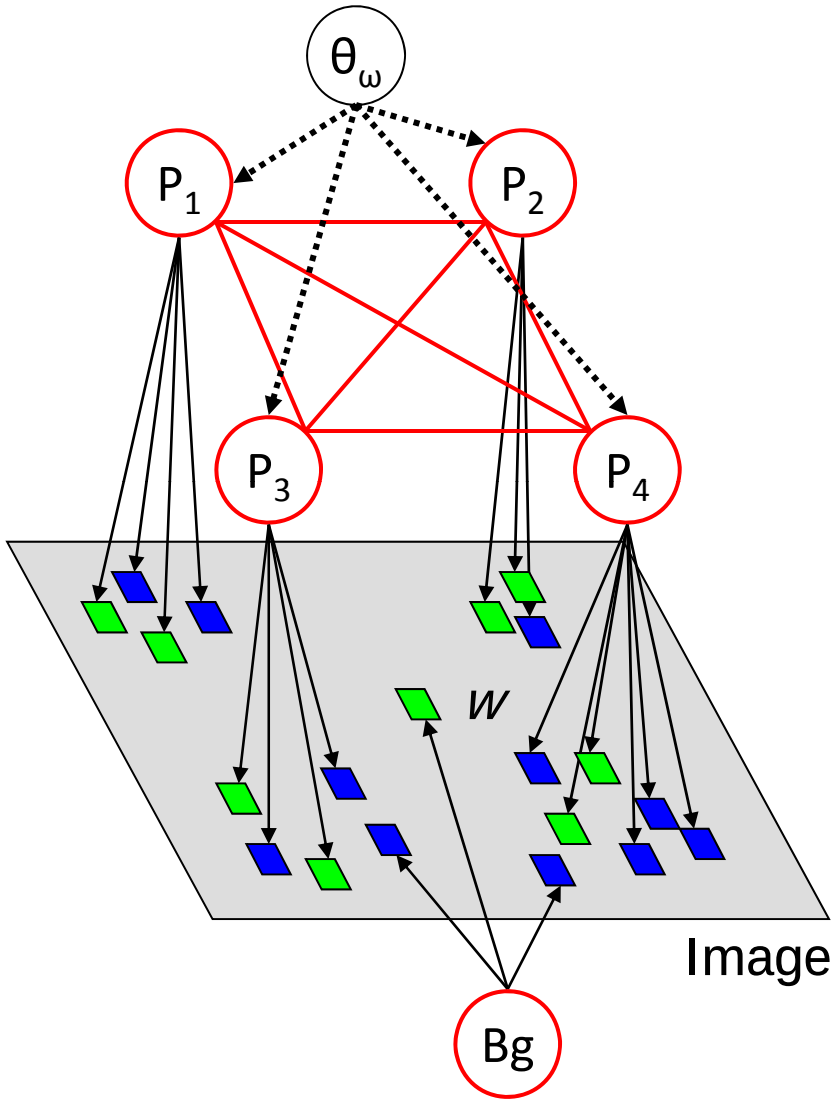
- Static features

  Edge map + Shape Context
  [Belongie '03]

- Motion features

  Interest Points + ST-gradients
  [Dollar '05]





J.C. Niebles, & L. Fei-Fei, CVPR, 2007

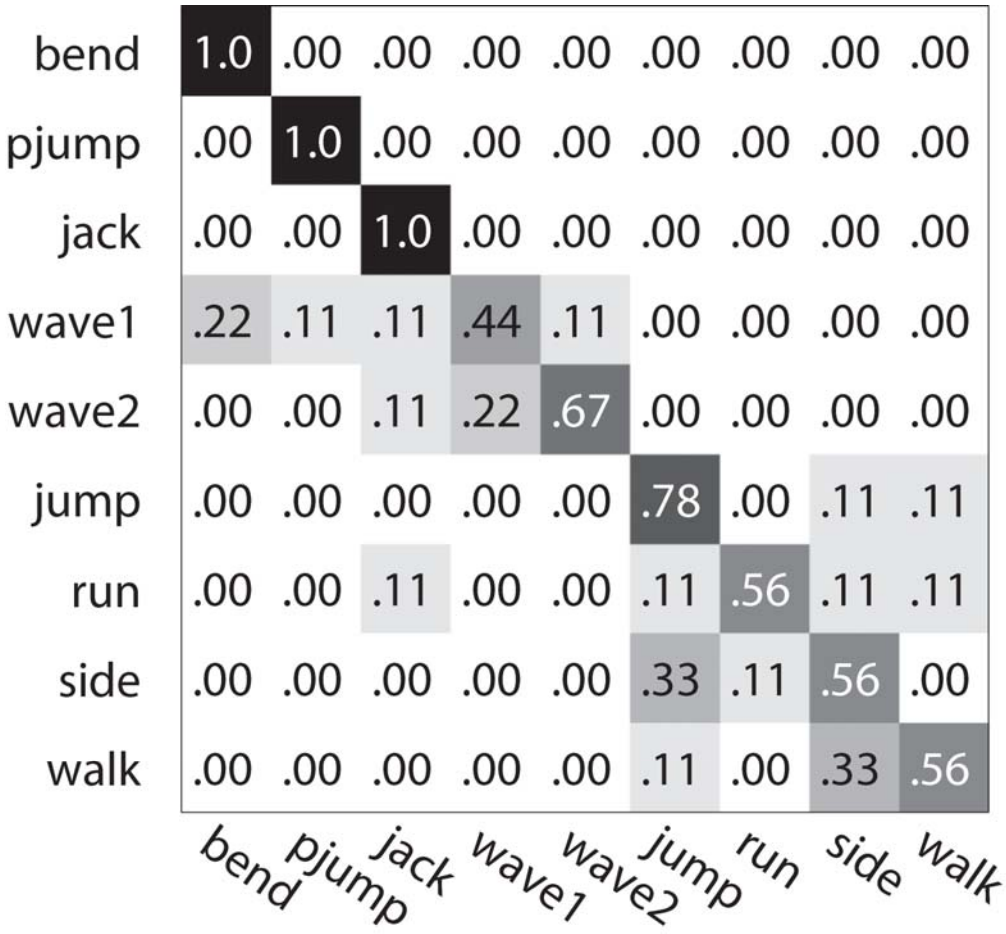
Dataset

“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.

J.C. Niebles, & L. Fei-Fei, CVPR, 2007

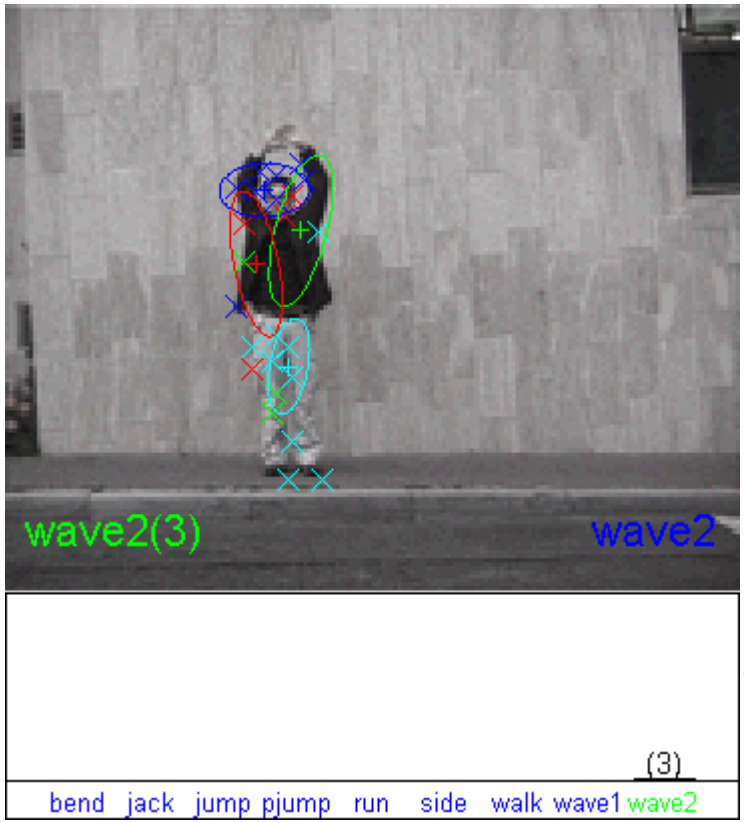
Experiment

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



J.C. Niebles, & L. Fei-Fei, CVPR, 2007

Recognition with constellation of bags of features



Predicted frame label

Ground truth frame label

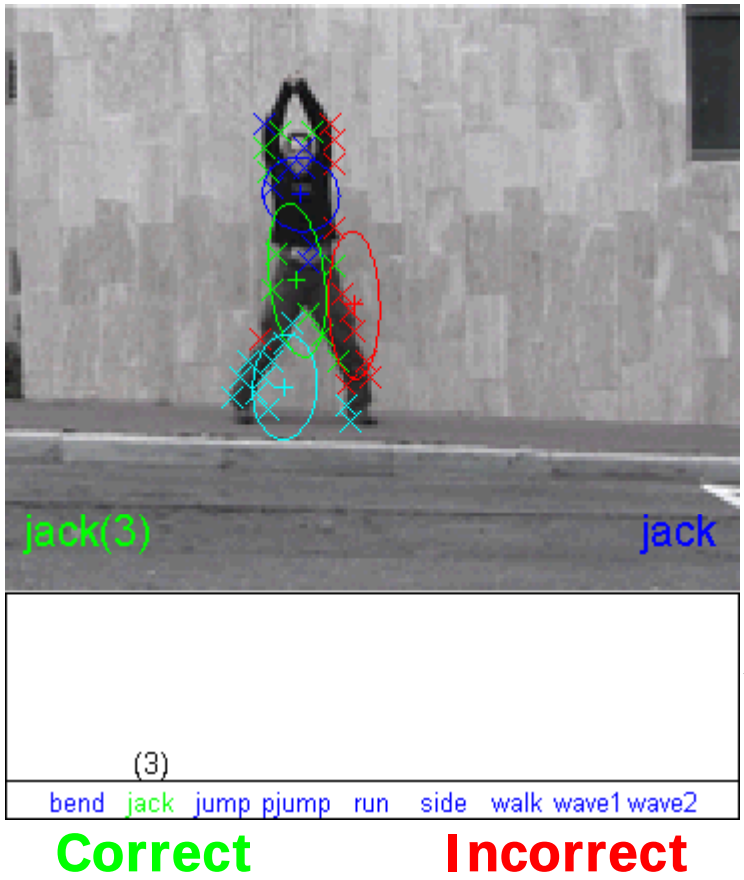
Accumulated votes

Predicted sequence label

Correct **Incorrect**

J.C. Niebles, & L. Fei-Fei, CVPR, 2007

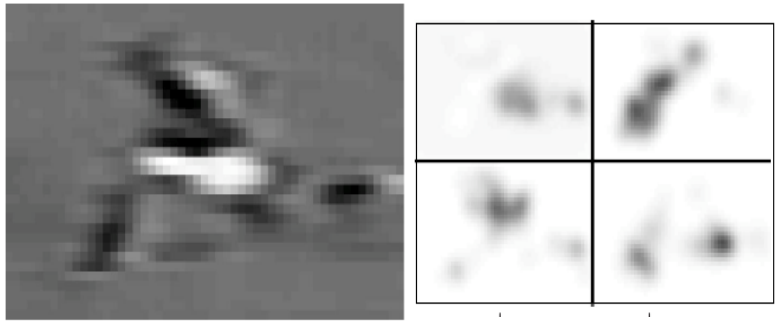
Recognition with constellation of bags of features



J.C. Niebles, & L. Fei-Fei, CVPR, 2007

Three representations

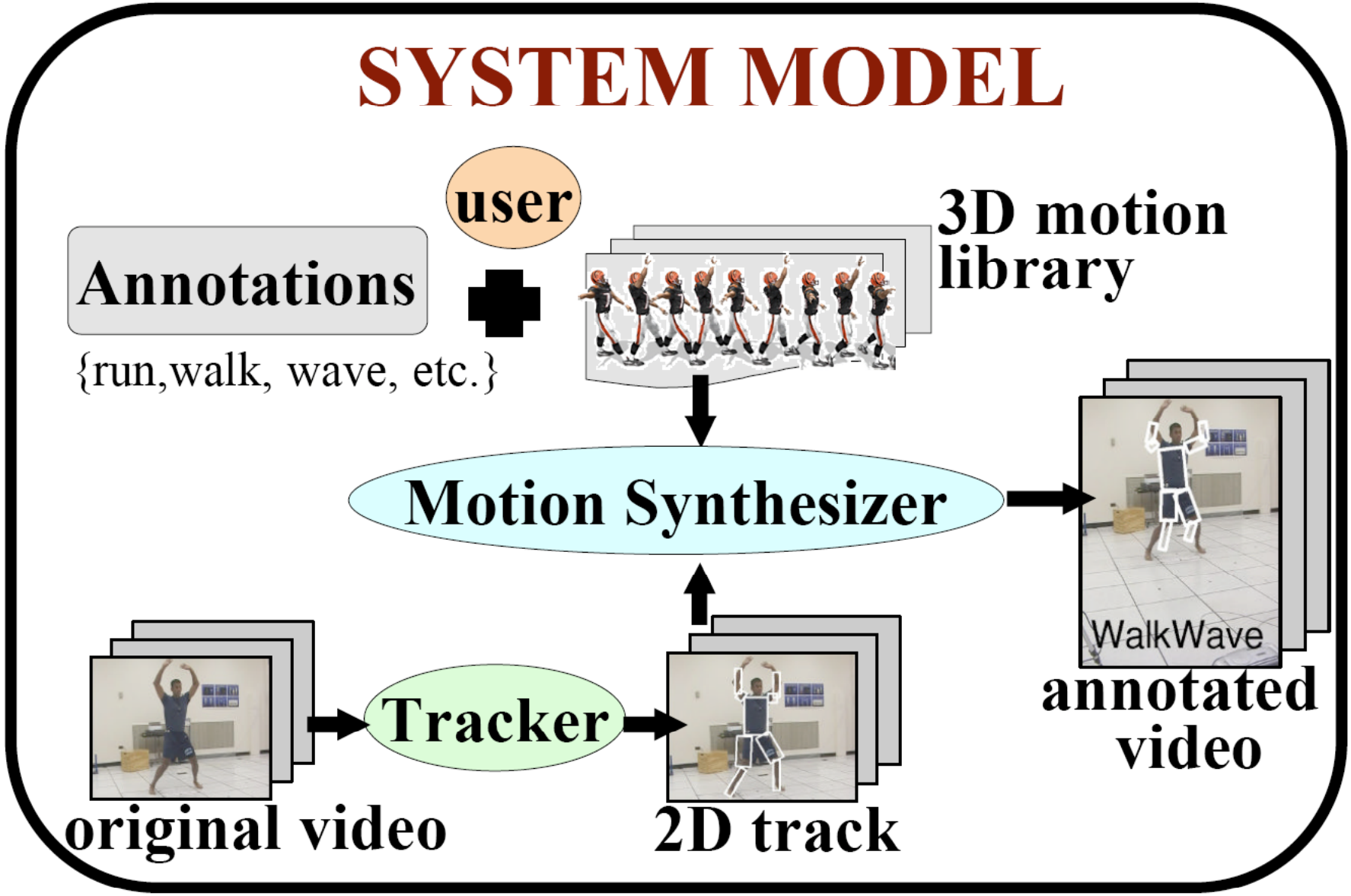
- Global template



- Local video patches

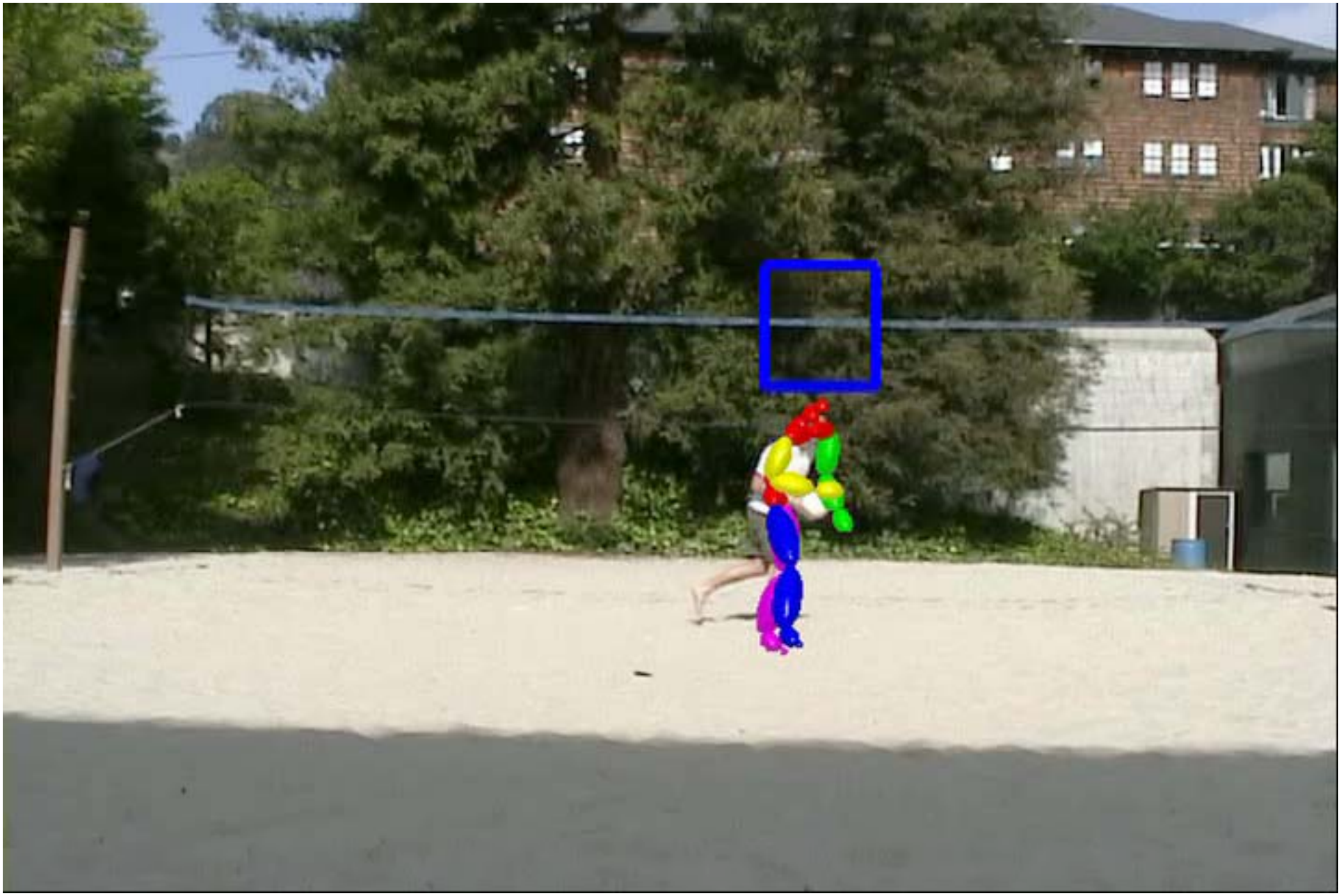
- **Kinematic/body pose**

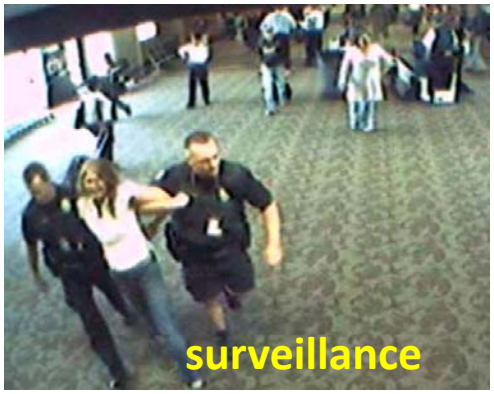
SYSTEM MODEL



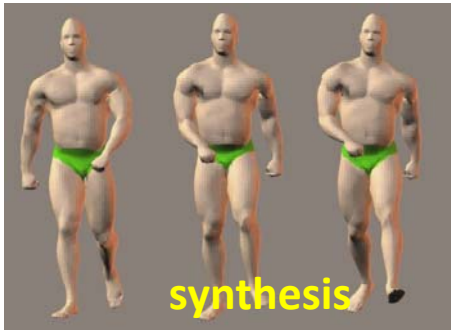
[Ramanan& Forsyth, NIPS 2003]


Run

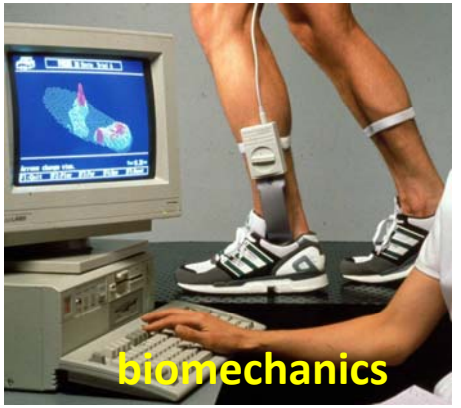




search



Human Motion Analysis



Useful Surveys

- Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O. Machine Recognition of Human Activities: A Survey CirSysVideo(18), No. 11, November 2008, pp. 1473-1488.

- David A. Forsyth, Okan Arikan, Leslie Ikemoto. Computational Studies of Human Motion: Tracking and Motion Synthesis, Part 1

What we have learned today?

- Introduction
- Motion classification using template matching
- Motion classification using spatio-temporal features
- Motion classification using pose estimation