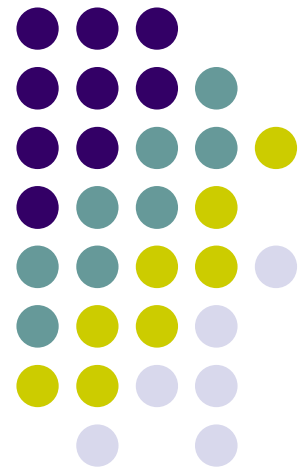


Unsupervised Learning of Human Actions Categories Using Spatial-Temporal Words

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



Outline

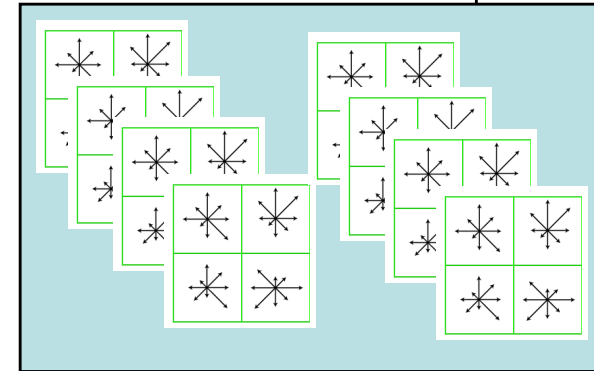


- Introduction
- Explaining the Model
- Results
- Discussion

Standard Approach (adopted from text IR)



Feature extraction
and representation
(e.g., SIFT)



Quantization
+ histogram



“Bag of visual words”

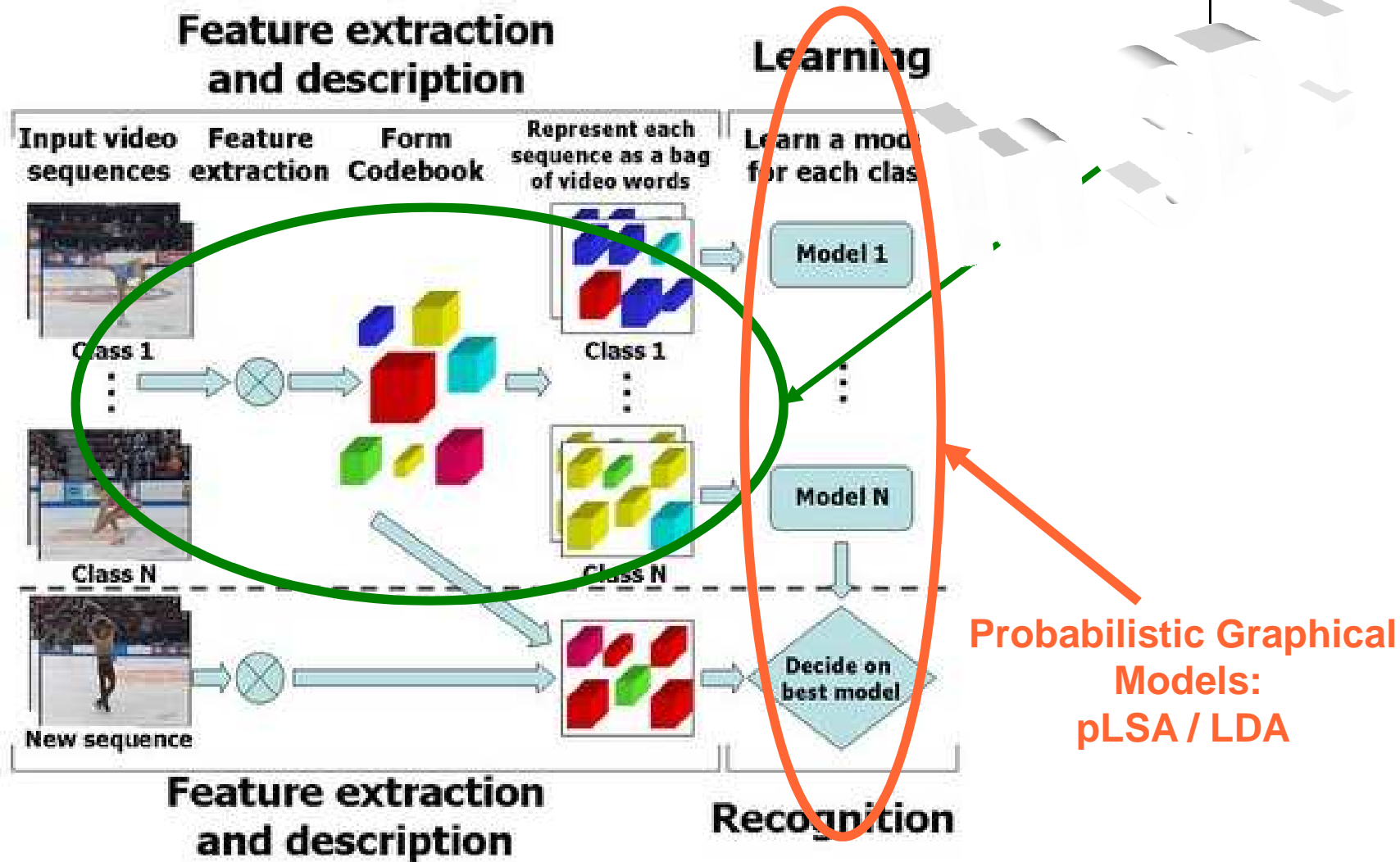


Classification
(e.g., SVMs)

| | | | |
|------------|---|---|-----|
| • sheep? | ✗ | ← | SVM |
| • bus? | ✓ | ← | SVM |
| • cat? | ✗ | | SVM |
| • bicycle? | ✓ | | SVM |
| • car? | ✓ | ⋮ | SVM |
| • cow? | ✗ | ⋮ | SVM |
| • dog? | ✗ | | SVM |
| • horse? | ✗ | | SVM |
| • mbike? | ✓ | | SVM |
| • person? | ✓ | ← | SVM |

[Fei-Fei *et al.*, 2005];
[Sivic *et al.*, 2005];
and many others

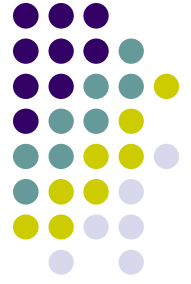
In our case:





Extending the Analogy

| | Text | Images | Videos |
|-----------------|------------------|-----------------|--------------------------|
| Building Blocks | Words | 2D visual words | 3D spatio-temporal words |
| Input | Documents | Images | Image Sequences |
| Latent Topics | Politics, Sports | Cars, Elephants | Walking, Clapping |



Goal

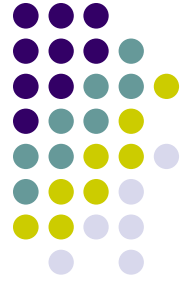
- Input: a collection of unlabeled videos
- Automatically learn classes of actions
- Apply learned model to categorize and localize actions in new video sequences

Outline

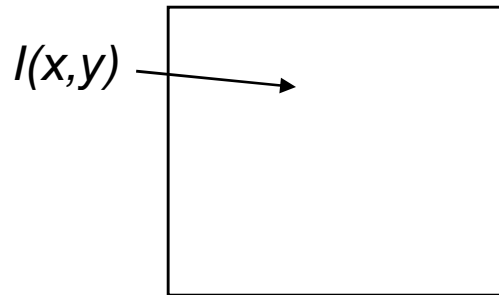


- Introduction
- Explaining the Model – ‘Bag of Video Words’
 - Codebook Generation
 - Learning the Action Models
 - Categorization and Localization
- Results
- Discussion

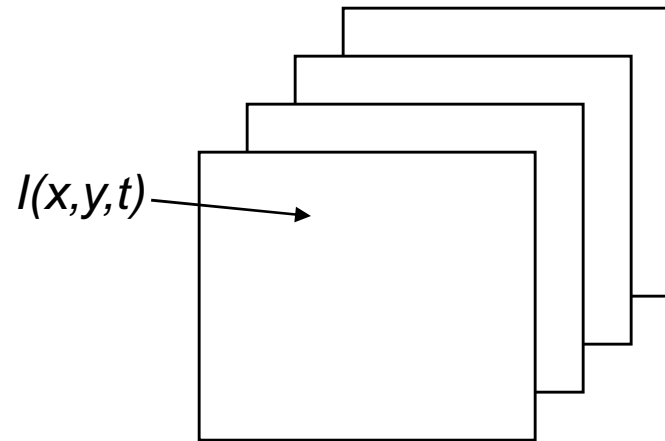
Digression – The Space Time Volume



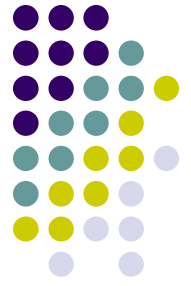
Static image:



Video:

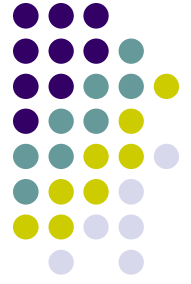


Codebook Generation – Space Time Interest Points



- What are we talking about?
- Space-Time Corners
 - Gradient can be taken over temporal dimension as well
 - Strong response where a spatial corner whose velocity is reversing direction
 - Well suited for recognizing human actions such as walking, clapping, hand waving, etc...
 - Too sparse for more articulated behaviours: facial expressions, rodent behaviour
 - Agnostic to periodic circular motion

Codebook Generation – Space Time Interest Points (2)



- Separable Linear Filter Method

(as suggested in [3] “*Behaviour Recognition via Sparse Spatio-Temporal Features*”, Dollar et al., 2005)

- Compute response function at different spatio-temporal scales:

$$R = \left[\mathbf{I} \square \mathbf{g} \square \mathbf{h}_{ev} \right]^2 \square \square \left[\mathbf{I} \square \mathbf{g} \square \mathbf{h}_{od} \right]^2$$

- ... and choose local maxima

Codebook Generation – Space Time Interest Points (3)



$$R = \left[I \otimes g \otimes h_{ev} \right]^2 + \left[I \otimes g \otimes h_{od} \right]^2$$

- Smooth each image with a gaussian kernel:

$$g(x, y; \sigma)$$

- Apply a quadrature pair of 1D Gabor filters along the temporal dimension:

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) e^{-t^2/\tau^2}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega) e^{-t^2/\tau^2}$$

- In all cases, $\omega = 4/\tau$ giving R two controlled parameters, responding to the scale of the detector

Codebook Generation – Space Time Interest Points (4)



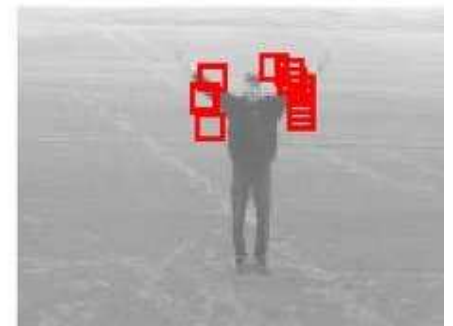
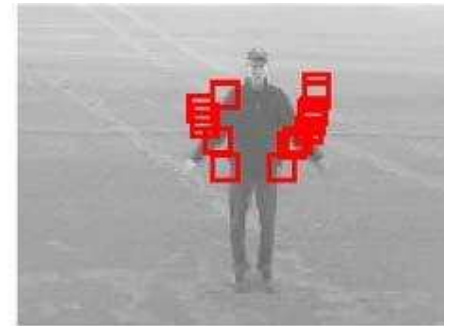
- Separable Linear Filter Interest Point Detector:
 - Assumes stationary camera or a process that can account for camera motion
 - Tuned to fire whenever intensity variations contain periodic frequency components
 - Responds strongly to other types of motion as well, including spatio-temporal corners
 - In general, any area containing interesting spatial features that undergoes complex motion will trigger a response
 - Pure translational motion will go unnoticed

Codebook Generation – Space Time Interest Points (5)

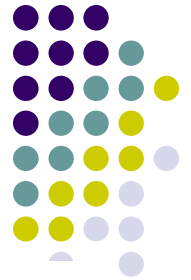


- Separable Linear Filter Interest Point Detector:
 - To handle multiple scales, a set of detectors must be used
 - In this work, only one scale is employed: “ ... [we] rely on the codebook to encode the few changes in scale that are observed in the dataset. “

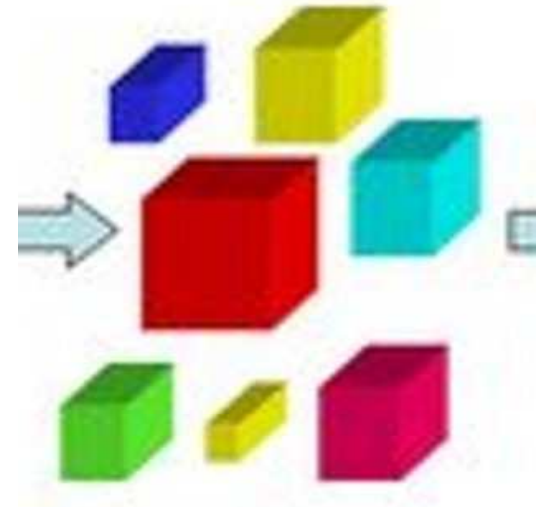
Codebook Generation – Space Time Interest Points



Codebook Generation – Space Time Descriptors



- The Cuboids:
 - Side length of six times the scale at which they were detected
- Descriptor:
 - Transform
 - Create feature vector
 - Reduce dimensionality (PCA)



Codebook Generation – Space Time Descriptors

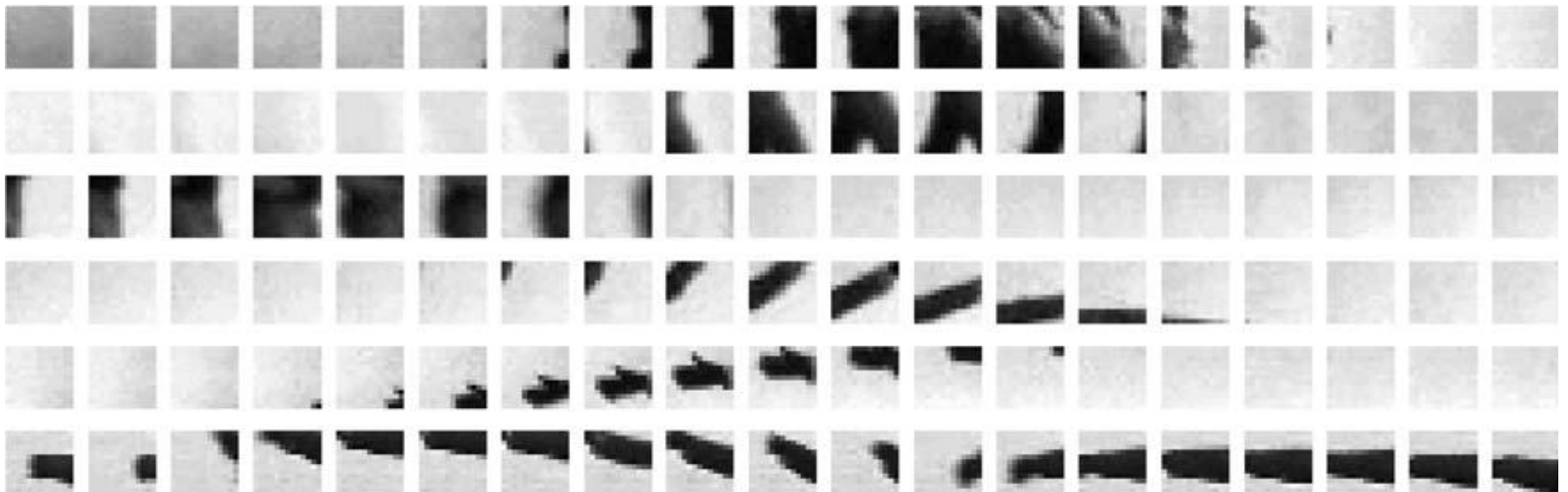


- The Descriptor:
 - Transform using one of three methods:
 - Normalized pixel values, Brightness Gradient, Windowed Optical Flow
 - Create feature vector, using one of three methods:
 - Vectorize, Histogram, Local Histograms (SIFT-style)
 - Nine combinations compared in [2]
 - Brightness gradient and local histograms performed best, and used in our paper.

Codebook Generation – That's It



- Once we have a set of descriptors extracted from the training videos, we create a codebook of size k using the k -means algorithm.

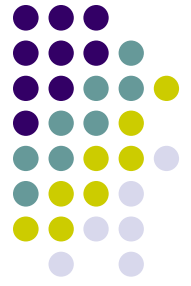


Outline



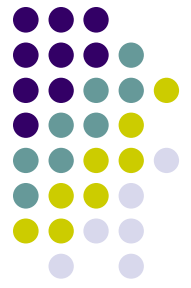
- Introduction
- Explaining the Model – ‘Bag of Video Words’
 - Codebook Generation
 - Learning the Action Models
 - Categorization and Localization
- Results
- Discussion

Learning the Action Models: pLSA

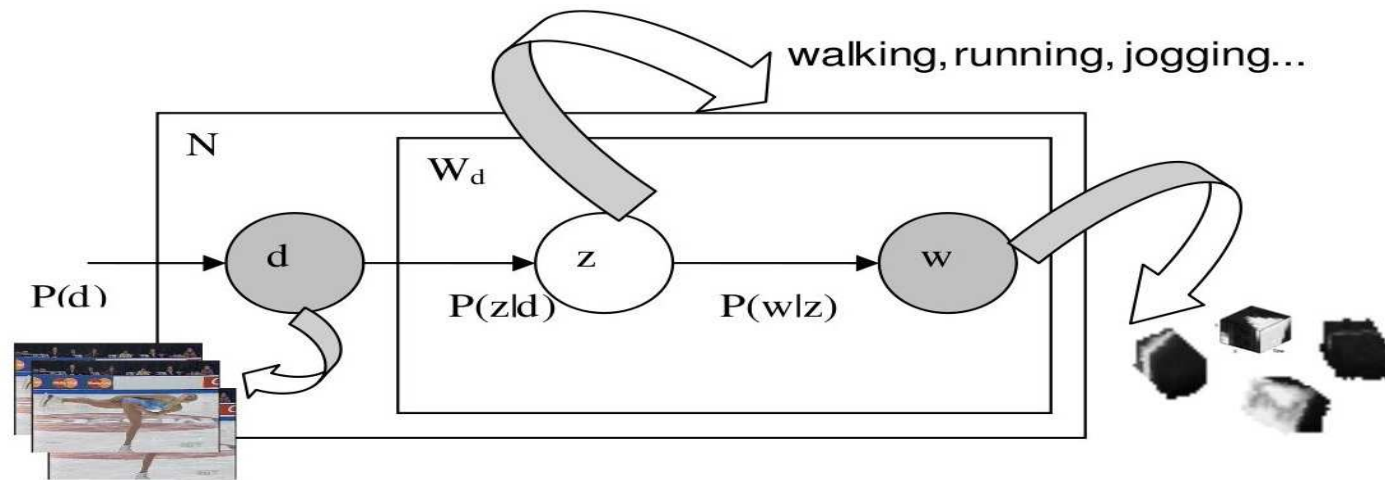


- N ($j = 1, \dots, N$) video sequences containing video words from vocabulary of size M ($i = 1, \dots, M$)
- Summarized as co-occurrence matrix \mathbf{N} , where $n(w_i, d_j)$ - number of occurrences of w_i in video d_j
- Assuming latent topics, the joint distribution is:

$$P(w_i, d_j, z_k)$$



Learning the Action Models: pLSA



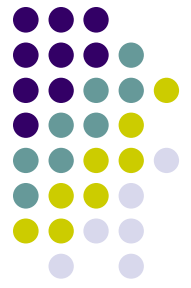
- Basic probability:

$$P(d_j, w_i) = P(d_j) P(w_i | d_j)$$

- Sum out latent

topics:

$$P(w_i | d_j) = \sum_{k=1}^K P(z_k | d_j) P(w_i | z_k)$$



Learning the Action Models: pLSA

- Intuitively, each video is expressed as a convex mixture of K action category vectors
- Videos are characterized with a specific mixture of factors with weights $P(z_k | d_j)$
 - => amounts to matrix decomposition, with probability constraints
- MLE of parameters is obtained through EM maximization of:

$$\prod_{i=1}^M \prod_{j=1}^N P(w_i | d_j)^{n_{w_i, d_j}}$$

Outline



- Introduction
- Explaining the Model – ‘Bag of Video Words’
 - Codebook Generation
 - Learning the Action Models
 - Categorization and Localization
- Results
- Discussion



Categorization of New Videos

- Project new video on the simplex projected by the learnt $P(w/z)$, and:

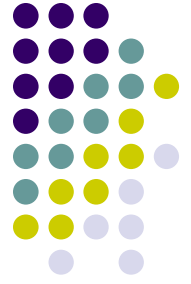
Find $P(z_k | d_{test})$ such that

$$KL(P(w | d_{test}), P(w | d_{test}))$$

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

- Solved by EM as well

$$\text{Action Category} = \operatorname{argmax} P(z_k | d_{test})$$



Localization in New Videos

- Already seen this, in Sivic et al. (2005)
- pLSA models posteriors as:

$$P(z_k | w_i, d_j) = \frac{P(w_i | z_k) P(z_k | d_j)}{\sum_{l=1}^K P(w_i | z_l) P(z_l | d_j)}$$

- MAP:

$$\text{Word's Topic} = \operatorname{argmax}_k P(z_k | w_i, d_j)$$

Outline

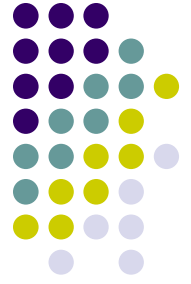


- Introduction
- Explaining the Model – ‘Bag of Video Words’
 - Codebook Generation
 - Learning the Action Models
 - Categorization and Localization
- Results
- Discussion

Experiments and Results: KTH database

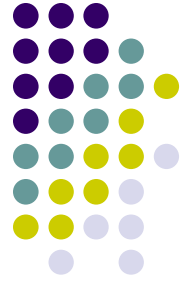


Experiments and Results: KTH database

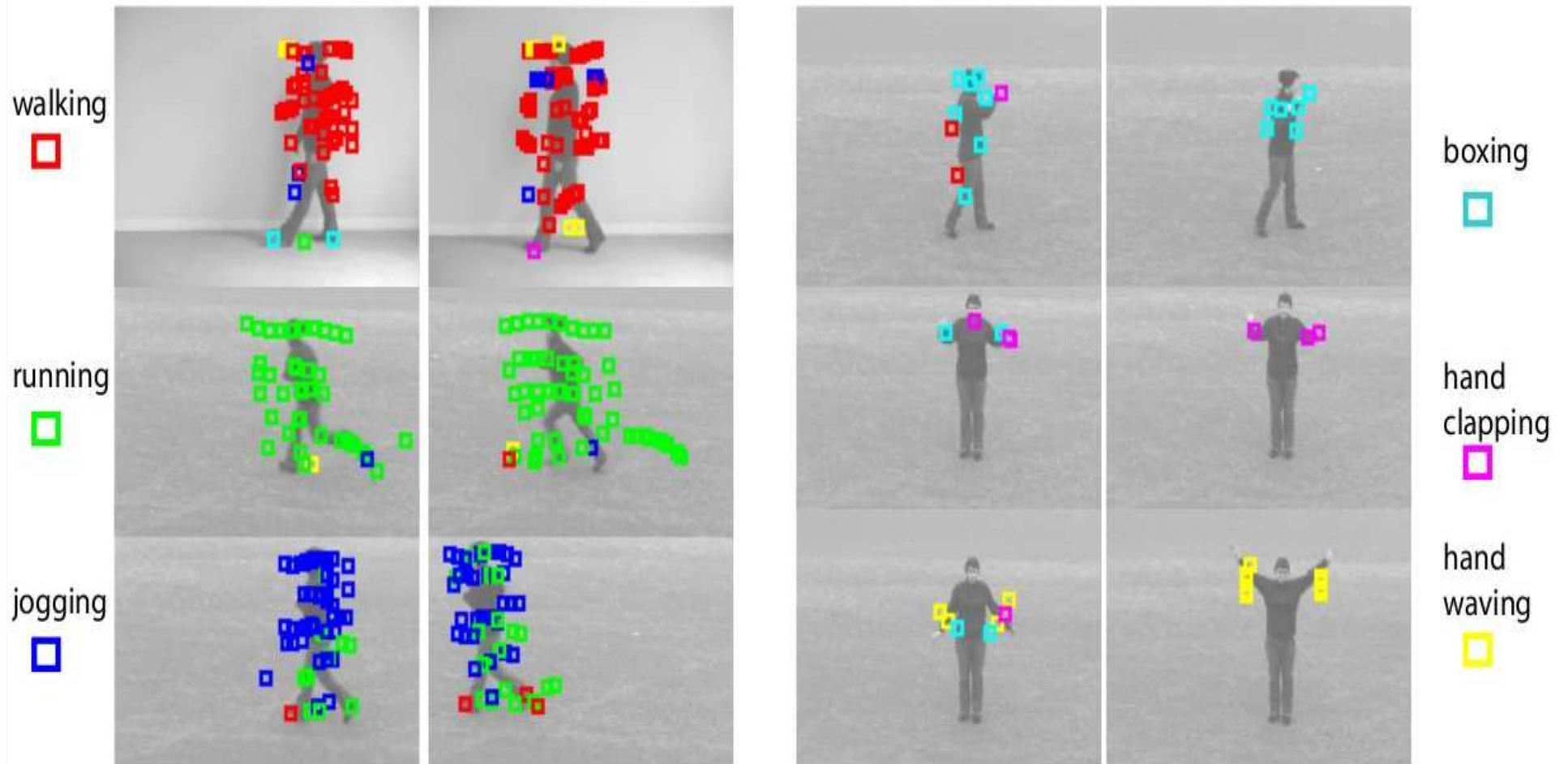


- Codebook built from two videos of each action, from three subjects
- Randomly select ~60,000 cuboids
- Leave-one-out Cross-validation: learn the model from 24 subjects, test on remaining subject

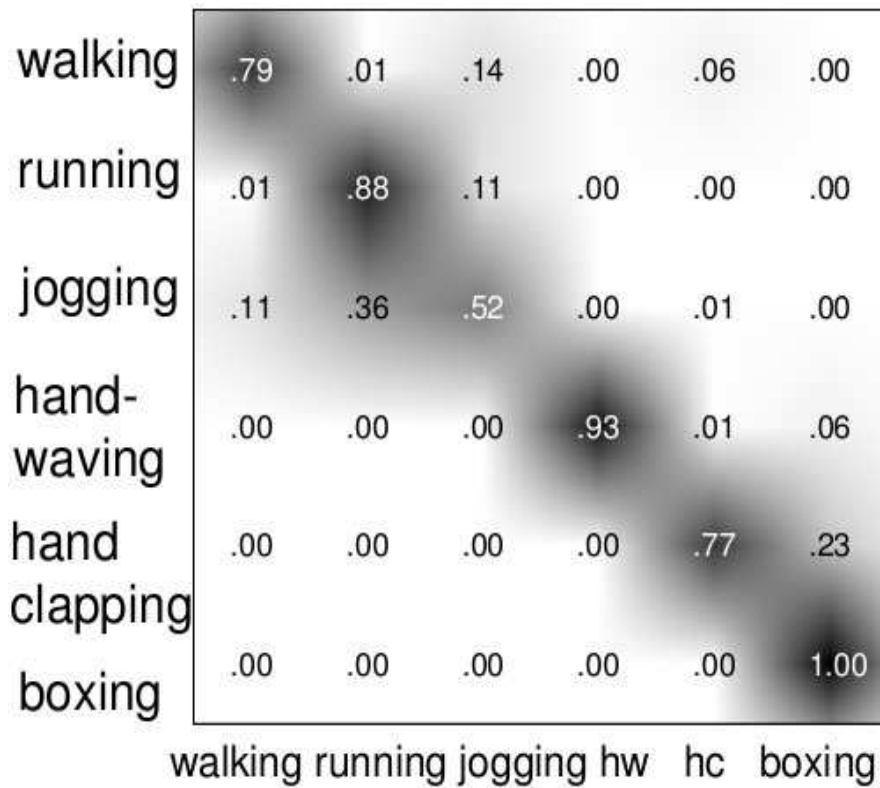
Experiments and Results: KTH database



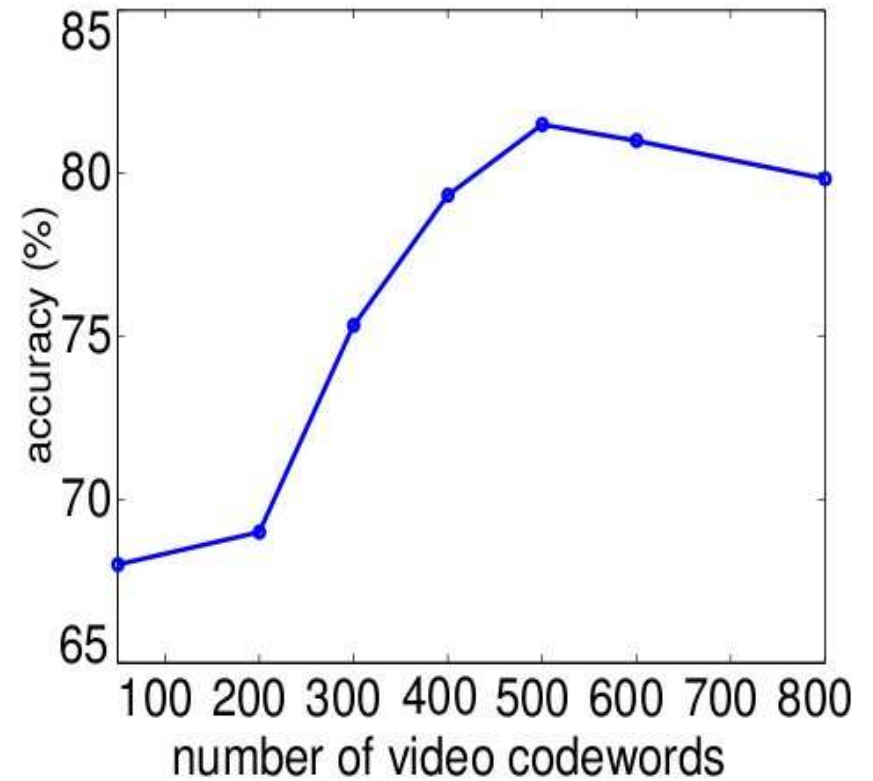
Experiments and Results: KTH database



Experiments and Results: KTH database

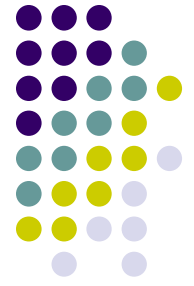


(a)



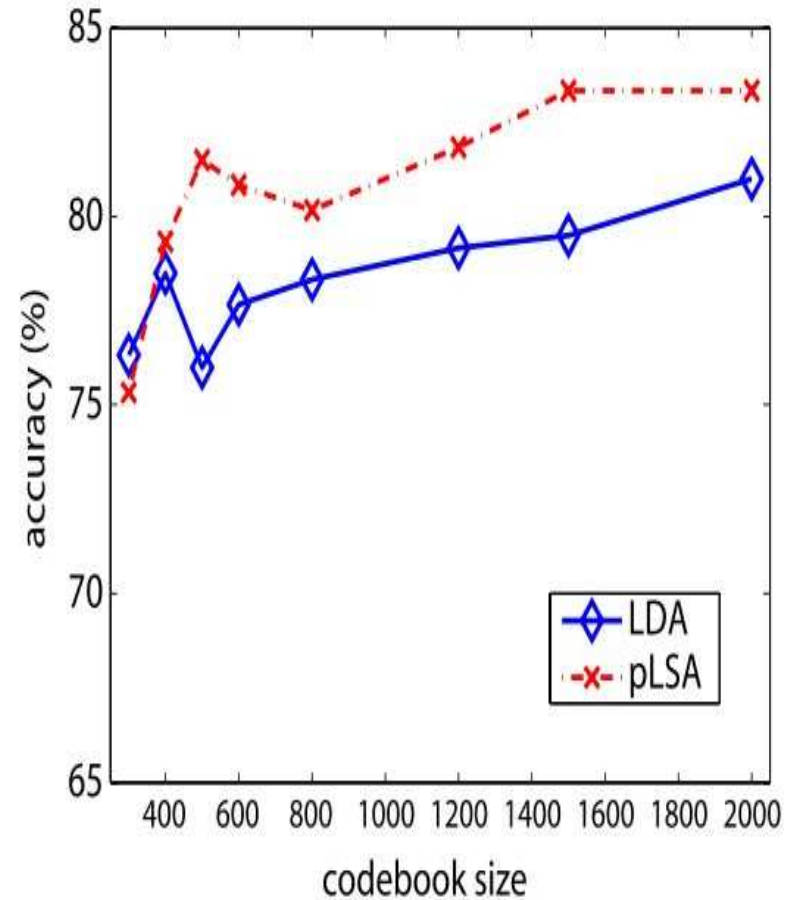
(b)

Experiments and Results: KTH database



| | | | | | | |
|--------------|-----|-----|-----|-----|-----|-----|
| walking | .82 | .00 | .17 | .00 | .00 | .01 |
| running | .01 | .88 | .11 | .00 | .00 | .00 |
| jogging | .09 | .37 | .53 | .00 | .01 | .00 |
| handwaving | .00 | .00 | .00 | .93 | .00 | .07 |
| handclapping | .00 | .00 | .00 | .00 | .86 | .14 |
| boxing | .00 | .00 | .00 | .00 | .02 | .98 |

(a)

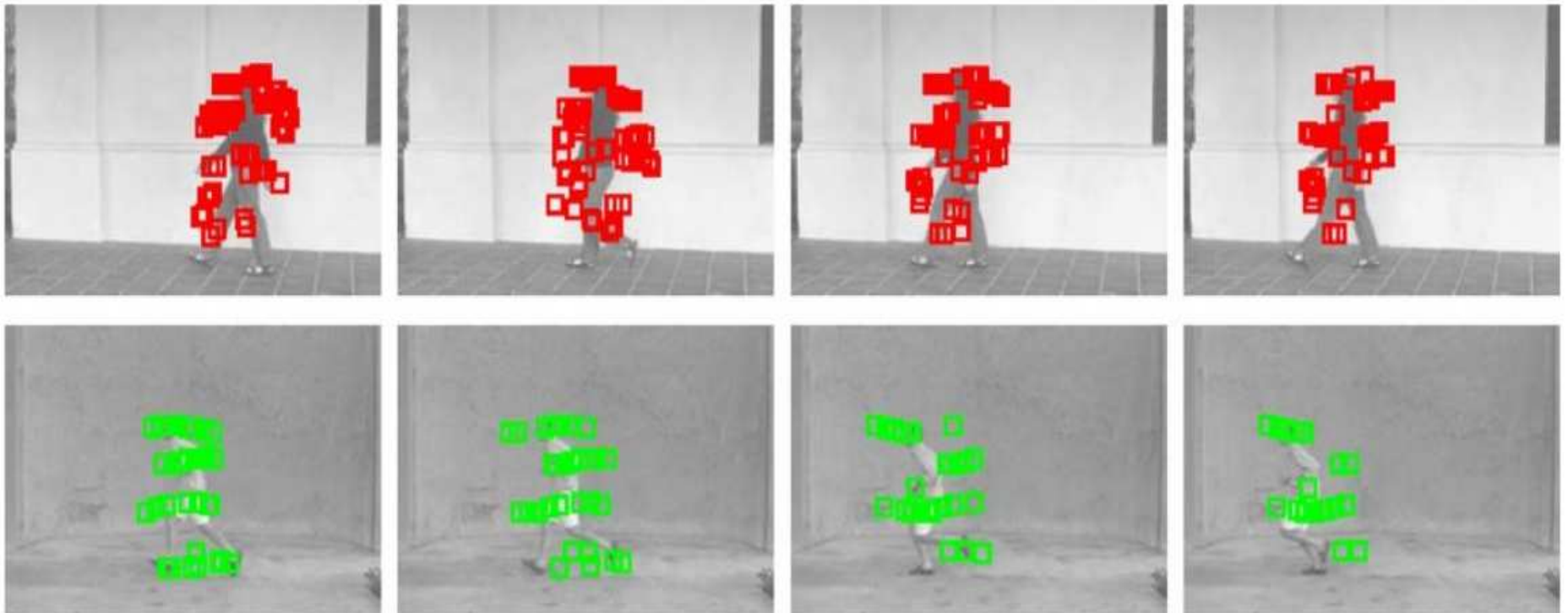


(b)

Experiments and Results: Caltech human motion database



- Learn the model using the KTH database, test it on Caltech's

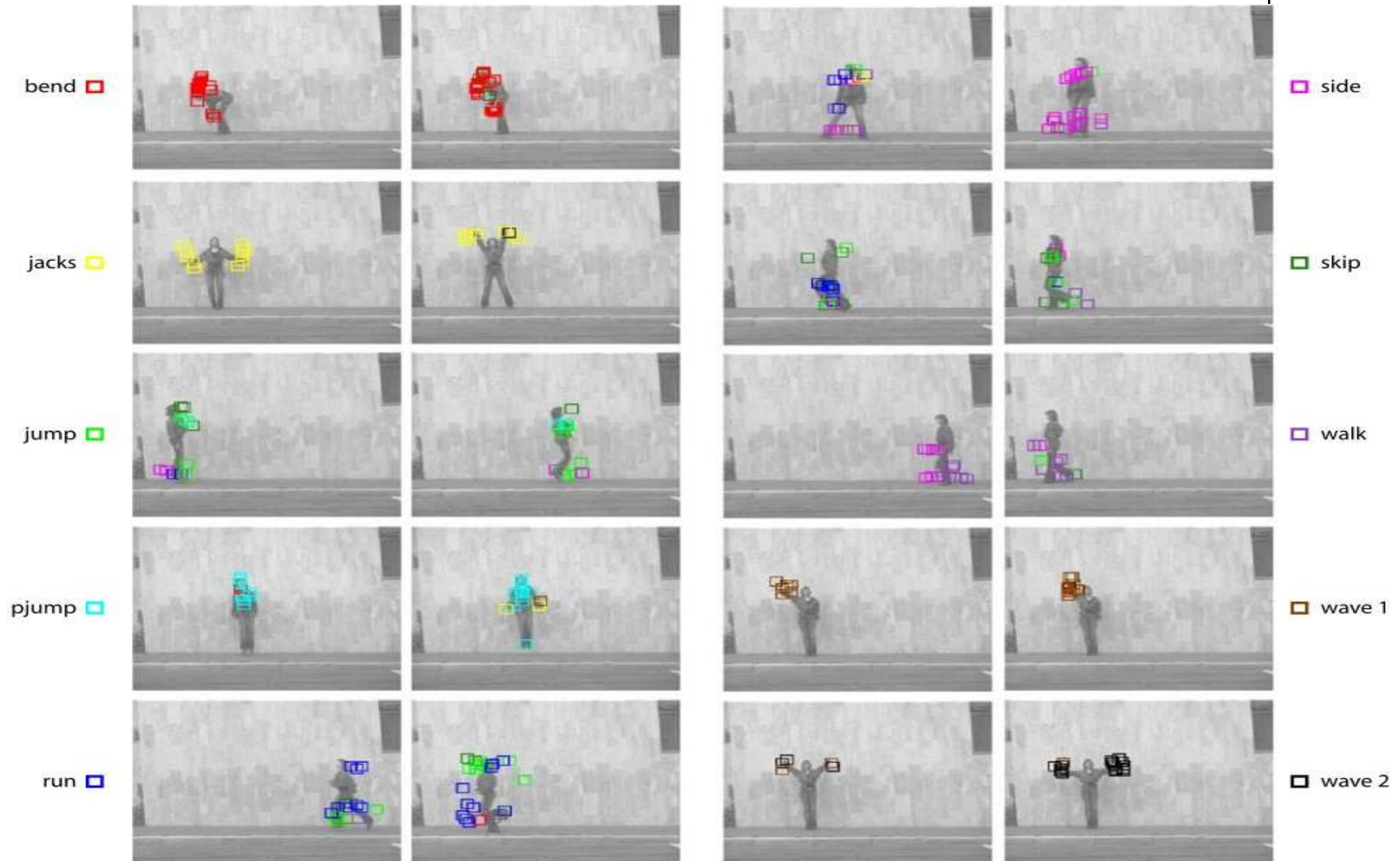


□ walking □ running □ jogging □ boxing □ hand clapping □ hand waving

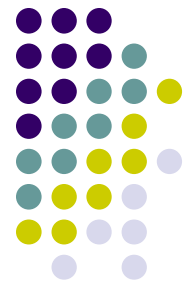
Experiments and Results: Weizmann Institute Human Action Dataset



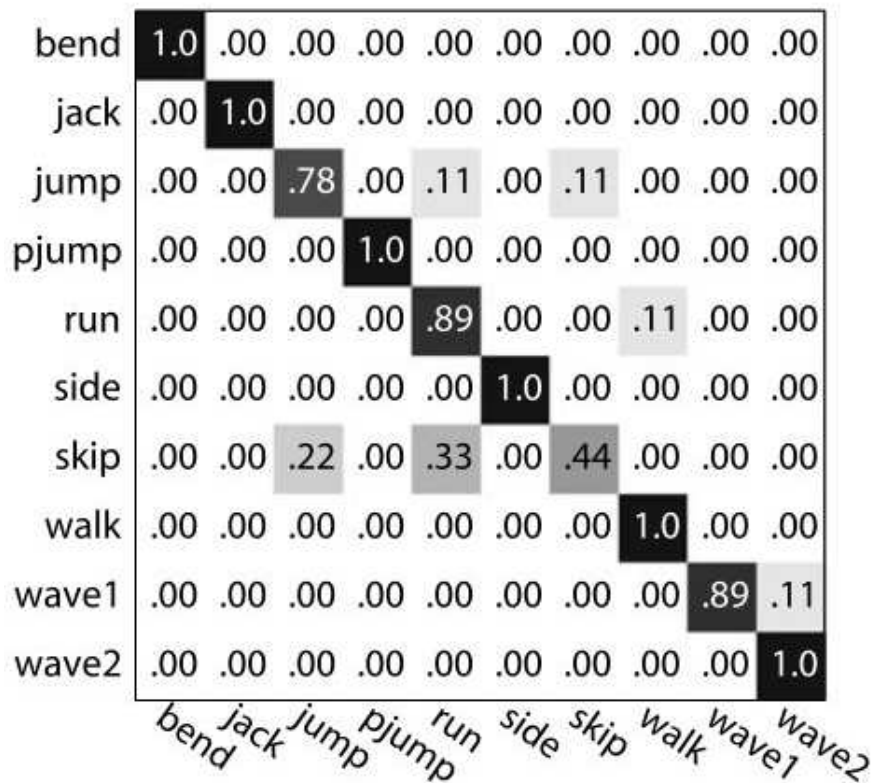
Experiments and Results: Weizmann Institute Human Action Dataset



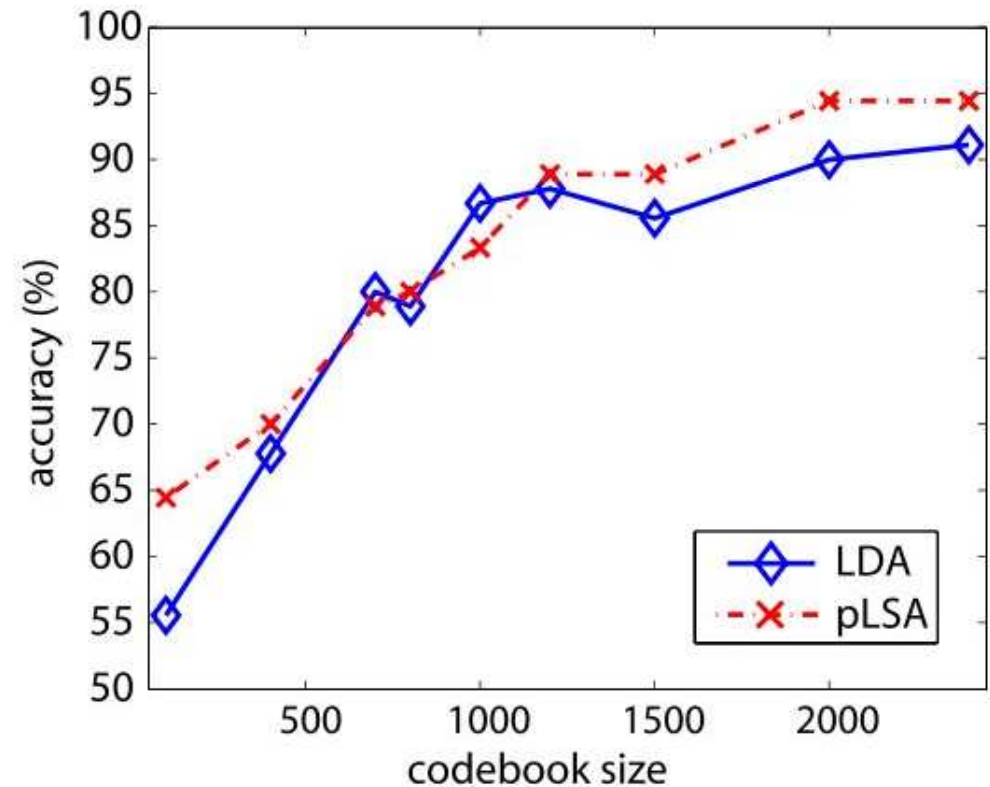
Experiments and Results: Weizmann Institute Human Action Dataset



- 10 action categories

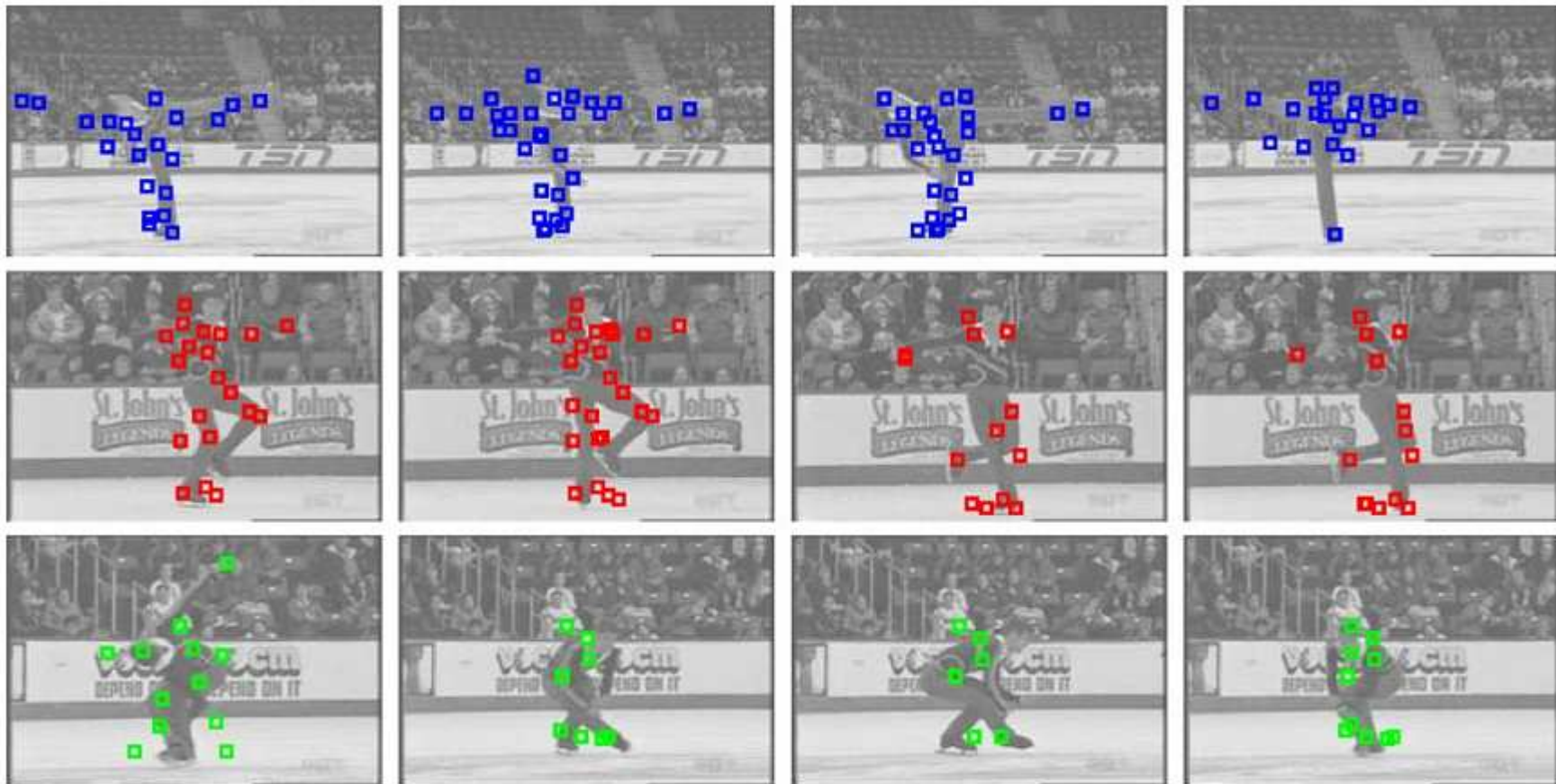
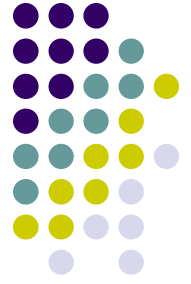


(a)



(b)

Experiments and Results: Figure Skating Database from Wang et al. 2006



□ stand-spin

□ sit-spin

□ camel-spin

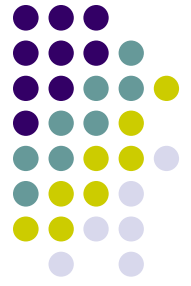
Experiments and Results :

Recognition and Localization of Multiple Actions in Long Videos

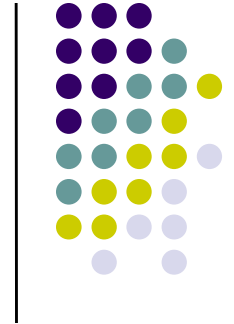


- Two scenarios:
 - Long sequence, one subject, several actions: sliding temporal window to create subsequences, determine action for each subsequence
 - Multiple simultaneous actions:
 - Identify most prominent actions (must be less than K , number of all learned actions)
 - Use k-means to cluster relevant words spatially
 - Each located word votes for its action, majority wins
 - Actions must be spatially separated!

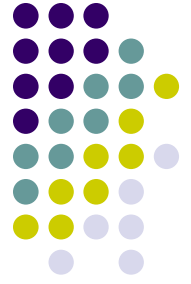
Experiments and Results : Recognition and Localization of Multiple Actions in Long Videos



Outline

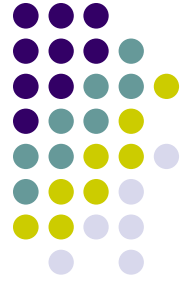


- Introduction
- Explaining the Model – ‘Bag of Video Words’
 - Codebook Generation
 - Learning the Action Models
 - Categorization and Localization
- Results
- Discussion



Recap

- A ‘bag of video words’ model
- Employs spatio-temporal detectors and descriptors
- pLSA / LDA models for unsupervised learning, recognition and localization



Future Directions

- Larger datasets and more complex videos
- Add discriminative model, to handle confusion between jogging and running
- Incorporate geometrical and temporal models knowledge

References



- [1] Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei, **Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words**, *Accepted for Oral Presentation At British Machine Vision Conference ([BMVC](#))*, Edinburgh, 2006.
- [2] Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei. **Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words**. *International Journal of Computer Vision (IJCV)*. September, 2008.
- [3] Piotr Dollár, Vincent Rabaud, Garrison Cottrell and Serge Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. VS-PETS 2005, Beijing, China.