Unsupervised Learning of Human Actions Using Spatial-Temporal Words

Juan Carlos Niebles[1,2], Hongcheng Wang[1], Li Fei-Fei[1]

[1]University of Illinois at Urbana – Champaign, Urbana, IL 61801, USA [2]Universidad del Norte, Barranquilla, Colombia

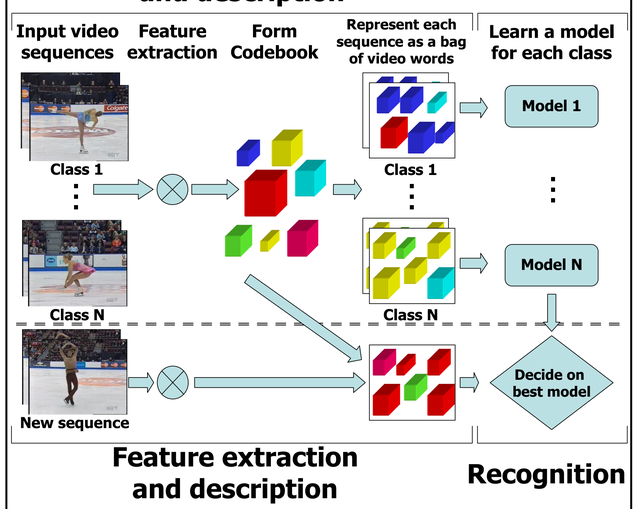
Summary

Problem statement: identifying and localizing different human actions in video sequences with moving background and moving camera.

Contributions:

- Unsupervised learning of actions using "bag of video words" representation
- Multiple action localization and categorization in a single video.
- Best reported performance on standard dataset.

Algorithm

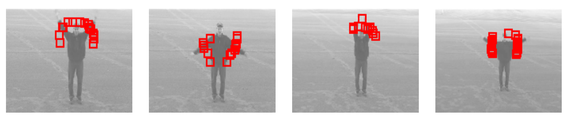


Training data

- KTH human motions data (6 classes)
- SFU figure skating data (3 classes)

Feature extraction

- Separable linear filters (2D Gaussian + 1D Gabor filters)
- A small video cube is extracted around each interest point



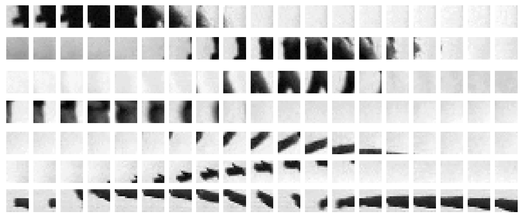
Feature representation

Feature description:

- Histogram of brightness gradient

Obtaining Codebook:

- K-means clustering of video word descriptors

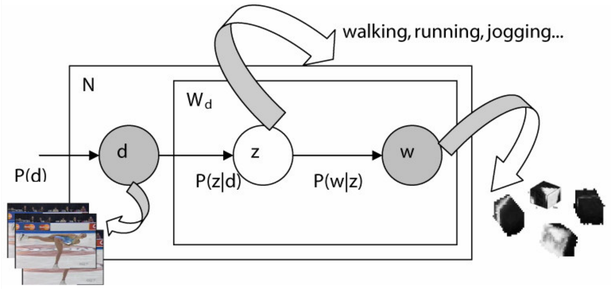


Representation:

- Histogram of video words from the codebook

Model

We deploy a pLSA model for video analysis.



d: input video z: action category w: video word

$$P(d_j, w_i) = P(d_j)P(w_i \mid d_j)$$

$$P(w_i \mid d_j) = \sum_{k=1}^{K} \underbrace{P(z_k \mid d_j)}_{\text{action category weights}} \underbrace{P(w_i \mid z_k)}_{\text{action category vectors}} \qquad K = \text{Number of Action Categories}$$

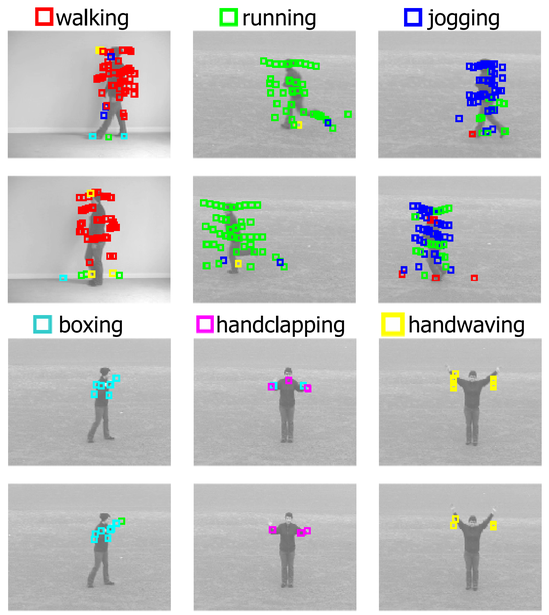
Classification

Given a new video and the learnt model, we can classify it as belonging to one of the action categories.

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

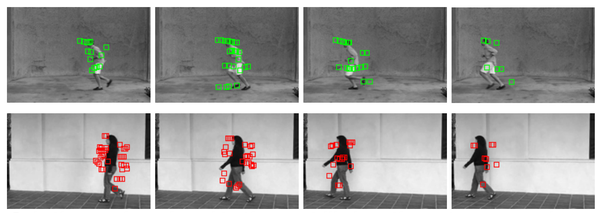
$$\text{action category} = \arg\max_k P(z_k \mid d_{test})$$

Exp I: 6-action classes of KTH dataset



Colored words according to their most likely action category.

Testing with the Caltech dataset



Only the words assigned to the most likely action category are shown.

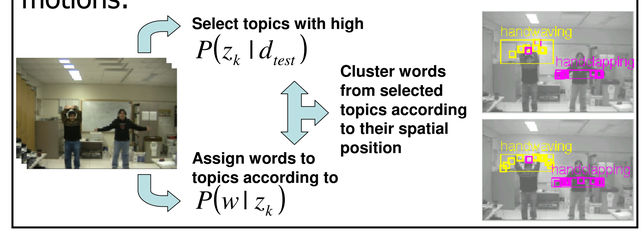
Performance:

Method	Recognition Accuracy %
Our method	81.50
Dollar et al.	81.17
Schuldt et al.	71.72
Ke et al.	62.96

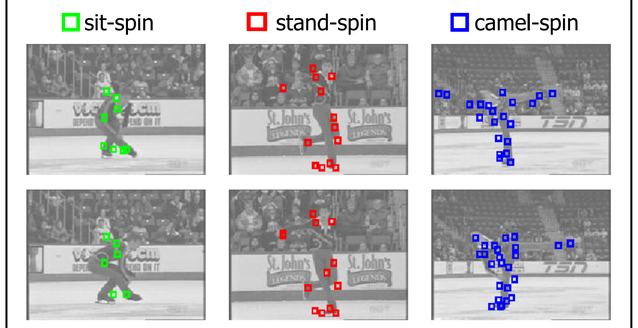
- Best Performance
- Multiple Actions
- Unlabeled training

Localization

Given a new video, we can localize multiple motions:

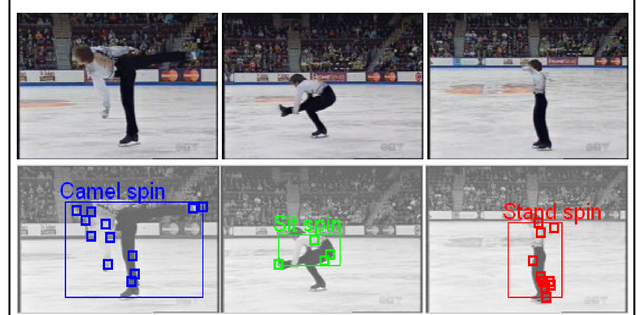


Exp II: 3-action classes of figure skating dataset

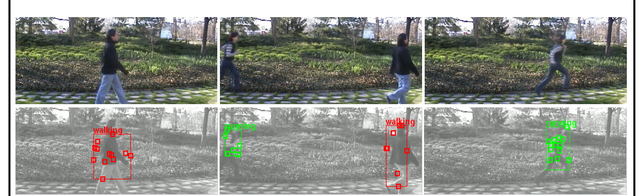


Only the words assigned to the most likely action category are shown.

Long and complex videos



3-class model: Results on a long video sequence



6-class model: Results on a complex natural video

Ref: [1] J.C. Niebles, H. Wang, L. Fei-Fei. Unsupervised Learning of Human Actions Using Spatial-Temporal Words. *Submitted*. 2006.

[2] T. Hofmann. Probabilistic latent semantic indexing. In SIGIR, 1999.