Combining Appearance and Motion for Human Action Classification in Videos

Paramveer S. Dhillon[1], Sebastian Nowozin[2], Christoph Lampert[2]

[1] University of Pennsylvania, Philadelphia, PA, U.S.A. [2] MPI for Biological Cybernetics, Tübingen, Germany

Overview

Open Problems in Human Action Recognition:

- Occlusion, Cluttered background, Camera motion
- **Strong Variations in Appearance of the Actors**

Observation:

> **Appearance and Motion are orthogonal concepts that can and should be modeled separately.**

Previous Work

Sparse spatio-temporal interest points and local descriptors that represent local motion and appearance, e.g.

- **Space-time interest points [2]**
  - Accelerating motion is *interesting*,
  - Match against walking model.
- **Behaviour recognition via sparse spatio temporal features [1]**
  - Periodic motion is *interesting*,
  - Classify histograms of sparse spatio-temporal interest points.
- **Unsupervised learning of human action categories using spatio-temporal words [3]**
  - Histograms of spatio-temporal words as [1]
  - Unsupervised model based on pLSA.

Our Approach

We extract *appearance* and *motion* into separate representations:

- **Motion: trajectories of particle filter cluster centers**
  - Means shift clustering of particles in a particle filter.
  - Particle filter setup:
    - Prior Distribution $p(x_0)$: squashed Difference of Gaussian filter.
    - Transition Model $(p(x_t|x_{t-1})$: second order autoregressive model.
    - Observation Model $p(y_t|x_t) = e^{-\lambda \sum_{i=1}^{n^2} |I_t^i - I_{t-1}^i|^2}$ Gaussian (particles follow local image regions)
- **Appearance: descriptors along motion trajectories.**
  - One log-polar binned histogram per cluster mode.
  - Histogram of number of particles in each bin.
- **Cluster descriptors and form "bag of words" histograms, classify with linear Support Vector Machine**

Our Approach (contd.)

Illustration of log-polar binned histogram descriptor:

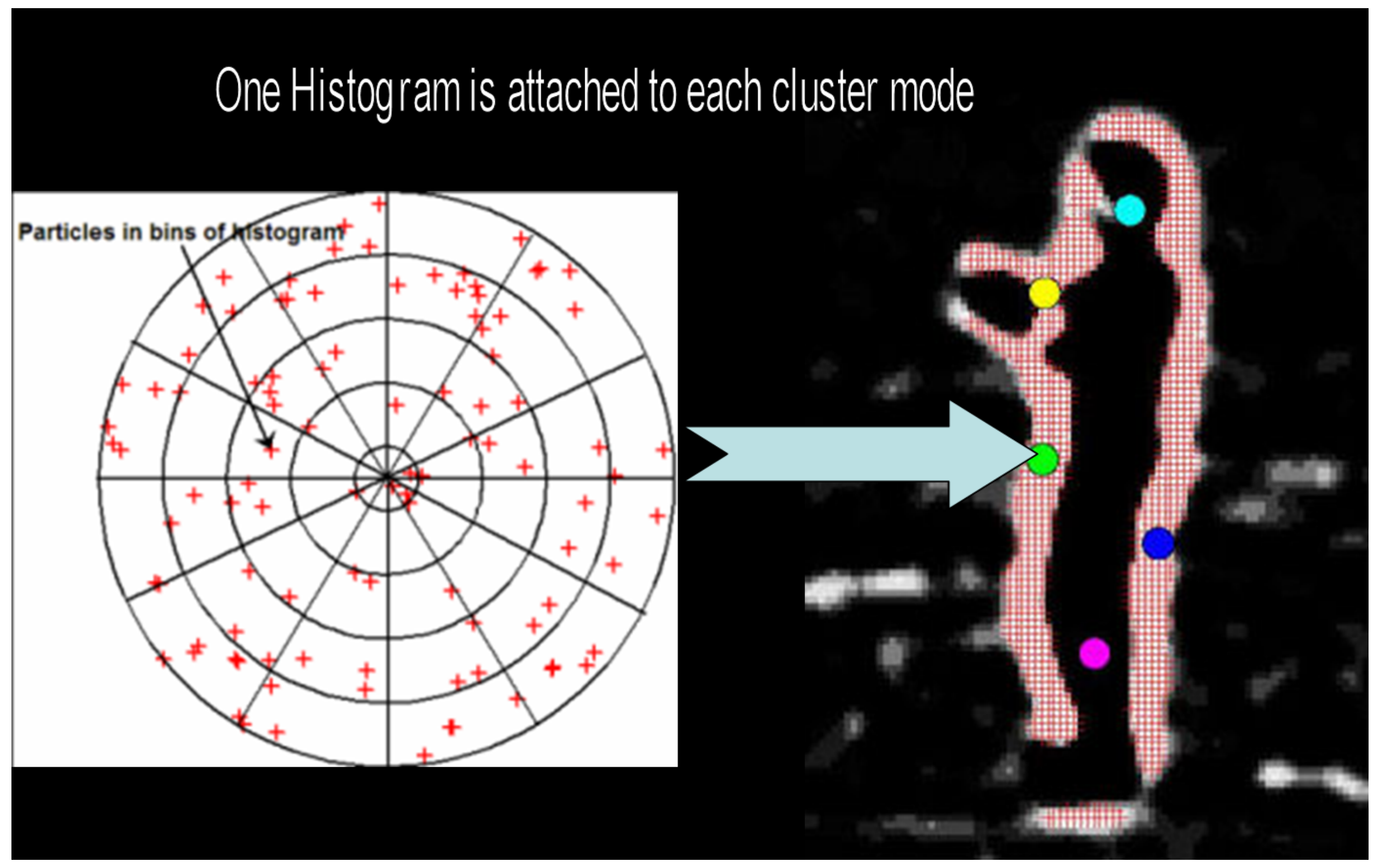
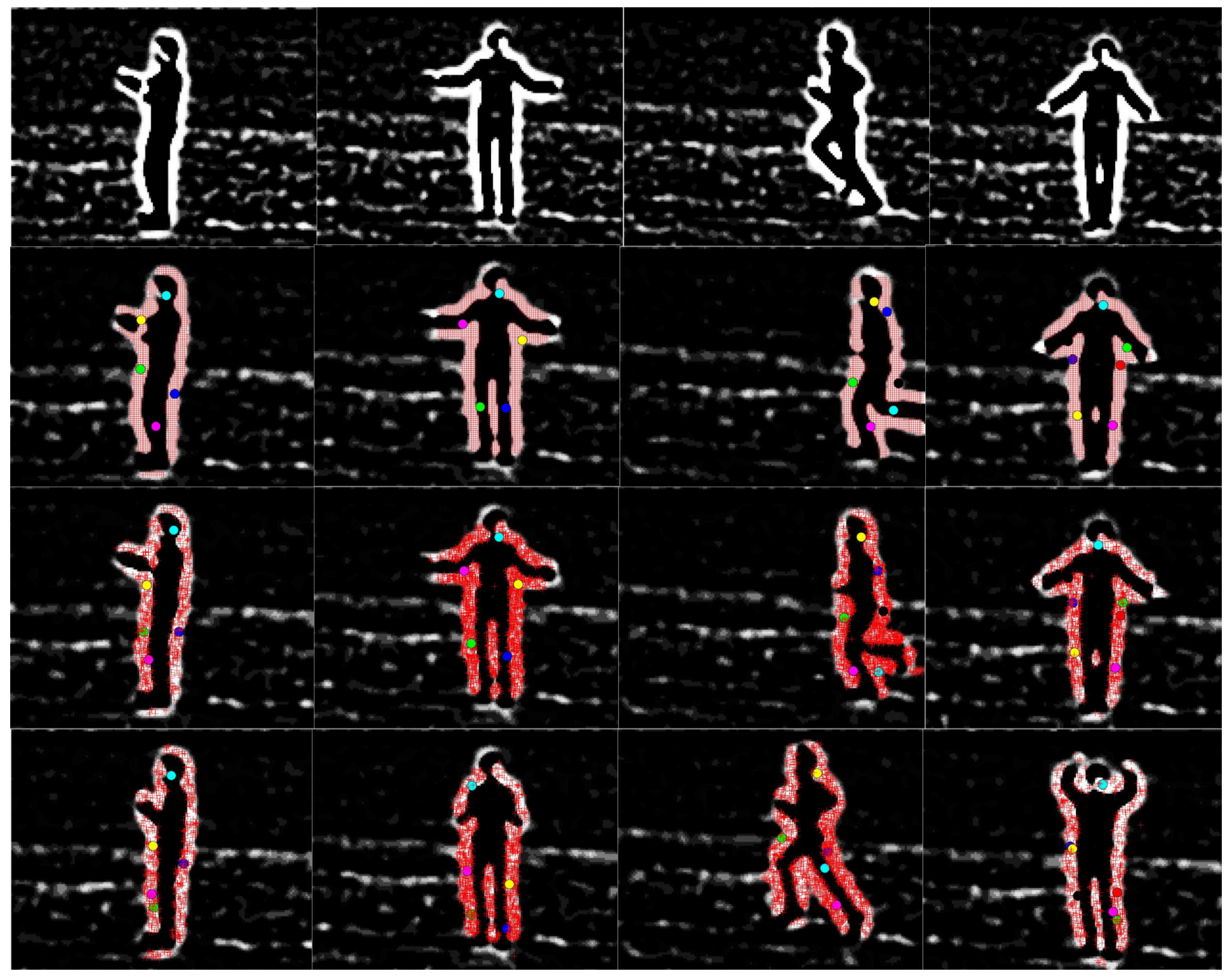

One Histogram is attached to each cluster mode
Particles in bins of histogram

Illustration of drifting particles and cluster modes:



Experimental Results

KTH Dataset with standard test-train splits:

- Accuracy: $84.67\% \pm 0.56\%$ averaged over 10 runs.
- No confusion between *local motion* patterns: *boxing*, *handwaving*, *handclapping* and *global motion* patterns: *walking*, *running*, *jogging*.

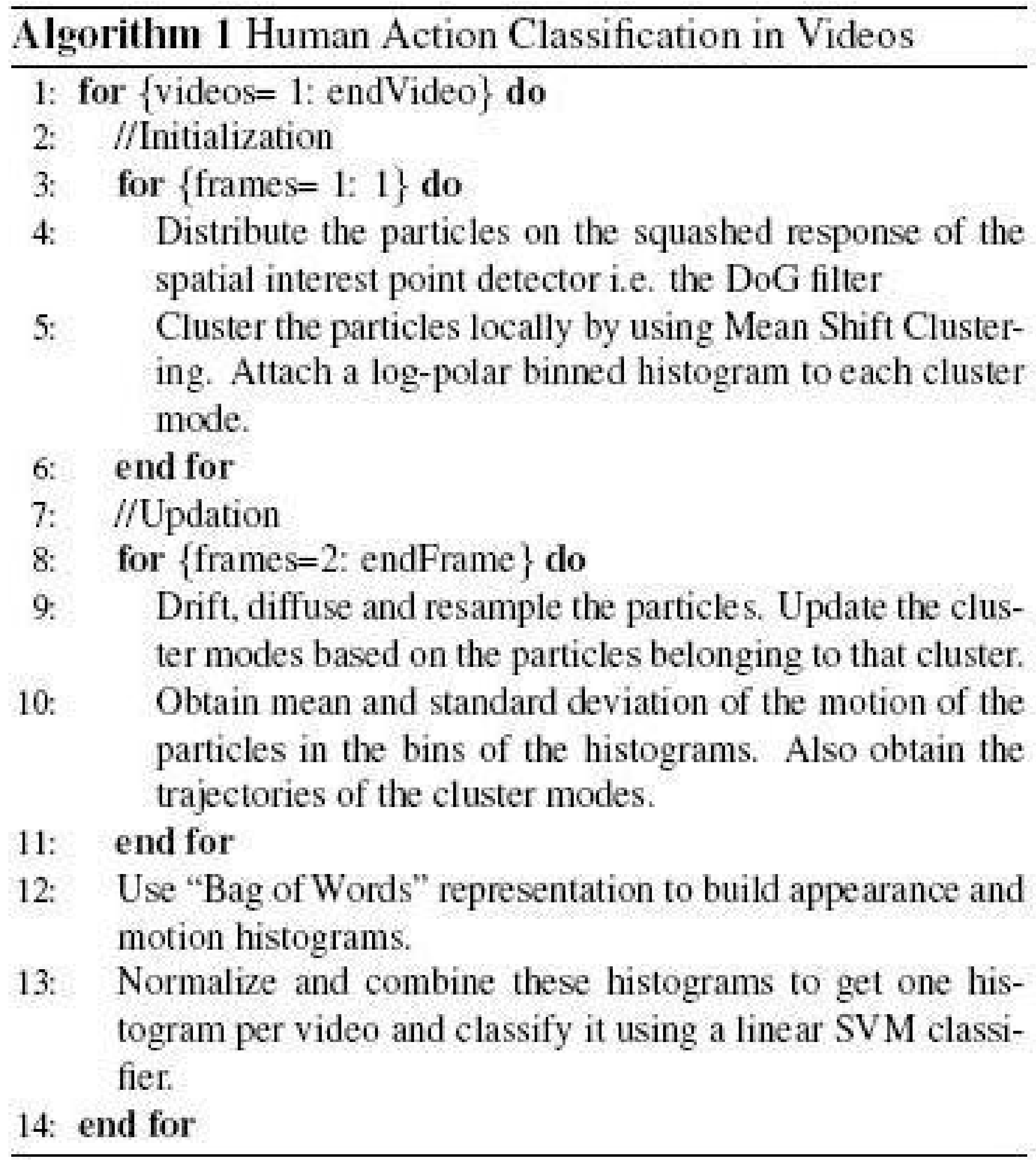
Weizmann Dataset with 70-30 split:

- Accuracy: $89.9\% \pm 1.61\%$ averaged over 10 runs.

Overall not as good as best published results, but promising given the simple appearance descriptor and classifier.

Algorithm

Our algorithm in psuedo-code:



Algorithm 1 Human Action Classification in Videos

```
for {videos= 1: endVideo} do
    //Initialization
    for {frames= 1: 1} do
        Distribute the particles on the squashed response of the spatial interest point detector i.e. the DoG filter
        Cluster the particles locally by using Mean Shift Clustering. Attach a log-polar binned histogram to each cluster mode.
    end for
    //Updation
    for {frames=2: endFrame} do
        Drift, diffuse and resample the particles. Update the cluster modes based on the particles belonging to that cluster.
        Obtain mean and standard deviation of the motion of the particles in the bins of the histograms. Also obtain the trajectories of the cluster modes.
    end for
    Use "Bag of Words" representation to build appearance and motion histograms.
    Normalize and combine these histograms to get one histogram per video and classify it using a linear SVM classifier.
end for
```

Summary

- **Separate Appearance and Motion for Action Classification**
  - goal 1: more invariant to actor appearance
  - goal 2: more discriminative to action performed
- Motion information abstracted by cluster modes trajectories
- Appearance information by local distribution of tracking particles
- Separated approach allows good discriminatation between activities with local body motion like (handclapping, handwaving etc.) and activities with global body motion (walking, running, ...)

References

[1] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS) at ICCV*, 2005.

[2] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)*, pages 432–439, 2003.

[3] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, 2008.