The paper provides an unsupervised learning approach of multiple human actions by using a bag of word representation. The goal is to localize and categorize multiple actions in a video. A good aspect of the paper is to provide a comprehensive background of human action recognition in video and comparison of different feature representations. Two methods, LDA and pLSA are used to learn and recognize the action in an unsupervised fashion.

A spatial-temporal interest point is extracted from video frame by a linear filter. A video patch is represented by six interest points to provide short range of local information of human action. So the representation is not totally spatial invariant. A spatial-temporal pattern has a short range because such representation only provides local information. The spatial-temporal cube is smoothed to normalize the variance in brightness.

pLSA is performed with an unobserved variable z. Z is estimated by producing the maximum e conditional probability given observed video and the number of occurrences of spatial-temporal word. Each interest point in the test data will be assigned to a spatial-temporal word. And a local topic is produced by counting the occurrence of spatial-temporal words. LDA introduces two variable γ and θ. Θ fixes the topic distribution of topic distribution over the video patches. γ and θ are free variables which are untraceable. They are selected by producing MLE of posterior probability. The idea is using the unobserved topics z to explain the observed video data.

The algorithm works well on the datasets. But it doesn’t show the power to show the power of detecting relative motion activities. Even in the skating dataset, the viewpoint of human action changes, but there is less horizontally relative motion between human and camera. Even the complex videos it classifies the walking and running but it still needs more explanation about how to detect or classify human action with relative motion respect to camera.