

# Unsupervised Learning of Human Action Categories

Juan Carlos Nieves<sup>1,2</sup>, Hongcheng Wang<sup>1</sup>, Li Fei-Fei<sup>1</sup>

<sup>1</sup>University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

<sup>2</sup>Universidad del Norte, Barranquilla, Colombia

Email: {jnieble2,hwang13,feifeili}@uiuc.edu

## DESCRIPTION

Imagine a video taken on a sunny beach, can a computer automatically tell what is happening in the scene? Can it identify different human activities in the video, such as water surfing, people walking and lying on the beach? To automatically classify or localize different actions in video sequences is very useful for a variety of tasks, such as video surveillance, object-level video summarization, video indexing, digital library organization, etc. However, it remains a challenging task for computers to achieve robust action recognition due to cluttered background, camera motion, occlusion, and geometric and photometric variances of objects. For example, in a live video of a skating competition, the skater moves rapidly across the rink, and the camera also moves to follow the skater. With moving camera, non-stationary background, and moving target, few vision algorithms could identify, categorize and localize such motions well. In addition, the challenge is even greater when there are multiple activities in a complex video sequence (Figure 1).

We present a video demo for our novel unsupervised learning method for human action categories [1]. A video sequence is represented as a collection of spatial-temporal words by extracting space-time interest points. The algorithm learns the probability distributions of the spatial-temporal words and intermediate topics corresponding to human action categories automatically using a probabilistic Latent Semantic Analysis (pLSA) model [4]. The learned model is then used for human action categorization and localization in a novel video, by maximizing the posterior of action category (topic) distributions. The contributions of this work are as follows:

- *Unsupervised learning of actions using ‘video words’ representation.* We deploy a pLSA model with ‘bag of video words’ representation for video analysis;
- *Multiple action localization and categorization.* Our approach is not only able to classify different actions, but also to localize different actions simultaneously in a novel and complex video sequence.

We test our algorithm on two datasets: the KTH human action dataset [3] and a recent dataset of figure skating actions [2]. These datasets contain videos of cluttered background, moving camera, and multiple actions. Our results on action recognition performance are on par or slightly better than the best reported results in the literature. In addition, our algorithm can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

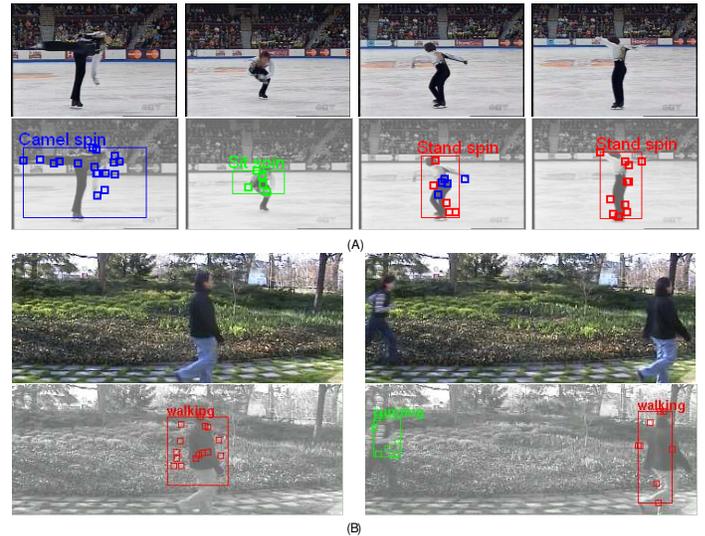


Fig. 1

(A) AN EXAMPLE OF A LONG FIGURE SKATING SEQUENCE; (B) AN EXAMPLE OF MULTIPLE ACTIONS RECOGNITION AND LOCALIZATION IN ONE VIDEO SEQUENCE. THE FIGURE IS BEST VIEWED IN COLOR AND WITH PDF MAGNIFICATION.

We tested several long figure skating sequences as well as our own complex video sequences as shown in Figure 1. For the long skating video sequences, we used a windowed sequence around each frame and identified significant actions using the learned three-class model of the figure skating dataset [2]. Then that frame was labeled using the action category identified. Another example is based on the six-class model learned from the KTH human action dataset, for multiple action recognition and localization in Figure 1 (B). We first identified how many action categories are significant. Then we applied K-means to find that number of clusters. By counting the number of video words within each cluster with respect to the action categories, we recognized and localized the actions within that video.

## REFERENCES

- [1] J. C. Nieves, H. Wang, L. Fei-Fei, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” *submitted*, 2006
- [2] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, “Unsupervised Discovery of Action Classes,” *CVPR (accepted)*, 2006
- [3] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach,” *ICPR*, 2004
- [4] T. Hofmann, “Probabilistic Latent Semantic Indexing,” *SIGIR*, 1999