Unsupervised Learning of Human Action Categories 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
Email: {jnieble2,hwang13,feifeili}@uiuc.edu

Abstract

We present a novel unsupervised learning method for human action categories. A video sequence is represented as a collection of spatial-temporal words by extracting space-time interest points. The algorithm automatically learns the probability distributions of the spatial-temporal words and intermediate topics corresponding to human action categories. This is achieved by using a probabilistic Latent Semantic Analysis (pLSA) model. Given a novel video sequence, the model can categorize and localize the human action(s) contained in the video. We test our algorithm on two challenging datasets: the KTH human action dataset and a recent dataset of figure skating actions. Our results are on par or slightly better than the best reported results. In addition, our algorithm can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

1 Introduction

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, beach volleyballs, or people taking a walk along the beach? To automatically categorize 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 cameras, non-stationary background, and moving target, few vision algorithms could identify, categorize and localize such motions well (Figure 1(b)). In addition, the challenge is even greater when there are multiple activities in a complex video sequence (Figure 1(d)). In this paper, we will present an algorithm that aims to account for both of these scenarios.

A lot of previous work has been presented to address these questions. One popular approach is to apply tracked motion trajectories of body parts to action recognition [15, 21, 1]. This is done with much human supervision and the robustness of the algorithm is highly dependent on the tracking system. Ke et al. [13] apply spatio-temporal volumetric feature that efficiently scan video sequences in space and time. Another approach is to use local space-time patches of videos [8]. Laptev et al. present a space-time

Figure 1: Example images from video sequences (a) KTH dataset; (b) Figure skating dataset; (c) Caltech dataset; (d) Our own complex video sequence.

interest point detector based on the idea of the Harris and Förstner [14] interest point operators. They detect local structures in space-time where the image values have significant local variations in both dimensions. The representation has been successfully applied to human action recognition combined with an SVM classifier [17]. Dollár et al. [7] propose an alternative approach to detect sparse space-time interest points based on separable linear filters for behavior recognition. Local space-time patches, therefore, have been proven useful to provide semantic meaning of video events by providing a compact and abstract representation of patterns. While these representations indicate good potentials, the modeling and learning frameworks are rather simple in the previous work [17, 7], posing a problem toward handling more challenging situations such as multiple action recognition.

Another category of work is based on a probabilistic graphical model framework in action categorization/recognition. Song et al. [19] and Fanti et al. [9] represent the human action model as a triangulated graph. Multiple cues such as position, velocities and appearance have been considered in learning and detection phases. Their idea is to map the human body parts in a frame-by-frame manner instead of utilizing space-time cubes for action recognition. Boiman and Irani [4] recently propose to extract ensemble of local video patches to localize irregular action behavior in videos. Dense sampling of the patches is necessary in their approach and therefore the algorithm is very time-consuming. It is not suitable for action recognition purpose due to the large amount of video data commonly presented in these settings. Another work named video epitomes is proposed by Cheung et al. [5]. They model the space-time cubes from a specific video by a generative model. The learned model is a compact representation of the original video, therefore this approach is suitable for video super-resolution and video interpolation, but not for recognition.

In this paper, we propose a generative graphical model approach to learn and recognize human actions in video, taking advantage of the robust representation of spatial temporal words and an unsupervised approach during learning. Our method is motivated by the recent success of object detection/classification [18, 6] or scene categorization [10] from unlabeled static images. Two related models are generally used, i.e., probabilistic Latent Semantic Analysis (pLSA) by Hofmann [12] and Latent Dirichlet Allocation (LDA) by Blei et al. [3]. In this paper, we choose to build a pLSA model for video analysis by taking the advantages of the powerful representation and the great flexibility of the generative graphical model. The contributions of this work are as follows:

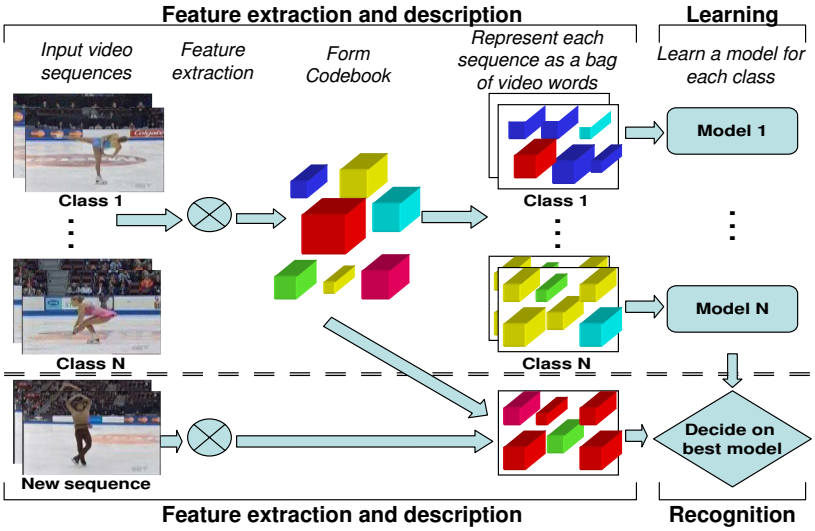

Figure 2: Flowchart of our approach. To represent motion patterns we first extract local space-time regions using the space-time interest points detector [7]. These local regions are then clustered into a set of video codewords, called codebook. Probability distributions and intermediate topics are learned automatically using a pLSA graphical model.

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

The rest of the paper is organized in the following way. In Section 2, we describe our approach in more details, including spatial-temporal feature representation, brief overview of the pLSA model in our context, and the specifics of the learning and recognition procedures. In Section 3, we present the experimental results on human action recognition using real datasets, and also compare our performance with other methods. Multiple action recognition and localization results are presented to validate the learned model. Finally, Section 4 concludes the paper.

2 Our Approach

Given a collection of unlabeled videos, our goal is to automatically learn different classes of actions present in the data, and apply the learned model to action categorization and localization in the new video sequences. Our approach is illustrated in Figure 2.

2.1 Feature Representation from Space-Time Interest Points

As Figure 2 illustrates, we represent each video sequence as a collection of spatial-temporal words by extracting space-time interest points. There is a variety of methods for interest points detection in images [16]. But less work has been done on space-time interest point detection in videos. Blank et al. [2] represent actions as space-time shapes and extracted space-time features such as local space-time saliency, action dynamics, shape structures and orientation for action recognition. Laptev and Lindeberg [14] propose an extended version of the interest points detection in the spatial domain [11] into space-time

domain by requiring image values in space-time to have large variations in both dimensions. As noticed in [7] and from our experience, the interest points detected using the generalized space-time interest point detector are too sparse to characterize many complex videos such as figure skating sequences in our experiments. Therefore, we use the separable linear filter method in [7]. Here we give a brief review of this method.

Assuming a stationary camera or a process that can account for camera motion, separable linear filters are applied to the video to obtain the response function as follows:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $g(x,y;\sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions (x,y), and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally, which are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. The two parameters σ and τ correspond to the spatial and temporal scales of the detector respectively. In all cases we use $\omega = 4/\tau$, effectively giving the response function R. To handle multiple scales, one must run the detector over a set of spatial and temporal scales. For simplicity, we run the detector using only one scale and rely on the codebook to encode the few changes in scale that are observed in the dataset.

It was noted in [7] that any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response. However, regions undergoing pure translational motion, or without spatially distinguishing features will not induce a strong response. The space-time interest points are extracted around the local maxima of the response function. Each patch contains the volume contributed to the response function, i.e., its size is approximately six times the scales along each dimension. To obtain a descriptor for each spatial-temporal cube, we calculate the brightness gradient and concatenate it to form a vector. This descriptor is then projected to a lower dimensional space using PCA. In [7], different descriptors have been used, such as normalized pixel values, brightness gradient and windowed optical flow. We found that both the gradient descriptor and the optical flow descriptor are equally effective in describing the motion information. For the rest of the paper, we will employ results obtained with gradient descriptors.

2.2 Learning the Action Models: Latent Topic Discovery

In this section, we will describe the pLSA graphical model in the context of video modeling. We follow the conventions introduced in [12, 18].

Suppose we have $N(j = 1,\cdots,N)$ video sequences containing video words from a vocabulary of size $M(i = 1,\cdots,M)$. The corpus of videos is summarized in an M by N co-occurrence table N, where $n(w_i,d_j)$ stores the number of occurrences of a word w_i in video d_j. In addition, there is a latent topic variable z_k associated with each occurrence of a word w_i in a video d_i. Each topic corresponds to a motion category.

The joint probability $P(w_i,d_j,z_k)$ is assumed to have the form of the graphical model shown in Figure 3.

$$P(d_j,w_i) = P(d_j)P(w_i|d_j) \tag{2}$$

Given that the observation pairs (d_j,w_i) are assumed to be generated independently, we can marginalize over topics z_k to obtain the conditional probability $P(w_i|d_j)$:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k) \tag{3}$$

where $P(z_k|d_j)$ is the probability of topic z_k occurring in video d_j; and $P(w_i|z_k)$ is the probability of video word w_i occurring in a particular action category z_k. K is the total number of latent topics, hence the number of action categories in our case.

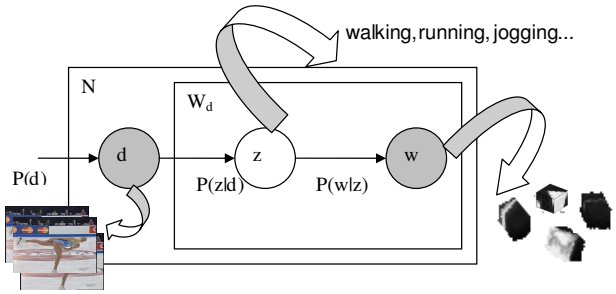

Figure 3: pLSA graphical model. Nodes are random variables. Shaded ones are observed and unshaded ones are unobserved. The plates indicate repetitions.

Intuitively, this model expresses each video sequence as a convex combination of K action category vectors, i.e., the video-specific word distributions $P(w_i|d_j)$ are obtained by a convex combination of the aspects or action category vectors $P(w_i|z_k)$. Videos are characterized by a specific mixture of factors with weights $P(z_k|d_j)$. This amounts to a matrix decomposition with the constraint that both the vectors and mixture coefficients are normalized to make them probability distributions. Essentially, each video is modeled as a mixture of action categories - the histogram for a particular video being composed from a mixture of the histograms corresponding to each action category.

We then fit the model by determining the action category vectors which are common to all videos and the mixture coefficients which are specific to each video. In order to determine the model that gives the high probability to the video words that appear in the corpus, a maximum likelihood estimation of the parameters is obtained by maximizing the objective function using an Expectation Maximization (EM) algorithm:

$$\prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)} \tag{4}$$

where $P(w_i|d_j)$ is given by Equation 3.

2.3 Categorization and Localization of Actions in a Testing Video

Our goal is to categorize new video sequences using learned action category models. We have obtained the action category specific video word distributions $P(w|z)$ from a different set of training sequences. When given a new video, the unseen video is 'projected' on the simplex spanned by the learned $P(w|z)$. We need to find the mixing coefficients $P(z_k|d_{test})$ such that the KL divergence between the measured empirical distribution $\tilde{P}(w|d_{test})$ and $P(w|d_{test}) = \sum_{k=1}^{K} P(z_k|d_{test})P(w|z_k)$ is minimized [12, 18]. Similarly to the learning scenario, we apply an EM algorithm to find the solution.

Furthermore, we are also interested in localizing multiple actions in a single video sequence. Though our 'bag-of-video-words' model itself does not explicitly represent spatial relationship of local video regions, it is sufficiently discriminative to localize different motions within each video. This is similar to the approximate object segmentation case in [18]. The pLSA model models the posteriors

$$P(z_k|w_i,d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)} \tag{5}$$

For the video word around each interest point, we can label the topics for each word by finding the maximum posteriors $P(z_k|w_i,d_j)$. Then we can localize multiple actions

Table 1: Comparison of different methods

methods	recognition accuracy (%)	learning	multiple actions
Our method	81.50	unlabeled	Yes
Dollár et al. [7]	81.17	labeled	No
Schuldt et al. [17]	71.72	labeled	No
Ke et al. [13]	62.96	labeled	No

corresponding to different action categories.

3 Experimental Results

We test our algorithm using two datasets: KTH human motion dataset [17], and figure skating dataset [20]. These datasets contain videos of cluttered background, moving cameras, and multiple actions. We can handle the noisy feature points arisen from dynamic background and moving cameras by utilizing the probabilistic graphical model (pLSA), as long as the background does not amount to an overwhelming number of feature points. In addition, we demonstrate multiple actions categorization and localization in a set of new videos collected by the authors. We present the datasets and experimental results in the following sections.

3.1 Recognition and Localization of Single Actions

3.1.1 Human Action Recognition and Localization Using KTH data

KTH human motion dataset is the largest available video sequence dataset of human actions [17]. Each video has only one action. The dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in different scenarios of outdoor and indoor environment with scale change. It contains 598 sequences. Some sample images are shown in Figure 1.

We build video codewords from two videos of each action from three subjects. Because the number of space-time patches used to extract the video codewords is usually very large, we randomly select a smaller number of space-time patches (around 60,000) to accommodate the requirements of memory. We perform leave-one-out cross-validation to test the efficacy of our approach in recognition, i.e., for each run we learn a model from the videos of 24 subjects (except those used to build codewords), and test the videos of the remaining subject. The three subjects used for forming the codebook are excluded from the testing. The result is reported as the average of 25 runs.

The confusion matrix for a six-class model for the KTH dataset is given in Figure 4(a) using 500 codewords. It shows large confusion between ‘running’ and ‘jogging’, as well as ‘handclapping’ and ‘boxing’. This is consistent with our intuition that similar actions are more easily confused with each other, such as those involving hand motions or leg motions. We test the effect of the number of video codewords on recognition accuracy, as illustrated in Figure 4(b). As the size of codebook increases, the classification rate peaks at around 500. We also compare our results with the best results from [7]) (performance average = 81.17%) using Support Vector Machine (SVM) with the same experimental settings. Our results by unsupervised learning are on par with the current state-of-the-art results obtained by fully supervised training. The comparison of different methods is listed in Table 1. We also test the LDA model [3] on this dataset, and find that pLSA is slightly better than LDA in recognition performance with the same number of codewords.

We apply the learned model to localize the actions for test videos in KTH dataset in Figure 5. We also test our action localization using the same model for the Caltech

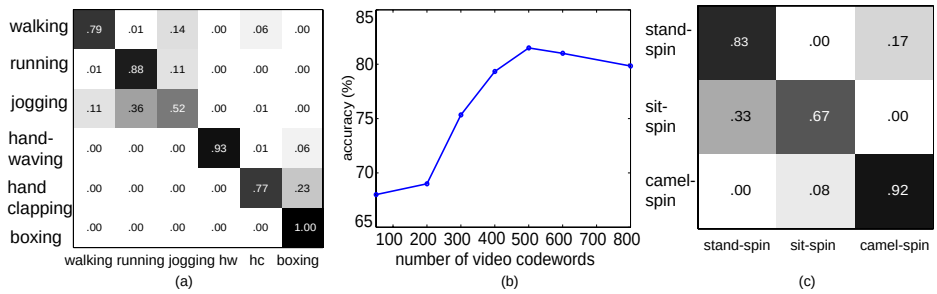

Figure 4: (a) confusion matrix for the KTH dataset using 500 codewords (performance average = 81.50%); horizontal lines are ground truth, and vertical columns are model results; (b) classification accuracy vs. codebook size for the KTH dataset; (c) confusion matrix for the figure skating dataset using 1200 codewords (performance average = 80.67%)

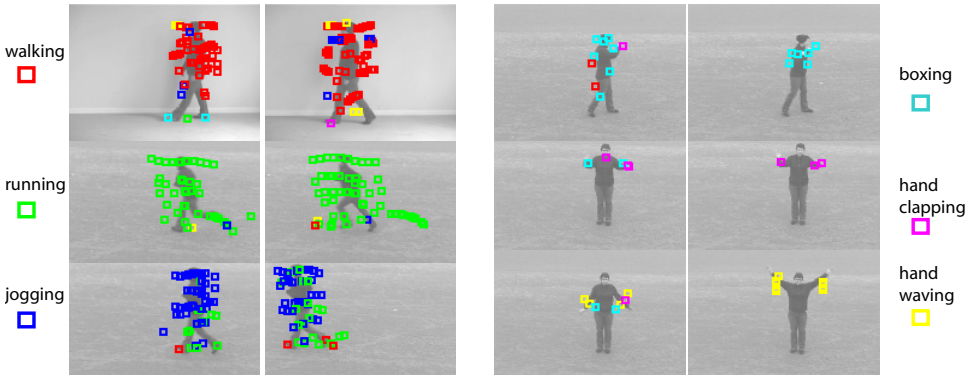

Figure 5: The action categories are embedded into video frames using different colors. Most spatial-temporal words are labeled by the corresponding action color for each video. The figure is best viewed in color and with PDF magnification.

human motion dataset [19] as shown in Figure 6(a). Most of the action sequences from this dataset can be correctly recognized. For the clarity of presentation, we only draw the video words of the most probable topic with their corresponding color.

3.1.2 Recognition and Localization of Figure Skating Actions

We use the figure skating dataset in [20][1]. We adapt 32 video sequences of 7 people each with three actions: stand-spin, camel-spin and sit-spin, as shown in Figure 1.

We build video codewords from the videos of six persons. We again perform leave-one-out cross-validation to test the efficacy of our approach in recognition, i.e., for each run we learn a model from the videos of six subjects, and test those of the remaining subject. The result is reported as the average of seven runs. The confusion matrix for a three-class model for the figure skating dataset is shown in Figure 4(c) using 1200 codewords. The larger size of codewords is useful to avoid overfitting of the generative model. The learned 3-class model is also used for action localization as shown in Figure 6(b).

[1]This work addresses the problem of motion recognition from still images. There is much other work to model motion in still images, which is out of the scope of this paper.

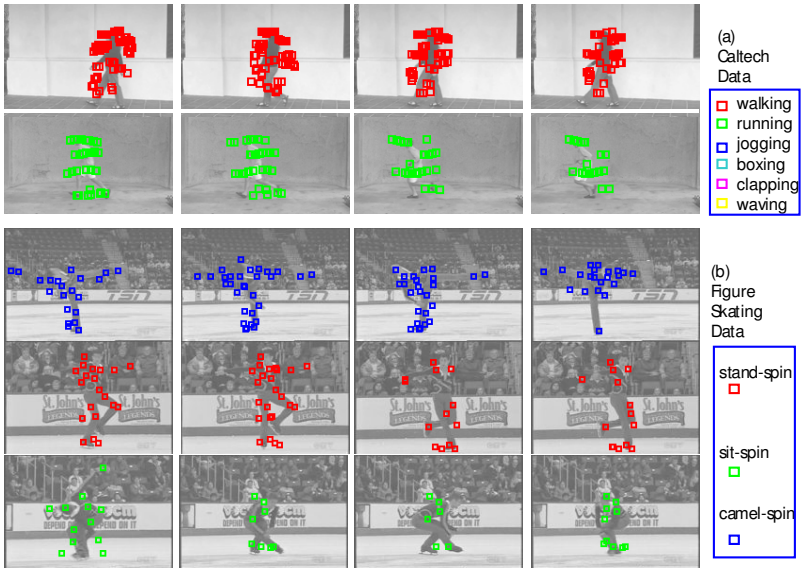

Figure 6: (a) Examples from Caltech dataset with color-coded actions; (b) Examples from figure skating dataset with color-coded actions. The figure is best viewed in color and with PDF magnification.

3.2 Recognition and Localization of Multiple Actions in a Long Video Sequence

One of the main goals of our work is to test how well our algorithm could identify multiple actions within a video sequence. For this purpose, we test several long figure skating sequences as well as our own complex video sequences.

For multiple actions in a single sequence, we first identify how many action categories are significantly induced by $P(z_k|w_i,d_j)$. Then we apply 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 recognize the actions within that video. The bounding box is plotted according to the principle axis and eigen-values induced by the spatial distribution of video words in each cluster. Figure 7 illustrates examples of multiple actions recognition and localization in one video sequence using the learned six-class model.

For the long skating video sequences, we extract a windowed sequence around each frame and identify significant actions using the learned three-class model. Then that frame is labeled as the identified action category. Figure 7 shows examples of action recognition in a long figure skating sequence. The three actions, i.e., stand-spin, camel-spin and sit-spin, are correctly recognized and labeled using different colors. (Please refer to the link of video demo: http://visionlab.ece.uiuc.edu/niebles/humanactions.htm)

4 Conclusion

In this paper, we have presented an unsupervised learning approach, i.e., a 'bag-of-video-words' model combined with a space-time interest points detector, for human action categorization and localization. Using two challenging datasets, our experiments validate the proposed model in classification performance. Our algorithm can also localize multiple actions in complex motion sequences containing multiple actions. The results are promising, though we acknowledge the lack of large and challenging video datasets to

thoroughly test our algorithm, which poses an interesting topic for future investigation.

References

[1] Ankur Agarwal and Bill Triggs. Learning to track 3d human motion from silhouettes. In *International Conference on Machine Learning*, pages 9–16, Banff, July 2004.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 2003.

[4] Oren Boiman and Michal Irani. Detecting irregularities in images and in video. In *Proceedings of International Conference on Computer Vision*, volume 1, pages 462–469, 2005.

[5] Vincent Cheung, Brendan J. Frey, and Nebojsa Jojic. Video epitomes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 42–49, 2005.

[6] Chris Dance, Jutta Willamowski, Lixin Fan, Cedric Bray, and Gabriela Csurka. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[7] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS 2005*, pages 65–72, 2005.

[8] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, Nice, France, 2003.

[9] Claudio Fanti, Lihi Zelnik-Manor, and Pietro Perona. Hybrid models for human motion recognition. In *ICCV*, volume 1, pages 1166–1173, 2005.

[10] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[11] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conferences*, pages 147–152, 1988.

[12] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, August 1999.

[13] Yan Ke, Rahul Sukthankar, and Martial Hebert. Efficient visual event detection using volumetric features. In *International Conference on Computer Vision*, pages 166–173, 2005.

[14] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *Proceedings of the ninth IEEE International Conference on Computer Vision*, volume 1, pages 432 – 439, 2003.

[15] Deva Ramanan and David A. Forsyth. Automatic annotation of everyday movements. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[16] Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 2(37):151–172, 2000.

[17] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *ICPR*, pages 32–36, 2004.

[18] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV)*, pages 370 – 377, October 2005.

[19] Yang Song, Luis Goncalves, and Pietro Perona. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(25):1–14, 2003.

[20] Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, and Greg Mori. Unsupervised discovery of action classes. In *CVPR*, 2006.

[21] Alper Yilmaz and Mubarak Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *IEEE International Conf. on Computer Vision (ICCV)*, volume 1, pages 150 – 157, 2005.

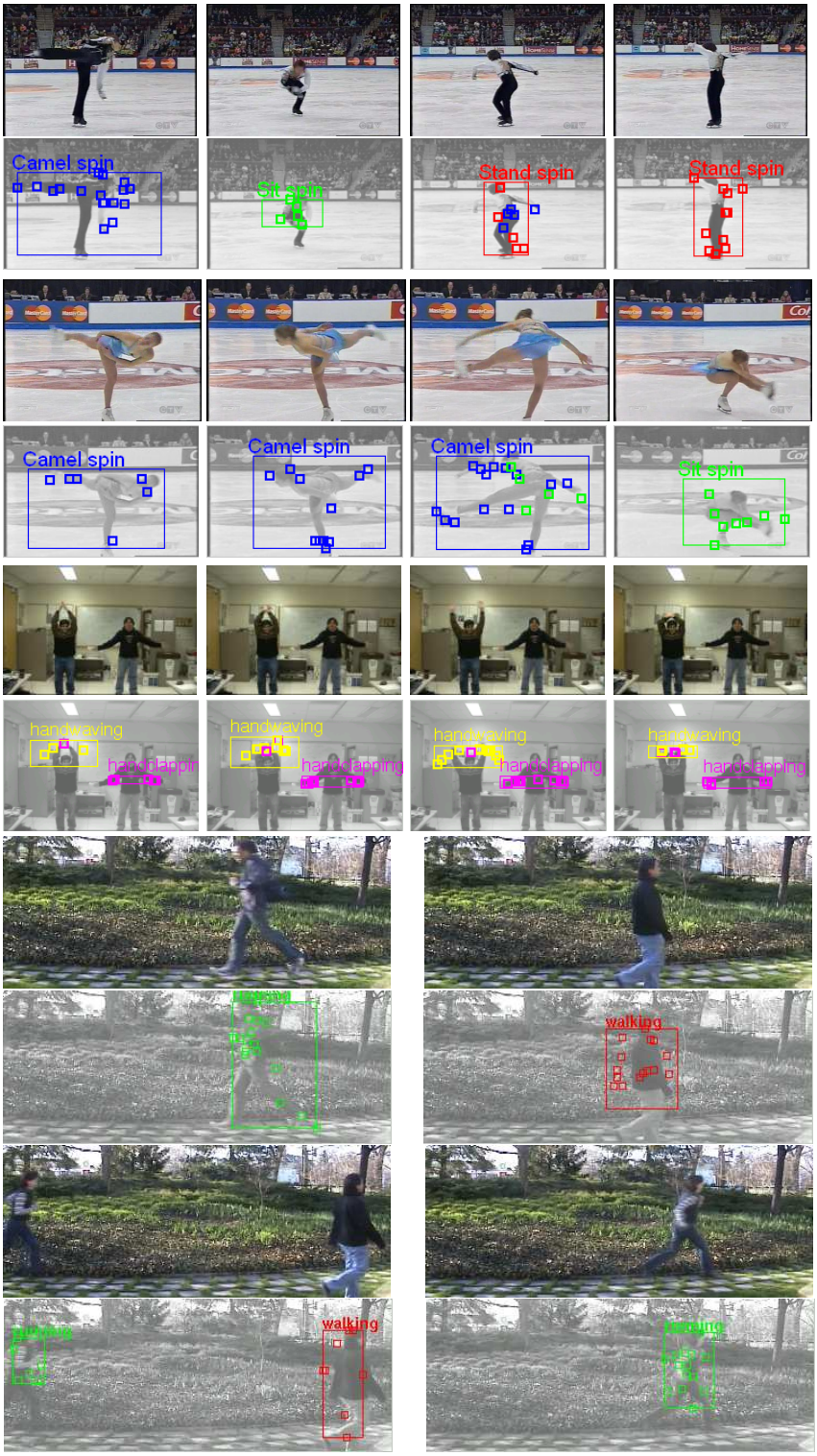

Figure 7: Multiple action recognition and localization in long and complex video sequences. The figure is best viewed in color and with PDF magnification.