Int J Comput Vis
DOI 10.1007/s11263-007-0122-4

# Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words

**Juan Carlos Niebles · Hongcheng Wang · Li Fei-Fei**

Received: 16 March 2007 / Accepted: 26 December 2007


**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 the intermediate topics corresponding to human action categories. This is achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA). Our approach can handle noisy feature points arisen from dynamic background and moving cameras due to the application of the probabilistic models. Given a novel video sequence, the algorithm can categorize and localize the human action(s) contained in the video. We test our algorithm on three challenging datasets: the KTH human motion dataset, the Weizmann human action dataset, and a recent dataset of figure skating actions. Our results reflect the promise of such a simple approach. In addition, our algorithm can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

J.C. Niebles (✉)
Department of Electrical Engineering, Princeton University, Engineering Quadrangle, Olden Street, Princeton, NJ 08544, USA
e-mail: jniebles@princeton.edu

J.C. Niebles
Robotics and Intelligent Systems Group, Universidad del Norte, Km 5 Vía Puerto Colombia, Barranquilla, Colombia

H. Wang
United Technologies Research Center (UTRC), 411 Silver Lane, East Hartford, CT 06108, USA

L. Fei-Fei
Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA

**Keywords** Action categorization · Bag of words · Spatio-temporal interest points · Topic models · Unsupervised learning

## 1 Introduction

Imagine a video taken on a sunny beach, where there are people playing beach volleyball, some are surfing, and others are taking a walk along the beach. Can a computer automatically tell us what is happening in the scene? Can it identify different human actions? We explore the problem of human action categorization in video sequences. Our interest is to design an algorithm that permits the computer to learn models for human actions. Then, given a novel video, the algorithm should be able to decide which human action is present in the sequence. Furthermore, we look for means to provide a rough indication of where (in space and time) the action is being performed.

The task of automatic categorization and localization of human actions in video sequences is highly interesting for a variety of applications: detecting relevant activities in surveillance video, summarizing and indexing video sequences, organizing a digital video library according to relevant actions, etc. It remains, however, a challenging problem for computers to achieve robust action recognition due to cluttered background, camera motion, occlusion, view point changes, and geometric and photometric variances of objects.

These challenges are common to a broad range of computer vision tasks. A cluttered background introduces information that is not relevant to the signal of interest, making the latter harder to isolate. Camera motion creates ambiguities in the motion patterns that are observed in the image


**Fig. 1** Example frames from video sequences in the figure skating dataset (Wang et al. 2006). We adapt 32 video sequences from the original dataset, to produce a subset which contains seven people executing three actions: camel-spin (*first row*), sit-spin (*second row*) and stand-spin (*third row*). The videos are taken with a moving camera and dynamic background

**Fig. 2** Example images from complex video sequences taken by the authors with a hand held camera. In these videos, there are multiple people performing different actions against a cluttered background

plane: it could make an object appear static when it is moving with the same speed and direction as the camera. In addition, human actions can also be observed only partially due to occlusions, thus the actual signal of interest can be dramatically reduced. Finally, view point changes as well as geometric and photometric variance produce very different appearances and shapes for the same category examples, resulting in high intra-class variances.

Consider for example, a live video of a figure skating competition, the skater moves rapidly across the rink and the camera also moves to follow the skater. With moving cameras, cluttered background, and moving target, few vision algorithms could identify, categorize and localize such motions well (Fig. 1). In addition, the challenge is even greater when there are multiple activities in a complex video sequence (Fig. 2). In this paper, we will present an algorithm that aims to account for these scenarios.

We propose a generative graphical model approach to learn and recognize human actions in video, taking advantage of the robust representation of sparse spatial-temporal interest points and an unsupervised learning approach. In the context of our problem, unsupervised learning is achieved by obtaining action model parameters from unsegmented and unlabeled video sequences, which contain a known number of human action classes. We advocate the use of an unsupervised learning setting because it opens the possibility to take advantage of the increasing amount of available video data, without the expense of detailed human annotation. Towards this end, a generative approach provides means to learn models in an unsupervised fashion; as opposed to discriminative models which generally require detailed labeled data.

Our method is motivated by the recent success of object detection/classification or scene categorization from unlabeled static images, using latent topic models (Sivic et al. 2005; Fei-Fei and Perona 2005). One key consideration in these works is known as the "bag of words" representation,[1] where the geometric arrangement between visual features is ignored. This is commonly implemented as a histogram of the number of occurrences of particular visual patterns in a given image. This representation

[1] Alternatively, some researchers refer to this representation as "bag of keypoints", see for example (Dance et al. 2004).

is a heritage from the text analysis domain, for which the latent topic models were first developed (Hofmann 1999; Blei et al. 2003). In spite of their simplicity, the latent topic models have been successfully applied to challenging computer vision tasks, which motivates us to explore their applicability in the human action categorization domain.

Two related models are generally used: probabilistic Latent Semantic Analysis (pLSA) by Hofmann (1999) and Latent Dirichlet Allocation (LDA) by Blei et al. (2003). In this paper, we investigate the suitability of both models for video analysis by exploring the advantages of the powerful representation and the great flexibility of these generative graphical models.

The contributions of this work are twofold. First, we propose an *unsupervised learning approach for human actions using a bag of words representation*. We apply two latent topic models, pLSA and LDA, to the problem of learning and recognizing human action categories, while adopting a "bag of spatial-temporal words" representation for video sequences. Second, our method can *localize and categorize multiple actions in a single video*. In addition to the categorization task, our approach can also localize different actions simultaneously in a novel and complex video sequence. This includes the cases where multiple people are performing distinct actions at the same time, and also situations where a single person is performing distinct actions through time.

In order to gather experimental evidence that supports our proposed approach, we train and recognize action models on three different datasets: the KTH human action database (Schuldt et al. 2004), the Weizmann human action dataset (Blank et al. 2005), and a figure skating dataset adapted from the dataset in (Wang et al. 2006). In addition, we used those models to perform recognition in videos from a different dataset (Song et al. 2003), as well as test sequences taken by ourselves (Fig. 2). Note that we use testing data that was collected in a totally different setting than that used for training. This will provide a proper out of sample testing scenario.

The rest of the paper is organized in the following way. We review previous related work in Sect. 2. In Sect. 3, we describe our approach in more details, including the spatial-temporal feature representation, a brief overview of the pLSA and LDA model in our context, and the specifics of the learning and recognition procedures. In Sect. 4, 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 learnt model. Finally, Sect. 5 concludes the paper.

A preliminary version of this paper appeared in BMVC 2006 (Niebles et al. 2006).

## 2 Background Work

A considerable amount of previous work has addressed the question of human action categorization and motion analysis. One line of work is based on the computation of correlation between volumes of video data. Efros et al. (2003) perform action recognition by correlating optical flow measurements from low resolution videos. Their method requires first segmenting and stabilizing each human figure in the sequence, as well as further human intervention to annotate the actions in each resulting spatial-temporal volume. Shechtman and Irani (2005) propose a behavior-based correlation to compute the similarity between space-time volumes which allows to find similar dynamic behaviors and actions. Their method requires to specify a query action template, which will be correlated to videos in database. At each pixel, the space-time gradients of the corresponding video patch must be computed and summarized in a matrix. The eigenvalues of the resulting matrices are used to compute similarity between two spatial-temporal patches. Therefore, this method requires significant computation due to the correlation procedure between every patch of the testing sequence and the video database.

Another popular approach is to first track body parts and then use the obtained motion trajectories to perform action recognition. This is done with much human supervision and the robustness of the algorithm is highly dependent on the tracking system. Ramanan and Forsyth (2004) approach action recognition by first tracking the humans in the sequences using a pictorial structure procedure. Then 3D body configurations are estimated and compared to a highly annotated 3D motion library. The algorithm permits assigning composed labels to the testing sequences; however, it relies heavily on the result of the tracker, and the estimation of the 3D pose may introduce significant errors due to hard-to-solve ambiguities. In Yilmaz and Shah (2005), human labeling of landmark points in the human body is first done at each frame in sequences from multiple moving cameras. Then actions are compared using their corresponding 4D $(x, y, z, t)$ trajectories. Thus, their approach can be applied to action recognition and retrieval, with the cost of a significant amount of human annotation. In the work by Song et al. (2003) and Fanti et al. (2005), feature points are first detected and tracked in a frame-by-frame manner. Multiple cues such as position, velocities and appearance are obtained from these tracks. Then human actions are modeled utilizing graphical models based on triangulated graphs. These models can be learnt in an unsupervised fashion. However, their methods cannot deal with dynamic backgrounds, since background features must be uniformly distributed and such assumption fails if a rigid object is moving and generating features in the background.

Alternatively, researchers have considered the analysis of human actions by looking at video sequences as space-time

intensity volumes. Bobick and Davis (2001) use motion history images that capture motion and shape to represent actions. They have introduced the global descriptors *motion energy image* and *motion history image*, which are used as templates that could be matched to stored models of known actions. Their method depends on background subtraction and thus cannot tolerate moving cameras and dynamic backgrounds. Blank et al. (2005) represent actions as space-time shapes and extract space-time features for action recognition, such as local space-time saliency, action dynamics, shape structures and orientation. Similarly, this approach relies on the restriction of static backgrounds which allows them to segment the foreground using background subtraction.

Other lines of work have been proposed for video analysis. Boiman and Irani (2005) propose composing the new observations as an ensemble of local video patches from previous examples in order 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 difficult to apply this method to action recognition or categorization due to the large amount of video data commonly presented in these settings. Another work known as video epitomes is proposed by Cheung et al. (2005). They model the space-time cubes from a specific video by a generative model. The learnt 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.

Some researchers have also explored unsupervised methods for motion analysis. Hoey (2001) applies a hierarchical dynamic Bayesian network model to unsupervised facial expression recognition. The approach relies on previously tracked and segmented faces whose motion is described using optical flow. Zhong et al. (2004) have proposed an unsupervised approach to detect unusual activity in video sequences. Using a global representation based on a simple descriptor vector per each frame, their method clusters video segments and identifies spatially isolated clusters as unusual activity. Therefore, the unusual activities must be observed during training. Xiang and Gong (2005) apply the Multi-Observation Hidden Markov model and spectral clustering to unsupervised training of behavior models that can detect abnormal activities.

Another approach uses a video representation based on spatial-temporal interest points. In spite of the existence of a fairly large variety of methods to extract interest points from static images (Schmid et al. 2000), less work has been done on space-time interest point detection in videos. Laptev (2005) presents a space-time interest point detector based on the idea of the Harris and Förstner interest point operators (Harris and Stephens 1988). They detect local structures in space-time where the image values have significant local variations in both dimensions. However, this method produces a small number of stable interest-points which are often non sufficient to characterize complex sequences. In addition, Dollár et al. (2005) propose a detector based on a set of separable linear filters, which generally produces a high number of detections. This method responds to local regions which exhibit complex motion patterns, including space-time corners. Also, a number of descriptors are proposed for the resulting video patches around each interest point. Ke et al. (2005) apply spatial-temporal volumetric features that efficiently scan video sequences in space and time. Their method builds on the rectangle features used by Viola and Jones (2001). Their approach detects interest points over the motion vectors, which requires dense estimation of the optical flow. Additionally, the method requires to calculate a significant number of features which are in the order of a million, even after discretizing and sampling the feature space. The detected interest points are then employed as features to perform human action categorization with a discriminative cascade classifier, which requires annotated positive and negative examples. Finally, a recent approach by Oikonomopoulos et al. (2006) extends the idea of saliency regions in spatial images to the spatiotemporal case. The work is based on the spatial interest points of Kadir and Brady (2003), which is extended to the space-time case. Two set of spatiotemporal salient points are compared based on the chamfer distance. Experimental results are promising based on their own video sequences captured by a stationary camera.

Interest points extracted with such methods have been used as features for human action classification. In (Schuldt et al. 2004; Dollár et al. 2005; Oikonomopoulos et al. 2006; Ke et al. 2005), the space-time interest points are combined with discriminative classifiers to learn and recognize human actions. Therefore, local space-time patches 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 (Schuldt et al. 2004; Dollár et al. 2005), posing a problem toward handling more challenging situations such as multiple action recognition.

Finally, we note the success of generative approaches based on latent topics models for object and scene recognition. Fei-Fei and Perona (2005) introduce the application of latent topic models to computer vision tasks, within the scope of natural scene categorization. Their models are inspired by the LDA model (Blei et al. 2003), and can learn intermediate topic distributions in an unsupervised manner. Also, Sivic et al. (2005) perform unsupervised learning and recognition of object classes by applying a pLSA model with the bag of visual words representation. The approach permits to learn object classes from images with no label and background clutter.

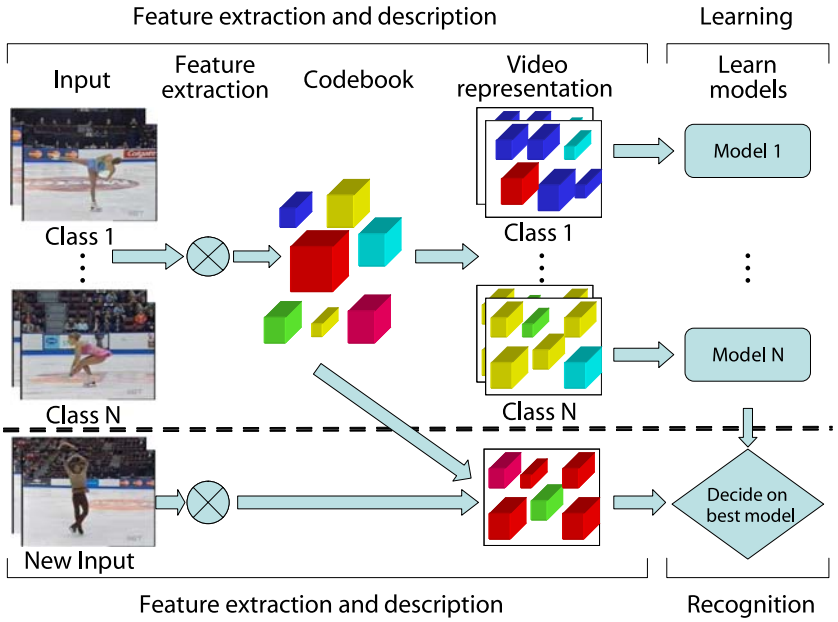


**Fig. 3** Flowchart of our approach. To represent motion patterns we first extract local space-time regions using the space-time interest point detector (Dollár et al. 2005). These local regions are then clustered into a set of spatial-temporal words, called *codebook*. Probability distributions and intermediate topics are learned automatically using one of the two models: *probabilistic Latent Semantic Analysis* (pLSA) or *Latent Dirichlet Allocation* (LDA). The learned models can then be used to recognize and localize human action classes in novel video sequences

All the previous works suggest that improvement can be made by relaxing assumptions of annotated data, stationary cameras and static backgrounds. Thus, we are interested in exploring the use of a generative approach where unsupervised learning methods can be applied, in conjunction with a representation based on local features. We present our proposed algorithm in the following section.

## 3 Our Approach

Given a collection of unlabeled videos, our goal is to automatically learn different classes of actions present in the data and to apply the learned model to perform action categorization and localization in the new video sequences. Our approach is illustrated in Fig. 3. We assume that the videos can contain some camera motion, for instance, the one observed in videos taken with a hand held camera. Also, we expect the videos to contain a dynamic background that might generate some motion clutter. In the training stage, we assume that there is a single person performing only one action per video. However, we relax this assumption at the testing stage, where our method can handle observations containing more than one person performing different actions.

We are given a set of unlabeled video sequences and we would like to discover a set of classes from them. Each of these classes would correspond to an action category, such that we can build models for each class. Additionally, we would like to be able to understand videos that are composed of a mixture of action categories, in order to handle the case of multiple motions. This resembles the problem of automatic topic discovery in text analysis (Hofmann 1999; Blei et al. 2003). Thus, we find a similar interpretation as that initially proposed by the use of latent topic models for object and scene classification (Fei-Fei and Perona 2005; Sivic et al. 2005). In our case, we would like to analyze video sequences instead of text documents; video sequences are summarized as a set of spatial-temporal words instead of text words; we seek to discover action categories instead of text topics; and we expect to explain videos as a mixture of actions instead of text documents as a mixture of topics. In this work, we investigate two models that were proposed in the text analysis literature to address the latent topic discovery problem. First, we employ the simpler *probabilistic Latent Semantic Analysis* (pLSA) proposed by Hofmann (1999). Second, we consider the *Latent Dirichlet Allocation* (LDA) model proposed by Blei et al. (2003), which provides a rigorous generative setting, permits the inclusion of priors in a Bayesian manner, and addresses the overfitting issues presented in the pLSA model. Both models permit to learn their parameters in an unsupervised fashion. Thus, these models provide an unsupervised learning framework that permits to automatically discover semantic clusters in the training data. Also, as opposed to discriminative methods such as Support Vector Machines, pLSA and LDA allow the algorithm to perform meaningful reasoning on the data beyond classification, for example topic localization. Furthermore, such localization can be realized without the need of scanning thousands or millions of windows per image. These models, however, do not provide spatial nor temporal scale invariances. Thus, they can only work within a small margin of the scales that have been observed in training. Alternative approaches that include such invariances might be based upon models such as those in (Fergus et al. 2005).

An important characteristic of the pLSA and LDA models is that they are based on the bag of words assumption, that is, the order of words in a text document can be neglected. This is equivalent to regarding the words in a document as *exchangeable*. In addition, the particular ordering of the documents in the document collection can also be neglected, yielding a further exchangeability assumption at the document level. In the context of human action classification, the bag of words assumption translates into a video representation that ignores the positional arrangement, in space and time, of the spatial-temporal interest points. Such assumption brings the advantages of using a simple representation that makes learning efficient. The lack of spatial information provides little information about the human body, while the lack of longer term temporal information does not permit us to model more complex actions that are not constituted by simple repetitive patterns. Alternative approaches might include structural information by encoding information of the human body using a pictorial structure model (Felzenszwalb and Huttenlocher 2005), by observing co-occurrences of local patterns such as those in (Savarese et al. 2006), or by modeling the geometrical arrangement of local features (Niebles and Fei-Fei 2007). In most cases, the trade off is an increased computational complexity.

### 3.1 Feature Representation from Space-Time Interest Points

There are several choices in the selection of good features to describe pose and motion. In general, there are three popular types of features: static features based on edges and limb shapes (Dalal et al. 2006; Feng and Perona 2002), dynamic features based on optical flow measurements (Dalal et al. 2006; Sidenbladh and Black 2003), and spatial-temporal features obtained from local video patches (Blank et al. 2005; Cheung et al. 2005; Dollár et al. 2005; Laptev 2005; Ke et al. 2005; Oikonomopoulos et al. 2006). In particular, features from spatial-temporal interest points have shown to be useful in the human action categorization task, providing a rich description and powerful representation (Dollár et al. 2005; Schuldt et al. 2004; Ke et al. 2005; Oikonomopoulos et al. 2006).

As Fig. 3 illustrates, we represent each video sequence as a collection of spatial-temporal words by extracting space-time interest points. Among the available interest point detectors for video data, the interest points obtained using the generalized space-time corner detector (Laptev 2005) are too sparse to characterize many complex videos. This was noted first in (Dollár et al. 2005), and confirmed in our experience with complex sequences such as the figure skating videos (Fig. 1). We choose to use the separable linear filter method in (Dollár et al. 2005), since it generally produces a high number of detections. Note, however, that our method does not rely on a specific interest point detector algorithm, as long as the detector produces a sufficiently large number of interest points. In the following, we provide a brief review of the detector proposed in (Dollár et al. 2005).

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 $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega) \times e^{-t^2/\tau^2}$ and $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. The two parameters $\sigma$ and $\tau$ correspond to the spatial and temporal scales of the detector respectively. In all cases we use $\omega = 4/\tau$, and thus reducing to two the number of parameters in 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 (Dollár et al. 2005) 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 that contributed to the response function, i.e., its size is approximately six times the scales along each dimension.

Figure 4 shows an example of interest point detection in a hand waving video sequence. Each colored box corresponds to a detected interest point, that is associated with a video patch. The neighborhood size is determined by the scale parameters $\sigma$ and $\tau$ of the detector. Interest points are correctly localized where significant motion occurs.

Note that by detecting spatial-temporal interest points, a sparse representation of the video sequences is produced. Small video patches are extracted from each interest point and constitute the local information that is used to learn and recognize human action categories. By employing local features, we intent to emphasize the importance and distinctiveness of the short range spatial-temporal patterns. We argue that the observed local patterns are discriminative enough across human action classes (refer to Fig. 9), and provide a reasonable feature space which allows to build good human action models. Additionally, this approach relaxes the need of previously common preprocessing steps in global approaches such as background subtraction in (Blank et al. 2005; Bobick and Davis 2001), or figure tracking and stabilization in (Efros et al. 2003).

**Fig. 4** Interest point detection using the method of separable linear filters in (Dollár et al. 2005). Each *red box* corresponds to a video patch that is associated with a detected interest point. The neighborhood size is determined by the scale parameters $\sigma$ and $\tau$ of the detector. The interest points are localized where significant motion occurs and can be used to generate a sparse representation of the video sequence. For a visualization of all the frames in particular spatial-temporal patches, please refer to Fig. 9. The figure is best viewed in color with PDF magnification

To obtain a descriptor for each spatial-temporal cube, we calculate its brightness gradients on $x$, $y$ and $t$ directions. The spatial-temporal cube is then smoothed at different scales before computing the image gradients. The computed gradients are concatenated to form a vector. The size of the vector is equal to the number of pixels in the cube times the number of smoothing scales times the number of gradients directions. This descriptor is then projected to a lower dimensional space using the principal component analysis (PCA) dimensionality reduction technique. In (Dollár et al. 2005), different descriptors have been used, such as normalized pixel values, brightness gradient and windowed optical flow. We find 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.

It is worth noting that a number of video patch descriptors have been proposed previously (Dollár et al. 2005; Laptev and Lindeberg 2006). As mentioned above, we have chosen a very simple descriptor based on image gradients (Dollár et al. 2005). Such descriptor does not provide scale invariance neither in the space nor the time domain. It does not capture relative camera motion. However, more complex descriptors that include small invariances to spatial scale and speed, as well as invariances to small camera motions, are available with the cost of more computational complexity (for instance, local position dependent histograms in Laptev and Lindeberg 2006). In our implementation, we rely on the codebook to handle scale changes and camera motions. As long as the newly observed local features do not contain patterns of scale change and camera motion that are extremely different from those observed in the data used to form the codebook, we expect that similar local features will be assigned to consistent memberships on the codebook.

### 3.2 Codebook Formation

The latent topic models pLSA and LDA rely on the existence of a finite vocabulary of (spatial-temporal) words of size $V$. In order to learn the vocabulary of spatial-temporal words, we consider the set of descriptors corresponding to all detected spatial-temporal interest points in the training data. This vocabulary (or codebook) is constructed by clustering using the $k$-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined to be a spatial-temporal word (or codeword). Thus, each detected interest point can be assigned a unique cluster membership, i.e., a spatial-temporal word, such that a video can be represented as a collection of spatial-temporal words from the codebook. The effect of the codebook size is explored in our experiments, and the results are shown in Fig. 13 and Fig. 8.

### 3.3 Learning the Action Models: Latent Topic Discovery

In the following, we will describe the pLSA and LDA models in the context of human action categories analysis, adapting the notation and terminology as needed from the ones introduced by (Hofmann 1999; Blei et al. 2003).

#### *3.3.1 Learning and Recognizing the Action Models by pLSA*

Suppose we have a set of $M(j = 1, \ldots, M)$ video sequences containing spatial-temporal words from a vocabulary of size $V(i = 1, \ldots, V)$. The corpus of videos is summarized in a $V$ by $M$ co-occurrence table $\bar{M}$, where $m(w_i, d_j)$ stores the number of occurrences of a spatial-temporal word $w_i$ in video $d_j$. In addition, there is a latent topic variable $z_k$ associated with each occurrence of a spatial-temporal word $w_i$

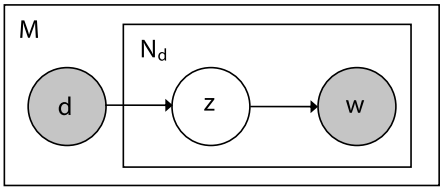


**Fig. 5** The *probabilistic Latent Semantic Analysis* (pLSA) graphical model. Nodes are random variables. *Shaded* ones are observed and unshaded ones are unobserved. The plates indicate repetitions. In the context of human action categorization, $d$ represents video sequences, $z$ are action categories and $w$ are spatial-temporal words. The parameters of this model are learnt in an unsupervised manner using an EM procedure. This figure is reproduced from (Blei et al. 2003)

in a video $d_j$. Each topic corresponds to an action category, such as walking, running, etc.

The joint probability $P(w_i, d_j, z_k)$ is assumed to have the form of the graphical model shown in Fig. 5:

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

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 spatial-temporal 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.

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 by a mixture of the histograms corresponding to each action category.

We then fit the model by determining the action category histograms $P(w_i|z_k)$ (which are common to all videos) and the mixture coefficients $P(z_k|d_j)$ (which are specific to each video). In order to determine the model that gives the highest probability to the spatial-temporal words that appear in the corpus, a maximum likelihood estimation of the parameters is obtained by maximizing the following objective function using an expectation-maximization (EM) algorithm:

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

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

Given that our algorithm has learnt the action category models, our goal is to categorize new video sequences. 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 learnt $P(w|z)$. 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 (Hofmann 1999). Similarly to the learning scenario, we apply an EM algorithm to find the solution. Thus, a categorization decision is made by selecting the action category that best explains the observation, that is:

$$\text{Action Category} = \arg\max_k P(z_k|d_{test}). \tag{5}$$

Furthermore, we are also interested in localizing multiple actions in a single video sequence. Though our bag of spatial-temporal words model itself does not explicitly represent the spatial or temporal relationships of the local video regions, it is sufficiently discriminative to localize different motions within each video. This is similar to the approximate object segmentation case in (Sivic et al. 2005). The pLSA model models the posteriors

$$P(z_k|w_i, d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)}. \tag{6}$$

Once each interest point has been assigned to a spatial-temporal word, we can label the corresponding word $w_i$ with a particular topic by finding the maximum of the posteriors $P(z_k|w_i, d_j)$ over $k$. Thus, we label the regions that support the set of detected interest points, effectively producing a topic localization, which corresponds to the localization of potentially multiple actions that can belong to different action categories.

### 3.3.2 Learning and Recognizing the Action Models by LDA

As noted in (Blei et al. 2003), pLSA is not a well-defined generative model of documents, since there is no natural way to use it to assign probability to a new testing observation. In addition, the number of parameters to be estimated in pLSA grows linearly with the number of training examples, which suggest that this model is prone to overfitting. LDA (Blei et al. 2003) addresses these weaknesses.

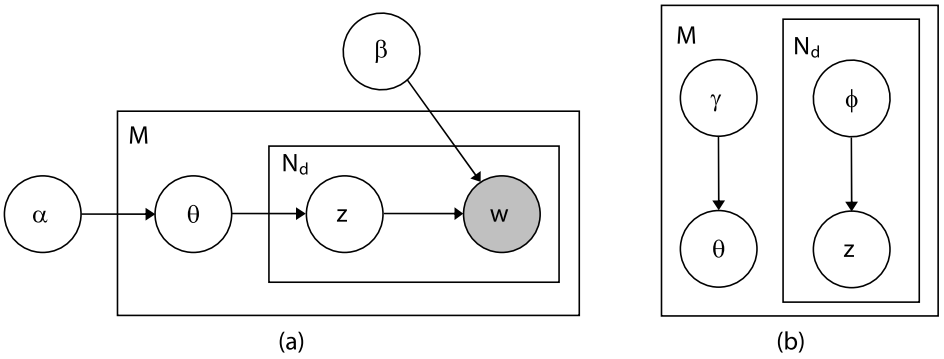


**Fig. 6** (**a**) *Latent Dirichlet Allocation* (LDA) graphical model (Blei et al. 2003). Nodes are random variables. Shaded ones are observed and unshaded ones are unobserved. The plates indicate repetitions. In the context of human action categorization, $\theta$ represents video sequences, $z$ are action categories and $w$ are spatial-temporal words. $\alpha$ is the hyperparameter of a Dirichlet distribution. (**b**) Graphical model that represents the variational distributions proposed in (Blei et al. 2003) to approximate the posterior probability in LDA. This figure is reproduced from (Blei et al. 2003)

Suppose we have a set of $M(j = 1, \ldots, M)$ video sequences containing spatial-temporal words from a vocabulary of size $V(i = 1, \ldots, V)$. Each video $d_j$ is represented as a sequence of $N_j$ spatial-temporal words $\mathbf{w} = (w_1, w_2, \ldots, w_{N_j})$. Then the process that generates each video $d_j$ in the corpus is:

1. Choose the number of spatial-temporal words: $N_j \sim \text{Poisson}(\xi)$
2. Choose the mixing proportions of the action categories: $\theta \sim \text{Dir}(\alpha)$
3. For each of the $N_j$ words $w_n$:
   - Choose an action category (topic): $z_n \sim \text{Multinomial}(\theta)$
   - Choose a spatio-temporal word $w_n$ from the multinomial distribution $p(w_n|z_n, \boldsymbol{\beta})$

Here we fixed the number of latent topics $K$ to be equal to the number of action categories to be learnt. Also, $\alpha$ is the parameter of a $K$-dimensional Dirichlet distribution, which generates the multinomial distribution $\theta$ that determines how the action categories (latent topics) are mixed in the current video. In addition, a matrix $\boldsymbol{\beta}$ of size $K \times V$ parameterizes the distribution of spatial-temporal words conditioned on each action category; each element of $\boldsymbol{\beta}$ corresponds to the probability $p(w_i|z_k)$.

The joint distribution of a topic mixture $\theta$, the set of words $\mathbf{w}$ observed in the current video, and their corresponding topic (action category) $\mathbf{z}$ can be written as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \boldsymbol{\beta}) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \boldsymbol{\beta}). \tag{7}$$

The probabilistic graphical model in Fig. 6 represents the LDA model.

In order to perform video classification with LDA, one must compute the posterior distribution of the hidden variables given a new input:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \boldsymbol{\beta}) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \boldsymbol{\beta})}{p(\mathbf{w}|\alpha, \boldsymbol{\beta})} \tag{8}$$

where $\theta$ is specific to each input and represents its latent topics distribution. Once $\theta$ is inferred, a classification decision can be made by selecting the most likely topic in the current testing video.

Although it is computationally intractable to perform inference and parameter estimation for the LDA model in general, several approximation algorithms have been investigated. A variational inference approach has been proposed in (Blei et al. 2003). The family of variational distributions that are considered can be represented by the model in Fig. 6(b), and are characterized by:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{9}$$

where $\gamma$ and $\theta$ are the free variational parameters. The corresponding optimization procedure produces the parameters $(\gamma^*, \phi^*)$ which are a function of $\mathbf{w}$.

Analogously to the pLSA case, the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ represent the projection of the new observed video into the simplex spanned by the latent topics. Thus, classification is performed by selecting the action category that corresponds to the maximum element in $\gamma^*(\mathbf{w})$.

Furthermore, the localization procedure can also be implemented using LDA. In this case, we can label each interest point with an action category, by selecting the topic that generates its corresponding spatial-temporal word with highest probability. That means, for a fixed $i$, we select $k$ such that $p(w_i|z_k)$ in $\boldsymbol{\beta}$ is maximum.


**Fig. 7** Example images from video sequences in the KTH dataset (Schuldt et al. 2004). The dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. These are performed several times by 25 subjects in different scenarios of outdoor and indoor environment. The camera is not static and the videos contain scale changes. This figure is reproduced from http://www.nada.kth.se/cvap/actions/

## 4 Experimental Results

We test our algorithm using three datasets: the KTH human motion dataset (Schuldt et al. 2004), a figure skating dataset (Wang et al. 2006), and the Weizmann human action dataset (Blank et al. 2005). These datasets contain videos of cluttered background, moving cameras, and multiple actions; as well as videos exhibiting a single action, with static camera and simple background. We can handle the noisy feature points arisen from dynamic background and moving cameras by utilizing the latent topic models pLSA and LDA, 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.

### 4.1 Recognition and Localization of Single Actions

#### *4.1.1 Human Action Recognition and Localization Using the KTH Dataset*

KTH human motion dataset is the largest available video sequence dataset of human actions (Schuldt et al. 2004). 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 short sequences. Some sample images are shown in Fig. 7.

We extract interest points and describe the corresponding spatial-temporal patches with the procedure described in Sect. 3.1. The detector parameters are set to $\sigma = 2$ and $\tau = 2.5$. Each spatial-temporal patch is described with the concatenated vector of its space-time gradients. Then, the descriptors are projected to a lower dimensional space of 100 dimensions. Examples of the detections for sequences in each category are shown in Fig. 10.

In order to build the codebook, we need to cluster the feature descriptors of all training video sequences. However, since the total number of features from all training examples is very large, we use only a subset of sequences to learn the codebook, in order to accommodate the requirements of memory. Thus, we build spatial-temporal codewords using only two videos of each action from three subjects. We keep these sequences out of the training and testing sets, to avoid contamination in the data.

In order to test the efficiency of our approach for the recognition task, we adopt the leave-one-out testing paradigm (LOO). Each video is labeled with the index of the subject performing the action but not with the action class label, so that the algorithm does not have information about the action class contained in the sequences. Thus, for each LOO run, we learn a model from the videos of 24 subjects (except those videos used to build the codebook) in an unsupervised fashion, test the videos of the remaining subject,

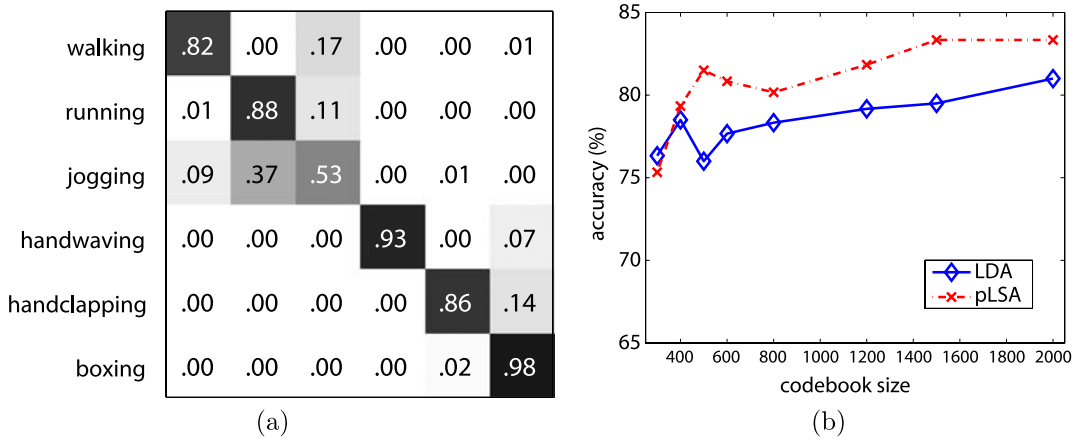


**Fig. 8** (**a**) Confusion matrix for the KTH dataset using 1500 codewords (performance average = 83.33%); *rows* are ground truth, and *columns* are model results; (**b**) Classification accuracy vs. codebook size for the KTH dataset. Experiments show that the results for the recognition task are consistently better when the pLSA model is adopted

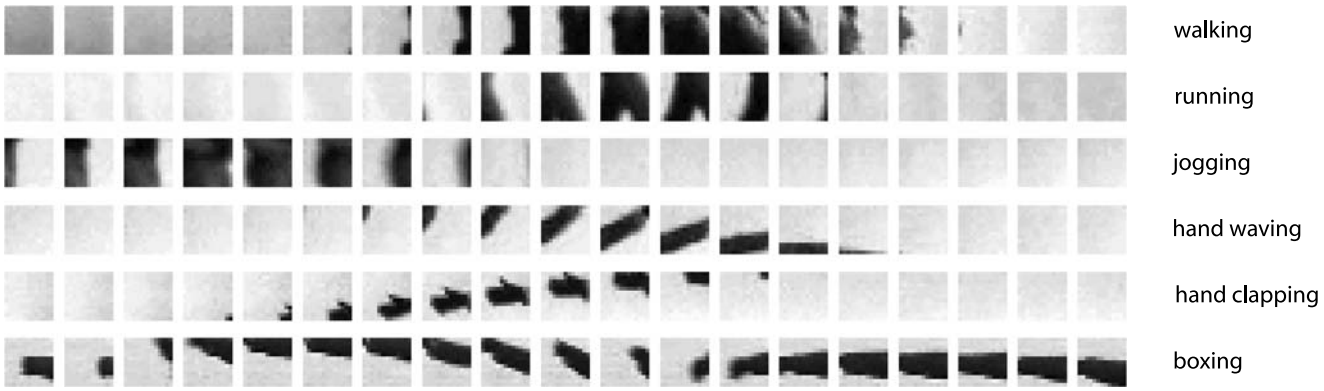


**Fig. 9** The latent topic models provide means to rank the spatial-temporal words given an action class. Here, we illustrate the top word from each category, in the KTH dataset, using a spatial-temporal patch. Each row contains the frames from the neighborhood of a single spatial-temporal interest point, which was assigned to a top word within the category on the right. The spatial-temporal patches clearly characterize each action class; for instance, the top interest point for hand-waving shows its signature of up-down arm motion

and compute a confusion table for evaluation. The results are reported as the average confusion table of the 25 runs.

Under these settings, we learn and recognize human action categories using the pLSA and LDA models. The confusion matrix for a six-class pLSA model for the KTH dataset is given in Fig. 8(a) using 1500 codewords. Our algorithm automatically assigns each test sequence to one of the action classes that were discovered during training. Each row in the confusion matrix corresponds to the ground truth class, and each column corresponds to the assigned cluster. Also, note that in order to maintain cluster correspondence between columns across different runs, each column is labeled with the majority label of videos that were assigned to the cluster.[2]

The confusion matrix shows the largest confusion between "jogging" and "running", "walking" and "jogging", and between "hand clapping" 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. Additionally, at the feature level, we note that the similarity across local patterns from different classes is highest between those categories where our method finds the largest confusion (please refer to Fig. 9).

We run our experiments on a Pentium 4 machine with 2 GB of RAM. The average times to train and test the pLSA and LDA models across the leave-one-out runs are reported in Table 1.

[2]Due to the unsupervised nature of our training procedure, each discovered cluster (i.e., action class) can only be automatically labeled with the names 'cluster 1', 'cluster 2', etc. Only by using ground truth labels, each cluster can then be named with the most popular action class label from the videos within the cluster. Alternatively, one can assign action names to each cluster by hand.

**Table 1** Learning and testing times for the KTH experiment

| Model | Codebook size | Learning time | Testing time |
|---|---|---|---|
| pLSA | 500 | 38.1 secs | 0.31 secs |
| LDA | 500 | 25.7 secs | 0.12 secs |

**Table 2** Comparison of different methods using the KTH dataset

| Methods | Recognition accuracy (%) | Learning | Multiple actions |
|---|---|---|---|
| Our method | 83.33 | unlabeled | Yes |
| Dollár et al. (2005) | 81.17 | labeled | No |
| Schuldt et al. (2004) | 71.72 | labeled | No |
| Ke et al. (2005) | 62.96 | labeled | No |

We test the effect of the number of video codewords on recognition accuracy on both models, as illustrated in Fig. 8(b). It shows some dependency of the recognition accuracy on the size of the codebook. Additionally, we can see that pLSA is slightly better than LDA in recognition performance with the same number of codewords. This is an interesting result. Our hypothesis for this outcome is that it is due to large variations and relatively small number of training samples in each action class, which may reduce the advantages of LDA.

We also compare our results with the best results from (Dollár et al. 2005) (performance average = 81.17%), which are obtained using a 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. Furthermore, our generative method provides better insight into the understanding of the action categories. Such analysis is not possible in the SVM discriminative approach. Additional comparison of recognition rates from different methods in the KTH dataset is given in Table 2. Please note that our experimental settings are equivalent to those in (Dollár et al. 2005). In (Ke et al. 2005) and (Schuldt et al. 2004), the training and testing sets are chosen by leaving out roughly half the data.

In order to obtain further insight into the model provided by the latent topic approach, we analyze the distribution of spatial-temporal words given a latent topic. In the pLSA case these distributions correspond to $p(w|z)$, and in the LDA case the distributions are given in $\boldsymbol{\beta}$. These parameters provide means to rank the spatial-temporal words according to their probability of occurrence within each action category. As a first exercise, it is interesting to observe which words are assigned the highest likelihood given an action category. Figure 9 shows example spatial-temporal patches that represent the top ranked word within each action category. These spatial-temporal patches clearly correspond to the correct human action class. Second, given a testing sequence, we can assign each of the observed interest points to a corresponding spatial-temporal word. This word in turn, can be assigned to the action class that generate it with highest probability, for example using (6) in the pLSA case. We show the result of this procedure in Fig. 10, using the distributions obtained with the pLSA model. Each interest point has been colored with the corresponding human action category. It is also clear how the model permits the mixture of action classes within a single sequence. Also, note that the dominant color correspond to the correct action category color.

Finally, we would like to use the models we have learned using the KTH dataset, to detect human actions in sequences from the Caltech human motion dataset (Song et al. 2003). We provide some examples frames from two of these video sequences in Fig. 11. There, the models learnt with a pLSA approach are used to detect the correct human action class. Most of the action sequences from this dataset can be correctly recognized. To provide further illustration, we have colored each spatial-temporal interest point according to its most likely action category. In the figure, we only draw the space-time features that were assigned to the action class that was detected by our model.

#### *4.1.2 Action Recognition and Localization Using the Weizmann Human Action Dataset*

In our second experiment, we employ the Weizmann human action dataset (Blank et al. 2005). It contains 10 action categories performed by 9 people, to provide a total of 90 videos. Example frames of the action categories are shown in Fig. 12. This dataset contains videos with static camera and simple background. However, it provides a good testing bed to investigate the performance of the algorithm when the number of categories is increased.

We detect and describe spatial-temporal interest points using the procedure detailed in previous sections. The detector parameters are set to $\sigma = 1.2$ and $\tau = 1.2$, and the dimensionality of the corresponding descriptors is reduced to 100. The codebook is learnt using all the feature descriptors obtained from all the training video sequences.

We again adopt a leave-one-out scheme to test the efficacy of our approach in recognition, i.e., for each run we learn a model from the videos of eight subjects, and test those of the remaining subject. The result is reported as the average of nine runs. The confusion matrix for a ten-class model is presented in Fig. 13(a) for a pLSA model learnt using a codebook of size 1200. The average performance of the pLSA model with this codebook size is 90%. Note that the confusion matrix shows how our model is mostly confused by similar action classes, such as "skip" with "jump" and "run", or "run" with "walk".

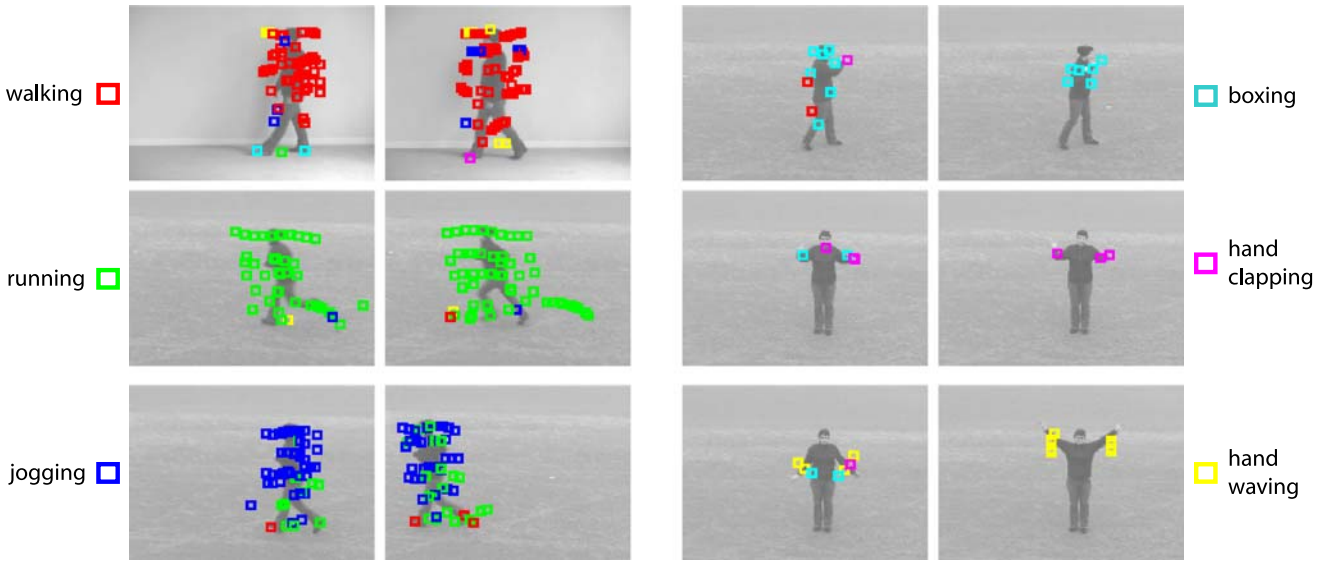


**Fig. 10** Example frames from testing sequences in the KTH dataset. The spatial-temporal patches in each sequence are automatically colored according to action class that most likely generated its corresponding spatial-temporal word. Although some of the words are assigned to the wrong topic, most interest points are assigned to the correct action for each video. Consistently, the predicted action class corresponds to the actual ground truth. In addition, we usually observe that the second best ranked action class corresponds to a similar action: in the "jogging" example of the figure, the second best label is "running". The figure is best viewed in color and with PDF magnification

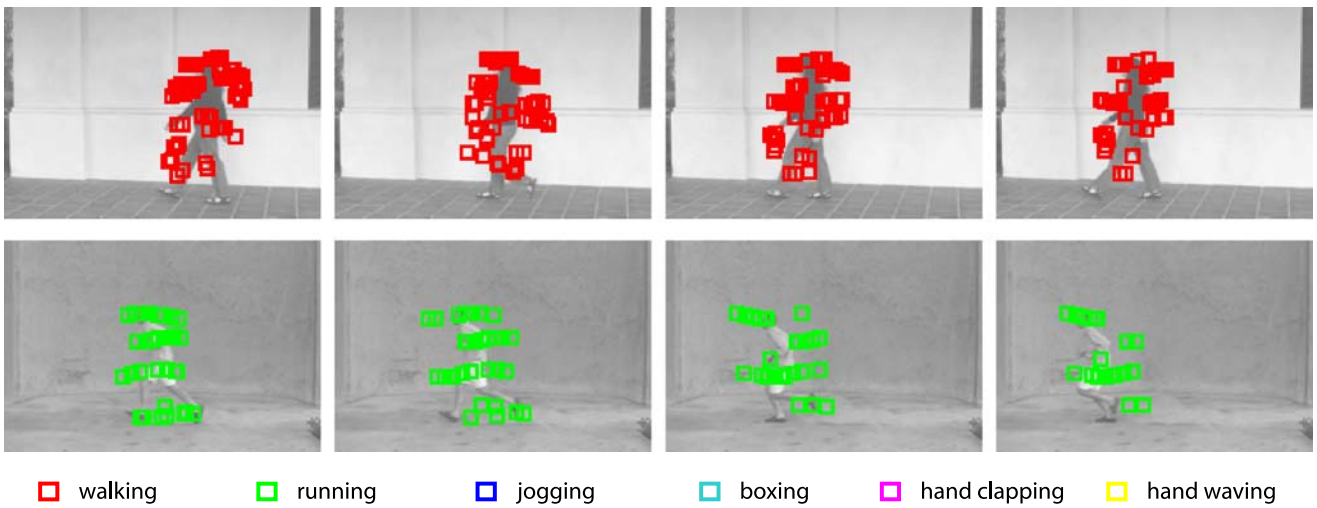


**Fig. 11** Examples frames from sequences in the Caltech dataset. Action category models were learnt using the KTH dataset, and tested again sequences in Caltech dataset. Each interest point is assigned to a action class, and only spatial-temporal interest points from the detected action category are shown. The figure is best viewed in color and with PDF magnification

We test the effect of the number of video codewords on recognition accuracy on the pLSA and LDA models, as illustrated in Fig. 13(b). It shows some dependency of the recognition accuracy on the size of the codebook.

Similarly to the previous experiment, we look for insight on what the latent topic model provides. Figure 14 illustrates sample frames from test sequences in each action class. We have colored each detected interest-point with its most likely action category. We observe how the model permits the mixture of action classes in each video; however, the actual action category dominates the coloring in all these cases. Additionally, it is also interesting to observe that those interest points that are not colored with the right action, are however assigned to a similar action. For instance, in the frames corresponding to the "jacks" category, there are some interest points assigned to "wave", and it is clear that both actions contain similar arm motion.


**Fig. 12** Example images from video sequences in the Weizmann human action dataset (Blank et al. 2005). The dataset contains 10 action categories, performed by 9 subjects. The videos are taken with static camera and static background

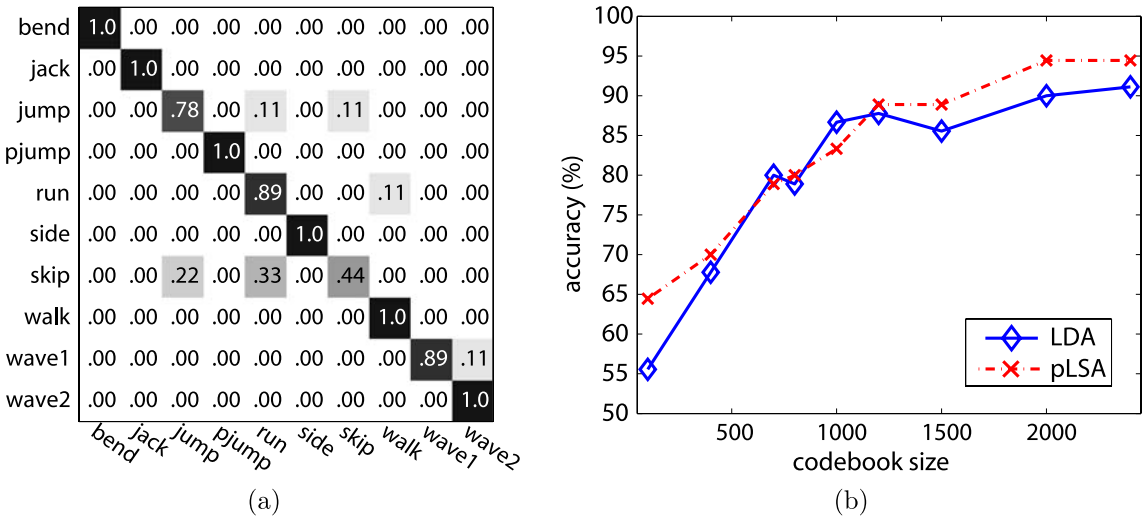


**Fig. 13** (**a**) Confusion matrix for the Weizmann human action dataset (Blank et al. 2005); *rows* are ground truth, and *columns* are model results. The action models learnt with pLSA and using 1200 codewords show an average performance of 90%. (**b**) Classification accuracy obtained using pLSA and LDA models vs. codebook size. Our results show that pLSA performs slightly better than LDA in the video categorization task

Finally, we note that in (Blank et al. 2005), experimental results were reported using 9 of the 10 action categories available in the dataset. Their classification task consisted on determining the action category of a set of space-time cubes, instead of classifying entire video sequences. Also, results on a clustering experiment were presented. These experiments differ from our task, which consist of categorizing complete video sequences. In addition, unlike our video sequence representation using local spatial-temporal words, their approach using space-time shape is sensitive to camera motion and dynamic background.

#### *4.1.3 Recognition and Localization of Figure Skating Actions*

As a third set of data, we use the figure skating dataset in (Wang et al. 2006).[3] We adapt 32 video sequences which contain seven people executing three actions: stand-spin, camel-spin and sit-spin, as shown in Fig. 1. The dataset con-

[3]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.

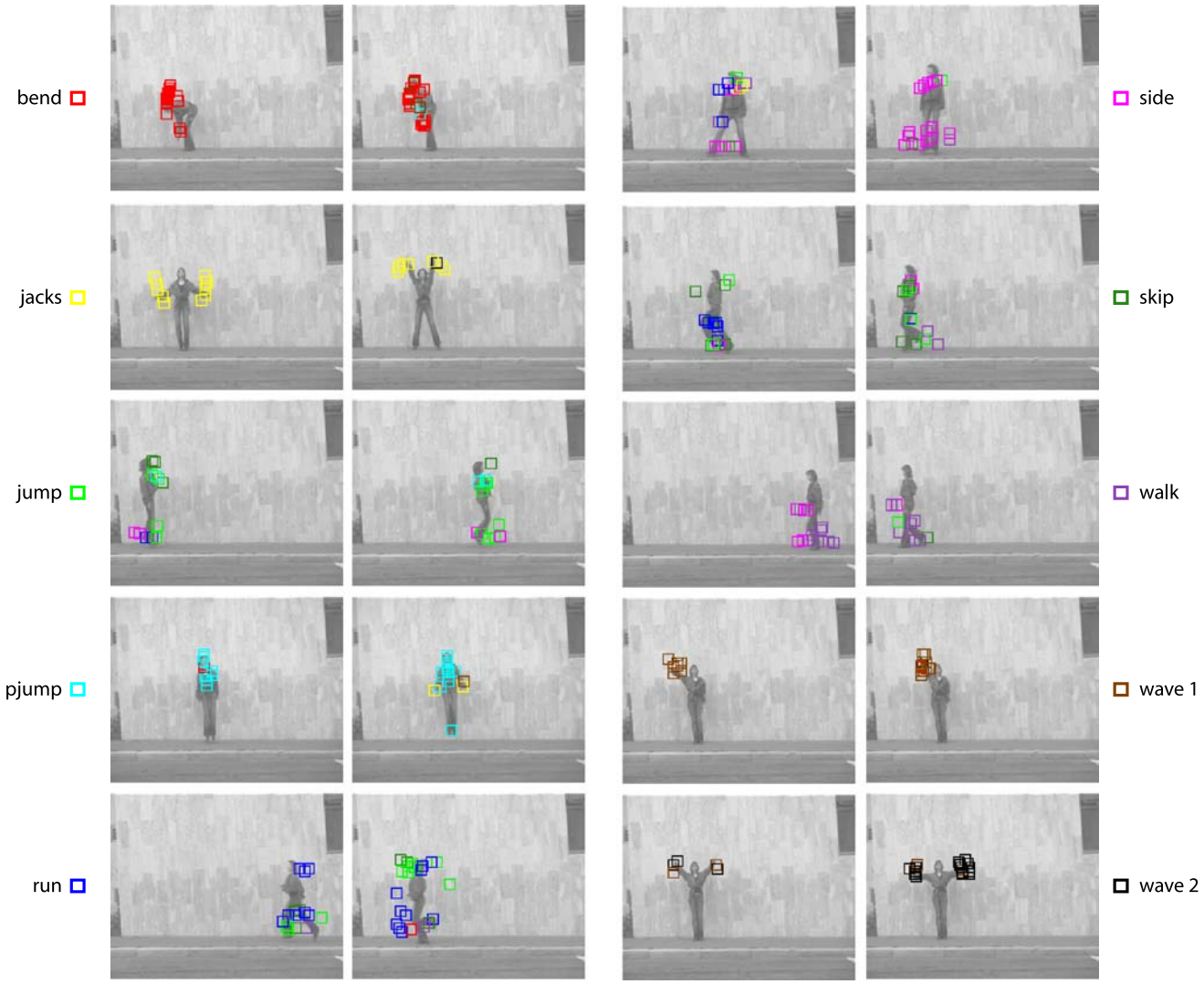


**Fig. 14** Example frames from testing sequences in the Weizmann Human Action dataset (Blank et al. 2005). The spatial-temporal patches in each sequence are automatically colored according to action class that most likely generated its corresponding spatial-temporal word. Although some of the words are assigned to the wrong topic, most interest points are assigned to the correct action for each video. Consistently, the predicted action class corresponds to the actual ground truth. The figure is best viewed in color and with PDF magnification

tains sequences with camera motion, background clutter and aggressive view point changes.

We detect and describe interest points using the procedure detailed in previous sections. The detector parameters are set to $\sigma = 2$ and $\tau = 1.2$, and the dimensionality of the corresponding descriptors is reduced to 100. We use all the videos available in training to build the codebook, using $k$-means.

We use the LOO procedure 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 pLSA model for the figure skating dataset is shown in Fig. 15 using 1200 codewords. The average performance of our algorithm is

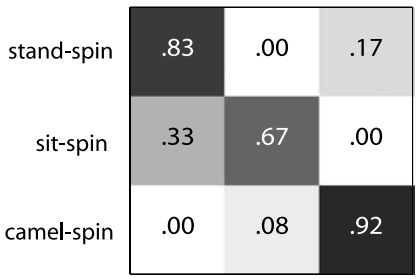


**Fig. 15** Confusion matrix for the figure skating dataset using 1200 codewords (performance average = 80.67%). Our algorithm can successfully categorize the figure skating actions in the presence of camera motion and cluttered background

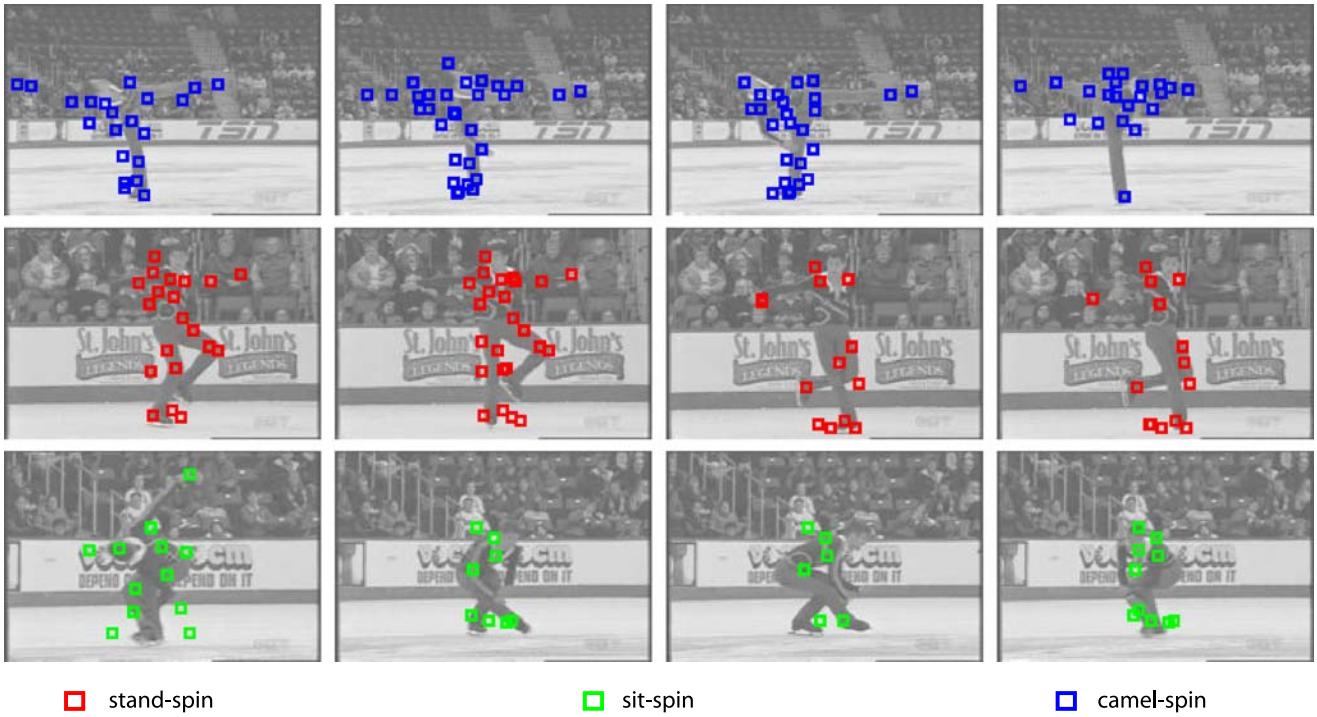


**Fig. 16** Example frames from testing sequences in the figure skating dataset. The interest points in each sequence are automatically colored according to the action class that most likely generated its corresponding spatial-temporal word. Note that only spatial-temporal interest points from the detected action category are shown. The figure is best viewed in color and with PDF magnification

80.67%. Note that in spite of the simple representation, our method can perform well in a very challenging dataset with camera motion, scale changes and severe occlusions.

Additionally, the learned 3-class pLSA model can be used for action localization as shown in Fig. 16.

### 4.2 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.

When the testing sequence is significantly long, we divide it into subsequences using a sliding temporal window. We process such subsequences independently and obtain classification decisions for each of them. This is necessary due to the nature of our representation: the lack of relative temporal ordering of features in our "bag of words" representation does not provide means to assign labels at different time instances within a video; instead, the analysis is made for the complete sequence. Thus, by dividing the original long video into subsequences, our method can assign labels to each subsequence within the long sequence.

First, suppose we encounter a testing video that contains multiple simultaneous human action categories. For multiple actions in a single sequence, and assuming we have learnt models employing the pLSA framework, we first identify how many action categories are significantly induced by $P(z_k|w_i, d_j)$. This is possible since $P(z_k|w_i, d_j)$ provides a measurement of the content of each action in the testing sequence. Thus, we allow the algorithm to select more than one action class if $P(z_k|w_i, d_j)$ is bigger than some threshold for more than one $k$. However, we need to assume that the number of actions present in the sequence is much less than the number of learnt actions categories $K$; in the extreme case that all action classes are present in the sequence, the distribution $P(z_k|w_i, d_j)$ should be very close to the uniform distribution and we cannot find salient action classes. Once the action categories of interest have been identified, the algorithm can select only the spatial-temporal interest points that are assigned to those classes, and apply $k$-means to the spatial position of these space-time patches. The number of clusters is set equal to the number of significant action categories. In order to label the resulting clusters with an action class, each word votes for its assigned action within its cluster. Finally a bounding box is plotted according to the principle axis and eigen-values induced by the spatial distribution of video words in each cluster. A further assumption that has to be made in order to use this procedure is that the actions must be performed in spatially distinct positions. Figure 17 illustrates examples of multiple

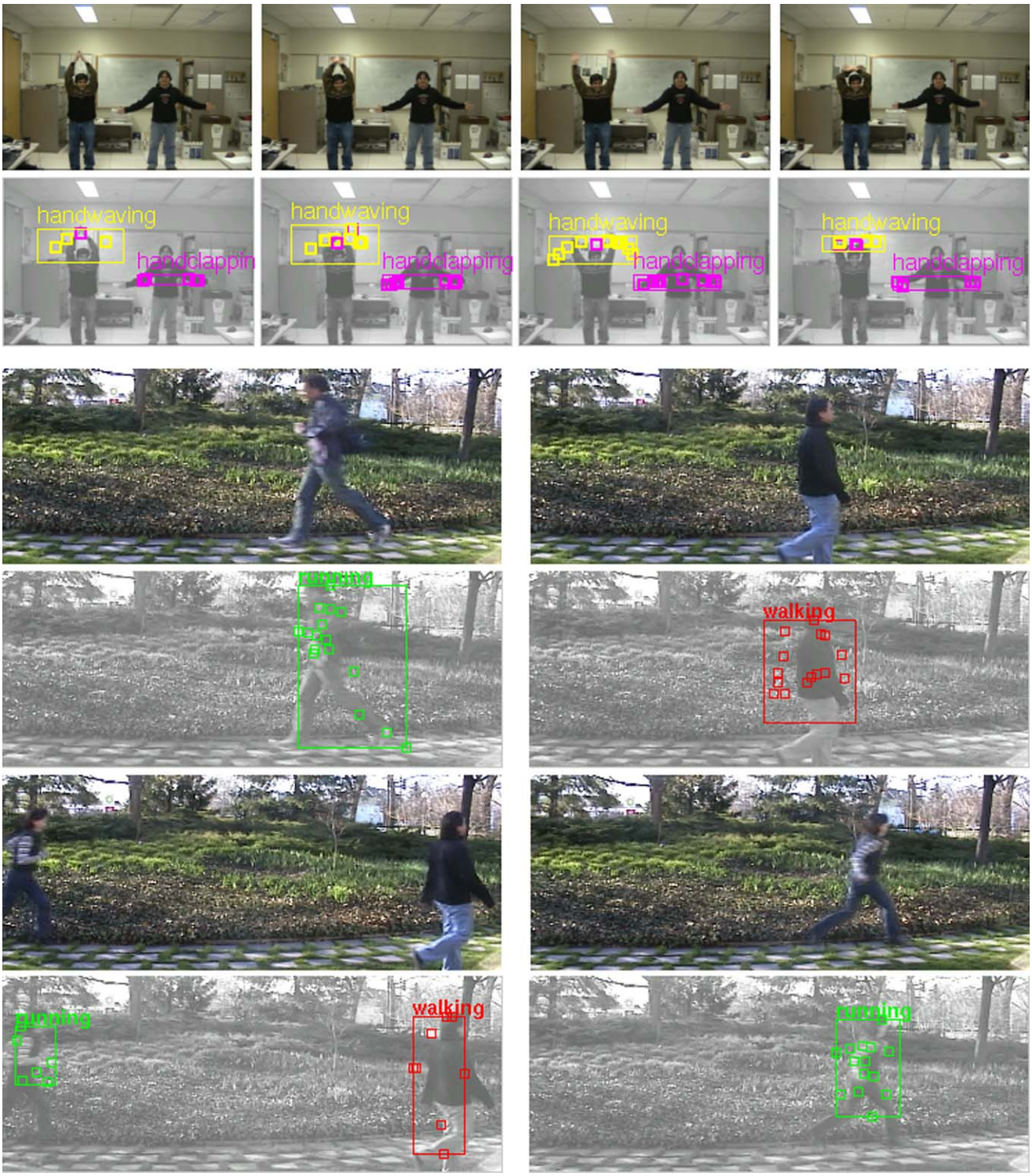


**Fig. 17** Multiple action recognition and localization in long and complex video sequences. The algorithm automatically detects the number of significant actions in a windowed subsequence around each frame. Then a clustering technique is used to group the interest points according to their spatial position. A bounding box is placed around each cluster with the automatically detected action label. The figure is best viewed in color and with PDF magnification

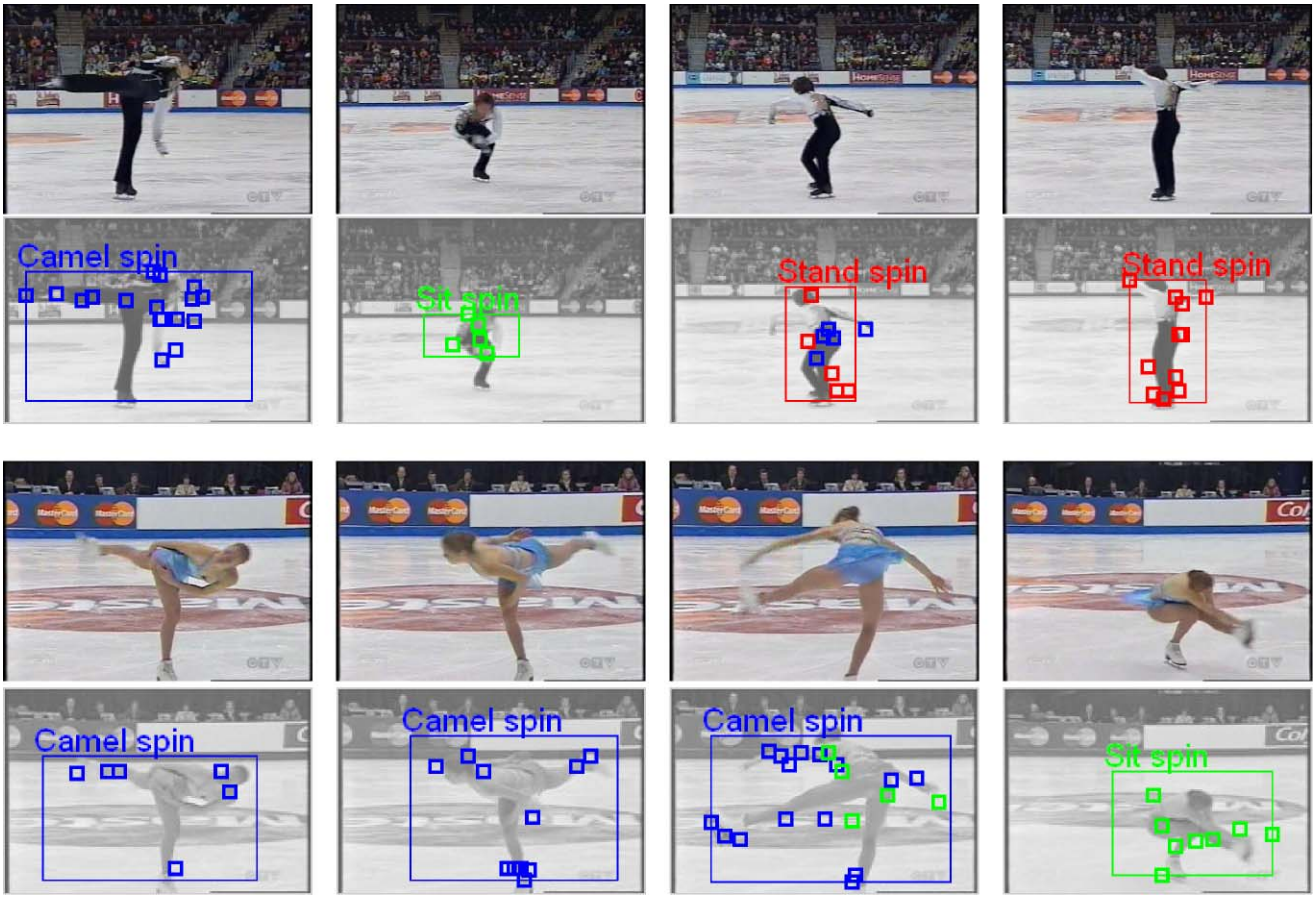


**Fig. 18** Multiple action recognition and localization in long and complex video sequences. The algorithm automatically detects the number of significant actions in a windowed subsequence around each frame. Then a clustering technique is used to group the interest points according to their spatial position. A bounding box is placed around each cluster with the automatically detected action label. The figure is best viewed in color and with PDF magnification

actions recognition and localization in one video sequence using a six-class pLSA model learnt from the KTH dataset (Sect. 4.1.1).

The second scenario we want to explore consists of a long testing video sequence that contains one subject performing different actions through time. Consider for example the long skating video sequences in Fig. 18. Assuming we have learnt models with pLSA, we perform recognition by extracting a windowed sequence around each frame, and identifying which actions receive a high weight according to $P(z_k|w_i, d_j)$. Thus the middle frame in the windowed sequence is labeled with the identified action category. Figure 18 shows examples of action recognition in a long figure skating sequence. Here we employ the three-class model learnt from figure skating sequences containing a single action (Sect. 4.1.3). The three actions (stand-spin, camel-spin and sit-spin), are correctly recognized and labeled using different colors. (Please refer to a video demo available at: http://vision.cs.princeton.edu/niebles/humanactions.htm.)

## 5 Conclusion

In this paper, we have presented an unsupervised learning approach, i.e., a "bag of spatial-temporal words" model combined with a space-time interest points detector, for human action categorization and localization. Using three challenging datasets, our experiments show that the classification performance using our unsupervised learning approach is on par with the current state-of-the-art results obtained by fully supervised training. 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. In addition, we plan to further investigate the possibilities of using a unified framework by combining generative and discriminative models for human action recognition. For similar actions (e.g., "running" and "walking"), the classification may benefit from a discriminative model. Finally, other interesting explorations include

richer models that can incorporate geometric information, such as the spatial and temporal arrangement of local features (Niebles and Fei-Fei 2007), as well as explicit models for the human body.

**Acknowledgements** The authors would like to thank Professor Greg Mori for providing the figure skating dataset. J.C.N. is supported by a Fulbright-COLCIENCIAS-DNP grant and Universidad del Norte. L.F.-F. is supported by the Microsoft Research New Faculty Fellowship.

## References

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the tenth IEEE international conference on computer vision* (Vol. 2, pp. 1395–1402). Los Alamitos: IEEE Computer Society.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267.

Boiman, O., & Irani, M. (2005). Detecting irregularities in images and in video. In *Proceedings of the tenth IEEE international conference on computer vision* (Vol. 1, pp. 462–469). Los Alamitos: IEEE Computer Society.

Cheung, V., Frey, B. J., & Jojic, N. (2005). Video epitomes. In *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition* (Vol. 1, pp. 42–49). Los Alamitos: IEEE Computer Society.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European conference on computer vision* (Vol. 2, pp. 428–441).

Dance, C., Willamowski, J., Fan, L., Bray, C., & Csurka, G. (2004). Visual categorization with bags of keypoints. In *ECCV international workshop on statistical learning in computer vision*.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2nd joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance* (pp. 65–72).

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the ninth IEEE international conference on computer vision* (Vol. 2, pp. 726–733). Los Alamitos: IEEE Computer Society.

Fanti, C., Zelnik-Manor, L., & Perona, P. (2005). Hybrid models for human motion recognition. In *Proceedings of the tenth IEEE international conference on computer vision* (Vol. 1, pp. 1166–1173). Los Alamitos: IEEE Computer Society.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition* (pp. 524–531). Los Alamitos: IEEE Computer Society.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, *61*(1), 55–79.

Feng, X., & Perona, P. (2002). Human action recognition by sequence of movelet codewords. In *1st international symposium on 3D data processing visualization and transmission (3DPVT 2002)* (pp. 717–721).

Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2005). Learning object categories from Google's image search. In *Proceedings of the tenth international conference on computer vision* (Vol. 2, pp. 1816–1823). Los Alamitos: IEEE Computer Society.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the fourth Alvey vision conference* (pp. 147–152).

Hoey, J. (2001). Hierarchical unsupervised learning of facial expression categories. In *IEEE workshop on detection and recognition of events in video* (pp. 99–106).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57), August 1999.

Kadir, T., & Brady, M. (2003). Scale saliency: a novel approach to salient feature and scale selection. In *International conference on visual information engineering* (pp. 25–28).

Ke, Y., Sukthankar, R., & Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Proceedings of the tenth IEEE international conference on computer vision* (pp. 166–173). Los Alamitos: IEEE Computer Society.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2–3), 107–123.

Laptev, I., & Lindeberg, T. (2006). Local descriptors for spatio-temporal recognition. In *Lecture notes in computer science* (Vol. 3667). *Spatial coherence for visual motion analysis, first international workshop, SCVMA 2004,* Prague, Czech Republic, 15 May 2004. Berlin: Springer.

Niebles, J. C., & Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In *Proceedings of the 2007 IEEE computer society conference on computer vision and pattern recognition*. Los Alamitos: IEEE Computer Society.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. In *Proceedings of British machine vision conference 2006* (Vol. 3, pp. 1249–1258), September 2006.

Oikonomopoulos, A., Patras, I., & Pantic, M. (2006). Human action recognition with spatiotemporal salient points. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *36*(3), 710–719.

Ramanan, D., & Forsyth, D. A. (2004). Automatic annotation of everyday movements. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in neural information processing systems* (Vol. 16). Cambridge: MIT Press.

Savarese, S., Winn, J. M., & Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlations. In *Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition*. Los Alamitos: IEEE Computer Society.

Schmid, C., Mohr, R., & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, *2*(37), 151–172.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *ICPR* (pp. 32–36).

Shechtman, E., & Irani, M. (2005). Space-time behavior based correlation. In *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition* (Vol. 1, pp. 405–412). Los Alamitos: IEEE Computer Society.

Sidenbladh, H., & Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, *54*(1-3), 181–207.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., & Freeman, W. T. (2005). Discovering objects and their location in images. In *Proceedings of the tenth IEEE international conference on computer vision* (pp. 370–377), October 2005. Los Alamitos: IEEE Computer Society.

Song, Y., Goncalves, L., & Perona, P. (2003). Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(25), 1–14.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*. Los Alamitos: IEEE Computer Society.

Wang, Y., Jiang, H., Drew, M. S., Li, Z.-N., & Mori, G. (2006). Unsupervised discovery of action classes. In *Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition*. Los Alamitos: IEEE Computer Society.

Xiang, T., & Gong, S. (2005). Video behaviour profiling and abnormality detection without manual labelling. In *Proceedings of the tenth IEEE international conference on computer vision* (pp. 1238–1245). Los Alamitos: IEEE Computer Society.

Yilmaz, A., & Shah, M. (2005). Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Proceedings of the tenth IEEE international conference on computer vision* (Vol. 1, pp. 150–157). Los Alamitos: IEEE Computer Society.

Zhong, H., Shi, J., & Visontai, M. (2004). Detecting unusual activity in video. In *Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition* (pp. 819–826). Los Alamitos: IEEE Computer Society.