Data Min Knowl Disc
DOI 10.1007/s10618-010-0195-5

Hierarchical visual event pattern mining and its applications

Peng Cui · Zhi-Qiang Liu · Li-Feng Sun · Shi-Qiang Yang

Received: 6 May 2009 / Accepted: 8 July 2010

Abstract In this paper, we propose a hierarchical visual event pattern mining approach and utilize the patterns to address the key problems in video mining and understanding field. We classify events into primitive events (PEs) and compound events (CEs), where PEs are the units of CEs, and CEs serve as smooth priors and rules for PEs. We first propose a tensor-based video representation and *Joint Matrix Factorization* (JMF) for unsupervised primitive event categorization. Then we apply frequent pattern mining techniques to discover compound event pattern structures. After that, we utilize the two kinds of event patterns to address the applications of event recognition and anomaly detection. First we extend the Sequential Monte Carlo (SMC) method to recognition of live, sequential visual events. To accomplish this task we present a scheme that alternatively recognizes *primitive* and *compound* events in one framework. Then, we categorize the anomalies into abnormal events (never seen events) and abnormal contexts (rule breakers), and the two kinds of anomalies are detected simultaneously by embedding a deviation criterion into the SMC framework. Extensive experiments have been conducted which demonstrate that the proposed approach is effective as compared to other major approaches.

Keywords Video mining · Matrix factorization · Event recognition · Anomaly detection

P. Cui (✉) · L.-F. Sun · S.-Q. Yang
Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China
e-mail: cuip@tsinghua.edu.cn

Z.-Q. Liu
School of Creative Media, City University, Kowloon, Hong Kong

Published online: 24 July 2010

1 Introduction

With the proliferation of video data captured by surveillance cameras or household devices, it would make most sense to develop approaches to automatically annotating video data and discovering knowledge from videos, for many fields like security, HCI, health care etc.. Our research presented in this paper is targeted to mining event patterns for both event recognition and anomaly detection from common stable-camera videos, and was motivated by the following questions:

(1) What should be the basic units of captured videos?

As the basic units of videos, events are defined as temporal objects (Zelnik-Manor and Irani 2001). However, events can be divided into higher level and low level events in terms of visual and semantic complexity. For ease of understanding without loss of generality, we would suggest two layers in this paper: the primitive event (such as walking, running, and waving) and compound event (such as shopping). The primitive event is the basic unit to organize the video content, and a compound event is constituted by a sequence of ordered primitive events.

(2) How to model the basic units so that we can decompose video content into low dimensional space?

As primitive events (also called actions without confusion in this paper) play the role of a bridge between low level visual features and high-level semantics, the modeling of primitive events is the most crucial step. In the literature, many supervised event recognition methods can be used for primitive event modeling. However, The ever-increasing video content makes the unsupervised methods more appealing in that the increasing amount of data can be exploited without the expense of detailed human annotation. In addition, this is a challenging problem due to the unconstrained environment, cluttered background, photometric variance of events, and executing rate variance of events. All these noises lead to a large intra class variance and weak separability between classes. It is therefore desirable to remove, if failing this, to reduce their effects. In this paper, we regard event as a composition of *dynamic pixels* (By dynamic pixels, we mean the pixel' value changes with time. In actual, each dynamic pixel correspond to a time series). These dynamic pixels are quantized into pixel prototypes, and then each action sample can be represented by a 2-D matrix, denoted as *action matrix* (AM), which is constituted by spatially distributed pixel prototypes. If we put all action samples together, the whole data set forms a 3-D tensor (denoted as *multi-action tensor* (MAT)) with the new dimension representing the sample index. The action categorization problem is then converted into a clustering problem on MAT.

(3) How to discover higher level event structures (compound events) from primitive event sequences?

Compound events are of great significance in two aspects: First, they contain higher-level semantics, which can better interpret the video content. Second, it serves as the context of primitive events. We regard compound events as repetitive primitive event patterns, where frequent pattern mining techniques are applied. In addition, we categorize compound events into Strong Correlation Pattern (SCP) and Weak Correlation Pattern (WCP) according to the inner correlations of primitive events, where WCPs are used as smooth prior, and SCPs are used as rules.

(4) How to utilize the primitive events and compound event patterns to address the problems in computer vision and video understanding?

As visual features are incomplete, visual ambiguity phenomena are inevitable in event recognition, especially in low-resolution surveillance video. To remedy this, we borrow the idea of smooth trajectory from motion tracking research field, where the smoothness of trajectory plays an important role in disambiguating nearby objects. Similarly, the trajectories of primitive events are smooth in the context of compound events. Therefore we may unify the recognition of primitive and compound events into one Sequential Monte Carlo (SMC) framework.

In addition, anomaly detection is the ultimate goal of most intelligent surveillance systems. We categorize anomalies into two classes: abnormal events and abnormal contexts. The former are events that look abnormal, without any need of context assistant, e.g. falling down, and fighting. The latter are events that look normal, but in actual abnormal given the context, e.g. a man walking away is normal with the context of 'Walk-LeftBag-PickUpBag', but abnormal with the context of 'Walk-LeftBag' without picking the bag up. In the literature, few researches have done on abnormal context. In our method, abnormal contexts are regarded as rule-breakers. As the rules (SCP) can be achieved in compound event discovery, the abnormal contexts are detected straightforwardly. We propose a criterion to evaluate the deviation from rules based on K-L divergence, by which the abnormal event and abnormal context can be detected in one framework.

In this paper, we present an effective and practical approach to these problems.

(1) We propose a novel visual representation of action, which naturally compact the video data from 3-D to 2-D with maintaining sufficient information. Also, we propose a joint matrix factorization method to simultaneously factorize a symmetric matrix and an asymmetric matrix for action categorization.

(2) We exploit pattern mining techniques to discover compound event structures, which provide powerful cues for event recognition and anomaly detection.

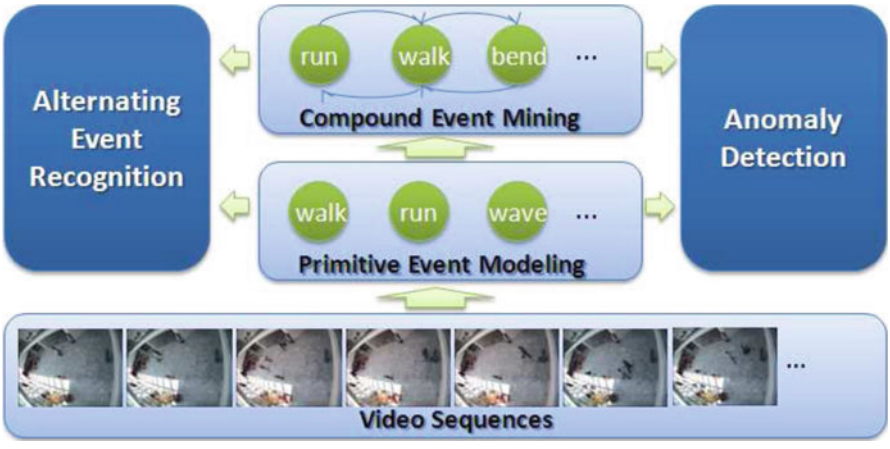

Fig. 1 The flowchart of the proposed methods

(3) We extend the SMC framework to simultaneously and interactively incorporate the primitive and compound event patters for better event recognition and anomaly detection.

The diagram of the proposed methods are shown in Fig. 1. After reviewing the related work in Sect. 2, we firstly model the primitive events in unsupervised fashion from the raw video sequences (in Sect. 3), and then discover the compound event patterns based on primitive event label sequences (in Sect. 4). After that, we exploit the two layers of events for two applications: alternating event recognition and anomaly detection (in Sect. 5), which makes full use of the duality between primitive events and compound events to alleviate the problems of lacking smoothing prior for event recognition and context for anomaly detection. Finally, we evaluate the proposed methods in Sect. 6 before drawing the conclusions in Sect. 7.

2 Related work

2.1 Primitive event recognition

Most published works on action/ activity/ behavior/ event recognition use object level features. Stauffer and Grimson (2000) and Makris and Ellis (2002) regarded pedestrians as rigid objects and considered only trajectory information. In order to describe human behaviors in more detail, Davis and Taylor (2002) and Haritaoglu et al. (2000) detected body parts and tracked them to discriminate different actions, which suffered from self-occlusion and view variance. Xiang and Gong (2005), Cui et al. (2007) and Zhong et al. (2004) used low level information directly to construct event models without the need of motion tracking and object segmentation, which were more robust. Efros et al. (2003) and Ke et al. (2005) proposed several sets of features based on optic flow, which were very sensitive to view variation. Recently, volumetric features for event recognition have become popular. Yilmaz and Shah (2005) proposed to use spatial-temporal volumes for action recognition by contour projection. Blank et al. (2005) generalized those techniques for the analysis of 2D shapes for describing spatial-temporal volumes.

Due to the difficulties of action recognition, most methods use supervised models to realize it (Gilbert et al. 2009; Sun et al. 2009; Lin et al. 2009). There are only a few researchers explored unsupervised methods. Niebles et al. (2008) introduce the generative models pLSA and LDA into action recognition field. In recent years, the matrix factorization theory (Li and Chris 2006; Chris et al. 2006) has been revisited for unsupervised categorization. It is widely applied in document clustering (Xu et al. 2003), EEG representation decomposition (Morup et al. 2006), social network analysis (Wang et al. 2008), and bioinformatics (Wang and Li 2007), etc. Most of these methods are performed on a single matrix or multiple independent matrices. How to simultaneously factorize dependent matrices is still challenging. Tensor is natural for high-dimensional dynamic data representation, and some pioneering works have been published on face and action representation and modeling by tensor decomposition (Tao et al. 2007a,b, 2008a,b), but the efficiency issue is what tensor methods have to face, especially for large scale video data.

2.2 Compound event recognition

More recently a few researchers started looking at higher level semantics and interactions. Xiang and Gong (2006) proposed to use multi-linked Hidden Markov Model and model selection techniques to learn the underlying event structure, which was theoretically solid and achieved promising results in correlation mining on multiple event sequences. Yamamoto et al. (2006) proposed a context free grammar method for action recognition, which fully exploited prior knowledge but was hard to be scalable. Ivanov and Bobick (2000) proposed a context-free parsing method combining with HMM for activity recognition. The context-free grammar was manually defined, which was infeasible for large event sets. Hakeem and Shah (2005) proposed an event correlation graph, on which frequent patterns were discovered by clustering on the graph. The most commonly used model for activity recognition is HMM (Robertson and Reid 2006; Duong et al. 2005; Du et al. 2006). However, to compute HMM one has to tune many parameters using a large data set, which makes it infeasible for online recognition.

2.3 Anomaly detection

There exist two major ideas for anomaly detection. One is to directly train or define event models for anomalies. This kind of methods is rarely adopted because of the sparsity of anomalies. Chan et al. (2004), manually defined HMMs for abnormal events, which was only practical in constrained scenes. The other is to train normal models, and anomalies are detected by an deviation measure. Zhong et al. (2004) built prototypes for normal events, and the distances between samples with the prototypes were used as the deviation measure. Boiman and Irani (2005) detected irregularity from a generative viewpoint, and highlighted the event that cannot be reconstructed by other samples as irregularities. Vaswani et al. (2003) detected the anomalies in motion trajectory by proposing a novel deviation criterion. These methods are for the detection of abnormal events. Few approaches have been proposed for detection of abnormal contexts.

3 Learning of primitive event models

3.1 Primitive event representation

The representation method in this paper is motivated by the fact that all video sequences can be displayed on a 2D screen which is constituted by a number of dynamic pixels, and all video content is understood through these dynamic pixels changes. Then we suppose that there exist dynamic pixel patterns which are related to high-level semantics, such as actions. Given a video sequence, each dynamic pixel corresponds to a time series. We observed that there are large correlations among different time series (as is shown in Fig. 2). For example, in Fig. 2a, the pixel curves of the first (or second or third) row pixels are similar. If we group these time series into several prototype pixels as visual words, the whole video sequence can be represented by a bag of

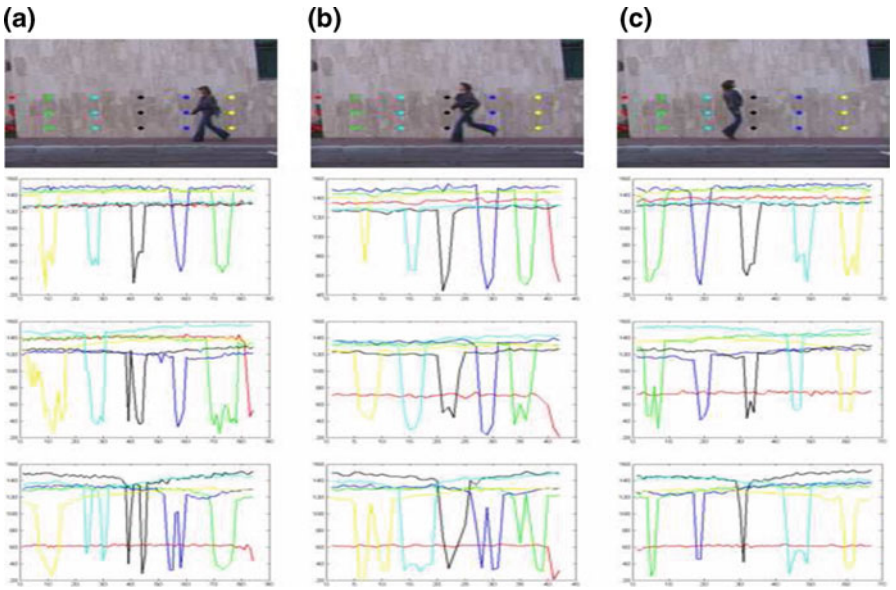

Fig. 2 Dynamic pixel curves. The horizontal axes represent the frame number, and the vertical axes indicate the grey-scale value. The three rows of dots on the frames are observing dynamic pixels, whose pixel curves are plotted on the bottom three rows of figures, with one row of figures corresponding to one row of pixels. The second row of figures correspond to the first row of observing pixels, and so forth. On each figure, the horizontal axis indicates the time, and the vertical one represent the pixel intensity.The primitive events are respectively (a) walking; (b) running; (c) jumping

spatially distributed words, which is convenient for pattern mining, and particularly, in our case, the action categorization. The technical details are specified in following sections.

3.1.1 Dynamic pixel descriptor

As mentioned, given a video sequence, each dynamic pixel corresponds to a time series. Features reflecting the characteristics of the time series should be extracted to form the descriptors of dynamic pixels. In our case, the Discrete Fourier Transformation (DFT) is adopted to form the dynamic pixel descriptors.

Given a video sequence $\mathbb{X}$, we denote $\mathbf{x}_{i,j} = [x_{i,j,t}]_{t=0,\ldots,n-1}$ as a dynamic pixel, where (i, j) indicates the spatial position of the dynamic pixel, t is the temporal index (i.e. the frame index in a video sequence), and n is the length of the time series (i.e. the total number of frames in a video sequence). We use $\widetilde{\mathbf{X}}_{i,j} = [X_{i,j,f}]_{f=0,\ldots,n-1}$ to denote the DFT of $\mathbf{x}_{i,j}$:

$$\widetilde{X}_{i,j,f} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_{i,j,t} e^{\frac{-i2\pi tf}{n}} \tag{1}$$

where $i = \sqrt{-1}$ is the imaginary unit.

In order to make the descriptor compact and noise robust, we select $k(k < n)$ coefficients from $\widetilde{\mathbf{X}}_{i,j}$. We select the first k coefficients in $\widetilde{\mathbf{X}}_{i,j}$ as the dynamic pixel descriptor $\mathbf{X}_{i,j} = [X_{i,j,f}]_{f=0,\ldots,k-1}$.

However, there are two problems with the DFT descriptors.

(1) Scaling Variance. As the DFT captures the global characteristics of time series, the pixel descriptors largely depend on the intensity value of dynamic pixels, which results in different descriptors for exact the same action performed by the same person wearing different colors, which is known as scaling variance. To remedy this, we realize the scaling invariance by normalizing the pixel time series $\mathbf{x}_{i,j}$

$$\mathbf{x}'_{i,j} = \frac{\mathbf{x}_{i,j} - \mu_{i,j}}{\sigma_{i,j}}, \tag{2}$$

where $\mu_{i,j}$ and $\sigma_{i,j}$ are respectively the mean value and standard deviation of $\mathbf{x}_{i,j}$.

(2) Phase Variance. Two asynchronous series (i.e. series with similar amplitudes but different phases) generate totally different DFT descriptors. However, in action video sequences, a simple action, e.g. raising an arm, would cause a batch of asynchronous series. These asynchronous series are in actual caused by the same action, so they should be regarded as similar ones. To realize this, we discard the phase information and only use the amplitude information to form the descriptor $\mathbf{X}'_{i,j} = [|X_{i,j,f}|^2]_{f=1,\ldots,k}$.

Hereto, the $w \times h \times n$ sized video sequences are transformed into $w \times h \times k$ sized action tensor.

3.1.2 Pixel prototype formation

To acquire a common set of pixel prototypes, we put all descriptors of all video sequences together. K-means is performed to find out R clusters. The centroids of clusters are used as pixel prototypes. Then the common pixel prototype set $\{\mathcal{X}_i | i = 1, \ldots, r\}$ is formed. When a new video sequence is input, the dynamic pixels are assigned with pixel prototype labels using K Nearest Neighbor (KNN):

$$label(\mathbf{x}_{i,j}) = \arg\min_r |\mathbf{X}'_{i,j} - \mathcal{X}_r|^2. \tag{3}$$

Then the original 3-D action tensor is represented as a 2D action matrix, which is a spatial distribution of R pixel prototype labels.

3.1.3 Multi-action tensor construction

In our case, by regarding the pixel prototypes as visual words, and actions as documents, we can also solve the action categorization problem by a Bag-of-Words (BOW) like method. However, there are some non-trivial differences between our method and textual BOW method due to the characteristics of video:

(1) The spatial distribution should be incorporated in visual word similarity measurement.

(2) The intrinsic similarities of different prototypes are measurable in feature space.

Considering the above issues, we calculate the distance of prototypes by three weighted components:

$$Dist(r_1, r_2) = \omega_1 D_c(r_1, r_2) + \omega_2 D_f(r_1, r_2) + \omega_3 D_{sp}(r_1, r_2), \tag{4}$$

where $Dist$ is $R \times R$ sized distance matrix; D_c, D_f, $and\, D_{sp}$, respectively represents the distance on prototype centroids, occurring frequency, and spatial distribution; and $\omega_1, \omega_2, \omega_3$ are all non-negative and meet $\omega_1 + \omega_2 + \omega_3 = 1$. The three distances are calculated as follows.

(1) **Spatial distribution distance**

We modify the Pyramid Match Kernel (Ling and Soatto 2007) for Geographical space to calculate this distance. Given a pair of prototypes $\mathcal{X}_{r_1}$ and $\mathcal{X}_{r_2}$, and an 2D action matrix Θ, we can derive two point sets Ω_{r_1} and Ω_{r_2}, where $\Omega_r = (i, j)|\Theta(i, j) = r; i = 1, \ldots, w; j = 1, \ldots, h$ contains all the positions of dynamic pixels labeled as prototype r. The feature extraction function on these point sets are defined as:

$$\Psi(\Omega_r) = [H_0(\Omega_r), H_1(\Omega_r), \ldots, H_L(\Omega_r)], \tag{5}$$

where i is the index of pyramid level, $H_i(\Omega_r)$ is a histogram which has $b_i = \frac{w \times h}{2^{2i}}$ bins, and the $H_i^j(\Omega_r)$ represents the points fall into the j^{th} bin. Note that the H_i is a one-dimensional histogram which is made by concatenating the two-dimensional histogram in book-order. Then the spatial distribution difference at level i is defined as

$$\mathcal{I}(H_i(\Omega_{r_1}), H_i(\Omega_{r_2})) = \sum_{j=1}^{b_i} |H_i^j(\Omega_{r_1}) - H_i^j(\Omega_{r_2})|. \tag{6}$$

For ease of denotation, we replace $\mathcal{I}(H_i(\Omega_{r_1}), H_i(\Omega_{r_2}))$ with $\mathcal{I}_i(\Omega_{r_1}, \Omega_{r_2})$. With the level being coarser, the difference gradually reduced. We define the new reduced difference at level i as $\mathcal{I}_{i-1} - (I)_i$, then the spatial distribution distance is defined as

$$\widetilde{D}_{sp}(r_1, r_2) = \sum_{i=1}^{L} \frac{1}{2^{L-i}} (\mathcal{I}_{i-1}(\Omega_{r_1}, \Omega_{r_2}) - \mathcal{I}_i(\Omega_{r_1}, \Omega_{r_2})). \tag{7}$$

The $\frac{1}{2^{L-i}}$ serves as a penalty factor which is higher for new reduced difference at coarser level. Finally, the distance is normalized by

$$D_{sp}(r_1, r_2) = \frac{\widetilde{D}_{sp}(r_1, r_2)}{\widetilde{D}_{sp}(r_1, r_1)\widetilde{D}_{sp}(r_2, r_2)}. \tag{8}$$

The main difference between spatial distribution distance and pyramid match kernel (Ling and Soatto 2007) lies in that the former is a distance matric defined in geographical space while the latter is a similarity matric defined in feature space.

(2) **Occurring frequency distance**

We define a function to calculate the occurring frequency of a prototype

$$\mathcal{F}(r) = \sum_{i,j} \delta_r(\Theta(i, j)), \tag{9}$$

where $\delta_a(b)$ is a delta function which is one when $a = b$, and zero otherwise. Then, the D_f is calculated by

$$D_f(r_1, r_2) = \frac{|\mathcal{F}(r_1) - \mathcal{F}(r_2)|}{max(\mathcal{F}(r_1), \mathcal{F}(r_2))}. \tag{10}$$

(3) **Centroid distance**

The centroid distance of two prototypes is measured by the Euclidean distance:

$$\widetilde{D}_c(r_1, r_2) = \sqrt{\sum_{k=1}^{K} (\mathcal{X}_{r_1,k} - \mathcal{X}_{r_2,k})^2}, \tag{11}$$

and then normalized as

$$D_c(r_1, r_2) = \frac{\widetilde{D}_c(r_1, r_2)}{\max\limits_{i,j}(\widetilde{D}_c(r_i, r_j))}. \tag{12}$$

Hereto, by substituting (8), (10) and (12) to (4), we acquire the distances of pixel prototype pairs in an action matrix. After that, we transform the distance into similarity by a Gaussian like function:

$$\mathbf{A}(r_1, r_2) = \exp\left(-\frac{Dist(r_1, r_2)}{\sigma}\right), \tag{13}$$

where $\mathbf{A}$ is the affinity matrix transformed from the 2D action matrix Θ, and σ is the deviation factor. All the diagonal elements in $\mathbf{A}$ are set to one, and if prototype r is absent from Θ, the elements in corresponding row and column in $\mathbf{A}$ are set to zero except the diagonal element. The achieved affinity matrices are guaranteed to be positive definite, which is convenient for further processing.

Actions (video sequences) are represented by a $R \times R$ sized affinity matrices, which are dimensionally aligned. By putting them together along a video index coordinate, the MAT (Multi-Action Tensor) is constructed, as is shown in Fig. 3.

3.2 Action categorization on MAT

In this section, we propose a matrix factorization method which simultaneously cluster pixel prototypes into signatures, and video sequences into action classes.

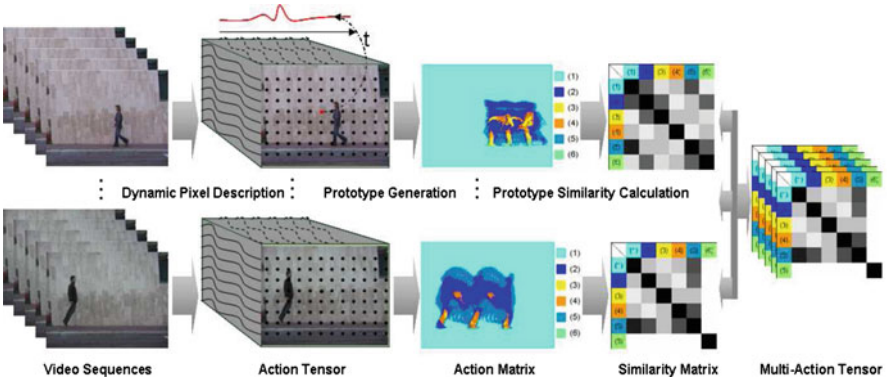

Fig. 3 The framework of the proposed method

3.2.1 Problem formulation

There are two types of matrices in MAT. One is the affinity matrices $\mathbf{A}_i$ corresponding to video sequences, which is symmetric; and the other type of matrices $\mathbf{B}_i$ are constituted by co-located columns of affinity matrices: $\mathbf{B}_i = [\mathbf{A}_1(i), \ldots, \mathbf{A}_n(i)]$, which is asymmetric. Thus, we define MAT as a case of semi-supersymmetric tensor. Let us denote the cluster indicators for pixel prototypes as $\mathbf{U} = [\mathbf{U}(1), \ldots, \mathbf{U}(\widetilde{R})]$, and the cluster indicators for video sequences as $\mathbf{V} = [\mathbf{V}(1), \ldots, \mathbf{V}(C)]$, where $\widetilde{R}$ and C are respectively the number of pixel signatures and action classes. Then the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ can be approximated by:

$$\mathbf{A}_i \approx \mathbf{U}\mathbf{S}\mathbf{U}^T, \tag{14}$$

$$\mathbf{B}_i \approx \mathbf{U}\mathbf{R}\mathbf{V}^T. \tag{15}$$

Our goal is to solve for cluster indicators which optimally approximate $\mathbf{A}_i$ and $\mathbf{B}_i$, i.e. we should solve the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n} \|\mathbf{A}_i - \mathbf{U}\mathbf{S}\mathbf{U}^T\|_F^2 + \lambda \sum_{i=1}^{n} \|\mathbf{B}_i - \mathbf{U}\mathbf{R}\mathbf{V}^T\|_F^2 \\ &s.t. \quad \mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}^T\mathbf{V} = \mathbf{I}, \end{aligned} \tag{16}$$

where $\| Q \|_F^2$ is the Frobenius norm of matrix Q.

3.2.2 Joint matrix factorization

In (16), there are two components. One is for pixel prototype clustering, and the other is for video sequence clustering based on the result of the former component. In order to simultaneously solve them, we propose an *joint matrix factorization* (JMF) method.

Directly solving the above optimization problem may cause the quartic form of $\mathbf{U}$, which makes the problem hard to solve. Therefore, we first apply Cholesky

decomposition on each $\mathbf{A}_i$ as $\mathbf{A}_i = \mathbf{G}_i\mathbf{G}_i^T$, and solve the following problem

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n} \|\mathbf{G}_i - \mathbf{U}\tilde{\mathbf{S}}\|_F^2 + \lambda \sum_{i=1}^{n} \|\mathbf{B}_i - \mathbf{U}\mathbf{R}\mathbf{V}^T\|_F^2 \\ &s.t. \quad \mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}^T\mathbf{V} = \mathbf{I} \end{aligned} \tag{17}$$

where $\tilde{\mathbf{S}} = \mathbf{S}^{1/2}$. Since $\sum_{i=1}^{n} tr\left(\mathbf{G}_i^T\mathbf{G}_i\right)$ and $\sum_{i=1}^{n} tr\left(\mathbf{B}_i^T\mathbf{B}_i\right)$ are constants, then our goal is to minimize

$$\begin{aligned} \mathcal{J} = &-2\sum_{i=1}^{n} tr\left(\tilde{\mathbf{S}}^T\mathbf{U}^T\mathbf{G}_i\right) + ntr\left(\tilde{\mathbf{S}}^T\mathbf{U}^T\mathbf{U}\tilde{\mathbf{S}}\right) \\ &-\sum_{i=1}^{n} 2\lambda tr\left(\mathbf{V}\mathbf{R}^T\mathbf{U}^T\mathbf{B}_i\right) + \lambda ntr\left(\mathbf{V}\mathbf{R}^T\mathbf{U}^T\mathbf{U}\mathbf{R}\mathbf{V}^T\right) \end{aligned}$$

The partial derivative of $\mathcal{J}$ with respect to $\tilde{\mathbf{S}}$ and $\mathbf{R}$ are

$$\frac{\partial \mathcal{J}}{\partial \tilde{\mathbf{S}}} = -2\mathbf{U}^T\sum_{i=1}^{n}\mathbf{G}_i + 2n\tilde{\mathbf{S}} \tag{18}$$

$$\frac{\partial \mathcal{J}}{\partial \mathbf{R}} = -2\lambda\mathbf{U}^T\sum_{i=1}^{n}\mathbf{B}_i\mathbf{V} + 2\lambda n\mathbf{R} \tag{19}$$

Therefore

$$\tilde{\mathbf{S}} = \mathbf{U}^T\mathbf{G} \tag{20}$$

$$\mathbf{R} = \mathbf{U}^T\mathbf{B}\mathbf{V} \tag{21}$$

where $\mathbf{G} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i$, $\mathbf{B} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_i$. Bringing $\tilde{\mathbf{S}}$ and $\mathbf{R}$ back to $\mathcal{J}$, we have

$$\mathcal{J} = (n-2)tr(\mathbf{G}^T\mathbf{U}\mathbf{U}^T\mathbf{G}) + \lambda(n-2)tr(\mathbf{V}^T\mathbf{B}^T\mathbf{U}\mathbf{U}^T\mathbf{B}\mathbf{V})$$

Fixing $\mathbf{U}$, we should solve the following optimization problem

$$\begin{aligned} &\min_{\mathbf{V}} tr(\mathbf{V}^T\mathbf{B}^T\mathbf{U}\mathbf{U}^T\mathbf{B}\mathbf{V}) \\ &s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I} \end{aligned} \tag{22}$$

which can be solved via eigenvalue decomposition.

Fixing $\mathbf{V}$, we should solve the following optimization problem

$$\begin{aligned} &\min_{\mathbf{U}} tr(\mathbf{U}^T\left(\mathbf{G}\mathbf{G}^T + \lambda\mathbf{B}\mathbf{V}\mathbf{V}^T\mathbf{B}^T\right)\mathbf{U} \\ &s.t. \quad \mathbf{V}^T\mathbf{V} = \mathbf{I} \end{aligned} \tag{23}$$

which can also be solved via eigenvalue decomposition. Therefore, the solution of (16) can be calculated by alternatively perform eigen decomposition on matrices $\mathbf{B}^T\mathbf{U}\mathbf{U}^T\mathbf{B}$ and $\mathbf{G}\mathbf{G}^T + \lambda\mathbf{B}\mathbf{V}\mathbf{V}^T\mathbf{B}^T$. After that, we get the cluster centers of primitive events from $\mathbf{S}$. Given a new sample, we calculate the probability of the sample belonging to a cluster by its distance from the cluster center. The probability will be used in Sect. 5.

4 Compound event structure discovery

The primitive event label sequences, generated from the annotation or the output labels from the previous section, are low level semantical interpretations of the video content. In this section, we model compound event on the basis of the label sequence for higher level semantics. We define compound events as encodings of structured patterns of primitive events. For example, 'shopping' is a compound event, which is usually constituted by the primitive event sequence of 'browse', 'take a product', 'pay' and 'exit'. The main issue is how to discover these structures from the label sequences.

In our case, we regard the compound events as primitive event patterns which satisfy the following requirements and constraints:

(1) Frequent. The pattern with higher occurrence frequency is more likely to be perceived as a whole on a higher level.

(2) Ordered. As compound events are time evolving, the pattern that consists of same primitive events with different order usually expresses different meanings.

(3) Correlated. The primitive events in a pattern should be correlated. Note that correlation is not proportional to co-occurring frequency.

These patterns can be discovered by the following three steps:

Step 1: Candidates selection. Assuming there are five primitive event classes respectively labeled by a, b, c, d, e. The candidate set is represented in a tree structure shown in Fig. 3. Note that the nodes of the tree need not to be traversed. The candidate set prune criterion is exploited in Step 2.

Step 2: Frequency counting. In order to incorporate intervention, we set a sliding window W with width w, within which a pattern count increase if all its components (primitive events) occur in W orderly. The window W slides along the label sequence with one label passed each time. The occurring frequency of a pattern Q is defined as the total number of windows where Q occurs, which is termed as $f(Q)$. If $f(Q) > \xi$, then Q is regarded as a frequent pattern, where ξ is the frequency threshold. As the search space is exponential to the number of label sequences, a pruning criterion is required. It has been demonstrated that if Q is a frequent pattern, all subsequence of Q are also frequent patterns. Reversely, if Q is not a frequent pattern, all patterns containing Q are not frequent patterns (Mannila et al. 1997). Therefore, the tree structure of patterns can be safely pruned as in Fig. 4.

Step 3: Correlation evaluation. The purpose of Correlation is to evaluate the dependencies of primitive events in a pattern. According to the pruning criterion, given a pattern, we only need to evaluate the correlation between the last component and the residual subsequent, and the complete correlation among components and subsequences can be evaluated recursively. Given a frequent pattern Q, we denote the last

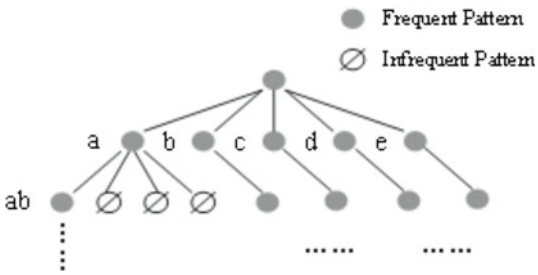

Fig. 4 Tree structured candidate set. When the parent node is a frequent pattern, the branch will go on for longer frequent patter. Else, the searching process along the branch is stopped

component of Q as q_E, and the residual subsequence as $(Q - q_E)$. Then we propose a forward correlation criterion to measure their correlation as:

$$C_f(Q) = \frac{f(Q)}{f(Q - q_E)} \tag{24}$$

where C_f represent forward correlation correlation. $C_f(Q)$ indicates the possibility of q_E occurs after $(Q - q_E)$ having occurred.

According to the correlation degree, we set a threshold ρ to categorize the frequent patterns into *Strong Correlation Pattern* (SCP) and *Weak Correlation Pattern* (WCP):

$$SCP(Q) = \delta(C_f(Q) > \rho) \tag{25}$$

$$WCP(Q) = \delta(C_f(Q) < \rho) \tag{26}$$

where $\delta()$ is the delta function.

The SCPs and WCPs not only contain high-level semantics, but also encode the knowledge on the transitions of primitive events. Both of them are regarded as compound events. The SCPs also serve as rules, i.e. the temporal context constraints on primitive event sequences. The interaction between primitive events and compound events are specified in the next section. Note that the compound events contain no duration information, i.e. the Walk-Walk-Walk-Inactive-Walk-Walk is represented as Walk-Inactive-Walk.

5 Applications

Based on Sect. 3, we have achieved the primitive event models and compound event models. On the one hand, compound events are higher level structures of primitive events. Thus, they are recognized on the basis of the primitive event recognition outputs. On the other hand, compound events encode smoothing priors on primitive event sequences (i.e. prior knowledge on primitive event transitions), which reversely influence the primitive event recognition results. In this section, we consider the duality to simultaneously recognize these two kinds of events and detect anomalies (at primitive event level) under the constraint of compound events.

5.1 Alternating event recognition

The interleaving relationship between primitive and compound events dictates that the recognition processes of these two kinds of events should interact with each other. Therefore, we propose an alternating event recognition method in this section.

In Kawanaka et al. (2006), Duong et al. (2005), and Du et al. (2006), the Hierarchical Hidden Markov Model (HHMM) is used to model the interaction between simple and complex activities. The estimation of the parameters by EM method in HHMM requires a complete data set for training. However, this is infeasible in practice sue to the fact that data arrive sequentially and incrementally. Thus an online event recognition method is required. *Sequential Monte Carlo* (SMC) framework has gained great success in online motion tracking research field (Doucet et al. 2001). But the traditional SMC assumes that there are only two layers: the observation layer and latent state layer. In our case, we extend it to a three-layer structure for alternating event recognition: observation layer (Z), primitive event state layer (X), and compound event state layer (Q). The events can be iteratively recognized as follows:

Step 1: Primitive event state prediction

Given the observation sequence $Z_{1:t-1}$, the posterior probability on compound event distribution $p(Q_{t-1}|z_{1:t-1})$ has been calculated in last iteration. The primitive event X_t is predicted by smoothing priors before Z_t arrives:

$$p(X_t|Z_{1:t-1}) = \int p(X_t|Q_{t-1})p(Q_{t-1}|Z_{1:t-1})d_{Q_{t-1}}, \tag{27}$$

where Q_{t-1} is the compound event at $t-1$. We define '+' as the the concatenation operator on primitive event sequence, and $Q_t = X_t + Q_{t-1}$. Then

$$p(X_t|Q_{t-1}) = \frac{p(Q_t)}{p(Q_{t-1})} = C_f(Q_t). \tag{28}$$

Step 2: Primitive event state verification

After the Z_t is observed, the predicted distribution is verified by

$$p(X_t|Z_{1:t}) = \frac{p(Z_t|X_t)p(X_t|Z_{1:t-1})}{\int p(Z_t|X_t)p(X_t|Z_{1:t-1})d_{X_t}}. \tag{29}$$

This raise up the probability of the state emitting high likelihood on Z_t and vice versa.

Step 3: Compound event recognition

We denote $Q_t = q_{t-K+1}q_{t-K+2}\ldots q_t$, where K is the length of the compound event. Then the probability of Q_t is calculated by

$$p(Q_t|Z_{1:t}) = \prod_{i=1}^{K} p(X_{t-i+1} = q_{t-i+1}|Z_{t-i+1}). \tag{30}$$

To implement the approach for sequential and online recognition, we approximate it by SMC. Given the weighted particles for primitive events and compound events at time $t-1$: $\{X^i_{t-1}, \omega^i_{p,t-1}\}_{i=1:N}$ and $\{Q^j_{t-1}, \omega^j_{c,t-1}\}_{j=1:M}$ (note that $\omega^i_{p,t-1}$ and $\omega^j_{c,t-1}$ are all normalized to one), new primitive event particles for time t are drawn from the proposal distribution:

$$\{X^i_t\} \sim \sum_{j=1}^{M} \omega^j_{c,t-1} p(X_t | Q^j_{t-1}), \tag{31}$$

which corresponds to the Step 1. Note that as the compound events are duration-eliminated, which means that $p(X_t = q_E^{t-1}) = 0$, the self-transition is forbidden. To remedy this, we only use the Eq. 31 to sample $3N/4$ particles. The remaining $N/4$ particles are for self-transition. Then the particles are weighted by

$$\omega^i_{p,t} = \omega^i_{p,t-1} p(Z_t | X^i_t). \tag{32}$$

After achieving the verified particles, the primitive event state posterior distribution can be approximated by

$$p(X_t | Z_{1:t}) = \sum_{i=1}^{N} \omega^i_{p,t} \delta(X_t = X^i_t) \tag{33}$$

which is the implementation of Step 2. The compound event recognition is based on the posterior distribution on primitive event states. Therefore, its distribution can be directly computed by (30). Then we sample compound event particles

$$Q^i_t \sim p(Q_t | Z_{1:t}) \tag{34}$$

and go back to (31).

Through recursive particle approximation, the posterior distribution at any time can be calculated. The primitive event is recognized by MAP:

$$\max_X p(X_t = X | Z_{1:t}). \tag{35}$$

The recognition of compound events needs double check, because compound events don't necessarily occur at any time:

$$\max_{Q \in \{\widetilde{Q} | p(Q_t = \widetilde{Q}) > \lambda\}} p(Q_t = Q), \tag{36}$$

where λ is the threshold indicating the minimum confidence for a compound event occurrence.

The above framework for event recognition combine both top-down (using compound events to smooth the primitive event sequences) and bottom-up (using primitive

events to recognize the compound events) methods. The two layer of events are recognized alternatively. The experiment section demonstrate its effectiveness.

5.2 Anomaly detection

In the literature, most anomaly detection approach are designed specifically for abnormal event detection, where either the model of abnormal events are learned from scarce training data, or the normal event models are built to detect the events with large deviation from all normal models. Few researches have been done on abnormal contexts. Although Chan et al. (2004) tried to go beyond individual events and recognize anomaly in higher level by using HMM, the structure of HMM is manually specified, which requires the prior knowledge about all kinds of possible anomalies. The variability and diversity of anomalies make it impossible to predefine the structures.

In our method, it is readily to realize abnormal context detection owing to the modeling of compound events. In Sect. 4, we categorized the compound events into WCP and SCP. If Q is an SCP, q_E is strongly expected if $(Q - q_E)$ have occurred. If q_E doesn't occur as expected, then it will be treated as a rule-breaker, and an alarm of abnormal context will be emitted. Thus, the detection of abnormal contexts can be realized by (1) expectation, and (2) a measure of deviation from expectation, which can be conveniently derived from the SMC framework for event recognition. The expectation is expressed by the prediction distribution $p(X_t|Z_{1:t-1})$, and the deviation from expectation by the difference between the prediction distribution and posterior distribution $p(X_t|Z_{1:t})$. In our case, the deviation indicates the *anomaly degree* (AD) of X_t which is measured by the Kullback-Leibler divergence:

$$\begin{aligned} AD(Z_t) &= D_{KL}(p(X_t|Z_{1:t-1})||p(X_t|Z_{1:t-1}, Z_t)) \\ &= \int p(X_t|Z_{1:t-1}) \log \frac{p(X_t|Z_{1:t-1})}{p(X_t|Z_{1:t})} d_{X_t}. \end{aligned} \tag{37}$$

By substituting (27) and (29) into (37), we get

$$\begin{aligned} AD(Z_t) &\propto -\int p(X_t|Z_{1:t-1}) \log p(Z_t|X_t) d_{X_t} \\ &= -\sum_{X_t} p(X_t|Z_{1:t-1}) \log p(Z_t|X_t). \end{aligned} \tag{38}$$

There are two terms in $AD(Z_t)$, where the first term is known in SMC framework, and the second term is derived from Sect. 3:

$$AD(Z_t) \propto \sum_{X_t} p(X_t|Z_{1:t-1})(Z_t - \mu_{X_t,j})^T \Sigma_{X_t,j}(Z_t - \mu_{X_t,j}), \tag{39}$$

where $\mu_{X_t,j}$ is the primitive event cluster center, and $\Sigma_{X_t,j}$ is the covariance of the samples in the cluster.

In order to investigate the characteristics of AD with different Z_t, we calculate the one-order and second-order derivative of AD:

$$AD'(Z_t) = \sum_{X_t} p(X_t|Z_{1:t-1})(Z_t - \mu_{X_t,j})^T((\Sigma_{X_t,j})^T + \Sigma_{X_t,j}) \tag{40}$$

$$AD''(Z_t) = \sum_{X_t} p(X_t|Z_{1:t-1})((\Sigma_{X_t,j})^T + \Sigma_{X_t,j}). \tag{41}$$

We can see that the second derivative is positive semi-definite. As the AD is in a quadratic form, it has a minimum, but has no maximum boundary. Setting $AD(Z_t) = 0$, from (39) we get

$$\widetilde{Z}_t = \frac{\sum_{X_t} p(X_t|Z_{1:t-1})\mu_{X_t,j}}{\sum_{X_t} p(X_t|Z_{1:t-1})}, \tag{42}$$

which means that the center of all prototypes achieves the lowest AD value. We are interested in knowing the observations that will result in higher AD values. According to the properties of quadratic functions, The further Z_t is away from $\widetilde{Z}_t$, the higher AD value it will be associated with. We measure the distance between Z_t and $\widetilde{Z}_t$ as

$$||Z_t - \widetilde{Z}_t|| \sim \sum_{X_t} ||Z_t - \mu_{X_t,j}||p(X_t|Z_{1:t-1}). \tag{43}$$

The distance between Z_t and $\widetilde{Z}_t$ is large, when $p(X_t|Z_{1:t-1})$ (which means a strong expectation) is large and meanwhile $\pi_i||Z_t - \mu^i_{X_t}||$ is large (which means a large deviation between observation and expectation) for a given X_t.

A high AD may signal two possibilities:

(1) abnormal context

Equation 27 shows that when the evidence of Q_{t-1} is high and at the same time $Q_{t-1}+X_t$ is an SCP, then the expectation term is strong. Meanwhile, if the observation Z_t has a large deviation from the expected X_t, then AD is high. If we regard the SCP as a rule, the Z_t is like a rule breaker. As a result, the primitive event sequence is an abnormal context.

(2) abnormal event

When the observation is a never seen event, which means the Z_t has a large deviation from all expected X_t (no matter strong or weak), then the AD is high. We regard this kind of never-seen events as abnormal events.

Hereto, the two kinds of anomalies are unified into one framework. These anomalies can be detected by either setting a threshold for $AD(Z_t)$ or search the local peak along AD_{Z_t}.

6 Experiments

In this section, we demonstrate the robust performance of the whole system. Three data sets are adopted for evaluation: Weizmann, KTH and PETS2004. Weizmann is composed of ten action categories (which is shown in Fig. 6) performed by 9 people, with a total of 90 videos. The actions are performed in different speeds, orientations, and scales with a fixed camera and consistent background. KTH contains six action categories (as shown in Fig. 7) performed by 25 people, with a total of 598 video sequences. It is a more challenging dataset compared with Weizmann in that the actions are performed in different scenarios (like indoor or out door), scales and with slight camera dithering. We use the two datasets for primitive event representation and modeling evaluation. The third one, PETS2004, contains six event scenarios, including walking, browsing, collapse, leaving object, meeting, and fighting, with a total of 28 video sequences, about 26,500 frames. As the event scenarios are composed of primitive event sequences, and can be divided into normal ones and abnormal ones, we use it for compound event structure mining and anomaly detection evaluation. The sampled frames of these datasets are shown in Fig. 5.

6.1 Primitive event representation and modeling

1. Experiments on Weizmann

In order to evaluate the effectiveness of the proposed action representation and modeling method, we firstly use a leave-one-out (LOO) procedure (which is similar with Niebles et al. 2008) to measure the degree of confusion of action categories. In each iteration, we separate the video sequences of a subject from the dataset, and use the remained videos to learn the models in an unsupervised way. Then, we use the separated sequences to test the learned models. For the parameters setting, the number of pixel prototypes and pixel signatures are respectively set to 50 and 10, which is determined by gradient search. As the number of action classes in Weizmann are known a priori, we set the number of action classes to 10, which is equal to the groundtruth number.

Fig. 5 Sample frames from Weizmann, KTH and PETS2004 datasets

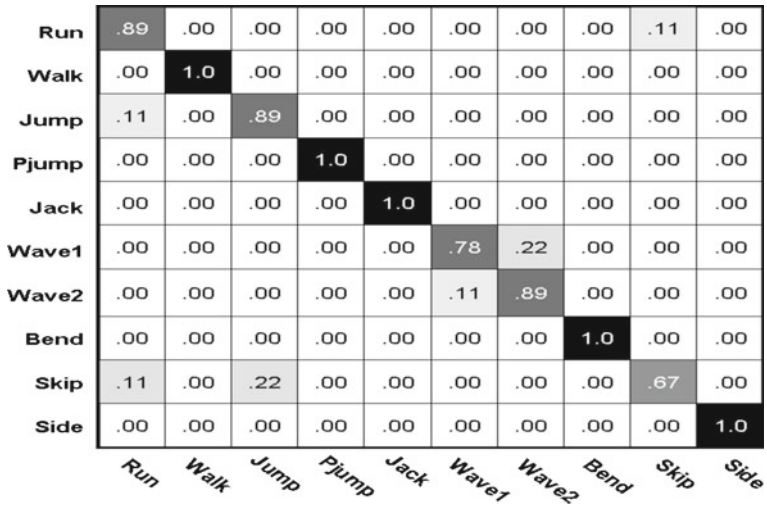

Fig. 6 Confusion matrix of action categories in Weizmann, with an average performance of 91.2%

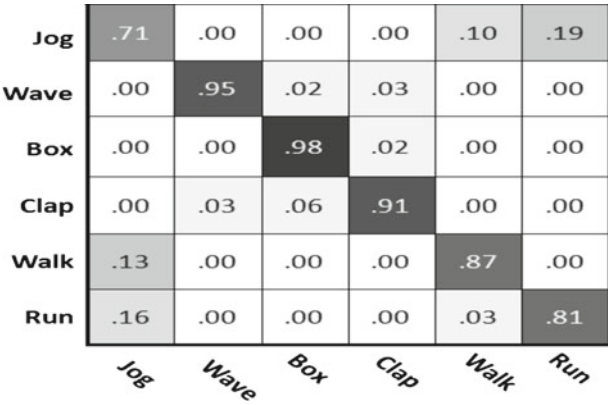

Fig. 7 Confusion matrix of action categories in KTH, with an average performance of 87.1%

We run this procedure nine times per action category, and calculate the average confusion among action categories. The result confusion matrix is shown in Fig. 6, with an average performance of 91.2%, which is higher than the result (90%) reported in Niebles et al. (2008). It shows that the confusion between Skip, Jump and Run is most obvious, which is consistent with our observation. In addition, the Wave1 and Wave2 are to some extent confused. The main reason is that the left and right arm perform similar and symmetric motion pattern, which makes the normalized features hard to discriminate in both centroid distance and frequency distance aspects. This can be solved by extending the Eq. 4 with unnormalized features like Niebles et al. (2008). Except these two cases, our method show very tiny confusion among action categories.

Table 1 Categorization accuracy comparison between JMF and Kmeans+MF

	VSN	JMF		Kmeans+MF	
		RCN	Acc	RCN	Acc
Run	10	9	0.9	9	0.9
Walk	10	10	1.0	9	0.9
Jump	9	8	0.89	8	0.89
Pjump	10	9	0.9	9	0.9
Jack	9	9	1.0	9	1.0
Wave1	9	7	0.78	7	0.78
Wave2	9	8	0.89	8	0.89
Bend	9	9	1.0	9	1.0
Skip	10	7	0.7	6	0.6
Side	9	9	1.0	8	0.89
Overall	93	86	0.92	83	0.89

Secondly, we directly categorize the whole dataset into action categories to test the overall performance. We denote the groundtruth number of videos in an action class as VSN, and the number of correctly categorized videos as RCN. Then the categorization accuracy (Acc) is calculated as RCV/VSN. The result of JMF is shown in Table 1. As we can see, 92% video sequences are correctly categorized. The Skip action is the worst categorized action, and the second to worst one is Wave1. This is consistent with the confusion of action representations, which demonstrates that except for the actions visually ambiguiate in feature space, JMF can successfully categorize all actions. In order to deal with the visually ambiguiate actions, many supervised methods have been proposed, and promising results have been achieved. But all these methods demand a large annotated dataset for training.

In order to evaluate the priority of JMF, we implement another algorithm which unidirectionally and orderly cluster prototypes and videos. We first cluster the 50 pixel prototypes into 10 pixel signatures by Kmeans, then we use the signatures to generate action matrices, and further MAT. After that, by setting $\mathbf{U}$ to an unit diagonal matrix, we perform eigen-decomposition on $\mathbf{B}^T\mathbf{U}^T\mathbf{UB}$. We term this algorithm as Kmeans + MF. Its results are shown in Table 1. We can see that ignoring the duality makes the overall performance degrade by 3 percents.

2. Experiments on KTH

We also evaluate the proposed primitive event representation and modeling methods on KTH. According to the confusion matrix shown in 7, we can see that Jog and Walk, Jog and Run, and Box and Clap are more easily to be confused, which are consistent with our observations. Actually some Jog and Run instances are even difficult to differentiate for people. Jog, Walk, and Run actions do have similar gaits and poses, but the speed is slightly different. Unfortunately, the speed information is difficult to be reflected from the pixel-wise and volume-wise (like Niebles et al. 2008) feature spaces. For Box and Clap, the main reason for confusion is that the salient parts for

Table 2 Comparison of different methods on KTH dataset with Jog and Run differentiated

Methods	Accuracy %	Learning
Our method	87.1	Unlabeled
Liu et al. (2010)	–	Unlabeled
Niebles et al. (2008)	83.33	Unlabeled

Table 3 Comparison of different methods on KTH dataset with Jog and Run fused

Methods	Accuracy %	Learning
Our method	92.9	Unlabeled
Liu et al. (2010)	91.3	Unlabeled
Niebles et al. (2008)	90.7	Unlabeled

differentiation are so tiny that they are prone to be submerged when extracting features from the whole images, which can be alleviated by feature weights assignment according to discriminability.

In order to demonstrate the advantages of the proposed methods, we compare the results with other reported ones, which is shown in Tables 2 and 3. We compare our method with two state-of-art unsupervised methods for human action categorization, among which Niebles et al. (2008) treated the Jog and Run as two categories, and Liu et al. (2010) fuse them into one category. We can see that the results of our method gain 3.7 percents improvement from Niebles et al. (2008), which used pLSA and LDA to model the extracted spatial-temporal visual words for action categorization. As Liu et al. (2010) only reported the results with Jog and Run fused, we compare the results of our method with Niebles et al. (2008) and Liu et al. (2010) in the case of fusing Jog and Run into one category. As shown in Table 3, our method compasses both state-of-art methods in accuracy. We can see that the improvement in this case is not that obvious as shown in Table 2, which implicates that our method has much better performance in discriminating visually ambiguiate actions like Jog and Run. The two main contributions for the improvement are that (1) the FFT-based dynamic pixel descriptors are more robust for high frequency noises, which is common in KTH dataset, and (2) the tensor based representation of video sequences and its matrix-based solution is more natural for 3-Dimensional video data without losing and segregating important spatial and temporal information.

6.2 Compound event discovery

In this experiment, the PETS2004 data set is used. To make sure that the modeling process are conducted on normal data, we firstly divide the PETS2004 into two subsets, where the subset1 only contains normal video sequences, and subset2 contains both normal and abnormal video sequences. As PETS2004 dataset has been annotated, we directly discover the compound event structure on the groundtruth primitive event label sequences in subset1. Some typical discovered compound events are listed in Tabel 4.

We can see that the discovered compound events are all interpretable. They are all common event patterns in real life scenarios. In the five WCPs, we can see that after the second to last event, there are more than one options. For example, after two man meet, they may shake hands or walk together. Therefore, it is discovered as WCP. In the two SCPs, the last event only have one choice. As the training data is the normal data, when a person drops a bag, he must be inactive and then pick it up. Although this event is not often (not a frequent pattern), but the correlations between the units of the pattern are strong, so that it can be discovered.

6.3 Event recognition using SMC

The PETS2004 is adopted in this experiment. We implement the SMC framework for event recognition using 200 particles for primitive events and 100 particles for compound events. As we link the state evolving probabilities with pattern correlations of compound events which have been calculated, the online and sequential recognition of primitive and compound events is able to perform efficiently.

On the testing data set, the precision rate of primitive event recognition reaches 98.2%, where only 11 out of 620 samples are wrongly recognized; the precision rate of compound event recognition reaches 94.4%, where only 2 out of 36 samples are erroneously recognized. Most of the recognition error on primitive events are because of the transition between two event classes, such as the transition from inactive to walk. The feature of these samples are ambiguous, which make the label decision process error prone. The recognition errors on compound events are caused by the error of primitive events.

In order to demonstrate the necessity of introducing the smooth prior into primitive event recognition, we use the primitive event categorization method directly for recognition on PETS2004. The same experiments are conducted, and the precision rate is degraded to 88.3%, where 72 out of 620 samples are wrongly recognized. The reason for the increased errors is that the visual ambiguities that are smoothed by high-level prior knowledge (encoded in compound events) are exposed. A typical example is shown in Fig. 8. Two 'Walk' events are erroneously recognized as 'Run' in primitive event categorization method, because the two samples have higher likelihood to 'Run'. However, as the transition probability from 'Walk' to 'Run' is rather small, the posterior probability for 'Run' is reduced down, while the second candidate 'Walk' wins. Therefore, the smooth prior which is incorporated in SMC plays an important role in smoothing state evolving process, especially when the events are visually ambiguous. Without the smooth prior, event recognition methods can hardly avoid the disturbances of noises and outliers.

6.4 Anomaly detection

We use PETS2004 for anomaly detection, where there are three classes of anomalies: One-Man-FallDown, Two-Man-Fighting, and LeftBag-without-Pickup, where the former two are abnormal events (never-seen events) and the last one is an abnormal context (rule breaker). The evaluation criterions are positive detection rate (PDR) and false alarm rate (FAR) which are defined as:

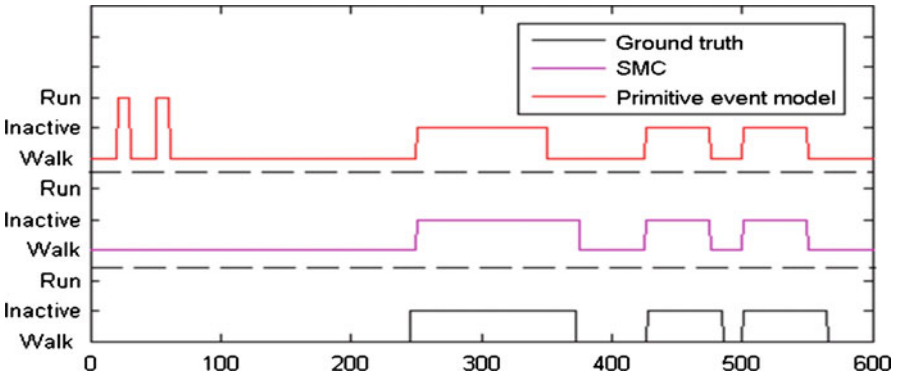

Fig. 8 Event recognition results comparison on different recognition methods

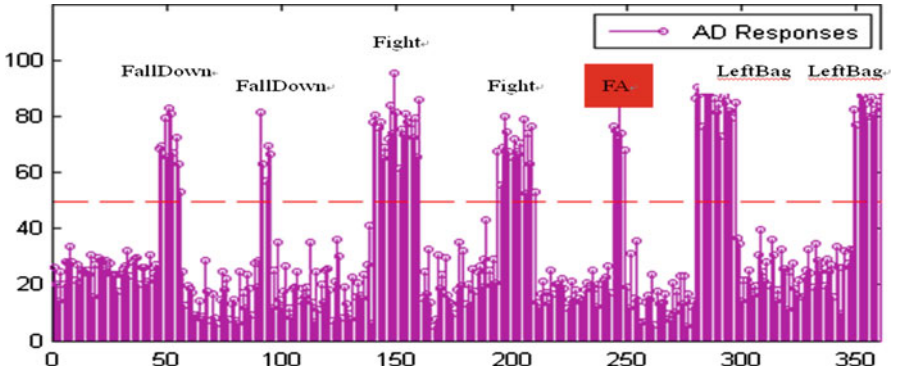

Fig. 9 The anomaly detection results. The dashed line is the threshold for alarm. The region of FA indicates false alarm

$$PDR = \frac{N_{pda}}{N_a}, \text{ and } FAR = \frac{N_{fda}}{N}, \tag{44}$$

where N, N_a, N_{pda} and N_{fda}, respectively represent number of samples, number of anomalies, number of positive detected anomalies and number of false detected anomalies. We append the testing sequences in PETS2004, and conduct the anomaly detection experiment on this sequence. The result is $PDR = 100\%$, and $FAR = 1.7\%$. The 6 false alarms out of 360 samples are caused by the visual ambiguity between Inactive and Man-Falldown, where a man stands without almost any motion. The AD responses to samples is shown in Fig. 9.

Among the detected anomalies, the LeftBag is detected as abnormal context, and the others are abnormal events. In our discovered compound events (listed in Table 4), the Walk-Inactive-LeftBag-Inactive-PickupBag is discovered as an SCP (i.e. a rule). In the test sequence, the primitive event sequence appears as Walk-Inactive-LeftBag-Inactive-Walk, which is a typical event of 'left a bag and leave'. When occurring the 'Walk' after 'Inactive', a large deviation from the strong expectation (PickupBag) triggerS a high AD response.

Table 4 Discovered compound event structures

Description	WCP	SCP
Walk-Inactive	√	
Walk-Inactive-Walk	√	
Walk-Inactive-LeftBag	√	
Walk-Meet-ShakeHands	√	
Walk-Meet-WalkTogether	√	
Walk-Inactive-LeftBag-Inactive		√
Walk-Inactive-LeftBag-Inactive-PickupBag		√

7 Conclusion

In this paper, we propose a hierarchical visual event (both primitive and compound events) pattern mining framework and its applications on recognition and anomaly (both abnormal events and abnormal contexts) detection, including a novel approach to unsupervised primitive event categorization, an extended Sequential Monte Carlo method for primitive and compound event recognition, and a unified method for abnormal events and contexts detection. The experiments demonstrate the good performance of our method.

Acknowledgment This work is supported by National Natural Science Foundation of China, No. 60933013 and No. 60833009; National Basic Research Program of China, No. 2006CB303103; China Postdoctoral Science Foundation, No. 20100470285; and Research Grants Council of the Hong Kong Special Administrative Region, China, under General Research Fund (CityU 117806, and 9041369).

References

Blank B, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. Internatinal conference on computer vision

Boiman O, Irani M (2005) Detecting irregularities in images and in video. Internatinal conference on computer vision

Chan MT, Hoogs A, Schmiederer J, Petersen M (2004) Detecting rare events in video using semantic primitives with HMM. International conference on pattern recognition

Chris D, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix tri-factorizations for clustering. International conference on knowledge discovery and data mining

Cui P, Sun LF, Liu ZQ, Yang SQ (2007) A sequential monte carlo approach to anomaly detection in tracking visual events. International conference on computer vision and pattern recognition

Davis JW, Taylor SR (2002) Analysis and recognition of walking movements. International conference on pattern recognition

Doucet A, Freitas N, Gordon N (2001) Sequential monte carlo methods in practice. Springer, New York

Du YT, Chen F, Xu WL, Li YB (2006) Recognizing interaction activities using dynamic bayesian network. International conference on pattern recognition

Duong TV, Bui HH, Phung DQ, Venkatesh S (2005) Activity recognition and abnormality Detection with the switching hidden semi-markov model. International conference on computer vision and pattern recognition

Efros AA, Berg AC, Mori G, Malik J (2003) Recognizing action at a distance. International conference on computer vision

Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. International conference on computer vision

Hakeem A, Shah M (2005) Multiple agent event detection and representation in videos. National conference on artificial intelligence
Haritaoglu I, Harwood D, Davis LS (2000) W4 real-time surveillance of people and their activities. IEEE Trans Pattern Anal Mach Intell 22(8):809–830
Ivanov YA, Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22:852–872
Kawanaka D, Okatani T, Deguchi K (2006) HHMM based recognition of human activity. IEICE transactions on information and systems, E89-D(7):2180–2185
Ke Y, Sukthankar R, Hebert M (2005) Efficient visual event detection using volumetric features. International conference on computer vision
Li T, Chris D (2006) The relationships among various nonnegative matrix factorization methods for clustering. International conference on data mining
Lin Z, Jiang ZL, Davis LS (2009) Recognizing actions by shape-motion prototype trees. International conference on computer vision
Ling HB, Soatto S (2007) Proximity distribution kernels for geometric context in category recognition. International conference on computer vision
Liu HW, Feris R, Krueger V, Sun MT (2010) Unsupervised action classification using space-time link analysis. EURASIP J Image Video Process
Makris D, Ellis T (2002) Spatial and probabilistic modeling of pedestrian behavior. British machine vision conference
Mannila H, Toivonen H, Verkamo AI (1997) Discovery of frequent episodes in event sequences. Data Min Knowl Discov 1(3):259–289
Morup M, Hansen LK, Arnfred SM (2006) Decomposing the time-frequency representation of EEG using nonnegative matrix and multi-way factorization. Technical report, Institute for Mathematical Modeling, Technical University of Denmark
Niebles JC, Wang HC, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79(3):299–318
Robertson N, Reid I (2006) A general method for human activity recognition in video. J Comput Vis Image Underst 104(2):232–248
Stauffer C, Grimson W (2000) Learning patterns of activity using real-time tracking. IEEE Trans Pattern Anal Mach Intell 22(8):747–758
Sun XH, Chen MY, Hauptmann A (2009) Action recognition via local descriptors and holistic features. International conference on computer vision and pattern recognition
Tao DC, Li XL, Wu XD, Hu WM, Maybank SJ (2007a) Supervised tensor learning. Knowl Inf Syst 13:1–42
Tao DC, Li XL, Wu XD, Maybank SJ (2007b) General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell 29:1700–1715
Tao D, Li X, Wu X, Maybank S (2008a) Tensor rank one discriminant analysis—a convergent method for discriminative multilinear subspace selection. Neurocomputing 71:1866–1882
Tao D, Song M, Li X, Shen J, Sun J, Wu X, Faloutsos C (2008b) Bayesian tensor approach for 3-D face modeling. IEEE Trans Circ Syst Video Tech 18(10):1397–1410
Vaswani N, Chowdhury A, Chellappa R (2003) Activity recognition using the dynamics of the configuration of interacting objects. International conference on computer vision and pattern recognition
Wang F, Li T (2007) Gene Selection via Matrix Factorization. International symposium on bioinformatics and bioengineering
Wang F, Li T, Zhang CS (2008) Semi-supervised clustering via matrix factorization. SIAM conference on data mining
Xiang T, Gong S (2005) Video behaviour profiling and abnormality detection without manual labelling. International conference on computer vision
Xiang T, Gong S (2006) Beyond tracking: modelling action and understanding behavior. Int J Comput Vis 67(1):21–51
Xu W, Liu X, Gong YH (2003) Document clustering based on non-negative matrix factorization. SIGIR conference on research and development in informaion retrieval
Yamamoto M, Mitomi H, Fujiwara F, Sato T (2006) Bayesian classification of task-oriented actions based on stochastic contextfree grammar. International conference on automatic face and gesture recognition
Yilmaz A, Shah M (2005) Actions sketch: a novel action representation. International conference on computer vision and pattern recognition

Zelnik-Manor L, Irani M (2001) Event-based video analysis. International conference on computer vision and pattern recognition
Zhong H, Shi J, Visontai M (2004) Detecting unusual activity in video. International conference on computer vision and pattern recognition