Human Activity Clustering for Online Anomaly Detection

Xudong Zhu
School of Computer Science and Technology, University of Xidian, Xi'an, China
Email: zhudongxu@vip.sina.com

Zhijing Liu
School of Computer Science and Technology, University of Xidian, Xi'an, China
Email: liuprofessor@163.com

Juehui Zhang
School of Computer Science and Technology, University of Xidian, Xi'an, China
Email: zhudong73@gmail.com

***Abstract*— This paper aims to address the problem of profiling human activities captured in surveillance videos for the applications of online normal human activity recognition and anomaly detection. A novel framework is developed for automatic human activity modeling and online anomaly detection without any manual labeling of the training dataset. The framework consists of the following key components: 1) A compact and effective activity representation method is developed based on a stochastic sequence of spatiotemporal actions. 2) The natural grouping of activities is discovered through a novel clustering algorithm with unsupervised model selection. 3) A runtime accumulative anomaly measure is introduced to detect abnormal activities, whereas normal human activities are recognized when sufficient visual evidence has become available based on an online Likelihood Ratio Test (LRT) method. This ensures robust and reliable anomaly detection and normal activity recognition at the shortest possible time. Experimental results demonstrate the effectiveness and robustness of our approach using noisy and sparse datasets collected from a real surveillance scenario.**

***Index Terms*—Computer Vision, Anomaly Detection, Hidden Markov Model, Latent Dirichlet Allocation**

I. Introduction

There is an increasing demand for automatic methods for analyzing an extreme number of surveillance video data produced continuously by video surveillance system. One of the key goals of deploying an intelligent video surveillance system (IVSS) is to detect abnormal activities. To achieve this objective, previously observed activities need to be analyzed and profiled, upon which a criterion on what is normal/abnormal is drawn and applied to newly captured patterns for anomaly detection. Due to the large amount of surveillance video data to be analyzed and the real-time nature of much surveillance applications, it is very desirable to have an automated system that requires little human intervention. In the paper, we aim to develop such a system that is based on fully unsupervised activity modeling and robust online anomaly detection.

Given an online Closed-Circuit Television (CCTV) input, the goal of automatic activity profiling is to learn a model that is capable of detecting unseen abnormal activities. In this context, we define an anomaly as an atypical activity that is not represented by sufficient samples in a training data set but critically satisfies the specificity constraint to an abnormal activity. The effectiveness of an activity profiling algorithm shall be measured by 1) how to measure specificity to expected patterns of activity and 2) how to maximize discrimination between normal activities.

To solve the problem, we develop a novel framework to automatically learn different classes of actions present in the data and to apply the learned model to perform anomaly detection in the new coming sequences.

1) A spatiotemporal action-based human activity representation. Due to the space-time nature of activities and their variable durations, we need to develop a compact and effective activities representation scheme and to deal with time warping. We start with the assumption that activities are sequences of discrete actions. Each action is represented as a feature vector comprising both trajectory information (position and velocity), and a set of local motion descriptors. Actions may have strong dependence on their preceding actions over multiple durations. So a temporal conjunction of such variable length action subsequences constitutes an activity. Motivated by the recent success of "bag-of-words" representations for object recognition problems in computer vision, we represent each human activity as a collection of action subsequence. This is different from most previous approaches such as [1], [2], [3] that tries to model the full dynamics of activity structures using sophisticated probabilistic models (e.g., hidden Markov models, dynamic Bayesian networks). Our activity representation aims to avoid the difficulties

doi:10.4304/jcp.6.6.1071-1079

associated with learning those models since there are usually a large number of parameters that need to be set.

2) Human activity clustering based on discovering the natural grouping of activity using Hidden Markov Model with Latent Dirichlet Allocation (HMM-LDA). A number of clustering techniques based on local word-statistics of a video have been proposed recently [4], [5], [6]. However, these approaches only capture the content of a video sequence and ignore its order. But generally activities are not fully defined by their action-content alone; however, there are preferred or typical action-orderings. This problem is addressed by the approach proposed in [5]. However, since discriminative prowess of the approach proposed in [5] is a function of the order over which action-statistics are computed, it comes at an exponential cost of computation complexity. In this work, we address these issues by proposing the usage of HMM-LDA to classify action instances of an activity into states and topics, constructing a more discriminative feature space based on the context-dependent labels, and resulting in potentially better activity-class discovery and classification.

3) Online anomaly detection using a runtime accumulative anomaly measure and normal human activity recognition using an online Likelihood Ratio Test (LRT) method. A runtime accumulative measure is introduced to determine an unseen normal or abnormal activity. The activity is then recognized as one of the normal activity classes using an online LRT method which holds the decision on recognition until sufficient visual features have become available. This is in order to overcome any ambiguity among different activity classes observed online due to insufficient visual evidence at a given time instance. By doing so, robust activity recognition and anomaly detection are ensured as soon as possible, as opposed to previous work such as [7], [8], which requires completed activity being observed. Our online LRT-based activity recognition approach is also advantageous over previous ones based on the Maximum Likelihood (ML) method [8], [9]. An ML-based approach makes a forced decision on activity recognition without considering the reliability and sufficiency of the visual evidence. Consequently, it can be error prone.

Note that our framework is fully unsupervised in that manual data labeling is avoided in both the feature extraction and the discovery of the natural grouping of activities. There are a number of motivations for performing activity clustering: First, manual labeling of activities is laborious and often rendered impractical given the vast amount of surveillance video data to be processed. More critically though, manual labeling of activities could be inconsistent and error prone. This is because a human tends to interpret activities based on the a priori cognitive knowledge of what should be present in a scene rather than solely based on what is visually detectable in the scene. This introduces a bias due to differences in experience and mental states.

The rest of the paper is structured as follows: Section 2 reviews related work to highlight the contributions of this work. Section 3 addresses the problem of activity representation. The activity clustering process is described in Section 4. Section 5 centers about the online detection of abnormal activity and recognition of normal activity. In Section 6, the effectiveness and robustness of our approach is demonstrated through experiments using noisy and sparse data sets collected from both indoor and outdoor surveillance scenarios. The paper concludes in Section 7.

II. RELATED WORK

A considerable amount of previous work has addressed the question of anomaly detection. One line of work is based on model-based anomaly recognition [10], [11]. The approach suggests a two-step solution for their detection. In the first step, on extracts image features from the video, typically achieved by detecting and tracking moving objects [12]. From tracked objects trajectory, speed, and possibly the shape descriptor of the moving objects can be computed [13]. In the second step the extracted features are used to develop models for the "normal" activity, either by hand or by applying supervised machine learning techniques [14]. A common choice is to use Hidden Markov Models [15],[16] or other graphical models [17] which quantize image features into a set of discrete states and model how states change in time. In order to detect unusual activities the video is matched against a set of normal models and segments which do not fit the models is considered unusual. This model-based approach can be quite effective in situations where "normal" activity is well-defined and constrained. However in a typical real-life video, like those used in our experiments, the number of different "normal" activity types observed can easily surpass the number of unusual types. Hence, defining and modeling what is the "normal" activity in an unconstrained environment can be more difficult than defining what is unusual. If the goal is to detect what unusual activity in a long video, the model-based approach is often over-killing.

Other line of work is based on anomalous activities' dissimilarity from regular activities [4], [5], [7]. They can be further categorized into two different types according to whether an explicit model is built. Approaches that do not model activity explicitly either perform clustering on observed patterns and label those forming small clusters as being abnormal [4], [5] or build a database of spatiotemporal patches using only regular/ normal activity and detect those patterns that cannot be composed from the database as being abnormal [7]. The approach proposed in [4] cannot be applied to any previously unseen activity patterns and therefore is only suitable for postmortem analysis but not for online anomaly detection. This problem is addressed by the approaches proposed in [5] and [7]. However, in these approaches, all the previously observed normal activity patterns must be stored either in the form of histograms of discrete events [5] or ensembles of spatiotemporal patches [7] for detecting anomaly from unseen data, which jeopardizes the scalability of these approaches.

To solve these problems, we construct an explicit generative model HMM-LDA in an unsupervised manner to learn specific activity classes for online automatic detection of abnormalities given unseen data. We also develop a more principled criterion for anomaly detection and normal activity recognition based on a runtime accumulative anomaly measure and an online LRT method originally proposed for keywords detection in speech recognition [18]. This makes our approach more robust to noise in activity representation.

Our work is similar to [4] in that activities are discovered using "video-visual word" co-occurrence matrix. However, the problem with most Cartesian space representation approaches for video like LSI or Bipartite graph co-clustering [4] is their inability to provide interpretable components. Our work is similar in spirit to [7] in that the activity model (constructed in [7] as a database of video patches) is able to infer and generalize from the training data to unseen data. However, apart from the scalability problem mentioned above, the approach in [7] has limitations in capturing the temporal ordering aspect of an activity pattern due to the constraint on the size of the video patches.

III. Action-Based ACTIVITY REPRESENTATION

First, using a standard mean shift tracking algorithm [19], we extract the following information for each target for each frame: position, velocity and a window around the target. Second, as Efros et al. [20], a local motion descriptor based on coarse optic flow is extracted from a target window I . In order to add temporal context and mitigate against confusion, we create a richer feature descriptor $(s_1^i, s_2^i, s_3^i, s_4^i)$ by concatenating the coarse motion descriptors from a number of consecutive frames, typically $T = 5$, to form a motion feature vector at each frame i, where

$$s_1^i = \sum_{t\in T} \sum_{x,y\in I} \widehat{Fb}_x^+ \quad (1)$$

$$s_2^i = \sum_{t\in T} \sum_{x,y\in I} \widehat{Fb}_x^- \quad (2)$$

$$s_3^i = \sum_{t\in T} \sum_{x,y\in I} \widehat{Fb}_y^+ \quad (3)$$

$$s_4^i = \sum_{t\in T} \sum_{x,y\in I} \widehat{Fb}_y^- \quad (4)$$

Four channels, $\widehat{Fb}_x^+$, $\widehat{Fb}_x^-$, $\widehat{Fb}_y^+$, $\widehat{Fb}_y^-$ of the motion descriptor for each frame are obtained by blurring with a Gaussian and normalizing four non-negative channels, F_x^+ , F_x^- , F_y^+ , F_y^- of optical flow vector field F . By integrating the position, velocity and motion descriptors, we define a spatiotemporal action as a target-centered action such as walking-left-to-right-on-nearside-pavement. Thus each spatiotemporal action can be represented as an eight-dimensional (8D) feature vector

$$\mathbf{f} = \left[\bar{x}, \bar{y}, v_x, v_y, s_1^i, s_2^i, s_3^i, s_4^i \right] \quad (5)$$

where $(\bar{x}, \bar{y})$ is the centroid of the target, and (v_x, v_y) is the qualitative direction.

Third, clustering is performed in the 8D spatiotemporal action feature space using a Gaussian Mixture Model (GMM). The number of spatiotemporal action classes V captured in the videos is determined by automatic model order selection based on the Bayesian Information Criterion (BIC). The learned GMM is used to classify each detected action into one of the V action classes.

Finally, the activity captured in the n th video $\mathbf{v}_n$ is represented as a feature vector $\mathbf{w}_n$, given as

$$w_n = (w_{n1}, ..., w_{nt}, ..., w_{nT_n}) \quad (6)$$

where T_n is the length of the n th video segment, and the t th element of $\mathbf{w}_n$ is a V -dimensional unit-basis vectors that have a single component equal to one and all other components equal to zero. w_{nt} corresponds to the t th image frames of $\mathbf{v}_n$, and $w_{nt}^k = 1$ if an action of the k th action class is detected in the frame, given the learned GMM; otherwise, $w_{nt}^k = 0$.

IV. ACTIVITY CLUSTERING

The activity clustering problem can now be defined formally. Consider a training data set $\mathbf{D}$ consisting of N feature vectors

$$\mathbf{D} = \{\mathbf{w}_1, ..., \mathbf{w}_n, ..., \mathbf{w}_N\} \quad (7)$$

where $\mathbf{w}_n$ is defined in (6), represents the activity captured by the n th video $\mathbf{v}_n$. The problem to be addressed is to discover the natural grouping of the training activities upon which a model for normal activity can be built. This is essentially a data clustering problem with the number of clusters unknown. There are a number of aspects that make this problem challenging: 1) Each feature vector $\mathbf{w}_n$ can be of different lengths. Conventional clustering approaches require that each data sample is represented as a fixed length feature vector. 2) Model selection needs to be performed to determine the number of cluster. To overcome the above mentioned difficulties, we propose a clustering algorithm with feature and model selection based on modeling each activity using HMM-LDA.

A. Hidden Markov Model with Latent Dirichlet Allocation (HMM-LDA)

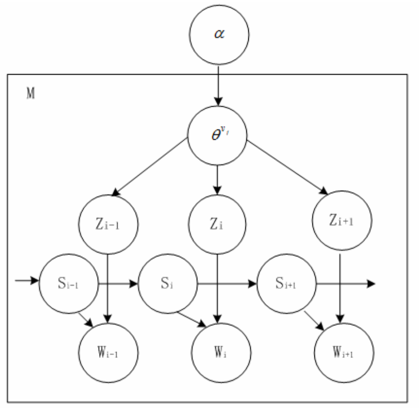


Figure 1. Graphical representation of HMM-LDA model.

Suppose we are given a collection of M video sequences $D=\{\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_M\}$ containing action words from a vocabulary of size V $(i=1,...,V)$. Each video $\mathbf{w}_j$ is represented as a sequence of N_j action words $\mathbf{w}_j=(w_1,w_2,...,w_{N_j})$, where w_i is the action word representing the i-th frame. Then the process that generates each video $\mathbf{w}_j$ in the corpus D is:

1. Draw topic weights $\theta^{(\mathbf{w}_j)}$ from $Dir(\alpha)$
2. For each word w_i in video $\mathbf{w}_j$
   a) Draw z_i from $\theta^{(\mathbf{w}_j)}$
   b) Draw c_i from $\pi^{(c_i-1)}$
   c) If $c_i=1$, then draw w_i from $\phi^{(z_i)}$, else draw w_i from $\phi^{(c_i)}$

Here we fixed the number of latent topic K to be equal to the number of activity categories to be learnt. Also, α is the parameter of a K-dimensional Dirichlet distribution, which generates the multinomial distribution $\theta^{(\mathbf{w}_j)}$ that determines how the activity categories (latent topics) are mixed in the current video $\mathbf{w}_j$. Each spatial-temporal action word w_i in video $\mathbf{w}_j$ is mapped to a hidden state s_i. Each hidden state s_i generates action words w_i according to a unigram distribution $\phi^{(c_i)}$ except the special latent topic state z_i, where the z_i th topic is associated with a distribution words $\phi^{(z_i)}$. $\phi^{(z_i)}$ corresponds to the probability $p(w_i \mid z_k)$. Each video $\mathbf{w}_j$ has a distribution over topic $\theta^{(\mathbf{w}_j)}$ and transitions between classes c_{i-1} and c_i follow a distribution $\pi^{(s_i-1)}$. The complete probability model is

$$\theta \sim Dirichlet(\alpha) \quad (8)$$

$$\phi^{(z)} \sim Dirichlet(\beta) \quad (9)$$

$$\pi \sim Dirichlet(\gamma) \quad (10)$$

$$\phi^{(c)} \sim Dirichlet(\delta) \quad (11)$$

Here, α, β, γ, and δ are hyperparameters, specifying the nature of the priors on θ, $\phi^{(z)}$, π and $\phi^{(c)}$.

B. Learning the Activity Models

Our strategy for learning topics differs from previous approaches [21] in not explicitly representing θ, $\phi^{(z)}$, π and $\phi^{(c)}$ as parameters to be estimated, but instead considering the posterior distribution over the assignments of words to topics, $p(\mathbf{z} \mid \mathbf{c},\mathbf{w})$. We then obtain estimates of θ, $\phi^{(z)}$, π and $\phi^{(c)}$ by examining this posterior distribution. Computing $p(\mathbf{z} \mid \mathbf{c},\mathbf{w})$ involves evaluating a probability distribution on a large discrete state space. We evaluate $p(\mathbf{z} \mid \mathbf{c},\mathbf{w})$ by using a Monte Carlo procedure, resulting in an algorithm that is easy to implement, requires little memory, and is competitive in speed and performance with existing algorithms.

In Markov chain Monte Carlo, a Markov chain is constructed to converge to the target distribution, and samples are then taken from Markov chain. Each state of the chain is an assignment of values to the variable being sampled and transitions between states follow a simple rule. We use Gibbs sampling where the next state is reached by sequentially sampling all variable from their distribution when conditioned on the current values of all other variables and the data. To apply this algorithm we need two full conditional distributions, $p(z_i \mid \mathbf{z}_{-i},\mathbf{c},\mathbf{w})$ and $p(c_i \mid \mathbf{c}_{-i},\mathbf{z},\mathbf{w})$. These distributions can be obtained by using the conjugacy of the Dirichlet and multinomial distributions to integrate out the parameters θ and ϕ, yielding

$$p(z_i \mid z_{-i},c,w) \propto \begin{cases} n_{z_i}^{w_j}+\alpha & c_i \neq 1 \\ (n_{z_i}^{w_j}+\alpha)\dfrac{n_{w_i}^{(z_i)}+\beta}{n^{(z_i)}+W\beta} & c_i=1 \end{cases} \quad (12)$$

where $n_{z_i}^{(\mathbf{w}_j)}$ is the number of words in video $\mathbf{w}_j$ assigned to topic z_i, $n_{w_i}^{(z_i)}$ is the number of words assigned to topic z_i that are the same as w_i, and all counts include only words for which $c_i=1$ and exclude case i.

$$p(c_i \mid \mathbf{c}_{-i}) = \frac{(n_{c_i}^{(c_{i-1})} + \gamma)}{(n_{.}^{(c_i)} + I(c_{i-1} = c_i) + C_\gamma)} \times \frac{(n_{c_{i+1}}^{(c_i)} + I(c_{i-1} = c_i) I(c_i = c_{i+1}) + \gamma)}{(n_{.}^{(c_i)} + I(c_{i-1} = c_i) + C_\gamma)} \tag{13}$$

$$p(c_i \mid \mathbf{c}_{-i}, \mathbf{z}, \mathbf{w}) \propto \begin{cases} \dfrac{n_{w_i}^{(c_i)} + \delta}{n^{(c_i)} + W\delta} p(c_i \mid \mathbf{c}_{-i}) & c_i \neq 1 \\ \dfrac{n_{w_i}^{(z_i)} + \beta}{n^{(z_i)} + W\beta} p(c_i \mid \mathbf{c}_{-i}) & c_i = 1 \end{cases} \tag{14}$$

where $n_{w_i}^{z_i}$ is as before, $n_{w_i}^{c_i}$ is the number of words assigned to class c_i that are the same as w_i, excluding case i, and $n_{c_i}^{(c_i)}$ is the number of transitions from class c_{i-1} to class c_i, and all counts of transitions exclude transitions both to and from c_i. $I(\cdot)$ is an indicator function, taking the value 1 when its argument is true, and 0 otherwise. Increasing the order of the HMM introduces additional terms into $p(c_i \mid \mathbf{c}_i)$, but does not otherwise affect sampling.

The z_i variables are initialized to values in $\{1, 2, ..., K\}$, determining the initial state of the Markov chain. We do this with an online version of the Gibbs samples, using Eq.12 to assign words to topics, but with counts that are computed from the subset of the words seen so far rather than the full data. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Eq.12. Because the only information needed to apply Eq.12 is the number of times a word is assigned to a topic and the number of times a topic occurs in a document, the algorithm can be run with minimal memory requirements by caching the sparse set of nonzero counts and updating them whenever a word is reassigned. After enough iteration for the chain to approach the target distribution, the current values of the z_i variables are recorded. Subsequent samples are taken after an appropriate lag to ensure that their autocorrelation is low.

With a set of samples from the posterior distribution $p(\mathbf{z} \mid \mathbf{c}, \mathbf{w})$, statistics that are independent of the content of individual topics can be computed by integrating across the full set of samples. For any single sample we can estimate θ, $\phi^{(z)}$, π and $\phi^{(c)}$ from the value $\mathbf{z}$ by

$$\hat{\phi}^{(z)} = \frac{n_{w_i}^{(z_i)} + \beta}{n^{(z_i)} + W\beta} \tag{15}$$

$$\hat{\phi}^{(c)} = \frac{n_{w_i}^{(c_i)} + \delta}{n^{(c_i)} + W\delta} \tag{16}$$

$$\theta = n_{z_i}^{w_j} + \alpha \tag{17}$$

$$\pi = \frac{(n_{c_i}^{(c_{i-1})} + \gamma)(n_{c_{i+1}}^{(c_i)} + I(c_{i-1} = c_i) I(c_i + c_{i+1}) + \gamma)}{n^{(c_i)} + I(c_{i-1} = c_i) + C_\gamma} \tag{18}$$

C. Model Selection

Given values of α, β and γ, the problem of choosing the appropriate value for K is a problem of model selection, which we address by using a standard method from Bayesian statistics. For a Bayesian statistician faced with a choice between a set of statistical models, the natural response is to compute the posterior probability of the set of models given the observed data. The key constituent of this posterior probability will be the likelihood of the data given the model, integrating over all parameters in the model. In our case, the data are the words in the corpus, $\mathbf{w}$, and the model is specified by the number of topics, K, so we wish to compute the likelihood $p(\mathbf{w} \mid K)$. The complication is that this requires summing over all possible assignments of words to topics $\mathbf{z}$. However, we can approximate $p(\mathbf{w} \mid K)$ by taking the harmonic mean of a set of values of $p(\mathbf{w} \mid \mathbf{z}, K)$ when $\mathbf{z}$ is sampled from the posterior $p(\mathbf{z} \mid \mathbf{c}, \mathbf{w}, K)$. Our Gibbs sampling algorithm provides such samples, and the value of $p(\mathbf{w} \mid \mathbf{z}, K)$ can be computed. Dates of manuscript submission, revision and acceptance should be included in the first page footnote.

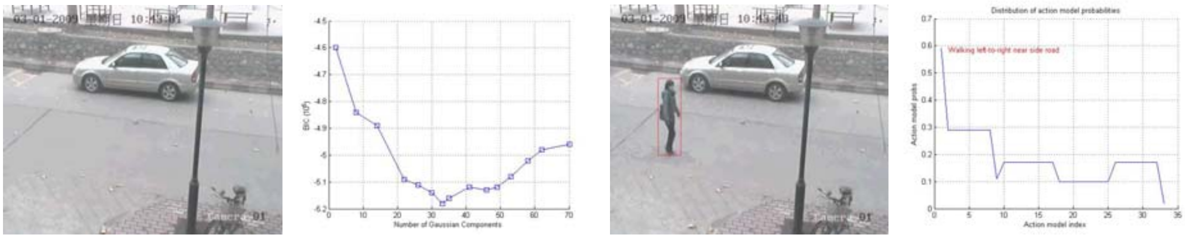
V. Online Anomaly Detection and Normal Activity Recognition

Given a unseen activity $\mathbf{P}$, we calculate the likelihood of $l(P; \alpha, \beta, \delta) = p(w \mid \alpha, \beta, \delta)$. The likelihood can be used to detect whether an unseen activity is normal using a runtime anomaly measure. If it is detected to be normal, the activity is then recognized as one of the K classes of normal activities using an online LRT method.

An unseen activity $\mathbf{P}$ including T clips is represented as $\mathbf{P} = [\mathbf{p}_1, ..., p_t, ..., p_T]$. At the t th clip, the accumulated visual information for the activity, represented as $\mathbf{w}_t$, is used for online reliable anomaly detection. First, the normalized likelihood of observing $\mathbf{w}_t$ at the t th clip is computed as

$$l_t = p(w_t \mid \alpha, \beta, \delta) \tag{19}$$

l_t can be easily computed online using the Gibbs sampling algorithm.



(a) (b) (c) (d)

Figure 2. Action-base activity representation. (a) show that 33 classes of actions are detected automatically using BIC. One typical action is highlighted in the image frame using bounding boxes in red color in (b). (c) show that the learned GMM is used to classify each detected action into one of the 33 action classes.

We measure the anomaly of $\mathbf{p}_{t+1}$ using an online anomaly

$$Q_t = \begin{cases} l_1 & if\ t = 1 \\ (1-\alpha)Q_{t-1} + \alpha(l_t - l_{t-1}) & otherwise \end{cases} \tag{20}$$

where α is an accumulating factor determining how important the visual information extracted from the current clip is for anomaly detection. We have $0 < \alpha \leq 1$. Compared to l_t as an indicator of normality/anomaly, Q_t could add more weight to more recent observations. Anomaly is detected at clip t if

$$Q_t < Th_A \tag{21}$$

where Th_A is the anomaly detection threshold. The value of Th_A should be set according to the detection and false alarm rates required by each particular surveillance application.

At each clip t, an activity needs to be recognized as one of the K activity categories when it is detected as being normal, that is, $Q_t > Th_A$. This is achieved by using an online LRT method. More specifically, we consider a hypotheses test between the following

H_k : $\mathbf{w}_t$ is from the hypothesized model z_k and belongs to k th normal activity category;

H_0: $\mathbf{w}_t$ is from a model other than z_k and does not belong to the k th normal activity category;

where H_0 is called the alternative hypothesis. Using LRT, we compute the likelihood ratio of the two hypotheses as

$$r_k = \frac{p(w_t; H_0)}{p(w_t; H_k)} \tag{22}$$

The hypothesis H_k can be represented by the model z_k , which has been learned in the activity-clustering step. The key to LRT is thus to construct the alternative model that represents H_0. In a general case, the number of possible alternatives is unlimited; $p(w_t; H_0)$ can thus only be computed through approximation. Fortunately, in our case, we have determined at the t th clip that $\mathbf{w}_t$ is normal and can only be generated by one of the K normal activity categories. Therefore, it is reasonable to construct the alternative model as a mixture of the remaining of $K-1$ normal activity categories. In particular, (13) is rewritten as

$$r_k = \frac{\sum_{i \neq k} p(z_i) p(\mathbf{w}_t \mid z_i)}{p(\mathbf{w}_t \mid z_k)} = \frac{\sum_{i \neq k} \frac{N_i}{N - N_k} p(\mathbf{w}_t \mid z_i)}{p(\mathbf{w}_t \mid z_k)} \tag{23}$$

where r_k is a function of t and computed over time, N is the total number of training activities and N_k is the number of patterns that belong to the k th activity category.

$\mathbf{w}_t$ is reliably recognized as the k th activity category only when $\chi^2 = -2\ln(r_k)$ is smaller than Th_r , where Th_r can be determined by the confidence level α as $Th_r = \chi^2_{1-\alpha}$. In our experiments, α is set to 0.05 . When there are more than one r_k lesser than Th_r , the activity is recognized as the class with the smallest r_k .

VI. EXPERIMENTS

In this section, we illustrate the effectiveness and robustness of our approach on activity clustering and online anomaly detection with experiments using data sets collected from the entrance/exit area of an office building.

A. Dataset and feature extraction

A CCTV camera was mounted on a on-street utility pole, monitoring the people entering and leaving the building (see Fig.2). Daily activities from 9a.m. to 5p.m. for 5 days were recorded. Typical activities occurring in

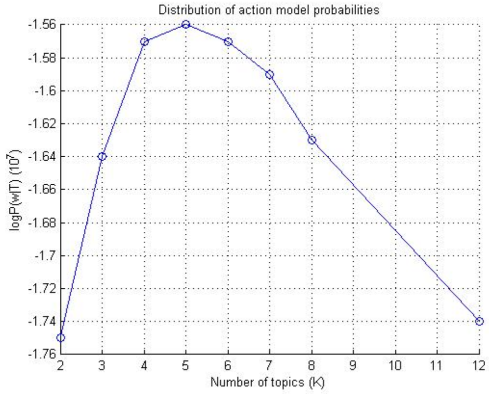
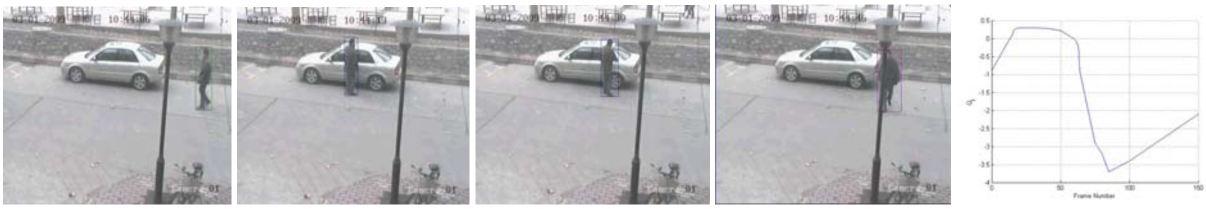


Figure 3. Model selection results

TABLE I.
THE FIVE CLASSES OF ACTIVITIES THAT MOST COMMONLY OCCURRED IN THE ENTRANCE/EXIT AREA OF AN OFFICE BUILDING

C1	Going into the office building.
C2	Leaving the office building.
C3	Passing by the office building.
C4	Getting off a car and entering the office building.
C5	leaving the office building and getting on a car.



35 62 70 90

(a) (b)

Figure 4. Example of anomaly detection in the entrance/exit area of an office building. (a) An abnormal activity where one people attempted to destroy the car parking the area. It resembles C3 in the early stage. (b) The activity was detected as an anomaly from Frame 62 till the end based on Q_t.

the scene would be people entering, leaving and passing by the building. Each activity would normally last a few seconds. For this experiment, a data set was collected from 5 different days consisting of 40 hours of video, totaling to 2880,000 frames.

To calculate the low-level feature descriptor, we first track and stabilize the persons in the video sequences using the algorithm in [19]. Then discrete actions were detected and classified using automatic model order selection in clustering, resulting in 33 classes of actions corresponding to the common constituents of all activity in this scene. By the action vocabulary of 33 actions, 947 instances of activities are collected from video data set. A training set consisting of 568 instances was randomly selected from the overall 947 instances without any activity class labeling. The remaining 379 instances were used for testing the trained model later.

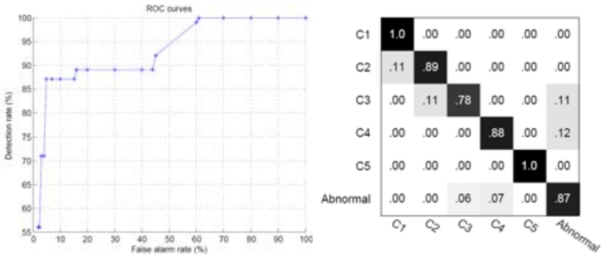
B. Activity Clustering

To evaluate the number of clusters K, we used the Gibbs sampling algorithm to obtain samples from the posterior distribution over $\mathbf{z}$ for K values of 3, 4, 5, 6, 7, 8, and 12. For all runs of the algorithm, we used $\alpha = \frac{50}{T}$, $\beta = 0.01$ and $\gamma = 0.1$, keeping constant the sum of the Dirichlet hyper-parameters, which can be interpreted as the number of virtual samples contribution to the smoothing of θ. We computed an estimate of $p(\mathbf{w} \mid K)$ for each value of K. For all values of K, we ran 7 Markov chains, discarding the first 1,000 iterations, and then took 10 samples from each chain at a lag of 100 iterations. In all cases, the log-likelihood values stabilized within a few hundred iterations. Estimates of $p(\mathbf{w} \mid K)$ were computed based on the full set of samples for each value of K and are shown in Fig.3.

The results suggest that the data are best accounted for by a model incorporating 5 topics. $p(\mathbf{w} \mid K)$ initially increases as function of K, reaches a peak at $K = 5$, and then decreases thereafter. By observation, each discovered data cluster mainly contained samples corresponding to one of five activity classes listed in Table I.

C. Anomaly detection

The activity model built using both labeled and unlabeled activities were used to perform online anomaly detection. To measure the performance of the learned models on anomaly detection, each activity in the testing sets was manually labeled as normal if there were similar activities in the corresponding training sets and abnormal otherwise. A testing pattern was detected as being abnormal when (18) was satisfied. The accumulating factor α for computing Q_t was set to 0.1. Fig.4. demonstrates one example of anomaly detection in the entrance/exit area of an office building. We measure the performance of anomaly detection using the anomaly detection rate, and the false alarm rate. The detection rate and false alarm rate of anomaly detection are shown in the form of a Receiver Operating Characteristic (ROC) curve by varying the anomaly detection threshold Th_A, as Fig.5.



	C_1	C_2	C_3	C_4	C_5	Abnormal
C1	1.0	.00	.00	.00	.00	.00
C2	.11	.89	.00	.00	.00	.00
C3	.00	.11	.78	.00	.00	.11
C4	.00	.00	.00	.88	.00	.12
C5	.00	.00	.00	.00	1.0	.00
Abnormal	.00	.00	.06	.07	.00	.87

(a) **(b)**

Figure 5. (a) the mean ROC curves for our dataset. (b)confusion matrix for our dataset; rows are ground truth, and columns are model results.

D. Normal Activity Recognition

To measure the recognition rate, the normal activities in the testing sets were manually labeled into different activity classes. A normal activity was recognized correctly if it was detected as normal and classified into a activity class containing similar activities in the corresponding training set by the learned activity model. Fig. 6(b) shows that when a normal activity was not recognized model trained using unlabeled data, it was most likely to be recognized as belonging to another normal activity class. On the other hand, for a model trained by labeled data, a normal activity was most likely to be wrongly detected as an anomaly if it was not recognized correctly. This contributed to the higher false alarm rate for the model trained by labeled data.

E. Result Analysis and Discussion

To compare our approach with six other methods, we use exactly the same experiment setup and list the comparison results in Table II. Each of these is a anomalous activity detection algorithm that is capable of dealing with low resolution and noisy data. We implement the algorithms of Xiang *et al.* [3], Wang *et al.* [6], Niebles *et al.* [22], Boiman *et al.* [7], Hamid *et al.* [5] and Zhong *et al.* [4]. The key findings of our comparison are summarized and discussed as follows:

1) Table II shows that the precision of our HMM-LDA is superior to the HMM method [3], the LDA method [6], the MAP-based method [7] and two co-clustering algorithms [4], [5]. HMM [3] outperforms the LDA [6] on our scenario, but HMM [3] require explicit modeling of anomalous activities structure with minimal supervision. Some recent methods ([4] using Latent Semantic Analysis, [22] using probabilistic Latent Semantic Analysis, [6] using Latent Dirichlet Allocation, [5] using n -grams) extract activity structure simply by computing local action-statistics, but are limited by their ability to capture activity structure only up to some fixed temporal resolution. Our HMM-LDA provided the best account, being able to efficiently extract the variable length action-subsequence of activity, constructing a more discriminative feature space, and resulting in potentially better activity-class discovery and classification.

2) Work done in [4] clusters activities into its constituent sub-class, labeling the clusters with low internal cohesiveness as anomalous cluster. This makes it infeasible for online anomaly detection. The anomaly detection method proposed in [5] was claimed to be online. Nevertheless, in [5], anomaly detection is performed only when the complete activity pattern is observed. In order to overcome any ambiguity among different activity classes observed online due to different visual evidence at a given time instance, our online LRT method holds the decision on recognition until sufficient visual features have become available.

TABLE II.
COMPARISON OF DIFFERENT METHODS

methods	**Anomaly Detection Rate (%)**
Our method	90.62
Xiang *et al.* [3]	85.76
Wang *et al.* [6]	84.46
Niebles *et al.* [22]	83.50
Boiman *et al.* [7]	83.32
Hamid *et al.* [5]	88.48
Zhong *et al.* [4]	85.56

VII. CONCLUSIONS

In conclusion, we have proposed a novel framework for robust online activity recognition and anomaly detection. The framework is fully unsupervised and consisted of a number of key components, namely, a activity representation based on spatial-temporal actions, a novel clustering algorithm using HMM-LDA based on action words, a runtime accumulative anomaly measure, and an online LRT-based normal activity recognition method. The effectiveness and robustness of our approach is demonstrated through experiments using data sets collected from real surveillance scenario.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No.60573139.

REFERENCES

[1] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden markov model,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992.

[2] A. F. Bobick and A. D. Wilson, “A state-based approach to the representation and recognition of gesture,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1325-1337, 1997.

[3] T. Xiang and S. Gong, “Beyond tracking: Modelling activity and understanding behaviour,” *International Journal of Computer Vision*, vol. 67, pp. 21-51, 2006.

[4] H. Zhong, J. Shi, and M. Visontai, “Detecting Unusual Activity in Video,” in IEEE Computer Society Conference

on Computer Vision and Pattern Recognition, pp. 819-826, 2004.

[5] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, “Detection and Explanation of Anomalous Activities: Representing Activities as Bags of Event n-Grams,” in IEEE computer society conference on Computer Vision and Pattern Recognition, pp. 1031-1038, 2005.

[6] Y. Wang and G. Mori, “Human Action Recognition by Semi-Latent Topic Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2009.

[7] O. Boiman and M. Irani, "Detecting irregularities in images and in video", in IEEE International Conference on Computer Vision, pp. 462-469, 2005.

[8] N. Oliver, B. Rosario and A. Pentland, "A Bayesian computer vision system for modeling human interactions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 831-843, 2000.

[9] L. Zelnik-Manor and M. Irani, "Event-based video analysis", in IEEE Conference on Computer Vision and Pattern Recognition, pp. 123-130, 2001.

[10] S. Hongeng and R. Nevatia, “Multi-agent event recognition,” in Proc. Eighth International Conference on Computer Vision, pp. 84-91, 2001.

[11] R. Russo and M. Shah, “A computer vision system for monitoring production of fast food,” in Proc. The 5th Asian Conference on Computer Vision, 2002.

[12] C. Wren, A. Azarbayejani, and T. Darrell, “Pfinder: Real-time tracking of the human body,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780-785, 1997.

[13] I. Haritaoglu, D. Harwood and L. S. Davis, “W4: Who? when? where? what? a real time system for detecting and tracking people,” in Proc. International Conference on Face and Gesture Recognition, Nara, 1998.

[14] N. Johnson and D. Hogg, “Learning the distribution of object trajectories for event recognition,” Image and Vision Computing, vol. 14, pp. 609-615, 1995.

[15] M. Brand, N. Oliver, and A. Pentland, “Coupled hidden markov models for complex action recognition,” in Proc. IEEE International Conference on Computer Vision, San Juan, 1997.

[16] G. Medioni, I. Cohen, and F. Bremond, “Event detection and analysis from video streams,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 873-889, 2001.

[17] M.R. Naphade and T.S. Huang. “A probabilistic framework for semantic indexing and retrieval in video,” in Proc. IEEE International Conference on Multimedia and Expo, pp.475-478, 2000.

[18] J. Wilpon, L. Rabiner, and C. Lee, “Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models,” IEEE TRANSACTIONS ON Acoustics, Speech, and Signal Processing, vol. 38, pp. 1870-1878, 1990.

[19] D. Comaniciu and P. Meer, “Mean Shift Analysis and Applications,” in Proceedings of the International Conference on Computer Vision, Kerkyra, pp. 1197-1203, 1999.

[20] A.A. Efros, A.C. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in IEEE International Conference on Computer Vision, pp.726-733, 2003.

[21] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[22] J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” in Proc. British machine vision conference, pp. 1249-1258, 2006.

Xudong Zhu Ph.D. candidate at school of Computer Science and Technology, Xidian University. He received his bachelor degree from Xidian University in 1996. His research interest covers video data mining.

Zhijing Liu Professor at school of Computer Science and Technology, Xidian University. He received his bachelor degree from Xidian University in 1982. His research interest covers computer vision and data mining.