Observe Locally, Infer Globally: a Space-Time MRF for Detecting Abnormal Activities with Incremental Updates

Jaechul Kim and Kristen Grauman
Department of Computer Sciences
University of Texas at Austin
{jaechul,grauman}@cs.utexas.edu

Abstract

We propose a space-time Markov Random Field (MRF) model to detect abnormal activities in video. The nodes in the MRF graph correspond to a grid of local regions in the video frames, and neighboring nodes in both space and time are associated with links. To learn normal patterns of activity at each local node, we capture the distribution of its typical optical flow with a Mixture of Probabilistic Principal Component Analyzers. For any new optical flow patterns detected in incoming video clips, we use the learned model and MRF graph to compute a maximum a posteriori estimate of the degree of normality at each local node. Further, we show how to incrementally update the current model parameters as new video observations stream in, so that the model can efficiently adapt to visual context changes over a long period of time. Experimental results on surveillance videos show that our space-time MRF model robustly detects abnormal activities both in a local and global sense: not only does it accurately localize the atomic abnormal activities in a crowded video, but at the same time it captures the global-level abnormalities caused by irregular interactions between local activities.

1. Introduction

Detecting unusual activities in video is of considerable practical interest. Algorithms able to single out abnormal events within streaming or archival videos would serve a range of applications—from monitoring surveillance feeds, or suggesting frames of interest in scientific visual data that an expert ought to analyze, to summarizing the interesting content on a day's worth of web-cam data. In any such case, automatically detecting anomalies should significantly improve the efficiency of video analysis, saving valuable human attention for only the most salient content.

Despite the problem's practical appeal, abnormality detection remains quite challenging technically, and intellectually it can even be hard to define. The foremost challenge is that "unusual" things naturally occur with unpredictable variations, making it hard to discriminate a truly abnormal event from noisy normal observations. Furthermore, the visual context in a scene tends to change over time. This implies that a model of what is normal must be incrementally updated as soon as new observations come in; a model requiring batch access to all data of interest at once would be useless in many real scenarios.

In this work, we introduce a space-time Markov Random Field (MRF) model that addresses these two primary challenges. To build a MRF graph, we divide a video into a grid of spatio-temporal local regions. Each region corresponds to a single node, and neighboring nodes are connected with links. We associate each node with continual optical flow observations, and learn atomic motion patterns via a Mixture of Probabilistic Principal Component Analyzers (MPPCA) [15]. Based on the learned patterns, we compute parameters for the MRF. Finally, by carrying out inference on the graph, we obtain probabilistic estimates of whether each node is normal or abnormal. To efficiently adapt the model as new video data streams in, we devise incremental updates for the MPPCA and associated MRF parameters. Figure 1 summarizes the approach.

The main advantages of our approach are twofold. First, we can detect abnormal activities both in a local and global context. Since we directly compute abnormality levels at each local node, we can provide "high-resolution" and localized estimates. This often aids in disambiguating abnormal activity within a crowded but otherwise normal scene—events that a global representation could easily miss. At the same time, our model accounts for space-time interactions between local activities, due to its global maximum a posteriori estimation with the MRF. This global context helps catch unusual interactions involving multiple local activities, which a purely local model may otherwise ignore. It also provides a smoothing effect that improves robustness in the face of noisy flow measurements.

Secondly, we show that the model parameters can be up-

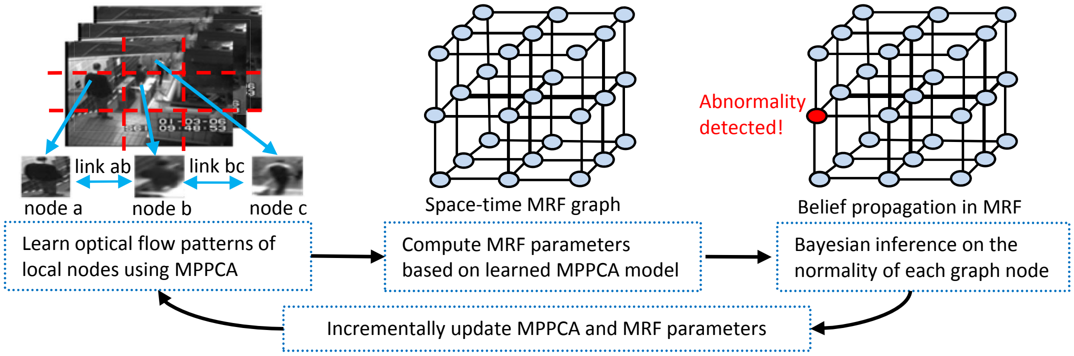

link ab
link bc
node a
node b
node c
Learn optical flow patterns of local nodes using MPPCA
Space-time MRF graph
Compute MRF parameters based on learned MPPCA model
Abnormality detected!
Belief propagation in MRF
Bayesian inference on the normality of each graph node
Incrementally update MPPCA and MRF parameters

Figure 1. Summary of our algorithm

dated incrementally whenever new observations come in. The MPPCA parameters permit a closed-form update for each mixture component, while the MRF parameters are designed to allow straightforward adjustments according to the new MPPCA values. The efficiency of the updates means that inference and revisions to the model can both be performed continuously in real-time.

We demonstrate the approach using hours of real videos collected at a subway station. When learning only from past (unlabeled) observations of the scene, our algorithm successfully detects many situations of interest for unmanned surveillance applications, such as loitering, passing through the gate without payment, dropping belongings, or getting stuck at the gate due to a malfunction.

1.1. Related work

To detect abnormal activities, most algorithms attempt to define normal activity patterns first, and then determine how much new observations deviate. Existing approaches vary in the amount of operator supervision entailed, ranging from rule-based approaches [6, 13] to unsupervised methods that directly learn normal activity patterns [19, 18, 2, 1, 16]. We take an unsupervised approach to handle abnormal activities with unpredictable variations. Broadly considered, previous unsupervised methods have explored explicit *tracking-based* methods based on typical trajectories [2, 14, 9], *activity learning* techniques based on more implicit *low-level measurements* [18, 1, 16, 11], *clustering-based* techniques [19, 8], and *indexing-based* methods that search for previously seen activity [4].

Tracking-based algorithms [2, 14, 9] determine the abnormality for each object's trajectory. While tracks directly capture an important semantic aspect of activity (where and how are people/vehicles moving), it is difficult to rely on tracks in crowded scenes with occlusions.

The approach of [18] builds a multi-observation Hidden Markov Model (HMM) and uses iterative EM for incremental updates. Similar to our approach, this work explores a graphical model and can account for where local activities typically happen in the video. However, while our MRF model captures space-time dependencies between local atomic activities, the HMM used in [18] deals with atomic activities independently due to complexity constraints. Furthermore, only a clip-level measure of abnormality is considered, whereas we localize events. In terms of evaluation, this is an advantage, in that we can more precisely say whether the detection is correct, or just a "lucky" hit due to some other noise in the clip.

Approaches using Bayesian topic models [16, 11] can also evaluate the normality of each local activity (i.e., word) while considering interactions (i.e., topic) between them. However, these methods do not impose explicit spatio-temporal dependencies between local activities, and only run in a batch mode. Clustering methods [19, 8] can automatically find outlier sequences and have shown good results, though the entire corpus is analyzed at once (in batch) to find the normal clusters.

The above methods [19, 18, 16, 11, 8] can be considered "global", in the sense that they typically attempt to find the abnormal global activity patterns in a video clip in which several local activities can co-occur. An alternative is to focus attention on individual local activities, as in [1], where typical flow directions and speeds are measured on a grid in the video frame. While efficient and simple to implement, such an approach fails to model temporal relationships between motions.

Rather than attempt to learn a model of normal variations, the method in [4] stores all previously seen spatio-temporal local patches in a database, so as to see if any configuration exists similar to a new observation. The method shows good performance in discriminating complex and detailed motions (such as a ballet performance) and makes incremental updates simple, yet it faces scalability issues once the database is very large.

Space-time MRFs have recently also been explored for some low-level video tasks, such as stereo matching [17] and image-denoising [5] with video sequences. The space-time MRF model we define is particularly well-suited for abnormality detection, as it can integrate the merits of both local and global approaches. By capturing both spatial and temporal interdependencies between local activity patterns, the proposed method translates low-level cues (motion) into a richer summary of activity. At the same time, it maintains efficiency with incremental updates fast enough to perform online with every frame.

2. Approach

Our goal is to infer when something unusual happens in a streaming video. The only data used to "train" the system is whatever initial video is captured to display the scene of interest, which is used to automatically initialize the model parameters. We extract optical flow features at each frame, use MPPCA to identify the typical patterns, and construct a space-time MRF to enable inference at each local site. For all subsequent video, we simultaneously infer abnormality levels, while incrementally updating the model.

2.1. Learning of local activity patterns

We use optical flow as a low-level measure of activity in local regions. We compute the flow with a multi-scale block-based matching between adjacent frames. Optical flows obtained at each scale are summed into a final flow vector, from which we compute a 9-dimensional optical flow vector (8 orientations + 1 speed) for every pixel. To construct a feature descriptor representing the atomic activity in each local region (corresponding to each node in Figure 1), we divide the region *L* into *u* by *v* sub-regions; each sub-region is represented by a 9-d vector obtained by summing the flow from all pixels within it. Finally, we concatenate the flow vectors of each sub-region into a 9*uv* dimensional activity descriptor for local region *L*. The number of sub-regions (i.e., *u* and *v*) is determined depending on how finely we want to capture the motion details.

After extracting descriptors for all local regions in the initial training video, we apply the Mixture of Probabilistic Principal Component Analyzers (MPPCA) algorithm to learn a generative model for local activity patterns. The dimensionality reduction offered by MPPCA gives us a compact representation of the high-dimensional descriptors. An MPPCA model is defined as follows:

$$p(t) = \sum_i \pi_i p_i(t|C_i, \mu_i), \tag{1}$$

where t is an activity descriptor, $p_i(t|C_i, \mu_i)$ is a probability density function of mixture component i, and C_i and μ_i denote the covariance matrix and mean vector of component i, respectively. The variable π_i is a mixing coefficient for component i. Expectation-Maximization (EM) is used to compute all MPPCA parameters [15].

Rather than fit one model per local region, we construct a common MPPCA over all local regions. This is due to the fact that some local regions do not have enough samples in the initial video to allow stable convergence in EM; that is, most of the observations are motion-free at some local regions. Essentially, the mixture model probabilistically encodes the "vocabulary" of low-level motions. From the learned MPPCA, we compute two histograms: a *frequency* histogram at each node, and a *co-occurrence* histogram at each link. The frequency histogram represents how often each MPPCA component is observed at each node; the co-occurrence histogram records how often two MPPCA components co-occur at neighboring nodes. Together these empirical distributions describe the typical local activities and their interactions, and are used to establish the space-time MRF (to be defined in the following section).

Let H_i denote the frequency histogram at node i, and let $H_{i,j}$ denote the co-occurrence histogram for neighboring nodes i and j, computed as follows:

$$\begin{aligned} H_i(l) &= \sum_{k=1}^{n} p(l|t_{i,k}), \\ H_{i,j}(l,m) &= \sum_{k=1}^{n} p(l|t_{i,k}) p(m|t_{j,k}), \end{aligned} \tag{2}$$

where $H_i(l)$ denotes the l^{th} bin of H_i, and $H_{i,j}(l,m)$ denotes the $(l,m)^{\text{th}}$ bin of $H_{i,j}$. The terms $p(l|t_{i,k})$ and $p(m|t_{j,k})$ are the posterior probabilities of the occurrence of MPPCA components l and m respectively, given activity descriptors $t_{i,k}$ and $t_{j,k}$ at nodes i and j at the k^{th} frame. Thus, $H_i(l)$ accumulates the posterior probability of component l over all previous activity descriptors observed at node i, thereby representing the likelihood of that low-level motion "type" occurring in that region of the video. Similarly, $H_{i,j}(l,m)$ represents the likelihood that components l and m co-occur at neighbor nodes i and j, thereby capturing the common interactions between nearby regions, whether spatially or temporally.

The posteriors are defined using Eq. (1):

$$\begin{aligned} p(l|t_{i,k}) &= \frac{\pi_l p_l(t_{i,k}|C_l, \mu_l)}{\sum_n \pi_n p_n(t_{i,k}|C_n, \mu_n)}, \\ p(m|t_{j,k}) &= \frac{\pi_m p_m(t_{j,k}|C_m, \mu_m)}{\sum_n \pi_n p_n(t_{j,k}|C_n, \mu_n)}. \end{aligned} \tag{3}$$

Having defined the distributions to capture local activity, we next show how to evaluate the normality of new observations using the learned MPPCA model and the established histograms. Then we describe our incremental learning strategy in section 2.3.

2.2. Bayesian inference on the space-time MRF

Whenever a new video frame comes in, we construct a space-time MRF in an online manner using the new frame and a fixed-length history of recently seen frames (we use 10 in our experiments). The MRF is defined in terms of two functions: the node evidence and the pair-wise potentials. We compute them both in terms of the learned MPPCA model defined above. Ultimately, inference on the graph will yield the maximum a posteriori (MAP) labeling that specifies which nodes are normal or abnormal, as computed by maximizing the following:

$$E(x) = \lambda \sum_{i} n(x_i) + \sum_{i,j \in neighbor} \rho(x_i, x_j), \qquad (4)$$

where $n(\cdot)$ is the node evidence function, and $\rho(\cdot,\cdot)$ is a pair-wise potential function. The value λ is a constant to weight the node evidence, and x_i denotes the label telling whether node i is normal or abnormal. ($x_i = 0$ signifies node i is abnormal; $x_i = 1$ means it is normal.)

The node evidence function itself consists of two terms: a *frequency* term $n_f(\cdot)$ and a *suitability* term $n_s(\cdot)$. The frequency term measures how often an activity pattern (i.e., a MPPCA component) similar to the current activity descriptor at the given node has been observed before at that node. The suitability term evaluates how likely it is that the current activity descriptor was generated by the existing MPPCA model.

The frequency term imposes a relational constraint on each node-component pair. Simply speaking, if the activity descriptor detected at node i belongs to one of the frequently observed components for node i, the value of $n_f(x_i = 1)$ becomes higher (or conversely, for a rarely observed component, it becomes lower). Complementarily, $n_f(x_i = 0) = 1 - n_f(x_i = 1)$. We compute the frequency term from each node's histogram H_i:

$$n_f(x_i = 1) = T_k\left(\sum_{c} H_i(c) p(c|t_i)\right), \qquad (5)$$

where $H_i(c)$ is a (normalized) frequency histogram for node i defined by Eq. (2), and $p(c|t_i)$ is the posterior probability of component c given activity descriptor t_i, as defined in Eq. (3). The function $T_k(\cdot)$ is a transformation function to control the degree of sensitivity to abnormalities, and is defined as:

$$T_k(x) = \begin{cases} \frac{0.5x}{k} & 0 \le x \le k, \\ 1 - \frac{0.5 \log x}{\log k} & k \le x \le 1. \end{cases} \qquad (6)$$

Lower values of the control parameter k will lead to fewer abnormal activity detections (i.e., less sensitivity to deviations from the model). This function is similar to those used for outlier rejection in robust statistics [3]. In sum, $n_f(x_i = 1)$ is the (transformed) normalized correlation between the frequency histogram H_i and the probability distribution of MPPCA components for a node i's current activity descriptor.

The suitability term reflects how well the current MPPCA model explains the new activity descriptor t. We compute it as: $n_s(x_i = 1) \propto p(t)$, where the term $p(t)$ denotes the pdf given in Eq. (1). For numeric stability we directly use the Mahalanobis distance to evaluate it. Thus, the suitability is defined as follows:

$$n_s(x_i = 0) = T_k\left(\sum_{c} d_c(t_i) p(c|t_i)\right), \qquad (7)$$

where $d_c(t_i) = F_c\left((t_i - \mu_c)^{\mathrm{T}} C_c^{-1} (t_i - \mu_c)\right)$ is the Mahalanobis distance between activity descriptor t_i and the MPPCA component c, normalized to be in $[0, 1]$. The normalization function, $F_c(\cdot)$ is the cumulative distribution of distances at the component c over all previous observations, which we implement using a cumulative histogram of the distances for all previous descriptors. $T_k(\cdot)$ is defined as above, and $n_s(x_i = 1) = 1 - n_s(x_i = 0)$.

Finally, we have the complete node evidence function:

$$n(x_i) = \begin{cases} (1-\tau) n_f(x_i) + \tau n_s(x_i) & \text{if } n_s(x_i = 0) > 0.5, \\ \tau n_f(x_i) + (1-\tau) n_s(x_i) & \text{otherwise,} \end{cases} \qquad (8)$$

where τ is a weighting constant and is always set with $\tau > 0.5$. Essentially, this serves to down-weight the frequency term should the activity descriptor at node i deviate significantly from the current MPPCA model (i.e., $n_s(x_i = 0) > 0.5$), which is important since the frequency term assumes that the observation can be explained well by the current model. Otherwise, we weight the frequency more than the suitability, as it is more discriminative in detecting abnormality as long as the activity descriptor's Mahalanobis distance is low. In short, the node evidence function measures the normality of an activity descriptor at each node, and it balances the frequency and suitability terms depending on how well the descriptor can be explained by the existing MPPCA model.

The pair-wise potential function, $\rho(\cdot,\cdot)$, consists of two terms: a *co-occurrence frequency* term $\rho_f(\cdot,\cdot)$ and a *smoothness term* $\rho_s(\cdot,\cdot)$. The co-occurrence frequency term evaluates how often we have observed two MPPCA components co-occurring at neighboring nodes i and j. If $x_i = 1$ and $x_j = 1$, then

$$\rho_f(x_i, x_j) = T_k\left(\sum_{c_i} \sum_{c_j} H_{i,j}(c_i, c_j) p(c_i|t_i) p(c_j|t_j)\right). \qquad (9)$$

Otherwise, $\rho_f(x_i, x_j) = 1 - \rho_f(x_i = 1, x_j = 1)$. This definition has a similar form to the frequency term of Eq. (5), except it uses the normalized co-occurrence histogram $H_{i,j}$ defined in Eq. (2). Here $p(c_i|t_i)$ and $p(c_j|t_j)$ denote the posterior probabilities of components c_i and c_j given activity descriptors t_i and t_j at nodes i and j, respectively.

This term will measure how normal it is for two motions to co-occur at neighboring nodes.

The smoothness term imposes smoothness on label assignments between neighboring nodes based on their motion similarity: more similar motions lead to more smoothing, which is based on the fact that similar motions at neighboring nodes are more likely to be involved in a common activity, so they have higher probability of the same labeling being assigned. We compute this term based on the normalized correlation between the two activity descriptors:

$$\rho_s(x_i, x_j) = \begin{cases} \frac{t_i \cdot t_j}{|t_i||t_j|} & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

From Eqs. (9) and (10), we can now define the complete pair-wise potential function, $\rho(\cdot, \cdot)$:

$$\rho(x_i, x_j) = \rho_f(x_i, x_j) + \alpha \rho_s(x_i, x_j), \tag{11}$$

where α is a constant to weight smoothness.

Given the MRF parameters in Eqs. (8) and (11) at every node and link of the graph, we carry out MAP inference to maximize the function defined in Eq. (4). We use loopy belief propagation with max-sum message passing, which provides the MAP labeling of whether each node is normal or abnormal. By capturing the spatial and temporal interdependency between local motion patterns, our MRF model enhances the descriptive power of what are initially purely local measurements.

2.3. Incremental updates for activity patterns

Having built an MPPCA model using a small amount of initial training video (section 2.1), we can continuously update its parameters using the new activity descriptors extracted at every video frame. All the histograms (i.e., frequency histograms, co-occurrence histograms, and cumulative histogram of Mahalanobis distances) and MRF parameters are straightforward to adjust according to the updated MPPCA parameters.

To update the MPPCA parameters given a new activity descriptor, we first pick the most likely component $c_{max} = \arg\max_c p(c|t)$ for the descriptor, and then update the covariance matrix C and mean vector μ of that component c_{max} using the algorithm given in [12]. The mixing coefficients π_i for all components are also adjusted:

$$\pi_{t+1,i} = \frac{N_{t+1,i}}{N_{t+1}}, \quad N_{t+1} = N_t + 1,$$

$$N_{t+1,i} = \begin{cases} N_{t,i} + 1 & \text{if } i = c_{max}, \\ N_{t,i} & \text{otherwise,} \end{cases}$$

in which N_t and N_{t+1} are the total numbers of activity descriptors observed until times t and $t+1$, and $N_{t,i}$ and $N_{t+1,i}$ are the total numbers of activity descriptors belonging to the component i until times t and $t+1$, respectively, and $\pi_{t+1,i}$ is the updated mixing coefficient of the component i at time $t+1$.

Our incremental algorithm is quite simple and easy to implement. However, we should note one necessary approximation that it makes: we assume that the posterior probability of each component is unchanged once the descriptor is inserted into the model. Since the MPPCA parameters change whenever a new input comes in, the posterior probabilities of all previous descriptors should also change in response. However, re-calculating all the posteriors would mean touching every previous observation, thus defeating the purpose of an incremental update. (The method given in [12] incrementally adjusts a single component, without such a backward computation.) This is a well-known issue with incremental learning and mixture models. Following [10], we assume that the posterior probability of an activity descriptor is fixed to the value computed at the time when the descriptor was first introduced to the model.

To choose the number of MPPCA components automatically, we empirically identify the minimum number of components that appear to account for most of the initial dataset. Starting with a single component, we increase the number of components until it happens that some trivial component is formed that accounts for only a very small number of activity descriptors (e.g., less than 5% of overall training data). In future implementations, model selection techniques for EM (such as [7]) could be used.

3. Experimental Results

We tested our algorithm using over two hours of surveillance videos from a subway station: one video monitors the entrance gates, and the other watches the exit gates. In both, there are typically one to 10 people moving in the scene at the same time. The videos are provided by courtesy of Adam *et al.* [1]. We discuss each one in turn below.

The frame size of the videos is 512 x 384. We divide the frames into 14 by 9 overlapping local regions of size 60 x 90 pixels each; each local region is further divided into 2 by 3 sub-regions, each of size 30 x 30 pixels. This yields a 54 (= 2 x 3 x 9) dimensional descriptor to represent activities at each local region.

For every input frame, we build a space-time MRF using the 10 most recent frames, and carry out MAP inference using belief propagation. After every MAP computation, the MPPCA parameters (see Eq. (1) and section 2.3), frequency and co-occurrence histograms (see Eq. (2)) and cumulative histograms (see section 2.2) are updated.

For all results, we use the following parameters: $\lambda = 1$ in Eq. (4), $\tau = 0.85$ in Eq. (8) and $\alpha = 0.5$ in Eq. (11). Below we evaluate the true-positive/false-positive tradeoff as a function of the control parameter k in Eq. (6), which dic-

Ground truth count	*Loitering (14/3)*	*No payment (13/-)*	*Wrong direction (26/9)*	*Irregular interaction (4/-)*	*Misc. (9/7)*	*Total (66/19)*	*False alarm*
Incremental	13/3	8/-	24/9	4/-	8/7	57/19	6/3
Batch	14/3	7/-	24/9	3/-	8/6	56/18	3/0

Table 1. Comparison of accuracy using incremental vs. batch learning for both subway videos. Numbers in parens denote count for each abnormal activity in the ground truth. The first number in the slash (/) denotes the entrance gate result; the second is for the exit gate result.

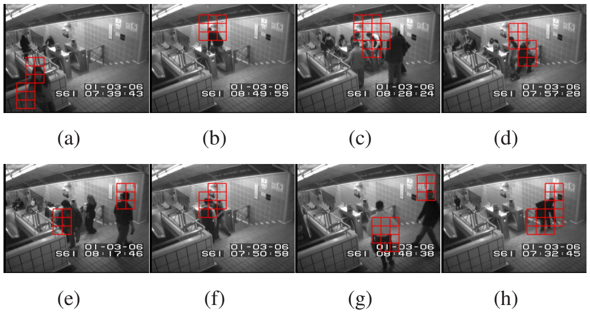


Figure 2. Example abnormal activities detected by our algorithm. Red rectangles indicate where abnormality is detected. (a) Loitering. (b)-(c) No payment. (d) Wrong direction. (e) Wrong direction and loitering: A left person is moving in the wrong direction. A person on the right is loitering as he sees the other going in the wrong direction. (f)-(g) Irregular interaction. In (f), two people are at the same gate. In (g), left person runs in hurry, and right person pauses to yield. (h) Misc.: woman drops her belonging. Both (d) and (e) illustrate detections in spite of crowded scenes and partial occlusions. Best viewed in the electronic version.

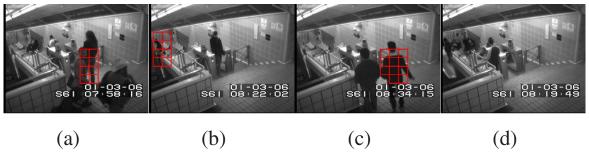


Figure 3. Examples of false alarms and abnormalities missed by our algorithm. (a)-(c) are false alarms; (d) is missed. (a) A person goes to the gate, and a person nearer to the camera is walking left to right. While this situation happens occasionally, here the nearer person slows down to talk, and the system raises an alert. (b) A person is getting on the train. Optical flows are often unreliable in the far-field areas, and it leads to a false alarm. (c) A person walks too fast. (d) Our method misses the "no payment action", perhaps because the motion is very similar to passing through the gate normally. Best viewed in the electronic version.

tates how selective we want the system to be about raising alerts for abnormalities detected.

Due to space limitations, we cannot display all of the detected abnormal activities in the paper. Interested readers may see our supplementary video which contains all of the detection results: `http://www.cs.utexas.edu/~jaechul/activity.html`.

3.1. Entrance gate

The entrance gate video is 1 hour 36 minutes long, with 144,249 total frames. We initially train the MPPCA model using the video clips containing normal activities in the first 15 minutes of video. The number of clusters in the initial MPPCA was automatically selected to be 10.

Results: In this video, the following types of abnormal activities occur: (i) Wrong direction: occasionally people exit through the entrance gate. (ii) No payment: some people sneak through or jump over the gate without tagging a payment card. (iii) Loitering: some people loiter for a long time at the station. (iv) Irregular interactions between persons: e.g. two people awkwardly zigzag to avoid each other. (v) Misc.: e.g., a person abruptly stops walking, or runs fast. We annotated the data to form a ground truth set of abnormal events, identifying a total of 66 unusual activities. We used the ground truth defined in [1] on the same data as a starting point, which marks 21 occurrences of abnormalities, primarily of the "wrong direction" event. We refined this to also capture all the more subtle abnormalities, such as "no payment" and "loitering"—in a sense, raising the bar for our algorithm, as the wrong direction events are easily detected. Admittedly, the very definition of abnormality is somewhat subjective; we took every effort to come up with the most accurate manual annotation possible.

Figure 2 shows examples of abnormal activities detected by our algorithm. Our method can identify an abnormal activity even within a crowded scene with significant occlusions (e.g. Figure 2 (d) and (e)), or cases where both normal activities and abnormal activities are mixed (e.g. Figure 2 (c) and (e)). Also, we can see that the algorithm captures the abnormality caused by irregular interactions between persons (e.g. Figure 2 (f) and (g)).

Table 1 summarizes the results of this experiment. Most errors occur when detecting the "no payment" behavior; about 40% of such actions are missed. This is largely due to poor optical flow measurements in the far-field area from the camera where the gate is located. Furthermore, some "no payment" actions can be too subtle to be recognized (see Figure 3 (d)). However, some errors are also due to our representation, which cares about the speed of motions; for example, our method issues a false alarm for slow walking in Figure 3 (a) or fast walking in Figure 3 (c), actions which deviate from existing local motions in the MPPCA model, but the ground truth says are normal.

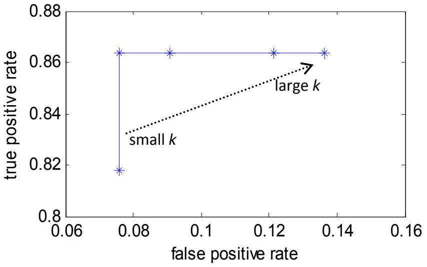
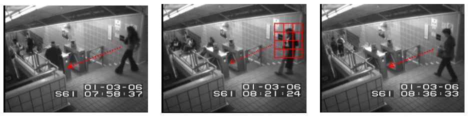


Figure 4. ROC curve with varying control parameter k.



(a) (b) (c)

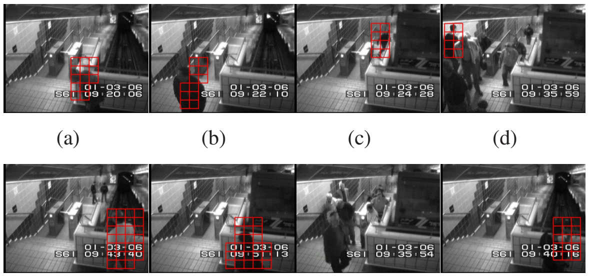
Figure 5. An effect of incremental learning. (a) Frame 41207. A movement from the right entrance to the gate (indicated by a red arrow) is detected as "normal". (b) Frame 75381. Later, the same type of movements are detected as "abnormal" because many of the most recent motions were along another path, making the motion in (a) a rarer event. (c) Frame 98104. A movement from the right entrance to the gate is again detected as "normal" because several similar observations are accumulated between (b) and (c).

The ROC curve in Figure 4 illustrates the false-positive/true-positive tradeoff, plotted as a function of k. For this data, the algorithm is quite robust to how the control parameter is set. Overall, smaller values of k will result in more selectivity in the detections (less false positives).

Incremental learning: Our method continuously updates the parameters for the MPPCA components and corresponding histograms using the new activity descriptors observed at every frame. Qualitatively, we can see how such an update influences the subsequent detections: Figure 5 shows how the same category of activity that is earlier detected as abnormal may later be detected as normal, depending on what kind of activities prevail at each observation period. Table 1 quantitatively compares our incremental algorithm's accuracy relative to a batch alternative, where all normal frames are observed at once. For the batch baseline, we trained MPPCA using all video clips containing normal activities. There are only minor differences in detection accuracy and false alarm rates, even though our incremental method has access only to a portion of the normal data and adjusts the model in real-time.

An "extreme" form of a batch model would re-compute the MPPCA model from scratch after each new frame, which is clearly not feasible computationally, since each update would require at least 10 minutes.

Run-times: We implemented our algorithm using C++ with a 2.4GHz CPU, 2Gbyte RAM machine. For the first video, the initial MPPCA training took about 10 minutes, while a batch training of MPPCA using the entire video took about 90 minutes. (The times were 3 vs. 26 minutes on the shorter second video.) Incremental parameter updates were done about at frame rate (25 frames per second). In the MAP computation, belief propagation usually converges within 10 iterations, or about 100 ms.



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

(e) (f) (g) (h)

Figure 6. Illustrative examples comparing our algorithm and baselines that are purely local or global (clustering-based). (a) - (e) Correctly detected results using our method. (f) False alarm by our method. (g) False alarm in the crowded scene by global baseline. (h) Missed abnormality by both baselines. (a) An employee cleans up the wall. (b) and (h) A person abruptly stops walking and changes his direction, which is missed by both baselines. (c) A person gets off the train and then gets on the train again very soon, which is missed by the global baseline. (d) Wrong direction: a person (left) is entering through the exit gate. Rare event: a child's movement is observed for the first time. (e) Loitering. (f) False alarm: a person is going from the right to the left, which is normal but very unusual, since most people exit the station from left to right. Best viewed in the electronic version.

3.2. Exit gate

The second video monitors an exit gate, and is 43 minutes long with 64,900 total frames. We use the first five minutes of video to train the initial MPPCA model, obtaining 8 total components.

Results: This video contains the following unusual events: (i) Wrong direction (ii) Loitering (iii) Misc.: e.g., an employee of the subway station is washing the wall. The ground truth data consists of 19 total abnormal activities. Table 1 summarizes our detection results. Figure 6 shows examples of detected abnormal activities and false alarms. We should note that all of the false alarms are raised due to the "from right exit to left exit" movement, which is "normal", but very rare compared to the "from left exit to the right exit" movement. Table 1 also compares the batch mode and incremental mode for this video. In batch mode, there are no false alarms: the batch mode can detect the "from right exit to left exit" movements as normal, since it can use all data to train.

Comparison to other types of methods: Finally, we compare against our own implementation of two existing methods: a local monitoring algorithm modeled after [1], and a "global" spectral clustering-based method modeled after [19]. Our goal is to demonstrate the relative strengths and weaknesses of local and global representations, and to see how well our MRF-based strategy can combine the strengths of both. Our implementation is necessarily a simplified version of the representative baselines.

Both approaches provide similar results in detecting conspicuous abnormalities such as "wrong direction", where the optical flow patterns are quite easy to discriminate. However, the local method fails to detect abnormal activities with irregular temporal orderings. In Figure 6 (b) and (h), people abruptly stop walking and turn around. Since a local method can only consider frame-by-frame individual actions, these kinds of actions cannot be detected. Also, the local method is sensitive to optical flow parameters, often resulting in a high false alarm rate, about an order of magnitude more false alarms than our method. Global MAP inference with a smoothness constraint in our MRF model helps reduce the number of false alarms caused by noisy local observations.

On the other hand, the global method fails to detect abnormal activity happening at a fine local scale. For example, in Figure 6 (c), the abnormal activity happens within a region so small that it is simply regarded as negligible noise in a global sense. At the other extreme, the global method generates false alarms in crowded scenes (Figure 6 (g)), where measurement noise accumulated over many normal activities is above the threshold. In contrast, our approach succeeds in these scenarios since we can localize individual abnormal activities at each node while taking into account spatio-temporal dependencies between them.

4. Conclusion

We proposed a space-time MRF for detecting abnormal activities that combines the advantages of both local and global approaches. Not only can the method localize abnormal activities even in crowded scenes, but it can also capture irregular interactions between local activities in a global sense. In addition, we demonstrated incremental real-time updates, which allow our algorithm to adapt to visual context changes over a long period of time. Experimental results on long surveillance videos show that our algorithm can work robustly in practical applications.

Acknowledgement

This research was supported in part by NSF CAREER 0747356, Microsoft Research, Texas Higher Education Coordinating Board award # 003658-01-40-2007, the DARPA VIRAT program, NSF EIA-0303609, and the Henry Luce Foundation. Jaechul Kim was also supported in part by a Fellowship from the ILJU Foundation, Korea.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors. *PAMI*, 30:555–560, Mar. 2008.

[2] A. Basharat, A. Gritai, and M. Shah. Learning Object Motion Patterns for Anomaly Detection and Improved Object Detection. In *CVPR*, 2008.

[3] M. J. Black and A. Rangarajan. On the Unification of Line Processes, Outlier Rejection, and Robust Statistics with Applications in Early Vision. *IJCV*, 19:57–91, July 1996.

[4] O. Boiman and M. Irani. Detecting Irregularities in Images and in Video. *IJCV*, 74:17–31, Aug. 2007.

[5] J. Chen and C. K. Tang. Spatio-Temporal Markov Random Field for Video Denoising. In *CVPR*, 2007.

[6] H. Dee and D. Hogg. Detecting Inexplicable Behaviour. In *BMVC*, 2004.

[7] M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *PAMI*, 24:381–396, Mar. 2002.

[8] 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 *CVPR*, 2005.

[9] W. Hu, X. Xiao, Z. Fu, and D. Xie. A System for Learning Statistical Motion Patterns. *PAMI*, 28:1450–1464, Sept. 2006.

[10] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust Online Appearance Models for Visual Tracking. *PAMI*, 25:1296–1311, Oct. 2003.

[11] J. Li, S. Gong, and T. Xiang. Global Behaviour Inference Using Probabilistic Latent Semantic Analysis. In *BMVC*, 2008.

[12] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental Learning for Robust Visual Tracking. *IJCV*, 77:125–141, May 2008.

[13] V. Shet, D. Harwood, and L. Davis. Multivalued Default Logic for Identity Maintenance in Visual Surveillance. In *ECCV*, 2006.

[14] C. Stauffer and E. Grimson. Learning Patterns of Activity Using Real-Time Tracking. *PAMI*, 22:747–757, Aug. 2000.

[15] M. Tipping and C. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 1999.

[16] X. Wang, X. Ma, and E. Grimson. Unsupervised Activity Perception by Hierarchical Bayesian Models. In *CVPR*, 2007.

[17] O. Williams, M. Isard, and J. MacCormick. Estimating Disparity and Occlusions in Stereo Video Sequences. In *CVPR*, pages 250–257, 2005.

[18] T. Xiang and S. Gong. Incremental and Adaptive Abnormal Behaviour Detection. *CVIU*, 111:59–73, June 2008.

[19] H. Zhong, J. Shi, and M. Visontai. Detecting Unusual Activity in Video. In *CVPR*, 2004.