Abnormality detection using low-level co-occurring events

Yannick Benezeth, Pierre-Marc Jodoin, Venkatesh Saligrama

Abnormality Detection Using Low-Level Co-occurring Events

Yannick Benezeth[a], Pierre-Marc Jodoin[b], Venkatesh Saligrama[c]

[a]*Orange Labs - France Telecom R&D*
4, rue du Clos Courtel - 35512 Cesson Sévigné - France
[b]*Université de Sherbrooke*
2500 bd. de l'Université Sherbrooke, J1K 2R1, Canada
[c]*Boston University - Department of Electrical and Computer Engineering*
8 Saint Mary's Street, Boston, MA 02215, USA

Abstract. We propose in this paper a method for behavior modeling and abnormal events detection which uses low-level features. In conventional object-based approaches, objects are identified, classified, and tracked to locate those with suspicious behavior. We proceed directly with event characterization and behavior modeling using low-level features. We first learn statistics about co-occurring events in a spatio-temporal volume in order to build the normal behavior model, called the *Co-Occurrence Matrix*. The notion of co-occurring events is defined using *Mutual Information* between motion labels sequences. Then, in the second phase, the co-occurrence matrix is used as a potential function in a *Markov Random Field* framework to describe, as the video streams in, the probability of observing new volumes of activity. The co-occurrence matrix is thus used for detecting moving objects whose behavior differs from the ones observed during the training phase. Interestingly, the Markov Random Field distribution implicitly accounts for speed, direction, as well as the average size of the objects without any higher-level intervention. Furthermore, when the spatio-temporal volume is sufficiently large, the co-occurrence distribution contains the average normal path followed by moving objects. Our method has been tested on various indoor and outdoor videos representing various challenges.

1. Introduction

In this paper, we present a low-level location-based approach for activity analysis and abnormal detection. In several traditional approaches (e.g. Hu et al., 2004), moving objects are first detected, analyzed and then tracked. Subsequently, behavior models are built based on object tracks and non-conformant ones are deemed abnormal. The main problem with this approach is that in case of complex environments, object extraction and tracking are performed directly on *cluttered* raw video or motion labels. We propose performing activity analysis and abnormal behavior detection first, followed possibly by object extraction and tracking. If the abnormal activity is reliably identified, then object extraction and tracking focus on *region of interest* (ROI) and thus is relatively straightforward. A question arises: *How to reliably identify abnormalities from a raw video?*

Some approaches have been proposed to perform such low-level abnormality detection (Adam et al., 2008; Jodoin et al., 2008). Nevertheless, we point out that these methods process each pixel independently and thus ignore spatial correlation across space and time. These correlations may not only be important in improving false alarms and misses but also in detecting abnormality of event sequences, such as a person in the act of dropping a baggage or a car making an illegal u-turn, etc. In our method, we account for these scenarios through spatio-temporal models. Although this model is simple, it nonetheless produces interesting results.

2. Previous work

Video analytics can be divided into two broad families of approaches namely *shape/pattern-recognition-based methods* and the *machine-learning-based methods*. The shape/pattern recognition approaches are typically those for

which the type of activity or object is known *a priori*. Examples of such methods include facial recognition systems (Zhao et al., 2003; Hu et al., 2009), restricted-area access detection (Konrad, 2005), car counting (Friedman and Russell, 1997), detection of people carrying cases (Haritaoglu et al., 2000), abandoned objects detection (Smith et al., 2006; Lim et al., 2006), action recognition (Ahmad and Lee, 2008), plate recognition, group detection, etc. These methods clearly focus on finding good matches between objects in a video and known templates stored in a database.

By their nature, shape recognition methods require a list of objects or behavior patterns that are anomalous. Unfortunately, this is not always possible, especially when suspicious activities cannot be known *a priori*. An alternative approach advocated in recent years is based on learning "normal" behavior from a video sequence exhibiting regular activity and then flag moving objects whose behavior deviates from normal behavior. In these methods, a learning phase serves as a behavior summarization step which is then used to discriminate between normal and abnormal patterns. As discussed in different review papers (Chen et al., 2005; Buxton, 2003; Hu et al., 2004), many such methods implement a general pipeline-based framework; moving objects are first detected in a motion detection step, then they are classified and tracked over a certain number of frames and finally, the resulting paths are used to distinguish "normal" objects from "abnormal" ones (Junejo et al., 2004; Stauffer and Grimson, 2000; Hu et al., 2006; Saleemi et al., 2009; Xiaogang et al., 2008). Although track-based methods have proven successful in different applications, they nevertheless suffer from fundamental limitations. First, implementing such pipeline methods can result in a fragile architecture which may suffer from a domino effect as an error can propagate to the subsequent processing stages. Secondly, tracking multiple objects at the same time is very demanding and is hardly efficient in crowded areas where objects merge or are partially occluded. Thirdly, tracking is efficient mostly with rigid moving bodies such as cars, trains, or pedestrians, and is not well suited to deal with unstructured motion such as waves on the water or tree shaking due to wind gusts.

To address these limitations, some authors have recently proposed learning methods based on characteristics other than motion paths. One such method is Pruteanu-Malinici and Carin (2008) 's approach which extracts features from each entire frames. The time-evolving properties of these features are thus modeled via an Infinite Hidden Markov Model (IHMM). Then, Boiman and Irani (2007) 's approach rebuilds observed sequences with small clips of videos taken from a database and exhibiting normal behaviors. In this case, abnormal activities are located whenever pieces of video cannot be rebuilt. While this method is mostly color-based, Adam et al. (2008) propose an optical-flow-based solution where pixel by pixel statistical distribution of motion vectors is learnt. Here suspicious activity is identified by detecting abnormal deviations from normal motion vectors. Jodoin et al. (2008) propose a pixel-by-pixel approach to learn patterns of activity. With their method, abnormalities are detected through a so called behavior subtraction procedure which amounts to flagging unusually high amounts of activity at each pixel. Unfortunately, both methods, (Adam et al., 2008; Jodoin et al., 2008) are only temporally sensitive and do not account for spatial abnormalities.

The main focus of this paper is to propose a simple low-level method for learning patterns of activity. As opposed to path-based approaches, we do not rely on tracking or any shape recognition procedure. The model accounts for spatial and temporal co-occurrences of activity and is robust to noisy motion label fields. Although the notion of co-occurrence of activity is not new, the notion of co-occurrence in previous work is accounted at a much higher level of abstraction (Xiang and Gong, 2006; Wang et al., 2009) or between trajectories observed across a network of cameras (Wang et al., 2010).

3. Context, Overview and Notation

3.1. Context

Although many video analytics methods use motion labels only in early stages of processing (mainly to locate moving objects) we argue that they carry fundamental information on the content of the scene and thus, can be used to perform high-level tasks. Motivated by this perspective, some authors have already shown that low-level motion labels can be used to summarize videos (Pritch et al., 2008), recognize human movements (Bobick and Davis, 2001) and detect abnormalities (Jodoin et al., 2008).

In general, motion labels sequences provide valuable information for characterizing "usual behavior" observed at each pixel. For instance, consider patterns associated with random activity (shaking tree), regular activity (highway traffic), bursty activity (due to traffic light), or simply inactivity. All of these scenarios are characterized by patterns of motion labels sequences at the pixel-level (or in general location). Consequently, abnormal behavior can be detected

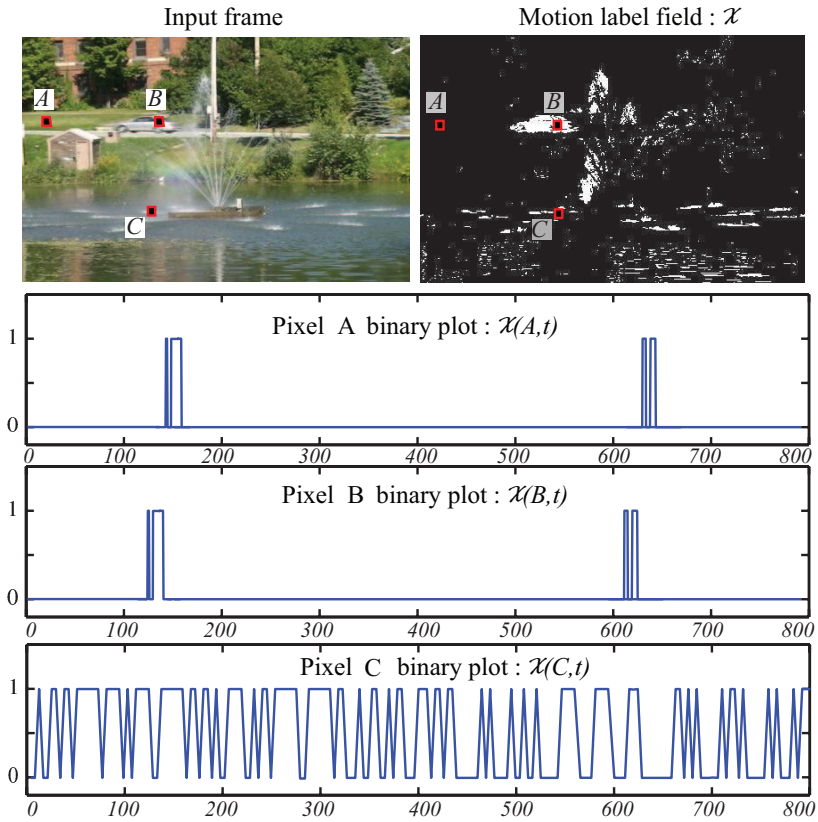


Figure 1: Binary signature for three pixels, two being highly correlated (A and B).

using low-level features whenever the observed pattern is unlikely under the normal activity model. In these cases, object identification and tracking can be circumvented for detecting abnormal behavior. As shown by Ermis et al. (2010), activity based on motion labels is particularly well-suited for modeling behavior since it exhibits invariance to geometry under general conditions.

However, the pure pixel-by-pixel approach is insufficient in applications where abnormality is manifested spatially as, for instance, cars running against traffic flow, cars making illegal u-turns, etc. Consequently, we need a strategy for incorporating spatial patterns in addition to the temporal patterns of motion labels sequences. The shortcomings of characterizing purely temporal behavior is further depicted in Fig. 1, which shows two pixels with identical signatures (except for a time-shift arising from cars going from right to left). Normal/Abnormal behavior arising from the pattern of activity between the two pixels cannot obviously be captured through a purely pixel-by-pixel analysis. For instance, a burst of activity occurring at pixel A before pixel B would mean that a car now runs from left to right.

3.2. Overview and Notation

The reader can follow the upcoming exposition through Fig. 2. Let $I_{\vec{x},k}$ be the luminance (or color) of a video sequence sampled on a 2-D lattice of size $Q_0 \times R_0$ at discrete time k, i.e., $\vec{x} \in Q_0 \times R_0 \subset R^2$, $k \in Z^+$. To simplify notation, we use s to denote the pixel location $\vec{x}$ at time t. X is a motion label field where $X_s \in \{0, 1\}$ specifies if a site s has an "inactive" or "active" state. Motion labels are obtained through a background subtraction procedure which subtracts and then thresholds a background image $B_{\vec{x}}$ to each frame $I_{\vec{x},k}$ (Benezeth et al., 2010). We also define the motion labels sequence centered at $s = (\vec{x}, t)$ as being $\vec{X}_s = [X_{\vec{x},t-\eta}, \ldots, X_{\vec{x},t+\eta}]$ where $2\eta + 1$ is the length of the vector $\vec{X}_s$. In short, $\vec{X}_s$ is a one-dimensional binary sequence at pixel $\vec{x}$ and time t as shown in Fig.1. A contiguous sequence of ones denotes a busy period and is associated with a passing object while a sequence of zeros corresponds to an idle period of activity. The entire spatio-temporal sequence can be alternatively defined over a 3D lattice $\mathcal{S}$ of size

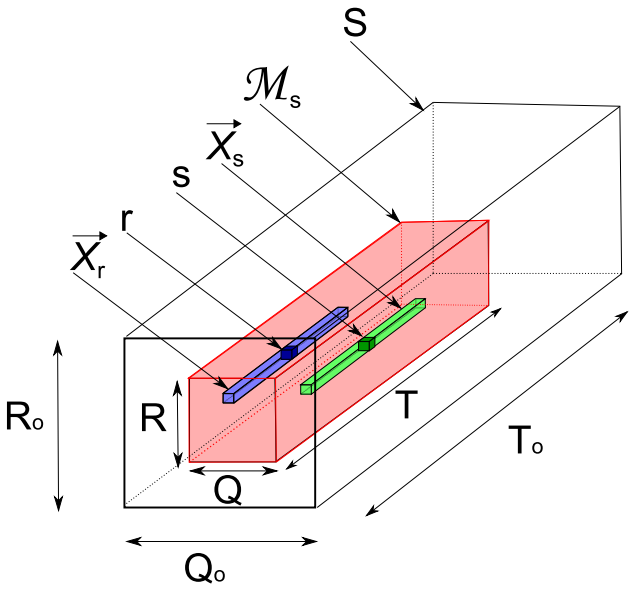


Figure 2: 3D lattice S with spatio-temporal neighborhood $\mathcal{M}_s$.

$Q_0 \times R_0 \times T_0$ with $s \in \mathcal{S}$ being a point in the spatio-temporal space, I_s being the corresponding luminance (or color) and X_s the corresponding motion label.

Now lets consider for each pixel $\vec{x}$ at time t, a spatio-temporal neighborhood centered on $s = (\vec{x}, t)$. This neighborhood is a 3D lattice $\mathcal{M}_s \subset \mathcal{S}$ with size $Q \times R \times T$, $Q < Q_0$, $R < Q_0$ and $T << T_0$, centered on $s \in \mathcal{S}$. Let us also consider a location $r = (\vec{y}, \tau) \in \mathcal{M}_s$ in the spatio-temporal neighborhood of $s = (\vec{x}, t)$. The spatial neighborhood of a pixel $\vec{x}$ is the set of all pixels $\vec{y}$ such that $s = (\vec{x}, t)$ and $r = (\vec{y}, \tau)$ are both in $\mathcal{M}_s$ for all t.

As we mentioned previously, whenever a moving object passes in front of $\vec{x}$ at time t, it leaves a spatio-temporal trace as some sites $r = (\vec{y}, \tau) \in \mathcal{M}_s$ co-occur with $s = (\vec{x}, t)$. Interestingly, several moving objects exhibiting regular behavior (think of cars on a highway going in the same direction) leave, after a while, similar traces in the neighborhood $\mathcal{M}_s$. Interestingly, the co-occurrence of two spatio-temporal neighbors s and r is not only due to the position and orientation of the camera in the scene, but also due to the shape, velocity and direction of the moving objects passing in front of a given spatial location $\vec{x}$. In this context, the goal of the co-occurrence matrix is to estimate how frequently a site r co-occurs with s given a training video sequence exhibiting normal activity.

Let us now define the notion of co-occurrence. A site $r \in \mathcal{M}_s$ co-occurs with s whenever their corresponding motion vector $\vec{X}_s$ and $\vec{X}_r$ exhibit a similar signature. The similarity between motion vectors at s and r is expressed using the mutual information defined as:

$$sim(\vec{X}_s, \vec{X}_r) = \sum_{m \in \{0,1\}} \sum_{n \in \{0,1\}} P_{\vec{X}_s \vec{X}_r}(m,n). \log\left(\frac{P_{\vec{X}_s \vec{X}_r}(m,n)}{P_{\vec{X}_s}(m) P_{\vec{X}_r}(n)} \right) \quad (1)$$

where $\vec{X}_s(i) = m$ and $\vec{X}_r(i) = n$, m and $n \in \{0,1\}$, $i = [t-\eta, \ldots, t+\eta]$, $P_{\vec{X}_s \vec{X}_r}(m,n)$ is the joint probability of discrete variables $\vec{X}_s$ and $\vec{X}_r$ and $P_{\vec{X}_s}(m)$ and $P_{\vec{X}_r}(n)$ are the marginal probabilities.

The mutual information is a useful tool for determining whether two motion labels sequences contain the same activity. For example, a temporal sequence of motion labels containing random values due to noise or false detections (caused, say, by an unstable background) will have a low mutual information with almost any other sequence. On the other hand, two sequences containing the trace left by the same moving object will have a large mutual information. In this way, the mutual information criteria minimizes the influence of spurious false detections and noisy environments.

4. Our Method

In this section, we present how, for a given site s, a co-occurrence matrix and its associated statistical model can be estimated from a training video sequence. Our statistical model is a Markov-Random Field (MRF) model that accounts for the likelihood of the co-occurrences. Since we account for normal scenarios in which objects follow typical paths, these paths manifest themselves as spatio-temporal dependencies across pixels as shown in Eq. 1. Our location-based approach for modeling normality uses a joint distribution of pixel-level activity modeled via a probabilistic model. To our knowledge, an MRF model is the simplest such model that accounts for correlation in space and time of pixel activity. We later present how abnormal events can be detected and how low-level connected graphs can be used to follow relevant moving objects.

4.1. Training Phase

Nominal Model. Let O_s denote a motion label volume in the spatial-neighborhood of location s, *i.e.* $O_s = (X_r : r \in \mathcal{M}_s)$. We are interested in modeling the likelihood of the normal observations, *i.e.*, $P_N(O_s)$. We do this using an MRF model parameterized through co-occurrences:

$$P_N(O_s) = \frac{1}{Z} \exp\left(\sum_{u,v \in \mathcal{M}_s} \alpha_{uv} sim(\vec{X}_u, \vec{X}_v) \right) \tag{2}$$

where $sim(\vec{X}_u, \vec{X}_v)$ is the mutual information between motion labels vectors $\vec{X}_u$ and $\vec{X}_v$ (as defined in Eq. 1). α_{uv} is the co-occurrence potential between site u and v determined in a learning phase as it will shortly be described (for the remainder of the paper, α_{uv} will be refereed to as the co-occurrence matrix). Z is the usual partition function, which is a normalization constant to ensure that the right hand side sums to one.

Learning the Co-Occurrence Matrix. As mentioned previously, the co-occurrence matrix α_{uv} accounts for how many times sites u and v co-occur during the training phase. Two sites are said to co-occur whenever their motion signature $\vec{X}_u$ and $\vec{X}_v$ exhibit a similar profile. In this paper, we measure the similarity between two sites based on their mutual information.

The co-occurrence matrix α_{uv} of two spatio-temporal locations, $u, v \in \mathcal{M}_s$ can be empirically computed as follows:

$$\alpha_{uv} = \frac{\beta_{uv}}{T_0 - T} \sum_{t=T/2}^{T_0 - T/2} sim(\vec{X}_u, \vec{X}_v) \tag{3}$$

where T_0 is the total number of frames in the training video sequence and β_{uv} is a constant that can depend on distance between the locations u and v (in this paper we assume $\beta_{uv} = 1$). Note that by definition, α_{uv} does not depend on the time index t. Therefore,

$$\alpha_{uv} = \alpha_{(\vec{y}_1, t+\tau_1),(\vec{y}_2, t+\tau_2)} = \alpha_{(\vec{y}_1, \tau_1),(\vec{y}_2, \tau_2)}. \tag{4}$$

A Specific Case for Co-Occurrence. Benezeth et al. (2009) show that the co-occurrence between two sites s and r can be determined by considering motion labels values X_s and X_r instead of the motion labels sequences $\vec{X}_s$ and $\vec{X}_r$. In this way, two sites co-occur whenever $X_s = X_r = 1$. In this case α_{uv} can be easily computed. However, this formulation is sensitive to noise and spurious false positives caused by unstable background. As can be seen in Fig. 11, accounting for plain co-occurrence between motion labels (third row) generates a large number of false positives and poor detection of true moving objects. This clearly shows how mutual information allows for *essential co-occurrences*, *i.e* co-occurrences caused only by real moving objects.

Complexity Issues & Conditional Independence. The main issue is the cost of computation of all of the edge potentials, since they are combinatorially many. In our practical implementations, we typically only consider a sparse number of well-separated locations for testing abnormalities. In many of our applications, abnormalities are typically associated with patterns of abnormal activity as opposed to inactivity. Motivated by this perspective, we make the following simplifying assumption: for any spatio-temporal neighborhood, $\mathcal{M}_s$ centered around $s = (\vec{x}, t)$, the co-occurrences are conditionally independent given X_s is active (namely $X_s = 1$). It will become clear why this assumption is not meaningful when $X_s = 0$. In other words, given X_s the values realized at the spatio-temporal locations X_v and X_u are statistically independent. Alternatively, one may think of this assumption as an instantiation of a naive Bayes perspective, namely, we assume that the pairwise co-occurrences in the spatial neighborhood of a location s are all independent. Practically, this assumption implies that we must have,

$$\alpha_{uv} = 0, \;\; u \neq s, \; v \neq s \tag{5}$$

In practice we have found this assumption does not severely degrade performance in our applications. Note that from a pure implementation perspective, the co-occurrence matrix $[\alpha_{uv}]$ is a 3D array with each component accounting for the number of times each site u co-occur with v while translating $\mathcal{M}_s$.

4.2. Observation Phase

Abnormal Model. It is generally difficult to describe an abnormality model except to say that abnormality is anything that does not look normal. However, from a classification perspective it becomes necessary to make some implicit assumptions about abnormality. Several researchers implicitly assume that abnormal observations are uniformly distributed in the feature space Polonik (1997). Our assumption is that abnormal observations are independent and identically distributed across the different pixels. This assumption amounts to a multinomial distribution. For simplicity, let $N_0 = |\mathcal{M}_s|$ be the total number of spatio-temporal locations and N_1 the total number of co-occurring pixels, i.e.,

$$N_1 = \sum_{u \in \mathcal{M}_s} f(\vec{X}_u, \vec{X}_s) \tag{6}$$

with

$$f(\vec{X}_u, \vec{X}_s) = \begin{cases} 1 & \text{if } \; sim(\vec{X}_u, \vec{X}_s) > \tau \\ 0 & \text{whereas} \end{cases} \tag{7}$$

then, the probability distribution of observations under the abnormal distribution is given by,

$$P_A(O_s) = p^{N_1}(1-p)^{N_0 - N_1} = \left(\frac{p}{1-p}\right)^{N_1} (1-p)^{N_0} \tag{8}$$

where, p is the probability that $f(\vec{X}_u, \vec{X}_s) = 1$

Abnormality Detection. Consider now a test video sequence $\mathcal{S}$ defined on a 3D lattice of size $Q_0 \times R_0 \times T_{test}$, a spatio-temporal neighborhood $\mathcal{M}_s$ with $s = (\vec{x}, t)$ in the test video, and its corresponding motion-label observations O_s. The goal now is to detect every time instant $t \in [0, T_{test}]$ for which the observations O_s has a low probability under the nominal distribution in comparison to likelihood of abnormality. It is well-known that the likelihood ratio test (LRT) is the optimal test for deciding between the two hypothesis: nominal vs. abnormal. The likelihood ratio $\ell(O_s)$ is the ratio of the probability of observations under nominal and abnormal hypothesis, from Eq. (2), (5) and (8), it follows:

$$\begin{aligned} \ell(O_s) &= \frac{P_N(O_s)}{P_A(O_s)} \\ &= \frac{1}{Z(1-p)^{N_0}} \exp\left(\sum_{r \in \mathcal{M}_s} \alpha_{sr} sim(\vec{X}_s, \vec{X}_r) - \log \frac{p}{1-p} \left(\sum_{r \in \mathcal{M}_s} f(\vec{X}_r, \vec{X}_s) \right) \right) \end{aligned} \tag{9}$$

where, as before, N_0 is the number of spatio-temporal locations and Z is a normalization constant.

The likelihood ratio test is to decide between nominal and abnormal hypothesis based on a global threshold η:

$$\ell(O_s) = \exp\left(\sum_{r\in\mathcal{M}_s} \alpha_{sr} sim(\vec{X}_s, \vec{X}_r) - \tau \sum_{r\in\mathcal{M}_s} f(\vec{X}_r, \vec{X}_s)\right) \underset{abnormal}{\overset{nominal}{\gtrless}} \eta \tag{10}$$

where $\tau = \log(p/1-p)$. Here we have absorbed Z, p^{N_0} into η. A related test obtained by choosing $\eta = 1$ above reduces to a test for positivity or negativity of the argument of the exponential function. This reduces to the following simple test:

$$\frac{\sum_{r\in\mathcal{M}_s} \alpha_{sr} sim(\vec{X}_r, \vec{X}_s)}{\sum_{r\in\mathcal{M}_s} f(\vec{X}_r, \vec{X}_s)} \underset{abnormal}{\overset{nominal}{\gtrless}} \tau. \tag{11}$$

4.3. Dealing with multiple moving objects

The test of Eq. (11) allows one to determine which observation O_s is normal and which one is not according to the co-occurrence matrix α_{sr} learned during the training phase. However, for any large $\mathcal{M}_s$, more than one object may leave a trace in O_s. Indeed, consider for example, a broken down car on a highway with parallel traffic. In this case, if $\mathcal{M}_s$ is large enough, both the abandoned car and the moving ones leave a trace O_s although only the broken down car is clearly of interest. One simple and efficient way of identifying only the moving objects which are associated with pixel $\vec{x}$ is by selecting every site $r \in \mathcal{M}_s$ which not only co-occurs with site s but also are connected to s (there is a connected graph of 1s which goes from r to s in O_s). This idea can be used for instance for tracking a person dropping a baggage (once a baggage drop has been identified as abnormal).

Another issue is what happens once an abnormality has been declared. To see this, consider the previous example of a car passing close to an abandoned car once the abandoned car has been declared as abnormal. With our algorithm their respective spatio-temporal traces will be fused into just one connected graph. Thus, the probability of the observed spatio-temporal trace will be modified by the abandoned car and every passing object can be declared abnormal. A simple way out of this situation is to compute a likelihood ratio test conditioned on observations generating the previous abnormality. If the trace $O_{\vec{x},t-1}$ has been declared as abnormal, one could compute the ratio of the intersection and union of past and current observations ($O_{\vec{x},t-1}$ and $O_{\vec{x},t}$). Our ratio amounts to:

$$\varepsilon = \frac{\sum_{r\in\mathcal{M}_s}(O_{\vec{x},t}(r) \ \wedge \ O_{\vec{x},t-1}(r))}{\sum_{r\in\mathcal{M}_s}(O_{\vec{x},t}(r) \ \vee \ O_{\vec{x},t-1}(r))}. \tag{12}$$

Thresholding ε provides a test for whether the observed spatio-temporal trace is composed of the union of the previous abnormal detection plus a new observation or just an update of $O_{\vec{x},t-1}$. If $\varepsilon < \gamma$, where γ is some threshold, we can then conduct a LRT on the innovation $O'_{\vec{x},t} = O_{\vec{x},t} - O_{\vec{x},t-1}$, where $O'_{\vec{x},t}$ represents the spatio-temporal trace of just the new observation. This LRT is precisely the LRT conditioned on the previously detected abnormal trace. In this way, one can ignore non-abnormal events once an abnormality is detected and update new abnormalities as they arise.

This is illustrated in the example presented in figure 6, when the man is passing in front of the abandoned luggage previously detected as abnormal, we compute the LRT of the spatio-temporal trace left by the walking man without the trace left by the bag.

5. Experimental results

We present in this section some results obtained on various indoor and outdoor sequences representing different challenges. For each sequence, a co-occurrence matrix of size ranging between $130 \times 70 \times 300$ and $210 \times 210 \times 150$ have been used. The size of the co-occurrence matrix is chosen so that a typical normal activity is entirely included in the volume. The reader shall note that since the matrix' size stay fix for the entire process, it has to be fixed only once while setting up the system. The number of frames T used to estimate P_N (Eq. 2) varies between 2000 and 7000 (*i.e.* from 1 and 4 minutes of video) depending on the sequence.

Note that results are presented in Fig. 4, 6, 8, 10 and 11. The green moving objects are ones classified as being normal and the red moving objects are those classified as being abnormal, i.e., whose trace is significantly different from the co-occurrence matrix Eq. (11).

The first example (see Fig. 4) shows normal traffic and a car making an illegal u-turns. In Fig. 3(a), a co-occurrence matrix associated with a normal traffic flow is presented. As shown in Fig. 3(c), the trace left by the u-turn significantly differs from the usual traffic flow illustrated in the Fig. 3(b). Cars following the regular path are tagged in green and cars making an illegal u-turn are tagged in red.

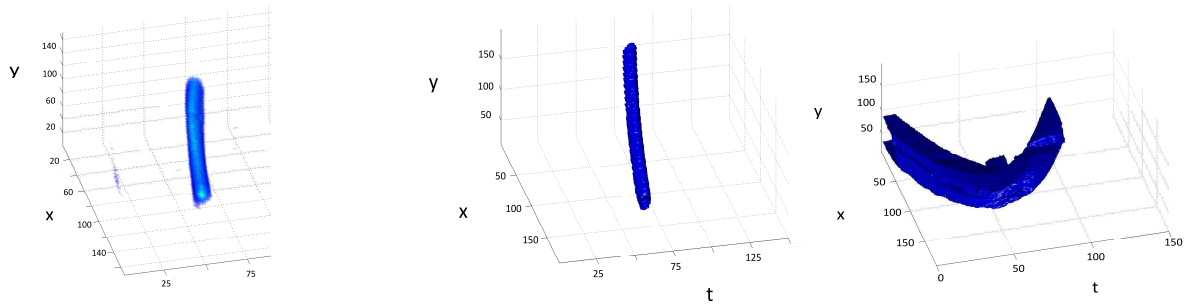


Figure 3: (a) Co-occurrence matrix of a regular traffic flow (b) one car moving along the regular path (c) the trace left by a car making an illegal u-turn.

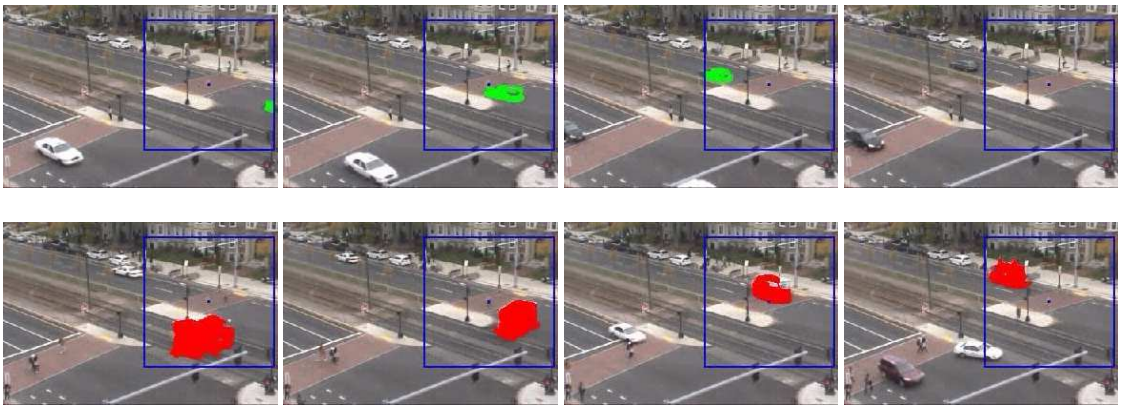
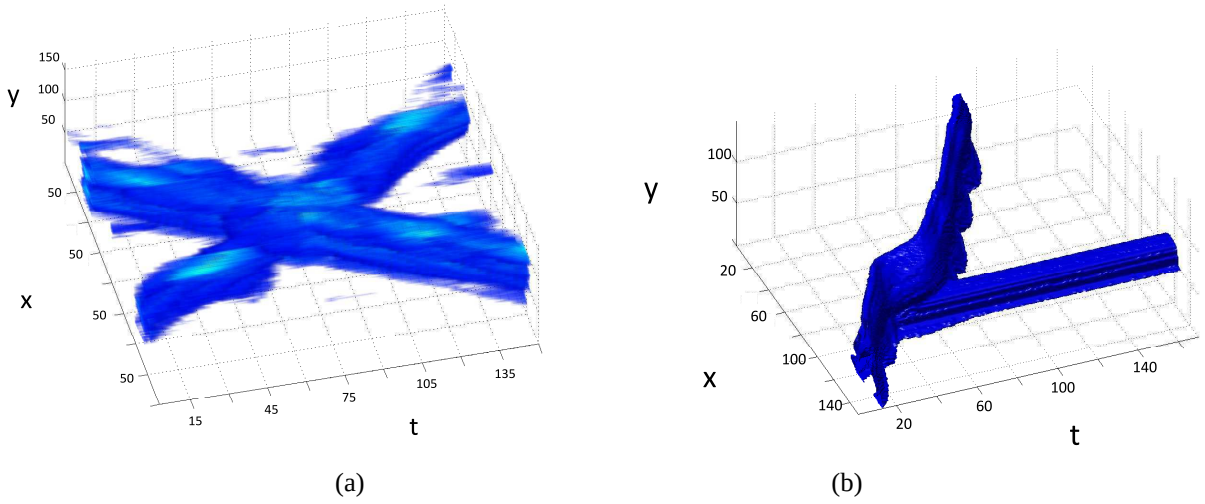


Figure 4: Example video in which cars following the regular traffic flow are tagged in green while the car making an illegal u-turn have been picked up by our algorithm and tagged in red.

The second example shows a person dropping a baggage and abandoning it. In this video, pedestrians usually walk from left to right and from right to left, hence the X shape of the co-occurrence matrix (see Fig. 5(a)). When the person drops the bag, the abandoned package leaves a straight elongated line which differs from the co-occurrence matrix and thus causes this situation to be suspicious (see Fig. 5(b) and Fig. 6). Interestingly, following the process described in section 4.3, all pedestrians walking across the bag are not recognized as being suspicious (see the green pedestrian and the red bag in Fig. 6). Note that the connected graph processing presented in Section 4.3 allows our method to track suspicious moving objects, even after the suspicious event occurred. This is shown for the person dropping the bag (Fig. 6) as well as for the car making an illegal u-turn (Fig 4).

The next two examples concern the detection of abandoned luggages in indoor places (a train station and a corridor). In figures 7(a) and 9(a), we present the regular activity model obtained from a training sequence. In both cases, we the co-occurrence matrix has two principal modes of activity (the X-shape). The trace left by a pedestrian crossing



(a) (b)

Figure 5: (a) co-occurrence matrix of pedestrians walking from left to right and from right to left and (b) the trace left by a person dropping a bag.

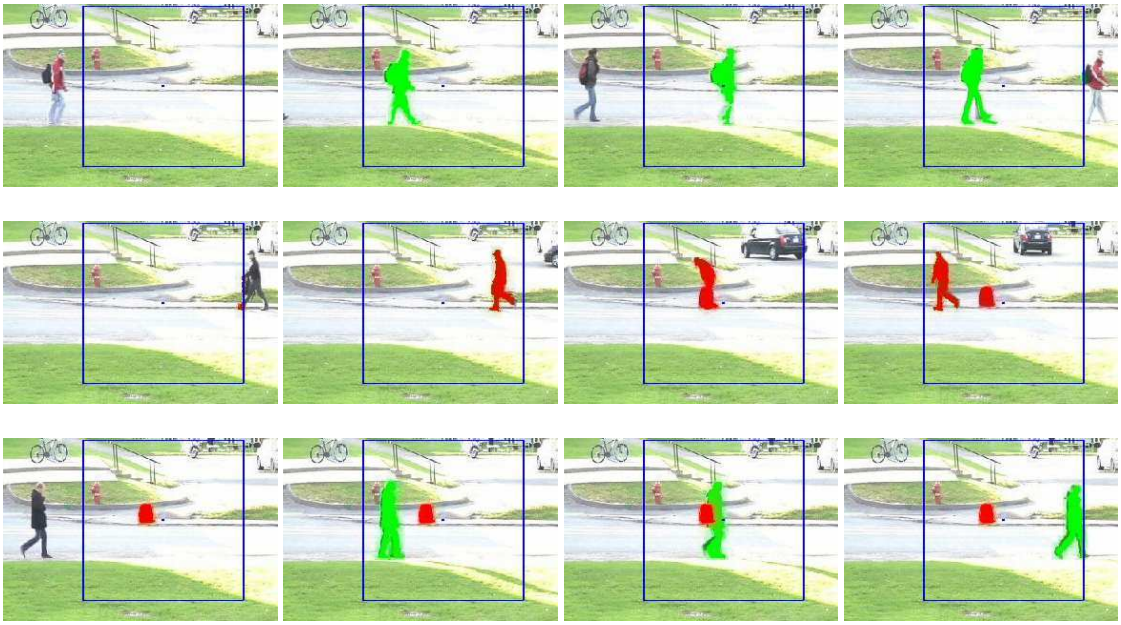
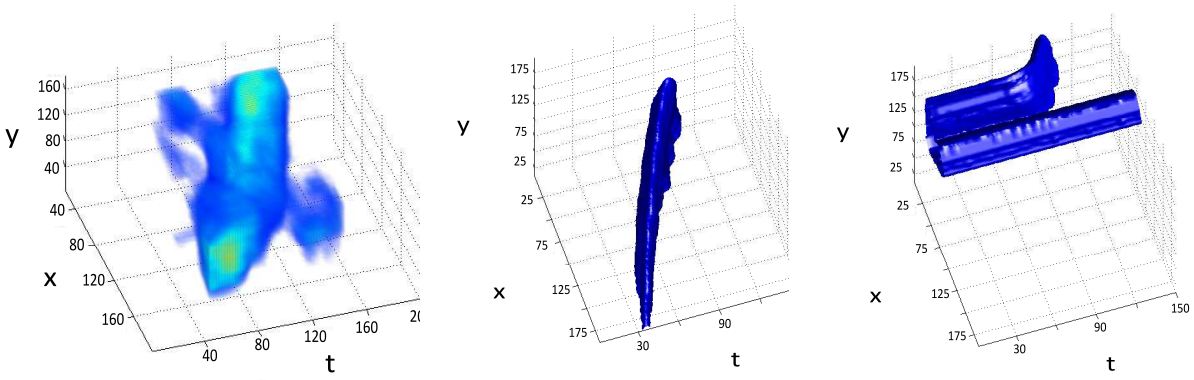


Figure 6: Example video in which people walking are tagged in green while the person dropping a bag is tagged in red.

the scene are shown in Fig.7(b) and 9(b) while the trace left by a person dropping a bag are shown in Fig.7(c) and 9(c). The difference between normal and abnormal traces being obvious, the abandoned bag has been identified as suspicious. Thumbnails of these two examples are presented in figures 8 and 10.

The fifth example, in Fig. 11, shows how our method deals with noisy environments. The third row presents results obtained considering co-occurrences with motion labels (Benezeth et al., 2009) while the fourth row presents results obtained considering motion labels vectors and mutual information. Clearly, the use of mutual information reduces the sensitivity to noise as the boat is clearly detected. The spatio-temporal trace left by the boat is shown in Fig. 12.

Unfortunately, few abnormality-detection papers quantitatively evaluate the strengths and the weaknesses of their method with videos compatible with our method. For example, Lim et al. (2006) present a qualitative evaluation of



(a) (b) (c)

Figure 7: (a) co-occurrence matrix at one location of a train station (b) trace left a pedestrian (c) trace left by a person dropping a bag.

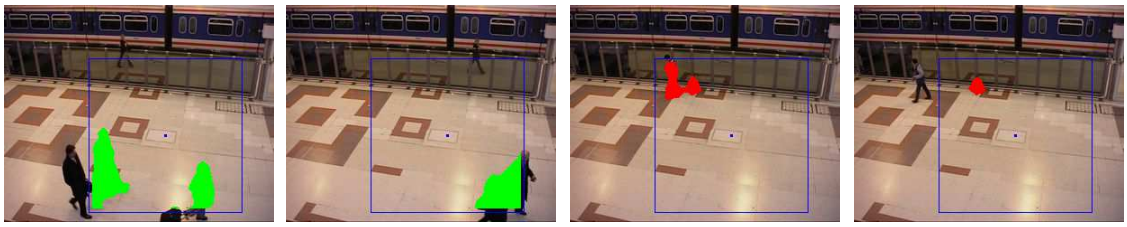
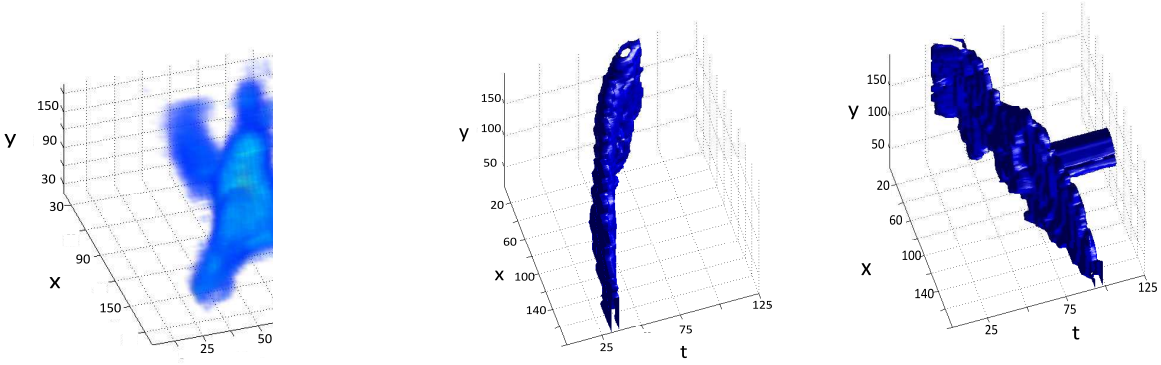


Figure 8: Example video - from (Thirde et al., 2006) - taking place in a train station in which a suspicious event has been tagged in red.



(a) (b) (c)

Figure 9: (a) co-occurrence matrix at one location of a corridor (b) trace left by a pedestrian (c) trace left by a person dropping a bag.

their method using only one video while Pruteanu-Malinici and Carin (2008) propose a quantitative evaluation using 70 small video clips each containing 20 frames approximatively. For obvious reasons, comparison with these methods is difficult if not impossible.

In order to limit the impact of the video dataset's heterogeneity, workshops and competitions such as PETS (Thirde et al., 2006) or ETISEO (Nghiem et al., 2007) put annotated video datasets on the web. Unfortunately, these datasets are often dedicated to specific tasks (pedestrian detection, object tracking on water, face localization, etc.). These goals are quite different from the objectives of our method.

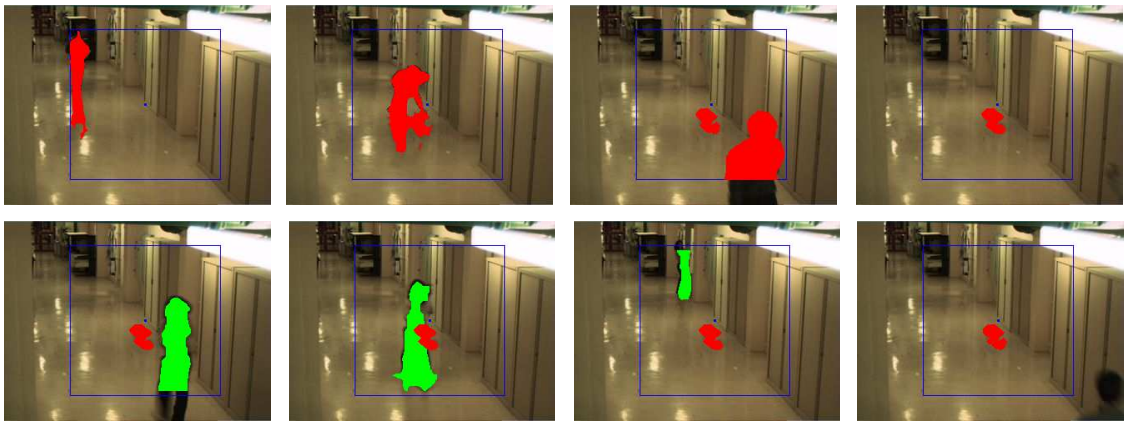


Figure 10: Example video - from (Nghiem et al., 2007) - taking place in a corridor in which suspicious event is tagged in red.

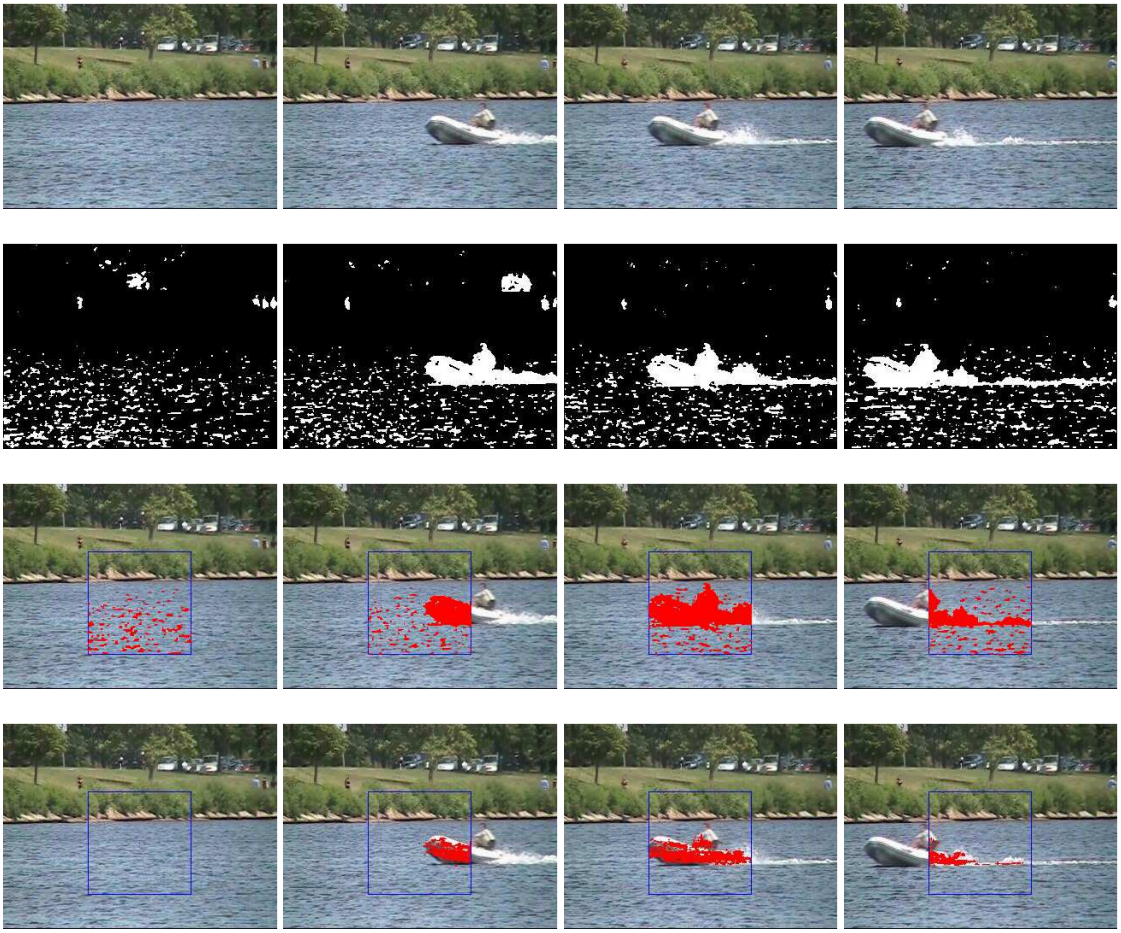


Figure 11: Illustration of detection in a challenging environment. The first row present the input images sequence, the second present the result of a background subtraction detection followed by the detection of abnormal activities (the boat displacement) considering co-occurrences with motion labels in the third row (Benezeth et al., 2009)or considering motion labels vectors in the fourth row.

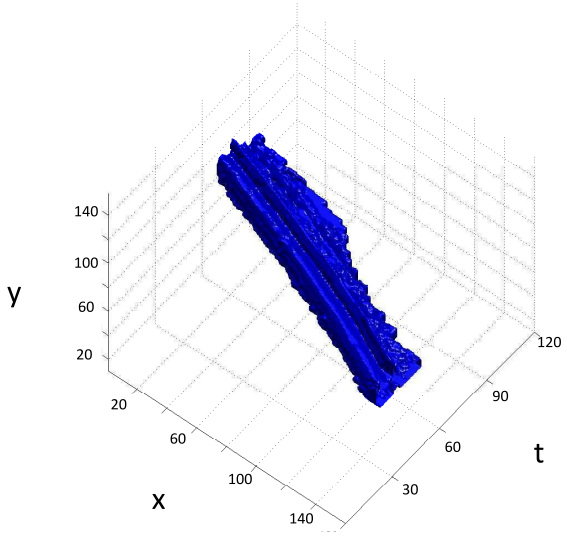


Figure 12: Trace left by the the boat presented in Fig.11.

In this paper, we use a video dataset containing different abnormal events recorded in various environments. Our dataset is made of five videos (see Fig. 4, 6, 8 and 10) containing roughly 100 normal events and 15 abnormal events. We compare our method with a track-based technique (Chen et al., 2005) and put the quantitative results in table 1. As one can see, our method detects all abnormal events and produces only 9.5% of false positive. Note that every false positive was caused by objects whose size strongly differed from that of the majority.

Nevertheless, some authors present methods with videos that are compatible with our technique. Such an evaluation has been performed by Adam et al. (2008) which is based on various real-life videos ranging from 10 minutes to 5 hours. The authors mention that their method detects roughly 90% of abnormal events and less than 6 false alarms per video. Such performance is roughly similar to that of our method. This being said, it is not clear how these actions can be considered as being abnormal considering the nature of the videos (people walking and dawdling in a mall and a subway).

	Normal	Abnormal
Normal	80.1	19.9
Abnormal	10.0	90.0

(a)

	Normal	Abnormal
Normal	90.5	9.5
Abnormal	0.0	100.0

(b)

Table 1: Confusion matrices illustrating results obtained with (a) an object-based method (Chen et al., 2005) and (b) our method. Horizontal rows are ground truth and vertical columns are observations.

6. Conclusion

We propose in this paper a method to perform behavior modeling and abnormality detection based on low-level characteristics. We use the spatial and temporal dependencies between motion labels vectors obtained with simple background subtraction. To do this, we built a Markov Random Field model parameterized by a co-occurrence matrix. Although simple, this matrix contains the average behavior observed in a training sequence. It also implicitly contains information about direction, speed and size of objects usually passing through one (or more) key-pixel(s). Equipped with the co-occurrence matrix, we can detect abnormal events by detecting traces which significantly differ from our nominal model following a likelihood ratio test.

The main advantages of our method are threefold. First, in contrast to conventional object-based approaches for which objects are identified, classified and tracked to locate those with suspicious behavior, we proceed directly with event characterization and behavior modeling using low-level characteristics and thus avoid the risk of errors propagation (e.g. due to the tracking algorithm limits in complex environments). Second, our method does not require any *a priori* knowledge about the abnormal event detection. We learn the usual behavior of moving objects in a scene and detect activity which significantly differ from usual ones. Third, our method is robust to noise and can detect unusual activities using very noisy background subtraction masks.

7. References

W. Hu, T. Tab, L. Wang, S. Maybank, A Survey on Visual Surveillance of Object Motion and Behaviors, Transactions on System Man and Cybernetics - Part C: Applications and Reviews 34 (3) (2004) 334–352.

A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors, Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008) 555–560.

P.-M. Jodoin, J.Konrad, V. Saligrama, Modeling Background Activity for Behavior Subtraction, International Conference on Distributed Smart Cameras (2008) 1–10.

W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.

H. Hu, P. Zhang, Z. Ma, Direct kernel neighborhood discriminant analysis for face recognition, Pattern recognition letters 30 (10) (2009) 902–907.

J. Konrad, Motion detection and estimation, in: A. Bovik (Ed.), Handbook of Image and Video Processing, 2nd Edition, chap. 3.10, Academic Press, 253–274, 2005.

N. Friedman, S. Russell, Image Segmentation in Video Sequences: A Probabilistic Approach, international conference on Uncertainty in Artificial Intelligence (1997) 175–181.

I. Haritaoglu, D. Harwood, L. Davis, W4: Real-Time Surveillance of People and Their Activities, Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 809–830.

K. Smith, P. Quelhas, D. Gatica-Perez, Detecting Abandoned Luggage Items in a Public Space, Performance Evaluation of Tracking and Surveillance Workshop (PETS) (2006) 75–82.

S.-N. Lim, H. Fujiyoshi, R. Patil, A One-Threshold Algorithm for Detecting Abandoned Packages Under Severe Occlusions Using a Single Camera, Tech. Rep. CS-TR-4784, University of Maryland, 2006.

M. Ahmad, S.-W. Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (7) (2008) 2237–2252.

T. Chen, H. Haussecker, A. Bovyrin, R. Belenov, K. Rodyushkin, A. Kuranov, V. Eruhimov, Computer vision workload analysis: case study of video surveillance systems, Intel Technology Journal 9 (2) (2005) 109–118.

H. Buxton, Learning and Understanding dynamic scene activity: A review, Image and Vision Computing 23 (2003) 125–136.

I. Junejo, O. Javed, M. Shah, Multi Feature Path Modeling for Video Surveillance, International Conference on Pattern Recognition (2004) 716–719.

C. Stauffer, E. Grimson, Learning Patterns of Activity Using Real-Time Tracking, Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, S. Maybank, A System for Learning Statistical Motion Patterns, Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1450–1464.

I. Saleemi, K. Shafique, M. Shah, Probabilistic Modeling of Scene Dynamics for Applications in Visual Surveillance, Transactions on Pattern Analysis and Machine Intelligence 31 (8) (2009) 1472–1485.

W. Xiaogang, T. Keng, N. Gee-Wah, W.Grimson, Trajectory analysis and semantic region modeling using a nonparametric Bayesian model, International conference on Computer Vision and Pattern Recognition (2008) 1–8.

I. Pruteanu-Malinici, L. Carin, Infinite Hidden Markov Models for Unusual-Event Detection in Video, Transactions in Image Processing 17 (5) (2008) 811–822.

O. Boiman, M. Irani, Detecting Irregularities in Images and in Video, International Journal on Computer Vision 74 (1) (2007) 17–31.

T. Xiang, S. Gong, Beyond Tracking: Modeling Activity and Understanding Behavior, International Journal on Computer Vision 67 (1) (2006) 21–51.

X. Wang, X. Ma, E.Grimson, Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models, Transactions on Pattern Analysis and Machine Intelligence 31 (3) (2009) 539–555.

X. Wang, K. Tieu, E. Grimson, Correspondence-Free Activity Analysis and Scene Modeling in Multiple Camera Views, Transactions on Pattern Analysis and Machine Intelligence 32 (1) (2010) 56–71.

Y. Pritch, A. Rav-Acha, S. Peleg, Non-Chronological Video Synopsis and Indexing, Transactions on Pattern Analysis and Machine Intelligence 30 (11) (2008) 1971–1984.

A. F. Bobick, J. W. Davis, The Recognition of Human Movement Using Temporal Templates, Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

E. Ermis, P. Clarot, P.-M. Jodoin, V.Saligrama, Activity Based Matching in Distributed Camera Network, IEEE Trans. on Image Processing 19 (10) (2010) 2595–2613.

Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, C. Rosenberger, Comparative Study of Background Subtraction Algorithm, Journal of Electronic Imaging 19 (3) (2010) .

Y. Benezeth, P.-M. Jodoin, V. Saligrama, C. Rosenberger, Abnormal Events Detection Based on Spatio-Temporal Co-occurences, international conference on Computer Vision and Pattern Recognition (2009) 2458–2465.

W. Polonik, Minimum volume sets and generalized quantile processes, Stochastic Processes and Applications 69 (1) (1997) 1–24.

D. Thirde, L. Li, J. Ferryman, An overview of the pets 2006 dataset, international Workshop on Performance Evaluation of Tracking and Surveillance (2006) 47–50.

A. Nghiem, F. Bremond, M. Thonnat, V. Valentin, Etiseo, performance evaluation for video surveillance systems, international conference on Advanced Video and Signal Based Surveillance (2007) 476–481.