Abnormal events detection based on spatio-temporal co-occurences

Yannick Benezeth, Pierre-Marc Jodoin, Venkatesh Saligrama, Christophe Rosenberger

Abnormal Events Detection Based on Spatio-Temporal Co-occurences *

Y. Benezeth[1] P.-M. Jodoin[2] V. Saligrama[3] C. Rosenberger[4]

[1]ENSI de Bourges
Institut PRISME
88 bd. Lahitolle
18000 Bourges, France

[2]MOIVRE
Université de Sherbrooke
2500 bd. de l'Université
Sherbrooke, J1K 2R1, Canada

[3]Boston University
Electrical & Computer Engineering
8 St. Mary's Street
Boston MA, 02215, USA

[4]GREYC, ENSICAEN
Université de Caen - CNRS
6 bd. Maréchal Juin
14000 Caen, France

Abstract

We explore a location-based approach for behavior modeling and abnormality detection. In contrast to the conventional object-based approach where an object may first be tagged, identified, classified, and tracked, we proceed directly with event characterization and behavior modeling at the pixel(s) level based on motion labels obtained from background subtraction. Since events are temporally and spatially dependent, this calls for techniques that account for statistics of spatio-temporal events. Based on motion labels, we learn co-occurrence statistics for normal events across space-time. For one (or many) key pixel(s), we estimate a co-occurrence matrix that accounts for any two active labels which co-occur simultaneously within the same spatio-temporal volume. This co-occurrence matrix is then used as a potential function in a Markov Random Field (MRF) model to describe the probability of observations within the same spatio-temporal volume. The MRF distribution implicitly accounts for speed, direction, as well as the average size of the objects passing in front of each key pixel. Furthermore, when the spatio-temporal volume is large enough, the co-occurrence distribution contains the average normal path followed by moving objects. The learned normal co-occurrence distribution can be used for abnormal detection. Our method has been tested on various outdoor videos representing various challenges.

1. Introduction

In this paper, we present a location-based approach for activity analysis and abnormal detection. In several traditional approaches, described later, motion in the scene is usually detected first followed by object extraction and object tracking [8]. 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 are relatively straightforward, both in terms of difficulty and computational complexity, on account of sparsity and absence of clutter. A question, however, arises: *How to reliably identify pixel-level abnormalities, or more generally activities, from raw video or motion labels?*

As will be discussed in Section 2, some approaches have been proposed to perform such low-level abnormality detection [1, 10]. Nevertheless, we point out that those 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, tracking the person who dropped the baggage, a car making an illegal U-turn, etc. In our method, we account for these scenarios through spatio-temporal models based on frequency of co-occurrences of spatio-temporal neighborhoods. Although, the spatio-temporal model presented in this paper is simple, it results in extremely interesting results on traffic monitoring videos, abandoning of baggages followed by tracking etc. Note that our scheme does not rely on object tagging, tracking or classification. Furthermore, the co-occurrence can be readily generalized to higher-dimensional co-occurrences and other interesting features can be augmented with our approach. However, to keep the development simple, we focus on the simpler model in this paper.

*This research was supported by the ONR Young Investigator Program and Presidential Early Career Award (PECASE) N00014-02-100362, NSF CAREER award ECS 0449194, the Department of Homeland Security, ALERT Program. This work was also financially supported by the Regional Council of Le Centre and the French Industry Ministry within the Capthom project of the Competitiveness Pole S^2E^2.

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 where the type of abnormal activity or object is known *a priori*. Examples of such methods include facial recognition systems [19], restricted-area access detection [13], car counting [5], detection of people carrying cases [6], abandoned objects detection [17, 14], 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.

Nevertheless, such shape recognition methods require a list of objects or behavior patterns that are anomalous. Unfortunately, this is not always possible, especially where 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 [4, 7, 8], 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 [11, 12, 18, 9]. 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 Boiman and Irani's approach [3] which 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.*[1] 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.* [10] also 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 usually high amounts of activity at each pixel. Unfortunately, both methods, [1, 10] are only temporally sensitive and do not take into account spatial abnormalities.

The main focus of this paper is to propose a simple low-level method for learning patterns of activities. As opposed to path-based approaches, we do not rely on tracking and can cope with unstructured activity and crowded scenes. However, as opposed to Adam *et al.* and Jodoin *et al.*'s approach, we incorporate spatio-temporal dependencies between events and thus, can detect objects following suspicious paths.

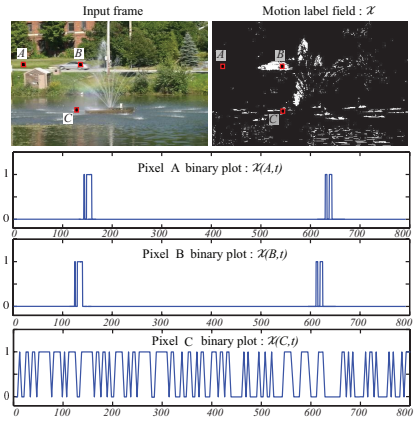


Figure 1. Binary signature for three pixels, two being highly correlated (A and B).

3. Context, Overview and Notations

3.1. Context

Let $I_{\vec{x},k}$ be the luminance (or color) of a video sequence spatially captured on a 2-D lattice of size $Q_0 \times R_0$ at discrete times k, i.e., $\vec{x} \in Q_0 \times R_0 \subset R^2$, $k \in Z^+$. Our atomic unit is the motion label, $X_{(\vec{x},k)} \in \{0, 1\}$, which is estimated through simple background subtraction, with 1 denoting moving object and 0 denoting static background. The temporal sequence of motion labels for each pixel is depicted 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 idle period of no activity. The entire spatio-temporal sequence can be alternatively defined over a 3D lattice $\mathcal{S}$ of size $Q_0 \times R_0 \times T_0$ with $s \in \mathcal{S}$ denoting a spatio-temporal location, I_s denoting the corresponding luminance (or color) and X_s the corresponding motion label.

Although many video analytics methods only use X_s in early stages of processing (mainly to locate moving objects) we argue in this paper that it nonetheless carries fundamental information on the content of the scene and thus,

can be used to perform high level tasks. Evidently, some have already shown this possibility by using it to summarize video [16], recognize human movements [2] and detect abnormally high activity [10].

In general, motion label sequences provides 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 label sequences at the pixel-level (or in general location). Consequently, abnormal behavior can be detected in a pixel-by-pixel manner 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.

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 label 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 right to left. To account for these situations we develop co-occurrence models as a function of location.

3.2. Co-occurrence Models

We develop co-occurrence models in this section. To this end, we consider for each pixel $\vec{x}$ at time t a spatio-temporal neighborhood centered at $s = (\vec{x}, t)$. The spatio-temporal neighborhood is a sub-video sequence $\mathcal{M}_s \subset \mathcal{S}$ centered at the spatio-temporal location, $s \in \mathcal{S}$ with size $Q \times R \times T$, $Q < Q_0$, $R < Q_0$ and $T << T_0$. These various quantities are all depicted in Fig. 2.

Consider a location $r = (\vec{y}, \tau) \in \mathcal{M}_s$ in the spatio-temporal neighborhood of $s = (\vec{x}, t)$. We say that two spatio-temporal neighbors r and s "co-occur" whenever their corresponding motion-labels are both active, namely, $X_s = 1$ and $X_r = 1$. The spatial neighborhood of a pixel, $\vec{x}$ is the set of all pixels $\vec{y}$ such that $s = (\vec{x}, t)$ and $r = (\vec{y}, t)$ are both in $\mathcal{M}_s$ for all t.

We compute the co-occurrences for each pixel $\vec{x}$ as follows. We consider each pixel $\vec{y}$ in the spatial neighborhood of $\vec{x}$. At each time t, we let $s = (\vec{x}, t)$ and $r = (\vec{y}, t + \tau)$ for a $\tau \in [-T/2, T/2]$. We count the number of times r co-occurs with s as t ranges over T_0. In this way, pairwise co-occurrences for neighboring pixels is derived. Note that the frequency of non-occurrence, namely, $X(s) = 1$ and $X(r) = 0$ can also be derived in a similar manner. Self-occurrences, namely, co-occurrences for $s = (\vec{x}, t)$ and its time-shift $r = (\vec{x}, t + \tau)$, reduces to characterizing activity counts for each individual pixel. The self-occurrences reduces to the pixel-by-pixel characterization with no spatial dependencies, which we described earlier.

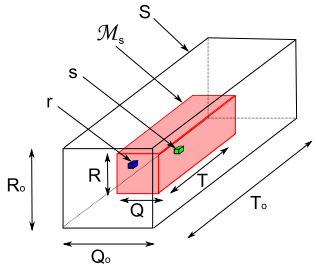


Figure 2. 3D lattice S with spatio-temporal neighborhood $\mathcal{M}_s$.

Note that two spatio-temporal neighbors co-occur 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 fact, whenever a moving object passes in front of $\vec{x}$ at time t, it leaves a spatio-temporal trace as some sites $r \in \mathcal{M}_s$ 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 spatial neighborhood of $\mathcal{M}_s$. Our method encapsulates such traces in terms of a co-occurrence frequency matrix, which accounts for the frequency with which two spatio-temporal neighbors co-occur. This matrix can serve as the basis for characterizing normal activity.

4. Our Method

In this section, we present how, for a given site s, a co-occurrence matrix and the associated statistical model can be estimated from a training video sequence. Our statistical model is a Markov-Random Field model that accounts for the likelihood of the co-occurrences. We later present how abnormal events can be detected and how connected graphs can be used to follow relevant moving objects.

4.1. Markov Random Field Models

Normal Model: Let O_s denote the set of observations in the spatio-temporal 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:

$$P_N(O_s) = \frac{1}{Z}\exp(\sum_{u,v\in\mathcal{M}_s}\alpha_{uv}\delta(X_u, X_v)) \quad (1)$$

where, $\delta(X_u, X_v) = 1$ if both X_u, $X_v = 1$ and zero otherwise. The α_{uv} are co-occurrence potentials which can be made to depend on pixel distances (for the remainder of the paper, α_{uv} will be refereed to as the co-occurrence matrix) . Note that the form of the probability expression favors co-occurrences over no co-occurrences. These potentials are estimated from training data. Z is the usual partition function, which is a normalization constant to ensure that the right hand side sums to one. Note that our MRF model is based on two-node cliques and the contributions from any two locations is zero unless two nodes co-occur. There is also a contribution from self-occurrences, α_{uu}.

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 [15]. 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 active pixels, i.e.,

$$N_1 = \sum_{u\in\mathcal{M}_s} X_u \quad (2)$$

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

$$P_A(O_s) = p^{N_1}(1-p)^{N_0-N_1} = \left(\frac{p}{1-p}\right)^{N_1}(1-p)^{N_0} \quad (3)$$

where, p is the probability that a site u is active (namely $X_u = 1$).

4.2. Training Phase: Learning Co-occurrence Matrix.

During the training phase, the co-occurrence matrix α_{uv} for two spatio-temporal locations, $u, v \in \mathcal{M}_s$ is empirically computed. Let,

$$s = (\vec{x}, t),\ u = (\vec{y}_1, t+\tau_1),\ v = (\vec{y}_2, t+\tau_2) \quad (4)$$

where $\tau_1, \tau_2 \in [-T/2,\ T/2]$. As mentioned previously, the co-occurence matrix (or potential) can be thought of as a summary of every trace (example of such trace is shown in Fig. 3(b) 4(b) and 5(b)) left by moving objects in the training sequence. We estimate the co-occurrence potentials as follows:

$$\alpha_{uv} = \frac{\beta_{uv}}{T_0 - T}\sum_{t=T/2}^{T_0-T/2}\delta(X_{(\vec{y}_1,t+\tau_1)}, X_{(\vec{y}_2,t+\tau_2)}) \quad (5)$$

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,

$$\alpha_{uv} = \alpha_{(\vec{y}_1,t+\tau_1),(\vec{y}_2,t+\tau_2)} = \alpha_{(\vec{y}_1,\tau_1),(\vec{y}_2,\tau_2)} \quad (6)$$

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,

$$\alpha_{uv} = 0,\ u \neq s,\ v \neq s \quad (7)$$

This is usually not a bad assumption if co-occurrence activity between the central pixel, s, and its neighborhood dominates other co-occurrences. 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 has been active simultaneously with v while translating $\mathcal{M}_s$.

An example of a simple co-occurence matrix is shown in Fig.4(a). From this figure, one can see that the moving objects passing in front of pixel $\vec{x}$ has an overall size of approximately 50×50 pixels and moves linearly from left to right at a pace of roughly 30 pixels per frame. Note that the co-occurrence matrix can be updated in time to account for changes in the behavior. This can be done by simply adding, in a linear fashion new traces O_s as they appear in a streaming video.

4.3. Observation Phase: Detecting Abnormalities

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 normal 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: normal vs. abnormal. The likelihood ratio $\ell(O_s)$ is the ratio of the probability of observations under normal and abnormal hypothesis, from Eq. 1, 3 and 7, it follows:

$$\ell(O_s) = \frac{P_N(O_s)}{P_A(O_s)} \tag{8}$$
$$= \frac{(1-p)^{N_0}}{Z} \exp\left(\sum_{r \in \mathcal{M}_s} \alpha_{sr} \delta(X_s, X_r) - \log \frac{p}{1-p} (\sum_{r \in \mathcal{M}_s} X_r) \right)$$

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 normal and abnormal hypothesis based on a global threshold η:

$$\ell(O_s) = \exp\left(\sum_{r \in \mathcal{M}_s} \alpha_{sr} \delta(X_s, X_r) - \tau \sum_{r \in \mathcal{M}_s} X_r \right) \underset{abnormal}{\overset{normal}{\gtrless}} \eta$$

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:

$$\frac{\sum_{r \in \mathcal{M}_s} \alpha_{sr} \delta(X_s, X_r)}{\sum_{r \in \mathcal{M}_s} X_r} \underset{abnormal}{\overset{normal}{\gtrless}} \tau. \tag{9}$$

Note that, following of Eq.7, this test is only performed when $X_s = 1$.

4.4. Dealing with multiple moving objects

The test of Eq. (9) allows one to determine which observation O_s is normal and which one is not according to the co-occurrence matrix 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. Rather than do this at every time step one could first compute ratio of the intersection and union of past and current observations $O_s = O_{\vec{x},t}$ and $A_{s'} = O_{\vec{x},t-1}$. Here we have used the symbol $A_{s'}$ to denote that an abnormality has been detected in the previous instant. Our ratio amounts to:

$$\varepsilon = \frac{\sum_{r \in \mathcal{M}_s} (O_s(r) \ \wedge \ A_{s'}(r))}{\sum_{r \in \mathcal{M}_s} (O_s(r) \ \vee \ A_{s'}(r))} \tag{10}$$

Threshold 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 $A_{s'}$. If $\varepsilon < \gamma$, where γ is some threshold, we can then conduct a LRT on the innovation $O'_s = O_s - A_{s'}$, where O'_s 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 7(c), 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 outdoor sequences representing different challenges. For each sequence, a co-occurence matrix of size ranging between $130 \times 70 \times 300$ and $210 \times 210 \times 150$ have been used. The number of frames T used to estimate P_N (Eq. (1)) varies between 2000 and 7000 (*i.e.* from 1 and 4 minutes of video) depending on the sequence. Note that results are presented in thumbnails of Fig. 6 and 7; 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-occurence

matrix (Eq. (9)).

The first example (see Fig. 6) is one which shows normal traffic and cars making illegal U-turns. As shown in Fig.3, the trace left by the U-turn significantly differs from the usual traffic flow. We can also observed that the co-occurrence matrix contains information about the regular traffic flow but also activities generated by pedestrians crossing the street. Cars following the regular path are tagged in green and cars making an illegal U-turn are tagged in red.

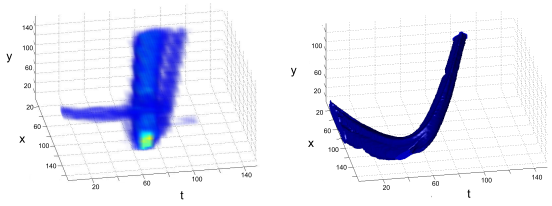
The second example is a boulevard with pixel $\vec{x}$ located on the side of the street where cars go from left to right. As shown in Fig. 4 (a), the co-occurrence matrix learned at that location clearly underscores the strong unimodality of motion at that position. For this scene, every car running from left to right at the average traffic speed is flagged as being normal (see the green car in Fig. 7 (a)). But after a while, a pedestrian shows up on the street, walking from right to left. Since the trace left by the pedestrian is significantly different than the co-occurrence matrix, the pedestrian is identified as an unusual moving object (see the red person in Fig. 7 (a)).

The third example shows a fountain on which two key pixels have been placed (see Fig. 7 (b)). The co-occurrence matrix of both pixels contains the average trace left by the fountain. Here, the motion induced by the fountain is considered as normal activity whereas any other activity such as a car or a pedestrian is seen as being different from the usual behavior. This example clearly underscores the fact that our approach can account for a large variety of motion, including those which cannot be handled by traditional tracking.

The last 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. 7 (c)). Interestingly, following the process described in section 4.4, all pedestrians walking across the bag are not recognized as being suspicious (see the green pedestrian and the red bag in Fig.7(c)).

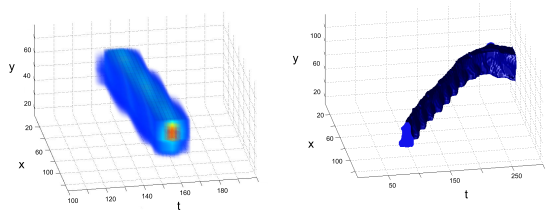
Note that the connected graph processing presented in Section 4.4 allows our method to track suspicious moving objects, even after the suspicious event occurred. This is shown for the person dropping the bag (Fig. 7 (c)) as well as the car making an illegal U-turn (Fig. 6).

Quantitative results illustrating the performance of our method are shown in table 1. We compare our method with an object-based one [4] in which blobs path is analyzed. As one can see, our method is able to detect most abnormal events. Note that every false positive was caused by objects whose size strongly differed from that of the majority.



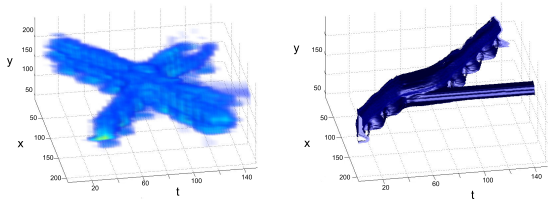
(a) (b)

Figure 3. (a) co-occurence matrix of regular traffic flow and pedestrians crossing the street and (b) the trace left by making an illegal u-turn.



(a) (b)

Figure 4. (a) Co-occurence histogram of regular traffic flow and (b) the trace left by a pedestrian walking against the traffic.



(a) (b)

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

A good example is for the boulevard sequence (Fig. 7 (a)) in which huge trucks and buses seldom pass by. Because of their size, these moving objects leave huge traces which differs significantly from the learned co-occurrence matrix. Since these objects are suspicious by their size and not their dynamics, simple heuristics can eliminate such false positives if needed. Moreover, since we only have to compare two 3D matrices and update one of this matrix, with a C++ implementation of our method, we are able to reach real-time performance (about 19 frames per second for a $210 \times 210 \times 150$ co-occurrence matrix).

6. Conclusion

We proposed in this paper a method to perform behavior characterization based on the spatial and temporal dependencies between motion labels obtained with simple background subtraction. To do this, we built an MRF model

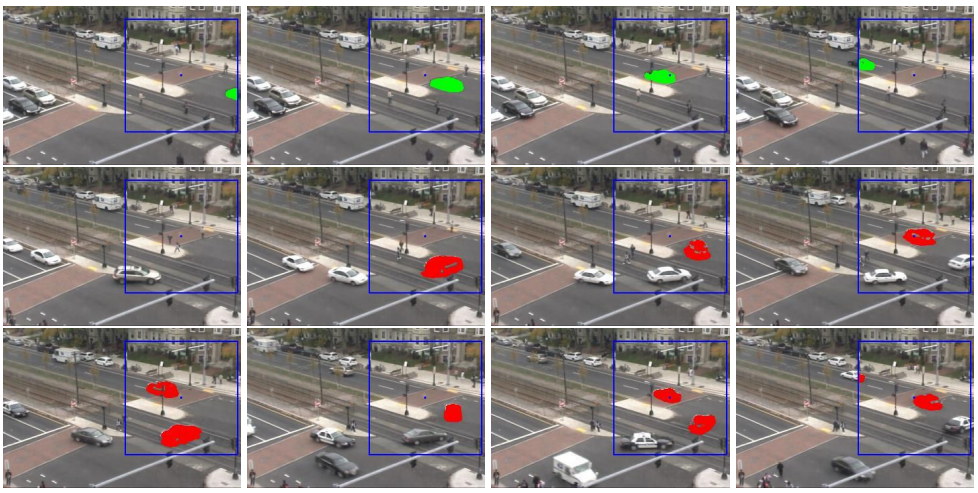


Figure 6. Example video in which cars following the regular traffic flow are tagged in green while car making an illegal U-turn have been picked up by our algorithm and tagged in red.

	Nor.	Abn.
Nor.	83.9	16.1
Abn.	9.1	90.9

(a)

	Nor.	Abn.
Nor.	91.25	8.75
Abn.	0	100

(b)

Table 1. Confusion matrices obtained with three different videos showing around eighty normal events and a dozen of abnormal events. Horizontal rows are ground truths and vertical columns are observations. (a) an object-based method [4] and (b) our method .

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) $\vec{x}$. Equipped with the co-occurrence matrix, we can detect abnormal events by detecting traces which significantly differ from our normal model following a likelihood ratio test.

We tested our approach over different videos containing various normal and abnormal activities (*i*.e. cars making illegal U-turns, abandoned baggage . . .).

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE PAMI*, 30(3):555–560, 2008.

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 23(3):257–267, 2001.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. *Int. Jour. of Compt. Vis.*, 74(1):17–31, 2007.

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

[5] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *UAI*, pages 175–181, 1997.

[6] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):809–830, 2000.

[7] H.Boxton. Learning and understanding dynamic scene activity: A review. *Image Vis. Comput.*, 23, 2003.

[8] W. Hu, T. Tab, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Sys. Man and Cyb. – Part C: App. and Reviews*, 34(3):334–352, 2004.

[9] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1450–1464, 2006.

[10] P.-M. Jodoin, J.Konrad, and V. Saligrama. Modeling background activity for behavior subtraction. In *proc. of IEEE ICDSC*, 2008.

[11] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Imag. and Vis. Compt.*, 14(8):609–615, 1996.

[12] I. Junejo, O. Javed, and M. Shah. Multi feature path modeling for video surveillance. In *proc. of ICPR '04*, pages 716–719, 2004.

[13] J. Konrad. Motion detection and estimation. In A. Bovik, editor, *Handbook of Image and Video Processing, 2nd Edition*, chapter 3.10, pages 253–274. Academic Press, 2005.

[14] S.-N. Lim, H. Fujiyoshi, and R. Patil. A one-threshold algorithm for detecting abandoned packages under severe occlusions using a single camera. Technical Report CS-TR-4784,, University of Maryland, 2006.

[15] W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.*, 1997.

[16] Y. Pritch, A. Rav-Acha, and S. Peleg. Non-chronological video synopsis and indexing. *IEEE PAMI*, 30(11):1971–1984, 2008.

[17] K. Smith, P. Quelhas, and D. Gatica-Perez. Detecting abandoned luggage items in a public space. In *IEEE Performance Evaluation of Tracking and Surveillance Workshop (PETS)*, pages 75–82, 2006.

[18] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):747–757, 2000.

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

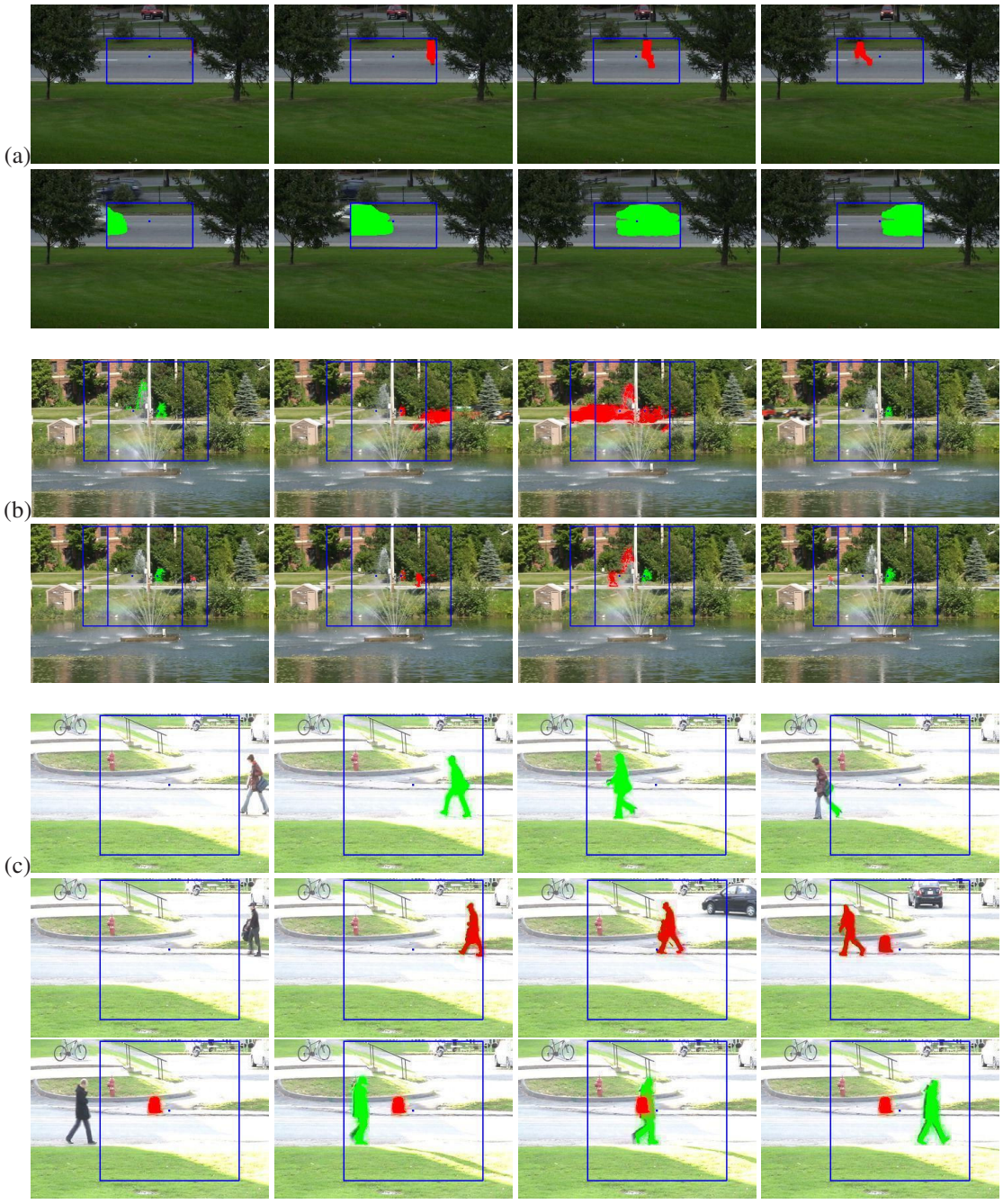


Figure 7. Three videos in which normal moving objects (in green) and abnormal events (in red) have been picked up by our algorithm.