Finding and Indexing of Eccentric Events in Video Emanates

Article *in* Journal of Multimedia · February 2010

DOI: 10.4304/jmm.5.1.22-35 · Source: DOAJ

3 authors:

Md. Haidar Sharif
University of Hail
60 PUBLICATIONS 518 CITATIONS
SEE PROFILE

Nacim Ihaddadene
Junia
37 PUBLICATIONS 714 CITATIONS
SEE PROFILE

Chaabane Djeraba
Université de Lille
273 PUBLICATIONS 3,716 CITATIONS
SEE PROFILE

Finding and Indexing of Eccentric Events in Video Emanates

Md. Haidar Sharif, Nacim Ihaddadene, Chaabane Djeraba
Computer Science Laboratory of Lille (LIFL)
University of Sciences and Technologies of Lille (USTL), France
Email: {md-haidar.sharif, nacim.ihaddadene, chabane.djeraba}@lifl.fr

***Abstract*— We propose a methodology that first extracts features of video emanates and detects eccentric events in a crowded environment. Afterwards eccentric events are indexed for retrieval. The motivation of the framework is the discrimination of features which are independent from the application domains. Low-level as well as mid-level features are generic and independent of the type of abnormality. High-level features are dependent and used to detect eccentric events, whereas both mid-level and high-level features are run through the indexing scheme for retrieval. To demonstrate the interest of the methodology, we primarily present the obtained results on collapsing events in real videos amassed by single camera installed an airport escalator exits to monitor the circumstances of those regions.**

***Index Terms*— Motion heat map, Entropy, Indexing, Search**

I. Introduction

The importance of detecting and retrieving eccentric (abnormal) events considerably increased recently, especially for security applications. The need of security teams to access informative data are important so that the safeguard team can fully cognisant and sedulous for recognizing perilous and inconsistent conditions and circumstances. We notice the explosion of the video surveillance systems in various environments such as airports, stations, banks, sporting events, shopping malls, parking places, hospitals, hotels, town centers, or other private and public spaces. Our objective is to develop a system that detects, indexes, searches, and more generally mines the video surveillance data to answer to the needs of security measures. For examples, a security team asks for video segments where a certain event happened (e.g., collapsing) during a week. Or the reassurance team asks questions, e.g., (i) What is the average frequency of abnormal events, during a day or a particular part of a day, considering all video emanates (streams)? (ii) What is the relationship between the crowd density and the abnormal event? (iii) What are the video streams with frequent abnormal events? (iv) What are the videos those are the most similar to the segment of video with specific abnormal event? (v) and so forth.

In this context, the first contribution of our work is to detect abnormal events in video streams. The contribution is articulated on a framework (see Figure 1) operating in three-level of features namely low-level, mid-level, and high-level. The primary goal of our framework is the discrimination of features, which are independent of the application domains from those which are dependent. Thereupon we use an approach, for which certain features are generic and independent of any application domain. As a consequence, it will be possible to reuse those features, and implement new ones. The features of the low and mid levels are computed as follows. Firstly, the approach starts by calculating the motion heat map during a period of time to extract the main regions of motion activity. The use of heat map improves the quality of the results and reduces the space of processing. Secondly, points of interest are extracted in the hot regions of the scene. Thirdly, optical flow is computed on these points of interest, marked off the boundaries by the hot areas of the scene. The optical flow information from video presents the crowd multi-modal behaviors as optical flow patterns make different in time. There is sufficient agitation in the optical flow pattern in the crowd in case of abnormal and/or emergencies situations [as compared to Fig. 3 (b) & (c)]. Finally, by reusing the features of low-level as well as mid-level; the high-level features, variations of motions, are estimated to make a distinction in favor of potential abnormal events. In the context of our need, we define specific abnormalities which are in relation with collapsing and panic-stricken.

Another contribution of our work concerns the indexing and search of abnormal events in video surveillance data. We present a representation model composed of units of search. The units of search, also called unit of representations u_r (where u_r is a vector that describes the properties of an abnormal event), are composed of high and mid levels of search (e.g., density, trajectory, time code of the video segment) plus annotations. Associations between features and annotations are extracted and saved when confidence is high. The confidence of associations is based on the combination of two measures, the conditional probability (probability to have the feature A when we have the feature B) and the intensity of implications (the degree of surprising to have the feature A and the feature B together). The intensity of implication is interesting for certain reasons, e.g., it is sensitive to the number of instances those validate the association. The indexing supports both search by content and keywords.

We experimented the proposed methodology mainly on a video surveillance data set. For the data set we predominantly concentrated on escalator videos, collected

doi:10.4304/jmm.5.1.22-35

by a single camera data-set placed in the escalator exits in an airport, to use in our applications. One practical application of our methodology is in the detection and indexing of real-time collapsing events, which could lead to potentially full of danger or risk situations in exit escalators. The videos used are from camera installed at the airport to monitor the situation of escalator exits. The impression of the application is to have exit escalators continuously perceived to respond efficiently in the event of any collapsing (breakdown). With this aim, cameras are installed in front of the exit locations to view and dispatch the video signal to a control room. In the control room consecrated employees can oversee, take control, and react favorably to a breakdown under the circumstances of anomalous and/or emergency.

In the remaining parts of the paper, we will start by introducing some related works in Section 2. In Section 3, we will confer our proposed frame work. In Section 4, we will go over low-level features. In Section 5, we will deliberate mid-level features. In Section 6, we will argue high-level features. Section 7 talks over indexing architecture. Section 8 reports the experimental results. Section 9 discusses a discussion with some clues for future work directions. Finally, Section 10 makes a conclusion.

II. Related Works

In the approaches of the state-of-the-art, various features have been proposed depending of the target abnormal events they deal with. Furthermore, learning, adaptive, and incremental processes have been studied. Notwithstanding, they did not discriminate clearly the three-level of features (low, mid, high). Also features are represented in specific ways so the projections in other applications need much adaptation works. We will focus with some details on the related works, where we have three-category: first category estimates crowd density, the second category engages the attention of abnormal event detection in crowd flows, and the third category concerns indexing of abnormal events.

Estimation of crowd density: The methods applied [1]–[4] are based on textures and motion area ratio and make an interesting static analysis for crowd surveillance, otherwise than detect abnormal situations. There exist some optical flow based methods [5], [6] to detect stationary crowds, or tracking methods by using multi-camera [7].

Detection of eccentric events: These works detect irregular events in crowd flows. The general approach consists of modeling normal behaviors, and then estimating the deviant behavior or attitudes between the normal behavior model and the observed behaviors. Those deviations are labeled as abnormal. The determining characteristic of the general approach is to exploit the fact that data of standard behaviors are generally available, and data of abnormal behaviors are reasonably less available. For this reason, the divergences from examples of standard behaviors are used to distinguish as the characteristic of abnormality. In this category, authors in [8] and [9] encoded optical flow features with Hidden Markov Models (HMM). The aim of both of their methods is to detect emergency or abnormal events in the crowd. The methods were experimented with simulated data only. Authors in [10] proposed a mathematical method based on Lagrangian particle dynamics for crowd flow segmentation and flow instability detection. The method is efficient for segmentation of high density crowd flows e.g., religious festivals, parades, concerts, football matches, etc. However, the method would not be so interest-bearing in the context of a crowd scene like airport, shopping malls, and so forth to detect abnormalities; more precisely, in case the of escalators where the density of people is not so high. In the same category, but for low crowded scenes, authors in [11] proposed a method which includes robust tracking, based on probabilistic method for background subtraction. The robust tracking method is not adapted to crowd scene, in which it is so multifarious that it cannot track objects. Authors in [12] formulated the problem of detecting regularities and irregularities as the problem of composing (explaining) the new observed visual data (an image or a video sequence, referred to below as query) using spatiotemporal patches extracted from previous visual examples (the database). Though the method is arousing the attention, it expects some degree of learning process and/or needs training data. Other works which are related to the same area of detecting abnormal events by incremental approaches [13]–[15].

Indexing of eccentric events: The state-of-the-art related to indexing of abnormal events is relatively weak (e.g., [16]), compared to works on abnormal event detection or crowd flow analysis. Authors in [16] proposed a semantic based retrieval framework for traffic video sequences. To estimate the low-level motion data, a cluster tracking algorithm is developed. A hierarchical self-organizing map is applied to learn the activity patterns. By using activity pattern analysis and semantic concepts assignment, a set of activity models (to represent the general and abnormal behaviors of vehicles) are generated, which are used as the indexing-key for accessing video clips and individual vehicles in the semantic level. The proposed retrieval framework supports various queries including query by keywords, query by sketch and multiple object queries. Authors in [17] developed a framework of human motion tracking and event detection, destined for indexing. Yet, no direct contribution to indexing has been realized.

Our work bears reference to the following specificities:

- It considers simultaneously density (motion area ratio) and direction. The previous works methods do not take into account those factors together.
- It detects all events where the motion variations are important as compared to previous events, with no restrictions of the number of people.
- It uses a mask to define a region of interest. This mask is obtained from the hot areas of the motion heat map image. Consequently, it is useless to analyze the whole scene particularly where there are few motion intensities or no motions. As a result, the approach is interesting for real-time applications.

- It concerns the framework specification that contains three-level of features namely low, mid, and high. Thus, it will make the framework general and adaptable to various situations, mainly in the first and second levels of features. In the second level of features, we compute a number of features to cover the large band of applications.
- It concerns the indexing and search of video segments with abnormal events. We support search by keywords, and by content (similarity).

III. Proposed Methodology

The hypothesis of our methodology is to consider the detection of abnormal events in a crowded context of various densities in video surveillance data. And the framework followed does not consider individual person tracking and consider the study of the general motion aspect and more particularly assesses sudden shift and strange locomotion discrepancy of a set of interest points discovered by Harris point of interest detector, instead of tracking persons one by one. The detection and tracking of individual persons are difficult in the case of crowded situations. The framework is composed of several steps:

1) The motion heat map is extracted. The heat map represents the motion intensities. For example, hot area corresponds to high motion intensities, cold areas represent to low motion intensities, etc. It is worth mentioning that the motion heat map could be calculated either online or off-line.
2) Harris points of interest are extracted in the hot regions of the scene. In the simplest case, it is applied in well limited areas. We consider a binary heatmap, white (movement), and black (no movement). In this case, the points of interest are applied into white regions. Furthermore, blobs are extracted.
3) Optical flows are computed on those points of interest, marked off the boundaries by the hot areas of the scene.
4) Mid-level features, the statistical scrutiny of the optical flow information, e.g., density, coefficient of direction variation, coefficient of distance variation, and direction histogram, are computed.
5) On the basis of mid-level features computed in the previous step, we define high-level features (e.g., entropy) which classify events in abnormal/normal and return different types of abnormality.

The 1-4 steps are generic and do not depend of a specific application domain. They concern the extraction of low and mid level of features. The fifth step is dependent of the application domain and requires a specific learning process. The flow diagram of our planned work has been depicted in Fig. 1, which is articulated on a framework operating in three-level of features as noted below.

Low-level: It bears reference to measurements those extracted directly from the signal (visual data), e.g., point of interests, region of interests (blobs), edges, ridges, optical flow, etc. We use mixture of Gaussian to detect foregrounds in case of low crowded (low density) areas, and optical flows on points of interest for high crowded (high density) areas.

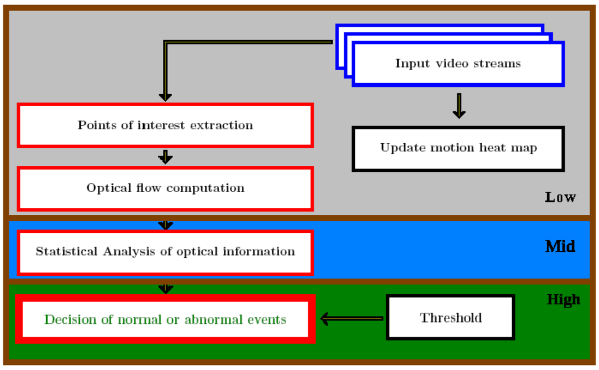

Input video streams
Points of interest extraction
Update motion heat map
Optical flow computation
Low
Statistical Analysis of optical information
Mid
High
Decision of normal or abnormal events
Threshold

Figure 1. Simple block diagram of the proposed framework.

Mid-level: It concerns of features those are generated after a learning process directly from the low-level features, and helps more to enhance the upper level features (semantics), e.g., crowd density (ratio of the blobs in the scene), trajectory, velocity, direction, acceleration, energy, and so forth. The mid-level features are computed on low-level features (e.g., interest regions, interest points) and are stored in basic structures.

High-level: It pertains to the features with more semantics than mid-level, and they are enough to take decision. Here we are in the possession of the normal/abnormal events.

IV. Low-level features

A. Motion heat map

A heat map is a graphical representation of data where the values taken by a variable in a two-dimensional ($2D$) map are represented as colors. Motion heat map is a $2D$ histogram expressing briefly the most important regions of motion activity. This histogram is built from the accumulation of binary blobs of moving objects, which were extracted following background subtraction method [18]. Assume symbols H and I indicate $heat\ map$ and $intensity$ respectively. The H is defined as:

$$H_n(i,j) = H_{n-1}(i,j) + I_n(i,j) \qquad (1)$$

$$H_0 = I_0(i,j) \qquad (2)$$

where n is the frame number ($n \geq 1$), and i & j are the coordinates (line and column) of the pixel (i,j) of a frame. The obtained map is used as a mask to define the $region\ of\ interest$ for the next step of the method. Figure 2 makes noticeable an occurrence of the obtained heat map from an escalator camera view. The use of heat map ameliorates the quality of the results and reduces processing time which is an important factor for real-time applications. The results are more significant when the video duration is long. In practice, even for the same place, the properties of abnormal events may vary depending on the context (day-night, indoor-outdoor, normal-peak time, occasion, vacation, etc.). We build a motion heat map for each set of conditions. It is not necessary to consider in detail the whole scene, and fastidiously the scene where there are few motion intensities or no motions. Thus, the approach directs the attention on the processing of specific regions where the density of motions is high. The threshold related to the density elevation is a contextual information.

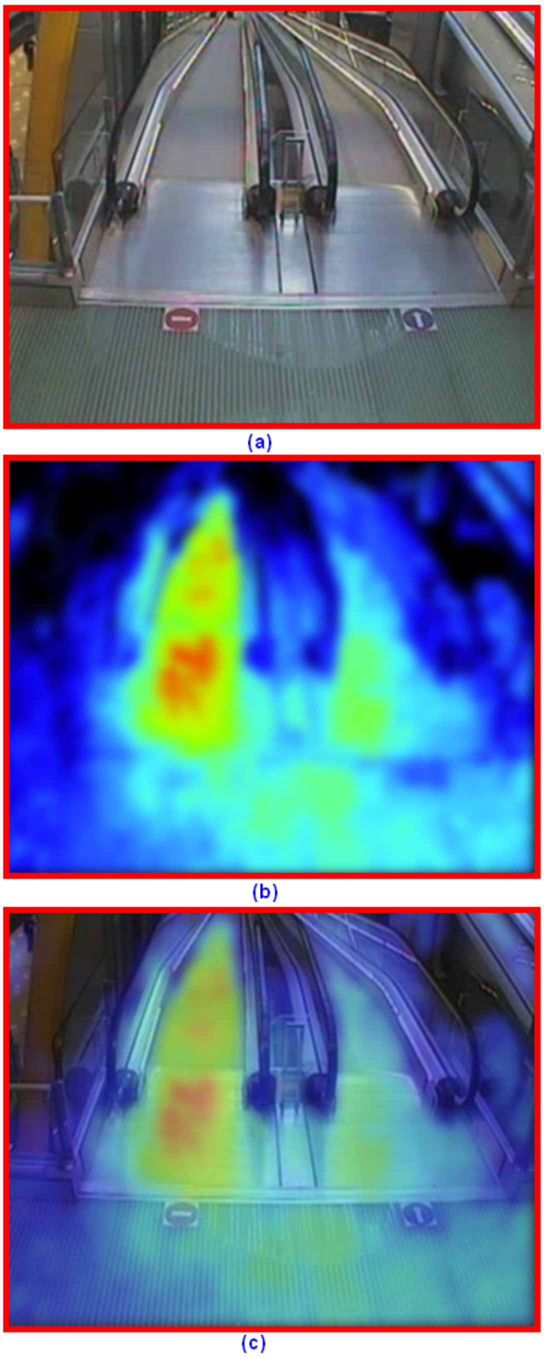


Figure 2. Motion heat map: (a) camera view, (b) generated motion heat map, and (c) masked view which recommends region of interest.

B. Points of interest extraction

Moravec's corner detector [19] is a relatively simple algorithm that was used by Moravec and others, but is now commonly considered out-of-date. It is not rotationally invariant (a property prevalent even in more modern corner detectors) as the response is not invariant with respect to direction (anisotropic), is considered to have a noise response, and is susceptible to reporting false corners along edges and at isolated pixels so is sensitive to noise. Nevertheless, it is computationally efficient which was critical for Moravec as he was interested in a real-time application and had minimal computational power at his disposal. The other way around the Harris corner detector [20] is computationally demanding, but directly addresses many of the limitations of the Moravec corner detector. In our approach, we consider Harris corner as a point of interest. The Harris corner detector is a famous point of interest detector due to its strong invariance to rotation, scale, illumination variation, and image noise [21]. It is based on the local auto-correlation function of a signal, where the local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions. Assume a point (x,y) and a shift (Δx,Δy), then the local auto-correlation function is defined as:

$$c(x,y) = \Sigma_w[I(x_i,y_i) - I(x_i+\Delta x, y_i+\Delta y)]^2 \quad (3)$$

where $I(.,.)$ denotes the image function and (x_i, y_i) are the points in the smooth circular window w centered on (x, y). The shifted image is approximated by a Taylor expansion truncated to the first order terms as:

$$I(X^i_\delta, Y^i_\delta) \approx I(x_i,y_i) + [I_x(x_i,y_i)I_y(x_i,y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (4)$$

where $X^i_\delta = x_i+\Delta x$, $Y^i_\delta = y_i+\Delta y$; and $I_x(.,.)$ & $I_y(.,.)$ denote the partial derivatives in x & y, respectively. Substituting the right hand site of Eq. 4 into Eq. 3 yields:

$$c(x,y) = \Sigma_w(\ [I(x_i,y_i)I_y(x_i,y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix})^2 \quad (5)$$

$$= [\Delta x \Delta y]M(x,y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (6)$$

where $M(x,y) =$

$$\begin{pmatrix} \Sigma_w(I_x(x_i,y_i))^2 & \Sigma_w I_x(x_i,y_i)I_y(x_i,y_i) \\ \Sigma_w I_x(x_i,y_i)I_y(x_i,y_i) & \Sigma_w(I_y(x_i,y_i))^2 \end{pmatrix}. \quad (7)$$

The 2×2 symmetric matrix $M(x,y)$ captures the intensity structure of the local neighborhood. Let λ_1 and λ_2 are the eigenvalues of matrix $M(x,y)$. The eigenvalues form a rotationally invariant description. There are three cases to be considered [20]:

- **No point of interest is found**: If both λ_1 & λ_2 are small, so that the local auto-correlation function is flat, i.e., little change in $c(x,y)$ in any direction, then the windowed image region is of approximately constant intensity.
- **An edge is found**: If one eigenvalue is high and the other is low, so the local auto-correlation function is

rigid shaped, then only shifts along the ridge (i.e., along the edge) cause little change in $c(x, y)$ and significant change in the orthogonal direction.
- **A point of interest is found**: If both λ_1 & λ_2 are high, so the local auto-correlation function is sharply peaked, then shifts in any direction result in a significant increase in $c(x, y)$.

Figure 3 (a) lets on an example of Harris corner detector. We deem that in video surveillance scenes, camera positions and lighting conditions admit to get a large number of corner features that can be easily captured and tracked.

C. Estimation of optical flow

The goal of optical flow technique is to compute an approximation to the 2D motion field, a projection of the 3D velocities of surface points onto the imaging surface, from spatiotemporal patterns of images intensity [22], [23]. To calculate the optical flow between successive video frames the well known combination of feature selection as introduced by Shi and Tomasi [24] and the algorithm of Lucas and Kanade for feature tracking [25] is used. Feature selection finds image blocks which are believed to allow the exact estimation of optical flow translation vector. The Shi-Tomasi algorithm makes use of the smallest eigenvalues of an image block as criterion to ensure the selection of features which can be tracked reliably by the Lucas-Kanade tracking algorithm. This algorithm matches the selected image blocks with blocks in the next frame using an efficient gradient descent technique. A pyramidal implementation of this algorithm is used to deal with larger feature displacements by avoiding local minima in a coarse to fine approach [26]. This combination has proven to allow fast and reliable computation of optical flow information [27]. In our optical flow computation step, we use a pyramidal implementation of this algorithm. Once we define the points of interest (features), we track those features over the next frames using the above combination feature tracker of Kanade-Lucas-Shi-Tomasi. On matching points of interest between frames, the result is a set of vectors:

$$V = \{V_1...V_N | V_i = (X_i, Y_i, D_i, \theta_i)\}$$

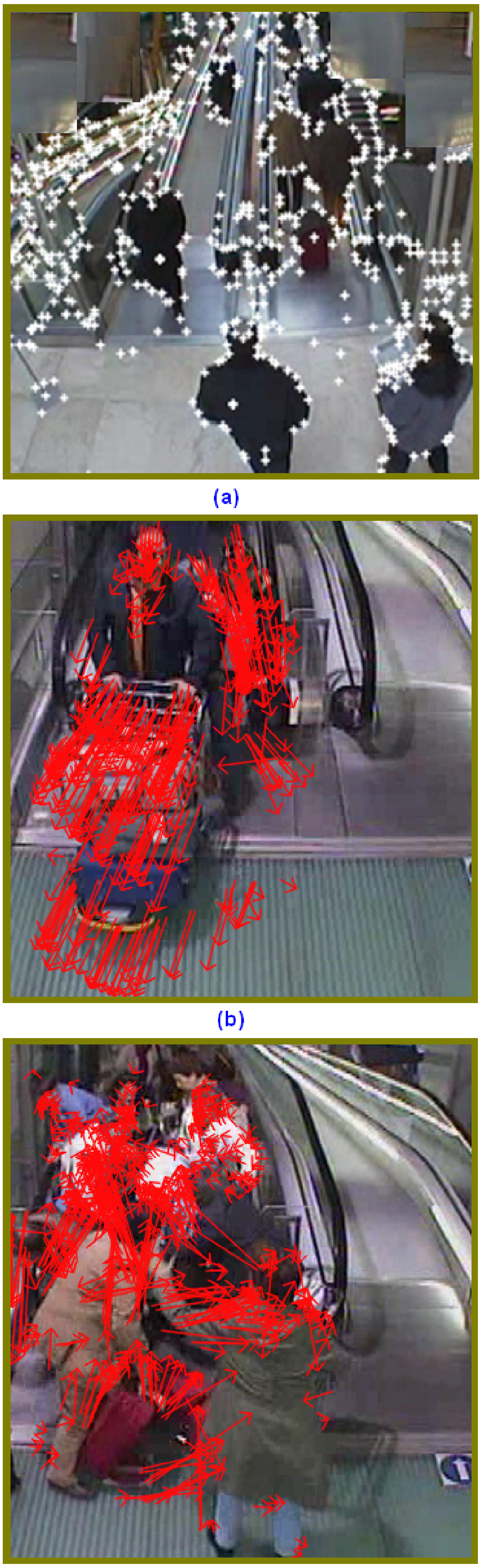
where

- $X_i \Rightarrow X$ coordinate of any feature i,
- $Y_i \Rightarrow Y$ coordinate of the feature i,
- $D_i \Rightarrow$ distance between the feature i in the frame f and its matched feature in frame $f + 1$,
- $\theta_i \Rightarrow$ direction of motion of the feature i.

Images in figure 3 (b) and (c) give evidence of the set of vectors obtained by optical flow feature tracking in two different situations. The image in 3 (b) divulges an orderly vector flow. The image in 3 (c) substantiates a littered vector flow due to the breakdown situation.

D. Estimation of displacement & direction of a feature

Figure 4 illustrates the feature i in the frame f with its coordinate $P(X_i, Y_i)$ and its matched in the frame $f + 1$


(a)

(b)

(c)

Figure 3. White points in (a) portray Harris point of interest. Red arrows are optical vector flows: (b) and (c) limn standard and aberrant situations in precisely the given order.

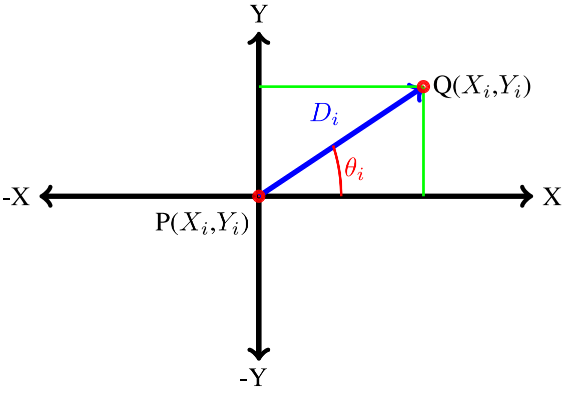


Figure 4. Cardinal point of the state of change of any feature i.

with coordinate $Q(X_i, Y_i)$. We can easily calculate the displacement or Euclidean distance of these two points using Euclidean metric as:

$$D_i = \sqrt{(Q_{X_i} - P_{X_i})^2 + (Q_{Y_i} - P_{Y_i})^2}. \qquad (8)$$

The direction of motion (θ_m) of the feature i can be calculated using the following trigonometric function:

$$\theta_m = atan(\frac{y_i}{x_i}) \qquad (9)$$

where $x_i = Q_{X_i} - P_{X_i}$ and $y_i = Q_{Y_i} - P_{Y_i}$. But there are several potential problems if we have a desire to calculate motion direction using Eq. 9, for instances:

Firstly: Eq. 9 does not show expected performance for a complete range of angles from $0\,°$ to $360\,°$. Only angles between $-90\,°$ and $+90\,°$ will be returned, other angles will be (say $180\,°$) out-of-phase. For example, let us consider two defined points $(x_1 = 3, y_1 = 3)$ and $(x_2 = -3, y_2 = -3)$. Using the Eq. 9, the point $(x_2 = -3, y_2 = -3)$ will produce the same angle as the point $(x_1 = 3, y_1 = 3)$ will do, but from Figure 4 we could consider that these are not in same quadrant.

Secondly: Points on the vertical axis have $x_i = 0$, hence, if we wish to calculate y_i/x_i we will get ∞ which will generate an exception when calculated on the computer.

To avoid these problems, we apply the $atan2(y_i, x_i)$ function which takes both x_i and y_i as arguments. Henceforth, the accurate direction of motion θ_i, where $\theta_i = atan2(y_i, x_i)$, of the feature i can be calculated as:

$$\theta_i = \begin{cases} \phi.\mathbf{sign(y_i)} & if\ x_i > 0, y_i \neq 0 \\ \mathbf{0} & if\ \ x_i > 0, y_i = 0 \\ \frac{\pi}{2}.\mathbf{sign(y_i)} & if\ x_i = 0, y_i \neq 0 \\ \mathbf{undefined} & if\ x_i = 0, y_i = 0 \\ (\pi - \phi).\mathbf{sign(y_i)} & if\ x_i < 0, y_i \neq 0 \\ \pi & if\ \ x_i < 0, y_i = 0 \end{cases}$$

where ϕ is the angle in $[0, \pi/2]$ such that $\phi = atan(|\frac{y_i}{x_i}|)$. The sign function $sign(y_i)$ can be defined as:

$$sign(y_i) \text{ = } \begin{cases} -1 & if\ \ y_i < 0 \\ 0 & if\ \ y_i = 0 \\ 1 & if\ \ y_i > 0. \end{cases}$$

As a consequence, the function $atan2(y, x)$ gracefully handles infinite slope, and places the angle in the correct quadrant [for instance: $atan2(0, 3) = 0$, $atan2(0, -3) = \pi$, $atan2(3, 3) = \pi/4$, $atan2(-3, -3) = -3\pi/4$, etc.].

V. Mid-level Features

In this step, we define some mid-level features those will be necessary to induce a specific abnormal event.

1) Motion area ratio M_R: In each video frame, the M_R estimates the ratio between the number of blocks containing motion and the total number of defined blocks. In crowded scenes the area covered by the moving blobs is important as compared to uncrowded scenes. We use this measure as a density estimator. To estimate M_R, we divide each video frame into $N \times M$ blocks, where N & M are number of columns & rows respectively. For any block (i, j), we define the moving block by means of:

$$movingblock(i, j) = \begin{cases} 1; & if\ movement\ \ exists \\ 0; & otherwise \end{cases}$$

If there are several movements exist in one block, then that block will be enumerated as one moving block. We count out the total number of moving blocks to define M_R as:

$$M_R = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} movingblock(i, j)}{N \times M}. \qquad (10)$$

2) Direction variance-mean ratio θ_R: To estimate direction variance (θ_V), it is important to estimate the mean direction $\overline{\theta}$ of the optical flow vectors in each video frame. The $\overline{\theta}$ is determined by:

$$\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \qquad (11)$$

where n is cardinality of the optical flow vectors in the frame and $360\,° \geq \theta_i > 0\,°$. Having firsthand knowledge of $\overline{\theta}$, we calculate the θ_V of those vectors as:

$$\theta_V = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\theta_i^2 - \frac{n}{n-1}\overline{\theta}^2. \qquad (12)$$

The direction variance-to-mean ratio (coefficient of direction variation) is defined as the ratio of the variance to the mean:

$$\theta_R = \frac{\theta_V}{\overline{\theta}}. \qquad (13)$$

3) Direction histogram θ_H: The θ_H gives directions to the direction tendencies and the number of peaks. In the histogram each column puts an address on the number of vectors in a given angle. The θ_H, which is affiliated to the frame, can be clearly characterized by the way of:

$$\theta_H = \{\theta_H(\theta_i), i = 1 \ldots s\} \qquad (14)$$

$$\theta_H(\theta_i) = \frac{\sum_{i=1}^{n} angle(i)}{s} \qquad (15)$$

$$angle(i) \text{ = } \begin{cases} 1, & if\ \ angle(i) = \theta_i \\ 0, & otherwise \end{cases}$$

where $\theta_H(\theta_i)$ is the normalized frequency of optical flow vectors those have the same angle θ_i. The θ_H is a vector of size s where s is the total number of angles considering the angle range between $-\pi$ and $+\pi$.

4) Distance variance-mean ratio D_V: Observation shows that distance variance (D_V) increases in abnormal situations. With one or many people walking even in different directions, they tend to have the same speed, which means a small value of the motion distance variance. But in case of abnormal observable activities (e.g., collapsing situations, a sudden overwhelming fear, escape circumstances, etc.) those often give rise to a big value for the D_V. The mean of distance variance $\overline{D}$ is clearly delimited by:

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{16}$$

where n is the number of optical flow vectors in the frame. Having $\overline{D}$ it is easy to ascertain D_V by:

$$D_V = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = \frac{1}{n-1}\sum_{i=1}^{n} D_i^2 - \frac{n}{n-1}\overline{D}^2. \tag{17}$$

The distance variance-to-mean ratio (coefficient of distance variation) is defined as the ratio of the variance to the mean:

$$D_R = \frac{D_V}{\overline{D}}, \quad where\ \overline{D} > 0. \tag{18}$$

VI. High-level features

High-level features concern the decision of event which is either normal or abnormal. These features are denoted as entropies. We developed a function entitled $Entropy$ that extracts the features at the frame f. The f is not explicated in the formula to keep the presentation simple.

A. Entropy estimation

The enumerated function $Entropy$, that depends on motion area ratio, coefficient of direction variation, coefficient of distance variation, and direction histogram characteristics at any frame f, is formulated as:

$$\{Entropy\}_f = P(E_f)log\frac{1}{P(E_f)} = -P(E_f)logP(E_f) \tag{19}$$

where $0 \leq P(E_f) \leq 1$. In case of $P(E_f) = 0$, the $P(E_f)logP(E_f)$ will be considered as 0. The E_f is defined as:

$$E_f = M_R \times \theta_R \times \theta_H \times D_R. \tag{20}$$

What we propose here is a way to detect collapsing event, which is an eccentric event in a crowded environment. The framework may be extended by any high-level features which are computed of mid-level features.

To calculate $P(E_f)$ we use cumulative distribution function (*cdf*) which has strict lower and upper bounds between 0 and 1. Deeming $\Omega_{\mu,\sigma}(E_f)$ denotes the *cdf* of E_f. Then $\Omega_{\mu,\sigma}(E_f)$ can be expressed in terms of a special function called the $error\ function$ (erf) or $Gauss\ error\ function$, as:

$$\Omega_{\mu,\sigma}(E_f) = \frac{1}{2}[1 + erf\{\frac{E_f - \mu}{\sigma\sqrt{2}}\}] \tag{21}$$

where $\sigma > 0$ is the standard deviation and the real parameter μ is the expected value. The erf can be defined as a $Maclaurin$ series:

$$erf(E_f) = \frac{2}{\sqrt{\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n\{E_f\}^{2n+1}}{n!(2n+1)} \tag{22}$$

$$= \frac{2}{\sqrt{\pi}}\{E_f - \frac{{E_f}^3}{3} + \frac{{E_f}^5}{10} - \frac{{E_f}^7}{42} + \frac{{E_f}^9}{216} - \ldots\}. \tag{23}$$

Since E_f is skewed to the right (positive-definite) and variances also large, we can use Log-normal distribution. Skewed distributions are particularly common when mean values are low, variances large, and values cannot be negative. Log-normal distributions are usually characterized in terms of the log-transformed variable, using as parameters the expected value, or mean, of its distribution, and the standard deviation. This characterization can be advantageous as log-normal distributions are symmetrical again at the log level [28]. The structure of log-normal distribution of the Eq. 21 yields:

$$P(E_f) = \frac{1}{2}[1 + erf\{\frac{log(E_f) - \mu}{\sigma\sqrt{2}}\}]. \tag{24}$$

By means of Eq. 24 & 23, and having extensive information of the values of μ and σ (say $\mu = 0$, $\sigma = 10$) we can explicitly estimate the value of $P(E_f)$ between 0 and 1.

B. Threshold estimation

To decide the normality or abnormality of the event on the basis of the function analysis, we examine and note the similarities or differences of each calculated value of $Entropy$ with a beforehand defined entropy threshold T_E, i.e., a deviant frame can be detected *if & only if* $Entropy < T_E$, otherwise standard frame. Hypothetical outlook of reckoning T_E is that we pay attention to the minimum number of entropies in large videos which keep under control snobbishly standard events:

$$T_E = \min_{k=1\ldots n}\{Entropy\}_k \tag{25}$$

where n is the number of frames of the video database. The T_E (also $Entropy$) depends on the controlled environment (video stream), specifically the aloofness of the camera to the scene, the orientation of the camera, the type and the position of the camera, lighting system, density of the crowd (e.g., motion ratio area), etc. The more is the remoteness between camera and the scene, the less is the considerable amount of optical flows and blobs. In case of escalator, T_E also places trust on the escalator type and position. Taking into account of these facts, we think carefully that we have at least one threshold by a video stream. If we have N video streams, which are the case in sites such as airport, shopping mall, bank, play ground, subway, concert, cinema hall, school, hospital, parking place, town center, political event, etc., then we put forward at least N thresholds. If the video stream leaves for another, then the threshold should be made over.

The theoretical principle is very time consuming, $O(n)$, if we consider that n corresponds to three months (or more) of video records, then n goes extremely high. That is why we propose the following approach. For each video stream, the input of the algorithm is the video database v_d, composed exclusively of normal events and the output is the set of potential thresholds. In a video stream, the threshold depends of the density of the movement (motion area ratio). In general, the higher is the density, the higher is the threshold. To simplify our implementation, mainly based on video duration along with motion area ratio, we consider three-category of densities namely low, medium, and high. And then we compute the thresholds associated to the three-category of densities of normal event. The algorithm relies on sampling. Finding representative frames of the three categories of density in normal situation, for the entire video database is extremely time consuming. For instance, 3 months of a video stream needs 3 months of processing. Therefore, the algorithm draws a sample of the database, applies K-medoid algorithm on the sample, and finds the medoids of the sample. Each medoid corresponds to a density. The point is that, if the sample is drawn in a sufficiently random way, the medoids of the sample would approximate the medoids of the entire video database. To come up with better approximations, the algorithm draws multiple samples and gives the best clustering as the output. For accuracy, the quality of a clustering is measured based on the average dissimilarity of all videos in the entire database also including those videos in the samples. The video database is composed of frames and each frame is represented by the motion area ratio (e.g., density). Experiments realized in more general way reported in [29] indicate that five samples of size less than hundred give satisfactory results.

```
Algorithm (input: $v_d$, output: $threshold - density - j$ $|j = 1..3$)
begin
1. For i=1 to 5, repeat the following steps:
2. Draw a sample of 50 frames randomly from the entire video database, and call Algorithm K-medoids to find $k$ medoids of the sample.
3. For each frame $f$ in the entire data set, determine which of the $k$ medoids is the most similar to $f$.
4. Calculate the average dissimilarity of the clustering obtained in the previous step. If this value is less than the current minimum, use this value as the current minimum, and retain the 3 medoids found in step 2 as the best set of medoids obtained so far.
5. Return to step 1 to start the next iteration.
6.
   $$Threshold - density - j = \min_{k=1...card(c_j)} \{Entropy\}_k \qquad (26)$$
   where $c_j$ is the cluster $j$ and $j = 1, 2, 3$ corresponding respectively to low, mid, and high densities.
end
```

Experiments showed that the clustering algorithm performs satisfactorily for large database. K-medoids of $O(n^2)$, by applying K-medoids just to the samples, each iteration is of $O(n)$. The complexity is reasonable, because the value of n (number of frames) is not high.

VII. Indexing Architecture

We will discuss compendiously different components of indexing architecture followed by the proposed extensions in the context of video surveillance indexing.

A. Components of indexing architecture

Indexing architecture is composed of three components specifically extraction, representation, and retrieval.

1) Extraction: This component extracts automatically in real-time low-level features and computes automatically mid-level features. Only abnormal events, represented by numeric vectors, are stored in the database. Our architecture gives to the user the possibility to annotate the abnormal events, particularly useful in keyword research. The architecture does not need methods, whether aided by the user or not, to identify relevant regions (regions of abnormal events) in videos. The motions point automatically and easily to relevant regions.

$$Database = \{Database(s), s = 1...n\} \qquad (27)$$

$$Database(s) = \{u_r(i), i = 1...m, a_r(j), j = 1...q\} \qquad (28)$$

The $Database(s)$ is the database associated to the video stream s where s, m, n, and q indicate respectively the identifier of the video stream, cardinality of the database s, cardinality of the set of video streams, and cardinality of the association set. The $a_r(j)$ presents the set of associations between features related to the database s.

2) Representation: This component corresponds to the abnormal events in the database. The unit of representation u_r, is a vector that describes the properties of an abnormal event, is composed of: identifier of the u_r, identifier of the video stream s, instant of begin in the video stream i_b, instant of the end i_e, i_b of the previous segment p_s, value of the entropy e, motion area ratio, direction variance-mean ratio, distance variance-mean ratio, direction histogram, abnormality a_b, and annotations a_n. We underline the fact that u_r may be extended by any mid-level of features. Those considered till now are enough to deal with our problem, however others may be added, when necessary.

$$u_r = (s, i_b, i_e, p_s, e, M_R, \theta_R, \theta_H, D_R, a_b, a_n) \qquad (29)$$

p_s is the previous normal segment to the current u_r. The structure of p_s is the same as u_r, except that $a_b = 0$. Nonetheless, in u_r, $a_b = 1$.

3) Retrieval: This component consists of specifying the user query, matching, and return of answers. The retrieval consists of searching similar abnormal events by selecting target abnormal event or content properties such as density, velocity, and direction or combinations of these. The architecture includes a classical query system, in relational model, that lets users form a query by specifying the interval of time when the event was produced or simply specify queries based on the key words, presented in the annotations. The system includes a visual query tool that lets users form a query by specifying parameters of u_r, such as entropy, motion area ratio, etc. Finally, the retrieval process computes distances between source and target features, and sorts the most similar vectors.

B. Proposed extensions

The basic properties of the architecture presented so far are classical and shared by lots of visual information systems. In the context of video surveillance indexing, we propose the following extensions:

1) On the basis of annotations, associated to abnormal events, the system extracts motion features and discovers the associations between mid-level features e.g., e, M_R, θ_R, θ_H, D_R, a_b, or a_n when available.
2) When an abnormal event is detected and inserted into the database of the stream s ($Database(s)$), we give to the user the possibility to annotate it. If the users do not want to annotate it, then the system suggests annotations on the basis of previously deployed associations. When the users validate the annotations, then they are saved in the database. To make the associations confident, we implement two confidence measures: intensity of implication and conditional probability. We keep in the database associations where the intensity of implication or conditional probability more than high threshold (e.g., 90/100). In the query, we can consider both. Given the probability space (F, P), and two features A and B belong to F, with $P(B) > 0$, the conditional probability of the mid-level feature A, respectively the annotation A, given the annotation B, respectively the mid-level of feature B, noted $P(A|B) = P(A \cap B)/P(B)$, is the probability of some features A, given the occurrence of the annotation B or conversely. The intensity of implication resolves two weaknesses of the conditional probability. The first weakness of conditional probability is that the conditional probability of an association between A and B is invariable when the size of the set B or D (total number of observations) varies. The confidence property is also insensitive to cardinal dilatation (i.e., the size of the subsets increases in the same proportion). The second drawback of the use of conditional probability is that when for a particular class, the cardinality is set to $1/100$ or even lower, it might very well happen that some associations have a high confidence parameter but on the other hand they might be confirmed by a very limited number of instances, and that those associations stem from noise only. This is why it is always dangerous to look for associations with small support even though these rules might look very interesting. Intensity of implication was used to perform the sorting. The intensity of implication, introduced by [30] measures the distance to random choices of small, even non-statistically significant, subsets. In other words, it measures the statistical surprise of having so few examples on a rule as compared with a random draw.
3) The architecture supports efficient retrievals. For example, the user may ask queries such as find videos similar to the source video corresponding to specific abnormal situation, or find all videos which illustrate the collapsing event with such density, direction or simply with such key words.

When the query specifies only features, the system has the possibility to extend the query to the videos which have similar features. Suppose the users ask for videos with annotations containing elderly. The system will search for videos those contain such annotation, and will also discover those associations between the annotation elderly and the features, and return the videos those share such associations. In the retrieval task, the features of the query specification are compared with features of the video databases to determine which videos correctly match (are similar to) the given features. The matching task based on computing the distance between the vectors of the target and the source videos. When mixing several features and annotations, the resulting distance is equal to the sum taking into account the confidence values of the considered features. The resulting videos are sorted; the shortest distance corresponds to the most similar videos. Another advantage of these extensions is the richness of the description contained in the results of queries since the system presents both similar videos and their associations and annotations, when available. For example, the user specifies the query: find abnormal events in videos which are similar to query specification. The query specifies luggage cart. The system matches the query features with video database. The system returns videos those contain the annotation luggage query. The first image was explicitly annotated by the user and the second one was deduced by the system. It contains valid association between the features and the annotation. When the videos are inserted in the database, as containing abnormal events, the system looks for similar videos which contain annotations and assigns the annotations to the inserted video after validation of the user.

VIII. Experimental Results

A. Data sets

The approach has been tested mainly on a set of real videos[1] taken from cameras installed in an airport to

[1] Videos have been provided by a video surveillance company. Images have been either cropped or shown in a short (e.g., escalator exits) for confidential reason.

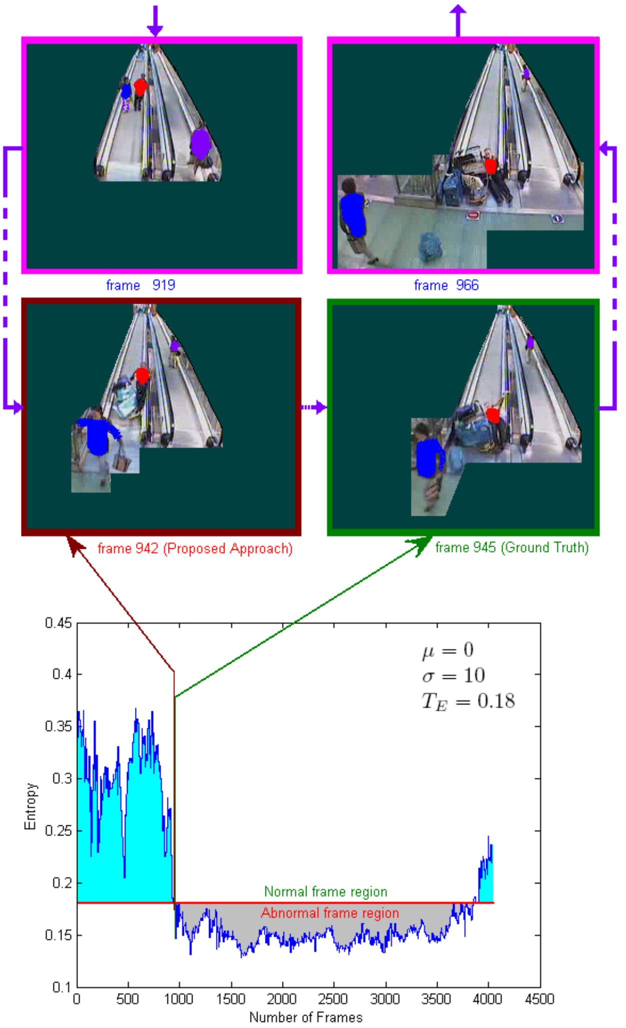


Figure 5. Two persons were riding on the moving escalator. One person escaped by running, while the non-escapee was rundown by the run-away trolley, and subsequently fell down at the exit point of the moving escalator. The blue colored curve is the output of the algorithm.

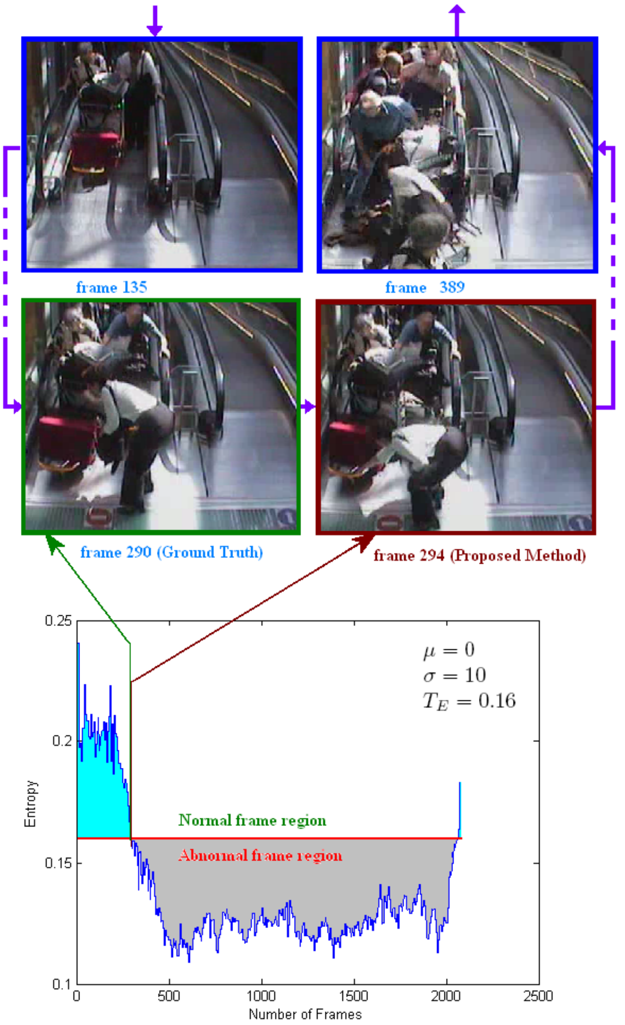


Figure 6. Suddenly the wheels of a trolley held firmly and tightly on the escalator egress and eventually as a consequence causing perilous and inconsistent circumstances on the egress. The blue colored curve indicates the output of the algorithm.

monitor the situation of escalator exits. The videos were used to provide informative data for the security team who may need to take prompt actions in the event of a critical situation such as collapsing or complete failure or breakdown. The data sets are videos from a video surveillance system on escalator exits. Data sets are 10 different length video streams taken in different duration of different time and seasons. Each video stream consists of standard and abnormal events. The standard situations correspond to crowd flows without breakdown in the escalator exits. Generally, in the videos we have two escalators corresponding to two-way-traffic of opposite directions. Abnormal situations correspond to videos which contain collapsing/breakdown events chiefly in escalator exists. The original video frame size is 640×480 pixels. For the features detection and tracking we extract about 1500 features per frame.

B. Results on eccentric event detection

Figures 5 (the 1st video stream listed on the Table I) and 6 (the 4th video stream listed on the Table I) demonstrate two different breakdown circumstances on the escalator exit points in the presence of few people and comparatively large amount of people respectively. Figure 5 substantiates the situation where two persons were standing on the moving escalator and suddenly a trolley rushed out toward them. One person escaped by running and was not descended under the force of trolley, while other did not. In the scene that followed the non-escapee was rundown by the run-away trolley, and subsequently fell down at the exit point of the moving escalator. Figure 6 manifests another aberrant situation where suddenly the wheels of a trolley clenched on the escalator egress and in the long run as a consequence causing perilous and inconsistent circumstances on the escalator egress. Both neurotic situations were successfully detected by the proposed algorithm. The detection results have been compared with ground truth as shown by the arrows. Ground truth is the process of manually marking what an algorithm is looked forward to the probable output. Different video frames in normal and curious/fortuitous circumstances have been differentiated by threshold labels

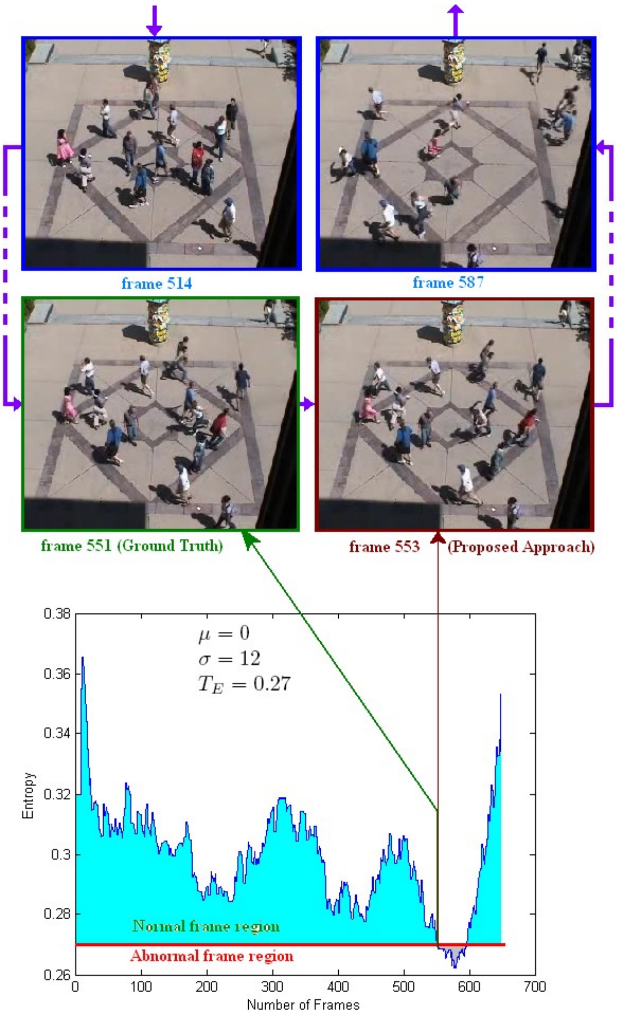


Figure 7. Aberrant event (canyon like part) has been detected by the proposed approach when the group of people has started rushing along random directions. The blue colored curve is the output of the approach.

(horizontal red lines in FIG. 5, 6, and 7). The exhaustive evaluation of the proposed algorithm for the provided data-set has been listed in Table I. The algorithm does not put into effective operation of one of the video streams (the 5th video stream listed on the Table I). Syntactically, the anomalous event of that video stream has taken place in significant far away from the camera, and hence the considerable amount of optical flow vectors is not sufficient to analyze abnormal frames. In Table I $root\ mean\ squared\ error$ (Ψ) has been also appraised. The symbol ∞ shows deliberately that we could not straightforwardly determine the $squared\ error$ by mathematical calculation in respective cases, because of end of video files. However, it is more important to detect the commencement of the atypical state of affairs than that of termination. The estimation of $\Psi = 0$ corresponds to the ideal algorithm or the state of perfect detection. Notwithstanding, the assessed $\Psi \approx 5$ performs appointively for many computer vision applications together with escalators.

Beyond the escalator unidirectional flow of mob videos, the approach has been tested on the videos where the movements of people are random directions, for instance the figure 7. This video consists of 657 frames with attribute 320×240 where both normal and abnormal motion exist. Abnormal motion includes a sudden situation when a group of people start running. From frame 1 to 550 the motion of people is normal. People tend to run from frame number 551. More precisely, the entropy of frame 551 will be lower than that of any other before encountered. Consequently, this frame can be considered as the ground truth frame. In figure 7 the blue colored curve is the output of the proposed approach. The canyon like region represents the abnormal motion activities when the group of people has started to leave their places with very quick motion. For clarity, the ground truth frame 551 and the output abnormal video frame 553, and their corresponding positions on the output curve have been indicated by arrows. In this case the estimation of Ψ also does not exceed the workable range of many computer vision applications.

On the basis of the data-set available here, one density was enough. In fact the clustering method presents three densities close to each other and the entropy of the clusters was not strong enough as a result one density was sufficient. The reason of this result is linked in the fact that the data-set obtained from a video stream of a short period (and the density of the movement, motion area ratio, was low). Consequently, it was not as much as necessary to obtain several data sets with several densities. One density means one threshold, instead of three thresholds in this respect. Long period videos with varying densities of the movement (motion area ratios), having the contextual information e.g., day-night, indoor-outdoor, normal-peak time, occasion, vacation, etc., will have three differing densities (three thresholds).

C. Indexing results

The retrieval system can be evaluated by considering its capacity to effectively retrieve information relevant to a user. It is called the retrieval efficiency. Retrieval efficiency is measured by recall and precision metrics. For a given query and a given number of videos retrieved, recall gives the ratio between the number of relevant images retrieved and the total number of relevant images in the collection considered. Precision gives the ratio between the number of relevant images retrieved and the number of retrieved images.

$$Precision = (Relevant_{videos} \cap Results) \div Results \tag{30}$$

$$Recall = (Relevant_{videos} \cap Results) \div Relevant_{videos} \tag{31}$$

The recall and precision values correspond to the following process: 10 reference (query) videos are selected from the test collection. The subset of videos is selected per type of abnormal event. The threshold used to select multiple associations is set at 80 percent and 90 percent, respectively, for conditional probability and implication intensity confidences. Since the data-set is not very voluminous, it is possible to retrieve all relevant videos.

TABLE I.

The achievement appraisal of the proposed algorithm based on the escalator data sets. The following alphabets and symbols are used for Contextual Information (ConInfo) and curtailment respectively: A $\mapsto$ Less illumination in escalator, B $\mapsto$ No shadowing, C $\mapsto$ Not varying illumination, D $\mapsto$ Varying illumination, E $\mapsto$ Light reflection, F $\mapsto$ Over head light, G $\mapsto$ Friday morning, H $\mapsto$ Thursday afternoon, I $\mapsto$ Saturday night, J $\mapsto$ Sunday night, K $\mapsto$ Wednesday afternoon, L $\mapsto$ Thursday afternoon, M $\mapsto$ Thursday night, N $\mapsto$ Sunday afternoon, O $\mapsto$ Monday night, P $\mapsto$ Spring, Q $\mapsto$ Winter; ∇ $\mapsto$ End of video file, ∞ $\mapsto$ Not straightforwardly determined by mathematical calculation.

Videos & ConInfo	Disruption Circumstances						
	Ground Truth (g_t)		Proposed Approach (p_a)				
	S_{g_t}: Start Frame	E_{g_t}: End Frame	S_{p_a}: Start Frame	E_{p_a}: End Frame	T_E: Threshold	$(S_{g_t} - S_{p_a})^2$	$(E_{g_t} - E_{p_a})^2$
1. C, E, G, P	**945**	**3886**	**942**	**3888**	**0.18**	**9**	**4**
2. D, E, N, P	**127**	**1275**	**130**	**1278**	**0.21**	**9**	**9**
3. E, F, O, P	**496**	∇	**492**	∇	**0.15**	**16**	∞
4. E, F, H, P	**290**	**2070**	**294**	**2075**	**0.16**	**16**	**25**
5. E, F, K, P	**1161**	**4158**	—	—	—	—	—
6. E, F, M, Q	**132**	**706**	**127**	**712**	**0.16**	**25**	**36**
7. D, E, L, Q	**103**	∇	**105**	∇	**0.21**	**4**	∞
8. A, C, J, Q	**100**	∇	**102**	∇	**0.27**	**4**	∞
9. E, F, I, Q	**199**	**1446**	**189**	**1442**	**0.15**	**100**	**16**
10. B, C, G, Q	**72**	**2359**	**78**	**2370**	**0.25**	**36**	**121**
Evaluation Study	*Root Mean Squared Error* $\Psi = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(g_t - p_a)^2}$					$\Psi_{start}=\sqrt{\frac{219}{9}} \approx 5$	$\Psi_{end}=\sqrt{\frac{211}{6}} \approx 6$

In reality the results obtained are vectors, and they are visualized at query of the user. The query is based on similarity. It is obvious that the use of similarity based measure leads to improvements in both precision and recall over the majority of queries tested, when they are correctly annotated. The average improvements of queries over mid-level of features queries are 10 percent for precision and 17 percent for recall. Precision and recall are better for combination of annotations and mid-level features (queries which mix mid-level of features and annotations) than for queries that use only features or annotations. We observed that the general principle of the larger the retrieved set, the higher the recall, and the lower the precision is observed.

IX. Discussion and Future work directions

We moved upwards with a methodology that detects the abnormalities in a crowd flow, with considering a framework of features composed of three levels namely low, mid, and high. It assesses sudden shift and strange locomotion discrepancy of a set of interest points discovered by Harris point of interest detector. Deeming that in video surveillance scenes camera positions and lighting conditions give access to get a large number of Harris points of interest those can be easily captured and tracked. Optical flow information is computed from those points of interest from the specific parts of the escalator using motion heat map, which primarily helps to make faster the calculation process by providing regions of interest. The low-level and mid-level features are generic and would be applied to any situation, considered as abnormal. The framework is designed to support extensive extensions of features in these two levels. The high-level features are dependent of the application domain. For example, in our experiments we defined an entropy function suitable to detect collapsing situations in an airport escalator exits. We defined in the high-level a measure that is sensitive to crowd density and direction. An important aspect of the detection is the automatic estimation of the thresholds. We considered three-level of thresholds corresponding to three-level of densities on the basis of video duration and motion area ratio. Thus depending on these pondering parameters, it is not always mandatory to have three explicit classifications; even one clear threshold is enough in some respects. The method developed is enough generic to deal with any number of densities. If we are confident that low-level features are enough to induce many mid-level features, it is not evident to define exhaustively all mid-level features to detect any abnormal events in a crowded scene. We expect to define libraries of mid-level features to optimize the detection of abnormal events. The framework defined is suitable to support such libraries. We developed an indexing method that supports both texture and similarity search. Furthermore, we developed an approach that extracts associations between mid-level features and annotations. At the end, the queries are not limited to the abnormal events but also to the segment of videos happened before the abnormal event. The achievement of the algorithm has been assessed mainly by a set of escalator videos obtained by a single camera installed in an airport as well as the videos where the movements of mob are in random directions. The methodology developed is promising on its robustness and has some degree of acceptance over the existing state-of-the-art, eventually expects further investigation to get comparative high degree of performance.

Future work direction would be expected to extend the estimation of the motion variations with factors such as acceleration by tracking the points of interest over multiple frames. The contextualization of the system has also a great significance. Introducing context information to optimize the system configuration allows using the same underlying detection algorithms in different locations and in an efficient way. The understanding of the system configuration is also suggesting a consciousness of high position for the security team to assess the current situation. The framework proposed is presented in the way that we can investigate many application domains and any benchmarks (e.g., TRECVID). Yet, we considered that in video surveillance scenes, camera positions, and lighting conditions allow getting a large number of Harris corners that can be easily captured and tracked. Since the mid-level features are extracted based on the result of Harris corner detection, these features might be sensitive to textures. For example, if a person wears a grid-dress-like cloth, there will be too many corners detected from the region of him/her so that most motion directions (e.g., 50% or more) in this frame are the same as the movement direction of the person. Features like direction histogram would be distorted in this situation. Further investigation would take into account this presupposition. A potential solution of this problem could be figured out as: each object can be characterized by a set of corners obtained with a color Harris detector. Each corner can be distinguished by its local appearance such as a vector of local characteristics. The use of a set of corners allows tracking the object through partial occlusion as long as one or more corners remain visible. To increase robustness, it could be important to exploit potential geometric relationships between the corners. Future work would also take into account the dedication of multiple-camera so that video could be conclusively broken down into its essential features properly in all parts (e.g., commencement, halfway point, and outlet) of an elongated escalator to proclaim the eccentric event if there will exist any. Future work would also carry on the algorithm (if something indispensable) to bring about the circumstances where very high dense crowds are milling like in retail or a surveillance video of a famous public/private place, etc. Although three-level of thresholds are computed based on video duration and motion area ratio, the threshold computation step is largely dependent on representative video libraries and sampling. This tedious threshold computation step would be a bottleneck of the system for quick adaptation to new context. Thus, finally, it would be worth investigating to include more high level features and using suitable classifiers to eliminate the threshold computing step and evaluating the technique with a wide range of data-set.

X. Conclusion

We keyed out a methodology that first extracts features of video streams and detects abnormal events in a crowded environment. Subsequently abnormal events are indexed for retrieval. The motivation of the methodology is the discrimination of features which are independent from the application domains. Low-level and mid-level features are generic and independent of the type of abnormality. High-level features are dependent and used to detect abnormal events, whereas both mid-level and high-level features are run through the indexing scheme for retrieval. The methodology has been verified mainly escalator videos and it is promising on its robustness and enough generic to deal with any number of densities. Yet, threshold computing step would be a bottleneck of the system for quick adaptation to new context, it would be one major concern among some future work directions.

Acknowledgment

This work was supported by the MIAUCE project, which is the 6^{th} Framework Research Programme of European Union (IST-2005-5-033715). *www.miauce.org*.

References

[1] A. N. Marana, S. A. Velastin, L. F. Costa, and R. A. Lotufo, "Estimation of crowd density using image processing," *Image Processing for Security Applications (Digest No.: 1997/074),IEE Colloquium*, pp. 11/1–11/8, Mar. 1997.

[2] H. Rahmalan, M. S. Nixon, and J. N. Carter, "On crowd density estimation for surveillance," *International Conference on Crime Detection and Prevention*, 2006.

[3] S. F. Lin, J. Y. Chen, and H. X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *Systems, Man and Cybernetics, Part A, IEEE Transactions*, vol. 31, no. 6, pp. 645–654, Nov. 2001.

[4] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," *IEEE Conference on Cybernetics and Intelligent Systems*, vol. 1, pp. 170–173, Dec. 2004.

[5] A. Davies, J. H. Yin, and S. Velastin, "Crowd monitoring using image processing," *Electronics and Communication Engineering Journal*, vol. 7, no. 1, pp. 37–47, 1995.

[6] B. Boghossian and S. Velastin, "Motion-based machine vision techniques for the management of large crowds," in *Proceedings of ICECS: 6th IEEE International Conference*, vol. 2, 1999, pp. 961–964.

[7] F. Cupillard, A. Avanzi, F. Bremond, and M. Thonnat, "Video understanding for metro surveillance," *Networking, Sensing and Control*, vol. 1, pp. 186–191, 2004.

[8] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Hidden markov models for optical flow analysis in crowds," *International Conference on Pattern Recognition*, vol. 1, pp. 460–463, 2006.

[9] ——, "Modelling crowd scenes for event detection," in *Proceedings of the 18th International Conference on Pattern Recognition*, vol. 1, 2006, pp. 175–178.

[10] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, June 2007.

[11] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[12] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, Aug. 2007.

[13] J. W. Davis and A. F. Bobick, "The representation and recognition of action using temporal templates," in *MIT Media Laboratory Perceptual Computing Section Technical Report Nr. 402. Appears in: CVPR'97*, 1997.

[14] T. Xiang and S. Gong, "Incremental and adaptive abnormal behaviour detection," in *IEEE International Workshop on Visual Surveillance*, 2006, pp. 65–72.

[15] A. B. Y. Ivanov, C. Stauffer, and W. E. L. Grimson, "Video surveillance of interactions," in *CVPR Workshop on Visual Surveillance*, 1999.

[16] D. Xie, W. Hu, and J. Peng, "Semantic-based traffic video retrieval using activity pattern analysis," *International Conference on Image Processing*, vol. 1, pp. 693–696, 2004.

[17] R. R. Wand and T. Huang, "A framework of human motion tracking and event detection for video indexing and mining," *DIMACS Workshop on Video Mining*, 2002.

[18] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 2–10.

[19] H. P. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover," in *Technical Report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doct. diss., Stanf. University*, 1980.

[20] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988, pp. 147–152.

[21] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[22] B. Horn, "Robot vision," in *MIT Press, Cambridge*, 1986.

[23] A. Verri and T. Poggio, "Against quantitative optical flow," in *International Conference on Computer Vision*, 1987, pp. 171–180.

[24] J. Shi and C. Tomasi, "Good features to track," in *Proceedings CVPR: Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[25] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

[26] J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature traker," in *A part of OpenCV Documentation, Intel Corporation, Microprocessor Research Labs*, 2000.

[27] J. L. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[28] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *BioScience*, vol. 51, no. 5, pp. 341–352, May 2001.

[29] L. Kaufman and P. Rousseeuw, *Fininding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[30] R. Gras and A. Lahrer, "L'implication statistique, une nouvelle methode pour l'analysedes donnes," *Mathematique, Informatique et Sciences humaines*, vol. 120, pp. 5–31, 1993.

Md. Haidar Sharif is currently a PhD candidate, supervised by Prof. Dr. Chaabane Djeraba, at University of Sciences and Technologies of Lille (USTL), France. He received his MSc in Computer Engineering and BSc in Electronics & Computer Science degrees from Duisburg-Essen University, Germany and Jahangirnagar University, Bangladesh in 2006 and 2001, respectively. Since 2002-2004, he had worked on a project concerning computer architecture, supervised by Dr. Christian Seidel, at the Department of Theory & Bio-Systems, Max Planck Institute of Colloids and Interfaces, Golm, Germany. His research interests include computer vision & pattern recognition, visual computing, computer architecture, and high-performance distributed computing.

Nacim Ihaddadene received his Master degree in Computer Science from Polytechnic School, University of Nantes, in 2002. He got his PhD in Computer Science from the University of Sciences and Technologies of Lille (USTL) in 2007. Dr. Nacim Ihaddadene is currently a Post-doctoral Research Engineer as well as the co-manager of the MIAUCE Project (Multimodal Interactions Analysis and exploration of Users within a Controlled Environment). His research interests include computer vision and pattern recognition.

Chaabane Djeraba received his PhD of Computer Science in 1993. He obtained, in 2002, the Habilitation thesis to supervise research, at Polytechnic School of Nantes University. He had been, 1994 to 2003, an associate/assistant professor at Polytechnic School of Nantes University. Since 2003, he has been a professor of Computer Science, at University of Sciences and Technologies of Lille (USTL). Prof. Dr. Chaabane Djeraba is leading a research group of Computer Science Laboratory of Lille (LIFL). His research background includes multimedia indexing, retrieving, and mining. His current research concerns video/web usage mining, and optimization of normalized multimedia descriptions. His research objective is human-centered multimedia models, with faceted view on capturing human experience, including both usage and multimodal information.