A Modular System for Global and Local Abnormal Event Detection and Categorization in Videos

Ahmed Chamseddine Ben Abdallah, Michèle Gouiffès, Lionel Lacassagne

Machine Vision and Applications manuscript No.
(will be inserted by the editor)

A Modular System for Global and Local Abnormal Event Detection and Categorization in Videos

Ahmed Chamseddine Ben Abdallah, Michèle Gouiffès, Lionel Lacassagne

the date of receipt and acceptance should be inserted later

Abstract This paper presents a modular system for both abnormal event detection and categorization in videos. Complementary normalcy models are built both globally at the image level and locally within pixels blocks. Three features are analyzed: 1) spatio-temporal evolution of binary motion where foreground pixels are detected using an enhanced background subtraction method that keeps track of temporarily static pixels, 2) optical flow, using a robust pyramidal KLT technique ; and 3) motion temporal derivatives. At the local level, a normalcy MOG model is built for each block and for each flow feature and is made more compact using PCA. Then, the activity is analyzed qualitatively using a set of compact hybrid histograms embedding both optical flow orientation (or temporal gradient orientation) and foreground statistics. A compact binary signature of maximal size 13 bits is extracted from these different features for event characterization. The performance of the system is illustrated on different datasets of videos recorded on static cameras. The experiments show that the anomalies are well detected even if the method is not dedicated to one of the addressed scenarios.

Keywords Abnormal Event Detection · Categorization · Video Analysis · Crowded Scenes

Ahmed Chamseddine
Univ. Paris-Sud, Universit Paris-Saclay, France

Michèle Gouiffès
LIMSI, CNRS, Univ.Paris-Sud, Universit Paris-Saclay, France
E-mail: michele.gouiffes@limsi.fr

Lionel Lacassagne
Sorbonne Universits, UPMC Univ Paris 06, CNRS UMR 7606, LIP6, France
E-mail: lionel.lacassagne@lip6.fr

1 Introduction

In order to enhance our sense of security, the use of video surveillance systems has grown very rapidly in the last few years. The computer vision and the multimedia computing communities have recently witnessed a surge of interest in automatic abnormal event detection and/or categorization in computer vision [1–4].

This task can be extremely challenging and has its own limitations. The main one is the lack of a universal and objective definition of *abnormalities*. They are usually defined in a subjective form and interpreted differently on a same dataset. In this paper, abnormal events are described as unusual events encountered at a specific context and may be reported for further investigations [5]. Other limitations are related to the sparseness, rarity, and discontinuity of abnormal events which limit the number of examples available to train an anomaly detection system. For crowded scenes, this difficulty is compounded by the complexity of normal crowd behaviors. On the one hand, unlike videos containing one or a few objects of interest, when dealing with large or massive groups of moving objects like dense crowds, individual object tracking is virtually impossible. On the other hand, it is infeasible to enumerate precisely the set of abnormalities that are possible in a given surveillance scenario.

One common solution to these issues is to define anomalies as events of low probability with respect to a probabilistic model of normal behavior. Then, a statistical analysis of anomaly detection can be performed, which conforms with the intuition of anomalies as events that deviate from the expected [4].

This paper proposes a modular system for abnormal event detection and characterization. It involves three categories of generic low-level features likely to

be available in any videos and therefore suitable for various applications, for scenes showing either small groups of objects or dense groups like crowds. First, the spatio-temporal evolution of binary motion pixels is analyzed both globally and locally. These motion pixels are detected using an enhanced background subtraction method that keeps track of temporarily static pixels. Then, optical flow and its temporal gradient are analyzed locally within pixels blocks. Optical flow is computed using a pyramidal KLT technique robust to photometric changes. Two models are used for these local features: a mixture of Gaussians (MOG) for quantitative analysis, which is made more compact using principal component analysis (PCA), and a set of compact hybrid histograms that embed both motion (or its gradient) orientation and binary motion statistics. On the basis of these simple features, a binary descriptor vector can be used to describe the event. The system is modular in the sense that the different anomaly detectors are independent and can be used solely for a given application.

The remainder of the paper is organized as follows. Section 2 reviews previous work on abnormal event detection. An overview on the proposal is provided in Section 3 and the main stages are detailed from Section 4 to Section 7. The algorithmic complexity is discussed in Section 8. Section 9 presents the experimental evaluation performed on different real datasets. Finally, conclusions are provided in Section 10.

2 Related Work

The procedure generally applied for anomaly detection consists in modeling normal behaviors and then estimating the deviations between the resulting model and the observed behaviors [5] since statistics are far easier to collect on normal behaviors than on anomalies.

Anomalies can be classified into three main categories: trajectory-based, motion-based, and anomalies detected jointly from motion and appearance. The trajectory-based approaches are founded on segmenting and tracking each object in the scene, either explicitly or implicitly, then on fitting models to the resulting object tracks [6–10]. These procedures are however computationally expensive and difficult to apply for crowded or cluttered scenes. To some extent, the trajectory-based method can be applied to crowd motion, for example in [11] where representative trajectories of a crowd flow are learned by clustering optical flow-based particle trajectories.

In motion-based techniques, the processing of each individual object is avoided by modeling motion patterns. The state-of-the-art methods differ mainly on the model used. With increasing complexity, it can be histograms of pixel changes [12], optical flow histograms [13, 14], or optical flow measures, mainly flow magnitude and direction [11,15–17]. In [12], the authors used a Markov Random Field (MRF) model to describe the probability of observations within the same spatio-temporal volume. In a different way, motion can be considered as a mixture of unitary events, such as [13] and [14] where unusual events are detected *via* a sparse reconstitution of query signals from a *normal* event dictionary. Local optical flow can also be modeled with a mixture of probabilistic principal component analyzers [15]. Social force models [18] have also been used in [16] and [17] to estimate the moving particles interaction forces. Note that all these approaches model dynamics and ignore anomalies of object appearance and thus anomalous behavior.

Some approaches use more comprehensive representations including appearance and motion [3,19–21]. By modeling motion variations of several space-time volumes as well as their spatial-temporal statistical behaviors, [19] proposes to characterize the overall behavior of the scene. In [20], the authors consider a given event as abnormal if the spatio-temporal patches cannot be composed with previously learned visual examples. In a more empirical way, [21] quantifies abnormality by creating rules that are based on score functions derived from local nearest neighbor distances across spatio-temporal locations and scales. Finally, [3] proposes a joint detector of temporal and spatial anomalies using a mixture of dynamic textures models (noted MDT). This detector uses dynamic textures to design models of normalcy over both space and time dimensions. A global analysis of the flow is however not included in the system.

Even if several approaches were proposed to detect abnormal events, it is usually quite difficult to objectively compare their results. Usually, these methods use different representations of motion and appearance with different graphical models of normalcy, which are usually dedicated to a given application and designed for specific scenes or specific definitions of anomaly. Abnormalities are themselves defined in a somewhat subjective form, sometimes according to what the algorithms can detect. Moreover, the same datasets are sometimes interpreted differently from one paper to the other. Finally, experimental results can be presented on datasets of very different characteristics, *e.g.*, traffic intersection or subway entrance, frequently proprietary, and with widely varying levels of crowd density [3]. To finish, most papers propose very interesting and powerful methodologies for anomaly detection [1–5, 9–16, 19–23] or for specific activities categorization [6, 18]. Very few

works attempt to both detect and characterize the abnormalities [24], except for specific applications like water analysis [25] or intrusion detection [26].

In the following, we propose a system for abnormal event detection and categorization that addresses anomaly in a quantitative and qualitative point of view, and which acts both globally and locally in the image.

3 Overview on the proposed framework

The proposed system aims to detect abnormal events at two different levels, on the whole frame (*global*) and on *local* areas. Figure 1 displays the architecture of the system, which consists of four main blocks, from left to right: *features extraction*, *model matching*, *abnormal event detection* (decision making) and *abnormal event categorization*.

1. **Features extraction.** From testing data, this block extracts foreground pixels as well as the motion flow and its temporal gradient. The information provided by motion analysis is used to enhance the quality of background subtraction algorithm. More details about this stage are provided in Section 4.
2. **Model matching.** The *model matching* block checks if the extracted features match the corresponding models built during the training stage (see Section 5) namely: appearance models based on foreground features, motion models based on velocity distribution and spatio-temporal models based on motion gradient distribution. The two latter models are then divided into *quantitative* and *qualitative* models. *Quantitative* models aim to detect abnormal feature values while *qualitative* models aim to detect abnormalities in the shape of the distributions.
   The choice of different features and a collection of models is motivated by the fact that different tasks may require different models of normalcy depending on the context and the application. For instance, a detector of freeway speed limit violations will rely on normalcy models based on speed features. On the other hand, global appearance is more important for the detection of an abnormal crowd behavior (gathering or dispersion for example).
3. **Abnormal event detection.** Based on models matching results, this block should decide whether the event is normal or not (see Section 6).
4. **Abnormal event categorization.** Finally, the *abnormal event categorization* block consists of a simple classifier that should decide, based on distances to models, if the abnormal event under consideration is similar to some of the known events. More details are provided in Section 7.

4 Features Extraction

Three types of features are detected. First, an enhanced background subtraction is proposed to detect both the moving and temporarily static foreground pixels. Then, optical flow and its temporal gradient are computed.

4.1 Enhanced Background Subtraction

Identifying moving objects from a video is a fundamental low-level task. Generally assuming a static camera, background subtraction (BGS) algorithms can fastly provide regions of interest that can serve as masks for more evolved (and possibly greedier) algorithms. Much research has been devoted to develop BGS algorithms that are robust against environmental changes (e.g. various levels of illumination, fog, rain, etc), and sensitive enough to identify all moving objects of interest [27]. Unfortunately, for most techniques, the objects that stop temporarily are assimilated as the background. The detection of these *static* objects may be of a high importance. For instance, crowd gathering in a walkway should be detected and reported as an usual event. However, merging the static crowd into the background makes them unnoticeable. On the other hand, disabling the adaptation factor of the BGS algorithms makes them sensible to environmental changes. A common approach to alleviate this issue is to use multiple background models running at different adaptation rates, and periodically cross-validate between different models to improve the foreground extraction performance [28–30].

In this paper, this issue is solved in a simple way by consolidating foreground over frames and detecting temporarily static foreground. It takes advantage of the robust optical flow described further in 4.2 which is also useful for abnormal event detection at pixel-block level. The method consists in integrating the new position of the previous foreground into the current foreground even if the whole or some of it is not detected by the BGS algorithm. In other words, if a foreground pixel of coordinates (x_t, y_t) at time t moves to the location (x_{t+1}, y_{t+1}) at time $t+1$ (in the next frame), the pixel (x_{t+1}, y_{t+1}) is included in the foreground mask, even if this point is not detected by the background subtraction algorithm or is not moving anymore (static foreground).

Figure 2 displays the diagram of the background subtraction enhancement. Using optical flow, a binary *flow mask* is obtained by calculating the new coordinates of each pixel of the previous foreground. An additional filtering stage can be performed to remove noise from the resulting mask, *e.g.* morphological opening

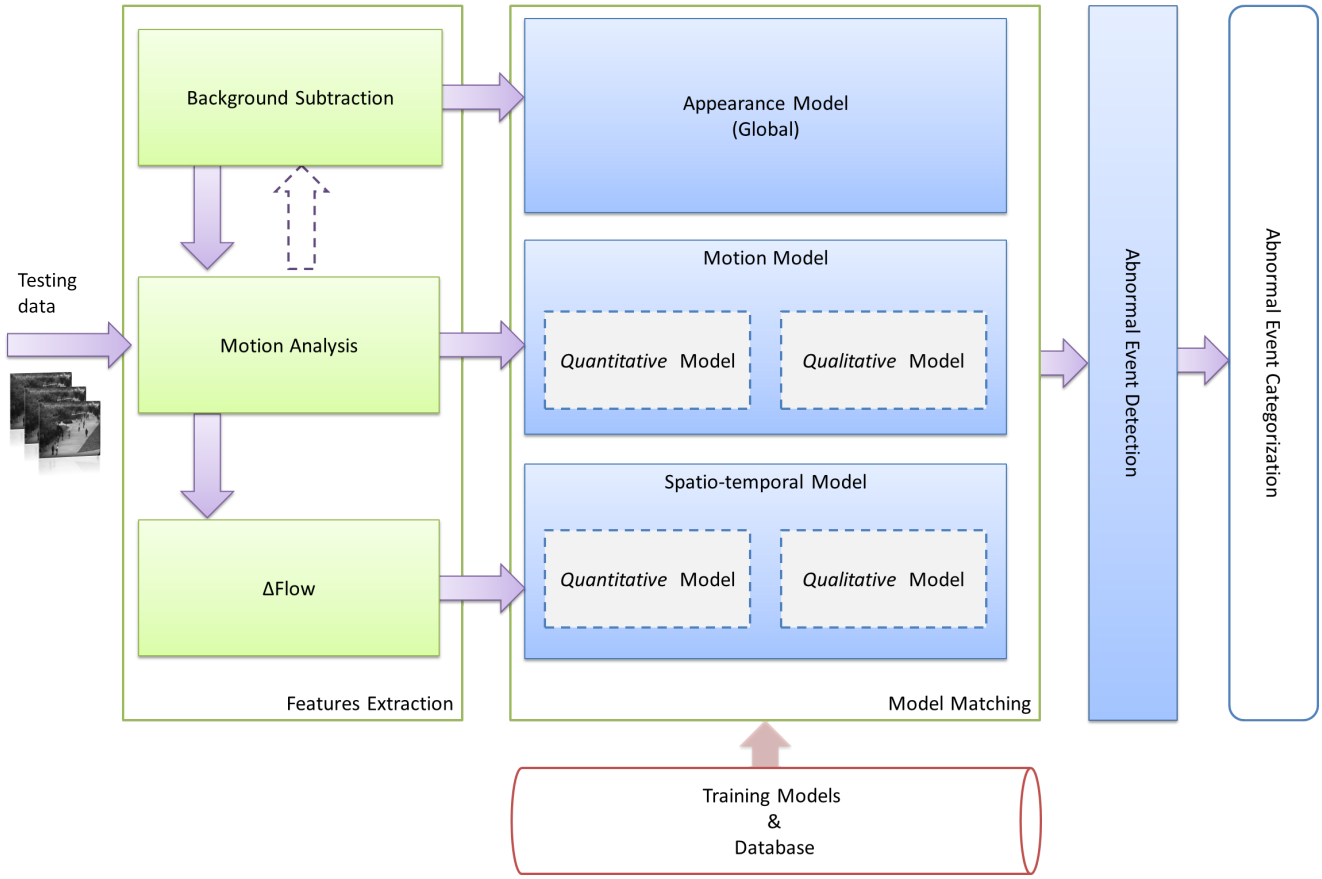


Fig. 1 Architecture of the proposed system in four main blocks, from left to right: 1) Features extraction: background/foreground pixels, motion vectors, temporal gradient of the motion flow; 2) Model Matching, involving a global appearance model as well as two local models, both of them providing a quantitative and qualitative analysis; 3) Decision making to classify the motion as normal or abnormal; 4) Categorization/classification of the detected anomaly.

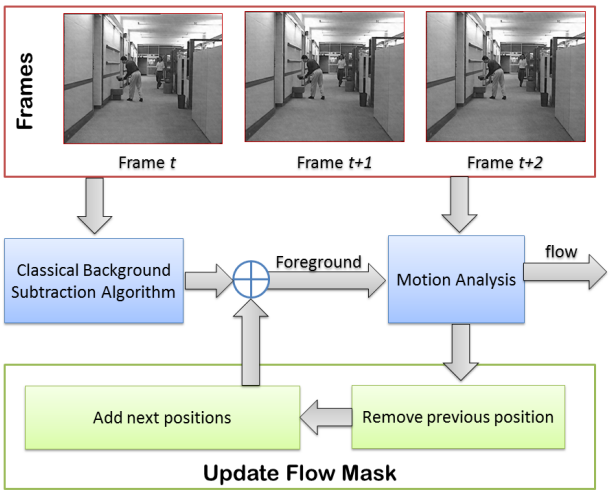


Fig. 2 Principles of the Enhanced Background Subtraction method.

and closing. In the experiments, a square structural element of size 7×7 has been used in all datasets. Then, the enhanced foreground mask is computed by merging the foreground mask obtained by a classical BGS method with the *flow mask*.

In the proposed system, the enhancement of the BGS does not require additional processing since it uses the optical flow that is also useful for local anomaly detection. Only a binary mask of the flow is used so the requirements in terms of memory are low.

4.2 Motion

For motion computation, we use the KLT method [31–33] in its pyramidal implementation [34] to approximate the optical flow of each pixel of the foreground. The parameters of a local photometric model are estimated jointly with the motion model [33], which improves the robustness against lighting changes. For each pixel, the magnitude and the direction of the optical flow are stored.

4.3 Motion Temporal Gradient

To detect anomalies in motion variation, the temporal gradient of the motion flow is computed. By considering the last $2 \times n$ frames and by noting $\mathbf{v}_t$ the flow at time t, its gradient $\dot{\mathbf{v}}_t$ is defined as follows:

$$\dot{\mathbf{v}}_t = \frac{1}{n} \sum_{i=0}^{n-1} (\mathbf{v}_{t-i} - \mathbf{v}_{t-n-i}) \,. \quad (1)$$

The choice of n depends on the frame rate and on the anomalies under consideration. A small value of n makes the algorithm very sensitive to noise and a big value of n may cause a big delay in the detection process; it may also attenuate the abnormality distortion and thus reduce the detection performance. In our experiments, n is set to 5, which corresponds to a period of 0.12s for 25 fps videos. Typically, considering a focal length of 18mm, a walking person located at 100 meters from the sensor will move of 3 pixels approximately in the image [35] during this period of time.

Figure 3 shows an example of feature extraction results on a sample frame from *UCSD ped 1 dataset* [3]. Figure 3(a) presents the original frame showing walking people in a walkway. The foreground extracted by the enhanced BGS algorithm (see Section 4.1) is shown in Figure 3(b), where the black pixels represent moving pixels. Figure 3(c) shows the optical flows computed on each foreground pixel, where each color is related to a different direction of the motion, as described by the color scale of Figure 3(e). To finish, Figure 3(d) shows the gradient of the motion in the given frame.

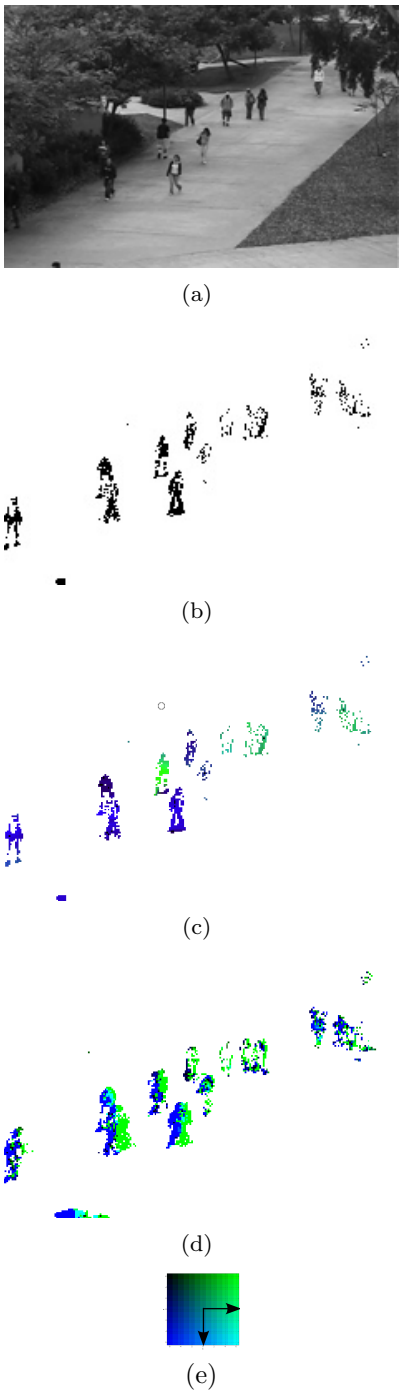


Fig. 3 Features extraction processes on a sample frame (a): (b) detected foreground, (c) the motion flow and (d) the motion gradient. Each color indicates a different direction and amplitude of the optical flow vectors as described by the color scale (e).

5 Global and local models of normalcy

On the basis of the three features introduced in the previous section, three categories of models are built: a global motion model (see 5.1) in order to detect anomalies at the image level, local motion models (see 5.2) and local motion gradients models (see 5.3). In addition, both latter categories comprise a quantitative and a qualitative model.

5.1 Global motion Model

Some simple statistical measures computed on the foreground mask can provide a first analysis of the movement in the scene, and can discriminate normal events from abnormal ones in a straightforward way. Four measures are considered here: the *area* (number of pixels), its *barycenter*, the spatial *variance* (computed as the mean of the variances in x and y directions) and the *percentage of the static area* within the foreground. These

values are easy to compute in any videos. For instance, an unusual gathering will lead to a variance decrease and to an area increase; on the contrary, a dispersion of the crowd may cause a variance increase; a mass movement of the crowd may be detected by a low variance and a significant move of the barycenter; and an unusual crowdedness may increase the area of the foreground. For consistency purpose, the three latter measures *i.e.* barycenter, variance and percentage of the static area, are considered only if the area of the foreground is large enough, above 2% of total number of pixels. Table 1 summarizes the types of abnormalities that can be detected using these four measures.

Table 1 Type of abnormal events that can be detected using a simple statistical analysis on the foreground.

Measure	Low value	High value
Area	Unusual low density	Unusual crowdedness
% static area	Unusual movement	Unusual stopping crowd
Barycenter	Concentration of the movement in an usual side of the frame	
Variance	Unusal gathering	Unusual dispersion

In the training stage, these four features are computed for all the frames, and the maximum and minimum values are collected for each measure. Then, for each testing frame, and for each measure m, the following ratio is computed:

$$r = \frac{\max(m_{min} - m, m - m_{max})}{m_{max} - m_{min}}. \quad (2)$$

An abnormal event is detected when the ratio r is higher than a given threshold (5% in our experiments) *i.e.* when the measure of m is either much higher than m_{max} or much lower than m_{min}. For better precision, m_{min} (resp. m_{max}) is the median value of the first (resp. last) decile.

5.2 Local motion models

In addition to the global model discussed in section 5.1, motion has also to be analyzed locally in the image, with different models in each area. In fact, a normal behavior at a large visual scale may be perceived as highly anomalous when a finer scale is considered. For instance, due to the perspective properties of images, the size of an object decreases when it gets further from the camera. As a result, a motion that can be usual when the object is in front of camera should be considered as abnormal when the object is far away. Also, the global model is able to detect speed limit violations on a freeway but not lane violators. However, a normal behavior in one side of the road should be considered as an abnormal event when perceived on the other side. So, in order to increase the detection sensitivity, the activity is characterized locally in pixel blocks of similar size w^2 using two models.

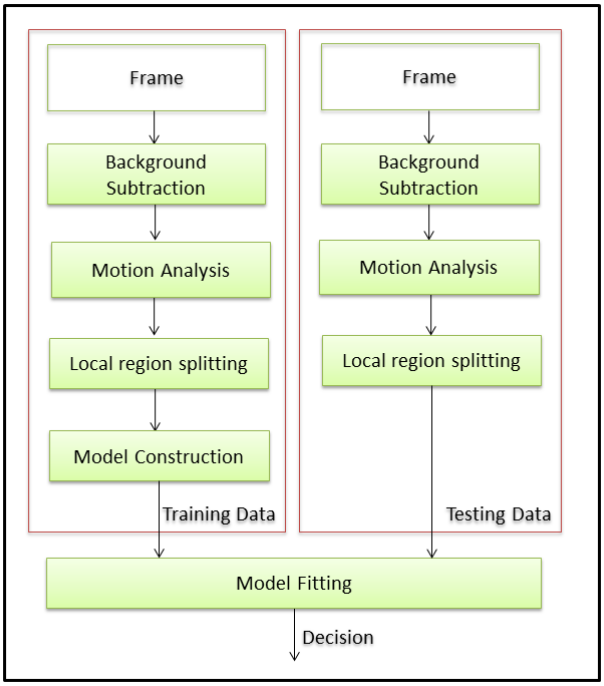
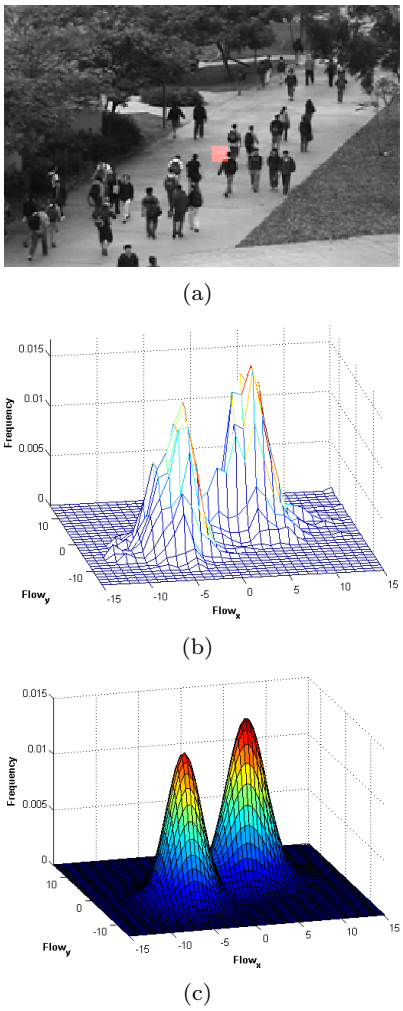


Fig. 4 Architecture of the Local Model Fitting Approach.

5.2.1 Quantitative model based on motion MOG

Figure 4 displays the architecture of the model fitting approach. Given all training frames, considered as normal, the 2D motion vector of each foreground pixel is computed. Then, the frame is split into small contiguous blocks, the width w of which is half the average height of the moving objects in the observed scene. Since the camera is static in most surveillance applications, this value is determined once for all, at the initialization of the system. After collecting the flow in each region, the expectation-maximization (EM) algorithm [36] is used to fit the collected values to a mixture of Gaussian models. For each Gaussian, the mean vector is stored, $\boldsymbol{\mu} = [\mu_x, \mu_y]$, as well as the covariance matrix $\boldsymbol{\Sigma}$.

Figure 5 shows an example of model fitting. The block under consideration, which comes from a sample frame of the *UCSD ped 1 dataset* [3], is highlighted in



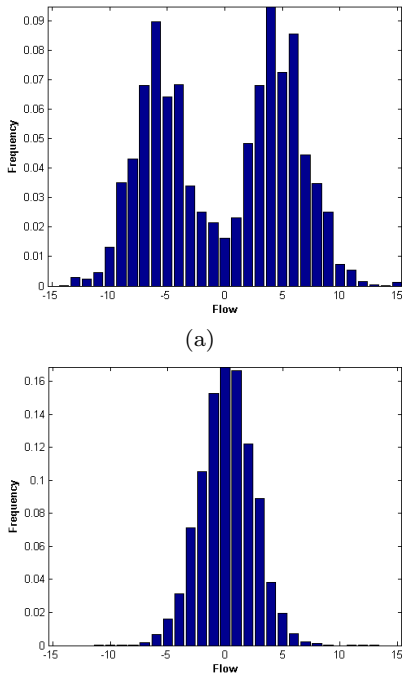
(a) (b) (c)

Fig. 5 An example of model fitting in (a) a local region of a sample frame : (b) the collected flow values and (c) the created models.

Figure 5(a). Figure 5(b) displays the flow values collected in this block. Using EM algorithm, these values are modeled by two Gaussians, as shown in Figure 5(c). They correspond to the two main directions of motion on the walkway. Of course, other scenes may show different models. For instance, a one-way street may show one Gaussian per speed class (for example cars/bikes/pedestrians).

In order to reduce calculation time and noise effect, all flow values are projected on the main direction using Principal Component Analysis (PCA) [37], where the eigenvectors correspond to the main motion directions and eigenvalues to the weights.

When the main eigenvalue explains 80% of the variance, the main eigenvector is assumed to be descriptive enough. Then, the data projected on the eigenvector are modeled by a mixture of 1D Gaussians. Generally, the projection on the secondary eigenvector represents



(a) (b)

Fig. 6 Projected flow values in a given local zone on (a) the main and (b) the secondary eigenvectors.

a white noise, and can be modeled using a single Gaussian model with a null mean.

For example, Figure 6 shows the projection of the flow values collected in the block highlighted in Figure 5(a). It can be seen in Figure 6(a) that the histogram of flow values projected on the main eigenvector can still be modeled with two Gaussian distributions. Figure 6(b) shows the projection of the histogram on the secondary eigenvector.

The local abnormality detection is explained in Figure 7. A normality MOG is built independently for each block in the training stage. In the testing stage, each of the w^2 foreground motion values within a block is assigned to the closest mode considering the Mahalanobis distance D:

$$D = \sqrt{(\mathbf{v} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{v} - \boldsymbol{\mu})}.$$

with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the mean and covariance matrix of the normal distribution (the mere variance when the PCA has lead to a 1D model). For each mode, the average D is computed and an anomaly is detected in the block when it is higher than a threshold (equal to 1 in the experiments).

For the 1D models, it is necessary to check models on both eigenvectors. A flow that does not fit any of the main vector model(s) may indicate a moving object/person with a different speed or direction; a flow

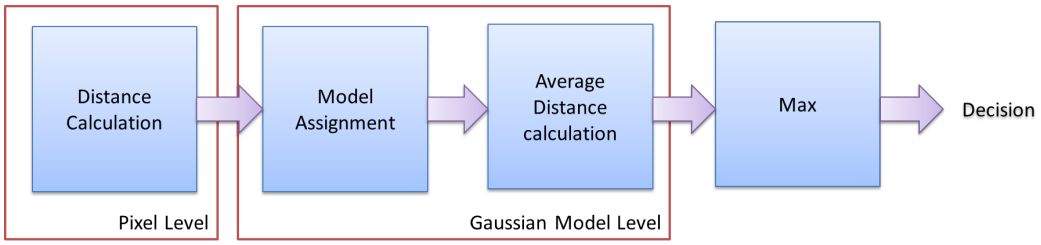


Fig. 7 Architecture of the Model Fitting Block.

that does not match the secondary vector model may indicate an abnormal movement on the transverse direction.

5.2.2 Qualitative modelling using a set of compact hybrid histograms

The *quantitative* model matching described previously can detect the pixels moving with unusual speed or direction but doesn't provide any information about the shape of the distribution. Let us consider a subway entrance where people can move in two opposite directions but not at the same time. Then, motion in the two directions should be considered as a usual event, but the presence of the two flows at the same time should be reported as an abnormal event. To solve this issue, the histograms of motion orientation are computed on the training images database of normal events. Figure 8 explains the training and testing stages required for histogram analysis. Then, an event the distribution of which does not match any of the training histograms will be reported as an unusual event. The histogram contains 10 bins. Each bin $k \in [1, 8]$ counts the number of pixels with the motion angle $k\pi/4 \pm \pi/8$. The ninth and tenth bin count respectively the static foreground pixels and the background pixels. As an example, figure 9(b) shows the histogram computed on the block highlighted in 9(a). For scenes with a principal direction of the motion (*e.g.* walkway, entrance), the dimensions of the space can be reduced as done previously, by projecting the flow on the first principal component provided by PCA. This allows to reduce the algorithm sensitivity to noise while accelerating the matching. The resulting 1D histogram has 5 bins, as illustrated by figure 9(c). The first two bins count the number of pixels on the two directions of the first principal component and the third bin counts the number of pixels on the second principal component. The last two bins count respectively the number of static foreground and background pixels.

One histogram is computed for each frame of the training dataset, and the size of the training samples set is reduced using complete linkage hierarchical clustering

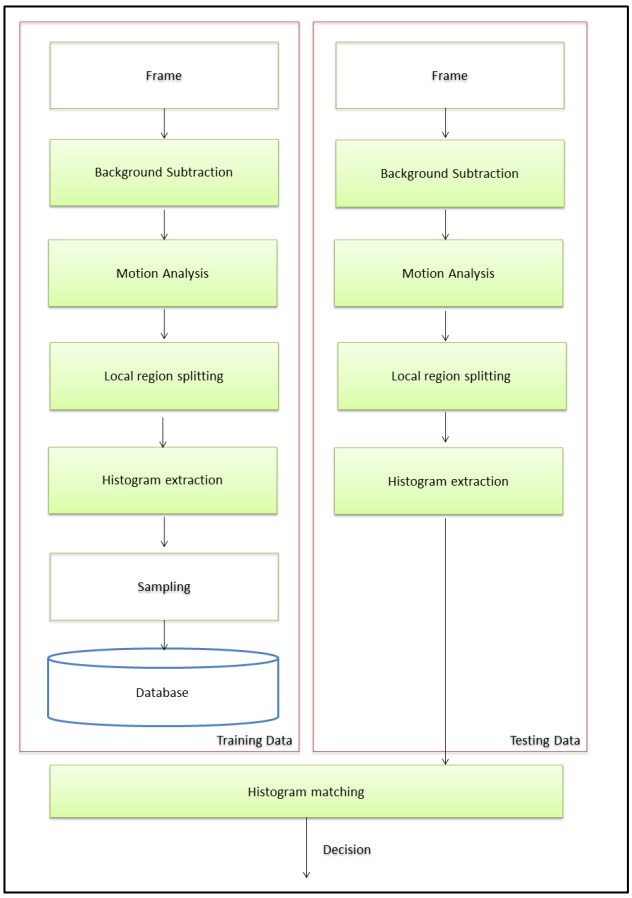


Fig. 8 Architecture of the Local Histogram Comparaison Approach

[38]. The Euclidean distance is used to compare the test histogram with the learned histograms. Note that the Earth Mover Distance has also been tested. The improvement is very limited and the complexity growth is significant.

5.3 Local motion gradient models

In this section, the similar *quantitative* and *qualitative* matching approaches described in Section 5.2.1 and Section 5.2.2 are applied on motion flow gradient. This

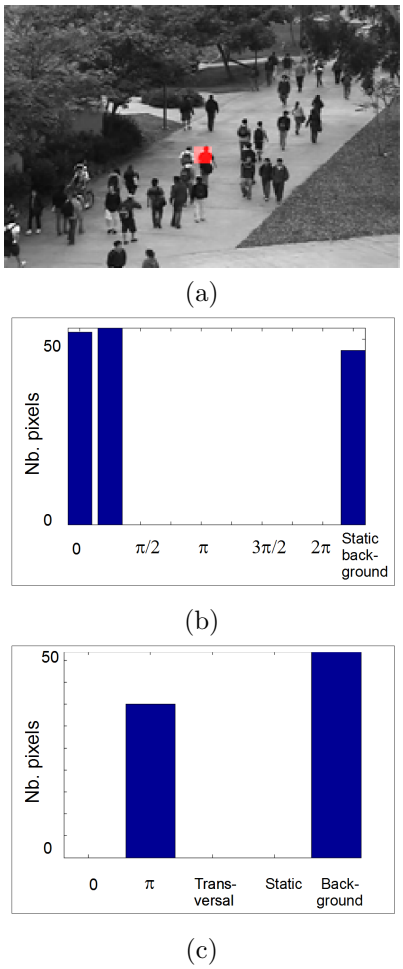


Fig. 9 An example of qualitative histogram in a local region of a sample frame displayed in (a): (b) the global histogram and (c) the reduced one (on 1D).

should allow detecting any abnormalities of flow variation, and detecting moving objects with unusual sizes, *e.g.* cars or bicycle running in a walkway.

6 Decision Making

The association of the different models of normal behaviors described previously should allow detecting different types of abnormalities, as illustrated by Figure 10. The global model allows detecting abnormalities at the frame level, for example for crowd motion (area, barycenter, variance, etc). The quantitative and qualitative local models are designed to detect respectively abnormal flow values and abnormal flow directions. If another type of abnormality occurs, it may be detected if it produces an abnormal variation of at least one of the used features. Consequently, an abnormal event has to be detected when at least one of the features extracted in the test scene does not match the corresponding trained models. Therefore, a simple logical *or* operator is used to report the presence of an abnormal event.

7 Abnormal Event Characterization

When an abnormal event is faced, it may be helpful for the human monitor to identify it. Through our multi-model matching process, thirteen binary features are collected to categorize the type of anomaly: (1) low foreground variance, (2) high foreground variance, (3) low foreground area, (4) high foreground area, (5) out of range foreground barycenter, (6) low static foreground area, (7) high static foreground area, (8) unmatched quantitative flow model, (9) unmatched transverse quantitative flow model, (10) unmatched qualitative flow histogram, (11) unmatched quantitative flow gradient model, (12) unmatched transverse quantitative flow gradient model and (13) unmatched qualitative flow gradient histogram. The 7 former features are specific to the whole frame and describe the *global* behavior of the crowd while the 6 latter features are specific to the given local block. Each abnormal event can be categorized based on one or more features. As a result, it will be categorized by a binary vector, called *signature*, which should be unique for each type of event.

For instance, dispersion events should provide a high variance signature, gathering should be related to a low variance signature and fighting should have simultaneously a low variance, a large static area and an unmatched qualitative flow histogram.

8 Implementation Complexity

By considering N the number of pixels of the given frame, M the number of local regions, and P the number of histogram models in the database, the complexity of our algorithm is $\mathcal{O}(N + M \times P)$. The architecture of our method, displayed in Figure 1, is based on different independent models and algorithms that have been chosen to be efficiently parallelized on a multi-core SIMD General Purpose Processor (GPP). Motion detection, morphological operators and KLT are embarrassingly parallel and can be easily accelerated using SIMD[1] and OpenMP[2]. Temporal gradient of equation (1) can be accelerated by using IIR recursive filters. Finally, the more time-consuming part of the proposed method is the histograms computation. As the regions do not have the same size, the histograms computation will not take the same execution time. This well-known issue can be

[1] Single Instruction Multiple Data
[2] http://openmp.org/

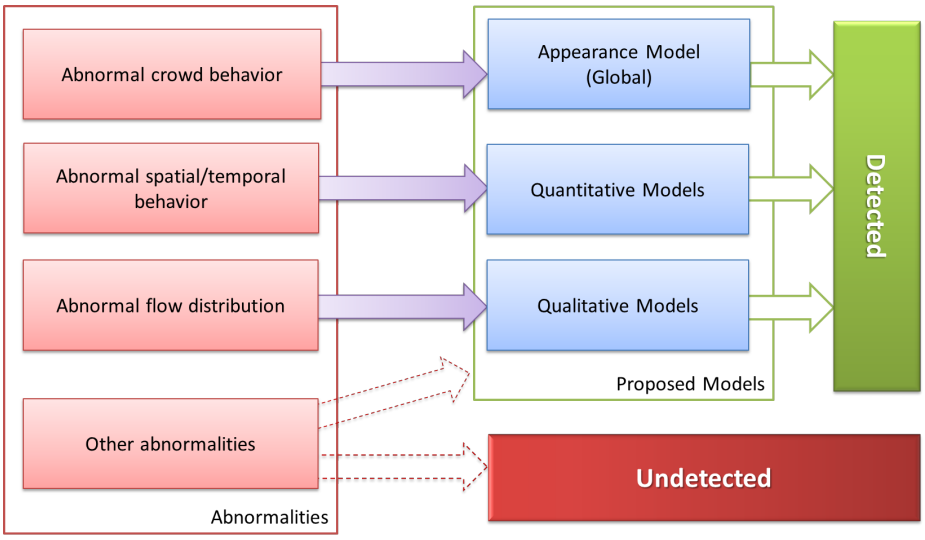


Fig. 10 Types of detected abnormalities.

addressed with `dynamic` scheduling within OpenMP to perform *farming* (one thread per histogram) and fix load-balancing problem, as long as there are more regions (a few hundreds) than processors (2×12 core for bi-socket Ivy-Bridge EP Xeon and 4×15 core for quad-socket Ivy-Bridge EX Xeon). We are hence confident in parallelizing this algorithm on a multi-core SIMD processor.

The acceleration results would probably be less satisfactory on GPU, mainly because histogram computation is an issue on such architectures. Indeed, there are not enough regions (hundreds) and histograms per region (maximum 4) to feed all the processing elements of the GPU with data. The latest ones have more than thousand cores: respectively 2034 and 2688 for Nvidia GTX 780 and GTX Titan. The problem remains the same with AMD GPU. Moreover, the PCIe bandwith to communicate between host and GPU is very slow, compared to internal bandwith of GPU and GPP.

9 Experimental Results

To evaluate the proposed system through different scenarios, three datasets showing different scenes are used, namely the UMN dataset[3], the UCSD dataset [3], and the BEHAVE dataset[4]. The UMN dataset is a set of escape events that allows us to test our global and local models for abnormal event detection. The UCSD dataset consists of videos of a crowded pedestrian walkway and is useful to test our PCA-based local model to detect local abnormal events. Finally, the BEHAVE dataset contains some labeled events which will be helpful for testing our event categorization component.

In all the experiments, foreground pixels are detected using the enhanced background subtraction described in Section 4.1 based on a Mixture of Gaussian Model [39]. The parameters used for the experiments have been discussed in the previous sections and remain the same during all the experiments.

9.1 UMN Dataset

This dataset, collected by University of Minnesota contains videos of 11 different scenarios of an escape event. The videos are shot in 3 different scenes, including both indoor and outdoor situations, and in different configurations of the sensor, as illustrated by the samples of Figure 11. Scenes in this dataset are relatively crowded, with about 20 people walking around. Each video clip starts with an initial part of normal behaviors and ends with sequences of abnormal behaviors. In the following, we use the half of the first part (normal behavior) to train the system and to create the normality models. Then, the algorithm is evaluated on the whole dataset.

Figure 12(a) displays the variance of the foreground on the eleventh scenario. The vertical red line indicates the starting of the escape event and the horizontal green lines show the minimum and the maximum variance values detected on the training frames (*i.e.* the first 367 frames). It can be seen that during the escape event, the variance increases and exceeds significantly the maximum value reached during the training phase. From these variance values, the discriminative ratio is

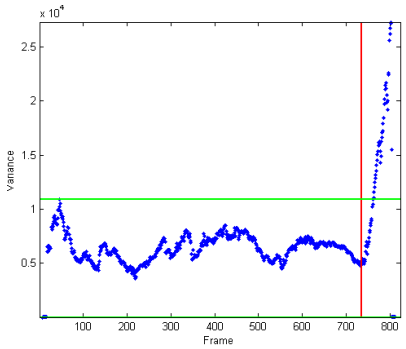
[3] Unusual Crowd Activity Dataset: http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi
[4] http://homepages.inf.ed.ac.uk/rbf/BEHAVE/

Fig. 11 Sample frames in three different scenes of the UMN dataset.

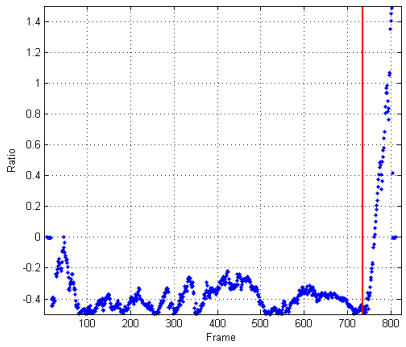
calculated as described in (2). The results are displayed in Figure 12(b).

Figures 12(c) and 12(d) are two sample frames from the considered scenario. The first one corresponds to a normal behavior of the crowd, with the lowest variance value 4397, which yield a ratio r of -0.4 (equation (2)). In this frame, the people are close to each other and are gathered in the center of the frame. The second one displays the dispersion event and shows a high variance of $1.625.10^4$ which corresponds to a ratio of 0.49. In this frame, people are spreading in different directions, the foreground barycenter is almost at the center of the frame and the escaping people are close to the frame edges, which make the variance increase.

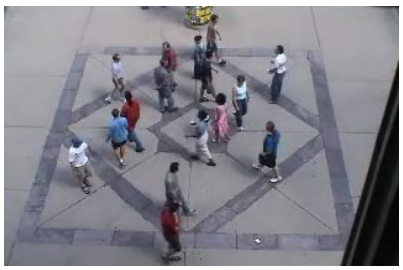
Figure 13 displays the local results on some escape event frames. The escaping individuals have a higher speed than in normal situations, and they can be easily detected using our local models. In the presented frames, even though our approach does not track people, the abnormal areas are correctly highlighted.



(a)



(b)



(c) frame 215

(d) frame 776

Fig. 12 The variance on the eleventh scenario: (a) its distribution, (b) the discriminative ratio and (c) (d) two sample frames

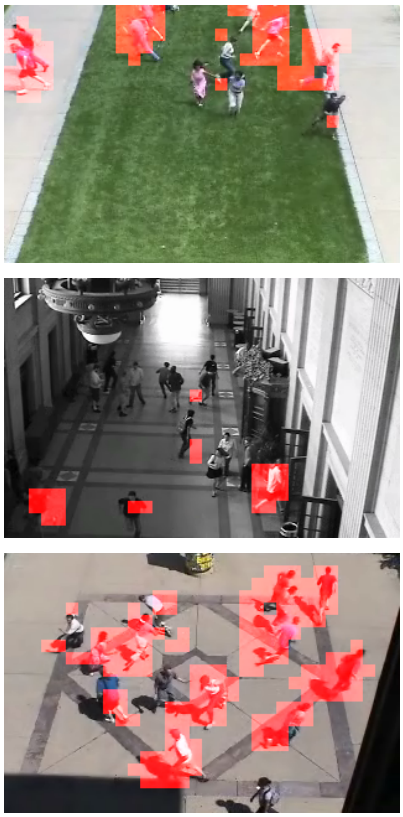


Fig. 13 The localization of the abnormal behaviors on different escape frames. Red blocks correspond to the highly probable abnormal regions.

The ROC curves in Figure 14 illustrate the performance of our approach on the UMN dataset. Figures 14(a), 14(b) and 14(c) show the ROC curves of our method on the 3 different scenes from the dataset and for all scenario. Obviously, the results obtained on scene 1 and scene 3 are much better than those obtained on scene 2. In fact, the performance of our system depends on the performance of the background subtraction and the motion analysis algorithms. Since scene 1 and scene 3 are outdoor scenes, they are well lighted which improves the quality of both the extracted foreground and the motion flow. To evaluate the overall performance of our approach on the UMN dataset, our results are compared to those provided by two referenced algorithms [16], that use social force flow (noted *Force Flow*) and optical flow (*Optical Flow*) to train a latent Dirichlet allocation (LDA) model.

Even if our algorithm has not been designed specifically for this scenario, it is competitive with these state-of-art methods, and it gets better performance for low false alarm probabilities.

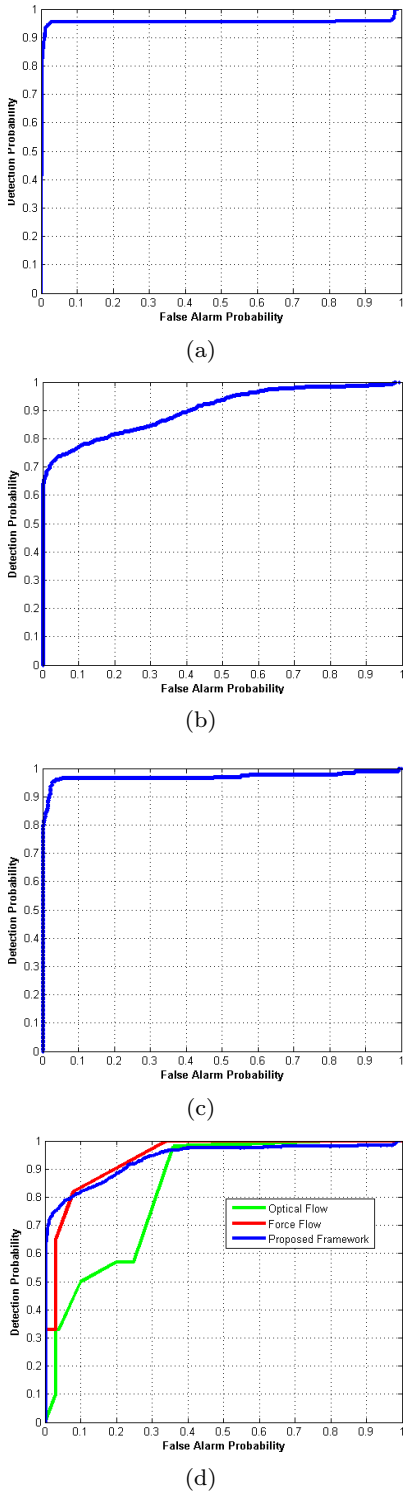


Fig. 14 The ROC curves for the detection of abnormal frames in (a) scene 1, (b) scene 2, (c) scene 3 and (d) the whole UMN dataset, with a comparison to state-of-the-art.

9.2 UCSD Ped Dataset

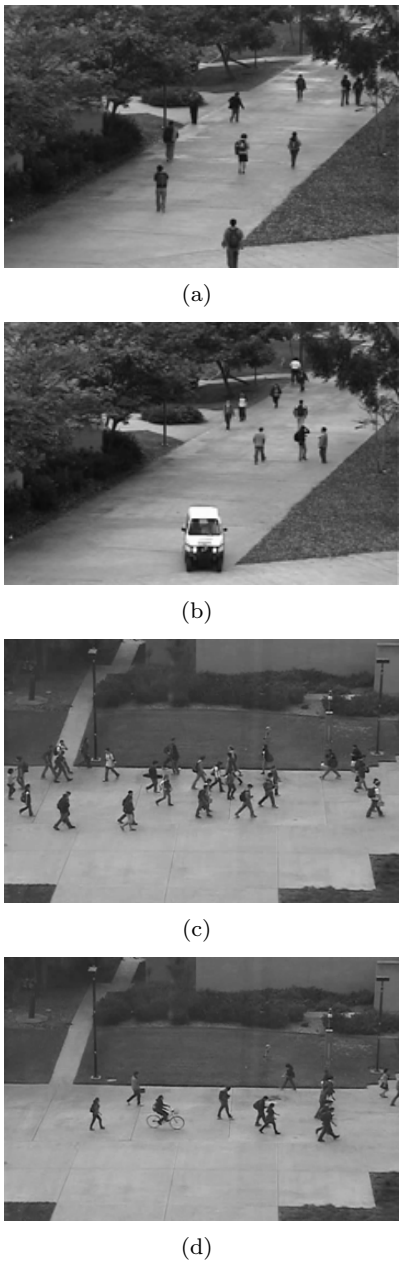
This dataset was acquired with a stationary camera mounted at an elevation, overlooking pedestrian walkways [3]. The crowd density in the walkways was variable, ranging from sparse to very crowded. In the normal setting, the video contains only pedestrians. Abnormal events are due to either 1) abnormal speed, 2) circulation of non pedestrian entities in the walkways, 3) anomalous pedestrian motion patterns or 4) circulation in forbidden areas. Commonly occurring anomalies include bikers, skaters, small carts, and people walking across a walkway or in the grass that surrounds it. A few instances of people in wheelchair were also recorded. The data was collected from two different scenes and split into 2 subsets. The first scene contains groups of people walking towards and away from the camera while the second contains scenes with pedestrian movement parallel to the camera plane. Figure 15 shows some sample frames of these two scenes.

In this dataset, the people actions are more realistic compared to the UMN dataset because all abnormalities occur naturally, they were not staged or synthesized for dataset collection. As for many real video-surveillance scenarios, the pedestrians' normal behavior can be modeled along one principal direction by getting the flow main direction using Principal Component Analysis as explained in Section 5.2.

Figure 16 displays the local results on some abnormal event frames. Figures 16(a), 16(c) and 16(d) highlight respectively a cart, a bicycle and a skateboard. These *vehicles* are detected since they are running faster than the pedestrians and they don't match the normal motion models. In Figure 16(b) a pedestrian who walks on the grass is signaled as an abnormality because the system has detected a motion in a motionless area.

The ROC curves in Figure 17 illustrate the performance of our approach in comparison to several other referenced approaches on the UCSD dataset, coming directly from [23] for the first scene and from [22] for the second scene. Table 2 provides a short explanation about the methods under consideration. The area under the ROC curve (AUC) and the equal error rate (EER) of the different approaches on the first scene are reported in Table 3. It is shown that our framework achieves satisfactory performance even if it has not been designed specifically for this scenario. It outperforms 6 up to 8 of them depending on the level of false alarm (see Figure 17(a)).

More precisely, our system outperforms the model of mixture of dynamic texture MDT [3] on the first scene, and the results are a little bit worse on the second scene. The relatively bad results of the second scene are probably due to the limited size of the training dataset (60% smaller than the dataset provided for first scene).



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

Fig. 15 Sample frames from the UCSD dataset. (a) Normal sample from the first scene. (b) Sample from the first scene with abnormalities. (c) Normal sample from the second scene. (d) Sample from the second scene with abnormalities.

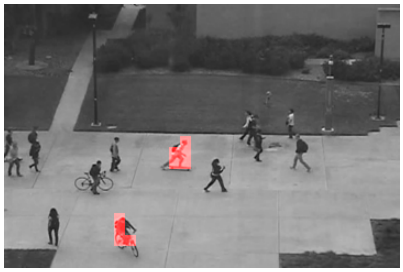
9.3 BEHAVE Dataset

To show the extension of the system to events categorization, experiments have been conducted on the BEHAVE Dataset which consists of a large number of

(a)

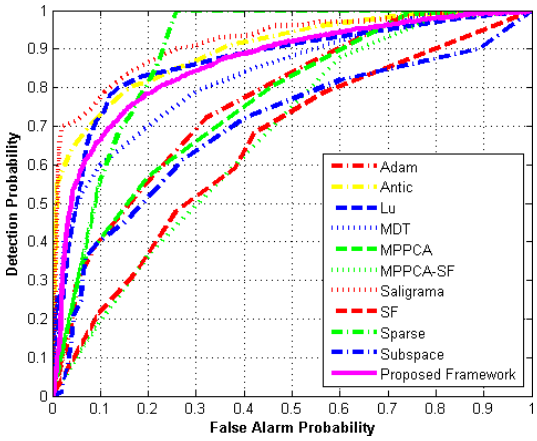
(b)

(c)

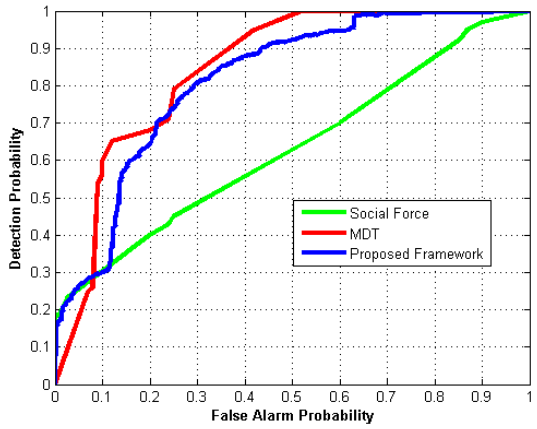


(d)

Fig. 16 Localization of the abnormal behaviors on different UCSD frames, where red blocks correspond to highly probable abnormal regions.



(a) Scene 1



(b) Scene 2

Fig. 17 The ROC curves for the detection of abnormal frames in the UCSD dataset.

complex group activities and is generally used for human activity classification. In this scenario, a abnormal event is considered when one of the group activity is detected: *InGroup* (IG), *Approach* (A), *WalkTogether* (WT), *Split* (S), *Ignore* (I), *Following* (FO), *Chase* (C), *Fight* (FI), *RunTogether* (RT), and *Meet* (M). Some samples are shown in Figure 18.

This dataset is useful to check the consistency of our categorization block. For display purpose, we use a simplified signature of four binary values. The first value, called **F**, is set to 1 if any unmatched flow model is detected, *i.e.* it aggregates the last six fields of the original signature. The value **LV** indicates a low variance and the value **HV** highlights a high variance. Finally, the value **S** stands for a high proportion of the static area in the foreground. The distribution of the seven most detected signatures is illustrated in Table 4.

To get a deep insight into the values presented in Table 4, they are displayed in pie graphs in Figure 19. As it can be seen, most of the presented events can be mainly characterized by one, two or three signatures. In fact, each event may be divided into different *sub-events*. For example, *Chase* event is mainly characterized by three signatures: **F**, **F-LV** and **F-HV**. All these events produce an abnormal event in the flow space. An abnormal variance can also be detected depending on the situation of the chased and chasing people. If the chased person or group is caught, we may detect an abnormal low variance feature. In the other case, if the chased person or group is faster than the chas-

ing group, an abnormal high variance is detected. Also, within this dataset, a frame can contain different events which make the characterization sometimes ambiguous. For instance, *InGroup* scenes, which should have a low variance and a high foreground static area, may correspond to *Approach* or *Split* events which may increase unexpectedly the variance and decrease the portion of the foreground static area.

9.4 SPY project

The system has been applied in the ITEA2 SPY European project[5] where several image processing methods have been ported to an embedded system (ARM Cortex A9) for video surveillance. Figure 20 shows a few snapshots of the system installed at two different locations and using two different sensors. In each case, the training has been performed on a 2 min video.

10 Conclusions

A modular system has been designed to detect and characterize abnormal events in videos. This method is based on a global analysis of the dynamic statistics of the foreground binary pixels and a spatio-temporal analysis of the motion in pixel blocks. Regarding the foreground detection, the background subraction method has been enhanced using optical flow to keep memory of temporarily static pixels. Considering local activity, a normalcy Mixture of Gaussians model is built on each block and for each flow feature and is made more compact using principal components analysis. A set of compact hybrid histograms is used to characterize qualitatively the local activity in each block. It embeds both optical flow orientation and foreground statistics.

An event is considered as normal if it fits all of the designed models, otherwise an abnormal behavior is reported. In addition, the method is able to precisely locate the abnormal area into the crowd, then it provides a binary vector of maximum 13 features that can be used for activity categorization. The system is modular in the sense that each module independently provides a normality/abnormality information. Due to the modularity of the system, the execution can be accelerated on multi-core architectures. The method has been succesfully used in a European project on two differents acquisition systems.

The method involves several parameters that have been chosen so as not to be critical. The values specified in the documents should applied to most videos. Even though the proposed approach provides promising results on different scenarios, improvements can be made. An abnormality detected on actions can depend on the size of the object in the image, which depends on its distance from the camera. In the context of a static camera, a calibration of the scene could be performed in order to adapt the decision rules depending on the location in the image. Furthermore, to categorize an abnormal event, a binary vector is used. It could be replaced by a vector of distances to the pre-defined normal events, which may improve the precision of the categorization results.

[5] Surveillance imProved sYstem https://itea3.org/project/spy.html

Acknowledgment

This research is supported by the European Project ITEA2 SPY Surveillance imProved sYstem https://-itea3.org/project/spy.html

References

1. T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, 2008.
2. B. Antic and B. Ommer, "Video parsing for abnormality detection," in *Int. Conf on Computer Vision (ICCV)*, 2011, pp. 2415–2422.
3. W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2013.
4. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.
5. H. Tabia, M. Gouiffès, and L. Lacassagne, "Motion modeling for abnormal event detection in crowd scenes," in *International Symposium on signal, Images, Video and Communications*, University of Valenciennes, France, July 2012.
6. C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.
7. T. Zhang, H. Lu, and S. Li, "Learning semantic scene models by object classification and trajectory clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1940–1947.
8. N. T. Siebel and S. J. Maybank, "Fusion of multiple tracking algorithms for robust people tracking," in *European Conference on Computer Vision-Part IV (ECCV)*. London, UK, UK: Springer-Verlag, 2002, pp. 373–387.
9. A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *IEEE conf. on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, 2008, pp. 1–8.
10. F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Comput. Vis. Image Underst.*, vol. 115, no. 3, pp. 323–333, Mar. 2011.

11. S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *IEEE Conf. On Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2054–2060.
12. Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *IEEE conf. on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, 2009, pp. 2458–2465.
13. Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *IEEE conf. on Computer Vision and Pattern Recognition (CVPR), 2011*, 2011, pp. 3449–3456.
14. B. Zhao, L. Fei-Fei, and E. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011*, 2011, pp. 3313–3320.
15. J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2921–2928.
16. R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conf. on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, 2009, pp. 935–942.
17. X. Cui, Q. Liu, M. Gao, and D. Metaxas, "Abnormal detection using interaction energy potentials," in *IEEE conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3161–3167.
18. D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, May 1995.
19. L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *IEEE Conf. On Computer Vision and Pattern Recognition, 2009. CVPR 2009*, 2009, pp. 1446–1453.
20. O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vision*, vol. 74, no. 1, pp. 17–31, Aug. 2007.
21. V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2112–2119.
22. V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE conf. on Computer Vision and Pattern Recognition (CVPR), 2010*, 2010, pp. 1975–1981.
23. C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *IEEE Int. Conf on Computer Vision (ICCV)*, Dec 2013, pp. 2720–2727.
24. R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: representing activities as bags of event n-grams," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 1031–1038 vol. 1.
25. D. Zhang, T. Sullivan, C. Briciu Burghina, K. Murphy, K. McGuinness, N. E. O'Connor, A. F. Smeaton, and F. Regan, "Detection and classification of anomalous events in water quality datasets within a smart city-smart bay project." 2014.
26. C. Kruegel, D. Mutz, W. Robertson, and F. Valeur, "Bayesian event classification for intrusion detection," in *Computer Security Applications Conference*. IEEE, 2003, pp. 14–23.
27. S.-C. S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," 2007.
28. M. Harville, "A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models," in *European Conference on Computer Vision-Part III (ECCV)*. London, UK: Springer-Verlag, 2002, pp. 543–560.
29. A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *European Conference on Computer Vision-Part II (ECCV)*. London, UK: Springer-Verlag, 2000, pp. 751–767.
30. T. E. Boult, R. Micheals, X. Gao, P. Lewis, C. Power, W. Yin, and A. Erkan, "Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded targets," in *IEEE Int. Workshop on Visual Surveillance*, 1999, pp. 48–55.
31. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Int. Joint Conference on Artifical Intelligence*, 1981, pp. 674–679.
32. C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University Technical Report, Tech. Rep. CMU-CS-91-132, April 1991.
33. M. Gouiffès, C. Collewet, C. Fernandez-Maloigne, and A. Trémeau, "A study on local photometric models and their application to robust tracking," *Computer Vision and Image Understanding*, vol. 116, pp. 896–907, Apr. 2012.
34. J.-Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," *Intel Corporation, Microprocessor Research Labs*, 2000.
35. M. Gouiffès, B. Planes, and C. Jacquemin, "Htri: High time range imaging," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 361 – 372, 2013.
36. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.
37. K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.
38. D. Defays, "An efficient algorithm for a complete link method." *Comput. J.*, vol. 20, no. 4, pp. 364–366, 1977.
39. P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *European Workshop on Advanced Video Based Surveillance Systems*, vol. 5308, 2001.
40. A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.
41. E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2790–2797.

Table 2 State-of-the-art methods used for comparison in the USCD dataset.

Abbreviation		Meaning of the abbrevation
Adam	[40]	Real-time detection of unusual events based on multiple local monitors which collect low-level statistics.
Antic	[2]	Scene parsing: localization of the abnormalities using statistical inference.
Lu	[23]	Fast sparse combination learning.
MDT	[3]	Mixture of Dynamic Textures models.
SF	[18]	Social force models.
MPPCA	[15]	Space-Time Markov Random Field. Atomic motion patterns are learnt via a Mixture of Probabilistic Principal Component Analyzers.
MPCCA-SF	[3, 15, 18]	Idem where MPCCA is combined with SF Normalized flow.
Saligrama	[21]	Some empirical rules are used to fuse local statistics across spatio-temporal locations and scales and produce a composite score for a video segment.
Sparse	[13]	Sparse reconstruction cost over a normal dictionary.
Subspace	[41]	Subspace clustering based on sparse representation.

Table 3 Equal Error Rate (EER) and Area Under the ROC Curve (AUC) on the UCSD Ped1 dataset.

	SF-MPPCA	SF	MDT	Sparse	Saligrama	Antic	Subspace	Lu	Ours
EER	40%	31%	25%	19%	16%	18%	29.6%	17%	20%
AUC	59%	67.5%	81.8%	86%	92.7%	91%	68.4%	91.8%	86.6%

(a) Fight (b) Approach (c) Split

(d) InGroup (e) WalkTogether (f) Chase

Fig. 18 Sample frames of the BEHAVE dataset.

Table 4 Distribution of the ten most repeated signatures among the ten labeled events.

Signature				Chase	Fight	Run together	In group	Approach	Walk together	Split	Ignore	Following	Meet
F	**LV**	**HV**	**S**										
0	1	0	0	0	82	15	268	110	488	0	52	0	0
0	0	1	0	0	3	0	2	68	227	75	0	0	0
0	1	0	1	0	0	0	321	0	0	0	0	0	0
1	0	0	0	55	117	50	0	2	7	3	1	21	0
1	1	0	0	121	0	0	24	7	21	8	12	0	0
0	0	0	1	0	16	0	149	7	0	0	0	0	0
1	0	1	0	30	43	0	0	3	2	7	0	0	0

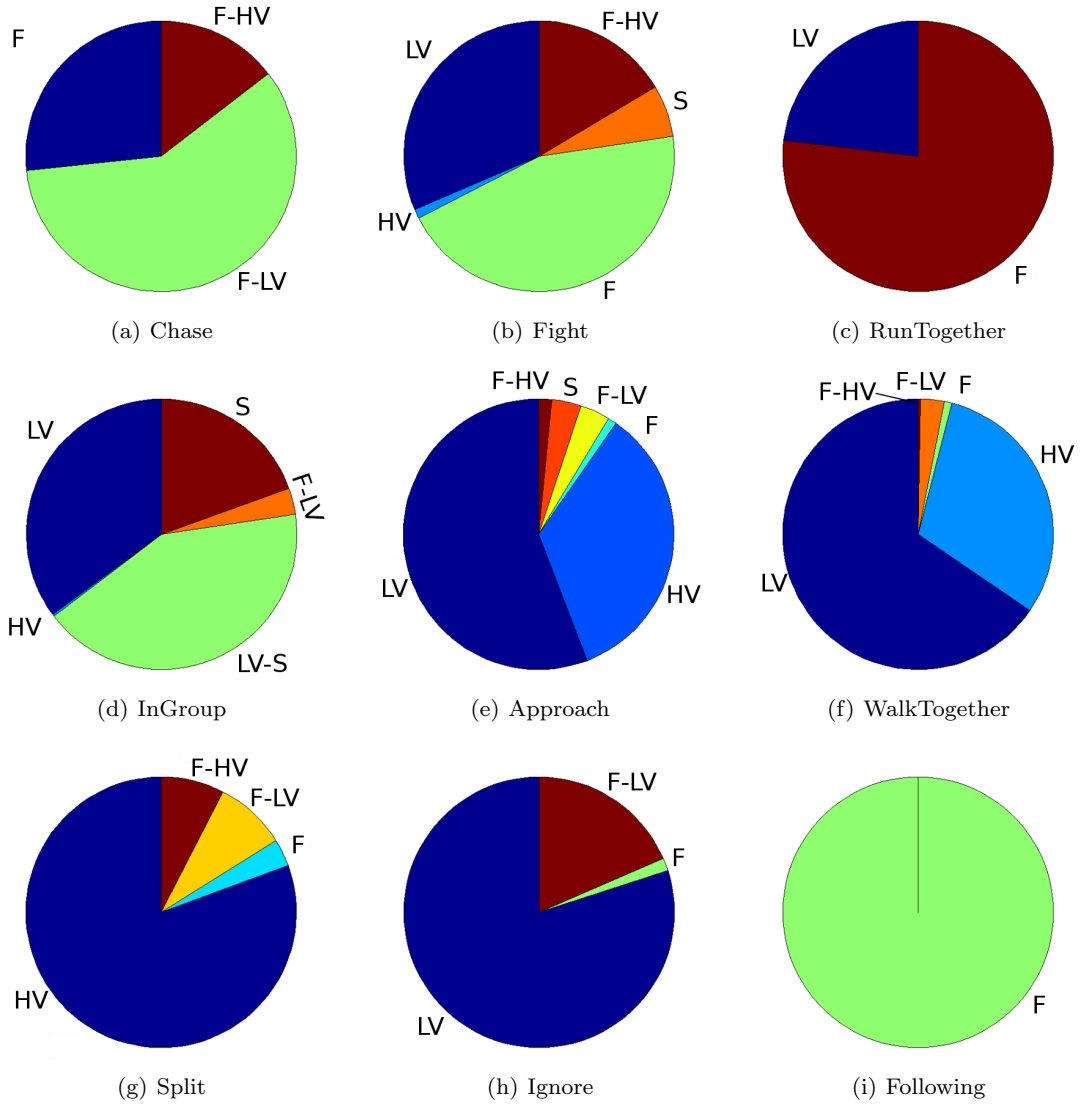
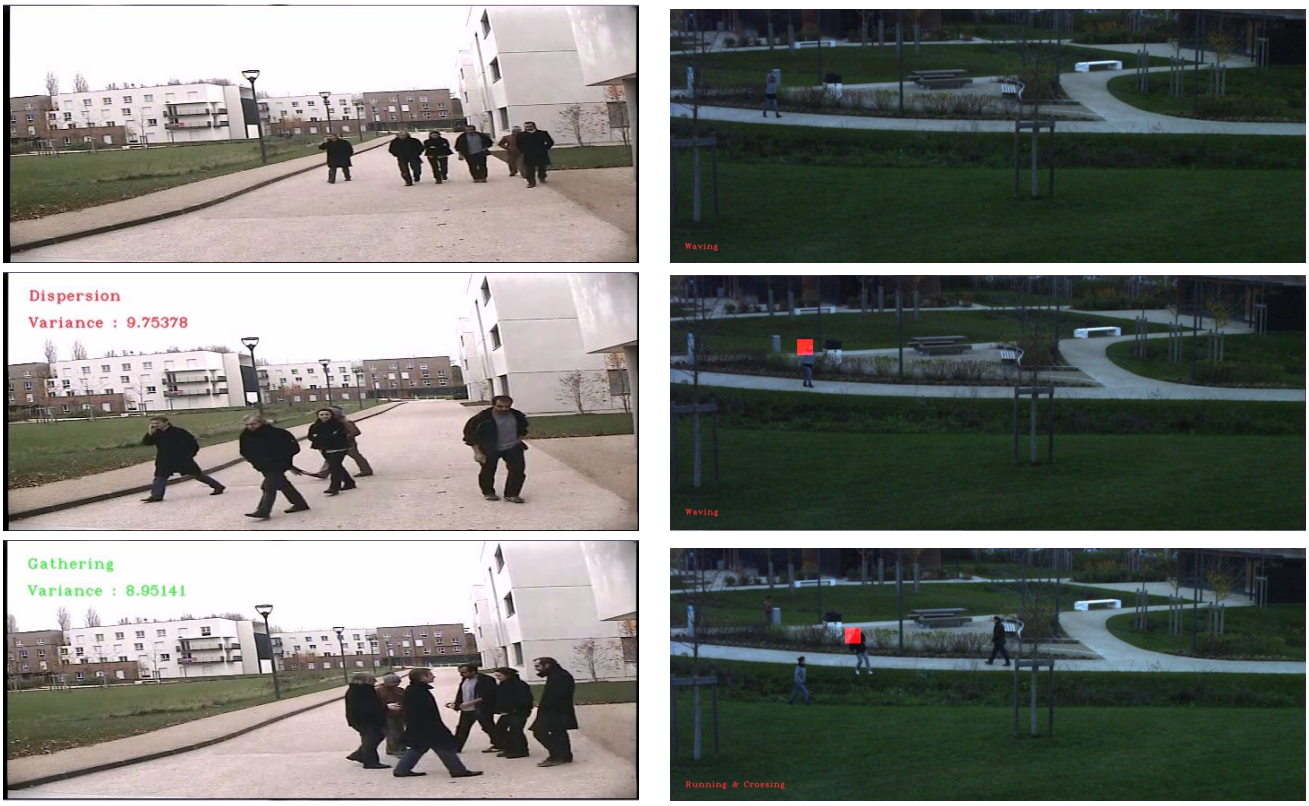


Fig. 19 Distribution of the ten most repeated signatures among nine events.



(a) (b)

Fig. 20 Examples of anomaly detection on images from ITEA SPY project. (a) IP camera 25 fps from a commercialized videosurveillance system, frame size 640 × 240 (1 row over two is not processed in order to reach real-time execution); (b) Camera GigaEthernet, 50 fps, frame of size 1280 × 1024.