Real-Time Abnormal Events Detection Combining Motion Templates and Object Localization

Thi-Lan Le and Thanh-Hai Tran

Abstract Recently, abnormal event detection has attracted great research attention because of its wide range of applications. In this paper, we propose an hybrid method combining both tracking output and motion templates. This method consists of two steps: object detection, localization and tracking and abnormal event detection. Our contributions in this paper are three-folds. Firstly, we propose a method that apply only HOG-SVM detector on extended regions detected by background subtraction. This method takes advantages of the background subtraction method (fast computation) and the HOG-SVM detector (reliable detection). Secondly, we do multiple objects tracking based on HOG descriptor. The HOG descriptor, computed in the detection phase, will be used in the phase of observation and track association. This descriptor is more robust than usual grayscale (color) histogram based descriptor. Finally, we propose a hybrid method for abnormal event detection this allows to remove several false detection cases.

Keywords Video analysis · Event recognition · Object detection and tracking

1 Introduction

Recently, abnormal event detection has attracted great research attention in computer vision because of its wide range of applications such as elderly surveillance, patient smart room. In general, previous approaches for abnormal event detection can be categorized into two groups: tracking-based and motion-based approaches. The tracking-based approaches [1] focus on the analysis of the trajectories of moving objects. Recently, motion-based approaches have been proposed to address the above problem [2, 3]. These approaches try to extract the motion features in order to recognize the events of interest. In this paper, we propose an hybrid method

T.-L. Le (✉) · T.-H. Tran
International Research Institute MICA, HUST - CNRS/UMI-2954 - Grenoble INP
and Hanoi University of Science and Technology, Hanoi, Vietnam
e-mail: Thi-Lan.Le@mica.edu.vn

Q.A. Dang et al. (eds.), *Some Current Advanced Researches on Information and Computer Science in Vietnam*, Advances in Intelligent Systems and Computing 341, DOI 10.1007/978-3-319-14633-1_2

combining both tracking output and motion templates. The contributions of this paper are three-folds. Firstly, we propose a method that apply only HOG-SVM detector on extended regions detected by background subtraction. This method takes advantages of the background subtraction method (fast computation) and the HOG-SVM detector (reliable detection). Secondly, we do multiple objects tracking based on HOG descriptor. The HOG descriptor, computed in the detection phase, will be used in the phase of observation track association. This descriptor is more robust than usual grayscale (color) histogram based descriptor. Finally, we propose a hybrid method for abnormal event detection this allows to remove several false detection cases. We have evaluated our system in person surveillance application and the experimental results obtained with 20 subjects are promising. The remaining of this paper is organized as follows. In the Sect. 2, we describe in detail our method from object detection, tracking to abnormal event detection. The experiment and the obtained results are discussed in Sect. 3. The Sect. 4 gives some conclusions and future works.

2 Proposed Abnormal Event Detection Systems

2.1 Overview

The objective of the project is to do daily surveillance for people with special need for example elderly people in their house/room by using different technology (vision, audio and RFID). The main aim of vision in this project is to track, localize the person in the room and to detect some events of interest. The flowchart of our work is illustrated in Fig. 1. This system has two main modules: person detection, tracking and localization and event detection. The intermediate results of person detection and event detection will be stored for further analysis. The result of event detection module can be used to provide different service (e.g. fall alarm by mobile phone). In this section, we will describe in detail these modules.

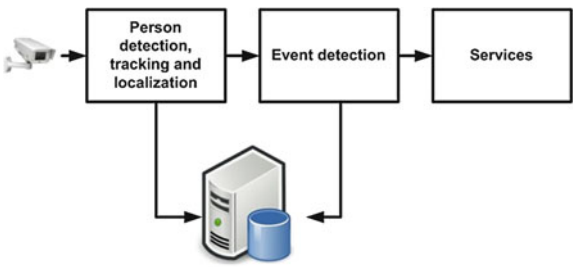

Fig. 1 Overall of vision-based person surveillance system

2.2 Object Detection and Localization

In this work, the term object means people in the room. Objective of this part is to detect the presence of people and their location in 2D image plane as well as 3D room space (real world). To be able to obtain this objective, the following works need to be carried out: background modeling, human detection and human tracking.

In the literature, there exists a lot of works for human detection. The simplest method is background subtraction. The main advantage of this method is that it is very fast in computation, so suitable for real-time application such video surveillance. However, this method gives sometime false alarms (due to the movement of some objects in the scene). Moreover, the localization of the detected objects (bounding box) is not precise.

Recently, Dalal and Triggs [4] has proposed a very efficient method for standing human detection that is considered nowadays a baseline method for comparison. This method introduced a new feature HOG (Histogram of Oriented Gradients) to represent a human and used SVM (Support Vector Machine) technique for learning the human model. At detection phase, the human non human classifier will be applied on each sliding window to finally detect human from background. This detector has been shown to be very efficient on several datasets. However, one of the biggest drawback of this method is it is very time consuming.

In the context of our work, camera and all background objects are fix, only human are moving objects in the scene. We then propose a method that combines advantages of two methods while reducing their drawbacks. Our proposed method composes of three main steps:

- Background modeling: learn fix background using codebook technique [5].
- Moving object detection: detect moving objects in the scene (human detection) by combining HOG-SVM detector and background subtraction technique [4].
- Human tracking: track human during time using Kalman filter [6].

In the following, we will describe in more detail each component in the module object detection and tracking (see Fig. 2).

Background modeling. The object detection assures an automatic initialization of the tracker as well as provides observations for data association. In our context, the camera is fix but scene can contain moving background (like waving curtain) and illumination variations. Using the simplest background subtraction technique can not to handle this problem. To handle to this, we proposed to use segmentation technique based on codebook technique [5].

The main idea of the method is to construct background model from long training sequences. Each pixel of background will be encoded by a codebook composed from several codewords computed from color distortion metric together with brightness bounds. By this way, it allows to build adaptive and compact background model that captures the structural background motion in a long time and the ability to cope with local and global illumination change. The reason that we choose this algorithm for background modeling is that this method is experimentally shown to be more efficient

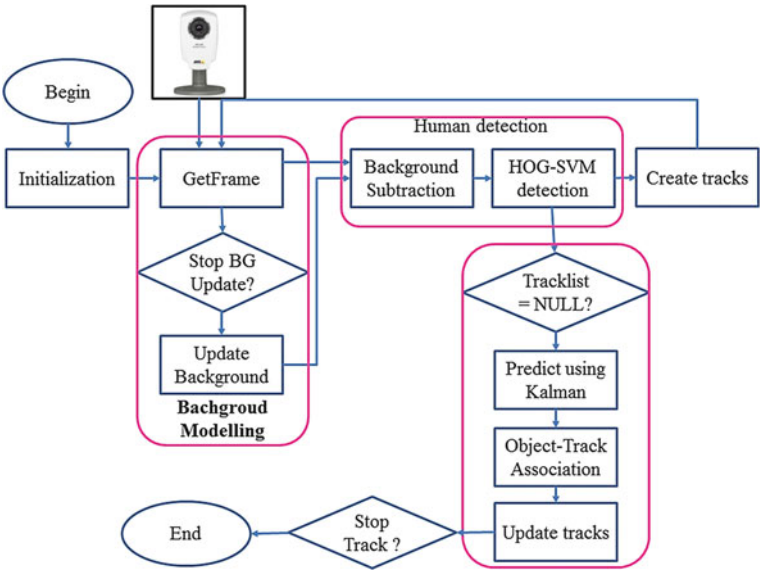

Fig. 2 Diagram of object detection and tracking

in time/memory and precision, meaning some moving elements in the background is considered as background. For technical detail, see the original paper [5].

We have compared the background subtraction technique with the codebook algorithm on a video (including 353 frames) captured from the indoor scene in which one person is walking under neon and daylight condition. The criteria for comparison are precision/recall and computational time. The precision/recall is computed in term of true positive, false positive, false negatives detected using the results of background and codebook algorithm. Table 1 shows the results of comparison. We notice that the codebook algorithm gives a little better precision and recall while it is significantly fast in computational time.

Moving object detection. Once the background model is built, given each video frame, the moving objects detection is carried out by differencing the current image with the background model. To remove noises, we threshold the different image. Morphological operators are then used, followed by connected component analysis to group pixels into blobs.

Table 1 Comparison of background subtraction and codebook algorithms

Algorithm	TP	FP	FN	Computational time per frame (ms)	Precision (P) % Recall (R) %
Codebook	130	2	27	13	P = 99 R = 83
Background subtraction	125	0	34	180	P = 100 R = 79

Fig. 3 Detection results obtained from differencing *current frame* with *background frame*

We can observe in Fig. 3 that using background subtraction, people are detected but their localization is not really perfect: the bounding boxes are mostly bigger than human. Sometime, a part of background is considered as a false alarm. To remove this kind of false alarms, first, we extend all bounding boxes then apply HOG-SVM based human detector on each extended bounding box for verification. We notice also that by this way, we can remove some false alarms occurred when we apply HOG-SVM detector on the whole image.

In addition, to avoid missed detection caused by HOG-SVM, we will keep detection that satisfy conditions to be still a human (ratio between width and height, percentage of foreground pixels and the bounding box) to keep tracking longer (Fig. 4).

Human Tracking. For tracking human, we propose to use the traditional Kalman filter that has been shown to be good enough in lot of surveillance applications. In our work, the state vector in Kalman model composes of 8 elements corresponding to: coordinates of the weight center (x, y), velocity of the weight center in two direction (v_x, v_y), the size of the bounding rectangle (b_w, b_h), and change of this size during time (v_{bw}, v_{bh}). The observation vector composes of 4 elements: coordinates of the weight center (x, y) and the size of the bounding rectangle (b_w, b_h). Observation and process noise are supposed as white noise with Gaussian distributions. The state

Fig. 4 **a** Detection results (*black rectangles*) by applying HOG-SVM on whole image. **b** Detection results (*red rectangle*) by applying HOG-SVM on the extended region (*green rectangle*). The false alarm in (**a**) (*smaller rectangle*) is now removed in (**b**), the localization of human is more precise

transformation of Kalman filter in our work is presented in following equation where $(x', y', v'_x, v'_y, b'_w, b'_h, v'_{bw}, v'_{bh})$ are the estimated values of the state vector.

$$\begin{pmatrix} x'_{t+1} \\ y'_{t+1} \\ vx'_{t+1} \\ vy'_{t+1} \\ bw'_{t+1} \\ bh'_{t+1} \\ vbw'_{t+1} \\ vbh'_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta_t & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta_t & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \Delta_t & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \Delta_t \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x'_t \\ y'_t \\ vx'_t \\ vy'_t \\ bw'_t \\ bh'_t \\ vbw'_t \\ vbh'_t \end{pmatrix} + N(0, Q) \tag{1}$$

The relation between observation and state vectors is presented as follows.

$$\begin{pmatrix} x_{t\,obs} \\ y_{t\,obs} \\ bw_{t\,obs} \\ bh_{t\,obs} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x'_t \\ y'_t \\ vx'_t \\ vy'_t \\ bw'_t \\ bh'_t \\ vbw'_t \\ vbh'_t \end{pmatrix} + N(0, R) \tag{2}$$

$N(0, Q)$ and $N(0, R)$ are process noise and observation noise respectively is assumed to be drawn from a zero mean multivariate normal distribution with covariance Q and R.

In our case, we would like to build a multiple human tracking, so we need to do a more complex track observation association. At a time t, we propose to rank tracks in function of its score (it is a function of track length and detection confidence). Then, an observation will be associated first to a track that has the biggest score remaining in the list. The association between a track and an observation will be selected based on a match measure that is the Euclidian distance between two HOG descriptors. After each association, the track and the observation have been pop out from the list. This is looped until all tracks find its observations. If a track does not find an observation (missed detection), we keep this track in several frames until it find an observation in the next frame. After important missed observations, we delete this track. For all remaining observations, we create new tracks.

2.3 Abnormal Event Detection

There are a lot of events of interest that are needed to recognize in our work. However, we focus on the following events: (1) Person falls from the bed or during walking; (2) Person lays motionless in the floor; (3) Person stays too long in the rest room;

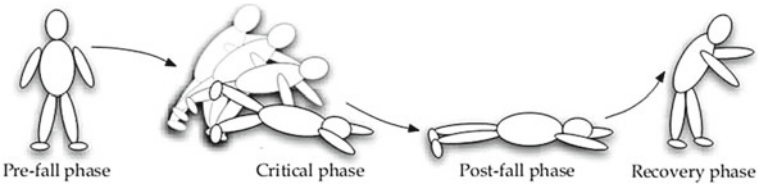

Fig. 5 Four phases of fall event [3]

Fig. 6 Person falls from the bed

(4) Person is out of the room too long. Among these four events, the first event (fall) has attracted many works in the computer vision community. According to [7], this event can be decomposed in four phases (see Fig. 5).

The pre-fall phase corresponds to daily life motions, with occasionally sudden movements directed towards the ground like sitting down or crouching down. The critical phase, corresponding to the fall, is extremely short. This phase can be detected by the movement of the body toward the ground or by the impact shock with the floor. The post-fall phase is generally characterized by a person motionless on the ground just after the fall. It can be detected by a lying position or by an absence of significative motion. A recovery phase can eventually occur if the person is able to stand up alone or with the help of another person. Figure 6 illustrates the fall event.

There are a number of works have been proposed for fall event detection. These works can be divided into two categories. The works belonging to the first category try to model and to recognize the fall events by using finite state machine, HMM (Hidden Markov Model) [8] while the second compute the motion templates such as MHI (Motion History Image) [3]. In this paper, for the fall event, inspired the work of Rougier et al. [3], we propose a fall event algorithm combining both object localization output and MHI.

Among 4 events of interest, the third and the fourth events ("Person stays too long in the rest room" and "Person is out of the room too long") are inferred directly from the output of object localization. The first and the second events ("Person falls from the bed or during walking" and "Person lays motionless in the floor") are recognized as described in Fig. 7.

For the fall event detection, the main difference of our method and that of Rougier et al. [3] is the Person detection, tracking and localization step. Based on this step, we verify the hypothesis: "is the person on the bed?". The result of this verification allows

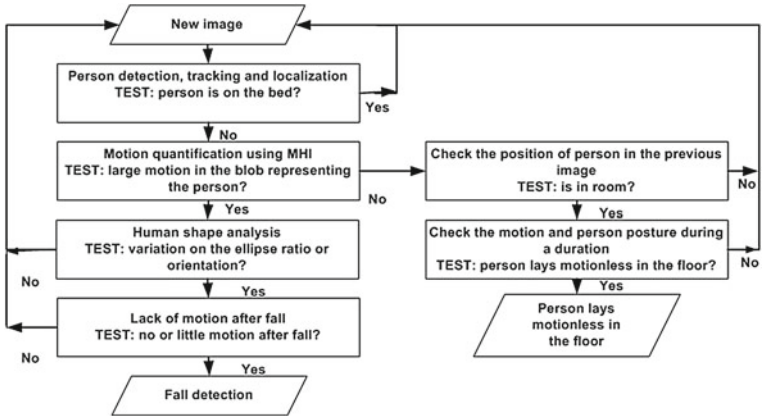

Fig. 7 "Person falls from the bed or during walking" and "Person lays motionless in the floor" recognition algorithm

to remove false detection because if person lays motionless in the bed, this is normal situation. Moreover, if the system knows that the person is on the bed, it does not need to do fall event detection. As discussed in the related work, MHI is introduced in [9]. The Motion History Image (MHI) is an image representing the recent motion in the scene, and is based on a binary sequence of motion regions $D(x, y, t)$ from the original image sequence $I(x, y, t)$ using an image-differencing method. Then, each pixel of the Motion History Image $H\tau$ is a function of the temporal history of motion at that point, occurring during a fixed duration τ (with $1 \leq \tau \leq N$ for a sequence of length N frames):

$$H\tau(x, y, t) = \begin{cases} \tau \text{ if } D(x, y, t) = 1 \\ \max(0, H\tau(x, y, t-1) - 1 \text{ otherwise} \end{cases} \tag{3}$$

The more recent moving pixels are seen brighter in the MHI image. Then, to quantify the motion of the person, we compute a coefficient C motion based on the motion history (accumulation of motion during 500 ms) within the blob representing the person (output of person detection) using:

$$C = \frac{\sum_{\text{pixel}(x,y) \in \text{blob}} H\tau(x, y, t)}{\#\text{pixels} \in \text{blob}} \tag{4}$$

Figure 8 illustrates the value of total intensity of MHI that is $\sum_{\text{pixel}(x,y) \in \text{blob}} H\tau(x, y, t)$ for a sequence of human activity: walking (from 0 to T1), falling (from T1 to T2) and being motionless on the floor (from T2 to T3). This figure shows that we can distinguish two events "Person falls from the bed or during walking" and "Person lays motionless in the floor" with other events of interest.

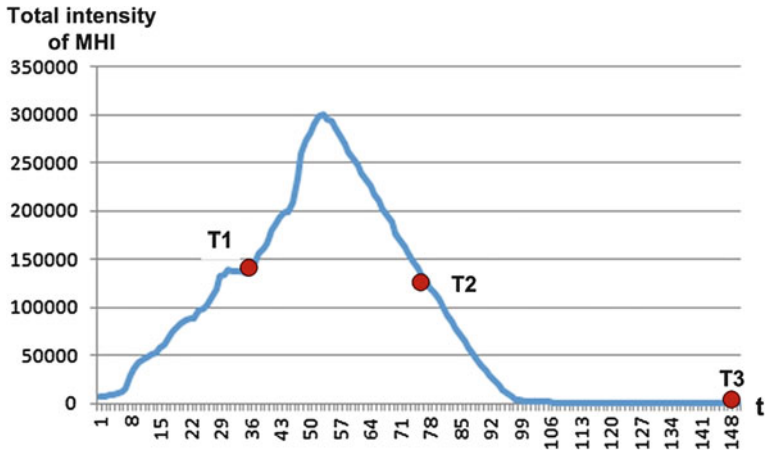

Fig. 8 Illustration of total value of MHI computed for a sequence of human activity: walking (from 0 to T1), falling (from T1 to T2) and being motionless on the floor (from T2 to T3)

The person is then approximated by an ellipse defined by its center $\overline{x}, \overline{y}$, its orientation θ and the length a and b of its major and minor semi-axes. The approximated ellipse gives us information about the shape and orientation of the person in the image. Two features are computed for a 1 s duration to analyze the human shape change:

- The orientation standard deviation δ_θ of the ellipse: If a person falls perpendicularly to the camera optical axis, then the orientation will change significantly and δ_θ will be high. If the person just walks, δ_θ will be low.
- The $\delta_{a/b}$ ratio standard deviation of the ellipse: If a person falls parallelly to the camera optical axis, then the ratio will change and $\delta_{a/b}$ will be high. If the person just walks, $\delta_{a/b}$ will be low.

3 Experimental Results

3.1 Experiment Description

To evaluate our algorithm, we need to setup environment, camera and define scenarios. We carry out experiments at the show room of MICA Institute. Table 2 gives some information of our testing environment while Fig. 9 shows the layout of the room.

With this layout, in order to have a good observation, we have installed two camera (see the position of Cam 1 and Cam 2 in Fig. 10). The camera 1 (Cam 1) allows to monitor the main door while the camera 2 (Cam 2) observe the region in the room

Table 2 Description of the experimental environment

Size (length × width × height)	9.2 m*8.8 m*3 m
Main door	01
Windows	02
Toilet inside	01
Objects	Bed, medical cabinet
Lighting condition	Neon and daylight through windows and floor

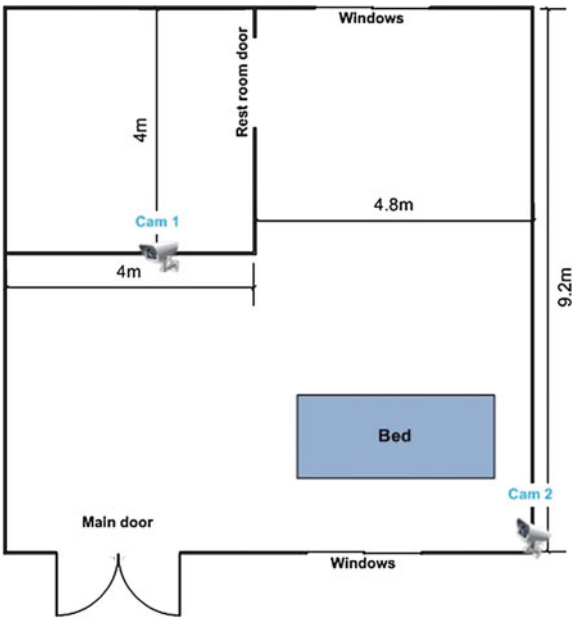

Fig. 9 Room layout

Fig. 10 Testing room with the position of cameras

from the bed to the rest room. Because of the privacy reason, we do not install the camera in the rest room. We perform the object detection and event recognition modules for each camera. The results obtained from two cameras will be fused.

The number of subjects participation to the experiments are 20, aging from 25 to 40 years old. Subjects are asked to do the following scenario 5 times. In the scenario, the person is asked to (1) Enter to the room; (2) Sit on the bed; (3) Lay on the bed; (4) Lying motionless on the bed; (5) Fall from the bed; (6) Lying motionless on the floor; (7) Get up and walk to the table; (8) Go toward the toilet; (9) Get into the toilet; (10) Staying long in the toilet; (11) Get out; (12) Fall on the floor; (13) Get out the room.

The main objective to do this scenario is to collect all our events of interest. The order of each steps in this scenario is not important since it does not influence the performance of our system. We perform the system in the computer with the following configuration: Intel(R) Core(TM) i5-2520M CPU @ 3.2 GHz × 2, RAM 4 GB. The results obtained with our system are stored in a log file. This file is used for performance analysis.

3.2 Object Detection and Localization Results

To evaluate the object detection and tracking module performance, we measure two criteria: Precision and Recall. These measures are computed as follows:

$$Precision = \frac{tp}{tp + fp} \quad \text{and} \quad Recall = \frac{tp}{tp + fn} \tag{5}$$

where tp (true positive) is the number of correct detections; fp (false positive) is the number of false alarms; fn (false negative) is the number of missed detections. We consider a correct detection if the intersection between its bounding box and the ground truth one is bigger than 50. We have compared our proposed method (HOG-SVM applied on extended bounding box provided by Background subtraction) with the original one (HOG-SVM applied on the whole image) in terms of computational time and precision/recall (see Table 3). Experiments are carried out on 353 frames containing a person walking in the room. The frame resolution is 640 × 480. We can observe that our proposed method has removed a lot of false detections while still keeping correct detections. In addition, the computational time is significantly reduced. The detection and localization results with 20 subjects playing the predefined scenario is shown in the Table 4. We can observe that this method give very high detection rate. In addition, the detection takes only 76.5 ms in average (13 fps) so ensure that our system could run in real-time. Figure 11 shows an example of human tracking. We can see that the track results is quite consistent. It still works well in case of multiple persons in the scene. When one person obscures other in a short time, our method still keep track both.

Table 3 Comparison of our detection method with HOG-SVM based method

Method	Precision (%)	Recall (%)	Computational time (ms)
Dalal et al.	72.5	88.5	376
Our proposed method	99.5	87.5	76.5

Table 4 Detection results of our proposed method obtained with a big dataset

TP	TN	FP	FN	Precision	Recall
27,870	9,328	6,279	856	0.86	0.96

Fig. 11 Object detection and tracking results. The *red line* presents the movement trajectory of the human

The proposed method for human detection and tracking can be applied in case of multiple people. Figure 12 shows the results when we apply the method on 3 sequences (two sequences are collected in MICA showroom, the third one comes from CAVIAR dataset (http://homepages.inf.ed.ac.uk/rbf/CAVIAR/).

Fig. 12 Object detection and tracking results with *one* and *multiple* people

3.3 Event Detection Results

To evaluate the event detection performance, we measure two criteria below:

$$F.A.R = \frac{fp}{tp + fp} \quad \text{and} \quad Sensitivity = \frac{tp}{tp + fn} \tag{6}$$

where tp (True Positive) is the number of correct events detected; fp (False Positive) is the number of wrong events detected, and fn (False Negative) is the number of lost events. The smaller F.A.R and the greater Sensitivity are, the better system is. The obtained results are shown in Table 5.

The total number of 4 interested events for each times is 120 (40 Event 1, 40 Event 2, 20 Event 3 and 20 Event 4). The experimental results show that our algorithm obtains the best result in term of Sensitivity with "Person lays motionless in the floor". However, the value of F.A.R obtained for this event is also high. This result is acceptable in the context of surveillance system for peoples with special need because this event is important and the lost of this event can cause the major heath problem. The result also shows that our algorithm detects well the fall event with a small value of F.A.R. Since our fall detection algorithm is based on MHI, it is invariant to human shape change in fall event. However, this algorithm decides the fall event by comparing the value of MHI with a threshold. If we set this threshold low, it may detect some like-fall event. The recognition results of "Person stays too long in the rest room" and "Person is out of the room too long" are relatively good. Since, these events are recognized by using the results of object localization and tracking module. The bad results of this module may lead the wrong recognition (Table 5).

Table 5 The obtained sensitivity and false alarm rate of 4 events of interest with 20 subjects in 5 times (Event 1: Person falls from the bed or during walking; Event 2: Person lays motionless in the floor; Event 3: Person stays too long in the rest room; Event 4: Person is out of the room too long)

Measure	Event 1	Event 2	Event 3	Event 4
Sensitivity at times #1	0.88	0.98	0.75	0.80
F.A.R at times #1	0.00	0.11	0.00	0.08
Sensitivity at times #2	0.88	0.9	0.75	0.85
F.A.R at times #2	0.07	0.15	0.00	0.13
Sensitivity at times #3	0.93	0.95	0.80	0.85
F.A.R at times #3	0.02	0.02	0.00	0.06
Sensitivity at times #4	0.93	0.98	0.95	0.80
F.A.R at times #4	0.00	0.03	0.00	0.00
Sensitivity at times #5	0.93	0.95	0.80	1.00
F.A.R at times #5	0.00	0.02	0.00	0.05
Average sensitivity	**0.91**	**0.95**	**0.81**	**0.86**
Average F.A.R	**0.018**	**0.067**	**0.00**	**0.064**

4 Conclusions and Future Works

In this paper, we have introduced a real-time abnormal events detection system combining motion templates and object localization. The proposed system is able to recognize four abnormal events. The experimental results with 20 subjects have been proved the robustness of the proposed system (high value of sensitivity and low value of false alarm rate). However, the object detection and localization part is still sensitive to illumination change and the abnormal event is based on the chosen threshold. Moreover, our system bases on the assumption that the surveillance room contains only one sole person. In the future, we would like extend our work by combining with other sensor types such as Kinect sensor and by recognizing other types of events.

Acknowledgments The research leading to this paper was supported by the National Project B2013.01.41 "Study and develop an abnormal event recognition system based on computer vision techniques". We would like to thank the project and people involved in this project.

References

1. Basharat, A., Gritai, A., Shan, M.: Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1–8 June 2008
2. Benezeth, Y., Jodoin, P.M., Saligrama, V., Rosen-berger, C.: Abnormal events detection based on spatio-temporal co-occurences. 2013 IEEE Conference on ComputerVision and Pattern Recognition, pp. 2458–2465 (2009)
3. Rougier, C., St-Arnaud, A., Rousseau, J., Meunier, J.: Video surveillance for fall detection. In: Lin, P.W. (ed.) Vid. Surveill. (2011)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 05), vol. **1**, pp. 886–893 June 2005
5. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-time foreground-background segmentation using codebook model. Real-Time Imaging **11**(3), 172–185 (2005)
6. Ribeiro, M.I.: Kalman and extended Kalman filters: concept, derivation and properties. Technical report (2004)
7. Noury, N., Rumeau, P., Bourke, A., Laighin, G., Lundy, J.: A proposal for the classification and evaluation of fall detectors. (IRBM) **29**(6), 340–349 (2008)
8. Vishwakarma, V., Mandal, C., Sural, S.: Automatic detection of human fall in video. In: Ghosh, A., De, R., Pal, S. (eds.) Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science, vol. 4815, pp. 616–623. Springer, Heidelberg (2007)
9. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)