ABNORMAL PEDESTRIANS ACTIVITIES RECOGNIZER AND TRACKER

V. VIRILI[1], F. GARZIA[1,2] & R. CUSANI[1]
[1]Department of Information, Electronics and Telecommunication Engineering,
SAPIENZA – University of Rome, Rome, Italy.
[2]Wessex Institute of Technology, Ashurst Lodge, Ashurst, Southampton, UK.

ABSTRACT

The purpose of the present work is to find out a new methodology to automatically detect abnormal situations in high risk places through a video surveillance system. The idea is to retrieve true and predicted movement of the people in the scene then, through a classifier, to map out different abnormal situations comparing proper vectors. To reach its purpose, the proposed methodology uses a multidisciplinary approach.

Keywords: Abnormal activities recognizer, automatic video surveillance, pedestrian tracker.

1 INTRODUCTION

Usually an operator works in front of a lot of monitors alone, and a decrease of concentration after a few hours of work in front of them is demonstrated. Public square, entrances of the underground or an embassy in enemy field need a timely response if something goes wrong. Obviously an abnormal situation depends on an abnormal pedestrian behavior, thus the idea is to classify and then detect these abnormal conducts. Different approaches to understand if something is going wrong in high risk places from a security point of view were studied. In [1, 2] the idea was to analyze sounds coming from the crew: this could be very useful in little/medium closed places where usually people do not shout or speak loudly. Surely this approach is also very useful where a camera (for different reasons) cannot see. Other approaches, instead, use images coming from a camera. In [3] the Helbing Model is used with a grid of particles placed over the images to study the interaction between pedestrians. This social force model represents a part of the proposed system.

More deeply the novelty of the proposed system is to use a classifier based on relationships between true pedestrian's movement and the predicted one. In fact, often, observing abnormal situations videos, a security operator can understand what is going on because someone in the scene does suddenly something of suspicious.

The type of abnormal situation depends on what the person is doing, but the fact that someone suddenly changes his own behavior turns on the operator attention. In the following a system which turns on the attention of an ad-hoc neural network which tries to understand the kind of the abnormal situation on the basis of its previous learning is illustrated.

2 THE PROPOSED SYSTEM

The approach is based on splitting the main challenge into three sub-challenges. The first one is represented by the detection of the people inside an image coming from a camera, the second one has the role to predict the movement of pedestrians in the environment and the third one is represented by an *ad hoc* classifier which compares two vectors: the measured speed and the predicted speed, for each pedestrian in the scene. For these tasks, the Abnormal Pedestrians Activities Recognizer and Tracker (APART) system was set up. In Fig. 1 the APART architecture is shown.

ISSN: 2041-9031 (paper format), ISSN: 2041-904X (online), http://journals.witpress.com
DOI: 10.2495/SAFE-V3-N4-278-289

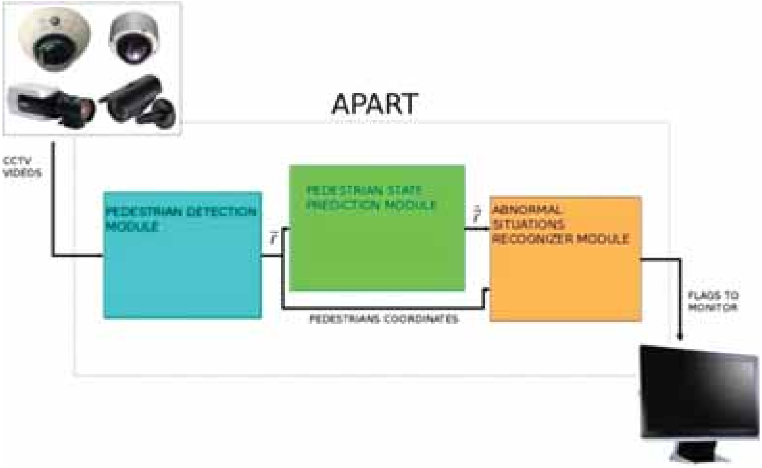
APART
CCTV VIDEOS
PEDESTRIAN DETECTION MODULE
$\vec{r}$
PEDESTRIAN STATE PREDICTION MODULE
$\hat{\vec{r}}$
ABNORMAL SITUATIONS RECOGNIZER MODULE
PEDESTRIANS COORDINATES
FLAGS TO MONITOR

Figure 1: APART Architecture.

Since there are three different goals to be reached, APART is properly divided into three different modules, represented by

1. pedestrian detection module;
2. pedestrian state prediction module;
3. abnormal situation recognizer module.

where

- in input there are images coming from the video surveillance system;
- the output of the first module is the measured position vector for each pedestrian in the scene;
- the second module output is a position vector prediction;
- in output from the system there are alarms to the video operator.

2.1 Pedestrian detection module

The first module is designed to detect people in the scene automatically. The image processing is based on the histogram of oriented gradients (HOGs) [4] that are feature descriptors used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. This method is similar to that of edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses overlapping local contrast normalization for improved accuracy. It is demonstrated that these descriptors are very convenient for pedestrian's detection. The basic idea, behind these descriptors, is the usage of distribution of intensity gradients or edge directions for each small region within the image. These small connected regions, called cells, coming from the splitting of the image, provide different histogram of gradient directions or edge orientations based on their internal pixels. Next, each of these histograms is contrast-normalized with the intensity of a larger region composed by the union of these cells. In this way, given an image, there are different regions, composed by different cells, each of these

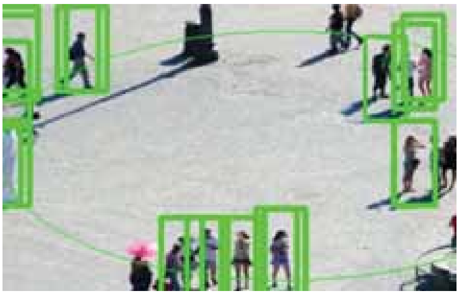

Figure 2: Image Processing Module Output without clustering.

associated to the normalized histograms taken as descriptors. In Fig. 2 an example of image processing module output without clustering is shown.

During the image processing, in a first moment, is important define a region of interest (ROI), because it is important the knowledge of the neighborhood for each pedestrian in the scene. If people around the edge of the image are analyzed, it is not possible to make a good prediction since just a part of its neighborhood is visible. Consequently the choice in this case was the use of an elliptical ROI (to get a good neighborhood it is necessary to use a smooth region), instead of a circle, because images coming from a camera are not squared, but have a 4:3 format. After this first step pedestrians are detected with the HOGs features with an Support Vector Machine (SVM) classifier. Thus the module proceeds to an unsupervised clustering of results. This is a critical point since the number of false positive could decrease, but in some cases true positive could decrease too. After these steps, since APART works with videos, and the previously points can have false negative, it is better to use a tracking features algorithm (like Optical Flow) which can track those pedestrians previously detected but not present in the current frame. In this way, if the previously points cannot detect some people (for different reasons) and if these ones were previously detected, through this step APART has acquired this information in its memory.

Last but not least, an homographic transformation is applied to the results. In fact, before this step, pixel coordinates for each pedestrian are available but APART must convert them in a metrical coordinate system. This is another critical point because an operator has to calibrate, in some way, each camera in the video surveillance systems.

Having these coordinates, APART can retrieve a measure about the pedestrians speed.

In output from this module the measured speed is available for each pedestrian, useful for the next and for the third module.

2.2 Pedestrian state prediction module

This module computes a good speed prediction for the next time step. We refer to the next time step, and not to the next frame, since it is supposed that a pedestrian cannot move in a relevant way between a frame and the next one. For this reason a time step of 0.5 sec is considered. Next the idea was to use a modern pedestrian model with a non-linear Kalman predictor and APART uses as model the Helbing ones [5–7]. In Fig. 3 an example of pedestrian state prediction module output is shown.

In this model each pedestrian is influenced by the force

$$\vec{f}_i(tot) = \vec{f}_i^{Int} + \sum\nolimits_{\forall j \neq i} \pm \vec{f}_{ij}^{Soc} + \sum\nolimits_{\forall w} \pm \vec{f}_{iw}^{Struct} \tag{1}$$

where

- $\vec{f}_i^{Int}$ is the will force;
- $\vec{f}_{ij}^{Soc}$ is the social forces;
- $\vec{f}_{iw}^{Struct}$ is the infrastructures forces.

The first one point out where pedestrian wants to go; thus it is easily (according to the Newton law):

$$\vec{f}_i^{Int} = \frac{m_i}{\tau_i}\left(\vec{v}_i^{desired} - \vec{v}_i(t)^{actual}\right) \tag{2}$$

where

- m_i is the individual mass;
- τ_i is the individual changing speed time;

The second and the third ones are based on the neighborhood instead. This means that during the movement a pedestrian is influenced by other pedestrians, walls and obstacles around him. In the last Helbing model [7] these forces are defined as

$$\vec{f}_{ij} = w\left(\varphi_{ij}(t)\right)\cdot \vec{g}\left(d_{ij}(t)\right) \tag{3}$$

where

- $w\left(\varphi_{ij}(t)\right)$ is the scalar factor;
- $\vec{g}\left(d_{ij}(t)\right)$ is the vector factor.

In particular the first term is based on the relative angle between us and pedestrians or obstacles around us

$$w\left[\varphi_{ij}(t)\right] = \lambda_a + \left[\left(1-\lambda_a\right)\frac{1+\cos(\varphi_{ij})}{2}\right] \tag{4}$$

where λ_a is a calibration parameter that changes the shape of the function from a circle $\lambda_a = 1$ (which means that the pedestrian takes care equally about ahead and behind pedestrians) to a cardioid function where $\lambda_a = 0$ (which means that it does not take care about pedestrians behind him). $\varphi_{ij}(t)$ is the angle between two pedestrians, or between the pedestrian and the obstacle.

The second term, instead, is based on the gradient of a potential field defined around each pedestrian

$$\vec{g}_{ij}\left(\vec{d}_{ij}\right) = -\vec{\nabla}_{\vec{d}_{ij}} V_{ij}(b_{ij}) = -\frac{dV_{ij}(b_{ij})}{db_{ij}}\vec{\nabla}_{\vec{d}_{ij}} b_{ij}\left(\vec{d}_{ij}\right) \tag{5}$$

where the potential field is simple defined as

$$V_{ij}(b) = ABe^{-b_{ij}/B} \tag{6}$$

where A and B are two constants and b_{ij} determines the shape of this field like an ellipse whose minor semi-axis is equal to

$$2b_{ij} = \sqrt{\left(\left\|\vec{d}_{ij}\right\| + \left\|\vec{d}_{ij} - \left(\vec{v}_j - \vec{v}_i\right)\Delta t\right\|\right)^2 - \left[\left(\vec{v}_j - \vec{v}_i\right)\Delta t\right]^2} \tag{7}$$

where $\vec{v}_i$ is equal to zero if infrastructure forces are considered, like wall and obstacles, and where

- $\vec{v}_j$ is the j-th individual speed;
- $\vec{d}_{ij}$ is the distance (as a vector) between pedestrian i and j.

As it is possible to see, a complex nonlinear system is described, thus a linear Kalman Filter could not be used directly as a predictor for this model, but the unscented Kalman filter (UKF) [8] could be used efficiently. This newer filter with respect to the Extended one, allows APART do not linearize the model, but use it as it stands. In this algorithm the inputs are represented by μ_{t-1}, Σ_{t-1}, u_t, z_t where

- μ_{t-1} is the position and speed mean at time t–1;
- Σ_{t-1} is the covariance at time t–1;
- u_t is the Helbing model input vector (relative distance and speed between the pedestrian and others or between him and in sight obstacles);
- z_t is the noise.

The UKF_algorithm ($\mu_{t-1}, \Sigma_{t-1}, u_t, z_t$) is represented by $\chi_{t-1} = \left(\mu_{t-1}\ \mu_{t-1} + \gamma\sqrt{\sum_{t-1}}\ \mu_{t-1} - \gamma\sqrt{\sum_{t-1}}\right)$

a. $\bar{\chi}_t^{*} = g\left(\chi_{t-1}, u_t\right)$

b. $\bar{\mu}_t = \Sigma_{i=0}^{2n} w_m^{[i]} \bar{\chi}_t^{*[i]}$

c. $\bar{\Sigma}_t = \Sigma_{i=0}^{2n} w_c^{[i]} \left(\bar{\chi}_t^{*[i]} - \bar{\mu}_t\right)\left(\bar{\chi}_t^{*[i]} - \bar{\mu}_t\right)^T + R_t$

d. $\bar{\chi}_t = \left(\bar{\mu}_t \bar{\mu}_t + \gamma\sqrt{\bar{\Sigma}}_t \bar{\mu}_t - \gamma\sqrt{\bar{\Sigma}}_t\right)$

e. $\bar{Z}_t = h\left(\bar{\chi}_t\right)$

f. $\hat{z}_t = \Sigma_{i=0}^{2n} w_m^{[i]} \bar{Z}_t^{[i]}$

g. $S_t = \Sigma_{i=0}^{2n} w_c^{[i]} \left(\bar{Z}_t^{[i]} - \hat{z}_t\right)\left(\bar{Z}_t^{[i]} - \hat{z}_t\right)^T + Q_t$

h. $\bar{\Sigma}_t^{x,z} = \Sigma_{i=0}^{2n} w_c^{[i]} \left(\bar{\chi}_t^{*[i]} - \bar{\mu}_t\right)\left(\bar{Z}_t^{[i]} - \hat{z}_t\right)^T$

i. $K_t = \bar{\Sigma}_t^{x,z} S_t^{-1}$

j. $\mu_t = \bar{\mu}_t + K_t(z_t - \hat{z}_t)$

k. $\Sigma_t = \bar{\Sigma}_t - K_t S_t K_t^T$

l. Return μ_t e Σ_t

So, the choice was to use it, with the Helbing model, in a predictor configuration.

Finally, it is important to underline that in this block there are some critical or heuristic variables that must be optimized. For this last purpose genetic algorithms [9] are used. Obviously, to apply an optimization algorithm to the previous steps it is necessary to define an error function to minimize. Given a ground truth, like in [7], the error function is defined as

$$e = \frac{\vec{r}_i^{SIM}(t+T) - \vec{r}_i^{REAL}(t+T)}{\vec{r}_i^{REAL}(t+T) - \vec{r}_i^{REAL}(t)} \tag{8}$$

based on real and simulated position vector $\vec{r}_i$ in two consequent time steps.

The results are

- $A = 1$;
- $B = 5$;
- $\lambda = 0$;
- $m = 70$ (Kg)

with the following options:

1. initial populations equals to 100 individuals;
2. roulette mode used to select individuals at each generation;
3. elitism applied just for two individuals at each generation;
4. cross-over with n cut-points.

Finally, once optimized, this block sends a predicted speed vector, for each pedestrian in the scene, as input to the last module.

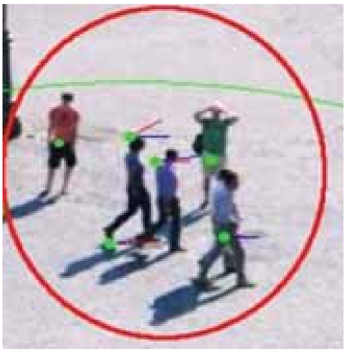

Figure 3: Pedestrian state prediction module output where the darker vector is the measured speed and the whiter vector is the predicted one.

2.3 Abnormal situation recognizer module

In the last block the system makes first a comparison between the two input vectors and then sends the results to a complex neural network. For this purpose the neural network uses

- $\hat{\theta}$ angle between these two vectors.
- v_p predicted speed norm.

Given these parameters, in fact, it is possible to describe different abnormal situations. In Table 1 some examples are shown.

As it is possible to see, different abnormal situations are mapped

- the first situation is trivial;
- the second is explained because a filter cannot predict if someone suddenly comes back and in this case the angle between these two vectors is more than 90 degrees;
- the third means a person which repeatedly change own direction, a zigzag behavior, but with a value greater than a certain threshold (to avoid noises), usually more than 45 degrees;
- the fourth is similar to the second one;
- the fifth instead can represent the faster effect described by Helbing in [6].

There are some situations where APART needs also the knowledge of previous time steps (like in situation #2 and #6); for others, instead, APART analyzes only what is going on in the actual timestamp. For this reason the choice was to use a classic feed-forward neural network [10] with a nonlinear autoregressive exogenous model (NARX) and to learn theme such that they could recognize the situations illustrated above. The NARX studies the behavior depending just on the angle between the input vectors at different time (in this way it has a memory about changing of direction for a person for four consecutive time-stamps) while the classic

Table 1: Some abnormal situations.

Abnormal situations	Typical scenario	Parameters
1 - Many running people	In a country at war	($v_p > \delta$) for many people (δ is a generic threshold and v_p is defined as $\frac{dx}{dt}$)
2 - Person that goes back and ahead continually	In front of an embassy	$(\hat{v} > 90°)$ many times
3 - Person moving at zigzag	In the night, in a park or in the underground	$\frac{d^2\hat{v}}{dt^2} > \delta_i$ and $\hat{v} > \delta_{ii}$
4 - Person that goes back suddenly	If police are present	$(\hat{v} > 90°)$
5 - In a crowd, many people try to go faster	Train station or airport exit	$(v_p > \delta) or \left(\frac{d^3x}{dt^3} > \delta^i\right)$
6 - Combinations of previous ones		

feed forward network, instead, tries to recognize situations based on pedestrian speeds and the changing of directions coming from the first neural network.

The network architecture for the third module is shown in Fig. 4.

This architecture is composed by a feed-forward neural network having 80 neurons in the middle layer and by an NARX having an input delay equals to 1 seconds and 20 neurons in the middle layer.

Next, there are five different abnormal outputs shown in Table 2.

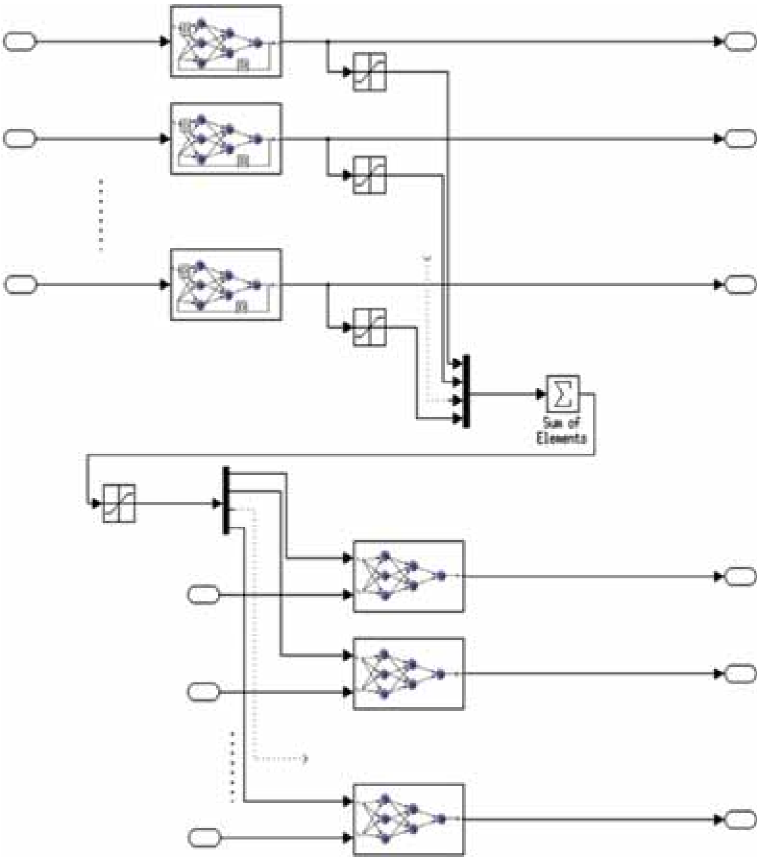

Figure 4: Neural brain architecture.

Table 2: Neural network outputs.

‘Bad’ parameter	Output	Description
Angle	1	Zigzag behavior
Angle	2	Someone (or more) suddenly comes back
Angle, Time	3	Peeping behavior
Speed	4	Someone (or more) is running
Angle, Speed, Time	5	Someone (or more) comes back and others are running

Once the architecture is outlined, the next step is to define how to train these neural networks. The choice consists in creating two different training sets with almost 90000 records for each neural network. In particular

- one data set for the NARX network is given in input as a record composed by 4 random angle values, include between 0° to 180°, representing different angles at different times (situation #1 and #2 described in Table 2);
- one data set for the feed-forward neural network is given in input as a record composed by 2 random values representing the logic OR operation between the outputs of the NARX network and a random pedestrian speed. In particular, the first value says if someone in the scene has changed own direction (the value 0 means no event while values 1 or 2 mean the same situations described in the Table 2) and the second value represents the pedestrian speed in m/s. In this data set a normal walking speed has a Gaussian distribution characterized by a mean of 1.4 m/s.

3 RESULTS

Once APART was configured, some videos about different abnormal situations were used as benchmark and testing phases.

During these testing phases different color videos, showing real critical situations, were used. The minimum video resolution was 480 × 360 pixels and the length of each one is approximately of 2 minutes. Obviously, the higher is the resolution the better the results.

In the following two examples in two different and real critical situations are described. The first scenario is a kamikaze attack to a public square (Fig. 5) and the second one is an attack to a police station in Pakistan (Fig. 6). However it is important to underline that, based on the scene to analyze, different choices are in place: the homographic transformation between camera e real coordinates, the mean for the pedestrian speed Gaussian distribution, which type of alert is expected to check, which type of situation (a market in a square or the entrance of an embassy) and so on.

In this cases, for example, APART does not care if in a public square someone suddenly change own direction or if just few people run. In this scenario could be important, instead, be aware if many people run or, in front of an embassy, could be important take care if a person running suddenly changes own direction instead of a simple zigzag behavior.

Let us explain the first scenario. Here APART works in a public square when, suddenly, many people after a kamikaze explosion, begin to run faster away. This scenario, is quite simple, since once APART retrieves the pedestrian predicted speeds, it is trivial to understand, for a neural network, what is going on.

This is because, regardless what the NARX network reports, the feed forward network acts like a threshold and reports that many people are running much faster than a walking speed mean.

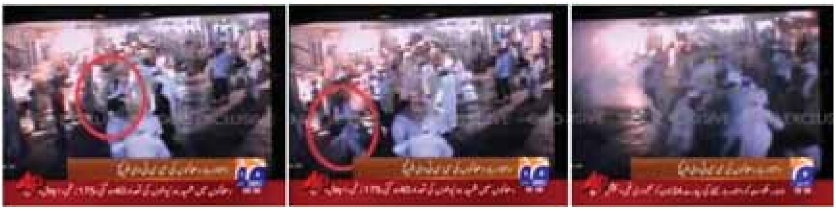

Figure 5: Kamikaze in action into red ellipse.

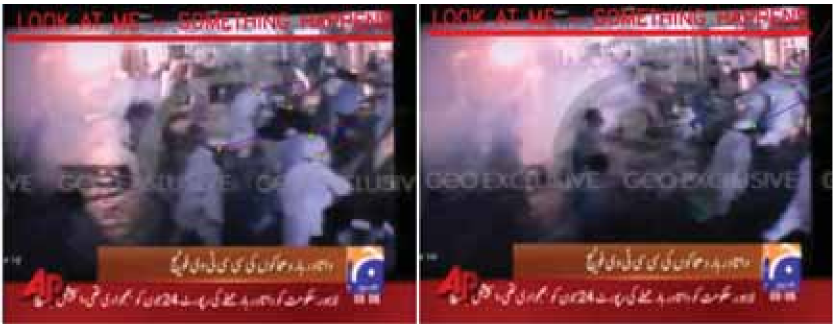
LOOK AT ME – SOMETHING HAPPENS
LOOK AT ME – SOMETHING HAPPENS
GEO EXCLUSIVE

Figure 6: Explosion, panic, many people running away and APART in action. In this case a proper message (Look at me – Something happens) appears on the monitor to alert the operator.

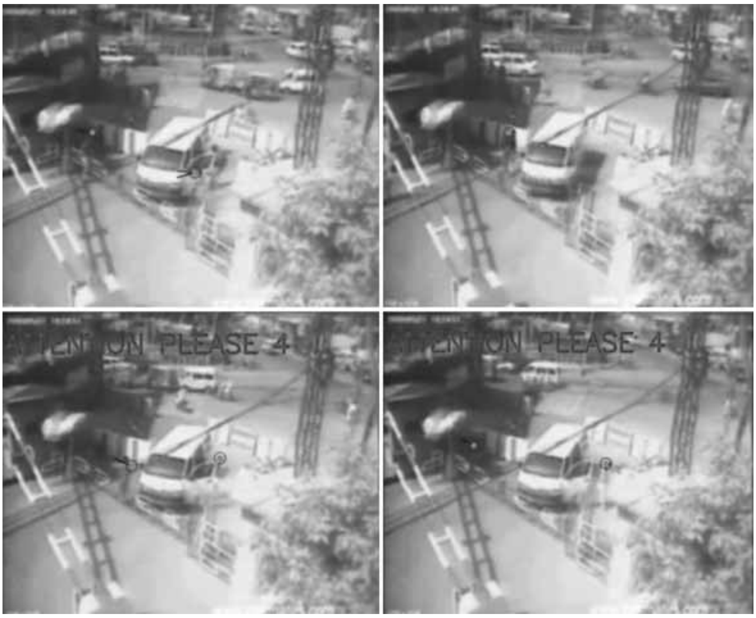

Figure 7: A guardian goes to inspect a van. Then an attacker goes out faster from the van meanwhile the guardian comes back running. In this case a proper message (Attention please) is shown for a few seconds on the monitor to alert the operator.

In the second scenario, instead, an attack to a police station by two gunmen with a white explosive van is present (Fig. 7). Here are considered situations like

- many pedestrians are running;
- one or more pedestrian changing their direction in a non-continuous way but many times;
- one or more pedestrians are running and others change their direction;
- combination of previous situations at different time.

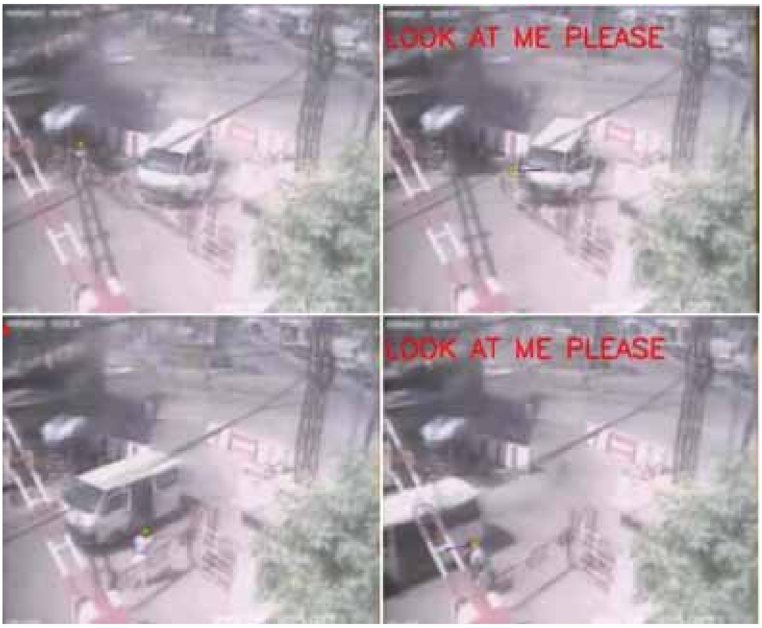

Figure 8: Attackers brake down entrance defenses going faster back and ahead continually trying to open the gate. In this case a proper message (Look at me please) is shown for a few seconds on the monitor to alert the operator.

With these ‘checks’ APART can understand if someone is spying the entrance (he changes own direction many times), if someone is escaping (one person changes his direction and runs away), etc.

In this first event the NARX network detects a significant change in the direction of one of the present pedestrians. This change, with a high walking speed, alerts the feed forward neural network that something is happening (alert with flag 5).

In this second event (Fig. 8), instead, continuous changes of direction, within a specific temporal slot (in this case 4 seconds) produce through the NARX network the alert with flag 2. This different alert type is due to the ‘normal’ walking speed of the attacker.

4 CONCLUSION

A new abnormal pedestrian behavior detection system has been presented. It represents a new way to face the automatic abnormal situation detection problems. Different approaches have already been studied, but the proposed system utilizes a multi-disciplinary approach using a nonlinear predictor, an heuristic pedestrian model and a complex neural network (HOG and SVM in the pedestrian detection module; Helbing model; the UKF for the prediction phase; feed forward and NARX Neural Networks).

The proposed system have demonstrated to be very efficient to understand if something is going wrong, from the security point of view, using an already existing video surveillance system.

REFERENCES

[1] Clavel, C., Devillers, L., Richard, G., Vasilescu, I. & Ehrette, T., Detection and analysis of abnormal situations through fear-type acoustic manifestations. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4, pp. 21–24, 2007.

[2] Clavel, C., Ehrette, T. & Richard, G., Events detection for an audio-based surveillance system. *Multimedia and Exp.*, **1**, pp. 1306–1309, 2005.

[3] Mehran, R., Oyama, A. & Shah, M., Abnormal crowd behavior detection using social force model. *Proc. of International Conference on Computer Vision & Pattern Recognitions*, pp. 935–942, 2009.

[4] Dalal, N. & Triggs, B., Histograms of oriented gradients for human detection. *Proc. of International Conference on Computer Vision & Pattern Recognitions*, Vol. 2, pp. 886–893, 2005.

[5] Helbing, D., A mathematical model for the behavior of pedestrians. *Behavioral Science*, **36**, pp. 298–310, 1991. doi: http://dx.doi.org/10.1002/bs.3830360405

[6] Helbing, D., Farkas, I.J., Molnàr, P. & Vicsek, T., *Simulation of Pedestrian Crowds in Normal and Evacuation Situations, Pedestrian and Evacuation Dynamics*, Springer: New York, USA, pp. 21–58, 2002. doi: http://dx.doi.org/10.1002/bs.3830360405

[7] Johansson, A., Helbing, D. & Shukla, P.K., Specification of a microscopic pedestrian model by evolutionary adjustment to video tracking data. *Advances in Complex Systems*, **10**, pp. 271–288, 2007. doi: http://dx.doi.org/10.1142/S0219525907001355

[8] Thrun, S., Burgard, W. & Fox, D., *Probabilistic Robotics*, The MIT Press: Boston, USA, 2005.

[9] Goldberg, D.E., *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley Professional: London, UK, 1989.

[10] Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press: Oxford, UK, 1996.