International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Overcoming Object Tracking Challenges for Abnormal Behavior Recognition

Khalid Abdul-Aziz Al-Shalfan

***Abstract* – This paper is about developing an Intelligent Video System which takes advantage of the latest technology and research enhancement. The goal is to let such a system automatically recognize an abnormal human behavior. We realized it and implemented it with a several scenarios using the SVM model. The results were excellent as we reached the rate of less than 1% of false positives in certain conditions. Many datasets have been developed for such activities like the UT-Interaction dataset and the UCR Video web dataset. As far as our work was concerned, we used a dataset we developed within the university premises containing both simple and complex activities. The experimental results on real-time video streams show the feasibility of our system and its effectiveness in human activity tracking and recognition.**

***Keywords* – Abnormal Behavior, Real-Time Detection, Multi-Camera Tracking, Video Surveillance.**

I. INTRODUCTION

Modeling activity and understanding behavior are fundamental for visual surveillance and for many other applications [1, 2]. However, it is unlikely that a global solution exists for all the event understanding problems. One can rather think of the many methods for abstraction and event modeling as a toolbox with each tool being called upon to address a specific type of problem [3]. For this analogy to be apt, we must have a good understanding of both our tools (i.e. methods for abstraction and event modeling) and our problems (i.e. various event domains). The major challenge in this research area (understanding video events) is translating low-level inputs into a semantically meaningful event description [4]. Robustness of activity detection, tracking and understanding modules, and occlusion handling [5]are crucial problems still to be investigated in a more systematic manner [6].The robustness of a recognition process under challenging conditions (lighting change, etc) remains one of the recurrent open problems. Also, understanding the scene context is essential as it helps considerably obtain a meaningful interpretation of a given video event [7]. In this context, we realized an IVSS and implemented it with a several scenarios using the SVM model. The results were excellent as we reached the rate of less than 1% of false positives in certain conditions. Many datasets have been developed for such activities like the UT-Interaction dataset and the UCR Video web dataset. Designing an activity recognition system which is able to compensate for such low-level failures is an extremely challenging task. It is relatively simple to classify a simple action in a limited video stream. By opposition, it is now very known in the scientific community that a more complex scenario in a long video sequence containing multiple activities is by far more difficult. In fact, detecting the starting and ending points of all occurring activities in the video is necessary but hard. As far as our work was concerned, we used a dataset we developed within the academic environment premises containing both simple and complex activities. The experimental results on real-time video streams show the feasibility of our system and its effectiveness in human activity tracking and recognition. This paper is organized as follows: Section 2 gives an overview on the state of the art. Section 3 will cover the system architecture of our IVSS. The Application of the SVM approach to the abnormal behavior recognition will be discussed in section 4. Finally, this paper will be closed by a conclusion and research directions for the future.

II. STATE OF THE ART

In the intelligent video surveillance systems, the interest is in identifying an abnormal activity. An abnormal activity could be a forbidden move in a given sport discipline; an unacceptable (rule violation for example) car move in traffic. Researchers, generally, divide activities to different categories based on the complexity or their recognition. For example, 'running' or walking could be classified as not complex as they do not contain extraneous variation, whereas activities like multi-people interaction are labeled as complex [7]. The main objective of Automated Surveillance Systems in public places requires detection of abnormal and suspicious activities as opposed to normal activities [8]. It should be able to alert the human operator of an existing suspicious activity like "A person leaving a bag in a subway station". Furthermore, the "false alerts" and "unresponsiveness to real threats" should be brought to minimum. During the last 2 decades, thankful efforts were made by the research and engineering communities in order to develop successful Intelligent Video Surveillance Systems (IVS) [9], either for commercial or research purposes, in order to answer to those specifications and challenges. The "Visual Surveillance and Monitoring (VSAM)" [10], the "Annotated Digital Video for Intelligent Surveillance and Optimized Retrieval(ADVISOR)" [11],or the "Smart Surveillance System (S3) of IBM" [2]are good examples of such systems. Unfortunately, these works and others in this field are generally seen as sets of pieces that act separately to detect specific events in particular scenarios far from a global analysis [12]. It is well established in the computer vision community that some outputs of research efforts in the topic of human behavior understanding could also be used in other interesting fields like healthcare (elderly and children real-time monitoring), human-computer interaction [13], intelligent environments [14], the entertainment industry and robot learning and control [15]. Researchers have focused on developing various recognition algorithms using space-time representations to

correctly match volumes, trajectories, or their features [4]. Activity recognition is done by matching the model with the volume constructed from inputs. Neighbor-based matching algorithms (i.e. discriminative methods) have also been applied widely. In the case of neighbor-based matching, the system maintains a set of sample volumes (or trajectories) to describe an activity. The recognition is performed by matching the input with all (or a portion) of them. Finally, statistical modeling algorithms have been developed, which match videos by explicitly modeling a probability distribution of an activity [16]. An interesting fact has to be pointed out though: Efforts have been and continue to be done towards liberation from segmentation and tracking in activity analysis [17]. If, for example, the tracking algorithm does not extract the object of the focus of attention, recognition of the activity being performed becomes more difficult.

III. System Architecture

Video surveillance is increasingly found in academic institutions [18]. It is used to oversee the safety of faculty members, staff and students, as well as to protect assets from vandalism and theft. Moreover, the campuses may be extensive, especially in the case of universities, and be comprised of several buildings, accesses and parking lots to monitor. Since educational institutions often have an IP network infrastructure, it is beneficial to set up digital video surveillance systems [19]. Due to the above reasons, we have implemented our IVSS in our University for testing. Basically, the system is composed of a set of wireless IP cameras plugged directly in the local network hub. The main advantage of such architecture is its flexibility. It enables a single human operator to monitor activities over a broad area using a distributed network of wireless IP-cameras.

Our approach is to model time and space linking multi-camera activities by tackling three problems in multi-camera activity understanding:

1. Estimating the spatial topology and importantly the temporal topology of a camera network.
2. Facilitating more robust and accurate person re-identification between different camera views, by resolving ambiguities and uncertainties that arise due to large and unknown separation between cameras both spatially and temporally.
3. Performing activity-based temporal segmentation by linking visual evidence collected from different camera views.

The architecture of our proposed system focuses on a reliable link between image processing and video content analysis as seen on Figure 1. Hence, integration of image processing within the digital video networked surveillance system itself is inevitable. The proposed IVSS system contains all the modules (video capture, image analysis, image understanding, event generator and field experience). Moreover, it contains an auto-learning module and another module about video retrieval.

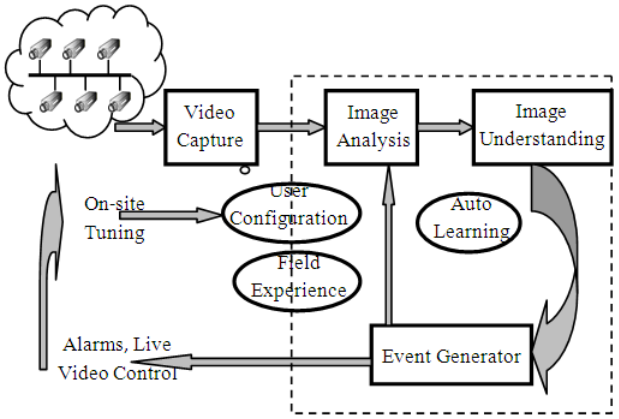


Fig.1. System Architecture

The video capture module is responsible of managing the video input data from different IP-cameras over a LAN where each camera can be accessed by its IP address. Accordingly, this module generates report about failures in the video capture process or in the network itself. Moreover, the image understanding module represent the master piece of the IVSS, it includes all AI techniques to figure out the meaning of the scene. Among its tasks: detecting abnormal behavior of human and other moving objects in the scene. The abnormal behavior is forwarded to the event generator module, which generates an alarm for the user and helps the image analysis module to tune the image processing tasks to enhance the behavior for easier perception and monitoring. The detected events based on abnormal behaviors can be modeled and stored in the field experience module for easier access and future detection [20]. This requires temporal and spatial coordination between the cameras to ensure that the same object is being tracked in each, as well as to merge statistical information about the tracked object into a coherent framework.

IV. SVM Approach to Abnormal Activity Recognition

The SVM model was finally selected to follow somehow the directions we reached in the literature review. Compared to other classification strategies, the SVM model was chosen based on the following reasons:

- So much work has been invested in the SVMs because of its earlier introduction.
- The SVMs are often called black-box classifiers. Compared to other classifiers such as Decision Trees for example, the user doesn't need to make many decisions. This could be seen as a short-coming point, but it is an advantage as the main objective is to reach the goal of discriminating some basic human actions.
- When combined with kernels, SVMs can learn a non-linear hypothesis while keeping the objective convex. This cannot be done with the Logistic Regression strategy for example.
- The decision boundary is determined only by a subset of the training data points whereas some other strategies like the Kernelized Logistic Regression, the boundary is

determined by all the training points. This affects the training and test time alike.

- SVMs are able to deal with very high-dimensional data.

Consequently, even if the SVMs have some shortcomings like the "blackboxiness" and not being suitable for "unsupervised learning", all the other positive aspects mentioned above made them the better choice for our work. The practical approach to implement the SVM on our application is by following these steps. We give a "falling person" event as an example here:

Decide about the features which will allow the differentiation between falling and non-falling events: f1 →fn. The most discriminating as found in the literature [21] are:

AR: Aspect Ration = (boundary rectangle width)/ (rectangle length)

FA: Fall Angle = between the "boundary ellipse long axis" and the "horizontal direction"

CS: Center Speed. Rather stable but change in case of occlusion

HS: Head Speed. More visible but less stable than CS.

Other features to consider eventually:

Post-fall information: The convenient example here is "lying on the ground".

Appearance-based features like posture.

Multiple cameras: info integration.

We can add another characteristic which can be quantified using a camera. A good example is "body shape change". In fact, the human shape will progressively and slowly change during usual activities, while during a fall, it will change drastically and rapidly [22].

Build a dataset

Training: Select 75% of the dataset to use in the training phase.

Test: The other 25% will be used for test. In the C++ Machine Learning API of OpenCV, training and test data is given as a cv::Mat matrix.

In the machine learning library of OpenCV each row or column in the training data is a n-dimensional sample. The default ordering is row sampling and class labels are given in a matrix with equal length (one column only, of course).

Falling Person Scenario

As an example of using the SVM in understanding the human behavior and specifically detecting abnormal ones, we used a well-known case of unusual event: a falling person. Even if the case looks trivial in the first glance, it is a lot more complex when it comes to let the machine understand it. In fact, many parameters could trigger false alarms or let pass some real falls. Many researchers tried, using different methods, to address this issue [23]. The results are globally related to the environment and the "falling type".

As per classifying cars and other objects other than humans (cars, bags, etc.), we developed the cars part but we decided not to include it and purposely omitted all the objects except the human as we concentrated on the indoor university environment where cars, for example, are inexistent. The tracking part, in contrary, has been largely applied in our project. Otherwise how can we understand the human behavior without tracking him?

Many approaches have been used to solve the problem of detecting a falling person. Some of them are listed in [24]. We focus here on how to solve this problem by using a single video camera. The disadvantage of this approach is the big number of false positives. This means that a big number of alerts are not correct. Figure 2 shows the proportion of each type of events in a dataset they used.

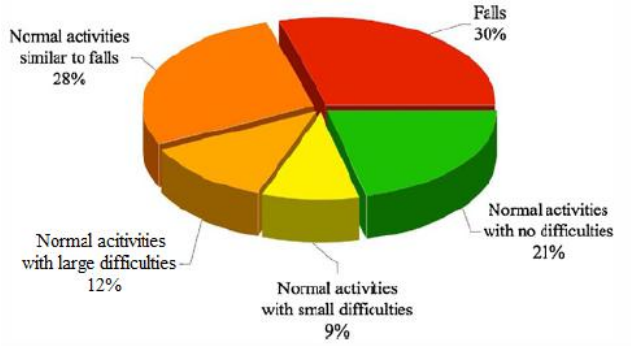


Fig.2. Some of the positions which can be falls or considered as such by an IVS.

Our idea in this study is to use the SVM statistical model to reduce the false positives. Our system detection accuracy using SVM goes from 0.92446 to 0.997602. This means that whatever strategy or combination of strategies we have chosen; the false positives are below 10% which is very respectable. The best one, which is obtained when combining the AR (Aspect Ration) and the FA (Falling Angle) strategies, goes even below 1%. Figure 3 shows some of the possible positions in the activity of falling.

(a) (b)

(c)

Fig.3.a-c: Some of the "fallings" positions detected by the IVS.

In this research work, 3 types of features have been used to give more chance for a fall to be detected.

- AR: Aspect Ratio
- CS: Center Speed
- FA: Falling Angle

Beside the detection of the falling action, we managed to get the directions of the movement. It could be "forward", "backward", "left" and "Right". In opposition to "Falling", we added also "standing". In the following, we will show some of the results we obtained as frames

depending on the directions and the position of the person (standing or falling). Later, we will show the very interesting accuracy results we got when using the SVM and depending of the feature extraction strategies we combined. The three figures around the person represent the three different strategies. The rectangle indicates the Aspect Ratio, the ellipse represents the Center Speed, and finally the cross represents the Falling Angle.

A. Accuracy

The following table shows the system accuracy when using the SVM. This is calculated when comparing the prediction to the training results. Notice the discrepancy in the results depending on the combination done between the strategies. The accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{Tue positives}}{\text{true positives} + \text{false positives}}$$

In the following, Table 1 gathers the system accuracy depending on the strategies and the combination chosen.

Table 1: The different accuracy results obtained when combining different strategies

Strategy	Accuracy
AR+CS	0.992806
AR+FA	0.997602
FA+CS	0.983213
AR+CS+FA	0.924460

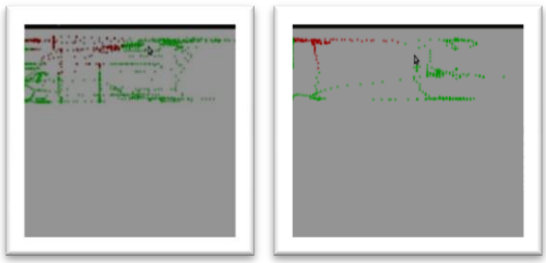
Of course and as the table shows, the best results are obtained when combining the AR and the FA strategies. The data used for our experiments are video shots we took in our university college.

B. Graphical Representations of the SVM results:

The graphics displayed in Figures 4, 5, and 6 below compare between training and prediction. 75% of the data was used for training whereas the remaining 25% was used for test and prediction. In all the cases, the SVM was able to categorize the data into 2 categories: The "Falling" in red and the "standing" in green. . The best result is obtained using the combination of AR and FA.

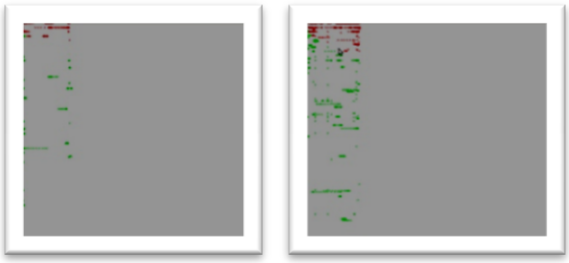
Training Data Predication SVM

Fig.4. AR and CS combination: Training vs Prediction



Training Data Predication SVM

Fig.5. AR and FA combination: Training vs Prediction



Training Data Predication SVM

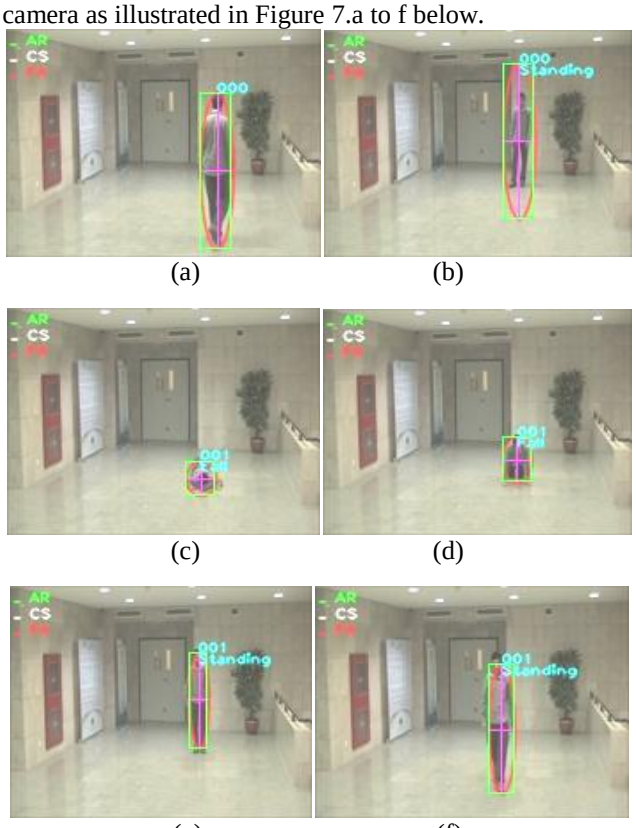
Fig.6. CS and FA combination: Training vs Prediction

C. Dataset

Many computer vision datasets exist on the Internet [25]. All of them are related to specific environments. As our system is mainly destined to academic institutions (universities, etc), we decided to develop our own dataset which summarizes the major falling positions which can occur in that specific environment. The area contains mainly class-rooms and corridors. The only detail which could be interesting to look at is detecting an abnormal human behavior in a crowd. Specialists know that is not an easy task and this can be considered as an independent project. In fact, there is an interesting number of investigators who focus on that part which is known as "human behavior understanding in a crowds" [26].

a) Backward-To-Forward falling:

This means that the person falls while coming towards the camera as illustrated in Figure 7.a to f below.



(a) (b)

(c) (d)

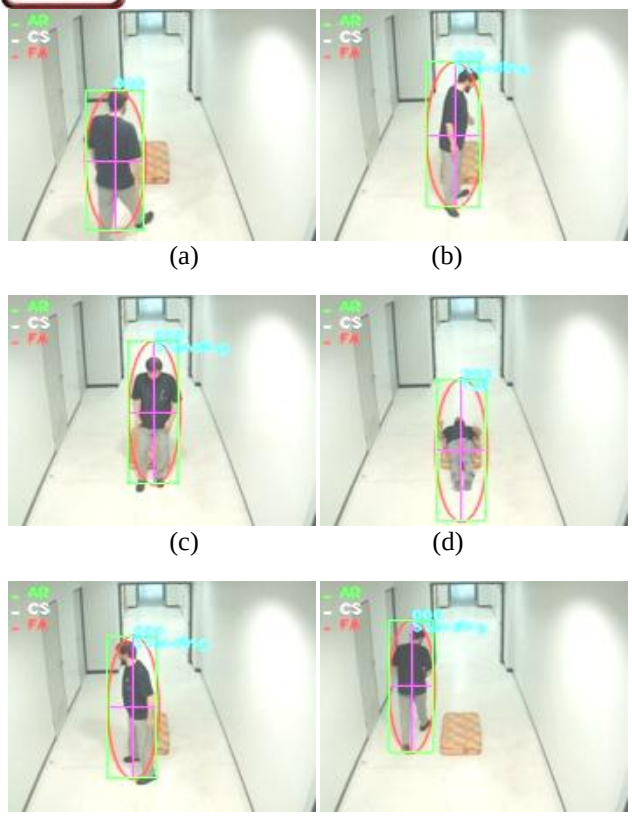
(e) (f)

Fig.7a-f: Backward To Forward Falling

b) Forward-To-Backward falling:

This means that the person falls while going away from the camera as illustrated in Figure 8.a to f below.

International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

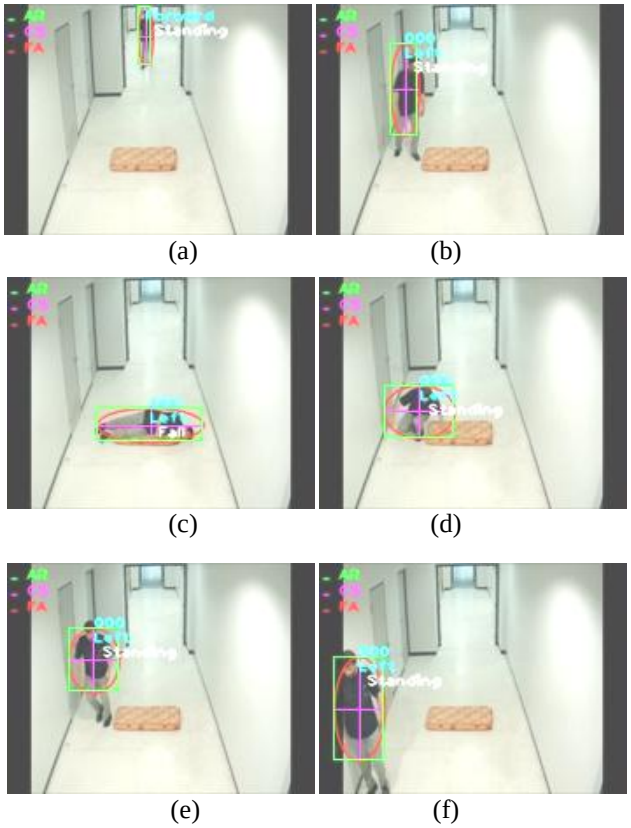


(a) (b) (c) (d) (e) (f)

Fig.8.a-f: Forward To Backward Falling

c) Left-To-Right falling:

This means that the person falls from the left to the right of the camera vision field as illustrated in Figure 9.a to f below.

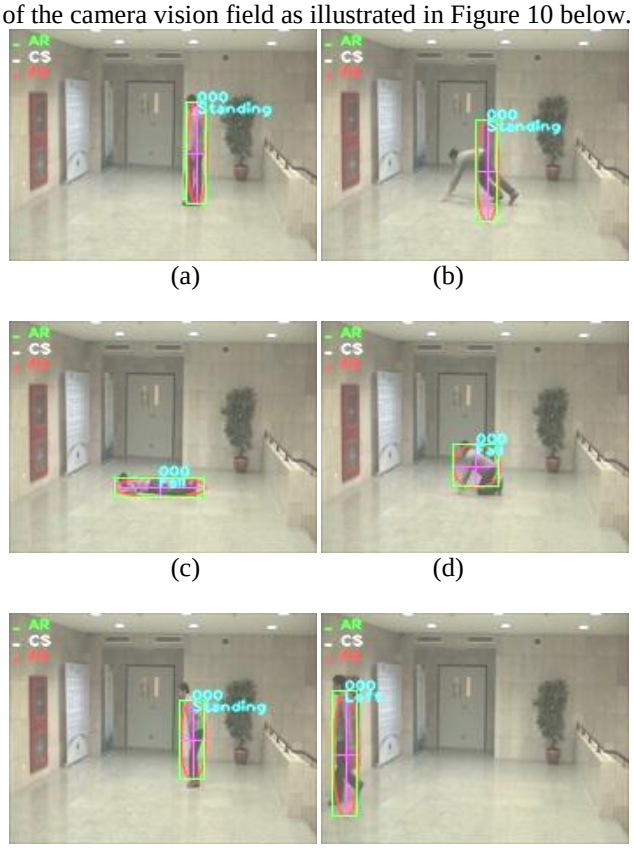


(a) (b) (c) (d) (e) (f)

Fig.9.a-f: Left To Right Falling

d) Right-To-Left falling:

This means that the person falls from the right to the left of the camera vision field as illustrated in Figure 10 below.



(a) (b) (c) (d) (e) (f)

Fig.10. Right to Left Falling

V. Overview of the System

Our system contains a few steps which can be summarized to detection of the Region of Interest (ROI), feature extraction and fall detection. Figure 11 gives a more detailed scheme.

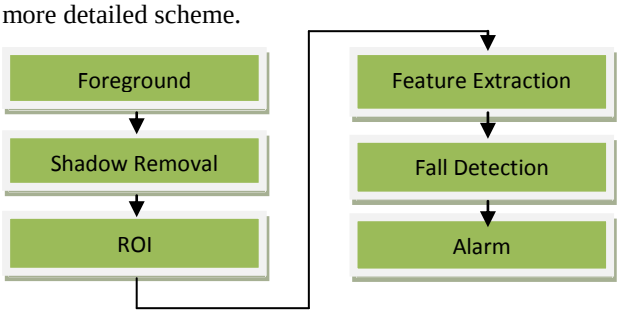


Fig.11. The different steps the system should go through to detect the falling event

In the following, we enumerate the different steps we have gone through to analyze, design and implement the IVS.

1. Model the background scene.
2. Obtain the existing blobs by separating the foreground.
3. Find the largest blobs in the scene (likely humans in our university environment), remove any noise (shadow, etc), and extract feature vector for each blob at time t.

4. Convert features into symbols by applying the k-means algorithm.
5. Aggregate symbol sequences and store for batch cluster analysis.

In the behavior clustering phase, for the set of all discrete symbol sequences, we performed the following steps:

1. Apply dynamic time warping to the set of sequences to construct a similarity-based distance matrix.
2. Run agglomerative hierarchical clustering on the distance matrix to get a tree diagram.
3. Perform SVM to model the typical behaviors in the scene.

VI. Conclusion

In this paper we presented a method for recognizing falling abnormal behavior in a wireless IP-multi-camera network given initial observations of activity in the monitored region. Both deterministic and stochastic methods were considered in order to understand the human activity. However, scientific communities in general and senior researchers in particular show more enthusiasm to using statistical methods. The same enthusiasm is shown towards the hierarchical approaches. For example, activities with structured scenarios (e.g. most of surveillance scenarios) require hierarchical approaches, and they are showing the potential to make a reliable decision probabilistically. Hierarchical approaches together with strong action-level detectors need to be explored more systematically for reliable recognition of complex activities. A good idea is to combine the advantages of these approaches and those of semantic models. In fact, the drawback of the statistical models is the complexity increase when they attempt to better capture the structure of the events being modeled. The semantic models do well in that job but lack uncertainty, thus inefficient in the recognition phase. Consequently, the combination of the two showed better event models. In fact, very interesting accuracy results were obtained when using the SVM and depending of the feature extraction strategies we combined. Our system detection accuracy using SVM goes from 0.92446 to 0.997602. This means that whatever strategy or combination of strategies we have chosen; the false positives are below 10% which is very respectable. The best one, which is obtained when combining the AR (Aspect Ratio) and the FA (Falling Angle) strategies, goes even below 1%. Future work is to generalize the approach to other abnormal behaviors.

Acknowledgment

The author expresses his deepest appreciation to King Abdulaziz City for Science and Technology for financial support of this research project (Grant ARP-29-314).

References

[1] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," ACM Comput. Surv., vol. 43, pp. 16:1–16:43, April 2010.

[2] H. M. Dee and S. A. Velastin, "How close are we to solving the problem of automated visual surveillance? : A review of real-world surveillance, scientific progress and evaluative mechanisms," Machine Vision and Applications, vol. 19, pp. 329–343, September 2008.

[3] I. S. Kim, H. S. Choi, Y. K. Moo, C. J. Young, and K. S. G., "Intelligent visual surveillance — a survey," International Journal of Control, Automation, and Systems, vol. 8, pp. 926–939, 2010.

[4] A Khalid Al-Shalfan, M. Elarbi-Boudihir. Tempo-Topographical Model Inference of a Camera Network for Video Surveillance. Accepted for publication in the International Journal of Computer and Electrical Engineering (IJCEE), ISSN: 1793-8163. 2013

[5] C. J. Lakhmi, A. Eugene, and A. Canicious, Innovations in Defence Support Systems - 3. Springer, 2011.

[6] C. Studios, "Background subtraction." Website, 2010. http://www.cutthroatstudios.com/blog/tag/running-gaussian-average/.

[7] R. A. C. Jimenez, "Event detection in surveillance video," Master's thesis, Florida Atlantic University, 2011.

[8] T. Ko, "A survey on behavior analysis in video surveillance for homeland security applications," Applied Image Pattern Recognition Workshop, vol. 0, pp. 1–8, 2008.

[9] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," ACM Comput. Surv., vol. 38, December 2006.

[10] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video," IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C, vol. 39, pp. 489–504, Sept. 2009.

[11] T. Xiang and S. Gong, "Beyond tracking: Modelling activity and understanding behaviour," Int. J. Comput. Vision, vol. 67, pp. 21–51, April 2006.

[12] J. Sherrah and S. Gong, "Vigour a system for tracking and recognition of multiple people and their activities.," in ICPR00, pp. 1179–1182, 2000.

[13] O. Boiman and M. Irani, "Detecting irregularities in images and in video.," in ICCV05, pp. 462–469, 2005.

[14] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 18, pp. 1473–1488, nov. 2008.

[15] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 844–851, 2000.

[16] A. Bobick and A. Wilson, "A state-based approach to the representation and recognition of gesture," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 1325–1337, 1997.

[17] Y. Ivanov and A. Bobick, "Recognition of visual activities and interactions by stochastic parsing," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 852–872, 2000.

[18] M. Elarbi-Boudihir, Khlaid Al-Shalfan. Intelligent Video Surveillance System Architecture for Abnormal Activity Detection. The International Conference on Informatics & Applications, Malaysia. June 3-5, 2012. Pp 102-111, http://sdiwc.net/digital-library/intelligent-video-surveillance-system-architecture-for-abnormal-activity-detection

[19] R. Bryll and F. Quek, "Agent-based gesture tracking," IEEE Transactions on Systems, Man and Cybernetics, Part A, vol. 35, pp. 795–810, 2005.

[20] S. Sarafijanovic and J. Leboudec, "An artificial immune system for misbehavior detection in mobile ad-hoc networks with virtual thymus," in Third Int. Conference on Artificial Immune Systems - Proceedings of ICARIS-2004 - Catania, Italy, vol. 4, pp. 342–356, 2004.

[21] A. Kundu, Y. He, and M.-Y. Chen, "Efficient utilization of variable duration information in hmm based hwr systems," in ICIP 3, pp. 304–307, 1997.

[22] T. Starner, J. Weaver, and A. Pentland, "A wearable computer based american sign language recognizer.," in Assistive Technology and Artificial Intelligence 98, pp. 84–96, 1998.

[23] N. Oliver, Towards Perceptual Intelligence: Statistical Modeling of Human Individual and Interactive Behaviors. PhD thesis, Massachusetts Institute of Technology, 2000.

[24] Mourad Bouzegza, M. Elarbi-Boudihir. Automatic Understanding of Human Behavior in Video Sequences: A Review. Accepted by the 8th IEEE International Workshop on Systems, Signal Processing and their Applications (WOSSPA-2013). Algiers, Algeria. May 12th-15th 2013

[25] S. Park, "A hierarchical bayesian network for event recognition of human actions and interactions," in Association For Computing Machinery Multimedia Systems Journal, pp. 164–179, 2004.

[26] M. S. Ryoo and J. K. Aggarwal, "Hierarchical recognition of human activities interacting with objects.," in CVPR'07, 2007.

AUTHOR'S PROFILE

Khalid A. Al-Shalfan

is currently Associate Professor at the department of computer science and information systems at Imam University. He received his M.Sc and Ph.D from the University of Bradford, England, in 1997 and 2001 respectively. His research interests include: Computer vision, Image processing and analysis, Image rectification, 3-D reconstruction, Target detection and tracking, Image coding and compression, Outdoor/Indoor scene interpretation for remote inspection and surveillance.