International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:8, 2012

ADABeV: Automatic Detection of Abnormal Behavior in Video-surveillance

Nour Charara, Iman Jarkass, Maria Sokhn, Elena Mugellini and Omar Abou Khaled

***Abstract*—**Intelligent Video-Surveillance (IVS) systems are being more and more popular in security applications. The analysis and recognition of abnormal behaviours in a video sequence has gradually drawn the attention in the field of IVS, since it allows filtering out a large number of useless information, which guarantees the high efficiency in the security protection, and save a lot of human and material resources. We present in this paper ADABeV, an intelligent video-surveillance framework for event recognition in crowded scene to detect the abnormal human behaviour. This framework is attended to be able to achieve real-time alarming, reducing the lags in traditional monitoring systems. This architecture proposal addresses four main challenges: behaviour understanding in crowded scenes, hard lighting conditions, multiple input kinds of sensors and contextual-based adaptability to recognize the active context of the scene.

***Keywords*—**Behavior recognition, Crowded scene, Data fusion, Pattern recognition, Video-surveillance

I. Introduction

THE past decade has witnessed the rise of video surveillance and its deployment spread around the world. The decreasing costs of video surveillance equipment have resulted in large volumes of video data. This excessive amount of information constitutes a significant challenge, even for a human observer, because it requires monitoring a huge number of individuals and their activities. Computational approaches that assist human security must be smart enough to be successful in real-world domains. Intelligent video surveillance (IVS) is a technology that uses software to identify automatically specific objects, behaviours or attitudes in video footage. It transforms the video into data to be transmitted or archived so that the video surveillance system can act accordingly. The increasing need for intelligent surveillance in the security domain, especially for public places and military applications, makes automated video surveillance systems to be unavoidable.

Nour Charara is Member of University of Applied Sciences of Western Switzerland▪ Perolles 80, CH-1705 Fribourg, Switzerland and the Lebanese University, Beirut , Lebanon (e-mail: name.surname@edu.hefr.ch).

Iman Jarkass is with LSI Department, Institute of Technology IUT, Lebanese University, Pb. 813, Saïda, Lebanon. (e-mail: ijarkass@ul.edu.lb).

Maria Sokhn, Elena Mugellini and Omar Abou Khaled are with the ICT Department, University of Applied Sciences of Western Switzerland, Perolles 80, CH-1705 Fribourg, Switzerland (e-mail: name.surname@hefr.ch).

We can group the research technologies for automatic video surveillance system into three groups: the video analysis, the video processing and filtering, and the video retrieval as video indexing and searching part..

Currently, we can discern that the visual surveillance in dynamic scenes is an active and important research area, strongly driven by many potential and promising applications. Nevertheless, more existing systems focus on analyzing recorded video using pre-defined hard rules. Consequently, they suffer from unaccepted false alarms. In our work, by choosing the context of crowded scenes in hard lighting conditions, we aim to develop an adaptive system for behaviour recognition, which it is able to provide an alert in case of abnormality detection. To achieve this objective, we will have to address the hard tasks of behaviour recognition to detect the abnormalities in crowded scenes. Several challenges such as the severe inter-object occlusion in crowded scene, the poor quality of recorded surveillance footage in hard lighting conditions, the exploration of multimodality video sensor and multimodal data fusion, and the adaptation to the change of contexts surrounding the target events or objects have to be resolved.

In order to validate the framework architecture proposal with its various modules, we test our proposals within a novel application environment: a multipurpose hall. A multipurpose hall offers many possible usages and it is designed to fit various kinds of events, such as a movie theatre (cinema), conference hall, or even as a room for parties' and ceremonies. This variety of uses also implies a variety of contexts and consequently a large number of possible events. In order to ensure safety of such halls, we need an adaptive system that can detect the type of the context.

The paper is structured as follows: in the next section we present the related works, section III introduces the framework ADABeV and gives an overview of the system and its different modules. In section IV we describe the context modeling and classification part of the framework and finally in section V we conclude the paper and present the future work.

II. Related Work

Several research works in the domain of video understanding have been carried out in the last decade. This section presents the ones which are more relevant to our work. Many issues as behavior understanding in crowded scenes,

hard lighting conditions, data fusion and contextual information exploitation are combined together in order to perform behavior recognition, so we review the representative papers and research works on these different topics respectively:

A. Behavior Understanding in Crowded Scenes

There are different research topics in video surveillance related to crowds: crowd density estimation, face detection and recognition in crowds [33]. We focus on the problem of crowd behaviour and abnormal event detection in crowded scenes. Compared to other video surveillance problems there are relatively few works related to crowds. In the past years, a variety of approaches were proposed to deal with crowd analysis and understanding. As noted in [3], there are two main approaches for crowd behaviour analysis. In the *object-based* approach, a crowd is treated as a collection of individuals; *holistic* approaches treat the crowd as a single entity, without the need of segmenting individually.

First, Cheriyadat and Radke [1] proposed an approach for clustering a set of low-level motion features into trajectories, similarly to [4], but by the usage of additional rules in the clustering process, such as the dominant movements that are computed based on the longest common sub-sequences. However, while the goal in [4] was to identify each member in the scene based on motion cues, the main goal in [1] was to extract dominant motion patterns in a crowded scene. In the same category, we can classify the work of Lin *et al.* [5] for automatic recognition of group activities for video surveillance applications. They proposed to use a group representative to handle the recognition with flexible or varying number of group members, and use an Asynchronous Hidden Markov Model (AHMM) to model the relationship between two people. A little far-away from abnormality detection, Hakeem and Shah [19] proposed a method to just detect events involving multiple agents in a video and to learn their structure in terms of temporally related chain of sub-events. Regarding the holistic approach, Davies *et al.* [10] presented a useful indicator of crowdedness or potential danger, he proposed an approach based on discrete Fourier transform, combined with a linear area transform algorithm to distinguish static from moving crowds. Andrade *et al.* [6] characterized a "usual behaviour" of a crowd based on the analysis of their optical flow, using hidden Markov models (HMMs). Mahadevan *et al.* [16] also present a framework for anomaly detection in crowded scenes. Their model for normal crowd behaviour is based on mixtures of dynamic textures; outliers under this model are labelled as anomalies.

The clear limitation of the current works is that there is no approach that combines between the holistic approach and the object based approach in crowded scenes, so that we can benefit from the important features of both approaches.

B. Hard Lighting Conditions

In fact, there are no much works that treat this problem, Ilie *et al.* [20] propose a new concept of image and video enhancement technique, CER (context enhanced rendering).

They incorporate context information of a night-time scene from one image with other important features from another daytime image at the same viewpoint. In a like manner, the authors of [7] and [8] have resolved this problem. In [7] they proposed an enhancement method that combines images taken at different times by using image fusion techniques. They decrease the aliasing by computing in a gradient domain, Cai *et al.* [8] combine daytime image and night-time image together using the image segmentation and the object extraction technique. A background modelling-based method suitable for the security surveillance application is proposed by Soumya [2], to segment the moving object from the poor illuminated area, using dynamic matrix.

It is still a challenging problem for night-time video applications since most of the previous work, mentioned above, focuses on the technical side of the problem i.e. image fusion, segmentation, etc, and does not concentrate on information processing to provide an analysis task such as detection and recognition.

C. Multiple Input Kinds

The work in this area can be divided into two categories, according to the fused modalities: one that uses fusion of visible and infrared images and the other that uses fusion of audio and video information. First, the operational requirement to use multiple inputs kinds or sensors is due to the limitations of individual sensor (the normal camera CCTV) to capture all available visual information about the scene. Some recent works have addressed the tracking of humans and vehicles with multiple sensors [9]. In [11], Torresan *et al* present their method in the context of surveillance and security, for efficient detection and tracking, they use the fusion of thermal infrared with visible spectrum video to track the blobs (regions), and the tracking is done first separately in the visible and thermal modality. Using thermal and visible imagery, Davis et al. [10] propose a new contour-based background-subtraction technique for persistent object detection in urban settings. Before fusion, the statistical background subtraction is performed in the thermal domain to identify the initial regions of interest, then colour and intensity information are used to obtain the corresponding regions of interest in the visible domain. In the second category, Gatica *et al.* [12] presented a method that fuses 2-D object shape and audio information.

D. Contextual-based Adaptability

The literature contains an important number of formal definitions of Context in vision understanding system, trying to answer to an essential problematic: how to define context in order to generate and use it automatically. A very broad definition is given by [25], who proposes that the context is any and all information that may influence the way a scene is perceived. More specifically, [26] brings a more detailed definition: a complex structure comprising generic descriptions for spatial structures, temporal changes […] and the intention of the action.

In [21] the context remains constant during processing, and describes the scene that is filmed. To model a context, information can come from one of the different types of information. First, the visual information which includes in primary position the description of static objects as the size, the color, and the geometry of their sub-parts; then comes the polygonal zones with semantic information from the entrance and exit zones to area of interest. In principal these zones need to be manually defined offline by a human operator. Furthermore, the camera characteristics as the focal length, the position and the direction affect the information. Consequently, the raw information standards (color, histograms, direction of edges and texture) have also some influence on the information. Second, the Context awareness, in [22] the authors use an approach inspired by Activity Theory to model context. In contrast to other approaches presented thus far [21]-[23]-[24]-[25] where the context is to be described in an objective manner, this model assumes a subjective view in the different situations by the different actors involved. Particularly, the proposed model uses a taxonomy with five different types of contextual information (environmental, personal, social, task, and spatiotemporal).

The representation of the context can use a map of the scene as a support. The representation of the scene as a space is used to gather all of the context information in one place. In [24], a 2D map with zones drawn offline by an operator is used. In previous work [27], the concept of using a scene map in order to improve the computation of spatiotemporal relations to mobile objects has been introduced. It has been found by the authors that such segmentation and spatial reasoning improves the computation of object detection and tracking, as well as behaviour analysis. However, manually drawing the plan of the scenes we want to observe is a time-consuming task. Other approaches detach themselves from the manual segmentation or context annotation and propose using graphical descriptors for automatic context representation. Based on the histogram of oriented gradients (HOG), which is a well-known object detection method [29], Heitz *et al.* [31] distinguish between rigid objects (e.g. vehicles) and other "stuff" like cars. Here, the things and stuff (TAS) context model allows clustering the different types of elements contained in a scene and thus improving the detection of objects. Expanding on the idea on [30], authors of [28] go further and present a new context risk function and a maximum margin context (MMC) model. This approach uses a polar geometric context descriptor and is able to take contextual information in a more unsupervised way. Experiments show that MMC[28] has a better detection rate than previous work (HOG [29] and TAS [30]), with less false positives.

III. Discussion

Moreover, some limitations of video processing technique in the real environment are not well declared in most of research. Here we propose solutions for resolving these following challenging ambiguous visual observations together and overcoming unreliability of conventional behaviour analysis methods:

a. Regarding the case of crowded scenes with all their challenging features.

b. Studying the problem in hard lighting conditions (in both night time and day time).

c. Exploiting an abnormal detection system based on a heterogeneous sensor network consisting of both CCTV cameras and thermal cameras with audio sensors. These networked heterogeneous sensors will function cooperatively to provide enhanced situation awareness.

d. Developing an intelligence system that is able to switch between different contexts at different times during the day, referring to predefined sufficient information to provide effective selection of the existing context.

Towards the goal of realizing an automatic and generic system that answers to the aforementioned challenges, a new formalism is proposed. ADABeV (Automatic Detection of Abnormal Behavior in Video-surveillance) is an integrated framework for event recognition in crowded scene surveillance videos. It aims at detecting the abnormal human behavior, automatically and in real time. Specifically, this framework focuses on two main challenges:

(1) Automatic event analysis in scenes monitored by video surveillance cameras.

(2) Detection of abnormal human behavior in crowded public scene.

Additional details about the framework concept will be presented in this paper (in section IV.A). Now we will briefly explain these two main issues.

A. Event Recognition

When talking about *Event*, we can find in the literature, many terms which denote similar meanings such as *activity* [13] and *action* [14]. In particular, action consists of whole body movement formed by a set of atomic movements. Nevertheless, we refer an Event to a set of related actions by time and targets that mostly involve several objects that interact in a common space monitored by one or more cameras. Our specification of Event is not so far from Kumar et al. [15]; they describe Events by the spatiotemporal relationship between targets and contextual elements or with other targets. Therefore, we denote *event recognition* the task of identification and classification of image sequence to have a high-level semantic interpretation of the scene.

B. Abnormal Behavior

In order to provide security, it is necessary to analyze the behaviors of people and determine whether these behaviors are normal or abnormal. Before we mention what is meant by abnormal behavior, we should notice that several keywords were used by many research works [3, 16, 31, 32] to refer to the same notion (*unusual, rare, atypical, interesting, suspicious, anomalous*). With Mahadevan *et al* [16], the *abnormalities* are defined as measurements whose probability is below a certain threshold under a normal model.

Breitenstein [32] differentiates between the *rare* behavior and the *unusual* one, he claims that the first means only events that have not been observed before; those that have been seen at least once are considered rare but not necessarily abnormal previously unseen or having low statistical representation in the dataset. Moreover, there are some abnormal behaviors that can be predicted in certain circumstances and consequent ly it can be modeled.

While it is hard to define individually the entirety of the different abnormal behaviors, we say about an event *abnormal behavior* when (1) it is done by a human operator and (2)

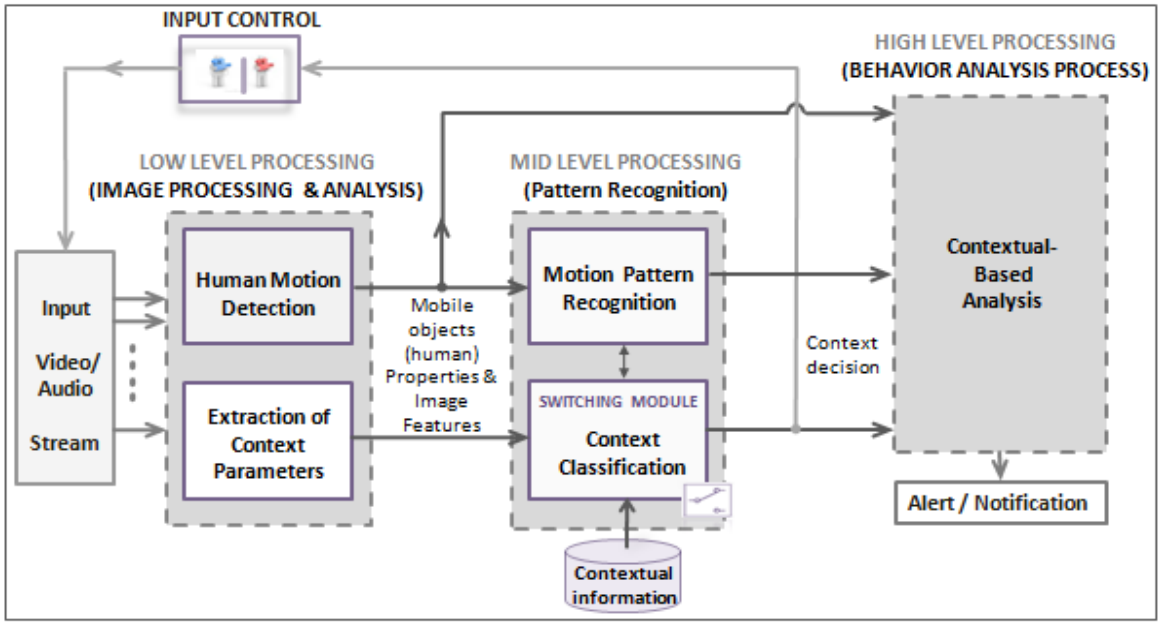


Fig. 1 Global ADABeV framework architecture

IV. ADABeV Framework

A. ADABeV Architecture

Based on our viewpoint of the two described issues, and respecting the workflow followed by the most existing vision-based human behavior recognition system (Fig. 2), we distribute our framework on three processing levels; each level is composed from one or more module: (Fig. 1)

- *The Low-Level Processing:* called 'Image Processing and Analysis'. Initially, this layer takes three types of input: RGB images (from video cameras), infrared spectrums (from infrared video camera) and audio (from equipped microphone), this input will be controlled later with the Input Control module to adapt and select just the needed input equipment according with the detected context.

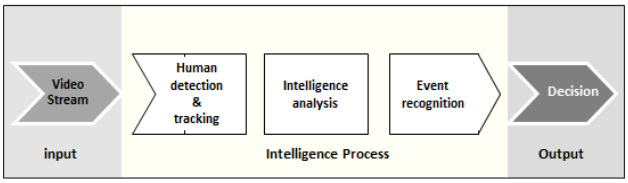


Fig. 2 Workflow of IVS system (Intelligent Video Surveillance)

The presence of a moving human body is detected by the 'Human Motion detection' module, and a set of features are computed into feature vectors, to be processed in the next level. The second module in this level is intended to extract the contextual parameters used in context classification in the mid-level.

- *The Mid-Level Processing:* for the principal pattern recognition task in this framework 'Pattern Recognition'; it is composed of the two following modules: the Motion Pattern Recognition module and the Context Classification module. The first aims to estimate, in principle, the holistic motion in the scene; it defines some properties about the entrance and exit movement. These holistic motion properties can also be defined with regard to the context, which is studied in the second module. Context Classification is a switching module; it focuses on context modeling in order to automatically detect the context of a scene, in order to carry out a hard decision about the context where the scene is filmed. At this stage, we consider a predefined set of Context models with their contextual information. This classification of context will later constitute the basis of behavior analysis in the highest level. More details on the context modeling are given in the next section.

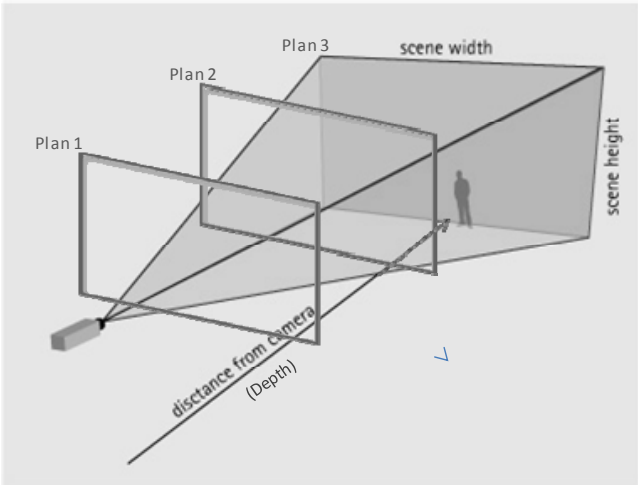
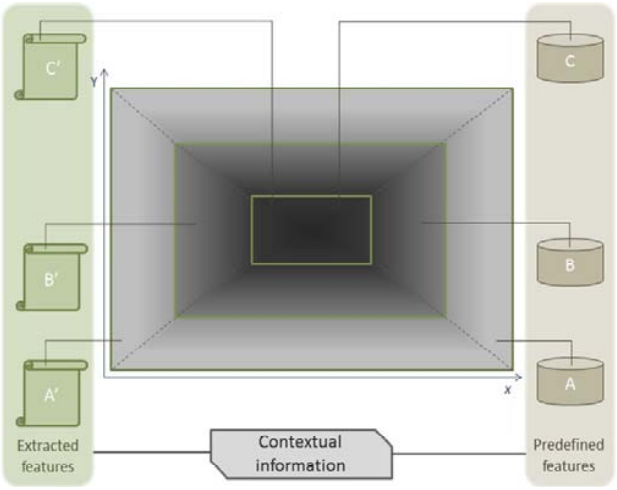


Fig. 3 (a) Segmentation of a filmed scene on several depth zones,



(b) Example of three zones of depth A, B and C with Camera 3D view

- *The High-Level Processing:* Finally, at the highest processing level, based on the context decision provided by the previous level, a pre-processing step is required to adapt and prepare the holistic motion properties with the mobile objects properties and the image features to have the best ground for learning method that must be integrated in this behaviour recognition module.

As final output of the system, an alarm or notification will be generated, in case of detection of abnormal behaviour according to certain criteria that must be predefined with regard to the existing context as mentioned above.

We propose this generic context definition: *Context is the environment where an event can occur. The isolation of this event, from its environment, leads to another interpretation of this event.* So the context is the environment that can strongly affect the interpretation of events occurring in it. Consequently, this method consists to segment the scene into different planes more or less far from the focal plane or in other terms with different depth values, this *Depth plans* should be detected automatically during the processing phase, according to certain predefined features. Once the plans are detected, the zones of depth are then determined and here start the features extraction for each zone separately (Fig. 3 (a)). Each zone has its special features to be extracted. We argue this segmentation that in our used case (Multipurpose hall); we can clearly observe that it can be segmented into three zones, from the camera's focal plane; first the spectator seats area, then, more deeply, the sidewalk and last, the theater or the onstage or the wide screen (according to the usage).

Each zone has a set of special features that can be essential in context classification decision. On the other hand, each zone also has its special predictive events and behaviors. In fact, we assume that there are some behaviors that can be considered as normal in certain zone, but abnormal in another. By this way, we will not confuse between the normal and abnormal behavior among the zones. Therefore, we obtain more accuracy in behavior recognition process.

Note that the depth data are provided by a stereo system (two calibrated video-cameras). After this description, we can point out the main advantage of this method. First, this method decreases the extracted data redundancies at the low level, when extracting just the needed features for each zone. Second, this leads to reduce the analysis complexity at the Mid-Level, and also to increase the robustness of behavior analysis at the highest level. After the extraction of several contextual data contained in the images according to the detected zones of depth, we must now be able to process these data in order to propose a classification. Thus, we must be able to link the extracted information from the different depth zones with the predefined information of each depth (Fig. 3 (b)). For this purpose, we use the Transferable Belief Model (TBM) [17, 18]; it provides a flexible and powerful representation for quantified beliefs associated with the extracted contextual information. It is perfectly adapted to our system application where data are collected from partially reliable sources.

B. Context Modeling and Classification

Many authors have used Context either implicitly or explicitly in their image understanding systems, but few have made the representation and use of context as central design feature, as we have proposed. The role of context modelling in this system is to define formalism on the context description, in order to enable the identification of a context in a given video through context classifier. First, the definition of the context depends on the process nature. Therefore, we need a more specific definition to the scope of this project.

V. Conclusion

In this paper we present a novel end to end framework called ADABeV. It aims to address the artificial intelligence issues in vision based systems. In particular, this framework is designed to detect automatically the abnormal behavior of human operator in video-surveillance of crowded scenes; especially in multipurpose hall. We putted ahead the notion of context modeling, we also present the global method based on depth zones segmentation. Finally, as a next step, we are currently implementing the designed concept.

References

[1] A. M. Cheriyadat and R. Radke, "Detecting dominant motions in dense crowds," IEEE J. Select. Topics Signal Process, vol. 2, no. 4, pp. 568–581, Aug. 2008.

[2] T. Soumya, "A Moving object segmentation method for low illumination night videos," in Proceeding of WCECS, October 22–24, San Francisco, USA, 2008.

[3] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[4] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Washington, DC, pp. 594–60, 2006.

[5] W. Lin, M.T. Sun, R. Poovendran, Z. Zhang, "Group Event Detection for Video Surveillance," in Proceedings of ISCAS'2009. pp.2830~2833.

[6] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in Proc. Int. Conf. Pattern Recognition, Washington, DC, pp. 175–178, 2006.

[7] R. Raskar, A. Ilie, and J. Yu, "Image fusion for context enhancement and video surrealism," in Proc. of the 3rd international symposium on Non-photorealistic animation and rendering (NPAR), pp. 85–152. Annecy, France, June 2004.

[8] Y. Cai, K. Huang, T. Tan, and Y. Wang, "Context enhancement of nighttime surveillance by image fusion," in Proceedings of ICPR, pp. 980–983, 2006.

[9] A. Nakazawa, H. Kato, and S. Inokuchi, "Human tracking using distributed vision systems," in Proceedings of the 14thICPR, pp. 593–596.

[10] J. W. Davis and V. Sharma, "Fusion-Based Background-Subtraction using Contour Saliency," Computer Vision and Pattern Recognition, 20-26 June, 2005.

[11] H. Torresan, B. Turgeon, C. Ibarra-Castanedo, P. Hébert, X. Maldague, "Advanced Surveillance Systems: Combining Video and Thermal Imagery for Pedestrian Detection," in Proc. of SPIE, Thermosense XXVI, volume 5405 of SPIE, pp. 506–515, April 2004.

[12] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filter," in IEEE International Conference on Image Processing (ICIP03), 2003.

[13] R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, 28(6):976–990, 2010.

[14] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," Computer Vision and Image Understanding, 104(2):90–126, 2006.
[15] P. Kumar, S. Ranganath, K. Sengupta, "Behavior Interpretation from Traffic Video Streams," in Proceedings of the IEEE International Conference on Intelligent Transportation Systems". October 12-15, 2003, Shanghai, China, volume 2:pp.1214-1219.
[16] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. "Anomaly detection in crowded scenes," in IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[17] P. Smets and R. Kennes, "The transferable belief model," Artificial Intelligence, vol. 66, no. 2, pp. 191–234, Dec. 1994.
[18] M. Guironnet, D. Pellerin, M. Rombaut, "Camera motion classification based on transferable belief model," European Signal Processing Conference (EUSIPCO'2006), Florence, Italy, September 2006.
[19] A. Hakeem, M. Shah, "Multiple Agent Event Detection and Representation in Videos," in Proceedings of American Association for Artificial Intelligence AAAI, 2005.
[20] A. Ilie, R. Raskar, J. Yu, "Gradient domain context enhancement for fixed cameras," in Proc. of ACCV. Jeju Island, Korea, January 2004.
[21] V. T. Vu, F. Bremond, M. Thonnat, "Human behavior visualisation and simulation for automatic video understanding," in Proc. of the 10th Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG–2002), Plzen–Bory, Czech Republic, 2002.
[22] J. Cassens, A. Kofod-Petersen, "Using activity theory to model context awareness: a qualitative study," in Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference, Florida, USA, AAAI Press, 2006.
[23] O. Brdicka, P. Reignier, J. L. Crowley, "Modéliser et faire évoluer le contexte dans des environnements intelligents," In Ingénierie des Systèmes d'Information (ISI), Lavoisier, Vol. 11, No. 5, December 2006.
[24] F. Bremond, M. Thonnat, "Issues of representing context illustrated by video-surveillance applications," in International Journal of Human-Computer Studies - Special issue: using context in applications, Volume 48 Issue 3, March 1998.
[25] T. Strat, "Employing contextual information in computer vision," in DARPA93, pages 217-229, 1993.
[26] H.H. Nagel, "From image sequences towards conceptual descriptions," Image and Vision Computing, 6(2):59-74, 1988.
[27] M. Mohnhaupt, B. Neumann, "Understanding object motion: Recognition, learning and spatiotemporal reasoning," research report FBI-HH-B-145/90, University of Hamburg.
[28] W-S. Zhen, S. Gong, T. Xiang, "Quantifying contextual information for object detection," in IEEE 12th International Conference on Computer Vision, pp.932-939, Sept. 29 2009-Oct. 2 2009. doi: 10.1109/ICCV.2009.545934
[29] N. Dalal, B. Triggs. "Histograms of oriented gradients for human detection," in CVPR, 2005.
[30] G. Heitz, D. Koller. "Learning spatial context: Using stuff to find things," in ECCV, 2008.
[31] A. Adam, E. Rivlin, I. Shimshoni and D. Reinitz, "Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 3, March 2008.
[32] M. D. Breitenstein, H. Grabner, and L. V. Gool, "Hunting Nessie - real-time abnormality detection from webcams," in IEEE International Workshop on Visual Surveillance, 2009.
[33] S. Saxena, F. Brémond, M. Thonnat, R. Ma. "Crowd Behavior Recognition for Video Surveillance," in Advanced Concepts for Intelligent Vision Systems (ACIVS 08), 2008.