TOWARD GENERIC INTELLIGENT KNOWLEDGE EXTRACTION FROM VIDEO AND AUDIO: THE EU-FUNDED CARETAKER PROJECT

C. Carincotte[1], X. Desurmont[1], B. Ravera[2], F. Bremond[3], J. Orwell[4], S.A. Velastin[4], J.M. Odobez[5], B. Corbucci[6], J. Palo[7], J. Cernocky[8]

[1] MULTITEL asbl, Mons - Belgium - carincotte@multitel.be
[2] Thales Communications, Colombes - France [3] INRIA, Sophia Antipolis - France
[4] Kingston University - United Kingdom [5] IDIAP, Martigny - Switzerland [6] ATAC, Roma - Italy
[7] Solid Information Technology, Helsinki - Finland [8] Brno University of Technology - Czech Republic

Keywords: Content analysis/retrieval, Knowledge extraction, Massive Recording.

Abstract

The CARETAKER[1] project, which is a 30-month project that has just kicked off, aims at studying, developing and assessing multimedia knowledge-based content analysis, knowledge extraction components, and metadata management sub-systems in the context of automated situation awareness, diagnosis and decision support. More precisely, CARETAKER will focus on the extraction of structured knowledge from large multimedia collections recorded over networks of cameras and microphones deployed in real sites. The produced audio-visual streams, in addition to surveillance and safety issues, could represent a useful source of information if stored and automatically analyzed, in urban/environment planning, resource optimization, disabled/elderly person monitoring, ...

1 Introduction

Advances in sensor devices, communication and storage capacities make it increasingly easier to collect large corpora of multimedia material. However, the value of this recorded data is only unlocked by technologies that can effectively exploit the knowledge it contains. It is thus the goal of the proposed project to investigate techniques allowing the automatic extraction of relevant semantic metadata from raw multimedia, to explore the value of the extracted information to relevant users, and to demonstrate this in a framework that preserves the privacy of the individual. In this context, we will focus on the extraction of structured knowledge from multimedia collections recorded over a network of camera and microphones. The motivation is that, despite the concerns over privacy issues, such networks are becoming more and more common in different environments such as public transportation premises, cities, public buildings or commercial establishments. The multimedia streams they produce, in addition to surveillance and safety issues, could thus potentially represent a useful source of information if stored and automatically analyzed, for instance in urban planning and resource optimization applications.

[1]The CARETAKER acronym stands for Content Analysis and REtrieval Technologies to Apply Knowledge Extraction to massive Recording.

We will consider two types of content knowledge: a first layer of primitive events that can be extracted from the raw data streams, such as ambient sounds, the degree of crowding present in the scene, and the routes taken by individual people. A second layer of higher semantic events is defined from longer term analysis and from more complex relationships between the primitive events. It is important to note that while few real systems are equipped with such content extraction and analysis tools, academic laboratories have developed many algorithms partially addressing these issues, but most of the time applied on toy problems, with very little actual or acted data. Thus, the overall goal of the project is to investigate current and novel technologies to extract and exploit this information, by evaluating them in a real test case, while exploring their added-value for real users.

The next section presents the CARETAKER project in terms of technical objectives; corresponding scientific research lines, and expected results and contributions are detailed in Sec. 3. Finally, the project objectives are summarized in Sec. 4.

2 Technological and scientific objectives

Based on the above short description, the project will address the following main issues:

2.1 End-user requirements and knowledge representation

A proper study of knowledge extraction and management system of multimedia data can not be performed without considering real data. Moreover, the design and the relevance of the system can only be investigated with the collaboration of a real end-user. This issue will be covered thanks to the participation of ATAC[2]. This partner will provide a real testbed site inside the metro of Roma, involving currently more than 20 sensors. Additionally, the identification of the real user needs and beneficial use-case scenarios will serve as a reference point for the correct framing of the semantic description scheme (i.e. the ontology), the knowledge extraction components, and the interface and demonstrator optimization.

[2]ATAC (Agenzia per i Transporti Autoferratranviari del Comune di Roma) is Rome Agency for Mobility, in charge of the planning, control and regulation of Public Transport in Roma.

2.2 Content extraction and knowledge discovery

Years of research have shown that bridging the semantic gap between the real sensory data and the high-level semantic entities which are of user-interest and embedded in the ontology, is still an open issue. In CARETAKER, we will clearly eliminate at the beginning of the project the user-wishful requirements which are out of reach with current technologies (e.g. pickpocket detection in low resolution images), and will focus on those concepts for which detection is potentially achievable.

Three lines of research will be conducted to address knowledge extraction: • At first, systematic evaluation of algorithms on the real-case data will be performed at each step of the extraction process, from both the algorithmic and the user point of view. This will apply to both existing algorithms (e.g. in tracking), and novel techniques developed within the project. • Secondly, we will conduct specific innovative research to address two essential issues of knowledge extraction from sensory data which are: 1. invariance recognition at the primitive level, and 2. partial detection and uncertainty handling at the composite event level. • Last, we will investigate new algorithms for knowledge discovery. Several directions will be considered to answer specific user needs: unsupervised sequence clustering techniques, outlier detection using robust statistics, and data mining techniques.

Last, we will develop two innovative user-centered ontology and event-driven demonstrators. The first one will provide web-service oriented access to low and mid-level semantic events detected in near real-time, demonstrating the operational mode capabilities of the project for information filtering (e.g. to help security agents focus on potential dangerous situations). A second offline retrieval system will allow users to query for combinations of higher semantic event information, statistics about transport and space usage, and detection of abnormal event. The two subsystems "Generic Event Recognition" and "Knowledge Modelling" are presented in Fig. 1.

3 Scientific research lines and contributions

The project essentially addresses three domains of research: knowledge modelling for scene understanding, surveillance systems and dedicated content extraction. We discuss the three of them below, presenting in each case the contribution of the project.

3.1 Knowledge modelling for scene analysis

All over the world, in many different contexts, video understanding systems have been used to extract knowledge about the scene they observe. In each domain, there is a need to establish a generalized "ontology" to describe what is observed in the scene [9]. An ontology is the set of all concepts and relations between concepts shared by the community of a given domain. The ontology is useful for experts of the application domain to use scene understanding systems in an autonomous way. It makes systems user-centered, and enables experts to fully understand the terms used to describe activity models. Moreover, the ontology is useful to evaluate scene-understanding systems, to understand exactly what types of events a particular system can recognize, and for developers desiring to share and reuse activity models dedicated to the recognition of specific events. However, most of the work in ontology deals with the structure of complex events (linguistic issues not addressing specifically video events) and with temporal reasoning. For the structure of complex events, Narayanan [20] has developed a formalism for the execution of actions and then applies it to several problems in linguistics. Several works have addressed the limitation of standard ontologies to represent time and temporal relationships. For instance, Hobbs [11] has developed a rich DAML ontology dedicated to time reasoning based on Allen temporal algebra [1]. A series of specific workshops sponsored by ARDA have been devoted to building ontologies of video events, for video understanding applications [4].

In CARETAKER, we will extend these video ontologies towards two directions: 1. the ontology will be adapted for urban surveillance applications; 2. the video ontology will be combined with an ontology for audio events, to enrich the description of scenarios of interest. Besides the knowledge representation issue, it is important to create user-friendly interfaces, to enable end-users to introduce context information about a new scene, add new scenarios adapted to a specific environment, and define the specification compliant with the given ontology. Only a few research have addressed this issue. Starting from the end-users requirements, such an interface will be developed based on an adapted temporal scenario representation.

3.2 Surveillance systems

In the past few years, many scene understanding systems have been developed in the computer vision community, mostly focusing on the tracking issue, e.g. through shape models [10, 16], or on human interaction [21]. Nevertheless, few scene understanding systems have been successfully applied to real world applications, due to a large variety of issues. First, typical image processing problems arise from shadows, illumination changes, over-segmentation or misdetection. Second, tracking remains a major issue, as the "loss" of a tracked object prevents the analysis of its behaviour. In addition, most of these systems address vision issues and few of them provide a true semantic scene understanding. Only a few research [21, 26] are able to perform complex (spatio-temporal) reasoning and to understand people interactions in real world applications. Still, the performances of these

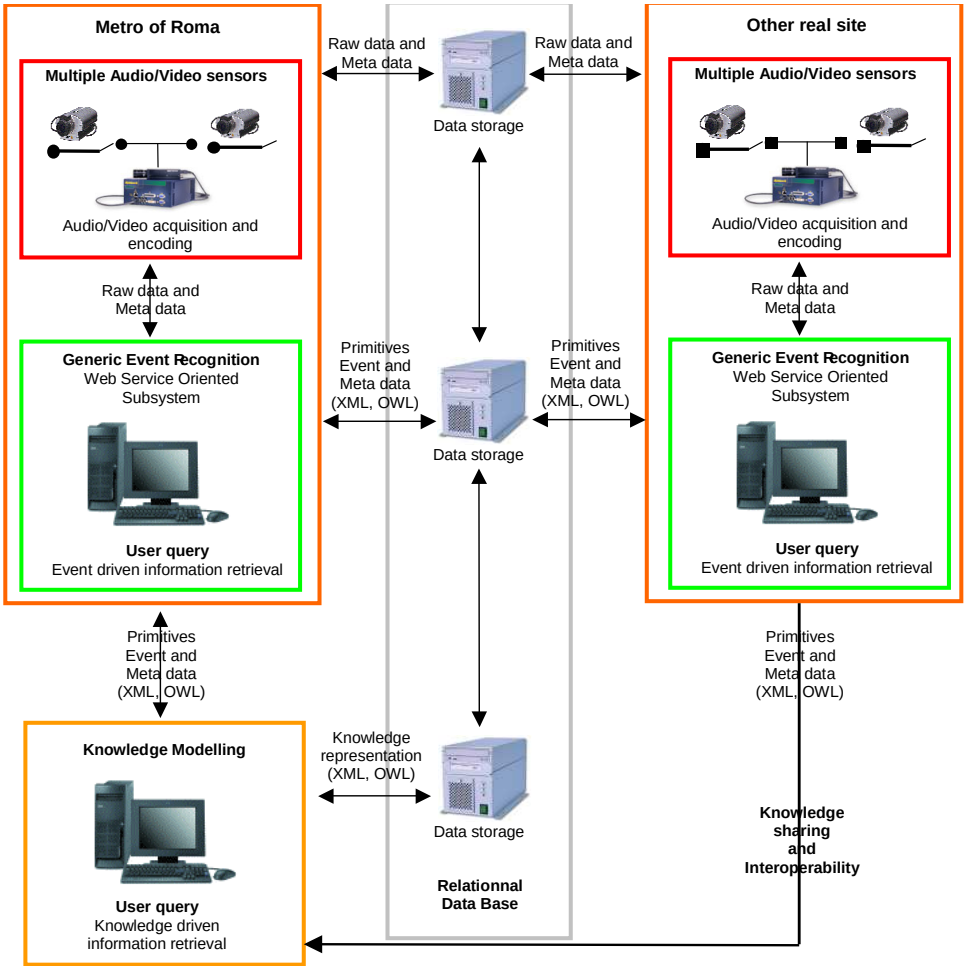

Figure 1: The CARETAKER project.

systems, which are usually good on small video sequence sets or in a well-constrained environment, significately degrade in real conditions.

A major innovation of CARETAKER in surveillance systems is to apply joint audio and video content analysis, knowledge modelling and extraction techniques in real-use cases with real data produced by a network of AV sensors. The project proposes a hierarchical modelling of the semantic content: a first layer of primitive events, extracted from the raw data streams and a second layer of higher semantic events from which rich meta-data can be produced. The semantic description will be used in two systems: a web service-oriented on-line prototype, demonstrating the validity of the primitive (and some mid-level) event extractor (the design of the on-line subsystem is illustrated in Fig. 2); and a second broader off-line system supporting spatio-temporal and semantic queries and statistical reporting will be developed with, and evaluated by, professional end-users to show the relevance of the targeted complete system (including the knowledge discovery components) on large amount of real test data from the metro of Roma.

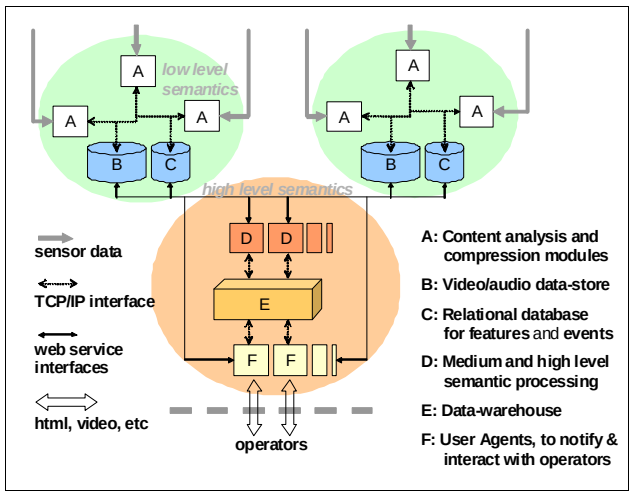

Figure 2: Diagram to illustrate the likely design configuration for the real-time subsystem.

3.3 Content extraction for scene understanding systems

In video-surveillance and monitoring applications, trajectories of people in the scene are one of the more obvious content to

extract. However, audio events, and video events that can be recognized without tracking, can bring significant advantages in the understanding of scene by bringing useful scene-specific context information. Finally, the recognition of more high-level events is highly desirable. In the subsequent paragraphs, we review the main issues and state-of-the-art techniques related to this important part of knowledge extraction.

3.3.1 Multi-object tracking (MOT)

Tracking is one of the most studied topics in dynamic scene analysis. Broadly speaking, there are three main components in the design of a tracking system. The first one is the observation model, i.e. the modelling of how well the data explains the presence of an object given its state (position, scale, ...). The second one is the prior model, which models the likeliness of object states. Although independent priors on single objects can be used, priors on the joint states are highly desirable in order to avoid two objects sharing the same physical part of the space, or to model that people tend to avoid each other. Finally, the third component is the search process, i.e. the mechanisms used to optimize a cost function with respect to the state values, or to search for the birth and death of new objects.

There is an abundance of literature devoted to MOT. In past years, state-space models [19, 24, 14, 25] have been shown to be the most successful. Although some methods choose to use a single-object state-space model [25], only a more rigorous formulation of the MOT problem [19, 24, 14, 18, 29] using a joint state space model allows object interactions and identity to be properly defined. In its simplest form, interactions can be defined based on proximity, occlusion being the so far most studied problem [19]. Recently proposed interaction models based on the use of priors over the joint state space constitute promising directions [18]. Tracking a variable number of objects with particle filters (PF) has been addressed in [24, 14]. Both works highlighted the need for a global observation model to deal with multi-object configurations varying in number. Algorithm efficiency has been addressed in [18], which proposed a model for tracking a fixed number of interacting objects using an efficient sampling method that combines a PF formulation with MCMC sampling. In [29], MCMC sampling with jump/diffusion dynamics was used to track multiple humans in a crowded scene.

To address the tracking issue, we will initially adopt a conservative approach, where the novel algorithms available in the consortium (e.g. [22] which combines the advantages of both [29] and [18]) will first be adapted to the specific environment (metro of Roma), and properly evaluated for the task. All the available context information will be employed (e.g. rough calibration of the ground-plane). In particular, one aim of CARETAKER is to ease and increase robustness with respect to system deployment: the addition of new cameras to the system should be doable by non tracking expert people.

3.3.2 Event recognition

Audio events The audio event detection problem is generally viewed as a classification task [27]. Such an approach needs to define all expected audio events, which is difficult to ensure in real cases. Another drawback of supervised classification is clearly the difficulty to collect enough samples for all the possible audio events. A closely-related issue is that sound signals of the same audio ontology event may strongly vary depending on the place where they are recorded (i.e. the same sound "train arriving" may depend on the microphone location or on the metro station). To handle these issues, CARETAKER proposes several original approaches. A first approach will consist in modelling first the normal audio ambiance related to a specific area [3], i.e. all the audio data that occurs with sufficient occurrences, and in restraining the problem of event detection to the detection of abnormal audio events (refer to knowledge discovery). A second approach will be based on probabilistic Maximum A Posteriori (MAP) techniques [8], which will allow to adapt a trained model to new data when only a very small number of samples are available. The method allows for both offline adaptation (e.g. using user feedback) and online adaptation. To increase invariance to spatial localization, the method will rely on robust audio feature units automatically detected using techniques derived from speech [5].

Video events In addition to trajectories, video events that can be recognized without tracking are important to provide contextual information (e.g. train arrives, doors open, occupancy of space) and to perform some task (e.g. counting). Moreover, it can be useful in detecting extreme cases where tracking may fail (too much crowd) and might be the only reliable solution to perform some analysis in these situations. Different techniques have been proposed in the past to perform such tasks, mainly through the analysis of spatio-temporal statistics [17, 7]. However, most of these research were not dedicated to the specific event types we will deal with in CARETAKER. Nevertheless, we will take advantage of them and focus our research on the recognition invariance of an ontological entity with respect to the viewed scene. This will be conducted by selecting low-level features presenting the most invariance to viewing conditions, and the use of MAP techniques [8] for adaptation, as explained for audio events.

3.3.3 Knowledge modelling and discovery

Significant progress has been made recently for the recognition of high-level semantic events from multiple streams of information. Broadly speaking, we can identify two main approaches: data-driven statistical methods, and rule-based (or scenario-based) systems.

Statistical approaches Dynamic Bayesian Networks, of which Hidden Markov Models (HMM) are a special case, have become the main framework within these data-driven approaches. They have proven robustness to data variations in the input streams and their generalization performance, as shown by their frequent use in activity recognition [21, 28]. A drawback, however, is that the amount of labelled training data they need usually increases exponentially with the complexity of the event. To handle this issue, several modelling techniques such as multi-streams HMM and layered-HMM [21, 28] have been proposed. In the latter case, the complexity is reduced by breaking the problem into a first layer, responsible for recognizing sub-events, and a higher layer, responsible for recognizing the complex event from the sub-event probabilistic output streams.

Ontology-based approaches Recognition engines based on ontology descriptions do not need training data in theory [13]. Moreover, they are better adapted to model the specification of an event defined by a user. However, most of today's available ontologies are based on the assumption of a strict classification system - an entity is either an instance of another entity or not - which makes them very sensitive to the quality of the primitive event detected from raw data. Even in the presence of low noise or ambiguity in the raw data, these approaches often break down, as they can not account for the partial detection of an event, and uncertainty handling becomes necessary [13].

Hybrid systems We currently observe a convergence of both frameworks in order to combine the advantages of the high-level modelling capabilities of the scenario-based approaches and the robustness of the statistical techniques [15, 12]. We will follow this highly promising research direction in event understanding, by exploiting the strength of HMM modelling approaches, and more specifically, the layered-HMM, in which one partner has a strong experience [2, 28], jointly with scenario-based modelling techniques developed by another partner [12].

Knowledge discovery Knowledge discovery is an essential function of a large scale video interpretation system. All events can not be predefined in advance, and the incorporation of new knowledge into a system must be feasible. This knowledge may come from the user, but not only. Data can also be used for such purpose. For instance, the unsupervised detection of the usual events from some data streams can be useful for pattern discovery (ontology refinement, trend identification containing relevant statistics for the user, such as the occupancy and use of specific spaces, or the flows of object), summarization, sequence indexing and retrieval. Few research have addressed this issue, and they were mainly applied to object trajectories (e.g. [23]). We plan to handle and combine more general data streams within the layered-HMM framework.

The automatic detection of unusual temporal events, which can be defined as seldom occurring (rarity) and not having been thought of in advance (unexpectedness), constitutes a problem which has recently attracted attention in computer vision and multimodal processing under a range of names (abnormal, unusual, or rare events) [23, 6]. It is clear from their definition that unusual event detection entails a number of challenges. The rarity of an unusual event means that collecting sufficient training data for supervised learning will often be infeasible. In addition, more than one type of unusual events may occur in a given data sequence, where the event types can be expected to differ markedly from one another. This implies that training a single model to capture all unusual events will generally be impossible. As well as such modelling problems due to rarity, the unexpectedness of unusual events means that defining a complete event lexicon will not be possible in general. In CARETAKER, we expect to address these issues by considering unusual event recognition as an outlier detection problem, and by training through MAP adaptation a statistical model on the data identified as the outlier.

4 Conclusion

CARETAKER will thus model and account for two types of knowledge: on one side, the multi-user knowledge (safety operators, decision makers), represented by their needs, their use-case scenarios definitions, and their abilities at providing context description for sensory data; on the other side, the content knowledge, characterized by a first layer of primitive events that can be extracted from the raw data streams, such as ambient sounds, crowd density estimation or object trajectories, and a second layer of higher semantic events, defined from longer term analysis and from more complex relationships between both primitive events and higher-level events. Both knowledge types will be modeled through ontologies and exploited by the content extraction methodologies. The latter will be based on an innovative approach, whereby probabilistic models will be associated with each ontological entity, allowing to take full advantage of both statistical data-driven and scenario-based reasoning approaches, allowing for effectiveness, flexibility, and robustness with respect to sensor deployment conditions. Extracted metadata will be incorporated in knowledge management systems providing web-based content access and automatic knowledge discovery retrieval capabilities.

Acknowledgements

This research project is supported by the European Commission under the 6th Framework Programme through the Activity: Semantic-based Knowledge and Content Systems (contract no.: FP6-027231). For further information about the CARETAKER project, please visit `http://www.ist-caretaker.org/`.

References

[1] J. Allen and G. Ferguson. *Actions and Events in Interval Temporal Logic*, pages 205–245. Spatial and Temporal Reasoning, Kluwer Academic Publishers, 1997.

[2] S. Bengio. Multimodal speech processing using asynchronous hidden Markov models. *Information Fusion*, 5(2):81–89, 2004.

[3] A. Bregman. *Auditory Scene Analysis*. 1990.

[4] F. Bremond, N. Maillot, M. Thonnat, and T. Vu. Ontologies for video event. Technical report 5189, INRIA, Sophia Antipolis, France, May 2004.

[5] J. Cernocky. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*. Phd. thesis, Université Paris XI, 1998.

[6] M. Chan, A. Hoogs, J. Schmiederer, and M. Perterson. Detecting rare events in video using semantic primitives with HMM. In *Int. Conf. on Pattern Recognition*, 2004.

[7] R. Fablet and P. Bouthemy. Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 25(12):1619–1624, December 2003.

[8] J. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains. *IEEE Trans. on Speech Audio Processing*, 2(2):291–298, April 1994.

[9] A. Hakeem and M. Shah. Ontology and taxonomy collaborated framework for meeting classification. In *Int. Conf. on Pattern Recognition*, August 23-26 2004.

[10] I. Haritaoglu, D. Harwood, and L. Davis. w^4: real-time surveillance of people and their activities. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 22:809–830, 2000.

[11] J. Hobbs. A DAML ontology of time. http://www.cs.rochester.edu/ ferguson/daml/.

[12] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, November 2004.

[13] R. Howart. Interpreting a dynamic and uncertain world: Task-based control. *Art. Intell.*, 100(1-2):5–86, 1998.

[14] M. Isard and J. MacCormick. BraMBLe: A bayesian multiple-blob tracker. In *Int. Conf. on Comp. Vis.*, 2001.

[15] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 22(8):852–872, 2000.

[16] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing*, 14(8):609–615, 1996.

[17] D. Keren. Recognizing image style and activities in video using local features and naive Bayes. *Pattern Recognition Letters*, 24(16):2913–2922, December 2003.

[18] Z. Khan, T. Balch, and F. Dellaert. An mcmc-based particle filter for tracking multiple interacting targets. In *European Conf. on Computer Vision*, 2004.

[19] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Int. Conf. on Comp. Vis.*, September 20-27 1999.

[20] S. Narayanan. *KARMA: Knowledge based Actions Representations for Metaphor and Aspect*. Phd. thesis, University of California at Berkeley, CA, USA, 1997.

[21] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 22(8):831–843, August 2000.

[22] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, June 20-25 2005.

[23] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 22:747–757, 2000.

[24] H. Tao, H. Sawhney, and R. Kumar. A sampling algorithm for detecting and tracking multiple objects. In *Int. Workshop on Vision Algorithms*, 1999.

[25] D. Tweed and A. Calway. Tracking many objects using subordinate condensation. In *British Machine Vision Conf.*, Cardiff, UK, September 2-5 2002.

[26] T. Vu, F. Brémond, and M. Thonnat. Automatic video interpretation: a novel algorithm for temporal scenario recognition. In *Int. Joint Conf. on Artificial Intelligence*, Acapulco, Mexico, August 9-15 2003.

[27] J. Woodard. Modelling and classification of natural sounds by product code hidden Markov models. *IEEE Trans. on Signal Processing*, 40(7):1833–1835, 1992.

[28] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: a two-layer HMM framework. In *IEEE Workshop on Event Mining at the Conf. on Computer Vision and Pattern Recognition*, 2004.

[29] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 27 June-2 July 2004.