Multisensor Fusion for Monitoring Elderly Activities at Home

Nadia ZOUBA, Francois BREMOND, Monique THONNAT
INRIA Sophia Antipolis - Mediterranee, PULSAR Team
2004 route des Lucioles, BP 93, Sophia Antipolis, France
Nadia.Zouba@sophia.inria.fr, Francois.Bremond@sophia.inria.fr, Monique.Thonnat@sophia.inria.fr

***Abstract*—In this paper we propose a new multisensor based activity recognition approach which uses video cameras and environmental sensors in order to recognize interesting elderly activities at home. This approach aims to provide accuracy and robustness to the activity recognition system. In the proposed approach, we choose to perform fusion at the high-level (event level) by combining video events with environmental events. To measure the accuracy of the proposed approach, we have tested a set of human activities in an experimental laboratory. The experiment consists of a scenario of daily activities performed by fourteen volunteers (aged from 60 to 85 years). Each volunteer has been observed during 4 hours and 14 video scenes have been acquired by 4 video cameras (about ten frames per second). The fourteen volunteers were asked to perform a set of household activities, such as preparing a meal, taking a meal, washing dishes, cleaning the kitchen, and watching TV. Each volunteer was alone in the laboratory during the experiment.**

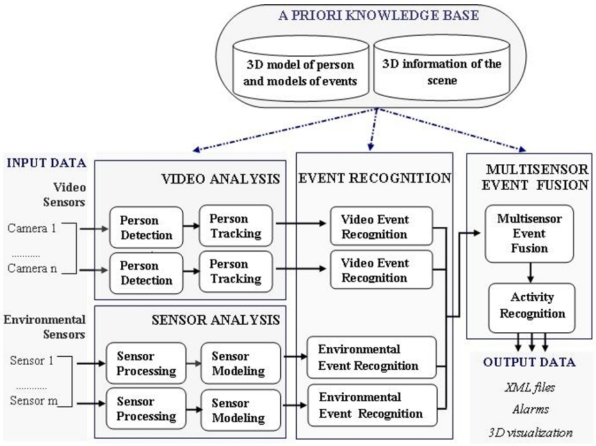

Figure 1. Architecture of the Proposed Approach

I. Introduction

Human activity recognition is an important part of cognitive vision systems because it provides accurate information about the behavior of the observed people. A major goal of current computer vision research is to recognize and understand human motion, short-term activities and long-term activities. The application areas for these vision systems are mostly surveillance and safety. Activity recognition is becoming also important in the application area of healthcare. In this paper, an approach combining heterogeneous sensor data for recognizing elderly activities at home is presented. In this approach we propose to combine data provided by video cameras with data provided by environmental sensors to monitor the interaction of people with the environment. We also propose an adapted description language to let users (i.e. medical staff) to describe the activities of interest into formal models. The proposed approach aims to recognize a large number of activities at home. The environmental sensors we use are attached to house furnishings. They are easy to install in home environments and removable without damage to the cabinets or furniture. The proposed sensors require no major modifications to existing homes and can be easily retrofitted in real home environments.

As described in Fig. 1, the input of the proposed multisensor approach consists in the data provided by the different sensors. Its output is a set of XML files and alarms stored in a database and also a 3D visualization of the recognized events. The proposed approach exploits three major sources of knowledge: the 3D model of person (e.g. 3D size of person), the models of events predefined in collaboration with gerontologists and the 3D information of the scene (e.g. position and size of furniture, zones of interest). The paper is organized as follows. In section II, we present the related work in the area of monitoring human activities at home. In section III, we briefly present the video analysis task. Section IV presents the environmental sensor analysis task. In this section we describe the proposed sensor model which is necessary to combine heterogeneous sensor data. Then section V describes the event recognition approach and how the events are modeled. Section VI presents the multisensor event fusion approach. Section VII presents our experiments and the obtained results. In this section we describe which sensors are used and why. Finally, section VIII presents our conclusion and the future work.

II. Related work

Over the last decade, much effort has been put into developing and employing a variety of sensors to monitoring activities at home, including camera networks for people tracking [1], cameras and microphones for activity recognition [2], [3], and embedded sensors for activity detection [4], [5].

There has also been a significant amount of research work in the area of recognition of Activities of Daily Living (ADLs). Recently, Jesse Hoey et al. [6] successfully used only cameras to assist person with dementia during handwashing. The system uses only video inputs, and combines a Bayesian sequential estimation framework for tracking hands and towel, with a decision using a partially observable Markov decision process. Several projects have investigated the use of different sensors to provide a "smart" environment for the observation of activities of daily living (ADLs). Examples include Georgia Tech's "Aware Home" [7], Imperial College's UbiMon system [8], SAPHE project [9], the Welfare-Techno house in Japan [10] and MIT's PlaceLab [11]. Most of these systems have been limited in the variety of activities they recognize and their robustness to noise. In particular, most of them use sensors that provide only a very coarse idea of what is going on. For example, by detecting only movement in a room, it is not possible to detect which activity occurs in the room. In this paper we propose an approach for activity recognition that addresses these problems by combining the use of video cameras with environmental sensors to determine when a person uses the household equipment and to detect most of the activities at home. Our work differs from the other works in two key ways. First, we propose to combine video cameras with environmental sensors embedded in the home infrastructure in order to recognize a set of interesting elderly activities at home. Second, we let users (i.e. medical staff) to describe the activities of interest into formal models by using an adapted description language [12].

III. Video Analysis

Video analysis aims at detecting and tracking people evolving in the scene. To achieve this task, we have used a set of vision algorithms coming from a video interpretation platform described in [13]. A first algorithm segments moving pixels in the video into a binary image by subtracting the current image with the reference image. The reference image is updated through out the time to take into account changes in the scene (e.g light, object displacement, shadows). The moving pixels are then grouped into connected regions, called blobs. Using calibrating information, a set of 3D features such as 3D position, width and height are computed for each blob. Then the blobs are classified into predefined classes (e.g. person). After that the tracking task associates to each new classified blob a unique identifier and maintains it globally throughout the whole video. Fig. 2 illustrates the detection, classification and tracking of a person.

IV. Sensor Analysis

In this section, we firstly describe the environmental sensor processing part, after that we describe the proposed sensor model which is necessary to fuse multisensor systems.

A. Environmental Sensor Processing

The physical sensor (e.g. electrical sensor) produces a response to the surrounding environment. For instance the electrical sensor triggers a signal when an appliance is used. The raw data collected by the physical sensors is processed to produce high-level representations of sensed object. This process converts the physical sensor response into a representative value of the raw environmental characteristics, such as electrical current.

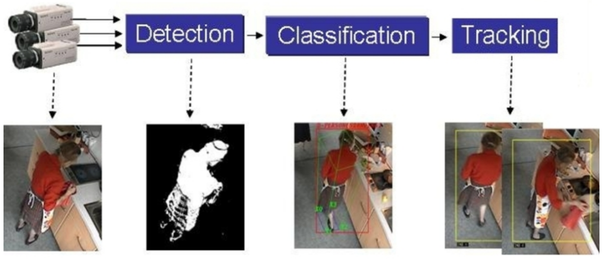

Figure 2. Detection, classification and tracking of a person.

B. Sensor Modeling

A sensor is characterized by various parameters such as the zone it covers, the precision of its measurement through this zone, its placement and the perturbations to which it is sensitive. The covered zone can be very variable depending on the sensors. For a camera, this zone is the field of view and for a contact sensor this zone is reduced to a point.

Because each sensor type has different characteristics and functional description, it is necessary to find a general model that is independent of the physical sensors, and that enables comparison of the performance and robustness of such sensors. For solving this issue we propose a generic sensor model in order to develop a coherent and efficient representation of the information provided by sensors of different types. This representation provides a means for recovery from sensor failure and also facilitates reconfiguration of the sensor system when adding or replacing sensors. In this work, we consider five attributes associated with each sensor observation:

- **Measurement Type M:** This includes the name of the physical property (e.g. sound, light, pressure) which is measured by the sensor and the units in which it is measured.
- **Sensor Location x:** This is the position of the sensor in the scene referential.
- **Time t:** This is the time when the physical property is measured. In real-time systems the timestamp of a measurement is often as important as the value itself.
- **Measurement y:** This is the value of the physical property as measured by the sensor. The physical property may have more than one dimension and this is the reason we represent it as a vector y.
- **Uncertainty Δy:** This is a generic term and includes many different types of errors relatively to y, including measurement errors, calibration errors and sensor

failure errors.

Symbolically we represent a sensor observation using the following 5-tuples: $O =< M, x, t, y, \Delta y >$.

Despite the growing research interest in monitoring activities at home, relatively little work has been carried out in extending them to encompass the management of uncertain information [14]. Sources of uncertainty include uncertain measurement, uncertain time and uncertain event. In the proposed sensor model we only take into account measurement uncertainty. In our approach we propose to model measurement uncertainty by using probability density function (pdf) of measurement, where the pdf's mean value and variance correspond to the measurement estimate and the measurement uncertainty respectively. To calculate the estimated value and variance of the measurement, we use the discrete probability density distribution. Under this model, the value of a given measurement is represented as a collection of alternative values, each with an associated probability.

V. Event Recognition

In order to express the semantics of the activities a modeling effort is needed. The models correspond to the modeling of all the knowledge needed by the system to recognize events occurring in the scene. To allow security operators to easily define and modify their models, the description of the knowledge is declarative and intuitive (in natural terms). In this work, we propose to represent the activities of interest into a formal model that satisfies a number of constraints by using the event description language proposed by Vu et al. [12]. We have extended this language to address complex activity recognition involving several physical objects of different types (e.g. person, chair) in a scene observed by video cameras and environmental sensors and over an extended period of time.

A. Model of Events

The event models correspond to the modeling of all the knowledge used by the system to detect events occurring in the scene. The description of this knowledge is declarative and intuitive (in natural terms), so that the experts of the application domain can easily define and modify it. Four types of event (called **components**) can be defined: primitive states, composite states, primitive events and composite events. A **primitive state** (e.g. a person is located inside a zone) corresponds to a perceptual property characterizing one or several physical objects (i.e. actors). A **composite state** is a combination of primitive states. A **primitive event** corresponds to a change of primitive state values (e.g. a person changes a zone). A **composite event** is a combination of primitive states and/or primitive events.

An event model M of an event E is composed of five parts: "**physical objects**" (a set of variables whose values correspond to the physical objects involved in E), "**components**" (a set of variables whose values correspond to the event instances composing E), "**forbidden components**" (a set of variables corresponding to all event instances that are not allowed to be recognized during the recognition of E), "**constraints**" (a set of conditions between the physical objects and/or the components to be verified for the recognition of E, they include symbolic, logical, spatial and temporal constraints (Allens interval algebra operators [15])), and "**alerts**" (an optional part of an event model which corresponds to a set of actions to be performed when E is recognized).

B. Examples of Event

We have modeled several primitive states, primitive events and composite events. In particular, we have modeled ten video events related to the location of the person in the scene (e.g. inside kitchen, inside livingroom, inside bedroom, inside entrance, stay inside kitchen). We have also modeled ten environmental events related to the status of various house furnishings (e.g. drawer is open, chair is pressed, stove is on).

C. Event Recognition

The event recognition process we used [12] is able to recognize which events are occurring in the scene at each instant. To benefit from all the knowledge, the event recognition process uses the coherent tracked mobile objects, the a priori knowledge of the scene and the predefined event models. To be efficient, the recognition algorithm processes in specific ways events depending on their type. Moreover, this algorithm has also a specific process to search previously recognized events to optimize the whole recognition. The algorithm is composed of two main stages. First, at each step, it computes all possible primitive states related to all mobile objects present in the scene. Second, it computes all possible events (i.e. primitive events then composite states and events) that may end with the recognized primitive states.

VI. Multisensor Event Fusion

The sensor fusion can be classified into different levels according to the input and output data types [16]. The fusion may take place at the data level, feature level and decision level. In the data level fusion, the raw output data of sensors are combined. In the feature level fusion, each sensor provides observational data from which a feature vector is extracted. These vectors are then concatenated together into a single feature vector. The decision level fusion involves combination of sensor high level output data (e.g. event). The use of sensor fusion at the decision level facilitates an extensible sensor system, because the number and types of sensors are not limited. In our approach, we use a fusion process at the decision level (i.e. event level) to address the problem of heterogeneous sensor system. For this, we combine the video events with the environmental events described above (section V-B) in order to detect more rich and complex events. The use of an heterogeneous sensor system involves a synchronization task to cope with the different output data frequencies of the sensors. To solve this issue, we currently use different configurations of delays between components composing a complex event. More precisely, we define different event models corresponding to variations of delays between environmental and video sensor outputs. The multisensor event recognition algorithm takes as input the sensor events and the a priori knowledge of complex events to be recognized. An event model M should be recognized at an instant t if all its components have been recognized, its last (using the temporal order) component being recognized at the given instant t.

For this experimentation, we have modeled twelve household activities (corresponding to ADLs) including: using the fridge, using the cupboards, using the drawers, using the microwave, using the stove, watching TV, washing dishes, taking a meal, and preparing a meal. An example of the modeling event ”taking a meal” is presented in table I.

In this example, ”Taking a meal” model contains four physical objects (i.e. person, zone1, zone2, equipment1 and equipment2), five components, eight constraints and one alert. The components include the location of the person in the livingroom, close to table, the pressed state of the chair and the sitting posture of the person in the livingroom. The constraints include 4 spatial constraints related to the zone and the equipments involved in the event, and 4 temporal constraints including the duration of the sub-events. When these components occurred and all the constraints are verified, the taking meal event is recognized and an alert is triggered.

Table I
TAKING A MEAL MODEL

CompositeEvent(TakingMeal,
PhysicalObjects((p : Person), (z1 : Zone), (z2 : Zone),
(eq1 : Equipment), (eq2 : Equipment))
Components((c1 : PrimitiveState InLivingroom(p, z2))
(c2 : PrimitiveState CloseToTable(p, eq1))
(c3 : CompositeState ChairPressed(p, eq2))
(c4 : CompositeState PersonSeatedInLivingroom(p, z2))
(c5 : CompositeEvent PreparingMeal(p, z1)))
Constraints ((z1's **Name =** Kitchen),
(z2's **Name** = Livingroom)
(eq1's **Name** = table),
(eq2's **Name** = chair)
(c2 **Duration** >= threshold1)
(c3 **Duration** >= threshold2),
(**Start** of c4 > **End** of c5),
(c4 **Duration** >= threshold3))
Alert(AText (”Person takes a meal”)
AType(”NOTURGENT”))

VII. Experimental Results

In this section, we describe which sensors we are used and why. First, we discuss the overall monitoring goals. After that, we list the sensors used in these experiments and present the experimental laboratory we have built for these experiments. Finally, we show and discuss the obtained results.

A. Monitoring Goals

Monitoring activities at home is predominantly composed of location and activity information. Below is a list of exactly what we wish to automatically recognize.

- **Presence:** Determine whether one or several individuals are present in the environment.
- **People Tracking:** Determine the location of each person (e.g. in the kitchen).
- **Motion:** Recognize whether and how a person is moving (e.g. walking).
- **Activities of Daily Living (ADLs):** Recognize daily activities such as cooking, eating, bathing, toileting [17], [18].

B. Sensor Choice and Placement

For the experiments, we choose to use commonly available sensors that they do not have to be worn or carried (non intrusive). The selected sensors can easily and quickly be installed in home environments and are removable without damage to the cabinets or furniture.

- **Video cameras:** These sensors were used to detect and track people evolving in the scene. They are installed in all rooms but bathroom to locate people at each time.
- **Contact sensors:** These inexpensive magnetic contact sensors indicate a closed or open status. They are placed on drawers, cupboards, cabinets and fridge. These sensors are also useful in determining the interaction with kitchen furnitures, such as cupboards, drawers, and fridge.
- **Pressure sensors:** These sensors are used to detect presence on chairs and beds. They are placed under chairs, armchairs, and bed.
- **Water flow sensors:** When placed in water pipes these sensors trigger a signal when flow exceeds some thresholds. They are placed on hot and cold water pipes and toilets.
- **Electrical sensors:** These sensors measure consumption of the current flow in a circuit, reporting when current exceeds some thresholds, e.g., whenever an appliance is used. They are placed on electrical outlets, to monitor the amount of current flowing to circuits.
- **Presence sensors:** These sensors are installed in front of the sink, the cooking stove and the washbowl to detect the presence of people nearby.

A laboratory (called GERHOME) equipped with the different sensors previously cited has been built to evaluate the performance of the multisensor system and to explore the activities that can be recognized by such a computer system. Fig. 3 shows some pictures, and a 3D visualization of the Gerhome laboratory. This laboratory looks like a typical apartment of an elderly person: 41m2 with entrance, livingroom, bedroom, bathroom, and kitchen. The kitchen includes an electric stove, a microwave, a fridge, cupboards, and drawers. See Fig. 4 for an overview of a typically instrumented home.

Figure 3. Gerhome laboratory. (a) The livingroom; (b) the kitchen; (c) the 3D visualization of the kitchen.

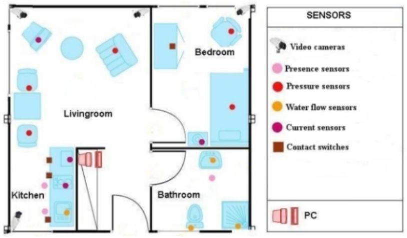

Figure 4. Overview of a typically instrumented home.

C. Experiment and Obtained Results

While evolving in the Gerhome laboratory, fourteen volunteers (aged from 60 to 85 years) have been observed, each one during 4 hours to measure the accuracy of the detected events recognized by the multisensor system. The fourteen volunteers were asked to perform a set of household activities, such as preparing a meal, taking a meal, washing dishes, cleaning the kitchen, and watching TV.

Among all analyzed data, preliminary results for one volunteer observed during 4 hours are shown in table II. This table summarizes the ground truth (GT), the true positive (TP), the false negative (FN), the false positive (FP), the precision (P) and the sensitivity (S) of the recognition of a set of states and events. $P = TP/(TP + FP)$ and $S = TP/GT$.

The primitive states ”in the kitchen” and ”in the livingroom” are well recognized by video cameras. The few errors in the recognition occur at the border between livingroom and kitchen. These errors are due to noise and shadow problems. In the other events the errors in the recognition are sometimes due to noise and shadow problems and sometimes due to the sensor measurement errors (e.g. pressure sensor active when a person puts something on the chair).

Table II
RESULTS FOR RECOGNITION OF A SET OF STATES AND EVENTS

States and events	GT	TP	FN	FP	P	S
In the kitchen	12	8	4	2	80%	66%
In the livingroom	20	17	3	4	81%	85%
Using fridge	10	8	2	3	72%	80%
Using stove	6	4	2	2	66%	66%
Preparing meal	1	1	0	1	50%	100%
Taking meal	1	1	0	1	50%	100%

Fig. 5 and Fig. 6 show respectively the recognition of ”preparing a meal” and ”taking a meal” activities and the 3D visualization of these recognitions.

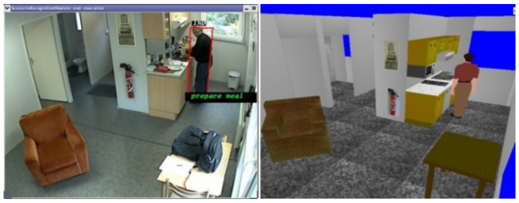

Figure 5. (a) Recognition of ”preparing a meal” activity. (b) the 3D visualization of this recognition.

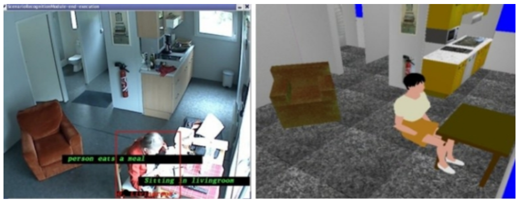

Figure 6. (a) Recognition of ”taking a meal” activity. (b) the 3D visualization of this recognition.

Preliminary results, comparing two elderly people (volunteer 1 and volunteer 2), observed during 4 hours are shown in table III. This table summarizes the Mean duration M1 and M2 (min:sec) of an instance, the Instance Number IN (N1 and N2), the Normalized Difference ND, the Normalized Difference of mean durations of Activities NDA, and the Normalized Difference of Instance number NDI.

Table III
RESULTS OF RECOGNITION OF A SET OF ACTIVITIES COMPARING TWO ELDERLY PEOPLE (VOLUNTEER 1 (64 YEARS) AND VOLUNTEER 2 (85 YEARS))

	Volunteer 1		Volunteer 2		ND	
Activity	**M1**	**IN (N1)**	**M2**	**IN (N2)**	**NDA**	**NDI**
Using fridge	0:12	14	0:13	5	4%	47%
Using stove	0:08	35	0:16	102	33%	49%
Sitting on chair	6:07	12	92:42	2	87%	71%
Standing	0:09	200	0:16	45	28%	63%
Bending	0:04	30	0:20	15	67%	33%

The Normalized Difference of mean durations of Activities NDA has been defined by the formula $NDA = |M1 - M2|/(M1 + M2)$; and the Normalized Difference of Instance number NDI has been defined by the formula $NDI = |N1 - N2|/(N1 + N2)$.

Table III shows some difference in the behavior of the two volunteers. For example, the volunteer 1 was sitting on chair more often than the volunteer 2 (for sitting on chair 12 vs. 2, NDI=71%) and the volunteer 2 was sitting on chair for a longer duration than the volunteer 1 (92:42 vs. 6:07, NDA=87%), showing a greater ability for the volunteer 1 to move in the apartment. Similarly volunteer 1 was bending twice more than volunteer 2 (for bending 30 vs. 15, NDI=33%) and in a quicker way (0:04 vs. 0:20, NDA=67%) showing greater dynamism for the volunteer 1. Also the volunteer 1 was more able to use the stove (i.e. less trials, for stove use 35 vs. 102, NDI=49%) than the volunteer 2.

All these measures show the greater ability, during the 4 hours period, of the volunteer 1, comparing to the volunteer 2, to live in the apartment and to use the house equipment. The obtained results demonstrate that the proposed method allows detecting and recognizing of a set of activities of a person by using the data provided by the selected sensors.

VIII. Conclusion

In this paper we have proposed an approach to recognize elderly activities at home based on multisensor data fusion. The approach combines video events with environmental events in order to recognize activities of interest and to optimize the use of sensors depending on the monitoring scenario. The main contribution of this work relies in the combination of the environmental and the video information at the event level. Another contribution consists in the adapted description language which allows users (i.e. medical staff) to describe activities of interest into a formal models.

More evaluation especially on long periods is required to assess the robustness of the proposed system. Future works also include learning event models and modeling their uncertainty. We also plan to improve time synchronization method in order to increase accuracy in sensor fusion systems.

References

[1] H. Sidenbladh and M. Black, "Learning image statistics for bayesian tracking." in *IEEE International Conference on Computer Vision (ICCV).*, 2001.

[2] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory context awareness via wearable computing." in *Proceedings of the Perceptual User Interfaces Workshop (PUI).*, 1998.

[3] D. Moore, I. Essa, and M. Hayes, "Exploiting human actions and object context for recognition tasks." in *Proceedings of IEEE International Conference on Computer Vision (ICCV99).*, 1999.

[4] S. Wang, W. Pentney, A.-M. Popescu, T. Choudhury, and M. Philipose, "Common sense joint training of human activity recognizers." in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI).*, January 2007.

[5] T. Moeslund, A. Hilton, and V. Krger, "A survey of advances in vision based human motion capture and analysis." *Behavior Research Methods, Instruments, and Computers*, vol. 32, no.3, 2000.

[6] J. Hoey, A. V. Bertoldi, and A. Mihailidis, "Assisting persons with dementia during handwashing using a partially observable markov decision process." in *International Conference on Computer Vision Systems (ICVS).*, March 2007.

[7] G. Abowd, A. Bobick, I. Essa, E. Mynatt, and W. Rogers, "The aware home: Developing technologies for successful aging." in *Proceedings of AAAI Workshop and Automation as a Care Giver: The Role of Intelligent Technology in Elder Care.*, July 2002.

[8] G. Yang, B. Lo, J. L. Wang, M. Rans, S. Thiemjarus, J. Ng, P. Garner, S. Brown, B. Majeed, and I. Neid, "From sensor networks to behavior profiling: A homecare perspective of intelligent building." in *The IEE Seminar for Intelligent Buildings.*, 2004.

[9] Saphe, "Saphe project," http://ubimon.doc.ic.ac.uk/saphe/, 2006.

[10] T. Tamura, "Biomedical engineering at the forefront in japan," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 24, no.4, pp. 23–26, 2005.

[11] D. Cook and S. Das, "How smart are our environments? an updated look at the state of the art." *Pervasive Mobile Computer*, vol. 3, no.2, pp. 53–73, 2007.

[12] V. Vu, F. Bremond, and M. Thonnat, "Automatic video interpretation: A novel algorithm based for temporal scenario recognition." in *The Eighteenth International Joint Conference on Artificial Intelligence.*, September 9-15 2003.

[13] A. Avanzi, F. Bremond, C. Tornieri, and M. Thonnat, "Design and assessement of an intelligent activity monitoring platform." *EURASIP Journal on Applied Signal Processing, Special Issue on 'Advances in Intelligent Vision Systems: Methods and Applications'.*, August. 2005.

[14] M. S. Ryoo and J. K. Aggarwal, "Human activities: Handling uncertainties using fuzzy time intervals." in *Proceedings of 19th International Conference on Pattern Recognition (ICPR).*, December 2008.

[15] J. F. Allen, "Maintaining knowledge about temporal intervals." in *In Communications of the ACM.*, 1983.

[16] B. V. Dasarathy, "Sensor fusion potential exploitation innovative architectures and illustrative applications." in *Proceedings of the IEEE*, January 1997.

[17] S. Katz, A. B. Ford, R. W. Moskowitz, B. A. Jackson, and M. W. Jaffe, "Studies of illness in the aged: The index of adl: A standardized measure of biological and psychosocial function." *Journal of the American Medical Association*, vol. 185, no.12, pp. 914–919, 1963.

[18] M. P. Lawton, "Aging and performance of home tasks." *Human Factors*, vol. 32, pp. 527–536, 1990.