Combining Multiple Sensors for Event Recognition of Older People

Carlos F. Crispim-Junior,
INRIA – Sophia Antipolis
2004 Route des Lucioles,
Sophia Antipolis, France
carlos-fernando.crispim_junior@inria.fr

Baptiste Fosty
INRIA – Sophia Antipolis
2004 Route des Lucioles,
Sophia Antipolis, France
baptiste.fosty@inria.fr

Rim Romdhane
INRIA – Sophia Antipolis
2004 Route des Lucioles,
Sophia Antipolis, France
rim.romdhane@inria.fr

Qiao Ma,
INRIA – Sophia Antipolis
2004 Route des Lucioles,
Sophia Antipolis, France
maqiao909@gmail.com

Francois Bremond
INRIA – Sophia Antipolis
2004 Route des Lucioles
Sophia Antipolis, France
francois.bremond@inria.fr

Monique Thonnat
INRIA – Sophia Antipolis
2004 Route des Lucioles
Sophia Antipolis, France
monique.thonnat@inria.fr

ABSTRACT

We herein present a hierarchical model-based framework for event recognition using multiple sensors. Event models combine *a priori* knowledge of the scene (3D geometric and semantic information, such as contextual zones and equipments) with moving objects (*e.g.,* a Person) detected by a monitoring system. The event models follow a generic ontology based on natural language; which allows domain experts to easily adapt them. The framework novelty relies on combining multiple sensors at decision (event) level, and handling their conflict using a probabilistic approach. The proposed approach for event conflict handling computes the event reliability for each sensor, and then combine them using Dempster-Shafer Theory with an alternative combination rule. The proposed framework is evaluated using multi-sensor recording of instrumental daily living activities (*e.g.*, watching TV, writing a check, preparing tea, organizing week intake of prescribed medication) of participants of a clinical trial for Alzheimer's disease. Two evaluation cases are presented: the combination of events (or activities) from heterogeneous sensors (RGB ambient camera and a wearable inertial sensor) by a deterministic fashion, and the combination of conflicting events recognized by video cameras with partially overlapped field of view (a RGB- and a RGB-D-camera, Kinect®). The results show the framework improves the event recognition rate in both cases.

Categories and Subject Descriptors

I.5.1 Models: Deterministic, Statistical; I.5.4 Applications: Computer vision

General Terms

Algorithms, Reliability, and Experimentation

Keywords

Event Recognition, Multi-sensor Fusion, Sensor Reliability

MIIRH'13, October 22, 2013, Barcelona, Spain.

1. INTRODUCTION

Human activity recognition research field has been experiencing a continuous evolution in the last decade. Computer Vision, Wearable and Ubiquitous computing research fields have proposed several methods to cope with the challenges brought by unconstrained environments of real life, such as illumination changes, moving cameras, and outdoor scenes. Activity (or Event) recognition has been studied for safety and security applications, such as older people monitoring at home, video surveillance and crime prevention; enablement and support of human tasks (e.g., in case of loss of a body limb function), and as tools to support objective assessment of emerging symptoms of diseases (medical diagnosis).

Lavee *et al*. [10] categorizes computer vision approaches for event recognition in three categories: State models, Pattern Recognition methods, and Semantic models. All three approaches are generally based on at least one of the following data abstraction levels: pixel-based, feature-based, or event-based. State models refer to techniques such as Conditional Random Fields, Dynamic Bayesian Networks, and Hidden Markov Models. Pattern Recognition methods are Artificial Neural Networks, Support-Vector Machines (SVM), Nearest Neighbor, etc. In this context, Le *et al.* [11] have presented an extension of the Independent Subspace Analysis algorithm applied at learning invariant spatio-temporal features from unlabeled video data for activity recognition. Wang *et al.* [17] have proposed new descriptors for dense trajectory estimation, which are later used as input for a non-linear SVM. Although these techniques have considerably increased the activity recognition performance in benchmark datasets, they extract information from pixel-based and feature-based abstractions, what poses limitations concerning their ability of describing the semantic and hierarchical nature of complex activities. Izadinia and Shah [9] have presented a method for learning low-level events from data, to later identify complex events from the joint relationship among the detected events by using a graph representation and a discriminative model.

Alternatively, Semantic (or Description-based) models use a descriptive language and logical operators to build event representations using domain expert knowledge. Its hierarchical nature allows the explicit modeling of semantic information, and they do not require as much data as Pattern Recognition and State models methods. Zaidenberg et al. [19] have presented a generic framework for activity recognition of group behaviors in an airport, a subway, and shopping center scenarios. However, one limitation of semantic models is their sensitivity to noise of underlying vision process, like image segmentation and people detection algorithms.

Ubiquitous and Pervasive computing fields have also been active at event recognition research. They have proposed data fusion of multiple sensors for the recognition task, such as inertial sensors (*e.g*., of accelerometers and gyroscopes), ambient sensors (*e.g.,* passive infrared sensors, change of state sensors, audio), with and without video cameras to monitor the daily living activities of a person. Gao *et al.* [8] have demonstrated the fusion of inertial sensors data worn at the waist, chest, thigh, and side of a person body using a Naïve Bayes Classifiers. See also Rong and Ming [15]. Disadvantages of inertial sensors approaches are motion noise and inter sensor-calibration, and the assumption that the sensors are always placed at the same body position, generally causing noise in large scale research studies.

Fleury *et al*. [6] have presented a multi-modal system using sensors such as Actimeter, Microphones, PIR (Passive Infrared), and Door contacts. Data fusion is performed using an SVM classifier. Medjahed and Boudy [12] have presented a smart-home setting which performs activity recognition relying on ambient sensors, such as infrared, change state sensors, audio, and physiological sensors fused by a Fuzzy Classifier.

A descriptive-based approach has been presented by Cao *et al.* [3] for event recognition. It models the context of a human (e.g., body posture) using data from a set of cameras, and of the environment (semantic information about the scene) using data of accelerometer devices attached to objects of daily living. The object sensors trigger events when manipulated (*e.g.*, TV remote control or doors use). A rule-based reasoning engine is used for processing and combining both model types at event detection level. Zouba *et al.* [5] have evaluated a video monitoring system at the identification of activities of daily living of older people on a model apartment equipped with home appliances. A set of environmental sensors (pressure, contact) is attached to home appliances, and their change of state is modeled using a description based approach. A video-camera is used to track the people over the environment and estimate their posture. Environmental sensors and video-camera data is combined using Dempster-Shafer theory.

Multi-sensor approaches for event recognition generally perform fusion at data or feature level by the use of State Models or Pattern recognition approaches. But, these approaches are whether too much complicate to be applied at real life context or sometimes too simple to cope with the challenges of real scenarios.

This paper extends the hierarchical model-based framework proposed by Vu *et al.* [18] to take into account multiple sensors at event recognition level. A generic ontology is used to describe the event models in terms of data coming from different sensors. This level is chosen due to the abstraction of sensor hardware and software implementation, which provides a flexible way to deal with sensor heterogeneity. A probabilistic approach is presented to handle event conflict among mutually exclusive events from different sensors.

We evaluate the proposed framework using multi-sensor recordings of real participants of a clinical protocol for Alzheimer disease study. Their activity dataset is chosen due to the growing applicability of monitoring systems for older people care, assisted living, and frailty diagnosis.

The paper is organized as follows: the Event recognition framework is described in section 2, the Evaluation procedure is described in section 3, the Results and Discussion are presented in section 4, followed by the Conclusion in section 5.

2. EVENT RECOGNITION FRAMEWORK

The framework is composed of two main components a hierarchical model-based framework for event modeling and a temporal event recognition algorithm [18]. The temporal algorithm takes as input the models developed by domain experts and evaluates whether their constraints are satisfied. This paper contribution extends the hierarchical model-based framework to take into account multiple sensor data, and to deal with mutually exclusive conflicting events of different sensors for people monitoring.

Figure 1 presents an example of architecture for the extended event recognition framework. It employs a wearable inertial sensor and two video-cameras as input sensors. These sensors are pre-processed and it is their processed output that is used as input for the Event recognition framework, represented as the Event Recognition Module. For instance, the output of the inertial sensor will consist of a set of postures of the person associated to a timestamp, while the output of video camera will consist of the set of Person detected in the scene and/or a set of primitive events associated to a timestamp.

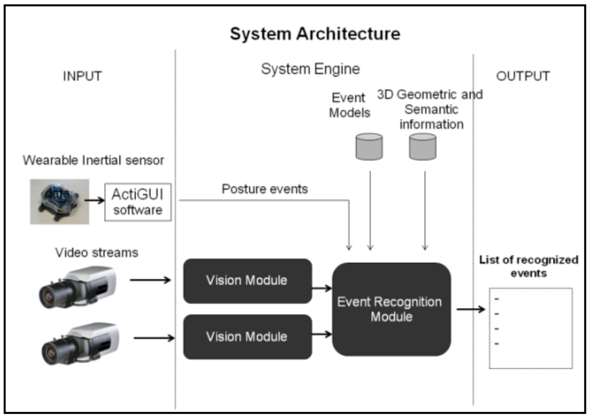

Figure 1. Overall Architecture of the Video Monitoring System

2.1 Hierarchical Model-Based Framework

The event models are described using a constraint-based ontology based on natural terminology to allow domain experts to easily add and change them.

An event model has six components [18]:

- Physical Objects refers to real objects involved in the recognition of the event modeled. Examples of physical object types are: mobile objects (e.g. person herein, or vehicle in another application), contextual objects (equipments) and contextual zones (chair zone);
- Components refer to sub-events that the model is composed of;
- Forbidden Components refer to events that should not occur in case of the event model is recognized;

- Constraints are conditions that the physical objects and/or the components should hold. These constraints could be logical, spatial and temporal;
- Alert describes the importance of a detection of the scenario model for a given specific treatment; and
- Action in association with the Alert type describes a specific action which will be performed when an event of the described model is detected (e.g. send a SMS to a caregiver responsible to check a patient over a possible falling down).

Two types of Physical Object are defined here: Person, Contextual Objects. The Person class is an extension of a generic class named mobile, which contains information of mobile objects (*e.g.*, 3D position, width, height). The Person class model has attributes like body posture, appearance, etc. Contextual Objects herein refer to *a priori* knowledge of the scene.

The *a priori* knowledge of the scene consists of a decomposition of a 3D projection of the scene floor plan into a set of spatial zones (e.g., TV zone, Armchair Zone), and relevant equipments (e.g., home appliances and furniture such as TV, armchair, Coffee machine) which hold semantic information relevant to the modeled events.

Constraints define conditions that physical object property (ies) and/or components should satisfy. They can be a-temporal, such as spatial and appearance constraints; or they could be temporal and specify two instances ordering which should generate a third event, for example, Person_crossing_from_Zone1toZone2 is defined as Person_in_zone1 before Person_in_zone2. Temporal constraints are expressed using Allen's interval algebra (*e.g.*, BEFORE, MEET, and AND) [2].

The ontology hierarchically categorizes models according to their complexity on (in ascending order):

- Primitive State models an instantaneous value of a property of a physical object (Person posture, or Person inside a semantic zone).
- Composite State refers to a composition of two or more primitive states.
- Primitive Event models a change in a value of physical object property (e.g., Person changes from Sitting to Standing posture).
- Composite Event refers to the composition of two previous event models which should hold a temporal relationship (Person changes from Sitting to standing posture before Person in Corridor Zone).

Figure 2 presents an example of a primitive state model for the recognition of sitting posture. This model checks whether the value of the attribute Posture is equal to the desired posture value (sitting).

```
PrimitiveState ( Person_sitting,
  PhysicalObjects( ( p1 : Person ) )
  Constraints ( ( p1->Posture = sitting) )
)
```

Figure 2. Primitive State of Person sitting

Figure 3 presents the Composite Event "Person sitting and using Office Desk". The model has two components and one constraint. The constraint establishes that the two components must be detected at the current time by using the AND operator of Allen's interval algebra. Briefly, the components define that for each physical object of type Person detected in the scene in the frame of evaluation, in which the current posture is sitting, and its 3D bounding box centroid projection into the ground is inside the *a priori* defined zone called "Office", an instance of the described event is created. If an instance of this event involving the same Person of the current instance exists in previous frames, these instances are merged if their time distance is lower than a given threshold.

```
CompositeEvent( Person_sitting_and_using_OfficeDesk,
PhysicalObjects( (p1:Person), (z1:Zone) )
Components(
    (c1:CompositeEvent  P_insideOfficeDeskZone(p1,z1))
    (c2:PrimitiveState  P_sitting (p1)))
Constraints( (c1 AND c2) )
)
```

Figure 3. Composite event "Person sitting and using OfficeDesk". The term Person is replaced within the model by the letter P to improve model visualization

2.2 Modeling Events from Different Sensors

The previous section has described how the hierarchical model-based framework categorizes and models events. Nevertheless, certain applications use multiple sensors to capture different aspects of a phenomenon, different phenomena, or even both to accomplish a given task.

To model events generated by different sensors we adopt Primitive State models. This event type is the basic building block of the event hierarchy and this choice allows the treatment of noise and false positive events early on the event hierarchy processing. It is also of particular usefulness at modeling events of heterogeneous sensors, since only the sensor output is considered at the event model, abstracting all the underlying process of acquisition and data processing. Consequently, hierarchically higher event model (like composite event) can be built without explicit knowledge of the primitive states (and the sensor that generated them) by relying in intermediate models (Primitive Event, Composite Event).

For instance, we present the modeling of a Person posture (e.g., Sitting, Standing) using events generated from a video-camera and a wearable inertial sensor. Figure 4 describes the Person model, where an attribute is added for the inertial and video sensor, respectively, Posture_WI and Posture_V.

```
class Person:Mobile
 {
       String PostureV;
       String PostureWI;
 }
```

Figure 4 Declaration of the Person Class

Figure 5 presents an example of declaration of Primitive state model which uses the attribute "posture" provided by the WI sensor.

```
PrimitiveState( Person_sitting_WI,
 PhysicalObjects(  (p1 : Person) )
  Constraints ( (p1->PostureWI = Sitting) )
)
```

Figure 5. Primitive state mapping a wearable sensor value

Figure 6 presents an example of Composite Event which combines (is composed of) the primitive states from the two sensors (WI and video camera), envisaging a situation where both sensors need to agree to have the final decision for a person sitting.

```
CompositeEvent( Person_Sitting_MS,
 PhysicalObjects(
                    (p1:Person), (z1:Zone), (eq1:Equipment))
 Components(
           (c1: PrimitiveState     Person_sitting_V (p1))
           (c2: PrimitiveState     Person_sitting_WI(p1)))
 Constraints( (c1 AND c2) )
)
```

Figure 6. Composite event "Person Sitting MS"; V: vision-system; WI: wearable inertial sensor, MS: multi-sensor

The described event model (Figure 6) has showed the combination of two sensors for the recognition of a posture event. This modeling is particularly useful when the developed system aims at a higher sensitivity (lower index of false positive events). Figure 7 presents an adapted version composite event of the model already presented in Figure 3, now using multi-sensor event model of Figure 6. This fact shows the flexibility provided by the adoption of a model-based approach for a multi-sensor context.

```
CompositeEvent(Person_Sitting_in_OfficeDeskZone,
 PhysicalObjects(
  (p1:Person), (z1:Zone), (eq1:Equipment))
 Components(
  (c1:CompositeEvent P_inside_OfficeDeskZone
                                         (p1, z1))
  (c2: CompositeEvent Person_sitting_MS(p1))
 )
 Constraints(  (c1 AND c2)  )
)
```

Figure 7. Composite Event Sitting in the zone Office Desk. Person term is replaced by P, to improve Figure visualization

Although we have presented the combination of two sensors related to the same attribute (or aspect) of an event model, domains experts are free to design the event models using one sensor per aspect. For instance, a model could have the Person posture described in terms of the posture provided by an inertial sensor, while the person position comes from a video camera data processing. Intermediate models can be added as needed to abstract the sources of information and create higher level representations of a person activity.

The described approach is most suited at combining distinct information obtained from different sensors, or in the case of dealing with same information a smaller set of sensors is recommended where it is feasible to define the full set of associations among the output of these sensors via the model constraints. Nevertheless, there are cases where it is necessary to take into account the output of several sensors about the same information. For such cases it would be impractical to define an attribute and the associated constraints for each case as the number of sensors raises. Moreover in real life scenarios it is common to have conflicting evidence among the set of sensor which can be occasioned by noise in underlying steps of data processing. To cope with both cases, we propose in the next section a probabilistic approach.

2.3 Event Conflict Handling

For cases where conflicting evidence arises amongst events detected by different sensors, we propose a probabilistic framework to assess event reliability, and based on it decide which of the events should be recognized. We herein propose the following framework for Event Conflict handling: firstly, the event instantaneous likelihood is computed;_secondly, the event instantaneous likelihood is combined with its previous values to generate a new probability, the event temporal reliability (see [14]); finally, Dempster-Shafer theory is applied to decide upon the event temporal reliability values of conflicting events from different which event is being performed by the person.

The event conflict handling framework is also applied at primitive state level, therefore allowing higher level models to be derived from them, and reducing the noise propagation to hierarchically higher models, a fact that tends to reduce the performance of model-based approaches.

2.3.1 Instantaneous likelihood of a Primitive State

The instantaneous likelihood is computed based on the feature used to generate the primitive state. For illustrative purposes, we will describe the posture events of sitting/standing. The person posture is going to be recognized based on a height threshold. If the height is below the threshold, the person is considered Sitting, otherwise Standing. For this case, we would consider that the posture information will be provided by pre-processing two cameras data, in which case, due to failures in underlying vision algorithms, the person height is affected, and consequently the posture identification.

We assume the features used to detect the Primitive states (*e.g.*, height) follow a Gaussian distribution. Therefore, a learning step is performed *a priori* to learn the distribution parameters mean (μ) and variance (σ2) of the height feature for Standing and Sitting primitive states for each sensor. Based on the obtained distribution parameters, the instantaneous likelihood of a given event for the current instant and a given sensor *i* is computed using Equation 1.

$$P^{inst}_{\Omega,k,i} = e^{-\frac{Height_{\Omega,k,i}-\mu_{\Omega,i}{}^2}{2\sigma^2_{\Omega,i}}} \qquad (1)$$

where,

k: video frame number (current instant), Ω: event model, i: sensor id

2.3.2 Temporal reliability of a Primitive State

The instantaneous likelihood of the Primitive State considers the probability of a given primitive state (e.g., sitting, standing) been recognized at the current frame. But, noise from underlying vision algorithms can compromised the feature value which a primitive state is based on for a short interval of time, (e.g., problems at image segmentation can harm the height estimation of a person). To cope with instantaneous value deviations we compute the event temporal reliability which takes into account current frame plus previous instants instantaneous likelihood values for a given time interval. Equation 2 and 3 present an adapted computation of temporal reliability using a time window of fixed size [14]. A cooling function is used to reinforce the information of near instants and lesser the one from farther ones.

$$P^{temp}_{\Omega,k,i} = \frac{P^{inst}_{\Omega,k,i} + M}{\sum_{t=k-w}^{t=k-1} e^{-(k-t)}} \quad (2)$$

$$M = \sum_{t=k-w}^{t=k-1} [e^{-(k-t)}(P^{temp}_{\Omega,k,i,} - P^{inst}_{\Omega,k,i,})] \quad (3$$

where,

k: video frame number (current instant), Ω: event model
i: sensor id, w: temporal window size

Concerning the window size parameter of these equations, and as primitive states are generally a continuous process which lasts for seconds or even minutes, the window size parameter should fit at least the minimum expected time interval for the modeled primitive state.

Gaussian distribution likelihood can be considered as a belief level value, and as we have assumed the feature values of Primitive states follow such distribution, the Primitive State Temporal Reliability is employed as "how strongly it is believed that the event generated by the sensor *i* is true at the evaluated time instant".

After computing the Primitive State's reliability (Event Temporal reliability), it is necessary to analyze these probabilities to decide which event is being performed.

2.3.3 Primitive State Conflict Handling

To decide upon the Event probabilities we have chosen Dempster-Shafer Theory (DS). DS theory was proposed by Dempster [5] and improved by Shafer [16]. It extends the Bayesian inference's application by allowing uncertainty reasoning based on incomplete information. The major components of evidence theory are the frame of discernment (Θ), and the basic probability assignment (BPA). The frame of discernment contains all possible mutually exclusive hypotheses.

$$\Theta = \{\text{Sitting}, \text{Standing}, \cdots\}$$

The BPA is a function m: $2^{\Theta} \rightarrow [0,1]$ related to a proposition satisfying conditions (X) and (Y) [1]:

$$m(\emptyset) = 0 \quad (X)$$

$$\sum_{A \in \Theta} m(A) = 1 \quad (Y)$$

where, *A* is any subset of the **frame of discernment**, and $\emptyset$ refers to the empty set.

For any $A \in 2^{\Theta}$, m(A) is considered as the subjective confidence level on the event A. Accordingly, the whole body of evidence of one sensor is the set of all the BPAs greater than 0 under one frame of discernment. The combination of multiple evidences defined on the same frame of discernment is the combination of the confidence level values based on BPAs (*e.g.*, pre-defined by experts). Given two sensors (1 and 2), where each sensor has its body of evidence (m_{s1} and m_{s2}), these are the corresponding BPA functions of the frame of discernment.

The combination rule of the classical DS theory can be implemented to fuse data from two sensors, but it can lead to illogical results in the presence of highly conflicting evidence [1]. We herein adapt the combination rule proposed by Ali et al. [1], as it has been demonstrated to provide more realistic results than the standard DS rule when combining conflicting evidence from multiple sources.

Equations 4 and 5 present the mass function for computing Sitting (Sit.) and Standing (Sat.) primitive states, respectively:

$$(m_{s1} \oplus m_{s2})(\text{Sit.}) = \frac{1 - (1 - m_{s1}(Sit.)) \times (1 - m_{s2}(Sit.))}{1 + (1 - m_{s1}(Sit.)) \times (1 - m_{s2}(Sit.))} \quad (4)$$

$$(m_{s1} \oplus m_{s2})(\text{Sat.}) = \frac{1 - (1 - m_{s1}(Sat.)) \times (1 - m_{s2}(Sat.))}{1 + (1 - m_{s1}(Sat.)) \times (1 - m_{s2}(Sat.))} \quad (5)$$

The combination rule can be iteratively used to combine more than two body of evidence.

3. EVALUATION

To evaluate the proposed framework we have used multi-sensor recordings of real participants of a clinical protocol for Alzheimer disease study. This dataset is chosen due to the growing applicability of monitoring systems for older people care, assisted living, and frailty diagnosis. Inertial sensor raw data is pre-processed using its (proprietary) software to generate the list of Person postures during the experimentation. Video streams are processed using a monitoring system. All the sensor recordings are time synchronized, and none spatial correspondence is performed among the cameras.

3.1 Performance Evaluation

The event recognition performance is evaluated in two scenarios: first, we compare a mono and multi-sensor approach using data from an RGB camera placed on one of the top corners of the observation room and a wearable inertial sensor. Event models only takes into account inertial sensor data for posture identification. Secondly, we evaluate the proposed probabilistic approach for conflict handling using events generated by two video cameras (RGB and RGB-D devices described in section 3.3.).

Event Recognition performance is evaluated using indices of sensitivity, precision, and F-score describe in Equations 6, 7, and 8, respectively.

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

where, TP: True Positive rate, FP: False Positive rate, FN: False Negative rate.

$$F - Score = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} \quad (8)$$

3.2 Monitoring System

The Monitoring System component herein used to test the proposed framework is a evaluation platform locally developed that allows the test of different algorithms for each step of the computer vision chain (*e.g.*, video acquisition, image segmentation, physical objects detection, physical objects tracking, actor identification, and actor events detection). The vision component extracts the objects to track from the current frame using an extension of the Gaussian Mixture Model algorithm for background subtraction proposed by [13]. People tracking is performed by an implementation of the multi-feature tracking algorithm proposed in [4], using the following features: 2D size, 3D displacement, color histogram, and dominant color. The vision component is responsible for detecting and tracking mobile objects on the scene. These objects (so-called physical objects) are classified according to a set of *a priori* defined classes, e.g., a person, a vehicle. The detected physical objects are then passed to the event recognition module which assess whether the actions/activities of these actors match the event models defined by the domain experts.

3.3 Dataset

Participants aged more than 65 years are recruited by the Memory Center (MC) of a collaborating Hospital. Inclusion criteria of the Alzheimer Disease (AD) group are: diagnosis of AD according to NINCDS-ADRDA criteria and a Mini-Mental State Exam (MMSE) [7] score above 15. AD participants which have significant motor disturbances (per the Unified Parkinson's Disease Rating Scale) are excluded. Control participants are healthy in the sense of behavioral and cognitive disturbances. The clinical protocol asks the participants to undertake a set of physical tasks and Instrumental Activities of Daily Living (IADL) in a Hospital observation room furnished with home appliances. Experimental recordings use a RGB video camera (AXIS®, Model P1346, 8 frames per second), a RGB-D camera (Kinect® sensor), and a wearable inertial sensor (MotionPod®).

The set of monitored IADLs is composed as follows:

1. Watch TV,
2. Make tea/coffee,
3. Write the shopping list of the lunch ingredients,
4. Write a check to pay the electricity bill,
5. Answer/Call someone on the Phone,
6. Read newspaper/magazine,
7. Water the plant
8. Organize the prescribed drugs inside the drug box according to the weekly intake schedule.

Figure 8 shows the recording viewpoint of the RGB and RGB-D cameras in A and B, where WI sensor is visible at image B.

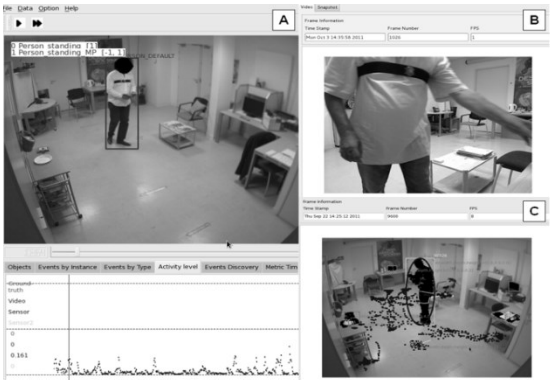

Figure 8. Participant' activity by the view point of different sensors: (A) RGB camera view and actimetry provided the inertial sensor (the bottom of image A); (B) RGB-D camera view of participant, which shows the inertial sensor worn by the participant; and (C) Drawn points on the ground represent the trajectory information of the participant during the experimentation.

3.4 Event Modeling

Each one of the eight IADL is modeled using two composite models and three primitive states. First composite model is composed of two of the primitive states: one for the recognition of the person position inside a contextual zone (*a priori* defined), and another for his/her proximity to a static object (equipment) located into the respective zone (also *a priori* defined, e.g., Phone station, Coffee machine). Second composite model is composed of the first composite model to include the recognition a given IADL, and a primitive state model related to the posture of the person. The posture primitive state uses the posture data obtained only from the inertial sensor. Temporal constraints are defined accordingly to each IADL. The activities "writing a check" and "writing a shopping list" are not differentiated, and are referred as "Person using Office Desk" due to the absence of object manipulation data from the monitoring system. The activity "Organize the prescribed drugs…" is shortened as Person using pharmacy basket.

4. RESULTS AND DISCUSSION

Table 1 presents the performance of the framework while recognizing the IADL a person is performing and his/her posture. Results are presented for a mono- and a multi-sensor approach (RGB camera and Inertial Sensor). Average performance is presented for the cases with and without posture recognition. The average value "IADL without Posture IADL" refers to the reference accuracy of event recognition framework without posture recognition, and it only takes as input the video-camera information; therefore no difference is expected between Mono- and Multi-sensor approaches.

Table 1. Comparison of Mono and Multi-sensor approaches

F – SCORE	Mono-	Multi-sensor
IADLs + Sitting posture	52.00 %	71.00 %
IADLs + Standing posture	73.15 %	71.00 %
Average of IADL with Posture	68.00 %	71.00 %
Average of IADL without Posture	81.22 %	81.22 %

N: 9; 15 min. each; total of 64800 frames (135 min).

Table 1 showed the average performance of event recognition decreases (see "Only IADLs" x "IADL + Posture") as the IADL models now take into account also the posture estimation. The Deterministic modeling of Multi-sensor events improve by ~19% the precision index value of Sitting. Recognition rate of model concerning Standing posture is slightly decreased, showing the inertial sensor could have a lower performance for this posture. The decrease in performance is explained by the fact the models have become more specific, and the lower performance of the posture recognition algorithms.

Table 2 presents the results of the proposed framework for conflict handling on the recognition of the Person posture using events from two different video-cameras (RGB and RGB-D). Individual performance of the description based approach using each camera is also presented for comparative purposes.

Table 2. Postures Recognition in Physical Tasks

Posture	Sitting	Standing

Sensor	Precision	Sensitivity	Precision	Sensitivity
RGB	84.29 %	69.41 %	79.82 %	91.58 %
RGB-D	100.00 %	36.47 %	86.92 %	97.89 %
Fusion	82.35 %	91.30%	91.04 %	95.31 %

N=10. A 5 second window is used for Temporal Probability

The results in Table 2 showed the proposed framework for event conflict handling improves the detection of the posture-related primitive states for both postures. The precision at standing recognition is higher than the one achieved individually by each video camera, suggesting the framework is able to assess (event) information gain and properly combine it.

5. CONCLUSIONS

We highlight as contributions of this paper a hierarchical model based framework for multi-sensor combination, and a probabilistic framework for event conflict handling and fusion.

The hierarchical model-based framework applied to event recognition using multiple sensors improves by ~ 19 % the F-Score of the recognition of a person sitting while performing IADLs with respect to the recognition using only a single camera. But, no improvement is obtained for models considering standing posture. The probabilistic approach for event conflict handling approach based on evidence theory and the alternative combination rule proved to be successful at handling multi-sensor event conflict for the tested situation. In two of four indices, it achieved a recognition rate higher than the one individually obtained by the two cameras, while for the others it had similar performance. The presented results indicate that information fusion cannot provide improvement for every case, being necessary to assess the gain provided by each sensor for a given task.

The presented framework provides a hybrid and intermediate approach between the modeling of semantic and hierarchical representation provided by description-based models and the need of a training of State Models and Classification (or Pattern Recognition) methods. Besides to the combination of different sensor at event level, the presented probabilistic approach also provides the basis to cope with errors of underlying process to event recognition, which are one of the major limitations of deterministic approaches such as description-based models.

Future work will extend the evaluation of the multi-sensor hierarchical model framework for a larger variety of primitive states and sensors (heterogeneous and homogeneous) with respect to their reliability at conflict handling and event recognition.

6. REFERENCES

[1] Ali T., Dutta, P. and Boruah H. 2012. A new combination rule for conflict problem of Dempster-Shafer evidence theory. *International Journal of Energy, Information and Communications*, 3, 1 (Feb. 2012).

[2] Allen J.F. 1983. Maintaining Knowledge about temporal intervals. *Communications of the ACM*, 26,11 (Nov. 1983), 832-843.

[3] Cao, Y., Tao, L., and Xu, G. 2009. An event-driven context model in elderly health monitoring. In *Proceedings of. Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing* (2009), 120-124.

[4] Chau, D. P., Bremond, F., and Thonnat, M. 2011. A multi-feature tracking algorithm enabling adaptation to context variations. In *Proceedings of International Conference on Imaging for Crime Detection and Prevention* (2011).

[5] Dempster, A. P., 1968. Generalization of Bayesian inference. *J.Royal Statist. Soc*., 30 (1968), 205–247.

[6] Fleury, A., Noury, N., Vacher, M. 2010. Introducing knowledge in the process of supervised classification of activities of Daily Living in Health Smart Homes. In *Proceedings of 12th IEEE International Conference on e-Health Networking Applications and Services* (Jul. 2010), 322-329.

[7] Folstein, M.F., Robins, L.M., and Helzer, J.E. 1983. The mini-mental state examination. *Arch Gen. Psychiatry*, 40, (1983), 812, 1983.

[8] Gao, L., Bourke, A.K, Nelson, J. 2011. A system for activity recognition using multi-sensor fusion. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Aug. 2011), 7869-7872.

[9] Izadinia, H., and Shah, M., 2012. Recognizing complex events using large margin joint low-level event model. In *Proceedings of the 12th European conference on Computer Vision*, 4, (Firenze, Italy, October 2012), 430-444.

[10] Lavee, G., Rivlin, E., Rudzsky, M., 2009. Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39, 5 (Sep. 2009), 489-504.

[11] Le, Q.V., Zou, W.Y, Yeung, S.Y., Ng, A.Y. 2011. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Jun. 2011) 3361-3368.

[12] Medjahed, H., Istrate, D., Boudy, J., Baldinger, J.-L., Dorizzi, B. 2011. A pervasive multi-sensor data fusion for smart home healthcare monitoring. In *Proceedings of IEEE International Conference on Fuzzy Systems*, (Jun. 2011), 1466-1473.

[13] Nghiem, A. T., Bremond, F., and Thonnat, M. 2009. Controlling background subtraction algorithms for robust object detection. In *Proceedings of 3rd International Conference on Imaging for Crime Detection and Prevention*, (London, UK, December, 2009), 1-6.

[14] Romdhane, R., Bremond, F., and Thonnat, M. 2010. Complex Event Recognition with Uncertainty Handling. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, (Boston, USA, Aug., 2010).

[15] Rong, L., and Ming, L. 2010. Recognizing Human Activities Based on Multi-Sensors Fusion. In *Proceedings of 4th International Conference on Bioinformatics and Biomedical Engineering*, (Jun. 2010), 1-4.

[16] Shafer, G., 1976. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[17] Wang, H., Klaser, A., Schmid, C., Liu, C. 2011. Action recognition by dense trajectories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Jun., 2011), 3169-3176.

[18] Vu, T., Bremond, F., and Thonnat, M. 2003. Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence* (Acapulco, Mexico, Aug. 9-15, 2003).

[19] Zaidenberg, S., Boulay, B., Bremond, F., and Thonnat, M. 2012, A generic framework for video understanding applied to group behavior recognition. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-based Surveillance*, (Sep. 2012), 136-142.

[20] Zouba, N., Bremond, F., and Thonnat, M. (2010). An Activity Monitoring System for Real Elderly at Home: Validation Study. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, (Boston, USA, August 29, 2010).