Combining Multiple Sensors for Event Detection of Older People

Carlos Crispim-Junior, Qiao Ma, Baptiste Fosty, Rim Romdhane, François Bremond, Monique Thonnat

Combining Multiple Sensors for Event Detection of Older People

Carlos F. Crispim-Junior, Qiao Ma, Baptiste Fosty, Rim Romdhane, Francois Bremond, Monique Thonnat

Abstract We herein present a hierarchical model-based framework for event detection using multiple sensors. Event models combine *a priori* knowledge of the scene (3D geometric and semantic information, such as contextual zones and equipment) with moving objects (*e.g.*, a Person) detected by a video monitoring system. The event models follow a generic ontology based on natural language, which allows domain experts to easily adapt them. The framework novelty lies on combining multiple sensors at decision (event) level, and handling their conflict using a probabilistic approach. The event conflict handling consists of computing the reliability of each sensor before their fusion using an alternative combination rule for Dempster-Shafer Theory. The framework evaluation is performed on multisensor recording of instrumental activities of daily living (*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 study. Two fusion cases are presented: the combination of events (or activities) from heterogeneous sensors (RGB ambient camera and a wearable inertial sensor) following a deterministic fashion, and the combination of conflicting events from video cameras with partially overlapped field of view (a RGB- and a RGB-D-camera, Kinect). Results showed the framework improves the event detection rate in both cases.

1 INTRODUCTION

Human Behavior (or event) monitoring has experienced continuous advances since last decade promoted by Computer Vision, Wearable and Ubiquitous Computing Fields. Examples of applications range from security field, such as video surveillance, crime prevention, and older people monitoring at home, to tools to support objective assessment of emerging symptoms of diseases (medical diagnosis), and even as a part of human-machine interfaces for game entertainment.

Wearable and Pervasive Computing communities have proposed multimodal event monitoring based on sensors such as, wearable inertial sensors, passive infrared presence sensors, change of state sensors, microphones. For instance, Gao *et al.* [8] and Rong and Ming [15] have demonstrated the fusion of wearable inertial sensors at the waist, chest, and sides of a person body for the detection of daily living activities, where data fusion was carried out by classification methods (*e.g.,* Naïve Bayes, C4.5). Although wearable inertial sensors provide a rich representation of body dynamics, they are subjected to problems such as motion noise, inter sensor-calibration, and in case of large scale research studies, the need of placing sensors in a relatively similar body position among monitored people, what may introduce noise in experimental data. Fleury *et al.* [6] have presented a multi-modal event monitoring system using actimeters, microphones, PIR (Passive Infrared) presence sensor, and door contact sensors. Data fusion is performed using a SVM classifier. Medjahed and Boudy [12] have proposed a smart-home setting which performs event detection relying only on ambient sensors like infrared, change of state sensors, and microphones, and physiological sensors; all fused by a Fuzzy classifier.

Computer Vision approaches for event detection may be summarized in three categories (adapted from [10]): classification methods, probabilistic graphical models (PGM), and semantic models, which rely on at least of the following data abstraction: pixel-level, feature-level, or event-level. Probabilistic Graphical Models refer to techniques such as Conditional Random Fields, Dynamic Bayesian Networks, and Hidden Markov Models. Kitani *et al.* [27] has proposed a Hidden Variable Markov Model approach for event forecasting based on people trajectories and scene features. Examples of classification methods are Artificial Neural Networks, Support-Vector Machines (SVM), and Nearest Neighbor. In this context, Le *et al.* [11] have presented an extension of the Independent Subspace Analysis algorithm applied for learning invariant spatio-temporal features from unlabeled video data for event detection. Wang *et al.* [17] have proposed new descriptors for dense trajectory estimation in action representation as input for non-linear support vector machines. Although PGMs and classification methods have considerably increased the event detection performance in benchmark data sets, as they focus on pixel-and feature-based representations, they have limitations at describing the scene semantics and the temporal dynamics and hierarchical structure of complex events. Moreover, these approaches only focus on video data, ignoring other modalities which could provide additional information in the presence of ambiguous data.

In the recent domain of video search in internet videos, multimodal event analysis have investigated event representations consisting of different image cues, like motion and appearance, combined with other modalities such as audio and text, and exploring fusion in different data abstraction levels. Jhuo *et al.* [22] introduced a feature-level representation which combines audio and video data by mapping the joint patterns among these two modalities. Myers *et al.* [24] have learned a set of base classifiers, each from a single data type/source (low-level vision, motion, audio, high-level visual concepts, or automatic speech recognition), and evaluated their fusion using different methods at event level (late fusion scheme). They report average output was one of the most effective fusion schemes. Similarly, Oh *et al.,*[25]

have presented a multimodal (audio and video) system, where base classifiers are learned from different subsets of features, and score fusion are used to combine them into complex events. Mid-level features, such as object detectors, were employed to enrich event model semantics. Even though multimedia event analysis approaches have demonstrated significant advances by seeking to capture the hierarchical nature of events and incorporating auxiliary sources of information, most methods rely on learning steps involving large amounts of training data.

Semantic (or Description-based) models make use of a description language and logical operators to build event representations incorporating knowledge of domain expert. These languages allows to explicitly model the semantic information and hierarchical structure of event, besides to not require as much data as PGMs and classification methods. For instance, Zaidenberg *et al.* [19] have presented a generic model-based framework for group behavior detection on surveillance applications such as airport, subway, and shopping center.

Cao *et al.* [3] proposed a multimodal event detection where two context model are defined: the human and the environment contexts. The human context (*e.g.,* body posture) is obtained from data of a set of cameras, while the environment context (semantic information about the scene) is based on accelerometer devices attached to objects of daily living which once manipulated trigger an event, (*e.g.,* TV remote control or doors use). A rule-based reasoning engine is used for combining both context types at event detection level. Although semantic models ability to easily incorporate scene semantics, they are sensitivity to noise of underlying process, like image segmentation and people tracking in vision systems. To overcome such limitations, probabilistic frameworks may be adopted to handle data uncertainty as in [20] and [12]. For example, Zouba *et al.* [20] have evaluated a multimodal monitoring system at the identification of activities of daily living of older people on a model apartment. Video-camera data was used to track people over the scene and environmental sensor to obtain complementary data on object interaction. Dempster-Shafer theory was employed for reasoning under imprecise data.

This paper presents hierarchical model-based framework to multiple sensor context. We extend the generic ontology proposed by Vu *et al.*[18] to describe event models in terms of elementary (low-level) events coming from different sensors, as a basis to infer Multimodal Complex Events. Event level fusion is chosen as it provides a flexible way to deal with sensor heterogeneity, and has been reported to presented a higher performance than early fusion schemes based on pixel- and feature-level representations [23] [24]. A Dempster-Shafer-based probabilistic approach is presented to handle event conflict using an adapted combination rule. The framework is evaluated on real multisensor recordings of participants of a clinical protocol for Alzheimer disease study.

2 HIERARCHICAL MODEL-BASED FRAMEWORK

The proposed framework is composed of two main components: an event ontology and a temporal event detection algorithm [18]. The temporal algorithm is responsible for event inference based on the event models defined by domain expert and available input data. The video event ontology proposed in [18] is extended for multiple sensor scenario (then referred as Event Ontology), and the temporal algorithm to deal with mutually exclusive conflicting events of different sensors during people monitoring.

Fig.1 presents the architecture of the extended event detection framework, where a wearable inertial sensor and two video-cameras are given as examples of sensors. Sensor data is individually processed and their resulting output is taken as input for the multisensor framework (Event Detection Module). For instance, inertial sensor data would consist of a set of attribute-based events (*e.g.,* from posture: person bending, person lying down), while for video camera data it would be a set of people detected in the scene and/or elementary events provided provided by vision module. All sensors are assumed to be time-synchronized.

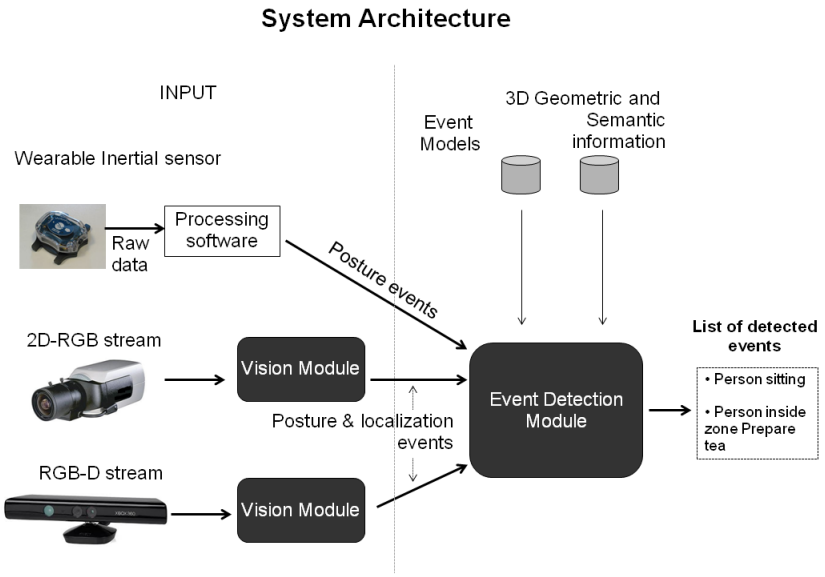

Fig. 1 Overall Architecture of the Video Monitoring System (adapted from [28])

2.1 Event Ontology or MultiModal

The event models are described using a constraint-based ontology language based on natural terminology to allow domain experts to easily add and change them. An event model is composed of up to six parts [18]:

- **Physical Objects** refer to real objects involved in the recognition of the modeled event. Examples of physical object types are: mobile objects (*e.g.* person herein, or vehicle in another application), contextual objects (equipment) 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).

Three types of Physical Object are defined: Mobile, Person, and Contextual Objects. Mobile class defines a set of attributes which be common to any mobile object (*e.g.,* height, width, position, speed). Person class extends Mobile by adding person-related attributes like body posture, appearance, *etc*. Contextual Objects refer to *a priori* knowledge of the scene. *A priori* knowledge refers to a decomposition of a 3D projection of the scene floor plan into a set of spatial zones (*e.g.*, TV zone, Armchair Zone), and equipment, (*e.g.*, home appliances and furniture such as TV, armchair, Coffee machine) which hold semantic information to the modeled events. Constraints define conditions that physical object property(-ies) and/or components should satisfy. They can be non-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).

Fig. 2 presents a description of Primitive State called Person sitting, which checks whether the attribute *posture* of a *person* object assumes the value *sitting*. Fig. 3 presents an example of Composite Event, called Person sitting and using Office Desk, which defines a constraint between two sub-events (components). First component checks whether the person position lies inside of *a priori* defined zone relative to an office desk, while second component verifies whether the person posture is sitting (using Fig. 2 model). Model constraint defines the model will be valid when both components should are recognized at the same time (c1 AND c2).

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

Fig. 2 Person_sitting

```
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->Interval AND c2->Interval) )
)
```

Fig. 3 Person_sitting_and_using_OfficeDesk

2.2 Modeling Events from Different Sensors

Previous section has described how the event ontology categorizes and models events. We have chosen to model events generated by different sensor data using Primitive States, since they are the most basic building block of the event ontology. Handling sensor input at an early stage in hierarchy level avoids the propagation of noise to high-level events, and also abstracts the derived models from the sensor data they are conditioned on.

Fig. 4 describes the class *Person* where an attribute is created for each posture estimation, *e.g., Posture_WI* for the estimation from wearable inertial sensor, and *Posture_V* for the estimation of the video-based algorithm.

Fig. 5 illustrates an example of Primitive state using the posture estimation from a inertial sensor. If one aims to increase system precision over recall, a Composite Event may be devised to combine (be composed of) both posture estimation

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

Fig. 4 Class Person

(primitive states) and to restrict the targeted posture detection to when all sensor estimations agree, see Fig. 6 for an example.

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

Fig. 5 Person_sitting_WI

```
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->Interval AND c2->Interval) )
)
```

Fig. 6 Person_Sitting_MS

Fig. 7 presents the event model "Person sitting and using Office Desk" which relies on a multisensor event for the detection of posture sitting. Using an ontology language for event modeling on multisensor scenarios allows to decompose event complexity and provides a flexible way to add or change sensor-based events.

The presented model examples described how to combine estimations of multiple sensor over the same attribute of Person class. But, there is not restriction on how event models from different sensors are combined. A complex event model may have a person posture estimated from an inertial sensor (*Posture_WI*) while his/her localization is provided by a vision system.

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

Fig. 7 Person_sitting_and_using_OfficeDesk

2.3 Event Conflict Handling

To address conflicting evidence among (mutually exclusive) events generated by different sensors, a probabilistic framework is proposed to assess event reliability for event fusion. The conflict handling framework works as follows: firstly, event instantaneous likelihood is computed; secondly, event temporal reliability is computed from the current and close past event instantaneous likelihood (see [14]); finally, a variant of Dempster-Shafer rule of combination is used to decide upon event reliability which of the events is being performed.

The event conflict handling framework is performed at primitive state level to reduce the propagation of noise from low-level components to hierarchically higher event models, abstract high-level events from the sensor estimated events, and derive semantically high-level event only from consolidated information.

2.3.1 Instantaneous likelihood of a Primitive State

The instantaneous likelihood of Primitive States is computed based on the feature(s) the event constraints are based on. Assuming the Primitive state feature (*e.g.,* height) follows a Gaussian distribution, a learning step is performed *a priori* to obtain the expected feature distribution parameters (mean, μ, and variance, σ^2) given a primitive state and a sensor. The learning procedure is performed for each mutually exclusive event model affected by the analyzed feature.

Learned distribution parameters are then used during event inference (detection) to compute the instantaneous likelihood of an event given the feature value and the sensor providing it using Equation 1.

$$P^{inst}_{\Omega,k,i} = \frac{\exp(-(Height_{\Omega,k,i} - \mu^2_{\Omega,i}))}{2\sigma^2_{\Omega,i}} \tag{1}$$

where,

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

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 compromise 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 deviations of primitive state probabilities we compute the event temporal reliability which considers the instantaneous likelihood of an event and its previous values for a given time interval (time window). Equations 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 frames and lessen the influence of farther ones. The window size parameter used in these equations was set to match the minimum expected duration of the modeled primitive states.

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

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

where,

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

Primitive State Temporal Reliability is then considered as a belief level value on "how strongly it is believed that the event generated by the sensor i is true at the evaluated time instant". From here on Primitive State Temporal Reliability will be referred as Primitive State Reliability.

2.3.3 Primitive State Conflict Handling

Once the reliability of all mutually exclusive Primitive States are computed it is then necessary to decide which events are being actually performed. To perform such task we have adopted Dempster-Shafer Theory (DS). DS theory was proposed by Dempster [5] and then improved by Shafer [16]. It extends the Bayesian inference by allowing uncertainty reasoning based on incomplete information. The major components of evidence theory are the frame of discernment (Θ, Equation 4), and the basic probability assignment (BPA). The frame of discernment contains all possible mutually exclusive hypotheses.

$$\Theta = \{Sitting, Standing, ...\} \tag{4}$$

The BPA is a function m: $2^{\Theta} \rightarrow [0,1]$ related to a proposition satisfying conditions (5) and (6) [1]:

$$m(\emptyset) = 0 \tag{5}$$

$$\sum_{A \in \Theta} m(A) = 1 \tag{6}$$

where, A is any subset of the frame of discernment, and 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 (zero) 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 7 and 8 present the mass function for computing Sitting (Sit.) and Standing (Sta.) primitive states, respectively:

$$(m_s1 \bigotimes m_s2)(Sit.) = \frac{(1-(1-m_s1(Sit.))(1-m_s2(Sit.)))}{(1+(1-m_s1(Sit.))(1-m_s2(Sit.)))} \tag{7}$$

$$(m_s1 \bigotimes m_s2)(Sta.) = \frac{(1-(1-m_s1(Sta.))(1-m_s2(Sta.)))}{(1+(1-m_s1(Sta.))(1-m_s2(Sta.)))} \tag{8}$$

Among a set of mutually exclusive events the framework chooses the event with the highest probability (mass function). The combination rule can be used on an iterative fashion to combine more than two body of evidence.

3 EVALUATION

To evaluate the proposed framework we have used multisensor recordings from real participants of a clinical protocol for Alzheimer disease study. This data set is chosen due to the growing applicability of monitoring systems for older people care, assisted living, and frailty diagnosis.

The event detection performance is evaluated in two scenarios: firstly, we compare the crisp multisensor approach using data from an 2D-RGB camera and a wearable inertial sensor to a mono-sensor (camera) approach. Inertial sensor raw data is pre-processed using its (proprietary) software to generate the list of postures performed by the participant during the experimentation. Multi-sensor event models use wearable inertial sensor data for posture-based events and video-based data for person localization in the scene. Second scenario evaluates the proposed probabilistic approach for event conflict handling on events generated by two vision modules

(the 2D-RGB camera vision system and a variant of it using a RGBD sensor). For this scenario, posture data is obtained per vision module and then propagated for fusion in the form of events.

All sensors are assumed to be time synchronized, but none spatial correspondence is computed among the cameras in the second scenario. Briefly, we assume the multi-sensor system does not know the transformation function amongst the coordinate-systems of the cameras.

3.1 Performance Evaluation

Event detection performance is measured using the indexes of sensitivity, precision, and F-Score described in Equations 9, 10, and 11, respectively. System event detection is compared to event annotation performed by domain experts.

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

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

$$F-Score = \frac{2*(Sensitivity*Precision)}{Sensitivity+Precision} \tag{11}$$

3.2 Vision Module

The Vision Module used to test the proposed framework is a evaluation platform locally developed that allows the testing 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). Image segmentation is performed by 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 detection module which assess whether the actions/activities of these actors match the event models defined by the domain experts.

3.3 Data set

Participants aged more than 65 years are recruited by the Memory Center (MC) of Nice 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. Prepare 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.

Fig. 8 shows the recording viewpoint of the 2D-RGB and RGB-D cameras in A and B, and WI sensor at image B.

3.4 Event Modeling

Each one of the eight focused 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 table, 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 primitive states for posture recognition used data from the inertial sensor only. The activities "writing a check" and "writing a shopping list" are not differentiated and are referred instead as "Person using Office Desk" due to the lack of information about the object been manipulated by the patient. The name of the activity "Organize the prescribed drugs" is shortened as "Person using pharmacy basket".

Fig. 8 Participant' activities by the point of view 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.

4 RESULTS AND DISCUSSION

Table 1 presents the performance of the framework at recognizing the IADLs a person is undertaking and his/her posture. Results are presented for mono- and multisensor approaches (2D-RGB camera and wearable inertial sensor). Average performance is presented for IADLs with and without posture sub-events. The row "Average of IADL without Posture" refers to event models based only on the person localization in the scene provided by the video-camera, therefore no difference is expected between Mono- and multisensor approach in this case.

Table 1 Comparison of Mono and multisensor approaches

F-SCORE	Mono-	multisensor
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: 64800 frames (135 min).

The deterministic (or crisp) modeling of multisensor events has improved by $\sim$ 19 % the precision index of IADLs involving sitting posture by the replacement of the vision system by an inertial sensor for posture estimation. However, the multisensor event models had a slightly lower performance on the detection of IADL involving standing posture than the mono-sensor approach. These results point that none of the two employed sensors can completely replace the other, and limiting the detection performance to the quality of the individual sensors output.

The difference in performance between IADL detection with and without posture component shows that by reducing the number of model constraints a higher detection performance can be achieved at the expenses of less information about how the event was performed. To tackle this problem, the probabilistic approach should be used to combine both posture estimations and also make models more robust to noise.

Table 2 presents the results of the proposed framework for conflict handling on the recognition of the Person posture using events from two different videocameras (2D-RGB and RGB-D). The individual performance of the hierarchical model-based framework per camera is 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 window of 5 second is used for Temporal Probability.

The results in Table 2 showed the conflict handling framework improves the detection of posture-related primitive states on both posture categories. The precision achieved at standing recognition is higher than the one achieved by each video camera individually, demonstrating the suitability of the conflict handling framework for the assessing of event reliability and the combination of multiple sensor events for a more accurate detection.

5 CONCLUSIONS

We highlight as contributions of this paper a hierarchical model based framework for multisensor combination and a probabilistic approach for event conflict handling and fusion. The hierarchical model-based framework following a crisp combination of events from different sensors improves the detection of people seated while undertaking IADLs, and present similar results to the mono-sensor approach in the other cases. Therefore in the crisp modeling case the detection performance

is limited to the quality of the output of individual sensors. However, with the event conflict handling approach we showed it is possible to obtain better results than the ones individually achieved by the combined sensors by measuring the event reliability before the fusion process. Moreover, the probabilistic approach would also reduce the influence of errors from low-level sensor in the inference of high-level events.

The hierarchical model based framework (event ontology + event conflicting handling) is a hybrid approach between the hand-crafted semantic-models and the completely learned parameters of Probabilistic Graphical Models, but requiring a much smaller amount of training data. Future work will extend the evaluation of the framework for a larger variety of sensors (heterogeneous and homogeneous) and types of primitive states, and verify possible alternatives to remove the learning step.

References

1. Ali T., Dutta, P. and Boruah H.: A new combination rule for conflict problem of Dempster-Shafer evidence theory. International Journal of Energy, Information and Communications **3:1**, (2012)
2. Allen J.F.:Maintaining Knowledge about temporal intervals. Communications of the ACM, **26,11**, 832-843, (1983)
3. Cao, Y., Tao, L., and Xu, G.: An event-driven context model in elderly health monitoring. *In* Proceedings of. Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 120-124, (2009)
4. Chau, D. P., Bremond, F., and Thonnat, M.: 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.: Generalization of Bayesian inference. J. Royal Statist. Soc., **30**, 205-247, (1968)
6. Fleury, A., Noury, N., Vacher, M.: 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, 322-329, (2010).
7. Folstein, M.F., Robins, L.M., and Helzer, J.E.: The mini-mental state examination. Arch Gen. Psychiatry, **40**, 812, (1983).
8. Gao, L., Bourke, A.K, Nelson, J.: A system for activity recognition using multi-sensor fusion. *In* Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 7869-7872, (2011).
9. Izadinia, H., and Shah, M.: Recognizing complex events using large margin joint low-level event model. *In* Proceedings of the 12th European conference on Computer Vision, 4, Firenze, Italy, 430-444, October, (2012).
10. Lavee, G., Rivlin, E., Rudzsky, M.: 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**, 489-504, (Sep. 2009).
11. Le, Q.V., Zou, W.Y, Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *In* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3361-3368, (2011)
12. Medjahed, H., Istrate, D., Boudy, J., Baldinger, J.-L., Dorizzi, B.: A pervasive multi-sensor data fusion for smart home healthcare monitoring. *In* Proceedings of IEEE International Conference on Fuzzy Systems, 1466-1473, (2011)

13. Nghiem, A. T., Bremond, F., and Thonnat, M.: Controlling background subtraction algorithms for robust object detection. *In* Proceedings of 3rd International Conference on Imaging for Crime Detection and Prevention, London, UK, 1-6, December, (2009).
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, August, (2010).
15. Rong, L., and Ming, L.: Recognizing Human Activities Based on Multi-Sensors Fusion. *In* Proceedings of 4th International Conference on Bioinformatics and Biomedical Engineering, June, 1-4, (2010)
16. Shafer, G.,: A mathematical theory of evidence. Princeton University Press, Princeton, NJ, (1976)
17. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. *In* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 3169-3176, , June, (2011).
18. Vu, T., Bremond, F., and Thonnat, M.: Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. **In** Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 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, September, 136-142, (2012)
20. Zouba, N., Bremond, F., and Thonnat, M.: 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, (2010).
21. Tong, Wei and Yang, Yi and Jiang, Lu and Yu, Shoou-I and Lan, ZhenZhong and Ma, Zhigang and Sze, Waito and Younessian, Ehsan and Hauptmann, AlexanderG. E-LAMP: integration of innovative ideas for multimedia event detection, *Machine Vision and Applications*, 25:1, (2014).
22. Jhuo, I-Hong and Ye, Guangnan and Gao, Shenghua and Liu, Dong and Jiang, Yu-Gang and Lee, D.T. and Chang, Shih-Fu, Discovering joint audiovisual codewords for video event detection, pages=33-47, *Machine Vision and Applications*, 25:1, (2014).
23. Snoek, Cees G. M. and Worring, Marcel and Smeulders, Arnold W. M., Early Versus Late Fusion in Semantic Video Analysis, *In* Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA 2005), 2005, pages 399-402.
24. Myers, GregoryK. and Nallapati, Ramesh and van Hout, Julien and Pancoast, Stephanie and Nevatia, Ramakant and Sun, Chen and Habibian, Amirhossein and Koelma, Dennis C. and van de Sande, Koen E.A. and Smeulders, Arnold W.M. and Snoek, Cees G.M., Evaluating multimedia features and fusion for example-based event detection, *Machine Vision and Applications*, 25:1, 2014, pages 17-32.
25. Oh, Sangmin and McCloskey, Scott and Kim, Ilseo and Vahdat, Arash and Cannons, KevinJ. and Hajimirsadeghi, Hossein and Mori, Greg and Perera, A.G.Amitha and Pandey, Megha and Corso, Jason J, Multimedia event detection with multimodal feature fusion and temporal concept localization, *Machine Vision and Applications*, 25:1, 2014, pages 49-69.
26. Dong, Yuan and Gao, Shan and Tao, Kun and Liu, Jiqing and Wang, Haila. Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications*, 17:1, 2014, pages 37-50.
27. Kris M Kitani and Brian D. Ziebart and J. Andrew (Drew) Bagnell and Martial Hebert, Activity Forecasting, European Conference on Computer Vision, 2012.
28. C. Crispim-Junior, V. Joumier and F. Bremond. A Multi-Sensor Approach for Activity Recognition in Older Patients. In the proceedings of the Second International Conference on Ambient Computing, Applications, Services and Technologies, AMBIENT 2012. Barcelona, September 23-28, 2012.