Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition

Journal:	*Transactions on Pattern Analysis and Machine Intelligence*
Manuscript ID	TPAMISI-2015-03-0205.R2
Manuscript Type:	Regular
Keywords:	I.2.4 Knowledge Representation Formalisms and Methods < I.2 Artificial Intelligence < I Computing Methodologies, I.2.3.l Uncertainty, “fuzzy,” and probabilistic reasoning < H.5.2 User Interfaces < H.5 Information Interfaces and Representation (HCI) < H Information Technology and Systems, Activity recognition, Concept synchronization, I.2.10 Vision and Scene Understanding < I.2 Artificial Intelligence < I Computing Methodologies, Multimedia Perceptual System

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. X, NO. Y, SEPTEMBER 2015

Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition

Carlos F. Crispim-Junior, Vincent Buso, Konstantinos Avgerinakis, Georgios Meditskos, Alexia Briassouli, Jenny Benois-Pineau, Yiannis Kompatsiaris, François Brémond

Abstract—Combining multimodal concept streams from heterogeneous sensors is a problem superficially explored for activity recognition. Most studies explore simple sensors in nearly perfect conditions, where temporal synchronization is guaranteed. Sophisticated fusion schemes adopt problem-specific graphical representations of events that are generally deeply linked with their training data and focused on a single sensor. This paper proposes a hybrid framework between knowledge-driven and probabilistic-driven methods for event representation and recognition. It separates semantic modeling from raw sensor data by using an intermediate semantic representation, namely concepts. It introduces a algorithm for sensor alignment that uses concept similarity as a surrogate for the inaccurate temporal information of real life scenarios. Finally, it proposes the combined use of an ontology language, to overcome the rigidity of previous approaches at model definition, and a probabilistic interpretation for ontological models, which equips the framework with a mechanism to handle noisy and ambiguous concept observations, an ability that most knowledge-driven methods lack. We evaluate our contributions in multimodal recordings of elderly people carrying out IADLs. Results demonstrated that the proposed framework outperforms baseline methods both in event recognition performance and in delimiting the temporal boundaries of event instances.

Index Terms—Knowledge representation formalism and methods, Uncertainty and probabilistic reasoning, Concept synchronization, Activity recognition, Vision and scene understanding, Multimedia Perceptual System.

1 Introduction

The analysis of multiple modalities for event recognition has recently gained focus, especially after the popularization of consumer platforms for video-content sharing, such as YouTube and Vimeo. The need to automatically analyze and retrieve subsets of video content according to textual or image queries has motivated research about ways to semantically describe videos.

This work focuses on a similar problem but different task: event recognition from heterogeneous sensor modalities, where we seek to recognize complex activities of daily living undertaken by people in ecological scenarios. This task requires us to accurately detect and track people over space and time, and recognize concepts and complex events across modalities. At the same time, it is necessary to handle the temporal misalignment of different modalities, and the different sources of uncertainty that intervene in them.

Combining multimodal, visual concept streams from heterogeneous sensors is a problem superficially explored for activity recognition. Single-sensor, data-driven studies have proposed rigid, problem-specific graph representations of an event model [17] [29]. But, once a new source of information is available, these models need to be re-designed from the scratch. On the other hand, knowledge-driven methods provide a generic formalism to quickly model and update events using heterogeneous sources of information [8] [10]. However, their performance degrades drastically in the presence of noise from underlying processes. Finally, most existing work on multimodal scenarios considers nearly perfect settings, where sensors and modalities are completely time synchronized. In real life settings, temporal misalignment among sensors is quite frequent, specially when heterogeneous sensors are combined. This misalignment is commonly aggravated by sensors with variable sampling rates, a characteristic that creates non-linear associations among the time points of different sensors.

In this paper, we propose two contributions for multi-modal event recognition. Firstly, we introduce an algorithm for aligning sensor data using semantic information as a surrogate for the inaccurate time synchronization of real life scenarios. Secondly, we propose a probabilistic, knowledge-driven framework, namely semantic event fusion (SEF), to combine multiple modalities for complex event recognition. The knowledge-driven aspect of our method eases model definition and update, avoiding the long training step required for pure data-driven methods. The probabilistic basis of our event models permits us to handle uncertain and ambiguous observations during event recognition, a limitation for other knowledge-driven methods.

We demonstrate the performance of SEF framework in the combination of different visual sensors (video camera, color-depth, wearable video camera) to recognize Instrumental Activities of Daily Living (IADL) of elderly people during clinical trials of people with dementia. In these settings event recognition needs to be accurate and event temporal intervals precisely assessed, since their results are

- *CFCJ and FB are with STARS team - INRIA Sophia Antipolis Méditerranée, Valbonne, France.*
  E-mail: carlos-fernando. crispim_junior@inria.fr, francois.bremond@inria.fr.
- *VB and JBP are with LABRI - University of Bordeaux, Talence, France*
  E-mail: vbuso@labri.fr, jenny.benois@labri.fr
- *AK, AB, GM, YK are with CERTH - ITI, Thessaloniki, Greece. E-mail: koafgeri@iti.gr, abria@iti.gr, gmeditsk@iti.gr, ikom@iti.gr*

Manuscript received March 6, 2015; revised MM DD, 2015.

used as indicators of a person's performance in such activities. This is the first time such diversity of visual sensors is deployed for this task.

1.1 Framework architecture

The semantic event fusion framework is structured in a hierarchical fashion where, firstly, we use a set of detectors to extract (interpret) low-level concepts from raw sensor data. Secondly, we align sensor concept streams using semantic similarity. Thirdly and finally, we initialize ontological event models with aligned concept observations, and then perform probabilistic event inference for complex event recognition.

The multimodal framework adopts the following definitions:

- **Concept**: any type of object from the real-world or derived from it that is modeled as a physical object or a atomic event (primitive state) in the ontology language.
- **Detector**: a process that provides an interpretation of raw sensor data to the conceptual world.
- **Instance**: an observed example of a concept.

Figure 1 illustrates the architecture of the SEF framework. Detectors (A-C) process their input sensor data ($S0 - S2$) and provide their results as an intermediate, conceptual representation for complex, high-level event inference. The conceptual representation forms the basis to build low-level event models and from their composite and temporal relationship the framework infers complex, composite activities.

The paper is organized as follows. Section 2 summarizes related work. Section 3 presents the methods used for multimodal concept recognition from heterogeneous visual sensors. Section 4 introduces the proposed framework for semantic event fusion. Section 5 presents the dataset and the baseline methods used for evaluation; and Sections 6, 7 and 8 presents Results, Discussion and Conclusions, respectively.

2 Related Work

Activity recognition methods have studied different sensor perspectives to model the semantic and hierarchical nature of daily living activities. Most approaches using heterogeneous sensors focus on simple sensors (*e.g.*, pressure, contact, passive infrared, RFID tags) spread over the targeted environment [14] [25] [19] [23]. Knowledge- and logic-driven methods have been extensively used in these settings [8] [12] [10] [2], as they facilitate the modeling of prior knowledge, sensor data, and domain semantics by means of rules and constraints.

For instance, Cao *et al.* [8] have proposed a multimodal event recognition approach, where they employ the notion of context to model human and environmental information. Human context (*e.g.*, body posture) is obtained from video cameras, while environmental context (semantic information about the scene) is described by inertial sensors attached to objects of daily living. A rule-based reasoning engine is used to combine both contexts for complex event recognition. Chen *et al.* [10] have proposed a hybrid approach between knowledge-driven (ontology-based) and data-driven methods for activity modeling and recognition. Domain heuristics and prior knowledge are used to initialize knowledge-driven event models, and then a data-driven method iteratively updates these models given the daily activity patterns of the monitored person. Even though simple sensors are easy to deploy and maintain, they limit activity recognition to simple phenomena (*e.g.*, opened/closed drawer, presence in the restroom, mug moved), thus limiting the system's ability to describe and recognize more complex and detailed human activities.

Moreover, despite the flexibility of deterministic logic-based methods for event definition, they are very sensitive to noisy observations from underlying components, and they demand the laborious manual definition of all sensor value combinations that satisfy the recognition of an activity. Existing work combining logic and probabilistic methods have proposed to formalize knowledge as weighted rules over raw sensor data [7] [4]. But, the lack of separation between raw-sensor data and event modeling makes these approaches very specific to the environments where they are deployed.

Approaches based on visual signals have focused on probabilistic, hierarchical representations of an event. These representations combine different types of features, from low-level motion and appearance patterns [35] [22] to more semantically rich features (*e.g.*, action segments, context information) [38] [36]. For instance, in [38] authors have proposed to first detect action segments from raw video data, and then use a two-layered Conditional Random Field to recognize activities from the segment patterns and context information (*e.g.*, boolean variables indicating object interaction). Despite the progress of these approaches at activity recognition, they still focus on a single modality, and tend to adopt rigid, problem-specific graph representations for an event. Moreover, to achieve their best performance with proper generalization, they require a large quantity of training data and a training step that may take days.

Studies on video content retrieval have investigated ways to extend the standard low-level, visual feature representations for actions [35] by aggregating other modalities commonly present in video recordings, such as audio and text [27] [29] [17]. In [17], authors have introduced a feature-level representation that models the joint patterns of audio and video features displayed by events. In [29], a multimodal (audio and video) event recognition system is presented, where base classifiers are learned from different subsets of low-level features, and then combined with mid-level features, such as object detectors [21] for the recognition of complex events. These studies have showed that by decomposing complex event representation into smaller semantic segments, like action and objects, inter-segment relations not attainable before can be captured to achieve higher event recognition rates. Nevertheless, these methods only recognize the most salient event in an entire video clip. The task targeted by this paper require us to precisely segment variable-length spatiotemporal regions along the multimodal recording, and accurately classify them into activities.

This paper proposes a hybrid framework between knowledge-driven and probabilistic-driven methods for event representation and recognition. It separates event semantic modeling from raw sensor data by using an intermediate semantic representation, namely concepts. An ontological language is used as a generic formalism to model complex events from their composite relations with concepts and domain knowledge, overcoming the rigidity of hierarchical, graph-based representations. Finally, we propose a probabilistic interpretation for the ontological event models, which equips the framework with a mechanism to handle noise and ambiguous observations.

None of the approaches described above addresses the temporal synchronization of multiple modalities. Most existing work considers nearly perfect settings, where all sensors are at least coarsely time synchronized and have a fixed sampling rate. Therefore, they adopt a sliding time window to accumulate information about event temporal components and to cope with small temporal misalignment between sensors [19] [2] [32]. This multipurpose use of a sliding time window tends to overestimates event duration, since window size is generally set to temporal lengths that are longer than typical event instances. In real-world settings, sensor synchronization is generally inaccurate, and sensors tend to have a variable data sampling rate. These conditions increase alignment complexity and make data fusion very challenging, since they create non-linear associations between the time points of different sensors.

To address the lack of time synchronization between sensors and cope with variable data acquisition rate, we propose a novel algorithm to temporally align sensors using semantic information as a surrogate for inaccurate or missing temporal information. Since the proposed algorithm seeks the global semantic alignment between sensor concept streams, it copes with non-linear associations between different sensor time points. Finally, it also translates all concept streams to the time axis of a reference stream, preserving not only concept temporal relations but also temporal information.

3 Multimodal Concept Recognition

To handle the complexity of real-world activities of daily living and abstract event model definition from low-level data, we adopt multimodal concept detectors to extract low-level concepts from raw sensor data [29] [17] [27]. Three types of concept detectors are used: knowledge-driven event recognition (KER, subsection 3.3), action recognition (AR, subsection 3.1), and object recognition (OR, subsection 3.2).

KER detector employs an off-the-shelf color-depth camera (Kinect, S_0,Fig.1), since this sensor provides real-time, 3D measurements of the scene. These measurements improve the quality of people detection and tracking algorithms by resolving 2D visual ambiguities with depth information, and making these algorithms invariant to light changes. AR detector employs a standard video camera (S_2,Fig.1) due to the broader field of view of this sensor when compared to the color-depth sensor. OR detector complements the previous detectors with a wearable video camera (S_1,Fig.1). This type of sensor has a closer view of the most salient object in the field of view of the person. Salient objects are a key piece of information to describe how activities are realized, and also to overcome situations where a person is occluded or too far from fixed cameras [37].

The novelty of this paper in terms of multimodal activity sensing refers to the variety (or heterogeneity) of visual concept modalities in use, i.e., the phenomena and points of view we use to describe the activities of daily living, and not to a specific choice of sensors. For instance, events from the global displacement patterns of a person, action from the local and finer motion patterns, and the different types of objects being that appear during an activity of daily living.

The choice of sensors that are going to feed the proposed concept detectors can be adapted to user needs. For instance, in a smaller scene than the one used for this work, one may choose to feed AR detector with the RGB image of Kinect instead of using an extra video camera. Alternatively, for the same size of scene, one could replace the Kinect sensor and the video camera with a stereo-camera system and then profit from both the 3D measurements and the scene coverage from a single sensor solution. In summary, the user of the system should select the sensors that provide the best trade-off between scene coverage, system setup complexity and solution cost that fits his/her needs.

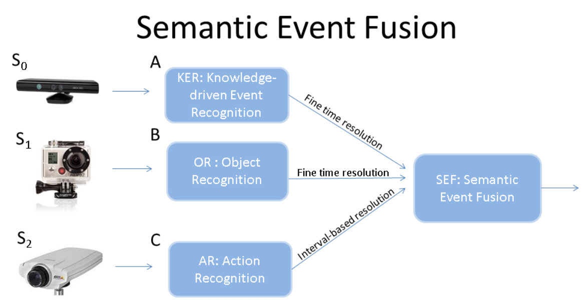

Fig. 1. Semantic event fusion framework: detector modules (A-C) process data from their respective sensors (S0-S2) and output concepts (objects and low-level events). Semantic Event Fusion uses the ontological representation to initialize concepts to event models and then infer complex, composite activities. Concept fusion is performed on millisecond temporal resolution to cope with instantaneous errors of concept recognition.

3.1 Action recognition from color images

Action recognition is usually addressed in the state of the art by localizing actions using a sliding spatiotemporal window [18]. However, these approaches entail a high computational cost due to the exhaustive search in space and time. Furthermore, activities are localized in rectangular spatial areas, which do not necessarily correspond to the area where they actually occur, increasing computational cost and false alarms due to search in irrelevant regions. Rectangular spatial search areas are most likely to contain both a moving entity - *e.g.*, human - and background areas, which both contribute with features to the overall scene descriptor. As a result, the feature vector describing the activity will contain erroneous, false alarm descriptors (from the background). The exhaustive search in time also increases computational cost due to the large number of features being compared and the overlapping sliding window that is usually to improve detection accuracy rates.

We propose a novel algorithm for spatiotemporal localization that overcomes the limitations of the current spatiotemporal sliding window based methods, which both succeeds in reducing the computational cost, while also achieving higher accuracy. To avoid the problems introduced by searching in rectangular spatial areas, we examine only pixels that are likely to contain activities of interest, so spatial localization examines regions of changing motion, the Motion Boundary Activity Areas (MBAAs). To avoid the high computational cost introduced by exhaustive search over time, temporal localization deploys statistical change detection, applied at each frame. Changes are detected in an online manner in the outcomes of a Support Vector Data Description (SVDD) classifier. The SVDD characterizes each activity by a hypersphere built from training data: as it is different for different human activities, changes in its outputs also correspond to different activities. The resulting method for sequential detection of changes between SVDD outcomes, where the latter use only data inside MBAAs, is thus called Sequential Statistical Boundary Detection. The sequential nature of the change detection results in a faster activity boundary detector, as sequential change detection has been proven to provide quickest detection.

Action cuboids are then extracted in the resulting subsequences. The action cuboids are much smaller in size than regions used for spatiotemporal activity localization and precisely localize pixels with the activity of interest, both in space in time. Thus, their motion and appearance properties are used for recognition in a multiclass SVM model. In concluding, the main novelty of our approach lies in the spatiotemporal activity localization. Spatial localization also takes place in an original manner, by isolating regions of changing activity, thus avoiding false alarms and increasing the system's accuracy, while temporal localization is accelerated as fewer subsequences need to be classified in order to detect the activities that occurs inside them [3]. This detector provides valuable cues about the actions taking place given its local motion patterns, but it does not identify the author of the action. For this reason, AR detector is a natural complement for the knowledge-driven event recognition that recognizes person-centered events (subsection 3.3).

3.2 Object recognition from egocentric vision

We employ several detectors of "active objects" (objects either manipulated or most salient in the field of view of the user), as we consider that the identification of these objects is a crucial step towards activity understanding. The recognition of activity-related objects adds more robustness to event models, especially when the emphasis is placed on activities of daily living. OR detector considers one concept detector per object category. The processing pipeline (Fig. 2) is shared by all detectors until the image signature step. A nonlinear classification model is learned for each object category.

We have built our model based on the well-known Bag-of-Words (BoW) paradigm [13] and used saliency masks as a way to enrich the spatial discrimination of the original BoW approach. Hence, for each frame in a video sequence, we extract a set of N SURF descriptors d_n [5], using a dense grid of circular local patches. Next, each descriptor d_n is

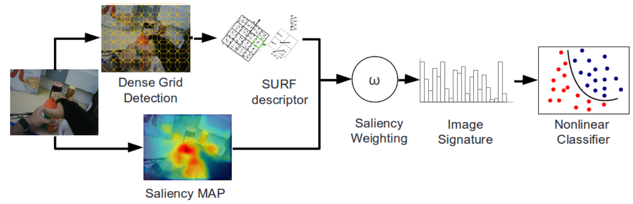

Fig. 2. Processing pipeline for saliency-based object recognition in first-person camera videos

assigned to the most similar word $j = 1..V$ in a visual vocabulary by following a vector-quantization process. The visual vocabulary is computed using k-means algorithm over a large set of descriptors of the training data set. We set the size of dictionary V to 4000 visual words. In parallel, our system generates a geometric spatiotemporal saliency map S of the frame with the same dimensions of the image and values in the range $[0, 1]$ (the higher the S the more salient a pixel is). Details about the generation of saliency maps can be found in [6]. We use the saliency map to weight the influence of each SURF descriptor in the final image signature, so that each bin j of the BoW histogram H is computed by the next equation:

$$H_j = \sum_{n=1}^{N} \alpha_n w_{nj}, \tag{1}$$

where the term $w_{nj} = 1$ if the descriptor or region n is quantized to the visual word j in the vocabulary and the weight α_n is defined as the maximum saliency value S found in the circular region of the dense grid.

Finally, the histogram H is L1-normalized to produce the image signature. A SVM classifier [11] with a non-linear χ^2 kernel [33] is then used to recognize the objects of interest over the weighted histogram of visual words. Using Platt approximation [30], we produce posterior probabilistic estimates O_k^t for each occurrence of an object k in frame t.

3.3 Knowledge-driven event recognition

KER detector equips the SEF framework with the ability to handle multiple people in the scene and derive person-centered events. Its processing pipeline is decomposed into people detection, tracking, and event recognition.

3.3.1 People Detection

People detection is performed using the depth-based framework of [28] that extends the standard detection range of color-depth sensors from 3-4 meters (Microsoft and PrimeSense) to 7-9 meters away. It works as follows: first, it performs background subtraction in the depth image to identify foreground regions that contains both moving objects and potential noise. These foreground pixels are then clustered into objects based on their depth values and neighborhood information. Among these objects, people are detected using a head and shoulder detector and tracking information about previously detected people.

3.3.2 People Tracking

People tracking [9] takes as input the video stream and the list of objects detected in the current and previous frames using a sliding time window. First, a link score is computed between any two detected objects in this time window using a weighted combination of six object descriptors: 2D and 3D positions, 2D object area, 2D object shape ratio, color histogram and dominant color. Then, successive links are formed to represent the several paths an object can follow within the temporal window. Each possible path of an object is associated with a score given by all the scores of the links it contains. The object trajectory is determined by maximizing the path score using Hungarian algorithm [20].

3.3.3 Event representation and recognition

We extend the declarative constraint-based ontology language proposed in [34] [12] to define event models based on *prior* knowledge about the scene, and real-world objects (*e.g.*, person) dynamically detected by underlying components (*e.g.*, people detection and tracking).

An event model is composed of three main parts:

- **Physical Objects** refer to real-world objects involved in the realization of the event (*e.g.*, person, kettle).
- **Components** refer to sub-events of which the model is composed of.
- **Constraints** are conditions that the physical objects and/or the components should satisfy.

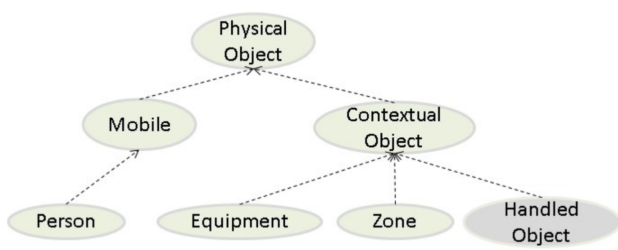

Fig. 3. Physical object sub-tree of the ontology language

KER detector uses three types of physical objects (Fig. 3): person, zones and equipment. Constraints are classified into non-temporal (*e.g.*, inter-object spatial relations, object appearance); and temporal (*e.g.*, time ordering between two event components). Temporal constraints are defined using Allen's interval algebra, *e.g.*, BEFORE, MEET, AND [1]. An alarm clause can be optionally defined to rank events by their importance for a sub-subsequent task, *e.g.*, to trigger an external process.

Events are hierarchically categorized by their complexity as (in ascending order):

- **Primitive State** models a value of property of a physical object constant in a time interval.
- **Composite State** refers to a composition of two or more primitive states.
- **Primitive Event** models a change in value of a physical object's property (*e.g.*, posture), and
- **Composite Event** defines a temporal relationship between two sub-events (components).

This detector provides person-centric events derived from knowledge about global spatiotemporal patterns that people display while performing activities of daily living. Example 1 illustrates the low-level, primitive state model $Person\_inside\_ZonePharmacy$ that maps the spatial relation between a person's position and the contextual zone $zPharm$. For instance, this zone may corresponds to the location of a medicine cabinet in the observed scene.

Example 1. Primitive state "Person inside Zone Pharmacy"

```
PrimitiveState(Person_inside_ZonePharmacy,
 PhysicalObjects( (p1: Person),(zPharm: Zone))
 Constraints(
   (p1->position in zPharm->Verticies)
 Alarm ((Level : NOTURGENT))
)
```

4 Semantic Event Fusion

The abovementioned concept detectors for actions, knowledge-based events and objects constitute the foundations of the semantic event fusion framework. They bridge the gap between the raw sensor data and the conceptual world and provide a natural separation between data specifics and event semantic modeling.

SEF takes place over concept observations and is responsible for linking these concept instances to related event models, and then infer whether the available evidence is sufficient to recognize one of the target events. To achieve this goal, SEF needs to handle the time misalignment among sensors and the different sources of uncertainty that intervene in concept and complex event recognition.

We divide SEF framework into four steps: model representation, semantic alignment, event probability estimation, and complex event probabilistic inference.

4.1 Model Representation

To represent the concept dependencies and semantics of complex events (*e.g.*, temporal order that involved concepts need to display), we extend the constraint-based ontology language used in KER detector to multimodal composite events.

The mapping between concept detector observations and the ontology language representation is performed as follows: actions from the AR detector are mapped to instances of primitive states. Objects from the OR detector are linked as instances of a new class of physical object, namely handled object. This class, as the name suggests, represents objects that can be manipulated with the hands (*e.g.*, kettle, teabag, pillbox, *etc*). Finally, events from the KER detector are mapped as instances of low-level, composite events.

Example 2 presents the ontological model of the multimodal, composite event "PreparePillBox_SEF". This model combines multimodal physical objects (person, zone, and handled object) and sub-events "PreparePillBox_KER" and "PreparePillBox_AR".

Example 2. Multimodal, Composite Event "Prepare pill box"

```
CompositeEvent(PreparePillBox_SEF,
 PhysicalObjects( (p1: Person),(zPharm: Zone),
                  (PillBox: HandledObject))
 Components(
```

```
   (c1: PrimitiveState PreparePillBox_AR() )
   (c2: CompositeEvent PreparePillBox_KER(
                              p1,zPharm)))
 Constraints(
   (c1->Interval AND c2->Interval)
   (duration(c2) > 3))
 Alarm ((Level : URGENT))
)
```

The concept dependencies of a complex event are the basis to quantify concept similarity across different visual modalities and hence align them, and to estimate composite event probability for probabilistic event inference. Figure 4 illustrates the concept dependencies extracted from the multimodal event "Prepare drink SEF".

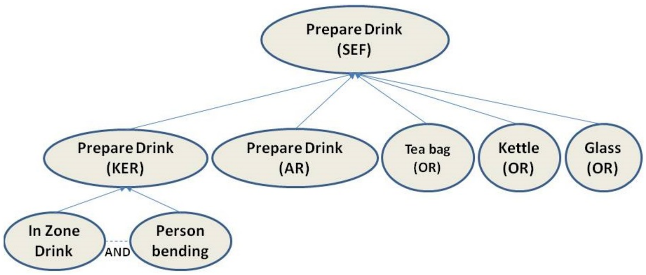

Fig. 4. Composite relations between concepts and event models. Multimodal Event "Prepare drink" is composed of conceptual events "prepare drink" from KER and AR detectors and conceptual objects "Tea bag", "Kettle" and "Glass" from OR detectors. For instance, the hierarchically lower event "Prepare drink" from KER detector can be further decomposed into two sub-events, while other detector concepts are atomic.

4.2 Semantic alignment

To align heterogeneous concept streams we propose a novel algorithm that uses concept similarity as surrogate for inaccurate temporal information. For instance, concepts are considered similar if they are part of the same complex event. However, semantic alignment is a complex problem on its own, since two concepts related to the same complex event may model very different aspects of the given event. For example, while the OR detector will generate fine-grained object-wise observations about the activity taking place (*e.g.*, telephone), KER detector will generate event-wise observations for the same period of time (*e.g.*,"talk on the telephone"). These conceptual differences create non-linear matches between concept streams. Similarly, some sensors might have variable sampling rates, a characteristic which may introduce non-linear time deformations in the derived concept stream.

To find the non-linear alignment between two concept streams we employ Dynamic Time Warping (DTW), an algorithm that seeks the optimal alignment between two time-dependent sequences [26]. By seeking for the global semantic alignment, we overcome both the coarse ontological alignment of concept detectors and the non-linear deformations introduced by the variable sampling rate of sensors.

Algorithm 1 describes the proposed method for semantic alignment. The algorithm starts by identifying each complex event with an unique code. Then, for each concept stream s_i, it creates an encoded concept stream c_i, where concepts are represented by the code of the complex event they belong to. Once all encoded streams are generated, they are aligned to the encoded reference stream (c_0, KER detector), in a pairwise manner, using the DTW variant proposed by [31]. For each warped concept stream $c_{w,i}$ generated by DTW, the temporal translation function Δ determines the warping deformations (position additions) that the alignment to c_i stream has introduced into c_0. By pruning the new positions in $c_{w,0}$ from $c_{w,i}$, function Δ projects $c_{w,i}$ into the time axis of the original reference stream c_0. Finally, we remove spurious, instantaneous concepts from the concept stream $c_{a,i}$ using median filtering.

Algorithm 1. Pseudo-code of the semantic alignment

//Shared semantic encoding
for each $s_i \in S$:
    for each $t \in s_i$:
        $c_i(\mathrm{t}) = \Omega(s_i(\mathrm{t}))$

//Semantic Alignment and Temporal Projection
$\mathrm{C} = \mathrm{C} \setminus c_0$
for each $c_i \in C$:
    $c_{w,0}$, $c_{w,i} = \Phi(c_0,c_i)$
    $c_{a,i} = \Delta(c_0$, $c_{w,0}$, $c_{w,i})$
    $c_{f,i} = \mathrm{medianFiltering}(c_{a,i})$

where,

- Ω : maps concepts to the composite event they are part of,
- s_i, S : concept stream i, and its set S,
- c_i, C : encoded concept stream i and its set C,
- t : time point t,
- $c_{w,i}$: warped version of c_i,
- $c_{a,i}$: aligned version of c_i,
- $c_{f,i}$: smoothed version of $c_{a,i}$,
- Φ: DTW function,
- Δ: temporal translation function.

The proposed algorithm assumes that the events and concepts used for the semantic alignment have an one-to-many relationship, respectively. To achieve the optimal alignment, the proposed algorithm requires that the streams have a sufficient amount of similar concepts, and that concept detectors have a reliable performance at the recognition of these concepts.

Concept similarity is extracted from the ontological representation of complex events (targeted activities). KER detector is chosen as the reference stream due to its sensor sampling rate be on an intermediate temporal resolution compared to other sensors, and due to its high performance at the recognition of different concept classes.

For probabilistic concept detectors that provide a confidence value for all their concept classes at every time point t, like OR detector, the alignment procedure implements two extra steps. Before the semantic alignment, we generate a concept stream s_g from the most likely concept of the detector at each time point t. Then, we semantically align the stream s_g and the reference stream. Once alignment is done, we use the temporal translation data found for s_g

and the reference stream to generate aligned, object-specific concept streams. These extra procedures are necessary since a single-object concept stream will most of the time lack enough semantics for accurate semantic alignment.

4.3 Event Probability Estimation

To estimate event probability from the combination of multi-modal sources of information is not a trivial task, since each modality carries different sources of uncertainty. For example, to accurately fuse multiple concepts it is necessary to consider not only the concept detector confidence on a given instance, but also its reliability as an source of information. Additionally, the relevance of a concept for an event model should be modeled to fully profit from the complementary nature of multimodal sources of information.

Studies in video content retrieval have mostly explored the complementary information provided by different modalities of raw video signals. Currently they lack mechanisms to handle other factors that interfere in real-world applications of event recognition, like information relevance and reliability. For instance, motion features should have a higher relevance than appearance features to discriminate walking from standing events. Similarly, a concept detector from a wearable camera should be more reliable in object recognition than one derived from a fixed camera attached to the ceiling of a room.

We formalize the probability of a composite event (cs) as a function of the probability of its concepts (ce) and the factors that affect them. We use a Countable Mixture Distribution (CMD, Eq. 4) to integrate the concepts' probability and the factors that intervene in them (concept weight). Concept weights are defined based on two factors: the concept relevance to the given event model and the concept reliability given a detector. Equation 5 presents the proposed CMD, which quantifies the probability of a composite event given its observed concepts. A partition function (Eq. 7) is adopted to normalize the weights of the CMD.

Reliability (RB) handles detector differences in concept recognition. It measures the detector precision at the recognition of each one of its concepts (Eq. 2). Relevance (RV) models the contribution of a concept to the recognition of a given event (Eq. 3). It also facilitates event modeling, since domain experts can focus on listing concepts they deem important for a complex event, and the framework will learn the degree of relevance of each assigned concept to the given event model.

$$P(ce_{i,j,k}|d_k) = \frac{|TP|}{|TP| + |FP|} \tag{2}$$

where,

- $P(ce_{i,j,k}|d_k)$: reliability of concept i part of composite event j given detector k,
- $|TP|$: number of times concept ce_i is correctly recognized by concept detector k during a true instance of composite event j,
- $|FP|$, number of times ce_i is observed by the concept detector k given there is no true realization of event j.

$$P(cs_j|ce_{i,j,k}) = \frac{|ce_{i,j,k} \cap cs_j|}{|ce_{i,j,k}|} \tag{3}$$

where,

- $P(cs_j|ce_{i,j,k})$: number of times composite event cs_j is detected during an instance of concept $ce_{i,j,k}$,
- $|ce_{i,j,k} \cap cs_j|$: number of times $ce_{i,j,k}$ is present during an instance of event cs_j,
- $|ce_{i,j,k}|$: number of times $ce_{i,j,k}$ is observed.

$$\begin{gathered} f(x) = \sum_{i=1}^{N} w_i \times P(x_i), \\ w_i >= 0, \\ \sum w_i = 1 \end{gathered} \tag{4}$$

$$P(cs_j) \quad = \quad \frac{\sum_{ce_{i,j,k} \in cs_j} \; w(ce_{i,j,k}) \times P(ce_{i,j,k})}{Z(cs_j)} \tag{5}$$

$$w(ce_{i,j,k}) \quad = \quad P(cs_j|ce_{i,j,k}) \; + \; P(ce_{i,j,k}|d_k) \tag{6}$$

$$Z(cs_j) = \sum_{ce_{i,j,k} \in cs_j} w(ce_{i,j,k}) \tag{7}$$

where:

- $P(cs_j|ce_{i,j,k})$: conditional probability of composite event j given concept k from detector i,
- $P(ce_{i,j,k})$: probability of concept k from detector i, part of composite event j,
- $P(ce_{i,j,k}|d_k)$: reliability of concept i from composite event j given detector k,
- $w(ce_{i,j,k})$: weight of concept ce_k from detector i, part of composite event j.

CMD models provide a compact representation of the different random variables that intervene in the estimation of the probability of the modeled event. It speeds up event inference, since the probability of an event probability is locally estimated based only on the probability of related concepts and uncertainties.

4.4 Probabilistic Inference

Event models guide the inference process considering evidence related only to the event model in analysis, then reducing the computational complexity of the inference process. Logic and temporal constraints can be then used throughout the event inference step to impose real-world constraints to event models. Probabilistic inference equips the framework with means to handle event ambiguity over mutually exclusive complex events, and to filter out events which are unlikely to correspond to real-world events.

The inference step takes as input the concepts extracted by the visual concept detectors at each time t, and links them as parts of related composite event models. For each event, it computes event probability using the corresponding CMD model (Eq. 5). *Maximum a posteriori* (Eq.8) is employed to retrieve the most likely event from a set of mutually exclusive candidates. Finally, probability thresholding is used over the most likely event to decide whether its probability corresponds to a real-life event. Probability thresholding provides an efficient way to find the probability level from where a complex event CMD has sufficient evidence to recognize a real-world event. Moreover, it can be easily translated into a supervised learning problem of parameter tuning, and it preserves semantic meaning for human analysis.

$$cs = \begin{cases} argmax_{cs_j} P(CS), & if\ P(cs_j) > th_{cs_j} \\ \emptyset, & otherwise \end{cases} \tag{8}$$

where,

- cs: most likely composite event,
- CS: set of mutually exclusive composite events,
- th_{cs_j}: probability threshold th_{cs_j} for the recognition of the composite event j.

4.5 Parameter Learning

The parameters of the SEF framework are determined using a supervised learning method (*maximum likelihood estimation*) in a 10-fold cross-validation scheme. Three main parameters are learned for the estimation of the probability of an event model: the RV and the RB of a concept, and the probability threshold of an event. These parameters are computed based on the overlap between instances of concepts and ground-truth annotations of composite events. Ground-truth instances are annotated by domain experts visualizing recordings of the color-depth sensor.

Finally, the composite relations between concepts and a complex event, which are necessary for semantic alignment and event probability estimation, are extracted from complex event models. Event models are provided by domain experts using the multimodal event ontology representation.

5 Experiments

The evaluation of the proposed framework for multimodal event recognition is performed as follows: firstly, we evaluate the effects of the semantic alignment over the performance of concept detectors. Secondly, we evaluate the overall semantic fusion by comparing its results to two baseline methods: Ontology-based Semantic Fusion (OSF, Subsection 5.2) and Support Vector Machine (subsection 5.3). All evaluations are run over multimodal recordings of elderly people carrying out activities of daily living (subsection 5.1). Results are reported for validation and test sets of a 10-fold cross-validation scheme.

5.1 Data set: monitoring activities of senior people

Participants aged 65 years and above were recruited by the Memory Center (MC) of Nice Hospital. The clinical protocol asks participants to undertake a set of physical tasks and IADLs in a hospital observation room, furnished with home appliances [15]. Experimental recordings used two fixed cameras: color-depth camera (Kinect ®, Microsoft ©, ~10 frames per second), standard color camera (AXIS®, Model P1346, 8 frames per second); and a wearable camera, GoPRO Hero - first generation.

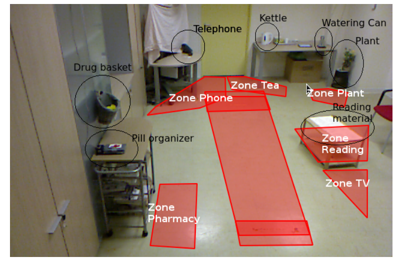

Fig. 5. Observation room where daily living activities are undertaken. Contextual zones are depicted as free-from closed polygons in red, and contextual objects as black ellipses.

Participants undertake IADLs for approximately 15 minutes, as the clinical protocol aim is to evaluate the level of autonomy of the participant by organizing and carrying out a list of these activities. Figure 5 illustrates the observation room where participants undertake IADLs, and the semantic zones that are annotated to incorporate *a priori* knowledge about the scene.

The clinical protocol IADLs are the following:

- Prepare drink (P. Drink, *e.g.*, prepare tea/coffee),
- Talk on the telephone (T. Telephone, *e.g.*, calling, answering),
- Read (*e.g.*, read newspaper, magazine),
- Prepare pill box (P. Pill box),
- Manage finances (M. Finances, *e.g.*, write a check, establish account balance),
- Search bus line (S. Bus line)
- Water the plant (W. Plant), and
- Watch TV (W. TV).

OR detector produces probability estimations [0,1] over 12 visual concepts: account, medication basket, checks, instructions (activities to perform), kettle, map, medical instructions, telephone, remote, TV, tablet, and watering can.

AR detector provides estimations about a set of mutually exclusive atomic actions: answer phone, call phone, look on map, pay bill, prepare drugs, prepare drink, read paper, water plant, and watch TV. KER detector generates events for all protocol activities, except for "watch TV" and "search bus line".

5.2 Baseline 1: Ontology-based Semantic Fusion

The ontology-based framework for semantic fusion (OSF) [24] is based on the use of RDF/OWL [16] ontologies to capture the dependencies among low-level domain observations and complex activities (events). More specifically, following a knowledge-driven approach, it defines the Context Dependency Models of the domain that captures the background knowledge required to detect the complex activities. The context dependency models serve as input to the semantic interpretation procedure for the recognition and classification of complex activities. The objective of the interpretation procedure is to analyze traces of observations provided by the various modules of the application domain and group them into meaningful situations, classifying them as complex activities. The interpretation algorithm consists of three steps: (a) definition of partial context, (b) identification of contextual links and (c) recognition and classification of situations. Details about the OSF approach are available in [24].

The ontology-based semantic fusion serves as a baseline for the delimitation of the temporal boundaries and the recognition of events if a holistic view of the concepts of the entire multimodal recording is employed. Its limitations are the following: it cannot handle interleaved activities, nor can it resolve conflicts after the recognition process. It also does not handle dynamic and incremental generation of partial contexts and context links in (near) real-time activity recognition, as it uses all recognized events. Finally, this baseline approach does not handle uncertainty in the input data, and assumes all observations (primitive and high-level) have the same confidence (100%).

5.3 Baseline 2: Support Vector Machine

The second baseline consists of linear SVM classifiers that learn to recognize activities of daily living from multimodal concept instances observed during a time-window. This method demonstrates the fusion performance of a fully supervised learning approach, which operates over a conceptual representation of raw sensor data (KER events, OR objects, and AR actions), and learns the best combination of concept observations from training data. The input for this baseline is a normalized histogram of concept observations across semantically aligned, concept streams. We compute a histogram for the concepts of each composite event across all concept streams during a time window. In the training set, time windows correspond to the exact time interval of the events from ground-truth data. For validation and test sets we browse the recording in a frame-wise fashion and compute histograms over a continuous sliding time window. The search for the most appropriate size for the time-window started with the average duration of activity classes in the training set. Model parameters and time-window size are learned and evaluated in the same 10-fold cross-validation scheme used to learn the parameters of the proposed approach. One-versus-all scheme is adopted to learn the classifier of each composite event. Model parameters are chosen based on the performance of the baseline method in the validation set.

5.4 Evaluation

To evaluate the proposed methods, we quantify the frame-wise agreement between the output of evaluated methods with event annotation provided by domain experts (ground-truth data). Frame-wise agreement may seem strict, but our goal is to achieve a high event recognition rate and a precise assessment of the temporal boundaries of event instances. Performance results are reported on the cross-validation scheme test sets, unless specified otherwise. F_1-score is employed as the performance index.

For the evaluation of the semantic concept synchronization method, we compare the performance of detectors AR, KER and OR before synchronization (NA), warped and smoothed (WS), and semantically synchronized (warped, backprojected and smoothed, WBS). Warped variant of concept streams are provided as a performance baseline to the temporal translation step of the semantic alignment.

To evaluate the semantic event fusion framework, we compare its results to the performance of two state-of- the-art baselines at two capabilities. Firstly, at the accurate fusion of concepts under the presence of ambiguous and noisy observations; and secondly, at the precise assessment of event time intervals. We also provide the performance of concept detectors as a reference to measure whether the proposed method can go beyond their individual performances by combining their complementary aspects.

6 Results

6.1 Semantic alignment

Figure 6 illustrates an example of semantic alignment between the concept stream of AR detector and a concept stream generated from the events annotated by a domain expert (color-depth sensor images are used as reference). We observe that the proposed technique accurately translates the AR detector stream from its original form - coarsely synchronized - to a new form that is optimally time-synchronized with the reference stream, and also preserves most shape characteristics of the original concept stream of AR detector.

Table 1 presents a quantitative evaluation of the gain in performance obtained by aligning the concept detector streams. To evaluate the improvement brought by the alignment, we assess the performance of each concept detector at individually recognizing the composite event they are part of. We present results for three cases: the original concept streams; the warped case, where both ground-truth and sensor stream are optimally aligned, and at last, the semantically aligned concept stream.

The semantic alignment improves the performance of the KER detector compared to its original version for all event classes, apart from “prepare pill box” event. It also displays a higher performance than the warped case in three out of seven classes, while being quite close for the remaining ones (*e.g.*, “prepare drink”, “talk on the telephone”, and “watch TV“ events). In AR case, the aligned streams perform better than the original stream for all cases, but worse than the warped streams for half of the events (“prepare drink”, “reading”, “talk on the telephone” and “watch TV” events). Finally, the aligned streams of OR detector outperform the original ones for all cases, except for two event classes: “prepare pill box” and “talking on the telephone”. Currently, the aligned concept streams of OR performs worse than their warped streams for the majority of cases.

6.2 Semantic Event Fusion

Figure 7 presents the performance of the semantic event fusion in the validation set and according to the probability threshold adopted. We observe that most event classes have their highest recognition rates adopting a probability threshold between 0.4 and 0.5. Exceptions are “search bus line” and “talk on telephone” events, where the threshold value of 0.1 achieves the highest performance.

Table 2 compares the performance of the SEF framework (with and without probability thresholding) to its individual

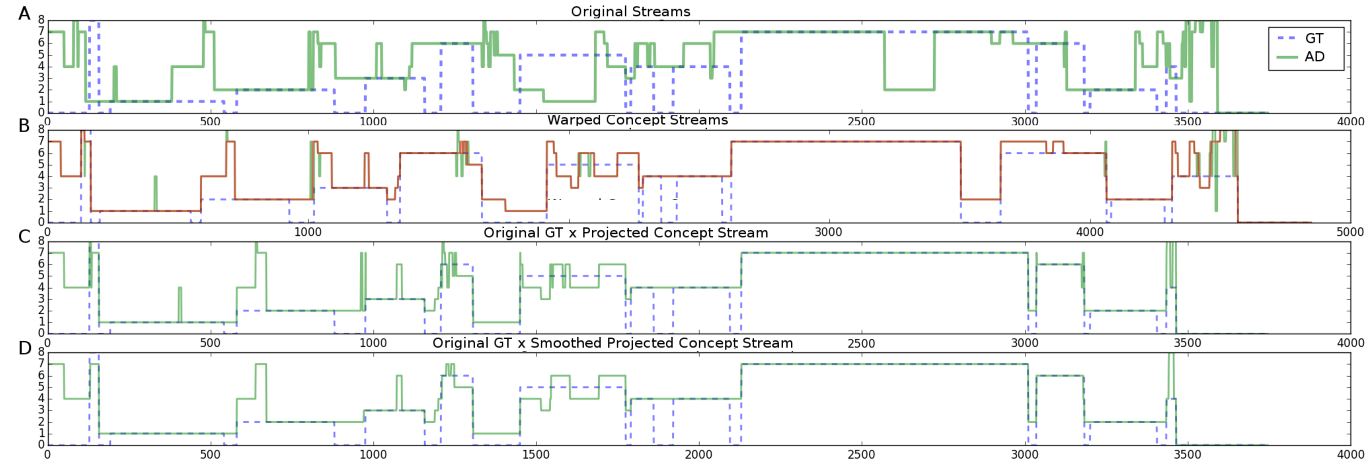

Fig. 6. Semantic alignment between the concept stream of the action recognition detector (AR) and a concept stream (GT) generated from events manually annotated by domain experts using the time axis of the color-depth camera. X-axis denotes time in frames, and Y-axis denotes activity code (1-8), respectively, search bus line on the map, establish bank account balance, prepare pill box, prepare a drink, read, talk on the telephone, watch tv, and water the plant. From top to bottom, images denote: (A) original GT and AR streams, (B) GT and AR streams warped, AR stream warped and smoothed (in red), (C) original GT and AR stream warped and then backprojected onto GT temporal axis, (D) original GT and AR warped, backprojected, and then smoothed with median filtering.

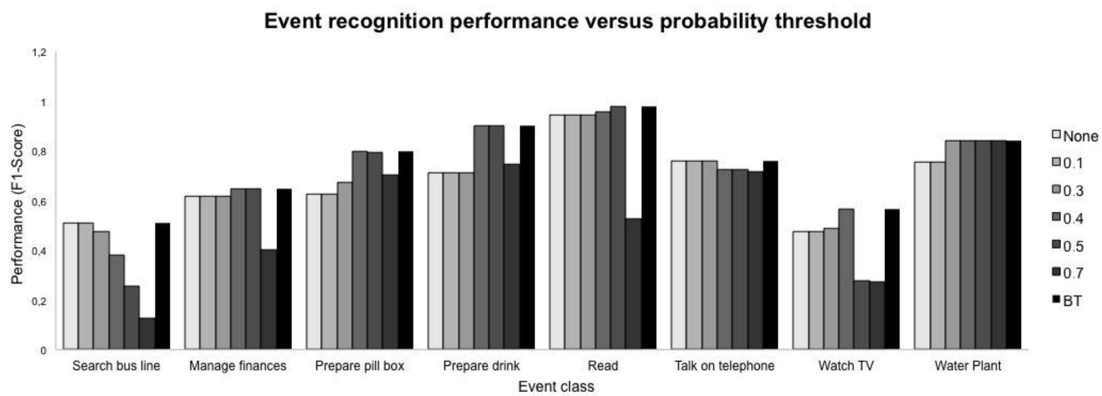

Fig. 7. Event recognition performance according to probability threshold. BT refers to the threshold with best performance for each event

TABLE 1
Semantic Alignment versus Event Recognition

mean F_1-score	Detector / Stream alignment								
	KER			AR			OR		
IADL	NA	WS	WBS	NA	WS	WBS	NA	WS	WBS
S. Bus line	16.6	27.7	27.8	40.9	44.3	45.0	11.7	13.7	13.7
M.Finances	0.0	0.0	0.0	61.7	60.9	62.1	26.7	30.9	28.7
P. Pill box	69.0	61.8	62.6	49.4	55.3	57.1	23.8	24.5	21.7
P. Drink	71.9	86.6	85.9	31.6	51.4	49.2	0.0	0.0	0.0
Read	73.8	97.9	98.2	50.8	62.9	56.8	0.1	8.3	7.0
T.Telephone	68.2	83.9	83.3	38.9	66.5	60.9	13.7	14.2	13.0
W. TV	9.9	30.5	27.3	17.2	42.9	36.5	10.1	17.1	14.7
W. Plant	47.4	86.4	86.4	9.0	21.4	21.9	0.0	0.0	0.0

N: 17 participants; 15 min. each; Total: 255 min.
(−) denotes concepts not available for the detector.
AR: action recognition, KER: Knowledge-driven event recognition, and OR: Object recognition.
NA: Not aligned, WS: warped and smoothed, and
WBS: warped, and backprojected and smoothed

concept detectors, before and after semantic alignment, on the validation set. Results demonstrate that the proposed framework has a performance higher than the semantically aligned versions of its individual detectors, with two exceptions: "managing finances" and "talking on the telephone" events. For the first event, the stream of the action detector without alignment has a performance 9% higher than the proposed method, while for the second event the aligned version of KER detector has a performance 14% higher. Probability thresholding improves the event recognition in the majority of cases.

Table 3 compares the performance of the proposed framework to the individual concept detectors in the test set, before and after semantic alignment. The proposed framework outperforms methods only using individual concept detectors in all cases and classes, with the exception of aligned KER in the events "talk on the telephone" (-17.5%), reading (-2.8 %), and search bus line (-1%).

Figure 8 illustrates the F_1-score of 12 classes of objects provided by OR concept detector. We observe that OR method has an average $F_1 - score$ performance of 56 % in 9/12 classes that appear in the test set recordings, and 43 % when considering all of them. The average precision of OR is 85.77 %, which demonstrates the high reliability of its

TABLE 2
Event recognition performance in the validation set

mean F_1-score	Stream alignment / Detector							
	None			Aligned			Proposed	
IADL	KER	AR	OR	KER	AR	OR	WT	BT
S. Bus line	13.1	47.3	8.3	13.9	45.3	17.8	51.1	51.1
M.Finances	0.0	73.0	24.0	0.0	66.7	27.0	61.8	64.7
P. Pill box	71.4	55.1	21.4	66.3	56.7	24.0	62.4	79.7
P. Drink	77.6	37.2	0.0	91.6	53.5	0.0	71.2	91.0
Read	73.2	49.9	0.0	98.2	54.8	0.5	94.5	97.7
T.Telephone	65.9	44.0	14.4	89.0	62.6	14.9	75.8	75.8
W. TV	13.0	22.4	11.9	30.0	44.9	17.6	47.6	56.6
W. Plant	45.3	11.0	0.0	83.4	26.7	0.0	75.6	84.2

WT: without probability thresholding
BT: event recognition performance of the best threshold values

TABLE 4
Event recognition performance versus concept detector composition

IADL	pairwise			All
	A+O	K+O	K+A	K+A+O
S. bus line	43.82	0	43.64	73.11
M.finances	43.68	35.99	43.68	43.73
P. pill box	58.75	55.84	60.31	65.02
P. drink	27.59	63.4	54.21	64.04
Read	68.86	97.6	93.94	95.22
T.telephone	54.17	74.18	92.48	75.58
W. TV	8.42	8.89	20.68	35.8
W. Plant	0	50	98.97	100
Average	38.16	48.24	63.49	69.06

A+O: AR and OR; K+O: KER and OR
K+A+O: KER and AR and OR

TABLE 3
Event recognition performance in the test set

mean F_1-score	Stream alignment / Detector						
	None			Aligned			Proposed
IADL	KER	AR	OR	KER	AR	OR	BT
S. Bus line	28.6	19.6	23.2	74.1	43.8	0.0	73.1
M.Finances	0.0	27.4	37.6	0.0	43.7	35.4	43.7
P. Pill box	60.6	28.6	32.4	49.1	58.7	24.7	65.0
P. Drink	43.7	4.0	0.0	57.5	27.6	0.0	64.0
Read	77.2	56.2	0.6	98.0	68.9	45.9	95.2
T.Telephone	77.6	18.7	10.7	93.1	54.2	5.2	75.6
W. TV	0.0	0.0	4.3	18.5	8.4	5.1	35.8
W. Plant	56.8	0.0	0.0	100.0	0.0	0.0	100.0

TABLE 5
Comparison to baseline methods in the test set

mean F_1-score	Fusion approach		
	Baselines		Ours
IADL	SVM	OSF	
S. bus line	44.19	31.36	73.10
M.finances	43.99	0.00	43.73
P. pill box	45.86	49.11	65.02
P. drink	20.02	24.29	64.03
Read	90.18	91.82	95.22
T.telephone	72.12	0.00	75.58
W. TV	2.32	0.00	35.80
W. Plant	0.00	0.00	100.00
Average	39.83	24.57	69.06

OSF: Ontology-based Semantic Fusion

observations.

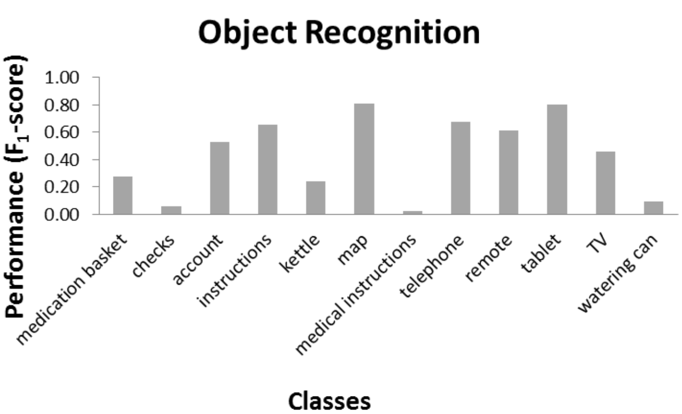

Fig. 8. Performance of OR concept detector per object class.

Table 4 presents the performance of the SEF framework at event recognition varying the concept detectors in use from a single detector to their pairwise combination, up to the full set. We observe that SEF presents the highest performance for six out of eight IADLs, and OR module has a complementary role to another detectors.

Table 5 compares the performance of the proposed approach to two baselines methods: OSF, and SVM. We observe that the proposed semantic event fusion outperforms all baseline approaches.

7 Discussion

7.1 Semantic alignment

We have proposed a method for heterogeneous visual sensor alignment based on semantic similarity. Results at event recognition level show that semantically aligned, concept detectors outperform their original form and their warped variant in the majority of cases. As such, our method is capable of accurately translate the optimal alignment achieved at warped space to the temporal axis of the reference concept stream.

Regarding the cases where the semantically aligned concept streams perform worse than their warped version, this behavior is mostly due to a loss of information during the temporal projection of the warped concept stream onto the temporal axis of the reference stream. This loss mainly happens when DTW removes time points from stream regions with a high variance in concept classes for a brief period of time. Changes in these regions severely penalize the performance of the aligned method if less-frequent, short-lengthened concepts are removed because they are temporally closer to longer concepts used for matching.

Finally, for the cases where the original concept stream outperforms both synchronized and warped streams, results suggest that this case is due to the DTW algorithm has not achieved the optimal alignment between the two streams.

7.2 Semantic Event Fusion

The evaluation of SEF framework performance according to the set of concept detectors used (Table 4) has shown out that all concept detectors provide meaningful information and are complementary. This is corroborated by the fact that the combination of the three concept detectors outperforms their pairwise combinations in six out of eight investigated IADLs. It has also shown that even if the observations of a given detector have a poor performance when directly mapped from concepts (*e.g.*, OR module, Figure 8) onto activity observations (*e.g.*, Table 3), SEF can still use them as a complementary source of information (*e.g.*, AR + OR improves AR individual recognition on five events, and KER + OR improves KER recognition on three events, see Table 4).

Regarding the performance of SEF compared to baseline methods, results demonstrate that SEF outperforms all of them in the test set of the 10-fold cross-validation scheme. This performance superiority is due to the framework capability of handling incomplete, ambiguous and noisy observations from heterogeneous concept detectors. The higher performance of the proposed framework compared to its individual detectors demonstrates its capability of exploring the complementary aspects of the detectors.

OSF baseline presents a performance close to the proposed approach on activities like “read”, “prepare pill box”, and “prepare drink”, and outperforms SVM baseline on the last two events. Its higher performance compared to SVM baseline is due to the existence of conceptual information from all detectors for the events in question. For instance, this behavior is not observed for “manage finances” and “watch tv“ events. “Manage finances” event has only concepts from AR and OR detectors, since this event happens most of the time outside of the field of view of KER sensor. Results demonstrate the lack of ability of OSF baseline in handling partial evidence. Similarly, the decrease of this baseline performance is observed for “watch tv“, and since this event is also undertaken at the border of the field of view of the color-depth sensor, the KER detector generates noisy observations in certain situations, which compromises OSF performance due to its lack of uncertainty handling.

SVM baseline gives better results than OSF for the events “read” and “talk on the telephone”, “search bus line”, “manage finances”, and “prepare pill box”. This superiority highlights this baseline’s capability of implicitly learn how to handle incomplete evidence, but still with less accuracy than the proposed approach. Both baselines underperform for brief activities, like “water plant”. For OSF this performance is attributed to noise and low reliability of the AR detector for the event in question. For SVM baseline, the low performance is mostly due to the reliance on a sliding time window, which provides less information for short events, compared to that obtained for event of longer duration.

From the described observations, we conclude the semantic fusion framework handles uncertain and incomplete evidence more accurately than baseline methods, especially when there is a disparity of reliability across intermediate detectors. It also goes beyond noise filtering, since it combines evidence from different sources in a complementary and semantically meaningful way.

8 Conclusion

This paper introduced a framework for semantic event fusion, composed of a novel probabilistic, knowledge-driven framework for event representation and recognition, and a novel algorithm for the semantic alignment of non-synchronized heterogeneous concept streams.

The knowledge-driven framework decomposes complex events into concepts, separating raw sensor data from event semantics modeling. Its main novelty lies in the combination of an ontological language for event modeling with a probabilistic inference method for uncertainty handling. This combination fosters more flexible event modeling than graphical model representations. At the same time it results in more reliable management of uncertainty than existing knowledge-driven methods.

The semantic alignment algorithm uses concept similarity across visual concept detectors as a surrogate for inaccurate temporal information. This method overcomes the limitation of state of the art approaches that require at least coarse time-synchronization among sensors and rely on a sliding time window for concept fusion.

As the extensive evaluation of our framework illustrates, the combination of these two contributions achieves a higher fusion performance in the presence of partial, complementary and uncertain information compared to baseline methods that uses supervised learning. Our method also delimits the temporal boundaries of activities more accurately than an ontology-driven approach over the entire set of observed concepts.

Future work will investigate ways to improve the performance of the semantic alignment algorithm on concept streams which contain regions featuring a high variance of concepts, and to adapt it to on-line scenarios, where not all concept stream information is available at once. Finally, it will also explore the dynamic estimation of concept reliability, *e.g.*, in response to observed changes on scene characteristics.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 288199 / DEM@CARE - Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support.

References

[1] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.

[2] A. Artikis, M. Sergot, and G. Paliouras. An event calculus for event recognition. *Knowledge and Data Engineering, IEEE Transactions on*, 27(4):895–908, April 2015.

[3] Konstantinos Avgerinakis, Alexia Briassouli, and Yiannis Kompatsiaris. Activity detection using sequential statistical boundary detection (ssbd). *Computer Vision and Image Understanding*, pages –, 2015.

[4] Tanvi Banerjee, James M. Keller, Mihail Popescu, and Marjorie Skubic. Recognizing complex instrumental activities of daily living using scene information and fuzzy logic. *Computer Vision and Image Understanding*, 140:68 – 82, 2015.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[6] Hugo Boujut, Jenny Benois-Pineau, and Remi Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In *ECCV 2012 - Workshops*, ECCV’12, pages 436–445, 2012.

[7] W. Brendel, A. Fern, and S. Todorovic. Probabilistic event logic for interval-based event recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3329–3336, June 2011.

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

[9] Duc Phu Chau, Franois Bremond, and Monique Thonnat. A multi-feature tracking algorithm enabling adaptation to context, 2011.

[10] Liming Chen, C. Nugent, and G. Okeyo. An ontology-based hybrid approach to activity modeling for smart homes. *Human-Machine Systems, IEEE Transactions on*, 44(1):92–105, Feb 2014.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[12] C.F. Crispim-Junior, V. Bathrinarayanan, B. Fosty, A. Konig, R. Romdhane, M. Thonnat, and F. Bremond. Evaluation of a monitoring system for event recognition of older people. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 165–170, Aug 2013.

[13] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[14] A. Fleury, N. Noury, and M. Vacher. 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*, pages 322 – 329, 2010.

[15] Marshal F. Folstein, Lee N. Robins, and John E. Helzer. The mini-mental state examination. *Archives of General Psychiatry*, 40(7):812, 1983.

[16] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. OWL 2: The Next Step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309–322, October 2008.

[17] I-H. Jhuo, G. Ye, S. Gao, D. Liu, Y.-G. Jiang, D. T. Lee, and S.-F. Chang. Discovering joint audio-visual codewords for video event detection. *Machine Vision and Applications*, 25(1):33–47, 2014.

[18] A. Klaser, M. Marszalek, C. Schmid, and A. Zisserman. Human focused action localization in video. *Proceedings of the 11th European Conference on Trends and Topics in Computer Vision - Volume Part I*, pages 219–233, 2012.

[19] Narayanan C. Krishnan and Diane J. Cook. Activity recognition on streaming sensor data. *Pervasive Mob. Comput.*, 10:138–154, February 2014.

[20] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[21] Li-Jia Li, Hao Su, Eric P. Xing, and Li Fei-fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, 2010.

[22] Guan Luo, Shuang Yang, Guodong Tian, Chunfeng Yuan, Weiming Hu, and S.J. Maybank. Learning human actions by combining global dynamics and local appearance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2466–2482, Dec 2014.

[23] Bayard E. Lyons, Daniel Austin, Adriana Seelye, Johanna Petersen, Jon Yeargers, Thomas Riley, Nicole Sharma, Nora C. Mattek, Hiroko Dodge, Katherine Wild, and Jeffrey A. Kaye. pervasive computing technologies to continuously assess alzheimers disease progression and intervention efficacy. *frontiers in aging neuroscience*, 7(102), 2015.

[24] Georgios Meditskos, Efstratios Kontopoulos, and Ioannis Kompatsiaris. Knowledge-driven activity recognition and segmentation using context connections. In *13th International Semantic Web Conference (ISWC'14)*, pages 260–275, 2014.

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

[26] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[27] Gregory K. Myers, Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ram Nevatia, Chen Sun, Amirhossein Habibian, Dennis C. Koelma, Koen E. A. van de Sande, Arnold W. M. Smeulders, and Cees G. M. Snoek. Evaluating multimedia features and fusion for example-based event detection. *Machine Vision and Applications*, 25(1):17–32, January 2014.

[28] A. T. Nghiem and F. Bremond. Background subtraction in people detection framework for rgb-d cameras. In *Proceedings of 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2014.

[29] Sangmin Oh, Scott McCloskey, Ilseo Kim, Arash Vahdat, Kevin J. Cannons, Hossein Hajimirsadeghi, Greg Mori, A. G. Amitha Perera, Megha Pandey, and Jason J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, 25(1):49–69, January 2014.

[30] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[31] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, October 2007.

[32] Juan C. SanMiguel and Jos M. Martnez. A semantic-based probabilistic approach for real-time video event recognition. *Computer Vision and Image Understanding*, 116(9):937 – 952, 2012.

[33] V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Generalized RBF feature maps for efficient detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

[34] T. Vu, F. Bremond, and M. Thonnat. Automatic video interpretation: A novel algorithm for temporal scenario recognition. In *The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003.

[35] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, June 2011.

[36] Xiaoyang Wang and Qiang Ji. A hierarchical context model for event recognition in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2561–2568, June 2014.

[37] Yan Yan, E. Ricci, Gaowen Liu, and N. Sebe. Egocentric daily activity recognition via multitask clustering. *Image Processing, IEEE Transactions on*, 24(10):2984–2995, Oct 2015.

[38] Yingying Zhu, N.M. Nayak, and A.K. Roy-Chowdhury. Context-aware activity modeling using hierarchical conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(7):1360–1372, July 2015.

Carlos F. Crispim-Junior received the degree of Bachelor in Computer Science from Vale do Itajaí University in 2006, and the Doctor degree in Electrical Engineering from Federal University of Santa Catarina in 2011. Since November 2011 he is working as a post-doctoral fellow at INRIA Sophia Antipolis in France. He is author/co-author of 3 book chapters, 4 papers in international journals (JAD, IRBM, CompBioMed, Gerontechnology), 11 papers in international conferences and workshops, 7 papers and 11 abstracts in Brazilian conferences and workshops, and 1 registered, free software for the video tracking and behavioral recognition of laboratory animals. He received the best paper award in the ISG*ISARC2012 conference.

Konstantinos Avgerinakis received his Diploma degree in computer and telecommunication engineering from the University of Thessaly in 2009, and the Phd degree in Electrical Engineering from University of Surrey in 2015. Since 2009 he has been a research assistant and PhD candidate at the Information Technologies Institute at Centre for Research and Technology Hellas - Information Technologies Institute (CERTH-ITI). His current research interests include computer vision and statistical video processing for event detection and recognition, human activity and facial expression recognition. He is the author of 13 conference publications, while he has also served as a reviewer for a number of international journals and conferences. He is Member of IEEE Technical Chamber of Greece.

Vincent Buso is a computer science PHD student at LaBRI (Laboratoire Bordelais de Recherche en Informatique) since October 2012. He graduated and got his master of science in Electrical Engineering from IIT (Illinois Institute of Technology) as well as an engineering diploma (master's degree equivalent) from ENSEIRB-MATMECA (École nationale Supérieure d'Électronique, Informatique, Télécommunications, Mathématique et Mécanique de Bordeaux). His fields of research are computer vision and image/video processing.

Georgios Meditskos received his PhD degree in Informatics from Aristotle University of Thessaloniki for his dissertation on “Semantic Web Service Discovery and Ontology Reasoning using Entailment Rules”. He also holds an MSc and a BSc degree from the same department. Since January 2012 he is working as a postdoctoral research fellow at the Information Technologies Institute (ITI) of the Center for Research and Technology Hellas (CERTH). His research interests include Knowledge Representation and Reasoning in the Semantic Web (RDF/OWL, rule-based ontology reasoning, combination of rules and ontologies) and Semantic Web Services (discovery, composition). Recently, his research interests have focused on context-aware multi-sensor reasoning and fusion in Pervasive Environments and the development of semantic interpretation frameworks for the high-level integration, analysis and preservation of contextual information.

Dr. Alexia Briassouli Dr. Alexia Briassouli (F) received the Diploma degree in electronic engineering from the National Technical University of Athens and the Ph.D. degree from the Department of Electrical and Computer Engineering at the University of Illinois in Urbana Champaign. She is currently working as a postdoctoral research fellow at CERTH. Her current research interests include statistical image and video processing, unusual or interesting event detection and activity recognition. She has authored over 50 publications in peer-reviewed journals and conferences and has participated in a number of European and National projects.

Jenny Benois-Pineau is a professor of Computer science at the University Bordeaux and chair of Video Analysis and Indexing research group in Image and Sound Department of LABRI UMR 58000 Universit Bordeaux/CNRS/IPB-ENSEIRB. She is also a deputy scientific director of theme B of French national research unity GDR CNRS ISIS, and Chair of International relations at College of Sciences et Technologies at University Bordeaux. She obtained her PhD degree in Signals and Systems in Moscow and her Habilitation Diriger la Recherche in Computer Science and Image Processing from University of Nantes France. Her topics of interest include image and video analysis and indexing, motion analysis and content description for content-based multimodal retrieval. She is the author and co-author of more than 110 papers in international journals, conference proceedings, book chapters. She is associated editor of EURASIP Signal Processing: Image Communication, Elsevier, Multimedia Tools and applications, Springer. She has served in numerous program committees in international conferences and workshops, *e.g.*, ACM MM and CIVR; and she has served as expert for European Commission since FP4.

Dr. Ioannis (Yiannis) Kompatsiaris is a Senior Researcher (Researcher A) with the Information Technologies Institute / Centre for Research and Technology Hellas, Thessaloniki, Greece. His research interests include semantic multimedia analysis, indexing and retrieval, social media and big data analysis, knowledge structures, reasoning and personalization for multimedia applications, eHealth and environmental applications. He received his Ph.D. degree in 3-D model based image sequence coding from the Aristotle University of Thessaloniki in 2001. He is the co-author of 69 papers in refereed journals, 35 book chapters, 8 patents and more than 240 papers in international conferences. He has been the co-organizer of various international conferences and workshops and has served as a regular reviewer for a number of journals and conferences. He is a Senior Member of IEEE and member of ACM.

François Brémond is a Reseach Director at INRIA Sophia Antipolis. He has conducted research work in video understanding since 1993 at Sophia-Antipolis. In 1997 he obtained his PhD degree from INRIA in video understanding and he pursued his research work as a post doctorate at USC (University of Southern California) on the interpretation of videos taken from UAV (Unmanned Airborne Vehicle). In 2007 he obtained his HDR degree (Habilitation à Diriger des Recherches) from Nice University on Scene Understanding. He created the STARS team on the 1st of January 2012. François Brémond is author or co-author of more than 140 scientific papers published in international journals or conferences in video understanding. He is a handling editor for MVA and a reviewer for several international journals (CVIU, IJPRAI, IJHCS, PAMI, AIJ, Eurasip JASP,) and conferences (CVPR, ICCV, AVSS, VS, ICVS,). He has (co-)supervised 13 PhD theses. He is an EC INFSO and French ANR Expert for reviewing projects.

Editor Comments

Associate Editor

Comments to the Author:

Reviewers of the paper have been received. One reviewer still points out some major concerns. However I am happy with the improved version of the manuscript. I agree that some of the reviewers comments have to be addressed before publication. Please note that although you have a minor revision you should carefully address all reviewers' comments.

We would like to thank you all again for the careful evaluation of our revised contribution for the special issue "Multimodal Human Pose Recover and Behavior Analysis" of the IEEE Transactions on Pattern Analysis and Machine Intelligence. You may find below our answers to the remaining questions of reviewers, which are addressed by the latest version of our paper.

Reviewer Comments

Reviewer: 1

Recommendation: Author Should Prepare A Major Revision For A Second Review

Comments:

Q1-a) "Regarding the effectiveness of the OR module using a wearable camera … What is not clear to me is that, if the design in the OR module, namely a BoW representation of SURF descriptors with a saliency mask and a SVM classifier, is sufficient to handle a seemingly difficult 18-class object recognition task. I was interested in the recognition accuracy of these 18 object classes, which is the direct output of the classifier."

Fig. A illustrates the F_1-score of 12 classes of OR which are the most relevant for the targeted activities of daily living. We observe that OR method has an average F_1-score performance of 56 % in 9/12 classes that appear in the test set recordings, and 43 % when considering all of them. The overall precision of OR for all classes is 85.77 %. This difference between F_1-score and precision denotes trustworthy observations, but places OR in a more complementary role than a standalone detector for activity recognition.

We have added Fig. A to the paper as Fig. 8 (see page 11) and the comments above are added to page 11, L895-901.

.

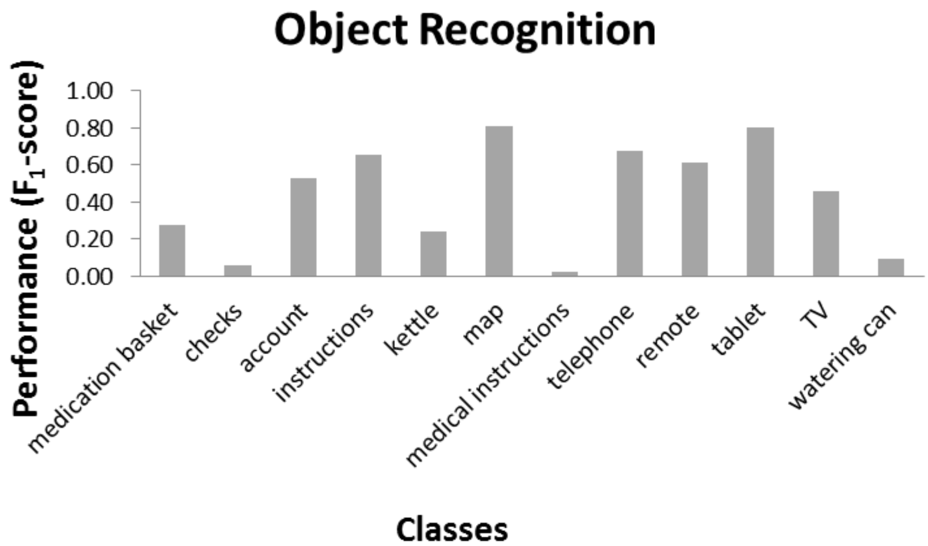

Figure A. Performance of OR concept detector per object class.

Q1-b) *It is possible that even if the recognition rate of the objects is not very high, the OR module is still helpful as its errors might be corrected by the results of the other two concept detectors, and the combination of the three sensors may lead to better results. ... But it would be more convincing to experimentally demonstrate that the performance using all three concept detectors is much better than the one when just combining KER and AR.*

As suggested by the reviewer, we have provided activity recognition results of SEF framework using as input pair-wise combinations of concept detectors (Table 1), and we compare them to the results of SEF framework using all detectors at once.

The combination of OR with other detectors in a pair-wise fashion shows its meaningful and complementary role. For instance, it improves AR recognition for 5 events, and KER recognition for 3 events. But more importantly, it is the combination of all concept detectors (AR+KER+OR) that has the highest average F_1-score (higher value in 6/8 activities).

.

Table 1. Performance of information fusion given different concept detectors

Events	Pairwise AR+OR	Pairwise KER+OR	Pairwise KER+AR	All
Search bus line	43.82	0.00	43.64	73.11
Manage finances	43.68	35.99	43.68	43.73
Prepare pill box	58.75	55.84	60.31	65.02
Prepare drink	27.59	63.40	54.21	64.04
Read	68.86	97.60	93.94	95.22
Talk on telephone	54.17	74.18	92.48	75.58
Watch TV	8.42	8.89	20.68	35.80
Water Plant	0.00	50.00	98.97	100.00

The top performer pairwise combinations are highlighted in yellow. The combination of all detectors is highlighted in blue when it outperforms the pairwise combinations.

These answers are added to the paper at page 11 lines 937-951.

Q2. It is not mentioned in the paper how the performance will change in a new environment after the retraining of some concept detectors and the adjustment of the reliability model during fusion. It is a little concerning that the performance for many activity classes on the test set and the validation set are quite different (above 15% gap in 6 out of 8 classes), even in the same room configuration. It is mentioned that the system has been deployed in three different locations. Did you get feedback from the customers regarding the performance of the system?

Currently, for every new environment where we install the system we first check if the performances of pre-trained concept detectors degrade (*e.g.*, AR and OR). If it degrades, we add video samples from the new scene into the previous training set of the detector and retrain it. Although we do not have a quantitative evaluation of detector performances in other environments, our observations have shown activity recognition performances in the range of test set results, if not higher (Table 3). We retrain concept detectors due to the large intra-class variance of daily living actions and objects, *e.g.,* the appearance and shape of a kettle may vary considerably in different real-world environments.

As we progress with the deployment of the system, we hope to acquire a sufficient amount of training data to overcome the need of retraining concept detectors. It should be emphasized that, for one of the major contributions of this work, the Semantic Event Fusion framework, nearly no changes are necessary when we deploy it to newer environments.

Q3. Minor issue: in Table 4, it is stated in the title that the comparison is on the test set, while the results of the proposed method are the same with the ones in the validation set.

By mistake we have included results from the validation set in the former version of Table 4. We have fixed it in the new revision of the paper and now it only contains results from the test set. We thank the reviewer again and apologize for any inconvenience.

.

Reviewer: 2

Comments:

Q1: "… The argument that the authors have provided for including a Kinect is that it can still capture the depth even if there is not light in the scene, and the argument for including a fixed RGB camera is that it can provide a better field of view compared to the Kinect that is already in the setup for extracting the depth and already provides RGB. "

> The novelty of this paper in terms of activity sensing refers to the variety (or heterogeneity) of visual concept modalities in use, i.e., the activity patterns and points of view we have used to recognize the activities of daily living, and less on optimization given a specific choice of sensors.
>
> For instance, we model the global displacement patterns of a person in the scene, his/her local and finer motion patterns, and the types of objects being handled during an activity. You can find a quantitative analysis of the benefit of employing these three types of concept detectors at Table 2 in page 3. Briefly, it is the combination of these patterns that permits a real-world, semantically rich description of activities of daily living, both in small and large rooms, which is robust and reliable enough to be deployed in practice.
>
> This paper focuses on studying the benefits of each concept detector to the overall performance of the Semantic Event Fusion framework. But, the final decision about which sensors to use will remain to the user, who should consider the combination of sensors that provides the best trade-off between scene coverage, system setup complexity and solution cost.
>
> Finally, the ever expanding proliferation of wearables and other ambient sensors will make such multimodal monitoring schemes very common in the future, so we consider our work very timely in this respect.

Q1.1 First of all, I don't agree with the first argument, because if the there is no light in the scene the other two cameras will not work anyway, so it is almost of no value if you still can extract the depth from the Kinect camera.

> We agree with the reviewer, our choice of sensors is mostly beneficial for daytime monitoring of activities of daily living, since only Kinect sensor works effectively at nighttime. Regarding Kinect, we have chosen this sensor due to its real-time, off-the-shelf 3D measurements of the scene and its objects, as stated in page 4, lines 47-60. The 3D measurements of this sensor improve the quality of people detection and tracking algorithms by resolving visual ambiguities with depth information, and make these algorithms invariant to light changes that occur during daytime.
>
> Currently, we have no setting that can provide multimodal data to SEF framework during nighttime period. Additional experiments have taken place beyond the scope of this paper (after its submission) in home environments (see Figure B), where there is only one camera, a color-infrared led camera. This camera provides RGB video in a lit environment during the day and infrared grayscale visual information in the absence of light. The resulting grayscale

.

images/videos are not as descriptive as the combined color-depth features, however they are still a useful source of information about the scene, which can be fused with other sensor measurements. Our initial experiments have shown that this fusion leads to satisfactory activity recognition accuracy even during nighttime. On a different site, composed of studio apartments in a nursing home, we have only been using the depth map of Kinect cameras to monitor people, due to privacy concerns. The data they provide, as in the case of the home environments, are still useful, and can lead to accurate activity recognition when fused with the other sensor data. Future work will investigate the findings of ongoing experiments on nighttime monitoring to extend SEF activity recognition for this period of the day.

Figure B. Example of night-time event monitoring with color-infrared camera

Q1.2 Regarding the second argument, there is not much discussion in the paper, except few places repeating the same argument that due to the better field of view, the authors have preferred the fixed RGB camera to the RGB data that they get from first version of the Kinect camera.

Does it mean that a newer version of Kinect can solve the problem? How much difference exactly are we talking about?

The important factor here is the coverage of the scene by the fixed sensor (camera). Indeed, we could have used the RGB image from Kinect 1 instead of the color camera, but the field of view of Kinect 1 does not cover the entire observation room in use.

Since the newer version of Kinect has a broader angle of field of view, we could use only Kinect 2 for AR and KER modules for new recordings in our observation room, however this was not tested in our experiments since the newer sensor was not yet available at the beginning of the clinical trials.

Q1.3 What is the effect of this much difference on the overall performance exactly in terms of the performance of the system?

We should not expect any difference in performance if we use the RGB signal of Kinect instead of a regular RGB camera to feed concept detectors, like AR. One can use Kinect sensor as input for both AR and KER detectors when this sensor covers the entire room to monitor. This is the case for some of our ongoing experiments in smaller rooms.

.

Q1.4 Why not to use a stereo setup instead of Kinect and the fixed RGB? A stereo setup can possibly provide a very good field of view.

> Yes, a set of stereo-cameras could also be put in place as suggested by the reviewer, however it would be more expensive and time consuming to set up than a Kinect. Some stereo-cameras could be an affordable solution such as Intel Real-sense, but this camera has a shorter field of view. The use of a Kinect sensor provides a good trade-off between cost and complexity to setup.

We have added the above answers to page 3, lines 234-253 of the paper.

.