Online recognition of daily activities by color-depth sensing and knowledge models

Carlos F Crispim-Junior, Alvaro Gómez Uría, Carola Strumia, Michal Koperski, Alexandra Konig, Farhood Negin, Serhan Cosar, Anh-Tuan Nghiem, Guillaume Charpiat, Francois Bremond, et al.

MDPI

Article

Online recognition of daily activities by color-depth sensing and knowledge models

Carlos Fernando Crispim-Junior [1,2]*, Alvaro Gómez Uría [1], Carola Strumia[1], Michal Koperski[1], Alexandra Konig[2,3], Farhood Negin[1], Serhan Cosar[1], Anh Tuan Nghiem[1], Duc Phu Chau[1], Guillaume Charpiat[1] and Francois Bremond [1,2]*

[1] INRIA Sophia Antipolis, 2004 route des Lucioles - BP 93 06902 Sophia Antipolis
[2] CobTek - Cognition Behaviour Technology - Université Nice Sophia Antipolis
[3] MUMC - School for Mental Health and Neuroscience - Alzheimer Center Limburg - Maastricht University
* Correspondence: carlos-fernando.crispim_junior@inria.fr; Tel.: +33-489-73-24-45

Academic Editor: name
Version April 29, 2017 submitted to Sensors; Typeset by LaTeX using class file mdpi.cls

Abstract: Visual activity recognition plays a fundamental role in several research fields as a way to extract semantic meaning of images and videos. Prior work has mostly focused on classification tasks, where a label is given for a video clip. However, real life scenarios require a method to browse a continuous video flow, automatically identify relevant temporal segments and classify them accordingly to target activities. This paper proposes a knowledge-driven event recognition framework to address this problem. The novelty of the method lies in the combination of a constraint-based ontology language for event modeling with robust algorithms to detect, track and re-identify people using color-depth sensing (Kinect sensor). This combination enables to model and recognize longer and more complex events and to incorporate domain knowledge and 3D information into the same models. Moreover, the ontology-driven approach enables human understanding of system decisions and facilitates knowledge transfer across different scenes. The proposed framework is evaluated with real-world recordings of seniors carrying out unscripted, daily activities at hospital observation rooms and nursing homes. Results demonstrated that the proposed framework outperforms state-of-the-art methods in a variety of activities and datasets, and it is robust to variable and low-frame rate recordings. Further work will investigate how to extend the proposed framework with uncertainty management techniques to handle strong occlusion and ambiguous semantics, and how to exploit it to further support medicine on the timely diagnosis of cognitive disorders, such as Alzheimer's disease.

Keywords: Activity recognition; activities of daily living; assisted living; color-depth sensing; complex events; people detection and tracking; knowledge representation; senior monitoring

1. Introduction

Research on technologies for assisted living has been growing on demand due to the aging of world population and the increasing number of elderly people living alone. The task of automatic recognition of daily living activities plays a fundamental role in this scenario, since it may provide doctors with a deeper glimpse of people's daily routine. However, this task is a challenging problem, far from being solved due to the unconstrained nature of real-life scenes, and the large intra-class variance of human activities (*e.g.*, each person may have their own way of preparing coffee). The recognition of human activities has been explored from different sensor perspectives over the years, *e.g.*, from ambient- [1,2] to visual-sensing [3,4], up to their combination [5]. Ambient sensing tends to equip the scene with several low-level sensors (*e.g.*, microphones, presence and door contact sensors) and to monitor people activities by their interaction with (or disturbances in) the sensor network [1,2].

Although ambient sensing by low-level sensors has its advantages, like preserving people privacy, it may undermine the recognition and detailed description of complex activities since complex events may become a function of relatively simpler sensor states (*e.g.*, kettle turned on, moved cup). As an alternative for low-level sensors, visual sensing focuses on the direct observation of people during the realization of activities [3,4,6], which fosters more detailed representations of activities. However, noise due to scene illumination changes and the estimation of 3D information from 2D data may degenerate the quality of vision systems using 2D video cameras and consequently degrade their performance.

This paper proposes a fully-working framework for event recognition based on color-depth sensing and ontological reasoning (Fig.1). It follows a person-centered pipeline (event recognition from people detection and tracking) to discriminate among the activities of different people and it explores the geometry of semantic zones to improve people detection and event recognition. The paper also extends the video event ontology language proposed by Vu *et al.* [7] from video surveillance to assisted living scenarios. Finally, it proposes an algorithm to improve people detection by coupling it with information about scene geometry (ground-plane estimation using semantic zones). The rest of the paper is structured as follows: Section 2 presents related work, Section 3 describes the proposed approach, Section 4 describes the experiments carried out, Sections 5 & 6 present the obtained results and discussion and Section 7 presents our conclusions.

2. Related work

Knowledge-driven methods, like first-order logic and description-based models, provide a formalism to systematically describe domain knowledge about real-world phenomena using rules or constraints. Constraints provide a generic basis to combine different sources of knowledge [8,9]. They can be handcrafted by domain experts [3,4,8], learned from data or obtained by a combination of both forms [9]. Knowledge-driven methods are generally associated to an ontological formalism to define domain concepts and their interrelations [10] [8]. Town [10] has introduced an ontological formalism for knowledge management and reasoning over raw visual data for video surveillance applications. Ceusters *et al.* [11] have proposed Ontological Realism to incorporate semantic knowledge into the recognition of high-level events using a video-analysis system supported by a human in the loop. Cao *et al.* [8] has used a rule-based engine to combine different sensing contexts (human and ambient) to monitor the daily activities of seniors. Human context (*e.g.*, postures like sitting, standing, walking) comes from video camera data, while ambient context comes from inertial sensors attached to objects of daily usage and home appliances (*e.g.*, TV remote control, and doors). In another direction, Chen *et al.* [12] have introduced a framework that combines ontology formalism for activity modeling with data-driven methods for model parameters update over time. Despite their representation power, knowledge-driven approaches are sensitive to noise due to their deterministic mechanism of reasoning. Therefore, these methods require that either their underlying modules for scene observation handle the sources of noise that intervene in the data [13][4] or that their reasoning mechanism is adapted to cope with noisy data at event level [14][15] [16][17].

This paper focuses on the conception of a framework that associates a color-depth sensing pipeline for people detection and tracking with an ontology-driven mechanism of reasoning. Prior work on knowledge-driven methods and color-depth sensing (*e.g.*, Asus Xtion PRO Live) have demonstrated the benefits of this sensing approach (3D information about the scene and invariance to illumination changes) to track the position of hands and facial features during psychomotor exercises (cognitive rehabilitation) [6], to recognize fall events in hospital rooms [13], and to recognize complex daily living activities of senior people (*e.g.*, making the bed). Finally, Crispim-Junior *et al.* [3] compared the performance of event recognition between two different vision pipelines: a standard, color video camera and a color-depth sensor (Kinect with PrimeSense library). They have demonstrated that a pipeline with a standard color camera demanded a finer parameter tuning to handle low-level noise and achieve a performance comparable to the color-depth one. But, although

the latter pipeline was less parameter-dependent, it could not detect people farther than 4 meters, a limitation that may undermine its applicability in real-world scenarios.

The proposed event recognition framework differs from prior work on the vision pipeline adopted (people detection and tracking approaches) and the formalism for event recognition. Firstly, color-depth sensing tends to be more robust to noise due to scene illumination changes than conventional video cameras, and it allow the subsequent modules to solve 2D ambiguities by using 3D measurements of the scene. Moreover, the proposed vision pipeline employs an algorithm for people detection in color-depth sensing that extends the range of people detection from 3-4m (*e.g.*, Microsoft and Primesense libraries) to 7-9 m, by handling noise at depth pixel-level. Finally, the ontology-driven mechanism of reasoning allows to incorporate different sources of information efficiently, from common sense knowledge and event semantics up to dynamic information about visual entities. The combination of both modules enables to model more complex and longer time-dependencies among events, barely explored before on online activity recognition.

3. Knowledge-driven event recognition

The proposed framework is divided into the following modules (Fig.1): ground-plane estimation (1), people detection (2), people tracking (3) and ontology-driven event recognition (4). Ground-plane estimation constructs a 3D estimation of the floor plane. People detection localizes people in every video frame. People tracking consists in finding appearance correspondences between people detected in the current and previous frames. Finally, event recognition combines the information of prior steps to infer which activities a person is performing. All steps follow an online fashion to address the task of continuous activity recognition in assisted living scenarios.

Next subsections describe the procedures employed to detect and track people in the monitored scene (ontology's physical objects, sub-sections 3.1, 3.2 and 3.3;) and how to model and recognize complex activities of daily living using the ontology-driven approach (Sub-section 3.4).

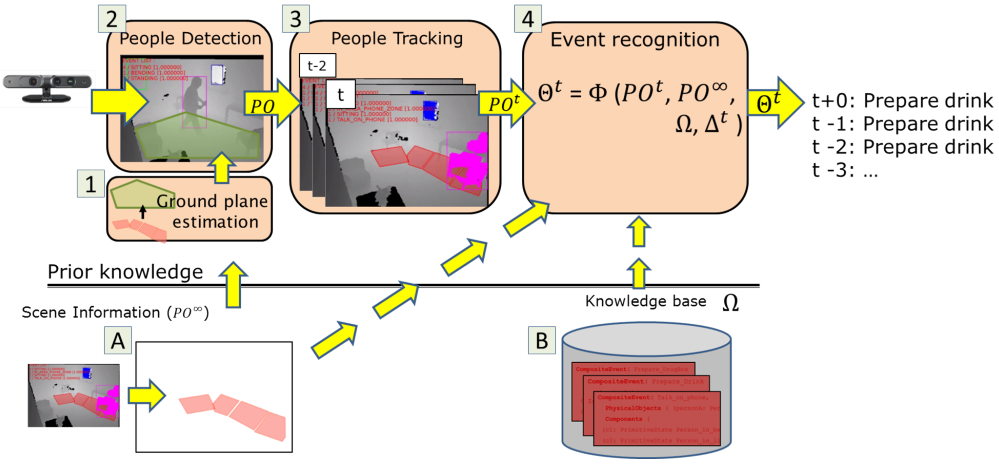

Figure 1. Knowledge-driven framework for visual event recognition. Firstly, 0) an estimation of the ground-plane is computed using the vertexes of semantic zones. 1) Video frame acquisition is performed using a color-depth sensor. Then, 2) people detection module analyzes the video frame for instances of physical objects of type person. For each instance found, it adjusts its height using ground-plane information. 3) Tracking step analyzes the set of detected people in the current and previous frames for appearance matching and trajectory estimation. 4) Event recognition takes as input the information from all previous steps and evaluates which event models in its knowledge base are satisfied. Recognized events are added to each person' history and the steps 2-4 are repeated for the next frame. Prior knowledge about the problem corresponds to semantic information about scene geometry and the events B) knowledge base.

3.1. Ground plane estimation

The estimation of a ground-plane is a key step for the vision pipeline, since its output is employed to improve the performance of the subsequent steps of people detection and tracking. The estimation process is made as follows: firstly, we search locally for pieces of planes, using the 3D-vertexes of the semantic zones. For each 3D vertex, we consider the cloud of points formed by its nearest neighbors and find the best plane which approximates it in the least square error sense (closed-form solution). When the approximation error is too high, *i.e.*, when the local cloud of points is not flat enough, the plane is discarded. The obtained planes are clustered into larger planes weighted by the number of 3D points they possess. We compare any two pieces of planes during the clustering step based on the angle between their Normals and on their alignment (distance between each center of mass and the other plane). We sort the newly obtained planes by their confidence (approximation error and number of points involved) and assign the first nearly horizontal plane to the ground plane.

3.2. People detection

The people detection step is performed by the depth-based algorithm proposed by [18]. The algorithm performs as follows: first, background subtraction is employed on the depth image provided by the color-depth sensor to identify foreground regions containing both moving objects and potential noise. Foreground pixels are clustered into objects based on their depth value and their neighborhood information. Among these objects, people are detected using head and shoulder detectors. After this step, noise is removed using information from people detection and tracking from previous frames. At last, the background model of the background subtraction algorithm is updated using current step results. Given that the raw depth-signal may be affected by the way some materials reflect infrared beams, like some clothing materials [19]; we re-estimate people's height by computing the Euclidean distance between the highest point in their silhouette's (3D cloud of points) and the estimated ground plane (Subsection 3.1). This procedure is needed since lower-limbs tend to be often missed due to either noise in depth measurement or to occlusion of the limbs by scene furniture (*e.g.*, desk). We have opted for a custom algorithm for people detection since the ones offered by the libraries of Microsoft and PrimeSense cannot detect people farther than 3-4 meters away from the sensor. With our own algorithm we extend people detection to 7-9 meters away, which is a more realistic distance for ambient assisted living scenarios. Finally, the people detection of background subtraction method does not make any assumption on people posture, and hence it can detect people in more unconstrained scenarios than the skeleton-based algorithm provided by Kinect(R) standard SDKs [20].

3.3. People tracking

The tracking algorithm takes as input the video stream and a list of detected people in a temporal sliding window. First, a link score is computed between any two detected people appearing in the 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. Hypothesis trajectories are built from links with scores greater than a pre-defined threshold. The reliability of each hypothesis trajectory is represented by the total score of its link scores. The trajectory of the objects are determined by maximizing objects' trajectory reliability using the Hungarian algorithm [21]. Since the descriptor weights generally depend on the content of the video being processed, we use the control algorithm proposed by [22] to tune the weights on an online manner.

3.4. Ontology-driven event recognition

The proposed framework extends the declarative constraint-based ontology proposed by [7] with knowledge about activities of daily living, scene information and domain physical objects. The video event ontology language (Fig. 2) employs three main conceptual branches: physical objects,

events and constraints. The first branch, physical objects, consists in the formalization - at conceptual level - of the observations of the vision pipeline, *i.e.*, the people and objects in the scene. The remaining two branches - video events and constraints - provide the basis for event modeling, *i.e.*, the types of event models and the possible relations between physical objects and sub-events (namely components) that characterize a composite activity (or event).

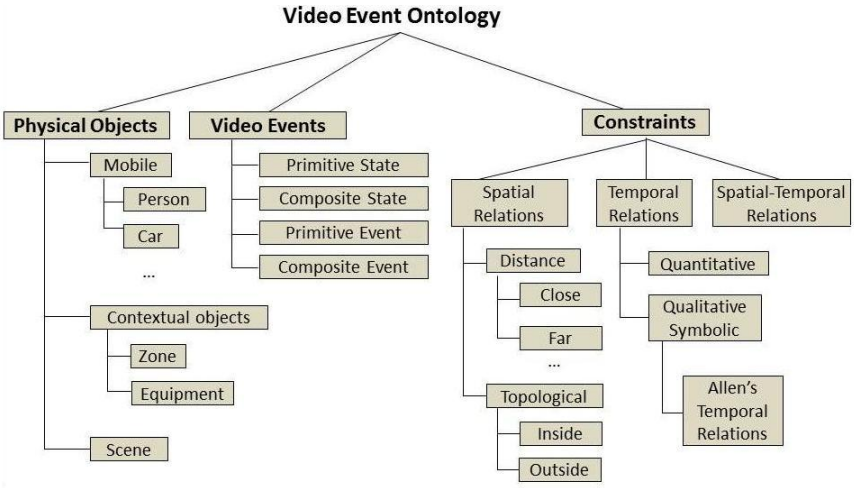

Figure 2. Video event ontology language. Three main concept branches are defined: physical objects, video events and constraints. Physical objects make abstractions for real-world objects. Video events describe the types of event templates available for activity modeling. Constraints describes the relations among physical objects and activities' components (sub-events).

Event models are defined by the triplet: physical objects, components (sub-events) and constraints; as described by Eq.1.

$$\omega_j = \langle PO_j, SE_j, CO_j \rangle \tag{1}$$

where,

- ω_j: event model j,
- PO_j: classes and number of physical objects involved in model j, where $PO_j = \{po_{j,1}, .., po_{j,m}\}$ and $m = |PO_j|$,
- SE_j: set of components of model j, where $SE_j = \{se_{j,1}, .., se_{j,k}\}$ and $k = |SE_j|$,
- CO_j: set of constraint of model j, where $CO_j = \{co_{j,1}, .., co_{j,l}\}$ and $l = |CO_j|$.

Physical object classes refer to abstractions of real-world objects that take part in the realization of target events. The possible types of physical objects depend on the domain for which the event modeling task is applied for. For assisted living settings, this paper defines five types of objects (Fig.1): mobile, person, contextual zone, contextual equipment and scene. Mobile is a generic class that contains the basic set of attributes for any moving object detected in the scene (*e.g.*, 3D position, width, height, length). It is represented as a 3D bounding box. Person is an extension of Mobile class whose attributes are "body posture", "speed" and "appearance signature". Scene class describes attributes of the monitored scene, like the number of people in the scene. Instances of mobile and person classes are provided to of event recognition step by underlying modules of the framework (Fig.1, steps 2 & 3). Physical objects which attributes evolve over time, like mobile and scene, are grouped together into the set $PO^t = \{po^{t,i}, .., po^{t,n}\}$.

Contextual object class corresponds to a 3D polygon of n-vertexes that describe a piece of semantic information about the scene. Zones and equipment extend contextual object class and refer to knowledge about the scene (*e.g.*, kitchen and couch zones or TV and table furniture, *etc*). They may be obtained automatically by algorithms for scene discovery or be provided based on human

knowledge. For instance, with the help of a software, one can easily define a 3D decomposition of the scene floor plane into a set of semantic regions, *i.e.*, spatial zones (*e.g.*, "TV", "armchair", "desk", "coffee machine"). In the context of this work, semantic zones are provided as prior knowledge about the scene (PO^{∞}) and their attributes are constant over time (non-temporal observations), since most semantic information about the target scenes refer to non-moving objects (*e.g.*, furniture). Figure 2 demonstrates how the proposed framework integrates 3D information about the scene (prior and dynamic) as instances of physical objects.

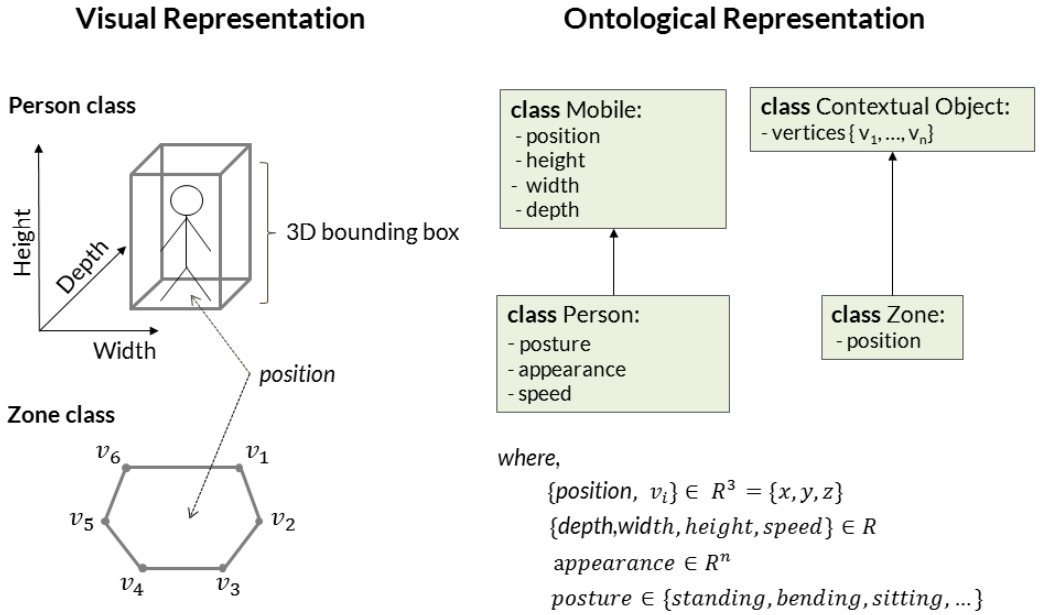

Figure 3. Physical objects integrate 3D visual information into the ontological events

Constraints are used to define conditions about attributes of physical object's or between the sub-events (components) of an event model. They are categorized into temporal and non-temporal constraints. Non-temporal constraints refers to conditions that do not directly depend on time, like spatial relations (*e.g.*, *in, close, out*) and posture values (*e.g.*, sitting, standing and bending). Temporal constraints refer to temporal relations between the time intervals of an event model's components. (*e.g.*, BEFORE, MEET and EQUAL) [23] or about their duration.

Event models are templates to describe relations between the elements of the event triplet (physical objects, components and constraints). The ontology language provides templates to support domain experts at modeling such relations. Templates are categorized according to the type of relations they model (in ascending order of complexity):

- **Primitive State** models the value of a attribute of a physical object (*e.g.*, person posture, or person inside a semantic zone) constant over a time interval.
- **Composite State** refers to a composition of two or more primitive states.
- **Primitive Event** models a change in the value of a physical object's property (*e.g.*, person changes from sitting to standing posture).
- **Composite Event** refers to a composition of two events of any type and it generally defines a temporal constraint about the time ordering between event components (sub-events).

Model 1 presents an example of composite event describing a temporal relations. The event model, "bed exit", is composed of three physical objects (a person and two semantics zones) and two components. The components of the event, c_1 and c_2, are, model respectively, "the person position lying on the bed" and "the person being outside of the bed" (out_of_bed). The abstraction p_1 corresponds to a person's instance dynamically detected by the underlying vision module.

Contextual zones z_B and z_{SB} are abstraction for the semantic zones "bed" and "side of the bed", which were *a priori* defined in the 3D coordinate system of the scene. The first constraint defines that the time interval of component s_1 must happen before the time interval of the component s_2, and in contrast to BEFORE relations, the relation MEET enforces that the boundaries of the time intervals must meet for a few frames. The second constraint defines a lower bound to the duration of the sub-event *out_of_bed*, 3 seconds. Parameter values, such as minimum duration of an event model instance, are computed based on event annotations provided by domain experts.

Model 1. *Composite Event bed exit*

```
CompositeEvent(BED_EXIT,
  PhysicalObjects((p1:Person),(zB:Zone),(zSB:Zone))
  Components(
    (s1: PrimitiveState in_zone_bed (p1,zB))
    (s2: PrimitiveState out_of_bed (p1,zSB)))
  Constraints((s1 meet s2) // c1
              (duration(s2) > 1)) //c2
  Alarm ((Level : URGENT))
)
```

Given that event models are defined at conceptual level (using the event ontology language), the underlying vision pipeline can be fine-tuned or replaced for a new scene without any changes to the models. The updated modules just need to keep providing the same type of physical objects expected by the model. Moreover, different from data-driven methods that require one to retrain (all/the) the event classifier(s) once a new class or input feature is added, the ontological formalism allows one to make as many changes as necessary to a single event model without requiring to visit the definition of other models. In short, the proposed framework eases model addition and update, and by consequence, it fosters knowledge transfer between different scenes (or datasets) with minimal changes.

Event inference (recognition) is performed at every frame t of a video sequence (or on the basis of a continuous video acquisition) and it relies on the temporal algorithm for event reasoning proposed by [7]. In short, for each time step t, the inference algorithm Φ takes as input the instances of physical objects present in t (PO^t), prior knowledge about the scene instances of events recognized at prior time steps (Δ^t), and the knowledge base (Ω). The algorithm Φ adopts an iterative, hierarchical fashion to generate the list of recognized events (θ^t), it first checks for the satisfaction of time independent events (primitive states). Then it, searches for all primitive and composite events that can be satisfied by recognized instances of primitive states. Inference is repeated until no composite event can be induced from the recognized events. The knowledge base corresponds to event models defined by domain experts or learned provided data.

4. Experiments

This paper has evaluated the proposed framework on three datasets: CHUN, GAADRD and Nursing home. The first two datasets have compared the proposed framework to three variations of a state-of-art baseline for visual action recognition. In CHUN dataset, it has also evaluated whether the recognition performance of the framework generalizes over a larger set of patients. The third dataset, Nursing home, has evaluated the performance of the framework on an unconstrained scenario: the continuous monitoring of a senior in her apartment, a scenario where only depth recording data is available. Next sub-sections describe in more details the baseline approach and datasets.

4.1. Performance baselines

To compare the performance of the proposed approach with the state of the art, we have chosen the action recognition pipeline described in [24]. Support Vector Machines (SVM) for action classification trained with a bag-of-visual-word embedding over descriptors of dense trajectories features. In short, for each video sequence we have first extracted local spatio-temporal patches using dense trajectories' detector. Then we have cut patches around each trajectory point as described in [24]. For each patch, we compute standard descriptors: trajectory shape, HOG, HOF and MBH. Then we have used each of the latter three descriptors to create a bag-of-words (BoW) representation as embedding function (with $k = 4000$). Finally, support vector machines with RBF kernel are used to classify the video representation as one of the target classes. Classifiers are learned on a supervised manner using video segments clipped from the original video sequence using ground-truth data. For online testing, the descriptors of a video are extracted over a temporal sliding window of size W (frames) with step size T and a minimal number of features extracted denoted as M. For each sliding window step we have extracted descriptors and apply BoW with SVM classifier given the number of detected features is equal or above M. Hyperparameters W, T and M were, empirically, set to 40, 15 and 20; respectively. A hold-out validation scheme is employed for training and testing the baseline classifiers. Baseline approaches were: Dense trajectories (DT) with Histogram of Gradients descriptor (HOG), DT with Histogram of Optical Flow (HOF) and DT with the y-component of Motion Boundary Histogram (MBH^y). All results are reported on the tested set of the respective baselines.

4.2. CHUN dataset

Participants aged of 65 years and above were recruited by the Memory Centre of Nice Hospital to participate on a clinical study about Alzheimer's disease. The study protocol asks the participants to carry out a set of physical tasks and Instrumental Activities of Daily Living (IADL) in a Hospital observation room equipped with home appliances (Fig. 4) [25]. Experimental recordings used a color-depth camera (Kinect®, Microsoft©). The activities in the experimental protocol are divided into two scenarios: guided- and semi-guided activities. Guided activities (10 minutes) intend to assess kinematic parameters about the participant's gait (*e.g.*, walking 8 m). Semi-guided activities ($\sim$ 15 minutes) aim to evaluate the level of autonomy of the participant by organizing and carrying out a list of IADLs. Semantic spatial zones are provided as prior knowledge about the geometry of the scene (Fig.4, red polygons): tea, telephone, plant, pharmacy, reading, TV, walking, stop/turn and counting. This evaluation focuses on the recognition of the following IADLs:

- Prepare drink (P. Drink, *e.g.*, prepare tea/coffee);
- Prepare drug box (organize medication);
- Talk on the telephone (calling, answering);
- Read article;
- Search bus line and;
- Water the plant.

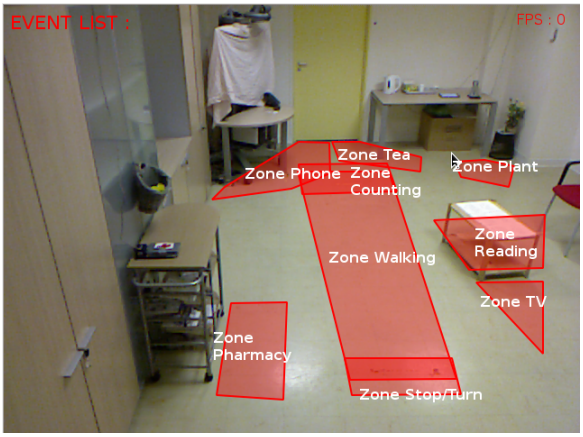

Figure 4. Contextual zones define geometric regions (red polygons, CHUN dataset) that carry semantic information about daily activities.

4.3. GAADRD dataset

Participants aged 65 years and above were recruited by a Greek Institute under the scope of a European project, called Dem@care, for the study of Alzheimer's disease [26]. This dataset contains recordings of seniors carrying out physical tasks and IADLs in an observation room with similar settings to those adopted in CHUN dataset. Experimental recordings have also adopted a color-depth sensor (here: Asus Xtion Pro Live®, ~10 frames per second). We have focused our evaluation in GAADRD subset called DS8 which contains recordings of 25 seniors. In this subset, participants are asked to perform the following IADLs:

- Establish account balance (M.Payment);
- Prepare drink (P. Drink, *e.g.*, prepare tea/coffee);
- Prepare drug box (P. Pill box);
- Read article;
- Talk on the telephone (T. Telephone, *e.g.*, calling);
- Turn radio on; and
- Water plant.

We highlight that there are a few differences between IADLs of these GAADRD and CHUN datasets. For instance, the "prepare pill box" activity in CHUN dataset consists in organizing a patient medication for a week, while in GAADRD dataset it corresponds to "taking the medicine". Moreover, GAADRD introduces the activity "Turn radio on". These differences have led to slightly different activity models between both datasets. Nevertheless, using the proposed ontological formalism we have swiftly adapted the event models' definition of CHUN dataset to GAADRD. From here on in the paper we will refer to GAADRD-DS8 subset as GAADRD dataset.

4.4. Nursing home dataset

This dataset consists of 72 hours of depth data recording about an 86 years old female, diagnosed with Alzheimer living at nursing home apartment. Her apartment is monitored by two partially overlapping color-depth sensors. She displays agitation and aberrant motor behavior and the nursing home staff is interested in finding out more about her night time behavior, *e.g.*, if she wanders during the night. In this evaluation we have evaluated thee performance of the proposed framework at describing common events in her daily routine: entering and exiting the bed, the restroom and the apartment, and sitting on the armchair. Figure 5 illustrates the monitored scene.

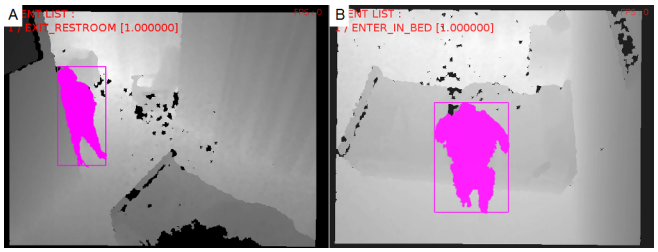

Figure 5. Monitored scene at the nursing home apartment: A) living area camera displays an "exit restroom" event and B) bed area camera displays an "enter in bed" event.

5. Results

This section summarizes the results of the evaluation carried out on CHUN (Subsection 5.1, GAADRD (Subsection 5.2) and the nursing home (Subsection 5.3) datasets.

5.1. CHUN dataset

This experiment have compared the performance of the proposed framework to baseline methods in the test set of CHUN dataset (Table 1). We have observed that the proposed approach has outperformed all variants of the baseline approach and it has also presented the performance with the smallest standard deviation of the mean. Among baseline approaches, DT-HOG has the best performance (3/6 events) followed by DT-HOF (2/6).

Table 1. Recognition of IADLs - CHUN dataset – F_1-score

Event	**DT-HOG**	**DT-HOF**	**DT-MBH**^{y}	**Proposed**
Prepare drink	58.61	47.33	63.09	74.07
Prepare drug box	60.14	70.97	27.59	90.91
Read	51.75	56.26	65.87	83.33
Search bus line	66.67	63.95	42.52	60.00
Talk on telephone	92.47	46.62	72.61	95.00
Water plant	42.58	13.08	24.83	72.22
Average ± SD	62.0 ± 17.0	49.7 ± 20.3	49.4 ± 20.6	79.3 ± 13.0

SD: standard deviation of the mean

Table 2. Recognition of a physical task in CHUN dataset

IADL	**Recall** (%)	**Precision** (%)	F_1**-score** (%)
Walking 8m	90.75	93.10	91.91

N : 58 *participants*; 7 *min*. *each*; *Total* : 406 *min*.

Table 3. Recognition of IADLs in CHUN dataset

IADL	**Recall** (%)	**Precision** (%)	F_1**-score**(%)
Prepare drink	89.4	71.9	79.7
Prepare drug box	95.4	95.4	95.4
Talk on telephone	89.6	86.7	88.1
Water plant	74.1	69.0	71.5
Average	87.1	81.0	85.3

N : 45 *participants*; 15 *min*. *each*; *Total* : 675 *min*.

5.2. GAADRD dataset

The second experiment has compared the performance of the proposed framework to baseline methods on GAADRD dataset (Table 4). The proposed framework has outperformed the baseline approaches in all event categories (Table 4). In addition, it has presented the smallest standard deviation of the mean in performance while baselines completely have failed to recognize some of targeted events (*e.g.*, prepare drug box, talk on the telephone and water the plant).

Table 4. Recognition of IADLs - GAADRD dataset - F_1-score

Event	**DT-HOG**	**DT-HOF**	**DT-MBH^y**	**Proposed**
Account Balance	44.96	34.71	42.98	66.67
Prepare Drink	81.66	44.87	52.00	100.00
Prepare Drug Box	14.19	0.00	0.00	57.14
Read Article	52.10	42.86	33.91	63.64
Talk on telephone	82.35	0.00	33.76	100.00
Turn on radio	85.71	42.52	58.16	94.74
Water Plant	0.00	0.00	0.00	52.63
Average $\pm$ SD	51.8 $\pm$ 34.4	23.6 $\pm$ 22.3	31.5 $\pm$ 23.3	76.4 $\pm$ 21.0

5.3. Nursing home dataset

Finally, the last experiment has evaluated the performance of the proposed method in the nursing home dataset. We have divided this evaluation according to the different point of views of the scene (bed or living room) and the days of evaluation (Table 5). In this experiment, event models make use of the physical object type "scene". The scene object carries global information about the monitored scene and to track its dynamics, like how the number of people varies over time. This type of concept is particularly useful to model the semantics of events related to entering/exiting the scene.

Table 5. Recognition of events in Nursing Home dataset

Day	D1		D2		D3	
Index	Recall	Precision	Recall	Precision	Recall	Precision
Camera at living area						
Enter restroom	100.0	100.0	100.0	84.2	61.7	100.0
Exit restroom	100.0	34.8	100.0	41.0	100.0	81.4
Leave room	91.1	100.0	63.0	100.0	96.7	100.0
Enter room	79.7	100.0	61.1	100.0	98.3	100.0
Sit in armchair	100.0	100.0	87.5	100.0	100.0	45.4
Average	94.2	87.0	82.3	85.0	91.3	85.4
Camera at bed area						
Enter bed	100.0	100.0	100.0	62.5	100.0	77.8
Bed exit	50.0	100.0	100.0	100.0	100.0	77.8
Average	75.0	100.0	100.0	81.2	100.0	77.8

N: 1 participant, 72 hours of recording per sensor.

6. Discussion

This paper presented a full-working framework for visual activity recognition using color-depth sensing and semantic events. This section summarizes the main findings of our evaluation ranging from the qualitative analysis of people tracking module up to the quantitative measurement of activity recognition performance on the three datasets depicting seniors carrying out activities of daily living.

6.1. Overall people tracking

A qualitative evaluation of people tracking performance has showed that in the short-term scenarios, such as the recognition of physical tasks, the tracking quality was nearly 100%. In mid-term scenarios, like daily living activities, the tracking quality dropped in cases of poor detection due to partial occlusion of a person's body (*e.g.*, person close to image borders or to scene furniture) or to the person be spending several minutes outside of the field of view of the sensor. The execution time of the event recognition framework is currently around 3.5 frames per second (people detection, tracking and event recognition), which enables a close to online monitoring of older people across most of the situations observed.

6.2. CHUN dataset

The proposed approach has outperformed all variants of the baseline approach and it also presented the performance with the smallest standard deviation of the mean. Among baseline approaches, DT-HOG has the best performance (3/6 events) followed by DT-HOF (2/6). The superior performance of the proposed method is mostly due to its capability to handle variable frame rate (here 4-15 frames per second) and to model the temporal dependencies between activity components. Since baseline methods rely on a temporal sliding window to capture temporal dependencies in test time, the information about short-duration activities is generally shadowed by the information of longer ones (*e.g.*, the short "water the plant" versus the long "prepare drink"). The performance of the proposed framework (Tables 2 and 3) can be also favorably compared to state-of-the-art approaches in a dataset with similar activities but different participants and camera setting [3]. Our framework has achieved a performance similar to prior work at the recognition of physical tasks (average recall: +1.12%, average precision: -4.6%). However, it had a higher precision for IADL recognition, which are more complex activities (av. precision: +4%). Finally, we have also observed that the performance of the proposed approach remains relatively stable as the size of the dataset increases (Table 1 x Table 3).

6.3. GAADRD dataset

The proposed framework has also outperformed baseline approaches in this dataset and, as in CHUN dataset, it has presented the smallest standard deviation of the mean in recognition performance. Baseline methods particularly failed to recognize the activities of "prepare drug box", "water the plant" and "talk on the telephone". This happens because the first two activities have an even briefer duration than in CHUN dataset. "Talk on the telephone" activity, on the other hand, is particularly challenging, because it takes place at the back of the scene and its most discriminative feature is its localization. Baseline methods have difficulty in capturing this subtle piece of information. Moreover, baseline methods were strongly affected by the low and variable frame rate of the dataset recordings (4 to 10 frames per seconds), a characteristic that the proposed framework can handle by focusing on relative temporal relations between events.

The ontology-driven component of the framework has made easy to port event models between the two datasets, since they contain similar activities. For instance, baselines approaches had to be re-trained from scratch to be tested on GAADRD dataset. However, to test the proposed method on the new dataset, we only had to update the geometry of the semantic zones and the minimum duration of activities to match the characteristics of the dataset (*e.g.*, "prepare drink" and "watering the plant" events only takes a few seconds in GAADRD contrary to CHUN dataset). The structure of event models and other semantics have remained unchanged. A trained expert only took a few minutes to carry these changes out

In summary, the ontology-driven formalism has a recognition performance that is superior to baselines with the great advantage of facilitating the transfer of event knowledge between different scenes, a important feature for real-world applications that baselines lack.

6.4. Nursing home dataset

In the nursing home dataset - long-term scenarios - the proposed approach has also presented a high recognition performance (mean recall and precision are, respectively, 89.27% and 85.78% for living area events, and 91.66% and 86.35% for bed area events). We have observed that this performance generalizes across the monitored days, a fact which highlights the robustness of the proposed approach for unconstrained environments. But, even though it has achieved a reasonable recognition performance in this unconstrained setting, a few challenges remain for future work. For instance, the low performance in "exit restroom", "enter and leaving room" and "bed exit" events. This problem happens due to strong occlusion of the person's body by either walls and door frames (Fig.5a) or scene furniture (Fig.5b), like the bed. Missed instances of "exit bed" (see Model 1) refer to failures at people detection step that harm the recognition of the transition between a person "lying on the bed" to "standing" in front of it with legs occluded. To solve the reported cases, it is necessary to consider uncertainty estimates for the different steps of the vision pipeline and then reason accordingly to the scene geometry, a characteristic that the proposed method and state-of-the-art methods still lack.

6.5. Summary

The proposed framework outperforms baselines approaches on a variety of activities of daily living and on different datasets. Results also demonstrate that the performance of the proposed framework scales both for a larger number of participants and for unconstrained scenarios, like nursing home apartments. The demonstrated improvements come from addressing previously described limitations of event recognition using color-depth sensing [3], like the short-range field of view of the depth sensor; the underestimation of people's body size - due to noisy depth signal and occlusion of body parts, and event reasoning on recordings with variable and low frame-rate. By handling these limitations, the proposed framework enables the modeling of longer temporal relations between events which are more natural of real-life scenarios. Finally, the proposed framework can also distinguish among the activities of different people in the scene, a feature that is very important for assisted living scenarios and that state-of-the-art methods lack [24].

7. Conclusion

This paper has introduced and extensively evaluated a fully working, knowledge-driven framework for the recognition of daily activities of senior people in assisted living scenarios. The framework combines a constraint-based ontology language to model daily living activities with a robust pipeline for people detection and tracking on color-depth signals. The proposed framework outperforms baseline approaches and enables the modeling and recognition of longer and more complex events, natural to real-life scenarios. The framework is currently used at a partner medical institute to support the daily evaluation of symptoms of Alzheimer's disease; and in a study of the daily activities of seniors at nursing home's apartments and at their domiciles.

Further work will investigate how to extend the framework to handle uncertainty, to fuse multiple sensor data, and to support the automatic diagnosis of cognitive disorders from event data.

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.

Author Contributions: CFCJ has designed, evaluated and supervised the development and tuning of the proposed framework for ambient assisted living. CFCJ and AK have developed and refined the IADL models. AGU has evaluated the proposed framework on CHUN dataset. CS has continued the work of AGU for nursing home settings. MK, FN and SC have provided the baseline method implementations, with CFCJ, FN and SC supporting the transfer of the proposed framework for GAADRD dataset. ATN has designed the people detection module, while DPC has designed and fine-tuned the tracking module. GC has developed the algorithm for ground-plane computation from 3D clouds of points which was adapted by CFCJ to work on semantic spatial

zones. FB has designed the people-centered architecture for event recognition framework and supervised all co-authors.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
IADL	Instrumental Activities of Daily Living

Bibliography

1. Fleury, A.; Noury, N.; Vacher, M. Introducing knowledge in the process of supervised classification of activities of Daily Living in Health Smart Homes. Proceedings of 12th IEEE International Conference on e-Health Networking Applications and Services, 2010, pp. 322 – 329.
2. Medjahed, H.; Istrate, D.; Boudy, J.; Baldinger, J.L.; Dorizzi, B. A pervasive multi-sensor data fusion for smart home healthcare monitoring. Proceedings of IEEE International Conference on Fuzzy Systems, 2011, pp. 1466–1473.
3. Crispim-Junior, C.; Bathrinarayanan, V.; Fosty, B.; Konig, A.; Romdhane, R.; Thonnat, M.; Bremond, F. Evaluation of a Monitoring System for Event Recognition of Older People. Proceedings of the 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance 2013, AVSS 2013, 2013.
4. Banerjee, T.; Keller, J.M.; Popescu, M.; Skubic, M. Recognizing Complex Instrumental Activities of Daily Living Using Scene Information and Fuzzy Logic. *Comput. Vis. Image Underst.* **2015**, *140*, 68–82.
5. Tasoulis, S.; Doukas, C.; Plagianakos, V.; Maglogiannis, I. Statistical data mining of streaming motion data for activity and fall recognition in assistive environments. *Neurocomputing* **2013**, *107*, 87 – 96. Timely Neural Networks Applications in Engineering Selected Papers from the 12th {EANN} International Conference, 2011.
6. Gonzalez-Ortega, D.; Díaz-Pernas, F.; Martínez-Zarzuela, M.; Antón-Rodríguez, M. A Kinect-based system for cognitive rehabilitation exercises monitoring. *Computer Methods and Programs in Biomedicine* **2014**, *113*, 620 – 631.
7. Vu, T.; Bremond, F.; Thonnat, M. Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), 2003.
8. Cao, Y.; Tao, L.; Xu, G. An event-driven context model in elderly health monitoring. Proceedings of Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, 2009.
9. Chen, L.; Hoey, J.; Nugent, C.; Cook, D.; Yu, Z. Sensor-Based Activity Recognition. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **2012**, *42*, 790–808.
10. Town, C. Ontological inference for image and video analysis. *Machine Vision and Applications* **2006**, *17*, 94–115.
11. Ceusters, W.; Corso, J.J.; Fu, Y.; Petropoulos, M.; Krovi, V. Introducing Ontological Realism for Semi-Supervised Detection and Annotation of Operationally Significant Activity in Surveillance Videos. Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense and Security (STIDS), 2010.
12. Chen, L.; Nugent, C.; Okeyo, G. An Ontology-Based Hybrid Approach to Activity Modeling for Smart Homes. *Human-Machine Systems, IEEE Transactions on* **2014**, *44*, 92–105.
13. Rantz, M.; Banerjee, T.; Cattoor, E.; Scott, S.; Skubic, M.; Popescu, M. Automated Fall Detection With Quality Improvement "Rewind" to Reduce Falls in Hospital Rooms. *J Gerontol Nurs.* **2014**, *40:1*.
14. Tran, S.D.; Davis, L.S. Event Modeling and Recognition Using Markov Logic Networks. ECCV '08: Proceedings of the 10th European Conference on Computer Vision; Springer-Verlag: Berlin, Heidelberg, 2008; pp. 610–623.
15. Kitani, K.M.; Ziebart, B.D.; Bagnell, J.A.D.; Hebert, M. Activity Forecasting. European Conference on Computer Vision. Springer, 2012.

16. Kwak, S.; Han, B.; Han, J.H. Scenario-based video event recognition by constraint flow. CVPR. IEEE, 2011, pp. 3345–3352.
17. Brendel, W.; Fern, A.; Todorovic, S. Probabilistic event logic for interval-based event recognition. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 3329–3336.
18. Nghiem, A.T.; Bremond, F. Background subtraction in people detection framework for RGB-D cameras, 2014. accepted paper in 11-th IEEE International Conference on Advanced Video and Signal-Based Surveillance.
19. Pramerdorfer, C. Evaluation of Kinect Sensors for Fall Detection. IASTED International Conference. Signal Processing, Pattern Recognition and Applications; , 2013; SPPRA 2013.
20. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time Human Pose Recognition in Parts from Single Depth Images. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition; IEEE Computer Society: Washington, DC, USA, 2011; CVPR 2011, pp. 1297–1304.
21. Kuhn, H.W. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83–97.
22. Chau, D. P.; Thonnat, M.; Bremond, F. Automatic parameter adaptation for multi-object tracking. Proceedings of International Conference on Computer Vision Systems (ICVS), 2013.
23. Allen, J.F. Maintaining Knowledge About Temporal Intervals. *Commun. ACM* **1983**, *26*, 832–843.
24. Wang, H.; Klaser, A.; Schmid, C.; Liu, C.L. Action Recognition by Dense Trajectories. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition; IEEE Computer Society: Washington, DC, USA, 2011; CVPR '11, pp. 3169–3176.
25. Folstein, M.F.; Robins, L.N.; Helzer, J.E. THe mini-mental state examination. *Archives of General Psychiatry* **1983**, *40*, 812.
26. Karakostas, A.; Briassouli, A.; Avgerinakis, K.; Kompatsiaris, I.; M., T. The Dem@Care Experiments and Datasets: a Technical Report. Technical report, Centre for Research and Technology Hellas, 2014.