



HAL
open science

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.

► **To cite this version:**

Carlos F Crispim-Junior, Alvaro Gómez Uría, Carola Strumia, Michal Koperski, Alexandra Konig, et al.. Online recognition of daily activities by color-depth sensing and knowledge models. *Sensors*, 2017, 17 (7), pp.1-15. 10.3390/s17071528. hal-01658438

HAL Id: hal-01658438

<https://inria.hal.science/hal-01658438v1>

Submitted on 7 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

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

Carlos Fernando Crispim-Junior^{1,2*}, Alvaro Gómez Uría¹, Carola Strumia¹, Michal Koperski¹, Alexandra Konig^{2,3}, Farhood Negin¹, Serhan Cosar¹, Anh Tuan Nghiem¹, Duc Phu Chau¹, Guillaume Charpiat¹ 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 L^AT_EX 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].

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

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

50 2. Related work

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

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

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

83 The proposed event recognition framework differs from prior work on the vision pipeline
 84 adopted (people detection and tracking approaches) and the formalism for event recognition. Firstly,
 85 color-depth sensing tends to be more robust to noise due to scene illumination changes than
 86 conventional video cameras, and it allow the subsequent modules to solve 2D ambiguities by using
 87 3D measurements of the scene. Moreover, the proposed vision pipeline employs an algorithm
 88 for people detection in color-depth sensing that extends the range of people detection from 3-4m
 89 (e.g., Microsoft and Primesense libraries) to 7-9 m, by handling noise at depth pixel-level. Finally,
 90 the ontology-driven mechanism of reasoning allows to incorporate different sources of information
 91 efficiently, from common sense knowledge and event semantics up to dynamic information about
 92 visual entities. The combination of both modules enables to model more complex and longer
 93 time-dependencies among events, barely explored before on online activity recognition.

94 3. Knowledge-driven event recognition

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

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

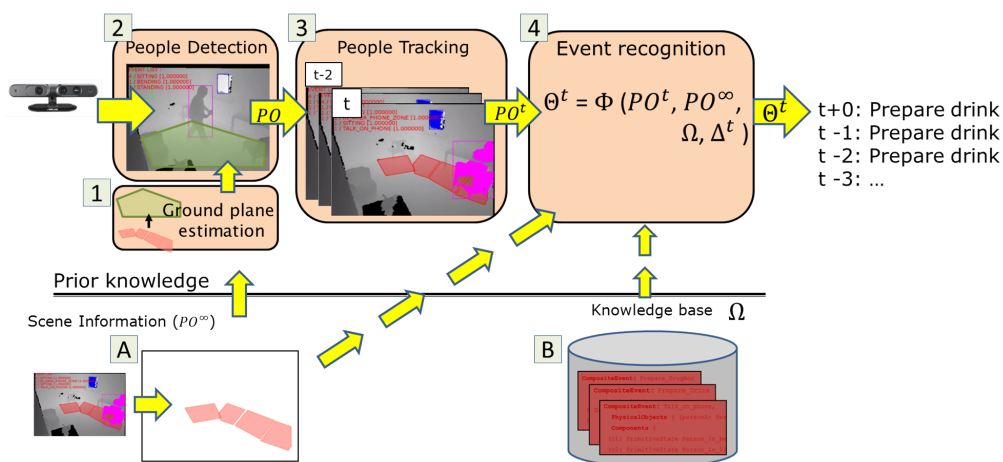


Figure 1. Knowledge-driven framework for visual event recognition. Firstly, 0) an estimation of the ground-plane is computed using the vertexes of semantic zones. 1) Video frame acquisition is performed using a color-depth sensor. Then, 2) people detection module analyzes the video frame for instances 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's 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.

105 3.1. Ground plane estimation

106 The estimation of a ground-plane is a key step for the vision pipeline, since its output is
107 employed to improve the performance of the subsequent steps of people detection and tracking.
108 The estimation process is made as follows: firstly, we search locally for pieces of planes, using the
109 3D-vertexes of the semantic zones. For each 3D vertex, we consider the cloud of points formed by
110 its nearest neighbors and find the best plane which approximates it in the least square error sense
111 (closed-form solution). When the approximation error is too high, *i.e.*, when the local cloud of points is
112 not flat enough, the plane is discarded. The obtained planes are clustered into larger planes weighted
113 by the number of 3D points they possess. We compare any two pieces of planes during the clustering
114 step based on the angle between their Normals and on their alignment (distance between each center
115 of mass and the other plane). We sort the newly obtained planes by their confidence (approximation
116 error and number of points involved) and assign the first nearly horizontal plane to the ground plane.

117 3.2. People detection

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

137 3.3. People tracking

138 The tracking algorithm takes as input the video stream and a list of detected people in a temporal
139 sliding window. First, a link score is computed between any two detected people appearing in the
140 time window using a weighted combination of six object descriptors: 2D and 3D positions, 2D object
141 area, 2D object shape ratio, color histogram and dominant color. Hypothesis trajectories are built from
142 links with scores greater than a pre-defined threshold. The reliability of each hypothesis trajectory
143 is represented by the total score of its link scores. The trajectory of the objects are determined by
144 maximizing objects' trajectory reliability using the Hungarian algorithm [21]. Since the descriptor
145 weights generally depend on the content of the video being processed, we use the control algorithm
146 proposed by [22] to tune the weights on an online manner.

147 3.4. Ontology-driven event recognition

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

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

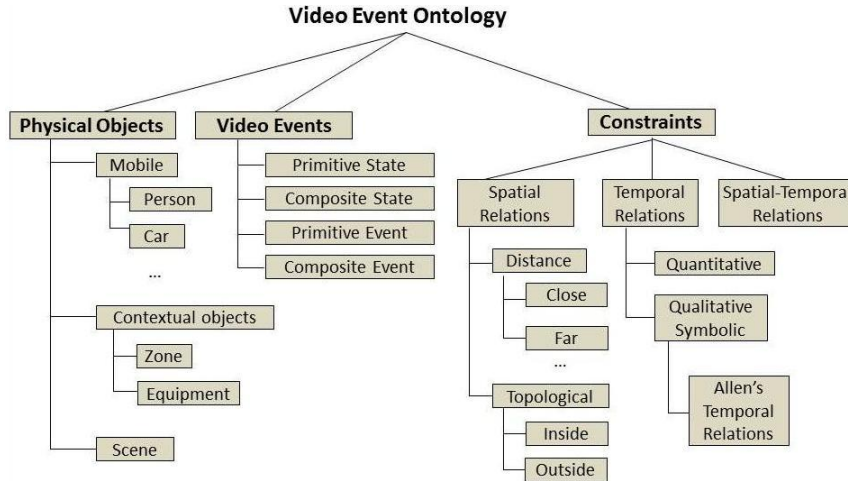


Figure 2. Video event ontology language. Three main concept branches are defined: physical objects, video events and constraints. Physical objects make abstractions for real-world objects. Video events describe the types of event templates available for activity modeling. Constraints describes the relations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

$$\omega_j = \langle PO_j, SE_j, CO_j \rangle \quad (1)$$

156 where,

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

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

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

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

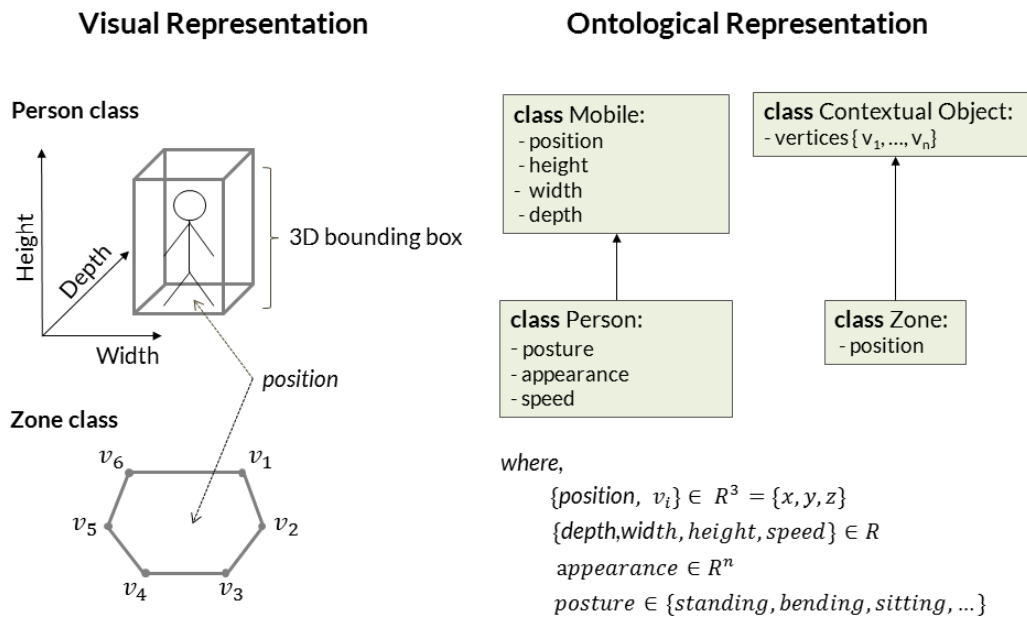


Figure 3. Physical objects integrate 3D visual information into the ontological events

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

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

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

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

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

213 **Model 1.** *Composite Event bed exit*

```

214
215 CompositeEvent(BED_EXIT,
216   PhysicalObjects(( $p_1$ :Person),( $z_B$ :Zone),( $z_{SB}$ :Zone))
217   Components(
218     ( $s_1$ : PrimitiveState in_zone_bed ( $p_1, z_B$ ))
219     ( $s_2$ : PrimitiveState out_of_bed ( $p_1, z_{SB}$ )))
220   Constraints(( $s_1$  meet  $s_2$ ) //  $c_1$ 
221     (duration( $s_2$ ) > 1)) //  $c_2$ 
222   Alarm ((Level : URGENT))
223 )

```

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

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

243 4. Experiments

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

251 4.1. Performance baselines

252 To compare the performance of the proposed approach with the state of the art, we have
253 chosen the action recognition pipeline described in [24]. Support Vector Machines (SVM) for action
254 classification trained with a bag-of-visual-word embedding over descriptors of dense trajectories
255 features. In short, for each video sequence we have first extracted local spatio-temporal patches using
256 dense trajectories' detector. Then we have cut patches around each trajectory point as described in
257 [24]. For each patch, we compute standard descriptors: trajectory shape, HOG, HOF and MBH. Then
258 we have used each of the latter three descriptors to create a bag-of-words (BoW) representation as
259 embedding function (with $k = 4000$). Finally, support vector machines with RBF kernel are used to
260 classify the video representation as one of the target classes. Classifiers are learned on a supervised
261 manner using video segments clipped from the original video sequence using ground-truth data. For
262 online testing, the descriptors of a video are extracted over a temporal sliding window of size W
263 (frames) with step size T and a minimal number of features extracted denoted as M . For each sliding
264 window step we have extracted descriptors and apply BoW with SVM classifier given the number
265 of detected features is equal or above M . Hyperparameters W , T and M were, empirically, set to
266 40, 15 and 20; respectively. A hold-out validation scheme is employed for training and testing the
267 baseline classifiers. Baseline approaches were: Dense trajectories (DT) with Histogram of Gradients
268 descriptor (HOG), DT with Histogram of Optical Flow (HOF) and DT with the y-component of
269 Motion Boundary Histogram (MBH^y). All results are reported on the tested set of the respective
270 baselines.

271 4.2. CHUN dataset

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

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

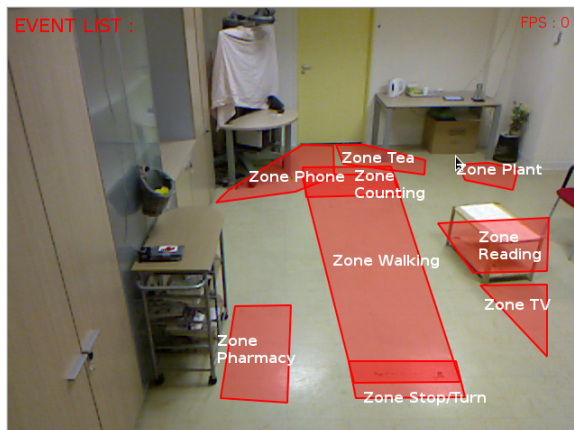


Figure 4. Contextual zones define geometric regions (red polygons, CHUN dataset) that carry semantic information about daily activities.

289 4.3. GAADR dataset

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

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

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

311 4.4. Nursing home dataset

312 This dataset consists of 72 hours of depth data recording about an 86 years old female, diagnosed
 313 with Alzheimer living at nursing home apartment. Her apartment is monitored by two partially
 314 overlapping color-depth sensors. She displays agitation and aberrant motor behavior and the nursing
 315 home staff is interested in finding out more about her night time behavior, *e.g.*, if she wanders during
 316 the night. In this evaluation we have evaluated the performance of the proposed framework at
 317 describing common events in her daily routine: entering and exiting the bed, the restroom and the
 318 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.

319 5. Results

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

322 5.1. CHUN dataset

323 This experiment have compared the performance of the proposed framework to baseline
324 methods in the test set of CHUN dataset (Table 1). We have observed that the proposed approach
325 has outperformed all variants of the baseline approach and it has also presented the performance
326 with the smallest standard deviation of the mean. Among baseline approaches, DT-HOG has the best
327 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 \pm SD	62.0 \pm 17.0	49.7 \pm 20.3	49.4 \pm 20.6	79.3 \pm 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.

328 5.2. GAARDR dataset

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

Table 4. Recognition of IADLs - GAARDR 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

334 5.3. Nursing home dataset

335 Finally, the last experiment has evaluated the performance of the proposed method in the
 336 nursing home dataset. We have divided this evaluation according to the different point of views of
 337 the scene (bed or living room) and the days of evaluation (Table 5). In this experiment, event models
 338 make use of the physical object type "scene". The scene object carries global information about the
 339 monitored scene and to track its dynamics, like how the number of people varies over time. This
 340 type of concept is particularly useful to model the semantics of events related to entering/exiting the
 341 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.

342 6. Discussion

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

348 6.1. Overall people tracking

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

357 6.2. CHUN dataset

358 The proposed approach has outperformed all variants of the baseline approach and it also
359 presented the performance with the smallest standard deviation of the mean. Among baseline
360 approaches, DT-HOG has the best performance (3/6 events) followed by DT-HOF (2/6). The superior
361 performance of the proposed method is mostly due to its capability to handle variable frame rate (here
362 4-15 frames per second) and to model the temporal dependencies between activity components. Since
363 baseline methods rely on a temporal sliding window to capture temporal dependencies in test time,
364 the information about short-duration activities is generally shadowed by the information of longer
365 ones (*e.g.*, the short "water the plant" versus the long "prepare drink"). The performance of the
366 proposed framework (Tables 2 and 3) can be also favorably compared to state-of-the-art approaches
367 in a dataset with similar activities but different participants and camera setting [3]. Our framework
368 has achieved a performance similar to prior work at the recognition of physical tasks (average recall:
369 +1.12%, average precision: -4.6%). However, it had a higher precision for IADL recognition, which
370 are more complex activities (av. precision: +4%). Finally, we have also observed that the performance
371 of the proposed approach remains relatively stable as the size of the dataset increases (Table 1 x Table
372 3).

373 6.3. GAADDRD dataset

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

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

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

395 6.4. Nursing home dataset

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

410 6.5. Summary

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

422 7. Conclusion

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

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

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

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

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

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

447 Abbreviations

448 The following abbreviations are used in this manuscript:

449 MDPI Multidisciplinary Digital Publishing Institute

450 IADL Instrumental Activities of Daily Living

451 Bibliography

- 452 1. Fleury, A.; Noury, N.; Vacher, M. Introducing knowledge in the process of supervised classification of
453 activities of Daily Living in Health Smart Homes. *Proceedings of 12th IEEE International Conference on*
454 *e-Health Networking Applications and Services*, 2010, pp. 322 – 329.
- 455 2. Medjahed, H.; Istrate, D.; Boudy, J.; Baldinger, J.L.; Dorizzi, B. A pervasive multi-sensor data fusion for
456 smart home healthcare monitoring. *Proceedings of IEEE International Conference on Fuzzy Systems*,
457 2011, pp. 1466–1473.
- 458 3. Crispim-Junior, C.; Bathrinayanan, V.; Fosty, B.; Konig, A.; Romdhane, R.; Thonnat, M.; Bremond, F.
459 Evaluation of a Monitoring System for Event Recognition of Older People. *Proceedings of the 10th IEEE*
460 *International Conference on Advanced Video and Signal-Based Surveillance 2013, AVSS 2013*, 2013.
- 461 4. Banerjee, T.; Keller, J.M.; Popescu, M.; Skubic, M. Recognizing Complex Instrumental Activities of Daily
462 Living Using Scene Information and Fuzzy Logic. *Comput. Vis. Image Underst.* **2015**, *140*, 68–82.
- 463 5. Tasoulis, S.; Doukas, C.; Plagianakos, V.; Maglogiannis, I. Statistical data mining of streaming motion
464 data for activity and fall recognition in assistive environments. *Neurocomputing* **2013**, *107*, 87 – 96. *Timely*
465 *Neural Networks Applications in Engineering Selected Papers from the 12th {EANN} International*
466 *Conference*, 2011.
- 467 6. Gonzalez-Ortega, D.; Díaz-Pernas, F.; Martínez-Zarzuela, M.; Antón-Rodríguez, M. A Kinect-based
468 system for cognitive rehabilitation exercises monitoring. *Computer Methods and Programs in Biomedicine*
469 **2014**, *113*, 620 – 631.
- 470 7. Vu, T.; Bremond, F.; Thonnat, M. Automatic Video Interpretation: A Novel Algorithm for Temporal
471 Scenario Recognition. *The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03)*,
472 2003.
- 473 8. Cao, Y.; Tao, L.; Xu, G. An event-driven context model in elderly health monitoring. *Proceedings of*
474 *Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, 2009.
- 475 9. Chen, L.; Hoey, J.; Nugent, C.; Cook, D.; Yu, Z. Sensor-Based Activity Recognition. *Systems, Man, and*
476 *Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **2012**, *42*, 790–808.
- 477 10. Town, C. Ontological inference for image and video analysis. *Machine Vision and Applications* **2006**,
478 *17*, 94–115.
- 479 11. Ceusters, W.; Corso, J.J.; Fu, Y.; Petropoulos, M.; Krovi, V. Introducing Ontological Realism for
480 Semi-Supervised Detection and Annotation of Operationally Significant Activity in Surveillance Videos.
481 *Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense and*
482 *Security (STIDS)*, 2010.
- 483 12. Chen, L.; Nugent, C.; Okeyo, G. An Ontology-Based Hybrid Approach to Activity Modeling for Smart
484 Homes. *Human-Machine Systems, IEEE Transactions on* **2014**, *44*, 92–105.
- 485 13. Rantz, M.; Banerjee, T.; Cattoor, E.; Scott, S.; Skubic, M.; Popescu, M. Automated Fall Detection With
486 Quality Improvement “Rewind” to Reduce Falls in Hospital Rooms. *J Gerontol Nurs.* **2014**, *40*:1.
- 487 14. Tran, S.D.; Davis, L.S. Event Modeling and Recognition Using Markov Logic Networks. *ECCV '08:*
488 *Proceedings of the 10th European Conference on Computer Vision*; Springer-Verlag: Berlin, Heidelberg,
489 2008; pp. 610–623.
- 490 15. Kitani, K.M.; Ziebart, B.D.; Bagnell, J.A.D.; Hebert, M. Activity Forecasting. *European Conference on*
491 *Computer Vision*. Springer, 2012.

- 492 16. Kwak, S.; Han, B.; Han, J.H. Scenario-based video event recognition by constraint flow. *CVPR. IEEE*,
493 2011, pp. 3345–3352.
- 494 17. Brendel, W.; Fern, A.; Todorovic, S. Probabilistic event logic for interval-based event recognition.
495 *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 3329–3336.
- 496 18. Nghiem, A.T.; Bremond, F. Background subtraction in people detection framework for RGB-D cameras,
497 2014. accepted paper in 11-th IEEE International Conference on Advanced Video and Signal-Based
498 Surveillance.
- 499 19. Pramerdorfer, C. Evaluation of Kinect Sensors for Fall Detection. *IASTED International Conference.*
500 *Signal Processing, Pattern Recognition and Applications; , 2013; SPPRA 2013.*
- 501 20. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time
502 Human Pose Recognition in Parts from Single Depth Images. *Proceedings of the 2011 IEEE Conference on*
503 *Computer Vision and Pattern Recognition; IEEE Computer Society: Washington, DC, USA, 2011; CVPR*
504 *2011*, pp. 1297–1304.
- 505 21. Kuhn, H.W. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly* **1955**,
506 *2*, 83–97.
- 507 22. Chau, D. P.; Thonnat, M.; Bremond, F. Automatic parameter adaptation for multi-object tracking.
508 *Proceedings of International Conference on Computer Vision Systems (ICVS)*, 2013.
- 509 23. Allen, J.F. Maintaining Knowledge About Temporal Intervals. *Commun. ACM* **1983**, *26*, 832–843.
- 510 24. Wang, H.; Klaser, A.; Schmid, C.; Liu, C.L. Action Recognition by Dense Trajectories. *Proceedings*
511 *of the 2011 IEEE Conference on Computer Vision and Pattern Recognition; IEEE Computer Society:*
512 *Washington, DC, USA, 2011; CVPR '11*, pp. 3169–3176.
- 513 25. Folstein, M.F.; Robins, L.N.; Helzer, J.E. THE mini-mental state examination. *Archives of General Psychiatry*
514 **1983**, *40*, 812.
- 515 26. Karakostas, A.; Briassouli, A.; Avgerinakis, K.; Kompatsiaris, I.; M., T. The Dem@Care Experiments and
516 Datasets: a Technical Report. Technical report, Centre for Research and Technology Hellas, 2014.

517 © 2017 by the authors. Submitted to *Sensors* for possible open access publication under the terms and conditions
518 of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).