D5.4

Multi-Parametric Behaviour Interpretation v2

Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support

Dem@Care - FP7-288199

Deliverable Information

Project Ref. No.		FP7-288199
Project Acronym		Dem@Care
Project Full Title		Dementia Ambient Care: Multi-Sensing Monitoring for Intelligence Remote Management and Decision Support
Dissemination level:		Public
Contractual date of delivery:		Month 30, 30 April 2014
Actual date of delivery:		Month 31, 28 May 2014
Deliverable No.		D5.4
Deliverable Title		Multi-Parametric Behaviour Interpretation v2
Type:		Report
Approval Status:		Approved
Version:		1.0
Number of pages:		59
WP:		WP5 Medical Ambient Intelligence
Task:		T5.3 Multi-parametric Patient Behaviour Interpretation
WP/Task responsible:		CERTH
Other contributors:		INRIA, UB1, CS
Authors (Partner)		Georgios Meditskos (CERTH), Efstratios Kontopoulos (CERTH), Carlos Crispim-Junior (INRIA), Serhan Cosar (INRIA), Francois Bremond (INRIA), Rémi Mégret (UB1), Gaelle Usseglio (UB1), Yannick Berthoumieu, Vincent Buso (UB1), Jenny Benois-Pineau (UB1), Dafni Stampouli (CS)
Responsible Author	**Name**	Georgios Meditskos (CERTH)
	Email	gmeditsk@iti.gr
Internal Reviewer(s)		Eamonn Newman (DCU)
EC Project Officer		Gerard Cultot
Abstract **(for dissemination)**		This deliverable reports on the second version (v2) of the multi-parametric interpretation framework of Dem@Care, describing the approaches that have been implemented for handling uncertainty and incomplete and noisy input. More specifically, a probabilistic framework is presented that extends the current deterministic constraint-based approach for event modelling based on low-level data and small differences in conditioned values that may be attributed to person-to-person differences when performing activities. In addition, a probabilistic activity recognition approach is presented that focuses on the management of uncertainty when fusing places and objects detected from wearable camera. In order to further handle the intrinsic challenges in multi-sensor fusion environments, such as imperfect information, noise, conflicts and inaccurate temporal correlations, a knowledge-driven fusion framework is described based on loosely coupled domain activity dependencies. The approach is based on the definition of context connections, i.e. links among relevant groups of observations that signify the presence of complex activities. Finally, further functional extensions to v1 are described, such as the support of questionnaire-related data and the incorporation of a Complex Event Processing engine to provide basic real-time interpretation services.

Version Log

Version	Date	Change	Author
0.1	09/03/2014	Deliverable outline	Georgios Meditskos (CERTH)
0.2	21/04/2014	UB1 contribution	Rémi Mégret, Gaelle Usseglio, Yannick Berthoumieu, Vincent Buso, Jenny Benois-Pineau (UB1)
0.3	30/04/2014	CS contribution	Dafni Stampouli (CS)
0.4	16/05/2014	Refinement in Section 5.3	Rémi Mégret (CS)
0.5	16/05/2014	CERTH contribution, first merging	Georgios Meditskos, Efstratios Kontopoulos (CERTH)
0.6	19/05/2014	Evaluation in Section 6.2.5	Georgios Meditskos (CERTH)
0.7	20/05/2014	INRIA contribution	Carlos Crispim-Junior, Francois Bremond (INRIA)
0.8	20/05/2014	Final draft for internal review	Georgios Meditskos (CERTH)
0.9	23/05/2014	Internal review feedback	Eamonn Newman (DCU)
0.10	24/05/2014	Address internal review comments	Georgios Meditskos (CERTH)
1.0	26/05/2014	Addressed review comments - Final version	Carlos Crispim-Junior (INRIA)

Executive Summary

The goal of multi-parametric behaviour interpretation in the Dem@Care project is to recognise the behaviour of the person with dementia (PwD) and discern traits that have been identified by the clinicians as relevant for diagnostic, status assessment, enablement and safety purposes. To this end, the information made available by WP3 and WP4 regarding physiological and lifestyle characteristics, as well as information regarding activities of daily living, is fused and aggregated in WP5 to derive high-level interpretations and decision support tasks.

The first version (v1) of the multi-parametric interpretation framework was presented in D5.2 [19], outlining the basic methods that were adopted by the two core modules of the framework, namely the Complex Activity Recognition (CAR) and Semantic Interpretation (SI) components. More specifically, CAR serves for identifying complex activities whose modelling is grounded on information at the level of person posture and location, whereas SI addresses situations that require encapsulating pieces of information of higher abstraction.

This document reports on the second version (v2) of the multi-parametric behaviour interpretation framework. It reviews state-of-the-art approaches relevant to the interpretation objectives of WP5 and proceeds by describing the extensions that have been implemented for supporting reasoning under uncertainty and handling incomplete and noisy input. The report closes by elaborating on further functional extensions to v1, such as the support of questionnaire-related data and the incorporation of a Complex Event Processing engine to provide basic real-time interpretation services, and with a discussion of future directions.

Abbreviations and Acronyms

ADL	Activities of Daily Life
AO	Active Object
VP	Visual Place
CAR	Complex Activity Recognition
DLs	Description Logics
DnS	Descriptions and Situations
DTI-2	Philips Discrete Tensions Indicator
DUL	DOLCE UltraLite
Gear4	Gear4 Renew Sleep Clock
HAR	Human Action Recognition from static and wearable cameras
IADL	Instrumental Activities of Daily Living
KB	Knowledge Base
KBM	Knowledge Base Manager
ORWC	Object Recognition from Wearable Camera
OSA	Offline Speech Analyser
OWL	Ontology Web Language
OWL	Ontology Web Language
OWL-DL	Ontology Web Language Description Language
PDT-PER	People Detection, Tracking and Primitive Events Recognition
PwD	Person with Dementia
RDF	Resource Definition Framework
RRWC	Room Recognition from Wearable Camera
SI	Semantic Interpretation
SPARQL	SPARQL Protocol And RDF Query Language
SPIN	SPARQL Inferencing Notation
SVM	Support Vector Machine
SWRL	Semantic Web Rule Language
TURTLE	Terse RDF Triple Language
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WCPU	Wearable Camera Processing Using
WP	Work Package
XML	eXtensible Markup Language

Table of Contents

List of Figures

List of Tables

1 Introduction

Sensor data are inherently imperfect. Inaccuracies may frequently arise, due to erroneous and/or missing sensor readings. Furthermore, when data is retrieved from multiple sources, components and modalities, ambiguities and conflicts may also arise. Under these circumstances, modelling and reasoning need to provide the means to cope with such imperfections and allow detecting possible errors, gracefully handling missing values, and deriving plausible conclusions, assessing the validity of available sensor data.

A significant challenge in activity recognition is the ability to identify and recognise the context signifying the presence of complex activities. As mentioned, an important factor to take into consideration is that contextual information is typically collected by multiple sensors and complementary modalities. For example, RGB-D video streams are used in WP5 to detect activities based on information regarding posture and location. In parallel, videos captured from wearable cameras are used in WP4 to detect the objects a person interacts with and his/her location that are further fused in WP5 to derive complex activities. Therefore, each sensing modality is used in a different way by each module, generating information from a different perspective. The challenge is to effectively fuse multiple sources of heterogeneous, noisy and potentially inconsistent information in a way that provides accurate and useful outputs.

The first version of the multi-parametric interpretation framework was presented in deliverable D5.2 [19], outlining the basic methods that were adopted by the two core modules of the framework, namely the Complex Activity Recognition (CAR) and Semantic Interpretation (SI) components. CAR recognises complex activities whose modelling is grounded on information at the level of person posture and location, whereas SI addresses situations that require encapsulating pieces of information of higher abstraction using rules.

This document reports on the second version of the multi-parametric behaviour interpretation framework describing the extensions that have been implemented to further address the medical ambient intelligence interpretation requirements. Section 2 provides an overview of the functionality provided in v1 and outlines the extensions implemented in v2, while Section 3 reviews state-of-the-art approaches relevant to the interpretation objectives of WP5. In Section 4, a probabilistic framework is presented that extends the generic constraint-based ontology language for event modelling in CAR. The framework is intended to overcome the deterministic nature of the constraint-based approach by handling uncertainty from low-level data, and consequently, being less sensitive to small deviations of the values on which the constraints are conditioned. A probabilistic activity recognition approach is presented in Section 5 that focuses on the management of uncertainty when fusing places and objects detected from wearable camera. In order to further handle the intrinsic challenges in multi-sensor fusion environments, such as imperfect information, noise, conflicts and inaccurate temporal correlations, a knowledge-driven fusion framework is described in Section 6 based on loosely coupled domain activity dependencies. The approach is based on the definition of context connections, i.e. links among relevant groups of observations that signify the presence of complex activities. Finally, further functional extensions to v1 are described in Section 7, such as the support of questionnaire-related data and the incorporation of a Complex Event Processing engine to provide basic real-time interpretation services. Section 8 concludes the deliverable discussing next steps.

2 Functionality Overview

In this section we summarise the functionality provided in the first version (v1) of the multi-parametric behaviour interpretation framework and we elaborate on the extensions that have been implemented in the second version (v2).

2.1 Overview of v1 Functionality

The first version of the multi-parametric behaviour interpretation framework consists of two components: the Complex Activity Recognition (CAR) component and the Semantic Interpretation (SI) component. The developed components support interpretation tasks at different levels of granularity. More specifically, CAR serves for identifying complex activities whose modelling is grounded on information at the level of person posture and location. The activity recognition task is supported by a hierarchical model-based approach that uses a generic constraint-based ontology language to describe event models in terms of a-priori knowledge of the scene (i.e. contextual objects and zones) and primitive events detected from RGB-D video streams (e.g. postures). CAR focuses primarily on the recognition of:

- The position of PwD with respect to predefined zones of interest and their moving from one zone to another (e.g. person inside the office desk zone)
- Elementary states and activities using posture and localisation information (e.g. person bending)
- Complex states and activities (e.g. person using office desk).

Figure 2-1 presents the model that describes the activity $PrepareDrink$. The activity is detected whenever the person is inside the $UseTeaCorner$ zone for more than 1 second.

In turn, the primary focus of SI is on the recognition of: i) complex situations by fusing the descriptions extracted by the various modules of the system (e.g. night bathroom visit after the person has gone to sleep), ii) functional problems as defined by clinicians (e.g. nocturia problem in case of more than two bathroom visits during the night and after the person has gone to sleep), and iii) summaries of key attributes of PwD behaviour with respect to the functional areas considered (e.g. for sleep, the number of awakening during night sleep and the number of naps the preceding daytime). SI espouses a hybrid approach that combines ontology- and rule-based reasoning. An OWL 2 ontology is used to model the domain

```
CompositeEvent(PrepareDrink,
  PhysicalObjects((p1 : Person), (z1 : Zone))
  Components((c1: PrimitiveState Person_Inside_Zone_UseTeaCorner(p1, z1)))
  Constraints((duration(c1) >= 1))
  Alarm ((Level : URGENT))
)

PrimitiveState(Person_Inside_Zone_UseTeaCorner,
    PhysicalObjects((p1 : Person), (z1 : Zone))
    Constraints ((p1->Position in z1->Vertices)
        (z1->Name = UseTeaCornerZone))
    Alarm ((Level : NOTURGENT))
)
```

Figure 2-1. Activity model for PrepareDrink

concepts (activities, situations, problems, etc.); SPARQL rules are used to enhance typical ontology-based reasoning with complex activity and problem detection, temporal reasoning and incremental knowledge updates.

Figure 2-2 presents the SPARQL rule that is used to detect the $PrepareTea$ activity. The rule combines (fuses) the result of CAR regarding the preparation of a drink and tea-related objects detected from the wearable camera in WP4. Similar SPARQL rules have been defined for the recognition of situations that indicate problems or possibly problematic behaviours that need to be highlighted to the clinician, e.g. a nocturia problem, and for aggregating and summarising the results offering a single point for collection of the PwD's contextual information.

2.2 Overview of v2 Functionality

The second version of the multi-parametric behaviour interpretation framework in WP5 extends the functionality provided in v1 with respect to the following directions:

```
CONSTRUCT{
    ?new a event:PrepareTea;
        event:startTime ?ta_start ;
        event:endTime ?ta_end ;
}
WHERE {
   ?ta a event:PrepareDrink ;
        event:startTime ?ta_start ;
    event:endTime ?ta_end .

    ?obj1 a event:UseObject ;
        event:startTime ?obj1_start ;
    event:endTime ?obj1_end ;
    event:relatedToObject event:Teabag.

    ?obj2 a event:UseObject ;
    event:startTime ?obj2_start ;
    event:endTime ?obj2_end ;
    event:relatedToObject event:Kettle.

    ?obj3 a event:UseObject ;
    event:startTime ?obj3_start ;
    event:endTime ?obj3_end ;
    event:relatedToObject event:Teabox.

    FILTER(cf:intervalIntersect(?ta_start, ?ta_end, ?obj1_start, ?obj1_end)).
    FILTER(cf:intervalIntersect(?ta_start, ?ta_end, ?obj2_start, ?obj2_end)).
    FILTER(cf:intervalIntersect(?ta_start, ?ta_end, ?obj3_start, ?obj3_end)).

    BIND (cf:newURI(?ta, "SEMI_TEA", str(anon:)) as ?new).
    FILTER NOT EXISTS {?new a [] . }.
}
```

Figure 2-2. SPARQL rule for detecting the PrepareTea activity

Uncertainly handling in event modelling: Uncertainty is present at different levels of the event modelling task, from the low-level data used as input (e.g. person dimension and position) to the intra-class variation of event models themselves (e.g. due to person-to-person differences when performing an event; event intra-class variation). We extend the current deterministic constraint-based approach for event modelling by proposing a probabilistic framework to handle uncertainty. In Section 4 we described the first part of this framework focusing on constraints based on low-level data and small differences in conditioned values that may be attributed to person-to-person differences when performing activities.

Probabilistic fusion of objects and locations: The recognition of concepts such as activities is best accomplished using classifiers, which provide the detection and recognition of activity related event. Since activity recognition is a difficult task, the classifier outputs can be noisy. Thus it is important to associate uncertainty measures to the detected events. We present in Section 5 a method for providing an evaluation of the uncertainty of the results obtained from activity recognition by fusing objects and location. It extends the classification approach by a confidence measure that transforms recognition results into probabilistic events suitable for reasoning under uncertainty.

Context-based high-level fusion: Due to the intrinsic characteristics of pervasive environments in real-world conditions, such as imperfect information, noise, conflicts or inaccurate temporal correlations, the use of strict contextual constraints to fuse information is not always a practical and flexible solution. Moreover, many activities are carried out differently even by the same person, e.g. the kettle during the make tea activity may be turned on before or after taking out a cup from the cupboard. Thus, the use of strictly structured background knowledge relevant to the order of activities or their temporal boundaries is not always able to effectively capture and reason about the context. We present in Section 6 the implementation of a fusion approach that detects complex situations based on loosely coupled domain activity dependencies rather than on strict contextual constraints.

Questionnaires: Questionnaires is an important tool for obtaining user-reported data about problems in the daily life, for example, mood and sleeping problems. We describe in Section 7.1 the knowledge structures and analysis procedures that have been developed for storing and calculating the scores of the questionnaires used in Dem@Care.

Complex Event Processing: Taking into account the fact that temporal relationships and real-time processing is of great importance in activity detection, we describe in Section 7.2 the incorporation of a Complex Event Processing engine in WP5. The purpose of the CEP engine is currently to provide real-time contextualised support of the patient via the coupling of the information made available through CAR with patient profile knowledge.

3 Related Work

This section reviews state-of-the-art approaches regarding reasoning under uncertainty and handling incomplete and noisy input for activity recognition and multi-sensor fusion in pervasive environments.

3.1 Uncertainty Modelling for Model-Based Event Detection

Automatic Event detection has become a very active area of research in the past years [27][20][48]. Event detection approaches can be divided into two main approaches: probabilistic approaches and description-based approaches. Main probabilistic approaches ranged are Bayesian Networks and Hidden Markov Models. Their main characteristic is to explicitly model the uncertainty of events. Bayesian Network have been applied to the detection of event such as person interaction [48] such as ‘shake hands’, events on parking lots [43], traffic monitoring [37], and detection of left luggage [39]. However, Bayesian Networks are not appropriated to model the temporal component of events. An alternative to them is the of HMM and its extensions [47][25][11] for time representation.

Model-based approaches have been largely used to detect activities for few decades, as they are suitable to for modelling and detecting high-level events, as they easily incorporate human knowledge into the models and require much less training data [47][44][58]. Constraint Satisfaction Problem (CSP) has been applied to model events as constraint networks [54]. Cao *et al.* [8] have proposed a model-based approach for older people monitoring whereas the human body context (*e.g.*, sitting, standing, walking) and the environment context are described in function of event models. Person body context data is provided by a set of cameras, while the environmental context is obtained of accelerometers attached to objects of daily living (e.g., TV remote control or doors use). A deterministic rule-based engine is used for reasoning and combining both context types. Zouba et al. [74] have evaluated a video monitoring system at the identification of activities of daily living of older people on a model apartment equipped with home appliances. Environmental sensors (pressure, contact) change of state was combined to video-based events using a hierarchical model-based approach. Banerjee et al. [2] have presented a fuzzy inference approach to reason over features extracted from a RGB-D camera to monitor fall events in hospital rooms.

Although it has its advantages, the deterministic nature of model-based approaches still lack a convenient mechanism to handle uncertainty and compensate for the failures of low-level components (e.g., people detection and tracking on a computer vision pipeline). Approaches combining logic and probabilistic reasoning have been proposed to overcome these two limitations. For instance, Ryoo and Aggarwal [59][60] have made use of the concept of the hallucinated time intervals, as in [42], to handle uncertainty. Tran and Davis [68] have adopted probabilistic graphical model, Markov logic networks (MLNs), to probabilistically infer events in a parking lot. In [7] a probabilistic approach is presented using weighted event-logic formulas to represent the probabilistic constraints among events. However, they did not handle low-level uncertainty but only consider the detection of primitive events of a basketball game. Kwak et al. [38] have made use of constraint flows to compose a complex event from the combination of primitive events.

Although the combination of logic and probability techniques has been a promising alternative to compensate for deterministic nature of model-based approach, there is still not a standard formalism that addresses the uncertainty of the whole hierarchy of events in model-based approaches.

The review of state of the above has been based on the published works of Romdhane et al. [55] and Crispim-Junior et al. [16].

3.2 Probabilistic Activity Recognition

Typical recognition of activities relies on classifiers, such as Support Vector Machines (SVM), which produce prediction scores. The design of successful classifiers in such applications relies on a combination of choices related to features, classifier architecture, parameters and various pre- and post-processing to take into account real data specificities such as noise, outliers and prior information such as regularization information. In typical scenarios, the complexity of the classifier system precludes the interpretation of the results as probabilistic elements, as they are defined on an arbitrary axis that is suitable for deciding of a best class, but not to associate a probabilistic interpretation to it. In this section, our objective is to study the problem of assigning appropriate probabilities to activity recognition and their suitability as probabilistic data within WP5. This problem corresponds to a calibration problem [30].

We consider here a two-class classifier that assigns to each observation x_k, in a multi-dimensional space, a predicted binary label y_n in {0, 1}. In practice the prediction is based on the thresholding of the result of a decision function $s_n = f(x_n)$, which we will call the classifier score. The calibration problem consists in finding a transformation $p_n=g(s_n)$ of these scores into a value in the interval [0,1] such that the result can be interpreted as the probability $p_k=P(y_k=1|x_k)$ that of a true positive conditioned on the observed sample. Therefore, the values need to have reasonable properties to be used in a fusion approach with other sources of information.

Some simple forms of calibration rely on ad-hoc approaches [57] to associate probabilities or on simple normalization of the values to the [0, 1] interval.

A more principled approach is to adopt a Bayesian point of view with a classifier that directly produces probabilities according to a generative model. Such classifiers can be used in a layered fashion, such as a final Bayesian classifier taking the scores of the first layer classifier as input, and producing the likelihoods $p(x_k|y_k=1)$ and $p(x_k|y_k=0)$ of a sample x_k to be positive or negative, and using Bayes rule to infer the *a posteriori* probability

$$P(y_k = 1 | x_k) = \frac{p(x_k \mid y_k = 1)P(y_k = 1)}{p(x_k \mid y_k = 1)P(y_k = 1) + p(x_k \mid y_k = 0)P(y_k = 0)}$$

Another approach is to train the mapping from scores to probability directly. The constraints on the form of the transformation and the criterion used to find the optimal one differentiate the approaches. One of the most common approaches, Platt scaling [51], fits a sigmoid function to the scores.

This shape is the theoretical transformation to use when the distribution of scores are two Gaussians, but may not be optimal in many other cases [46]. A more generic transformation can be trained by binning, which is the assignment of fixed output probabilities to a set of predefined intervals on the score values. The choice of the number of intervals needs to be defined using cross-validation. The more general isotonic regression [73] tries to find a non-parametric *g*, with the only constraint that it is monotonically increasing. It was extended in

[71] to include regularization parameters and multi-classifier fusion. Since the constraint is less stringent than for simpler models, such models are more prone to over fitting that the Platt model for instance [73]. Depending on the application, the performances may therefore vary, depending on the complexity of the classification task and the availability of training data.

These various approaches show that the choice of a calibration approach should be evaluated for the application at hand. We will therefore present the study of the calibration issue for activity recognition in Section 5.

3.3 Ontology-based Reasoning with Imperfect Information

Because the components in a pervasive environment deal with the real world, they come with certain caveats: sensors in the field could break down, or they could report inaccurately because they come up against an unusual phenomenon, i.e. one for which they have not been designed.

Since these issues must be taken into account when dealing with multi-sensor systems, it should be possible to describe the concepts of accuracy, uncertainty and provenance with respect to sensed data and represent them as part of its ontological description. With these descriptions in place, particular reasoning mechanisms on ontologies need to be designed to support efficient and precise reasoning on the data.

Gaia, a representative of the early works, tries to capture and make sense of the impreciseness and conflicts inherent when dealing with real-world data [53]. An uncertainty model is developed based on a predicate-based representation of contexts and associated confidence values. The predicate structure and semantics are specified in ontologies that benefit in: (a) checking the predicates' validity; (b) simplifying the definition of context predicates in rules; (c) facilitating interoperation between different systems; and (d) further reducing the possibility of uncertainty when interpreting context information. To reason about uncertainty, Gaia employs mechanisms such as probabilistic logic, fuzzy logic, and Bayesian networks, each of which is advantageous under different circumstances. For instance, Bayesian networks are used for identifying causal dependencies (represented as edges) between different events (represented as nodes). The networks are trained with real data, in order to get more accurate probability distributions for their event nodes.

This above approach uses ontologies syntactically as a vocabulary for exchanging knowledge base specified in a probabilistic model. Responding to the need for modelling imperfect knowledge in the Semantic Web, much research has been devoted to extending formalisms and reasoning services, so as to handle uncertain and/or vague information. Representative examples include - among others - fuzzy extensions of DLs [64][66], OWL [5][65] and SWRL [72], and probabilistic extensions such as PR-OWL [9][14] and BayesOWL [21]; for an extensive overview the reader is referred to [67]. Further relevant proposals include a pattern-based approach for representing and reasoning with fuzzy knowledge [69], and a generic, formalised approach for managing uncertainty [22]. Few works, however, have explored the applicability of such initiatives in the domain of pervasive applications; an example is the approach presented in [12], where fuzzy reasoning is used to provide personalised mobile services based on situation awareness.

Missing data is another source of uncertainty when reasoning about context: a missed (or inaccurate) detection of low-level context information may easily lead to irrecoverable failures in the inference of higher-level context abstractions. One possible solution is to model the interpretation of perceptual data as inference to the best explanation using abductive reasoning [49][62]. Romero et al. [31][32] investigate this idea in the context of an ontology-based surveillance application. A set of ontologies is used to capture context at increasing levels of abstractions, including tracking knowledge, scene objects and activities. Once the low-level context acquired from visual sensors is translated into ABox assertions, abductive rules are applied to derive missing facts and trigger the derivation of higher-level context descriptions. No information is provided whatsoever about the computational framework used to implement the abductive reasoning and the preference criteria used for selecting explanations. Acknowledged as a mode of reasoning that is inherent in a plethora of tasks, much research has been devoted to understanding abduction. For a detailed account on the potential of abductive reasoning in DLs see [24][35].

A formal model based on defeasible logic is proposed by Bikakis et al. to support reasoning with imperfect context in ambient computing environments [4]. Extending the Multi-Context Systems model with non-monotonic features, the proposed framework supports reasoning in cases of missing context knowledge. Potential inconsistencies are resolved by means of an argumentation framework that exploits context and preference information that expresses confidence on the contexts considered. The propositional representation of context knowledge may not allow a direct integration with ontology-based context reasoning frameworks; yet possibilities for interesting hybrid architectures emerge, where contextual assertions can be selectively translated into equivalent grounded formulas.

In Section 6 we propose a knowledge-driven framework for activity recognition and fusion, coupling ontology models of abstract domain activity dependencies with a context-aware approach for multi-sensor fusion and monitoring. Our objective is to provide a lightweight context-aware framework towards handling the intrinsic characteristics of pervasive environments in real-world conditions, such as imperfect information, noise, conflicts or inaccurate temporal correlations. We formalise activity dependencies, capitalising upon the Situation conceptualisation of the DnS ontology pattern in DUL [28] for defining generic context descriptors, whereas activity segmentation and recognition is reduced in linking and classifying meaningful contextual segments.

4 Uncertainty Modelling for Low-Level Event Detection

A constraint-based approach following an ontology language allows a straightforward modelling of events by domain experts as it uses natural terms. Nevertheless, the deterministic nature of its constraints makes the models susceptible to noise from its underlying components (e.g., people detection and tracking components in a pipeline of computer vision system). Additionally, model-based frameworks are sensitive from any deviation from the defined constraints, which may come even from event intra-class variability.

We present a probabilistic framework to extend the generic constraint-based ontology language for event modelling proposed by Vu et al. [70], which has been extensively evaluated for event detection on older people monitoring domain [74][34][17].

4.1 Constraint-based Approach for Event Detection

The constraint-based framework is composed of a temporal scenario (event) recognition algorithm and an event modelling framework. The event models follow a declarative and intuitive ontology-based language that uses natural terminology to allow end users (e.g., medical experts) to easily add and change event models of a system. The models are built taking into account *a priori* knowledge of the experimental scene, and attributes of objects (herein called Physical Objects, e.g. a person, a car, etc.) detected and tracked by the vision components. *A priori* knowledge consists of the decomposition of a 3D projection of the scene floor plan into a set of spatial zones which carry semantic information about the monitored scene (e.g., zones like "TV", "armchair", "desk", "coffee machine"). The temporal algorithm is responsible for the inference task, where it takes as input low-level data from underlying vision components, and evaluates whether these objects (or their properties) satisfy the constraints defined in the modelled events.

An event model is composed of (up to) six components [70]:

- **Physical Objects**, refer to real objects involved in the recognition of the event modelled. Examples of physical object types are: mobile objects (e.g. person herein, or vehicle in another application), contextual objects (equipment) and contextual zones (chair zone).
- **Components** refer to sub-events of which the model is composed.
- **Forbidden Components** refer to events that should not occur when a certain event model is recognized.
- **Constraints** are conditions that the physical objects and/or the components should hold. These constraints could be logical, spatial and temporal.
- **Alert** describes the importance of a detection of the scenario model for a given specific treatment, and
- **Action** in association with the Alert type describes a specific action which would be performed when an event of the described model is detected (e.g. send a SMS to a caregiver responsible to check a patient over a possible falling down).

The physical object type depends on the domain in which the event modelling is been applied and may be expanded accordingly. Three basic types are defined by default: Person, Contextual Zones and Contextual Objects. Person type is an extension of a generic type called Mobile, which defines basic information (e.g. 3D position, width, height) that mobile objects should have. Examples of “Person” type attributes are body posture and appearance signature(s). Contextual Zone and Object types refer to *a priori* knowledge on the scene (e.g. contextual zone mostly refers to environment furniture). Constraints define conditions that physical object properties and/or components must satisfy. They can be non-temporal, such as spatial and appearance constraints; or temporal such as the time ordering between two sub-events (components). For instance, the model $Person\_changing\_from\_Zone1\_to\_Zone2$ was modelled as $Person\_in\_zone1\ BEFORE\ Person\_in\_zone2$. Temporal constraints are expressed using Allen's interval algebra (e.g., BEFORE, MEET, and AND) [1].

The ontology hierarchically categorizes event models according to their complexity as follows (in ascending order):

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

4.2 Uncertainty Handling

Uncertainty is present at different levels of the event modelling task, from the uncertainty on the low-level data used as input for the task (e.g. values of attributes of people detected in the scene) to the event models themselves (e.g. person-to-person differences when performing an event, called event intra-class variation). It would be desirable that an event model handles small deviations of defined constraints due to both event intra-class variability and low-level noise. Finally, uncertainty may also come from a semantic gap between the event model and the real event, for instance, on cases where the model is based on a correlated but indirect measurement of the targeted real-world event.

We propose a probabilistic framework to handle uncertainty from low-level data. The framework builds conditional probability distributions (CPD) for each constraint of an event model using a learning step. The CPD are then used on an inference step to handle small deviations of model constraints. Low-level uncertainty is associated to event models, herein called Elementary Scenarios, as their constraints are mostly based on low-level data.

4.2.1 Notation

We have grouped the event model types of the constraint-based ontology into two categories to treat their uncertainty: elementary and composite scenarios. The term *scenario* is used to differentiate the modelling task from the inference task. Elementary Scenario corresponds to the type primitive state of the ontology, while Composite Scenario corresponds to all other event types. The framework for uncertainty modelling follows the concepts below:

- **Elementary Scenario (ES)** is composed of physical objects and constraints. The constraints are related to instantaneous values (e.g. current frame) of physical object attribute (s).
- **Composite Scenario (CS)** is composed of physical objects, sub-scenarios (components) and constraints (generally on components). The constraints generally refer to temporal relations among sub-scenarios or to a sub-scenario itself.

- **Constraint** is a condition that physical object(s) or sub-scenarios must satisfy. They are categorized into two types: non-temporal and temporal.
- **Attributes** correspond to the properties (characteristics) of real world objects measured by the underlying vision components.
- **Observation** corresponds to the amount of evidence regarding a scenario model.
- **Instance or Solution** refers to an individual detection of a given scenario.

4.2.2 Elementary Scenario Uncertainty

The uncertainty of an Elementary Scenario is formalized as function of its constraints. Equation (1) presents its uncertainty using Bayes Rule.

$$p(E_i\,|C_i) = \frac{p(C_i|E_i) * p(E_i)}{p(C_i)} \tag{1}$$

where,

$P(E_i\,|C_i)$ = Conditional Probability of Event E_i given its observed Constraints C_i;

$P(C_i)$ = Probability of constraints which intervene on E_i at the current frame; and

$P(E_i)$ = Prior Probability of Event E_i.

The Conditional Probability of Event E_i given its set of observed Constraints C_i is obtained by the multiplication of the conditional probability of all constraints of E_i. Consequently, this approach assumes that all constraints contribute equally to the event model probability (2).

$$P(C_i|E_i) = \prod_{C_{j,i}\,\in\,C_i}^{N_j} P\left(C_{j,i}\middle|E_i\right) \tag{2}$$

where $C_{j,i}$: Conditional probability of Constraint *j* of given event *i*.

In the case of Elementary Scenarios, most constraints are a function of low-level data with respect to the involved physical objects, referring to a specific or a range of values an attribute of single physical object assumes, or to a relation amongst the attributes of two or more of them (e.g. person position inside a zone). Therefore, we learnt the conditional probabilities of the constraints on a training step using the event models and annotated video recordings both provided by domain experts.

The prior probability of Elementary Scenarios is assumed to be 1.0 strengthening their dependency on the conditional probabilities at the current time, and also establishing that all ES are equally probable.

In order to avoid computing $P(C_i)$ which can be costly as the number of constraints of a scenario increases, we opted to use the unnormalized probability of $P(E_i\,|C_i)$, $\tilde{P}(E_i\,|C_i)$. Equation (3) presents the final form of the equation for probability computation of Elementary Scenario. To obtain a probability value in the interval [0, 1] we restrict the framework to Conditional probabilities with co-domain range in the interval [0, 1].

$$\tilde{P}(E_i\,|C_i) = \prod_{C_{j,i}\,\in\,E_i}^{N_j} P\left(C_{j,i}\middle|E_i\right) \tag{3}$$

Equation (3) addresses both small deviations on attribute values of physical objects constraints due to noisy from underlying components, and by consequence, event intra-class variations (e.g. caused by person-to-person differences on executing the targeted events). Basically, the proposed approach quantifies the confidence on the constraint satisfaction, and then propagates it to high-level scenarios.

4.2.3 Learning Constraint Conditional Probabilities

The probability of a constraint was addressed by associating it to a Probability Density Function (PDF) which quantifies how likely is the constraint to be satisfied given the current evidence on the event (e.g. the current value of the attribute Height of the physical object involved in the constraint for bending posture detection). The adoption of PDFs provides a flexible way to model the uncertainty process that governs the probability distribution of a given constraint as it decomposes in a modular fashion the complexity of elementary scenario models. Different elementary scenarios may then have different PDFs according to the low-level data they are conditioned on. The use of PDFs also avoids the need of fully specifying all the possible assignments of the conditional probability table of a given scenario model.

Figure 4-1 presents an example of an elementary scenario, the model $Person\_in\_Zone\_Tea$. This scenario is based on the physical objects $Person$ and the semantic zone $zoneTea$ (e.g. a polygon drawn on the ground close/around the table where there were kitchen tools to prepare tea). The model has only two constraints: the relational constraint which enforces the involved zone to be $zoneTea$; and a spatial operator called *In* which verifies whether the person position lies inside a given zone.

```
ElementaryScenario(Person_in_Zone_Tea,
  PhysicalObjects( (per:Person), (zT: Zone) )
  Constraints(
           (per->Position In zT->Vertices)
           (zT->name = "zoneTea")
  ))
```

Figure 4-1. Elementary Scenario Person_in_zone_Tea

The deterministic version of the operator *In* is susceptible to different kinds of uncertainty. Firstly, malfunction of low-level computer vision components may deviate the estimated person position from its actual position. Likewise, the semantic zone $zoneTea$ which is *a priori* defined by an expert may not accommodate the complete floor surface where people may stand to prepare tea (e.g. by accommodating only the front face of the table and not its corners). The above-mentioned cases will certainly invalidate the detection of the event model if a deterministic approach is in use.

To address such cases, we propose and evaluate two alternatives for the deterministic *In*. The goal is to devise an alternative that quantifies how likely is the person position to be inside (or how close to) the zone. Briefly, a high probability would be assigned to the constraint when the person is inside the zone, and as the distance of the person from the zone increases, the given probability is proportional to this distance among objects.

The first operator, called fully probabilistic *In*, defines a PDF with respect to the relative distance between the projected centroid of the person onto the floor and the semantic zone centre. The PDF converts the observed distance among objects into a uniform Gaussian distribution and then applies its result to an exponential function. Briefly, this function provides a probability curve with maximum value around the mean distance of the person to the zone, with a monotonically decreasing behaviour as the observed values distance from the mean.

$$P\left(C_{j,i}\right) = e^{\frac{1}{2} * \frac{observed\ value - \bar{x}^2}{s}} \qquad (4)$$

where $\bar{x}$: sample mean and *s:* standard deviation of $\bar{x}$.

The second operator, hybrid probabilistic *In*, is a hybrid approach. It provides maximum probability (100%) when the person is anywhere inside the defined semantic zone (same as the deterministic approach), and the proportional probability is given when the person is outside (Equation (4)). Differently from the fully probabilistic *In*, the distance used to computed the proportional probability is the relative distance between the person centroid (also projected onto the floor) and the *closest edge* of the zone polygon (adapted from [56]).

The Gaussian distribution parameters of each operator are computed by a learning step based on deterministic event models using the constraint-based ontology and RGB-D recordings fully annotated in terms of events.

The learning step proceeds as follows: firstly, an event detection process is performed using the deterministic event models. For each time the deterministic *In* is evaluated, the distance between person and the semantic zone is kept. Secondly, using the event annotation we collect the values frequently assumed by the deterministic *In* when event annotation is present (independent of constraint satisfaction). Thirdly, we compute $\bar{x}$ and s parameters for each contextual zone.

By using event models combined with event annotation (both provided by domain experts) in the learning step, it allows us to capture not only the true distribution of the low-level parameters, but also to prune parameter values according to the event model semantics.

4.3 Evaluation

Evaluation was performed using RBG-D recordings of the pilot@lab of Dem@care project. Each RBG-D recording corresponded to a patient recording. The number of recordings in each evaluation varied from 4 to 49 according to the hypothesis in test.

4.3.1 Performance Measurement

The prototype accuracy was evaluated using the indices of recall and precision described in Equations (5) and (6), respectively, using as ground-truth events annotated by domain experts.

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$Precision = \frac{TP}{TP + FP} \qquad (6)$$

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

4.3.2 Dataset of RGB-D Recordings

Participants of 65 years and over were recruited by the Memory Center (MC) of a collaborating Hospital. Inclusion criteria of the Alzheimer Disease (AD) group are: diagnosis of AD according to NINCDS-ADRDA criteria and a Mini-Mental State Exam (MMSE) [26] score above 15. AD participants who have significant motor disturbances (per the Unified Parkinson’s Disease Rating Scale) are excluded. Control participants are healthy in the sense of behavioural and cognitive disturbances. The clinical protocol asks the participants to undertake a set of physical tasks and Instrumental Activities of Daily Living in a Hospital observation room furnished with home appliances [40][36]. Experimental recordings used a RGB-D camera (Kinect®, Microsoft©).

The activities of the clinical protocol are divided into three scenarios: Guided, Semi-guided, and Free activities. Guided activities (10 minutes) intend to assess kinematic parameters of the participant gait profile (e.g. static and dynamic balance test, walking test):

- Balance testing: the participant should keep balance while performing actions such as keeping both feet side by side stand, standing with the side of the heel of one foot touching the big toe of the other foot.
- Walking Speed test (WS): the assessor asks the participant to walk through the room, following a straight path from one side of the room to the other (chair side to video camera side, outward attempt, 4 meters), and then to return (return attempt, 4 meters).
- Repeated Transfer test: the assessor asks the participant to make the first posture transfer (from sitting to standing posture) without using help of his/her arms. The assessor will then ask the participant to repeat the same action five times in a row, and
- Time Up and Go test (TUG): participant start from the sitting position, and at the assessor's signal he/she needs to stand up, to walk a 3 meters path, to make a U-turn in the centre of the room, return and sit down again.

Semi-guided activities (15 minutes) aim to evaluate the level of autonomy of the participant by organizing and carrying out a list of instrumental activities of daily living (IADL) within 15 minutes. The participant is alone in the room with the list of activities to perform, and he/she is advised to leave the room only when he/she has felt the required tasks are completed.

- Watch TV,
- Make tea/coffee,
- Write the shopping list of the lunch ingredients,
- Answer the Phone,
- Read the newspaper/magazine,
- Water the plant,
- Organize the prescribed drugs inside the drug box according to the daily/weekly intake schedule,
- Write a check to pay the electricity bill,
- Call a taxi,
- Get out of the room.

For this work, we have focused on recordings of patients performing the semi-guided activities.

4.3.3 RGB-D Monitoring System

This approach has been evaluated using a RGB-D monitoring system which builds on the event detection framework proposed by [17]. In the base work a hierarchical model-based approach is used to model and detect activities of daily living. The monitoring system is composed of three main steps: people detection, people tracking, and event detection.

For the present evaluation, people detection step is replaced by the depth-based algorithm proposed in Nghiem et al.[45], since we have used a RGB-D sensor instead of a 2D-RGB camera. The depth-based algorithm performs as follows: first, background subtraction is employed on the depth image provided by the RGBD camera to identify moving regions. Then, region pixels are clustered in objects based on their depth and neighbourhood information. Finally, head and shoulder detectors are employed to detect people amongst other types of detected objects.

The set of detected person in the analysed frame is then passed to the tracking step. This module is based on a multi-feature algorithm proposed in Chau et al. [10], and it employs features such as 2D size, 3D displacement, colour histogram, and dominant colour to discriminate amongst tracked people.

Event detection step is based on the deterministic event modelling framework (described in Section 4.1) and the temporal algorithm proposed by Vu et al. [70] and evaluated for older people monitoring in Crispim-Junior et al. [17]. This step analyses whether the set of 3D tracked people satisfies the constraints defined into the set of event models in use.

The next section provides a comparison of the uncertainty modelling framework proposed here to the deterministic event modelling approach. Both frameworks were evaluated under the same configuration of the underlying components (e.g. people detection and tracking).

4.4 Results and Discussion

The proposed approach results are presented for two event detection configurations: firstly, primitive state (elementary scenario) detection; and secondly, composite event detection with the uncertainty modelling framework as basis for elementary scenario. Although we focus on uncertainty modelling for low-level events, we have evaluated the improvement brought for the deterministic composite scenario level when using as basis the low-level uncertainty modelling. Table 4-1 presents the results on primitive state detection. “Deterministic” stands for the standard constraint-based approach; “FP In” for the fully probabilistic version of the spatial operator In; and “Hybrid In” to the mixed version between “FP In” and the Deterministic”. These results correspond to 10 videos fully annotated in terms of elementary scenarios in a 3-fold cross-validation scheme. Results are reported as the average performance of the approaches in the three validation sets of 3-fold cross-validation scheme.

Table 4-1. Average Performance of Primitive State Detection on 3-fold-cross-validation

	Deterministic		**Hybrid In**		**FP In**	
IADL	**Rec.**	**Rec.**	**Prec.**	**Prec.**	**Rec.**	**Prec.**
Inside zone Phone	85.23 ± 15.00	90.00 ± 17.32	70.90 ± 11.91	88.89 ± 11.11	67.62 ± 29.32	96.29 ± 6.41
Inside zone Tea	100 ± 0.00	100.00 ± 0.00	38.81 ± 27.07	58.65 ± 25.63	100.00 ± 0.00	55.63 ± 29.45
Inside zone Pharmacy	100 ± 0.00	100.00 ± 0.00	75.55 ± 21.43	83.33 ± 28.86	100.00 ± 0.00	70.45 ± 32.06

Inside zone Plant	100 ± 0.00	100.00 ± 0.00	14.12 ± 2.10	61.66 ± 37.53	100.00 ± 0.00	24.58 ± 7.84
Inside zone Reading	86.66 ± 23.10	100.00 ± 0.00	18.26 ± 2.43	45.08 ± 15.77	100.00 ± 0.00	26.40 ± 5.11

Rec.: Recall, Prec.: Precision

The two proposed probabilistic constraints (operators) outperformed the deterministic approach in the recall index for person inside zone reading. The *Hybrid In* outperformed all other approaches on the recall index. The remaining events have been detected with equal performance rate. The deterministic approach still presents the highest performance in the precision index for most of the primitive states, except on Inside Zone phone event. In this case, *FP In* constraint outperformed the other approaches.

Lower precision values are observed in at least two events for the best performing probabilistic approach (FP In). These low precision values are mostly due to the fact that event models were not posing any restriction regarding whether a person is static or not, walking closely to the zone, or having any filtering step. These restrictions would be part of higher-level models built on the primitive states reported. Table 4-2 presents the results on the use of uncertainty modelling for elementary scenarios as a basis for a deterministic detection of composite events. To accomplish this, we have performed the learning step on 10 videos annotated with elementary scenarios (N: 10), and then evaluated the proposed approach performance in a second dataset composed of 45 recordings annotated only in terms of Composite Events.

Table 4-2. Extended Evaluation of Framework Performance on Composite Event Detection

	Deterministic		**Hybrid In**		**FP In**	
IADL	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**	**Prec.**
Using Phone	89.65	86.66	91.86	71.81	88.50	81.05
Preparing Tea/ Coffee	93.93	65.95	98.46	35.36	98.48	49.24
Using Pharmacy Basket	95.45	93.33	100.00	86.53	100.00	86.27
Watering plant	88.88	70.58	100.00	19.58	100.00	20.45
Average Performance	91.97	79.13	97.58	53.32	96.74	59.25

N: 45 participants, 15 min. each.

In this extended evaluation (45 participants), the probabilistic constraints have outperformed the deterministic approach in the recall index on all cases. The only exception was the *FP In* on the detection of the activity *using phone*. For this activity, *FP In* had a slightly smaller precision than the deterministic approach, but higher than the one presented at the detection of primitive states.

Concerning precision index, the deterministic approach has still the highest values. From the two probabilistic constraints, *FP In* has shown the highest precision. For instance, *Hybrid In* has improved its precision for the watering plant activity, but this activity remained with the worst performance on both probabilistic approaches. This low precision may be explained by the short length of this activity in association with its closeness to the activity Preparing Tea. Nevertheless, new methods will be investigated to model conditional probabilities of constraint whose low-level data may not follow a univariate Gaussian distribution.

4.5 Conclusions

We have presented a framework for uncertainty modelling of low-level events (elementary scenarios). The results show that the uncertainty framework has increased the recall index of the event detection of elementary scenario by handling uncertainty on constraints based on low-level data.

The benefit of the framework has been further demonstrated in the second evaluation which was performed on a larger dataset with the detection of high-level (composite events) instead of primitive states. In this second evaluation there was a greater difference in recall index values between the probabilistic approaches and the deterministic ones. Despite the improvement brought by the probabilistic approach (as *FP In*), further development is being performed to improve the precision. For instance, new methods to model conditional probabilities are under investigation to model constraints based on low-level data, for which Gaussian-based models are not appropriate. Concerning the overall framework, the uncertainty modelling will be also further extended to handle high-level scenarios by modelling time constraints among models.

The proposed uncertainty framework extended the deterministic constraint-based framework in use for the modelling of activities of daily living of the CAR component. Currently, a supervised learning step is necessary to learn the conditional probabilities associated to the event model constraints. This step may be *a priori* addressed during the installation of the system and still remain transparent to the end user that will be in charge of designing the event models. But, in order to ease the system deployment, we intend to also investigate possible alternatives to assess the confidence on constraint satisfaction without the need of the learning step.

5 Probabilistic Activity Recognition and Confidence Values

This contribution focuses on the management of uncertainty when dealing with the output of recognition modules such as activity recognition from the wearable camera, taken as input to the behaviour interpretation modules. The review of related work was presented in Section 3.2. This section details a study that evaluates the various calibration approaches for classifiers output. The section is organized as follows:

- In Section 5.1, we present the recognition framework from which the prediction scores are obtained and discuss its suitability to score calibration.
- In Section 5.2, we present in detail the calibration approaches we have considered.
- In Section 5.3, we present the experimental setup and results associated to the problem of calibrating the probabilities

5.1 Description of the Baseline Recognition Framework

For this study, we focus on the recognition of Instrumental Activities of Daily Living (IADLs) by analysing human-object interactions and also the contextual information surrounding them. We will consider the example of the supervised activity pattern recognition approach proposed in Section 4 of deliverable D5.3, which merges the information from an ensemble of specialized classifiers. Our hierarchical approach has two connected processing layers. The first layer contains a set of Active Object (AO) detectors and Visual Place (VP) detector. An AO is an object with which the subject/patient wearing the camera interacts. Here the interaction is understood as manipulation or observation. A VP is defined as a semantic place in which the patient is standing, which can be either a room (kitchen) or a more specific place such as “in front of the sink”. The second layer uses the outputs of the first layer to perform the activity recognition task. In the following sections, we introduce the complete pipeline.

5.1.1 Active Object Recognition

In general, we consider one individual detector for each object category although the nonlinear classification stage is the only step that is specific for each category. We have built our model on the well-known Bag-of-Words (BoW) paradigm [18] and added saliency masks as a way to provide spatial discrimination to the original Bag-of-Words approach. Hence, for each frame in a video sequence, we extract a set of N SURF descriptors d_n [3], using a dense grid of circular local patches. Next, each descriptor d_n is assigned to the most similar word $j=1…V$ in a visual vocabulary by following a vector-quantization process. The visual vocabulary, computed using a k-means algorithm over a large set of descriptors in the training dataset (about 1M descriptors in our case), has a size of V=4000 visual words.

In parallel, our system generates a geometric-spatio-temporal saliency map S of the frame with the same dimensions of the image and values in the range [0, 1] (the higher the value, the more salient a pixel is). Details about the generation of saliency maps can be found in [6].

We use this saliency map to weight the influence of each descriptor in the final image signature, so that each bin j of the BoW histogram H is computed following the next equation:

$$H_j = \sum_{n=1}^{N} a_k w_{nj}$$

where the term w_{nj}=1 if the descriptor or region n is quantized to the visual word j in the vocabulary, and zero otherwise, and the weight α_n is defined as the maximum saliency value S found in the circular local region of the dense grid. Finally, the histogram H is L1-normalized in order to produce the final image signature.

Once each image is represented by its weighted histogram of visual words, we use a SVM classifier [13] with a nonlinear χ^2 kernel, which has shown good performance in visual recognition tasks working with normalized histograms such as those ones used in the BoW paradigm [63].

Since there are multiple classes, we adopt a 1-vs-all approach where one SVM detector is trained for each specific class. Applying this procedure to the test frames produces the raw multi-class prediction scores: a vector is associated to each frame, where each coefficient corresponds to the score of the frame with respect to a single class detector.

5.1.2 Visual Place Recognition

The general framework can be decomposed in three steps. First of all, for each image, a global image descriptor is extracted. We choose the Composed Receptive Field Histograms (CRFH) [52] since it was proven to produce good performances for indoor localization estimation [23]. Then a non-linear dimensionality reduction method is employed. In our case, we use a Kernel Principal Component Analysis (KPCA) [61]. The purpose of this step is twofold: it reduces the size of the image descriptor, alleviating the computational burden of the rest of the framework, and it provides descriptors on which linear operations can be performed. Finally, based on these features, a linear Support Vector Machine (SVM) [13] is applied to perform the place recognition, and the result is regularized using temporal accumulation [23].

In a similar way as for AO detection, a 1-vs-all approach is used to produce the raw multi-class prediction scores.

5.1.3 Activity Recognition

Before injecting the scores of AO and VP into the second layer of the activity recognition the scores need to be normalized. This ensures that the scores associated to the various objects and various sources (AO and VP) have a comparable scale before merging them.

We use the Platt approximation [51] to produce posterior probabilistic estimates O_k^t and $P_{\mathrm{j}}^{\mathrm{t}}$ for the respective occurrences of the object and places of class k and j in the frame t.

Our activity recognition module uses the temporal pyramid of features presented in [50], which allows exploiting the dynamics of user's behaviour in egocentric videos. However, rather than combining features for active/non-active objects, we represent activities as sequences of AOs and VPs. For instance, cooking may involve the user's interaction with various utensils whereas cleaning the house might require a user to move around various places of the house.

In particular, for each frame t being analysed, we consider a temporal neighbourhood Ω_{t} corresponding to the interval [t-Δ/2, t+ Δ/2]. This interval is then iteratively partitioned into two sub-segments following a pyramid approach, so that at each level l=0...L-1 the pyramid contains 2^l sub-segments. Hence, the final feature of a pyramid with L levels is defined as:

$$F_t = \left[F_t^{0,1} \dots F_t^{l,1} \dots F_t^{l,2^l} \dots F_t^{L-1,2^{L-1}}\right]$$

where $F_t^{l,m}$ represents the feature associated to the sub-segment *m* in the level *l* of the pyramid and is computed as:

$$F_t^{l,m} = \frac{2^l}{\Delta} \sum_{s \in \Omega_{tm}^l} f_s$$

Where Ω_{tm}^l represents the *m* temporal neighbourhood of the frame *t* in the level *l* of the pyramid and f_s is the feature computed at frame *s* in the video. In the experimental section, we will assess the performance of our approach using the outputs of *K* object detectors [O_1^s ... O_K^S], the outputs of *J* place detectors [P_1^s ... P_J^s], or the concatenation of both, as features f_s.

In this work, we have used a sliding window method with a fixed window of size Δ and a pyramid with *L*=2. Finally, the temporal feature pyramid has been used as input for a linear multiclass SVM in charge of deciding the most likely action for each frame.

The multiclass SVM is again trained in a 1-vs-all fashion. For the baseline, its scores have not been calibrated, as they are used to only detect the best activity, by selecting the activity with the highest raw score. Therefore, in order to feed the rest of the WP5 inference system with events with probabilistic interpretation of occurrence, the score need to be calibrated.

5.2 Prediction score calibration

In the following, we will consider methods from different approaches. The simplest approach consists in normalizing the scores to [0, 1] using a sigmoid:

$$g(s) = \frac{1}{1 + \exp(s)}$$

Concerning direct calibration approach, we have used the Platt calibration method [51]:

$$g(s) = \frac{1}{1 + \exp(As + B)}$$

The real coefficients A and B are estimated by fitting the sigmoid *g*(*s*) to modified targets t_i:

$$t_i = \begin{cases} \dfrac{N_+ + 1}{N_+ + 2} & \text{if } y_i = +1 \\ \dfrac{1}{N_- + 2} & \text{if } y_i = 0 \end{cases}$$

where N_+ is the number of positive samples and N_- the number of negative samples. This is done by minimizing:

$$-\sum_{i=1}^{N} t_i \log(g(s_i)) + (1 - t_i)\log(1 - g(s_i))$$

A more general calibration function is given by monotonic functions. Their shape is not parameterized, as they only satisfy $r < s \Rightarrow g(r) \leq g(s)$.

The underlying assumption is that the two-class classifier ranks the samples correctly. Hence calibrating the scores consists in finding the monotonic mapping from score space to probability space. This can be done using isotonic regression and implemented using the efficient Pairwise Adjacent Violators algorithm (PAV) [73].

Finally, we also consider a Bayesian generative model on the classifier scores. The raw scores are used to train a Gaussian Mixture Model (GMM), under two options. For each individual activity class detector, a fitting is done as follows:

- The positive class is fitted with one Gaussian, the negative samples are fitted with one Gaussian. We call this model the Gaussian Model (GM_2_gauss). For a test sample x_k, the likelihoods $p(x_k|y_k=1)$ and $p(x_k|y_k=0)$ are computed using each Gaussian. They are combined using Bayes rule and uniform prior to deduce the posterior $P(y_k=1|x_k)$.
- The positive class is fitted with one Gaussian, the negative samples are divided into their respective activity classes, each of which is fitted with one Gaussian. For the dataset considered 19 classes are defined, therefore 19 Gaussians are trained for each of the 19 activity detectors. We call this model the Gaussian Mixture Model (GMM_19_gauss). For a test sample x_k, the likelihood $p(x_k|y_k=1)$ is computed using the corresponding Gaussian, and $p(x_k|y_k=0)$ is computed as the additive combination of the 18 likelihoods with uniform prior. They are combined using Bayes rule and uniform prior to deduce the posterior $P(y_k=1|x_k)$.

Choosing a different calibration measure for each individual activity detector has an effect on both the quality of the final confidence, but also on the relative ranking of the detections when dealing with multi-class classification. Therefore, a decision based on first ranked class can be influenced by the calibration.

Hence, in the following section, we will compare these approaches from an experimental point of view. We will evaluate both the influence of the calibration on the classification performance, as well as the reliability of the probabilistic scores, and decide which calibration approach is best suited in the specific context of activities recognition from wearable camera.

Automatic activity recognition from wearable camera is a difficult problem so it is very important to be able to assign confidence measures to the predictions, in order to keep the uncertainty under control when feeding the higher level inference. Even though the automatic detection of all possible events is not possible in all cases, computing relevant confidences can mitigate this by trusting the prediction only when the system is confident.

5.3 Experimental evaluation

We have assessed our model in the publicly available IADL dataset, proposed by the authors of [50] that contains videos captured by a chest-mounted GoPro camera on 20 users performing various daily activities at their homes. This dataset was already annotated for 44 object-categories and 18 activities of interest recognition and we have additionally labelled 5 rooms and 7 places of interest.

This dataset is very challenging since both the environment and the object instances are completely different for each user, thus leading to an unconstrained scenario we have in Dem@care project: @Home scenario at DCU in particular. It is therefore well adapted to the objective of evaluating the quality of the transformation of scores to probabilities in a realistic dataset. This is important, as score calibration methods have been usually presented in academic contexts, and our objective here is to evaluate their suitability in the much more complex setup of activity recognition.

For the experiments, the first 6 users have been used for the cross-validation of the classifier parameters and the estimation of calibration parameters, whereas the remaining users (7-20) have been used to train and test the models following a leave-1-out approach.

The raw performances of the OA and VP modules have already been presented in D5.3. In summary:

- Active Object recognition has a 11% Mean Average Precision (AP) with performance varying a lot from one class to the other (ranging from 1% to 54%)
- Visual Place recognition has a 68.4% accuracy, using 7 place classes (in front of the bathroom sink, in front of the washing machine, in front of the kitchen sink, in front of the television, in front of the stove, in front of the fridge and outside)

In this section we show our results in IADL recognition in egocentric videos. Our system identifies the activity at every frame of the video using a sliding window that allows us to compute an average Frame level classification accuracy. For that end, we have also included a new class “no activity/reject” associated to frames that are not showing any activity of interest. It is also worth noting that the global performance is computed by averaging the particular accuracies for each class (rather than simply counting the number of correct decisions) and, thus adapts better to highly unbalanced sets as the one being used (where most of the time there is no activity of interest).

The window-size Δ has been fixed from previous experiments at 1200 frames (40 seconds of video). The fusion of AO and VP was also shown to provide the best results in terms of accuracy and outperform previous works [50] on this dataset. They are fused using a mixing coefficient, α=0.65, computed by cross-validation, as shown in Figure 5-1.

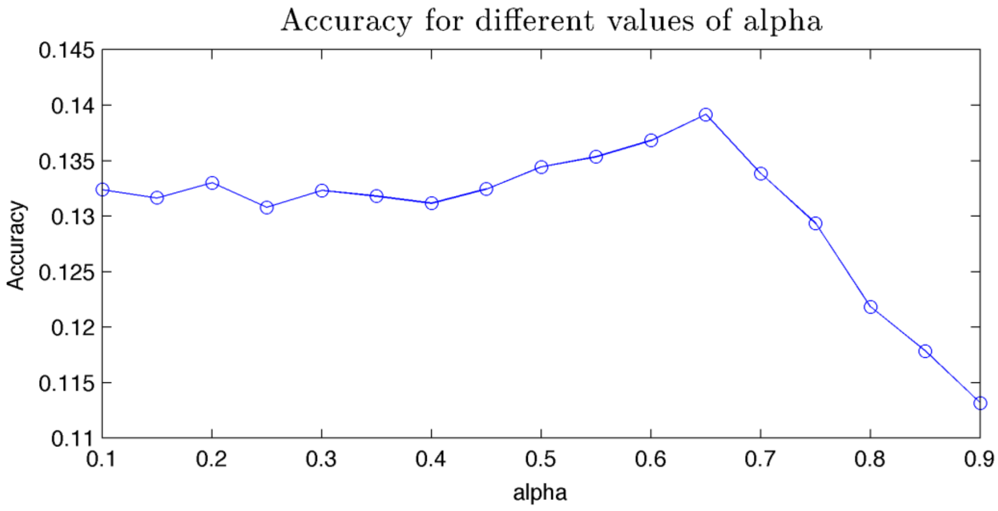


Figure 5-1. Choice of alpha parameter for fusing object and place recognition outputs.

The overall performance on the various approaches considered is presented in Figure 5-2, using classical ROC curve and Precision-Recall (PR) curve. The associated quantitative metrics are compared in Figure 5-3: Area under the ROC curve (AUC) and Mean Average Precision (MAP), which corresponds to the area under the PR-curve. Those results are computed by averaging the individual results for each class.

Under both metrics, the Normalized (ad-hoc sigmoid), PAV and Platt methods perform better than the GM and GMM methods. According to the AUC metric, the best methods are PAV, Normalized, Platt, respectively. According to MAP, the ranking is Normalized, Platt, PAV, respectively.

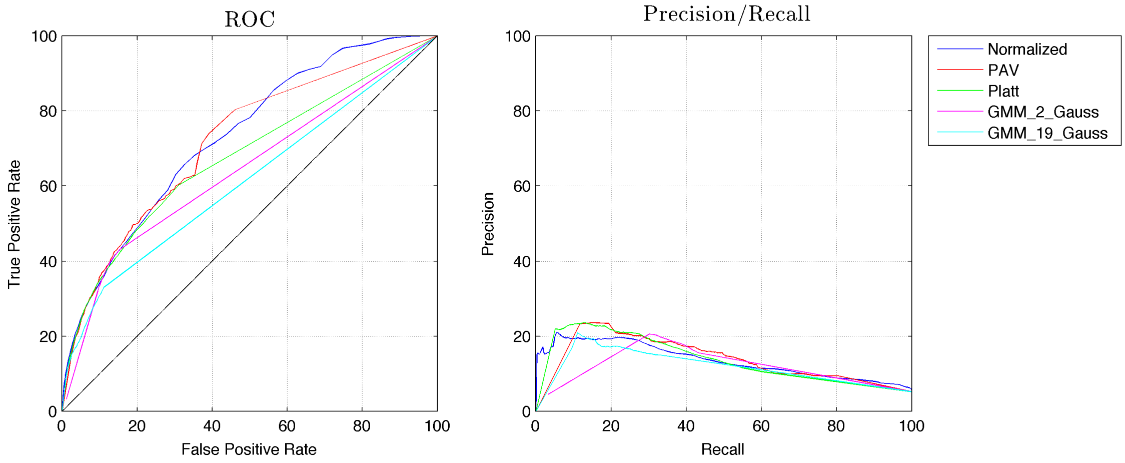


Figure 5-2. Global performance of activity recognition: ROC curve (left) and precision-recall curve (right)

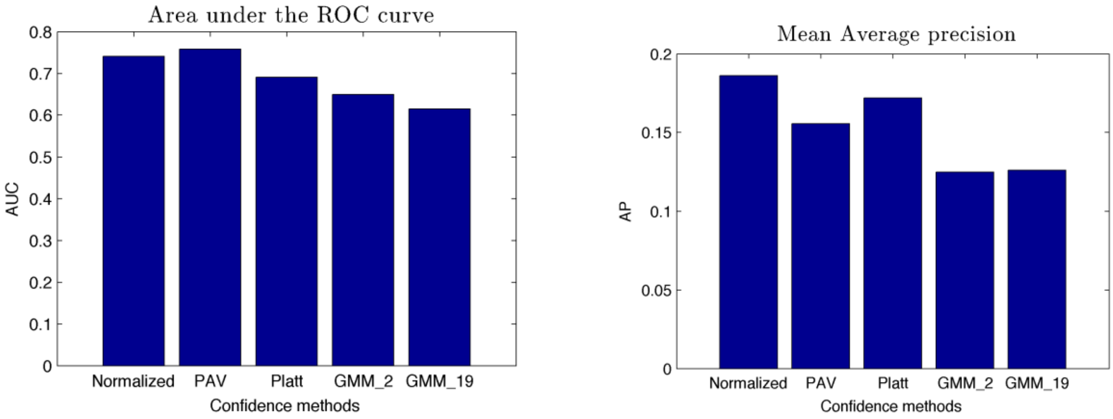


Figure 5-3. Average performance of activity recognition: AUC and AP (average on all classes)

Next, we present the reliability plot of the methods. The graph in Figure 5-4 represents, for each interval of calibrated score (on x-axis), the probability of correct estimation (on y-axis) for all events having a calibrated score in this interval. The interval [0, 1] is divided into 10 subintervals of the form [0.1 k, 0.1 (k+1)]. For all events whose calibrated score belongs to one subinterval, the empirical frequency of correct estimation is computed. When an interval of calibrated score contains very few events, its frequency of correct estimate is not shown. An ideal curve should be on the diagonal, which means that even though all samples cannot be classified correctly, the calibrated score reflects the actual probability that the estimate is correct.

On this plot we notice that apart from the Normalized method, which has a globally fixed score mapping, all other approaches tend to be overconfident on the [0.5, 1] interval. This may be related to a case of over fitting to the training data. In Figure 5-5, the number of occurrences for each class of activity is shown. Many classes have less than 300 events in the dataset, which is then divided in training and testing. To examine this analysis in detail, we will focus on the 4 most frequently occurring classes.

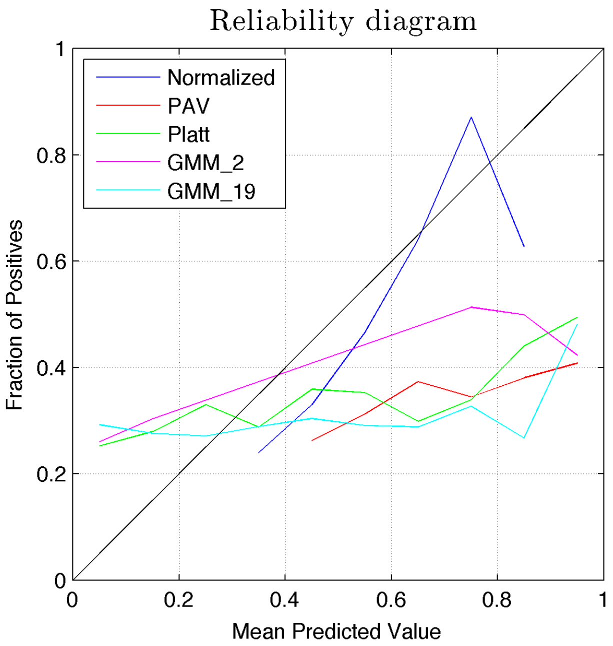


Figure 5-4. Reliability plot (average on all classes)

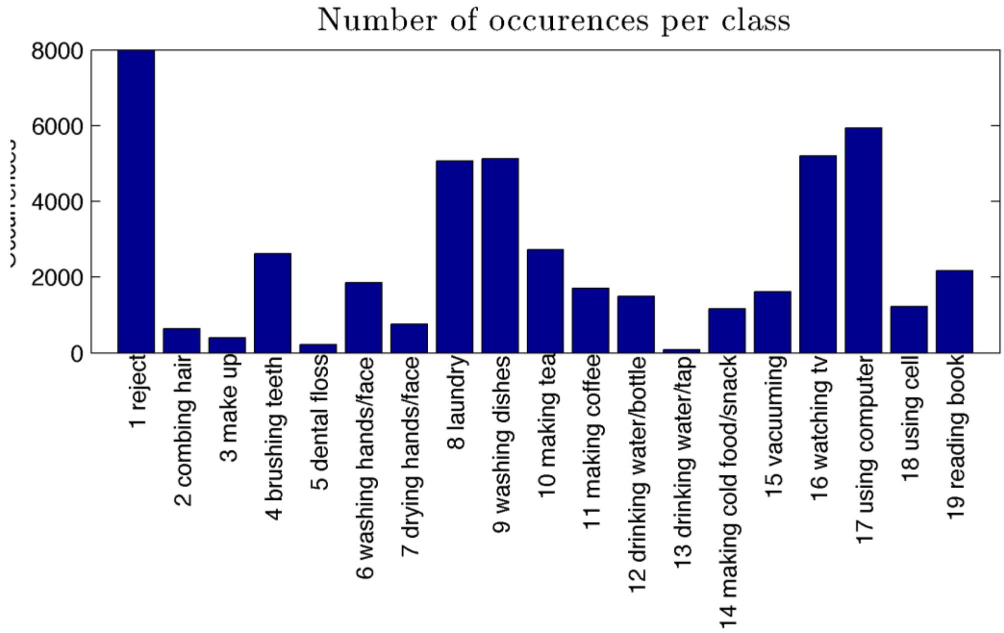


Figure 5-5. Occurrence of classes within the training/testing dataset

In Figure 5-6, the AUC and MAP metrics are shown for the 4 most frequent classes in the dataset. Similar ranking of the methods can be observed as the global analysis. We can nevertheless observe that the class “watching TV” has consistently lower performances than the other classes “laundry”, “washing dishes” and “using computer”. In terms of MAP, the Normalized approach performs best, followed by Platt, then PAV. It is interesting to note that this corresponds to the ordering of the approaches in terms of number of parameters to estimate: more complex models yield lower performance.

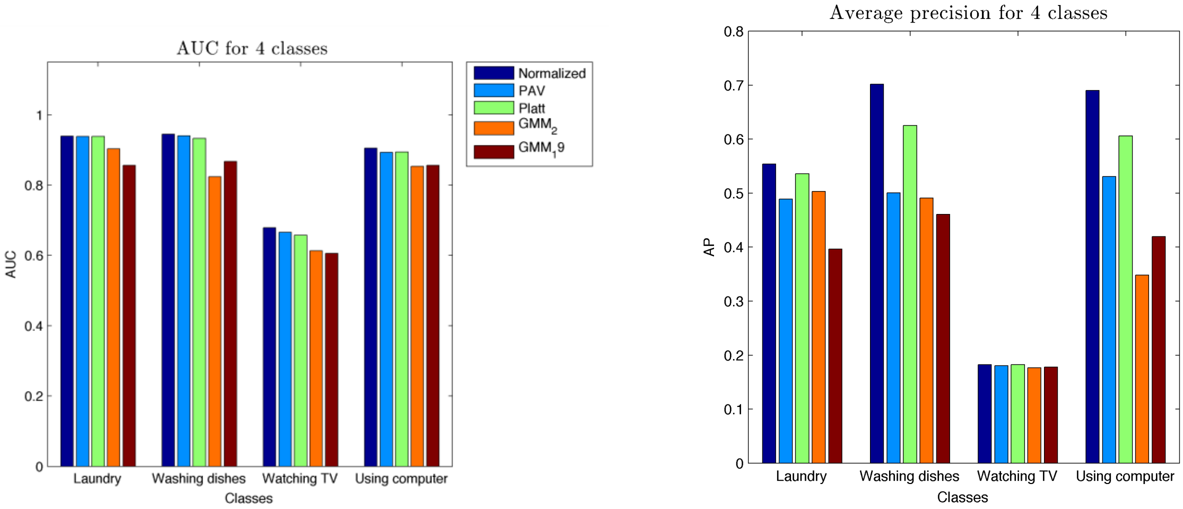


Figure 5-6. AUC and AP for 4 most frequent classes

This analysis is complemented by the reliability plot in Figure 5-7. The poor performance on the “watching TV” shows that this specific class is difficult to recognize in this dataset. On the other hand, on the three other classes, the Normalized approach is over-estimating much the confidence of almost all incorrect predictions. Platt and GMM model seems to have more balanced estimation of confidence, which is closer to a linear curve. Nevertheless, their estimate is almost always over-confident. Finally, the PAV approach is systematically over-confident, which could be associated to over-fitting, due to its very flexible non-parametric model.

Overall this dataset, which is representative of real life activity recognition from wearable camera, is prone to some over-fitting by most models. Therefore simpler models should be preferred. Obtaining a large amount of annotated data seems also an important point to be able to capture the true variability of each class, thus reducing the over-confidence observed here with more complex models.

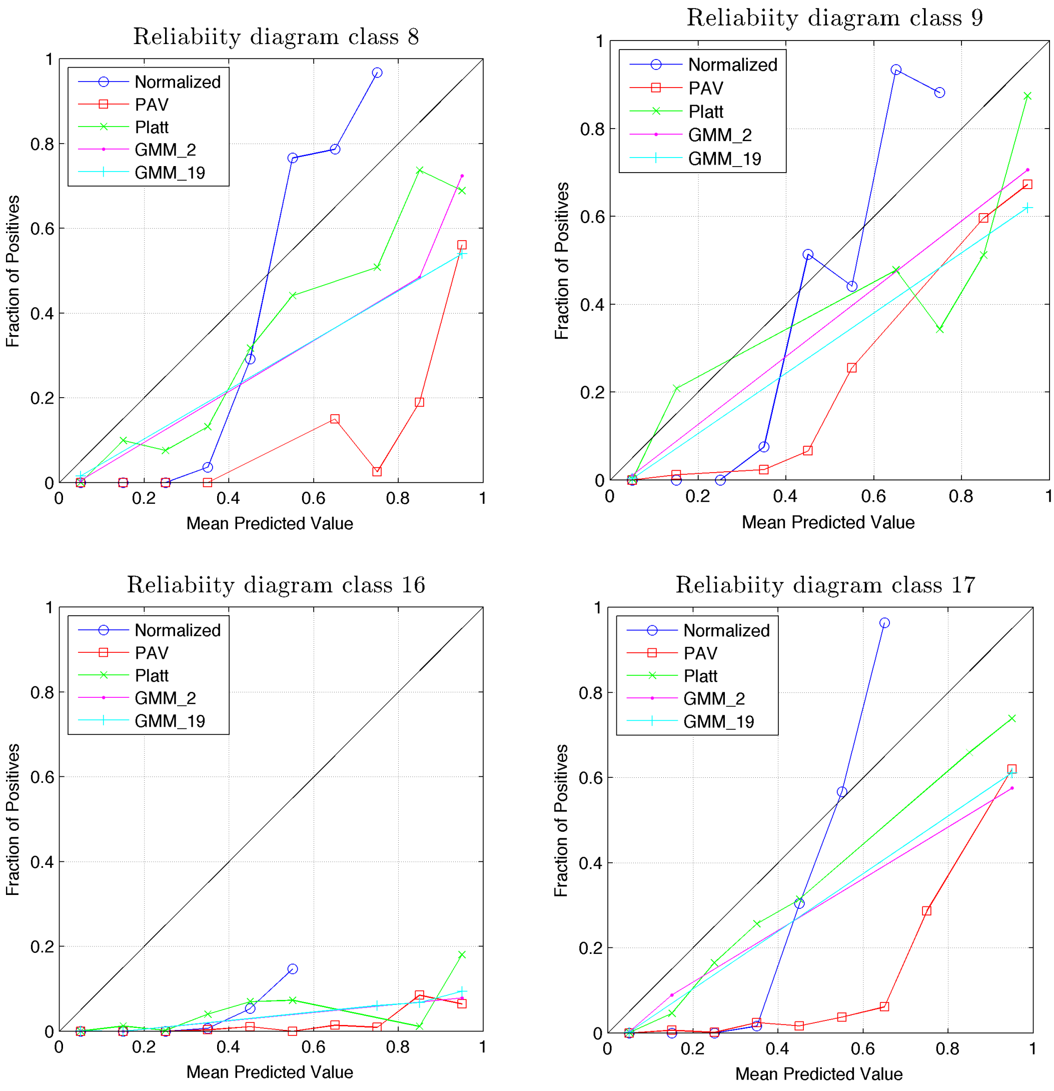


Figure 5-7. Reliability plots for each of the most frequent classes. From left to right: (top-row) laundry, washing dishes, (bottom-row) watching TV, using computer.

5.4 Conclusion

In this chapter, we have evaluated various strategies for probabilistic classifier score calibration in the context of activity recognition. This is a very difficult problem, as noted by the performances of some activities. Nevertheless, some activities show good performance. We have selected 4 classes for detailed study based on the amount of available training data, three of which achieve over 50% AP performance. It turns out that quite good performances are achieved using an ad-hoc normalization, although probabilities are over-estimated in the low probability intervals and underestimated in the higher probability intervals. The reliability diagram shows the best overall reliability for the Platt calibration, both in the low and high probability intervals. Less constrained approaches such as PAV and GMMs yield overall overconfident estimates, which could be interpreted as over-fitting to the training data. For the activity recognition application, we can therefore conclude that given the limited data available for training (as it is also the case in @Home scenario) simpler calibration models should be preferred.

6 Context-based Fusion in Multi-Sensor Environments

Due to the intrinsic characteristics of pervasive environments in real-world conditions, such as imperfect information, noise, conflicts or inaccurate temporal correlations, the use of strict contextual constraints to fuse information from multiple sources is not always a practical and flexible solution. Consider, for example, the rule in Figure 2-2 that fuses four inputs: the $PrepareDrink$ activity that is detected by CAR and three tea-related objects (kettle, tea box and tea bag) detected in WP4 from the wearable camera. In this case, if the video analysis in WP4 fails to detect the tea bag during tea preparation, then the rule would not trigger, and thus, the system will fail to derive the $PrepareTea$ activity. Moreover, many activities are carried out differently even by the same person, e.g. the kettle during the make tea activity may be turned on before or after taking out a cup from the cupboard. Thus, the use of strictly structured background knowledge relevant to the presence and order of activities or their temporal boundaries is not always able to effectively capture and reason about the context.

In an effort to overcome the above limitations, the second version of the multi-parametric interpretation framework introduces a more flexible high-level fusion approach that detects complex situations based on loosely coupled domain activity dependencies rather than on strict contextual constraints. More specifically, given an RDF dataset of observations from WP4 and WP5 components, we define a procedure for assigning *context connections*, i.e. links among relevant groups of observations that signify the presence of complex activities. The connections are determined by semantically comparing *local contexts*, i.e. the type and number of neighbour observations, against context descriptors, i.e. background knowledge about domain activity dependencies. We formalise these descriptors by capitalising on the Situation concept of the DnS (Descriptions and Situations) pattern [29] of the DOLCE+DnS Ultralite (DUL) ontology [28], exploiting the OWL 2 meta-modelling capabilities for defining generic relations among classes.

It should be noted that the context-based fusion approach presented in this deliverable only substitutes the strict complex activity SPARQL rules described in D5.2 (Section 4.2.3, page 38), such as the SPARQL rule in Figure 2-2.

6.1 Domain Context Descriptors

In order to describe the context pertinent to each complex activity in an abstract yet formal way, we reuse the $Situation$ concept of the DnS pattern of DUL. The aim is to provide the conceptual model for annotating domain activity classes with lower-level observation types. Figure 6-1 (a) shows the specialisation of the $Situation$ class, along with two sub-properties of the $isSettingFor$ upper-level property.

Our aim is to define relations among domain activity classes, therefore, the proposed ontology treats classes as instances, allowing property assertions to be made among domain concepts.

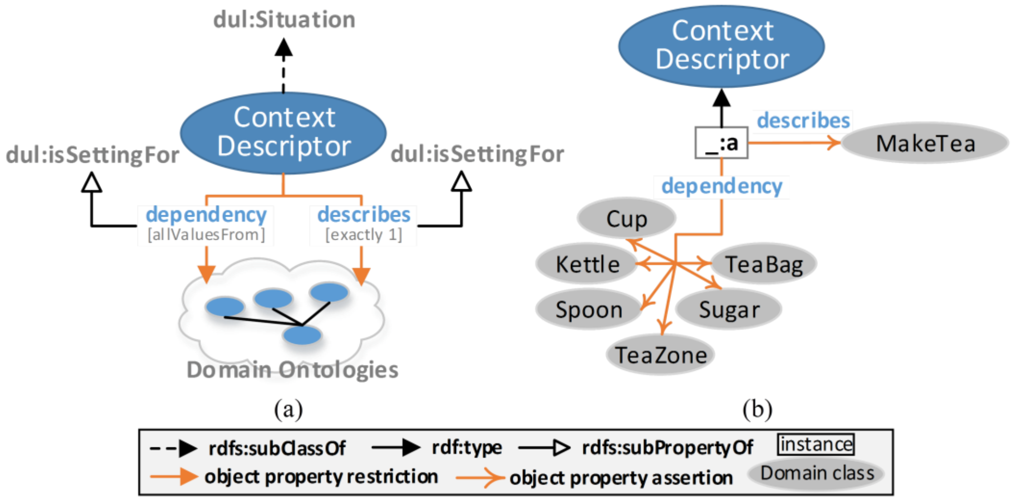


Figure 6-1. (a) The ContextDescriptor class, (b) Example annotation of the MakeTea activity

Intuitively, the ontology can be thought of as a conceptual (meta) layer that can be placed on top of any domain activity ontology. This way, instances of the $ContextDescriptor$ are used to link domain activities ($describes$ property) with one or more lower-level conceptualisations through dependency property assertions. Figure 6-1 (b) presents an example of annotating class $MakeTea$ with class types relevant to objects (e.g. Cup) and location (e.g. $TeaZone$).

The model also allows annotated classes to inherit the context dependencies of the superclasses through the following property chain axiom:

$$describes \circ rdfs{:}subClassOf \circ isDescribedBy \circ dependency \sqsubseteq dependency$$

We use the term "*context descriptor*" to refer to the set of classes, denoted as d_C, that a domain activity C has been annotated with. For example, the context descriptor of $MakeTea$ is denoted as $d_{MakeTea}$ and is equal to the set $\{Cup, Kettle, Spoon, TeaZone, Sugar, TeaBag\}$.

6.2 Context Connections and Activity Recognition

Given a set $O = \{o_1, o_2, \dots, o_n\}$ with RDF instances representing low-level observations, e.g. objects, locations, postures, etc., and a set of domain context descriptors $D = \{d_{C_1}, d_{C_2}, \dots, d_{C_k}\}$, we describe in this section the steps involved in identifying meaningful contexts in O for recognizing higher level activities. The confidence value of an observation is denoted as $P(o_i)$.

6.2.1 Local Contexts

The first step is to define the *local contexts* of each observation $o_i \in O$ that capture information relevant to the neighbour observations of o_i and the most plausible domain activities that o_i can be part of, based on the context descriptors $d_k \in D$ the local context similarity φ.

Definition 1. *A local context* l_i *of an observation* $o_i \in O$ *is defined as the tuple* $\langle o_i, N_i^r, C \rangle$, *where* $N_i^r = \{o_j \mid \forall o_j \in O, o_i \circ o_j \vee n(o_i, o_j) \leq r\}$ *and* C *is the high-level class of the most plausible classification of* l_i, *such that* $\nexists\, d_A \in D : \varphi(N_{i,T}^r, d_A) > \varphi(N_{i,T}^r, d_C)$, *where* $d_C \in D, d_C \neq d_A$, $N_{i,T}^r = \{t^p \mid \forall o_j \in N_i^r, t = T(o_j), p = P(o_j)\}$.

More specifically, N_i^r is the set of observation instances o_j in the neighbourhood of o_i that either overlap with o_i ($o_i \circ o_j$) or are the r-nearest to o_i ($n(o_i, o_j) \leq r$), based on their temporal ordering. Class C denotes the most plausible domain activity classification of l_i, derived by computing the φ similarity between the set with the most specific observation classes $N_{i,T}^r$ and the domain context descriptors $d_k \in D$. $T(o_i)$ denotes the most specific class for instance o_i and we use multisets (duplicates are allowed), since the number of observations with similar class types in the neighbourhood of o_i is important. Moreover, the class types t in $N_{i,T}^r$ are associated with the confidence values p of the corresponding observations.

The φ measure captures the similarity between the multiset N_T^r of a local context against the context descriptor set d_C of a class C. It is defined as

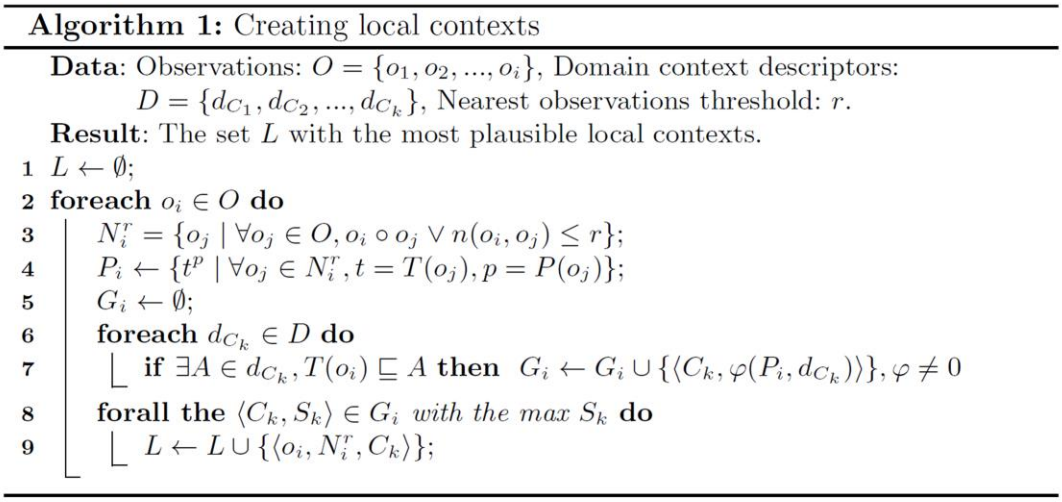
$$\varphi(N_T^r, d_C) = \frac{\sum_{\forall n^p \in N_T^r} \max\limits_{\forall c \in d_C} (\delta(n, c) \cdot p)}{|N_T^r|}$$

where N_T^r is the multiset with neighbour observation class types and d_C is the context descriptor of C. φ is computed as the mean value of the sum of the maximum δ similarities for each concept $n \in N_T^r$, since each n may have more than one relevant concepts in d_C. Intuitively, φ captures the local plausibility of an observation o_i to be part of a complex activity C. If $\varphi = 1$, then all the classes in N_T^r appear in d_C and, therefore, it is very likely that the corresponding local context is part of the complex activity C.

The φ measure uses the δ function that computes the similarity of a neighbour observation class $n \in N_T^r$ against a context descriptor class $c \in d_C$ as

$$\delta(n, c) = \begin{cases} 1, & if\ n \sqsubseteq c \\ \dfrac{|U(n) \cap U(c)|}{|U(n)|}, & if\ c \sqsubseteq n \\ 0, & otherwise \end{cases}$$

where $U(C)$ is the set of the superclasses of C, excluding the $owl{:}Thing$ concept, such that $U(C) = \{A \mid C \sqsubseteq A, A \neq \top\}$. Intuitively, an observation class n in the neighbourhood of o_i exactly matches a class c in the context descriptor set d_C, if it is equivalent to or a subclass of c. In this case, n is subsumed by c and, thus, fully satisfies the contextual dependency imposed by d_C that there should be at least one observation of type c. On the other hand, if c is subsumed by n ($c \sqsubseteq n$), then n is a more general concept than the one required by the context descriptor and the similarity is computed based on the rate of the superclasses of n that are also superclasses of c. For example, if $Spoon$ is a direct subclass of y ($Spoon \sqsubseteq Cutlery$), $n = Spoon$ and $c = Cutlery$, then $\delta(Spoon, Cutlery) = 1$, since $Spoon$ is subsumed by $Cutlery$. If $n = Cutlery$ and $c = Spoon$, then $\delta(Cutlery, Spoon) < 1$, depending on their superclasses.



```
Algorithm 1: Creating local contexts
  Data: Observations: $O = \{o_1, o_2, ..., o_i\}$, Domain context descriptors:
        $D = \{d_{C_1}, d_{C_2}, ..., d_{C_k}\}$, Nearest observations threshold: $r$.
  Result: The set $L$ with the most plausible local contexts.
1 $L \leftarrow \emptyset$;
2 foreach $o_i \in O$ do
3     $N_i^r = \{o_j \mid \forall o_j \in O, o_i \circ o_j \vee n(o_i, o_j) \leq r\}$;
4     $P_i \leftarrow \{t^p \mid \forall o_j \in N_i^r, t = T(o_j), p = P(o_j)\}$;
5     $G_i \leftarrow \emptyset$;
6     foreach $d_{C_k} \in D$ do
7         if $\exists A \in d_{C_k}, T(o_i) \sqsubseteq A$ then $G_i \leftarrow G_i \cup \{\langle C_k, \varphi(P_i, d_{C_k})\rangle\}, \varphi \neq 0$
8     forall the $\langle C_k, S_k\rangle \in G_i$ with the max $S_k$ do
9         $L \leftarrow L \cup \{\langle o_i, N_i^r, C_k\rangle\}$;
```

Algorithm 1 describes the procedure for creating set L with the most plausible local contexts for each $o_i \in O$. The algorithm begins by defining set N_i^r with the neighbour observations of o_i (line 3). Then, the partial context set P_i is created as the multiset of the most specific class types of the observations in N_i^r (line 4), together with the corresponding confidence values p. The algorithm then computes the φ similarity S_k of P_i against each context descriptor d_{C_k}, creating the set G_i with tuples of the form $\langle C_k, S_k\rangle$ (lines 5 to 7). If the class type of o_i does not semantically belong to class descriptor d_{C_k}, then the corresponding similarity tuple is omitted (line 7), ignoring noisy observations. Finally, a tuple $\langle o_i, N_i^r, C_k\rangle$ is created for all $\langle C_k, S_k\rangle$ with the maximum similarity in G_i and inserted into L. Note that G_i may contain more than one $\langle C_k, S_k\rangle$ tuples with the maximum similarity, and, therefore, more than one local contexts can be generated for o_i.

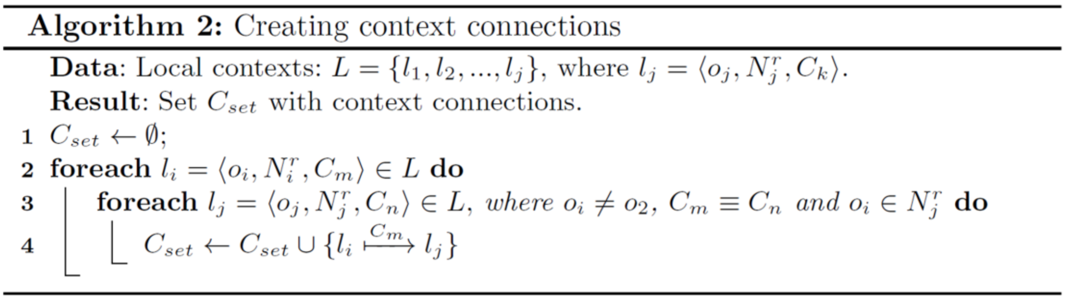
6.2.2 Context Connections

Based on the local contexts obtained in the previous section, the next step is to define *context connections*, that is, links among relevant local contexts that will be used to create the final segments for activity recognition.

Definition 2. *Two local contexts $l_i = \langle o_i, N_i^r, C_m\rangle$ and $l_j = \langle o_j, N_j^r, C_n \rangle$ are linked with a context connection, denoted as $l_i \xrightarrow{C_m} l_j$, if $o_i \in N_j^r$ and $C_m \equiv C_n$.*

Intuitively, a context connection captures the contextual dependency between two neighbour observations o_i and o_j with respect to a common high-level classification activity C_m ($C_m \equiv C_n$). Note that symmetry and transitivity do not hold. For example, the fact that an observation o_i belongs to the neighbours of o_j does not impose that o_j also belongs to the neighbours of o_i.

Algorithm 2 describes the process for creating the set of context connections C_{set}. Two local contexts $l_i = \langle o_i, N_i^r, C_m \rangle$ and $l_j = \langle o_j, N_j^r, C_n \rangle$ are retrieved from L, such that o_i belongs to the neighbourhood of o_j ($o_i \in N_j^r$) and $C_m \equiv C_n$ (lines 2 and 3), and the context connection $l_i \xrightarrow{C_m} l_j$ is added to C_{set} (line 4).



```
Algorithm 2: Creating context connections
Data: Local contexts: $L = \{l_1, l_2, ..., l_j\}$, where $l_j = \langle o_j, N_j^r, C_k \rangle$.
Result: Set $C_{set}$ with context connections.
$C_{set} \leftarrow \emptyset$;
foreach $l_i = \langle o_i, N_i^r, C_m \rangle \in L$ do
    foreach $l_j = \langle o_j, N_j^r, C_n \rangle \in L$, where $o_i \neq o_2$, $C_m \equiv C_n$ and $o_i \in N_j^r$ do
        $C_{set} \leftarrow C_{set} \cup \{l_i \xmapsto{C_m} l_j\}$
```

6.2.3 Activity Situations and Recognition

The last step is to create *activity situations*, i.e. subsets of the initial set of observations O, and to compute the similarity σ to the context descriptor d_C.

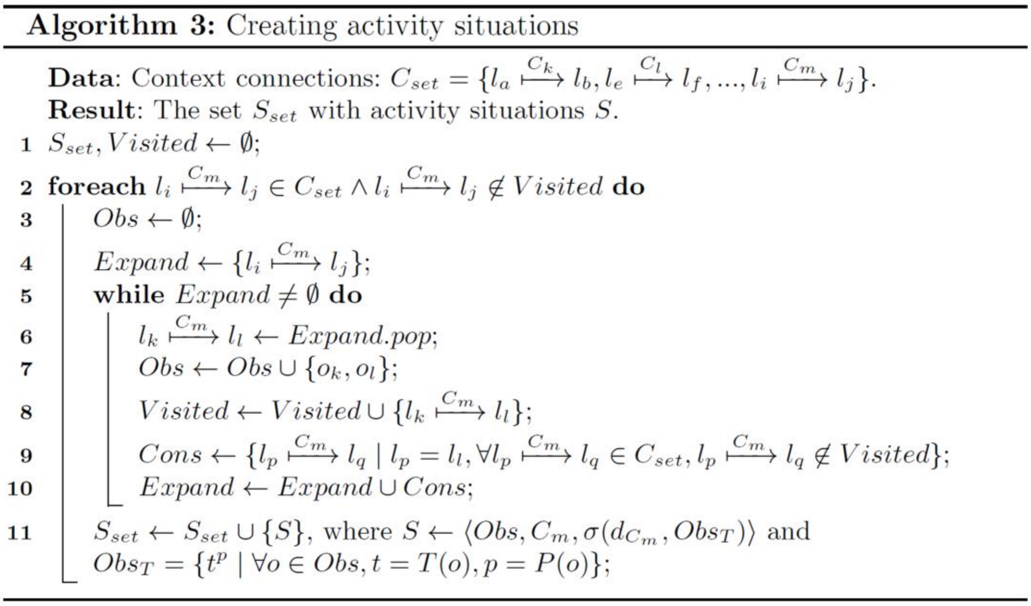
Definition 3. *An activity situation S is defined as the tuple $\langle Obs, C, V \rangle$, where $Obs \subseteq O$ is the set of the observations that belong to the activity situation and V denotes the similarity of S to the context descriptor d_C, such that $V = \sigma(d_C, Obs_T)$, where $d_C \in D$ and $Obs_T = \{t^p \mid \forall o_i \in Obs, t = T(o_i), p = P(o_i)\}$.*

The σ measure captures the similarity between the domain context descriptor of class C, namely d_C, and set Obs_T with the most specific classes of the observations in a situation.

$$\sigma(d_C, Obs_T) = \frac{\sum_{\forall n \in d_C} \max_{\forall c^p \in Obs_T} \delta(c,n) \cdot p}{|d_C|}$$

Similarly to φ, σ denotes the similarity of two sets of concepts. However, φ aims to capture the local (partial) similarity of neighbourhood class types (N_T^r) against the context descriptor d_C. In contrast, σ captures the similarity of the context descriptor d_C against the set of situation observation class types (Obs_T), in order to derive the final plausibility for the corresponding situation. If $\sigma = 1$, then all the classes in d_C appear in Obs_T, meaning that the situation can be considered identical to the context descriptor d_C, and, therefore, to class C.

An activity situation is derived by simply traversing the path defined by context



```
Algorithm 3: Creating activity situations
Data: Context connections: $C_{set} = \{l_a \xmapsto{C_k} l_b, l_e \xmapsto{C_l} l_f, ..., l_i \xmapsto{C_m} l_j\}$.
Result: The set $S_{set}$ with activity situations $S$.
$S_{set}, Visited \leftarrow \emptyset$;
foreach $l_i \xmapsto{C_m} l_j \in C_{set} \wedge l_i \xmapsto{C_m} l_j \notin Visited$ do
    $Obs \leftarrow \emptyset$;
    $Expand \leftarrow \{l_i \xmapsto{C_m} l_j\}$;
    while $Expand \neq \emptyset$ do
        $l_k \xmapsto{C_m} l_l \leftarrow Expand.pop$;
        $Obs \leftarrow Obs \cup \{o_k, o_l\}$;
        $Visited \leftarrow Visited \cup \{l_k \xmapsto{C_m} l_l\}$;
        $Cons \leftarrow \{l_p \xmapsto{C_m} l_q \mid l_p = l_l, \forall l_p \xmapsto{C_m} l_q \in C_{set}, l_p \xmapsto{C_m} l_q \notin Visited\}$;
        $Expand \leftarrow Expand \cup Cons$;
    $S_{set} \leftarrow S_{set} \cup \{S\}$, where $S \leftarrow \langle Obs, C_m, \sigma(d_{C_m}, Obs_T) \rangle$ and
    $Obs_T = \{t^p \mid \forall o \in Obs, t = T(o), p = P(o)\}$;
```

connections $l_a \xrightarrow{C_m} l_b \xrightarrow{C_m} \ldots \xrightarrow{C_m} l_e$, collecting the observations o_i of the local contexts l_i found in the path. The collected observations constitute set Obs of a situation $S = \langle Obs, C_m, V \rangle$. Algorithm 3 describes the procedure. It begins by selecting a context connection $l_i \xrightarrow{C_m} l_j$, which has not been visited yet (line 2), as the root of the current path, adding it to the $Expand$ set (line 4). In each iteration, a context connection $l_k \xrightarrow{C_m} l_l$ is selected from the $Expand$ set and: (a) the observations of the pertinent local contexts are added to Obs (line 7), (b) the current context connection is added to the $Visited$ set (line 8), and (c) the context connections $l_p \xrightarrow{C_m} l_q$ are retrieved from C_{set} and added to the $Expand$ set, such that $l_p = l_l$ (lines 9, 10). An empty $Expand$ set denotes that there are no other context connections in the current path. In this case, the context descriptor of C_m (d_{C_m}) is compared against the set Obs_T with the most specific types of observations in Obs to compute the σ similarity of S (line 11).

6.2.4 Example

In order to better illustrate the basic notions that underpin our approach, we present an example regarding the detection of the $PrepareTea$ and $DrinkTea$ activities. The context descriptors (see Section 6.1) that are used for the two activities are:

$$d_{PrepareTea} = \{PrepareDrink_s, \quad TeaBag_w, \quad Kettle_w, \quad Spoon_w, \quad TeaCup_w, \quad TeaZone_w, \quad PrepareTea_w\}$$

$$d_{DrinkTea} = \{TeaCup_w, Spoon_w, Drink_s, Sitting_s, TableZone_w\}$$

More specifically, the $PrepareTea$ complex activity involves the fusion of:

- $PrepareDrink_s$ activities detected in WP5 from RGB-D streams (static camera - s).
- $TeaBag_w$, $Kettle_w$, $Spoon_w$ and $TeaCup_w$ objects detected in WP4 based on object recognition from wearable camera (w).
- $TeaZone_w$ observations derived in WP4 regarding the location of the person based on place recognition from wearable camera.
- $PrepareTea_w$ activities detected in WP5 by fusing objects and locations (Section 5).

Similarly, the $DrinkTea$ activity involves the fusion of:

- $TeaCup_w$, $Spoon_w$ objects detected in WP4 based on object recognition from wearable camera.
- $TableZone_w$ observations derived in WP4 regarding the location of the person based on place recognition from wearable camera.
- $Sitting_s$ observations derived in WP5 from RGB-D streams.
- $Drink_s$ activities detected in WP4 based on activity recognition from static camera.

The aim of the context-based fusion is to effectively integrate the complementary information detected by the various components of the system, in order to identify contexts that signify the presence of complex situations and to derive the most plausible classification to the domain activity classes. In our example, for instance, the goal is to combine $PrepareTea_w$ activities that are derived by fusing objects and locations from wearable camera, with $PrepareDrink_s$ activities that are detected from RGB-D video streams, as well as, with contextual information regarding objects and locations in order to take a final decision about the occurrence of the $PrepareTea$ activity.

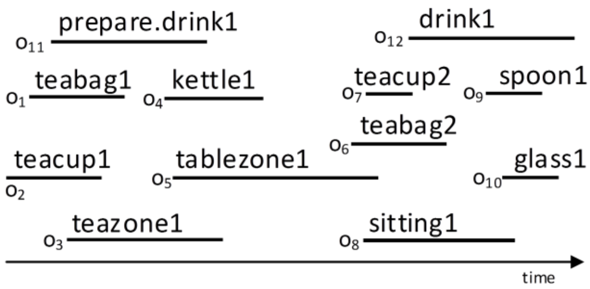


Figure 6-2. Example observations for preparing and drinking tea

Following the above description, the context descriptors of the $PrepareTea$ and $DrinkTea$ activities are defined in our example as: $d_{PrepareTea} = \{\, PrepareDrink_s,\ TeaBag_w,\ Kettle_w,\ Spoon_w,\ TeaCup_w,\ TeaZone_w,\ PrepareTea_w \,\}$ and $d_{DrinkTea} = \{TeaCup_w,\ Spoon_w,\ TableZone_w,\ Sitting_s,\ Drink_s\}$. In the following, we use the observations depicted in Figure 6-2 relevant to making and drinking tea and we describe the procedure for identifying and classifying the corresponding complex activities. For simplicity, we assume in this example that the confidence values of all the observations is 1.

Local Contexts (Algorithm 1). We describe the definition of the local context for the $teacup2_w$ observation (o_7) in Figure 6-2, using $r = 1$. Observations o_5, o_6, o_8 and o_{12} overlap with o_7, whereas o_9 and o_4 are the 1-nearest to o_7. Thus, $N_7^1 = \{o_7, o_5, o_6, o_8, o_9, o_4, o_{12}\}$ (line 3) and $P_7 = \{TeaCup, TableZone, TeaBag, Sitting, Spoon,\ Kettle, Drink\}$ (line 4). Our example involves two context descriptors and therefore $D = \{d_{PrepareTea}, d_{DrinkTea}\}$. The class type for o_7 is $TeaCup$ that exists in both context descriptors, therefore φ will be computed for both of them. We have that $\varphi\left(P_7, d_{PrepareTea}\right) = \frac{1+0+1+0+1+1+0}{7} =\ 0.57$ and $\varphi(P_7, d_{DrinkTea}) = \frac{1+1+0+1+1+0+1}{7} = 0.71$. Thus, we have a single local context for o_7 with maximum plausibility, denoted as $l_7 = \langle o_7, N_7^1, DrinkTea\rangle$. Similarly, we have the following local contexts for the other observations:

$l_1 = \langle o_1, N_1^1, PrepareTea\rangle^{1.0}$

$l_2 = \langle o_2, N_2^1, PrepareTea\rangle^{1.0}$

$l_3 = \langle o_3, N_3^1, PrepareTea\rangle^{0.85}$

$l_4 = \langle o_4, N_4^1, PrepareTea\rangle^{0.8}$

$l_5 = \langle o_5, N_5^1, DrinkTea\rangle^{0.42}$

$l_6 = \langle o_6, N_6^1, PrepareTea\rangle^{0.57}$

$l_8 = \langle o_8, N_8^1, DrinkTea\rangle^{0.625}$

$l_9 = \langle o_9, N_9^1, DrinkTea\rangle^{0.6}$

$l_{10} = -$

$l_{11} = \langle o_{11}, N_{11}^1, PrepareTea\rangle^{0.83}$

$l_{12} = \langle o_{12}, N_{12}^1, DrinkTea\rangle^{0.71}$

Context Connections (Algorithm 2). 36 context connections are created among the local contexts of our example. For instance, o_7 belongs to the neighbourhood of the local contexts l_5, l_6, l_8, and l_{12}. As described, the classification class of l_5, l_8 and l_{12} is Tea (DT), whereas the classification class of l_6 is $PrepareTea$ (PT). Therefore, l_7 will form context connections only with l_5, l_8 and l_{12}, i.e. $l_7 \xrightarrow{DT} l_5, l_7 \xrightarrow{DT} l_8, l_7 \xrightarrow{DT} l_{12}$. The other context connections that are generated are:

$l_{12} \xrightarrow{DT} l_7$	$l_4 \xrightarrow{PT} l_2$	$l_7 \xrightarrow{DT} l_{12}$	$l_{11} \xrightarrow{PT} l_1$	$l_4 \xrightarrow{PT} l_3$	$l_8 \xrightarrow{DT} l_5$	$l_{11} \xrightarrow{PT} l_4$	$l_4 \xrightarrow{PT} l_{11}$
$l_3 \xrightarrow{PT} l_2$	$l_3 \xrightarrow{PT} l_{11}$	$l_8 \xrightarrow{DT} l_{12}$	$l_8 \xrightarrow{DT} l_7$	$l_2 \xrightarrow{PT} l_{11}$	$l_8 \xrightarrow{DT} l_9$	$l_5 \xrightarrow{DT} l_7$	$l_2 \xrightarrow{PT} l_3$
$l_3 \xrightarrow{PT} l_4$	$l_4 \xrightarrow{PT} l_6$	$l_{11} \xrightarrow{PT} l_2$	$l_2 \xrightarrow{PT} l_1$	$l_5 \xrightarrow{DT} l_{12}$	$l_7 \xrightarrow{DT} l_8$	$l_1 \xrightarrow{PT} l_{11}$	$l_3 \xrightarrow{PT} l_1$
$l_9 \xrightarrow{DT} l_7$	$l_{12} \xrightarrow{DT} l_9$	$l_5 \xrightarrow{DT} l_8$	$l_9 \xrightarrow{DT} l_{12}$	$l_9 \xrightarrow{DT} l_8$	$l_{12} \xrightarrow{DT} l_8$	$l_1 \xrightarrow{PT} l_3$	$l_1 \xrightarrow{PT} l_2$
$l_4 \xrightarrow{PT} l_1$	$l_7 \xrightarrow{DT} l_5$	$l_1 \xrightarrow{PT} l_4$	$l_{11} \xrightarrow{PT} l_3$				

Activity Situations (Algorithm 3). By applying Algorithm 3 over the 36 context connections, two activity situations are generated: $S_1 = \langle Obs_1, PrepareTea, 0.714\rangle$ and $S_2 = \langle Obs_2,\ DrinkTea, 1.0\rangle$, where $Obs_1 = \{o_1, o_2, o_3, o_4, o_6, o_{11}\}$ and $Obs_2 = \{o_5, o_7, o_8, o_9, o_{12}\}$ (Figure 6-3 (a)). Despite the overlapping and noisy nature of the observations in the example (e.g. the location-related observations o_3 and o_5 overlap), the algorithm is able to discriminate the two situations of preparing and drinking a tea by connecting also the relevant observations. The observation $glass1$ is considered as noise and it is ignored. It is worth noting that the $PrepareTea$ activity is recognised (with lower plausibility) even if the situation S_1 partially matches the corresponding context descriptor. For example, there are no observations of type $PrepareTea_w$ and $Spoon_w$ in Obs_1.

The nearest observations threshold r in the running example was set to 1, meaning that, apart from overlapping observations, the 1-nearest observations were also taken into account to define neighbourhood relations. If we instead use $r = 0$, then we get the result of Figure 6-3 (b). In this case, o_6 ($teabag2$) is not connected with observations relevant to the $PrepareTea$ activity, and is considered as noise, breaking also the connection of o_7 and o_9 with the $PrepareTea$ activity that are classified instead in the $DrinkTea$ activity. Intuitively, r allows control of the amount of contextual information to be taken into account during the definition of the neighbourhood sets and local contexts of observations. Currently, r is defined manually based on domain knowledge regarding the quality and temporal characteristics of the data used and, in principle, datasets with highly overlapping and incoherent observations need small r values.

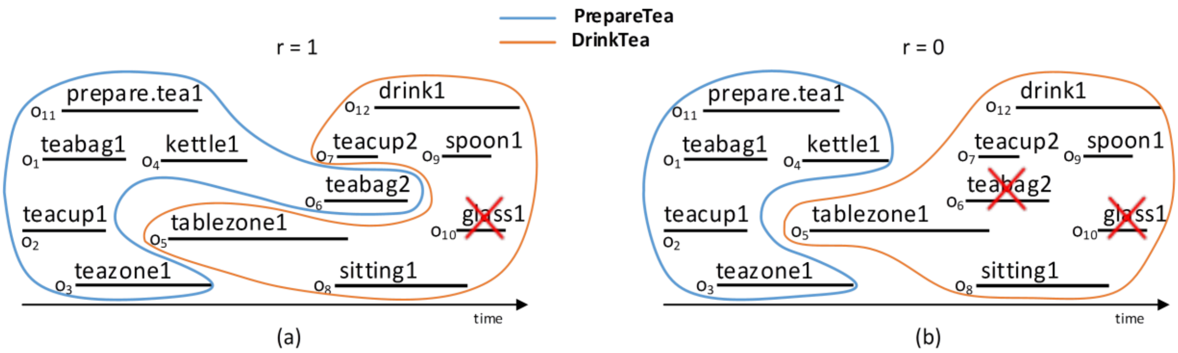


Figure 6-3. Fusion results according to the r value

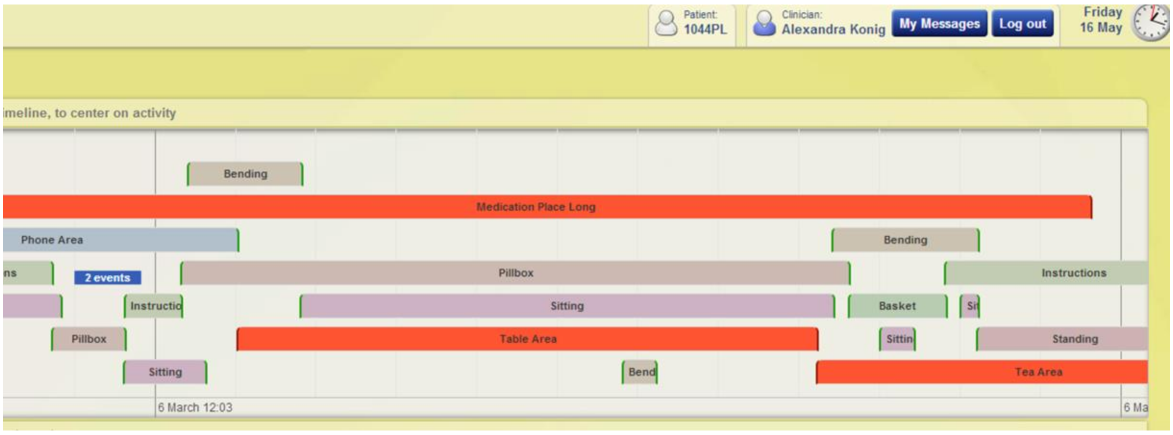


Figure 6-5. Example visualisation of objects, locations and activities detected at Lab

6.2.5 Evaluation

Given the limited availability of datasets with observations from the various Dem@Care components, we have evaluated the context-based fusion approach on synthetic data. Figure 6-4 depicts the visualisation by the Clinician Interface of results obtained by Dem@Care components (v1) during a Lab session. As illustrated, the generated observations are highly overlapped, noisy and incomplete, with inaccurate temporal correlations. For example, the patient is detected at the medication and phone areas at the same time.

In order to evaluate our approach with synthetic data that capture the aforementioned intrinsic properties and characteristics of Dem@Care data, we developed a synthetic data generator tool that can be used to generate observations for IADLs. Each IADL is described in terms of its relevant primitive observations regarding objects, locations, postures and actions. For each observation, the probability of occurrence in each relevant IADL is also specified, in order to simulate missing information and the fact that an activity can be performed in a variety of ways. Moreover, there is a fixed probability of each observation to appear in non-relevant IADLs, generating in that way noise. Figure 6-5 presents observations that have been generated for the activity $PrepareDrugBox$.

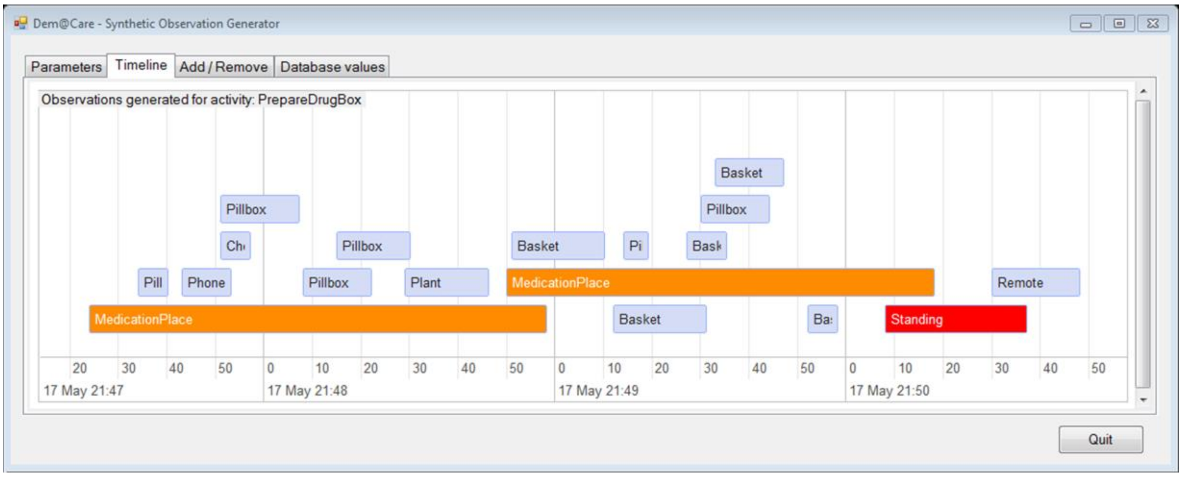


Figure 6-4. Sample synthetic data generated by the Dem@Care observation generator

Table 6-1. Contexts descriptors of high-level activities

IADL	Context Descriptors
PrepareDrugBox	Pillbox, Basket, MedicationPlace
PrepareHotTea	Kettle, TeaArea, TeaBag, Cup, Sugar, TeaBox
MakePhoneCall	Phone, PhoneZone, PickUpPhone, Talk
WatchTV	Remote, TV, TVZone, Sitting
WaterPlant	WateringCan, PlantZone, Bending, Plant
WriteCheck	Sitting, Pen, Check, TableZone, Table
WashHands	Soap, SinkZone, Tap, Hands
ReadBook	Sitting, ChairZone, Book

We generated 50 datasets using the context descriptors of the 8 activities depicted in Table 6-1. Each dataset contains all nine activities and their order, while the number and type of the respective observations have been randomly defined. Moreover, the observations of an activity have been defined to start at most 5 seconds before or after the last observation of the previous activity. In that way, we incorporate temporal inaccuracies among high-level activities allowing observations from different activities to overlap.

Table 6-2 summarises the performance on the synthetic datasets, where True Positives (TP) is the number of IADLs correctly recognised, False Positives (FP) is the number of IADLs incorrectly recognised as performed and False Negatives (FN) is the number of IADLs that have not been recognised. The recall (Rec) and precision (Pre) are defined in Section 4.3.1.

We used $r = 0$, since the synthetic datasets contain highly overlapping observations, and we set a minimum threshold on σ ($\sigma \geq 0.65$), so as to ignore activities with low plausibility. As described, the optimal r value depends on the data quality and temporal characteristics and, in principle, datasets with highly overlapping and incoherent observations need small r values. Our approach achieves the best accuracy for activities “Prepare hot tea”, “Make a phone call”, “Watch TV”, “Water the plant”, “WashHands” and “WriteCheck”, whose context descriptors encapsulate richer domain contextual information, compared to “Prepare drug box” and “ReadBook”. On the other hand, the recall of these activities is relatively low, since they are

Table 6-2. Recall and precision results of context-aware fusion

	TP	FP	FN	Rec	Pre
PrepareDrugBox	45	10	5	90.00	81.82
PrepareHotTea	38	3	12	76.00	92.68
MakePhoneCall	36	4	14	72.00	90.00
WatchTV	41	3	9	82.00	93.18
WaterPlant	41	3	9	82.00	93.18
WriteCheck	40	4	10	80.00	90.91
WashHands	42	4	8	84.00	91.30
ReadBook	45	8	5	90.00	84.91

more susceptible to false negatives, requiring richer contextual dependencies to be present.

Our framework achieves an average precision close to 90%, demonstrating the feasibility of our approach. However, there are still certain limitations, which we consider as very important research directions for future work. First, our approach cannot handle interleaved activities, nor can it resolve conflicts after the recognition process. We are investigating the use of defeasible reasoning on top of the framework for further enhancing the activity recognition capabilities. Second, our next step is to provide context-aware real-time assistance to Alzheimer's patients. To this end, we are currently investigating adaptations of our algorithms to allow the dynamic and incremental generation of local contexts and context connections for real-time fusion, using a CEP engine (see Section 7.2).

6.3 Summary

We presented a knowledge-driven framework towards activity recognition, coupling ontology models of abstract domain activity dependencies with a context-aware approach for fusing observations coming from multiple sources. We formalise activity dependencies capitalising upon the $Situation$ conceptualisation of the DnS ontology pattern in DUL, whereas fusion is reduced in linking and classifying meaningful contextual segments. We elaborated on the obtained results from the evaluation of our approach in a synthetic dataset. The use of generic context descriptors in representing activity models achieves very promising results, leading to handling the intrinsically noisy and imperfect information in multi-sensory environments, beyond strict activity patterns and background knowledge.

The key directions that underpin our ongoing research involve: (a) introducing an additional layer for detecting interleaved activities and resolving conflicts, (b) adapting our algorithms for supporting real-time context-aware monitoring, and, (c) patient profiling through the extraction and learning of behavioural patterns from the detected activity situations. In addition, we are investigating extensions to the $Situation$ model for capturing richer contextual dependencies, such as compositions of context descriptors.

7 Functional Extensions

This section presents service-related contributions that extend the first version of the multi-parametric interpretation framework presented in D5.2.

7.1 Support of Questionnaires

Questionnaires are an important tool for obtaining user-reported data about problems in the daily life, for example, mood and sleeping problems. However, the information provided by the patients is subjective and in many cases incomplete, without reflecting their actual state, progress and functioning in daily life. The standard doctor’s questionnaire-based assessment of the person with dementia in Dem@Care can be greatly reinforced by the multi-sensor processing results, behavioural profiling and interpretation, so as to deliver through WP6 the appropriate clinical feedback and at-home treatment recommendations (WP2), closing the clinician’s loop.

The second version of the multi-parametric behaviour interpretation framework provides the necessary knowledge structures and analysis procedures for storing and calculating the scores of the questionnaires that are used in Dem@Care, as these have been reported in the respective deliverables (Functional Requirements & Scenarios v1 [33] and v2 [41]). In the following subsections, we briefly describe: (a) the updates performed on the representation layer of SI to adequately capture questionnaire-related data, and (b) the SPARQL-based implementation of the scoring algorithms.

7.1.1 Questionnaire Ontology

The basic class relationships of the questionnaire-related ontology are depicted in Figure 7-1. Seven subclasses of the *Questionnaire* class have been defined for the representation of the seven questionnaire types that are currently supported by the ontology. Each Questionnaire can be associated with one or more *Questions* through the *hasQuestion* property; each *Question* has an *id* and it can be associated with an *Answer* through the *hasAnswer*

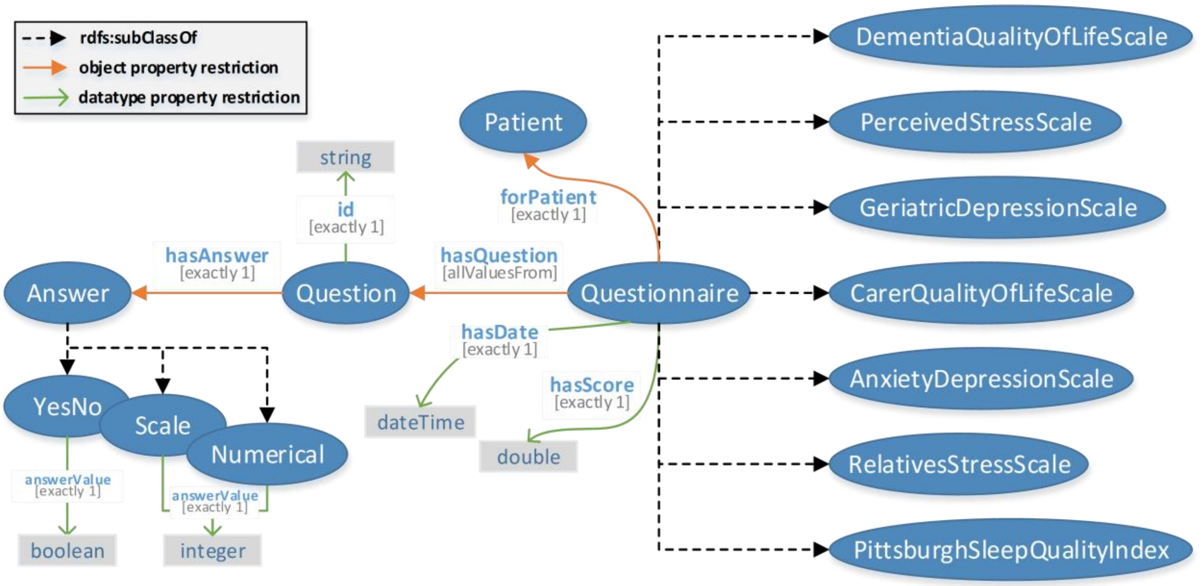


Figure 7-1. Questionnaire ontology

property. Moreover, each $Questionnaire$ is linked to a $Patient$ ($forPatient$ property), it has a date ($hasDate$ property) and it has a score that is computed by the SI module, as we describe in the following section. Towards standardisation, the questionnaire ontology has been aligned with the conceptual schema of the DOLCE+DnS Ultralite (DUL) ontology [28].

7.1.2 Calculating Scores

In order to compute the scores of the questionnaires, we have used SPARQL rules in terms of SPARQL `CONSTRUCT` query graph patterns to implement the interpretation logic of each questionnaire type, reusing the SPARQL rule execution framework of SI for activity interpretation (see D5.2). For example, the steps needed to calculate the sleep latency score (**PSQILATEN** value) based on the answers of the second (**Question2**) and fifth (**Question5a**) queries of the Pittsburgh Sleep Quality Index (PSQI) questionnaire are the following (for the complete list of steps refer to deliverable D8.2 Evaluation Protocols [36]):

Question 2: *During the past month, how long (in minutes) has it usually taken you to fall asleep each night?*

Step1:

IF **Question2** ≥ 0 and ≤ 15, THEN set value of **Question2_new** to 0

IF **Question2** > 15 and ≤ 30, THEN set value of **Question2_new** to 1

IF **Question2** > 30 and ≤ 60, THEN set value of Question2_new to 2

IF **Question2** > 60, THEN set value of **Question2_new** to 3

Question 5: *During the past month, how often have you had trouble sleeping because you …*

a) Cannot get to sleep within 30 minutes (select one of the following)

Not during the past month__, Less than once a week__, Once or twice a week__, Three or more times a week__

Step2

IF **Question5a** + **Question2_new** = 0, THEN set **PSQILATEN** to 0

IF **Question5a** + **Question2_new** ≥ 1 and ≤ 2, THEN set **PSQILATEN** to 1

IF **Question5a** + **Question2_new** ≥ 3 and ≤ 4, THEN set **PSQILATEN** to 2

IF **Question5a** + **Question2_new** ≥ 5 and ≤ 6, THEN set **PSQILATEN** to 3

The following SPARQL query implements the first IF statement of Step1, setting the `Question2_new` property equal to 0, when the answer is between 0 and 15 minutes.

```
CONSTRUCT {
    ?qs :Question2_new 0;
}
WHERE {
    ?qs a :Questionnaire;
        :hasQuestion [:id “2”; :hasAnswer [:answerValue ?Question2]].
    FILTER (?Question2 >= 0 && && ?Question2 <= 15) .
}
```

Similar SPARQL rules have been defined for the rest of the IF statements of Step1. Regarding Step2, the following SPARQL rule implements the first IF statement, adding the `Question2_new` property value to the answer value (0, 1, 2 or 3) of the fifth question.

```
CONSTRUCT {
    ?qs :PSQILATEN ?sum;
}
WHERE {
    ?qs :Question2_new ?v;
        :hasQuestion [:id “5a”; :hasAnswer [:answerValue ?Question5a]].
    BIND(?v + ?Question5a as ?sum) .
    FILTER(?sum = 0) .
}
```

Similar SPARQL rules have been defined for the rest of the IF statements of Step2. The final score of the PSQI questionnaire is derived by the aggregation of the scoring values for other sleep-related attributes, such as subjective sleep quality, sleep duration, sleep efficiency, sleep disturbances, daytime dysfunctions and use of sleep medications. The scores are stored in the Knowledge Base (KB) and can be retrieved by WP6 in order to deliver the appropriate feedback and at-home treatment recommendations, taking also into account other multi-sensor processing results, for example, additional sleep-related problems that might have been detected by WP5 (e.g. nocturia incidents).

7.2 Complex Event Processing

In the second version of the multi-parametric behaviour interpretation framework, WP5 has been enhanced with Complex Event Processing (CEP) capabilities. The objective is to take advantage of the native temporal reasoning capabilities of CEP to provide contextualised real-time support of patients, caregivers and clinicians via coupling profile and clinical knowledge with results made available by the other components of the system.

In the current version of the Dem@Care system, however, only CAR is able to deliver real-time analysis results, delineating the possible realm of WP5 real-time interpretation. For example, the location, object and activity recognition modules in WP4 (ORWC, RRWC, WCPU, HAR) analyse video data in an offline mode. Similarly, the processing of the DTI-2 data for extracting moving intensity measurements, as well as, the processing of audio data (OSA module) for extracting speech-related measurements, such as verbal reaction time, verbal participation, voice rating during conversation, are offline procedures. As a result, there is limited availability of online data in WP5 for real-time fusion.

Following the above discussion, we present in this section preliminary investigations on providing basic real-time services within WP5 regarding the detection of potentially critical situations of patients, coupling activities detected by CAR with patient profile knowledge. More elaborate real-time interpretation tasks will be tackled as the Dem@Care system will be gradually enriched with real-time sensors and analysis components.

7.2.1 Modelling Profile Knowledge

Profile knowledge is represented using the ontology patterns described in D5.3 for modelling basic behaviour aspects of patients, such as typical night sleep duration, the amount of times the patient visits the bathroom during the night, the daily frequency of medicine intake, the frequency of meals per day, etc. Figure 7-2 presents instantiations of the duration and frequency patterns to model patient profile knowledge regarding the duration of night sleep and the number of bathroom visits during the night.

The instantiated patterns are stored in the KB of WP5, allowing SPARQL queries to be executed for retrieving profile information about the patient. For example, the following SPARQL query retrieves from the KB the typical frequency of a patient’s night bathroom visits.

```
SELECT ?frequency
WHERE {
    ?p a :ActivityFrequency;
       :hasDescription
           [:definesActivityType
              [:classifiesActivity :NightBathroomVisit;
                 :period "daily";
                 :value ?frequency]
           ].
}
```

It should be noted that the profile information is currently defined manually as part of the initialisation of the system with patient background knowledge. Next steps include the development of algorithms for their automated population and enrichment, i.e. the learning and evolution of pattern descriptions. First results will be reported in the upcoming deliverable D5.5 Contextualised Knowledge Enrichment.

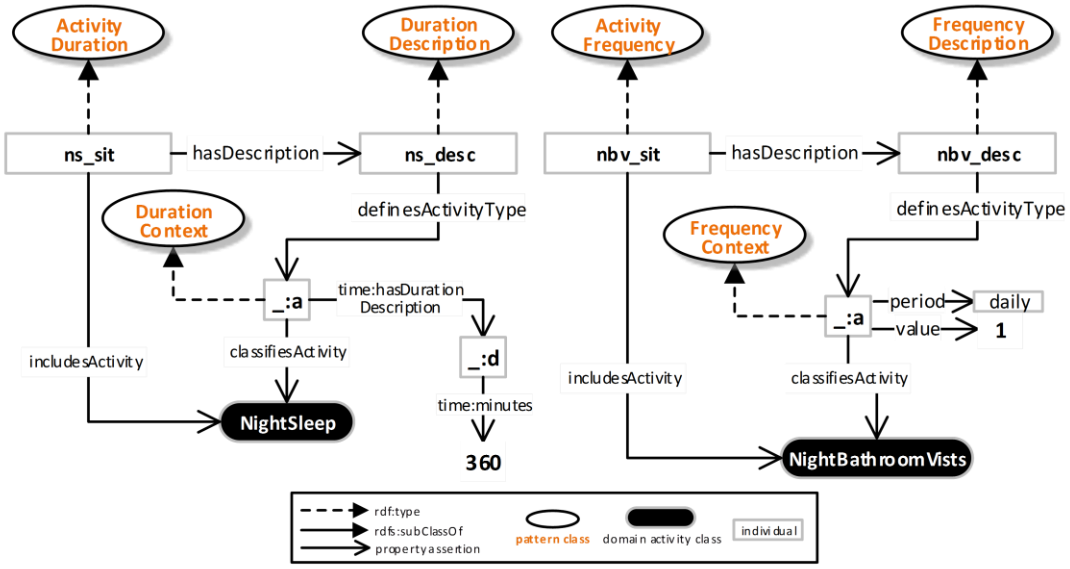


Figure 7-2. Example pattern instantiations for representing profile knowledge.

7.2.2 CEP Patterns

The purpose of the CEP patterns is to detect (near) real-time deviations of patients from normal behaviours that indicate problems or possibly problematic situations, triggering the respective feedback/alert services in WP6. To this end, the CEP engine of Drools (Drools

Fusion[1]) is used within WP5 to correlate activities detected by CAR with patient-pertinent profile knowledge by means of ontology patterns.

Figure 7-3 presents the interactions of the WP5 components in v2. The CAR component sends the detected activities directly to the CEP module for real-time fusion with profile knowledge. The CEP engine queries the KB to retrieve behaviour patterns and the events that are detected are sent to the alert and feedback services of WP6. Note that the activities sent by CAR are also stored in the KB for further offline processing and fusion with other observations by SI.

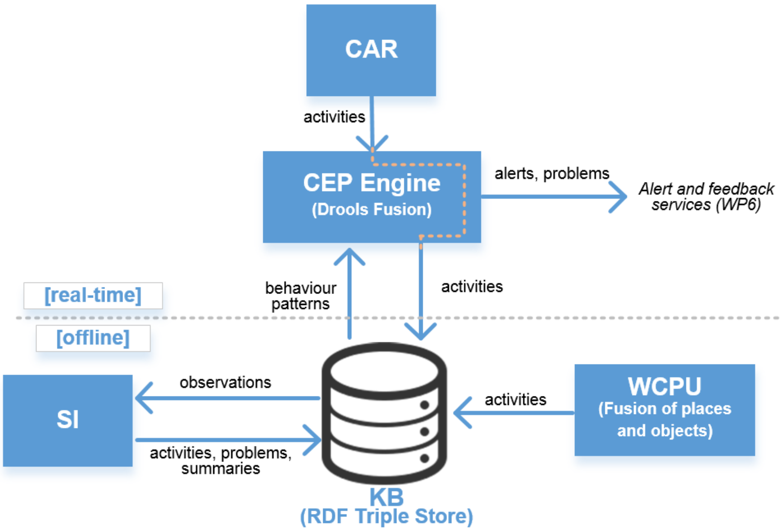


Figure 7-3. Logical component architecture within WP5

Drools provides a declarative, rule-based language for the definition of CEP patterns. Figure 7-4 presents the CEP rule that is used to detect situations where patients go to the bathroom at night more often than usual. This rule actually fuses the two profile patterns in Figure 7-2. More specifically, the rule queries the KB to retrieve profile information regarding the frequency of the night bathroom visits as well as the typical duration of night sleep. The duration is used to define the window size that is used to count the number of the already detected night bathroom visits. If this number is greater than the typical frequency, then an alert is generated that is sent to WP6 for further processing.

```
rule: "NightBathroomVisitAlert"
when
    NightBathroomVisit()
    Frequency($f : Value) from KB.GetNightBathroomVisitFrequency()
    Duration($d : Value) from KB.GetNightSleepDuration()
    Number(intValue > $f) from accumulate (
        $nb: NightBathroomVisit() over window:time($d),
        count( $nb ))
then
    //send alert to WP6 (feedback & alert manager)
```

Figure 7-4. Example CEP rule

[1] http://drools.jboss.org/drools-fusion.html

8 Conclusions

This deliverable presented the second version of the multi-parametric behaviour interpretation framework, focusing on the extensions that have been implemented for supporting reasoning under uncertainty and handling incomplete and noisy input.

We have presented a framework for modelling uncertainty of low-level events (elementary scenarios). The results show that the framework has increased the recall index of the detection of elementary scenarios whose constraints are based on low-level data. Further development is being performed to improve the approach precision and extending it to all hierarchy levels of event modelling. The uncertainty framework intends to extend the deterministic constraint-based framework used in CAR.

We have also studied the prediction of calibrated probability scores for activity recognition using the wearable camera. The fusion of active object recognition and location features is used as input to classify the activities. It was demonstrated that, given the complexity of the task and the available training data, models with fewer parameters such as the Platt model should be used to avoid over-fitting. The reliability of these probabilities is indeed important for ensuring the quality of the interpretation when fusing multi-parametric data.

In order to handle the intrinsic characteristics in pervasive environments, such as imperfect information, noise, conflicts or inaccurate temporal correlations, we defined a practical ontology-based framework for fusing observations from multiple heterogeneous sources. The framework detects complex activities based on loosely coupled domain activity dependencies rather than on strict contextual constraints in the form of rules.

In addition, further functional extensions to v1 have been described regarding the support of questionnaires and the incorporation of a CEP engine for providing (initially basic) real-time services within WP5 regarding the detection of potentially critical situations of patients, coupling activities detected by CAR with profile knowledge.

Future work will focus on the learning of semantic spatial zones on an unsupervised manner and the use of activity probabilities with temporal and auxiliary information to detect salient temporal events characterizing the behaviour of the person. We are also investigating methods for extracting behaviour patterns and detecting behaviour changes from activity situations that are generated based on the abstract context descriptors presented in this deliverable towards patient profiling. Regarding the representation of these patterns, our objective is to take full advantage of the DnS pattern, associating each situation to one or more behavioural description instantiations pertinent to patients' idiosyncratic and habitual information. First results will be reported in the upcoming deliverable D5.5 Contextualised Knowledge Enrichment.

9 References

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

[2] T. Banerjee, M. Rantz, M. Popescu, E. Stone, M. Li and M. Skubic. Monitoring Hospital Rooms for Safety Using Depth Images. AI for Gerontechnology, Arlington, Virginia, US, November 2012.

[3] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool. Speeded-Up Robust Features (SURF), Comput. Vis. Image Underst., vol. 110, pp. 346-359, June 2008

[4] A. Bikakis and G. Antoniou. Contextual argumentation in ambient intelligence. In Proceedings of the 10th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR ’09), pages 30–43, Potsdam, Germany, 2009.

[5] F. Bobillo and U. Straccia. Fuzzy ontology representation using OWL 2. Int. J. Approx. Reasoning, 52(7):1073–1094, 2011

[6] H. Boujut, J. Benois-Pineau, R. Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion, ECCV 2012 - Workshops, 2012

[7] W. Brendel, A. Fern, and S. Todorovic. Probabilistic Event Logic for Interval-Based Event Recognition. In CVPR, p. 3329 - 3336, 2011.

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

[9] R.N. Carvalho, K. B. Laskey, and P. C. G. da Costa. PR-OWL 2.0 - Bridging the gap to OWL semantics. In URSW, pages 73–84, 2010

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

[11] C.L. Chiang, C. C. Lien, and C. H. Lee. Scene-based event detection for baseball videos. In Journal of Visual Communication and Image Representation, pp. 1-14, February, 2007.

[12] A. Ciaramella, M. G. C. A. Cimino, F. Marcelloni, and U. Straccia. Combining fuzzy logic and semantic web to enable situation-awareness in service recommendation. In Proceedings of the 21st international conference on Database and expert systems applications: Part I, DEXA’10, pages 31-45, Berlin, Heidelberg, 2010. Springer-Verlag

[13] C. Cortes, V. Vapnik. Support-vector networks, Machine Learning, vol. 20, pp. 273-297, 1995

[14] P.C. Costa, K. B. Laskey, and K. J. Laskey. PR-OWL: A bayesian ontology language for the semantic web. In P. C. Costa, C. D’Amato, N. Fanizzi, K. B. Laskey, K. J. Laskey, T. Lukasiewicz, M. Nickles, and M. Pool, editors, Uncertainty Reasoning for the Semantic Web I, pages 88–107. SpringerVerlag, Berlin, Heidelberg, 2008

[15] C. Crispim-Junior, S. Cosar, F. Bremond, G. Meditskos, S. Dasiopoulou, C. Doulaverakis, A. Bugeau, V. Buso, J. Benois-Pineau, D5.3 Behavioural Profile Learning, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199

[16] C. Crispim-Junior, B. Fosty, R. Romdhane, F. Bremond and M. Thonnat. Combining Multiple Sensors for Event Recognition of Older People. In the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare, MIIRH 2013, Copyright 2013 ACM 978-1-4503-2398-7/13/10, October 22, 2013.

[17] C. Crispim-Junior, V. Bathrinarayanan, B. Fosty, A. Konig, R. Romdhane, M. Thonnat and F. Bremond. Evaluation of a Monitoring System for Event Recognition of Older People. In the 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance 2013, AVSS 2013, Krakow, Poland on August 27-30, 2013.

[18] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray. Visual categorization with bags of keypoints, In Workshop on Statistical Learning in Computer Vision, 2004

[19] S. Dasiopoulou, V. Efstathiou, G. Meditskos, C. Crispim-Junior, A.T. Nghiem, V. Buso, “D5.2 Multi-parametric Behaviour Interpretation v1”, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199

[20] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. *In* Proc. 12th European Conference on Computer Vision (ECCV), 2012.

[21] Z. Ding, Y. Peng, R. Pan, Z. Ding, Y. Peng, and R. Pan. BayesOWL: Uncertainty modeling in semantic web ontologies. Soft Computing in Ontologies and Semantic Web, pages 3–29, 2006

[22] R. Dividino, S. Sizov, S. Staab, and B. Schueler. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. Web Semant., 7:204–219, September 2009

[23] V. Dovgalecs, R. Megret, Y. Berthoumieu. Multiple Feature Fusion Based on Co-Training Approach and Time Regularization for Place Classification in Wearable Video, Advances in Multimedia, 2013

[24] J. Du, G. Qi, Y.-D. Shen, and J. Z. Pan. Towards practical ABox abduction in large OWL DL ontologies. In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, pages 1160–1165, San Francisco, California, USA, 2011. AAAI Press

[25] T. Duong, H. Bui, D. Phung, and S. Venkatesh. Activity y recognition and abnormality detection with the switching hidden semi-markov model. In CVPR, 2005.

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

[27] D.F. Fouhey, V. Delaitre, A. Gupta, A. Efros, I. Laptev, and J. Sivic. People Watching: Human Actions as a Cue for Single-View Geometry. *In* Proc. 12th European Conference on Computer Vision (ECCV), pp. 732-745, 2012.

[28] A. Gangemi. DOLCE+DnS Ultralite (DUL) ontology. http://www.loa.istc.cnr.it/ontologies/DUL.owl, July 2012.

[29] A. Gangemi, P. Mika. Understanding the semantic web through descriptions and situations. In: Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics. pp. 689-706 (2003)

[30] M. Gebel. Multivariate Calibration of Classifier Scores into the Probability Space. PhD thesis, Dortmund, Techn. Univ., Diss., 2009.

[31] J. Gómez-Romero, M. A. Patricio, J. García, and J. M. Molina. Context-based reasoning using ontologies to adapt visual tracking in surveillance. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS ’09, pages 226–231, Washington, DC, USA, 2009. IEEE Computer Society.

[32] J. Gómez-Romero, M. A. Patricio, J. García, and J. M. Molina. Ontology-based context representation and reasoning for object tracking and scene interpretation in video. Expert Syst. Appl., 38(6):7494–7510, 2011

[33] K. Irving, D. O’Mahoney, V. Joumier, S. Sävenstedt, D2.2 Functional Requirements and Clinical Scenarios v1, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199.

[34] V. Joumier, R. Romdhane, F. Bremond, M. Thonnat, E. Mulin, P.H. Robert, A. Derreumeaux, J. Piano and L. Lee. Video Activity Recognition Framework for assessing motor behavioural disorders in Alzheimer Disease Patients. In the International Workshop on Behaviour Analysis, Behave 2011, Sophia Antipolis, France on the 23rd of September 2011.

[35] S. Klarman, U. Endriss, and S. Schlobach. ABox Abduction in the Description Logic ALC. J. Autom. Reasoning, 46(1):43–80, 2011

[36] A. König, C. Fernando Crispim-Junior, J. McHugh, S. Sävenstedt, D8.2 Evaluation Protocols, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199

[37] P. Kumar, S. Ranganath, H. Weimin, and K. Sengupta. Framework for real-time behavior interpretation from traffic video. *In* IEEE Trans. on Intelligent Transportation Systems, 6, p. 43-53, 2005.

[38] S. Kwak, B. Han, and J. H. Han. Scenario-Based Video Event Recognition by Constraint Flow. In CVPR, p. 3345–3352, 2011.

[39] F. Lv, X. Song, V. Wu, B. Kumar, and R. Nevatia. Left luggage detection using Bayesian inference. *In* Proceedings of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance, pages 83–90, 2006.

[40] J. McHugh, E. Murphy, K. Irving, L. Hopper, A. König, S. Sävenstedt, D2.6 Functional Requirements and Scenarios v2, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199.

[41] J. McHugh, E. Murphy, K. Irving, L. Hopper, A. König, S. Sävenstedt, D2.2 Functional Requirements and Clinical Scenarios v2, Dementia Ambient Care: Multi-

Sensing Monitoring for Intelligent Remote Management and Decision Support, Dem@Care – FP7 288199

[42] D. Minnen, I. Essa, and T. Starner. Expectation grammars: Leveraging high-level expectations for activity recognition. *In* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 626–632, 2003.

[43] R. Nevatia, S. Hongeng, and F. Bremond. Video-based event recognition: activity representation and probabilistic methods. *In* Computer Vision and Image Understanding, 2, pp. 129-162, 2004.

[44] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In In IEEE Workshop on Event Mining, 2003.

[45] A.T. Nghiem, E. Auvinet, J. Meunier. Head detection using Kinect camera and its application to fall detection. In Proceedings of 11th International Information Science, Signal Processing and their Applications, ISSPA, Montreal, July 2012, p. 164-169.

[46] A. Niculescu-Mizil, C. Rich. Predicting Good Probabilities with Supervised Learning. In Proceedings of the 22nd International Conference on Machine Learning, 625–32. ACM, 2005

[47] N. Oliver, E. Horvitz, and Garg. Layered representations for human activity recognition. *In* Proceedings of IEEE International Conference on Multimodal Interfaces (ICMI), pp. 3-8, 2002.

[48] S. Park and J. Aggarwal. A Hierarchical Bayesian Network for Event Recognition of Human Actions and Interactions. Multimedia Systems, 10:2, pp. 164–179, 2004.

[49] I.S.E. Peraldi, A. Kaya, and R. Möller. Formalizing multimedia interpretation based on abduction over description logic ABoxes. In Description Logics, volume 477 of CEUR Workshop Proceedings, 2009

[50] H. Pirsiavash, D. Ramanan. Detecting Activities of Daily Living in First-person Camera Views, IEEE Conference on Computer Vision and Pattern Recognition, 2012

[51] J.C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, Advances in Large Margin Classifiers, 1999

[52] A. Pronobis, O. M. Mozos, B. Caputo, P. Jensfelt. Multi-modal Semantic Place Classification, The International Journal of Robotics Research (IJRR), Special Issue on Robotic Vision, vol. 29, n°12-3, pp. 298-320, 2010

[53] A. Ranganathan, J. Al-Muhtadi, and R. H. Campbel. Reasoning about uncertain contexts in pervasive computing environments. IEEE Pervasive Computing, 3(2):62–70, 2004

[54] S. Reddy, Y. Gal, and S. Shieber. Recognition of Users Activities Using Constraint Satisfaction. In Springer Berlin/Heidelberg, pages 415–421, 2009.

[55] R. Romdhane, C. Crispim-Junior, F. Bremond and M. Thonnat. Activity Recognition and Uncertain Knowledge in Video Scenes. In the 10th IEEE International Conference on Advanced Video and Signal-Based Surveillance 2013, AVSS 2013, Krakow, Poland on August 27-30, 2013.

[56] R. Romdhane, F. Bremond and M. Thonnat. Complex Event Recognition with Uncertainty Handling. In the 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2010, Boston, 29 August 2010.

[57] S. Rüping. A Simple Method for Estimating Conditional Probabilities for SVMs. LWA 2004:206-210

[58] M.S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. *In* CVPR, pages 1709–1718, 2006.

[59] M.S. Ryoo and J. K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. International Journal of Computer Vision (IJCV), p. 1–24, 2009.

[60] M.S. Ryoo and J. K. Aggarwal. Stochastic Representation and Recognition of High-Level Group Activities, International journal of computer Vision, 2010.

[61] B. Scholkopf, A. Smola, K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem, 1996

[62] M. Shanahan. Perception as abduction: Turning sensor data into meaningful representation. Cognitive Science, 29(1):103–134, 2005.

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

[64] G. Stoilos, G. B. Stamou, J. Z. Pan, V. Tzouvaras, and I. Horrocks. Reasoning with very expressive fuzzy description logics. J. Artif. Intell. Res. (JAIR), 30:273–320, 2007

[65] G. Stoilos, G. B. Stamou, and J. Z. Pan. Fuzzy extensions of owl: Logical properties and reduction to fuzzy description logics. Int. J. Approx. Reasoning, 51(6):656–679, 2010

[66] U. Straccia. Towards a fuzzy description logic for the semantic web (preliminary report). In Proceedings of the Second European conference on The Semantic Web: research and Applications, ESWC’05, pages 167–181, Heraklion, Greece, 2005.

[67] U. Straccia. Managing uncertainty and vagueness in description logics, logic programs and description logic programs. In C. Baroglio, P. A. Bonatti, J. Maluszy ´ nski, M. Marchiori, A. Polleres, and S. Schaffert, editors, Reasoning Web, pages 54–103. Springer-Verlag, Berlin, Heidelberg, 2008

[68] S. Tran and L. S. Davis. Event modelling and recognition using Markov logic networks. In European Conference on Computer Vision, ECCV 08, October 2008.

[69] M. Vacura, V. Svátek, and P. Smrž. A pattern-based framework for uncertainty representation in ontologies. In Proceedings of the 11th international conference on Text, Speech and Dialogue, TSD ’08, pages 227–234, Brno, Czech Republic, 2008. Springer-Verlag

[70] T. Vu, F. Brémond and M. Thonnat, Automatic Video Interpretation: A Novel Algorithm for Temporal Scenario Recognition. The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico, 9-15 August 2003.

[71] Z.L. Wenliang, J.T. Kwok. Accurate Probability Calibration for Multiple Classifiers. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 1939–45. AAAI Press, 2013

[72] T.W. Wlodarczyk, C. Rong, M. O'Connor, and M. Musen. SWRL-F: a fuzzy logic extension of the semantic web rule language. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11, pages 39:1–39:9, Sogndal, Norway, 2011

[73] B. Zadrozny, C. Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 694–99. ACM, 2002

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

A Appendix

A.1. Illustration of activity classes for experiments in Chapter 4

Combing hair, make up, brushing teeth, dental floss, washing hands/face, drying hands/face, laundry, washing dishes, making tea, making coffee, drinking water/bottle, drinking water/tap, making cold food/snack, vacuuming, watching tv, using computer, using cell, reading book.