Noname manuscript No.
(will be inserted by the editor)

Culture as a Sensor? A novel perspective on Human Activity Recognition

Ting-Chia Chiang · Barbara Bruno · Roberto Menicatti · Carmine Tommaso Recchiuto · Antonio Sgorbissa

Received: date / Accepted: date

Abstract Human Activity Recognition (HAR) systems are devoted to identifying, amidst the sensory stream provided by one or more sensors located so that they can monitor the actions of a person, portions related to the execution of a number of a-priori defined activities of interest. Improving the performance of systems for Human Activity Recognition is a long-standing research goal: solutions include more accurate sensors, more sophisticated algorithms for the extraction and analysis of relevant information from the sensory data, and the enhancement of the sensory analysis with general or person-specific knowledge about the execution of the activities of interest.

Following the latter trend, in this article we propose the association and enhancement of the sensory data analysis with cultural information, that can be seen as an estimate of person-specific information, relieved of the burden of a long/complex setup phase.

We propose a culture-aware Human Activity Recognition system which associates the recognition response provided by a state-of-the-art, culture-unaware HAR system with culture-specific information about where and when activities are most likely performed in different cultures, encoded in an ontology. The merging of the cultural information with the culture-unaware responses is done by a Bayesian Network, whose probabilistic approach allows for avoiding stereotypical representations. Experiments performed offline and online, using images acquired by a mobile robot in an apartment, show that the culture-aware HAR system consistently outperforms the culture-unaware HAR system.

Keywords Culture-aware Robotics · Human Activity Recognition · Ontology · Bayesian Network · Vision-based HAR

T. Chiang · B. Bruno · R. Menicatti · C. Recchiuto · A. Sgorbissa
Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, via Opera Pia 13, 16145 Genoa, Italy.
E-mail: antonio.sgorbissa@unige.it

1 Introduction

Human Activity Recognition (HAR) describes the problem of automatically recognising the activities performed by a person from a series of observations of the person's actions and/or environmental conditions [23].

In most cases, the recognition is expected to occur exclusively at the person's home and the activities of interest are related to the so-called Activities of Daily Living (ADL), which are daily activities identified by gerontologists as indicative of the level of autonomy of a person and thus tightly related to his/her quality of life. Commonly considered ADL are those included in the Katz Index of Independence in Activities of Daily Living [22], which addresses basic person needs exclusively (e.g., bathing, eating, drinking, walking, getting up), and in the Instrumental Activities of Daily Living Scale [26], which focuses on the usage of devices and tools of common use (e.g., placing a telephone call, doing the laundry, preparing food).

From a technical point of view, the automated recognition of a set of human activities requires to find structures and methods for describing them in terms of the data provided by available sensors, to make them distinguishable one another [9]. In turn, this requires to identify suitable sensors for the task (i.e., sensors whose data allow for representing the activities of interest in a way that makes them recognisable and distinguishable) and to define how to place and configure them [6].

The identification of the sensor types better suited for the recognition of a given set of human activities is a complex and still unsolved issue. Commonly adopted sensing strategies can be grouped in three categories:

- *Vision*-based HAR systems describe activities in terms of information captured by one or more RGB [32] or RGB-D cameras [2], statically placed in the environment or mounted on a mobile robot [15].
- Body-worn *inertial sensors*, embedded in smartphones [35] and smartwatches [5, 39] and possibly complemented by other wearable sensors [9], have gained increasing popularity for the recognition of activities mostly characterised by gestures, such as sitting down on a chair, getting up from the bed, drinking or brushing one's teeth.
- HAR systems based on distributed *environmental sensors* usually aim at developing cost-effective [12, 34] or unobtrusive [17] solutions for detecting anomalous behaviours or activities requiring the interaction with home appliances and furniture, such as cooking, watching TV or doing the laundry [3].

Many traditional HAR systems only describe activities of interest in terms of the information provided by the chosen sensing approach and implicitly assume that they all always have equal probability of occurring. However, it is intuitively clear that this is not true: bathing, for example, is much more likely to occur in the bathroom than in any other area of the house, while sleeping is much more likely to happen during night hours than at any other time.

In line with this intuition, Coppola et al. [13] propose to exploit temporal and spatial long-term patterns of recurring activities to improve the performance of any state-of-the-art HAR system, independently of the adopted sensing approach. In their framework, activity recognition is formulated as a Bayesian decision making problem, where the response of the HAR system is coupled with the probability of that activity to be occurring at the given time and in the given location.

Information about the relation between activities and context can either be automatically extracted from training data (data-driven approach) [13], or provided by experts (knowledge-driven approach). Following the latter approach, Chen et al. [11] propose a HAR system which relies on an ontology to encode experts' knowledge about the relation between activities, time, location and household objects. In their framework, distributed environmental sensors are linked to specific concepts and properties in the ontology, while subsumption-based reasoning is used to infer the performed activity from the sensors readings.

Let us again consider the relation between activities and contextual information, for example for the bathing activity: one person might prefer baths, while another showers (i.e., use different household objects to perform the same activity), one might perform this activity in the early morning, shortly after waking up, while another in the late evening, just before going to bed (i.e., perform the same activity at different times). Setting aside solutions tailored onto the specific lifestyle of one person only, the management of all variants of a same activity is non trivial: modelling an activity as the *intersection* of all its variants might lead to the creation of very simple models, possibly lacking representative features; conversely, modelling an activity as the *union* of all its variants might lead to the creation of very complex models, possibly hardly distinguishable one from the other. In both cases, the outcome can be a loss in the recognition performance of the HAR system.

Literature has long identified the relation between human activities and *culture* [25]: beside bathing habits [36], meal habits, for example, are known to differ from one country to another in physical (different objects and tools), procedural (different organisation) and social (different conventions and norms) aspects [16, 20]. Intuitively, similar conclusions can be drawn about dressing, self-care habits, etc. We argue that cultural variants of a same activity can thus be considered as a reasonable generalisation of personal variants, since the latter are strongly influenced by the former.

The above intuition is drawn from the definition of culture adopted in the field of transcultural health and social care [31]: "Culture is the shared way of life of a group of people that includes beliefs, values, ideas, language, communication, norms and visibly expressed forms such as customs, art, music, clothing, food, and etiquette. Culture influences individuals' lifestyles, personal identity and their relationship with others both within and outside their culture. Cultures are dynamic and ever changing as individuals are influenced by, and influence their culture, by different degrees".

Concretely, in this article we investigate whether information about customs and habits concerning daily life activities in a given culture can be a reasonable first approximation of a person's lifestyle, and therefore help improve the performance of any HAR system.

To the best of our knowledge, the first attempt at using culture to enhance the recognition of daily-life human activities is the work of Menicatti et al. [28]. Their results suggest that variants of the same HAR system that explicitly take cultural information into account (i.e., that are aware of the variants of activities which are more common in different countries) are more accurate than culture-unaware solutions, and the best performance is obtained when (i) variants of a same activity which are common in different coun-

tries (specifically, *sleeping on a bed*, as usually done in Western countries, and *sleeping on a tatami*, which is common in Japan) are modelled as separate activities and labelled with the corresponding national-level culture, and (ii) information about the user's cultural background (e.g., Japanese) is provided at run-time to the HAR system, leading same-culture variants of an activity to weight more, in the recognition process, than other-culture variants. The latter, albeit deemed unlikely, are not considered impossible, which allows the system to avoid stereotypes.

Building upon the above findings, the main contribution of this article is a modular architecture to equip any state-of-the-art HAR system with national-level cultural information, to be matched online with the user's cultural background. Our architecture relies on: (i) an ontology to link activities with culture-specific contextual information provided by experts; (ii) a Bayesian Network to extend the ontology with probabilistic reasoning and link the knowledge therein with the recognition results provided by the adopted state-of-the-art HAR system. In our experiments we adopt the Microsoft Azure Custom Vision Service[1], to ensure its independence from our framework.

The article is organised as follows. Section 2 discusses relevant works in the literature, with a specific focus on the methodologies adopted by the proposed culture-aware HAR system, while Section 3 defines the context of the work. The proposed system is presented in Section 4 and its subsections. Sections 5 and 6 are devoted to the experimental evaluation, while Section 7 discusses the obtained results. Conclusions follow.

2 Related Work

2.1 Encoding of Culture-related Knowledge in Automated Systems

Beside the afore-mentioned work of Menicatti et al. [28], there is little research done on the modelling of the influence of a person's cultural background over his/her daily activities to the purposes of enhancing their recognition by an automated system. However, a number of studies exploring the interaction between humans and virtual or embodied agents focus on the relation between a person's culture and his/her expectations on the agent's behaviour, and allow for identifying a categorisation in: (i) *bottom-up*, data-driven approaches that aim at extracting national-level information from personal-level data [37]; and (ii) *top-down*, knowledge-driven approaches that aim at encoding available experts knowledge about the relation between culture and human activities [7, 27, 33].

An example of the bottom-up approach is the framework for the learning and selection of culturally appropriate greeting gestures and words proposed by Trovato et al. [37], where an initial set of gestures and words is extracted from video and text corpora, and initial associations between gestures, words and cultural factors are drawn from literature in social studies and expressed as conditional probabilities in a Naive Bayes classifier. At run-time, the user's cultural background, stored as a vector of cultural factors, is used to identify the greeting gestures and words which better match his/her profile. A post-interaction questionnaire is then used as a feedback for the classifier, to allow for an online update of the association between cultural profiles and greeting gestures and words.

Bottom-up approaches require enormous quantities of data to extract meaningful culture-related preferences, and the process of ensuring that the generalisation from person-specific knowledge is correct is not trivial and, at the best of our knowledge, unexplored. Conversely, top-down, knowledge-driven approaches are faced with the problem of identifying suitable structures for the representation of the often heterogeneous and sparse relevant knowledge and its bridging to the chosen application domain. Among the most popular metrics for the description of culture at national level, Hofstede's Dimensions for the Cultural Categorisation of Countries are six scales in which the relative positions of different countries are expressed as a score from 0 to 100 [21]. Knowledge encoded in Hofstede's dimensions has been used for the cultural customisation of the gestures and facial expressions of a virtual agent [33], first-encounter situations [27], and the personal distance to keep from a person during a conversation [7].

An alternative approach is adopted in the framework proposed by Bruno et al. [8] for the encoding of heterogeneous cultural information in the knowledge base of an in-home assistive robot for elderly people, which relies on an ontology with an associated Bayesian Network. Ontologies, which are among the most commonly adopted structures for knowledge representation, allow for the naming and definition of the types, properties, and interrelationships of the entities relevant for the chosen domain. The terminology defining the domain, including classes and their general properties, is stored in the terminological box (TBox) of the ontology, while knowledge that is specific to instances belonging

[1] https://azure.microsoft.com/en-us/services/cognitive-services/custom-vision-service/

to the domain is stored in the assertional box (ABox) of the ontology[2] [19].

In the framework presented by Bruno et al. [8], the TBox contains all relevant classes, regardless of the culture they best (or exclusively) relate to, while instances in the ABox are of one of two types: culture-specific instances encode national-level information, while person-specific instances encode information that is valid for a single user. A special property, called *likeliness* and associated with all classes in the ontology, allows for specifying how likely it is for that concept to hold true for a given cultural group (in the case of culture-specific instances) or individual (in the case of person-specific instances). The Bayesian Network, which is used to discover person-specific knowledge, is initialised with appropriate culture-specific likeliness values and used to propagate the effects of the acquisition of one information onto interconnected instances.

2.2 Knowledge Representation with Ontologies in Human Activity Recognition systems

Ontologies have long been used in the context of the automated recognition of human activities, for a variety of purposes. Ontologies have been used in "smart home" applications to provide a unifying framework for human activities, physical and cognitive diseases, hazards and emergencies, and the responses of a monitoring system [24], as well as to serve as a bridge between the low-level information provided by distributed binary sensors and the high-level descriptions of human activities [29, 34]. The possibility of ontologies to be organised in hierarchical structures, representing knowledge at different levels, has also been exploited to provide the foundation of solutions for concurrent activity recognition [40].

In a survey on the use of knowledge in vision-based HAR systems, Onofri et al. [30] propose the distinction between *a-priori information* about the entities and the structure of the activities of interest (e.g., general knowledge on the human body shape that might be used for a more robust detection of people in video streams), and *contextual information* (such as the aforediscussed time and location that might be used to support the analysis and interpretation of images to the purposes of activity recognition). Examples of the first trend include the use of a hierarchy of ontologies for the fusion of multi-modal vision-based information [14], while, following the latter approach, Banerjee et al. [4] use an ontology to decompose activities of interest into actions and contextual information, and rely on a hierarchy of Fuzzy Inference Systems to climb from the raw visual data to the activities of interest.

[2] The OWL-2 terminology for describing ontologies [38] defines *classes*, *properties* and *individuals*. However, we prefer the term *instance* to *individual* because the latter is commonly used as a synonym of person, which might lead to confusion in this article.

2.3 Probabilistic Reasoning with Bayesian Networks in Human Activity Recognition systems

In the context of activity recognition it often happens that available information does not allow for exactly pinpointing which, if any, of the activities of interest is being performed, rather only inferring that a number of activities of interest might be the one currently performed. Similarly, however rich the representation of a human activity might be, it will never be able to encompass all its possible variants, thus resulting in some occurrences of that activity being closer to the model than others. In both cases, endowing the HAR system with probabilistic reasoning capabilities is crucial to enhance its performance.

In the context of knowledge-driven HAR, since standard ontologies do not allow for probabilistic reasoning, a number of solutions have been devised to overcome this limitation. Beside approaches aiming at extending the ontology itself with mechanisms for dealing with probability, such as PR-OWL [10], a large corpus of literature relies on complementing a standard ontology with a probabilistic reasoner. As an example, Gayathri et al. [18] propose a framework in which activity models described in an ontology are converted into the corresponding first order rules, which are then used to train a Markov Logic Network, while Latfi et al. [24] use the information encoded in the ontology for the initialisation of a Bayesian Network, which performs the analysis of input data to recognise occurrences of the activities of interest. The combination of ontologies and Bayesian Networks is also a common choice in domains other than activity recognition, such as disease diagnosing [1] and assistive robots for elderly care [8].

2.4 Contribution

In our case, the corpus of knowledge describing the influence of culture over daily activities and related contextual elements is heterogeneous, incomplete and typically provided in formats and structures which are not immediately compatible with the data extracted from sensors. As discussed in Sections 2.1 and 2.2, the article proposes ontologies as a particularly fitting solution to model and bridge between the two domains.

The modelled knowledge is independent from the chosen sensing strategy, but necessary for the analysis

of the acquired sensory data: as discussed in Section 2.3, the article proposes Bayesian Networks as a powerful tool to satisfy both requirements, allowing for a smooth and meaningful transition between natural language information and numerical data.

The adoption of a probabilistic approach for the modelling of cultural information also provides another crucial feature: by setting variants which are supposedly very far from the user's cultural background as less probable, but not impossible, we ensure that the proposed system, while relying on national information for its analysis of a person's activities, does not fall into stereotypical representations.

3 Culture-specific Knowledge Retrieval

In this work we rely on the culture-specific knowledge acquired in the course of project CARESSES and encoded in a number of scenarios and guidelines publicly available on the project's website[3]. In line with the CARESSES project, the cultures we consider are Japanese, English and Hindu Indian.

We have identified five cross-cultural activities of interest: `Eating`, `Cooking`, `Sleeping`, `Showering` and `Reading`, plus `Puja praying`, which is a Hindu praying ritual and therefore culture-specific. In the experiments, the tag Others is added to denote any activity which is not among the considered ones. All of the above activities are common daily activities that usually take more than 15 minutes to be completed, which is important to ensure that any chosen HAR system has enough time to perform tasks related to perception and reasoning.

Culture-specific information about the location where activities of interest are usually performed, and the objects they require interaction with, are extracted from the guidelines and used to identify rooms and items of interest. Specifically, the considered rooms are `Kitchen`, `LivingRoom`, `Bathroom`, `Bedroom`, `DiningRoom`, `PujaRoom`, with the latter being specific of the Hindu culture and used for performing the puja praying ritual.

Lastly, since time-related knowledge concerning daily activities is usually expressed in terms of *morning, afternoon* and *evening* [36], for which there are no absolute agreed-upon definitions[4], we follow the *de-facto* rule that organises the 24 hours composing a day in seven periods (see Figure 2). Culture-specific information about typical meal times (e.g., *Lunchtime is usually in between noon and 1.30 p.m. in England, Dinnertime is usually in between 6 p.m. and 8 p.m. in England*, and therefore *Afternoon* can be defined as the time in between 1.30 p.m. and 6 p.m.) is used to define the relation between the periods composing a day and the 24 hours for the three cultures considered, and this framework is used to express time-related knowledge concerning all other activities of interest.

[3] http://caressesrobot.org/en/2018/03/08/caresses-scenarios-and-guidelines-available/

[4] https://en.wikipedia.org/wiki/Afternoon

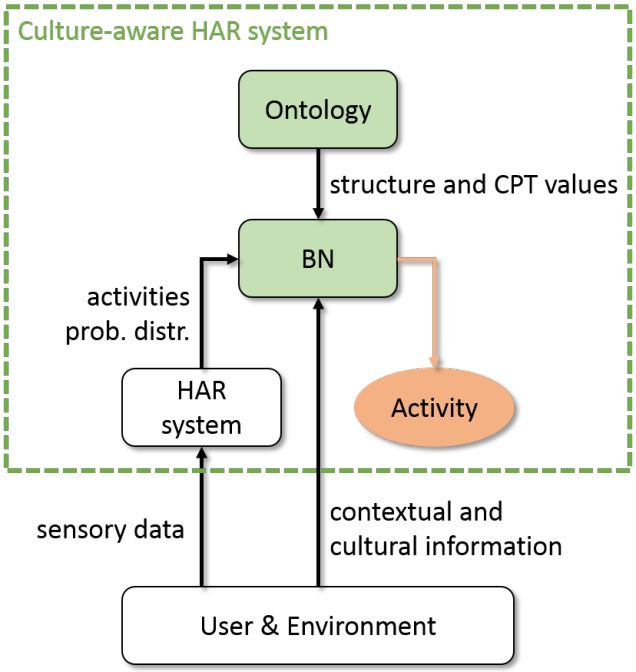

Fig. 1: System architecture. The dashed box denotes the proposed culture-aware HAR system. Green boxes denote the proposed modules for the management of cultural knowledge. White boxes denote the system's input sources, while the orange oval denotes the system's output.

4 System Architecture

The architecture of the proposed framework for culture-aware HAR is sketched in Figure 1. The system is composed of three main modules:

- a state-of-the-art, culture-unaware *HAR system* that, regardless of its chosen sensing strategy, provides in output the probability of each activity of interest to be the one performed by the user, given the available sensory input. This module, although embedded in the culture-aware HAR system, is used as a black-box by the other modules.
- an *ontology* modelling the influence of culture over the activities of interest and related contextual elements. Concretely, the ontology encodes the relation

between culture, time of the day, location within the house and activities of interest, specifying how likely it is for each *culture* × *time* × *location* × *activity* combination to hold true.
– a *Bayesian Network* (BN) providing the culture-aware response on the activity currently performed by the user, having acquired information about the user's cultural background and location, the current time and the culture-unaware guess provided by the state-of-the-art HAR system.

4.1 Ontology for Culture-aware Human Activity Recognition

As anticipated in Section 2, ontologies are composed of *classes* and *properties*, in the TBox, and their *instances* in the ABox. Properties are divided in two categories: *data properties* relate instances of the class to literal data (e.g., strings, numbers), while *object properties* relate an instance to another instance.

In this work, we follow the rationale for the encoding of cultural knowledge in an ontology proposed in [8], which assumes cultural knowledge to be typically provided in natural language and envisions the TBox to define the grammar and vocabulary of the domain of discourse, while the ABox includes all possible "statements" that the TBox allows for composing, each annotated with how likely it is to hold true.

To explain the approach, let us assume that we want to encode the information: *Japanese people usually shower in the evening, while English people usually shower in the morning* [36]. Starting from the above information given in natural language, the TBox of the ontology is designed so that nouns typically correspond to hierarchical classes (e.g., `User`, `Shower`, which is a sub-class of the more general class `Activity`, `Evening`, `Morning`, which are sub-classes of the more general class `PeriodOfTheDay`) and verbs and prepositions to object properties (e.g., `hasActivity <domain=User> <range=Activity>`, linking people to activities, and `hasTime <domain=Activity> <range=PeriodOfTheDay>`, linking activities to moments of the day). A special data property, `hasLikeliness`, describes the relationship between any culture of interest and the concept it refers to.

Once the structure of the information to be encoded has been defined in the TBox, assertions about habits can be added to the ABox as culture-specific instances. We first create the instances `SJP_GEN` (Japanese) and `SEN_GEN` (English) of the class `User`, in line with the naming convention introduced in [8], and then moving to the classes `Activity` and `PeriodOfTheDay` we generate the instances `SJP_SHOWER_EVENING` and `SEN_SHOWER_MORNING`, as well as `SJP_SHOWER_MORNING` and `SEN_SHOWER_EVENING`. By annotating the former two instances with a "high likeliness" and the latter two with a "low likeliness", we encode in the ontology the fact that it is more likely, for a Japanese person, to have a shower in the evening rather than in the morning, while the opposite holds true for an English person.

Concretely, this rationale requires to:

1. encode as top class in the TBox the generic class `Entity`, associated with the data property `hasLikeliness <type=xsd:decimal>`, which allows for annotating each instance of `Entity` or any of its sub-classes with a value in the range $[0, 1]$ representing how likely it is for it to hold true;
2. encode in the TBox the generic class `User`;
3. considering the corpus of relevant knowledge, encode all nouns/concepts as a hierarchy of classes in the TBox. In our case, available knowledge includes information such as *Most people eat their meals in the dining room or in the kitchen*, or *Dinnertime is usually in between 6 p.m. and 8 p.m. in England, and in between 7 p.m. and 9 p.m. in India*. As a consequence, our TBox (see Figure 2) includes all activities of interests, organised as sub-classes of a generic class `Activity`, the different areas of the house in which they take place, organised as sub-classes of a generic class `Room`, and the different moments within a day in which they occur, organised as sub-classes of a generic class `PeriodOfTheDay`. Time expressed in hours is represented by the class `Hour` and its 24 sub-classes `OneAM`,`TwoAM`,`ThreeAM`... not shown in the Figure;
4. considering the corpus of relevant knowledge, encode all predicates, complements and attributes linking one concept to another as object properties in the TBox. In our ontology relevant conceptual links are modelled by: (i) the object property `hasActivity <domain=PeriodOfTheDay,Room> <range=Activity>`, which associates instances of the class `Activity` (or any of its sub-classes) to instances of the class `PeriodOfTheDay` (or any of its sub-classes) and of the class `Room` (or any of its sub-classes), and (ii) the object property `hasPeriod <domain=Hour> <range=PeriodOfTheDay>`, which associates instances of the class `PeriodOfTheDay` (or any of its sub-classes) to instances of the class `Hour` (or any of its sub-classes). In Figure 2, these object properties are shown as orange arrows linking the domain class to the range class.

Concerning the populating of the ABox, the rationale outlined in [8] requires to:

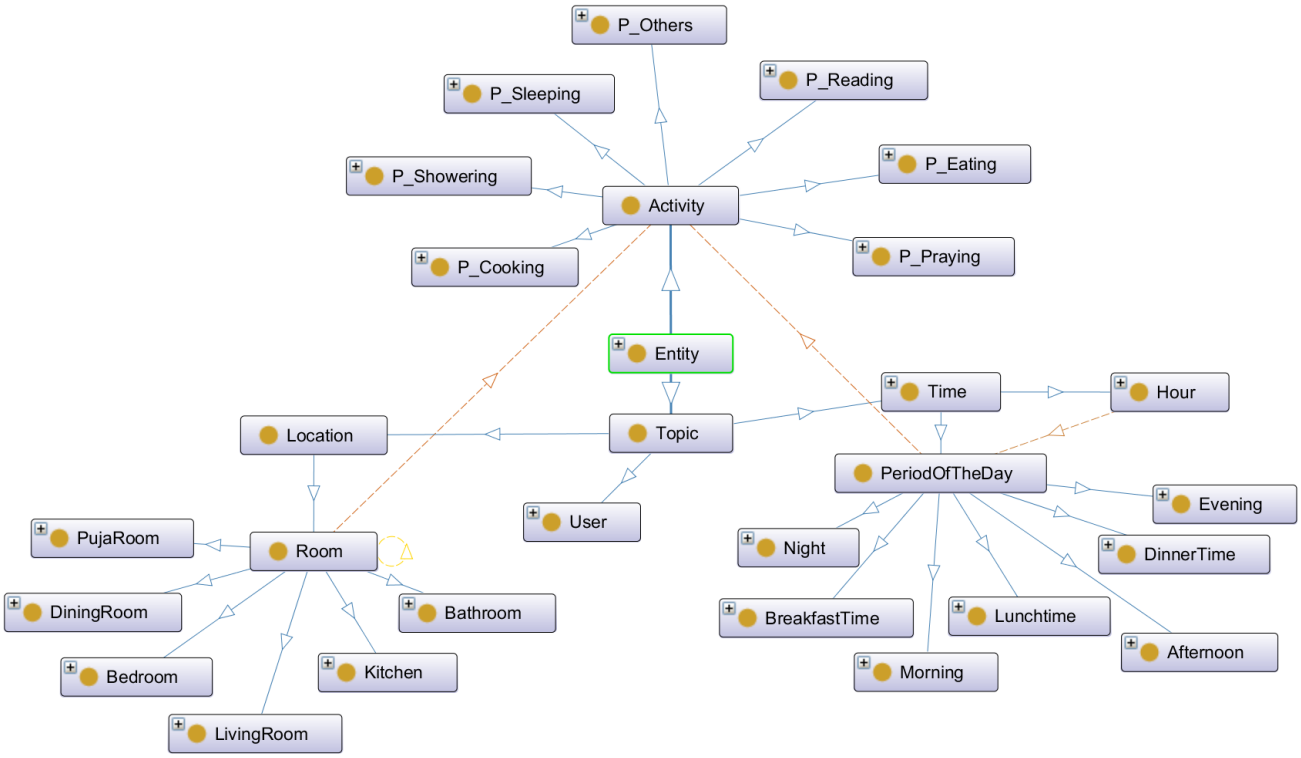

Fig. 2: Key concepts in the TBox of the ontology. Classes (e.g., `Activity`) appear as boxes and hierarchical relationships among them are denoted by blue arrows (e.g., `Cooking` *is a* `Activity`). Object properties (e.g., `hasActivity`, linking `Room` or `PeriodOfTheDay` to `Activity`) appear as dashed arrows, while data properties (e.g., `hasLikeliness`) are not shown.

1. populate the ABox with one instance of `User` per culture of interest (e.g., `SJP_GEN` for the Japanese culture, `SEN_GEN` for the English culture and `SIN_GEN` for the Indian culture, in line with the naming convention introduced in [8]);
2. for each class in the TBox:
   (a) identify all of its callers, i.e., all other classes that have the considered class as filler along an object property, plus the virtual caller *culture*;
   (b) populate the ABox with instances of the considered class corresponding to unique and relevant elements of the cartesian product *culture* $\times$ *(sub-)class* $\times$ *(sub-)caller*$_1$ $\times$ *(sub-)caller*$_2$... In our case, the class `PeriodOfTheDay` and its sub-classes are fillers of the class `Hour` for the object property `hasPeriod`, which is used to model the different meanings given to concepts such as *morning, afternoon* and *evening* by different cultures, while the class `Activity` and its sub-classes are fillers for the object property `hasActivity` both for the `PeriodOfTheDay` and the `Room` classes. Let us consider the class `Morning`, which is a subclass of `PeriodOfTheDay` and therefore filler of the class `Hour` for the culture-related object property `hasPeriod`, and let us assume that our ontology includes the Japanese culture. Possible combinations for instances of `Morning` include `SJP_MORNING_ONEAM`, `SJP_MORNING_TWOAM`, `SJP_MORNING_THREEAM` (according to the naming convention introduced in [8]), respectively representing the concepts of "1 a.m. being considered morning in the Japanese culture", "2 a.m. being considered morning in the Japanese culture", and "3 a.m. being considered morning in the Japanese culture". Similarly, the influence of culture over the rooms composing a person's house is grasped by the culture-specific instances of the class `Room` and its sub-classes, named according to the *culture* $\times$ *(sub-)class* combination they refer to (e.g., `SJP_KITCHEN`, `SEN_BATHROOM`...).
   Figure 3 shows a subset of the instances of `Cooking` whose name specifies the combination they refer to. Concretely, instances of the class `Cooking` represent variants of the activity, per-

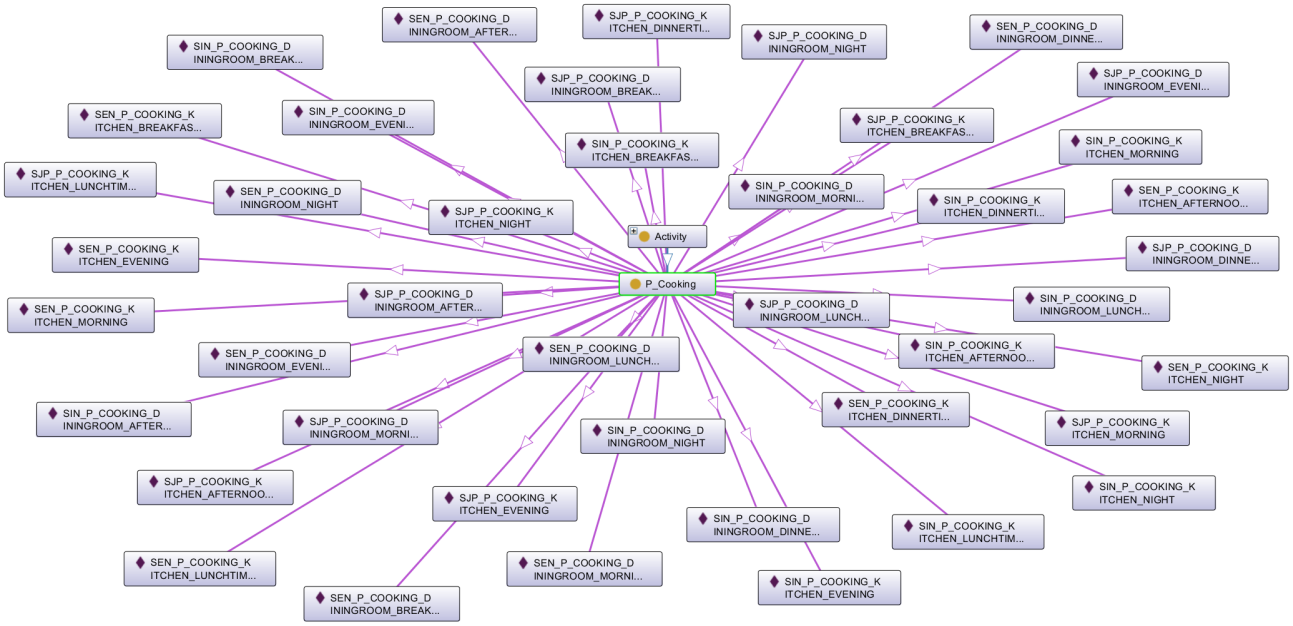

Fig. 3: ABox instances of the `Cooking` class. The names of the instances report the *culture* × *class* × *caller(s)* combination they refer to.

formed at different periods of the day, in different locations and by people of different cultures;

3. fill for each culture-specific instance the `hasLikeliness` property with an estimate of how likely it is for the corresponding *culture* × *(sub-)class* × *(sub-)caller*$_1$ × *(sub-)caller*$_2$... combination to hold true. As an example, the *likeliness* of instance `SJP_MORNING_ONEAM` represents how likely it is for "1 a.m. being considered morning in the Japanese culture", the *likeliness* of instance `SJP_COOKING_KITCHEN_LUNCHTIME` represents how likely it is to cook in the kitchen during lunchtime in the Japanese culture, while the *likeliness* of instance `SJP_KITCHEN` represents how likely it is to have a kitchen in the house in the Japanese culture.

A notable feature of the proposed method for the encoding of knowledge in an ontology is that the influence of culture over other concepts is hidden in the TBox and only made explicit in the ABox. This fact not only allows for the re-use of existing ontologies and vocabularies, with little or no modifications to their TBoxes, but also for the co-existence, in the ABox, of culture-specific and culture-independent instances, which might prove crucial for the cultural enhancing of existing systems.

A last consideration concerning the rationale we followed in organizing culture-specific knowledge in the ontology, which is borrowed from [8], is that it allows for the co-existence in the ABox of culture-specific instances (describing customs and habits related to the activities of interest at a national level) and person-specific instances (describing a person's customs and habits related to the activities of interest, i.e., encoding information related to that person's lifestyle). In [8] we present a framework which, building on an ontology structured as described above, uses the culture-specific information therein encoded to drive the discovery of person-specific information, thus avoiding stereotypes. Future work will be devoted to integrating that framework in the culture-aware HAR system presented in this article, to ensure that the description of customs and habits encoded in the ontology can always converge to the user's lifestyle, however far it is from the starting culture-specific description.

4.2 Bayesian Network for Culture-aware Human Activity Recognition

Figure 4 shows the structure of the Bayesian Network built on the basis of the ontology for culture-aware HAR: classes identified for the creation of culture-specific ABox instances are mapped onto the random variables associated with the culture-related nodes in the Bayesian Network (those above the dotted line in Figure 4), while their sub-classes become the variables' possible states. The class `User` and its culture-specific instances are mapped onto the random variable C representing the user's culture and its possible states.

Culture	Room					
	Bathroom	Bedroom	DiningRoom	Kitchen	LivingRoom	PujaRoom
SIN	0.188	0.197	0.104	0.198	0.188	0.125
SJP	0.204	0.204	0.183	0.204	0.194	0.011
SEN	0.202	0.202	0.182	0.202	0.191	0.021

Table 1: CPT of the R (`Room`) node

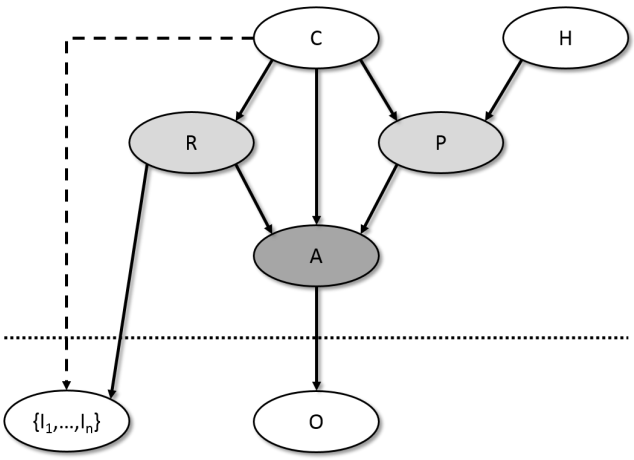

Fig. 4: Structure of the Bayesian Network adopted for culture-aware Human Activity Recognition. The node O below the dotted line corresponds to the output of the culture-unaware HAR system. Nodes above the dotted line encode culture-related information about human activities extracted from the associated ontology. At each time step, the network updates its belief about the performed activity (node A), given evidence about the person's background culture (node C), the current time (node H) and the presence/absence of items in the environment (nodes $\{I_1, ..., I_n\}$), which allow for better assessing the type of room where the person is. Evidence propagates from lighter to darker nodes, i.e., first influencing the nodes describing the user's location (node R) and the current period of the day (node P) and then the node describing the performed activity.

The Conditional Probability Tables (CPT) of nodes C, H, R, P, A are filled with the normalised values encoded in the `hasLikeliness` data property of the corresponding combinations. As an example, Table 1 reports the CPT entries for the class `Room` = R node, computed from the *likeliness* values of all the instances of relevance (e.g., `SIN_BATHROOM`, `SIN_BEDROOM`,... `SEN_PUJAROOM`) and normalised so that each row sums up to 1. Concretely, CPT entries represent the a-priori probability of a room to be each one of the relevant rooms, for the three considered cultures.

As shown in Figure 4, the goal of the Bayesian Network is to provide an updated belief on the activity currently performed by the user (node A - `Activity`), given a culture-unaware estimate provided by any state-of-the-art HAR system (node O, that stands for *Observation*, below the dotted line in the Figure) combined with evidence about the current time (nodes H - `Hour` and P - `PeriodOfTheDay`), the person's background culture (node C - `Culture`) and his/her current location (nodes $I_1, ..., I_n$ - `Items` and R - `Room`).

White nodes in Figure 4 are those for which we assume evidence to be attainable and those which, once the evidence is collected, drive the update of the probability of the other nodes (shown in shades of grey in the Figure). As the Figure shows, we assume evidence about the current time (node H), which influences the current period of the day, to be easily directly accessible, and accurate information about the person's background culture (node C) to be attainable by considering nationality and country of residence. Although state-of-the-art localisation systems allow for obtaining reliable estimates of a person's position within a house, we propose to rely on an object-detection system for on-the-fly identification of the type of room where the user is. Concretely, rooms are defined in terms of the items they contain, and each item of interest is associated with a binary node I_i, which is set to $I_i = true$ if the object is detected in the person's surroundings, and to $I_i = false$ if not. Similarly, we assume the estimate on the current user's activity provided by the chosen culture-unaware HAR system (node O), trained to recognise all of the activities of interest, to be directly accessible and in the form of a probability distribution over all possible activities. The filling of the CPT associated with the I nodes and with the O node is discussed in Section 5.3.

The culture-related knowledge encoded in nodes C, R and P can be interpreted as a *descriptor of culture-specific habits*, that the Bayesian Network combines with run-time, person-specific observations of the user's actions to provide a final response on the user's current activity. The hypothesis driving this work and evaluated in our experiments is that the response informed by the culture-specific knowledge is more accurate (i.e.,

more often correct) than the culture-unaware response initially provided by the state-of-the-art HAR system.

4.3 Computational Complexity

The space complexity of the approach can be computed according to the following rationale.

The proposed procedure to populate the ABox of the ontology with proper instances requires to consider, for each class (e.g., `PeriodOfTheDay`), its N subclasses (e.g., `Morning`, `Afternoon`, ...), its M callers (e.g., `Hour`), the P_i sub-classes of each $caller_i$ (e.g., `1pm`, `2pm`, ...), and finally the number of cultures Q encoded in the ontology. By defining P as the upperbound of P_i, this yield a polynomial complexity $O(NMPQ)$, which determines the number of instances and hence the memory required to encode the desired knowledge in the ontology. The space complexity of the Bayesian Network is $O(NMPQ)$ by construction, since each instance in the ontology has a 1-to-1 correspondence with a CPT entry in the Bayesian Network. This concept is evident in Figure 4: the dimensions of the CPT of a node (e.g., node A) can be computed by multiplying the number of possible events of that node and all its parents.

Finally, it should be noticed that while the structure of the Bayesian Network may change when considering growing ontologies with more concepts and details about the cultural context, its space complexity does not change, since it depends on the procedure according to which the ABox and the CPTs are built. This fact allows for soft real-time Bayesian inference under all the conditions considered in experiments up to now.

5 Experimental Setup

5.1 Rationale

In our experiments, we rely on the Cloud-based, vision-based service provided by Google (Google Vision Services[5]) for the recognition of items of interest, and on the Cloud-based, vision-based HAR system provided by Microsoft (Microsoft Azure Custom Vision Service[1]), which is by design independent from our framework and very easy to train and use.

Figure 5 shows the run-time behaviour of the proposed culture-aware HAR system[6].

[5] https://cloud.google.com/vision/

[6] Publicly available on Github at: https://github.com/TingChiaChiang/Culturally-Competent-Human-Activity-Recognition-with-a-Pepper-Robot.

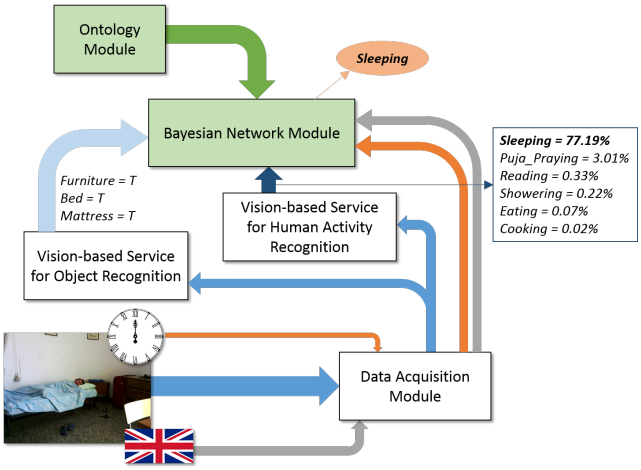

Fig. 5: Example showing the run-time behaviour of the culture-aware HAR system used in our experiments, which relies on a state-of-the-art vision-based culture-unaware HAR system. Images are fed to the Google Vision Service for the recognition of items of interest, and to the Microsoft Azure Custom Vision Service for the culture-unaware activity estimates.

The Data Acquisition Module acquires an image of the user performing an activity and information about the time and the user's nationality. In the case shown in the Figure, the module is fed with the image of a person sleeping in a bed (blue arrow), together with the information that it is midnight (orange arrow) and the user is English (grey arrow). The image is independently analysed by the Google Vision Service to identify all known items of interest (light blue arrow) and by the Microsoft Azure Custom Vision Service to determine the probability of each activity of interest of being the one shown in the image (dark blue arrow). The responses of the two services (reported in *italics* in the Figure) are fed to the Bayesian Network, as evidence for, respectively, the I nodes and the O node, while the information about the user's background culture and the time at which the picture is taken are provided as evidence for the C and H nodes.

As anticipated above, the hypothesis evaluated in our experiments is that the culture-aware response ultimately produced by the A node of the Bayesian Network is more often correct than the culture-unaware observation provided by the external HAR system and mapped onto the Network's O node. Concretely, in our experiments we consider as response, for both systems, the activity with highest probability. In the case of Figure 5, for example, both HAR systems successfully label the input image as an execution of the *sleeping* activity.

The proposed system has been tested in two different conditions: (i) off-line, with images collected from the internet; (ii) online, with images acquired by a Pepper robot (a mobile robot equipped with a camera) in an apartment in Genova, Italy. For this latter case, we have designed a simple program that exploits Pepper's NAOqi APIs[7] for avoiding obstacles and identifying human faces to have it locate the user and move towards him/her, while computing suitable distance and orientation to stand at to take pictures that capture the person and immediate surroundings.

5.2 Training Datasets

The training of the system includes: (a) the specific training of the chosen vision-based modules for the recognition of items of interest and the culture-unaware recognition of human activities, and (b) the training of the Conditional Probability Tables of the I nodes, that relate the items of interest with rooms and culture, and of the O node, that relates the labels provided by the culture-unaware HAR system with the activities of interest. While the former task is dependent on the specific solutions adopted, and possibly unnecessary, the latter, discussed in Section 5.3, is a mandatory key step for the setup of the proposed system.

Two types of training datasets were collected:

1. *Rooms&Items* (10 images per room per culture, collected from the internet via Google search). This dataset is used for the training of the CPTs of the I nodes with the Google Vision Service performance in recognising each item of interest. Cultures have been explicitly taken into account in the collection of this dataset, since one of its purposes is to highlight differences in the objects and furniture that are commonly found within a house in different cultures and encode this information in the culture-aware HAR module. Concretely, we have selected for inclusion in the dataset only those images which closely matched the culture-specific descriptions of rooms given in the guidelines, specifically in terms of presence of items of interest.
2. *Activities* (21 images per activity, collected from the internet via Google search). This dataset is used in k-fold for the training of the culture-unaware HAR system (Microsoft Azure Custom Vision Service) and the training of the CPT of node O with the system's recognition performance. Please notice how cultures have not been taken into account in the collection of this dataset, since this module is oblivious of cultural aspects. Concretely, we have selected images which matched common-sense descriptions of the activities of interest (e.g., the selection criteria for images related to the `Reading` activity is that they include a person with an open book in hand).

[7] http://doc.aldebaran.com/2-4/naoqi/motion/index.html

5.3 Training and Validation of the BN

Figure 6 shows the confusion matrices describing the recognition performance for rooms for the three considered cultures, on the basis of the items of interest identified by the Google Vision Service in the *Rooms&Items* dataset, which allows us to determine whether the service is truly capable of recognising them. As shown in Figure 4, the Bayesian Network requires evidence about both the items of interest and the person's culture to estimate what type of room the user is in. Since all images in the *Rooms&Items* dataset are associated with the corresponding culture, during validation we have set the evidence for the culture node C to the correct one for each image (e.g., *English* for images showing rooms in English houses).

In the matrices of Figure 6, columns represent the actual class of an image (target class), while rows represent the predicted class (output class), i.e., the room given highest probability by the R node of the Bayesian Network, when the Google Vision Service is fed with the image and its output is used to set the evidence for the I nodes. The cells of the matrices contain the number of images with the actual label specified by the column and the predicted label specified by the row, and in brackets, the percentage this number corresponds to over the whole dataset. Green cells denote correct classifications (i.e., the output class corresponds to the target class), while red cells denote incorrect classifications. The grey cells on the bottom of the matrices denote the *recall* rate of each room (i.e., the number of true positive predictions over all true images for that class), while the grey cells on the right of the matrix denote the *precision* rate of each room (i.e., the number of true positive predictions over all predictions of that class). The light blue cell in the bottom-right corner denotes the overall *accuracy* (i.e., the number of true positive predictions over all images).

As the confusion matrices report, `Kitchen`, `DiningRoom`, `Bedroom` and `Bathroom` have good recall in all cultures (specifically, `Kitchen` has perfect recall, `Bathroom` has an average recall rate above 95%, `Bedroom` has an average recall rate above 85% and `DiningRoom` has an average recall rate above 75%), while the recall rate for the `PujaRoom` (only present for the Indian culture) and the `LivingRoom` are relatively poor (both 30%). The reason of this result is that there

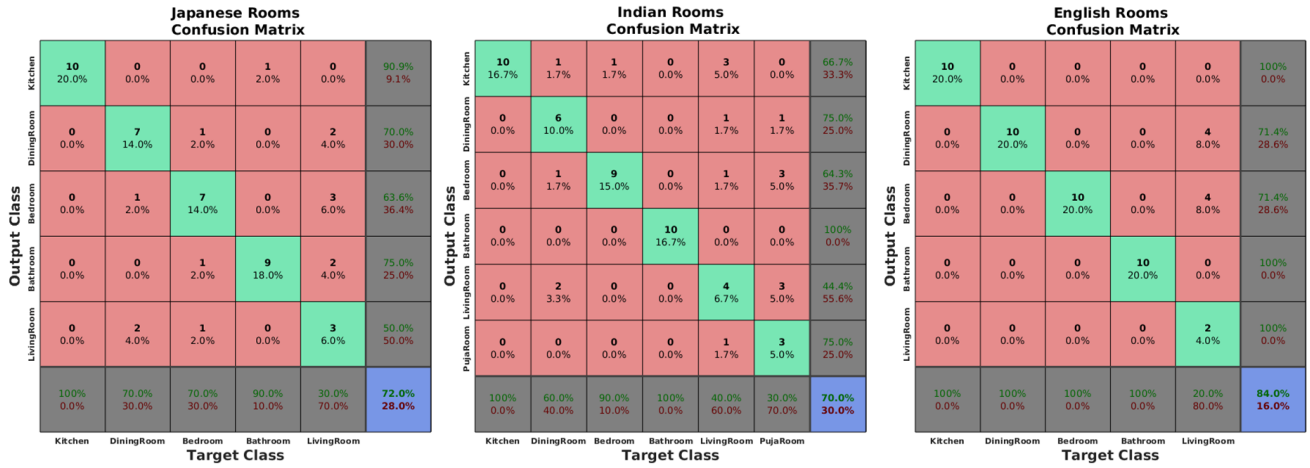

Fig. 6: Recognition performance for rooms on the basis of the items of interest identified by the Google Vision Service in the *Rooms&Items* dataset.

are few highly representative objects reliably associated with these two rooms. Specifically, living rooms are mostly associated by the object recognition system we adopt with furniture items such as chairs, tables, or windows, which however are also present in many other rooms. Conversely, while the Puja room has highly distinctive objects, they are very often not recognised by the object recognition system and therefore only rarely considered by the Bayesian Network in its inference. Concerning precision, it is easy to see that the above discussed problems of the classes `LivingRoom` and `PujaRoom` also affect the precision rate of all other classes (specifically, `Kitchen` has an average precision rate of 85.87%, `DiningRoom` 72.13%, `Bedroom` 66.43%, `Bathroom`, the highest, 91.67%, `LivingRoom`, the lowest, 64.8% and `PujaRoom` 75%).

The overall accuracy for the R node, averaged across the three considered cultures, is 75%.

Table 2 reports the CPT of the O node, built on the basis of the confusion matrix describing the performance of the Microsoft Azure Custom Vision Service trained and validated with k-fold over the *Activities* dataset. As the Table shows, the culture-unaware system achieves very good recognition performance (100% accuracy for all activities).

The services and Bayesian Network thus trained have been used for both the offline and online tests.

6 Experimental Evaluation

6.1 Offline testing

The purpose of the offline testing is to preliminary evaluate and compare the performance of the proposed culture-aware HAR system and the chosen culture-unaware HAR module in the recognition of images describing the six in-home activities of interest. Two types of test sets were collected for the offline testing:

1. *Clean images* (5 images per activity per culture, collected from the internet via Google search). Inclusion criteria for this dataset are the same described for the *Activities* training dataset.
2. *Varied images* (5 images per activity per culture, collected from social media, extracted from YouTube videos, or taken in real life). The purpose of this dataset is to bridge between the "clean" conditions of the still images composing the first test set, and the real-life conditions that are likely to be found in online tests. In particular, images included in this set present (partial) occlusion and high variance in illumination, which are known hindrances for vision-based HAR systems. Examples of such images, collected from Instagram, are shown in Figure 7. Please notice that in this dataset, due to difficulties in finding representative images, images about the `Showering` activity refer to the English culture only.

In all our experiments, we have set the evidence for the H node as the most likely time at which the activity shown in each test image is performed.

Activity	Observation						
	Cooking	Eating	OTHERS	Praying	Reading	Showering	Sleeping
Cooking	0.994	0.001	0.001	0.001	0.001	0.001	0.001
Eating	0.001	0.994	0.001	0.001	0.001	0.001	0.001
Others	0.001	0.001	0.994	0.001	0.001	0.001	0.001
Praying	0.001	0.001	0.001	0.994	0.001	0.001	0.001
Reading	0.001	0.001	0.001	0.001	0.994	0.001	0.001
Showering	0.001	0.001	0.001	0.001	0.001	0.994	0.001
Sleeping	0.001	0.001	0.001	0.001	0.001	0.001	0.994

Table 2: CPT of the O (`Observation`) node

(a) (b) (c) (d) (e)

Fig. 7: Images of the `Reading` activity for the Indian culture, included in the *Varied images* test set.

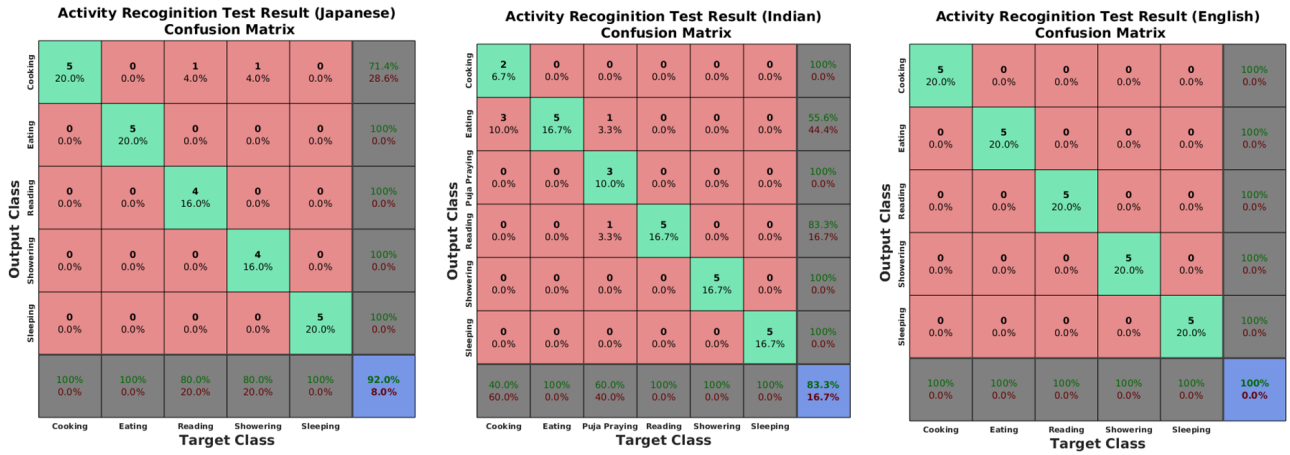

Fig. 8: Confusion matrices of the proposed culture-aware HAR system on the *Clean images* test set.

Figure 8 shows the confusion matrices of the proposed culture-aware HAR system on the *Clean images* test set. As the Figures show, the proposed system achieves good performance with all cultures and activities, in terms of accuracy, precision and recall rates. The worst results are obtained with images of `Cooking` for the Indian culture, which are often misclassified as occurrences of the `Eating` activity (and this yields a low recall rate), and images of `PujaPraying`, which are misclassified as a consequence of the system's poor performance in distinguishing the `PujaRoom` from other rooms (as shown in Figure 6).

Table 3 compares the recognition performance (in terms of overall accuracy) of the proposed culture-aware HAR system and the chosen culture-unaware HAR system over the *Clean images* test set. Concretely, for each execution of the analysis process described in Figure 5 we record the activity with highest probability according to the Microsoft Azure Custom Vision Service (i.e., in the `Observation` node) and the one with highest probability according to the proposed system (i.e., in the Activity node) and mark down whether they are correct or not. As the Table shows, in this case the addition of culture-related information does not appear to

		Culture-unaware HAR system		
		Correct	Incorrect	TOTAL
Culture-aware HAR system	Correct	86.25%	5%	91.25%
	Incorrect	5%	3.75%	8.75%
	TOTAL	91.25%	8.75%	

Table 3: Comparison between the recognition performance of the proposed culture-aware HAR system and the chosen culture-unaware HAR system over the *Clean images* test set.

be very advantageous, since both the culture-unaware HAR system and the proposed culture-aware HAR system achieve a total accuracy of 91.25%.

Figure 9 shows the confusion matrices of the proposed culture-aware HAR system on the *Varied images* test set. As the Figures show, the proposed system retains good performance with all cultures and activities, in terms of accuracy, precision and recall rates. As for the *Clean images* test set, the worst results are obtained with images of `Cooking` for the Indian culture, which are often misclassified as occurrences of the `Eating` activity. The overall accuracy of the system, averaged over all cultures, is 91.43%. However, as shown in Table 4, also in this case it seems that adding culture-related information does not have a significant impact on the performance, since the culture-unaware HAR system achieves an overall accuracy of 92.85%.

6.2 Online testing

Online testing took place in an apartment in Genova, Italy, where volunteers belonging to different cultures performed a number of repetitions of the considered activities. For each activity, the starting location of the mobile robot Pepper is chosen so that the user is always visible, since the robot is not equipped with a map of the environment nor localization functionalities. At start-up the robot relies on audio and visual stimuli coming from the environment (sound, faces and bodies, or movements) to detect the user and orientate towards him/her, then computes the initial distance to the person using the information provided by a depth camera and moves towards him/her along a straight line. Once the distance to the user is below $1.5m$, the robot keeps approaching him/her at a constant speed of $0.1m/s$, stopping every $1.75s$ to take a picture. The maximum distance and the temporal interval have been set experimentally, as suitable trade-offs between collecting as many images as possible, ensuring that there is a perceivable difference between two consecutive images, and ensuring that the image only shows the person and the immediate surroundings. As a consequence, in our experiments an average of 4 images were taken by the robot while approaching the person (see Figure 10).

Differently from the offline testing, in which each image is processed and yields a recognition result, in the online test one result is given by the analysis of all consecutive images acquired by the robot during one approach of the user. Concretely, each image acquired by the robot produces a probability distribution over all activities of interest, and the final result is given by the sum of the probability values given by all images acquired during one approach.

In line with the offline test sets, we analysed five executions of each activity per culture, and only considered the `Showering` activity for the English culture.

Figure 11 shows the confusion matrices of the proposed culture-aware HAR system on the online test. As the Figures show, the proposed system achieves good performance with most activities, in terms of accuracy, precision and recall rates, for all considered cultures. The worst results are obtained with occurrences of the `Eating` activity, which are consistently misclassified as occurrences of the `Cooking` activity, due to the fact that the two activities occur in the same room and therefore share many items of interest. Conversely, occurrences of the `Reading` activity obtain very high precision and recall rates.

Table 5 compares the recognition performance (in terms of overall accuracy) of the proposed culture-aware HAR system, with those of the chosen culture-unaware HAR system. As the Table shows, in this case the cultural information is particularly beneficial for the recognition performance: while the culture-unaware HAR module alone correctly labels only 42.3% of all executions, the proposed system correctly labels 70.4% of all executions, suggesting that the culture-dependent nodes can grasp hints related to the performed activity that can lead to a correct recognition even when the main recognition system fails (as the Table shows, in 30.1% of the executions although the culture-unaware HAR module is wrong, the culture-aware HAR system is able to provide a correct classification). Conversely, only in 2% of all executions the analysis of culture-dependent information actually worsens the perfor-

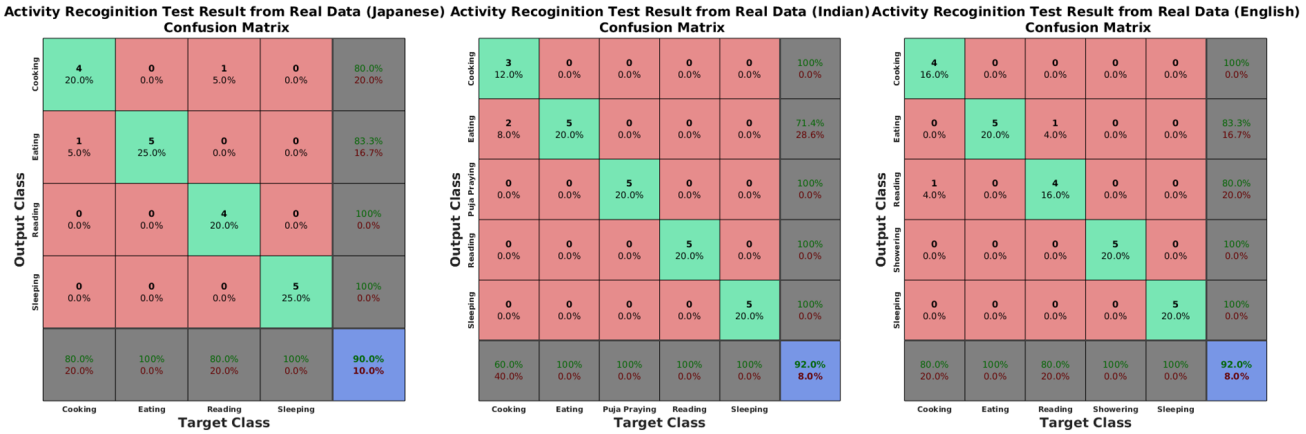

Fig. 9: Confusion matrices of the proposed culture-aware HAR system on the *Varied images* test set.

		Culture-unaware HAR system		
		Correct	Incorrect	TOTAL
Culture-aware HAR system	Correct	87.14%	4.29%	91.43%
	Incorrect	5.71%	2.86%	8.57%
	TOTAL	92.85%	7.15%	

Table 4: Comparison between the recognition performance of the proposed culture-aware HAR system and the chosen culture-unaware HAR system over the *Varied images* test set.

(a) (b) (c) (d)

Fig. 10: Sequence of pictures taken by the robot while approaching a person performing the `Eating` activity.

mance, turning a correct classification from the culture-unaware HAR module into an incorrect overall result of the culture-aware HAR system.

7 Discussion

A number of considerations arise from the analysis of the offline and online experiments.

First of all, while a general performance drop was expected in the online test, due to the significant differences between the pictures acquired by the robot and those used for the training of the HAR systems, the drop is much bigger for the culture-unaware HAR system. The performance gap between the two systems in the online test (+30% overall accuracy for the culture-aware HAR system) can be viewed as a measure of the importance that accurate a-priori information about context can assume to retain good performance even in unforeseen situations, and we believe that culture-specific knowledge about customs and habits concerning daily life activities (including, but not limited to, the times and locations in which they are commonly performed) can provide such accurate background.

At the same time, the performance of the two systems are nearly identical in the offline case. We hypothesise that this similarity might have two causes: 1) the limited size of the *Activities* dataset used for the training and validation of the culture-unaware HAR system and the filling of the CPT associated with the O node (see Table 2), which led the culture-unaware HAR sys-

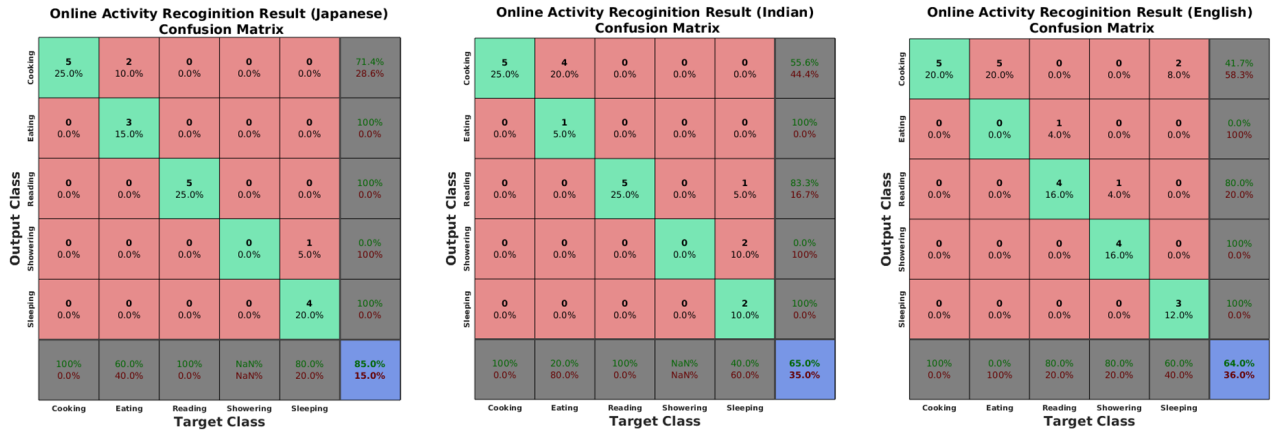

Fig. 11: Confusion matrices of the proposed culture-aware HAR system on the online test.

		Culture-unaware HAR system		
		Correct	Incorrect	TOTAL
Culture-aware HAR system	Correct	40.3%	30.1%	70.4%
	Incorrect	2%	27.6%	29.6%
	TOTAL	42.3%	57.7%	

Table 5: Comparison between the recognition performance of the proposed culture-aware HAR system and the chosen culture-unaware HAR system in the online test.

tem to achieve perfect recognition performance over the validation dataset and therefore the O node and the whole culture-aware HAR system to heavily rely on the initial guess it provides; 2) the limited size of both testing datasets, which, together with the above consideration, does not allow for differences between the culture-aware and the culture-unaware HAR system to emerge. To overcome both issues, we plan to expand the training and testing datasets, also including anonymised images taken in a number of care homes in the UK and in Japan, with English, Indian and Japanese volunteers, in the context of the experiments performed within the CARESSES project.

The online experiments have also highlighted the importance of selecting suitable distances and orientations for the robot to stand at to take pictures during the approach to the user. While the sequence of snapshots shown in Figure 10 suggests that moving along a straight line allows for meaningful images in some cases, it has the limitation of constraining all acquired images to show the same background, and therefore possibly leave out important contextual element. To overcome this issue, we are exploring the use of a parabolic approach trajectory in place of the straight one, so that the sequence of images taken during the approach would better span the background on the left and on the right of the user.

On a more general level, this work and the reported experiments confirm the intuition discussed in the Introduction that contextual information can play a crucial role to the purposes of Human Activity Recognition, and take a step further along that line suggesting that such contextual information is even more effective if described taking *culture* into account.

Reliable a-priori estimates of a person's habits can be crucial to allow for good recognition performance when little user-specific information is available, and there is abundant Literature evidence on the influence of one's background culture on his/her lifestyle, especially when home activities are concerned. In this respect, a limitation of this work is the use of static labels to identify the user's background culture: in our experiments, users are set to be English, or Indian, or Japanese, according to their nationality, and assumed to perfectly match their background culture. This assumption has two problems: 1) information about the user's cultural background might not always be available at system setup; 2) people rarely perfectly match their background culture, since one's lifestyle is influenced, beside culture, by person-specific factors such as personality and life experiences.

As anticipated in Section 4.1, the rationale for managing cultural knowledge adopted in this article is taken from [8], which presents a framework allowing a robot

to estimate a person's culture on the basis of national-level cultural information and verbal interaction. The compatibility between the two works is a precise design choice. The integration of that framework in the proposed culture-aware HAR system would allow for solving both of the above problems, and truly let culture-related person-specific information, captured over time by the robot, act as a sensor providing more and more precise information about the user's habits and preferences, which in turn make the HAR system better and better in recognizing the performed activities.

The culture-aware HAR system proposed in this article proves that it is possible to capture and model culture-related knowledge relevant for the description of daily-life activities in a way that is compatible with state-of-the-art HAR system, and the reported preliminary tests performed suggest that the use of such information, even at a simplistic level, allows for an improvement in the recognition performance.

8 Conclusions

In this article we investigate whether and how the explicit modelling of culture-specific information related to daily in-home activities can improve the performance of state-of-the-art Human Activity Recognition systems. We propose to encode the cultural information in an ontology whose structure, inspired by [8], allows for defining concepts of relevance, regardless of their relation with culture, in the TBox of the ontology, and for specifying the relevance of such concepts for each considered culture with instances in the ABox. The ontology is associated with a Bayesian Network, which performs a culture-aware activity recognition by combining the stored culture-specific information with the preliminary assessment provided by a culture-unaware HAR module.

To test our hypothesis we have acquired information about the execution of five common daily-life activities (`Eating`, `Cooking`, `Sleeping`, `Showering` and `Reading`) in the English, Japanese and Hindu Indian cultures, together with an activity (`Puja praying`, a Hindu praying ritual) which is specific of one culture only. Collected culture-specific information include information about the *time* at which activities are typically performed, and the *location* where they are performed, defined by the items that characterise it.

In our tests we assume information about the user's background culture and the current time to be directly accessible by the proposed culture-aware HAR system, while we rely on two external modules for the identification of the user's location and for the preliminary, culture-unaware assessment. Concretely, we rely on the vision-based, Cloud-based Google Vision Service for the recognition of items of interest, and on the vision-based, Cloud-based Microsoft Azure Custom Vision Service for the preliminary culture-unaware labelling.

We have compared the performance of the chosen culture-unaware HAR system and the proposed culture-aware HAR system both offline, using images taken from the web, and online, using images acquired by a mobile robot equipped with a camera in an apartment in Genova, Italy. In all tests the proposed culture-aware HAR system achieves good performance and in the online test it is significantly better than the culture-unaware HAR system, thus suggesting that the enhancement of Human Activity Recognition systems, regardless of the sensing strategy they adopt, with culture-specific information, is a simple and effective method for improving the recognition performance.

Compliance with Ethical Standards

Funding This work has been partially supported by the European Commission Horizon2020 Research and Innovation Programme under grant agreement No. 737858 (CARESSES), and by the Erasmus+ programme under grant agreement No. 2014-2616/001-001 (EMARO+).

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Agarwal P, Verma R, Mallik A (2016) Ontology based disease diagnosis system with probabilistic inference. In: Information Processing (IICIP), 2016 1st India International Conference on, IEEE, pp 1–5
2. Aggarwal JK, Xia L (2014) Human activity recognition from 3d data: A review. Pattern Recognition Letters 48:70–80
3. Bakar U, Ghayvat H, Hasanm S, Mukhopadhyay S (2016) Activity and anomaly detection in smart home: A survey. In: Next Generation Sensors and Systems, Springer, pp 191–220
4. Banerjee T, Keller JM, Popescu M, Skubic M (2015) Recognizing complex instrumental activities of daily living using scene information and fuzzy logic. Computer Vision and Image Understanding 140:68–82, DOI 10.1016/j.cviu.2015.04.005
5. Bruno B, Mastrogiovanni F, Sgorbissa A, Vernazza T, Zaccaria R (2013) Analysis of human behavior recognition algorithms based on acceleration data. In: ICRA 2013, pp 2293–2299
6. Bruno B, Mastrogiovanni F, Sgorbissa A (2014) A public domain dataset for ADL recognition using

wrist-placed accelerometers. In: 23rd IEEE International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN 2014)
7. Bruno B, Mastrogiovanni F, Pecora F, Sgorbissa A, Saffiotti A (2017) A framework for culture-aware robots based on fuzzy logic. In: Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on, IEEE, pp 1–6
8. Bruno B, Recchiuto CT, Papadopoulos I, Saffiotti A, Koulouglioti C, Menicatti R, Mastrogiovanni F, Zaccaria R, Sgorbissa A (2019) Knowledge Representation for Culturally Competent Personal Robots: Requirements, Design Principles, Implementation, and Assessment. International Journal of Social Robotics DOI 10.1007/s12369-019-00519-w
9. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR) 46(3):33
10. Carvalho RN, Laskey KB, Costa PC (2017) Prowl–a language for defining probabilistic ontologies. International Journal of Approximate Reasoning 91:56–79
11. Chen L, Nugent CD, Wang H (2012) A Knowledge-Driven Approach to Activity Recognition in Smart Homes. IEEE Transactions on Knowledge and Data Engineering 24(6):961–974
12. Cook DJ, Crandall AS, Thomas BL, Krishnan NC (2013) Casas: A smart home in a box. Computer 46(7):62–69
13. Coppola C, Krajńík T, Duckett T, Bellotto N (2016) Learning temporal context for activity recognition. Frontiers in Artificial Intelligence and Applications 285:107–115
14. Crispim-Junior CF, Buso V, Avgerinakis K, Meditskos G, Briassouli A, Benois-Pineau J, Kompatsiaris I, Bremond F (2016) Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(8):1598–1611, DOI 10.1109/TPAMI.2016.2537323
15. Faria DR, Vieira M, Premebida C, Nunes U (2015) Probabilistic human daily activity recognition towards robot-assisted living. In: Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, IEEE, pp 582–587
16. Fjellstrm C (2004) Mealtime and meal patterns from a cultural perspective. Scandinavian Journal of Nutrition 48(4):161–164, DOI 10.1080/11026480410000986
17. Froehlich JE, Larson E, Campbell T, Haggerty C, Fogarty J, Patel SN (2009) Hydrosense: infrastructure-mediated single-point sensing of whole-home water activity. In: Proceedings of the 11th international conference on Ubiquitous computing, ACM, pp 235–244
18. Gayathri K, Easwarakumar K, Elias S (2017) Probabilistic ontology based activity recognition in smart homes using Markov Logic Network. Knowledge-Based Systems 121:173–184, DOI 10.1016/j.knosys.2017.01.025
19. Guarino N, et al (1998) Formal ontology and information systems. In: Proceedings of FOIS, pp 81–97
20. Guptill AE, Copelton DA, Lucal B (2017) Food and society: Principles and paradoxes. John Wiley & Sons
21. Hofstede G, Hofstede GJ, Minkov M (1991) Cultures and organizations: Software of the mind, vol 2. Citeseer
22. Katz S, Chinn A, Cordrey L (1959) Multidisciplinary studies of illness in aged persons: a new classification of functional status in activities of daily living. Journal of Chronic Disease 9(1):55–62
23. Kim E, Helal S, Cook D (2010) Human activity recognition and pattern discovery. IEEE Pervasive Computing/IEEE Computer Society & IEEE Communications Society 9(1):48
24. Latfi F, Lefebvre B, Descheneaux C (2007) Ontology-based management of the telehealth smart home, dedicated to elderly in loss of cognitive autonomy. In: OWLED, vol 258
25. Law M (1993) Evaluating activities of daily living: directions for the future. American Journal of Occupational Therapy 47:233–237
26. Lawton M, Brody E (1969) Assessment of older people: self-maintaining and instrumental activities of daily living. The Gerontologist 9:179–186
27. Lugrin B, Frommel J, André E (2015) Modeling and evaluating a bayesian network of culture-dependent behaviors. In: Culture Computing 2015, pp 33–40
28. Menicatti R, Bruno B, Sgorbissa A (2017) Modelling the influence of cultural information on vision-based human home activity recognition. In: 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp 32–38
29. Okeyo G, Chen L, Wang H, Sterritt R (2014) Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. Pervasive and Mobile Computing 10(PART B):155–172, DOI 10.1016/j.pmcj.2012.11.004
30. Onofri L, Soda P, Pechenizkiy M, Iannello G (2016) A survey on using domain and contextual

knowledge for human activity recognition in video streams. Expert Systems with Applications 63:97–111, DOI 10.1016/j.eswa.2016.06.011
31. Papadopoulos I (2006) Transcultural health and social care: development of culturally competent practitioners. Elsevier Health Sciences
32. Poppe R (2010) A survey on vision-based human action recognition. Image and vision computing 28(6):976–990
33. Rehm M, Bee N, Endrass B, Wissner M, André E (2007) Too close for comfort?: adapting to the user's cultural background. In: HCM 2007, pp 85–94
34. Scalmato A, Sgorbissa A, Zaccaria R (2013) Describing and Recognizing Patterns of Events in Smart Environments With Description Logic. IEEE Transactions on Cybernetics 43(6):1882–1897, DOI 10.1109/TSMCB.2012.2234739
35. Shoaib M, Bosch S, Incel O, Scholten H, Havinga P (2015) A survey of online activity recognition using mobile phones. Sensors 15(1):2059–2085
36. Soo-Hoo F (2016) How women around the world get clean. URL `https://www.refinery29.com/en-us/2016/01/101925/cultural-differences-women-showering`
37. Trovato G, Ham JR, Hashimoto K, Ishii H, Takanishi A (2015) Investigating the effect of relative cultural distance on the acceptance of robots. In: ICSR 2016, pp 664–673
38. W3C Owl Working Group and others (2009) OWL 2 web ontology language document overview
39. Weiss GM, Timko JL, Gallagher CM, Yoneda K, Schreiber AJ (2016) Smartwatch-based activity recognition: A machine learning approach. In: Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on, IEEE, pp 426–429
40. Ye J, Stevenson G, Dobson S (2015) KCAR: A knowledge-driven approach for concurrent activity recognition. Pervasive and Mobile Computing 19(2):47–70, DOI 10.1016/j.pmcj.2014.02.003