# On-line Human Activity Recognition from Audio and Home Automation Sensors: comparison of sequential and non-sequential models in realistic Smart Homes

Pedro Chahuara, Anthony Fleury, François Portet, Michel Vacher

# On-line Human Activity Recognition from Audio and Home Automation Sensors: comparison of sequential and non-sequential models in realistic Smart Homes [1]

Pedro Chahuara [a,*], Anthony Fleury [b] François Portet [a,**] and Michel Vacher [a]
[a] *Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France*
*CNRS, LIG, F-38000 Grenoble, France*
*E-mail: {pedro.chahuara,francois.portet,michel.vacher}@imag.fr*
[b] *Univ. Lille, F-59000 Lille, France*
*Mines Douai, IA, F-59508 Douai Cedex, France*
*E-mail: anthony.fleury@mines-douai.fr*

**Abstract.** Automatic human Activity Recognition (AR) is an important process for the provision of context-aware services in smart spaces such as voice-controlled smart homes. In this paper, we present an on-line Activities of Daily Living (ADL) recognition method for automatic identification within homes in which multiple sensors, actuators and automation equipment coexist, including audio sensors. Three sequence-based models are presented and compared: a Hidden Markov Model (HMM), Conditional Random Fields (CRF) and a sequential Markov Logic Network (MLN). These methods have been tested in two real Smart Homes thanks to experiments involving more than 30 participants. Their results were compared to those of three non-sequential models: a Support Vector Machine (SVM), a Random Forest (RF) and a non-sequential MLN. This comparative study shows that CRF gave the best results for on-line activity recognition from non-visual, audio and home automation sensors.

Keywords: Activity Recognition, Markov Logic Network, Statistical Relational Learning, Smart Home, Ambient Assisted Living

## 1. Introduction

Automatic human Activity Recognition (AR) is an important process for human behaviour monitoring but it is also extensively studied for the provision of context-aware services for smart objects (smartphones, robots…) and smart spaces (smart homes, smart rooms, public spaces…)[20]. Smart Homes in particular have become a topic of increasing interest since they are a promising way to improve the daily life of people with loss of independence (elderly people or people with physical or cognitive disabilities) so that they always keep control over their lives and continue to live independently, to learn and to stay involved in social life. These technologies can also improve the life of the carers (who are often close relatives) by reducing the human and financial burden of such situations [71,30,57,65].

Many projects related to Smart Homes have been supported by national and international research foundations to address the challenges imposed by a growing elderly population such as ADAPTIVE HOUSE [51], AWAREHOME [31], C@SA [22], CIRDO [9], GER'-HOME [88] MAVHOME [19], PLACELAB [35], or

[1] This work is part of the SWEET-HOME project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09-VERS-011).

[*] Pedro Chahuara is now at European Commission Joint Research Centre, Ispra,Italy

[**] Corresponding author. Tel.: +33457421544

SWEET-HOME [77]. All of them have integrated human activity modelling and recognition in their systems.

Most of the progress made in the AR domain came from the computer vision domain [1]. However, the installation of video cameras in the user's home is not only raising ethical questions [72], but is also rejected by some of the targeted population [59][1]. Moreover, video processing is highly sensitive to light conditions which can dramatically vary in a home. Other approaches rely on information from RFID tags [32] and wearable devices [87]. In the first case, putting RFID tags on objects makes the maintenance of Smart Homes burdensome since any new object implies technical manipulations to fix the corresponding sensor and to configure it. The case of wearable sensors is sometimes not applicable when inhabitants do not want (or forget) to wear sensors all the time. Moreover, cost and dissemination of assistive technologies would be better if they were built on standard home automation technologies with minimal technical additions. This is why there is an increasing interest in automatic human activity recognition from home automation sensors [72,83,70,14,13,63,85,55]. This type of environment imposes constraints on the sensors and the technology used for recognition. Indeed, information provided by the sensors for activity recognition is indirect (no worn sensors for localisation), heterogeneous (numerical or categorical, continuous or event based), noisy, and non-visual (no camera). This application setting calls for new methods for activity recognition which can deal with the poverty and unreliability of the provided information and can process streams of data. Moreover, these models should be checkable by humans and linked to domain knowledge.

In home automation sensor based AR, the problem has often been approached using off-line machine learning methods on pre-segmented activity intervals [6,54,26]. In that case the entire information (past, present, future) is considered to be accessible and the detection problem is ignored (i.e. detecting when an activity starts and ends). If such an approach is valid for off-line analyses of human behaviour, many real-world applications will need real-time or at least on-line AR. For instance, context aware systems must know, at the time of the user's interaction, which activities the user is performing. This task is more difficult than the off-line one as only present and past information can be used and classification must be provided within a reasonable time. Another issue is that the system must deal with activities that are not known a priori to avoid undesirable behaviours.

In this paper, we present an on-line activity recognition method for AR within homes in which multiple sensors, actuators and home automation equipments coexist. This research is carried out as part of the SWEET-HOME [77] project which aims at developing a complete framework to enable voice command in Smart Homes. In this framework, the interpretation of the commands and the decisions to be made depend on the context in which the interaction occurs. This context is composed, among other information, of the user's current activity. For instance, if the user utters "Turn on the light", the best action, if she is awaking in the middle of the night (respectively if she is dressing in the morning), could be to provide low intensity light using the bedside lamp (respectively high intensity light using the ceiling lamp). To perform on-line AR in Smart Homes from audio and home automation sensors, a framework was developed to summarise the stream of data into temporal windows and classify each window into one known class, or into a specifically defined `Unknown` class. This research brings the following contributions:

1. The integration of audio signals with home automation sensors for AR is an understudied area. This work not only demonstrates the interest of such fusion but also brings the first complete datasets for AR that contain home automation data as well as audio signals with multiple users. These datasets are available to the community [80,27]. Some of them were acquired during experiments in a realistic smart home involving elderly and visually impaired people [76].
2. The framework for on-line AR makes it possible to summarise asynchronous as well as continuous sampled signals into temporal windows.
3. The paper introduces a recent model for AR — Markov Logic Network— in both sequential and non-sequential versions. Moreover three sequential and two other non-sequential models for the AR task were tested and compared.
4. These models were evaluated on the above-mentioned datasets in a realistic way since windows of unknown class are fed to the classifiers. Indeed, in real world setting not all possible ac-

[1] As for any technology, video cameras can be very well accepted if the benefit is perceived to be higher than the feeling of intrusion.

tivities can be learned thus applications must be able to handle unforeseen situations. Moreover, to avoid overfitting, a cross validation technique was designed so as to exclude from the learning set, the participants' records used for testing.

The paper is organised as follow. After a short description of the AR classification techniques in Section 2, the framework for on-line AR is detailed in Section 3. In particular this section introduces three sequence based models namely the Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) and finally the Markov Logic Networks (MLNs), a statistical relational method that combines high expressibility (first order logic) with the handling of uncertainty. The methods were tested in two experiments performed in two real Smart Homes involving more than 30 participants. Moreover, their results were compared to those of state of the art non-sequential models such as are Support Vector Machine (SVM) and Random Forests (RF). These experiments and the corresponding results are described in Section 4. The paper ends with a discussion in Section 5 and a short conclusion in Section 6.

## 2. Related Work

In the literature, Activity Recognition (AR) has been defined differently according to the level of granularity under consideration. In some works, for instance, a movement such as standing up, running or walking is considered as an activity [39]. As the activity to be recognized depends on the movement of the body, worn sensors are often used. This can be found in research concerning medical assessment [45] or daily activity interpretation [58]. Some other works consider the variation of a certain task: making tea, coffee, or preparing a meal [56]. In such cases, each activity is a specialization of a general task, and frequently the accuracy of the recognition is related to the number and type of the applied sensors since some subtask can only be recognized by the use of a particular sensor. In some applications of surveillance in public places, activities are considered as interactions among people. For instance, complex activities such as fighting and stealing are identified by means of video recognition techniques [47,67,3].

Besides the level of granularity of the activity, the way to perform the recognition can be divided into off-line and on-line. The former case consists of the analysis of a static set of data [6]. For example, when assessing the health state of a patient in a hospital, the sensor data of a previous time-span can be used to recognize the corresponding activities or to identify a change of behaviour. The advantage of such an analysis is that all temporal relations can be exploited allowing better accuracy since for every instance past and future events are available. In on-line recognition [34,40], the case we focus on, the analysis is done from a data stream while the subject is performing the activity. In this case the aim is to identify as quickly as possible the current activity at a certain instant relying only on past and present information.

Approaches for activity modeling can be divided mainly into two categories: knowledge-driven and data-driven. In the former category, a logic-based approach offers an ideal framework to model explicit knowledge which can be provided by an expert of the domain. Ontologies have been widely used for AR [16] since they provide readability and formal definitions while the inference can be performed by an ontology reasoner as a problem of satisfiability. Moreover, under a description-based approach, logic rules facilitate the implementation of expert knowledge within a model [70]. For instance, Augusto and Nugent [5] used logical models to represent the temporal relations among events to recognise activities. In Artikis *et al.* [3], Event Calculus (EC) has been used for AR because of its capacity to model complex activity and temporal relations. EC has also been used for behaviour reasoning by Chen *et al.* [15] in a framework aiming at assisting a person in a smart environment. Though logical approaches are highly expressive, they do not handle uncertainty whereas input data in smart home are highly noisy.

Data-driven approaches can be either unsupervised or supervised. Unsupervised activity recognition is pertinent when it is not required to recognize specific activities; for instance, in applications intended to recognize a change in the daily pattern of the inhabitant. Some relevant works [49,18] have studied methods to discover recurrent patterns, or motifs, from a stream of sensor data; other approachs consider the segmentation and clustering of the data in order to create models that can subsequently label a segment in one of the clusters [62,24].

In the case of supervised learning methods, the AR model is learnt by means of an annotated corpus. In most cases the training corpus is exploited in order to find the best parameters of the model. However the structure of the model can also be inferred automatically, for instance, by the induction of logical rules [4]. Many works have applied statistical methods in order to classify sets of sensor data produced over a short time interval as belonging to a particular activity [26,11]. As information in pervasive environments is uncertain in most cases, probabilistic approaches are suitable candidates to be applied for AR, although they assume a probabilistic independence between consecutive time intervals, which is often a false assumption. One of the most applied methods to include temporal relations in the model is dynamic Bayesian networks [83,86]. Activity recognition has also been treated like a problem of sequence labeling: to label a segment of sensor data into the most probable activity performed. Thus, modeling activities by Hidden Markov Models (HMMs) is extensive [23,52,84]. For instance, Duong *et al.* [23] extended a conventional HMM to model the duration of an activity, and Naeem *et al.* [52] defined activities as a composition of tasks modeled by hierarchical HMMs. During recent years, conditional random fields (CRFs) [42] have also been widely applied to AR. In particular, Chieu *et al.* [17] presented an application of CRFs for AR using physiological data. Nazerfard *et al.* [54] and Vail *et al.* [81] showed that CRFs can give better results than HMMs since they do not assume the probabilistic independence of the observation variables. Tong and Chen presented a method using Latent-Dynamic CRF for recognizing activities in smart homes [74].

Recently, Statistical Relational Learning (SRL) [29], a sub domain of machine learning, has gained much attention as it integrates elements of first order logic and probabilistic models. Under the SRL scheme, models are defined in a formal logical language that makes them reusable and easy to verify, that systematically takes uncertainty into account, and that allows easy inclusion of *a priori* knowledge. SRL has recently attracted attention in the domain of human activity modelling and recognition. For instance, Logic HMMs [38] and relational Markov networks [73] are both SRL methods that were considered for AR [53,46,58,33]. In our work, we applied Markov Logic Networks (MLN) [66], which become Markov networks when their predicates are grounded during the inference process. It is also possible to define a MLN which is equivalent to a dynamic model such as a linear CRF.

Some other researchers have carried out comparative studies on the application of machine learning methods [81,2]. However these works have focused mainly on the properties of the methods that make some of them more appropriate for AR than others. We consider it essential to extend these studies through the analysis of the inherent characteristics of the problems relative to this recognition task, such as the most influential sensor information for AR, or the importance of historical information in this specific task. Moreover, an analysis of state-of-the-art sequential methods compared to non-sequential methods for modelling historical information statically can shed light on the AR problem. Another original aspect of the present work with regard to the state of the art is that our evaluation is done under the assumption of on-line recognition, where future information is not available.

## 3. Method

Our approach for on-line activity recognition from audio and home automation sensors is detailed in this section. In Smart Homes, AR can be performed from a set of very heterogeneous raw data streams of various sensors, such as binary presence detectors (Presence Infra-Red sensors or PIR), continuous microphone signals or temperature measurement. To handle this heterogeneity, the overall strategy we adopted is to summarise data from these sensors within temporal sliding windows to generate vectors of attributes that will feed into an activity classifier. This approach relies on the hypothesis that each instance of any activity is composed of a set of events whose observations are captured by the set of sensors. These observations are signatures of the activities and they can be described by statistics of predefined variables computed over temporal windows shorter than the minimal activity duration. Although activities captured in this manner might be large scale activities, we showed that they can provide sufficient contextual information to an home automation decision module [13].

The method to recognise activities from streams of raw sensor data goes through different levels of abstraction, as depicted in Figure 1. The raw data are composed of symbolic timestamped values (from, e.g., infra-red sensors), state values (from e.g., switches), irregularly sampled signals (e.g., temperature) and equidistantly sampled signals (from, e.g., microphones). Some of these data are pre-processed to extract higher-level information such as speech, non-speech sounds and the location of the inhabitant. This step is detailed in Section 3.1. Then, all the raw and inferred information is summarised as vectors of features $V_n$, each of which corresponds to a temporal window $W_n$ of du-

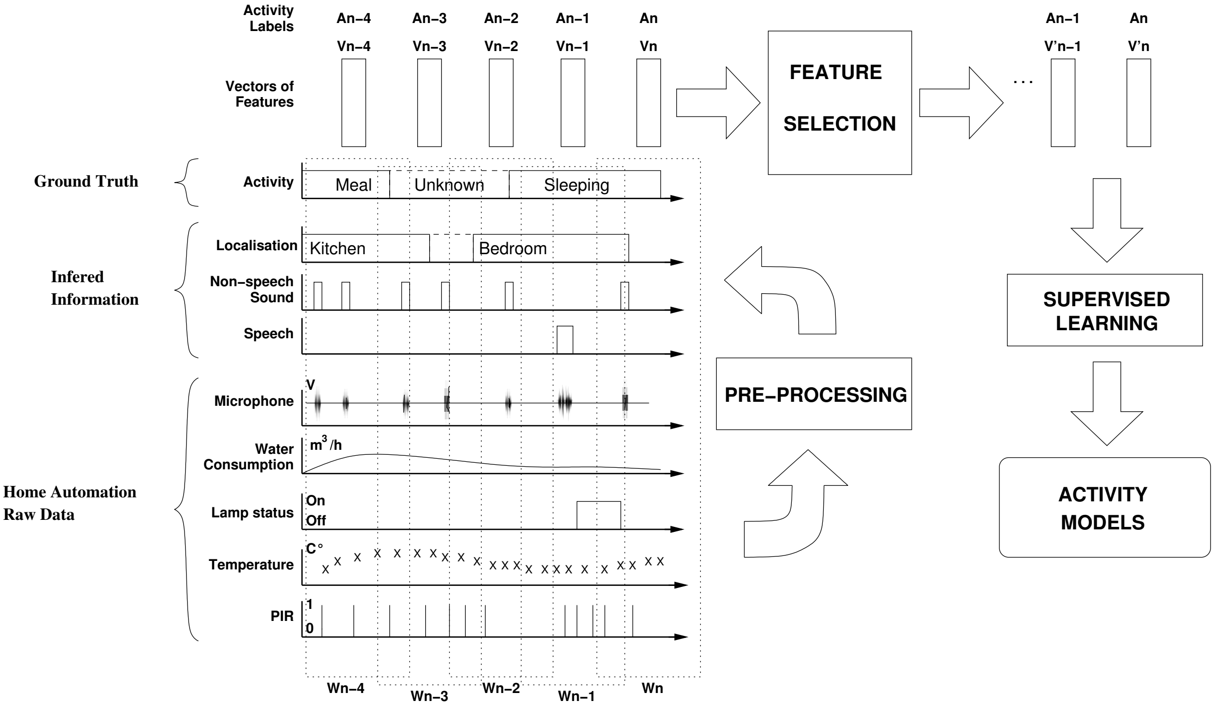

Fig. 1. Diagram of the overall methodology for activity model determination.

ration $T$. The feature vector comes together with an activity label $A_n$ that is generated from the ground truth by taking the activity having the longest duration within $W_n$ as the label. As the number of features can be very large, a feature selection step is performed to discard redundant and uninformative features, resulting in feature vectors $V'_n$. $V'_n$ together with $A_n$ are used as input to the activity model learning schemes. All of the classification models are trained using supervised machine learning techniques described in Sections 3.3 to 3.6.

This section summarises the pre-processing stage, and details the attributes and the classifier models.

### 3.1. Generating the Vectors of Attributes

The raw data captured within the Smart Home (see bottom of Fig. 1) are summarised by features computed over a temporal window. This section details the windowing strategy applied and the features computed.

#### 3.1.1. Windowing strategy

In this paper, the aim is to build classification models for on-line processing. In on-line processing, only current and past information is available. This means that, for each current time $t$, the temporal windows $W$ will cover the interval $]t - T, t]$. For the sake of clarity, we will call $W_1$ the temporal window representing the interval $]0, T]$, $W_2$ the temporal window representing the interval $]T, 2 * T]$, $W_n$ the temporal window representing the interval $](n - 1) * T, n * T]$, etc.

Given the dynamic nature of the activities, $T$ must be chosen to be shorter than the minimal duration of an activity instance, but should be long enough to benefit from the past history. A problem with fixed-size windows is that an activity can be under-represented due to the windows boundaries (e.g., an activity covered by half each of two temporal windows). To solve this problem, overlapping of intervals in 0% to 50% of $T$ may be used. In case of the intersection rate $\alpha$ between two consecutive windows then, $\forall n > 1,\ W_n$ covers the following interval:
$](n - 1) * (1 - \alpha) * T\ ,\ (n * (1 - \alpha) + \alpha) * T]$.

Finding the best values for $T$ and $\alpha$ is a tedious task that requires testing each value for classification method on datasets partitioned with different combinations of $T$ and $\alpha$. In a previous work [11], we tested values for $T$ of 60 and 120 seconds, with values for $\alpha$ of 0, 0.33 and 0.5 and we found that $T = 60$s and $\alpha = 0.5$ gave the best results. Based on these results, all the experiments reported in this study were conducted using $T = 60$s and $\alpha = 0.5$.

#### 3.1.2. Localisation and Speech/Non-Speech sound detection

The raw data contains information that must be extracted to enhance activity recognition. Two types of information are considered: speech/non-speech sound event — which are important for activities of communication — and localisation of the inhabitant — which is of primary importance for activity recognition.

*Speech/non-speech sound detection* In this approach, sound events are detected in real-time by the AUDITHIS system [78]. Briefly, the audio events are detected in real-time by an adaptive threshold algorithm based on a wavelet analysis and an SNR (Signal-to-Noise Ratio) estimation. The events are then classified into speech or everyday life sound by a Gaussian Mixture Model. The microphones being omnidirectional, a sound can be recorded simultaneously by multiple microphones; AUDITHIS identifies these simultaneous events. For more details about the audio processing, the reader is referred to [78].

*Localisation* In Smart Homes, localisation can be performed using cheap infra-red sensors detecting human movements but these sensors can lack sensitivity. To improve this, our approach fuses information coming from different data sources, namely infra-red sensors, door contacts and microphones. The data fusion model is composed of a two-level dynamic network [12] whose nodes represent the different location hypotheses and whose edges represent the strength (i.e., certainty) of the relation between nodes. This method has demonstrated a correct localisation rate between 63% and 84% using uncertain data from several sensors.

#### 3.1.3. Computed features

The traces generated from human activities are difficult to generalise, even in a given setting, due to the high inter and intra-person variability of realisations of a same task. This is why statistical attributes and inferred information were chosen to summarise the content of each window.

For all the binary sensors (e.g., infra-red motion detectors, switches), the number of firings in a time frame was computed. For all the contact-door sensors (e.g., doors, windows, furniture, curtains), the number of state changes was computed for each temporal window. For all events for which the duration is important (e.g., speech occurrences), the number of detections and their duration as a percentage of the temporal window were computed. For all signals ($CO_2$ level, temperature, humidity, brightness, water or electricity), the difference of mean value between time frames was computed. Regarding location, the percentage of time of occupation of each room was computed for each time window. Moreover, to add past information, the previous main occupied room is added as a feature. Finally, to account for the level of "activeness"[2] of the person within the home, the number of events per temporal window for each of the categories: room, doors, electricity, water, non-speech and speech sounds were summed up and divided by the frame duration.

Most of the activities under consideration in this study have a sequential pattern (e.g., sleeping implies going to the bedroom, then to lie down on the bed and to make no or infrequent large movements, dressing implies to get clothes and to make movements to put them on, etc.). However, in this windowing approach most of the temporal information within the temporal window is lost. But, given the high variability in the sequence of events for a simple activity even by the same person, we claim that such abstraction is a way to eliminate intra-class variations and noise in order to obtain a better generalization. Moreover, the duration of the windows being short, the hypothesis is that the sequential nature of the activities can be captured through sequence based models. Another advantage of these features is their very low computational cost.

### 3.2. Known and unknown activities

The activities under consideration in the study are inspired by the well known Activities of the Daily Living introduced by Katz [37] which are often used in geriatric assessments (dressing, feeding, toileting, etc.). The chosen activities, slightly different in the two experiments, are detailed in Sections 4.2.1 and 4.2.2. They were chosen mainly to provide contextual information for decision making (e.g., for a voice-based home automation system [13]) but also to provide relevant information about the behaviour of the user.

Another class was also considered in the study: the `Unknown` class (also called the `NULL` hypothesis) that represents periods during which it is not currently known which activity is being performed (or labelled in the case of the training dataset). Indeed, as reported in other studies [27,40], a large number of the activ-

[2] In this paper, we distinguish the activity –i.e., the task being performed– from the activeness –i.e., the state of being active. It is also called 'total agitation' in the paper from the French *agitation* equivalent to *bustle* in English.

ities recorded in the user's home are either transient activities not identifiable by the classifier (e.g., movement between rooms, wandering) or irrelevant activities. In our approach, we handle the `Unknown` class by considering it as a class label. Although acquiring a unique model of such a mixture of situations is of low interest for knowledge acquisition, its inclusion challenges the other learned classes and leads to more accurate learning of the "known" classes. It must be emphasized that, in further experiments, when the dataset is considered without the `Unknown` class, all the windows that are labelled as `Unknown` class are excluded from the training and testing set.

### *3.3. Activity Modelling by Hidden Markov Model (HMM)*

Hidden Markov Models (HMMs) [61] (Figure 2) are extensively used in activity recognition, for which it has become a "standard" approach [23,52,84]. One use of HMM in AR is to compute the most probable sequence of hidden states $Y = \{y_1, y_2, \ldots, y_n\}$ given a sequence of observations $X = \{x_1, x_2, \ldots, x_n\}$. In this paper we focus on ergodic HMMs, that is HMMs with fully connected hidden states (i.e., it is possible to reach any hidden state from any other hidden state).

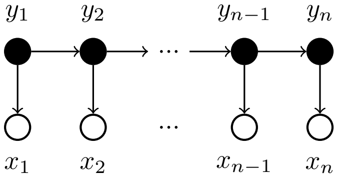

Fig. 2. Representation of a classical HMM for labelling elements in a sequence: $y_i$ are the hidden states, $x_i$ are the observations

Ergodic HMM models are based on two assumptions. The first one, which is true for any HMM, is the conditional independence of the observations. An observation $x_t$, emitted at time $t$, does not depend on any other observation, given the state that generated it. In many cases this assumption is false, but still works well in practice. For example, if one observation is the location of the subject, this variable is not independent between consecutive states. Indeed, depending on the organization of the habitation, when we have a location at a time $t$, the one at time $t + 1$ is in a restricted subset that depends on the first observation (the same location and the location that are just near this one). The second assumption, which is prevalent in first order HMM (such as the ergodic HMMs considered in this study) follows the Markov principle that the probability of the HMM being in state $i$ at time $t$ depends only on the state value at time $t - 1$. That means that the activity performed within a certain temporal window is independent of all the other previous temporal windows except the preceding one. This is considered to be an acceptable assumption in our AR application.

To model activities, a separate model was trained for each activity. Each hidden state was modelled by a Gaussian Mixture Model. The learning process was consequently carried out for each of the different activities (eating, dressing, etc.). This consisted in estimating the initial probabilities, the parameters of the GMMs (using the EM algorithm), and probabilities of the observations for each state and the state transition matrix. Convergence to the final parameters was obtained via the Baum-Welch algorithm. Finally, models with 2 hidden states and a GMM with 3 Gaussians for each state were obtained.

The AR was then performed by computing the log-likelihood of each of the $N$ activity models with an input vector. We consider that we are handling the data as a datastream, so we also do not try to determine the frontiers of each activity performance. Sequencing the datastream and adapt the model could be part of future works. During these tests, only the current and previous windows were considered. The HMM with the maximum likelihood was retained as the most probable class of the input sequence.

### *3.4. Activity Modelling by Conditional Random Fields (CRFs)*

Conditional Random Fields (CRFs) are graph based models to perform discriminative probabilistic inference over a set of variables in classification tasks [42].

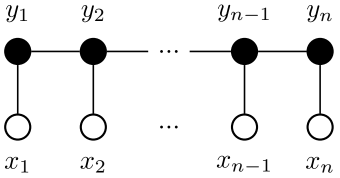

Fig. 3. Representation of a CRF for labelling elements in a sequence

Similarly to HMMs, CRFs can classify a sequence of variables $Y = \{y_1, \ldots, y_n\}$ for a given sequence of observations $X = \{x_1, \ldots, x_n\}$. However, CRFs are not generative models, so they are not intended to model the joint distribution $p(X, Y)$. CRFs are discriminative models instead, they model the conditional distribution $p(Y|X)$, but without the requirement to model the distribution of the variable $X$. Graphically, a CRF is represented by an undirected graph, as shown in figure 3. In the case of activity recognition, we can consider $X$, to be a set of vectors describing temporal windows, and hidden variables $Y$ whose inferred value corresponds to the most probable activity which generated the observations.

Lafferty *et al.* [42] defines CRF as follows:
Let $G = (V, E)$ be a graph such that $Y = (Y_\nu)_{\nu \in V}$, so that $Y$ is indexed by the vertices of $G$. Then $(X, Y)$ is a conditional random field if, when conditioned on $X$, the random variables $Y_\nu$ obey the Markov property with respect to the graph, that is:

$$p(Y_\nu \mid X, Y_\omega, \omega \neq \nu) = p(Y_\nu \mid X, Y_\omega, \omega \sim \nu)$$

where $w \sim v$ means that $w$ and $v$ are neighbours in $G$.

Therefore, the probability of a node is conditioned by its neighbours and by the set of observations. CRFs are generally implemented as log linear models by means of feature functions $f_k$ . In the case of linear conditional random fields, the simplified equation for estimating $p(Y \mid X)$ is the following:

$$p(Y \mid X) = \frac{1}{Z} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \right\} \quad (1)$$

where $f_k$ are feature functions defined on subsets of $Y$ and $X$, $Z$ is a normalization factor, and $\lambda_k$ is a parameter to assign a weight to the feature function $f_k$. These weights are estimated during the learning phase. The algorithm that was used is L-BFGS, a quasi-newton method that aproximates the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

When considering a model to label temporal windows as performed activities, having $y_t$ as the activity at time $t$ and only one evidential variable $x_t$ representing the location of the user at time $t$, an example of feature function can be:

$$f_k(y_t, y_{t-1}, x_t) = \begin{cases} 1 \text{ if } (y_t = \text{"sleep"}, \\ \qquad y_{t-1} = \text{"clean"}, \\ \qquad x_t = \text{"bedroom"}) \\ 0 \text{ Otherwise} \end{cases}$$

The feature functions $f_k$ take the value of 1 if their variables are set to the values specified in the function, 0 otherwise.

In our implementation of CRF for activity recognition the evidential variable $x_t$ is not a single value but a vector $V_t' = x_t = \{x_t^1, ..., x_t^m\}$ where each $x_t^i$ is the value of attribute $i$ at time $t$. Thus, $m$ feature functions $f_1(y_t, y_{t-1}, y_{t-2}, x_t^1, x_{t-1}^1, x_{t-2}^1)$, $\ldots$, $f_m(y_t, y_{t-1}, y_{t-2}, x_t^m, x_{t-1}^m, x_{t-2}^m))$, one for each attribute, were defined. These features function are dependend on the values of the current windows and the two previous ones.

### *3.5. Activity Modelling by dynamic Markov Logic Network (MLN)*

A Markov Logic Network (MLN) is a generative statistical relational model that combines First Order Logic (FOL) and probabilistic inference. A MLN model is expressive enough to include explicitly, by means of FOL, the main relations that exist among the elements of the smart environment involved in the activity recognition. In addition, every logical formula is given a numerical weight indicating a degree of truth. This logical representation along with its set of weights can be considered as a meta model that, during the inference process, allows the construction of a Markov network, a pure probabilistic model that can deal with uncertain variables. In this section, we introduce formally the MLN model and our implementations for activity recognition.

A MLN is composed of a set of FOL formulae, each one associated with a weight that expresses a degree of truth. This approach softens the assumption that a logic formula can only be true or false. A formula $f$ is grounded by substituting each variable in $f$ by a constant. A grounded formula that consists of a single predicate is a *ground atom*. A set of ground atoms is a *possible world*. All possible worlds in a MLN are true with a certain probability which depends on the number of formulae satisfied and the weights of these formulae. Let's consider $F$ a set of first-order logic formulae, with $w_i \in \mathbb{R}$ the weight of the formula $f_i \in F$, and $C$ a set of constants. During the inference process, each MLN predicate is grounded and the Markov network $M_{F,C}$ is constructed where each random variable is instanced with a ground atom. An edge is created for every pair of variables representing predicates that appear in the same formula. The obtained Markov network allows the estimation of the probability of a possible world $P(X = x)$ by the equation 2:

$$P(X=x) = \frac{1}{Z} exp\left(\sum_i w_i n_i(x)\right) \quad (2)$$

where $Z = \sum_{x' \in \chi} exp\left(\sum_i w_i n_i(x')\right)$ is a normalisation factor, $\chi$ the set of the possible worlds, and $n_i(x)$ is the number of true groundings of the i-th clause in the possible world $x$. When the number of predicates and the size of the domain of the variables grows, exact inference in MLN becomes intractable, so Markov Chain Monte Carlo methods are applied to approximate $P(X = x)$ [66]. In our case, recognizing an activity consists of finding the activity $a$ in $A$ that maximises $P(X = a, e)$, where $e$ is the evidence represented by the value of the attribute in each window (e.g., the values of the $V'$ vector). Learning a MLN consists of two independent tasks: weight learning and structure learning. Weight learning can be achieved by maximizing the likelihood with respect to a training set. If the $i^{th}$ formula is satisfied $n_i(x)$ times in $x$, then by using equation (2), the derivative of the log-likelihood with respect to the weight $w_i$ is given by equation (3):

$$\frac{\partial}{\partial w_i} \log P_w(X=x) = n_i(x) - \sum_{x' \in \chi} P_w(X=x') n_i(x) \quad (3)$$

where $x'$ is a possible world in $\chi$. The sum is thus performed over all the possible worlds $x'$ and $P_w(X = x')$ is $P(X = x')$ computed using the vector $w = (w_1, \ldots, w_i, \ldots)$. The maximisation of the likelihood is performed by an iterative process converging towards an optimal $w$. Unfortunately, doing this maximisation (3) is intractable in most cases. Thus, approximation methods are used in practice such as the *Scaled Conjugate Gradient* method [48].

The implementation proposed for Activity Recognition uses a set of rules which models the relationship between each feature and the activity independently from the other features. Formally, if $N$ discrete features are used for classification, the possible values for a feature $i$ is given by the set $Values_i = \{V_{i,1}, ..., V_{i,|Values_i|}\}$, and the activities considered are $Classes = \{A_1, ..., A_c\}$.

The rules used to classify activities have the following structure $feature_i(W, V_{i,j}) \Rightarrow class(W, A_k)$ where the variable $W$ is the temporal window to be classified. The following rules are examples of this pattern:

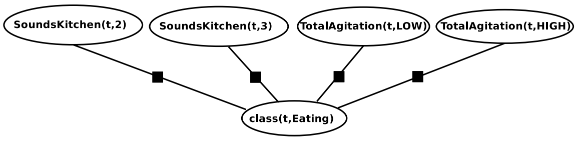

Fig. 4. Ground Naive MLN.

$SoundsKitchen(W_1, I_2) \Rightarrow class(W_1, Eating)$
$SoundsKitchen(W_1, I_3) \Rightarrow class(W_1, Eating)$
$TotalAgitation(W_1, LOW) \Rightarrow class(W_1, Eating)$
$TotalAgitation(W_1, HIGH) \Rightarrow class(W_1, Eating)$

The total number of rules in the model is given by $\sum_{i=1}^{N} |Values_i|.|Classes|$. As two grounded predicates $class(W_i, A)$ and $class(W_j, A)$, for $i \neq j$ , never appear in the same formula, in the resulting Markov network the probability of an activity being performed in a certain temporal window is independent from the other windows. This structure is then similar to a logistic regression model as shown in figure 4. This model is called the Naive MLN in reference to the Naive Bayes network.

In addition, we implemented a dynamic model that represents the activity recognition problem as a stochastic process in time. In this model, we use the same predicates, but the identifiers of the temporal windows become time arguments whose values are positive integers. We also introduce a temporal predicate $Previous$ that defines the order of the time arguments; for instance, $Previous(W_2, W_3)$, $Previous(W_3, W_4)$, $\neg Previous(W_5, W_4)$. The basic rules composing the Naive MLN remain the same with the addition of the following rules:

$$Previous(W_1, W_2) \wedge class(W_1, A_i) \Rightarrow class(W_2, A_j) \ \forall A_i, A_j \in Classes$$

The purpose of these rules is to establish a sequential relation between consecutive temporal windows. In the ground Markov network two predicates $class(W_i, A)$ and $class(W_j, A)$ can belong to the same clique, i.e. there is a probabilistic dependency among them. In the case of learning the weights of these formulae are learned from the ground truth, in the case of inference, the value of the previous class is provided by the preceding inference. This model is called the dynamic MLN or more simply **MLN**.

### 3.6. *Activity modelling by non-sequential methods: SVMs and Random Forests*

The two last methods considered in this study were Support Vector Machines (SVMs) and Random

Forests. They are classification algorithms that have been executed on each temporal window independently. These two algorithms have previously been used for activity recognition and have demonstrated good performances (e.g. [26,25,21]).

The processing behind a SVM is to project an input vector into a feature space using a kernel (we choose here a Gaussian kernel, for which we have first to determine the standard deviation $\sigma$ – the first parameter of the model). From the projected vectors, the learning algorithm determines the best possible separation hyperplane between the individuals of two classes, that is the hyperplane at the largest distance from all the points belonging to each class, called margin. A second parameter, C, controls the trade-off between the size of this margin and the number of possibly misclassified training samples. This algorithm, originally developed by Vapnik *et al.* [8] has demonstrated a very good efficiency on different kinds of classification tasks.

A Random Forest (RF) [10] is an ensemble classifier composed of several decision trees. For a new input, each decision tree decides a class and a voting strategy is used to determine, among the several trees, which class to attribute to the input vector. The induction of a RF combines random subspaces and bagging. It constructs a decision tree using a randomly selected reduced number of attributes (the number of trees created is a parameter of the algorithm).

For more details about these well known and documented models the reader is referred to the previously cited papers. The description of the determination of each of the parameters is provided in Section 4.3.2.

## 4. Experiments and Results

The methods were applied on data collected in two Smart Homes during two experiments involving respectively 21 persons and 15 persons. This section describes the Smart Homes (Sec. 4.1), the data sets (Sec. 4.2) that were acquired and the attribute selection and model parametrisation (Sec. 4.3). At the end, the results of the activity recognition are presented in Section 4.4.

### 4.1. Pervasive Environments

#### 4.1.1. The Domus Smart Home

The first pervasive environment considered is the Domus Smart Home that was designed by the *Laboratoire d'Informatique de Grenoble* (LIG) [28]. This flat was extensively used in the Sweet-Home project for experiments. Figure 5 shows the details of the flat. It is a thirty square meters flat including a bathroom, a kitchen, a bedroom and a study room, all equipped with sensors and actuators such as infra-red movement detectors, contact sensors, video cameras (used only for annotation purposes), etc. In addition, seven microphones were set in the ceiling for audio analysis. The flat is fully usable and can accommodate a dweller for several days. The technical architecture of Domus is based on the KNX bus system (`www.knx.org`), a worldwide ISO standard (ISO/IEC 14543) for home and building control. Besides KNX, several field buses coexist in Domus, such as UPnP (Universal Plug and Play) for the audio video distribution or X2D for the detection of the opening and closing of doors, windows, and cupboards. More than 150 sensors, actuators and information providers are managed in the flat (e.g., lighting, shutters, security systems, energy management, heating, etc.). Sounds are recorded independently to other sensors data thanks to a National Instrument mutichannel acquisition board and analyzed by the AuditHIS software [78].

The Domus flat was designed to be as normal as a standard flat, so that the participants moving in the smart home would behave as naturally as possible, performing activities in as close as possible to their usual manner.

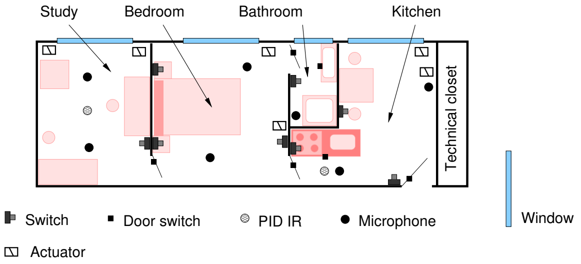

Fig. 5. The Domus Smart Home of the LIG

#### 4.1.2. The HIS Smart Home

The second Smart Home has been set up inside the Faculty of Medicine of Grenoble by the researchers of the TIMC-IMAG laboratory. This 47m$^2$-flat is composed of a bedroom, a living-room, a hall, a kitchen (with cupboards, fridge. . . ), a bathroom with a shower and a toilet. It was equipped with: 1) *infra-red presence sensors* (PIR), placed in each room to sense specific locations in the flat; 2) *door contacts* for the detection of the use of some of the commodities (fridge, cupboard and chest of drawers); 3) *microphones*, also in each room, to monitor, record and process all the sounds inside the flat and classify them into sounds of daily living or speech; and (4) large angle webcams (for annotation purposes only).

All the sensors, their location and also the organization of the flat are presented in Figure 6. The basis of the flat is the wireless presence infra-red (PIR) sensor, used in the AILISA project to monitor the level of activity of the person [44]. The other sensors (i.e. the microphones, webcams, environmental and contact sensors), that have been added to the initial AILISA platform, are optimally distributed to the four computers of the technical room (to optimize both resources and processing time use). This room, next to the Health Smart Home (HIS), contains these computers and electronic devices that receive and store, in real time, the information from the HIS. These computers are from standard ones. Apart from the microphones that need a National Instrument multichannel acquisition board for the analog to digital conversion of the signals from the microphones, the other connections are done with serial or USB ports.

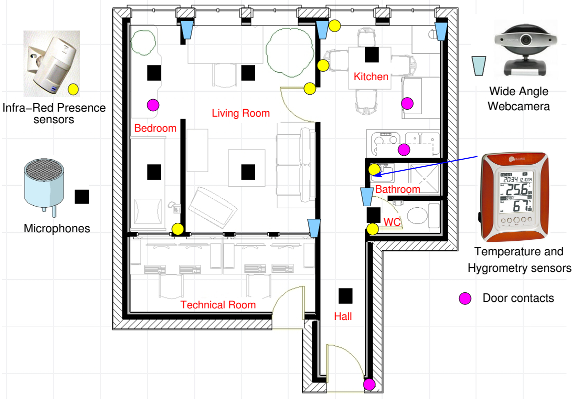

Fig. 6. Equipment and layout of the Health Smart Home (HIS) of the TIMC-IMAG Laboratory in Grenoble

## *4.2. Experimental Data*

This section details the acquisition and the characteristics of each of the two datasets and the procedures followed to record them before describing the manner they were used in this study. In all cases, cameras were used for video recording in each room for annotation purpose only, except in the toilet and bathroom in which there were no cameras in order to respect privacy.

### *4.2.1. The multimodal* SWEET-HOME *(SH) dataset*

The multimodal SWEET-HOME dataset is part of the SWEET-HOME corpus [80], in the following we will refer it by SH dataset. For the record, 21 persons (including 7 women) participated to the experiment to record all sensor data in a daily living context. The average age of the participants was 38.5 ± 13 years (min 22, max 63), height 1.72 ± 0.13m (min 1.43, max 2.10), and weight 70 ± 14kg (min 50, max 105). Participants were asked to enter the flat and behave as if they were in their own home. Before this, the experimenter organised a detailed visit of the flat for each participant to make sure that the individual would not search for everyday items and would feel like being at home. Then, the individual was asked to perform each of the previously defined activity at least once. No time constraint was given to perform these activities. The list of activities the person was asked to perform is the following: (1) to close the door with the electrical lock, (2) to undress/dress, (3) to wash hands, (4) to prepare a meal and to eat, (5) to wash dishes, (6) to brush his/her teeth, (7) to have a nap, (8) to rearrange the bed, (9) to do the housework, (10) to read a book and to listen to music, (11) to have a phone call, (12) to go out of the flat for shopping and to come back, (13) to use the PC and to call a relative, (14) to undress, take a shower and go to the bed for sleeping.

In total, thanks to the 21 participants, more than 26 hours of data have been acquired with an average scenario duration of one hour. About 18 hours were kept for an activity recognition experiment while the remaining time was retained for specific audio analysis. Data were annotated with the 7 classes of activity using the Advene software[3]. This corpus is freely available at `http://sweet-home-data.imag.fr/`. For our study, this corpus was divided into two independent parts: the tuning set and the train-test set. The tuning set was used to train the localisation algorithms as well as to learn a discretisation model to provide discrete values to the MLN classifier. This tuning set was excluded from the train-test set to avoid overfitting. It is important to note that the tuning set was also used to tune the parameters of the classifiers.

### *4.2.2. The HIS corpus (HIS)*

The HIS corpus [27] was acquired to monitor the activity of a person living alone at home, with the aim of helping geriatricians to evaluate the dependency level of various elderly people [26]. Seven activities were

[3] `liris.cnrs.fr/advene`

selected to be classified automatically: Preparing and having a meal, Performing hygiene activities, Dressing and undressing, Sleeping or having a nap, Resting and Communicating with relatives on the phone, and finally the Elimination activity (the fact to be in the toilet and using it). Each person was asked to perform the different activities for as long as they wanted and as often as they wanted. They were only instructed of the different activities to perform but not of the order in which to do them or of the way to perform them. The activities were performed by 15 healthy and non-elderly subjects (six women and nine men).

In total, about 13 hours of data have been acquired in the HIS flat. The average age of participants was 31.9 ± 9.0 years (min 24, max 57), the average height 1.74 ± 0.11m (min 1.62, max 1.92), and the average weight 68.5 ± 9.11kg (min 50, max 81). The mean execution time of each experiment was 51min 40s for a single participant (min 23min 11s, max 1h 35min 44s). This corpus is freely available at `http://getalp.imag.fr/HISData`. Data were annotated using the same conventions as for the SH corpus. For our study, in the same way as for the previous corpus, it was divided into two independent parts: the tuning set and train-test set.

#### 4.2.3. *Implementation for evaluation purpose*

From the raw data recorded by the different sensors of the flat, a vector $V$ of attributes was extracted for each temporal window as described in section 3.1. The duration $T$ of the window $W$ was set to 60s. The attributes were inferred or computed from the different signals. This resulted in an attribute vector $V$ of 94 elements in the case of the SH dataset and of 26 elements in the case of the HIS corpus.

Table 1 and Table 2 present the distribution of the activity classes for 60 second frames for the SH corpus and the HIS corpus. The first column represents the activity classes, the second column shows the percentage of data that has been put apart for preprocessing and tuning, the third column is the percentage of data used for training and testing the activity classification models and the last column presents the total. For SH, the preprocessing and tuning set was composed of the data from participants 8,10,11,13,14,15 while for HIS it consisted of data from the last four participants.

The distribution of classes is unbalanced due to the natural differences in the duration of each daily activity and to the fact that the scenarios were different in the HIS and SH cases. For each experiment, participants were recruited to play a scenario in one of the two smart homes but the scenarios were different. The following section details attribute selection and model parametrisation.

Table 1
Distribution of time windows for each activity in each SH dataset part ($T = 60s$).

| Class | Tuning set | Train-Test set | Both sets |
|---|---|---|---|
| Cleaning | 19.4% | 20.6% | 20.2% |
| Dressing /Undressing | 2% | 2.6% | 2.4% |
| Eating | 31.2% | 27.2% | 28.3% |
| Hygiene | 6.5% | 6.8% | 6.7% |
| Phone | 2.3% | 2.9% | 2.7% |
| Reading /Computer /Radio | 23.5% | 22.9% | 23.1% |
| Sleeping | 10.8% | 12.8% | 12.3% |
| `Unknown` | 4.3% | 4.2% | 4.2% |
| Number of participants | 7 | 14 | 21 |
| Number of windows | 603 | 1526 | 2129 |
| Duration | 5h01m30s | 12h43m00s | 17h44m30s |

### 4.3. *Attribute selection and model parametrisation*

This section details the pre-processing that has been applied to reduce the set of attributes and to tune the classification algorithms. The intervals that were not identified as one of the 7 specified activities were considered as belonging to the `Unknown` class.

#### 4.3.1. *Attribute selection*

Performing attribute selection is a necessary step in data mining both to reduce the size of the data and to improve performance [50]. Moreover, for some of our algorithms, the number of features is important and its reduction is crucial for two reasons: (1) the speed of both training and testing can grow exponentially with the number of features and (2) the curse of high dimensionality makes difficult to interpret differences in distances in high dimensional spaces.

Information Gain Ratio (IGR) has been chosen for feature selection because it usually performs well in

Table 2
Distribution of time windows for each activity in each HIS dataset part ($T = 60s$).

| Class | Tuning set | Train-Test set | Both sets |
|---|---|---|---|
| Dressing /Undressing | 3.2% | 2.1% | 2.4% |
| Eating | 19.4% | 15.7% | 16.6% |
| Elimination | 4.5% | 4.6% | 4.6% |
| Hygiene | 3.5% | 5.2% | 4.8% |
| Phone | 7.1% | 6.1% | 6.2% |
| Reading /Computer /Radio | 22.1% | 26.5% | 25.5% |
| Sleeping | 22.9% | 21.2% | 21.6% |
| Unknown | 17.3% | 18.6% | 18.3% |
| $N^o$ Participants | 4 | 11 | 15 |
| $N^o$ Windows | 376 | 1221 | 1597 |
| Duration | 3h08m00s | 10h10m30s | 13h18m30s |

practice [50] and because it is independent of the classification model (by contrast with wrapping attribute selection methods [68]). IGR is the basis criterion of some decision tree algorithms (e.g., C4.5 [60]) which progress by selecting the best attributes at each decision step from the remaining set of attributes. Recall that information gain is defined considering the entropy and the probability of each values for this attribute currently under consideration. The entropy $H(V)$ of a variable $V$ taking the values $v$ is defined by:

$$H(V) = -\sum_{v} p(v) \cdot log_2(p(v))$$

while the entropy of a value $V$, given a variable $X$ (with its possible values $x$) is defined by:

$$H(V|X) = -\sum_{x} p(x) \cdot \sum_{v} p(v|x) \cdot log_2(p(v|x))$$

The IGR for an attribute $a \in A$, considering the class $c \in C$, is then obtained as:

$$IGR(a, c) = \frac{H(c) - H(c|a)}{H(a)} \tag{4}$$

The formula (4) is applied to each attribute to obtain the score, then a threshold can be chosen to retain the $k$ best attributes.

The computation of the IGR is done on the complete dataset across classes. This gain is determined for each attribute and each class and for each attribute a weighted mean is computed to obtain its final value.

At the end, only those features with non-zero IGR (features including some information) were retained.

*SH dataset* The feature selection method was applied to the multimodal SWEET-HOME corpus (SH dataset). The attribute vector $V$ originally composed of 94 features was reduced to $V'$ of 66 features by using IGR.

Table 3 shows the 66 obtained attributes. In that case, only the attributes that have a non-null information gain were kept. In this table, the 20 attributes having the highest IGR scores are highlighted. It suggests that among the selected attributes those that provide the best information to classify activities are the attributes related to the location of the inhabitant and the acoustic features.

*HIS dataset* Following the same method as for the SH dataset, the HIS corpus, with data vectors originally composed of 27 features (26 plus the class) was reduced to 24 attributes with IGR. For this dataset, the number of features originally available was really small. That explains why only a few attributes were eliminated by the attribute selection process.

Table 4 sums up the reduced dataset.

### *4.3.2. Model tuning*

*HMM, SVM and Random Forest tuning* A 10-fold cross-validation on each tuning set was performed to optimize several parameters of the classifiers. For the Random Forest, the number of trees has been optimized, for the SVM, the pair $(C, \sigma)$ has been searched using a grid search and finally for the HMM, the number of states and the number of Gaussians has been determined. For this last one, the optimization has been done on the whole set of activities. This optimization is not done for each individual activity separately (for which an optimal topology of the HMM could perhaps improve the results). The parameter search was performed for each dataset (HIS and Sweet-Home) and the optimal values found for the parameters were kept for the classification.

Table 3

Attributes selected for the SH dataset using Information Gain Ratio for each attribute (66 attributes out of 94). The best 20 attributes are highlighted.

| Type | Attributes Names for GainRatio |
|---|---|
| Location | **PercentageLocationRoom1**, **PercentageLocationRoom2**, **PercentageLocationRoom3**, **PercentageLocationRoom4**, **PredominantRoom**, **LastRoomBeforeWindow**, NumberOfDetectionPIROffice, NumberOfDetectionPIRKitchen, **TimeSinceInThisRoom**, **PercentageAgitationRooms** |
| Switches | SwitchBathroomUse, SwitchBedroomBed, SwitchBedroom, SwitchOffice, SwitchSinkKitchen |
| Lights | **PercentageTimeLightBathroomOn**, ActivationDeactivationLightBathroomSink, ActivationDeactivationLightBedLeft, ActivationDeactivationLightBedRight, ActivationDeactivationLightKitchenSink, PercentageLightOfficeOn, PercentageLightKitchenSinkOn, PercentageLightKitchenTableOn, PercentageLightBedLeftOn, PercentageLightBedRightOn |
| Shutter | **PercentageShutterBedroom**, **PercentageShutterBedroom2**, PercentageShutterDesk2, PercentageShutterKitchen, ActivationDeactivationShutterBedroom, ActivationDeactivationShutterDesk, **PercentageCurtain**, PercentageShutterOffice |
| Power | PowerLastUse, PowerLastLastUse, PowerLastLastLastUse |
| Doors and Windows | ActivationDeactivationNumberOfDoorBedroom, ActivationDeactivationNumberOfDoorBathroom, ActivationDeactivationDoorCupboardKitchen, ActivationDeactivationDoorFridge, ActivationDeactivationNumberOfWindowBedroomBathroom, PercentageAgitationDoors |
| Sounds | **SoundsKitchen**, **SoundsDinningRoom**, SoundsBathroom, SoundsOfficeDoor, SoundsBedroomWindow, SoundsOfficeWindow, SoundsBedroomDoor, SpeechBedroomDoor, SpeechBedroomWindow, SpeechBathroom, **SpeechKitchen**, SpeechOfficeDoor, SpeechOfficeWindow, SpeechDinningRoom, **PercentageAgitationSounds**, **PercentageAgitationSpeech**, **PercentageTimeSound**, **PercentageTimeSpeech** |
| Divers | ColdWaterTotal, HotWaterTotal, **TotalAgitation**, AmbientSensorCO2Bedroom, AmbientSensorTemperatureOffice, AmbientSensorTemperatureBedroom |
| Class | One of: cleaning, dressing up, eating, hygiene, phone, sleeping, reading/computer/radio, unknown activity/transition |

Table 4

Attributes selected for HIS using retained non-zero Information Gain Ratio for each attribute (24 attributes out of 26).

| Type | Attributes Names for GainRatio |
|---|---|
| Location | PredominantRoom, LastRoomBeforeWindow, PercentageLocation (in every room), TimeSinceInThisRoom |
| Doors | ActivationDeactivationCupboardDoor, ActivationDeactivationDressingDoor |
| Sounds | Sound on all the microphones |
| Speech | Speech in Entrance, Hall, Shower, WC, Kitchen |
| Class | One of: dressing up, eating, elimination, hygiene, phone, sleeping, reading/computer/radio, unknown activity/transition |

*CRF and MLN tuning* The feature functions designed for the CRF model consider the evidential information of the current temporal window and also the two previous ones. We found that using the two previous windows instead of only one, slightly improves the accuracy of the algorithm while keeping an acceptable processing time.

In the cases of the MLN and the CRF, all the continuous numerical variables were discretised. A supervised method for discretisation, CAIM (Class-Attribute Interdependence Maximization) [41], has been run on the tuning set. It resulted in a set of discretisation intervals for each continuous attribute that were applied as a preprocessing stage to the input data of the CRF and the MLN. This algorithm works individually on each feature without the need to fix the number of discrete intervals as parameter. CAIM's optimization goal is to maximize the class-attribute interdependence while minimizing the number of intervals. The number of intervals found in the datasets was always between 3 and 8. Once again, only the tuning set was used to avoid overfitting.

### *4.4. Results*

#### *4.4.1. Performance evaluation*

The method used to evaluate the classifier was based on Cross-Validation but used a specific type namely Leave-One-Subject-Out-Cross-Validation (LOSOCV). If the dataset is composed of records[4] from $N$ participants, for each fold, records from $N-1$ participants were used to train the model, while the remaining record was used for evaluating the learned model. Consequently, testing was performed on different individuals from training, and thus LOSOCV prevents participant overfitting.

Performance was assessed using the accuracy measure over the full dataset, defined as:

$$Acc_{Global} = \frac{\sum_i V_i}{\sum_i S_i}$$

where $V_i$ is the number of windows of class $i$ correctly classified as $i$ and $S_i$ is the total number of windows of

[4] Here 'record' means the full record for a single participant

Table 5

Overall accuracy (%) results on the two datasets with and without the `Unknown` class

| | SH dataset | | | HIS corpus | | |
|---|---|---|---|---|---|---|
| Model | Without `Unknown` | With `Unknown` | diff | Without `Unknown` | With `Unknown` | diff |
| SVM | 75.00 | 71.90 | 3.10 | 74.86 | 64.90 | 9.96 |
| Random Forest | 82.96 | 80.14 | 2.82 | 70.72 | 62.32 | 8.40 |
| MLN naive | 79.20 | 76.73 | 2.47 | 75.45 | 66.81 | 8.64 |
| HMM | 74.76 | 72.45 | 2.31 | 77.26 | 67.11 | 10.15 |
| CRF | 85.43 | 83.57 | 1.86 | 75.85 | 69.29 | 6.56 |
| MLN | 82.22 | 78.11 | 4.11 | 75.95 | 65.82 | 10.13 |

class $i$. The average accuracy per class was also computed to assess the capacity of the learning method to model each class independently. This was defined as

$$Acc_{Class} = \frac{\sum_i Acc_i}{N_c}$$

where $N_c$ is the total number of classes and $Acc_i = \frac{V_i}{S_i}$, i.e., the accuracy $A_i$ for the $i^{th}$ class.

In all the results presented in the tables 6–9, the overall accuracy is given as well as the mean accuracy and standard deviation, computed over the participants, in brackets.

### 4.4.2. Preprocessing performance

As presented in section 3.1.2, two kinds of information were inferred from the raw data: location of the dweller and speech/non-speech sound events.

We adapted a dynamic network for multisource fusion with the aim of locating a participant in the smart home [12]. This process contains two levels: the first corresponds to generating location hypotheses from an event; and the second represents the context for which the activation indicates the most probable location given the previous events. Training was achieved separately on the two tuning sets, SH and HIS datasets (cf. Section 4.2) and gave 84% correct location for each 1 second windows of the Train-Test set of SH and 96% correct with HIS dataset. Thus, though the accuracy is acceptable for SH and excellent for HIS, the activity models are trained on imperfect data that may impact on the learning.

As far as sound processing is concerned, the discrimination module was a Gaussian Mixture Model (GMM) which classified each audio event as either an everyday life sound or a speech sound. The discrimination module was trained with an everyday life sound corpus [36] and with the Normal/Distress speech corpus recorded in our laboratory [79]. Acoustic features were Linear-Frequency Cepstral Coefficients (LFCC) with 16 filter banks and the classifier was made of 24 Gaussian models. Acoustic features were computed for every frame using a size of 16 ms, with an overlap of 50%. On the HIS Train-Test set, the global accuracy of the speech discrimination was 84.61%. 25% of the sounds classified as speech were actually "non-speech sounds" and 13% of the sounds classified as non-speech were actually "speech-sounds". So the classifier is again imperfect regarding speech/non-speech sound related features.

### 4.4.3. Global results

Table 5 shows the overall accuracy results for all the classification models and datasets both with and without including the `Unknown` class. Let's recall that the case without the `Unknown` class means that these windows were excluded from the datasets both for the learning and testing stages.

It can be observed that the CRF approach has the highest accuracy in 3 out of 4 conditions but the HMM approach shows the best accuracy for the HIS without including the `Unknown` class. MLN is always the second or third ranked method. The worst classifiers are the SVM under all conditions and the HMM on the SH dataset and the Random Forest for the HIS dataset (even if this was amongst the best for the SH dataset). For the SH dataset without `Unknown` class condition, a Kruskal-Wallis test revealed a significant effect for dependency of accuracy on the model ($\chi^2(5) = 16.22, p = 0.006$). A post-hoc test using pairwise Wilcoxon summed rank tests with Bonferroni correction showed that this dependency is mostly driven by the difference between the CRF and the HMM ($p = 0.032$). When the `Unknown` class is included, the significance increases ($\chi^2(5) = 17.78, p = 0.003$), still driven by the difference between the CRF and the HMM ($p = 0.028$) with the difference between the CRF and the SVM just short of significance

($p$ = 0.082). None of the HIS results show a significant difference, probably due to the high variability between subjects. When analysing the difference between the conditions both without and including the Unknown class, it can be seen that the CRF has the smallest decrease of performance resulting from including the Unknown class for both datasets, while MLN, HMM and SVM show the biggest decreases. The importance of the different decreases between the two datasets can be explained by the high proportion of Unknown class windows in the HIS dataset (more than 18% of the total dataset) compared with the SH set (about 4%). Overall, CRF seems to be the method with the best performance in most of the conditions. In the remaining sections we focus on the CRF and other dynamic models (HMM, MLN) to study their behaviour in each condition.

#### *4.4.4. Results on the SH dataset*

Detailed results per class both without and with the Unknown class are given in Table 6 and Table 7. Without the Unknown class, CRF has the overall best accuracy (85.43%) and averaged over classes (76.26%), closely followed by the MLN performance (82.22% globally and 75.41% per class) and both greatly outperform HMM (74.8% globally and 63.25% per class). CRF shows the best accuracy for most classes (Cleaning, Dressing, Eating, Hygiene, Sleeping) while the MLN has particularly good results for Phone and Reading. Clear superiority of the CRF method is exhibited for Dressing (56.11±31.88%) and Cleaning (84.25±12.93%) while the MLN shows significant superiority in the Phone class (79.76±23.6%). HMM shows good results on Hygiene and on Reading classes but is very poor on Dressing.

When the Unknown class is considered, the pattern remains the same. All the accuracy measures decrease except for both MLN and HMM in the case of the Cleaning class, where the results were slightly improved and the MLN case outperformed the CRF results for that one. The MLN shows again a significant superiority in the classification for the Phone class (80.1±23.56%) over both the HMM (49.8 ±35.8%) and the MLN (51.63±38.74%). In all other cases, CRF demonstrates the best accuracy.

#### *4.4.5. Results on HIS dataset*

Detailed results per class both without and with the Unknown class being included are given in Table 8 and 9. Without the Unknown class, the HMM has the best accuracy overall (77.3%) and averaged over class (71.0%), being slightly better than the MLN performance (75.95% globally and 68.99% per class) and that of the CRF (75.85% globally and 66.71% per class). The statistical tests did not reveal any significant difference between the models. Moreover, the highest performances for each class are well distributed over the methods, with HMM the best for Dressing, Sleeping and Elimination, with the CRF best for Eating and Reading and the MLN the best for Hygiene and Phone.

When the Unknown class is considered, the pattern changes slightly. CRF gives the highest accuracy globally (69.29%) but not per class (59.07% vs 60.2% for HMM). The best performance per class remains the same with HMM, being still the best for Dressing (equals with MLN), Elimination and Sleeping, CRF for Eating, Phone, Reading and Unknown and the MLN best for Hygiene (equal with HMM). Thus, the overall improvement of CRF is mostly driven by its good classification of the Unknown class, which represents 18% of the HIS dataset. Again the HMM exhibits clearly the best performance for Elimination compared with CRF and the MLN.

#### *4.4.6. CRF performance in discriminating activities*

Some classes were more difficult to discriminate between than others, Tables 10 and 11 [5] present the confusion matrices for the CRF for the SWEET-HOME and HIS datasets in the case where the Unknown class is included. Without surprise, in both corpus, the Unknown class is very uniformly confused with other classes, with stronger consequences for the HIS corpus since instances of the Unknown class constitute a big part of the dataset. For the SWEET-HOME dataset, Eating and Cleaning are confused with each other. It should be noted that these two activities were often performed in the same room. The Reading/Computer class exhibits a low specificity, with a lot of confusion with Phone, Dressing and Sleeping. This is not surprising since the Reading/Computer class is composed of different sub-classes which share common characteristics with classes which it gets confused with. For the HIS corpus, Elimination and Hygiene are confused with each other. Again, it should be noted that these two actives were performed in the same area. For the Reading/Computer class a similar trend as for SWEET-HOME is observed. This class shares many properties

[5] In these tables are also given sensitivity and specificity. As a reminder, let's consider **TP** the True Positive rate, **TN** the True Negative rate, **FP** the False Positive rate and **FN** the False Negative rate. Sensitivity $= \frac{TP}{TP+FN}$ and Sensitivity $= \frac{TN}{TN+FP}$

Table 6
Classification accuracy using the SH dataset without `Unknown` class: overall (per participant record $\pm SD$).

| Class | HMM | CRF | MLN |
|---|---|---|---|
| Cleaning | 64.8% (66.1 ±18.9%) | 82.80% (84.25±12.93%) | 75.16%(76.71±12.39%) |
| Dressing/Undressing | 2.7% (7.1 ±26.7%) | 53.84% (56.11±31.88%) | 30.77%(28.33±32.4%) |
| Eating | 76.9% (76.1 ±28.5%) | 85.43% (87.76±16.49%) | 83.37%(82.67±18.52%) |
| Hygiene | 79.1% (77.6 ±25.7%) | 79.80% (79.78±23.84%) | 78.85%(76.44±26.4%) |
| Phone | 54.8% (55.5 ±33.8%) | 50% (51.97±37.72%) | 81.82%(79.76±23.6%) |
| Reading/Computer/Radio | 92.1% (90.2 ±13.1%) | 91.14 % (91.08±8.67%) | 91.71%(92.76±7.15%) |
| Sleeping | 72.4% (72.1 ±25.8%) | 90.81% (88.81±13.29%) | 86.22%(87.29±12.62%) |
| Global | 74.8% | 85.43 | 82.22 % |
| Class | 63.25% | 76.26% | 75.41% |

Table 7
Classification accuracy using the SH dataset: overall (per participant record $\pm SD$).

| Class | HMM | CRF | MLN |
|---|---|---|---|
| Cleaning | 70.8% (72.4 ±13.9 %) | 79.82% (82.65±14.22%) | 81.85%(84.08±14.84%) |
| Dressing/Undressing | 2.7% (7.1 ±26.7%) | 50%(52.78±26.66%) | 25.64%(23.89±28.67%) |
| Eating | 77.4% (76.7 ±23.9%) | 83.57%(87.30±15.86%) | 77.35%(79.13±22.65%) |
| Hygiene | 68.1% (65.8 ±25%) | 87.88%(77.32±25.29%) | 78.85%(77.4±21.61%) |
| Phone | 50% (49.8 ±35.8%) | 54.34% (51.63±38.74%) | 79.55%(80.1±23.56%) |
| Reading/Computer/Radio | 92.1% (86.6 ±19.8%) | 91.66%(90.88±7.86%) | 87.71%(90.16±10.34%) |
| Sleeping | 66.7% (67.2 ±26.6%) | 89.62%(87.56±13.87%) | 84.69%(85.25±14.4%) |
| Unknown | 24.6% (23.2 ±18.8%) | 59.42%(50.37±36.85%) | 21.88%(18.96±23.36%) |
| Global | 72.45% | 83.57 | 78.11% |
| Class | 56.55 % | 74.54 % | 67.19% |

Table 8
Classification accuracy using the HIS dataset without `Unknown` class: overall (per participant record $\pm SD$).

| Class | HMM | CRF | MLN |
|---|---|---|---|
| Dressing/Undressing | 46.2 % (30.8 ±40%) | 38.46%(23.73±39.97%) | 30.77%(26.98±39.55%) |
| Eating | 90.6 % (90.2 ±17.7%) | 95.31%(94.81±9.5%) | 93.23%(93.43±12.86%) |
| Elimination | 85.7 % (65.5 ±46.7%) | 64.29%(46.61±43.35%) | 48.21%(33.72±44.45%) |
| Hygiene | 36.5 % (35.6 ±45.1%) | 36.51%(41.44±36.28%) | 73.02%(52.27±50.56%) |
| Phone | 83.8 % (76.6 ±36.3%) | 81.08%(72.09±34.8%) | 91.89%(83.93±31.65%) |
| Reading/Computer/Radio | 75.9 % (73.8 ±38%) | 77.16%(75.4±35.64%) | 75.93%(68.96±39.53%) |
| Sleeping | 78.4 % (61.4 ±41.5%) | 74.13%(64.77±32.35%) | 69.88%(65.13±37.93%) |
| Global | 77.3% | 75.85% | 75.95% |
| Class | 71.0 % | 66.71 % | 68.99 % |

Table 9
Classification accuracy using the HIS dataset: overall (per participant record $\pm SD$).

| Class | HMM | CRF | MLN |
|---|---|---|---|
| Dressing/Undressing | 26.9 % (10.1 ±18.8%) | 15.38%(4.25±9.74%) | 26.92%(18.36±32.73%) |
| Eating | 88 % (87.6 ±20%) | 89.58%(87.37±20.25%) | 85.94%(84.52±19.05%) |
| Elimination | 76.8 % (61 ±46.9%) | 62.5%(42.07±38.89%) | 57.14%(47.69±45.89%) |
| Hygiene | 25.4 % (32.3 ±45.9%) | 20.63%(26.36±40.56%) | 34.92%(34.55±48.24%) |
| Phone | 77 % (69.8 ±32.4%) | 78.38%(69.68±35.54%) | 60.81%(55.65±36.62%) |
| Reading/Computer/Radio | 74.7 % (72.6 ±37.4%) | 75.93%(73.04±38.06%) | 69.44%(61.46±44.44%) |
| Sleeping | 79.9 % (68.1 ±37.7%) | 70.66%(65.44±36.59%) | 71.81%(64.2±39.11%) |
| Unknown | 32.9 % (33.9 ±7.8%) | 59.47%(60.33±14.72%) | 52.86%(53.43±14%) |
| Global | 67.1% | 69.29% | 65.82% |
| Class | 60.2 % | 59.07% | 57.48% |

with other classes, notably sleeping. Indeed, Reading consisted if sitting on the sofa and reading a magazine while the Sleeping activity was to lie on the bed doing nothing. These two activities were thus very quiet, evolved very little motion and generated a very low amount of information. Moreover they occurred in the same area (only an open partition separating the bed and the sofa). It is also worth noticing the very low sensitivity for Dressing. This activity was very short and performed between the living room and the bed area. The small amount of examples of this class explains the low performance in learning it.

### *4.4.7. Subsequent analyses*

To assess the impact of including the `Unknown` class on the learning, the training was performed on the HIS dataset with the best and worst classifiers, CRF and Random Forest, whilst varying the percentage of examples of the `Unknown` class in the dataset. Figure 7 shows a rapid decrease of Random Forest performance up to 8% of the total dataset being `Unknown` when it reaches a plateau, while CRF shows a sharp decrease in performance when the `Unknown` class is being introduced (even by 1% of `Unknown` examples among all classes), but then the performance decreases very slowly until 10% of exemaples being `Unknown` when it reaches a plateau. Thus, although such behaviour calls for further investigation, it seems that the CRF is more robust to the inclusion of the perturbing `Unknown` class than the Random Forest approach. This is in line with some studies reporting a decrease of performance when a RF is learned from noisy datasets [69].

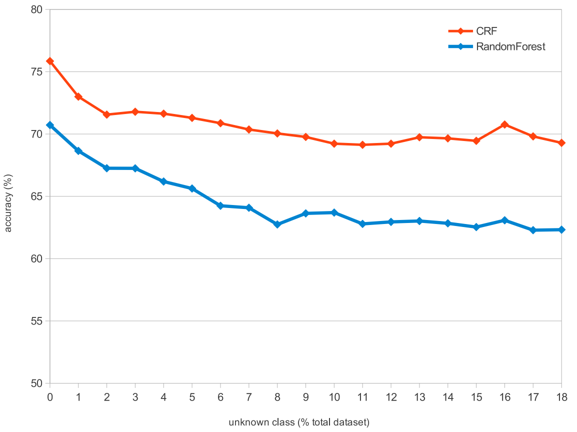

Fig. 7. Accuracy against percentage of windows belonging to the `Unknown` class for Random Forest and CRF applied to the HIS database.

## 5. Discussion

Automatic recognition of human activities in smart spaces is an important challenge for Ambient Assisted Living (AAL). In real-world applications, this task would often have to be performed on-line using data from cheap, distant and noisy home automation sensors. In this paper, we present a study to recognize on-line the activities of one dweller from distant (i.e., not worn on the user's body) home automation sensors (not including any video camera) and microphones using 6 different models: a SVM, Random Forest, dynamic/non-dynamic Markov Logic Networks, a M and CRF. This study, performed on 2 realistic and publicly available datasets, sheds light on the limita-

Table 10

Confusion Matrix for CRF – SH dataset

| | Cleaning | Dressing<br>Undressing | Eating | Hygiene | Phone | Reading/<br>Computer | Sleeping | Unknown |
|---|---|---|---|---|---|---|---|---|
| Cleaning | 255 | 2 | 39 | 1 | 0 | 4 | 1 | 6 |
| Dressing/Undressing | 0 | 20 | 1 | 1 | 0 | 3 | 2 | 2 |
| Eating | 48 | 1 | 364 | 8 | 0 | 1 | 0 | 6 |
| Hygiene | 3 | 3 | 3 | 80 | 0 | 1 | 2 | 4 |
| Phone | 1 | 0 | 0 | 0 | 23 | 3 | 0 | 0 |
| Reading/Computer/Radio | 4 | 9 | 1 | 9 | 19 | 319 | 14 | 8 |
| Sleeping | 1 | 2 | 6 | 4 | 1 | 16 | 175 | 2 |
| Unknown | 2 | 2 | 1 | 1 | 1 | 3 | 2 | 36 |
| Sensitivity | 81.21% | 51.28% | 87.71% | 76.92% | 52.27% | 91.14% | 89.29% | 56.25% |
| Specificity | 95.63% | 99.40% | 94.24% | 98.88% | 99.73% | 94.56% | 97.59% | 99.18% |

Table 11

Confusion Matrix for CRF – HIS corpus

| | Dressing<br>Undressing | Eating | Elimination | Hygiene | Phone | Reading/<br>Computer | Sleeping | Unknown |
|---|---|---|---|---|---|---|---|---|
| Dressing/Undressing | 4 | 0 | 0 | 0 | 3 | 1 | 7 | 8 |
| Eating | 1 | 172 | 0 | 0 | 0 | 3 | 2 | 23 |
| Elimination | 0 | 0 | 35 | 37 | 0 | 0 | 0 | 7 |
| Hygiene | 0 | 0 | 18 | 13 | 0 | 1 | 1 | 7 |
| Phone | 0 | 0 | 0 | 0 | 58 | 1 | 0 | 7 |
| Reading/Computer/Radio | 5 | 1 | 0 | 4 | 4 | 246 | 48 | 23 |
| Sleeping | 7 | 2 | 0 | 0 | 0 | 59 | 183 | 17 |
| Unknown | 9 | 17 | 3 | 9 | 9 | 13 | 18 | 135 |
| Sensitivity | 15.38% | 89.58% | 62.50% | 20.64% | 78.38% | 75.93% | 70.66% | 59.47% |
| Specificity | 98.41% | 97.18% | 96.22% | 97.67% | 99.30% | 90.52% | 91.16% | 92.15% |

tions and advantages of these models for the activity recognition tasks which are discussed below.

To achieve the on-line real-time classification requirement of the study, the sequential models (HMM, CRF and MLN) were only learned and applied using the past history, meaning that no future data are known at the time of making the decision. Moreover, on-line and realistic activity recognition in the home must deal with `Unknown` activities [40] as well as taking into account the transitions between activities (i.e., segmentation of the activities). This is a radical difference from many off-line studies [6,54,26,85] that induce models using both the past and future history, sometimes apply preprocessing to the overall dataset and often use a cross-validation over all the activity windows without considering the detection problem. In our case, both preprocessing and classification is performed without the use of any information on future values.

The results of this study shows that the sequential models, as a group (HMM, MLN, CRF), do not significantly outperform the non-sequential models (SVM, Random Forest). This can be explained by two reasons. Firstly, AR is highly dependent on the location, the presence of this information alone in the temporal windows is important enough to allow accuracy classification in instance-based models. Secondly, the design of the features for classification in our method allows the inclusion of historical information in the temporal window. For instance, the time the person has spent in the same room is accumulative from one window to the next one, as long as the person does not change location. However, it must be emphasized that a sequential model is always ranked first in all conditions (CRF three times, HMM once). It can then be concluded that CRF is generally the best suited algorithm for on-line human activity recognition from simple non visual sensors, as it consistently outperforms its best non-sequential competitor, namely the Random Forest (RF). While a CRF has already been reported as outperforming a HMM in human activity classification tasks [81,82], the competition between CRF and RF has not been previously reported, mostly because these models do not belong to the same type.

The difference in performance between CRF and the HMM/MLN is due to their discriminative or generative natures. While a CRF is trained to maximize the likelihood over the whole dataset, the HMM/MLN are trained by maximizing the likelihood of each class independently. Thus, a CRF biases its learning towards the most dominant classes, as do the non-sequential discriminative schemes (SVM, Random Forest, Naïve MLN). On the contrary, the MLN and HMM model the classes independently, and this explains why they performed better on some activities. For instance, MLN had the best performance for recognizing the phone activity for 3 out of 4 conditions (cf. Tables 6–9), while the HMM showed the best general performance by class for the HIS dataset (cf. Table 8 and 9).

The inclusion of the `Unknown` class decreased all the performances of all the models. However, it is the generative approaches that are the most visibly affected. Indeed the MLN and HMM exhibit the most notable decrease in performance, and this is due to their inability to model the `Unknown` class. Since, in our case, the `Unknown` class was not an "other" class, because it contained instances very similar to the classes to be found (transitions between activities and activities of no interest to the study), the SVM also showed difficulties in finding hyperplanes separating the `Unknown` class from the others. In the HIS dataset, since the `Unknown` class represented more than 18% of all examples, most of the models were highly perturbed. For instance, the Random Forest, being composed of decision trees, tried to generate a set of trees leading to leaves of consistent classes. However, due to the diverse nature of the instances of the `Unknown` class, generating consistent leaves became very hard. Conversely, the CRF, by considering all the classes, captured complex dependencies in the feature window to give better classification of `Unknown` instances.

The two datasets used in this study, though being of the same nature and comparable, were not acquired with the same participants or in the same smart home. But the most prominent difference between them is the amount of information each of them provides. The HIS dataset is far less informative than the multimodal SWEET-HOME one, due to a lower number of sensors. That explains why the non-sequential models were so competitive in the SWEET-HOME case, being able to benefit from informative data, such as door and window activations, as well as temporal data (the previously occupied room). In the HIS case, the number of features being smaller, the non-sequential models exhibited the lowest performance, while the sequential models (HMM, CRF, MLN) stayed competitive. While classical sequential models such as HMM and CRF benefited from the history described within the sequence, the MLN-based approaches took advantage of their high expressibility. For instance the MLN induced the following rules:

1. SWEET-HOME dataset:

```
 1.66047 percentageagitationroom(window,HIGH)
                          -> class(window,PHONE)
 1.13709 speech_studywindow(window,HIGH)
                          -> class(window,PHONE)
-1.24098 totalagitation(window,LOW)
                          -> class(window,PHONE)
-0.175539 previous(window1,window2) and
       class(window1,EATING) -> class(window2,PHONE)
```

2. HIS dataset:

```
 1.50854 percentageoc_livingroom(window2,MEDIUM)
                          -> class(window,PHONE)
 1.25924 timesinceinthisroom(window2,LOW)
                          -> class(window,PHONE)
-1.00466 timesinceinthisroom(window2,HIGH)
                          -> class(window,PHONE)
0.92 previous(window1,window2) and
       class(window1,PHONE) -> class(window2,PHONE)
```

where the head of each rule $class(w, c)$ takes $w$, the current window and predicts $c$, the class of the window. Each predicate in the body indicates the value of a feature in the window. For instance `speech_studywindow(window,HIGH)` indicates that there was a high amount of speech detected in the study room. Positive values to the left side of each rule indicate that, when the rule is fired, it adds confidence to the class, while a negative value shows that this rule removes credence from the class. For the particular class "Phone", MLN was able to translate the fact that when someone is talking a lot close to the phone for a short time, then s/he is most likely phoning (recall we assume a single person occupation hypothesis). Thus, despite the relative low performance of the MLN, this model seems to be a good candidate to represent higher-level activities such as iADL [43] as it is able to express complex semantic relations that purely probabilistic models cannot express.

Regarding speech/non-speech audio information, the results of the feature selection performed on both datasets during the tuning phase suggested that the most important features for activity recognition were those related to the location of the inhabitant followed by those related to speech/non-speech sound occurrences. This can also explain the role of acoustic information on the final accuracy. Even when the most important aspect was the location of the inhabitant since all the activities were performed in at most two rooms, it was also difficult to disambiguate two different activities that took place in the same room. The total agitation in a room, which was highly dependent on the number of sound events, was helpful to differentiate between eating and cleaning, both performed in the kitchen. In this particular case, the agitation produced by room doors and windows were very similar, however it was the number of sounds which helped classifiers to differentiate the activities. Likewise, reading and communication, when both performed in the study, had similar settings on door contacts and light states, but the number of speech events detected was informative enough for good classification. Also, in the MLN model, the weights of the rules relating acoustic information to some activities were large when the association was relevant, as in the following examples:

```
-1.572  percentagetime_sound(win,LOW) -> class(win,EATING)
 1.095  totalagitation(win,LOW) -> class(win,READING)
 1.148  totalagitation(win,MEDIUM) -> class(win,PHONE)
```

In this example, the first rule indicates that an eating activity is unlikely to generate a low amount of sound. The second rule expresses the fact that a reading activity is likely to generate a low amount of sound while the third rule shows that a phoning activity is expected to generate many sound events. These rules are further evidence that audio information is important for activity recognition.

## 6. Conclusion and future work

The study presented in this paper brings the following contributions:

1. The paper presents a complete framework for online AR, making it possible to summarise asynchronous as well as continuous sampled signals into temporal windows.
2. This framework has been evaluated on two smart home datasets, available to the community [80, 27], integrating acoustic information, a kind of information which has been rarely included in previous studies of the domain. This evaluation shows the interest of these acoustic features for AR since they relate to the agitation level of the occupants (noise) as well as their social interactions (speech).
3. The AR task in the framework has been implemented with different sequential and instance-based models. This includes a recent model for AR —the Markov Logic Network— in both sequential and non-sequential versions. The evaluation exhibited strengths and weaknesses of each of the models for the AR task.
4. The models were evaluated on the datasets mentioned above in an realistic way since windows of unknown class are fed to the classifiers. Moreover, to avoid overfitting, a cross-validation technique was designed so as to exclude from the learning set one of the participant records used for testing.

Overall, Conditional Random Fields (CRF) are very competitive for on-line activity recognition from non-visual, audio and home automation sensors. Even though non-sequential models such as Random Forests show good performance on some datasets, the CRF approach is more robust to the presence of activities of `Unknown` class, since it showed the least decrease in performance between the datasets with and without the `Unknown` activities included. The performance of each method was assessed in a Leave-One-Subject-Out-Cross-Validation (LOSOCV), so that no data from the same participant was used both in the training and testing sets in the same trial. Hence, the genericity of the model was not biased by the presence of data relating to the participant being used in that test.

Although the CRF has the best performance overall, generative models such as the HMM and the MLN also perform well. These models show interesting features as they are able to model each class independently, and thus do not bias their learning towards the largest class. Moreover, the Markov Logic Network approach (MLN) is a statistical-relational model, and so its logical structure could be learned in conjunction with *a priori* knowledge provided by expert rules, so that the model can benefit from highly expressive previous knowledge whilst also being able to handle uncertainty.

The results presented in this study are based on two different datasets that have not been acquired in the same environment and that work with different sensors. Although the datasets are different in terms of activities considered, sensors and quantity of data, it has been shown that the trend in the results from the two sets are similar.

Finally, it has to be noted that, for some of the results, the standard deviation is quite large between subjects. Some of the activities were represented with few samples and the difference between subjects is then more predominant. For future models, a generic model that adapts to a participant with the first samples would be a very good direction of research.

We plan to extend our work in two directions. On the one hand, we would like to compare a window-based approach, which loses semantics and temporality but summarises the data well, against an event/state-based approach, which keeps semantic and time information but necessitates the use of even more robust models to handle errors in the data stream. It would be interesting to study the behaviour of the CRF and MLN approaches in these two cases. On the other hand, one of the main problems in human activity learning is the lack of annotated data. Indeed, in-lab recording of scenarios allow an accurate annotation with many participants, but it is not the case for real-world data. Field experiments in real homes do provide more realistic data but annotation is often performed by the participants themselves and cannot easily be verified [7]. Moreover, it is difficult to recruit participants who would be willing to have surveillance technology set up in their own home for experimental purposes. Besides, collecting real-world data is highly expensive in terms of time and resources. This is why we intend to use learning methods that either deal with partially labelled data [75] or use a Universal Background Model [64] so that a large amount of data, of which only a small portion is annotated, can be used for classification.

## Acknowledgements

This work is a par of the SWEET-HOME project supported by the Agence Nationale de la Recherche (ANR-09-VERS-011). Many thanks to the anonymous reviewers for their very constructive comments that helped improve the manuscript.

## References

[1] Aggarwal, J. and Ryoo, M. (2011). Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16:1–16:43.

[2] Artikis, A., Paliouras, G., Portet, F., and Skarlatidis, A. (2010a). Logic-based representation, reasoning and machine learning for event recognition. In *the 4th ACM International Conference on Distributed Event-Based Systems (DEBS)*, pages 282–293.

[3] Artikis, A., Skarlatidis, A., and Paliouras, G. (2010b). Behaviour recognition from video content: a logic programming approach. *International Journal on Artificial Intelligence Tools*, 19(2):193–209.

[4] Artikis, A., Skarlatidis, A., Portet, F., and G., P. (2012). Logic-Based Event Recognition. *Knowledge Engineering Review*, 27(4):469–506.

[5] Augusto, J. C. and Nugent, C. D. (2004). The use of temporal reasoning and management of complex events in smart homes. *European Conference on Artificial Intelligence*, pages 778–782.

[6] Bao, L. and Intille, S. (2004). Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, volume 3001 of *Lecture Notes in Computer Science*, pages 1–17. Springer.

[7] Blachon, D., Portet, F., Besacier, L., and Tassart, S. (2014). RecordMe: A Smartphone Application for Experimental Collections of Large Amount of Data Respecting Volunteer's Privacy. In *UCAmI 2014*, pages 345–348, Belfast, United Kingdom.

[8] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory, ACM*, pages 144 – 152, Pittsburgh.

[9] Bouakaz, S., Vacher, M., Bobillier-Chaumon, M.-E., Aman, F., Bekkadja, S., Portet, F., Guillou, E., Rossato, S., Desserée, E., Traineau, P., Vimon, J.-P., and Chevalier, T. (2014). CIRDO: Smart companion for helping elderly to live at home for longer. *Innovation and Research in BioMedical engineering (IRBM)*, 35(2):101–108.

[10] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[11] Chahuara, P., Fleury, A., Vacher, M., and Portet, F. (2012). Méthodes SVM et MLN pour la reconnaissance automatique d'activités humaines dans les habitats perceptifs: tests et perspectives. In *Actes de la conférence RFIA 2012*, pages 340–347, Lyon, France.

[12] Chahuara, P., Portet, F., and Vacher, M. (2011). Location of an Inhabitant for Domotic Assistance Through Fusion of Audio and Non-Visual Data. In *Pervasive Health*, pages 1–4, Dublin, Ireland. European Alliance for Innovation. http://www.pervasivehealth.org/.

[13] Chahuara, P., Portet, F., and Vacher, M. (2013). Making Context Aware Decision from Uncertain Information in a Smart Home: A Markov Logic Network Approach. In *Ambient Intelligence*, volume 8309 of *Lecture Notes in Computer Science*, pages 78–93, Dublin, Ireland. Springer.

[14] Chen, L., Hoey, J., Nugent, C., Cook, D., and Yu, Z. (2012a). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(6):790–808.

[15] Chen, L., Nugent, C., Mulvenna, M., Finlay, D., Hong, X., and Poland, M. (2008). A logical framework for behaviour reasoning and assistance in a smart home. *International Journal of Assistive Robotics and Mechatronics*, 9(4).

[16] Chen, L., Nugent, C., and Wang, H. (2012b). A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):961–974.

[17] Chieu, H., Lee, W., and Kaelbling, L. (2006). Activity recognition from physiological data using conditional random fields. In *Proc. of the Singapore-MIT Alliance Symposium (SMA)*.

[18] Clarkson, B. (2003). *Life patterns : structure from wearable sensors*. PhD thesis, Massachusetts Institute of Technology, USA.

[19] Cook, D. J., Youngblood, M., III, E. O. H., Gopalratnam, K., Rao, S., Litvin, A., and Khawaja, F. (2003). Mavhome: An agent-based smart home. *IEEE International Conference on Pervasive Computing and Communications*, page 521.

[20] Coutaz, J., Crowley, J. L., Dobson, S., and Garlan, D. (2005). Context is key. *Communications of the ACM*, 48(3):49–53.

[21] Dalal, S., Alwan, M., Seifrafi, R., Kell, S., and Brown, D. (2005). A rule-based approach to the analysis of elders activity data: Detection of health and possible emergency conditions. In *AAAI Fall 2005 Symposium*.

[22] de Carolis, B. and Cozzolongo, G. (2004). C@sa: Intelligent home control and simulation. In *International Conference on Computational Intelligence (ICCI)*, pages 462–465, Istanbul, Turkey.

[23] Duong, T., Phung, D., Bui, H., and Venkatesh, S. (2009). Efficient duration and hierarchical modeling for human activity recognition. *Artificial Intelligence*, 173(7-8):830–856.

[24] Flanagan, J. A., Mantyjarvi, J., and Himberg, J. (2002). Unsupervised clustering of symbol strings and context recognition. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 171–178, Washington, DC, USA. IEEE Computer Society.

[25] Fleury, A., Noury, N., and Vacher, M. (2011). Improving supervised classification of activities of daily living using prior knowledge. *International Journal of E-Health and Medical Communications (IJEHMC)*, 2(1):17–34.

[26] Fleury, A., Vacher, M., and Noury, N. (2010). SVM-based multi-modal classification of activities of daily living in health smart homes: Sensors, algorithms and first experimental results. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):274 –283.

[27] Fleury, A., Vacher, M., Portet, F., Chahuara, P., and Noury, N. (2012). A French corpus of audio and multimodal interactions in a health smart home. *Journal on Multimodal User Interfaces*, 7(1):93–109.

[28] Gallissot, M., Caelen, J., Jambon, F., and Meillon, B. (2013). Une plate-forme usage pour l'intégration de l'informatique ambiante dans l'habitat : Domus. *Technique et Science Informatiques (TSI)*, 32/5:547–574.

[29] Getoor, L. and Taskar, B., editors (2007). *Introduction to Statistical Relational Learning*. The MIT Press.

[30] Haigh, K. Z. and Yanco, H. (2002). Automation as caregiver: A survey of issues and technologies. In *Proceedings of the AAAI-02 Workshop Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, pages 39–53.

[31] Hamid, M. (2008). *A Computational Framework For Unsupervised Analysis of Everyday Human Activities*. PhD thesis, Georgia Institute of Technology, USA.

[32] Helaoui, R., Niepert, M., and Stuckenschmidt, H. (2010). A statistical-relational activity recognition framework for ambient assisted living systems. In *Ambient Intelligence and Future Trends-International Symposium on Ambient Intelligence (ISAmI 2010)*, pages 247–254, Guimarães, Portugal.

[33] Helaoui, R., Niepert, M., and Stuckenschmidt, H. (2011). Recognizing interleaved and concurrent activities: A statistical-relational approach. In *International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9.

[34] Htike, Z. Z., Egerton, S., and Kuang, Y. C. (2010). Real-time human activity recognition using external and internal spatial features. In *Sixth International Conference on Intelligent Environments*, pages 52–57, Kuala Lumpur, Malaysia. IEEE Computer Society.

[35] Intille, S. S. (2002). Designing a home of the future. *IEEE Pervasive Computing*, 1(2):76–82.

[36] Istrate, D., Castelli, E., Vacher, M., Besacier, L., and Serignat, J.-F. (2006). Information extraction from sound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions on*, 10(2):264–274.

[37] Katz, S. (1983). Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, 31(12):721–727.

[38] Kersting, K., Raedt, L. D., and Raiko, T. (2006). Logical hidden markov models. *Journal of Artificial Intelligence Research*, 25:425–456.

[39] Khan, A., Lee, Y.-K., Lee, S., and Kim, T.-S. (2010). A tri-axial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Transactions on Information Technology in Biomedicine*, 14(5):1166–1172.

[40] Krishnan, N. C. and Cook, D. J. (2014). Activity recognition on streaming sensor data. *Pervasive and Mobile Computing*, 10:138–154.

[41] Kurgan, L. A. and Cios, K. J. (2004). CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):145–153.

[42] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning,ICML'01*, pages 282–289.

[43] Lawton, M. and Brody, E. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. *Gerontologist*, 9:179–186.

[44] Le Bellego, G., Noury, N., Virone, G., Mousseau, M., and Demongeot, J. (2006). A model for the measurement of patient activity in a hospital suite. *IEEE Transactions on Information Technologies in Biomedicine*, 10(1):92 – 99.

[45] Lee, M.-W., Khan, A. M., and Kim, T.-S. (2011). A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation. *Personal Ubiquitous Computing*, 15(8):887–898.

[46] Liao, L. (2006). *Location-based Activity Recognition*. PhD thesis, University of Washington, USA.

[47] Lin, W., Sun, M.-T., Poovendran, R., and Zhang, Z. (2008). Activity recognition using a combination of category components and local models for video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1128–1139.

[48] Lowd, D. and Domingos, P. (2007). Efficient weight learning for markov logic networks. In *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 200–211.

[49] Minnen, D. (2008). *Unsupervised Discovery of Activity Primitives from Multivariate Sensor Data*. PhD thesis, Georgia Institute of Technology, USA.

[50] Molina, L. C., Belanche, L., and Nebot, À. (2002). Feature selection algorithms: A survey and experimental evaluation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2002)*, pages 306–313.

[51] Mozer, M. C. (2005). Lessons from an adaptive home. In Cook, D. and Das., R., editors, *Smart Environments: Technologies, protocols, and applications*, pages 271–294. John Wiley & Sons, Inc.

[52] Naeem, U. and Bigham, J. (2008). Activity recognition using a hierarchical framework. In *Second Int. Conf. on Pervasive Computing Technologies for Healthcare*, pages 24 – 27.

[53] Natarajan, S., Bui, H. H., Tadepalli, P., Kersting, K., and Wong, W. (2008). Logical hierarchical hidden markov models for modeling user activities. In *Proceedings of the 18th international conference on Inductive Logic Programming*, pages 192–209, Prague, Czech Republic. Springer-Verlag.

[54] Nazerfard, E., Das, B., Holder, L. B., and Cook, D. J. (2010). Conditional random fields for activity recognition in smart environments. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 282–286, New York, NY, USA. ACM.

[55] Ni, Q. N., García Hernando, A., and Pau de la Cruz, I. (2015). The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development. *Sensors*, 15:11312–11362.

[56] Okeyo, G., Chen, L., Wang, H., and Sterritt, R. (2012). Dynamic sensor data segmentation for real-time knowledge-driven activity recognition. *Pervasive and Mobile Computing*. (in press).

[57] Patterson, D. J., Etzioni, O., Fox, D., and Kautz, H. (2002). Intelligent ubiquitous computing to support alzheimer's patients: Enabling the cognitively disabled. In *Fourth International Conference on Ubiquitous Computing*, pages 21–22, Göteborg, Sweden.

[58] Pentney, W., Popescu, A.-M., Wang, S., Kautz, H., and Philipose, M. (2006). Sensor-based understanding of daily life via large-scale use of common sense. In *Proceedings of the 21st National Conference on Artificial intelligence - Volume 1*, AAAI'06, pages 906–912. AAAI Press.

[59] Portet, F., Vacher, M., Golanski, C., Roux, C., and Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17:127–144.

[60] Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, pages 725–730. AAAI/MIT Press.

[61] Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[62] Rasanen, O. (2012). Hierarchical unsupervised discovery of user context from multivariate sensory data. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2105 – 2108.

[63] Rashidi, P. and Mihailidis, A. (2013). A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics*, 17(3):579–590.

[64] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3):19 – 41.

[65] Rialle, V. (2007). Rapport sur les technologies nouvelles susceptibles d'améliorer les pratiques gérontologiques et la vie quotidienne des malades âgés et de leur famille. Technical report, Rapport remis à M. Philippe Bas, Ministre de la Santé et des Solidarités, République Française. 74p.

[66] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1-2):107–136.

[67] Rota, N. and Thonnat, M. (2000). Activity recognition from video sequences using declarative models. In Horn, W., editor, *European Conference on Ambient Intelligence (ECAI)*, pages 673–680. IOS Press.

[68] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517.

[69] Segal, M. (2004). Machine learning benchmarks and random forest regression. Technical report, Center for Bioinformartics and Molecular Biostatistics, University of California, San Francisco, CA, USA.

[70] Storf, H., Becker, M., and Riedl, M. (2009). Rule-based activity recognition framework: Challenges, technique and learning. In *3rd International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2009*, pages 1–7, London, UK.

[71] Tang, P. and Venables, T. (2000). Smart homes and telecare for independent living. *Journal of Telemedicine and Telecare*, 6(1):8–14.

[72] Tapia, E. M., Intille, S. S., and Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, 2:158–175.

[73] Taskar, B., Abbeel, P., and Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 485–492, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[74] Tong, Y. and Chen, R. (2014). Latent-Dynamic Conditional Random Fields for recognizing activities in smart homes. *Journal of Ambient Intelligence and Smart Environments*, 6:39–55.

[75] Truyen, T., Bui, H., Phung, D., and Venkatesh, S. (2008). Learning discriminative sequence models from partially labelled data for activity recognition. In Ho, T.-B. and Zhou, Z.-H., editors, *PRICAI 2008: Trends in Artificial Intelligence*, volume 5351 of *Lecture Notes in Computer Science*, pages 903–912. Springer Berlin Heidelberg.

[76] Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., and Chahuara, P. (2015). Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation. *ACM Transactions on Accessible Computing*, 7(issue 2):1–36.

[77] Vacher, M., Chahuara, P., Lecouteux, B., Istrate, D., Portet, F., Joubert, T., Sehili, M., Meillon, B., Bonnefond, N., Fabre, S., Roux, C., and Caffiau, S. (2013). The SWEET-HOME Project: Audio Technology in Smart Homes to improve Well-being and Reliance. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, pages 7298–7301, Osaka, Japan.

[78] Vacher, M., Fleury, A., Portet, F., Serignat, J.-F., and Noury, N. (2010). Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living. In Campolo, D., editor, *New Developments in Biomedical Engineering*, pages pp. 645 – 673. In-Tech. ISBN: 978-953-7619-57-2.

[79] Vacher, M., Fleury, A., Serignat, J.-F., Noury, N., and Glasson, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. In *Interspeech'08*, pages 496–499, Brisbane, Australia. 4p.

[80] Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

[81] Vail, D. L., Veloso, M. M., and Lafferty, J. D. (2007). Conditional random fields for activity recognition. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '07, pages 235:1–235:8, New York, NY, USA.

[82] van Kasteren, T., Englebienne, G., and Kröse, B. (2010). An activity monitoring system for elderly care using generative and discriminative models. *Personal and Ubiquitous Computing*, 14(6):489–498.

[83] van Kasteren, T. and Krose, B. (2007). Bayesian activity recognition in residence for elders. In *3rd IET International Conference on Intelligent Environments*, pages 209–212, Ulm, Germany.

[84] van Kasteren, T. L. M., Englebienne, G., and Kröse, B. J. A. (2011). Hierarchical activity recognition using automatically clustered actions. In *Proceedings of the Second international conference on Ambient Intelligence*, AmI'11, pages 82–91, Berlin, Heidelberg. Springer-Verlag.

[85] Velik, R. (2014). A brain-inspired multimodal data mining approach for human activity recognition in elderly homes. *Journal of Ambient Intelligence and Smart Environments*, 6(4):447–468.

[86] Wang, X. and Ji, Q. (2012). Learning dynamic bayesian network discriminatively for human activity recognition. In *21st International Conference on Pattern Recognition (ICPR 2012)*, pages 3553–3556.

[87] Zappi, P., Lombriser, C., Stiefmeier, T., Farella, E., Roggen, D., Benini, L., and Tröster, G. (2008). Activity recognition from on-body sensors: Accuracy-power trade-off by dynamic sensor selection. In Verdone, R., editor, *European Conference on Wireless Sensor Networks*, volume 4913 of *Lecture Notes in Computer Science*, pages 17–33. Springer.

[88] Zouba, N., Bremond, F., Thonnat, M., Anfosso, A., Pascual, E., Mallea, P., Mailland, V., and Guerin, O. (2009). A computer system to monitor older adults at home: Preliminary results. *Gerontechnology Journal*, 8(3):129–139.