Disseminating Synthetic Smart Home Data for Advanced Applications

Andrea Masciadri
Politecnico di Milano
Como 22100, Italy
andrea.masciadri@polimi.it

Fabio Veronese
Politecnico di Milano
Como 22100, Italy
fabio.veronese@polimi.it

Sara Comai
Politecnico di Milano
Como 22100, Italy
sara.comai@polimi.it

Ilaria Carlini
Politecnico di Milano
Como 22100, Italy
ilaria.carlini@mail.polimi.it

Fabio Salice
Politecnico di Milano
Como 22100, Italy
fabio.salice@polimi.it

Abstract

The research in IoT and Smart Homes fields is rapidly growing, leading to the emergence of new services to improve the health and lifestyle of people based on the analysis of data that they produce performing their daily activities. However, researchers report a lack of high-quality publicly-available datasets: conducting experiments gathering such data is long and expensive, especially if the annotation of meaningful information (environment, person's activity, health status) is required. Moreover, there are even more specific settings (e.g., dementia detection) where data must be related to a change in inhabitants' behavior. We present a collection of new publicly-available datasets generated with the SHARON simulator. Thanks to this software, researchers can obtain synthetic data suiting their specific requirements. Two classes of datasets are described: one extends existing datasets preserving the original statistical properties, the other is composed of simulations of virtual inhabitant-environment systems. Moreover, we induced behavioral drifts compatible with dementia symptoms, generating further datasets. We believe that these resources may help the progress of research, as long as new real-life high-quality datasets are not available.

1 Introduction

The possibility of gathering large amounts of data from Smart Home environments is a valuable opportunity for the development of numerous applications, like, e.g., security, home automation, remote monitoring, etc.

Data are collected by using different types of sensors, connected to a home (usually wireless) network and stored in a central database. Localization of the inhabitants, state of the house such as brightness, temperature, humidity, doors and windows opening, as well as the activation of household appliances can be a source of knowledge for advanced analytics.

Moreover, in addition to the mentioned data, there is often a need for extensive descriptions of the context in which the data were collected: the so-called "ground-truth". For example, much attention has been dedicated to the research in the Activity Recognition (AR) field – that is the task of identifying the ongoing Activity of Daily Living (ADL) from sensors data. As highlighted by Sprint et al. [SCFSE16], in order to access the Health Events related to a person living in a Smart Environment, supervised machine-learning algorithms are commonly used. Usually, AR requires a set of labels related to the performed ADLs: these data are provided by external annotators (often called oracles) which look at them and utilize extra information (such as videos, the house floor-plan, the resident profile, etc.) to generate corresponding ground-truth labels.

It comes clear that creating Smart Home datasets with ground-truth information related to the inhabitant's activities and well-being status is a long and costly operation that often slows down the progress of research and advanced applications. In the next section we provide an overview of the currently publicly-available datasets, highlighting the strengths and weaknesses of the various resources. In Section 3 we describe a new set of resources, their peculiarities and how have they been generated. Finally, in Section 4 we conclude this work discussing future challenges about the dissemination of Smart Home datasets.

2 Background

In recent years, many papers have been discussing the importance of the continuous monitoring of the person's behavior as a source of information concerning his/her well-being [RBC+15, PKL+05, PLJ+15]. According to Saives et al. [SPF15], improving the life of the inhabitant with new technological services makes a house "Smart"; those applications cover several interesting research fields, all of them sharing the same need to collect Home Automation datasets. A literature review by Rashidi et al. from 2013 reports 18 noticeable projects in Ambient Assisted Living, and confirms that "one of the most important component is Human Activity Recognition" [RM13]. Despite the great interest in research concerning Activity Recognition (AR) and Behavioral Drift Detection (BDD), the amount of publicly-available high-quality datasets is particularly small. Indeed, the collection of Home Automation data in controlled settings, with good annotation, is a hard and resource demanding task.

Table 2 summarizes the features of the most widely used datasets in the literature to evaluate AR and BDD research; as reported by Benmansour et al. [BBF16], AR and BDD with multiple residents introduce complexity in identifying the dwellers and disassociating data and activities.

ARAS (Activity Recognition with Ambient Sensing) is a project developed aiming at ADL recognition [AEIE13]. The authors have published their datasets, that comprise data collected from two houses with two inhabitants, for a duration of one month each. The deployed sensors set was composed of 20 boolean sensors, and data were annotated with 27 different ADLs. The dataset however reports erratic routine of the inhabitants (unusual meal times, unexpected behavior during the ADL, etc.), specifies only one activity at a time (even when two happens concurrently), and reports ADLs which cannot be identified due to sensor lack (e.g., no sensor to detect *"using internet"* and *"reading"* activities were present).

CASAS (Center for Advanced Studies in Adaptive Systems) is a research project and a department of the Washington State University very active in AR studies. Their focus is to design a smart home "small in form, lightweight in infrastructure, extendable with minimal effort, and ready to perform key capabilities out of the box", through their *Smart Home in a Box* project [CCTK13]. The success of this project enabled the collection and publication of several datasets, so that many AR research studies worked using CASAS data [CSEC+09]. Nonetheless, the annotation of the datasets is restricted to a reduced subset of the freely available data, and in most of these cases it was obtained thanks to an automatic labeling method rather than using a personal diary or an oracle. Finally, the variety of the installed sensors is often restricted to two different types (motion and temperature), reducing the possibility of advanced data analysis.

Table 1: Datasets comparison.

	# Houses	Multiple residents	Duration	# Sensors	# Activities
Aras	2	y	2 m	20	27
Casas	>38	y	2-8 m	20-86	11
MIT	2	n	2 w	77-84	13
Kasteren	3	n	>1 m	21	27

Tapia et al. [TIL04] presented two datasets related to two houses with a single resident each collected by *MIT*. They comprise data collected from many Boolean sensors (up to 85) for two weeks each. Activity annotation was achieved asking the inhabitant to use a Person Digital Assistant (PDA). Every 15 minutes candidates were reminded by the PDA to record the performed activities. Even if this methodology is less intrusive and less demanding than spontaneous annotation, it resulted to be less accurate probably because it is not spontaneous. Moreover, the reduced duration makes it less relevant for traditional machine learning methods.

T.Van Kasteren [VKNEK08], working at an Activity Recognition project at University of Amsterdam, has collected a dataset concerning two houses with single inhabitant. The volunteers houses were instrumented with 20 boolean sensors, collecting data for 28 days. The annotation was done directly by the inhabitant, but it reports some inaccurate entries, as well as some unexpected data (e.g., sensors always

on/off).

Referring to the reported projects, we can subsume the weak points of publicly-available datasets as follows:

- Limited Sensor Variety: many projects use few sensors or a limited variety in sensed quantity.
- Limited Extension: projects involving several volunteers, present short duration per-person; conversely, long lasting collections refer only to few participants;
- Limited Annotation Reliability: inhabitants and automatic methods could lead to insufficient results in terms of accuracy and, in some cases, the single activity annotation is not sufficient to describe properly the experimental settings;
- Heterogeneity: every project defines its set of activities, sensors, standards, and protocols, resulting in non-comparable datasets;
- Specificity and Applicability: most of the projects report data collected with a specific intent, not necessarily matching the aim of other research groups; dually, if a dataset is collected in generic settings, it might not contain some specific situations required by other research groups.

Moreover, we would like to emphasize the lack of attention devoted to the behavioral change annotation. Indeed, all the mentioned datasets have a too short time duration and/or have no annotation concerning such modifications in the inhabitant behavior.

Alternative approaches for the dataset collection phase consist in substituting the real world system with a simulation software producing synthetic data [Mas, MN06, AR07]. In this paper we present a collection of datasets generated with SHARON simulator, which can be tuned to produce highly customized synthetic home automation data for advanced applications.

3 Synthetic Smart Home Datasets

The datasets we present have been obtained exploiting SHARON's sensor data generation algorithms, with different environments and inhabitant behaviors.

The resources are accessible at the persistent URL *http://www.purl.org/synthetic_sh_dataset*, and are available under the Creative Commons Attribution 3.0 CC-BY License; when exploiting the hereby included data, please cite the work of Veronese et al. [VMT+16]. The resources and the software to generate further data are also available at the institutional website of the Assistive Technology Group ATG [b1115].

3.1 SHARON simulator

SHARON is a tool developed in the BRIDGE project [MSV+15] to face the lack of data for advanced Smart Home applications such as Activity Recognition. The simulator is structured in two main layers: the **top layer** generates the daily activity schedule, the **bottom layer** translates them into sensor activations. The software can be tuned designing the dwelling characteristics, the virtual sensors models, and a set of parameters representing the inhabitant response to needs (e.g., hunger, tiredness, boredom, stress, etc.). The activity schedule attempts to satisfy the person needs in relation to the time of the day, the weekly cycle, the weather conditions and other non-deterministic components. The bottom level relies either on a *virtual agent*, performing the scheduled action in the environment and activating the sensors following a set of alternative patterns, or on a *statistical module*, reproducing the activations of sensors given an activity as performed in an available training dataset.Finally it is possible to program a change in the simulation parameters so that the inhabitant behavior is affected accordingly.

All the details about the data generation model implemented in SHARON to produce Synthetic Smart Home Data are available in the work of Veronese et al. [VPC+14]. The evaluation of the simulator has already been performed through a cross-validation process applied on a real world dataset (ARAS [AEIE13]); the work in Veronese et al. [VMT+16] reports the results for both the layers of the SHARON simulator:

- **Top layer validation (ADL scheduling):** three different validation metrics (Bhattacharyya distance [Bha46], Earth Mover Distance [Hit41] and Kullback-Leibler divergence [KL51]) have been used to evaluate the difference between activity distributions in the generated dataset with respect to a test set extracted from the original dataset. The same distance has been computed between a training set of the above mentioned real world dataset (*original dataset*) and the test set; Figure 1 shows that the ADL scheduling generated by the SHARON simulator is compatible with the schedule of the original dataset.
- **Bottom layer validation (Sensor activations):** the Bhattacharyya distance have been computed to compare the sensor activation distributions in the ARAS dataset with respect to

Table 2: Bhattacharyya distance of the sensor activation distributions in the generated datasets with respect to the original dataset; smaller values represent closer distributions. A: Agent, S: Statistical

ADL	Lunch		Shower		Cleaning	
DATASET	**A**	**S**	**A**	**S**	**A**	**S**
Couch	0.34	0.06	-	-	0.79	0.43
Chair 1	0.38	0.29	-	-	-	-
Chair 2	0.25	0.47	-	-	0.47	0.59
Fridge	0.66	0.41	-	-	0.54	0.54
K. Drawer	0.74	0.60	-	-	-	-
B. Door	-	-	0.16	0.11	-	-
Shower	-	-	0.63	0.34	-	-
Hall	0.83	0.89	-	-	0.36	0.77
K. Mov.	0.22	0.20	-	-	0.33	0.21
Tap	-	-	-	-	0.94	0.54
K. Temp.	0.18	0.19	-	-	-	-

the generated datasets (using both the agent module and the statistical module). Table 2 reports the results for three relevant activities: *Cleaning* (where the sequence of sensor activations is almost random), *Lunch* (where several executions are different, but keeping an overall procedure), and *Having Shower* (where the procedural connotation is strong).

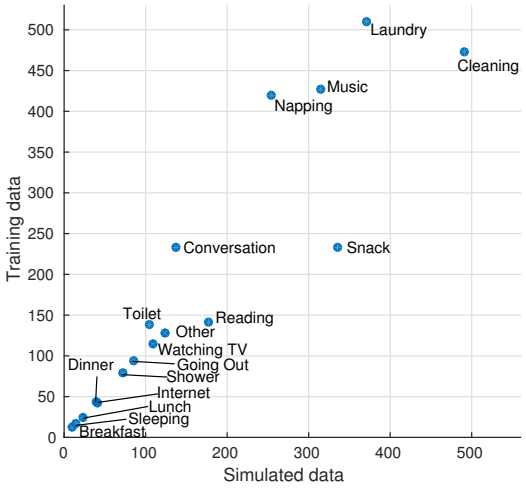

Figure 1: Comparison of the Earth Mover Distance between the activity distributions in simulated data and training data as reported by Veronese et al. [VMT+16].

3.2 Dataset description

The generated dataset has been obtained using the SHARON simulator; every day of simulation is represented by two text files: one describing the ADL scheduling and one describing sensor activations. The former contains all the performed activities - one activity per line - with the starting time, the activity identifier, and the activity name in a comma separated format. The latter contains 86400 lines - one for every second of the day - reporting the boolean status of every sensor of the house separated by blank characters.

The proposed datasets refer to three different house models. Each dataset comprises 90 days of the virtual inhabitant life, and has an alternative version comprising an injected behavioral drift compatible with dementia symptoms, that can be used for comparison. In the following, the characteristics of different classes of datasets are described; they are summarized in Tables 3 and 4.

Table 3: New datasets comparison. A: ARAS, K: Van Kasteren, V: V-House.

	House Map	Type	Drift	Days	Annot.
A-ext-norm	A	Statistical	no	90	yes
A-ext-dem	A	Statistical	yes	90	yes
A-agn-norm	A	Agent-based	no	90	yes
A-agn-dem	A	Agent-based	yes	90	yes
K-ext-norm	K	Statistical	no	90	yes
K-ext-dem	K	Statistical	yes	90	yes
V-agn-norm	V	Agent-based	no	90	yes
V-agn-dem	V	Agent-based	yes	90	yes

Table 4: New datasets performed ADLs.

	Sleeping	Breakfast	Lunch	Dinner	Toilet	Shower	Working	Cleaning	Internet	Relax	Reading	Watching TV	Going Out	Other
A-*	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
K-*	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓
V-*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

3.2.1 ARAS dataset extension

This first group of datasets comprises four synthetic home automation datasets (their names start with *A-**) based on a virtual reproduction of the ARAS project test environment [AEIE13]. Two of them (*A-ext-**) have been obtained by training SHARON over the behavior of one of the original ARAS project in-

habitants, resulting in an *extension* of the original data. The other two (*A-agn-**) have been obtained using the same ADL scheduling but with an *agent-based* simulation. Two variants with behavioral drift due to dementia (**-dem*) are also available.

Environment

The house environment exploited for simulation comprises 20 binary home automation sensors. The location is a simple apartment with four main spaces: bedroom, bathroom, and an openspace with kitchen and living room. Most common sensors are motion detectors, but in this environment there are also tap, toilet, and shower sensors, pressure detectors on chairs, sofa and bed.

Inhabitant

The inhabitant routine comprises two different patterns for weekdays and weekends. During the weekdays the inhabitant spends a daily amount of time outside the dwelling (for working activities), while during the weekend leisure is the main occupation (relax, reading, internet, etc.). There are 13 performed activities, as described in Table 4, plus an unqualified activity "Other".

3.2.2 Van Kasteren dataset extension

The second dataset group (*K-**) is related to the research project home by Van Kasteren et al. [VKNEK08]. In this case the virtual environment reproduces the experimental house, as well as the sensor activations, which are produced after a training on the original data. The results are two datasets: one with the extension of the real dataset (*K-ext-norm*), the other with the superimposed behavioral drift (*K-ext-dem*).

Environment

The house environment exploited for simulation comprises 21 binary home automation sensors. The location is a two-storey apartment: on the first floor there are a bathroom and an open-space with kitchen and livingroom; the second floor is composed by two bedrooms, a bathroom, and a study room. Installed sensors include motion sensors to detect doors, drawers and cupboards openings, tap and shower sensors, sensors to detect appliances uses, pressure detectors on chairs, sofa and bed.

Inhabitant

The dataset describes 12 activities, the same of the ARAS datasets, except for "Working" and "Internet" activities that are missing (Table 4). The inhabitant routine comprises two different patterns for weekdays and weekends, mainly by differentiating the time and duration of meals.

3.2.3 V-Home dataset

This last group of datasets are fully virtual (V-*). The authors designed a simple four room house, and programmed an easy routine for a virtual inhabitant. The obtained datasets are based on an agent based sensor activation simulation, one with plain routine (V-agn-norm), the other with the injected drift (V-agn-dem).

Environment

The virtual designed environment includes 11 binary sensors. The house is designed with four main rooms: kitchen, bedroom, bathroom, and livingroom. Most devices are movement sensors, with open-close detectors on main door and bathroom cabinet.

Inhabitant

The inhabitant routine represents a remote-worker, working 8 hours at home in weekdays, and relaxing in the weekends. The activities are 14, with the addition of an unqualified "Other".

3.3 Behavioral Drift Description

Alzheimer's Disease (AD) is becoming widespread as reported by AD International [WJB+13]: there will be up to 65.7 million people living with dementia worldwide by 2030 and up to 115.4 million by 2050. The typical symptoms of AD involve the daily routine, concerning: forgetfulness, difficulty performing ADL, incontinence, speech problems, wandering and getting lost, depression, sleep disorders. In the provided dataset (**-dem*) this condition is simulated by replicating part of the symptoms. The time taken to perform complex tasks such as "Take a shower" is increased by 20%, its rate is decreased by 15%. The duration of nighttime sleep passes from an average of 8 uninterrupted hours to 4.5 hours fragmented up to 5 times, while naps appear during the day. The frequency of activities such as "Dinner" and "Going out" slightly decreases.

4 Discussion and Future Work

The presented datasets, generated with SHARON, are a support resource for research groups working on smart home data processing for advanced applications. Even if with some limitations, the proposed data are a resource to foster such research, avoiding the costs of creating a real world testbed. Moreover, the software SHARON is publicly-available, enabling to generate further different data with particular conditions

and behavioral drifts, and overcoming the lack of high-quality real world datasets. The quality of the data generated by the simulator has been discussed in the work of Veronese et al. [VMT+16], which has already attracted the attention of the scientific community that has expressed a willingness to access data. We believe that this could be used as a tool to provide early tests for new methods development (e.g., Activity Recognition and Behavioral Drift Detection), before allocating time and financial resources. The provided datasets are only a possible application of the simulation software, whose next releases will include further features and a user friendly web interface to allow the generation of high quality synthetic datasets.

References

[AEIE13] Hande Alerndar, Halil Ertan, Ozlem Durmaz Incel, and Cem Ersoy. Aras human activity datasets in multiple homes with multiple residents. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pages 232–235. IEEE, 2013.

[AR07] Ibrahim Armac and Daniel Retkowitz. Simulation of smart environments. In *IEEE International Conference on Pervasive Services*, pages 257–266. IEEE, 2007.

[b1115] Assistive Technology Group (ATG) of Politecnico di Milano. http://www.atg.deib.polimi.it, 2015.

[BBF16] Asma Benmansour, Abdelhamid Bouchachia, and Mohammed Feham. Multioccupant activity recognition in pervasive smart home environments. *ACM Computing Surveys (CSUR)*, 48(3):34, 2016.

[Bha46] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946.

[CCTK13] Diane J Cook, Aaron S Crandall, Brian L Thomas, and Narayanan C Krishnan. Casas: A smart home in a box. *Computer*, 46(7), 2013.

[CSEC+09] Diane Cook, M Schmitter-Edgecombe, Aaron Crandall, Chad Sanders, and Brian Thomas. Collecting and disseminating smart home sensor data in the casas project. In *Proceedings of the CHI Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research*, 2009.

[Hit41] Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys*, 20(2):224–230, 1941.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

[Mas] Mason project website. http://cs.gmu.edu/eclab/projects/mason. Accessed: 2015.

[MN06] Miquel Martin and Petteri Nurmi. A generic large scale simulator for ubiquitous computing. In *2006 Third Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services*, pages 1–3. IEEE, 2006.

[MSV+15] Simone Mangano, Hassan Saidinejad, Fabio Veronese, Sara Comai, Matteo Matteucci, and Fabio Salice. Bridge: Mutual reassurance for autonomous and independent living. *Intelligent Systems, IEEE*, 30(4):31–38, 2015.

[PKL+05] Paula Paavilainen, Ilkka Korhonen, Jyrji Lötjönen, Luc Cluitmans, Marja Jylhä, Antti Särelä, and Markku Partinen. Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy. *Journal of sleep research*, 14(1):61–68, 2005.

[PLJ+15] Kirsten KB Peetoom, Monique AS Lexis, Manuela Joore, Carmen D Dirksen, and Luc P De Witte. Literature review on monitoring technologies and their outcomes in independently living elderly people. *Disability and Rehabilitation: Assistive Technology*, 10(4):271–294, 2015.

[RBC+15] Daniele Riboni, Claudio Bettini, Gabriele Civitarese, Zaffar Haider Janjua, and Viola Bulgari. From lab to life: Fine-grained behavior monitoring in the elderly's home. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE*

International Conference on, pages 342–347. IEEE, 2015.

[RM13] Parisa Rashidi and Alex Mihailidis. A survey on ambient-assisted living tools for older adults. *IEEE journal of biomedical and health informatics*, 17(3):579–590, 2013.

[SCFSE16] Gina Sprint, Diane Cook, Roschelle Fritz, and Maureen Schmitter-Edgecombe. Detecting health and behavior change by analyzing smart home sensor data. In *Smart Computing (SMARTCOMP), 2016 IEEE International Conference on*, pages 1–3. IEEE, 2016.

[SPF15] Jérémie Saives, Clément Pianon, and Gregory Faraut. Activity discovery and detection of behavioral deviations of an inhabitant from binary sensors. *IEEE Transactions on Automation Science and Engineering*, 12(4):1211–1224, 2015.

[TIL04] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. *Activity recognition in the home using simple and ubiquitous sensors*. Springer, 2004.

[VKNEK08] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9. ACM, 2008.

[VMT+16] Fabio Veronese, Andrea Masciadri, Anna A Trofimova, Matteo Matteucci, and Fabio Salice. Realistic human behaviour simulation for quantitative ambient intelligence studies. *Technology and Disability*, 28(4):159–177, 2016.

[VPC+14] F Veronese, D Proserpio, S Comai, M Matteucci, and F Salice. Sharon: a simulator of human activities, routines and needs. *Studies in health technology and informatics*, 217:560–566, 2014.

[WJB+13] Anders Wimo, Linus Jönsson, John Bond, Martin Prince, Bengt Winblad, and Alzheimer Disease International. The worldwide economic impact of dementia 2010. *Alzheimer's & Dementia*, 9(1):1–11, 2013.