

Analysis and effects of smart home dataset characteristics for daily life activity recognition

Iram Fatima · Muhammad Fahim ·
Young-Koo Lee · Sungyoung Lee

© Springer Science+Business Media New York 2013

Abstract Over the last few years, activity recognition in the smart home has become an active research area due to the wide range of human centric-applications. With the development of machine learning algorithms for activity classification, dataset is significantly important for algorithms testing and validation. Collection of real data is a challenging process due to involved budget, human resources, and annotation cost that's why mostly researchers prefer to utilize existing datasets for evaluation purposes. However, openly available smart home datasets indicate variation in terms of performed activities, deployed sensors, and environment settings. Unfortunately, the analysis of existing datasets characteristic is a bottleneck for researchers while selecting datasets of their intent. In this paper, we develop a Framework for Smart Homes Dataset Analysis (FSHDA) to reflect their diverse dimensions in predefined format. It analyzes a list of data dimensions that covers the variations in time, activities, sensors, and inhabitants. For validation, we examine the effects of proposed data dimension on state-of-the-art activity recognition techniques. The results show that dataset dimensions highly affect the classifiers' individual activity label assignments and their overall performances. The outcome of our study is helpful for upcoming researchers to develop a better understanding about the smart home datasets characteristics with classifier's performance.

I. Fatima · M. Fahim · Y.-K. Lee (✉) · S. Lee
Department of Computer Engineering, Kyung Hee University, Seoul, South Korea
e-mail: yklee@khu.ac.kr

I. Fatima
e-mail: iram.fatima@oslab.khu.ac.kr

M. Fahim
e-mail: fahim@oslab.khu.ac.kr

S. Lee
e-mail: sylee@oslab.khu.ac.kr

Keywords Smart home · Pervasive sensing · Dataset analysis · Classifiers performance · Activity recognition

1 Introduction

The knowledge convergence in sensing environments and pervasive computing has created massive interest in the development of smart homes [1–3]. A smart home is an intelligent agent that perceives state of resident and the physical environments using sensors. Recent advancements in the field of machine learning and data mining have enabled activity recognition research using smart homes sensing data to play a direct role in improving the quality of health care. All techniques for activity recognition depend on various types of methods, parameters, and diverse characteristics of sensing data. Availability of real time data is significantly important for testing and evaluation of methods and parameters. However, feasibility of new dataset collections is inadequate due to involved budget, human resources, and annotation cost. Due to these reasons, mostly researchers prefer to utilize existing datasets for the testing and validation of their techniques.

A large number of smart home datasets are publicly available and activities performed by individuals can be characterized by duration, frequency, sequential order, temporal order, and other factors such as age, and gender [4]. Important attributes that must be considered while analyzing smart home datasets include the time scale that is necessary to estimate the times of different activities. For example, eating requires more time than taking medication. Periodic variations may occur in daily, weekly, monthly, annual, and even seasonal activities [5]. For example, cleaning is more likely to occur on weekends than on weekdays. Parallel activities may occur, in which one user can perform more than one activity in a single time unit [6]. Sequential activities are characterized by preceding and following activities to identify the influences of different activities on each other. For example, taking medicine is very likely followed by eating. Therefore, the acquisition of large datasets poses several challenges for their use and analysis. The existing machine learning algorithms utilize smart home datasets in terms of performed activities, deployed sensors, allocated time, and number of inhabitants. The development of a standard framework for data analysis is significantly important for realistic performance evaluation of the methods. Unfortunately, there are no standardized methods available that can provide detail analysis of datasets.

In this paper, we develop a Framework for Smart Home Datasets Analysis (FSHDA) to represent them in a predefined format. Our proposed method analyzes the time duration, activity count, activity type, sensor count, sensor type, sensor location, activity occurrence, activity sensor events, activity time, parallel activity count, parallel activity type, and sequences of parallel activities. These dimensions can provide detailed views regarding differences in environments, residents, and activities and provide similar information about every dataset so that researchers can choose datasets according to algorithm requirements and application domains. A particular dataset cannot be classified with the same accuracy from all classifiers. To select an appropriate classifier for a certain type of data, there is a need to understand the behavior of classifiers on different data characteristics. Suppose there are four classifiers

$k_1, k_2, k_3,$ and $k_4,$ and three classes $c_1, c_2,$ and $c_3.$ Let us assume that depending on the training set and the set of features used, classifier k_1 is more efficient to classify given points in class $c_1.$ Similarly, depending on the configuration, classifiers $k_2, k_3,$ and k_4 are more efficient in classifying points in classes $c_2, c_3,$ and $c_1,$ respectively. During our analysis, we experience a lot of variation in annotations of activities along their sensor events. In all datasets, few activities are incomplete which means they start but end annotation tag is missing. In some cases, activities are complete however, there is no sensor event fired during that time. In this case, no algorithm can detect such activities because there is no sensor information in sensory data. We find out the relationship between the distribution of data, on the one hand, and classifier performance, on the other. We analyze the effects of data dimensions on state-of-the-art classifiers such as Artificial Neural network (ANN) [7], Hidden Markov Model (HMM) [8], Conditional Random Field (CRF) [9], and Support Vector Machines (SVM) [10]. These four selected schemes are applied on four analyzed datasets selected from three most significant smart home projects such as CASAS [11], ISL [1], and House_n [12] smart homes. The experimental results provide an evaluation of the existing activity classification schemes to resolve the uncertainties associated with the choice of classifier and the nature of smart home dataset. The results show that neither of the classifier is best for all datasets; the classification accuracy of each classifier depends on the underline data characteristics.

The rest of the paper is organized as follows. We briefly describe related work in Sect. 2. In Sect. 3, we introduce proposed framework for datasets analysis. The analyzed results of the CASAS, ISL, and House_n smart home datasets are presented in Sect. 4. In Sect. 5, we discuss the effects of analyzed data dimensions on state-of-the-art machine learning techniques. Finally, conclusion and future works are given in Sect. 6.

2 Related work

Smart homes can be divided into two types of environments according to the interactions and types of deployed sensors: event based and continuously sensing environments [1–3]. An event based environment can be built by using binary sensors that observe ON/OFF states when inhabitants interact with appliances (i.e., cabinet, stove, chair) [1]. A continuously sensing environment can be built using environmental sensors (i.e., temperature, humidity, light) and sending the sensed data periodically to the server [5]. Advances in wireless communications enabled multiple sensor devices to send and receive data over long distance [13, 14]. In such environments, sensory data have different levels of abstraction, different representations, and are diverse in nature [15] that significantly affects the performance of activity recognition techniques.

The aim of smart home technology is to provide ambient assisted living for care delivery, remote monitoring, and promotion of residential safety and quality of life [16–20]. Several studies have been conducted to better understand the smart homes and their applications based on activity recognition. In [16], authors studied the impact of few dataset features on classification accuracy of machine learning algorithm

based on semi-Markov models. They considered the availability of labeled data, importance of required training time, and speedy inference of ISL dataset for experimental purpose. In their analysis, they showed that CRF outperforms among semi-Markov models. The authors in [10] apply SVM to identify daily living activities on their smart home dataset. They selected a set of features from dataset according to their domain of interest before using multi-SVM for effective activity classification as compare to other classifiers. The work in [17] applied the ANN to cluster human activities of daily living within their own developed smart home environment. They proposed GSOM (Growing Self Organizing maps) based approach to utilize the appropriate data dimensions for cluster analysis of human activities effectively.

The authors in [18] synthesizes the sensor information collected from CASAS smart home and extracted the wide range of useful features. They compared several machine learning algorithms on the selected features to compare the performance of activity recognition. They discussed the performance of the machine learning algorithm based on information gain and mRMR. The authors in [19] resolved the problem of sensor selection for activity recognition in smart homes along with classifier selection. They examine the issue of selecting and placing sensors in a CASAS smart home in order to maximize activity recognition accuracy. In [8], authors used ISL smart home dataset to show the potential of generative and discriminative models for recognizing activities. They presented that CRFs are more sensitive to overfitting on a dominant class than HMM. The work in [20] assisted to understand the boundaries of context-aware computing. They helped application designer to decide which features of data are important to implement. A survey about major techniques related to activity recognition work can be found in [21]. The underneath of these contexts are deployed sensors and algorithms that played an important role for high performance and accuracy. Some of the work [22–25] has been done for data analysis of sensory data, but none of the work has focus on analysis of data based on smart home sensor.

The commonly observed methodologies in literature for smart home datasets are to propose machine learning algorithms and select the one, which gives relatively better results for their particular domain. According to best of our knowledge, no existing work has intentions to develop a system that can analyze the different data dimensions for better understanding of smart home datasets so application developer can know what features are important for particular classification task. In this study, we measure the predefined data dimensions and analyze the classifiers performance to show the effects of data characteristics. Our study will help the researchers in choosing an appropriate dataset according to intended requirements of classifiers and associated parameters.

3 Proposed framework for dataset analysis

In smart homes, datasets are collected in the form of raw data files and intended to support the diverse applications for ambient assisted living. Dataset is composed of “Log Files” *LF* to keep the record of active sensors at different time slots and “Annotation Files” *AF* to maintain the list of performed activities in temporal manner. Our proposed Framework for Smart Homes Dataset Analysis (FSHDA) process the raw data to compute the data dimensions as (a) *total time duration* for collected dataset

time, (b) *activity type and count* for annotated activities with their labels, (c) *sensor type and count* for deployed sensors with their labels, (d) *smart home characteristics* for dataset characteristics in terms of dataset time, annotated activities and deployed sensors. (e) *activity occurrences* for number of unique annotated activities, (f) *activity time* for time duration of individual activity, (g) *activity sensor event* for the number of generated sensor events for a particular activity, (h) *idle activity time* for the total time of unannotated activities, (i) *idle activity sensor events* for generated sensor events during idle activity time, (j) *parallel activity type and count* for total number of parallel activities with their labels, (k) *sequence of parallel activity* for the start and end sequences of parallel activities. These dimensions assist the researchers to better understand the nature of performed activities before finalizing the classification technique.

To define the proposed framework, we assume $A_n^D = \{a_1, a_2, a_3, \dots, a_n\}$ are n annotated activities in the dataset “ D ”, and $S_m^D = \{s_1, s_2, s_3, \dots, s_m\}$ are m deployed sensors in the smart home for a duration of $T : T_d^D = \{t_1, t_2, t_3, \dots, t_d\}$. The set of unique time stamps $U_T = \text{unique}(T_d^D)$, unique activities $U_A = \text{unique}(A_n^D)$, unique deployed sensors $U_S = \text{unique}(S_m^D)$ and $S_{act} = \text{active sensor events}(S_n^D)$ are the set of sensor values and events. The proposed data dimensions for uniform representation and compact analysis of smart home datasets are computed in Table 1 to generalize the characteristics of all datasets. The pseudo code for the computation of data dimension from (a) to (c), (e) to (f), and (i) to (j) are given in Algorithms 1, 2, and 3, respectively.

Algorithm 1 Computation of data dimensions (a), (b) and (c)

Input: *LF*: Log File
AF: Annotation File

Output: *TD*: Total Time Duration
ATC: Activity Type and Count
STC: Sensor Type and Count

Begin

- 1 $TD = \text{endTime}(LF) - \text{startTime}(LF)$
- 2 $actCount = \text{unique}(AF(:, (\text{size}(AF)-1)))$
- 3 **for** $i = 1:\text{length}(actCount)$
- 4 $actType[i] = \text{map}(actCount[i], label)$
- 5 **end**
- 6 $ACT = \{actType, actCount\}$
- 7 $senCount = \text{unique}(LF(:, (\text{size}(AF)-1)))$
- 8 **for** $j = 1:\text{length}(senCount)$
- 9 $senType[i] = \text{map}(senCount[i], label)$
- 10 **end**
- 11 $STC = \{senType, senCount\}$

End

The proposed framework¹ is publically available to compute data dimensions of other smart home datasets in addition to already evaluated datasets. It has been im-

¹<http://uclab.khu.ac.kr/ext/iram/FSHDA>.

Table 1 The FSHDA data dimension calculations

Data Dimension	Function for calculations
Total time duration	$= \sum_{i=1}^n t_{d_i}^D \forall t' \in U_T$
Activity type and count	$= \sum_{i=1}^n a_i^D \forall a' \in U_A$
Sensor type and count	$= \sum_{i=1}^n s_i^D \forall s' \in U_S$
Smart home characteristics	$= [\sum_{i=1}^n t_{d_i}^D \sum_{i=1}^n a_i^D \sum_{i=1}^n s_i^D]$
Activity occurrences	$= \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_i, a_j)) \forall \begin{cases} 1 & a_i = a_j \\ 0 & a_i \neq a_j \end{cases}$
Activity time	$= \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_{ij}, \Delta t_{ij}))$ $\Delta t_{ij} = a_{ij\text{end}} - a_{ij\text{start}}$ and $(a_{ij\text{end}} \gg a_{ij\text{start}})$
Activity sensor events	$= \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_i', s_j')) \forall s_j' \in S_{act}$
Idle activity time	$= \sum_{i=1}^n t_{d_i}^D - \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_{ij}, \Delta t_{ij}))$
Idle activity sensor event	$= \sum_{i=1}^n s_i^D - \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_i', s_j'))$
Parallel activity type and count	$= \sum_{i=1}^n a_i' (\sum_{j=1}^m (a_j, t_j))$ $\forall \begin{cases} a_j & a_{i\text{start}} < a_{j\text{start}} \wedge a_{i\text{end}} > a_{j\text{end}} \\ & a_{i\text{start}} < a_{j\text{start}} \wedge a_{i\text{end}} < a_{j\text{end}} \\ & a_{i\text{start}} > a_{j\text{start}} \wedge a_{i\text{end}} < a_{j\text{end}} \\ & a_{i\text{start}} > a_{j\text{start}} \wedge a_{i\text{end}} > a_{j\text{end}} \\ & a_{i\text{start}} = a_{j\text{start}} \wedge a_{i\text{end}} = a_{j\text{end}} \\ 0 & \forall \text{others} \end{cases}$
Sequence of parallel activities	$= \sum_{i=1}^n a_i' \Delta t_i (\sum_{j=1}^m (a_j, t_j))$ $\forall \begin{cases} a_j & a_{i\text{start}} < a_{j\text{start}} \wedge a_{i\text{end}} > a_{j\text{end}} \\ & a_{i\text{start}} < a_{j\text{start}} \wedge a_{i\text{end}} < a_{j\text{end}} \\ & a_{i\text{start}} > a_{j\text{start}} \wedge a_{i\text{end}} < a_{j\text{end}} \\ & a_{i\text{start}} > a_{j\text{start}} \wedge a_{i\text{end}} > a_{j\text{end}} \\ & a_{i\text{start}} = a_{j\text{start}} \wedge a_{i\text{end}} = a_{j\text{end}} \\ 0 & \forall \text{others} \end{cases}$

plemented in MATLAB 7.6 on an Intel Pentium(R) Dual-Core 2.5 GHz with 3 GB of memory and Microsoft Windows 7. The analysis of three most significant smart home datasets is presented in the preceding section.

4 Results of data dimensions analysis

In this section, we experimentally validate the proposed framework of dataset analysis. We calculate a list of data dimensions from a set of diverse datasets in order to express consistency and scalability of the proposed framework. During analysis, we keep the same activity labels as mentioned in the collected datasets. The details of the datasets and the results of analysis are shown in the following subsections. However, the scope of proposed framework is not limited to the following datasets and researchers can analyze other datasets with the help of publically available version of the proposed framework.

Algorithm 2 Computation of data dimensions (e), (f), (g) and (h)

Input: *LF*: Log File
AF: Annotation File

Output: *AO*: Activity Occurrences
AT: Activity Time
ASE: Activity Sensor Events
IAT: Idle Activity Time
IASE: Idle Activity Sensor Events

Begin

```

1 actCount = unique(AF(size)-1)
2 for i= 1: length (actCount)
3   AO[i] = count(AF(actCount))
4   AT[i] = sum(AF(endTime) - AF(startTime), actCount(i))
5   index=1
6   for j= 1: actCount(i)
7     sensEvent=0
8     for k= 1: length (LF)
9       if (startTime(AF,j) >= startTime(LF,k) && endTime (AF,j) <= endTime(AF,k))
10        |   senEveTemp[index] = sensEvent+1
11        |   end
12        |   end
13        |   index = index+1
14      end
15      ASE [i] = senEveTemp
16    end
17 IAT = (sum(LF(end) - sum(LF(start)) - sum(AT)
18 IASE = count(AF(sensor)) - sum(ASE)

```

End

Algorithm 3 Computation of data dimensions (i) and (j)

Input: *LF*: Log File
AF: Annotation File
AO: Activity Occurrences

Output: *PATC*: Parallel Activity Type and Count
SPA: Sequence of Parallel Activities0

Begin

```

1 for i= 1 : length (AF)
2   for j= 1: length (AF)
3     if (AF (i,startTime) < AF(j,startTime)) && (AF(i,endTime) < AF (j, endTime))
4       |   parallelAct1 = AF(j,activity)
5       |   parallelAct2 = AF(i,activity)
6     elseif (AF (i,startTime) < AF(j,startTime)) && (AF(i,endTime) > AF (j, endTime))
7       |   parallelAct1 = AF(i,activity)
8       |   parallelAct2 = AF(j,activity)
9     end
10    end
11  end

```

End

Table 2 The FSHDA calculation of CASAS smart home dataset characteristics

Datasets	Activity count	Sensor count	Duration	Inhabitants
Tulum2010	16	37	149 days	2
Tworsummer2009	12	86	55 days	2
Tulum2009	10	20	83 days	2
Twor2009	13	71	46 days	2

4.1 CASAS smart home dataset analysis

The CASAS smart home project is a research project at Washington State University (WSU), we selected the four most recent datasets for analysis of our proposed framework. Different kinds of sensors, such as temperature sensors, motion sensors, and binary sensors, are deployed at various locations. These sensors are placed on multiple objects like doors, kitchen stove burners, TV lounge, and other places in the home environment [26]. The datasets are collected for long durations of time, and thousands of sensor events are generated while single or multiple inhabitants perform daily life activities. In Table 2, the results of computed data dimension activity count, sensor counts, duration, and inhabitant information are given to outline the characteristics of the datasets. It is obvious from Table 2 that *Tulum2010* is the largest dataset of CASSAS smart home in terms of activity count and the duration of data collection compared to others.

The detailed characteristics of four CASAS datasets are computed under our developed framework and results are shown in Table 3. During analysis, we identify some common activities that are annotated by almost all datasets are “toileting”, “meal preparation”, “sleeping”, “watching TV”, “personal hygiene”, and “work”. Among these activities, the most commonly annotated activity is “meal preparation”, which is annotated 1791 times in *Tulum2010*. Our computed dimensions show that in some datasets, a few activities are annotated at the macrolevel, while in others the same activities are annotated at microlevels. For example, cooking is annotated as “meal preparation” in all datasets except *Tulum2009*, where it is annotated as “cook breakfast” and “cook lunch”. During analysis, the other noticeable activities are those that are only annotated one or two datasets, such as “yoga” (*Tulum2010*), “study” (*Tulum2009*), and “R1 snack” (*Tulum2009*), which occur 24, 9, and 491 times, respectively.

During the dataset collection period, some activities are performed in parallel and our framework compute these activities by considering as separate data dimensions and results are shown in Table 4 with their counts and dataset descriptions. The pairs of parallel activities do not always have the same start and end sequence. For example, “clean” and “R1 work” are parallel activities in *Twor2009*. But their start and end sequences are not same for all occurrences. Similarly, “cleaning” and “grooming” are parallel in *TwoSummer2009* with different start and end sequences.

Table 3 The FSHDA computed characteristics of CASAS smart home datasets *Tulum2010*, *TwoSummer2009*, *Twor2009*, and *Tulum2009*. The “Num” column shows activity count, the “Time” column shows activity time in minutes, and the “Sensor” column shows activity sensor events

Activities	Tulum2010			TwoSummer2009			Twor2009		
	Num.	Time	Sensor	Num.	Time	Sensor	Num.	Time	Sensor
Idle	–	230654.4	39450	–	18550.79	267222	–	8240.938	73043
Bed Toilet Transition	153	241.18	2277	2	3.36	63	39	94.55	2241
Meal Preparation	1791	5036.2	91817	199	1788.32	41744	118	6207.32	41730
R1 Sleeping in Bed	137	74230.52	74710	52	18202.24	43914	35	18428.36	29503
R2 Sleeping in Bed	131	76716.67	79300	26	9723.08	29762	35	18572.11	29604
Bathing	556	47159.95	17039	–	–	–	–	–	–
Eating	625	4198.58	68255	–	–	–	–	–	–
Personal Hygiene	1312	3037.66	24189	–	–	–	–	–	–
Work Bedroom 1	1174	10713.91	52107	–	–	–	–	–	–
Work Bedroom 2	1924	17732.21	84742	–	–	–	–	–	–
Work LivingRoom	1016	9893.18	39782	–	–	–	–	–	–
Work Table	1191	2899.09	52589	–	–	–	–	–	–
Yoga	24	330.44	3312	–	–	–	–	–	–
Cleaning	–	–	–	28	143.31	3375	2	49.75	1540
R1 Work	–	–	–	443	22262.50	284570	59	5902.51	45675
R2 Work	–	–	–	352	7245.24	84965	44	2530.03	17955
Grooming	–	–	–	47	351.14	7172	–	–	–
R1 Shower	–	–	–	41	556.52	4125	–	–	–
R1 Wakeup	–	–	–	48	15.84	289	–	–	–
R2 Shower	–	–	–	28	321.52	4950	–	–	–
R2 Wakeup	–	–	–	25	36.11	393	–	–	–
Tulum2009									
Idle	–	–	–	–	102986.4	203408	–	–	–
Wash Dishes	197	388.24	9530	71	1204.84	24869	–	–	–
Watch TV	2174	13132.59	91075	528	4955.43	52222	31	3228.77	17879
Enter Home	114	179463.8	288	73	119.42	604	–	–	–
Leave Home	116	39.92	386	75	101.58	1854	–	–	–
Cook Breakfast	–	–	–	80	1440.31	33435	–	–	–
Cook Lunch	–	–	–	71	972.74	24527	–	–	–
Group Meeting	–	–	–	11	1847.04	31084	–	–	–
R1 Eat Breakfast	–	–	–	66	932.87	20077	–	–	–
R1 Snack	–	–	–	491	4461.85	81183	–	–	–
R2 Eat Breakfast	–	–	–	47	497.06	13649	–	–	–
R1 Bed to Toilet	–	–	–	–	–	–	34	337.06	2298
R1 Personal Hygiene	–	–	–	–	–	–	45	663.32	5818
R2 Personal Hygiene	–	–	–	–	–	–	39	1029.47	8237
Study	–	–	–	–	–	–	9	922.71	8133
Wash_Bathtub	–	–	–	–	–	–	1	33.09	219

Table 4 The FSHDA computed list of parallel activities for CASAS smart home datasets

Dataset	Parallel activities		Num.
	Activity 1	Activity 2	
Twor2009	Clean	R1 Work	2
	Meal Preparation	R1 Bed to Toilet	1
	Meal Preparation	R1 Personal Hygiene	1
TwoSummer2009	Bed toilet transition	R1 work	1
	Cleaning	Grooming	4
	Cleaning	R2 work	1
Tulum2009	Cook Breakfast	R1 Eat Breakfast	4

Table 5 The FSHDA calculation of ISL smart home dataset characteristics

Datasets	Activity count	Sensor count	Duration	Inhabitants	Environment
House A	16	14	24 days	Male, 26 years	3 room apartment
House B	25	23	15 days	Male, 28 years	2 room apartment
House C	17	21	20 days	Male, 57 years	2 storey home with 6 rooms

4.2 ISL smart home dataset analysis

The Intelligent System Laboratory (ISL) is a research group from the University of Amsterdam, recorded three single subject datasets² using three different smart homes, *House A*, *House B*, and *House C* [27, 28]. The data dimensions related to number of deployed sensors, performed activities, and inhabitants of each house are presented in Table 5. From computed results, we can infer that the data collection duration is longest in *House A*, while *House B* has larger activity and sensor counts. *Houses A* and *B* are one-room apartments, while *House C* data is collected in a two-story home [27].

The detailed data dimensions in terms of activity count, time durations, and sensor events of three smart homes are computed and results are depicted in Table 6. From these statistics, we identify the most commonly annotated activities in the datasets are “leaving”, “toileting/toilet downstairs”, “showering”, “brush teeth”, “sleeping/go to bed”, and “prepare dinner”. The occurrences of “toileting/toilet downstairs” are more numerous in *House A* than in the others. Activities that are only annotated in *House A* are “load dishwasher”, “unload dishwasher”, “store groceries”, “unload washing machine”, and “receive guest”. Among these activities, “store groceries” occurs only once, while the occurrences of the remaining activities are similar to each other. The annotated activities in *House B* are higher as compared to *House A* and *House C*, but the number of occurrences of these activities is very low. For example, “shaving”, “unpacking”, and “on phone” are annotated only once each. The only activities annotated for *House C* are “relax”, “use toilet upstairs”, “take medication”,

²<https://sites.google.com/site/tim0306/datasets>.

Table 6 The FSHDA computed characteristics of ISL smart home datasets *House A*, *House B*, and *House C*. The “Num” column shows activity count, the “Time” column shows activity time in minutes, and the “Sensor” column shows activity sensor events

Activities	House A			House B			House C		
	Num.	Time	Sensor	Num.	Time	Sensor	Num.	Time	Sensor
Idle	–	5817.23	23	–	3691.50	6589	–	14496.4	1176
Leaving	33	19664.2	84	23	11183.5	1065	29	4200.83	147
Toileting/toilet Downstairs	114	155.38	402	26	31.52	157	21	12.42	51
Showering	23	136.38	59	11	104.92	2242	14	173.15	996
Brush Teeth	10	9.78	22	12	17.67	308	25	51.25	231
Sleeping/Go to bed	24	7914.37	183	13	5812.2	5531	19	7460.46	18805
Prepare Dinner	9	306.47	128	6	82.99	1068	11	269.12	272
Snack	12	24.33	50	–	–	–	5	1.37	14
Prepare Breakfast	20	39.42	122	–	–	–	10	57.65	112
Eating	1	22.56	0	–	–	–	23	162.02	144
Drink	20	12.23	63	7	4.62	77	7	4.38	25
Load Washing Machine	3	4.01	7	–	–	–	1	41.12	18
Load Dishwasher	5	31.85	15	–	–	–	–	–	–
Unload Dishwasher	4	15.23	27	–	–	–	–	–	–
Store Groceries	1	1.183	3	–	–	–	–	–	–
Unload Washing Machine	4	3.27	9	–	–	–	–	–	–
Receive Guest	3	424.93	65	–	–	–	–	–	–
Shaving	–	–	–	1	8.99	111	7	45.55	133
Get Dressed	–	–	–	12	17.92	88	21	36.27	181
Brunch	–	–	–	9	75.33	592	–	–	–
Wash Dishes	–	–	–	6	22.32	163	–	–	–
Answering Phone	–	–	–	1	0.29	2	–	–	–
Eat Dinner	–	–	–	5	6.12	22	–	–	–
Eat Brunch	–	–	–	3	0.78	9	–	–	–
Setting up Sensors	–	–	–	1	32.58	77	–	–	–
Unpacking	–	–	–	1	35.83	52	–	–	–
Install Sensor	–	–	–	2	75.07	408	–	–	–
On phone	–	–	–	1	14	0.3	–	–	–
Fasten Kitchen Camera	–	–	–	3	5.90	81	–	–	–
Play Piano	–	–	–	23	355.79	271	–	–	–
Gwenn Searches Keys	–	–	–	1	8.48	56	–	–	–
Prepare for Leaving	–	–	–	8	15.37	90	–	–	–
Drop Dish (No dish wash)	–	–	–	3	1.04	16	–	–	–
Relax	–	–	–	–	–	–	36	1742.28	298
Use Toilet Upstairs	–	–	–	–	–	–	29	13.54	95
Take Medication	–	–	–	–	–	–	3	4.79	12
Lunch	–	–	–	–	–	–	8	27.33	71

Table 7 The FSHDA calculation of House_n smart home dataset characteristics

Datasets	Activity count	Sensor count	Duration	Inhabitants
Subject 1	13	28	16 days	30 year old woman
Subject 2	9	20	16 days	80 year old woman

and “lunch”. Among these activities, “relax” and “use toilet upstairs” are significantly more frequent than the other two activities.

Proposed framework found a set of parallel activities in all three datasets. In the case of *House A*, “get drink” and “receive guest” are parallel activities and annotated three times in the dataset. Similarly, for *House B* “eat brunch” and “prepare brunch” are annotated two times as parallel activities. In the case of *House C* “eatubg” is parallel to “prepare dinner” and “relax” activities. During analysis, discrepancies are found between the reported and actual statistics of activity count, and sensor count of these datasets. The reported numbers of activities are 10, 14, and 16, while the actual activity counts are 16, 25, and 17 for *Houses A, B, and C*, respectively. In *House A*, “eating” is annotated one time in the dataset, but no sensor event is fired within that period so it is not detectable during computation. That is why the total count for actual detected activities is reduced from 16 to 15. The most variations are found in the analysis of *House B* data dimensions, “prepare breakfast” and “get drink” activities are annotated in reported statistics; however, in actual statistics, we did not find even a single occurrence for these activities. Furthermore, the “wash toaster” activity is annotated once for 2.45 minutes in *House B*, but no sensor event is generated at that time.

4.3 House_n smart home dataset analysis

The House_n smart homes project from University of MIT, collect data by using a set of simple state change sensors [29]. Two datasets³ are collected for two different subjects. Both individuals live alone in one bedroom apartments. The number of performed activities, deployed sensors and duration with inhabitants’ information are shown in Table 7.

The detail characteristics of the performed activities are computed by predefined data dimensions as depicted in Table 8. From computed statistics, we identify the most common activities of both datasets are “toileting”, “washing dishes”, “preparing breakfast”, “preparing lunch”, “preparing dinner”, and “preparing a snack”. Among these activities, the “toileting” is annotated almost twice as often for *Subject 1* as for *Subject 2* (82 vs. 33 times). The durations of data collection and sensor events also vary accordingly. The number of occurrences of “washing dishes” for *Subject 2* is three times that for *Subject 1* (20 vs. 7). The occurrences of “preparing dinner” are almost double for *Subject 2* as for *Subject 1*. Activities that are only annotated for *Subject 1* are “preparing a beverage”, “dressing”, “bathing”, “grooming”, “cleaning”, “doing laundry”, and “going to work”. Among these activities, “grooming” has the

³<http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/>.

Table 8 The FSHDA computed characteristics of House_n smart home datasets *Subject 1*, and *Subject 2*. The “Num” column shows activity count, the “Time” column shows activity time in minutes, and the “Sensor” column shows activity sensor events

Activities	Subject 1		Subject 2			
	Num.	Time	Sensor	Num. Time	Sensor	
Idle	–	20930.6	731	–	19624.6	814
Toileting	82	161.5	323	33	158.98	167
Washing Dishes	7	42.23	67	20	125.82	89
Preparing Breakfast	14	182.42	147	18	276.63	197
Preparing Lunch	17	524.37	497	20	662.57	339
Preparing Dinner	8	136.72	122	14	445.57	228
Preparing a Snack	13	58.43	66	16	157.4	106
Preparing Beverage	15	55.47	77	–	–	–
Dressing	24	88.7	121	–	–	–
Bathing	18	343.93	224	–	–	–
Grooming	37	216.98	302	–	–	–
Cleaning	8	149.67	145	–	–	–
Doing Laundry	19	146.12	172	–	–	–
Going out to Work	12	2.87	25	–	–	–
Taking Medication	–	–	–	14	26.42	39
Watching TV	–	–	–	15	571.7	177
Listening to Music	–	–	–	14	990.32	143

most occurrences, while “cleaning” has the least. Activities that are only annotated for *Subject 2* are “taking medication”, “watching TV”, and “listening to music”; the number of occurrences of these activities are all similar.

During analysis, parallel activities are found in case of *Subject 2*, where “listening to music” is parallel to “preparing a snack” for three times, “preparing lunch” for two times, and once for “toileting” and “washing dishes” at different time slots. In the reported statistics, for *Subject 1*, 77 sensors are deployed, however, only 28 of those sensors are found to generate events after analysis from the proposed framework. Similarly, activity count also varies such as “toileting” is reported 85 times, but its actual count is 82 times because no sensor event is fired during start and end duration of remaining three activities. In the case of *Subject 2*, the reported number of deployed sensors is 84, however, analysis results show that only 20 of those sensors are activated during the data collection period. Similarly, the reported count for “listening to music”, “toileting”, and “washing dishes” are 18, 40, and 21, however, their actual count is 14, 33, and 20 times, respectively.

5 Effects of proposed data dimensions on activity recognition

In this section, we validate the effectiveness of the proposed data dimensions through demonstrating how performances of activity recognition techniques are influenced by the data dimensions calculated by the proposed framework. The proposed methodology of validation consists of three major modules. (1) Data preprocessing: to represent the sensory data as an observation vector for machine learning algorithms. (2) Classifiers for activity recognition: to introduce the applied classifier for the domain of activity recognition with preferred settings for our experiments. (3) Results

of variations in classifiers performances and discussion: to show the effect of data dimensions through demonstrating how performances of activity recognition techniques are influenced by the dataset characteristics. The detail of each module is described in the following subsections.

5.1 Data preprocessing

Data preprocessing is an important step towards accurate training of machine learning techniques [30]. Data collected from ubiquitous sensors based on subject interactions are stored in sensor logs and annotation files with attributes start time, end time, sensor id, sensor value, and activity label. In order to recognize the performed activities, recorded dataset is preprocessed into the form of $\{(x_1, y_1), \dots, (x_n, y_m)\}$. The x_i is the vectors whose components are the values of embedded sensors $\{S_1, \dots, S_n\}$ such as stove-sensor, refrigerator-sensor, and door-sensor. The values of “y” are drawn from a discrete set of classes $\{c_1, \dots, c_m\}$ such as a “leave home”, “read”, and “sleep”. Furthermore, excessive information such as multiple header lines is also removed from the sensor logs and annotation files.

5.2 Classifiers for activity recognition

In this section, we briefly introduce the applied classifier⁴ for the domain of activity recognition. We applied the various types of state-of-the-art techniques including probabilistic and statistical approaches (i.e., ANN, HMM, CRF, and SVM) on the analyzed datasets. However, these methods cannot handle the parallel activities due insufficient training instances of parallel activities in the dataset. Brief description of each classifier with preferred settings for our experiments is given as the following.

Artificial Neural Networks (ANN): It is an information processing network of artificial neurons connected with each other through weighted links. In activity recognition, multilayer neural network with the back propagation learning algorithm is utilized to recognize the human activities [8]. The structure of the network, number of hidden layers, and number of neuron in each layer affects the learning of different activities. The activation of the neurons in the network depends on the activation function [8]. We train multilayer neural network through the back propagation learning method and weights are updated by the following equation:

$$\Delta w_{ki} = -c \left[-2 \sum_j \{ (y_{j(\text{desired})} - y_{j(\text{actual})}) f'(\text{act}_j) w_{ij} \} f'(\text{act})_i x_k \right] \quad (1)$$

where Δw is the weights adjustment of the network links and c is the learning rate of neural network such that $(0 < c \leq 1)$. We set the value of $c = 0.1$ to learn the performed activities more rapidly. In our network, we used one hidden layer, twenty neurons, tangent sigmoid function as an activation function as given below:

$$\varphi(x) = \tanh\left(\frac{x}{2}\right) = \frac{1 - \exp(-x)}{1 + \exp(-x)} \quad (2)$$

⁴The scope of variables used to describe the functionality of the classifiers is limited to this section.

Learning of the network is limited to maximum 1000 epochs. The multilayer neural network can be seen as an intuitive representation of a multi layer activity recognition system. The number of correctly classified activities depends on the number of training instances during the learning phase.

Hidden Markov Model (HMM): It is a generative probabilistic graph model that is based on the Markov chains process [9]. For activity recognition, HMM is based on the number of states and their transition weight parameters. In our experimental setting, parameters are learned thorough observation and following parameters are required to train the model:

$$\lambda = \{A, B, \pi\} \tag{3}$$

where λ is graphical model for activity recognition, A is a transition probability matrix, B represents the output symbol probability matrix, and π is the initial state probability [9]. We used Baum–Welch algorithm to determine the states and transition probabilities during training of HMM. The i th classification of an activity is given as

$$\lambda_i = \{A_i, B_i, \pi_i\}, \quad i = 1, \dots, N \tag{4}$$

Conditional Random Fields (CRF): It is a discriminative probabilistic graph model for labeling the sequences. The structure of the CRF is similar to HMM for activity recognition but learning mechanism is different due to absence of the hidden states [10]. In our experimental settings, the conditional probabilities of activity labels with respect to sensor observations are calculated as follows:

$$p(y_{1:T} | x_{1:T}) = \frac{1}{Z(x_{1:T}, w)} \exp \left\{ \sum_{j=1}^{N_f} w_j F_j(x_{1:T}, Y_{1:T}) \right\} \tag{5}$$

In Eq. (5), Z denotes normalized factor and $F_j(x_{1:T}, Y_{1:T})$ is a feature function. To make the inference in the model, we compute the most likely activity sequence as follows:

$$y_{1:T}^* = \operatorname{argmax}_{y_{1:T}'} p(y_{1:T}' | x_{1:T}, w) \tag{6}$$

Support Vector Machine (SVM): SVM is statistical learning method to classify the data through determination of a set of support vectors and minimization of the average error [11]. It can provide a good generalization performance due to rich theoretical bases and transferring the problem to a high dimensional feature space. In our experimental setting, for a given training set of sensors value and activity pairs, the binary linear classification problem requires the following maximum optimization model using the Lagrangian multiplier techniques and Kernel functions as

$$\text{Maximize (w.r.t } \alpha) \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=0}^n \sum_{j=1}^n \alpha_i y_i \alpha_j y_j K(x_i, x_j) \tag{7}$$

$$\text{Subject to: } \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \tag{8}$$

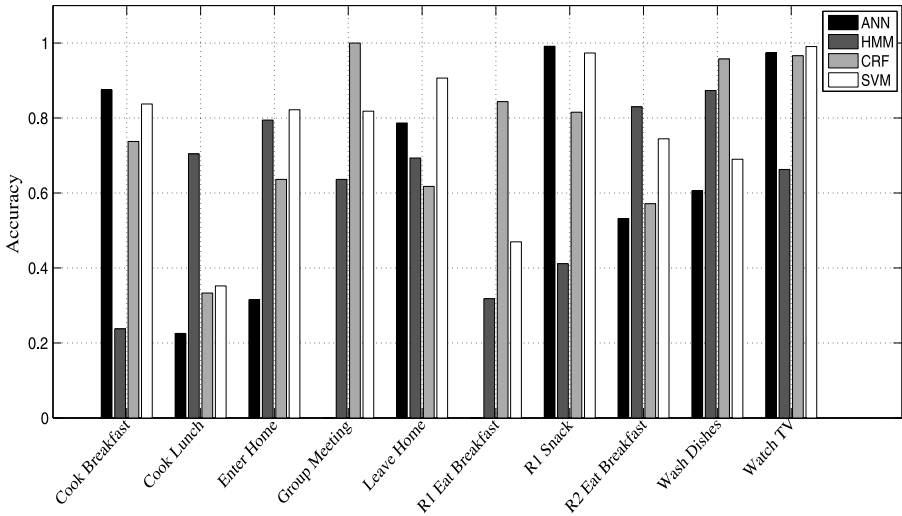


Fig. 1 Tulum2009 activity based accuracy of classifiers

where K is the kernel function that satisfies $K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j)$. We used the radial basis function (RBF) for recognizing the activities.

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{(2\sigma^2)}\right) \quad (9)$$

Activity recognition is a multiclass problem so we adopt the “one-versus-one” method to classify the different activities. Classification of the final activity class is based on the voting mechanism and maximum vote of a class determined the activity label.

5.3 Results of variations in classifiers performances and discussion

In this section, results for the effects of smart home dataset dimensions on classifiers performance are presented for recognizing the daily life activities. The approach has been implemented in MATLAB 7.6. The configuration of the computer is Intel Pentium(R) Dual-Core 2.5 GHz with 3 GB of memory and Microsoft Window 7. The standard metric of accuracy is used as performance measure by using the values of the confusion matrix [31] and computed as

$$Accuracy = \frac{\sum_{i=1}^Q TP_i}{T} \quad (10)$$

where Q is the number of performed activities, TP is the number of true positives, and “ T ” is total number of activities. We evaluated Tulum2009 and Twor2009 datasets from CASAS smart home. From ISL and House_n smart homes experiments are performed on House A and Subject 1 datasets, respectively. We split the datasets using the “leave one day out” approach; therefore, the sensor readings of one day are

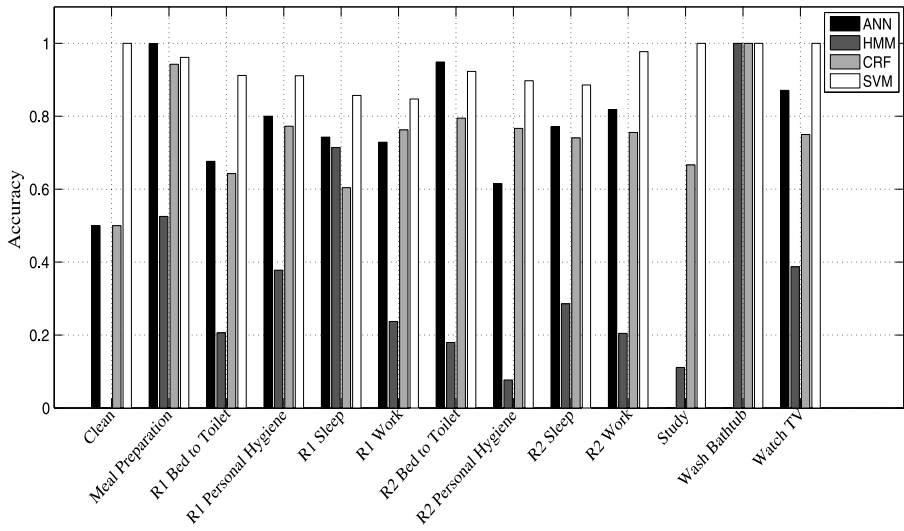


Fig. 2 Twor2009 activity based accuracy of classifiers

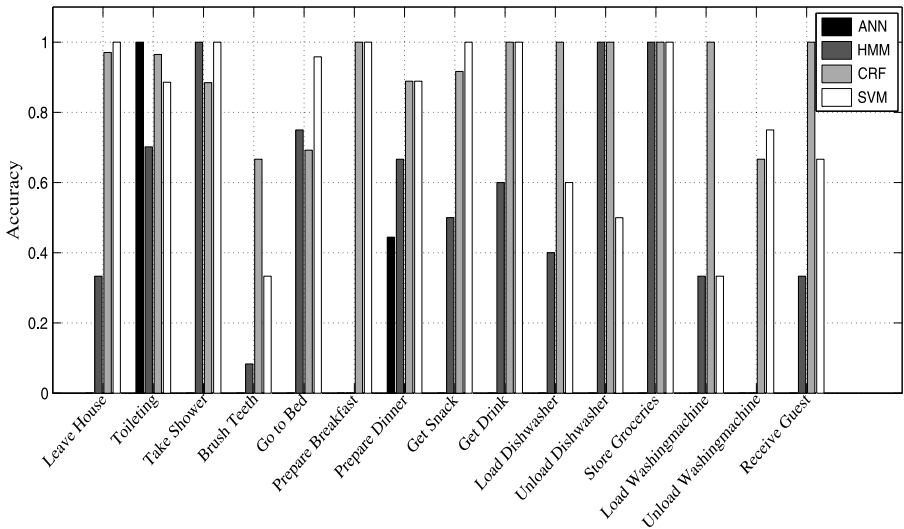


Fig. 3 House A activity based accuracy of classifiers

used for testing and the remaining days for training. Figures 1, 2, 3, and 4 show the experimental results for the Tulum2009, Twor2009, House A, and Subject 1 datasets characteristics, respectively. In each dataset, for each activity, accuracies of ANN, HMM, CRF, and SVM are illustrated. In the following paragraphs, we will discuss how each technique has the performance variations over the different datasets by using the proposed data dimensions.

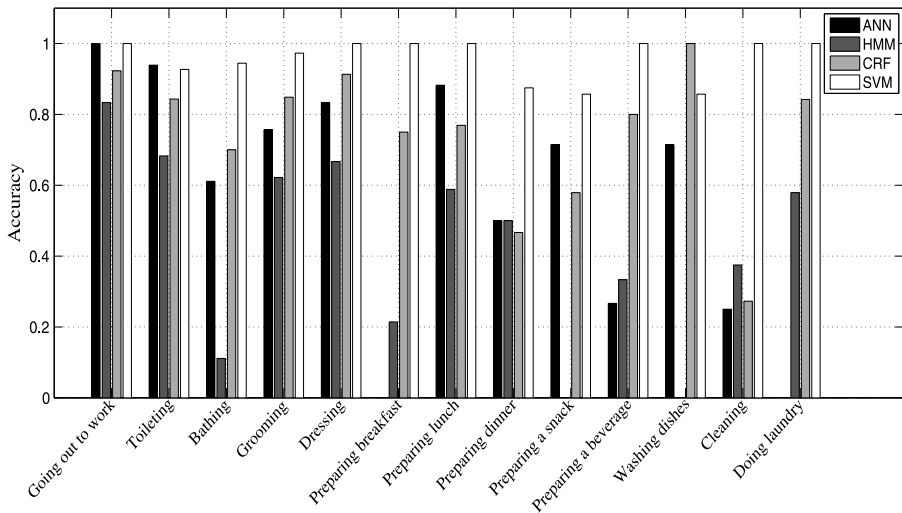


Fig. 4 Subject 1 activity based accuracy of classifiers

In our experiments, ANN shows high diversity in its performance. In order to learn the activities correctly, it requires a sufficient amount of training instances so hidden neurons can learn precisely from training data. Due to this reason, ANN performs better on the set of activities whose training instances are high in the dataset while its performance is insignificant for the recognition of those activities whose training examples are few in the dataset. Overall performance of ANN on *Tulum2009* is 81.09 %, and it correctly classified “R1 snack” activity; however, it could not recognize the “group meeting” activity. The training instances for these activities are 491 and 11, respectively, as shown in Table 3 that affects the ANN classification process. In the case of *Twor2009*, training instances for “meal preparation” are 118, and it outperforms all other classifiers on the identification of this activity. While on the same dataset, it could not classify the “study” activity due to less training instances. For *House A*, the overall performance of ANN is a low 41.11 % as compared to other datasets. However, it is the only classifier that 100 % classifies the “toileting” activity as its training instances (i.e., 114 samples) are very high as compare to other activities as shown in Table 6. Similarly for *Subject 1*, training instances of “toileting” activity are more (i.e., 82 samples) as shown in Table 8, and ANN performance is better than other classifiers for the recognition of this activity. Although the overall accuracy of ANN varies from dataset to dataset, however, its better performance is consistently dependent on the large number of training instances in the dataset for a particular activity.

In the case of HMM, the number of deployed sensors affects the activity class distribution by observing their variation during the performed activities. For instance, HMM achieves 57.83 % accurate results in case of *House A* due to a small number of deployed sensors (i.e., 14 sensors) as shown in Table 5. It correctly classified “take shower”, “unload dishwasher”, and “store groceries”. In *Tulum2009*, it outperforms other classifier for “cook lunch” and “R2 eat breakfast” activities with overall accuracy 56.84 %, the number of deployed sensors is 20 in this case as shown in Table 2.

Table 9 Overall classifiers accuracy

Dataset	Classifier			
	ANN	HMM	CRF	SVM
<i>Tulum2009</i>	0.8109	0.5684	0.8374	0.8889
<i>Twor2009</i>	0.7983	0.3421	0.7780	0.9307
<i>Hosue A</i>	0.4111	0.5783	0.9230	0.8919
<i>Subject 1</i>	0.6836	0.5200	0.7745	0.9563

For *Subject 1* and *Twor2009*, accuracy of the HMM is not significant; the number of deployed sensors in these smart homes is 28 and 71 as shown in Tables 7 and 2, respectively. The large number of missed classified activities is the result of HMM distributions modeling as they are observed in the dataset.

The performance of CRF is affected by a set of data dimensions that are mentioned in Table 1. Its performance does not depend only on a single data dimension like sensor count or activity occurrences; however, its internal processing is based on conditioning of a set of data characteristics. CRF outperforms all classifiers in case of *Tulum2009* for “group meeting”, “R1 eat breakfast”, and “wash dishes”. However, other classifiers are better for “cook lunch”, “enter home” and “leave home”, “R1 snack” and “R2 eat breakfast”. In the case of *House A*, CRF is superior for “brush teeth”, “load washing machine”, and “receive guest”. For *Subject 1*, its performance is high only for “washing dishes”, however, for “cook lunch”, “enter home”, and “leave home”, “R1 snack” and “R2 eat breakfast” other classifiers performed better. In the case of *Twor2009*, CRF shows low performance for all activities except “wash bathtub”.

Support Vector Machine (SVM) efficiently identified activities in the case of *Subject 1* and *Twor2009*; it outperforms all classifiers in these datasets except for “washing dishes” and “R2 bed to toilet”, respectively. The performance of SVM is high if performed activities in the dataset are highly discriminative, however, it is hard for SVM to differentiate between activities that are closely correlated in various data dimensions. Due to this reason in *House A* and *Tulum2009*, for some activities other classifiers are better than SVM as discussed in the above paragraphs. For example, in *Tulum2009*, it confused “R1 eat breakfast” with “R1 snack”. Both activities are very similar to each other, the second most confused activity is “cook lunch”. SVM performance is affected if the performed activities are closely interrelated in respect to data dimensions.

The overall comparison results of different classifiers are presented in Table 9. It specifies the overall accuracy associated with each of the dataset over the four learning techniques. ANN performs better for *Tulum2009* due to the large number of training examples for each activity in the dataset that finally results in its overall high performance. The performance of HMM is better in the case of *House A*, where the number of deployed sensors are small as compared to other datasets that facilitates it to maintain clear transitions between adjacent activities. CRF outperforms other classifiers for *House A* with 90.30 % accurate results while the performance of SVM is superior for remaining datasets. The highest accuracy of 95.63 % is achieved in the case of *Subject 1* as the performed activities are not correlated with respect to data

dimensions. The above results and statistics clearly show that the proposed dataset dimensions highly affect the classifiers' individual class level assignments, and thus their overall performances.

6 Conclusion

Technological advances in pervasive sensing play important roles in leveraging the use of smart home datasets for different applications of ambient assisted living. The underlying logic of various classifications methods depends on diverse data characteristics so the dataset is significantly important for their evaluation. One of the main challenges is to compute and analyze data dimensions and variations. So, accurate data analysis is necessary to understand and reuse the datasets. Nevertheless, the information provided with smart home datasets is not always sufficient to explore the possible dimensions of analysis. We developed a framework to analyze the smart home datasets on predefined data characteristics. It enables the researchers to compute data dimensions that cover variations in time, activities, sensors, and inhabitants. To evaluate its effectiveness, we showed the influence of dataset characteristics on the performance of the classifiers (i.e., ANN, HMM, CRF, and SVM). We applied each classifier on four different datasets from three smart home projects. The result shows that usually classifiers perform complementary to each other based on dataset characteristic for the recognition of activities. Therefore, it is imperative to choose an appropriate dataset for a particular algorithm. Hence, the impact of the proposed framework is to provide a valuable and better understanding of data for the domain of activity recognition.

In the future, we are planning to extend this framework as a recommender system, such that a user can provide information about dataset requirements and the proposed framework will recommend a list of candidate datasets along most suitable classifiers from its data repository. The proposed framework will check the availability of datasets by analyzing their compatibility with provided requirements.

Acknowledgements This research was supported by the MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-(H0301-13-2001)).

References

1. Kasteren T, Noulas A, Englebienne G, Krose B (2008) Accurate activity recognition in a home setting. In: *Proceeding of 10th international conference on ubiquitous computing*, pp 1–9
2. Singla G, Cook DJ, Maureen S (2010) Recognizing independent and joint activities among multiple residents in smart environments. *J Ambient Intell Humaniz Comput* 1(1):57–63
3. Silas S, Ezra K, Rajsingh EB (2012) A novel fault tolerant service selection framework for pervasive computing. *Human-Centric Comput Inf Sci*. doi:[10.1186/2192-1962-1962-2-5](https://doi.org/10.1186/2192-1962-1962-2-5)
4. Cuntoor NP, Yegnanarayana B, Chellappa R (2008) Activity modeling using event probability sequences. *IEEE Trans Image Process* 17(4):594–607
5. Parisa R, Diane JC, Lawrence BH, Maureen S (2011) Discovering activities to recognize track in a smart environment. *IEEE Trans Knowl Data Eng* 23(4):527–539
6. Cook DJ, Crandall A, Singla G, Thomas B (2010) Detection of social interaction in smart spaces. *Cybern Syst* 2(41):90–104

7. Helal S, Kim E, Hossain S (2010) Scalable approaches to activity recognition research. In: Proceedings of the workshop of how to do good activity recognition research? Experimental methodologies, evaluation metrics, and reproducibility issues, pp 450–453
8. Kasteren TLM, van Englebienne G, Krose BJA (2010) An activity monitoring system for elderly care using generative and discriminative models. *Pers Ubiquitous Comput* 14(6):489–498
9. Krose BJA, Kasteren TLM, Gibson CHS, Dool T (2008) CARE: context awareness in residences for elderly. In: Proceeding of 6th international conference of the international society for gerontechnology, pp 101–105
10. Fleury A, Vacher M, Noury N (2010) SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *IEEE Trans Inf Technol Biomed* 14(2):274–283
11. Cook D (2012) CASAS smart home project. [Online]. <http://www.ailab.wsu.edu/casas/> [June 10, 2012]
12. Larson K (2012) House_n. [Online]. http://architecture.mit.edu/house_n/ [June 10, 2012]
13. Sarkar P, Saha A (2012) Security enhanced communication in wireless sensor networks using Reed–Muller codes and partially balanced incomplete block designs. *J Converg* 2(1):23–30
14. Huang C, Cheng R, Chen S, Li C (2010) Enhancing network availability by tolerance control in multi-sink wireless sensor networks. *J Converg* 1(1):15–22
15. Rashidi P, Cook DJ (2011) Activity knowledge transfer in smart environments. *Pervasive Mob Comput* 7(3):331–343
16. van Kasteren TLM, Englebienne G, Krose BJA (2010) Activity Recognition Using Semi-Markov models on real world smart home datasets. *Ambient Intell Smart Environ* pp 311–325
17. Zheng H, Wang H, Black N (2008) Human activity detection in smart home environment with self-adaptive neural networks. In: Proceedings of the IEEE international conference on networking, sensing and control, pp 1505–1510
18. Chen C, Das B, Cook DJ (2010) A data mining framework for activity recognition in smart environments. In: Proceeding of the international conference on intelligent environments, pp 80–83
19. Cook D, Holder L (2011) Sensor selection to support practical use of health-monitoring smart environments. *Data Mining Knowl Discov* 1(4):339–351
20. Gregory DA, Anind KD, Peter JB, Nigel D, Mark S, Pete S (1999) Towards a better understanding of context and context-awareness. In: Proceedings of the 1st international symposium on handheld and ubiquitous computing (*HUC'99*), pp 304–307
21. Liming C, Jesse H, Chris DN, Diane JC, Zhiwen Y (2012) Sensor-based activity recognition: a survey. *IEEE Trans Syst Man Cybern* 99:1–19
22. Krishnan NC, Panchanathan S (2008) Analysis of low resolution accelerometer data for continuous human activity recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3337–3340
23. Teraoka T (2012) Organization and exploration of heterogeneous personal data collected in daily life. *Human-Centric Comput Inf Sci*. doi:[10.1186/2192-1962-2-1](https://doi.org/10.1186/2192-1962-2-1)
24. Singh RR, Banerjee R (2010) Multi-parametric analysis of sensory data collected from automotive drivers for building a safety-critical wearable computing system. In: IEEE international conference on computer engineering and technology (ICCET), pp 355–360
25. Busemann C, Behrensen S, Nicklas D (2010) ALYZE—an analysis tool for wireless sensor networks with a direct physical interaction interface. In: IEEE international conference on pervasive computing and communications workshops, pp 823–825
26. Liming C, Jesse H, Chris DN, Diane JC, Zhiwen Y (2012) Sensor-based activity recognition: a survey. *IEEE Trans Syst Man Cybern* 99:1–19
27. Du Y, Chen F, Xu W, Li Y (2011) Behavior metrics for multiple residents in a smart environment. PhD dissertation, Washington State University
28. Kasteren T, Englebienne G, Krose B (2010) Transferring knowledge of activity recognition across sensor networks. In: Proceeding of 8th international conference on pervasive computing, pp 283–300
29. Tapia EM (2003) Activity recognition in the home setting using simple and ubiquitous sensors. M.S. Thesis Media Arts and Sciences, Massachusetts Institute of Technology
30. Mitchell T (1997) *Machine learning*. McGraw Hill, Columbus
31. van Kasteren TLM, Alemdar H, Ersoy C (2011) Effective performance metrics for evaluating. In: Proceeding of second workshop on context-systems design, evaluation and optimisation