

ARAS Human Activity Datasets in Multiple Homes with Multiple Residents

Hande Alemdar*, Halil Ertan*, Ozlem Durmaz Incel[†] and Cem Ersoy*

*NETLAB, Department of Computer Engineering
Bogazici University, Istanbul/Turkey

Email: {hande.alemdar, halil.ertan, ersoy}@boun.edu.tr

[†]Peralab, Department of Computer Engineering
Galatasaray University, Istanbul/Turkey,

Email: odince1@gsu.edu.tr

Abstract—The real world human activity datasets are of great importance in development of novel machine learning methods for automatic recognition of human activities in smart environments. In this study, we present the details of ARAS (Activity Recognition with Ambient Sensing) human activity recognition datasets that are collected from two real houses with multiple residents during two months. The datasets contain the ground truth labels for 27 different activities. Each house was equipped with 20 binary sensors of different types that communicate wirelessly using the ZigBee protocol. A full month of information which contains the sensor data and the activity labels for both residents was gathered from each house, resulting in a total of two months data. In the paper, particularly, we explain the details of sensor selection, targeted activities, deployment of the sensors and the characteristics of the collected data and provide the results of our preliminary experiments on the datasets.

I. INTRODUCTION

Daily human activities and long term monitoring of those activities are important indicators of health status. The changes and the patterns in these changes help diagnosing several diseases at early stages and let the people take the necessary precautions. In order to make the long term health monitoring systems sustainable, we need smart environments in which the human activities hence the human behavior are recognized automatically. Advanced machine learning methods are used for inferring human activities using the data collected from the smart environment but those methods require annotated datasets to be trained on. The quality of the dataset in terms of accuracy and the total number of data points greatly affects the prediction power of the machine learning model. Therefore, annotated datasets are required to develop novel machine learning models with higher prediction accuracies. However, recording and annotating these datasets are costly since they require time and human effort. Hence, the publicly available annotated datasets are of significant importance to the research community and this demand tends to increase.

The datasets recorded in real world conditions rather than laboratory settings are even more difficult to obtain and also more valuable since they reflect the natural human behavior. When the subjects in laboratories are replaced by the real residents living in their own houses, the value of the information increases and the world wide deployments of smart health monitoring systems will become possible. For this purpose, we introduce publicly available ARAS (Activity Recognition with Ambient Sensing) datasets that are collected from two real houses with multiple residents. The datasets are available

for download at <http://www.cmpe.boun.edu.tr/aras/>.

The rest of the paper is organized as follows. In Section II, we summarize and compare several previous datasets to ARAS dataset. Section III presents the design details of the ambient sensing system and the data collection phase together with our preliminary results on the datasets. Section IV provides a discussion on the challenges and key lessons learned during the study. Finally, in Section V, we provide the conclusions.

II. RELATED STUDIES

The research efforts on smart homes mainly focus on two different fields. In the first group, there are studies where hundreds of sensors and sensor equipped home appliances are deployed in a smart laboratory houses [1], [2], [3]. Those studies generally focus on human interactions with the future smart environments and do not necessarily have a healthcare focus. The second group of studies focus on human activity detection for health status monitoring [4], [5], [6], [7]. Our study also focuses on making today's houses smart enough to provide long term health monitoring for not only people who live alone but also with a spouse or a flat mate. Also, we believe that the studies in that area can be improved by the community with the use of publicly available datasets for human activities gathered from real environments. In Table I, we summarize the main attributes of several publicly available datasets. Most of the earlier studies consider a single resident situation. While collecting the ARAS dataset we relaxed that assumption and collected the data from multi-resident homes. Another important feature of ARAS datasets is that they contain a larger variety of human activities and a larger number of activity occurrences, as shown in Table I.

III. ARAS DATASETS

In this section, we present the details of ARAS datasets, outlining the sensor hardware used, targeted activities and the deployment of the system in the target houses.

A. System Overview

The basic properties of the houses, demographic information about the residents and the details about the infrastructures are provided in Table II. We used the Arduino platform to build our ambient sensors. Arduino is an open source, cost and power efficient hardware platform which enables quick prototyping and flexibility. The sensors were chosen to communicate wirelessly using low power ZigBee protocol. This enabled easy

TABLE I: List of similar publicly available annotated datasets

| | # of Houses | Mult. Residents | Duration | # of Sensors | # of Activities | Activity Occurrences |
|-----------|-------------|-----------------|-------------------------|--------------|-------------------|----------------------|
| ARAS | 2 | Yes | 2 months | 20 | 27 | 1023 - 2177 |
| Casas [4] | 7 | Yes | 2-8 months | 20-86 | 11 | 37-1513 |
| UvA[5] | 3 | No | 2 months | 14-21 | 10 - 16 | 200-344 |
| Domus[6] | 1 | No | 1.5 hours (24 subjects) | 78 | 0 (user feelings) | NA |
| MIT [7] | 2 | No | 2 weeks | 77-84 | 9 - 13 | 176- 278 |

deployment of the sensors since completely wiring an entire house would be extremely difficult and impractical. The the low-power ZigBee protocol enabled the sensor batteries to last longer. Also, the sensors were programmed to send information only when an event is detected, for instance, the contact sensor on the door sends data only when someone opens it. When there are no events, the sensors were put to sleep mode in order to save energy. All sensor data was accumulated in a central processing unit where the synchronization between the labels and the sensor data was handled.

Due to the differences between the houses, the topologies of the Wireless Ambient Sensor Networks (WASN) were different. The layouts of the houses together with the locations of the sensor nodes are provided in Fig 1. In House A, there were two Personal Area Networks (PANs) in order to cover the entire house, whereas in House B, using only one PAN was sufficient. Although House A is smaller, due to the obstructions in the environment, we needed to setup the second PAN and synchronize the information flow which is shown in Fig 2 for House A.

B. Sensor Selection and Deployment

One of the most important issues we considered while selecting the sensors is privacy. For this reason, we did not use any sort of cameras or video recorders in the data collection. Moreover, we did not place any identification tags on the residents mostly because not to reveal their identities and also it is not realistic to assume that the residents will carry identification tags on them all the time. Therefore, we preferred to use ambient binary sensors which are unobtrusive and user-friendly compared to the alternatives.

The binary sensors used in the datasets are force sensitive resistors (FSR), pressure mats, contact sensors, proximity sensors, sonar distance sensors, photocells, temperature sensors and infrared (IR) receivers. A total of 20 sensors of 7 different types for House A and 6 different types for House B were used. Some of the deployed sensors in House B are presented in Fig 3. We selected the most suitable sensors which enable us to understand the targeted activities in a more robust and efficient manner. For this purpose, we experimented with several types of sensors before the actual data collection phase.

TABLE II: Properties of the Datasets

| | House A | House B |
|----------------------|---|--|
| # of PANs | 2 | 1 |
| # of Ambient Sensors | 20 of 7 different types | 20 of 6 different types |
| Size of the House | 50m ² | 90m ² |
| House Information | One bedroom, one living room, one kitchen, one bathroom | 2 bedrooms, one living room, one kitchen, one bathroom |
| Residents | 2 males both aged 25 | Married couple, age average 34 |
| Duration | 30 full days | 30 full days |
| # of Activities | 27 | 27 |

We considered the sensor and activity matching while selecting the sensors and the locations for deployment. Consider for example the cooking activity. When a person is cooking, it is expected that several actions occur at the same time or in succession during the whole activity. Some of the expected actions are opening of the refrigerator, taking spoons, forks and knives from the drawer, heating the oven, and move around the kitchen. Although, the order can change according to the person performing the activity, the actions are the same most of the time. Therefore, while deploying the sensors, we aimed to recognize the several actions related with the activities. The aimed actions and the approximate locations of the used sensors are provided in Table III.

C. Collected Data

During the data collection phase the residents were not required to follow a specific scenario. We strongly asked them to behave naturally. Since the ambient sensors used are unobtrusive, they generally do not interfere with the daily activities of the residents. The only interaction with the sensors happens while changing the batteries which happened 2-8 times during 30 days. The most challenging and tedious part for the residents is the annotation. In our case, we have a greater variety of activities making the process even more challenging. The ground truth activity labels were obtained from the residents themselves using a simple graphical user interface (GUI). In

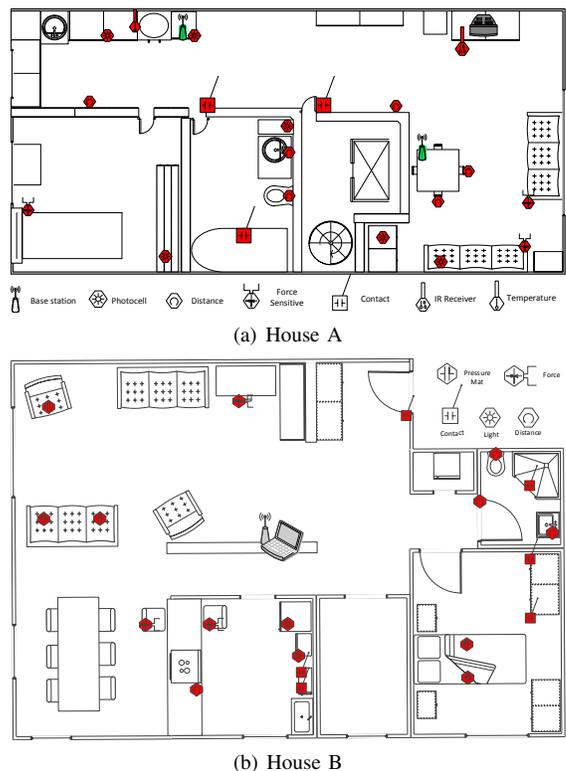


Fig. 1: Layouts of the Houses

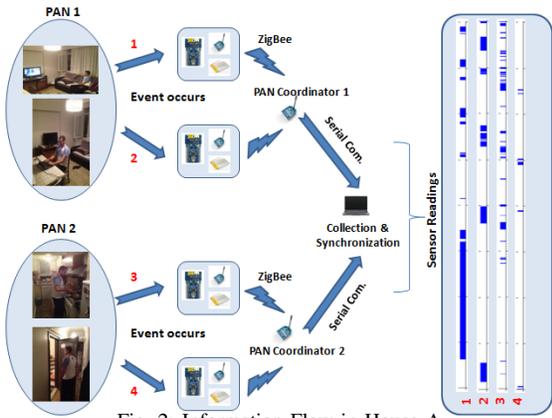


Fig. 2: Information Flow in House A

order to ease up the annotation process, several instances of GUI applications were placed to the most convenient places in the houses. In this way, the annotation process were handled without interrupting the actual activities. Overall, we could manage to obtain between 60-100 activity labels for each day. 27 different activities were concerned during data collection including every day activities like sleeping, brushing teeth, watching TV, toileting, preparing a meal, eating, etc. and also activities that are not performed every day such as hanging out laundry, having a guest, doing cleaning, and having a nap. The rare activities are also important when inferring the health status of the residents. Therefore, we gathered information about those activities unlike previous studies.

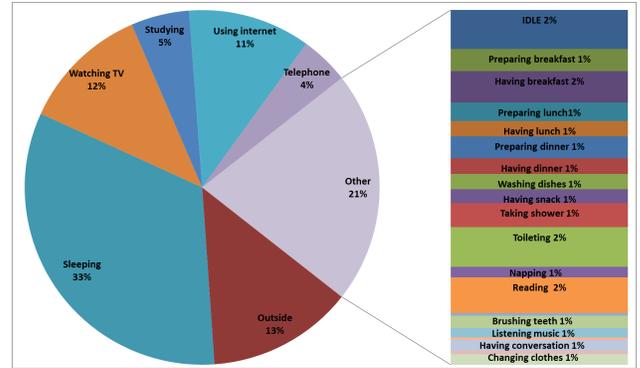
We developed a software that handles the synchronization of the sensor values and activity labels, converts sensor values to binary format and merges the sensor data and activity labels from different instances. The residents labeled their activities by using the simple interface at the beginning of the activity. It was assumed that the resident performed the same activity until the next label is provided. Previous studies use methods like keeping a diary or using Bluetooth headsets for annotation. Our method is more accurate than manually keeping a diary and more user friendly than wearing a headset all the time. The total number of activity occurrences in House A and House B are 2177 and 1023 respectively. The total number of sensor readings are around 26 million. The sensors are sampled at 10 Hz (except the IR sensor which is sampled at 100 Hz) in order to make sure that no sensor events are missing. In order to make the handling easier without losing any information, the raw data was formatted in seconds resolution after that. In total, there were 86400 data points for each day comprising



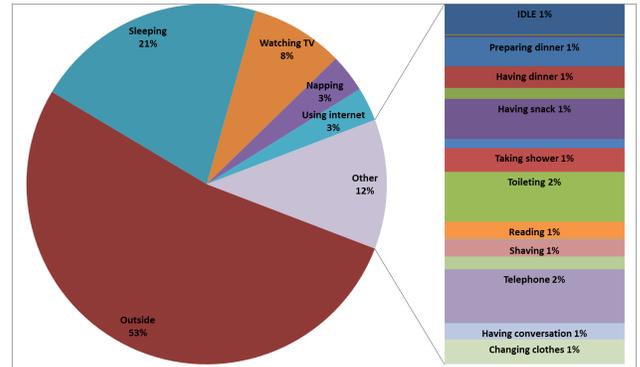
Fig. 3: Sample Deployments of Ambient Sensors in House B

TABLE III: Targeted Actions and The Locations of The Ambient Sensors

| Ambient Sensors | Actions | Location |
|------------------------------|---|--|
| Force Sensor or Pressure Mat | Sleeping, sitting, napping | Under the beds and the couches |
| Photocell | Opening the drawers and the wardrobes | In the drawers, the wardrobes and the refrigerator |
| Contact Sensors | Opening and closing of the doors, cupboards | On door frames, shower cabin, cupboards |
| Proximity Sensors | Detecting close distance objects | On the chairs, on the closets and on the taps |
| Sonar Distance Sensors | Detecting presence | On the walls, door frames |
| Temperature Sensors | Cooking | Near the oven in the kitchen |
| Infrared Receiver | Watching TV | Near the TV |



(a) Resident 1



(b) Resident 2

Fig. 4: Activity Duration Distributions for 30 days for House A

of the time stamp, 20 sensor values, either 0 or 1 for each second, and the activity labels for both residents. Since human activities cannot last shorter than 1 second but the sensor events can, we took this approach to make sure that no sensor events or activities are missing.

In order to give insights about the datasets, we provide the duration distributions of activities for both residents in House A during 30 days in Fig 4. The figure reflects the different life styles of the residents even though they share the same house. The first resident spends more time in the house on the other hand the second resident has a day job and is out of the house for longer durations. Also, the second resident sleeps significantly less than the first resident which can lead to health issues in the long term.

D. Preliminary Experiments

In order to infer activities from the data, we use probabilistic machine learning methods and perform preliminary

experiments on the datasets. We consider the frequent activities namely sleeping, eating, personal hygiene, going out and relaxing. All the other activities are grouped into the ‘other’ class, yielding a total of six activities. Since there are two residents in each house, we use combinations of these activities as class labels. For example, we have a class label for the first resident is sleeping while the second resident is eating, and we have a different class label for the opposite case. In this way, we manage to infer which activity is being performed by whom. We have a total of 25 combinations of the activities for House A and 20 combinations for House B. We use a hidden Markov model (HMM) for modeling the activities and sensor data since it is well suited for sequential data. The observations correspond to the sensor readings and the hidden states correspond to the activities. We use leave-one-out cross validation in the experiments and report the average accuracy (the percentage of correctly classified labels) of 30 days. According to the results, we obtain an average accuracy of 61.5% for House A (with min=46.3% and max=88.4%) and 76.2% (with min=31.1% and max=96.7%) for House B.

IV. CHALLENGES AND KEY LESSONS

In this section, we summarize some of the challenges we faced and key lessons learned during the study.

Sensor Selection and Deployment: Finding the correct sensors and suitable locations are of great importance for identifying the activities. In order to make decisions, we conducted several interviews with the residents about their habits at home which helped us in choosing sensor types and deployment locations. Moreover, we experimented with different types of sensors for several days before deciding on the final set of sensors. As an example, we experimented with a humidity sensor to detect having a shower activity. However, it was not possible to grasp the exact duration of the activity since the sensor value does not change immediately with the changing humidity value in the bathroom, instead it continues to give high values after the activity is completed. Therefore, we installed a contact sensor on the door of the shower cabin to detect whether the door is open or not.

Sensor Data Quality: Data quality depends on several factors such as the data losses due to the wireless transmission, exhausted batteries, or sensor hardware failures. In order to minimize these failures, we took some precautions during the study. To begin with, although the transmission range of the wireless modules are given as 30m in the data sheets, due to the obstructions in the layout of the first house, we needed to deploy two different networks in order to cover the entire house. Moreover, the ZigBee network uses the same frequency range that of WiFi networks therefore, we first gathered information about the WiFi networks around the house and decided on the best channel for communication in order to minimize any possible interference. In order to increase the lifetime of the batteries, we programmed the hardware nodes such that they continuously sample their sensors and only when an event is detected they use the wireless channel to send the data, since the wireless transmission is the main energy consuming task. In order to capture all the sensor value changes, the sampling rate was configured to 10Hz except for the IR sensor which required 100Hz sampling rate. This scheme enabled us to detect all the sensor events and use the wireless link only when a real sensor event is detected. Finally, in order to prevent hardware failures, we experimented with

several sensors and favoured the use of more robust ones in the actual deployment.

Annotation Label Accuracy: In terms of the annotation accuracy, there is a tradeoff between the accuracy of the labels and the natural behavior. Under laboratory settings or when predetermined scenarios are used, the accuracy of the labels can be increased yet the knowledge obtained is hardly applicable in real world deployments. In natural settings, due to the complex nature of human activities, it is extremely difficult to come up with a precise label describing the activity. There may be concurrent activities such as listening to music while washing the dishes and also there may be interrupting activities such as answering the phone while watching TV. In order to come up with a balanced solution, we provided the residents with a very detailed list of activities and asked them to label the most dominant activity they are engaged in.

V. CONCLUSION

In this study, we presented the details of publicly available, annotated, real world human activity datasets we collected from two real houses with multiple residents during a full month for each house. The annotated datasets are extremely essential for developing novel machine learning models for human activity recognition which will pave the way to continuous and long-term healthcare monitoring systems that will be deployed in future smart environments. Since collecting this kind of human activity data under realistic conditions rather than laboratory settings is extremely costly and difficult, the community can greatly benefit from the publicly available datasets. In that respect, ARAS dataset is among those that contain the largest number of different activities, with a diversity in residents’ demographic information and house layouts. The results of our preliminary experiments indicate a potential increase in recognition performance when more sophisticated machine learning methods are used.

ACKNOWLEDGMENT

This work is supported partially by the Bogazici University Research Fund under the grant numbers 11A01P7 and 11A01D4, and by the Galatasaray University Research Fund under the grant number 13.401.002.

REFERENCES

- [1] J. A. Kientz, S. N. Patel, B. Jones, E. Price, E. D. Mynatt, and G. D. Abowd, “The Georgia Tech Aware Home,” in *HI '08 Extended Abstracts on Human Factors in Computing Systems*, 2008.
- [2] S. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen, “The Gator Tech Smart House: A Programmable Pervasive Space,” *Computer*, vol. 38, no. 3, pp. 50–60, 2005.
- [3] S. S. Intille, K. Larson, J. S. Beaudin, J. Nawyn, E. M. Tapia, and P. Kaushik, “A living laboratory for the design and evaluation of ubiquitous computing technologies,” in *In Extended Abstracts of the 2005 Conference on Human Factors in Computing Systems*. ACM Press, 2005, pp. 1941–1944.
- [4] D. J. Cook, “Learning setting-generalized activity models for smart spaces,” *IEEE Intelligent Systems*, vol. 27, no. 1, pp. 32–38, 2012.
- [5] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, “Accurate activity recognition in a home setting,” in *Proceedings of the 10th international conference on Ubiquitous computing*, ser. UbiComp '08. New York, NY, USA: ACM, 2008, pp. 1–9.
- [6] M. Gallissot, J. Caelen, N. Bonnefond, B. Meillon, and S. Pons, “Using the multicom domus dataset,” LIG, Grenoble, France, Research Report RR-LIG-020, 2011.
- [7] E. M. Tapia, S. S. Intille, and K. Larson, “Activity recognition in the home using simple and ubiquitous sensors,” in *Proceedings of International Conference on Pervasive Computing*, 2004, pp. 158–175.