A Long-Term Evaluation of Sensing Modalities for Activity Recognition

Beth Logan[1], Jennifer Healey[1], Matthai Philipose[2], Emmanuel Munguia Tapia[3], and Stephen Intille[3]

[1] Intel Digital Health, One Cambridge Center 11FL,Cambridge MA 02139, USA
Beth.Logan@intel.com
http://www.intel.com/healthcare/
[2] Intel Research Seattle, 1100 NE 45th Street, Seattle WA 98105, USA
[3] MIT House_n, One Cambridge Center 4FL, Cambridge MA 02142, USA

Abstract. We study activity recognition using 104 hours of annotated data collected from a person living in an instrumented home. The home contained over 900 sensor inputs, including wired reed switches, current and water flow inputs, object and person motion detectors, and RFID tags. Our aim was to compare different sensor modalities on data that approached "real world" conditions, where the subject and annotator were unaffiliated with the authors. We found that 10 infra-red motion detectors outperformed the other sensors on many of the activities studied, especially those that were typically performed in the same location. However, several activities, in particular "eating" and "reading" were difficult to detect, and we lacked data to study many fine-grained activities. We characterize a number of issues important for designing activity detection systems that may not have been as evident in prior work when data was collected under more controlled conditions.

1 Introduction

Computer sensor systems able to reliably identify activities of daily living would enable novel ubiquitous computing applications for health care, education, and entertainment. For example, in long term home health monitoring, automatic detection of activity may allow people to receive continuous care at home as they age, thus reducing health care costs, improving quality of life, and enabling independence. Recent work on automatic detection of activities in the home setting has shown promising results using machine learning algorithms and data from embedded sensors (e.g., with RFID tags[3] or switch sensors[2]), mobile devices (e.g., accelerometers[5]), or combinations of these[7].

In this paper, we extend these prior results by studying activity detection from sensor data generated by subjects living for a relatively long period in a real but highly instrumented home. Our experiment studied a married couple living for 10 weeks in this home. We annotated and then analyzed a 104 hour sub-set of the data, comprised of data collected on 15 separate days. Neither of the participants was affiliated with the authors, and the ground truth video annotations were

J. Krumm et al. (Eds.): UbiComp 2007, LNCS 4717, pp. 483–500, 2007.

provided by a third party also not part of the research team. Only the activities of the male subject were annotated, due to financial constraints and the original intent of our experiment as discussed in Section 6.5.

We report on: (1) the impact on recognition performance when several different types of sensors are considered (wired switches, RFID tags, wireless object usage detectors, electrical current and water flow detectors, etc.), and (2) difficult cases that were encountered when doing this work that impact how one might design a home activity recognition system. In addition, we highlight some evaluation issues that arose when testing algorithms and that may be of interest for others studying home activity recognition inference systems. Finally, we describe some of the challenges we encountered as we extended previously reported lab experiments (e.g. [11]) to more realistic, real-world conditions.

2 Related Work

Although there are different commercial systems available for activity monitoring in the home, such as Quiet Care Systems[13] and e-Neighbor[14], these provide only a limited analysis of activity. So-called "smart" appliances such as the Japanese "i-pot"[15] only detect one activity.

In this paper we examine methods of monitoring many activities within the home using dense object-based sensing with low cost sensors. In particular, we focus on work using object usage (motion) detectors placed on large objects and furniture [1], body-worn accelerometers [5,4], magnetic reed switches [6,2], water/power system monitors [10] and wrist-worn RFID readers [8]. Prior work shows promising results for these methods. However, it is difficult to extrapolate from reported experiments how these sensors would perform under more real-world conditions for the following reasons.

First, in some cases, test data was generated by having the researchers who developed the system or their affiliates perform activities. Although such data are often sufficient as an early proof of concept for a system, there exists the obvious problem of potential bias. A researcher who knows how the system works may perform an activity in a manner favoring the recognition system. We address this problem by using subjects completely unaffiliated with any of the researchers.

Second, in many cases, the techniques are evaluated on datasets collected when subjects are asked to perform activities collected in short recording sessions, where subjects repeatedly perform random sequences of activities. In each case, the data may not represent the variety of ways in which a subject may perform activities outside of these artificial conditions over a long period of time. For instance, a single subject may ordinarily eat in many locations around the home, but in a time-bounded experiment he or she may always do so at the dining table. In our study, we collect data over a period of 10 weeks and draw our evaluation set from a 15 day subset of this. Additionally, unlike many previous long-term studies, we record full video and audio for annotation rather than relying on self-reporting.

Third, the techniques are often evaluated on data collected in lab settings that are intended to mimic the home, but are not the actual residence of the subjects. In the unfamiliar confines of a lab setting, subjects will likely perform activities in much more restricted ways. For instance, it is unlikely that a subject will sprawl on a couch and eat dinner over a few hours in a lab setting. Although in this work our participants moved into an instrumented home that is not their own, they lived in the home for 10 weeks, allowing time for them to acclimate to the environment.

Fourth, much prior work focuses on proving the effectiveness of one particular type of sensor. Further, each technique is validated on a data set that may be (unintentionally) biased toward the capabilities of the sensor type selected. As a result, it is difficult to gage the relative efficacy of the sensors for recognizing activities. In this work, we employ a number of sensors simultaneously in a single apartment. Our goal is not to introduce interesting new sensors, but rather to allow comparison of previously proposed sensors using a common baseline.

Finally, in most cases, labeling of the data was performed by researchers and their affiliates. Such self-labeling may be unfairly biased towards the system being tested since researchers may favor labels that they expect their system to produce. For instance, when the subject eats in short spurts over a period of hours, there may be a temptation to label the whole period as a single long period of eating. To avoid biases stemming from self-labeling, we employ a professional coder to label our data.

To summarize, we describe in this paper a carefully constructed experiment to compare recently proposed sensors for activity recognition while avoiding a number of limitations in their evaluations. We believe that such apples-to-apples analyses are essential if activity recognition sensor technology is to move beyond the "interesting concept prototype" stage.

3 Description of the Experiment

The aim of our study was to provide guidance for the development of home activity recognition systems by testing under more realistic conditions than may have been achieved in past work. We did this by exploiting an instrumented home environment that permitted multiple modes of sensor data to be collected simultaneously.

3.1 Sensing Environment

We obtained access to the PlaceLab, an instrumented home environment operated as a shared research facility [9], and collected and analyzed data from a couple who lived at the home for a period of 10 weeks. The home is a custom built condominium instrumented with several hundred sensors, including an audiovisual recording system that captures ground truth of the participants' activities. The environment contains the following built-in wired sensors that were used in this work: 101 reed switch sensors installed on doors, cabinets,

drawers and windows, electrical current flow sensors on 37 residential circuits, 36 temperature sensors, 10 humidity sensors, 6 light sensors, 1 barometric pressure sensor, 1 gas sensor and 14 water flow sensors. We also used 277 wireless object usage (motion) detection sensors [1] of three types: 265 "stick-on" object usage sensors that measure when objects move, 2 3-axis accelerometer sensors that are worn on limbs and measure limb movement at 20+Hz, and 10 wireless infra-red motion sensors that detect when there is motion in various regions of the condominium. The object usage sensors were placed on nearly all objects that might be manipulated ranging from doors and cabinets to remote controls for appliances. In some cases these were redundant with wired sensors. The 3-axis accelerometer sensors were worn on the dominant wrist and dominant hip of the male subject. The infra-red motion detector sensors were placed around the apartment to cover each room.

The home's audio-visual recording infrastructure was used to record the behavior of the participants as they lived in the home. The audio-visual record shows all views of the apartment except for the bathrooms, with a limited view of the bedroom. All data is relayed to a central processing and storage facility in the apartment where it is time-stamped on arrival. Figure 1 shows an image of the living room of the home. Despite the ubiquity of the sensor infrastructure, the majority of sensors are embedded in cabinetry or hidden from sight.

Fig. 1. View of the PlaceLab living room taken before the participants moved in

We further augmented the existing sensors in the home by installing 435 RFID tags. We obtained access to an RFID reader in a bracelet form factor [8] and requested that the male subject wear it whenever he was awake and in the home. Ideally, we would have had the female subject wear a bracelet also, but lack of hardware and financial resources for annotation of her activities prevented this.

Readings from the RFID bracelet were sent wirelessly to the central processing and storage facility for logging. Three types of 13.56MHz RFID tags were used: 309 55mmx55mm stickers, 78 86mmx54mm "credit card" style tags and 48 22mm diameter "button" tags. These different form factors trade off parameters such as range, obtrusiveness, cost and durability. Tags were placed on all objects in the home that could be tagged without impacting use of the object in an obvious way and that appeared to permit the tag to be read if the object was

handled normally. Tags were placed on food items, on major kitchen objects such as handles of cooking knives, on appliances and devices (computer mouse-pad/keyboard), under shelf paper at the edge of shelves, inside couch armrests and pillows, inside the front cover of books, etc. In some cases, due to the shape of objects, the makeup of the objects (e.g., metal), or the usage of objects (i.e., might be put in the microwave), it was not possible to place tags. Examples where RFID tags were not placed include the television remote, metallic kitchen appliances, and cups and plates which might be put in the microwave. Once a week when the participants were at work, researchers entered the apartment and added tags to new objects found in the apartment, such as food and magazines.

We believe this is one of the largest and richest continuous datasets collected of its kind, and certainly it is the only one that combines embedded sensors such as switch and flow meters with wearable accelerometer data and RFID readings and contains full video and audio. It is our intent to release as much of this dataset as possible without violating privacy. Researchers interested in using this dataset should contact the authors.

3.2 Participants and Data Collection

A major goal of our experiment was to analyze behavioral data that was as natural as possible. We recruited our participants from a pool of individuals who had responded to advertisements for a study on how to make technology easier to use in the home. The participants were a married couple: a woman, age 31, working in the publishing industry, and a man, age 29, a high school science teacher. Although they both worked in science-related fields, they did not have advanced knowledge of computer science or sensor technology.

The participants were encouraged to maintain as normal a routine as possible. They went to work, had visitors over, cooked meals, and worked on projects and leisure activities according to their own preferences. They brought objects such as small appliances, clothing, bedding, boxes of books and audio tapes, and food from their own home when they moved in. Although they were living away from home, the relatively long duration of the experiment allowed the residents to acclimate to the apartment.

Both participants were interviewed together after the study about the experience of living in the home; the interview was audio recorded and transcribed. Identity-masked interview transcripts are available from the authors. The post study interviews with these and other participants who have lived in the facility indicate that after a few days the sensors do not impact most of the residents' everyday behavior. For example, the male subject reported that, "We weren't as conscious as I thought we would be, it was actually kind of natural being here ... I didn't notice some things as much as I thought I would, like the cameras." The female felt similarly stating, "I wasn't bothered by it, really at all, I thought I might get weirded out every once in a while, but there were very few times where I was totally tired of being in the project, and I felt pretty comfortable here." The participants were asked to wear sensors, which obviously impacted

behavior, but study of the video suggests that the activities that our team was interested in, such as eating, hygiene and grooming, are performed in a natural way. Additional institutional review board procedures and options for deletion of sensitive data may have also helped the participants to feel comfortable in the sensorized environment. Over a ten week period, data was collected for several experiments in addition to this one.

3.3 Data Annotation

At the conclusion of the experiment, an anthropology student unaffiliated with the investigative team was hired to perform annotation. A custom tool was used to annotate the data using the audio and video record. The annotator averaged about 1.5 hours of real-time annotated for each hour of effort. Within our financial constraints, 104 hours of data could be annotated.

Since only the male was wearing the RFID bracelet, only his activities were annotated, and we chose data from a series of days where he was in the instrumented home for the longest periods of time. Data was used from 15 days in total. These days consist of 4 days preceding and 11 days following a 10 day vacation in the middle of the experiment. The first day begins after the couple had been living in the home for 3 weeks, hopefully allowing them time to acclimatize to the environment.

A detailed activity ontology was used for labeling and is available from the authors on request. The activity information is quite fine-grained. For example, activities include "sweeping", "folding laundry" and "brushing teeth". The annotator was given instructions to make reasonable judgments about the start and end times of the activities. In some cases, "foreground" and "background" activities were labeled. For example, "actively watching tv" occurred when the subject appeared to be paying attention, while "watching tv in background" occurred when the TV was on but was only being selectively attended to. The annotator was extremely precise, for instance labeling a "misc. hygiene" activity each time the male subject wiped his face with a napkin during eating. While spot checking the annotator's work, no errors have been found.

3.4 Limitations

Our data collection process has several limitations, some mentioned above that we reiterate here. The instrumented home was not the participants' real home, and although the experiment was much longer than most in prior work, ideally a longer time-frame would be observed and annotated. As mentioned shortly, our dataset is missing some activities and has only limited data on others. Due to the tedious and therefore costly nature of annotation, our results use a 104 hour subset of the collected data. The full bathroom, powder room, and part of the bedroom were not observable by the annotator, so many activities of potential interest related to sleep, personal hygiene, and grooming are not labeled. Finally,

the subjects were participating in other experiments that may have made them somewhat more conscious of the sensors, because, for instance, they were wearing sensors on their bodies. Nevertheless, in comparison with data collection methods and lengths of observation time reported in prior work, the data we collected may be more natural and, as will become apparent, may be more challenging to analyze.

4 Characterization of Data Collected

In this section, we present some initial statistics of our dataset. Unless otherwise stated, this and all subsequent sections will refer only to the 104 hours of data that has been annotated.

4.1 Activity Frequency and Length

We first examined which activities were most common by time and by number. Our ontology contains 98 different activities which cover most aspects of home life. In our dataset, we only have examples of 43 of these. There were no annotations for many of the cleaning, laundry, cooking or yard-work tasks. These tasks were either not performed, or were performed by the subject's spouse whose activities were not annotated because she did not wear an RFID bracelet.

Widely varying amounts of data were collected for each activity. Table 1 shows the amount of data collected for the 5 most and least often observed activities by cumulative time. We see that the amount of sensor data available for each activity can be severely limited by how often the activity is performed and the typical length of the activity. Additionally, we see that several of the infrequently observed activities likely take place in the bedroom or bathroom where we have limited or no video for annotations. Table 1 illustrates that even the 104 hours of data annotated in this work may be too little to build data-driven models of some activities of interest.

Table 1. Most and least often observed activities performed by the male subject

Activity	Total cumulative time (min)
using a computer	1866
listening to music or radio in the background	813
actively watching tv or movies	732
sleeping deeply	728
reading paper/book/magazine	359
preparing a snack	0.74
leaving the home	0.70
making the bed	0.56
washing hands	0.40
drying dishes	0.10

4.2 Statistics of Activities Studied

For the remainder of this paper we will focus on activities or groupings of activities for which we have at least 10 minutes of data. We studied a range of activities that may be useful input to a home health monitoring system. In some cases, this meant defining the activity at a higher level of our ontology. Table 2 lists various statistics of the activities studied. This table highlights the detail of our annotations. For example, there are 197 instances of the "eating" activity. This does not mean the subjects had 197 meals. Rather, there were 197 bouts of eating or drinking instances.

Table 2. Mean, variance, total time and number of instances of the activities studied with sub-activities in italics if applicable. Time units are in minutes unless otherwise noted. The number of instances of sub-activities is shown in parentheses.

Activity	Mean Time	Var Time	Total Time	Number Instances
actively watching tv or movies	33.29	2613	732	22
dishwashing	$23s$	$820s^2$	11	30
-putting away dishes (0), loading the dishwasher (1), hand washing or rinsing dishes (28), drying dishes (1), dishwasher on in the background (0), soaking dishes (0), unloading the dishwasher (0), dishwashing misc (0).				
eating	1.58	25	311	197
- eating a meal (11), eating a snack (35) , drinking (151).				
grooming	1.05	4.4	50	48
- drying hair (0), brushing hair (0), shaving (0), getting undressed (25), applying makeup (0), putting up clothes (0), getting dressed (23), grooming misc (0).				
hygiene	3.05	36	116	38
-brushing teeth (2), washing hands (2), flossing (0), washing face (0), bathing or showering (2), toileting (1), hygiene misc (31).				
meal preparation	$27s$	$2954s^2$	59	132
- cooking or warming food on microwave (3), retrieving ingredients/cookware (39), measuring (0), chopping/slicing/grating food (1), preparing a drink (54), preparing a snack (4), preparing a meal (2),cooking or warming food on stove-top (0), preparing a meal in background (0), cooking or warming food on oven (0) mixing/stirring food (10), combining/adding (15), washing ingredients (0), meal preparation misc (4).				
reading book/paper/magazine	14.36	443	359	25
using a computer	19.24	1068	1866	97
using a phone	2.02	23	204	101

4.3 Complex Behavior

Video footage reveals various complexities in the way activities are performed. Here we describe an episode of the "hand washing or rinsing dishes". We originally watched this sequence to investigate the performance of the RFID reader so the description contains a number of references to tag firings.

The male starts in the office using the computer for a few minutes and apparently wants a drink. The RFID tag under the keyboard fires. The male turns out the light and goes to the kitchen, where he opens the cup cabinet with his right hand (wearing bracelet), but reaches in with the left hand. The tags under the shelves usually fire when the bracelet reaches in, but the participant used the "wrong" hand to grab his cup. Cups don't have tags because of the microwave. He puts the cup on counter and opens the fridge with his right hand. No tags are on the front of the fridge because they did not work due to the metal surface. He reaches in with his right hand and a tag on one of the shelves fires. He grabs a bottle, which is untagged because it was recently purchased, and puts it on counter next to the cup. He leaves the fridge door open and walks out of kitchen into the hallway to speak to his spouse. He comes back and closes the fridge with his right hand and then walks to the living room to get a key chain that has a bottle opener on it. He reaches down to the table with his right hand, at which time a tag for another object on the table might have fired if he were just centimeters closer. He returns to the kitchen, opens the bottle and pours a glass. He takes the bottle to the untagged metal sink and rinses several times holding the bottle in his right hand, without using the tagged soap. He takes a drink and then puts the glass down and carries the bottle down the hall to the recycling area. A tag could fire at the recycle bin, but the area is large and even with 2 tags nearby, his hand does not get sufficiently close. He walks back to the living room and starts cleaning up, leaving the full cup in the kitchen. His spouse is in the apartment activating other sensors the entire time.

The example behavior above is not atypical, and it only takes a few minutes of watching any sequence of the video to encounter examples of behavior that either defy common assumptions about how people will behave or create difficult activity labeling and detection challenges. Examples include eating dinner in several different locations in the home that are not the dining room table, brushing teeth while walking all around the home, eating and snacking for extended lengths of time in front of the television with no clear start and end time, and multiple behaviors that are very similar in how they appear to the sensor stream (e.g. eating vs. watching TV).

These examples and others we have identified highlight several complexities in our dataset that may not be present in datasets collected in less natural conditions: interruptions (e.g. talking to spouse while fridge is open), task abandonment (e.g. leaving the cup on the counter while going to do another task), lack of location specificity for many activities (e.g. eating dinner at the office computer), and interleaved multi-tasking and overlapping multi-tasking (e.g. snacking, doing laundry, watching TV and talking at the same time).

Also common is having two people in the same space, both doing independent activities but also cooperating on activities. Finally, although the RFID sensors that fire are person-specific, the rest of the sensors in the unit fire due to the actions of either of the participants, creating a data analysis challenge. We will return to such challenges as we discuss our results.

5 Activity Classification

Having made some initial observations about the dataset, we now investigate how well we can recognize a set of common activities. We follow standard activity classification procedure and convert the data to a series of feature vectors, each covering a fixed period of time. We then conduct a series of "activity" vs. "the rest of the world" experiments. We use binary classifiers for each activity rather than considering all activities together because very few of the activities are mutually exclusive. We report results for the three main sensor categories studied: RFID, "built-in", which covers all the wired sensors, and "motion", which covers the on-body and on-object accelerometers and the infra-red sensors.

5.1 Data Preparation

We converted the sensor data to a series of vectors formed by concatenating all of the data observed in 30s windows overlapped by 15s. All but three of the activities studied have average durations on the order of a minute.

Different types of sensors require different processing to convert them to features. We assigned one component of the feature vector to each sensor input. For all sensor types, if readings were observed during the time window, we stored the average value. If no readings were observed, for the RFID and motion sensors, we set the sensor value to zero. For the remainder of sensors such as continuous valued sensors (e.g., current flow, water flow) and switches with an on/off state (wired switches in cabinets), if a sensor value was not seen in the current window we used the value from the previous window, assuming that the state of the sensor had not changed. The three sensor types, built-in, motion and RFID, generated 206, 281 and 435 component feature vectors respectively. Experiments involving all sensor types use 926-dimensional feature vectors.

5.2 Label Assignment

We assigned each feature vector to Class 0 or Class 1 according to whether the activity of interest occurred at any time during the 30s window covered. We thus took a very conservative approach to annotation. For activities with typical durations of less than 30s, we expect some error in the class assignments.

5.3 Classification Results

We experimented with two types of static classifiers, naive Bayes and C4.5 decision trees, using the implementations in the WEKA [16] software package. The

decision tree classifiers had consistently superior performance so we restrict our discussion to these in the interest of space. Decision trees have the added advantage of relative transparency of which sensors inputs contribute to classification.

We conducted "leave one day out" cross validation-experiments for each activity using the various input sensor categories. Using folds of one full day of data was chosen as the best method for generating reasonably independent test data that would best reflect a classifier's real world performance. In the discussion section, we address how more simplistic sampling methods for choosing folds can lead to over-fitting. In our method, we designate one day's worth of data as testing and train on data from all the other days in our dataset. Because the training sets were highly unbalanced for many activities, we balanced them by uniformly sampling features from Class 0 (i.e. times when the activity was not being performed) to match the number of features in Class 1. We did not balance the test sets.

Our figure of merit was area under the ROC curve for each cross validation experiment averaged over all the folds. This was chosen as the best overall figure of merit because it gives a measure of goodness at all possible thresholds of a binary classifier and is invariant to class skew. An ROC curve plots the true positive rate vs. the false positive rate. Figure 2 (a) shows an ideal classifier with area 1.0, where every "operating point" on the curve gives only true positives and no false positives. Figure 2 (b) shows a worthless classifier with area 0.5. This is comparable to pure chance where every operating point gives an equal number of false positives for every true positive. A good rule of thumb states that any classifier with ROC area less than 0.7 is poor while any classifier with ROC area greater than 0.9 is excellent (e.g. [17]). Note though that ROC area is an overall measure of goodness. An application can choose to operate at any point on the curve. For example, if false positives are more costly than false negatives, an operating point toward the left hand side would be chosen.

For each activity, we calculated the ROC area over each of the cross-validation tests and averaged the result. Figure 2 (c) shows the curves for classifying the "eating" activity using all sensors. The average curve is shown in bold. This experiment yielded poor performance with an average ROC area of 0.587. Conversely, Figure 2 (d) shows the curves for classifying the activity "dishwashing" using only motion sensors. In this case the average classifier resembles the ideal and has an ROC area of 0.937. Space limitations preclude showing the many ROC curves and detailed analysis generated by all the classifiers studied. We therefore instead report the area under each ROC curve averaged over all the folds/days for each experiment. Figure 3 shows results for the activities and sensor input subsets studied.

A few trends stand out. First, for every activity except "actively watching tv" and "reading," motion-based sensors, which comprise on-object and on-body accelerometers and infra-red sensors, are the best sensor category. In fact, except for "actively watching tv", motion sensors outperform the classifier that results from combining all sensors. This indicates that we have insufficient data to learn how the other sensors contribute to detecting most activities. Thus in most

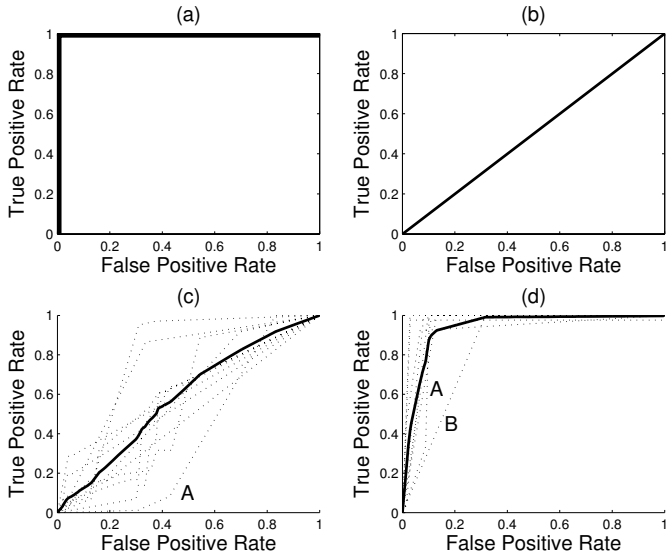

Fig. 2. ROC curves for different classifiers: a) ideal b) worthless and each cross-validation for c) "eating" using all sensors and d) "dishwashing" using only motion-based sensors with the average curve in bold. Labeled curves are discussed in the text.

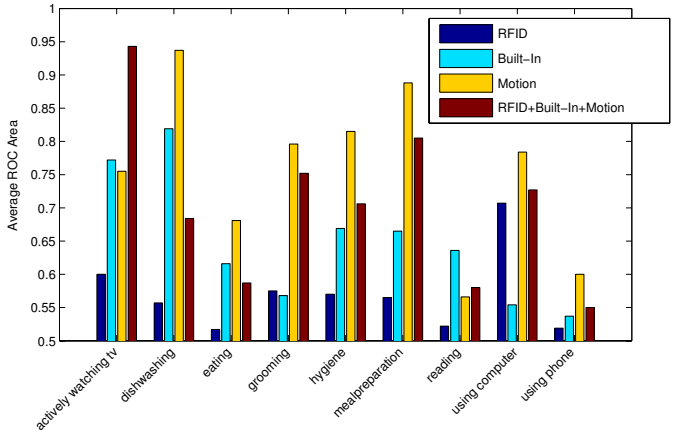

Fig. 3. "Leave one day out" cross validation results for decision tree classifiers for various sensor categories: RFID, built-in (wired switches, current flow etc), motion (object motion, on-body accelerometers and infra-red) and all sensors.

cases, the other sensors add little to motion sensors for this dataset. When motion sensors are outperformed by any other kind of sensor, it is not by a great deal. Except for "reading" and "using phone", motion sensors yield ROC areas greater than 0.7; they are therefore not just better performers than the other

other sensors, but at least moderately good performers according to the rule of thumb mentioned above. For the "dishwashing" activity, the performance is excellent. Given the set of sensors we used and the activities we observed, motion-based sensors appear to be the most promising. We discuss the reasons for this in Section 6.1.

Second, RFID has less than acceptable performance on every activity except "using computer". This can be explained by data sparsity. Examination of the data collected showed that there were relatively few RFID firings. On average, a typical five minute time slice contained less than 1 tag firing. Fewer than 10% of all episodes yielded *any* RFID data. We discuss the reasons for this in Section 6.2.

Third, although built-in sensors perform better than RFID sensors, for the most part they fail our 0.7 rule of thumb. This is likely because many of the wired sensors, such as cabinet doors, are not tied to specific objects but rather groups of objects.

6 Discussion

To determine the causes of success and failure of our experiments, for many outlier cases, we watched the video ourselves. Our analysis highlighted a number of considerations that we discuss below.

6.1 Why Motion-Based Sensors Perform Well

Figure 4 shows the performance of the different kinds of motion-based sensors. It is clear that the infra-red detectors have the best performance for almost every activity. The only exceptions are "eating", where on-body accelerometers achieve acceptable performance, and "reading" and "using phone" for which no motion-based sensor, or indeed any sensor, succeeds.

Looking closer at our list of activities, it is clear why the infra-red sensors are so successful; There is a one-to-one mapping between activities we detect acceptably using infra-red sensors and the location where they are almost always performed in the house: "watching tv" happens in the couch area, "dishwashing" in front of the sink, "grooming" in the bed room, "hygiene" in the bathroom, "meal preparation" in the kitchen, but not in front of the sink, and "using computer" in the study. Conversely, the three activities with no fixed location, "eating", "reading" and "using phone" fare poorly with motion sensors.

Overall, it seems that if the performance of activities is strongly correlated with locations in the home and since these locations do not overlap, a few infra-red sensors at these locations can yield excellent performance.

6.2 Why RFID Performs Poorly

The primary reason RFID proved to be a poor sensor was that for most activities, it detected very few objects being touched. For example, during the "eating and meal preparation" activity, only 26 distinct tags were ever observed to fire in

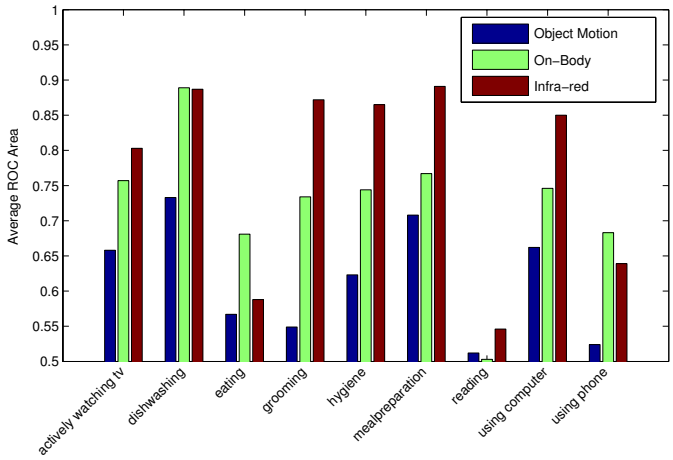

Fig. 4. 'Leave one day out' cross validation results for decision tree classifiers for the different kinds of motion sensors

the kitchen despite 205 tags being installed there. For the "hygiene" activity, only 7 distinct tags of the 50 installed in the bathroom and powder room were ever observed to fire. This lack of tag firing means that even using a temporal classifier ([11]) does not improve the results since there is simply too little data for training.

In order to understand why the bracelet detected so few tags, we identified 10 activity episodes where we expected objects to be used, but did not see any tag firings. These spot-checked activities checked included "hand washing or rinsing dishes", "hygiene miscellaneous" and "using phone". In each case, we examined video of the episodes. An example of our notes was described in Section 4.3.

Our findings from the video may be summarized as follows:

1. Some activities,(e.g., sleeping), do not involve interactions with objects
2. Many activities (e.g., dishwashing) involve objects that could not be tagged because of the object is metallic or went in the microwave.
3. Some activities involve objects that were too small to tag, and
4. Some activities involved tagged objects that were missed because the bracelet was on the opposite hand.
5. For certain activities (e.g., toileting aspects of hygiene), the bracelet was removed, according to followup interviews.

In this work we tagged as many objects as we could, doing our best to place tags so they would fire during normal object use but not impede normal usage of a device or be too visibly noticeable. Based on our results, however, that strategy was insufficient. Instead, it may be more useful to ensure that a few specific objects of interest have one or more tags that will definitely fire under normal use. Objects may require either multiple tags, highly visible tags, or even changes to the objects themselves (e.g., metal to plastic) to overcome some of the

limitations of RFID technology. In practice, it may be useful to observe subjects performing activities before attaching RFID tags. Improving bracelet range and requiring bracelets on both arms may also improve RFID performance.

6.3 Cross-Validation and Over-Fitting

Initially, we conducted a series of 10-fold cross validation tests in which for each fold, we constructed the testing and training sets by assigning every 10th positive and negative data point to the positive and negative test sets respectively. Since little variability is seen across the adjacent time slices, we observed good results. Indeed, the average ROC area for all sensor types was close to or above 0.9 except for RFID which again performed poorly due to data sparsity. However, analysis of the "excellent" decision trees learned showed that they were over-fitted, evidenced by deep trees with little overlap between the sensors used for decisions between folds. This highlights an important issue to consider when building statistical models for home activities. Data is often collected in one session, but lack of variability in the way activities are performed and the lack of major changes in the environment may promote algorithm over-fitting, locking on to possibly spurious features that just happened not to change much. As more sensors are used, the chance of modeling a sporadic correlation may increase. For instance, if a sensor is on a kitchen appliance that happens to be near another sensorized appliance, firings from both appliances might be built into a model. But the minute one appliance is moved that correlation will no longer exist and the model will fail. We believe we have reduced this problem by using one day cross-validation and that these results are a more robust estimate of real world performance.

6.4 Performance vs. the Marginal Model

To illustrate the benefits of our classifiers over more naive schemes, we compared one of our classifiers to the marginal model classifier that takes maximum advantage of the unbalanced set. Specifically we chose the example of dishwashing where a classifier that always chose "not dishwashing" would be correct nearly 100% of the time (0.003% error). Table 3 compares this classifier to our trained "dishwashing"classifier for day "9-11-2006," shown as curve A in Figure 2(d). We see that at the (0,0) operating point, the trained classifier has the same accuracy. However this is a useless operating point since no dishwashing events are ever reported. By operating at 90% accuracy our model is able to report all instances of dishwashing with only 10% false positives.

6.5 Other Considerations

In addition to the lessons learned about the various sensor types, our experiments highlighted the following points that are of interest to anybody designing or experimenting with home activity recognition systems.

Table 3. Performance of a "dishwashing" classifier trained on motion-based sensor data vs. the marginal model that always chooses "not dishwashing"

Classifier	True Positive Rate	False Positive Rate	Accuracy
Trained	1.00	1.00	0.00
	1.00	0.90	0.11
	1.00	0.10	0.90
	0.33	0.09	0.91
	0.00	0.00	1.00
Marginal	0.00	0.00	1.00

Lack of Data. Despite annotation of 104 hours of the male's activity, we experienced two types of lack of data. The first is a lack of sufficient number of observed examples of specific activities to use for training. We have less than a minute's worth of data for activities such as "drying dishes" and "making the bed" which is likely unreasonable for training feature vectors of length $O(100)$ or $O(1000)$, given the observed variability in the way that some activities are performed. The second type of lack of data is the number of sensors of particular types that fired within each bout of activity. The RFID example in Section 6.2 highlights this problem; despite a very high density of RFID tags, a sufficient number of tags was not detected for most activities.

Data at Transitions. A related problem to lack of sensor data within a bout is that for some activities and types of sensors, sensor firings may cluster at the beginning and end of the activity, with a sparse or non-existent signal throughout the activity itself. An example is eating, when RFID tags or object motion detectors may fire during food preparation, but then once the participants sit down on the couch to eat and watch television, no more eating-related sensors may fire until they go back to the kitchen to wash dishes. Those that might fire (e.g., an RFID sensor on a remote control) may only indicate another activity, such as television watching.

Multiple Subjects and Incomplete Annotations. The presence of a second subject whose actions were not annotated in our ground truth was a definite source of error. For example, we attempted to train a "cooking or warming food on microwave" classifier using only the 5 sensors directly related to the microwave (outlet current, object motion sensor on the microwave etc) and tested it against our ground truth. Contrary to common sense, the results showed that this classifier performed poorly. This is due to the fact that the times when the female used the microwave were not marked as true positives. Similarly, when we examined video to understand why classifier labeled B in Figure 2(d) performed relatively poorly, we saw that the this was due to the female subject using the sink to wash objects. Although we had originally intended to conduct experiments using only RFID data and thought that only the male's activities

would be of interest, this assumption failed as soon as we took all sensors into account. The main reason the results were reasonable for many activities is that often the couple did things together in the same location.

Annotating Events with Privacy Concerns. Although we had the luxury of full video and audio for annotating purposes in much of the instrumented home, these facilities were limited for privacy reasons in the bedroom and bathroom. Unfortunately though, many health-related activities take place in these rooms. Our annotator had only the audio to guide her here, so she was unable to annotate many of these activities. For example, 31 of the 38 observed "hygiene" activities are labeled "hygiene misc".

Behavioral Factors. We cannot stress enough that this experiment highlighted that the way people behave when they live somewhere for a while is likely very different from the way they might simulate an activity in the lab or self-report how they perform it. We saw many examples of interrupted activities and multitasking in our dataset. Eating, in particular, was performed in several places and in a variety of ways but usually not at the dining room table. For example, we examined the video for the worst eating classifier, labeled A in Figure 2(c). On this day, "eating" consisted of the subjects "grazing" on food for several hours in front of the television. Differentiating "eating" from "non-eating" in this scenario is difficult because the sensor firings are practically identical for both classes.

7 Conclusions

In this paper, we described experiments performed on 104 hours of annotated activity data collected from a person living in a home instrumented with over 900 sensor inputs. These included built-in wired sensors, motion-detection sensors and RFID tags. The subject wore an RFID reader in a bracelet form factor. Neither the subject nor the annotator were affiliated with the authors.

We found that 10 infra-red motion detectors outperformed the other sensors on many of the activities studied, especially those which were typically performed in the same location. However, several activities, in particular "eating" and "reading" were difficult to detect and will likely require the use of additional sensors and improved algorithms We may have found different results had we had sufficient data to analyze fine-grained activities.

Although some of our classifiers may be sufficiently good for development of some ubiquitous computing applications, on the whole we found this dataset to present a challenge for automatic activity recognition. Some of the problems that we have characterized may not have been as evident in prior work when data was collected under more controlled conditions. This work highlights the importance of studying real-world behavior in home settings when proposing and evaluating home-based activity recognition algorithms.

References

1. Munguia Tapia, E., Intille, S.S., Lopez, L., Larson, K.: The design of a portable kit of wireless sensors for naturalistic data collection. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 117–134. Springer, Heidelberg (2006)
2. Munguia Tapia, E., Intille, S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
3. Philipose, M., Smith, J.R., Jiang, B., Mamishev, A., Roy, S., Sundara-Rajan, K.: Battery-free wireless identification and sensing. IEEE Pervasive Computing 4(1), 37–45 (2005)
4. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannaford, B.: A hybrid discriminative/generative approach for modeling human activities. In: IJCAI, pp. 766–772 (2005)
5. Bao, L., Intille, S.S.: Activity recognition in the home setting using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
6. Wilson, D.H., Atkeson, C.G.: Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 62–79. Springer, Heidelberg (2005)
7. Wang, S., Pentney, W., Popescu, A.-M., Choudhury, T., Philipose, M.: Common sense joint training of human activity recognizers. In: Proceedings of IJCAI 2007 (2007)
8. Fishkin, K.P., Philipose, M., Rea, A.D.: Hands-On RFID: Wireless wearables for detecting use of objects. In: ISWC, pp. 38–43 (2005)
9. Intille, S.S., Larson, K., Munguia Tapia, E., Beaudin, J.S., Kaushik, P., Nawyn, J., Rockinson, R.: Using a live-in laboratory for ubiquitous computing research. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 349–365. Springer, Heidelberg (2006)
10. Fogarty, J., Au, C., Hudson, S.E.: Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In: UIST, pp. 91–100 (2006)
11. Patterson, D., Fox, D., Kautz, H., Philipose, M.: Fine-Grained Activity Recognition by Aggregating Abstract Object Usage. In: ISWC (2005)
12. Placelab Data website. Available: `architecture.mit.edu/house_n/data/PlaceLab/PlaceLab.htm`
13. Quiet care systems. Available: `www.quietcaresystems.com`
14. E-Neighbor system from Healthsense. Available: `www.healthsense.com`
15. i-pot from Zojirushi Corporation. Available: `www.mimamori.net`
16. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
17. Streiner, D.L., Cairney, J.: What's under the ROC? An introduction to receiver operating characteristic curves. Canadian Journal of Psychiatry 52(2) (2007)