Acquiring In Situ Training Data for Context-Aware Ubiquitous Computing Applications

Stephen S. Intille and Ling Bao and Emmanuel Munguia Tapia and John Rondoni
Massachusetts Institute of Technology
1 Cambridge Center, 4FL
Cambridge, MA 02142 USA
intille@mit.edu

ABSTRACT

Ubiquitous, context-aware computer systems may ultimately enable computer applications that naturally and usefully respond to a user's everyday activity. Although new algorithms that can automatically detect context from wearable and environmental sensor systems show promise, many of the most flexible and robust systems use probabilistic detection algorithms that require extensive libraries of training data with labeled examples. In this paper, we describe the need for such training data and some challenges we have identified when trying to collect it while testing three context-detection systems for ubiquitous computing and mobile applications.

Author Keywords

Context-aware, ubiquitous, computing, supervised learning, experience sampling, user interface design

ACM Classification Keywords

H5.m Information interfaces and presentation (e.g. HCI): Miscellaneous.

INTRODUCTION

Traditional desktop computing applications are reactive – they wait until the user expresses an intent via a keyboard and mouse before taking action. *Context-aware* applications, however, can use sensors to infer a user's activity to automatically determine good times and places to proactively present or request information [1]. One of the key difficulties in creating useful and robust ubiquitous, context-aware computer applications is developing the algorithms that can detect context from noisy and often ambiguous sensor data. To date, most context-aware prototype systems have assumed a one-to-one mapping from one sensor reading (e.g. GPS location) to some action (e.g. display place-specific information). A challenge is to create context recognition algorithms that detect complex activities such as "cooking" and "walking" that differ greatly across users and that are not defined by a single sensor activation.

Researchers in computer vision, speech processing, machine learning, and other computational perception domains have found supervised learning algorithms to be highly effective at complex recognition of activity by computer. These algorithms use training sets of sensor data that have been annotated with activity labels to construct computational models that capture the variability in the examples and the uncertainty in the sensor measurements. New sensor readings are then compared with the models and classified using maximum likelihood reasoning or other statistically-based matching techniques. With good training sets, supervised learning techniques can be significantly less brittle than context-detectors that employ hand-constructed rule-based models. Further, supervised learning techniques can be used to create context detection algorithms that are customized to individual users in the field by training the algorithms on user-specific datasets.

Although they can be effective, the supervised learning context recognition algorithms raise new user interface design challenges. Much user activity is situated – the setting will influence behavior. Therefore, algorithms that detect this behavior must be trained using representative examples of the activity acquired *in situ*. This paper is about the challenge of acquiring these examples. We argue that context-aware applications will need to be extended so that the labeled training sets required by the context-detection algorithms can be obtained from the users themselves after the applications are deployed. We report on observations made while conducting three studies. Each study attempted (with varying degrees of success) to acquire supervised learning training data from users in non-laboratory environments. Our studies were with subjects who were participating for compensation and to advance engineering research. However, even with this highly motivated group, the difficulty of developing strategies for acquiring good training data from users quickly became apparent. We suggest some design guidelines and flag *in situ* collection of training data from end users of context-aware ubiquitous computing applications as an important research topic that merits future work.

CONTEXT FROM PHYSICAL ACTIVITY

The type of context information an algorithm requires impacts the type of supervised training data a system might

CHI 2004, April 24–29, 2004, Vienna, Austria.

need to collect. In this paper we focus on context information that can be acquired from a person's *physical activity*, as opposed to other types of context (e.g. location-based, emotional). Most everyday physical activities roughly fall into one of two categories. Some activities, such as walking, running, scrubbing, and vacuuming, require a user to engage in highly repetitive motion of the body and limbs. For these activities, wearable mobile computing sensors such as accelerometers can feed data into classification algorithms that can often robustly detect the activities. Because the user's movement is dictated by the structure of the human body, classification algorithms that interpret the sensor data may require less user-specific supervised data to be collected to provide accurate activity recognition. Our results discussed shortly support this conclusion [3]. Other types of activities may be easier to detect by modeling repetition in the use of objects in the environment rather than in the movement of the user's body. For instance, to detect "cooking" it may be easier to learn that cabinet 5 typically opens before cabinet 3, and sometime approximately 45 minutes after that, appliances 2 and 5 will run. The user's limb and body movement may be highly variable, but if sensors can detect the use of specific doors, drawers, switches, and appliances, classifiers can be trained to detect some common household activities [16]. Because sensor readings for some types of user activity will be highly dependent upon the user, the user's location, the structure and sensors in the user's environment, and other contextual factors (e.g. time, date, history of activities), user-specific training data for supervised learning algorithms for many activities will be required each time a new user initializes a context-aware system. Acquiring training data for physical activity is particularly challenging, in part, because acquiring a single example of a particular activity can take a substantial amount of time.

If applications will ultimately require that supervised learning training data be acquired from end users in the field, two problems must be addressed. First, researchers must begin to make efforts to test context-detection algorithms on data that is acquired and annotated not by the researchers themselves but by the end users. Second, researchers must identify user-acceptable design strategies for *in situ* acquisition of training examples, even for activities of interest that take a long time to complete or that are intimately tied to the environment. We use case studies from our research focused on the first challenge [9] to report on some observations relevant to the second.

SPEECH RECOGNITION ANALOGY

The use of supervised learning algorithms for context recognition has been motivated, in part, by the success of continuous speech recognition systems. These systems are powerful because of the training data exploited by the classification algorithms. State-of-the-art speaker-dependent speech recognition systems such as desktop dictation software use large labeled speech corpi in combination with additional training sets acquired from the intended user that customize the models to the individual [20]. Speaker independent systems must be trained on even larger corpi of labeled speech data.

The user training for speaker-dependent speech recognition systems typically consists of a known passage that must be read by the user, usually taking about 30 minutes. The user also loads previous text documents into the system so the speech recognition algorithm can learn about the user's particular vocabulary. Accurate speech recognition relies upon this training on individual speech to account for individual variation in word pronunciation and to reduce uncertainty in sentence-level recognition [20].

Compare a speech recognition system with a context-aware system that must detect which activities a user is performing: walking, sitting, cooking, vacuuming, etc. What is the analogous user training step? The equivalent of having a user read a passage might be to require a user perform a series of actions in a particular order. Realistically, however, because some of these activities may be complex and time consuming (e.g. cooking), a user cannot be expected to repeat long sequences of activities in a script-like fashion. Even if the user could be asked to do this, the behavior observed by the sensors will not be as complex as natural behavior. The analogy to loading previous text documents would be to provide lists of typical activities and their temporal sequencing. Users will not have this information, except at the coarsest level (e.g. "I eat dinner about 6:30 every night"). In short, the context-aware user interface designer must create not only the context-aware application, but the application used to gather training data required by the supervised learning context-detection algorithms. A design challenge is that this software must acquire examples without becoming disruptive.

Some activities such as standing, walking, running, jumping, and sitting may be analogous to "words," because the structure of the human body may constrain the way the activities can be accomplished so that they can be detected from a large person-independent training set using on-body sensors. As with word pronunciation, some variability in style (e.g. differences in walking due to aging) may be captured in large person-independent training sets. Sequences of word-level activities can be thought of as "sentences." "Going to work" could be a sentence level activity involving sequences of walking, standing, sitting, and climbing stairs. Given a sufficiently large training set, transition probabilities between the word level activities could be computed much as bi-grams and tri-grams are computed for continuous speech recognition. Although algorithms such as hidden Markov models may work well for recognizing discrete words and continuous speech [20] and show some promise for activity recognition, collecting sufficient data to train the Markov models to work in non-speech domains is a significant user interface challenge for ubiquitous, context-aware systems. Whereas sentences in speech take seconds to utter, sentence-level physical activities may take minutes or hours to complete.

Beyond the size of the data set, the quality of the data is also critical for effective supervised learning. Specifically, it is essential to verify activity recognition systems on data collected under naturalistic circumstances because laboratory

environments may artificially constrict, simplify, influence, or change activity patterns. For instance, laboratory acceleration data of walking displays distinct phases of a consistent gait cycle that can aide recognition of pace and incline [2]. However, acceleration data from the same person outside of the laboratory may display marked fluctuation in the relation of gait phases and total gait length. In a naturalistic environment, the person is less aware of the fact that his walking is monitored and may assume a more relaxed, unconventional gait. Furthermore, traffic patterns and mood may frequently alter the person's walking pace and style. Consequently, a highly accurate activity recognition algorithm trained on laboratory data may rely too heavily on distinct phases and periodicity of an activity. The accuracy of such a system may suffer when tested on naturalistic data, where there is greater variation in behavior.

EXAMPLES FROM NATURAL SETTINGS

Most work on detection of user activity relies on data collected from subjects under artificially constrained laboratory settings to validate recognition results. For instance, researchers testing systems that detect activity from mobile sensors typically test the systems on just a few people, usually the researchers themselves or their colleagues [7, 15, 24, 2]. Researchers evaluating data collected in natural, non-laboratory settings typically only use limited data sets collected from one individual [23]. In some cases prototypes are developed and run in natural settings, but quantitative performance is not reported because supervised training data is not obtained [12, 14, 13]. Research using naturalistic data collected from multiple subjects has generally focused on recognition of a limited subset (e.g. nine or fewer) of everyday activities. For recognition of activity from mobile computers, these consist largely of ambulatory motions and basic postures such as sitting and standing [8, 6] that may require limited training data because they are strongly constrained by human anatomy. Consequently, it is not clear how well any of these systems will perform recognizing a variety of everyday activities for a diverse sample population under real-world conditions. Training data for particular individuals may be required to obtain good recognition performance outside of the laboratory.

While exploring these issues, we have implemented technology for acquiring user activity sensor data from natural settings [9], and we have deployed the systems in three types of studies to acquire data for training supervised learning algorithms. We briefly describe each study type and use the examples to discuss the challenges we have encountered when trying to obtain training data from end users *in situ*.

CASE STUDY 1

Mobile computing devices such as phones, watches, and PDAs are becoming increasingly powerful computers. Augmented with sensors such as accelerometers they can be used to support context-aware mobile applications that respond to the activities of the user. Past work has demonstrated 85% to 95% recognition rates for ambulation, posture, and other activities using acceleration data [6, 14, 15, 8, 23, 2, 13]. Activity recognition has been performed on acceleration data collected from the hip [15, 23, 2] and from multiple locations on the body [6, 14, 8]. The energy of a subject's acceleration can discriminate sedentary activities such as sitting or sleeping from moderate intensity activities such as walking or typing and vigorous activities such as running [18, 17]. See [3] for a complete review of recognition of activities from accelerometer data.

Although the literature supports the use of acceleration for physical activity recognition, most prior work on activity recognition using acceleration relies on data collected in controlled laboratory settings [7, 15, 24, 2]. Additionally, prior work focuses on recognizing a special subset of physical activities such as ambulation, with the exception of [8] which examines nine everyday activities. Interestingly, [8] demonstrated 95.8% recognition rates for data collected in the laboratory but only 66.7% recognition rates for data collected outside the laboratory in naturalistic settings. These results suggest that more work is warranted on recognizing a broad array of everyday activities in naturalistic, uncontrolled environments. In addition, more work is needed on the type of training data required for useful activity recognition performance. Recognition accuracy rates of 80% to 95% can be achieved for postures and ambulatory activities using accelerometer data without individual training – where datasets from multiple people are aggregated into one training set [15, 8, 23, 21]. However, this may not be the case for more complex everyday activities. Although comparisons of activity recognition performance with and without the use of individual training is lacking, recognition accuracy for certain activities such as ascending stairs has been shown to improve through use of individual training [11]. Additionally, for an activity such as tooth brushing, individuals may display significant variations in brushing vigor, duration, and posture. In such cases, individual training could improve discrimination of tooth brushing from other similar motions such as window scrubbing. If individual training is required, then the human-computer interface must be adapted to acquire such information without burdening the user.

Study design

In our activity recognition study, elastic medical bandages were used to secure five data collection boards each with a 2-axis accelerometer to five different points on the body [4]. Placement of the sensors is shown in Figure 1a. The devices were light and did not require any wires, minimizing restrictions on subject movement.

Twenty activities were studied.[1] To address ambiguities in activity labels, subjects were provided short sentence descriptions of each activity. For example, walking was described as "walking without carrying any items in your hand or on your back heavier than a pound" and scrubbing was described as "using a sponge, towel, or paper towel to wipe

[1]Walking, walking while carrying items, sitting and relaxing, working on computer, standing still, eating or drinking, watching television, reading, running, bicycling, stretching, strength-training, scrubbing, vacuuming, folding laundry, lying down and relaxing, brushing teeth, climbing stairs, riding an elevator, and riding an escalator.

a window" (for a complete list, see [3]). Subjects first participated in a semi-naturalistic data collection session and then participated in a laboratory data collection session.

For semi-naturalistic data collection, subjects completed an "activity obstacle course": a series of activities listed on a worksheet. These activities were disguised as goals in an obstacle course to minimize subject awareness of how they performed specific activities. For instance, subjects were asked to "use the web to find out what the world's largest city in terms of population is" instead of being asked more directly to "use a computer." Subjects recorded the time they began each obstacle and the time they completed each obstacle. Subjects completed 20 obstacles, one per activity. Acceleration data collected between the start and stop times was labeled with the name of the activity. Subjects were free to rest between obstacles and proceed through the worksheet at their own pace as long as they performed obstacles in the order given. Furthermore, subjects had freedom in how they performed each obstacle. For example, one obstacle was to "read the newspaper in the common room. Read the entirety of at least one non-front page article." The subject could choose which and exactly how many articles to read. There was no researcher supervision of subjects while they collected data under the semi-naturalistic collection protocol. Many activities were performed outside of the lab, but some activities, such as watching TV, vacuuming, lying down and relaxing, could be performed in a common room within the lab equipped with a television, vacuum, sofa, and reading materials.

The obstacle course is somewhat analogous to the speech recognition training situation where a user is asked to read a passage. A preferable but impractical strategy for acquiring training data would have been to allow subjects to do whatever they would have done for several days and to label their natural activity. Unfortunately, this requires direct observation by a researcher, which is a difficult and invasive task. Further, direct observation is only possible for research studies, not for deploying context-aware applications. Alternatively, subject self-report of activities via the experience sampling method [10] could be used, but this can be burdensome for the user and can miss important activities, as we discuss in the next case study. Most problematic, however, is that some activities such as folding laundry, riding escalators, and scrubbing windows occur infrequently. Weeks of observation time might be required to obtain just a few examples, and many examples may be needed to train a recognition system. In this work we compromised by using the obstacle course data collection method.

After obstacle course data collection was completed, subjects were asked to perform a randomized sequence of specific activities listed on a worksheet. For example, the first 3 activities listed on the worksheet might be "bicycling," "riding elevator," and "standing still." As subjects performed each of these activities in order, they labeled the start and stop times for that activity and made any relevant notes about that activity such as "I climbed the stairs instead of using the elevator because the elevator was out of service." Acceler-

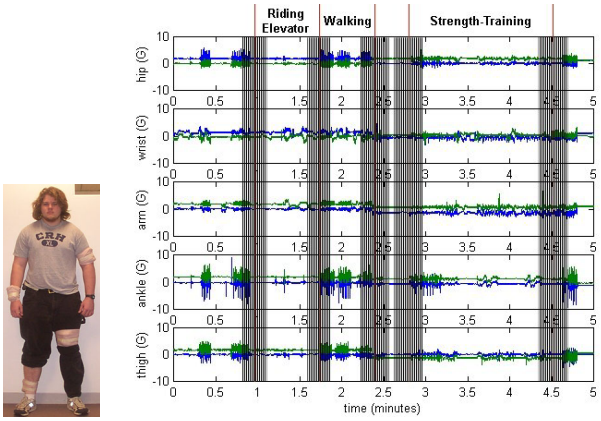

Figure 1: (a) Accelerometers were attached to 20 subjects on the 4 limb positions shown here (bands), plus the right hip. (b) Five min. of 2-axis acceleration data annotated with subject self-report activity labels. Data within 10 s of self-report labels is discarded (masking).

ation data collected between the start and stop times were labeled with the name of the activity. We call this the activity diary data collection step because the subjects were explicitly told which activity to perform and for how long. Even in this step, however, the subjects were not observed or supervised by laboratory personnel.

One goal of this study was to develop and test algorithms using training from both the obstacle course data collection method and the diary data method and to compare the performance of activity recognition algorithms on each.

Algorithm evaluation

Subjects were recruited with fliers seeking research study participants for compensation. Each subject participated in two sessions of the study. Data was collected from 13 males and 7 females ranging in age from 17 to 48. For laboratory data, each subject collected between 54 and 131 minutes of activity diary data (M=96 min., SD=16.7 min.). Eight subjects skipped between one to four activities during laboratory data collection because of factors such as inclement weather, time constraints, or problems with equipment including the television, vacuum, computer, and bicycle. Each subject collected between 82 and 160 minutes of semi-naturalistic, obstacle course data (M=104 min., SD=13.4 min.). Subjects performed each activity on their obstacle course for an average of 156 s (SD=50 s). Six subjects skipped between one to two obstacles during semi-naturalistic data collection due to factors listed earlier.

In the activity diary data, start and stop times for labels were (imprecisely) hand annotated by the subject. Therefore, data within 10 s of the listed start and stop times was discarded, as shown by the masking in Figure 1b.

Mean, energy, entropy, and correlation features were extracted from the acceleration data [3]. Activity recognition on these features was performed using machine learning al-

gorithms such as the C4.5 decision tree classifier [19]. Classifiers were trained and tested using two protocols. Under the first protocol, classifiers were trained on each subject's activity diary data and tested on that subject's obstacle course data. This individual training protocol was repeated for all twenty subjects. The activity diary data acts as the "ground truth" labeled data against which the performance of the algorithms trained with the obstacle course data is evaluated. Under the second protocol, classifiers were trained using both the activity diary and obstacle course data for all subjects except one. The classifiers were then tested on obstacle course data for the only subject left out of the training data set. This leave-one-subject-out validation process was repeated for all twenty subjects. Mean and standard deviation for classification accuracy under both protocols using the C4.5 decision tree classifier was as follows: 71.58 ± 7.438 for individual training and 84.26 ± 5.178 for leave-one-subject-out training. The overall mean recognition accuracy of 84.26% was achieved for the 20 activities using five accelerometers without training on an individual's own data. In fact, the recognition accuracy was significantly higher for all the classifier algorithms we tested when using the leave-one-subject-out validation process [3].[2] This suggests that the effects of individual variation in body acceleration may often be dominated by strong commonalities between people in activity pattern, at least for activities with repetitive limb movements. The leave-one-subject-out validation may have resulted in more generalized and robust activity classifiers because a large training set from 19 subjects was used. The markedly smaller training sets used for the individual training protocol may have limited the accuracy of classifiers.

To control for the effects of sample size in comparing leave-one-subject-out and individual training, preliminary results were gathered using a larger training data set collected for three subjects. These subjects were researcher affiliates and are not included in the results reported for the 20 subjects. Each of these subjects participated in one obstacle course and five activity diary data collection sessions. The C4.5 decision tree algorithm was trained for each individual using data collected from all five of the subjects' activity diary sessions and tested on the obstacle course data. The algorithm was also trained on five activity diary data sets from five random subjects other than the individual and tested on the individual's obstacle course data. The results are compared in Figure 2. In this case, individual training resulted in an increase in recognition accuracy of 4.32% over recognition rates for leave-one-subject-out-training. This difference shows that given equal amounts of training data, individual training can result in better performance than leave-one-subject-out training. However, the certainty of this conclusion is limited by the low number of subjects used for this comparison and the fact that the three individuals studied were all affiliates of the researchers. Nonetheless, these initial results support the need for further study comparing the accuracy of recognizers trained on data specific to an individual versus data collected from many individuals.

[2]The algorithm was evaluated only against known activities, not activities outside of the 20 used for training.

Classifier	Individual Training	Leave-one-subject-out Training
C4.5	77.31 ± 4.328	72.99 ± 8.482

Figure 2: Summary of preliminary classifier results (mean $\pm$ standard deviation) using individual training and leave-one-subject-out training where both training data sets use equivalent amounts of data (each from 5 data collection sessions).

Data collection challenges

We achieved classification accuracy rates between 80% to 95% for walking, running, climbing stairs, standing still, sitting, lying down, working on a computer, bicycling, and vacuuming, These are comparable with recognition results using laboratory data from previous work where training data was collected under significantly more controlled laboratory conditions [15, 11, 24, 21, 8, 23]. We were able to achieve 84.26% overall recognition in this work acquiring datasets from subjects with no researcher supervision.

The recognition results from this study suggest that real-world activity recognition systems that rely on mobile accelerometer sensor data to recognize some everyday activities may perform well using classifiers that are pre-trained on large activity data sets rather than on user-specific training data. Although preliminary results show that individual training could lead to more accurate activity recognition given large training sets, pre-trained systems offer greater convenience by simplifying deployment in non-laboratory settings. Nonetheless, there may be limitations to using pre-trained algorithms. Although "word-level" activities such as running, walking, and scrubbing were accurately recognized, higher level "sentence-level" activities show more variation among individuals. Ambulation and posture are usually similar across able-bodied individuals due to shared physiology, but higher level activities such as "taking a coffee break" or "walking the dog" are more subject to personal behavioral patterns that are strongly tied to the user's environment.

Further, leave-one-subject-out validation still shows deficiencies in recognizing certain activities. Recognition accuracies for stretching and riding an elevator were below 50%. Recognition accuracies for "watching TV" and "riding escalator" were 77.29% and 70.56%, respectively. These activities have motion characteristics that are easily confused with those of other activities. User-specific training sets of sufficient size could lead to improved recognition for these activities.

In this study we had subjects collect their own training data. Although the procedure for training data collection was more free-form than most prior work, some of the activities were still performed in and around the lab. Subjects may, therefore, have performed the actions more consistently than they would have in their own homes. This motivates our other work, where all data collection occurs outside of the lab, such as in subject homes.

OTHER CASE STUDIES

We have conducted two other types of studies where subjects have been asked to self-report on their activities. In the first type, small state-change sensors were installed in the homes of subjects, and subjects were asked to use experience sampling to self-report their activities [9]. The goal was to develop algorithms that can detect activities such as "cooking," "grooming," and "dressing." In the second method, context-aware experience sampling was used to collect data used to develop an algorithm that can detect transitions between physical activities [22]. Both methods have provided insight into how one might develop a human-computer interface for collection of training data.

Using ESM for self report of activity

In the home sensor studies, small sensors are placed on any objects in an apartment that are manipulated by subjects – light switches, appliances, doors, drawers, etc. [9]. The sensors, which are literally taped onto objects, measure movement of the physical objects. About 100 are installed in a typical one-bedroom apartment. The devices are then left to passively collect data, usually for two weeks. To date, we have installed the sensors in 5 homes (2 researcher homes and 3 homes of subjects not affiliated with our research project).

Some of the sensor data collected has been used to develop supervised learning activity recognition algorithms [16]. Because the sensor firings change dramatically when collected from different homes, the activity recognition algorithms *must* use user-specific data for training. In most of our work to date, we have used training data that is hand labeled from researchers [16]. However, we have also begun to explore the collection of user-labeled data. In preliminary experiments using this sensor system, we attempted to use experience sampling. The subjects, who lived alone, were given a PDA running experience sampling software [10]. As the state-change sensors recorded data about movement of objects, the subjects used experience sampling to record information about their activities. We used a high sampling rate; the subjects were beeped via a PDA once every 15 minutes for the two weeks the sensors were in their homes. At the beep, the subjects received the following series of questions. First, the user was asked "What were you doing at the beep?". The subject selected the activity that best matched the one that he/she was doing from a menu listing 33 activities. Next, the PDA asked "For how long have you been doing this activity?" The subject selected from four choices: less than 2 min., less than 5 min, less than 10 min., and more than 10 min. Then, the user was asked, "Were you doing another activity before the beep?". If the user responded positively, the user was presented with another menu of 33 activities to select from.

The self-reported activities can be correlated with the sensor data. Figure 3 shows all the sensor activations for one non-researcher subject one day at about the time she responded to an experience sampling prompt and reported she was "cooking breakfast." Unfortunately, we did not receive a sufficient density of user self reports in this study to train the supervised learning algorithms. Instead we had to defer to data

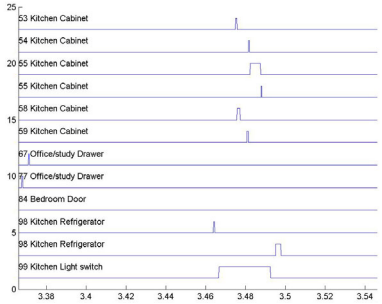

Figure 3: All the sensors that triggered in one 45 minute window for one subject at the time when that subject self-reported she was "cooking breakfast" using experience sampling software.

labeled jointly after the experiment by the investigator and the subject [16].

The reasons we could not use the self-report data for training (based on observation of the data and discussion with the subjects) included the following. *Human error.* The subjects specified an activity that they were not actually carrying out by clicking on the wrong answer box. *False starts.* The subjects specified an activity that they began to carry out but then did not finish. *Multitasking.* While multitasking, the subjects reported the primary activity, but the sensor firings for secondary activities were also recorded by the state-change sensors. *Short duration activities not captured.* Activities with duration shorter than the ESM sampling rate (15 minutes) such as grooming and preparing a beverage proved to be difficult to capture. One subject reported she could wake up and leave for work between sample times. *Number of labels collected.* Tired of being interrupted once every 15 minutes, the subjects sometimes did not answer the ESM questions at the beep, finding a way to ignore the persistent PDA device or to rapidly hit a "mute" option.

For instance, two (non-researcher) subjects, ages 31 and 80, answered an average of 18.7 and 20.1 prompts per day (approximately 33% of total opportunities). However, our research team was able to easily identify more than twice that number of activities from simple inspection of the data. The subjects were graciously volunteering for a scientific study and had agreed in advance to the aggressive 15 min. prompting schedule. Even so, both subjects were frustrated by the ESM within several days. Developers of context-aware systems will not have the luxury of developing systems that are so intrusive. Yet, even with highly-motivated subjects and frequent sampling, we did not obtain a sufficient density and quality of labels for supervised learning algorithm training.

Using CAES for self report of activity

In another study, we used context-aware experience sampling (CAES) to collect training data for a system designed to detect when people are transitioning between office activities such as walking, sitting, standing still, etc. [22]. CAES extends electronic experience sampling to include a set of sensors that both collect data and proactively trigger data

collection self report [10]. In this case, 18 subjects wore a wireless Polar heart rate chest-strap monitor that transmitted data to a PDA in real time. The PDA prompted users to self-report their activity, selecting from a short list of possibilities. The sampling occurred randomly but was also triggered by large changes detected in the subject's heart rate. Our goal was to trigger the most prompts at the times we were most interested in for algorithm development – transitions between physical activities that often cause changes in heart rate.

This technique of triggering questions based upon sensor data holds promise for reducing the burden of user self-report by only triggering prompts at "important" times. In current work, we are studying how to create context detection algorithms that proactively request data from the user for training supervised learning algorithms at times specific activities are detected using other previously trained learning algorithms.

KEY DESIGN OBSERVATIONS

End-users will ultimately balance the inconvenience caused by training set data collection against the value of the context-aware application itself. Applications that have low perceived value will require less-invasive data collection methods. In our experiments we have used subjects who volunteer for studies where some amount of inconvenience is expected. These subjects are probably much more forgiving of interruptions and self-reporting of activities than end-users of ubiquitous computing systems will be. Nevertheless, even with highly tolerant subjects, the data collection methods we have tried require improvement. The three studies mentioned in this paper have led us to identify five design goals for context-aware systems that must acquire training data from users.

Account for the user's mental model. In qualitative interviews, users of experience sampling reported that they did not mind "telling the computer" what they were doing when they were doing something new. However, they found ESM frustrating when they had to "teach" the device "what it should already know."

Use context-aware sampling. One way to avoid asking users what they are doing when the users have already indicated particular activities is to use sensors to monitor for large activity changes. Context-aware experience sampling (e.g. using changes in motion or heart rate) could dramatically reduce disruptive sampling during many activities, clustering prompts for self-report at activity transitions.

Mix prompting and self report. Another way to minimize subject annoyance is to allow the users to proactively indicate when they begin and end activities. In post-session interviews, our subjects expressed an interest and willingness to proactively label in order to reduce random prompting. However, over time subjects will forget to proactively label activity. Therefore, the sampling interface must balance proactive reporting and prompted reporting to minimize user frustration.

Show users their own data. We have also found that by showing subjects their own data, they become more personally invested in the data collection process [5]. A user interface for collecting supervised training examples may collect higher quality data if the interface affords opportunities for the users to study the labels they have generated and "fill in the gaps."

Make it easy for the user. Our subjects expressed frustration at having to search long lists to answer questions. A speech interface with keyword detection, at least for users who live alone, may allow subjects to more quickly respond to prompts thereby increasing the response rate.

IMPLICATIONS FOR CONTEXT-AWARE DESIGNERS

The case studies suggest that more research is needed to determine the type of supervised training datasets that will be required to train context-aware systems that detect a user's activities. However, activities consisting of highly repetitive body motions such as walking, running, and scrubbing may require only small amounts of training data to get good recognition performance. If that is true, user interfaces could be designed that, out of the box, ask users to complete a small set of scripted activities from a worksheet. Users might use a mobile computer device, for example, to manually indicate to the computer system roughly when the activities begin and end. We have shown that such data collected without researcher supervision can be used to train good activity detection algorithms. In combination with larger training sets acquired by professional developers in the laboratory, the small user datasets may permit the context-aware system to begin operating with little time and effort on the part of the end user.

Activities such as cooking, however, may require a different training approach. Activities that do not involve highly repetitive body motions that are easy to sense, that are recognized by sensors placed in the environment, and/or that are performed differently from person to person and environment to environment may require a more carefully designed interface to collect training examples that respect the design guidelines mentioned in the previous section. In summary, automating the data collection that may be necessary to train context-aware algorithms in many cases may be as significant a user interface design challenge as creating the context-aware applications themselves.

ACKNOWLEDGMENTS

This work was supported, in part, by National Science Foundation ITR grant #0112900 and the Changing Places/House_n Consortium. The authors thank Jennifer Beaudin and the anonymous reviewers.

REFERENCES

1. G.D. Abowd and E.D. Mynatt. Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction*, 7(1):29–58, 2000.

2. K. Aminian, P. Robert, E. Jequier, and Y. Schutz. Estimation of speed and incline of walking using neural

network. *IEEE Transactions on Instrumentation and Measurement*, 44(3):743–746, 1995.

3. L. Bao. *Physical Activity Recognition from Acceleration Data under Semi-Naturalistic Conditions*. M.Eng. Thesis, Massachusetts Institute of Technology, 2003.

4. L. Bao and S.S. Intille. Activity recognition from user-annotated acceleration data. In *Proceedings of Pervasive 2004: the Second International Conference on Pervasive Computing*. Springer, 2004.

5. J. Beaudin. *From Personal Experience to Design: Externalizing the Homeowner's Needs Assessment Process*. S.M. Thesis, Massachusetts Institute of Technology, 2003.

6. J.B. Bussmann, W.L. Martens, J.H. Tulen, F.C. Schasfoort, H.J. van den Berg-Emons, and H.J. Stam. Measuring daily behavior using ambulatory accelerometry: the Activity Monitor. *Behavior Research Methods, Instruments, & Computers*, 33(3):349–56, 2001.

7. G.S. Chambers, S. Venkatesh, G.A.W. West, and H.H. Bui. Hierarchical recognition of intentional human gestures for sports video annotation. In *International Conference on Pattern Recognition*, pages 1082–5, Quebec City, 2002.

8. F. Foerster, M. Smeja, and J. Fahrenberg. Detection of posture and motion by accelerometry: a validation in ambulatory monitoring. *Computers in Human Behavior*, 15:571–583, 1999.

9. S.S. Intille, E. Munguia Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, L. Bao, and K. Larson. Tools for studying behavior and technology in natural settings. In A.K. Dey, A. Schmidt, and J.F. McCarthy, editors, *Proceedings of UbiComp 2003: Ubiquitous Computing*, volume LNCS 2864, pages 157–174. Springer, Berlin Heidelberg, 2003.

10. S.S. Intille, J. Rondoni, C. Kukla, I. Anacona, and L. Bao. A context-aware experience sampling tool. In *Proceedings of the Conference on Human Factors in Computing Systems: Extended Abstracts*, pages 972–973. ACM Press, NY, NY, 2003.

11. S.-W. Lee and K. Mase. Recognition of walking behaviors for pedestrian navigation. In *Proceedings of 2001 IEEE Conference on Control Applications (CCA01)*, pages 1152–5. IEEE Press, 2001.

12. P. Lukowicz, H. Junker, M. Stager, T.V. Buren, and G. Troster. WearNET: a distributed multi-sensor system for context aware wearables. In G. Borriello and L.E. Holmquist, editors, *Proceedings of UbiComp 2002: Ubiquitous Computing*, volume LNCS 2498, pages 361–70. Springer-Verlag, Berlin Heidelberg, 2002.

13. M. Makikawa and H. Iizumi. Development of an ambulatory physical activity monitoring device and its application for categorization of actions in daily life. In *Proceedings of the 8th World Congress on Medical Informatics: MEDINFO 95*, pages 747–750. North-Holland, Amsterdam, 1995.

14. M. Makikawa, S. Kurata, Y. Higa, Y. Araki, and R. Tokue. Ambulatory monitoring of behavior in daily life by accelerometers set at both-near-sides of the joint. In V. Patel, R. Rogers, and R. Haux, editors, *Proceedings of MedInfo 2001*, volume 10(Pt) 1, pages 840–3. IOS Press, Amsterdam, 2001.

15. J. Mantyjarvi, J. Himberg, and T. Seppanen. Recognizing human motion with multiple acceleration sensors. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 747–52. IEEE Press, 2001.

16. E. Munguia Tapia, S.S. Intille, and K. Larson. Activity recognition in the home setting using simple and ubiquitous sensors. In *Proceedings of Pervasive 2004: the Second International Conference on Pervasive Computing*. Springer, 2004.

17. S.M. Patterson, D.S. Krantz, L.C. Montgomery, P.A. Deuster, S.M. Hedges, and L.E. Nebel. Automated physical activity monitoring: validation and comparison with physiological and self-report measures. *Psychophysiology*, 30(3):296–305, 1993.

18. M.R. Puyau, A.L. Adolph, F.A. Vohra, and N.F. Butte. Validation and calibration of physical activity monitors in children. *Obesity Research*, 10(3):150–7, 2002.

19. J.R. Quinlan. *C4.5 : Programs for Machine Learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

20. L.R. Rabiner and B.-H. Juang. *Fundamentals Of Speech Recognition*. PTR Prentice Hall, Englewood Cliffs, N.J., 1993.

21. C. Randell and H. Muller. Context awareness by analysing accelerometer data. In B. MacIntyre and B. Iannucci, editors, *The Fourth International Symposium on Wearable Computers*, pages 175–176. IEEE Press, 2000.

22. J. Rondoni. *A Context-Aware Application for Experience Sampling and Machine Learning*. M.Eng. Thesis, Massachusetts Institute of Technology, 2003.

23. M. Uiterwaal, E.B. Glerum, H.J. Busser, and R.C. van Lummel. Ambulatory monitoring of physical activity in working situations, a validation study. *Journal of Medical Engineering & Technology.*, 22(4):168–72, 1998.

24. K. Van Laerhoven and O. Cakmakci. What shall we teach our pants? In *The Fourth International Symposium on Wearable Computers*, pages 77–83. IEEE Press, 2000.