Using Context History for Data Collection in the Home

Daniel H. Wilson
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dwilson@cs.cmu.edu

Danny Wyatt
Dept. of Computer Science
& Engineering
University of Washington
Seattle, WA 98195
danny@cs.washington.edu

Matthai Philipose
Intel Research Seattle
1100 NE 45th St., 6th Floor
Seattle, WA 98105
matthai.philipose@intel.com

ABSTRACT

Practical in-home health monitoring technology depends upon accurate activity inference algorithms, which in turn often rely upon labeled examples of activity for training. In this position paper, we describe a technique called the context-aware recognition survey (CARS) – a game-like computer program in which users attempt to correctly guess which activity is happening after seeing a series of symbolic images that represent sensor values generated during the activity. We describe our own implementation of the CARS, introduce preliminary results, and discuss the first steps toward a completely unsupervised system.

INTRODUCTION

Pervasive computing applications implicitly gather *context history* as they collect and store sensor data over time. In this position paper, we describe the context-aware recognition survey (CARS), which employs context history to help users label anonymous activity episodes. User-labeled examples of activity are valuable because they can 1) improve pervasive computing design decisions and 2) be used to train machine learning algorithms that recognize activities.

Drawing on recent research in practical home monitoring systems, game-based image-labeling techniques, and data visualization techniques [2,6,7], we designed a game-like multiple choice test that displays low-level sensor readings as colorful symbols and descriptive text. Users answer the questions with the goal of correctly labeling the activity being depicted. We report a study in which users (N=10) performed a subset of tasks in an instrumented environment and completed a context-aware recognition survey approximately one week later.

RELATED WORK

Several standard classes of methods exist for collecting data about daily activities, including one-on-one or group interviews, direct observation, self report recall surveys, time diaries, and the experience sampling method (ESM) [1, 4]. While direct observation is often reliable, it is prohibitively time-consuming. In interviews and recall surveys, users often have trouble remembering activities and may censor what they do report. Cognitively enhanced recall surveys mitigate forgetfulness by using cues such as photo snapshots. Time diaries also reduce recall and selective reporting bias, but require a commitment from the user to carry around (and use) the diary. Experience sampling uses a prompting mechanism (e.g., a beep) to periodically ask the user for a self-report. These prompts may interrupt activities and must be carefully delivered in order to avoid annoying the user [4]. All of these methods require the participation of the person who performed the activity and others may require outside help as well (e.g., interviewers).

CONTEXT AWARE RECOGNITION SURVEY

The key idea of the context-aware recognition survey is to use contextual information collected by ubiquitous sensors to provide an augmented recall survey that can be performed by anyone at any time, regardless of who performed the activity or how the sensors were configured. The technique consists of the following steps: 1) sensor readings are collected over time and stored, 2) sensor readings are automatically segmented by activity into episodes (called *episode recovery*), 3) episodes are converted into a series of generic, highly descriptive images, and 4) episodes are labeled by users in a game-like computer-based recognition survey. Afterwards, the labeled episodes may be used to train machine learning algorithms or to improve design decisions for pervasive computing applications.

Initial Study

We performed an experiment in which we designed, implemented, and tested a context-aware recognition survey. We now briefly describe the study.

Subjects. We recruited 10 adult volunteers from the university and from the community. Subjects ranged in age from 25 to 32 years, and the sample was 50% female and 50% male. Subject background varied, ranging from librarians to engineers.

Instrumented environment. This study occurred in the author's home. A kitchen and bathroom were instrumented with two types of anonymous, binary sensors: magnetic contact switches and pressure mats. Contact switches were placed on doors and drawers (e.g., refrigerator door, cabinet door, kitchen drawers). Pressure mats were placed in front of important areas (e.g., in front of the sink). Sensors were polled every second and values were stored in a mySQL database.

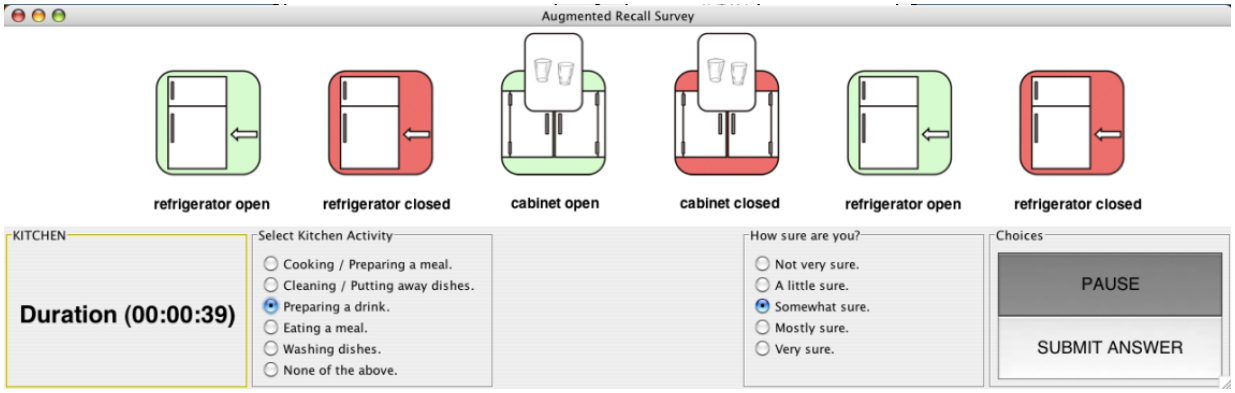

Figure 1. Screenshot of program.

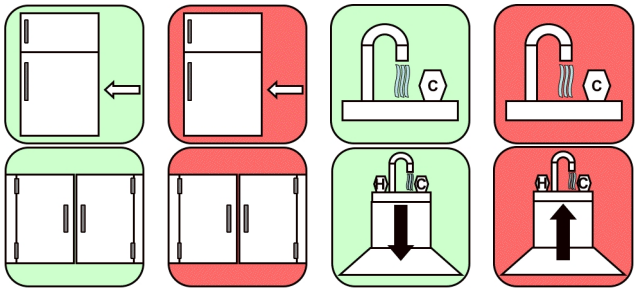

Figure 2. From left to right, top to bottom: (a) Refrigerator open, (b) refrigerator close, (c) cold water on, (d) cold water off, (e) cabinet open, (f) cabinet closed, (g) stand near sink, (h) leave sink.

Activity recording. Subjects were instructed to choose and perform a subset of several kitchen tasks. The kitchen tasks were: prepare a cold drink, prepare either a sandwich, a fried egg, or a microwave pizza, eat the meal, wash dishes and put them away, and throw away any trash. During the bathroom portion, subjects were given a toothbrush and were instructed to brush their teeth and then perform two of three tasks: washing their face, washing their hands, and combing their hair. An observer time-stamped the start and end points of each activity using a laptop computer. Subjects participated one at a time.

Context-Aware Recognition Survey. We presented our computer-based recognition survey as a "game" in which the goal was to correctly guess which activities were happening given only the sensor readings collected from the kitchen and bathroom environments. The contextual information gathered by the sensors was hand-segmented into episodes and converted into a series of images via the Narrator program [7].

See Figure 1 for a screenshot of the computer program. Each episode consisted of a series of scrolling images that had red or green backgrounds, depending on whether that object was turned on or off (see Figure 2). The word "kitchen" or "bathroom" was presented with each episode to indicate the location of the episode. The only timing information included was the total duration of the episode. Subjects were able to pause the scrolling pictures, but were not able to replay an episode. After viewing an episode, subjects were asked to select from a multiple choice list of every possible kitchen or bathroom activity (depending on which room the activity occurred in) plus a "None of the Above" answer. Subjects were also asked to rate how confident they were about their choice on a scale of one to five.

Subjects were administered the CARS on a laptop computer a mean of 5 days following the activity recording. Each subject was presented with two sets of 12 activity episodes, which we call the self set and the other set. The self set contained 8 episodes from the subjects own activities and 4 counterfeit episodes which did not correspond to any activity. The other set contained 8 episodes of someone else's activities and 4 counterfeit episodes. Subjects were informed of which sets were self or other. The survey administration was counterbalanced, with half of the subjects presented the self set first, and the other half with the other set first.

Results

Here, we discuss selected results of our study. See [9] for a more detailed discussion of results.

- Subjects successfully identified 82% of the 24 total episodes (M = 19.60, SD = 3.47). This indicates that **context history is useful for data collection in the home.** Indeed, subjects were able to successfully label most activities with confidence: on the Likert scale of 1-5 (1=Not Sure and 5=Very Sure), subjects reported being Mostly Sure (M = 3.96, SD = 1.03) across all of the episodes. Futhermore, user confidence ratings were significantly related to whether the episode was actually rated correctly, with a significant difference between mean confidence level on correct (M = 3.03, SD = 1.03) vs. incorrect (M = 2.61, SD = 1.06) selections, $t(238) = 2.39, p < .01$.

- Overall, **subjects were equally good at labeling their own or other people's activities**. Ignoring counterfeit episodes, performance on the self section (M = 7.10, SD = 1.29) and the other section (M = 7.10, SD = .99) was identical, with subjects correctly identifying 89% of the 8 possible episodes.

Figure 3. The iBracelet, a wearable RFID reader.

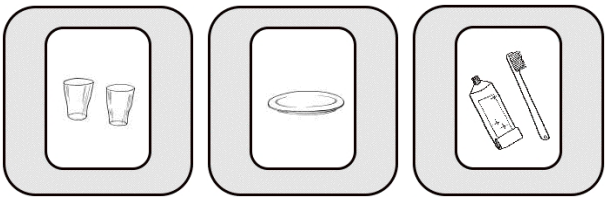

Figure 4. From left to right: (a) Cups, (b) plate, (c) toothbrush & toothpaste.

- The number of days between activity performance and activity recall ranged from 2 to 7 (M = 5.00, SD = 1.63) and was not significantly correlated with total performance scores, $r(8) = .27, p = .44$. This indicates that **context history may help mitigate recall bias.**

- We found that the order of test administration (self then other, or vice versa) impacted performance on the identification of counterfeit episodes. Subjects who completed the self section first were significantly better at detecting fake episodes in the other section ($t(8) = 2.36, p < .05$), indicating that **as subjects gained more practice their performance improved.**

- Subjects reported that they enjoyed using the program, calling the symbols "cute," and "easy to understand." Subjects reported that the symbolic images were "pretty easy" to "very easy" to understand on a Likert scale of 1-5 (*M* = 4.70, *SD* = .48). Thus, we found that using **a scrolling set of symbolic images was a useful approach for displaying context history.**

CURRENT WORK

We identified two main weaknesses in our CARS implementation: 1) we used low-granularity sensors (e.g., contact switches), and 2) we depended on a human to hand-segment the data into episodes. In this section we describe our current solutions in these areas.

Higher Granularity Sensors

In our study, we found that our choice of simple sensors did not provide sufficient granularity for users to confidently label certain activities. For example, it was particularly difficult to tell the difference between washing hands and face. To remedy this situation, we have begun to integrate higher granularity RFID sensors, specifically the iBracelet [5].

Figure 3 illustrates the RFID infrastructure that we assume. On the left is a bracelet which has incorporated into it an antenna, battery, RFID reader and radio. On the right are day-to-day objects with RFID tags (battery-free stickers that currently cost 20-40 cents apiece) attached to them. The reader constantly scans for tags within a few inches. When the wearer of the bracelet handles a tagged object, the tag on the object modulates the signal from the reader to send back a unique 96-bit identifier (ID). The reader can then ship the tag ID wirelessly to a base computer which can map the IDs to object names. We currently assume that subjects or their caregivers will tag objects; we have tagged over a hundred objects in a real home in a few hours. The corresponding CARS symbols are images of the objects being manipulated. We assembled several dozen prototypical object-symbols using the image search function of the Google search engine. See Figure 4 for example symbols.

Automatic Episode Recovery

An attractive aspect of the context-aware recognition survey is the fact that it is completely unsupervised (aside from the user labeling step). In our previous study, however, we hand-segmented the stream of sensor readings generated by the user. In a first step towards automating this step, we conducted a small study that used HMMs bootstrapped with common sense information mined from the Internet. The key idea is to train rough HMM models with information "scraped" from instructional web pages, and then to use these models to identify the segments between activity episodes.

We conducted an experiment to test the usefulness of bootstrapped HMMs for automatic episode recovery. We used data from a previous study in which over 100 RFID tags were deployed in a real home. Objects as diverse as faucets and remote controls were tagged. We had 9 non-researcher subjects with a wearable RFID reader perform, in any order of their choice, 14 ADLs each from a possible set of 65; in practice they restricted themselves to 26 activities over a single 20 to 40 minute session. There were no interleaved activities and a written log was used to establish ground truth.

An HMM was trained on information gathered from the Internet. The datamining process used word appearances on "how to" websites to compute the probability that an object was used during each activity. From this mined information we assembled an HMM with one state for each activity, and a set of observations composed of the set of mined objects, pruned to include only those which we know are in our set of deployed tags. The observation probabilities were set to normalized values of the mined probabilities. We set the HMM's transition probabilities to reflect an expected number of observations (5) for each activity, as well as a uniform probability of switching to any other activity. See [5] for a thorough description of the datamining process.

Next, for each of the 9 sensor traces (one for each subject) we used the Viterbi algorithm to compute the most likely sequence of labels for each object (i.e., sensor reading). We then simply segmented the labeled trace into contiguous sequences of the same label. To measure accuracy of the segmentation we used the P_k metric [3]. The P_k metric is the probability that two observations at a distance of k from one

another are incorrectly segmented. As such, it can be thought of as the error rate for the segmentation and 1 - P_k can be thought of as the segmentation's accuracy. k is set to one half of the average segment length (3 in our case). The P_k score for our segmentation using only the mined parameters is 29.7, indicating that we should expect to be able to segment sensor traces in a completely unsupervised manner with higher than 70% accuracy. This indicates that **bootstrapped HMM models can potentially perform unsupervised episode recovery**.

EXPECTATIONS FOR THE WORKSHOP

Context history is a powerful source of information with many exciting applications. The ECHISE workshop provides the first author an opportunity to meet other researchers who are using similar technologies and approaching similar issues. Moreover, it offers a valuable opportunity to achieve consensus among other researchers as to problem areas and promising avenues of future research.

We are interested in determining how other researchers are using context history in terms of pervasive computing. Specifically, we are interested in sharing tips and techniques for using context history in the domain of automatic health monitoring – an increasingly important application of pervasive technology. How other researchers collect context history, what they choose to collect, and how they present it is of interest. Finally, we are particularly interested in learning how other researchers are dealing with privacy constraints.

CONCLUSION

In this position paper, we described current work with the context-aware recognition survey, an approach for labeling activities that uses contextual information collected by sensors. We presented results from a recent user study, indicating that such an approach can be effective. We discussed improvements being incorporated into the next generation our own CARS. Finally, we described what we hope to get out of the workshop.

AUTHOR BIOGRAPHIES

Daniel H. Wilson is a fourth year Ph.D. candidate in the Robotics Institute of Carnegie Mellon University, where he has received masters degrees in robotics and data mining. His research goal is to provide simultaneous tracking and activity recognition for multiple occupants in the home.

Danny Wyatt is a Ph.D. student in the Department of Computer Science & Engineering at the University of Washington. His research interests include sensing and modeling human behavior.

Matthai Philipose is a researcher at Intel Research Seattle. His primary areas of interest are programming languages and probabilistic reasoning. He is currently working on sensors, data modeling, and statistical reasoning techniques for recognizing human activities.

REFERENCES

1. L. F. Barrett and D. J. Barrett. An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, 19(2):175-185, 2001.
2. C. Beckmann, S. Consolvo, and A. LaMarca. Some assembly required: Supporting end-user sensor installation in domestic ubiquitous computing environments. In *Proc. of UBICOMP 2004*, 2004.
3. D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. Machine Learning. 34(1-3):177-210, 1999.
4. S. Intille, E. M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, and L. Bao. Tools for studying behavior and technology in natural settings. In *Proc. of UBICOMP 2003*, 2003.
5. M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. IEEE Pervasive Computing Magazine 3(4):5057, 2004.
6. L. V. Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of CHI 2004*, pages 319-326, 2004.
7. D. H. Wilson and C. Atkeson. The narrator: A daily activity summarizer using simple sensors in an instrumented environment. In *Adjunct Proc. of UBICOMP 2003: Demonstrations*, pages 141-144, 2003.
8. D. H. Wilson. Simultaneous tracking and activity recognition (STAR) using many anonymous, binary sensors. *Ph.D. Thesis Proposal*, CMU, June 2004.
9. D. H. Wilson, A. C. Long, and C. Atkeson. A context-aware recognition survey for data collection using ubiquitous sensors in the home. In *Proceedings of CHI 2005: Late Breaking Results*, pages 1856-1857, 2005.