ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior

Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, Nadir Weibel
University of California, San Diego, USA
{yvaizman, kellis, glanckriet, weibel}@ucsd.edu

ABSTRACT

We introduce a mobile app for collecting in-the-wild data, including sensor measurements and self-reported labels describing people's behavioral context (*e.g.* driving, eating, in class, shower). Labeled data is necessary for developing context-recognition systems that serve health monitoring, aging care, and more. Acquiring labels without observers is challenging and previous solutions compromised ecological validity, range of behaviors, or amount of data. Our user interface combines past and near-future self-reporting of combinations of relevant context-labels. We deployed the app on the personal smartphones of 60 users and analyzed quantitative data collected in-the-wild and qualitative user-experience reports. The interface's flexibility was important to gain frequent, detailed labels, support diverse behavioral situations, and engage different users: most preferred reporting their past behavior through a daily journal, but some preferred reporting what they're about to do. We integrated insights from this work back into the app, which we make available to researchers for conducting in-the-wild studies.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords

Activity tracking; Behavioral monitoring; Self-reporting; Data collection.

INTRODUCTION

The ability to automatically recognize people's behavioral context (the activities they're doing, where they are, their body posture, *etc.*) is desirable for many domains, such as health management [10, 12], aging care [14, 16] and office assistant systems [30]. Machine learning methods to train and test context-recognition systems require data, including sensor measurements and labels describing the actual context of real people. Many activity-recognition studies validated their systems with data collected in a lab [21, 3, 9] . However, in order to develop ecologically valid systems that work well in the real world, the data used for development should be collected in-the-wild — capturing people's authentic behavior in their regular environments.

Data collection in-the-wild raises technical difficulties related to interruptions in sensor recording and diversity in phone-devices [24] and device placement [15]. The harder challenge, however, is acquiring labels when there is no researcher-observer present with study participants. Previously suggested solutions involved unnatural equipment [20, 7, 25, 2] or simple self-reporting interfaces [8, 13] and resulted in data that had limited ecological validity and labels that describe behavior in a single-dimensional manner and cover a small portion of everyday life.

Recently, we have collected the *ExtraSensory Dataset* from 60 participants using everyday devices [27]. To maintain ecological validity, participants used their *own personal phones*, without restricting phone placement, contributed data from their natural environments (home, work, commute, *etc.*), while they engaged in their natural (unscripted, unobserved, and without a prescribed list of tasks to perform) behavior, and described their own behavior in an authentic, subjective manner. We applied simple machine learning methods to the data and demonstrated successful recognition of a wide variety of everyday contexts, like sleeping, shower, on a bus, *etc.*

In this paper, we present the tool we used to collect the data — the *ExtraSensory App*, a mobile app designed to collect sensor data and engage participants to contribute detailed and frequent labels describing their behavioral context. To evaluate how the user interface enabled and affected data collection, we analyze the quantitative data from the 60 participants, as well as the qualitative feedback that they gave about their experience using the app.

The contribution of this paper is fourfold:

- Design. Our rich user interface enables self-reporting both in-situ (active-feedback and notifications) and recall-based (daily history) and has additional features to facilitate detailed-labeling with little interaction.
- Validation. The app enabled collecting data *in-the-wild*. The resulting *ExtraSensory Dataset* is larger than previous datasets in scale (over 300,000 labeled minutes), range of behaviors (more than 50 diverse context-labels), and detail (combinations of more than three relevant labels per minute).

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

 ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3174128

This data was successfully used to train and test context-recognition systems [27, 28].

- Insights. Our combined analysis of the quantitative data collected in-the-wild and qualitative user-experience reports from our participants helps understand the effectiveness of the various design features. Among our findings: the rich history page facilitated reporting about long behavioral time with detail, using a watch for single-click confirmation of notifications was very helpful, and active-feedback engaged people who preferred reporting about their immediate future rather than recalling their past behavior.
- Open source code. With this paper, we also make the complete source code of the *ExtraSensory App* freely available (`http://extrasensory.ucsd.edu/ExtraSensoryApp`). The published app includes improvements based on the analysis in this paper and can be used either to collect labeled data or as a black-box tool for real-time behavioral context recognition.

RELATED WORK

Previous data collection studies in-the-wild exploited a variety of different approaches to acquire context labels.

Camera-based Approaches

In several studies, participants wore a camera that took snapshots of the scene, enabling context labels to be assigned to different times throughout the day based on the captured images. In some studies research assistants annotated the images [20, 7], compromising the privacy of the participants and their surrounding. In other studies, the participants annotated their own images, which resulted in limiting the range of targeted behaviors (like eating detection [25]) or the number of participants (*e.g.* single person in [2]). Relying solely on camera can miss situations where context is not visible (*e.g.* phone in pocket, singing) or private situations like shower.

Self-Reporting In-Situ

In in-situ self-reporting, participants report their own context (*e.g.* location, activity, emotion, *etc.*) in real-time. For instance, the Experience Sampling Method (ESM) is a technique where the participant is prompted at different times to fill a short form and report their context [23]. This method samples time to estimate statistics of well-being, time-usage, or relations between activities and feelings [5]. The CrowdSignals project [29] aims to collect phone-sensor data from large crowds of users, with additional sparse probing for labels by using quick multiple-choice questions whenever the user unlocks their phone, combined with more in-depth (and less frequent) ESM questionnaires. In [31], whenever the user selected a music playlist, she was prompted to report one out of 13 activities and one out of 10 moods.

In context-recognition studies that target a specific list of activities, researchers often used in-situ self-reporting, but instead of sampling reports in arbitrary times, they let participants actively initiate reporting at relevant times. Studies that tracked a single activity (*e.g.* eating detection [6]) used a simple interface with a single button for the user to mark the start and stop times of eating. Works that targeted multiple activities (like watching TV, driving, *etc.*) added to the interface a selection of a single activity from a list [8, 13]. Commercial systems like Toggl[1] offer similar timer-based reporting.

Self-Reporting by Recall

An alternative to in-situ self-reporting is reporting after-the-fact, by recalling. When the required time resolution is daily, silent notifications may be helpful to remind people to answer simple questions (*e.g.* "how much did you eat today?") every day [1], but for more detail, it can be hard for people to remember their daily events. The Day Reconstruction Method (DRM [11]) is a survey-based method that requires the participant to arrange the previous day in a short diary, as a sequence of episodes. By thinking of each episode as a holistic scene with different contextual aspects (location, activity, interaction with others, emotion), the person can better recall specific variables of interest, like tiredness or joy. In sensor-based context recognition studies, accurate timing of the context is important in order to align the labels with the sensor-data. DRM was used in [26] for eating detection, but the participants struggled remembering when they were eating. The researchers then listened to audio recordings and they reported that annotating eating periods based on audio was difficult.

Mixed Self-Reporting Approaches

Mark *et al.* [18] explored multitasking at work. They assessed productivity with end-of-day surveys, sleep using an actigraph, and monitored computer activity with a custom software.

Rahman *et al.* [22] dealt with self-assessment of stress-level and discussed the trade-off between in-situ reporting (more ecologically valid but disruptive and may cause stress) and recall-based reporting (non disruptive but introduces memory bias). They proposed a compromise solution, where participants could report on their own time, but with the aid of contextual cues like location and ambient sound level, to help them remember how they felt at specific times of the day.

Mehrotra *et al.* [19] explored people's receptivity to phone notifications. Their study combined ESM with cue-assisted recall. Four times a day, a questionnaire presented a selected notification that the phone received in the past four hours, and asked the person what they were doing at the time, how disruptive the notification was, *etc.* They showed higher likelihood to dismiss a notification during complex ongoing tasks, and apparent connection between personality traits and responsiveness to notifications.

Consolvo *et al.* [4] designed and validated UbiFit Garden, an application to promote physical exercise, with both sensor-based automated activity recognition and user manual labeling. The activity recognition component, which was trained on controlled, scripted, and observed data [3], ran in the background and recognized activities like walking and cycling. The user could view the recognized events in a daily journal and delete, add, or change today's and yesterday's events. In addition, the visual appearance of the phone's wallpaper (graphics of flowers and butterflies) was adjusted according to the user's exercise events and was designed to incentivize the user to engage in physical activity or to correct the recognized events.

[1] `https://toggl.com`

THE EXTRASENSORY MOBILE APP

The solution we present in this paper — ExtraSensory — is a mobile app that automatically collects data from a range of sensors built into popular smart phones and a dedicated smart watch. In addition, it provides a rich labeling interface.

Our labeling approach uniquely combines the advantages of multiple existing solutions for self-reporting. Similar to [8, 13], our users can actively report that they are starting an activity. Similarly to the DRM, the users can look at the previous day (or today) as a journal of events and as in UbiFit Garden, this journal is filled by both automated recognition and manual editing [4]. As in the ESM studies, our app also triggers pre-scheduled prompts to ask the user to report labels [23]. Much like survey-based studies with ESM or DRM [23, 11], we address the multi-aspect nature of behavioral context and allow users to report combinations of activities, as well as location, company, body posture and more.

All the sensor-based studies mentioned above provided a study-phone to their participants and constrained the position of the phone. Contrary to that approach, to support ecological validity, we evaluated ExtraSensory with participants that used their own personal phone, in any way convenient to them. In order to broaden the options for participants, we implemented our app for both iPhone and Android. Additionally, we added support for the optional pairing of a Pebble-watch,[2] which can interact with both phone devices, and adds more sensing and user-interaction capabilities to the data collection solution.

Recording Sensors

When ExtraSensory is running (in either the foreground or background of the smartphone), it records a 20-second window of sensor measurements every minute and sends the measurements to a dedicated server. The measurements include 40Hz 3-axial motion sensors (accelerometer, gyroscope, and magnetometer), location coordinates, audio (the app processes the raw audio on the phone to produce 13 Mel Frequency Cepstral Coefficients [17]), and phone-state indicators (app-state, WiFi availability, time-of-day, *etc.*). During the 20-second window, the app also collects measurements from the optional watch, if it is available and used by the participant (25Hz 3-axial accelerometer and compass heading updates).

Communication with the ExtraSensory server is encrypted and users have the option to allow cellular communication or, as all our participants chose, communicate via WiFi only. In case no network is available, measurements are stored until they can be transmitted. The server has a basic activity-classifier that was trained on preliminary data from two iPhone users. When the server receives the sensor data, it responds with a guessed activity (the body posture/movement state), which in turn helps the user report their own subjective labels.

The app has a "data-collection" switch, which is on by default whenever the app is launched. The user can decide, for any reason (low battery, privacy, *etc.*), to temporarily turn data-collection off, in which case new recordings are suspended, but the label-reporting interface is still available.

[2] https://www.pebble.com

Reporting Context Labels

In ExtraSensory, the description of behavior is based on two label components: *main activity* and *secondary activities*.

"Main activity" refers to the body posture/movement state — a single value out of the mutually-exclusive states: *lying down*, *sitting*, *standing in place*, *standing and moving*, *walking*, *running*, and *bicycling*. We included the label "standing and moving" with the intention of describing intermediate situations — not exactly standing in the same position and not exactly walking towards a destination, but something in between (*e.g.* when cooking or cleaning at home).

"Secondary activities" refer to any additional fine-grained attributes that apply to a situation, in a multi-label formulation (multiple labels can apply simultaneously). This includes specific sport activities, work or home activities, transportation modes, as well as other non-activity descriptors for location, phone position, and more. We also defined a multi-label "moods" component but we did not focus on collecting mood labels. The app lets the user decide which labels best describe their own behavior and the goal is to later train classifiers that are able to predict those subjective labels.

The flexible user interface provides a variety of mechanisms to help make label-reporting quick and easy, and has two modes of reporting: *past* and *near-future*.

History page (past). The main route for past reporting is through ExtraSensory's history page (Fig. 1 (A)) — it allows users to engage in some behavior (*e.g.* sleep, drive) and then report about it later. This page displays a daily calendar, where each row represents an "event" — a continuous time segment where the context stayed the same. The server guesses of body state appear with a question mark, to signal to the user that their own labels for this time-segment were not yet provided (*e.g.* "07:52 Walking?"). In case the server guessed the same body state for several consecutive minutes, these minutes appear in the history as merged to a single event (*e.g.* "08:08 – 08:11 Sitting?") and the user can report the same labels to all these minutes simultaneously. By clicking on an event, the app opens the labeling form, where the user can edit the context-labels (Fig. 1 (B)). If the event was already labeled by the user, the existing labels are loaded and can be edited; otherwise, the server-guess is loaded to the "main activity" field and the other fields start blank. After selecting the context-labels in the labeling form and pressing "send feedback", the labels are sent to the server (or queued, waiting for network connection) and the history now displays the time-segment without a question mark, and with the added secondary labels in parenthesis (*e.g.* "07:54 – 08:00 Sitting (At home, Eating)").

The colors of the history rows correspond to the main activity (body-state), ranging from a cold blue for "lying down" to a warm red for "bicycling". The color-code was designed to roughly illustrate the intensity level of movement and help the user visually see when their activity might have changed. With finger-swipe gestures, the user can split a time-segment to separate minutes or merge consecutive rows to a longer event with constant context. Additionally, users can view previous days (by clicking the "previous day" button, at the top left), but they can only edit labels for today and yesterday, to avoid memory bias of looking back too-long ago.

Active feedback (near-future). For cases where users already know what behavior they are going to engage in, ExtraSensory enables pre-labeling immediate-future context. The main route for near-future reporting is the "active feedback" feature: at any time, users can press the green plus-symbol (bottom center, see Fig. 1 (A)); this opens the labeling form in the near-future mode (Fig. 1 (C)), where users can report their current or upcoming context. For example, a user can report that she is going to be driving a car, with family, and that this context is going to stay relevant for the next 25 minutes. After pressing "send feedback", at every new recorded minute, the same labels will automatically be sent to the server, and the user can attend to the actual activity (*e.g.* driving, without distractions). We limit the foresight time (the "valid for" field) to a maximum of 30 minutes in the future.

Selecting the labels – From the labeling form (Fig. 1 (B)–(C)), clicking the "main activity" field opens a simple menu to select a single body-state out of the seven options (in the past-mode, there is an additional "I don't remember" option — in case the user just wants to report secondary activities). Clicking the "secondary activities" field opens a richer menu that allows selecting multiple labels from a list of over 100 labels (Fig. 1 (D)). To make it easier for the user to find the relevant labels quickly, the menu is organized by topics (with quick-link index in the side), like "basic needs" or "transportation." A "frequently-used" section (indexed by the link "frequent") displays the labels that the individual user previously applied, in order of usage frequency, making it quicker to find personalized relevant labels after a day or two of participation.

Notifications (past or near-future) – In addition to the participants' initiated reports, the app also triggers notifications at constant intervals (the default is every 10 minutes, but the user can increase this up to 45 minutes). These notifications remind users to report labels and they provide a direct connection to the labeling form, in either the past or near-future modes, depending on whether the user reported any labels for the recent 20 minutes (see flow diagram in Fig. 2). After reporting near-future context, the next notification is re-scheduled to appear after the reported near-future period is over.

Watch Notifications and quick responses – In addition to the increased sensor recording, the optional smart watch also contributes to the interaction with the user. When a notification is triggered on the phone, it also appears on the the watch. In case the system asks whether the recent context is still the same, and if the answer is "correct", the user can actually respond on the watch by pressing the right top button on the side of the watch (see Fig. 3). In case of an open-ended notification (when there is no user-provided recent context), the notification on the watch serves merely as a reminder (Fig. 4). The visual indication is complemented with a vibration when every new notification appears on the watch (users can disable vibrations, *e.g.* when going to sleep).

Additional Visual Features

Besides the label-reporting mechanisms, ExtraSensory provides additional supporting features. During every 20-second recording window a red dot appears on the control bar of the app (see Fig. 1 (A)) and a "REC" text appears on the watch

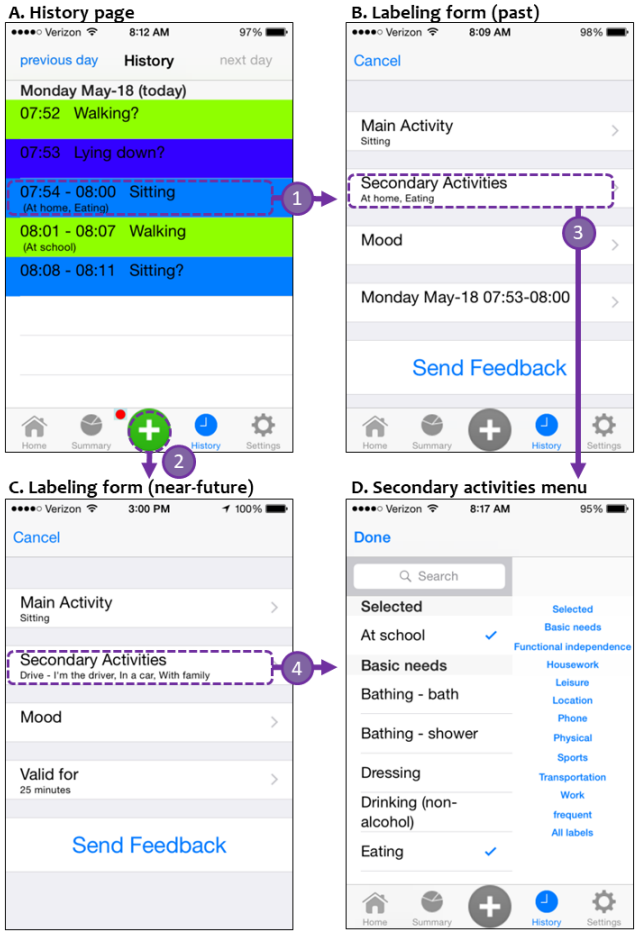

Figure 1: Label-reporting user interface, with flow marked in purple shapes and arrows. In the history page (A), each row represents a segment of time with constant context-labels, either with question mark (server-guess) or without (user-reported). 1) By clicking a row, the app opens the Labeling form in the past-mode (B), where the user can edit the context labels for a specific time-segment in the past. 2) By pressing the active-feedback button (green with plus symbol), the app opens the labeling form in the near-future mode (C), where the user can initiate a report of what they are about to do. 3–4) From the labeling form, pressing the "secondary activities" field opens a rich menu (D), where the user can select multiple labels, jump to a relevant topic, and see personalized frequently-used labels.

(see Fig. 4). Additionally, the app has a home page that acts as a dash-board to keep users informed and to help debug possible problems. The page specifies how many minutes currently have data awaiting to be sent to the server and has an icon indicating whether or not the watch is currently paired with the phone. In the iPhone version, there is an additional cartoon image that symbolizes the latest guessed main activity. This feature was originally designed to attract the user's attention and encourage them to provide their own labels. However, in preliminary experimentation, it became clear that it was more useful to keep the app on the history page rather than the home page, so we did not include the cartoon in the Android version.

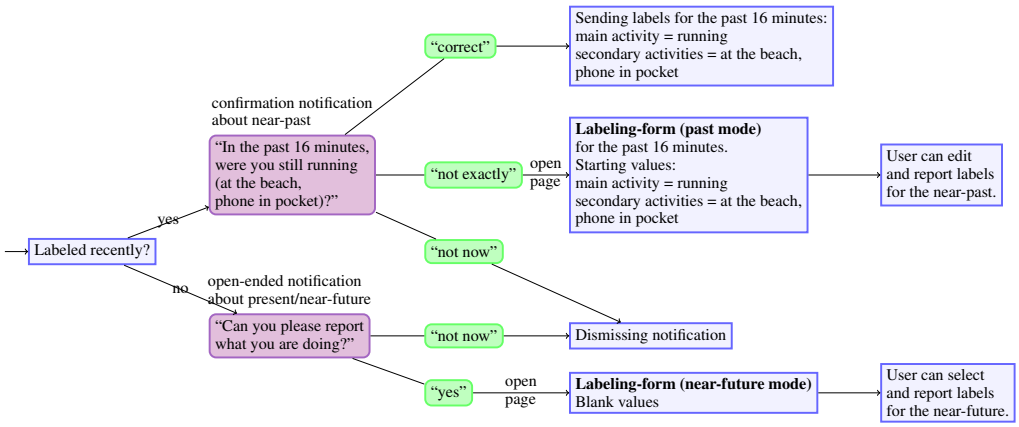

Figure 2: Notification flow with possible example scenarios. The flow starts in periodic intervals and first checks if there are any reported labels for any minute in the past 20 minutes. The purple rounded boxes present the notification messages displayed to the user. The green rounded boxes present the optional buttons for the user. For a confirmation-notification, the user-answer "correct" is a way to send labels for up to 20 minutes with a single click. Two possible routes lead to opening the labeling-form in two different modes: the "not exactly" user-answer enables adjusting the labels for the near-past and the "yes" user-answer enables reporting near-future context (like when pressing the active-feedback plus-symbol button).

Similarly, the app has an additional "summary" page, which displays minute counts of each of the main activity labels, in a pie chart (iPhone) or bar plot (Android), with the same color-code as in the history page. Similar to UbiFit-Garden [4], the user can take a quick glance at this visual summary and possibly decide to report more labels, to update this visualization.

USER DEPLOYMENT, ANALYSIS AND RESULTS

To evaluate ExtraSensory as a solution for data collection in-the-wild, and to collect data to develop context-recognition systems, we conducted an in-the-wild study. We recruited 60 participants (34 female, 26 male). They were mostly students and research assistants at our local university, averaged 25 years in age, and had diverse ethnic backgrounds. With each participant (user), we conducted two meetings, approximately seven days apart.

Figure 3: Watch — confirmation notification. The same notification from the phone scrolls on the top half of the watch app. If the user's context remained the same, they can reply "correct" by pressing the top-right button.

Figure 4: Watch — open-ended notification. This is only useful as a reminder; to initiate reporting labels the user has to go to the phone. During a 20-second recording window, the text "REC" is shown in the bottom half.

In the first meeting, we installed the app on the user's personal phone (34 were iPhone users, 26 were Android users) and provided them with a Pebble smart watch (56 users agreed to wear the watch). The user read and signed the consent form. We explained how to use the app and requested that the users keep the app running (with data-collection on) as much as convenient. We also requested that they use the different label-reporting mechanisms to provide as many labels as convenient (and as much as they can remember) without interfering too much with their natural behavior. We did not specify any targeted activities, but rather asked that they engage in their routine, and report any labels that they believe appropriately describe their context. We explained that the collected data will be de-identified and published and will serve for training systems that can measure people's activities using sensors (but we did not specify any particular application).

In the second meeting, we uninstalled the app from the user's phone, collected the watch back, and asked the user to fill out a short survey about the experience. We also compensated users for their participation with a basic amount of US$40, plus an incentive amount of $0–35, depending on the amount of labeled data that the user contributed. Although we did not explicitly examine the compensation's impact, we can report that 39 users contributed more than enough data to reach the maximum total of $75 and the other 21 averaged $60.

We present results first from the quantitative data that was collected and then from the qualitative user-experience surveys. For both aspects, we examine how the user interface of ExtraSensory influenced the study.

Quantitative Analysis

During the six months of the study, we collected over 300,000 minutes from the 60 users, labeled with combinations of

over 50 diverse context-labels. On average, each minute was assigned more than three labels. These detailed contexts describe over 14,000 distinct "events" (segments of constant context), with median duration of nine minutes. Although not a direct contribution of this paper, we made this dataset, titled the *ExtraSensory Dataset*, publicly available at `http://extrasensory.ucsd.edu`. In this section we analyze these data to gain insight about the usage of our app.

Turning on data-collection. The users had control and could decide when to turn off data-collection (*e.g.* when battery is too low or to maintain privacy). Figure 5 shows two users who participated for approximately seven days and had different patterns of data collection. Some users (like the one presented on the left) kept the app running and data-collection on almost continuously throughout their participation days. Other users (like the one on the right) collected data in many separate segments with gaps.

Figure 5 also shows that during the times that data-collection was on, not all sensors were available all the time. Most notably, the users were free to remove the watch (and turn off the watch app) so there are times when data collection was on but there are no measurements from the watch accelerometer. Similarly, users sometimes turned off location services on their phone or location had a weak signal.

Keeping the app running caused faster draining of the battery. According to users' estimation, on average, they charged the phone 2.3 times a day and the watch once every 1.75 days.

Reporting labels. Figure 6 shows the distribution of label reporting mechanisms over the total recorded minutes in the dataset. For 82% of the recorded minutes, the users provided context-labels. History was the reporting mechanism that yielded most of the label coverage (85% of the labeled minutes). Out of the minutes that were labeled via notification, the vast majority were cases of confirmation-notification (asking whether the near-past context remained the same), and in most of those cases, the user replied "correct" using the watch.

The different reporting mechanisms helped support a variety of situations. Figure 7 shows the relative minute-coverage achieved by the three reporting mechanisms for reporting specific labels. Understandably, the history page was almost the only feasible way to report sleeping. Similarly, laughing was mainly report retroactively (with the history page), which is fitting for a spontaneous action that is typically not predictable in advance. Contexts like "on a bus", "running", and "Yoga" were more planned, so users utilized active feedback more to report starting these contexts. During Yoga, users were not available to reply to notifications, so they had to either use active feedback before starting or history after they were done. When the phone was held in hand, users were more easily available to report in-situ contexts (using active feedback and notifications), but when the phone was in a bag, they had to rely more on the history and report about it after the fact.

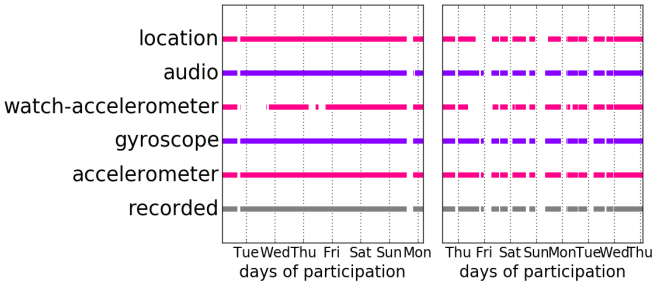

Figure 5: Sensor recordings for two users. The "recorded" row describes when data-collection was on and the other rows refer to recording of specific sensors. The bars indicate when, during the days of participation, data was collected. The vertical dotted grid lines indicate the time 6AM in the participation days. Watch and location were sometimes unavailable.

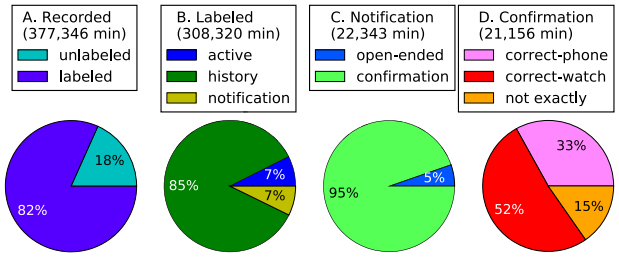

Figure 6: Distribution of label-reporting over the minutes in the dataset. Above each pie chart is the total number of minutes in the whole pie. A) Most of the recorded minutes were labeled. B) The vast majority of labeled minutes were labeled via the history page but active-feedback and notifications still contributed significantly. C) Almost all of the minutes that were labeled via notification were from confirmation-notifications. D) For more than half of the minutes that were labeled via confirmation-notification, the user replied "correct" through the watch. C–D) Relatively few minutes (4,500) were labeled as the result of editing the labeling form triggered by notification (either open-ended notification or when replying "not exactly" to a confirmation-notification).

Figure 8 shows the label reporting patterns across different days of the week and hours of the day. Overall, users turned

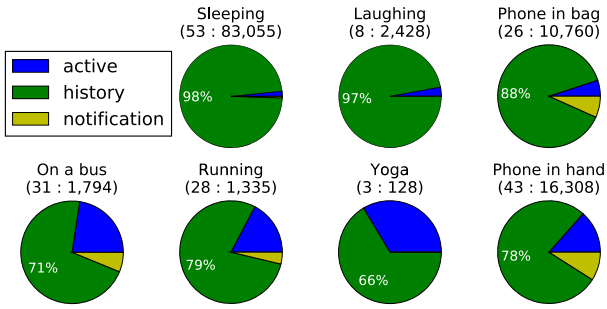

Figure 7: Distribution of label-reporting mechanisms for selected labels. For each label, the title above the pie indicates how many users reported this label followed by how many minutes were annotated with the label (the whole pie). The percentage of minutes for which the label was reported via history is also indicated numerically inside the pie. The top-row contexts were mostly reported via history. For the bottom-row contexts, active-feedback was utilized more.

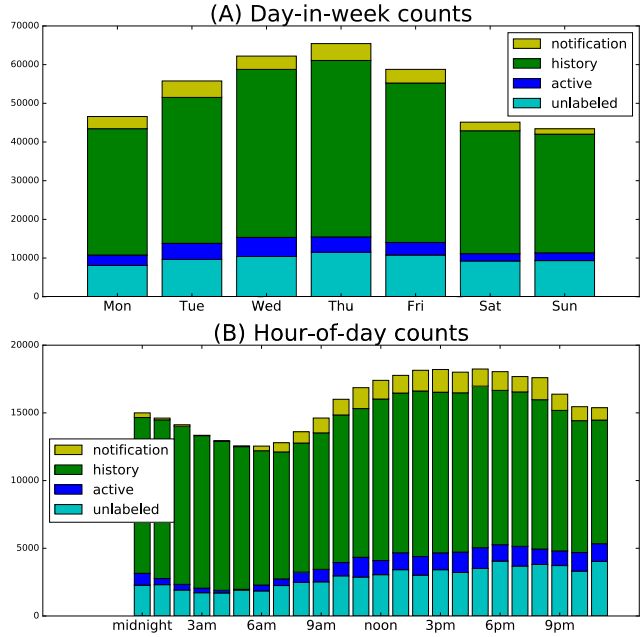

Figure 8: Label-reporting mechanisms over time. The color bars illustrate how many recorded minutes were unlabeled and how many were labeled via the different reporting mechanisms, across (A) days of the week and (B) hours of the day. History contributed the majority of labeled minutes in all days and hours. Night-time was covered very little by in-situ reporting (active-feedback or notification).

on data-collection more during the work week and less in the weekends. Similarly, more data was recorded in the afternoon and evening hours. These peak hours also show increased usage of notifications and active-feedback. This makes sense, given that people are not available to interact with the app while they sleep, but they can report about it later through the history. History covered the vast majority of labeled minutes in all days of the week and all hours of the day.

Users had different approaches to reporting labels. Figure 9 shows labeling patterns (including the label-reporting mechanism and the reported label combinations) from two example users, over their entire participation periods. The first user (top subplot) tended to use the history page much more than notifications or active-feedback. The reported labels seem to form long time-segments of continuous context (especially at nights, when the context involved "lying down" and "sleeping"). The reported daytimes were comprised of long, continuous contexts, like "sitting" and "at school", with additional details that changed more frequently, like short periods of walking or changing phone positions (sometimes in the pocket, sometimes on table). The second user (bottom subplot) used active-feedback more than history. The labeling of this user is comprised of many short time-segments and labels that change more frequently (compared to the first user).

Qualitative analysis

In this section, we summarize the feedback gathered in the surveys that the users filled after their use of the ExtraSensory App, and highlight common themes with selected quotes (with participant number in brackets). Among the questions, we asked each user to discuss their preferred and least preferred method of reporting labels, among active-feedback, history, and notifications.

Using active-feedback. Nine users selected active-feedback as their preferred method. A common reason was that it was "more accurate" or as one stated: *"It is easier to say what I'm about to do than try to recall what I did"* [P20]. Some of these users also stated that it was easier for them to remember to provide labels using active-feedback.

Nine other users had an opposite view of active-feedback and selected it as their least preferred method. Some explained it was hard for them to predict exactly what they were about to do or how long it will take them; one stated *"Most of my activities are spontaneous. I found myself edit again what I reported in active-feedback"* [P6, sic]. Some users said they were too busy to use active-feedback or simply forgot to use it. In addition, active-feedback lacked desirable features that other mechanisms had, including adapting to changes in activity and chronological view of the whole day.

Using the history page. 31 users preferred the history page for reporting labels. Multiple themes arose from these users, highlighting different features of the history page:

- Server-guesses. *"Easier to see and confirm or change what was predicted"* [P45].
- Batch-report. *"Could batch-edit entries and change tasks easier"* [P51]; *"You can combine intervals, which made it easy for me"* [P56].
- Free-time interaction. *"I didn't have to be constantly on my phone. Instead I could report activities all at once"* [P23].
- Reporting fine details. *"I felt I could really pin down everything, even if it required more time to accomplish"* [P21]; *"My activity was very varied minute by minute so it was easier to adjust my data"* [P46].
- Easier to recall. *"Most convenient for retrieving relevant memories"* [P27]; *"I do almost the same things everyday so it's convenient looking through history view"* [P34]; *"It was easiest to record my activity when all the data was in front of me"* [P50]; *"I was able to manage my time by viewing history"* [P54]; *"Much easier to remember what I'd done within the view of past events"* [P58].

For 13 users, the history page was the least preferred mechanism. The most prevalent reason was that it is *"less accurate"* [P1], with specific explanations like: *"Hard to remember precise minute labels"* [P33]; *"I wasn't always certain on the minutes and changed activities so I was afraid to give incorrect data"* [P18]; *"Hard to remember when I was doing a lot of things quickly and could not use active feedback"* [P39]. Some users did not like the interaction with the history interface, stating *"Minute-by-minute is too specific. Every five minutes would be better"* [P5]; *"tedious and inconvenient"* [P48]; or *"takes forever because it guessed something different every minute and the guesses were rarely accurate"* [P60]. These inconveniences were partly due to the real-time classifier on the server that was very basic (trained on little data from only two iPhone users); in sedentary behavior it sometimes al-

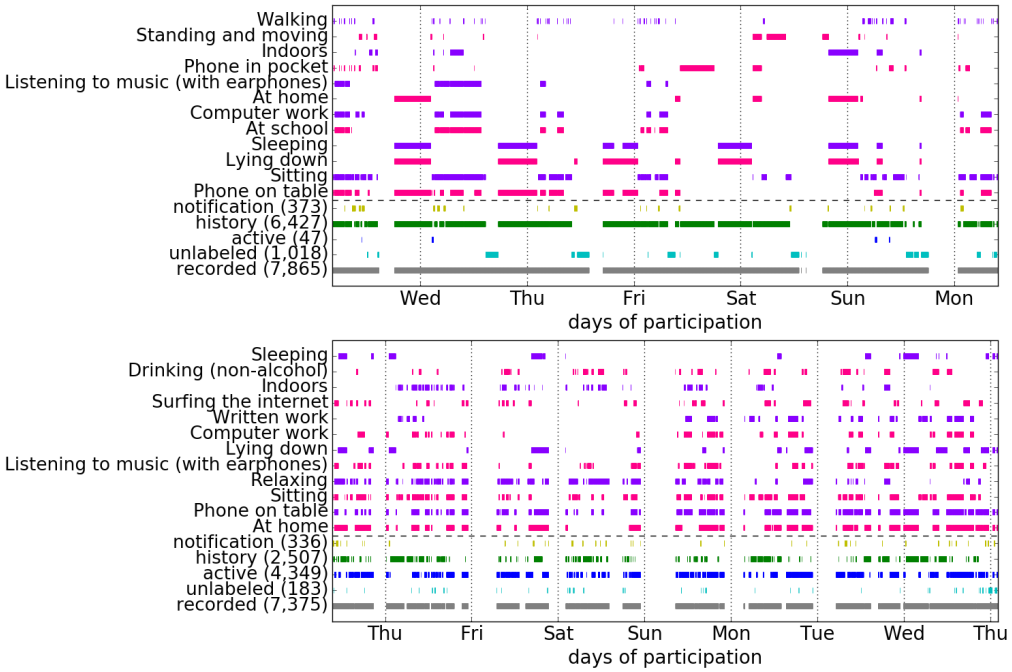

Figure 9: Label reporting. Each subplot shows the participation time of a single user. The bottom "recorded" row shows when data-collection was on. The next four rows show for each minute which reporting mechanism was used to report labels for this minute (or if it is unlabeled). For the rows below the horizontal dashed line, the count of relevant minutes is presented in parenthesis. The rows above the horizontal dashed line show the user's most commonly used labels, with the color bars indicating the minutes for which each label was reported. The vertical dotted grid lines indicate the time 6AM in the participation days. The two users demonstrate different styles of label reporting (long time-segments *vs.* fluctuating contexts).

ternated between guessing "sitting" and "lying down", which made consecutive minutes appear in the history as separate, un-merged events.

Using notifications. Seven users noted notification as their preferred label-reporting method. Most of them specifically referred to confirmation-notifications (asking whether their context remained the same): *"I mostly do the same thing for extended periods of time and the watch made it easy to respond without using my phone"* [P36]; *"Watch-confirm was very easy to press and confirm"* [P8]; *"When doing the same thing it is less intrusive"* [P28]; *"I mostly forgot or was too busy to change (correct) the context at the moment, but when the question was accurate the notification helped"* [P49].

Three users selected notification as both preferred and least preferred; P26 explained preferring it *"because I had the chance to remember to update the app"* and least preferring it *"because it was stressful."* 34 users least preferred the notification. These users did not benefit from the notifications as a reminder; many of them explained that when they were not reporting labels for a while it was because they were too busy (not because they forgot to report). Some complained about the timing of notifications being inopportune, too frequent, and not based on changes in behavior. This caused negative perception of notifications, ranging from low utility (*"Felt rather useless"* [P11]; *"Not needed much thanks to history view"* [P59]) to different levels of annoyance and stress (*"Somewhat annoying, especially while I was working"* [P41]; *"It was disturbing sometimes"* [P54]; *"Extremely annoying and I find them unkind"* [P29]; *"It was stressful"* [P26]).

Mixed preferences. Most users actually utilized the combination of reporting mechanisms and some reported two mechanisms as preferred, like five users who selected both active-feedback and history, explaining *"active for short duration events, history for long duration events"* [P14] or *"used active to enter main activities and later whenever I got home I filled the secondary details in history"* [P57].

Uncomfortable situations. We asked the users "Were there any situations where you did not feel comfortable using the app?" Most users (33) answered "no" (one specified *"no. I was open to use the app anywhere I was"* [P13]). From the users who did raise issues, several themes arose:

- Distraction from work. *"My supervisor knows about my participation, but otherwise it would be a problem to be distracted with the phone during work"* [P4]; *"during exam time, notifications were disturbing"* [P48].
- Social politeness. *"In a meeting or out with friends, because did not want to be rude"* [P24]; *"in class — notifications would vibrate loudly"* [P31].
- Privacy concerns. *"conversations felt a bit strange when I knew it was recorded"* [P11]; *"during intimate settings (sex), but other than that no problem"* [P21]; *"I don't usually have my phone in my hand or use my phone at all times, so documenting everything was a little invasive to me"* [P39].

- Practical inconvenience. *"During the weekend, too busy to tag all the activity labels"* [P5]; *"going to sleep — kept getting notifications and had urge to either update or deal with the notification"* [P30]. For some users, the inconvenience was physical so they avoided using the watch or the app altogether: *"some nights, while sleeping, I put off the watch since it was not very comfortable for me"* [P59]; *"sleeping (I don't like sleeping with accessories)"* [P49]; *"didn't use it during the race because it is less aerodynamic"* [P60].

The label menus. To assess the coverage of behaviors that we pre-defined in the label menus, we asked "Were there any situations where you did not know what labels to select? Are there any labels you think are missing from the lists on the app?" 26 replied "no" (some described the label lists as *"comprehensive"* [P4,P6]). Many users suggested specific activities that were missing (*e.g.* "brushing teeth" or "playing squash"), but they could find more general labels (*e.g.* "grooming" or "exercising", respectively). Some of the suggested labels never crossed our minds when we initially composed the label menus, like "my kids are using my phone" or differentiating between being "in class" as a student *vs.* as a teacher.

After getting such feedback, we added new labels to the "secondary activities" menu, so they became available to the following participants. Also, after about thirty participants, we noticed reoccurring cases of users suggesting labels they could not find although these labels were already in the menu; P30 also mentioned that the topics in the side-bar index were not clear/intuitive. Following that realization, we decided to adjust our protocol for the first meeting and dedicate a few minutes for the new participant to go over the "secondary activities" menu, including looking at the index-topics of the menu. The purpose was to keep the list in the back of their minds, possibly already noticing specific labels that they are likely to use, so that during the study it would be quick and easy for them to find the relevant labels.

In the "main activity" menu, the two versions of standing — "standing in place" and "standing and moving" — caused some confusion; some users described selecting "standing and moving" in situations that involved alternating between standing and shifting from place to place; some said it was hard to distinguish "standing and moving" from "walking." One user indicated that the list did not cover all postures, lacking "crouching" or "kneeling" and some users expected labels like "driving" and "skateboarding" to be in the "main activity" menu instead of the "secondary activities" menu.

DISCUSSION

Collecting self-reported behavioral data in-the-wild raises the challenge of how to get plenty, detailed, and reliable labels, with minimal interference to natural behavior? We tried to overcome this challenge with the design of ExtraSensory's flexible label-reporting interface. In this section, we discuss how our solution addressed data-collection trade-offs and outline how this can guide future designs of in-the-wild studies.

Behavioral time *vs.* interaction time. To cover plenty of behavior-time with little interaction-time, we provided mechanisms that allow reporting labels for time segments of variable durations, either in-situ (up to 30 minutes in the future or 20 minutes in the past) or by recall (for today or yesterday). Indeed, the users utilized batch-reporting of whole time segments, and managed to provide labels for the vast majority of their recorded time. Unfortunately, we did not log the duration of interaction with the app, so it is hard to measure how efficient the reporting-interface was.

Detailed labeling in-situ. We asked users for *detailed* labeling (with diverse aspects of behavior). In-situ reporting has a trade-off between labeling with detail and interfering with natural behavior (it takes time to add all the relevant specific labels). This was, however, mitigated by the confirmation-notifications, especially in cases when the recent context remained the same and the user could easily and quickly respond "correct" (either on the phone or the watch). The difficulty in entering detailed-labels was also mitigated by the "frequently-used" link, which made it easier to find labels after a few days. Users who liked the history page also used it in-situ (reporting about a few minutes ago).

Reliable labels with recall. Recall-based reporting has the risk of poorer reliability of the reported labels. To mitigate this problem, the history page combined various features to help the user recall their past context, including server guesses, visually organizing the day chronologically, and the multi-label details in the labeled events. These features, along with the ability to quickly cover long segments of time, made the history page by far the users' most preferred mechanism, which also yielded the majority of labeled minutes.

Flexible interface. Multiple reporting mechanisms significantly contributed to the labeled data, throughout all days of the week and most hours of the day. Active-feedback was used to track quick or temporary changes in behavior, like switching posture or going to the restroom. Some users used active-feedback to mark time in-situ, and later went over the history page to fill in the details. Using the watch was very popular for confirmation-notifications: while users could also use the history page for such reporting, pressing a watch button is much less intrusive. The flexibility in the interface helped cover different situations and the results show interaction patterns with mixed mechanisms. In addition, the surveys confirm that this flexibility was important to engage users with different preferences and styles of daily behavior.

Open-ended notification perceived negatively. The reporting option of open-ended notification was especially disliked, and correspondingly, it yielded very few labeled minutes. With the advantage of the other mechanisms, there was little use for a blank-notification that acts merely as a reminder, comes at inopportune timings, and sometimes causes stress.

Types of users. Some users were very meticulous and wanted to provide the best data, so they made sure to keep reporting up-to-date labels. While this contributed much labeled data, this may have also affected the authentic nature of their behavior (it is typically not natural for a person to interact with such an app every few minutes). Other users dedicated less effort to labeling, so they contributed less detailed or less accurate labeling, but their recorded behavior was more authentic. The

frequently-used labels section seems to have eased this trade-off after a few days of participation.
Our validation users were mostly students and university workers. People with attention-demanding jobs (*e.g.* child care, construction) may tend to use in-situ reporting less but can still report by-recall using the history page. However, generalization to other occupations has to be validated empirically.

Label structure. The labeling form we designed was semi-structured: it had dedicated fields for body-state (main activity) and moods, but it also included the less structured secondary-activities menu. While forcing a single body-state value helped generate a consistent behavioral dimension, it had some disadvantages: in some situations it was confusing and users were not sure which value to select; also, forcing a dedicated field for this dimension made interaction sometimes inconvenient (more clicks), especially in cases the user did not remember the exact body-state. On the other hand, the multi-label approach that we used in the secondary-activities menu may produce labeling that is inconsistent (a user may accidentally mark both "indoors" and "outside") or incomplete (*e.g.* reporting activity-labels while ignoring environment-labels). However, multi-label can be more convenient and it enables reporting simultaneous activities (like watching TV while eating) and situations that the researcher did not have in mind — thus promoting individual authentic behavior.

Label accuracy. When relying solely on self-reporting, the labeling may be noisy and there is no direct way to assess how accurate are the labels that participants report about themselves. Furthermore, the accuracy of the labels can be inconsistent, because of multiple mechanisms of label reporting and users with different levels of rigorousness. In our previous works [27, 28], we specified simple methods to clean the labels and demonstrated successful machine learning experiments with the data. In [28], we specifically addressed training classifiers with in-the-wild data, which may be inconsistent and have highly unbalanced labeling and occasions of missing labels or sensors. The results were encouraging and facilitate future data collections that are less strict and easier on the participants.

Revised ExtranSensory App

Following our experience in data collection and the user-experience feedback, we revised the mobile application to addresses some of the insights raised in this discussion. We make this revised version of our mobile app publicly available at `http://extrasensory.ucsd.edu/ExtraSensoryApp`. This public code package includes the full code to the Android phone app, the Pebble watch component, and the server-side code, as well as a fully detailed guide for users and for researchers. We provide this package to allow other researchers to use the app (with or without adjusting its code) for further data collection.

In order to provide better inference, this revised version is based on a server-side real-time classifier that is now trained on the full data from the 60 users, so its guesses are now more accurate and more detailed — specifying probability values for 51 diverse context-labels. These detailed guesses now appear on the history, providing more cues to recall the true context. The predicted labels also appear sorted by probability in a new "server guess" section in the secondary-activities menu, making it easier to find the relevant labels to select. The researcher can decide to combine all the study's labels (including body-states and moods) in the secondary-activities menu to make the interaction flow easier and allow for more subjective definitions of the user's behavior. Researchers can also easily edit a text file to customize the label menu and its organization by topics.

In addition, the ExtraSensory App now allows disabling open-ended or confirmation notifications, selecting which sensors to record (to reduce battery consumption and communication), selecting classifier to be used on the server, and more.

Future directions

More detailed logging of the user-app interaction can improve assessment of time-efficiency of the interface and of accuracy of the reported labels. Daily self-audit can help users correct their own labeling mistakes. Light-weight body-worn cameras and user-taken phone pictures can augment the self-reporting to help users remember their past context and to allow for external validation of label accuracy.

Further improvements can come from enhancing features of our app, like utilizing real-time guesses to cleverly trigger opportune notifications, or adding more functionality to the watch. Speech-recognition engines will enable voice dictation self-reporting with structured instructions (*e.g.* "start: running, with pet, at the beach, valid for: 30 minutes") and eventually intelligently processed free text (*e.g.* "I'm taking Barkley out for a walk around the block"). These additions and more will reduce the load on users and make interaction smoother and less intrusive.

CONCLUSION

In this paper, we introduce the *ExtraSensory App*, a mobile application for collecting data in-the-wild, including sensor measurements and self-reported detailed labels of behavioral context. We validated this app in an in-the-wild study with 60 users. The app's rich label-reporting interface was important to engage users with different behavior styles and phone-interaction preferences and to acquire detailed labels for over 300,000 minutes of diverse behavioral contexts.

ExtraSensory's history page showed to be very useful and the features it offered helped users recall their past context. The additional watch component turned out to be very helpful to keep the user-interaction from interfering with natural behavior. To maximize the utility of the watch, its prompts should be cleverly timed and require minimal reaction (single button press). Ongoing data collection and re-training of real-time classifiers will improve the server guesses and notifications and make user interaction easier and less time consuming.

We believe that the insights that we describe in this paper will inspire future designs of in-the-wild data-collection studies. The public version of the ExtraSensory App that we provide will allow for further collections of data and studies that use real-time context-recognition in-the-wild for various applications in health-monitoring, aging-care, and other domains.

REFERENCES

1. Frank Bentley and Konrad Tollmar. 2013. The Power of Mobile Notifications to Increase Wellbeing Logging Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1095–1098. DOI: http://dx.doi.org/10.1145/2470654.2466140
2. Daniel Castro, Steven Hickson, Vinay Bettadapura, Edison Thomaz, Gregory Abowd, Henrik Christensen, and Irfan Essa. 2015. Predicting daily activities from egocentric images using deep learning. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 75–82.
3. Tanzeem Choudhury, Sunny Consolvo, Beverly Harrison, Jeffrey Hightower, Anthony LaMarca, Louis LeGrand, Ali Rahimi, Adam Rea, G Bordello, Bruce Hemingway, and others. 2008. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing* 7, 2 (2008).
4. Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, and others. 2008. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1797–1806.
5. Mihaly Csikszentmihalyi and Reed Larson. 2014. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*. Springer, 35–54.
6. Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover. 2014. Detecting periods of eating during free-living by tracking wrist motion. *IEEE journal of biomedical and health informatics* 18, 4 (2014), 1253–1260.
7. Katherine Ellis, Suneeta Godbole, Jacqueline Kerr, and Gert Lanckriet. 2014. Multi-Sensor physical activity recognition in free-living. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 431–440.
8. Raghu Kiran Ganti, Soundararajan Srinivasan, and Aca Gacic. 2010. Multisensor fusion in smartphones for lifestyle monitoring. In *2010 International Conference on Body Sensor Networks*. IEEE, 36–43.
9. John J Guiry, Pepijn van de Ven, and John Nelson. 2014. Multi-Sensor Fusion for Enhanced Contextual Awareness of Everyday Activities with Ubiquitous Devices. *Sensors* 14 (2014), 5687–5701.
10. Erik B Hekler, Predrag Klasnja, Vicente Traver, and Monique Hendriks. 2013. Realizing effective behavioral management of health: the metamorphosis of behavioral science methods. *IEEE pulse* 4, 5 (2013), 29–34.
11. Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780.
12. Jacqueline Kerr, Ruth E Patterson, Katherine Ellis, Suneeta Godbole, Eileen Johnson, Gert Lanckriet, and John Staudenmayer. 2016. Objective Assessment of Physical Activity: Classifiers for Public Health. *Medicine and science in sports and exercise* 48, 5 (2016), 951–957.
13. Adil Mehmood Khan, Ali Tufail, Asad Masood Khattak, and Teemu H Laine. 2014. Activity recognition on smartphones via sensor-fusion and kda-based svms. *International Journal of Distributed Sensor Networks* 2014 (2014).
14. Abby C King, Eric B Hekler, Lauren A Grieco, Sandra J Winter, Jylana L Sheats, Matthew P Buman, Banny Banerjee, Thomas N Robinson, and Jesse Cirimele. 2013. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PloS one* 8, 4 (2013), e62613.
15. Kai Kunze and Paul Lukowicz. 2014. Sensor placement variations in wearable activity recognition. *IEEE Pervasive Computing* 13, 4 (2014), 32–41.
16. Matthew L Lee and Anind K Dey. 2015. Sensor-based observations of daily living for aging in place. *Personal and Ubiquitous Computing* 19, 1 (2015), 27–43.
17. Beth Logan and others. 2000. Mel Frequency Cepstral Coefficients for Music Modeling.. In *ISMIR*.
18. Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics Can'T Focus: An in Situ Study of Online Multitasking in the Workplace. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1739–1744. DOI: http://dx.doi.org/10.1145/2858036.2858202
19. Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1021–1032. DOI: http://dx.doi.org/10.1145/2858036.2858566
20. Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2847–2854.
21. Susanna Pirttikangas, Kaori Fujinami, and Tatsuo Nakajima. 2006. Feature selection and activity recognition from wearable sensors. In *International Symposium on Ubiquitious Computing Systems*. Springer, 516–527.

22. Tauhidur Rahman, Mi Zhang, Stephen Voida, and Tanzeem Choudhury. 2014. Towards Accurate Non-Intrusive Recollection of Stress Levels Using Mobile Sensing and Contextual Recall. In *International Conference on Pervasive Computing Technologies for Healthcare*.

23. Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180.

24. Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart Devices are Different: Assessing and MitigatingMobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 127–140.

25. Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015a. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1029–1040.

26. Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd. 2015b. Inferring Meal Eating Activities in Real World Settings from Ambient Sounds: A Feasibility Study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 427–431.

27. Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. 2017a. Recognizing Detailed Human Context In-the-Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing* 16, 4 (2017). DOI: **http://dx.doi.org/10.1109/MPRV.2017.3971131**

28. Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2017b. Behavioral Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification. *Under review for Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2017).

29. Evan Welbourne and Emmanuel Munguia Tapia. 2014. CrowdSignals: a call to crowdfund the community's largest mobile dataset. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 873–877.

30. Hao Yan and Ted Selker. 2000. Context-aware office assistant. In *Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, 276–279.

31. Yi-Hsuan Yang and Yuan-Ching Teng. 2015. Quantitative Study of Music Listening Behavior in a Smartphone Context. *ACM Trans. Interact. Intell. Syst.* 5, 3, Article 14 (Sept. 2015), 30 pages. DOI: **http://dx.doi.org/10.1145/2738220**