Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches

To promote context recognition in the wild, the authors set out to capture people's authentic behavior in their natural environments using everyday devices—smartphones and smartwatches. The authors address difficulties related to in-the-wild conditions and show how multimodal sensors can help.

Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet
University of California, San Diego

The ability to automatically recognize a person's behavioral context (including where they are, what they're doing, and who they're with) is greatly beneficial in many domains. Health monitoring applications have traditionally been based on manual, subjective reporting,[1] sometimes using end-of-day recalling.[2] These applications could be improved with the automatic (frequent, effortless, and objective) detection of behaviors, especially behaviors related to exercise, diet, sleep, social interaction, and mental states (such as stress). Aging-care programs could use automated logging of older adults' behavior to detect early signs of cognitive impairment, monitor functional independence, and support aging at home.[3]

Similarly, just-in-time interventions could benefit from automatic context recognition. Such interventions often prompt patients at arbitrary times of day, missing when the patient is most in need of support.[4] A system that could detect such needs could intervene at more appropriate times. For example, an alcoholic patient might be most at risk of relapse when the context is "at a bar, with friends."

The biomedical research community has acknowledged the effects of behavior, lifestyle, and environment on health, disease, and treatments.[5] Automatic context-recognition tools are essential for incorporating behavioral aspects and exposure into large-scale studies and for tailoring treatment options to address individual patient needs. The range of measured exposures should be broad, covering diverse lifestyle and environmental conditions. Commercial tools that offer superficial recognition (of walking, running, and driving, for example) will not suffice.

For all of these different applications to succeed on a larger scale, the context-recognition component must be unobtrusive and work smoothly, without making people adjust their behaviors. Research thus must emulate the real-world settings in which such applications will eventually be deployed. Here, we promote context recognition in the wild, capturing people's authentic behavior in their natural environments using everyday devices. In particular, we use smartphone and smartwatch sensors to recognize detailed situations of people performing

Published by the IEEE CS ■ 1536-1268/17/$33.00 © 2017 IEEE

their natural behaviors, to explore the challenges presented by in-the-wild conditions, and to show how multimodal sensors can help.

Promoting Real-Life Applications

People usually have their phones close to them.[6] This growing trend, combined with a variety of built-in sensors, makes phones popular agents for automatically recognizing human behavior. Smartwatches are also a useful sensing tool, because they can capture informative signals about hand and arm motion, yet they're natural to wear and don't additionally burden the user.

Previous works have shown the advantage of fusing sensors of different modalities from smartphones and smartwatches to improve the recognition of basic movement activities[7] and of more complex activities, such as smoking or drinking coffee.[8] However, most past works have collected data under heavily controlled conditions, with researchers instructing subjects to perform scripted tasks. Fitting models to recognize prescribed activities can result in poor generalization to real-life scenarios.[9]

Understanding In-the-Wild Conditions

To promote real-life working applications, we argue that research must be done in natural and realistic settings, satisfying the following in-the-wild conditions.

Naturally used devices. Introducing a foreign device to the user adds a burden and can affect natural behavior. Ideally, subjects should use their own phone and possibly additional convenient devices, such as watches.

Unconstrained device placement. Sensor placement and orientation greatly influence recognition success.[10] However, this doesn't mean we should force specific placements; a practical real-world application shouldn't require the phone to remain in the user's pocket for proper recognition. Instead, research should address the variability in device placement and find ways to overcome it.

Natural environment. The recorded behavior should be in the subjects' natural environment and on their own schedule. Subjects shouldn't be instructed where or when to perform the activity.

Natural behavioral content. Researchers often instruct subjects to perform scripted tasks,[7,8] leading to simulated rather than natural behavior. Others let the subjects behave on their own time but still prescribe a list of targeted activities,[11,12] which can cause the subject to perform actions they wouldn't typically perform, such as "vacuum cleaning." In-the-wild studies should record behavior that is natural to each individual subject.

Our Work

A major challenge is acquiring labels of the behavioral context. Attaining in-the-wild conditions usually trades off with other aspects of the data collection effort, resulting in fewer labeled examples, a smaller range of interest labels, or compromised privacy of the subjects.

Previous research has addressed some aspects of in-the-wild data collection in different ways (see the "Related Work in Data Collection" sidebar). In this work, we used smartphone and smartwatch sensors to collect labeled data from over 300,000 minutes from 60 subjects. Every minute has multisensor measurements and is annotated with relevant context labels. To the best of our knowledge, this dataset, which is publicly available, is far larger in scale than others collected in the field.

Similar to earlier works,[11,13,14] we relied on self-reporting. Unlike those works, however, our data collection app offered an extensive menu of more than 100 context labels and the ability to select combinations of relevant labels. This facilitated natural behavior—for example, subjects were free to "run on a treadmill" while "watching TV," rather than being forced to choose only one activity.[11,14] This also allowed for rich descriptions of context through combinations of different aspects—such as the environment, activities, company, and body posture.

Similar multi-aspect representations were previously used to describe objects and actions in images[12] and to identify locations, objects, humans, and animals in sound clips.[15] In those cases, annotation was done offline, but in our case, attaining detailed labeling through self-reporting required attention and effort from the subjects. To mitigate this, our app's interface offered many reporting mechanisms to minimize interaction time: subjects could report the start of an activity (as in other projects[11,13,14]) and could manually edit events in a daily calendar that included automatically recognized contexts (similar to work by Sunny Consolvo and her colleagues[16]). We treated only the manual corrections or additions as ground truth; the automated real-time predictions acted as cues to help the subjects recall the context (as suggested by Tauhidur Rahman[17]).

The main contribution of this work is our emphasis on the in-the-wild conditions mentioned earlier:

- Naturally used devices—subjects used their own personal phones and a smartwatch that we provided.
- Unconstrained device placement—subjects were free to carry their phones in any way convenient to them.
- Natural environment—subjects collected data in their own regular environment for approximately one week.
- Natural behavioral content—no script or tasks were given, and we didn't target a specific set of activities. Instead, the context labels we analyzed came from the data—the subjects engaged in their routines and

Related Work in Data Collection

Researchers have addressed in-the-wild data collection in different ways. Manhyung Han and his colleagues designed a decision-tree architecture that activates predetermined sensors to differentiate eight ambulatory and transportation states.[1] Such a hand-crafted system is hard to scale to more contexts. They validated their system with an observer that followed a single user.

Javier Ordonez and his colleagues installed a set of state-change sensors around a home to detect daily home activities.[2] Although such sensors are unobtrusive and maintain natural behavior, the complicated device setup limits the deployment of data collection and practical applications. It also cannot track the person outside of the monitored environment.

Yujie Dong and his colleagues targeted eating periods and used an unnatural setup of having a smartphone bound to the wrist.[3] Subjects had to mark start times of eating and, after data collection, review and correct their markings. This resulted in 449 hours of data with 116 eating periods from 43 subjects.

Tauhidur Rahman and his colleagues compared different approaches for subjects to self-report their stress level (immediately or via recall).[4] They suggested a compromise approach where the subjects could report on their own time but with the help of cues (such as the location or surrounding sound level) to remember how they felt at specific times of day.

Tanzeem Choudhury and her colleagues designed a system to address the requirements for a practical context recognition system, including unobtrusive lightweight devices, long battery life, and multimodal sensing.[5] However, most of their validation was done on controlled data, collected in specific locations, with constrained positioning of device, and with a sequence of eight activities that were scripted, observed, and repeated by 12 subjects. Sunny Consolvo and her colleagues used the same system (trained on the controlled data) in a field study of an application to promote physical activity.[6] The mobile app used a combination of the automated recognition (of walking, cycling, and so on) and manual editing of a daily journal.

Samuli Hemminki and his colleagues targeted transportation modes and specifically designed features that would be less sensitive to phone placement.[7]

Raghu Ganti and his colleagues gave eight subjects a Nokia N95 phone for eight weeks and asked them to go about their regular routines and use the phone for recording whenever they could, in any location or time of day.[8] The phone was constrained to a pocket or pouch. The interface allowed selecting an activity from a set of eight activities and marking when the activities started and finished. They collected a total of 80 hours. Adil Khan and his colleagues targeted 15 activities.[9] To collect measurements and annotations, they handed a NEXUS phone to subjects for a month. Subjects were free to perform the prescribed activities on their own time and they used the phone to mark the beginning and end of the selected activity. They collected approximately 3,000 examples per activity from 30 subjects, plus a follow-up validation with eight subjects using the trained real-time recognition system.

Two of us (Katherine Ellis and Gert Lanckriet) worked with colleagues to collect data from 40 subjects who recorded hip-mounted accelerometer and GPS data from routine behavior in a natural environment for several days.[10] The subjects wore a SenseCam device around their neck, which periodically took snapshots of the scene. The thousands of images were later used by research assistants to annotate the activity. Hamed Pirsiavash and his colleagues used a GoPro video camera for both sensor measurements and ground truth labels.[11] The subjects wore the device around the chest for a single morning at their own home, and were prescribed a list of home activities to perform with no extra specifications. They recorded more than 10 hours of video from 12 people and later annotated household objects and activities for approximately 30,000 frames (every second). Although the camera approach might generate more reliable labels in certain cases, the unnatural and uncomfortable equipment compromises natural behavior. Furthermore, offline annotation of images is costly, making it hard to scale, and it violates the privacy of the subjects and people around them. The alternative—self-reporting—has the advantage of collecting labels when a camera isn't present (such as in the shower), when the context isn't clearly visible in the image (such as when the user is singing), or when the subject knows best what is happening (such as whether he or she is with family or friends).

REFERENCES

1. M. Han et al., "Comprehensive Context Recognizer based on Multimodal Sensors in a Smartphone," *Sensors*, vol. 12, no. 9, 2012, pp. 12588–12605.
2. F.J. Ordonez, P. de Toledo, and A. Sanchis, "Activity Recognition Using Hybrid Generative/Discriminative Models on Home Environments Using Binary Sensors," *Sensors*, vol. 13, no. 5, 2013, pp. 5460–5477.
3. Y. Dong et al., "Detecting Periods of Eating During Free-Living by Tracking Wrist Motion," *IEEE J. Biomedical and Health Informatics*, vol. 18, no. 4, 2014, pp. 1253–1260.
4. T. Rahman et al., "Towards Accurate Non-Intrusive Recollection of Stress Levels Using Mobile Sensing and Contextual Recall," *Proc. Int'l Conf. Pervasive Computing Technologies for Healthcare*, 2014, pp. 166–169.
5. T. Choudhury et al., "The Mobile Sensing Platform: An Embedded Activity Recognition System," *IEEE Pervasive Computing*, vol. 7, no. 2, 2008, pp. 32–41.

6. S. Consolvo et al., “Activity Sensing in the Wild: A Field Trial of Ubifit Garden,” *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2008, pp. 1797–1806.

7. S. Hemminki, P. Nurmi, and S. Tarkoma, “Accelerometer-Based Transportation Mode Detection on Smartphones,” *Proc. 11th ACM Conf. Embedded Networked Sensor Systems*, 2013, article no. 13.

8. R.K. Ganti, S. Srinivasan, and A. Gacic, “Multisensor Fusion in Smartphones for Lifestyle Monitoring,” *Proc. 2010 Int’l Conf. Body Sensor Networks*, 2010, pp. 36–43.

9. A.M. Khan et al., “Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs,” *Int’l J. Distributed Sensor Networks*, vol. 10, no. 5, 2014; doi: 10.1155/2014/503291.

10. K. Ellis et al., “Multi-Sensor Physical Activity Recognition in Free-Living,” *Proc. 2014 ACM Int’l Joint Conf. Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 431–440.

11. H. Pirsiavash and D. Ramanan, “Detecting Activities of Daily Living in First-Person Camera Views,” *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), 2012, pp. 2847–2854.

selected any relevant labels (from the large menu) that fit what they were doing.

Recognizing context in the wild is more challenging compared to controlled conditions because of the large variability in real life. Diversity in phone devices and sensor hardware affects the measurements.[18] Our data came from both iPhone and Android devices, with a wide variety of models. Variability in behavioral content was clearly visible in the ground truth labels of our data, including, for example, the following combinations:

- {running, outside, exercise, talking, with friends},
- {running, indoors, exercise, at the gym, phone on table},
- {sitting, indoors, at home, watching TV, eating, phone on table},
- {sitting, at a restaurant, drinking (alcohol), talking, eating},
- {sitting, on a bus, phone in pocket, talking, with friends}, and
- {on a bus, standing}.

Such variability was missed in works that defined behavior with a small set of mutually exclusive activities. Variability in manner or style (such as different gaits) was less visible but still captured in our sensor measurements. Such variability could easily be missed in scripted experiments or with restrictions on how devices should be used. Our analysis demonstrates the difficulty in resolving context in the wild and the importance of using complementary sensing modalities. We show that everyday devices, in their natural usage, can capture information about a wide range of behavioral attributes.

The Context Recognition System

Figure 1 illustrates the flow of our recognition system. The system is based on measurements from five sensors in a smartphone: accelerometer, gyroscope, location, audio, and phone-state sensors, as well as accelerometer measurements from a smartwatch. For a given minute, the system samples measurements from these six sensors, and the task is to detect the combination of relevant context labels (see Figure 1a)—that is, the system must declare for each label l a binary decision: $y_l = 1$ (the label is relevant to this minute) or $y_l = 0$ (not relevant).

For this article, we opted for simple computational methods based on linear classifiers and basic heuristics for sensor fusion. We modelled each label separately and treated every minute as an independent example. We included the time of day as part of the phone-state features, but we didn’t model the behavioral time series throughout the day. Our goal was to show the potential of context recognition in the wild and establish a baseline. Future work will use nonlinear methods, dynamic-context models, and interaction among labels.

Single-Sensor Classifiers

Single-sensor classifiers use sensor-specific features and help us understand how informative each sensor can be, independent of the other sensors, for a given context label (Figure 1b). We used logistic regression—a linear classifier that outputs a continuous value (interpreted as probability) in addition to the binary decision. This is helpful for sensor fusion. The following procedure was performed for a given sensor s and a given label l:

- For each example, compute a d_s-dimensional feature vector x_s. Each sensor has a different set of relevant features.
- Standardize each feature by subtracting the mean and dividing by the standard deviation (these statistics are estimated based on the training set).
- Learn a d_s-dimensional logistic regression classifier from the training set.
- Apply the logistic regression classifier to a test example to obtain a binary classification y_l and probability value $P(y_l = 1 | x_s)$.

To overcome the imbalance between the positive class and negative class, we applied balanced class weights (inversely proportional to the class frequency in the training set).

At this point, it’s possible to introduce some domain knowledge and assign appropriate sensors to certain labels. For example, the watch accelerometer can be a good indicator for specific hand-motion activities, such as “washing dishes,” while audio might

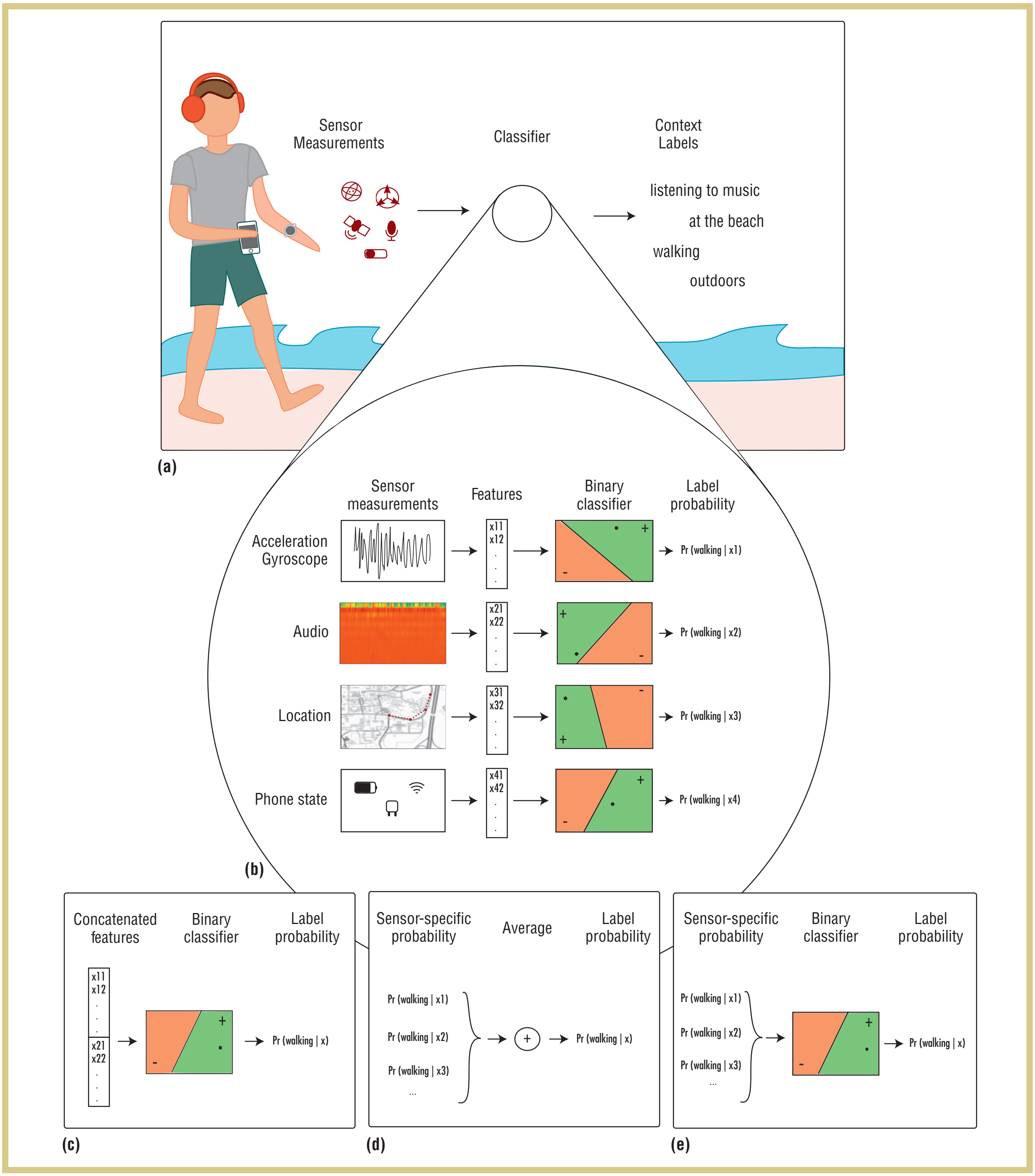


Figure 1. Our context-recognition system. (a) While a person is engaged in natural behavior, the system uses sensor measurements from the smartphone and smartwatch to automatically recognize the person's detailed context. (b) Single-sensor classifiers: Appropriate features are extracted from each sensor. For a given context label, classification can be done based on each sensor independently. (c) In early fusion (EF), features from multiple sensors are concatenated to form a long feature vector. (d) Late fusion using averaging (LFA) simply averages the output probabilities of the single-sensor classifiers. (e) Late fusion with learned weights (LFL) learns how much to "listen" to each sensor when making the final classification.

better predict environmental contexts, such as "in class" or "at a party." These design decisions aren't always obvious, so we used sensor-fusion methods that could learn the best predictors from data.

Sensor Fusion

Our system further combines information from N different sensors, and we propose three alternative methods for such fusion.

The first method is to use *early fusion* (EF) classifiers, which combine information from multiple sensors prior to the classification stage (see Figure 1c). The following procedure was performed for a given label l:

- Start with the sensor-specific feature vectors $\{x_s\}_{s=1}^{N}$.
- Concatenate the (standardized) sensor-specific feature vectors into a single vector $\mathbf{x}$ of dimension $d = \sum_{s=1}^{N} d_s$.
- Learn a d-dimensional logistic regression classifier from the training set.
- Apply the logistic regression classifier to a test example to obtain a binary classification y_l and probability value $P(y_l = 1 | x)$.

We also explored two *late fusion* classifiers. We used ensemble methods to combine the predictions of the N single-sensor classifiers. We chose to combine the probability outputs $P(y_l = 1 | x_s)$, and not the binary decisions, to take into account the "confidence" of each of the N classifiers and to avoid too much influence from irrelevant sensors.

So the second fusion method is *late fusion using average probability* (LFA), shown in Figure 1d. LFA applies a simple bagging heuristic and averages the probability values from all the single-sensor classifiers to obtain a final "probability" value—that is,

$$P\left(y_l = 1 \middle| x_1, x_2, \ldots, x_N\right) = \frac{1}{N} \sum_{s=1}^{N} P\left(y_l = 1 \middle| x_s\right).$$

LFA declares "yes" if the average probability is larger than 0.5. No additional training is performed after the single-sensor classifiers are learned. This method grants equal weight to each sensor, hoping that informative sensors will classify with higher confidence (probability close to 0 or 1) and will influence the final decision more than irrelevant sensors (which will hopefully predict with probability close to 0.5).

As mentioned earlier, some sensors might be consistently better suited for certain labels. As a flexible alternative to deciding a priori how to assign sensors to labels, we can let sensor weights be learned from data. So the third fusion method is to use *late fusion using learned weights* (LFL), shown in Figure 1e. This second type of late fusion places varying weight on each sensor. This method learns a second layer—an N-dimensional logistic regression model. The second layer's input is the N probability outputs of the single-sensor models, and the output is a final decision y_l.

Data Collection

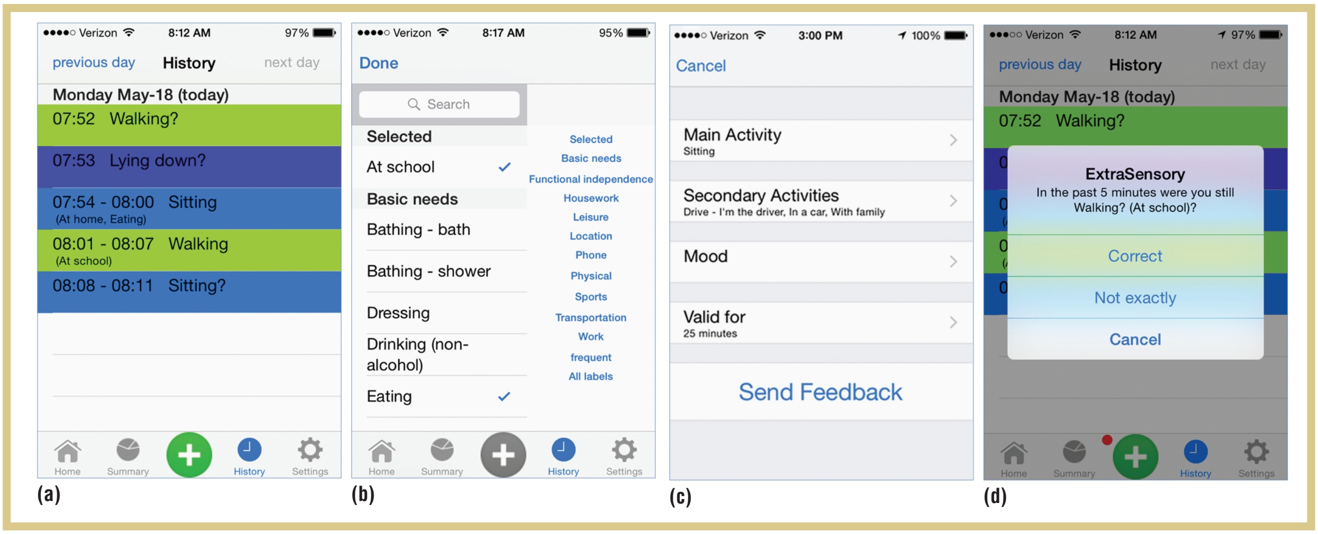
For large-scale data collection, we developed a mobile app called *ExtraSensory App*, with versions for both iPhone and Android smartphones and a companion application for the Pebble smartwatch that integrates with both. We used the app to collect both sensor measurements and ground truth context labels. Every minute, the app recorded a 20-second window of sensor measurements from the phone and watch. Within that window, the time samples of different sensors weren't guaranteed to be exactly aligned. The flexible user interface provided the user with many mechanisms to self-report the relevant context labels and cover long behavioral times with minimal effort and interaction with the app (see Figure 2).

As Figure 2a shows, the history tab showed the user a daily log of activities and contexts. The server would send real-time body-state predictions (based on preliminary training data from two iPhone users—the researchers). These predictions appeared with question marks to help the user organize the log and recall when the activity might have changed. The user could update the history records' labels, add secondary labels (such as "at home" and "eating"), merge consecutive records into a longer period, and split records.

Figure 2b shows the label selection menu, indexed by topics and a "frequently used" link to make it easier for the user to select quickly, and Figure 2c shows the "active feedback" page, which let the user report that he or she was engaging in a specific context (starting immediately and valid for a selected period of time). Figure 2d demonstrates the periodic notifications, which would remind the user to provide labels. If the user was engaged in the same context as recently reported, he or she simply replied "correct." If any element of the context changed, the user could press "not exactly" and be directed to a screen for updating labels of the recent time period. These notifications appeared on the watch as well to enable easier responses.

We recruited 60 subjects (34 female and 26 male users) using fliers posted around the UC San Diego campus and using campus-based email lists. Thirty-four of the subjects were iPhone users (iPhone 4 to iPhone 6; iOS versions 7, 8, and 9). Twenty-six subjects were Android users, with various devices (Samsung, Nexus, Motorola, Sony, HTC, Amazon Fire-Phone, and Plus-One). The subjects were from diverse ethnic backgrounds (self-defined), including Chinese, Mexican, Indian, Caucasian, African-American, and more. Most of the subjects (93 percent) were right-handed and chose to wear the smartwatch on their left wrist, and almost all were students or research assistants. Table 1 presents additional subject characteristics.

We installed the app on each subject's personal phone and provided the watch (56 out of the 60 agreed to wear the watch). The subjects then engaged in



In the past 5 minutes were you still Walking? (At school)?

Figure 2. Screenshots from the *ExtraSensory App* (iPhone version): (a) the history tab, which shows a daily log of activities and contexts as well as predictions (shown with question marks); (b) the label selection menu, indexed by topics and a "frequently used" link; (c) the "active feedback" page, which lets the user report that he or she will be engaged in a specific context, starting immediately and valid for a selected time period; and (d) periodic notifications, which remind the user to provide labels.

TABLE 1
Statistics for the 60 users in the dataset.

	Range	Mean (standard deviation)
Age (years)	18–42	24.7 (5.6)
Height (cm)	145–188	171 (9)
Weight (kg)	50–93	66 (11)
Body mass index (kg/m^2)	18–32	23 (3)
Labeled examples	685–9,706	5,139 (2,332)
Additional unlabeled examples	2–6,218	1,150 (1,246)
Average applied labels per example	1.1–9.7	3.8 (1.4)
Participation duration (days)	2.9–28.1	7.6 (3.2)

their usual behavior for approximately one week, while keeping the app running in the background on their phone as much as possible. The subjects were asked to report as many labels as possible without interfering too much with their natural behavior. They were free to remove the watch whenever they wanted and were asked to turn off the watch app when they weren't wearing the watch. Basic compensation was US$40 for each subject, with an additional incentive of up to $35 depending on the amount of labeled data provided.

The resulting *ExtraSensory dataset* contains 308,320 labeled examples (minutes) from the 60 users. Table 2 specifies details about the sensors recorded (though not all sensors were available at all times). The dataset is publicly available and researchers are encouraged to use it to develop and compare context recognition methods (http://extrasensory.ucsd.edu).

Evaluation and Results

We evaluated classification performance using fivefold cross-validation: each fold had 48 users in the training set; the other 12 users were in the test set. We also conducted leave-one-out (LOO) experiments (leaving one user out). When measuring performance, classification accuracy can be a misleading metric because of imbalanced data; for a rare label that appears in 1 percent of the test set, a trivial classifier that always declares "no" will achieve 99 percent accuracy but is completely useless. It's important to consider competing metrics, such as sensitivity and specificity.

A common approach is to observe sensitivity (recall) against precision or to calculate their harmonic mean (F1). However, precision and F1 are less fitting, because they are very sensitive to rare labels. Chance level of precision or F1 can be arbitrarily small, and when averaging them over many labels,

TABLE 2
The sensors in the dataset. "Core" represents examples that have measurements from all six core sensors analyzed in this article (shown here in bold and italic font).

Sensor	Raw measurements	No. of examples	No. of users
Accelerometer	***3-axis (40Hz)***	***308,306***	***60***
Gyroscope	***3-axis (40Hz)***	***291,883***	***57***
Magnetometer	3-axis (40Hz)	282,527	58
Watch accelerometer	***3-axis (25Hz)***	***210,716***	***56***
Watch Compass	Heading angle (variable*)	126,781	53
Location	***Long-lat-alt (variable*)***	***273,737***	***58***
Location (precomputed)	***Location variability (once per example)***	***263,899***	***58***
Audio	***13 MFCC (46-ms frames)***	***302,177***	***60***
Audio power	Once per example	303,877	60
Phone state	***Once per example***	***308,320***	***60***
Low-frequency sensors	Once per example	308,312	60
Core		176,941	51

*Variable sampling rate—gathering updates whenever the value changes.

certain labels will unfairly dominate the total score. Additionally, the self-reported data might be noisy, possibly including cases where a label was actually relevant but not reported by the subject. Precision and F1 are too sensitive in such cases. Unlike F1, the balanced accuracy, BA = 0.5 (sensitivity + specificity), doesn't suffer from these issues and can serve as a convenient objective that fairly balances two complementary metrics.

We first assessed the potential of single sensors. Table 3 shows some specific context labels for which relatively few examples were collected. Our first (and sometimes second) guess regarding the relevant sensor achieved reasonable context recognition.

Next, to see if we could do better, we evaluated the three sensor-fusion methods discussed earlier (EF, LFA, and LFL) and compared them to the single-sensor classifiers. Figure 3 shows the performance for 25 labels from diverse context domains. In most cases, sensor fusion managed to match the best fitting single sensor. The system learned from data how to best use the different sensors without requiring human guidance, which can be useful for scalable systems, where the researcher doesn't necessarily know which sensor to trust for which label. Furthermore, in many cases, sensor fusion improved the performance compared to the best single sensor—there was complementary information in different sensors.

We see the overall advantage of multisensor systems over single-sensor systems, shown by the average performance of the different systems in Table 4. The three sensor-fusion alternatives seem to perform similarly well, with LFL slightly ahead. The selection of a sensor-fusion method can be guided by the training data available to the researcher. When there are plenty of labeled examples that have all six sensors available, the simple EF system can work. Otherwise, late fusion would be more fitting, and there would be plenty of data to train each single-sensor classifier alone.

The leave-one-user-out results are consistent with the fivefold evaluation (Table 4 shows LOO results for the EF system). For some labels, such as "running," the system benefited from the larger training set in the LOO evaluation. For the full per-label results, see https://extras.computer.org/extra/mpc2017040062s1.pdf.

Why Sensor Fusion Helps

The performance of single-sensor classifiers on selected labels (see Figure 4a) demonstrates the advantage of having sensors of different modalities. As expected, for detecting sleep, the watch was more informative than the phone's motion sensors (accelerometer and gyroscope)—because the phone might be lying motionless on a nightstand, whereas the watch could record wrist movements. Similarly, contexts such as "shower" or "in a meeting" have unique acoustic signatures (running

TABLE 3
Predicting specific labels with single-sensor classifiers. The first and second guess of sensors that intuitively seemed relevant, and the balanced accuracy (BA) score of the corresponding single-sensor classifiers.

Label	No. of examples	First-guess sensor (BA)	Second-guess sensor (BA)
Stairs - going up	399	Gyroscope (.73)	Accelerometer (.70)
Stairs - going down	390	Gyroscope (.73)	Accelerometer (.71)
Elevator	124	Gyroscope (.76)	Audio (.71)
Cleaning	1,839	Watch accelerometer (.71)	Gyroscope (.64)
Laundry	473	Watch accelerometer (.66)	Accelerometer (.65)
Washing dishes	851	Watch accelerometer (.70)	Audio (.60)
Singing	384	Audio (.68)	Location (.61)
At a party	404	Audio (.81)	Accelerometer (.74)
At the beach	122	Location (.72)	Phone state (.70)
At a bar	520	Phone state (.93)	Gyroscope (.66)

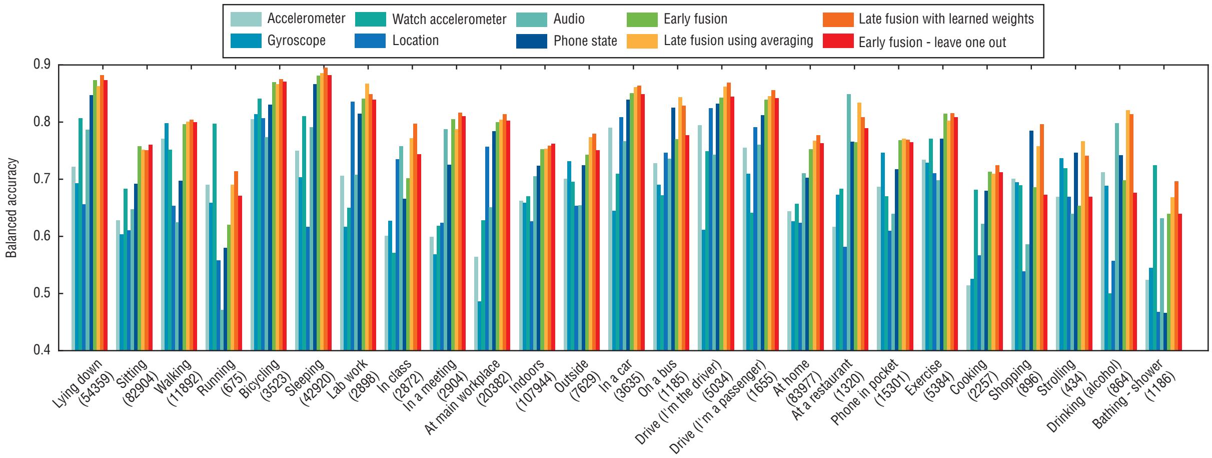


Figure 3. The balanced accuracy scores for selected labels from diverse domains (the number of examples appears in parenthesis).

water, voices) that allowed the audio-based classifier to perform well. When showering, the phone would often be in a different room, leaving the watch as an important indicator of activity. Figure 4a demonstrates that the LFL method assigns reasonable weights to the six sensors—sensors that perform better for a given label were given higher weights.

Investigating where misclassifications occurred helped us understand the system's predictive ability. Figure 4b–g shows confusion matrices that depict misclassification rates between related context labels. For example, a classifier using the phone's motion sensors—accelerometer and gyroscope (see Figure 4b)—to discriminate between body states had confusion between dissimilar labels ("running" versus "lying down"). Such errors arise in natural, unconstrained behavior; in the wild, people don't always carry their phones in their pockets—subjects were sometimes running on a treadmill with their phone next to them, motionless. Adding the watch accelerometer features to the classifier (Figure 4c) helped reduce confusion regarding the activities.

The audio signal from the smartphone (Figure 4d) was informative for labels related to the environmental context. We see a hierarchy of misclassification: while there was some confusion

TABLE 4
The average performance metrics over the 25 context labels from Figure 3. All average scores were well above the p99 value, which marks the 99th percentile of random scores.

Classifier		Accuracy	Sensitivity	Specificity	Balanced accuracy	Precision	F1
Single sensor	Accelerometer	.73	.64	.73	.68	.17	.22
	Gyroscope	.70	.64	.69	.66	.16	.20
	Watch accelerometer	.73	.67	.72	.70	.18	.22
	Location	.71	.63	.70	.67	.17	.22
	Audio	.75	.65	.75	.70	.18	.22
	Phone state	.76	.74	.76	.75	.20	.24
Sensor fusion	EF	.87	.67	.87	.77	.24	.30
	LFA	.84	.76	.83	.80	.23	.29
	LFL	.85	.76	.85	.80	.24	.30
	EF-LOO	.86	.69	.86	.78	.24	.30
p99*		.50	.50	.50	.50	.11	.13

*Scores above the p99 value have less than 1 percent probability of being achieved randomly (p99 was estimated from 100 random simulations).

between labels that shared similar acoustic properties ("toilet" versus "shower" or "class" versus "meeting"), there was a sharper distinction between label groups from different domains ("toilet or shower" versus "class or meeting" versus "restaurant").

The phone placement itself provides cues about the user's activity; when the phone is lying on a table, it is more likely the user is showering than walking to work. The ability to recognize the phone's position can improve overall context recognition. However, a single modality is not sufficient to fully identify the phone position. A classifier based on motion sensors is sensitive to movement, so when the phone was in a bag (possibly motionless), it was often mistaken for being on a table (Figure 4e). On the other hand, a classifier using audio alone is more sensitive to whether the phone is enclosed or exposed to environmental sounds, so with this classifier, cases of "phone in bag" were mistaken for "phone in pocket," and "phone in hand" was often mislabeled as "phone on table" (Figure 4f). By combining motion and audio modalities, the classifier synthesized these two dimensions of discrimination to better recognize phone position (Figure 4g). These examples demonstrate the large variability in behavior in the wild and highlight the benefit of fusing multimodal sensors.

Exploring User Personalization

People move, behave, and use their phones in different manners. A system that is fine-tuned to its specific user might outperform a more general model. To explore the potential of personalization, we performed experiments with a single test user. We compared three models: *universal* (trained on data from other users), *individual* (trained on half of the data from the same test user), and *adapted* (merged both). We tested the three models on the same unseen data.

Figure 5 shows the results of these experiments. The universal model demonstrates a good starting point. This shows the basic ability of a trained system to work well for a new unseen user. As suspected, the individual model performed better than the universal model for labels that had many individual examples ("lying down," "sitting," "sleeping," "at home," "computer work," and "at main workplace"). However, the individual user was missing data for many context labels. For some labels, a new user can acquire only a limited number of examples during the few "training" days, which risks over-fitting to these few examples. In such cases, a universal model is better, having been trained on plenty of data from many users.

The optimal solution is to benefit from both universal and individual data: the user-adapted model shows overall improvement in recognition performance, even among the labels

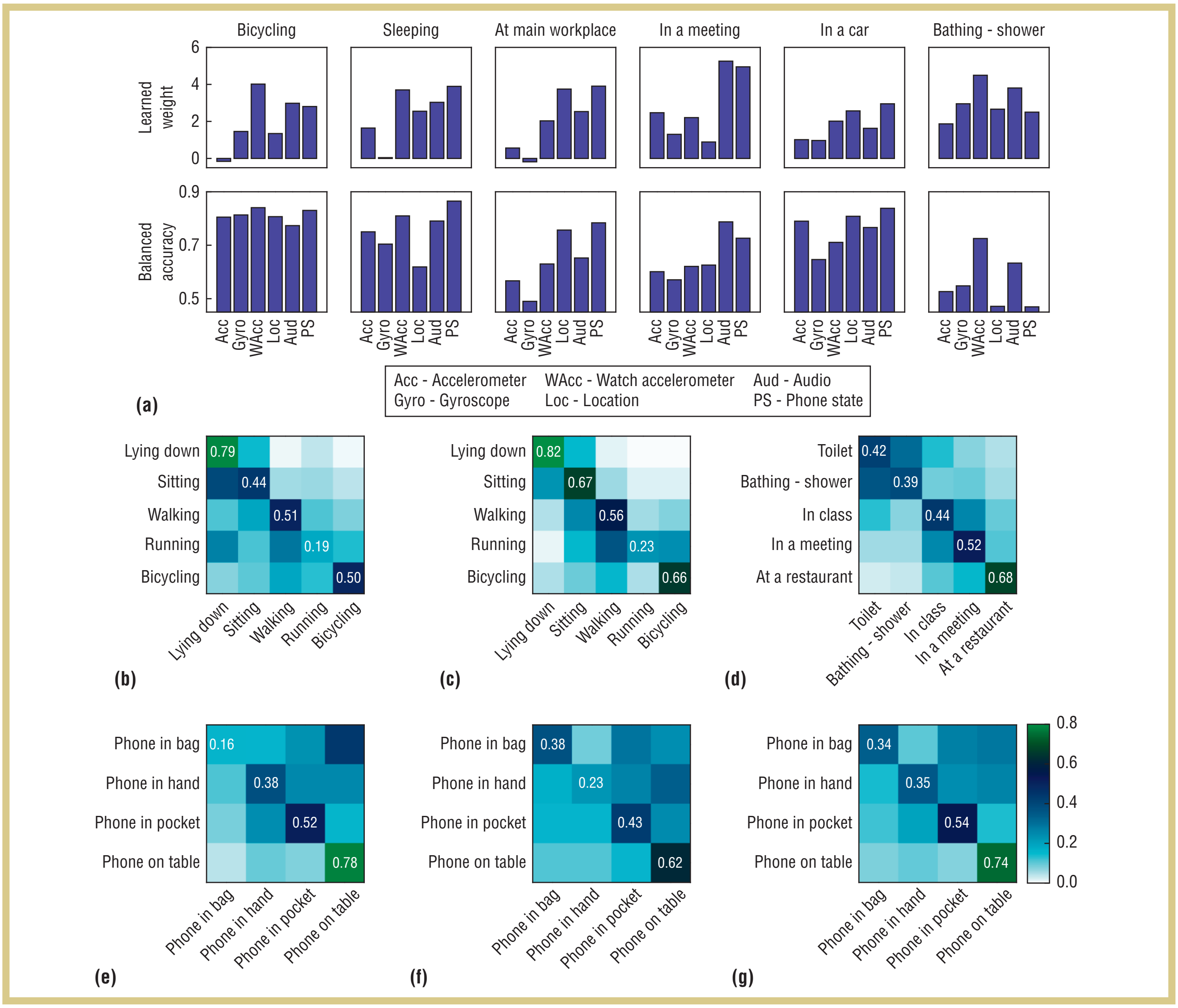


Figure 4. Why sensor fusion helps recognition: (a) The bottom row shows the overall performance (balanced accuracy) of each single-sensor classifier, and the top row shows the weights that the LFL classifier learned to assign to each sensor (taken from the first cross-validation fold). We present confusion matrices for subsets of mutually exclusive body-states using the following sensors: (b) accelerometer and gyroscope; (c) accelerometer, gyroscope, and watch accelerometer; and (d) for different environments using audio. Then we present confusion matrices for phone placement labels using the following sensors: (e) accelerometer, gyroscope, and watch accelerometer; (f) audio; and (g) accelerometer, gyroscope, watch accelerometer, and audio. Rows represent ground truth labels and columns represent predicted labels. Rows are normalized so that a cell in row *i* and column *j* displays the proportion of examples of class *i* that were assigned to class *j*. The correct classification rates (main diagonal) are also marked numerically.

that had more than 300 examples from the test user. LFA is a simple heuristic that manages to demonstrate this advantage. For each label, when there wasn't enough data to train an individual model, the adapted model relied only on the universal model. When there was enough data to train an individual model, the adapted model "listened" to the universal and individual models, and in some cases achieved better performance than either model on its own (for example, for "sleeping" and "at home").

In practical systems, the logistics of implementing personalization might not be an obvious task. For medical applications, the clinician or patient

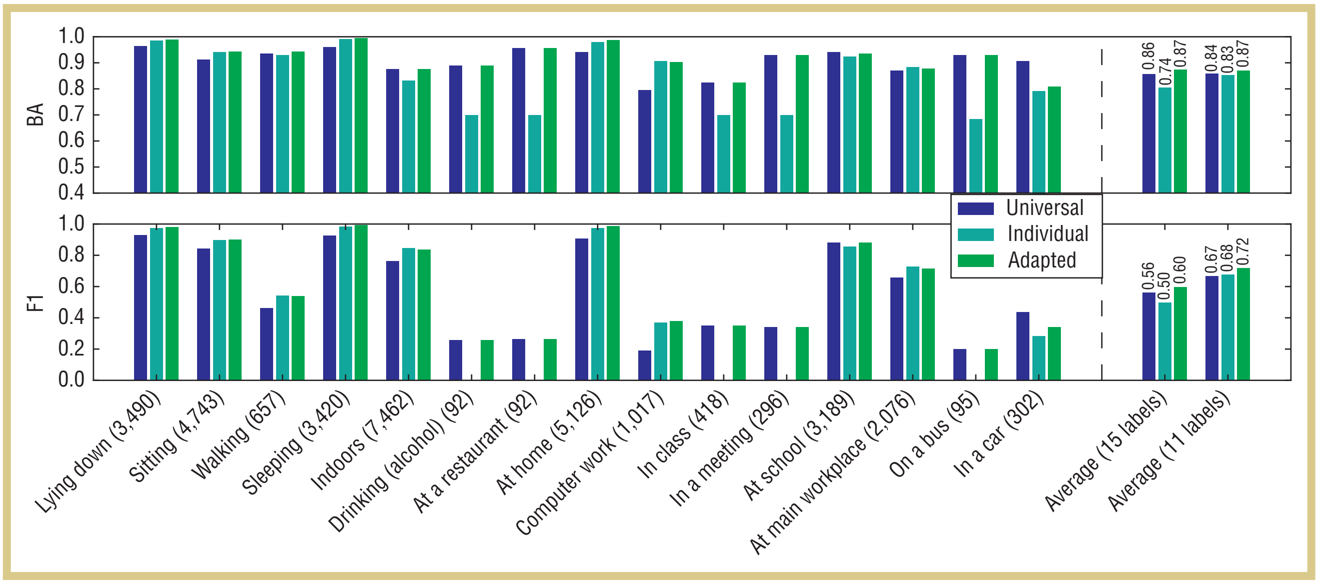


Figure 5. User adaptation performance. The balanced accuracy (BA) and harmonic mean (F1) scores were assessed for a single user. The *x*-axis denotes the labels with the total number of examples from the user. The "universal" model was trained on data from other users, the "individual" model was trained on data from the same user, and the "adapted" model combined the universal and individual models using LFA. The bars on the right-hand side of each plot present the scores averaged over the 15 tested labels and averaged over the 11 labels that had more than 300 examples.

might decide that the cause is important and worth dedicating some effort to provide individual labeled data for a few days to better adapt the model. However, in commercial applications, the users (clients) might not be motivated to invest the extra effort in labeling. In such cases, semi-supervised methods can still be used to make the most of unlabeled data from the individual user to personalize the model.

Our novel dataset reveals behavioral variability in the wild that was underrepresented in controlled studies, but we demonstrate how sensing modalities can complement each other and how fusing them helps resolve contexts that arise with uncontrolled behavior.

Combinatorial representation of behavior is very flexible. A well-trained system has the potential to correctly recognize a new specific situation (combination of labels) that didn't appear in the training. To broaden the range of contexts, researchers can use supervised methods and focus on newly added target labels when collecting extra data. An alternative is to use unsupervised methods to discover complex behaviors in the form of common combinations or sequences of simpler contexts.[19] The labels in our work were interpreted in a subjective manner. The same location might be considered "school" for one subject and "workplace" for another. We didn't tell the subjects how we defined "walking" or "eating" in order to capture the full scope of what people considered "walking" or "eating." Domain-expert researchers might decide to define labels clearly to subjects or use more specific labels, such as "eating a meal" or "snacking."

New technologies and clever solutions for collecting labels in the wild are required to reduce annotation load from study subjects to increase the reliability of labeling. Online learning can be used to keep improving real-time recognition, resulting in less label-correcting effort by new research subjects. Active learning can be used to collect data in scale, while sparsely probing subjects to provide annotations. In parallel, semi-supervised methods can be used to make the most of unlabeled data (which is easy to collect), reducing dependence on labeled examples.

The public dataset we collected provides a platform to develop and evaluate these methods, as well as explore feature extraction, inter-label interaction, time-series modeling, and other directions that will improve context recognition. 🅿

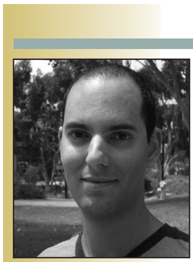
ACKNOWLEDGMENTS

We thank Rafael Aguayo and Jennifer Lu for their help implementing the ExtraSensory App.

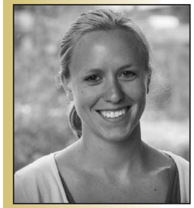
REFERENCES

1. K. Servick, "Mind the Phone," *Science*, vol. 350, no. 6266, 2015, pp. 1306–1309.
2. M. Basner et al., "American Time Use Survey: Sleep Time and Its Relationship to

the AUTHORS



Yonatan Vaizman is a PhD candidate in the Department of Electrical and Computer Engineering at UC San Diego. His fields of interest are machine learning, signal processing, and artificial intelligence. In his research, he applies methods from these fields to music recommendation, mobile sensor processing, and human behavior recognition. He received an MS in electrical engineering from UC San Diego. Contact him at yvaizman@eng.ucsd.edu.



Katherine Ellis is a research scientist at Amazon (but was at UC San Diego when writing this article). Her research interests are in applications of machine learning to physical activity measurement, mobile health, social network analysis, and music recommendation. She received a PhD in electrical engineering from UC San Diego. She is an IEEE member and an ACM member. Contact her at kkatellis@gmail.com.

Gert Lanckriet is a principal applied scientist at Amazon and a professor of electrical and computer engineering at UC San Diego. His interests are in data science, on the interplay between machine learning, applied statistics, and large-scale optimization, with applications to music and video search and recommendation, multimedia, and personalized, mobile health. He received a PhD in electrical engineering and computer science from UC Berkeley. He is a senior IEEE member and an ACM member. Contact him at gert@ece.ucsd.edu.

Waking Activities,” *Sleep*, vol. 30, no. 9, 2007, p. 1085.

3. M.L. Lee and A.K. Dey, “Sensor-Based Observations of Daily Living for Aging in Place,” *Personal and Ubiquitous Computing*, vol. 19, no. 1, 2015, pp. 27–43.

4. I. Nahum-Shani et al., *Just in Time Adaptive Interventions (JITAIs): An Organizing Framework for Ongoing Health Behavior Support*, tech. report 14-126, Methodology Center, 2014.

5. S. Intille, “The Precision Medicine Initiative and Pervasive Health Research,” *IEEE Pervasive Computing*, vol. 15, no. 1, 2016, pp. 88–91.

6. A.K. Dey et al., “Getting Closer: An Empirical Investigation of the Proximity of User to Their Smart Phones,” *Proc. 13th Int’l Conf. Ubiquitous Computing*, 2011, pp. 163–172.

7. J.J. Guiry, P. van de Ven, and J. Nelson, “Multi-Sensor Fusion for Enhanced Contextual Awareness of Everyday Activities with Ubiquitous Devices,” *Sensors*, vol. 14, no. 3, 2014, pp. 5687–5701.

8. M. Shoaib et al., “Towards Detection of Bad Habits by Fusing Smartphone and Smartwatch Sensors,” *Proc. 2015 IEEE Int’l Conf. Pervasive Computing and Communication Workshops* (PerCom Workshops), 2015, pp. 591–596.

9. J. Kerr et al., “Objective Assessment of Physical Activity: Classifiers for Public Health,” *Medicine and Science in Sports and Exercise*, vol. 48, no. 5, 2016, pp. 951–957.

10. K. Kunze and P. Lukowicz, “Sensor Placement Variations in Wearable Activity Recognition,” *IEEE Pervasive Computing*, vol. 13, no. 4, 2014, pp. 32–41.

11. R.K. Ganti, S. Srinivasan, and A. Gacic, “Multisensor Fusion in Smartphones for Lifestyle Monitoring,” *Proc. 2010 Int’l Conf. Body Sensor Networks*, 2010, pp. 36–43.

12. H. Pirsiavash and D. Ramanan, “Detecting Activities of Daily Living in First-Person Camera Views,” *Proc. 2012 IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), 2012, pp. 2847–2854.

13. Y. Dong et al., “Detecting Periods of Eating During Free-Living by Tracking Wrist Motion,” *IEEE J. Biomedical and Health Informatics*, vol. 18, no. 4, 2014, pp. 1253–1260.

14. A.M. Khan et al., “Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs,” *Int’l J. Distributed Sensor Networks*, vol. 10, no. 5, 2014; doi: 10.1155/2014/503291.

15. M. Rossi, G. Troster, and O. Amft, “Recognizing Daily Life Context Using Web-Collected Audio Data,” *Proc. 16th Int’l Symp. Wearable Computers* (ISWC), 2012, pp. 25–28.

16. S. Consolvo et al., “Activity Sensing in the Wild: A Field Trial of Ubifit Garden,” *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2008, pp. 1797–1806.

17. T. Rahman et al., “Towards Accurate Non-Intrusive Recollection of Stress Levels Using Mobile Sensing and Contextual Recall,” *Proc. Int’l Conf. Pervasive Computing Technologies for Healthcare*, 2014, pp. 166–169.

18. A. Stisen et al., “Smart Devices Are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition,” *Proc. 13th ACM Conf. Embedded Networked Sensor Systems*, 2015, pp. 127–140.

19. J. Seiter et al., “Discovery of Activity Composites Using Topic Models: An Analysis of Unsupervised Methods,” *Pervasive and Mobile Computing*, Dec. 2014, pp. 215–227.