Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data

MARK MIRTCHOUK, DREW LUSTIG, ALEXANDRA SMITH, IVAN CHING, MIN ZHENG, and SAMANTHA KLEINBERG, Stevens Institute of Technology

Automated dietary monitoring solutions that can find when, what, and how much individuals consume are needed for many applications such as providing feedback to individuals with chronic disease. Advances in body-worn sensors have led to systems with high accuracy for finding meals and even which foods are consumed in each bite. However, most tests are done in controlled lab settings with restricted meal choices, little background noise, and subjects focused on eating. For these systems to be adopted by users it is critical that they work well in realistic situations and be able to handle confounding factors such as background noise, shared meals, and multi-tasking. Work in realistic environments usually has lower accuracy, but has challenges in determining ground truth. Most critically, there has been a significant gap between lab and free-living environments. This is compounded by data usually being collected for different individuals in each setting, making it difficult to determine how the accuracy gap can be closed. We present a multi-modality study on eating recognition, using body-worn motion (head, wrists) and audio (earbud microphone) sensors for 12 participants (6 from the lab study, 6 new to test generalizability). In contrast to the lab, where audio alone has the highest accuracy, we find now that a combination of sensing modalities (audio, motion) is needed; yet sensor placement (head vs. wrist) is not critical. We further find that lab data does generalize to other participants, but while personal free-living data improves accuracy, more data from others can actually lead to worse performance.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**;

Additional Key Words and Phrases: Nutrition; Eating recognition; Acoustic and motion sensing

ACM Reference format:
Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. 2017. Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 85 (September 2017), 20 pages.
DOI: http://doi.org/10.1145/3131894

1 INTRODUCTION

Nutrition is key to managing and preventing many chronic diseases, such as diabetes and obesity. Dietary monitoring at the population level can provide insight into changes in behavior that may lead to global health problems such as the obesity epidemic [16] and nutritional deficiencies even in resource rich nations [24]. At the individual level, tracking food consumption can lead to increased and longer lasting weight loss [9]. Unlike physical activity, which can be tracked with consumer-grade devices (from pedometers to activity monitors that count repetitions of exercises), food logging is still mainly done with paper logs or apps where users select foods consumed from a database or log photos of their meals. This logging is subject to bias and error and is extremely

Authors' address: M. Mirtchouk, D. Lustig, A. Smith, I. Ching, M. Zheng, and S. Kleinberg, Stevens Institute of Technology, Computer Science Department, 1 Castle Point on Hudson, Hoboken, NJ 07030. contact email: samantha.kleinberg@stevens.edu.

2474-9567/2017/9-ART85
DOI: http://doi.org/10.1145/3131894

labor intensive to complete, so users rarely stick with such systems longterm [10]. Photos are a faster input method, but there are not yet fully automated solutions for turning photos from the start and end of a meal into a set of foods and amount of each consumed.

There have been many steps toward automating dietary monitoring, including using body-worn audio and motion sensors [3] or environmental sensors [8] to find meals and in some cases foods consumed [4]. However, the most detailed studies with the highest accuracy have been in controlled laboratory settings, often with few foods and little background noise or activity, but with objective ground truth. Real-world studies provide insight into performance in daily life, but often there are large differences between the environments (e.g. full meals vs. discrete foods, loud restaurants vs. quiet labs, and shared meals vs. eating alone) that make it difficult to determine the reasons for lower precision in the wild. Some works examined performance in both conditions [25], but we are not aware of any that do so for the same participants.

We aim to understand how and why accuracy for identifying eating differs between laboratory and free-living settings, and particularly to gain insight into how this gap can be narrowed. We previously developed the ACE (accelerometer and audio-based calorie estimation) dataset, a set of finely annotated data for a group of 6 individuals wearing 4 sensors (Google Glass for head motion, smart watches on each wrist for wrist motion, and an earbud to capture chewing sounds) over 12 hours (2 6-hour sessions) each as they consumed the meals of their choice in a lab outfitted with video cameras [14]. By comparing the sensors in the same setting, we found that audio is surprisingly robust. However, meals were still consumed in a relatively quiet and controlled environment and without the variety of eating behavior encountered in daily life. Thus it is not clear whether the sensors that perform best in the lab will still do so in general, and whether models learned in the lab are transportable to daily life. We address this knowledge gap by collecting two sets of free-living data that provide complementary insight: 1) an additional 24 hours (two 12-hour sessions) of data for 5 of the ACE participants in free-living conditions, and 2) two days of data from 5 individuals who were not involved in the initial lab study plus 5 days for a 6th individual. By collecting data for the same individuals, and allowing free choice of foods in both settings, we can better disentangle which differences in performance are due to differences between individuals versus environmental differences. By collecting data from a second cohort of more diverse individuals we can assess the generalizability of both lab and free-living data.

1.1 Contributions

Our key contributions are 1) evaluation of how well an eating recognition system trained in the lab performs in the wild for the same and different participants; 2) quantitative assessment of the value of different sensors (audio, motion) and weakly labeled but personalized data, and 3) development of a publicly available data set with around 12 hours of laboratory data and 24 hours of free-living data for the same 5 participants, 24 hours of free-living data for 5 other participants, and 60 hours for another.[1]

2 RELATED WORK

Research on automated eating recognition can be categorized in two main ways: study environment (controlled lab vs. free living), and sensing modality (e.g. motion, audio). We aim to understand the relationship between activity in the two environments, and review work with a focus on these.

2.1 Laboratory studies

We define laboratory conditions as any set-up where the environment is the same between all participants and participants are observed (by researchers or video). Some laboratory environments are a single participant eating

[1]The dataset is available at: http://healthailab.org/data.html

in a silent room, while others allow multiple people to eat at once. However, all introduce homogeneity and involve observation during eating, which may influence behavior.

Work on recognizing eating activity automatically has mainly focused on lab environments to enable controlled assessment of novel sensing technologies. Some of the earliest and most accurate approaches with wearable sensors used earbud microphones to detect chewing sounds with 99% accuracy [3] and to identify amount of food consumed in each bite [2]. However these works involved 3 to 5 discrete foods (e.g. potato chip, apple) consumed in a quiet room. Later work developed specialized microphones [18] and microphones placed on the neck [20]. Since all evaluations were on small pieces of food with distinct sounds eaten in quiet environments, it is unclear what accuracy can be obtained in realistic settings, where meals may be shared or eaten in noisy restaurant environments.

Motion sensors have also been used to identify eating in the lab. Google Glass was used to find meal periods using head motion [19], with ground truth coming from researchers marking meal times during observation. Results varied significantly by individual, with 11/38 participants having 100% precision for identifying meal periods and 9 participants having 0% precision. Other work used sensors in the ear to measure jaw movement [5] while eating three different foods and performing other activities, and classified meal periods with high accuracy. Chewing motion has been captured with EMG equipped eyeglasses [29]. Finally, wrist and arm motions have been used to identify eating-related gestures, such as bringing food to the mouth. One study of two participants achieved 87% accuracy for identifying eating gestures in the lab [1]. More recently, Ramos-Garcia et al. [22] collected wrist motion data for 25 participants at an instrumented table in a university cafeteria to allow a more realistic environment with wider food choices. However, no data was collected outside of meals (where other wrist motions may be confounded with eating) and using a single environment limits understanding of the method's generalizability.

2.2 Free-living environments

Ultimately, we aim to detect eating periods and foods consumed during daily life. As in the lab, work has mainly used a single body-worn sensor at a time. Dong et al. [12] used wrist motion to find meal times in a study of 43 individuals for one day each, with 79-81% accuracy when using a 20:1 weighting of true positives to true negatives. At the level of meal segments, 100 activities were correctly recognized, 16 were missed, and there were 379 false positives. Ground truth came from participants noting meal times in a written log or using a button on a phone. Data on food type and amount was not collected and participants must remember to mark the start and end of meals. One of the challenges is that only a single day of data was collected for most individuals, making it difficult to determine what variations exist between and within individuals. Other work by Scisco et al. [23] collected data for 77 participants over 2-weeks each, finding a correlation of r=0.44 between bite count and calories consumed, with ground truth from 24-hour daily recalls. However, our goal is to understand accuracy on a per meal and per bite level (as real-time information is needed for applications such as insulin dosing guidance), and to gain insight into the relationship between behavior in lab versus daily life.

While audio sensors have mainly been used in controlled environments, wrist-based audio has been used to capture environmental sounds related to eating. Thomaz et al. [27] collected 4-7 hours of data (one day) for each of 21 participants, with ground truth based on participants' activity recall and researchers coding the audio data. Unlike most audio-based methods, quieter environments proved more challenging for recall, as they did not have distinct eating-related noises. Other work collected data from a body-worn microphone and physiological sensors (Microsoft Band, Affectiva Q), but the goal was to predict future eating moments and the audio data was not used [21].

Image-based methods have been primarily used to identify food type and quantity consumed in an environment independent manner, such as using crowd-workers to annotate images [17]. Recently, deep-learning based

methods have been used to automatically recognize food types by linking images and GPS location information to restaurant menus [6, 7]. With the exception of continuous first-person point of view cameras (POV), though, these require users to actively take images of their food, prohibiting a fully automated solution, and will be less accurate for foods consumed outside of restaurants (as they use menu items to reduce the set of food classes). Further, photo-based methods use a single photo of the meal (taken prior to consumption) and assume that the entire quantity is consumed, while this was only true for 67% of meals in our free-living datasets. Other work has used POV cameras to find periods of eating, though not in real time or with the fine-grained timing possible with motion and audio sensing [26].

2.3 Combining lab and free-living data

While most studies have tested systems in either laboratory or free-living conditions, Thomaz. et al [25] used wrist-based motion sensing, trained on lab data from 20 subjects and evaluated on data collected in-the-wild with 7 participants for 1 day each (31.5 hours of data total) and 31 days for 1 participant. Participants wore POV cameras that took photographs each minute and later reviewed the photos to find meal times. This work is unique in that it trained a classification system on controlled data and evaluated it on free-living data. By relying on automated cameras, it also avoided the common problems of participants forgetting to write down meals, though it is possible some quick snacks may not be captured or photos may not clearly depict meals. One of the limitations is that the set of participants differed between the two settings as did the foods (with a controlled set of 5 meals used in the lab), so it is not clear whether training on more food types or some data from each individual would be better for improving future performance in the wild. Dong et al. [11] also used wrist motion from laboratory (51 participants in lab eating waffles, 47 in lab eating their own food) and free-living (4 participants over 54 total meals) settings, to count food intakes. It is not clear how many free-living days were captured, and bite count accuracy was not evaluated in that setting (only relationship between inferred bites and calories from food logs). Most recently, Zhang and Amft [28] developed EMG sensing smartglasses customized to each participant. The glasses were evaluated with 10 participants, over 2 controlled lab sessions and one free-living day each. While this is the first work we are aware of that collects data for the same participants in both settings, the lab data does not contain full meals (it is used to study EMG signals based on food hardness), and only one day is collected for each participant, so it is not known how much variation there is in individual behavior.

3 STUDY DESIGN

In this paper, we aim to bridge the gap between laboratory and free-living environments by 1) collecting data from the same set of participants in both settings over multiple days in each to disentangle person-level variation from variation due to environment or foods consumed, and 2) collecting data for a second group of participants to test generalizability. All data collection was approved by our university IRB.

Our goal is to understand the relative value of different sensing modalities (audio, head motion, wrist motion), and determine the true difference between lab and real-world performance. To this end, our study has three main components: lab data collection (ACE), which has been described in prior work (comparing sensing modalities [14], and estimating food type and quantity [15]) and new data introduced in this paper from free-living environments for the same people (ACE-FL) and a new group of individuals (ACE-E). In both environments data is collected using multiple sensing modalities (audio, motion), with no restrictions on meal choice. An overview of the study components and some of the key differences between lab and free living protocol are shown in figure 1.

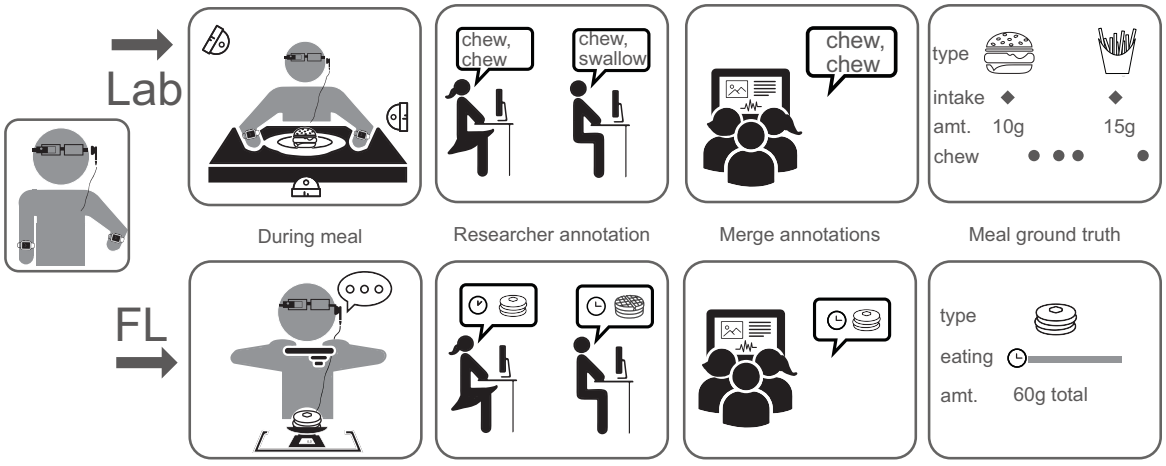


Fig. 1. Overview of data collection and annotation differences between lab and free-living (FL) data. In lab, annotation is based on video data and is at the level of chews and swallows, and foods consumed in each bite. In FL data, participants mark meal start and end, which are verified by the researchers and provide an overview of the foods consumed in the whole meal.

3.1 ACE (laboratory) study design

We begin with a brief review of the ACE study, before describing the ACE-FL and ACE-E additions. ACE aimed to create data that could be used to evaluate sensors rigorously in the same setting, with detailed and objective ground truth, while also capturing realistic behavior.

3.1.1 Sensors and Devices. Wearable sensors were used for eating recognition, while others were used for ground truth.

Audio We augmented an earbud with internal and external microphones, to enable removal of environmental noise, and recorded at 44.1kHz using a pocket audio recorder worn in a running belt.

Wrist motion Participants wore an LG G-Watch with 9-axis inertial motion sensor on each wrist, so that we could determine how much eating activity is done with the dominant versus non-dominant hand. The watches recorded at a rate of 15Hz, due to battery life.

Head motion is captured with the 9-axis motion sensor in Google Glass. While Glass was attached to a small external battery due to its battery life, we recorded at the 15Hz frequency of the watches. The battery fit in the running belt with the audio recorder, and its flexible wire did not impede movement.

Ground truth Data was collected from video cameras and a continuous scale. Three cameras captured top, side, and front views of participants eating, recording at a rate of 30fps. To obtain information on food weight a weigh scale embedded in the table used for meals recorded continuously at 10Hz, yielding granular information at the level of bites rather than amount consumed for a full meal.

3.1.2 Data collection. Data was collected from 6 people in two 6-hour sessions each (around 72h total), as they ate the meals of their choice at the times of their choice. There were no food restrictions and meals included complex foods such as a burger and fries, sushi and dumplings, and yogurt with candy. Outside of meals, participants were free to do the activity of their choice, as long as they remained in the lab space. Activities included reading, using a computer or smartphone, doing pushups, walking around, and napping.

3.1.3 Ground truth annotation. Ground truth of eating activities came from annotation of video data. After regularizing the frame rate and synchronizing the three cameras, the video data was annotated by two researchers for the following eating activities: chew, swallow, mouthing (manipulation of food with the tongue), preparation of food or drink (with one or both hands), delivery of food or drink to the mouth (with one or both hands), and touching face with napkin. Codes that agreed within 500ms (250ms for chews) were merged automatically by averaging their times, and all others were resolved in a collaborative process with a third researcher. Food weight and type for each intake were independently annotated by two researchers using the video and data from the scale.

3.2 ACE-FL (free-living) study design

In the ACE free-living (FL) component, we aim to understand how behavior for the same set of individuals may differ across environments, and how well classification methods trained on laboratory data work on free-living data. By allowing unrestricted foods in the laboratory setting and collecting data for the same participants, we remove two of the usual differences between environments. We aimed to collect data covering all eating in each study day (rather than only two primary meals as in the lab). This lets us identify differences in eating patterns based on context such as slower eating due to talking with friends or waiting for food at a restaurant and how this affects inference accuracy when algorithms are trained on lab data (where participants are more focused on food consumption). Beyond environment, the primary difference between the two studies is in how ground truth is obtained.

3.2.1 Sensors. Participants wore the same set of sensors as in the lab (Glass, earbud, smartwatch on each wrist). Since video cameras were not used in the free-living environment, and eating could be done anywhere (not just at the instrumented table), participants were given a portable food scale and smartphone for recording images of foods along with food weights at the start and end of each meal.

3.2.2 Data collection. ACE-FL data collection spanned two 12-hour days per participant (24h total), though the days were normally not consecutive. For each session, participants picked up the sensors at the lab the evening before data collection and returned them the day after so that data would be collected for most waking hours, capturing all meals. When devices were returned, there was an unstructured debriefing interview, where we walked through the day to identify potential omissions in the food logs, asked about problems encountered such as uncomfortable social situations (none reported), and found out about activities other than eating. In some cases one or both watch batteries died in under 12 hours. FL data was collected for 5 of the ACE participants, as one was unavailable due to scheduling reasons. Participants were aged 18-35, with one female and were white (3) and Asian (2).

Participants were instructed to remove Glass before driving, and to remove all sensors during aquatic activities such as bathing. They were also told to notify anyone they live with or planned to interact with for extended periods during the study, in case they objected to the audio recording (as the body-worn microphone may pick up speech from the wearer and those around her). In one case, a participant's friend was uncomfortable with the continuous audio collection and rescheduled a shared meal. Participants were allowed to redact any audio segments that either they or their companions wished to redact, though only one redacted a brief conversation. There were no other restrictions on activities.

At the beginning and end of each data collection day, participants performed the same synchronization procedure as in the lab, to ensure all time series can be aligned. Glass, the earbud, and the two watches were held together in a bundle by the participant and three slow controlled knocks on a table were performed. This produces clear spikes in audio and motion signals that can be used to align the signals.

3.2.3 Ground truth annotation. The key challenge of FL data collection is obtaining accurate times for each activity, while ensuring that this documentation process does not itself impede activities and movement or introduce bias. There is a trade off between a precise account of each individual meal (e.g. through photographs and participant logging of each meal component using a mobile app), and the potential for lower compliance due to the time burden. Estimating not only meal time, but also food type and amount is important for automating nutrition logging, though, and collecting this data can also lead to better understanding of errors during classification (e.g. failure with soft foods). We combine multiple types of annotation to avoid a single point of failure in obtaining ground truth.

We only require eating to be annotated, as recording a full day's activities is a time consuming process, and may lead to gaps or omissions. By focusing solely on meals we aim to encourage more accurate annotation of these activities. While we do not know the activities between meals, we can be nearly guaranteed that they are negative examples of eating. Unlike the controlled laboratory data, annotations are now at the meal level, rather than individual chews and swallows. Thus, some of the times marked will not be eating per se, but may include talking with friends at a shared meal, a time gap between courses at a restaurant meal, or food preparation (e.g. adding condiments to the meal).

Participant annotation Participants collected photos of meals, written food logs, and voice notes. To do this, at the start and end of each meal participants were instructed to tap the annotation button on one watch to indicate there will be a spoken annotation, or record the time in a paper or electronic journal. Then participants verbally announce when they are starting to eat and when they are done, which is recorded by the earbud. Times marked by button, journal, or photo timestamp are used to determine where to listen for spoken annotation or, in the absence of that, eating noises.

To determine food type and quantity in FL, we provided participants with a small portable kitchen scale. They photographed each component of their meal atop the scale at the start and end of each meal (e.g. separate photograph for a hamburger and a side of fries). In addition, they completed daily journals with free-text descriptions of each meal. The photographs serve a few purposes. They document the weight of each food or drink consumed, and their timestamps can be used as back-up documentation when journals are incomplete or audio annotations of meal times are missing.

Researcher annotation Two authors independently found meal times for each session and the times were reconciled in a collaborative process involving a third author, as shown in figure 1. Researchers first looked for meal times recorded with the annotation button or in the day's journal. Using these times, they listened to the audio data for spoken annotation. In cases where the participant did not announce the start or end of a meal, researchers listened for signs of eating, and also compared the number of meals logged to the number of meal photos. In cases with photos but no times logged, the procedure was to use the image times as outer bounds on meal time and then listen for eating sounds around that time to find the start and end of eating. Of the 5 participants, 2 consistently had spoken meal annotations, 1 had spoken annotations and journal entries, and 2 did not have announcements (necessitating the use of photo time stamps). To identify foods and amounts consumed, food logs were matched to image files. For each pair of weights for each component (before and after the meal), we assign the difference to the set of food objects present in the photo. For example, one image contained bread and salad (two distinct food objects), but since they were photographed together, we only know the total amount of bread and salad combined and cannot determine the amount of each component.

3.3 ACE-E study design

As a test of generalizability, we collect data for a second group of individuals who were not involved in the ACE and ACE-FL studies, aiming to recruit a more diverse set of participants and test how well models learned on

ACE generalize to a new population. Participants were compensated $50 for two days of data collection or $150 for 5 days, since they had to return to the lab each day.

3.3.1 Sensors. We used the same sensors as in ACE-FL, with the omission of Google Glass. Thus the sensors were the two smartwatches, and our acoustic sensing earbud. Glass was omitted due to our preliminary findings showing wrist motion may be adequate for recognizing food type [15] and the added challenge of participant recruitment and data collection with Glass.

3.3.2 Data collection. Protocol closely matched ACE-FL, with 2 or 5 days of data collected. The only difference was that each participant had breakfast in the lab, then left for free-living data collection, spanning all waking hours (or until device batteries died). There were 8 participants (three 5-day, five 2-day), but two had device malfunctions leading to unusable data (one 5-day, one 2-day). A third 5-day participant had only two days of usable data. In this work we analyze the FL data as in the lab participants ate a standardized breakfast with the ACE camera and scale set-up for future use in understanding chewing patterns, while in all other data collection participants had free choice of meals.

3.3.3 Participants. The usable ACE-E dataset has 15 days of data from 6 people (2 days for 5 people, 5 days for 1 person), more than doubling the 10 FL days of ACE-FL. Participants were 3 male and 3 female; self-identified as white (2), Asian (1), Black or African-American (2), and Hispanic (1); and were 20 to 63 years old (mean 36, s.d. 19) with a mean BMI of 22.3 (s.d. of 2.7, range 19.6 to 27.4).

3.3.4 Ground truth annotation. The protocol was as in ACE-FL, and most meals were logged with photos and a paper food log. In one case we also identified a meal that was not logged but was clearly described on the audio (Participant announced "I'm going to eat Doritos." and was heard chewing).

4 DATA CHARACTERISTICS AND COMPARISON

4.1 Environment and activities

4.1.1 ACE lab. In the lab, participants spent 5.4h eating out of 59 hours of data collection, with a mean of 2.5 meals consumed per 6-hour data collection day (s.d. 0.7). During lab data collection, participants were asked to eat however they wished. Some chose to multitask while eating (working on a laptop or using a smartphone), others had conversations with research group members, and some shared a meal with the group. Thus the lab data include some background noise and external chewing. However, any controlled environment cannot reproduce the full variety of real-world eating behavior.

4.1.2 ACE-FL. For the 5 ACE participants who participated in free-living data collection, we collected 112.5 hours of data, with a mean of 22.5h (s.d. 2.4h) per person. After excluding periods where all four devices were not recording, we have 110.5h (mean 11.1h, s.d. 1.6h per session). The total eating time was 8.4h, with a mean of 1.7h of eating (s.d. 0.6h) per person across the two days of data collection. A mean of 3.1 meals were consumed per day (s.d. 1.4). Meals were consumed at home (2 participants), at work (1 participant), 60% at home and rest at work (1 participant), and half at home and half at restaurants (1 participant). Participants did not provide detailed activity annotations, but mentioned engaging in: running, walking, biking, being a passenger in a car, watching a movie, working at a computer, making phone calls, sawing, sleeping, reading, and shopping for food.

4.1.3 ACE-E. From the 6 participants with usable data, we collected a total of 144.2h of data where all devices were recording, with a mean of 9.6h per session (s.d. 2.7h). The mean is similar to ACE-FL, but there was more variation in session length. The total time spent eating was 11.5h, with a mean of 1.7h for the five two day participants (s.d. 0.7h) and a total of 3.1h for the five-day participant. The daily eating time is similar to the ACE-FL cohort, though overall there were somewhat more meals of shorter duration, as participants ate a mean

Fig. 2. Each column shows a sample of two foods consumed by one of the five participants in ACE-FL.

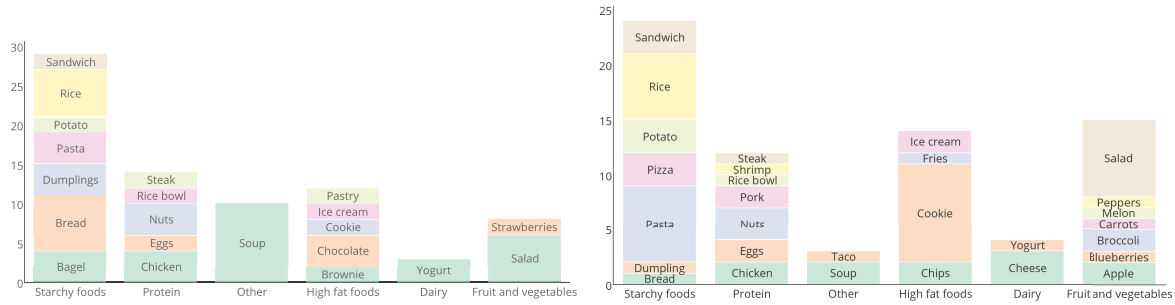
of 3.7 meals per day (s.d. 1.1), though the mean using only periods where all devices were recording is 3.2 meals per day (s.d. 1.3). Most meals were consumed at home, though one participant ate half their meals at restaurants and another ate a third of meals outdoors. During discussion when returning the devices participants reported activities including skateboarding, playing a guitar, playing sports, cooking, picnicking, and working.

4.2 Food and drink consumed

Foods consumed in all phases of data collection spanned a large range of intake methods and food textures. For example, ACE-FL meals included fruit (strawberries, watermelon, cantaloupe, cherries, blueberries, pears), soups (both broths and soups with solids and noodles), handheld foods (e.g. taco, sandwich, hamburger, bagel), complex food types (e.g. stir fry, salads, pasta with sauce and meat), crunchy foods (e.g. carrots, crackers, peanuts), meats, and tacky foods (peanut butter, yogurt). Drinks were consumed from cups and bottles, and via straw. In ACE-E, participants consumed a similarly wide range of foods. A sample of foods from each ACE-FL participant are shown in figure 2.

First, we examine how food choices in the lab and in FL compare for the same group of individuals. In ACE, a total of 44 unique foods were consumed plus drinks (combining all drink types into a single "drink" class), and 36 foods after further excluding the individual who did not participate in ACE-FL. In ACE-FL, the 5 participants consumed 43 unique items (again plus a single "drink" category). Foods consumed in both ACE and ACE-FL are shown in figure 3a, grouped by nutritional categories. In all, 21 food types were present in both environments, and if we include the participant who did not participate in ACE-FL, this rises to 26 foods (adding burger, chips, fries, nut butter, and tacos). Nearly half the food types in ACE-FL were thus also consumed in ACE (60% if the individual who only participated in ACE is included). This suggests both that participants did not alter their diet for the lab setting, and that a relatively small number of meals can capture most of the variation in diet.

Next, we compare how foods consumed by the individuals in ACE-E compare to those from the ACE-FL cohort. In ACE-E, participants consumed a total of 43 unique foods, plus various drinks (44 classes total). Despite the substantially different demographics, a majority of these foods overlapped with foods consumed in ACE or ACE-FL. Figure 3b shows foods from ACE-E that overlapped ACE or ACE-FL. A total of 29 foods (plus drink) in ACE-E were consumed by the 5 ACE-FL participants (either in lab or free living data collection). Thus despite the different set of participants, 67% of foods overlapped with the two prior rounds of data collection, suggesting that



(a) Number of times foods from each category were consumed in ACE or ACE-FL, for foods consumed in both settings.

(b) Number of times each food was consumed by ACE-E participants, for foods also consumed in ACE or ACE-FL.

Fig. 3. Comparison of foods consumed (left) in ACE lab and ACE-FL and (right) between ACE-E and ACE lab/FL participants.

we are capturing a representative set of foods. While the food choices were similar, there were some differences in the distribution of meals, such as more soup in ACE-FL/lab and more cookies and pasta consumed in ACE-E. The remaining 14 foods from ACE-E with no overlap were meats (bacon, beef, fish, meatloaf, mussels, salami), fruits and vegetables (banana, blackberries, corn, grapes, plantains, spinach) and other items (pretzels, whipped cream). This suggests that rather capturing more foods, the primary value of data collected from more participants would be capturing more variety in eating movements.

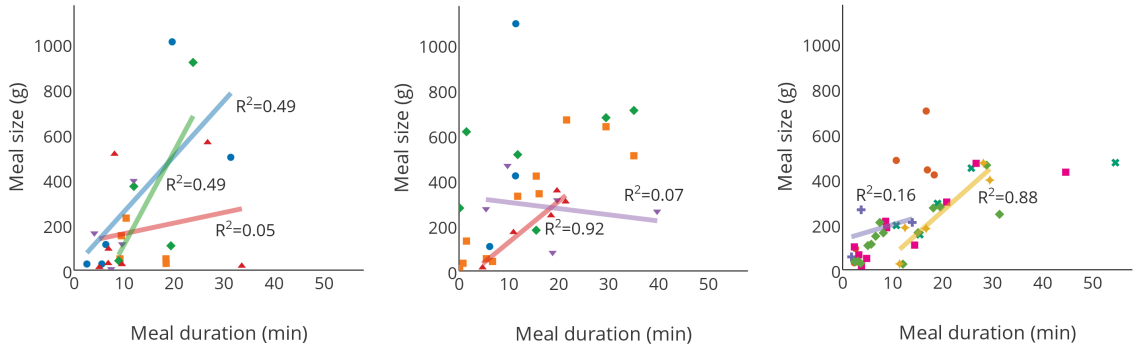
4.3 Eating behavior

We now compare eating behavior between lab, FL, and our generalizability test. References to "meals" encompass all eating episodes, as the boundary between meals and snacks can be unclear.

In ACE-FL, mean meal duration was 16.2min (s.d. 10.3min), while lab meals had a mean duration of 13.1min (s.d. 8.6min). This difference however was not statistically significant, with $p > .2$ using an unpaired t-test. Given the similarity in duration and food choice, this also suggests that the meal logging procedure in FL did not lead to a change in behavior relative to the lab (where participants did not log meals, as we had video). Meals in ACE-E were more similar to those in ACE-FL, lasting a mean of 14.4min (s.d. 11.3min) ($p > 0.46$). On the other hand, meals in ACE-FL had a mean size of 341.1g (s.d. 255.6g), while ACE-E meals had a mean size of 233.1g (s.d. 171.1g). Thus on average ACE-E meals were significantly smaller ($p < .022$ using an unpaired t-test), despite being only slightly shorter in duration, supporting our hypothesis that meal size and duration are weakly related.

Since meal duration is sometimes used as a proxy for amount consumed, we examined this relationship in more depth. Meal duration relative to amount consumed is shown in figure 4, with each participant depicted with a different color and shape. The best (lab $R^2 = 0.49$, FL $R^2 = 0.92$, E $R^2 = 0.88$) and worst (lab $R^2 = 0.05$, FL $R^2 = 0.07$, E $R^2 = 0.16$) fits of a regression of amount on duration for each participant's data are indicated. For all data pooled together $R^2 = 0.20$ in lab, $R^2 = 0.26$ in ACE-FL, and $R^2 = 0.49$ in ACE-E. Thus eating time is likely not a reliable proxy for food consumption. Further, these values may diverge even more strongly with behaviors such as binge eating or frequent snacking.

It is unsurprising that lab meals were generally shorter with less variation in duration. First, lab data was collected for 6 hours each day and meals were mainly breakfast and lunch (with some participants having lunch and dinner) with some snacks while ACE-FL data was collected for ~12-hours each day and included all meals. Second, while participants multitasked and talked with researchers in the lab, eating was still likely more focused than in totally unconstrained environments, where people may share meals, go to restaurants, or eat while



(a) ACE (lab) (b) ACE-FL (c) ACE-E

Fig. 4. Meal duration relative to meal size in each study component. Each participant is represented with a different color and shape. For the ACE lab and ACE-fl group, each participant is represented by the same color/shape across the two plots. Fit lines are shown for highest and lowest R^2 values of a regression of food amount on meal duration for each individual.

watching TV. The difference in meal duration is not explained by differences in food choices given the overlap observed. The same meal can have different duration by context (e.g. rushing to work versus leisurely meal with friends).

As many image-based methods estimate amount consumed using only a photo at the start of a meal, we examined how often people consumed all food served. Out of 31 meals in ACE-FL, only 55% (17) were fully consumed while 75% (41) were fully consumed in ACE-E. We cannot draw conclusions about tendency to consume all food served, but can say that in a number of cases in data from 11 people over at least two days of daily life, this assumption did not hold and it may be important to evaluate accuracy for estimating the amount of food remaining, which may be more difficult.

5 METHODS

We now discuss the process for data processing and classification. Since we aim to compare laboratory and free-living data, we keep the procedure as similar as possible to our lab study [14].

5.1 Data processing

5.1.1 Audio. Audio data underwent an initial noise-cancellation procedure that was described in [14]. This procedure subtracts from the earbud's internal microphone signal the portion that is well-predicted by the earbud's external microphone. This enables us to remove environmental noise, including eating noises from other people. We then segment the audio data into 200ms frames with a 20ms step size. These values are chosen to capture the entirety of a chew, while being small enough to not span multiple chews. For each window we extract a standard set of signal processing features using the Yaafe toolbox: energy, spectral flux, zero-crossing rate, and 11 MFCC coefficients.

5.1.2 Motion. All feature extraction from motion data proceeds on raw data. Since this data is primarily useful for picking up longer eating gestures, such as movement of food to the mouth, rather than brief activities such as chewing, we segment all motion signals (head and wrist) into 5-second frames with 100ms offsets. From all motion sensors we compute the following 32 features: statistical features (mean, covariance, derivative), temporal

shape features (coefficients of 4th order polynomial fit to acceleration values), and frequency features (zero crossing rate and its standard deviation).

5.2 Classification and evaluation

We focus on finding meal periods while varying the training data to understand how data from each setting affects accuracy in daily life. We classified each frame (combining feature vectors from each sensor) using random forest with 100 trees.

Lab data is annotated in detail, so classifiers are trained for chew and intake. Since free-living data is labeled at the meal level (with only start and end), the only label that can be used for classification on that data is eating. The classification leads to a probability of each event during each frame (based on the frequency of the label among the trees). The threshold is that labels must be in the 99.9th percentile (computed for each session individually), followed by a drop to below 50%.

We use the following training data. First, for ACE-FL we vary training data to understand portability: train only on lab (TOOL), trains on laboratory data (10 sessions) and evaluates on each of the ten ACE-FL sessions. Train on lab plus personal free living (TOFL) trains on the 10 lab sessions plus one of an individual's ACE-FL sessions and evaluates on the other. This is repeated for all 10 ACE-FL sessions (so each is alternately training and testing). Leave one free-living person out (LOFPO) leaves out both of an individual's ACE-FL sessions, training on the 10 lab sessions and 8 remaining FL sessions, testing on the held out two. Finally, leave one free-living session out (LOFSO) trains on all lab sessions and all but one ACE-FL session, evaluating on the held out one, repeating this 10 times. For the ACE-E generalizability test, we use a similar approach, with TOOL-E being training on ACE lab data and testing on ACE-E; TOAFL training on ACE lab and ACE-FL; and leave one E session out (LOESO) training on all ACE lab, ACE-FL, and ACE-E sessions but one and evaluating on the held out one. For the experiments with just the five day participant we leave one session out (LOSO), and add the following other training data: ACE lab (TAT5), ACE-FL and ACE-E (TFET5), and ACE-E (TFT5).

For evaluation, we construct meals from the classified events, combining those with a gap of less than 1 minute between them to form a meal. Since the eating label is noisier than chew/intake, we require meals based on it to be >5 minutes long or contain at least one inferred chew or intake. For inferred chews and intakes, meals are constructed by combining events with a gap of less than 1 minute, ensuring that all meals have at least one chew and intake. We then take the union of all inferred meals. We evaluate precision and recall by testing for exact matches between ground truth and inferred eating times at the level of deci-seconds (rather than coarser windows). For meal duration: Precision= $\frac{\#\text{correct eating found}}{\#\text{all eating found}}$ and Recall = $\frac{\#\text{correct eating found}}{\#\text{all actual eating}}$. We evaluate accuracy with a 20:1 weighting of true positives due to the severe class imbalance, and to compare with prior work [12].

6 RESULTS

6.1 Portability of lab-generated models

First, we examine how well classifiers trained on lab data fare on detecting the same people eating during daily life. This evaluation, TOOL, tells us about how well lab data generalizes to real-world environments. For the same task, using LOSO evaluation in ACE, we previously achieved 88% precision and 87% recall with all sensors, and 92% and 89% with audio alone [14]. As shown in table 1, ACE-FL sessions have lower average precision (45%), though recall is similar at 85%. Results are based on percent of time that is correctly classified, so any portion of a meal not found is a false negative, and any extra minutes of eating are false positives. Precision is lower in part due to imbalanced classes. At the meal level, 2 of 31 meals had less than 50% overlap with the true meal, both being snacks of chocolate that were entirely missed (one was 0.2min and below our meal duration threshold, and the other 4.8min). It may be that brittle, melting, foods are harder to recognize, particularly when consumed quickly. While we expected restaurant meals to be more difficult to recognize given the potential for

Table 1. Results for ACE-E/ACE-FL participants, varying the amount of FL and personal training data, given in percent. Note that LOFPO and LOFSO leave out only FL sessions. From left to right, each study increases the amount of FL training data.

		Only lab (TOOL)		Lab + 1 FL (TOFL)			Leave one FL person out (LOFPO)		Leave one FL session out (LOFSO)		
Subj.	Sess	Precision	Recall	Precision	Recall	Acc.	Precision	Recall	Precision	Recall	Acc.
2	0	86	96	63	100	98	51	96	50	96	94
2	1	48	91	27	91	90	31	91	29	91	90
3	0	27	75	32	100	92	22	77	24	78	74
3	1	40	80	39	87	86	38	85	37	82	82
4	0	70	86	55	99	96	46	86	44	86	85
4	1	40	98	31	98	92	27	98	29	98	90
5	0	36	66	46	100	96	24	66	23	66	70
5	1	48	82	31	95	91	30	95	30	95	90
6	0	22	86	25	100	93	14	86	15	86	81
6	1	28	88	28	89	86	27	96	27	96	89
Mean		45	85	38	96	92	31	87	31	87	85

background noise and different behavior than in the lab, this was not the case. In fact, all restaurant meals had at least 92% overlap with the true meals and none were missed. Entirely false meals (i.e. inferred meals with zero overlap with a true meal) tended to be short, with an average duration of 4.6min (s.d. 3.2min). We did not set a stricter minimum meal duration, as we aim to infer snacks and the other brief snacks (<6min) also had 100% recall. However, 67% of falsely detected meals were under 5 minutes, so results may be improved with a minimum meal duration and a separate classifier for snacks. Our data, though, did not contain enough snacks to partition this way.

The standard deviation of precision is 20% and recall 10%. Rather than all sessions proving equally difficult for eating recognition, some achieve results on par with ACE, while others such as for participants 3 and 6 consistently have lower precision. Both participants had multiple extended shared meals with family or friends, suggesting that this behavior is different enough that training data specific to such situations is needed. We also require each meal to have at least one intake and at least one chew, and for participant 3, few intakes were recognized, further hampering performance. Participants chose their data collection days, which may affect accuracy, as they contain a mix of weekend and weekday activities (participant 3 did two consecutive weekend days). While most sessions were over the same summer period, participant 6 did the second day of FL data collection a few months later.

6.2 Utility of FL training data

Next we examine how adding free-living training data affects inference. TOFL adds one FL session to the training data for each individual and evaluates on the participant's other session. While this is a relatively small amount of additional data, it led to an increase in recall (moving from 85 to 96% on average), with only a 7% decrease in precision. Personalized data is known to benefit performance, but it is unexpected how little data is needed to achieve this. In fact, 40% of sessions have 100% recall in TOFL, and 48% of meals do. The FL data was particularly beneficial for participants 3 and 5, who had lower accuracy in TOOL, suggesting that they have unique FL behavior. Neither ate in restaurants, suggesting there is indeed a difference between eating at home and in lab.

For 3, this is consistent with their long and shared meals in FL. At the meal level, now only one meal is missed, the 0.2min chocolate. As in TOOL, false meals were short, averaging 5.3min (s.d. 4.5min).

LOFPO conversely examines the impact of non-personal free living data, adding 8 FL sessions from other participants. Despite the larger amount of data (8 vs 1 session), recall improves only 2% from TOOL while precision drops for every session (14% on average). The improvement in recall came from three subjects (3, 5, and 6), who had the most shared meals. In our lab data, since there is already free choice of foods, there is a smaller gap between lab and real life, so simply adding more data from others was only beneficial when the eating environment differed substantially from the lab.

Finally, LOFSO includes all training data of TOFL and LOFPO, merging personal and group FL data (9 sessions total). We find that results are on average the same as LOFPO, and only a 2% improvement in recall over TOOL, at the expense of a 14% drop in precision. This is our most unexpected result, as it indicates that while personalized data helps (TOFL), non-personalized data can actually hurt if all are pooled together. With 8 non-personal FL days and 1 personal, it appears that the personalized data is drowned out. This suggests that future work should aim to reduce the amount of other data used as more personal data is collected, or to learn general models from the larger dataset, with parameters potentially personalized from the individual data.

It is difficult to compare our results directly to prior work, as most evaluations with personalized data train on data from the same session, which may create an overly optimistic picture [13], while evaluation leaving one session out means impersonal data in most works (as data is rarely collected for individuals across multiple sessions). We also evaluate timings exactly, allowing no leeway if a meal is detected early, and further use multiple sensing modalities while the prior work uses a single modality at a time. Unweighted F-measures for TOOL of 65%, TOFL of 67%, LOFPO of 59%, and LOFSO of 59% compare favorably to FL work that found LOPO F-measures of 28.7% (79.8% when using 10-fold cross validation, and not leaving any whole sessions out) [27]. The closest approach, using time correctly classified, is that of Dong et al. [12]. Using that paper's 20:1 weighting of true positives to true negatives (as there are few eating instances throughout a day), accuracy is 86% for TOOL, 92% for TOFL, 85% for LOFPO, and 85% for LOFSO, which are all above their reported 79-81% and significantly so in the case of TOFL. We report per-session accuracy with this measure in table 1 for TOFL (highest values) and LOFSO (most training data). Values for TOFL are all over 90% except for a session for participants 6 and 3. Session 1 for participant 6 did not have unusual foods, but involved extended conversation and multitasking (doing homework as a group) that may have affected results. Additionally, there is no relationship between complexity of meals (number of food types/textures) and recall or precision (R^2 for recall/number of food types is 0 for all experiments and R^2 for precision ranges from 0.09 to 0.16). Moving forward, rather than expanding the set of foods captured, it may be necessary to make efforts to intentionally capture more variety in eating context.

6.3 Comparison of sensors

Our hypothesis was that audio would perform better in lab than in FL, while motion sensors would be more consistent across environments, due to interference from background noise. This is confirmed, though we also find unexpectedly that while audio and motion sensors combined are most accurate, the exact motion sensor used does not make a large difference. In ACE lab, audio had 92% precision and 89% recall for identifying meals, while Glass alone had 73% precision and 48% recall, and the right watch had 93% precision and 42% recall. Figure 5 shows results comparing all combinations of sensors with LOFSO evaluation for ACE-FL. Even though it was not the best performer, we used LOFSO as it has the most training data, and will be less sensitive to outlying sessions. The figure highlights combinations with audio in darker green. We now find that for single sensors, audio has lower accuracy than the right watch, with a different tradeoff on precision and recall (audio 12% and 99% and right watch 47% and 72%). Glass was more balanced, with precision and recall of 55% and 54% respectively. Thus while audio retains high recall in FL, finding most eating, precision drops dramatically from lab, due to factors such as

Fig. 5. Comparison of each modality (A=Audio, G=Glass, R=right watch, L=Left watch) for ACE-FL LOFSO evaluation. Highest and lowest weighted accuracy, and those for key combinations are highlighted. All participants in ACE-FL are right-handed.

pauses in meals with no eating sounds. Recall for the motion sensors actually increases in FL, though precision drops. Overall, audio combinations had the highest recall, while precision was higher for motion sensing. Any combination of audio and motion sensing, though, achieves a weighted accuracy of over 80%, as the audio boosts recall while motion boosts precision. By a slight margin, audio plus the two watches was the best performing combination, though the watches were essentially interchangeable, with no difference at all when combined with audio, and only 1% difference when combined with audio and Glass. Note that this was only true for the combination, whereas the right watch alone was much better than the left watch alone. This is encouraging, as it is thought to be best to instrument the dominant wrist, while users generally prefer to wear devices on their non-dominant wrist. The worst performer was Glass plus the left watch (all ACE-FL subjects were right-handed, though in the lab data we observed and previously reported that both hands were used in eating [14]).

These results suggest that while audio is sufficient in the lab, combinations of audio and motion sensing – no matter where the motion sensors are placed – are necessary to maintain performance outside the lab.

6.4 Generalizability

The ACE-FL dataset provided insight into portability of models learned in the lab for recognizing eating in the same individuals in the wild. We now test generalizability: how well do models trained on one population work for a second very different population? In this section we use all ACE-E data, while in the following one, we focus solely on the 5-day participant. First, TOOL-E replicates TOOL, except we now test on ACE-E rather than ACE-FL. Results in table 2 show that precision drops compared to TOOL (from 45 to 31%) while recall drops only 4%. Thus models trained in the lab capture a substantial amount of eating behavior for other populations. While 38% of meals have 100% recall, there were 9 meals with low recall (under 50%), with 7 of those missed entirely and the other two having 49% and 34% recall. The missed meals included 3 meals under 3min (whipped cream, cookies, blackberries), 2 meals of 5-6min (cantaloupe and grapes, trail mix), a 9min meal (apple), and an 18min meal (2 hardboiled eggs). All but one missed meal was consumed at home, and three of the missed meals plus one with low recall came from participant 104. Even though she consumed all meals at home, the slower meal pace (e.g. 2 eggs consumed over 18min) may have made recall more challenging due to infrequent intakes and

Table 2. Results for ACE-E participants, comparing lab (TOOL-E), impersonal FL data (TOAFL), and all data (LOESO).

	Only lab (TOOL-E)			Lab + ACE-FL (TOAFL)			Leave one ACE-E session out (LOESO)		
Value	Precision	Recall	Acc.	Precision	Recall	Acc.	Precision	Recall	Acc.
Mean	31	81	80	28	84	80	25	83	79
S.D.	17	14	10	17	13	10	14	14	10

chews. As in ACE-FL, we find restaurant meals do not have lower accuracy, and it is instead home environments, which may involve more multitasking, that prove more challenging.

Next, training on ACE lab plus ACE-FL (TOAFL eval), akin to LOFPO, increases recall 3%, while reducing precision by the same amount. One difference is that here there is no personal data at all, since ACE-E participants are not in the lab dataset. The increase in recall came from three sessions (two participants), including 104. By incorporating FL data in training, we now recover 42% of the hardboiled egg meal, and move from 56% to 96% recall on a 31min meal (pizza eaten outside) by another participant. Thus while free-living data from others is not beneficial in all cases, it can make a difference when meal context differs substantially from lab, even if the specific foods were not recorded in the FL data (pizza was consumed in ACE lab, but not ACE-FL).

Finally, we increase the amount of FL data, training on all data minus one session and evaluating on the held out session (leave one ACE-E session out, LOESO). As for the lab participants, we find that more data does not actually improve results: from TOOL-E to TOAFL to LOESO, precision drops at every step. For LOESO, recall also drops 1% from TOAFL. As before, the impersonal data tends to reduce the benefit of the personal session. Further, LOESO includes 24 days of FL data in training, so even substantially more training data does not improve results. The hardboiled egg meal, for example, drops to 12% recall, while some others have slight increases. As for ACE-FL, we did not see any reduction in performance in restaurants, and again all restaurant meals have over 79% recall. There was more variation in outdoor meals, with one missed entirely and the 31min pizza meal having 56% recall in LOESO and TOOL-E, while the other outdoor meals had 100% recall. Based on this, we believe that with more training data in different environments, it may be possible to use GPS or other information to determine location, and then select between classifiers trained by environment.

6.5 Personalization

The experiments described thus far are designed to evaluate the impact of environment and sensor on results, and determine how portable results are to a new population. We now further examine the role of personalized training data in more depth, focusing on the 5-day ACE-E subject. With 5 days of data, we can test fully personalized models with leave-one-session-out (LOSO) evaluation. As shown in table 3, despite the small amount of data, LOSO precision is 80%, and recall is 40%, leading to a weighted accuracy of 64%. While this is still a relatively small amount of data (the participant had 3.1h eating total), with noisy labels (participant logged meals such as ice cream whose times could not be found, and had audible gaps between eating during a picnic), it achieves competitive results, with three sessions having over 99% precision. Further, this is the only FL experiment where precision was higher than recall. All but one missed meal was consumed at home, though one meal with low recall (22% of 31min) was eating pizza outside. There were two meals of pizza consumed outdoors on different days (the other having 73% recall of 15min), so it is possible there are long gaps in the meal that are properly being excluded as eating activity. There was no pattern to the missed meals by food texture, as they included crunchy (e.g. cookies, chips), soft (e.g. banana), and wet crisp (e.g. grapes, salad) foods. While some were consumed on multiple days (cookies), the meals may have been too brief to provide sufficient training data. This confirms findings in the experiments above, showing that with a wide variety of meals in lab, environment rather than food type appears to be responsible for accuracy in FL environments.

Table 3. Results for ACE-E 5-day participant. Here each experiment trains on 4 of the participant's 5 days, and tests on the other. The first three experiments use data from the ACE participants: both FL and lab (TAT5), just lab (TAT5), ACE-E FL (TFT5), and all FL (TFET5). The last uses only 4 days from the same participant and tests on the 5th (LOSO).

Session	Lab + 4 days (TAT5)			ACE-(FL+E) +4 days (TFET5)			ACE-E + 4 days (TFT5)			4 days (LOSO)		
	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.
0	19	87	79	65	53	68	96	45	63	100	25	49
1	21	96	88	27	70	79	91	68	82	99	52	74
2	28	70	75	40	42	61	55	25	52	52	21	49
3	16	85	78	10	35	54	22	45	64	47	29	58
4	8	100	77	41	75	87	100	76	90	100	73	89
Mean	18	88	79	37	55	70	73	52	70	80	40	64

As before, we also examine how other types of impersonal training data may improve results, with the difference here that there is much more personal data (4 days vs 1 training day). Each addition of more free-living data from others increases recall (to 52% by adding the other ACE-E data, and 55% using all ACE-FL and ACE-E data), with a significant penalty in precision. While the rest of the ACE-E data (10 more days of FL data) leads to only a 7% reduction in precision, adding ACE-FL as well (20 days total FL data from others) drops precision to 37%. This suggests the ratio of personal to impersonal data is critical. Finally, we test combining ACE lab with the 4 personal days (TAT5). This is akin to TOFL, but with 4 days of personal data, rather than one, and with fully impersonal lab data. Even though the lab data was collected for a different set of participants, once again this increased recall, to 88%, similar to our findings in TOFL. Only one meal was missed (cantaloupe and grape snack) and no others had under 50% recall. Comparing these results to LOESO only for this subject, two sessions have 2% lower precision, while one meal goes from 75 to 100% recall by using only the personalized FL data. Thus with four days of personal data, rather than only one as in ACE-FL, the benefit of personalization is retained. Our results also suggest that lab data is highly generalizable, but more work is needed to retain the boost to recall, without incurring a cost in precision. Work on activity recognition generally shows that personalized data is highly beneficial, but we aimed to also test whether non-personal data can be beneficial even when personal data is available. Ultimately we find there is a small benefit to including free-living data from others, and the highest precision comes from solely personalized data, while lab data can have a large impact on recall. This suggests that in practice it may be possible to train highly accurate systems by having individuals log a small number of their meals, though more work is still needed to determine exactly how much data is required.

7 DISCUSSION

7.1 Sensors and training data

Prior work has noted the value of personalized training data for activity and eating recognition, but it is also often assumed that more data is better. In contrast, we find that extremely small amounts of personalized data (a single day) can improve accuracy more than substantially larger non-personalized data sets (24 days), even when annotations are noisy. Further, we find that not only is personalized data beneficial, but that including non-personalized data can actually lead to worse results. By collecting data for the same individuals in lab and daily life, over multiple days, we avoid potentially overstating the benefit of personal data. For example, food choice and context (e.g. stress, eating environment, social factors) may play a role, in theory making a rushed breakfast for person A more similar to person B's quick breakfast, than to A's usual eating style. When only a single day is available for A, we cannot distinguish between person and setting. In our studies, individuals generally ate completely distinct foods during each data collection day, so the value of personal data is not

explained by potentially learning an individual's behavioral patterns. Further, we find that while lab and restaurant meals were easier to recognize, meals at home had higher variance in results, which could not be explained by meal duration or food type. For future work, it is necessary to conduct longer-term studies on individuals (rather than increasing participant number only) to understand at what point accuracy plateaus, and to explore whether non-personalized data can be used to identify general model structures for which personalized parameters can be learned.

A second key finding is that while multiple sensing modalities (audio, motion) are needed to achieve the highest accuracy, the specific location of the motion sensor (head, dominant wrist, non-dominant wrist) was not a factor. In contrast, while audio sensing alone was best in the lab, it had the largest drop in precision in daily life, despite our use of two microphones and successful noise cancellation. Thus, studies evaluating single sensing modalities in the lab may not be sufficient for gaining insight into how sensors compare in daily life. Our findings on sensor combinations suggest that it may be possible for people to wear sensors in their preferred location, rather than having to trade accuracy versus comfort. However, for practical future use, it will be necessary to integrate the multiple sensing modalities into a single device, to avoid requiring users to wear multiple sensors.

7.2 Evaluation

One of the key challenges for training a classifier based on free-living data is lower quality ground truth, and this is similarly a problem for evaluating eating recognition methods. Many meals involve gaps due to talking or other factors, so participants are not continuously chewing or manipulating food. Thus we expect that recall will often be less than 100% even for a perfect system when applied to FL data, since it will be penalized for not finding those eating periods – which are not actually eating. POV cameras may help somewhat, but can only show whether the environment looks like a meal – not whether the individual is still eating at each timepoint. More work is needed to understand the right metrics for comparison, and whether eating duration (rather than foods consumed or their quantity) is the right level of granularity. We evaluate inferences strictly, at the level of seconds, as our ultimate application goal is to advance real-time decision-support systems. However, for other applications different metrics may be appropriate, and more work is needed to determine how best to compare systems.

7.3 Limitations

The main limitation of our work is that the data comes from only 11 participants over a short timespan, so we cannot draw conclusions about the behavior of all adults (or even the same adults longterm), and it is unknown how eating behavior differs in populations with chronic disease (e.g. diabetes, obesity). Annotation time was the primary factor limiting sample size (annotating each lab session took each researcher around 8 hours and two researchers annotated each). Thus while we need to increase sample size (person-days recorded), we also need to find strategies to reduce the burden of human labeling, such as with active learning or video analysis. Our findings in the LOSO 5-day experiment suggest that noisier labels can still be useful, so one strategy may be using more participant logs, with only spot checking rather than confirming all meals exactly.

8 CONCLUSION

We described a novel set of multimodality data for identifying eating activities and investigating the differences between performance in laboratory and real-world environments. We collected acoustic (earbud microphone) and motion (head with Google Glass, wrist with smartwatch on each) data for 5 people in the lab and over two 12-hour days each in their natural environments, and tested generalizability with a second set of 5 2-day participants and one 5-day participant. We provide a first step toward truly reconciling the difference between accuracy in controlled and realistic environments, and ultimately bringing real-world performance closer to that

achieved in lab. This work led to three key findings. First, small amounts of well-annotated laboratory data can be used successfully for classification in free-living settings, and can generalize to new participants in unconstrained environments. Second, while personalized data was key to high precision and successful with even small amounts of such data, impersonal FL data can actually worsen results even when mixed with personal data. Finally, while using only audio sensors was best in the lab, we find a combination of motion and audio sensing is needed for the best accuracy in FL, but multiple types of motion sensing give similar results. Overall by combining laboratory data with one personal free-living day, we achieve an average weighted accuracy of 92%.

A key advance of our work over the state of the art is the collection of lab and free-living data for the same set of individuals using multiple sensing modalities, and test of generalizability on a second more diverse sample. Prior work has focused on lab or free-living environments, and usually only a single sensor at a time. While it is expected that accuracy will decrease moving from lab to the real world, the gap between studies in each has made it difficult to determine why exactly this is. By making the data publicly available, we also aim to enable others to add to and build upon this work. For automated dietary monitoring to be a consumer good like fitness monitoring, we need more data from more diverse individuals and better machine learning methods that can make use of weakly labeled (or even unlabeled) datasets. More work is needed on personalized modeling, and especially on eating in different environments, while also reducing the burden of collecting and labeling such data.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under award number 1347119. We also thank the research participants who made this study possible.

REFERENCES

[1] Oliver Amft, Holger Junker, and Gerhard Tröster. 2005. Detection of Eating and Drinking Arm Gestures Using Inertial Body-worn Sensors. In *Proceedings of the Ninth IEEE International Symposium on Wearable Computers (ISWC '05)*. 160–163. https://doi.org/10.1109/ISWC.2005.17

[2] Oliver Amft, Martin Kusserow, and Gerhard Tröster. 2009. Bite Weight Prediction From Acoustic Recognition of Chewing. *IEEE Transactions on Biomedical Engineering* 56, 6 (2009), 1663–1672. https://doi.org/10.1109/TBME.2009.2015873

[3] Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. 2005. Analysis of Chewing Sounds for Dietary Monitoring. In *Proceedings of the 7th International Conference on Ubiquitous Computing (UbiComp'05)*. 56–72. https://doi.org/10.1007/11551201_4

[4] Oliver Amft and Gerhard Tröster. 2009. On-Body Sensing Solutions for Automatic Dietary Monitoring. *IEEE Pervasive Computing* 8, 2 (April 2009), 62–70. https://doi.org/10.1109/MPRV.2009.32

[5] Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. Detecting Mastication: A Wearable Approach. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction (ICMI '15)*. 247–250. https://doi.org/10.1145/2818346.2820767

[6] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-Match: Restaurant-Specific Food Logging from Images. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV '15)*. 844–851. https://doi.org/10.1109/WACV.2015.117

[7] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D. Abowd, and Irfan Essa. 2015. Leveraging Context to Support Automated Food Recognition in Restaurants. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV '15)*. 580–587. https://doi.org/10.1109/WACV.2015.83

[8] Michael Buettner, Richa Prasad, Matthai Philipose, and David Wetherall. 2009. Recognizing Daily Activities with RFID-based Sensors. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp '09)*. 51–60. https://doi.org/10.1145/1620545.1620553

[9] Lora E Burke, Jing Wang, and Mary Ann Sevick. 2011. Self-monitoring in weight loss: a systematic review of the literature. *Journal of the American Dietetic Association* 111, 1 (2011), 92–102. https://doi.org/10.1016/j.jada.2010.10.008

[10] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. 2015. Barriers and Negative Nudges: Exploring Challenges in Food Journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. 1159–1162. https://doi.org/10.1145/2702123.2702155

[11] Yujie Dong, Adam Hoover, Jenna Scisco, and Eric Muth. 2012. A new method for measuring meal intake in humans via automated wrist motion tracking. *Applied Psychophysiology and Biofeedback* 37, 3 (2012), 205–215. https://doi.org/10.1007/s10484-012-9194-1

[12] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover. 2014. Detecting Periods of Eating During Free-Living by Tracking Wrist Motion. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (July 2014), 1253–1260. https://doi.org/10.1109/JBHI.2013.2282471

[13] Nils Y. Hammerla and Thomas Plötz. 2015. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. 1041–1051. https://doi.org/10.1145/2750858.2807551

[14] Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. 2016. Multimodality Sensing for Eating Recognition. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '16)*. 130–137. http://dl.acm.org/citation.cfm?id=3021319.3021339

[15] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. 2016. Automated Estimation of Food Type and Amount Consumed from Body-worn Audio and Motion Sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. 451–462. https://doi.org/10.1145/2971648.2971677

[16] Marie Ng, Tom Fleming, Margaret Robinson, et al. 2014. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 384, 9945 (2014), 766–781. https://doi.org/10.1016/S0140-6736(14)60460-8

[17] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. 2011. Platemate: Crowdsourcing Nutritional Analysis from Food Photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. 1–12. https://doi.org/10.1145/2047196.2047198

[18] Sebastian Päßler and Wolf-Joachim Fischer. 2011. Acoustical method for objective food intake monitoring using a wearable sensor system. In *Proceedings of the 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '11)*. 266–269. http://ieeexplore.ieee.org/document/6038810

[19] Shah Atiqur Rahman, Christopher Merck, Yuxiao Huang, and Samantha Kleinberg. 2015. Unintrusive Eating Recognition using Google Glass. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '15)*. 108–111. http://dl.acm.org/citation.cfm?id=2826165.2826181

[20] Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. 2014. BodyBeat: A Mobile System for Sensing Non-speech Body Sounds. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*. 2–13. https://doi.org/10.1145/2594368.2594386

[21] Tauhidur Rahman, Mary Czerwinski, Ran Gilad-Bachrach, and Paul Johns. 2016. Predicting "About-to-Eat" Moments for Just-in-Time Eating Intervention. In *Proceedings of the 6th International Conference on Digital Health (DH '16)*. 141–150. https://doi.org/10.1145/2896338.2896359

[22] Raul Ramos-Garcia, Eric Muth, John Gowdy, and Andrew Hoover. 2015. Improving the Recognition of Eating Gestures Using Intergesture Sequential Dependencies. *Journal of Biomedical and Health Informatics* 19, 3 (2015), 825–831. http://ieeexplore.ieee.org/document/6826479/

[23] Jenna L. Scisco, Eric R. Muth, and Adam W. Hoover. 2014. Examining the Utility of a Bite-Count–Based Measure of Eating Activity in Free-Living Human Beings. *Journal of the Academy of Nutrition and Dietetics* 114, 3 (2014), 464–469. https://doi.org/10.1016/j.jand.2013.09.017

[24] Sherry A Tanumihardjo, Cheryl Anderson, Martha Kaufer-Horwitz, et al. 2007. Poverty, obesity, and malnutrition: an international perspective recognizing the paradox. *Journal of the American Dietetic Association* 107, 11 (2007), 1966–1972. https://doi.org/10.1016/j.jada.2007.08.007

[25] Edison Thomaz, Irfan Essa, and Gregory D. Abowd. 2015. 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 (UbiComp '15)*. 1029–1040. https://doi.org/10.1145/2750858.2807545

[26] Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D. Abowd. 2013. Feasibility of Identifying Eating Moments from First-person Images Leveraging Human Computation. In *Proceedings of the 4th International SenseCam and Pervasive Imaging Conference (SenseCam '13)*. 26–33. http://doi.acm.org/10.1145/2526667.2526672

[27] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D. Abowd. 2015. 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 (IUI '15)*. 427–431. http://doi.acm.org/10.1145/2678025.2701405

[28] Rui Zhang and Oliver Amft. 2017. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE Journal of Biomedical and Health Informatics* (April 2017). https://doi.org/10.1109/JBHI.2017.2698523

[29] Rui Zhang, Severin Bernhart, and Oliver Amft. 2016. Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses. In *Proceedings of the 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN '16)*. 7–12. https://doi.org/10.1109/BSN.2016.7516224

Received November 2016; revised May 2017; accepted June 2017