Automated Estimation of Food Type from Body-worn Audio and Motion Sensors in Free-Living Environments

Mark Mirtchouk MMIRTCHO@STEVENS.EDU
Computer Science
Stevens Institute of Technology
Hoboken, NJ, USA

Dana L. McGuire DMCGUIR1@STEVENS.EDU
Computer Science
Stevens Institute of Technology
Hoboken, NJ, USA

Andrea L. Deierlein ALD8@NYU.EDU
College of Global Public Health
New York University
New York, NY, USA

Samantha Kleinberg SAMANTHA.KLEINBERG@STEVENS.EDU
Computer Science
Stevens Institute of Technology
Hoboken, NJ, USA

Abstract

Nutrition is fundamental to maintaining health, managing chronic diseases, and preventing illness, but unlike physical activity there is not yet a way to unobtrusively and automatically measure nutrition. While recent work has shown that body-worn sensors can be used to identify meal times, to have an impact on health and fully replace manual food logs, we need to identify not only when someone is eating, but what they are consuming. However, it is challenging to collect labeled data in daily life, while lab data does not always generalize to reality. To address this, we develop new algorithms for semi-supervised hierarchical classification that enable higher accuracy when training on data with weak labels. Using this approach, we present the first results on automated classification of foods consumed in data collected from body-worn audio and motion sensors in free-living environments. We show that by exploiting a mix of lab and free-living data, we can achieve a classification accuracy of 88% on unrestricted meals (e.g. stir fry, pizza, salad) in unrestricted environments such as home and restaurants. Ultimately, this lays the foundation for body-worn devices that can calculate calories and macronutrients by identifying food type and quantity.

1. Introduction

Nutrition is key to preventing and managing many chronic diseases, but it can be challenging to gain objective insight into food consumption. While physical activity can be easily quantified, existing solutions for monitoring nutrition place a large burden on the user. Currently, a consumer who wants to track their food consumption must input information themselves using a database app, such as MyFitnessPal, if they want detailed information

on macronutrients and calories consumed, or they may take photos of their meals, which again requires a user's action. In addition to the time burden, self-reporting is susceptible to reporting errors (e.g. misjudging portion size, forgetting to record a meal). Instead, an automated approach could lead to many mobile health interventions such as for managing weight (e.g. food suggestions when a user is not meeting targets for macronutrients or calories) and diabetes (reminders for missed insulin with meals, estimation of insulin needs, automated input to an artificial pancreas system). However, these applications require detailed information about each meal's contents. For example, insulin needs depend not only on carbohydrates but also fat and protein and the order in which these macronutrients are consumed affects glycemic response (Shukla et al., 2015).

Automated dietary monitoring aims to develop systems that can objectively assess nutrition, with little or no user input. While there is not yet a commercially available automated solution, recent work has shown that eating can be recognized with body-worn sensors: audio to identify chewing noises (Amft et al., 2005b), motion sensors to infer intakes of food or drink from wrist movements (Scisco et al., 2014), and combinations of these sensors to provide more robust inferences in daily life (Mirtchouk et al., 2017). However, these works have focused on identifying meal periods, rather than meal contents. Efforts to find foods consumed using body-worn sensor data have been done in the lab with small sets of constrained foods rather than full meals (Rahman et al., 2014), with one exception (Mirtchouk et al., 2016). There have been image-based approaches for food detection (Meyers et al., 2015), however these require users to photograph meals, and are thus not automated.

While prior work has shown the possibility of using body-worn sensors to identify foods consumed in a lab environment, this has not been done in the unconstrained environments of daily life. A core challenge is obtaining training data: lab data can be annotated in detail from video but lacks the breadth of activity found in daily life, while free-living (FL) data is representative but comes with noisy labels. To address this, we introduce new algorithms for semi-supervised hierarchical classification that provide higher accuracy when training on data with a mix of both trustworthy and uncertain/high level labels, by exploiting the structure between class labels. This also allows output at varying levels of granularity depending on the classifier's confidence (e.g. meat, protein, chicken). On FL data (body-worn audio, motion sensors) from 11 people over 2-5 days each consuming unrestricted meals we achieve a food type classification accuracy of 87.7% in unrestricted environments (e.g. home, picnic, restaurant). This surpasses the prior best of 82.7% in the lab with similarly complex foods (e.g. salad, stir fry) (Mirtchouk et al., 2016).

Technical significance We introduce a novel semi-supervised approach to hierarchical classification with a mix of strong and weakly labeled data. Unlike the state of the art in semi-supervised learning, we are able to exploit aggregate labels (e.g. a meal contained steak and potato) and strongly labeled data even with no overlap in the weak/unlabeled data (e.g. instances of chicken, beef) to create fine-grained training labels (e.g. bite at time 10 is steak). This approach may be applicable to many problems in healthcare where data is labeled at varying levels of granularity, such as ICD10 diagnosis codes, where we must decide whether to aggregate for example controlled and uncontrolled diabetes.

Clinical relevance We present the first study showing that foods consumed in each bite can be reliably inferred in unrestricted environments with unrestricted foods using data from body-worn sensors. While our classifier may output general categories (e.g. pork,

meat) rather than specific foods (e.g. bacon) when less certain, this is still a significant advance over approaches that only count bites or identify meal times. Food groups are often examined in nutrition epidemiology research into diet-disease associations such as in Type 2 diabetes (Schwingshackl et al., 2017), so our output is highly clinically relevant. Currently, dietary self-reports are a primary tool to assess diet long-term in nutritional epidemiology (Subar et al., 2015) and our approach could be used to augment self-reports, by prompting individuals to fill in missing details or corroborating logs. Further, being able to measure nutrition automatically could have significant impacts on real-time health interventions in diabetes (reminders for missed insulin boluses, estimation of insulin needs) and research into determinants of health (e.g. dietary risk factors for disease).

2. Related Work

Our goal in this paper is to classify foods consumed in each intake during meals in daily life. This is critical for our application areas, such as feedback on macronutrient deficiencies, detection of overeating, and decision-support for individuals with diabetes. Given the difficulty in obtaining a large amount of high quality labels, we need methods that can make use of a mix of strong and weakly labeled data. Thus we review two primary lines of related work: advances in automated dietary monitoring, and classification methods relevant to our problem structure (weak and strong labels, exploiting shared information between classes).

2.1. Automated Dietary Monitoring

Most works, regardless of output (e.g. meal times or foods consumed), use a single sensor in a constrained environment, so we review related work by sensor.

Audio sensors Acoustic sensors have been used to recognize chewing sounds, but the majority of work has focused on identifying whether or not someone is eating, rather than what they are eating. Some works have attempted to classify food intakes in lab environments where participants all ate the same set of foods, though these generally used a small sample of food types (as few as 2 (Yatani and Truong, 2012) or 4 (Rahman et al., 2014)) or more food types but fewer participants (3 subjects and 19 foods (Amft and Tröster, 2009)). Since data was collected in quiet lab environments with simple foods, accuracy with realistic background noise and full meals (where multiple foods are combined) has been unknown.

Motion sensors While audio sensors can fail for soft foods and noisy environments, motion sensors (on the head, wrist, or throat) may be usable in a wider range of settings. Wrist-based sensors are the most frequently used for detecting eating in FL (Thomaz et al., 2015), and have been used to identify both eating gestures (Amft et al., 2005a) and drinking episodes (Schiboni and Amft, 2018). While motion alone has not been used to identify *what* is being consumed, wrist-motion has been used to estimate calorie intake (Scisco et al., 2014), by identifying intakes and assuming a fixed calorie content per bite. However, for our applications, it is critical to identify the specific foods consumed rather than only estimating calorie content, and calories further vary significantly by food and bite size.

Multimodality sensing Due to the limits of individual sensors, a combination of modalities (audio, wrist motion, head motion) has been used to identify foods consumed in each

bite in the lab with 82.7% accuracy for unrestricted full meals (e.g. burger, tacos, salad) (Mirtchouk et al., 2016). A multimodality approach has also improved accuracy for recognizing meal times in FL environments (Mirtchouk et al., 2017) suggesting that this may be important for food type detection in FL. Fontana et al. (2015) used a collar plus sensor behind the ear to sense chewing and swallowing and correlated these features with energy intake, however that work did not identify foods consumed.

Image-based approaches Photos taken by individuals at the start of each meal have been used to estimate intake with manual annotations by nutritionists (Martin et al., 2012) and crowd-workers (Noronha et al., 2011), and automated classification using GPS information and a CNN to match images to restaurant menu items (Meyers et al., 2015). While the latter approach is automated, it relies on users being at an identifiable restaurant for the highest accuracy. Most critically, image-based methods have only been used with images from the start of a meal. Yet in a sample of FL data other work found only 55% of meals were fully consumed (Mirtchouk et al., 2017), so this could lead to dangerous overestimates of insulin needs if used in decision support.

2.2. Classification of data with shared structure

Given the lack of generalizability of laboratory data, and challenges of obtaining granular labels in FL data, we propose that food type classification can be formulated as a semi-supervised learning problem. Foods also naturally form a hierarchical structure. Thus, we review related works in semi-supervised learning and hierarchical classifiers.

Semi-supervised learning (SSL) leverages large amounts of unlabeled data (which provide insight into the data distribution), and small amounts of labeled data (which are used for generating labels). Transductive approaches, such as label propagation (Zhu and Ghahramani, 2002), propagate labels from nodes to their unlabeled neighbors in a graph. Like self-training (Scudder, 1965), though, this assumes the same classes are present in both labeled/unlabeled sets. Inductive approaches aim to learn both labels for unlabeled data and also a classifier (Vapnik, 1998). In both these and recent deep learning approaches (Rasmus et al., 2015), labels are either present or absent, but in much health and activity data, labels may instead be noisy or too coarse (e.g. label for foods in meal, but not each bite). Thus a closer task to ours is when the prevalence of each label for a group of datapoints is known, but labels are not available for individual instances (Quadrianto et al., 2009). However, this would require us to know what proportion of a meal's intakes are from each food. While SSL is a promising approach, there is not yet a method that addresses the problems we face in food type classification from a mix of lab and FL data: we aim to identify foods in each FL intake (when only meal-level labels are available), and to do this even when the FL classes do not exist in the strongly labeled lab data.

Hierarchical Classification To address this challenge, we propose to use shared information between classes. Even with no labeled instances of pork, data from intakes of chicken should allow us to identify bites of pork as meat, since proteins have similar properties. Hierarchical classifiers leverage a structure (e.g. tree) showing the relationship between classes in the data. In our task, leaves are individual foods such as blackberries and strawberries, which share a parent "berries." Top-down approaches train a classifier at each level (Clare

and King, 2003) or node (Wu et al., 2005), and then iteratively traverse the hierarchy, outputting a leaf or, in some cases, internal node (Sun and Lim, 2001). However, early errors can propagate. Global methods train, and minimize loss over, a single classifier (Cesa-Bianchi et al., 2006), but have higher complexity and are sensitive to the loss function. While prior approaches have been effective on large datasets (Gopal and Yang, 2015), in automated dietary monitoring datasets are smaller and classes are highly imbalanced. In the data we analyze here > 12 meals include soup, yet only one has strawberries. Hierarchical decomposition methods were proposed for imbalanced classes (Beyan and Fisher, 2015), but do not address our task, where we aim to exploit rather than discover latent structure. We are not aware of prior work that unites SSL and hierarchical classifiers, but we propose that this is a natural solution to food type classification and may be applicable to many other problems with a mix of label granularities and quality.

3. Methods

We now aim to recognize the type of food consumed in each intake, or whether the intake is a drink, in FL environments with unconstrained food choices and activities. A core challenge is the lack of rigorous ground truth in FL, due to the expense of obtaining labels and difficulty of obtaining continuous video to be annotated. FL data is primarily annotated at the meal level, with meal times and foods consumed in the meal – rather than each bite. While lab data has accurate intake labels, it does not contain the variety of eating behaviors and environments observed in FL settings. We propose that lab data with granular labels can be combined with FL data to leverage the advantages of both. We develop a semi-supervised classifier that first infers intake-level labels for FL data to improve training, and then uses a new approach to top-down hierarchical classification that better allows uncertainty. Unlike existing SSL methods, we also allow the unlabeled data to be labeled at any level of a hierarchy. The following sections introduce the two main components of our algorithm: 1) inferring intake-level labels for training on a mix of strongly (lab) and weakly (FL) labeled data, and 2) test-time classification using the hierarchical structure.

3.1. Labeling FL Intakes

For lab data we are able to annotate the specific foods consumed in each bite since we have video ground truth, while for FL data we only know what was consumed in the entire meal (from photos taken before and after consumption). This gap creates challenges for learning from a mix of these data types. However, we show that intake-level labels can be inferred for FL meals for use in training. This is a classification problem, but it differs from test-time in two important ways: 1) we have a set of potential labels for the meal, drastically reducing the number of possible classes, and 2) intakes can be left unlabeled if there is not sufficient confidence. That is, we do not necessarily want to label every single intake as some may be false positives (not eating). Further, we allow labels at all levels of granularity (e.g. meat instead of chicken wing). By leveraging the meal-level labels we are able to infer granular (intake-level) labels for foods that have never been seen before in lab data, which improves training. For example, pork was not in the lab dataset used in this paper, but because of the hierarchical approach, we are able to infer that it is meat, and in some cases at test time, as we describe in results, we are able to infer specifically that intakes are of pork.

We use a running example of labeling one intake from FL data, and then describe how to do this for all FL data. Note that this is for an intake that will be part of the training set (not the meal to be labeled at test time). FL data is labeled at the meal level, so a meal M may have a photo and food log indicating that it contains pork, salad, and rice. We assume that we have already identified the meal and its intakes, as has been done previously in FL data (Mirtchouk et al., 2017). We now identify which of the meal level labels (pork, salad, rice) applies to intake i at minute 12 of meal M. The output for i may be at the finest level of granularity from the FL label (pork, salad, rice), or at a higher level such as meat.

Correcting class imbalance Before labeling i, we first address the data imbalance, as we may have different amounts of training data on each class. As we traverse the hierarchy top-down, we begin by correcting imbalances in the top level categories for the foods in M. In our running example, using the hierarchy introduced later in the paper, those categories are vegetable, starchy foods, and meat. Say we have 400 lab intakes of starchy foods but only 80 vegetable and 150 protein intakes. We iteratively generate synthetic intakes of each food with fewer than N (parameter for minimum per class) intakes.

As long as there are foods $r \in M$ with fewer than N intakes, we randomly choose a food r and generate a new intake i_{new}. The intake has a set of features, generated from existing samples of r, which includes both real data and previously generated synthetic intakes. We assume each feature $f_k \in \{f_1, f_2, \ldots, f_o\}$ among the existing n samples of food r is normally distributed with mean $\bar{f}_k$ and standard deviation σ_k defined as:

$$\bar{f}_k = \frac{\sum_{j=1}^{n} f_{kj}}{n} \tag{1}$$

$$\sigma_k = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(f_{kj} - \bar{f}_k)}, \tag{2}$$

where f_{kj} is the kth feature of the jth intake. Then the new intake simply selects from these distributions for each feature:

$$i_{new}[k] = \mathcal{N}(\bar{f}_k, \sigma_k). \tag{3}$$

The output is a set of features for the new instance $i_{new} = \{f_1, f_2, \ldots, f_o\}$. To ensure the intake is valid, we apply a random forest (RF) classifier trained on all foods in M. If i_{new} is correctly classified as r, it is added to the dataset, otherwise it is discarded. The classifier is retrained every R intakes, and by generating intakes in random order, we avoid skewing it. See algorithm 1 in the Appendix for details.

In our running example with intake i, we correct the imbalance between protein (150 intakes), vegetable (80 intakes), and starchy foods (400 intakes). We use $N = 250$ (based on the average number of intakes across various foods) and $R = 50$ based on how often there is relevant new information that will improve the classifier. Thus the result at the top level is an addition to the training data of 100 new protein and 170 new vegetable intakes, generated from the relevant distributions and which the classifier labeled as being each type.

Hierarchical classification of FL training data The core of our approach is the hierarchical classification, which determines the best label for FL training intake i. Input to this step is 1) the set of training data, X, where each $x \in X$ is an intake with features extracted from the sensor data or generated in the first step, and label that indicates the food type of x; and 2) a hierarchy of foods, which is a tree structure that need not be balanced. Output is the most specific label that meets our significance criteria.

We use a top-down approach, so we first train a classifier at at the top level of the hierarchy ($l = 1$), using only the categories of the foods known to be in M. We use a separate RF classifier trained on each level, with $2 * z$ trees, where z is the number of unique food types at level l. Thus for the node "meat," the classifier uses data from all descendants of that node (e.g. pork, chicken). The output is a classification (food) from each tree. We then use the percentage of trees predicting a food as the probability of that food type. For a forest of T trees, the probability for a particular food, x, is: $|[t \in T : t = x]|/|T|$.

As we aim to output the most specific labels possible, we do a breadth first traversal of the hierarchy. If a category x at level l meets our criteria for significance, then in the next iteration $l = l+1$ and the root node becomes x (meaning we only consider its children as labels). If one category is not sufficiently significant, the next iteration still begins at $l = l + 1$, but with multiple roots (all nodes at level l). Note that it is possible there will be no labeled data for the food in M at level l (e.g. lab data contains other meats, but not pork). In that case, when there are zero intakes for a category, training data is composed of siblings (e.g. other meats) or children of its parents' siblings (e.g. other proteins if there is no meat). This is one of the fundamental differences between training and test time and allows us to label new foods.

The output of this step can be no label (if none are significant). Otherwise we select 1) the most significant leaf node (if any are significant), or 2) otherwise the most significant internal node. Our significance criteria aims to balance the classifier's confidence with our desire for the most specific label. To be considered significant, a label $x_i \in X$, the set of labels for intake i, must satisfy:

$$\forall j \neq i, p(x_i) - p(x_j) > \delta \tag{4}$$

and

$$p(x_i) > \overline{p(x)} + \sigma(p(x)). \tag{5}$$

Label x_i must be have at least δ higher probability than the other labels for the intake, and the probability must be at least one standard deviation away from the mean of the probability of the set of candidate labels. We do not want to introduce ambiguous labels in training, and this prevents close calls from being accepted (e.g. [0.51 steak, 0.49 chicken]). We chose parameters empirically, and use $\delta = 1/3$.

Thus, to summarize, we do a top-down traversal of the hierarchy, where at each level if a label is significant, we only explore its children on the next iteration. Finally, if there are significant labels we output the most significant leaf node (if there are significant leaves), or else the most significant label at any level. The leaf node exception is because if we use maximal significance, upper level nodes will frequently have higher probability.

We continue the example of labeling intake i from a meal containing pork, rice, and salad. First, we use the generated and actual intakes to train an RF classifier with 6 trees

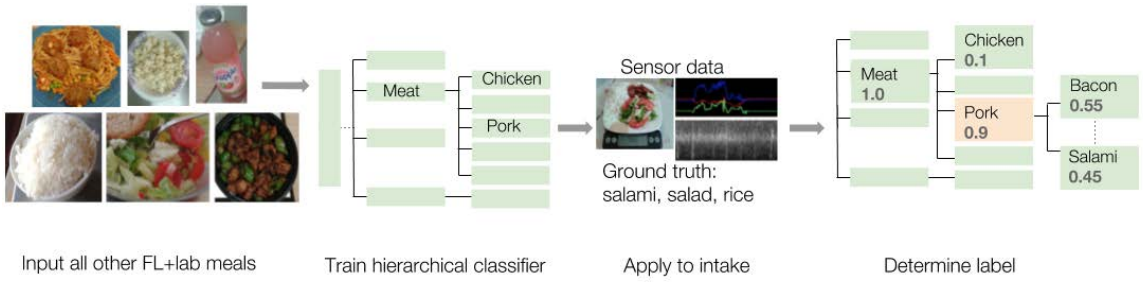

Figure 1: Test-time classification of an intake from an FL meal. The meal contains salami, salad, and rice, and the intake is labeled as pork.

(2 x classes). This classifier aims to select between protein, vegetable, and starchy foods. If the output probabilities are [0.83 protein, 0.17 vegetable], protein is significant so we now repeat the process for its children. At the next step we correct the class imbalance among proteins (e.g. meats, tofu) and would again classify i. If no particular protein is significant, we use the children of all proteins at the next step. Assume those children are all leaf nodes. If the output is [0.40 pork, 0.35 chicken, . . .] then no label meets our criteria for significance, so the output would be protein, which *was* significant. Alternatively, if the values were [0.75 pork, 0.25 chicken], even though the parent of both (meat) does not meet our significance criteria, pork does, so the intake would be labeled as pork.

3.2. Classifying Intakes

The first step of our approach is to infer more granular labels for the weakly-labeled FL data. In the second step, we now use these intakes to train a classifier and apply it to the test data. The overall method, as shown in Figure 1 is similar to that of the labeling component. Again our approach proceeds top down in the hierarchy. Unlike in the training step, the actual labels are unknown, so we begin at the highest level of the hierarchy using all food classes (not a subset as in the previous step). Once again we determine the probability of each food and assess whether the classifier is sufficiently confident. When an inferred label meets our criteria we only explore its descendants as potential labels, while using all paths when there is uncertainty. As before, this process is repeated until a leaf node is reached.

Differences In summary, there are three key differences during the initial FL labeling step and the final testing steps: 1) the input data in the final stage contains both the lab and all labeled FL intakes, 2) we no longer restrict the potential classes considered, and 3) since meal-level labels are unknown at test time, we do not use the approximation used in training when there are zero intakes of a food. This means that at test time, if there is no training data on say, bacon, the most specific possible correct answer would be pork (if there were training data on other types of pork).

4. Data

We aim to evaluate the algorithm developed, and test our hypothesis that training on a mix of lab and FL data will lead to accurate food type inferences in FL. We use the multimodality ACE lab (Merck et al., 2016) and free-living (Mirtchouk et al., 2017) datasets, which have previously been used to identify eating periods (in lab and FL). The datasets are publicly available at: http://www.healthailab.org/data.html. The lab data was also used to identify food type and amount consumed (Mirtchouk et al., 2016), but it has not been known whether such classification will succeed on the more difficult problem of food type recognition in daily life. Collection of the datasets was approved by the Stevens IRB.

4.1. Sensors and devices

Both the lab and free-living studies aimed to balance rigorous ground truth with realistic behavior. The same body-worn sensors were used in all experiments.

Audio Audio signals were recorded from a customized earbud with an internal and external microphone, to allow for noise cancellation (Merck et al., 2016). The internal microphone captures chewing and other eating-related noises, and the external microphone captures environmental sounds that can then be subtracted from the internal mic. The microphones recorded at 44.1 kHz, and the data was later downsampled to 16 kHz.

Motion Participants wore an LG G Watch on each wrist. The watches recorded at a frequency of 15Hz using a 9-axis motion sensor. The sampling rate was intended to allow a full day of data collection without a need for recharging in between, though the watches did not always last a full day. Head motion was captured with Google Glass's 9-axis IMU, but as the data is not available for all FL participants, we omit it.

Video camera In the lab, ground truth at the level of bites was obtained by instrumenting the lab space with 3 video cameras positioned above, in front of, and perpendicular to the study participant. The cameras recorded at 30fps, and were synchronized to the sensor data, allowing precise timing of each food or drink intake.

Phone In free-living environments, participants were provided with an Android phone for photographing meals and a scale for recording meal quantity. At the beginning and end of each meal or drink they photographed the items or leftovers on top of the scale.

4.2. Data collection

Laboratory The dataset includes 6 subjects (2 female) aged 18-35 who participated in two 6-hour days of data collection, for a total of 12 days (72h) of data. The study period included at least two meals for most participants, and they had free choice of what to consume and when. Participants consumed all meals at the instrumented table.

Free-living In the FL dataset introduced by Mirtchouk et al. (2017), data was collected for 5 of the original lab participants, plus a new cohort of 6 individuals to test generalizability. Participants were aged 18-63, with 4 female and 7 male. Each day of data collection was 12h, to capture all meals. For 10 participants, the data collection period was 2 days,

while the 11th participant provided 5 days of data. Five participants wore all sensors (earbud, Glass, both smartwatches), while the other six (including the 5-day participant) did not wear Glass. This resulted in 27 free-living days, with 256.7h of usable data (batteries from some sensors died before the end of a day, and usable data is when all sensors are present). Participants consumed the meals of their choice, and went about their normal daily activities in their normal environments. Meals were thus consumed at restaurants, at home with family, in parks, and on the go. Participants engaged in many other non-eating activities, such as driving, running, playing sports, and so on.

4.3. Ground truth annotation

The core difference between lab and FL is in obtaining accurate ground truth.

Laboratory Each session was annotated independently by two researchers. Annotations that matched closely were automatically merged, while those outside a tolerance range were discussed and resolved collaboratively with a third researcher. All eating-related events were annotated including: chew, swallow, discrete intake (e.g. bite of sandwich), continuous intake (e.g. sipping soup from bowl), delivery of food or drink to the mouth, preparation of food or drink, and mouthing (manipulation of food with the tongue).

To identify food type, again two researchers used the video data to determine what was consumed in each intake. Annotations for each intake are an ordered list, with the most prominent food consumed listed first. Thus [guacamole, chips] means that there was more guacamole than tortilla chip in that bite. Unlike work on classification with discrete samples (e.g. bite of apple, piece of potato chip), this leads to complex labels. An intake can be labeled "soup," even though some intakes of soup may have pieces of vegetables and others may be primarily liquid. Thus this classification task is more challenging, as each type is not necessarily homogeneous. We use the intake event times and the most prominent food (first in list) during the intake as a label.

Free living In FL data collection, participants kept paper logs, used the annotation button on the smartwatch, and verbally annotated meal start and end (using the earbud). There were also unstructured interviews, after participants returned the sensors, which clarified omissions or discrepancies. Meal times were obtained using the logs as a starting point, and then listening to the audio data to identify more specific start and end times. While some participants photographed each meal component in detail (weighing and photographing each element of a dish), this still only records that these items were consumed at some point during the meal – not which bites at which times. The annotations in this environment are meal times and the items consumed in the meal. Amount consumed was generally only available for the meal (e.g. 70g of burger and fries) rather than its components.

4.4. Data characteristics

The lab data contain 7.5h of eating, with 1483 food and 285 drink intakes, while the FL data contain 19.9h of eating. Both datasets include a range of unique foods. In lab, after excluding foods with fewer than 4 intakes (6 intakes excluded) and combining all beverages into a single "drink" class, there were 40 unique food types consumed in 30 meals. In the FL data there were 81 meals and 71 unique foods. Our evaluation used 68 meals with

Figure 2: Examples of meals consumed by participants during free-living data collection.

65 unique foods (described further in the following section). We do not make a formal distinction between meals and snacks, and refer to all eating episodes as meals. Combining the two datasets, we have a total of 27.4h of eating and 111 meals.

A sample of foods consumed in FL are shown in Figure 2. Classes are imbalanced, with categories such as salad, soup, and rice being well represented (in 12, 11, and 8 meals respectively), while other foods like salami were consumed in a single meal (though pork in any form was in 5 meals). Out of the 40 and 65 unique foods for Lab and FL respectively, there are 28 that overlap at the lowest level of the hierarchy (including sandwich, soup, salad, yogurt, cookies, pizza, nuts, broccoli, and shrimp). However, we are able to identify foods even without an exact overlap due to our hierarchical approach. The hierarchy is shown in Appendix Figure 3. Defining such a structure is challenging when some foods involve multiple components that can also be consumed on their own (e.g. salad). We aim for output to be nutritionally informative at each level (e.g. distinguishing between protein and vegetable at the highest level), but future work may investigate other categorizations.

5. Experiments

5.1. Data processing

We used the same data processing pipeline as in prior work on this dataset (Mirtchouk et al., 2017), and briefly outline it below.

Audio processing We applied the noise cancellation approach of Merck et al. (2016), which uses the external mic to remove background noise from the internal one. We then extract standard signal processing features (energy, spectral flux, zero-crossing rate, and 11 MFCC coefficients) and temporal shape features (centroid, spread, skewness, kurtosis) at a time window of 200ms with a 20ms step size to capture chews.

Motion processing The motion data (accelerometer, magnetometer, and gyroscope from both watches, and the lab ablation experiment) was partitioned into 5 second windows with a 100ms step size to capture eating motions. We used the same 32 statistical features as in prior work on this data (Mirtchouk et al., 2017).

Extracting intakes We averaged the data for each feature for 15sec time windows. This duration was the average time between intakes in the lab data, and thus allows us to capture relevant detail before and after each intake. We then determine which windows are eating (rather than pauses during a meal to talk) by applying a random forest classifier trained on both the lab and FL data as in Mirtchouk et al. (2017). This led to 13 meals (with 6 unique foods) being excluded from classification, as no chews and intakes were inferred during them. We have a total of 68 FL meals (84% of the original 81) with 65 unique foods (92% of the original 71). The average FL meal length was 13.9min, and there was a total of 15.8h of eating. Combined with the lab data, this yields 23.3h of eating and 98 meals.

5.2. Evaluations

Our evaluations address three areas 1) importance of personalized data, 2) relative value of sensors, and 3) comparing components of our algorithm.

LOMO We first evaluate accuracy with leave one meal out (LOMO), iteratively training on 67 FL meals plus the lab data (97 meals total) and testing on one FL meal. We average results across all meals. For the upsampling to correct class imbalance we set $N = 250$. We follow the procedure described in section 3.1 to label the 67 FL meals at the intake level. Intakes can be labeled at different granularities depending on confidence so a meal that is known to contain salad and steak may have a list of intakes labeled: salad, vegetable, steak, beef, meat, protein. Only 5 intakes out of 3784 were not labeled at all.

LOHMO While LOMO ensures separation between training and test data, some foods may only be present in a single meal. We now leave one half meal out (LOHMO), as in Rahman et al. (2015). The process is the same as LOMO, but now training data includes either the first or second half of the test meal. We divide meals evenly by intakes rather than duration, as intakes can become sparse toward the end of a meal, leading to an imbalance if split based on duration. Results are averaged across all half meals.

Ablation study We now isolate each component of our approach to investigate its utility. The comparisons are 1) train lab test FL using RF, 2) train all (lab + FL) and test FL using RF, and 3) train lab test FL with our semi-supervised hierarchical classifier. These show what portion of accuracy comes from including FL training data (comparing 2 to 1, and LOMO to 3) and what portion comes from our classifier (comparing 3 to 1, and LOMO to 2). We also apply our approach to only lab data, as ground truth is known for each intake.

Table 1: LOMO and LOHMO results. % Total is what percent of intakes are classified at that level, and accuracy is percentage of those intakes correctly classified.

	LOMO		LOHMO	
	% Total	Level Accuracy	% Total	Level Accuracy
Level 1	35.8	86.4	36.0	97.5
Level 2	30.2	85.5	31.9	89.0
Level 3	28.7	77.8	27.2	73.1
Level 4	5.4	74.9	4.9	88.9
Overall (intakes)		83.1		87.7
Meal-level (unweighted)		76.5		91.2
Meal-level (weighted)		80.1		92.3

We conduct the same leave-one-intake-out (LOIO) analysis as Mirtchouk et al. (2016) did on this data with 1) training only on lab and 2) training on lab and FL data.

Sensor Comparison We re-ran LOMO using individual sensors (audio–A, right watch–R, left watch–L), and a combination of motion sensors (RL). This shows how a multimodality approach compares to single modality sensing for this task.

5.3. Accuracy

Intake accuracy is evaluated by level of the hierarchy. For example in a meal of "pork, salad, rice" the label at level 1 is "protein, vegetable, starchy foods." If any of these nodes are chosen, the intake is considered correct, just at a higher level than the original intake. The evaluation becomes stricter further down, as a correct classification at level 2 requires identifying the intake as meat, salad, or rice. Drinks are combined into a single class.

Meal accuracy evaluates what percentage of foods in the meal are accounted for. If the meal "pork, salad, rice" had intakes that were classified as: pork, protein, starchy foods, and fruit, then the meal accuracy would be 67% as pork and rice are covered by pork and protein, and starchy foods, respectively, but salad was missed. We compute both weighted and unweighted accuracy. In the unweighted version each meal contributes equally to the results and the other evaluation weights each meal by the number of intakes. Thus a long meal with many intakes will contribute more to the overall evaluation than a brief snack.

6. Results

LOMO and LOHMO As shown in table 1, our main accuracy results are 83.1% for LOMO and 87.7% for LOHMO. While it is not possible to compare directly with prior work, as this is the first study to infer food-type from body-worn sensors in FL, we note that this is on par with the 82.7% achieved in prior work on the lab dataset used here (Mirtchouk et al., 2016). That work used LOIO (leave one intake out) evaluation due to the smaller number of meals, so the classifier used data where the same food was consumed by the same individual in the same meal, which is closer to our LOHMO evaluation (leave half the meal out). In general, personalized data is known to improve activity recognition

performance, and this is what we observe comparing LOMO and LOHMO. Accuracy at both the highest and lowest levels of the hierarchy increased substantially.

Table 1 shows the percentage of intakes classified using labels at a given level (% Total in the table), and then within that level, what percent of those intakes were correctly classified (Level accuracy). Most intakes were classified at the top two levels. Level 1 (35.8% of classifications) includes categories such as protein and fruit. The tree is not balanced, and for example all vegetables are leaves at level 2. In contrast, for beef a correct level 4 classification would have to distinguish between steak and meatloaf. While there were few level 4 classifications, accuracy for these is still high. Both the weighted and unweighted meal-level accuracy suggest that the majority of foods in each meal are inferred. Comparing LOMO and LOHMO, we can see that with data from the specific food consumed, we now capture nearly all meal components regardless of meal duration (91.2% unweighted, 92.3% weighted), with 13 more distinct foods now identified. This improvement from half a meal suggests that after a cold-start, a live system can be rapidly improved as it adapts to a user's usual foods. Foods that were missed in LOMO are primarily salad (6 meals), and cookie (3, second highest). Since salad is a heterogeneous class, each instance is different enough that despite the large number of meals it was in, it may still be misclassified. As shown in figure 2, some salads do not have lettuce, making them quite different from those that do. Other missed foods are mainly those that appear in a single FL meal and for which there is no lab training data, and thus no training data for that food type in LOMO.

Two advantages of our approach are leveraging small amounts of lab data, and bootstrapping from foods solely in FL. First, broccoli was consumed as part of a stir fry in lab, and was leveraged to label intakes of an FL meal with pasta, carrots, and broccoli. Using these and the intakes generated to correct class imbalance, we correctly found 2 vegetable and 2 broccoli intakes during a different FL meal of pizza and broccoli. Second, we can learn foods that are observed solely in FL data. In the lab data, no one ate pork, while in the FL data there were five meals with pork, salami, and bacon. For LOMO, we identified pork or its ancestor in 3 of 5 meals, while for LOHMO it was identified in all 5 meals.

Ablation To understand how each component of our approach influences accuracy, we systematically varied which parts were used with LOMO evaluation. We find that for the overall (intake) evaluation accuracy is 61.7% for train lab test FL using RF and 68.4% with the addition of FL training data, showing that FL data is beneficial. However, our semi-supervised hierarchical classifier outperforms standard RF regardless of training data, with 77.3% accuracy when training on lab, and 83.1% training on all data (already reported in Table 1). Thus while the FL training data contributes to about 6% of accuracy regardless of classifier, our proposed classifier has around 15% more correct output than RF regardless of the training data set. Further, it was more accurate training only on lab (77.3%), than RF was training on all data (68.4%). As shown in detailed results in Appendix Table 4 our classifier output more specific foods than RF. Notably, not all FL participants were in the lab data, so this demonstrates that the approach can generalize well to new participants.

Finally, we use the same evaluation of Mirtchouk et al. (2016) on lab data (audio, motion from both wrists, head motion from Glass). Since lab was annotated from video as described, we have intake by intake ground truth. For the LOIO evaluation where Mirtchouk et al. had 82.7% accuracy using RF, our approach has 84.4% accuracy training on lab only, and

Table 2: Comparison of sensor accuracy (Acc) using LOMO: audio (A), right wrist motion (R), and left wrist motion (L). RL combines data from both watches.

	A		R		L		RL	
	% Total	Acc (%)	% Total	Acc (%)	% Total	Acc (%)	% Total	Acc (%)
Level 1	32.8	90.1	36.9	84.4	54.4	83.5	49.9	91.9
Level 2	28.0	74.1	36.8	71.7	16.9	66.5	24.0	62.0
Level 3	36.2	60.8	22.1	63.3	23.8	59.7	20.9	62.9
Level 4	3.0	72.4	4.1	70.1	4.9	66.6	5.2	87.8
Overall		74.5		74.4		74.1		78.4

88.0% when training on lab and FL. This shows that our method can leverage noisy data and use it to improve results beyond training and testing on a single clean dataset.

Sensor comparison Results for the sensor comparison are shown in Table 2. This is for the same LOMO evaluation for which ARL (audio, right and left watch) accuracy was 83.1%. There was a significant drop in accuracy when using a single sensor, with each sensor around 74% accuracy. The combination of watches (RL) was better, at 78.8%, but accuracy was still below the combination of audio and motion. Classifications at the top level had higher accuracy, but accuracy generally dropped significantly after that. This suggests that while audio may be able to distinguish between vegetables and protein, it cannot alone identify what foods within these categories were inferred. While the watches combined (RL) had higher accuracy than either alone, the combination also led to more classifications at level 1, so accuracy came at the expense of less specificity.

We investigated the relative contribution of features for ARL (all sensors), and found that audio (particularly energy and MFCC) was most useful at the top two levels of the hierarchy, while motion from the dominant hand (primarily the temporal shape features of the accelerometer) was most useful at the bottom two levels. This confirms our hypothesis that while audio is useful for distinguishing between high level food groups (e.g. protein, starch), motion aids in discriminating between lower level categories. Intuitively, the acoustic properties of vegetables and meats differ significantly, but audio cannot capture the difference between grilled chicken and chicken wings. On the other hand, the wrist motion during consumption of each differs significantly. By using a top-down hierarchical approach, our classifier can exploit the strengths of each modality at different stages.

7. Conclusion

Automated dietary monitoring has advanced considerably, with body-worn sensors being used to identifying meal times both in the lab and in the wild. However, foods consumed have only been automatically identified using body-worn sensors (rather than images) in the lab. To our knowledge, ours is the first study to demonstrate that body-worn acoustic and motion sensors can be used to identify the specific foods consumed in each bite in completely unrestricted environments with free-choice of meals. We introduce a semi-supervised hierarchical classifier that exploits shared structure between foods, and allows us to infer labels at varying levels of granularity. We achieve 83.1% accuracy for identifying

foods consumed in each intake using leave-one-meal-out evaluation, and 87.7% with leave-one-half-meal-out. We ultimately aim to link the inferred foods to nutrition databases to provide nutrient information, and to infer quantity consumed to create automated calorie counts and food logs. Future work is needed to determine how best to scale up data collection and to determine the optimal structure of foods for both classification and final use. Code is available at https://github.com/health-ai-lab/SSH

Acknowledgments

This work was supported in part by the NSF under award number 1347119, NIH under award number R01LM011826, the James S. McDonnell Foundation, and the LoPorto Graduate Fellowship.

References

Oliver Amft and Gerhard Tröster. On-body sensing solutions for automatic dietary monitoring. *IEEE Pervasive Computing*, 8(2):62–70, April 2009. doi: 10.1109/MPRV.2009.32.

Oliver Amft, Holger Junker, and Gerhard Tröster. Detection of eating and drinking arm gestures using inertial body-worn sensors. In *ISWC*, 2005a. doi: 10.1109/ISWC.2005.17.

Oliver Amft, Mathias Stäger, Paul Lukowicz, and Gerhard Tröster. Analysis of chewing sounds for dietary monitoring. In *UbiComp*, 2005b. doi: 10.1007/11551201_4.

Cigdem Beyan and Robert B. Fisher. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48:1653–1672, 2015.

Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Hierarchical classification: combining bayes with svm. In *ICML*, 2006.

Amanda Clare and Ross D King. Predicting gene function in saccharomyces cerevisiae. *Bioinformatics*, 19(suppl_2):ii42–ii49, 2003.

Juan M. Fontana, Janine A. Higgins, Stephanie C. Schuckers, France Bellisle, Zhaoxing Pan, Edward L. Melanson, Michael R. Neuman, and Edward Sazonov. Energy intake estimation from counts of chews and swallows. *Appetite*, 85:14–21, 2015.

Siddharth Gopal and Yiming Yang. Hierarchical bayesian inference and recursive regularization for large-scale classification. *TKDD*, 9:18:1–18:23, 2015.

Corby K. Martin, John B. Correa, Hongmei Han, et al. Validity of the Remote Food Photography Method (RFPM) for Estimating Energy and Nutrient Intake in Near Real-Time. *Obesity*, 20(4):891–899, 2012.

Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. Multimodality sensing for eating recognition. In *Pervasive Health*, 2016.

Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2calories: towards an automated mobile vision food diary. In *ICCV*, 2015.

Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In *UbiComp*, 2016. doi: 10.1145/2971648.2971677.

Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. Recognizing eating from body-worn sensors: Combining free-living and laboratory data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):85:1–85:20, September 2017. doi: 10.1145/3131894.

Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. Platemate: Crowdsourcing nutritional analysis from food photographs. In *UIST*, 2011.

Novi Quadrianto, Alexander J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

Shah Atiqur Rahman, Christopher Merck, Yuxiao Huang, and Samantha Kleinberg. Unintrusive Eating Recognition using Google Glass. In *Pervasive Health*, 2015.

Tauhidur Rahman, Alexander T. Adams, Mi Zhang, Erin Cherry, Bobby Zhou, Huaishu Peng, and Tanzeem Choudhury. Bodybeat: A mobile system for sensing non-speech body sounds. In *MobiSys*, 2014. doi: 10.1145/2594368.2594386.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.

Giovanni Schiboni and Oliver Amft. Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors. In *ISWC*, 2018. doi: 10.1145/3267242.3267253.

Lukas Schwingshackl, Georg Hoffmann, Anna-Maria Lampousi, et al. Food groups and risk of type 2 diabetes mellitus: a systematic review and meta-analysis of prospective studies. *European Journal of Epidemiology*, 32(5):363–375, May 2017.

Jenna L. Scisco, Eric R. Muth, and Adam W. Hoover. 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):464–469, 2014. doi: 10.1016/j.jand.2013.09.017.

H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Alpana P. Shukla, Radu G. Iliescu, Catherine E. Thomas, and Louis J. Aronne. Food order has a significant impact on postprandial glucose and insulin levels. *Diabetes Care*, 38(7): e98–e99, 2015. doi: 10.2337/dc15-0429.

Amy F Subar, Laurence S Freedman, Janet A Tooze, et al. Addressing current criticism regarding the value of self-report dietary data, 2. *The Journal of Nutrition*, 145(12): 2639–2645, 2015.

Aixin Sun and Ee-Peng Lim. Hierarchical text classification and evaluation. In *ICDM*, 2001.

Edison Thomaz, Irfan Essa, and Gregory D. Abowd. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *UbiComp*, 2015.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

Feihong Wu, Jun Zhang, and Vasant Honavar. Learning classifiers using hierarchically structured class taxonomies. In *SARA*, 2005.

Koji Yatani and Khai N. Truong. Bodyscope: A wearable acoustic sensor for activity recognition. In *UbiComp*, 2012.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, 2002.

Appendix A. Algorithms

Algorithm 1 Generating Intakes

Input:

$X = [x_1, x_2 \ldots, x_m]$, an array of vectors of features where $x_i = [f_1, f_2 \ldots, f_{amtFeat}]$, a vector of features

$F = [f_1, f_2 \ldots, f_m]$, an array of food categories associated with a corresponding instance of X

N, the minimum amount of intakes for each category

R, the number of iterations before regenerating a new Random Forest

Output:

X_{new}: X concatenated with newly generated intake features

F_{new}: F concatenated with newly generated intake categories

The length of each distinct type of F_{new} is greater than or equal to N

$count = 0$
while $\exists f(f \in F, \|f\| < N)$ **do**
  $d \in F < N$
  $x_d = x \in X \mid F = d$
  for $i = 1$ to $amtFeat$ **do**
    $I_{new_i} = \mathcal{N}(\overline{x_{d_i}},\ \sigma(x_{d_i}))$
  if $count \mod R = 0$ **then**
    $F_{train} = [f : f \in F, \|f\| < N]$
    $X_{train} = x \in X \mid F \in F_{train}$
    rf = Random Forest (X_{train}, F_{train})
  else
    Use previous rf
  if rf.predict(I_{new}) $= d$ **then**
    X.append(I_{new})
    F.append(d)
    $count+ = 1$
return X,F

Appendix B. Results

Table 3: Accuracy for tests on lab data. Leave one intake out (LOIO) using our classification approach training on lab only or both lab and FL data.

	LOIO (lab)		LOIO (lab+FL)	
	% Total	Level Accuracy	% Total	Level Accuracy
Level 1	20.1	89.8	27.9	91.9
Level 2	54.3	87.1	61.5	89.0
Level 3	22.0	75.0	8.8	74.0
Level 4	3.6	70.2	1.8	62.5
Overall (intakes)		84.4		88.0
Meal-level (unweighted)		80.2		82.9
Meal-level (weighted)		83.5		87.4

Table 4: Detailed accuracy for ablation experiments using LOMO evaluation. All tests are on FL data. Last column repeats LOMO results from Table 1 to facilitate comparison. RF is random forest, SSH is our semi-supervised hierarchical classifier.

	Train lab, RF		Train lab+FL, RF		Train lab, SSH		Train lab+FL, SSH	
	% Total	Acc	% Total	Acc	% Total	Acc	% Total	Acc
Level 1	40.8	68.2	45.4	70.3	29.9	82.2	35.8	**86.4**
Level 2	36.7	61.8	30.5	71.2	35.1	80.1	30.2	**85.5**
Level 3	15.5	50.3	17.5	60.9	26.2	69.9	28.7	**77.8**
Level 4	7.0	48.9	6.7	61.1	8.9	70.4	5.4	**74.9**
Overall (intakes)		61.7		68.4		77.3		**83.1**
Meal-level (unweighted)		59.3		68.9		71.9		**76.5**
Meal-level (weighted)		59.9		68.2		76.7		**80.1**

Appendix C. Food Ontology

- protein
  - bean
  - dairy
    - cheese
    - yogurt
  - egg
  - meat
    - beef
      - meatloaf
      - steak
    - chicken
      - grilled chicken
      - wing
    - fish
      - mussel
      - salmon
      - shrimp
      - sushi
    - pork
      - bacon
      - salami
  - nuts
    - almonds
    - nut butter
    - peanuts
  - rice bowl
  - tofu
- starchy foods
  - bread
  - dumpling
  - oatmeal
  - pasta
    - noodle
  - pizza
  - potato
    - fries
    - hash
    - potato salad
  - rice
  - sandwich
    - burrito
    - hamburger
    - tacos
- vegetable
  - avocado
  - broccoli
  - carrot
  - celery
  - cooked vegetables
  - corn
  - eggplant
  - green bean
  - lettuce
  - mushroom
  - pea
  - peppers
  - salad
  - spinach
  - tomato
- fruit
  - apple
  - berry
  - cherry
  - citrus
  - grape
  - melon
  - musaceae
  - pear
- dessert
  - baked goods
  - candy
  - chocolate
  - cookie
  - ice cream
  - whipped cream
- snack
  - chips
  - cracker
  - energy bar
  - popcorn
  - pretzel
  - trail mix
- liquid
  - drink
  - sauce
  - soup
    - ramen

Figure 3: Food type ontology.