Automated Estimation of Food Type and Amount Consumed from Body-worn Audio and Motion Sensors

Mark Mirtchouk **Christopher Merck** **Samantha Kleinberg**

Stevens Institute of Technology
Hoboken, USA
{mmirtcho, cmerck, samantha.kleinberg}@stevens.edu

ABSTRACT

Determining when an individual is eating can be useful for tracking behavior and identifying patterns, but to create nutrition logs automatically or provide real-time feedback to people with chronic disease, we need to identify both what they are consuming and in what quantity. However, food type and amount have mainly been estimated using image data (requiring user involvement) or acoustic sensors (tested with a restricted set of foods rather than representative meals). As a result, there is not yet a highly accurate automated nutrition monitoring method that can be used with a variety of foods. We propose that multi-modal sensing (in-ear audio plus head and wrist motion) can be used to more accurately classify food type, as audio and motion features provide complementary information. Further, we propose that knowing food type is critical for estimating amount consumed in combination with sensor data. To test this we use data from people wearing audio and motion sensors, with ground truth annotated from video and continuous scale data. With data from 40 unique foods we achieve a classification accuracy of 82.7% with a combination of sensors (versus 67.8% for audio alone and 76.2% for head and wrist motion). Weight estimation error was reduced from a baseline of 127.3% to 35.4% absolute relative error. Ultimately, our estimates of food type and amount can be linked to food databases to provide automated calorie estimates from continuously-collected data.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI)

Author Keywords

Nutrition; Eating recognition; Acoustic and motion sensing

INTRODUCTION

While eating recognition has been an active area of research (leading to a number of solutions using acoustic sensors, continuous image capture, and motion sensing), we need to know not only *that* someone is eating, but *what* they are eating and *how much* they consume to develop truly automated dietary monitoring. A system that could determine that an individual consumed 3oz of steak, 5oz of potato, and 2oz of salad could use this information to track nutrition (identifying deficiencies and unhealthy behaviors), provide feedback to individuals with chronic diseases such as diabetes to improve self-management (suggesting insulin doses), and determine adherence to dietary guidelines. Other applications include providing personalized food suggestions based on dietary goals and prior consumption, for example by pushing menu suggestions to a user's smartphone when location data says they have entered a restaurant.

Body sounds and movements have been used to recognize eating episodes, but have been under-explored for estimating food type and quantity. Instead, most approaches to tracking nutrition in such detail require extensive user involvement (e.g. manual input with paper logs and mobile apps), or have low accuracy. Automated solutions mainly use only audio, classifying a small set of simple foods; or only motion sensors, tracking solely the number of bites from wrist motion. However, acoustic sensing fails for soft foods, and counting bites does not provide the nutrition information needed to fully replace user-generated food logs.

We propose that audio and motion data can be combined to accurately estimate food type and amount for each intake. We hypothesize that while acoustic sensors can distinguish between different food textures (crisp vs. tacky), motion sensors can help discriminate between soft foods such as ice cream and a milkshake based on head or wrist position. Similarly, features such as the number of chews after a bite are likely related to the size of the bite in a food-dependent way due to physical properties, but finding individual chews requires acoustic data and we need motion to accurately estimate type.

We use data collected from 6 individuals wearing motion (head, both wrists) and acoustic sensors (customized earbud), with ground truth from detailed video annotation (e.g. individual chews, swallows). Food choice was unconstrained, and participants ate meals such as tacos, sushi, and steak. Even with 40 food classes we achieve high classification accuracy (82.7%). Further, when combining knowledge of food type with features extracted from sensor data, amount consumed can be estimated with error similar to that of human annotators in prior work: 35.4% error for solid food, 47.2% for drinks. In contrast, without type information error is high (127.3% for solid foods, 60.4% for drinks). While we use multiple sensors, these modalities may ultimately be combined into a single wearable (e.g. motion/audio sensing earbud).

Our primary contributions are: 1) rigorous comparison of sensing modalities in a single setting with realistic meals,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
UbiComp '16, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4461-6/16/09.
http://dx.doi.org/10.1145/2971648.2971677

showing that motion and audio data can be combined to accurately estimate intake content and amount; and 2) a unique publicly available dataset with ground truth of food type and weight for each bite. Data can be obtained at http://www.skleinberg.org/data.html.

RELATED WORK

Eating recognition has been a significant topic of research, with many methods developed using acoustic and motion sensing, image processing, and environmental sensors, but these modalities have mainly been investigated separately.

Acoustic sensing

Audio sensors, such as earbud microphones, have been highly accurate for recovering activities such as chewing [3], and intuitively such sounds should help to discriminate between foods such as carrots and tortilla chips. Thus, methods for automatically estimating the type of food consumed have generally focused on acoustic sensing, though have mainly classified intakes from among a small group of foods. Amft and Tröster [4] used a set of 19 foods (including crisp and soft), with classification accuracy of 80% with an earpad sensor (70-75% with an earbud form factor). However this work involved only 3 male participants and more critically evaluated discrete foods (e.g. potato, orange) rather than meals, which normally combine multiple food types. Later work developed specialized wearable microphones, though these were evaluated on fewer foods: Päßler et al. [22, 23] used 7 foods (potato chip, peanut, walnut, carrot, apple, chocolate, pudding) in 10 pieces, Bodybeat [25] used 4 (cookie, apple, bread, banana), and Bodyscope [32] 2 (cookie, bread). Rather than discrete sequential bites (e.g. piece of apple, then piece of cookie) our data is from full meals with complex foods such as salad or a stir fry where every bite may be somewhat different.

While audio sensors are mostly placed on the ear or around the throat, recent work has used wrist-mounted audio to identify eating behavior, such as using environmental sounds to identify meal periods [31]. Other work attempted to use such data to classify food type, but tested only apple and potato chips and participants were told how many bites to take [17]. Sensors further away from the ear may also face more challenges with background noise.

To automate nutrition logging we need to estimate amount consumed in addition to food type, though this problem has been less studied than classification. One seminal work extracted features from audio data to estimate intake size for potato chips, lettuce and apple [2] with 8 participants, with error ranging from 19-31% depending on food type. Despite the relationship between food sounds and quantity consumed, we are not aware of more recent advances in finding bite weight.

Motion sensing

Motion sensing raises fewer privacy concerns than continuous audio collection, though it is generally less accurate than audio-sensing. Wrist motion, recorded with smart watches, has been used to detect a range of eating behaviors, such as when a person is eating [28, 12] and what utensils are used (e.g. chopsticks, spoon) [1, 27]. Thus far, wrist motion has not been used to detect food type. Other work used head motion captured with Google Glass to identify meal periods [24], though this data is not specific enough to identify individual chews (as is needed to characterize eating speed and time between chews to identify food quantity).

In addition to accelerometers, motion has also been captured with capacitive sensors, primarily placed around the neck with collar-like devices [9]. These have been primarily used to identify eating behavior such as individual swallows, but have also been used to estimate amount of fluid intake [8], with AUC of .73 to .76 for classifying drinks as 5 or 15ml. Other work used proximity sensors to measure ear deformation [5], though this did not aim to classify food type or quantity and had lower accuracy in the wild than in lab settings (where foods were restricted to M&Ms, apples, and bananas).

Wrist-based motion sensors have been used to indirectly estimate calories consumed, by counting bites [11, 26] – assuming a fixed number of calories per bite. However, bite weight and calorie content vary in a food type and user-dependent manner, and may change over the course of a meal. While food-type and user-specific features can be estimated, they were not used in that work. In our data, intake size varied greatly, with food intakes from 0g to 43g (mean 6g, s.d. 20g) and drink intakes from 2.5g to 168g (mean 31g, s.d. 25.5g). Prior work has also tended to focus on either solids or liquids, while we aim to estimate the quantity of both.

Image-based methods

Another key approach to objective nutrition assessment is based on image analysis. Platemate used crowdsourcing to label images, and found the accuracy of this approach was comparable to that of experts (mean absolute error of 198 calories, 33.2%) [21]. However, this approach requires a user to remember to take a photo and requires the labor of crowd-workers to label the images. The mean time to complete the task was 94.14 minutes. While this is not an impediment for daily food logs, it does not allow real-time interventions, such as adjusting dosing of insulin for a person with diabetes or helping someone eat more mindfully. Menu-match, Im2Calories, and others provide a more automated solution by combining image analysis with GPS to map photos taken by users to restaurant menu items [6, 20, 7]. However, these require users to take action (taking photo and using app), rather than enabling the fully automated sensing needed to identify events such as mindless snacking. On the other hand, continuously collected photos from first-person point-of-view (POV) cameras have been used to identify eating episodes [29, 30], but have not been used for nutritional analysis and the angle and photo quality make this challenging.

Environmental sensors

The related work described thus far has relied on sensors worn or actively used by an individual (in the case of a smartphone camera). While these approaches are the most mobile, they require users to wear a device such as a microphone, accelerometer, or POV camera. An alternate approach is to instrument the environment where people eat, such as with

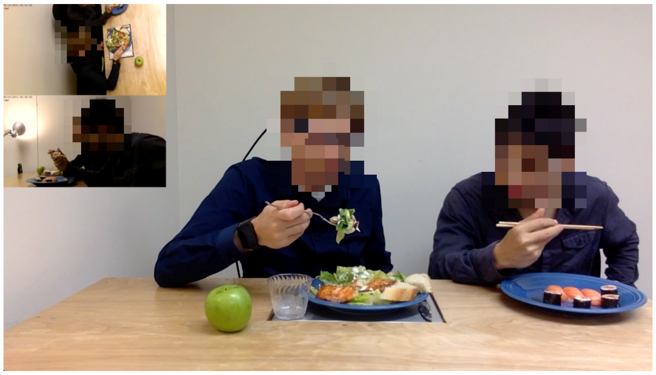

Figure 1: Screenshot from the 3-camera setup (top and side views are shown in upper left corner) during a shared meal, as the participant on the left lifts food to their mouth.

smart utensils, surfaces, or kitchens outfitted to track calories while cooking [10]. Sensor-embedded forks can be used to count bites, and one such fork also used color sensors in the tip to classify among 17 food types with an F-measure of 87.5% [16]. While this approach could be expanded to other utensils (spoon, knife, chopsticks) to allow more eating methods, it cannot identify hand-held food such as sandwiches. Sensor-augmented cups have also been used to classify liquids based on their pH [18]. Pressure-sensitive surfaces can be used to identify eating behaviors such as cutting food or stirring it. For example, work by Zhou et al. [33, 34] classified between four food categories with high accuracy, though the sensor was less precise at measuring weight. However both the surface and cup would require users to bring the devices with them for meals outside the home.

Existing work has mainly studied a single body-worn sensor at a time for a single task. Image-based and environmental sensors have been used for more complete nutrition-monitoring solutions, but these are less mobile and either require human labor (in the case of human annotated and captured images) or create privacy concerns (for POV cameras). In contrast we show that food type is critical for estimating amount consumed, and that estimating type accurately with a wide range of foods requires multiple sensing modalities (audio, motion).

STUDY DESIGN AND SENSORS

We aim to investigate the use of multiple sensing modalities for recognizing food type and amount for each intake. One challenge for translation to real-world use is that controlled lab studies often use few foods eaten separately (bite of apple, piece of chocolate), while in-the-wild studies lack the ground truth needed for rigorous evaluation (relying primarily on user self reports of eating activity or the sensor signals themselves to provide ground truth). We further aim to estimate food type and amount consumed on a per intake basis, rather than per meal, as this is ultimately needed to provide real-time guidance such as feedback on speed or content of consumption, or automated adjustment of insulin pump doses in diabetes.

We previously developed the ACE (accelerometer and audio-based calorie estimation) dataset [19], which combined multimodality sensing with detailed annotation (at the level of chews and swallows) from video data. Here we describe the study design and sensors, and discuss the previously unreported food weight and food type measurements and annotations developed in this paper.

Body-worn sensors

To capture signals that may indicate eating activities, participants wore four sensors. While this is somewhat unrealistic for real-life use, it uniquely lets us compare each sensing modality in the exact same situation (same food, user, and setting), in contrast to prior work that has mainly evaluated each separately. Future work may combine multiple sensors into a single housing (e.g. audio and motion sensing earbud).

Acoustic sensor

We customized a standard earbud with internal and external microphones and recorded at 44.1Khz with a pocket audio recorder. The external microphone enables us to remove most speech (external and from the user) and non-eating sounds, as these are captured on both microphones while the subtle eating signal is captured mainly on the internal microphone.

Wrist motion

An LG G watch was worn on each wrist, and recorded from its 9-axis motion sensor at 15Hz. This sampling rate was chosen to balance recording frequency and battery life and ensured we could record a full day of data. Deploying a smartwatch on each wrist enables comparison of accuracy when instrumenting the dominant versus non-dominant hand, and the value of having accelerometer data for both wrists, which may be able to help discriminate between food type (e.g. holding sandwich versus a piece of fruit).

Head motion

Previous work has shown that Google Glass's 9-axis motion sensor can detect eating-specific motion [24]. We recorded from Glass using the same 15Hz sampling rate as for the watches. Glass, which has a form factor similar to glasses without lenses, was connected to a small external battery pack to ensure sufficient battery life. The battery and audio recorder were worn in a small running belt to allow free motion.

Each sensor records at a different rate, so we developed a synchronization procedure. A controlled tap of all body-worn devices on the scale at the beginning and end of recording provided a spike in motion and audio data and was visible on the video, enabling synchronization with high accuracy.

Ground truth sensors

Video

To be able to identify the composition and timing of each bite, we instrumented the lab space with 3 IP-based, 1280x720 pixel resolution, H.264 encoding video cameras. These were positioned on the wall (top view) and clamped to the table where eating occurred (front, side view). The cameras provided clear views of each participant's mouth and throat (to identify chewing and swallowing) and the scale, as shown in figure 1. We recorded at 30fps and previously regularized the recordings to ensure exactly this recording rate, as the actual mean frame rate was 30.08fps (min 8, max 32fps).

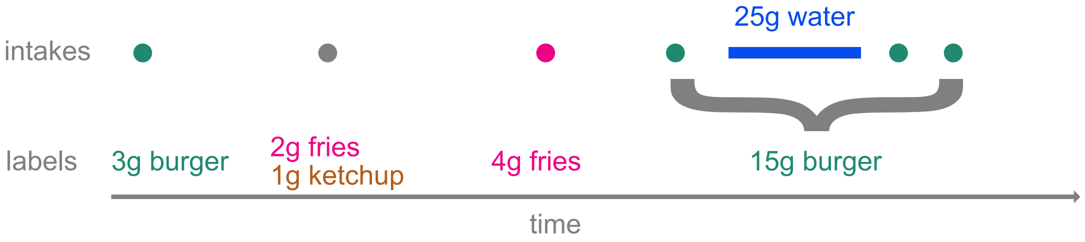

Figure 2: Representative example of intake annotations. Circles indicate food intakes, and the bar shows duration of a fluid intake. When the burger is not placed on the scale between bites, a weight is assigned to the group of 3 intakes.

Food weight

To record weight of each intake (rather than total consumed over an entire meal), we embedded a balance in the table used for eating (visible in figure 3 with food on top). The weigh scale (Sartorius 12.5x9.5 inch Midrics weighing platform with Combics 2 IP44 indicator) recorded continuously at 10Hz with 0.5g resolution. The large platform and wooden surround we built to be nearly flush with top of the weighing platform ensured that participants could eat naturally and would not accidentally rest their arms on the scale.

Data collection

With approval from the university IRB, data was collected from 6 participants (4 male, 2 female), in two ~6hr sessions for each person for a total of ~72hrs of data. Participants did not receive incentives for participation. The data collected allow us to obtain ground truth about what food type was consumed and how much was eaten in each bite, while ensuring behavior was as realistic as possible. To that end, participants ate the foods of their choosing in the quantities of their choosing (rather than selecting from a limited set of foods). Each participant had at least two meals of their choice and some also chose to consume snacks. Note that we do not make a formal distinction between meals and snacks, as some people simply ate small meals. Roughly, snacks here are eating episodes outside of usual meal times that were not in place of a meal. Outside of eating sessions participants were free to move about, work at a computer, nap or do the activity of their choice (one did push ups, others played games on their phones). Participants remained in the lab space to ensure activities were recorded on video. However behavior during meals was completely unrestricted, so participants could multitask (e.g. eating and working on a laptop or using a mobile phone), share meals with research group members, and talk while eating. All meals and snacks were eaten at the scale-instrumented table. Outside of meal periods, drink containers were weighed on the scale after each drink to determine intake weight.

Annotation and ground truth

Eating activities

Annotation of the video data does not provide true ground truth, as some events such as swallows may be missed, but it is the best approximation and ensures no bias (as when using one of the sensors to find the activities) and unrestricted movement (in comparison to having an individual or researcher mark events in real-time). The video was annotated independently by two researchers. Most annotations matched (and could be merged automatically), but all were discussed and resolved in a collaborative process involving a third researcher. Rather than coarse activities such as meals, we focused on fine-grained annotation of eating as information about individual chews (including their frequency) can provide insight into the amount and type of food consumed. The annotations included the following italicized activity codes, where brackets indicate events with a duration, and asterisks those that can be performed with one or both hands:

- *preparation*[]* of food, such as stirring, cutting, or scooping
- *delivery*[]* of food or drink to the mouth
- fluid *intake*[], solid *intake*, marked when food or drink first passes the lips. Fluid intake means continuous intake from a vessel or straw, and a solid intake is a discrete quantity. A food such as soup can be consumed multiple ways in a single meal.
- *chew*, annotated when jaw first closes
- *mouthing*[] or manipulating food with the tongue
- *swallow*, annotated when visible in the throat
- touching the face with a *napkin*[]*

Other activities such as walking or typing were not annotated, but are guaranteed to be negative examples (not eating or drinking). Annotation of each session using vCode [14] took each researcher around 8 hours and merging 2-3 hours.

Food type and amount

In this work we augment the eating activity annotations with type and quantity annotations for each intake of food or drink. Annotations were once again done by two researchers, who then discussed and combined their annotations. Due to the smaller number of events (intakes rather than chews), there were few disagreements and little ambiguity, so a third person was not needed to help merge. For example, a difference in annotation may be due to one annotator forgetting to subtract the weight for a utensil, which is clear upon video review. However, as a precaution we recorded video and audio of the annotation merging process in case the rationale for an annotation needed to be examined later.

We developed a tool to visualize annotation sequences (described in the previous section) along with the video data, and

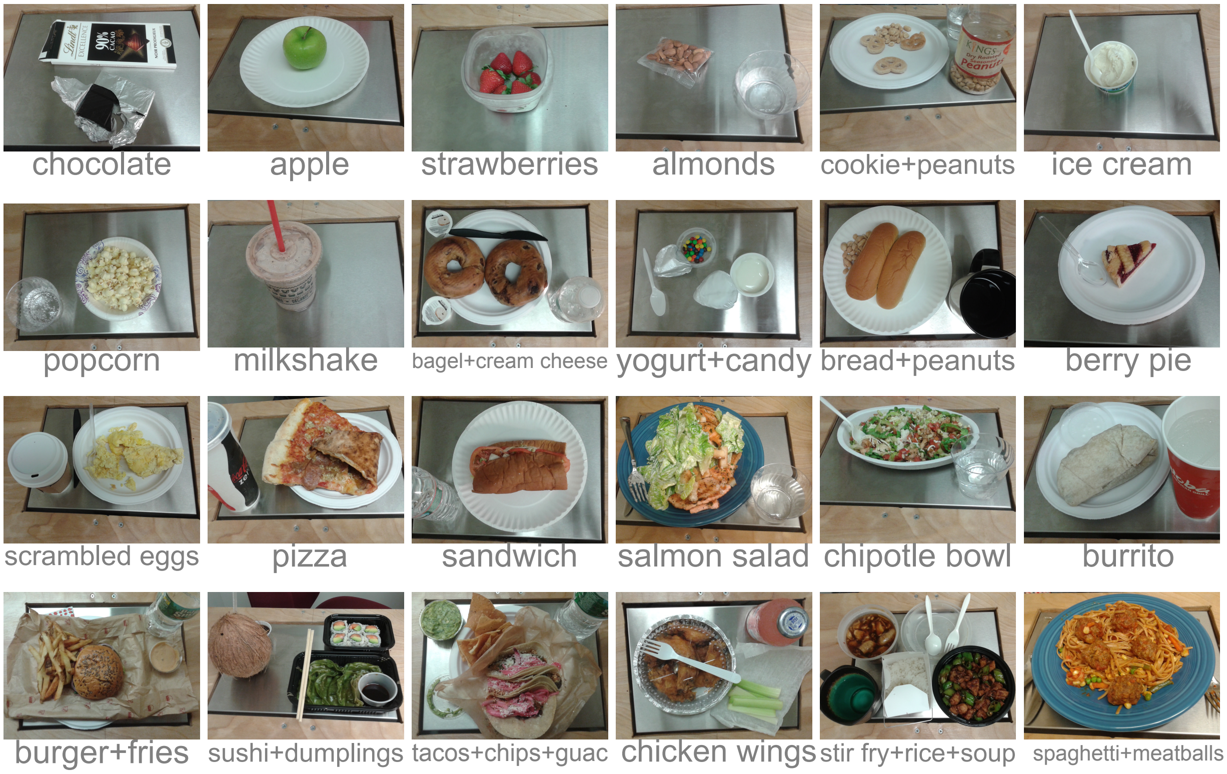

Figure 3: Sample of meals consumed by participants. Top two rows are mainly snack and breakfast foods, and bottom two rows primarily depict lunch and dinner. Many meals involved combinations of foods.

raw scale data. Annotators first labeled the set of foods in each meal, then assigned one or more food types to each intake.

We treat complex foods (e.g., burger) as a single food type rather than attempt decomposition into components. Thus if a participant removes a pickle from a burger and eats that by itself, that intake is still annotated with the burger type. For foods that are dipped or otherwise combined by participants in simple ways, such as chips and guacamole, intakes could be labeled with each food individually (e.g. if participant clearly licks guacamole off the chip, and does not consume the chip in an intake). The set of foods associated with each intake is ordered from most to least prominent (e.g. [bagel, cream cheese] indicates more bagel than cream cheese in the bite).

Weight of food or drink consumed in each intake was derived from the scale time series, after accounting for utensils and unconsumed food (e.g. bite of a tortilla chip that is then replaced on scale). When food was not replaced on the scale in between intakes, such as when a participant took multiple bites before putting a sandwich down, a total weight is assigned to the group of intakes. While this does not yield true ground truth for the size of these intakes, we aimed to allow natural eating behavior, rather than force participants to eat in a predefined way, such as putting down a slice of pizza after every bite.

Figure 2 shows an example annotation sequence. One intake combines multiple foods (fries, ketchup), and later there is a fluid intake of water. Before and after the water the participant consumes a burger, but does not put it back down on the scale between bites. As a result, a weight is assigned to the group.

Data characteristics

The raw data included a total of 1489 food and 285 drink intakes, 17,080 chews, and 1422 swallows.[1] A total of 51 unique foods and drinks were consumed. We exclude 4 foods (pocky, brownie, cheese, sour cream) with fewer than 4 intakes (most with only a single intake), leading to a total of 6 intakes excluded from further analysis. We also combined all drinks (coconut water, hot and iced coffee, various sodas, Snapple, tea, water, and sparkling water) into a single drink class. Milkshake remained as its own food type as it contained chunks of chocolate and was sometimes consumed with a spoon. No drink intakes were excluded.

This leaves a total of 1768 total intakes (1483 food + 285 drink), and 40 unique food types. In all, over 450 minutes was spent eating across 30 meals (mean 2.5 per person, s.d. 0.52). Participants ate a mean of 8.5 different foods (s.d. 2.9) over their two sessions, and had a mean of 2.3 drinks (s.d. 1.2).

A sample of foods consumed are shown in figure 3. Only three foods (popcorn, water, and yogurt) were consumed by multiple participants. The yogurt also varied, with one being smooth and the other having candy pieces added. Foods spanned a range of textures and intake methods, including

[1] Note that these were the total annotated, and the true number of swallows may be greater than the amount visible on video.

sushi (chopsticks), tacos and pizza (handheld), and steak (cut with fork and knife). Further, many meals combined multiple foods or were composed of many parts, such as spaghetti with meatballs or tacos eaten with tortilla chips and guacamole. Participants drank from a variety of containers, including mugs and cups, bottles, and to-go cups with and without straws. Eating similarly used a variety of utensils and containers (bowls, ceramic and paper plates, food wrappers, take-out containers).

While this does not capture all possible foods, it is a more realistic sample than a controlled set of foods eaten in single bites. Further, an individual does not regularly consume hundreds of different foods. Diet diversity has been examined as an indicator for nutrition and health, and one study found a mean dietary variety score (number of unique foods consumed) of 64 after 15 days among adults, with this score increasing rapidly over the first few days of the study and then plateauing [13]. In a practical system, we propose that population data can provide a starting point, and over time, the system would adapt to a user's particular set of most likely food choices.

Mean intake weight and standard deviation varied considerably by food type, as shown in figure 4, which depicts foods with more than 10 intakes. Note that as mean weight increases, standard deviation does too (Pearson's r=0.60 across all foods with >10 intakes and usable weights). Thus using only average weight may lead to considerable bias, particularly for foods with large intake sizes. Mean intake size (weighted by number of intakes for the food) across all foods with more than four intakes was 6g (unweighted mean of 11g), with a standard deviation of 20g. The mean intake size for fluid intakes was 31g with a standard deviation of 25.5g. The data are highly imbalanced, with some foods such as a banana consumed in very few intakes while others such as popcorn were consumed across a large number of intakes (comprising 11% of all intakes). The number of samples for all foods used for the weight estimation task are shown in table 1. Prior work [11] that examined which hand was used for eating found 86% of intakes were done with the dominant hand, while we find that 69% of intakes of food or drink were delivered to the mouth with the dominant hand, 18% with the non-dominant hand, and 13% with both hands.

METHOD

We now discuss how these data were used for food-type classification and amount estimation, starting with the data preprocessing and then discussing the experimental evaluation.

Data processing and feature extraction

Noise cancellation

Before any further processing, we applied our previously developed [19] noise cancellation procedure. Essentially, this procedure removes the sounds from the internal microphone that are well-predicted by the external microphone signal. This removed nearly all participant and external speech and noise, leaving only a participant's eating noises.

Feature extraction

In this work our goal is to assign a type and weight to each of the 1768 food or drink intakes automatically. Rather than process the entire time series or whole meal periods, we hypothesize that the time shortly before an intake until the end of chewing is the most informative. This time contains interaction with the food or drink (e.g. picking it up), moving it to the mouth, and food-specific chewing noises (in contrast to right before swallowing, when food is softened and less likely to have a distinct sound).

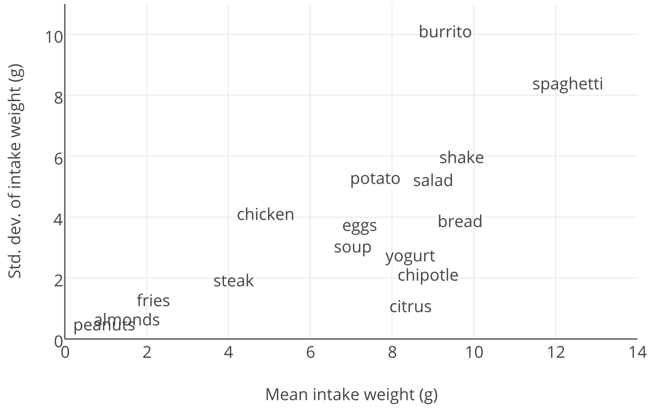

Figure 4: Intake sizes for foods with more than 10 intakes. Drink (30g mean) and broccoli (26g mean) are not pictured.

Thus we limit the data used for classification to the following time windows. For audio and solid intakes, the time window is from the intake until the last chew before the next intake. For drink intakes, the window is that of the continuous fluid intake. For motion data, both solid and liquid intakes are expanded to include the delivery of food or drink prior to the intake, to capture the movement of food or drink to the mouth.

For example, in figure 2 classification for the first bite of burger will use data from the bite until the last chew before the intake of fries (for audio features) and the same window plus the delivery of the burger to mouth (for motion features). For the intake of water in that figure, the time period shown in blue will be used (audio), and will be augmented with the delivery of the drink to the mouth (motion).

We extract three types of features: audio and motion features (based on the sensor data) and annotation features (which are higher order features).

Audio features After noise cancellation, we divide each intake window (as defined above) into 200ms-long frames with an offset of 20ms. This frame length is chosen so as to capture a full chew without including multiple chews in a single frame. For each frame we compute the following features: energy, spectral flux, zero-crossing rate, and 11 MFCC coefficients. Then, the mean and standard deviation of the frame features forms the feature vector for the whole intake window.

Motion features We similarly segment each intake into 5-second frames with a 100ms offset. The larger frame and offset for motion data (head and wrists) is intended to capture a complete movement such as of the wrist toward the mouth or the head toward a drink, without capturing multiple intakes in a single frame. We then compute the following features:

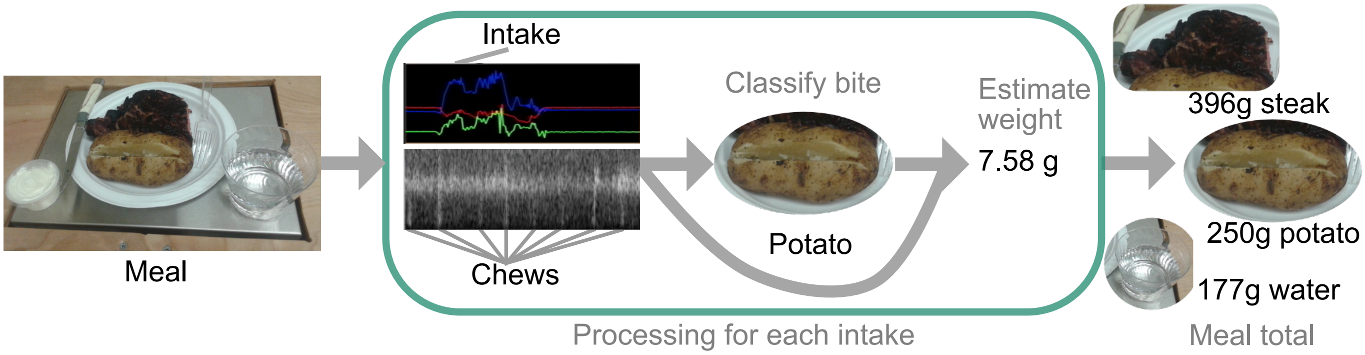

Figure 5: Overview of pipeline. Each intake is classified based on motion and sensor data, and the raw data plus food type information form input to estimate of amount consumed. Results are aggregated across all intakes to form meal total.

11 statistical features (mean, covariance, and derivatives), 15 temporal shape features (coefficients of polynomial fit to acceleration components), and 2 frequency features (zero crossing rate and its standard deviation). We use mean and standard deviation of features from each intake's frames.

Annotation features In addition to features of the raw motion, higher order features are also used in weight estimation. We use number of chews after an intake, average period between chews, and duration of intake window. We hypothesize that due to physical constraints, number of chews will be related to intake size (with more food requiring more chewing). Time between chews may not be linearly related to intake size, as both very small and large pieces of food may lead to gaps between bites, but for different reasons. However this feature may interact with food type. Finally, duration of intake window captures overall how much time is spent manipulating food, whether by chewing or mouthing.

Estimation and evaluation

To evaluate classification, we used leave-one-intake-out (LOIO) cross-validation, meaning N-folds where N is the number of intakes. Leave-one-person out could not be used, as foods were rarely shared across individuals. We did not use leave-one-sample out as this may overstate true accuracy, and we ensured that windows for adjacent intakes did not overlap.

Food type

To estimate food type, we trained a random forest classifier with 40 trees,[2] using most combinations of sensors (Glass, right/left watch, audio) to compare their contributions. We evaluate results for all 1768 intakes (using LOIO for each) using accuracy, defined as percent of intakes correctly classified. We compare to the baseline of assigning each intake to the most common class (popcorn), which has an accuracy of 11%.

Weight estimation

Weight estimation is treated as a regression problem with the intake weights to be predicted. The features are as for food type (audio, motion features), with the addition of annotation features. While we excluded foods with fewer than four intakes at the beginning, we now also exclude foods with fewer than 10 singleton intakes to provide enough samples for the regression, and omit non-singleton intakes. That is, we exclude those where a weight is assigned to a group of intakes (such as the last burger intakes in figure 2, or when multiple bites of pizza are taken before replacing the pizza on the scale). One could divide total weight for the group by number of intakes, but it is not obvious that intake sizes do not decrease or increase systematically in such situations. We further excluded foods where weight annotations were not usable: popcorn, as intake size was near the scale resolution; and a Chinese meal involving a stir fry and rice, because the food containers were not entirely on the scale during the eating session. A small number of fluid intakes could not be used because they were outside the meal period and a researcher forgot to place the container on the scale after consumption.

Thus, weight classification is done on a total of 17 foods plus drinks (18 classes total), over a total of 500 solid intakes and 171 drink intakes (671 intakes total), again using LOIO.

We aim to understand the impact of knowing food type on estimating food weight, and what improvements can be made over assuming that each intake will have the mean weight for that food type. We compare the following approaches:

Mode uses mean weight of the intake mode (solid, liquid).

Type assigns the mean weight for the intake mode and food type. Note that this is not the same as mean for the food type, since foods such as soup and milkshake were consumed in multiple ways (sometimes with a spoon, others from a straw or sipping from a container).

Full uses random forest regression with 40 trees, and includes all features (audio, motion, annotation) and ground truth for food type and intake mode. This lets us determine what portion of errors are due to inference of food type.

Full$_I$ uses all features (audio, motion, annotation) as above, but instead of using ground truth, uses inferred food type and intake mode. This lets us determine what performance can be achieved with current type inference accuracy.

An overview of the estimation process is shown in figure 5, where food type is determined on a per-intake basis, and then forms input to amount estimation.

[2] Accuracy increased up to 40 trees (# of classes) then plateaued.

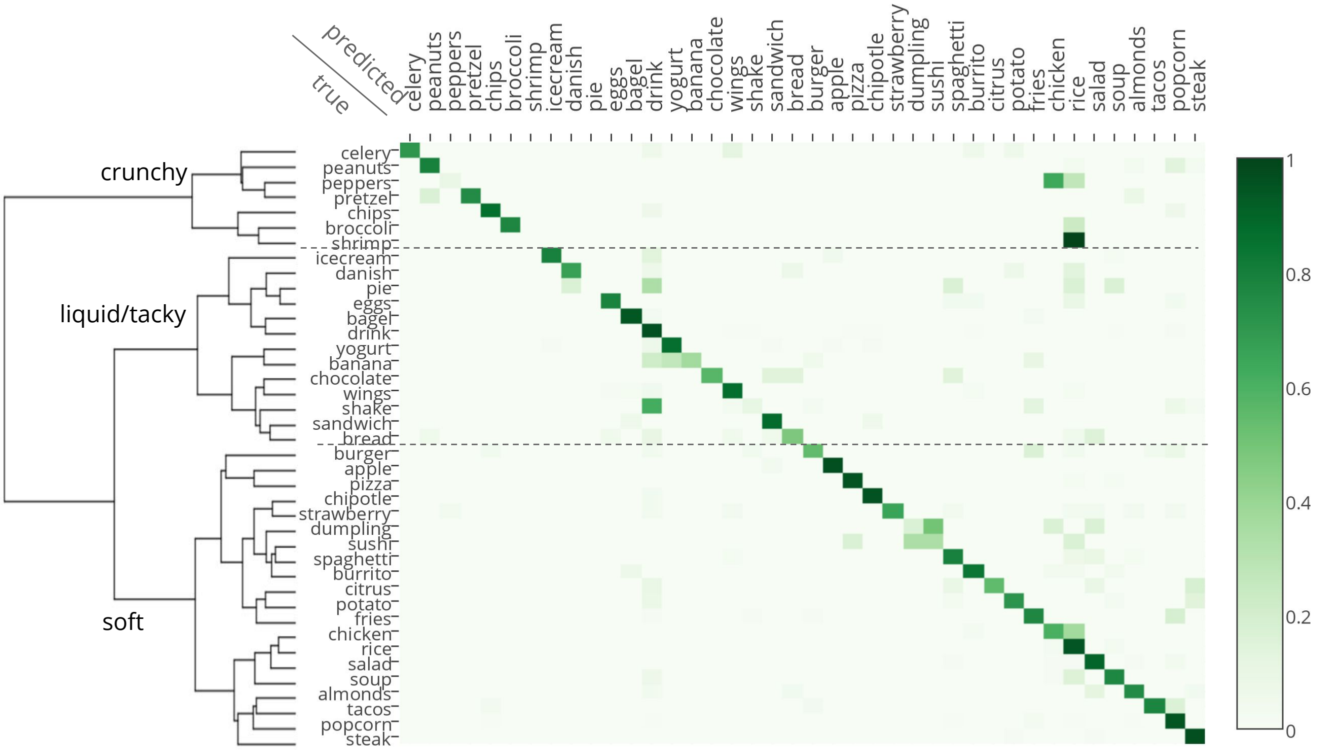

Figure 6: Hierarchical clustering by sensor features and confusion matrix for food type classification with all sensors (AGRL).

For each food, we calculate accuracy using the relative mean absolute percentage error. That is, we take the mean of the absolute value of the differences between actual and estimated weight for each intake, and then divide by mean intake size for that food. To calculate overall values for solid foods, we use the food averages weighted by number of intakes for each.

RESULTS

Food type classification

Accuracy for different combinations of motion and audio sensors is shown in figure 7, indicated with abbreviations (A=acoustic, G=Glass, R=right watch and L=left watch). The best combination is all audio and motion sensors (AGRL), with 82.7% accuracy. The three motion sensors (GRL: both wrists, head motion) achieved 76.2% accuracy, while audio alone had 67.8% accuracy. Note that there was a substantial boost going from one to two sensors, with the difference between the best single sensor (A) and worst pair (RL) being 4.6%. By chance all participants were right handed, so combinations with the right watch always indicate using data from the dominant hand. Our results suggest using multiple sensors can improve accuracy substantially, but multiple types of combinations may be feasible. For users who find audio sensing too invasive, the tradeoff between 82.7 and 76.2% accuracy may be acceptable. Further, while multiple sensors may be obtrusive, when higher food type accuracy is needed these may be combined into a single housing.

For comparison, prior work with the largest number of foods (19 foods) achieved an accuracy of 80% using an audio sensor

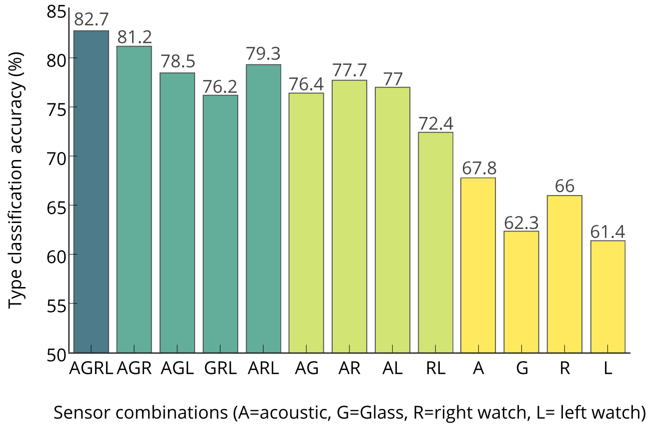

Figure 7: Food type classification accuracy by sensor (# of sensors is highlighted with color grouping).

[4], though all 3 participants ate the same foods and accuracy was 70-75% with an earbud design like ours. One reason is that motion and audio data are each insufficient (celery and chicken wings may sound alike if eaten by the same person; crunchy foods like pretzels and nuts may have similar noises), and motion alone may not be able to tell us if a person is eating yogurt or soup, but the two modalities together provide information on complementary aspects of eating.

Detailed results using the best combination of sensors (AGRL: head and wrist motion, audio) are shown in a confusion matrix

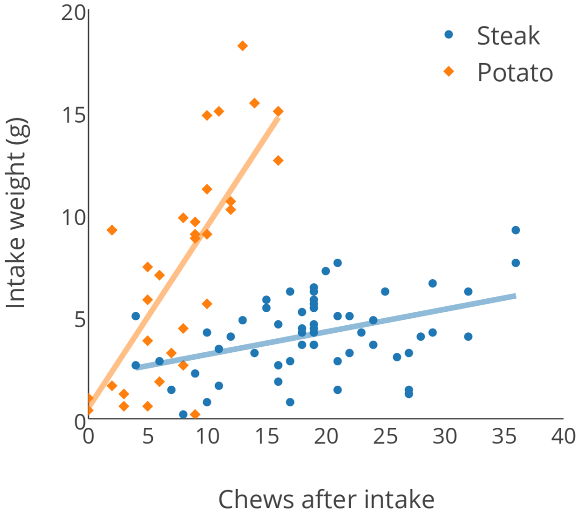

Figure 8: Food weight versus chews after intake.

in figure 6. To order the food types, we applied complete-link clustering to the mean feature vectors for each food, shown in the dendrogram at left. We find that crunchy and soft foods cluster well (top and bottom sections), while liquids and tacky foods (e.g. bagel with cream cheese) are grouped together in the middle. Despite the large number of classes, the results lie strongly along the diagonal. Note that all beverages aside from a milkshake were combined into a single "drink" class. Yet the matrix shows that indeed the shake is mostly a drink. This is supported by how it was consumed: mostly with a straw, but sometimes with a spoon (as there were chunks of chocolate in the shake). We classify each intake as a single type, and use the first (i.e. most dominant) annotation as the ground truth. In the case of shrimp and rice, a review of the video shows that the participant did not mix the foods but instead while still chewing one, often took a large bite of the other, so foods may be mixed in the mouth even though the intakes are distinct.

While most foods were only consumed once, popcorn and water were both consumed by two people, enabling us to conduct a small between-subjects evaluation. We train solely on data from one participant, and test on the other. Accuracy was 89%, significantly higher than the chance baseline of 52%, showing that results can generalize across individuals.

Weight estimation

Initially, we hypothesized that food type would be key to accurate estimation of amount consumed. Figure 8 illustrates the connection for one of the features we extract for a meal of steak and potato. The number of chews after intake and intake size are strongly correlated, yet this is only apparent once the intakes are separated by type.

Weight estimation error for each food and solids and drinks overall is shown in table 1. For the individual foods, we report mean absolute error per intake as a percentage of the mean weight for that food type. As we hypothesized initially, knowledge of food type is critical. Error is reduced by 85.4% using mean weight for each food (Type) rather than only mean of the intake modality (Mode). For foods with unusually small intakes, such as peanuts, the gain is even greater, with a reduction from 717.2% to 35.7% error. As shown in figure 4, mean intake weight varied considerably by food type (notably it also varied significantly by food type within individuals). Thus, assuming a standard intake size will lead to considerable error, but methods focused on counting bites (e.g. using wrist motion or smart utensils) can be substantially improved by incorporating food type information – whether inferred automatically or provided by a user.

Table 1: Weight estimate error for each food, solids overall, and drink in percent. Intakes are the number of singletons.

Class	Intakes	Mean	Type	Full	Full_I
salad	60	48.3	47.9	34.5	36.5
steak	58	87.7	36.0	31.5	43.3
soup	51	35.1	33.1	30.2	39.7
chipotle	44	22.3	19.2	19.1	19.4
spaghetti	41	56.6	52.6	43.3	47.9
chicken	38	85.9	61.1	66.6	102.3
peanuts	38	717.2	35.7	35.6	36.3
potato	33	61.1	61.1	42.5	43.6
shake	20	47.3	47.1	35.0	47.3
yogurt	20	28.3	27.0	21.8	24.8
eggs	19	46.4	45.8	32.9	32.9
almonds	15	435.2	31.8	27.9	29.8
burrito	15	86.8	95.8	62.5	63.3
fries	14	273. 0	47.1	47.2	73.2
broccoli	13	76.7	23.9	27.2	32.3
citrus	11	14.0	9.9	10.2	49.7
bread	10	41.9	37.0	34.1	48.2
All solids	500	127.3	41.9	35.4	45.3
drink	171	60.4	60.4	47.2	47.2

Results were further improved by leveraging annotation, audio, and motion features (Full) with a 6.5% error reduction over Type (total of 91.9% over Mode baseline). The improvement is expected for a food such as the salad with salmon (13.4% improvement in error) and burrito (33.3% improvement), as both have a lot of variation between bites. However, there is also considerable error reduction on potato (18.6%) and eggs (13.0%), which have little variation in content or texture, though both had a high standard deviation in intake size. Countable foods such as broccoli florets and fries, which have consistent intake sizes, benefit the least from adding sensor data.

As illustrated in figure 9, there was no correlation between mean intake weight and accuracy ($r = -0.17$), or meal duration and accuracy ($r = 0.20$), though standard deviation of intake size and accuracy were weakly correlated ($r = 0.50$). This suggests that few samples may be needed for foods with little variation, while accuracy for those where intake sizes vary considerably can be improved with more training data. Importantly, the lack of correlation with intake weight and meal duration suggests that our accuracy is the same for both meals and shorter duration snacks. This is vital for nutrition monitoring, as such snacks may be dense in calories. Varia-

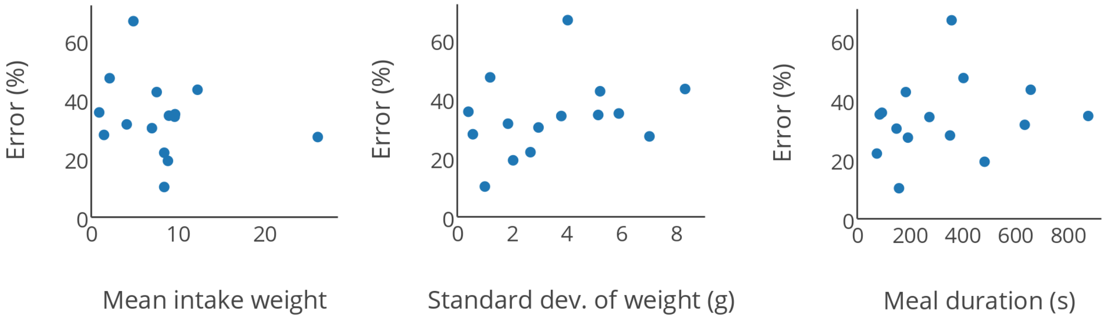

Figure 9: Weight estimation error as a function of three meal features. Each point indicates a single food.

tions in accuracy between individuals cannot be disentangled from variation due to food type, since few types were consumed by multiple participants.

While our weight and type estimates are imperfect, we aim primarily to match human performance without the need for human labor. For example, our average absolute error on solid foods of 35.4% is comparable to crowdsourced image annotation (Platemate) at 33.2% [21] error in calorie estimate, where it also took on average an hour and a half for the task to be completed. Neither our approach nor Platemate systematically over or underestimate quantity consumed. On the other hand, self reports of energy intake consistently under estimate consumption by about 20% [15].

Combined system performance

To understand how the type and amount estimates can provide calorie counts, consider the running example of a meal of 396g steak (at 2.5kcal/g) and a 250g baked potato (at 1.0kcal/g), for a total of 1240kcal. Using food type and sensor features (Full), each food type (steak, baked potato) would be mapped to items in a food database. We used the USDA Nutrient Database, which can also be queried with Google searches. Our worst case estimate would be 1240 ± 439kcal, where every intake is an over- or underestimate. In reality, meal error will be lower as variance in per-intake estimates cancel out.

For applications such as automated calorie counting, both food type and amount must be estimated. While type accuracy was imperfect, at 83%, using inferred food type (Full_I) only modestly increased weight estimate error (mean 9.9% for solid), compared to using ground truth of food type. In practice, sequence voting among intakes [2] could reduce type error and achieve weight estimation error closer to the Full result.

DISCUSSION

A key implication of our work is the need for devices with multiple sensing modalities (motion, audio). While a combination of modalities led to significant improvements in accuracy, this comes at the cost of users having to wear multiple devices. In the future, more work is needed on larger and more diverse populations (including individuals with chronic disease) to understand generalizability and how to best adapt the classification system to individuals. Further, there is a need to balance detailed ground truth (at the level of chews), with weakly labeled data collected in the wild. Annotation from video (rather than marking meal start and end) let us determine how much meal time was spent chewing versus talking or between bites, but this approach does not scale or allow more diverse eating contexts.

One limitation is that we did not distinguish between types of liquids, and that can have a significant impact on caloric intake. However, once the eating modality (liquid, solid) and amount (e.g. ounces of fluid) are inferred, personalized training data such as knowledge that a user usually drinks coffee or juice in the morning and other external cues (such as GPS data indicating that user is at a coffee shop) may help fill in this gap. Alternatively, sensor data may be augmented with food and drink images that could allow liquids to be identified.

Similarly, we did not decompose complex foods into their parts, and did not attempt to recognize each component of each bite. Despite the video data, it is difficult to create such annotations with high accuracy. For example, a user may add guacamole to a chip but eat the combination in multiple bites, making it difficult to determine the proportion of guacamole and chip in each bite. Our aim is primarily to achieve performance comparable to human annotators without the human labor, but in the future users may be able to provide high level annotation on combined foods (correcting automated logs).

CONCLUSIONS

To create fully automated nutrition logs, we need methods that can automatically determine what and how much a person consumes, rather than identifying only *if* they are eating. Such a solution may ultimately lead to computer-generated food logs (removing the need for human labor, and making logs objective), identification of nutritional deficiencies, and better management of diseases like diabetes. We show that 1) motion and audio sensing together lead to significantly more accurate estimates of food type (82.7% accuracy) than either modality alone, and 2) with knowledge of food type, food quantity can be estimated with 35.4% error (reduced from a 127.3% baseline error), on par with human annotators. Our publicly available dataset with detailed annotations will enable further work to reduce these error rates (`http://www.skleinberg.org/data.html`).

ACKNOWLEDGMENTS

This work was supported in part by NSF Award #1347119.

REFERENCES

1. Oliver Amft, Holger Junker, and Gerhard Troster. 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)*. IEEE Computer Society, Washington, DC, USA, 160–163. DOI: http://dx.doi.org/10.1109/ISWC.2005.17

2. Oliver Amft, Martin Kusserow, and Gerhard Troster. 2009. Bite weight prediction from acoustic recognition of chewing. *IEEE Transactions on Biomedical Engineering* 56, 6 (2009), 1663–1672.

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. http://dx.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 (2009), 62–70.

5. Abdelkareem Bedri, Apoorva Verlekar, Edison Thomaz, Valerie Avva, and Thad Starner. 2015. Detecting Mastication: A Wearable Approach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 247–250. DOI: http://dx.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. http://dx.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)*. IEEE Computer Society, Washington, DC, USA, 580–587. DOI:http://dx.doi.org/10.1109/WACV.2015.83

8. Jingyuan Cheng, Oliver Amft, and Paul Lukowicz. 2010. Active Capacitive Sensing: Exploring a New Wearable Sensing Modality for Activity Recognition. In *Proceedings of the 8th International Conference on Pervasive Computing (Pervasive'10)*. Springer-Verlag, Berlin, Heidelberg, 319–336. DOI: http://dx.doi.org/10.1007/978-3-642-12654-3_19

9. Jingyuan Cheng, Bo Zhou, Kai Kunze, Carl Christian Rheinländer, Sebastian Wille, Norbert Wehn, Jens Weppner, and Paul Lukowicz. 2013. Activity Recognition and Nutrition Monitoring in Every Day Situations with a Textile Capacitive Neckband. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp '13 Adjunct)*. 155–158. http://doi.acm.org/10.1145/2494091.2494143

10. Pei-Yu . Y. Chi, Jen-Hao . H. Chen, Hao-Hua . H. Chu, and Jin-Ling . L. Lo. 2008. Enabling Calorie-Aware Cooking in a Smart Kitchen. In *Proceedings of the 3rd International Conference on Persuasive Technology (PERSUASIVE '08)*. Springer-Verlag, Berlin, Heidelberg, 116–127. DOI: http://dx.doi.org/10.1007/978-3-540-68504-3_11

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.

12. Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Andrew Hoover. 2014. Detecting periods of eating during free-living by tracking wrist motion. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (2014), 1253–1260.

13. Adam Drewnowski, Susan Ahlstrom Renderson, Alissa Driscoll, and Barbara J Rolls. 1997. The Dietary Variety Score: assessing diet quality in healthy young and older adults. *Journal of the American Dietetic Association* 97, 3 (1997), 266–271. DOI: http://dx.doi.org/10.1016/S0002-8223(97)00070-9

14. Joey Hagedorn, Joshua Hailpern, and Karrie G. Karahalios. 2008. VCode and VData: Illustrating a New Framework for Supporting the Video Annotation Workflow. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '08)*. 317–321. http://doi.acm.org/10.1145/1385569.1385622

15. R. J. Hill and P. S. W. Davies. 2001. The validity of self-reported energy intake as determined using the doubly labelled water technique. *British Journal of Nutrition* 85, 04 (2001), 415–430.

16. Azusa Kadomura, Cheng-Yuan Li, Koji Tsukada, Hao-Hua Chu, and Itiro Siio. 2014. Persuasive Technology to Improve Eating Behavior Using a Sensor-embedded Fork. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. 319–329. http://doi.acm.org/10.1145/2632048.2632093

17. Haik Kalantarian and Majid Sarrafzadeh. 2015. Audio-based detection and evaluation of eating behavior using the smartwatch platform. *Computers in Biology and Medicine* 65 (10 2015), 1–9. DOI: http://dx.doi.org/10.1016/j.compbiomed.2015.07.013

18. J. Lester, D. Tan, S. Patel, and A. J. B. Brush. 2010. Automatic classification of daily fluid intake. *Pervasive Health* (3 2010), 1–8. DOI:http://dx.doi.org/10.4108/ICST.PERVASIVEHEALTH2010.8906

19. Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. 2016. Multimodality Sensing for Eating Recognition. In *Proceedings of the 10th International Conference on Pervasive Computing Technologies for Healthcare*.

20. Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*. 1233–1241.

21. 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. http://doi.acm.org/10.1145/2047196.2047198

22. Sebastian Päßler and Wolf-Joachim Fischer. 2011. Acoustical method for objective food intake monitoring using a wearable sensor system. In *Pervasive Health*.

23. Sebastian Päßler, Matthias Wolff, and Wolf-Joachim Fischer. 2012. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological measurement* 33, 6 (2012), 1073.

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

25. 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. http://doi.acm.org/10.1145/2594368.2594386

26. 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.

27. Sougata Sen, Vigneshwaran Subbaraju, Archan MISRA, Rajesh Krishna Balan, and Youngki Lee. 2015. The Case for Smartwatch-based Diet Monitoring. In *Workshop on Sensing Systems and Applications Using Wrist Worn Smart Devices*.

28. 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)*. ACM, New York, NY, USA, 1029–1040. DOI:http://dx.doi.org/10.1145/2750858.2807545

29. Edison Thomaz, Aman Parnami, Jonathan Bidwell, Irfan Essa, and Gregory D. Abowd. 2013a. Technological Approaches for Addressing Privacy Concerns when Recognizing Eating Behaviors with Wearable Cameras. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. 739–748. http://doi.acm.org/10.1145/2493432.2493509

30. Edison Thomaz, Aman Parnami, Irfan Essa, and Gregory D. Abowd. 2013b. 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

31. 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

32. Koji Yatani and Khai N. Truong. 2012. BodyScope: A Wearable Acoustic Sensor for Activity Recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. 341–350. http://doi.acm.org/10.1145/2370216.2370269

33. Bo Zhou, Jingyuan Cheng, P. Lukowicz, A. Reiss, and O. Amft. 2015a. Monitoring Dietary Behavior with a Smart Dining Tray. *IEEE Pervasive Computing* 14, 4 (10 2015), 46–56. DOI:http://dx.doi.org/10.1109/MPRV.2015.79

34. Bo Zhou, Jingyuan Cheng, Mathias Sundholm, Attila Reiss, Wuhuang Huang, Oliver Amft, and Paul Lukowicz. 2015b. Smart table surface: A novel approach to pervasive dining monitoring. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 155–162. DOI: http://dx.doi.org/10.1109/PERCOM.2015.7146522