Multimodality Sensing for Eating Recognition

Christopher Merck
cmerck@stevens.edu

Christina Maher
cmaher@stevens.edu

Mark Mirtchouk
mmirtcho@stevens.edu

Min Zheng
mzheng3@stevens.edu

Yuxiao Huang
yhuang23@stevens.edu

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Stevens Institute of Technology
Hoboken, NJ

ABSTRACT

While many sensors can monitor physical activity, there is no device that can unobtrusively measure eating at the same level of detail. Yet, tracking and reacting to food consumption is key to managing many chronic diseases such as obesity and diabetes. Eating recognition has primarily used a single sensor at a time in a constrained environment but sensors may fail and each may pick up different types of eating. We present a multi-modality study of eating recognition, which combines head and wrist motion (Google Glass, smartwatches on each wrist), with audio (custom earbud microphone). We collect 72 hours of data from 6 participants wearing all sensors and eating an unrestricted set of foods, and annotate video recordings to obtain ground truth. Using our noise cancellation method, audio sensing alone achieved 92% precision and 89% recall in finding meals, while motion sensing was needed to find individual intakes.

CCS Concepts

•Human-centered computing → Ubiquitous and mobile computing;

Keywords

Eating recognition; Acoustic and motion sensing

1. INTRODUCTION

Nutrition and physical activity are key components of maintaining health and treating disease, but while physical activity measurement has reached widespread consumer adoption, similar devices to track eating have lagged behind. Understanding the quantity and type of food eaten is necessary to manage chronic diseases such as diabetes and obesity and knowing when someone is eating can improve interaction between people and their environment (e.g. silencing phone alerts during a meal or giving medication reminders). The most common method for longterm tracking of food consumption is with food logs using paper journals or smartphone apps but these face low adherence, put the burden on the user, and have only been evaluated during short-term tracking. In comparisons against doubly labeled water (a gold standard for estimating energy intake), self reports led to overestimation of more than 20% in many studies [10].

While there have been computational advances in this area, moving from small-scale studies to real-world environments remains a challenge. The comparative accuracy of different sensing modalities is unknown since prior work has evaluated each individually (e.g. only acoustic or motion sensing), often with constrained food choices or simplified environments (e.g. no multitasking, no background noise). Work in free-living environments on the other hand, has not had the level of detail needed for ground truth (e.g. relying on users to remember when they had a meal).

The main contributions of this paper are 1) rigorous comparison of acoustic and motion (wrist and head) sensors individually and in combination; 2) a unique publicly available data resource,[1] annotated from video at the level of chewing and swallowing (rather than meals); and 3) demonstration of the feasibility of eating recognition with completely unconstrained foods and multitasking.

We develop the ACE (accelerometer and audio-based calorie estimation) dataset, which includes 6 participants (2 ~6 hr sessions each, total ~72 hrs) wearing audio and motion (head, wrist) sensors simultaneously. Using a customized earbud with two microphones we show that our noise cancellation procedure can remove nearly all external noise and user and external speech. Acoustic sensors worked for more foods than expected, but failed completely with soft foods like bananas. Motion sensors capture more activities (e.g. moving food to the mouth) but require more feature engineering to fully exploit. We have also collected free-living data from the same participants and continuously weighed food during consumption for future use.

2. RELATED WORK

Automated eating recognition can be categorized in three main ways: environment (laboratory, free-living), sensors used, and output (identifying meal periods, foods consumed, or individual chews and swallows), summarized in table 1.

2.1 Acoustic Sensors

One of the earliest works used acoustic sensors to detect chewing and food type [2]. While accuracy was high,

[1]Avialable at: http://www.skleinberg.org/data.html

PervasiveHealth 2016, May 16-19, Cancun, Mexico
Copyright © 2016 EAI
DOI 10.4108/eai.16-5-2016.2263281

Study	Sensors	Ground truth	Foods	# People	Duration	Environment	Output
Amft et al. [2]	Audio	Not reported	4	4	1 hour	Lab	Chew times
BodyBeat [15]	Audio	Not reported	4	14	3.5 hrs	Lab	Eating times
Thomaz et al. [17]	Single wrist motion	Video	5	21	10.5 hrs	Lab†	Eating times
Dong et al. [8]	Single wrist motion	Participant log	Free	43	449 hrs	Free-living	Eating start
Rahman et al. [14]	Head motion	Real-time log	Free	38	76 hrs	Lab	Eating times
This study	Motion (Head, both wrists), Audio	Video	Free	6	72 hrs	Lab*	Chew, intake

Table 1: Comparison of related work. †Data collected for 9 others in free-living conditions. *While we report only lab data in this paper, we have collected approximately 2 days of free-living data for the same participants.

only four foods were tested and participants were told how to chew (with mouth closed) and how large a bite to take (small enough to be consumed at once). Later work developed specialized microphones to better capture non-speech sounds [15] and calculated bite weight from chewing sounds [1]. While these works achieved high accuracy, participants ate a limited set of foods in quiet laboratory environments, and accuracy with realistic background noise and representative food samples is unknown. In-ear and reference microphones have been used for comparing sound levels to improve eating recognition [13] though that work did not remove external noise from the in-ear signal. More recently, ambient sounds were used to detect eating from audio recorded at the wrist [19], though this detected meals rather than chews, which are needed to estimate food type and quantity.

2.2 Motion Sensors

Sensors placed on the wrist and arm have been used to identify gestures related to food and drink intake [16]. Wrist-based sensors are the most frequently used for detecting periods of eating in free-living environments, and have also been used to identify individual bites of food, though accuracy decreased when food type was not controlled [7]. In free-living conditions, the primary obstacle is determining ground truth for training and evaluation. In [8] this was based on participant logs (which may contain inaccuracies and omissions). Since participants used their hands to mark the start and end of eating (using written logs or a button on the watch), this motion precedes all eating and may be responsible for some of the inference accuracy. In other work [17], a first-person camera automatically captured images to allow objective ground truth in free-living conditions, though accuracy was lower than in other lab-based studies (66.7% precision, 88.8% recall). Google Glass allowed recognition of periods of eating based on head movement [14], though this work did not detect individual chews and required personal training data to achieve higher accuracy. Other work with smart eyewear identified activities based on blinking, but had lower accuracy than other sensing modalities [11]. Finally, sensors have been placed around the throat to detect the specific motion of swallowing and eating [3, 6], but these may be uncomfortable or too invasive for long-term use.

2.3 Environmental Sensors

In contrast to body-worn sensors, external cues such as instrumented environments (e.g. RFID tagging [5], pressure sensitive table surfaces [21]) have been used, but reduce mobility. Another approach is using a body-worn camera to automatically take photos at regular intervals, though this requires some photos to be discarded or edited to preserve privacy [18]. In other cases individuals intentionally capture images of meals, which are annotated by crowd-workers to determine nutrition content [12], but this does not yield the real-time information on chewing and swallowing needed for classifying food type and quantity automatically.

3. STUDY DESIGN AND SENSORS

Our overall goal is to determine the relative contributions of audio and motion sensors (mounted on the head and both wrists) for recognizing eating in realistic scenarios, while balancing the need for ground truth, as the accuracy of sensors in the same scenario is not yet known. Some prior work has reported on multiple sensors, but each was evaluated in a separate experiment [4]. To address this, we allow free choice of food, unconstrained activity sequences and multitasking, while collecting data primarily in our laboratory, which was outfitted with video cameras. We discuss each device used, then the noise cancellation method, data collection, and annotation. An overview is shown in figure 1.

3.1 Sensors

3.1.1 Audio

Experiments have avoided capturing background noise by requiring a quiet environment or no conversation, but in reality many meals are shared or take place in loud restaurants. Instead, we augmented a standard earbud with internal and external microphones to enable noise cancellation. Most eating-related sounds are audible only on the internal mic, while talking and noises external to a participant (e.g. other people's speech) are captured by both mics. We remove the portion of the internal signal that is well-predicted by the external recordings, meaning background noise and the eating noises of others, and noises that may be confounded with eating. For example, walking over gravel may sound like chewing, but because our audio cancellation does not filter based on the type of noise (e.g. eating vs speech) but rather its source, these will be removed. The noise cancellation process is described in depth later in the paper.

Figure 2 shows the earbud. After removing the speakers, we used wires from one side for the internal microphone (which was then sealed inside the earbud), and from the other for the external microphone, which was glued to the back of the earbud shell. Audio was recorded using a pocket voice recorder with uncompressed 16-bit 44.1kHz samples.

3.1.2 Wrist Motion

Many eating activities, such as lifting food to the mouth or

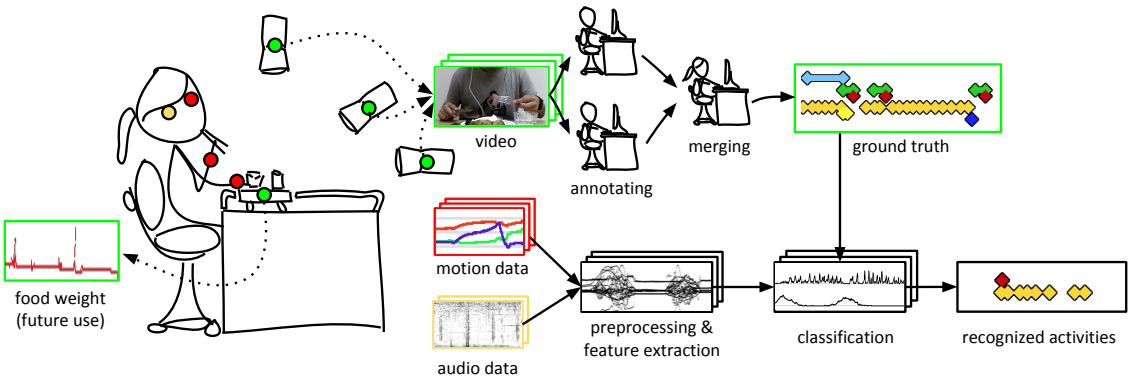


Figure 1: Overview of study design, showing ground truth (green circles), motion sensors (red circles) and acoustic sensor (yellow circle). Video is annotated by two researchers and a third participates in merging annotations (colored diamonds).

Figure 2: Earbud modified to record audio from two mics, one inside the earbud and one attached to the back (shown).

cutting food involve wrist movement. However, most work has used a single motion sensor placed on the dominant arm, while watches are usually worn on the non-dominant side. We aim to understand how this affects accuracy.

When possible we used consumer devices to determine what performance can be achieved in real-world settings. We used the LG G watch (W100) due to its Android Wear OS, 9-axis inertial motion sensor (accelerometer, gyroscope, and magnetometer), and resizable wristband. We recorded at 15Hz as a trade off between battery life and recording rate, as we wanted to be certain that the full 6 hours of data would be recorded. In lab conditions both watches always recorded for the full time period, while in free-living conditions, the watches had 8–12 hours of battery life. We developed an Android app to log the sensor data.

3.1.3 Head Motion

Head motion was captured using Google Glass, which has no lenses but has a similar form factor to glasses along with a small display. Glass contains a 9-axis inertial motion sensor situated near the right temple. We hypothesized that this would capture the motion of the head toward food or drink, and that the sensor's location near the ear may allow it to pick up subtle motions of the jaw during chewing.

Prior work using Glass for this purpose recorded at ~2.5 Hz [14], due to the task (classifying 1-minute windows as eating or not) and battery life concerns. We developed Glassware to record at 15Hz as for the watches. Since Glass has limited battery life and could not record at this rate for a full session, we connected it to a small external battery.

3.1.4 Video

We used three IP-based, 1280x720 pixel resolution, H.264 encoding cameras attached to software-controllable pan-tilt mounts: one near the ceiling for a top view, and two clamped to the front and side of the table. While the participant was seated at the table, the three viewing angles ensured mouth and throat were visible even while hands were raised to the face. During other times the cameras were panned to track the participant when they moved around the lab space.

We recorded at 30 frames per second (fps). As the cameras are not real-time devices, actual frame-rate varied between 8 and 32fps (mean 30.08fps). While only 0.26% faster than desired, this yields over 50s of timing discrepancy after 6h. Because the timing of events and synchronization of all recording modalities was critical for high-quality annotation and inference, we regularized the video to 30fps using the 1-sec resolution timestamp imposed on each frame by the camera software. We modified a video processing program[2] using optical character recognition[3] to recover the timestamps and duplicate or remove frames from the video as needed to maintain 30fps without re-encoding.

3.1.5 Sensor Synchonization

Data were not extracted in real-time, and the device clocks may not remain in sync, so we developed a synchronization procedure. At the beginning and end of each session we held all three sensors together and performed three controlled taps on the table. This generated strong spikes in the motion and audio data and was clearly visible on the cameras. The tap times in each data stream were used to correct for timing offset and slight differences in device clock rates.

3.2 Noise Cancellation

While acoustic sensors can capture data specific to eating (e.g. sounds of individual chews and swallows), they are often used in quiet laboratory settings, without background noise. To enable real-world use, we develop an approach for filtering out all sounds except for the user's eating, using the idea that unwanted signals will be recorded by both micro-

[2]http://www.avidemux.org/
[3]https://github.com/tesseract-ocr/tesseract

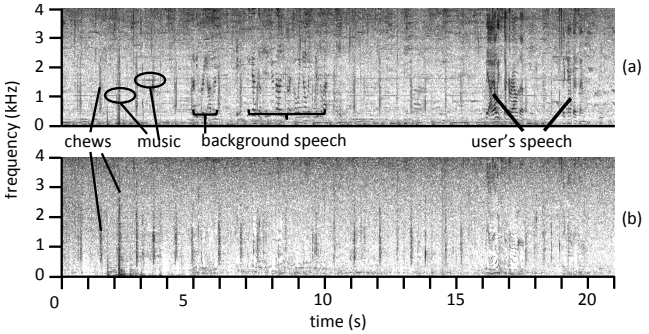


Figure 3: Spectrograms of (a) raw inner microphone audio and (b) recovered eating sound signal.

phones. While we cannot guarantee that all speech is unrecoverable, this processing provides some privacy protection and the approach is fast enough to be run in real-time on a low-power digital signal processor embedded in an earbud.

The outer microphone of our earbud measures the background noise signal, n, while the inner microphone measures a noisy eating signal, x, which is a function of the pure eating signal, s, and the noise, n. Noise here means non-eating sounds and may include background music, the user's speech, or background speech. We cannot simply subtract n from x because there is a frequency-dependent phase shift in x due to the ~1cm distance between the microphones, and the earbud enclosure attenuates the noise in a frequency-dependent manner. We model x as $s + E(n)$, where E is the unknown transfer function of the earbud enclosure.

We performed a static characterization of the transfer function E, but it depends on the earbud construction, user, and environment. Using an adaptive filter [20] we continuously update an estimate of E and use it to remove an estimate of the noise from the inner microphone signal. Intuitively, the filter subtracts out the information in n from x, leaving only the desired information s. For each sample of audio, the filter computes the dot product of the previous k samples from the outer microphone n with a k-element weight vector w to yield an estimate of the inner microphone signal, $\hat{x} = w \cdot n$. We then subtract this estimate of the transformed noise from the actual inner microphone signal x to obtain the recovered eating sound signal, $\hat{s} = x - \hat{x}$.

Finally we update filter weights to adapt to changes in the earbud transfer function: $w := w + \eta \hat{s} \cdot n$, where η is an adaptive learning rate parameter set by $\eta = \min(\frac{\eta_0}{|n|}, \frac{1}{|n|^2})$ and η_0 is the base learning rate. This provides rapid adaptation to changing environmental signals (via the $\eta_0/|n|$ term) and avoids filter instability by limiting the learning rate to the maximum theoretical limit ($1/|n|^2$) [20].

We use a filter length of 50ms and set $\eta_0 = 0.6$ after empirically tuning for maximum rejection of background noises and minimum distortion. This technique rejects background music and speech signals while preserving eating sounds. Figure 3 shows a typical example of audio from a busy fast food restaurant. The sample includes background music from the restaurant, activities of other customers, and a shared meal. After filtering, chewing sounds remain while background music and speech are removed. While the data in this paper were collected in the lab, this demonstrates that the approach can handle more realistic environments.

3.3 Data Collection

With IRB approval we collected data from 6 participants (4 male; all aged 18-35) in two ~6hr lab data collection sessions for each participant.[4] Each session contained at least two meals, chosen by the participants. Ten sessions included breakfast and lunch while two others included breakfast and dinner, and six sessions had additional snacks. While eating in the laboratory always happened at the same table (instrumented with a scale), the shared space allowed for social meals and natural conversation (participant and research group members could eat at the same time). Fluids could be consumed anywhere in the room.

During each session one researcher was responsible for ensuring consistency in procedures, calibrating and synchronizing equipment, starting and ending recording, monitoring video and other sensors during data collection, and transferring data off the devices. Due to the complexity and many potential points of failure, we did test runs of data collection and developed a ~40 item checklist for the procedure.

3.4 Annotation and Ground Truth

Prior work has mainly labeled activities in real-time [14] or used a constrained sequence of activities [15]. Less frequently, video has been used to annotate the start and end times of eating and other activities [17]. In contrast, we allow free choice of activities (e.g. simultaneous talking and eating), and annotate at a finer level of granularity than meals. This enables future work estimating food type and amount, and provides more precise meal times. Individuals do not chew continuously, they may pause to stir their soup, talk, or answer email so many non-eating activities may be erroneously labeled as eating when using coarse labels such as only the beginning and end of a meal. Further, some of these other activities, such as stirring or preparing food, may be useful for recognizing precursors to food intake.

Video annotation is time consuming, but it is the closest to true ground truth, and most activities were easily observed. We labeled the video using vCode [9], and to avoid biasing the results, no sensor data was consulted during annotation. It is possible that events like swallows, which are more apparent on the audio, may be missed. Activities annotated and their definitions are as follows. Events with duration are marked with [] and L,R subscripts denote those where the left and right hands may be separately involved.

Preparation$_{L,R}$[] Interacting with food or drink other than raising it to the mouth (e.g. stirring soup, adding croutons to a salad, and moving food from takeout containers to a plate). The rationale is that these activities occur prior to consumption and may have identifiable wrist movement.

Delivery$_{L,R}$[] Continuous motion of bringing food or drink to the mouth using one or both hands. We considered "lift," which encompasses the full cycle of drawing the hand towards and away from the mouth, but in preliminary tests this cycle was often interrupted (e.g. when talking). We require that each delivery event ends in an intake.

Drink[] and **Intake** We distinguish between intakes with a duration (sipping from a straw) and consuming a discrete amount of food, corresponding to drinking and eating respectively. Note that the intake modality and not the food creates the distinction. Soup may have a continuous intake

[4] A single session was recorded for one more participant who was not available for further data collection. Due to our leave one session out evaluation we exclude this participant.

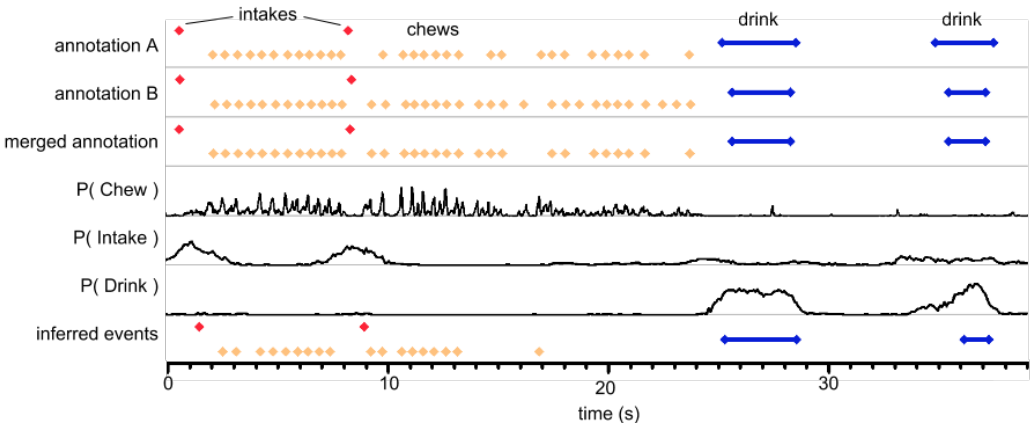


Figure 4: Example of merging annotations, along with event probability calculations and final inferred event sequence.

when drinking it from a cup and a milkshake may have a discrete intake when consumed with a spoon.

Chew This is the manipulation of food with the teeth. The time used is that when the jaw first closes.

Mouthing[] Manipulating food with the tongue, including cleaning the mouth and moving food around in the mouth.

Swallow This is the most difficult annotation using only video, but can often be heard clearly on the audio. We annotate only when swallowing is clearly visible in the neck.

Napkin$_{L,R}$[] Touching the face with a napkin. This may be conflated with delivering food to the mouth and may occlude chews. The time range begins when the napkin touches the face and ends when the napkin loses contact.

We enforced the following constraints: each hand does only one activity at a time; mouthing, swallowing, and chewing cannot be simultaneous; each intake is either continuous or discrete; and events of the same type may not overlap.

The data outside of meals include activities such as talking, napping, and doing push-ups, but we do not annotate these. Since our focus is eating detection, we aim to ensure eating and drinking are annotated accurately, while unlabeled activities provide negative examples, improving the robustness of classification. In practice, classification of swallowing was poor due to the difficulty of its annotation.

Each session was annotated independently by two annotators. This reduced the chance that codes would be missed and ensured consistency in their use. To combine annotations we developed a tool that resolved minor disagreements automatically and let us visualize and manually resolve the remainder. We used conservative thresholds for automatic merging, requiring that intake and swallow events be within 500ms and chews within 250ms. The stricter timing for chews is because two distinct events may be close in time. For ranged events, the tolerance was 1000ms, and we averaged the annotated times. Intakes were automerged 89% of the time, while chews were automerged 77% of the time. All annotations not automatically combined were discussed by three people (two annotators plus a third researcher), and automerged annotations were also reviewed during this process. Annotation of each session took ~8 hours per annotator and merging took around 2–3 hours per session.

Figure 4 illustrates the process. The events shown are mainly chews (orange) with some intakes (red), and drinking (blue bar). The annotations agree closely on the first intake, of peanuts. There are some disagreements about the second intake, part of a cookie. The first chew from B is added to the merged annotation, but some others are dropped. There is slight disagreement on the duration of the drinking episode and B's timings are used. During classification, the period of chewing disagreement had a lower probability of containing chews than the time period with higher agreement.

4. DATA PREPARATION AND ANALYSIS

We now describe the processing of raw sensor data and classification procedure. We train binary classifiers for each event (chew, intake, drink) using combinations of sensors to determine how their inclusion affects precision and recall.

4.1 Feature Extraction

4.1.1 Audio

After noise cancellation, we down-sample the audio from 44.1 kHz to 16 kHz to speed up processing. We choose this sample rate to capture chewing noises, which are present up to 4 kHz, plus the absence of signal from 4 kHz to 8kHz, to allow differentiation from more broadband non-chew signals.

Next we segment the audio into 200ms windows with a 20ms step size. The windows are large enough to capture a whole chew signal plus silence before and after, while being small enough to avoid capturing multiple chews. For each window we then extract a set of 14 features often used in speech processing using the Yaafe toolbox: energy, spectral flux, zero-crossing rate, and 11 MFCC coefficients.

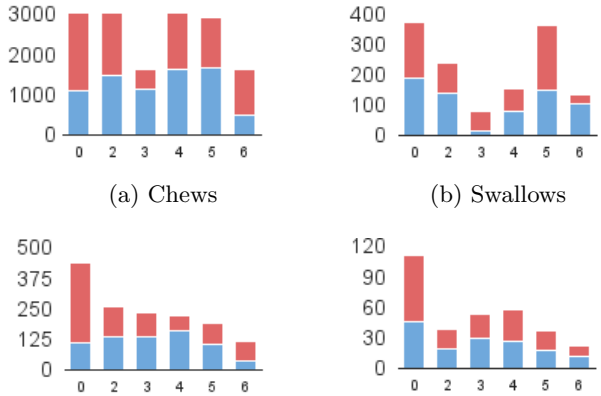
4.1.2 Motion

For each motion sensor, we segment the raw data into 5s windows with a 100ms step size. This captures movement to and from the mouth for intakes, yet contains only a single intake per window. We extract 32 features per window.

Statistical features are means of the accelerometer and gyroscope axes (6 features), to capture orientation; mean magnitude of the gyroscope vector and mean magnitude of derivative of acceleration vector (2 features), which measure total amount of linear and twisting motion; and covariance of acceleration components (3 features) to capture changes in device orientation. To capture **temporal shape** we use

(a) Salad with Salmon (b) Bagels, cream cheese (c) Sushi and dumplings (d) Steak and potato

Figure 5: Examples of meals from study.



(a) Chews (b) Swallows

(c) Food intake (d) Fluid intake

Figure 6: Number of events per participant and session (shown in red and blue).

coefficients of 4th order polynomial fits to each acceleration component with hamming window weighting (15 features). To measure **frequency of oscillation** in motion we use zero-crossing rate of high-pass filtered acceleration components and standard deviation of the zero-crossing intervals.

4.2 Classification

Each feature window is labeled with zero or more events. For instantaneous events (chew, intake) this happens when the event occurs in the center half of the window. For events with duration (drink) a window is labeled when the event's duration includes the window's center time.

In leave one session out (LOSO) validation, we omit one session during training and evaluate on that session (training on 11 sessions from 6 participants and evaluating on the 12th). Leave one participant out (LOPO) trains on 10 sessions from 5 participants and evaluates on the 6th participant. We do not evaluate on finer partitions (e.g. leave one window out) to avoid bias. We used a random forest classifier, as it outperformed other methods in prior comparisons [14]. We use the implementation in scikit-learn with 100 trees to obtain high precision event probabilities. For each fold of each experiment, one classifier is trained per event type. We chose several binary classifiers over a single multi-class classifier as multiple events may occur at one time. When training on an event-type, we only include features from relevant sensors: chew (audio, Glass), intake (watches, Glass), and drink (watches, Glass).

To get event sequences as in the annotations, we find all intervals where the probability for an event type exceeds a threshold and then drops by 50%. A 99.9th percentile threshold, recomputed for each session, optimized meal recognition while trading off recall of individual events. For instantaneous events (intake, chew) we use the midpoint time for each interval, while for events with duration (drink) the entire interval is used. Figure 4 shows example output.

4.3 Evaluation

We evaluate precision and recall of individual events and of meal periods. To determine if an inference matches the ground truth, we use the same tolerances as for merging: 250ms for chew, 500ms for intake, and 1000ms for drink.

To evaluate performance on detection of meal periods, we define a meal as a cluster of intakes or chews. Treating each intake as a vertex of a graph, we add an edge between any pair that is less than 2min apart and exclude components with under 2min of total duration. Start and end times are the first and last intake of the component. This captures consecutive intakes and chewing in between as a single meal while breaking up multi-part meals that have long pauses. The same definition is used to determine ground truth and to identify meals from inferred chews and intakes. Precision and recall of meal periods is defined in terms of amount of time correctly/incorrectly detected as a meal period.

5. EXPERIMENTAL RESULTS

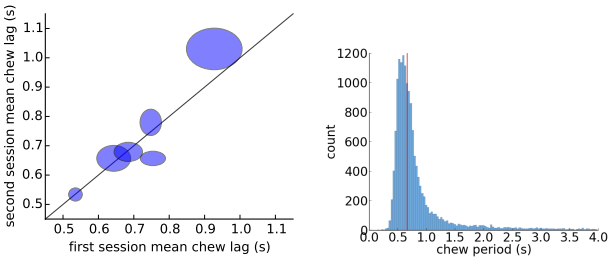
5.1 Data Characteristics

We recorded a total of 71.53 hours of data. We aimed for 6 hours per session (mean 5.96, sd 0.22) to capture multiple meals and snacks. On average each participant had 2.5 meals (sd 0.52), with a meal defined as clusters of intakes with a maximum separation of 15min. Meal duration ranged from 1.65 to 58.66 minutes (mean 15.05, sd 13.51). We identified a total of 1492 food and 329 drink intakes. The same food may be consumed in multiple ways, as with a milkshake having a continuous drink-type intake when consumed with a straw, but a discrete food-type intake when consumed with a spoon. The data contain 17,080 chews and 1422 swallows. Since swallowing is not always visible, this number is lower than the true number of swallows that occurred. The number of events per participant and session are shown in figure 6. Participants spent a significant amount of meal time talking (avg. 33%, range 8-79%), and ate a variety of foods including soft tacos, popcorn, pizza, fruits, soup, chips, and ice cream. Multiple foods were often consumed in a meal (e.g. sushi and dumplings) and in a bite (e.g. chips with guacamole). Examples are shown in figure 5.

While prior work found that time between chews depended

	Chew		Intake		Drink		Meal	
Sensor	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
AGRL (LOPO)	65	21	37	20	47	11	83	90
AGRL	72	23	42	21	49	12	88	87
AGR	72	23	**41**	**24**	47	10	86	88
AGL	72	23	16	9	41	10	81	91
GRL	-	-	42	21	49	12	95	43
ARL	72	25	**42**	**24**	46	10	86	93
AG	72	23	11	9	30	6	73	93
RL	-	-	**42**	**24**	46	10	94	45
A	**72**	**26**	-	-	-	-	**92**	**89**
G	-	-	11	9	30	6	73	48
R	-	-	38	24	**50**	**14**	93	42
L	-	-	8	7	30	7	61	36

Table 2: System performance averaged over all sessions. Results are for LOSO unless indicated. Letters denote combinations of A (audio), G (Glass), R (right) and L (left) watch sensors. For each task, the result with the highest F-measure is bolded.



(a) Time between chews (b) Chew lag histogram

Figure 7: Chew timing detail: (a) each participant is represented by an ellipse with width in each direction showing the sd for that axis; and (b) combined time between chews.

on food type [2], we found this was consistent within an individual while varying between people. Participants ate very different foods in their two sessions (e.g. salad one day, pasta another), so this is not explained by a preference for similar foods. Figure 7a summarizes this result, comparing the mean time between chews for each user's two data collection sessions using the 2nd and 3rd quartiles. We plot the mean for session one against that of session two for each user.

5.2 Sensor Comparison

Results of event and meal detection for various sensor combinations are given in table 2. Unexpectedly, given the range of foods and amount of talking during meals, audio has the highest precision (92%) and recall (89%) on meals. Our annotation of individual chews and noise cancellation procedure may have enabled this as while some chews are missed, those recovered have high precision. This is also why performance degrades in some cases with more sensing modalities. All combinations with audio have F1-scores above 80% on meal detection, with the best being all sensors (AGRL) or audio plus the two watches (ARL). LOSO outperformed LOPO along every dimension, except recall of meals, possibly due to the individuality of chew timing.

Without audio, individual chews are not detected (only periods of chewing) and meal recall suffers, as it is based on intakes. High precision on meal recognition is still achieved with any combination of motion sensors that includes the right watch (i.e. the dominant wrist for all participants).

Performance varied across individuals. Using all sensors and LOSO, 6 sessions had 99% precision in detecting meals, while one participant had two sessions with low precision (69% and 35%). This participant was an outlier, potentially due to food choices (chicken wings, scrambled eggs, ice cream). Chew precision in most sessions was over 80% but that same participant's sessions were 22% and 27%.

Chew performance was best for crisp foods like salad, rather than soft ones like bread and rice. However, in these cases enough chews were still detected to recover the meal period from audio. In one case (banana with almond butter) chew detection failed entirely, yet intakes were inferred and the meal period was detected with every combination that included at least two motion sensors. This demonstrates the added value of motion sensors. In contrast, chewing of strawberries was easily detected, suggesting such variation within food groups may enable recognition of food type.

Figure 8 shows an example of false negatives. Only two of five intakes were inferred. The three others show clear peaks in the intake probability time series but did not meet the threshold (dashed line). The earbud initially came loose from the participant's ear, and chews were only detected after the participant readjusted the earbud (blue arrow). More work is needed to better capture patterns in the event probabilities to recover event timings.

The annotations were largely adequate, however two drink events were found that were missed in the annotation due to poor visibility in the video, and periods of eating yogurt that contained candy pieces were frequently annotated as mouthing while chewing was detected and may be a more appropriate description of the activity.

5.3 Dominant Versus Non-dominant Hand

Wrist-based sensors are often placed on the dominant hand, while watches are normally worn on the non-dominant, leading to a need to understand how much each hand is used for eating. By chance, all participants were right handed. Food preparation (stirring, cutting, opening containers and so on) was done mainly with the dominant hand (55.28% of total prep time) or both hands (41.23%), with the non-dominant hand rarely used alone (3.43%). In contrast, delivery, bringing food or drink to the mouth was most often done with the dominant hand (69.3%), with the non-dominant hand

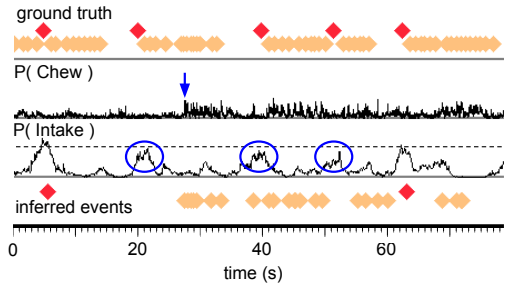


Figure 8: Example of false negatives (circled in blue). Intakes are shown in red and chews in yellow.

used much more frequently by itself than was the case during preparation (17.57%). Both hands were used simultaneously (such as when lifting a sandwich) 12.82% of the time. The difference between the use of the non-dominant hand during the two activities was statistically significant ($p = 0.0280$ with a paired t-test). We found the non-dominant hand was used for intake when the dominant hand was in use for an activity requiring more precision (using a mobile phone) or strength (opening a bottle), or when eating a more formal meal (e.g. cutting steak with knife in dominant hand, without switching hands for intake).

6. CONCLUSIONS AND FUTURE WORK

Identifying eating activities in an automated way is a core problem for promoting health and treating chronic disease. Despite much work on this problem, little has been known about how sensors compare since they have been evaluated separately. We present the first comparison of acoustic and motion sensors (head and both wrists), with evaluation against finely annotated video ground truth. Using a second microphone and noise cancellation, audio sensing achieves high levels of precision and recall in detecting meals (92% and 89%), but misses some individual chews. Motion sensors can fill this gap by identifying food and drink intake, but the lower specificity of this signal requires more advanced modeling, such as modeling the dependency between events and their temporal sequencing. Data are available at http://www.skleinberg.org/data.html

7. ACKNOWLEDGMENTS

This work was supported in part by NSF Award #1347119 (SK, CM, MZ) and NIH Award Number R01LM011826 (YH).

8. REFERENCES

[1] O. Amft, M. Kusserow, and G. Troster. Bite weight prediction from acoustic recognition of chewing. *IEEE Trans Biomed Eng*, 56(6):1663–1672, 2009.

[2] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster. Analysis of Chewing Sounds for Dietary Monitoring. In *UbiComp*, 2005.

[3] O. Amft and G. Tröster. Methods for detection and classification of normal swallowing from muscle activation and sound. In *Pervasive Health*, 2006.

[4] O. Amft and G. Tröster. On-body sensing solutions for automatic dietary monitoring. *IEEE Pervasive Computing*, 8(2):62–70, 2009.

[5] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall. Recognizing Daily Activities with RFID-based Sensors. In *UbiComp*, 2009.

[6] J. Cheng, B. Zhou, K. Kunze, C. C. Rheinländer, S. Wille, N. Wehn, J. Weppner, and P. Lukowicz. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *UbiComp Adjunct*, 2013.

[7] Y. Dong, A. Hoover, J. Scisco, and E. Muth. A new method for measuring meal intake in humans via automated wrist motion tracking. *Applied psychophysiology and biofeedback*, 37(3):205–215, 2012.

[8] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover. Detecting periods of eating during free-living by tracking wrist motion. *IEEE J Biomed Health Inform*, 18(4):1253–1260, 2014.

[9] J. Hagedorn, J. Hailpern, and K. G. Karahalios. Vcode and vdata: Illustrating a new framework for supporting the video annotation workflow. In *AVI*, 2008.

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

[11] S. Ishimaru, K. Kunze, Y. Uema, K. Kise, M. Inami, and K. Tanaka. Smarter eyewear: Using commercial eog glasses for activity recognition. In *UbiComp Adjunct*, 2014.

[12] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos. Platemate: Crowdsourcing nutritional analysis from food photographs. In *UIST*, 2011.

[13] S. Passler and W.-J. Fischer. Acoustical method for objective food intake monitoring using a wearable sensor system. In *Pervasive Health*, 2011.

[14] S. A. Rahman, C. Merck, Y. Huang, and S. Kleinberg. Unintrusive Eating Recognition using Google Glass. In *Pervasive Health*, 2015.

[15] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury. Bodybeat: A mobile system for sensing non-speech body sounds. In *Mobisys*, 2014.

[16] R. Ramos-Garcia, E. Muth, J. Gowdy, and A. Hoover. Improving the recognition of eating gestures using intergesture sequential dependencies. *IEEE J Biomed Health Inform*, 19(3):825–831, 2015.

[17] E. Thomaz, I. Essa, and G. D. Abowd. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *UbiComp*, 2015.

[18] E. Thomaz, A. Parnami, I. Essa, and G. D. Abowd. Feasibility of identifying eating moments from first-person images leveraging human computation. In *SenseCam*, 2013.

[19] E. Thomaz, C. Zhang, I. Essa, and G. D. Abowd. Inferring meal eating activities in real world settings from ambient sounds: A feasibility study. In *IUI*, 2015.

[20] B. Widrow and S. D. Stearns. *Adaptive signal processing*. Prentice-Hall, Englewood Cliffs, NJ, 1985.

[21] B. Zhou, J. Cheng, M. Sundholm, A. Reiss, W. Huang, O. Amft, and P. Lukowicz. Smart table surface: A novel approach to pervasive dining monitoring. In *PerCom*, 2015.