AUTOMATIC EATING DETECTION IN REAL-WORLD SETTINGS WITH COMMODITY SENSING

A Thesis
Presented to
The Academic Faculty

by

Edison Thomaz, Jr.

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
May 2016

AUTOMATIC EATING DETECTION IN REAL-WORLD SETTINGS WITH COMMODITY SENSING

Approved by:

Gregory D. Abowd, Advisor
School of Interactive Computing
Georgia Institute of Technology

Irfan Essa, Advisor
School of Interactive Computing
Georgia Institute of Technology

Thad Starner
School of Interactive Computing
Georgia Institute of Technology

Elizabeth Mynatt
School of Interactive Computing
Georgia Institute of Technology

Tanzeem Choudhury
Department of Information Science
Cornell University

David Conroy
College of Health and Human Development
Penn State University

Date Approved: January 5th 2016

To Vovó and my family,

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my family, and especially Andrea, Lucas and Stella for their unwavering love, patience and support throughout my Ph.D. journey. Having little ones to care for and hug at the end of each day was incredibly comforting, kept me grounded and put everything else in perspective. I love them all and always will.

This dissertation would not have come to life without the support of my advisors. I am eternally grateful to Gregory Abowd and Irfan Essa for their friendship and endless supply of guidance, attention, knowledge, and encouragement over the last five years. Having the privilege to work with two outstanding mentors was the equivalent of winning the academic lottery. I hope to be able to pay their generosity forward by teaching and inspiring students and scientists for years to come.

I had the opportunity to form an absolute all-star committee for my dissertation. Thad Starner, Beth Mynatt, Tanzeem Choudhury and David Conroy are all leaders in their respective fields who were fantastic at providing thoughtful feedback and pushing me towards scientific research excellence. I owe much of the quality of this dissertation to them.

A welcoming, stimulating, and collaborative academic environment is key to success and this is exactly what I found at Georgia Tech and GVU. I am thankful to the faculty and staff in the School of Interactive Computing and everyone in the Ubicomp and i-team labs, including Caleb, Yi, Aman, Gabriel, Cheng, Hwajung, Ivan, Daniel, Vinay, Steve, Unaiza, Julia and so many other friends and collaborators. I would also like to thank Rosa Arriaga, Agata Rozga and Thomas Ploetz for their continuous advisement and support.

Finally, studying human behaviors takes time and resources. I would like to thank the dozens of participants of my user studies and the hard working IRB team led by Melanie Clark that facilitated my experiments. Additionally, I would also like to acknowledge the companies and organizations that financially supported my research, including Humana, Intel and the National Institutes of Health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Motivated by challenges and opportunities in nutritional epidemiology and food journaling, ubiquitous computing researchers have proposed numerous techniques for automated dietary monitoring (ADM) over the years. Although progress has been made, a truly practical system that can automatically recognize what people eat in real-world settings remains elusive. This dissertation addresses the problem of ADM by focusing on practical eating moment detection. Eating detection is a foundational element of ADM since automatically recognizing when a person is eating is required before identifying what and how much is being consumed. Additionally, eating detection can serve as the basis for new types of dietary self-monitoring practices such as semi-automated food journaling.

In this thesis, I show that everyday eating moments such as breakfast, lunch, and dinner can be automatically detected in real-world settings by opportunistically leveraging sensors in practical, off-the-shelf wearable devices. I refer to this instrumentation approach as "commodity sensing". The work covered by this thesis encompasses a series of experiments I conducted with a total of 106 participants where I explored a variety of sensing modalities for automatic eating moment detection. The modalities studied include first-person images taken with wearable cameras, ambient sounds, and on-body inertial sensors. I discuss the extent to which first-person images reflecting everyday experiences can be used to identify eating moments using two approaches: human computation, and by employing a combination of state-of-the-art machine learning and computer vision techniques. Furthermore, I also describe privacy challenges that arise with first-person photographs. Next, I present results showing how certain sounds associated with eating can be recognized and used to infer eating activities. Finally, I elaborate on findings from three studies focused on the use of on-body inertial sensors (head and wrists) to recognize eating moments both in a semi-controlled laboratory setting and in real-world conditions. I conclude by relating findings and insights to practical applications, and highlighting opportunities for future work.

CHAPTER I

INTRODUCTION AND MOTIVATION

"Without proper diet, medicine is of no use. With proper diet, medicine is of no need."
–ancient Ayurvedic proverb

1.1 Monitoring Eating Activity

Eating is one of the most fundamental human activities. Satisfying the hunger urge is essential for survival and sharing a meal has been one of the most enduring social practices for thousands of years [47]. Because of the important role eating plays in our lives, it has been extensively studied. Anthropologists have investigated the relationship of eating behavior to culture and society and have claimed that learning how food is eaten is to learn how a society functions [40]. Food consumption has been shown to be tied to rituals, symbols, belief systems and identities [95].

For several decades health researchers have also been deeply interested in studying eating habits and its impact on human health. It is now understood that good nutrition is vital for optimal growth and development, and prevention of disease [46, 72]. Dietary intake has been widely examined as it relates to cardiovascular disease, hypertension, obesity, diabetes, cancer, osteoporosis and many other medical conditions [17].

Despite the importance of eating as an activity, keeping track of what, where, how much and with whom people eat remains a significant challenge, particularly in naturalistic settings. As described by Jacobs:

> *"A full characterization of a person's diet would consist of a large number of discrete pieces of information. There are thousands of foods, prepared in myriad ways, and eaten in various amounts and combinations. Even a single food such as a carrot or an onion presents a challenge, as there are many varieties and genetic variations; growing*

conditions are influential in food composition. The timing and context of eating, as well as the number of meals eaten, may all contribute to metabolism of food" [58].

Nutritional epidemiologists have typically relied on validated dietary assessment instruments driven by self-reported data including food frequency questionnaires and meal recalls [141]. Unfortunately, these instruments suffer from several limitations, ranging from biases to memory recollection issues [58, 93].

Over the last 15 years, a large body of research has aimed at automating the task of food intake monitoring. This research has been possible thanks to advances in mobile, wearable and sensing technologies. Despite significant progress, most proposed systems have required individuals to wear specialized devices such as neck collars for swallow detection [6], or microphones inside the ear canal to detect chewing [80]. These form-factor requirements have severely limited the immediate practicality of automated food intake monitoring in health research.

Another factor that has hampered progress in automated nutrition monitoring has been the way the recognition problem has been represented. There are at least two key technical challenges in building a fully automated food intake monitoring system: (1) recognizing *when* an individual is performing an eating activity, and then (2) inferring *what* and *how much* the individual eats. Historically, methods devised to automate dietary tracking have largely ignored the distinction between these challenges. Failing to acknowledge the many facets of the problem, I would argue, is an important reason why previous automated dietary assessment research efforts did not meet expectations in terms of practical applicability and deployment.

My work is particularly concerned with the challenge of eating detection. The aim of this dissertation is to defend the thesis that it is possible to automatically detect eating moments, such as breakfast, lunch, dinner and snacks by opportunistically leveraging sensors embedded in practical, off-the-shelf wearable devices that have become increasingly popular with the general population. I call this sensing approach "*commodity sensing*".

Figure 1: Wearable cameras, smartphones, activity trackers and smartwatches are examples of popular consumer electronic devices that can be used for sensing everyday human activities.

1.2 Commodity Sensing

When implementing activity recognition systems, researchers have traditionally designed custom devices or employed dedicated data collection methods. Although this approach makes it possible to experiment with new types of sensing technologies and conduct user studies with participants, it is not practical when capturing sensor data in real-world settings and in longitudinal studies. Participants are often unwilling to wear custom devices and sensors for days or weeks at a time no matter how motivated they are about the research goal. Fortunately, over the last few years, we have seen the emergence of a wide range of wearable devices such as smart watches, activity trackers, and wearable cameras, with extensive computation and sensing capabilities. Many of these devices have been widely embraced by individuals already while some are becoming increasingly more popular.

By leveraging the capabilities of these devices, it is possible to capture sensor data

referring to people's everyday activities over the long term without requiring the utilization of any other custom device. This strategy, which is central to this thesis, is referred to as "*commodity sensing*". It is derived from a concept introduced by Lukowicz et al. called opportunistic sensing [82]. These researchers argued that to recognize complex activities in real-world environments and realize universal context awareness, it is imperative that systems take advantage of sensors and devices that just happen to be in the environment. This approach is in contrast to systems that require specialized sensors deployed for specific applications. The idea of readily-available sensors and devices that can be leveraged by applications when necessary is well aligned with Weiser's vision of omnipresent computing, where a sensing layer is seamlessly overlayed in the physical environment, becoming in effect invisible [135].

While opportunistic sensing has been proposed as a comprehensive approach that encompasses automatic discovery, configuration and even on-demand exchange of signals and algorithms as appropriate, commodity sensing is a simpler construct. I define it as the utilization of sensors in off-the-shelf devices which happen to be ubiquitous and omnipresent by virtue of having been adopted by the general population. Despite its conceptual simplicity if compared to opportunistic sensing, it satisfies the requirement of supporting activity recognition systems without custom or application-specific sensors, a key ingredient for scalability.

1.3 Application Domains

Eating is a universal, multi-faceted activity that is strongly tied to the everyday human experience. It is also one of the most important health determinants. Therefore, it is no surprise that understanding people's eating habits and its consequences at the individual and societal levels is incredibly valuable. The approaches I explore in this dissertation are motivated by applications of eating detection along four inter-related dimensions: population health, nutritional epidemiology, dietary self-monitoring, and patient and elder care.

1.3.1 Population Health

Between 2006 and 2008, on a typical day, Americans age 15 or older dedicated 67 minutes to eating and drinking activities. An additional 23.5 minutes were spent in secondary eating, that is, eating while engaged in another activity. Men spent more time eating and drinking (69 minutes) than women (65 minutes). Individuals who snack all day long, reporting at least 4.5 hours of primary or secondary eating and drinking, are called constant grazers [66].

These types of findings emerge out of population health studies. Population health is defined as "*the health outcomes of a group of individuals, including the distribution of such outcomes within the group*" [70]. In the context of eating activities, population health means gathering information on eating patterns to understand key issues related to nutrition and health. As described by Andrews et al, "*A better understanding of American eating patterns, including the context of their food consumption, can improve programs and policies targeted at reducing obesity and improving overall nutrition and, more generally, inform consumer education, food assistance programs, and product development/ marketing.*"

In the U.S., a large portion of population health data, including nutritional habits, is gathered through the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey focused on health-related risk factors[1]. Although it is the largest continuously conducted survey system in the world, it is based on self-reported data, which is prone to inaccuracies and biases. The ability to collect objective data about people's activities at scale would be a breakthrough in population health. In fact, it would represent a development as significant as the development of the BRFSS itself. In the context of nutrition, this vision is becoming a reality and the work outlined in this document demonstrates technical approaches for realizing this future.

1.3.2 Nutritional Epidemiology

Nutritional epidemiological findings form the backbone of public health policy, nutritional guidelines and even agricultural subsidies. One of the key reasons why health researchers are interested in how people eat is to elucidate the mapping between dietary habits and

[1] http://www.cdc.gov/brfss/about/index.htm

disease. Finding out what people eat and the broader context of eating activities has been of interest to epidemiologists for many decades. For example, cohort studies have shown an inverse association between adherence to the Mediterranean diet and cardiovascular risk [39]. Moreover, researchers are beginning to explore the impact of time-restricted diets on human health [49].

One disease, cancer, has been extensively studied alongside dietary impact. Cancer is considered a chronic disease of the genome that may be influenced at many stages by nutritional and metabolic factors. It is estimated that up to 80% of colon, breast, and prostate cancer cases and one third of all cancer cases may be influenced by diet and associated lifestyle factors [46]. Unfortunately, there is much we do not know about the mechanisms underlying lifestyle-disease relationships. This challenge is exemplified by an analysis by Schoenfeld and Ioannidis titled "Is everything we eat associated with cancer? A systematic cookbook review" [119]. In this work, the relative risk of different foods with respect to cancer are analyzed based on a review of published work. While some studies show that items such as eggs and coffee act in cancer prevention, others claim they are cancer risk factors.

The source of divergent findings in nutritional epidemiology stems in large part from the use of flawed measurement tools. As with population health studies, epidemiological research is also based on self-reported data. In fact, population health data sometimes drive explorations in the space of nutrition. And as previously mentioned, self-report based instruments have many flaws, which are detailed in Chapter 2. Recently, there has been a strong sentiment in the health research community that more resources need to be allocated towards the development of more objective and precise measures, which includes the ability to detect eating activities [32, 94].

1.3.3 Dietary Self-Monitoring

The need for improved dietary tracking is also shared by individuals interested in meeting health goals. Recently, health concerns linked to dietary behaviors such as obesity and

diabetes have fueled demand for dietary self-monitoring, one of the most effective methods for weight control [20, 12]. It is characterized by systematic self-observation, periodic measurement and recording of target behaviors with the goal of increasing self-awareness [64, 140]. However, adherence to dietary self-monitoring is poor and generally wanes over time, even with modern smartphone-based systems such as MealSnap[2] and MyFitnessPal[3] [30, 19]. Individuals must remember to log meals and snacks throughout the day, and then manually record eating activities, a tedious and time-consuming task.

Semi-automated food journaling, a technique that hinges on eating detection, is a promising new approach where the food tracking task is split between individuals and an automated system, thus reducing the burden of self-monitoring while keeping individuals involved in the process. In essence, it is an attempt to reach a compromise between manual and automatic nutrition tracking. The key aspect of the approach is the splitting of the food journaling task into two sub-tasks. The first sub-task, which is completed first and is fully automated, centers on detecting when an eating activity is taking place. Examples of eating activities include breakfast, lunch, dinner and snacking. Upon eating detection, the other sub-task is triggered, requiring individuals to manually provide some information pertaining to what was consumed.

A practical instantiation of this approach starts with an on-body sensor that automatically infers when a person is eating. In my work, this has been done in a number of ways with varying degrees of accuracy: by recognizing eating moments from images taken with wearable cameras [129], presented in Chapter 3; by identifying acoustic signatures associated with eating from environmental sounds [130], presented in Chapter 4; and by recognizing food intake gestures with inertial sensors [127], presented in Chapter 5.

Once eating is taking place and the eating activity has been detected, several courses of action could be pursued to prompt the individual for more information. In one scenario, the individual's smart-watch could softly vibrate to remind and nudge the individual to add an entry to a food log. In some cases it might be undesirable or not socially acceptable

[2]http://www.mealsnap.com
[3]http://www.myfitnesspal.com

to document a meal while it is taking place. Instead, the individual could receive a text message later in the day as a reminder to log at an opportune time in the near future.

1.3.4 Patient and Elder Care

While healthy eating habits are important in the prevention of a large number of medical illnesses for the general population, it is particularly critical for the elderly and individuals with chronic diseases [89]. For older adults, poor nutritional intake is linked to increased morbidity and mortality due to energy deficiencies, low-body mass, cognitive decline, and many other factors [92, 110, 88]. In particular, a deeper understanding of the impact of poor dietary habits on individuals 75 years old and older is needed.

Poor dietary habits are also common for individuals with chronic diseases such as mental illnesses [87]. For example, individuals with schizophrenia have a 20% shorter life expectancy than the population at large [97] and are vulnerable to lifestyle diseases including diabetes, coronary heart disease, and hypertension. To make matters worse, some of the medications used to treat schizophrenia have been associated with weight gain, the onset of diabetes and other problems. The combined effect of these risks suggests that physical activity and nutrition monitoring as a means of health promotion would be beneficial to this population.

1.4 Thesis and Research Contributions

In this dissertation, I defend the thesis that **everyday eating moments can be automatically detected in real-world settings by opportunistically leveraging sensors in practical, off-the-shelf wearable devices**. I define eating moments as eating activities such as breakfast, lunch, and dinner.

My work encompasses a wide span of research contributions around the study and evaluation of different sensing modalities for eating moment detection, starting with first-person point-of-view photographs taken with wearable cameras, progressing to ambient sound sensing and concluding with a detailed examination of inertial sensing. The specific research contributions of this work are enumerated below:

1.4.1 Activity Recognition with Human Computation

Human computation is an approach that combines humans and computers to solve large-scale problems that neither can solve alone. Services such as Amazon Mechanical Turk[4] have popularized human computation by creating work marketplaces where any individual can sign up to perform computer-based tasks and be compensated accordingly. Luis von Ahn, the pioneer of human computation, originally demonstrated the value of this technique by applying it towards image recognition problems [132, 133]. Today, human computation is used for a variety of tasks, including gathering labels that can be used to train machine learning classifiers [75, 121].

In chapter 3, I present an approach where human computation is used in the recognition of eating moments from first-person images. This work demonstrates that human computation can be used not only to gather data to train a classifier, but act as the classifier itself. One of the significant challenges of the task is that human computation workers are not simply identifying objects in photographs, but reasoning about whether the individual the photos refer to is in the middle of an eating moment. In other words, the task involves recognizing an activity from a photographic scene, and not just an object. This method was validated with photographs taken by multiple individuals over several days in real-world settings [129].

1.4.2 Privacy-Saliency Matrix

A difficulty of analyzing first-person images taken in real-world settings with human computation is that privacy concerns arise. The reason for this is because human computation workers are unknown individuals, and thus untrustworthy. In truth, anyone can sign up to become a worker, regardless of profile or background. Giving these individuals access to images portraying family, friends, bystanders, personal habits, and locations could be characterized as a threat. Researchers have proposed ethical guidelines for dealing with first-person photos, but these have had limited practical utility [68].

Although computer vision techniques can support the mitigation of privacy concerns in

[4] http://www.mturk.com

first-person photographs (e.g., face detection), it is not possible to guarantee that they can eliminate all privacy threats altogether. Additionally, a privacy-preserving approach might end up flagging photos whose content is relevant, such as evidence of eating activity. To understand and quantify the balance between privacy threat mitigation and the need to preserve certain photos due to the presence of relevant content, I developed a framework called the Privacy-Saliency Matrix [128]. I evaluated the framework by testing four computational techniques on a dataset of images collected in real-world settings. The techniques were face detection, image cropping, location filtering and motion filtering.

1.4.3 Identifying Eating with Computer Vision Techniques

First-person images offer the possibility of capturing a person's activities throughout the day objectively. But computer vision techniques have traditionally fallen short when it comes to recognizing objects and scenes in a photograph without any type of human input. Recently, however, convolutional neural networks (CNN) have recently been used with success on single image classification with a vast number of classes [73] and have been effective at learning hierarchies of features [144].

In light of these promising results, I conducted an experiment and showed that CNN can be used not only to identify images, but also to classify everyday activities into 19 categories including eating, working, driving, biking and cooking [22]. As part of this research, I compiled the largest annotated dataset of first-person images in everyday, real-world settings. In total, more than 40,000 images were collected over a period of 6 months. In the specific context of eating detection, I showed the extent to which a model generalizes to other individuals. Also significantly, I quantified how much a general model's performance improves when re-trained with a small amount of data for one individual. This research was the first to analyze model generalizability and personalization for eating moment classification with data collected in naturalistic settings and over a period of time spanning several months.

1.4.4 Ambient Sounds as Evidence of Eating

Activity recognition researchers have investigated acoustic sensing for a variety of applications, including eating detection [103, 1]. Traditionally, the sensing takes place with on-body

microphones capturing internal body sounds [143, 107]. In a feasibility study in real-world settings with 20 participants, I demonstrate that ambient, environmental sounds recorded around and outside the body, can also be a powerful predictor of eating activity. This is possible because there are several acoustic signatures tied to eating activities that can be recognized, such as the clicking of utensils on bowls, the unwrapping sound of foods coming out of packages and containers, and the background noise of certain eating environments (e.g., restaurant background music or chatter).

1.4.5 Commodity Inertial Sensing for Eating Detection

Body-worn inertial sensors have been extensively employed in activity recognition and eating detection. But until recently, inertial sensors were specialized devices and only instrumented for research studies in laboratory settings [3, 60]. Today, it is possible to explore the problem of eating detection with off-the-shelf devices such as smartwatches, activity tracking devices and wearable computers (i.e., Google Glass). In my work, I pioneered the notion of piggybacking on the sensing capabilities of devices that individuals have already adopted for their own personal use for the purpose of eating detection. In this document, I present results that validate this methodology.

1.4.6 Inferring Eating Moments from Food Intake Gestures

Activity recognition researchers are often interested in identifying when people perform certain gestures with arms and hands. Therefore, gesture spotting techniques have been developed for a number of applications over the years including eating recognition, when it is desirable to identify food intake gestures. However, for automatic eating moment detection, pinpointing intake gestures is not enough; the goal is to recognize an eating moment such as breakfast, lunch and dinner. Using a density-based unsupervised learning technique, I present an approach where eating moments can be inferred from predicted food intake gestures [127].

1.4.7 Transfer Learning from Lab to Real-World

Model learning is often a difficult and resource-intensive task involving data collection and a significant amount of tuning in terms of features and parameter optimizations. Once a model has been built and meets certain specifications, it is highly desirable to be able to use it in a variety of settings. When the model is reutilized, the accumulated "knowledge" of the model is applied towards solving a different but related problem; this is referred to as transfer learning.

Using a wrist-mounted inertial sensor, I show that transfer learning for an eating detection model trained in the lab and deployed in real-world settings is possible. I built a model with data collected in a laboratory study with 20 participants and validated it in real-world conditions in two studies. The first study involved 7 participants for a period of 1 day, and the second study had one participant collecting data for an entire month. While 1-day studies set in naturalistic conditions are common, in-the-wild studies for a period of several weeks are rare. For both studies, results proved to be highly encouraging.

1.4.8 Impact of One vs. Two-Handed Inertial Sensing

Eating is an activity that often requires the use of both hands. Therefore, instrumenting both hands with inertial sensors might seem like the best approach to detect eating events. However, in practice, individuals only wear one wristwatch or activity tracking band. Additionally, these types of devices are often placed on the non-dominant hand, the one that tends to be used less often while eating.

To understand the impact of dominant versus non-dominant inertial sensing for eating moment detection, I conducted a study with 4 participants in a laboratory setting and compiled preliminary results that address some of these empirical questions.

1.4.9 Activiome: A Platform for Activity Recognition Research

A large portion of my dissertation work focuses on recognizing eating behaviors in real-world settings. This endeavor entails collecting sensor data and estimating ground truth labels "in the wild", a task that is known to be challenging. To facilitate this process, I developed

a platform called Activiome that is composed of a web server backend, a web application, a mobile application, and connectivity support for a range of activity tracking devices and smartwatches.

The mobile application is programmed to function as a wearable camera taking first-person photos at pre-defined intervals throughout the day; participants wear the phone on a lanyard around the neck. The recorded photos capture objective evidence of participants throughout their day. The mobile app also collects sensor data from tracking devices (e.g., inertial sensor data from a smartwatch) and uploads it to the Activiome web server in near real-time. The web application provides an interface that allows study participants to review all collected photographs and annotate them by activity category. By having participants review their own content, privacy concerns are greatly minimized.

Although this platform was designed for the purposes of this dissertation work, it is not tied to the recognition of eating moments in any way. It will be made available to researchers and represents, in my opinion, a compelling tool for activity recognition research in real-world settings.

1.5 Dissertation Overview

Automated eating detection has been a topic of study within the ubiquitous computing research community for many years. Chapter 2 presents an analysis of relevant background material and related work. Chapter 3 focuses on eating detection with wearable cameras. Aside from direct observation, a video recording of an individual's life experience represents one of the best ways to capture the richness of everyday activities. Unfortunately, there are many technical challenges associated with continuous video recording, ranging from battery life and storage to data processing. An alternative to continuous video capture is the shooting of photographs at regularly-spaced time intervals throughout the day. Although not continuous, first person point-of-view photographs also provide a good representation of one's daily activities. This technique is one of the approaches I explored for eating detection. In addition to presenting the wearable system developed for photo capture, I discuss two methods used for inference, one based on human computation and another

based on computer vision and machine learning techniques.

A difficulty that emerges with first-person photographs taken in naturalistic settings is privacy. Pictures taken automatically with on-body cameras might result in the recording of undesirable moments and scenes. To make matters worse, photos taken of computer screens might also capture sensitive information such as computer passwords and credit card numbers. These problems are amplified when these photographs are examined with human computation services like Amazon Mechanical Turk, which are populated by individuals whose real identities are unknown. A detailed discussion of privacy challenges and techniques I employed to better understand and mitigate privacy concerns can also be found in chapter 3.

The second method I investigate for identifying eating moments in first-person photographs uses a combination of metadata and computer vision features. In particular, it leverages a machine learning method, convolutional neural networks (CNN), that has been lately shown to perform well at image recognition tasks. In this case a performance analysis was done for eating detection while also examining the approach's ability to recognize a much larger set of everyday activities in real world settings.

Is it possible to recognize eating moments by the sounds that people make when they eat, such as chewing noises, and the acoustic signature of people's eating environments? This is the question I address in chapter 4. I conducted an experiment with participants in-the-wild where the audio of their everyday experiences was captured with a wrist-mounted recorder. This approach is promising because it relies on a simple, and arguably ubiquitous sensor: a microphone. After presenting results, I discuss future directions for this work.

Chapter 5 is dedicated to the use of single-point inertial sensing in eating detection. The first part of the chapter focuses on how food intake gestures and eating moments can be detected with one wrist-mounted inertial sensor placed on the wrist. It begins with the description of a system I built for this task, ranging from data collection to high-level activity inference. The system was evaluated both in a laboratory setting and also in the wild. One of the highlights of the analysis was the exploration of whether a model trained in the lab can be successfully used in naturalistic conditions. A discussion section follows

the presentations of the study results. The second part of the chapter presents results of a lab study where a head-mounted inertial sensor was used to detect eating activity.

In Chapter 6, I discuss two studies where participants wore wrist-mounted inertial sensors on both wrists. In the first study, I analyze the impact of detecting food intake gestures when both arms are instrumented. In the second study, I present preliminary results of fully personalization eating detection model. The dissertation concludes in Chapter 7 with an overview of the all the work encompassing this dissertation. Notably, It includes insights gained from conducting user experiments and building practical systems in the context of eating detection. The chapter ends with future directions for this line of research.

Finally, in the interest of completeness, I included two appendices in this document. Appendix A describes Activiome, a system I built for sensor and meta data aggregation, visualization and annotation. Activiome was employed in many of the user studies, and played a key role in allowing me to evaluate some of this dissertation's systems and approaches in real-wold settings. Appendix B collects forms and materials that supported my user studies.

1.6 Peer-Reviewed Publications

The work I present in this document explores three sensing modalities for eating moment detection: first-person images, acoustic sensing and inertial sensing. In total, I conducted 2 laboratory studies and 6 in-the-wild studies with 106 participants, which resulted in 5 conference publications:

- "Predicting Daily Activities From Egocentric Images Using Deep Learning". Daniel Castro, Steve Hickson, Vinay Bettadapura, Edison Thomaz, Gregory D. Abowd, Henrik Christensen, Irfan Essa. Proceedings of the International Symposium on Wearable Computers (ISWC) 2015.

- "A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing". Edison Thomaz, Irfan Essa, Gregory D. Abowd. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp) 2015.

- "Inferring Meal Eating Activities in Real World Settings from Ambient Sounds: A Feasibility Study". Edison Thomaz, Cheng Zhang, Irfan Essa, Gregory D. Abowd. Proceedings of ACM Conference on Intelligence User Interfaces (IUI) 2015.

- "Feasibility of Identifying Eating Moments from First-Person Images Leveraging Human Computation". Edison Thomaz, Aman Parnami, Irfan Essa, Gregory D. Abowd. Proceedings of International SenseCam and Pervasive Imaging Conference 2013.

- "Technological Approaches for Addressing Privacy Concerns When Recognizing Eating Behaviors with Wearable Cameras". Edison Thomaz, Aman Parnami, Jonathan Bidwell, Irfan Essa, Gregory D. Abowd. Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp) 2013.

CHAPTER II

BACKGROUND AND RELATED WORK

In this chapter, I present the most relevant work that pertains to tracking eating habits while leveraging sensors and ubiquitous computing technologies. It begins with sensing approaches leveraging environmental resources and sensor-embedded utensils, discusses eating detection using acoustic means and with inertial sensors, and concludes with an analysis of strategies for inferring people's routines. Since building eating detection classifiers require annotated data, obtaining a reliable measure of ground truth is critical. Therefore, the final section is dedicated to methods for annotating human activities in real-world settings.

2.1 Dietary Assessment Methods

It has been more than 70 years since researchers first became interested in understanding the science of measuring dietary intake [118]. Bingham traced the first attempts to perform this measurement outside of a controlled setting to the 1930s and 1940s [15]. Widdoson et al., for instance, presented an examination of English diets using the weighted food record in 1936 [136, 137]. The process involved recording the weight of each item of food and beverage consumed. Soon thereafter, Wiehl, Turner and Reed pioneered interview-based dietary recall and food frequency methods, with the goal of estimating energy intake[138, 131, 139].

Dietary recalls, food records and food frequency questionnaires (FFQ) remain the primary dietary assessment mechanisms in use today. In dietary recall, an interviewer assists an individual in remembering what was eaten over a period of time, typically 24 hours. Dietary records are different in that participants are asked to write down what is consumed shortly after the eating moment. Jacobs observed that in practice people often wait until the end of the day to record what they ate [58]. In this case, the dietary record becomes a self-administered recall.

When it comes to the level of detail that is logged in a dietary recall or record, it varies depending on the end goal. It might be necessary to weigh the food before eating,

collect food samples for chemical analysis, gather detailed information about the foods (e.g., brand, whether it was eaten with condiments or paired with a beverage, etc.), specific timing information and more.

With food frequency questionnaires (FFQ), which come in many flavors in terms of the number and specificity of questions, the objective is to obtain more general dietary knowledge and habits. For instance, a question in a FFQ might be "How often do you eat pizza, and if so, how often and how many slices do you typically consume?". More detailed questions might be asked, such as "When you drink milk, is it typically fat free, 1%, or whole-milk?" or "Do you prefer white or whole-wheat bread?".

Despite the use of these self-report methods for several decades, observations have shown that people tend to forget items that were eaten, underestimate large portion sizes, overestimate small ones and, in general, be susceptible to a large variety of errors and biases, some of which are shown in Table 1. Recently it has became possible to measure the accuracy of dietary recalls, records and FFQs thanks to the doubly-labeled water technique [79]. Findings confirmed the weaknesses of these assessment methods.

Table 1: Some sources of error or bias in dietary intake estimates from FFQ. A complete list can be found in Coulston and Boushey [17].

Type of Error	**Reason for Error**
Memory	Unable to recall food consumption
Frequency judgment	Cognitive difficulty in providing information (low-literacy)
Question comprehension	Not able to understand which foods are being talked about
Response errors	Mistankely codes incorrect frequency
Social desirability bias	Misrepresent dietary intake to please investigators

In light of these limitations, researchers have begun to question the validity of the data collected by these methods. Archer et al. focused on the National Health and Nutrition Examination Survey (NHANES), stating that "*methodological limitations compromise the validity of U.S. nutritional surveillance data and the empirical foundation for formulating*

Figure 2: The top image shows a version of a nutrition monitoring necklace by Kalantarian et al. [63]. The bottom image depicts a sensor embedded in a tooth for oral activity recognition [78].

dietary guidelines and public health policies" [8]. Dhurandhar et al. believe traditional instruments like dietary recalls and records should not be used at all for energy intake (EI) and physical activity energy expenditure (PAEE) assessment. In their own words, "*...it is time to move from the common view that self-reports of EI and PAEE are imperfect, but nevertheless deserving of use, to a view commensurate with the evidence that self-reports of EI and PAEE are so poor that they are wholly unacceptable for scientific research on EI and PAEE.*" [32].

2.2 Automated Dietary Monitoring

Dietary assessment challenges and limitations have fueled interest in automated processes starting in the 1980s. At the time, researchers tried to detect chews and swallows using oral sensors in order to measure the palatability and satiating value of foods [124]. The desire to automate nutrition monitoring persists to this day, with researchers developing and evaluating practical and experimental systems spanning many different types of techniques. Cheng-Yuan Li et al. recently revisited oral activity detection with a wearable

system, shown in Figure 2 [78]. Sounds from the users mouth and on-body sensing approaches have also been suggested as ways to detect when and what individuals are eating [4]. Other approaches have explored a variety of sensing modalities and computational methods, including the use of crowdsourcing techniques [99], shopping receipts [86], and neckband wearables [26, 63]. A key finding from this body of research is that no single sensor can capture all dimensions of eating behavior.

2.2.1 Sensing with Objects, Places and Artifacts

Several techniques for tracking, recording, and even modifying nutrition patterns have been put forth through the instrumentation of everyday home environments and objects. In 2006, Chang et al. presented a dining table that could track various eating scenarios, from afternoon teas to Chinese-style dinners [23]. It was designed with the goal of tracking what and how much individuals ate. Automated food logging was achieved using two layers of sensing surface, one with RFID sensing and another with weighting cells. The RFID surface identified tabletop objects and tracked their location while the weighting surface helped recognize food transfers between containers. The authors claimed that their approach allowed them to track entire food movement paths and validated their system with results showing accuracy in the range of 80%. Recently, Zhou et al. also experimented with a smart table surface [147] (Figure 3). It performed favorably when evaluated with 5 subjects across 40 meals, recognizing the spotting and recognition of food intake related actions such as cutting, scooping and stirring.

Macaw built a system for automatically photographing meal eating activities from a camera mounted on a dining room ceiling light [83] (Figure 3) . According to the author, one of the difficulties of relying on individuals for documenting meals is that people often forget to take pictures while eating. In previous work exploring barriers to food journaling, colleagues and I identified that remembering to track foods is indeed a problem [30]. In this implementation, Maekawa configured the camera to turn on and off with the ceiling light under the assumption that eating always takes place with the lights on.

Figure 3: The top row images show the ceiling light camera by Maekawa and a typical photo taken with the camera [83]. The bottom-left image is showing the smart table surface by Zhou et al. [147]. The HAPIfork is shown in the bottom-right picture.

Kadomura et al. explored a sensor-embedded fork around an interactive mobile application with the goal of monitoring and possibly modifying a child's eating behavior. [61]. The fork was instrumented with motion sensors for detecting changes in eating behavior state and a single-pixel color sensor to determine food colors. By tracking different foods by color, the system attempted to encourage children to eat a variety of food items. Another instrumented utensil is the HAPIfork [1], shown in Figure 3. It was designed to sense and control the pace of eating, delivering vibrations when it identifies that the person is eating too fast. Fluid intake tracking through specialized and instrumented cups has also been a focal point of researchers. Lester et al. developed a method that uses optical, ion selective electrical pH, and conductivity sensors to sense and classify liquid in a cup. Accuracies of up to 79% were obtained for 68 different types of drinks [77].

[1] https://www.hapi.com/product/hapifork

Despite the promise and constant improvement of approaches around instrumented objects and locations, their practicality is severely limited by the fact they are often not portable enough, if at all. From the point of view of activity tracking, the key advantage of wearable sensors is that individuals are free to move amongst different locations and eat anywhere since they are carrying the system with them at all times. In other words, they are not restricted to the infrastructure in the built environment or having to remember to always carry a sensing object with them (e.g., a "smart" fork).

2.2.2 Acoustic Sensing for Eating Detection

Sound is a contextually-rich source of information that can be easily recorded using one of the simplest and most ubiquitous sensors; a microphone. Hence, a large body of work at the intersection of acoustic sensing and activity recognition has emerged over the last decade. Clarkson and Pentland were able to infer environmental and situational context through audio classification many years before smartphones and wearable sensors became widely popular [28, 29]. Soon thereafter, Stager examined a low-power implementation of a sound recognition system and evaluated the tradeoff between classification parameters (e.g., features, feature selectors) and performance [123]. Ward et al. explored the use of on-body microphones and accelerometers to recognize activities involved in an assembly task in a wood workshop, where hand and machine tools are typically used interchangeably [134]. And framed in the context of Activities of Daily Living (ADL) recognition, Chen examined bathroom sounds recorded with a microphone and obtained 84% accuracy when identifying six activities including taking a shower and hand washing [25].

Mobile phones have been explored extensively in auditory scene recognition and analysis. Rossi et al. implemented a system called AmbientSense as a mobile phone application for recognizing user context from ambient sounds [114]. AmbientSense operated in two modes; it could perform audio recognition on the mobile device in real-time or by sending audio features for classification to a server. Using Mel-frequency cepstral coefficients (MFCC) as features, the system was able to identify 23 ambient sound classes including "beach", "forest", "phone ring", and "street" with 58% accuracy. Given the difficulty of

collecting training data for a system like AmbientSense, Rossi et al. also examined the use of web-collected audio data to build a context recognition system [113]. Through the compilation of 114 hours of audio data from the FreeSound database, they obtained practical recognition rates between 50% and 80% for the same 23 classes studied in AmbientSense. Another implementation of an audio-centric activity recognition system on mobile devices is SoundSense [81]. It combines supervised and unsupervised machine learning techniques with a hierarchical classifier to perform varying levels of audio classification, and discover novel sound events specific to individual users. Its coarse category classifier had an accuracy range between 80% and 90% for three types of sounds (ambient, sound and speech).

There is no question mobile phones are ubiquitous, but most of the time they are inside pockets and pursues. This raises an important question; how practical are mobile phone-based activity recognition systems that rely on environmental sounds? Franke et al. partially addressed this point by showing that a mobile phone can successfully infer ambient sounds even when inside clothing [42]. An alternative to mobile phones are wearable devices that are directly placed on the body; these are not limited by the constraint of being inside some other object (e.g., purse or bag). A system that illustrates this approach is BodyScope, a wearable acoustic sensor attached to the user's neck [143]. Its goal was to explore how accurately a large number of activities could be recognized with a single acoustic sensor. The system was able to recognize twelve activities at 79.5% F-measure accuracy in a lab study and four activities (eating, drinking, speaking, and laughing) in an in-the-wild study at 71.5% F-measure accuracy.

One of the most explored applications of sound-based activity recognition with wearable devices has been dietary intake tracking. Sazonov et al. proposed a system for monitoring swallowing and chewing through the combination of a piezoelectric strain gauge positioned below the ear and a small microphone located over the laryngopharynx [116, 84]. Passler investigated the problem of intake monitoring using microphones in the outer ear canal [104]. A promising and comprehensive approach to automated dietary monitoring was proposed by Amft et al. [5]. It involves having individuals wear sensors in the wrists, head and neck and automatically detect food intake gestures, chewing, and swallowing from accelerometer and

acoustic sensor data. Liu et al. developed a food logging application based on the capture of audio and first-person point-of-view images [80]. The system processes all incoming sounds in real time through a head-mounted microphone and a classifier identifies when chewing is taking place, prompting a wearable camera to capture a video of the eating activity. The authors validated the technical feasibility of their method with a small user study.

Bai et al. developed a wearable computer called eButton with the goal of "evaluating the human lifestyle" [11]. Similar to the SenseCam in terms of functionality and capabilities, eButton was designed to be worn like a chest button instead of around the neck with a lanyard. It houses a CPU, storage components, a wide-angle digital camera module, and an array of sensors in a small form factor. Sun et al. suggested the use of the eButton for objective dietary assessment [126], and Zhang et al. implemented an activity recognition system from video segments captured with the eButton [145].

Recently, Liu et al. developed a food logging application based on the capture of audio and first-person point-of-view images [80]. The system processes all incoming sounds in real time through a head-mounted microphone and a classifier identifies when chewing is taking place, prompting a wearable camera to capture a video of the eating activity. The authors validated the technical feasibility of their method with a small user study, so it is unclear how their system performs in real world settings.

2.2.3 Recognizing Eating with On-Body Inertial Sensing

The widespread availability of small wearable accelerometers and gyroscopes has opened up a new avenue for detecting eating activities through *on-body inertial sensing* [6]. Amft et al. have detected eating gestures with a measurement system comprised of five inertial sensors placed on the body (wrists, upper arms and on the upper torso) [5, 60]. Recognition of four gesture types resulted in recall of 79% and precision of 73% in a study with four participants.

Zhang et al. investigated an approach for eating and drinking gesture recognition using a kinematic model of human forearm movements [146]. With accelerometers located on the wrists, features were extracted using an extended Kalman filter, and classification was done

with a Hierarchical Temporal Memory network. Results showed a "successful rate" around 87% for repetitive eating activities. The authors were not explicit about which performance measures they used in their evaluation (i.e., what they meant by "successful rate"), how many participants took part in the study, and whether the results reflected person-dependent or person-independent findings. Additionally, the study focused exclusively on eating and drinking activities so the system's ability to differentiate between eating and drinking versus other activities is unclear.

Also with wrist-based inertial sensors, Kim et al. proposed an approach for recognizing "Asian-style" eating activities and food types by estimating 29 discrete sub-actions such as "Taking chopsticks", "Stirring", and "Putting in mouth" [69]. In a feasibility study with 4 subjects, the authors obtained an average F-measure of 21% for discriminating all sub-actions. The system performed better when considering only certain classes of sub-actions, but hand actions could not be identified at all. These measurements led the authors to state that the 29 pre-defined sub-actions may not be suitable for the recognition of meals.

Recently Dong at al. put forth a method for detecting eating moments in real-world settings based on a wrist-motion energy heuristic [35, 34]. They evaluated it with participants wearing a smartphone on the wrist. The phone collected continuous inertial sensor data reflecting people's arm and hand gestures. One possible concern with this setup is that it is unclear how much the placement and weight of the phone influenced intake gesture movements. Precision and recall measurements were in the range of 20% and 80% respectively. Finally, Amft et al. proposed a system for spotting drinking gestures with one wrist-worn acceleration sensor. Based on a study with six users that resulted in 560 drinking instances, the system performed remarkably well, with average of 84% recall and 94% precision[2]. In this work, the authors also attempted to recognize container type and fluid level, and achieved recognition rates over 70% in both cases.

2.2.4 Identifying Daily Routines and Patterns

Discovering daily routines in human behavior from sensor data has been an active area of research. With a dataset of 46 days of GPS sensor data collected from 30 volunteer

subjects, Biagioni and Krumm demonstrated an algorithm that uses location traces to assess the similarity of a person's days [14]. Blanke and Schiele explored the recognition of daily routines through low-level activity spotting, with precision and recall results in the range of 80% to 90% [16]. Other proposed techniques for human activity discovery have included non-parametric approaches [125], and topic modeling [55].

One of the most comprehensive analysis of human behaviors in naturalistic settings was done by Eagle and Pentland [36]. By collecting data from 100 mobile phones over a 9-month period, they were able to recognize social patterns in daily user activity, infer relationships, identify socially significant locations, and model organizational rhythms. Their work was based on a formulation for identifying structure in routines called *eigenbehaviors* [37]. By examining a weighted sum of an individual's eigenbehaviors, the researchers were able to predict behaviors with up to 79% accuracy. This approach also made it possible to calculate similarities between groups of individuals in terms of their everyday routines. With data collected in the wild over 100 days, Clarkson also presented an approach for the discovery and prediction of daily patterns from sensor signals [29].

In 2014, Chen et al. showed that it is possible to leverage these kinds of daily routines and patterns in service of eating detection. The researchers built an eating prediction model based on location histories and behavior data such as user activity (e.g., stationary, walking, running, driving, cycling), sleep duration and sociability (i.e., the number of independent conversations and their durations). Since the study was conducted on a college campus, predictions were compared against actual food purchases logged through student identification cards. The system was able to predict eating with 74% accuracy [24].

2.2.5 Techniques for Estimating Ground Truth in Real World Settings

One element of eating detection that has been prevalent throughout the years and across different sensing modalities is the use of statistical machine learning techniques for inference and modeling. The fundamental challenge with this approach is that obtaining labeled ground truth examples for real-world activity recognition requires interrupting individuals as they are performing everyday tasks. This is often achieved by constantly prompting

Figure 4: Example first-person point-of-view photos taken with a wearable camera.

people to self-report what they are doing on a journal or logbook [57]. A popular self-report technique is the experience sampling method (ESM), first suggested by Larson and Csikszentmihalyi [74].

Over the years, a variety of strategies have been created for including individuals in the process of activity labeling. A variation of ESM called Context-Aware Experience Sampling (CAES) attempts to reduce the frequency of interruptions by prompting individuals to log their activities only when a significant change in context occurs, such as a sudden change in heartbeat rate [111]. An alternative to ESM is the Day Reconstruction Method (DRM), which helps participants re-construct activities and experiences of the preceding day using a procedure designed to mitigate recall biases [62]. Despite the shortcomings of self-reporting, numerous mobile and web-based systems have been developed to facilitate this process in the last few years such as AndWellness [51], MyExperience [43], and Ohmage [108].

Recently, the idea of directly observing individuals from egocentric cameras for overall lifestyle evaluation has been gaining appeal [33]. In this approach, individuals wear cameras that take first-person point-of-view photographs at regular intervals throughout the day

(e.g., every 30 seconds), documenting one's everyday activities including dietary intake, as shown in Figure 4 [101]. One of the first cameras used in this context was SenseCam, a lightweight digital camera worn around the neck that passively captures first-person point of view images and sensor readings at regular intervals throughout the day [53]. One of the most unique characteristics of SenseCam is that it doesn't require wearers to perform any action, since images are taken completely automatically. Since its introduction, the SenseCam device has enabled a wide range of applications. Kelly et al. investigated the potential of SenseCam to infer travel research, and in particular evaluate modes and volumes of active versus sedentary travel [67]. Byrne et al. explored SenseCam as a collector of observational data and found it to be complementary to traditional methods. Among other findings, they reported that the passive nature of SenseCam is particularly well-suited for task observations since it doesn't intrude into people's environment [21]. In the domain of eating activities, the capture and categorization of environmental and social context was explored by Gemming et al. [44].

Image-Diet Day is another system that automatically captures first-person images [7]. Fourteen participants wore the mobile phone-based device during eating periods for three days and the captured images assisted participants in completing a 24-hour recall procedure. In terms of their value for recall, the images were regarded as helpful, but participants did report technical and perception issues wearing the phone camera device.

Although first-person point-of-view images offer a viable alternative to direct observation, a fundamental problem remains. All captured images must be manually coded for lifestyle indicators, and even with supporting tools such as ImageScape [109], the process tends to be tedious and time-consuming. To address this challenge, human computation-based methods around the Amazon's Mechanical Turk infrastructure have been developed, such as Platemate [99]. Crowdsourcing has matured in the last five years to become an attractive approach to researchers in many fields, including nutritional analysis and activity recognition. One emerging way to leverage human computation is to use the crowd not only to annotate images and other forms of media but also to provide training data for machine learning classifiers [121].

CHAPTER III

FIRST-PERSON POINT-OF-VIEW PHOTOGRAPHS

With the advent of small wearable cameras such as the Narrative Clip[1] and the GoPro[2], it has become possible to capture everyday experiences with unprecedented richness in detail. A head or chest-mounted camera is configured to take first-person point-of-view (FPPOV) images automatically throughout the day (e.g. every 30 seconds), and the resulting snapshots capture people performing a wide range of everyday activities, from socializing with friends to having meals with family members.

I explored this approach to infer eating moments in naturalistic settings. This technique is particularly promising because it is completely passive; it does not require individuals to do any extra work. Moreover, the capture images reflect people's eating activities and the surrounding context of those activities truthfully. Despite these advantages, one of the major challenges of this technique is that only a small portion of the total number of automatically-captured images from a wearable camera depicts an eating activity. Therefore, before these images can be examined from an nutritional perspective or saved in a food journal, it is necessary to devise a mechanism to sift through thousands of FPPOV images and discover the ones that pertain to eating. The sheer volume of images generated per day makes it impractical to annotate them manually.

I pursued two research directions to identify eating moments with FPPOV images, one using human computation and one combining computer vision and convolutional neural network techniques. Important privacy considerations arise out of the use of FPPOV images, and these issues are also discussed in this chapter.

[1] http://www.getnarrative.com
[2] http://www.gopro.com

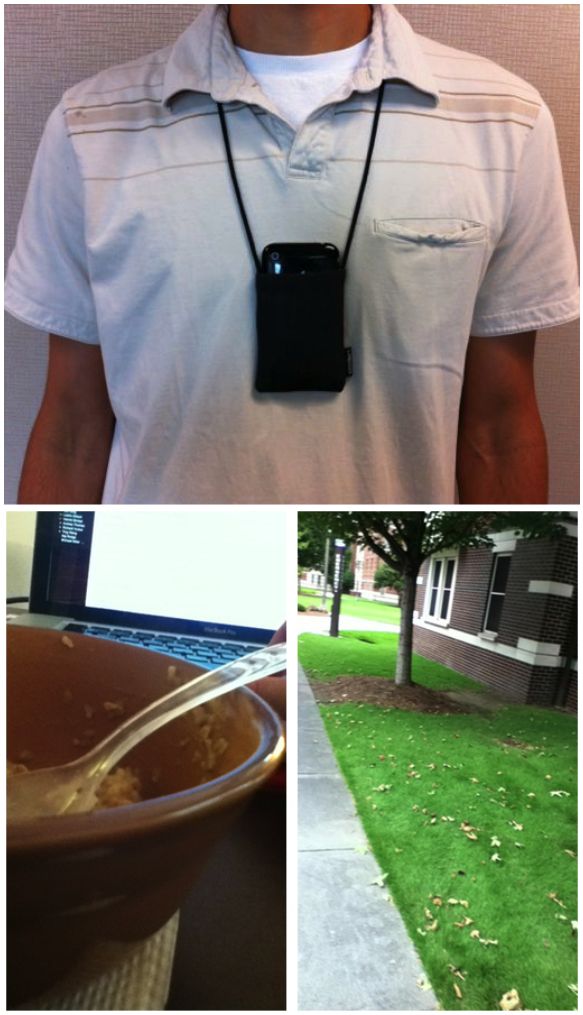


Figure 5: I implemented an application on a standard mobile phone to passively capture first-person point-of-view images (FPPOV).

3.1 Collecting First-Person Point-of-View Photos

Before FPPOV images can be analyzed for evidence of eating activity, they must be captured. Researchers have used a number of tools for taking FPPOV images in the past, such as SenseCam [53]. Because I was interested in using mobile phones for this task and ran into performance issues when testing existing applications that promise this functionality, I chose to implement a mobile photo capture application targeting the iOS platform. The additional motivation for having my own implementation was that it could serve as a platform for future experiments and prototypes.

The application, called WAID, took photos automatically every 30 seconds using either

the front or back phone camera. People wear the phone as a pendant around the neck with its back-camera facing forward, as shown in Figure 5. All images were saved on the device itself and were immediately visible through the built-in "Photos" application. The application was optimized to conserve battery life; it didn't provide any user interface when running, except for displaying a gray logo on an otherwise completely black background. The only feedback people got from the application was the system's default image snapshot sound effect whenever a picture was taken. If people chose to suppress or minimize this sound effect, they could mute the phone or turn down the volume.

By turning off certain features of the phone, such as Wifi and Bluetooth, and setting the brightness of the screen to its lowest level, WAID ran for up to 10 hours on a single battery charge for different iPhone models (iPhone 4S, iPhone 4 and iPhone 3G), all running the most recent version of the iOS supported by each device (iOS 6.0.1 and iOS 4.3) at the time of the study.

3.2 Method I: Human Computation

Human computation has emerged as a viable way to tackle problems that can't be presently solved by computers. Although human computation has been validated as a technique for image labeling [132, 133, 122, 115], identifying health-specific activities in photos through crowdsourcing techniques has not been explored with much depth. I devised a method where human computation was applied towards identifying eating moments in FPPOV images. The method is comprised of 3 stages, where images are first collected and filtered for privacy protection, formatted into temporal groups, and finally presented to a group of trusted and human computation workers as part of an evaluation (Figure 6).

3.2.1 Excluding Images for Privacy Protection

First-person point-of-view images captured every 30-seconds might depict a day in an individual's life with an unprecedented level of detail. But there is a good chance that these images also reflect aspects of one's life that might be embarrassing or compromising. Therefore, an important step of the method was the exclusion of images that posed a privacy threat to the individuals wearing the camera and to individuals who, knowingly or not,

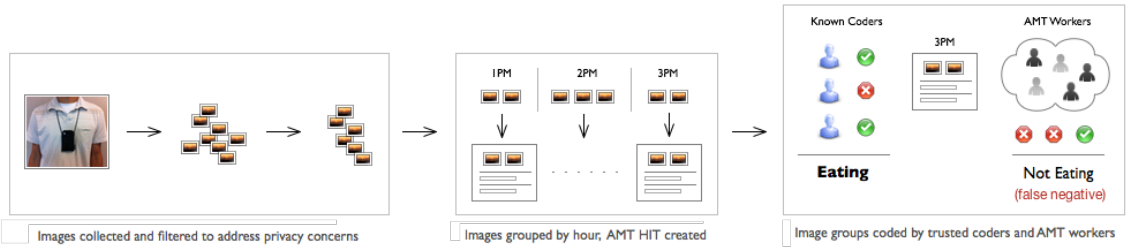


Figure 6: The pipeline for recognizing eating moments from first-person point-of-view (FPPOV) images leveraging human computation and evaluating the performance of the system. It is comprised of 3 stages, where images are first collected and filtered for privacy protection, formatted into temporal groups as a web-based user interface, and finally presented to a group of trusted and human computation workers.

were captured in the images.

After transferring all images from the phone to a computer, participants were given the opportunity to review all photos taken by their device and delete any images they did not like to share. Additionally, I reviewed the images and deleted any photo that either captured other individuals, or that could reveal sensitive information of the individual who wore the camera. These privacy measures were established by the Institutional Review Board (IRB) at Georgia Tech.

3.2.2 Coding Images in AMT

In this method, the task of recognizing eating moments in thousands of FPPOV images was performed by human computation coders. The human computation platform I chose to use was Amazon's Mechanical Turk (AMT). It is described as a "a marketplace for work that requires human intelligence". It exists on the premise that a large number of tasks that computers aren't good at, such as identifying objects in photographs, can be easily carried out by people. Through Mechanical Turk, companies or individuals (called "requesters"), post well-defined tasks ("human intelligence tasks" or HITs) that are matched with, and executed by "workers". Workers signup on the site to perform HITs in exchange for rewards, which range from $0.01 to $1. Requesters can specify a number of parameters for HITs, such as the number of workers that are allowed to perform the task, the qualification of those workers, and the reward amount for tasks completed. Workers are paid only after

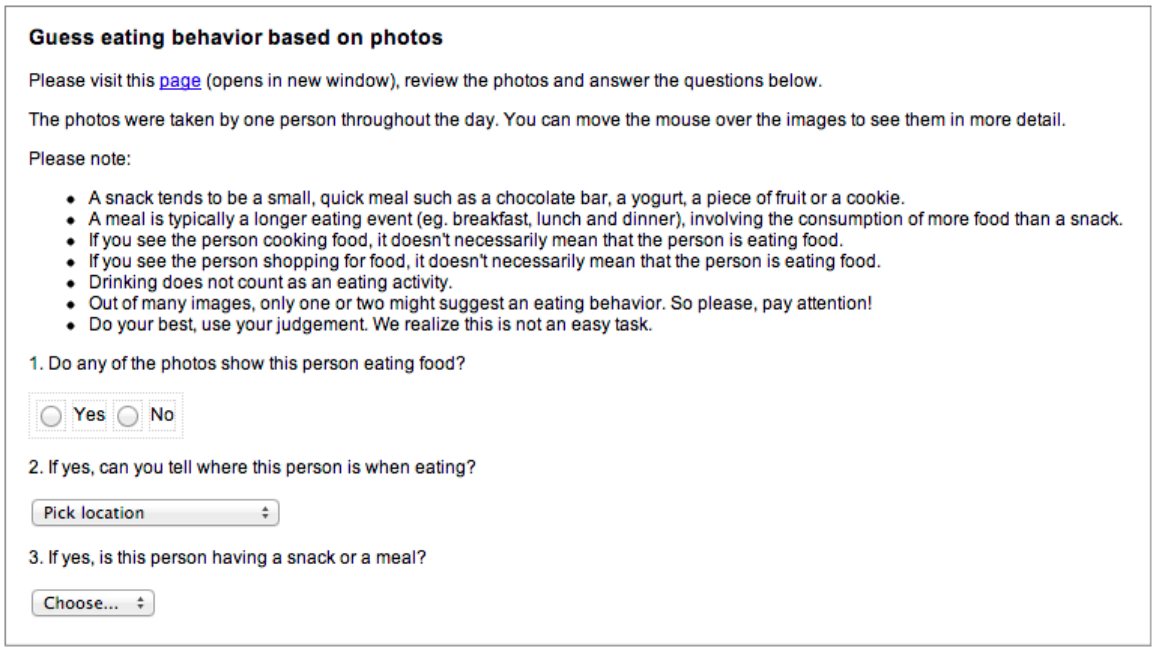


Figure 7: The layout of the human intelligence task (HIT) posted at Amazon's Mechanical Turk for the study. I included a set of guidelines to help workers perform the task successfully. The choices for meal location were: at home, at work or school, at a fast-food restaurant, at a sit- down restaurant, in the car, somewhere else. The choices for meal type were: meal, snack.

HITs have been completed and approved by requesters.

3.2.3 Generating and Assigning HITs

I created a human-intelligence task on AMT that asked workers to examine a group of photos and indicate whether any photo showed an eating activity. If positive, I asked workers for additional information (i.e. meal location and type). The images were grouped by hour, and formatted into a web-based mosaic-like interface (Figure 8). In order to fit a large number of images on the grid, the images were reduced in size, which lowered the amount of activity detail that could be seen. To counter the effect of smaller image sizes, I implemented a script that enlarged the photo underneath the cursor, on hover.

Once a HIT was created, it had to be assigned to workers. On AMT, it is possible to specify exactly how workers are matched to tasks. To improve the validity of workers' results, I assigned each HIT to three unique workers, and coalesced their votes on each question by taking a majority vote. With this method, depending on the number of workers

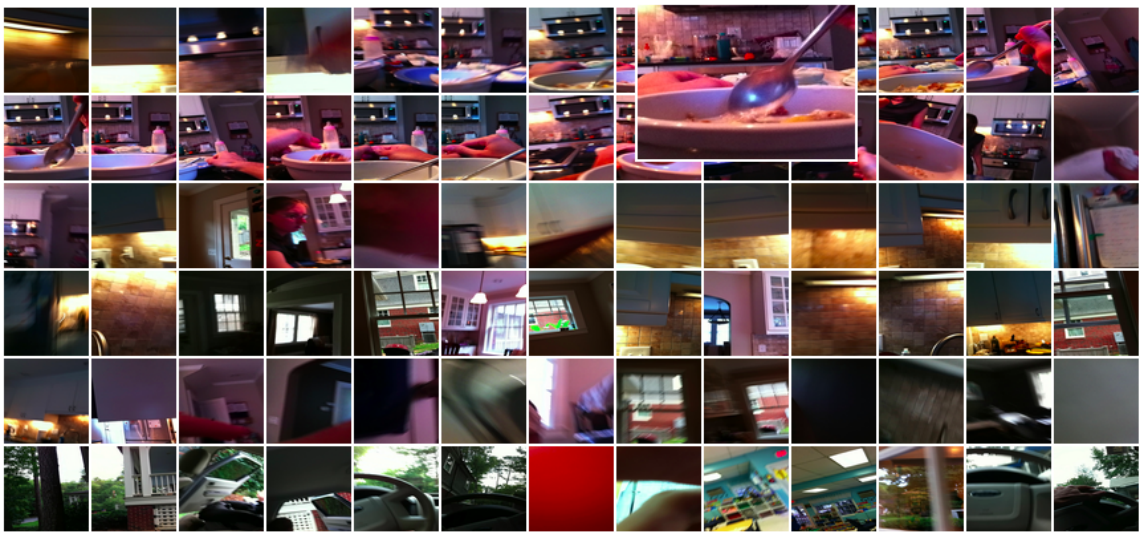


Figure 8: The image grid interface was designed to help Amazon's Mechanical Turk workers browse a large number of photos more efficiently. Hovering the cursor over an images expanded it such that it can be examined in more detail, as shown in the middle of the first row.

and valid answers per question (e.g. for meal location), there was a possibility that a majority vote might not be obtained. If and when this condition occurred, the HIT was resubmitted until a majority vote was reached. A completed HIT assignment consisted of the answers to the three questions, the photo group examined, and an identifier for the workers who completed the task.

3.2.4 Deployment and Evaluation

I conducted a feasibility study with a non-random convenience sample of participants (n = 5) over 3 days. The only requirement for being in the study was familiarity with the basic operations of a smartphone device. There were 3 females and 2 males, and they ranged in age from 23 to 35 years old and were either graduate students or research scientists at Georgia Tech. With the exception of one married participant, all other participants were single and either lived alone or with roommates.

Participants were provided with a smartphone preloaded with the WAID application and were instructed to wear the device as much as possible, ideally from the moment they woke up until when they went to sleep. It would be impractical for subjects to wear the

smartphone continuously for hours at a time, so I gave them complete latitude to turn the device off, or take it off if they wanted to or needed to. Due to limited battery life, participants were asked to recharge the device every night.

On average, each participant provided 3,509 photos. The image exclusion step where participants reviewed their own images lasted about 15 minutes per participant and led to the removal of up to 200 images. Going through the remaining images and deleting photos that included secondary participants took us at least 45 minutes per subject, and resulted in the deletion of an additional 700 images on average. in total, forty nine instances of eating activity were recorded in the photos.

One important aspect of Mechanical Turk is that it makes it possible to select workers based on a number of qualifications tied to cost. For instance, it costs more to recruit so-called "master" workers because they have been identified by Amazon as proficient at categorization tasks. I hypothesized that performance would be significantly affected by workers' level of qualifications. Therefore, I created identical categorization tasks for masters and regular workers and compared their results. I rewarded all workers $0.15 per assignment and, for regular workers, I indicated that they should have a HIT approval rate greater than 98%.

3.2.5 Results

To assess the performance of Mechanical Turk workers at recognizing eating activities in photos, I had to estimate a measure of ground truth for the image data collected. This was accomplished by having three trusted coders answer the three questions posed in the AMT tasks for each one of the photo groups. The trusted coders used the same web-based interface to examine and browse images as the AMT workers, and their inter-rater reliability was calculated to be 0.65 (Fleiss' kappa).

Table 2 shows how AMT workers performed at identifying eating activities in participants' photos in relation to the estimated ground truth. I calculated recognition accuracy, precision and recall for each participant and across all participants. The results are broken down by worker type to highlight the performance impact of hiring master versus regular

Table 2: Individual and aggregate performance measures showing how well the system was able to identify eating moments from first-person point-of-view (FPPOV) images and human computation. The TP, FP, TN and FN abbreviations refer to true positive, false positive, true negative, and false negative results, respectively.

Participant	Worker Type	TP	FP	TN	FN	Precision	Recall	Accuracy
P1	regular	5	0	33	9	100%	35.71%	80.85%
	master	10	0	33	4	100%	71.42%	91.48%
P2	regular	1	2	59	10	33.34%	9.09%	83.34%
	master	6	1	60	5	85.71%	54.54%	91.67%
P3	regular	1	1	24	7	50%	12.5%	75.75%
	master	5	0	25	3	100%	62.5%	90.90%
P4	regular	2	2	25	8	50%	20%	72.97%
	master	7	3	24	3	70%	70%	83.78%
P5	regular	1	0	28	5	100%	16.67%	85.29%
	master	3	1	27	3	75%	50%	88.23%
All	**regular**	**10**	**5**	**169**	**39**	**66.67%**	**20.4%**	**80.26%**
	master	**31**	**5**	**169**	**18**	**86.11%**	**63.26%**	**89.68%**

workers on AMT. As expected, I saw improved results across all measures when the tasks were assigned to master workers, with overall eating behavior recognition accuracy reaching 89.68% accuracy in the best case scenario. With master workers, overall precision was 86.11% and overall recall was 63.26%.

Inferring meal type and location from FPPOV images is desirable since it might provide additional information that is valuable from a health perspective. However, achieving this from images alone proved to be challenging. Only 19% of meal locations and 24% of meal types were correctly recognized. However, as will be discussed in the next sections, these numbers bear little practical significance since meal location can be often obtained through other means in real-world applications, such as GPS, and meal type is open to interpretation based on time of day and other factors.

3.2.6 Discussion

One of the most salient results from the evaluation was the low overall recall of AMT master workers (63.26%), indicating that they missed many instances of eating activities. Since each photo group contained upwards of 50 images, it is reasonable that a human might miss important details in the images when constrained by time. This was validated when I confirmed that recall was worse when only one or two photos in a group showed participants eating. This often occurred when the food eaten was consumed quickly, within a minute or two, resulting in the eating behavior being captured in only a small number of photos. I found this to be the case with at least one of the participants, who replaced meals with energy bars.

Overall precision (86.11%) was much closer to overall accuracy for master workers. There were many photos where participants were clearly around food items, such as when shopping for food, in line at a cafe or cooking at home. In most of these cases, one could be easily led to believe that eating was also taking place. This was a common source of false positives in the data. A particularly noticeable result was the disparity in the overall precision measure between regular and master workers. The results provide evidence that master workers are indeed better at categorization tasks than regular workers, as Amazon claims. This justifies the higher cost paid to AMT to recruit master workers. Overall, for the reasons mentioned above, recognizing eating moments from FPPOV images proved to be a difficult task. This had a direct effect on precision, recall and explains the relatively low agreement reliability amongst coders.

It is important to note that the results only refer to eating activities that were photographed by participants' cameras. Some eating activities might not have been captured. However, given the perspective from which the photos were captured, the largest majority of participants' eating activities was documented.

3.2.6.1 Meal Location and Type

An individual's location can be often obtained from sensors in mobile phones and other wearable devices. Since there are circumstances when a location sensor is not present or

can't be used (e.g. to preserve battery life), I felt that it would be valuable to understand the extent to which meal location could be inferred from images alone. Upon analysis, I was able to attribute the low recognition rates for meal location to two factors. Firstly, because participants wore a phone as a pendant around the neck, all photos were taken at chest-level, pointing directly forward. When participants were sitting at a table and eating, the field of view of the camera was often obstructed by objects in the scene (e.g. body parts, table, chairs, dish-ware, food). This made it difficult to examine the background of the photos and determine participants' whereabouts. I suspect that this issue would have been greatly minimized with the use of a wide-angle camera lens. Secondly, to protect the privacy of secondary participants, I had to discard all photos showing people other than study participants. More often than not, eating is a social activity, with people congregating around a physical space, therefore many of the deleted photos provided rich contextual information about the meal, such as where it took place and with whom. Without these deleted images, it became significantly harder to determine the physical context of the meal.

In terms of meal type, there is a significant amount of ambiguity in what one refers to as a snack or as a meal. Given a photo of a participant eating an energy bar, it is unclear if it should be categorized as a snack or a meal (e.g. lunch). Time of day could be used to help with this differentiation, but ultimately it is a matter of personal interpretation. This interpretive flexibility was reflected in the results for meal type, since the methodology for measuring performance was based on response agreement amongst trusted coders and AMT workers.

3.2.6.2 Multiple Eating Activities in Photo Group

In the experiment, each photo group included all images captured within a 1 hour interval per participant. I never saw more than one eating activity per photo group. If there had been multiple eating activities within the hour, the exact activity AMT workers based their answers on would have been ambiguous. Spreading all captured photos into more photo groups, each with an interval window of 15 or 30 minutes, would have been a way to address this issue. As previously mentioned, this is an area I plan to explore in future work since

I expect that a shorter window might also improve the workers' ability to recognize eating moments.

3.2.6.3 Mechanical Turk Worker Qualifications

Although the human computation approach offers advantages if compared to a computer vision technique in estimating eating moments from real-world everyday images, it has limitations of its own. One of the characteristics of the method is that people with a wide range of skills and backgrounds are the ones ultimately accepting and completing tasks [112]. Consequently, there is a certain level of variability and non-determinism in human computation that might be unacceptable in certain applications. A set of workers recruited now is always likely to be different from another set of workers recruited just five minutes later.

For a price, it is possible to benefit from a categorization scheme set by Amazon where certain workers are considered to be more proficient at certain tasks than others. I employed both "categorization masters" and regular workers in the study and could verify that results improved significantly with experts. In my experience, seemingly simple parametric modifications in the HIT can have a dramatic impact on performance. There is a large body of research that corroborates this finding, indicating how various factors, from pricing to qualifications, affect the timeliness and quality of the work performed by workers on Mechanical Turk [71, 90, 122].

3.2.6.4 Annotation Quality Control

The strategy of labeling images through majority vote is the only crowdsourcing quality control used in this work. It is certainly an effective one, as it accounts for occasional human errors and variability in human performance [122]. Hara et al. studied the impact of accuracy in majority group size and determined that performance gains diminish significantly as group size grows beyond 5 AMT workers [48]. For cost reasons, I kept majority vote group size to 3 workers in this feasibility study. One opportunity that exists involves putting in place additional quality measures, such as validation or Find-Fix-Verify [13]. With validation, a set of AMT workers evaluate the classification of images that have already been

labeled.

3.2.7 Privacy Considerations

Privacy arouse as an important element of this work, and privacy-related constraints dictated important aspects of the methodology. One of the challenges of continuous and automatic capture of FPPOV images is that these images may, in some circumstances, pose a privacy concern. Privacy is an area that deserves special attention when dealing with wearable cameras, particularly in public settings. A body of research work has explored this area. Kelly et al. proposed an ethical framework to formalize privacy protection when wearable cameras are used in health behavior research and beyond [68]. People's perceptions of wearable cameras are also very relevant. Nguyen et al. examined how individuals perceive and react to being recorded by a wearable camera in real-life situations [98], and Hoyle et al. studied how individuals manage privacy while capturing lifelong photos with wearable cameras [54].

In the specific case of this study, a large number of photos of non-study participants ended up being captured (Figure 5). These included participants' family members, colleagues, neighbors and many other individuals that participants did not know, such as people who happened to be sharing public transportation with participants, visiting the same coffee shop or eating at the same restaurant. Since these individuals were not in the study, they did not consent to their pictures being taken and reviewed by Amazon Mechanical Turk workers. In order to approve this research, the IRB requested that I delete all such images, which led to the removal of an average of 700 photos per participant (20% of the total). Importantly, the elimination of these photos had a detrimental impact on the performance of the system, since so many photos of eating activities included secondary subjects. In some cases, more than 90% of a set of images depicting an eating activity had to be deleted.

In light of these privacy concerns and methodological restrictions, one might question whether the benefits gained by crowdsourcing the identification of eating moments in FPPOV photos is outweighed by the effort involved in having to manually review and delete

photos for privacy reasons. One way to address this question is by considering whether privacy threats can be automatically and computationally mitigated. If this is indeed possible, then outsourcing photo annotation and identification represents a clear advantage. Through an empirical study with 5 participants over 3 days, I quantified the extent to which four techniques could be used to reduce the privacy-infringing content of images. The techniques were face detection, cropping, location filtering, and motion filtering. To perform this analysis, I developed a framework called *Privacy-Saliency Matrix* for understanding the balance between the eating information in an image and its potential privacy concerns.

3.2.7.1 The Privacy-Saliency Matrix

One of the most constructive ways to address privacy and technology is to make explicit the balance between the positive value proposition of a technology and the negative impact on privacy concerns. Iachello and Abowd portrayed this kind of analysis in Ubicomp as a proportionality argument [56]. For FPPOV imagery, the balance is between whether an image contains information considered to be a privacy concern and if that image contains information salient to a particular task at hand, such as eating. For a set of images, one can visualize this balance in a 2-by-2 matrix, the privacy-saliency matrix (see Figure 9).

The two dimensions of the matrix, as the name suggests, reflect the presence of privacy concerns and content salience. In this work, content salience corresponds to evidence of eating behavior or not. Any FPPOV image taken throughout the daily life of an individual can be uniquely placed into a single quadrant of the matrix. Images in *Quadrant 1 (Q1)* contain evidence of eating and exhibit no privacy concerns. For example, these images show people eating by themselves or the camera only captures evidence of the food in front of a person and not any evidence of others who might be around. Images in *Quadrant 2 (Q2)* contain evidence of eating behavior but also exhibit some information that would be considered a privacy concern. Usually, these photos capture people eating with others who can be identified (e.g. friends or family also eating across the table, or strangers who are nearby). Images in *Quadrant 3 (Q3)* do not reveal any eating behavior, nor do they pose any privacy threat. Sending these images to a human computation service is not a problem

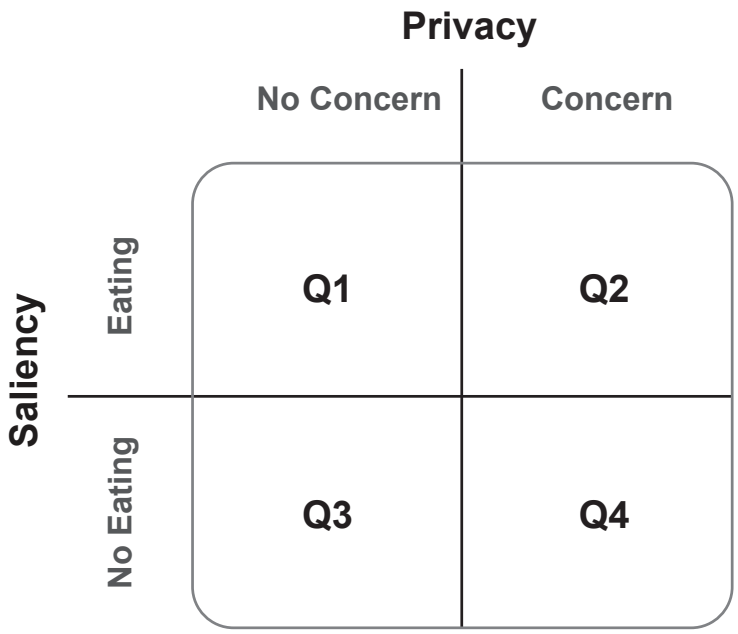


Figure 9: Privacy-saliency matrix provides a framework for studying the balance between privacy concerns and evidence of eating in images.

for privacy reasons, but having too many of them makes the human computation task more expensive and, depending on the information task being presented to the workers, more susceptible to misclassifications. Images in *Quadrant 4 (Q4)* similarly do not reveal any eating behavior, but they do pose a privacy threat.

The privacy-saliency matrix makes it clear how one can understand the opportunities for technology to address the privacy concern for using human computation to identify eating for FPPOV imagery. It also provides a way to quantitatively assess the impact of any given technique or set of techniques. In the context of eating activities, these techniques can be assessed by the following guidelines:

- Keep images in Q1: I would like to keep as many images in Q1 as possible, since they show an eating activity without privacy concerns.
- Eliminate images in Q3 and Q4: Images in Q3 and Q4 can be eliminated completely since they do not depict an eating activity. As I described above, it is important to

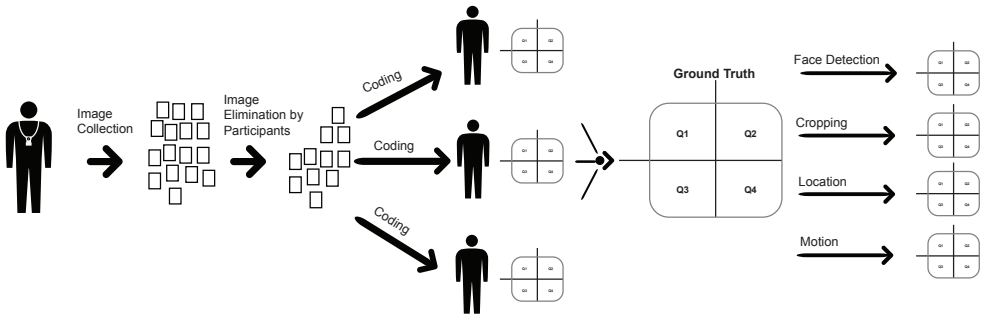


Figure 10: A high-level view of the user study, image coding, and evaluation process. Once participants reviewed and released their images for analysis, the images were coded for evidence of eating behaviors and privacy concerns. Four privacy mitigation techniques were applied on the images separately, and each of the resulting matrices were compared to the privacy-saliency matrix reflecting the images' ground truth.

remove Q4 images because of privacy concerns. Removing images from Q3 has other non-privacy advantages.

- Move images from Q2 to Q1: It would be advantageous to keep the images in Q2, since they also capture an eating activity. The issue with Q2 images is that they contain one or more elements that pose a privacy risk. The ideal scenario would be to purge the visual component that constitutes that privacy risk while keeping the rest of the image, and thus the evidence of eating behavior, intact. In effect, this corresponds to moving the image from Q2 to Q1.

- Eliminate images in Q2: Depending on the approach, it might not be possible to fully suppress the privacy risks of images in Q2 and move them to Q1. A less desirable alternative is to simply delete these images, since they cannot be reviewed by human computation workers. In this case, I want some assurance that the episode of eating evidence by that image removed from Q2 is reflected by an image in Q1 already. For example, if taking pictures every 30 seconds during a meal, it is likely that images within some temporal window of another image might reveal the same eating behavior. This may not hold for shorter duration eating activities, like a snack.

It is important to note that since the ultimate goal is to optimize the multi-variate balance between privacy and content salience for a given application, single-objective measures such as precision and recall are not adequate. The field of multi-objective optimization, also known as Pareto optimization, is concerned with reaching optimality of more than one objective function, and thus comes closest to addressing this privacy-saliency compromise.

Table 3: I recruited 5 participants to be part of the study. A total of 14,422 first-person point-of-view (FPPOV) images were captured and analyzed.

Participant	Age	Gender	# of Images
P1	31	Male	1230
P2	24	Male	5360
P3	21	Male	2528
P4	23	Male	1958
P5	25	Male	3346

3.2.7.2 User Study

To assess how face detection, cropping, location filtering, and motion filtering could be applied to mitigate privacy concerns in FPPOV images, an IRB-approved user study was conducted with graduate student participants (n = 5, all male). The only criteria that I set for participating in the study was that participants had to be familiar with the operation of a smartphone device and be able and willing to recharge the phone every night. Participants were asked to wear the phone for 3 days.

After going over the study protocol, participants were provided with an iPhone 3GS smartphone preloaded with the previously described Waid application (Figure 5). Participants were asked to wear the device as much as possible for the duration of the study; I told them that they could turn off the phone, or take it off, if they did not feel comfortable wearing the device in certain places or situations. All images captured by the mobile application were saved in the phone's default photo library, so participants could review and delete photos whenever they wished. Finally, at the end of the study, participants had the opportunity to review, delete, and get a copy of all captured photos before releasing the images to us. In total the number of FPPOV images collected across all participants was

14422.

3.2.7.3 Method

The methodology for evaluating the privacy mitigating techniques for FPPOV images in the context of eating activity recognition was comprised of two phases. Figure 10 shows the overall workflow. In the first phase, the images were individually coded for evidence of eating behavior and also for privacy threats using the privacy-saliency matrix. The goal was to establish a ground truth baseline for the image set so that I could confidently measure the impact of each automated technique on an image-by-image basis. In the second phase, all images were processed with one of the 4 techniques proposed (i.e. face detection, image cropping, location filtering and motion filtering), and results were compared to the baseline.

The images were reviewed by 3 coders. To reduce the learning effect caused by reviewing FPPOV images in sequential order, I developed a custom image annotation application that arranged images randomly. Coders viewed images on a grid, and tagged them according to privacy and saliency (as defined on a codebook) using keyboard shortcuts for efficiency. The criteria for a privacy concern was the presence of a human head in the image or any body part thereof (e.g. hair, eye, nose). The head could belong to the participant himself or someone else who happened to be photographed. Evidence of eating behavior was determined to be one or more visual cues that indicated that the participant was engaged in an eating activity, such as the presence of silverware, food on a plate, food in hand, others eating nearby, the identification of a restaurant, etc.

The inter-rater agreement amongst coders on the total of 14,422 images was calculated to be 0.73 (Fleiss' kappa), indicating general agreement. In the case of disagreement, I treated privacy and saliency differently. If any one of the three coders thought that there was a privacy concern in the image, the image was considered to have a privacy concern. The overall categorization on the eating dimension was based on a majority vote by the coders.

3.2.7.4 Privacy Mitigation Techniques

In this section, I describe in more detail the four techniques that I implemented with the goal of automating balancing privacy against saliency: face detection, cropping, location filtering and motion filtering.

Face Detection

It is relatively common for faces to be captured in FPPOV images. When this occurs, the identity of the individuals whose faces were recorded is completely revealed, a worst-case scenario in terms of privacy. Ideally, I would like to be able to flag all FPPOV images that contain faces, the images found in Q2 and Q4 in the privacy-saliency matrix, so that they can be either deleted or filtered further. For the analysis in this paper, I simply assume all flagged images are deleted.

I evaluated the performance of two face detection algorithms with respect to its impact in the distribution of images in the privacy-saliency matrix, (1) the one available in the Core Image framework of Mac OS X (10.7 and above), and (2) the set of Haar's cascade classifiers available through the OpenCV library [18]. For the Core Image detector, I implemented an application that leveraged the framework's API. The Haar classifiers consisted of groups of Haar-like features that were learned using Viola and Jones' boosted cascade approach (AdaBoost) for encoding the contrast and spatial relationship of facial features within a window. The Haar Cascade Classifiers were trained on hundreds of face images at similar orientations. Following training, the classifiers were applied to images at multiple scales using a sliding window.

Image Cropping

Recognizing eating behavior in a passive, objective and automated fashion is a hard problem amplified by the fact that eating is often a social activity. Taking photos from a first-person perspective will generally result in images that include other people, such as those sitting across the table or sharing the same environment (e.g. restaurant), a clear privacy risk. This is a typical case where it would be desirable to crop FPPOV images to exclude undesirable elements in the scene (e.g. faces) while retaining the salient content (e.g. evident of eating activity). In the matrix representation previously discussed, this

Figure 11: Several images that contain evidence of eating behavior might pose a privacy concern. By cropping a portion of the image, it is often possible to eliminate privacy issues.

corresponds to the "Move images from Q2 to Q1" scenario.

The cropping technique I considered is perhaps the simplest, and hinges on the observation that when people eat, they usually have a plate or food container right in front of them. Thus, when taking photos from a first-person perspective, the bottom-half region of the images is likely relevant to the evidence of eating (Figure 11). The top-half region of the image is usually where faces are located, and can be discarded.

I implemented an application in Objective-C for Mac OS X that cropped the bottom-half of participants' images, shrinking the image height in half. Image cropping not only has a desirable effect of eliminating privacy risks, it also has an undesirable potential side effect of deleting the evidence of eating behavior. Therefore, to calculate exactly how this technique performed, all cropped images were coded again for evidence of eating behavior and privacy. Like before, 3 coders reviewed and tagged the images, two of whom are authors.

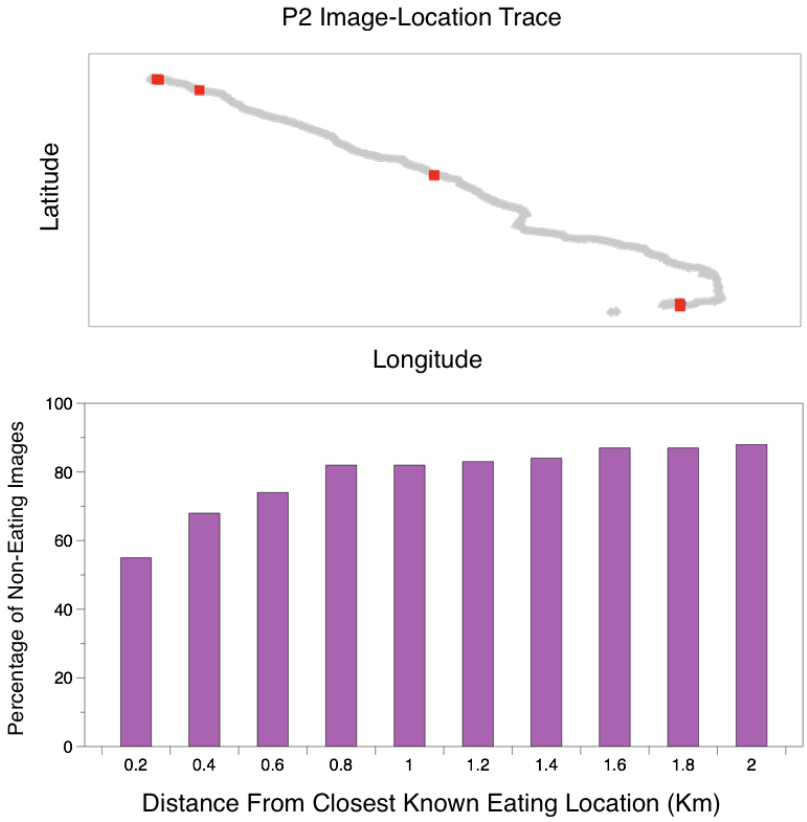


Figure 12: The top chart shows a location trace of one of the participants in the study. Each point in the trace corresponds to a FPPOV image automatically taken with the wearable camera. From the distribution of photos, it is possible to see that photos with evidence of eating activity (red squares) are clustered around a few locations only. The bottom chart illustrates the positive correlation between the number of images depicting non-eating activities and the distance between the location the image was taken and the closest known eating location.

The inter-rater agreement amongst coders in this session was calculated to be 0.8 (Fleiss' kappa).

Location Filtering

The top of Figure 12 shows the geo-location distribution of images for one participant. Red areas of the graph indicate where eating behavior was found in the ground truth coding, and gray areas of the graph are images with no eating behavior. What this plot suggests is that eating activity is localized in space, and this is evident from all of the participants in the study. This empirical evidence reinforces the intuition that routines such as eating can often be inferred from location data [9, 65]. Most eating behaviors can be mapped to

a small number of locations, such as home and work. Naturally, presence in locations such as restaurants and to a lesser degree bars, are highly correlated with the activity of eating as well. The central idea of this technique is to reduce privacy exposure by considering only the photos that maximize the chance of an eating behavior being recorded. In the privacy-saliency matrix, this technique is aligned with the goal of eliminating photos in Q3 and Q4, whose images do not show evidence of eating activity.

This approach leverages the latitude and longitude metadata embedded in each one of the images captured by participants over the duration of the study. To demonstrate the value and performance of this technique, I show how I can eliminate a significant number of images simply on the basis of their geo-spatial physical distance from the closest image that depicts an eating activity. This distance is calculated from the latitude and longitude of two points using the Haversine formula:

$$d = 2r \arcsin(\sqrt{\sin^2(\frac{\Delta\phi}{2}) + \cos(\phi_1)cos(\phi_2)\sin^2(\frac{\Delta\lambda}{2})})$$

In a practical application of location filtering, I would infer the likely locations of eating in two ways. First of all, when collecting location and FPPOV images for a longer period of time, previous work shows that it is possible to infer where home and work are for an individual based on location traces alone [9, 65]. Secondly, discovering that an individual is or was at a restaurant can be easily done by looking up the individual's coordinates on a location database. By combining these two methods, I argue that further locations could be feasibly inferred through a semi-supervised learning approach.

Motion Filtering

It is more likely that people are eating when they are not moving. Based on this insight, I implemented a filter that disregards images when the level of motion of the individual wearing the camera around the time the images were taken exceeds a predefined threshold. The objective was to eliminate images from Q3 and Q4 in the privacy-saliency matrix, which do not convey any information as far as eating activities are concerned.

To collect movement data at the time FPPOV photos were shot, I instrumented the image capture application to continuously log the stream of accelerometer events for as

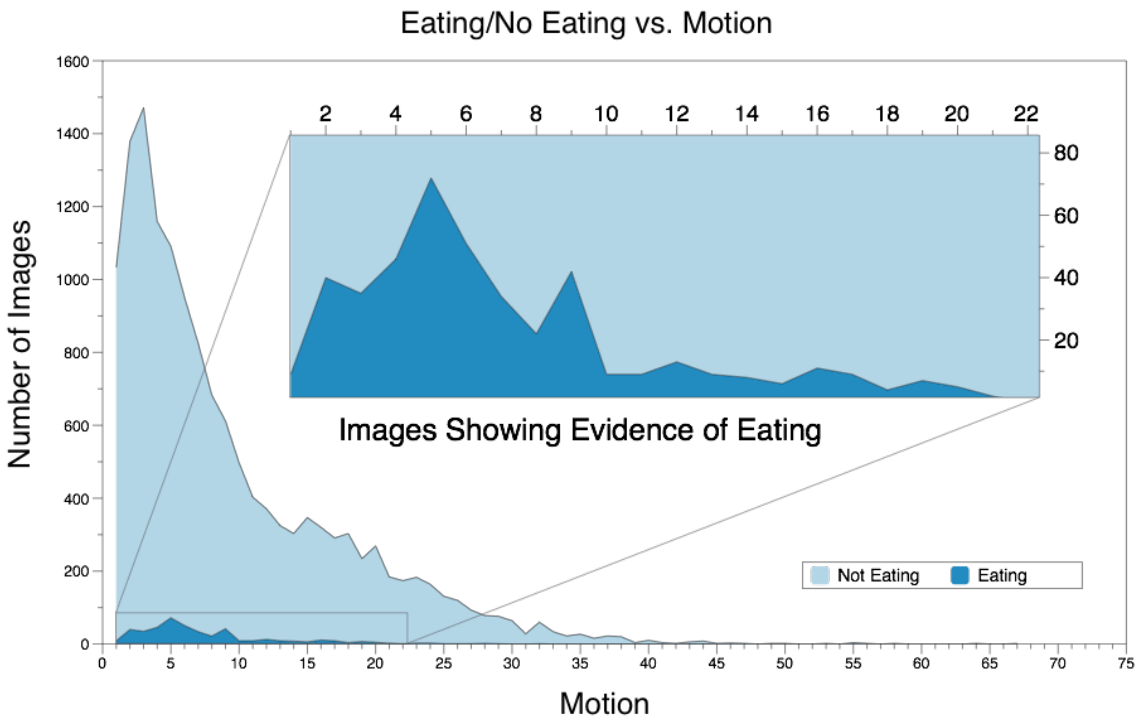


Figure 13: I computed a measure of human motion intensity by leveraging accelerometer data from the mobile phone camera. By adding up the number of images in each quadrant of the privacy-saliency matrix by level of motion, it is possible to see that the most eating activities are contained within a region of motion that range from 1 to 21.

long as the application was running. This enabled us to compile sensor data at the moment images were captured and also several seconds before and after. The level of motion, set for each image, was calculated to be the standard deviation of the composite 3-axis accelerometer data (i.e. x, y, and z) over the minute the photo was taken:

$$M_s = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}|(|x_n| + |y_n| + |z_n|) - \mu|^2} * 100$$

where N is sampling rate times number of seconds in a minute. The normalized score value M_s ranged from 0 to 65 and the threshold for eating activities was set to 8. This was determined empirically, based on the distribution of FPPOV images of the study participants. As shown in Figure 13, the distribution of motion intensity for eating images has a range of 1-21 only, which is distinct from the distribution of motion intensity seen in non-eating images. Additionally I verified that these distributions are significantly different with a Kolmogorov-Smirnov test (p ¡ 0.001).

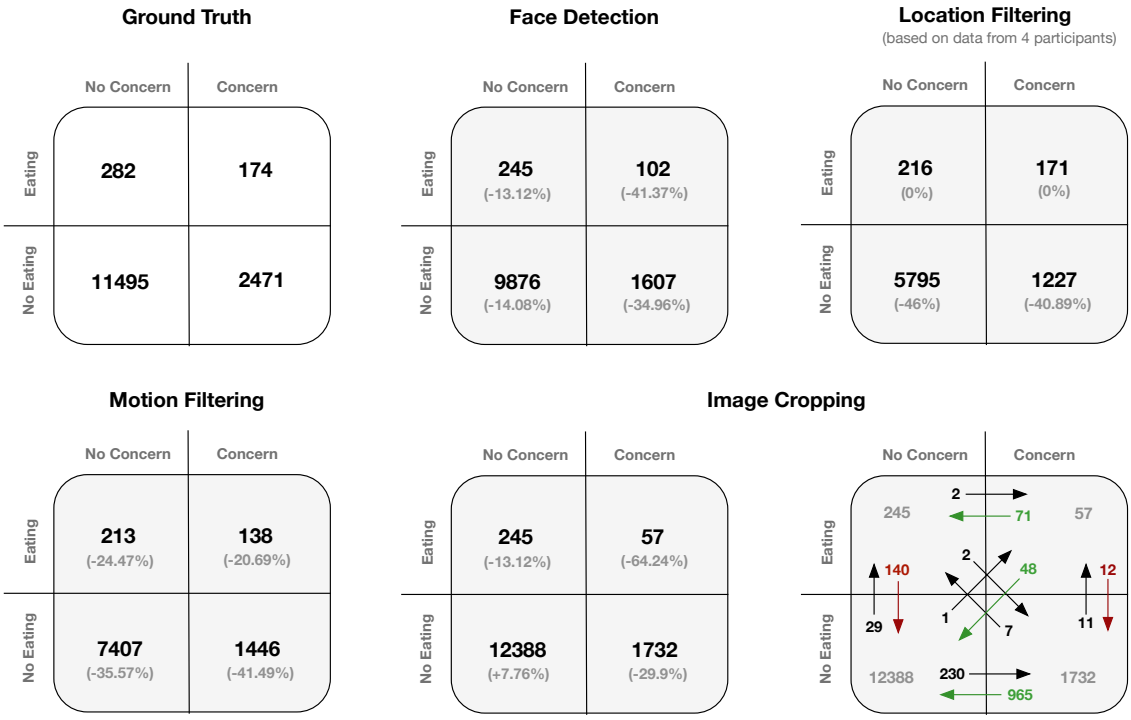


Figure 14: The privacy-saliency matrices showing the coded distribution of images before the application of the privacy mitigation techniques (ground truth) and after. Note that due to corrupted data, the location filter could be applied to images from 4 participants only. The matrix in the bottom-right corner shows how images transitioned from one quadrant to another after cropping. The arrows in green show transitions that I consider "good" (e.g. reduction of images with privacy concerns), while red arrows highlight transitions that I consider "bad" (e.g. removal of evidence of eating behavior).

3.2.7.5 Results

A total of 14,422 images were captured in the 5-person study. Figure 14 shows the ground truth coding in terms of the priv-acy-saliency matrix of the raw FPPOV images. I show the resulting privacy-saliency matrix after each of the four automated techniques are applied to those images.

I ran two face detection algorithms on the participants' images, the one available through the Mac OS X's Core Image framework and the set of of Haar's cascade classifiers available through the OpenCV library. The Haar classifiers outperformed the Core Image detector by an order of magnitude, therefore I am just reporting results with respect to this classifier. As shown in Figure 14, Q2 and Q4 in the privacy-saliency matrix saw the largest decrease

in the number of images, in the range of 35% to 42%. Around 13% to 14% of the images in Q1 and Q3 were flagged for containing faces, which is indicative that the face detection algorithm generated false positives, since the images in these quadrants were previously screened for faces by human coders.

Note that I did not measure the performance of the algorithm with respect to its ability to recognize faces. Instead, by assuming the removal of images from the quadrants when the algorithm detected faces in them, I measured how the application of the algorithm modified the distribution of images in the privacy-saliency matrix. One of the reasons why the face detection method did not perform better is because FPPOV images are often blurry and do not capture faces looking directly at the camera frequently. Nevertheless, as FPPOV images become more popular, it is likely that we will see the development of face detection and other computer vision techniques that are optimized for this type of photography. Also, the privacy criteria that I employed while coding the images was the presence of a human head or any visible part thereof, such as hair, nose, eyes, etc, and not a face. In light of this, many of the images assigned to Q2 and Q4 in the matrix could have never been flagged by face detectors.

With regards to cropping the bottom-half region of the images, it had a positive effect in that it reduced the number of photos with privacy concerns. The number of images in Q2 and Q4 fell around 67% and 30% respectively, as shown in Figure 14. More importantly, the intended effect of having images transition from Q2 to Q1 materialized. Out of 174 images in Q2, 75 moved to Q1. This represents a best case scenario since many images depicting eating activities but compromised by privacy threats had those threats removed with cropping. A smaller but still significant number of images (48) moved from Q2 to Q3. This can be interpreted from two perspectives. On one hand, 48 images that presented privacy issues before no longer did after cropping. This meant that they could be examined by human computation workers without the risk of a privacy violation, for example. On the other hand, the evidence of eating activities in the images is no longer present, so from the point of view of eating behavior recognition, these images do not hold any useful information anymore.

Location filtering proved to be an effective approach for removing images that do not include evidence of eating activity. When considering photos within a radius of 0.2 km of a known eating location, images in Q3 and Q4 fell by 46% and 40.89% respectively. However, as previously discussed, the condition under which these results were obtained is when all eating locations are known. If that is the case, all instances of eating activity are accounted for, and thus there is no loss of images in Q1 and Q2 (no percentage change in the number of images). Unfortunately, all collected location data for one of the participants was corrupted and had to be discarded. This required us to generate ground truth quadrant numbers for the privacy-saliency matrix with 4 participants instead of 5. This is the reason why the numbers in Q1 and Q2 differ from those in the ground-truth privacy-saliency matrix in Figure 14.

Motion filtering performed similarly to location filtering in terms of the reduction of images in Q3 and Q4. Q3 saw a decrease of 35.57% in its images and the number of images in Q4 fell by 41.49%. Because of the need to establish a range in the motion score under which an eating behavior is most likely to occur, it is always the case that some images representing eating activities end up outside of that range and are disregarded. This is why the privacy-saliency matrix for motion filtering shows a decrease in the number of images in Q1 (24.47%) and Q2 (20.69%). Without a doubt, this decrease is undesirable, but it is less pronounced than the loss of images in Q3 and Q4. Overall, the collective loss of images in all quadrants, affecting Q3 and Q4 to a higher degree, underscores the trade-off between capturing activities of interest and mitigating privacy concerns that lies at the core of this paper.

3.2.7.6 Additional Privacy Risks

Though I followed a strict criteria of marking all the images that had any part of the head as a privacy threat, I discovered several other categories of threats while coding the images. In some instances, information captured in an image could be linked back to an individual. For example, personal id, credit card number, cell phone usage, email screen. In other cases, the display of jewelry, tattoos, clothes could help an acquaintance identify

an individual. Furthermore, a silhouette could provide enough information for a friend or family member to infer identity. A non-obvious threat emerged as a result of analysis of one participant's images of a meeting where under-table shots had potential of providing compromising information about secondary participants.

The IRB mandated us to mark all images that contained any personally identifiable information like face, accessories, and tattoos. Although I found the IRB requirements to be too restrictive, the findings suggested a more complex definition of privacy, one that begs understanding of the relationship between the secondary participants and third party that looks at the images. For example, an email of a person becomes more important than the jewelry or tattoo when an image is shown to a third person. However, it is not easy to establish that relationship when an image becomes publicly available. Hence most stringent rules should be imposed in those situations. But in the cases where access is limited to a set of third party members such as coders or Mechanical Turkers, some criteria could be overlooked without compromising privacy.

An important and somewhat paradoxical condition that this work does not take into account is when the recording of an eating activity represents a privacy violation. In a survey focusing on the activities and habits that people do at home that they would not want recorded, Choe et al. found that the "cooking and eating" category ranked third, behind the self-appearance and intimacy categories [27]. This finding underscores the complexity of the privacy-saliency balance, in particular when there is an overlap between the two.

3.3 Method II: Convolutional Neural Network (CNN)

Two high-level insights emerged out of the study aimed at identifying eating moments using human computation (i.e., Method I above). The first one was that there is a positive correlation between the skill and cost of AMT workers and the quality of inferences. Although the best case scenario in terms of performance resulted in overall accuracy in the range of 90%, this could only be achieved when hiring the most expensive workers. Therefore, it is likely that for most applications, this approach will not scale. Secondly, and more importantly, it is practically impossible to guarantee the level of privacy protection that individuals

demand with a photographic method that also makes use of human computation.

In light of these findings and limitations, I explored another approach for identifying eating activities with FPPOV images. The approach, which is entirely computational, does not make use of external and untrustworthy annotators. Instead, it leverages state-of-the-art methodologies in machine learning and computer vision to automatically infer everyday activities from FPPOV photos.

In contrast to state-of-the-art methods that use hand-crafted features with traditional classification approaches on FPPOV images and videos, the approach is based on Convolutional Neural Networks (CNNs) combining image pixel data, contextual metadata (time) and global image features. Convolutional Neural Networks have recently been used with success on single image classification with a vast number of classes [73] and have been effective at learning hierarchies of features [144]. However, little work has been done on classifying activities on single images from a wearable device over extended periods of time.

To test and evaluate the method, I compiled a dataset of 40,103 images representing everyday human activities. The dataset has 19 categories of activities and were collected by one individual over a period of six months "*in the wild*". The classification method uses a combination of a Convolutional Neural Network (CNN) and a Random Decision Forest (RDF), using what I refer to as a CNN late-fusion ensemble. It is designed to work on single images captured over a regular interval.

3.3.1 Data Collection and Annotation

Over a period of 26 weeks, 40,103 FPPOV images of activities of daily living were collected for one subject. These photos were annotated into 19 activity classes such as cooking, eating, cleaning and playing with kids. The images were aggregated and manually annotated using a tool I developed to facilitate this daily task. The web-based tool, called Activiome, is described in more detail in Appendix A. The activity classes were defined by the subject at their discretion prior to data collection.

The FPPOV photo collection setup used in this study was the same one that was employed for the Method I experiment; The participant wore a phone as a pendant around

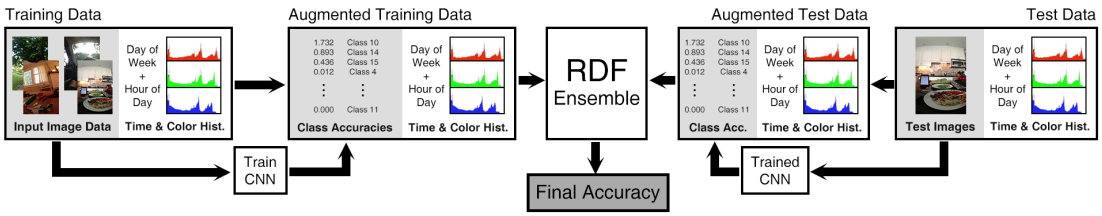


Figure 15: Overview of the Convolutional Neural Network Late Fusion Ensemble for predicting activities of daily living.

the neck with its back-camera facing forward, as shown in Figure 5. An application running on the phone took photos automatically every 30 seconds and uploaded them in real-time to the Activiome system.

At the end of the day, the participant could filter through the images in order to remove unwanted and privacy-sensitive images and annotate the remaining images. The distribution of annotated photos into activity classes is shown in Table 4. The "Working" and "Family" are the top two dominant classes due to the participant's lifestyle. The participant was free to collect and annotate data at their disclosure. The subject was also free to leave ambiguous images (i.e. going from work to a meeting) unannotated. Any unlabeled and deleted images were not included in the dataset.

3.3.2 Description of Dataset

As shown in Table 4, the distribution of tasks was represented by a few common daily tasks followed by semi-frequent activities with fewer instances. It is important to note the difficulty of categorizing certain classes due to their inherent overlap (e.g., socializing vs. chatting, chores vs. family, cleaning vs. cooking, etc). This class overlap is due to the inherent impossibility of describing a specific moment with one label (the participant could be eating and socializing).

The bi-weekly breakdown of data collection is shown in Table 5. It is possible to see a general increase in the number of annotated samples later in the collection process. Some of this was due to increasing the interval at which the application captured images up to once a minute from once every five minutes. The rest of the increase can be attributed to the

Table 4: The distribution of the 19 different classes in the dataset.

Classes	Number of Images	Percent of Dataset
Chores	725	1.79
Driving	1031	2.54
Cooking	759	1.87
Exercising	502	1.24
Reading	1414	3.48
Presentation	848	2.09
Dogs	1149	2.83
Resting	106	0.26
Eating	4699	11.58
Working	13895	34.24
Chatting	113	0.28
TV	1584	3.90
Meeting	1312	3.23
Cleaning	642	1.59
Socializing	970	2.39
Shopping	606	1.49
Biking	696	1.71
Family	8267	20.37
Hygiene	1266	3.12

participant becoming more comfortable with the data collection and annotation process, and over time, successfully incorporating this process into their day-to-day routine.

The participant collected the majority of the data from approximately 7-8am to 7-8pm. Most of the data that was not captured took place during the participant's sleep cycle. On an average day, 80% of the photos were kept; the participant removed approximately 20% of the photos due to privacy concerns and uncertainty about category assignment. The participant handled null classes, such as blurry images, by leaving them unlabeled. These images were then removed prior to assembling the dataset.

3.3.3 Implementation

Recently, Convolutional Neural Networks (CNNs)[76] have been shown to be effective at modeling and understanding image content for classification of images into distinct, pre-trained classes. I used the Caffe CNN framework [59] to build the model since it has achieved good results in the past and has a large open-source community. Since the dataset has a small number of images, I fine-tuned the CNN using the methodology of Hinton et al. [52].

Table 5: The bi-weekly distribution of the number of images in the dataset.

Classes	Number of Samples	Percent of Dataset
Week 1&2	553	1.40
Week 3&4	814	2.07
Week 5&6	69	0.18
Week 7&8	216	0.55
Week 9&10	239	0.61
Week 11&12	2586	6.58
Week 13&14	5858	14.90
Week 15&16	6268	15.94
Week 17&18	2903	7.38
Week 19&20	3417	8.69
Week 21&22	6465	16.45
Week 23&24	4695	11.94
Week 25&26	5229	13.30

It uses the ImageNet [31] classification model introduced by Krizhevsky et al. [73], which was trained on over a million images in-the-wild. I retrained the last layer using the collected data with 19 labels for daily activity recognition. I set the base learning rate to 0.0001 in order to converge with the added data and used the same momentum of 0.9 and weight decay of 0.0005 as Krizhevsk et al. [73] with up to 100,000 iterations as shown in Figure 16. The CNN had five convolutional layers, some max-pooling layers, and three fully-connected layers followed by dropout regularization and a softmax layer with an image size of 256x256. I split the data by classes into 75% training, 5% validation, and 20% testing; the classifier was never trained with testing data on any of the experiments. The parameters were chosen using the validation set and the fine tuning in all of the experiments was only done with the training set. It is interesting to note that the algorithm achieved almost 78% accuracy after only 20,000 to 30,000 iterations and converged around 50,000 iterations due to fine tuning. Despite a high total accuracy, the class accuracy of a CNN alone was hindered due to the lack of contextual information and global image cues.

3.3.3.1 Classic Ensemble

One method to combine the CNN output with non-image data is a classic ensemble method. Training a classifier such as a RDF on the contextual metadata can yield a probability distribution which can be combined with the CNN probability distribution to produce a

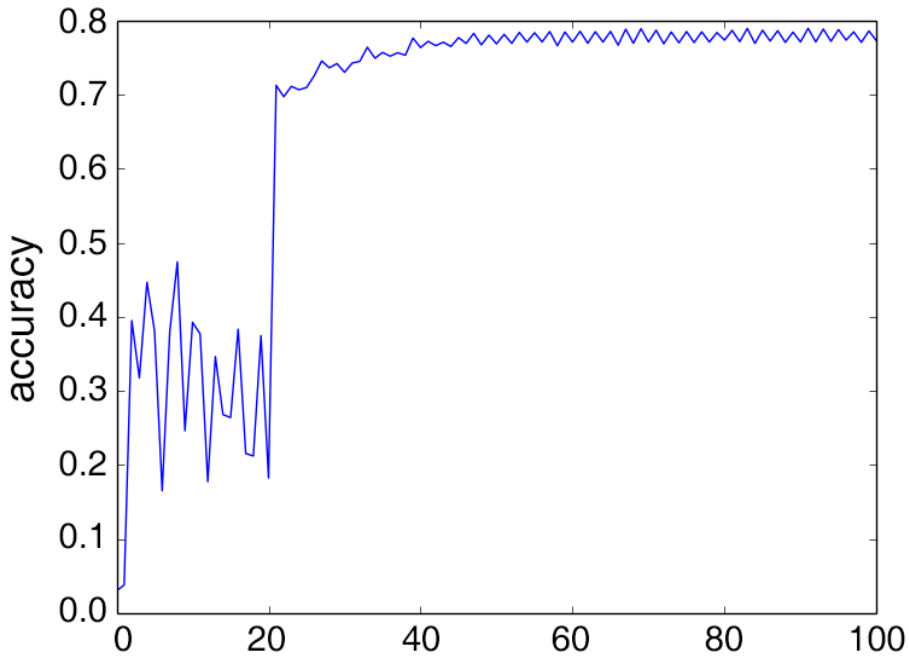


Figure 16: A Convolutional Neural Network trained for 100,000 iterations. Accuracy converges after 20,000 to 30,000 iterations.

Table 6: A comparison of the baselines using RDF trained on contextual metadata, color histograms and a combination of both.

	RDF Metadata	**RDF Hist**	**RDF Metadata+Hist**
Avg. Class Accuracy	15.51	40.43	50.71
Total Accuracy	52.50	68.89	76.06

final probability. This equally weighs the CNN output and the RDF output in order to get the best output possible. This can prevent overfitting from the CNN but doesn't necessarily increase the prediction accuracies since it doesn't leverage which classifier is better at which classes or which information from the classifiers is important.

3.3.3.2 Late Fusion Ensemble

To solve the problem of combining a CNN with a classic ensemble, I developed a late-fusion ensemble technique. I used a RDF trained on the CNN soft-max probabilities along with the contextual metadata (day of week and time of day) and the global image information

Table 7: A comparison of the baselines using kNN trained on contextual metadata, color histograms and a combination of both.

	kNN Metadata	**kNN Hist**	**kNN Metadata+Hist**
Avg. Class Accuracy	15.51	44.23	54.72
Total Accuracy	52.50	65.62	73.07

(histograms of color), each being separate features for the RDF. This allowed for a good combination of outputs that could be learned rather than naively combined. This outperformed the classic ensemble and the normal CNN model by approximately 5%. The pipeline for the method is shown in Figure 15.

3.3.4 Results and Baseline Comparison

In this section I present a comparison of baseline machine learning techniques against the different convolutional approaches for the classification of daily living activities. As shown in Tables 6 and 7, RDF and kNN performed well with contextual metadata (day of the week and time of day) and color histograms. RDFs marginally outperform the kNN methods, particularly with the use of color histograms. It is worth mentioning that other global features (such as GIST [100]) were tested on the same baseline methods and obtained negligible changes in accuracy.

In order to improve the performance of the activity prediction, I leveraged the use of local image information. With a regular CNN, there was a minor increase in total accuracy (+2%) over the baseline (see Table 9), and a more substantial increase in average class accuracy (+7%). There was an even greater increase in accuracy when incorporating both contextual metadata and global image information (color histograms). This motivated the development of the CNN late fusion ensemble that leveraged the metadata and global and local image features. This configuration resulted in a total accuracy of 83.07% with an average class accuracy of 65.87%, showing an impressive increase over the baseline and the other methods. A confusion matrix of the final method's results is shown in Figure 17. In particular, eating activities were recognized with 83.12% accuracy.

I ran evaluations using k-Nearest Neighbor (kNN) and Random Decision Forest (RDF)

Table 8: A comparison of the best of all methods (using contextual metadata, color histograms and pixel data) for all the 19 activity classes. CNN+LF is CNN with Late Fusion Ensemble

	kNN	**RDF**	**CNN**	**CNN+LF**
Chores	**33.10**	17.24	00.69	20.00
Driving	55.07	60.87	**98.55**	96.62
Cooking	25.66	35.53	47.37	**60.53**
Exercising	44.00	63.00	69.00	**73.00**
Reading	**68.55**	49.12	30.04	53.36
Presentation	80.00	72.35	80.59	**87.06**
Dogs	62.17	44.35	55.65	**66.09**
Resting	**72.73**	54.55	27.27	45.45
Eating	77.14	75.75	82.05	**83.12**
Working	91.10	**96.42**	93.49	95.19
Chatting	**21.74**	04.35	00.00	17.39
TV	77.38	75.79	**81.75**	**81.75**
Meeting	68.73	61.00	73.36	**81.47**
Cleaning	26.56	30.47	38.28	**46.09**
Socializing	**52.85**	37.31	31.60	45.08
Shopping	40.16	27.87	63.93	**64.75**
Biking	19.57	23.19	78.26	**81.88**
Family	70.82	87.42	86.69	**90.15**
Hygiene	52.36	46.85	51.57	**62.60**
Avg. Class Accuracy	54.72	50.71	57.38	**65.87**
Total Accuracy	73.07	76.06	78.56	**83.07**

Table 9: A comparison of different CNNs and CNN ensembles using contextual metadata, global features (color histograms), raw image pixels and their combinations. LF is short for "Late Fusion".

	Average Class Accuracy	**Total Accuracy**
CNN	57.38	78.56
CNN Classic (Pixel + Metadata)	53.48	78.47
CNN Classic (Pixel + Metadata + Hist)	59.72	81.49
CNN LF (Pixel)	63.22	80.94
CNN LF (Pixel + Metadata)	65.29	82.45
CNN LF (Pixel + Metadata + Hist)	**65.87**	**83.07**

classifiers in order to adequately fine-tune the best accuracy for the baseline. I parametrized the dataset using contextual metadata (day of the week (as a nominal value from 0 to 6) and time of day) and global image features (color histograms). I found that a kNN classifier (with a k-value of 3) trained on the metadata and the color histograms (with 10 bins) gave an accuracy of 73.07% which was better than training a kNN trained on the metadata alone or the color histograms alone. I tested the classifier at incremental parameters of k (until 50) and found that performance slowly degraded as I increased k beyond 3. I further tested the time metadata at three granularities (the hour, hour + minutes (i.e. 7:30am = 7.5), and hour and minute as separate features) and found the difference in prediction accuracy to be negligible. As a result, I selected to keep the hour and minute as separate features as this led to the highest accuracy. Further, I found that a RDF classifier with 500 trees trained on the metadata and color histograms (with 10 bins) yielded best overall accuracy of 76.06%. As a point of comparison, random chance for this dataset, by picking the highest prior probability, was 34.24%).

Training the RDF with more than 500 trees had a negligible effect on the total accuracy. The baseline results can be seen in Table 6. It is important to note that a high total accuracy was driven by the distribution of the data amongst the classes. Since a majority of the data was in two classes ("Working" and "Family"), a classifier could achieve a high total accuracy by accurately classifying only those two classes. Average class accuracy is also shown to highlight how well the baseline classifier does for all classes distributed evenly.

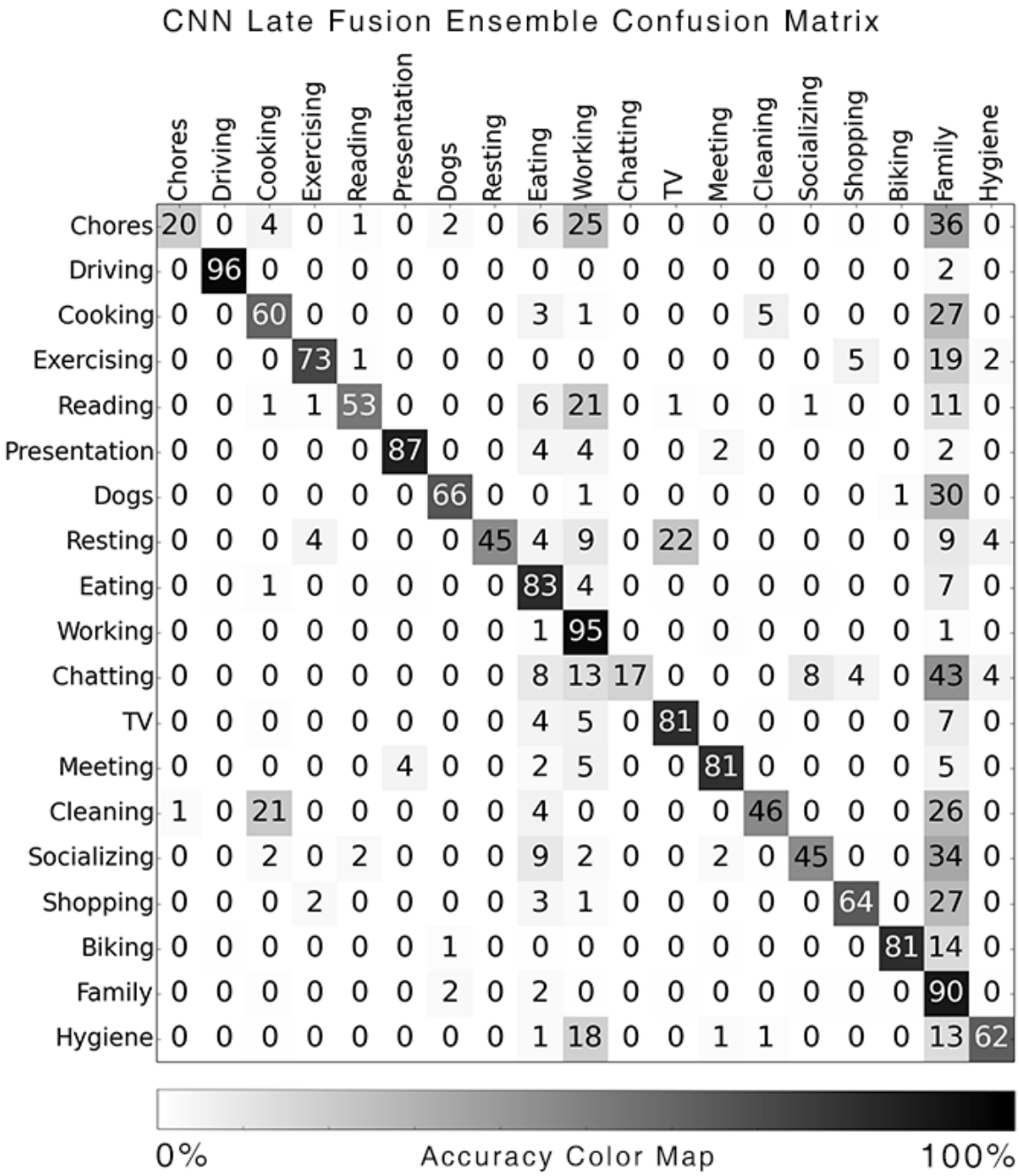


Figure 17: Confusion Matrix for the 19 classes of the dataset with columns as the predicted labels and rows as the actual labels.

3.3.5 Discussion

As shown in Table 9, the CNN late-fusion ensemble method outperformed both the CNN along and the CNN classic ensemble configuration. Training an RDF with extra features and the CNN probabilities allowed the RDF to find what was important for each individual

class. It also allowed for the other types of data to be effectively added, in a framework that prevented some of the overfitting that CNNs typically have. This shows how the novel ensemble method effectively combined local pixel-level information, contextual information, and global image-level information. Because it relied on a CNN running on a GPU, the system used a large amount of power and was not well suited for embedded devices. On an ARM device, testing each image would take more than 15 seconds. However, the method could be run on a server that an embedded device could query.

Figure 18: An example of a classification error on an image from the class "Chores" (class 0). The presence of the kitchen environment in the image led to confusion against other classes including "Eating" (class 8), "Socializing" (class 14) and "Family" (class 17).

Many of the classification failures of the method had to do with classes being inter-related. The worst results were with the "Chores" and "Chatting" activities. An example of a "Chores" misclassification can be seen in Figure 18. In this example, the image has erroneous probability peaks for "Eating", "Socializing" and "Family" classes due to the presence of the kitchen environment in the image, a place where the family meets, socializes and eats together.

In a second experiment, a positive correlation was found between the amount of training data and the algorithms' test accuracy. I highlight two hypotheses for the increase in accuracy over time. The first is that the algorithm was adequately learning the participants' schedule and frequented activities, which allowed it to improve the model. The second plausible hypothesis is that the algorithm was adapting to general human behavior and learning the overall characteristics of specific classes. This presents two interesting questions for the applications of this research. Firstly, how much data is required to train a generic model and secondly, how much data is required to "fine-tune" said generic model to a specific user.

To answer the first question, I trained the model with varying amounts of data points to observe the number of days/samples a person is required to collect in order to train a good generic model. The top 7 classes are shown in Figure 19, with the other 12 classes omitted to maintain clarity). It is possible to see that class accuracies improve as more data is captured, with a significant increase in accuracy after the first 4 weeks.

In order to address the second question, I performed a final experiment in which two volunteers (V1 and V2) wore the wearable device for 48 hours in order to collect images. The data was divided equally into a training and test set (Day 1 for training and Day 2 for testing) in order to test the validity of the model trained by the original participant's data. The results of this experiment are demonstrated in Table 10. As you can see, for some classes that involve a similar viewpoint and environment, like reading, the model generalized well. However, for many others such as driving and chatting, where volunteers were going different places and talking to different people, the model did not generalize well. It is worth noting that the initial accuracy prior to fine-tuning performed worse than the highest prior probability of the original model (34.24%). I reason that this is due to the difference in habits between participants, which requires fine-tuning to adapt to one's specific daily schedule.

Different individuals also have different activities and one set of class labels from one individual might not fit another individual's lifestyle. A valid question to ask is, given the model trained for one person, is it possible to fine-tune the classifier to yield good results

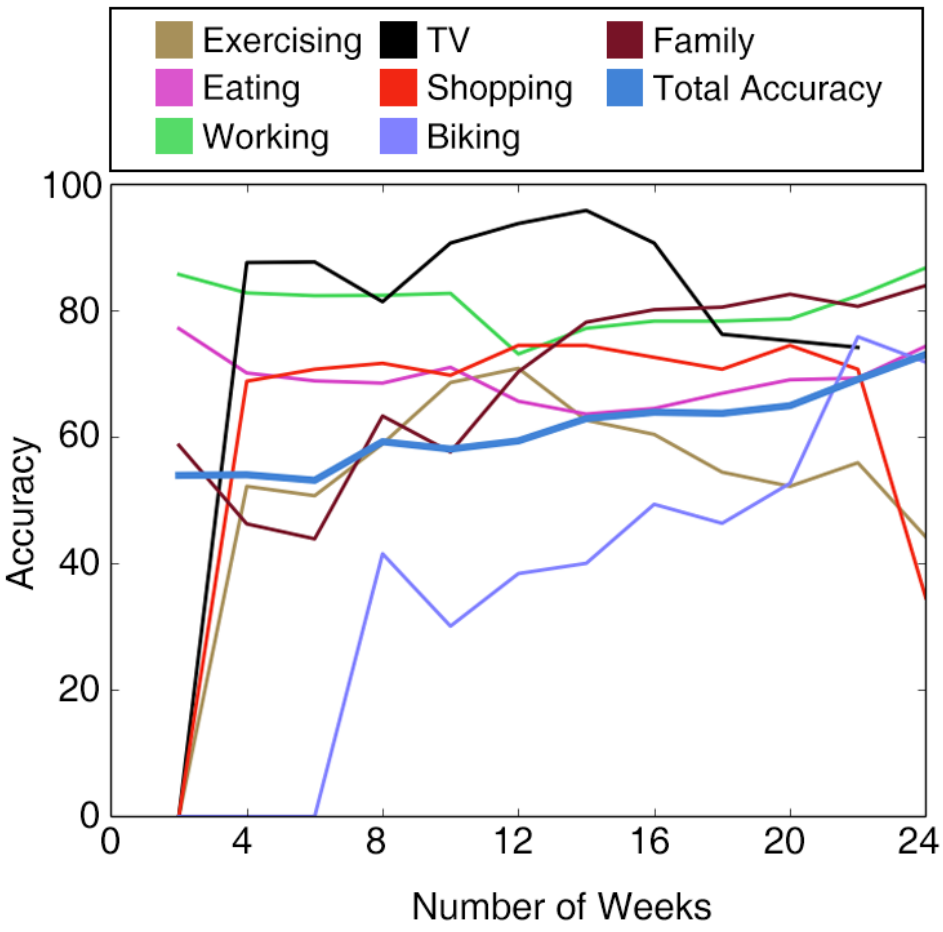


Figure 19: A plot of class accuracies vs. the number of weeks of training samples. A general trend is visible where the class accuracies increase as the amount of training samples increase. A significant increase in accuracy is seen after training on the first 4 weeks of data.

for a different person, even with different classes? At its core, this addresses the question of whether a classifier is learning the schedule and habits of one person or if the learning is inherently adapting to common human behavior. As seen in Table 10, the classifier trained on the original participant was not very successful. However, fine-tuning that model with just one day of data from the new user yielded very good accuracy. Not only did this achieve great accuracy, but the CNN converged in less than 5,000 iterations, whereas the original CNN took more than 50,000 iterations to converge.

3.4 Comparing Method I vs. Method II

In the best-case scenario, when master AMT workers were used, a study employing Method I showed that it is possible to detect eating moments using FPPOV and human computation with 89.68% accuracy. In contrast, experiments with Method II demonstrated that analyzing FPPOV with a state-of-the-art machine learning approach resulted in accuracy of 83.12%.

Clearly, in terms of performance, Method I is superior to Method II. However, its 6% performance gain over Method II comes at a cost. First of all, there is the financial cost associated with the use of human computation. Even though the cost of completing one human computation task is low, the need to review thousands of images and validate annotations causes the overall operational cost to climb rapidly. Secondly, there is the challenge of addressing privacy concerns when making FPPOV photos available to human computation workers. As previously stated, even when employing the most advanced techniques for identifying faces and other possible sources of privacy threats, it is currently not possible to guarantee that all privacy concerns can be addressed computationally. These limitations directly impact the method's scalability and viability for practical, real-world deployments.

With regards to Method II, it is purely computational. As a result, it sidesteps the key scaling limitations of Method I: financial cost and privacy. On the other hand, Method II is centered around training a classifier for identifying eating activities, which also comes at a cost. The model building process requires the acquisition of training data under a variety of real-world settings. However, my experiments showed evidence that it would be possible to build a general classifier for eating detection that could be personalized to individuals without too many additional examples. Under these circumstances, performance results climbed significantly, highlighting the promise of this approach.

Table 10: A comparison of the original model tested on two volunteers and the fine tuned model. "Original" is the original applicants data and model. "V1" and "V2" are the results from the original model tested on volunteers 1 and 2 data respectively. "V1 Fine" and "V2 Fine" are the results from the fine-tuned models trained on volunteers 1 and 2 data respectively. The results that are not available are classes that the two volunteers did not perform when collecting their data.

	Original	**V1**	**V1 Fine**	**V2**	**V2 Fine**
Chores	20.00	5.56	25.0	N/A	N/A
Driving	96.62	18.6	100.0	0.0	100.0
Cooking	60.53	0.0	25.0	N/A	N/A
Exercising	73.00	0.0	50.0	N/A	N/A
Reading	53.36	77.78	75.0	N/A	N/A
Presentation	87.06	N/A	N/A	N/A	N/A
Dogs	66.09	N/A	N/A	N/A	N/A
Resting	45.45	N/A	N/A	N/A	N/A
Eating	83.12	11.48	76.92	30.68	100.0
Working	95.19	31.59	98.32	39.14	94.44
Chatting	17.39	0.0	86.67	0.0	96.72
TV	81.75	0.0	33.33	N/A	N/A
Meeting	81.47	0.0	100.0	0.0	60.0
Cleaning	46.09	0.0	0.0	N/A	N/A
Socializing	45.08	0.0	0.0	0.0	83.33
Shopping	64.75	40.0	50.0	N/A	N/A
Biking	81.88	N/A	N/A	N/A	N/A
Walking	N/A	0.0	57.14	N/A	N/A
Family	90.15	N/A	N/A	N/A	N/A
Hygiene	62.60	13.33	0.0	27.78	81.82
Class Acc	65.87	10.56	51.83	13.94	88.05
Total Acc	83.07	23.58	86.76	27.06	91.23

CHAPTER IV

AMBIENT AUDIO

There are many sounds associated with, and indicative of eating activities. These include the background noise of restaurant environments, the opening and closing of food containers and wrappers, the sound of a microwave oven warming up food, and the softer but highly distinguishable sounds generated by the mouth when chewing and biting. In light of the existence of such audible patterns, I built and evaluated a system to explore whether an eating activity can be detected exclusively from acoustic signatures.

4.1 Method and Implementation

The sound identification task presents two technical challenges: the extraction of information-rich features from ambient audio collected with a microphone, and the design of a binary classifier with the ability to distinguish eating sounds from non-eating sounds from audio features. The next sections describe the entire activity recognition pipeline, from data collection to classification.

4.1.1 Audio Data Collection

Practicality was of utmost priority in terms of audio data collection, therefore my system did not rely on any specialized sensors. Audio was captured by a smartphone attached to the wrist running an off-the-shelf audio recording mobile application. I chose to collect data from the wristThe implementation run on a smartphone device and was evaluated on the wrist in an effort to simulate a smart watch device or some other wearable piece of technology designed for everyday use.

4.1.2 Audio Frames and Features

Audio was recorded at a sample rate of 11,025Hz (16 bits per sample), and audio frames with size 50ms were extracted using a Hanning-filtered sliding window with an overlap of

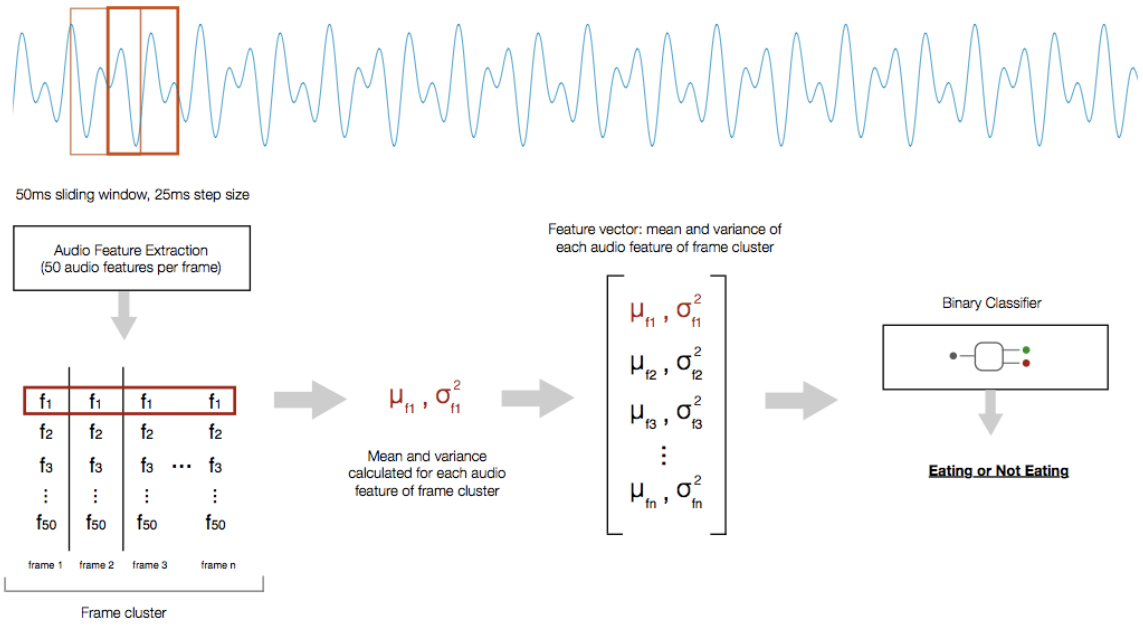


Figure 20: The audio processing pipeline consists of audio framing, audio feature extraction, frame clustering, frame clustering, and classification.

50% (block size=552, step size=276). This audio frame size is larger than what is typically chosen for speech recognition applications but adequate to capture environmental sounds.

I extracted 50 features from each frame, using the Python-based Yaafe tool [91]. Based on previous work that also attempted to recognize human activities from audio [81, 114], I chose the following time and frequency domain features: Zero-Crossing Rate [117], Loudness [96], Energy, Envelope Shape Statistics, LPC [85], LSF [10, 120], Spectral Flatness, Spectral Flux, Spectral Rolloff [117], Spectral Shape Statistics [45], and Spectral Variation.

4.1.3 Clustering and Classification

Because many ambient sounds that characterize eating activities are often much longer than a single audio frame, I clustered 400 consecutive frames and calculated the mean and variance of each feature across these frames (Figure 20). This step also reduced feature "noise" that could be introduced if I had accounted for the acoustic characteristics of every single audio frame.

For clustering, I applied a sliding window over the audio frame stream, also with 50% overlap. This resulted in a frame cluster vector of size 100 (mean and variance of 50

features). I chose 400 frames for each cluster because that is equivalent to a total of 10 seconds of audio, a duration that can encapsulate sounds of interest that are both short (e.g., the clicking sound of utensils hitting plates or bowls), and long (e.g., background noise in a restaurant). I performed classification with the Random Forest classifier available in the Scikit-learn Python package [105].

4.2 Deployment and Evaluation

To evaluate the system, I conducted an IRB-approved in-the-wild study, where I recruited participants and examined how the system performed when classifying ambient sounds collected in the real-world, as individuals performed their normal everyday activities. I recruited 21 participants (15 males and 6 females) between the ages of 21 and 55 through my social network, word-of-mouth, flyers and mailing lists. For joining the study, they received $20 as compensation. Participants included students, research scientists, designers, entrepreneurs and other professionals.

The study lasted between 4 and 7 hours on a single day; for 17 participants, the study began in the morning sometime between 8AM and 11AM and ended between 3PM and 4PM, while for 3 participants it began between 4PM and 7PM and ended before 10PM. This time period was enough to guarantee that all study participants had at least one meal (lunch or dinner).

Subjects wore an audio recording device on the wrist. I chose this placement for the collection of ambient sounds because I anticipate that smart watch-type devices will become popular in the near future. It is very likely that these devices will be capable of recording and even analyzing audio, despite their compact size.

The audio recorder registered sounds continuously throughout the study. At the end of the study, participants were given the opportunity to review their audio file, and delete any audio segment that they did not want to share with us. After this initial step, I performed a walkthrough of the 4-7 hour study period with participants using the Day Reconstruction Method (DRM) [62]. At the end of this process, I was able to discover when individuals ate during the study interval and segmented and labeled their audio clips accordingly.

4.2.1 Day Reconstruction and Verification

To obtain ambient audio ground truth for the eating activities, I asked participants to recall their activities for the day and list them in order, indicating an estimated beginning and end time for each activity. This activity list in chronological order allowed us to discover if and when the participant had a meal. To make sure that time periods indicated by participants were in fact eating activities, two of the authors coded the audio files independently after agreeing on a guideline and then compared results. Disagreements beyond a range of 5 minutes at the beginning or end of an eating activity audio segment were discussed; there were 5 disagreements in total. The final set of ground truth data for each participant included the audio clip referring to the reported eating activity, and another clip with all the audio except for the eating activity segment. As expected, the eating activity audio clip was always much shorter in duration than the audio clip of non-eating activities.

4.3 Results

To reiterate, the high-level goal is to develop and evaluate a *practical* approach to detect when meals are being consumed in the wild. In this work, the primary performance metric I wished to assess was whether the system could identify meal eating activities from ambient sounds. This assessment was driven by collecting data in real situations and learning models from the data to test the approach.

I evaluated the models using a person-dependent technique and reported results in terms of precision, recall and F-score metrics (Table 11); I performed 10-fold cross-validation on each study participant's data and then averaged the results across all participants to obtain an overall result. For comparison, I tested three different classifiers: Support Vector Machines (SVM), Nearest Neighbors (n=5), and Random Forest. The Random Forest classifier proved to be vastly superior to the other two classifiers, yielding an F-score of 79.8%. As a means of comparison, this result is equivalent to what Yatani et al. achieved with BodyScope [143]. On one hand, BodyScope was able to recognize multiple activities. On the other hand, the system does not require any specialized sensor, and can run in any off-the-shelf device that is capable of recording and processing audio, such as smartphones

Table 11: Person-dependent, 10-fold cross-validation results for each classified I evaluated. The Random Forest classifier performed significantly better that the SVM and Nearest Neighbors classifiers.

Classifier	Precision	Recall	F-score
SVM	47.5%	50.5%	48.9%
5-NN	53.3%	51.9%	51.4%
Random Forest	**89.6%**	**76.3%**	**79.8%**

and smart watches.

A LOPO (leave-one-participant-out) cross-validation resulted in an F-score of 28.7%, suggesting that this approach would greatly benefit from personalization. It is important to note that F-measures below 50% are not uncommon in LOPO evaluations, particularly in the context of free-living studies [143].

4.4 Discussion

The ambient audio dataset included meal eating activities in a wide variety of contexts. Participants ate alone and with friends; they ate at home, at work, at school and in the classroom. Although desirable, this level of variety in the data made the classification task particularly challenging.

One factor that hampered the classifier's ability to identify meal eating was the short duration of meal events, which were shorter than 12 minutes in some cases. This resulted in a small number of frame clusters for the classifier to examine, and a misclassification proved very costly. Another difficulty was that some of the participants had their meals while performing other activities such as attending a class or working in the computer, which were not labeled as meal eating activities. It is likely that additional examples would help with activity class separation in this case. Finally, classifying meal-eating in quiet environments, such as one's office or home, has obvious challenges. This suggests a design rationale for training the classifier while emphasizing the specific characteristics of different sounds environments (e.g. home, school, restaurant).

Despite these difficulties, it is worth noting that it would have been impractical to evaluate the system in a controlled lab setting, since it would have been devoid of most

of the natural environmental sounds that individuals are enveloped in when in real world settings and conditions.

4.4.1 Ground Truth Annotation

Estimating ground truth from the audio files proved to be a challenging undertaking. Individuals were asked to recall the exact time they had meals, but often could not do so accurately. In some cases, finding this segment proved particularly difficult, especially when the length of the meal was under 10 minutes. Moreover, while in some audio clips it was possible to hear that participants were eating or were in a restaurant environment, in other clips this was not clear at all. For instance, participants P9 and P14 ate in a classroom or classroom-like environment, whose sounds could not be easily identified as those that are characteristic of an eating activity. In these situations I had to rely on subtle cues, such as the sound of a food container coming out of a brown bag.

Another difficulty I faced in obtaining ground truth had to do with the characterization of an eating activity. Some participants had hour-long lunches, where they chatted with friends extensively before, during and after the meal. On the other hand, some participants had very short meals, eating uninterruptedly for 10 or 15 minutes. In the case of the long lunch, a question might be raised as to whether the whole meal event should be labelled as "eating" or only the period when individuals were actively eating.

4.4.2 Data Collection

Although the feasibility study represents a large ecologically-valid data collection effort, it is limited in two important ways. First of all, since participants joined the study for 4-6 hours in a single day, ambient audio data was recorded for only one meal of their day. For most participants the recorded meal was lunch. The system was evaluated on a per-participant basis through cross-validation, but having just one example of a meal eating activity per participant lowers the confidence that the results generalize over several days. In the future, I plan to address this weakness by collecting data for multiple days per participant. Additionally, the lack of multi-day audio data makes it unlikely that the system's capability to infer eating activities generalizes across individuals. Although I plan

to evaluate the system using a person-independent metric in the future, I believe that most applications and interfaces built on top of the implementation will be personalized (e.g., a just-in-time intervention tailored to address an individual's specific challenges).

Secondly, snacking behavior was not the focus of this study. The duration of data collection per day combined with the times when the study began and ended precluded us from capturing ambient audio around snack-eating activities. However, there is no question that snacking is a highly relevant behavior, and I plan to improve the study design and techniques to account for it in the future. Having said this, a few of the meal eating activities logged in the feasibility study were shorter than 10 minutes, which more closely matches snack eating duration than a "traditional" meal eating duration. The truth is that there is a great deal of ambiguity when it comes to characterizing an eating activity as meal eating versus snack eating.

One of the key issues in audio-based activity recognition is privacy. Understandably, most people object to the recording and analysis of audio of their everyday lives, particularly if it is done completely autonomously and without human input. In the implementation I did not address this challenge, although techniques for protecting privacy in audio streams, and conversational speech in particular, have been proposed [142].

4.5 Conclusion

Based on the results, and despite the limitations of the study, it is clear that acoustic sensing represents a promising opportunity. The system was able to identify meal eating with 89.6% precision and 76.3% recall in a person-dependent evaluation. Although the focus in this work is on the binary presence of eating moments in an audio stream, there are many other dimensions of eating that are relevant from a diet and behavior change perspective. With audio, it might be possible to determine whether individuals are eating alone or with friends, and whether they are eating while working (e.g. typing in a computer) or watching television. I hope to extend the audio-based activity classification platform in the future to capture these additional contextual parameters.

CHAPTER V

SINGLE-POINT INERTIAL SENSING

Considering all human activities, perhaps the most distinguishable characteristic of eating is the set of physical body movements involved in food intake, so called *hand-to-mouth* gestures. These gestures are the ones involved in picking up food, with or without utensils, and bringing it to the mouth. The first study described in this chapter hinged on the recognition of such *food intake gestures* as a foundation to infer eating moments (e.g., breakfast, lunch, dinner, and snacking). In this study, I leveraged the inertial sensing capabilities of wrist-mounted commodity devices for data collection.

The second study also focused on measuring body movements caused by eating, but with inertial sensors placed on a different part of the body: the head. The hypothesis underlying this study was that it is possible to recognize eating from naturally occurring head movements caused by chewing and swallowing. In this final study, participants wore a Google Glass device while performing eating and non-eating activities.

After presenting the systems used for eating moment detection and describing the data collection processes and results for each one of the studies, the chapter concludes with a discussion of several issues and opportunities that emerged from the experiments.

5.1 Dominant Wrist-Mounted Sensing

The aim with this work was to explore a *practical* solution for eating moment detection leveraging the inertial sensor (3-axis accelerometer) contained in a popular off-the-shelf smartwatch. This approach contrasts with methods that require either multiple sensors or specialized forms of sensing.

The eating moment recognition method consists of two steps. First, I perform food intake gesture spotting on the stream of inertial sensor data coming from the smartwatch, which correlate with arm and hand movements. Secondly, I cluster these gestures across the time dimension to unearth eating moments. To evaluate the approach, I first ran

a formative study with 20 participants to validate the experimental design protocol and instrumentation. Informed by this pilot, I conducted user studies that resulted in three datasets: (1) a laboratory semi-controlled study with 20 participants; (2) an in-the-wild study with 7 participants; and (3) 422 hours of in-the-wild data for one participant collected over the course of 31 days.

The approach for estimating eating moments was evaluated in two contexts, in the lab and in-the-wild. The questions I explored in the analysis were:

- How well does the model recognize food intake gestures and eating moments with data collected in a controlled setting?
- How does a model trained with lab data perform at recognizing eating moments in unseen in-the-wild data?
- What is the temporal stability of eating moment recognition in-the-wild using a model trained with laboratory data?

5.1.1 System Implementation

The system was designed to learn to identity moments when individuals are eating food. The sensor data processing pipeline consists of data capture and pre-processing, frame and feature extraction, food intake gesture classification, and eating moment estimation (Figure 21).

5.1.1.1 Sensor Data Capture

Practicality was one of the key driving forces guiding this work. Thus, for data capture I relied on a non-specialized, off-the-self device with inertial sensing capabilities: the Pebble Watch[1]. I wrote custom logging software for capturing continuous 3-axis accelerometer sensor data from the device. The version of the smartwatch I employed did not contain a gyroscope. I also developed an iOS smartphone companion application for data storage and retrieval. Subjects wore the smartwatch on the wrist of their dominant hand. Sensor data was captured at 25Hz.

[1] http://www.getpebble.com

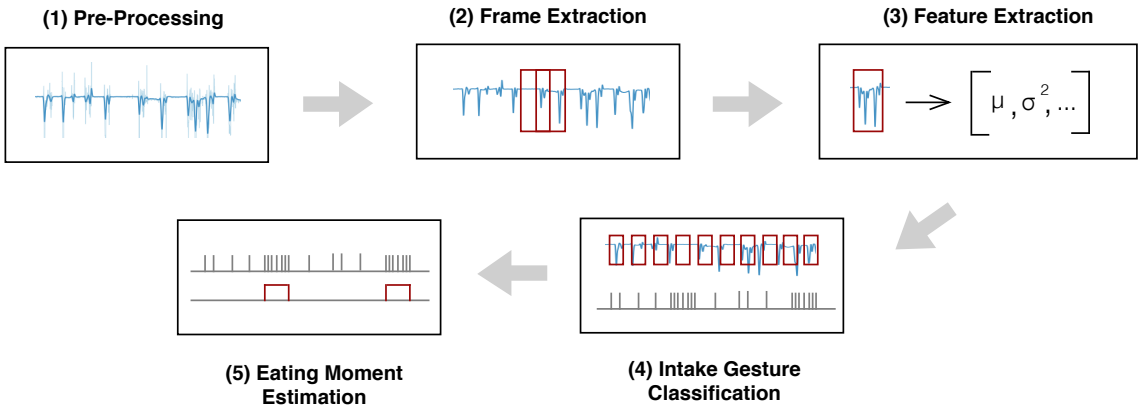


Figure 21: The data processing pipeline of the eating moment detection system. In the approach, food intake gestures are firstly identified from sensor data, and eating moments are subsequently estimated by clustering intake gestures over time.

5.1.1.2 Frame & Feature Extraction

The first steps in the data processing pipeline involved filtering the sensor streams using an exponentially-weighted moving average (EMA) filter and scaling the resulting data to unit norm (l2 normalization).

I extracted frames from the pre-processed data streams using a traditional sliding window approach with 50% overlap. The frame size plays an important role in classification since it needs to contain an entire food intake gesture. The gesture duration is determined by many factors, such as individuals' eating styles and whether they are multitasking (e.g., reading a book, socializing with friends) while eating. Based on data observed in the laboratory user study, I noticed that an intake gesture might last between 2 and 10 seconds. An analysis examining the sensitivity of window size suggested best classification results when the frame size was close to the mid-point of this range, around 6 seconds.

I computed five statistical functions for each frame, shown in Table 12: the signal's *mean*, *variance*, *skewness*, *kurtosis*, and *root mean square (RMS)*. These frame-level features comprise a concise and commonly used representation for the underlying inertial sensor data. This resulted in 5-dimensional feature vectors for each axis of the accelerometer.

Table 12: Feature definitions used for food intake gesture classification

Feature	Description	Definition
mean	average value of the samples of signal $\mathbf{x}$	$\mu_{\mathbf{x}} = \frac{1}{N}\sum_{n=0}^{N-1} x_n$
variance	power of values of signal $\mathbf{x}$ with mean removed	$\sigma_x^2 = \frac{1}{N}\sum_{n=0}^{N-1}\lvert x_n - \mu_x\rvert^2$
skewness	measure of (lack of) symmetry in data list.	$\frac{\sum_{n=1}^{N}(x_n - x)^3}{(N-1)s^3}$
kurtosis	measure of the shape of the data distribution	$\frac{\sum_{n=1}^{N}(x_n - x)^4}{(N-1)s^4}$
RMS	square root of the average power of signal $\mathbf{x}$	$\sqrt{P_x}$, where $P_x = \frac{E_x}{N} = \frac{1}{N}\sum_{n=0}^{N-1}\lvert x_n\rvert^2$

5.1.1.3 Food Intake Gesture Classification

The first classification task in the system is the identification of food intake gestures, which I define as the arm and hand gestures involved in bringing food to the mouth from a resting position on a table, for instance, and then lowering the arm and hand back to the original resting position. In practice, this task is made much harder by intra-class diversity. For example, individuals eat differently if compared to each other and different types of food consumption require different gestures. Additionally, an individual might perform other tasks while eating, such as gesticulate when talking to others, hold a mobile phone or magazine, etc.

For food intake gesture classification, I evaluated classifiers using the Scikit-learn Python package [105]. Best results were obtained with the Random Forest learning algorithm (Figure 24); Random Forests typically perform well with non-linearly separable data, such as the data in this study.

5.1.1.4 Eating Moment Estimation

I estimated eating moments by examining the temporal density of observed food intake gestures. When a minimum number of inferred intake gestures were within a certain temporal distance of each other, I called this event an eating moment. I employed the DBSCAN clustering algorithm for this calculation [38]. DBSCAN has three characteristics that make it especially compelling for this scenario; there is no need to specify the number of clusters

Table 13: To evaluate the system, I conducted laboratory and in-the-wild studies that resulted in three datasets. The duration for the Lab-20 and Wild-7 datasets above represent average duration across all participants.

Dataset	# Participants	Avg Duration	% Eating
Lab-20	20	31m 21s	48%
Wild-7	7	5hrs 42m	6.7%
Wild-Long	1	31 days	3.7%

Table 14: In the laboratory study, participants were assigned to one of two activity groups. Some of the activities involved eating different types of food items while others required participants to perform non-eating tasks. The food eating activities were categorized according to eating style, and utensil type.

	P1-P12	P13-P21
Eat (Fork & Knife)	Lasagna	-
Eat (Hand)	Popcorn	Popcorn, Sandwich
Eat (Spoon)	Breakfast Cereal	Rice & Beans
Non-Eating	Watch Trailer Conversation Take a Walk Place Phone Call	Watch Trailer Conversation Take a Walk Brush Teeth Comb Hair

ahead of time; it is good for data that contains clusters of similar density; and it is capable of identifying outliers (i.e., food intake gestures) in low-density regions. A well-defined method for pinpointing outliers is important because there are many gestures that could be confused with intake ones throughout one's day. Once areas of high intake-gesture densities have been identified as clusters in the time domain, I calculate their centroids and report them as eating moment occurrences.

5.1.2 Deployment and Evaluation

I conducted three user studies, a laboratory semi-controlled study with 20 participants (Lab-20), an in-the-wild study with 7 participants over the course of one day (Wild-7), and a naturalistic study with one participant where I collected 422 hours of in-the-wild data over a month (Wild-Long). More details about these details are available in Table 13.

Table 15: This table is showing the average duration of each activity in the laboratory user study across all participants (dominant wrist-mounted sensing).

Activity	Avg Duration
Eat (Fork & Knife)	5m 1s
Eat (Fork/Spoon)	5m 48s
Eat (Hand)	5m 54s
Watch Movie Trailer	3m 47s
Chat	5m 3s
Take a Walk	2m 18s
Place Phone Call	1m 28s
Brush Teeth	3m 54s
Comb Hair	39s

To evaluate the approach to eating moment detection with wrist-mounted inertial sensors, I first ran a formative study with 20 participants to validate the experimental design protocol and instrumentation for the semi-controlled laboratory study. Participants were asked to eat a variety of foods including fruits (e.g., apple), pizza, and snacks of varying sizes and shapes, such as cookies and M&Ms. To test the feasibility of food intake gesture spotting from a wrist-mounted inertial sensor, I collected data from a smartphone attached to participants' arm, the same setup employed by Dong et al. [35]. A custom application logged all the sensor data on the phone, and all individuals were continuously video-recorded as they ate the food provided.

The pilot study helped us address a number of issues in the experimental procedures, such as the foods offered to participants, the types of non-eating activities I asked participants to perform, the amount of time in-between activities, and the data annotation process. In particular, after observing participants wearing a smartphone attached to their wrists, it became clear that the device's weight and size could affect participants' arm and hand movements, and thus influence the study results. As a result, I transitioned to a smartwatch platform for data collection.

5.1.2.1 Laboratory Study (Lab-20)

I conducted a user study in the laboratory and examined how the method performed when discriminating between eating and non-eating moments. I recruited 21 participants (13 males and 9 females) between the ages of 20 and 43. All participants were right-handed. Due to a data collection error, I had to discard the data for one of the participants.

The study lasted an average of 31 minutes and 21 seconds and participants were invited to arrive around lunch time, between 11AM and 1PM. Participants were asked to wear the smartwatch on the arm they deemed dominant for eating activities. I did not compensate subjects monetarily, but provided them lunch, which they ate as part of the study itself. Before the activities began, I told them the foods I would be serving and gave them the freedom to eat as much as they wanted. I never had more than one subject participating in the study at a time.

The study was designed so that participants performed a sequence of activities (Table 15). Participants were assigned to one of two activity groups (Table 14), which contained a mix of eating moments and non-eating activities. The order in which subjects performed these activities varied depending on the activity group. There were no time constraints, and activities were performed in succession without a significant pause in-between. At the end of each activity, except for the last one, the experimenter instructed participants on what to do next. Although this study was scripted and took place in a lab, participants were free to eat completely naturally. Some participants chose to check news and messages on their phone while eating; others were more social, and ate the food provided while having a conversation with the experimenter and others non-participants who happened to be in the lab.

The eating moments involved eating different kinds of food, such as rice and beans, and popcorn. For consistency, all foods offered were vegetarian, even though many participants did not have any food restrictions. Subjects were provided with utensils for the activities that required them, and a water-filled cup and napkins were made available to them throughout the study. Although drinking is often linked with food consumption, it was not annotated as an eating moment in this study.

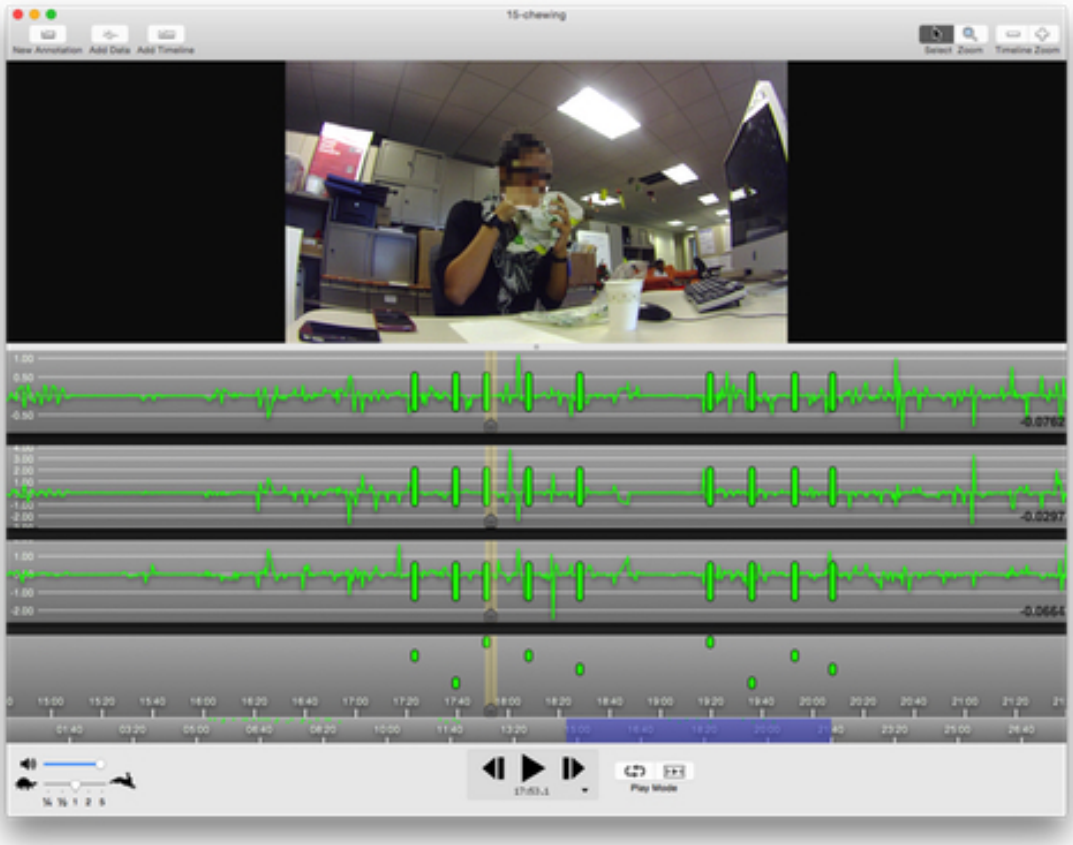


Figure 22: I estimated ground truth by recording each study session with a video camera and then coding the data with the ChronoViz tool [41].

The non-eating activities either required physical movement, or made participants perform hand gestures and motions close to or in direct contact with the head. These activities typically lasted no more than a few minutes, and as little as a few seconds, and were chosen because they are typically performed in daily life and could be confused with food intake in terms of the gestures associated with them. For the "Walking" activity, I asked participants to walk down a hallway, take the stairs down to the floor below, turn around and come back to the study area. The "Phone Call" task involved placing a phone call and leaving a voice message. For the "Comb Hair" and "Brush Teeth" activities, I provided each participant with a hair brush, a tooth brush, toothpaste and they performed these tasks on the spot, with the exception of teeth brushing, which took place in the bathroom.

Participants were continuously audio and video recorded during the study as they performed their assigned activities (Figure 22). The only exceptions were the "Walking" and "Brushing Teeth" activities, when subjects left the user study room momentarily. The acquired video footage served as the foundation for the ground truth I estimated; all coding was performed using the ChronoViz tool [41].

For eating activities, I coded every food intake gesture and differentiated between gestures made with the instrumented arm versus the non-instrumented arm. For food intake, I marked the absolute time the food reached the mouth, and then added a fixed pre and post offset of three seconds to each intake event. This offset made it possible to model the entirety of food intake gestures, which often begin and end moments before and after the food is placed in the mouth. A three-second offset was chosen empirically based on observations of participants' eating gestures. Non-eating activities were coded from the moment they began until their conclusion. In other words, coding for non-eating activities was not focused on modeling any specific gesture.

The reliability of the ground truth estimation scheme was verified by having an external coder review 15% of the recorded audio and video. This was equivalent to 3 study sessions. To account for minor temporal differences in the assigned codes, I established that as long as they were within 3 seconds of each other, the codes referred to the same activity. By following this protocol, there was agreement in 96.7% of the coded gestures.

5.1.2.2 In-the-Wild Studies

To evaluate the ecological validity of the method, I conducted two in-the-wild studies. For the first one, I recruited 7 participants (2 males and 5 females, between the ages of 21 and 29), who did not participate in the laboratory study. They were asked to wear the smartwatch on their dominant arm for an average of 5 hours and 42 minutes for one day while performing their normal everyday activities, which included taking public transportation, reading, walking, doing computer work, and eating. Four participants started the study in the morning and 3 in the afternoon and at least one eating moment was documented for each participant. Of a total data collection time of 31 hours and 28 minutes, 2 hours and

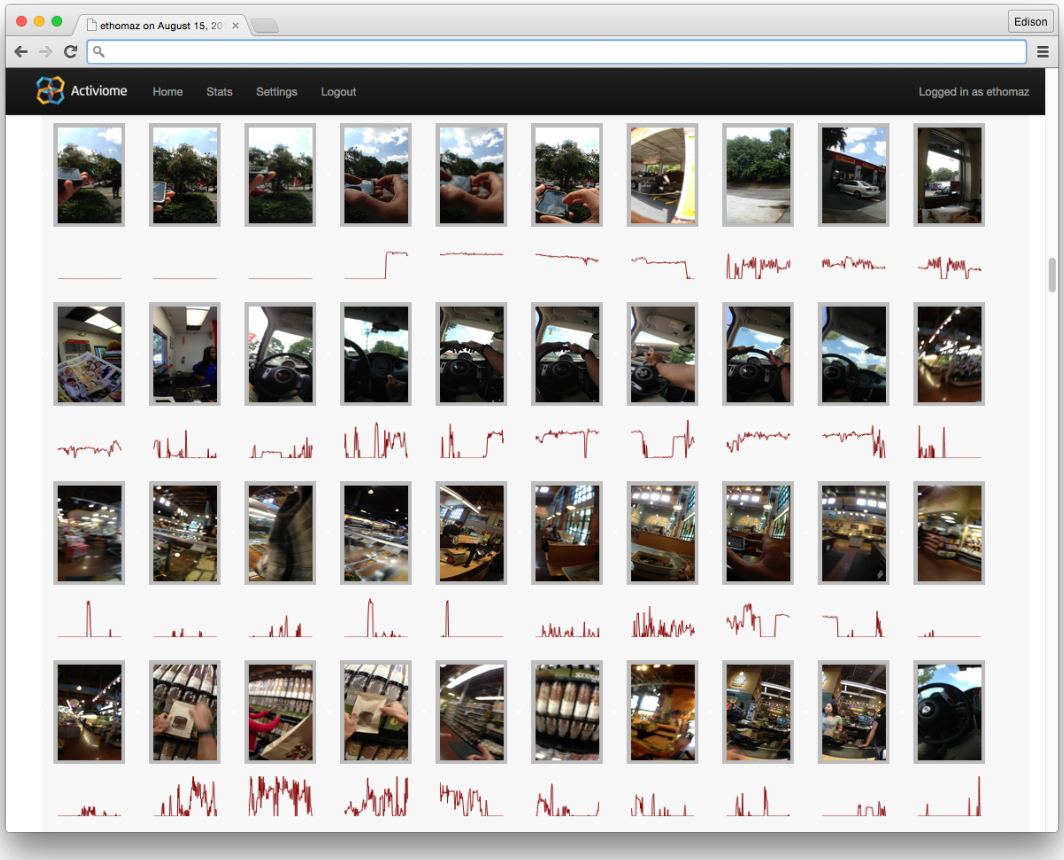


Figure 23: Participants of the in-the-wild study wore a wearable camera that captured photos automatically every minute. After the study, participants were asked to review the photographs and label all eating moments using a web tool specifically designed for this purpose.

8 minutes corresponded to eating activities (6.7% of the total).

In the second study, I (male, 38 years of age) collected and annotated free-living inertial sensor data for 31 days. I wore the smartwatch throughout the entire day, accumulating a total of 422 recorded hours during this period. For this dataset, 3.7% of all sensor data collected reflected eating activities; non-eating activities spanned personal hygiene (e.g., brushing teeth), transportation (e.g., driving), leisure (e.g., watching tv), and work (e.g. computer typing).

In the field of activity recognition, one of the critical challenges of in-the-wild studies is collecting reliable ground truth data for model training and evaluation. Self-reports are

Table 16: Confusion matrix showing the percentage of actual vs. predicted activities by the Random Forest model. The FK and FS acronyms refer to eating activities employing fork and knife, and fork or spoon, respectively.

	Other	Eat FK	Eat FS	Eat Hand	Movie	Walk	Chat	Phone	Comb	Brush	Wait
Other	26%	6.6%	4%	13.2%	13.7%	1.5%	28.5%	3%	0%	3%	0%
Eat FK	2.4%	35.6%	34.2%	14.3%	1.6%	0.2%	10.2%	0.2%	0.2%	0.7%	0%
Eat FS	0.2%	6.2%	74.7%	7.1%	1.1%	0.6%	7.5%	0.5%	0%	1.7%	0%
Eat Hand	1%	4.2%	9.6%	72.9%	1.7%	0.9%	8.8%	0.2%	0%	0.1%	0.1%
Movie	2.2%	0.8%	2.9%	4.7%	77.3%	0.82%	10.1%	0.6%	0%	0%	0.2%
Walk	0.3%	0.3%	0.3%	0.7%	0%	91.3%	5.5%	0%	0%	1.3%	0%
Chat	2.6%	4.5%	15.9%	10.7%	6.9%	1.5%	53%	0.8%	0.3%	3.1%	0.3%
Phone	2.4%	2.4%	24.7%	14%	1.6%	0%	5.7%	47.1%	0%	1.6%	0%
Comb	7.1%	14.2%	17.8%	3.5%	0%	0%	7.1%	0%	39.2%	10.7%	0%
Brush	1.4%	3.3%	16.8%	16.8%	0%	11%	11%	0.9%	0.9%	37.5%	0%
Wait	3%	5.1%	17.3%	5.1%	5.1%	4%	9.1%	0%	0%	6.1%	44.9%

typically used for this purpose, but they are known to be susceptible to biases and memory recollection errors. To improve the reliability and objectivity of ground truth for the in-the-wild studies, I built an annotation platform around first-person images called Activiome, described in detail in Appendix A. In addition to the smartwatch, participants wore a wearable camera on a lanyard that captured photographs automatically every 60 seconds, depicting participant's activities throughout the day (Figure 5 in page 30). These images were uploaded in real-time to a server, and participants could access and review them at any time by logging into a password-protected web application. With this system, participants were able to indicate when they were engaged in eating moments from photographic evidence without having to share their photos with the research team, mitigating privacy concerns.

This method offered greater confidence for the ground truth labels, because the annotation was based on picture evidence. The camera was outfitted with a wide-angle lens to maximize the field-of-view and capture food and eating-related activities and objects even if they were not directly in front of the individual. However, since photos were taken only every 60 seconds, there is a small possibility that a short eating moment (e.g., a snack) occurred in-between two photos and was not recorded. I set the interval to 60 seconds as a compromise between maximizing battery life and photo capturing for as long as possible on a given day.

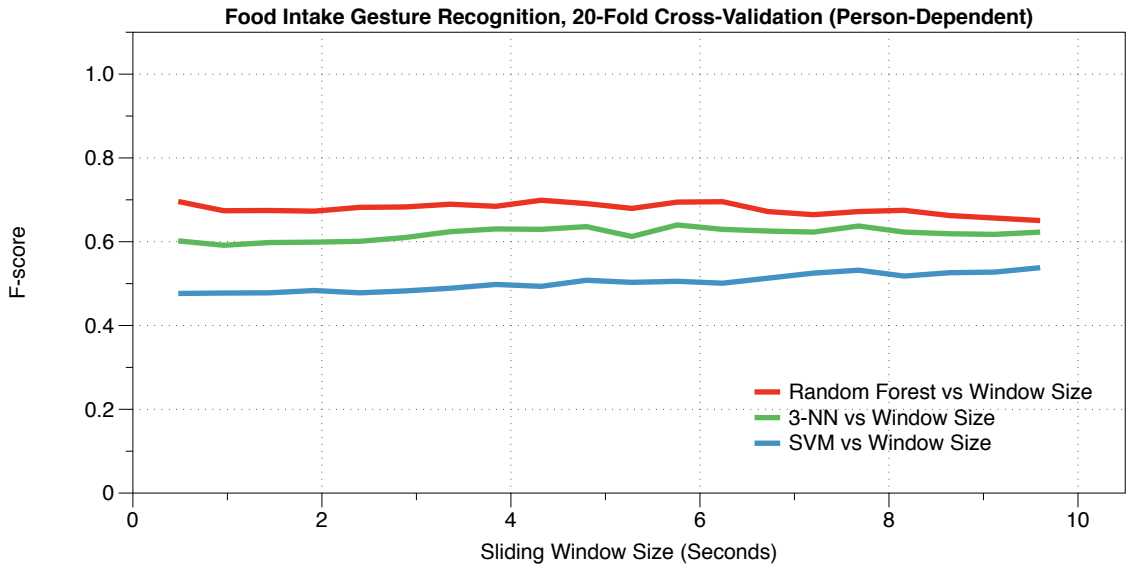


Figure 24: I evaluated the person-dependent performance of three food intake gesture classifiers with respect to window size (Lab-20 dataset). Each classifier was trained with a different learning algorithm: Random Forest, SVM (RBF kernel), and 3-NN. I achieved best results with the Random Forest classifier.

5.1.3 Results

To reiterate, my goal is to develop and evaluate a *practical* approach to detect eating moments, using sensor data from an off-the-shelf smartwatch. To that end, the primary performance metric I wished to assess was whether the system could distinguish eating moments from non-eating moments. In this section I first review the eating gesture classification findings and then discuss the eating moment recognition results.

5.1.3.1 Recognizing Eating Gestures

In the system, predicting eating moments hinges on the detection of food intake gestures. Using the Lab-20 data, I evaluated the performance of three food intake gesture classifiers (Random Forest, SVM, and 3-NN) as a function of sliding window size for the person-dependent (Figure 24) and person-independent cases. The Random Forest classifier outperformed the SVM and 3-NN classifiers using the F-score measure for comparison. I attribute this result to the Random Forest's powerful nonlinear modeling capability. This learning algorithm was also appealing to us because it does not require much parameter tuning.

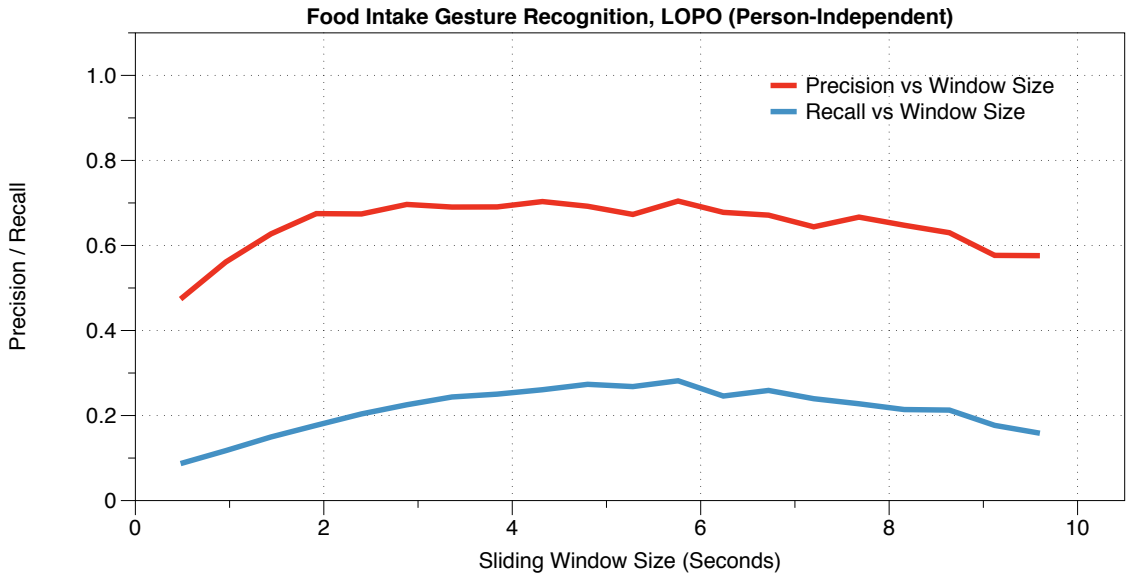


Figure 25: I performed a leave-one-participant-out (LOPO) evaluation of the food intake gesture classifier trained with the Random Forest learning method. The figure shows its sensitivity to window size.

A person-independent evaluation of the Random Forest classifier using the leave-one-participant-out strategy (LOPO) is shown in Figure 25. Note that the reported precision, recall and F-score measurements in Figures 24 and 25 reflect the classifiers' ability to spot intake gestures at the frame level, and best performance was achieved with a frame size of just under 6 seconds.

Table 16 provides a detailed picture of how the Random Forest model performed at classifying eating gestures in relation to non-eating activities. The data for all laboratory study participants was combined and randomly split into one training and one test set; approximately one third of the data was held out for testing. This procedure was performed with Scikit-learn's train-test-split cross-validation function [105]. For purposes of reporting results, I further distinguish 3 different eating gestures to gain a richer understanding of model classification and error rates: eating with fork and knife (i.e., Eat FK), eating with fork or spoon only (i.e., Eat FS), and eating with hands (i.e., Eat Hand).

Figure 26: F-score results for a model trained with lab data (Lab-20 dataset) and tested with in-the-wild data, Wild-7 (red), and Wild-Long (blue). The x-axis correspond to time segment size, in minutes.

5.1.3.2 Estimating Eating Moments

As previously described, the approach for inferring eating moments depends on the temporal density of observed food intake gestures; I cluster these intake gestures over time using the DBSCAN algorithm, which takes two parameters, a minimum number of intake gestures (minPts), and a distance measure given as a temporal neighborhood (eps). To assess how well eating moments were recognized, I compared ground truth and predictions over a time window that is longer than a frame size. This is necessary because an eating moment is in the range of minutes, not seconds. In this paper, I refer to this longer time window for eating moment recognition as a *time segment*, shown in Figure 27. When one or more eating moments are recognized within a time segment, the entire time segment is assigned the eating label.

One of the questions this work explores is whether it is feasible to build a model for eating moment recognition based on semi-naturalistic behavior data captured in a laboratory. To answer this question, I trained a model with the Lab-20 dataset and tested it on both in-the-wild datasets (Wild-7 and Wild-Long). Figure 26 plots F-scores as a function of time

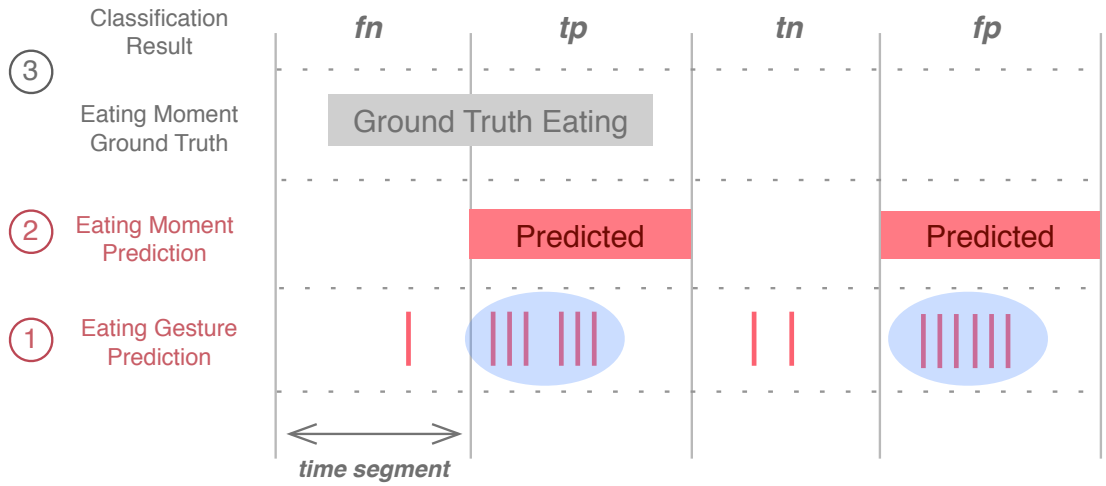


Figure 27: Going from bottom to top, the first step to eating moment recognition involves recognizing eating gestures (1). These are clustered temporally to identify eating moments (2). Finally, estimated eating moments are compared against ground truth in terms of precision and recall measurements at the level of time segments ranging from 3 to 60 minutes (3).

segment size ranging from 5 to 60 minutes (DBSCAN parameters set to minPts=1, eps=10, meaning at least 1 intake gesture that is within 10 seconds from another recognized intake gesture). The charts show an upward trend in recognition performance as time segment duration increases. This is because more data points become available in terms of recognized and non-recognized food intake gestures, leading to improved density estimation, and thus better eating moment recognition results. When the time segment size is set to 60 minutes, the F-scores are 64.8% and 56.8%.

The intuition guiding eating moment recognition is that making a prediction about a 60-minute time segment would suffice for most practical applications of the work. Given that intuition, it is valuable to understand how much one can optimize the classifier when the time segment is fixed at 60 minutes. Varying the minPts and eps parameters of the DBSCAN algorithm, but still using the Lab-20-trained intake gesture recognition model, (shown in Figures 28 and 29), F-scores of 76.1% (66.7% Precision, 88.8% Recall) and 71.3% (65.2% Precision, 78.6% Recall) could be achieved when evaluating the classifier with the Wild-7 and Wild-Long datasets, respectively.

5.1.4 Discussion

In this section, I discuss the classification results, the instrumentation strategy I chose, characteristics of the data collected, and the practical implications of the findings.

5.1.4.1 Classification Challenges

To more realistically assess the system's classification performance in the lab study, I purposely included gestures that required arm movements similar to food intake gestures. Activities such as placing a phone call, combing hair and brushing teeth are all similar to eating in that they all require hand-arm motions around the head and mouth areas. Other observed movements that occurred in the laboratory study closely matching eating gestures included wiping the face with a napkin, scratching the head, and assuming a resting position by supporting the head and chin with the instrumented hand and wrist. Because of the semi-controlled nature of the laboratory study, these movements occurred naturally during sessions, and did not have to be scripted.

Based on the results, shown in the confusion matrix in Table 16, I found that one of the most challenging activities to discriminate from eating was "Chat". This is because when people are having a conversation, they typically gesticulate. This effect varies in intensity amongst individuals but it was significant enough across all participants in the laboratory study that between 7.5% and 10% of each eating intake class (Eat FK, Eat FS, Eat Hand) was misclassified as "Chat".

In Table 16, it is also possible to see false positives originating from the "Phone", "Comb", and "Brush" activities. This is not surprising since these activities were specifically included to induce misclassifications. Common to these non-eating activities gestures was a movement bringing the hand close to the head; the *temporality* of subsequent movements was one of the key characteristic differentiating them. In the "Phone" activity, the hand stayed up holding the phone close to the ear; in effect there is no subsequent "hand down"

Figure 28: F-score results for estimating eating moments given a time segment of 60 minutes as a function of DBSCAN parameters (minPts, and eps). Tested on the Wild-7 dataset, eating moments can be estimated with an F-score of up to 76.1% when minPts=2 and eps=80 (at least 2 intake gestures that are within 80 seconds from another intake gesture).

gesture in this case. For the "Comb" activity, the hand was lifted up and remained in motion, moving slowly in a pattern that depended on the hairstyle of the participant. The "Brush" activity pattern was distinguished by quick-moving hand gestures while holding a toothbrush. I believe the rate of false positives can be lowered by incorporating time-dependent features that can better characterize these types of non-eating activities.

5.1.4.2 Intra-Class Diversity

I observed a large amount of variability in participants' eating styles. Some held a sandwich with two hands, others with one hand, sometimes alternating between them. A minority of participants took bites of their food at regular intervals (P4 in Figure 30). Others were not so regular; they gesticulated more while talking and eating (P5 in Figure 30).

When using utensils, and in the short intervals between bites, some participants kept mixing their food in a regular pattern. This could be attributed to an individual's own

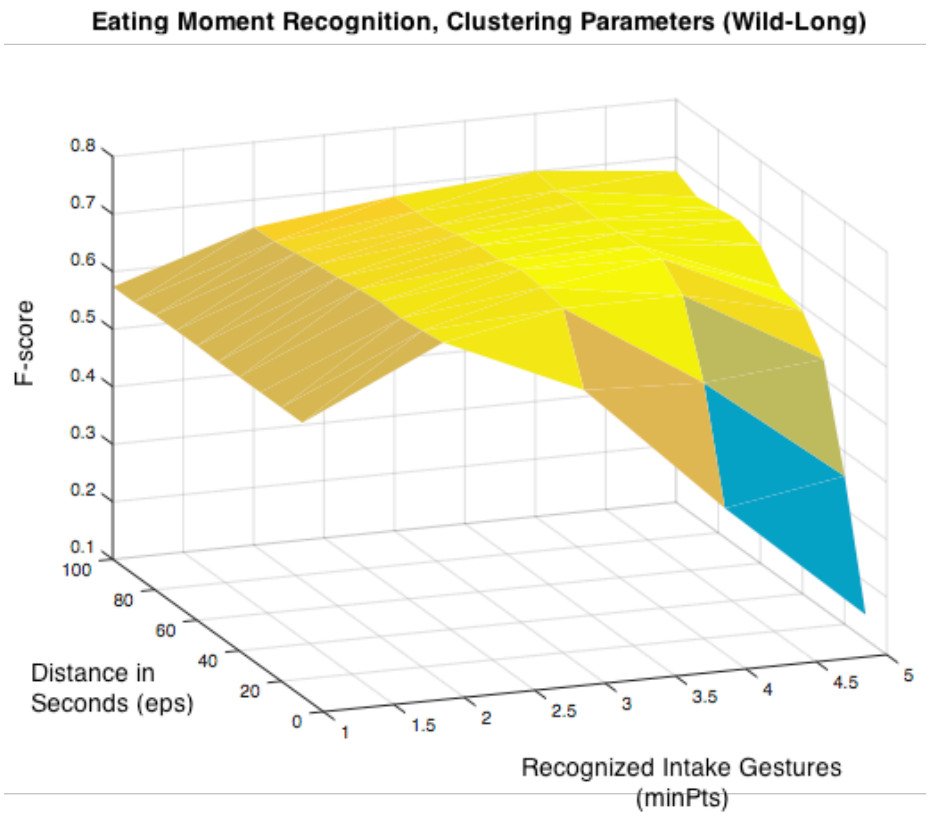


Figure 29: F-score results for estimating eating moments given a time segment of 60 minutes as a function of DBSCAN parameters (minPts, and eps). Tested on the Wild-Long dataset, eating moments can be estimated with an F-score of up to 71.3% when minPts=3 and eps=40 (at least 3 intake gestures that are within 40 seconds from another intake gesture).

eating style or an attempt to cool off the food, for example. There was significant variation in the way participants ate smaller foods as well. Several participants held several kernels of popcorn in hand and ate them continuously until they were gone. Others liked to eat more than one popcorn at a time.

While many participants performed the "traditional" food intake gesture of bringing food to the mouth using utensils, hands, or by lifting a bowl, I noticed that many participants did the opposite; they bent over their plate, brought their head close to the food and then moved their arm in a modified, shorter and subtler version of the traditional intake gesture. This was particularly common when participants were trying to avoid food spillage (P1 in Figure 30).

In this study I did not create a separate model for each observed eating style; all intake

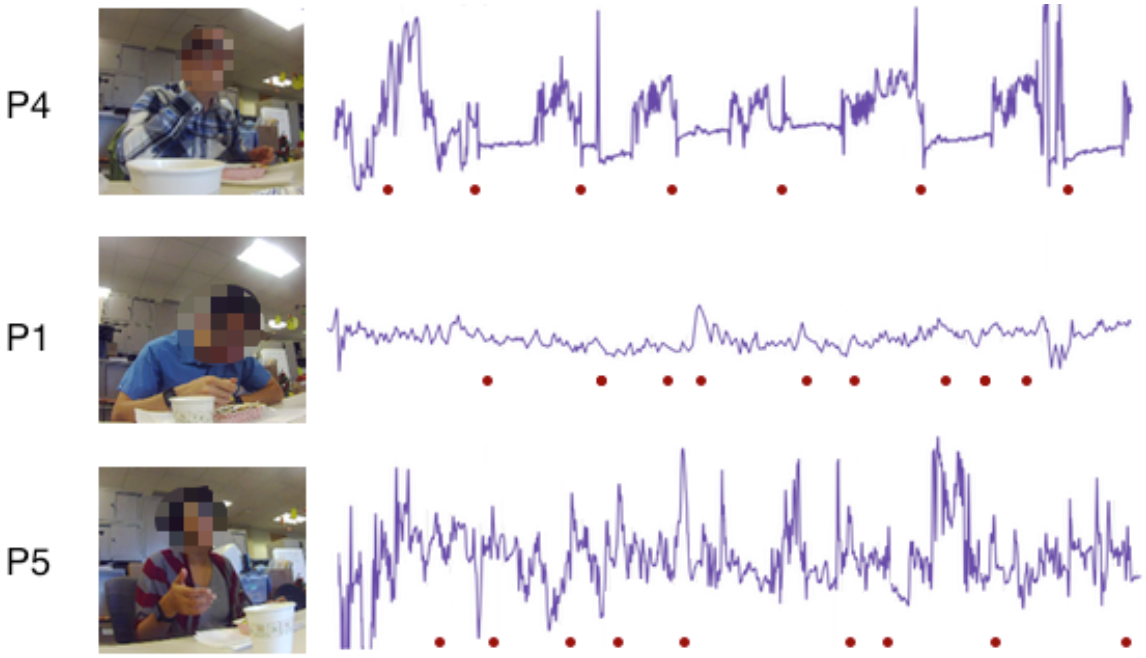


Figure 30: The accelerometer data (x-axis) of three participants as they ate a serving of lasagna depicts personal variation in eating styles and makes intra-class diversity evident. The red dots are intake gesture markers.

gestures were given one label: "eating". Without any question, this posed an additional challenge to the classification task. Fitting a model to user-specific data might be the most effective way to address intra-class diversity, and I hope to explore this in future work. Also, face-mounted wearable computing systems like Google Glass are becoming more popular; these devices offer the opportunity to capture inertial sensing data reflecting head movements, which might contribute significantly to the identification of eating and chewing activities despite individual differences.

5.1.4.3 Instrumentation

I provided participants with one wrist-worn device, a smartwatch, and placed it on their dominant hand. There are two key reasons why I decided on a strategy of minimal instrumentation. Firstly, in real-world settings, people wear only one smartwatch at a time. In this context, with an eye towards the practical applicability of this research, I was interested in the extent to which eating moments can be estimated with just one sensor data capture device. Secondly, I felt that asking participants to wear one additional device would be

unnatural, and thus result in a level of discomfort that could compromise the validity of the data.

I chose participants' dominant hand because it is the one that is typically used in food intake gestures. However, the dominant hand might play different roles while eating, such as cutting with a knife, and this has an effect in modeling intake gestures; it is possible to observe in Table 4 that the "eating with a fork and knife" class was misclassified as "eating with fork or spoon only", and with "eating with hand". This is inconsequential if the goal is to identify "whether" eating is taking place, but it presents modeling opportunities for characterizing "what" is being eaten.

5.1.4.4 Ecological Validity

The evaluation results demonstrate the promise of a minimally-instrumented approach to eating moment detection. However, it is important to situate the findings in light of the study design and aspects of the system implementation. An issue that might arise in practice while collecting data with only one device is that certain eating gestures might not get captured. For instance, a person might be wearing a smartwatch on the non-dominant hand while eating with a fork held by the dominant hand. Although this scenario represents a challenge, I believe it can be addressed in two ways: by modeling non-eating gestures performed by the non-dominant hand during eating, and by leveraging additional modalities such as ambient sounds. In future work, I plan to explore the combination of these two different paths.

With regards to the validity of the results, the types of foods that I served participants and the enforcement of which utensils they were allowed to use, if any, were in line with current western eating traditions. I aimed for a representative sample of eating activities and styles by picking foods such as rice, popcorn, and sandwiches apples but the scientific claims do not and cannot generalize to all populations and cultures. For instance, none of participants in the study ate with chopsticks.

5.1.4.5 Practical Applications

Despite the importance of high precision and recall measures for both benchmarking and practical applications, the experiments showed that since there are usually many intake gestures within one eating moment, a slightly lower recall in food intake gesture classification does not have a large effect in the results. In contrast, consecutive false positives have a direct effect in the misclassification of eating moments. With respect to the applications I envision leveraging this work, there are two paths to consider. In a system designed to facilitate food journaling, lower precision means that individuals might be frequently prompted to provide details about meals that did not occur, which is undesirable. However, as a tool for health researchers to determine when individuals eat meals, what is critically important is to not miss any eating activities. In this case, false positives are preferable to false negatives.

5.2 Head-Mounted Sensing

It is possible to observe a variety of head motions as an individual performs an eating activity. Many of these motions are subtle and caused by the biomechanics of chewing and swallowing food. Others are more noticeable, such as when the head tilts up or down to place the mouth in the trajectory of an incoming fork or spoon. Despite the existence of head movement and patterns that seem linked to food consumption, head motions are a constant in daily life. To understand whether it is possible to uniquely identify eating head motions from non-eating motions, I conducted a study with 20 participants in a laboratory setting. The following sections describe the study and how the data was collected and analyzed.

5.2.1 Laboratory Study

The study and data collection effort for the head-mounted sensing of eating took place together with the dominant wrist-mounted eating detection experiment. In other words, the 20 participants in the dominant wrist-mounted sensing laboratory study wore two devices with inertial sensing capability: a smartwatch and a Google Glass device.

5.2.2 Data Capture and Analysis

Data was collected with a standard Google Glass device that participants wore throughout the experiment. An application written for the Android OS captured 6 streams of sensor data in real-time at 45Hz: 3-axis of accelerometer data and 3-axis of gyroscope data. The data was saved locally on the device and downloaded at the end of each study session for analysis.

Similarly to how I processed the inertial sensor data obtained from the wrist-mounted device, I first filtered and scaled the data; an exponentially-weighted moving average (EMA) filter was used to smooth the data and l2 normalization was applied to bring it to unit norm. A sliding window extracted frames from the sensor streams (50% overlap), and 5 statistical measures were calculated for each frame (Table 12): *mean*, *variance*, *skewness*, *kurtosis*, and *root mean square (RMS)*. For classification, the Random Forest learning algorithm was used. It took as input a vector with 30 features (5 statistical measures for each one of the 6 streams of inertial data), and output an eating detection model.

5.2.3 Results

To assess the extent to which the aforementioned approach worked for eating detection, I performed a person-independent (leave-one-participant-out) evaluation with the laboratory data and calculated precision and recall measures. With a sliding window size of 35 seconds, precision and recall were 72.9% and 66.5% respectively (69.6% F-Score). I experimented with different window sizes, ranging from 10 to 50 seconds and did not observe a significant difference in results, as shown in Figure 31.

This result compares very favorably to other efforts focused on eating detection with head-mounted inertial sensing. Rahman et al. obtained a LOPO $F_{0.5}$-Score of 49.73% in a lab-setting where 38 participants ate their own food and performed other activities of their choice for a total period of 2 hours [106]. The LOPO $F_{0.5}$-Score of my classifier was 71.52%, but with a smaller number of participants and a shorter lab session.

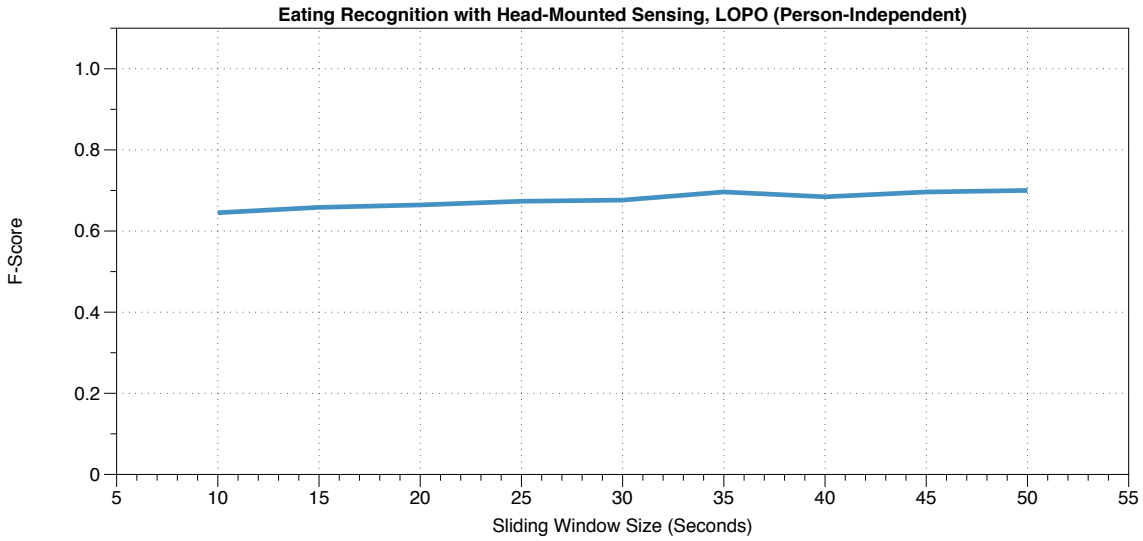


Figure 31: I performed a leave-one-participant-out (LOPO) evaluation of the activity classifier. The Random Forest classifier was trained with inertial sensor data captured with Google Glass. The figure shows its sensitivity to window size.

5.2.4 Conclusion

The two studies described in this chapter aimed at investigating the performance of eating detection classifiers based on inertial sensor data. The first study hinged on the recognition of food intake gestures and eating moments with a wrist-mounted device. The results obtained were promising for three reasons. Firstly, they represent a baseline for practical eating detection using a device with very limited sensing capabilities: the Pebble watch. I anticipate performance gains when employing additional inertial sensing modalities, or using a device with a more powerful IMU. As a means of comparison, Amft et al. obtained 84% recall and 94% precision with accelerometer and gyroscope in drinking gesture spotting [2]. Secondly, the dominant hand study explored one type of sensing modality, inertial sensing, but many other contextual cues could be utilized to improve eating moment detection, such as location and perhaps even ambient sounds [130]. And thirdly, this work suggests that it might be possible to build ecologically valid models of complex human behaviors while minimizing the costly acquisition of annotated data in real-world conditions.

The second study also focused on measuring body movements caused by eating, but with inertial sensors placed on a different part of the body: the head. The hypothesis

underlying this study was that it is possible to recognize eating from naturally occurring head movements caused by chewing and swallowing. In this study, participants wore a Google Glass device while performing eating and non-eating activities. The results proved to be on par, if not better, against comparable efforts aimed at detecting eating with head-mounted inertial sensing.

Despite the promise of this method, it was evaluated in a laboratory setting; I would expect lower performance overall in real-world conditions. One of the challenges of a model created around the recognition of head movements is the poor signal-to-noise ratio of head-acquired inertial sensor data. In naturalistic settings, the head is constantly moving even when the person is performing just one activity. To make matters worse, eating is often a social activity, with individuals often turning their heads to face each other to talk or looking away from their food if something catches their attention. These and other causes for head movements while eating result in *noise* in the data. The existence of this *noise* masks the much smaller variations and patterns in the data that correspond to chewing and swallowing motions. As a point of comparison, Hernandez et al. were successful in estimating vital measures such as pulse and respiratory rate of 12 participants using Google Glass, but only in a controlled experiment where participants had to be still while measurement were taken[50].

CHAPTER VI

TWO-HANDED INERTIAL SENSING

As noted, one of the limitations of the dominant wrist-mounted sensing study was that gestural data was captured on only one arm. Although it is reasonable to assume that most eating gestures engage the dominant hand, there are situations when that is not the case. If the only instrumented arm is the dominant one, missing an eating gesture due to the use of the non-dominant hand results in a false negative. These types of false negatives indicate that an eating gesture went undetected, but not because the food intake gesture classifier produced an incorrect result. From a scientific perspective, it is valuable to measure the performance of an inertial sensor-based food intake gesture classifier irrespective of which arm performs the gesture. I addressed this research question by conducting a laboratory study where participant wore a wrist-mounted inertial sensor on each wrist.

Additionally, while the dominant wrist-mounted sensing experiment examined the performance of eating moment detection in real-world settings for multiple individuals, it did so for just one day. There was one exception; data was collected for one participant for a month, but additional validation with more participants is warranted. Considering that individuals have unique eating styles, it would be valuable to know if an eating moment classifier can be tailored to a person from data compiled in laboratory and real world conditions. This was the motivation for the in-the-wild study presented in this section.

6.1 Implementation and Data Capture

In the dominant wrist-mounted sensing experiment, participants wore a Pebble watch. By means of the watch's accelerometer, inertial data was recorded as participants performed eating gestures and engaged in other activities. Recently, wrist-mounted consumer devices with more powerful inertial measurement units have become available. For this study, I

Table 17: This table is showing the average duration of each activity in the laboratory user study across all participants (double wrist-mounted sensing).

Activity	Avg Duration
Eat (Fork & Knife)	12m 41s
Eat (Spoon)	5m 39s
Eat (Hand)	5m 28s
Drink (Hand)	0m 44s
Watch Movie Trailer	2m 19s
Read Magazine	6m 38s
Take a Walk	4m 14s
Use Mobile Phone	5m 34s
Place Phone Call	1m 26s
User Computer	6m 14s
Brush Teeth	4m 5s

relied on one of these newer devices, the Microsoft Band[1]. It contains both an accelerometer and a gyroscope, thus providing 6 DoF inertial sensor data. I adapted the Activiome iOS smartphone companion application to work with the Microsoft Band, and captured sensor data at 30Hz. The Activiome system is described in detail in Appendix A.

The pipeline used for data processing was exactly the same as the one I employed for the dominant wrist-mounted sensing experiment, with one exception. Participants wore two Bands; each Band recorded 6 channels of inertial data: 3 for accelerometry and 3 for gyroscopic data. Therefore the data processing pipeline was modified to take in these additional data channels.

6.2 Food Intake Gesture Spotting

I conducted a laboratory study to compare the performance of an inertial sensor-based food intake gesture classifier with gestural data from both hands, only the dominant hand, and only the non-dominant one. Like previous lab experiments, it centered on collecting behavioral sensor data as participants ate a variety of foods and performed non-eating activities in a semi-controlled environment. Four participants (3 males, 1 female) were

[1]http://www.microsoft.com/microsoft-band

Figure 32: Participants were video-recorded as they performed eating and non-eating activities in the laboratory study.

recruited for the studies; they were graduate students between the ages of 19 and 26, and all of them claimed to be right-handed.

The protocol I employed was very similar to the one used in the dominant wrist-mounted sensing experiment. The study lasted an average of 55 minutes and took place around lunchtime. I instrumented participants with two Microsoft Bands, one on each wrist, for collecting accelerometer and gyroscope inertial sensor data. A video camera was setup in front of participants and video recordings were used for annotating gestures and activities (Figure 32).

Participants performed eating and non-eating activities (Table 17) and there were no time constraints for completing them. The eating activities revolved around a pre-defined set of foods which included popcorn, a serving of lasagna, and yogurt (Figure 33). All participants were offered the exact same food types and amount for each food. Some eating activities required the use of utensils and some did not. Participants were told which foods would be served and allowed to eat as much as they wanted, and drinking activities were

Figure 33: A participant in the study wearing two Microsoft Bands, one on each wrist, and eating a serving of lasagna with fork and knife.

coded as separate from eating activities.

A variety of non-eating activities were included in the study. One of them required physical movement (i.e., walking), some were mundane everyday tasks (i.e., use computer or mobile phone), and some involved performing hand gestures and motions close to or in direct contact to the head (i.e., brush teeth). These activities involving hand gestures were included because they could be confused with food intake gestures and pose an additional challenge to the intake gesture classifier.

The annotation process involved coding food intake gestures using the same method employed in the the dominant wrist-mounted sensing experiment. Based on empirical observations, I considered each intake event to last 9 seconds, 2 seconds before and ending 7 seconds after the time the food reached the mouth. Like before, I used the ChronoViz tool [41] for annotating the video and data streams.

6.2.1 Results & Discussion

In total, participants performed 295 food intake gestures in the laboratory study. Of the total, 161 were performed with the right-hand, and 134 with the left hand (Figure 36). Interestingly, all participants claimed to be right-handed prior to the study. The charts in Figure 36 show, for each participant, the distribution of intake gestures by hand and eating activity type. While P1 and P2 were clearly right handed, P3 and P4 made extensive use of their left hand while eating. For P4, the left hand was used in hand-to-mouth gestures while the right, and dominant hand, was dedicated to cutting with a knife. Although the experiment was limited to only 4 participants, the right-hand–left-hand utilization ratio reveals how much the non-dominant hand gets used during eating. This suggests that monitoring food intake gestures by tracking the non-dominant hand with a smartwatch device is feasible.

The balanced use of both hands while eating is also reflected in the recognition of intake gestures. The graph in Figure 36 shows the effect of sliding window size on food intake gesture recognition performance (F-Score) as a function of wrist instrumentation across all participants in the lab study (leave-one-participant-out cross-validation). When best performance is achieved, with a window size around 60 seconds, instrumenting the left and right wrists proves to be superior to instrumenting only either one of the wrists, but only marginally.

It is worth noting that with a window size of 60 seconds, what gets modeled is not a single intake gesture, but a period of eating activity that encompasses multiple intake gestures and other non-eating gestures as well. When performing inference with a larger window size, examining eating moments from a more "holistic" perspective, it is to be expected that analyzing hand gestural data from both hands would lead to better results. On the other hand, with a shorter window size, the model that gets created with data from either one of the wrist-mounted devices is a better fit for individual intake gestures.

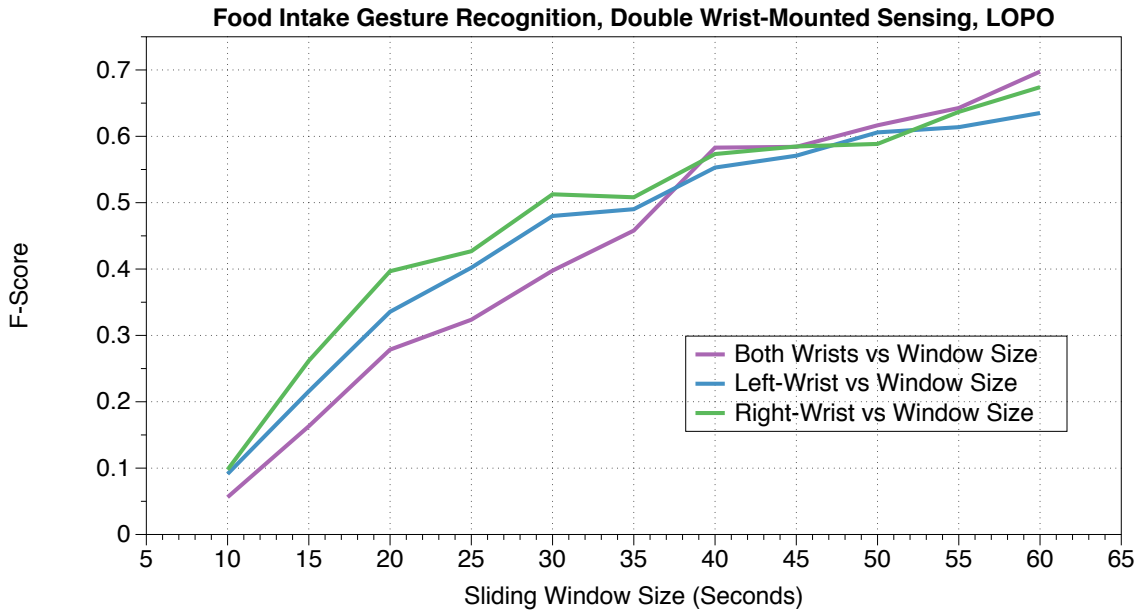


Figure 34: In this graph, it is possible to see the effect of sliding window size (SWS) on food intake gesture recognition performance (F-Score) as a function of wrist instrumentation across all participants in the lab study. When best performance is achieved, with SWS above 50 seconds, instrumenting the left and right wrists proves to be superior to instrumenting only either one of the wrists.

6.3 Fully Personalized Eating Detection Model

To reiterate, one the goals of this work was to assess whether an eating moment classifier can be tailored to a person from data compiled in laboratory and real world conditions. This was motivated by the observation that eating styles vary greatly between people. For example, some individuals bend down to eat, bringing their head close to the plate of food. Others eat while sitting upright, requiring their arm to traverse a longer distance to bring food to the mouth.

To answer this question, I conducted an in-the-wild study. The experiment had three phases. In the first phase, which was described in the previous section, 4 participants wore two wrist-mounted devices, one on each arm, and collected inertial sensor data in the lab while performing eating and non-eating activities. An eating activity detector was created for each participant. In the second phase of the study, the same 4 participants who completed the laboratory study were asked to wear the sensing devices on their wrists for

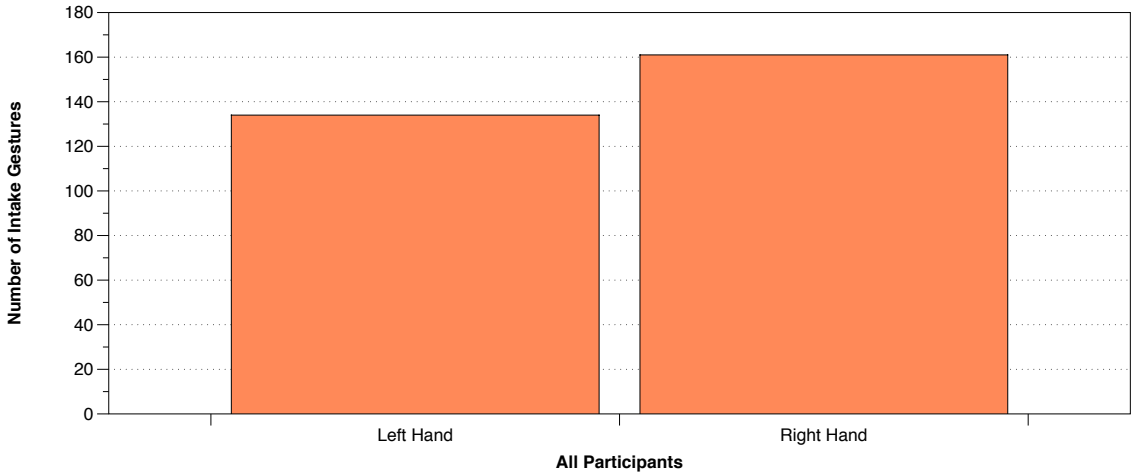


Figure 35: As shown in this chart, out of 295 food intake gestures performed by 4 participants in the laboratory study, 161 were performed with the right-hand, and 134 were performed with the left hand. Prior to the study, all participants claimed to be right-handed.

several days while they performed their normal everyday activities. Following this period of data collection in naturalistic settings, each participant's lab-trained eating activity detector was tested on the in-the-wild dataset. Finally, the third phase of the experiment involved building and evaluating personalized eating detection classifiers by combining lab and in-the-wild data in different proportions.

The Activiome system, described in the appendix of this thesis document, was used for ground truth data collection in-the-wild. Participants wore a wearable camera with a wide-angle lens that was programmed to take first person photos every 60 seconds. At the end of each photo capture cycle, the camera uploaded the images in real-time to the Activiome server. The photos portrayed participant's activities throughout the day, and were used as a memory aid to participants as they recollected and annotated their activities.

Participants were instructed to label the photographs and associated sensor data as eating when the image provided enough evidence that an eating activity was taking place. This could have been because a plate of food was visible in the image, or participants simply recalled eating food during that time. We provided some directions for how to complete the annotation process (available in Appendix B), but participants were asked to use their own judgment as needed. To minimize privacy concerns, study participants were the only ones

to see their own first-person point-of-view images, therefore their annotations could not be externally validated.

Table 18: The amount of time participants performed eating vs. non-eating activities in the wild according to their own photo-assisted annotations.

Participant	Non-Eating Time	Eating Time	% of Eating Time
P2	49hrs 26mins	1hr 42mins	3.43
P3	59hrs 4mins	55mins	1.55
P4	49hrs	6hrs 4mins	12.37

6.3.1 Results & Discussion

In total, 3 participants collected 166 hours and 12 minutes of inertial sensor data in real-world settings over a period of at least 5 days per person (Table 18). One of the participants failed to complete the annotation of images and his data was excluded from the study. The photo-aided annotations of eating versus non-eating activities were assigned with a resolution of one minute, the interval at which photographs were taken.

As described, the first phase of the study was completed in the laboratory study and consisted of building a personalized eating detector for each participant. To evaluate the eating detectors, the in-the-wild data for each participant was split into 5 segments. Precision, Recall and F-Score measures were then calculated for the eating detectors under 5 evaluation sessions as shown in Table 19.

The evaluation was structured this way to test how a model trained with increasingly more data from one participant performs at recognizing eating moment for the same participant. To illustrate this process and starting with Session 0, the model is trained with lab data only and is evaluated with all in-the-wild data segments. In Session 1, the model is trained with lab data plus one of the in-the-wild segments and is tested on the remaining in-the-wild segments. This pattern repeats, with the trained model incorporating increasingly more data for one participant, until there is only one segment left for evaluation.

Figure 37 shows how the personalized model for P4 performed when trained with increasingly more personal data acquired in-the-wild. Precision followed an upward trajectory

up to Session 3 at slightly over 40%, and then dropped close to 10% in Session 4. Recall never rose above 20%.

Table 19: The amount of time participants performed eating vs. non-eating activities in the wild according to their own photo-assisted annotations.

Session	Training Data	Evaluation Data
0	Lab	Segments 1 + 2 + 3 + 4 + 5
1	Lab + 1	Segments 2 + 3 + 4 + 5
2	Lab + 1 + 2	Segments 3 + 4 + 5
3	Lab + 1 + 2 + 3	Segments 4 + 5
4	Lab + 1 + 2 + 3 + 4 + 5	Segments 5

Except for P4, the personalized models performed very poorly, with F-Scores under 10% throughout all sessions. I hypothesize that the subpar P2 and P3 results were due in large part to the low ratio of eating vs. non-eating activities, as shown in Table 18. For example, only 1.55% of the data compiled in the field by P3 was annotated as an eating activity; considering that eating-related sensor data was acquired in the laboratory for no more than 22 minutes on average (Table 17), it is possible to see that there is in fact very little data for training a personalized model.

Another explanation for the underwhelming results is that participants might have annotated the ground truth labels incompletely. Due to privacy concerns, the study protocol prevented me from seeing the first-person photographs captured by participants, and used for ground truth annotation. Therefore, there could have been problems in this stage of the process.

Finally, I also examined how a personalized model trained with lab data for one participant compares to a model trained with lab-data for all participants (personalized+all). As shown in Figure 38, the F-Scores for both models is low, but the personalized model consistently outscores the personalized+all model. This result suggests that in the context of eating detection, with reduced training data, it might be best to train a model with personal data only versus with all the training data available. But additional studies with a larger number of participants must be conducted to validate this hypothesis.

In conclusion, due to the poor results in these experiments, it is difficult to derive

any significant findings with regards to model personalization with wrist-mounted inertial sensors. This is especially true considering the small number of participants in the studies and the singular laboratory session per participant whose data led to the creation of the models.

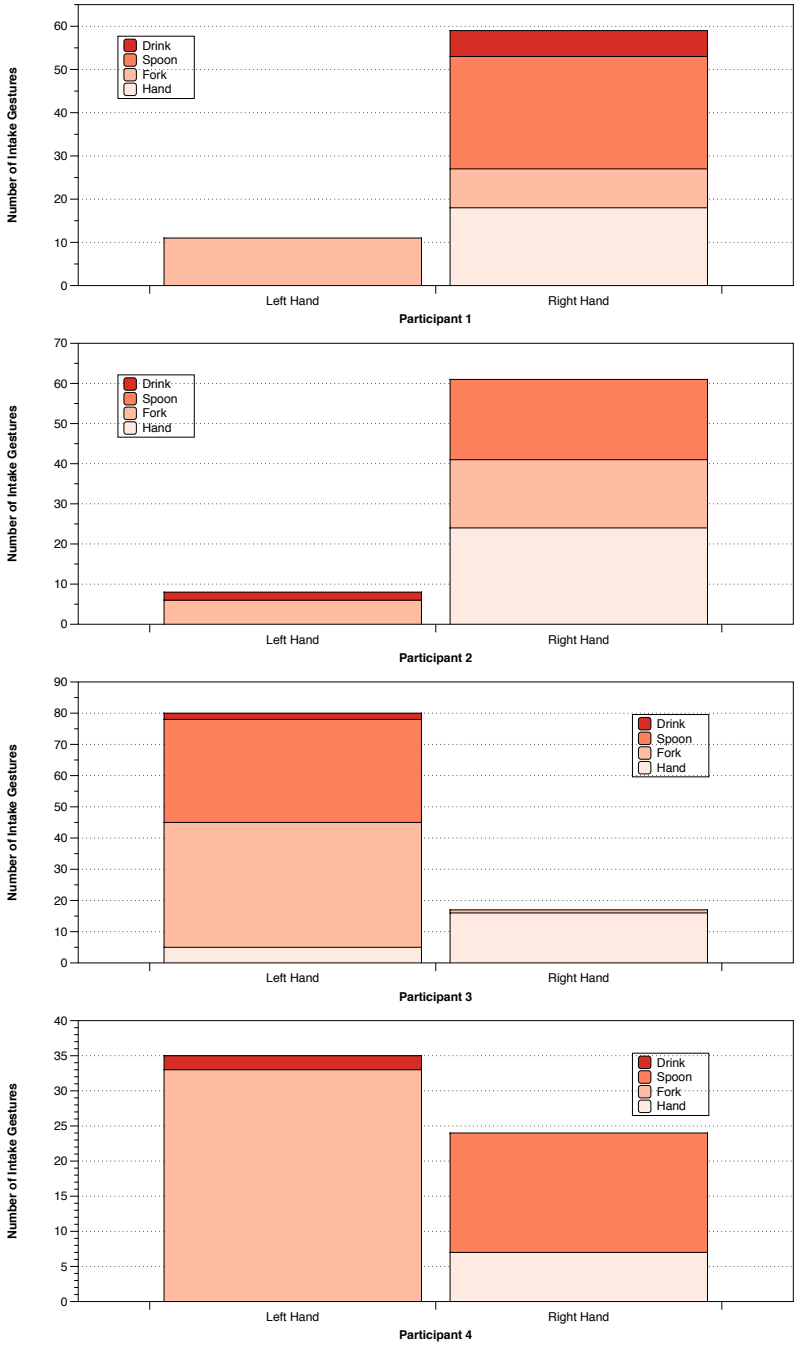


Figure 36: The charts above show, for each participant, the distribution of intake gestures by hand and eating activity type. While P1 and P2 were clearly right handed, P3 and P4 made extensive use of their left hand while eating. For P4, the left hand was used in hand-to-mouth gestures while the right, and dominant hand, was dedicated to cutting with a knife.

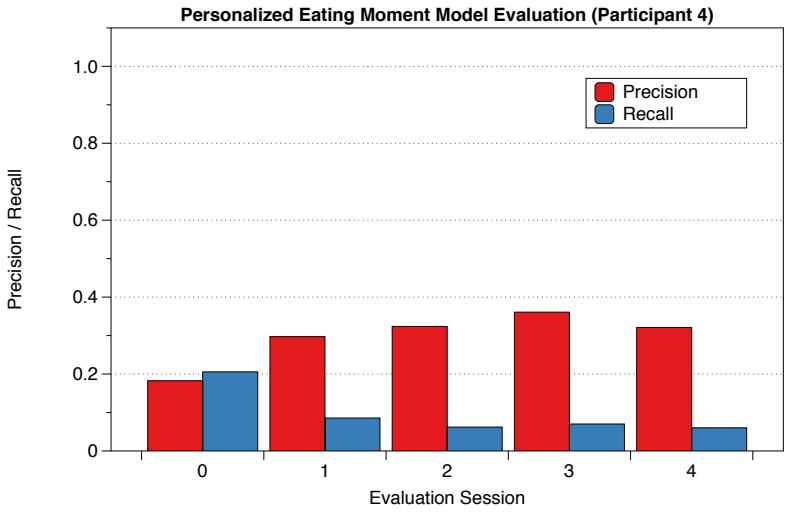


Figure 37: Precision and recall measures for the personalized eating detection model for P4. The data combination used for each evaluation session can be found in Table 19.

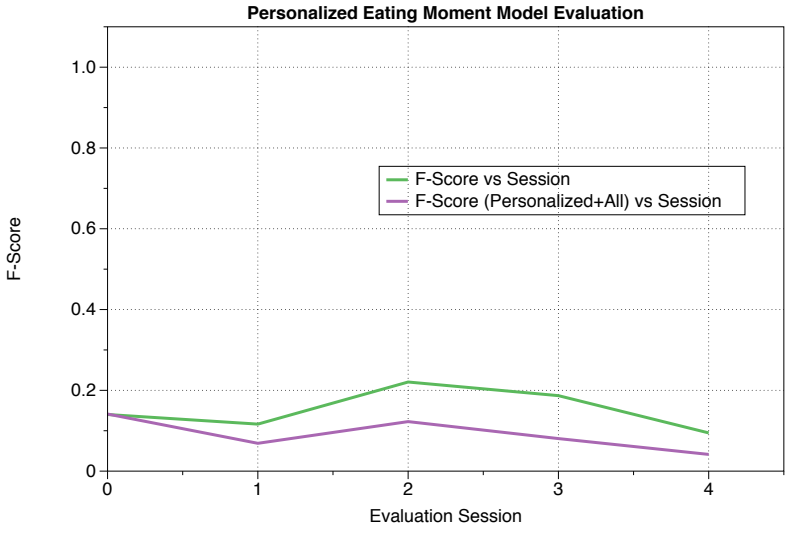


Figure 38: Performance comparison between P4's personalized eating detection model versus a model trained with all participants' lab data.

CHAPTER VII

CONCLUSION AND FUTURE OPPORTUNITIES

With the goal of defending the thesis that **everyday eating moments can be automatically detected in real-world settings by opportunistically leveraging sensors in practical, off-the-shelf wearable devices**, the work I present in this document touched on a variety of research contributions around the study and evaluation of three sensing modalities for eating moment detection: first-person images, acoustic sensing and inertial sensing. In total, I conducted 2 laboratory studies and 6 in-the-wild studies with 106 participants, which resulted in 5 conference publications and the release of public datasets that other researchers can leverage to both validate and extend my work [127, 129, 130, 22, 128].

I first discussed eating moment detection with first-person point-of-view images taken with wearable cameras. Photographs automatically shot at regularly-spaced time intervals throughout the day represents one of the best ways to capture the richness of everyday activities without requiring direct human feedback. To examine the potential of first-person point-of-view images in eating detection, I conducted two studies. In the first study, I used human computation to identify eating activities in photographs. The second study had the same objective, eating detection, but I employed a combination of computer vision and machine learning techniques as opposed to human computation.

Despite promising results, a difficulty that emerges with first-person photographs taken in naturalistic settings is privacy. Pictures taken automatically with on-body cameras might result in the recording of undesirable moments and scenes. To make matters worse, photos taken of computer screens might capture sensitive information such as computer passwords and credit card numbers. These problems are amplified when these photographs are examined with human computation services like Amazon Mechanical Turk, which are populated by individuals whose real identities are unknown. The process of understanding the privacy implications of these types of images and evaluating computational techniques for

minimizing them led to the development of the Privacy-Saliency Matrix framework.

Leveraging context-rich first-person images while minimizing privacy concerns motivated he study of an alternative inference technique; the method uses metadata and computer vision features to classify images without human input. In particular, the technique leverages a machine learning method, convolutional neural networks (CNN), that has been lately shown to perform well at image recognition tasks. A performance analysis was done for eating detection while also examining the approach's ability to recognize a much larger set of everyday activities in real world settings. This method proved to work quite well both in the general and in the personalized cases.

Images reflecting everyday experiences are compelling, but one must continuously wear a camera in order to compile a meaningful set of photographs portraying daily life. In the interest of practicality, I investigated whether eating moments can be inferred through the sensing capabilities of more practical devices such as mobile phones, smartwatches, and other wearable technologies. In a feasibility study in real world settings, I implemented and evaluated a system that recognized eating moments from ambient audio. Participants wore a wrist-mounted audio recorder that captured audio of their everyday experience throughout the day. Results were positive, and demonstrated that identifying certain acoustic signatures of eating might be one way to infer eating moments, while making use of one of the most ubiquitous sensors: a microphone.

Over the last decade, inertial sensors have become commonplace and are now an integral part of personal devices, from phones to activity trackers. A large portion of my dissertation work focused on the use of devices imbued with inertial sensors to detect food intake gestures and eating moments. I built recognition systems for detecting eating activities and evaluated them with a series of studies with human subjects. One experiment looked at the system's ability to detect eating from a head-mounted inertial sensor. Others centered on intake gesture and eating moment detection from the wrist. Inference with wrist-mounted devices was evaluated both in a laboratory setting and also in the wild, and I also examined the impact of having gestural data from one wrist (more practical) versus both wrists (less practical). One of the highlights of the inertial sensing analysis was the exploration of

whether a model trained in the lab can be successfully used in naturalistic conditions. This strategy is highly compelling since acquiring and annotating real world data is a difficult and time-consuming undertaking. Although more studies are needed, my results showed that this is indeed possible.

Overall, I found inertial sensing to be a highly desirable and practical modality for eating detection. There is a very direct link between the physical body movements involved in eating (e.g., hand-to-mouth gestures), and the types of measures that can be obtained with inertial sensors. Moreover, it is likely that we will see a rapid evolution in inertial sensing technology in the next several years, as more powerful inertial measurement units (IMUs) become available and are miniaturized to be integrated in personal devices.

The wrist proved to be a particularly good location for sensing eating activity, since that is exactly where most people wear smartwatches and activity tracking devices today. Despite promising results with head-mounted sensing, I found it challenging to discriminate eating-related head movements from other types of head motions. Another limitation of head-mounted sensing is the need for instrumenting the head with sensors; there is currently not a practical and socially-acceptable way to realize this instrumentation.

7.1 Performance Results and Applications

In chapter 1, I motivated the need for automatic eating detection with applications in four domains: population health, nutritional epidemiology, dietary self-monitoring, and patient and elder care. After exploring a variety of sensing modalities and approaches, it is now possible to ask whether the performance of the eating detection systems presented in this dissertation satisfy the requirements imposed by these applications.

To answer this question, it is useful to distill the motivating applications in two categories. The first category is characterized by applications of eating detection where eating is a matter of concern for the individual who is performing the eating activities, such as dietary self-monitoring. The second category includes applications where eating detection is applied to assist one or more individuals, typically researchers, track and understand the eating habits of other individuals. Population health and nutritional epidemiology are some

of the motivating applications.

Starting with the first category, it is widely known that dietary self-monitoring is one of the most effective methods for weight control [20, 12]. As individuals recall and log what they consume, they become aware of foods eaten and often change their eating habits towards healthier food choices. Therefore, the fundamental element of dietary self-monitoring is the self-reflection process triggered by the journaling task, which cannot be measured in terms of accuracies and F-scores the same way as an automatic eating detection system or approach. However, automatic eating detection is not irrelevant to dietary self-monitoring. In fact, the opposite is true; in a study examining barriers to food journaling, we noted that forgetting to journal is one of the major barriers to reliable journals [30]. Food journalers reported that missed entries discourages logging and causes them to abandon journaling altogether.

Emerging, semi-automated journaling approaches that combine manual logging with automated support (i.e., eating detection) show promise as a way to facilitate dietary self-monitoring. For these applications, it is undeniable that eating detection performance close to 100% would be preferred. However, since these semi-automated approaches feature some level of end-user involvement, a degree of inaccuracy might be tolerated if individuals can be prompted for verification and correction of low-certainty inferences.

The second category of applications is centered on a model where researchers or caregivers track the eating habits of individuals and populations. As previously discussed, this tracking model is based on survey instruments that have been deemed unreliable. As a result, tools and methods that can inject some level of objectivity into the food tracking process are desirable. In terms of performance, the perfect scenario would again involve an approach that can detect eating moments without any errors. Unfortunately, this is unrealistic in practice. To make matters worse, having participants correct recognition errors might not be an option since individuals are not personally invested in the data collection process for this class of applications.

Reviewing study results, best eating detection performance was observed with first-person images and inertial sensors. With first-person images, eating moment detection

with 83.12% accuracy was achieved. Although this result was obtained using data for just one person, I showed evidence that the corresponding technique can be successfully generalized to other individuals. One advantage of this method is that it operates on data that is immediately available once an image is captured: the image itself and timestamp information. Leveraging this data makes the eating detector suitable for real-time or near real-time applications, such as just-in-time interventions. However, a significant limitation is the need for individuals to continuously wear a portable camera. This requirement lowers the practicality of the approach.

Inertial sensors on the wrist offer a good balance in terms of performance and usability. With wrist-mounted devices, best eating detection performance was achieved with a F-score of 76.1% in real-world settings. This result is very promising considering that it requires only one off-the-shelf smartwatch device placed on individuals' non-dominant hand. However, a caveat of this result is that it hinges on the examination of one hour of sensor data preceding the moment of inference. This limitation renders the approach unsuitable for situations when it is critically important to detect eating during or immediately after an eating moment, such as for just-in-time interventions and also when researchers are evaluating dietary self-tracking techniques (e.g., to study how the timing of journaling affects the effectiveness of food logging).

One strategy for applying imperfect automated eating detectors in nutritional epidemiology and population health studies is to combine them with other validated instruments. By triangulating results originating from multiple survey methods, it is often possible to attenuate inaccuracies. This technique is already used today, and could be expanded to include computational instruments and classifiers such as the ones described in this document. In the machine learning community, this approach is known as an "ensemble" method and hinges on the implementation of "weak" classifiers. These classifiers perform poorly, with results just above average, but it has been shown that ensembles of weak classifiers can often perform better than any single classifier. Most of the eating detection classifiers presented in this dissertation performed significantly better than average, but they can be adapted or "weakened" as required to fit into an ensemble configuration with other classifiers and

survey methods. Therefore, these eating detection classifiers presented can indeed add value to health applications today and continue to do so over time as their performance improves.

7.1.1 Snacking Behavior

One topic that often arises in discussions involving automatic dietary monitoring is how systems perform at detecting snacking activity. Before this question can be answered, it is useful to first establish what a snack is. Unfortunately, even nutritional epidemiologists disagree on an exact definition. The dictionary[1] describes it as "a small amount of food eaten between meals". However, in practice, eating small amounts of foods throughout the day has become a habit for many individuals to the detriment of traditional sit-down meals.

For thousands of people, if not millions, snacking has become synonymous with eating. What makes snacking particularly hard to identify is that it occurs within a short time window; snacks are often bite-sized foods and consumed relatively quickly. A cup of yogurt, an apple, and a granola bar are often considered to be snack-type foods. In my laboratory studies, I incorporated these kinds of foods in eating sessions with the goal of modeling a variety of forms of eating. Thus, my study design took snacking into account, and my results incorporate the impact of snacking on eating detection.

One scenario I did not investigate in my eating detection studies is the consumption of very small quantities of food, such as grabbing and eating only a handful of popcorn kernels or M&Ms. To identify these very short eating moments with inertial sensing alone, it is necessary to spot individual food intake gestures with high accuracy. Since the inertial-based eating moment detection approach I proposed hinges on the discovery of *clusters* of food intake gestures, it was not designed for this condition. This insight is evident from the clustering algorithm parameters that optimized results for the Wild-Long study; at least 3 predicted intake gestures within 40 seconds were needed for an eating moment to be recognized. Due to the difficulty of detecting every single intake gesture with high certainty, I believe the key to recognizing very short eating moments lies in multimodal sensing. Although further research is necessary, the combination of audio and inertial

[1] http://www.oxforddictionaries.com/

sensing seems particularly promising for this scenario.

7.2 Future Opportunities

In this section, I outline opportunities that I have identified for expanding and improving upon the automatic eating detection work presented in this document.

7.2.1 Multimodal Sensing

This dissertation presents results showing how different sensing modalities fare at the task of eating moment detection. Amongst others, I showed how wrist-mounted inertial sensing and first-person photographs can be successfully used as indicators that an eating activity is taking place. However, one direction I did not explore in my work, which represents a large opportunity to improve eating detection inference, is multimodal sensing. In other words, combining eating evidence from gestures, environmental sounds and other contextual sources such as location, time of day and even appointments in an individual's electronic calendar, is very likely to result in estimates that surpass those obtained with individual sensing modalities alone. Albeit simple conceptually, much work lies ahead when it comes to understanding how to best model activities in light of multiple streams of data, and whose types are so different from each other.

7.2.2 Personalized Models

Although one might be led to believe that the problem of eating detection can be solved by identifying the canonical hand-to-mouth gesture, in reality there is enormous variability in how individuals eat. This problem of intra-class diversity in food intake gestures is illustrated in Figure 30 of Chapter 5. Given the existence of intake gesture styles that diverge from person to person but that are stable for an individual, building eating moment detection models that are personalized is a natural way to proceed. This personalization might be implemented either at the level of one person, or by eating style cohort.

Additionally, and beyond intake gestures, individuals typically adopt habits that also remain stable over time. People usually eat around the same times and in a relatively small number of locations. When considering the opportunity to personalize in a multimodal

context, the notion of tailoring a model to a person or group of individuals becomes even more powerful. Clearly, one of the challenges of personalization is acquiring enough data for just one person to make personalization possible. Many opportunities related to acquiring training data exist, some of which are described in the sections below.

7.2.3 Modeling Full Set of Eating Gestures

In the work presented in this dissertation, I gave the problem of eating detection the same treatment that is commonly applied to identifying other activities such as standing, running and sitting. In reality, eating is a more complex, multifaceted activity. While there might only be a frequency difference in the sensor signal whether a person is running quickly or slowly, it is possible to consider two activities that are widely regarded as "eating" but that are very different from each other. My studies made evident how arm, wrist and hand gestures differ significantly whether people are eating with a spoon, fork or holding food with their bare hands.

Therefore, there is an opportunity to piece apart the many gestures commonly associated with eating and build specific classifiers for them. In this scenario, "eating" is not modeled as one activity but at a lower-level and through multiple classifiers, each taking a "gestural" aspect of eating into account. In practice, recognizing an eating activity would involve querying various classifiers and combining their output. This strategy could be put in practice through majority voting or by combining probability distributions. I believe this approach could lead to improved eating detection accuracy since classifiers would be specifically tuned to relevant eating gestures.

7.2.4 More Powerful Features and Representations

When converting the stream of inertial sensor data into a feature representation, I employed a sliding window and extracted frames from the signal. I calculated a traditional set of statistical features for each frame, including mean, variance and kurtosis. Although these features have been successfully used in activity recognition and been shown to serve as a compact representation of the underlying data, I am certain that using more sophisticated, and domain-specific features will have a positive impact in performance. For instance, I

expect that methods such as Dynamic Time Warping (DTW) and features that involve wrist rotation will positively impact the accuracy of eating detection.

In the context of features and representations, one area that is also worth noting is sensor complexity. As more powerful sensors become available, and continue to be embedded in consumer electronics devices, sensor data streams are likely to change as well. For example, the Invensense MPU-9150 IMU used in RisQ [102] outputs a 3D orientation in the form of a quaternion. These new representations will certainly contribute to improved inference, but they will also demand new features and new ways to process the data.

7.2.5 Improved Annotation Methods

There is no question that one of the most significant challenges of automatic eating detection is to build a system that works in real-world settings. However, in order to train a system to work well in naturalistic settings, it is necessary to compile realistic training data, ideally from real world conditions as well. The hurdles of obtaining annotated ground truth data in the field has been discussed in this document. In fact, this issue is what motivated the development of the Activiome system and the use of wearable cameras in many of the in-the-wild studies.

Unfortunately, capturing photographs every 30 seconds or so is not enough for finely-grained annotations at the level of food intake gestures, for example. A continuous video recording would be more effective for annotation, but at the expense of much more storage capacity and battery consumption. These increased resource demands render video capture impractical for continuous data collection in real-world conditions. Even if continuous video capture were possible, it would require tremendous effort during annotation, since it would take significantly longer to review and label video events than a set of photographs. Thus, improved data collection and annotation methods are needed.

7.2.6 New Model Learning Approaches

The previous section discussed the issue of collecting and annotating training data. Training data is required to build a model using supervised machine learning techniques, the predominant approach for building activity recognition classifiers. However, I am confident

there is a large opportunity in exploring how one might build an activity classifier, such as a food intake gesture recognizer, without relying so much on previously acquired training data. Techniques such as semi-supervised machine learning and active learning might offer an alternative to the traditional supervised method. One direction that is also worth investigating with more depth is that of transfer learning. The dominant wrist-mounted sensing study results showed that it is possible to train a classifier with data compiled in a semi-controlled laboratory setting, instead of having to acquire all training data in real-world conditions.

7.3 Final Thoughts

Building a truly generalizable system for eating moment detection, and automatic food intake monitoring in general, represents a significant challenge. I believe such a system could provide the foundation for a new class of practical applications, benefiting individuals and health researchers. Despite limitations and opportunities for improvement, I believe the work outlined in this document provides compelling evidence that a practical solution based on commodity sensing can play an important role towards this vision.

APPENDIX A

THE ACTIVIOME SYSTEM

One of the biggest challenges of building activity recognition systems that work in real world settings is that training these systems using supervised machine learning techniques requires large amounts of labeled ground truth data. The difficulty of collecting and annotating said data is well known, and has been discussed in the "Techniques for Estimating Ground Truth in Real World Settings" section of this dissertation.

Directly observing an individual is considered to be the best method for compiling a log of activities performed in-the-wild, but it is often not practical and could, in principle, alter the person's natural behavior. An alternative method involves instrumenting individuals with a wearable camera that captures front-facing photographs at regular intervals (e.g., every minute) throughout the day. In this configuration, the photos taken by the camera are rich in contextual detail, showing individuals perform everyday tasks. This is the primary method I chose for estimating ground truth in naturalistic settings for many of the studies described in this document.

To facilitate the acquisition of sensor data and ground truth labels based on images taken with first-person cameras, I developed a system around a mobile phone application, a backend server database, and a web application called Activiome. Although there are commercial wearable cameras designed for capturing everyday experiences, they do not offer programmatic access and configurability. Moreover, photo capture constitutes just the first step of the annotation process. A mechanism that allows individuals to review and label their own photos is as critical as the photo taking process itself. The sections that follow describe the Activiome system in detail.

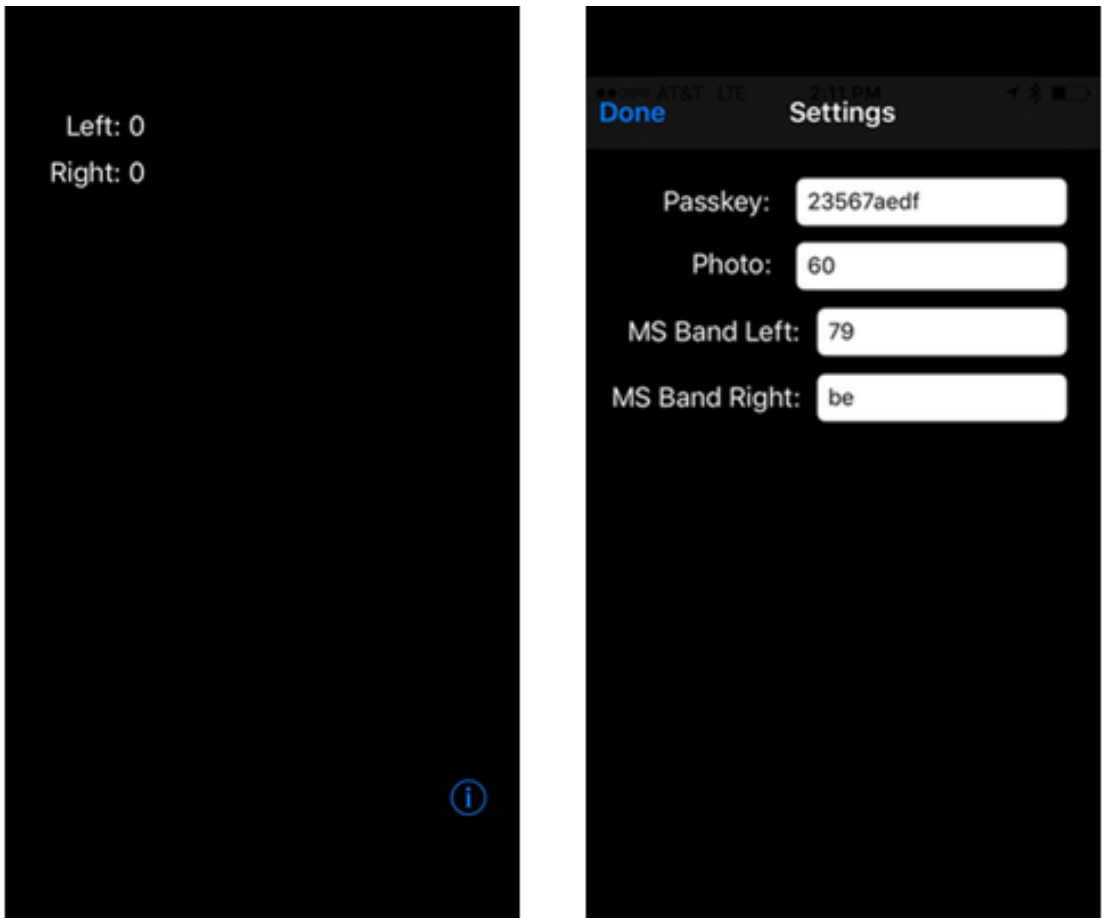


Figure 39: The Activiome mobile application user interface. The main screen, on the left, has a dark background and features an indicator of how much inertial sensor data was received in the last cycle. The settings screen on the right is used to configure parameters and sync up the mobile phone, the sensing device and the Activiome user account.

A.1 Mobile Application

The mobile application, designed for the iOS and thus iPhone, is a key element of the Activiome system and accumulates many functions. These functions are performed as part of a data acquisition cycle that is illustrated in Figure 40:

- **Captures first-person point-of-view photographs at regular intervals**: The mobile application is programmed to take a photo with the back-facing camera at a user-configurable interval. In other words, the mobile application repurposes the phone as the wearable camera, and captures the first-person point-of-view photographs that are later used for estimating ground truth activity labels. Consequently the phone

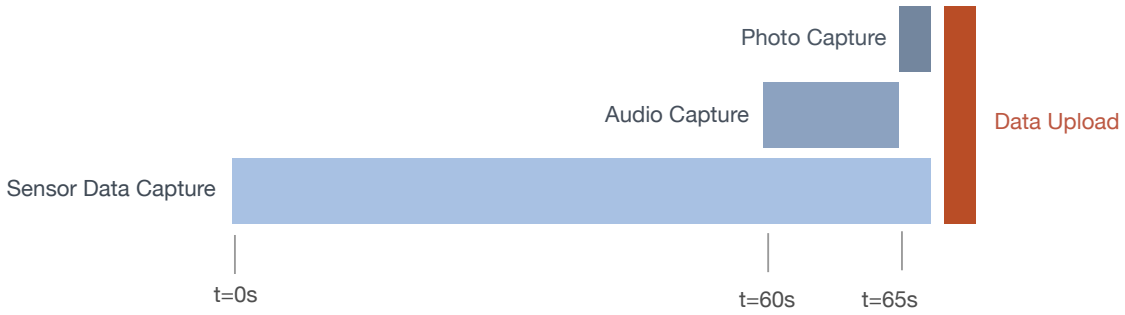


Figure 40: The data acquisition cycle of the Activiome mobile app. In this example, the cycle is set to 60 seconds. After 60 seconds, the app first captures 5 seconds of audio and then takes a picture. Inertial sensor data is captured throughout the entire cycle. At the completion of data acquisition, the data is packaged as a HTTP POST request and uploaded to the Activiome server. This cycle repeats until the application quits.

running the application should be the one that individuals wear on a lanyard around the neck. Additionally, the mobile application leverages the location tracking capability of the phone itself to geotag the images with latitude and longitude metadata.

- **Record a 5-second audio clip prior to photo capture**: To provide more context about the activity and setting recorded by the photograph, the mobile app also records a 5-second audio clip immediately prior to capturing an image.

- **Collect and store inertial sensor streams**: The mobile application interfaces with devices such as the Pebble watch and the Microsoft Band to record inertial data wirelessly using the Bluetooth protocol. This interface is possible through the SDKs provided by the developers of these respective devices. Data collection begins immediately after the application launches and takes place uninterruptedly even while the 5-second audio clip is being recorded. The sensor data is sent to the server and also saved locally on the mobile device as flat files following the iOS Property List format.

- **Upload data to server**: At the end of each cycle, the sensor data, metadata, photo and audio clip are uploaded to the server as one HTTP POST request.

The interface of the mobile application can be seen in Figure 39. The UI was intentionally designed to be black with the goal of reducing battery consumption, since the application needs to be running continuously throughout the day to perform the aforementioned tasks. The main screen features a settings button and two indicators that provide feedback about how much sensor data is being recorded by the phone. The settings screen is used to set up communication between the sensing device, the mobile application, and the Activiome server.

A.2 Backend Server Database

The backend infrastructure of Activiome was designed around well-established web technologies; it centers around a set of PHP scripts and a MySQL database. A PHP script is called by the mobile application with all the collected data in the HTTP POST, and proceeds to parse and validate it. A new entry is created on the database and is populated with the sensor data, metadata (i.e., geo-location and timestamps), and links to the audio and image files.

A.3 Web Application

Individuals interact with their data using the Activiome web application. Prior to data collection, study participants create an account on the system with a username and password such that they are the only ones with access to their own sensor data and, more importantly, their photographs (Figure 41). Once individuals log in, they see a list of their most recent data entries for the day. Each entry on the web app interface, which maps to an entry on the database, includes a first-person photo, the audio clip and a graphical representation of the sensor data (Figure 42). It is also possible to browse activities of previous days by changing the date.

Typically, the reason why participants log onto the system is to perform annotations of the data. Using the Activiome web application interface, study participants can review the images, listen to the audio clips and recall their activities at the time. A drop-down menu is available for each entry, and participants select an item to indicate an activity out of a pre-defined activity list. For eating moment detection, for instance, all participants needed

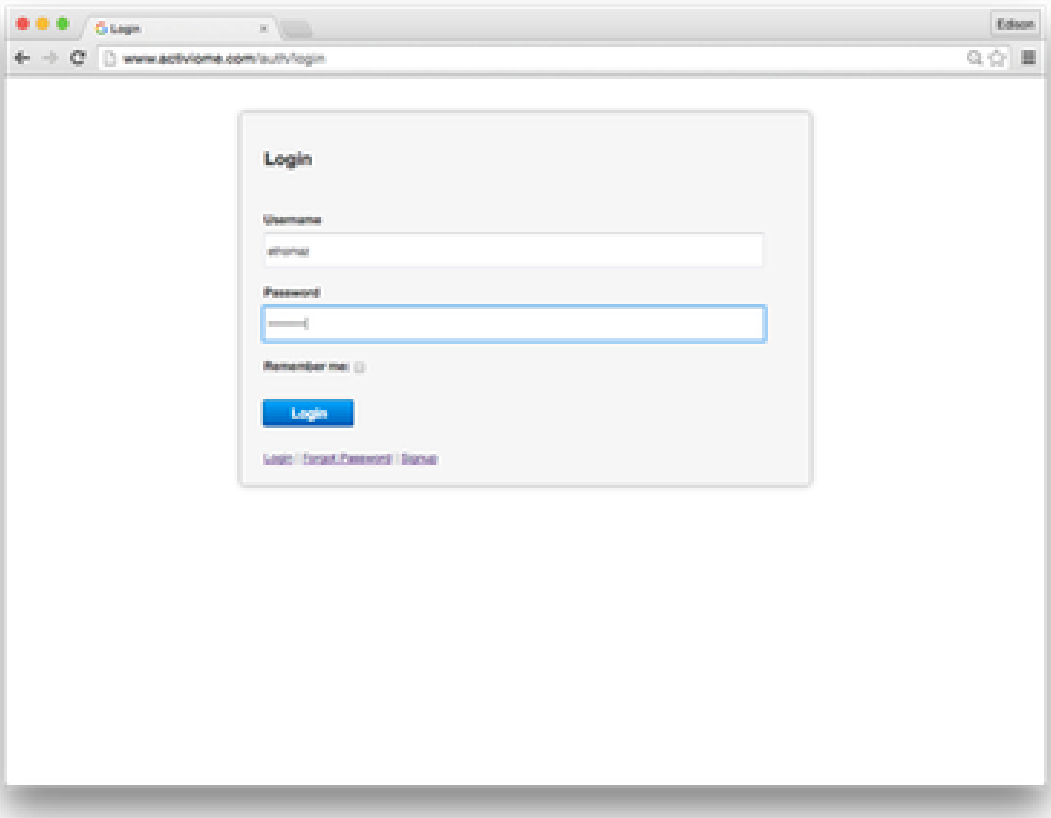


Figure 41: The Activiome web application login screen. Each participant creates a personal account on the Activiome system and can login to review and annotate the acquired data.

to do was to indicate which images portrayed themselves during an eating activity.

When the Activiome mobile application is running and set to record photographs every minute, a large number of images and associated data are recorded every single day. Associating activities with individual images becomes a time-consuming and tedious process. To aid the ground truth labeling process of large photo collections, the Activiome web application also offers a mosaic view, where thumbnails of all first-person point-of-view photographs taken on a given day are shown together (Figure 43). Using this view, participants can select multiple photos at a time using the keyboard or mouse and annotate all of them at once, reducing the time required for image labelling significantly.

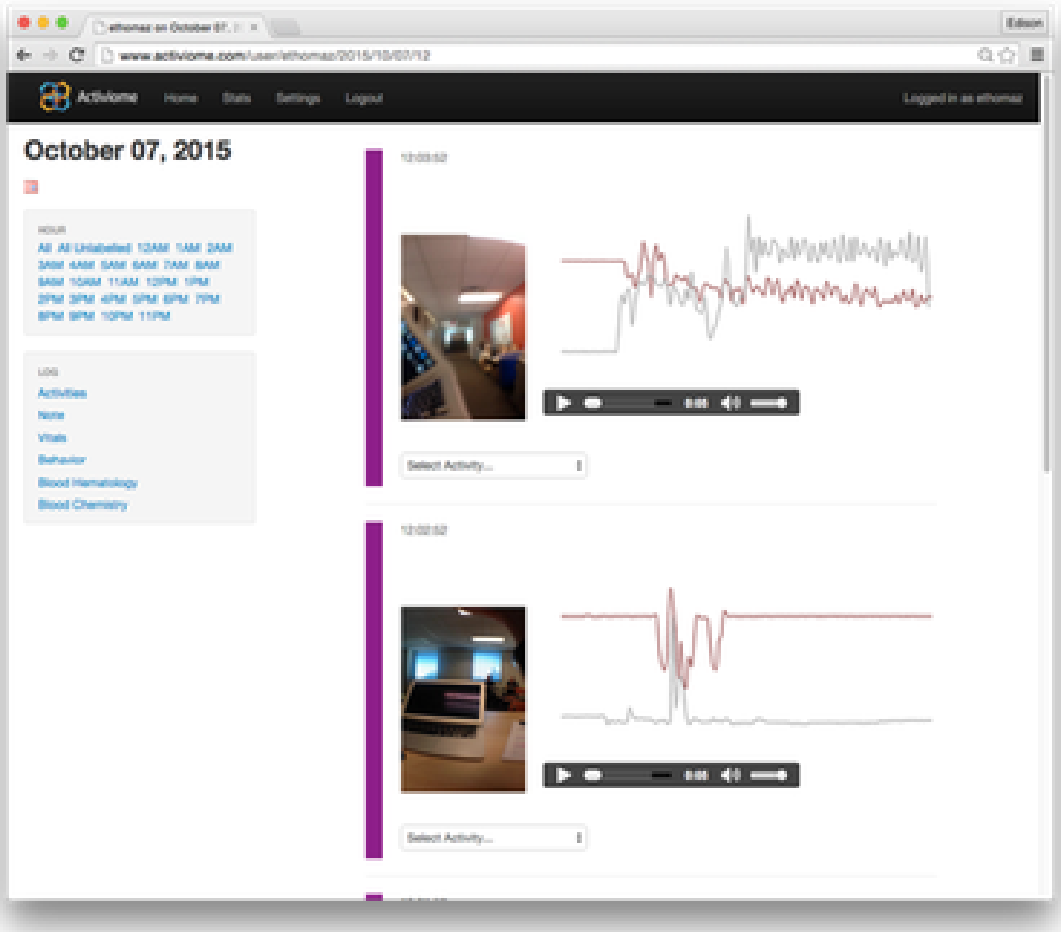


Figure 42: The Activiome main screen. Once participants log in, they are shown a detailed list of recorded activities for the most recent hour

A.4 Performance Considerations

Considering the workload of the Activiome mobile application, it is not surprising that battery consumption is a serious concern when it comes to recording first-person photos, audio clips and other forms of data for many hours at a time. My experiments showed that lab studies lasting shorter than one hour did not present any battery utilizations problems. For instance, one lab study wherein sessions lasted 31 minutes and 21 seconds on average, battery performance was never an issue. In this case, the data capture setup employed a Pebble smartwatch and an iPhone 4S. Smartwatch accelerometer data was captured at 25Hz and transmitted to the smartphone every second using Bluetooth. The sensor data

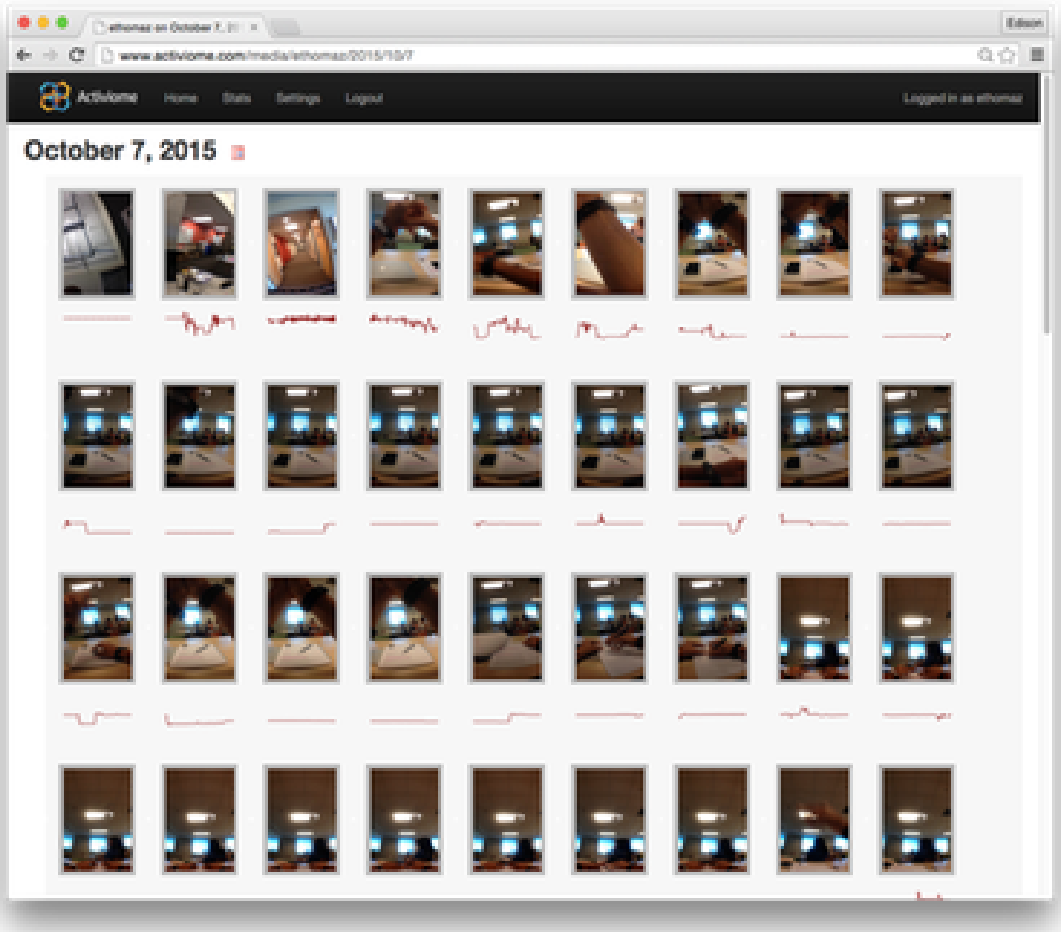


Figure 43: The Activiome mosaic screen. To facilitate annotating large numbers of images, I created a photo view that shows thumbnails of the captured first-person point-of-view images and makes it easy to select and label multiple images at a time.

was saved locally on the phone and retrieved at the end of each session.

On the other hand, long in-the-wild studies posed a significant challenge in terms of power consumption. In one study, starting on a full charge, the smartphone collected data continuously for an average of 5 hours and 42 minutes without problems. However, for a 31-day in-the-wild study that aimed at recording data for as long as possible during the day, an additional battery pack had to be connected to the phone. Carrying the battery pack proved to be an additional inconvenience, but it allowed data collection to take place for the entire day. Throughout these studies, the smartwatch, the smartphone and the battery

pack were restored to full charge overnight and used again the following day. The Pebble watch never represented a limiting factor in data collection; this was most likely due to the low power consumption of its e-ink display and lack of a more sophisticated inertial measurement unit (IMU).

APPENDIX B

STUDY MATERIALS AND PROTOCOLS

The documents that follow represent supporting materials produced in support of the user studies presented in this dissertation. Some of these documents, and particularly the consent forms, were prepared to be applicable and shared across studies.

- Day reconstruction form for ambient audio sensing study
- Instructions for operating iPhone and Pebble watch devices
- Instructions for operating iPhone and Microsoft Band devices
- Instructions for annotating first-person images using Activiome system
- Consent form for one-day sensor data collection in-the-wild study
- Consent form for multiway sensor data collection in-the-wild study

Participant #:

Age: _____ Profession: _____

Please describe your activities since the study began. Think of your activities as a continuous series of scenes or episodes in a film. Give each episode a brief name, for example, “commuting to work”, or “at lunch with B”, where B is a person or a group of people. Write down the approximate times at which each episode began and ended.

The episodes people identify usually last between 15 minutes and 2 hours. Indications of the end of an episode might be going to a different location, ending one activity and starting another, or a change in the people you are interacting with. There is room to list 20 episodes , although you may not need that many, depending on your day.

	Episode Name	Began	Ended
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			
11.			
12.			
13.			
14.			
15.			
16.			
17.			
18.			
19.			
20.			

Operating the Pebble watch and iPhone

Prerequisites

1. Knowing how to operate a Pebble watch, including launching apps
2. Knowing how to operate an iPhone, including launching and quitting apps

To start a new logging session (e.g., at the beginning of the day)

1. Make sure the Activiome app on the phone is *not* running. If so, quit it.
2. Launch the Pebble app on the phone
3. Launch the Activiome app on the phone
4. Start the pebble_live app on the Pebble watch.

If the “ack” number on the pebble_live increments every second, then everything is working correctly.
If instead the “faild” number is going up, finish the logging session as described below and try again.

If you continue having problems, reboot the iPhone, the Pebble and try again.

(To reboot the Pebble, hold the button on the left and the middle button on the right for 10 seconds. When the Pebble reboots you will see the logo (just “Pebble” really) on the screen for a second or less. The booting process is quick)

If that doesn’t do it, please get in touch with Edison at 617-733-6215 or ethomaz@gatech.edu.

To finish a logging session (e.g., at the end of the day)

1. Quit the Activiome app on the phone
2. Quit the Pebble watch app on the phone if it is running
3. Quit the pebble_live app on the Pebble watch (press the button on the left side of the watch)

What if the phone or Pebble runs out of battery mid-day?

No problem. If you are going to be collecting more data, plug in the device(s).
When the device(s) come back to life, finish the logging session as described above.

What if the “ack” number on the pebble_live stops incrementing?

The app on the phone and the Pebble are having problems communicating with each other. Follow the instructions for finishing a logging session above and immediately start a new one (as described above as well).

Operating the Microsoft Bands and iPhone

Pre-requisites

1. Knowing how to operate an iPhone: including powering on/off phone, launching and quitting apps, check Bluetooth settings
1. Knowing how the basics of how to navigate the MS Band watch interface, turn band on and off

At the beginning of the day

1. Turn phone on
2. Launch the Activiome app on the phone
3. Make sure the MS Bands are on (press one of the button to light up the display)
4. Wait 1 or 2 minutes
5. Check the Left/Right numbers on the main app screen
    - If they are both higher than 0, you are set.
    - If they are still at zero, try:
        (1) Reboot phone, turn bands off and on and repeat steps 2-5. If it works, you are set. If not, read below.
        (2) Check if bands and phone are connected:
            1. Go to phone Settings>Bluetooth, make sure there are four connected items under 'My Devices'. If yes, try steps 2-5 one more time.
            2. If there are unconnected devices, tap on them to connect. After connecting them, try steps 2-5.

If you are still having problems, please get in touch with Edison at 617-733-6215 or ethomaz@gatech.edu.

At the end of each the day

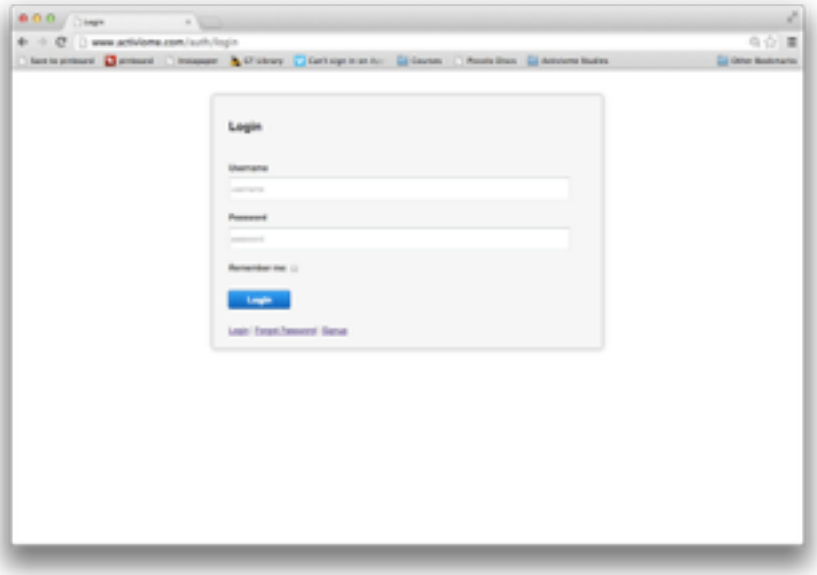
1. Quit the Activiome app on the phone
2. Turn phone off (you don't need to turn bands off)
3. Recharge the devices
    - Phone
    - 2 Microsoft Bands
    - Battery pack

What if the phone or bands runs out of battery mid-day?

You are done collecting data for the day. Follow the steps under 'At the end of each the day' above.

Annotating Images

1. Login to web application at **http://www.activiome.com/auth/login** with your username/password



2. Click on the “Activities” link on the left sidebar

2. Select active eating images per eating session

Throughout the day, you probably eat multiple times - breakfast, lunch, snack, dinner, etc. We call each one of these an “eating session”. You will probably see between 2-10 images per eating session. In these images, there might a plate or bowl in front of you, or you will be holding some food with your hand(s), which could be a snack or fruit.

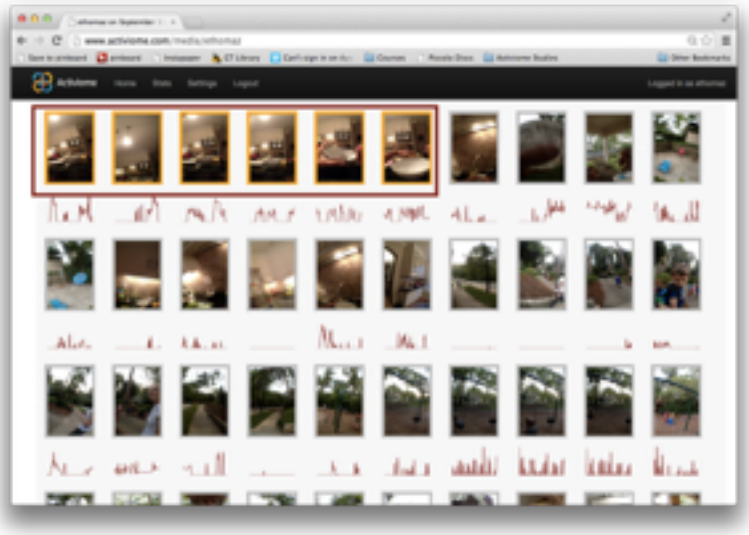
We would like you to:

1. Select all the images that correspond to an eating moment (i.e., lunch)
2. Annotate them with a label (described in more detail below)
3. Repeat 1 & 2 until all eating moments have been labeled

Note that there might be times during an eating session when you are not actually eating. For example, if you go to a restaurant, you will order some food and then wait until it comes. You might chat with friends or use your mobile phone in the meantime. When the food comes you will start eating per se - this is what we call the active eating session. We would like you to select the images that show you actively eating. Some of the images might depict you moving food towards your mouth, but most of the images will probably be of the food in front of you, most likely in a plate, bowl, or in your hands. Once you are done actively eating, you might chat with friends a bit more, wait for the check, etc. Again, we want you to select only the images that show you in the active eating session.

How to select images

To select one image just click on it. When it is selected, the image’s border becomes yellow. You can use the mouse to select multiple images (multi-select). To add/remove images to a selection, you can use the command key (on the Mac).

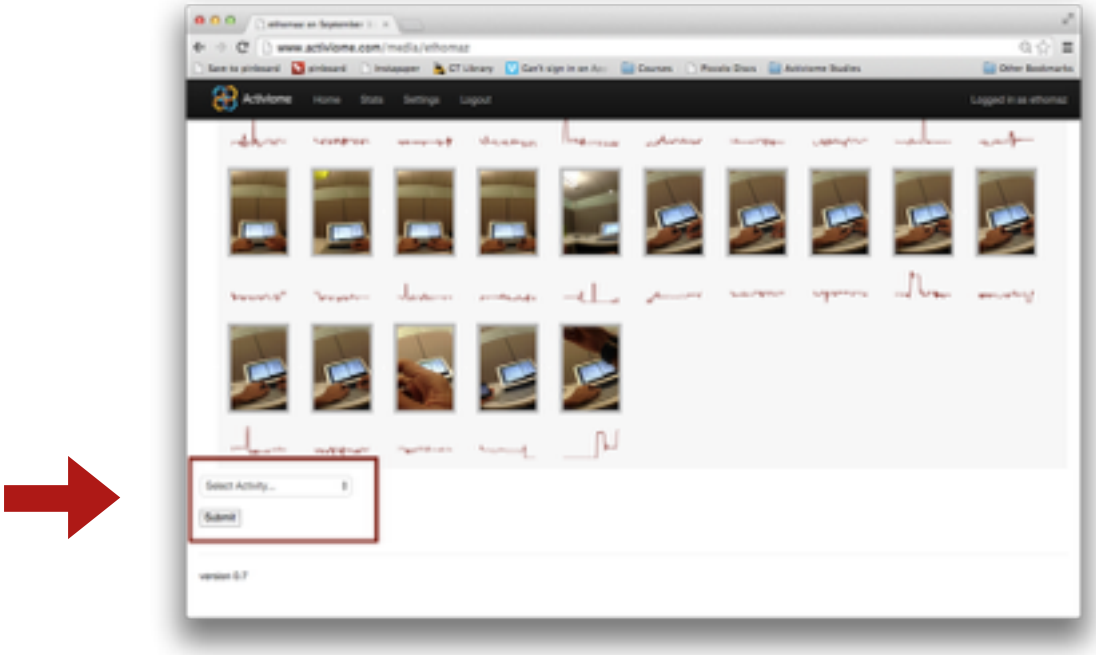


3. Annotate eating session (selection of images)

With eating images selected (corresponding to one eating session), scroll down towards the end of the web page, choose one of the “Eating” options and hit the submit button. The options are: Eating Fork, Eating Fork Knife, Eating Spoon, Eating Hand, Eating Other.

Eating Fork: Eating with a fork
Eating Fork Knife: Eating with fork and knife
Eating Spoon: Eating with a spoon
Eating Hand: Eating while holding the food with one or two hands (e.g., sandwich)
Eating Other: Eating in some other way or with some other utensil (e.g., chopsticks)

If the eating session involved a combination of these (e.g., eating with utensils and hand), pick the one that best represents, in your view, the way you ate. Optionally, you could sub-divide the eating session even more and accurately label each image of set of images. For us, the more accurate the annotation the better.



CONSENT DOCUMENT FOR ENROLLING ADULT PARTICIPANTS IN A RESEARCH STUDY

Georgia Institute of Technology
Project Title: Everyday Activity Recognition with Multimodal Sensing
Investigators: *Gregory Abowd, Irfan Essa, Edison Thomaz*

You are being asked to be a volunteer in a research study.

Purpose:

In our research we are exploring the use of multimodal sensing approaches to recognize people's everyday activities, such as sleeping, exercising, socializing, eating, etc. The ultimate goal is to build systems that can recognize what people are doing in real-time and act on that knowledge, either by providing relevant information or nudging people for behavior change purposes (e.g. help them eat healthier meals).

The study consists of providing participants with (a) lightweight wearable sensor(s) (e.g. activity tracker) and/or (a) smartphone(s) and asking them to perform their normal activities for **one day** while wearing these devices. At the end of the study we will collect the sensor data and use it to build systems that can classify activities based on sensor signals.

Exclusion/Inclusion Criteria:

- You must be willing to wear the sensor(s) and/or device(s) continuously throughout the duration of the study.
- You must agree that we will collect and examine images and sensor data reflecting your everyday activities.

Procedures:

- If you agree to be in this study, we will provide you with one or more wearable sensors and possibly (a) smartphone(s) as well.
- We will assist you with the setup of the sensor(s) and/or device(s).
- One of the smartphones, if any, might be setup to be worn around the neck and take photos automatically every 30 seconds. All images taken will be stored in the device and you will be able to see and delete them at any point. Also, at the end of the study you will be able to review all images and have an opportunity to delete any of them before returning the device to us.
- We understand that it may not be feasible to wear the sensor(s) and/or device(s) continuously for hours at a time. When not possible or desirable to wear them, you may take them off.

- We will collect the sensor(s) and/or device(s) at the end of the study.

Risks or Discomforts:

- Wearing the wearable sensor(s) and/or device(s) might prove uncomfortable.
- You may be concerned about the images that we might collect. You will be able to review and delete images before you make them available to researchers.

Benefits:

- You are not likely to benefit in any way from joining this study. We hope that what we learn will someday help you and others.

Compensation to You:

- We will give you $10 as compensation for being in the study

Confidentiality:

- We will not share any of your sensor data with anyone.
- To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB may review study records. The Office of Human Research Protections may also look over study records during required reviews.

Costs to You:

- There are no costs to you, other than your time, for being in this study.

In Case of Injury/Harm:

- If you are injured as a result of being in this study, please contact Prof. Gregory Abowd at telephone (404) 385-5055 or via email at abowd@gatech.edu. Neither the Principal Investigator nor Georgia Institute of Technology has made provision for payment of costs associated with any injury resulting from participation in this study.

Participant Rights:

- Your participation in this study is voluntary. You do not have to be in this study if you don't want to be.

- You have the right to change your mind and leave the study at any time without giving any reason and without penalty.
- Any new information that may make you change your mind about being in this study will be given to you.
- You will be given a copy of this consent form to keep.
- You do not waive any of your legal rights by signing this consent form.

Questions about the Study:

- If you have any questions about the study, please contact Prof. Gregory Abowd at telephone (404) 385-5055 or via email at abowd@gatech.edu.

Questions about Your Rights as a Research Participant:

If you have any questions about your rights as a research participant, you may contact:

Ms. Melanie Clark, Georgia Institute of Technology
Office of Research Compliance, at (404) 894-6942.
or
Ms. Kelly Winn, Georgia Institute of Technology
Office of Research Compliance, at (404) 385- 2175.

If you sign below, it means that you have read (or have had read to you) the information given in this consent form, and you would like to be a volunteer in this study.

_____
Participant Name (printed)

_____ _____
Participant Signature Date

_____ _____
Signature of Person Obtaining Consent Date

CONSENT DOCUMENT FOR ENROLLING ADULT PARTICIPANTS IN A RESEARCH STUDY

Longitudinal Tracking and Inference of Everyday Activities with Multimodal Sensing

Investigators: *Gregory Abowd, Irfan Essa, Edison Thomaz, Rushil Khurana*
Georgia Institute of Technology

You are being asked to be a volunteer in a research study.

Purpose:

In our research we are exploring the use of multimodal sensing approaches to recognize people's everyday activities, such as sleeping, exercising, socializing, and eating. The ultimate goal is to build systems that can recognize what people are doing in real-time and act on that knowledge, either by providing relevant information or nudging people for behavior change purposes (e.g. help them eat healthier meals).

The study consists of providing participants with wearable sensor(s) (e.g. activity trackers, physiological sensors, wearable cameras) and asking them to perform their normal activities while wearing these devices over multiple days. Participants will be asked to login to a web site on a regular basis (e.g., every evening) and annotate their sensor data for the day, indicating their activities. We will collect the annotated sensor data and use it to build systems that can classify activities based on sensor signals.

We will delete all the raw data, including images and sensor data after a period of two weeks from data collection.

Exclusion/Inclusion Criteria:

- You must be willing to wear the sensor(s) and/or device(s) throughout the duration of the study as much as possible.
- You must agree to care for and recharge the sensor(s) and/or device(s) throughout the duration of the study.
- You must agree that we will collect sensor data reflecting your everyday activities.
- You must agree to annotate the sensor data collected, indicating your everyday activities.
- You must be willing to receive short text messages or notifications on your phone asking for confirmation about activities you are performing.

Procedures:

- If you agree to be in this study, we will provide you with one or more wearable sensors and a smartphone.
- We will assist you with the setup of the sensor(s) and/or device(s).
- We understand that it may not be feasible to wear the sensor(s) and/or device(s) continuously. When not possible or desirable to wear them, you may take them off.
- You will need to recharge the sensor(s) and smartphone every night.
- During the study, the sensors will be uploading data to a server in real-time through a cellular connection mediated by the phone we will provide.
- All the sensor data collected will be available to you through a password-protected web site that only you will have access to. By logging in you will be able to visualize and/or delete any of the data.
- You will be asked to annotate the sensor data using the password-protected web site on a regular basis (e.g., every evening). You will be given specific instructions for how to perform said annotation.
- You might be asked through a text message or phone notification to confirm whether you are performing a specific activity at a particular time during the day.
- We will collect the sensor(s) and/or device(s) at the end of the study.

Risks or Discomforts:

- Wearing the wearable sensor(s) and/or device(s) might prove uncomfortable.
- We tried to make the sensor data annotation process as efficient as possible but it might still prove tedious over multiple days.

Benefits:

- One of the domains this study aims to impact is that of automated dietary assessment; being able to automatically detect when and what people are eating. For decades, researchers have been trying to build systems that automatically recognize what people eat. A system like this would enable health researchers to develop a better understanding of dietary habits at the population level. Additionally, as has been shown in numerous studies, people also benefit from food journaling; by becoming more aware of what they eat, people tend to eat better.

Compensation to You:

- We will give you $10 per day as compensation for being in the study. We will consider a day if data collection has been collected for at least 5 hours.

Confidentiality:

- All of you sensor data will be uploaded to a server and you will be the only one with access to it. You will be able to review all the data and/or delete it if you wish.

- To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology IRB may review study records. The Office of Human Research Protections may also look over study records during required reviews.

Costs to You:

- There are no costs to you, other than your time, for being in this study.

In Case of Injury/Harm:

- If you are injured as a result of being in this study, please contact Prof. Gregory Abowd at telephone (404) 385-5055 or via email at abowd@gatech.edu. Neither the Principal Investigator nor Georgia Institute of Technology has made provision for payment of costs associated with any injury resulting from participation in this study.

Participant Rights:

- Your participation in this study is voluntary. You do not have to be in this study if you don't want to be.
- You have the right to change your mind and leave the study at any time without giving any reason and without penalty.
- Any new information that may make you change your mind about being in this study will be given to you.
- You will be given a copy of this consent form to keep.
- You do not waive any of your legal rights by signing this consent form.

Questions about the Study:

- If you have any questions about the study, please contact Prof. Gregory Abowd at telephone (404) 385-5055 or via email at abowd@gatech.edu.

Questions about Your Rights as a Research Participant:

If you have any questions about your rights as a research participant, you may contact:

Ms. Melanie Clark, Georgia Institute of Technology
Office of Research Compliance, at (404) 894-6942.
or
Ms. Kelly Winn, Georgia Institute of Technology
Office of Research Compliance, at (404) 385- 2175.

If you sign below, it means that you have read (or have had read to you) the information given in this consent form, and you would like to be a volunteer in this study.

_____
Participant Name (printed)

_____ _____
Participant Signature Date

_____ _____
Signature of Person Obtaining Consent Date

REFERENCES

[1] Amft, O., "A wearable earpad sensor for chewing monitoring," *Sensors, 2010 IEEE*, pp. 222–227, 2010.

[2] Amft, O., Bannach, D., Pirkl, G., Kreil, M., and Lukowicz, P., "Towards wearable sensing-based assessment of fluid intake," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pp. 298–303, 2010.

[3] Amft, O., Junker, H., and Tröster, G., "Detection of eating and drinking arm gestures using inertial body-worn sensors," in *ISWC '05: Proceedings of the Ninth IEEE International Symposium on Wearable Computers*, IEEE Computer Society, Oct. 2005.

[4] Amft, O., Stäger, M., Lukowicz, P., and Tröster, G., "Analysis of chewing sounds for dietary monitoring," in *UbiComp'05: Proceedings of the 7th international conference on Ubiquitous Computing*, Springer-Verlag, Sept. 2005.

[5] Amft, O. and Tröster, G., "Recognition of dietary activity events using on-body sensors," *Artificial Intelligence in Medicine*, vol. 42, pp. 121–136, Feb. 2008.

[6] Amft, O. and Tröster, G., "On-Body Sensing Solutions for Automatic Dietary Monitoring," *IEEE pervasive computing*, vol. 8, Apr. 2009.

[7] Arab, L., Estrin, D., Kim, D. H., Burke, J., and Goldman, J., "Feasibility testing of an automated image-capture method to aid dietary recall," *European Journal of Clinical Nutrition*, vol. 65, pp. 1156–1162, May 2011.

[8] Archer, E., Hand, G. A., and Blair, S. N., "Validity of U.S. Nutritional Surveillance: National Health and Nutrition Examination Survey Caloric Energy Intake Data, 1971–2010," *PLoS ONE*, vol. 8, p. 76632, Oct. 2013.

[9] Ashbrook, D. and Starner, T., "Using GPS to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 275–286, 2003.

[10] Bäckström, T. and Magi, C., "Properties of line spectrum pair polynomials—A review," *Signal Processing*, vol. 86, pp. 3286–3298, Nov. 2006.

[11] Bai, Y., Li, C., Yue, Y., Jia, W., Li, J., Mao, Z.-H., and Sun, M., "Designing a wearable computer for lifestyle evaluation," in *Bioengineering Conference (NEBEC), 2012 38th Annual Northeast*, pp. 93–94, 2012.

[12] Baker, R. C. and Kirschenbaum, D. S., "Self-monitoring may be necessary for successful weight control," *Behavior Therapy*, vol. 24, no. 3, pp. 377–394, 1993.

[13] Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K., "Soylent: a word processor with a crowd inside.," *UIST*, pp. 313–322, 2010.

[14] Biagioni, J. and Krumm, J., "Days of Our Lives: Assessing Day Similarity from Location Traces," *User Modeling*, 2013.

[15] Bingham, S. A., *The dietary assessment of individuals; methods, accuracy, new techniques and recommendations*. 1987.

[16] Blanke, U. and Schiele, B., "Daily routine recognition through activity spotting," *International Symposium on Location and Context Awareness (LoCA)*, pp. 192–206, 2009.

[17] Boushey, C. J., Coulston, A. M., Rock, C. L., and Monsen, E., *Nutrition in the Prevention and Treatment of Disease*. Academic Press, 2001.

[18] Bradski, G., "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[19] Burke, L. E., Swigart, V., Warziski Turk, M., Derro, N., and Ewing, L. J., "Experiences of Self-Monitoring: Successes and Struggles During Treatment for Weight Loss," *Qualitative Health Research*, vol. 19, pp. 815–828, May 2009.

[20] Burke, L. E., Wang, J., and Sevick, M. A., "Self-Monitoring in Weight Loss: A Systematic Review of the Literature," *YJADA*, vol. 111, pp. 92–102, Jan. 2011.

[21] Byrne, D., Doherty, A. R., Jones, G. J. F., Smeaton, A. F., Kumpulainen, S., and Järvelin, K., "The SenseCam as a tool for task observation," in *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*, British Computer Society, Sept. 2008.

[22] Castro, D., Hickson, S., Bettadapura, V., Thomaz, E., Abowd, G., Christensen, H., and Essa, I., "Predicting daily activities from egocentric images using deep learning," in *the 2015 ACM International Symposium*, (New York, New York, USA), pp. 75–82, ACM Press, 2015.

[23] Chang, K., Liu, S., Chu, H., Hsu, J., Chen, C., Lin, T., and Huang, P., "The diet-aware dining table: Observing dietary behaviors over a tabletop surface," *Pervasive Computing*, pp. 366–382, 2006.

[24] Chen, F., Wang, R., Zhou, X., and Campbell, A. T., "My smartphone knows i am hungry," in *the 2014 workshop*, (New York, New York, USA), pp. 9–14, ACM Press, 2014.

[25] Chen, J., Kam, A., Zhang, J., Liu, N., and Shue, L., "Bathroom activity monitoring based on sound," *Pervasive Computing*, pp. 65–76, 2005.

[26] Cheng, J., Zhou, B., Kunze, K., Rheinländer, C. C., Wille, S., Wehn, N., Weppner, J., and Lukowicz, P., "Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband," in *the 2013 ACM conference*, (New York, New York, USA), p. 155, ACM Press, 2013.

[27] Choe, E. K., Consolvo, S., Jung, J., Harrison, B., and Kientz, J. A., "Living in a glass house: a survey of private moments in the home," *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 41–44, 2011.

[28] Clarkson, B., Basu, S., Eagle, N., Choudhury, T., and Pentland, A., "Learning your life: wearables and familiars," in *Development and Learning, 2002. Proceedings. The 2nd International Conference on*, 2002.

[29] Clarkson, B. P. ., "Life patterns : structure from wearable sensors," *Thesis (Ph. D.) Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences*, 2005.

[30] Cordeiro, F., Epstein, D. A., Thomaz, E., Bales, E., Jagannathan, A. K., Abowd, G. D., and Fogarty, J., "Barriers and Negative Nudges: Exploring Challenges in Food Journaling ," in *the 33rd Annual ACM Conference*, (New York, New York, USA), pp. 1159–1162, ACM Press, 2015.

[31] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, IEEE, 2009.

[32] Dhurandhar, N. V., Schoeller, D., Brown, A. W., Heymsfield, S. B., Thomas, D., Sorensen, T. I. A., Speakman, J. R., Jeansonne, M., and Allison, D. B., "Energy balance measurement: when something is not better than nothing," *International Journal of Obesity*, Nov. 2014.

[33] Doherty, A. R. A., Hodges, S. E. S., King, A. C. A., Smeaton, A. F. A., Berry, E. E., Moulin, C. J. A. C., Lindley, S. S., Kelly, P. P., and Foster, C. C., "Wearable cameras in health: the state of the art and future possibilities.," *American journal of preventive medicine*, vol. 44, pp. 320–323, Mar. 2013.

[34] Dong, Y., "Tracking Wrist Motion to Detect and Measure the Eating Intake of Free-Living Humans," *Thesis (Ph. D.) Clemson University*, pp. 1–106, May 2012.

[35] Dong, Y., Scisco, J., Wilson, M., Muth, E., and Hoover, A., "Detecting periods of eating during free living by tracking wrist motion," *IEEE Journal of Biomedical Health Informatics*, Sept. 2013.

[36] Eagle, N. and Pentland, A., "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[37] Eagle, N. and Pentland, A. S., "Eigenbehaviors: identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[38] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.," *KDD*, pp. 226–231, 1996.

[39] Estruch, R., Ros, E., Salas-Salvadó, J., Covas, M.-I., Corella, D., Arós, F., Gómez-Gracia, E., Ruiz-Gutiérrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R. M., Serra-Majem, L., Pintó, X., Basora, J., Muñoz, M. A., Sorlí, J. V., Martínez, J. A., and Martínez-González, M. A., "Primary Prevention of Cardiovascular Disease with a Mediterranean Diet," *New England Journal of Medicine*, vol. 368, pp. 1279–1290, Apr. 2013.

[40] Farb, P. and Armelagos, G., *Consuming passions, the anthropology of eating.* Houghton Mifflin, 1980.

[41] Fouse, A., Weibel, N., Hutchins, E., and Hollan, J. D., "ChronoViz: a system for supporting navigation of time-coded data.," *CHI Extended Abstracts*, pp. 299–304, 2011.

[42] Franke, T., Lukowicz, P., Kunze, K., and Bannach, D., "Can a Mobile Phone in a Pocket Reliably Recognize Ambient Sounds?," *IEEE International Symposium on Wearable Computers. Proceedings*, pp. 161–162, Sept. 2009.

[43] Froehlich, J., Chen, M., Consolvo, S., Harrison, B., and Landay, J., "My-Experience: a system for in situ tracing and capturing of user feedback on mobile phones," *Proceedings of the 5th international conference on Mobile systems, applications and services*, pp. 57–70, 2007.

[44] Gemming, L., Doherty, A., Utter, J., Shields, E., and Mhurchu, C. N., "The use of a wearable camera to capture and categorise the environmental and social context of self-identified eating episodes," *Appetite*, vol. 92, pp. 118–125, Sept. 2015.

[45] Gillet, O. and Richard, G., "Automatic transcription of drum loops," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. iv–269–iv–272, IEEE, 2004.

[46] Go, V. L. W., Nguyen, C. T. H., Harris, D. M., and Lee, W.-N. P., "Nutrient-gene interaction: metabolic genotype-phenotype relationship.," *The Journal of nutrition*, vol. 135, pp. 3016S–3020S, Dec. 2005.

[47] Gowdy, J., *Limited wants, unlimited means: A reader on hunter-gatherer economics and the environment.* Island Press, 1997.

[48] Hara, K., Le, V., and Froehlich, J., "Combining crowdsourcing and google street view to identify street-level accessibility problems," in *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Request Permissions, Apr. 2013.

[49] Hatori, M., Vollmers, C., Zarrinpar, A., DiTacchio, L., Bushong, E. A., Gill, S., Leblanc, M., Chaix, A., Joens, M., Fitzpatrick, J. A., and others, "Time-restricted feeding without reducing caloric intake prevents metabolic diseases in mice fed a high-fat diet," *Cell metabolism*, vol. 15, no. 6, pp. 848–860, 2012.

[50] Hernandez, J., Li, Y., Rehg, J. M., and Picard, R. W., "BioGlass: Physiological parameter estimation using a head-mounted wearable device," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pp. 55–58, IEEE, 2014.

[51] Hicks, J., Ramanathan, N., Kim, D., Monibi, M., Selsky, J., Hansen, M., and Estrin, D., "Andwellness: An open mobile system for activity and experience sampling," *Wireless Health 2010*, pp. 34–43, 2010.

[52] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012.

[53] Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., and Wood, K., "SenseCam: a retrospective memory aid," in *UbiComp'06: Proceedings of the 8th international conference on Ubiquitous Computing*, Springer-Verlag, Sept. 2006.

[54] Hoyle, R., Templeman, R., Armes, S., Anthony, D., Crandall, D., and Kapadia, A., "Privacy behaviors of lifeloggers using wearable cameras," in *the 2014 ACM International Joint Conference*, (New York, New York, USA), pp. 571–582, ACM Press, 2014.

[55] Huỳnh, T., Fritz, M., and Schiele, B., "Discovery of activity patterns using topic models," *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 10–19, 2008.

[56] Iachello, G. and Abowd, G. D., "Privacy and proportionality: adapting legal evaluation techniques to inform design in ubiquitous computing," in *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 91–100, ACM, 2005.

[57] Intille, S., Bao, L., Tapia, E., and Rondoni, J., "Acquiring in situ training data for context-aware ubiquitous computing applications," *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1–8, 2004.

[58] Jacobs, D. R., "Challenges in research in nutritional epidemiology," *Nutritional Health*, pp. 29–42, 2012.

[59] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T., "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, pp. 675–678, 2014.

[60] Junker, H., Amft, O., Lukowicz, P., and Tröster, G., "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognition*, vol. 41, pp. 2010–2024, June 2008.

[61] Kadomura, A., Li, C.-Y., Tsukada, K., Chu, H.-H., and Siio, I., "Persuasive technology to improve eating behavior using a sensor-embedded fork," in *the 2014 ACM International Joint Conference*, (New York, New York, USA), pp. 319–329, ACM Press, 2014.

[62] Kahneman, D., Krueger, A. B., Schkade, D. A., and Schwarz, N., "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science*, 2004.

[63] Kalantarian, H., Alshurafa, N., and Sarrafzadeh, M., "A Wearable Nutrition Monitoring System," in *Wearable and Implantable Body Sensor Networks (BSN), 2014 11th International Conference on*, pp. 75–80, 2014.

[64] Kanfer, F. H., "Self-monitoring: Methodological limitations and clinical applications," *Journal of Consulting and Clinical Psychology*, vol. 35 (2), pp. 148–152, Oct. 1970.

[65] Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G., "Extracting places from traces of locations," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, (New York, NY, USA), pp. 110–118, ACM, 2004.

[66] Karen S Hamrick, M. A. J. G. D. H. and McClelland, K., "How Much Time Do Americans Spend on Food?," pp. 1–64, Nov. 2011.

[67] Kelly, P., Doherty, A., Berry, E., Hodges, S., Batterham, A. M., and Foster, C., "Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 8, p. 44, May 2011.

[68] Kelly, P., Marshall, S. J., Badland, H., Kerr, J., Oliver, M., Doherty, A. R., and Foster, C., "An ethical framework for automated, wearable cameras in health behavior research.," *American journal of preventive medicine*, vol. 44, pp. 314–319, Mar. 2013.

[69] Kim, H.-J., Kim, M., Lee, S.-J., and Choi, Y. S., "An analysis of eating activities for automatic food type recognition," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pp. 1–5, 2012.

[70] Kindig, D. and Stoddart, G., "What is population health?," *American Journal of Public Health*, vol. 93, no. 3, pp. 380–383, 2003.

[71] Kittur, A., Chi, E. H., and Suh, B., "Crowdsourcing user studies with Mechanical Turk," in *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM Request Permissions, Apr. 2008.

[72] Kleitman, N., *Sleep and wakefulness*. Chicago: The University of Chicago Press, July 1963.

[73] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105, 2012.

[74] Larson, R. and Csikszentmihalyi, M., "The Experience Sampling Method," in *Flow and the Foundations of Positive Psychology* (Reis, H. T., ed.), pp. 21–34, Dordrecht: Springer Netherlands, 2014.

[75] Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P., "Real-time crowd labeling for deployable activity recognition," *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 1203–1212, 2013.

[76] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[77] Lester, J., Tan, D., Patel, S., and Brush, A., "Automatic classification of daily fluid intake," *Audio, Transactions of the IRE Professional Group on*, pp. 1–8, Mar. 2010.

[78] Li, C.-Y., Chen, Y.-C., Chen, W.-J., and Huang, P., "Sensor-Embedded Teeth for Oral Activity Recognition," *ISWC 2013*, vol. 42, Sept. 2013.

[79] Lifson, N. and McClintock, R., "Theory of use of the turnover rates of body water for measuring energy and material balance," *Journal of theoretical biology*, vol. 12, no. 1, pp. 46–74, 1966.

[80] Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., and Yang, G.-Z., "An Intelligent Food-Intake Monitoring System Using Wearable Sensors," in *Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on*, pp. 154–160, IEEE Computer Society, 2012.

[81] Lu, H., Pan, W., Lane, N., Choudhury, T., and Campbell, A., "SoundSense: scalable sound sensing for people-centric applications on mobile phones," *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pp. 165–178, 2009.

[82] Lukowicz, P., Pentland, A. S., and Ferscha, A., "From Context Awareness to Socially Aware Computing," *IEEE pervasive computing*, vol. 11, no. 1, pp. 32–40, 2012.

[83] Maekawa, T., "A sensor device for automatic food lifelogging that is embedded in home ceiling light: A preliminary investigation," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 405–407, 2013.

[84] Makeyev, O., Lopez-Meyer, P., Schuckers, S., Besio, W., and Sazonov, E., "Biomedical Signal Processing and Control," *Biomedical Signal Processing and Control*, vol. 7, pp. 649–656, Nov. 2012.

[85] Makhoul, J., "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[86] Mankoff, J., Hsieh, G., Hung, H. C., Lee, S., and Nitao, E., "Using Low-Cost Sensing to Support Nutritional Awareness," in *UbiComp '02: Proceedings of the 4th international conference on Ubiquitous Computing*, Springer-Verlag, Sept. 2002.

[87] Marder, S. R., Essock, S. M., Miller, A. L., Buchanan, R. W., Casey, D. E., Davis, J. M., Kane, J. M., Lieberman, J. A., Schooler, N. R., Covell, N., and others, "Physical health monitoring of patients with schizophrenia," *American Journal of Psychiatry*, 2014.

[88] Markson, E. W., "Functional, social, and psychological disability as causes of loss of weight and independence in older community-living people.," *Clinics in geriatric medicine*, vol. 13, no. 4, pp. 639–652, 1997.

[89] Marshall, T. A., Stumbo, P. J., Warren, J. J., and Xie, X. J., "Inadequate nutrient intakes are common and are associated with low diet variety in rural, community-dwelling elderly.," *The Journal of nutrition*, vol. 131, pp. 2192–2196, Aug. 2001.

[90] Mason, W. and Watts, D. J., "Financial incentives and the "performance of crowds"," in *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation*, ACM Request Permissions, June 2009.

[91] Mathieu, B., Essid, S., Fillon, T., Prado, J., and Richard, G., "*YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software*," in *proceedings of the 11th ISMIR conference, 2010*, Sept. 2010.

[92] McCormack, P., "Undernutrition in the elderly population living at home in the community: a review of the literature," *Journal of advanced nursing*, vol. 26, no. 5, pp. 856–863, 1997.

[93] Michels, K. B., "A renaissance for measurement error.," *International journal of epidemiology*, vol. 30, pp. 421–422, June 2001.

[94] Michels, K. B., "Nutritional epidemiology–past, present, future," *International journal of epidemiology*, vol. 32, pp. 486–488, Aug. 2003.

[95] Mintz, S. W. and Du Bois, C. M., "The anthropology of food and eating," *Annual review of anthropology*, pp. 99–119, 2002.

[96] Moore, B. C. J., Glasberg, B. R., and Baer, T., "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[97] Newman, S. C. and Bland, R. C., "Mortality in a cohort of patients with schizophrenia: a record linkage study.," *The Canadian Journal of Psychiatry/La Revue canadienne de psychiatrie*, 1991.

[98] Nguyen, D. H., Marcu, G., Hayes, G. R., Truong, K. N., Scott, J., Langheinrich, M., and Roduner, C., "Encountering SenseCam: personal recording technologies in everyday life," pp. 165–174, 2009.

[99] Noronha, J., Hysen, E., Zhang, H., and Gajos, K. Z., "Platemate: crowdsourcing nutritional analysis from food photographs," *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 1–12, 2011.

[100] Oliva, A. and Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[101] O'Loughlin, G., Cullen, S. J., McGoldrick, A., O'Connor, S., Blain, R., O'Malley, S., and Warrington, G. D., "Using a wearable camera to increase the accuracy of dietary analysis.," *American journal of preventive medicine*, vol. 44, pp. 297–301, Mar. 2013.

[102] Parate, A., Chiu, M.-C., Chadowitz, C., Ganesan, D., and Kalogerakis, E., "RisQ: recognizing smoking gestures with inertial sensors on a wristband," in *MobiSys '14: Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, ACM Request Permissions, June 2014.

[103] Passler, S. and Fischer, W., "Acoustical method for objective food intake monitoring using a wearable sensor system," in *Pervasive Computing Technologies for*

Healthcare (PervasiveHealth), 2011 5th International Conference on, pp. 266–269, 2011.

[104] Pässler, S., Wolff, M., and Fischer, W.-J., "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food," *Physiological Measurement*, vol. 33, pp. 1073–1093, May 2012.

[105] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[106] Rahman, S. A., Merck, C., Huang, Y., and Kleinberg, S., "Unintrusive eating recognition using Google glass," in *PervasiveHealth '15: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, May 2015.

[107] Rahman, T., Adams, A. T., Zhang, M., Cherry, E., Zhou, B., Peng, H., and Choudhury, T., "BodyBeat: a mobile system for sensing non-speech body sounds.," *MobiSys*, pp. 2–13, 2014.

[108] Ramanathan, N., Alquaddoomi, F., Falaki, H., George, D., Hsieh, C., Jenkins, J., Ketcham, C., Longstaff, B., Ooms, J., Selsky, J., Tangmunarunkit, H., and Estrin, D., "ohmage: An open mobile system for activity and experience sampling," *PERVASIVE*, pp. 203–204, 2012.

[109] Reddy, S., Parker, A., Hyman, J., Burke, J., Estrin, D., and Hansen, M., "Image browsing, processing, and clustering for participatory sensing: lessons from a DietSense prototype," in *EmNets '07: Proceedings of the 4th workshop on Embedded networked sensors*, ACM Request Permissions, June 2007.

[110] Ritchie, C. S., Burgio, K. L., Locher, J. L., Cornwell, A., Thomas, D., Hardin, M., and Redden, D., "Nutritional status of urban homebound older adults.," *The American journal of clinical nutrition*, vol. 66, no. 4, pp. 815–818, 1997.

[111] Rondoni, J., "Context-Aware Experience Sampling for the Design and Study of Ubiquitous Technologies," *M.Eng. Thesis, Massachusetts Institute of Technology, 2003*, pp. 1–84, June 2003.

[112] Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B., "Who are the crowdworkers?: shifting demographics in mechanical turk," *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pp. 2863–2872, 2010.

[113] Rossi, M., Tröster, G., and Amft, O., "Recognizing Daily Life Context Using Web-Collected Audio Data," *Wearable Computers (ISWC), 2012 16th International Symposium on*, pp. 25–28, 2012.

[114] Rossi, M., Feese, S., Amft, O., Braune, N., Martis, S., and Tröster, G., "AmbientSense: A real-time ambient sound recognition system for smartphones," in

Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on, pp. 230–235, 2013.

[115] Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T., "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, May 2008.

[116] Sazonov, E., Schuckers, S., Lopez-Meyer, P., Makeyev, O., Sazonova, N., Melanson, E. L., and Neuman, M., "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior," *Physiological Measurement*, vol. 29, pp. 525–541, Apr. 2008.

[117] Scheirer, E. and Slaney, M., "Construction and evaluation of a robust multifeature speech/music discriminator," *IEEE Internation Conference on Acoustics, Speech and Signal Processing, p.1331-1334, 1997.*, vol. 2, pp. 1331–1334, 1997.

[118] Schoeller, "Limitations in the assessment of dietary energy intake by self-report," *Metabolism*, vol. 44, pp. 5–5, Feb. 1995.

[119] Schoenfeld, J. D. and Ioannidis, J. P., "Is everything we eat associated with cancer? A systematic cookbook review," *The American journal of clinical nutrition*, vol. 97, pp. 127–134, Jan. 2013.

[120] Schussler, H., "A stability theorem for discrete systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 87–89, Feb. 1976.

[121] Song, Y. C., Lasecki, W. S., Bigham, J. P., and Kautz, H., "Training Activity Recognition Systems Online Using Real-time Crowdsourcing," *Ubicomp '12: Proceedings of the 14th ACM International Conference on Ubiquitous Computing*, 2012.

[122] Sorokin, A. and Forsyth, D., "Utility data annotation with Amazon Mechanical Turk," *Audio, Transactions of the IRE Professional Group on*, pp. 1–8, June 2008.

[123] Stager, M., Lukowicz, P., Perera, N., von Buren, T., Tröster, G., and Starner, T., "SoundButton: design of a low power wearable audio classification system," *Proceedings. Sixth International Symposium on Wearable Computers,*, pp. 12–17, Oct. 2003.

[124] Stellar, E. and Shrager, E. E., "Chews and swallows and the microstructure of eating," *The American journal of clinical nutrition*, vol. 42, no. 5, pp. 973–982, 1985.

[125] Sun, F.-T., Yeh, Y.-T., Cheng, H.-T., Kuo, C., and Griss, M. L., "Nonparametric discovery of human routines from sensor data.," *PerCom*, pp. 11–19, 2014.

[126] Sun, M., Fernstrom, J. D., Jia, W., Hackworth, S. A., Yao, N., Li, Y., Li, C., Fernstrom, M. H., and Sclabassi, R. J., "A wearable electronic system for objective dietary assessment," *Journal of the American Dietetic Association*, vol. 110, no. 1, p. 45, 2010.

[127] Thomaz, E., Abowd, G., and Essa, I., "A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing," in *UbiComp '15: Proceedings of the 2015 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 1–12, July 2015.

[128] Thomaz, E., Parnami, A., Bidwell, J., Essa, I. A., and Abowd, G. D., "Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras.," *UbiComp*, pp. 739–748, 2013.

[129] Thomaz, E., Parnami, A., Essa, I. A., and Abowd, G. D., "Feasibility of identifying eating moments from first-person images leveraging human computation.," *SenseCam*, pp. 26–33, 2013.

[130] Thomaz, E., Zhang, C., Essa, I., and Abowd, G. D., "Inferring Meal Eating Activities in Real World Settings from Ambient Sounds," in *the 20th Intelligent User Interfaces Conference (IUI)*, (New York, New York, USA), pp. 427–431, ACM Press, 2015.

[131] Turner, D. and others, "The estimation of the patient's home dietary intake.," *Journal of the American Dietetic Association*, vol. 16, pp. 875–881, 1940.

[132] von Ahn, L. and Dabbish, L., "Labeling images with a computer game," in *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Request Permissions, Apr. 2004.

[133] von Ahn, L., Liu, R., and Blum, M., "Peekaboom: a game for locating objects in images," in *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM Request Permissions, Apr. 2006.

[134] Ward, J. A., Lukowicz, P., Tröster, G., and Starner, T. E., "Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 10, pp. 1553–1567, 2006.

[135] Weiser, M., "The computer for the 21st Century," *Pervasive Computing, IEEE*, vol. 99, no. 1, pp. 19–25, 2002.

[136] Widdowson, E., "A study of english diets by the individual method: Part i. men," *Journal of Hygiene*, vol. 36, no. 03, pp. 269–290, 1936.

[137] Widdowson, E. and McCance, R., "A study of english diets by the individual method: Part ii. women," *Journal of Hygiene*, vol. 36, no. 03, pp. 293–307, 1936.

[138] Wiehl, D. G., "Diets of a group of aircraft workers in southern california," *The Milbank Memorial Fund Quarterly*, pp. 329–366, 1942.

[139] Wiehl, D. G. and Reed, R., "Development of new or improved dietary methods for epidemiological investigations," *American Journal of Public Health and the Nations Health*, vol. 50, no. 6_Pt_1, pp. 824–828, 1960.

[140] Wilde, M. H. and Garvin, S., "A concept analysis of self-monitoring," *Journal of Advanced Nursing*, vol. 57, pp. 339–350, Feb. 2007.

[141] Willett, W., *Nutritional Epidemiology*. Oxford University Press, Oct. 2012.

[142] Wyatt, D., Choudhury, T., and Bilmes, J., "Conversation detection and speaker segmentation in privacy-sensitive situated speech data.," *Proceedings of Interspeech*, pp. 586–589, 2007.

[143] Yatani, K. and Truong, K. N., "BodyScope: a wearable acoustic sensor for activity recognition," *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 341–350, 2012.

[144] Zeiler, M. D. and Fergus, R., "Visualizing and understanding convolutional networks," in *ECCV*, pp. 818–833, Springer, 2014.

[145] Zhang, H., Li, L., Jia, W., Fernstrom, J. D., Sclabassi, R. J., and Sun, M., "Recognizing physical activity from ego-motion of a camera.," *Proceedings of IEEE EMBS*, vol. 2010, pp. 5569–5572, 2010.

[146] Zhang, S., Ang, M. H., Xiao, W., and Tham, C. K., "Detection of activities by wireless sensors for daily life surveillance: eating and drinking," *Sensors*, vol. 9, no. 3, pp. 1499–1517, 2009.

[147] Zhou, B., Cheng, J., Sundholm, M., Reiss, A., Huang, W., Amft, O., and Lukowicz, P., "Smart table surface: A novel approach to pervasive dining monitoring," in *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*, pp. 155–162, IEEE, 2015.