Sensor-Based Analysis of Object-Use Patterns for the Automatic Assessment of Cognitive Status

by

Mark Richard Hodges

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2010

Doctoral Committee:

Professor Martha E. Pollack, Chair
Professor Satinder Singh Baveja
Associate Professor Ned L. Kirsch
Associate Professor Brian D. Noble
Assistant Professor Mark W. Newman

To my wonderful wife, Abigail Hodges,
for her encouragement and support and, most of all, for her love.
It is my great honor and great joy to be your husband.
I love you.

ACKNOWLEDGEMENTS

First and foremost, thank you to Martha Pollack, my advisor, whose has worked tirelessly to support and encourage me, to teach me and to improve my work. From the moment I started working with her, she seemed convinced that I would succeed and it has been her patience and kindness, as well as her fierce determination that have made it possible for me to do so.

Thank you also to the rest of my committee. I have appreciated the genuine interest in my topic from Satinder Baveja Singh, the conversations with Brian Noble that helped to direct the research in this dissertation as well as my future research, and the support and mentorship from Mark Newman in our many meetings together. I am particularly thankful to Ned Kirsch whose insight and guidance have been invaluable, and without whose expertise this research would not have been possible.

Thank you to all of the staff of MedRehab at the University of Michigan for their immense help with subject recruitment and data collection, with special thanks to Erin Spirl, Michelle Shenton and Pippin Gilbert.

I am very appreciative of the great advice, support, and entertainment I've received from others in the research group: Amy Wester, Julie Weber, Brett Clippingdale, Mike Moffitt, Peter Schwartz, and Mark Schaller, as well as others in the computer science program including Mark Liffiton, Nate Derbinsky, Benji Wester, James Boerkoel, Shruti Gupta, Kaushik Veeraraghavan, Arnaub Nandi, and Kyla McMullen. Thanks also to Tershia Pinder-Grover and those I worked with as part

of the Graduate Student Mentorship program, especially Emine Cagin, Sarah Root, Eugene Daneshvar, Amor Menezes, and Marcial Lapp; I have learned so much from all of you.

Lastly, thank you to my wonderful family and friends, especially to my wife, Abigail, to my parents Richard and Carol Hodges, and to my sister Janet Smith for your constant love and support, in this as in all things.

If I take the wings of the morning
and settle at the farthest limits of the sea,
even there your hand shall lead me,
and your right hand shall hold me fast.
I praise you, for I am fearfully and wonderfully made.
Wonderful are your works; that I know very well.

Psalm 139: 9-10, 14

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Indications of cognitive impairments such as dementia and traumatic brain injury (TBI) are often subtle and may be frequently missed by primary care physicians. These impairments are not uncommon: approximately 0.46% of Americans are hospitalized for brain injury each year and it is estimated that by 2050 up to 13 million Americans may have Alzheimer's disease—the most common form of dementia—quadruple the number that did in 2002.

This dissertation proposes and investigates ways in which machine inference and wireless sensors can be used to support the assessment of cognitive functioning. The central hypothesis is that object-usage data collected from wireless sensors during the performance of daily activities are sufficient to assess cognitive impairment.

I first investigate the ability to recognize individuals based on their sensed object-usage patterns during a simple task. This experiment constitutes an initial step in understanding how well object-use patterns can be automatically observed and analyzed. A preliminary study, using the simple task of preparing a cup of coffee, demonstrated the ability to correctly recognize ten individuals with 77% accuracy based on nine trials from each individual as training data.

The dissertation then directly addresses assessment of cognitive impairment with a study in which individuals with TBI made a pot of coffee. Four features were derived from the sensed data and compared to to the subjects' scores on standard neuropsychological evaluations. A key result is that Edit Distance, the most knowledge-

rich feature, significantly correlates with an apparent indicator of general neuropsychological integrity, namely the first principal component of the neuropsychological assessments. Since cognitive impairments are measured along many dimensions, suggestive correlations between the computed features and individual assessments are also presented.

Lastly, I present a preliminary study that investigates the possibility of differentiating impaired individuals from unimpaired individuals. Data was collected from five subjects with TBI and five matched, unimpaired subjects; analysis was done using the same set of computed features. Although the study is preliminary, it is interesting that Edit Distance is able to perfectly differentiate the impaired subjects from the unimpaired and that full results are consistent with those from the assessment study.

CHAPTER I

Introduction

Mark Weiser, the father of Ubiquitous Computing wrote in his seminal 1991 paper that, "the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it" [62]. The paradigm of Ubiquitous Computing foresees the "third wave" of computing: following the move from mainframes to personal computers, with many users and one computer giving way to one user and one computer, Ubiquitous Computing predicts a move to one user-many computer relationships. Many of these computers may individually perform very simple operations, but collectively they provide incredible value to the user, "weaving" themselves into daily life.

Passive observation of a user has been an important addition to the field of ubiquitous computing. Unlike previous forms of human-computer interaction, the ability of a computer to passively observe the user allows it to reason about the user and her needs without the user needing to be actively engaged. The analysis of data collected by sensors has made possible a broad range of prototype applications including a number of assistive technologies, which support users with physical, cognitive, or other types of disabilities. Examples include applications that observe and interact with a user with dementia to facilitate handwashing [20], provide navigational

guidance when a cognitively impaired user makes an error [34], and act as assurance systems that can monitor movement and activity of an older adult living at home [37].

Assistive technologies such as these often depend on the computer's ability to recognize a user's activities. In the field of activity recognition, one tactic is to attach to frequently used objects sensors that can detect human interaction with those objects. Using object-use patterns, these detected patterns of interaction, it is possible to identify basic activities that the user is performing [47]. Rather than inferring the activities of the user, however, this dissertation focuses on making inferences about the user herself: specifically, the identity of the user and whether the user has a cognitive impairment and, if so, how severe it is. The main contribution is a set of novel techniques for the automatic assessment of cognitive impairments through the passive observation of an individual performing typical daily activities.

1.1 Research Hypothesis

The central hypothesis of this dissertation is that important inferences about an individual can made by using sensor technology to observe the individual's object-use patterns during the performance of a simple task, significantly including the assessment of an individual's cognitive impairment. This hypothesis is broken into three key research questions:

1. Can an individual be recognized from her object-use patterns in the performance of a task?

2. Can the severity of an individual's impairment be assessed by electronically observing the individual's object-use patterns from the performance of a task?

3. Is it possible to recognize whether or not an individual has a cognitive impairment based on sensor-derived object-use patterns?

The first research question is an important starting point for the hypothesis: it is difficult to envision that assessment of impairment would be possible if regular patterns of behavior could not be observed using sensors and recognized automatically. Further, the question of whether people can be distinguished on the basis of their object-use patterns is interesting in its own right, suggesting the possibility of several new applications. This question will be examined in Chapter II.

The second and third research questions are the two important sub-tasks of assessing a cognitive impairment: assessing the severity of an impairment in an individual known to have a cognitive impairment, and determining whether an individual is impaired or not. These questions will be examined in Chapters III and IV, respectively. The impaired subjects in the two studies presented here are limited to individuals with Traumatic Brain Injury (TBI). Although it seems plausible that the results presented in this dissertation would generalize to other forms of impairment, further studies are needed to confirm this.

More generally, these three research questions also begin to address an important larger question: what useful inferences about an individual can be made by automatically observing the individual perform a simple task? This larger question is interesting because automatic task observation is possible in an unobtrusive manner; important inferences that can be made in this way would allow additional services to be provided without demanding time or attention from the individual being observed.

1.2 Motivation

The research presented here is motivated both by foundational questions and by a range of applications that it can support.

1.2.1 Motivation for Recognition of Individuals

The study of whether object-use patterns can be used to automatically recognize individuals has three direct motivations, as well as one indirect one—the latter is as a precursor to the study of assessing cognitive impairments.

The first direct motivation is that the research is interesting in its own right: the ability to do so indicates an individual's object-use patterns can be recognized by computers using existing technology.

There are also important privacy concerns that motivate this research: failing to understand how well individuals can be recognized from object-use patterns could lead to accidental release of private information. If recognition is possible, object-use data collected for other purposes should be held more carefully to prevent mistakes such as AOL's release of 'anonymized' search data that still allowed some searchers to be identified [5].

The final direct motivation is that the ability to recognize individuals based on their object-use patterns may allow for unobtrusive biometric security. While many forms of security are like fences–preventing entry to a system by unauthorized users but providing no security once the user has accessed the system–security based on usage patterns can continually confirm a user's identity without interrupting the user and while the user performs tasks that she would be performing anyway. A similar idea is studied by Clarke and Furnell who authenticate users of mobile phones using keystroke analysis since mobile phones are frequently stolen and are generally not

protected by passwords or non-physical methods [11].

The indirect motivation is to provide preliminary evidence of the feasibility of assessing cognitive impairment. As noted earlier, if recognition of individuals is not possible—if sensors are not able to detect regular patterns of behavior from subjects that are distinct from one another—it is difficult to envision that the recognition and assessment of cognitive impairments would be feasible.

1.2.2 Motivation for Recognition and Assessment of Cognitive Impairments

The ability to assess impairments using object-use patterns adds an interesting, objective metric to the set of assessment tools available to health care professionals. There are several advantages to this form of metric. First, it can be performed automatically, in a home environment without the need for direct oversight by health care professionals. This allows the observation of subjects over a longer period of time and in the home environment, without additional effort from the health care professionals. Second, with this passive approach, frequent observation and re-evaluation is possible without disrupting the life or schedule of the individual. Any sudden changes in impairment, for example improvement caused by successful medication or treatments, or sudden degradation resulting from a side-effect of medication, could be quickly detected and acted upon. Additionally, frequent re-assessment of cognitive status can potentially be used to track changes in status with a fine granularity. Finally, it can automatically provide feedback to systems that are assisting the subject. For example, the system designed by Hoey, et al. to assist during handwashing uses the cognitive status of the individual as one parameter to decide when to make recommendations [20]. Additionally, reminder systems such as Autominder [48] could use cognitive status to decide when to issue reminders that might be important for

a more severely cognitively impaired individual but unnecessary and obnoxious for a less impaired person.

Traumatic Brain Injury is not uncommon: approximately 0.46% of Americans are hospitalized for brain injury each year and individuals aged 15-24 are far more likely than any other age group with over 0.9% hospitalized each year for brain injury [54]. Of the Americans hospitalized each year, 50,000 will die from the brain injuries and an additional 80,000–90,000 experience the onset of long-term disability [58]. Adding to the importance of this work is the fact that cognitive impairments are frequently seen in wounded veterans returning from the Iraq War. Improved body armor has helped soldiers survive explosions that they might not have survived before, but the soldiers are suffering brain damage as a result of the blasts. The increase in Traumatic Brain Injury has been so dramatic that it has been called the "signature wound" of the Iraq War [66]—in one study of servicemembers arriving at Walter Reed Army Medical Center with injuries caused by explosions, 59% of the soldiers were found to have TBI and 56% of those are considered moderate or severe [38]. Like dementia, the existence and severity of TBI can be difficult to assess, in part because it cannot always be detected with imaging tests [4].

Recognizing whether or not an individual has a cognitive impairment is motivated by the fact that dementia can be very difficult to assess: primary care physicians may often miss indications of cognitive impairment during patient visits, resulting in between 29% and 76% of cases of dementia or probable dementia going undiagnosed by primary care physicians [21]. Additionally, cognitive ability may vary from day to day and, since case managers typically cannot observe patients on a daily basis, they are often forced to rely on questioning and required to "play detective" [64].

Additionally, this work is motivated by the dramatically aging population both

nationally and around the world. Dementia becomes more common with age, affecting fewer than 1% of individuals in North America aged 60-64, but more than 30% of those over age 85 [13]. Additionally, up to 50% of all individuals over 85 are found to have measurable decline in cognitive function [2]. As the population ages, the number of cases of Alzheimer's disease and dementia in general is poised to increase significantly. In 2000, 12.4% of the U.S. population was aged 65 and older, and it is predicted to increase to 19.6% by 2030 and 20.6% by 2050. The oldest subgroup, that of individuals aged 80 and older, is expected to rise even more dramatically, more than doubling from 3.3% of the population in 2000 to 5.4% in 2030 and 8.0% in 2050 [59]. Without scientific advances to lower the incidence rates or the progression of Alzheimer's, it is estimated that between 7.98 and 12.95 million people in the United States will have Alzheimer's Disease in 2050, four times the number that did in 2002 [51]. Even among younger individuals, certain subgroups of the population may be more likely to develop dementia. For example, a recent survey showed that nearly 2% of retirees from the National Football League between the ages of 30 and 49 had dementia or another memory-related disease compared to just 0.1% of the general population of US men that age. Likewise, 6.1% of NFL retirees over the age of 50 had memory problems, compared with 1.2% of the general population [61].

1.3 Background

In this section, I review related work in the areas of automatic detection of cognitive impairments, activity recognition, and biometric identification. Section 1.3.1 explores attempts to automatically detect cognitive impairment as well as other conditions. Section 1.3.2 presents activity recognition, focusing on techniques which apply object-use interactions. Finally, section 1.3.3 looks at other biometric mea-

sures that have been used for recognition or identification of individuals.

1.3.1 Automatic Detection of Cognitive Impairments

Due to the difficulty of evaluating dementia and cognitive decline in limited-duration, periodic visits to a health-care provider [21], several projects have focused on diagnosing dementia and cognitive impairments automatically.

A group at the Oregon Health & Science University has developed several techniques to observe computer performance in order to assess impairment. One study focuses on keyboard and mouse usage of elderly individuals, for logging in to the computer and for playing the solitaire game FreeCell, respectively [25]. Another study focuses on the performance of those individuals in the FreeCell game, comparing the number of moves used by an individual to the number needed by an automated solver to account for games of varying difficulty. Based on the performance results, it was possible to differentiate the three mildly cognitively impaired subjects from the six others [26]. Work with several other computer games, specially created to perform assessments of cognitive impairments is underway [27, 23] including some promising early results, showing a correlation between performance in a word game and scores on a standard neuropsychological assessment of verbal ability [24].

Other work has studied automatically monitoring mobility because slowed mobility may be a predictor of future cognitive decline. The time to answer a phone call was used to measure mobility, making the assumption that the user's location distribution in the home would remain relatively stable over time [41]. Another study by the same team used passive infrared detectors and several models to infer the mobility of subjects more directly as they move about a residence, either by observing the amount of time after activity ceased in one room room for it to begin in another, or

by observing travel time down a hallway [18].

Both of these are very promising approaches towards the automatic assessment of cognitive impairment. I chose a different approach—observation of object usage during the performance of a task—for this dissertation in large part because observation of task performance is commonly used by occupational therapists to assess impairment and to make decisions about whether an impaired individual should be living independently, driving, or working. An advantage of this approach compared to observing performance on a computer game is the ability to observe an individual performing a task she would regularly perform—a recent study showed that fewer than one quarter of individuals 65 and older play video games at all [46] and even those who do play regularly may not ordinarily choose to play the particular games that are designed to measure impairment.

This dissertation goes further than previous research to automatically assess an individual's impairment. While multiple studies involve impaired individuals, only [26] attempts to individually differentiate impaired subjects from unimpaired subjects (for example, [18] shows that the change in walking times from the morning to the evening are significantly different between the impaired and unimpaired individuals, but only examines them as an average of the two groups, not on an individual basis). Additionally, none of these studies attempt to assess the severity of an impairment—[24] shows a correlation between computed features and average verbal fluency, a neuropsychological assessment, but none of the subjects had been diagnosed with a cognitive impairment.

Automatic Assessment of Other Conditions

Sensors have also been used to perform automatic assessments of conditions other than cognitive impairments. This section gives three examples of systems that demonstrate the range of conditions that are being investigated.

Albinali has used wireless accelerometers to detect stereotypical repetitive motor movements (body rocking and hand flapping) that are common in individuals with Autism Spectrum Disorders (ASD). If unregulated, these stereotypical motor movements can interfere with the development of new skills, the performance of established skills, and even result in self-injurious behavior in certain conditions. The accelerometers were placed on the chest and on each wrist of six children and young adults with ASD. The system then used decision trees to achieve an accuracy of 89% in real-time evaluations of subjects in a test lab, and 83% in a classroom setting [1].

Another system is the Gesture Pendant designed by Starner, a worn camera that tracks hand movement. While the main goal of the system is to recognize participants' "control gestures" to control a television, stereo, or other devices around the home, the system was also designed to track some tremors, potentially allowing it to monitor tremors that can be symptoms of certain medical conditions such as Parkinson's Disease and pathological tremors, or that can be a side effect of medicine, or even a warning sign for emergencies such as insulin shock in a diabetic [55].

Westeyn has also has also performed a preliminary study of adding sensors to toys to support assessment of a child's development. The toys are modeled after typical toys for small children and have several sensors to measure touch, motion, and sound in order to measure 25 distinct classes of play including rolling, grasping, stacking, and knocking down [63].

1.3.2 Activity Recognition

Activity recognition is an active field of research that uses various sensors and algorithms to observe subjects and recognize the activities they are performing. Different applications in activity recognition focus on a wide range of activities. Recognizing whether a subject is moving in ways such as jumping or walking [49], identifying a user's common destinations in a city [34], and differentiating whether a user is taking medication, making cereal, or eating cereal [45] are all examples of tasks distinguished by activity recognition systems. Although the work proposed in this dissertation does not lie in the field of activity recognition, the two nonetheless have ideas and techniques in common.

Several types of sensors can be used to observe interactions with objects, such as RFID readers [8], motion detectors and accelerometers [57], as well as electric current and water flow detectors [35]. In each case, the sensors measure approximations of object usage: with RFID readers, for example, proximity of a hand and an object is used as a proxy for object interaction; with accelerometers, movement of the object serves as a proxy.

Because the starting and stopping points of tasks typically must be inferred, various techniques are used to determine when the individual has changed tasks. One way to do this is to perform the activity recognition based on sensor readings during sliding windows of time [39, 35]; an alternative is to include task changing as part of the model, such as in a Hidden Markov Model where one or more nodes represent each task, and transitions are possible between the nodes of the different tasks [39].

Various techniques have been used to analyze the collected sensor data, including probabilistic methods and decision trees [39, 47, 56, 15, 35]. These vary in the types

of features that are analyzed. Within a 30-second window, the RFID data collected in [35] is examined to determine whether or each tag was detected, like the Detected feature that is be presented in Chapter II. The eleven single-state HMMs (one for each task) or single eleven-state HMM presented in [39] really examine how frequently a sensor is detected, which the Count feature in Chapter II and the Object Misuse feature in Chapter III are similar to, while the much larger HMM or the DBN used in [39] additionally look at the order in which a task is performed like the Order feature in Chapter II and the Edit Distance feature in Chapter III.

The differences between features used in this dissertation and those used in previous research reflect the fact that this dissertation has a different goal than activity recognition research—inferring information about the individual performing a known task rather than inferring which task is being performed.

Other recent work has automatically developed models for activity recognition from instructional web pages [65] or common sense databases such as the Open-Mind Indoor Common Sense database [45]. In [45], for example, the information mined from the database was converted into Markov Random Fields (MRFs); temporal relationships were represented using chain graphs in which a series of MRFs represented a time slice and links are created between nodes in different slices. These models have also been improved using sparsely labeled sensor data, to address gaps in the common sense databases [43, 44]. This approach is interesting for future research in the assessment of cognitive impairment, to replace the manual creation for each task of a partial-order plan and manual identification of appropriate numbers of object interactions that were used in the Edit Distance and Object Misuse features presented in Chapter III.

Note that there are also approaches to activity recognition research that are not

based on the analysis of interactions with objects. One example is Opportunity Knocks, a system that uses GPS to track a user as she uses public transportation to travel around a city. The system learns the user's common destinations and uses knowledge of the mass transit system to identify situations where it believes the user may have gotten on the wrong bus or be traveling in the wrong direction. When this occurs, the system makes a "knocking" noise to alert the user and then gives information about the presumed destination, the problem detected, and a remedy for that problem (what steps to take to get to the destination) [40]. This is still within the field of activity recognition because these locations are often associated with the activities that will be performed there. Other work on activity recognition uses small sensing platforms that collect accelerometer data and various other types of sensing data sometimes including light, barometric pressure, temperature, compass direction or sound readings. This data is then used to distinguish between activities such as riding up an elevator, jogging, and riding a bicycle [31] or between types of exercise such as running, cycling, and using an elliptical trainer or stair machine [10]. Another approach is to use a smaller number of extremely data-rich sensors such as video cameras or microphones. For example, Zhang and Gong employ video cameras to distinguish among ten basic actions including running, walking, and bending, as well as jumping jacks, jumping forward, and jumping in place [22].

Although there exist a wide range of paradigms for observation of activity performance, I focus on object-usage for two reasons. First, object usage has been employed to perform activity recognition at the scope of the activities that should be observed for the assessment of cognitive impairments. That is, object usage is appropriate for observing users perform daily tasks rather than the way in which a user is moving. Second, object usage patterns give insight to the specific processes

undertaken by a subject when she performs activities. If a subject forgets certain steps in the performance of a task or orders them oddly, object use patterns present an excellent and natural way to observe this.

Other approaches involving activity recognition have been motivated by assessing cognitive impairments. These include the use of motion detectors and contact switches to broadly observe several activities of daily living (ADLs) [16, 6], and using sensors on pillboxes to specifically observe medication taking, an instrumental ADL (IADL) [28]. Although the motivation for these studies is the assessment of impairment, they have primarily focused on the feasibility of observing the individual, and no assessment of the individual performing the task has been reported. Additionally, Hoey, et al use an estimate of the subject's functionality in their system observes handwashing in order to assist a user with dementia complete the task correctly. This estimate is updated over time as the user completes the handwashing task [20]. While this approach seems very promising, there has been no analysis of the accuracy of the estimate.

1.3.3 Biometric Identification

Although object-use analysis is frequently used in activity recognition, to the best of my knowledge, there has not been other work done on recognizing or identifying individuals—as opposed to their activities—based on their object-use patterns. However, significant researched has focused on identifying individuals using other biometrics. In this section, I review previous work studying two related forms of biometric identification, typing patterns and gait recognition. Unlike other forms of biometric identification such as fingerprint reading or retinal scanning, these two forms of biometric identification use behavioral patterns.

Keystroke Dynamics

Keystroke dynamics is a form of biometrics that analyzes users' typing patterns. It is used in computer security, most commonly to 'harden' passwords, or add a layer of biometric security to confirm a user's identity. Peacock [42] provides a good overview and comparison of the work in this area. While most systems analyze the timing of a user's keystrokes, modified keyboards may also be made pressure-sensitive to add additional feature to the data [12]. By using a combination of three methods—average and the standard deviation of feature times, rhythm of striking, and a comparison of the ordering—Hocquet, et al. have been able to achieve an error rate of just 1.8% [19]. Additional research has been to apply keystroke dynamics to mobile phones to identify users entering phone numbers and writing text messages, in particular because mobile phones are frequently stolen [11].

Gait Recognition

In comparison to keystroke dynamics, which is used to confirm the identity of a user, gait recognition aims to supplement the physical security of environments such as airports and banks by identifying individuals who may pose a threat. Gait recognition identifies individuals using patterns of movement [60]. It is frequently performed in conjunction with automatic face recognition, which uses a photograph or a series of photographs of a face to identify individuals [3].

Gait recognition is often approached by transforming the image of the individual into a binary silhouette and then creating a model of the individual or measuring several features such as silhouette width [7]. Performance in gait recognition systems is frequently compared using the data from the HumanID Gait Challenge Problem, data which was collected by the University of South Florida in a variety of conditions,

including with two different kinds of shoes, and carrying or not carrying a briefcase [7]. Using this data, the Gait Challenge Problem consisted of several problems of varying difficulty, based on how the test cases are different from the training data. For example, one of the test sets presents individuals who are walking on grass when all the training data had individuals walking on concrete. Currently, one of the best systems correctly identifies the individual 100% of the time on simpler test sets, even with false negative rates of just 1%. On the harder problems, the correct verification rates drop considerably, down to 36% when the same false negative rate is used [7].

CHAPTER II

Recognition of Individuals

In this chapter, I study the first research question: can an individual be recognized by her object-use patterns from the performance of a task? I address this question with a study in which I used electronic sensors to observe ten individuals perform a basic task (making a cup of coffee). I then developed features that I hypothesized would correspond to individuals' object usage patterns and I learned decision trees from those features. This chapter presents that study, including the features that were developed for it, and the results of applying those features to the collected data.

2.1 Experimental Methodology

2.1.1 Selection of a Task

Several criteria were used in the selection of a task for subjects to perform. The task should be an activity of daily living that is performed by many people on a regular basis, for the long term goal of identifying pattern changes that may indicate a cognitive impairment. An ideal task would also be relatively constrained in how it may be performed, but would also have some natural variance in its performance. Finally, it should be possible to perform the task in an instrumented laboratory. The task of making a cup of coffee was chosen for this experiment because it is an


Figure 2.1: Equipment Used in the Study. The Glove with an RFID Reader Attached (l) and the Coffee Grinder with Several RFID Tags Attached (r).

excellent fit for all of these criteria.

2.1.2 Selection of Technology

As discussed earlier in section 1.3.2, Radio Frequency Identifier (RFID) technology has been used successfully to study object-use interactions in several activity recognition projects. RFID uses tags placed throughout an environment and readers which detect nearby tags. An important advantage of this technology is that one can use passive RFID tags which, while not as accurate as active tags, require no battery source, meaning they can be placed throughout an environment without need for cords or batteries that need to be replaced. Other advantages to using RFID are that it has a false positive rate of essentially 0% [39], tags are available in small sizes (approximately the size of a postage stamp) and they are inexpensive (less than \$0.20) [47].

In earlier work done by Intel, an RFID reader placed on a glove or bracelet was used to detect tags that are in close proximity to the hand (within 10cm) [14, 39, 47, 53]. Activity recognition is performed with the assumption that detected objects

are the ones with which a user is interacting. Because Intel's wireless iBracelet was not available in time for use in this study, a wired system was used consisting of a commercially available RFID reader and tags created by Phidgets, Inc. ®.[1] This glove is shown in Figure 2.1, along with the grinder to show the RFID tags. Because the glove has a very short range, several tags were needed on most objects to allow reliable detection of object usage.

2.1.3 Experimental Setup

Ten subjects were recruited to participate in the experiment. For each trial, the subject was instructed to make a cup of coffee as if about to drink it, including adding sugar and creamer as preferred. Subjects wore a glove outfitted with an RFID reader on their right hand, but were told to ignore it as best they could. Each subject participated in ten trials, spaced out with generally at most one per day, so that the trials would reflect normal patterns of use, rather than artificial patterns created by performing trials repeatedly one after another[2].

Subjects were given a brief tour of the instrumented lab before their first trial, and those who did not know how to make coffee were given basic instructions. These instructions were as general as possible, and no physical demonstration of the coffee-making process was given. The experimental setup consisted of a coffee maker, one cup, one spoon, a coffee bean grinder, and a cabinet containing a bag of filters, a bag of ground coffee, a bag of coffee beans, and a canister each of creamer and sugar. Each item was tagged with multiple RFID tags and was put in the same location

[1] The studies presented in Chapters III and IV use one of the bracelets from Intel. This bracelet overcomes several of disadvantages of the previously used glove, including having a longer range and operating wirelessly.

[2] In some cases, the availability of the subject required more than one trial per day; five subjects performed two trials on the same day at least once and one of those performed six trials on the last day of the subject's availability.

before each trial, oriented in the same direction. (The bag of filters did not have an obvious front and thus may have been reversed between trials.)

2.2 Machine Learning Approach

The data collected by the sensors during the trials was then analyzed by an algorithm that learns decision trees. Decision trees are the simplest form of classifier, and were used in this experiment because they form a reasonable starting point for my investigation. A decision tree takes a set of properties as input and outputs a "decision" by following the path down a tree from the root. This path is determined by tests which are performed at each internal node encountered on the path, and the decision is dictated by the leaf node that is reached on this path. I used the C4.5 decision-tree algorithm, which is based on the ID3 tree-inducing algorithm; C4.5 is modified to avoid overfitting the data, the condition where an overly complex tree is created that is less accurate than a simpler one would be [50]. Additionally, C4.5 has been used successfully in activity recognition [35].

A key question for this type of study is the proper definition of a feature set. In this paper, I use a layered approach with five types of feature at three levels of granularity. Formally, each feature is defined as $< T, G, E >$ where T is one of the five feature types defined in section 2.2.2 and G is one of the three levels of granularity defined in section 2.2.1. E is an ordered set of entities (tags, groups, or objects) the feature is applied to. As described below, for most feature types E is a set with a single entity, though with Order it is an ordered set of two or three entities.

2.2.1 Observation Granularity

Observation granularity refers to how specific or general an observation is. The classifier may use the more general fact that a subject touched the coffee-maker or

more specifically that she touched the left-most sensor on the lid of the coffee-maker. Because many of the objects used in the study had multiple tags affixed to them, I considered three levels of abstraction:

1. Tag: Interaction with an individual tag affixed to an object.

2. Group: Interaction with any of a group of tags that are equivalent except for the orientation of the object (e.g., the tag on the left side of the bag of ground coffee and the tag on the right)

3. Object: Interaction with an object; that is, any of the tags on the object were detected.

At times, these are functionally equivalent. The carafe, for example, has only a single tag, so there is no difference between tag, group, and object interactions on the carafe. For several objects like the bag of ground coffee, though, several tags divided into multiple groups allow for different patterns to be detected at each granularity. Table 2.1 gives a list each object used in the experiment and all the groups of tags on that object.

2.2.2 Feature Type

Five types of features are used to measure the subject's interaction with each entity (tag, group, or object). The five types of features are:

1. Detected: A binary feature that is positive iff there was any interaction with an entity.

2. Count: A scalar feature that records the number of times interaction with an entity was observed.

Objects	Groups
Coffee maker	Lid, power switch
Carafe	Carafe
Mug	Mug
Spoon	Spoon
Coffee grinder	Top row of tags, middle row, bottom row
Left cabinet door	Left cabinet door
Right cabinet door	Right cabinet door
Ground coffee	Top tags on front and back, bottom tags on front and back, tag on bottom, tags on sides
Coffee beans	Top tags on front and back, bottom tags on front and back, tag on bottom, tags on sides
Filters	Tags on sides, tag on bottom
Creamer	Tags on top row, tags on bottom row
Sugar	Tags on top row, tags on bottom row
Faucet	Faucet

Table 2.1: List of Tag Groups and Objects

3. Total Duration: A scalar feature that records the total amount of time interaction occurred with an entity.

4. Average Duration: A scalar feature representing the average time of interaction with an entity: this is a computed feature, equal to the Total Duration divided by the Count.

5. Order: A binary feature that is positive iff an arbitrary two- or three-entity ordering is observed.

With the exception of Order, each feature type is calculated while considering the interactions with a single entity during a single trial. For example, Detected may be calculated using the tag underneath the coffee maker's power switch ($<\mathit{Detected}, \mathit{Tag}, \{\mathit{Coffee\text{-}Maker\text{-}Power\text{-}Switch}\}>$). This will calculate whether that tag was detected during a given trial. Detected would be calculated again for the tag on the mug ($<\mathit{Detected}, \mathit{Tag}, \{\mathit{Mug}\}>$), and for every other tag. A specific mea-

surement, then, consists of applying one feature type to one particular entity at one level of granularity while analyzing one trial. When analyzing the accuracy of using specific sets of measurements to classify trials, the specified feature types will be applied to all of the entities at the specified levels of observation granularity.

Order is the one slightly different feature type because it is applied to two or three entities, rather than to a single entity. It determines whether a specific ordering of interactions is observed during a trial. This ordering may consist of two or three entities, for example ($<$*Order*,*Object*,$\{$*Ground-Coffee*,*Filters*,*Spoon*$\}$ $>$) For simplicity, only one level of granularity is used within a specific ordering. Even with this simplification, however, considering the 70 tags, 25 groups, and 13 objects in the experiment, allows over 300,000 possible orderings. As a result, considering all possible orderings of tags comes at a significant cost of performance. This will be discussed further in sections 2.3.4 and 2.4.

2.2.3 Pre-Processing

Although the RFID reader and tag system provides accurate and generally reliable results, an individual tag is sometimes found and lost in quick succession, either when it is near the maximum distance from the reader at which it can be sensed, or if the reader moves rapidly. In order to smooth the data, a pre-processing step can be performed on each trial prior to analysis. This step looks for consecutive accesses to the same tag, group, or object within 0.5 seconds. When this is found, the records of the two accesses at that level are merged into one, hopefully providing a more accurate model of the subject's actual behavior. This means that when a subject moves her hand over several tags on the same object, the action will be interpreted as one continuous interaction at the object level (which it is) though no change will be

made at the tag level. Using pre-processing, a tag near the maximum range at which it can be picked up, causing it to be lost and found repeatedly, will be interpreted as one interaction (which it again is); it also means that when she quickly draws her hand away from a tag and then puts her hand on the tag, the action will be interpreted as one continuous interaction (even though it's not a continuous interaction). Section 2.3.5 discusses the effect of this pre-processing.

2.2.4 Simplifying Behavioral Differences

I performed one additional type of pre-processing on the collected data. The task of making coffee was chosen in part because there is a natural variance in how the task may be performed. However, looking at whether a subject puts creamer or sugar in her coffee has the potential to simplify recognition too much by partitioning the subjects into four sets. A similar concern is identified using domain-specific knowledge: using sensor data from the creamer and sugar may be misleading if an individual is preparing coffee for someone else; while most actions should not be affected, different patterns of the use of creamer and sugar may confuse a system that could otherwise correctly recognize the subject. For that reason, in most of the analysis, all data collected about creamer and sugar interactions are removed before the analysis is performed (section 2.3.6 includes this data for comparison of the results).

Like the first concern given, a subject's use of either whole beans and a grinder or of ground beans may potentially have over-simplified the recognition task. However, because disregarding the data collected from tags on the ground coffee, whole beans, and grinder would remove one-fifth of the data collected, I did not eliminate it. I note, though, that one subject used whole beans, and that subject used whole beans

in every trial. For that reason, one of the ten subjects can be distinguished very easily from the other nine.

2.3 Results

By using ten subjects who performed ten trials each, I obtained 100 trials for analysis. I then used a ten-fold cross-validation process, repeatedly using 90 trials as training data for C4.5 and reserving the remaining 10 trials as test data. In each iteration, the training data contained 9 trials for each subject, with the tenth reserved for testing data; however, the learning system did not use the information that there is exactly one trial per subject during the classification process.

2.3.1 Full Feature Set

Using the full feature set without the creamer and sugar data, the system achieves a 73% accuracy. Three of the ten subjects are correctly recognized in all ten of their trials, and eight of the ten are correctly recognized in at least half of their trials. The remaining two are correctly recognized in only 2 and 4 of their ten trials, respectively. A confusion matrix is shown in table 2.2.

The ten decision trees produced (one for each fold of the cross-validation) have an average of 12.2 internal nodes and an average maximum depth of 7.8. All three levels of observation granularity are used at least once, as well as all five feature types, though Order features make up well over half the internal nodes.

2.3.2 Influence of Observation Granularity

While the first experiment considered all features, I now consider the possibility of using a subset of the features. There are at three reasons for investigating the success when only certain subsets of the features are used in the recognition process.

	Inferred Subject									
Truth	A	B	C	D	E	F	G	H	I	J
A	7					1	2			
B		8		1						1
C	2		2		3		1	1		1
D				10						
E		1	1		4	1		1		2
F	1			1		7	1			
G							10			
H			1					9		
I									10	
J		1	1		1	1				6

Table 2.2: Confusion Matrix of the Full Feature Set

Observation Granularity	Accuracy	Subjects Correctly Recognized in ≤ 5 Trials
Tag	72%	C,E
Group	72%	C
Object	69%	C,E

Table 2.3: Accuracy at Each Level of Granularity

First, observing which features are more or less accurate individually can be used as a heuristic to determine their relative importance to the system. Second, the impact of the Order features deserves particular attention because, as will be discussed in section 2.3.4, using Order features increases run time drastically. Finally, this information may influence engineering decisions made in future systems—the effort to identify groups and objects, for example, may not be necessary if the system is just as accurate using only observations at the tag granularity.

To begin, I consider the influence of the different levels of observation granularity. Table 2.3 shows the accuracy of the system with each level of granularity. This analysis uses all five feature types.

The system's accuracy is only slightly affected by using tag interactions only or group interactions only rather than interactions at all three levels of granularity. At

Interaction Measure	Accuracy	Subjects Correctly Recognized in ≤ 5 Trials
Detected	59%	A,C,E,F,J
# of Times Detected	65%	C,E,J
Total Duration	75%	E,F
Average Duration	58%	C,D,E,F
Order	70%	C,E,J

Table 2.4: Accuracy of Individual Feature Types

69%, object-level interactions perform the worst, but still only 4% below the accuracy set by all three interaction types together. These results indicate that the effort of combining tags into groups and objects may not have much of a payoff in terms of increased accuracy. Also given for each level are the subjects who were correctly recognized in five or fewer trials.

2.3.3 Influence of Interaction Measure

I next consider the success of the system using only a single interaction measure, each computed using all three levels of observation granularity. Table 2.4 shows the results. As with the levels of granularity results, all subjects who were correctly recognized in five or fewer of the trials are also listed in the table.

The accuracy of the individual interaction measures clearly varies with a seventeen percent gap from the highest accuracy to the lowest, and several measures distributed between them. Once again subjects C and E are difficult to recognize correctly, with subject E listed in all of the feature types and subject C being listed in all but one. A very surprising result is that using the Total Duration feature type alone actually has a slightly higher accuracy than the full system using all five feature types. This result will be discussed further in section 2.4.

2.3.4 "All But Order"

The subset of using all feature types except for Order was originally looked at because performance of the system when Order is used is not ideal. With the creamer and sugar data ignored, the system takes an average of 8 minutes to compute the features and learn one tree, then perform subject recognition for ten trials. Including the creamer and sugar data degrades performance significantly, requiring an average of 40 minutes for the same task. Since performing recognition is still very quick, taking less than one second, this performance may be considered acceptable. However, even learning time must be bounded; moreover, the performance problems are likely to be exacerbated by more complex environments. Observing subjects performing larger and more complex tasks may require several times as many sensors as were used in this experiment. Using much larger time frames (for example, a year's worth of data instead of just ten trials) complicates this problem further. Additionally, allowing a user to interleave actions from multiple tasks may prevent a system from simplifying the learning process by only considering sensors relevant to a single task.

Although the number of features computed for other interaction measures grow linearly, Order undergoes cubic growth since using Order involves generating and considering a large number of possible two- and three-step sequences. I thus repeated the analysis, using the full feature set except for the three Order features (Order applied to each level of observation granularity). This analysis also adds understanding to which features are important in performing recognition.

As expected, processing time decreases in this case. It takes, on average, less than five seconds to learn one tree and perform recognition ten times, a speed up of two orders of magnitude from the eight minutes when Order is also considered. Surprisingly, however, accuracy increases to 77%, the highest level observed without

Interaction Measure	Accuracy Before Pre-Processing	Accuracy After Pre-Processing	Subjects Correctly Recognized in ≤ 5 Trials
All Features	73%	70%	A,C,E,F,J
Detected	59%	59%	A,C,E,F,J
# of Times Detected	62%	65%	C,E,J
Total Duration	75%	75%	E,F
Average Duration	58%	58%	C,D,E,F
Order	70%	70%	C,E,J
All But Order	77%	77%	C

Table 2.5: Effects of Pre-Processing

using creamer and sugar data. This also has the best worst-case performance, missing only half of the trials of subject C, its worst subject. Like the result from the previous section, the fact that this outperforms the system when all five features types are used will be discussed in section 2.4.

2.3.5 Influence of Pre-Processing

As discussed in section 2.2.3, I created a pre-processing step that smooths the data in situations where a single interaction with an entity might be recorded as several interactions that occur in rapid succession. The results of using this pre-processing (which was not used in any previous results) are shown in Table 2.5.

It turns out that this pre-processing technique has little effect on the accuracy. By definition, both the detected and Order features are unaffected by pre-processing. Of the other three feature types, the accuracy of two remain unchanged while using # of Times Detected becomes only slightly more accurate using pre-processing. When all features but Order are used, the accuracy remains the same, while all five features used together actually becomes slightly less accurate.

Interaction Measure	Accuracy Without Creamer and Sugar	Accuracy With Creamer and Sugar	Subjects Correctly Recognized in ≤ 5 Trials
All Features	73%	70%	C
Detected	59%	59%	A,B,C,E,F
# of Times Detected	62%	62%	C,E,J
Total Duration	75%	79%	
Average Duration	58%	66%	C,D,E
Order	70%	67%	C,J
All But Order	77%	78%	

Table 2.6: Effects of Including Creamer and Sugar

2.3.6 The Use of Creamer and Sugar

Creamer and sugar data have been ignored, primarily because including them should simplify the problem, partitioning the full group of ten subjects into four smaller groups. To see how much of an impact this data would have, several trials were repeated now including the creamer and sugar data. The results are shown in table 2.6.

The system does generally improve its accuracy, although not by much. Using the total duration feature type alone records the highest accuracy achieved by the system, going up four percent to 79%. All but Order also improves slightly to 78%. In both cases, the worst subject is now correctly recognized in 6 of the 10 trials.

2.4 Discussion

This work provides preliminary evidence that individuals can be fairly consistently recognized based on their object-use fingerprint, with the system accurately recognizing individuals more than three-quarters of the time with the best-performing feature set. Additionally, with this feature set, the system recognized every subject in at least half of the subject's trials.

There were some surprising results, however, particularly that using all five fea-

ture types together did not achieve the best results. Most notably, removing Order and using the other four feature types produced the highest accuracy. I hypothesize that this is a result of the sheer number of Order features causing overfitting. The results using creamer and sugar appear to support this hypothesis; although most feature types achieved slightly higher accuracy when creamer and sugar data was added, Order had slightly lower accuracy as did all five features used together. The additional entities considered when using creamer and sugar cause a large increase in the total number of Order features, and the increased overfitting may have been more negative than the additional information about creamer and sugar was positive. One possibility for simultaneously addressing the overfitting and performance problems of Order while still using the feature is to identify a smaller number of interesting orderings for the system to use. This could be done manually using domain-specific knowledge, by pre-testing or by data-mining from the web (the latter of which would be interesting in its own right).

Other potential areas for future study include replicating the experiment described here on more and different types of subjects, on a broader range of activities, and in naturalistic settings, so as to validate the generality of my preliminary results.

CHAPTER III

Assessing Cognitive Impairments

This chapter investigates the second research question of this dissertation: can the severity of an individual's impairment be assessed by observing the individual's object-use patterns from the performance of a task?

I developed four metrics that could be automatically computed from the collected sensor data and that I hypothesized would correlate with cognitive impairments. I hypothesized that these patterns of errors made in the performance of such activities are associated with the severity and type of a patient's cognitive impairment and further, that wireless sensors could be used to accurately detect those errors. I was concerned both with predicting overall neuropsychological integrity, and with identifying more specific neuropsychological profiles, such as isolated difficulties with memory, attention, or executive reasoning.

3.1 Experimental Methodology

3.1.1 Selection of Task and Technology

For the same reasons listed in sections 2.1.1 and 2.1.2, I selected the task of preparing coffee and decided to use RFID to observe the task.

Two main changes were made in this study relative to the previous one: the

Figure 3.1: RFID Equipment Used in the Experiment. The Intel iBracelet RFID Reader (l) and Canister of Ground Coffee (r).

preparation of a pot of coffee rather than a cup and the use of the Intel iBracelet as the worn RFID reader rather than the RFID glove. In the previous study, subjects both prepared a pot of coffee and, once it was ready, poured themselves a cup and added cream or sugar as if they were about to drink it. In this study, subjects only prepared the pot of coffee, so far as to turn the coffee maker on to begin brewing—this change is discussed more in the following section. The Intel iBracelet is an improvement from the previously used RFID glove because it does not decrease the subjects' fine dexterity as the glove did and because it is wireless, not limiting mobility or creating a tripping hazard as the glove did. The Intel iBracelet has a range of about 10cm [14, 53] and is shown in Figure 3.1 with the canister of ground coffee.

3.1.2 Experimental Setup

Twenty-five subjects with traumatic brain injuries were recruited to participate in the study. Participants were asked to perform five trials each, with no more than one trial per day, and the iBracelet RFID reader was used along with many RFID tags to

observe the individual's activity performance. The results presented in this chapter are from the sixteen of those subjects for whom full neuropsychological evaluations were available. Of these, thirteen completed at least three trials and ten completed all five trials.

For each trial, the subject was asked to start a pot of coffee brewing—putting in water, a filter and ground coffee and turning the coffee maker's power on. Subjects were each asked to perform five trials on five different days. The subjects performed the trials in a kitchen at the medical center where they were receiving care for their cognitive impairments. The coffee pot and all supplies were placed on a counter in the kitchen, next to a sink for water. Subjects were asked if they knew how to make coffee and given basic instructions if they did not. No physical demonstrations were given. If subjects asked how much material to put in, they were told to use enough for six cups of coffee (about half the capacity of the coffee pot).

The material that was set out included the coffee pot and carafe, a canister of ground coffee, a bag of filters, a mug and a spoon. Twelve tags were used: four on the coffee pot, four on the canister of ground coffee, and one each on the other objects. Multiple tags are needed for some objects to reliably detect interaction due to the range of the iBracelet (the shorter range is desirable to avoid a higher rate of false positives).

The experimental setup was influenced by the fact that the subjects had cognitive impairments and were performing the task within the clinic. I placed the supplies on the counter, rather than away in cabinets, to make the task easier for the subjects to complete in order to avoid causing frustration by having subjects searching in an unfamiliar kitchen if they forgot where a supply was located. This also should not be necessary when observing subjects in a home environment.

Time Stamp	Tag Detected
1200503935	Carafe
1200503935	Filters
1200503936	Filters
1200503937	Filters
1200503939	Ground Coffee 3
1200503955	Ground Coffee 2
1200503956	Ground Coffee 1
1200503989	Coffee Maker Lid 3
1200503990	Coffee Maker Lid 3

Figure 3.2: Stream of Time-Stamped Interactions from a Portion of a Trial. When Multiple Tags are Placed on One Object, a Number Is Given Indicating Which Tag Has Been Detected.

Out of an abundance of caution and on the advice of the clinic staff, I also did not have subjects pour a cup of coffee once the pot had been brewed. This was to ensure that the individuals would not be handling hot liquids and decrease the potential of injuring patients at the medical facility. This should not be a barrier to using a similar system in-home since I expect that many cognitively impaired individuals regularly make coffee and, anecdotally, several participants in the study noted that they regularly made coffee at home.

3.2 Automatic Assessments

The sensor data collected in each trial consists of a series of time-stamped interactions with RFID tags, a sample of which is shown in Figure 3.2. From the collected sensor data, I computed four features that I hypothesized might correlate with the subjects' cognitive impairments. The features are increasingly representative of the task, ranging from very simple—how long it takes the subject to complete the trial—to much more detailed—how "far off" the subject's behavior is from a correct instance of task performance.

3.2.1 Trial Duration

I hypothesized that an impaired individual would take longer to prepare the pot of coffee than an unimpaired individual, as a result of confusion, mistakes, or simply performing steps more slowly. Therefore, the first feature I computed is the duration of the trial: how long it takes the subject to complete the task (measured in seconds).

Given a trial with n detected interactions, I define this feature using the following formula: $TrialDuration(t) = EndTime_n - StartTime_1$ where $StartTime_i$ and $EndTime_i$ indicate the start and end times of the i^{th} action in the temporal sequence of trial t. That is, the feature is simply measured as the time between the first interaction that is detected and the last.

3.2.2 Action Gaps

Note that the trial duration feature is extremely simple and has very limited representational power: it would not distinguish between two people who are behaving in very different ways, provided only that the total amount of time for each trial was the same. I next moved to a somewhat more representational feature, which is based on the hypothesis that more severely impaired individuals might have more periods during which they were not interacting with any objects, on the assumption that during those periods they are considering what step to take next. The second feature measures these periods of inactivity during the trial which I call Action Gaps. I define the number of Action Gaps with length g of trial t:

$$ActionGaps_g(t) = \sum_{i=1}^{n-1} \begin{cases} 1, & \text{if } StartTime_{i+1} - EndTime_i \geq g \\ 0, & \text{otherwise} \end{cases}$$

I examine the number of brief action gaps using $g = 3$ seconds and the number

of longer action gaps using $g = 10$ seconds.

3.2.3 Object Misuse

I next moved to a feature that accounts for the specific objects used in task performance. One way of determining whether someone is being effective in carrying out a task is to examine the number of times he or she interacts with each object used in the task. I thus hypothesized that failure to interact with a required object—e.g. to "touch" the coffee filters—indicates a problem, as does an excessive number of interactions. For the simple task of making coffee, I manually determined a reasonable range of interactions with each object, shown in Table 3.1. The filters, for example, may be used once or twice—once to open the bag of filters and perhaps again if the user closes the bag in a separate interaction (remember that the tag is on the bag of filters, not individual filters themselves). Note that I do not state a maximum number of accepted interactions with the Ground Coffee or the Mug or Carafe (to get water) because these are difficult to define—unlike closing the lid which is one distinct interaction, putting ground coffee in the coffee pot may involve touching the ground coffee multiple times to get several scoops and filling the water may require using the mug multiple times to fill the coffee pot. The Spoon is not included in this feature because it was rarely detected—it would also be difficult to use since it is not required but, like the ground coffee, may be used multiple times.

For each trial, I then computed the number of times the subject interacted with each object b ($touch_b$) and determined whether that number was outside the accepted range and, if so, how far outside the range it was.[1]

[1] I also investigated a few variations of the Object Misuse metric, to address the possibility that touching an object too many times could have a disproportionately large impact compared with touching too few times. These variations were approximately as successful as the basic metric here; because the variations and results did not appear to be interesting, they are not presented in this dissertation.

Object b	min_b	max_b
Lid	2	2
Ground Coffee	1	∞
Filters	1	2
Mug or Carafe (Getting Water)	1	∞
Power Switch	1	1

Table 3.1: Number of Accepted Interactions for Each Object

$$ObjectMisuse(t) = \sum_{b \in Objects} \begin{cases} 0, & min_b \leq touch_b \leq max_b \\ min_b - touch_b, & touch_b < min_b \\ touch_b - max_b, & touch_b > max_b \end{cases}$$

3.2.4 Edit Distance

My final approach to automatically measuring performance moves even further in the direction of matching the subject's performance to an explicit model of correct performance. With this approach, I begin with a representation of how to make coffee—a "plan" for the task. The plan I used in my analysis is a partial order over object interaction, depicted in Figure 3.3, with "Water" indicating using the carafe, mug, or both to get water from the sink and put it into the coffee maker. Note that the use of the partial order allows me to score as "correct" alternative task executions that are reasonable: I score as correct both executions in which water is added before the filter and ground coffee and those in which those actions are reversed. However, I judge to be incorrect executions in which the power switch is turned on before the filters are used.

I then further constrain my plan for correct executions to those in which object interactions are not interleaved and using filters is followed directly with using ground

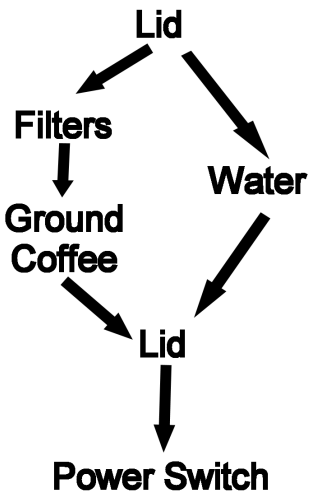

Figure 3.3: Partially Ordered Plan of Object Interaction for Making Coffee

coffee. These two criteria are added for the same reason: for a basic task like making coffee, I hypothesize that it is more likely that a mistake occurred than that the individual chose to interleave actions (like getting ground coffee, water, and then ground coffee again). Using filters and ground coffee are kept together because I view them as really one general action: putting ground coffee in the coffee maker.[2]

Although I manually created the plan to represent making coffee, other research on activity recognition has addressed the question of automatically constructing plans for everyday activities by mining the web for descriptions of these activities [65]. Such an approach could be adopted for extensions of this work.

Once I had a plan that models correct task executions in terms of object interactions, I next needed to define a measure of deviations from that plan. For that, I adopted the notion of edit distance, which is frequently used in the literature on

[2]The assumptions made in this model may be too constraining—perhaps many unimpaired individuals do interleave using filters and ground coffee with getting water, for example. This suggests a further elaboration, in which the plans are probabilistic—with the probabilities representing the plausibility of certain sequences being performed. This elaboration, however, is outside the scope of this dissertation.

natural-language processing [29], but which has also been used in prior work on activity recognition [39]. More specifically, I made use of the Levenshtein distance which allows the insertion, deletion, or substitution of a character [32]. I computed the distance between the sequence of observed object interactions and each of the legal executions of the plan for the task and used the smallest of these distances (assuming that the subject intended to use the ordering to which she was the closest).

Note that to compute the edit distance, I merged consecutive interactions with the same object (for example, multiple usages of the ground coffee were just shown once as long as no other objects are used in between). I then defined the Edit Distance of a trial t:

$$EditDistance(t) = min_{e \in \{LegalExecutions\}}(Distance(observed, e))$$

With the very simple plan, there are only two legal executions: the one in which placement of the filters and the ground coffee precedes the filling of the water canister, and the one in which these occur in the reverse order. Hence the Edit Distance feature is easy to compute, involving determination of just two distances.

The edit distance is intended to provide a fairly fine-grained measure of the relationship between the "correct" task performance, at least as modeled in my plan, and the subject's actual performance.

3.3 Neuropsychological Assessment

Neuropsychological impairments are assessed with a battery of tests that sample a broad range of cognitive domains. Many of these tests assess general functioning, such as intellectual ability. Others are very specific, having been chosen because they are known to be associated with functioning that is mediated by a specific brain

locus (e.g., left versus right hemisphere, anterior or lateral frontal lobe, specific subregions of the areas that mediate expressive or receptive language), or because they provide critical information about a cognitive domain that is central to performance of everyday activities (e.g., attentional shifting). The measures employed for this type of assessment are meticulously normed, often in the context of multiple samples, such that statements can often be made about a patient's performance relative to the population at large, to specific cohorts (e.g., those of same gender and similar age or education), or relative to specific clinical comparison groups (e.g., is the profile most consistent with a cerebro-vascular accident, dementia, or depression) [52, 17, 33]. The neuropsychological assessments I used are given in Table 3.2.

I obtained the results of neuropsychological tests from the patient records of the 16 subjects in this study to use as ground truth. I then computed the correlations of the computed features with the individual neuropsychological assessments listed, using an individual's average value over the five trials for each computed feature and applying one-tailed non-parametric evaluation. In addition to the individual neuropsychological assessments, I applied principal component analysis (PCA) to the complete set of neuropsychological assessment data for the subjects in order to examine how well the computed features correlate with larger trends in the assessment data. PCA is a standard statistical technique that finds linearly independent components that explain as much of the variance in the data as possible. Each component is a linear combination of the assessments where the sum of the squares of the component coefficients is one. The first principal component is the linear combination that has the largest possible variance; the second principal component is the linear combination that has the largest possible variance and is uncorrelated with the first principal component; the third is uncorrelated with either of the first two

Wechsler Adult Intelligence Scale (WAIS) III Verbal Comprehension (1)
WAIS III Perceptual Reasoning (1)
WAIS III Working Memory (1)
WAIS III Processing Speed (1,3)
Wechsley Memory Scale-Revised (WMS-R) Logical Memory I (3)
WMS-R Logical Memory II (3)
WMS-R Visual Reproduction I
WMS-R Visual Reproduction II
California Verbal Learning Test II (CVLT II) Total (1)
CVLT II Long Delay Free Recall (4)
CVLT II Recall Discriminability (4)
Trails A
Trails B (2,5)
Booklet Category Test (BCT) Error Total
Wisconsin Card Sorting Test (WCST) Concepts (3)
WCST Perseverative Errors (5)
Controlled Oral Word Association Test (COWAT-FAS) Total (1)
Animals (1)
Wide Range Achievement Test (WRAT) 4 Reading (1)
WRAT 4 Mathematics (5)
Peabody Individual Achievement Test-Revised (PIAT-R) Reading Comprehension
Peabody Picture Vocabulary Test-Revised (PPVT-R) (1)
Tactual Performance Test (TPT) Total (2)
TPT Memory (1)
TPT Location (2)
Finger Tapping Test - Dominant
Finger Tapping Test - Non-Dominant
Grooved Pegboard (GPB) - Dominant (2)
GPB - Non-Dominant (2)

Table 3.2: List of Standard Neuropsychological Assessments Used. Parentheses Indicate Principal Components in Which the Assessment has a Loading of 0.6 or More.

Component	% of Variance	Cumulative %
1	26.4	26.4
2	15.0	41.5
3	12.8	54.2
4	9.0	63.3
5	8.7	72.0

Table 3.3: Principal Component Analysis of Neuropsychological Assessments

components, and so on. To perform the PCA, some of the summary assessments in Table 3.2 were replaced with their component scores, a standard statistical practice.

The first principal component of the neuropsychological data accounts for 26.4% of the total variance in the data and the top five principal components together account for 72.0% of the total variance. Table 3.3 shows the first five principal components and the variance explained by each component. Table 3.2 indicates which factors, if any, each assessment (or any subtest of that assessment) has a loading of 0.6 or higher.

After the principal components were computed, a domain expert and member of the dissertation committee, Ned Kirsch, interpreted them. The first principal component includes a diverse set of measures of general intelligence. It appears to be a good proxy for general neuropsychological integrity, including measures of intellectual functioning, verbal and nonverbal reasoning, memory, and complex attention. The interpretation of the lower-order components is less clear, although the second could be seen as a measure of general motor integrity; the third as representing verbal memory and concept formation; the fourth, the ability to retain verbal information over time; and the fifth, strategy formation and modification.

3.4 Results

3.4.1 Assessing Neuropsychological Integrity

Recall that the main question I ask in this study is whether I can assess a patient's cognitive status by observing performance of an everyday activity using wireless sensor networks. The main result is quite promising: there is a statistically significant correlation ($p < 0.01$) between the Edit Distance feature and the first principal component of the neuropsychological assessments, which, as just described, can serve as a proxy for overall generalized neuropsychological integrity. Importantly, I did not find such a correlation with any of the simpler features (Trial Duration, Action Gaps, or Object Misuse). The ability to predict neuropsychological integrity, at least within the scope of this experiment and in particular for the population of TBI patients involved, is a strong indication that it is possible to conduct the types of automatic assessments that motivate this work. Figure 3.4 shows the plot of Edit Distance and the first principal component, with the regression line.

3.4.2 Assessing Other Metrics of Impairment

Although general neuropsychological integrity is a very important metric, it is also interesting to see how the features assess other metrics of cognitive impairments. The reason for doing this is based on domain practice—in addition to a concern with overall neuropsychological integrity, it is often important for a rehabilitation team to understand different aspects of a patient's impairment: does it involve memory impairment? Problems with focus of attention? Decreased motor coordination? There is a potential for these to be blurred in a single measure of overall integrity, important though that summary measure is. For instance, an individual whose impairment involves decreased motor coordination or processing speed may have

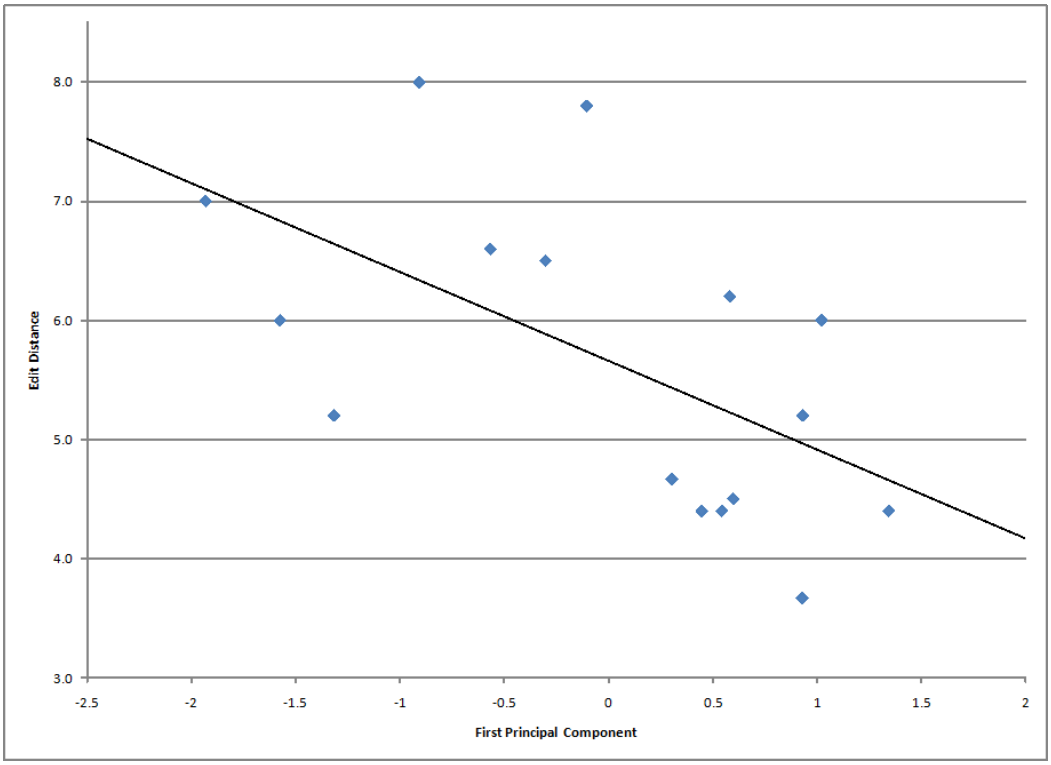

Figure 3.4: Plot and Regression Line of Edit Distance with the First Principal Component of the Neuropsychological Assessments

unimpaired executive function, and thus still be able to follow a "plan" for making coffee successfully, but perform the task more slowly. An increased Trial Duration might help tease out the types of cognitive difficulties facing this patient. To address this, I also look at the next four (the second through fifth) principal components as well as the twenty-nine individual assessments.

Additional statistical analysis such as a Bonferroni correction is required to state that a correlation between two variables exists with statistical significance. However, because of the large number of assessments and features, achieving statistical significance at this strict level would require the collection of data from a huge number of subjects—many more than were in the scope of this project. Nonetheless, the results on the individual tests in this section are important as exploratory data analysis and as a foundation for further research in the area. Although I recognize

that correlations at the variable level are of questionable significance because of the number of analyses performed, I am nevertheless presenting these findings because the coherence of the relationships and the number of associations I found provide important direction for future research. Additionally, while only suggestive, the variable level correlations provide tentative guidance in regard to further refinement of "markers" that clinicians can use when attempting to make a determination of the mechanisms for a patient's failure. Analysis of mechanisms (e.g., decreased processing speed vs. executive functioning) may then lead, in turn, to choices on the clinician's part about interventions that are specifically targeted to the underlying impaired cognitive mechanism.

Beyond this, it is noteworthy that I saw more correlations than would have been expected by chance (22 actual correlations for the individual assessments versus 14.5 expected by chance), especially when looking at what would be strict p-values if the Bonferroni correction were not required ($p < .01$: 8 actual correlations versus 2.9 by chance). Moreover, many of the identified correlations "make sense" from a neuropsychological standpoint, in a manner similar to the example in the previous paragraph. The results from correlations with principal components will also be presented, although the number of correlations with the second through fifth principal components (2) is what was expected by chance.[3]

Edit Distance

The Edit Distance feature achieved the best results with the individual evaluations as well, having a suggestive correlation with the fourth principal component

[3]These numbers include the results from the additional five variations of Object Misuse noted in Section 3.2.3 although those results are not presented here.

Feature	Correlations with Principal Components	# Suggestive Correlations with Individual Features
Edit Distance	1st ($p < 0.01$) Suggestive: 4th	7
Trial Duration	-	6
Action Gaps ($\geq$ 3s)	-	5
Object Misuse	-	3

Table 3.4: Summary of Results from Each Feature

	Computed Feature			
Assessment	Edit Distance	Trial Duration	Action Gaps $\geq$ 3s	Object Misuse
WAIS III Processing Speed		* #		
WMS-R Visual Reproduction II	*	*		*
CVLT II Total	*			
CVLT II Long Delay Free Recall	*	*		*
CVLT II Discriminability	*			*
Trails B			* #	
Animals	*	*	*	
WRAT 4 Reading			* #	
TPT Memory	*			
Finger Tapping - Dominant	*			
Finger Tapping - Non-Dominant		* #	* #	
GPB - Non-Dominant		* #	* #	

* indicates a suggestive correlation

notes additional coverage: a metric not also correlated with Edit Distance

Table 3.5: Suggestive Correlations Between Neuropsychological Assessments and Computed Features. Assessments with No Suggestive Correlations Are Not Shown.

as well as with 7 of the 29 (24%) neuropsychological assessments.[4] Recall that the fourth principal component appeared to represent the ability to retain verbal information over time. The correlations with individual evaluations are predominantly and compellingly with memory features; speculatively they could also be said to measure the integrity of the left-cerebral hemisphere and the capacity to engage in sequential and logical thinking. Generally the assessments that have suggestive correlations with Edit Distance are also factors with a high loading in the principal components with which Edit Distance is correlated but the slightly weaker interpretation is likely due to the less efficient analysis of individual assessments. Table 3.4 summarizes the Edit Distance results and compares them to the other features. Table 3.5 shows more detail, giving the assessments with which Edit Distance had a suggestive correlation.

Trial Duration and Action Gaps

Trial Duration and Action Gaps also proved to be promising features. Though neither had suggestive correlations with any of the principal components, they did with a number of neuropsychological assessments. Trial Duration had a suggestive correlation with 6 (21%) of the 29 neuropsychological assessments. Similarly, Action Gaps of 3 seconds or greater suggestively correlated with 5 (17%) of the neuropsychological assessments. These results are less coherent from a neuropsychological perspective than the Edit Distance results but the correlation between processing speed and Trial Duration is very logical. And while the results are not as good as the Edit Distance results, they are still valuable: between the three features presented thus far, there are suggestive correlations with over 12 (40%) of the 29 neuropsycho-

[4] I identify a suggestive correlation whenever there would be a statistically significant correlation if the Bonferroni correction were not needed. Because it is necessary, these correlations are not significant but are still of interest for their value in guiding future studies.

logical tests, including 5 that did not have suggestive correlations with Edit Distance. Suggestive correlations in Table 3.5 that provide additional coverage—correlations with a metric that did not have a correlation with Edit Distance—are marked with a #. I also tested Action Gaps of 10 seconds or greater but this only had a suggestive correlation with one metric (GPB - Non-Dominant which also correlated with two other features); I hypothesize that the poor result for this feature is due to the low frequency of gaps that long.

Object Misuse

The results from the Object Misuse feature were the least successful—as shown in Table 3.5 the feature had fewer suggestive correlations than Edit Distance, Trial Duration or Action Gaps of 3 Seconds, and none with assessments that were not also correlated with Edit Distance.

3.5 Conclusion and Discussion

I have presented an approach to using RFID-based sensing of individuals as they perform a simple task, with the aim of assessing their level of cognitive impairment. I presented four features, with increasingly representational power, that can be computed from the collected sensor data, and evaluated them using the results of the subjects' performance on standard neuropsychological assessments as well as with the principal components of those assessments. The most knowledge-rich feature I computed, Edit Distance, had a statistically significant correlation with the meaningful first principal component, a measure of general neuropsychological integrity. I also presented the results of exploratory analysis of the correlations between the four types of features and the individual assessments; these results are helpful to guide future research into other metrics of impairment without the need for a massive

amount of data collection.

It is interesting to note that the vast majority of trials resulted in coffee being made successfully. Although mistakes were frequently made, the subject almost always recognized and corrected them (for example, the lid of the coffee pot was often closed before all the water and ground coffee was in, but this was almost always recognized eventually and the lid reopened so the additional material could be put in). Therefore, if a computed feature could be created that determined whether or not the individual succeeded in making coffee correctly, it would have been of very little use—it might be able to assess the most impaired individuals but would be unable to differentiate the rest. I hypothesize that the computed features in this study were successful because they measured how efficiently an individual performed the task. Although Edit Distance would be able to detect forgotten steps, these forgotten steps rarely occurred since coffee was almost always made correctly. Instead, Edit Distance appears to have detected inefficiencies: extra steps that were caused by a poor decision at the time of the extra step or to correct for earlier mistakes. Likewise, it would be possible to have a very short Trial Duration if several necessary steps were omitted during the trial. If this had happened with any frequency, Trial duration would likely have been less successful, unless the possibility could be accounted for (such as interpreting a very short Trial Distance as a poor result). However, since almost all trials were completed successfully, Trial Duration was more simply a measure of how quickly an individual could correctly make coffee—a shorter trial indicated more efficiency in making decisions and performing the actions necessary to complete the task.

There are many practical concerns for the in-home implementation of a system that could automatically assess impairments. Compliance with the system is im-

portant since the user must wear the bracelet and complete the task to be assessed; individuals with an impairment may be particularly forgetful about doing this. Other sensor modalities, such as accelerometers placed on the objects, motion detectors, or current or water-flow sensors might be considered which do not have this drawback. On the other hand, the privacy implications of observing individuals in a home environment are important to address and these may be somewhat alleviated by using a system which can clearly be prevented from observing an individual's behavior (by taking the bracelet off).

3.5.1 Future Work

A great deal of future work remains, including collecting additional data and performing further analysis to investigate the suggestive individual correlations identified in this study. Additionally, observation of other kinds of impairments (particularly dementia) and longitudinal studies are necessary to understand the ability of these techniques and potentially to develop new techniques to observe change in an individual's performance over time. There are a number of ways in which the scope of the research can be expanded, particularly applying these assessment techniques to other activities beyond coffee making and using them in a home environment.

CHAPTER IV

Recognizing Cognitive Impairments

This chapter continues to investigate how object-use patterns can be used to learn about an individual's cognitive impairment. While the previous chapter addressed the question of whether an impairment could be assessed using object-usage patterns for an individual known to be impaired, this chapter addresses the third research question: is it possible to recognize whether or not an individual has a cognitive impairment based on those object-use patterns?

To answer this question, I asked individuals with Traumatic Brain Injury (TBI) and unimpaired individuals to perform the same task of preparing a pot of coffee while wearing electronic sensors to observe their object usage. I used the same four metrics developed in the previous chapter and I then investigated how successful those metrics were at recognizing which individuals were impaired and which were not. This chapter presents the results of that study which, though very preliminary due to the limited number of subjects, are promising for being able to distinguish between impaired and unimpaired individuals based on their object-usage patterns.

	Unimpaired Subject		Impaired Subject	
Pair	Age	Yrs. Educ.	Age	Yrs. Educ.
1	49.8	13	49.4	12
2	38.1	16	33.1	17
3	45.4	17	45.0	18
4	21.3	16	25.7	14
5	20.7	15	24.8	14

Table 4.1: Matched Unimpaired and Impaired Subjects

4.1 Experimental Methodology

4.1.1 Experimental Setup

The experimental setup used here is very similar to the setup used in Chapter III; subjects were asked to make coffee five times while wearing the Intel iBracelet, with each trial on a different day. Thirteen unimpaired subjects were recruited, to be compared with the twenty-five subjects with TBI. To address the concern that age and intelligence might have an impact on how individuals prepared the coffee, I selected the five pairs of subjects (each pair having one impaired individual and one unimpaired individual) that most closely matched each other in age and education (which is used as a proxy for intelligence). Information about the five pairs of individuals is shown in Table 4.1. Three of the impaired subjects and three of the unimpaired subjects completed all five coffee trials. One of the impaired subjects completed only two trials, and the remaining subjects (one impaired, two unimpaired) completed three trials. Although promising, this study is preliminary due to the limited number of subjects (no additional good matches were possible).

The unimpaired individuals were drawn from two different pools of subjects. Unimpaired subjects 1 through 3 performed the trials at the same MedRehab kitchen as the 25 impaired subjects. Unimpaired subjects 4 and 5 performed the trials at the Computer Science and Engineering building at the University of Michigan, in a

kitchenette that was set up to mirror the kitchen at MedRehab. Although these two subjects were asked to complete all the steps in preparing a cup of coffee, including pouring it into a cup for this study, I only used the data up to the point at which the subject had finished preparing the pot of coffee (the point at which the subject had finished preparation of the pot of coffee and started the coffee maker was manually recorded in each trial; this corresponded to the point at which the other subjects' trials were stopped).

4.1.2 Analysis

The four features presented in Chapter III are used again in this study. Based on the results from that chapter, only Action Gaps with a duration of 3 seconds are considered, not Action Gaps with a duration of 10 seconds. Likewise, only the standard measure of Object Misuse is considered, not the other variations mentioned in Section 3.2.3. In this study, these features were used to attempt to classify the individuals as either impaired or unimpaired.

4.2 Results

4.2.1 Edit Distance

I first examine Edit Distance, the most knowledge-rich feature proposed and the most successful feature in the experiments described in Chapter III. Figure 4.1 plots the average Edit Distance for each subject, with error bars showing the standard error for each subject. As can be seen, the impaired and unimpaired subjects can be perfectly separated: the largest average Edit Distance for an unimpaired subject is 4.0 and the smallest average Edit Distance for an impaired subject is 4.4.

Although the results are very good, this study is preliminary since the sample size is quite small. With more subjects, it would be unlikely that the results would

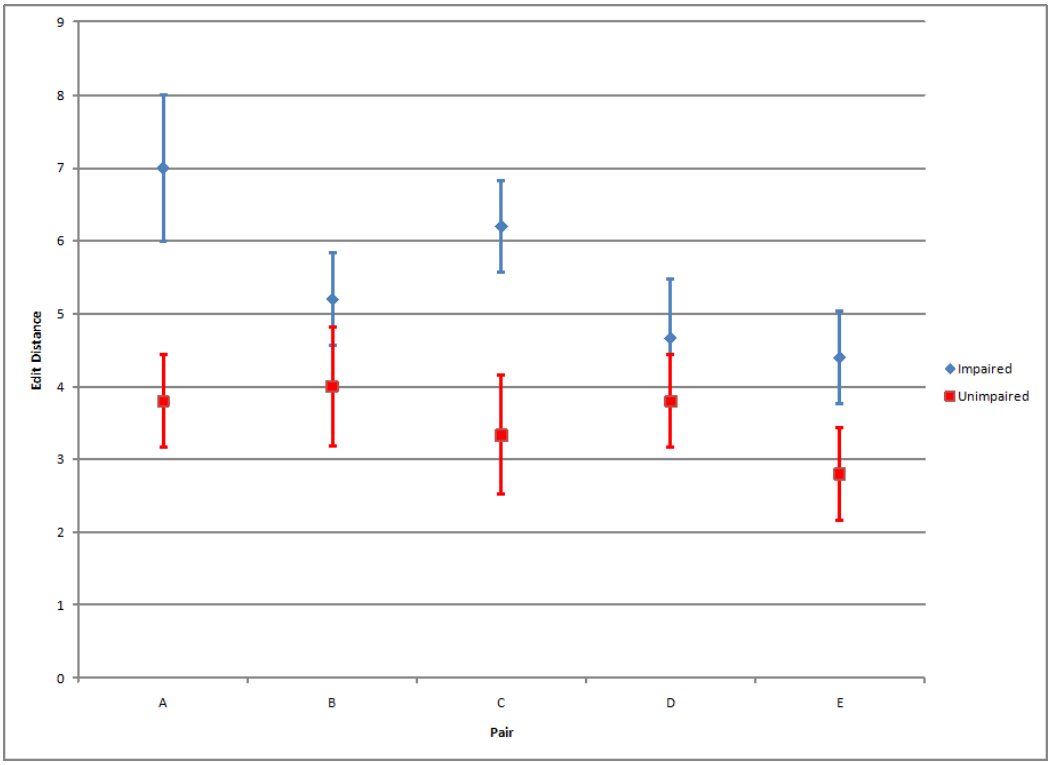

Figure 4.1: Plot of Edit Distance for Impaired and Unimpaired Subjects

continue to remain perfect—even if repeated with the same subjects, an error would not be unexpected since several subjects have error bars which cross the gap between the impaired and unimpaired subjects. Nonetheless, the fact that the subjects can be perfectly split based on the data that was collected is a very positive result and is evidence that differentiating impaired and unimpaired subjects may be possible with very good accuracy.

4.2.2 Other Features

Since it is not expected that Edit Distance would continue to perfectly classify subjects, it is interesting to examine the performance of the other features to direct future research in recognizing cognitive impairments. Figure 4.2 shows the plot of average Trial Duration and Figures 4.3 and 4.4 show average Action Gaps of 3 seconds or longer and average Object Misuse, respectively. For clarity, error bars are

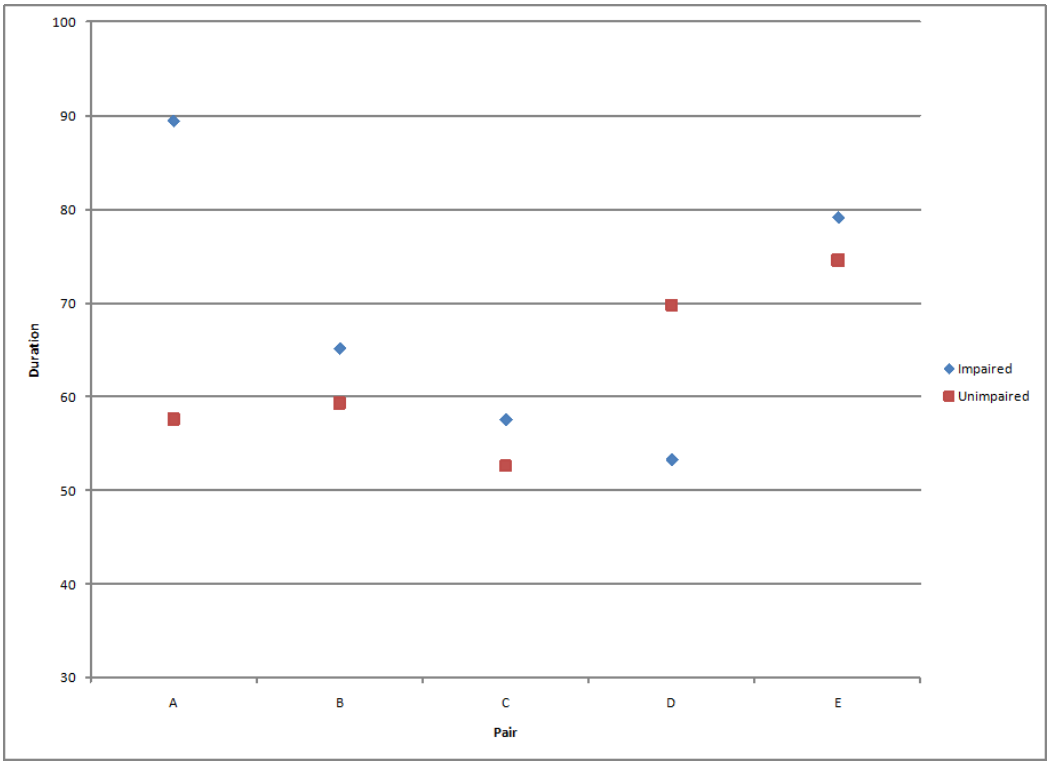

Figure 4.2: Plot of Trial Duration for Impaired and Unimpaired Subjects

not shown in these graphs.

None of these three features can be used to perfectly classify subjects, at least not individually. As a result, there's a tradeoff between correctly identifying more impaired individuals and correctly identifying more unimpaired individuals. A balanced outcome, where both sets of subjects are identified with similar success rates is one solution, but it might not be ideal. In particular, a false positive may be preferable to a false negative (since a likely response to a positive result would be to further investigate whether an impairment exists), so a lower threshold may be better. Alternatively, a higher threshold—where false positives are minimized—might be preferred since the subject can be evaluated repeatedly over time.

These tradeoff can be visualized using a Receiver Operating Characteristic (ROC) curve. An ROC curve plots the true positives (sensitivity) against the false positives

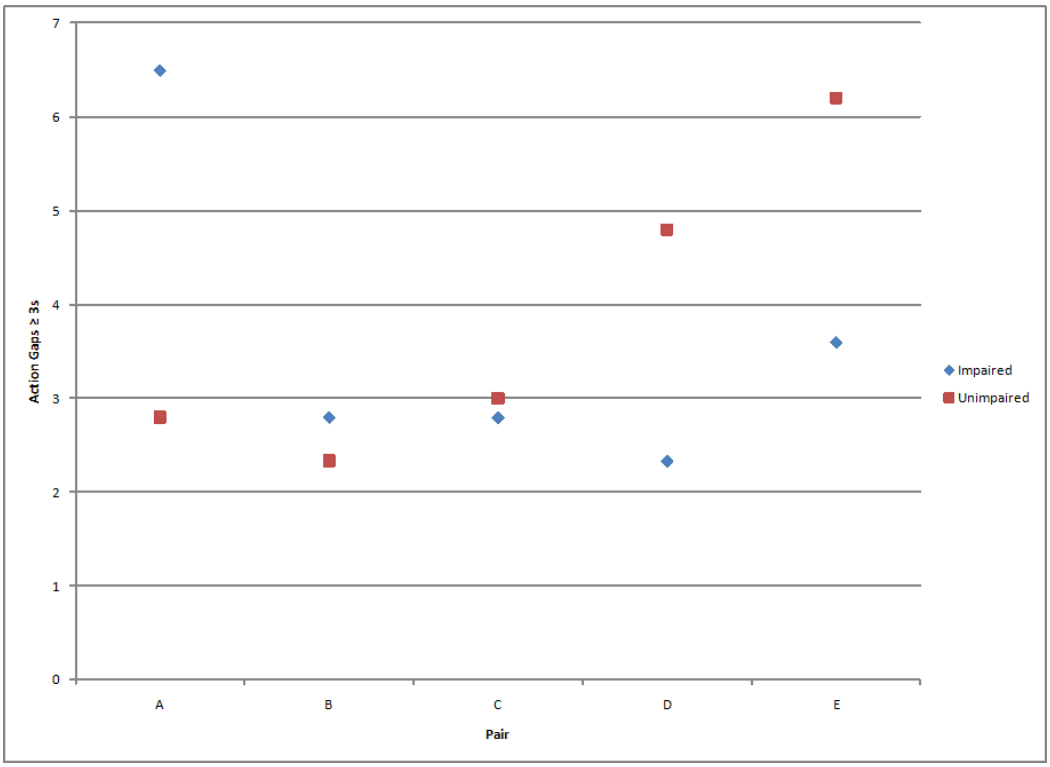

Figure 4.3: Plot of Action Gaps $\geq$ 3 Seconds for Impaired and Unimpaired Subjects

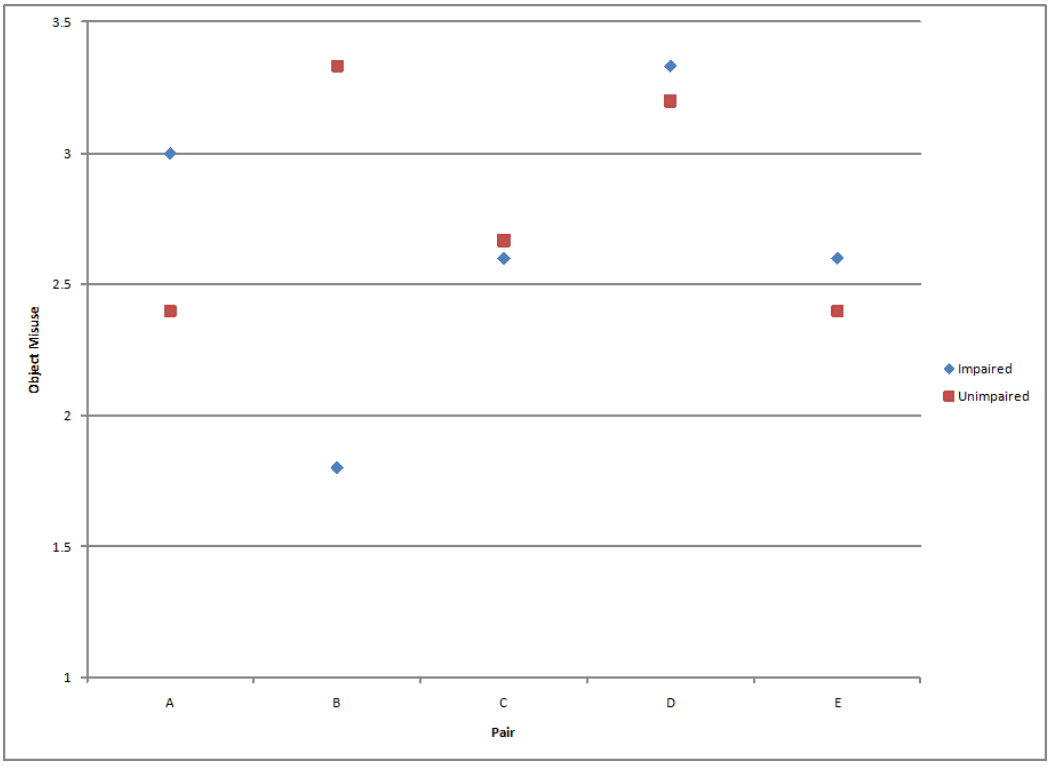

Figure 4.4: Plot of Object Misuse for Impaired and Unimpaired Subjects

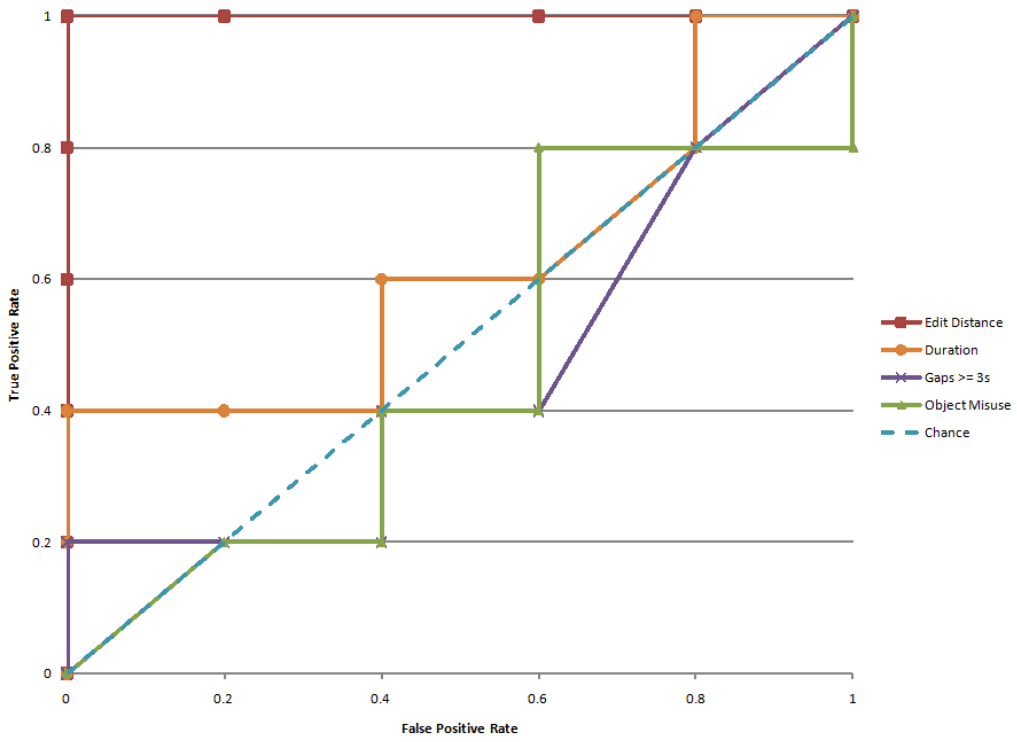

Figure 4.5: ROC Curve with All Four Features

(1 - specificity). By chance (if the feature has no impact on whether a subject is impaired or not), these two values should be the same, creating the 45° line shown for reference in Figure 4.5. A feature performing better than chance will have a curve above that line, demonstrating that true positives are more common than false positives. Since Edit Distance is able to perfectly differentiate the subjects, it has an ideal ROC curve along the left and top sides of the graph.

As seen in Figure 4.5, Trial Duration, the second best feature from the results of Chapter III, has a curve that is always at or above the line for chance, although the difference is not very large. It is able to identify 2 of the 5 impaired subjects without misidentifying any of the unimpaired subjects or correctly identify 3 of the 5 subjects in both data sets. The last two features, Action Gaps greater than 3 seconds and Object Misuse do not perform well, being mainly at or below the line of chance.

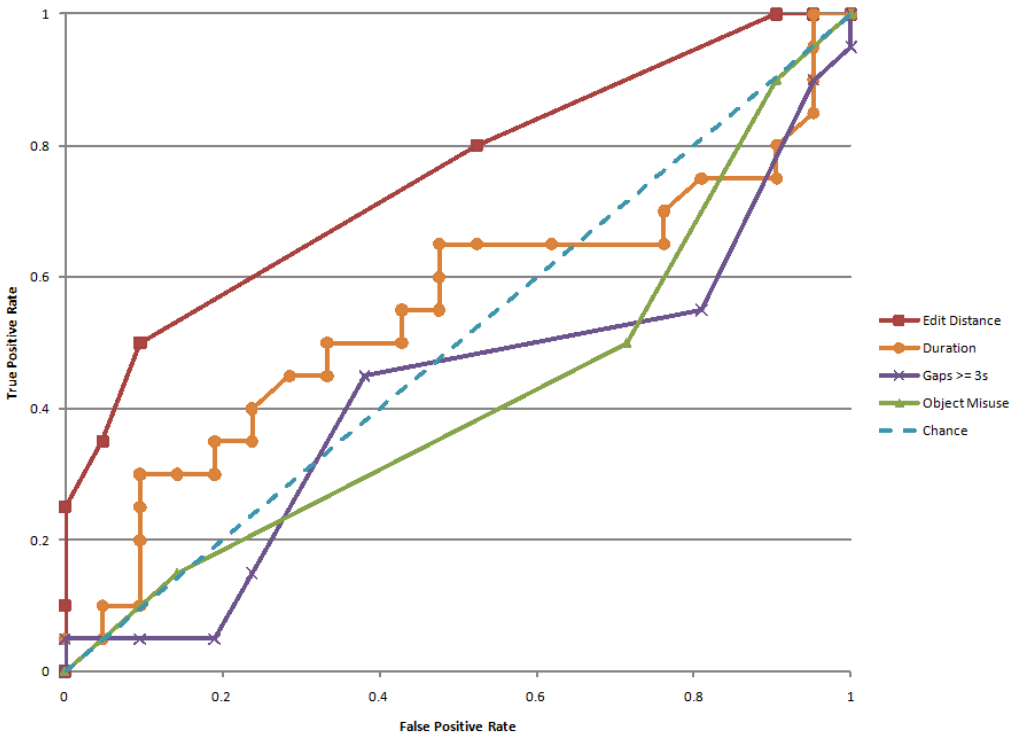

Figure 4.6: ROC Curve using Individual Trials

Again, remember that these results are preliminary due to the small sample size.

4.2.3 Analysis of Individual Trials

Finally, I study the influence of variation across individual trials, which is important to the question of whether it is important to use an average of multiple trials or if a single trial can provide similar accuracy. A related question is whether the higher average values for impaired individuals are the result of a small number of outliers for each subject or if the values are consistently higher in each trial.

Figure 4.6 shows the ROC curve from applying the four features to the individual trials. Not surprisingly, the success of Edit Distance decreases dramatically from the perfect accuracy achieved before, but it is still the most successful feature. The other features have similar ROC curves as before: Trial Duration is mostly above the line of chance while Action Gaps and Object Misuse are generally below that line.

Due to the small sample size, it's difficult to make meaningful comparisons between the performance using a single trial and that of using an average of multiple trials. Nonetheless, the fact that Edit Distance and Trial Duration both perform reasonably well using just one trial is further support for the earlier results suggesting that those two features can be successfully used to recognize cognitive impairments.

With Edit Distance still fairly successful using only a single trial, it's clear that the impaired subjects' trials had higher Edit Distance scores than the trials from the unimpaired subjects with some consistency. The outliers in Figure 4.7 are also notable, however: 4 of the 5 impaired individuals had at least one trial with an Edit Distance of 8 or higher, while no trial from an unimpaired subject had an Edit Distance over 6. These outliers suggest that there may be an important advantage to computing an average of multiple trials since impaired individuals are able to perform trials with low Edit Distances at least some of the time but none of the impaired subjects in this study did so with the same consistency as the unimpaired subjects.

There is an additional disadvantage to using just one trial that is noticeable in Figure 4.6: the small number of feasible thresholds for the Edit Distance feature. Edit Distance values are integers and are typically small, even for most of the trials from impaired subjects, meaning there is not a great deal of variety: over 80% of trials by unimpaired subjects have Edit Distances of 3 or 4, along with nearly 50% of trials by impaired subjects (the features Action Gaps and Object Misuse have the same problem, thought it's less relevant due to their lower success rates). While the available thresholds have fairly good success rates (two likely choices have a 50% true positive rate with just 10% false positive rate or an 80% true positive rate with a 50% false positive rate), there is no way to have rates between those two points

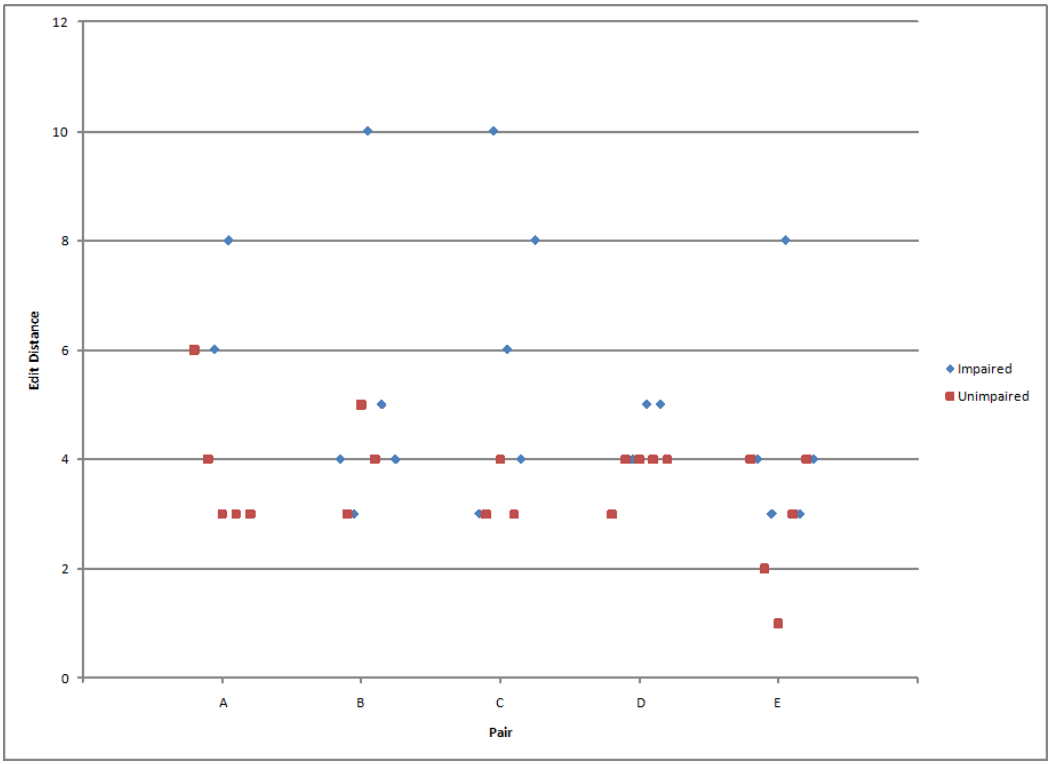

Figure 4.7: Plot of Edit Distance for Individual Trials

with just a single trial.

4.3 Conclusion and Discussion

I have presented the results of a preliminary study using RFID-based observation of individuals' object usage in order to determine whether each individual has a cognitive impairment or not. I applied the four features presented in Chapter III and evaluated them on five pairs of impaired and unimpaired subjects, with each member of a pair having a similar age and level of education as the other.

Although very preliminary, these results are promising: one feature (Edit Distance) is able to successfully differentiate the five impaired subjects from the five unimpaired subjects using a simple threshold value. Results from the other three features were presented as well. These results are consistent with the results from

Chapter III: Edit Distance is definitively the best feature, Trial Duration is the next best, and the Action Gaps and Object Misuse features do not perform as well. The results of applying the features to a single trial were also presented, with similar accuracy except for the Edit Distance feature which, while still the best feature, was well below the perfect success rate it had using the average of the subjects' trials. The single trial results also discussed the problem of having few reasonable choices for thresholds in Edit Distance, since a small number of values are the result of a large number of trials for both unimpaired and impaired subjects.

A first step in future research is to collect data from more subjects, in order to validate the results. Although Edit Distance and Trial Duration are clearly the best two features in these results, Action Gaps and Object Misuse did have more suggestive correlations than would be expected by chance in Chapter III, so they may still prove valuable in larger studies of recognizing cognitive impairments. Additionally, it's very unlikely that Edit Distance would continue to perfectly differentiate impaired subjects from unimpaired in larger data sets, so it may be interesting to consider multiple features together: for example, by classifying an individual as impaired if her scores are above thresholds for a majority of the features.

The practical concerns raised in Section 3.5 are also concerns for a system differentiating impaired and unimpaired individuals: an individual may not comply with the system if she does not want to know (or to have others know) if she develops a cognitive impairment. Again, using other types of sensors may help to increase compliance but likely also have privacy drawbacks to consider.

Other important areas for future work include observation of other forms of impairment, particularly dementia, since that is where the highest demand for this type of system may be. Additionally, longitudinal studies of individuals at risk for impair-

ment are important to understand how well a system could detect impairment at its onset, and particularly how quickly after onset that determination could be made, as well as to potentially develop new techniques that use an unimpaired individual's past behavior to assist in that determination.

Finally, as in the previous chapter, this research could be usefully expanded by applying it to other activities beyond coffee making and using them in a home environment.

CHAPTER V

Conclusion

This dissertation explored a novel area for computer science research: the automatic assessment of cognitive impairments from object-use interactions. This research was undertaken as a result of practical, real-world concerns stemming from the increased number of cases of cognitive impairment likely to result from an aging population and the difficulty in assessing those impairments.

The main contribution of this thesis is the demonstration that inexpensive, unobtrusive, and privacy-preserving technologies can be used to assess the level of cognitive impairment in patients performing routine activities of daily living. Additionally, I developed a set of features that may be used to perform this assessment and studied their effectiveness at both recognizing whether an individual has a cognitive impairment and assessing the impairment of an individual known to have one. The most knowledge-rich of these features, Edit Distance, correlates significantly with the meaningful first principal component of the neuropsychological evaluations for individuals with traumatic brain injury and, in a preliminary study, was able to differentiate impaired individuals from unimpaired individuals using a simple threshold.

In this chapter, I first review the research hypothesis and three key research ques-

tions presented in Chapter I, summarizing the answers to them that were obtained in my research and explaining the research contributions of this dissertation. I then discuss directions for future work, limitations and potential for the research presented in the dissertation, and finally important lessons I learned that may be valuable to others conducting similar research.

5.1 Key Research Results

The central hypothesis of this dissertation is that important inferences about an individual can made by using sensor technology to observe the individual's object-use patterns during the performance of a simple task, significantly including the assessment of an individual's cognitive impairment. This hypothesis was divided into three key research questions:

1. Can an individual be recognized from her object-use patterns in the performance of a task?

This first question, addressed in Chapter II, introduces a novel application of sensor-derived object-usage patterns: recognizing an individual. That chapter presents preliminary evidence that the answer to this question is "yes": the subjects in the study were correctly recognized in more than three-quarters of the trials.

This application was chosen as an important starting point for the larger hypothesis in order to learn about individuals' patterns of object usage and whether those can be detected by passive observation and analyzed automatically. I presented an experiment to study the ability to observe individuals and extract relevant features that can be used for recognition. Five features were introduced, ranging from a very simple metric of whether an individual interacted with an object during the course of

the trial to a more complex measure of the order in which the individual interacted with objects. I also introduced the concept of levels of granularity: the ability to consider the interactions at the object level, or at more detailed tag or group levels.

Finally, I presented the results of this study, including analysis of the different features and levels of granularity. These results showed that, using just nine trials as training data, ten individuals could be correctly recognized up to 77% of the time based on their object-use patterns.

This research laid a foundation for the research in the two chapters following it and for future research in object-usage patterns, showing that these patterns can be observed and analyzed automatically. The research also had direct significance, showing that privacy must be an important consideration in the design of systems that collect and use object-usage data. Finally, the research provided evidence that biometric security based on object usage may be feasible, potentially allowing smart homes or other monitored spaces to provide an additional level of security based on the tasks performed by individuals in the space.

2. Can the severity of an individual's impairment be assessed by electronically observing the individual's object-use patterns from the performance of a task?

This is, of course, a complex question, but the results in my research point to an affirmative answer.

In Chapter III, I addressed a second novel application of object-usage patterns: the automatic assessment of cognitive impairments. To study this application, I recruited 16 individuals with traumatic brain injuries for whom a full set of neuropsychological evaluations had been performed and I collected sensor data as they each performed a basic task, preparing a pot of coffee. I then introduced four types

of features which, as measures of how well the subject performed the task, I hypothesized would correlate with the severity of a cognitive impairment. I compared these features to the subjects' scores on the neuropsychological evaluations.

The most important result in this chapter is that Edit Distance, the most knowledge-rich feature I computed, had a statistically significant correlation with the first principal component of a standard suite of neuropsychological tests; moreover, this component was judged by a domain expert to be a meaningful measure of general neuropsychological integrity. Additionally, although the results are preliminary, there were many more suggestive correlations between the computed features and neuropsychological evaluations than would be expected by chance, showing that they can potentially provide important information to caregivers and medical professionals about the individual's cognitive impairment. Forty percent of the standard neuropsychological assessments had a suggestive correlation with one of the features, evidence that the computed features can be used to evaluate a broad range of neuropsychological abilities. This result is notable because cognitive impairments are measured along many dimensions (e.g. memory, attention and executive reasoning) and it is important for a rehabilitation team to understand different aspects of a patient's impairment.

As strong evidence that the automatic assessment of cognitive impairments is possible based on individuals' object usage patterns, the results presented here provide a basis for further development of this novel application: the success of the features, particularly Edit Distance, is a motivation for larger studies with subjects in lab environments, as well as for studies in subjects' homes. All four features also provide a basis for comparison with features that are developed in the future. The suggestive correlations presented can also direct future research to study these types

of features without the need for a massive amount of data collection. An important additional result from this study is that, since the vast majority of trials resulted in coffee being made correctly, the success of these features appears to be due to their ability to measure efficiency in the performance of a task, rather than simply whether or not the task was performed correctly.

3. Is it possible to recognize whether or not an individual has a cognitive impairment based on sensor-derived object-use patterns?

Although the results in Chapter IV are preliminary, they point to an affirmative answer for this question as well. Studying five pairs of impaired and unimpaired subjects matched by age and education, I was able to perfectly separate the impaired subjects from the unimpaired subjects using a simple threshold value of the Edit Distance feature.

I also examined the performance of the three other types of features using an ROC curve, which showed Duration generally performing better than chance and the other two features slightly worse than chance.

Although preliminary, this research suggests that the same features that were applied to assessing how severe an impairment is can also be applied to recognizing whether an individual has an impairment or not. Because these results are consistent with the results from Chapter III, it would appear that techniques that are useful for one part of the assessment (recognition of the impairment or assessment of its severity) are also useful for the other. It further bolsters the evidence that Edit Distance and Duration are useful features for assessing cognitive impairments using object-usage observation and that Edit Distance is a particularly strong feature for this purpose.

5.2 Future Work

5.2.1 Recognition of Individuals

There are several avenues for future work to specifically address the limitations of the study in Chapter II. These include developing new techniques to increase accuracy, conducting additional studies in more realistic conditions, generalizing the results to other tasks, and investigating the implications of these results.

There are three key ways that the accuracy in this study could potentially be improved: by developing new features that can be extracted from the sensor data, by using different sensing technologies to observe the subjects, and by applying different learning techniques.

Further research should address the ways in which the realism was limited in the study. An important step is to replicate the study in subjects' homes, where subjects have more freedom in how they perform a task and where it may be difficult to tag all relevant objects. Another way to address the limitations of the study is to involve other, distracting, activities in the study: so that the starting and ending points of the observed task must be determined automatically, and allowing the possibility of interleaved or interrupting tasks. Additionally, because the subjects from the study in Chapter II were mostly graduate students in engineering, further work should include subjects across a broader range of ages and backgrounds.

Studying the recognition of individuals using other tasks is important in order to understand the generalizability of these results, and it is also interesting to study whether similar results can be obtained using a broad range of behavior (e.g. behavior in an office, or during a day spent at home) rather than a specific task that may or may not be performed when a subject's identification is desired.

Finally, further analysis of the privacy implications of object-usage data includes

studying various scenarios—for example, how much data is needed for a reasonably accurate system when there are large numbers of subjects, or how accurate a system can be when less training data is available.

5.2.2 Recognition and Assessment of Cognitive Impairments

Future work in the recognition and assessment of cognitive impairments includes several similar directions as the future work in the recognition of individuals, as well as several new directions including confirming preliminary results from Chapters III and IV, studying the generalizability of the research to other kinds of cognitive impairment, and examining the acceptance of this kind of technology by those at risk for impairments.

There are several areas of future work in the recognition and assessment of cognitive impairments that are similar to those in the recognition of individuals: studying subjects in home environments, studying other tasks, and using other technology to observe the individuals or creating new computed features from the collected sensor data.

Studying subjects in home environments is probably the most important of these avenues for future work. Although the assessment could be used in controlled environments at rehabilitation centers, this form of assessment is particularly interesting because of the potential for regular, unobtrusive observation in the home. There are several additional challenges in a home environment including detecting when the observed activity is being performed, and determining when it started and when it stopped so that the features are only computed from relevant sensor data. Additionally, in a home environment, subjects may be interrupted during the performance of a task, or may interleave the performance of multiple tasks, and computed features

would need to take these complications into account in order to provide a meaningful assessment of the individual.

Also very important is generalizing the results by studying other tasks, or potentially broader categories of behavior rather than specific tasks. Obviously, many individuals do not drink coffee or make it frequently, so additional tasks would allow more subjects to be assessed in an unobtrusive way, but there are important advantages to being able to observe a range of tasks, even for people who would make coffee regularly. First, since performing different tasks require different sets of skills, some tasks may be better suited for measuring a particular type of impairment than others; this would also help assess different types of impairment more precisely since coffee-making had suggestive correlations with a large number of neuropsychological assessments. Second is that this would allow the use of a range of "graded" tasks of varying difficulty. A rehabilitation team typically uses graded tasks to observe impairment since some tasks may be too difficult for a more severely impaired individual or too easy for a less severely impaired individual.

Further research into using other sensor technologies and developing new features from the collected data may be able to improve the accuracy and reliability of the assessments; it may also facilitate the observation of tasks or behavior for which RFID sensors or the computed features presented are not ideal. To observe meal preparation, for example, Edit Distance may be infeasible due to the huge number of different foods that could be made, but other measures of efficiency may still be valuable, such as the number of cabinets that were opened in order to find a desired item (recall from section 3.5 that due to the vast majority of trials being performed correctly, the computed features appear to be successful because they are measuring efficiency in various ways).

After studying subjects in a home environment, I expect that observing a set of tasks would be the most likely of these to substantially improve an assessment system, by adding redundancy to the evaluation and by including tasks that require other kinds of abilities to complete. Although their impact is more difficult to predict, the creation of new features could also be a very positive improvement, especially if a new feature can build on the research here and be designed more explicitly to measure efficiency, or if a successful new feature was found that measured a very different attribute of task performance and, therefore, provided a different type of insight into an individual's impairment. Although there was room for improvement, the RFID sensors used in this dissertation generally seemed to perform well; a change in sensing technology would be particularly helpful only if a new task wasn't well-suited for observation with RFID or if needing to wear an RFID bracelet caused too much intrusion or low compliance.

Several suggestive correlations were identified between neuropsychological evaluations and the automatic features presented in Chapter III; further study is important to learn which of those relationships are statistically significant. Similarly, the results presented in Chapter IV are preliminary due to the small sample size and a study with a larger subject population is needed to confirm these results.

Further study is necessary to understand the generalizability of these results to other types of cognitive impairments, particularly dementia, including Alzheimer's disease. Of particular interest is studying how quickly the features presented in this dissertation or other computed features could detect the onset of impairment in older individuals at risk for those impairments, so longitudinal studies are important for future work in this area. The current version of the Intel iBracelet is likely to be considered too large and obtrusive and have too short a battery life to be ideally used

in longitudinal studies, so other technology for object usage sensing may be better to encourage higher participation, particularly over time. Longitudinal studies will also face the challenges of in-home environments, addressed previously in this section.

Finally, further research is needed to examine how well this kind of technology would be accepted by individuals at risk for impairments, particularly senior citizens. While privacy implications of sensors in the home have already been explored, including among senior citizens, subjects may be more hesitant to install a system that could identify an impairment. In this dissertation, a decision was made to use technology that subjects can clearly prevent from observing their behavior (by removing the iBracelet RFID reader); this further research could also address questions of whether this kind of concession is successful in alleviating many privacy concerns and, alternatively, what the impact is on the amount of data collected from subjects and whether it is still sufficient to assess impairments well.

Occupational Therapy Assessments

Neuropsychological assessments, which were the focus of the study in Chapter III, are only a portion of the assessments performed for many cognitively impaired patients. Occupational therapy (OT) assessments are of particular interest for future work because they play an important role in making recommendations as to whether an individual should be living independently, driving, or working; determining whether an impairment prevents an individual from doing any of those tasks could be more important than just determining if an impairment exists. Additionally, automatic assessment based on task performance seems well-suited since an important part of occupational therapists' assessments are their own observation of patients' task performance.

Motor-Free Visual Perception Test-3 and MVPT-Revised
Test of Visual-Perceptual Skills (Non-Motor) (Upper Level and Lower Level)
Hooper Visual Organization Test
O'Connor Wiggly Block Test
Contextual Memory Test

Table 5.1: Occupational Therapy Assessments

Using the features described in Chapter III, I have performed some early exploratory analysis with several standard occupational therapy assessments, listed in Table 5.1. These assess a wide range of skills including visual memory, perceptual and spacial reasoning, and the perception of visual figure ground, visual closure, and of spacial relationships. From this preliminary analysis, there appear to be some interesting trends, including multiple potential suggestive correlations. This analysis is not developed to a point that it is ready to be reported, however, but remains a promising area for future work.

5.3 Limitations and Potential of this Research

In this section, I address two important limitations to the research presented in this dissertation that are not identified elsewhere, as well as two related potential strengths of this type of research.

As with much research involving human subjects, one limitation of this work is the potential for subject bias, that subjects may attempt to provide "positive" results in order to aid the researcher [36]. For example, in Chapter II, studying the recognition of individuals, subjects may have attempted to have more consistent task performance than normal in order to assist recognition. Beyond an explanation of the sensors necessary for informed consent, I declined to show the raw data collected or discuss features being studied until a subject had completed participation in the study, to reduce the ability of a subject to know what behavior might directly affect

their results, but the risk still exists since subjects understood that their task performance was being measured. This bias is difficult to explicitly avoid or measure, so it is important to acknowledge the potential for it. This limitation directly relates to a potential strength of this type of research, however, since the potential for this effect exists in nearly all kinds of human subjects research, including neuropsychological research. By using unobtrusive sensors in longitudinal studies, this form of assessment has the potential to observe subjects without researchers present and over a period of time such that this effect is likely to disappear.

A second limitation of this approach to the assessment of cognitive impairments is the need to develop appropriate models for the computed features for each task that is being observed. This is clear for the Edit Distance and Object Misuse features, but even for the simpler features Trial Duration and Action Gaps, the typical range for unimpaired individuals must be established for a task before recognizing or assessing an impairment is possible. Previous work in activity recognition, discussed in Section 1.3.2, has studied automatically creating models for activity performance from web sites or common sense databases [65, 45]; this work could potentially be adapted to create models for Edit Distance and Object Misuse and potentially even to understand appropriate values for all four types of computed features. In contrast to the case of activity recognition, an ideal assessment system would not need to work for every activity performed, but using a large set of tasks to form the assessment has several advantages, discussed in section 5.2.2.

The second strength is the potential to improve understanding of the ecological validity of neuropsychological assessments. Neuropsychological assessments have typically been developed to diagnose a patient, in order to understand whether a patient has one or more brain lesions and, if so, where the lesions are. More recently,

however, there has been increased interest in using those neuropsychological assessments to understand how capable the impaired individual is of working, driving, and living independently [9, 30]. This interest has led to investigating the ecological validity of neuropsychological assessments, the degree to which results on an assessment relate to real-world performance. One reason ecological validity is challenging is that, although actual ability should increase with rehabilitation and as the patient develops compensatory strategies, diagnostic tests are designed to remain stable measures of the actual brain damage itself. As a result, diagnostic assessments are evaluated in a quiet, supportive environment over short periods of time, designed to measure the best that a patient can do, in order to avoid falsely diagnosing a brain lesion. Ecological concerns, however, focus on how the patient does in real-world environments that are often distracting and unsupportive and this change in environment may impact individuals differently [9]. By studying the relationship between real-life task performance and neuropsychological assessments, this type of research has the potential to improve the understanding of neuropsychological assessments and potentially to assist in the development of assessments with higher ecological validity.

5.4 Lessons Learned

There are several valuable lessons I've learned while doing this research—especially from the mistakes I've made along the way. Such mistakes are all too easy to make when doing research involving a field well outside computer science. I share them to help others pursuing similar lines of inquiry since, while many seem obvious in hindsight, they were not so at the time.

The most important lesson is the value of a domain expert at all stages of the research, especially throughout the experimental design and with the interpretation

of results. In several cases, there were significant concerns with the validity of the research designs that I created which would have led to a great deal of lost effort if not for the involvement of a neuropsychologist, Dr. Ned Kirsch, in the design process. Additionally, discussing the best approach to data analysis and interpreting the results would have been impossible without the perspective of a domain expert.

A second lesson is the difficulty of performing research with human subjects. An Institutional Review Board (IRB) oversees any research with human subjects and, while this protection is very important, beginning the research is slowed dramatically by needing to go through the approval process. Anecdotally, this process takes even longer when the study includes protected subjects (such as individuals with cognitive impairments) and when a medical school IRB is involved (as it was in the studies for Chapters III and IV since the subjects were patients at a medical facility at the university). Writing a detailed research protocol and discussing it amongst the research team members and with individuals familiar with IRB-approved studies makes this process go much more smoothly and helps to ensure that the application is internally consistent. The protocol should cover the recruitment of subjects as well as the experimental design, with particular concern for any safety precautions that will be taken. Safety includes concerns such as how a study will be stopped if there appears to be any danger to the subjects (in this study, I was particularly alert for the possibility that a subject might be burned by hot coffee even though the trial should end before any is produced—I would have intervened if there appeared to be a situation where there was hot coffee or hot parts of the coffee maker that the subject might touch). Other considerations in the experimental design were how much instruction to give before the start of the task (both on how to complete the task, and if the subjects need to do anything special to use the sensors—I always

instructed them to ignore the sensors as best they could), how to answer questions about the task during a trial, and how to respond to conversation from the subject during the task. Though several of those details were not included in the design given to the IRB, they were important to consider ahead of time in order to make the best decision and to be consistent with all subjects. Since we used a worn sensor as part of our study, that device needed to be separately approved by the IRB, including a subcommittee specifically for biomedical engineering devices.

Subject recruitment and actual data collection can also be very slow. Without the active support of Dr. Kirsch and the other staff at the University of Michigan's MedRehab clinic, a medical clinic for individuals with cognitive impairments, the recruitment of impaired individuals would have been almost impossible. The data collection was done before or after existing appointments at MedRehab to avoid imposing on the subjects, but even though subjects were generally very willing to participate, there were a large number of forgotten appointments. Since data collection is so slow, a pilot study is highly recommended to find problems with the experimental setup before real data collection begins. This helped to address reliability concerns with the sensors, make sure the information and prompts I was giving to the subjects were appropriate, and prompted me to think about how to respond to questions from the subjects. Once data collection for the full study has started, it's very difficult to modify any part of the study since even moving a sensor that wasn't often being picked up would have made it very difficult to compare data collected before the change to data collected after it. Since a pilot study does not need to be with impaired subjects, it may be possible to quickly submit a more broad application for it to the IRB (or even one that will cover several pilot studies); by the time the details have been figured out for the full study, the IRB application for

the pilot study will hopefully have been approved.

A further complication with subject recruitment was the need to match subjects for the study presented in Chapter IV. Dr. Kirsch identified two factors that might particularly impact how an individual performs the task of preparing a pot of coffee: age and intelligence (we used the number of years of education as a proxy for intelligence since intelligence itself is difficult to measure). For example, a younger individual would be likely to prepare coffee more quickly than an older individual even if they were both unimpaired. The original set of unimpaired subjects I collected data from, university students who were mostly undergraduates, were generally much younger and better educated than the impaired subjects that I had collected data from. As a result, I was unable to say whether the results showing differences between unimpaired and impaired subjects were due to the difference in impairment or simply the difference in demographics. In fact, neuropsychological assessments are generally studied very carefully for affects of age (and potentially gender) so that an individual's assessment is made relative to what is normal for her age group. It was only by recruiting a second set of unimpaired subjects and then matching the subjects (choosing pairs of subjects who had similar ages and levels of education) that I was able to draw meaningful conclusions from the different results. It's worth adding that I did not actually compare the subjects in each pair to one another (though that may have had some value with a larger sample size), but that simply choosing pairs of subjects helped protect my results from any effects of age and education. Recruiting unimpaired subjects was actually harder than recruiting the impaired subjects, mainly because there was a clear pool of impaired subjects with whom the MedRehab staff already had an established relationship, and partly because I only wanted unimpaired individuals who had demographics that were similar

to an impaired subject.

Careful consideration of what data should be collected is very important. In this study, the sensor data was obviously central to the research, as were some medical records about each patient's impairment, but I wish I had collected more information about the subjects themselves and it's often very difficult or impossible to get information about subjects after their participation in the study has ended. In particular, I would have liked to have recorded how often each subject made coffee previously and which hand the subject's dominant hand was, since those could impact how well a subject performed and potentially the performance of the sensors, respectively. I had also not originally planned to collect information about each subject's age and level of education which ended up being very important to the study in Chapter IV—without it, appropriate analysis would have been impossible. Since gender plays a role in some neuropsychological assessments, it also would have been good to record. This is not to suggest that huge amounts of demographic information should be blindly collected, but that the kinds of information that might be relevant should be carefully considered during the research design phase.

Lastly, I would like to discuss task selection. Almost certainly, making coffee is not the only task for which this form of assessment could have succeeded. Nonetheless, it has two attributes that I believe helped to make it a successful choice: it is a good fit for the sensing technology and a good fit for testing the hypothesis.

Making coffee is easy to observe using RFID sensors. This is not true of all tasks because some types of objects are more difficult to observe than others: RFID tags on metallic objects often cannot be detected due to interference and I had trouble with tags on smaller objects like spoons because the tag needed to be bent more tightly than the circuitry allowed. I also wanted to avoid putting RFID tags in the oven or

microwave, or having them submerged in water in order to keep them functioning properly. Tags can't go directly onto all objects of interest—a piece of bread, for example, can't have a tag on it, though putting tags on the bag of bread may be enough to allow reasoning about the bread.

The task is a good fit for testing the hypothesis as well. There are only a few different ways to most directly complete the task and most individuals seem to try to perform the task in a direct way (inefficiencies that I observed generally appeared to have been mistakes). It is also a task for which it is fairly easy to make small mistakes—the fact that unimpaired individuals often unintentionally performed it inefficiently gave me hope that impaired individuals would also make small mistakes (and hopefully more of them!). A task that was simple enough that mistakes were rarely made by unimpaired individuals might also rarely involve mistakes from impaired individuals. Finally, these mistakes were typically related to object interaction: for example, two common mistakes are closing the lid too soon and starting to get the ground coffee before putting the filter in, and both of those mistakes are related to object usage that the sensors could pick up, and relatively simple logic could determine from the sensor data if those mistakes had occurred.

Two tasks that were considered were making a sandwich and doing a load of laundry; both were eventually discarded because they did not fit either criterion well. They both included items that were hard to tag—bread and utensils for the sandwich and clothing that will be going into water and a metallic washer for doing laundry. I also envision that it would be harder to detect mistakes with both of those tasks by observing object usage because I cannot easily write out a series of object interactions that clearly involves a mistake or inefficiency and is also likely to occur during the performance of one of those tasks.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] F. Albinali, M. Goodwin, and S. Intille. Recognizing stereotypical motor movements in the laboratory and classroom: A case study with children on the autism spectrum. *Ubicomp*, 2009.

[2] R. N. Anderson. Method for constructing complete annual U.S. life tables. *National Center for Health Statistics. Vital and Health Stat*, 2, 1999.

[3] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 581–588, 2005.

[4] L. J. Bagley, J. C. McGowan, R. I. Grossman, G. Sinson, M. Kotapka, F. J. Lexa, J. A. Berlin, and T. K. McIntosh. Magnetization transfer imaging of traumatic brain injury. *Journal of Magnetic Resonance Imaging*, 11:1–8, January 2000.

[5] M. Barbaro and J. Tom Zeller. A face is exposed for AOL searcher No. 4417749. *The New York Times*, August 2006.

[6] T. S. Barger, D. E. Brown, and M. Alwan. Health-status monitoring through analysis of behavioral patterns. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 35:22–27, January 2005.

[7] N. Boulgouris and Z. Chi. Gait recognition using radon transform and linear discriminant analysis. *IEEE Transactions on Image Processing*, 16(3):731–740, March 2007.

[8] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall. Recognizing daily activities with rfid-based sensors. *Ubicomp*, 2009.

[9] N. Chaytor and M. Schmitter-Edgecombe. The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, 13:181—197, December 2003.

[10] T. Choudhury, S. Consolvo, B. Harrison, J. Hightower, A. LaMarca, L. LeGrand, A. Rahimi, A. Rea, G. Borriello, B. Hemingway, P. P. Klasnja, K. Koscher, J. A. Landay, J. Lester, D. Wyatt, and D. Haehnel. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, 7:32–41, 2008.

[11] N. L. Clarke and S. M. Furnell. Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security*, 6(1):1–14, 2006.

[12] W. E. Eltahir, M. Salami, A. F. Ismail, and W. Lai. Dynamic keystroke analysis using AR model. *IEEE International Conference on Industrial Technology*, 3:1555–1560, 2004.

[13] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. R. Menezes, E. Rimmer, and M. Scazufca. Global prevalence of dementia: a delphi consensus study. *The Lancet*, 366:2112–2117, 2006.

[14] K. P. Fishkin, B. Jiang, M. Philipose, and S. Roy. I sense a disturbance in the force: Unobtrusive detection of interactions with RFID-tagged objects. *Ubicomp*, pages 268–282, 2004.

[15] K. P. Fishkin, M. Philipose, and A. Rea. Hands-on RFID: Wireless wearables for detecting use of objects. *IEEE International Symposium on Wearable Computers*, pages 38–43, 2005.

[16] A. P. Glascock and D. M. Kutzik. The impact of behavioral monitoring technology on the provision of health care in the home. *Journal of Universal Computer Science*, 12:59–79, 2006.

[17] I. Grant and K. M. Adams. *Neuropsychological Assessment of Neuropsychiatric Disorders*. Oxford University Press, Oxford, 2008.

[18] T. L. Hayes, F. Abendroth, A. Adami, M. Pavel, T. A. Zitzelberger, and J. A. Kaye. Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimer's & Dementia*, November.

[19] S. Hocquet, J.-Y. Ramel, and H. Cardot. Fusion of methods for keystroke dynamic authentication. *IEEE Workshop on Automatic Identification Advanced Technologies*, pages 224–229, October 2005.

[20] J. Hoey, A. von Bertoldi, P. Poupart, and A. Mihailidis. Assisting persons with dementia during handwashing using a partially observable markov decision process. *International Conference on Computer Vision Systems*, 2007.

[21] T. Holsinger, J. Deveau, M. Boustani, and J. John W. Williams. Does this patient have dementia? *JAMA*, 297:2391–2404, June 2007.

[22] S. G. Jianguo Zhang. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43:197–203, 2010.

[23] H. Jimison and M. Pavel. Embedded assessment algorithms within home-based cognitive computer game exercises for elders. *International Conference of the*

IEEE Engineering in Medicine and Biology Society, pages 6101–6104, September 2006.

[24] H. B. Jimison, M. Pavel, and T. Le. Home-based cognitive monitoring using embedded measures of verbal fluency in a computer word game. *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008.

[25] H. B. Jimison, M. Pavel, and J. McKanna. Unobtrusive computer monitoring of sensory-motor function. *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5431–5434, September 2005.

[26] H. B. Jimison, M. Pavel, J. McKanna, and J. Pavel. Unobtrusive monitoring of computer interactions to detect cognitive status in elders. *IEEE Transactions on Information Technology in Biomedicine*, 8(3):248–252, 2004.

[27] H. B. Jimison, M. Pavel, K. Wild, P. Bissell, J. McKanna, D. Blaker, and D. Williams. A neural informatics approach to cognitive assessment and monitoring. *IEEE EMBS Conference on Neural Engineering*, pages 696–699, May 2007.

[28] J. Kaye. Home-based technologies: A new paradigm for conducting dementia prevention trials. *Alzheimer's & Dementia*, 4, 2008.

[29] K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439, 1992.

[30] J. Lengenfelder, M. T. Schultheis, T. Al-Shihabi, R. Mourant, and J. DeLuca. Divided attention and driving: A pilot study using virtual reality technology. *Journal of Head Trauma Rehabilitation*, 17:26—37, February 2002.

[31] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A hybrid discriminative/generative approach for modeling human activities. *Proceedings of International Joint Conference on Artificial Intelligence*, July 2005.

[32] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, February 1966.

[33] M. D. Lezak. *Neuropsychological Assessment, Fourth Edition*. Oxford University Press, Oxford, 2004.

[34] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 787–794. MIT Press, Cambridge, MA, 2006.

[35] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille. A long-term evaluation of sensing modalities for activity recognition. *Ubicomp*, September 2007.

[36] M. L. Mitchell and J. M. Jolley. *Research Design Explained*. Wadsworth Publishing, 2009.

[37] M. Ogawa, R. Suzuki, S. Otake, T. Izutsu, T. Iwaya, and T. Togawa. Long term remote behavioral monitoring of elderly by using sensors installed in ordinary houses. *IEEE EMBS Special Topic Conference on Microtechnologies in Medicine & Biology*, pages 322–325, 2002.

[38] S. Okie. Traumatic brain injury in the war zone. *New England Journal of Medicine*, 352:2043–2047, May 2005.

[39] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. *IEEE International Symposium on Wearable Computers*, pages 44–51, 2005.

[40] D. J. Patterson, L. Liao, K. Gajos, M. Collier, N. Livic, K. Olson, S. Wang, D. Fox, and H. A. Kautz. Opportunity knocks: A system to provide cognitive assistance with transportation services. *Ubicomp*, pages 433–450, 2004.

[41] M. Pavel, A. Adami, M. Morris, J. Lundell, T. L. Hayes, H. Jimison, and J. A. Kaye. Mobility assessment using event-related responses. *Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare*, pages 71–74, 2006.

[42] A. Peacock, X. Ke, and M. Wilkerson. Typing patterns: A key to user identification. *IEEE Security and Privacy*, 2(5):40–47, 2004.

[43] W. Pentney, M. Philipose, and J. Bilmes. Structure learning on large scale common sense statistical models of human. *AAAI*, July 2008.

[44] W. Pentney, M. Philipose, J. Bilmes, and H. Kautz. Learning large scale common sense models of everyday life. *AAAI*, July 2007.

[45] W. Pentney, A.-M. Popescu, S. Wang, H. A. Kautz, and M. Philipose. Sensor-based understanding of daily life via large-scale use of common sense. *AAAI*, July 2006.

[46] Pew Internet & American Life Project. Video games: Adults are players too. December 2008.

[47] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, 2004.

[48] M. E. Pollack. Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. *AI Magazine*, 26:9–24, 2005.

[49] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, June 2010.

[50] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[51] P. D. Sloane, S. Zimmerman, C. Suchindran, P. Reed, L. Wang, M. Boustani, and S. Sudha. The public health impact of Alzheimer's disease, 2000-2050: Potential implication of treatment advances. *Annual Review of Public Health*, 23:213–231, 2002.

[52] G. E. Smith, R. J. Ivnik, and J. Lucas. Assessment techniques: Tests, test batteries, norms and methodological approaches. In J. M. . J. Ricker, editor, *Textbook of Clinical Neuropsychology*, pages 38–58, New York, 2008. Taylor & Francis.

[53] J. R. Smith, K. P. Fishkin, B. Jiang, A. Mamishev, M. Philipose, A. D. Rea, S. Roy, and K. Sundara-Rajan. RFID-based techniques for human-activity detection. *Communications of the ACM*, 48(9):39–44, 2005.

[54] D. M. Sosin, J. E. Sniezek, and D. J. Thurman. Incidence of mild and moderate brain injury in the united states, 1991. *Brain Injury*, 10:47–54, 1996.

[55] T. Starner, J. Auxier, D. Ashbrook, and M. Gandy. The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. *IEEE International Symposium on Wearable Computers*, 2000.

[56] E. M. Tapia, S. S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. *Pervasive*, pages 158–175, 2004.

[57] E. M. Tapia, S. S. Intille, L. Lopez, and K. Larson. The design of a portable kit of wireless sensors for naturalistic data collection. *Pervasive*, pages 117–134, 2006.

[58] D. J. Thurman, C. Alverson, K. A. Dunn, J. Guerrero, and J. E. Sniezek. Traumatic brain injury in the united states: A public health perspective. *Journal of Head Trauma Rehabilitation*, 14:602–615, 1999.

[59] United States Census Bureau. International data base. July 2007.

[60] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003.

[61] D. R. Weir, J. S. Jackson, and A. Sonnega. Study of retired NFL players. *National Football League Player Care Foundation*, September 2009.

[62] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3):66–75, September 1991.

[63] T. L. Westeyn, J. A. Kientz, T. E. Starner, and G. D. Abowd. Designing toys with automatic play characterization for supporting the assessment of a childs development. *Workshop on "Designing for Children with Special Needs" at the Conference on Interaction Design and Children*, 2008.

[64] D. Wilson, S. Consolvo, K. Fishkin, and M. Philipose. In-home assessment of the activities of daily living of the elderly. *Workshop on "HCI Challenges in Health Assessment" at CHI*, page 2130, April 2005.

[65] D. Wyatt, M. Philipose, and T. Choudhury. Unsupervised activity recognition using automatically mined common sense. *AAAI*, July 2005.

[66] G. Zoroya. Scientists: Brain injuries from war worse than thought. *USA Today*, September 2007.