Summer 2015

Automatic recognition of healthcare worker hand hygiene

Valerie Galluzzi
University of Iowa

AUTOMATIC RECOGNITION OF HEALTHCARE WORKER HAND HYGIENE

by
Valerie Galluzzi

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science in the Graduate
College of the University of Iowa

August 2015

Thesis Supervisor: Professor Ted Herman

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Valerie Galluzzi

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the August 2015 graduation.

Thesis Committee: _____
Ted Herman, Thesis Supervisor

_____
Octav Chipara

_____
Sriram Pemmaraju

_____
Philip Polgreen

_____
Alberto Segre

This thesis is dedicated to the family, friends, and community who have supported me throughout my scholastic career. Some would say that the completion of this thesis marks the end of my education. For me it marks the start of me educating myself. Thank you for giving me the tools I needed to never stop learning.

ACKNOWLEDGEMENTS

First I would like to acknowledge my advisor, Professor Ted Herman, for his support and guidance.

I would also like to thank the other members of my thesis committee, Professor Alberto Segre, Professor Sriram Pemmaraju, Professor Phil Polgreen, and Professor Octav Chipara for their advice and helpful criticism.

I would like to thank the Graduate College for their support through the Dean's Graduate Fellowship.

I would like to thank Gojo Industries for funding this work. In particular I would like to thank Dr. Arbogast and Dr. Macinga for their interest and support.

I would like to thank Dr. Phil Polgreen, Elena Segre, and Rachel Butler for their help with collecting data in the hospital. Without your exacting care and relentless recruitment of participants I would never have got this far.

I would also like to thank Sheryl Semler, our Academic Services Coordinator, and Catherine Till, our Departmental Administrator, for their tireless work.

ABSTRACT

Hand hygiene is an important part of preventing disease transmission in the hospital. Due to this importance, electronic systems have been proposed for automatically monitoring healthcare worker adherence to hand hygiene guidelines. However, these systems can miss certain hand hygiene events and do not include quality metrics such as duration or technique. We propose that hand hygiene duration and technique can be automatically inferred using the motion of the wrist. This work presents a system utilizing wrist-based 3-dimensional accelerometers and orientation sensors, signal processing (including novel features), and machine learning to detect healthcare worker hand hygiene and report quality metrics such as duration and whether the healthcare worker used recommended rubbing technique. We validated the system using several different types of data sets with up to 116 healthcare workers and activities ranging from synthetically generated hand hygiene movements to observation of healthcare worker hand hygiene on the hospital floor. In these experiments our system detects up to 98.4% of hand hygiene events, detects hand hygiene technique with up to 92.1% accuracy, and accurately estimates hand hygiene duration.

PUBLIC ABSTRACT

Hand hygiene is an important part of preventing disease transmission in the hospital. Due to this importance, electronic systems have been proposed for monitoring whether hospital healthcare workers are cleaning their hands when they are supposed to according to the World Health Organization. However, *how* hand hygiene is performed is just as important as *when* hand hygiene is performed. Current systems do not examine how hand hygiene is performed and do not include quality metrics such as duration or technique. We propose that hand hygiene duration and technique can be automatically inferred using the motion of the wrist. In this work we monitor healthcare workers with two sensor-equipped wristbands, one on each wrist. We use machine learning to detect healthcare worker hand hygiene and report quality metrics such as duration and whether the healthcare worker used recommended rubbing technique. We validate the system using several different types of data sets with up to 116 healthcare workers and activities ranging from healthcare workers performing hand hygiene and non-hand hygiene movements on command to observation of healthcare worker hand hygiene on the hospital floor. In these experiments our system detects up to 98.4% of hand hygiene events, detects hand hygiene technique with up to 92.1% accuracy, and accurately estimates hand hygiene duration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

HAND HYGIENE MONITORING SYSTEMS

Hospital healthcare worker hand hygiene is a first line of defense against the transmission of pathogens in the hospital [3]. Due to its importance, best practice guidelines have been created by multiple agencies. These guidelines inform both how and when hand hygiene is practiced by healthcare workers in the hospital. In addition to providing direction for best practice of hand hygiene, it is also recommended that hospitals monitor hand hygiene in order to reinforce its importance and direct interventions (e.g., placing posters reminding health care workers about hand hygiene, talking with healthcare workers, installing different hand sanitizer pumps or changing their locations). The metric that is generally monitored is the *hand hygiene rate*, defined as the ratio of hand hygiene events (defined as a session of hand hygiene consisting of cleaning the hand with either alcohol-based hand rub or soap and water) to total hand hygiene opportunities.

Hand hygiene opportunities are defined differently by different agencies. The World Health Organization has produced a guideline called the "Five Moments of Hand Hygiene" as shown in Figure 1.1. This guideline outlines five different scenarios in which a healthcare worker has a hand hygiene opportunity– before patient contact, before an aseptic task, after a body fluid exposure risk, after patient contact, and after contact with patient surroundings. However these opportunities are often difficult to observe as they occur within the patient room. A more common method is known as "Wash In, Wash Out" hand hygiene, wherein a healthcare worker has an opportunity to perform hand hygiene before entering and after leaving a patient room [4]. This metric is more commonly used in hospitals as it is easier to observe once alcohol-based hand rub dispensers have been placed outside of rooms.

Hand hygiene rates have traditionally been recorded by human observers. Another traditional approach is to measure the amount of hand rub used on the hospital floor as a proxy for hand hygiene activity [5]. Both of these approaches have their drawbacks. The human observer approach has proved to be problematic as the observers can affect the behavior of healthcare workers, causing observed hand hygiene rates to be much higher than they should be [6, 7, 8]. The hand rub volume approach is a

Figure 1.1: Five Moments of Hand Hygiene. From [1].

noisy metric for hand hygiene rates among healthcare workers as the hand rub dispensers can be accessed by anyone on the hospital floor, including visitors, and the metric itself may be affected by the type of dispenser or hand rub product used.

Various types of electronic systems have been proposed for assisting with monitoring of hand hygiene rates in hospitals. Several approaches use cameras to observe hand hygiene. One system, Arrowsight, combines cameras that are focused on hand rub dispensers and sinks with motion sensors to capture video of potential hand hygiene opportunities. These opportunities were viewed by remote auditors who marked whether the workers performed hand hygiene [9].

While monitoring the rate of hand hygiene is important, the duration of and technique used during a hand hygiene events are important components of the efficacy of hand hygiene [10]. The World Health Organization's recommendations for effective hand hygiene using alcohol-based hand rub can be seen in Figure 1.2. Various motions are recommended in order to ensure that alcohol-based hand rub is distributed evenly throughout the hand, particularly in the nail beds which are often the dirtiest [11]. In addition the World Health Organization recommends that hand hygiene be performed for a certain amount of time, since hand hygiene of longer duration has been shown to be more effective. The previously listed systems only provide rates of hand hygiene compliance without information on duration or technique.

Electronic Hand Hygiene Technique Recognition

Many have considered the question of automatically tracking the number of hand hygiene attempts ([12, 13, 14, 15] are just a few commercial systems that offer this functionality), but few have tackled the automatic recognition of hand hygiene technique. Predominant approaches to this problem have utilized vision-based recognition. SureWash is a commercial system which uses vision-based recognition as part of a training system for healthcare workers and visitors [16]. However, this system is not meant to perform recognition of hand hygiene technique in situ. A vision-based system which could be mounted over a sink for technique recognition in situ is proposed in [17]. They report detection rates between 65.99% and 97.81% for various hand hygiene movements.

The major drawback of vision-based hand hygiene technique recognition is that they require that the observed hand washing occur in a certain small area because the hands must stay clearly in view of the camera. SureWash enforces this by showing video of the hands to the user so they can make sure to stay in frame, and in [17] they

Figure 1.2: WHO-recommended Alcohol-Based Hand Rub scrubbing motions. From [2].

ensure that the user is in frame by having them wash using soap and water at a sink. These assumptions are quickly violated if we attempt to extend vision-based methods to hand hygiene techniques using alcohol-based hand rub because users may place themselves arbitrarily in relation to the pump and may move away from the pump while they wash, removing themselves from frame. In addition vision-based methods suffer under poor lighting conditions which makes it difficult to extend their use in a general hospital setting.

While current approaches are insufficient for use in a general hospital setting, improvements in technology have given rise to smaller processors, batteries, and sensors. It is now possible to place wristbands on the wrists of healthcare workers and monitor them continually on the hospital floor, avoiding the restrictions of vision-based methods.

CHAPTER 2:

ACTIVITY RECOGNITION

Activity recognition is a field in the crossroads of sensor networks and machine learning. In this field subjects of interest are instrumented with sensors and machine learning methods are used on the sensor data to determine which activities are taking place. This chapter provides some background on machine learning metrics and techniques and also reviews special applications in activity recognition and gesture detection related to recognition of hand hygiene technique using wrist-based sensors.

Supervised Machine Learning Methods and Metrics

Before discussing the various results in the field of activity recognition it is useful to have an understanding of the methods and metrics we will discuss. This section gives a general overview of supervised machine learning and introduces several metrics that will be used later in the thesis. In addition common trade-offs are discussed.

It is first necessary to explain the general concept of a supervised machine learning problem. In machine learning a data set is considered to be embedded in a *feature space*. As an example consider Figure 2.1. In this figure we have a data set consisting of snowy days in Chicago, IL. This data set has been embedded in a two-dimensional feature space–one feature being the low temperature of the day and the other being the date. The number of features determines the dimensionality of the feature space–so for instance, if Figure 2.1 also included the humidity that would be a third dimension in the feature space. Each point in Figure 2.1 represents an *instance* in the data set. In this example the instances are days, but an instance could be a participant or (in this thesis) a half-second of data.

In Figure 2.1 each point also has a color which reflects whether the day was snowy or not. In general this can be thought of as a class label. Figure 2.1 shows a problem with two possible classes, but there can be many more. Machine learning methods find ways to divide the feature space to create a *classifier* that can predict the class label of an unknown instance. An example of a naive classifier that predicts that the day will be snowy if the low temperature is below 1.1 degrees C is depicted in Figure 2.1. In general the division of space does not have to be linear and can take on many

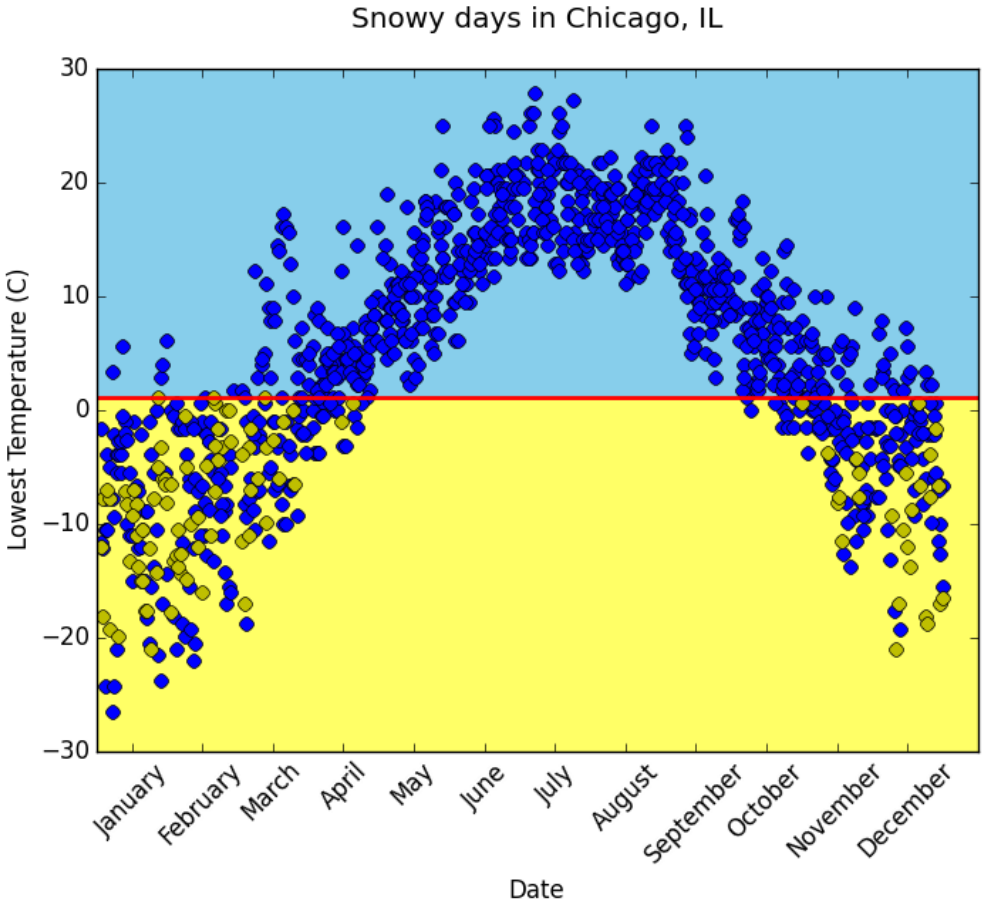

Figure 2.1: Example Data Set Embedding: Snowy Days in Chicago. Snowy days are indicated with yellow points. A naive classifier divides the space with the red line. The classifier would guess that the day would be snowy if the low temperature was below 1.1 degrees C. Data is weather at Chicago O'hare International Airport from January 1, 2012 through December 31, 2014 as provided by the National Centers for Environmental Information.

different forms depending on the method used.

A good classifier will correctly predict the class label of an unknown instance. In order to form an idea of how a given classification method will work in a deployment the data set is split into two sets: a *training set* and *test set*. The training set is the set of instances that the machine uses to create the division in space–to train the classifier. The test set is the set of instances that the machine will treat as unknown and predict class labels for. Since the test set consists of instances where the true class is known we can then compare the classifier's prediction to the truth. However, this approach is vulnerable to variation depending on the choice of training and test set. Using the example in Figure 2.1 imagine that the training set contained only snowy days in December–the class dividing line could look very different and then we could think that a linear classifier performs poorly when in fact the training set was chosen poorly. In order to avoid this problem results are generally reported using 10-fold cross validation. In 10-fold cross validation the data set is split into 10 parts (or folds). One part is held out as a test set and the rest used for training. Each part has its turn as the test set. In the end results are averaged over all 10 folds to create a metric that better reflects the average performance of the classification method on this problem.

The basis for the calculation of most metrics is the confusion matrix. A confusion matrix provides an organized way to examine the errors made by a classifier. Table 2.1 illustrates a standardized confusion matrix for a binary classification test. In this example we can see that the classifier is detecting a positive and negative class. The positive class could represent anything that we want to detect with the system–for example, the presence of hand hygiene (in our system). As we can see in Table 2.1 we can have two types of errors: a false positive or a false negative. A false positive can be seen as a false alarm or a type I error in statistics. A false negative can be seen as a miss or a type II error. When a confusion matrix is created from an example classification the cells will be filled in based on the number of instances that apply to each case. An example of a two class confusion matrix can be seen later in Table 4.5.

Various metrics can be calculated using a confusion matrix. A metric commonly used in the medical field is sensitivity and specificity. Sensitivity (also known as hit rate, recall, or true positive rate) is defined as $TP/(TP + FN)$. It is the ratio of correctly detected positive instances to the total number of positive instances. In our application we can think of this as representing how frequently hand hygiene was correctly detected. Specificity (also known as true negative rate) is defined as

		Classified As	
True Class		+	-
	+	True Positives (TP)	False Negatives (FN)
	-	False Positives (FP)	True Negatives (TN)

Table 2.1: A standardized confusion matrix.

Metric	Equation
Sensitivity, recall, hit rate, true positive rate	$TP/(TP+FN)$
Specificity, true negative rate	$TN/(TN+FP)$
Positive predictive value, precision	$TP/(TP+FP)$
Accuracy	$Total\ Correct/All\ Instances$

Table 2.2: Machine Learning Metrics Used in this Thesis. Bold terms indicate those in thesis.

$TN/(TN+FP)$. It is the ratio of correctly detected negative instances to the total number of negative instances. In our application we can think of this as representing how frequently non-hand hygiene was correctly detected.

Sensitivity and specificity are good for a binary classification problem, but specificity cannot be calculated for a classification problem with more than two classes because the number of true negatives is unclear. We can see a confusion matrix for one such classification problem with ten classes in Table 4.6. As an example, consider calculating the number of true negatives using that confusion matrix when the positive class is PR (Palm Rub). Do we simply sum the number of correctly classified instances that are not PR? That ignores the fact that some of the incorrect classifications in those other classes were not classified as PR. Do the instances that were truly LTS (Left Thumb Scrub) and were incorrectly classified as RTS (Right Thumb Scrub) count as true negatives because they were correctly not classified as PR or are they false negatives because they were incorrectly classified? Because of these difficulties the metrics of positive predictive value and accuracy are used for classification problems with more than two classes.

Accuracy is defined as $Total\ Correct/All\ Instances$. This can be thought of as an average of the sensitivity of each class, weighted by the prevalence of that class. Positive predictive value (also known as precision) is defined as $TP/(TP+FP)$. In our application we can think of this as how often our system was correct when it detected hand hygiene. This is different from specificity, which represented how often our system

was correct when it did not detect hand hygiene.

Accuracy and positive predictive value are the metrics used in this thesis because many of the classification problems consist of more than two classes.

In a detection task like the one discussed in this thesis there are two competing goals: 1) the machine should detect when the class of interest (i.e., hand hygiene) is present and 2) when the machine detects that the class of interest is present it should be correct. In most cases there is a trade-off between goals 1 and 2. As a trivial example, imagine that we only cared about goal 1, detecting when hand hygiene is present. In that case we could easily fulfill that goal by constantly guessing that hand hygiene is present–we would perform perfectly on goal 1 but not on goal 2 because we would be frequently wrong. In reality one can think of goals 1 and 2 as being opposing points on a spectrum. Most machine learning methods try for a happy medium, but have various parameters that can change in order to accommodate moving toward one goal or the other depending on preference.

In order to examine the trade-off one can use a Receiver Operating Characteristic (ROC) curve. This plots the true positive rate (increase this to meet goal 1) for a given class against the false positive rate (decrease this to meet goal 2) for the same class. An example ROC curve can be seen later in Figure 5.8. In it we can see that as the true positive rate increases the false positive rate also increases, as expected. Given a curve like this one could choose a certain point with the trade-off desired for a given application.

Activity Detection Using Wrist-Based Accelerometers

Wrist-based accelerometers have been used as components in many activity detection systems. Not all of them are closely applicable to the problem of hand hygiene recognition. For instance, gait detection, a research area with a large body of work containing many systems which incorporate wrist-based accelerometers, is mostly concerned with large changes in the frequency of an action (e.g. the change from sitting to standing to walking). However, these large differences in frequency do not exist between different types of hand hygiene technique. In this section a few selected examples where wrist-based accelerometers were used to detect everyday activities are reviewed in order to facilitate understanding of results presented later in this thesis.

One class of activity recognition applications that has made use of wrist-based accelerometers involves recognizing exercise activities (e.g., calisthenics, weight training). Most of these systems have made use of multiple sensors throughout the body

and observe movements that are simpler than hand hygiene. In [18] a wrist-based accelerometer is combined with sensors on the upper arm and hip to detect exercise motions such as dumbbell curls or push ups. Cheng et al. work on recognizing classes that have no examples in the training set and report recall from 45.6% to 95.9% on a data set consisting of 20 subjects performing 10 motions. In [19] the duration and number of repetitions of an exercise are detected using a wrist worn sensor that collected accelerometer and gyroscope values from 94 participants for an average of 38 minutes each. Morris et al. report recall of above 85.6% for the duration of exercise and 98.2% for the number of repetitions using a dataset consisting of 14 different exercises (e.g., push ups and squats). In [20], Velloso et al. distinguish between the dumbbell curl and four common incorrect variations of the dumbbell curl using a wristband, belt, and glove equipped with an accelerometer, gyroscope, and magnetometer. They report precision from 84.9% to 98.2% and recall from 85% to 98.2% using a data set containing 6 subjects.

Another class of activity recognition applications that has made use of wrist-based accelerometers involves recognizing activities of daily living (e.g., brushing teeth, making coffee). All of these systems made use of multiple sensors placed throughout the body. In [18] Cheng et al. use a public dataset of 34 daily life activities collected using an accelerometer worn on the wrist and the hip from one subject over seven days. They report a mean precision of 52.3% and recall of 73.4%. In [21] Hong et al. use accelerometers on the thigh, waist, and wrist and an RFID reader equipped glove to classify daily activity from 16 hand activities (e.g. shaking hands) and 5 body states (e.g. sitting). They report recall of 56.35% to 94.57% using only the accelerometer values and 94.38% to 98.34% when the RFID data is included using a data set containing 15 subjects.

Gesture Recognition

Since the introduction of the Nintendo Wii (an accelerometer equipped gaming control) and increased availability of accelerometers on smartphones, accelerometers have been frequently utilized in gesture recognition. In this section examples of gesture recognition using wrist-based accelerometers are reviewed as a basis for understanding the work presented in this thesis.

A few papers have explored the use of wrist-based accelerometers for gesture detection. In [22] a wrist-based accelerometer is used to detect the motions associated with smoking. Parate et al. report precision and recall of 72% and 85% when using

random forests and 91% and 81% when using a conditional random field. In [23] a wrist-based accelerometer is used to count the number of bites in a meal as part of a calorie counter. Ramos et al. report an accuracy of 71.7% with a K-Nearest Neighbors classifier and 84.3% with a Hidden Markov Model that took into account the ordering of activities within a meal. These examples are notable for their use of wrist-based sensors to detect movements of the hand, the motions that they detect are quite large scale. Small scale motions are more difficult to detect with a wrist-based accelerometer, so others have introduced extra sensors in order to obtain good results. In [24] Li et al. use both a wrist-based accelerometer and an surface electromyographic (EMG) sensor which detects muscle activation. Using this system they were able to recognize a vocabulary of 120 signs from Chinese sign language with an accuracy of 96.5%.

A more substantial body of work exists in the realm of identifying gestures using handheld accelerometers. Most of the gesture sets in the literature consist of very simple motions made on a plane ([25, 26, 27, 28] are only a few examples of handheld accelerometer based gesture recognition using extremely simple motions like circles or swiping motions) but some have considered larger and more complex sets of gestures.

In [29] an accelerometer equipped pen is used to recognize a set of gestures and digits. Wang et al. use a probabilistic neural network and report recognition rates of up to 98% in the case of digit recognition and up to 98.75% in the case of simple gesture recognition. In [30] Agrawal et al. seek to recognize English characters using a mobile phone to "write in air". In order to make this problem tractable they had to ask users to use one corner of the phone to write with (since using the accelerometer alone it is impossible to differentiate between a rotation of the phone and a linear movement that would have resulted in the same acceleration values) and to pause slightly between strokes in a character (to enable them to take a quick snapshot of the accelerometer values and infer the phone rotation from the force of gravity). They decompose every character into a series of common simple strokes and treat this as a grammar. Agrawal et al. report character recognition rates of 91.9% for trained users and 78.2% for novice users (human readers had 83% and 85.4% recognition). Classification accuracy improved when writing was on a table instead of in the air, presumably because this reduced variability in at least one plane of motion. An interesting facet of this paper was their studies in the hospital with impaired patients (e.g. someone suffering partial paralysis from a stroke). On a doctor's advice they attempted to simulate performance on a patient that was only slightly impaired by having a small set of participants write with their non-dominant hand. On this set

they report accuracy of 81.73%.

This literature can provide good insights into possible avenues of inquiry, but it is important to remember that there are differences between gesture recognition and hand hygiene recognition. Perhaps the largest difference is the location of the accelerometer. In the case of a pen or joystick the system can detect motions made with the hand with high accuracy because it is aligned with the hand, not just the wrist. By contrast approaches which use wrist-based accelerometers confine themselves to predicting simple gestures like those associated with smoking or eating. Inferring more intricate hand movements solely from wrist accelerometery is more difficult.

CHAPTER 3:

SYSTEMS USED FOR DATA COLLECTION

The quality of hand hygiene recognition is affected by the system chosen for data collection. In this thesis various systems were used to collect data sets consisting of hand hygiene motions in an attempt to explore the utility of different sensing modes and other system considerations. The sensors evaluated consist of accelerometers and MEMS-factor gyroscopes. One of the platforms is a custom platform constructed for these experiments. We also evaluate a "fitness wristband" which can be programmed to record accelerometer data to flash memory.

3.1 Custom System

The remainder of this section describes our custom platform. Our platform design aims to satisfy some general constraints: the device should be small and relatively unobtrusive so that healthcare workers will tolerate wearing it; and the device needs to be low-cost, have some signal processing and storage capabilities, wireless connectivity, and adequate power for at least eight hours of deployment (before replacing/recharging battery). We constructed a small circuit board hosting a CC2531 SoC [31], an MMA7455L accelerometer [32], two LEDs, and a PCB antenna. The board's size, 50mm x 32mm, includes some extra conveniences: an on/off switch, one mini-USB port, one micro-USB port (used for programming), and a recharging chip/circuit for a lithium-polymer battery, recharged via USB-supplied power. It is possible to trim these extra conveniences, leaving a working dimension of 36mm x 16mm, small enough to fit in some commercial watch cases. Our experiments used the larger form factor fit into a 3d-printed ABS plastic case fitted with a cloth and velcro wristband.

The CC2531 has the equivalent of a CC2520 radio, which supports IEEE 802.15.4 messaging. We chose the CC2531 due to its low cost, power-conserving sleep mode, and radio compatibility existing devices. A disadvantage of the CC2531's MCU is the 8KB RAM, which limits flexibility for our applications; newer ARM-based SoC's would offer better signal-processing and memory at somewhat higher cost. To avoid some timing complications, experiments in this thesis first record sensor values into a 6KB

buffer, and in a second step offload the buffered values in a burst of messages. Because both wrists carry such devices, we used two radio frequencies, one for each wrist, so that offloading proceeded in parallel. Recording values from the accelerometer at 125 Hz is driven by a timer in the accelerometer, which interrupts the MCU to yield each (x,y,z) reading.

Another component of the experimental system is a gateway board that forwards 802.15.4 messages to a PC via a serial to USB converter (we used two of these, one for each programmed radio frequency). The final device in the setup is an alcohol-rub dispenser modified to trigger a microswitch on each dispensing event, and an attached board and battery that transmits 802.15.4 messages. The scenario of an experiment is thus: (*i*) a user dispenses from the bottle, which transmits messages received by both wrists (both frequencies) and a basestation PC; (*ii*) the user performs some hand washing activity for around sixteen seconds (also known by LED flashes on the wrists), which fills the 6KB buffers with sensor values; (*iii*) sensor data then streams to the PC in a sequence of messages.

Later experiments on the custom platform attached another small sensor board to the wrist device, a "motion processing unit" similar to the Invensense family of MPU chips [33]. We used a board with a signal processing API that fuses and filters sensed values from an accelerometer, gyroscope and compass [34]. This API simplified obtaining Euler angles (orientation) from the gyroscope. The custom sensor board functioned as an Inertial Measurement Unit (IMU) and the resulting Euler angles represent pitch, roll, and yaw of the wrist.The driver on the CC2531 converted floats into 8-bit integers representing the angles, so that the same message format could be used as in earlier experiments.

3.2 Geneactiv Wristband

The Geneactiv Wristband [35] is equipped with a 3-axis accelerometer, photometer, and thermometer. This device records data to flash memory, for up to one week, at 100 Hz sampling rate including timestamps. We found that wristband time can drift from initial synchronization to a reference clock by up to 1-2 seconds after five hours, so we limited some experiments to several hours within clock synchronization. Programming the wristbands and offloading data is done by USB connection to a workstation.

3.3 Observer Marking Systems

In addition to collecting measurements from the wristbands, it was frequently necessary in our experiments to record marking signals sent by operators. In initial pilot studies this marking was done using a simple script which recorded when a key was pressed on a computer. In later experiments more mobility was necessary and a simple iPhone app was developed to take recordings.

This iPhone app consisted of three buttons which the operator could press, one to indicate the start of a hand hygiene event, one to indicate the end of a hand hygiene event, and a third small button to indicate an event with an unknown start or end point (e.g., if the healthcare worker began a hand hygiene event but then turned in such a way that the end of the hand hygiene event could not be observed).

Time synchronization between the wristbands and marking systems was obtained by utilizing a shared time server. The time server was used by the iPhone for timestamping observer markings. The time server was also used by the laptop used to program the Geneactiv wristbands. In this way observer markings and wristband measurements were synchronized within 1-2 seconds of each other as long as observations were taken within a few hours of wristband programming.

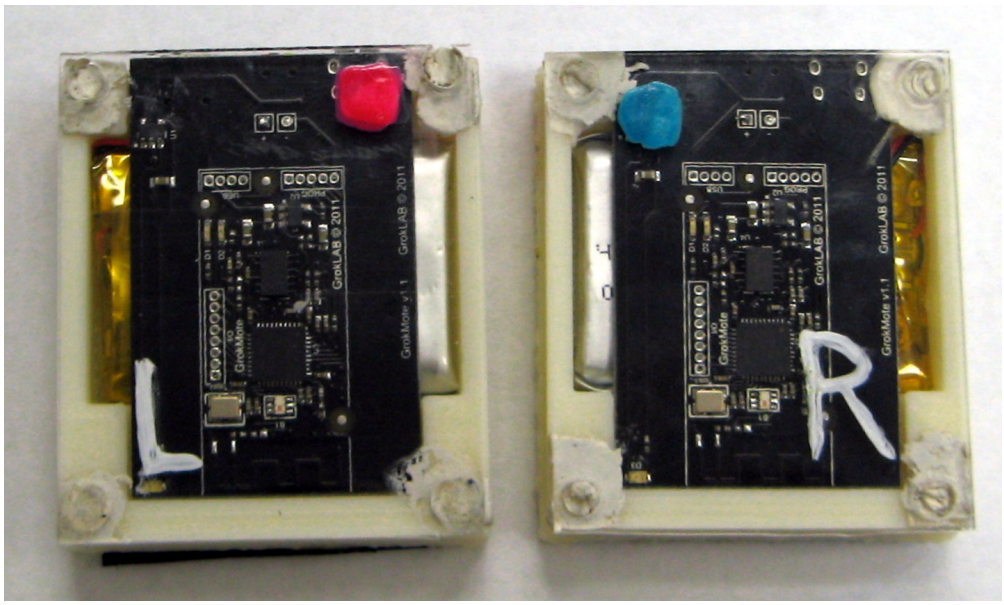

Figure 3.1: First custom-built wristband. Takes 3D-accelerometer measurements at 125 Hz for approximately 16 seconds, then sends data to base station.

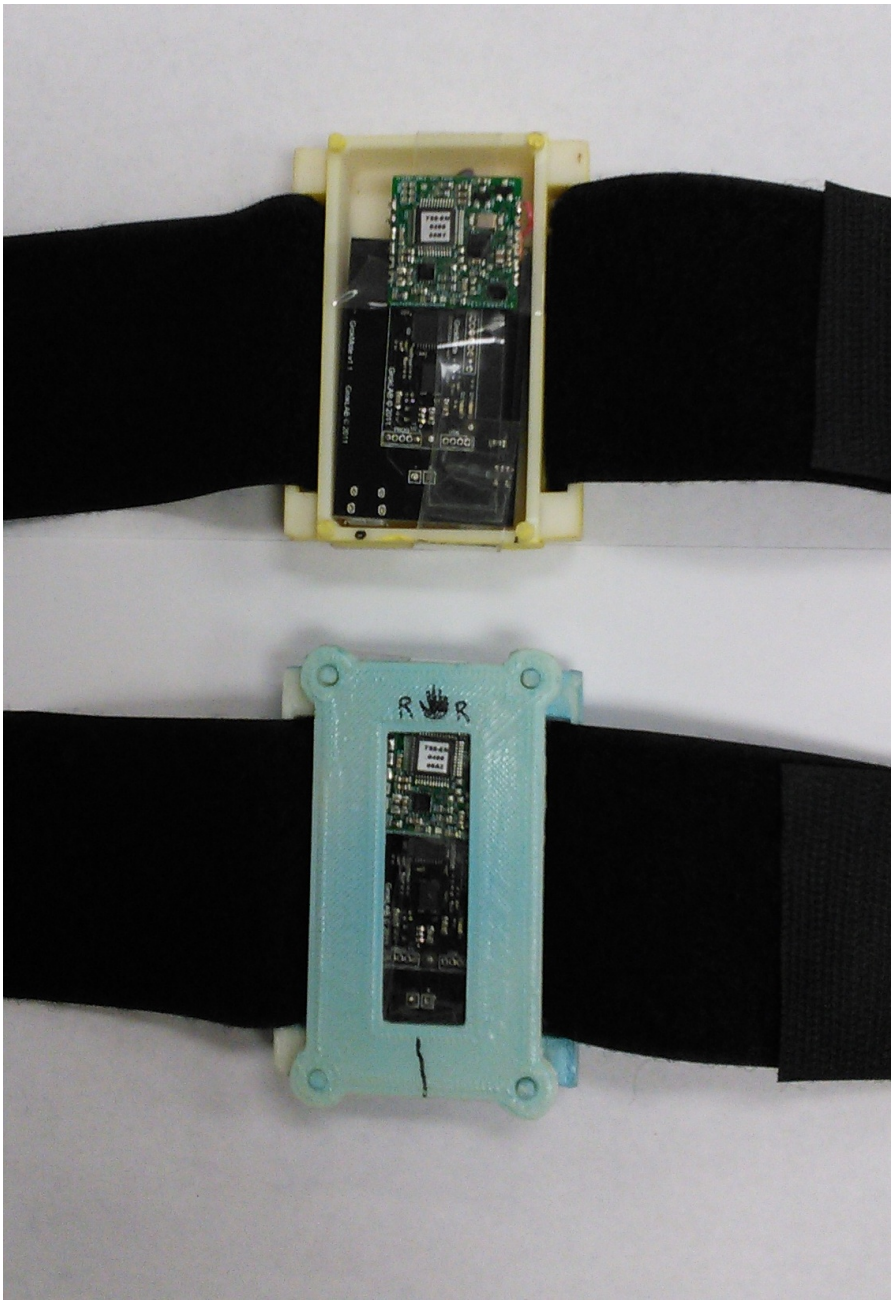

Figure 3.2: Second custom-built wristband. Takes 3D-accelerometer and orientation measurements at 125 Hz for approximately 8 seconds, then sends data to base station.

Figure 3.3: Geneactiv wristband system. Records 3D-accelerometer measurements at up to 100 Hz for up to one week. Does not have wireless capabilities.

CHAPTER 4:

RECOGNIZING HAND HYGIENE DURATION AND TECHNIQUE: PILOT STUDIES

4.1 Data Sets

This section introduces the vocabulary of specific hand hygiene motions and explains the operational protocol used to direct subjects in our studies, which range from ten subjects up to over one hundred subjects for one experiment. For smaller studies, we found volunteers (chiefly students); the largest experiment was in a working hospital and included 116 health-care workers.

4.1.1 Hand Hygiene Technique

Different hand hygiene protocols vary in the use of cleaning or disinfecting agents; pre-surgery scrubbing is more intense than typical scrubbing before entry and after exit from a patient area. Currently there is no monolithic technique for hand hygiene and different organizations may specify different protocols. In this thesis, we concentrate on a World Health Organization (WHO) recommendations for alcohol-based hand rub. The WHO recommended protocol contains six (or ten, counting symmetric actions) hand motions [36], reproduced in Figure 1.2. In addition to such diagrams, there are training videos demonstrating proper technique[37]. The various motions aim at cleansing different areas of each hand.

While health-care workers are trained in proper technique, the full WHO protocol is rarely used in practice. The true measure of effectiveness of hand hygiene is the amount of the hand covered with alcohol-based hand rub. It is conjectured that normal or "in the wild" hand scrubbing reasonably covers hands with AHBR, though some crucial areas, such as the beds of fingernails, may not be sufficiently scrubbed. The area of the hand covered by alcohol-based hand rub is difficult to measure in the field, but two measurements are considered to be proxies for coverage: duration of hand hygiene and adherence to WHO technique. These are difficult for a human observer to measure in practice as they would need to be quite close to the healthcare worker, so observation of duration or technique by an automated system is of interest.

	Accelerometer & Orientation Data set	116 HCW Data set	Geneactiv Data set
Number of Participants	10	116	30
Palm Rub	X		
Fingertip Scrub (R)	X	X	
Fingertip Scrub (L)	X	X	
Interlocking Fingers	X		
Thumb Scrub (R)	X		
Thumb Scrub (L)	X		
Knuckle Twist (R)	X		
Knuckle Twist (L)	X		
Back of Hand (R)	X		
Back of Hand (L)	X		
Wrist Rub (R)	X		
Wrist Rub (L)	X		
Wild		X	X
Walking		X	
Confounders			X

Table 4.1: Motion Types and Presence in Pilot Data Sets: (R) or (L) indicates that the right or left hand is the one being cleaned or the one on top of the other.

4.1.2 Data Collection

The platforms described in Chapter 3 provide different sensor data formats at different sampling rates. Because some of our research questions compare platform outputs, we generally reduce data to eight-bit samples per $\{x, y, z\}$ channel, coupled with a sequence number to detect dropped samples due to lost messages and other transfer errors. Sampling rate is either 125 Hz or is scaled/interpolated to 125 Hz if needed. Following trials with subjects, scripts or operational procedures verified that complete payloads (using sequence numbers and CRC fields) were received, either discarding imperfect runs or requesting that subjects repeat experiments.

For both custom platform and Geneactiv wristbands, subjects wore the sensors on both wrists. Typical experiments with Geneactiv units, which do not have radio, are preceded by clock synchronization via vendor-provided PC software; we tested the synchronization and found drift to be sub-second for a few hours. The Geneactiv experiments were synthetic: subjects were asked to engage in requested hand motion activities at marked times for some specified duration while an observer marked the time and duration of activities. The experiments with our custom platform were also synthetic: subjects were instructed on various scrubbing techniques, taken from Figure 1.2, then performed the requested motion for approximately sixteen seconds, after which the data downloaded wirelessly from the wrists.

The data sets collected by these procedures vary in the known motions and raw data formats. Table 4.1 shows which motions are present in various data sets. The largest of our trials, with 116 healthcare workers, was collected using the first version of the custom-built platform and included three motions: hand-washing "in the wild" (labeled as Wild HH in later sections of the thesis), the WHO recommended fingertip scrub, shown as panel 7 of Figure 1.2, engaged as separate motions once for each hand, and walking without any scrubbing, which represents the typical behavior of a health care worker after a hand hygiene event. In addition to samples consisting of only these motions, a fourth sample was taken in which each participant performed Wild HH and walked after finishing, with an observer marking when the participant finished washing their hands. This experiment was designed to be collected within five minutes (because healthcare workers were volunteers, we limited this larger trial's scope to a few motions, so as to minimize interference with job duties).

Another data set on small trials of 10 subjects used ten motions taken from Figure 1.2; the figure shows only six scrubbing motions, but four of these should be performed with both left- and right-handed orientation. These were collected using the second version of the custom-built platform which sensed orientation (3.2). Two smaller trials using the same motions were also collected. A trial of 5 subjects was collected using the first generation of the custom-built system, and a trial of 7 subjects was collected using the second generation of the system using only the orientation readings and no accelerometer readings. These trials took approximately 20 minutes for each subject; participants were not healthcare worker volunteers.

Data from the Geneactiv trials consists of various durations of hand hygiene (from ten seconds to one minute) and confounder motions designed to engage the hands in quick, correlated motions: opening a jar containing candy, opening and eating the candy; tying shoes; and applying bandages. Such motions were collected from 30 healthcare workers. This was designed to be collected within seven minutes per participant.

4.2 Feature Set Choice

For each type of sensor, accelerometer or angle, the software samples eight bit values on each of the $\{x, y, z\}$ channels. Figure 4.1 shows "raw" data for a hand hygiene motion. Several typical aspects are observable in the plots. The large swings in accelerometer readings are due to clipping: the underlying signed accelerometer values

are ten bits but are saved as eight bit values when the platform records them into a buffer for later transmission. This clipping effect can be removed by filtering. Another phenomenon within small sample windows is periodic behavior typically with most energy along one or two axes (consider the case where hands rub flatly, almost entirely on the z-axis, with little movement on other axes). Initially, we suspected that transforming signals to frequency domain could produce a useful feature for classifying different motions. It turned out that different people have different rhythms of scrubbing, change speed during washing, and rotate arms or hand angles while washing. Thus periodicity varies and signal energies project onto different axes over the course of one scrubbing episode and we were unable to exploit the frequency domain transform.

In order to train the classifier the raw data must be transformed into a set of features S. We define the raw data as $R = (R^{rx}, R^{ry}, R^{rz}, R^{lx}, R^{ly}, R^{lz})$, where rx, ry, rz are the x, y, z axes on the right wrist and lx, ly, lz are the x, y, z axes on the left wrist. We define the set of axes as $A = rx, ry, rz, lx, ly, lz$. Each axis in R consists of readings from time 0 to time t. R_i^a is the raw reading on axis a at time $i, 0 \leq i \leq t$. $R_{i,j}^a$ is the set of readings on axis a from time i to time $i + j$, $0 \leq i < i + j \leq t$.

In all cases, the raw data was resampled if it was not collected at the target sampling rate. To do this the signal was estimated by interpolation and then resampled at the target sampling rate. The default setting was a target sampling rate of 125 Hz. The resampled raw data was then split into windows of length w. These windows were not overlapping. The default window length of the system was .5 seconds.

Once the windows of data had been collected, each window of raw data was transformed into a vector of features. Two types of novel features are introduced in our system. The first is the *crossing rate*, which can be intuitively thought of as a change in the dominating axis of movement. This feature was originally derived from the well known zero-crossing rate [38], except that the reading on another axis is used in place of zero. The crossing rate of an axis is the sum of the *crossings* between that axis and all other axes in A. A *crossing* is detected between two axes a_a and a_b at time i according to the following equation:

$$c(R_i^{a_a}, a_b) = \begin{cases} 1, & if\ (R_i^{a_a} > R_i^{a_b} \wedge R_{i+1}^{a_a} < R_{i+1}^{a_b}) \\ & \quad \vee (R_i^{a_a} < R_i^{a_b} \wedge R_{i+1}^{a_a} > R_{i+1}^{a_b}) \\ 0, & otherwise \end{cases} \tag{4.1}$$

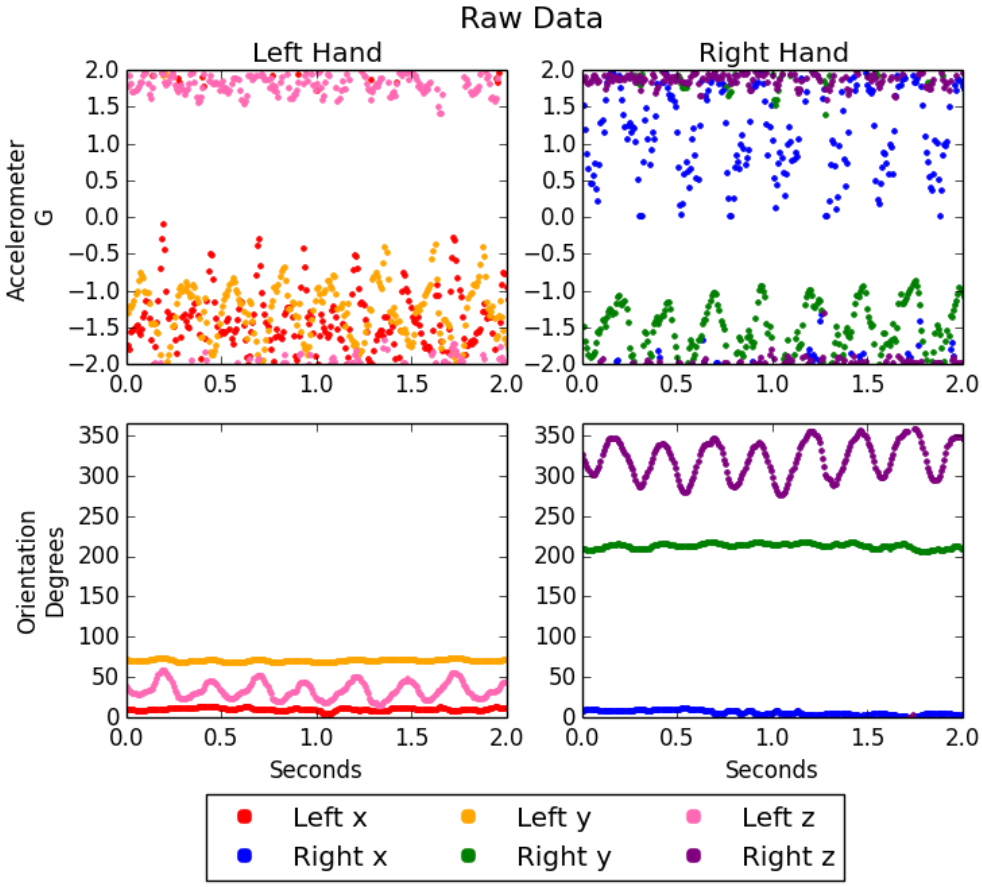

Figure 4.1: Example Raw Data From Accelerometer and Orientation Sensors. Eight bit values were converted into units of G for accelerometer and degrees for orientation.

The second is a peak, which determines if the rate of change of the axis has changed beyond the norm. Originally this feature was a simple threshold metric which was later refined to self-adjust to the current characteristics of the signal by using the mean of the recent readings from the signal. A peak is detected in an axis a at time i according to the following equation:

$$p(R^a, i) = \begin{cases} 1, & if\ R_i^a - R_{i-1}^a > P * \overline{x}(R_{i,i-M}^a) \\ 0, & otherwise \end{cases} \tag{4.2}$$

where $\overline{x}$ is the mean of the indicated data. Two types of peaks were detected: peaks and soft peaks. Soft peaks were only counted if the reading in question was not a peak. For peaks the value of P was set to 1.5, for soft peaks the value of P was set to $.25$. The value of M was set to 12. All of these settings were experimentally determined.

The features used in the system were as follows: the mean of each axis, the standard deviation of each axis, the crossing rate for each axis, the total number of peaks across all axes, the total number of soft peaks across all axes, the sum of all axis crossing rates, the sum of all data in the window, and the standard deviation of all axes.

When the features from the accelerometer dataset were naively applied to the orientation data results were not as good as we had expected. Some features had to be transformed to account for the Euler angles being a circular quantity which require special handling to improve results. A simple approach was used which implemented a changed distance function for finding the distance between two points,

$$d(a, b) = min(a - b, 360 - (a - b)) \tag{4.3}$$

which was used to update the mean, standard deviation, and peak functions. For the mean both the arithmetic mean and the angle directly opposite were obtained. The value that was the smallest distance from the input angles was returned as the mean. For standard deviation the distance function replaced the standard subtraction from the mean. For the peak the distance function replaced the previous difference calculation.

A similar process of feature selection was performed for the orientation feature set. In addition to the standard features the maximum and minimum values of the axis were found to positively affect classification so they were added to the orientation feature set.

Feature	Abbreviation
All Time-based Features	T
Maximum	MAX
Minimum	MIN
Mean	$\bar{x}$
Standard Deviation	σ
Cross Correlation	CC
Zero Crossing Rate	Z
All Frequency-based Features	F
Discrete Fourier Transform (DFT)	DFT
Mean of DFT	$\bar{x}_{DFT}$
Cross Correlation of DFT	CC_{DFT}

Table 4.2: Features Explored for Initial Feature Set. Results of this exploration can be seen in Figure 4.2.

4.3 Results

The default sampling rate in these results is 125 Hz and the default window length is .5 seconds.

The data was classified using a multiclass k-nearest neighbors classifier [39], which labels an instance by the majority label of the k nearest instances in the feature space, with k=3. Unless labeled otherwise, all results were obtained using disjoint windows of data. Unless marked otherwise, the metrics reported consist of the classification accuracy obtained using 10-fold cross validation. Classification accuracy is the number of correctly classified windows of data divided by the total number of windows of data.

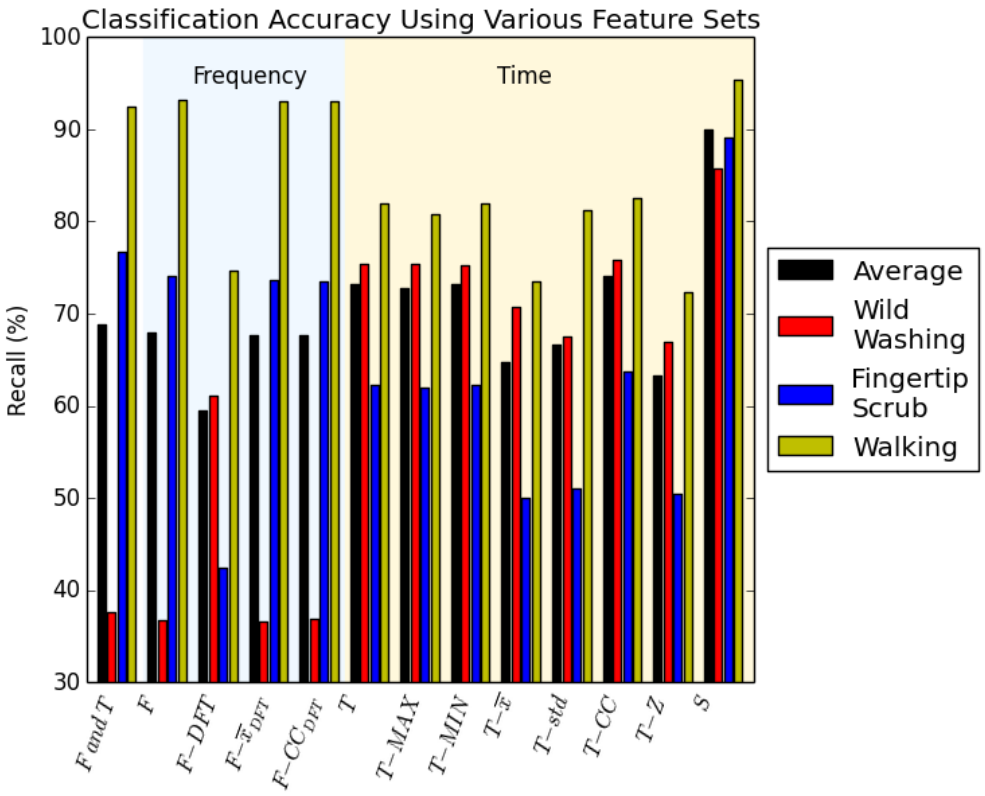

Figure 4.2: Pilot Studies: Classification Accuracy Using Different Feature Sets. Time-based features worked better than frequency-based features on average. In addition time-based features produced good classification results for all classes while frequency-based features worked well at discriminating walking from hand hygiene motions, but did not discriminate well between different hand hygiene motions. While the frequency-based features consistently outperformed time-based features when detecting walking, the final feature set which consists of time-based features that have been further refined classifies walking better than any of the preliminary frequency-based feature sets.

4.3.1 Feature Choice

The choice of feature set is key to the proper function of any classification method. A good feature set is one which transforms readings from the raw data space into a feature space where the various classes are easily separable. The selection of such a set is non-trivial and can make the difference between the success and failure of any classification scheme.

Features can be divided into two general categories: time-based and frequency-based. Time-based features are calculated using the signal as it is in the time domain while frequency based features are calculated using the signal after it has been transformed into the frequency domain, (e.g through a Fourier transform). As a preliminary step we determined which features were useful by examining a sample set of common time and frequency-based features listed in Table 4.2.

Figure 4.2 shows that the time-based features consistently outperform the frequency-based features. The average performance is similar for all feature sets, but the frequency based features' accuracy is inflated by accurate classification of walking while other classes are poorly classified. By comparison, the time-based feature set classifies all classes with similar accuracy. This suggests that the time-based features provide more generalizable results than the frequency-based features. We added more time-based features to the set used in 4.2, including features already known in the domain and features created after manual examination of the signal (e.g., axis crossing rate). These time-based features were further refined using Correlation-based Feature Subset Selection [40] to produce our final feature set discussed in Section 4.3.1.

The main difference between the frequency-based and time-based feature sets was that frequency-based features performed better when classifying the walking motion. After refinement the final feature set, composed entirely of time-based features, classifies walking better than any of the frequency-based sets.

The final feature set includes two custom features, peaks and crossings. In Figure 4.3 we can see that each feature type chosen for the final feature set improves classifier performance.

4.3.2 Classifier Choice

It is also worthwhile to show why a k-nearest neighbors classifier was used. The machine learning software Weka [41] was used to evaluate several different classifiers. Table 4.3 lists the performance and training time of various classifiers that could have

	116 HCW Data Set	
Classifier	Accuracy	Time (s)
K-Nearest Neighbors	90.2%	0.03
Decision Tree	86.7%	3.44
Neural Network	88.7%	118.03
Naive Bayes	78.1%	0.22
	Geneactiv Data Set	
K-Nearest Neighbors	93.2%	0.01
Decision Tree	92.4%	3.39
Neural Network	93.5%	152.17
Naive Bayes	90.0%	.17
	10 Motion Data Set	
K-Nearest Neighbors	89.5%	0
Decision Tree	83.2%	.55
Neural Network	92.1%	78.3
Naive Bayes	70.0%	.04

Table 4.3: The Accuracy and Training Time of Various Classifiers on Pilot Data Sets. Accuracy was similar across different classifiers. A K-Nearest Neighbors approach was selected to produce the results in this section because of its short training time.

been used. Observe that the performance of k-nearest neighbors is comparable to the performance of other classifiers, while at the same time taking the least time to train. The speed of training the k-nearest neighbors method made it the best choice.

While k-nearest neighbors was the best choice, a decision tree or neural network would also be good choices, and once trained have representations simple enough for classification to occur on the sensor platform.

4.3.3 Sampling Rate

A sampling rate of 125Hz is high enough to correctly record the hand hygiene signal, but correct classification can be performed using a lower sampling rate. Reducing the sampling rate is desirable because it uses less memory and enables data processing between samples, making feature extraction and data compression possible on the sensor. Literature on the minimum sampling rate for accurate recognition of hand motions is sparse. Figure 4.4 shows classification performance after data sets were resampled at lower sampling rates.

The quality of classification declines with the sampling rate, but it is not until the sampling rate falls below 60 Hz that the classification accuracy is too low for effective prediction. This drop off occurs in all data sets, although it is more severe in the Accelerometer and Orientation data set than it is in the other sets, probably due to

the greater number of closely related classes in this set.

The sampling rate could be adjusted to improve classification accuracy for different tasks. For instance, if a system first determined whether hand hygiene was occurring (i.e., Geneactiv data set) and then assessed whether the proper technique was being used (i.e., Accelerometer and Orientation data set) the sampling rate could be set around 60 Hz for the first task and 100 Hz for the second in order to provide results with at least 80% accuracy.

4.3.4 Window Length

The length of the window of data used to calculate the features has an effect on system accuracy and responsiveness. Balancing these two goals is crucial when identifying hand hygiene technique, as each individual hand hygiene motion will only be performed for a short time during a hand hygiene session. Figure 4.5 shows classification accuracy using different window lengths across all data sets.

The classification behavior is different depending upon the data set. For the 116 HCW and Geneactiv data sets, performance improves with window length but quickly levels off after the windows reach .5 seconds in length. In comparison increasing the window length reduces the classification accuracy in the Accelerometer and Orientation data set. As the window length increases above one second performance on the Accelerometer and Orientation data set drops noticeably.

There are many possible reasons for this difference. One could be that the motions are more complex and therefore more difficult to classify. Another could be differences between the two systems. For instance, only 8 seconds of data on the motion could be recorded in the Accelerometer and Orientation data set because both accelerometer and orientation readings were being recorded in limited memory. Another possibility is that the readings were coming from a different sensor, as both the accelerometer and orientation values were sampled from the orientation sensor board.

To explore this more, we examine two other data sets that had been collected as a baseline–one using the original system that had been used for the 116 HCW data set collection (Original Accelerometer), and the other using only the orientation values (Original Orientation). During data collection participants first performed all 10 motions with the original system and then performed them with the new system with the added sensor board. In these two datasets the sensor can be sampled for the full 16 seconds as only one value was being recorded.

Figure 4.6 compares the Accelerometer and Orientation data set with these other

data sets. The lines do not all deteriorate similarly–as we would expect if the difference were due to using 10 motions instead of the small number in the 116 HCW or Geneactiv data sets. The Original Orientation Data Set deteriorates in a similar manner to the other data sets collected using the same sensor board, so the difference cannot be due to the shorter sampling time. The behavior of the Original Accelerometer Dataset is similar to the behavior of the 116 HCW data set in Figure 4.5.

A limitation of this paper is that the cause of this difference could not be completely narrowed down using the existing data. There remain several possibilities, including differences in the system and differences in collection methods/motions as the Original Accelerometer data set collection preceded collection of the Accelerometer and Orientation data set samples, so participants could have changed their motions due to fatigue.

Determination of whether hand hygiene is occurring (i.e., Geneactiv data set) can be done using a longer window of 2 seconds. This opens up the possibility of additional data compression by calculating features from 2 second windows of data when the task is determining whether hand hygiene is occurring.

4.3.5 Unknown Subject

In a deployment, this system would need to recognize hand hygiene duration and technique in previously unseen subjects. In order to simulate a deployment scenario the machine was trained using all subjects except for one who was held out. Sliding windows were used in order to train the machine on all possible shifts of the training set. Then the held out subject was used as the test subject. Figure 4.7 shows a Cumulative Distribution Function (CDF) of the results on the 116 HCW set.

Figure 4.7 shows that most of the error is due to a small number of subjects. 91.3% of subjects have a classification error lower than 27%. No subject had a classification error greater than 60%, which is still better than chance. This suggests that our model generalizes to correctly classify unknown subjects.

4.3.6 One Wrist vs. Both Wrists

The current system requires the healthcare worker to wear the sensor devices on each wrist. This is different from commercial systems [15] which monitor only one wrist. To explore whether monitoring both wrists increases our classification accuracy, the data sets were resampled, this time creating three data sets: one containing only data gathered on the left wrist, a second containing only data gathered on the right

wrist, and a third containing data from both the left and right wrists. The results of classification on these data sets can be seen in Figure 4.9.

Figure 4.9 shows that in all data sets using data from both wrists is better than using only one wrist. This is consistent with results from other papers [42, 43] which suggest that using multiple sensing locations improves classification results. This difference becomes more pronounced as the classification tasks become more difficult. In particular performance on the 10 Motion Orientation data set drops precipitously, changing from 86.7% when using both wrists to 75.2% and 68.7% using the left and right wrists respectively.

It is interesting to note that classification accuracy using the right wrist is consistently lower than classification accuracy using the left wrist. However, the motions used in this study were theoretically symmetric for both hands. This difference may be due to hand dominance causing difference in the performance of hand hygiene between the two hands. Unfortunately efforts to confirm this were inconclusive since performance on left handed and right handed participants could not be compared due to the small numbers of left handed participants (e.g., only 7 participants were left handed in the 116 HCW data set). An open question is whether there is a transformation which would make readings from left handed subjects like those from right handed subjects, which would remove the need to create a special classifier for left handed people.

4.3.7 Sensor Fusion

In all previous experiments the combination of the accelerometer and orientation sensor produces better accuracy than only using the accelerometer or orientation sensor alone. Recall from Figure 4.2 that overall performance alone may not tell the entire story. In this experiment we compare using just the accelerometer, just the orientation sensor, and using both under the default parameters of the system. Figure 4.10 shows the resulting classification accuracy for each motion.

Figure 4.10 shows that the increase in accuracy is not due to increasing the classification accuracy of one class at the expense of all others. In fact using both the accelerometer and orientation sensor is better for all motions except for the left and right fingertip scrubs, where using only the orientation sensor is better, and the palm rub, where using only the accelerometer is better. This is to be expected as these motions in particular play to those sensor's strengths. The fingertip scrub motions consist of one wrist twisting around 180 degrees while the other remains stationary–

Classified As			
Wild HH	R & L FS	Walking	True Class
86	12	2	Wild HH
10	**89**	1	R & L FS
2	2	**96**	Walking

Table 4.4: Confusion Matrix of 116 HCW Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. Numbers given as percentage of true class. Overall accuracy is 90.1%.

Classified As		
Wild HH	Non-HH	True Class
93	7	Wild HH
7	**93**	Non-HH

Table 4.5: Confusion Matrix of Geneactiv Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. Numbers given as percentage of true class. Overall accuracy is 93.2%.

something which the orientation sensor excels at detecting so it is unlikely that adding the accelerometer measurement will increase performance. Similarly the palm rub is detected almost solely by the accelerometer, since the orientation of the wrists is similar to that of many other movements and does not change throughout, so it is unlikely that adding the orientation measurements will improve performance.

4.3.8 Duration Estimation

A goal of this experiment was to accurately estimate the duration of healthcare worker hand hygiene. In order to test this, the 116 HCW data set was used. In this experiment the training set consisted of the wild wash and the walking movements. The test set consisted of the mixed wild and walking movement. To determine duration of hand hygiene the test movement was split up into windows of .5 seconds and each window was classified as either wild washing or walking. When two windows in a

row had been classified as walking the hand hygiene event was marked as finished at the start of the first walking window. Samples where there was no observer marking (the hand hygiene duration was longer than the 16 second sampling period, accounting for a large number of samples) or an observer marking and no machine marking (usually when hand hygiene ends close to the end of the sampling period) were eliminated, leaving 38 samples with both machine and observer markings. The machine detected duration of hand hygiene was compared to the observer marked duration of hand hygiene, and a difference calculated in seconds. A histogram of those differences appears in Figure 4.11. While the classifier does not accurately classify every window of data, the misclassified windows are infrequent enough that duration can be accurately obtained.

In Figure 4.11 we can see that the differences between the machine prediction and the observer's prediction are normally distributed about a mean of -.78 seconds. This is consistent with a delay due to human reaction time. Overall the duration estimates provided by the system are close to those provided by the observer, being between -2.1 and 1.1 seconds from the observer's marking.

The system estimated duration of hand hygiene ranged from 3.3 to 16 seconds, with a mean of 9.7 seconds; the observer estimated duration of hand hygiene ranged from 4.1 to 16.8 seconds, with a mean of 11.1 seconds. Recall that the WHO recommends scrubbing with alcohol-based hand rub for 20-30 seconds, and that duration is a proxy for coverage of the hand with alcohol-based hand rub. These results show that many of the healthcare workers in the data set did not reach the minimum recommended duration even when an observer was standing next to them recording the duration of their scrubbing, a factor which would normally increase compliance. Every single one of these hand hygiene events would be judged equally as being "in compliance" using the normal hospital hand hygiene monitoring metric of compliance rates. These results show the importance of monitoring the duration of hand hygiene.

4.3.9 Technique Classification

While the overall classification accuracy is a helpful metric it is important to also examine the performance on individual motions. Table 4.6 contains a confusion matrix that shows how each motion in the Accelerometer and Orientation data set was classified by the system. No single motion was classified poorly, with the lowest accuracy being 84% in the Fingertip Scrub motions.

In Table 4.6 we can see that in fact most errors are caused by confusion between

Classified As										
PR	L TS	R TS	L KT	R KT	L FS	R FS	R BH	L BH	IF	True Class
93	0	0	0	0	0	1	0	1	5	PR
0	**87**	8	0	0	0	1	1	4	0	LTS
1	7	**88**	0	0	0	1	1	2	0	RTS
0	2	1	**93**	0	1	0	1	3	0	LKT
0	0	0	0	**98**	0	1	0	1	0	RKT
2	2	1	2	0	**84**	4	3	2	0	LFS
2	1	1	0	3	5	**84**	1	3	0	RFS
1	0	0	0	0	1	0	**91**	7	0	RBH
1	5	1	1	0	1	1	6	**84**	0	LBH
4	0	0	0	0	1	0	0	1	**94**	IF

Table 4.6: Confusion Matrix of Ten Motion Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. Numbers given as percentage of true class. Overall accuracy is 89.6%. Percentages may not sum to 100 due to rounding error. Motions are abbreviated using the first letters of the motion name to conserve space. R or L indicates that the right or left hand is the one being cleaned or the one on top of the other. See Table 4.1 for the full list of motions.

the left and right handed versions of the same motion. For example, 7% of Right Thumb Scrub samples were classified as Left Thumb Scrub samples. Some exceptions occur when the movements produce similar wrist motions. For example, the Interlocking Fingers and Palm Rub motions are confused with each other because they have the same wrist movement if the participant does not coordinate both hands with each other in the Interlocking Fingers motion or the participant moves both hands in the same direction at the same time during the Palm Rub.

4.4 Conclusion

Hand hygiene duration and technique can be accurately recognized using data from wrist-worn commodity sensors and machine learning techniques. Reported classification accuracies range from 89.6% for a data set consisting of 10 students performing 10 motions from the WHO hand hygiene guidelines to 90.1% for a data set consisting of 116 health care workers to 93.2% for a data set consisting of 30 health care workers performing hand hygiene and confounder motions. These accuracies are consistent despite changes in motion type classified, sensors used, and variation in technique across subjects.

Experimental results were also presented which reinforce the importance of choosing sampling rate and window size carefully. In many cases reducing the sampling rate or window size may be possible without seriously impacting the overall quality of classification. Time-based features were found to be more useful than frequency based features for this application due to the natural differences in scrubbing frequencies between participants and rotation involved during washing. The effects of sensor fusion from multiple sensor locations (one wrist vs. both) and multiple sensor modes (accelerometer vs. gyroscope) was experimentally determined, and in both cases adding more sensors/sensing modes improved results.

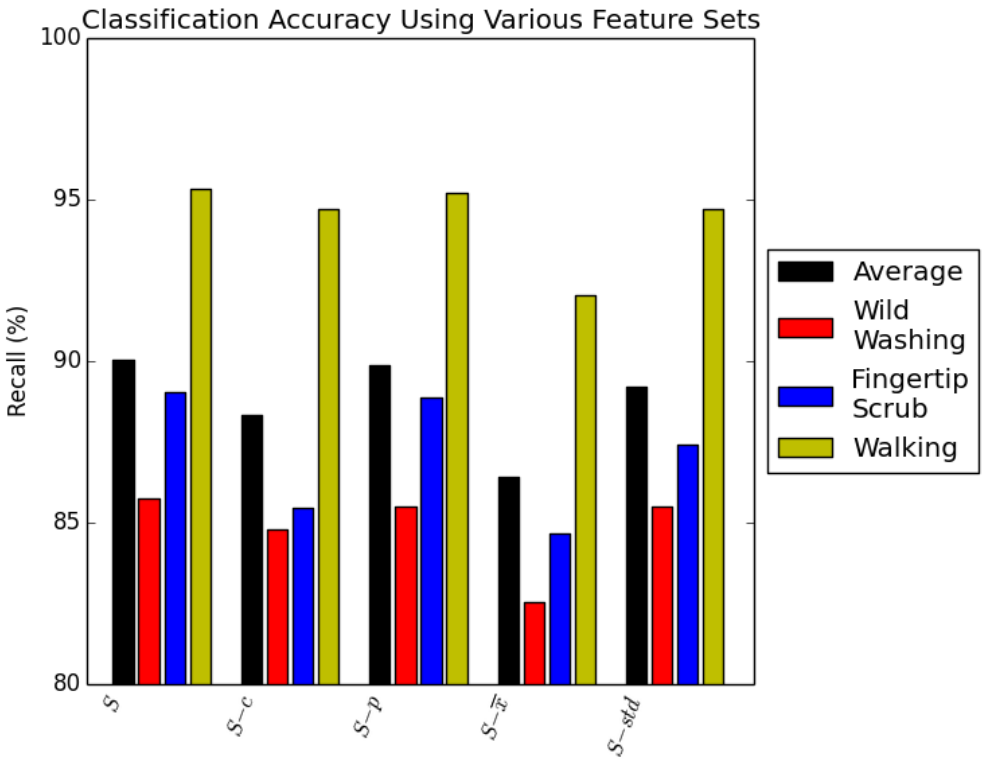

Figure 4.3: Classification Accuracy Using Subsets of Final Feature Set. When any feature is removed from the feature set classification performance declines. Every feature in the final set is useful.

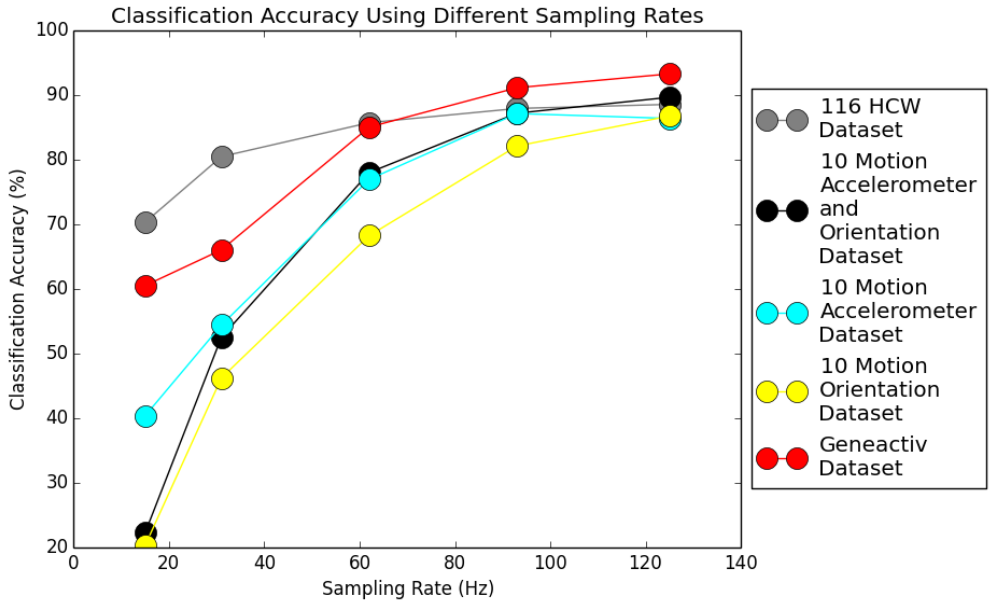

Figure 4.4: Classification Accuracy on Pilot Data Sets Using Different Sampling Rates. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Window size is .5 seconds, and windows do not overlap. Increasing the sampling rate improves classification accuracy, but shows diminishing returns as the sampling rate grows faster.

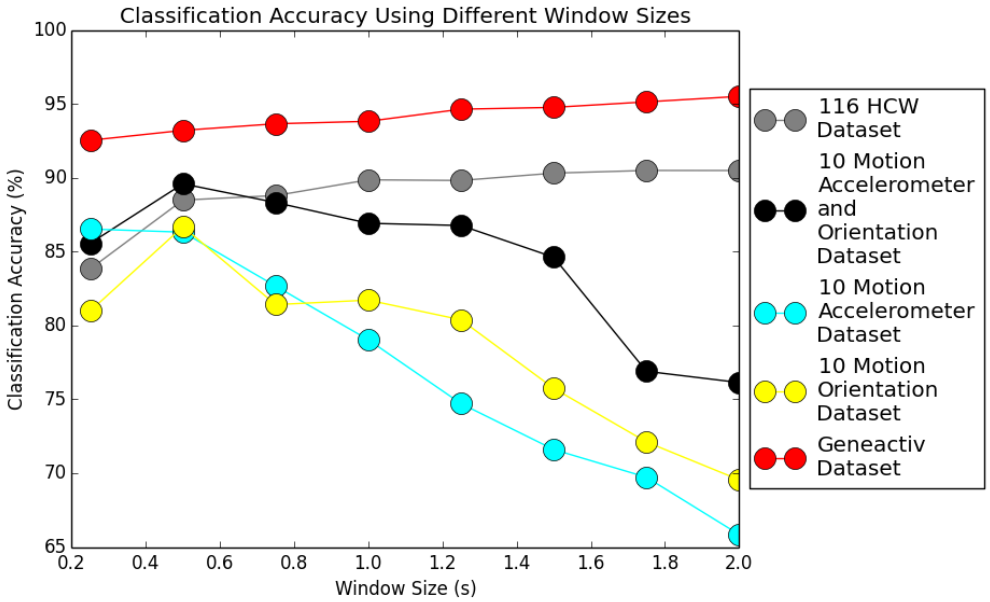

Figure 4.5: Classification Accuracy on Pilot Data Sets Using Different Window Sizes. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows do not overlap. As the window size increases the performance of the classifiers initially improves, then levels off or declines (depending on the data set).

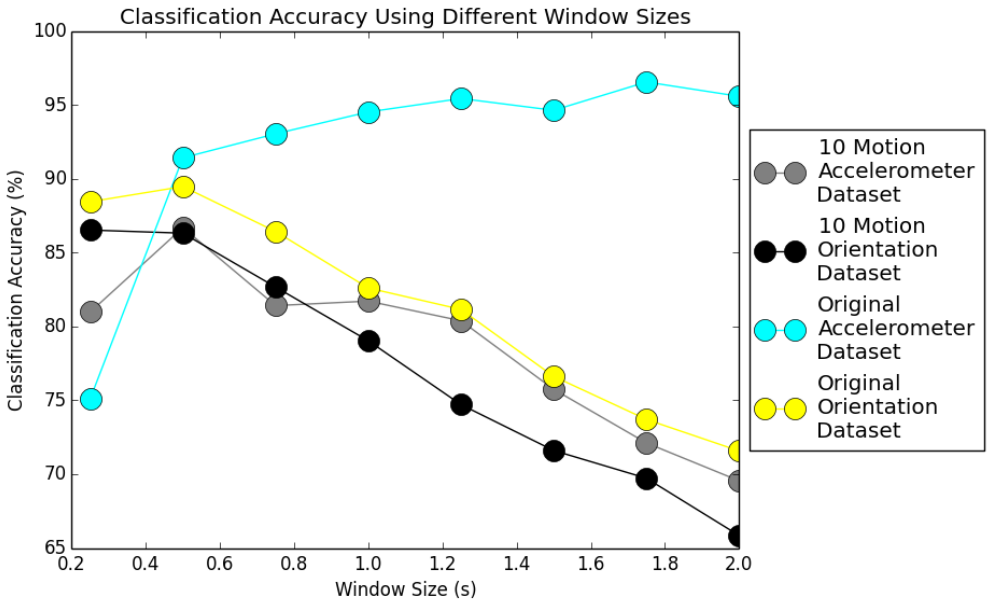

Figure 4.6: Classification Accuracy on Pilot Data Sets Using Different Window Lengths: Comparing sensor types. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows do not overlap. The classification performance on all systems does not deteriorate similarly. This could be due to many causes, including differences in the system and differences in collection methods. The cause of this difference could not be completely determined using existing data.

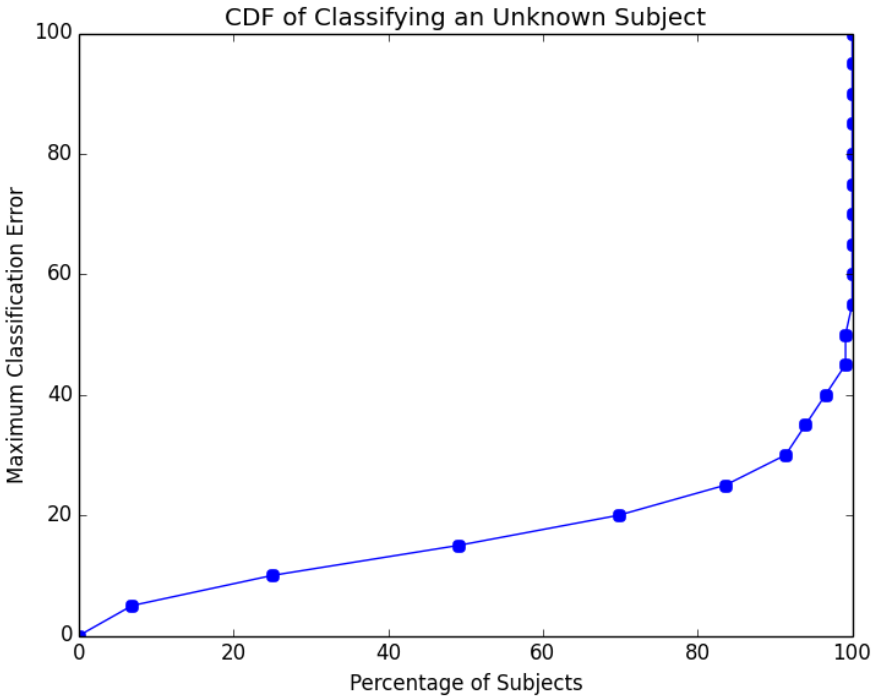

Figure 4.7: Cumulative Distribution Function of Classification Errors Using an Unknown Subject in the 116 HCW Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Sliding windows were used in the training set. The machine was trained on all subjects but one and tested on the held out subject. Error is 1-Accuracy. Most errors occur in a small number of subjects.

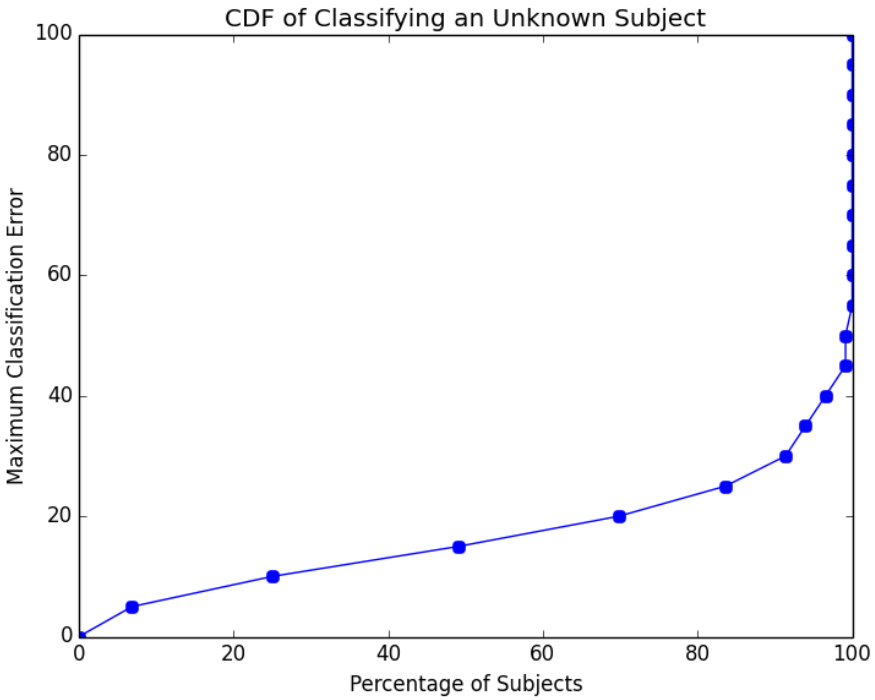

Figure 4.8: Accuracy Using an Unknown Subject in the Pilot Geneactiv Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. Sliding windows were used in the training set. The machine was trained on all subjects but one and tested on the held out subject. Most errors occur in a small number of subjects.

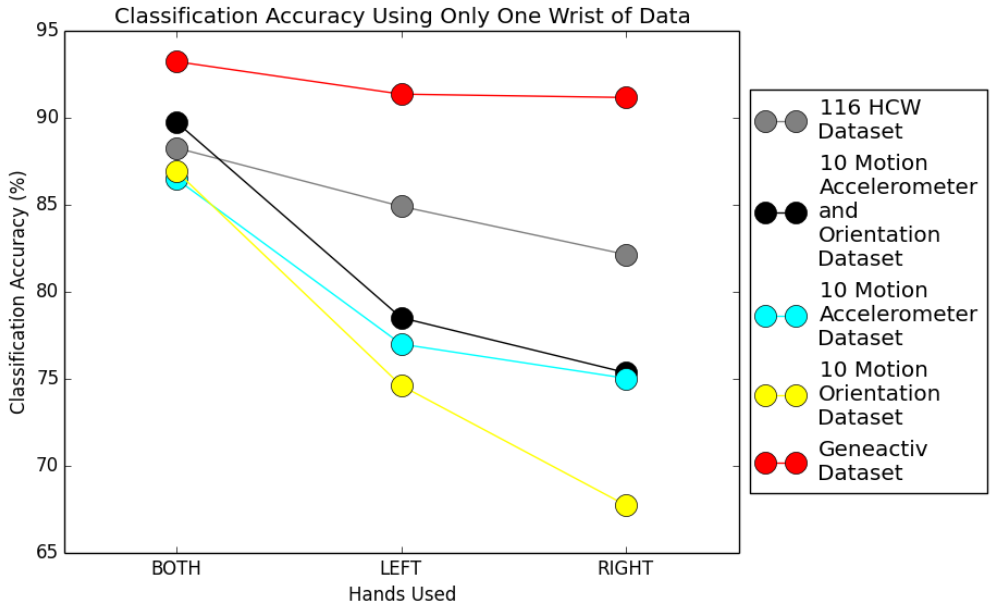

Figure 4.9: Classification Accuracy on Pilot Data Sets Using Different Hands. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. "BOTH" contains results using data from both wrists, while "LEFT" and "RIGHT" use the same features calculated using only data from one wrist. Using data from both wrists outperforms using data from only one wrist every time

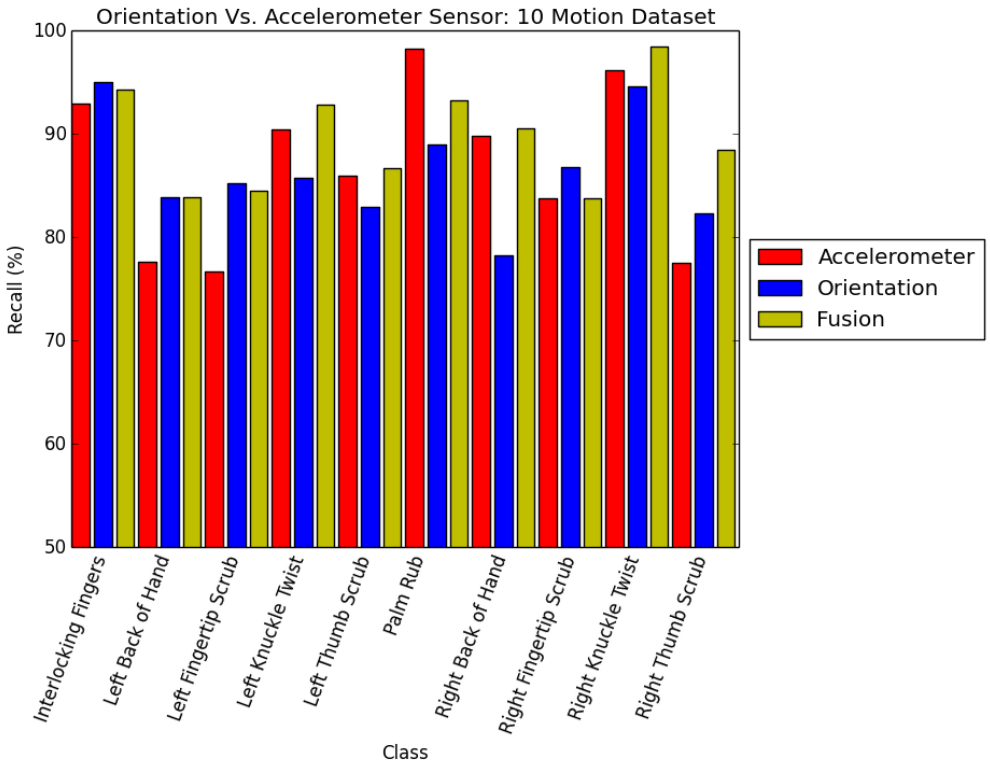

Figure 4.10: Classification Accuracy Using Different Sensor Types. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. The fusion of both sensor types outperforms using only one sensor for most motions. Those motions which favor one sensor over the fusion of sensors (e.g., Palm Rub or Fingertip Scrub) are either strongly rotational (favoring the orientation sensor–Fingertip Scrub) or consist of movement of the wrist along one plane (favoring the accelerometer–Palm Rub).

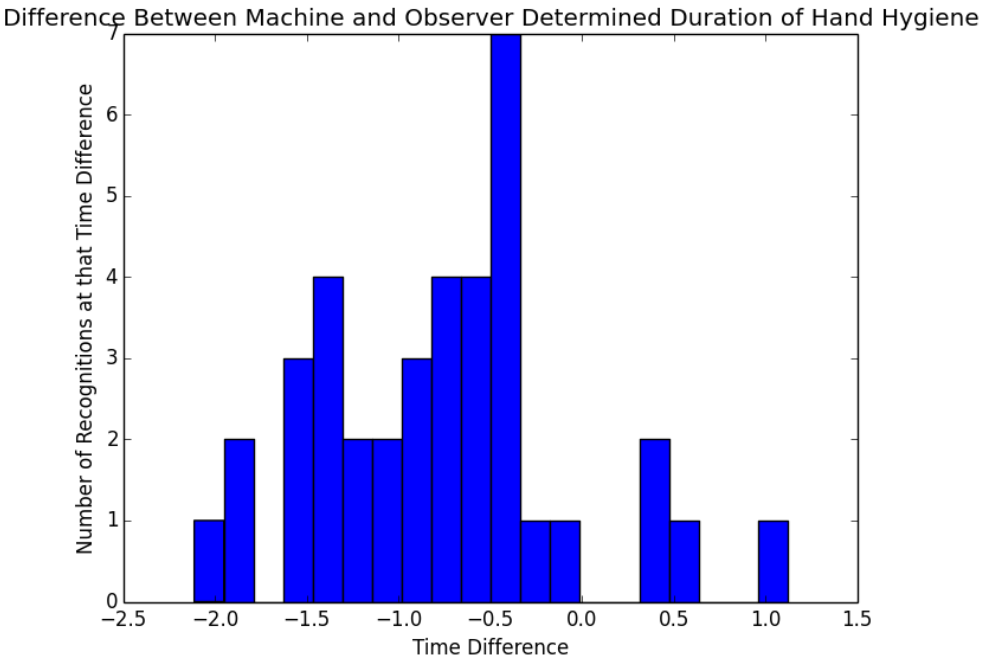

Figure 4.11: Difference Between Machine and Observer Hand Hygiene Durations in 116 HCW Data Set. Results shown are obtained using a K-Nearest Neighbors classifier with K=3. Sampling rate is 125 Hz. Windows are .5 seconds long. Hand hygiene was considered complete when two consecutive windows were classified as walking.

CHAPTER 5:

HAND HYGIENE RECOGNITION ON THE HOSPITAL FLOOR

Work on the pilot data sets showed promise, but movements in the pilot data sets were largely synthetically obtained. These movements were designed to mimic confounding motions that could occur during a healthcare worker's day, but it is unknown whether other motions occur in the hospital that would cause the system to perform poorly. In addition the amount of time spent performing hand hygiene is unknown, making it difficult to correctly construct an artificial test data set. In order to resolve these issues, it is essential to collect data consisting of an ordinary healthcare worker's routine on the hospital floor and note hand hygiene events when they occur.

5.1 Geneactiv Shadowing Data Set

For the "Geneactiv Shadowing" data set, data was collected from 22 healthcare workers using the Geneactiv wristbands described in Chapter 3. Healthcare workers at the University of Iowa Hospitals and Clinics wore the wristbands on each wrist for approximately an hour during their work day. During that hour they were shadowed by observers who marked when the instrumented healthcare worker performed hand hygiene using the iPhone app described in Section 3.3.

Two separate observers were involved in collecting this data set. In order to ensure that both observers were marking events similarly some participants were followed by both observers who marked events at the same time.

For clarity in this section we will refer to the previous, synthetically generated Geneactiv data set as "Geneactiv Pilot Data Set".

Observers made their best effort to observe the health care worker at all times, but close observation was not always possible due to many factors in the busy hospital environment (e.g., if a team of doctors was in the patient room the room may be too crowded and observation would have to occur outside the door). Due to those factors we may not be as sure of the accuracy of the designation of periods of time as hand

hygiene or non-hand hygiene in comparison to earlier, synthetically generated data sets.

5.2 Methods

The data set was processed and marked according to the data provided by the observers. In the case where a participant was followed by two observers, the reading in question was marked as containing hand hygiene as long as at least one observer marked it as a hand hygiene event.

Every second of data was marked with one of three possible observer markings: hand hygiene, non hand hygiene, or unknown (in the case that the observer was unable to observe the healthcare worker).

Three different training sets were used. In the first, data from the portion of the Geneactiv Shadowing data set where participants were followed by both observers was used as the training set. A second training set which balanced the prevalence of hand hygiene and non-hand hygiene was created from this first training set by resampling with replacement, a technique where the created data set consists of instances randomly sampled from the original data set with a bias for uniform class and without eliminating an instance from the sampling pool when it is sampled. In the second, data from the Geneactiv Pilot data set was used to train the system. The test set in both cases consisted of data from the participants in the Geneactiv Shadowing data set that were not followed by both observers.

Hand hygiene events consist of k seconds and therefore 2k different classifications. These classifications can indicate that a period of time belongs to a hand hygiene event even if the classifier has erroneously marked it as non-hand hygiene. To reflect this, a second stage of processing was done on the classification outputs to produce the final set of duration results. In this stage, consecutive hand hygiene detections were considered to be part of the same hand hygiene event if they were less than 10 seconds apart. Hand hygiene event detections of fewer than 3 seconds in duration were discarded. Hand hygiene event detections of greater than 60 seconds were cut off at 60 seconds.

In the results section, an observed hand hygiene event was considered as a matching candidate for a machine hand hygiene event detection if they occurred within one minute of each other. Once all candidates were found, the closest matching hand hygiene event candidate was considered the ”match” to the hand hygiene detection.

5.3 Results

5.3.1 Characteristics of the Data Set

One of the goals of collecting the Geneactiv Shadowing data set was to better understand the characteristics of hand hygiene on the hospital floor. The data set revealed some interesting aspects of hand hygiene that were not previously known.

Inter-rater Agreement

Data from the participants that both observers observed was examined and an inter-rater agreement of 95% was obtained.

Observation Time

Effort was made to observe each healthcare worker for an hour, but it was not possible to observe every healthcare worker for the same amount of time due to constraints imposed by the clinical setting. For example, a patient may request that the observer leave the room or the healthcare worker may need to leave the observer's field of view (e.g., when using the bathroom). In such cases, while an hour of time was allotted for observation there may not be an hour of data.

In Figure 5.1 we can see the distribution of time spent observing each subject. Most subjects were observed for approximately an hour, with a low of observation for 28 minutes and a high of observation for an hour and 20 minutes.

Prevalence of Hand Hygiene

Overall, hand hygiene makes up approximately 3.5% of the data set. This equates to healthcare workers washing their hands for 2.1 minutes for every hour. However, this is a mean value.

In order to better understand the range of hand hygiene prevalence between health care workers, an estimate of hourly hand hygiene rate was computed by calculating hand hygiene as a percent of all observations. This percentage was then multiplied by an hour to determine how many minutes of hand hygiene each healthcare worker might perform. Figure 5.2 shows that different healthcare workers can have very different rates of hand hygiene, ranging between 5.4 seconds for the healthcare worker who performed the least hand hygiene in the observed hour to 4.5 minutes for the healthcare worker who performed the most hand hygiene in the observed hour.

Distribution of Observer Marked Duration

One might theorize that healthcare workers who perform hand hygiene frequently would have greater consistency in technique due to their more frequent practice. In Figure 5.3 we can see that this is not true of all healthcare workers. Subjects who performed hand hygiene more frequently were not more likely to wash with more consistent duration. Instead it appears that generally there is a trend of hand hygiene of less than 20 seconds in duration, with a few outliers of longer hand hygiene. This is a concern because the World Health Organization recommends that hand hygiene using an alcohol-based hand rub be 20-30 seconds in duration, and most healthcare workers were not washing for that long in our observations.

Separation of Hand Hygiene Events

The World Health Organization has issued a hand hygiene recommendation known as the "Five Moments of Hand Hygiene" [1], shown in Figure 1.1. If healthcare workers perform hand hygiene according to these guidelines, we should expect that many hand hygiene events would occur within close proximity of each other. As an example scenario, imagine a healthcare worker who enters a patient room in order to physically interact with the patient in some way (e.g., take a temperature, perform an examination, or even just chat with them and shake their hand). That healthcare worker would perform hand hygiene four separate times–once before entering the room, once before touching the patient, once after touching the patient, and once after leaving the room. That healthcare worker would perform hand hygiene again before entering the room of another patient.

In Figure 5.4 we can see the amount of time that passed between consecutive hand hygiene events. Most events occurred within five minutes of a previous event, showing that hand hygiene events do occur within close proximity of one another as the World Health Organization guidelines would suggest.

While the World Health Organization guidelines are the recommended way to perform hand hygiene, they are difficult to enforce because they would require monitoring of healthcare workers in patient rooms. This has led to most hospitals monitoring only moments one and five–when the healthcare worker enters or leaves the patient room–because they can be monitored from the hallway. After examining Figure 5.4 more closely, note two smaller peaks in hand hygiene frequency, one at around 10 minutes and another at around 15. These suggest hand hygiene events that occurred before and after entering a patient room.

We can not know for sure where hand hygiene events occurred, but an effort was made to estimate which hand hygiene events were due to "Wash In, Wash Out" hand hygiene. A "Wash In, Wash Out" hand hygiene pattern was considered to be one that is consistent with a healthcare worker leaving patient room A, entering patient room B for some time, and then leaving again for patient room C. This would be characterized by two closely spaced hand hygiene events (one when leaving room A, and another when entering room B) followed by a long time between hand hygiene events (due to dwell time in room B) subsequently followed by another closely spaced pair of hand hygiene events (one when leaving room B and another when entering room C). In our analysis, two hand hygiene events are defined as being closely spaced when they are less than 5 minutes apart, and a long time is defined as being longer than 10 minutes. The results can be seen in Figure 5.5.

This method for finding events due to "Wash In, Wash Out" hand hygiene is very crude. For instance, it would not correctly categorize an event where a healthcare worker only saw one patient since there would only be one pair of hand hygiene events separated by a long time. It would also not correctly categorize a healthcare worker who only scrubs out or scrubs in, which would be characterized by continually having hand hygiene events with long time separation. In addition to those two common scenarios that would cause undercounting of "Wash In, Wash Out" events, there are some scenarios that would cause overcounting. For instance it could miscategorize an event as "Wash In, Wash Out" if the long time separation was due to some other factor such as spending time at the nurse's station or on the computer. A hand hygiene event could also be miscategorized if the healthcare worker needed to do a lengthy non-aseptic task on the patient that involved touching the patient but no risk of exposure to bodily fluids (then a healthcare worker in compliance with the WHO's five moments of hand hygiene would perform hand hygiene upon entering the patient room, before touching the patient, perform the procedure for over 10 minutes, and then perform hand hygiene after touching the patient and after leaving the patient room). Overall the risk of undercounting hand hygiene events is greater than the risk of overcounting hand hygiene events using this method, so the red bars seen in Figure 5.5 can be understood as a conservative estimate of the amount of hand hygiene due to "Wash In, Wash Out" hand hygiene adherence.

In Figure 5.5 we can see that hand hygiene events due to "Wash In, Wash Out" hand hygiene patterns explain a large portion (80%) of the bump at 15 minutes that was seen in 5.4. In addition these patterns explain approximately 15% of the hand

hygiene events that are less than five minutes apart. Overall these patterns account for 19.5% of hand hygiene events in the hospital. This "Wash In, Wash Out" behavior is not in compliance with the WHO guidelines for hand hygiene, so this percentage is not simply a reflection of the prevalence of moments 1 and 5 in comparison to other WHO moments. This shows that solely employing "Wash In, Wash Out" hand hygiene monitoring incentivizes "Wash In, Wash Out" hand hygiene and that healthcare workers will therefore neglect other, unmonitored hand hygiene moments that are also important to preventing the spread of disease in the hospital.

5.3.2 Detection Accuracy

Detection Accuracy Using Different Training Data and Classification Methods

Figure 5.6 shows the classification accuracy using different training sets and classifiers. All combinations of training set and classifier seem to produce high accuracy. The results using a training set generated from the Geneactiv Shadowing data set using replacement to obtain an equal mix of hand hygiene and non-hand hygiene motions performs the worst. This is probably due to a mix of two factors, the first being the low prevalence of hand hygiene in the data set and the second being a low certainty in the existence of hand hygiene as compared to the synthetically generated Geneactiv Pilot data set.

The general accuracy of the methods seems good, but it is worthwhile to remember that only 3% of the data set consists of hand hygiene, and therefore guessing the majority class of non-hand hygiene should produce an accuracy of 97%. As we are interested in our ability to find hand hygiene instances, it is useful to examine the positive predictive value as well. In Figure 5.7 we can see the positive predictive value for hand hygiene.

These figures are for classification done on .5 seconds of data, and do not include the second processing step which will be discussed in the next section.

Hand Hygiene Event Detection

Classification accuracy is helpful, but it does not tell us whether hand hygiene events are correctly detected. In Figure 5.9 we can see how many hand hygiene events are detected by each classification method. While all methods detect almost all hand hygiene events, they do this at the expense of a low positive predictive value. In Figure 5.10 we can see that the extra processing step produces gains in the positive

Classifier	Average Undetected Event Duration (s)	Average Detected Event Duration (s)
Naive Bayes	10.8	18.2
K-Nearest Neighbors	10.4	17.9
Neural Network	10.11	18.5
Decision Tree	10.55	18.16

Table 5.1: Hand Hygiene Event Duration: Effect on Detection Probability. Longer hand hygiene events are more likely to be correctly detected. The World Health Organization recommends rubbing for 20-30 seconds, so the undetected events that average 13 seconds or less in length would all be under the recommended duration of hand hygiene.

	Detections		Missed HHE		False Detections		Duration Error(s)	
Classifier	Before	After	Before	After	Before	After	Before	After
Naive Bayes	452	131	3	11	226	36	14.4	11.7
KNN	713	216	2	7	456	108	14.0	11.1
Neural Network	533	150	3	13	305	57	14.8	10.6
Decision Tree	586	164	0	10	337	66	15.3	10.2

Table 5.2: Number of Detections Before and After Processing. The processing step brings down the number of hand hygiene event detections considerably. There are 85 hand hygiene events in the data set. The duration error is the average of the absolute value of the difference between the duration predicted by the machine and the observed duration of the matching hand hygiene event.

predictive value at the expense of detecting fewer hand hygiene events.

These figures are based on all hand hygiene events in the data set. However, many of those events are very short. If the hand hygiene events are restricted to those that are compliant with the World Health Organization recommended duration of longer than 20 seconds then 100% of hand hygiene events are detected with every method. As shown in Table 5.1, hand hygiene events with longer duration are more likely to be detected, and a large difference can be seen between the average duration of detected vs. undetected hand hygiene events. At this point it may seem that based on the trade-off between the number of hand hygiene events detected and the positive predictive value results there is no clear benefit to including the processing step. However, those numbers can be misleading due to the small number of hand hygiene events. To better understand the effects of the second processing step, see Table 5.2. As we can see in this table, many false detections are eliminated in the processing approach. Some detections of true hand hygiene events are also eliminated, (generally because the events are of short duration and therefore resemble false detections). After examining Table 5.2 one can understand why a drop in the number of detected hand hygiene events can be well worth a rise in the positive predictive value of the test.

	Before Processing		After Processing	
Classifier	Mean Error(s)	Median Error(s)	Mean Error(s)	Median Error(s)
Naive Bayes	14.4	10.0	11.7	7.5
KNN	14.0	9.5	11.1	8
Neural Network	14.8	10.3	10.6	7.0
Decision Tree	15.3	10.5	10.2	6.0

Table 5.3: Mean and Median Duration Error. The duration error is the average of the absolute value of the difference between the duration predicted by the machine and the observed duration of the matching hand hygiene event. The processing step reduces the duration error in all cases.

The processing step also improves the estimate of duration, as shown in Table 5.2. In the unprocessed form it is common to have many separate detections during each hand hygiene event, so there is a large error between the duration reported by the detection and the duration of the associated event. Combining the separate detections which are all associated with the same hand hygiene event creates a better estimate of the hand hygiene event duration.

5.3.3 Duration Estimation Accuracy

In Table 5.2 we observe that the mean duration error is approximately 10-11 seconds. However, means can be influenced by outliers. In fact, the median duration error ranges between 6 and 7 seconds, as shown in Table 5.3 In Figure 5.11 we can see the distribution of the difference between the machine estimate of duration and the observed duration. Most estimates of duration are close to the observed duration of the event, but there are several outliers which can be different from the observed duration by as much as a minute.

In Figure 5.12 we can see the difference between the machine estimate of the start of the hand hygiene and the true starting point of the associated event. Again most detections place the starting point close to the true starting point. There are some notable exceptions, some missing the true start point by well over a minute. Figure 5.13 exhibits similar trends. One difference is that the peak near 0 is not as pronounced, so it is more likely to err when predicting the end of hand hygiene compared to predicting the start of hand hygiene. Multiple factors could drive this difference, but one possibility is that it is easier for an observer to be sure about the start of a hand hygiene event (since a dispenser is being activated) than it is to be sure about the end of a hand hygiene event, as one must be sure that the healthcare worker is done washing and not simply pausing. Healthcare workers may also engage in various confounding activities at the end of a hand hygiene event–as an example,

some healthcare workers do not wash until their hands are dry. In that case they may wave their hands in the air to dry them, which could cause confusion in the system. Another possibility is that hand hygiene changes as the hands become drier toward the end of the hand hygiene event, and therefore the motion becomes more difficult to recognize correctly.

5.4 Conclusion

The system performs well on data taken in the field on the hospital floor. In addition to correctly detecting 100% of World Health Organization-compliant hand hygiene events, the duration of events is also accurately calculated.

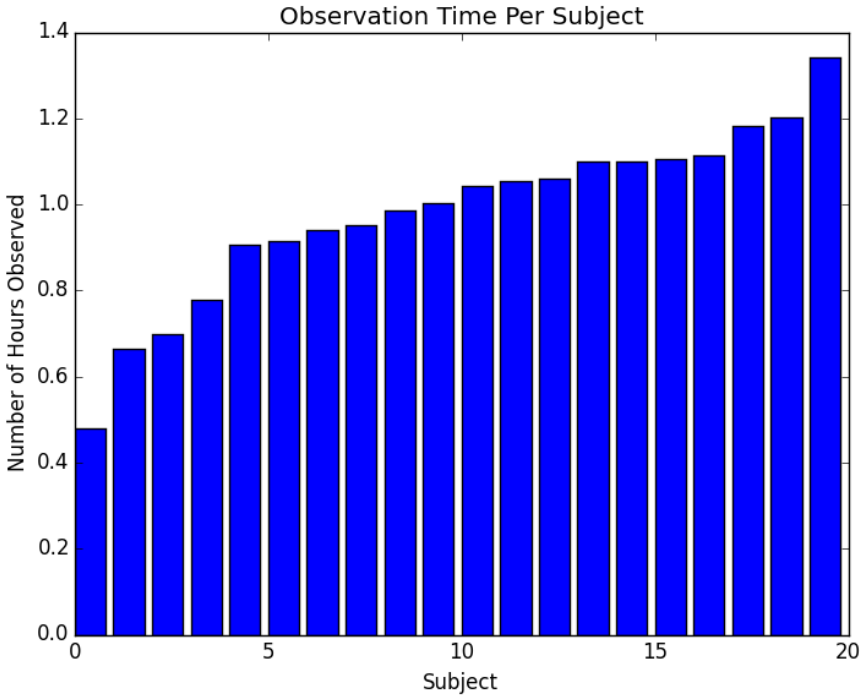

Figure 5.1: Amount of Time Each Healthcare Worker was Observed in Geneactiv Shadowing Data Set.

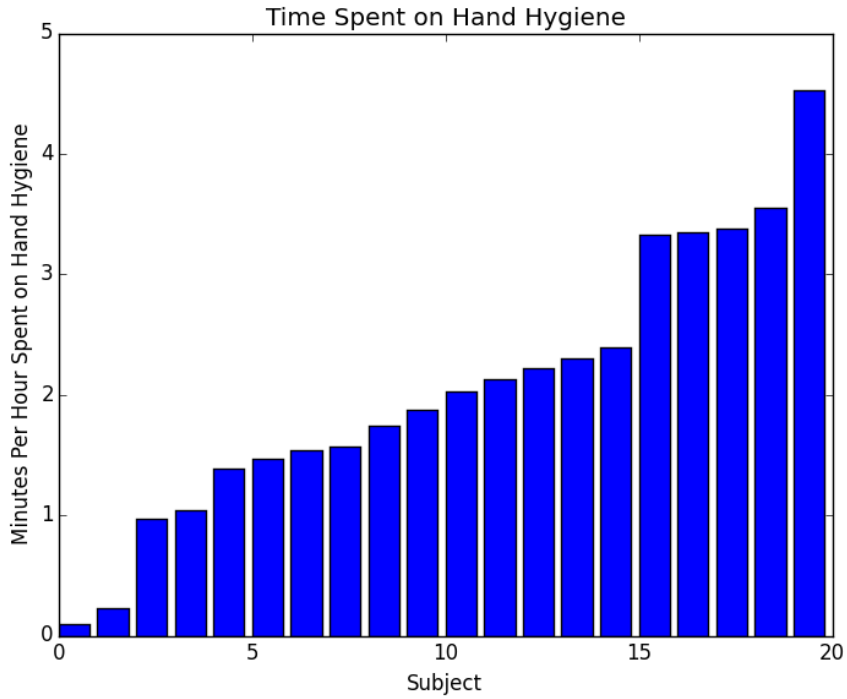

Figure 5.2: Minutes Each Healthcare Worker Spent per Hour on Hand Hygiene in Geneactiv Field Data Set.

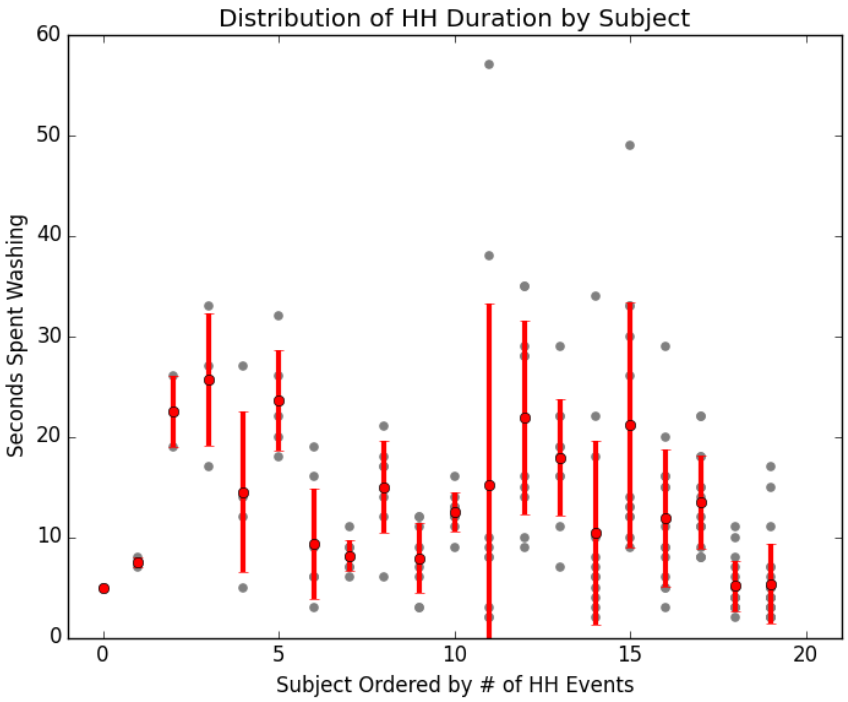

Figure 5.3: Distribution of Hand Hygiene Duration by Subject in Geneactiv Shadowing Data Set. Subjects are ordered by the number of observed hand hygiene events in the data set. The mean duration and standard deviation are shown by the red points and error bars. Healthcare workers who wash more frequently do not necessarily develop a routine and wash for the same amount of time every time.

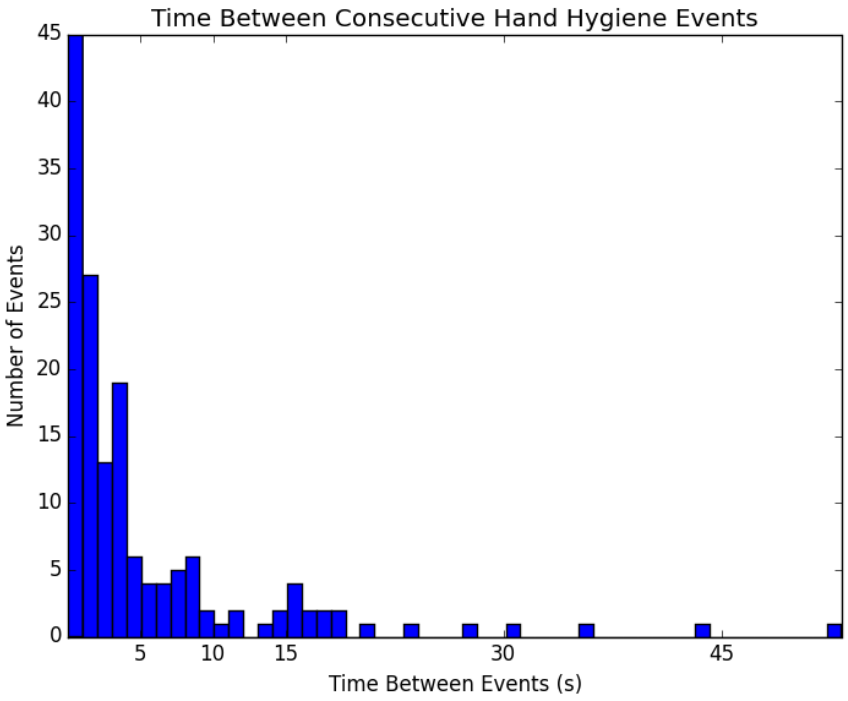

Figure 5.4: Separation of Consecutive Hand Hygiene Events in Geneactiv Shadowing Data Set. Bins are one minute in size. Note the small rises in frequency at roughly 10 and 15 minutes, which possibly suggest hand hygiene events upon entering and leaving patient rooms (the times hand hygiene is normally monitored).

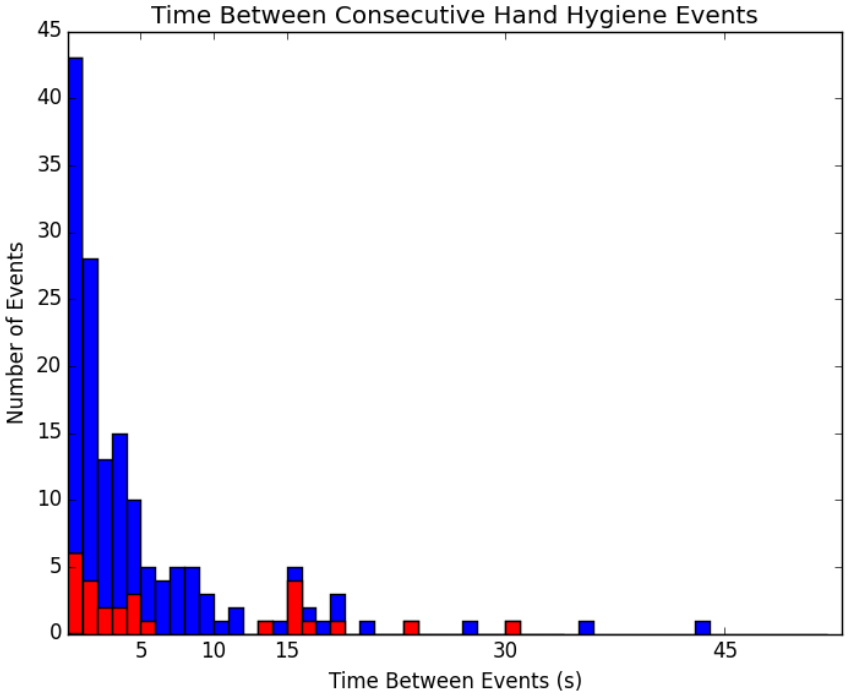

Figure 5.5: Hand Hygiene Events Due to "Wash In, Wash Out" Monitoring in Geneactiv Shadowing Data Set. Bins are one minute in size. The red portion of the bar indicates hand hygiene events which are possibly caused by healthcare workers performing hand hygiene upon patient room entry and exit. These events are a large portion of the "bump" observed at 15 minutes in Figure 5.4.

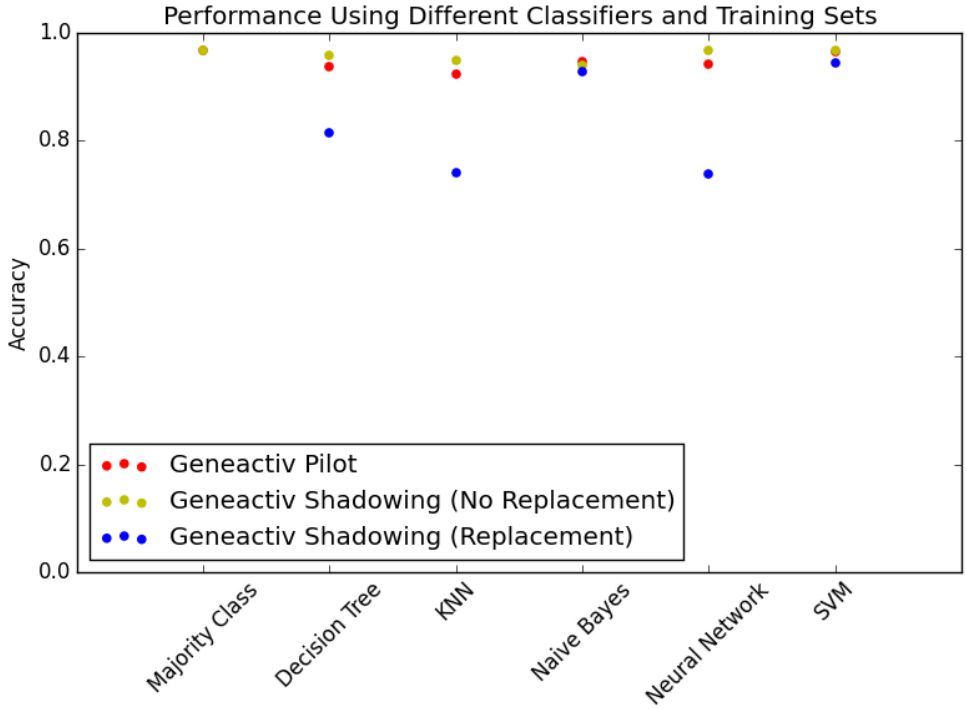

Figure 5.6: Different Training Sets and Classifier Performance on Geneactiv Shadowing Dataset. These results are from the classifications before the second processing step. The classifiers have better accuracy when trained using the Geneactiv Pilot data set and Geneactiv Shadowing (Replacement) data set.

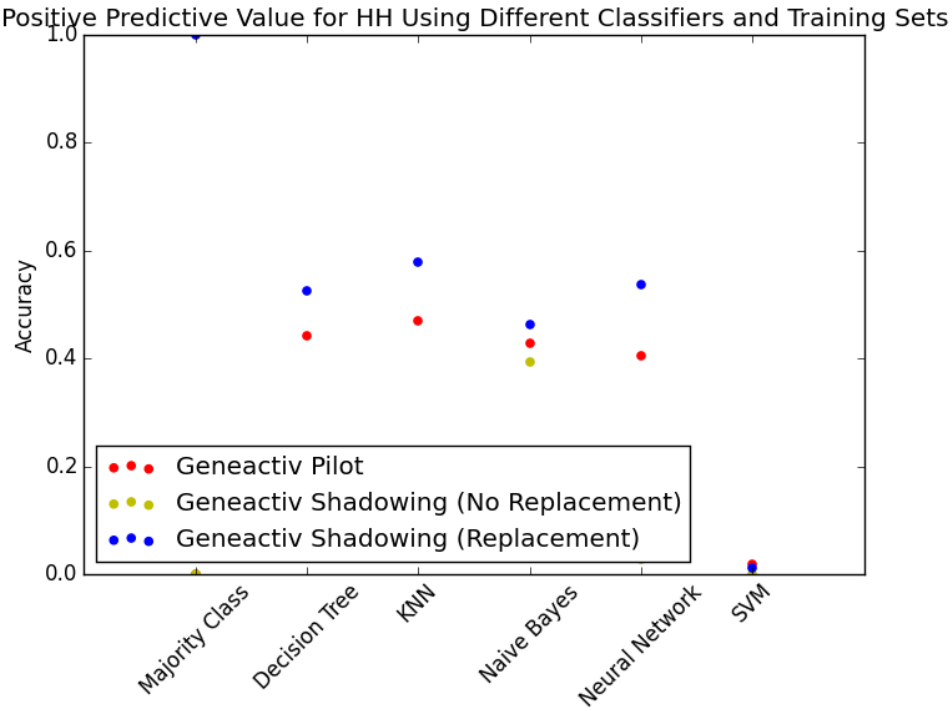

Figure 5.7: Different Training Sets and Classifier Effect on Positive Predictive Value for Hand Hygiene on Geneactiv Shadowing Dataset. These results are from the classifications before the second processing step. The classifiers have the best positive predictive value when trained using the Geneactiv Pilot data set. Because of these results and results from Figure 5.6 all future results are presented using the Geneactiv Pilot data set for training.

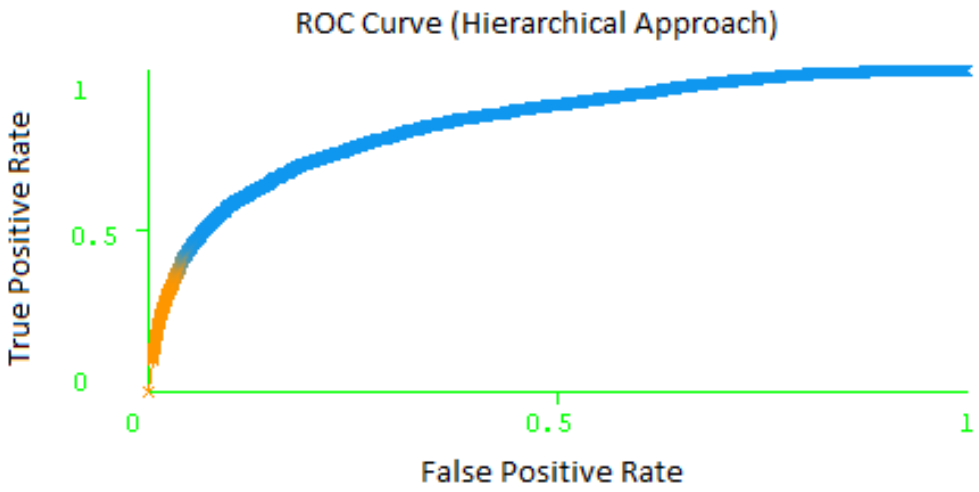

Figure 5.8: ROC Curve for Geneactiv Shadowing Dataset. This is the curve generated by the naive bayes classifier. These results are from the classifications before the second processing step. The x axis represents the false positive rate (the ratio of false positives to negative detections) and the y axis represents the true positive rate (a.k.a. sensitivity or recall). As the true positive rate increases the number of false positives also increases.

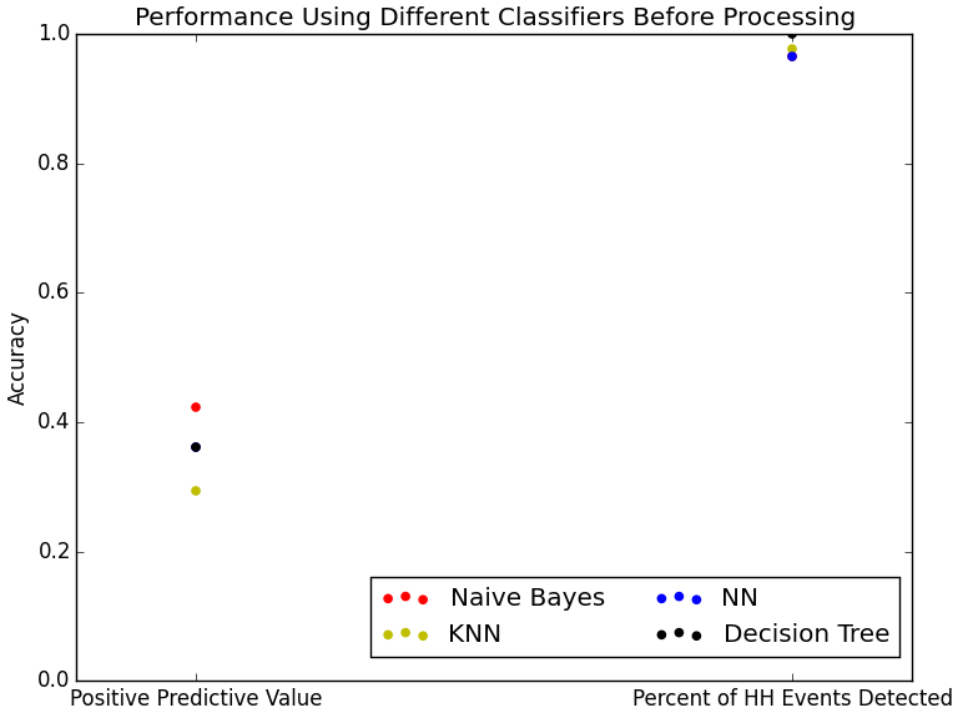

Figure 5.9: Multiple Classifiers Machine vs. Observer Detections Before Extra Processing. Most hand hygiene events are correctly detected. The positive predictive value is low because there are many spurious detections.

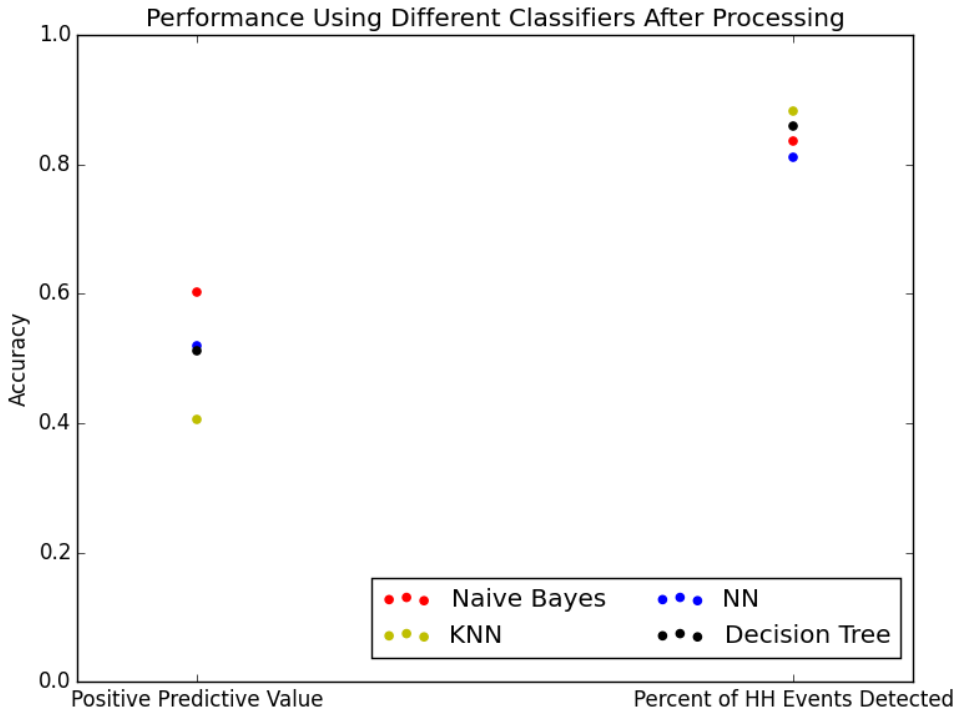

Figure 5.10: Multiple Classifiers Machine vs. Observer Detections After Extra Processing. The positive predictive value is higher than that seen in Figure 5.9 after the processing step as the number of spurious detections is decreased. This increase in positive predictive value is paired with a decrease in the percent of hand hygiene events detected as correct detections of hand hygiene events of short duration closely resemble spurious detections. Recall that the shortest hand hygiene event in the data set is only three seconds long, so it can be difficult to differentiate between true and false detections in hand hygiene events of such short duration.

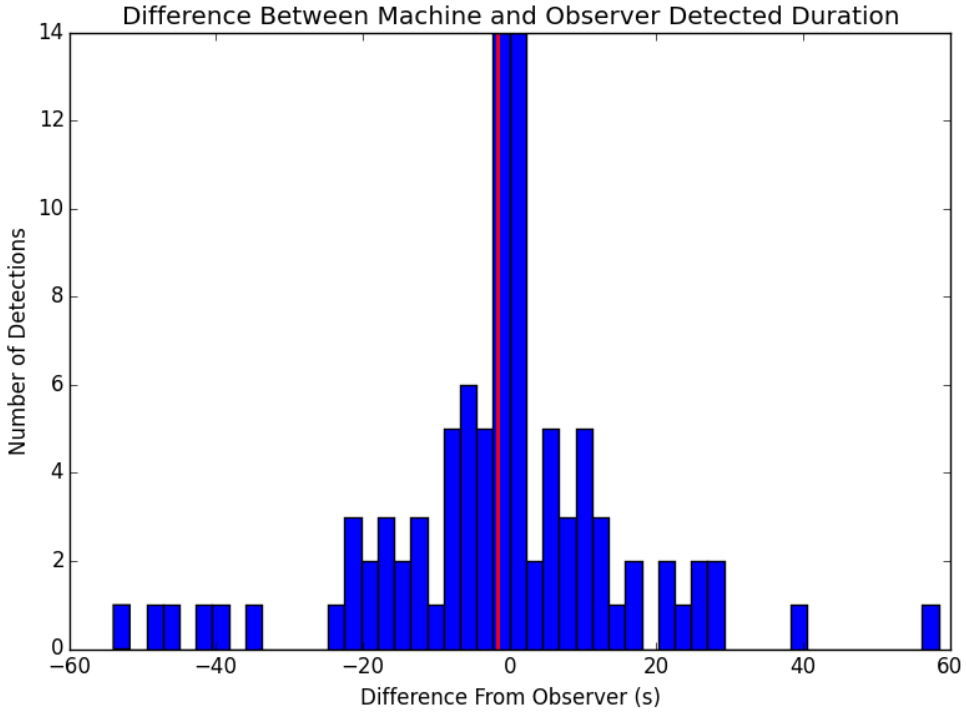

Figure 5.11: Difference Between Machine and Observer Duration. Results using predictions from Naive Bayes classifier. Red line indicates mean. The duration estimates provided by both the machine and the observer match closely, with a small tendency for machine estimates to be slightly shorter than observer estimates.

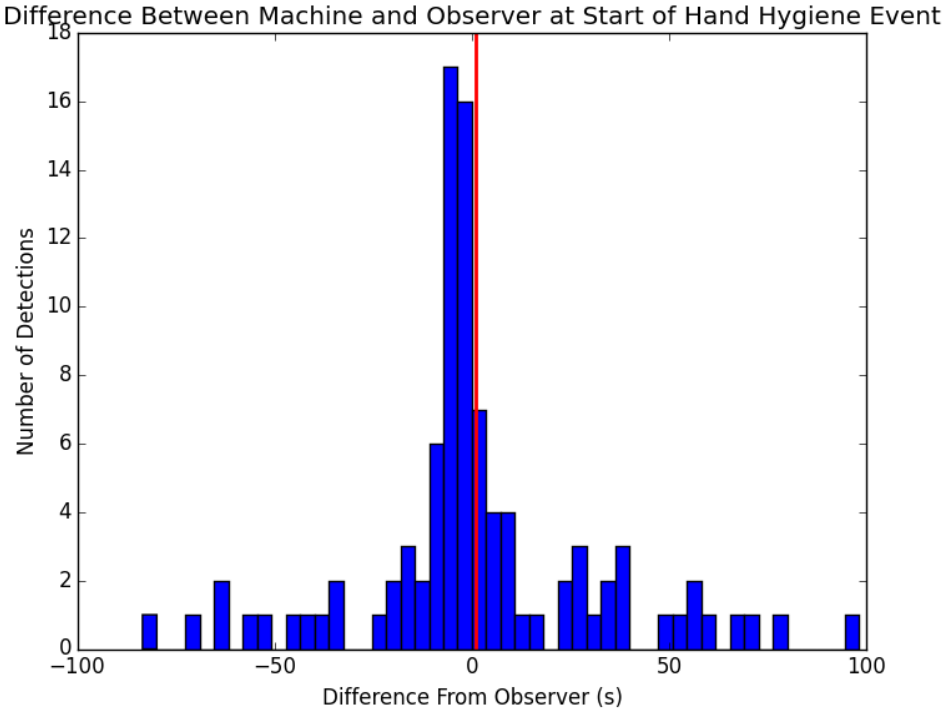

Figure 5.12: Difference Between Start of Machine and Human Observed Hand Hygiene Events. Results using predictions from Naive Bayes classifier. Red line indicates mean. The machine and observer marking of the start of a hand hygiene event match closely.

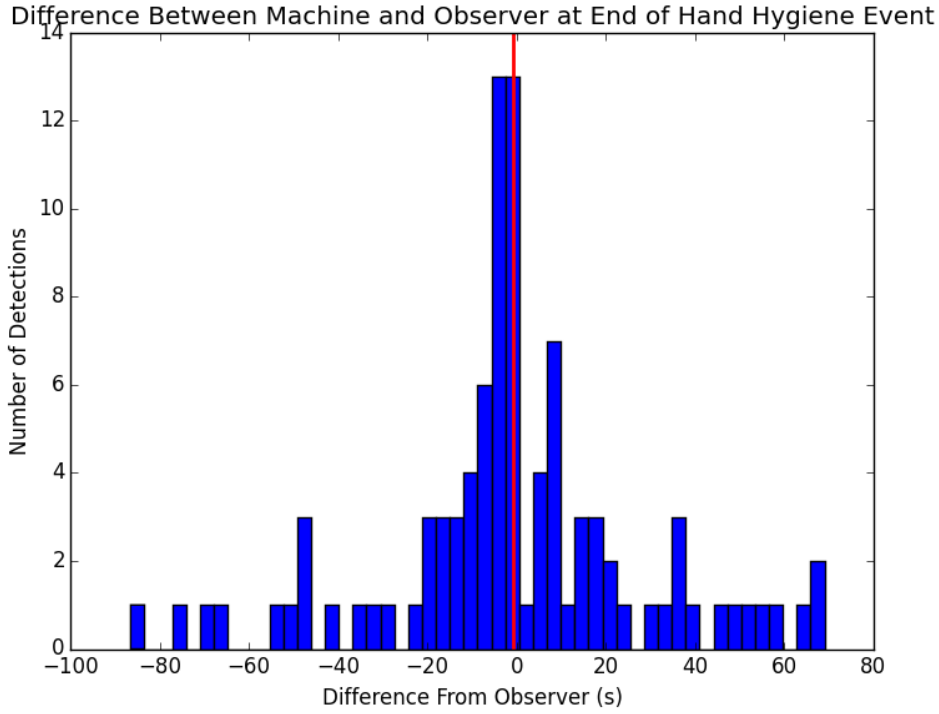

Figure 5.13: Difference Between End of Machine and Human Observed Hand Hygiene Events. Results using predictions from Naive Bayes classifier. Red line indicates mean. The machine and observer marking of the start of a hand hygiene event match closely.

CHAPTER 6:

HIERARCHICAL RECOGNITION FOR HAND HYGIENE ON THE HOSPITAL FLOOR

The system described in Chapter 5 works well, but there is little effort to improve the initial classification results before the final processing step. In addition the system relies on knowledge of the readings from both wrists. In a true deployment scenario it would be desirable to limit wireless communication by devices, and communicating all accelerometer readings from a wrist, either to a base station or to the paired wrist, would be power intensive and may be unnecessary.

This chapter describes a system which would instead do most classification on the wrists. The wrists would only need to communicate their most recent classification result. A second classifier is then run on those classification values to determine whether hand hygiene is occurring.

This system would both save power and possibly provide more accurate classifications for the final processing step.

6.1 Methods

For these results all training was done using data from the first Geneactiv data set and testing was done on the second Geneactiv data set (described in Section 5.1).

Classification is first performed on each wrist. The classification method is the same as described in Chapter 4, the only difference being that the features have been modified to be calculated using only the information available on the wrist. As an example, the axis crossing rate was calculated using only the three axes on the wrist, not all six axes on both wrists.

Every .25 seconds classification is performed on each wrist. The results are then used as inputs to a second classification scheme. This classifier uses the last five seconds of classifications on the left and right hand as input to predict whether hand hygiene is currently occurring.

The results from this second phase of classification were then processed in the same manner described in 5.2.

Tier 1	Tier 2	PPV(%)	HHE Detected(%)	Avg Duration Difference(s)
Naive Bayes	Naive Bayes	66.7	93.7	1.5
Naive Bayes	KNN	56.9	87.3	0.8
Naive Bayes	NN	66.0	76.7	-0.3
Naive Bayes	Decision Tree	64.6	94.7	2.1
KNN	Naive Bayes	72.5	90.5	0.6
KNN	KNN	48.1	96.8	4.0
KNN	NN	73.7	68.8	-2.1
KNN	Decision Tree	56.2	95.8	5.5
NN	Naive Bayes	67.1	90.5	2.3
NN	KNN	40.3	98.4	7.6
NN	NN	59.7	81.5	0.3
NN	Decision Tree	44.7	98.4	13.3
Decision Tree	Naive Bayes	61.9	93.7	2.6
Decision Tree	KNN	43.7	97.9	8.0
Decision Tree	NN	61.5	83.1	0.2
Decision Tree	Decision Tree	48.5	97.4	9.4

Table 6.1: Accuracy of Hierarchical Approach Using Different Classifiers.

6.2 Results

6.2.1 Detection Accuracy

In Figure 6.1 we can see that in the first phase accuracy is high while positive predictive value is low. Results obtained using only one wrist worth of data are also not as good as those obtained using both wrists (shown in Figure 5.9). Figure 6.2 shows that after the second phase of classification both accuracy and positive predictive value improve, becoming comparable with those obtained using both wrists.

In Table 6.1 we can see the performance after the final processing step. Overall the accuracy is quite high, in many cases beating the performance of the previous method as shown in Figure 5.10. The duration estimates are accurate as well, with many falling within a second of the observed duration.

6.2.2 Duration Accuracy

It is useful to look at the distribution of duration estimate error in order to determine whether the average is a good representation of classifier performance. Figure 6.4 shows the distribution of duration estimation error. Most estimates are within a few seconds of the true duration, however there are outliers, and even a few duration estimates that were more than 40 seconds different than the true duration.

Figure 6.5 shows the difference between the estimated start of a hand hygiene event and the true start of a hand hygiene event. Again most estimates are very close

to the actual starting time. In Figure 6.6 we can see that the end point estimation performs similarly.

Overall the estimation of duration is very accurate, and in addition both start and end points are accurately located.

6.3 Conclusion

It is possible to classify hand hygiene events accurately without knowing the accelerometer readings from each wrist and instead knowing only the classification results from each wrist. Not only were hand hygiene events detected, but their duration is accurately estimated. The hierarchical approach explored performed better than the simpler, single-tiered classification approach explored in Chapter 5, recognizing more hand hygiene events while at the same time producing a higher positive predictive value.

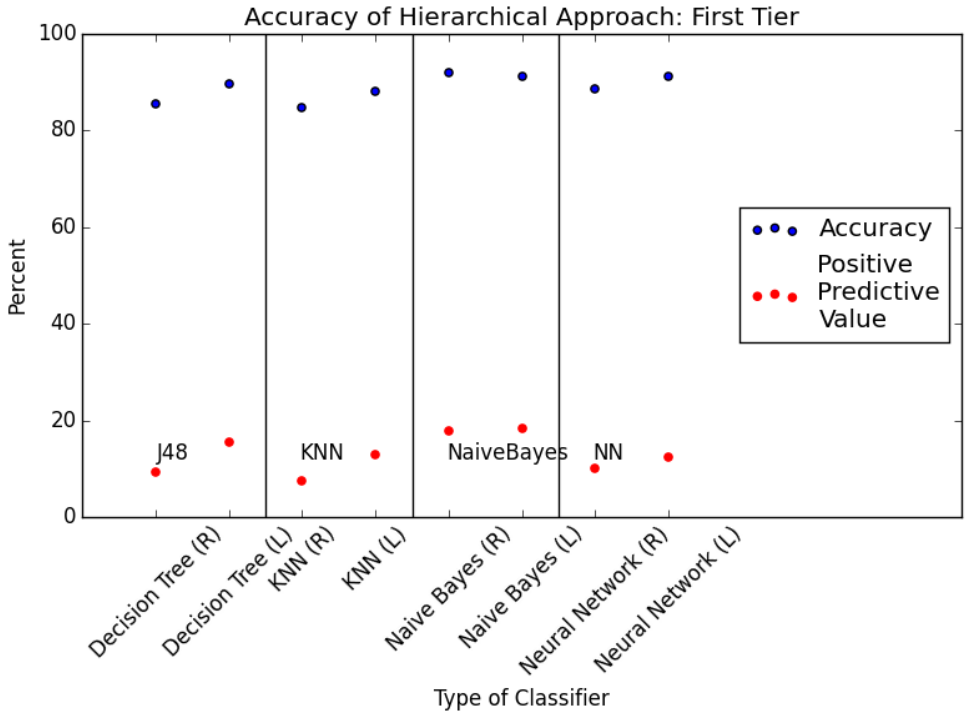

Figure 6.1: Hierarchical Approach: Phase 1 Accuracy. The labels in the figure indicate the phase 1 classifier used. The accuracy and positive predictive value is lower than that obtained when using data from both hands as shown in Figure 5.9.

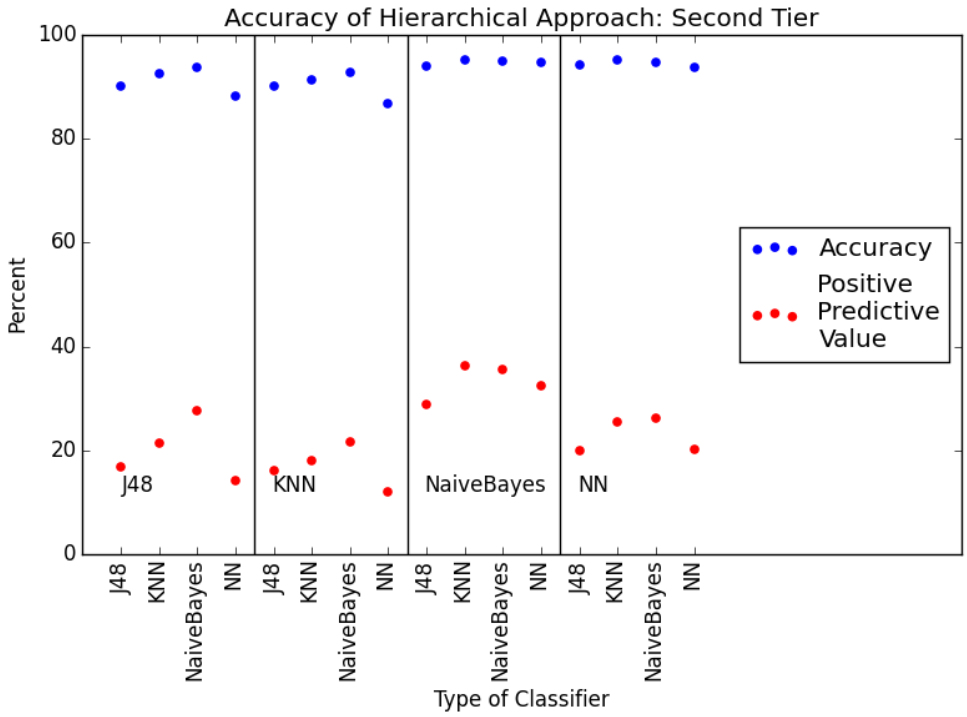

Figure 6.2: Hierarchical Approach: Phase 2 Accuracy. The labels on the x-axis represent the second phase classifier used while the labels on regions in the figure note the first phase classifier used. After the second tier of classification accuracy and positive predictive value is much higher, in many cases beating the performance of the previous method (Figure 5.10) that had access to both wrists of data.

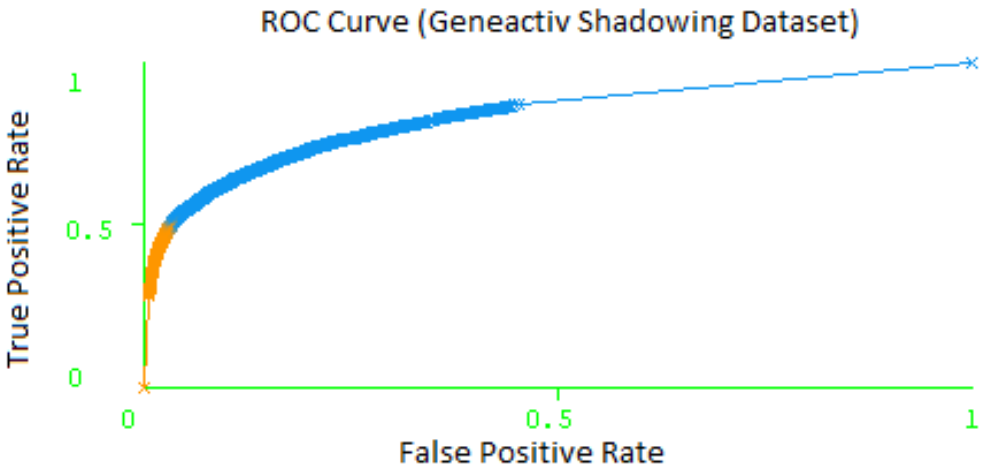

Figure 6.3: ROC Curve for Hierarchical Approach. This is the curve generated by using the naive bayes classifier in both tiers. These results are from the classifications before the second processing step. The x axis represents the false positive rate (the ratio of false positives to negative detections) and the y axis represents the true positive rate (a.k.a. sensitivity or recall). As the true positive rate increases the number of false positives also increases.

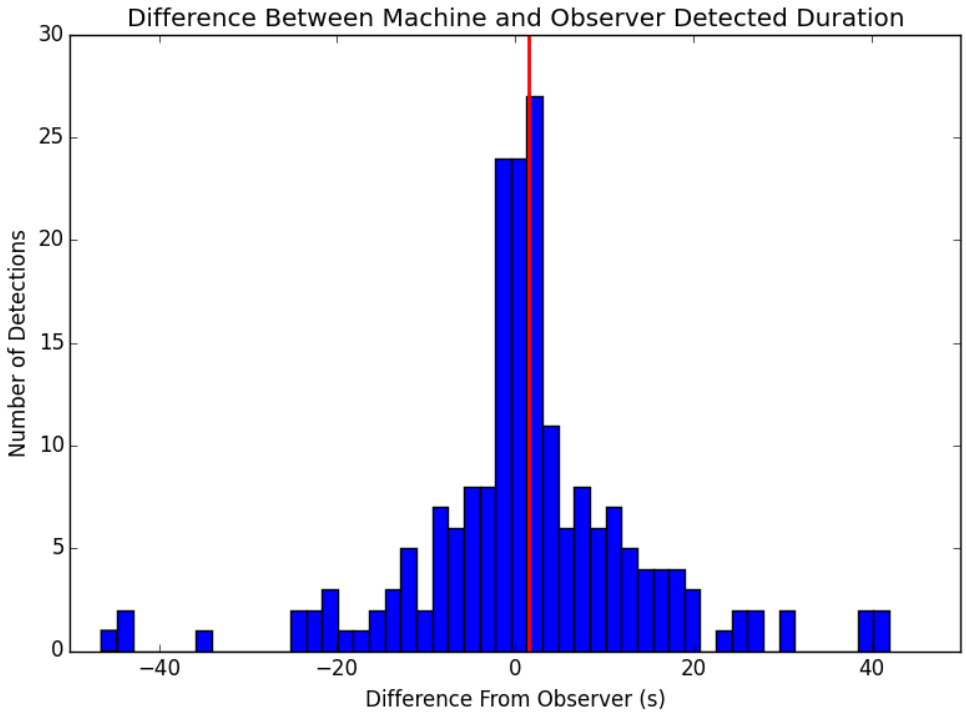

Figure 6.4: Hierarchical Approach: Duration Estimate Difference. Graph uses duration estimates from using a Naive Bayes classifier in both the first and second phases. The red line indicates the mean duration estimate difference. The observer and machine estimates of duration were frequently quite close, with the average difference in duration being 1.5 seconds.

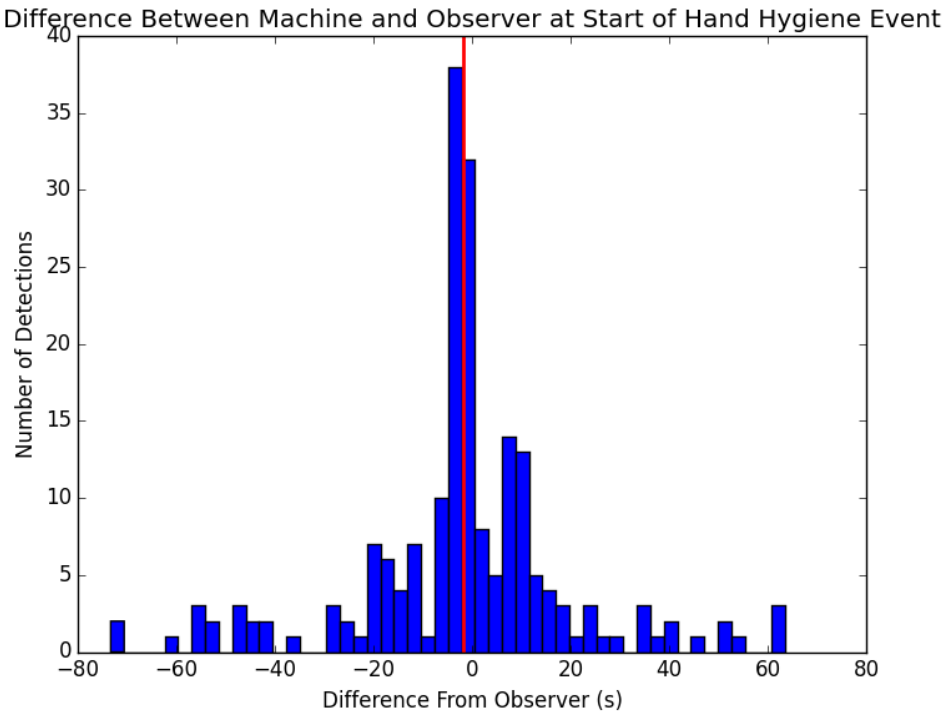

Figure 6.5: Hierarchical Approach: Start Estimate Difference. Graph uses duration estimates from using a Naive Bayes classifier in both the first and second phases. The red line indicates the mean start estimate difference. The observer and machine estimates of start time were frequently quite close.

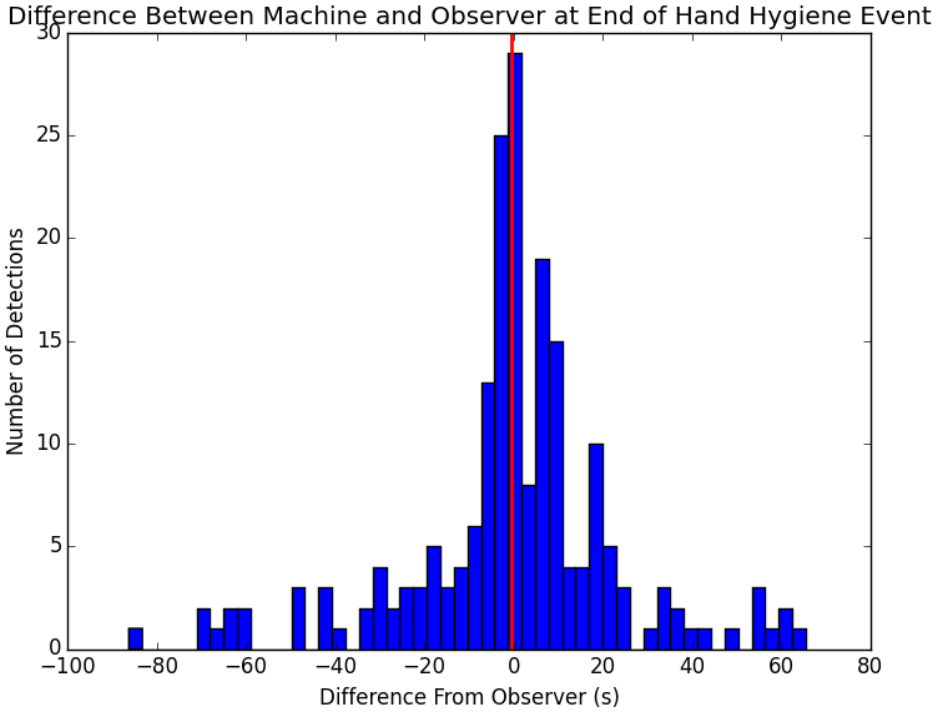

Figure 6.6: Hierarchical Approach: End Estimate Difference. Graph uses duration estimates from using a Naive Bayes classifier in both the first and second phases. The red line indicates the mean end estimate difference. The observer and machine estimates of end time were frequently quite close.

CHAPTER 7:

CONCLUSION

Electronic recognition of hand hygiene using only inexpensive, accelerometer-equipped wristbands is both accurate and feasible over large populations of subjects in the hospital. With hand hygiene event detection accuracies of over 90% in most cases, it is possible to detect hand hygiene events in the hospital without instrumented dispensers or location beacons. The system is accurate even over small windows of time, so the duration of an event can be closely estimated, in many cases to within a second of the observed duration.

The system has been tested extensively, from synthetic data sets consisting of 116 healthcare workers to data sets consisting of hours of healthcare workers observed on the hospital floor. Not only is the system accurate, but it is robust to changes in motion type and setting.

7.1 Open Questions

Every effort has been made to test this system extensively, but there are several open questions remaining. One is whether handedness has an effect on classifier performance. This was investigated but could not be answered due to the small numbers of left-handed individuals. A future trial consisting of left-handed subjects could help reveal whether handedness has an effect and possible ways to address it.

Another open question is whether frequency-based features could be made more effective.

It is possible that utilizing the proximity/existence of a triggered pump as another feature would increase the accuracy of a system that included instrumented pumps. This was not able to be explored using the current data because it would require a full-scale deployment in the hospital. In future a data set consisting of such measurements combined with wrist-based accelerometer readings could be used to determine whether pump triggering events can help in correctly identifying hand hygiene events.

A limitation of this system is that there is no microbiological confirmation of the quality of hand hygiene. The quality of hand hygiene is assumed to be good if it is

of sufficient duration or follows the correct technique. In reality the quality of hand hygiene is good if the hands have been cleansed of bacteria, and the duration or technique are only proxy measurements of this.

BIBLIOGRAPHY

[1] World Health Organization. Your 5 moments for hand hygiene. http://who.int/gpsc/tools/Five_moments/en, Oct 2006.

[2] World Health Organization. Who guidelines on hand hygiene in health care: First global patient safety challenge clean care is safer care. *Hand hygiene practices among health-care workers and adherence to recommendations*, 16, 2009.

[3] Fred M Gordin, Maureen E Schultz, Ruth A Huber, and Janet A Gill. Reduction in nosocomial transmission of drug-resistant bacteria after introduction of an alcohol-based handrub. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America*, 26(7):650–3, July 2005.

[4] Sunkesula VC, Meranda D, Kundrapu S, Zabarsky TF, McKee M, Macinga DR, and Donskey CJ. Comparison of hand hygiene monitoring using the 5 moments for hand hygiene method versus a wash in-wash out method. *American Journal of Infection Control*, 43(1):16–19, Jan 2015.

[5] John M Boyce. Measuring healthcare worker hand hygiene activity: current practices and emerging technologies. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America*, 32(10):1016–28, October 2011.

[6] Tim Eckmanns, Jan Bessert, Michael Behnke, Petra Gastmeier, and Henning Ruden. Compliance with antiseptic hand rub use in intensive care units: the Hawthorne effect. *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America*, 27(9):931–4, September 2006.

[7] Erol Kohli, Judy Ptak, Randall Smith, Eileen Taylor, Elizabeth A Talbot, and Kathryn B Kirkland. Variability in the Hawthorne effect with regard to hand hygiene performance in high- and low-performing inpatient care units. *Infection*

control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America, 30(3):222–5, March 2009.

[8] Mauricio N. Monsalve, Sriram V. Pemmaraju, Geb W. Thomas, Ted Herman, Alberto M. Segre, and Philip M. Polgreen. Do peer effects improve hand hygiene adherence among healthcare workers? *Infection Control and Hospital Epidemiology*, 35(10):pp. 1277–1285, 2014.

[9] Donna Armellino, Erfan Hussain, Mary Ellen Schilling, William Senicola, Ann Eichorn, Yosef Dlugacz, and Bruce F. Farber. Using high-technology to enforce low-technology safety measures: The use of third-party remote video auditing and real-time feedback in healthcare. *Clinical Infectious Diseases*, 54(1):1–7, 2012.

[10] S Tschudin Sutter, R Frei, M Dangel, and A F Widmer. Effect of Teaching Recommended World Health Organization Technique on the Use of Alcohol Based Hand Rub by Medical Students. *Infection Control and Hospital Epidemiology*, 31(11):1194–1195, 2010.

[11] McGinley KJ, Larson EL, and Leyden JJ. Composition and density of microflora in the subungual space of the hand. *Journal of Clinical Microbiology*, 26(5):950–953, May 1988.

[12] Gojo Industries. GOJO USA: GOJO SMARTLINK Hand Hygiene Solutions. http://www.gojo.com/united-states/market/healthcare/smartlink-compliance-solutions.aspx.

[13] HyGreen. HyGreen System. http://hygreen.com/HandHygieneMonitor/How.asp.

[14] Hill-Rom. Hill-Rom Hand Hygiene Compliance Solution — hill-rom.com. http://www.hill-rom.com/usa/Products/Category/Clinical-Workflow-Solutions/Hill-Rom-Hand-Hygiene-Compliance-Solution1/.

[15] Hyginex. Hyginex Electronic Hand Hygiene Monitoring Solution for Hospitals. http://www.hyginex.com/.

[16] SureWash. SureWash automated hand hygiene training and compliance measurement for wards and surgical hand preparation. http://www.surewash.com/.

[17] David Fernández Llorca, Ignacio Parra, Miguel Ángel Sotelo, and Gerard Lacey. A vision-based system for automatic hand washing quality assessment. *Machine Vision and Applications*, 22(2):219–234, December 2011.

[18] Heng-Tze Cheng, Feng-Tso Sun, Martin Griss, Paul Davis, Jianguo Li, and Di You. NuActiv: Recognizing Unseen New Activities Using Semantic Attribute-Based Learning. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13*, page 361, New York, New York, USA, June 2013. ACM Press.

[19] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. Recofit: Using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3225–3234, New York, NY, USA, 2014. ACM.

[20] Eduardo Velloso, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. Qualitative activity recognition of weight lifting exercises. *Proceedings of the 4th Augmented Human International Conference on - AH '13*, pages 116–123, 2013.

[21] Yu-Jin Hong, Ig-Jae Kim, Sang Chul Ahn, and Hyoung-Gon Kim. Mobile health monitoring system based on activity recognition using accelerometer. *Simulation Modelling Practice and Theory*, 18(4):446–455, April 2010.

[22] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services - MobiSys '14*, pages 149–161, New York, New York, USA, June 2014. ACM Press.

[23] Raul I. Ramos-Garcia and Adam W. Hoover. A Study of Temporal Action Sequencing During Consumption of a Meal. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*, pages 68–75, New York, New York, USA, September 2013. ACM Press.

[24] Yun Li, Xiang Chen, Xu Zhang, Kongqiao Wang, and Z Jane Wang. A sign-component-based framework for Chinese sign language recognition using accelerometer and sEMG data. *IEEE transactions on bio-medical engineering*, 59(10):2695–704, October 2012.

[25] Mridul Khan, Sheikh Iqbal Ahamed, Miftahur Rahman, and Ji-Jiang Yang. Gesthaar: An accelerometer-based gesture recognition method and its application in NUI driven pervasive healthcare. In *2012 IEEE International Conference on Emerging Signal Processing Applications*, pages 163–166. IEEE, January 2012.

[26] Sven Kratz, Michael Rohs, and Georg Essl. Combining acceleration and gyroscope data for motion gesture recognition using classifiers with dimensionality constraints. In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*, page 173, New York, New York, USA, March 2013. ACM Press.

[27] Ruize Xu, Shengli Zhou, and Wen J. Li. MEMS Accelerometer Based Nonspecific-User Hand Gesture Recognition. *IEEE Sensors Journal*, 12(5):1166–1173, May 2012.

[28] Liang Yin, Mingzhi Dong, Ying Duan, Weihong Deng, Kaili Zhao, and Jun Guo. A high-performance training-free approach for hand gesture recognition with accelerometer. *Multimedia Tools and Applications*, March 2013.

[29] Jeen-Shing Wang and Fang-Chen Chuang. An Accelerometer-Based Digital Pen With a Trajectory Recognition Algorithm for Handwritten Digit and Gesture Recognition. *IEEE Transactions on Industrial Electronics*, 59(7):2998–3007, July 2012.

[30] Sandip Agrawal, Ionut Constandache, Shravan Gaonkar, Romit Roy Choudhury, Kevin Caves, and Frank DeRuyter. Using mobile phones to write in air. In *Proceedings of the 9th international conference on Mobile systems, applications, and services - MobiSys '11*, page 15, New York, New York, USA, June 2011. ACM Press.

[31] *CC2530 Second Generation System-on-Chip Solution for 2.4 GHz IEEE 802.15.4, RF4CE, ZigBee.*

[32] *MMA7455L Datasheet.*

[33] *Invensense Motion Processing Unit.*

[34] Yei 3-space sensor. http://www.yeitechnology.com.

[35] http://www.geneactiv.org.

[36] D. Pittet, B. Allegranzi, and J. Boyce. The world health organization guidelines on hand hygiene in health care and their consensus recommendations. *Infection Control and Hospital Epidemiology*, 30(7):611–622, 2009.

[37] Tools for training and education. Web, Aug 2009.

[38] Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, Nov 1986.

[39] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[40] M.A. Hall. Correlation-based feature subset selection for machine learning, 1998.

[41] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[42] HT Cheng. Learning and Recognizing The Hierarchical and Sequential Structure of Human Activities. 2013.

[43] H. Ghasemzadeh, V. Loseu, and R. Jafari. Collaborative signal processing for action recognition in body sensor networks: a distributed classification algorithm using motion transcripts. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks - IPSN '10*, pages 244–255. ACM Press, April 2010.