A Practical Approach to Recognizing Physical Activities

Jonathan Lester[1], Tanzeem Choudhury[2], and Gaetano Borriello[2,3]

[1]Department of Electrical Engineering, University of Washington, Seattle WA 98195, USA
[2]Intel Research Seattle, Seattle, WA 98105, USA
[3]Department of Computer Science, University of Washington, Seattle WA 98195, USA

Abstract. We are developing a personal activity recognition system that is practical, reliable, and can be incorporated into a variety of health-care related applications ranging from personal fitness to elder care. To make our system appealing and useful, we require it to have the following properties: (i) data only from a single body location needed, and it is not required to be from the same point for every user; (ii) should work out of the box across individuals, with personalization only enhancing its recognition abilities; and (iii) should be effective even with a cost-sensitive subset of the sensors and data features. In this paper, we present an approach to building a system that exhibits these properties and provide evidence based on data for 8 different activities collected from 12 different subjects. Our results indicate that the system has an accuracy rate of approximately 90% while meeting our requirements. We are now developing a fully embedded version of our system based on a cell-phone platform augmented with a Bluetooth-connected sensor board.

1 Introduction

The task of recognizing human activities from body worn sensors has received increasing attention in recent years. With a growing demand for activity recognition systems in the health care domain, especially in elder care support, long-term health/fitness monitoring, and assisting those with cognitive disorders [1, 2, 3]. For an automatic activity recognition system to be useable in these domains it is important for it to be practical as well as accurate.

Current methods for tracking activities in the healthcare field are time and resource consuming manual tasks, relying on either paid observer (i.e. a job coach who periodically monitors a cognitively disabled person performing their job or a nurse monitoring an elderly patient) or on self-reporting, namely, having patients complete an activity report at the end of the day. However, these methods have significant deficiencies in cost, accuracy, scope, coverage, and obtrusiveness. Paid observers like job coaches and nurses must typically split their time among several patients at different locations, or the patients must be clustered together. Self-reporting is often inaccurate and of limited usefulness due to patient forgetfulness and both unintentional and intentional misreporting, such as a patient reporting more fitness activities than they actually completed.

An automatic activity recognition system would not only help reduce the errors that arise from self-reporting and sparse observational sampling, but hopefully also improve the quality of care for patients as caregivers spend less of their time performing bookkeeping duties. In addition, unobtrusive monitoring enables people to go about their daily lives in an unimpeded manner while providing their caregivers with a more accurate assessment of their real life activities rather than a small sample. While the full extent to which patients would benefit from such a device is not currently known, an accurate automated system does have a clear benefit over existing methods such as surveys in providing a continuous activity log along with times and durations for a wide range of activities.

Moreover, if monitoring can be accomplished using a personal device that looks like any other common consumer device (rather than a device with many wires to all the limbs, for example) then there is less resistance to wearing it as it does not cause a social stigma that would identify a person's ailment to others. Acceptability is further enhanced by the embodiment in a personal device over which the user has complete control, and may choose to share (or choose not to share) their data with health care professionals or choose to not wear the device (or turn it off) for some occasions.

1.1   Related Work

A majority of the research using wearable sensors for activity recognition has so far focused on using a single sensor modality, typically accelerometers, placed in two or more (up to 12) locations on the body [4, 5, 16]. Placing sensors in multiple pre-defined locations or even a single, fixed, location every time can be quite cumbersome when one has to collect data on a daily and continuous basis. Work by [5] has showed that placing an accelerometer at only two locations (either the hip and wrist or thigh and wrist) did not affect activity recognition scores significantly (less than 5%) when compared to a system with five sensors; whereas the use of a single sensor reduced their average accuracy by 35%. In our recent work [7], we showed that we could compensate for the accuracy lost using a single sensing location by using a single sensing location with multiple sensor modalities (in this case 7 different sensor types).

This is a promising line of investigation as it is much more comfortable for the user to wear a single device at only one location. Moreover, because we only need a single sensing location these sensors could be incorporated into existing mobile platforms, such as cell phones or wristwatches. Integrating sensors into devices people already carry is likely to be more appealing to users and garner greater user acceptance as these consumer devices do not make them look "different". Greater user acceptance would hopefully also mean that users would be more inclined to wear the device more often and for a larger part of the day, allowing it to capture a greater range of activities than a device worn only sparingly.

1.2   Our Hypotheses

In this paper, we investigate several practical aspects of creating an automatic, personal activity recognition system. More specifically, there are three aspects we want to understand in more detail: location sensitivity, variations across users, and the required sensor modalities. Through our experiments, we seek to answer the following questions:

- Does it matter where on their person the user carries the device? If we have training examples that contain data from multiple body locations, will the recognition algorithm generalize such that we do not have to learn location specific models?
- How much variation is there across users? Does the device need to be customized to each individual for reliable accuracy or can it be trained more generally so that it works "out-of-the-box"?
- How many sensors are really needed to recognize a significant set of basic activities? Are 7 really necessary or can a cheaper, lower-power system be built from fewer sensors but still have similar recognition accuracy?

The remainder of the paper presents answers to these questions by providing evidence based on data collected from 12 different individuals performing 8 different activities over several days, carrying a collection of sensors worn in three different locations on the body. The activities include: sitting, standing, walking, walking up/down stairs, riding elevator up/down, and brushing teeth. These activities were selected because they correspond to basic and common human movements and will be useful for elder care and personal fitness applications. These physical tasks are also similar to the tasks previous activity recognition systems have attempted to recognize.

The three locations where volunteers wore the sensors correspond to locations where people already carry devices – (i) wrist (e.g., wristwatch), (ii) waist (e.g., cell phone or pager clip), and (iii) shoulder (e.g., cell phone pouch on a bag's strap). Data was collected from these three locations simultaneously by using a wired set of three sensor boards, one at each location. However, we currently have a wireless implementation that sends data from a single sensing unit via Bluetooth to any commodity Bluetooth device, like a cell phone. In the future, we expect that the sensors will be part of the cell phone itself and the challenge lies in implementing all of our classification algorithms on that platform. Wristwatch type platforms (e.g. Microsoft SPOT) are also becoming more capable but are likely to only provide a glanceable user interface rather than a capable computational resource in the short to medium-term future.

To accurately track various activities using a single wearable unit, we use the activity classification algorithm we developed in [7]. This algorithm employs an ensemble of very simple static classifiers to select the most useful features and then uses those features to recognize a set of basic human movements (walking, sitting, going down stairs, etc.). Each simple static classifier in the ensemble operates on a single feature, giving the system the flexibility to use a varying number of features. A second layer of hidden Markov models (HMMs) combines the outputs of the classifiers into an estimate of the most likely activity while providing temporal smoothing.

Figure 1. The multimodal sensor board *(top)*, a Bluetooth iMote *(lower left)*, and USB rechargeable battery board *(lower right)*. A Nokia 6600 series phone with a sensor board on the back for scale *(middle image)*. The data collection setup consisted of three sensor boards and a small Vaio Notebook for data logging (inside the pouch on the user's right side)

The results presented in this paper show that the classification algorithms are robust to the realistic variations that appear in the data. Furthermore, they are well behaved with respect to different locations, different people, and a smaller set of sensors. Of course, further testing across a larger population is still needed to fully understand the limits of this approach. This paper makes the case that the approach is worth investigating further and meets many of the practicality requirements of the real world.

In the following sections we will describe in more detail the multi-modal sensor platform we use, the machine learning algorithms for robustly inferring activities, and our experimental methodology and results. We will conclude with a summary of our contributions and directions for future work.

2   Experimental Methodology

To gather data simultaneously from a large set of sensors and to better understand the usefulness of different sensor modalities in inferring human activities, we used the multi-modal sensor board (MSB) (shown above in Figure 1). The MSB is designed to attach to the Intel Mote (iMote, also shown in Figure 1), a Bluetooth/32-bit ARM7-based wireless sensor node, but can also communicate with handheld, desktop computers, and cell phones via Bluetooth, serial, or USB. The iMote allows us to send the MSB's sensing data to any commodity Bluetooth device like a cell phone[1] and our other interfaces allow us to connect to handhelds, laptops, and other devices.

The sensor board contains seven different sensors and is capable of sampling them all simultaneously at fairly high sampling rates (see Table 1). The sensors on the MSB were selected for their general usefulness (as evidenced by related work in activity inference [6, 8, 9]), small footprint, low power consumption, and availability of digital

---

[1] Due to the limitations of Bluetooth and cell phones we can only transmit audio data at 8kHz all other sensor can function at their normal rates.

Table 1. Sensors on the multi-sensor board and the sampling rates used for the experiments in this paper

Manufacturer	Part No.	Description	Sampling Rate
Panasonic	WM-61A	Electric Microphone	~ 16000 Hz
Osram	SFH-3410	Visible Light Phototransistor	~ 550 Hz
STMicro	LIS3L02DS	3-Axis Digital Accelerometer	~ 550 Hz
Honeywell	HMC6352	2-Axis Digital Compass	30 Hz
Intersema	MS5534AP	Digital Barometer / Temperature	15 Hz
TAOS	TSL2550	Digital Ambient (IR and Visible+IR) Light	5 Hz
Sensirion	SHT15	Digtial Humidity / Temperature	2 Hz

interfaces. Sensors such as those used on the MSB are already being incorporated into cell phones and similarly equipped wristwatches are likely to follow soon [10].

The MSB is small and light enough (9.2g) to wear comfortably for long periods of time. Even with an iMote (5.7g) and battery board (9.6g including a 200mAh battery), it only weighs 25g. While streaming data to the cell phone the iMote+MSB can run for approximately 4 hours on a single 200mAh Li-Polymer battery. The battery board can also use one or two 1800mAh Li-Ion batteries allowing for a runtime of more than 36 hours (72 hours with two batteries).

2.1 Data Collection from Multiple MSBs

As the basis for this and future research we have created a multi-person dataset from 12 individuals (two female ten male) containing 8 basic physical activities. Eight of the volunteers in the study were graduate students in their mid 20s and four were older in their 30s. Two-thirds of the data was collected from a computer science building and the other third was collected in an office building. Volunteers collected data wearing three MSBs: one on a shoulder strap, one on the side of their waist, and one on their right wrist. The volunteers were given a sequence of activities to perform, like sitting on a couch for a few minutes before walking upstairs to brush their teeth. A miniature notebook (a Sony Vaio U750 measuring 16.8cm×6.6cm×2.5cm and weighing 560g) was carried in a small pouch to collect the data from the three MSBs

Table 2. Data collected for the second data set, consisting of 8 basic activities collected by 12 volunteers. A total of 12 hours of data was recorded, 7 hours of which was labeled as corresponding to any one of our set of 8 activities

		Duration		Instances
	Sitting		56 mins	22
	Standing	1 hr	13 mins	135
	Walking	1 hr	43 mins	215
Labeled Activities	Walking up stairs		19 mins	34
	Walking down stairs		14 mins	30
	Riding elevator down		16 mins	42
	Riding elevator up		15 mins	39
	Brushing Teeth		20 mins	12

Average Duration:		40 mins	
Total Labeled Data:	6 hrs	55 mins	

Table 3. An example subset of some of the features calculated with descriptions. Note, that a single feature type, such as frequency bands, can have multiple outputs in the feature vector, i.e. one feature per band

Feature	Description
Cepstral Coefficients	The FFT of the log FFT spectrum, that is FFT(log(FFT(x)))
Log FFT Frequency Bands	Real valued FFT values grouped into logarithmic bands
Spectral Entropy	Measure of the distribution of frequency components
Energy	The sum of the real FFT spectrum
Mean	The average value of the time series
Variance	The square of the standard deviation
Linear FFT Frequency Bands	Real valued FFT values grouped into linear bands from 100Hz - 2kHz
Correlation Coeffs	Correlation between axis pair, XY, XZ, YZ
Integration	Integration of the timer series over a window

and an observer annotated the data in real time by using a simple annotation program on an iPAQ. The observer also provided cues to the volunteers about the sequence of activities they should be performing. Figure 1 shows a picture of the setup the volunteers wore while collecting their data. And Table 2 lists the activities collected from this experiment along with the amount of data recorded and the number of times a label appeared in our annotations.

2.2   Ground Truth

For our data set, ground truth was obtained while the experiments were being carried out. An observer carrying an iPaq marked the start and end points of the activities as the volunteer performed them by clicking on a set of labels on the iPaq. This on-line ground truth collection eliminated the need for the volunteers to annotate their data after they had completed the experiment and helped to reduce timing and annotation errors. If the user was performing an activity that wasn't in our list of activity classes the segment was automatically marked as null/do-not care and was ignored when we trained and tested our classifiers.

2.3   Classification Features

As we collect approximately 18,000 samples of data per second we do not use the samples directly for classification, but instead we compute features to summarize the data, reduce the dimensionality of our data, and to bring out important details from the data. We currently compute a total of 651 features; which include linear and log-scale FFT frequency coefficients, cepstral coefficients, spectral entropy, band-pass filter coefficients, correlations, integrals, means, and variances. Table 3 gives a description of the various types of features used – note that a feature type (e.g., linear FFT bands is based on FFT magnitudes grouped in linear bands) can account for multiple features in the feature vector. Also, we do not compute every type of feature for every sensor, e.g., an FFT is only computed for the audio, accelerometer and high frequency light sensors, as it does not make sense to compute it for the other sensors. We combine the

features from various sensors to produce a 651 dimensional feature vector at 4Hz. Due to the fact that we have sensors with different sampling rates; there are multiple instances of some features within each 0.25 second window. Furthermore, when calculating some features (e.g., the integral features) we incorporate a longer time window that varies from several seconds to as long as a minute. For those features, we restrict the time windows to only use data from the past, so that our system functions without a time lag.

It might be intuitively clear which features our algorithms should use for some activities (e.g., FFT coefficients of acceleration will likely capture walking gait); while for others it might not be as clear (e.g., for riding an elevator). Using all 651 features might solve the feature selection problem; however, the disadvantage of this approach is that we might not have enough data to reliably learn all the parameters of our models and some features might even confuse the classification algorithms rather than help discriminate between the various activity classes. In section 3, we briefly describe the algorithm developed in [7] that automatically picks the most useful features and learns classifiers from these.

2.3   Classification

To train our classifiers we needed to separate our data into training and testing sections. To do this we divided up our data set into a 4-folded test and training set by segments, or continuous blocks of data that were classified as the same activity. We randomly selected segments from our available data and placed them into our folds until we had about the same number of segments in each fold. We then train our static and HMM classifiers using 3 of the 4 folds and then test on the remaining fold (75% for training and 25% for testing), performing this same training/testing operation four times, using the different combinations of our 4 folds. Note that all the feature extraction, classification, and data analysis presented in this paper was performed offline.

The static decision stumps classifiers we learned were all trained using our 651 features computed at 4Hz. To prevent our classifier from over fitting to properties of the locations where our data was collected, we did not allow it to use the temperature or humidity sensors. These sensors should have very little to do with the actual recognition of the physical activities and could potentially cause the classifier to report results that were good classifications; but, were influenced by the locations where the data was collected and not the actual activities. For example, if you collected a lot of data of a person sitting in a cool air conditioned room and then standing in a warm heated room. During the training phase the classifier would see temperature as being the most telling feature in determining whether you were sitting or standing, even though temperature clearly has no direct bearing on whether one is sitting or standing. Of course, in the real data similar effects may be more subtle, nonetheless to avoid any obvious problems temperature and humidity are not used as sensor inputs for our activity recognition.

The HMM classifier was trained using the margin output of the decision stumps classifier (calculated at 4Hz) and uses a 15 second sliding classification window with a 5 second overlap between windows (a 10 second window advancement).

3   Learning Activity Models

The two principal approaches used for classification in machine learning are: (i) generative techniques that model the underlying distributions of the data classes and (ii) discriminative techniques that only focus on learning the class boundaries [11]. Both of these approaches have been used extensively in the vision and wearable-sensing communities for recognizing various human behavior and activities. In [7], we developed a hybrid approach that combines the two techniques and demonstrated it to be quite effective. We only provide a brief summary of the techniques here, readers are referred to [7] for more details.

First, a modified version of AdaBoost proposed by [12], was used to automatically select the best features and rank them based on their classification performance. Given the maximum number of features the activity recognition system can use, the system automatically chooses the most discriminative sub-set of features and uses them to learn an ensemble of discriminative static classifiers for the activities that need to be recognized. As the features are selected incrementally based on their usefulness, desirable classifier performance was achieved while using less than 10% of the possible features (i.e., the top 50 features). Including additional features beyond the top 50 had very little impact on the classification performance. On average using classifiers with 600 features improves training error by less than 1% compared to classifiers with 50 features.

Second, the class probabilities estimated from the static classifiers are used as inputs into hidden Markov models (HMMs). The discriminative classifiers are tuned to make different activities more distinguishable from each other, while the HMM layer on top of the static classification stage ensures temporal smoothness and allows for continuous tracking of the activities. Using HMMs to recognize activities in continuous time chunks has the following advantages: (i) the classification output will incorporate history information to smooth out much of the sporadic errors that occur during the static classification step and (ii) we can learn how people transition between activities and thereby more about people's composite activity routines.

4   Location Sensitivity

In [7], we colleted a large data set of about 30 hours[2] of data from two volunteers wearing a MSB on their shoulder performing various activities. From this data set, features and classifiers were developed that could robustly detect 10 activities with

[2] We collected approximately 30 hours of data; however, due to memory limitations we could only train our classifiers using approximately 12 hours of this dataset

greater than 90% accuracy: sitting, standing, walking, jogging, walking up/down stairs, riding a bicycle, driving a car, and riding an elevator up/down. This larger dataset served to verify that the algorithms would work on a variety of data using a single multi-modal sensor placement trained with a large amount of data. In this paper we build on this result. It should be noted that even though the results in this paper are slightly lower than those in [7], we would expect that with more training data from each individual we would approach the results in [7].

Previous methods have shown that by using sensors placed on different parts of the body one can classify activities fairly well. However, the problem with multiple sensor placements is of course that they can be quite obtrusive; but, similarly a single sensor placement can also be obtrusive (although to a lesser extent) if the user is required to carry it in the same location all of the time. Ideally, we would like the classification algorithms to work accurately with data from different locations on the body. This would allow the user to carry the device in a location that is the most convenient for a given context. For example, although a majority of men do not object to wearing devices on their waist belts, it is not always practical to do so; and women often do not wear a belt.

To determine the role sensor placement in recognition accuracy, we trained four sets of classifiers using data from the three locations (i) trained using data from all three locations, (ii) trained using data from the shoulder, (iii) trained with data from the waist, and (iv) trained with data from the wrist.

Note that the first classifier (using data from all three locations) does not combine all the locations together and require the user to simultaneously wear three sensors. Instead it treats each location as a separate input; so that when a user wears a single sensing device, they may wear this device at any of the three locations and use this classifier to classify that data. Table 4, shows the confusion matrix for this more location-independent first classifier and Table 5 lists the overall precision and recall[3] for each of our four classifiers (note that since the HMM uses the output of the static classifier as input, it is the final classification and the static classifier is only an intermediate stage – we show both to highlight the effect of the combining the two methods).

As we can see from Tables 4 and 5 the more generic classifier trained on data from all three locations does quite well over all. The classifiers trained for specific locations achieve slightly higher overall precision and recall scores. As mentioned in section 3,

---

[3] Overall precision and recall are calculating by normalizing the confusion matrix so that each labeled class has the same number of examples (i.e. all rows sum to 1).
Precision is defined as the proportion of the data labeled as activity N that actually was from data labeled as activity N in the ground truth:
  True Positive / (True Positive + False Positive)
While recall is defined as the proportion of the data originally labeled as activity N that was correctly classified as activity N:
  True Positive / (True Positive + False Negative)
Overall accuracy is defined as:
  True Positive + True Negative / Total Number of Examples
  Where the True Negative and False Positive counts are normalized by dividing by the (number of classes – 1)

the algorithm uses the training data to select the most discriminative subset of features. So the first classifier (which uses training data from all three locations) succeeds in finding features that are common to the data from all three locations and builds classifiers that work reliably on data from all three locations.

These results show that it is possible to train a generalized classifier that can recognize our 8 activities when worn in any of the three locations and that if we train for a specific location we can achieve slightly higher classification results. In addition it is also possible for us to train several classifiers (i.e., a generic one and more

Table 4. Confusion matrix for the static and HMM classifier trained using a single stream of sensor data from all three locations on the body

Static Classifier Confusion Matrix		Recognized Activity							
		Sitting	Standing	Walking	Walking up stairs	Walking down stairs	Riding elevator down	Riding elevator up	Brushing Teeth
Labeled Activities	Sitting	19245	17941	501	82	39	30	34	267
	Standing	6118	37602	3154	245	87	520	331	747
	Walking	998	7280	57658	1965	1907	771	666	1004
	Walking up stairs	1	297	1481	11277	35	23	188	39
	Walking down stairs	7	139	1422	59	7959	294	1	28
	Riding elevator down	68	2326	328	9	451	8023	214	11
	Riding elevator up	24	1936	332	267	4	332	7817	31
	Brushing Teeth	231	5252	1052	65	50	20	21	7859

Static+HMM Classifier Confusion Matrix		Recognized Activity							
		Sitting	Standing	Walking	Walking up stairs	Walking down stairs	Riding elevator down	Riding elevator up	Brushing Teeth
Labeled Activities	Sitting	28526	9173	404	0	0	0	0	36
	Standing	14308	26956	2791	261	204	2106	1241	937
	Walking	2933	4719	57278	2195	2770	941	766	647
	Walking up stairs	0	71	467	12632	10	29	132	0
	Walking down stairs	20	110	728	42	8821	188	0	0
	Riding elevator down	0	189	156	0	936	9992	157	0
	Riding elevator up	0	194	221	648	0	295	9385	0
	Brushing Teeth	252	1523	399	0	0	0	65	12311

Table 5. Overall precision/recall for the static and HMM classifiers trained/tested on all locations *(top row)* and a single location *(bottom rows)*. The classifier trained on all three locations does not combine all three sensing positions to perform its classification; rather it is a generalized classifier, using a single sensor placement, which works at all three locations. The overall accuracies for all cases were approximately 90%

	Static Classifier		HMM Classifier	
	Overall Precision	Overall Recall	Overall Precision	Overall Recall
Trained on Location 1,2,3 *(all locations)* **Tested on Location 1,2,3 *(all locations)***	79.18%	71.14%	82.07%	81.55%
Trained on Location 1 *(shoulder)* **Tested on Location 1 *(shoulder)***	79.37%	71.26%	83.84%	82.64%
Trained on Location 2 *(waist)* **Tested on Location 2 *(waist)***	81.83%	77.05%	85.87%	84.85%
Trained on Location 3 *(wrist)* **Tested on Location 3 *(wrist)***	81.01%	68.66%	87.18%	87.05%
Single Location Average:	80.74%	72.32%	85.63%	84.85%

location-specific ones) and determine the most appropriate classifier to use at runtime. There are some existing techniques that have shown promise in determining where a device is being carried on the body [6] and using the output from these techniques we could select location specific classifiers to boost our accuracy when the device is carried in a well-characterized location.

While the locations here do not represent all the possible locations where one could carry the sensing device they do cover the most common ones where people carry existing devices. It is also unlikely that subtle variations in the actual locations will change the classification significantly. Intuitively, there should be much larger differences between the waist and the wrist than there would be between the waist and a pocket. When we take our location specific classifiers and test them on data from another location (for example when classifiers are trained on the shoulder and tested on data from the waist), we suffer a 20% reduction in precision/recall. And when trained on data from the shoulder and waist and tested on data from the wrist there is about a 10-15% reduction in precision/recall.

5   Variation Across Users

An important practical issue with any device based on statistical inference is whether it can be useful immediately to the end user or whether it has to go through a training period. Clearly most users will want any device to work immediately upon purchase. If it gets better over time, that is a plus, but it must perform reasonably out of the box. Ideally, we would want to collect a lot of training data from a large diverse group of individuals to train a generic classifier, and then apply this classifier to classify a new individual's data without having to collect any new training data or retrain the classifier.

To test how well our classifiers could handle the variations that exist across different users we created several combinations of the 12 individuals' data on which to train. These combinations were created by randomly selecting N individuals' data for training, where we varied N = 1 (training data from one individual) to N = 12 (training data from all of our 12 test subjects). For each combination we performed four folded cross validation where we used 75% of our data to train with and 25% to test. The data used for this experiment uses data from the shoulder location only, so we would expect similar results to those we saw in row two of Table 5.

We measured the performance of our classifiers under two test conditions: (i) in the first case we tested on data from all 12 individuals but trained on data from [1, … ,N] individuals (ii) in the second case we tested on data from individuals who had not been in the training set, i.e. if we trained on individuals [1,…,N] we would test on data from individuals [N+1, … , 12].

The objective of the first test case was to determine if training on an increasingly larger subset would improve recognition accuracy. We would expect that each additional individual we add to our training data would improve our recognition accuracy. The objective of the second test case is to ensure that any improvement in recognition accuracy comes solely from the classifier being more "generalized" and
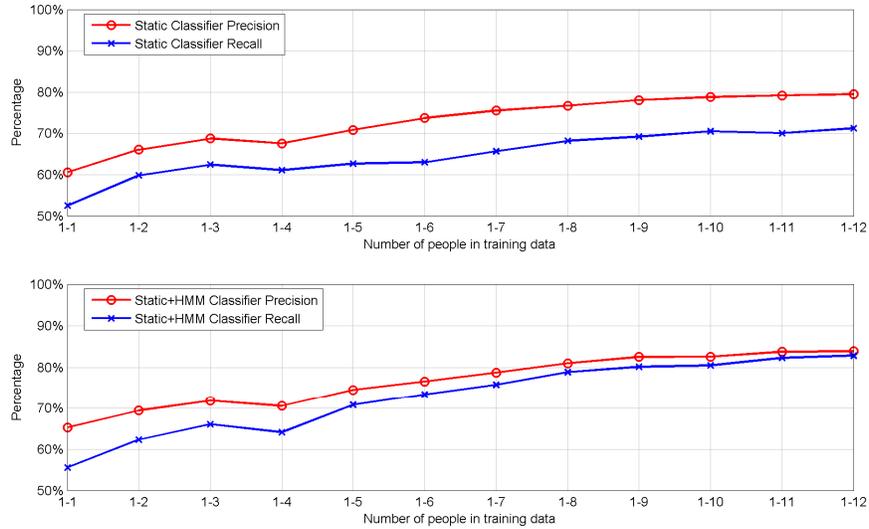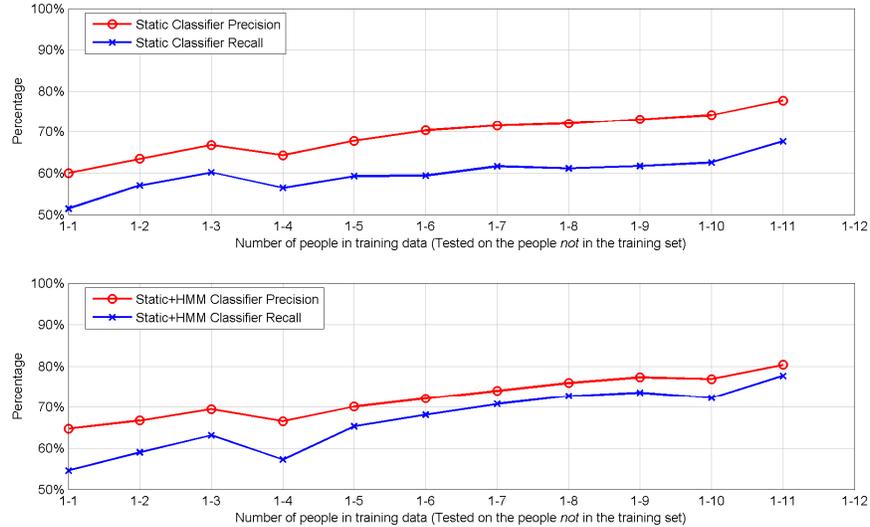
Figure 2. The overall precision and recall graphs for the static and static+HMM classifier. As we add in more individuals to our training set the precision and recall increase. The classifiers trained at each point were tested on test data from all 12 volunteers. E.g. data point [1-10] corresponds to training on 10 individuals and testing on the test data from all 12. The overall accuracy for this test case reaches approximately 95%

not because data from an increased fraction of individuals is used during training as in the first test case.

Figure 2 shows the overall precision and recall for the first test where we add in more and more people into the training set while always testing on all the test data from the 12 individuals. As we can see from Figure 2 the overall precision and recall increase as we add more and more individuals into our training set, eventually reaching about 84%. The [1-12] data point Figure 2 also corresponds very closely to the results we saw in row two of Table 5 as they both represent very similar tests.

Figure 3 shows the overall precision and recall for the second test case where we again add more and more people into the training set but test using test data from people who were not in our training set. And again we see an increase in overall precision and recall, eventually reaching around 80%. As figures 2 and 3 indicate, the more people we include in our training data the better our classifiers perform and that they are able to operate on a diverse set of individuals. In practice, manufacturers of an activity recognition system should be able to pre-train a set of classifiers so the device will work well for most users right away.

As you may have noticed in [7] we achieved ~90% accuracy for our classifiers and here we're only obtaining about 80-84%. The reason for this difference is that the data used in [7] contains a large amount of data (~30 hours) from two individuals over a period of six weeks, whereas the data set here contains a total of 6 hours of data from 12 individuals. We would expect that with more data these results would approach those in [7].

Figure 3. The overall precision and recall graphs for the static and static+HMM classifier. As we add in more individuals to our training set the precision and recall increase. The classifiers trained at each point were tested on test data from individuals who were not in the training data. E.g. data point [1-10] corresponds to training on 10 individuals and testing on the test data from the other two (of our 12) individuals who were not in the training dataset. The overall accuracy reaches approximately 87%

6   Sensors Necessary for Classification

While all the sensors on the MSB platform are interesting and provide some useful information, we do not necessarily need all the sensors to perform accurate classification. In fact, the results presented here and in [7] discarded information from the temperature and humidity sensors, simply because the classifiers would often incorrectly choose them as being indicators of when a physical activity was performed (sitting and standing shouldn't be strongly correlated to the temperature).

In addition, sensors such as light can often be obscured because the device is being carried in a pocket or obscured by clothing. By examining the features selected by the classifier we can see that the three most important sensors for the classifiers are the accelerometer, audio, and barometric pressure sensors. These correspond to basic motion of the user (acceleration) and the changing properties of their environment (sound profile and altitude). Table 6 contains the results of re-training the four classifiers (as described in section 4) using only the accelerometer. As we can see, using a single modality, even though it is the most important one, causes a rather large decline in precision/recall for all the different locations. However, if we include

accelerometer, audio, and barometric pressure sensors (as in Table 7) we obtain precision/recall numbers that are very comparable to the results using more of the sensors shown in Figures 2 and 3.

Reducing the number of sensors not only makes the system less susceptible to environmental changes but also enables us to make more practical devices. As we add more and more sensors to our device the system complexity, computational/power requirements, and cost increase. Reducing the number of sensors also reduces the size requirements for incorporating these sensors into a mobile platform, such as a cell phone. Many cell phone manufacturers are starting to add accelerometers to their devices to enable games, interfaces, and basic pedometer functionality, and they already contain an audio sensor. The barometric pressure sensor is typically found in GPS units, where it is used to adjust altitude calculations. However, given the usefulness of barometric pressure sensors in activity recognition and their potential usefulness in localization tasks, they could easily become a common feature of cell phones.

Table 6. Summary statistics for the static and HMM classifiers trained/tested on all locations *(top row)* and a single location *(bottom rows)*. The classifiers trained here only used a single sensor, the accelerometer. The overall accuracies here were all approximately 65%

	Static Classifier		HMM Classifier	
	Overall Precision	Overall Recall	Overall Precision	Overall Recall
Trained on Location 1,2,3 *(all locations)* **Tested on Location 1,2,3 *(all locations)***	40.49%	32.63%	41.15%	38.96%
Trained on Location 1 *(shoulder)* **Tested on Location 1 *(shoulder)***	41.12%	*36.15%*	45.78%	*42.46%*
Trained on Location 2 *(waist)* **Tested on Location 2 *(waist)***	39.61%	*33.67%*	38.67%	*38.30%*
Trained on Location 3 *(wrist)* **Tested on Location 3 *(wrist)***	39.58%	*34.41%*	45.81%	*45.10%*
Single Location Average:	*40.10%*	*34.74%*	*43.42%*	*41.95%*

Table 7. Summary statistics for the static and HMM classifiers trained/tested on all locations *(top row)* and a single location *(bottom rows)*. The classifiers trained here use three sensors: the accelerometer, audio, and barometric pressure. The overall accuracies here were all approximately 90%

	Static Classifier		HMM Classifier	
	Overall Precision	Overall Recall	Overall Precision	Overall Recall
Trained on Location 1,2,3 *(all locations)* **Tested on Location 1,2,3 *(all locations)***	75.53%	68.45%	81.97%	81.38%
Trained on Location 1 *(shoulder)* **Tested on Location 1 *(shoulder)***	76.26%	*69.48%*	83.52%	*82.49%*
Trained on Location 2 *(waist)* **Tested on Location 2 *(waist)***	78.29%	*73.88%*	85.50%	*84.60%*
Trained on Location 3 *(wrist)* **Tested on Location 3 *(wrist)***	76.44%	*56.33%*	85.94%	*85.76%*
Single Location Average:	*77.00%*	*66.56%*	*84.99%*	*84.28%*

7 Conclusion

The work presented here further demonstrates that accurate recognition of a range of physical activities can be achieved by a light-weight and unobtrusive multi-modal wearable device. The experimental results presented in sections 4, 5, and 6 demonstrate the promise of this approach. Specifically, our contribution is that the activity recognition approach is accurate even under practical considerations such as where the device is worn, that it can be pre-trained by other individuals, and that we can use fewer and cheaper sensors. We have answered the three questions we posed in the beginning:

(i) Our single-board activity recognition system generalizes well and we do not need to learn location-specific activity models. As long as we have training data from the different locations we expect to carry the sensing device, the algorithm can pick the right subset of discriminatory features that will work for all those locations (as in [6]).

(ii) For our dataset, we do not require customization to specific individuals and the system works reliably when tested on data from a novel individual. However, we need activities collected over longer periods of time and over people of differing ages, body types, and in more varied settings to get a better understanding of how broadly this result will hold.

(iii) Although our sensor-board was equipped with seven different modalities, we found that three modalities in particular yielded the most discriminative information for our activities: the audio, barometric pressure, and accelerometer sensors. These three modalities provide complementary information about the environment and the wearer. The audio captures the sounds produced during the various activities, whereas the accelerometers data is sensitive to the movement of the body, and the barometric pressure provides important movement queues, such as detecting the activity of riding in an elevator or moving up and down stairs.

Some important things to keep in mind about the results are that the 12 individuals in this paper were all young healthy individuals, so while the results can generalize to wider populations they do not necessarily cover all possible groups, like elderly patients or people with health impairments. In addition, of the 12 hours of data we collected, only 7 hours of it was labeled as belonging to our 8 classes; which raises an interesting problem of how to handle activities that don't quite fall into our labeled activities. We could add an 'other' class to handle these activities; however, to handle ambiguities associated with compound activities like 'sitting' and 'driving a car' we would need to create a hierarchy of activities. This would allow us to recognize an activity like 'driving a car' as both 'driving a car' and 'sitting' and more gracefully handle unrecognized activities.

The fact that a small, single sensor device can classify activities reliably opens the doors to several lines of further exploration. We are currently working on an embedded version of our system that incorporates the sensing and the feature processing on a cell-phone like platform. Embedding the device in such widespread consumer electronics would make such computing truly ubiquitous. We believe the most promising potential applications of such activity recognition systems are in

health-care and health maintenance. For many years, a successful and often life-saving device for the elderly has been a simple "help button" that can be used to initiate a phone call after a disabling fall. For health maintenance and personal fitness, commercially available devices like pedometers are also gaining popularity, but they often lack the subtlety or range of activities performed throughout the day. For example, a pedometer displays the same step count for steps taken on a flat surface as it does for steps walked up the stairs, despite the fact that walking upstairs takes much more effort. The fact that such a simple devices can still be beneficial gives a glimpse of the potentially vast health-care benefits that wireless, unobtrusive activity-recognizing sensors could bring about.

Acknowledgements

References

1. Morris, M., Lundell, J., Dishman, E., Needham, B.: New Perspectives on Ubiquitous Computing from Ethnographic Study of Elders with Cognitive Decline. In: Proc. Ubicomp (2003)
2. Lawton, M. P.: Aging and Performance of Home Tasks. Human Factors (1990)
3. Consolvo, S., Roessler, P., Shelton, B., LaMarcha, A., Schilit, B., Bly, S.: Technology for Care Networks of Elders. In: Proc. IEEE Pervasive Computing Mobile and Ubiquitous Systems: Successful Aging (2004)
4. Kern, N., Schiele, B., Schmidt, A.: Multi-Sensor Activity Context Detection for Wearable Computing. In: Proc. Proc. EUSAI, LNCS (2003) 220-232
5. Bao, L., Intille, S.: Activity Recognition from User-Annotated Acceleration Data. In: Proc. Proc. Pervasive (2004) 1-17
6. Lukowicz, P., Junker, H., Stäger, M., T. von, B., Tröster, G.: WearNET: A Distributed Multi-sensor System for Context Aware Wearables. In: Proceedings of the 4th international conference on Ubiquitous Computing. Springer-Verlag (2002) 361-370
7. Choudhury, T., Lester, J., Kern, N., Borriello, G., Hannaford, B.. A Hybrid Discriminative/Generative Approach for Modeling Human Activities. 19th International Joint Conference on Artificial Intelligence (IJCAI), Edinburgh, Scotland. (2005)
8. Park, S., Locher, I., Savvides, A., Srivastava, M., Chen, A., Muntz, R., Yuen, S.: Design of a Wearable Sensor Badge for Smart Kindergarten. In: Proc. 6th International Symposium on Wearable Computers (2002) 231-238
9. Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D., Anderson, J.: Wireless Sensor Networks for Habitat Monitoring. In: Proc. Proceedings of the 1st ACM

International Workshop on Wireless Sensor Networks and Applications (2002) 88-97

10. Smailagic, A., Currens, B., Maurer, U., Rowe, A.: eWatch.[Online]. Available: http://flat-earth.ece.cmu.edu/~eWatch/
11. Rubinstein, Y. D., Hastie, T.: Discriminative vs. informative learning. In: Proc. In the Proceedings of Knowledge Discovery and Data Mining (1997) 49-53
12. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proc. Computer Vision and Pattern Recognition (2001)
13. Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S.: Boosting the margin: a new explanation for the effectiveness of voting methods. In: Proc. (1997) 322--330
14. Oliver, N., Horvitz, E.: Selective Perception Policies for Limiting Computation in Multimodal Systems: A Comparative Analysis. In: Proc. Proceedings of Int. Conf. on Multimodal Interfaces (2003)
15. Jaakkola, T., Haussler: Exploiting generative models in discriminative classifiers. In: Proc. In Advances in Neural Information Processing Systems (1999)
16. Zhang, F., Pi-Sunyer, F. X., Boozer, C. N.: Improving Energy Expenditure Estimation for Physical Activity. In: Medicine and Science in Sports and Exercise (2004) 883-889.