Activity Recognition from User-Annotated Acceleration Data

Ling Bao and Stephen S. Intille

Massachusetts Institute of Technology
1 Cambridge Center, 4FL
Cambridge, MA 02142 USA
intille@mit.edu

Abstract. In this work, algorithms are developed and evaluated to detect physical activities from data acquired using five small biaxial accelerometers worn simultaneously on different parts of the body. Acceleration data was collected from 20 subjects without researcher supervision or observation. Subjects were asked to perform a sequence of everyday tasks but not told specifically where or how to do them. Mean, energy, frequency-domain entropy, and correlation of acceleration data was calculated and several classifiers using these features were tested. Decision tree classifiers showed the best performance recognizing everyday activities with an overall accuracy rate of 84%. The results show that although some activities are recognized well with subject-independent training data, others appear to require subject-specific training data. The results suggest that multiple accelerometers aid in recognition because conjunctions in acceleration feature values can effectively discriminate many activities. With just two biaxial accelerometers – thigh and wrist – the recognition performance dropped only slightly. This is the first work to investigate performance of recognition algorithms with multiple, wire-free accelerometers on 20 activities using datasets annotated by the subjects themselves.

1 Introduction

One of the key difficulties in creating useful and robust ubiquitous, context-aware computer applications is developing the algorithms that can detect context from noisy and often ambiguous sensor data. One facet of the user's context is his physical activity. Although prior work discusses physical activity recognition using acceleration (e.g. [17,5,23]) or a fusion of acceleration and other data modalities (e.g. [18]), it is unclear how most prior systems will perform under real-world conditions. Most of these works compute recognition results with data collected from subjects under artificially constrained laboratory settings. Some also evaluate recognition performance on data collected in natural, out-of-lab settings but only use limited data sets collected from one individual (e.g. [22]). A number of works use naturalistic data but do not quantify recognition accuracy. Lastly, research using naturalistic data collected from multiple subjects has focused on

recognition of a limited subset of nine or fewer everyday activities consisting largely of ambulatory motions and basic postures such as sitting and standing (e.g. [10,5]). It is uncertain how prior systems will perform in recognizing a variety of everyday activities for a diverse sample population under real-world conditions.

In this work, the performance of activity recognition algorithms under conditions akin to those found in real-world settings is assessed. Activity recognition results are based on acceleration data collected from five biaxial accelerometers placed on 20 subjects under laboratory and semi-naturalistic conditions. Supervised learning classifiers are trained on labeled data that is acquired without researcher supervision from subjects themselves. Algorithms trained using only user-labeled data might dramatically increase the amount of training data that can be collected and permit users to train algorithms to recognize their own individual behaviors.

2   Background

Researchers have already prototyped wearable computer systems that use acceleration, audio, video, and other sensors to recognize user activity (e.g. [7]). Advances in miniaturization will permit accelerometers to be embedded within wrist bands, bracelets, adhesive patches, and belts and to wirelessly send data to a mobile computing device that can use the signals to recognize user activities.

For these applications, it is important to train and test activity recognition systems on data collected under naturalistic circumstances, because laboratory environments may artificially constrict, simplify, or influence subject activity patterns. For instance, laboratory acceleration data of walking displays distinct phases of a consistent gait cycle which can aide recognition of pace and incline [2]. However, acceleration data from the same subject outside of the laboratory may display marked fluctuation in the relation of gait phases and total gait length due to decreased self-awareness and fluctuations in traffic. Consequently, a highly accurate activity recognition algorithm trained on data where subjects are told exactly where or how to walk (or where the subjects are the researchers themselves) may rely too heavily on distinct phases and periodicity of accelerometer signals found only in the lab. The accuracy of such a system may suffer when tested on naturalistic data, where there is greater variation in gait pattern.

Many past works have demonstrated 85% to 95% recognition rates for ambulation, posture, and other activities using acceleration data. Some are summarized in Figure 1 (see [3] for a summary of other work). Activity recognition has been performed on acceleration data collected from the hip (e.g. [17,19]) and from multiple locations on the body (e.g. [5,14]). Related work using activity counts and computer vision also supports the potential for activity recognition using acceleration. The energy of a subject's acceleration can discriminate sedentary activities such as sitting or sleeping from moderate intensity activities such as walking or typing and vigorous activities such as running [25]. Recent work

Ref.	Recognition Accuracy	Activities Recognized	No. Subj.	Data Type	No. Sensors	Sensor Placement
[17]	92.85% to 95.91%	ambulation	8	L	2	2 thigh
[19]	83% to 90%	ambulation, posture	6	L	6	3 left hip, 3 right hip
[10]	95.8%	ambulation, posture, typing, talking, bicycling	24	L	4	chest, thigh, wrist, forearm
[10]	66.7%	ambulation, posture, typing, talking, bicycling	24	N	4	chest, thigh, wrist, forearm
[1]	89.30%	ambulation, posture	5	L	2	chest, thigh
[12]	N/A	walking speed, incline	20	L	4	3 lower back 1 ankle
[22]	86% to 93%	ambulation, posture, play	1	N	3	2 waist, 1 thigh
[14]	≈65% to ≈95%	ambulation, typing, stairs shake hands, write on board	1	L	up to 36	all major joints
[6]	96.67%	3 Kung Fu arm movements	1	L	2	2 wrist
[23]	42% to 96%	ambulation, posture, bicycling	1	L	2	2 lower back
[20]	85% to 90%	ambulation, posture	10	L	2	2 knee

Fig. 1. Summary of a representative sample of past work on activity recognition using acceleration. The "No. Subj." column specifies the number of subjects who participated in each study, and the "Data Type" column specifies whether data was collected under laboratory (L) or naturalistic (N) settings. The "No. Sensors" column specifies the number of uniaxial accelerometers used per subject.

with 30 wired accelerometers spread across the body suggests that the addition of sensors will generally improve recognition performance [24].

Although the literature supports the use of acceleration for physical activity recognition, little work has been done to validate the idea under real-world circumstances. Most prior work on activity recognition using acceleration relies on data collected in controlled laboratory settings. Typically, the researcher collected data from a very small number of subjects, and often the subjects have included the researchers themselves. The researchers then hand-annotated the collected data. Ideally, data would be collected in less controlled settings without researcher supervision. Further, to increase the volume of data collected, subjects would be capable of annotating their own data sets. Algorithms that could be trained using only user-labeled data might dramatically increase the amount of training data that can be collected and permit users to train algorithms to recognize their own individual behaviors. In this work we assume that labeled training data is required for many automatic activity recognition tasks. We note, however, that one recent study has shown that unsupervised learning

can be used to cluster accelerometer data into categories that, in some instances, map onto meaningful labels [15].

The vast majority of prior work focuses on recognizing a special subset of physical activities such as ambulation, with the exception of [10] which examines nine everyday activities. Interestingly, [10] demonstrated 95.8% recognition rates for data collected in the laboratory but recognition rates dropped to 66.7% for data collected outside the laboratory in naturalistic settings. These results demonstrate that the performance of algorithms tested only on laboratory data or data acquired from the experimenters themselves may suffer when tested on data collected under less-controlled (i.e. naturalistic) circumstances.

Prior literature demonstrates that forms of locomotion such as walking, running, and climbing stairs and postures such as sitting, standing, and lying down can be recognized at 83% to 95% accuracy rates using hip, thigh, and ankle acceleration (see Figure 1). Acceleration data of the wrist and arm are known to improve recognition rates of upper body activities [6,10] such as typing and martial arts movements. All past works with multiple accelerometers have used accelerometers connected with wires, which may restrict subject movement. Based on these results, this work uses data collected from five wire-free biaxial accelerometers placed on each subject's right hip, dominant wrist, non-dominant upper arm, dominant ankle, and non-dominant thigh to recognize ambulation, posture, and other everyday activities. Although each of the above five locations have been used for sensor placement in past work, no work addresses which of the accelerometer locations provide the best data for recognizing activities even though it has been suggested that for some activities that more sensors improve recognition [24]. Prior work has typically been conducted with only 1-2 accelerometers worn at different locations on the body, with only a few using more than 5 (e.g. [19,14,24]).

3   Design

Subjects wore 5 biaxial accelerometers as they performed a variety of activities under two different data collection protocols.

3.1   Accelerometers

Subject acceleration was collected using ADXL210E accelerometers from Analog Devices. These two-axis accelerometers are accurate to ±10 G with tolerances within 2%. Accelerometers were mounted to *hoarder boards* [11], which sampled at 76.25 Hz (with minor variations based on onboard clock accuracy) and stored acceleration data on compact flash memory. This sampling frequency is more than sufficient compared to the 20 Hz frequency required to assess daily physical activity [4]. The hoarder board time stamped one out of every 100 acceleration samples, or one every 1.31 seconds. Four AAA batteries can power the hoarder board for roughly 24 hours. This is more than sufficient for the 90 minute data collection sessions used in this study. A hoarder board is shown in Figure 2a.
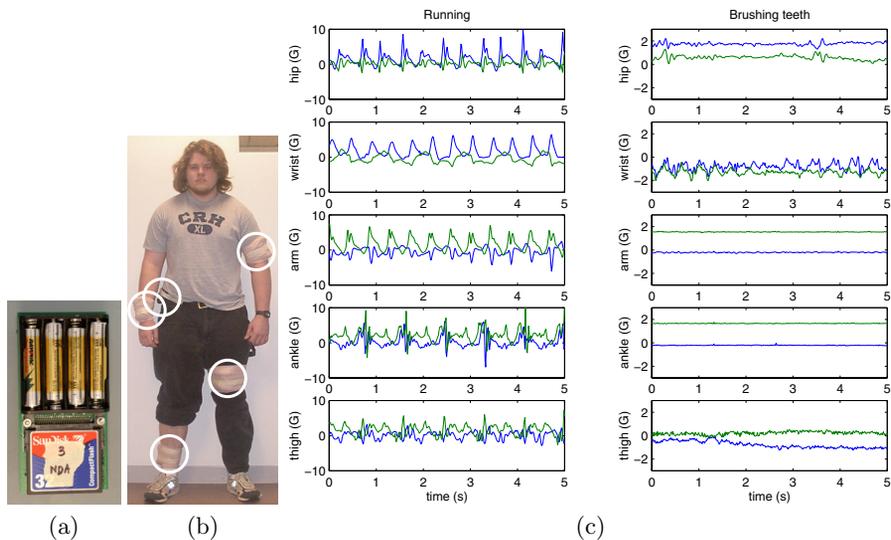
Fig. 2. (a) Hoarder data collection board, which stored data from a biaxial accelerometer. The biaxial accelerometers are attached to the opposite side of the board. (b) Hoarder boards were attached to 20 subjects on the 4 limb positions shown here (held on with medical gauze), plus the right hip. (c) Acceleration signals from five biaxial accelerometers for walking, running, and tooth brushing.

Previous work shows promising activity recognition results from ± 2 G acceleration data (e.g. [9,14]) even though typical body acceleration amplitude can range up to 12 G [4]. However, due to limitations in availability of ± 12 G accelerometers, ± 10 G acceleration data was used. Moreover, although body limbs and extremities can exhibit a 12 G range in acceleration, points near the torso and hip experience a 6 G range in acceleration [4].

The hoarder boards were not electronically synchronized to each other and relied on independent quartz clocks to time stamp data. Electronic synchronization would have required wiring between the boards which, even when the wiring is carefully designed as in [14], would restrict subject movements, especially during whole body activities such as bicycling or running. Further, we have found subjects wearing wiring feel self-conscious when outside of the laboratory and therefore restrict their behavior.

To achieve synchronization without wires, hoarder board clocks were synchronized with subjects' watch times at the beginning of each data collection session. Due to clock skew, hoarder clocks and the watch clock drifted between 1 and 3 seconds every 24 hours. To minimize the effects of clock skew, hoarder boards were shaken together in a fixed sinusoidal pattern in two axes of acceleration at the beginning and end of each data collection session. Watch times were manually recorded for the periods of shaking. The peaks of the distinct sinusoidal patterns at the beginning and end of each acceleration signal were

visually aligned between the hoarder boards. Time stamps during the shaking period were also shifted to be consistent with the recorded watch times for shaking. Acceleration time stamps were linearly scaled between these manually aligned start and end points.

To characterize the accuracy of the synchronization process, three hoarder boards were synchronized with each other and a digital watch using the above protocol. The boards were then shaken together several times during a full day to produce matching sinusoidal patterns on all boards. Visually comparing the peaks of these matching sinusoids across the three boards showed mean skew of 4.3 samples with a standard deviation of 1.8 samples between the boards. At a sampling frequency of 76.25 Hz, the skew between boards is equivalent to $.0564 \pm .0236$ s.

A T-Mobile Sidekick phone pouch was used as a carrying case for each hoarder board. The carrying case was light, durable, and provided protection for the electronics. A carrying case was secured to the subject's belt on the right hip. All subjects were asked to wear clothing with a belt. Elastic medical bandages were used to wrap and secure carrying cases at sites other than the hip. Typical placement of hoarder boards is shown in Figure 2b. Figure 2c shows acceleration data collected for walking, running, and tooth brushing from the five accelerometers.

No wires were used to connect the hoarder boards to each other or any other devices. Each hoarder in its carrying case weighed less than 120 g. Subjects could engage in vigorous, complex activity without any restriction on movement or fear of damaging the electronics. The sensors were still visually noticeable. Subjects who could not wear the devices under bulky clothing did report feeling self conscious in public spaces.

3.2   Activity Labels

Twenty activities were studied. These activities are listed in Figure 5. The 20 activities were selected to include a range of common everyday household activities that involve different parts of the body and range in level of intensity. Whole body activities such as walking, predominantly arm-based activities such as brushing of teeth, and predominantly leg-based activities such as bicycling were included as were sedentary activities such as sitting, light intensity activities such as eating, moderate intensity activities such as window scrubbing, and vigorous activities such as running. Activity labels were chosen to reflect the content of the actions but do not specify the style. For instance, "walking" could be parameterized by walking speed and quantized into slow and brisk or other categories.

3.3   Semi-naturalistic, User-Driven Data Collection

The most realistic training and test data would be naturalistic data acquired from subjects as they go about their normal, everyday activities. Unfortunately,

obtaining such data requires direct observation of subjects by researchers, subject self-report of activities, or use of the experience sampling method [8] to label subject activities for algorithm training and testing. Direct observation can be costly and scales poorly for the study of large subject populations. Subject self-report recall surveys are prone to recall errors [8] and lack the temporal precision required for training activity recognition algorithms. Finally, the experience sampling method requires frequent interruption of subject activity, which agitates subjects over an extended period of time. Some activities we would like to develop recognition algorithms for, such as folding laundry, riding escalators, and scrubbing windows, may not occur on a daily basis. A purely naturalistic protocol would not capture sufficient samples of these activities for thorough testing of recognition systems without prohibitively long data collection periods.

In this work we compromise and use a semi-naturalistic collection protocol that should permit greater subject variability in behavior than laboratory data. Further, we show how training sets can be acquired from subjects themselves without the direct supervision of a researcher, which may prove important if training data must be collected by end users to improve recognition performance.

For semi-naturalistic data collection, subjects ran an obstacle course consisting of a series of activities listed on a worksheet. These activities were disguised as goals in an obstacle course to minimize subject awareness of data collection. For instance, subjects were asked to "use the web to find out what the world's largest city in terms of population is" instead of being asked to "work on a computer." Subjects recorded the time they began each obstacle and the time they completed each obstacle. Subjects completed each obstacle on the course ensuring capture of all 20 activities being studied. There was no researcher supervision of subjects while they collected data under the semi-naturalistic collection protocol. As subjects performed each of these obstacles in the order given on their worksheet, they labeled the start and stop times for that activity and made any relevant notes about that activity. Acceleration data collected between the start and stop times were labeled with the name of that activity. Subjects were free to rest between obstacles and proceed through the worksheet at their own pace as long as they performed obstacles in the order given. Furthermore, subjects had freedom in how they performed each obstacle. For example, one obstacle was to "read the newspaper in the common room. Read the entirety of at least one non-frontpage article." The subject could choose which and exactly how many articles to read. Many activities were performed outside of the lab. Subjects were not told where or how to perform activities and could do so in a common room within the lab equipped with a television, vacuum, sofa, and reading materials or anywhere they preferred. No researchers or cameras monitored the subjects.

3.4   Specific Activity Data Collection

After completing the semi-naturalistic obstacle course, subjects underwent another data collection session to collect data under somewhat more controlled conditions. Linguistic definitions of activity are often ambiguous. The activity
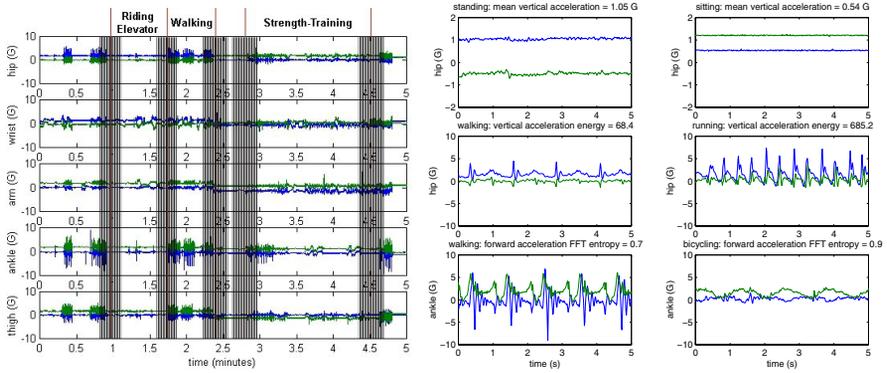
Fig. 3. (a) Five minutes of 2-axis acceleration data annotated with subject self-report activity labels. Data within 10s of self-report labels is discarded as indicated by masking. (b) Differences in feature values computed from FFTs are used to discriminate between different activities.

"scrubbing," for example, can be interpreted as window scrubbing, dish scrubbing, or car scrubbing. For this data collection session, subjects were therefore given short definitions of the 20 activity labels that resolved major ambiguities in the activity labels while leaving room for interpretation so that subjects could show natural, individual variations in how they performed activities. For example, walking was described as "walking without carrying any items in you hand or on your back heavier than a pound" and scrubbing was described as "using a sponge, towel, or paper towel to wipe a window." See [3] for descriptions for all 20 activities.

Subjects were requested to perform random sequences of the 20 activities defined on a worksheet during laboratory data collection. Subjects performed the sequence of activities given at their own pace and labeled the start and end times of each activity. For example, the first 3 activities listed on the worksheet might be "bicycling," "riding elevator," and "standing still." The researcher's definition of each of these activities was provided. As subjects performed each of these activities in the order given on their worksheet, they labeled the start and stop times for that activity and made any relevant notes about that activity such as "I climbed the stairs instead of using the elevator since the elevator was out of service." Acceleration data collected between the start and stop times were labeled with the name of that activity. To minimize mislabeling, data within 10 s of the start and stop times was discarded. Since the subject is probably standing still or sitting while he records the start and stop times, the data immediately around these times may not correspond to the activity label. Figure 3a shows acceleration data annotated with subject self-report labels.

Although data collected under this second protocol is more structured than the first, it was still acquired under less controlled conditions than in most prior work. Subjects, who were not the researchers, could perform their activities

anywhere including outside of the laboratory. Also, there was no researcher supervision during the data collection session.

3.5   Feature Computation

Features were computed on 512 sample windows of acceleration data with 256 samples overlapping between consecutive windows. At a sampling frequency of 76.25 Hz, each window represents 6.7 seconds. Mean, energy, frequency-domain entropy, and correlation features were extracted from the sliding windows signals for activity recognition. Feature extraction on sliding windows with 50% overlap has demonstrated success in past works [9,23]. A window of several seconds was used to sufficiently capture cycles in activities such as walking, window scrubbing, or vacuuming. The 512 sample window size enabled fast computation of FFTs used for some of the features.

The DC feature is the mean acceleration value of the signal over the window. The energy feature was calculated as the sum of the squared discrete FFT component magnitudes of the signal. The sum was divided by the window length for normalization. Additionally, the DC component of the FFT was excluded in this sum since the DC characteristic of the signal is already measured by another feature. Note that the FFT algorithm used produced 512 components for each 512 sample window. Use of mean [10,1] and energy [21] of acceleration features has been shown to result in accurate recognition of certain postures and activities (see Figure 1).

Frequency-domain entropy is calculated as the normalized information entropy of the discrete FFT component magnitudes of the signal. Again, the DC component of the FFT was excluded in this calculation. This feature may support discrimination of activities with similar energy values. For instance, biking and running may result in roughly the same amounts of energy in the hip acceleration data. However, because biking involves a nearly uniform circular movement of the legs, a discrete FFT of hip acceleration in the vertical direction may show a single dominant frequency component at 1 Hz and very low magnitude for all other frequencies. This would result in a low frequency-domain entropy. Running on the other hand may result in complex hip acceleration and many major FFT frequency components between 0.5 Hz and 2 Hz. This would result in a higher frequency-domain entropy.

Features that measure correlation or acceleration between axes can improve recognition of activities involving movements of multiple body parts [12,2]. Correlation is calculated between the two axes of each accelerometer hoarder board and between all pairwise combinations of axes on different hoarder boards.

Figure 3b shows some of these features for two activities. It was anticipated that certain activities would be difficult to discriminate using these features. For example, "watching TV" and "sitting" should exhibit very similar if not identical body acceleration. Additionally, activities such as "stretching" could show marked variation from person to person and for the same person at different times. Stretching could involve light or moderate energy acceleration in the upper body, torso, or lower body.

As discussed in the the next section, several classifiers were tested for activity recognition using the feature vector.

4   Evaluation

Subjects were recruited using posters seeking research study participants for compensation. Posters were distributed around an academic campus and were also emailed to the student population. Twenty subjects from the academic community volunteered. Data was collected from 13 males and 7 females. Subjects ranged in age from 17 to 48 (mean 21.8, sd 6.59).

Each subject participated in two sessions of study. In the first session, subjects wore five accelerometers and a digital watch. Subjects collected the seminaturalistic data by completing an obstacle course worksheet, noting the start and end times of each obstacle on the worksheet. Each subject collected between 82 and 160 minutes of data (mean 104, sd 13.4). Six subjects skipped between one to two obstacles due to factors such as inclement weather, time constraints, or problems with equipment in the common room (e.g. the television, vacuum, computer, and bicycle). Subjects performed each activity on their obstacle course for an average of 156 seconds (sd 50).

In the second session, often performed on a different day, the same subjects wore the same set of sensors. Subjects performed the sequence of activities listed on an activity worksheet, noting the start and end times of these activities. Each subject collected between 54 and 131 minutes of data (mean 96, sd 16.7). Eight subjects skipped between one to four activities due to factors listed earlier.

4.1   Results

Mean, energy, entropy, and correlation features were extracted from acceleration data. Activity recognition on these features was performed using decision table, instance-based learning (IBL or nearest neighbor), C4.5 decision tree, and naive Bayes classifiers found in the Weka Machine Learning Algorithms Toolkit [26].

Classifiers were trained and tested using two protocols. Under the first protocol, classifiers were trained on each subject's activity sequence data and tested on that subject's obstacle course data. This user-specific training protocol was repeated for all twenty subjects. Under the second protocol, classifiers were trained on activity sequence and obstacle course data for all subjects except one. The classifiers were then tested on obstacle course data for the only subject left out of the training data set. This leave-one-subject-out validation process was repeated for all twenty subjects. Mean and standard deviation for classification accuracy under both protocols is summarized in Figure 4.

Overall, recognition accuracy is highest for decision tree classifiers, which is consistent with past work where decision based algorithms recognized lying, sitting, standing and locomotion with 89.30% accuracy [1]. Nearest neighbor is the second most accurate algorithm and its strong relative performance is

Classifier	User-specific Training	Leave-one-subject-out Training
Decision Table	36.32 ± 14.501	46.75 ± 9.296
IBL	69.21 ± 6.822	82.70 ± 6.416
C4.5	71.58 ± 7.438	84.26 ± 5.178
Naive Bayes	34.94 ± 5.818	52.35 ± 1.690

Fig. 4. Summary of classifier results (mean $\pm$ standard deviation) using user-specific training and leave-one-subject-out training. Classifiers were trained on laboratory data and tested on obstacle course data.

Activity	Accuracy	Activity	Accuracy
Walking	89.71	Walking carrying items	82.10
Sitting & relaxing	94.78	Working on computer	97.49
Standing still	95.67	Eating or drinking	88.67
Watching TV	77.29	Reading	91.79
Running	87.68	Bicycling	96.29
Stretching	41.42	Strength-training	82.51
Scrubbing	81.09	Vacuuming	96.41
Folding laundry	95.14	Lying down & relaxing	94.96
Brushing teeth	85.27	Climbing stairs	85.61
Riding elevator	43.58	Riding escalator	70.56

Fig. 5. Aggregate recognition rates (%) for activities studied using leave-one-subject-out validation over 20 subjects.

also supported by past prior work where nearest neighbor algorithms recognized ambulation and postures with over 90% accuracy [16,10].

Figure 5 shows the recognition results for the C4.5 classifier. Rule-based activity recognition appears to capture conjunctions in feature values that may lead to good recognition accuracy. For instance, the C4.5 decision tree classified sitting as an activity having nearly 1 G downward acceleration and low energy at both hip and arm. The tree classified bicycling as an activity involving moderate energy levels and low frequency-domain entropy at the hip and low energy levels at the arm. The tree distinguishes "window scrubbing" from "brushing teeth" because the first activity involves more energy in hip acceleration even though both activities show high energy in arm acceleration. The fitting of probability distributions to acceleration features under a Naive Bayesian approach may be unable to adequately model such rules due to the assumptions of conditional independence between features and normal distribution of feature values, which may account for the weaker performance. Furthermore, Bayesian algorithms may require more data to accurately model feature value distributions.

Figure 6 shows an aggregate confusion matrix for the C4.5 classifier based on all 20 trials of leave-one-subject-out validation. Recognition accuracies for stretching and riding an elevator were below 50%. Recognition accuracies for

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	< classified as
942	46	0	0	2	0	0	0	8	3	8	1	4	2	7	0	3	8	8	8	a = walking
83	**1183**	9	0	3	2	0	0	8	1	3	8	14	1	16	0	8	53	38	11	b = walking/carry
0	9	**762**	11	0	1	17	3	0	0	0	0	0	0	0	1	0	0	0	0	c = sitting relaxed
0	0	10	**893**	9	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	d = computer work
0	0	0	7	**774**	11	0	0	0	6	1	2	2	0	4	0	2	0	0	0	e = standing still
0	2	1	0	12	**712**	9	1	0	0	2	1	10	1	18	0	26	1	4	3	f = eating/drinking
0	0	42	21	0	1	**320**	28	0	0	0	0	0	0	0	0	0	0	0	1	g = watching TV
0	0	23	1	1	6	16	**961**	9	0	2	0	0	1	0	1	2	0	2	22	h = reading
14	12	0	0	1	0	0	0	17	**491**	10	1	1	1	1	0	1	3	4	1	i = running
0	1	0	0	5	0	0	0	8	**830**	10	0	1	0	3	0	2	1	0	1	j = bicycling
9	3	2	16	30	22	45	9	3	35	**309**	37	26	21	99	1	38	12	3	26	k = stretching
4	10	0	0	6	5	2	7	0	6	23	**500**	13	2	9	3	6	5	3	2	l = strength train
1	7	0	0	5	10	0	0	0	3	9	1	**403**	11	10	1	26	1	6	4	m = scrubbing
1	0	0	0	0	3	1	0	0	2	0	1	9	**885**	11	0	1	0	2	2	n = vacuuming
1	1	0	0	1	6	0	0	0	1	4	1	4	7	**822**	8	4	0	1	3	o = folding laundry
0	0	4	9	0	2	1	7	0	0	0	0	1	0	10	**791**	8	0	0	0	p = lying down
1	2	0	0	3	32	0	0	0	1	5	0	18	7	10	9	**637**	10	2	10	q = brushing teeth
7	14	0	0	1	1	0	0	0	3	2	1	1	0	2	0	12	**351**	10	5	r = climbing stairs
84	70	0	7	20	60	0	0	8	40	33	11	24	34	40	0	0	59	**502**	160	s = riding elevator
5	2	0	0	5	6	0	1	0	1	0	3	3	1	0	0	3	7	16	**127**	t = riding escalator

Fig. 6. Aggregate confusion matrix for C4.5 classifier based on leave-one-subject-out validation for 20 subjects, tested on semi-naturalistic data.

"watching TV" and "riding escalator" were 77.29% and 70.56%, respectively. These activities do not have simple characteristics and are easily confused with other activities. For instance, "stretching" is often misclassified as "folding laundry" because both may involve the subject moving the arms at a moderate rate. Similarly, "riding elevator" is misclassified as "riding escalator" since both involve the subject standing still. "Watching TV" is confused with "sitting and relaxing" and "reading" because all the activities involve sitting. "Riding escalator" is confused with "riding elevator" since the subject may experience similar vertical acceleration in both cases. "Riding escalator" is also confused with "climbing stairs" since the subject sometimes climbs the escalator stairs.

Recognition accuracy was significantly higher for all algorithms under the leave-one-subject-out validation process. This indicates that the effects of individual variation in body acceleration may be dominated by strong commonalities between people in activity pattern. Additionally, because leave-one-subject-out validation resulted in larger training sets consisting of data from 19 subjects, this protocol may have resulted in more generalized and robust activity classifiers. The markedly smaller training sets used for the user-specific training protocol may have limited the accuracy of classifiers.

To control for the effects of sample size in comparing leave-one-subject-out and user-specific training, preliminary results were gathered using a larger training data set collected for three subjects. These subjects were affiliates of the researchers (unlike the 20 primary subjects). Each of these subjects participated in one semi-naturalistic and five laboratory data collection sessions. The C4.5 decision tree algorithm was trained for each individual using data collected from all five of his laboratory sessions and tested on the semi-naturalistic data. The algorithm was also trained on five laboratory data sets from five random subjects other than the individual and tested on the individual's semi-naturalistic data. The results are compared in Figure 7. In this case, user-specific training resulted in an increase in recognition accuracy of 4.32% over recognition rates for leave-one-subject-out-training. This difference shows that given equal amounts of training data, training on user-specific training data can result in classifiers

Classifier	User-specific Training	Leave-one-subject-out Training
C4.5	77.31 ± 4.328	72.99 ± 8.482

Fig. 7. Summary of classifier results (mean $\pm$ standard deviation) using user-specific training and leave-one-subject-out training where both training data sets are equivalent to five laboratory data sessions.

that recognize activities more accurately than classifiers trained on example data from many people. However, the certainty of these conclusions is limited by the low number of subjects used for this comparison and the fact that the three individuals studied were affiliates of the researchers. Nonetheless, these initial results support the need for further study of the power of user-specific versus generalized training sets.

The above results suggest that real-world activity recognition systems can rely on classifiers that are pre-trained on large activity data sets to recognize some activities. Although preliminary results show that user-specific training can lead to more accurate activity recognition given large training sets, pre-trained systems offer greater convenience. Pre-trained systems could recognize many activities accurately without requiring training on data from their user, simplifying the deployment of these systems. Furthermore, since the activity recognition system needs to be trained only once before deployment, the slow running time for decision tree training is not an obstacle. Nonetheless, there may be limitations to a pre-trained algorithm. Although activities such as "running" or "walking" may be accurately recognized, activities that are more dependent upon individual variation and the environment (e.g. "stretching") may require person-specific training [13]).

To evaluate the discriminatory power of each accelerometer location, recognition accuracy using the decision tree classifier (the best performing algorithm) was also computed using a leave-one-accelerometer-in protocol. Specifically, recognition results were computed five times, each time using data from only one of the five accelerometers for the training and testing of the algorithm. The differences in recognition accuracy rates using this protocol from accuracy rates obtained from all five accelerometers are summarized in Figure 8. These results show that the accelerometer placed on the subject's thigh is the most powerful for recognizing this set of 20 activities. Acceleration of the dominant wrist is more useful in discriminating these activities than acceleration of the non-dominant arm. Acceleration of the hip is the second best location for activity discrimination. This suggests that an accelerometer attached to a subject's cell phone, which is often placed at a fixed location such as on a belt clip, may enable recognition of certain activities.

Confusion matrices resulting from leave-one-accelerometer-in testing [3] show that data collected from lower body accelerometers placed on the thigh, hip, and ankle is generally best at recognizing forms of ambulation and posture. Ac-

Accelerometer(s) Left In	Difference in Recognition Accuracy
Hip	-34.12 ± 7.115
Wrist	-51.99 ± 12.194
Arm	-63.65 ± 13.143
Ankle	-37.08 ± 7.601
Thigh	-29.47 ± 4.855
Thigh and Wrist	-3.27 ± 1.062
Hip and Wrist	-4.78 ± 1.331

Fig. 8. Difference in overall recognition accuracy (mean $\pm$ standard deviation) due to leaving only one or two accelerometers in. Accuracy rates are aggregated for 20 subjects using leave-one-subject-out validation.

celerometer data collected from the wrist and arm is better at discriminating activities involving characteristic upper body movements such as reading from watching TV or sitting and strength-training (push ups) from stretching. To explore the power of combining upper and lower body accelerometer data, data from thigh and wrist accelerometers and hip and wrist accelerometers were also used and results are shown in Figure 8. Note that recognition rates improved over 25% for the leave-two-accelerometers-in results as compared to the best leave-one-accelerometer-in results. Of the two pairs tested, thigh and wrist acceleration data resulted in the highest recognition accuracy. However, both thigh and wrist and hip and wrist pairs showed less than a 5% decrease in recognition rate from results using all five accelerometer signals. This suggests that effective recognition of certain everyday activities can be achieved using two accelerometers placed on the wrist and thigh or wrist and hip. Others have also found that for complex activities at least one sensor on the lower and upper body is desirable [14].[1]

4.2   Analysis

This work shows that user-specific training is not necessary to achieve recognition rates for some activities of over 80% for 20 everyday activities. Classification accuracy rates of between 80% to 95% for walking, running, climbing stairs, standing still, sitting, lying down, working on a computer, bicycling, and vacuuming are comparable with recognition results using laboratory data from previous works. However, most prior has used data collected under controlled laboratory conditions to achieve their recognition accuracy rates, typically where data is hand annotated by a researcher. The 84.26% overall recognition rate achieved in this work is significant because study subjects could move about freely outside the lab without researcher supervision while collecting and annotating their own

---

[1] Only the decision tree algorithm was used to evaluate the information content of specific sensors, leaving open the possibility that other algorithms may perform better with different sensor placements.

semi-naturalistic data. This is a step towards creating mobile computing systems that work outside of the laboratory setting.

The C4.5 classifier used mean acceleration to recognize postures such as sitting, standing still, and lying down. Ambulatory activities and bicycling were recognized by the level of hip acceleration energy. Frequency-domain entropy and correlation between arm and hip acceleration strongly distinguished bicycling, which showed low entropy hip acceleration and low arm-hip correlation, from running, which displayed higher entropy in hip acceleration and higher arm-hip movement correlation. Both activities showed similar levels of hip acceleration mean and energy. Working on a computer, eating or drinking, reading, strength-training as defined by a combination of sit ups and push-ups, window scrubbing, vacuuming, and brushing teeth were recognized by arm posture and movement as measured by mean acceleration and energy.

Lower recognition accuracies for activities such as stretching, scrubbing, riding an elevator, and riding an escalator suggest that higher level analysis is required to improve classification of these activities. Temporal information in the form of duration and time and day of activities could be used to detect activities. For instance, standing still and riding an elevator are similar in terms of body posture. However, riding an elevator usually lasts for a minute or less whereas standing still can last for a much longer duration. By considering the duration of a particular posture or type of body acceleration, these activities could be distinguished from each other with greater accuracy. Similarly, adults may be more likely to watch TV at night than at other times on a weekday. Thus, date and time or other multi-modal sensing could be used to improve discrimination of watching TV from simply sitting and relaxing. However, because daily activity patterns may vary dramatically across individuals, user-specific training may be required to effectively use date and time information for activity recognition.

The decision tree algorithm used in this work can recognize the content of activities, but may not readily recognize activity style. Although a decision tree algorithm could potentially recognize activity style using a greater number of labels such as "walking slowly," "walking briskly," "scrubbing softly," or "scrubbing vigorously," the extensibility of this technique is limited. For example, the exact pace of walking cannot be recognized using any number of labels. Other techniques may be required to recognize parameterized activity style.

Use of other sensor data modalities may further improve activity recognition. Heart rate data could be used to augment acceleration data to detect intensity of physical activities. GPS location data could be used to infer whether an individual is at home or at work and affect the probability of activities such as working on the computer or lying down and relaxing. Use of such person-specific sensors such as GPS, however, is more likely to require that training data be acquired directly from the individual rather than from a laboratory setting because individuals can work, reside, and shop in totally different locations.

5   Conclusion

Using decision tree classifiers, recognition accuracy of over 80% on a variety of 20 everyday activities was achieved using leave-one-subject-out-validation on data acquired without researcher supervision from 20 subjects. These results are competitive with prior activity recognition results that only used laboratory data. Furthermore, this work shows acceleration can be used to recognize a variety of household activities for context-aware computing. This extends previous work on recognizing ambulation and posture using acceleration (see Figure 1).

This work further suggests that a mobile computer and small wireless accelerometers placed on an individual's thigh and dominant wrist may be able to detect some common everyday activities in naturalistic settings using fast FFT-based feature computation and a decision tree classifier algorithm. Decision trees are slow to train but quick to run. Therefore, a pre-trained decision tree should be able to classify user activities in real-time on emerging mobile computing devices with fast processors and wireless accelerometers.

References

1. K. Aminian, P. Robert, E.E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon. Physical activity monitoring based on accelerometry: validation and comparison with video observation. *Medical & Biological Engineering & Computing*, 37(3):304–8, 1999.
2. K. Aminian, P. Robert, E. Jequier, and Y. Schutz. Estimation of speed and incline of walking using neural network. *IEEE Transactions on Instrumentation and Measurement*, 44(3):743–746, 1995.
3. L. Bao. *Physical Activity Recognition from Acceleration Data under Semi-Naturalistic Conditions.* M.Eng. Thesis, Massachusetts Institute of Technology, 2003.
4. C.V. Bouten, K.T. Koekkoek, M. Verduin, R. Kodde, and J.D. Janssen. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Transactions on Bio-Medical Engineering*, 44(3):136–47, 1997.
5. J.B. Bussmann, W.L. Martens, J.H. Tulen, F.C. Schasfoort, H.J. van den Berg-Emons, and H.J. Stam. Measuring daily behavior using ambulatory accelerometry: the Activity Monitor. *Behavior Research Methods, Instruments, & Computers*, 33(3):349–56, 2001.
6. G.S. Chambers, S. Venkatesh, G.A.W. West, and H.H. Bui. Hierarchical recognition of intentional human gestures for sports video annotation. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 2, pages 1082–1085. IEEE Press, 2002.
7. B.P. Clarkson. *Life Patterns: Structure from Wearable Sensors.* Ph.D. Thesis, Massachusetts Institute of Technology, 2002.
8. M. Csikszentmihalyi and R. Larson. Validity and reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, 175(9):526–36, 1987.

9. R.W. DeVaul and S. Dunn. Real-Time Motion Classification for Wearable Computing Applications. Technical report, MIT Media Laboratory, 2001.
10. F. Foerster, M. Smeja, and J. Fahrenberg. Detection of posture and motion by accelerometry: a validation in ambulatory monitoring. *Computers in Human Behavior*, 15:571–583, 1999.
11. V. Gerasimov. Hoarder Board Specifications, Access date: January 15 2002. http://vadim.www.media.mit.edu/Hoarder/Hoarder.htm.
12. R. Herren, A. Sparti, K. Aminian, and Y. Schutz. The prediction of speed and incline in outdoor running in humans using accelerometry. *Medicine & Science in Sports & Exercise*, 31(7):1053–9, 1999.
13. S.S. Intille, L. Bao, E. Munguia Tapia, and J. Rondoni. Acquiring in situ training data for context-aware ubiquitous computing applications. In *Proceedings of CHI 2004 Connect: Conference on Human Factors in Computing Systems*. ACM Press, 2004.
14. N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *European Symposium on Ambient Intelligence (EUSAI)*. 2003.
15. A. Krause, D.P. Siewiorek, A. Smailagic, and J. Farringdon. Unsupervised, dynamic identification of physiological and activity context in wearable computing. In *Proceedings of the 7th International Symposium on Wearable Computers*, pages 88–97. IEEE Press, 2003.
16. S.-W. Lee and K. Mase. Recognition of walking behaviors for pedestrian navigation. In *Proceedings of 2001 IEEE Conference on Control Applications (CCA01)*, pages 1152–5. IEEE Press, 2001.
17. S.-W. Lee and K. Mase. Activity and location recognition using wearable sensors. *IEEE Pervasive Computing*, 1(3):24–32, 2002.
18. P. Lukowicz, H. Junker, M. Stager, T.V. Buren, and G. Troster. WearNET: a distributed multi-sensor system for context aware wearables. In G. Borriello and L.E. Holmquist, editors, *Proceedings of UbiComp 2002: Ubiquitous Computing*, volume LNCS 2498, pages 361–70. Springer-Verlag, Berlin Heidelberg, 2002.
19. J. Mantyjarvi, J. Himberg, and T. Seppanen. Recognizing human motion with multiple acceleration sensors. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pages 747–52. IEEE Press, 2001.
20. C. Randell and H. Muller. Context awareness by analysing accelerometer data. In B. MacIntyre and B. Iannucci, editors, *The Fourth International Symposium on Wearable Computers*, pages 175–176. IEEE Press, 2000.
21. A. Sugimoto, Y. Hara, T.W. Findley, and K. Yoncmoto. A useful method for measuring daily physical activity by a three-direction monitor. *Scandinavian Journal of Rehabilitation Medicine*, 29(1):37–42, 1997.
22. M. Uiterwaal, E.B. Glerum, H.J. Busser, and R.C. van Lummel. Ambulatory monitoring of physical activity in working situations, a validation study. *Journal of Medical Engineering & Technology.*, 22(4):168–72, 1998.
23. K. Van Laerhoven and O. Cakmakci. What shall we teach our pants? In *The Fourth International Symposium on Wearable Computers*, pages 77–83. IEEE Press, 2000.
24. K. Van Laerhoven, A. Schmidt, and H.-W. Gellersen. Multi-sensor context aware clothing. In *Proceedings of the 6th IEEE International Symposium on Wearable Computers*, pages 49–56. IEEE Press, 2002.
25. G. Welk and J. Differding. The utility of the Digi-Walker step counter to assess daily physical activity patterns. *Medicine & Science in Sports & Exercise*, 32(9):S481–S488, 2000.
26. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.