

Using Wearable Activity Type Detection to Improve Physical Activity Energy Expenditure Estimation

Fahd Albinali Stephen S. Intille
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
albinali | intille @mit.edu

William Haskell Mary Rosenberger
Stanford Medical School
Stanford, CA 94305 USA
whaskell | maryr @stanford.edu

ABSTRACT

Accurate, real-time measurement of energy expended during everyday activities would enable development of novel health monitoring and wellness technologies. A technique using three miniature wearable accelerometers is presented that improves upon state-of-the-art energy expenditure (EE) estimation. On a dataset acquired from 24 subjects performing gym and household activities, we demonstrate how knowledge of activity type, which can be automatically inferred from the accelerometer data, can improve EE estimates by more than 15% when compared to the best estimates from other methods.

Author Keywords

Wearable, accelerometer, wireless, activity recognition, physical activity, energy expenditure, health.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Algorithms, Human Factors, Measurement.

INTRODUCTION

The impact of physical activity on health is well documented. Physical inactivity, for example, may contribute to the onset of chronic diseases, such as heart disease and diabetes, as well as conditions such as overweight and obesity that may exacerbate a host of health problems. Because of the important relationship between physical activity and health, the physical activity and exercise science research communities are in need of better tools to study how and when people engage in physical activity and/or sedentary behaviors.

A large volume of work in the literature has shown the value of using objective accelerometer-based activity monitors [6,7,10,15]. Typically, subjects are asked to wear a monitor at the hip that measures hip motion. This motion is then

mapped onto an energy expenditure (EE) estimate using models based on EE measurements taken with indirect calorimeters that compute EE from respiratory gases [6,7,15]. Linear regression is often used. We show that by detecting activity *type*, these estimates can be improved by more than 15% on our test dataset. Our focus is on improving EE estimation rather than improving the classification of activities. On a dataset collected from 24 subjects engaged in a variety of activities including postures, ambulation, and gym and household activities, we demonstrate that state-of-the-art algorithms for EE estimation overestimate sedentary activities and underestimate vigorous activities. The proposed activity-based approach, however, is less sensitive to these deviations.

The algorithm we describe can be run in real-time on a mobile phone using a system of miniature wireless sensors we developed. The improved estimation of EE could be used by physical activity researchers, but also by a growing number of mobile applications intended to help people continuously measure or motivate physical activity for chronic disease prevention and management.

RELATED WORK

Physical activity outside of the laboratory is typically measured using either self-report surveys or, more recently, accelerometer-based activity monitors. Data from either method are often converted into EE estimates, which are typically measured in kilocalories (Kcals). Different people performing the same activity can expend varying amounts of energy based on exercise capacity and body size, and small changes in sensor placement can impact the precision and accuracy of conversion of motion to EE.

Measurement Tools

Early physical activity measurement studies used self-administered or interviewer-administered questionnaires to collect detailed information about both occupational and leisure-time activities. Self-report methods have been conducted using a variety of diaries, logs and more recently mobile devices such as PDAs where participants record their activities on hourly, daily or weekly basis. This approach suffers from a number of limitations, including burden and inconvenience in entering data, poor compliance, and inaccuracy due to bias or poor memory.

Objective tools that use accelerometers for measuring activities have been developed in research labs as well as in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

UbiComp '10, September 26–29, 2010, Copenhagen, Denmark.
Copyright 2010 ACM 978-1-60558-843-8/10/09...\$10.00.

commercial products. The Actigraph GT1M and GT3X [1] devices are widely used by researchers and clinicians. These and similar devices measure two or three-axis acceleration in a band-limited frequency. An 8-bit analog to digital converter samples the data at 10Hz, and these values are then integrated over a specific time period (epoch), usually 1 s or 1 min. This “activity count” is then mapped to energy expenditure (e.g. $\text{VO}_2/\text{kg}/\text{min}$ or MET). The limitation is that activity counts aggregate information that otherwise might be used to discriminate between activity types.

Energy Expenditure Estimation Methods

Early studies in exercise physiology [7,8,16] used a single regression model and one accelerometer placed at the hip for estimating EE across different intensities of walking and running activities. These regression models provide close estimates for moderate intensity activities but often underestimate high intensity and overestimate low intensity physical activities. The cumulative EE estimates tend to be reasonable when equal amounts of time are spent in high and low intensity activities, but skewed otherwise.

Some of the most popular models are the Freedson [7], Swartz [16], and Hendleman [8] regressions. Freedson’s model was developed using data from 50 adults during treadmill exercise at three different speeds. Swartz’s model is based on data from 70 participants who completed less-structured activities within the categories of yardwork, housework, family care, occupation, recreation, and conditioning. Hendleman’s “Lifestyle Regression” is based on data from 25 subjects who completed four bouts of walking at a range of self-selected speeds, played two holes of golf, and performed indoor (window washing, dusting, vacuuming) and outdoor (lawn mowing, planting shrubs) household tasks.

Crouter et al. proposed an improvement that uses a two-regression model [6]. This method computes the coefficient of variation on the integrated accelerometer signal for a 10 s window. The value is then used to distinguish lower intensity activities from higher ones. The approach improves EE estimates by first classifying activity into one of two classes – high or low intensity – and then applying a *type-specific* model.

More recently, supervised pattern recognition techniques have been applied to map from the activity count data to EE, in particular, using neural networks [14,15]. Neural networks have also been used to estimate activity type from the count data [15], and decision trees have been used to estimate activity type from vector magnitude data [4].

Within ubiquitous and wearable computing, much work has been proposed to detect activity type from wearable accelerometers using features computed from the raw accelerometer signal (versus summarized activity counts output by Actigraphs) (e.g., [11,12,17]). Assuming activity type is inferred from such methods, EE can be estimated using the Compendium of Physical Activities [3] (“the

Compendium”). This provides person-independent mean metabolic equivalent (MET) values, which can be converted to EE. This method does not account for differences in age, adiposity (i.e. accumulation of fat tissue) and sex. Lester et al. demonstrated that activity inference from a single accelerometer combined with regression equations from the American College of Sports and Medicine can be used to estimate calories expended during ambulatory activities. The study was validated using indirect calorimetry across 51 subjects on 3 types of activities including sitting, walking and jogging [10]. The approach has not been applied to less structured home activities (e.g. washing windows).

MOTIVATION

We combine techniques from the fields of ubiquitous/wearable computing and exercise science to create a tool that improves EE estimation by extending Crouter’s approach. Instead of using the coefficient of variation to differentiate two activity classes, we use multiple wearable sensors to detect specific activity types and then apply a type-specific or type-and-person-specific model to estimate EE.

We show that this technique offers improvements over prior approaches when tested on a comprehensive dataset of postures, ambulatory activities, gym activities, and other everyday activities. To our knowledge, this is the first time such an approach has been used. Moreover, the algorithms we present are amenable to real-time implementation and could enable new wearable computing applications that use real-time energy expenditure estimation to help with health-related tasks.

We note that our goal is not only to generate a good overall EE estimate, but also to generate estimates that will not consistently overestimate or underestimate particular activities. Much of the prior regression-mapping research, used extensively in the medical community, produces good overall estimates (assuming a distribution of physical activities similar to the training set) but systematically overestimates or underestimates EE of particular activities. The models are heavily dependent on the balance of activities used as training data, a limitation avoided in the approach presented here.

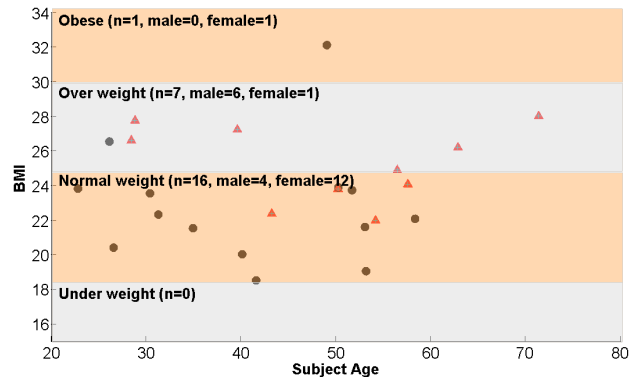


Figure 1: 24 subjects participated in the experiment (circle=female, triangle=male)

EXPERIMENT

Twenty-four healthy participants (10 male, 14 female) aged 18-75 years (mean=44.2, standard deviation=12.8), with Body Mass Index (BMI) (mean=23.8, standard deviation=3.1) from diverse ethnic backgrounds were recruited from the San Francisco Bay area to participate in the study. Based on answers to a physical activity screening questionnaire, people with medical conditions known to increase their risk of a medical complication during physical activity or a blood pressure measurement over 130/90 mm Hg were excluded. Each participant received \$50 U.S. for participation. The study procedures were approved by a human subjects review board. Figure 1 shows the age and BMI distribution of the study population.

Participants were scheduled on different days to perform two different types of routines, one in the lab and one inside and outside of the lab, while wearing an Oxycon Mobile indirect calorimeter [2], an Actigraph GT1M [1] at the dominant hip, miniature wireless accelerometers at three locations, and some other small sensors not discussed in this work. Participants were asked to perform one or both of two activity routines. Nineteen participants performed routine 1 and eighteen participants performed routine 2 for a total of 37 sessions with 15 participants performing both routines on separate days. The routines are defined in Table 1. Each routine took about an hour for participants to complete. Participants were given instructions on the activities to be performed while being monitored by trained personnel. Each participant's heart rate was continuously monitored and

Activity	R	Code	MD (min)	METS
Sitting slouching	2	SS	5.5	1.0
Sitting at desk	1,2	SI	2.6	1.0
Lying down	1,2	LD	4.6	1.0
Sitting fidgeting hands	1	SH	2.8	1.5
Sitting fidgeting feet	1	SF	2.0	1.5
Standing	1,2	ST	8.9	2.0
Standing folding laundry	2	FL	9.3	2.5
Walking 2mph	1	WT2	9.2	2.5
Stretching sitting	2	T	1.5	2.5
Washing windows	2	WW	5.1	3.0
Stretching standing	2	A	2.3	2.3
Cycle erg. 75 rpm/25 watts	1	CYC-25	6.1	3.1
Tread. walk 3mph/0% grade	1,2	TW3	9.5	3.3
Vacuuming	2	VA	8.6	3.5
Mopping	2	MO	9.7	3.5
Airdyne 30 rpm	2	AIR-30	10.4	4.2
Tread. walk 3mph/9% grade	1	WT3-9	10.1	4.5
Cycle erg. + Airdyne 20 rpm	1	AIR-20	5.8	4.5
Stepping on platform	2	P	1.5	4.9
Cycle erg. 75 rpm/50 watts	1	CYC-50	5.9	5.2
Cycle erg. + Airdyne 50 RPM	1	AIR-50	5.4	7.0
Jog 5 mph	1	JG	1.5	8.0

Table 1: Activity types observed in both routines (R) of the experiment, duration of data collected, and corresponding MET levels from the Compendium of Physical Activities [3].

perceived exertion was assessed. The staff member stopped exercise if the heart rate exceeded 85% of age-predicted maximal heart rate (Maximal HR=220-age) or if the participant's rating of perceived exertion exceeded 17 on the Borg scale of 6 to 20.

The dataset acquired and used in this work contains 35.7 hours of annotated laboratory data and 12.3 hours of annotated field data. The activities performed during the routines are listed in Table 1 with the mean duration (MD) of the activities and their MET levels from the Compendium [3].

Oxygen consumption (VO_2 /kg/min) was measured on a breath-by-breath basis using the well-validated second generation Oxycon Mobile device from Viasys Healthcare [2]. The Oxycon consists of a sensing unit, a wireless transmitter unit, a mask and a shoulder-belt system with a total weight of approximately 950 g including batteries. Before each test, the Oxycon oxygen/carbon dioxide sensor was calibrated according to the manufacturer's instructions. Although the device is relatively comfortable, session length was limited to approximately an hour to ensure that participants did not become uncomfortable. A recent study demonstrated that the Oxycon Mobile can be used to measure a wide range of VO_2 values at different work rates (50, 100 and 150 Watts) accurately, consistently and reliably [13].

Three miniature wireless accelerometers were worn at the following locations: the dominant hip to capture overall body motion, the dominant thigh to capture lower body motion and posture, and the dominant upper arm to capture upper body motion and posture. The sensors were small enough to be worn on these locations comfortably and without restricting movement. The accelerometers transmitted 45 Hz 3-axis +/- 10g data to two receivers connected to a lightweight laptop worn in a backpack. The Actigraph GT1M was set to a 1 s epoch and its start time was synchronized with the backpack computer that recorded the acceleration data. The laptop was also synchronized with the Oxycon mobile device. All sensors were tightly placed on the body and inspected regularly to ensure that no displacement occurred during data collection.

Research staff carried a mobile device with custom touch-screen annotation software, which allowed easy entry of activity start/stop times. This device was synchronized with all the other data collection devices. Figure 2 shows a person wearing the experimental setup.

ANALYSIS

Although the results presented in this section were computed offline for analysis purposes, all components of our system also run in real-time on Windows Mobile phones using data transmitted from miniature Bluetooth accelerometers.

At the beginning of each data session, clocks on all devices were synchronized within 1 s accuracy. The Oxycon Mobile software does not store its reference time and therefore we

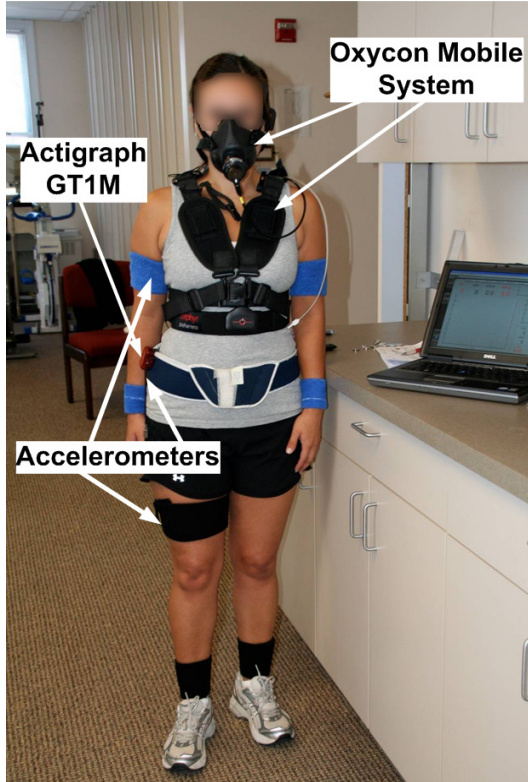


Figure 2: A person wearing the experimental setup.

created a special application to record the start time of the device when the researcher initiates the data collection.

Once the session completed, data were downloaded from each device. The Actigraph recorded activity counts with 1 s epochs. For the Oxycon, the sampling rate depended on the respiratory rate of the subject and the type of activity. The average respiratory rate ranged from 21 to 35 breaths/min. To synchronize the measured volume of oxygen, we applied a cubic spline interpolation and resampled the data at 60 Hz. For the accelerometer data, we computed an activity count similar to the Actigraph activity count. First, we applied a two step filtering process to smoothen the accelerometer signal and to remove its static component. Specifically, we used a Butterworth band pass filter (0.1 Hz-20 Hz band, 4 poles) and a baseline correction to each acceleration value by removing the mean calculated using data from the 5 s that precede the acceleration value. Second, we integrated the resulting signal over a window of 1 s and synchronized it with the rest of the data.

We describe Figure 3 and Table 2, because we will refer to them frequently in the rest of this section. Figure 3 shows four plots of an annotated EE trace from 2 different subjects performing routines 1 and 2. The x axis represents time in minutes. The y axis represents EE in $\text{VO}_2/\text{kg}/\text{min}$. Figures 3A and 3B are traces from routine 1 that include 12 exercise activities annotated on the horizontal bar plot shown between the 2 figures and described in Table 1. Figures 3C and 3D are traces from routine 2 that include 12 exercise and household activities that are annotated on the horizontal bar plot shown between the two figures. All plots display the ground truth measured by the Oxycon as a solid line. Both Figures 3A and 3C compare the ground truth to state-of-the-art in EE estimation in the exercise physiology field including Freedson regression [7], Hendelman regression [8], Swartz regression [16], and Crouter’s refined 2-regression model [6]. These methods use Actigraph data for EE estimation. Figures 3B and 3D compare the Oxycon measurements to new approaches for EE estimation that are based on a combination of activity type detection and either regression modeling for EE estimation or direct lookup of MET values in the Compendium.

Table 2 provides more detailed comparisons of the overall results across all subjects. First, results from the worst and best sessions with respect to EE estimation using our proposed method are listed with EE estimates in Kcals, error in Kcals, and the percentage of error. *Mean session errors* describes the mean of the percentage errors in energy expenditure estimates across all sessions and all subjects. As we will discuss, overall results are biased based on the relative percentage of types of activity in the test sessions. However, the relative percentage experienced by people in everyday life may be quite different. Therefore, we also present *projected daily errors*, which is the mean of percentage of errors in energy expenditure for all subjects based on 15.7 waking hours of a hypothetical weekday.¹

Performance of regression methods

We begin by describing the performance of state-of-the-art methods in the exercise physiology field on our test data.

¹ The hypothetical day, which was constructed, in part, using the American Time Use Survey, includes the following activities: 30 min of self-care in the morning, 45 min of purposeful exercise, 15 min walking the dog, 12 min eating breakfast, 21 min driving to work, 24 min eating lunch, 225 min working while standing, 21 min driving back home, 15 min walking the dog, 30 min eating dinner, 210 min watching TV and reading and the remaining 105 min doing miscellaneous house activities while standing. To compute an estimated expenditure, each activity was mapped to a similar activity in our test set (e.g., dinner was mapped to *sitting fidgeting*, and walking dog was mapped to *treadmill walking at 2mph*). This is an approximation, but the total estimated waking day expended Kcals using the hypothetical day across subjects is 1911, which seems reasonable.

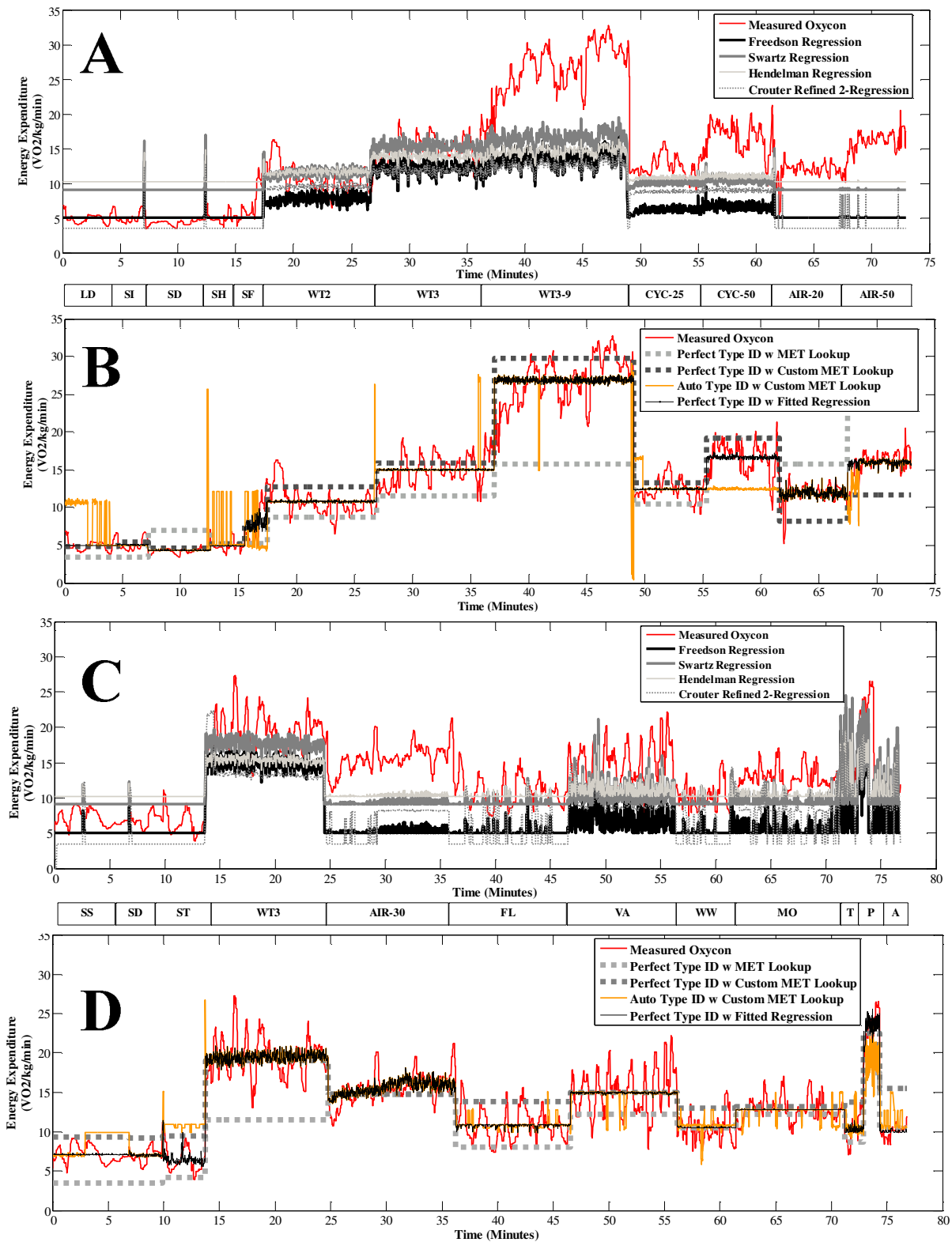


Figure 3: Example EE (measured and estimates) for 2 subjects performing routine 1 (A and B) and routine 2 (C and D). Figures 3A and 3C compare measured VO₂ to Freedson, Swartz, Hendelman and Crouter regressions. Figures 3B and 3D compare measured VO₂ to subject-specific perfect classification with fitted regression, perfect and automatic classification with custom MET lookup and perfect classification with MET lookup.

Method	Worst Session (EE=299Kcals)			Best Session (EE=428Kcals)			Mean Session Errors	Projected Daily Errors
	EE (Kcals)	Error (Kcals)	Error (%)	EE (Kcals)	Error (Kcals)	Error (%)		
Freedson	163	-136	-45%	281	-147	-34%	33.5%	12.7%
Swartz	251	-48	-16%	370	-57	-13%	29.2%	50.9%
Hendelman	259	-39	-13%	349	-79	-18%	34.2%	58.3%
Crouter	159	-140	-47%	350	-78	-18%	38.5%	23.8%
PC w/ MET lookup	236	-63	-21%	359	-68	-15%	88.9%	24.7%
Auto ID w/ MET lookup	186	-113	-38%	193	-235	-55%	37.3%	26.5%
PC w/ fitted regression	299	-0.3	-0.1%	431	3	0.6%	1.5%	1.3%
Auto ID w/ fitted regression	306	7.8	2.3%	435	7	2%	3.9%	2.1%
PC w/ custom MET lookup	210	-89	-30%	425	-3	-0.6%	13.6%	3.2%
Auto ID w/ custom MET lookup	219	-81	-27%	405	-23	-5%	13.1%	3.3%

Table 2: Comparison of different methods to estimate EE. *Mean session errors* is the mean of the percentage errors of energy expenditure across all sessions and all subjects. *Projected daily errors* is the mean of percentage of errors in energy expenditure for all subjects based on 15.7 waking hours of a typical waking day with an average projected EE per subject of 1911 Kcals. *Auto ID* is short for automatic type identification. *PC* is short for perfect type classifier.

These methods use a single accelerometer at the hip, such as the Actigraph, and map their integrated output to energy expenditure using regression models.

Our first observation is that all regression equations consistently underestimated moderate to vigorous physical activities in routine 1. Figure 3A shows a clear disparity between the regression models and the measured Oxycon values during treadmill walking at 3 mph with 9% grade (WT3-9) and during cycle ergometer (CYC-25, CYC-50) and Airdyne activities (AIR-20, AIR-50). The problem is influenced by at least two factors. First, the accelerometer at the hip does not effectively capture the movement in the lower and the upper areas of the body (e.g. during the Airdyne biking exercise) so it underestimates the EE for those activities. Second, the activity count from an accelerometer does not capture the inclination of the surface and therefore the fitted regression cannot estimate the increase in EE due to inclination. Figure 3A shows little difference between the estimated EE for treadmill walking 3 mph (WT-3) and treadmill walking 3 mph at 9% grade (WT3-9) unlike the measured ground truth. Similarly, all regression equations consistently underestimated common daily activities in routine 2. Figure 3C shows all regressions underestimating almost all the activities starting from the Airdyne activity (AIR-30) until the end of the session. We see slight improvements during vacuuming (VA) and stepping on platform (P), most likely because the hip accelerometer registered higher activity counts.

Our second observation is that both the Swartz and Hendelman regressions consistently overestimated sedentary and low intensity activities. Figure 3A demonstrates the problem during the first 5 activities of *lying down* (LD), *sitting* (SI), *sitting at desk* (SD), *sitting hand fidgeting* (SH) and *sitting feet fidgeting* (SF). Increasing the slope of the regression can improve the EE estimates for the sedentary

portion of the routine but will increase the error for well-fitted regions such as *treadmill walking 2 mph* (WT2) and *treadmill walking* (WT3). When regressions consistently overestimate sedentary activities and underestimate vigorous activities by equal EE amounts and for equal intervals of time, it is possible to have reasonable cumulative estimates of EE because the underestimated portion compensates for the overestimated portion of the data. During typical daily activities, it is unlikely that a person will have equal intervals of underestimated EE (e.g. during vigorous activities) and overestimated EE (e.g. during sedentary activities). Further, this approach does not provide accurate estimates in real-time, as might be desired for wearable, ubiquitous computing systems.

Our third observation is that using two regressions to separately model sedentary and non-sedentary activities as proposed by Crouter et al. has the potential to improve EE estimates but still suffers from the same limitations of a linear regression model. Specifically, the coefficients of a regression can vary arbitrarily depending on the type of activities in the dataset and/or the number of examples in each different activity type. The regressions are tuned to the specific types of activities that were used to generate the coefficients; the regression minimizes the residuals by fitting at the middle of the data. Figure 3A shows significant errors in EE estimates when the regression is used to estimate EE for activities that were not included in the original data. Unless the data used for training are comprehensive and all EE estimates follow a linear trajectory, a regression model can deviate significantly from the true EE.

Based on our observations, we postulate that by generalizing Crouter's approach to use a separate regression for each activity in combination with activity type detection, we can further improve EE estimates. We examine this approach in the next section.

Performance of the type classifier

Generalizing Crouter’s approach requires automatic detection of physical activity type.

Automatic Detection of Physical Activity Type

Previous work [17] demonstrates that decision tree classifiers with multiple accelerometers can be used to effectively recognize a variety of postures and physical activities and are conducive to real-time implementation, even on mobile devices. They have performed well in prior work, and we use them here.

The body can block the 2.4 GHz range low-power radio signal used by the accelerometers deployed in this experiment, so to compensate for signal loss due to body blockage, we resampled the data using cubic spline interpolation at 128 Hz. If there was significant signal loss in a window exceeding 50% of the expected sampling rate then we omitted that window from the analysis. This amounted to less than 4% of the data.

To recognize users’ activities, we calculated time and frequency domain features for each acceleration stream from 3 sensors: the dominant hip, the dominant thigh and the dominant upper arm. Previous work [17] has shown that these locations worked well for a variety of physical and daily living activities. The data were broken into 50% overlapping sliding windows of length 1 s. The features that we used are: (1) The distances between the means of the axes of each accelerometer to capture sensor orientation for posture; (2) Variance to capture the variability in different directions; (3) Correlation coefficients to capture the simultaneous motion in each axis direction; (4) Entropy to capture the type of physical activity; and (5) FFT peaks and frequencies to differentiate between different intensities of the movements. The features are then assembled into a vector and used as input to the C4.5 classifier in the WEKA toolkit [18]. WEKA is then used to evaluate classification performance using within-subject 10-fold cross validation.

Table 3 shows the overall performance results of the activity recognition averaged across 37 sessions that included 24 participants. Accuracy (A), Precision (P) and Recall (R) are the average accuracies, precisions and recalls across all sessions.

At first glance, the performance of the algorithm on subject dependent training is not as high as we might expect. The mean accuracy, precision and recall are 49%, 0.46 and 0.49 respectively, where random guessing is 4.5%. Examination of the confusion matrices, however, shows that incorrect predictions are mostly classified into similar activities. For instance, *sitting slouching* (1.0 MET) was frequently misclassified as *sitting at desk* (1.0 MET). Similarly, *stepping on platform* (5.0 METs) is consistently misclassified as *treadmill walking at 3mph* (3.3 METs). There are at least 3 reasons for the suboptimal performance of the classifier. First, the datasets are relatively small. Second, the number of examples varies greatly across different classes. For

Activity	Subject Dependent		
	A	P	R
Sitting slouching	45%	0.68	0.45
Sitting at Desk	56%	0.39	0.56
Lying down	20%	0.19	0.19
Sitting fidgeting hands	0%	0.00	0.00
Sitting fidgeting feet	99%	0.97	0.99
Standing	99%	0.90	0.99
Standing folding laundry	55%	0.44	0.54
Walking 2mph/0% grade	75%	0.76	0.74
Stretching sitting	0%	0.00	0.00
Washing windows	53%	0.52	0.53
Stretching Standing	0%	0.00	0.00
Cycle erg. 75 rpm/25 watts	53%	0.35	0.53
Tread. walk. 3mph/0%grade	80%	0.67	0.80
Vacuuming	72%	0.71	0.72
Mopping	31%	0.27	0.31
Airdyne 30 rpm	95%	0.93	0.95
Tread. walk. 3mph/9%grade	72%	0.71	0.72
Cycle erg.+Airdyne 20 rpm	70%	0.59	0.70
Stepping on platform	0%	0.00	0.00
Cycle erg. 75 rpm/50 watts	29%	0.22	0.29
Cycle erg.+ Airdyne 50 rpm	55%	0.55	0.56
Jog 5 mph	29%	0.33	0.30
Mean	49%	0.46	0.49

Table 3: Activity recognition results averaged across 37 sessions, where subject-specific training data was used.

example, *treadmill* activities can have 10 min of data while *stepping on platform* can have 1.5 min of data. Third, most errors occur on activities with short training data samples (e.g. *stepping on platform* or *standing stretching*) and our performance results are averaged across all activities without weighting them. Fortunately, as we show shortly, because the type estimation errors tend to be confused between activities with similar EE levels, we can use the activity type to improve EE estimation.

We also ran the activity type detection on data using a leave-one-subject-out cross validation. The performance of the algorithm in this subject-independent case is much lower. The mean accuracy, precision and recall are 26%, 0.30 and 0.26 respectively. Upon inspecting the decision trees, it becomes clear that the dataset is overfit to specific subjects and that more training data are needed. Notably though, we observed many cases where the algorithm made errors confusing similar activities such as *vacuuming* and *mopping*. Further improving the type classification, as well as making the type classification subject independent (a topic for future work), would only improve the results we present here on the use of type detection to improve EE estimation.

Performance of auto-detected type + regression

In this section, we examine a generalization to Crouter’s approach that uses two regressions to estimate EE for sedentary and non-sedentary activities. The proposed approach uses C4.5 to recognize activity type and then uses a different regression line for each activity to estimate EE. This

approach overcomes the non-linearity due to different EE levels of activities. The classifier is trained using subject-specific data and the regression coefficients are estimated using subject-specific data. The results are generated using 10-fold cross validation.

We first present results when having a perfect classifier to serve as a benchmark for comparison. Figures 3B and 3D show the case for a perfect classifier with a regression trained on subject-specific data. The EE estimates for all activities are lined up close to the mean of the measured EE. Table 2 shows that the average error in EE across all subjects and sessions in the case of a perfect classifier is 1.5%. In the two sample sessions in Table 2, errors in EE with a perfect classifier are negligible: -0.1% (-0.3Kcals) during a 299 Kcal session and 0.6% (3 Kcals) during a 428 Kcal session.

When using the automatic type identification with subject-specific fitted regression (Auto ID with fitted regression), the mean error in EE estimates across all sessions and subjects increases to 3.9%. For a projected day of a typical person without accounting for sleep time, the error drops to 2.1%. Figures 3B and 3D show that the EE estimates from the classifier with fitted regression often overlap with the perfect case. Note that when misclassification errors occurred, activities with similar EE levels were picked. Figure 3D shows a slightly underestimated EE for *stepping on platform* (P). The activity is misclassified for *treadmill walking 3 mph* (WT3), which still has a regression with a high EE. The spikes shown on the Auto ID EE line in Figures 3B and 3D are intermittent misclassifications that result in picking an inappropriate regression for the activity.

While this approach is accurate, it is impractical because the fitted regression requires measuring EE values to estimate the coefficients for each subject, and it is only possible to collect these values in lab settings. An interesting observation, however, is that all fitted regressions had slopes close to 0 and the intercept determined the estimated EE value. This suggests that the regression can be replaced with a simple lookup of an offset that should provide a good EE estimate given activity type. We explore this idea in the following section.

Performance of auto-detected type + MET Lookup

The Compendium of Physical Activities [3] consists of a list of the most common everyday activities and their associated energy expenditure in metabolic equivalents (METs) where 1 MET=3.5 VO₂/kg/min. Here we consider the possibility of using automatic detection of the activity and then estimating the EE by looking up the appropriate MET value in the Compendium.²

² cycle ergometer + Airdyne 50 rpm was not in the Compendium; we consulted with experts in exercise physiology to assign it a proper value.

Figure 3B shows the perfect classifier with MET lookup underestimating the measured EE values on most activities. The mean session error is 88.9% and the mean error for a projected day is 24.7%. Figure 3D shows a similar pattern with the EE being underestimated most of the time. Results from individual sessions in Table 2 are consistent with our observation that this approach underestimates EE values (Session 1, Error=-63Kcals (-21%) and Session 2 Error= -68Kcals (-15%)).

An unexpected result is that the mean error for the automatic type detection (37.3%) is significantly smaller than the mean error for the perfect recognition case (88.9%). With perfect classification, the EE estimates are mostly underestimating the measured value, as demonstrated earlier, whereas misclassifications (shown as spikes in Figures 3B and 3D) pick higher EE estimates that partially compensate for the under estimation demonstrated in the perfect classification case.

The primary limitation of the Compendium is that it does not account for differences among individuals that impact EE, such as age, adiposity or sex. The authors of the Compendium acknowledge that individual differences in energy expenditure for the same activity can be large [3].

Performance of auto-detected type + Custom Lookup

To overcome the deviations in MET values from measured EE values, we explore the possibility of customizing MET values to account for individual differences. The proposed approach requires EE measurements for different activities from a subsample of individuals and information about their age, height, weight and resting heart rate. Data for each activity across the subsample are then combined and used to fit a separate linear regression that uses age, height, weight and resting heart rate to estimate the measured EE. The developed regressions for different activities are then used on other individuals to estimate EE given their activity. We evaluate the viability of this approach using leave-one-subject-out validation where we generate EE estimation equations for each activity from 23 subjects and use the equations to predict EE for the left-out subject.

Figure 3B and 3D show significant improvements using a custom equation over lookup from the Compendium. The mean session error using a perfect classifier is reduced for the sessions and the projected day from 88.9% and 24.7% to 13.6% and 3.2%, respectively. Similarly, the mean session error using automatic classification is reduced for the sessions and the projected day from 37.3% and 26.5% to 13.1% and 3.3%, respectively.

Table 2 shows the best and worst case sessions for this approach. In the worst session, the error for the perfect classifier is -89 Kcals (-30%) and -81 Kcals (-27%) using automatic classification. In this case, the regression equation improved the MET estimates relative to a Compendium MET lookup but still underestimated the values. Notably, this particular subject had the lowest resting heart rate among our

subjects (47 bpm, mean= 65 bpm), which may partly explain why the estimation was suboptimal. Estimating our regression from larger subject samples can potentially account for such differences.

The best case session shows negligible deviations from the measured EE. The mean error for the perfect classifier is -3Kcals (-0.6%) and -23Kcals (-5%) using automatic classification. Errors from other methods range between -57Kcals (-13%) and -147Kcals (-34%). Importantly, we emphasize that our approach as shown in Figures 3B and 3D closely matches the measured VO_2 values in real-time. Other approaches might have cumulative EE estimates that are close to the measured EE, but they consistently overestimate and underestimate sedentary and vigorous activities respectively. Matching the true EE values is particularly important for ubiquitous computing applications that require good real-time estimates, perhaps to drive novel health interventions. Although this approach is not as accurate as the approach that uses a regression fitted to subject-specific VO_2 data, it is more amenable to real deployments because no training data from an indirect calorimeter are needed.

DISCUSSION

In this paper, we compare a variety of energy expenditure estimation methods to several existing approaches. The analyses show that automatic detection of physical activity type followed by use of activity-specific regressions provides the best cumulative estimation of EE (session error=3.9%, daily error=2.1%). The regressions rarely overestimate or underestimate and map closely to the mean of the measured EE. This approach, however, is impractical because it requires subject-specific VO_2 measurements that can only be gathered in lab settings. A good compromise is to use automatic detection of physical activities and then to lookup an activity-specific regression that maps age, height, weight and resting heart rate to an EE value. Using leave-one-subject-out validation, we have shown that this approach can result in good EE estimates (session error=13.1%, daily error=3.3%).

The proposed approach that uses custom MET lookup outperformed all regressions that were evaluated in this work. Comparing our approach to the best performing regression, we see an improvement in the mean session error of 15.6% and 15.1% using perfect and automatic classification respectively. For the projected day, we see an improvement of 9.5% and 9.4% using perfect and automatic classification respectively. Using our approach, we do not see dramatic deviations from the measured EE as shown in Figure 3. This is attributed to the use of custom regressions that fit to individual activities. Further, having separate regressions for each activity is conducive to incremental learning, where adding activities does not require recalculating regression coefficients as would be the case in other regression methods. Importantly, our approach does not require EE measurements from individuals; we have shown

that reasonable EE estimates can be predicted using subject-independent data given the activity type.

Although the type classifier misclassified a significant portion of the data, the type estimation errors tend to be confused between activities that are similar and that have similar EE levels. For type recognition, our dataset is challenging relative to many others reported in prior work, with many similar activities, such as *sitting* and *sitting and fidgeting*. The fact that the technique works well *despite* the imperfections of the type classifier, which can most likely be improved, is a strength of the type-to-EE mapping approach. A variation of our approach would train the classifier on clusters of activities with similar EE levels. While this may improve the accuracy of the classifier, it probably would have minimal impact on the EE estimation because misclassifications resulted in similar EE estimates.

For a typical waking day, the estimated average caloric expenditure of a subject was 1911Kcals. For such a day, our approach introduced an error of 63 Kcals using the automatic type classification approach. The errors using other approaches are 235 Kcals for Freedson, 974 Kcals for Swartz, 1108 Kcals for Hendelman, 455 Kcals for Crouter's and 506 Kcals for Compendium lookup. A 100 Kcal/day imbalance is believed to be sufficient to lead to weight gain [9], so an overall improvement of 172 Kcals (in comparison to Freedson) over a waking day would be a meaningful one.

It might be possible with more refinement of our regressions and larger datasets to reduce these errors. The number of wearable sensors can possibly be reduced if we take advantage of accelerometers in handheld devices such as phones. Compared to recent work from Lester et al. [10] that uses a single accelerometer and looks at 3 types of activities – walking, sitting and running – our approach outperforms their field results by 7% and underperforms their lab results by 3%. We emphasize that direct comparisons between studies are difficult to interpret since the results are highly dependent on the range of activities and types of sensors in the study.

An alternative to our proposed approach can use non-linear models to estimate EE. There is, however, an important advantage to detecting activity type and then mapping it to an EE estimate. When errors are made on activity type or when the regression consistently overestimates or underestimates EE levels, users may be able to correct the system. Previous work [5] has shown that enabling end-users to correct the system when it makes errors is important.

Our results are consistent with intuition: that type detection using a system that captures lower and upper body motion might work well on many common activities, and that a type-specific EE estimation approach – essentially a generalization of the Crouter approach – should outperform other linear methods. Widely popular methods that rely on using a single regression model (e.g. Freedson's and Hendelman's equations) cannot account for the variability in everyday

activities. By using a combination of pattern recognition algorithms and regression models that are specifically tuned to individual activities, the estimate of EE is improved by 15% using subject-dependent activity-type training data.

Based on this work, we therefore make the following recommendations for researchers developing ubiquitous computing health systems to estimate energy expenditure:

- Use more than a single sensor at the hip. A single sensor will not properly capture some activities, such as cycling. A sensor placed at the upper and lower body will improve type detection and type-based EE estimation.
- Avoid using a single regression model, which will systematically underestimate and overestimate certain activities. If possible, use type-specific regression models. This is particularly important for real-time systems, where end-users may be interested in accurate estimates of what is being expended *right now*, rather than simply overall estimates for the day.
- Evaluate the proposed system's error not only on collected data sessions with balanced activities but also in "simulated days" where activities are weighted towards common sedentary behavior. There is a fundamental limitation to much of the prior work because the makeup of the training data may skew performance of the model towards high-intensity activities that are uncommon for most people.

Our work is based on testing with 24 healthy subjects and 22 activity types. Despite the limited data set, we have shown that using activity type to EE estimation is a promising approach that can result in significant improvements to current EE estimation methods that are used in the exercise physiology field and that may enable new types of wearable fitness monitoring systems. In future work, we will expand both the size of our test set and activity types and deploy the algorithm as part of a real-time feedback system. We will also explore the applicability of our results to populations with different fitness levels and BMIs. Finally, we emphasize that the system discussed in this work is fully implemented to run in real-time on Windows Mobile phones.

ACKNOWLEDGMENTS

We thank Jeff Myll and Selene Mota for assistance with data collection and processing. This work was funded by the National Institutes of Health grant #5U01HL091737. The development of some sensors was funded by National Science Foundation grant #0313065.

REFERENCES

1. Actigraph. <http://www.theactigraph.com/>.
2. Oxycon Mobile. *Viasys Healthcare*. <http://www.viasyshealthcare.com/>.
3. Ainsworth, B., Haskell, W., Whitt, M., et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc* 32, 9 (2000), 498-516.
4. Bonomi, A., Goris, A., Annelies, H.C., Bin, Y., and Westerterp, K. Detection of type, duration, and intensity

- of physical activity using an accelerometer. *Med Sci Sports Exerc* 41, 9, 1770-1777.
5. Consolvo, S., Klasnja, P., McDonald, D.W., et al. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. *Proc UbiComp*, (2008), 54-63.
6. Crouter, S.E., Kuffel, E., Haas, J.D., Frongillo, E.A., and Bassett, D.R.J. A refined 2-regression model for the Actigraph accelerometer. *Med Sci Sports Exerc*, , Publish Ahead of Print.
7. Freedson, P.S., Melanson, E., and Sirard, J. Calibration of the computer science and applications, Inc. accelerometer. *Med Sci Sports Exerc* 30, 5 (1998), 777-781.
8. Hendelman, D., Miller, K., Bagget, C., Debold, E., and Freedson, P.S. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med Sci Sports Exerc* 32, 9, 442-449.
9. Hill, J.O., Wyatt, H.R., Reed, G.W., and Peters, J.C. Obesity and the Environment: Where Do We Go from Here? *Science* 299, 2003, 853-855.
10. Lester, J., Hartung, C., Pina, L., Libby, R., Borriello, G., and Duncan, G. Validated caloric expenditure estimation using a single body-worn sensor. *Proc UbiComp*, (2009), 225-234.
11. Lukowicz, P., Ward, J.A., Junker, H., et al. Recognizing workshop activity using body worn microphones and accelerometers. *Proc Pervasive*, (2004), 18-32.
12. Minisun. The intelligent device for energy expenditure (IDEEA). <http://www.minisun.com/>.
13. Rosdahl, H., Gullstrand, L., Salier-Eriksson, J., Johansson, P., and Schantz, P. Evaluation of the Oxycon Mobile metabolic system against the Douglas bag method. *Eur J Appl Physiol*, (2009), Published online.
14. Rothney, M.P., Neumann, M., Beziat, A., and Chen, K.Y. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *J Appl Physiol* 103, (2007), 1419-1427.
15. Staudenmayer, J., Pober, D., Bassett, D., Crouter, S.E., Bassett, D.R.J., and Freedson, P.S. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *J Appl Physiol* 107, (2009), 1300-1307.
16. Swartz, A., Strath, S.J., Bassett, D.J., O'Brien, W., King, G.A., and Ainsworth, B.E. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc* 32, 9, 450-456.
17. Tapia, E.M., Intille, S.S., Lopez, L., and Larson, K. The design of a portable kit of wireless sensors for naturalistic data collection. *Proc Pervasive*, (2006), 117-134.
18. Witten, I.H. and Frank, E. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, CA.