

Development of Novel Techniques to Classify Physical Activity Mode Using Accelerometers

DAVID M. POBER^{1,2}, JOHN STAUDENMAYER², CHRISTOPHER RAPHAEL², and PATTY S. FREEDSON¹

¹Department of Exercise Science, Exercise Physiology Laboratory, University of Massachusetts, Amherst, MA; and ²Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA

ABSTRACT

POBER, D. M., J. STAUDENMAYER, C. RAPHAEL, and P. S. FREEDSON. Development of Novel Techniques to Classify Physical Activity Mode Using Accelerometers. *Med. Sci. Sports Exerc.*, Vol. 38, No. 9, pp. 1626–1634, 2006. **Purpose:** Use of accelerometers to assess physical activity (PA) is widespread in public health research, but their utility is often limited by the accuracy of data-processing techniques. We hypothesized that more sophisticated approaches to data processing could distinguish between activity types based on accelerometer data, providing a more accurate picture of PA. **Methods:** Using data from MTI Actigraphs worn by six subjects during four activities (walking, walking uphill, vacuuming, working at a computer), quadratic discriminant analysis (QDA) was performed, and a hidden Markov model (HMM) was trained to recognize the activities. The ability of the new analytic techniques to accurately classify PA was assessed. **Results:** The mean (SE) percentage of time points for which the QDA correctly identified activity mode was 70.9% (1.2%). Computer work was correctly recognized most frequently (mean (SE) percent correct = 100% (0.01%)), followed by vacuuming (67.5% (1.5%)), uphill walking (58.2% (3.5%)), and walking (53.6% (3.3%)). The mean (SE) percentage of time points for which the HMM correctly identified activity mode was 80.8% (0.9%). Vacuuming was correctly recognized most frequently (mean (SE) percent correct = 98.8% (0.05%)), followed by computer work (97.3% (0.7%)), walking (62.6% (2.3%)), and uphill walking (62.5% (2.3%)). In contrast to a traditional method of data processing that misidentified the intensity level of 100% of the time spent vacuuming and walking uphill, the QDA and HMM approaches correctly estimated the intensity of activity 99% of the time. **Conclusion:** The novel approach of estimating activity mode, rather than activity level, may allow for more accurate field-based estimates of physical activity using accelerometer data, and this approach warrants more study in a larger and more diverse population of subjects and activities. **Key Words:** ACTIGRAPH, CLASSIFICATION, HIDDEN MARKOV MODEL, QUADRATIC DISCRIMINANT ANALYSIS

Assessment of physical activity (PA) behavior in a free-living environment is an important component of many scientific investigations. Accurate PA measurements are necessary for understanding individuals' behavior, determining the effectiveness of interventions designed to increase PA, and quantifying the relationship between PA and disease. Of particular interest in the assessment of PA is the ability to capture moderate-intensity activities highlighted in the recommendations of the Surgeon General and the U.S. Centers for Disease Control and Prevention/American College of Sports Medicine (26,33). Commensurate with the widespread interest in PA assessment, there are many instruments available for the assessment of free-living PA. These instruments can be broadly categorized as either subjective or objective measures of PA.

Subjective measures include self-report or interview-based methods of recalling PA over a given time period. These measures are subject to a variety of biases, primarily due to difficulty in the subjects' reporting of their activity or due to desires to conform to social norms. These problems seem to affect estimates of light and moderate PA most strongly (7,15). In contrast, measures of PA from doubly labeled water and accelerometer-based methods provide an objective measure of PA. Doubly labeled water provides a very accurate estimate of energy expenditure over a given period of time, but it is often prohibitively expensive, and it provides little information regarding the pattern or source(s) of energy expenditure. Accelerometer-based methods of PA assessment provide information about duration and intensity of PA and can generally be implemented at a reasonable cost. Thus, the advantages of accelerometer-based methods have led to the adoption of these methods by many researchers.

In general, accelerometer-based approaches to quantifying PA "calibrate" the accelerometer by simultaneously recording accelerometer output (usually a function of the total acceleration over time, e.g., counts or mG) and some physiological variable (e.g., $\dot{V}O_2$ or METs) in a laboratory setting, typically during locomotion. The relationship between these variables is then determined using linear regression, and ranges of accelerometer output corresponding to different levels of PA are established. The end points of these ranges (often called cut points) are then applied to

Address for correspondence: John Staudenmayer, Department of Mathematics and Statistics, Lederle Graduate Research Tower, University of Massachusetts, Amherst, MA 01003; E-mail: jstauden@math.umass.edu.
Submitted for publication April 2005.
Accepted for publication March 2006.

0915-9131/06/3809-1626/0

MEDICINE & SCIENCE IN SPORTS & EXERCISE®

Copyright © 2006 by the American College of Sports Medicine

DOI: 10.1249/01.mss.0000227542.43669.45

data collected in the field to estimate the amount of time spent at various intensities of PA (9,14).

Although an accelerometer-based approach to PA assessment has several advantages over other methods, the technique described above has an associated intrinsic misclassification error. This approach is unable to correctly distinguish two activities that produce similar total acceleration over time but that have different energy costs (e.g., walking at a given speed over a level surface (approximately 3 METs) versus walking at that same speed over an inclined surface (approximately 6 METs). Further, this method would incorrectly place into different categories two activities that have similar energy requirements but different total acceleration over time (e.g., walking over level ground vs vacuuming (both activities approximately 3 METs) (14)).

The deficiency of the cut-point approach stems from the fact that the cut-point approach only uses the mean counts per minute to classify the intensity of PA. A potential solution is to adopt a new methodology for data processing: one that uses more of the information available from the accelerometer to determine the PA mode that produced the data. In general, this is the goal of the statistical subfield of classification.

For instance, a simple example of a classification approach would be to use not only the mean counts per time unit as is usually done in cut-point methods, but also to use the standard deviation of counts per time unit to classify PA mode. David Bassett suggested this general approach in personal communication (2005) with the first author concomitantly with our pursuit of this methodology. Figure 1 shows data from MTI Actigraphs worn by six subjects performing four activities. The top panel plots the mean and standard deviation of the counts for each 15-s interval of data. Whereas the mean count provides the largest differences between the activities, the standard deviation of the count can also help. For instance, the figure suggests that the boundary between level walking and walking uphill cannot be simply defined with a single mean count cutoff. In addition, even the boundary between vacuuming and deskwork is slightly improved by using the standard deviation. Figure 1 suggests that this general approach holds promise as a way to use the accelerometer data to classify PA mode. Quadratic discriminant analysis (QDA) is a straightforward, well-developed statistical methodology that makes this approach more formal (13). Intuitively, QDA works by finding the PA mode that is closest to the observed data. QDA is optimal when the data are normally distributed, but we will use it simply as a method to discriminate between activities that produce counts with different means and variances.

There are also a variety of pattern recognition or machine learning approaches that may be taken to address the problem of classifying PA from accelerometer data. In particular, there are several classes of stochastic models available to identify patterns in data, and these patterns were used to provide information about the underlying process that generated the data. As early as 1970, it was suggested that stochastic models should be applied in

processing data from studies of human movement (31). However, because of their computational complexity, these types of methods have been implemented infrequently in the applied context of estimating physical activity.

Kiani and colleagues have reported the use of an automated probabilistic neural network approach to processing motion data in a clinical setting (17,18). They were able to extract features of accelerometer and goniometer signals to determine whether a patient was supine, seated, standing, or locomoting. More recently, Zhang et al. used probabilistic neural networks to extract information about the type, duration, and intensity of activity from a new type of accelerometer-based PA monitor (35,36). Devices and methods to measure human activity are also an active area of research in the machine learning community (27). Whereas the aforementioned investigations relied on data from fairly complex instruments that are currently not practical for application in large field settings, the results suggest that stochastic modeling techniques may be applicable to the problem of quantifying PA in the field.

A particular class of stochastic model for which the theory and application are well developed in a pattern-recognition setting is the hidden Markov model (HMM). HMM are part of a broad class of probabilistic pattern-recognition algorithms similar to the neural networks that have proved successful in other PA assessment applications (18,35). These models have been used with considerable success in the speech-recognition literature (28) and also appear in analyses of neuron-firing patterns, DNA sequences, analysis of viral mutations, and many other natural phenomena (5,8,30,34); thus, they are good candidates for application to the problem of classifying PA from accelerometer signals.

The methods and approaches we propose are to be used in a two-stage procedure. First, a study would be done to produce labeled data, where the accelerometer counts and the actual activities were observed. The data from that study would be used to estimate the parameters in the classification models or to train the models. After that, the trained models could be applied to accelerometer counts that were unlabeled by activity, and the unobserved activities could be estimated. Note that the unlabeled data could come from a different study or from different subjects. In spirit, this is similar to the way accelerometer count cut points are used currently, but our procedure results in a model that predicts activity mode rather than cutoff numbers that predict activity energy expenditure.

The purpose of the present study was to provide proof of concept that using data from a commonly used accelerometer, a classification algorithm, could classify several common types of PA more accurately than the standard analysis techniques that are currently in use. Because this proof of concept uses a small number of subjects and activities, it will not definitively prove that the proposed methodology is optimal. Instead, it is intended to illustrate that the methodology holds considerable

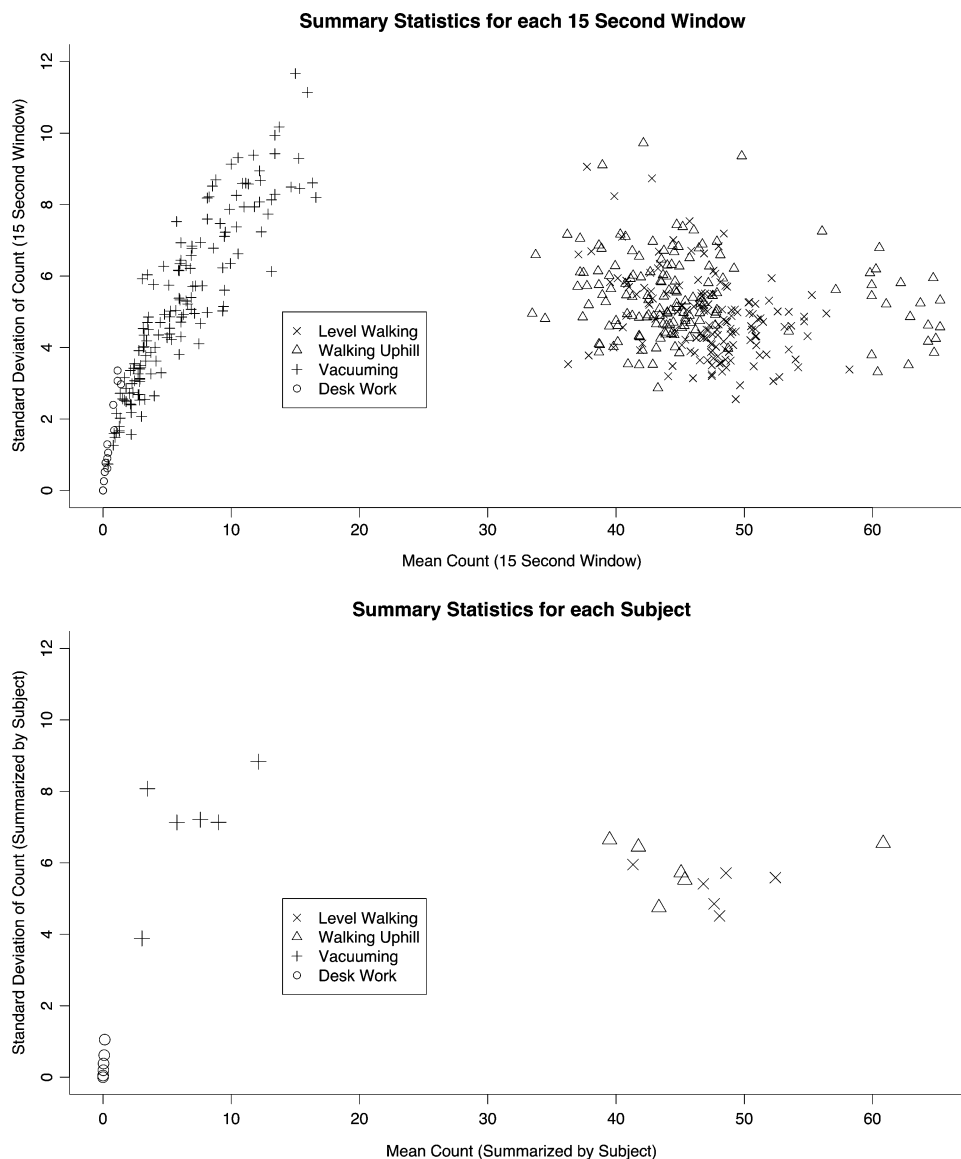


FIGURE 1—This plot is based on MTI Actigraph accelerometer count data from six people who did four activities for 7 min each. The *top panel* plots the sample mean and standard deviation for each 15-s interval of second-by-second MTI Actigraph accelerometer data. The *bottom panel* illustrates the similarity between the means and standard deviations within each activity across subjects. Even with this simple method, activities tend to cluster. The ovals are drawn by hand.

promise and deserves to be evaluated on a larger, more diverse population of subjects and activities. We chose to illustrate this with a simple and well-studied generic method (QDA) and a method that is designed to address the time-series nature of accelerometer counts (HMM). (We list other potential classification methods in this paper's conclusion.) The QDA method uses the mean and standard deviation of the counts to classify activity; it is optimal when the data are distributed normally. We considered a simpler linear discriminant, but it did not work as well as the QDA in a preliminary investigation (see Fig. 1). A potential benefit of the HMM approach is that the autocorrelation built into the HMM allows one to share classification strength across observations that are close together in time. For instance, if a person is walking at one point in time, it is likely that he or she will be walking 1 s in the future; the HMM automatically uses that

information. These new methods were evaluated for accuracy, and their accuracy was compared with the accuracy of a traditional approach to classify PA using accelerometer data. We used cross-validation to do all the evaluations; the model was always trained and evaluated on separate individuals.

MATERIALS AND METHODS

Subjects and Data Collection

Six subjects (four males and two females) who were healthy and free from any musculoskeletal limitations to moderate exercise (mean (SD) age = 24.8 (4.2) yr, height = 1.81 (0.48) m, mass = 82.3 (16.3) kg) participated in this study. Data were recorded on each subject at a sampling frequency of 1 Hz (epoch = 1 s) using an MTI Actigraph (Model 7164, Shalimar,

FL) affixed with an elastic waist belt on the right hip as described previously (9,24). This sampling frequency was chosen to provide more finely grained data than the typical recording epoch of 1 min while being sufficiently parsimonious with memory that the monitor could still record data for an entire day (8 h) and remain practical for use in field studies. Researchers wishing to use this model of Actigraph for field studies could provide subjects with multiple accelerometers initialized to begin recording data on successive days. Newer Actigraph versions (Model GT1M) have sufficient memory to record for nearly 6 d at 1 Hz.

Subjects performed a set of activities in the exercise physiology laboratory at the University of Massachusetts, Amherst, MA. The activities consisted of walking on a level treadmill at $1.25 \text{ m}\cdot\text{s}^{-1}$, walking on an inclined treadmill (7.5% grade) at $1.25 \text{ m}\cdot\text{s}^{-1}$, vacuuming, and simulated computer work. Each activity was performed for 7 min. These activities were selected because under some traditional methods for analyzing activity monitor data (e.g., Freedson cut points, (9)), walking uphill would be classified as moderate activity (≥ 3 METs) when it is in fact vigorous (≥ 6 METs). Vacuuming produces relatively low counts per minute and thus would typically be classified as light (< 3 METs), when, in fact, it requires approximately 4 METs and should be classified as moderate (14). Walking on level ground (approximately 3 METs) and computer work (approximately 1 MET) are usually correctly classified by traditional methods and were included to ensure that our approach worked at least as well as traditional approaches. The estimated MET requirements of the activities selected were based upon the compendium of physical activities, previous investigations in our laboratory (9,14,24), and the ACSM energy cost prediction equations for walking (1–3).

No instructions were provided to the subjects regarding their performance of the activities, other than the restriction that once they seated themselves at the computer, they were not permitted to rise until that stage of the data collection was complete. Subjects completed the activities in random order. When the data collection was complete, data were downloaded via the MTI reader interface unit to a personal computer and saved as ASCII text files for processing. Before participating in the data collection, all subjects read and signed an informed consent document according to the guidelines set forth by the University of Massachusetts institutional review board.

Modeling Approach

Quadratic discriminant analysis. The fitting of the QDA model (the training step) consists of estimating means and variances/covariances from nonoverlapping intervals of accelerometer data from each activity. We used 15-s intervals of second-by-second data, which provided four samples per subject-minute of activity, or approximately 168 samples per activity. Performance of the classification method was found to be relatively insensitive to variations in interval length. These means

and variances/covariances define a separate multivariate normal model for the intervals of data from each activity. This step uses labeled data where both the accelerometer count and the PA mode are known. After the model has been trained, it can be used for classification based on accelerometer count data only (unlabeled data). In this step, the models that were fit in the training step are used to estimate the probability that the data in each interval came from one of the PA modes. The PA mode with the highest probability can be assigned to an interval, or an unknown mode can be assigned if none of the probabilities exceed a given threshold. Mathematical details are in the Appendix. QDA was implemented in this study using the MASS library R. Fitting and validating the model required less than a minute on a 2-GHz dual-processor Power Mac G5.

Hidden Markov model. HMM are based on a theoretical model for the generation of a time series of data. Several key features characterize the model. The first component of an HMM is a system of interest (e.g., a human subject in a PA research study). This system can be in any one of a finite number of states (in the present example, a collection of states was used to represent each of the possible activities). The “hidden” part of the appellation HMM refers to the fact that the true state of the system at any time cannot be observed directly (e.g., the subject is in a free-living situation). The investigator is, however, able to record observations over time that are somehow related to the states (e.g., counts from an accelerometer). It is, perhaps, worth noting that the states are statistical modeling constructs and that formal, physical interpretations of them (e.g., heel-strike and midsupport phases of locomotion for the model of walking) should be avoided. The problem is that because the physical characteristics other than the accelerometer counts are neither measured nor used in the model, multiple interpretations would be in equal agreement with the data at hand.

We modeled the time series of accelerometer counts using a Poisson-based HMM as summarized in MacDonald and Zucchini (20). See the Appendix of the present article for the mathematical details of our model. The first part of the model is based on the idea that an unobserved (hidden) Markov chain with a finite number of states progresses from one state to another in parallel with the observed time series. The second part of the model links the Markov chain with the observed data. In our application, a different Poisson rate is associated with each state of the Markov chain, and, conditional on the state of the Markov chain, the observed counts are from a Poisson distribution with the associated rate. Future work will explore the validity and consequences of the assumption of a conditional Poisson distribution and consider alternatives (22). In the present study, we assumed three states were associated with each activity. We also tried two and four states per activity and achieved similar results. Methods to objectively determine the number of states have been proposed and reviewed by R. J. MacKay (21). The hidden Markov

chain used in this study was fully ergodic (i.e., all state transitions were possible), but transitions between states within an activity were much more probable than transitions between states in different activities.

We estimated the Poisson rate parameters and the Markov chain transition probabilities by maximum likelihood using the so-called Baum–Welch algorithm (23). This is referred to as training the model. We did this using a labeled dataset where we observed both the actual activity (but not the state) and the accelerometer counts (Fig. 2). After the model has been trained, it can be applied to unlabeled accelerometer data to estimate the unknown activity that produced the count at each point. We implemented this classification step through a localized decoding procedure (20) that estimates the most likely activity at each time point (in this case, at each second) given the data. As the HMM approach used the second-by-second data, we were able to test its performance on 60

samples per subject-minute of activity, or more than 2500 samples per activity. The HMM was implemented in this study using R, but could be successfully implemented in any software that has basic probability and matrix functions. Fitting and validating the model required less than 2 min on a 2-GHz dual-processor Power Mac G5.

Traditional approach. Additionally, the raw accelerometer counts were evaluated using a traditional method (Freedson cut points (9); $< 1952 \text{ counts} \cdot \text{min}^{-1} = < 3 \text{ METs}$, $1952\text{--}5724 \text{ counts} \cdot \text{min}^{-1} = 3\text{--}6 \text{ METs}$, $> 5724 \text{ counts} \cdot \text{min}^{-1} = > 6 \text{ METs}$ (9)) for classification of PA data that assigned a MET value to each minute of data based on the total acceleration during that minute. QDA- and HMM-based estimates of activity at each time point were converted to MET values based on the compendium of physical activity (vacuuming and computer work) or the ACSM prediction equations (walking and uphill walking). The accuracy of the traditional method was assessed and

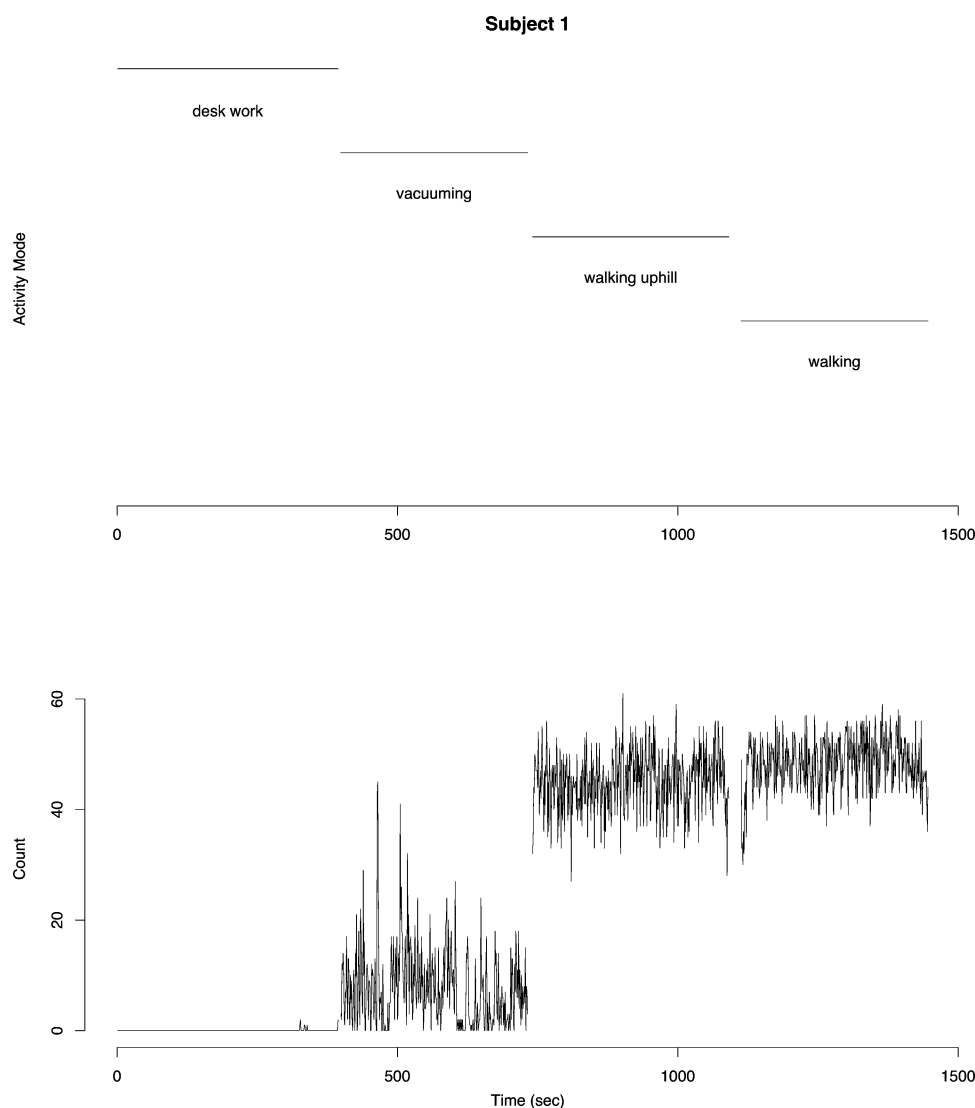


FIGURE 2—This plot contains data from one person. The *top plot* describes the person's activity mode at each time point, and the *bottom plot* contains the associated time series of second-by-second MTI Actigraph accelerometer counts. We say that activity mode labels the accelerometer counts, and we call this training data. This is the type of data used to train the models used to determine activity mode when only the accelerometer counts are observed.

TABLE 1. Proportion of seconds for each activity that were correctly classified by the QDA.

Activity	% Correct	SE (%)
Total	70.9	1.2
Walking	53.6	3.3
Walking uphill	58.2	3.5
Vacuuming	67.5	1.5
Computer work	100.0	0.1

compared with the accuracy of the QDA- and HMM-based approaches.

RESULTS

QDA-based identification of activity. On average, QDA was able to correctly identify activity from the accelerometer data in 70.9% of the seconds for which data were recorded. Specifically, walking was correctly identified in 53.6% of the appropriate seconds, walking uphill was correctly identified in 58.2% of appropriate seconds, vacuuming was correctly identified in 67.5% of appropriate seconds, and computer work was correctly identified 100% of the time (Table 1). When walking seconds were categorized incorrectly, it was almost always as uphill walking. When uphill walking was categorized incorrectly, it was nearly always misidentified as level walking. When vacuuming was misidentified, it was confused for computer work (Table 2). These error rates were computed using ordinary leave-one-out cross-validation (29), a method of performance assessment that is standard in the field of pattern recognition. This procedure consists of repeatedly training the model on all but one of the subjects, testing it on the one that was left out of the training set, and averaging the resulting scores. As the predictive ability of the classification method is tested on data that were not used to train or create the model, the resulting estimates of classification accuracy are valid estimates of those that would be obtained from a true external validation (32).

HMM-based identification of activity. On average, the HMM was able to correctly identify activity from the accelerometer data in 80.8% of the seconds for which data were recorded. Specifically, walking was correctly identified in 62.6% of the appropriate seconds, walking uphill was correctly identified in 62.5% of appropriate seconds, vacuuming was correctly identified in 98.8% of appropriate seconds, and sitting was correctly identified 97.2% of the time (Table 3). When walking seconds were categorized incorrectly, it was almost always as uphill

TABLE 3. Proportion of seconds for each activity that were correctly classified by the HMM.

Activity	% Correct	SE (%)
Total	80.8	0.9
Walking	62.6	2.3
Walking uphill	62.5	2.3
Vacuuming	98.8	0.5
Computer work	97.3	0.7

walking. When uphill walking was categorized incorrectly, it was nearly always misidentified as level walking. Similarly, when vacuuming and sitting were misidentified, they were confused for each other (Table 4).

These error rates were also computed using ordinary leave-one-out cross-validation (29). This procedure consists of repeatedly training the model on all but one of the subjects, testing it on the one that was left out of the training set, and averaging the resulting scores. Because we train and evaluate the model on different individuals, the resulting estimate of predictive performance is a valid estimate of what would have been obtained using true external validation. For the HMM we also evaluated the performance of the algorithm by dividing the subjects into a training set and a test set where we varied the size of each. The results were nearly identical (not shown).

Identification of energy expenditure level: QDA and HMM compared with the cut-point method.

The QDA- and HMM-based classifications of activities can be combined with existing estimates of the energy costs of those activities to produce estimates of the fraction of time that the subjects spent at a given level of energy expenditure. By design in this experiment, the subjects spent 25% of their time at < 3 METs (computer work), 50% of their time at 3–6 METs (level walking and vacuuming), and 25% of their time at > 6 METs (uphill walking). The QDA-based estimate is that the subjects spent 24.7% of their time at < 3 METs, 50.5% at 3–6 METs, and 24.8% at > 6 METs. The HMM-based estimate is that the subjects spent 24.7% of their time at < 3 METs, 50.5% at 3–6 METs, and 24.8% at > 6 METs. Note that this is an aggregate summary; the information in Table 1 and 2 shows that some of the 3–6 METs were misclassified as > 6 METs, and vice versa.

Regression cut points (9,14) provide another way to estimate the fraction of time that the subjects spent at a given level of energy expenditure (< 1952 counts·min⁻¹ = < 3 METs, 1952–5724 counts·min⁻¹ = 3–6 METs, and > 5724 counts·min⁻¹ = > 6 METs (9). Applying this

TABLE 2. Confusion matrix for QDA.

Classification	Truth			
	Walking (%)	Walking Uphill (%)	Vacuuming (%)	Computer Work (%)
Walking	53.6	46.4	0.0	0.0
Walking uphill	41.8	58.2	0.0	0.0
Vacuuming	0.0	0.2	67.5	32.5
Computer work	0.0	0.0	0.0	100.0

The ith row and jth column of this table contains the mean percentage of seconds in which activity j was classified as activity i by the model.

TABLE 4. Confusion matrix for the HMM.

Classification	Truth			
	Walking (%)	Walking Uphill (%)	Vacuuming (%)	Computer Work (%)
Walking	62.6	37.3	0.0	0.0
Walking uphill	36.9	62.5	0.0	0.0
Vacuuming	0.3	0.2	98.8	2.7
Computer work	0.2	0.0	1.2	97.3

The ith row and jth column of this table contains the mean percentage of seconds in which activity j was classified as activity i by the model.

method to our experiment, estimates that the subjects spent 50% of their time at 1–3 METs and 50% of their time at 3–6 METs. All of the walking and uphill-walking minutes produced counts per minute between 1952 and 5724 (3–6 METs), and all of the vacuuming and computer work minutes produced counts per minute that were less than 1952 (< 3 METs). Table 5 reflects the accuracy of the cut-point, QDA, and HMM methods for detecting minutes spent at various activity intensities. On a minute-by-minute basis, the QDA and HMM produced very accurate estimates of time spent at the different intensities, whereas the cut-point method failed to capture any vigorous (walking uphill) minutes and misclassified all of the vacuuming minutes as sedentary.

DISCUSSION

Based on the results of this investigation, it appears that improved analytical techniques show promise in reducing the misclassification error associated with traditional methods of analyzing PA monitor data. In particular, the use of QDA- and HMM-based methods in the present investigation resulted in estimates of time spent at given intensity levels that were more accurate than estimates derived from the cut-point method of processing accelerometer data. Additionally, the use of QDA or an HMM yielded estimates of the actual activity being performed at each time point.

There are at least two benefits of this additional information. First, it will enhance assessments of PA by providing a context for energy expenditure, which has been suggested as an important component of PA investigations (25). Second, this information may be used to detect a wide variety of the types of lifestyle activities that have been identified as important to the accumulation of energy expenditure due to PA (PAEE). There are, however, several considerations that should be discussed regarding the application of these techniques to the problem of PA assessment.

In addition to testing the method on, and perhaps customizing it to, a larger and more diverse population of subjects, the range of activities included in the model should be expanded. In particular, because a wide variety of activities are important in the accumulation of daily PAEE, future work should be directed towards expanding the algorithm to recognize many activities. Future research should include identification of activities in a controlled

setting and subsequent validation of the method on data from subjects in free-living environments.

Although successful in discriminating among the list of activities in the present study, the extent to which these methods will be effective is unclear because the list of activities continues to grow. In our investigation, the HMM outperformed the QDA in correctly distinguishing activities, and thus may prove more fruitful as a means of ultimately distinguishing many activities based on accelerometer data. As we attempt to increase the size of the list of identifiable activities, it is likely that the increased flexibility of the HMM will allow it to continue to outperform QDA. Put another way, it seems likely that there will be more activities with similar means and variances but different time series of data, and that this will favor the HMM for distinguishing among activities. In addition, the QDA method treats the accelerometer data from different intervals as if they were mutually independent. This is an unrealistic assumption.

Several improvements to the HMM approach can be suggested beyond simply increasing the number of activities that the HMM is trained to recognize. For more successful use of an HMM to classify PA data, it is likely that attributes of the subject could be used to customize the model and thus improve the accuracy of classification. For instance, height and weight are likely to have an impact on the pattern of accelerometer counts over time for a given activity. Other factors that contribute to interindividual variability in the relationship between accelerometer output and activity, such as age or handedness, could be explored as well in larger and more diverse populations. Future work could include studies designed to determine whether demographic or descriptive variables may be used to enhance the predictive accuracy of classification-based approaches to PA assessment.

There are also at least three technical aspects of our HMM implementation that could be improved. First, in our HMM we assume that the counts conditional on the hidden Markov state have a Poisson distribution. Although this performs relatively well in the current study, alternative distributions such as the multinomial (on bins of counts) or even the normal distribution for activities that produce high counts on average might yield improved performance.

A second technical aspect of the HMM implementation that could be improved is choosing the number of states per activity. In the present study, we found that the accuracy of the model was essentially unchanged when we used two, three, or four states per activity. As the list of activities in the model expands, however, it is likely that some activities will be more accurately represented by a larger number of states. By optimizing the number of states representing each activity, it is likely that we can reduce the proportion of seconds that are incorrectly classified. Objective methods for selecting the optimum number of hidden states for an HMM have been described in the literature (21).

A third technical aspect of the implementation that could be modified is the development and customization of an

TABLE 5. Comparison of the ability of the HMM, QDA, and cut-point methods to assess the percentage of time the subjects spent at different levels of energy expenditure.

Activity Level	Truth (%)	QDA (%)	HMM (%)	Cut-Point Method (%)
Less than 3 METs	25.0	33.1	24.7	50.0
3–6 METs	50.0	40.7	50.5	50.0
More than 6 METs	25.0	26.2	24.8	0.0

The HMM is nearly perfect in aggregate for the six subjects, but this does not mean that it correctly classified each time point. Rather, it made nearly symmetric misclassifications as was shown in Table 3. For instance, approximately equal numbers of level-walking seconds (3–6 METs) were misclassified as uphill walking (> 6 METs), and vice-versa.

efficient global decoding algorithm for this problem. We used local decoding in the present investigation. This chooses an activity for each time point by finding the mostly likely activity at each time point given all of the accelerometer data. A global decoding approach (19) would also use the estimated activities in the previous seconds to estimate each activity. Intuitively, this approach would make use of the fact that a subject is unlikely to alternate between walking, sitting, walking, and sitting in four sequential seconds. The global decoding algorithm estimates the entire sequence of activities simultaneously rather than one at a time as in the local decoding approach. The global decoding problem of determining the likely sequence of states is typically achieved by the Viterbi algorithm (20), but the inclusion of multiple states per activity complicates the necessary computations.

CONCLUSION

In summary, QDA- and HMM-based data analysis methods hold promise as a tool for improving the classification of PA data from accelerometers. Specifically, they may be able to correctly identify a variety of household and lifestyle activities that contribute to the measurement error associated with traditional data-processing methodologies. Perhaps more importantly, QDA and HMM performed substantially better than the traditional cut-point method in quantifying minutes spent at given intensity levels based on the activities studied in this investigation. We chose the QDA and HMM classification tools because one (QDA) is quite simple and easily understood, and the

other (HMM) is designed explicitly to exploit the time series nature of our data. As an anonymous reviewer pointed out, it would be possible to extend or modify the QDA by using summary statistics other than the mean and the standard deviation. Many modifications of the HMM are possible. In fact, there is a large and growing list of classification tools that could be applied to this problem (e.g., classification trees (4), other discriminant analyses (DA, (10–12)), neural networks (NN, (29)), regression splines (MARS, (19)), and support vector machines (SVM, (6,16)). However, in general, the relatively simple but somewhat dated QDA method can be quite competitive with these more modern methods (13).

Additionally, we have demonstrated that this approach is feasible using simple activity monitors that many researchers already use and that are readily applicable to field use, using software that can easily be implemented and shared using any commercially available packages that have matrix-language capabilities (e.g., SAS, S-Plus/R, Matlab). For this approach to be useful in epidemiological studies, more work is necessary to develop and refine the model's classification ability over a wider range of activities and subject characteristics. Finally, whereas the present methods proved successful, they by no means represent an exhaustive list of the methods available to address the problem of processing accelerometer data and have only been tested on a limited range of activities. Progress in this area is likely to be rapid, particularly if a wide variety of new methods are adapted and applied to the task of objectively estimating PA behavior from accelerometer data derived from a wide range of activities.

REFERENCES

1. AINSWORTH, B. E., W. L. HASKELL, A. S. LEON, et al. Compendium of physical activities: classification of energy costs of human physical activities. *Med. Sci. Sports Exerc.* 25:71–80, 1993.
2. AINSWORTH, B. E., W. L. HASKELL, M. C. WHITT, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med. Sci. Sports Exerc.* 32:S498–S504, 2000.
3. KENNEY, W. L., R. H. HUMPHREY, C. X. BRYANT, and D. A. MAHLER. *ACSM's Guidelines for Exercise Testing and Prescription*. Baltimore, MD: Williams & Wilkins, p. 303, 1995.
4. BREIMAN, L. *Classification and Regression Trees*. Belmont, Calif: Wadsworth International Group, pp. 20–23, 1984.
5. COOPER, B., and M. LIPSITCH. The analysis of hospital infection data using hidden Markov models. *Biostatistics.* 5:223–237, 2004.
6. CRISTIANINI, N., and J. SHAWE-TAYLOR. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge; New York: Cambridge University Press, 2000.
7. DURANTE, R., and B. E. AINSWORTH. The recall of physical activity: using a cognitive model of the question-answering process. *Med. Sci. Sports Exerc.* 28:1282–1291, 1996.
8. FOULKES, A. S., and V. DE GRUTTOLA. Characterizing the progression of viral mutations over time. *J. Am. Stat. Assoc.* 98:859–867, 2003.
9. FREEDSON, P. S., E. MELANSON, and J. SIRARD. Calibration of the Computer Science and Applications, Inc. accelerometer. *Med. Sci. Sports Exerc.* 30:777–781, 1998.
10. HASTIE, T., A. BUJA, and R. TIBSHIRANI. Penalized discriminant analysis. *The Annals of Statistics* 23:73–102, 1995.
11. HASTIE, T., and R. TIBSHIRANI. Discriminant analysis by Gaussian mixtures. *JRSS-B* 58:155–176, 1996.
12. HASTIE, T., R. TIBSHIRANI, and A. BUJA. Flexible discriminant analysis by optimal scoring. *J. Am. Stat. Assoc.* 89:1255–1270, 1994.
13. HASTIE, T., R. TIBSHIRANI, and J. H. FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-Color Illustrations*. New York: Springer, 2001.
14. HENDELMAN, D., K. MILLER, C. BAGGETT, E. DEBOLD, and P. FREEDSON. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med. Sci. Sports Exerc.* 32:S442–S449, 2000.
15. JACOBS, D. R. Jr, B. E. AINSWORTH, T. J. HARTMAN, and A. S. LEON. A simultaneous evaluation of 10 commonly used physical activity questionnaires. *Med. Sci. Sports Exerc.* 25:81–91, 1993.
16. KARATZOGLOU, A., A. SMOLA, K. HORNIK, and A. ZEILEIS. Kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Soft* 11:1–20, 2004.
17. KIANI, K., C. J. SNIJDERS, and E. S. GELSEMA. Computerized analysis of daily life motor activity for ambulatory monitoring. *Technol. Health Care* 5:307–318, 1997.
18. KIANI, K., C. J. SNIJDERS, and E. S. GELSEMA. Recognition of daily life motor activity classes using an artificial neural network. *Arch. Phys. Med. Rehabil.* 79:147–154, 1998.
19. KOOPERBERG, C., S. BOSE, and C. J. STONE. Polychotomous regression. *J. Am. Stat. Assoc.* 92:117–127, 1997.

20. MACDONALD, I. L., and W. ZUCCHINI. *Hidden Markov and Other Models for Discrete-Valued Time Series*. London; New York: Chapman & Hall, p. 66, 1997.
21. MACKEY, R. J. Estimating the order of a hidden Markov model. *Can. J. Stat.* 30:573–589, 2002.
22. MACKAY-ALTMAN, R. J. Assessing the goodness-of-fit of hidden Markov models. *Biometrics* 60:444–450, 2004.
23. McLACHLAN, G. J., and D. PEEL. *Finite Mixture Models*, New York: Wiley, pp. 329–332, 2000.
24. MELANSON, E. L. Jr, and P. S. FREEDSON. Validity of the Computer Science and Applications, Inc. (CSA) activity monitor. *Med. Sci. Sports Exerc.* 27:934–940, 1995.
25. PAFFENBARGER, R. S. Jr, S. N. BLAIR, I. M. LEE, and R. T. HYDE. Measurement of physical activity to assess health effects in free-living populations. *Med. Sci. Sports Exerc.* 25:60–70, 1993.
26. PATE, R. R., M. PRATT, S. N. BLAIR, et al. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *JAMA* 273:402–407, 1995.
27. PENTLAND, A. Healthwear: medical technology becomes wearable. *IEEE Computer* 37:34–41, 2004.
28. RABINER, L. R., and B. H. JUANG. *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: PTR Prentice Hall, 1993, pp. 321–386.
29. RIPLEY, B. D. *Pattern Recognition and Neural Networks*. Cambridge; New York: Cambridge University Press, pp. 69–71, 1996.
30. SCHLIEP, A., C. STEINHOFF, and A. SCHONHUTH. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics* 20(Suppl 1):I283–I289, 2004.
31. SCHUTZ, R. W. Stochastic processes: their nature and use in the study of sport and physical activity. *Res. Q.* 41:205–213, 1970.
32. STONE, M. Cross-validatory choice and assessment of statistical predictions. *J. Royal. Stat. Soc. Ser. B.* 36:111–147, 1974.
33. United States. Public Health Service. Office of the Surgeon General, National Center for Chronic Disease Prevention and Health Promotion (U.S.), and President's Council on Physical Fitness and Sports (U.S.). Physical activity and health: a report of the Surgeon General. Atlanta, GA. U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, President's Council on Physical Fitness and Sports; For sale by the Supt. of Docs. 1996.
34. WU, W., M. J. BLACK, D. MUMFORD, Y. GAO, E. BIENENSTOCK, and J. P. DONOGHUE. Modeling and decoding motor cortical activity using a switching Kalman filter. *IEEE Trans. Biomed. Eng.* 51:933–942, 2004.
35. ZHANG, K., F. X. PI-SUNYER, and C. N. BOOZER. Improving energy expenditure estimation for physical activity. *Med. Sci. Sports Exerc.* 36:883–889, 2004.
36. ZHANG, K., P. WERNER, M. SUN, F. X. PI-SUNYER, and C. N. BOOZER. Measurement of human daily physical activity. *Obes. Res.* 11:33–40, 2003.

APPENDIX: MATHEMATICAL DETAILS

This appendix contains mathematical definitions of quadratic discriminant analysis and the hidden Markov model. In general, we use y_t to denote the accelerometer count at time t , $t = 1, \dots, T$, and a_t to represent the activity at time t where the possible activities are $1, \dots, A$.

Quadratic Discriminant Analysis

Let m_a and S_a^2 denote the mean and covariance matrix for activity a ($a = 1, \dots, A$) that are estimated from the training data. Suppose we observe unlabeled accelerometer counts, $y_t = (y_{t1}, \dots, y_{t14})'$, and suppose that all A activities are equally likely, *a priori*. Using Bayes' rule, the probability that y_t comes from activity a is

$$\begin{aligned} & |S_a^2|^{-1/2} \exp\{-(y_t - m_a)'(S_a^2)^{-1}(y_t - m_a)/2\} / \\ & [|S_1^2|^{-1/2} \exp\{-(y_t - m_1)'(S_1^2)^{-1}(y_t - m_1)/2\} + \dots \\ & + |S_A^2|^{-1/2} \exp\{-(y_t - m_A)'(S_A^2)^{-1}(y_t - m_A)/2\}] \end{aligned}$$

Hidden Markov Model

Let $1, \dots, S$ index the states of the hidden Markov chain, where $z_t = s$ means that the chain is in state s at time t . We relate the activities to the states by defining the following partition of the states: $(1, \dots, s_1)$, $(s_1 + 1, \dots, s_2)$, ..., $(s_{A-1} + 1, \dots, s_A)$ where $a_t = a$ if and only if z_t is in the set $(s_{a-1} + 1, \dots, s_a)$.

(We let $s_0 = 1$.) Finally, we use $\lambda_s > 0$ for the Poisson rate that corresponds to state s . The likelihood for the observed data, $\{y_t\}_{t=1, \dots, T}$ can be defined from the following three specifications: $y_t | z_t = s$ approximately Poisson (rate = λ_s), $t = 1, \dots, T$, independent, $\text{Prob}(z_t = s | z_{t-1} = r) = \pi_{rs} \geq 0$, $t = 2, \dots, T$, $r = 1, \dots, S$, $s = 1, \dots, S$, with $\pi_{r1} + \dots + \pi_{rS} = 1$ for all r , and $\text{Prob}(z_1 = s) = \delta_s \geq 0$, $s = 1, \dots, S$, with $\delta_1 + \dots + \delta_S = 1$.

The training problem corresponds to using observations of $\{y_t\}_{t=1, \dots, T}$ and $\{a_t\}_{t=1, \dots, T}$ to estimate the parameters $\{\lambda_s\}_{s=1, \dots, S}$, $\{\pi_{rs}\}_{r=1, \dots, S, s=1, \dots, S}$, and $\{\delta_s\}_{s=1, \dots, S}$. The classification problem corresponds to using a model with estimated parameters to estimate a_1, \dots, a_U from new observations y_1, \dots, y_U . In our application, there are four activities ($A = 4$), and we use three states for each activity. In other words, $a_t = 1$ corresponds to z_t in the set $(1-3)$, ..., $a_t = 4$ corresponds to z_t in the set $(10-12)$.

We train the model one activity at a time. Specifically, we use the y_t s that correspond to $a_t = 1$ to estimate $\{\lambda_s\}_{s=1, \dots, 3}$, and $\{\pi_{rs}\}_{r=1, \dots, 3, s=1, \dots, 3}$, the y_t s that correspond to $a_t = 2$ to estimate $\{\lambda_s\}_{s=4, \dots, 6}$, and $\{\pi_{rs}\}_{r=4, \dots, 6, s=4, \dots, 6}$, etc. To allow transitions from one activity to another, we set the π_{rs} s that correspond to those state changes to 0.01 and renormalize so that $\pi_{r1} + \dots + \pi_{rS} = 1$ for all r . We let $\delta_s = 1/12$. We found the performance of the algorithm to be insensitive to these choices. We classify by finding a_t to maximize $\text{Prob}\{z_t \text{ is in the set } (s_{a_t-1}+1, \dots, s_{a_t}) \mid y_1, \dots, y_U\} = \text{Prob}(z_t = s_{a_t-1} + 1 \mid y_1, \dots, y_U) + \dots + \text{Prob}(z_t = s_{a_t} \mid y_1, \dots, y_U)$.