Published in final edited form as:
Med Sci Sports Exerc. 2016 May ; 48(5): 933–940. doi:10.1249/MSS.0000000000000840.

Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification

Katherine Ellis[1], **Jacqueline Kerr**[2], **Suneeta Godbole**[2], **John Staudenmayer**[3], and **Gert Lanckriet**[1]

[1]Department of Electrical and Computer Engineering, University of California, San Diego, CA

[2]Department of Family Medicine and Public Health, University of California, San Diego, CA

[3]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA

Abstract

Purpose—Accelerometers are a valuable tool for objective measurement of physical activity (PA). Wrist-worn devices may improve compliance over standard hip placement, but more research is needed to evaluate their validity for measuring PA in free-living settings. Traditional cut-point methods for accelerometers can be inaccurate, and need testing in free-living with wrist-worn devices. In this study we developed and tested the performance of machine learned (ML) algorithms for classifying PA types from both hip and wrist accelerometer data.

Methods—Forty overweight or obese women (mean age = 55.2 ±15.3 yrs; BMI = 32.0 ± 3.7) wore two ActiGraph GT3X+ accelerometers (right hip, non-dominant wrist) for seven free-living days. Wearable cameras captured ground truth activity labels. A classifier consisting of a random forest and hidden Markov model classified the accelerometer data into four activities (sitting, standing, walking/running, riding in a vehicle). Free-living wrist and hip ML classifiers were compared to each other, to traditional accelerometer cut points, and to an algorithm developed in a laboratory setting.

Results—The ML classifier obtained an average of 89.4% and 84.6% balanced accuracy over the four activities using the hip and wrist accelerometer, respectively. In our dataset with an average of 28.4 minutes of walking or running per day, the ML classifier predicted an average of 28.5 minutes and 24.5 minutes of walking or running using the hip and wrist accelerometer, respectively. Intensity-based cutpoints and the laboratory algorithm significantly underestimated walking minutes.

Conclusions—Our results demonstrate the superior performance of our PA type classification algorithm, particularly in comparison to traditional cut-points. While the hip algorithm performed better, additional compliance achieved with wrist devices might justify using a slightly lower performing algorithm.

Corresponding author: Katherine Ellis, 9500 Gilman Dr 0409, University of California, San Diego, La Jolla, CA 92093-0811, Phone: 360-608-4265, kellis@ucsd.edu.

CONFLICT OF INTEREST
None of the authors have conflicts of interest and the results of the present study do not constitute endorsement by ACSM.

Keywords

Random forest; Hidden Markov Model; Classification; Physical activity

INTRODUCTION

Objective measurement of Physical Activity

Accelerometers are a valuable tool for objective measurement of physical activity (PA), and accelerometer use in PA research has increased exponentially in recent years (1). Objective PA measurement is an important tool for examining the relationship between PA and various health outcomes, assessing the effects of behavioral interventions, and studying the relationships between PA and built environments. Traditional analysis of accelerometer data in PA research has been based on proprietary manufacturer-defined second-level summary measures known as “activity counts.” Researchers have established cut-points, based on laboratory-collected accelerometer data from healthy mostly young adults, to group activity count data into intensity categories of sedentary, light or moderate to vigorous PA (MVPA) (11).

Tri-axial accelerometry and machine learning

More recently, accelerometer manufacturers have been providing access to raw tri-axial acceleration data in units of gravity (*g*). This has allowed researchers more flexibility in data processing. In particular, machine learning (ML) methods to analyze raw accelerometer data have gained popularity. Machine learning methods can capture complex dependencies and nonlinearities in accelerometer data to improve recognition of PA. A particular advantage is that ML methods can be used to classify PA behaviors, in addition to intensity of acceleration (29). The ability to measure specific behaviors can be useful for PA researchers; for example, a researcher interested in measuring the effect of interventions might like to know that participants are increasing time spent in targeted behaviors such as walking or cycling as opposed to general MVPA. In Active Living Research, specific behaviors such as cycling and walking are more likely to be related to built environments than general MVPA (27). Additionally, advancing the understanding of the relationship between behaviors and health outcomes can allow for public health guidelines based on specific behaviors, which may be more easily interpreted by members of the public who might struggle to process and enact terms such as “moderate physical activity”.

Importance of free-living assessment

A limitation of traditional cut-point methods and emerging computation techniques is that they have been calibrated or trained on data collected in controlled laboratory-based settings. Recent research has shown that traditional cut-point methods misclassify behaviors in free-living data with high error rates (16). Bastian et al. demonstrated differences in classifiers developed in the laboratory and in observational protocols (2). Machine learning methods developed on free-living data have been shown to outperform methods based on observational protocols (9). While it is expensive and time-consuming to collect free-living accelerometer data with corresponding ground-truth activity labels, new camera technologies have made this more feasible. Recent work has shown that wearable cameras can clearly be used to assess free-living behaviors (7).

Growing popularity of wrist accelerometers

Traditionally, a hip-mounted accelerometer has been used to measure PA (1). However, the use of wrist-worn accelerometers to measure PA has been gaining popularity in recent years. For example, the large-scale National Health and Nutrition Examination Survey (NHANES) changed its accelerometer measurement location from the hip to the wrist in 2011 (30). An advantage of wrist accelerometers is that they may improve compliance, and can be worn 24 hours without removal for sleep (33). Previous studies have shown that a wrist accelerometer can be used to classify PA behaviors (12, 19) and estimate overall PA (10) and intensity thresholds (13, 23, 25). More research is needed, however, to evaluate the validity of wrist-worn devices for measuring PA in free-living. Intuitively, wrist acceleration will have a more noisy relationship with PA than hip acceleration, as extraneous arm movements can create high accelerations without correspondingly high energy expenditure. A wrist accelerometer will, therefore, record higher overall levels of acceleration than a hip accelerometer (13). Machine learning methods for classifying PA have been shown to perform well on free-living hip-worn accelerometer data (9, 18). However, these methods have not been tested in participants with multiple days of entirely free-living wrist worn accelerometer data.

The aims of this study were to compare ML behavior classifications made from wrist-worn accelerometer data and hip-worn accelerometer data on a free-living population. We also compare our ML behavior classifications with traditional intensity cut-points, to highlight differences that might impact comparisons across future studies of population estimates using different techniques. There have been calls for greater standardization in accelerometer methods and the current findings will inform consensus development for accelerometer wear and data processing protocols in future studies (34). Further, we compare our new free-living algorithms to a published computation technique developed in a laboratory setting. The two-level ML behavior classification algorithm presented in this paper is available as an R package that is free for other researchers to download and use on their own data (cran.r-project.org/web/packages/TLBC).

METHODS

Dataset collection

Subjects—Forty overweight or obese women (mean age = 55.2 ±15.3 yrs; BMI = 32.0 ± 3.7) performed their normal daily activities for seven days. This population was recruited as part of the NCI funded Transdisciplinary Research in Energetics and Cancer (TREC) (21). Almost half of these women were breast cancer survivors. Written informed consent was provided by all participants. Details regarding participant eligibility are published elsewhere (22).

Accelerometers—Participants were asked to wear two ActiGraph GT3X+ (ActiGraph, Pensacola, FL) accelerometers: one on a belt on their right hip and one on their non-dominant wrist. The ActiGraph GT3X+ captured raw triaxial acceleration (± 6 *g*) at 30 Hz

(Previous literature (4, 20) has shown 30Hz to be sufficient for human behavior classification). Participants were instructed to wear the hip accelerometer during waking hours and the wrist accelerometer 24 hours/day, including overnight.

Ground truth capture—In order to capture ground truth information about participant behavior, participants also wore a small camera called a SenseCam (Vicon Revue) on a lanyard around their neck. The SenseCam captured first-person images approximately every 20 seconds. Participants were trained on IRB-approved procedures for ensuring privacy and confidentially for themselves and others. Participants wore the camera for the full seven days of data collection. SenseCam image data were downloaded and imported into the Clarity SenseCam Browser (8), and researchers annotated the SenseCam images with ground truth behavior labels. A standardized annotation protocol was developed, and inter-rater reliability was established. More details on SenseCam image annotation can be found elsewhere (16) and the complete annotation protocol is available from the authors upon request. The SenseCam annotation protocol assigns one of six mutually exclusive posture labels to each image: sitting, standing still, standing moving, walking/running, and bicycling. For this analysis we combined the two types of standing, ignored bicycling (as only participant bicycled), and separated sitting in a vehicle from other types of sitting (This separation was done for ease of classification only, as the accelerometer measurements look different in vehicle travel. Note that sitting in a vehicle and other types of sitting can be combined after behavior classification). This resulted in four behavior labels: sitting, standing, walking/ running, and riding in a vehicle. The resulting dataset contains over 100,000 annotated minutes of data, comprising a free-living dataset much larger than comparable literature (19).

Behavior classification algorithm

We designed a behavior classification system that uses machine learning algorithms to predict behaviors from raw triaxial accelerometer data. Our system predicts a behavior label for each minute of accelerometer data. A one-minute window was chosen because we believe it is a sufficiently detailed interval by which to represent public health relevant behaviors on a daily level, and allows comparison to traditional counts-per-minute based methods. The behavior classification process is comprised of three steps: feature extraction, minute-level classification, and time-smoothing. A detailed description of these three steps follows:

Feature extraction—A sliding one-minute window was used to convert raw (unfiltered) triaxial accelerometer data into a sequence of forty-one dimensional feature vectors. The features extracted include basic time-domain descriptors, correlation features, angular features, and frequency-domain features.

For each sample in a data window, the vector magnitude (VM) of the acceleration signal was calculated *i.e.*, $v = (x^2 + y^2 + z^2)^{1/2}$. The following basic statistical descriptors of the VM were calculated over the data window: *mean,* standard deviation *(sd),* coefficient of variation (*coefvariation*)*,* minimum (*min*)*,* maximum (*max*)*,* 25th, 50th and 75th percentile (*25thp*, *median*, *75thp*). The 1-s lag autocorrelation (*autocorr*) of the VM and the correlation between each axis were computed (*corrxy*, *corrxz*, *corryz*). Many of these features have been shown to be useful for identifying behaviors in previous literature (28). For each sample in the window, the roll, pitch and yaw angles of the direction of acceleration were computed, as roll = $\tan^{-1}(y, z)$, pitch = $\tan^{-1}(x, z)$ and yaw = $\tan^{-1}(y, x)$. The average (*avgroll*, *avgpitch*, *avgyaw*) and standard deviation (*sdroll*, *sdpitch*, *sdyaw*) of these angles were computed over the window. A low-pass filter with a cutoff frequency of 0.5 Hz (preliminary experiments tested a few cutoff frequencies and found 0.5Hz to perform best) was applied to the data window to estimate the average direction of gravity, and the roll, pitch and yaw angles of this direction were computed (*rollg*, *pitchg*, *yawg*) (14).

The fast Fourier transform (FFT) was applied to the VM to decompose the time domain signal v_t to its frequency components. The resulting power spectrum describes the contribution of a given frequency to the measured acceleration signal. The dominant frequency of the signal (*fmax*), *i.e.*, the frequency with highest power, and corresponding maximal power (*pmax*) were computed from the power spectrum. A similar calculation was done between the frequency bands of 0.3 Hz and 3Hz (*fmaxband*, *pmaxband*). The *entropy* of the frequency domain signal was computed. Finally, the power in each frequency band between 1 and 15 Hz (*fft1*-*fft15*) was computed.

Minute-level classification—Next, each feature vector was input into a random forest classifier that produces a predicted behavior label. A random forest classifier is an ensemble of randomized decision trees. Each decision tree is learned from a random sample of training examples and a random sample of features. Test examples are classified by averaging the output of each decision tree in the forest. The random forest algorithm was chosen because it has a number of advantageous properties: (i) feature selection is encompassed in the training phase, and non-informative features are easily ignored, (ii) it can model interactions between features, (iii) it is a type of ensemble learning, which makes it more robust to noise than a single tree, (iv) it does not require much fine-tuning of parameters, and (v) it is relatively fast to train (5). The importance of each feature to the random forest classifier can be calculated by iteratively holding out each feature and calculating the change in accuracy of the resulting classifier.

Time-smoothing—Finally, the sequence of probabilities output by the random forest classifier was smoothed over time by a hidden Markov model (HMM) to produce the final sequence of predicted behaviors. The HMM is a probabilistic model defined over a sequence of observations o_t, $t = 1, 2, \ldots T$. In our model the observation o_t represents the behavior prediction by the random forest classifier at time t. The HMM assumes that the observation at time t was generated by some process whose state h_t is hidden from the observer. In our model the hidden state h_t represents the true behavior at time t. The outputs were assumed to be independent given the hidden states, *i.e.* given h_t, o_t is independent of the hidden states and observations at all other time indices. Additionally, the sequence of hidden states was assumed to be a Markov process, *i.e.*, given the value of state h_{t-1}, the current state h_t is independent of all states prior to time $t-1$.

Hence the model was defined by three parameters. The observation matrix D represents the probabilities of each observation value given each hidden state, *i.e.*, the probabilities that the random forest will classify an example of a given behavior with each behavior label. The transition matrix B represents the probabilities of transitioning between behaviors in the hidden state sequence. Finally, the vector of initial probabilities π represents the a priori probability of each given behavior. The model parameters B, D and π were learned from the training data using maximum likelihood estimation. Test sequences of predictions from the random forest classifier were smoothed by using the Viterbi algorithm to infer the most likely sequence of hidden states. (24)

The HMM can smooth errors made by the random forest by explicitly modeling the probability of transitioning between activities. For example, it is very unlikely to transition directly from sitting to riding in a vehicle without a small bout of walking in between. The HMM can also model the duration of an activity bout via the self-transition probability, *i.e.* the probability that the hidden state will remain in a given behavior. Similarly, modeling the random forest errors through the observation probabilities allows the HMM smoother to place higher confidence on behaviors with good classification accuracy, while behaviors with lower classification accuracy might be more likely to be changed through the smoothing process.

Evaluation—The performance of our behavior classification algorithm was evaluated using leave-one-participant-out cross-validation.

Standard cut-points

Standard cut-points were used to categorize hip accelerometer counts per minute (CPM) into sedentary (< 100 CPM), light (100 – 1951 CPM) and MVPA (≥ 1952 CPM) (11). These analyses were processed using the ActiLife software (ActiGraph, Pensacola, FL).

GGIR

A new method called GGIR calibrates and analyzes raw triaxial accelerometer data. The R package GGIR (31) was used to calibrate accelerometer data, compute the metric Euclidean norm minus one (ENMO) for each minute of data, and estimate the number of minutes in MVPA (32). For hip data the threshold ENMO > 69 and for wrist data the threshold ENMO >100 was used to define minutes of MVPA. These thresholds were developed in a laboratory-based study of 30 healthy adults aged 18–65 (13).

Statistical Analysis

Hip accelerometer non-wear time was assessed using the Choi algorithm (6). Ground truth behavior labels, ML behavior predictions, CPM and ENMO values were matched by time on the minute-level at the hip and wrist. We compared minutes under the sedentary cutpoint to minutes of sedentary behaviors (sitting and riding in a vehicle), and compared minutes classified as MVPA to minutes of walking.

Since the hip accelerometer and SenseCam were not worn overnight, only comparisons for the available matched minutes were performed. Minutes per wear day in each behavior or intensity category were computed. Mixed effects linear regression analyses, adjusting for wear time and the nesting of days within participants, were used to compare the number of minutes/day in each method. Mixed effects linear regression was performed using the R package lme4 (3). All statistical analyses were performed in R (http://www.R-project.org/).

RESULTS

ML behavior classification

The sensitivity, specificity, and balanced accuracy (the mean of sensitivity and specificity) of the ML behavior classifiers using hip and wrist accelerometers are presented in Table 1. The ML behavior classifier using the hip accelerometer obtained on average 5% higher accuracy than the classifier using the wrist accelerometer. The mean minutes per day of behaviors predicted using each classifier are presented in Table 2. Predicted minutes of sitting and riding in vehicle were not significantly different from ground truth for either accelerometer position. Both hip and wrist ML classifiers significantly over-predicted minutes of standing. Bland-Altman plots (Figure 1) showing the difference between predicted and ground truth minutes per day of walking (on the participant level) indicate good agreement and that there is no bias with increasing time. The root mean squared error (RMSE) for the hip classifier was 11.5 min/day. For the wrist classifier, the RMSE was 13.3 min/day. Standing was most commonly confused for sitting by the hip classifier and most commonly confused for walking by the wrist classifier (Table 3). The wrist classifier under-predicted walking by an average of 4 minutes, while the hip classifier showed no significant difference in minutes of walking predicted. Figure 2 shows the importance of each feature used by the ML classifier. Differences between feature importance for the wrist and hip classifiers demonstrate that each algorithm is unique to the wear location. Notably, the hip classifier puts more trust in the features such as the standard deviation, coefficient of variation, and maximum, which might be less informative in noisier wrist data. Conversely, a feature such as power at the dominant frequency, which can extract information from a noisy signal, is more important to the wrist classifier.

Traditional hip accelerometer cut-points

The mean minutes per day of each intensity category are presented in Table 2. The 100 CPM cut-point applied to the hip accelerometer under-predicted minutes in sedentary behaviors by an average of 47 minutes. The majority of the CPMs (Table 4). The 1952 CPM cut-point for MVPA applied to the hip accelerometer under-predicted walking by 16 minutes. In fact, as Table 4 shows, the majority of walking minutes fell under the 1952 CPM cut-point. The Bland-Altman plots (Figure 1) show the under prediction of walking minutes, and increasing discrepancies as participants' walking time increases missed sedentary time was due to vehicle time with high. The RMSE was 20.4 min/day.

GGIR

The mean minutes per day classified as MVPA are presented in Table 2. Minutes of MVPA estimated by the hip and wrist accelerometer under-estimate walking by 11 and 10 minutes, respectively. The estimates are fairly consistent between the hip and wrist devices, with only one minute difference between the two estimates. No comparison for sitting time was available. Bland-Altman plots show that estimated MVPA and true minutes of walking are in relatively good agreement, although there is a slightly increasing underestimation with increasing walking minutes). The Hip accelerometer had a RMSE of 15.5 min/day and the wrist accelerometer had a RMSE of 17.5 min/day.

DISCUSSION

In this study, we presented a method to classify free-living behaviors from a hip-mounted or wrist-mounted accelerometer. This two-layer machine learning method improved classification over traditional cutpoint methods. The hip based classifier demonstrated higher minute-level accuracy than the wrist based classifier, although both classifiers correctly estimated minutes in sedentary behaviors and walking time.

The performance of both hip and wrist free-living ML classifiers was comparable to previous studies (9). Variations in performance can depend on the number of activities being classified and the amounts of signal and noise in the training data. For example, laboratory data can be classified with higher accuracy than observation protocol data (2), and observational protocol data can be classified with higher accuracy than free-living data (9, 17). However, classifiers trained on both laboratory and observational protocol data perform poorly when applied to free-living data. This study demonstrated that free-living data can be successfully classifier using a ML trained on free-living data.

These results demonstrate some limitations of traditional cut-point methods for categorizing hip accelerometer data into intensity categories. A sizeable portion of vehicle travel (29%) fell above the threshold for sedentary activity. This is unsurprising, as the cutpoints were developed in laboratory datasets that did not include vehicle time. However, in daily life vehicle time often constitutes a significant portion of sedentary time – in fact, the population of women measured in this study spent on average more than an hour a day in a vehicle. Researchers investigating the relationship between commuting and sedentary behavior will be especially interested in correctly capturing sedentary behavior in a vehicle. Previous work with wearable cameras showed high errors with self-reported travel diaries (15)—reliable detection of travel behaviors using accelerometers would allow larger-scale studies. Further, studies of commuting and pollution that compared sitting in vehicle to riding in traffic could benefits from these estimates. Additionally, the majority of standing time fell under the threshold for sedentary activity; an incorrect classification particularly for researchers interested in measuring standing as breaks from sitting.

A significant portion of walking fell under the cut-point for MVPA. The GGIR algorithm developed in the laboratory showed a similar pattern of predicting fewer minutes of MVPA than walking time. This is not surprising as much walking is not of moderate intensity. However, intensities of 1952 are not appropriate reflections of moderate in all populations, especially middle aged to older adults. Given the prevalence of walking/running as the preferred PA of many populations and the importance of walking for public health (e.g. Surgeon General’s Everybody walk! Initiative), it is important that we can accurately measure it. In this population of overweight and obese women, other forms of PA were seldom observed.

The ML classifier trained on free-living data accurately estimated the average minutes of sitting and walking behaviors over the population, with no significant difference from ground truth. Higher levels of arm movement than torso movement lead to wrist data being noisier than hip data and emphasize the need for more sophisticated data processing methods. While the hip ML algorithm performed better on the available data, additional wear time achieved with wrist devices might justify use of a slightly lower performing algorithm. Because we only had ground truth data for a limited wear time (on average 7 hours per day), we were not able to present the performance of the wrist algorithms over the full wear time.

This study demonstrated the validity of a ML behavior classifier trained on a homogeneous sample of 40 overweight and obese women. It should be noted that while the ML algorithms presented in this paper were tuned for the appropriate population, the cutpoint and GGIR methods were developed on data from laboratory studies of healthy adults. Additional work aims to understand the generalizability of these algorithms between heterogeneous populations (17). But this work provides a strong proof of concept and advances our understanding of the differences between wrist and hip placement of accelerometers for behavior monitoring. An additional source of limitation is the choice of a one-minute resolution for behavior classification. Minutes that consist of a combination of behaviors (*i.e.*, a transition from sitting to standing) will presumably be classified as the single behavior most represented (*i.e.*, either as sitting or standing). (Note that performance measures presented in this paper were assessed only on minutes containing a single behavior). Although traditionally count-based intensity measures have also been assessed on the minute level, future work should investigate methods to identify specific transition times between behaviors. An additional limitation was the use of a combination of walking and running into a single behavior—this was necessitated by the use of wearable cameras as a ground truth measure, as speed cannot be accurately assessed from still images, as well as a lack of running behaviors in our population. Future work can attempt to differentiate between these behaviors by including data from GPS devices to estimate speed.

The ML algorithms presented in this paper are available in an R package (cran.r-project.org/web/packages/TLBC). The package contains both functions to train behavior classification models on your own dataset, and models that have been pre-trained on our free-living dataset and are ready to apply to raw accelerometer data.

Intensity and behaviors are two different windows into physical activity measurement. While intensity has been the paradigm of choice for years past, the ability to measure behaviors has several advantages. Behavior-specific interventions require the ability to measure the specific behaviors being targeted. Additionally, such methods enable research on dose-response relationships between specific behaviors and health outcomes (26). Such findings can inform behavior-based public health guidelines that may be more understandable to the public. For example, a recommendation of 30 min/day of walking might be more easily understood and actionable than 30 min/day of MVPA as the public may struggle with intensity monitoring. Walking may need to be of a higher intensity or longer duration to impact health outcomes, but objective measures of walking have yet to be tested against outcomes in large cohort studies.

Acknowledgments

This study was supported by NIH grant 1R01CA164993-01A1.

References

1. Bassett DR Jr, Rowlands A, Trost SG. Calibration and validation of wearable monitors. Med Sci Sports Exerc. 2012; 44(1 Suppl 1):S32–S38. [PubMed: 22157772]
2. Bastian T, Maire A, Dugas J, et al. Automatic identification of physical activity types and sedentary behaviors from 3-axial accelerometer: lab-based calibrations are not enough. J Appl Physiol. 2015:jap-01189.
3. Bates D, Maechler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015; 67(1):1–48.
4. Bonomi AG, Plasqui G, Goris AH, Westerterp KR. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol. 2009; 107(3):655–61. [PubMed: 19556460]
5. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.
6. Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of Accelerometer Wear and Nonwear Time Classification Algorithm. Med Sci Sports Exerc. 2011; 43(2):357–364. [PubMed: 20581716]
7. Doherty AR, Kelly P, Kerr J, et al. Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. Int J Behav Nutr Phys Act. 2013; 10(22.10):1186.
8. Doherty AR, Moulin CJa, Smeaton AF. Automatically assisting human memory: a SenseCam browser. Memory. 2011; 19(7):785–95. [PubMed: 20845223]
9. Ellis, K.; Godbole, S.; Chen, J.; Marshall, S.; Lanckriet, G.; Kerr, J. Physical activity recognition in free-living from body-worn sensors. Proceedings of the 4th International SenseCam & Pervasive Imaging Conference; 2013 Nov; San Diego. 2013. p. 88-89.
10. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA Accelerometer. Med Sci Sports Exerc. 2011; 43(6):1085–93. [PubMed: 21088628]
11. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc. 1998; 30(5):777–781. [PubMed: 9588623]
12. He B, Bai J, Zipunnikov V, et al. Predicting human movement with multiple accelerometers using movelets. Med Sci Sports Exerc. 2014; 46(9):1859–66. [PubMed: 25134005]
13. Hildebrand M, Van Hees VT, Hansen BH, Ekelund U. Age-group comparability of raw accelerometer output from wrist-and hip-worn monitors. Med Sci Sports Exerc. 2014; 46(9):1816–24. [PubMed: 24887173]
14. Karantonis DM, Narayanan MR, Mathie M, Lovell NH, Celler BG. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans Inf Technol Biomed. 2006; 10(1):156–67. [PubMed: 16445260]
15. Kelly P, Doherty A, Mizdrak A, et al. High group level validity but high random error of a self-report travel diary, as assessed by wearable cameras. Journal of Transport & Health. 2014; 1(3):190–201.
16. Kerr J, Marshall SJ, Godbole S, et al. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. Am J Prev Med. 2013; 44(3):290–6. [PubMed: 23415127]
17. Kerr J, Patterson R, Ellis K, Godbole S, Johnson E, Lanckriet G, Staudenmayer S. Objective Assessment of Walking and Sitting: Classifiers for Public Health. Med Sci Sports Exerc. In press.
18. Lyden K, Kozey-Keadle S, Staudenmayer J, Freedson PS. A method to estimate free-living active and sedentary behavior from an accelerometer. Med Sci Sports Exerc. 2014; 46(2):386–97. [PubMed: 23860415]
19. Mannini A, Intille S, Rosenberger M, Sabatini AM, Haskell W. Activity recognition using a single accelerometer placed at the wrist or ankle. Med Sci Sports Exerc. 2013; 45(11):2193–203. [PubMed: 23604069]

20. Oshima Y, Kawaguchi K, Tanaka S, et al. Classifying household and locomotive activities using a triaxial accelerometer. Gait & posture. 2010; 31(3):370–4. [PubMed: 20138524]

21. Patterson RE, Colditz GA, Hu FB, et al. The 2011–2016 Transdisciplinary Research on Energetics and Cancer (TREC) initiative: rationale and design. Cancer Causes Control. 2013; 24(4):695–704. [PubMed: 23378138]

22. Patterson RE, Rock CL, Kerr J, et al. Metabolism and breast cancer risk: frontiers in research and practice. J Acad Nutr Diet. 2013; 113(2):288–96. [PubMed: 23127511]

23. Phillips LR, Parfitt G, Rowlands AV. Calibration of the GENEA accelerometer for assessment of physical activity intensity in children. J Sci Med Sport. 2013; 16(2):124–8. [PubMed: 22770768]

24. Rabiner LR, Juang BH. An introduction to hidden Markov models. ASSP Magazine, IEEE. 1986; 3(1):4–16.

25. Rosenberger ME, Haskell WL, Albinali F, Mota S, Nawyn J, Intille S. Estimating Activity and Sedentary Behavior From an Accelerometer on the Hip or Wrist. Med Sci Sports Exerc. 2013; 45(5):964–75. [PubMed: 23247702]

26. Sallis JF, Kerr J. Physical activity and the built environment. President’s Council on Physical Fitness and Sports Research Digest. 2006; 7(4):1–8.

27. Sallis JF, Owen N, Fotheringham MJ. Behavioral epidemiology: a systematic framework to classify phases of research on health promotion and disease prevention. Ann Behav Med. 2000; 22(4):294–8. [PubMed: 11253440]

28. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An Artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009; 107(4):1300–7. [PubMed: 19644028]

29. Strath SJ, Kaminsky LA, Ainsworth BE, et al. Guide to the assessment of physical activity: Clinical and research applications A scientific statement from the American heart association. Circulation. 2013; 128(20):2259–79. [PubMed: 24126387]

30. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008; 40(1):181–8. [PubMed: 18091006]

31. Van Hees, VT. GGIR: Raw Accelerometer Data Analysis. R package version 1.1–5. 2015. http://CRAN.R-project.org/package=GGIR

32. Van Hees VT, Gorzelniak L, Dean Leon EC, et al. Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. PLoS one. 2013; 8(4):e61691. [PubMed: 23626718]

33. Van Hees VT, Renström F, Wright A, et al. Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. PloS one. 2011; 6(7):e22922. [PubMed: 21829556]

34. Wijndaele K, Westgate K, Stephens SK, et al. Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. Med Sci Sports Exerc. 2015 Oct; 47(10):2129–39. [PubMed: 25785929]

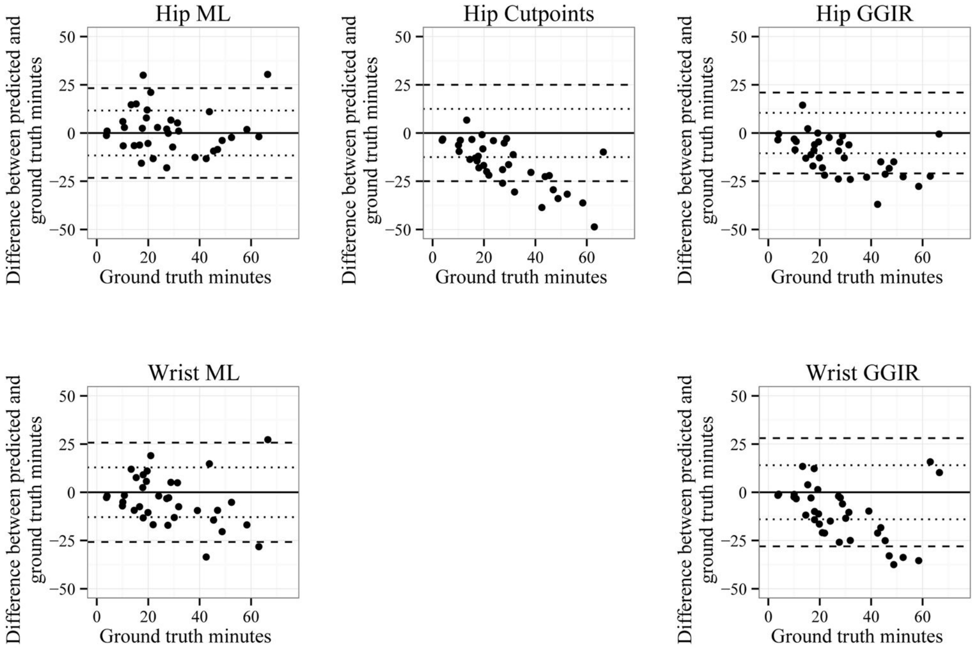

Figure 1.

Bland-Altman plots showing the difference between minutes of predicted walking/running or MVPA and ground truth minutes of walking/running, by participant, for the (a) hip accelerometer and (b) wrist accelerometer.

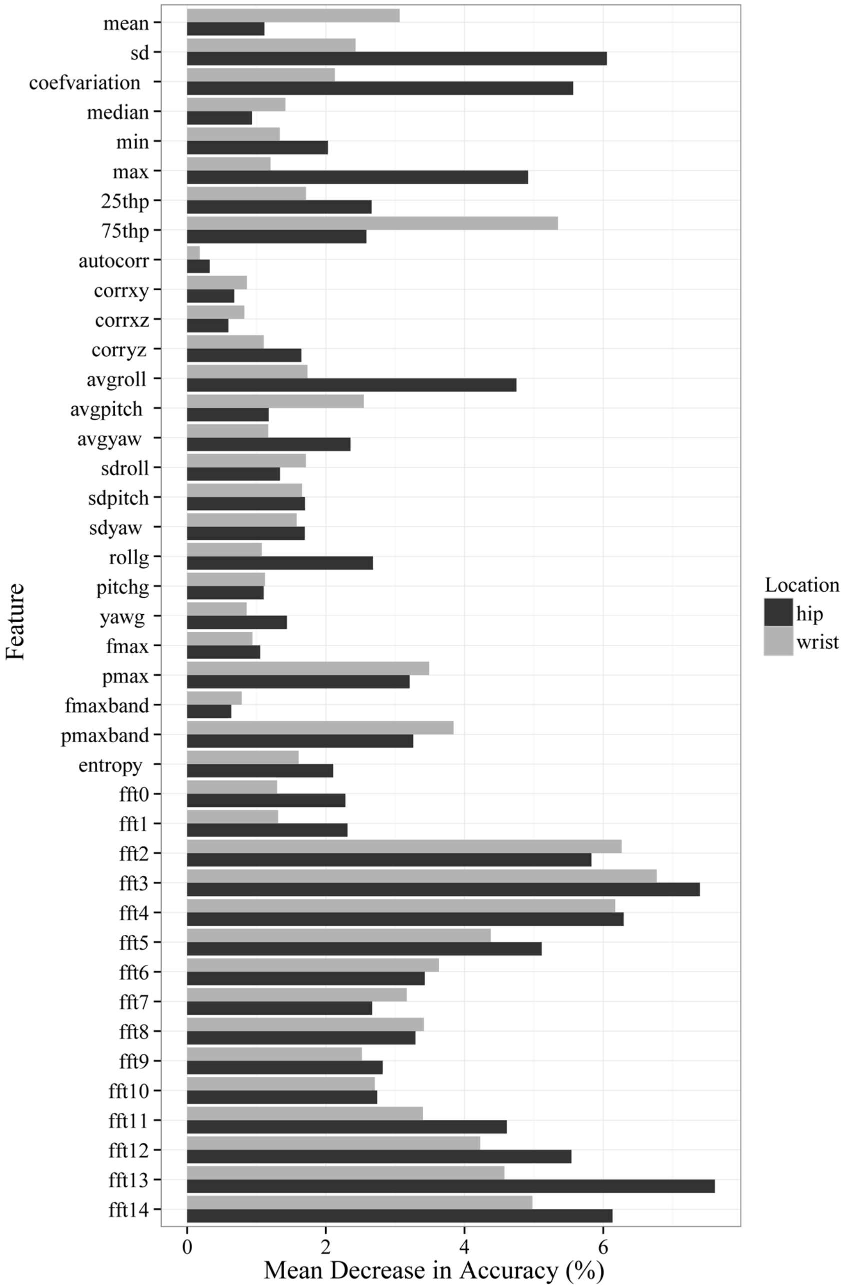

Figure 2.

Feature importance measured by mean decrease in accuracy of the random forest classifier

Table 1

Performance of machine learned classifiers, using leave-one-participant-out cross-validation

	Hip			Wrist		
	Sens	**Spec**	**BA**	**Sens**	**Spec**	**BA**
Sitting	0.894	0.923	0.908	0.883	0.870	0.876
Vehicle	0.870	0.987	0.929	0.823	0.964	0.893
Walking/Running	0.687	0.981	0.834	0.574	0.983	0.779
Standing	0.797	0.929	0.851	0.687	0.904	0.795
Average	0.812	0.955	0.881	0.742	0.930	0.836

*Sensitivity (sens), specificity (spec) and balanced accuracy (BA)

Table 2

Mean minutes per day of behaviors and intensities with various classification methods

		Mean (SE) Minutes per day[a,b]			
Ground truth behaviors		**Sitting** **274.4 (10.7)**	**Vehicle** **61.0 (5.5)**	**Standing** **64.0 (6.1)**	**Walking/Running** **28.4 (2.8)**
Hips	Cutpoints	**Sedentary** 288.8 (8.0)*			**MVPA** 12.0 (2.0)*
	GGIR				**MVPA** 17.1 (2.5)*
	ML behaviors	**Sitting** 262.4 (12.6)	**Vehicle** 56.3 (5.6)	**Standing** 81.0 (8.0)*	**Walking/Running** 28.5 (3.5)
Wrist	Cutpoints	**Sedentary** 118.4 (7.0)*			**MVPA** 72.0 (6.5)*
	GGIR				**MVPA** 18.1 (1.5)*
	ML behaviors	**Sitting** 264.6 (11.8)	**Vehicle** 64.8 (6.6)	**Standing** 75.6 (8.5)*	**Walking/Running** 24.5 (3.4)*

[a] Mixed model adjusted for wear time and nesting of days in participants

[b] Mean wear time = 428.9 min

*Significantly different from ground truth (Chi-square test, $p < 0.01$)

Table 3

Confusion matrix: ML behaviors classified by hip (H) and wrist (W) accelerometer as a percentage of minutes of true behaviors

ML Predicted behaviors	Ground truth behaviors							
	Sitting		Vehicle		Standing		Walking/Running	
	H	W	H	W	H	W	H	W
Sitting	89	88	6	9	11	21	3	5
Vehicle	1	3	87	82	1	4	2	6
Standing	9	8	5	8	80	69	27	32
Walking	1	1	2	1	8	6	69	57

Table 4

Confusion matrix: Intensity categories from hip accelerometer cut-points as a percentage of minutes of true behaviors

Cut-point intensities	Ground truth behaviors			
	Sitting	**Vehicle**	**Standing**	**Walking/Running**
Sedentary	90	70	44	6
Light	10	29	55	55
MVPA	0	0	0	39