HHS Public Access

Author manuscript

Med Sci Sports Exerc. Author manuscript; available in PMC 2017 May 01.

Published in final edited form as:

Med Sci Sports Exerc. 2016 May ; 48(5): 941–950. doi:10.1249/MSS.0000000000000844.

Performance of Activity Classification Algorithms in Free-living Older Adults

Jeffer Eidi Sasaki[1], **Amanda Hickey**[1], **John Staudenmayer**[2], **Dinesh John**[3], **Jane A. Kent**[1], and **Patty S. Freedson**[1]

[1]Department of Kinesiology, University of Massachusetts, Amherst, MA

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

[3]Department of Health Sciences, Northeastern University, Boston, MA

Abstract

Purpose—To compare activity type classification rates of machine learning algorithms trained on laboratory versus free-living accelerometer data in older adults.

Methods—Thirty-five older adults (21F and 14M ; 70.8 ± 4.9 y) performed selected activities in the laboratory while wearing three ActiGraph GT3X+ activity monitors (dominant hip, wrist, and ankle). Monitors were initialized to collect raw acceleration data at a sampling rate of 80 Hz. Fifteen of the participants also wore the GT3X+ in free-living settings and were directly observed for 2-3 hours. Time- and frequency- domain features from acceleration signals of each monitor were used to train Random Forest (RF) and Support Vector Machine (SVM) models to classify five activity types: sedentary, standing, household, locomotion, and recreational activities. All algorithms were trained on lab data (RF_{Lab} and SVM_{Lab}) and free-living data (RF_{FL} and SVM_{FL}) using 20 s signal sampling windows. Classification accuracy rates of both types of algorithms were tested on free-living data using a leave-one-out technique.

Results—Overall classification accuracy rates for the algorithms developed from lab data were between 49% (wrist) to 55% (ankle) for the SVM_{Lab} algorithms, and 49% (wrist) to 54% (ankle) for RF_{Lab} algorithms. The classification accuracy rates for SVM_{FL} and RF_{FL} algorithms ranged from 58% (wrist) to 69% (ankle) and from 61% (wrist) to 67% (ankle), respectively.

Conclusion—Our algorithms developed on free-living accelerometer data were more accurate in classifying activity type in free-living older adults than our algorithms developed on laboratory accelerometer data. Future studies should consider using free-living accelerometer data to train machine-learning algorithms in older adults.

Keywords

Machine learning algorithms; pattern recognition; wearable activity monitor; ActiGraph GT3X+

Corresponding author: Patty Freedson, Mailing address: Department of Kinesiology, University of Massachusetts, 30 Eastman Lane, 111 Totman Building, Amherst, MA – 01003, Phone: 413-545-2620, Fax: 413-545-2906, psf@kin.umass.edu.

Conflicts of Interest: There are no conflicts of interest to disclose.

The results from the present study do not constitute endorsement by the American College of Sports Medicine.

Introduction

Classification of activity type from acceleration sensors is not a new concept in computer science (4, 5). However, it has only been recently explored in the physical activity assessment field (15, 19). Technological advances in activity monitors have allowed researchers to collect raw acceleration data (g force) at high sampling rates (e.g., 100Hz), bringing new data analytic possibilities for physical activity assessment. High-resolution acceleration data reveal signal patterns for different activities and increase the potential for using machine-learning algorithms to classify activity type (12).

Currently, most studies developing algorithms for classification of activity type have been conducted in laboratory settings and have generally used semi-structured activities (fixed duration and standard instructions) (2, 15, 18, 19). These conditions facilitate the achievement of high algorithm accuracy and provide evidence for the proof of concept. However, there is little evidence about algorithm performance in free-living settings where they will ultimately be implemented. In fact, there has been little progress in developing and validating activity type classification algorithms in free-living conditions.

A segment of the population for which activity type classification may be important is in older adults to identify activity type behaviors that may be linked to risk of disability. To date, physical performance battery tests have been used to identify older adults with functional limitations and, thus, at increased risks for physical disability (6, 13). While important, these tests describe the performance of specific activities (e.g., walking, sit-to-stand movements, balance stances) but not the daily activity and sedentary behaviors that lead to better or worse scores. Improving physical activity assessment methods in free-living older adults will lead to a better understanding of real functional abilities in older adults. Obtaining information about activity type, as well as frequency and duration of certain activity types (e.g., locomotion, household tasks, sit-to-stand transitions) will provide better means for assessing the risks for physical disability. Moreover, activity classification algorithms may help to characterize patterns of physical activity in older adults, generating empirical evidence about activity types that should be included in behavioral lifestyle interventions for older adults.

Thus, the purposes of this study were to: 1) develop and test the accuracy of lab-based algorithms in detecting activity type in free-living older adults, and 2) to develop and evaluate algorithms from free-living accelerometer data.

Methods

The study was conducted in two parts. In part 1, we developed and tested laboratory-based algorithms for classification of activity type from raw acceleration signals in older adults. Part 2 was used to test the lab-based algorithms in free-living older adults and to develop and evaluate algorithms trained using free-living accelerometer data. All study methods and procedures were approved by the Institutional Review Board of the University of Massachusetts Amherst.

Part 1: Laboratory protocol

Participants—Thirty-five healthy older adults were recruited from Amherst, MA and surrounding areas. Exclusion criteria for this study included: 1) age <65 or >85 years, 2) use of any ambulatory assistive device, 3) lower extremity mobility impairment, or 4) presence of any condition that significantly affected ambulation.

Informed consent visit—Volunteers visited the Physical Activity and Health Laboratory at the University of Massachusetts Amherst and provided written informed consent. They completed a health history questionnaire and the Physical Activity Readiness Questionnaire (PAR-Q). Participants then completed the Short Physical Performance Battery test (SPPB) and were required to achieve a score of 12 in order to rule out the presence of lower extremity mobility impairment. Detailed description of the SPPB test is provided elsewhere (6). Medical clearance was obtained from each participant's physician before scheduling the activity routine trial.

Activity monitor—Participants were fitted with three ActiGraph GT3X+ (ActiGraph, LLC, Pensacola, FL) activity monitors positioned at the dorsal and distal aspects of the dominant wrist, aligned with the anterior axillary line of the dominant hip, and above the lateral malleolus of the dominant ankle. The monitors were secured to the body locations using an elastic belt (hip) and two cotton velcro straps (wrist and ankle). The ActiGraph GT3X+ is a lightweight monitor (4.6cm × 3.3cm × 1.5cm, 19g) that measures triaxial acceleration ranging from -6g to +6g. We initialized the monitors to sample triaxial acceleration signals at a rate of 80 Hz, which is similar to what is used in the NHANES activity monitoring study (9, 11).

Activity Routine Visit—Once fitted with the monitors, participants performed 30 s of each of the following postures: a) stationary standing, b) sitting, and c) lying down. There were no specific instructions provided about how these postures were performed. Subsequently, participants performed one of two activity routines:

Activity routine 1: (a) Crossword puzzles (sitting), (b) miscellaneous self-care, (c) organizing the room, (d) gardening, (e) carrying groceries, (f) 400 m walk, and (g) Tai-chi

Activity routine 2: (a) Playing cards while sitting, (b) folding laundry, (c) dusting, (d) vacuuming, (e) slow walk (0.8 $m \cdot s^{-1}$), (f) 400 m walk, and (g) simulated bowling.

Each activity was performed for five minutes. Immediately after data collection, accelerometer data were downloaded to a laptop using ActiLife 5.0 software (ActiGraph, LLC, Pensacola, FL).

Part 2: Free-Living Protocol

Participants—Fifteen individuals who completed the laboratory protocol volunteered to participate in the free-living protocol. After a verbal explanation about the protocol and its objectives, participants provided written informed consent.

Direct Observation System—A Personal Digital Assistant (PDA) programmed for continuous focal sampling direct observation (CFS-DO) (The Observer®; Noldus Information Technology, Wageningen, The Netherlands) was used to code the activities performed by the participants in the free-living environment. The PDA and CFS-DO software were used to capture two activity variables:

1. Activity type – The menu of activities for the PDA comprised five group categories: a) standing, b) sedentary, c) locomotion, d) household, and e) recreational.
2. Activity duration: The software recorded the activity duration when a participant changed the activity type as noted by the observer.

Activity Monitors—Similar to the lab-based protocol, three ActiGraph GT3X+ (ActiGraph Inc, Pensacola, FL) activity monitors were synchronized to a laptop computer and initialized using the ActiLife 5 software to collect raw triaxial acceleration signals (g) at a sampling rate of 80 Hz.

Observers—Three observers were trained to use the PDA and the continuous focal sampling software. They received instructions on coding activity type and duration during face-to-face training sessions and group discussion meetings. At the end of the training period, the observers completed a test to examine inter-observer reliability. The test consisted of coding activity type and duration of twenty activity video clips. The *Cohen's kappa coefficient* for inter-observer agreement for coding activity type and duration was 0.89.

Direct Observation—A single daytime block of 2-3 hours of direct observation (DO) was carried out for each participant. The observers met the participants at the pre-determined time and location (e.g., home, laboratory, senior center, gym) and assisted participants with placement of the monitors (dominant wrist, hip and ankle) before starting the DO session. Participants were then instructed to perform their daily routine as normally as possible. Observers coded activities as either 1) standing, 2) sedentary, 3) locomotion, 4) household, 5) recreational, or 6) private time. The latter was used when participants went to the bathroom/restroom and/or needed time alone. During the training process, observers were instructed to code free-living activities following the same standards of activity categorization adopted for laboratory data described below.

Feature Extraction, Data Processing and Algorithm Development

Laboratory algorithms—Raw acceleration signals from the three activity monitors (hip, wrist, and ankle) were labeled by one of the investigators (JES) according to the individual activity type and activity category (described below). A start and stop marker was used to label signals corresponding to the exact times of the activities. Individual activities were categorized as:

Standing: (a) stationary standing

Sedentary: (a) lying down, (b) sitting, (c) crossword puzzles and (d) playing cards

Locomotion: (a) slow walk, (b) self-paced 400 m walk and (c) carrying groceries

Household: (a) dusting, (b) gardening, (c) vacuuming, (d) self-care (miscellaneous), (e) laundry and (f) organizing the room

Recreational: (a) Tai-chi and (b) simulated bowling

These categories were based on body posture, movement patterns (e.g., intermittent, continuous), and the amount of movement observed. Thus, activities such as crossword puzzles and playing cards that contextually are classified as 'recreational activities' were classified as 'sedentary activities' due to predominance of the sitting posture. Similarly, while 'household' and 'recreational' activities might have similar movement patterns, we classified activities as 'recreational' if they were sport-related or structured activities (i.e., simulated bowling and Tai-chi).

Following data reduction and labeling, visual inspection was performed to ensure alignment of signals to the corresponding activities. Examples of acceleration signals for lab-based and free-living activities with similar coding for a single participant are shown in Figure 1. Time- and frequency-domain features (obtained using a Fourier transform) for acceleration signals were extracted for each orthogonal axis (vertical, antero-posterior, medio-lateral) and also for the vector magnitude in 20 s windows.

Time-domain features included the 1) 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} percentiles of acceleration signals (g), 2) mean acceleration, 3) standard deviation of acceleration.

Frequency-domain features included: 1) 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} percentiles of signal frequency, 2) range of frequency distribution, 3) total signal power, 4) mean frequency, 5) dominant frequency, 6) power of dominant frequency, 7) second dominant frequency, 8) power of second dominant frequency, 9) dominant frequency between 0.6 – 2.5 Hz (df625), 10) power of df625, 11) entropy, 12) entropy density, and 13) ratio noise/signal.

Acceleration features were used as the input features while activity labels were used as the outcome variables for training 'Support Vector Machines' and 'Random Forest' algorithms to classify activity type (SVM_{Lab} and RF_{Lab}). Previous studies have used these techniques and reported high recognition rates for activity type (12, 15, 19). A description of these machine-learning techniques can be found elsewhere (10, 12). Algorithms for each technique were developed using hip, wrist and ankle data. We developed algorithms using a combination of time- and frequency- domain features to classify activity type in sequential windows of 20 s (sliding windowing technique), with no detection of activity transitions.

Free-living algorithms—Similar to laboratory algorithms, time- and frequency- domain features for the acceleration signals were extracted for each orthogonal axis (vertical, antero-posterior, medio-lateral) and also for the vector magnitude in 20 s windows. These features were used with the direct observation labels to train Support Vector Machines and Random Forest algorithms to classify activity type from free-living accelerometer data (SVM_{FL} and RF_{FL}). Algorithms were developed for hip, wrist, and ankle data and were trained to classify activity type in sequential windows of 20 s, with no detection of activity transitions. As oppose to lab-based activities that lasted for a fixed duration and for which each minute of activity resulted in three sequential windows of 20 s, free-living activities were variable in

duration and some did not last for a minimum of 20 s. Therefore, activities lasting less than 20 s were excluded from the training set. Figure 1 demonstrates acceleration signals for laboratory and free-living activities with similar coding for a single participant. It is worth noting that we plotted signals for activities that were as similar as possible between the laboratory and free-living conditions. However, given the broader scope of activities that may occur in free-living conditions and the general classification adopted in DO (sedentary, standing, household, locomotion, and recreational), it was only possible to match laboratory and free-living acceleration signals based on activity category (e.g., lab-based household vs. free-living household) but not for specific activities (e.g., lab-based vacuuming vs. free-living vacuuming).

Software for Developing and Testing the Algorithms—The open source *R statistical software package*, version 3.0.1 - "Good Sport" (www.r-project.org) was used for developing and evaluating the algorithms. Packages 'e1071′, and 'Random Forest' were used for developing the SVM and RF algorithms. For the SVM models, a radial-basis kernel was used and all other parameters from the 'e1071′ package were kept in their initial default configuration. For the RF models, the number of trees was set to 500.

Statistical Evaluation

The 'leave-one-out' validation technique was used to determine percent correct classification rates for the algorithms. Performance of lab-based algorithms was first evaluated in laboratory and subsequently in free-living conditions. Algorithms developed from free-living data were only tested in free-living conditions.

Overall percent correct classification by the algorithms was calculated as follows: *% correct = (minutes in different activity categories that are correctly identified by the algorithm ÷ total minutes of the protocol) × 100.* To calculate percent correct classification by the algorithms for a specific activity category (e.g. locomotion), the equation was as follows: *% correct for 'locomotion' = (minutes in locomotion that are correctly identified by the algorithm ÷ total minutes in locomotion) × 100.* Similar to previous studies, 80% of correct classification rate was considered an acceptable accuracy level for our algorithms (15, 17, 19). In addition to percent correct classification, the *Kappa Statistic* was used to verify agreement between actual and predicted activity classification taking into account agreement that occurs by chance alone (16). Common interpretation of *Kappa* values (*k*) is usually based on the scale by Landis and Koch (8), where the following categorization is used: 1) <0 = less than chance agreement, 2) 0.01 - 0.20 = slight agreement, 3) 0.21 – 0.40 = fair agreement, 4) 0.41 – 0.60 = moderate agreement, 5) 0.61 – 0.80 = substantial agreement, 6) 0.81 – 0.99 = almost perfect agreement. Confusion matrices for identification of misclassified minutes across the different activity groups were tabulated for the free-living algorithms (SVM_{FL} or RF_{FL}) with the best classification rate for each monitor placement. Sensitivity (proportion of true events that are correctly classified) and specificity (proportion of false events that are correctly classified) for classifying the different activity groups were also calculated for the algorithms with the best performances. Recognition rates of the algorithms were also tested using sampling windows varying from 5 to 30 s.

Results

Participant characteristics (n = 35; 21F and 14M) for the laboratory protocol were: age=70.6±5.0 years, body mass=76.4±14.4 kg, height=168.6±9.9 cm, BMI=26.8±4.2 $kg{\cdot}m^{-2}$, and 400 m walk speed= 1.17±0.18 $m{\cdot}s^{-1}$. Participant characteristics (n = 15, 9F and 6M) for the free-living protocol were: age=70.0±4.3 years, body mass=74.5±11.5 kg, height=169.8±9.9 cm, BMI=26.0±4.3 $kg{\cdot}m^{-2}$, and 400 m walk speed= 1.20±0.2 $m{\cdot}s^{-1}$. There were no significant differences between those participating in the laboratory protocol and the subsample that completed the free-living protocol.

Laboratory Algorithms

In laboratory conditions, the overall activity type classification accuracy rates of the SVM_{Lab} hip, SVM_{Lab} wrist, and SVM_{Lab} ankle algorithms were 87%, 96%, and 90%, respectively. The overall accuracy rates of the RF_{Lab} hip, RF_{Lab} wrist, and RF_{Lab} ankle algorithms, were 87%, 94%, and 89%, respectively. *Kappa* values ranged from 0.82 (SVM_{Lab} hip and RF_{Lab} hip) to 0.94 (SVM_{Lab} wrist). Table 1 provides correct algorithm classification rates for the different activity categories. The lowest recognition rates were for standing and recreational activities. The SVM_{Lab} hip and RF_{Lab} hip algorithms correctly classified 0% of the standing minutes (total of nine minutes) while the recognition rates by the SVM_{Lab} ankle and RF_{Lab} ankle algorithms for standing were 20% and 40%. Conversely, correct recognition rates by the SVM_{Lab} wrist and RF_{Lab} wrist algorithms for standing were 80% and 82%, respectively. For recreational activities, the highest recognition rates were 92% (RF_{Lab} wrist) and 94% (SVM_{Lab} wrist); with the remaining recognition rates ranging from 53-64% for this activity type category. Locomotion, sedentary behavior, and household activity were correctly classified over 90% of the time.

In free-living conditions, overall accuracy rates of the lab-based SVM algorithms were substantially lower than in laboratory. The recognition rates by the SVM_{Lab} hip, SVM_{Lab} wrist, and SVM_{Lab} ankle algorithms were 49%, 49%, and 55% (Table 2). Similarly, recognition rates by the RF_{Lab} hip, RF_{Lab} wrist, and RF_{Lab} ankle algorithms were 51%, 49%, and 54% (Table 2), respectively. The RF_{Lab} and SVM_{Lab} algorithms performed extremely poorly for standing (accuracy range: 0-1%) and recreational activity (13-26%), poorly for locomotion (33-52%), and reasonably well for sedentary behavior (62-79%) and household activity (71-87%). Overall agreement between algorithms and DO were at best ‘fair’, according to *Kappa* values (Table 2, $k = 0.34 – 0.39$).

Free-living Algorithms

Descriptive summaries for DO data and labeled accelerometer data are provided in Table 3. A total of 1770.9 min of DO (∼29.5 h) was used, with sedentary, household, and locomotion activities accounting for ∼74% (∼22 h) of the total DO time. Household activity occurred most frequently (Total number of observations = 549; 36.6 ± 18.1 observations per participant), while private time was the least frequent code used by observers (Total number of observations = 10; 0.7 ± 0.8 observations per participant). When labeling accelerometer data, there were small declines in the number of minutes for each activity category (e.g., locomotion: 349.6 min from DO vs. 341.9 min from accelerometer). These reductions were

caused by elimination of observations that lasted for less than the selected window size for training the algorithms. Once trained, the SVM_{FL} algorithms performed substantially better than the SVM_{Lab} algorithms, with recognition rates of 64% ($k=0.53$), 58% ($k=0.44$), and 69% ($k=0.59$) for the SVM_{FL} hip, SVM_{FL} wrist, and SVM_{FL} ankle algorithms, respectively (Table 2). Recognition rates of the RF_{FL} algorithms were 66% ($k=0.56$), 61% ($k=0.48$), and 67% ($k=0.57$) for RF_{FL} hip, RF_{FL} wrist and RF_{FL} ankle algorithms (Table 2). Table 2 displays performance of free-living RF and SVM algorithms for each activity group. Overall, the algorithms trained with free-living data (SVM_{FL} and RF_{FL}) improved the detection for standing (10-52%), recreational activity (20-41%), locomotion (65-80%), and for sedentary behavior (75-87%). However, there was a small decrease in accuracy for household activity (63-73%).

When classification interval was increased to 30 s, the recognition rates of the SVM_{FL} hip, SVM_{FL} wrist, and SVM_{FL} ankle algorithms increased to 65% ($k=0.56$), 59% ($k=0.46$), and 71% ($k=0.62$) (Table 4). Reducing classification interval to 10 s or 5 s resulted in consistent reduction in recognition rate of the SVM_{FL} algorithms (Table 4). For 30 s classification intervals, the SVM_{FL} algorithms performed poorly for classifying standing and recreational activity. The SVM_{FL} hip, SVM_{FL} wrist, and SVM_{FL} ankle correctly classified standing only 45%, 10% and 53% of the time. For recreational activity, classification accuracy was 22%, 21% and 40% for SVM_{FL} hip, SVM_{FL} wrist, and SVM_{FL} ankle. Accuracy rates for the other activity groups ranged from 66% to 87% (Table 4).

For the RF_{FL} algorithms, increasing classification interval to 30 s resulted in improved classification accuracy for RF_{FL} ankle algorithms (+3%) and there were very small increases for RF_{FL} hip and RF_{FL} wrist algorithm (+1% and +3%) compared to 20 s classification interval (Table 4). Reducing classification interval to 10 or 5 s resulted in lower accuracy rates for all three RF_{FL} algorithms (Table 4). The RF_{FL} algorithms accuracy rates were low for recreational activity and standing. For 30 s classification intervals, recreational activities were correctly classified only 25%, 24%, and 39% of the time by the RF_{FL} hip, RF_{FL} wrist, and RF_{FL} ankle algorithms (Table 4). Similarly, standing was correctly classified only 40%, 10%, and 50% of the time by the RF_{FL} hip, RF_{FL} wrist, and RF_{FL} ankle algorithms (Table 4).

Table 5 presents the confusion matrices as well as sensitivity and specificity analyses for the SVM_{FL} ankle, RF_{FL} hip, and RF_{FL} wrist algorithms (30 s classification interval), which were the algorithms exhibiting the highest overall correct classification rates (71%, 68% and 63%). The lowest accuracy rate of the SVM_{FL} ankle was for recreational activity. Of a total of 134 min of recreational activity, the algorithms correctly classified only 53 min and misclassified 21, 12, 31 and 17 min as household activity, locomotion, sedentary behavior and standing, respectively. The highest algorithm accuracy was for sedentary behavior with 412 min correctly classified and only 62 min of misclassifications (Table 5, panel a). Sensitivity of the SVM_{FL} ankle algorithm for the different activity groups varied from 54% (standing) to 86% (locomotion). Specificity of the algorithm ranged from 89% (household) to 95% (recreational) (Table 5, panel a). Similar patterns of activity misclassifications were observed for the RF_{FL} hip algorithm, which only correctly classified 34 min of recreational activity (from a total of 136 min), misclassifying 32, 31, 16, and 23 min as household

activity, locomotion, sedentary behavior and standing, respectively (Table 5, panel b). The highest algorithm accuracy was in identification of sedentary behavior with 391 min being correctly classified, while 82 min were misclassified either as recreational, household, or standing activities. Sensitivity of the RF_{FL} algorithm for the different activity groups varied from 47% (recreational and standing) to 82% (locomotion), while specificity ranged from 89% (household and standing) to 95% (locomotion) (Table 5, panel b).

The lowest classification accuracy for the standing category was observed for the RF_{FL} wrist algorithm. The RF_{FL} wrist algorithm only correctly classified 24 min (from a total of 251 min), whereas the SVM_{FL} ankle and RF_{FL} hip algorithms correctly classified 134 and 102 min of standing (from totals of 251 and 252 min), respectively (Table 5, panel C).

Discussion

This is the first study to train machine learning algorithms on laboratory and free-living accelerometer data and test their accuracy in classifying activity type in free-living older adults. Our results demonstrated that lab-based algorithms performed poorly in free-living conditions while algorithms developed with free-living accelerometer data improved activity type recognition rates. However, none of the algorithms achieved our preset acceptable accuracy level of 80%, which has been reported in previous studies that have developed activity classification algorithms from laboratory data (15, 17, 19).

Our results indicate that high accuracy for activity recognition from wearable accelerometer algorithms developed in laboratory settings does not translate into high accuracy in free-living conditions. Previous studies have reported high recognition rates for their algorithms for pre-defined activity routines comprising activities of fixed durations in laboratory conditions (1, 2, 15, 19). These studies have defined classification intervals (e.g., 30 s, 1 min) according to the pre-defined duration of the activities (e.g., 5 min, 6 min), which means that an algorithm that classifies activity type for 1-min intervals will have a perfect match of five classification events for a 5-min activity. This approach usually results in significant inflation of classification accuracy of algorithms. As demonstrated in the current study, substantial degradation of algorithm accuracy occurs in free-living conditions where activities are not fixed in duration. Acceleration signals from free-living conditions are substantially different than those collected in lab environment (Figure 1), providing evidence for the inaccuracy of lab-based algorithms applied to free-living accelerometer data.

Past studies have already reported similar trends in their results, demonstrating that accuracy of algorithms developed on laboratory data decay when applied to free-living conditions. Gyllensten and Bonomi (7) showed reductions of ∼16-20% in accuracy of three machine learning models when they were tested in free-living conditions. The decay in accuracy of our algorithms was ∼40-46% and some factors may have contributed to the greater decline observed in our study, including type of activity monitor, monitor placement, training dataset (theirs: >246 hours, ours: ∼29 hours), and the criterion measure used. An important observation made by Gyllensten and Bonomi (7) was that activities in free-living conditions exhibit a higher degree of overlapping characteristics in their acceleration features when

compared to activities performed in the lab. This may partially explain why locomotion in our study was markedly misclassified as household activity in free-living conditions.

The use of training data from free-living conditions may offer a possible solution to improve algorithms accuracy. This was highlighted in a previous study by Ermes et al. (3), who reported substantial improvement (∼17%) in the accuracy of their machine learning algorithms when free-living data were included in the training dataset. We employed this same approach and observed similar improvements in accuracy of our SVM and RF algorithms (9-14%). Overall, this approach resulted in more balanced recognition rates for sedentary, locomotion, and household activities, ranging from 62% to 87%. However, our algorithms yielded high misclassification rates of household activities with locomotion, standing with household activities, and recreational activities with all other categories. Perhaps, future studies will need to develop models that use transition points to define sampling window boundaries rather than sliding windows (12). In addition, hybrid models that first detect transition points and then apply probabilistic techniques, such as hidden Markov modeling, may also be beneficial for improving recognition rates of the algorithms (9, 12). These and other types of approaches will need to be verified in further studies in free-living settings. Classifying short duration activities will require more than just reducing classification intervals as we have shown that reducing fixed intervals from 20 s to 10 or 5 s did not improve recognition rate of algorithms in free-living conditions.

Finally, the results from this study reveal that monitor placement may influence correct classification rates of the algorithms in free-living conditions. Although classifying standing and recreational activities was problematic for all monitor placements (Tables 2, 4 and 5), hip and ankle placements appear slightly better than wrist placement for overall classification of activity type. However, the differences in recognition rates ranged from 2 to 11%, precluding any definitive conclusion as to which placement is superior for classifying activity type in free-living conditions. In 2011, the NHANES adopted the wrist as the standard monitor placement for monitoring physical activity in Americans (11); however, no studies have validated methods to analyze raw acceleration data from the wrist in free-living conditions, only in laboratory (9, 14, 19). Future studies will need to further examine if activity type is accurately classified from wrist accelerometer data in free-living settings.

This study is not without limitations. Our current DO system does not allow for post observation coding, which is important for correcting miscoded data or for improving labeling of acceleration signals. The possibility of recoding activities using video records may be beneficial to address this limitation. With our system, very short duration activities and transitions may have been missed whereas a video system would allow these activities and transitions to be identified and coded. As mentioned in the methods section, the activity categorization employed in the DO was very general (sedentary, standing, household, locomotion, and recreational) and could be expanded and refined if we had used a video system. In addition, video systems provide the possibility of ‘acquiring’ data for longer periods and with lower burden to researchers. Aside from allowing free-living data collection for prolonged periods, video systems may also result in less reactivity from participants, which may facilitate collecting data for a broader scope of activities. Thus, future studies need to consider this alternative when collecting free-living accelerometer data

for training activity type classification algorithms. Another limitation of this study was the number of participants in the free-living portion of this study. We observed 15 participants for a total of 29.5 h. It may be necessary to collect data on a larger number of participants in order to increase representativeness of activities that the algorithms can classify. In addition, testing algorithms in an independent sample may be required, as part of the reason for the higher accuracy rates of the free-living algorithms may have been due to testing the algorithms on the same sample they were developed on. Nevertheless, the free-living participants were a subsample of the participants enrolled in the laboratory protocol, meaning that their data were also used for developing the laboratory algorithms. This fact provides an indication that the superiority of the free-living algorithms is not solely due to dependence/independence of the samples used for developing and testing algorithms. Furthermore, we minimized the 'sample dependence' effect by using a leave-one-out validation technique.

In conclusion, the results of this study suggest that our algorithms are currently not sufficiently accurate for assessment of free-living PA in older adults. Our algorithms presented overall activity recognition rates lower than 80%, which is the acceptable accuracy level suggested in previous studies (15, 17, 19). Thus, it is necessary to further improve the accuracy of the algorithms, which may be possible by implementing a modified direct observation system that allows for recoding of data. This may allow for applying algorithms that detect point of transition to improve differentiation of the start and end of activities, and thus, minimizing confusion by the algorithms. Based on the current study, we recommend that future studies develop activity type classification algorithms using free-living accelerometer data and also employ a more sophisticated direct observation system as a criterion measure.

Acknowledgments

This research was funded in part by a grant from the National Institutes of Health R01 CA121005.

The authors would like to thank all the participants from the study and the Senior Centers from Amherst, Belchertown, Erving, Northampton, South Hadley, and South Deerfield.

References

1. Bonomi AG, Goris AHC, Yin B, Westerterp KR. Detection of type, duration, and intensity of physical activity using an accelerometer. Med Sci Sports Exerc. 2009; 41(9):1770–7. [PubMed: 19657292]
2. DeVries SI, Garre FG, Engbers LH, Hildebrandt VH, Van Buuren S. Evaluation of Neural Networks to Identify Types of Activity Using Accelerometers. Medicine & Science in Sports & Exercise. 2011; 43(1):101–7. [PubMed: 20473226]
3. Ermes M, Parkka J, Mantyjarvi J, Korhonen I. Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions. IEEE Transactions on Information Technology in Biomedicine. 2008; 12(1):20–6. [PubMed: 18270033]
4. Fahrenberg J, Foerster F, Smeja M, Müller W. Assessment of posture and motion by multichannel piezoresistive accelerometer recordings. Psychophysiology. 1997; 34(5):607–12. [PubMed: 9299915]
5. Foerster F, Fahrenberg J. Motion pattern and posture: correctly assessed by calibrated accelerometers. Behav Res Methods Instrum Comput. 2000; 32(3):450–7. [PubMed: 11029819]

6. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. J Gerontol. 1994; 49(2):M85–94. [PubMed: 8126356]
7. Gyllensten IC, Bonomi AG. Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. IEEE Trans Biomed Eng. 2011; 58(9):2656–63. [PubMed: 21712150]
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. [PubMed: 843571]
9. Mannini A, Intille SS, 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]
10. Mannini A, Sabatini AM. On-line classification of human activity and estimation of walk-run speed from acceleration data using support vector machines. Conf Proc IEEE Eng Med Biol Soc. 2011; 2011:3302–5. [PubMed: 22255045]
11. NHANES. [2013 Oct 23] NHANES 2011-2012 - Manuals, Brochures, and Consent Documents. [Internet][date unknown]Available from: http://www.cdc.gov/nchs/nhanes/nhanes2011-2012/current_nhanes_11_12.htm
12. Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R. Activity identification using body-mounted sensors--a review of classification techniques. Physiol Meas. 2009; 30(4):R1–33. [PubMed: 19342767]
13. Rikli R, Jones C. Development and Validation of a Functional Fitness Test for Community-Residing Older Adults. J Aging Phys Activ. 1999; 7:129–61.
14. Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. J Appl Physiol. 2015; 119(4):396–403. [PubMed: 26112238]
15. 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. Journal of Applied Physiology. 2009; 107(4):1300–7. [PubMed: 19644028]
16. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37(5):360–3. [PubMed: 15883903]
17. Welch WA, Bassett DR, Thompson DL, et al. Classification accuracy of the wrist-worn gravity estimator of normal everyday activity accelerometer. Med Sci Sports Exerc. 2013; 45(10):2012–9. [PubMed: 23584403]
18. Zhang S, Murray P, Zillmer R, Eston RG, Catt M, Rowlands AV. Activity classification using the GENEA: optimum sampling frequency and number of axes. Med Sci Sports Exerc. 2012; 44(11): 2228–34. [PubMed: 22617400]
19. Zhang S, Rowlands AV, Murray P, Hurst TL. Physical activity classification using the GENEA wrist-worn accelerometer. Med Sci Sports Exerc. 2012; 44(4):742–8. [PubMed: 21988935]

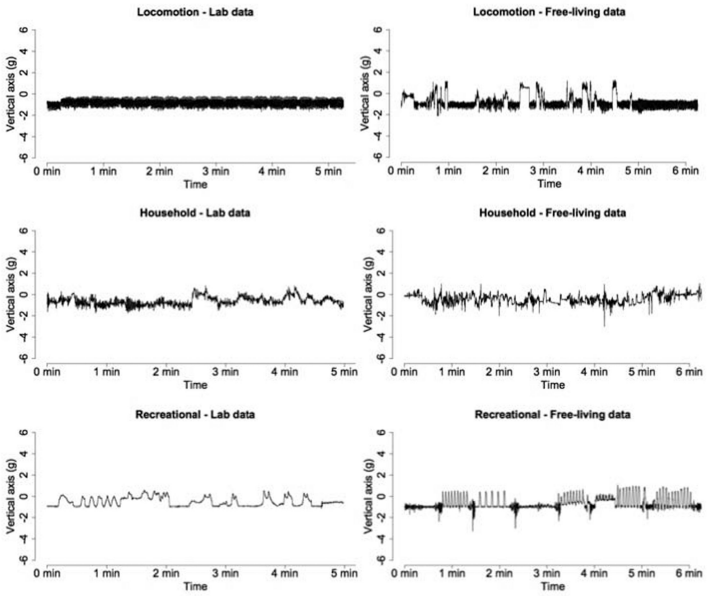

Figure 1. Laboratory and free-living accelerometer signals from the wrist for locomotion, household and recreational activities for a single participant

Signals are raw acceleration (g) from vertical axis collected at a sampling rate of 80 Hz. Each panel shows acceleration for a different activity category (see panel title). *Y-axis* of each graph depicts acceleration and *x-axis* of each graph shows time in minutes. Note: 1) Lab-based accelerometer signals for locomotion, household, and recreational activities correspond to self-paced 400 m walk, vacuuming, and Tai-chi, respectively. 2) Free-living accelerometer signals were broadly coded as sedentary, standing, household, locomotion, and recreational. It is possible and quite likely that free-living accelerometer signals depicted in this figure do not perfectly correspond to accelerometer signals from the same specific activity performed in the laboratory.

Table 1

Percent correct classification by the lab-based algorithms in laboratory conditions

	Activity Category (% correct)						Agreement Statistic
	Loc	**Sed**	**House**	**Rec**	**Stand**	**Overall**	***kappa***
SVM_{Lab} hip	98	92	91	55	0	87	0.82
SVM_{Lab} wrist	97	97	96	94	80	96	0.94
SVM_{Lab} ankle	99	92	93	64	20	90	0.86
RF_{Lab} hip	99	92	91	53	0	87	0.82
RF_{Lab} wrist	96	93	95	92	82	94	0.92
RF_{Lab} ankle	99	89	92	61	40	89	0.84

Values are percent of total time correctly identified for each activity group across all participants. Overall percent correct classification (second *column* from right to left) is percent correct classification across all activities and participants. First cluster of rows displays recognition accuracy for support vector machine algorithms developed with laboratory accelerometer data (SVM_{Lab}). Last cluster of rows displays recognition accuracy for random forest algorithms developed with laboratory accelerometer data (RF_{Lab}). Within each cluster, each row represents performance of an algorithm developed with accelerometer data from a different monitor placement. Loc: Locomotion, Sed: Sedentary, House: Household, Rec: Recreational, Stand: Standing.

Note: Overall correct classification is not simply the mean of the classification rates for the different activity categories because the number of minutes in each category was not the same.

Med Sci Sports Exerc. Author manuscript; available in PMC 2017 May 01.

Table 2

Performance of lab-based and free-living SVM and RF algorithms in free-living conditions

		Activity Category (% correct)						Agreement Statistic
		Loc	Sed	House	Rec	Stand	Overall	*kappa*
Laboratory Algorithms	SVM_{Lab} hip	49	68	71	13	0	49	0.34
	SVM_{Lab} wrist	33	73	72	21	1	49	0.33
	SVM_{Lab} ankle	37	79	87	20	0	55	0.39
	RF_{Lab} hip	52	62	82	17	0	51	0.36
	RF_{Lab} wrist	34	71	73	26	1	49	0.33
	RF_{Lab} ankle	39	76	87	19	0	54	0.39
Free-living Algorithms	SVM_{FL} hip	76	82	63	24	38	64	0.53
	SVM_{FL} wrist	65	75	73	20	10	58	0.44
	SVM_{FL} ankle	72	87	65	41	52	69	0.59
	RF_{FL} hip	81	81	68	25	37	66	0.56
	RF_{FL} wrist	74	81	70	22	22	61	0.48
	RF_{FL} ankle	70	84	64	40	51	67	0.57

Values are percent of total time correctly identified for each activity group across all participants. Overall percent correct classification (second column from right to left) is percent correct classification across all activities and participants. Lab: laboratory, FL: free-living. First two clusters of rows display recognition accuracy for algorithms developed with laboratory accelerometer data (SVM_{Lab}, RF_{Lab}). Last two clusters of rows display recognition accuracy for algorithms developed with free-living accelerometer data (SVM_{FL}, RF_{FL}). Within each cluster, each row represents performance of an algorithm developed with accelerometer data from a different monitor placement. Loc: Locomotion, Sed: Sedentary, House: Household, Rec: Recreational, Stand: Standing.

Note: Overall correct classification is not simply the mean of the classification rates for the different activity categories because the number of minutes in each category was not the same.

Med Sci Sports Exerc. Author manuscript; available in PMC 2017 May 01.

Table 3

Descriptive information from direct observation and labeled accelerometer data

	Direct Observation (DO)				Labeled accelerometer data	
	Total minutes	**Total number of events**	**Mean duration per event (min)**	**Frequency per participant (events)**	**Total minutes**	**No. of 20s windows**
Standing	264.5	94	2.81 ± 2.11	6.3 ± 4.4	254.9	765
Sedentary	503.3	109	4.62 ± 6.39	7.3 ± 5.9	476.7	1430
Locomotion	349.6	177	1.98 ± 5.08	11.8 ± 9.5	341.9	1025
Household	455.4	549	0.83 ± 0.92	36.6 ± 18.1	404.6	1213
Recreational	141.0	52	2.71 ± 2.34	3.5 ± 7.3	135.7	407
Private	57.1	10	5.72 ± 7.49	0.7 ± 0.8	NA	NA

Note: For each activity category, the difference between total minutes from DO and total minutes from labeled accelerometer data is the amount of data that were lost in the data reduction process.

Table 4

Performance of SVM_{FL} and RF_{FL} algorithms in free-living conditions according to different window length for classification

		Activity Category (% correct)						Agreement Statistic
	Window	Loc	Sed	House	Rec	Stand	Overall	*kappa*
SVM algorithms								
SVM_{FL} hip	5 s	64	79	59	19	26	56	0.42
SVM_{FL} wrist		62	70	68	13	9	53	0.38
SVM_{FL} ankle		54	78	59	39	43	59	0.48
SVM_{FL} hip	10 s	70	82	59	23	32	61	0.49
SVM_{FL} wrist		64	74	71	19	10	56	0.41
SVM_{FL} ankle		60	82	63	38	48	63	0.53
SVM_{FL} hip	20 s	76	82	63	24	38	64	0.53
SVM_{FL} wrist		65	75	73	20	10	58	0.44
SVM_{FL} ankle		72	87	65	41	52	69	0.59
SVM_{FL} hip	30 s	75	82	66	22	45	65	0.56
SVM_{FL} wrist		67	78	75	21	10	59	0.46
SVM_{FL} ankle		77	87	70	40	53	71	0.62
RF algorithms								
RF_{FL} hip	5 s	71	72	61	16	28	57	0.44
RF_{FL} wrist		69	73	67	16	12	56	0.41
RF_{FL} ankle		53	75	59	38	43	58	0.46
RF_{FL} hip	10 s	73	77	65	17	36	62	0.50
RF_{FL} wrist		71	77	69	22	12	59	0.44
RF_{FL} ankle		58	81	61	38	46	62	0.51
RF_{FL} hip	20 s	81	81	68	25	37	66	0.56
RF_{FL} wrist		74	81	70	22	22	61	0.48

Med Sci Sports Exerc. Author manuscript; available in PMC 2017 May 01.

		Activity Category (% correct)						Agreement Statistic
	Window	**Loc**	**Sed**	**House**	**Rec**	**Stand**	**Overall**	***kappa***
RF_{FL} ankle		70	84	64	40	51	67	0.57
RF_{FL} hip		81	83	71	25	40	68	0.58
RF_{FL} wrist	30 s	76	84	73	22	10	63	0.50
RF_{FL} ankle		78	86	68	39	50	70	0.61

Recognition rates for the Support Vector Machine (top) and Random Forest (bottom) algorithms developed with free-living accelerometer data (SVM_{FL}, RF_{FL}). Values are percent of total time correctly identified for each activity group across all participants. Overall percent correct classification (second column from right to left) is percent correct classification across all activities and participants. Each cluster of rows displays performance of the algorithms according to a different classification interval. Within each cluster, each row represents performance of an algorithm developed with accelerometer data from a different monitor placement. Loc: Locomotion, Sed: Sedentary, House: Household, Rec: Recreational, Stand: Standing.

Note: Overall correct classification is not simply the mean of the classification rates for the different activity categories because the number of minutes in each category was not the same.

Med Sci Sports Exerc. Author manuscript; available in PMC 2017 May 01.

Table 5

Confusion matrices along with sensitivity and specificity values for free-living algorithms presenting the highest recognition rate for each monitor placement

a. SVM_{FL} Ankle Algorithm

		Predicted				
		Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**
Actual	**Recreational**	53	21	12	31	17
	Household	6	282	28	24	60
	Locomotion	2	73	260	1	2
	Sedentary	10	17	2	412	33
	Standing	2	81	0	34	134

Overall accuracy: 71%
(95% CI: 70% - 73%)

	Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**
Sensitivity	74%	59%	86%	82%	54%
Specificity	95%	89%	94%	94%	91%

b. RF_{FL} Hip Algorithm

		Predicted				
		Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**
Actual	**Recreational**	34	32	31	16	23
	Household	10	283	26	32	50
	Locomotion	4	58	274	0	2
	Sedentary	7	35	0	391	40
	Standing	16	75	1	58	102

Overall accuracy: 68%
(95% CI: 66% - 69%)

	Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**

	Sensitivity	47%	59%	82%	79%	47%
	Specificity	93%	89%	95%	93%	89%

c. RF_{FL} Wrist Algorithm

		Predicted				
		Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**
Actual	**Recreational**	33	28	16	42	16
	Household	4	293	21	57	26
	Locomotion	2	76	256	0	3
	Sedentary	6	52	1	396	18
	Standing	11	133	3	80	24

Overall accuracy: 63% (95% CI: 61% - 65%)

	Recreational	**Household**	**Locomotion**	**Sedentary**	**Standing**
Sensitivity	58%	50%	85%	69%	32%
Specificity	93%	89%	94%	92%	85%

Panel a: SVM_{FL} ankle algorithm (30 s classification interval); **Panel b:** RF_{FL} Hip algorithm (30 s classification interval); **Panel c:** RF_{FL} Wrist algorithm (30 s classification interval). **Top part of each panel:** Rows are actual activities and columns are predicted activities. Values are in minutes and combined for all participants. Shaded values are correctly classified minutes and values outside the diagonal line (shaded) are misclassified minutes. **Middle part of each panel:** Overall accuracy indicates the percent correct classification of the algorithm for combined data of all activities and participants. 95% CI indicates the upper and lower bound of correct classification for 95% of the participants. **Lower part of each panel:** Values are percent of detection by the algorithm. **Note:** Sensitivity identifies the number of true events that are correctly classified as such. Specificity identifies the number of false events that are correctly classified as false events. SVM: support vector machine; RF: Random Forest; FL: Free-living.