Original Article

Comparison of Several Algorithms to Estimate Activity Counts with Smartphones as an Indication of Physical Activity Level

V.H. Rodriguez [a,*], C. Medrano [a,b], I. Plaza [a,b], C. Corella [c], A. Abarca [c], J.A. Julian [c]

[a] EduQTech, E.U. Politecnica, c/ Atarazana 2, 44003, Teruel, Spain
[b] Instituto de Investigación Sanitaria de Aragón, Universidad de Zaragoza, Zaragoza, Spain
[c] EFYPAF, Facultad de Ciencias Sociales y Humanas, c/ Atarazana 2, 44003, Teruel, Spain

HIGHLIGHTS

- Four algorithms to get GT3X+ activity counts from accelerometer signals are compared.
- The area under the rectified curve shows a strong linear relation with counts.
- Other algorithms with more elaborate filtering do not achieve better results.
- Obtaining the area under the rectified curve leads to the lower power consumption.
- Smartphones could be used to test fulfillment of physical activity recommendations.

GRAPHICAL ABSTRACT
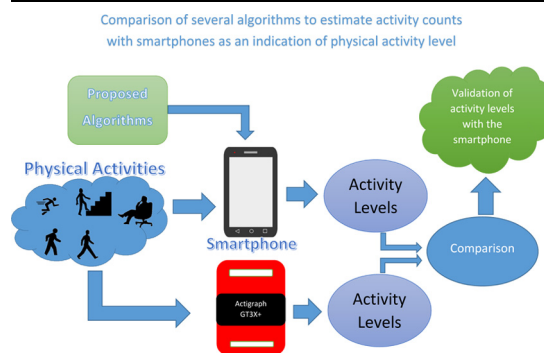
Comparison of several algorithms to estimate activity counts with smartphones as an indication of physical activity level

ABSTRACT

Background: Dedicated devices like GT3X+, Actical or ActivPal have been widely used to measure physical activity (PA) levels by using cut-points on activity counts. However, the calculation of activity counts relies on proprietary software. Since smartphones incorporate accelerometers they are suitable candidates to determine PA levels in a wider population.

Objective: Our aim was to compare several algorithms so that smartphones can reproduce the results obtained with GT3X+. The influence of smartphone location was also investigated.

Methods: Volunteers participated in the experiment performing several activities carrying two smartphones (hip and pocket) and one GT3X+ (hip). Four algorithms (A1–A4) were considered to obtain GT3X+ counts from smartphone accelerometer signals. A1 was based on a traditional filtering on temporal domain and a posterior calculation of the area under the curve. A2 was based on computing histograms of acceleration values, which were used as independent variables in a standard linear regression procedure. A3 also used a linear regression, but in this case the independent variables were power spectrum bands, leading to a kind of filtering in the frequency domain. A4 was based on a direct measure of area under the rectified curve of the raw accelerometer signal. Performance was measured in terms of raw activity counts or the corresponding PA level classification. The influence of the algorithm was tested with a Quade test. Multiple comparisons were performed with Wilcoxon test with Bonferroni's correction. Besides, battery consumption was also measured as a secondary parameter. The output of the selected algorithm was compared with GT3X+ counts using correlation (pearson and spearman) and agreement (Intra-Class Coefficient, ICC and Bland–Altmann plots for raw counts, and weighted kappa for activity levels). Several experimental conditions regarding smartphone location were compared with Wilcoxon tests.

* Corresponding author.
E-mail addresses: victor_rod_ont@hotmail.com (V.H. Rodriguez), ctmedra@unizar.es (C. Medrano), inmap@unizar.es (I. Plaza), criscor@unizar.es (C. Corella), aabarca@unizar.es (A. Abarca), jajulian@unizar.es (J.A. Julian).

Results: Thirty-two volunteers participated in the experiment. More refined algorithms based on filtering techniques did not prove to achieve better performance than A2 or A4. In terms of classification of PA level, A4 got the lowest error rate, although in some cases the differences with other algorithms were not statistically significant (p-value > 0.05). A4 is also the simplest and the one that implies less battery depletion. The comparison of A4 with GT3X+ gave good agreement (ICC = 0.937) and correlation (spearman = 0.927) for raw counts and good agreement when classifying four or two PA levels (weighted kappa = 0.874 or 0.923 respectively). Besides, in real situations, activity classification into four levels was significantly improved (p-value < 0.05) if data from several body locations were used to find model parameters.

Conclusions: Simple algorithms can reproduce the results of GT3X+. Thus, smartphones could be used to control the fulfillment of PA recommendations previously validated with cut-points. However, it must be acknowledged that accelerometers are not the gold standard to measure PA.

1. Introduction

The appearance of cardiovascular diseases and other diseases such as diabetes, colon cancer, obesity, hypertension, osteoporosis and depression has been related to the level of physical activity (PA) [1]. Thus, measuring PA in real life is a relevant research topic. Ideally, methods for evaluating PA should measure energy expenditure (EE) during PA, as well as the type, duration and intensity of these activities. Usually EE is determined by the consumption of oxygen and exhaled carbon dioxide. However, the equipment required for this method is usually very costly and it is found only in specialized areas [2]. Considering everyday environments, researchers have opted for the use of inertial sensors such as accelerometers or gyroscopes, which can record body movements. Due to relation between PA and movement, accelerometers can be used to measure EE [2]. They present several advantages: low cost, low battery consumption, small size or integration into wearable devices, making them ideal to measure activities of daily living [3,4]. Nevertheless, they present several disadvantages: they underestimate EE at higher intensities, they generally do not identify the type of physical activity and they cannot directly derive energy consumption or activity levels. For that purpose, it is necessary to use an extrapolation formula [5].

Dedicated devices based on accelerometers have been developed to measure PA. The accelerometers of Actigraph are one of the most used to measure PA, together with Philips (Actical) and PALtechnologies (ActivPal) [6,7]. Given the wide use of Actigraph devices and the previous experience of the authors with them, the Actigraph GT3X+ was selected for this study [8]. The accelerometer signal is converted by the companion software into activity counts, which can be used to determine activity levels using validated cut-points, for instance following the values proposed by Sasaki et al. [9] for adults (labeled as Freedson VM3 in ActiLife software). Determining activity levels is of great importance because several PA recommendations have been developed based on them. The Centers for Disease Control and Prevention, the American College of Sports Medicine, the American Heart Association and the World Health Organization have used the daily minutes spent in moderate to vigorous physical activity as measurement criterion to establish the minimum levels of PA to obtain health benefits [10,11]. Therefore, most PA instruments, subjective as questionnaires and objective as accelerometers try to extract Moderate-Vigorous Physical activity (MVPA) minutes.

On the other hand, most commercial smartphones include accelerometers. Thus, they can be used to monitor activity and several applications have been developed in this field: pedometers [12], PA level and EE estimation [13–16] and human activity recognition [17]. In particular, in [14] a high correlation was found between raw activity counts measured with Actigraph and several smartphones in laboratory test. In free-living the correlation was lower, but the minutes in MVPA per day still showed a high correlation. In [15] an application to measure PA was developed, Cal Fit. After 5 days of wearing GT3X+ and the smartphone running the application, a high correlation was found between vertical axis counts and a high agreement too between duration and energy expenditure, using an equation to transform counts to metabolic equivalents (METS).

From a wider perspective, tracking physical activity with smartphones or dedicated devices connects with a new trend called the Quantified Self (QS). It refers to engagement in the self-tracking of several information sources (biological, physical, behavioral, or environmental), either as individuals or in groups, with the aim of improving self-sensing, self-awareness, and human performance within the digital health industry [18]. The last research shows that QS can positively influence the performance, improve awareness about nutrition and facilitate learning insights about one's body and mind [19]. In the case of smartphones, Bert et al. [20] highlighted positive aspects like the possibility of accessing health information and allowing communications between the health care world and the population. Potential relevant applications include the management of chronic diseases, fight against obesity and bad habits or the promotion of healthy lifestyles. However, QS has also some drawbacks. Bert et al. [20] warned about problems like control of contents with respect to scientific validity, gap in digital literacy and privacy data protection, which is also pointed out in [19]. Shin and Biocca [18] indicate problems with poor user experience, motivation or accuracy. If the information shown is not adequate, a negative effect with the follow-up of the activity is expected. Therefore, it is sometimes difficult to provide engagement. These authors recommend presenting the information in comparative and textual formats.

This work was motivated by the possibility of measuring PA continuously in an intervention to promote PA among university students lasting several months. This measure should complement a PA monitoring using GT3X+ at some specific weeks. However, when facing this problem, several questions were raised. The definition of activity counts is not clear. The Actigraph web page includes general ideas about activity counts, but no detailed description. Activity counts seem to be obtained from a band pass filter and an integration [21–23]. In [24] a study was carried out to get filter characteristics from the action of Actilife software on raw acceleration signals. In [4] a mechanical oscillator was used to find the filter shape. In other studies, activity counts were extracted by a low pass filter and a calculation of the area under the curve [14]. In [25] the mean amplitude deviation (MAD) of raw acceleration was proposed as a method to better quantify PA regardless of the accelerometer brand. In [26] a new index was proposed, the activity index (AI), which is an additive and rotational invariant metric. It was shown to be a better classifier of physical activity intensity

than activity counts. However, the sample population in which it was tested was restricted to women over 60 years old.

Therefore, we decided to carry out the present study with the goal of checking several algorithms to process smartphone accelerometer data and compare them in order to better reproduce results from Actigraph accelerometer. As a secondary goal, the effect of the smartphone placement on the body was be investigated from a practical point of view.

This paper extends our preliminary work [27] in the following aspects: i) One more activity is presented; ii) three more algorithms to estimate PA have been implemented and tested; iii) the algorithms were actually implemented in android applications and iv) more volunteers were involved in the experiment.

2. Materials and methods

2.1. Subjects

The subjects for this experiment were thirty-two volunteers aged from 18 to 45 years old (19 males, 13 females, age 27.88 ± 12.90, height 1.68 ± 0.12 meters, weight 66.26 ± 21.55 kg). All subjects were healthy following Chen and Sun sample selection [28], with no evidence of past or present different disorders or injuries. The study protocol was approved by the Ethical Committee for Clinical Research of Aragon (CEICA). All subjects received oral and written information about the study, and written informed consent was obtained from them.

2.2. Experiments

Volunteers were asked to wear a smartphone in the right pocket and another on the right hip while doing some activities under the guidance of the research team. Just one GT3X+ was placed on the right hip, next to the smartphone to validate and compare the results of the tests performed (see Fig. 1). Each activity lasted 10 min and was performed only once in a random order. Nine of the volunteers avoided running. Then, the corresponding file was labeled with the date, time and type of activity.

Specifically, the activities performed by the subjects were the following:

- Walking at low speed (self-selected speed)
- Walking at high speed (self-selected speed)
- Going up and down stairs (self-selected speed)
- Running (self-selected speed)
- Working at office (mainly being sitting in front of a computer but with small movements allowed, selected by each volunteer; thus is equivalent to a sedentary activity most of the time)

Walking and running activities were carried out in the athletics track facility of our university, while going up and down stairs was performed in the building of our university center. Working at office was simulated in our laboratory. Speed was not measured in the experiment. By allowing each volunteer to select a different speed we expected to obtain a larger variety of movements and counts.

To record the activities, we used the SAMSUNG GALAXY TREND PLUS GT-S7580 smartphone with range ± 2 g. We developed a specific app to record accelerometer values. When configuring the accelerometer in the app, the option "SENSOR_DELAY_GAME" was selected, resulting in a sampling period of 20 ms (50 Hz) but with some variability since smartphones cannot deal with strict timing restrictions. The information obtained was dumped to a text file in the smartphone that was then transferred to a PC, in which algorithms were implemented for off-line analysis using Matlab.

Fig. 1. Location of the devices.

Smartphone acceleration was interpolated at 30 Hz for comparison with Actigraph. At the same time, the GT3X+ device also recorded the information with a sampling frequency of 30 Hz and a range of ± 6 g, which served to validate the samples captured by the smartphone. Some of the Actigraph devices can only sample at 30 Hz in raw mode and this is the frequency used in [9] for stablishing cut-points. The devices were manually synchronized by adjusting smartphone date and time with that of the GT3X+, which in turn was synchronized with a PC through the ActiLife software. The algorithms were also implemented in Android (see section 2.5).

2.3. Calculating counts

2.3.1. Algorithm 1 (A1): filtering on temporal domain

In the first step, the vector magnitude was calculated from each three-axial acceleration sample Eq. (1). It contains information about the intensity of the activity carried out by the subject in a rotational invariant magnitude.

$$Vm_i = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{1}$$

Subsequently, the vector magnitude fed a band pass filter [21], which removes the gravitational component and the high frequency noise. The cutoff frequencies were 0.25 and 2.5 Hz for the high pass and low pass respectively. These values correspond to the range in which most human activities are performed [24]. In this way, we have tried to follow as closely as possible the known information about the calculation of counts in devices like GT3X+ (ActiLife 6 User's Manual Appendix A) [22].

A recursive filter with infinite impulse response (IIR) of order 2 was used to implement the filter digitally. The filter coefficients that best fit the counts provided by the GT3X+ were obtained by scanning different attenuation values at high and low cutoff frequencies (f_{c1} and f_{c2}) and band pass frequency (fp).

Once the signal was filtered, it was divided into 1 second epochs (30 samples in total, as in the configuration used by the GTX3+) resulting in an array Epc_{ij}, where j represents the epoch, and i is the sample number in that epoch.

Taking the 1 second vector magnitude in epochs (Epc), the area under the curve (AUC) for each epoch was calculated as:

$$AUC_j = \frac{1}{L} \sum_{i=1}^{L} |Epc_{ij}| \qquad (2)$$

where L = size of the window, i = sample between the epoch, j = Epoch number of 1 sec, AUC_j = area under the curve for the j-th epoch.

In [21] it is indicated that, once the area under the curve for each epoch is calculated, the number of counts are obtained by dividing this value by the accelerometer resolution. However, as we have no information about the range and resolution of the accelerometer used in a generic smartphone, this part has been omitted. Instead, to compare with the output of Actigraph, the counts returned by GT3X+ were multiplied by 0.002929 (which is the resolution of this device, since its accelerometer has a range of ±6 g and 12 bits for analog conversion, $0.002929 \approx 12/2^{12}$) [21], thus making the measure independent of the accelerometer resolution. In the remaining of this paper, GT3X+ counts will be always subjected to this processing. Finally, the counts per epoch (second) were added in groups of 60 to obtain counts per minute (cpm), since this is the official measure provided by Actigraph.

2.3.2. Algorithm 2 (A2) – fit to histogram values

We have also tried some algorithms that allow a standard fitting procedure, without relying on a previous assumed filter shape. In this case, histograms of acceleration values were used as input variables, while cpm provided by GT3X+ devices were considered as the output variables. Histograms are one of the features used in activity recognition [29], so it seems reasonable that they can include a clue about the activity level. The process to obtain these histograms was the following. Acceleration magnitude values were split into 60 s windows. Then, the mean value of the acceleration in the window was subtracted, to avoid the influence of accelerometer bias, which it was noticed to be high in some smartphone models. The acceleration axis was divided into several levels and the number of acceleration samples in each one was extracted.

The histogram bins were normalized by dividing them into the total number of samples. Then, some of the bins were used as input parameters of the model (h_i, $i = 1 \ldots 15$, 15 intervals in the range from -0.9 g to 2.1 g). Only a subset of the bins were used to avoid perfect collinearity in the model input variables since their sum was 1 after normalization. The Actigraph counts per minute were fitted as:

$$cpm = b_0 + \sum b_i h_i \qquad (3)$$

In this way, no other restriction is set on coefficients b_i ($i = 0 \ldots 15$) and we let a standard linear fit procedure to find them.

2.3.3. Algorithm 3 (A3) – fit to power spectrum bands

As in the previous case, a standard fitting algorithm was tested, in which the input variables were taken as the power contained in bands, while the output variable was the cpm value returned by the Actigraph. Thus, the smartphone acceleration magnitude samples contained in windows of 60 s were subjected to a Fourier transform in order to obtain the power spectrum. Then, seven bands contained in the interval [1 Hz, 3.3 Hz] were calculated (P_i, $i = 1 \ldots 7$) by integrating power in each interval. Fig. 2 shows an example. The left figure is an inset of a 60 s acceleration window. The right is the corresponding power spectrum. Band limits are shown as red dashed lines. The value of one of the bands is obtained as the area highlighted in blue.

Finally, Actigraph counts were fitted as a function of band values:
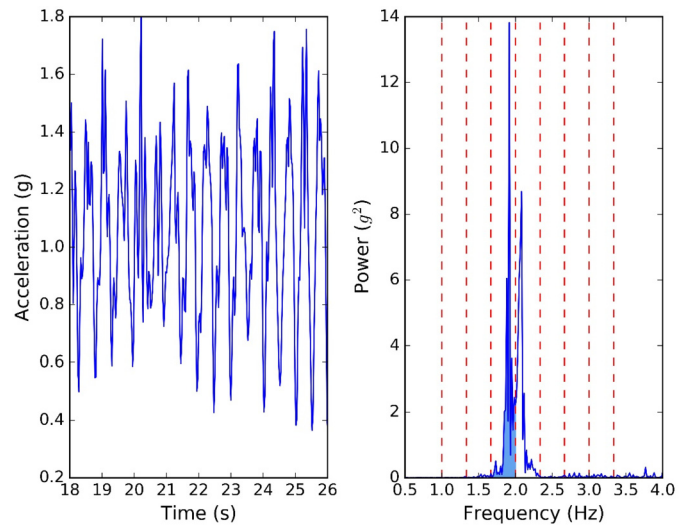
$$cpm = c_0 + \sum c_i P_i \qquad (4)$$

Fig. 2. Example of extraction of power bands. Left: acceleration signal; right: power spectrum with bands. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 1
Different conditions regarding smartphone location used to fit model parameters and test the models.

	C1	C2	C3	C4
Data used to fit	Hip	Pocket	Hip + Pocket	Hip
Data used to test	Hip	Pocket	Hip + Pocket	Hip + Pocket

2.3.4. Algorithm 4 (A4) – area of the accelerometer signal

In this case, the algorithm was based on a simple calculation of the area under the signal, similar to [14]. The signal magnitude acceleration was calculated and split into 60 s windows. Then, the mean value in each window was subtracted, and the resulting signal rectified. The area under the curve of the rectified signal (AUCR) was obtained as a measurement of activity. Finally, a linear relationship was found between area and Actigraph cpm:

$$cpm = d_0 + d_1 AUCR \qquad (5)$$

2.4. Smartphone location

The comparison of the four algorithms was performed using data from both locations, the hip and the pocket. After selecting one algorithm, the influence of the phone location was investigated for it.

This influence can be studied in several ways. Thus, we created a set of conditions to perform the comparison. These conditions differ in the smartphone raw acceleration data used to fit and test the model. Table 1 indicates these conditions, which are referred to as C1 to C4.

The set C1 (only hip) corresponds to the conditions in which the GT3X+ is used. The conditions C2 (only pocket) are more common in everyday use of the phone. It is expected that the performance would improve if the phone is located on the hip (C1) since our reference values are GT3X+ counts. Thus, we compared C1 and C2 to find the location with higher performance.

However, in our opinion the above comparison is more academic than practical. We feel that nobody would wear the smartphone only on the hip. For practical purposes, the smartphone is likely to be worn on several locations or used to perform tasks other than measuring physical activity. Thus, we also compared models with conditions C3 and C4. C3 are the conditions utilized to compare algorithms A1 to A4. On the other hand, in C4 only data from the hip were used to fit the parameters of the model. In

Table 2
Performance of different algorithms. The values are given as mean and standard deviation.

Performance parameter	Algorithm			
	A1	A2	A3	A4
RMSE (cpm)	4.16 ± 3.55	2.54 ± 1.73	3.40 ± 2.52	2.64 ± 1.48
Error rate classification (four levels)	0.28 ± 0.36	0.18 ± 0.25	0.18 ± 0.25	0.17 ± 0.25
Error rate classification (two levels)	0.077 ± 0.225	0.030 ± 0.105	0.029 ± 0.084	0.029 ± 0.105

both cases, data from hip and pocket were used for testing, which is close to a real situation. The comparison between the performance of C3 and C4 can give a clue about whether the model requires data from several locations in the fitting procedure or just one location is enough.

The performance of the model was always estimated with a cross-validation strategy, as explained in the next section (2.5).

2.5. Performance evaluation

The performance estimation was based on a cross-validation strategy by person-activity, as in [14]. In each run of the cross-validation, data from an activity of a given person was left out (validation set). Then, the algorithms were optimized to predict Actigraph counts in the remaining data. For A2, A3 and A4 this optimization followed a standard fitting procedure, while for A1 it was the brute force search of filter parameters explained in section 2.3.1. Then, the models were used to predict counts in the person-activity that was left out. From them, three performance variables were obtained: the root mean square error (RMSE) of raw counts and the ratio of misclassified minutes using four or two activity levels, selected as follows. The four activity levels were taken using Actigraph cut-points as explained in Sasaki et al. [9] as light, moderate, vigorous and very vigorous. For many physical activity recommendations, the difference between the three highest levels is irrelevant [10,11]. Thus, we also considered the option of merging them into a single level.

The experimental data turned out not to be normal, so that non-parametric tests were selected. The influence of the algorithm was tested with a Quade test. Multiple comparisons were performed with Wilcoxon test with Bonferroni's correction. However, no algorithm stood out clearly as shown below. Thus, battery life was also measured as a secondary performance parameter. For that purpose, algorithms were also implemented in Android applications. Battery level was measured with a dedicated application which took samples of the battery charge every half hour while the selected algorithm was running continuously. This process lasted until the battery was fully discharged. It is well-known that keeping sensors leads to higher battery depletion.

Once an algorithm was selected, the correlation (spearman, ρ_s, and pearson, ρ_p) and agreement (Intra-Class Coefficient, ICC and Bland–Altmann plots) between calculated and Actigraph counts or between classification of activity levels (weighted kappa) were also obtained allowing the comparison with previous studies. A Wilcoxon test was used to compare the performance of different locations (C1 versus C2 and C3 versus C4).

3. Results

Table 2 shows RMSE and the error rate of classification in different levels for the four algorithms. Table 3 provides a summary of p-values. The results gave a significant effect of the algorithm in terms of both RMSE and error classification with two levels but there were not significant differences in error classification with four levels (although the p-value is close to the 0.05 limit). The pairwise comparisons for RMSE is shown in Table 4. According to it, two groups can be discriminated: (A1, A3) and (A2, A4). Algorithms A1 and A3 gave the highest RMSE in Table 2. The pairwise

Table 3
Summary of p-values obtained from Quade tests.

	RMSE	Error classification (4 levels)	Error classification (2 levels)
Algorithm	0.000	0.065	0.002

Table 4
Paired comparison with Bonferroni's correction for RMSE.

	A1	A2	A3	A4
A1	—	0.000	0.150	0.000
A2	0.000	—	0.000	0.140
A3	0.150	0.000	—	0.000
A4	0.000	0.140	0.000	—

Table 5
Paired comparison with Bonferroni's correction for error classification (two levels).

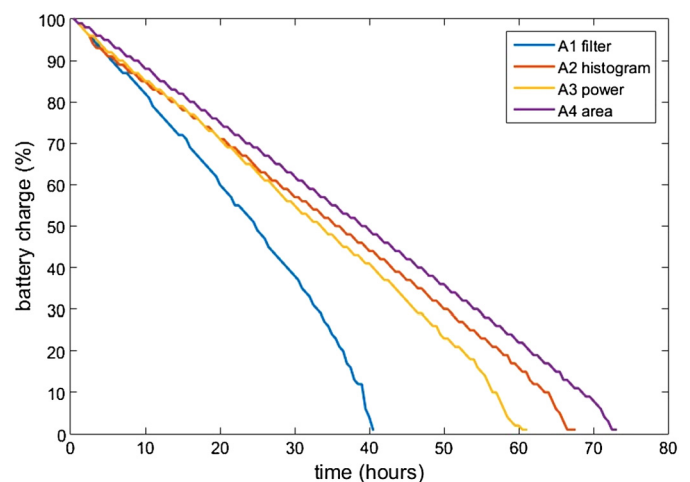|	A1	A2	A3	A4
A1	—	0.070	0.031	0.078
A2	0.070	—	1.000	1.000
A3	0.031	1.000	—	1.000
A4	0.078	1.000	1.000	—

Fig. 3. Battery level as a function of time for the four algorithms considered. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

comparisons for error classification with two levels are presented in Table 5: A1 differs significantly from A3. A1 is again the worse one in error classification (Table 2).

As an overall view of performance parameters related to counts, we decided to rule out algorithm A1. The choice between the remaining ones is not clear, although A3 is significantly worse in terms of RMSE. Turning to battery lives, they were found to be 73, 67.5, 61 and 40.5 hours for algorithms A4, A2, A3 and A1 respectively (Fig. 3). Thus, A1 is clearly the worse. The differences between the other ones, although smaller, are significant. For instance, the difference between battery levels while running A4 and A2 as a function of time showed a high correlation ($\rho_p = 0.838$, p-value < 0.001). Thus, the algorithm based on a simple and almost
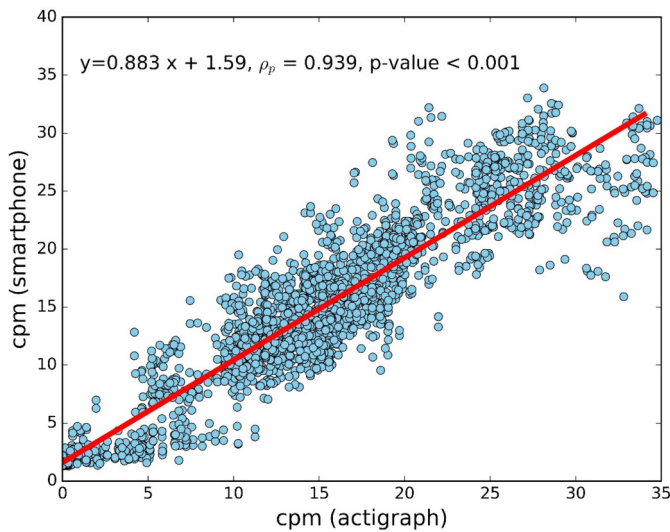
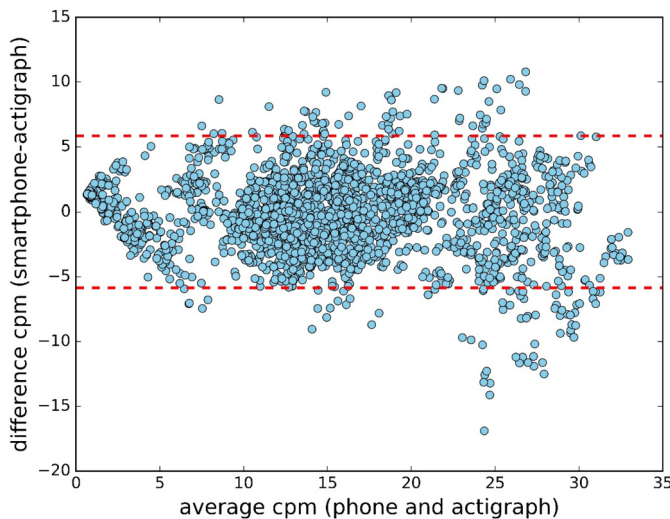**Fig. 4.** Smartphone counts vs. GT3X+ Actigraph counts.

Fig. 5. Bland–Altmann plot of actigraph and smartphone counts.

direct calculation of area under the curve is noticeably the best. Therefore, there is no reason to select A2 or A3, and we performed a more detailed analysis of A4. It should be highlighted that in real life the battery life would be lower since the smartphone would run also many other applications.

Fig. 4 shows a fit of the overall data, in which the counts calculated with the smartphone (algorithm A4) are plotted versus the counts obtained by the GT3X+. The raw count values showed an agreement of ICC = 0.937 (p-value < 0.001). The correlation was strong (ρ_p = 0.939, p-value < 0.001; ρ_s = 0.927, p-value < 0.001). With respect to classification with four PA levels, the agreement according to weighted kappa values was 0.874 (p-value < 0.001). The corresponding kappa value with two levels was 0.923 (p-value < 0.01).

Fig. 5 shows the Bland–Altmann plot comparing the Actigraph counts and the counts obtained by algorithm A4. The limits of the difference are ±5.84. Recovering original Actigraph counts by taking into account its accelerometer resolutions, this accounts for ±5.84/0.002929 ≈ ±2000 counts. Although this value is large, the influence on the classification of PA level is not so strong as shown in Table 2.

Table 6 shows the confusion matrix for algorithm A4 and four levels of PA. Light, moderate and vigorous levels are accurately

Table 6
Confusion Matrix of A4.

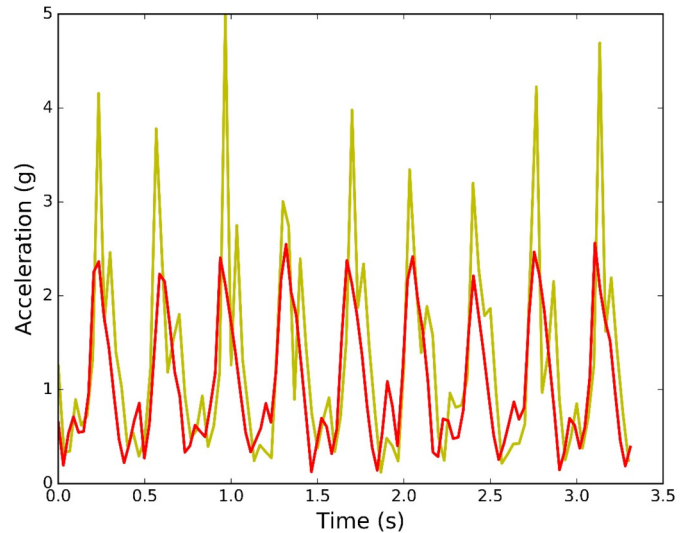| Smartphone\GT3X+	Light	Moderate	Vigorous	Very vigorous
Light	92.1%	7.9%	0%	0%
Moderate	1.8%	89.4%	8.8%	0%
Vigorous	0%	15.3%	75.7%	8.9%
Very Vigorous	0%	2.3%	68.5%	29.1%

Fig. 6. Raw value of the acceleration magnitude for a snapshot of "running". Yellow signal: GT3X+; red signal: smartphone. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 7
Performance under location conditions C1 and C2. The values are given as mean and standard deviation.

	RMSE	Error rate classification (four levels)	Error rate classification (two levels)
C1	2.29 ± 1.30	0.16 ± 0.26	0.033 ± 0.114
C2	2.64 ± 1.98	0.18 ± 0.28	0.030 ± 0.124
p-value	0.156	0.378	0.284

Table 8
Performance of location conditions C3 and C4. The values are given as mean and standard deviation.

	RMSE	Error rate classification (four levels)	Error rate classification (two levels)
C3	2.64 ± 1.48	0.17 ± 0.25	0.029 ± 0.105
C4	2.72 ± 1.42	0.19 ± 0.25	0.035 ± 0.127
p-value	0.055	0.003	0.172

classified (success higher than 70%), whereas the very vigorous level is not well discriminated. The same trend was observed in algorithms A3 and A2.

The differences between smartphone and GT3x+ for intense physical activity also stands out in raw values as Fig. 6 shows for an example of the "running" activity. The smaller range of smartphone accelerometer implies that it does not reach the values of GT3x+. This has also an influence in the area under the curve and, thus, in the activity counts.

With respect to the study of the phone location, the comparison between conditions C1 and C2 (using only the hip and the pocket respectively) is shown in Table 7. With the phone on the hip, the performance is better for RMSE, better also for the error rate with four levels although less clearly, and almost the same for error rate with two levels. Nonetheless in any of the comparisons the difference is statistically significant.

The comparison between conditions C3 and C4 is shown in Table 8. This comparison is more realistic as explained above, since in both cases the system is tested with signals in both locations and the model parameters are obtained using both locations (C3) or only the hip (C4). The performance under conditions C3 is better in all cases, but significant only for error rate with four levels and very close to the limit in RMSE.

4. Discussion and conclusions

In this paper, four different algorithms were proposed to calculate activity counts. Despite being more refined and being based on discriminating frequencies, A1 and A3 are not the best. More simple algorithms such as those based on area under the curve (A4) or histograms (A2) of acceleration values give lower values of RMSE. Given that A4 is the simplest to implement and the one with lower battery consumption, the direct area-based algorithm (A4) would be the choice following our results. The importance of battery consumption could not be undermined since this is a relevant parameter for end users.

Algorithm A4 agrees with the recommendation found in [25] in which the mean amplitude deviation (MAD) was proposed to get similar results irrespective of accelerometer brand. The MAD is in fact proportional to an area of the rectified curve after subtracting the mean, although in [25] a different epoch was used and the study targeted adolescents instead of adults.

When considering "running" or very vigorous activities algorithm A4 has a bad classification rate into four PA levels. In the work of Van Hees et al. [4] a similar result was found, in which the best levels classified correctly were light and moderate in terms of counts too. One of the possible causes could be the range of the smartphone's accelerometer, as it is limited to ± 2 g in the smartphone we have used, while it is ± 6 g in the GT3X+. Thus if some PA recommendation relied on achieving very vigorous activities, this point should be addressed, for instance using smartphones with a wider range. Another option would be to find another variable that could inform about intense activities. On the other hand, in [14] light activities were not properly classified. Perhaps the difference in the results of the classification are the cut-points used in each work, while in [14] the cut points are from Freedson et al. [30], we used those of Sasaki et al. [9].

When considering the difference between light and MVPA, error rates are below 10%, suggesting that smartphones are suitable to monitor the fulfillment of typical PA recommendations in terms of MVPA. This has several implications. For instance, they allow monitoring continuously the levels of PA with a very popular device. In addition, the development of smartphone applications can include feedback elements, attractive user interfaces and other motivational elements. This is especially true in some groups of people, like young people, who are used to utilize smartphones massively.

The correlation between raw counts in A4 and Actigraph GT3x+ is strong ($\rho_s = 0.939$) and similar to that found in [14] (ρ_s from 0.77 to 0.82 depending on smartphone model). When considering the values of agreement (ICC or weighted kappa), the values we have found are conventionally considered as acceptable or substantial [31]. Some results suggest that differences between different brands can be lower than differences between Actigraph GT3x+ and smartphones. Vanderloo et al. [32] compared the rate of MVPA in free living conditions, considering minutes per hour of wear time to account for participant's varied adherence. They found significant differences between Actigraph and Actical methods (p-value < 0.00, Cohen's $d = 2.93$). A somehow similar calculation has been performed with our data taking the number of minutes in MVPA in each of the person-activity bouts used for cross validation. In this case, Actigraph and smartphone minutes in MVPA are

not significantly different and, in any case, with a lower size effect (p-value $= 0.19$, Cohen's $d = 0.11$).

With respect to the phone location, our results indicate that this is not a critical parameter. As expected, using the phone on the hip would allow to better predict the counts of the Actigraph in terms of RMSE, since this device is also worn on the hip, but the difference is not even statistically significant. Moreover, the difference is almost zero when classifying PA with two levels. In real life, it is expected that the phone would be worn in different locations and subjected to different movements. In this sense, it would be convenient to use data from all the locations to find model parameters since this seems to improve error classification into four levels (Table 8). However, differences in error rate are only about 2% in average, which appears to be low from a practical point of view. The results that we have obtained are coherent with those found in [14], in which it is briefly stated that the location of the smartphone has no influence on the counts; however, in [14] no specific analysis is devoted to this aspect.

One remarkable limitation of this study is that accelerometers generally do not identify the type of physical activity and cannot directly derive energy consumption or activity levels. They require a translation formula or thresholds to calculate the PA level or METS. In addition, accelerometers have been shown to underestimate EE at higher intensities due to a plateau around ten METS and to inaccurately assess EE during incline walking [5]. Besides, it should be acknowledged that there is not still a hardware standard both for Actigraphs and smartphones. Different brands present differences in range and sensitivity, so that the results presented in this paper cannot be generalized and the parameters of the linear fit to Actigraph counts can depend on the accelerometer model.

To sum up, the present study indicates that physical activity levels as defined conventionally in the literature can be measured with smartphones. Future research will focus on determining the influence of the smartphone application on the adherence of participants in interventions to promote physical activity. The recommendations found in [18] will be taken into consideration.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

Disclosure of interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Acknowledgements

References

[1] Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. Can Med Assoc J 2006;174(6). https://doi.org/10.1503/cmaj.051351.

[2] Andrew PH, Najat M, Nuala MB. Assessment of physical activity and energy expenditure: an overview of objective measures. Front Nutr (Nutrition Methodology) 2014;1(5):1–16.

[3] Lowe SA, ÓLaighin G. Monitoring human health behaviour in one's living environment: a technological review. Med Eng Phys 2013;36:147–69.

[4] Van Hees VT, Pias M, Taherian S, Ekelund U, Brage S. A method to compare new and traditional accelerometer data in physical activity monitoring. In: World of Wireless Mobile and Multimedia Networks (WoWMoM); 2010.

[5] Ugarriza R, Ayuso J, Benito PJ, Chamizo R, Ara I, González-Gross M. Physical activity assessment in the general population; instrumental methods and new technologies. Nutr Hosp 2015:219–26. https://doi.org/10.3305/nh.2015.31.sup3.8769.

[6] An H-S, Kim Y, Lee J-M. Accuracy of inclinometer functions of the activPAL and ActiGraph GT3X+: a focus on physical activity. Gait Posture 2016;51:174–80.

[7] Murphy SL. Review of physical activity measurement using accelerometers in older adults: considerations for research design and conduct. Prev Med 2008:108–14. https://doi.org/10.1016/j.ypmed.2008.12.001.

[8] Rodriguez S, Corella C, Abarca A, Zaragoza J. Validation of three short physical activity questionnaires with accelerometers among university students in Spain. J Sports Med Phys Fit 2017;57(12):1660–8. https://doi.org/10.23736/S0022-4707.17.06665-8.

[9] Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sports 2011;14(5):411–6.

[10] Oja P, Bull FC, Fogelholm M, Martin BW. Physical activity recommendations for health: what should Europe do?. BMC Public Health 2010;10(1):1–10. https://doi.org/10.1186/1471-2458-10-10.

[11] Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007;39(8):1423–34.

[12] Gowda A, Babu S, Sekaran D. UMOISP: usage mode and orientation invariant smartphone pedometer. IEEE Sens J 2017;17(3):869–81. https://doi.org/10.1109/JSEN.2016.2635691.

[13] Duclos M, Fleury G, Lacomme P, Phan R, Ren L, Rousset S. An acceleration vector variance based method for energy expenditure estimation in real-life environment with a smartphone/smartwatch integration. Expert Syst Appl 2016;63:435–49.

[14] Hekler EH, Buman PM, Grieco L, Rosenberger M, Winter JS, Haskell W, et al. Validation of physical activity tracking via Android smartphones compared to actigraph accelerometer: laboratory-based and free-living validation studies. JMIR mHealth and uHealth 2015;3(2):36.

[15] Donaire-Gonzalez D, De Nazelle A, Seto E, Mendez M, Nieuwenhuijsen MJ, Jerrett M. Comparison of physical activity measures using mobile phone-based CalFit and Actigraph. J Med Internet Res 2013;15(6):111.

[16] Lee JM, Kim Y, Kwon YS, Derrick TR, Welk GJ. Calibration of built-in accelerometer using a commercially available smartphone. American College of Sports Medicine; 2014.

[17] Wang A, Chen G, Yang J, Zhao S, Chang C. A comparative study on human activity recognition using inertial sensors in a smartphone. IEEE Sens J 2016;16(11):4566–77. https://doi.org/10.1109/JSEN.2016.2545708.

[18] Shin D, Biocca F. Health experience model of personal informatics: the case of a quantified self. Comput Hum Behav 2017;69:62–74. https://doi.org/10.1016/j.chb.2016.12.019.

[19] Maltseva K, Lutz C. A quantum of self: a study of self-quantification and self-disclosure. Comput Hum Behav 2018;81:102–14. https://doi.org/10.1016/j.chb.2017.12.006.

[20] Bert F, Giacometti M, Gualano M, Siliquini R. Smartphones and health promotion: a review of the evidence. New York: Springer Science & Business Media; 2013.

[21] What are Counts? Available from: https://actigraph.desk.com/customer/en/portal/articles/2515580-what-are-counts (accessed on 5 December 2017).

[22] ActiLife manual. Available from: http://actigraphcorp.com/support/manuals/actilife-6-manual/ (accessed on 5 December 2017).

[23] Brond JC, Arvidsson D. Sampling frequency affects the processing of Actigraph raw acceleration data to activity counts. J Appl Physiol 2015:362–9. https://doi.org/10.1152/japplphysiol.00628.2015.

[24] Peach D, Van Hoomissen J, Callender HL. Exploring the ActiLife® filtration algorithm: converting raw acceleration data to counts. Physiol Meas 2014;35:2359–67.

[25] Aittasalo M, Vähä-Ypyä H, Vasankari T, Husu P, Jussila AM, Sievänen H. Mean amplitude deviation calculated from raw acceleration data: a novel method for classifying the intensity of adolescents' physical activity irrespective of accelerometer brand. BMC Sports Sci Med Rehabil 2015. https://doi.org/10.1186/s13102-015-0010-0.

[26] Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. PLoS ONE 2016;11(8):e0160644. https://doi.org/10.1371/journal.pone.0160644.

[27] Rodriguez VH, Medrano C, Plaza I, Corella C, Abarca A, Julian JA. Estimating the physical activity with smartphones: analysis of the device position and comparison with GT3X+ actigraph. In: Advances in intelligent systems and computing (AISC), vol. 476. 2016. p. 49–56.

[28] Chen KY, Sun M. Improving energy expenditure estimation by using a triaxial accelerometer. J Appl Physiol 1997;83(6):2112–22.

[29] Weiss GM, Lockhart JW. The impact of personalization on smartphone-based activity recognition. AAAI technical report WS-12-05, workshops at the twenty-sixth AAAI conference on artificial intelligence, 2012.

[30] Freedson PS, Melanson E, Sirard J. Calibration of the computer science and applications. Med Sci Sports Exerc 1998;30(5):777–81.

[31] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge University Press. ISBN 9780521133852, 2011.

[32] Vanderloo LM, Di Cristofaro NA, Proudfoot NA, Tucker P, Timmons BW. Comparing the actical and ActiGraph approach to measuring young children's physical activity levels and sedentary time. Pediatr Exerc Sci 2016;28(1):133–42. https://doi.org/10.1123/pes.2014-0218.