Study on the Generalization Ability of Accelerometer Threshold-based Methods for Noise Detection in PPG Signals

S Mula [1], R Zangróniz [1], JJ Rieta [2], R Alcaraz [1]

[1] Research Group in Electronic, Biomedical and Telecommunication Engineering, University of Castilla-La Mancha, Cuenca, Spain
[2] Universitat Politecnica de Valencia, Valencia, Spain

Abstract

Continuous heart rate monitoring through wearable devices incorporating photoplethysmogram (PPG) sensors commonly provides very noisy signals, especially during daily activities and physical exercise. Since motion artifacts are one of the main sources of this noise, it is a common approach to use an accelerometer to detect movements and discard periods when the acceleration exceeds a certain threshold. To quantitatively assess the performance of these methods on different datasets, all recordings from the public datasets PPG DaLiA and WESAD were segmented into 5 second-length intervals, then labeled as clean or noisy on the difference in the inter-beat interval between the available synchronous ECG and PPG signals, and finally compared with the results of several acceleration thresholds.

The results show that, while accuracies of 77.3% can be achieved, those results fail to generalize across datasets. While it is already known that acceleration threshold-based methods show poor performance, even methods as simple as these might seem accurate in a particular dataset, while being useless in others. This might happen to other noise detection methods and serves as a remainder that an external validation with varied datasets necessary for a rigorous evaluation of any noise detection method.

1. Introduction

Photoplethysmography (PPG) is a technique in which light is used to measure variations in the size of a tissue, and can be used to measure heart rate when used to measure changes in blood volume on the skin. This method is non-invasive and can be embedded in wearable devices, making it a promising choice for continuous heart rate monitoring.

PPG signals are very noisy, especially during daily activities and physical exercise. Motion artifacts are one of the main sources of this noise, and share the same frequency band with the PPG signal. In such cases, motion artifacts cannot be removed by filtering without discarding the desired signal as well [1]. One of the many approaches to deal with this kind of noise is the use of additional sensors, such as an accelerometer, to discard periods where movement exceeds a certain threshold [1]. However, little research can be found in the literature about how extrapolable the accelerometer threshold-based methods are to successfully work with different datasets. It had already been suggested in 2006 that statistical approaches on the raw PPG data could outperform accelerometer or external sensor based methods [2]. Some of the works which study noise detection assert that the accelerometer signal does not contain enough information to detect all motion artifacts, such as [3]. Casson shows that gyroscopes perform better than accelerometers in half the cases and suggests both signals should be studied [4].

2. Methods

In order to quantify the performance of any noise detection method, a way to measure signal quality is necessary. There are several ways to achieve that, from manually annotating datasets to the use of reference signals assumed to be accurate. With the intention of objectively evaluating signal quality, the ECG signal has been used as reference source of heart rate. This approach is deemed to be more objective than manual annotation. Public datasets with simultaneous ECG and PPG have been used to evaluate the performance of simple accelerometer threshold-based methods.

2.1. Public datasets with simultaneous PPG and ECG in daily activities

While there are many public datasets containing PPG and ECG signals, most contain recordings of ICU patients and sleeping subjects. It is relatively uncommon to find datasets with simultaneous ECG and PPG in daily activities and physical exercise. This work has focused on those

ISSN: 2325-887X DOI: 10.22489/CinC.2023.410

Name	Subjects	References
MAXREFDES100	7	[5]
PPG-DaLiA	15	[6]
WESAD	15	[7]
WPPG	8	[8][9]
CIME-PPG	10	[10] [11]
IEEE SPC 2015	12	[12]
LTAF	8	[13]

Table 1. Main public datasets with simultaneous ECG and PPG in daily activities

activities, since they would be encountered by any noise detection method applied in continuous heart rate monitoring.

Hence, the datasets to evaluate must contain long term recordings, if possible continuous to be able to control the process of segmenting the recordings. The candidate datasets are listed on table 1. Of these, PPG-DaLiA, WESAD, WPPG and LTAF are not segmented and recordings have subject identifier. WPPG has short-term recordings and was not used. In LTAF, while containing days-long recordings, the ECG signal is not continuos and the dataset was not used. This work has been done with the PPG-DaLiA and WESAD datasets, with a total of 30 subjects. The PPG and accelerometer signal of both have been recorded with the Empatica E4 wristband.

2.2. Signal quality evaluation

The first phase of the signal quality evaluation is to measure the reference heart rate, in this case from an ECG signal. The ECG signal might contain noise as well. To avoid the effects of the ECG noise, an ECG noise detection algorithm was used to exclude the noisy segments [14]. Then, the Pan-Tompkins algorithm was used to locate the ECG R peaks, from which the heart rate was calculated as 60 divided by the time difference between peaks in seconds, obtaining the heart rate in beats per minute, or bpm.

The signals were divided in 5 second segments, which were classified as noisy if the maximum error exceeds a given threshold, and clean otherwise. Two thresholds have been used: 3 bpm and 8 bpm.

2.3. Noise detection

For noise detection, two acceleration metrics are compared: the acceleration vector module of the raw signal of each axis, and the acceleration vector module of the band-pass filtered signals of each axis on the PPG frequency range. The band-pass filtering was performed with an order 3 Butterworth filter with cutoff frequencies of 0.5Hz and 12.0 Hz. Then, the signal was evaluated for each acceleration threshold in the range between 0 g and 1.5 g in 75 steps of 0.02 g.

The signals were divided in the same 5 second segments as the quality evaluation phase, which were classified as noisy if the average acceleration exceeds the threshold, and clean otherwise. Then, the accelerometer threshold labels are compared with the quality evaluation labels, obtaining metrics like accuracy, false negative ratio or false positive ratio. It is considered a true positive, or TP, when both the accelerometer label and the reference quality label state that the signal is clean, and a false positive, or FP, when the accelerometer label is clean buy the reference label is noisy. False negatives, or FN, represent the opposite case, when the accelerometer labels state that the signal is noisy while it is not. Accuracy is the amount of correct labels divided by tha total amount of labels, and can be expressed as $(TP + TN)/(TP + TN + FP + FN)$.

3. Results

As the maximum error threshold grows the proportion of valid signal increases, but this process is not linear. In figure 1 that difference in the valid ratio shape is visible.

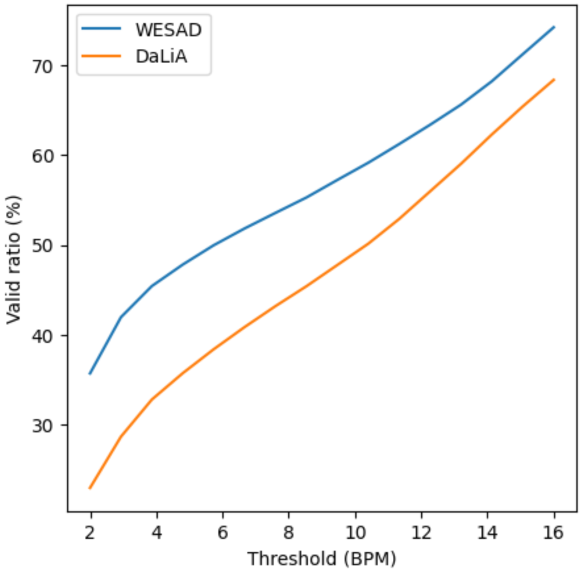

Figure 1. Ratio of valid segments by maximum error threshold

In all cases, the false negative ratio, which is the proportion of clean segments labeled as noisy, decreases as the acceleration threshold grows. The opposite happens with the false positive ratio, which increases with the threshold. This is expected, as the larger the threshold the more segments are labeled as clean. The accuracy peak is near the point where the false negative ratio and the false positive ratio are equal. The main differences between datasets are the location of this point and the maximum accuracy. This differences are visible in figures 2, 3, 4 and 5, with the highest accuracy in figure 2, with an accuracy of 77.3% at 0.02 g with a reference error threshold of 3 bpm in the PPG_DaLiA dataset. With that same configuration, WESAD only reaches a 71.4%, and at 0.28 g.

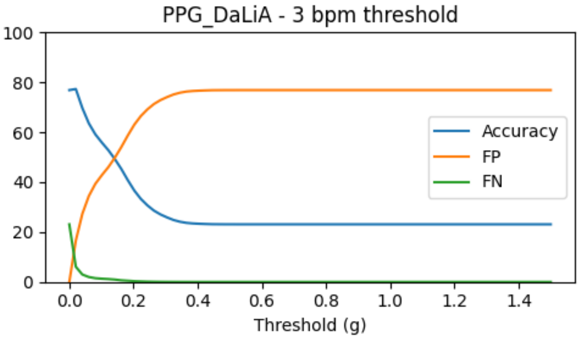

Figure 2. Accuracy, false positives and false negatives for the filtered acceleration threshold method in PPG_DaLiA by acceleration threshold. The error threshold in signal quality evaluation is 3 bpm.

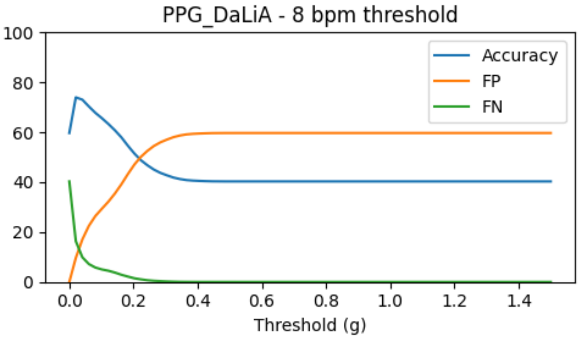

Figure 3. Accuracy, false positives and false negatives for the filtered acceleration threshold method in PPG_DaLiA by acceleration threshold. The error threshold in signal quality evaluation is 8 bpm.

4. Conclusion

This study is limited by the nature of the datasets employed, particularly by the fact that the same PPG sensor was used in both datasets. This makes these datasets more similar, something that favours generalization. The possible sources of the differences between the datasets are the differences in the study population, the use of a different ECG sensor and the variation of the activities and environment.

Even with two datasets recorded with the same wristband device, the results show that there are significant differences in signal quality. The accelerometer threshold-based noise detection approach proved to perform at its peak at a different threshold for each dataset, showing that the generalization ability of threshold-based methods is almost non existent for this type of PPG signal. These methods should be expected to generalize worse with very different datasets, such as datasets recorded with transmissive PPG, common in pulse-oximeters, instead of the reflective PPG found on wristbands.

That these accelerometer threshold-based methods have a poor performance and do not generalize is nothing new, but looking only at one of the datasets, it would seem as a method with over 75% accuracy, while that same threshold is useless in a distinct but similar dataset, and would almost certainly be so in the real world. This effect, in this case observed with these extremely simple methods, might happen on more sophisticated methods as well. In order to assess the performance of a PPG noise detection method, an external validation with varied datasets is necessary to

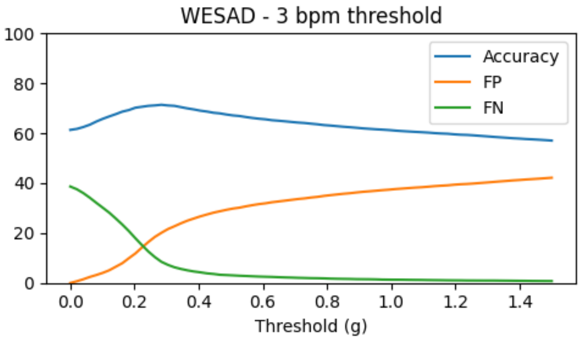

Figure 4. Accuracy, false positives and false negatives for the filtered acceleration threshold method in WESAD by acceleration threshold. The error threshold in signal quality evaluation is 3 bpm.

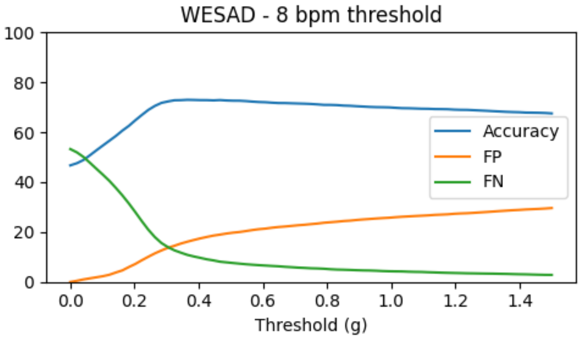

Figure 5. Accuracy, false positives and false negatives for the filtered acceleration threshold method in WESAD by acceleration threshold. The error threshold in signal quality evaluation is 8 bpm.

avoid a possible misleading appearance of higher performance.

In summary, these results show that these simple methods can be moderately effective on a single dataset, if used with the appropriate parameters, but fail to generalize even on similar datasets. The performance of acceleration threshold-based methods is too limited for critical uses such as real-world medical applications.

Acknowledgments

This research has received financial support from Daiichi Sankyo SLU and from public grants PID2021-00X128525-IV0, PID2021-123804OB-I00, and TED2021-130935B-I00 of the Spanish Government 10.13039/501100011033 jointly with the European Regional Development Fund (EU), and SBPLY/21/180501/000186, and SBPLY/21/180501/000122 from Junta de Comunidades de Castilla-La Mancha, and AICO/2021/286 from Generalitat Valenciana.

References

[1] Fine J, Branan KL, Rodriguez AJ, Boonya-ananta T, Ajmal, Ramella-Roman JC, McShane MJ, Coté GL. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. Biosensors 2021;11(4). ISSN 2079-6374.

[2] Foo JYA, Wilson SJ. A computational system to optimise noise rejection in photoplethysmography signals during motion or poor perfusion states. Medical ampmathsemicolon Biological Engineering ampmathsemicolon Computing January 2006;44(1-2):140–145.

[3] Ahn J, Ra HK, Yoon HJ, Son SH, Ko J. On-device filter design for self-identifying inaccurate heart rate readings on wrist-worn ppg sensors. IEEE Access 2020;8:184774–184784.

[4] Casson AJ, Galvez AV, Jarchi D. Gyroscope vs. accelerometer measurements of motion from wrist PPG during physical exercise. ICT Express December 2016;2(4):175–179.

[5] Biagetti G, Crippa P, Falaschetti L, Saraceni L, Tiranti A, Turchetti C. Dataset from ppg wireless sensor for activity monitoring. Data in Brief 2020;29:105044. ISSN 2352-3409.

[6] Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. Sensors 2019;19(14). ISSN 1424-8220.

[7] Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356923, 2018; 400–408.

[8] Jarchi D, Casson A. Description of a database containing wrist ppg signals recorded during physical exercise with both accelerometer and gyroscope measures of motion. Data Dec 2016;2(1):1. ISSN 2306-5729.

[9] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet. Circulation 2000;101(23):e215–e220.

[10] Xu K, Jiang X, Ren H, Liu X, Chen W. Deep recurrent neural network for extracting pulse rate variability from photoplethysmography during strenuous physical exercise. In 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS). 2019; 1–4.

[11] Xu K, Jiang X, Chen W. Photoplethysmography motion artifacts removal based on signal-noise interaction modeling utilizing envelope filtering and time-delay neural network. IEEE Sensors Journal 2020;20(7):3732–3744.

[12] Zhang Z, Pi Z, Liu B. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Transactions on Biomedical Engineering 2015;62(2):522–531.

[13] Bacevičius J, Abramikas Z, Badaras I, Butkuvienė M, Daukantas S, Dvinelis E, Gudauskas M, Jukna E, Kiseliūtė M, Kundelis R, Marinskienė J, Paliakaitė B, Petrėnas A, Petrylaitė M, Pilkienė A, Pluščiauskaitė V, Rapalis A, Sokas D, Sološenko A, Staigytė J, Stankevičiūtė G, Taparauskaitė N, Aidietis A, Marozas V. Long-term electrocardiogram and wrist-based photoplethysmogram recordings with annotated atrial fibrillation episodes, 02 2022.

[14] Huerta A, Martinez Rodrigo A, Gonzalez V, Quesada A, Rieta J, Alcaraz R. Quality assessment of very long-term ecg recordings using a convolutional neural network. In 2019 E-Health and Bioengineering Conference (EHB). 11 2019; 1–4.

Address for correspondence:

Santiago Mula Muñoz
Campus Universitario S/N. 16071 – Cuenca (España)
santiago.mulamunoz@uclm.es