Physiological Measurement

PAPER

Comparison between electrocardiogram- and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
14 February 2018

REVISED
9 July 2018

ACCEPTED FOR PUBLICATION
11 July 2018

PUBLISHED
8 August 2018

PAPER

Comparison between electrocardiogram- and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions

Linda M Eerikäinen[1,2], Alberto G Bonomi[2], Fons Schipper[2], Lukas R C Dekker[1,3], Rik Vullings[1], Helma M de Morree[2] and Ronald M Aarts[1,2]

[1] Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ, Eindhoven, Netherlands
[2] Philips Research, Eindhoven, Netherlands
[3] Department of Cardiology, Catharina Hospital, Eindhoven, Netherlands

E-mail: L.M.Eerikainen@tue.nl

Keywords: atrial fibrillation, electrocardiography, photoplethysmography, movement artefacts

Abstract

Objective: Atrial fibrillation (AF) is the most commonly experienced arrhythmia and it increases the risk of stroke and heart failure. The challenge in detecting the presence of AF is the occasional and asymptomatic manifestation of the condition. Long-term monitoring can increase the sensitivity of detecting intermittent AF episodes, however it is either cumbersome or invasive and costly with electrocardiography (ECG). Photoplethysmography (PPG) is an unobtrusive measuring modality enabling heart rate monitoring, and promising results have been presented in detecting AF. However, there is still limited knowledge about the applicability of the PPG solutions in free-living conditions. The aim of this study was to compare the inter-beat interval derived features for AF detection between ECG and wrist-worn PPG in daily life. *Approach*: The data consisted of 24 h ECG, PPG, and accelerometer measurements from 27 patients (eight AF, 19 non-AF). In total, seven features (Shannon entropy, root mean square of successive differences (RMSSD), normalized RMSSD, pNN40, pNN70, sample entropy, and coefficient of sample entropy (CosEn)) were compared. Body movement was measured with the accelerometer and used with three different thresholds to exclude PPG segments affected by movement. *Main results*: CosEn resulted as the best performing feature from ECG with Cohens kappa 0.95. When the strictest movement threshold was applied, the same performance was obtained with PPG (kappa = 0.96). In addition, pNN40 and pNN70 reached similar results with the same threshold (kappa = 0.95 and 0.94), but were more robust with respect to movement artefacts. The coverage of PPG was 24.0%–57.6% depending on the movement threshold compared to 92.1% of ECG. *Significance*: The inter-beat interval features derived from PPG are equivalent to the ones from ECG for AF detection. Movement artefacts substantially worsen PPG-based AF monitoring in free-living conditions, therefore monitoring coverage needs to be carefully selected. Wrist-worn PPG still provides a promising technology for long-term AF monitoring.

1. Introduction

Atrial fibrillation (AF) is the most commonly experienced arrhythmia affecting 1%–2% of the general population and its prevalence is expected to increase in the coming years (Camm *et al* 2010). AF increases the risk of stroke, heart failure, hospitalization, and death (Camm *et al* 2010).

AF is a progressive disease that starts with occasional events, called paroxysmal AF, and slowly progresses to persistent and permanent AF (Camm *et al* 2010). The challenge in the early detection of AF is the occasional nature of the events, but also that AF can be asymptomatic. In a group of patients with paroxysmal AF, the episodes were more often asymptomatic than accompanied by symptoms (Page *et al* 1994).

The standard practice for diagnosing AF is with electrocardiography (ECG). However, ECG has its limitations. The sensitivities of 12-lead ECGs and transtelephonic ECGs are between 30%–40% whereas for 24/48 h

Holters it is 44%–60% (Rosero *et al* 2013). The added value of prolonged continuous monitoring for diagnosing AF has been shown when monitoring survivors from cryptogenic stroke either continuously with an insertable cardiac monitor (ICM) compared to Holter screening (Sanna *et al* 2014, Brachmann *et al* 2016). In a three year period, eight times more patients were diagnosed with AF with an ICM compared to the Holter control group. Implantable devices are costly and require surgical procedures whereas Holter monitors can cause irritation from the electrodes and are cumbersome to wear, and thus are not suitable for long-term monitoring. Therefore, there is a demand for more convenient long-term monitoring solutions for diagnosing AF.

Photoplethysmography (PPG) is an unobtrusive measurement modality that enables the measurement of different physiological parameters, such as heart rate (Allen 2007, Valenti and Westerterp 2013, van Andel *et al* 2015). Use of PPG for AF detection has been studied with different technologies, such as with smartphones (Lee *et al* 2012, McManus *et al* 2013, Chong *et al* 2015, Chan *et al* 2016, Schäck *et al* 2017), with finger probes in a clinical environment (Shan *et al* 2016, Tang *et al* 2017) and with wrist-worn devices (Bonomi *et al* 2016, Lemay *et al* 2016, Nemati *et al* 2016, Corino *et al* 2017, Pantelopoulos *et al* 2017, Shashikumar *et al* 2017).

Wrist-worn PPG devices are easy to use and comfortable to wear and therefore provide a promising solution for long-term monitoring. Although the wrist-worn PPG based AF detection algorithms showed promising classification performance, so far there is limited knowledge about their applicability to free-living conditions where the measurements are affected by various types of movement artefact. The majority of the studies have been conducted in fairly controlled conditions and with short measurements up to 10 min (Lemay *et al* 2016, Nemati *et al* 2016, Corino *et al* 2017, Shashikumar *et al* 2017) with only two exceptions. Pantelopoulos *et al* (2017) have presented results with overnight measurements and in our previous study we presented a Markov-model approach when using 24 h data (Bonomi *et al* 2016).

The aim of this study is to compare state-of-the-art inter-beat interval (IBI) derived features commonly used for AF detection from ECG and PPG in free-living conditions. Information about body movement is used to investigate the effect of movement artefacts to their discriminative power during daily living.

2. Methods

2.1. Data

The dataset for the analysis was collected in patients scheduled for 24 h Holter measurement. Patients were contacted by a cardiologist and given at least one week to consider participation in the study. The participants gave a written informed consent before the start of the measurements. The dataset was collected in the Catharina Hospital, Eindhoven, The Netherlands.

The data consisted of 24 h ECG measurement with a 12-lead Holter monitor (H12+, Mortara, Milwaukee, WI, USA), PPG, and three-axis accelerometer measurements from the non-dominant wrist with a data logging device equipped with the Philips Cardio and Motion Monitoring Module (CM3 Generation-3, Wearable Sensing Technologies, Philips, Eindhoven). The PPG sensor was based on reflective mode using two green light LEDs. The sampling frequency of both PPG and accelerometry was 128 Hz and the dynamic range of the accelerometer was ±8 g.

For synchronization purposes, at the start of the measurement the event button of the Holter monitor was pressed and the data logger tapped at the same time instant. During the recording period patients marked in a diary the daily activities, possible symptoms, and medication intake. At the end of the measurement, patients returned to the hospital and the same synchronization procedure was repeated. Recording devices were detached and the diary was handed in. Information about the daily activities in the diary was not used in this study. In addition, baseline characteristics, medical characteristics, and information about medication were collected.

The ECG data were visually analyzed by a clinical expert using an automated rhythm detection software (Veritas, Mortara, Milwaukee, WI, USA). The software extracted beat times from the ECG and identified every beat either to normal, supraventricular premature beat, ventricular premature beat, AF, paced, artefact, or unknown. The rhythm was then confirmed or corrected by the expert. The raw ECG data was not available for further research purposes, and therefore the beat times and beat labels were used in the data analysis.

In total 30 patients were recruited. Eight patients had continuous AF, 19 patients normal rhythm with premature beats, two patients atrial flutter, and one patient had a very noisy ECG reference. The patients with atrial flutter and very noisy ECG reference were excluded from the analysis. The patient characteristics of the remaining 27 patients are presented in table 1.

2.2. Preprocessing and data synchronization

The raw PPG data was downsampled from 128 Hz to 64 Hz and bandpass filtered to range from 0.3 to 5 Hz. The pulses were detected by identifying fiducial points in the PPG waveform, i.e. the troughs, by detecting local minima. For finding the local minima, the points where the first derivative goes from negative to positive were selected. To prevent detecting too many local minima, an adaptive threshold was used to exclude locally

Table 1. Patient characteristics.

Baseline characteristics	AF ($N = 8$)	Non-AF ($N = 19$)
Sex, male, *n* (%)	5 (62.5)	10 (52.6)
Age, years, M ± SD (range)	69 ± 11 (43–79)	67 ± 13 (34–87)
Height, cm, M ± SD (range)	166.5 ± 8.6 (152–179)	171.8 ± 8.7 (151–185)
Weight, kg, M ± SD (range)	86.5 ± 26.6 (71–149)	83.2 ± 20.3 (52–113)
BMI, kg m^{-2}, M ± SD (range)	30.9 ± 7.5 (24.6–48.1)	27.9 ± 5.5 (20.2–39.3)
Medical characteristics		
Structural heart disease		
Coronary artery disease, n (%)	1 (12.5)	3 (15.8)
Heart failure, *n* (%)	1 (12.5)	1 (5.3)
Heart valve disease, *n* (%)	0 (0)	1 (5.3)
Risk factor		
Hypertension, *n* (%)	4 (50.0)	4 (21.0)
Hyperlipidemia, *n* (%)	0 (0)	0 (0)
Diabetes mellitus, *n* (%)	2 (25.0)	0 (0)
Obstructive sleep apnea, *n* (%)	1 (12.5)	0 (0)
Medication		
Beta-blocker, *n* (%)	6 (75.0)	9 (47.4)
Calcium channel blocker, *n* (%)	4 (50.0)	3 (15.8)
Statin, *n* (%)	3 (37.5)	8 (42.1)
Anti-arrhythmic drug class I, *n* (%)	1 (12.5)	5 (26.3)
Anti-arrhythmic drug class III, *n* (%)	1 (12.5)	2 (10.5)
Digoxin, *n* (%)	2 (25.0)	1 (5.3)
Anticoagulation, *n* (%)	8 (100)	15 (79.0)

insignificant ones. The threshold was obtained by filtering the bandpass-filtered PPG signal with a first order lowpass filter (time constant 125 ms). The search for the minima was enabled only when the PPG signal was below the threshold. Additionally, maximum magnitude of acceleration, after removal of gravity, was assessed every second from the accelerometer. If a threshold of 0.1 g was exceeded, the local minima in that period were not considered.

The time between the PPG fiducial points was used to calculate the inter-pulse intervals (IPI). Similarly, the IBIs were calculated from the ECG as the time difference between the beat times given by the Holter software. The raw ECG signal was not available for further data analysis purposes.

The IBI and IPI series were recorded with different devices each having their own clock. The clocks may exhibit a time-offset and may run at different speeds, i.e. there might be a drift. For synchronizing the IBI and IPI series, first a set of short IPI subsequences was taken from the full IPI sequence. For each IPI subsequence from this set, the best matching position in the IBI sequence was determined. Each match gave a time-offset between the start time of the IPI subsequence and the start time of the best matching position in the IBI sequence. From these time-offsets the clock offset and drift were determined. In addition, the accelerometer signal was aligned based on the offset defined by the fit. In figure 1 example sequences of 30 s of PPG signal and corresponding IBIs and IPIs during sinus rhythm, AF, and movement are presented.

2.3. Features

In total seven IBI derived features for AF detection from the literature were compared in this study. The features are pNN40, pNN70, Shannon entropy (ShE), root mean square of successive differences (RMSSD), normalized RMSSD (nRMSSD), sample entropy (SampEn), and coefficient of sample entropy (CosEn). Prior to computing the features, outlier IBIs/IPIs were discarded by removing the ones <200 ms and >2200 ms.

In the study of Corino *et al* (2017) a wide range of features from PPG were analyzed for AF detection. The percentage of differences of successive IBIs that exceeded 40 ms or 70 ms (pNN40 and pNN70) were found to be the best discriminative feature combination.

Shannon entropy is an entropy estimate often used to determine regularity of an IBI sequence to distinguish AF (Lee *et al* 2012, McManus *et al* 2013, Chong *et al* 2015, Schäck *et al* 2017). The values in the sequence, in this case IBIs or IPIs, are divided into bins and the probability $p(i)$ of a value being in the bin i is

$$p(i) = \frac{n_{(i)}}{l - n_{outliers}}. \tag{1}$$

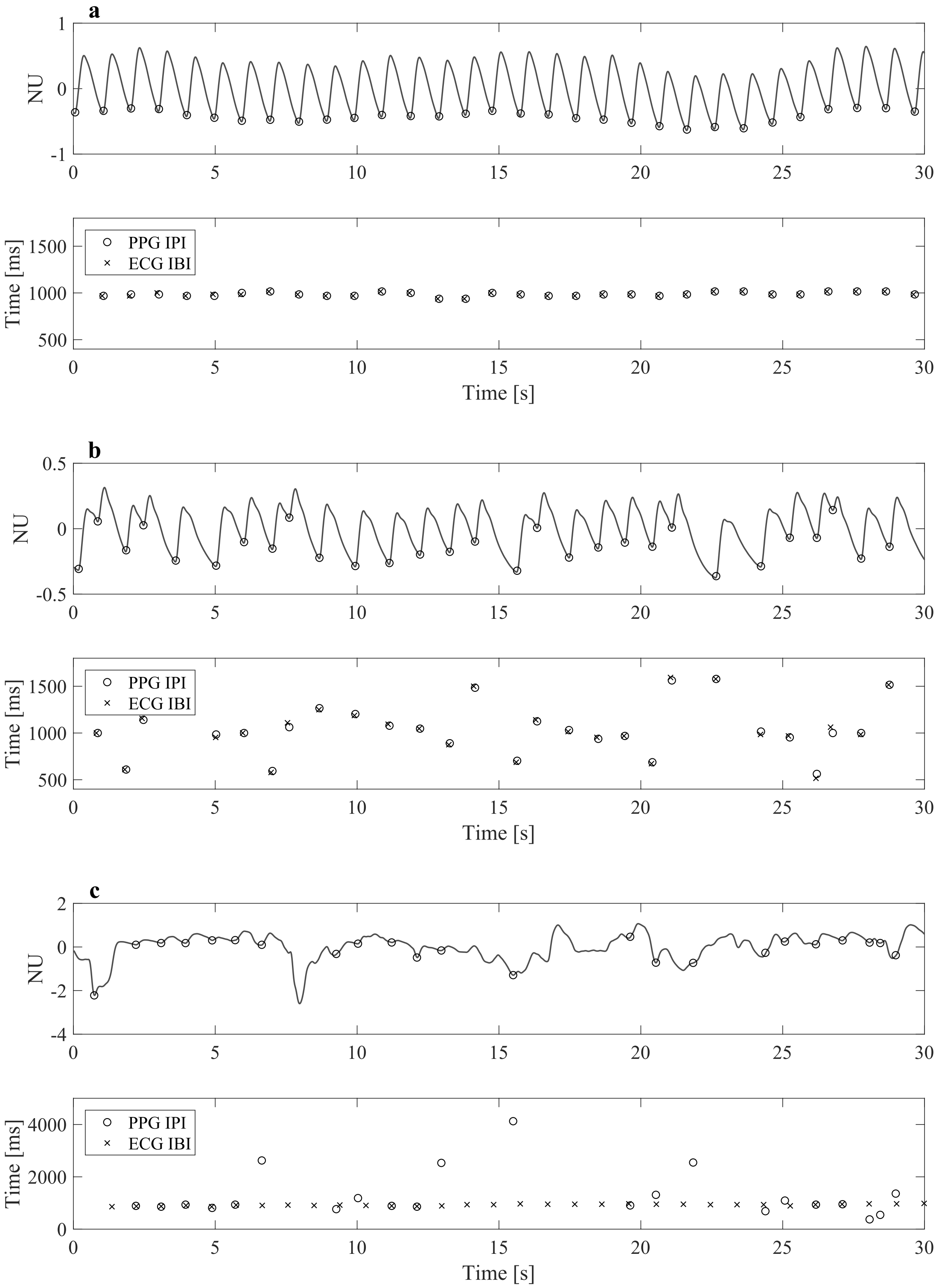

Figure 1. Examples of 30 s segments of PPG signals (above) and corresponding IBIs and IPIs (below) during sinus rhythm (a), AF (b), and movement (c).

$n(i)$ is the number of values in the bin i, l is the length of the sequence, and $n_{outliers}$ the number of outliers in the sequence. The bins were equally spaced in the range from 200 ms to 2200 ms. After having the probabilities for every bin, ShE can be calculated as follows:

$$\mathrm{ShE} = -\sum_{i=1}^{N} p(i)\frac{\log(p(i))}{\log(N)}, \tag{2}$$

where N is the number of bins. We used 16 bins as that has been shown to be the minimum number of bins to obtain reasonable accuracy (Dash *et al* 2009).

RMSSD and nRMSSD are features used to assess the beat-to-beat variability and have been studied for AF detection from PPG (Lee *et al* 2012, McManus *et al* 2013, Chong *et al* 2015, Corino *et al* 2017, Schäck *et al* 2017). The RMSSD of an IBI sequence of a length *l* is

$$\mathrm{RMSSD} = \sqrt{\frac{1}{l-1}\sum_{j=1}^{l-1}(IBI(j+1) - IBI(j))^2}. \tag{3}$$

The nRMSSD is the RMSSD divided by the mean IBI (or IPI) of the sequence.

SampEn assesses the similar patterns in a sequence, a lower value indicating more self-similarity. SampEn is the negative natural logarithm of the conditional probability that two sequences that match with each other at m points, i.e. the difference between the two sequences of length m is smaller than tolerance r, they also match when $m + 1$ points are compared. SampEn was calculated according to Richman and Moorman (2000):

$$\mathrm{SampEn} = -\ln(A/B) = -\ln(A) + \ln(B), \tag{4}$$

where A is the number of matches with template length $m + 1$ and B is the number of matches with length m. m was set to 1, and r was 0.25 times the standard deviation of the sequence in line with (Corino *et al* 2017).

CosEn is an entropy estimate proposed by Lake and Moorman (2011) that is optimized for AF detection and calculated as

$$\mathrm{CosEn} = \mathrm{SampEn} + \ln(2r) - \ln(\mathrm{mean}(IBI)), \tag{5}$$

where r is the tolerance used for computing SampEn.

The features were computed in sliding time windows of 30 s, 60 s, and 120 s with a 30 s shift. Windows that had less than 20, 40, and 80 intervals for 30 s, 60 s, and 120 s windows, respectively, were excluded from the analysis. For the computation of SampEn and CosEn, the window needed to contain at least nine consecutive IBIs after removing outliers, otherwise the window was excluded.

2.4. Movement intensity

The information from the accelerometer was used to evaluate the movement of the wrist. Movement intensity was defined as

$$Movement\ intensity = \sum_{ax=1}^{3}\left[\frac{1}{l_{acc}}\sum_{i=1}^{l_{acc}}(acc(i)_{ax} - m_{ax})^2\right] \tag{6}$$

where ax is the axis of the accelerometer, l_{acc} the length of acceleration sequence, and m_{ax} the mean acceleration over the sequence on that axis.

Movement intensity was used in the feature computation to discard windows exceeding a predefined movement threshold. Three different thresholds were set for comparison: 75%ile, 50%ile, and 25%ile of the movement distribution of all patients.

2.5. Performance metrics

The discriminative power of the features was determined with the following performance metrics: sensitivity, specificity, accuracy, positive predictive value (PPV), F_1-score and Cohen's kappa. Where *TP* are true positives, *TN* true negatives, *FP* false positives, and *FN* false negatives, sensitivity, specificity, accuracy, and PPV are calculated as follows:

$$Sensitivy = \frac{TP}{TP + FN}, \tag{7}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

and

$$PPV = \frac{TP}{TP + FP}. \tag{10}$$

The F_1-score is a harmonic mean of precision (PPV) and recall (sensitivity) and is based on the efficiency score of van Rijsbergen (1979):

$$\text{F}_1\text{-score} = 2 \times \frac{PPV \times sensitivity}{PPV + sensitivity}. \tag{11}$$

Cohen's kappa (Cohen 1960) is a measure describing the inter-rater agreement of two categorical variables. Kappa is calculated with the following formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}, \tag{12}$$

where p_o is the observed agreement and p_e the chance agreement. p_o is calculated the same way as accuracy and is the percentage of true observations from all the observations. The chance agreement p_e is

$$p_e = \frac{1}{N_{all}^2} \sum_k n_{k1} n_{k2}, \tag{13}$$

where k is the class, N the number of observations, and n_{ki} the number of times rater i predicted class k. In our case, the two raters are the reference and the output of the automatic classification based on the feature.

2.6. Cross-validation

The cut-off values for AF detection for every feature were determined by leave-one-subject-out cross-validation. The data from one subject were held for testing whereas the data from the remaining 26 subjects were used for training. Due to the imbalance in the number of subjects between the AF and non-AF groups, the data from AF group were upsampled in the training set to balance the class distribution.

The cut-off value for every feature with every window length and movement intensity threshold was determined in the training set by using a receiver operating characteristics (ROC) curve and the Youden index (Youden 1950). Figure 2 is an example of the ROC curves and operative points defined by the Youden index during the training phase with features computed from ECG with 30 s window and when data of one patient have been left for testing. For every defined cut-off value, the procedure of holding data for testing from one patient was repeated 27 times.

3. Results

The AF classification performance of every feature was calculated by aggregating the results obtained with the test data from every round of cross-validation. The results were computed with ECG by varying the window length and with PPG by varying both the window length and movement threshold. Cohen's kappa was selected as the metric to compare the performance of individual features, since it is a metric not affected by the imbalance between the two classes, i.e. AF and non-AF. For the comparison between features computed from ECG, figure 3 shows a histogram of kappa for every feature when varying the window length. Based on this comparison, CosEn is the strongest feature from ECG with kappa 0.901–0.950. For every feature the longest window gave the best performance.

The effect of the movement threshold to PPG derived features was compared using a 120 s window. Furthermore, for comparing the effect of the window length, the strictest movement threshold (25%ile) was used. Figure 4 shows on the left a histogram of kappa of the features when window length is kept constant but movement threshold varies. On the right, there is a histogram when movement threshold is kept constant and window length varies. When the movement threshold is set to reject more movement, the performance increases for all the features. The results of varying the window length are in line with the results from ECG and with a longer window length better kappa is obtained. CosEn is again one of the strongest features with $\text{kappa}_{120\text{s}(25\%\text{ile})}$ 0.956, but additionally pNN40 and pNN70 appear as strong features for AF classification from PPG with $\text{kappa}_{120\text{s}(25\%\text{ile})}$ 0.953 and 0.945, respectively.

Restricting sufficiently the accepted amount of movement resulted in an increase in the performance. However, when windows for feature computation are discarded from the analysis due to movement, the coverage, which is defined as the percentage of 30 s instants with a feature value, also decreases. In figure 5, the mean coverage of all patients with ECG and PPG when varying the movement threshold is presented with all the different window lengths. On average, the coverage with ECG with different window lengths is 92.1%. With PPG the average coverage with no movement threshold, the 75%ile, 50%ile, and 25%ile thresholds are 57.6%, 54.9%, 45.4%, and 24.0%, respectively. The movement thresholds are determined separately for different window lengths and therefore varying the window length does not significantly influence the coverage.

CosEn resulted as the best feature from both from ECG and PPG with 25%ile threshold when kappa was compared. Table 2 shows sensitivity, specificity, PPV, accuracy, kappa, and F_1-score of CosEn with both ECG and PPG when varying window length and movement threshold. When the 25%ile threshold is used, the classifica-

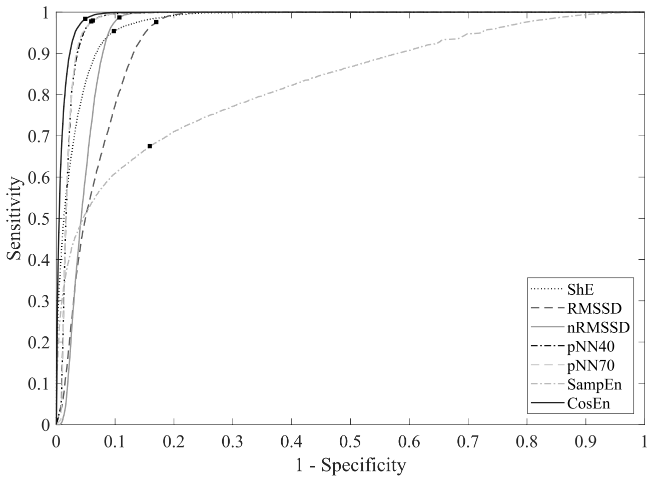

Figure 2. An example of ROC curves during the training phase with features computed from ECG with a 30 s window. The black squares on the curve are the operative points defined by the Youden index.

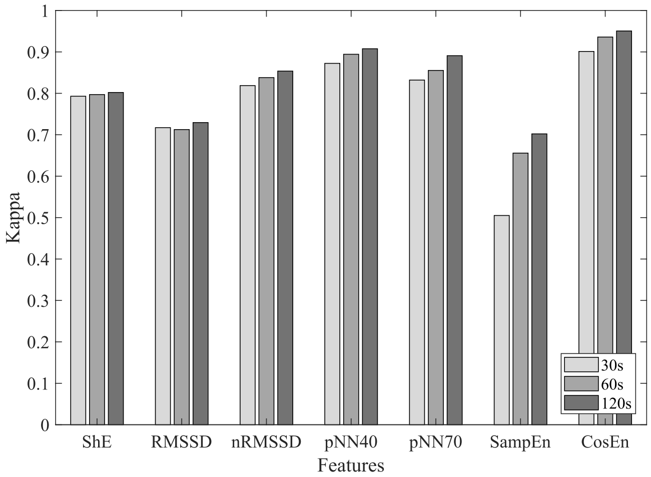

Figure 3. Cohen's kappa of features computed from ECG with different window lengths.

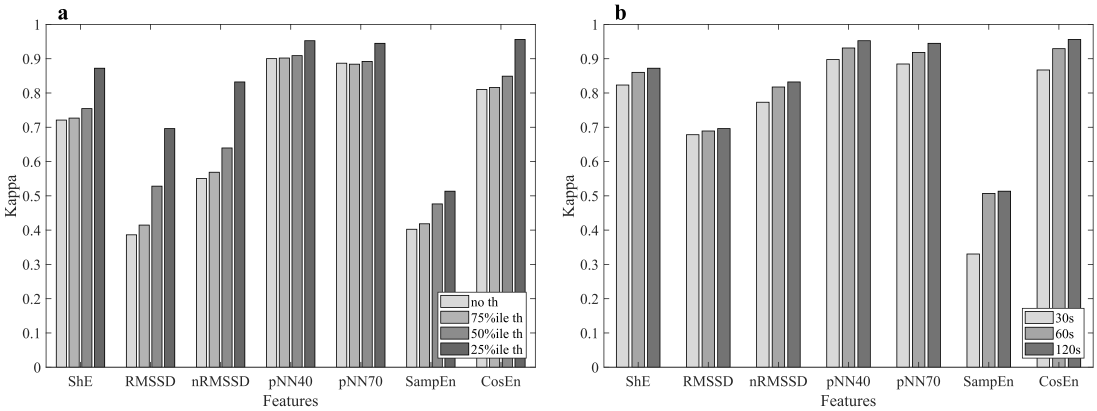

Figure 4. Cohen's kappa of features computed from PPG with different movement thresholds using a 120 s window (a) and window lengths using 25%ile threshold (b).

tion performance with PPG approaches the results with ECG with all the metrics. With 60 s and 120 s windows, the results with PPG are at the level of ECG. Figure 6 shows kappa with CosEn and pNN40 from PPG as a function of movement threshold compared to how the coverage changes when the movement threshold is changed. Kappa with CosEn from ECG is marked as a reference, since it was the highest kappa obtained with ECG. It is visible how kappa increases when the movement threshold is stricter and both CosEn and pNN40 from PPG

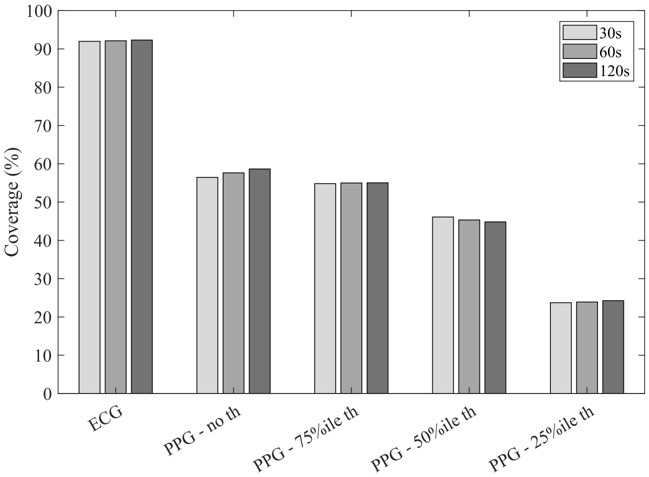

Figure 5. Mean coverage of ECG and PPG with different movement thresholds and window lengths.

Table 2. Performance with CosEn from PPG and ECG with different window lengths and movement thresholds.

Window		Sensitivity	Specificity	PPV	Accuracy	Kappa	F_1-score
30 s	ECG	*0.981*	*0.950*	*0.886*	*0.958*	*0.901*	*0.931*
	PPG—no th	0.952	0.819	0.684	0.857	0.691	0.796
	PPG—75%ile th	0.950	0.829	0.695	0.864	0.703	0.803
	PPG—50%ile th	0.960	0.866	0.746	0.893	0.761	0.839
	PPG—25%ile th	0.968	0.934	0.855	0.944	0.867	0.908
60 s	ECG	*0.980*	*0.971*	*0.931*	*0.973*	*0.936*	*0.955*
	PPG—no th	0.942	0.884	0.773	0.901	0.777	0.849
	PPG—75%ile th	0.948	0.890	0.784	0.907	0.791	0.858
	PPG—50%ile th	0.959	0.916	0.829	0.929	0.837	0.889
	PPG—25%ile th	**0.975**	**0.968**	**0.928**	**0.970**	**0.929**	**0.951**
120 s	ECG	*0.983*	*0.978*	*0.948*	*0.980*	*0.951*	*0.965*
	PPG—no th	0.959	0.898	0.798	0.916	0.810	0.871
	PPG—75%ile th	0.960	0.901	0.805	0.919	0.816	0.876
	PPG—50%ile th	0.966	0.919	0.839	0.934	0.849	0.898
	PPG—25%ile th	**0.984**	**0.980**	**0.955**	**0.981**	**0.956**	**0.970**

eventually reach the same kappa as with ECG. With higher movement thresholds pNN40 performs better, thus being more robust against movement artefacts. In contrast to kappa, coverage decreases when excluding more movement from the analysis.

4. Discussion

In this study, we compared for the first time commonly used IBI features derived from ECG and PPG for AF detection in free-living conditions with 24 h measurements. CosEn resulted as the most powerful individual feature from ECG and with strict movement threshold from PPG reaching high sensitivity, specificity, and kappa with both measurement modalities. With the 25%ile threshold for movement, pNN40 calculated from PPG gave similar kappa (0.952) compared to CosEn (0.956). When accepting more movement, pNN40 performed better, therefore it is more robust. Even without using the movement information the coverage was substantially reduced by the movement artefacts, being on average 58% compared to 92% with ECG. During high intensity movement the pulses were not detected from the PPG and these segments were excluded from the analysis even without using any movement threshold. This also explains why the coverage and performance remain at the same level when 75%ile threshold was applied.

The results indicate that when periods of PPG data affected by movement are discarded from the analysis, i.e. when we expect stable measurement conditions and better signal quality, the PPG works equally well as an ECG Holter measurement. The impact of presence of simulated muscle artefact noise on the performance of IBI-based AF detection algorithms for ECG has been previously studied by Oster and Clifford (2015). They showed a linear increase in the performance when SNR increased. In the large part our results are in line with their findings.

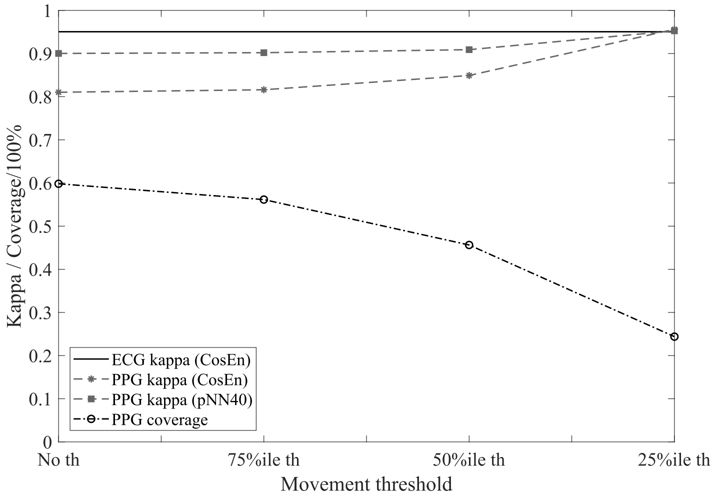

Figure 6. Increase of kappa with CosEn and pNN40 from PPG and decrease of coverage as a function of movement threshold. No movement threshold was applied when computing kappa with CosEn from ECG, but for easier comparison it is marked as a line of reference.

Reducing the movement artefacts had the highest impact on specificity and PPV, thus reducing the false positives. However, in contrast to the results of Oster and Clifford (2015) sensitivity also improved. In our current study, the movement thresholds were not optimized in terms of trade-off between classification performance and coverage. That is left for further research. One option could also be to incorporate movement or signal quality information to the classification model as a feature (Nemati *et al* 2016, Shashikumar *et al* 2017) to further improve accuracy and increase coverage.

We compared only individual features to make a more objective comparison between measurement modalities, i.e. ECG and PPG. Alternatively, we could have compared the AF detection of a classification model combining either ECG or PPG derived features. With the current way, the comparison is independent from the choice of the feature combination and classification model. These choices might be different when optimized for ECG and PPG, depending also on whether PPG is affected by movement artefacts or not. Better classification performance could be possibly obtained when more features are combined. Therefore the classification performance obtained with an individual feature is not intended to reflect the maximum performance that is possible to obtain with PPG in free-living conditions. In particular, adding information beyond IBI-derived features, such as morphology features (Pantelopoulos *et al* 2017, Schäck *et al* 2017) and spectral features (Shashikumar *et al* 2017), could possibly further boost the performance.

There are some limitations in the study. The dataset was not large enough to divide the data into a separate training and an unseen test set, and therefore results with cross-validation are presented. There is an imbalance between the two classes that affects some of the performance metrics, such as accuracy, PPV, and F_1-score. Therefore, these metrics are not comparable to the results obtained in other studies with balanced class distributions. In addition, the division into two groups was made solely based on the rhythm, i.e. whether AF was present or not, and the patient characteristics between these groups resulted in being slightly different. However, in such a small dataset, the possible influence of these differences to the results is difficult to assess.

Another limitation of the study is that all the patients with AF had continuous AF. Ideally, the aim was to measure events of paroxysmal AF, but no paroxysmal AF was detected in our study population. It was not possible to determine before the measurement if a patient would have a paroxysmal event during the measurement period. This also reflects the current problem with 24 h Holter monitoring that rare events are missed if they occur outside the monitoring period (Rosero *et al* 2013). Therefore, with the current dataset it is not possible to assess how accurately paroxysmal events are detected and whether the window length influences that. Nevertheless, this is the first study comparing ECG and PPG for AF detection during daily life and the results, even with only continuous AF, are promising. Further studies with prolonged PPG measurements to multiple days or weeks, which is difficult and uncomfortable to measure with a Holter, can most likely better capture subjects with paroxysmal events to the study population and give information about their detection with PPG.

As mentioned earlier, even when the movement intensity was not considered in the analysis, the coverage of the rhythm classification with PPG was on average 58% due to the inability to detect pulses. When adding the assessment of movement intensity, the coverage reduced even more. This causes a limitation for the use of IBI-based methods for continuous monitoring to detect paroxysmal events that might occur during the periods when coverage is lost. The detection of these AF events would be therefore partly dependent on the frequency, duration, and daily distribution of the events. However, the lost coverage could be compensated for with a prolonged monitoring period up to weeks or even months. The impact of lost coverage on the sensitivity of detecting paroxysmal events and the added value of prolonged monitoring with PPG should be assessed with further studies. Moreover, development of methods to improve the coverage could help to overcome the issue with data loss.

In general, the comparison between the performance of different algorithms developed for AF detection from PPG is difficult. Algorithms are developed in different settings, and with datasets having different characteristics, e.g. AF versus subjects with sinus rhythm and AF versus subjects with other rhythms, such as the presence of premature contractions. We have previously shown that results from one measurement setting and patient group are not directly applicable to another setting and patient group with different characteristics (Eerikäinen *et al* 2017). In addition, the amount of data points to compute a feature, i.e. window length, which is not equal between different solutions, influences the results. This was also shown in the work of Tang *et al* (2017) when comparing models using 1 min, 2 min, and 10 min data.

5. Conclusion

Comparable results in AF detection are possible to obtain with PPG and ECG when using a single feature and when discarding PPG signals during movement identified with the accelerometer. On the one hand this leads to a limited coverage, but on the other hand PPG devices can be worn for much longer periods than ECG recorders compensating for the lost coverage. The prolonged monitoring period might have an added value in detecting paroxysmal AF. Therefore, wrist-worn PPG devices provide a promising solution for long-term monitoring of AF. Future studies should be performed to assess the impact of the coverage loss on the sensitivity of detecting paroxysmal AF events.

Acknowledgments

This research was performed within the framework of the strategic joint research program on Data Science between TU/e and Philips Electronics Nederland BV. The authors would like to thank N Sturkenboom (MD), L Verborg (MD), R Eerdekens (MD), L van den Heuvel, the personnel of the Holter department, and the Holter analysts of the Catharina Hospital for their help and contributions in the data collection.

References

Allen J 2007 Photoplethysmography and its application in clinical physiological measurement *Physiol. Meas.* **28** R1–39

Bonomi A G, Schipper F, Eerikäinen L M, Margarito J, Aarts R M, Babaeizadeh S, Morree H M D and Dekker L 2016 Atrial fibrillation detection using photo-plethysmography and acceleration data at the wrist *Comput. Cardiol.* **43** 277–80

Brachmann J, Morillo C A, Sanna T, Di Lazzaro V, Diener H C, Bernstein R A, Rymer M, Ziegler P D, Liu S and Passman R S 2016 Uncovering atrial fibrillation beyond short-term monitoring in cryptogenic stroke patients: three-year results from the cryptogenic stroke and underlying atrial fibrillation trial *Circ.: Arrhythmia Electrophysiol.* **9** 1–10

Camm A J *et al* 2010 Guidelines for the management of atrial fibrillation: the task force for the management of atrial fibrillation of the European Society of Cardiology (ESC) *Eur. Heart J.* **31** 2369–429

Chan P H, Wong C K, Poh Y C, Pun L, Leung W W C, Wong Y F, Wong M M Y, Poh M Z, Chu D W S and Siu C W 2016 Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting *J. Am. Heart Assoc.* **5** 1–8

Chong J W, McManus D D and Chon K H 2015 Arrhythmia discrimination using a smart phone *IEEE J. Biomed. Health Inform.* **19** 1–4

Cohen J 1960 A coefficient of agreement for nominal scales *Edu. Psychol. Meas.* **20** 37

Corino V D A, Laureanti R, Ferranti L, Scarpini G, Lombardi F and Mainardi L T 2017 Detection of atrial fibrillation episodes using a wristband device *Physiol. Meas.* **38** 787–99

Dash S, Chon K H, Lu S and Raeder E A 2009 Automatic real time detection of atrial fibrillation *Ann. Biomed. Eng.* **37** 1701–9

Eerikäinen L M, Dekker L, Bonomi A G, Vullings R, Schipper F, Margarito J, Morree H M D and Aarts R M 2017 Validating features for atrial fibrillation detection from photoplethysmogram under hospital and free-living conditions *Comput. Cardiol.* **44** 3–6

Lake D E and Moorman J R 2011 Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices *AJP: Heart Circ. Physiol.* **300** H319–25

Lee J, Reyes B A, McManus D D, Mathias O and Chon K H 2012 Atrial fibrillation detection using a smart phone *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 1177–80

Lemay M, Fallet S, Renevey P, Sol J, Pruvot E and Vesin J M 2016 Wrist-located optical device for atrial fibrillation screening: a clinical study on twenty patients *Comput. Cardiol.* **43** 681–4

McManus D D, Lee J, Maitas O, Esa N, Pidikiti R, Carlucci A, Harrington J, Mick E and Chon K H 2013 A novel application for the detection of an irregular pulse using an iPhone 4S in patients with atrial fibrillation *Heart Rhythm* **10** 315–9

Nemati S, Ghassemi M M, Ambai V, Isakadze N, Levantsevych O, Shah A and Clifford G D 2016 Monitoring and detecting atrial fibrillation using wearable technology *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 3394–7

Oster J and Clifford G D 2015 Impact of the presence of noise on RR interval-based atrial fibrillation detection *J. Electrocardiol.* **48** 947–51

Page R L, Wilkinson W E, Clair W K, McCarthy E A and Pritchett E L 1994 Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia *Circulation* **89** 224–7

Pantelopoulos A, Faranesh A, Milescu A, Hosking P, Venkatraman S and Heneghan C 2017 Screening of atrial fibrillation using wrist photoplethysmography from a fitbit tracker *Iproceedings* 3 e17

Richman J S and Moorman J R 2000 Physiological time-series analysis using approximate entropy and sample entropy *Am. J. Physiol. Heart Circ. Physiol.* **278** H2039–49

Rosero S Z, Kutyifa V, Olshansky B and Zareba W 2013 Ambulatory ECG monitoring in atrial fibrillation management *Prog. Cardiovascular Dis.* **56** 143–52

Sanna T *et al* 2014 Cryptogenic stroke and underlying atrial fibrillation *New Engl. J. Med.* **370** 2478–86

Schäck T, Harb Y S, Muma M and Zoubir A M 2017 Computationally efficient algorithm for photoplethysmography-based atrial fibrillation detection using smartphones *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 104–8

Shan S M, Tang S C, Huang P W, Lin Y M, Huang W H, Lai D M and Wu A Y A 2016 Reliable PPG-based algorithm in atrial fibrillation detection *IEEE Biomedical Circuits and Systems Conf. (BioCAS)* (https://doi.org/10.1109/BioCAS.2016.7833801)

Shashikumar S P, Shah A J, Li Q, Clifford G D and Nemati S 2017 A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology *IEEE EMBS Int. Conf. on Biomedical and Health Informatics* pp 141–4

Tang S C, Huang P W, Hung C S, Shan S M, Lin Y H, Shieh J S, Lai D M, Wu A Y and Jeng J S 2017 Identification of atrial fibrillation by quantitative analyses of fingertip photoplethysmogram *Sci. Rep.* **7** 45644

Valenti G and Westerterp K R 2013 Optical heart rate monitoring module validation study *Digest of Technical Papers—IEEE Int. Conf. on Consumer Electronics* pp 195–6

van Andel J, Ungureanu C, Aarts R, Leijten F and Arends J 2015 Using photoplethysmography in heart rate monitoring of patients with epilepsy *Epilepsy Behav.* **45** 142–5

van Rijsbergen C J 1979 *Information Retrieval* 2nd edn (London: Butterworth)

Youden W J 1950 Index for rating diagnostic tests *Cancer* **3** 32–5