Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

A Method for Assessing the Reliability of Heart Rates obtained from Ambulatory ECG

Christina Orphanidou[1], Timothy Bonnici[2], David Vallance[3], Alexander Darrell[1], Peter Charlton[4] and Lionel Tarassenko[1]

[1]Institute of Biomedical Engineering , University of Oxford, UK, christina.orphanidou@eng.ox.ac.uk
[2]Nuffield Department of Medicine, University of Oxford, UK.
[3]Nuffield Department of Surgery, University of Oxford, UK.
[4]Kings College London, King's Health Partners, UK.

***Abstract*—In this paper we present a method of assessing the reliability of heart rates (HRs) obtained from ambulatory ECGs. Our method assigns a Reliability Index (RI) to ECG segments based on a set of physiologically relevant rules prior to using a template matching approach. We validated the algorithm on 1500 manually annotated samples of ECG taken from two different studies and using three different sensors at different sampling rates. The sensitivity of our method was 98% and the specificity was 94%. Our method matched or was more conservative than the human annotations in 99.4% of the samples, making it a promising tool for inclusion in next-generation wearable sensors.**

Keywords-signal quality, signal reliability, heart rate, ambulatory ECG.

I. INTRODUCTION

Of the abnormal vital signs linked to the occurrence of adverse events in hospital patients, both an increased or decreased heart rate (tachycardia and bradycardia, respectively) have been suggested as important signals to trigger interventions for the prevention of cardiopulmonary arrests [1,2]. As patient monitoring is progressing towards future of ubiquitous monitoring of ambulatory patients with wearable sensors, physiological data is becoming noisier and harder to interpret (because of movement artifact and sensor noise) leading to higher false alarm rates [3] which not only add to the already high workload of clinical staff, but may also lead to the phenomenon of "alarm fatigue" whereby nursing staff ignore all alarms assuming that they are artifactual. An effective system, therefore, must reduce false alarm rates whilst maintaining a high sensitivity so that adverse events are not missed. False alarm rates can be reduced and the probability of identifying precursors to adverse events can be increased through the use of signal quality indices (SQIs) which indicate the degree of trust that can be placed on a particular section of a physiological signal, such as the ECG.

In recent years, sensor fusion approaches have been proposed for improving the reliability of Heart Rates (HR) derived from physiological signals. These approaches use SQIs to qualify multiple simultaneously recorded physiological signals (or more than one channel of ECG) and then combine them optimally in order to obtain a more reliable value [6, 10]. While these methods have demonstrated clear improvement in the reliability of the HRs obtained, the new generation of wearable sensors currently emerging makes it necessary to be able to qualify single-channel ECG data. Methods for qualifying HRs obtained from single-channel ECGs have been proposed in the literature, based on techniques such as fuzzy logic [11] and Support Vector Machines [12]. These methods, however, appear to have high sensitivity to noise and produce rather conservative HR qualification results. In addition, many of the proposed algorithms require tuning to different sensors and datasets, which restrict their application. Data collected using wearable sensors tend to be corrupted by movement artifact. It is important to use as much of the high-quality data as possible. With an SQI based on stringent morphological rules, segments of the signal which would provide a useful HR tend to be rejected, which can lead to extended periods of time without an HR output from the monitor (if for instance the patient is continually moving, such as during a physiotherapy session in hospital).

In this paper, we present an approach to automatically qualifying segments of ECG for reliability in HR estimation, using a Reliability Index (RI). In our proposed approach, R-peak detection is performed using the well-known Hamilton and Tompkins algorithm [7]. We then apply a series of physiologically-relevant rules before using an adaptive template matching approach in order to obtain a "reliable" or "unreliable" label for each ECG segment. The template used is derived from the QRS complexes (the three main deflections seen on a typical ECG trace). In order to provide a generalized tool, which would not require tuning to a specific dataset, we trained and validated our RI on data obtained from a large number of subjects, using three different ECG sensors, all with different sampling rates.

II. METHODS

A. Data Selection

To develop and validate our proposed algorithm, we used 1500 ECG samples, each 10 seconds in duration, taken from two different clinical studies. The choice of 10 seconds allows physiologically relevant rules to be applied without any localized artifacts resulting in significant data loss. 500 samples were randomly selected from the Sana/Physionet

database made available freely as part of the Physionet/Computing in Cardiology Challenge 2011 [4]. The original database consisted of 1500 segments of 12-lead ECG, sampled at 500 Hz with 16-bit resolution, collected using a mobile phone. For the current study, we used the first lead of ECG data. The other 1000 segments were taken from another study, carried out in our research group, whose purpose was to assess the performance of different ECG and pulse oximetry wearable sensors [5]. In that study, four ECG sensors were tested, two of which provided continuous ECG. We used data collected from those two sensors. Sensor D was attached to the patients using conventional ECG leads and the sensor data was sampled at 100 Hz with 12-bit resolution. Sensor E was part of a belt worn around the chest and the sensor data was sampled at 256 Hz with 10-bit resolution. Each sensor was worn by 12 subjects for an average of 11 hours for sensor D and 14 hours for sensor E. To ensure a balanced training set, we randomly selected 500 10-second segments from data acquired with sensor D and 500 from data acquired with sensor E.

B. Labeling

Labeling was carried out in two stages. In the first stage, two human experts visually examined and categorized the 1500 samples based on the following rule:

An ECG segment is labeled as "reliable" if a human expert can confidently derive a reliable HR from it. Otherwise it is labeled as "unreliable".

When there was disagreement between the two experts, a third human expert reviewed the ambiguous samples and gave the decisive label. In total, 12.2% of the data (183 samples) had to be reviewed by the third human expert. In the first stage, 991 samples (66%) were ranked as "reliable" and 509 (34%) were ranked as "unreliable".

In the second stage, the R-peaks of the 991 "reliable" samples were manually identified and marked by the two human experts. As in the first stage of labeling, when there was disagreement between the two human annotators, the third human expert reviewed the sample and gave a third annotation which was taken as the decisive one. At this second stage, 103 samples (10.4%) had to be reviewed by the third human expert. Because our RI is intrinsically linked to the QRS detector, we then proceeded to run the QRS detector on the 991 samples and re-labeled the samples by comparing them to the human annotations, based on the following rule:

An ECG segment is labeled as "unreliable" if more than one R-peak is missed by the QRS detector or if more than one instance of artifact is mistakenly identified as an R-peak by the QRS detector. Otherwise it is labeled as "reliable".

Because our samples are only 10 seconds in duration, the rule is deliberately conservative: if more than one R-peak in a 10 second sample is missed or more than one T-wave or noisy peak is mistakenly identified as an R-peak, the HR extrapolated from this sample will have a large error.

This work was supported by the RCUK Digital Economy grant award, EP/H0199944/1.

C. Reliability Index

The proposed algorithm is illustrated in Fig. 1. Once QRS detection has been performed on a sample, a set of physiologically relevant rules are used. If any of the rules is not satisfied, the sample is labeled as "unreliable". If a rule is satisfied, the next rule is tested and so on.

- **Rule 1**: The HR extrapolated from the 10-second sample is between 40-180 beats per minute (bpm) (physiologically probable range of HR for healthy adults).
- **Rule 2**: The maximum acceptable gap between successive R-peaks is 3 seconds. (This rule ensures no more than one beat is missed. The value of 3 seconds follows on from the choice of 40 bpm for the minimum acceptable HR).
- **Rule 3**: The ratio of the maximum RR interval to the minimum RR interval within the sample should be less than 2.2 (This is a rather conservative limit since we would not expect the HR to change by more than 10% in a 10-second sample. We use a limit of 2.2 to allow for a single missed beat).

If all three rules are satisfied, an adaptive QRS template matching approach is then used, as explained next.

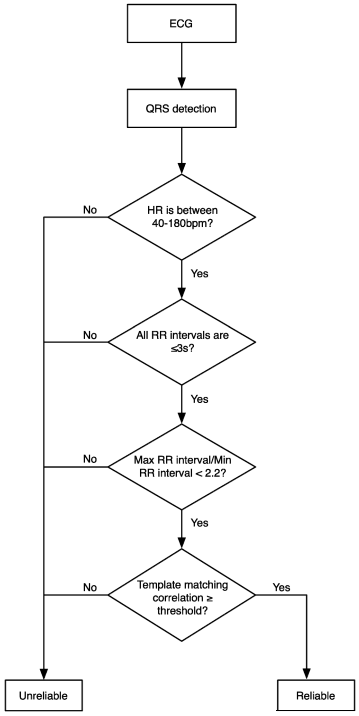

Figure 1: Flowchart of Reliability Index (RI) approach. Once QRS detection is performed, a set of physiologically relevant rules are applied. If any of the rules is not satisfied the sample is classified as unreliable. If all rules are satisfied, a QRS template matching approach is employed to qualify the segment.

Adaptive QRS template matching

Template matching approaches have been used in the past for identifying ventricular ectopic beats [8] and heartbeats [13] in the ECG and for signal quality assessment of the PPG [9]. Regardless of the actual morphology of the QRS complexes in a particular ECG sample, QRS template matching searches for regularity in a segment which is an indicator of reliability (since a segment contaminated by movement artifact is irregular in morphology). Our approach is as follows:

- Using the R-peaks of each ECG sample, the median RR interval is calculated.
- Individual QRS complexes are then extracted by taking a window, the width of which is the median RR interval, centered on each detected R-peak.
- The average QRS template is then obtained by taking the mean of all QRS complexes in the sample.
- The correlation coefficient of each individual QRS complex with the average QRS template is then calculated.
- The average correlation coefficient is finally obtained by averaging all correlation coefficients over the whole ECG sample.

In the training stage of our algorithm development, the average correlation coefficient was optimized in order to determine the threshold which gives the best differentiation between "reliable" and "unreliable" segments.

Fig. 2 shows an example of the average QRS template creation from a morphologically regular sample of ECG.

D. Training and Validation

We used 1000 out of the 1500 ECG samples (chosen at random) to optimize the threshold of the average correlation coefficient. The remaining 500 samples were used in the testing stage. We tested values of the average correlation coefficient ranging from 0.5 to 1 and averaged the performance of the algorithm over 5 cycles of cross-validation. The optimum threshold for the average correlation coefficient was found to be 0.66 although there was little difference between the performance on the training set for values between 0.58 and 0.7.

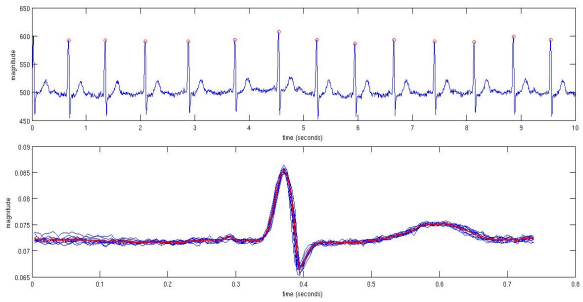

Figure 2: The top plot shows a 10-second ECG sample with the R-peaks marked by red circles and the bottom plot shows individual QRS complexes in blue, with the average template taken to be the mean of all those, shown in red.

III. RESULTS AND DISCUSSION

We tested our RI on the 500 ECG samples of the test set. The sensitivity of our method was 98% and the specificity was 94% (where a "false positive" is a valid segment being labeled as "unreliable"). Overall, our method matched or was more conservative than the human labeling in 99.4% of the cases. In order to investigate whether the sensor has any effect on the performance of the RI we also calculated the performance of our method for the three different sensors separately, as shown in table 1.

Table 1: Performance of RI based on data source. The third column gives the percentage of samples for which the RI either matched or was more conservative than the human labeling.

Database	Sensitivity	Specificity	Performance
Sana/Physionet (500Hz/16-bit)	94.6%	91.6%	99.1%
Sensor D (100Hz/12-bit)	97.1%	82.7%	99.3%
Sensor E (256Hz/10-bit)	100%	98.5%	99.9%

The performance of the RI on data from the different sensors is independent of the sampling rate and excellent results are obtained for all sensors. Sensor D has the worst specificity (i.e. a high number of false positives). Closer inspection of the data reveals that a large proportion of the data collected with this sensor was clipped either at the R-peak or more often at the S-trough of the QRS complex. Since the R-peaks were clearly visible, the human annotators labeled those segments as "reliable" but the algorithm labeled them as "unreliable", resulting in a high number of false positives and hence a low specificity value.

The overall results justify the conservative design of the algorithm, giving a higher sensitivity than specificity which is desirable for a method to be used as part of a patient monitoring system. The performance of our algorithm can also be illustrated using the ambulatory ECG recording of Fig. 3, which contains short periods of artifact. The shaded areas are those which the RI labels as "unreliable".

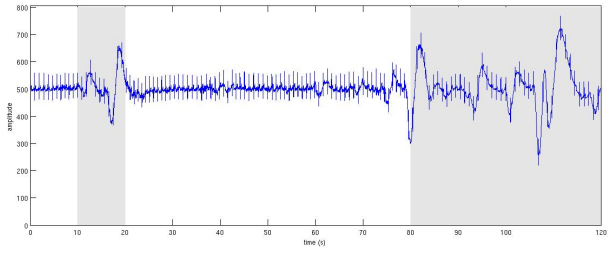

Figure 3: A two-minute ECG recording from an ambulatory patient using sensor E. The shaded areas indicate the 10-second segments which were labeled as "unreliable" by the Reliability Index.

It is evident that our RI has successfully identified the segments of artifact which would be likely to produce an erroneous estimate of HR.

Fig. 4 shows the distribution of HRs obtained from all the ambulatory hospital patients in [5] from segments labeled as "reliable" and those labeled as "unreliable".

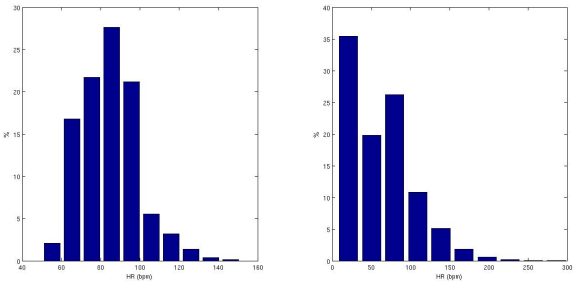

Figure 4: histograms of HRs obtained from segments labeled as "reliable" (left) and "unreliable" (right). The data used is from ambulatory hospital patients, taken from the study presented in [5].

IV. CONCLUSIONS

This paper presented an automated way of assessing the reliability of HRs obtained from a single channel of ECG . Our method is based on a set of physiologically relevant rules and employs an adaptive QRS template matching approach to check for morphological regularity in a 10-second segment of ECG. We validated and tested our approach on 1,500 10-second segments collected using three different sensors and annotated by human experts and we obtained a sensitivity of 98% and specificity of 94%. Our method either matched or was more conservative than the human annotations for 99.4% of the time. Our results are promising and may be further improved by considering longer segments of data so that more physiologically relevant rules can be added based on the heart rate variability characteristics of the ECG. In addition, unlike other approaches proposed in the literature, our approach was developed as a generalized tool and does not require tuning to a specific dataset.

V. REFERENCES

[1] J. F. Fieselmann, M. S. Hendryx, C. M. Helms and D. S. Wakefield, "Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients", J. Gen. Intern. Med. 8, 1993, pp. 354-360.

[2] C. W. Seymour, J. M. Kahn, C. R. Cooke, T. R. Watkins, S. R. Heckbert and T. D. Rea, "Prediction of critical illness during out-of-hospital emergency care", JAMA, J. Am. Med. Assoc. 304, 2010, pp. 747-754.

[3] G.D. Clifford and D. Clifton, "Wireless Technology and Disease Management in Medicine", Ann. Rev. Med. 63, 2012, pp. 479-492.

[4] I. Silva, G. B. Moody and L. Celi, "Improving the Quality of ECGs Collected Using Mobile Phones: The Physionet/ Computing in Cardiology Challenge 2011", Computing in Cardiology 2011.

[5] T. Bonnici , C. Orphanidou, D. Vallance, A. Darrell and L. Tarassenko "Testing of Wearable Monitors in a Real-World Hospital Environment: What Lessons Can Be Learnt?" in Body Sensor Networks 2012, pp. 79-84.

[6] Q. Li , R. G. Mark and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter", Physiological Measurement 29, 2008, pp. 15-32.

[7] P.S. Hamilton and W. J. Tompkins, "Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database", IEEE Transactions on Biomedical Engineering, 33(12) , 1986, pp. 1157-1165.

[8] V. Krasteva and I. Jekova, "QRS Template Matching for Recognition of Ventricular Ectopic Beats," in Annals of Biomedical Engineering, 35:12, 2007, pp. 2065-2076.

[9] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals", Physiological Measurement, in press, 2012.

[10] M. H. Ebrahim, J. M. Feldman and I. Bar-Kana., "A robust sensor fusion method for heart rate estimation", Journal of Clinical Monitoring, 13, 1997, pp. 385-393.

[11] J. Liu , T. M. McKenna, A. Gribok, B. A. Beidleman.,W. J. Tharion and J. Reifman, "A fuzzy logic algorithm to assign confidence levels to heart and respiratory time series", Physiological Measurement 29, 2008, pp. 81-94.

[12] C. Yu, Z. Liu, T. McKenna, A. T. Reisner and J. Reifman , "A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms", Journal of the American Medical Informatics Association, 13-3, 2006, pp. 309-320.

[13] H. L. Chan, G. U. Chen, M. A Lin. and S. C Fang., "Heartbeat Detection Using Energy Thresholding and Template Match", in Proceedings of EMBC 2006, pp. 6668-6670.