Hardware-Efficient Robust Biometric Identification from Amplitude and Interval Features of 0.58 Second Limb (Lead I) ECG Signal Using Logistic Regression Classifier

Md Nazmus Sahadat, *Student Member*, *IEEE, EMBS,* Eddie L. Jacobs, *Senior Member, IEEE* and Bashir I. Morshed, *Member*, *IEEE, EMBS*

***Abstract*— The electrocardiogram (ECG), widely known as a cardiac diagnostic signal, has recently been proposed for biometric identification of individuals; however reliability and reproducibility are of research interest. In this paper, we propose a template matching technique of 12 features through fiducial point detection and logistic regression classifier that achieved high reliability and accuracy. Non-invasive ECG signals were captured using our custom-built ambulatory EEG/ECG embedded device (NeuroMonitor). ECG data were collected from healthy subjects (10), between 25-35 years, for 10 seconds per trial. The number of trials from each subject was 10. From each trial, only 0.58 seconds of Lead I ECG data were used as template. Data were randomly separated into training and testing sets at a ratio of 80:20. Test data was used to find the classification accuracy. New hardware-efficient fiducial point detection for feature extraction was implemented. ECG template data with 12 extracted features provided the best performance in terms of accuracy (up to 100%) and processing complexity (computation time of 1.2ms). This work shows that a single limb (Lead I) ECG can robustly identify an individual quickly and reliably with minimal contact and minimal data processing using the proposed algorithm.**

I. Introduction

Biometric identification methods such as face, iris, voice, and fingerprint detection are currently in use [1]. However, these identification methods have practical limitations due to high computational complexity, low reliability and reproducibility, or the potential of a security breach with artificial or forged features [2]. The electrocardiogram (ECG), known as a cardiac diagnosis signal for medical applications, might provide higher security for biometric identification [3-11]. One of the major concerns with this method is the variation of ECG signals with stress, anxiety, and the time of day [3-5]. Several recent works have shown that heartbeat contains only the scalar differences under stress [3-6]. Researchers have also shown the uniqueness of an individual's cardiac signals [7]. In fact, the ECG signal not only uniquely identifies an individual, but also can act as a living biometric [6].

Various methods have been proposed for biometric identification with ECG. In 2001, Biel *et al.* demonstrated the feasibility of ECG-based human identification by supervised classification over significant principal components of several morphological features including amplitudes, durations and areas of the P, Q, R, S, T waves and the ST segment [7]. In some reports, only durations of characteristic waves and intervals between characteristic points were selected as the discriminating features [3,9]. Kyoso and Uchiyama applied discriminate analysis with two features from P duration, PQ interval, QRS duration and QT interval to identify the registered ECGs from nine subjects by selecting the smallest Mahalanobis distance [9]. Improved classification was achieved using the combination of QRS duration and QT interval. Israel *et al.* extracted 15 time-intervals from a heart beat and further reduced feature dimensionality to 12 by the Wilke's lambda method [3]. Shen *et al.* proposed a two-step identification scheme where a template match method was first applied to find possible candidates followed by a decision-based neural network with inputs of seven temporal and amplitude features to complete final verification [8]. Shen applied quartile discriminant measurement to reduce the number of ECG features from 17 to 11, thereby achieving an identification rate of 95% for a large (169) subject pool [10]. Wubbeler *et al.* proposed a two dimensional heart vector determined from amplitude values of leads I–III composition [11].

Most of the methods listed above use multiple features from the ECG signal to classify individuals. Multiple feature extraction from the ECG signal is time consuming and hardware inefficient. This work shows a simplified approach to extract those features using only a window based max/min method. The novelty of this work is in the use of a simplified feature extraction technique from a brief ECG data (Lead I only) to robustly classify individuals with high degree of reliability.

II. Setup and Procedure

ECG data are captured from subjects using our custom-built ambulatory EEG/ECG embedded device (NeuroMonitor) [12, 13]. Captured data are wirelessly (Bluetooth) transmitted to a remote computer for processing and analysis. The NeuroMonitor device contains a C program for data capture and transmission, while MATLAB is used in the computer for data processing and classification. A block diagram of the entire process is shown in Fig. 1.

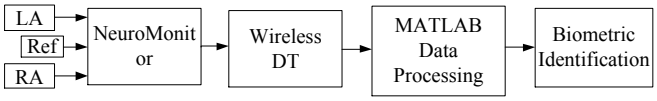

Figure 1. Block diagram of the overall system

A. Device

The NeuroMonitor device is a small (5.58 cm x 2.03 cm

Md Nazmus Sahadat, Eddie Jacobs and Bashir I. Morshed are with the Department of Electrical and Computer Engineering, The University of Memphis, Memphis, TN 38152 USA. (E-mails: mnshadat@memphis.edu; eljacobs@memphis.edu; bmorshed@memphis.edu).

Preprint submitted to 36th Annual International IEEE EMBS Conference.
Received March 15, 2014.

x 0.91 cm) and light-weight (41.8 g), and contains hardware for ECG/EEG data collection [12, 13]. For ECG data collection, an overall gain of A_v = 93.28 and a bandwidth of 0.5 – 126 Hz are used for the analog front end. The sampling rate is 256 samples/sec and ADC resolution is 16 bit. Lead I (Limb) configuration is chosen to capture data from each individual. Three electrodes are attached to left hand, right hand and a reference to right mastoid bone.

B. Data Collection

Subjects were 10 healthy individuals (25-35 years, 6 men and 4 women). Subjects were asked to relax, and 10 seconds of ECG data were captured for each subject per trial. Each subject repeated 10 trials at different time-of-day to incorporate ECG variability. The rationale of selecting similar age-group healthy subjects is to study the identification performance with similar ECG patterns.

C. MATLAB Data Processing

MATLAB is used to receive, and analyze the data, and display the results. A simulated UART port is used to communicate with the Bluetooth module of NeuroMonitor in SPP profile at a baud rate of 115.2 kbps. The channel data is then converted to mV using Expression (1).

$$\text{Lead I} = ((\text{ChannelData}) \times V_{range}) / (2^{16} \times A_v) \quad (1)$$

where $V_{range} = 3.3 \times 10^3$. Lead I ECG is filtered in MATLAB using an IIR (Infinite Impulse Response) notch filter (f_c = 60 Hz) with a Q factor of 1 to reduce utility line noise. Then, a Parks-McClellan optimal 3rd order FIR (Finite Impulse Response) low-pass filter (f_c = 40 Hz) is applied. The processed ECG data is utilized for further analysis. The MATLAB processing flow diagram is depicted in Fig. 2.

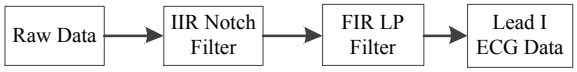

Figure 2. MATLAB preprocessing of Lead I ECG

D. Template and Fiducial Point Extraction

Various features have been suggested in the literature as the inputs to the classifier (such as [7]). In this work, 15 datasets were used to find the optimal performance with minimum complexity. To obtain 15 datasets, interval and amplitude information are extracted from the 10 seconds of ECG data. At first, the ECG signal is normalized using Expression (2).

$$\text{Normalized ECG} = (\text{I} - \min(\text{I})) / (\max(\text{I}) - \min(\text{I})) \quad (2)$$

where, I is the Lead I ECG data. These 10 seconds of normalized ECG data are given as an input to a window based maximum point detector. This detects the peaks of 10 seconds of ECG data (i.e. R points). Data is then cropped from one R peak ± 300 points (1.17 second). The cropped ECG data is used to find the R-R and S-S intervals for feature extraction. To get one template for each trial of an individual, the signal is cropped further to 150 points (0.58 second) using the same technique. Thus 10 templates for 10 trials of each person are obtained. Template extraction is shown graphically in Fig. 3 by steps from the Lead I ECG to a 0.58 second ECG signal.

Template data along with the features are used to train the classifier. In this work only 12 features are used, which are listed in Table I. With this set of features we achieved 100% classification accuracy.

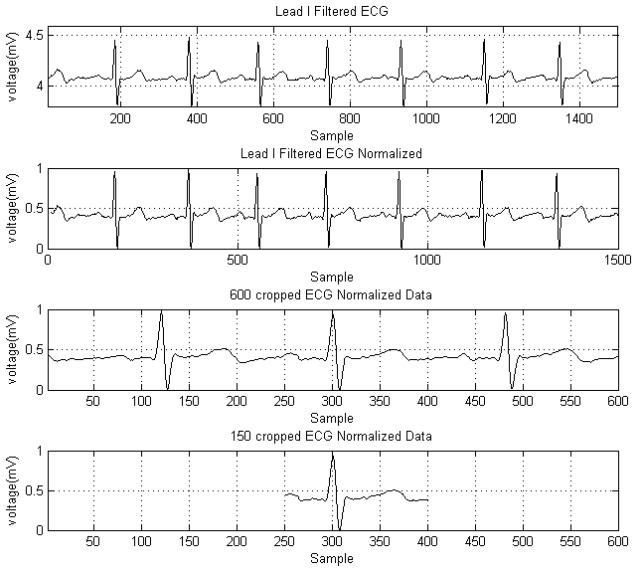

Figure 3. Template extraction steps from ECG data.

TABLE I. LIST OF FEATURES FOR CLASSIFICATION

Feature Number	Feature Description
1	R peak voltage
2	R-R interval
3	S peak voltage
4	S-R interval
5	P peak voltage
6	T peak voltage
7	P-S interval
8	R-T interval
9	PS voltage
10	TS voltage
11	RP voltage
12	TR amplitude

To obtain these features, a two-step procedure is used. In the first step, the R-R interval, the S-S interval and their amplitudes are found from 2.34 seconds of ECG data. To extract these fiducial points, a maximum/minimum window technique is used. For each individual, the maximum and the minimum are detected using sliding windowing technique throughout the ECG interval. The window size is based on the heart rate of the individual. When R and S peaks are needed, a larger window size is used. A smaller sized window is used to find the P and T peaks from one template of 0.58 second ECG data. Finally, the features are obtained from the amplitude and interval fiducial points. This two staged feature extraction technique is shown in Fig. 4.

E. Classification

ECG data is highly correlated from one subject to another. Linear classifiers might not achieve the best performance with a small training and test data set. The rationale for selecting a logistic regression (LR) classifier is that it uses the nonlinear logistic function to classify the dataset. In this work, 80% of the data is used to train the classifier and 20%

Preprint submitted to 36th Annual International IEEE EMBS Conference.
Received March 15, 2014.

to identify. Using those training data, an optimized classifier model is built from 100 iterations of the LR classifier. Then test data are applied to find the model accuracy. For each individual, 10 templates from 10 trials, and 12 features from each trial are used. The block diagram of this procedure is outlined in Fig. 5. Template data with two types of features are used as inputs to the classifier: amplitude features and interval features.

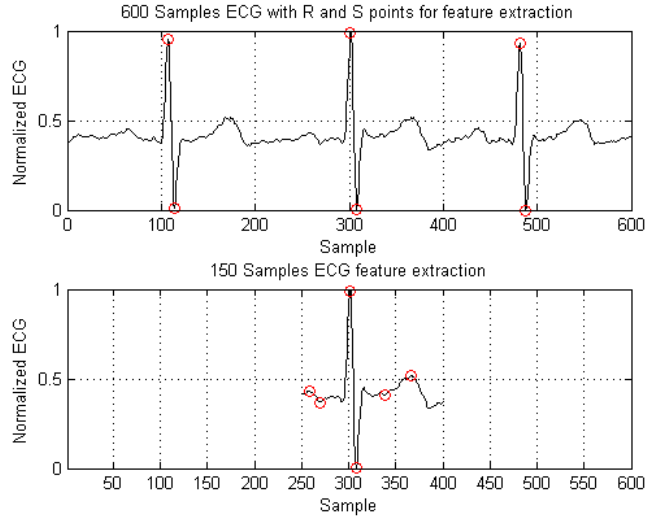

Figure 4. Fiducial point extraction in two stages.

For classification, 15 sets of data are used. Template data of each trial for each person is down-sampled by factors of 5, 10, 15, and 25. These sampled data are used independently and also with the combination of features for classification.

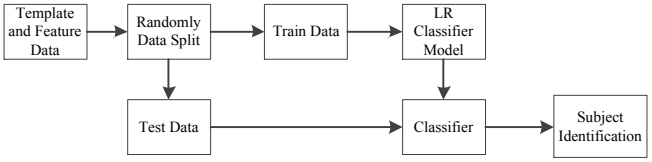

Figure 5. Classification block diagram

III. Results

The results of various stages of data preprocessing are given in Fig. 6. Filter responses are shown as frequency spectrum, while the filtered ECG data is presented in time domain. After getting the templates using above method, 15 sets of data and their combinations are supplied as inputs to the classifier. To test the average accuracy, classification is performed for 100 times using each dataset. Different combinations are made by down sampling the template. For example, if a down sampling factor is 25, the template size is reduced from 150 to 6. Only the template, features and combination of features and template are used to calculate mean, standard deviation and highest classification accuracy. Two types of features (interval and amplitude) are used with and without the template. A summary of the results is presented in Table II. It can be observed that for the interval feature, the accuracy is greater than the amplitude feature because of the larger variance. A full template with features achieves the best performance in terms of the mean, standard deviation (STD), and accuracy. The accuracy using template data with and without features for 100 different classifiers are shown in Fig. 7.

CPU execution time for training and testing using this method was also evaluated. An Intel Core i7 2.93GHz CPU with 8 GB memories was used to run MATLAB for data processing. The timings for 3 cases are shown in Table III.

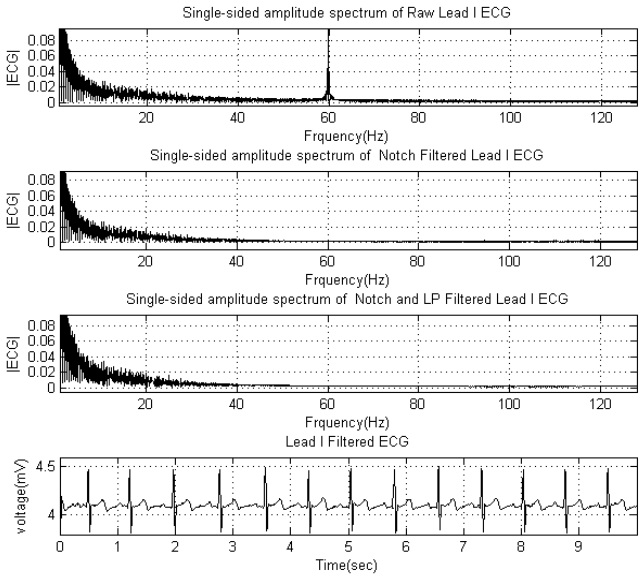

Figure 6. ECG data during different stages of processing

TABLE II. CLASSIFICATION RESULTS FOR DIFFERENT DATASET

Item	Feature	Mean and STD of accuracy for 100 trials	Highest Accuracy
1.	Template only with DSF=25	52.3 ± 5.05	65
2.	Only features	71.2 ± 7.94	95
3.	Combining 1 and 2	77.85 ± 8.23	100
4.	Template only with DSF=15	81.1 ± 5.79	95
5.	Combining 4 and 2	84.55 ± 6.96	100
6.	Template only with DSF=10	82.55 ± 7.01	95
7.	Combining 6 and 2	88.6 ± 6.32	100
8.	Template only with DSF=5	89.6 ± 5.35	100
9.	Combining 8 and 2	91.95 ± 5.5	100
10.	Template data only	93.2 ± 5.05	100
11.	Combining 10 and 2	97.9 ± 3.35	100
12.	Only interval features	53.3 ± 9.02	85
13.	Combining 10 and 12	94.7 ± 4.43	100
14.	Only amplitude features	48.85 ± 7.03	65
15.	Combining 14 and 10	95.45 ± 4.32	100

To reduce the data dimensionality, principal component analysis (PCA) is used. Only the 10 largest variance principal components are chosen. Training data are only used to find the PCA projection matrix. Test data are projected using that matrix. Projected test data are then used to test the accuracy of the classifier. Principal components (PC) 1 vs 2 are plotted in Fig. 8 for the 10 subjects.

TABLE III. CLASSIFICATION TRAINING AND TEST TIMING

CPU Execution Time	150 Template Data Only	12 Features Only	Combination of Template and Features
Training Time (sec)	2.5371	2.045	2.497
Testing Time (sec)	12×10^{-4}	2.29×10^{-4}	2.43×10^{-4}

Preprint submitted to 36th Annual International IEEE EMBS Conference.
Received March 15, 2014.

Using PCA, the data dimensionality is reduced from 162 features to 10 per trial. 800 features are used to train the classifier and 200 to find the accuracy. The highest achieved accuracy after data dimension reduction is also 100%. Training time of the classifier is 2.1806 seconds and it takes 1.2ms to identify 10 subjects from 200 data. Classification is run for 100 times, a mean and standard deviation accuracy of 90.85 ± 5.32 is obtained. The accuracy plot for 100 different classifiers is shown in Fig. 9.

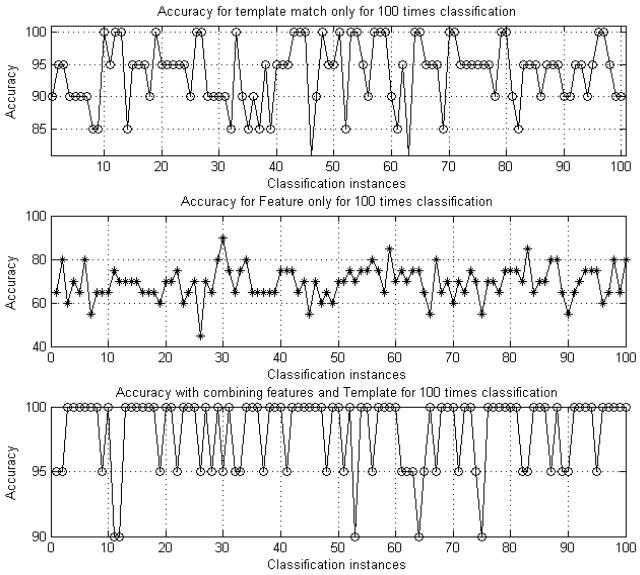

Figure 7. 100 different classifier accuracy results

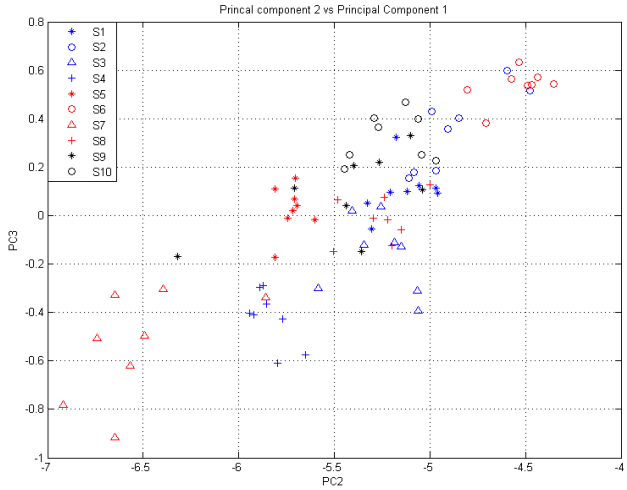

Figure 8. PC2 vs PC1 graph for 10 subjects

IV. Conclusion

This work shows a new method to identify a particular human subject using only 0.58 second Lead I ECG data. It can be implemented in a small, embedded hardware platform (such as NeuroMonitor) due to its use of the Lead I configuration. Data dimensionality is greatly reduced to train the classifier. It requires only 80 features per subject to train the classifier. A simplified technique for fiducial point detection is also implemented to extract features in a hardware-efficient manner rather than using a complex algorithm. The future directions of this research include implementation of the algorithm in the ambulatory NeuroMonitor hardware for real-time classification.

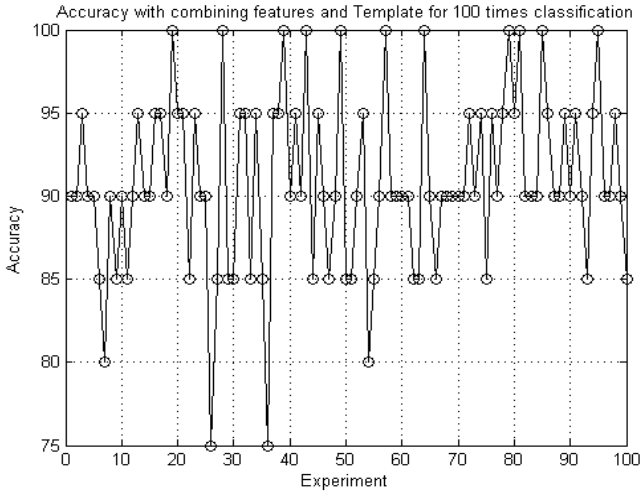

Figure 9. 100 times classification accuracy results after PCA

Acknowledgment

Authors would like to thank Jayanta Kumar Dutta for some critical comments and suggestions about this work.

References

[1] J. D. Woodward, N. M. Orlans, and P. T. Higgins, *Biometrics:[identity assurance in the information age]*. McGraw-Hill/Osborne New York, 2003.

[2] D. Hawkins, Body of evidence, *US News and World Report*, pp. 60-62, 18 February 2002.

[3] S. A. Israel, *et al.*, "ECG to identify individuals," *Pattern recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[4] Irvine, J. M., et al. "Heart rate variability: a new biometric for human identification." *Proceedings of the International Conference on Artificial Intelligence (IC-AI'01)*. 2001.

[5] J. Irvine, *et al.*, "A new biometric: human identification from circulatory function," *in Joint Statistical Meetings of the American Statistical Association,* San Francisco, 2003.

[6] S. A. Israel, *et al.*, *The heartbeat: the living biometric*. Wiley-IEEE Press, New York, NY, USA, 2009.

[7] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: a new approach in human identification," *IEEE Transactions on Instrumentation and Measurement,* vol. 50, no. 3, pp. 808–812, 2001.

[8] Shen, T. W., W. J. Tompkins, and Y. H. Hu. "One-lead ECG for identity verification." *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*. Vol. 1. IEEE, 2002.

[9] Kyoso, Masaki, and Akihiko Uchiyama. "Development of an ECG identification system." *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*. Vol. 4. IEEE, 2001.

[10] T. Shen, "Quartile discriminant measurement (qdm) method for ECG biometric feature selection," *in Proceedings of International Symposium of Biomedical Engineering*, Taiwan, no. 10394, 2006.

[11] G. Wübbeler, *et al.*, "Verification of humans using the electrocardiogram," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1172–1175, 2007.

[12] Sahadat, Md Nazmus, S. Consul-Pacareu, and Bashir I. Morshed. "Wireless ambulatory ECG signal capture for HRV and cognitive load study using the NeuroMonitor platform." *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. IEEE, 2013.

[13] Mahajan, Ruhi, et al. "Ambulatory EEG NeuroMonitor platform for engagement studies of children with development delays." *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2013.