Hardware-Efficient Robust Biometric Identification from 0.58 Second Template and 12 Features of 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 with 12 features using logistic regression classifier that achieved high reliability and identification 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. Hardware-efficient fiducial point detection technique was implemented for feature extraction. To obtain repeated random sub-sampling validation, data were randomly separated into training and testing sets at a ratio of 80:20. Test data were used to find the classification accuracy. 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 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 is 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 were captured from subjects using our custom-built ambulatory EEG/ECG embedded device (NeuroMonitor) [12, 13]. Captured data were 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.

A. Device

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

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).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

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 was chosen to capture the data from each individual. Three electrodes were attached to left hand, right hand and a reference to right mastoid bone.

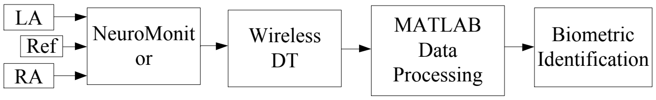

Figure 1. Block diagram of the overall system

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).

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

where $V_{range} = 3.3 \times 10^3$ and $A_v = 93.28$. 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.

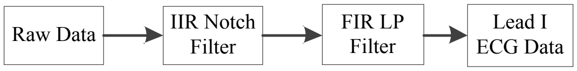

Figure 2. MATLAB data preprocessing steps of Lead I ECG signal

The results of various stages of data preprocessing are shown graphically in Fig. 3. Filter responses are shown as frequency spectrum, while the filtered ECG data are presented in time domain.

D. Template and Fiducial Point Extraction

ECG data have a tendency to be corrupted by different artifacts, noise and other interferences. Same subject had different DC offset ECG signal for different trial because of muscle artifact at different time. So filtered ECG signal was first normalized using Expression (2).

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

where, I is the Lead I ECG data. These 10 seconds of normalized ECG data were given as an input to a window based maximum point detector. It detected the R peaks of the 10 seconds of ECG data. Data was then cropped from R peak ± 1.17 second (300 points).

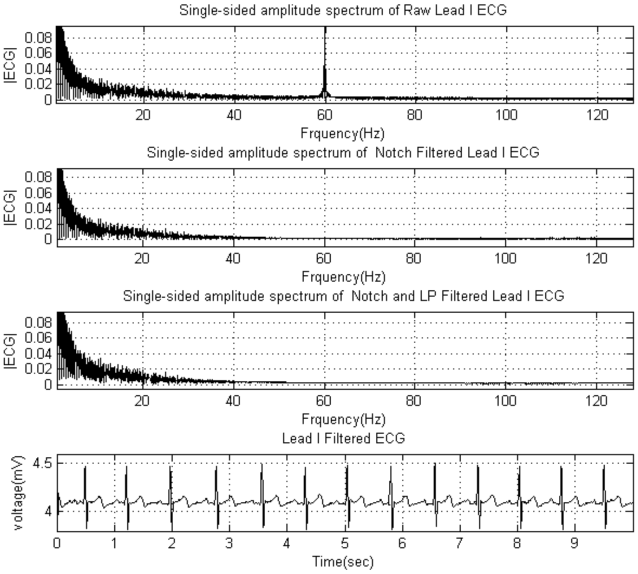

Figure 3. ECG data during different stages of processing

Here any R peak (but the first and last) can be selected within the 10 s ECG data per trial. The cropped 2.34 seconds (600 sample points) ECG data were used to find the R and S fiducial points. R peak voltage, S peak voltage, R-R interval, S-S interval features were detected from those fiducial points. Data were further cropped to 0.58 seconds (150 sample points). This 0.58 second ECGs were used as a template per trial. The template detection steps are shown graphically in Fig. 4. ECG templates (0.58 second signal) were given as an input to sliding window based maximum-minimum point detection algorithm to find P and T fiducial points. P peak voltage and T peak voltage features were found from these fiducial points. P-S interval, R-T interval, PS voltage, TS voltage, RP voltage and TR voltages were computed by combining all fiducial point information.

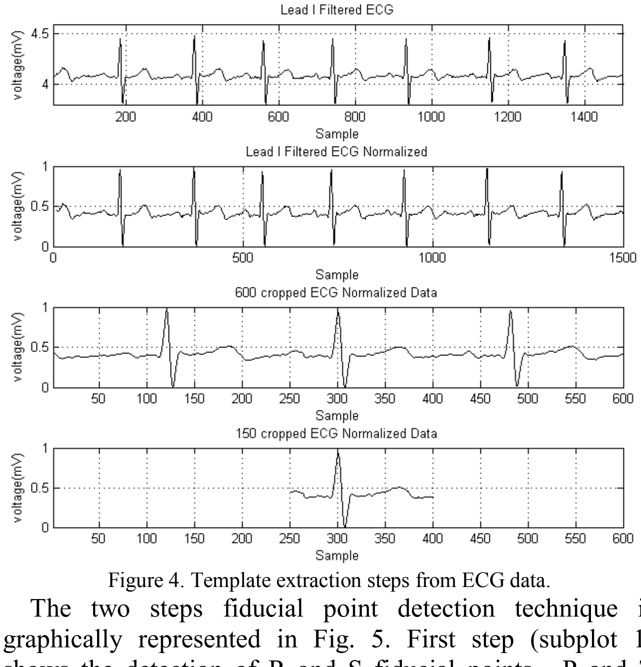

Figure 4. Template extraction steps from ECG data.

The two steps fiducial point detection technique is graphically represented in Fig. 5. First step (subplot 1) shows the detection of R and S fiducial points. P and T fiducial point detection are shown in second step (subplot 2). The different extracted features are summarized in Table I.

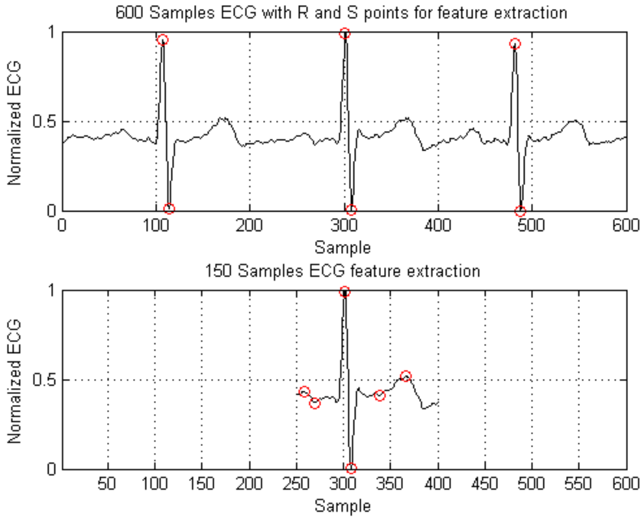

Figure 5. Fiducial point extraction is shown in two stages.

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

E. Classification

Filtered ECG data were found to be highly correlated from one subject to another (correlation coefficient = 0.90 ~ 0.94). Linear classifiers might not achieve the best performance with a small training and test data set. But Logistic regression (LR) classifier which uses the nonlinear logistic function to classify the dataset might provide better results. In this work, 80% of the data are used to train the classifier and 20% to identify. Using those training data, an optimized classifier model was built. Then test data were applied to find the accuracy of the classification. The block diagram of this procedure is outlined in Fig. 5.

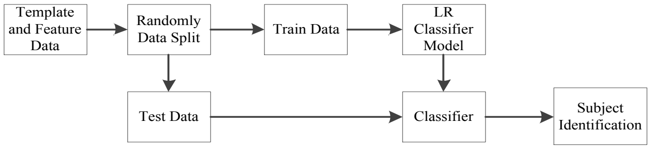

Figure 6. Classification block diagram

For each individual, 10 templates from 10 trials, and 12 features from each trial were available as the input to the classifier. Different combinations of dataset were made to find the best dataset for high accuracy classification. Template data of each trial for each person were down-sampled by the factors of 5, 10, 15, and 25. For example, if a down sampling factor is 25, the template size is reduced from 150 samples to 6. Feature data were separated into amplitude and interval features. By combining different down sampled template data and features, 15 datasets were obtained as shown in Table II. Each type of datasets was separated randomly into train and test data to build classifier model and test the accuracy of the classification.

III. RESULTS

Classification is performed for 100 times using each dataset to obtain repeated random sub-sampling validation. Only the template, features and combination of features and template were used to calculate mean, standard deviation and highest accuracy for 100 classifications. Summary of the results are presented in Table II. It can be observed that for the interval feature, the accuracy is greater than the amplitude feature because interval features contain exclusive information for different individuals. Template (0.58 second ECG data) with features achieved the best performance in terms of the mean, standard deviation (STD), and accuracy of the classification. The accuracies of the 100 classifications using template data, features and combinations of them are shown in Fig. 7. Template data with 12 features obtained an average accuracy of 97.9 ± 3.35 which means, this algorithm can identify an average of 19.5 (~20) subjects correctly from 20 individuals for 100 different classifications.

TABLE II. CLASSIFICATION RESULTS FOR DATASETS

Dataset	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

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

Data dimension reduction is necessary if we want to use this technique for large population. To reduce the data dimensionality, principal component analysis (PCA) was used. Only the 10 largest variance principal components were chosen as an input to the classifier. Training data were only used to find the PCA projection matrix. Test data were projected using that matrix. Projected test data were used to test the accuracy of the classifier. Fig. 8 shows the principal components (PC) 1 vs. 2 for the 10 subjects.

Using PCA, the data are reduced from 162 to 10 features 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% here. Training time of the classifier is 2.1806 seconds and it takes 1.2ms to identify 10 subjects from 200 data. Classification was run for 100 times to obtain the repeated random sub-sampling validation. Mean and standard deviation accuracy of 90.85 ± 5.32 is obtained. The accuracy plot for 100 different classifiers is shown in Fig. 9.

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}

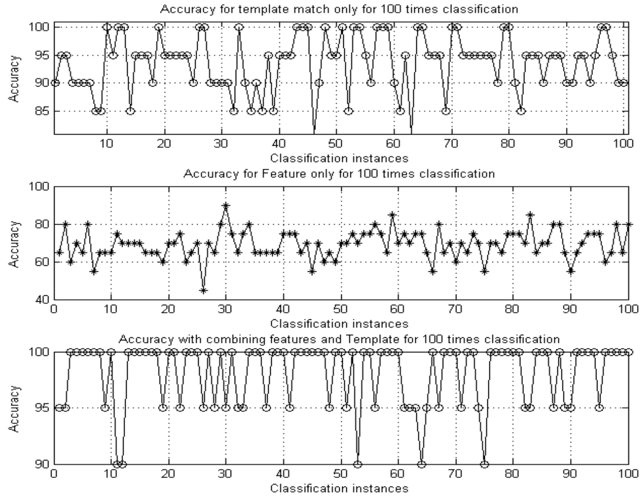

Figure 7. 100 different classifier accuracy results

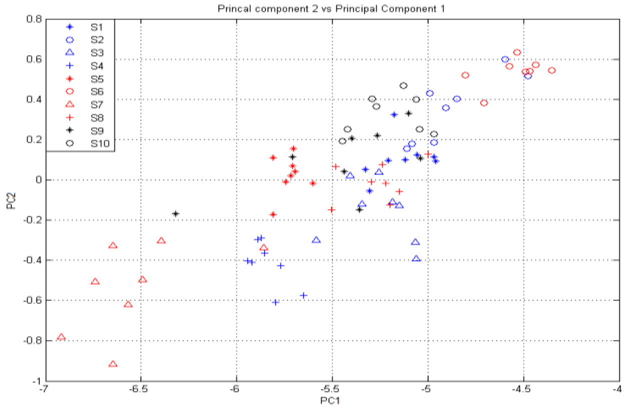

Figure 8. PC2 vs PC1 graph for 10 subjects

IV. Conclusion

This work shows a new method to identify a human subject using 0.58 s template excerpted from 2.34 s 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 fiducial point detection technique is also implemented to extract the features in 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.

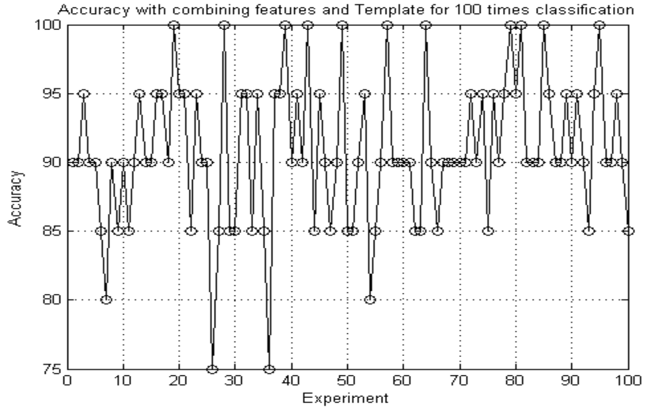

Figure 9. 100 times classification accuracy results after PCA

Acknowledgment

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

References

[1] J. D. Woodward, *et al.*, *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, *et al.*, “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., *et al.*, "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, *et al.*,"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.