PLOS ONE

Normalizing Electrocardiograms of Both Healthy Persons and Cardiovascular Disease Patients for Biometric Authentication

Meixue Yang[1], Bin Liu[3], Miaomiao Zhao[1], Fan Li[2], Guoqing Wang[2], Fengfeng Zhou[1]*

1 Shenzhen Institutes of Advanced Technology, and Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong, P.R. China, 2 Key Laboratory of Zoonosis, Ministry of Education, Norman Bethune College of Medicine, Jilin University, Changchun, Jilin, P.R. China, 3 First Hospital of Jilin University, Changchun, Jilin, P.R. China

Abstract

Although electrocardiogram (ECG) fluctuates over time and physical activity, some of its intrinsic measurements serve well as biometric features. Considering its constant availability and difficulty in being faked, the ECG signal is becoming a promising factor for biometric authentication. The majority of the currently available algorithms only work well on healthy participants. A novel normalization and interpolation algorithm is proposed to convert an ECG signal into multiple template cycles, which are comparable between any two ECGs, no matter the sampling rates or health status. The overall accuracies reach 100% and 90.11% for healthy participants and cardiovascular disease (CVD) patients, respectively.

Introduction

Biometric features have been widely used in multiple identity authentication areas, *e.g.* access control system and communication authentication, etc [1–5]. The biometric features used for authentication include fingerprint, face and voice [6]. It was demonstrated that the integration of multiple biometric features could significantly increase the authentication accuracy [7,8]. However, these individualized features have the possibility to be duplicated or faked to bypass the authentication system. So other more secure individualized features are sought for the biometric authentication problem. The electrocardiogram (ECG) signal is individual-specific in the sense of amplitude, peak and other characteristics, and difficult to be faked [9,10]. ECG also represents a versatile factor that can derive a number of personal health measurements, *e.g.* dynamic heart rate variation, and acts as the diagnostic basis of various cardiovascular diseases (CVDs), *e.g.* arrhythmia [11,12]. So it is becoming one of the major new features for the biometric authentication problem [1,6,9,10,13].

The major problem of ECG-based biometric applications is the dynamic temporal changes of ECG signals due to the physiological activities of the human subject. The majority of currently available algorithms extract measurements between the peaks and valleys of ECG signals. Generally, ECG based biometric algorithms can be categorized as fiducial point dependent or independent, according to whether they detect the ECG sensor's physical position on the subject body [14,15]. The ECG-based biometric authentication (EBA) problem for healthy persons can be satisfyingly solved with

accuracy$> = 95\%$, but a much lower accuracy may be achieved for CVD patients. A discrete wavelet transformation algorithm achieved accuracies of 100% and 81% on 35 healthy persons and 10 arrhythmia samples, respectively [16]. The discrete cosine transformation based algorithm achieved accuracies of 84.61% and 100% on 13 healthy samples on the PTB and MIT-BIH databases, respectively [10]. Another recent study applied the cross correlation algorithm on the EBA problem, and achieved 80%, 70% and 80% in accuracies on the chosen 10 arrhythmia samples from the databases AFPDB, SVDB and TWADB, respectively [17]. Furthermore, most of the current algorithms only work well when the training and testing data were detected by the same experiment [18].

With the increased algorithm accuracy of the EBA problem, this technique is becoming more of interest to industries in biomedical engineering and mobile personal identification. Min *et al* filed patents KR2006082677-A and KR750662-B1 in 2007 to recognize different persons based on the limb lead III ECG signal [19]. Sun *et al* proposed a biometrics information system to store and measure biomedical signals including electrocardiograms in patent KR2011099197-A in 2011 [20]. And Apple Inc recently also filed a patent (20100113950) for an ECG-based biometrics application [21].

We propose in this study a novel normalization and interpolation algorithm to transform the ECG signal of any person collected at any time into a comparable template ECG cycle (TEC). The algorithm features a high similarity among the TECs of the same person, either healthy or with cardiovascular disease,
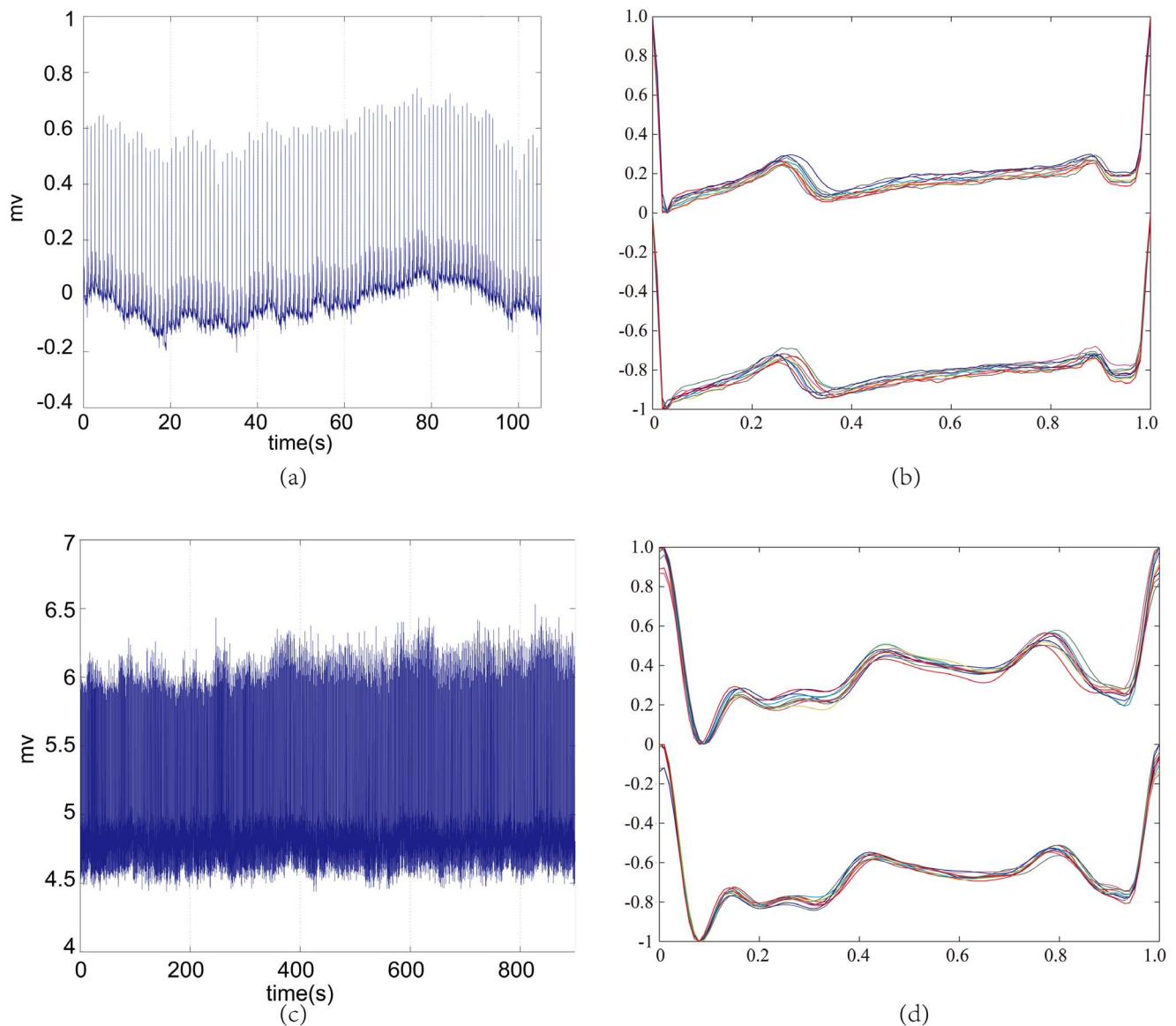
Figure 1. The ECG curves of the samples s0306lre and sel100 m, respectively. (a) The original curve and (b) the 10 representing training and testing ECG cycles of s0306lre from the PTB database. And the same datasets (c) and (d) for sel100 m from the QT database. For (a) and (c), the horizontal axis is in seconds, and the vertical axis is in mV.
doi:10.1371/journal.pone.0071523.g001

but a much lower similarity between two TECs of different persons. This study also represents one of the most comprehensive investigations of the EBA (ECG-based biometric authentication) problem. A real-world EBA problem with no prior knowledge of sample health status was also investigated in a mixture of healthy persons and CVD patients. A consistently high accuracy suggests that our algorithm could greatly facilitate the EBA model in clinical cases. Finally we discussed two potential future directions for improving our algorithm.

The major contributions of this study include a description of the first biometric algorithm based on direct comparison of the ECG curve itself, and the consistently high accuracies of this algorithm on different data sources, which is a major problem faced by the other algorithms.

Materials and Methods

Data sources and preprocessing

The ECG data of 52 healthy persons were downloaded from the PTB database in the PhysioNet Diagnostic ECG Database [22]. Abnormal ECG data of 105 persons (including 14 healthy persons and 91 patients with cardiovascular diseases (CVDs)) were downloaded from the QT database in PhysioNet [23]. Database QT mainly consists of ECG signals with varied QT or ST intervals, including arrhythmia [23,24]. Both datasets were downloaded on March 5, 2012. The list of sample IDs used in this study can be found in Table S1. The ECG signals were de-noised using a 0.5–45 Hz Butterworth Band-Pass (BBP) filter [25].

Given an ECG signal E, we detect the R waves using the So-Chan method [13], and split E into ECG cycles between two neighboring R waves. Although some studies generated the ECG cycles with fixed-length flanking regions of R waves, it's possible to
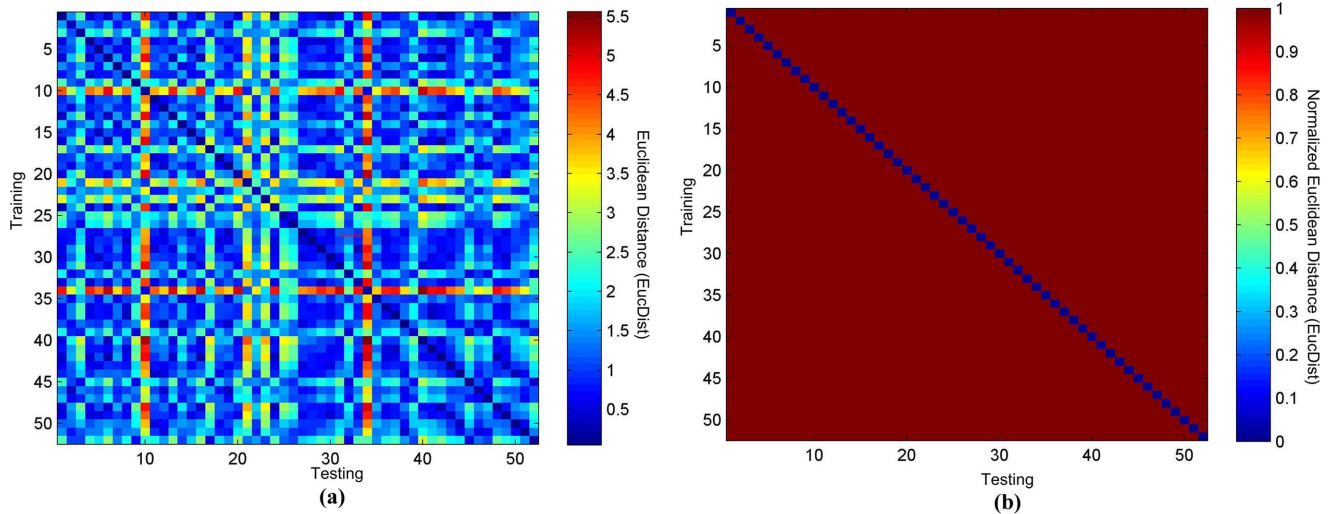
Figure 2. The heatmaps of TEC matching of 52 healthy persons. (a) The original heatmap and (b) the binary heatmap.
doi:10.1371/journal.pone.0071523.g002

include multiple neighboring R waves in one cycle, due to the varying heart rates. Considering the high accuracy in detecting R waves, this study considers the region between two R peaks as one ECG cycle.

ECG cycle interpolation (ECI) algorithm

Two normalized ECG cycles are not comparable due to the following two reasons. Firstly, the two cycles may be detected using different sampling frequencies, which make them have different numbers of data points. Secondly, even with the same sampling frequency or within one ECG signal, the two cycles may have different lengths due to various reasons, e.g. running or arrhythmia. This also makes the two cycles have different numbers of data points.

We interpolate the normalized ECG cycle by fitting it with a cubic spline curve between any pair of neighboring data points , as similar in [9]. Given the normalized ECG cycle with $X = [x_0, x_1, \cdots, x_n](x_0 = 0, and \quad x_n = 1)$ and $Y = [y_0, y_1, \cdots, y_n])$, the cubic spline curve $S(x)$ is defined as:

$$S(x) = \begin{cases} S_0(x), & x_0 \leq x \leq x_1, \\ S_1(x), & x_1 \leq x \leq x_2, \\ \vdots & \vdots \\ S_{n-1}(x), & x_{n-1} \leq x \leq x_n, \end{cases} \quad (1)$$

where $S_i(x) = A_i x^3 + B_i x^2 + C_i x + D_i (i \in [0:1:n])$, $S(x_i) = y_i$ $(i \in [0:1:n])$, $\lim_{x \to x_i^-} S(x) = \lim_{x \to x_i^+} S(x)$, $\lim_{x \to x_i^-} S'(x) = \lim_{x \to x_i^+} S'(x)$, and $\lim_{x \to x_i^-} S''(x) = \lim_{x \to x_i^+} S''(x)$ where $i \in [1:1:n-1]$.

The interpolation interval is $X = [X_0, X_1, \cdots, X_n] = [0:0.01:1]$, and the interpolated ECG cycle is $Y_i = S(X_i)$, where $X_i \in [0:0.01:1]$. In this study, the difference between two ECG cycles is measured by the Euclidean Distance (*EucDist*) [26] of their normalized and interpolated vectors. The Euclidean Distance between two vectors $A = [a_0, a_1, \cdots, a_m])$ and $B = [b_0, b_1, \cdots, b_m])$ is defined to be:
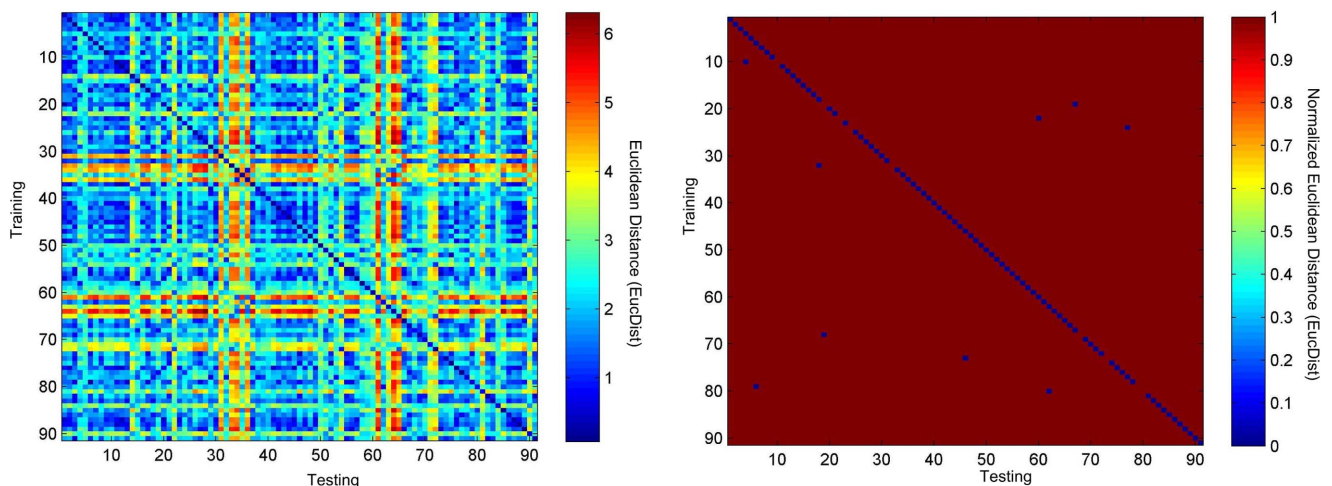
Figure 3. The heatmaps of TEC matching of 91 CVD patients. (a) The original heatmap and (b) the binary heatmap.
doi:10.1371/journal.pone.0071523.g003
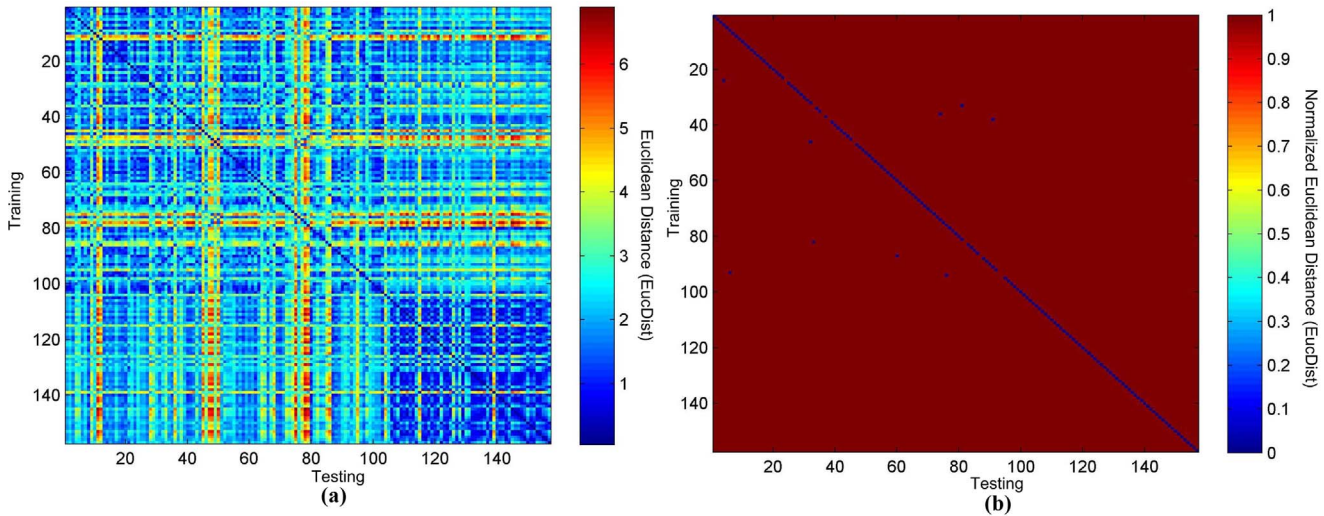
Figure 4. The heatmaps of TEC matching of 52+14 = 66 healthy persons and 91 CVD patients. (a) The original heatmap and (b) the binary heatmap.
doi:10.1371/journal.pone.0071523.g004

$$EucDist(A,B) = \sqrt{(a_0 - b_0)^2 + (a_1 - b_1)^2 + \cdots + (a_m - b_m)^2},$$

where the smaller the *EucDist* between two ECG cycles is, the more similar the two ECG cycles are to each other.

ECG cycle normalization (ECN) algorithm

Given an ECG cycle $EC = [v_0, v_1, \cdots, v_n]$) with the sampling times $T = [t_0, t_1, \cdots, t_n]$), let the minimum and maximum voltages (in mV) of $v_i (i \in [0 : 1 : n])$ be v_{min} and v_{max}, respectively. EC is linearly scaled by the following formula:

$$y_i = \frac{n-i}{n} \times \frac{v_i - v_{min}}{v_0 - v_{min}} + \frac{i}{n} \times \frac{v_i - v_{min}}{v_0 - v_{min}}, \quad where \quad i \in [0 : 1 : n] \quad (2)$$

The sampling times are also linearly scaled to be between 0 and 1, as follows:

$$x_i = \frac{t_i - t_0}{t_n - t_0}, \quad where \quad i \in [0 : 1 : n] \quad (3)$$

ECG cycle normalization and interpolation algorithm (ECOIL)

ECG **C**ycle **N**ormalization and **I**nterpolation **A**lgorithm (ECOIL)

Input: an ECG signal E

Output: an ECG cycle M with its data points sampled at $[0 : 0.01 : 1]$ and the curve within $[0,1] \times [0,1]$.

Procedure:

1. Denoise E with the BBP filter;
2. Detect the R waves in E;
3. Split E into ECG cycles between neighboring R waves;
4. Interpolate the ECG cycles using the algorithm ECI;
5. Normalize the ECG cycles using the algorithm ECN;

6. Cluster the cycles into two clusters based on the pair-wise Euclidean distances, using the k-means clustering algorithm;
7. Randomly choose 10 cycles $C_j (j \in [0 : 1 : 10])$ from the larger cluster;
8. For each $X_i \in [0 : 0.01 : 1]$, calculate the medium value $M(X_i) = Medium(C_j(X_i))$, where $X_i \in [0 : 0.01 : 1]$ and $j \in [0 : 1 : 10]$.
9. Output the ECG cycle $M = M(X_i)$, where $X_i \in [0 : 0.01 : 1]$.

We derive a template **E**CG **c**ycle (*TEC*) from an ECG signal by the above procedure. The random choice of 10 representative TECs is based on the observation that the majority of the generated TECs are highly individual-specific, as shown in Figure 1 (b) and (d). The classification performance of the 10-TEC based biometric algorithm also suggests that the choice of a limited number of TECs is both consistently accurate and calculation efficient, as shown in Figures 2, 3 and 4.

Performance evaluation strategies

We investigated the ECOIL algorithm on solving the biometric authentication problem, which seeks a differentiating measurement of two biometric data, so that the data from the same person are more similar to each other than to those from different ones. We measured the difference between two ECG signals by the Euclidean Distance (*EucDist*) of their TECs. The function *EucDist* is defined in the above section of *ECG Cycle Interpolation (ECI) algorithm*. The larger the $EucDist(TEC_1, TEC_2)$ is, the less similar the two ECG signals are to each other.

An ECG signal was split into two halves with equal lengths, and the training and testing TEC data of this ECG signal were generated on the first minute of each of the two halves. There were denoted as $fTEC(k)$ and $sTEC(k)$, respectively , where $k \in [1 : 1 : K]$. Some ECG data in the PTB database are shorter than two minutes, and the training and testing TECs were generated from the full first and second halves of the data. Next we built the template library $LibTEC = \{fTEC(k)\}$, where $k \in [1 : 1 : K]$. The identity of a testing data $sTEC(a)$ is defined as:
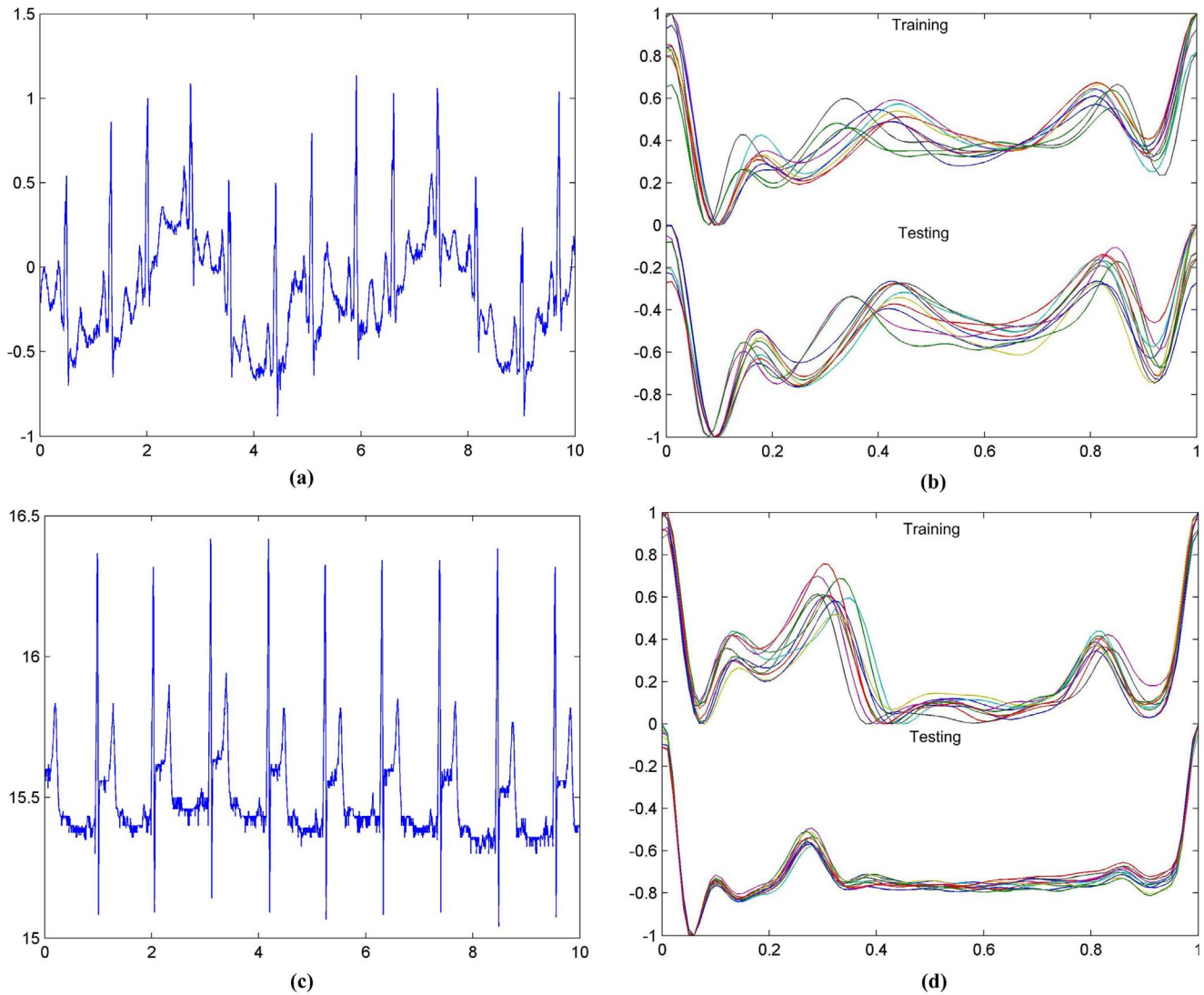
Figure 5. The ECG signal and processed ECG cycles of two samples. (a)–(b) sel39 m, and (c)–(d) sele0136 m. The ECG cycles in the first 10 seconds of the two signals were plotted, to make a clearer curve of the two signals. For (a) and (c), the horizontal axis is in seconds, and the vertical axis is in mV.
doi:10.1371/journal.pone.0071523.g005

$$P(a) = b, \ if \quad EucDist(sTEC(a), \ LbTEC(b))$$
$$= \min_{1 \leq k \leq K} \ EucDist(sTEC(a), \ bTEC(k)) \tag{4}$$

The numbers of both training and testing datasets are K, and the number of persons with correct identity predictions is TP. The testing dataset can be regarded as an independent test set, because the training and testing data of the same persons were collected at different times. The algorithm performance was evaluated by the accuracy $Acc = TP/K$. The ECG-based biometric authentication (EBA) problem consists of the training dataset $LibTEC = \{fTEC(k)\}$, and testing dataset $TestTEC = \{sTEC(k)\}$. A prediction of $sTEC(i) = fTEC(j)$ is correct, only if both TECs come from the same person, *i.e.* the label of host $Person(sTEC(i)) = Person(fTEC(j))$. The label of each person represents a class, and there are 52 and 105 classes for the EBA problem on the PTB and QT databases, respectively. Since

classification measurements, sensitivity and specificity, can only be calculated on a binary classification problem, only recognition accuracy is investigated in this study.

Due to the fact that the ECG-based biometric authentication problem determines which sample in the training dataset a query signal belongs to, we did not train on one dataset, and test on another, with no overlapping samples between the two datasets.

Results and Discussion

Self similarity of the processed ECG cycles

We investigated the self similarity of the ECG cycles before and after the processing of the ECN and ECI algorithms, using the samples s0306lre and sel100 m from the PTB and QT databases, respectively. For sample s0306lre, the voltage of the ECG cycle baselines varies with sampling time, as shown in Figure 1 (a). The 10 representative training ECG cycles after the ECN and ECI algorithms perfectly fit each other, as in the top part of Figure 1 (b). The 10 representative testing ECG cycles fit with each other,
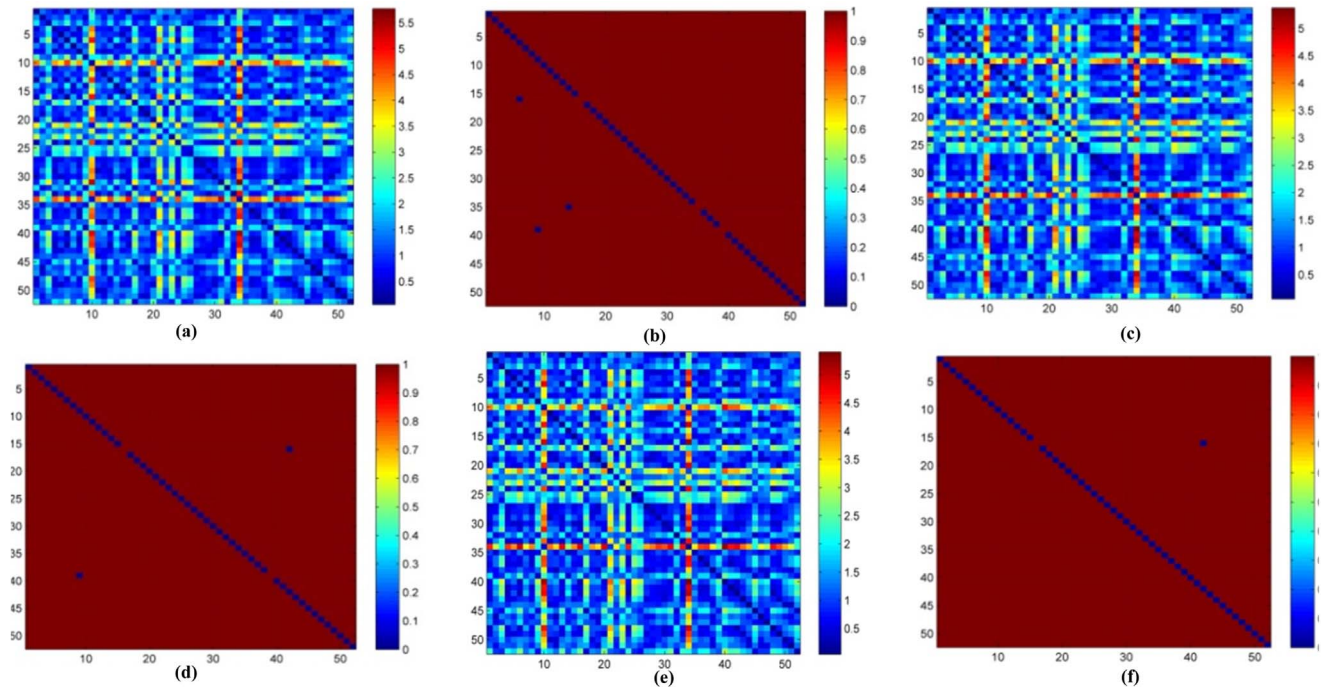
Figure 6. The heatmaps of TEC matching of 52 healthy persons in PTB database. (a) The original heatmap with 5 s and (b) binary heatmap with 5 s. (c) The original heatmap with 10 s and (d) binary heatmap with 10 s. (e) The original heatmap with 20 s and (f) binary heatmap with 20 s.
doi:10.1371/journal.pone.0071523.g006

too, as in the bottom part of Figure 1 (b). And there is also a very high similarity between the 10 training and 10 testing ECG cycles. As for QT sample sel100 m, the voltage of the QRS waves varies with sampling time, but similar consistency can still be observed in the ECG cycles processed by the ECN and ECI algorithms, as shown in Figure 1 (c) and (d). The same pattern holds for all the other samples in the PTB and QT databases.

ECG-based biometric authentication of healthy persons

Firstly, we studied the similarities of the TECs from the testing dataset to the training dataset of 52 healthy persons in the PTB database. As shown in Figure 2, the colors blue and red represent the smallest and largest *EucDist* between two TECs. The heatmap shows a consistently matching TEC from the testing dataset to that of the same person in the training dataset. To make the illustration clearer, we changed all the values but the smallest one in each column to 1, and the smallest value in each column to 0, which makes the heatmap have two colors, called a binary heatmap, as shown in Figure 2 (b). Figure 2 (b) clearly shows the accuracy is 100% for the PTB database using the ECG signals of 52 healthy persons.

ECG-based biometric authentication of CVD patients

We further evaluated the similarities of the TECs from the testing dataset to the training dataset of 91 CVD patients in the QT database. As shown in Figure 3 (a), there is a consistent self similarity between the training and testing TECs of the same patient. The accuracy 90.11% (~82/91) suggests that our algorithm outperforms the others on the CVD patients by at least 10%, as shown in the binary heatmap in Figure 3 (b) [16,17]. The incorrect predictions for some samples could be due to the significant shifting of P and T waves between different cycles of ECGs, as shown in Figure 5 (a)–(d).

Combined performance of both healthy persons and CVD patients

We also produced one dataset from the ECGs of all the 52 healthy persons in the PTB database, and 14 healthy persons and 91 CVD patients in the QT database. Since the two databases, *i.e.* PTB [22] and QT [23], were independently collected and curated, we believe that this dataset of mixed samples well represents a practical ECG-based biometric authentication problem. Next, we investigated the biometric authentication accuracy of the 157 persons, with no prior knowledge about the diseases of the persons for the algorithm. Only the same 9 CVD patients received incorrect predictions, and our algorithm produced a satisfying accuracy of 94.27% (~(157–9)/157). Hence our algorithm works reasonably well on independently collected datasets, and does not produce confused results among different datasets.

Longer signal duration increases authentication accuracy

We further investigated whether a shorter ECG signal duration produces a reasonable biometric authentication accuracy, based on the PTB database. As illustrated in Figure 6, when the ECG signal duration is 5, 10 and 20 seconds, the numbers of incorrect authentication cases are 3, 2 and 1, respectively. Overall authentication accuracies are 94.23%, 96.15% and 98.07%, respectively. The data in Figures 2 and 6 suggests that a longer ECG signal detection time leads to better ECG-based biometric authentication accuracy. Considering that the rest ECG signal is usually taken for a period of 10 seconds, *e.g.* in an office visit during an annual physical examination, the overall accuracy of 96.15% is acceptable.

Conclusions

Our experimental data shows that the TECs of the same persons are highly similar to each other, irrespective of ECG

variations induced by the factors such as health status and the ECG electrode placement. The TECs of different persons also show significant differences to each other, compared to self similarity. This feature merits the TEC a good biometric signal for the purpose of authentication. Our algorithm works very well on both datasets of the same health status, and the real-world dataset with no prior knowledge of the health status, by comprehensive evaluations in this study.

Further exploration of our algorithm in a larger and more genetically diversified population [27,28] will be conducted. The algorithm will also be revised to accommodate the 9 CVD patient samples in Figures 4 and 5 and the 3 healthy individuals in Figure 6 with incorrect predictions in this study.

Supplementary materials

The algorithm source code and the supplementary materials can be found in the Supplements section of http://www.HealthInformaticsLab.org/supp/. The ECG data may be obtained from the MIT PhysioNet database.

Supporting Information

Table S1   Sample IDs used in this study.
(PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: FZ FL. Performed the experiments: MY BL GW. Analyzed the data: MY BL. Contributed reagents/materials/analysis tools: BL GW. Wrote the paper: FZ MY MZ.

References

1. Biel L, Pettersson O, Philipson L, Wide P (2001) ECG analysis: a new approach in human identification. Instrumentation and Measurement, IEEE Transactions on 50: 808–812.
2. Di Crescenzo G, Graveman R, Ge R, Arce G (2005) Approximate Message Authentication and Biometric Entity Authentication Financial Cryptography and Data Security. In: Patrick A, Yung M, editors: Springer Berlin/Heidelberg. 578–578.
3. Sandhu R, Samarati P (1996) Authentication, access control, and audit. ACM Comput Surv 28: 241–243.
4. Dai H, Jiang S, Li Y (2013) Atrial activity extraction from single lead ECG recordings: evaluation of two novel methods. Comput Biol Med 43: 176–183.
5. He C, Fan X, Li Y (2013) Toward ubiquitous healthcare services with a novel efficient cloud platform. IEEE Trans Biomed Eng 60: 230–234.
6. Can Y, Coimbra MT, Kumar BVKV (2010) Investigation of human identification using two-lead Electrocardiogram (ECG) signals. 27–29 Sept. 2010, BTAS: 1–8.
7. Poh N, Korczak J (2001) Hybrid Biometric Person Authentication Using Face and Voice Features Audio- and Video-Based Biometric Person Authentication. In: Bigun J, Smeraldi F, editors: Springer Berlin/Heidelberg. 348–353.
8. Snelick R, Uludag U, Mink A, Indovina M, Jain A (2005) Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 450–455.
9. Sidek K, Sufi F, Khalil I, Al-Shammary D (2010) An efficient method of biometric matching using interpolated ECG data.IECBES: 2010 Nov. 30 2010-Dec. 2 2010. 330–335.
10. Wang Y, Agrafioti F, Hatzinakos D, Plataniotis KN (2008) Analysis of human electrocardiogram for biometric recognition. EURASIP J Adv Signal Process 2008: 19.
11. Ma Q, Luo Y, Guo P, Gao G, Yang M, et al. (2013) Clinical effects of xinmailong therapy in patients with chronic heart failure. Int J Med Sci 10: 624–633.
12. Schechtman VL, Harper RM (1991) Time of night effects on heart rate variation in normal neonates. J Dev Physiol 16: 349–353.
13. So HH, Chan KL (1997) Development of QRS detection method for real-time ambulatory cardiac monitor.EMBS: 30 Oct-2 Nov 1997. vol. 1, 289–292.
14. Sidek KA, Khalil I, Smolen M (2012) ECG biometric recognition in different physiological conditions using robust normalized QRS complexes. CinC:2012 9–12 Sept. 2012. 97–100.
15. Plataniotis KN, Hatzinakos D, Lee JKM (2006) ECG Biometric Recognition Without Fiducial Detection. BSYM:2006 Sept. 19 2006-Aug. 21 2006. 1–6.
16. Chiu C-C, Chuang C-M, Hsu C-Y (2009) Discrete Wavelet Transform Applied on Personal Identity Verification with ECG Signal. International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP) 7: 341–355.
17. Sidek KA, Khalil I (2011) Person identification in irregular cardiac conditions using electrocardiogram signals. IEMBS:2011 Aug. 30 2011-Sept. 3 2011. 3772–3775.
18. Odinaka I, Po-Hsiang L, Kaplan AD, O'Sullivan JA, Sirevaag EJ, et al. (2012) ECG Biometric Recognition: A Comparative Analysis. Information Forensics and Security, IEEE Transactions on 7: 1812–1824.
19. Min CH, Kim TS, Lee JH (2006) Biometrics system and method using electrocardiogram, capable of recognizing users using neural network. INHA IND PARTNERSHIP INST (INHA-Non-standard).
20. Sun HJ, Sun KY(2011)Biometrics information measuring and storing system, has memory unit storing biometrics information, and main controller recording transmission log and number of transmission of biometrics information in log file. Univ Yonsei Ind Academic Coop Found (Uyia).
21. Gloria L, Taido N, Pareet R, Andrew H (2009) Seamlessly Embedded Heart Rate Monitor In: Office UPT, editor. Patent Application Full Text and Image Dataabse. United States of America: Apple Inc. Cupertino, CA.
22. Moody GB, Mark RG, Goldberger AL (2011) PhysioNet: Physiologic signals, time series and related open source software for basic, clinical, and applied research. Conf Proc IEEE Eng Med Biol Soc 2011: 8327–8330.
23. Laguna P, Mark RG, Goldberg A, Moody GB (1997) A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. Computers in Cardiology: 1997 7–10 Sep 1997. 673–676.
24. Taddei A, Distante G, Emdin M, Pisani P, Moody GB, et al. (1992) The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. Eur Heart J 13: 1164–1172.
25. Butterworth S (1930) On the Theory of Filter Amplifiers. Wireless Engineer 7: 536–541.
26. Danielsson P-E (1980) Euclidean distance mapping. Computer Graphics and Image Processing 14: 227–248.
27. de Wit T, Dekker S, Maas A, Breedveld R, Knoch TA, et al. (2010) Tagged mutagenesis by efficient Minos-based germ line transposition. Mol Cell Biol 30: 68–77.
28. Chen Y, Zhou F, Li G, Xu Y (2008) A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4. Genetics 179: 2291–2297.