Finding Similar ECGs in a Large 12-lead ECG Database

Richard E Gregg[1], Sophia H Zhou[2], Saeed Babaeizadeh[1]

[1]Philips Healthcare, Andover, USA
[2]Philips Research North America, Cambridge, USA

Abstract

Automated matching of a patient's 12-lead ECG within a large 12-lead ECG database to find similar ECGs has many potential applications including searching for examples, diagnosing ECG by probability estimation, or confirming patient identification. We created a morphology-similarity matching algorithm and reported the performance in this study.

The study set consisted of 24,262 ECGs from 8,663 subjects. Similar ECGs were found by two methods, exhaustive search of pair-wise template matching and fast query using a k-dimensional tree architecture and a processed version of the ECG signal as feature vector. Two ECGs were similar if they came from the same patient. For each ECG in the study set, 20 nearest neighbors were extracted from the database. Sensitivities were calculated for finding any and all of the ECGs from the same patient in the set of 20 nearest neighbors.

In the exhaustive search, sensitivities were 68% and 37% for finding any and all of the ECGs from the same patient respectively. With the fast query, sensitivities were 48% (any) and 30% (all). Sensitivity for the fast query increased to 56% (any) and 37% (all) when extracting 50 nearest neighbors.

We conclude that a low complexity but fast query can be used to find similar 12-lead ECGs from a large database.

1. Introduction

Fast query of similar ECGs has many potential applications including the following 1) assistance with ECG interpretation by viewing examples, 2) statistical diagnosis from the probabilities given by the matching ECGs and 3) finding previous ECGs of a patient with incorrect identifying information.

The idea of similarity in ECG by morphology has been used mainly within a patient's recording to find similar beats for averaging and to reject or classify ectopic beats (1). Cross correlation or template matching is typically used to decide if a new heart beat matches normal or abnormal templates constructed from previous beats. Similarity of ECG has also been studied as a biometric like finger prints or retina scan (2).

Bousseljot et al. introduced statistical interpretation of ECGs not by statistical rules but by direct estimation of probability of diagnosis from the interpretation of a set of matching ECGs (3). Bousseljot used cross correlation to test each new ECG to every ECG in the 10,000 ECG database. Since cross correlation is a computationally expensive operation and testing for a match against every ECG in the database limits the size of the database, our goal is to find an inexpensive way to match ECGs by waveform morphology to allow fast query with large databases.

2. Methods

2.1. Study population

From a single teaching hospital, patients were selected over a three year period. Patients with multiple 12-lead ECGs were included. Excluded were patients with error in study identification number. The resulting test set consisted of 24,262 ECGs from 8,663 patients. The training set came from the publicly available PhysioNet PTB dataset (4). The PTB dataset consisted of 549 15-lead ECG recordings from 294 subjects.

2.2. Algorithm development

The basis for the similar ECG query is a k-dimensional (KD) tree (5). The first step is to build a tree or set of trees from an initial database. For each new ECG in question, the KD tree is queried for nearest neighbors. Two different methods were employed for feature selection to use in the KD tree. The first method involved selecting common ECG features like QRS axis and QRS duration that had a statistically significant contribution to a linear regression estimate of template match score. The lower the template match score, the better the match because template match involves subtracting the waveform in question from the template. For a perfect match, the difference is zero. The second method for feature selection involved using a

processed version of the actual signal as the feature vector. This second method was far superior. To reduce the feature vector size, several steps were required to reduce the number of points while maintaining the shape information. An average beat was generated from the dominant morphology of each ECG. The average beat was transformed from 12-leads to 3 orthogonal Frank leads (vectorcardiogram or VCG) with the Kors transformation for a 3:8 reduction in samples (6). To reduce the number of samples further while retaining the shape information, the approximation was taken from the 4th level wavelet decomposition (7). The X, Y and Z leads were concatenated to make the final feature vector. Figure *1* below shows an example feature vector where the QT interval is clearly visible in the concatenated X, Y and Z leads.

Since a patient's heart rate may vary among the patient's multiple ECGs, QT interval heart rate correction was applied to normalize the STT region to a heart rate of 60 bpm. The corrected QT interval was calculated according to the Hodges formula (8) and the difference between QT and corrected QT was used as a scale factor to linearly correct the waveform from the end of the QRS beyond the end of the T-wave. QT interval correction was applied in a similar way for the pairwise template match comparison of ECGs except that the correction was based on the heart rate difference between the ECG pair.

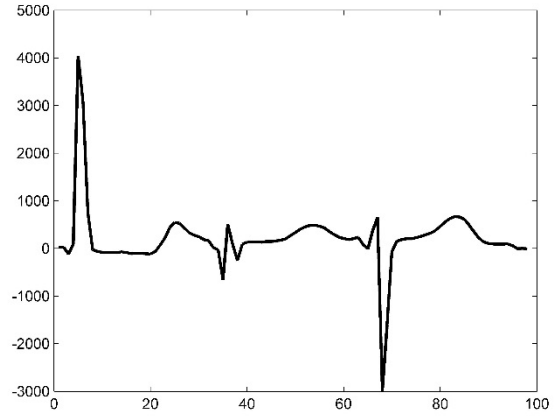

Figure 1 Example Frank lead feature vector from subject s0001 of the PhysioNet PTB database. The X, Y and Z signals are concatenated into one vector. Only the QT region is used.

2.3. Test method

To test the ability of the similar-ECG algorithm to return ECGs that are truly similar in shape, 20 nearest neighbor ECGs were retrieved from the database for each ECG in the test set. The first step was to build the KD tree from the test set. The number 20 was chosen for the number of nearest neighbors because it is an empirical upper limit for a reasonable number of ECGs a user might review when using similar ECGs as helpful examples. The returned set from a query was compared to that patient's list of ECGs to arrive at two sensitivity numbers, a sensitivity to detect any of the patient's other ECGs and a sensitivity to detect all of the patient's other ECGs.

As a performance reference, the same procedure was performed using template matching and exhaustive search. Each ECG was compared to every other ECG to get an N x N matrix of template match scores. For each ECG, the 20 best matches were returned as nearest neighbors.

In addition, the test was repeated 200 times with a random number of patients to see the effect of the number of patients on the sensitivity for finding a patient's other ECGs. The choice of patients was also random for each resampling.

3. Results

Table 1 below shows the sensitivity for detecting the patient's other ECGs when performing similar-ECG queries. The main result is found in the top two rows of the table, the KD tree fast query and the template match exhaustive search. The additional rows show how the sensitivity increases by increasing the number of nearest neighbors in the KD tree query. While the exhaustive search gives the highest sensitivity to find any other ECG, the KD tree method approaches closely by increasing the number of nearest neighbors in the query.

Table 1. Sensitivity (SE) for detecting the subject's other ECGs. N stands for the number of nearest neighbors requested in the similar-ECG query.

Query method	N	SE (%) Any ECG	SE (%) All ECGs
Exhaustive search	20	68	37
KD tree	20	48	28
KD tree	30	52	33
KD tree	40	55	35
KD tree	50	56	37

Figure 2 shows an example of the output from a similarity search of the test database. The example ECG is shown in the left column and a similar ECG from a different patient is shown on the right.

4. Discussion

Several important questions need to be answered in retrieval of similar ECGs from a database. How similar is similar? How to measure ECG similarity? How much computational resource is needed for such an application? Is the application practical? In this paper, we attempt to answer the degree of similarity question not by a

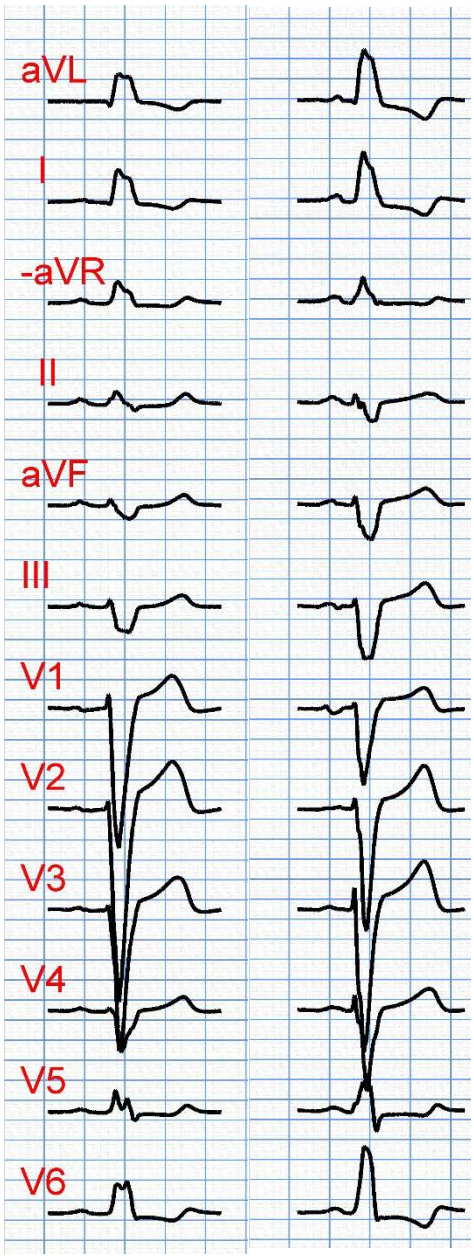

Figure 2. Subject ECG (left) and matching ECG (right) from a different subject. Limb leads are shown in Cabrera order. The ECGs are not exactly the same, just similar.

morphology metric but by an objective measure, that is whether the ECGs belonging to the same patient. We believe this approach is better because it is not an arbitrary choice of shape metric and threshold. In addition, it is more challenge since the ECGs were recorded at different times so there are a variety of shape-based differences due to quality of recording technique and physiological changes. The computational requirement was not tested specifically but elements of the algorithm were designed with focus on low computational alternatives by design, reduced feature vector size, and efficient comparisons due to KD tree architecture.

Template matching of 500 sps data over the entire QT interval was used as the reference for waveform morphology matching of 12-lead ECG in the exhaustive search. Template matching is an expensive operation in CPU cycles. Cross correlation is even more expensive since it involves multiplies in addition to subtraction. KD tree query is quite different however. Muja found the speed-up factor for KD trees compared to linear search was somewhere between 57 and 110 (9). Muja's result is based on a single feature vector comparison. Template match often involves many time shift and feature vector compare operations. The results by tree based query versus exhaustive template match search are close enough to say the fast tree based query is equivalent given the fact that exhaustive search is not practical.

The clean definition of "similar shape" - the patient's other ECGs – comes with a downside. As can be seen in Figure 3, the sensitivity when using this similar shape criterion decreases as the number of patients in the database increases. This does not mean the algorithm gets worse with more patients. It just means that with more patients comes a higher probability of similar shape from other patients. Intuitively, we expect similarity in shape to increase or at least stay the same when the number of patients increases. This is not a problem with the nearest neighbor algorithm, just a problem with the definition of similar shape used in this study. On the other hand, search complexity increases with longer feature vectors. There is a tradeoff between shape information and feature vector length. As we reduce the feature vector size by higher levels of wavelet approximation, we reduce shape information. The fine detail is smoothed out.

The sensitivity for shape match against a reference of same patient's ECG was not as high as expected by the authors. Several factors contributed to the modest sensitivity. Primarily, the ECGs for each patient were not taken at the same time with the same electrode positions. In many cases, a patient's ECGs were spaced by months and even years. In addition, no correction was made for change in health status which can change the ECG morphology drastically. In an analysis of ECG reconstruction from reduced leads, Gregg et al. found that time duration had a small contribution but the remaining contribution to ECG variation was much larger (10). The remaining variation could be due to potentially electrode placement deviation or health status change.

As seen in the performance results of Table 1, increasing the number of nearest neighbors in the KD tree query increases the sensitivity for finding a patient's other ECGs. Presumably, this allows a design tradeoff between sensitivity and complexity. More nearest neighbors can be queried and then winnowed with a pairwise shape based operation or other features.

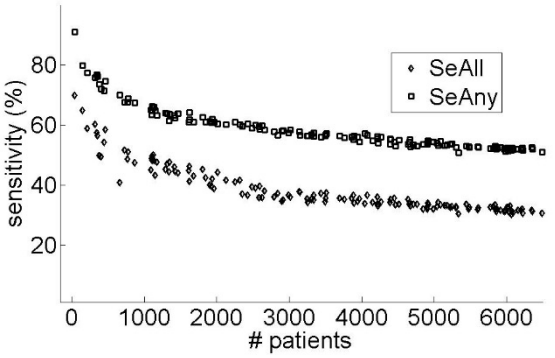

Figure 3. Sensitivity versus number of patients in the database by bootstrap. The sensitivity of finding a patient's other ECGs falls as the number of patients grows.

A comparison of results between this study and that of Bousseljot (3) is not possible because the strategies were completely different. While Bousseljot used a fixed shape metric and threshold and a variable number of nearest neighbors, we used a fixed number of nearest neighbors and a variable level of similarity by shape.

Since we used the other ECGs of a patient as the similarity reference, it makes sense to compare to studies using ECG as an identifying biometric. However, we want to find ECGs with a similar shape purposely focusing what is similar about different patients rather than focusing on what makes different patients unique. The error rate of the ECG biometric techniques is lower than ours in general but those great results come from training and testing with healthy subjects during the same session (2). For an equivalent comparison, the error rate of our technique is 1 – sensitivity and the number of patients must be reduced to 200. For equivalence, we must compare to Odinaka's results for 16 beats, training in one session and testing in another session. Our error rate of 10 to 30% (see Figure 3, lowest number of patients) is in the range of methods reported by Odinaka where the average error rate was 23% (16 to 47%) across 20 different methods.

Since our morphology-similarity metric was based on a patient's other ECGs, and ECGs in the test set were recorded over a period of several years in a hospital population, the validation was designed to account for ECG morphology variations due to time, recording techniques such as electrode placement, and health status changes.

5. Conclusion

We conclude that using a low complexity similarity query for 12-lead ECG can achieve similar performance level as an expensive shape metric and exhaustive search.

References

[1] Laguna P, Jane R, Caminal P. Adaptive feature extraction for QRS classification and ectopic beat detection. Computers in Cardiology. 1991;: p. 613-616.

[2] Odinaka I, Lai PH, Kaplan AD, O'Sullivan JA, Sirevaag EJ, Rohrbaugh JW. ECG Biometric Recognition: A Comparative Analysis. IEEE Transactions on Information Forensics and Security. 2012; 7(6): p. 1812-24.

[3] Bousseljot R, D K. Waveform recognition with 10,000 ECGs. Computers in Cardiology. 2000;: p. 331-334.

[4] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000; 101(23): p. e215-e220.

[5] Vedaldi A, Fulkerson B. VLFeat: An Open and Portable Library. [Online].; 2008 [cited 2016 June 10. Available from: http://www.vlfeat.org.

[6] Kors JA, van Herpen G, Sittig AC, van Bemmel JH. Reconstruction of the Frank vectorcardiogram from standard electrocardiographic leads: diagnostic comaprison of differnt methods. Euro Heart J. 1990; 11: p. 1083-1092.

[7] MATLAB 82. Ensemble learning. Natick: MathWorks Inc.; 2013.

[8] Hodges M, Salerno D, Erlien D. Bazett QT correction reviewed - evidence that a linear QT correction for heart-rate is better. JACC. 1983; 1(2): p. 694-694.

[9] Muja M, Lowe DG. Fast approximate nearest neighbors with automatic algorithm aonfiguration. In International Conference on Computer Vision Theory and Application; 2009. p. 331-340.

[10] Gregg RE, Zhou SH, Lindauer JM, Helfenbein ED, Feild DQ. Limitations on the re-use of patient specific coefficients for 12-lead ECG reconstruction. IEEE Computers in Cardiology. 2008;: p. 209-212.

[11] Moody GB, Mark RG. Development and evaluation of a 2-lead ECG analysis program. Computers in Cardiology. 1982;: p. 39–44.

Address for correspondence.

Richard Gregg
Philips Healthcare, MS4201
3000 Minuteman Road
Andover, MA USA 01886