DESIGN OF A HAMMING-DISTANCE CLASSIFIER FOR ECG BIOMETRICS

Siddarth Hari, Foteini Agrafioti, Dimitrios Hatzinakos

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering,
University of Toronto,
10 King's College Road, Toronto, ON, Canada, M5S 3G4
{shari,foteini,dimitris}@comm.utoronto.ca

ABSTRACT

In existing ECG-based biometric recognition systems, the feature extraction and matching are performed in Euclidean spaces. However, there are many scenarios (e.g., biometric template encryption for privacy protection, or low-complexity classification in an identification mode of operation) in which it is useful to binarize the feature vectors. The main contribution of this paper is a Hamming-distance classifier for ECG biometrics based on SPEC-Hashing. The proposed system was evaluated over a database of ECG signals from 52 different subjects that were collected at the Biometrics Security Laboratory of the University of Toronto. The EER of the Hamming-distance classifier was found to be 5.5% for closed-set matching and 14.82% for open set matching.

***Index Terms*—** Autocorelation, electrocardiogram, SPEC-Hashing

1. INTRODUCTION

1.1. Medical Biometrics

Biometric recognition systems are rapidly replacing traditional identification systems based on PIN numbers, tokens or passwords. Fingerprint, face and iris biometric systems are increasingly being used to secure passports and grant access to high security environments.

However, as the technology for biometric recognition advances and its use becomes more widespread, so too do the methods and technology for biometric falsification. Commercial fingerprint sensors are defeated with off-the-shelf components that can be used to create identical copies of a person's fingerprint features. Replay attacks are already a credible threat to voice and fingerprint biometrics. Biometric obfuscation, i.e. removal of biometric features to avoid establishment of one's true identity, is another prominent challenge (for example, asylum-seekers in Europe intentionally damaged their fingerprints). With the wide deployment of biometrics, these attacks are becoming more frequent and concerns are being raised about the kind of security that biometric technologies can offer.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

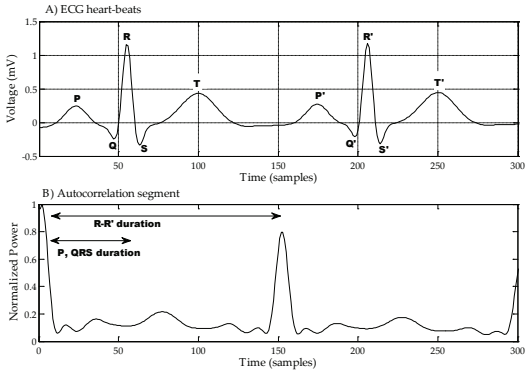

Fig. 1. (a) Basic components of ECG heart-beats (b)Autocorrelation (zoomed-in) of an ECG signal

The next generation of biometric systems and modalities are being developed in light of the threats listed above. Medical biometrics, a new but promising category of biometric features, is one such example. Medical biometrics utilize physiological signals of the human body (vital signals) that contain subject-specific characteristics. Since these vital signals are internal to the human body, medical biometrics offer an inherent robustness to circumvention, replay and obfuscation attacks. Furthermore, biometric liveness is inherently guaranteed.

1.2. ECG Biometrics

The focus of this paper is on ECG-based biometric recognition. The ECG is a vital signal of cardiovascular origin which reflects the cardiac electrical potential of the heart. It is formed during the depolarization and repolarization of specific parts of the myocardium and can be measured at the surface of the body using electrodes which can be placed in various configurations. From an engineering point of view, the ECG has a non-periodic but highly repetitive pattern. Every quasi-period of the ECG corresponds to a pulse (or heartbeat) and it signifies a full cycle of the cardiac function. There are particular points of interest on a heart-beat called *fiducial points*, such as the ones shown in Figure 1. These are primarily related to the main ECG waves, namely the *P* wave, the

QRS complex and the *T* wave.

There are two main approaches for feature extraction from ECG signals, namely fiducial points-dependent or independent. Fiducial-dependent approaches extract features based on the local characteristics of heart-beats [1, 2]. For example, temporal and amplitude distances between consecutive fiducial points on the ECG were proposed in [1, 3]. On the other hand, non-fiducial approaches explore the ECG waveform as a whole thereby eliminating the need for heart-beat segmentation and fiducial-point detection [4, 5, 6]. In the above works, the resulting feature vectors lie in Euclidean spaces. To the best of our knowledge, there hasn't been significant prior work that addresses the problem of binarization (i.e. transformation of feature vectors from Euclidean into binary Hamming spaces) for ECG signals. In [7], a quantization approach was suggested (using a Max-Lloyd quantizer), together with a Gray-code mapping. However, quantization-based approaches do not necessarily preserve distance properties, i.e two feature vectors that are separated by large Euclidean distances are not guaranteed to be mapped to two binary vectors that have a large Hamming-distance separation.This observation has motivated the subsequent analysis for ECG biometric processing in the Hamming space.

2. OVERVIEW OF THE PROPOSED SYSTEM

The proposed method for ECG analysis in the hamming space is based on the Autocorrelation/ Linear Discriminant Analysis (AC/LDA) [6] in conjunction with SPEC-Hashing. We use the following notation : we denote vectors using bold lower-case $\mathbf{x}$. x_m denotes the m-th component of the vector $\mathbf{x}$, whereas $\mathbf{x}^{(i)}$ denotes the i-th vector in a set. We use bold upper-case $\mathbf{X}$ for matrices. $\mathbf{f}(\cdot)$ denotes that the output of the function is a vector.

2.1. Feature extraction

The AC/LDA is a two step procedure in which the autocorrelations of 5-sec ECG windows are first estimated, and then subjected to dimensionality reduction using the LDA.

Figure 1 shows an example of AC for an ECG reading. The autocorrelation is computed as:

$$\widehat{R}_{ee}[m] = \sum_{i=0}^{N-|m|-1} e[i]e[i+m] \qquad (1)$$

where $e[i], i = 0, 1...(N-1)$, are samples of the ECG signal, m is the time lag, and N is the length of the signal. Out of $\widehat{R}_{ee}$ only a segment $\phi[m], m = 0, 1...M$, starting from the zero lag instance and extending to approximately the length of *P* and *QRS* waves (as shown in Figure 1) is input to the LDA training block. These waves are the least affected by heart rate changes, and consequently, utilizing only this segment for discriminant analysis makes the system robust to heart rate variability.

The output of the LDA-training is a transformation matrix $\mathbf{W}$. The AC feature vectors are projected with $\mathbf{W}$ into the Euclidean space: $\mathbf{y} = \mathbf{W}\phi^T$. Details pertaining to the AC/LDA can be found in [6] and [8].

2.2. Binarization

In [7], a quantization approach was suggested (using a Max-Lloyd quantizer), together with a Gray-code mapping. While Gray-mapping is effective when the number of bits is small ($<$ 5bits), this approach does not scale well. In addition, biometric feature vectors have unique properties, such as small intra-class variability and high inter-class variability, which are not taken into account in ML quantization.

This motivates the need for class-specific or similarity-preserving binarization techniques, which is the focus of this work.

There has been significant amount of work on discovering good projections from arbitrary feature spaces (metric spaces) into binary vector-spaces. This interest has been motivated by the need for fast nearest neighbor search in face-recognition and image-retrieval applications. Some methods include the traditional KD-tree, and the more recent Locality Sensitive Hashing, Parameter Sensitive Hashing, RBM-based learning, Spectral Hashing, SPEC-Hashing, random quantization using Fourier features etc. Some qualitative discussion and relevant references can be found in [9].

In this paper, the possibility of adapting the SPEC-Hashing algorithm to the needs of biometric systems is evaluated. The purpose is to map (component wise) feature vectors from the space containing the image of the LDA-based transform (i.e., $Im(\mathbf{W})$) into binary vectors in such a way that two feature vectors which have a small Euclidean distance between them are mapped to two binary vectors that are separated by small Hamming distance. With this manner, the intra-class and inter-class variabilities will be carried to the new space. SPEC-Hashing is a similarity preserving algorithm for entropy-based coding that was developed by Lin *et al.*[9] for fast nearest neighbor search in high-dimensional feature spaces, primarily in the context of image retrieval and celebrity face recognition applications. The nearest neighbors are defined according to the semantic similarity between objects in the feature space.

In this work, the SPEC-Hashing approach is evaluated for closed-set and open-set biometric matching. In the first case, a fraction of each user's enrollment data is used for training of the SPEC-Hashing algorithm. For the open-set scenario, a generic database of non-enrollees is input for SPEC-Hashing training while the output is applied to a new set of users i.e., the enrollees of the system under evaluation.

The input to the SPEC-Hashing block is the set of LDA-feature vectors $\{\mathbf{x}^{(i)}\} \subset \mathbb{R}^M$ corresponding to the training data, along with a similarity matrix $\mathbf{S}$. S_{ij}, the (i,j)-th entry of the similarity matrix, denotes the semantic similarity between the i-th and j-th feature vectors in the training set.

The SPEC-Hashing algorithm is run separately for each of the M components of the LDA- feature vectors. The output is a list of M threshold-vectors $\{\lambda_1, \lambda_2, \cdots, \lambda_M\}$. Each

of these threshold-vectors corresponds to a component of the LDA- feature vectors, and they can have varying lengths (their length is optimally determined by the algorithm).

Once the list of thresholds is obtained from the SPEC-Hashing algorithm during the training phase, the binarization of the ECG feature vectors is done as follows :

Let us first look at the m-th component. Suppose $\lambda_m = [\lambda_{m,1}, \lambda_{m,2}, \cdots, \lambda_{m,n}]$. Then, the m-th component x_m of a feature vector $\mathbf{x}$ is mapped to an n-bit vector $\mathbf{b}_m(x_m) = [b_{m,1}(x_m) b_{m,2}(x_m) \cdots b_{m,n}(x_m)]$ where

$$b_{m,i}(x_m) = \begin{cases} 0 & x_m \leq \lambda_{m,i} \\ 1 & x_m > \lambda_{m,i} \end{cases} \quad (2)$$

Once we have the binary vectors for each component x_m, we simply form the binary representation of $\mathbf{x}$ by concatenation, i.e., $\mathbf{b}(x) = [\mathbf{b}_1(x_1)\mathbf{b}_2(x_2) \cdots \mathbf{b}_M(x_M)]$.

In the enrollment phase, for each feature vector $\mathbf{x}^{(i)}$ we compute the corresponding binary representation $\mathbf{b}\left(\mathbf{x}^{(i)}\right)$. Here on, we will denote this as $\mathbf{b}^{(i)}$ for simplicity, and refer to these as the binary feature vectors.

During the authentication phase, the feature vector $\mathbf{y}$ extracted from the user's ECG reading is mapped to the binary vector $\mathbf{b} = \mathbf{b}(\mathbf{y})$ using the same binary decision stumps $\mathbf{b}(\cdot)$.

We now comment briefly on the operation of the SPEC-Hashing algorithm (a detailed treatment can be found in [9]). Suppose we have a list of thresholds $\mathcal{L}$, and the corresponding binary mappings obtained using Eq.(3). Let $d_H(i,j)$ denote the Hamming distance between the binary vectors $\mathbf{b}^{(i)}$ and $\mathbf{b}^{(j)}$ (corresponding to the feature vectors $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$). We can define the matrix $\mathbf{T}$ (corresponding to list $\mathcal{L}$)as follows

$$T_{ij} = \frac{1}{Z} e^{-d_H(i,j)} \quad (3)$$

where $Z = \sum_{i,j} e^{-d_H(i,j)}$. Also, let us denote the normalized similarity matrix by $\tilde{\mathbf{S}}$. We can then view the operation of the SPEC-Hashing algorithm as finding an optimum list of thresholds that minimizes the Kullback-Liebler divergence between the distribution $\mathbf{T}$ and the target distribution $\tilde{\mathbf{S}}$. The closeness of the two distributions represents how closely the similarities are preserved by the mapping from the Euclidean space to the Hamming feature space.

The SPEC-Hashing algorithm builds the list $\mathcal{L}$ in an incremental fashion. Let $\mathcal{L}^* = [\mathcal{L}, \lambda]$. Define

$$\Delta(\lambda) = KL(\tilde{\mathbf{S}} || \mathbf{T}^*) - KL(\tilde{\mathbf{S}} || \mathbf{T}) \quad (4)$$

If $\Delta(\lambda)$ is negative, then adding λ to the list reduces the divergence from the target distribution, and the list is updated to include λ.

Algorithm 1 SPEC-Hashing

Inputs : LDA-feature vectors (for training set), Similarity matrix

Outputs : an ordered list of thresholds

```
Set L = ∅
for each component m = 1, 2, ··· , M do
    Set flag = 0.
    while flag = 0 do
        Find λ_opt that minimizes Δ(λ).
        if Δ(λ) < 0 then
            Update L = [L, λ]
        else
            Set flag = 1.
        end if
    end while
end for
```

3. EXPERIMENTAL RESULTS

The proposed system was evaluated over ECG signals from 52 different volunteers that were collected at the Biometrics Security Laboratory of the University of Toronto. Two signal collection sessions took place, scheduled a few weeks apart in order to evaluate the stability of the signal with time. Every recording is 3 minutes long and the lead orientation matches that of Lead II of the standard 12-lead ECG system. 16 out of 52 volunteers were recorded in both sessions and 36 in just one session. The sampling frequency is 200Hz.

Open-set recognition setup. To simulate an open-set biometric system the dataset was split between a *training* and an *evaluation* set. The training set included the 36 subjects for who just one ECG reading is available.The proposed algorithm was trained on signals from this set. The evaluation set included signals from the 16 volunteers for which two recordings are available. The earlier recording was used for enrollment and the later for testing.

Closed-set recognition setup. To simulate the closed-set biometric system the proposed algorithm was trained on ECG signals from all users i.e., the enrollees of the system. The enrollment data includes the first ECG readings from the 16 volunteers and the first half of the ECG reading from the 36 volunteers.

The resulting binary vectors, obtained using the list of binary decision-stumps learnt by the SPEC-Hashing algorithm were 345 bits long for the open-set setup and 468 bits long for the closed-set setup (we set the maximum number of bits per dimension to be 25 bits). The difference of the bit length between the two setups is due to the dimensionality after LDA projection.

Figure 2 shows the histograms for the intra-subject and inter-subject Euclidean and Hamming distances for the closed-set recognition setup. For the Euclidean space, the average inter-subject distance is 0.2656 and the average intra-subject distance is 0.0757. For the Hamming space, the average inter-subject distance is 0.2496 (116 bits) and the intra-subject 0.0719 (33 bits).

Figure 3 shows the performance of three open-set sys-

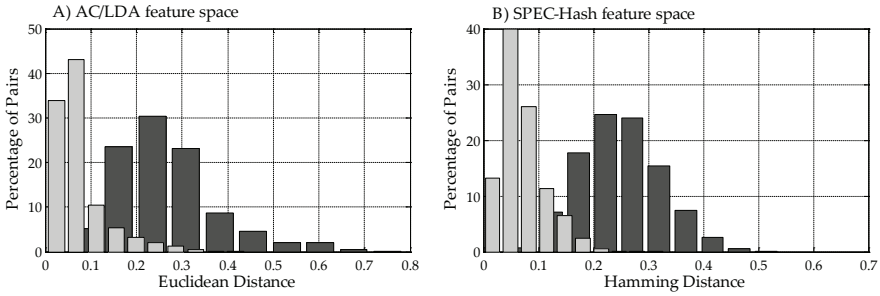

Fig. 2. Histogram of intra-class and inter-class Euclidean and Hamming distances

System	Setup	EER
AC/LDA Euclidean Classifer	Open-set	18.6%
Hamming Classifier (SPEC-Hashing)	Open-set	14.82%
AC/LDA Euclidean Classifer	Closed-set	11.4%
Hamming Classifier (SPEC-Hashing)	Closed-set	5.5%

Table 1. System equal-error-rate (EER)

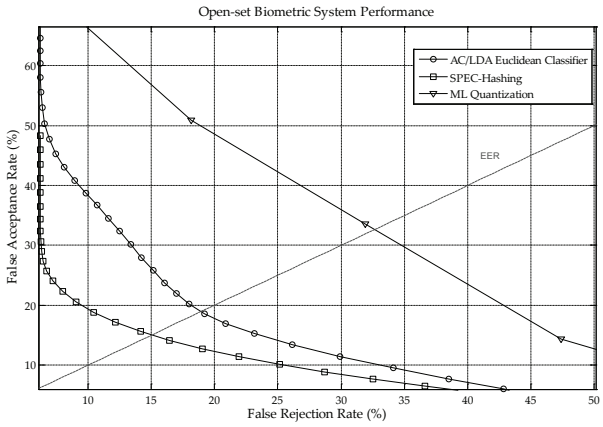

Fig. 3. ROC: FAR versus FRR for the Euclidean and Hamming (SPEC-Hash and ML) spaces

tems: the AC/LDA Euclidean-distance classifier (which is the baseline system), the Max-Lloyd quantization approach (for the sake of completeness), and the proposed Hamming-distance classifier with binary feature vectors designed using the SPEC-Hashing algorithm. The Max-Lloyd quantization approach does not perform well, because binarization does not take into consideration class information. The performance of the proposed Hamming-distance classifier using binarization based on SPEC-Hashing outperforms all cases. As stated in Table 1, the Equal Error Rate (EER) for the AC/LDA based Euclidean-distance classifier was found to be 18.6% (Open-set), and the EER of the proposed Hamming-distance classifier is 14.82% (Open-set). The EER in the closed set setup of the proposed system is 5.5%.

4. CONCLUSION

As ECG biometric systems begin to scale, it is important to study binarization of the ECG feature vectors and design of Hamming-distance classifiers, in order to enable important applications such as template protection using biometric encryption, efficient low-complexity matching for identification mode of operation etc. A method for binarization of the ECG feature vectors based on SPEC-Hashing was proposed and evaluated over ECG signals from 52 subjects. With the proposed treatment there is virtually no loss in performance compared to nearest neighbor classification in the Euclidean LDA-projection space.

5. REFERENCES

[1] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, "ECG to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

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

[3] S.I. Safie, J.J. Soraghan, and L. Petropoulakis, "Electrocardiogram ECG biometric authentication using pulse active ratio PAR," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 4, pp. 1315–1322, dec. 2011.

[4] G. G. Molina, F. Bruekers, C. Presura, M. Damstra, and M. van der Veen, "Morphological synthesis of ECG signals for person authentication," in *Proceedings of 15th European Signal Proc. Conf.*, Poland, Sept. 2-7 2007.

[5] I. Odinaka, Po-Hsiang Lai, A.D. Kaplan, J.A. O'Sullivan, E.J. Sirevaag, S.D. Kristjansson, A.K. Sheffield, and J.W. Rohrbaugh, "ECG biometrics: A robust short-time frequency analysis," in *Proceedings of IEEE International Workshop on Information Forensics and Security*, Dec. 2010, pp. 1–6.

[6] F. Agrafioti and D. Hatzinakos, "Fusion of ECG sources for human identification," in *Proceedings of 3rd Int. Symp. on Communications Control and Signal Processing*, Malta, March 2008, pp. 1542–1547.

[7] F. Agrafioti, F. M. Bui, and D. Hatzinakos, "Medical biometrics in mobile health monitoring," *Security and Communication Networks*, vol. 4, no. 5, pp. 525–539, July 2010, .

[8] Foteini Agrafioti and Dimitrios Hatzinakos, "ECG biometric analysis in cardiac irregularity conditions," *Signal, Image and Video Processing*, pp. 1863–1703, 2008.

[9] Ruei-Sung Lin, D.A. Ross, and J. Yagnik, "Spec hashing: Similarity preserving algorithm for entropy-based coding," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 848–854.