OPTIMAL THRESHOLDING FOR KEY GENERATION BASED ON BIOMETRICS

Wende Zhang[+]*, Yao-Jen Chang*[*]*, and Tsuhan Chen*[+]

[+]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[*]Industrial Technology Research Institute, Chutung, Hsinchu, Taiwan 310, R.O.C.

ABSTRACT[1]

In this paper, we introduce a novel method for key generation based on biometrics. Given the biometrics, a set of reliable features are extracted. Each feature is compared with multiple thresholds to generate a multi-bit key. By cascading the multi-bit keys, we obtain one bio-key that can be used for security applications. In order to generate a reliable bio-key, an optimal thresholding method is proposed to minimize the authentication error rate, in terms of the false accept rate (*FAR*) and the false reject rate (*FRR*). The experimental results show that the proposed approach of key generation is user-friendly and reliable.

1. INTRODUCTION

Many access control applications, such as ATM access and building access, demand secret information from users to authenticate the users' identities. One example of such secret information is a digital key. A digital key is also used in many encryption / decryption applications.

A long key can improve the security of the system and decrease the possibility of an exhaustive attack (i.e. trying all possible keys). However, a long key is often associated with a high false reject rate (FRR), since it is easily forgotten and not user-friendly.

In the research for the replacement of long keys, researchers have moved to key-generation based on biometric authentication, which is more user-friendly. Given a user's biometric data, such as face images, fingerprints, etc., a biometric authenticator validates the user's identity. If the access to the link table between the claimed identity and the stored key is granted to the user, the system yields the corresponding key. However, such a system is secure only if the key generation system is protected perfectly, which may not be true in some cases. For example, an attacker might break into the key generation system by bypassing the authenticator to access the link table directly, and extract the key.

Some key generation systems have been proposed recently to generate the bio-key directly instead of having a biometric authenticator followed by the link table. Soutar *et al.* [1] proposed the biometric encryption method. In this method, the input biometric image is correlated with a pre-designed filter to create the correlation output. The key is then generated based on this binarized output pattern.

Monrose *et al.* [2] proposed using a secret sharing method to generate the bio-key. First, the distinguishing biometric features are selected based on the separation between the authentic and imposter values, and then binarized by some thresholds. The key is then released under the secret sharing scheme by matching all the bits of the input biometrics with the authentic bits, which are computed from the distinguishing features.

Based on Monrose *et al.*'s [2] bio-key generation approach, Chang *et al.* [3] extended the distinguishing feature selection to feature transformation in order to generate more distinguishing features using the cascaded two-class classification scheme. They also extended the binary values of each feature to multiple values so that each feature may contribute multiple bits rather than one bit to improve the performance of the bio-key generation system.

However, both Monrose *et al.* [2] and Chang *et al.* [3] did not address the issue of setting the thresholds for the features. In this paper, we propose a method to minimize the authentication error rate in terms of the false accept rate (*FAR*) and the false reject rate (*FRR*) of the bio-key generation system by setting optimal thresholds for each feature.

This paper is organized as follows: In Section 2, we introduce the framework of key generation based on biometrics. In Section 3, we detail the optimal thresholding method for the bio-key generation system. In Section 4, we describe the biometric database used in the experiment and present the performance of the proposed

[1] This work is a partial result of Project B33BCM1100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

bio-key generation system. Our conclusions are given in Section 5.

2. THE FRAMEWORK OF KEY GENERATION

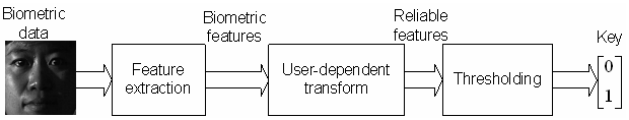

Figure 1. Key generation based on biometrics

A flowchart of the bio-key generation system is shown in Figure 1. First, the biometric features are extracted from the biometric data based on the feature extraction module. For example, we use Principal Component Analysis (PCA) [4] to extract the eigen-coefficients from the data as the biometric features. Next, the user-dependent transform module produces the reliable features, where the authentic values are more clearly separated from the imposter values. Finally, the reliable features are thresholded into the key by the thresholding module.

3. OPTIMAL THRESHOLDING

In this paper, we propose a new method for determining the optimal thresholds for each feature, thereby minimizing the authentication error rate in terms of the *FAR* and the *FRR*. For key generation, *FAR* is defined as the rate that imposter users generate the same bit-sequence as claimed users. *FRR* is defined as the rate that an authentic user generates a bit-sequence other than his key.

The objective of the optimization is to find a set of thresholds on each feature to minimize the *FRR* for any given *FAR* on the Receiver Operation Characteristic (ROC) curve as follows.

$$\max(1-FRR), \text{ given } FAR=\alpha,\ 0\le\alpha\le 1 \qquad (1)$$

where $FAR=\int_{R_A} p(\mathbf{x}\mid w_I)d\mathbf{x}$, $FRR=\int_{R_I} p(\mathbf{x}\mid w_A)d\mathbf{x}$, and $p(\mathbf{x}\mid c)$ is the probability density function of the feature vector $\mathbf{x}=[x_1,x_2,...,x_n]$, and n is the number of the features, given the class label c. R_A is the classification decision region for the authentic w_A, and R_I is the classification decision region for the imposter w_I.

The authentic decision region R_A is defined by the thresholds for each feature x_i in the feature vector $\mathbf{x}$. The decision region R_A can be rewritten as (2), since it is separated from the imposter decision region R_I by thresholding each feature individually, and the final bit-sequence is cascaded by the bits from all features.

$$R_A=\{x_1\in R_1\wedge x_2\in R_2\ \ldots\wedge x_n\in R_n\} \qquad (2)$$

For simplicity, we assume that the authentic decision region R_i of the feature x_i is defined by its boundary thresholds $T_{left,i}$ and $T_{right,i}$ as follows.

$$R_i=[T_{left,i},T_{right,i}]=[m_{authen,i}-k_i\sigma_{authen,i},m_{authen,i}+k_i\sigma_{authen,i}] \qquad (3)$$

where $m_{authen,i}$ and $\sigma_{authen,i}$ are the mean and the standard deviation of the authentic values on the feature x_i. The value of k_i controls the boundary thresholds $T_{left,i}$ and $T_{right,i}$ as shown in Figure 2.

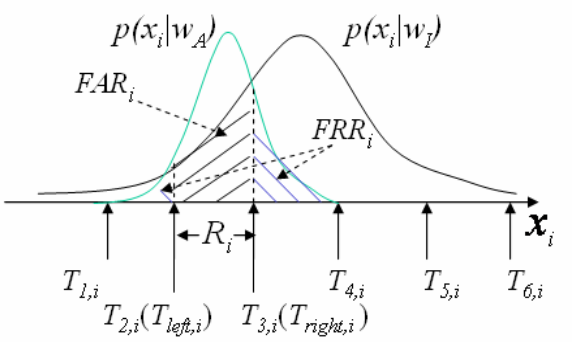

Figure 2. The i^{th} reliable feature

In order to generate the bits from each feature for bio-key generation, the feature x_i is divided equally into t_i key-generation-segments, each having the same range as the authentic decision region R_i, by the thresholds, $T_{r,i}$, $1\le r\le t_i+1$, to cover all the values.

Since the value of k_i controls the thresholds on the feature x_i, to optimize the combined (*FAR, FRR*) we search for an operating point (FAR_i, FRR_i) on each ROC curve corresponding to feature x_i; each point (FAR_i, FRR_i) corresponds to one k_i value.

Assuming all the features are conditionally independent from each other given the class labels, we have

$$p(\mathbf{x}\mid w_I)=p(x_1\mid w_I)p(x_2\mid w_I)...p(x_n\mid w_I)$$
$$p(\mathbf{x}\mid w_A)=p(x_1\mid w_A)p(x_2\mid w_A)...p(x_n\mid w_A)$$

Therefore, the overall *FAR* and *FRR* can be explicitly derived as a function of the individual FAR_i and FRR_i as follows.

$$\begin{aligned} FAR &= \int_{R_A} p(\mathbf{x}\mid w_I)d\mathbf{x} \\ &= \int_{R_1} p(x_1\mid w_I)dx_1\int_{R_2} p(x_2\mid w_I)dx_2...\int_{R_n} p(x_n\mid w_I)dx_n \\ &= FAR_1\cdot FAR_2\cdots FAR_n \end{aligned} \qquad (4)$$

$$\begin{aligned} 1-FRR &= \int_{R_A} p(\mathbf{x}\mid w_A)d\mathbf{x} \\ &= \int_{R_1} p(x_1\mid w_A)dx_1\int_{R_2} p(x_2\mid w_A)dx_2...\int_{R_n} p(x_n\mid w_A)dx_n \\ &= (1-FRR_1)\cdot(1-FRR_2)\cdots(1-FRR_n) \end{aligned} \qquad (5)$$

One way to solve the optimization is to perform an exhaustive search for the operating point (FAR_i, FRR_i) on each ROC curve. For example, given two ROC curves as shown in Figure 3a and Figure 3b, first, we plot the possible operating points on the combined ROC curve by trying all the possible combination of (FAR_1, FRR_1) and (FAR_2, FRR_2) based on equations (4) and (5). Using this brute-force approach, we can identify a feasible region of the combined ROC curve as shown in Figure 3c. Then,

given an *FAR*, we choose the minimal *FRR* in this region as the optimal combined curve as shown in Figure 3d. As a result, for each point on the optimal combined ROC curve, we have also identified corresponding optimal (FAR_i, FRR_i) on individual ROC curves.

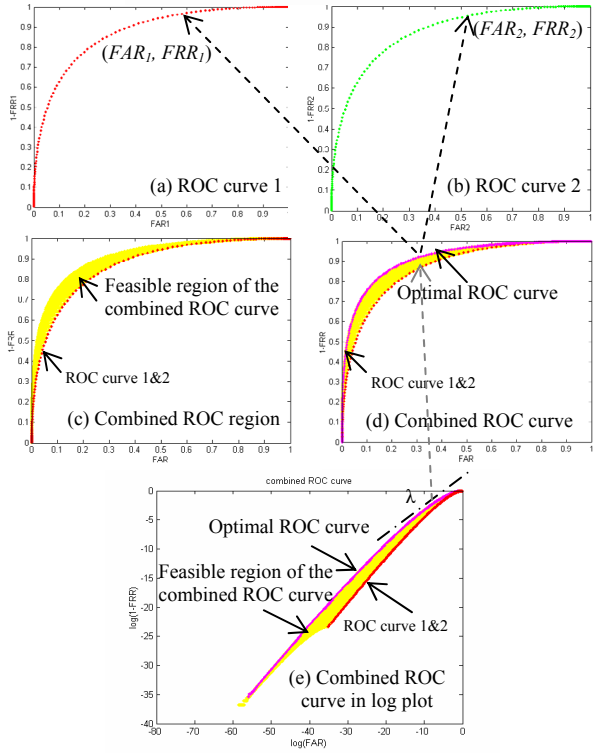

Figure 3. ROC curve combination

If we further assume that the curve $\log(1-FRR)$ vs. $\log(FAR)$ is concave, we can use a simpler method to determine the optimal combined ROC curve without an exhaustive search as follows.

Since $\log(x)$ is a monotonic increasing function, we can form the optimization in (6) instead of (1).

$$\max \log(1-FRR) = \max \sum_{i=1}^{n} \log(1-FRR_i)$$
$$\text{given } \log(FAR) = \sum_{i=1}^{n} \log(FAR_i) = \log \alpha \qquad (6)$$

Since we assume that the curve $\log(1-FRR)$ vs. $\log(FAR)$ is concave, the combined ROC curve, which is the boundary of the feasible region as shown in Figure 3e, can be carved out by a set of straight lines with different slopes λ. Therefore, the solution of (6) can be found by solving the following equivalent optimization problem.

Given $\lambda \geq 0$,

$$\max[\log(1-FRR) - \lambda \log(FAR)]$$
$$= \max \sum_{i=1}^{n} [\log(1-FRR_i) - \lambda \log(FAR_i)] \qquad (7)$$
$$= \sum_{i=1}^{n} \max \{\log(1-FRR_i) - \lambda \log(FAR_i)\}$$

By maximizing the value of $\log(1-FRR_i) - \lambda \log(FAR_i)$ with the same slope λ on each curve individually, we can easily find the solution of the optimal point ($\log(FAR_i)$, $\log(1-FRR_i)$) on each curve, which corresponds to one optimal point on the combined curve. By tuning different λ's, we can get the solution for any optimal point on the combined ROC curve. Actually, similar curve-combining methods can be found in many other fields, such as choosing the threshold optimally for a multiple-subject authentication system [5] and the rate-distortion techniques used for video coding [6].

4. EXPERIMENTAL RESULTS

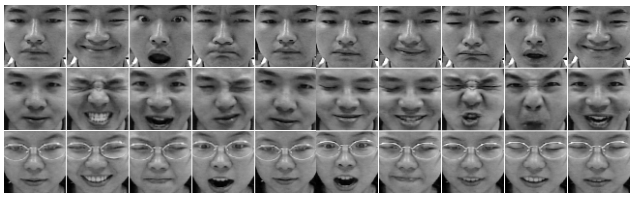

Figure 4. Sample images of AMP face database with expression and registration error

We conducted experiments on the AMP face database with expression and registration error to verify the performance of the bio-key generation system using the proposed method. This database has 30 subjects. Each subject has 137 images. The face region is 64×64 pixels for each image. Sample images are shown in Figure 4.

We propose a function to evaluate the key generation system based on biometrics. The key generation system must have a low authentication error rate in terms of the *FAR* and the *FRR* for the users in the training database (closed-set verification [7]). This system also must have a small error rate for the unknown users outside the training database (open-set verification [7]), since an unknown user should not be allowed to access the system by chance. This type of error is denoted as *E_U*. In addition, the chance that an attacker provides the key directly without the knowledge of the key generation system should be small. We denote such chance as *E_K*. If the attacker does not know any biometric information about the subjects or the system, he has to make an exhaustive search in the key space. Therefore, $E\_K = 1/2^{B_{effective}}$, where $B_{effective} = \log_2 \prod_{i=1}^{n} t_i$, and t_i is the number of the segments in the i^{th} feature.

We combine all the above terms with the trade-off parameters c_1, c_2, c_3, and c_4 to define the evaluation function for the bio-key generation system, as described in (8).

$$f_{evaluate} = c_1 * FAR + c_2 * FRR + c_3 * E\_U + c_4 * E\_K \qquad (8)$$

We randomly assign 20 subjects as the users in the training database ($S_{training}$) and 10 as the unknown users. 25 images of each user of $S_{training}$ are for training the feature extraction module. Principal Component Analysis is performed on all the training images to reduce the features' dimensionality. The first 100 eigen-coefficients are taken as the biometric features as shown in Figure 1. The user-dependent transform is also trained based on the same set of the training images. In the experiment, we trained an L-dimensional ($1 \le L \le 100$) G-SMMS subspace [8] as the user-dependent transform. We use another 25 images of each user of $S_{training}$ to determine the thresholds for each feature. The remaining 87 images of each user of $S_{training}$ are used as test images to evaluate *FAR*, *FRR* and *E_K*. All the images of the unknown users are used as test images for evaluating the error of unknown users, *E_U*.

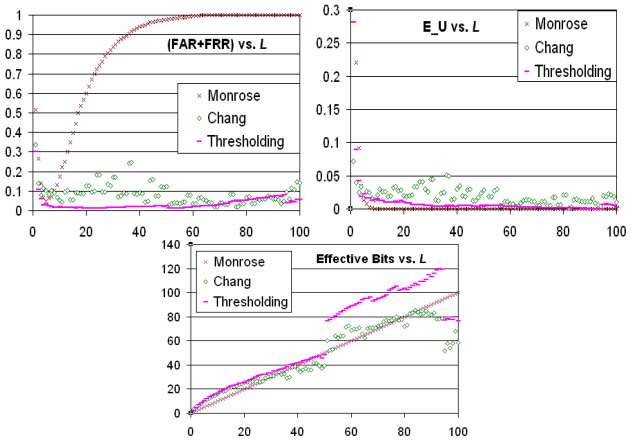

Figure 5. Performance comparison for different values of L

We compare the proposed (Thresholding) bio-key generation method with Monrose's method proposed in [2] and Chang's method proposed in [3]. Monrose *et al.* [2] determine the thresholds for the features by setting them at the global means. The *Effective Bits* ($B_{effective}$) of their approach is the number of the distinguishing features. Chang *et al.* [3] determine the thresholds for authentic region by a k value in (3), which is fixed for all the features, while the proposed method searches for an optimal k_i value for each feature in order to minimize the authentication error rate. We evaluate the bio-key generation system performance of these three methods among the different L values, which are the number of distinguishing features [2] in 100-dimensional PCA subspace for Monrose's method, and the number of dimensions of the G-SMMS subspace for Chang's method and the proposed method, by choosing the best operating point where $f_{Evaluate}$ is the minimal with $c_1 = c_2 = c_3 = c_4 = 0.25$. The results of (*FAR*+*FRR*) vs. L, *E_U* vs. L, and *Effective Bits* ($B_{effective}$) vs. L are shown in Figure 5. We also show the results of the best operating point among all the L values in Table 1. The results indicate that the optimal thresholding approach is better than Monrose's and Chang's approaches in term of improving the overall performance of the bio-key generation system by reducing the authentication error rate significantly while improving the security in term of the key space, which is indicated by the *Effective Bits*.

	$f_{Evaluate}$	*FAR*	*FRR*	*E_U*	$B_{effective}$
Monrose	0.0156	2.0%	2.8%	1.3%	5
Chang	0.0082	0.5%	1.7%	1.1%	70.6
Thresholding	**0.0041**	**0.1%**	**1.0%**	**0.5%**	**79.1**

Table 1. Performance comparison of three methods

5. CONCLUSIONS

In this paper, we propose a reliable bio-key generation system with an optimal thresholding approach. The experimental results show that the proposed approach improves the performance of the original bio-key generation system by reducing the authentication error rate significantly while improving the security in term of the key space.

6. REFERENCES

[1] C. Soutar, D. Roberge, A. Stoianov, R. Gilroy and B.V.K. Vijaya Kumar, "Biometric Encryption™," Chapter 22 in *ICSA Guide to Cryptography*, edited by R.K. Nicholls, pp.649-675, 1999.

[2] F. Monrose, M.K. Reiter, Q. Li, and S. Wetzel, "Cryptographic key generation from voice," *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pp. 202-213, 2001.

[3] Y. Chang, W. Zhang and T. Chen, "Biometric-based cryptographic key generation," submitted to *IEEE Conference on Multimedia and Expo,* 2004.

[4] I.T. Jolliffe, *Principle Component Analysis*, Spring-Verlag, New York, 1986.

[5] X. Liu, T. Chen and B.V.K. Vijaya Kumar, "*Face Authentication for Multiple Subjects Using Eigenflow*," Pattern Recognition, special issue on Biometric, Volume 36, Issue 2, pp. 313-328, 2003.

[6] A. Ortega, K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, 15(6), pp. 23-50, 1998.

[7] E. Bailly-Bailliere, and *et. al*, "The BANCA database and evaluation protocol," *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*. Springer-Verlag, 2003.

[8] W. Zhang and T. Chen, "Personal authentication based on generalized symmetric max minimal distance in subspace," *Proceedings of IEEE Conference on Multimedia and Expo*, Baltimore, MD, 2003.