Securing Biometric Templates where Similarity is Measured with Set Intersection

Daniel Socek[1], Vladimir Božović[2], and Dubravko Ćulibrk[1]

[1] CoreTex Systems LLC
2851 S Ocean Blvd. 5L, Boca Raton, Florida 33432, USA
[2] Department of Mathematical Sciences
Florida Atlantic University
777 Glades Road, Boca Raton, FL 33431, USA

Abstract. A novel scheme for securing biometric templates of variable size and order is proposed. The proposed scheme is based on a new similarity measure approach, namely the set intersection, which strongly resembles the methodology used in most of the current state-of-the-art biometrics matching systems. The applicability of the new scheme is compared with that of the existing principal schemes, and it is shown that the new scheme has definite advantages over the existing approaches. The proposed scheme is analyzed both in terms of security and performance.

1 Introduction

Authentication systems based on user's biometric data have several advantages over other authentication methods. The main advantages of biometric-based authentication is the simplicity of use and a limited risk of losing, stealing, or forging users' biological identifiers. On the other hand, the major disadvantage of biometrics-based authentication is the *non-renewability* of biological identifiers. This is a particularly significant issue regarding the identity theft problem, the fastest growing type of fraud in the United States [1]

Biometric-based authentication with the same biometrics is likely to be used in multiple application systems. For example, a fingerprint-based authentication could be used to gain access to multiple systems or facilities. If a biometric template is stolen from a authentication system, criminals can abuse it in the present or future time in multiple venues. In addition, to respect valid privacy concerns by the users, such as corrupt employees at the trusted institutions that have access to a database of biometric templates, the templates should not be stored as plaintext (in its clear form). One solution to the problem is to make use of tamper resistant systems; however, the use of such systems could be infeasible in a given system setup.

Biometric templates often contain condensed discriminatory information about the biometric uniqueness of the user. For instance, in case of fingerprints, the system often stores the discriminatory set of minutiae points. With this information, an adversary can bypass the access control system or extract certain system-specific keys provided that tampering with the system at that level is feasible. In addition, this information could potentially also be used to perform attacks even from the topmost sensory level by creating fake biometric identifiers with the same discriminatory biometric features, a method often referred to as *spoofing*. For instance, given fingerprint minutiae, an attacker can construct a fake fingerprint that has the same discriminatory information as the stolen template. Methods for creating fake fingerprints such as SFINGE by Cappelli, Miao and Maltoni [2] or synthetic generation technique by Araque et al. [3] can be used for exactly that purpose. Uludag and Jain [4] described many attacks on fingerprint-based identification systems using a fake fingerprint such as rubber or silicon finger, and alike. Similar considerations are also applicable to other types of biometrics.

Some biological identifiers are prone to so-called *side-channel attacks*. Side channel attacks refer to attacks on a security system that are performed outside of the system itself. In particular, stealing one's biometric data could be performed by acquiring an imprint of one's fingerprint from a glass, or by acquiring person's face image or voice recording. There are several biological identifiers that are less prone to such attacks as complex equipment is needed for their acquisition (e.g. hand vein pattern). Biometric systems based on such identifiers are still in immaturity; however, they offer tremendous security potential in respect to the side-channel attacks. In practice, even for systems prone to the side-channel attacks, overall security risk to users is still very limited as isolated biometric information cannot fully reproduce one's entire identity. Clearly, in case of a stolen database of users containing their biometric templates as well as additional personal information (such as name, address, phone numbers, email etc), situation is quite the opposite as such information allows for a new, more dangerous form of identity theft.

Another concern is that some biometric data may reveal certain disorders or predisposition to certain disorders, an information which would clearly violate users' privacy. For example, biometric fingerprints and palmprints contain certain genetic information (such as race or presence of Down's syndrome), while retina reveals susceptibility for strokes and diabetes.

Standard cryptographic one-way primitives are not suitable for protecting biometric templates since the biometric identifiers are fuzzy (not exactly reproducible). This fuzziness is an artifact of a limited accuracy of current biometric sensors, such as sensory resolution and 3D-to-2D projection errors, as well as the errors introduced by the feature extraction algorithms that are often used to concisely describe the discriminatory information of the particular biometric. Additionally, if the biometric template consists of a set of elements (as is the case for fingerprint templates), applying a cryptographic one-way function element-wise to conceal the values is also out of the question since the universe of possible point values is typically very small and a simple membership test would suffice to determine the concealed elements.

As a result, one-way transformations that are robust to fuzziness are being studied, and several prominent schemes for storing biometric templates securely were recently proposed. As we shall see, most of the proposed schemes are based on an artificial but mathematically more suitable similarity measures to overcome design difficulties posed by the native similarity measures that are not metrics (e.g. set intersection). We overcome this design difficulty and propose a scheme that relies on the set intersection-based similarity measure which is a natural way of measuring similarity for many biometric templates, including fingerprint minutiae.

The rest of this paper is organized as follows. In Section 2 we present a brief summary of principal work in this area and point out a number of limitations of several state-of-the-art methods for securing biometric templates. In Section 3 we propose a novel approach to securing biometric templates that is based on a novel similarity measure which allows for the practical applicability of the scheme in many real-world scenarios. Security aspects of the proposed scheme are addressed in Section 4. Finally, conclusions and a number of topics for further research are given in Section 5.

2 Related Work

Before describing and analyzing properties of the principal schemes that have been proposed up to date, and also to set the stage for later discussion, several preliminary definitions and concepts are presented next.

2.1 Basic Definitions

The design of a scheme for securing biometric templates is constrained with a type of biometric feature vector that is extracted from the sensory information. Properties of feature vectors representing biometric templates heavily depend on the type of biometric data involved, capability of a sensor, and the corresponding feature extraction algorithm. These properties include the types of errors introduced during data acquisition process, as well as the expected range of values and similarity thresholds.

Typically, two types of biometrics templates (feature vectors) often appear in practice: (1) templates with points that have constant size and order, here denoted by *type I* templates, and (2) templates with points having variable size and order, denoted by *type II* templates. For example, type I biometric templates often appear in face recognition systems where feature vectors are singular value decomposition of a face image, or in iris recognition systems such as IrisCode [5]. Fingerprint and palm print minutiae-based recognition systems, which constitute what are the most common biometric systems [6] work with type II templates. Schemes for securing biometric templates are in general designed for a particular template type.

In terms of application requirements, there are several types of schemes for securing biometric templates. In work by Dodis et al. [7] and [8], two types of schemes are defined:

1. *Secure sketch* – This scheme essentially allows for the precise reconstruction of a noisy input. Given an input x, the scheme produces a public value $f(x)$, called secure sketch, from which no information about x can be deduced (i.e. f is a one-way function). The scheme can recover the original value of x solely from $f(x)$ and y if and only if y is similar to x according to some similarity measure, denoted with $y \sim x$.
2. *Fuzzy extractor* – For a given input x this scheme produces a public value $f(x)$ and a secret value k. Function f is a one-way map so that no information about x can be deduced from $f(x)$. The scheme is able to recover k solely from y and $f(x)$ if and only if $y \sim x$. In practice, k is often used as a secret key for further cryptographic processing.

In [8], it was also shown that it is always possible to construct fuzzy extractors from secure sketches. Intuitively this means that secure sketches comply with a stronger condition (or requirement) than fuzzy extractors do. However, in a number of biometrics-based security applications, even fuzzy extractors comply to a stronger requirement than what suffices in practice.

When concerned with pure verification or identification applications, ability to determine whether a new template matches the stored one is a sufficient requirement. In general, a match is declared when two templates are similar, or, in other words, with similarity measure greater than some threshold t (also referred to as the *similarity bound*). Note that the similarity function is not necessarily a metric. We define a *threshold-based similarity measuring scheme* S to be a scheme that for given one-way transformed value $f(x)$ and a template y determines whether the original template x and y are similar or not:

$$S(f(x), y) = \begin{cases} \textit{similar}, & \text{if } s(x, y) > t; \\ \textit{not similar}, & \text{if } s(x, y) \leq t, \end{cases}$$

where $s(x, y)$ denotes a similarity measure of x and y. Strictly speaking, this kind of scheme is slightly more limited than a scheme that can compute the actual value of $s(x, y)$ from $f(x)$ and y; however, almost all biometrics security systems are based on a threshold similarity measure approach.

It is not too difficult to observe that both secure sketches and fuzzy extractors are also threshold-based similarity measuring schemes. It may be of interest to have schemes which are threshold-based similarity measuring schemes that are strictly not secure sketches.

2.2 Previously Proposed Schemes and Certain Security Considerations

To secure biometric templates of type I, Juels and Wattenberg proposed a scheme called *fuzzy commitment*. This conceptually simple scheme is based on error correcting codes. Let $\mathcal{F}$ be a field, and $\mathcal{C}$ the set of vectors of some t-error correcting code. Let $x \in \mathcal{F}^n$ denote a biometric feature vector. Assuming that all codewords lie in $\mathcal{F}^n$, a codeword c is selected uniformly at random from $\mathcal{C}$ and difference $\epsilon = c - x$ is computed. Next, a suitable one-way function h is selected, and the pair $(\epsilon, h(c))$ is published, representing the output of fuzzy commitment scheme.

To reconstruct the original feature vector x, a similar vector y is required, where the measure of similarity is given by a certain metric. If the usual Hamming distance between $c' = \epsilon + y$ and c is less than t, the error correcting capability of the code $\mathcal{C}$, then it is possible to reconstruct c and consequently x. Since the feature vectors are required to be from $\mathcal{F}^n$, the scheme can be applied only to type I feature vectors, where constant size and order is assumed. Fuzzy commitment is a secure sketch scheme. A scheme based on fuzzy vault principle was constructed and successfully applied for securing a particular type of iris templates, called IrisCode, as described in [5].

Juels and Sudan in [9, 10] proposed a scheme, called *fuzzy vault*, that slightly extends the applicability of a scheme from [11] by allowing for the order invariance of feature vector coordinates. This scheme substantially relies on Reed-Solomon error correcting codes, where the codewords are polynomials over a finite field $\mathcal{F}$. Given a feature vector (set) $x \subset \mathcal{F}$ and a secret value k, a polynomial $p \in \mathcal{F}[X]$ is selected so that it encodes k in some way (e.g., has an embedding of k in its coefficients). Then an evaluation of the elements of x against p is computed and, along with these points, a number of random *chaff* points that do not lie on p is added to a public collection R.

To recover k, a set y similar to x must be presented. If $y \sim x$, then y contains many points that lie on p. Using error correction procedure, it is possible to reconstruct p exactly, and thereby k. If y is not similar to x, it does not overlap substantially with x and thus it is not possible to reconstruct p using the error correction mechanism of Reed-Solomon code. By observing the public value R, it is infeasible to learn k due to the presence of many chaff points. This is also a secure sketch scheme. While fuzzy vault does allow for a variable order, it does require feature vector sizes to be of the fixed length, thus still not fully supporting biometrics feature vectors of type II. Several schemes based on fuzzy vault principle were reported for fingerprint data in [12] and [4].

One of the most serious attacks considered for fuzzy vault-based schemes is the *multiple-use attack* that the original authors did not consider in their security model. Under the multiple-use attack, the adversary has public information obtained from multiple authentication systems regarding user U. The multiple-use attack is successful if it is possible to compromise the secret information about U (in whole or in part) from analyzing the public information about U from multiple systems. Schemes based on fuzzy vault and generally any schemes that are based on the principle of *chaffing and winnowing* [13] are weak against multiple-use attack.

Suppose the same user is enrolled in $k > 1$ authentication systems which are all based on the same kind of biometric (e.g. fingerprint) and which all use the fuzzy vault scheme for securing biometric feature vectors. For simplification, let us assume that the user's biometric feature vector in all systems was $x = \{x_1, \ldots, x_t\}$, since almost the same arguments apply when these vectors are *similar*. Recall that the public information that is stored in system i is a collection $R^{(i)}$ that contains t points $(x_1, p^{(i)}(x_1)), \ldots, (x_t, p^{(i)}(x_t))$ and $m^{(i)}$ chaff points $(r_1^{(i)}, s_1^{(i)}), \ldots, (r_{m^{(i)}}^{(i)}, s_{m^{(i)}}^{(i)})$. According to the fuzzy vault specification chaff points are selected uniformly at random from $\mathcal{U} - x$, where $\mathcal{U}$ denotes the universe of feature vector coordinates. If $R_{\mathrm{x}}^{(i)}$ denotes the restriction of $R^{(i)}$ to the x-axis, then

$$\lim_{k \to \infty} (R_{\mathrm{x}}^{(1)} \cap R_{\mathrm{x}}^{(2)} \cap \ldots \cap R_{\mathrm{x}}^{(k)}) = x$$

unless chaff points always entirely cover the remaining universe $\mathcal{U} - x$ or some fixed parts of it. Moreover, if we take a simple case when $r = |R_{\mathrm{x}}^{(i)}| - t \ll q$ for $i = 1, 2$, then

$$Prob(R_{\mathrm{x}}^{(1)} \cap R_{\mathrm{x}}^{(2)} = x) = \frac{\binom{q-t-r}{r}}{\binom{q-t}{r} + 1} \approx 1,$$

where q denotes the cardinality of set $\mathcal{U}$. In other words, if the number of randomly selected chaff points is much smaller than the size of the universe $\mathcal{U}$, the intersection of chaff points of the same person taken from two authentication systems will almost certainly be empty.

In [9, 10] it is shown that the number of different polynomials that agree on t is small if the size of collection R is small. Thus, in order to ensure security from that point of view, the authors recommend taking a large number of chaff points. Yet, the authors do not *require* to always cover the entire remaining universe $\mathcal{U} - x$ with chaff. Indeed, this is probably infeasible when dealing with larger universes. However, to avoid the multiple-use attack as described here, the entire remaining universe or fixed part of it must be covered by chaff. That is, $R_{\mathrm{x}}^{(i)} = \mathcal{U}'$ for all i where $\mathcal{U}'$ is a subset of $\mathcal{U}$ (likely $\mathcal{U}' = \mathcal{U}$) that provides a large number of polynomials that agree on t points and also a computationally infeasible search space.

In [14], Boyen showed how careless construction of fuzzy extractor from a secure sketch is prone to the multiple-use attack. Namely, using the fuzzy sketch based on the Juels-Wattenberg scheme [11], Boyen derived a construction that is in all respects a fuzzy extractor according to the definition given in [7], yet substantially insecure. It has been shown that sufficient number of calls to Gen', the generating algorithm of so constructed fuzzy extractor, on the same secret input w^*, leads to the complete reconstruction of w^*. In addition, Boyen also showed that Juels-Wattenberg scheme is breakable when used with the *biased* error correcting code, if used multiple times. Here, biased code refers to a nonlinear binary code where, on average over all codewords, the value 0 is more likely to appear than the value 1 at every coordinate of the code space. Finally, Boyen pointed out a third source of potential vulnerability in the abstractions used in generic fuzzy sketches and extractors, such as the permutation based construction. He showed that a poor implementation of a particular abstraction can reveal secret information, if used multiple times.

Thus, in the security model of fuzzy extractors and secure sketches the multiple-use attacks must be addressed and related security risks accounted for, as such attacks are often feasible to launch. Earlier schemes, such as the ones found in [11, 9, 10, 15, 7], did not consider multiple-use attack in their security models.

Dodis et al. in [7, 8] proposed a scheme that allows for securing biometric feature vectors of type II. This scheme, called *PinSketch*, relies on t-error correcting (BCH) code C. In order to simplify description, let us assume H to be a parity check matrix of the code C over some finite field $\mathcal{F}$. For a given feature vector x which belongs to $\mathcal{F}^n$, the scheme computes output $syn(x) = Hx$, which is referred to as the *syndrome* of vector x.

In the reconstruction phase, $syn(y)$ is computed for a given vector y. Let $\delta = syn(x) - syn(y)$. It is easy to see that there exists at most one vector v such that $syn(v) = \delta$ and $weight(v) \leq t$. One of the nice features of binary BCH codes is possibility of computing $supp(v)$ given $syn(v)$ and vice versa, where $supp(v)$ represents the listing of positions where v has nonzero coordinate. Computing of $supp(v)$ for a given $syn(v)$ is the key step in the reconstruction phase. If a distance metric $d(x, y) \leq t$ then $supp(v) = x \triangle y$, and in that case the original set could be reconstructed by $x = y \triangle supp(x)$. PinSketch is a secure sketch scheme that supports biometrics feature vectors of type II.

2.3 Applicability Critique of Error Correcting-Based Schemes for Securing Type II Templates

From the mathematical point view, the most suitable method for measuring *similarity* between two sets is by their symmetric set difference. However, this quite reasonable mathematical choice is often a limitation for practical use. Let us try to illustrate this problem in the case where it is needed to measure closeness between two sets A and B that represent biometric (fingerprint) personal data, of not necessarily different persons. This is an inevitable step in the process of verification or identification. Reconstruction of A, using *similar* set B will be successful if and only if $|A \triangle B| \leq t$, where t is a given parameter that controls the closeness between sets. It seems that error correcting codes are a suitable choice for reconstructing A from a noisy input B. Here, t is the error correcting bound of the chosen code.

We argue that the use of error correcting codes and consequently the Hamming distance as a measure of similarity between type II feature vectors is not an adequate choice. For instance, in the PinSketch scheme [8], templates are represented as characteristic vectors with respect to universe $\mathcal{U}$. Therefore, the symmetric difference is simply related to the Hamming distance between characteristic vectors. In a typical application of PinSketch, such as fingerprint identification, the scheme has a substantial applicability issue. The number of minutiae, according to many statistical analyses of fingerprints lies with high probability in the interval between 20 and 80 [16]. Thus, the choice of the error correcting bound t that is used in this scheme seems to be its main shortcoming.

Considering that the size of the universe is not large, t must be chosen in a way not to compromise security. For instance, if a template set is of size 15, then setting $t > 12$ would not be an adequate choice, since an adversary could test all elements or 2-subsets of the universe (which is feasible for a universe of fingerprint minutiae) and use the error correction to obtain the template set. On the other hand t must be set to provide proper authentication. Due to imperfections in the template extraction it is common to have spurious minutiae and some real minutiae that are not recognized. Thus, symmetric difference between newly presented and stored template could became relatively large, yet the intersection could still be large enough for authentication of B as A with high confidence. For example, suppose $|A| = 20$ and $q \approx 10^6$. Therefore, t could be at most 17. If we accept *twelve point matching rule* as valid, and if $|B| = 22$ and $|A \cap B| = 12$ then B will not be authenticated as A although intersection is large enough to confirm the identity. Even if we do not accept *twelve point matching rule*, it is possible to construct many examples where symmetric difference does not appear as an adequate choice for *similarity* measure. In most minutia-based authentication systems similarity is measured using the number of points that agree in the best possible alignment of two sets of minutiae using translation, rotation and potentially scaling. Therefore, the set intersection is a more appropriate similarity measure in practice.

The authors of fuzzy vault [9, 10] indicated that the scheme is applicable to feature vectors with fixed size and variable ordering which limits the practical use of the scheme to type I vectors. Even if it is possible to extend the fuzzy vault scheme to work with the type II feature vectors, the scheme would face the similar applicability issues since it is based on error correction approach. As an artifact of fuzzy vault where the entire universe is covered by chaff due to multiple use attack and the requirement about the minimal number of different polynomials that agree on t points, the similarity measure is not achieved with symmetric set difference but with ordinary set difference $B - A$. This slightly better scenario is still inappropriate since it is possible to have cases where both $A \cap B$ and $B - A$ are relatively large, in which case the fuzzy vault scheme would give a false rejection.

In this work we design a scalable secure scheme applicable to type II biometric templates, such as fingerprint minutiae which are currently the most common biometric templates [6].

3 The Proposed Approach

Let $\mathcal{F}$ be a finite field of size that is sufficient to provide for computationally infeasible search space. Typically, by modern standards, the size of $\mathcal{F}$ should be at least 2^{80}. We consider biometric templates of type II as subsets of $\mathcal{F}$. Let $\mathcal{U}$ be the union of all biometric templates, and $|\mathcal{U}| = q$. It is common to refer to $\mathcal{U}$ as the *universe* of all template point values.

The key observation is that the size of the universe is typically much larger than the size of a biometric template, but still in a range that allows feasible exhaustive search. For instance, the size of the universe representing fingerprint minutiae is approximately in the range of 10^5-10^7, depending on technical characteristics of the sensor, yet the size of a biometric template is between 20 and 80 with high probability. In further analysis, we will assume $q \gg |A|$, where A represents a template set.

Accuracy of the extraction of biometric data depends on several factors, but mostly on the sensory technology for data acquisition and image processing algorithms for biometric template extraction. Due to these imperfections, it cannot be expected that newly submitted templates perfectly match the stored ones. It is not uncommon to have, under certain scenarios, just part of the fingerprint that needs to be identified. Therefore, a scheme for secure authentication needs to have a necessary level of tolerance with respect to possible incompleteness and inaccuracy of submitted templates. The tolerance threshold for our scheme can be easily customized regarding the particular application.

3.1 Scheme Description

Let m_1 and m_2 be integers such that $m_1 \leq |A| \leq m_2$ for all templates A. Suppose that ℓ is an integer chosen such that $\ell \leq m_1$, and

$$\binom{m_2}{\ell} \leq 2^{k_1} \ll 2^{k_2} \leq \binom{q}{\ell},$$

for some positive integers k_1 and k_2.

In general, it is required for k_1 to be small enough to allow for a feasible search through the set of ℓ-subsets of any given template A. On the other hand, it is required for k_2 to be large enough, making it infeasible to search through all ℓ-subsets of the universe $\mathcal{U}$. As an illustration, if $q \approx 10^6$ and $m_2 = 100$, even with a choice of $\ell = 3$ the size of $\binom{q}{\ell}$ is approximately 2^{60} which is a larger search space than that of DES. For the same parameters, the size of $\binom{m_2}{\ell}$ is just 161700. The generation of public one-way transformation of the given template in the proposed scheme is as follows:

1. Let $A = \{a_1, a_2, \ldots, a_n\}$ be the input biometric template. Randomly choose $s \in \mathcal{F}$. Using an ℓ-out-of-n perfect secret sharing scheme, create n shares of s denoted by $s_1, \ldots, s_n$.
2. Choose a secure cryptographic hash function h and obtain set $\{h(a_1 s), h(a_2 s), \ldots, h(a_n s)\}$, where $a_i s$ means concatenation of s and a_i. It is required that the chosen hash function is both preimage resistant and collision-resistant.
3. Define a discrete function $f_A : \mathcal{U} \to \mathcal{F}$ in the following way

$$f_A(x) = \begin{cases} s_i, \text{ if } x = a_i; \\ r_x, \text{ if } x \notin A, \end{cases}$$

   where the values r_x are chosen uniformly at random.
4. Store $f_A(x)$, $H_A = \{h(a_1 s), h(a_2 s), \ldots, h(a_n s)\}$ and $h(s)$ as a one-way public transformation of A.

The recovery process in our scheme is performed in the following way:

1. For a given set $B = \{b_1, \ldots, b_m\}$, for all ℓ-subsets of B, denoted by B_1, $B_2, \ldots, B_{\binom{m}{\ell}}$, do the following:
   (a) Evaluate $f_A(B_i)$.
   (b) Using the reconstruction method provided by the secret sharing scheme, obtain s' from $f_A(B_i)$.
   (c) Compute $h(s')$; if $h(s') = h(s)$, then assume $s = s'$, compute $H_B = \{h(b_1 s'), \ldots, h(b_m s')\}$, and then output $|H_A \cap H_B| = |A \cap B| \geq \ell$ and terminate.
2. If for all ℓ-subsets of B no termination was reached, output $|A \cap B| < \ell$ and terminate.

In our scheme, s corresponds to the extracted key from the definition of fuzzy extractor. Moreover, with minor modifications the proposed scheme can also be turned into a secure sketch scheme where original set A can be completely reproduced. The algorithm determines a threshold-based similarity of templates A and B using set intersection as a similarity measure, which reflects the same principle used in most minutia-based recognition methods. The algorithm outputs $|A \cap B|$ if $|A \cap B| \geq \ell$. Once $|A \cap B|$ has been obtained, it is to be decided if the authentication threshold has been achieved.

The authentication bound is not substantially involved in our scheme, which is not the case in the previous schemes. The only requirement related to the authentication bound is that it must be greater than or equal to the security bound ℓ.

One drawback of the aforementioned recovery algorithm is its complexity. Namely, the number of ℓ-subsets of probe template B could be significantly large. However, the proposed recovery algorithm can be run probabilistically to accommodate a feasible performance.

3.2 Probabilistic Recovery

In our scheme, for the enrollment template A and a probe B that originates from the same subject as A, we can assume without loss of generality that $|A \cap B| = \lceil t|A| \rceil$ for $t \in (0, 1)$.

Let X be a random variable that describes the number of unsuccessful attempts before getting a *qualified* subset, i.e. a set from $A \cap B$. Clearly X has a negative hypergeometric distribution. If $a^{(b)} = a(a-1)\cdots(a-b+1)$ than the distribution of X is

$$Prob(X = r) = \frac{bw^{(r-1)}}{c^{(r)}}, \tag{1}$$

where $b = \binom{\lceil t|A| \rceil}{\ell}$, $c = \binom{|B|}{\ell}$ and $w = c - b$.

Then, the mathematical expectation of X is given by

$$EX = \frac{c+1}{b+1}. \tag{2}$$

Next, we show some concrete parameters that give a clear view of the computational complexity of the searching process for an ℓ-subset in $A \cap B$. In Table 1 we fix parameter $t = 0.5$, i.e. B contains at least 50% of the points from A. For simplicity, we fix the sizes of A and B to be equal although this is not required by our construction.

Table 1. The expected number of attempts needed to find an ℓ-subset of $A \cap B$ for various sizes of A and B when $t = 0.5$.

t	0.5							
ℓ	10				12			
$\|A\| = \|B\|$	80	60	40	30	80	60	40	30
EX	1943	2510	4588	10002	10784	16179	44351	189679

If we set t to be slightly higher, for example $t = 0.6$, then the expected values significantly change, as depicted in Table 2. For many authentication systems it is not unreasonable to expect that set B, which originates from the same subject as A, have at least 60% common points with A.

Table 2. The expected number of attempts needed to find an ℓ-subset of $A \cap B$ for various sizes of A and B when $t = 0.6$.

t	0.6							
ℓ	10				12			
$\|A\| = \|B\|$	80	60	40	30	80	60	40	30
EX	252	297	433	687	865	1118	2067	4659

Although parameter t is not included in the construction of the scheme, it is useful to have a presumption on the expectation for t. Taking into consideration the particular application and by doing a preliminary statistical analysis on the accuracy of the template extraction system, an estimation for t can be achieved. When higher level of security is required, t generally must be higher. Consequently, it is possible to choose larger ℓ and still have a high efficiency in the task of finding ℓ-subset from $A \cap B$.

For instance, for certain high-security authentication, the threshold of common points between the new and stored template could be set to at least 80% of the stored template set. In that case, even setting $\ell \geq 20$ results in efficient performance of our scheme. Table 3 shows the case when $\ell = 20$.

The probability that the probe template B contains a qualified subset after the number of iterations in the reconstruction algorithm surpasses the expected value EX is clearly decreasing. For example, when $t = 0.6$, $|A| = |B| = 40$, $\ell = 10$, if qualified subset is not found after 5000 iterations, the probability that B contains such subset is less than 2.17×10^{-8}.

Table 3. The expected number of attempts needed to find 20-subset of $A \cap B$ when $t = 0.8$ and $|A| = |B| = n$.

n	30	40	60	80
EX	2828	611	251	181

For every choice of parameters in the proposed scheme, using equation 1, it is possible to select a reasonable bound for the number of iterations that provides negligible false rejection rate (FRR). Thus, it is possible to significantly reduce the number of iterations in the reconstruction algorithm while allowing for a negligible FRR.

4 Security Considerations

If a uniform distribution is assumed, the adversary does not know if a certain subset is more likely the subset of a template than not.

We consider computationally bounded adversarial model. According to this it follows that the size of the adversary search space is equal to $\binom{q}{\ell}$. It is reasonable to hypothesize that adversary does not have computational power that exceeds 2^{80}. Since we accept that the size of the universe is fixed, then the adversary search space depends only on the choice of parameter ℓ. For example, even if q is as small as 10^4, when ℓ is set to 7 then $\binom{q}{\ell} > 2^{80}$. Thus, by changing the parameter ℓ, it is possible to adapt model to the desired security level. It should be noted that by increasing the parameter ℓ, efficiency of searching through ℓ subsets inside of submitted template is decreased. Therefore, a change of parameter ℓ represents the trade off between efficiency and security of the proposed model.

Although the uniform distribution assumption does not correspond to the actual distribution of biometric data in nature, we believe that by increasing the parameter ℓ, it is possible to annulate potential advantages of nonuniform distribution for an adversary.

To address the security of our method, it is essential to discuss issues regarding the distribution of the source data. The attacker's goal is to learn information about the original template A given only the public values $f_A(x)$, H_A and $h(s)$. In a model where an adversary has bounded computational power, finding s from $h(s)$ is not possible due to properties of h and size of the search space. Note that the multiple use attack is not applicable to our scheme since the entire universe $\mathcal{U}$ is covered by uniformly random values according to f_A.

There have been a number of attempts to explain the minutiae distribution. Most recent papers tracking this subject come from the Michigan State University group [17] which mainly dealt with the questions of individuality of fingerprints and how similar two randomly chosen fingerprint templates could be. This problem was partially inspired by a recent challenge to the generally accepted *twelve points matching rule* in some US courts.

The statistical model of distribution of minutiae points has not been established due to very complex nature of the problem. The distribution of minutiae that has been proposed in [17] is a so-called *mixed distribution*. This distribution appears to be more appropriate than the uniform distribution regarding the statistical data collection taken from three large publicly available databases of fingerprints [17]. However, note that all results heavily depend on the quality of acquired fingerprint data and the extraction method used in the experiments.

The result which could be of particular importance for our security model is a result about the probability that two random fingerprint templates of 36 minutiae share more than 12 points. If $P(36, 36, 12)$ denotes this probability and assuming the mixed distribution, it can be shown that $P(36, 36, 12) \approx 6 \times 10^{-7}$. In our scheme, if $\ell = 12$ then an attacker could try to get stored set A of 36 minutiae by choosing a random subset B of 36 elements of the universe $\mathcal{U}$, hoping that $|A \cap B| \geq 12$. However, the only way the attacker can know if the chosen subset B contains more than 12 elements of the stored template A is by running through all 12-subsets of B. Thus, the probability of an attacker's success is $\approx 6 \times 10^{-7} \times \frac{1}{\binom{36}{12}} \approx 2^{-56}$. That makes this kind of attack inefficient especially if we set ℓ to be higher than 12.

We would like to stress that the previously mentioned results are dependent on the effectiveness of the automated minutiae extraction methods which are only of moderate reliability.

It must be understood that the nonuniformity of the universe of certain biometrics influences all proposed schemes regarding security issues. For the schemes based on error correction codes, nonuniformity affects the error correction bound. Consequently, it produces an increase of the false rejection rate (FRR). In our scheme, it induces an increase of the parameter ℓ that causes a higher computational cost.

5 Conclusions

We proposed a novel scheme for securing biometric templates of variable size and order. Unlike previously proposed schemes, our scheme uses set intersection as the similarity measure between the enrollment template and a probe. This principle reflects matching criteria used in most minutia-based authentication systems, and as such offers better applicability than the schemes based on error correcting approach. We showed that the scheme is scalable and has a relaxed dependency on the similarity bound. Finally we demonstrated how to set the parameters of the proposed scheme in order to achieve both high security and broad applicability even when the minutiae distribution is nonuniform.

Acknowledgements

The authors would like to express their gratitude to Prof. Dr. Spyros S. Magliveras and Prof. Dr. Rainer Steinwandt at Florida Atlantic University for their helpful comments and suggestions.

References

1. Elbirt, A.J.: Who are you? how to protect against identity theft. IEEE Technology and Society Magazine (Summer 2005)
2. Cappelli, R., Maio, D., D.Maltoni: Synthetic fingerprint-database generation. In: 16th International Conference on Pattern Recognition. Volume 3. (2002) 744–747
3. Araque, J., Baena, M., Chalela, B., Navarro, D., Vizcaya, P.: Synthesis of fingerprint images. In: 16th International Conference on Pattern Recognition. Volume 2. (2002) 422–425
4. Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.: Biometric cryptosystems: Issues and challenges. IEEE Special Issue on Enabling Security Technologies for Digital Rights Management **92**(6) (2004) 948–960
5. Hao, F., Anderson, R., Daugman, J.: Combining crypto with biometrics effectively. IEEE Transactions on Computers **55**(9) (2006) 1081–1088
6. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer-Verlag (2003)
7. Dodis, Y., Reyzin, L., Smith, A.: Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In: EUROCRYPT 2004, Interlaken, Switzerland. (2004) 523–540
8. Dodis, Y., Ostrovsky, R., Reyzin, L., Smith, A.: Fuzzy extractors: How to generate strong keys from biometrics and other noisy data (April 28, 2006)
9. Juels, A., Sudan, M.: A fuzzy vault scheme. In: IEEE International Symposium on Information Theory (ISIT 2002), Lausanne, Switzerland. (2002)
10. Juels, A., Sudan, M.: A fuzzy vault scheme. Designs, Codes and Cryptography **38**(2) (2006) 237–257
11. Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: ACM Conference on Computer and Communications Security. (1999) 28–36
12. Clancy, T.C., Kiyavash, N., Lin, D.J.: Secure smartcard-based fingerprint authentication. In: ACM SIGMM Workshop on Biometrics Methods and Applications (WBMA '03), ACM Press (2003) 45–52
13. Rivest, R.L.: Chaffing and winnowing: Confidentiality without encryption (April 24, 1998)
14. Boyen, X.: Reusable cryptographic fuzzy extractors. In: ACM Conference on Computer and Communications Security (CCS 2004), New-York: ACM Press (2004) 82–91
15. Sutcu, Y., Sencar, H.T., Memon, N.: A secure biometric authentication scheme based on robust hashing. In: ACM Multimedia and Security Workshop (MM&Sec'05), New York, NY. (2005) 111–116
16. Amengual, J., Juan, A., Pérez, J., Prat, F., Sáez, S., Vilar, J.: Real-time minutiae extraction in fingerprint images. In: International Conference on Image Processing and Its Applications (IPA97). Volume 2. (1997) 871–875
17. Dass, S.C., Zhu, Y., Jain, A.K.: Statistical models for assessing the individuality of fingerprints. In: IEEE Workshop on Automatic Identification Advanced Technologies (AUTOID '05), IEEE Computer Society (2005) 3–9