Binary Biometrics: An Analytic Framework to Estimate the Bit Error Probability under Gaussian Assumption

E.J.C. Kelkboom, G. Garcia Molina, T.A.M. Kevenaar, R.N.J. Veldhuis and W. Jonker

Abstract— **In recent years the protection of biometric data has gained increased interest from the scientific community. Methods such as the helper data system, fuzzy extractors, fuzzy vault and cancellable biometrics have been proposed for protecting biometric data. Most of these methods use cryptographic primitives and require a binary representation from the real-valued biometric data. Hence, the similarity of biometric samples is measured in terms of the Hamming distance between the binary vector obtained at the enrolment and verification phase. The number of errors depends on the expected error probability P_{e} of each bit between two biometric samples of the same subject. In this paper we introduce a framework for analytically estimating P_{e} under the assumption that the within- and between-class distribution can be modeled by a Gaussian distribution. We present the analytic expression of P_{e} as a function of the number of samples used at the enrolment (N_{e}) and verification (N_{v}) phases. The analytic expressions are validated using the FRGC v2 and FVC2000 biometric databases.**

I. Introduction

With the increased popularity of biometrics and its application in society, privacy concerns are being raised by privacy protection watchdogs. This has stimulated research into methods for protecting the biometric data in order to mitigate these privacy concerns. Numerous methods such as the *helper data system* [9], [10], [11], *fuzzy extractors* [2], [5], *fuzzy vault* [8] and *cancellable biometrics* [15] have been proposed for transforming the biometric data in such a way that the privacy is safeguarded. Several of these privacy or template protection techniques use some cryptographic primitives (e.g. hash functions) and error correcting codes (ECC) and require a binary representation of the biometric sample, referred to as the *binary vector*.

Fig. 1 shows a high level overview of a biometric system that extracts a binary vector from a biometric sample, e.g. a fingerprint image. In the enrolment phase, where the subject presents itself to the biometric system, a biometric sample is obtained and sent to the feature extraction module. The biometric sample is preprocessed (enhancement, alignment, etc.) and a real-valued *feature vector* $\mathbf{f}^{\mathrm{e}}_{\mathrm{R}} \in \mathbb{R}^{N_{\mathrm{F}}}$ is extracted, where N_{F} is the number of feature components. In the verification phase, another biometric sample is taken from which its feature vector $\mathbf{f}^{\mathrm{v}}_{\mathrm{R}}$ is extracted. In a classical biometric system, the matcher would base its decision on the similarity between the feature vectors $\mathbf{f}^{\mathrm{e}}_{\mathrm{R}}$ and $\mathbf{f}^{\mathrm{v}}_{\mathrm{R}}$. Because of the binary vector requirement, the real-valued feature vectors are quantized, i.e. bits are extracted from each feature component, obtaining the binary vectors $\mathbf{f}^{\mathrm{e}}_{\mathrm{B}}$ and $\mathbf{f}^{\mathrm{v}}_{\mathrm{B}}$. The quantization process turns into a binarization process when only a single bit is extracted from each real-valued component of the biometric sample. In the literature, various quantization and binarization schemes have been presented [3], [4], [9], [10]. In this paper we focus on the binarization scheme based on thresholding, which is used in the helper data system schemes [9], [10]. When multiple samples are taken, a feature vector is extracted from each sample and sent to the quantization block, which quantizes the average feature vector.

The transition from a real-valued to a binary representation of the biometric sample according to [9], [10] implies that the similarity between two biometric samples can be measured in terms of the Hamming distance, i.e. the number of bit errors between the binary vectors. The number of bit errors depends on the probability of each bit to change between two biometric samples of the same subject. Each subject will have a different error probability and we are interested in the average error probability seen over the whole population, referred to as the *expected error probability* P_{e}. Because the classification performance of a biometric system depends on the P_{e} of each component, the performance could be estimated if we can estimate P_{e}. In this paper we introduce a framework for analytically estimating the P_{e} of each bit in the threshold-based binarization scheme under the assumption that the real-valued features are distributed according to Gaussian models characterized by the within-class variance σ^2_{w} and the between-class variance σ^2_{b}. The Gaussian assumption is used as the basis of our analytic framework, because due to the central limit theorem we can assume that the real-valued features will tend to approximate a Gaussian distribution when they are obtained by a linear

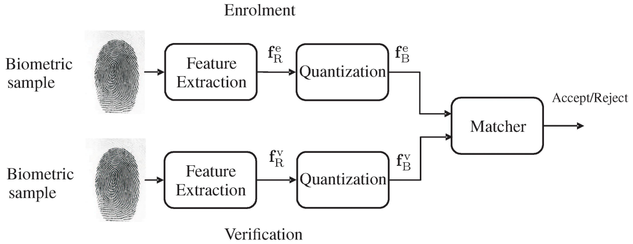

Enrolment
Biometric sample
Feature Extraction
$\mathbf{f}^{\mathrm{e}}_{\mathrm{R}}$
Quantization
$\mathbf{f}^{\mathrm{e}}_{\mathrm{B}}$
Matcher
Accept/Reject
Biometric sample
Feature Extraction
$\mathbf{f}^{\mathrm{v}}_{\mathrm{R}}$
Quantization
$\mathbf{f}^{\mathrm{v}}_{\mathrm{B}}$
Verification

Fig. 1. A high level overview of a biometric system where binary vectors are extracted.

E.J.C. Kelkboom, G. Garcia Molina, T.A.M. Kevenaar, and W. Jonker are with Philips Research, The Netherlands {Emile.Kelkboom, Gary.Garcia, Tom.Kevenaar, Willem.Jonker}@philips.com

R.N.J. Veldhuis is with the University of Twente, Fac. EEMCS, The Netherlands R.N.J.Veldhuis@utwente.nl

978-1-4244-2730-7/08/$25.00 ©2008 IEEE

combinations of many components, e.g. feature extraction techniques based on the principle component analysis (PCA) or linear discriminant analysis (LDA). This assumption is motivated and made plausible in Section IV. Secondly, the Gaussian assumption makes it possible to obtain an analytical closed-form expression for $P_{\rm e}$. PCA or LDA techniques are often being used to perform dimension reduction in order to prevent overfitting or to simplify the classifier [7]. In the field of template protection, PCA is also used to uncorrelated the features to guarantee uniformly distributed keys extracted from the biometric sample [2]. *Our objective is to obtain an analytical closed-form expression for $P_{\rm e}$ as a function of the number of enrolment $N_{\rm e}$ and verification $N_{\rm v}$ samples given the ratio $\sigma_{\rm b}/\sigma_{\rm w}$.*

This paper is organized as follows, we first present the real-valued Gaussian model assumption of the biometric distribution and the binarization method in Section II. Using this model, we formulate the analytic expression of $P_{\rm e}$ in Section III for three cases, namely (i) the known-reference-template case, (ii) the single sample case for both the enrolment and verification phase, and (iii) the multiple sample case. In Section IV we validate these analytic expressions with two different real biometric databases consisting of FRGC v2 3D face images [13] and the FVC2000 fingerprint images [12]. We finalize with the conclusions in Section V.

II. Distribution Model Assumption and Binarization Method

We assume that over the whole population each component of the real-valued biometric sample $\mathbf{f}_{\rm R}$ has a Gaussian distribution $N(\mu_{\rm t}, \sigma_{\rm t}^2)$ with mean $\mu_{\rm t}$ and variance $\sigma_{\rm t}^2$ where t stands for total distribution, see Fig. 2. The total distribution is a combination of the within- and between-class distributions, which we assume to be Gaussians $N(\mu_{{\rm w}_i}, \sigma_{{\rm w}_i}^2)$ and $N(\mu_{\rm b}, \sigma_{\rm b}^2)$, respectively. The within-class distribution characterizes the variability of multiple biometric samples of subject i, whose mean is $\mu_{{\rm w}_i}$ with variance $\sigma_{{\rm w}_i}^2$. The between-class distribution is the probability density function (pdf) of the means $\mu_{{\rm w}_i}$ of all subjects. For simplicity but without loss of generality we consider $\mu_{\rm t} = \mu_{\rm b} = 0$. We further assume that the within-class variance is the same for each subject, i.e. $\sigma_{{\rm w}_i} = \sigma_{\rm w}$. Henceforth, the subject subindices (i) are omitted for notation convenience unless stated otherwise.

To binarize the components of $\mathbf{f}_{\rm R}$, we use the thresholding method [9], [10], in which the threshold δ is equal to the mean of the between-class distribution $\mu_{\rm b}$. If the value of a component of $\mathbf{f}_{\rm R}$ is smaller than δ then it is set to "0" otherwise it is set to "1", see Fig. 2.

III. Analytic Expression for the Error Probability

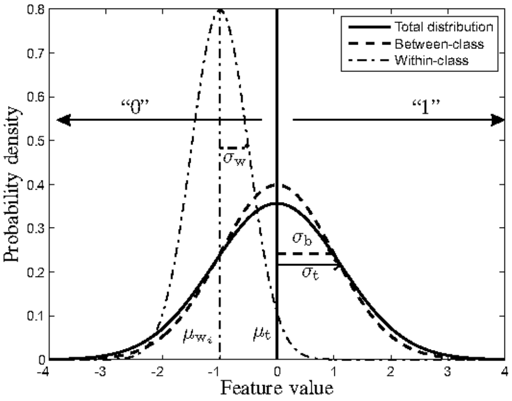


Fig. 2. Distribution model of a single component of the real-valued biometric.

Using the distribution model defined in Section II the expected error probability $P_{\rm e}$ over the whole population is defined by

$$\begin{aligned} P_{\rm e} &= E\left[P_{\rm e}(\mu_{\rm w})\right], \\ &= \int_{-\infty}^{\infty} p_{\rm b}(\mu_{\rm w}) P_{\rm e}(\mu_{\rm w})\,{\rm d}\mu_{\rm w}, \end{aligned} \tag{1}$$

where $P_{\rm e}(\mu_{\rm w})$ is the *error probability* given $\mu_{\rm w}$ and $p_{\rm b}$ is the pdf of the between-class distribution. With the binarization threshold $\delta = \mu_{\rm b} = 0$, this problem becomes symmetric with respect to δ. Consequently, (1) becomes

$$\begin{aligned} P_{\rm e} &= 2\int_{-\infty}^{0} p_{\rm b}(y) P_{\rm e}(y)\,{\rm d}y, \\ &= 2\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma_{\rm b}} {\rm e}^{-\left(\frac{y}{\sqrt{2}\sigma_{\rm b}}\right)^2} P_{\rm e}(y)\,{\rm d}y, \\ &= \frac{2\lambda}{\sqrt{\pi}} \int_{-\infty}^{0} {\rm e}^{-(\lambda y)^2} P_{\rm e}(y)\,{\rm d}y, \end{aligned} \tag{2}$$

where $\lambda = \frac{1}{\sqrt{2}\sigma_{\rm b}}$. In subsequent subsections, the integral is solved by defining $P_{\rm e}(y)$ for three different cases:

i Known reference template with a single verification sample,
ii Single enrolment and verification sample,
iii Multiple enrolment and verification samples.

These cases are related to each other as will be explained in Section III-C. The known reference template case has a simple intuitive solution and serves as a framework for solving the other two cases. The second case is an extension of the first one when a single sample is used at the enrolment and verification phase. In the third case, the final analytic expression of $P_{\rm e}$ is obtained by extending the single sample case to the multiple sample case.

A. Known Reference Template

In this case, the reference template obtained in the enrolment phase is assumed to be known and therefore $\mu_{\rm w}$ is known. Hence, the individual error probability for the known-reference-template case $P_{\rm e}^{\rm r}$ is the probability that the sample, in the verification phase, is on the other side of the threshold δ compared to $\mu_{\rm w}$. This probability is depicted by

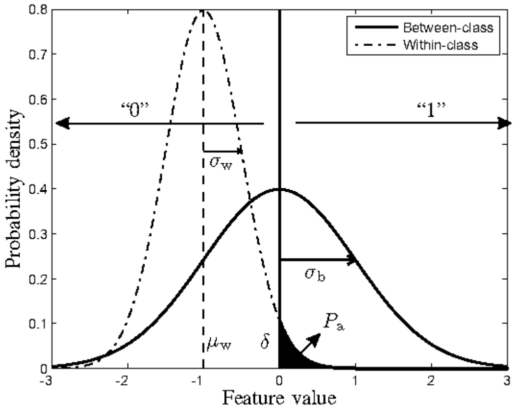


Fig. 3. The error probability $P_{\rm e}^{\rm r}(y)$ given by $P_{\rm a}(y)$.

the shadowed area in Fig. 3 and referred to as $P_{\rm a}$, where

$$P_{\rm e}^{\rm r}(y) \quad = P_{\rm a}(y) = \int\limits_{\delta=0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_{\rm w}} {\rm e}^{-\left(\frac{x-y}{\sqrt{2}\sigma_{\rm w}}\right)^2} {\rm d}x. \tag{3}$$

The easiest way to solve (2) is to write $P_{\rm a}(y)$ in terms of the error function

$$\texttt{erf}(z) \quad = \frac{2}{\sqrt{\pi}} \int\limits_0^z {\rm e}^{-t^2}\,{\rm d}t. \tag{4}$$

By defining $\eta = \frac{1}{\sqrt{2}\sigma_{\rm w}}$, $P_{\rm a}(y)$ can be rewritten as

$$\begin{aligned} P_{\rm a}(y) \quad &= \frac{\eta}{\sqrt{\pi}} \int\limits_0^{\infty} {\rm e}^{-(\eta(x-y))^2} {\rm d}x, \\ &= \frac{1}{\sqrt{\pi}} \int\limits_{-\eta y}^{\infty} {\rm e}^{-z^2} {\rm d}z, \text{ with } z = \eta(x-y), \\ &= \frac{1}{\sqrt{\pi}} \left[\int\limits_0^{\infty} {\rm e}^{-z^2} {\rm d}z - \int\limits_0^{-\eta y} {\rm e}^{-z^2} {\rm d}z \right], \text{for } y \leq 0, \\ &= \frac{1}{\sqrt{\pi}} \left[\frac{\sqrt{\pi}}{2} - \frac{\sqrt{\pi}}{2} \texttt{erf}(-\eta y) \right], \\ &= \frac{1}{2} \left[1 - \texttt{erf}(-\eta y) \right]. \end{aligned} \tag{5}$$

By using $P_{\rm e}^{\rm r}(y) = P_{\rm a}(y)$ and substituting (5) into (2) yields

$$\begin{aligned} P_{\rm e}^{\rm r} \quad &= \frac{\lambda}{\sqrt{\pi}} \int\limits_{-\infty}^{0} {\rm e}^{-(\lambda y)^2} \left[1 - \texttt{erf}(-\eta y) \right] {\rm d}y, \\ &= \frac{\lambda}{\sqrt{\pi}} \int\limits_0^{\infty} {\rm e}^{-(\lambda y)^2} \left[1 - \texttt{erf}(\eta y) \right] {\rm d}y, \\ &= \frac{\lambda}{\sqrt{\pi}} \left[\int\limits_0^{\infty} {\rm e}^{-(\lambda y)^2} - \int\limits_0^{\infty} {\rm e}^{-(\lambda y)^2} \texttt{erf}(\eta y) \right] {\rm d}y, \\ &= \frac{\lambda}{\sqrt{\pi}} \left[\frac{\sqrt{\pi}}{2\lambda} - \int\limits_0^{\infty} {\rm e}^{-\lambda^2 y^2} \texttt{erf}(\eta y) \right] {\rm d}y, \end{aligned} \tag{6}$$

where we used the known result $\int_0^\infty \lambda {\rm e}^{-(\lambda y)^2}\,{\rm d}y = \frac{\sqrt{\pi}}{2}$. The second term with the `erf` function can be rewritten using the general solution of `erf` integrals given as [14]

$$\int\limits_0^{\infty} {\rm e}^{-px^2} \texttt{erf}(ax)\texttt{erf}(bx)\,{\rm d}x = \frac{\arctan\left(\frac{ab}{\sqrt{p(a^2+b^2+p)}}\right)}{\sqrt{p\pi}}. \tag{7}$$

Using (7) and by setting $p = \lambda^2$, $a = \eta$, and $b = \infty$, (6) can be simplified to

$$\begin{aligned} P_{\rm e}^{\rm r} \quad &= \frac{1}{2} - \frac{\lambda}{\sqrt{\pi}} \frac{\arctan\left(\frac{\eta}{\lambda}\right)}{\lambda\sqrt{\pi}}, \\ &= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\eta}{\lambda}\right), \\ &= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\sigma_{\rm b}}{\sigma_{\rm w}}\right). \end{aligned} \tag{8}$$

B. Single Enrolment and Verification Sample

The previous case is now extended by taking into account that the sample obtained in the enrolment phase could also be on the other binarization side of $\mu_{\rm w}$. We restrict this problem to the single sample case for both the enrolment and verification phase, i.e. $N_{\rm e} = N_{\rm v} = 1$. Hence, the individual error probability for the single sample case $P_{\rm e}^{\rm s}$ becomes

$$P_{\rm e}^{\rm s}(y) = (1 - P_{\rm a}(y))P_{\rm a}(y) + P_{\rm a}(y)(1 - P_{\rm a}(y)), \tag{9}$$

where the first term on the right side is the probability of the sample to be on the same binarization side as $\mu_{\rm w}$ in the enrolment phase multiplied by the probability of being on the other side in the verification phase. The second term is the probability for the sample to be on the other side in the enrolment phase multiplied by the probability to be on the same side in the verification phase. Since both terms are equal, (2) becomes

$$P_{\rm e}^{\rm s} \quad = \frac{4\lambda}{\sqrt{\pi}} \int\limits_{-\infty}^{0} {\rm e}^{-(\lambda y)^2} \left[P_{\rm a}(y)(1 - P_{\rm a})(y) \right] {\rm d}y. \tag{10}$$

Substituting (5) into (10) yields

$$\begin{aligned} P_{\rm e}^{\rm s} \quad &= \frac{\lambda}{\sqrt{\pi}} \int\limits_{-\infty}^{0} {\rm e}^{-(\lambda y)^2} \left[1 - \texttt{erf}^2(-\eta y) \right] {\rm d}y, \\ &= \frac{\lambda}{\sqrt{\pi}} \int\limits_0^{\infty} {\rm e}^{-(\lambda y)^2} \left[1 - \texttt{erf}^2(\eta y) \right] {\rm d}y, \\ &= \frac{1}{2} - \frac{\lambda}{\sqrt{\pi}} \int\limits_0^{\infty} {\rm e}^{-\lambda^2 y^2} \texttt{erf}^2(\eta y)\,{\rm d}y, \end{aligned} \tag{11}$$

where the integration of the $\texttt{erf}^2$ function can be solved using (7) with $p = \lambda^2$, and $a = b = \eta$. The analytic expression of $P_{\rm e}^{\rm s}$ becomes

$$\begin{aligned} P_{\rm e}^{\rm s} \quad &= \frac{1}{2} - \frac{\lambda}{\sqrt{\pi}} \frac{\arctan\left(\frac{\eta^2}{\sqrt{\lambda^2(\eta^2+\eta^2+\lambda^2)}}\right)}{\lambda\sqrt{\pi}}, \\ &= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\eta^2}{\lambda\sqrt{2\eta^2+\lambda^2}}\right), \\ &= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\eta}{\lambda\sqrt{2+\left(\frac{\lambda}{\eta}\right)^2}}\right), \\ &= \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{\sigma_{\rm b}}{\sigma_{\rm w}\sqrt{2+\left(\frac{\sigma_{\rm b}}{\sigma_{\rm w}}\right)^{-2}}}\right). \end{aligned} \tag{12}$$

C. Multiple Enrolment ($N_{\rm e}$) and Verification ($N_{\rm v}$) Samples

In this section, the analytic expression in (12) is further extended by considering $N_{\rm e}$ samples in the enrolment and $N_{\rm v}$ in the verification phase. The effect of taking the average of multiple samples is that the variance of the within-class $\sigma_{\rm w}$ decreases according to

$$\sigma_{{\rm w},N}^2 = \frac{\sigma_{\rm w}^2}{N} \Rightarrow \sigma_{{\rm w},N} = \frac{\sigma_{\rm w}}{\sqrt{N}}. \tag{13}$$

The definition of the individual error probability is analogous to the single sample case by taking N_e and N_v into account

$$P_\mathrm{e}^\mathrm{m}(y;N_\mathrm{e},N_\mathrm{v}) = \begin{aligned}[t] &(1-P_\mathrm{a}(y;N_\mathrm{e}))P_\mathrm{a}(y;N_\mathrm{v}) \\ &+ P_\mathrm{a}(y;N_\mathrm{e})(1-P_\mathrm{a}(y;N_\mathrm{v})),\end{aligned} \tag{14}$$

where the error probability given by the area in Fig. 3 now depends on the number of samples as

$$\begin{aligned} P_\mathrm{a}(y;N) &= \int_0^\infty \frac{\sqrt{N}}{\sqrt{2\pi}\sigma_\mathrm{w}} e^{-\left(\frac{\sqrt{N}(x-y)}{\sqrt{2\pi}\sigma_\mathrm{w}}\right)^2} \mathrm{d}x, \\ &= \frac{1}{2}\left[1-\mathtt{erf}(-\sqrt{N}\eta y)\right]. \end{aligned} \tag{15}$$

Hence, (14) can be expanded into

$$\begin{aligned} P_\mathrm{e}^\mathrm{m}(y;N_\mathrm{e},N_\mathrm{v}) =& \tfrac{1}{4}\left[(1+\mathtt{erf}(-\eta_\mathrm{e}y))(1-\mathtt{erf}(-\eta_\mathrm{v}y))\right. \\ &\left.+ (1-\mathtt{erf}(-\eta_\mathrm{e}y))(1+\mathtt{erf}(-\eta_\mathrm{v}y))\right], \\ =& \tfrac{1}{2}\left[1-\mathtt{erf}(-\eta_\mathrm{e}y)\mathtt{erf}(-\eta_\mathrm{v}y)\right], \end{aligned} \tag{16}$$

where $\eta_\mathrm{e} = \sqrt{N_\mathrm{e}}\eta$ and $\eta_\mathrm{v} = \sqrt{N_\mathrm{v}}\eta$. The total error probability for multiple samples is obtained by substituting (16) into (2) as

$$\begin{aligned} P_\mathrm{e}^\mathrm{m}(N_\mathrm{e},N_\mathrm{v}) =& \frac{\lambda}{\sqrt{\pi}} \int_{-\infty}^{0} \mathrm{e}^{-(\lambda y)^2} \left[1-\mathtt{erf}(-\eta_\mathrm{e}y)\mathtt{erf}(-\eta_\mathrm{v}y)\right] \mathrm{d}y, \\ =& \frac{\lambda}{\sqrt{\pi}} \int_{0}^{\infty} \mathrm{e}^{-(\lambda y)^2} \left[1-\mathtt{erf}(\eta_\mathrm{e}y)\mathtt{erf}(\eta_\mathrm{v}y)\right] \mathrm{d}y, \\ =& \frac{1}{2} - \frac{\lambda}{\sqrt{\pi}} \int_{0}^{\infty} \mathrm{e}^{-\lambda^2 y^2} \mathtt{erf}(\eta_\mathrm{e}y)\mathtt{erf}(\eta_\mathrm{v}y)\, \mathrm{d}y, \end{aligned} \tag{17}$$

which can be solved with the use of (7) with $p=\lambda^2$, $a=\eta_\mathrm{e}$, and $b=\eta_\mathrm{v}$ as

$$\begin{aligned} P_\mathrm{e}^\mathrm{m}(N_\mathrm{e},N_\mathrm{v}) =& \tfrac{1}{2} - \tfrac{\lambda}{\sqrt{\pi}} \frac{\arctan\left(\frac{\eta_\mathrm{e}\eta_\mathrm{v}}{\sqrt{\lambda^2\left(\eta_\mathrm{e}^2+\eta_\mathrm{v}^2+\lambda^2\right)}}\right)}{\lambda\sqrt{\pi}}, \\ =& \tfrac{1}{2} - \tfrac{1}{\pi}\arctan\left(\frac{\eta\sqrt{N_\mathrm{e}N_\mathrm{v}}}{\lambda\sqrt{N_\mathrm{e}+N_\mathrm{v}+\left(\frac{\lambda}{\eta}\right)^2}}\right), \\ =& \tfrac{1}{2} - \tfrac{1}{\pi}\arctan\left(\frac{\sigma_\mathrm{b}\sqrt{N_\mathrm{e}N_\mathrm{v}}}{\sigma_\mathrm{w}\sqrt{N_\mathrm{e}+N_\mathrm{v}+\left(\frac{\sigma_\mathrm{b}}{\sigma_\mathrm{w}}\right)^{-2}}}\right). \end{aligned} \tag{18}$$

The relationship between the three cases is now evident. For $N_\mathrm{e} \to \infty$ and $N_\mathrm{v} = 1$, the solution of the multiple sample case (18) converges to the solution of the known-reference-template case (8). Indeed, since $N_\mathrm{e} \to \infty$, the position of μ_w with respect to the binarization threshold is precisely known. Note that the same convergence occurs when $N_\mathrm{e} = 1$ and $N_\mathrm{v} \to \infty$. In addition, for $N_\mathrm{e} = N_\mathrm{v} = 1$ the multiple sample case (18) becomes the single sample case (12).

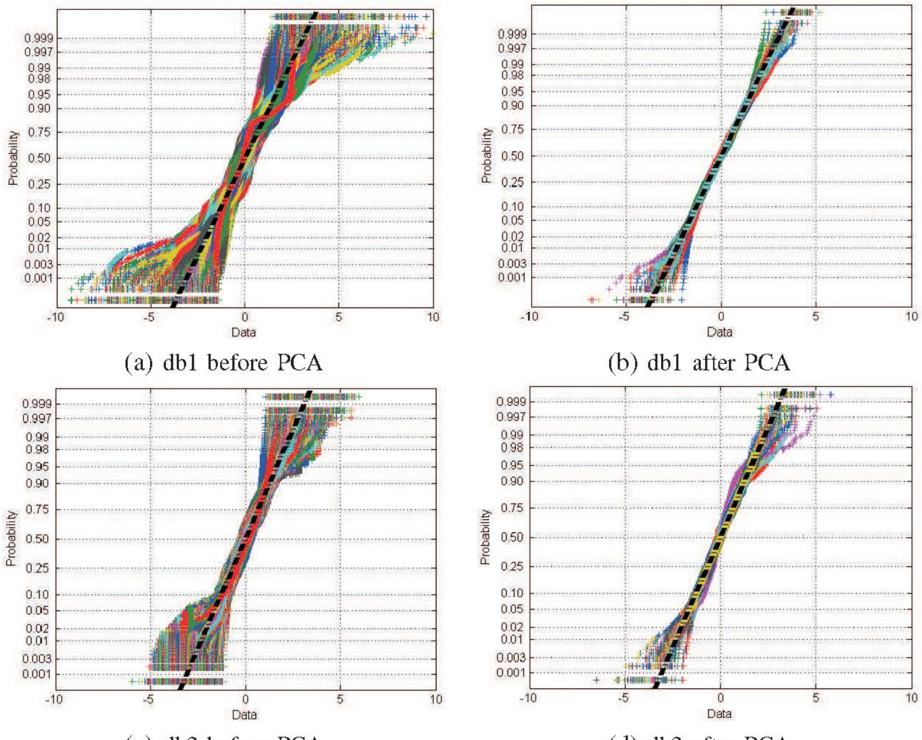
IV. Experimental Validation

In this section, the analytic expressions and the effect of the assumption of Gaussian distributions are validated using two real biometric databases. The first one (db1) consists of 3D face images from the FRGC v2 database [13], where we used the shape-based 3D face recognizer of [6] to extract feature vectors of dimension $N_\mathrm{F} = 696$. Subjects with at least 8 samples are selected resulting into 230 subjects with a total of 3147 samples. The second database (db2) consists of fingerprint images from Database 2 of FVC2000 [12], and uses the feature extraction algorithm based on Gabor filters and directional fields [1] where $N_\mathrm{F} = 1536$. There are 110 subjects with 8 samples each. We applied the PCA dimension reduction technique on each database. We computed the PCA transformation matrix by using all the samples of the database, and we reduced the dimension of the feature vectors until the optimum performance in terms of the equal error rate (EER) was obtained. The EER is defined at the point where the false acceptance rate (FAR) equals the false rejection rate (FRR). For db1, the optimal reduced dimension is at 88 components while 83 for db2.

As described in Section II, the analytic framework is based on the Gaussian model assumption. Fig. 4(a)(c) show the normal probability plot of each component of the feature vectors of db1 and db2 respectively, before PCA has been applied. The normal probability plot is a graphical technique for assessing whether or not a data set is approximately Gaussian distributed. Prior to comparing, we normalized each component to a unit variance and subtracted its mean to make it zero-mean. For both databases it is evident that the distributions are not Gaussian, because they significantly deviate from the dashed-dotted line that depicts a perfect Gaussian distribution. Fig. 4 (b)(d) depict the normal probability plot of the components of db1 and db2 respectively, after applying PCA. For both databases the figures show that after applying PCA the features tend to behave more like Gaussians. Nevertheless, the tails of the distribution deviate the most from being Gaussian.

To assess the model assumptions, we compare the estimated bit error probability of the feature components of the biometric database $\hat{P}_\mathrm{e}^\mathrm{db}$ with the corresponding analytically obtained P_e^m given by (18). The test protocol per feature component is as follows: $\hat{P}_\mathrm{e}^\mathrm{db}$ is calculated as the average over the error probability P_{e_i} of each subject. The subject error probability P_{e_i} is calculated by performing 200 matches and determining the relative number of errors. For each match, N_e distinct features are randomly selected, averaged and binarized (enrolment phase). The obtained bit is compared against bit obtained from averaging and binarizing N_v different randomly selected features of the same subject (verification phase). To evaluate P_e^m, the parameters σ_b, σ_w, N_e and N_v are needed for each component. The estimated standard deviation of the between-class distribution $\hat{\sigma}_\mathrm{b}$ is calculated as the standard deviation over the average feature vector of all subjects. For the within-class distribution $\hat{\sigma}_\mathrm{w}$ is computed as the average standard deviation of each subject.

The comparison between $\hat{P}_\mathrm{e}^\mathrm{db}$ and P_e^m is shown in Fig. 5 for both db1 and db2. Fig. 5(a)(d) show the comparison for the features prior to PCA for the $N_\mathrm{e} = N_\mathrm{v} = 1$ case, while Fig. 5(b)(e) show the features after PCA for the $N_\mathrm{e} = N_\mathrm{v} = 1$ case, and Fig. 5(c)(f) $N_\mathrm{e} = N_\mathrm{v} = 3$ show the features after PCA for the $N_\mathrm{e} = N_\mathrm{v} = 3$ case. In all figures, three lines are given. The solid line represents the average bit error probability as a function of $\sigma_\mathrm{b}/\sigma_\mathrm{w}$ according to the analytical model of Section II with known Gaussian distributions. From these Gaussian distributions, for a given $\sigma_\mathrm{b}/\sigma_\mathrm{w}$, two synthetic databases db1* and db2* are drawn with the same number of



(a) db1 before PCA

(b) db1 after PCA

(c) db2 before PCA

(d) db2 after PCA

Fig. 4. Normal probability plot of each feature vector component of db1 and db2 before and after applying PCA.

samples as db1 and db2, respectively (i.e. db1* contains 230 individuals and db2* contains 110 individuals with 8 samples each). For each subject, we randomly generate μ_{w_i} according to $N(\mu_{\mathrm{b}}, \sigma_{\mathrm{b}}^2)$ and generated 8 samples from $N(\mu_{\mathrm{w}_i}, \sigma_{\mathrm{w}}^2)$. For each synthetic database db1* and db2*, the average bit error probabilities are estimated using the same test protocol as used to estimate $\hat{P}_{\mathrm{e}}^{\mathrm{db}}$. For the same value of σ_{b} and σ_{w}, 500 synthetic databases are generated which allows us to estimate the distribution of the average error probability for a given $\sigma_{\mathrm{b}}/\sigma_{\mathrm{w}}$ as well as the 95 percentile interval of the distribution. By repeating this process for a range of $\sigma_{\mathrm{b}}/\sigma_{\mathrm{w}}$ values, the overall 95 percentile area can be determined which is indicated by the dashed and dashed-dotted lines. This indicates the area where 95% of the estimated average bit error probabilities using the databases db1 and db 2 *would* fall if they would satisfy the assumptions of the analytical model. Because of its smaller size, db2 has a larger 95 percentile area than db1.

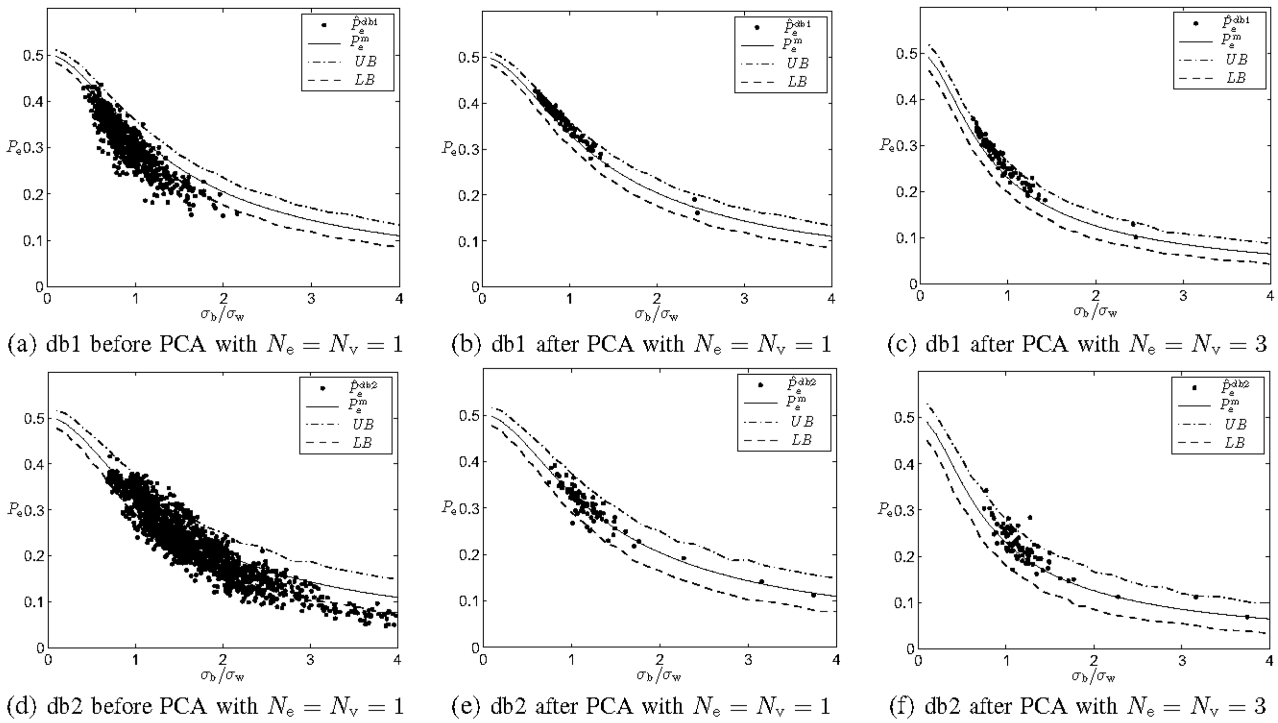
Fig. 5 shows that there is a smaller difference between $P_{\mathrm{e}}^{\mathrm{m}}$ and $\hat{P}_{\mathrm{e}}^{\mathrm{db}}$ when PCA has been applied, which strengthens the assumption that by applying PCA the features will tend to approximate a Gaussian distribution on which this analytic framework is based. When PCA has been applied, in most cases $\hat{P}_{\mathrm{e}}^{\mathrm{db}}$ is estimated within the 95 percentile boundaries. In the $N_{\mathrm{e}} = N_{\mathrm{v}} = 1$ case, the $\hat{P}_{\mathrm{e}}^{\mathrm{db}}$ of three features for both db1 and db2 slightly fall outside the boundaries, which is well within the 95 percent of 88 and 85 respectively. In the $N_{\mathrm{e}} = N_{\mathrm{v}} = 3$ case, there are six cases outside the boundary for db1 and four for db2, where four would be allowed by the 95 percent boundary.

Possible causes for the observed deviations are (i) the limited size of the database which also has an influence on the estimation error of $\hat{\sigma}_{\mathrm{b}}$ and $\hat{\sigma}_{\mathrm{w}}$, (ii) the Gaussian assumption of the feature vector distribution, and (iii) the outliers in the biometric samples of the databases.

V. Conclusions

We have presented a framework for analytically estimating the bit error probability when comparing binary vectors extracted in the enrolment and the verification phase. Having the error probability of each bit, the expected Hamming distance between the enrolment and verification binary vectors can be determined. Hence, a bound on the performance of the biometric system can be formulated. We focused on formulating the analytical error probability for the thresholding binarization method under the assumption that the real-valued biometric sample is distributed according to Gaussian models characterized by the within-class variance σ_{w}^2 and the between-class variance σ_{b}^2.

We derived the analytic expression of P_{e} for three cases, namely (i) the known reference template with a single verification sample $P_{\mathrm{e}}^{\mathrm{r}}$, (ii) the single enrolment and verification sample $P_{\mathrm{e}}^{\mathrm{s}}$, and (iii) the multiple enrolment and verification sample $P_{\mathrm{e}}^{\mathrm{m}}$. The latter case has resulted in a general analytical expression from which case (i) and (ii) can be derived. We validated the analytic expressions using two real biometric database: i) 3D face images from a subset of the FRGC v2 database with 230 subjects and a total of 3147 samples, and ii) fingerprint images from the FVC2000



(a) db1 before PCA with $N_e = N_v = 1$ (b) db1 after PCA with $N_e = N_v = 1$ (c) db1 after PCA with $N_e = N_v = 3$

(d) db2 before PCA with $N_e = N_v = 1$ (e) db2 after PCA with $N_e = N_v = 1$ (f) db2 after PCA with $N_e = N_v = 3$

Fig. 5. Comparison between $\hat{P}_e^{db}$ and P_e^m for the feature components of databases db1 and db2. The lower and upper boundaries (LB and UB) indicate the 95 percentile area of the estimated error probability when using a synthetic database of similar size.

database with 110 subjects and 880 samples in total. By using normal probability plots, we have shown that when applying PCA, the resulting features tend to better approximate a Gaussian distribution and thus the analytic framework has a much better fit with real biometric data if PCA has been applied. Furthermore, a good fit was observed between the analytic and calculated error probability for both biometric databases for both the $N_e = N_v = 1$ and $N_e = N_v = 3$ settings. In all but one setting, the number of occurrences where the analytic error probability P_e^m was outside the given 95 percentile boundary was smaller than what was allowed.

References

[1] A. M. Bazen and R. N. J. Veldhuis. Likelihood-ratio-based biometric verification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):86–94, 2004.

[2] E.-C. Chang and S. Roy. Robust extraction of secret bits from minutiae. In *Int. Conf. on Biometrics*, pages 750–759, Seoul, South Korea, August 2007.

[3] Y.-J. Chang, W. Zhang, and T. Chen;. Biometrics-based cryptographic key generation. In *IEEE Int. Conf. on Multim. and Expo*, volume 3, pages 2203 – 2206, June 2004.

[4] C. Chen, R. Veldhuis, T. Kevenaar, and A. Akkermans. Multi-bits biometric string generation based on the likelihood ratio. In *IEEE Conf. on Biometrics: Theory, Applications and Systems*, Washington DC, September 2007.

[5] Y. Dodis, L. Reyzin, and A. Smith. Fuzzy extractors: How to generate strong secret keys from biometrics and other noisy data. In *Advances in Cryptology - Eurocrypt 2004, LNCS 3027*, pages 532–540, 2004.

[6] B. Gökberk, M. O. Irfanoglu, and L. Akarun. 3D shape-based face representation and feature extraction for face recognition. *Image and Vision Computing*, 24(8):857–869, August 2006.

[7] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):4–37, January 2000.

[8] A. Juels and M. Sudan. A fuzzy vault scheme. In *Proc. of the 2002 International Symposium on Information Theory (ISIT 2002)*, Lausanne, 2002.

[9] E. J. C. Kelkboom, B. Gökberk, T. A. M. Kevenaar, A. H. M. Akkermans, and M. van der Veen. "3d face": Biometric template protection for 3d face recognition. In *Int. Conf. on Biometrics*, pages 566–573, Seoul, Korea, August 2007.

[10] T. A. M. Kevenaar, G.-J. Schrijen, A. H. M. Akkermans, M. van der Veen, and F. Zuo. Face recognition with renewable and privacy preserving binary templates. In *4th IEEE workshop on AutoID*, pages 21–26, Buffalo, New York, USA, October 2005.

[11] J.-P. Linnartz and P. Tuyls. New shielding functions to enhance privacy and prevent misuse of biometric templates. In *4th Int. Conf. on AVBPA*, 2003.

[12] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2000: fingerprint verification competition. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 24(3):402–412, 2002.

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE CVPR*, volume 2, pages 454–461, June 2005.

[14] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev. *Integrals and Series, Vol. 2: Special Functions.* Gordon and Breach, New York, 1990.

[15] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001.