Accelerating Biometric Identification

David Naccache, Zdenek Riha

***Abstract*—By opposition to biometric matching, biometric identification is a relatively costly process. Let $B = \{b_1, \ldots, b_n\}$ be a database of n biometric templates and let b be a given individual biometric acquisition. The biometric identification problem consists in finding the most likely b_i corresponding to b. This paper assumes the existence of an oracle $\mathfrak{A}$ taking as b and b_i, and responding with true or false. Considering $\mathfrak{A}$ as an atomic operation, any system-level optimization must necessarily minimize the number of calls to $\mathfrak{A}$ per identification session. This is the parameter that we optimize in this paper.**

***Index Terms*—biometrics, biometric identification, correlation**

I. Introduction

By opposition to biometric matching, biometric identification [2], [3] is a relatively costly process. Let $B = \{b_1, \ldots, b_n\}$ be a database of n biometric templates and let b be a given individual biometric acquisition. The biometric identification problem consists in finding the most likely b_i corresponding to b [1].

Whilst in reality matching algorithms return a score compared to a threshold, for the sake of simplicity this paper assume the existence of an oracle $\mathfrak{A}$ taking b and b_i as input, and responding with true or false:

$$\mathfrak{A}(b, b_i) \in \{\mathsf{T}, \mathsf{F}\}$$

Considering $\mathfrak{A}$ as an *atomic* operation, any system-level optimization must necessarily minimize the number of calls to $\mathfrak{A}$ per identification session. This is the parameter that we attempt to optimize in this paper.

For doing so, we assume that every user i has a collection of k additional biometric parameters $m_{i,1}, \ldots m_{i,k}$. An $m_{i,j}$ can be either derived from the template b_i (*i.e.* $m_{i,j} = \text{function}_j(b_i)$) or measured independently. For instance if b_i is a fingerprint then $m_{i,7}$ can be the density of minutiae (the number of minutiae per unit of surface) or an additional parameter, such as the person's height, which is not correlated to b_i.

We will use the $m_{i,j}$ to accelerate identification by applying $\mathfrak{A}$ to the *most probable candidates first*. We denote by σ_j the standard deviation of the $m_{i,j}$'s, for all users i.

The proposed identification process is:

1) Acquire the biometric candidate information b and the additional information $m_1, \ldots, m_k$.
2) Compute for every user i the score:

$$t_i = \sum_{j=1}^{k} \frac{(m_j - m_{i,j})^2}{{\sigma_j}^2} \tag{1}$$

3) Try $\mathfrak{A}(b, b_i)$ by order of *increasing* t_i values.

Given that $\mathfrak{A}$ will be applied to the most promising candidates first (the ones with the lowest t_i), this is likely to result in a significantly faster identification procedure.

However, the comparison of the t_i's assumes that the $m_{i,j}$ are independent. This is not always the case. For instance a tall person is likely to be heavier. In other words, height (*e.g.* $m_{i,2}$) and weight (*e.g.* $m_{i,5}$) are *correlated*.

The process described so far did not take such correlations into account.

II. Analysis of the Procedure

We start by analyzing the proposed procedure without taking correlations into account.

The computation of the t_i's as given by equation (1) rests on the assumption that the measurements m_j each follow an independent normal distribution. More precisely, assuming that every measurement m_j follows a normal distribution with mean μ_j and standard deviation σ_j, the density function can be expressed as:

$$f_{m_j}(x) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_j)^2}{2{\sigma_j}^2}\right)$$

When the m_j's are independently distributed, the probability density of all measurements m_j for $1 \leq j \leq k$ can be expressed as a k-dimensional multivariate distribution:

$$f_{\vec{m}}(\vec{x}) = \prod_{j=1}^{k} f_{m_j}(x_j) = \frac{1}{(2\pi)^{k/2} \prod_{j=1}^{k} \sigma_j} \exp\left(-\sum_{j=1}^{k} \frac{(x_j - \mu_j)^2}{2{\sigma_j}^2}\right)$$

where $\vec{x} = (x_1, \ldots, x_k)$.

Note that in the previous equation μ_j and σ_j are the mean and standard deviation of m_j for *all* users i. For a measurement m_j corresponding to a specific user i, we can also assume that m_j follows a normal distribution with mean $\tilde{\mu_j} = m_{i,j}$ and standard deviation $\tilde{\sigma_j}$; we also assume that the standard deviation $\tilde{\sigma}_j$ around $m_{i,j}$ is the same for all users. In this case, the measurement m_j corresponding to user i has the following distribution:

$$f_{\vec{m}}(\vec{x}) = \frac{1}{(2\pi)^{k/2} \prod_{j=1}^{k} \tilde{\sigma}_j} \exp\left(-\sum_{j=1}^{k} \frac{(x_j - m_{i,j})^2}{2{\tilde{\sigma}_j}^2}\right)$$

D. Naccache is a researcher at the École normale supérieure's Cryptography Group and a professor at the University of Paris II (email: david.naccache@ens.fr).

Z. Riha is with the Masaryk University, Faculty of Informatics, Brno, Czech Republic (email: zriha@fi.muni.cz).

Additionally, we assume that the standard deviation $\tilde{\sigma}_j$ of m_j around $m_{i,j}$ is proportional to the standard deviation σ_j of m_j when all users are considered, *i.e.* we assume $\tilde{\sigma}_j = \alpha \cdot \sigma_j$ for all $1 \leq j \leq k$ for some $\alpha \in \mathbb{R}$. In this case, the probability density function of the m_j's for user i can be written as:

$$f_i(\vec{m}) = \frac{1}{(2\pi)^{k/2}\alpha^k \prod_{j-1}^{k} \sigma_j} \exp\Big(-\sum_{j=1}^{k} \frac{(m_j - m_{i,j})^2}{2\alpha^2 {\sigma_j}^2}\Big) =$$

$$= \frac{1}{(2\pi)^{k/2}\alpha^k \prod_{j-1}^{k} \sigma_j} \exp\Big(-\frac{t_i}{2\alpha^2}\Big)$$

where t_i is precisely the quantity given by equation (1). The probability to obtain measurements m_j from user i is thus a decreasing function of t_i. Given $\vec{m}$, the most probable candidate is hence the one with the lowest t_i.

III. Taking Correlation into Account

The comparison of the t_i's assumes that the different biometric measurements $m_{i,j}$ are independent. This is not necessarily the case since (for instance) a tall person is likely to be heavier; in other words, height and weight are correlated. In this section we the definition of t_i to take correlation into account.

A. Multivariate Normal Distribution

We denote by Σ the covariance matrix of the measurements m_j, defined as follows:

$$\Sigma = \mathrm{var}(\vec{m}) = \mathrm{var}\begin{pmatrix} m_1 \\ \vdots \\ m_k \end{pmatrix} =$$

$$= \begin{pmatrix} \mathrm{var}(m_1) & \mathrm{cov}(m_1 m_2) & \cdots & \mathrm{cov}(m_1 m_k) \\ \mathrm{cov}(m_1 m_2) & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(m_1 m_k) & \cdots & \cdots & \mathrm{var}(m_k) \end{pmatrix}$$

where $\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$.

We assume that the measurements m_j follow a k-dimensional multivariate distribution with mean $\vec{\mu}$ and covariance matrix Σ; the probability density function can then be expressed as:

$$f_{\vec{m}}(\vec{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\Big(-\tfrac{1}{2}(\vec{x} - \vec{\mu})'\Sigma^{-1}(\vec{x} - \vec{\mu})\Big)$$

where $|\Sigma|$ is the determinant of Σ. Note that mean $\vec{\mu}$ is a k-dimensional vector and Σ is a $k \times k$-matrix.

Note that in the previous equation $\vec{\mu}$ and Σ are the expected value and covariance matrix of measurements m_j for *all* users i. As in Section II, for measurements m_j's corresponding to a specific user i, we also assume that the m_j's follow a k-multivariate normal distribution with mean $\tilde{\mu}_j = m_{i,j}$ and covariance matrix $\tilde{\Sigma}$; we also assume that $\tilde{\Sigma}$ is the same for all users. In this case, the measurement $\vec{m}$ for user i follows the multivariate distribution:

$$f_{\vec{m}}(\vec{x}) = \frac{1}{(2\pi)^{k/2}|\tilde{\Sigma}|^{1/2}} \exp\Big(-\tfrac{1}{2}(\vec{x} - \vec{m_{i,\cdot}})'\tilde{\Sigma}^{-1}(\vec{x} - \vec{m_{i,\cdot}})\Big)$$

As in Section II we additionally assume that the covariance matrix satisfies $\tilde{\Sigma} = \alpha \cdot \Sigma$ for some $\alpha \in \mathbb{R}$. In this case, the probability density function can be written as:

$$f_{\vec{m}}(\vec{x}) = \frac{1}{(2\pi\alpha)^{k/2}|\Sigma|^{1/2}} \exp\Big(-\tfrac{1}{2\alpha}(\vec{x} - \vec{m_{i,\cdot}})'\Sigma^{-1}(\vec{x} - \vec{m_{i,\cdot}})\Big)$$

which gives:

$$f_{\vec{m}}(x) = \frac{1}{(2\pi\alpha)^{k/2}|\Sigma|^{1/2}} \exp\Big(-\frac{t_i}{2\alpha}\Big)$$

where:

$$t_i = (\vec{m} - \vec{m_{i,\cdot}})'\Sigma^{-1}(\vec{m} - \vec{m_{i,\cdot}}) \tag{2}$$

Therefore we obtain that equation (2) is a generalization of equation (1) when taking correlations into account.

B. The New Identification Procedure

The new algorithm is:

1) Collect from the user the biometric information b and the additional information $m_1, \ldots, m_k$.
2) Compute for every user i the value:
$$t_i = (\vec{m} - \vec{m_{i,\cdot}})'\Sigma^{-1}(\vec{m} - \vec{m_{i,\cdot}})$$
3) Sort the t_i's by increasing values and apply $\mathfrak{A}(b, b_i)$ to user i by increasing t_i values.

C. Bivariate Case

To illustrate the algorithm we first restrict ourselves to the bivariate case. In this case, the covariance matrix between variables X and Y can be written:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

where $\mathrm{var}(X) = \sigma_x^2$, $\mathrm{var}(Y) = \sigma_y^2$ and $\mathrm{cov}(X,Y) = \rho\sigma_x\sigma_y$ where ρ is the correlation between X and Y. In this case, we have:

$$\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} \frac{1}{\sigma_x^2} & \frac{-\rho}{\sigma_x\sigma_y} \\ \frac{-\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{bmatrix}$$

and the probability density function can be written:

$$f(x,y) =$$

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right]\right)$$

In this case, equation (2) gets simplified as follows:

$$t_i = \frac{(m_1 - m_{i,1})^2}{\sigma_1^2} + \frac{(m_2 - m_{i,2})^2}{\sigma_2^2} -$$

$$- \frac{2\rho(m_1 - m_{i,1})(m_2 - m_{i,2})}{\sigma_1\sigma_2}$$

where $\sigma_1 = \mathrm{var}(m_1)$, $\sigma_2 = \mathrm{var}(m_2)$ and ρ is the correlation between m_1 and m_2.

D. Illustration

We illustrate this with a set of simulated measurements: height, weight and number of collected minutiae, for 13 users.

User	1	2	3	4	5	6	7
Height	178	165	190	176	174	192	182
Weight	71	66	82	80	76	85	76
Minutiae	14	15	14	27	15	25	14

User	8	9	10	11	12	13
Height	162	168	175	187	195	168
Weight	65	80	77	68	92	72
Minutiae	22	23	24	23	19	25

We obtain the following correlation matrix:

$$\Sigma = \begin{bmatrix} 104.9 & 52.9 & -5.2 \\ 52.9 & 56.3 & 3.9 \\ -5.2 & 3.9 & 22.8 \end{bmatrix}$$

which can be written as:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix}$$

where $\sigma_1 = 10.2$, $\sigma_2 = 7.5$, $\sigma_3 = 4.8$, and $\rho_{12} = 0.688390$, $\rho_{13} = -0.107015$, $\rho_{23} = 0.109587$.

Since ρ_{13} and ρ_{23} are small, for simplicity we consider only correlations between the first and second variables (height and weight). More precisely we consider the simplified covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \\ & & \sigma_3^2 \end{bmatrix}$$

with the same previous values of σ_1, σ_2, σ_3 and $\rho = \rho_{12}$. This gives:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{(1-\rho^2)\sigma_1^2} & \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} & \\ \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{(1-\rho^2)\sigma_2^2} & \\ & & \frac{1}{\sigma_3^2} \end{bmatrix}$$

This gives the following formula for t_i which takes into account correlations between height and weight:

$$t_i = \frac{(m_1 - m_{i,1})^2}{(1-\rho^2)\sigma_1^2} + \frac{(m_2 - m_{i,2})^2}{(1-\rho^2)\sigma_2^2} - \\ - \frac{2\rho(m_1 - m_{i,1})(m_2 - m_{i,2})}{(1-\rho^2)\sigma_1\sigma_2} + \frac{(m_3 - m_{i,3})^2}{\sigma_3^2}$$

IV. Conclusions

In the paper we have presented an approach to accelerate the biometric identification process. The algoritm is based on the basic principle of testing the most probable candidates first. We started with assumption that set of measurements of a user are considered to be independent and later we introduced correlations into the scheme.

One drawback of the previous technique is that given a measurement $\vec{m} = (m_1, \ldots, m_k)$ the t_i's must be computed for all users i. A possible speed-up could be to select only those users i for which $|m_1 - m_{i,1}|$ is relatively small. This can be done efficiently if the values $m_{i,1}$ are pre-sorted. Another refinement consists computing all the t_i's simultaneously (*i.e.* compute j-wise rather than i-wise), progressively delay the computation of "heavier" t_i's and start the comparison of the "lighter" ones as soon as these become available.

References

[1] Michael E. Schuckers. Computational Methods in Biometric Authentication. Springer, London, 2010. ISBN 978-1-84996-201-8.
[2] Herve Jarosz and Jean-Christophe Fondeur. Large-Scale Identification System Design. In *Biometric Systems Technology, Design and Performance Evaluation*. Springer, 2005. ISBN: 978-1-85233-596-0.
[3] Michael Brauckmann and Christoph Busch. Large Scale Database Search. In *Handbook of Face Recognition*. 2011, pp 639-653. ISBN: 978-0-85729-931-4.

David Naccache is a researcher at the École normale supérieure's Cryptography Group and a professor at the University of Paris II. His research interests include public-key cryptography and mobile code security. Naccache has a PhD in cryptology from the École nationale supérieure des télécommunications Paris. Contact him at david.naccache@ens.fr.

Zdenek Riha a is an Assistant Professor at the Masaryk University, Faculty of Informatics, in Brno, Czech Republic. He received his PhD degree from the Faculty of Informatics, Masaryk University. In 1999 he spent 6 months on an internship at Ubilab, the research lab of the bank UBS, focusing on security and usability aspects of biometric authentication systems. Between 2005 and 2008 he was seconded as a Detached National Expert to the European Commission's Joint Research Centre in Italy. Zdenek can be contacted at zriha@fi.muni.cz.