Searching methods for biometric identification systems: Fundamental limits

Searching Methods for Biometric Identification Systems: Fundamental Limits

Frans M.J. Willems
Eindhoven University of Technology, Electrical Engineering Department, Eindhoven, The Netherlands.

***Abstract*—We study two-stage search procedures for biometric identification systems in an information-theoretical setting. Our main conclusion is that clustering based on vector-quantization achieves the optimum trade-off between the number of clusters (cluster rate) and the number of individuals within a cluster (refinement rate). The notion of excess rate is introduced, a parameter which relates to the amount of clusters to which the individuals belong. We demonstrate that noisier observation channels lead to larger excess rates.**

I. Introduction

Biometric identification systems rely on the physiological and/or behavioral characteristics of individuals. Examples of these characteristics are face, fingerprint, hand-geometry, iris, retina, keystroke, signature, and voice, see Uludag et al. [7]. An identification system operates in two modes. In the first mode, the enrollment mode, the biometric data of all individuals are observed, and maybe after some pre-processing, the system stores in a database an enrollment vector (record) for each individual. When at some later time an individual shows up for identification, this corresponds to the second mode of the system, the individual is observed again and this results, possibly after some post-processing, in an identification vector (record). The system then searches the database for the enrollment vector that gives the best match with the observed identification vector. It should be noted that in the enrollment mode and the identification mode, the observed vectors are in general noisy versions of the "real" feature vectors (records).

In principle the system can perform an exhaustive search on all the enrollment records to find the best match. Chavez et al. [2] give an extensive overview of methods that intend to reduce the number of enrollment records that are actually accessed. Weber et al. [9] compare indexing techniques to methods based on what they call vector-approximations (VA). Similar to these VA methods are the fingerprinting techniques that used in content-based audio identification, see Haitsma and Kalker [3], and Cano et al. [1]. In an information-theoretical context such methods would be referred to as quantization methods. Weber et al. [9] observe that for searching high-dimensional spaces quantization methods like VA outperform indexing methods.

Quantization can also be used in the enrollment mode with the objective to compress the database. Tuncel et al. [5], the first authors that investigated the rate-distortion approach to database searching, apply quantization during enrollment and consider the fundamental trade-off between compression rate and reconstruction distortion. Later Tuncel [6] also considered the trade-off between enrollment compression rate and identification rate. This extends a result of Willems et al. [8] showing that the maximum identification rate of a biometrical system is equal to the mutual information between the enrollment and identification observations, see also [4]. A crucial observation to obtain this result is that a set of biometric enrollment vectors can be regarded as a random channel code.

In the current manuscript we focus on speeding up the search process, as in [9]. We are not interested in compressing the database as in [5], [6]. We will show that in an information-theoretical setting quantization methods are optimal.

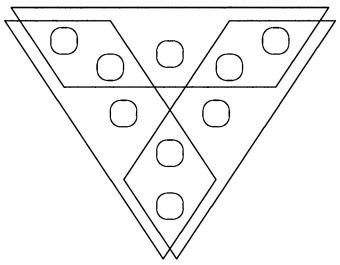

Fig. 1. Nine individuals in three clusters. Three cluster-checks and five refinement-checks. Eight checks in total.

To demonstrate what we mean by quantization, suppose that the system upon observing an individual, first detects to which cluster the individual belongs, and after that decides about the individual itself (two-stage identification). If there are M individuals, an ideal systems will have $\sqrt{M}$ clusters each containing $\sqrt{M}$ individuals. To determine the cluster index $\sqrt{M}$ candidate-clusters can be checked, and then to determine the individual within the cluster, $\sqrt{M}$ refinement-checks are needed. This results in $2\sqrt{M}$ checks in total, considerably less than the M checks that are required for exhaustive search. In general however individuals can be in more than one cluster, see Fig. 1, and then the number of cluster-checks times the number of refinement-checks exceeds the number of individuals. Here we investigate the fundamental trade-off between cluster-check rate and refinement-check rate.

An important point is what we mean by a cluster-check. In principle a cluster-check could correspond to $\sqrt{M}$ sub-checks, one for each individual within the cluster. To prevent this, we require the device that makes the cluster-decision to be "ignorant" of the biometric enrollment vectors. Under this assumption an optimal system contains an ignorant device that acts as a vector quantizer.

In the next section we present our model of a biometrical identification system based on two-stage identification and we

978-1-4244-4313-0/09/$25.00 ©2009 IEEE

will state our main result. Section III contains the proof of this result. In section IV we consider as an example a binary symmetric system and we introduce the notion of excess rate there. Concluding remarks will follow in Section V.

II. Model Description and Statement of Result

A. Model Description

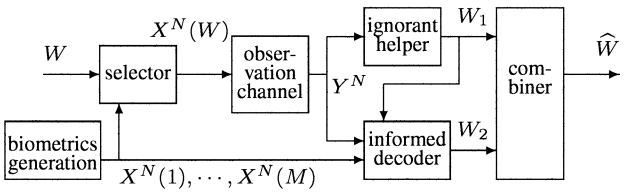

Fig. 2. Model of a two-stage biometric identification system.

In a biometric identification system, see Fig. 2, there are M individuals indexed $w \in \{1, 2, \cdots, M\}$ that are to be identified. To each such individual there corresponds a randomly generated biometric sequence (vector) of length N. This sequence has symbols $x_n, n = 1, 2, \cdots, N$ taking values in the discrete alphabet $\mathcal{X}$, and the probability that sequence $x^N = (x_1, x_2, \cdots, x_N)$ occurs as biometric sequence for individual w is

$$\Pr\{X^N(w) = x^N\} = \Pi_{n=1}^N Q_b(x_n), \tag{1}$$

hence the components $X_1, X_2, \cdots, X_N$ are independent and identically distributed according to $\{Q_b(x), x \in \mathcal{X}\}$. Note that this probability does not depend on the index w. We assume that all biometric sequences are generated prior to the identification procedure. They form what we call the "code" here. This code C is the list of biometric sequences, hence

$$C = \left(x^N(1), x^N(2), \cdots, x^N(M)\right). \tag{2}$$

In the identification process the probabilities for the individuals to show up for identification all equal, hence

$$\Pr\{W = w\} = 1/M \text{ for } w \in \{1, 2, \cdots, M\}. \tag{3}$$

When individual w shows up for identification, its biometric sequence $x^N(w)$ is "selected" from the code C and presented to the system, hence

$$x^N = s(w, C). \tag{4}$$

The system observes x^N via a memoryless observation channel $\{Q_c(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$, with discrete alphabet $\mathcal{Y}$, and the resulting channel output sequence is $y^N = (y_1, y_2, \cdots, y_N)$, where $y_n \in \mathcal{Y}$ for $n = 1, 2, \cdots, N$. Now

$$\Pr\{Y^N = y^N | X^N(w) = x^N\} = \Pi_{n=1}^N Q_c(y_n|x_n). \tag{5}$$

After observing y^N identification starts by making a first decision (cluster decision). This decision with outcome $w_1 \in \{1, 2, \cdots, M_1\}$ is taken by a so-called "ignorant" helper, a device that has no knowledge of the biometric sequences that were generated, hence

$$w_1 = h(y^N). \tag{6}$$

Then a second decision is made (refinement decision), based on the first decision w_1 and the list of generated biometric sequences. This decision with outcome $w_2 \in \{1, 2, \cdots, M_2\}$ is taken by a so-called "informed" decoder, hence

$$w_2 = d(y^N, w_1, C), \tag{7}$$

where C is the code. Finally a combiner forms an estimate of the index of the individual that presented its biometric sequence for identification, hence

$$\widehat{w} = c(w_1, w_2). \tag{8}$$

We assume that $\widehat{w} \in \{1, 2, \cdots, M\}$. The reliability of our identification system is measured by the error probability

$$P_{\mathcal{E}} = \Pr\{\widehat{W} \neq W\}. \tag{9}$$

B. Statement of Result

We now say that rate triple (R_1, R_2, R) with $R \geq 0$ is achievable if for all $\epsilon > 0$ there exist for all N large enough mappings $h(\cdot)$, $d(\cdot, \cdot, \cdot)$, and $c(\cdot, \cdot)$ such that

$$\begin{aligned} \log_2(M_1) &\leq N(R_1 + \epsilon), \\ \log_2(M_2) &\leq N(R_2 + \epsilon), \\ \log_2(M) &\geq N(R - \epsilon), \text{ and} \\ \Pr\{\widehat{W} \neq W\} &\leq \epsilon. \end{aligned} \tag{10}$$

We call R the identification rate, and R_1 and R_2 resp. cluster and refinement rate. We are now ready to state the main result of this submission, the proof follows in section III.

Theorem 1: The region of achievable rate triples for our biometric identification system is given by

$$\begin{aligned} \{(R_1, R_2, R) \quad : \quad & R_1 \geq I(Y; U), \\ & R_2 \geq \max(0, R - I(X; U)), \\ & 0 \leq R \leq I(X; Y), \\ & \text{for } P(x, y, u) = Q_b(x) Q_c(y|x) P(u|y), \\ & \text{where } |\mathcal{U}| \leq |\mathcal{Y}| + 1\}. \end{aligned} \tag{11}$$

III. Proof

The proof consists of the achievability part, a converse, and a cardinality bound part. We start with the converse.

A. Converse Part

For the range M_1 of the first decision we find that:

$$\begin{aligned} \log_2(M_1) &\geq H(W_1) \geq I(Y^N; W_1) \\ &= \sum_{n=1}^N I(Y_n; W_1 | Y^{n-1}) \overset{(a)}{=} \sum_{n=1}^N I(Y_n; W_1, Y^{n-1}) \\ &\overset{(b)}{=} \sum_{n=1}^N I(Y_n; U_n), \end{aligned} \tag{12}$$

where (a) follows from the fact that $H(Y_n|Y^{n-1}) = H(Y_n)$ since $Y_1, Y_2, \cdots$, and Y_N are independent of each other, and (b) from definition $U_n \triangleq (W_1, Y^{n-1})$ for $n = 1, 2, \cdots, N$. Next let N be a random variable taking values in $\{1, 2, \ldots, N\}$

with equal probability, and let $X = X_n$ and $Y = Y_n$, when $N = n$. Then

$$\begin{aligned}\sum_{n=1}^{N} I(Y_n;U_n) &= NH(Y_N|N) - NH(Y_N|U_N,N)\\ &\overset{(c)}{=} NH(Y) - NH(Y|U_N,N) = NI(Y;(U_N,N))\\ &\overset{(d)}{=} NI(Y;U), \end{aligned} \tag{13}$$

where step (c) follows since $Y_1, Y_2, \cdots$, and Y_N are identically distributed and $Y_N = Y$, and (d) from $U \triangleq (U_N, N)$.

Since $M_2 \geq 1$ we obtain for the range M_2 of the second decision that:

$$\log_2(M_2) \geq 0. \tag{14}$$

Moreover consider, using $F \triangleq 1 + \Pr\{\widehat{W} \neq W\}\log_2(M)$, the series of (in)equalities:

$$\begin{aligned}\log_2(M) = H(W) &\leq H(W) - H(W|\widehat{W}) + F\\ &\leq I(W;\widehat{W},W_1,W_2) + F\\ &\overset{(e)}{=} I(W;W_1) + I(W;W_2|W_1) + F\\ &\leq I(W,X^N;W_1) + \log_2(M_2) + F\\ &\overset{(f)}{=} I(X^N;W_1) + \log_2(M_2) + F\\ &= \sum_{n=1}^{N} I(X_n;W_1|X^{n-1}) + \log_2(M_2) + F\\ &\leq \sum_{n=1}^{N} I(X_n;W_1,X^{n-1},Y^{n-1}) + \log_2(M_2) + F\\ &\overset{(g)}{=} \sum_{n=1}^{N} I(X_n;W_1,Y^{n-1}) + \log_2(M_2) + F\\ &\overset{(h)}{=} NI(X;U) + \log_2(M_2) + F. \end{aligned} \tag{15}$$

where (e) follows from the fact that $I(W;W_1,W_2,\widehat{W}) = I(W;W_1,W_2)$, (f) since $W - X^N - W_1$, (g) since $X^{n-1} - Y^{n-1} - X_n, W_1$, and (h) similar to how (13) was obtained.

Finally consider the number M of individuals:

$$\begin{aligned}\log_2(M) = H(W) &\leq I(W;\widehat{W}) + F\\ &\overset{(i)}{\leq} I(X^N;Y^N) + F \overset{(j)}{=} \sum_{n=1}^{N} I(X_n;Y_n) + F\\ &= NI(X_N;Y_N|N) + F \overset{(k)}{\leq} NI(X;Y) + F. \end{aligned} \tag{16}$$

where (i) follows from $I(W;\widehat{W}) \leq I(W;Y^N,C,\widehat{W}) = I(W;Y^N,C) = I(W;Y^N|C) = I(W,X^N;Y^N|C) \leq H(Y^N) - H(Y^N|X^N) = I(X^N;Y^N)$, and (j) from the fact that $(X_1,Y_1),(X_2,Y_2),\cdots,(Y_N,Y_N)$ are independent, (k) since these pairs are identically distributed and since $(X,Y) = (X_n,Y_n)$ for $N = n$.

Note that $X^{n-1} - Y^{n-1} - X_n, W_1$ (and (g)) follows from

$$\begin{aligned}&p(x^{n-1},y^{n-1},x_n,w_1)\\ &= \sum_{w,y_n,x_{n+1}^N,y_{n+1}^N} p(w)[\Pi_{i=1}^N p(x_i)p(y_i|x_i)]p(w_1|y^N)\\ &= [\Pi_{i=1}^{n-1}p(x_i)p(y_i|x_i)]p(x_n)\sum_{y_n,x_{n+1}^N,y_{n+1}^N} p(y_n|x_n)\\ &\quad\cdot[\Pi_{j=n+1}^N p(x_j)p(y_j|x_j)]p(w_1|y^N)\\ &= p(y^{n-1})p(x^{n-1}|y^{n-1})p(x_n)p(w_1|x_n,y^{n-1}), \end{aligned} \tag{17}$$

where we use the extra notation $x_a^b \triangleq x_a, x_{a+1}, \cdots, x^b$. Furthermore Y_n is independent of Y^{n-1}, used in (a), since

$$\begin{aligned}p(y_n|y^{n-1}) &= \frac{\sum_{w,x^N,y_{n+1}^N} p(w)p(x^N)p(y^N|x^N)}{\sum_{w,x^N,y_n,y_{n+1}^N} p(w)p(x^N)p(y^N|x^N)}\\ &= \frac{\sum_{x^N,y_{n+1}^N} p(x^N)p(y^N|x^N)}{\sum_{x^N,y_n,y_{n+1}^N} p(x^N)p(y^N|x^N)}\\ &= \frac{\Pi_{i=1}^n \sum_{x_i} p(x_i)p(y_i|x_i)}{\Pi_{i=1}^{n-1}\sum_{x_i} p(x_i)p(y_i|x_i)}\\ &= \sum_{x_n} p(x_n)p(y_n|x_n) = p(y_n). \end{aligned} \tag{18}$$

Assume that (R_1,R_2,R) is achievable. Then for all block-lengths N and small enough $\epsilon > 0$, using $F \leq 1 + \epsilon\log_2(M)$, we obtain from (12) and (13), (14) and (15), and (16) that

$$\begin{aligned}N(R_1+\epsilon) &\geq \log_2(M_1) \geq NI(Y;U),\\ N(R_2+\epsilon) &\geq \log_2(M_2) \geq 0,\\ N(R_2+\epsilon) &\geq \log_2(M_2) \geq \log_2(M) - NI(X;U) - F,\\ &\geq (1-\epsilon)N(R-\epsilon) - 1 - NI(X;U),\\ N(R-\epsilon) &\leq \log_2(M) \leq \frac{1}{1-\epsilon}(NI(X;Y)+1), \end{aligned} \tag{19}$$

for some $p(x,y,u) = Q_b(x)Q_c(y|x)P(u|y)$. Note that this follows from

$$\begin{aligned}&p(x,y,u) = p(x_n,y_n,w_1,y^{n-1},n)\\ &= \frac{1}{N}\sum_{w,x_1^{n-1},x_{n+1}^N,y_{n+1}^N} p(w)p(x^N|w)p(y^N|x^N)p(w_1|y^N)\\ &= p(x_n)p(y_n|x_n)\frac{1}{N}\sum_{x_1^{n-1},x_{n+1}^N,y_{n+1}^N} \Pi_{i=1}^{n-1}[p(x_i)p(y_i|x_i)]\\ &\quad\cdot\Pi_{j=n+1}^N[p(x_j)p(y_j|x_j)]p(w_1|y^N)\\ &= p(y_n)p(x_n|y_n)p(y^{n-1},w_1,n|y_n). \end{aligned} \tag{20}$$

From (19) the converse to Thm. 1 now follows after letting $\epsilon \downarrow 0$ and $N \to \infty$.

B. Achievability

We can only give an outline of the achievability proof here. Fix an $0 < \epsilon < 1$, a distribution $p(x,y,u) = Q_b(x)$

$Q_c(y|x)P(u|y)$, and identification rate $0 \leq R \leq I(X;Y)$. Now we define the sets $\mathcal{B}_\epsilon^{(N)}(YU)$ as

$$\mathcal{B}_\epsilon^{(N)}(YU) \stackrel{\Delta}{=} \{(\underline{y},\underline{u}) : \Pr\{(\underline{X},\underline{y},\underline{u}) \in \mathcal{A}_\epsilon^{(N)}(XYU) \mid (\underline{Y},\underline{U}) = (\underline{y},\underline{u})\} \geq 1-\epsilon\}, \quad (21)$$

where $\underline{X}$ is the output of a "backward" channel $Q_{X|Y}(x|y) = Q(x,y)/\sum_x Q(x,y)$, with $Q(x,y) = Q_b(x)Q_c(y|x)$, having input $\underline{y}$. Typical set $\mathcal{A}_\epsilon^{(N)}(XYU)$ corresponds to $p(x,y,u)$.

We first use a random coding argument to construct a collection of covering sequences $\underline{u}(1), \underline{u}(2), \cdots, \underline{u}(M_1)$, where we take $M_1 = 2^{N(I(Y;U)+4\epsilon)}$. Averaged over the random covering code, the probability that a sequence $\underline{y}$, i.i.d. according to $p(y) = \sum_{x,u} p(x,y,u)$ occurs, such that $(\underline{y},\underline{u}(w_1)) \notin \mathcal{B}_\epsilon^{(N)}(YU)$ (not jointly $\mathcal{B}$-typical) for all $w_1 \in \{1,2,\cdots,M_1\}$, can be made $\leq 3\epsilon$ letting $N \to \infty$. Consequently there exists a covering code with probability that at least one of the covering sequences is jointly $\mathcal{B}$-typical with an i.i.d. $\underline{y}$ of at least $1-3\epsilon$.

During enrollment, after biometric sequence $\underline{x}(w)$ was generated, for $w = 1,2,\cdots,M$, the system finds out which $\underline{u}(w_1)$ are jointly typical with $\underline{x}(w)$ for $w_1 \in \{1,2,\cdots,M_1\}$. In this way the system creates index-lists $\mathcal{L}(w_1) = \{w : (\underline{x}(w),\underline{u}(w_1)) \in \mathcal{A}_\epsilon^{(N)}(XU)\}$, one for each w_1. These index-lists are available to the informed decoder and the combiner.

During identification, the ignorant helper upon receiving $\underline{y}$ chooses list-index $\widehat{w_1}$ such that covering sequence $\underline{u}(\widehat{w_1})$ is jointly $\mathcal{B}$-typical with $\underline{y}$ i.e. $(\underline{u}(\widehat{w_1}),\underline{y}) \in \mathcal{B}_\epsilon^{(N)}(YU)$. If such a list-index cannot be found, an error is declared. Note that the ignorant helper makes at most M_1 cluster-checks. The corresponding error probability is not larger than 3ϵ. If no error is declared the ignorant helper sends the index $\widehat{w_1}$ to the informed decoder and the combiner.

Next the informed decoder chooses a unique index $\widehat{w}$ from list $\mathcal{L}(\widehat{w_1})$ such that $(\underline{x}(\widehat{w}),\underline{y},\underline{u}(\widehat{w_1})) \in \mathcal{A}_\epsilon^{(N)}(XYU)$. If such a unique index cannot be found, an error is declared. Note that the informed decoder makes at most M_2 refinement-checks.

It follows from the definition of $\mathcal{B}_\epsilon^{(N)}(YU)$ that the probability, that the actual index w doesn't lead to joint typicality, is smaller than ϵ. Note that this typicality also implies that the actual index is in the list $\mathcal{L}(\widehat{w_1})$.

The probability that some "other" index $w' \neq w$ results in joint typicality (and is in the list $\mathcal{L}(\widehat{w_1})$) can be made $\leq \epsilon$ for $M = 2^{N(R-4\epsilon)}$ and N large enough. The informed decoder sends the rank of $\widehat{w_2}$ within the list $\mathcal{L}(\widehat{w_1})$ to the combiner only if it is not larger than $M_2 = 2^{N(R-I(X;U))}$. Otherwise an error is declared. It can be shown that also this probability is not larger than ϵ for N large enough. When no errors occurred the combiner will reconstruct the actual individual-index $\widehat{w} = w$ from both the list index $\widehat{w_1}$ and rank $\widehat{w_2}$.

This demonstrates the achievability part corresponding to Thm. 1.

C. Cardinality Bounds for Auxiliary Random Variable U

To find a bound on the cardinality of the auxiliary variable U let $\mathcal{D}$ be the set of probability distributions on $\mathcal{Y}$ and consider the $|\mathcal{Y}|+1$ continuous functions of $P \in \mathcal{D}$ defined as

$$\begin{aligned} \phi_y(P) &= P(y) \text{ for all but one } y, \\ \phi_Y(P) &= H_P(Y), \\ \phi_X(P) &= H_P(X), \end{aligned} \quad (22)$$

where in the last equation we use $\Pr\{X = x\} = \sum_y P(y)Q_{X|Y}(x|y)$ where $Q_{X|Y}(x|y) = Q_b(x)Q_c(y|x)/\sum_x Q_b(x)Q_c(y|x)$. By the Fenchel-Eggleston strengthening of the Caratheodory lemma (see Wyner and Ziv [11]) there are $|\mathcal{Y}|+1$ elements $P_u \in \mathcal{D}$ and α_u that sum to one, such that

$$\begin{aligned} P(y) &= \sum_{u=1,|\mathcal{Y}|+1} \alpha_u \phi_y(P_u) \text{ for all but one } y, \\ H(Y|U) &= \sum_{u=1,|\mathcal{Y}|+1} \alpha_u \phi_Y(P_u), \\ H(X|U) &= \sum_{u=1,|\mathcal{Y}|+1} \alpha_u \phi_X(P_u). \end{aligned} \quad (23)$$

The entire probability distribution $\{Q(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ and consequently the entropies $H(X)$ and $H(Y)$ are now specified and therefore also both $I(U;Y)$ and $I(U;X)$. This implies that cardinality $|\mathcal{U}| = |\mathcal{Y}|+1$ suffices.

IV. Example, Excess Rate

We consider here a system with binary uniform biometric sequences hence $Q_b(x) = 1/2$ for $x \in \{0,1\}$ and a binary symmetric observation channel, thus $Q_c(y|x) = q$ if $y \neq x$ and $Q_c(y|x) = 1-q$ if $y = x$ where $y \in \{0,1\}$. Parameter $0 \leq q \leq 1/2$ is called the crossover probability. Note that $Q_Y(y) = 1/2$ for $y \in \{0,1\}$.

It is important to observe that the "backward" channel from Y to X is also binary symmetric with crossover probability q since

$$Q_{X|Y}(x|y) = \frac{Q_b(x)Q_c(y|x)}{\sum_x Q_b(x)Q_c(y|x)} = Q_c(y|x). \quad (24)$$

Therefore $X = Y \oplus Z$ where $\oplus$ denotes modulo-2 addition and Z is additive noise independent of Y with $\Pr\{Z=1\} = q$.

We can write

$$\begin{aligned} I(U;Y) &= 1 - H(Y|U), \\ I(U;X) &= 1 - H(X|U). \end{aligned} \quad (25)$$

Since the channel from Y to X is binary additive with crossover probability q Mrs. Gerber's Lemma [10] tells us that if $H(Y|U) = v$ then $H(X|U) \geq h(q * h^{-1}(v))$, where $h(a) \stackrel{\Delta}{=} -a\log_2(a) - (1-a)\log_2(1-a)$ for $0 \leq a \leq 1$ denotes the binary entropy function. If now $0 \leq p \leq 1/2$ is such that $h(p) = v$ then $H(Y|U) = h(p)$ and $H(X|U) \geq h(q * p)$.

When we take the "channel" from Y to U binary symmetric with crossover probability p the minimum $H(X|U)$ is achieved and consequently the region of achievable rate

triples for binary uniform biometrics and a binary symmetric observation channel is given by

$$\{(R_1, R_2, R) \quad : \quad R_1 \geq 1 - h(p), \qquad (26)$$
$$R_2 \geq \max(0, R - 1 + h(p * q),$$
$$0 \leq R \leq 1 - h(q), \text{ for } 0 \leq p \leq 1/2\}.$$

Fig. 3 contains the optimal cluster-refinement rate-pairs (R_1, R_2) for three values of the identification rate R for an observation channel with crossover probability $q = 0.1$.

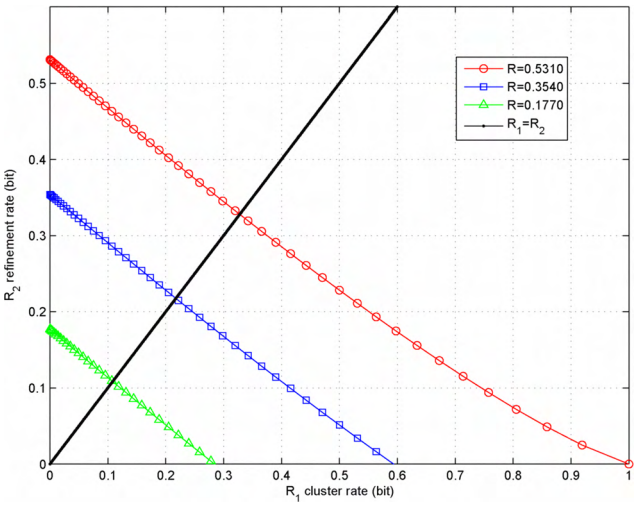

Fig. 3. Optimum cluster-refinement rates-pairs (R_1, R_2) for a system with uniform biometric sequences and a binary symmetric observation channel with crossover probability $q = 0.1$, for biometric rates $R = 0.5310$ (maximum), 0.3640, and 0.1770.

Note that the number of cluster-checks that have to be made by the ignorant helper is roughly 2^{NR_1} and the number of refinement-checks made by the informed decoder is approximately 2^{NR_2}. Minimizing the total number of checks is therefore roughly equivalent to minimizing $\max(R_1, R_2)$. The figure therefore shows the line $R_1 = R_2$.

It is interesting to observe that there is always an "excess rate", in the sense that

$$R_1 + R_2 \geq 1 - h(p) + R - 1 + h(p * q) = R + h(p * q) - h(p). \qquad (27)$$

The excess rate $\Delta \triangleq R_1 + R_2 - R$ for maximum identification rate $R = 0.5310$ is equal to 0.1248.

In the general case we can write for the excess rate

$$\begin{aligned} \Delta = R_1 + R_2 - R &\geq I(U;Y) - I(U;X) \\ &= H(U|X) - H(U|Y,X) \\ &= I(U;Y|X) \\ &= H(Y|X) - H(Y|X,U). \end{aligned} \qquad (28)$$

For U such that $R \geq I(X;U)$ and for optimum cluster-refinement rate-pairs (R_1, R_2) we get

$$\Delta = H(Y|X) - H(Y|X,U) \leq H(Y|X). \qquad (29)$$

This maximum excess rate is achieved for $U = Y$, and this results in refinement rate $R_2 = 0$. Note that the upper bound on the excess rate is larger for more noisy observation channels. Noise-free observation channels allow for a zero excess rate.

V. Concluding Remarks

We have investigated the fundamental trade-off for a two-stage search procedure in a biometric identification system. Our main conclusion is that clustering based on vector-quantization achieves optimum cluster-refinement rate-pairs. We have introduced the notion of excess rate and demonstrated that noisier channels lead to a larger excess rate.

Although our investigation suggests that our random covering code does not contain structure we could use a structured vector quantizer in practise. In such a situation the search complexity of this code (i.e. the cluster rate) is not relevant, however the refinement rate remains significant.

We have only considered a two-step system here. It is not so difficult hoever to find the fundamental limits for multi-stage systems.

The concept of an ignorant helper turns out to be crucial here. We anticipate that the notion of ignorant devices can lead to interesting statements about other information processing systems.

Acknowledgment

The author thanks Ton Kalker and Jaap Haitsma for introducing him to audio fingerprinting, and Michael Gastpar for discussions on ignorant devices.

References

[1] P. Cano, E. Battle, T. Kalker, and J. Haitsma, "A Review of Algorithms for Audio Fingerprinting, in *Proc. 5th IEEE Workshop MMSP*, St. Thomas, Virgin Islands, 2002, pp. 196 - 173.

[2] E. Chávez, G. Navarro, R. Baeza-Yates, J. Marroquin, "Searching in Metric Spaces," *ACM Comput. Surv.,* vol. 33, No. 3., pp. 273 - 321, 2001.

[3] J. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprintng System," *Proc. 3rd Int. Conf. on Music Inform. Retriev., ISMIR,* Paris, France, Oct. 13-17, 2002, pp. 107 - 115.

[4] J.A. O'Sullivan and N.A. Schmidt, "Large Deviation Performance Analysis for Biometrics Recognition," *Proc. 40th Ann. Allerton Conf. Comm. Control, and Comput.,* Oct. 2-4, 2002, Monticello, Ill., pp. 1482 - 1492.

[5] E. Tuncel, P. Koulgi, and K. Rose, "Rate-Distortion Approach to Databases: Storage and Content-Based Retrieval," *IEEE Trans. Inform. Th.,* Vol. IT - 50, No. 6, pp. 953 - 967, June 2004.

[6] E. Tuncel, "Capacity/Storage Tradeoff in High-Dimensional Identification Systems," *IEEE Int. Symp. Inform. Th.,* Seattle, July 9-14, 2006. pp. 1929 - 1933.

[7] U. Uludag, S. Pankanti,, S. Prabhakar, amd A.K. Jain, "Biometric Cryptosystems: Issues and Challenges," *Proc. IEEE,* Vol. 92, No. 6, June 2004, pp. 948 - 960.

[8] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the Capacity of a Biometrical Identification System," *IEEE Int. Symp. Inform. Th.,* Yokohama, June 29 - July 4, 2003, p. 82.

[9] R. Weber, H.-J. Schek, S. Blott, "A Quantitative Analysis and Performance Study for Similarity Search in High-Dimemsional Spaces," *Proc. 24th VLDB Conf.,* New York, 1998, pp. 194 - 205.

[10] A.D. Wyner and J. Ziv, "A Theorem on the Entropy of Certain Binary Sequences and Application: Part I," *IEEE Trans. Inform. Th.,* Vol. IT - 19, No. 6, pp. 769 - 773, November 1973.

[11] A.D. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Trans. Inform. Th.,* Vol. IT - 22, No. 1, pp. 1 - 10, January 1976.