Fundamental Limits of Identification: Identification rate, search and memory complexity trade–off

Farzad Farhadzadeh
Dep. of Computer Science
University of Geneva
Geneva, Switzerland
Email: farzad.farhadzadeh@unige.ch

Frans M.J. Willems
Dep. of Electrical Eng.
Eindhoven University of Technology
Eindhoven, The Netherlands
Email: f.m.j.willems@tue.nl

Sviatoslav Voloshynovskiy
Dep. of Computer Science
University of Geneva
Geneva, Switzerland
Email: svolos@unige.ch

***Abstract*—In this paper, we introduce a new generalized scheme to resolve the trade–off between the identification rate, search and memory complexities in large–scale identification systems. The main contribution of this paper consists in a special database organization based on assigning entries of a database to a set of predefined and possibly overlapping clusters, where the cluster representative points are generated based on statistics of both entries of the database and queries. The decoding procedure is accomplished in two stages: At the first stage, a list of clusters related to the query is estimated, then refinement checks are performed to all members of these clusters to produce a unique index at the second stage. The proposed scheme generalizes several practical searching in identification systems as well as makes it possible to approach a new achievable region of search–memory complexity trade–off.**

I. Introduction

The identification or the nearest neighbor search is a research problem that simultaneously has emerged in a number of applications such as human biometrics [1], content management (multimedia retrieval) [2], multimedia security (copy detection, content identification and tracking) [3] as well as physical object security [4].

An identification system [1] consists of two main phases: *enrollment* and *identification*. In the first phase, the enrollment, feature vectors representing digital contents, humans or physical objects are extracted and stored in a database. In the identification phase, a noisy (degraded) counterpart of an enrolled data, defined as query, is presented to the identification system to identify the query by comparing to feature vectors stored in the database.

In modern applications, the size of a database might be of order of several billions. Therefore, theoretical investigation and development of practical methods achieving identification capacity [1] is of great interest. An efficient approach should satisfy several important requirements. First, users should be able to identify the objects or individuals reliably (reliability). Secondly, the decoding method should be as fast as possible in time (search complexity). Finally, it should require the least possible amount of memory for both the items and the indexing structure (memory complexity). These triple conditions require to solve an information-theoretical problem that considers maximization of identification rate, minimization of computational complexity and memory complexity. It should be pointed out that all these requirements contradict each other, and in fact this triple trade-off is still an open and emerging research problem.

In principle, an identification system can perform an exhaustive search on all entries of the database to find the best match. [5] gives an extensive overview of methods to reduce search complexity in metric spaces. [2] compares 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 [6] observed that for searching in high-dimensional spaces quantization methods like VA outperform indexing methods. In an information-theoretical context such methods would be referred to as quantization methods.

Quantization can also be used in the enrollment phase with the objective to compress the database. [7] exploits quantization during enrollment and consider the fundamental trade-off between compression rate and reconstruction distortion. Later [8] considered the trade-off between enrollment compression rate and identification rate. [4] exploits a search scheme based on Hamming sphere around the noisy feature vector, that can reduce search complexity and simultaneously achieves the identification capacity. However, it should be noted that this scheme is efficient only for low degradations between queries and enrolled data.

This paper is a generalization of the scheme introduced by Willems in [9] to speed up the search process by means of clustering. Where the system upon observing a query, first detects to which cluster the related item belongs, and after that decides about the item itself (two-stage identification). The main differences between the current manuscript and [9] are: (a) generalization of cluster representative points, considered as auxiliary random variables, based on statistics of both entries of the database and queries, while in [9] the cluster representative points have been generated only based on statistics of queries; (b) estimation of a list of clusters versus unique estimation in [9], upon observing a query at the first stage of decoding; (c) the memory complexity was not addressed directly in [9], for the analysis of the triple trade–off; (d) a new result on search–memory complexity region of capacity achieving identification systems.

In the next section we present our model of an identification system based on two-stage identification and we will state our main result. Section III contains the proof of this result. In Section IV we investigate search–memory complexity of the proposed scheme and in Section V we consider as an example a binary symmetric system. Concluding remarks will follow in Section VI.

II. Model Description and Statement of Result

A. Model Description

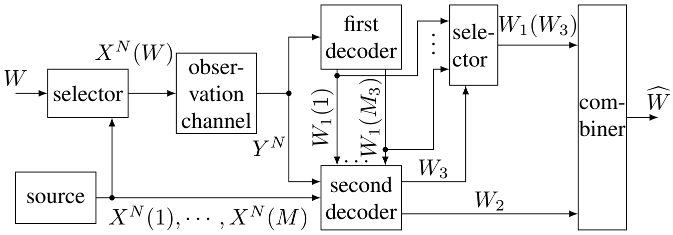

Fig. 1. Model of a two-stage biometric identification system.

In an identification system, see Fig. 1, there are M items indexed $w \in \{1, 2, \cdots, M\}$ that are to be identified. A randomly generated sequence (vector) of length N corresponds to each such item. 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 for item 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 sequences are generated prior to the identification procedure. They form a codebook that we call the "database" here. This database C consists of the list of entries, hence

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

In the identification process the probabilities for the items to be presented for identification are all equal, hence

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

When item w is presented for identification, its corresponding sequence $x^N(w)$ is "selected" from the database 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 constructing a list of indices with cardinality M_3. This index list with outcome $\underline{w}_1 = (w_1(1), w_1(2), \ldots, w_1(M_3))$, $w_1(i) \in \{1, 2, \cdots, M_1\}, 1 \leq i \leq M_3$, is constructed by a so-called "first decoder" $d_1 : \mathcal{Y}^N \rightarrow \mathcal{M}_1^{M_3}$, a device that has no knowledge of the entries that were generated, hence

$$\widehat{\underline{w}_1} = d_1(y^N). \tag{6}$$

Then, at the second decoding stage, a decision (refinement decision) is made, based on the index list $\widehat{\underline{w}_1}$ and the corresponding list of generated sequences. This decision consisting of $w_3 \in \mathcal{M}_3 = \{1, 2, \cdots, M_3\}$ and $w_2 \in \mathcal{M}_2 = \{1, 2, \cdots, M_2\}$ is taken by a so-called "second decoder", $d_2 : \mathcal{Y}^N \times \mathcal{M}_1^{M_3} \times \mathcal{C} \rightarrow \mathcal{M}_2 \times \mathcal{M}_3$, hence

$$(\widehat{w_2}, \widehat{w_3}) = d_2(y^N, \widehat{\underline{w}_1}, C). \tag{7}$$

Finally a combiner, $c : \mathcal{M}_1 \times \mathcal{M}_2 \rightarrow \mathcal{M}$, based on w_2 and the selector output $\widehat{w_1}(\widehat{w_3})$, an index indicated by the second decoder from the index list $\underline{w}_1$, forms an estimate of the index of the item, hence

$$\widehat{w} = c(\widehat{w_1}(\widehat{w_3}), \widehat{w_2}). \tag{8}$$

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 quadruple (R_1, R_2, R_3, R) with $R \geq 0$ is achievable if for all $\epsilon > 0$ there exist for all N large enough mappings $d_1(\cdot)$, $d_2(\cdot, \cdot, \cdot)$, and $c(\cdot, \cdot)$ such that

$$\begin{aligned} \log_2(M_1) &\leq N(R_1 + \epsilon), \\ \log_2(M_2) &\leq N(R_2 + \epsilon), \\ \log_2(M_3) &\leq N(R_3 + \epsilon), \\ \log_2(M) &\geq N(R - \epsilon), \text{ and} \\ \Pr\{\widehat{W} \neq W\} &\leq \epsilon. \end{aligned} \tag{10}$$

We call R the identification rate, and R_1 and R_2 respectively cluster and refinement rate, and R_3 cluster list 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 quadruples $\mathcal{R}$ for our biometric identification system is given by*

$$\begin{aligned} \{(R_1, R_2, R_3, R) : & R_1 \geq I(X, Y; U), \\ & R_2 \geq \max(0, R - I(X; U)), \\ & R_3 \geq I(X; U|Y), \\ & 0 \leq R \leq I(X; Y), \\ & \textit{for } P(x, y, u) = Q_b(x) Q_c(y|x) P(u|x, y), \\ & \textit{where } |\mathcal{U}| \leq |\mathcal{Y}| \cdot |\mathcal{X}| + 2\}. \end{aligned} \tag{11}$$

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:

$$\log_2(M_1) \geq H(W_1(W_3)) \geq I(X^N, Y^N; W_1(W_3))$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} I(X_j, Y_j; W_1(W_3)|X^{j-1}, Y^{j-1})\\
&\overset{(a)}{=} \sum_{j=1}^{N} I(X_j, Y_j; W_1(W_3), X^{j-1}, Y^{j-1})\\
&\overset{(b)}{=} \sum_{j=1}^{N} I(X_j, Y_j; U_j),
\end{aligned} \tag{12}
$$

where $W_1(W_3) \in \{1, \ldots M_1\}$, (a) follows from the fact that $H(X_j, Y_j|X^{j-1}, Y^{j-1}) = H(X_j, Y_j)$ since $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ are independent of each other, and (b) from definition $U_j \overset{\Delta}{=} (W_1(W_3), X^{j-1}, Y^{j-1})$ for $j = 1, 2, \cdots, N$. Next let J be a random variable taking values in $\{1, 2, \ldots, N\}$ with equal probability, and let $X = X_j$ and $Y = Y_j$, when $J = j$. Then

$$
\begin{aligned}
\sum_{j=1}^{N} I(X_j, Y_j; U_j) &= N[H(X_J, Y_J|J) - H(X_J, Y_J|U_J, J)]\\
&\overset{(c)}{=} N[H(X, Y) - H(X, Y|U_J, J)]\\
&= NI(X, Y; (U_J, J)) \overset{(d)}{=} NI(X, Y; U),
\end{aligned} \tag{13}
$$

where step (c) follows since $(X_1, Y_1), (X_2, Y_2), \cdots,$ and (X_N, Y_N) are identically distributed and $X_J = X, Y_J = Y$, and (d) from $U \overset{\Delta}{=} (U_J, J)$.

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 $F \overset{\Delta}{=} 1 + \Pr\{\widehat{W} \neq W\} \log_2(M)$, the series of (in)equalities:

$$
\begin{aligned}
&\log_2(M) = H(W) \leq H(W) - H(W|\widehat{W}) + F\\
&\leq I(W; \widehat{W}, W_1(W_3), W_2) + F\\
&\overset{(e)}{=} I(W; W_1(W_3)) + I(W; W_2|W_1(W_3)) + F\\
&\leq I(W, X^N; W_1(W_3)) + \log_2(M_2) + F\\
&\overset{(f)}{=} I(X^N; W_1(W_3)) + \log_2(M_2) + F\\
&= \sum_{j=1}^{N} I(X_j; W_1(W_3)|X^{j-1}) + \log_2(M_2) + F\\
&\leq \sum_{j=1}^{N} I(X_j; W_1(W_3), X^{j-1}, Y^{j-1}) + \log_2(M_2) + F\\
&\overset{(g)}{=} \sum_{j=1}^{N} I(X_j; U_j) + \log_2(M_2) + F\\
&\overset{(h)}{=} NI(X; U) + \log_2(M_2) + F.
\end{aligned} \tag{15}
$$

where (e) follows from the fact that $I(W; W_1(W_3), W_2, \widehat{W}) = I(W; W_1(W_3), W_2)$, (f) since $W - X^N - W_1(W_3)$, (g) from definition $U_j \overset{\Delta}{=} (W_1(W_3), X^{j-1}, Y^{j-1})$, and (h) similar to how (13) was obtained.

For a given query y^N, the first decoder constructs the list of clusters $\{W_1(1), \ldots, W_1(M_3)\}$ and therefore $W_1(W_3) \in \{w_1(1), \ldots, w_1(M_3)\}$. Consequently, we can obtain for the range M_3 of the second decision that:

$$
\begin{aligned}
&\log_2(M_3) \geq H(W_1(W_3)|Y^N) = H(W_1(W_3)) - I(W_1(W_3); Y^N)\\
&= H(W_1(W_3)) - I(W_1(W_3); X^N, Y^N) + I(W_1(W_3); X^N|Y^N)\\
&\geq I(W_1(W_3); X^N|Y^N) = \sum_{j=1}^{N} I(W_1(W_3); X_j|X^{j-1}, Y^N)\\
&= \sum_{j=1}^{N} H(X_j|X^{j-1}, Y^N) - H(X_j|W_1(W_3), X^{j-1}, Y^{j-1}, Y_j, Y_{j+1}^N)\\
&\overset{(a)}{=} \sum_{j=1}^{N} H(X_j|Y_j) - H(X_j|W_1(W_3), X^{j-1}, Y^{j-1}, Y_j, Y_{j+1}^N)\\
&\overset{(b)}{\geq} \sum_{j=1}^{N} H(X_j|Y_j) - H(X_j|U_j, Y_j)\\
&\overset{(c)}{=} NI(X; U|Y)
\end{aligned} \tag{16}
$$

where (a) follows since $(X_1, Y_1), (X_2, Y_2), \cdots,$ and (X_N, Y_N) are independent, (b) results from definition $U_j \overset{\Delta}{=} (W_1(W_3), X^{j-1}, Y^{j-1})$ and conditioning reduces entropy, and (c) similar to how (13) was obtained.

Finally consider the number M of individuals:

$$
\begin{aligned}
\log_2(M) = H(W) &\leq I(W; \widehat{W}) + F\\
&\overset{(a)}{\leq} I(X^N; Y^N) + F \overset{(b)}{=} \sum_{n=1}^{N} I(X_n; Y_n) + F\\
&= NI(X_N; Y_N|N) + F \overset{(c)}{\leq} NI(X; Y) + F.
\end{aligned} \tag{17}
$$

where (a) 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 (b) from the fact that $(X_1, Y_1), (X_2, Y_2), \cdots, (Y_N, Y_N)$ are independent, (c) since these pairs are identically distributed and since $(X, Y) = (X_j, Y_j)$ for $J = j$.

Assume that (R_1, R_2, R_3, R) is achievable. Then for all blocklengths N and small enough $\epsilon > 0$, using $F \leq 1 + \epsilon \log_2(M)$, we obtain from (12) and (13), (14) and (15), (16) and (17) that

$$
\begin{aligned}
N(R_1 + \epsilon) &\geq \log_2(M_1) \geq NI(X, Y; U),\\
N(R_2 + \epsilon) &\geq \log_2(M_2) \geq 0,\\
N(R_2 + \epsilon) &\geq \log_2(M_2) \geq \log_2(M) - NI(X; U) - F,\\
&\geq (1 - \epsilon)N(R - \epsilon) - 1 - NI(X; U),\\
N(R_3 + \epsilon) &\geq \log_2(M_3) \geq NI(X; U|Y),\\
N(R - \epsilon) &\leq \log_2(M) \leq \frac{1}{1 - \epsilon}(NI(X; Y) + 1),
\end{aligned} \tag{18}
$$

for some $p(x, y, u) = Q_b(x)Q_c(y|x)P(u|x, y)$. From (18) 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 $P(x, y, u) = Q_b(x)$ $Q_c(y|x)P(u|x, y)$, and identification rate $0 \leq R \leq I(X;Y)$.

We first use a random coding argument to construct a collection of covering sequences $u^N(1), u^N(2), \cdots, u^N(M_1)$, where we take $M_1 = 2^{N(I(U;X,Y)+5\epsilon)}$. Averaged over the random covering code, the probability that a pair of sequences (x^N, y^N), i.i.d. according to $P(x, y) = \sum_u P(x, y, u)$ occurs, such that $(x^N, y^N, u^N(w_1)) \notin \mathcal{A}_\epsilon^{(N)}(XYU)$ 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 typical with an i.i.d. pair (x^N, y^N) of at least $1 - 3\epsilon$.

During enrollment, after sequence $x^N(w)$ was generated, for $w = 1, 2, \cdots, M$, the system finds out which $u^N(w_1)$ are jointly typical with $x^N(w)$ for $w_1 \in \{1, 2, \cdots, M_1\}$. In this way the system creates index-lists $\mathcal{L}(w_1) = \{w : (x^N(w), u^N(w_1)) \in \mathcal{A}_\epsilon^{(N)}(XU)\}$, one for each w_1. An error occurs if the cardinality of the list $\mathcal{L}(w_1)$ that contains $x^N(w)$ is larger than M_2. Based on the Markov inequality, it can be shown that this probability is not larger than 2ϵ for $M_2 = 2^{N(R-I(X;U)+4\epsilon)}$ and N large enough. These index-lists are available to the second decoder and the combiner.

During identification, the first decoder upon receiving y^N chooses list-indexes $\widehat{w_1} = \{\widehat{w_1}(1)), \ldots, \widehat{w_1}(M_3)\}$, $\widehat{w_1}(i) \in \{1, \ldots, M_1\}, 1 \leq i \leq M_3$ such that covering sequences $u^N(\widehat{w_1}(1)), \ldots, u^N(\widehat{w_1}(M_3))$ are jointly typical with y^N i.e. $(y^N, u^N(\widehat{w_1}(i)) \in \mathcal{A}_\epsilon^{(N)}(YU), 1 \leq i \leq M_3$. An error occurs if there are more than M_3 sequences $u^N(w_1)$ jointly typical with y^N. Based on the Markov inequality, it can be shown that this probability is not larger than 2ϵ for $M_3 = 2^{N(I(X;U|Y)+9\epsilon)}$ and N large enough. Note that the first decoder makes at most M_1 cluster-checks. If no error is declared the first decoder sends the indeces $\widehat{w_1}$ to the second decoder and the combiner.

Next the second decoder chooses a single index-pair $(\widehat{w_3}, \widehat{w_2})$ from lists of list $\mathcal{L}(\widehat{w_1}(1)), \ldots \mathcal{L}(\widehat{w_1}(M_3))$ such that $(x^N(\widehat{w_1}(\widehat{w_3}), \widehat{w_2}), y^N, u^N(\widehat{w_1}(\widehat{w_3}))) \in \mathcal{A}_\epsilon^{(N)}(XYU)$. If such an index cannot be found, an error is declared. Note that the informed decoder makes at most $M_3 M_2$ refinement-checks.

We have seen that the probability that the actual sequence $x^N(w)$ doesn't lead to joint typicality with y^N and some $u^N(w_1)$ is smaller than 3ϵ. The probability that some "other" index $w' \neq w$ results in joint typicality (and is in the lists $u^N(\widehat{w_1}(1)), \ldots, u^N(\widehat{w_1}(M_3))$) can be made $\leq \epsilon$ for $M = 2^{N(R-4\epsilon)}$ and N large enough.

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 $\mathcal{X} \times \mathcal{Y}$ and consider the $|\mathcal{X}| \cdot |\mathcal{Y}| + 2$ continuous functions of $P \in \mathcal{D}$ defined as

$$\begin{aligned} \phi_{x,y}(P) &= P(x, y) \text{ for all but one } (x, y), \\ \phi_{X,Y}(P) &= H_P(X, Y), \\ \phi_X(P) &= H_P(X), \\ \phi_Y(P) &= H_P(Y), \end{aligned} \tag{19}$$

where in the last two equations we use $\Pr\{X = x\} = \sum_y P(x, y)$ and $\Pr\{Y = y\} = \sum_x P(x, y)$. By the Fenchel-Eggleston strengthening of the Caratheodory lemma (see Wyner and Ziv [10]) there are $|\mathcal{X}| \cdot |\mathcal{Y}| + 2$ elements $P_u \in \mathcal{D}$ and α_u that sum to one, such that

$$\begin{aligned} P(x, y) &= \sum_{u=1}^{|\mathcal{X}||\mathcal{Y}|+2} \alpha_u \phi_y(P_u) \text{ for all but one } (x, y), \\ H(X, Y|U) &= \sum_{u=1}^{|\mathcal{X}||\mathcal{Y}|+2} \alpha_u \phi_{X,Y}(P_u), \\ H(X|U) &= \sum_{u=1}^{|\mathcal{X}||\mathcal{Y}|+2} \alpha_u \phi_X(P_u), \\ H(Y|U) &= \sum_{u=1}^{|\mathcal{X}||\mathcal{Y}|+2} \alpha_u \phi_Y(P_u). \end{aligned} \tag{20}$$

The entire probability distribution $\{Q(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ and consequently the entropies $H(X, Y)$, $H(X)$ and $H(Y)$ are now specified and therefore also $I(X, Y; U)$, $I(Y; U)$ and $I(X; U)$. This implies that cardinality $|\mathcal{X}| \cdot |\mathcal{Y}| + 2$ suffices.

Remark 1. It can be shown that under Markov chain $X \leftrightarrow Y \leftrightarrow U$ condition, i.e., $P(x, y, u) = Q_b(x)Q_c(y|x)P(u|y)$, the region of achievable rate $\mathcal{R}$ reduces to the following triples for the biometric identification system

$$\begin{aligned} \{(R_1, R_2, R) : & R_1 \geq I(Y; U), \\ & R_2 \geq \max(0, R - I(X; U)), \\ & 0 \leq R \leq I(X; Y)\}, \end{aligned} \tag{21}$$

that coincides to the results shown by Willems [9]. It should be noted that under this Markov condition $R_3 = 0$, which means that only a single u^N is sent to the second decoder.

IV. Search–Memory complexity

In this section, we consider the search–memory complexity of the two–stage decoding scheme explained in Section II-A.

A. Memory–complexity exponent

Using the proposed scheme, we try to cover the space $\mathcal{X}^N \times \mathcal{Y}^N$. Fixing a covering–channel $P(u|x, y)$, we generate a covering–code $C_u = \{u^N(1), \ldots, u^N(M_1)\}$ of $M_1 \approx 2^{NI(U;X,Y)}$ codewords $u^N(w_1), 1 \leq w_1 \leq M_1$, according to

$$P(u) = \sum_{x,y} Q_b(x)Q_c(y|x)P(u|x, y). \tag{22}$$

The memory–complexity exponent related to the covering–code C_u is approximately $I(U; X, Y)$.

In the enrolment phase, for covering code $u^N(w_1) \in C_u$ the list $\mathcal{L}(w_1)$ will be constructed. From Theorem 1

- if $R > I(U;X)$ there are approximately $2^{N[R-I(U;X)]}$ sequences $x^N(w) \in C$ jointly typical with a $u^N(w_1) \in C_u$. The cardinality of the set $\mathcal{L}(w_1)$ is approximately $2^{N[R-I(U;X)]}$,
- if $R < I(U;X)$ there is at most a single sequence $x^N(w) \in C$ jointly typical with a $u^N(w_1) \in C_u$. The cardinality of the set $\mathcal{L}(w_1)$ is at most 1.

Consequently, it can be shown that the memory–complexity exponent of the scheme is

$$\max\{R + I(U;Y|X), I(U;X,Y)\}. \tag{23}$$

B. *Search–complexity exponent*

According to the proposed decoding procedure, for a given query y^N, the first decoder constructs the list $\underline{w}_1$ with the cardinality approximately $M_3 \approx 2^{NI(U;X|Y)}$ using the following strategy

- checks all codewords $u^N(w_1) \in C_u$ and finds out what codeword $u^N(\widehat{w_1})$ is jointly typical with y^N, i.e., $(y^N, u^N(\widehat{w_1})) \in \mathcal{A}_\epsilon^{(N)}(UY)$,

It should be noted that the search–complexity exponent related to the first decoder is approximately $\approx I(U;X,Y)$.

Next, the second decoder for each $\widehat{w_1}(w_3) \in \underline{w}_1$ considers the set of sequences $x^N(w) \in \mathcal{L}(\widehat{w_1}(w_3))$. All these sequences should be checked to determine the one that $(x^N(\widehat{w_1}(\widehat{w_3}), \widehat{w_2}), u^N(\widehat{w_1}(\widehat{w_3})), y^N) \in \mathcal{A}_\epsilon^{(N)}(XYU)$. This requires,

- when $R > I(U;X)$, $2^{NI(U;X|Y)}.2^{N[R-I(U;X)]} = 2^{N[R+I(U;X|Y)-I(U;X)]}$ refinement checks,
- when $R < I(U;X)$, $2^{NI(U;X|Y)}$ refinement checks.

Therefore, the search–complexity exponent corresponding to the second decoder turns out to be approximately $\max\{R + I(U;X|Y) - I(U;X), I(U;X|Y)\}$.

Consequently, it can be shown that the search–complexity exponent related to the two–stage decoding scheme is approximately

$$\max\{I(U;X,Y), \\ \max\{R + I(U;X|Y) - I(U;X), I(U;X|Y)\}\}. \tag{24}$$

V. Binary Example

We consider here a system with binary uniform 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 $y \in \{0,1\}$.

Let $u \in \{0,1\}$, we have generated $10,000$ trials of the covering–channel $P(u|x,y)$ at random, the four relevant probabilities are uniformly distributed over $[0,1]$. Fig. 2 (red o' s) shows the exponent of the search–memory complexity of the binary biometric system with the rate $R = 0.5$ and $q = 0.1$, using (23) and (24). On the other hand, Fig. 2 (blue ×' s) shows the exponent of the search–memory complexity under Markov chain $X \leftrightarrow Y \leftrightarrow U$ condition similarly to [9].

From Fig. 2, one can conclude that the proposed scheme in this context can achieve a new search-complexity region with less search–memory complexity compared to the scheme proposed in [9].

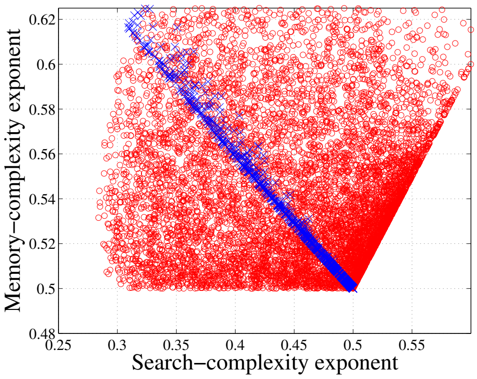

Fig. 2. (red o) indicate the exponent of the search–memory complexity of the proposed scheme using the covering–channel $P(u|x,y)$, vs. (blue ×) show the complexity using the covering–channel $P(u|y)$ under the condition $X \leftrightarrow Y \leftrightarrow U$.

VI. Concluding Remarks

We have investigated the triple trade-off identification–rate, search and memory complexities for a two-stage search procedure in a biometric identification system. We have shown a new result on search–memory complexity region of capacity achieving identification systems. We have only considered a two-step system here. It is not so difficult however to find the fundamental limits for multi-stage systems.

Acknowledgment

This paper was partially supported by SNF projects 200020-134595.

References

[1] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, “On the capacity of a biometrical identification system,” in *IEEE Int. Symp. Inform. Th.*, june-4 july 2003, p. 82.

[2] R. Weber, H.-J. Schek, and S. Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” in *24th VLDB Conf.*, August 24-27 1998, pp. 194–205.

[3] F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, “Performance analysis of content-based identification using constrained list-based decoding,” *IEEE Trans. on Inf. Foren. and Sec.*, vol. 7, no. 5, pp. 1652–1667, 2012.

[4] S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak, “Information-theoretical analysis of private content identification,” in *IEEE Inf. Th. Work.*, Aug.30-Sep.3 2010.

[5] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín, “Searching in metric spaces,” *ACM Comput. Surv.*, vol. 33, no. 3, pp. 273–321, Sep. 2001.

[6] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in *3rd Int. Conf. on Music Inform. Retriev., ISMIR*, Oct. 13-17 2002, pp. 107 – 115.

[7] E. Tuncel, P. Koulgi, and K. Rose, “Rate-distortion approach to databases: storage and content-based retrieval,” *IEEE Trans. Inform. Th.*, vol. 50, no. 6, pp. 953 – 967, june 2004.

[8] E. Tuncel, “Capacity/storage tradeoff in high-dimensional identification systems,” in *IEEE Int. Symp. Inform. Th.*, july 2006, pp. 1929 –1933.

[9] F. Willems, “Searching methods for biometric identification systems: Fundamental limits,” in *IEEE Int. Symp. Inform. Th.*, july 2009, pp. 2241 –2245.

[10] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Trans. Inform. Th.*, vol. 22, no. 1, pp. 1 – 10, jan 1976.