Information-Theoretic Analysis of Content Based Identification for Correlated Data

Farzad Farhadzadeh, Sviatoslav Voloshynovskiy, Oleksiy Koval and Fokko Beekhof
Computer Science Department, University of Geneva, Geneva, Switzerland
Email: {Farzad.Farhadzadeh, svolos, Oleksiy.Koval, Fokko.Beekhof}@unige.ch

***Abstract*—A number of different multimedia fingerprinting algorithms and identification techniques were proposed and analyzed recently. This paper presents a content identification setup for a class of multimedia data that can be modeled by a Gauss-Markov process. We advocate a constrained order statistics decoding scheme based on digital fingerprints extracted from correlated data to identify contents. Finally, we investigate the fundamental limits of the proposed setup by deriving bounds on the miss and false acceptance probabilities.**

I. Introduction

In today's world, digital reproduction tools and user generated content (UGC) websites such as Youtube, Flicker, etc., have performed an impressive evolution, providing professional solutions to various groups of users. Besides these obvious advantages, at the same time these tools have raised concerns about copyrighted content protection. Thus, content based identification (CBI) becomes a critical issue.

Multimedia applications use high-dimensional data that are frequently privacy-sensitive. The data are also highly correlated in spatial and time coordinates. Moreover, multimedia data might be severely distorted due to the habitual chain of processing, transcoding, communication and storage. Therefore, in order to design a robust CBI system, one must consider not only its ability to handle high-dimensional, correlated and privacy-sensitive data but also its performance under strong distortions.

There exist several approaches to deal with the former problems, such as robust hashing and digital fingerprinting. A *digital fingerprint* represents a short, robust and distinctive content description. The main idea behind digital fingerprinting approaches is to extract digital fingerprints of a lower dimensionality with a maximum possible entropy, i.e., in the binary case the bits of digital fingerprints should be independently and equally likely 0s and 1s. However, since multimedia data are correlated, one of the principle tasks of a dimensionality reduction transform is to eliminate correlation between the samples. A mapper that possesses such properties is the Karhunen-Loève transform (KLT) [1]. However, the price that must be paid for this optimality is its data dependence and the necessity of updating the transform matrix for new entries. In order to allay this dependence, several approximations of the KLT were proposed such as the Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) [1]. The basis vectors of these transforms are fixed and independent of the statistics of their inputs. Actually, the basis vectors of DCT and DWT are optimized for locally correlated data. However, the main drawback of such fixed basis transforms consists in the public disclosure of the basis vectors, which is rarely acceptable for multimedia security applications [2].

One solution to overcome this privacy/security shortcoming is a randomized mapper that can be designed based on random projections (RP) [3]. The RP have been the object of much interest due to their ability for distance preservation, which is also recognized in the Compressed Sensing community for sparse data [4]. However, the drawback of this approach is that multimedia data are real valued correlated signals but not sparse samples. Although the decorrelation property of orthogonal transforms is well-known [1], the RP are based on approximately orthogonal bases. Therefore, the statistics of projected data, i.e., the covariance matrix, are not well justified. On the other hand, prior knowledge of the statistics of extracted digital fingerprints is crucial for evaluation of the performance of CBI systems. As mentioned above, the second important issue of CBI systems is their ability to deal with highly distorted data where the performance of unique decoding is characterized by a high probability of error. As a possible solution, one can envision the use of the Forney [5] list decoding approach as mentioned in [6]. However, in many identification problems, the final sink of information will be a human being. This restriction makes this type of list decoding undesirable, due to the high variability of the list size, i.e., for very noisy environment the list can be exceedingly long. Another solution, which is proposed by the authors in [7], is the constrained *Order Statistics List Decoding* (OSLD) approach. In the constrained OSLD, which is indeed a combination of Elias [8] and Forney list decoding approaches in communication setups, a limited number of candidates with the largest likelihood functions that can satisfy a specific threshold is selected. The performance analysis accomplished in [7] is based on the assumption that contents are generated independently and identically, which is not true for multimedia data [1]. Moreover, one is often interested in choosing system parameters, i.e., the length of digital fingerprints, the threshold and the maximum number of final candidates, to ensure that the probabilities of miss and false acceptance are below certain bounds. Hence, in this paper, we derive bounds on the probabilities of miss and false acceptance using fingerprints of a given length.

The main contribution of this paper can be summarized as follows: we introduce an identification setup by using a

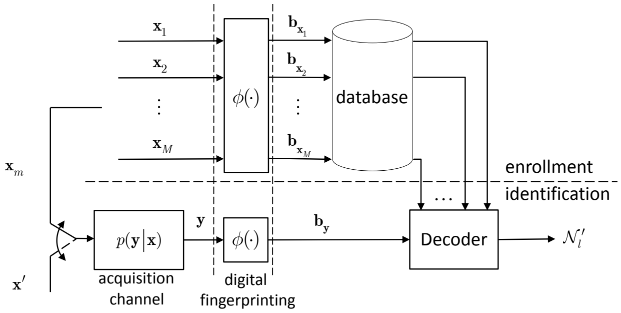

Fig. 1. The general setup for CBI based on binary fingerprints.

constrained OSLD for a class of multimedia data that can be modeled by a correlation-based model like a Gauss-Markov process, which captures image pixel dependencies directly in the coordinate domain [1]. Then, we analyze the fundamental performance limit by deriving bounds on the miss and false acceptance probabilities.

The outline of this paper is as follows. In Section II, we introduce the identification system. In Section III, we analyze the statistics of extracted digital fingerprints. The fundamental limits of the proposed identification setup is considered in Section IV. Finally, the conclusions are presented in Section V.

Notations: We use capital letters X to denote scalar random variables and $\mathbf{X} = \{X[i]\}_{i=1}^N$ to denote vector random variables. Corresponding small letters x and $\mathbf{x} = \{x[i]\}_{i=1}^N$ denote realizations of scalar and vector random variables, respectively. $\mathbf{x}^\dagger$ denotes the transpose of $\mathbf{x}$. We use $H_2(\cdot)$ to denote a binary entropy. $\mathcal{N}(\mu, \sigma_X^2)$ stands for the Gaussian distribution with mean μ and variance σ_X^2. $\mathcal{B}(N, p)$ denotes the Binomial distribution with N trails and probability of success p. $E[\cdot]$ designates the expectation. $\mathcal{D}(\tau \| \delta)$ stands for the divergence between $0 \leq \tau \leq 1$ and $0 \leq \delta \leq 1$.

II. Identification Setup

The identification setup under analysis shown in Fig. 1 consists of two main phases: *content enrollment* and *content identification*.

Regarding storage requirements and computational complexity, the cost of identification could be enormous for large databases, especially in multimedia applications. Therefore, in the content enrollment phase, the digital fingerprints are extracted from contents to be identified and stored in a *Database*. The database is a collection of M binary vectors denoted by $\mathbf{b}_{\mathbf{x}_m} \in \{0,1\}^L, m \in \{1, \ldots, M\}$, where $\mathbf{b}_{\mathbf{x}_m} = \phi(\mathbf{x}_m)$ is a digital fingerprint extracted from the content $\mathbf{x}_m, \mathbf{x}_m \in \mathcal{X}^N$, which is drawn from a common stationary distribution $p(\mathbf{x})$. Here $\phi(\cdot)$ is a digital fingerprint extraction function that can be key-dependent.

In the content identification phase, for a given query $\mathbf{y}$ the digital fingerprint is extracted following the same approach as in the enrollment phase, i.e., $\mathbf{b}_\mathbf{y} = \phi(\mathbf{y})$. Then, the decoder constructs a list of indices of entries $\mathcal{N}_l'$ which are the most likely related to the query. Otherwise, it produces an erasure, $\mathcal{N}_l' = \emptyset$.

a) Identification Problem: In the event the query digital fingerprint $\mathbf{b}_\mathbf{y}$ is related to some element $\mathbf{b}_{\mathbf{x}_m}$ of the database, one can assume that this relationship can be modeled by a binary channel with the transition probability $p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m})$. If the query digital fingerprint $\mathbf{b}_\mathbf{y}$ is unrelated to any database entry, we assume that $\mathbf{b}_\mathbf{y}$ is drawn from $p(\mathbf{b}_\mathbf{y}) = \sum_{\mathbf{b}_{\mathbf{x}_m} \in \{0,1\}^L} p(\mathbf{b}_\mathbf{x}) p(\mathbf{b}_\mathbf{y}|\mathbf{b}_\mathbf{x})$. Therefore, we can define the content identification problem as a composite hypothesis test:

$$\begin{cases} \mathcal{H}_0 \ : \mathbf{B}_\mathbf{y} \sim p(\mathbf{b}_\mathbf{y}) \\ \mathcal{H}_m : \mathbf{B}_\mathbf{y} \sim p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m})), \end{cases} \quad (1)$$

where $\mathcal{H}_0$ and $\mathcal{H}_m$ correspond to the cases that the query digital fingerprint $\mathbf{b}_\mathbf{y}$ is unrelated to any database entry, and the query digital fingerprint $\mathbf{b}_\mathbf{y}$ is related to the m^{th} entry of database, respectively.

b) Decoder: We define the constrained OSLD as follows:

1) The likelihood functions, $p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m}), 1 \leq m \leq M$, for all entries of the database are evaluated.
2) The N_l indices with the largest likelihood functions are chosen which form a set $\mathcal{N}_l$. The parameter N_l is referred to as the primary list size.
3) The final output set of the decoder is defined by $\mathcal{N}_l' = \{m \in \mathcal{N}_l : p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m}) \geq e^{\gamma L}\}$, where the parameter γ controls the number of final candidates.

The performance metrics of the identification setup are defined by the probability of miss:

$$P_m = \sum_{m=1}^{M} \Pr\{(m \notin \mathcal{N}_l) \cup p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m}) < e^{\gamma L} | \mathcal{H}_m\} \Pr\{\mathcal{H}_m\}, \quad (2)$$

and the probability of false acceptance:

$$P_f = \Pr\{\mathcal{N}_l' \neq \varnothing | \mathcal{H}_0\}. \quad (3)$$

III. Digital Fingerprint Extraction and Statistical Analysis

The digital fingerprint extraction function $\phi(\cdot)$ works as follows:

1) The dimensionality of some content $\mathbf{x}_m$ or a query $\mathbf{y}$ is reduced from N to L by applying random projections (RP) [9], which are approximately *orthoprojectors*, i.e., $\mathbf{w}\mathbf{w}^\dagger \approx \mathbf{I}_L$ where $\mathbf{w} \in \frac{1}{\sqrt{N}}\{\pm 1\}^{L \times N}$ with $W_{ij} \sim$ Bernoulli$(\frac{1}{2})$, $1 \leq i \leq L$ and $1 \leq j \leq N$. For a given $\mathbf{w}$, the projected $\tilde{\mathbf{x}}_m$ and $\tilde{\mathbf{y}}$ are obtained by $\tilde{\mathbf{x}}_m = \mathbf{w}\mathbf{x}_m$ and $\tilde{\mathbf{y}} = \mathbf{w}\mathbf{y}$.
2) L-length binary digital fingerprints, $\mathbf{b}_\mathbf{y}$ and $\mathbf{b}_{\mathbf{x}_m}$, are derived by taking the sign of the projected data, i.e., $\mathbf{b}_{\mathbf{x}_m} = \{\text{sign}(\tilde{x}_m[i])\}_{i=1}^L$ and $\mathbf{b}_\mathbf{y} = \{\text{sign}(\tilde{y}[i])\}_{i=1}^L$.

A. The Statistics of Content Digital Fingerprints

In this Section, we investigate the statistics of content digital fingerprints obtained by the RP. We assume that the content $\mathbf{X}$ is a Gauss-Markov process with the covariance matrix $\mathbf{K}_{\mathbf{x}\mathbf{x}}$. This is a simple but often-used model in image processing [1]. Then, the covariance matrix of the projected data is obtained by:

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = E[\mathbf{w}\mathbf{X}\mathbf{X}^\dagger\mathbf{w}^\dagger] = \mathbf{w}\mathbf{K}_{\mathbf{x}\mathbf{x}}\mathbf{w}^\dagger, \quad (4)$$

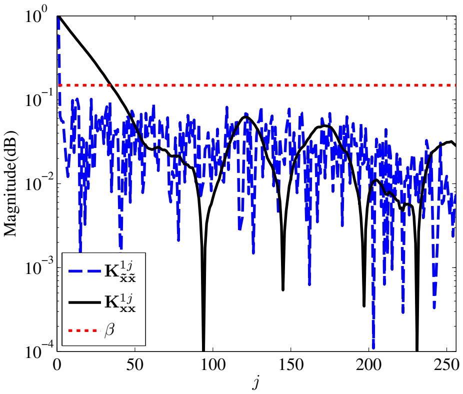

Fig. 2. The first 256 elements of the first row of $\mathbf{K}_{\mathbf{x}\mathbf{x}}$ and $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, where $\mathbf{x}$ is generated from the Gauss-Markov process with $\rho = 0.95$ and $\sigma_X^2 = 1$. $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have the length of $N = 2^{15}$ and $L = 2^8$, respectively.

where $\mathbf{K}_{\mathbf{x}\mathbf{x}}$ is defined by [1]:

$$\mathbf{K}_{\mathbf{x}\mathbf{x}} = \sigma_X^2 \begin{bmatrix} 1 & \rho & \ldots \rho^{N-1} \\ \rho & 1 & \ldots \rho^{N-2} \\ \vdots & \vdots & \ddots \quad \vdots \\ \rho^{N-1} & \rho^{N-2} & \ldots \quad 1 \end{bmatrix}, \tag{5}$$

where σ_X^2 and $0 \leq \rho < 1$ are the variance and the normalized correlation coefficient, respectively. We use the following proposition for statistical modeling of projected data.

Proposition 1 (*Decorrelation property of RP*)**.** Let the elements of the RP matrix, $\mathbf{W}$ of size $L \times N$ and $L > 1$, be generated by the probability mass function (PMF) $\Pr\{W_{ij} = +\frac{1}{\sqrt{N}}\} = \Pr\{W_{ij} = -\frac{1}{\sqrt{N}}\} = \frac{1}{2}$, and $\mathbf{X}$ be a real zero-mean random vector modeled as the Gauss-Markov process with variance σ_X^2 and normalized correlation coefficient ρ. Then, we have:

$$\Pr\left\{\max_{i \neq j} |\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}| > \beta\sigma_X^2\right\} < \frac{1}{L}, \tag{6a}$$

$$\Pr\left\{\max_{i} |\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \sigma_X^2| > \alpha\sigma_X^2\right\} < \frac{2}{L^{\left(\frac{1}{\rho}\right)}}, \tag{6b}$$

where $\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ denotes the $(i,j)^{\text{th}}$ element of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, $\beta = \sqrt{\frac{6}{N}\left(\frac{1+\rho^2}{1-\rho^2}\right)\ln L}$, $\alpha = \sqrt{\frac{4}{N}\left(\frac{\rho}{1-\rho}\right)\ln L}$ and $\alpha < \beta$.

Proof: Appendix A. ■

Remark 1. For a sufficiently large N and L, $L \leq N$, $\alpha \to 0$ and $\beta \to 0$, $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ asymptotically converges to $\sigma_X^2\mathbf{I}_L$ with high probability. Moreover, from the fact that the content source is a Gauss-Markov process, which implies that the content vector $\mathbf{x}$ is jointly Gaussian, and RP is a linear transform, the projected data $\tilde{\mathbf{x}}$ follows the jointly Gaussian distribution, i.e., $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})$. Therefore, since elements of $\tilde{\mathbf{x}}$ are asymptotically uncorrelated, $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \approx \sigma_X^2\mathbf{I}_L$, one can conclude that $\tilde{\mathbf{x}}$ are asymptotically independent and identically distributed (i.i.d.). In addition, the digital fingerprint extracted from $\tilde{\mathbf{x}}$ asymptotically follows $\mathcal{B}(L, \frac{1}{2})$ due to symmetry of the Gaussian distribution function.

The decorrelation property of the RP is illustrated in Fig. 2. All off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ are below the evaluated threshold β.

B. The Statistics of Query Digital Fingerprint

Consider the query $\mathbf{y}$ to be a noisy version of a content that can be modeled as a Gauss-Markov process and is observed through an Additive White Gaussian Noise (AWGN) channel, $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2\mathbf{I}_N)$ and σ_Z^2 is the variance of the noise. At the output of the first step of the digital fingerprinting, we have $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} + \tilde{\mathbf{Z}}$. From Proposition 1, we can asymptotically assume that the projected content part of the query, $\tilde{\mathbf{X}}$, follows the distribution $\mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I}_L)$. To justify the statistics of $\tilde{\mathbf{Z}}$, we have the following collary.

Collary 1 (*i.i.d. preservation property of RP*)**.** Let the RP matrix, $\mathbf{W}$, be generated the same as in Proposition 1, and $\mathbf{Z}$ are drawn i.i.d. from a common stationary distribution with variance σ_Z^2. Then, the diagonal elements of covariance matrix of the projected noise $\tilde{\mathbf{Z}} = \mathbf{W}\mathbf{Z}$ are equal to σ_Z^2, i.e., $\forall i, \mathbf{K}^{ii}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} = \sigma_Z^2$, and all off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$ satisfies:

$$\Pr\left\{\max_{i \neq j} |\mathbf{K}^{ij}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}| > \delta\sigma_Z^2\right\} < \frac{1}{L}, \tag{7}$$

where $\delta = \sqrt{\frac{6}{N}\ln L}$.

Proof: Appendix B. ■

Remark 2. For a sufficiently large N and $L, L \leq N$, $\delta \to 0$, $\mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$ asymptotically converges to $\sigma_Z^2\mathbf{I}_L$ with high probability. Moreover, $\mathbf{Z}$ is i.i.d. Gaussian and RP is a linear transform, $\tilde{\mathbf{Z}}$ is jointly Gaussian whose elements are asymptotically uncorrelated, i.e., $\tilde{\mathbf{Z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}), \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} \approx \sigma_Z^2\mathbf{I}_L$, thus $\tilde{\mathbf{Z}}$ follows asymptotically i.i.d. Gaussian. Consequently, the transformed channel is a discrete memoryless channel, i.e., $p(\mathbf{b}_\mathbf{y}|\mathbf{b}_\mathbf{x}) = \prod_{i=1}^{L} p(b_y[i]|b_x[i])$.

Remark 3. From Proposition 1 and Collary 1, $\tilde{\mathbf{Y}}$ is the summation of two independent random vectors $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Z}}$ where $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}), \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \approx \sigma_X^2\mathbf{I}_L$ and $\tilde{\mathbf{Z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}), \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} \approx \sigma_Z^2\mathbf{I}_L$. Thus, $\tilde{\mathbf{Y}}$ is a jointly Gaussian distributed random vector with asymptotically uncorrelated elements which implies their independence. Moreover, one can conclude that $\mathbf{B}_\mathbf{y}$ asymptotically follows $\mathcal{B}(L, \frac{1}{2})$ due to symmetry of the Gaussian distribution function. Conditioned on $\mathcal{H}_m$, the relation between $\mathbf{b}_{\mathbf{x}_m}$ and $\mathbf{b}_\mathbf{y}$ can be modeled by the Binary Symmetric Channel (BSC) with crossover probability $P_b = \frac{1}{\pi}\arctan\left(\frac{\sigma_Z}{\sigma_X}\right)$ [9].

IV. Bounds on Error Probabilities

In this section, we derive bounds on the miss and false acceptance probabilities based on the obtained results.

A. Miss Probability Bound

From Remark 3, conditioned on $\mathcal{H}_m$, the transition probability of the BSC is given by $p(\mathbf{b}_\mathbf{y}|\mathbf{b}_{\mathbf{x}_m}) = P_b^{d_m}(1-P_b)^{L-d_m}$ that is a decreasing function of the Hamming distance $d_m \triangleq d_H(\mathbf{b}_\mathbf{y}, \mathbf{b}_{\mathbf{x}_m})$ for $P_b \in [0, 0.5]$, which is a realization of D_m and can be considered as a sufficient statistic. From Remark 1, all entries of the database are i.i.d., and since they can be queried equally likely, i.e., $\Pr\{\mathcal{H}_m\} = \frac{1}{M}$, the overall probability of miss does not depend on the particular index and hence for $m = 1$:

$$
\begin{aligned}
P_m &= \Pr\{(m_1 \notin \mathcal{N}_l) \cup p(\mathbf{b}_{\mathbf{y}}|\mathbf{b}_{\mathbf{x}_1}) < e^{\gamma L}|\mathcal{H}_1\} \\
&= \Pr\{(1 \notin \mathcal{N}_l) \cup (D_1 > \eta L)|\mathcal{H}_1\} \\
&\overset{(a)}{=} \Pr\{(1 \notin \mathcal{N}_l) \cap (D_1 \leq \eta L)|\mathcal{H}_1\} + \Pr\{D_1 > \eta L|\mathcal{H}_1\},
\end{aligned}
\tag{8}
$$

where $\eta = \frac{\gamma - \ln(1-P_b)}{\ln(P_b/(1-P_b))}$ and (a) follows from the addition rule of probability . The first term in (8) is referred to as the *miss probability of the first kind*, P_m^{I}, and the second term is the *miss probability of the second kind*, P_m^{II}.

By using Remarks 1 and 3, conditioned on $\mathcal{H}_1$, the sufficient statistics mentioned above have the following distributions for m, $1 \leq m \leq M$:

$$
D_m \sim \begin{cases} \mathcal{B}(L, P_b), \text{for } m = 1, \\ \mathcal{B}(L, \frac{1}{2}), \text{ for } m \neq 1. \end{cases} \tag{9}
$$

Proposition 2 (*Miss probability bound*)**.** For the binary symmetric channel with the crossover probability P_b, the probability of miss of the constrained OSLD, for any η, $P_b < \eta < \frac{1}{2}$, is bounded by:

$$
\begin{aligned}
P_m &= P_m^{\mathrm{I}} + P_m^{\mathrm{II}} \leq \{\exp[-L(\ln 2 - R - H_2(\eta))]\}^{N_l} \\
&\quad + \exp[-L\mathcal{D}(\eta \| P_b)].
\end{aligned}
\tag{10}
$$

Proof: Appendix C. ■

Remark 4. For the case $N_l = 1$, i.e., Maximum Likelihood (ML) decoding, the obtained miss probability bound coincides with the result achieved in [2]. If $N_l > 1$, i.e., list decoding, P_m^{I} converges to 0 faster than for ML decoding.

Remark 5. For $P_b < \eta < \frac{1}{2}$ and if $R < \ln 2 - H_2(\eta)$ there exist fingerprints with the rate R and miss probability P_m such that $\lim_{L \to \infty} P_m = 0$.

B. False Acceptance Probability Bound

From Remark 3 , conditioned on $\mathcal{H}_0$, the sufficient statistics follows $D_m \sim \mathcal{B}(L, \frac{1}{2}), 1 \leq m \leq M$.

Proposition 3 (*False acceptance probability bound*)**.** For the binary symmetric channel with the crossover probability P_b, the average probability of false acceptance of the constrained OSLD, for any η, $P_b < \eta < \frac{1}{2}$, is bounded by

$$
P_f \leq \exp[-L(\ln 2 - R - H_2(\eta))]. \tag{11}
$$

Proof: Appendix D. ■

Remark 6. For $P_b < \eta < \frac{1}{2}$ and $R < \ln 2 - H_2(\eta)$ there exist fingerprints with the rate R and false acceptance probability P_f such that $\lim_{L \to \infty} P_f = 0$.

Remark 7. From Remarks 5, 6, both P_m and P_f go to zero as L goes to ∞. Moreover, it holds for η arbitrarily close to P_b. Therefore, the identification capacity $C_{\text{id}} = I(\mathbf{B}_{\mathbf{x}}, \mathbf{B}_{\mathbf{y}}) = \ln 2 - H_2(P_b)$ [10] is achievable.

V. Conclusions

In this paper, we present a theoretical analysis of the proposed CBI system in multimedia applications. A quite simple approach is introduced to extract digital fingerprints from multimedia data that can be modeled by a Gauss-Markov process.

Acknowledgment

The authors would like to thank Prof. M. Meckes and V. Balakirsky for the stimulating discussions. This paper was partially supported by SNF projects 200020-134595.

References

[1] A. K. Jain, *Fundamentals of digital image processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[2] S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh and T. Holotyak, “ Information-Theoretical Analysis of Private Content Identification ”, In Proc. of ITW, Dublin, Ireland, 2010.

[3] J. Fridrich, “Robust bit extraction from images”, In Proc. of MCS, 1999.

[4] M. Davenport, P. Boufounos, M. Wakin, R. Baraniuk, “Signal Processing With Compressive Measurements”, IEEE Journal of Selected Topics in Sig. Proc., vol. 4, no. 2, pp. 445–460, 2010.

[5] G. D. Forney, Jr, “Exponential error bounds for erasure, list and decision feedback schemes”, IEEETrans Inf. Theory, vol. IT-14, no. 2, pp. 206-220, Mar. 1968.

[6] P. Moulin, “Statistical modeling and analysis of content identification”, In. Proc. ITA, San Diego, CA, 2010.

[7] F. Farhadzadeh, S. Voloshynovskiy and Oleksiy Koval, “Performance Analysis of Identification System Based on Order Statistics List Decoder”, In Proc. of IEEE ISIT, Austin, TX, 2010.

[8] P. Elias, “List decodeing for noisy channels”, Tech. Rept. 335, Research Labratoary of Electronics, M.I.T, 1955.

[9] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun, “Conception and limits of robust perceptual hashing: toward side information assisted hash functions”, SPIE Photonics West, San Jose, USA, 2009.

[10] F. Willems, T. Kalker, J. Goseling, and J. Linnartz, “On the capacity of a biometrical system”, in Proc. 2003 IEEE ISIT, Yokohama, Japan.

[11] G. Caraux, O. Gascuel, “Bounds on distribution functions of order statistics for dependent variates”, Statistics & Probability Letters, vol. 14, no. 2, pp. 103–105, 1992.

[12] W. Hoeffding, “Probability Inequalities for Sums of Bounded Random Variables”, Journal of the American Statistical Association, vol. 58, no. 301, pp. 13–30, 1963.

Appendix A
Proof of Proposition 1

Proof: The elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ can be expanded as follows:

$$
\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{ij} = \sum_{r=1}^{N} \sum_{c=1}^{N} w_{ir} \mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc} w_{jc}. \tag{12}
$$

At first, we investigate upper off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{ij}, 1 \leq i < j \leq L$. One can represent these elements as a sum of N^2 independent and zero-mean random variables, $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{ij} = \sum_{r=1}^{N} \sum_{c=1}^{N} V_{rc}^{ij}$, where $V_{rc}^{ij} \in \frac{\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N}\{+1, -1\}, 1 \leq r \leq N, 1 \leq c \leq N$, which are almost surely bounded, i.e., $\Pr\{V_{rc}^{ij} \in [\frac{+\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N}, \frac{-\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N}]\} = \Pr\{V_r^{ij} = \frac{+\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N}\} + \Pr\{V_r^{ij} = \frac{-\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N}\} = 1$. In order to bound these elements, we compute the probability that the largest upper off-diagonal element of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ is greater than $\beta\sigma_X^2$, where β is a positive real value. This probability is given by:

$$
\begin{aligned}
&\Pr\left\{\max_{i \neq j} |\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{ij}| > \beta\sigma_X^2\right\} \overset{(a)}{\leq} \frac{L(L-1)}{2} \Pr\{|\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{ij}| > \beta\sigma_X^2\} \\
&\overset{(b)}{\leq} \frac{L(L-1)}{2} 2 \exp\left(-\frac{2 \frac{\beta^2\sigma_X^4}{N^4} N^4}{\sum_{r,c}(2 \frac{\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc}}{N})^2}\right) \\
&= L(L-1) \exp\left(-\frac{\beta^2\sigma_X^4}{\frac{2}{N^2} \sum_{r,c}(\mathbf{K}_{\mathbf{x}\mathbf{x}}^{rc})^2}\right),
\end{aligned}
\tag{13}
$$

where (a) follows from the fact that there are only $\frac{L(L-1)}{2}$ such random variables which are identically distributed [11] and (b) follows from the Hoeffding's inequality [1][12]. In order to bound the $\frac{1}{N^2}\sum_{r,c}(\mathbf{K}^{rc}_{\mathbf{xx}})^2$, from (5) one obtains:

$$\begin{aligned}&\frac{1}{N^2}\sum_{r=1}^{N}\sum_{c=1}^{N}(\mathbf{K}^{rc}_{\mathbf{xx}})^2\\ =&\frac{\sigma_X^4}{N^2}\left(N+2((N-1)\rho^2+\ldots+\rho^{2(N-1)})\right)\\ <&\frac{\sigma_X^4}{N}(1+2(\rho^2+\ldots+\rho^{2(N-1)}))\overset{(a)}{\leq}\frac{\sigma_X^4}{N}\left(\frac{1+\rho^2}{1-\rho^2}\right),\end{aligned} \tag{14}$$

where (a) follows from the summation of the first N terms of geometric series and the fact that $0 \leq \rho < 1$. Using the inequality (14), we have:

$$\Pr\left\{\max_{i\neq j}|\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}|>\beta\sigma_X^2\right\}<L(L-1)\exp\left(-\frac{\beta^2N(1-\rho^2)}{2(1+\rho^2)}\right). \tag{15}$$

By setting $\beta=\sqrt{\frac{6}{N}\left(\frac{1+\rho^2}{1-\rho^2}\right)\ln L}$, (6a) is obtained.

For the diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ we have:

$$\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}=\sigma_X^2+\sum_{\substack{r=1\\r\neq c}}^{N}\sum_{c=1}^{N}w_{ir}\mathbf{K}^{rc}_{\mathbf{xx}}w_{ic}. \tag{16}$$

Similarly to (12), all diagonal elements can be modeled as a sum of $N(N-1)$ independent random variables with the mean σ_X^2, i.e, $\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}=\sum_{\substack{r=1\\r\neq c}}^{N}\sum_{c=1}^{N}P^{ij}_{rc}$, where $P^{ij}_{rc}\in\{\sigma_X^2+\frac{\mathbf{K}^{rc}_{\mathbf{xx}}}{N},\sigma_X^2-\frac{\mathbf{K}^{rc}_{\mathbf{xx}}}{N}\}, 1\leq r\leq N, 1\leq c\leq N$, which are almost surely bounded, i.e., $\Pr\{P^{ij}_{rc}\in[\sigma_X^2+\frac{\mathbf{K}^{rc}_{\mathbf{xx}}}{N},\sigma_X^2-\frac{\mathbf{K}^{rc}_{\mathbf{xx}}}{N}]\}=1$. Then, the probability that the largest of all $|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}-\sigma_X^2|$ exceeds $\alpha\sigma_X^2$ satisfies:

$$\begin{aligned}&\Pr\Big\{\max_i|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}-\sigma_X^2|>\alpha\sigma_X^2\Big\}\overset{(a)}{\leq}L\Pr\{|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}-\sigma_X^2|>\alpha\sigma_X^2\}\\ &\overset{(b)}{\leq}2L\exp\left(-\frac{\alpha^2\sigma_X^4}{\frac{2}{N^2}\sum\limits_{r,c,r\neq c}(\mathbf{K}^{rc}_{\mathbf{xx}})^2}\right),\end{aligned} \tag{17}$$

where (a) follows from the fact that there are only L such identically distributed random variables [11] and (b) follows from Hoeffding's inequality [12]. In order to bound the term $\frac{1}{N^2}\sum\limits_{r,c,r\neq c}(\mathbf{K}^{rc}_{\mathbf{xx}})^2$, from (5), one obtains:

$$\begin{aligned}&\frac{1}{N^2}\sum_{\substack{r=1\\r\neq c}}\sum_{c=1}(\mathbf{K}^{rc}_{\mathbf{xx}})^2=\frac{2\sigma_X^4}{N^2}\left((N-1)\rho^2+\ldots+\rho^{2(N-1)}\right)\\ <&\frac{2\sigma_X^4}{N}\left(\frac{\rho^2-\rho^{2N}}{1-\rho^2}\right)\leq\frac{2\sigma_X^4\rho^2}{N(1-\rho^2)}.\end{aligned} \tag{18}$$

By using (18), (17) can be bounded as follows:

$$\Pr\left\{\max_i|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}-\sigma_X^2|>\alpha\sigma_X^2\right\}<2L\exp\left(-\frac{\alpha^2N(1-\rho^2)}{4\rho^2}\right). \tag{19}$$

By setting $\alpha=\sqrt{\frac{4}{N}\left(\frac{\rho}{1-\rho}\right)\ln L}$, (6b) is obtained. ■

Appendix B
Proof of Collary 1

Proof: This is a corollary of Proposition 1, where $\rho\to 0$. For the off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$, we can easily derive (7) by substituting $\rho=0$. For the diagonal elements, $\alpha|_{\rho=0}=0$. Thus, $\Pr\{\max_i|\mathbf{K}^{ii}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}-\sigma_Z^2|>0\}<\lim_{\rho\to 0}\frac{1}{L^{\left(\frac{1}{\rho}\right)}}=0$ for all $L>1$, which implies that $\forall i, 1\leq i\leq L, \mathbf{K}^{ii}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}=\sigma_Z^2$. ■

Appendix C
Proof of Proposition 2

Proof: Conditioned on $\mathcal{H}_1$, we define the event $E_{\mathcal{D}_j}$, as the event that there exists a subset of Hamming distances $\mathcal{D}_j\subset\mathcal{D}=\{D_1,\ldots,D_M\}$ with $|\mathcal{D}_j|=N_l$, including N_l of D_ms for each of them $m\neq 1$, $D_m\leq D_1$ and $D_m\leq\eta L$. P^{I}_m can be bounded as follows:

$$\begin{aligned}P^{\mathrm{I}}_m=&\Pr\left\{\bigcup_j E_{\mathcal{D}_j}\middle|\mathcal{H}_1\right\}\overset{(a)}{\leq}\sum_j\Pr\left\{E_{\mathcal{D}_j}\middle|\mathcal{H}_1\right\}\\ &=\sum_j\Pr\{(D_{j(1)}\leq D_1\cap D_1\leq\eta L)\cap\\ &\cdots\cap(D_{j(N_l)}\leq D_1\cap D_1\leq\eta L)|\mathcal{H}_1\}\\ &\overset{(b)}{=}\binom{M-1}{N_l}\Pr\{D_{m\neq 1}\leq D_1\cap D_1\leq\eta L|\mathcal{H}_1\}^{N_l}\\ &\overset{(c)}{\leq}(M-1)^{N_l}\Pr\{D_{m\neq 1}\leq\eta L|\mathcal{H}_1\}^{N_l}\end{aligned}$$

where $D_{j(i)}, j(i)\neq 1, 1\leq j\leq\binom{M-1}{N_l}, 1\leq i\leq N_l$ denotes the i^{th} element of the set $\mathcal{D}_j$, (a) follows from union bound, (b) from the fact that the events are independent and $D_{j(i)}$ are i.i.d. random variables and (c) follows from the inequality $\binom{M-1}{N_l}\leq(M-1)^{N_l}$ and the fact that

$$\begin{aligned}&\Pr\{D_{m\neq 1}\leq D_1\cap D_1\leq\eta L|\mathcal{H}_1\}\leq\Pr\{D_{m\neq 1}\leq D_1|D_1\leq\eta L,\mathcal{H}_1\}\\ \leq&\Pr\{D_{m\neq 1}\leq\eta L|\mathcal{H}_1\}.\end{aligned}$$

By using the Chernoff bound for any $P_b<\eta<\frac{1}{2}$, one obtains:

$$P^{\mathrm{I}}_m\leq\left\{M\exp\left[-L\mathcal{D}\left(\eta\|\tfrac{1}{2}\right)\right]\right\}^{N_l}\leq\{\exp[-L(\ln 2-H_2(\eta)-R)]\}^{N_l},$$

where $M=\exp\lfloor LR\rfloor$. By using the Chernoff bound, P^{II}_m can be bounded as:

$$P^{\mathrm{II}}_m\leq\exp[-\mathcal{D}(\eta\|P_b)].$$

By combining the bounds on P^{I}_m and P^{II}_m, P_m can be bounded by (10). ■

Appendix D
Proof of Proposition 3

Proof: Conditioned on $\mathcal{H}_0$, the probability of false acceptance can be bounded as follows:

$$\begin{aligned}P_f=&\Pr\left\{\bigcup_{m=1}^{M}D_m\leq\eta L\middle|\mathcal{H}_0\right\}\overset{(a)}{\leq}\sum_{m=1}^{M}\Pr\left\{D_m\leq\eta L\middle|\mathcal{H}_0\right\}\\ \overset{(b)}{=}&M\Pr\{D_m\leq\eta L|\mathcal{H}_0\}\leq M\exp\left[-\mathcal{D}(\eta\|\tfrac{1}{2})\right]\end{aligned} \tag{20}$$

where (a) follows from union bound and (b) holds because all D_m are i.i.d. random variables. By using Chernoff bound for any $P_b<\eta<\frac{1}{2}$, we have (11). ■

[1] If $X_1, X_2, \ldots, X_N$ are independent and $\Pr\{X_i\in[a_i,b_i]\}=1$, $(\forall i, 1\leq i\leq N)$, then for $t>0$, $\Pr\{|\bar{X}-E[\bar{X}]|\geq t\}\leq 2\exp\left(-\frac{2t^2N^2}{\sum_{i=1}^{N}(b_i-a_i)^2}\right)$, where $\bar{X}=\frac{X_1+X_2+\cdots+X_N}{N}$.