Performance Analysis of Content-Based Identification using Constrained List-Based Decoding

Farzad Farhadzadeh, *Student Member, IEEE,* Sviatoslav Voloshynovskiy, *Senior Member, IEEE,* and Oleksiy Koval, *Member, IEEE*

Abstract

This paper is dedicated to the performance analysis of content-based identification using binary fingerprints and constrained list-based decoding. We formulate content-based identification as a multiple hypothesis test and develop analytical models of its performance in terms of probabilities of correct detection/miss and false acceptance for a class of statistical models, which captures the correlation between elements of either the content or its extracted features. Furthermore, in order to determine the block/codeword length impact on the identification's accuracy, we analyse exponents of these probabilities of errors. Finally, we develop a probabilistic model, justifying the accuracy of identification based on list decoding by evaluating the position of the queried entry on the output list. The obtained results make it possible to characterize the performance of traditional unique decoding, based on the maximum likelihood for the situations when the decoder fails to produce the correct index. This paper also contains experimental results that confirm theoretical findings.

Index Terms

Content-based identification, digital fingerprint, constrained list-based decoding, order statistics, miss error exponent, false acceptance error exponent.

I. Introduction

In today's world, digital reproduction tools and user generated content (UGC) websites, such as Youtube, which enable massive distribution, sharing and storage of multimedia contents, have undergone an impressive evolution, providing professional solutions to various groups of users. Besides these obvious

Preliminary results from this work were presented in the IEEE Information Theory Symposium, 2010 [1] and the IEEE International Workshop on Multimedia and Signal Processing, 2010 [2].

F. Farhadzadeh, S. Voloshynovskiy and O. Koval are with the Department of Computer Science, University of Geneva, Geneva, Switzerland (Email: {Farzad.Farhadzadeh, svolos, Oleksiy.Koval}@unige.ch) and S. Voloshynovskiy is the corresponding author.

advantages, these tools offer, at the same time, unprecedented possibilities for counterfeiters to virtually reproduce any physical or digital items, i.e., images, videos, audiofiles, documents in electronic or printed form, fake biometrics or any luxury goods or art objects. Thus, the issue of integrity in content identification becomes a critical one demanding an urgent solution for various applications.

The content based identification (CBI) problem can be considered as a multiple hypothesis testing problem based on the Neyman–Pearson criterion [3], [4], [5], while the cost for making the wrong decision should be adjusted for each particular application. Since most CBI systems deal with critical and sensitive decisions in security applications, such as biometrics, content identification for copyright protection and illegal copy detection, etc., this cost is relatively high. No less important are the consequences of the wrong identification of physical objects such as over-terminated or fake medications, objects of art or luxury goods [6]. Therefore, under these conditions, the *identification problem* is defined as the multiple hypothesis test with $M+1$–alternatives, where M is the number of contents to be identified and the additional hypothesis stands for the erasure, if no match can be found. The performance of the CBI system is characterized by the probability of miss, i.e., when the genuine content is wrongly rejected, and the probability of false acceptance, when the faked or content-independent entry is falsely accepted as one of M genuine contents. In each considered application, both probabilities should be very small, which makes it similar to the classical digital communication setup.

On the other hand, the CBI systems are facing numerous additional requirements related to such issues as identification complexity, privacy, security as well as memory storage [5]. The trade-off between these requirements is a quite complex problem that still remains unsolved. To address this trade–off *digital fingerprints* are used [7], [8]. A digital fingerprint represents a short, robust and distinctive content description. The main idea behind digital fingerprinting consists in the extraction of a lower dimensional content representation that is usually accomplished as follows [6], [7], [8]. First a lower dimensional data representation from a content or its extracted feature is obtained (dimensionality reduction). Secondly, to address complexity, security, privacy and memory storage requirements, the transformed data are converted to a binary format. At the identification stage, either binary (hard decoding) or real valued query (soft decoding) can be used [5].

One key factor that restricts the progress in this direction is related to the analysis of the CBI system performance. This in turn requires to introduce tractable analytical models for CBI. Moreover, in many applications, data can be severely distorted and the classical unique decoding might not be capable of reasonably handling noisy inputs, thus resulting in a high rate of erroroneus decisions. However, it is known in digital communication that replacement of the unique decoding decision rule by the list

decoding with variable [9] or fixed [10] list size might help in such a situation. The reason for this enhancement is due to the fact that content degradation might change the order of the likelihood of the correct content. Since most of the identification techniques using unique decoding are based on the maximum likelihood (ML) principle, the change of the order of the correct likelihood will incur an error. However, this change might only cause the flip of the correct likelihood position to the nearest positions in its sorted list. Consequently, providing the list of most probable likelihoods of candidates might resolve the problem as soon as the correct candidate is on the list. Such a situation is mostly acceptable for the above-mentioned multimedia security, biometrics and physical object security applications, where the final decision is made by human means. Obviously, the change of decoding rule from 'unique' to 'list' decoding should be considered along with the relaxation of a constraint on the probability of false acceptance. Nevertheless, the potential help of list decoding in the CBI systems is little investigated and remains largely undiscovered with a few exceptions [1], [2], [11]. Therefore, an investigation of the impact of list decoding in the CBI applications is of great theoretical interest and practical importance. In this paper, we analyze the CBI for still images.

II. State-of-the-art

One of the first attempts to establish the theoretical limits of the CBI systems in biometrics applications was performed by Willems et al. [12]. The authors demonstrated that by using unique decoding under the assumption of an infinite length of sequences, one can attain the upper achievable rate given by the mutual information between outputs of the enrollment and identification channels in the class of discrete memoryless channels (DMC)s. This result was derived using the concept of typicality [13]. The false acceptance event was not considered in [12], due to the fact that the probability of two independent sequences being jointly typical is asymptotically vanishing. However, the obtained result can not be directly applied to the correlated contents as that would violate the principle of independence in the concept of typicality. To address this problem, as well as to relax the typicality constraint on the infinite length of sequences, Varna et al. [4] considered the CBI problem based on the ML criterion with the fidelity constraint for the images possessing local correlations and finite length fingerprint representations. The preservation of correlation in the binary data representation unavoidably leads to a decrease in entropy of fingerprints and thus to a decrease in identification rate as well as privacy leakage. Moreover, distortions should be also treated with special care due to their dependence upon original data. These factors considerably impact the accuracy of the conveyed analysis that is performed under certain assumptions. Independently, Voloshynovskiy et al. [5] and Willems [14] considered the CBI

for the independent and identically distributed (i.i.d.) binary data with finite length based on a bounded distance decoder (BDD) that can operate in erasure or list decoding modes similarly to [9]. However, the main focus of the above mentioned papers was on the analysis of unique decoding under privacy and complexity constraints for finite length sequences. Thus, the impact of real data statistics still remains uncovered.

Therefore, targeting an accurate performance analysis of the CBI systems, we will consider the performance of the CBI based on digital fingerprints taking into account the statistics of real images. Our analysis is accomplished in several stages. First, in order to guarantee the optimal discriminative power of binary fingerprints, one should maximize the entropy of the fingerprinting output that requires independence between fingerprint bits. Usually, such a property is satisfied by the proper selection of a linear mapper that is followed by binarization. Selection of such a mapper plays a crucial role at this stage due to the following argument: if the input to binarization procedure is a vector with uncorrelated components, the output is composed of pairwise independent bits [15]. Moreover, if this input has the jointly Gaussian distribution, the elements of the output are mutually independent. The mapper that possesses such properties is the Karhunen-Loève transform (KLT) [16] that optimally decorrelates its input for a given covariance matrix as well as optimally compacts its energy into a few components, making dimensionality reduction a straightforward process. 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. The latter issue gains importance due to the high computational complexity of this transform that can be evaluated as $\mathcal{O}(N^3)$, where N is the dimension of its input [17]. Additionally, the estimation of covariance matrices for large databases can be prohibitively expensive. Besides the drawbacks indicated above, the public disclosure of the basis vectors for a given class of data models makes this transform undesirable in the secure identification applications.

In order to ameliorate the issue of complexity, several approximations of the KLT were proposed. These include, for example, the Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) [16], which demonstrate a nearly optimal decorrelation of locally correlated data. The basis vectors of these transforms are fixed and independent of the statistics of their inputs. Due to their decorrelation and energy compaction capabilities, as well as the existence of fast implementation algorithms, they are used as a common tool in various signal and image processing applications. 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 [5].

One possible solution to this privacy/security shortcoming is a mapper that can be designed, based

on random projections (RP) [7]. The RP have been the object of much interest due to the fact that they are capable of providing an approximate distance preservation, something also recently recognized in the Compressed Sensing community for sparse data [18], [19]. While the decorrelation property of orthogonal transforms is well-known [16], the RP are based on approximately orthogonal bases. Therefore, the statistics of the projected data, i.e., the covariance matrix, are not well justified. On the other hand, prior knowledge of the statistics of the extracted digital fingerprints is crucial for the evaluation of the performance of the CBI systems. It is also interesting to explore the possibility of combining the DCT with the RP to benefit from both energy compaction and decorrelation, as well as security.

As mentioned above, the other important issue of the CBI systems is their ability to deal with highly distorted data. As a possible solution, one can envision the use of Forney's [9] list decoding approach as mentioned in [5]. However, in many identification applications, the final sink of information will be a human being. This constraint makes this type of list decoding undesirable, due to the high variability of the list size. Another solution, which is proposed by the authors in [1], is the *Constrained List-Based* (CLB) decoding approach. In the CLB decoding, which is a combination of Elias [10] and Forney's list decoding techniques in information transmission and coding applications, a limited number of candidates with the largest likelihood functions that can satisfy a specific threshold is selected. The analysis accomplished in [1] is based on the assumption that the contents are generated independently and identically. Thus, one of the main goals of this paper consists in the extension of this analysis to a broader class of statistical models with correlation. Moreover, one is often interested in choosing system parameters, i.e., the length of digital fingerprints, the decision threshold and the maximum number of candidates, to ensure that the probabilities of miss and false acceptance are below certain bounds. Hence, in this paper, besides computing the exact probabilities of correct detection and false acceptance, we derive bounds on the probabilities of miss, the complement of the probability of correct detection, and false acceptance for the digital fingerprints of a given length. Further yet, to show the impact of the list decoding, we investigate the probability that the correct entry of a database might fall in some position of the list, depending on the level of query degradation.

A. Contribution to the state-of-the-art

The main contribution of this paper can be summarized as follows: we analyze an identification setup based on binary i.i.d. fingerprints. In this identification setup, we exploit the CLB decoding in the binarized projected domain for either contents or their extracted features that can be modelled by a correlation-based model like a first order autoregressive (AR(1)) process, which captures correlation between elements of data [16]. Then, we investigate the fundamental performance limits in this setup by analysing probabilities

of errors and establishing the error exponent bounds as well as deriving achievable identification rates. Finally, we consider order statistics of the correct entry appearance on the list in order to justify the optimal list size for various operational modes. These results extend and deepen our preliminary findings [1] and [2] in regard to the analysis of the CBI based on the CLB as well as the previously considered contribution of [12].

To the best of our knowledge, the only work dealing with list decoding in the content fingerprinting applications is [11]. The closest relevant work addressing the theoretical analysis of correlated contents and binary fingerprints under the unique decoding is [4]. The principal differences with these papers can be summarised as follows:

- the CLB decoding proposed in this paper differs from the one analysed in [11] in two ways:[1]
  - *the type of list decoding*: the list decoder proposed in [11] produces the variable list size based on thresholding of likelihood functions computed for all items while the list decoder considered in this paper always outputs the list of candidates that does not exceed the predefined list size. The list decoding analysed in [11] represents better performance in terms of probability of miss in exchange for the unbounded list size that is not always desirable in those applications where the final sink is a human being;
  - *prior knowledge about channel statistics*: the decoder considered in [11] is based on some generic distance, which can be matched with the channel statistics, while the CLB considered in this paper is based on the Hamming distance deduced for the binary fingerprints.
- contrarily to [4], we consider a decorrelation approach based on the RP which makes it possible to generate binary fingerprints with asymptotically independent and equal likely distributed bits; this property could be of advantage for the maximization of the achievable rate of binary fingerprint identification, efficient fingerprint storage, privacy-preserving as well as extension of unique decoding to more general list decoding rules[2].

The main extension of the results earlier published by the authors [1], [2] consists in:

- in [1], [2] we have assumed that the contents to be identified can be modeled as an i.i.d. Gaussian process. Moreover, the impact of RP which are approximate ortho-projectors was not considered. In

[1]It should be pointed out that due to the different decoding strategies, i.e., constrained list size in the CLB case and variable list size in [11], the performance measure in terms of probability of miss is different and that makes a direct comparison unfeasible.

[2]In some applications, the extra correlation between fingerprint bits is favoured to strengthen the method with geometrical transforms or to avoid computational complexity of decorrelation in large-scale applications.

this manuscript, we extend this assumption from an i.i.d. process to an AR(1) Gaussian process and we analyze the impact of RP on the statistics of the projected data by deriving upper bounds;

- the performance analysis of identification systems proposed in [1] and [2] was based only on exact formulae of probabilities of miss and false acceptance, where the distortion channel was assumed to be a BSC. In this manuscript, we derived upper bounds on the probabilities of miss and false acceptance for a more general DMC distortion model;
- the numerical evaluations in [1] and [2] were based on synthetic data generated by an i.i.d. Gaussian process, however, in this context we extend our validation to simulations using a real image database, Uncompressed Colour Image Database (UCID) [20].

The outline of this paper is as follows. In Section III, we introduce notations and definitions exploited through this paper. Section IV defines the structure of the identification setup. In Section V, we consider the statistics of data used in the identification setup and demonstrate decorrelation and independence preserving properties of RP. Section VI elaborates the fundamental limits of the introduced identification setup. Finally, the conclusions are presented in Section VIII.

III. Notations, Definitions and Preliminaries

A. General notations

Throughout this paper, we adopt the convention that a scalar random variable is denoted by a capital letter X, a specific value it may take is denoted by the lower case letter x, and its alphabet is designated by the script letter $\mathcal{X}$. As for vectors, a boldface capital letter $\mathbf{X}^N$ with a corresponding superscript will denote an N-dimensional random vector $\mathbf{X}^N = \{X[i]\}_{i=1}^N$, a boldface lower case letter $\mathbf{x}^N$ will represent its particular realization $\mathbf{x}^N = \{x[i]\}_{i=1}^N$, and the respective superalphabet is the N^{th} Cartesian power of $\mathcal{X}$, i.e., $\mathcal{X}^N$. $\mathbf{x}^{N^\dagger}$ stands for the transpose of $\mathbf{x}^N$. The expectation operator is designated by $\mathbb{E}[\cdot]$. We use $H_2(\cdot)$ to denote the entropy of a binary random variable. $\mathcal{N}(\mu, \sigma_X^2)$ stands for the Gaussian distribution with mean μ and variance σ_X^2. $\mathrm{Q}(\cdot)$ designates the Q-function. $\mathcal{B}(N, p)$ denotes the Binomial distribution with N trails and probability of success p. The Kullback-Leibler divergence between two distributions $p(x)$ and $q(x)$ on $\mathcal{X}$ is defined as, $\mathcal{D}(p(x)\|q(x)) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)}$, with the conventions that $0 \ln 0 = 0$, and $p \ln \frac{p}{0} = \infty$ if $p > 0$.

B. Order statistics

Let $V_1, V_2, \ldots, V_M$ be M i.i.d. random variables, each with a cumulative distribution function (CDF) $F(v)$. The r-th order statistic of these M i.i.d. random variables is denoted by $V_{(r:M)}$, which corresponds

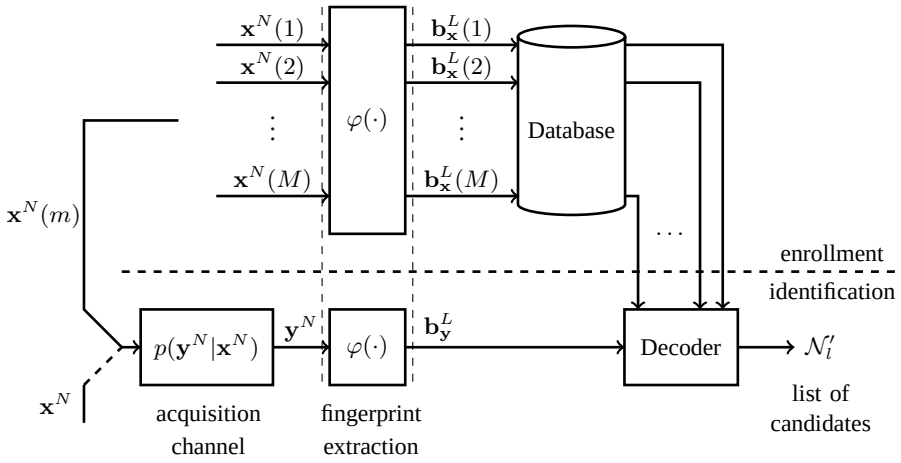


Fig. 1. The identification setup for CBI based on binary fingerprints.

to the r-th position of $v_{(1:M)} \leq v_{(2:M)} \leq \ldots \leq v_{(r:M)} \leq \ldots \leq v_{(M:M)}$ for a specific outcome. $F_{(r:M)}(v)$, the CDF of $V_{(r:M)}$, is given by [21]:

$$\begin{aligned} F_{(r:M)}(v) &= \Pr\left\{V_{(r:M)} \leq v\right\} = \Pr\left\{\text{at least } r \text{ of } V_i \text{ are less than or equal to } v\right\} \\ &= \sum_{i=r}^{M} \binom{M}{i} F^i(v)[1-F(v)]^{M-i}, \end{aligned} \tag{1}$$

since the term in the summand is the binomial probability that *exactly* i of $V_1, V_2, \ldots, V_M$ are less than or equal to v.

C. Random Projections

In RP, the original N-dimensional data are projected to an L-dimensional ($L \leq N$) subspace, by a linear mapper $\mathbf{w}^{L\times N}$ drawn from a specified probability distribution. We further use $\mathbf{w}$ rather than $\mathbf{w}^{L\times N}$ for convenience. The key idea behind the dimensionality reduction using RP is based on the Johnson-Lindenstrauss lemma [18]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. The choice of the random matrix $\mathbf{W}$ is very important for satisfying the conditions of this lemma. The elements W_{ij} of $\mathbf{W}$ are often Gaussian distributed, but Achlioptas [22] has shown that the Gaussian distribution can be replaced by a much simpler Bernoulli distribution $\Pr\{W_{ij} = \pm\frac{1}{\sqrt{N}}\} = \frac{1}{2}$. We also consider the RP based on the above Bernoulli distribution due to the simplicity of statistical analysis of projected data.

IV. The Identification Setup

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

In the content enrollment phase, the digital fingerprints are extracted from either contents or their extracted features and stored in a *Database*. The Database is a collection of M labelled binary vectors denoted by:

$$\mathbf{b}_{\mathbf{x}}^{L}(m) \in \{0,1\}^{L}, m \in \{1,\ldots,M\},$$

where $\mathbf{b}_{\mathbf{x}}^{L}(m) = \varphi(\mathbf{x}^{N}(m))$ is a digital fingerprint extracted from either the content or its extracted feature $\mathbf{x}^{N}(m) \in \mathcal{X}^{N}$, which is drawn from a common stationary distribution $p(\mathbf{x}^{N})$. $\varphi(\cdot)$ is a digital fingerprint extraction function that can be key-dependent. Conversion to binary in the fingerprint extraction is applied so as to cope with storage, privacy, security and complexity constraints. However, since the use of the secret key does not impact statistical analysis of the setup due to its symmetrical presence at enrollment and identification stages, we consider only a key-independent digital fingerprint generation in this paper.

In the content identification phase, for a given query $\mathbf{y}^{N}$ the digital fingerprint is extracted similar to the enrollment phase, i.e., $\mathbf{b}_{\mathbf{y}}^{L} = \varphi(\mathbf{y}^{N})$. Then, the decoder decides whether the query is relevant to some entries of the Database, and if so, decides to which ones. Otherwise, it produces an erasure.

A. Identification Problem as a Decoding Problem

In the case the query digital fingerprint $\mathbf{b}_{\mathbf{y}}^{L}$ is related to some element $\mathbf{b}_{\mathbf{x}}^{L}(m)$ of the Database, this relationship can be modeled as a binary channel with the transition probability $p(\mathbf{b}_{\mathbf{y}}^{L}|\mathbf{b}_{\mathbf{x}}^{L}(m))$. If the query digital fingerprint $\mathbf{b}_{\mathbf{y}}^{L}$ is unrelated to any entry of the Database, we assume that $\mathbf{b}_{\mathbf{y}}^{L}$ is drawn from $p(\mathbf{b}_{\mathbf{y}}^{L}) = \sum_{\mathbf{b}_{\mathbf{x}}^{L} \in \{0,1\}^{L}} p(\mathbf{b}_{\mathbf{x}}^{L}) p(\mathbf{b}_{\mathbf{y}}^{L}|\mathbf{b}_{\mathbf{x}}^{L})$. Therefore, we can define the content identification problem as a statistical test with $M+1$ hypotheses:

$$\begin{cases} \mathcal{H}_0 \ : \mathbf{B}_{\mathbf{y}}^{L} \sim p(\mathbf{b}_{\mathbf{y}}^{L}) \\ \mathcal{H}_m : \mathbf{B}_{\mathbf{y}}^{L} \sim p(\mathbf{b}_{\mathbf{y}}^{L}|\mathbf{b}_{\mathbf{x}}^{L}(m)), \end{cases} \tag{2}$$

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

B. Constrained List Based Decoder

We define the CLB decoding as follows:

1) For each entry $\mathbf{b}_{\mathbf{x}}^{L}(m), 1 \leq m \leq M$, of the Database, the decoder computes log-normalized-likelihoods $\mathcal{L}_m = \ln \frac{p(\mathbf{b}_{\mathbf{y}}^{L}|\mathbf{b}_{\mathbf{x}}^{L}(m))}{p(\mathbf{b}_{\mathbf{y}}^{L})}$.
2) The computed log-normalized-likelihoods are sorted in ascending order.
3) The N_l indices with the largest log-normalized-likelihood functions are chosen. Then, their indices are put in the primary list $\mathcal{N}_l$ one-by-one, i.e., the first index in $\mathcal{N}_l$ corresponds to the largest one and so forth. Parameter N_l is referred to as the primary list size.

4) The final list of candidates is defined as:

$$\mathcal{N}_l' = \{m \in \mathcal{N}_l : \mathcal{L}_m \geq \gamma L\}, \tag{3}$$

where γ controls the number of final candidates and defines the rejection option.

The performance metrics of the CBI are defined by the probability of correct detection, P_d:

$$P_d = 1 - P_m = \sum_{m=1}^{M} \Pr\{(m \in \mathcal{N}_l) \cap (\mathcal{L}_m \geq \gamma L) | \mathcal{H}_m\} \Pr\{\mathcal{H}_m\}, \tag{4}$$

where P_m denotes the probability of miss, and the probability of false acceptance:

$$P_f = \Pr\{\mathcal{N}_l' \neq \varnothing | \mathcal{H}_0\}. \tag{5}$$

V. The Statistical Analysis of Digital Fingerprint Extraction

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

1) The dimensionality of a content or its extracted feature $\mathbf{x}^N(m)$ and a query $\mathbf{y}^N$ is reduced from N to L by applying the RP operator, $\mathbf{w}^{L \times N}$ [6]. Note that RP 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 the probability mass function (PMF) $\Pr\{W_{ij} = \pm\frac{1}{\sqrt{N}}\} = \frac{1}{2}$, $1 \leq i \leq L$ and $1 \leq j \leq N$. For a given $\mathbf{w}$, the projections $\tilde{\mathbf{x}}^L(m)$ and $\tilde{\mathbf{y}}^L$ are obtained by $\tilde{\mathbf{x}}^L(m) = \mathbf{w}\mathbf{x}^N(m)$ and $\tilde{\mathbf{y}}^L = \mathbf{w}\mathbf{y}^N$.
2) L-length binary digital fingerprints, $\mathbf{b}_{\mathbf{y}}^L$ and $\mathbf{b}_{\mathbf{x}}^L(m)$, are derived by taking the sign of the projected data, i.e., $\mathbf{b}_{\mathbf{x}}^L(m) = \{\text{sign}(\tilde{x}[i](m))\}_{i=1}^L$ and $\mathbf{b}_{\mathbf{y}}^L = \{\text{sign}(\tilde{y}[i])\}_{i=1}^L$, $\text{sign}(a) = 1$, if $a > 0$, and 0, otherwise.

A. The Statistics of Digital Fingerprints Extracted from Correlated Data

In this Section, we investigate the statistics of digital fingerprints obtained by the RP. We assume the input $\mathbf{X}^N$ can be modelled as an AR(1) Gaussian process. The justification of the use of this model is two-fold. First, $\mathbf{X}^N$ can be considered as an image that is characterized by local correlations between neighbouring pixels. To capture these correlations, a number of statistical models such as autoregressive and Markov random field are proposed [16]. The AR(1) Gaussian process is considered as one comprised of simple yet powerful models that accurately represent the local correlations present in images $\mathbf{X}^N$ [16]. Secondly, in the case $\mathbf{X}^N$ represents some robust features extracted from an original content to cope with potential malicious attacks, such an assumption that might yet be valid. For example, SIFT [23], SURF [24] or Fourier-Mellin [25] transform used for image description includes a certain level of correlation among samples of extracted features that can be modelled as an AR(1) model. Finally, many fingerprinting algorithms operates in decorrelation domains such as DCT or DWT [26], where the residual

correlation among components of transformation coefficients can be modelled as AR(1) [27] . Assuming this model, the covariance matrix in the projected domain is given by:

$$\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathbb{E}[\mathbf{w}\mathbf{X}^N\mathbf{X}^{N^\dagger}\mathbf{w}^\dagger] = \mathbf{w}\mathbf{K}_{\mathbf{xx}}\mathbf{w}^\dagger, \tag{6}$$

where $\mathbf{K}_{\mathbf{xx}}$ is defined by [16]:

$$\mathbf{K}_{\mathbf{xx}} = \sigma_X^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & 1 & \dots & \rho^{N-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \dots & 1 \end{bmatrix}, \tag{7}$$

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

Proposition 1. Let the elements of the RP matrix, $\mathbf{w}$ of size $L \times N$ and $1 < L \leq N$, be drawn from PMF $\Pr\{W_{ij} = \pm\frac{1}{\sqrt{N}}\} = \frac{1}{2}$, and $\mathbf{X}^N$ be a real zero-mean random vector modelled as the AR(1) Gaussian process with variance σ_X^2 and normalized correlation coefficient ρ. Then, we have:

$$\Pr\left\{\max_{i\neq j}|\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}| > \beta\sigma_X^2\right\} < \frac{1}{L}, \tag{8a}$$

$$\Pr\left\{\max_{i}|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \sigma_X^2| > \alpha\sigma_X^2\right\} < \frac{2}{L^{\left(\frac{1}{\rho}\right)}}, \tag{8b}$$

where $\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ denotes the $(i,j)^{\text{th}}$ element of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$, $\beta = \left(\frac{1-\rho^N}{1-\rho}\right)\sqrt{\frac{12}{N}\ln L}$, and $\alpha = \left(\frac{1-\rho^{N-1}}{1-\rho}\right)\sqrt{\frac{8}{N}\rho\ln L}$.

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 source is the AR(1) Gaussian process, which implies that $\mathbf{X}^N$ is jointly Gaussian, and RP is a linear transform, the projected data $\tilde{\mathbf{x}}^L$ follow the jointly Gaussian distribution, i.e., $\tilde{\mathbf{X}}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}})$. Therefore, since elements of $\tilde{\mathbf{X}}^L$ are asymptotically uncorrelated, $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \approx \sigma_X^2\mathbf{I}_L$, one can conclude that $\tilde{\mathbf{X}}^L$ are asymptotically i.i.d. In addition, the elements of the digital fingerprint extracted from $\tilde{\mathbf{X}}^L$ asymptotically consist of L i.i.d. $\text{Bernoulli}(\frac{1}{2})$ bits due to symmetry of the Gaussian distribution function.

Remark 2. In this case $\mathbf{w}$ is chosen to consist of the eigenvectors of $\mathbf{K}_{\mathbf{xx}}$, i.e., in the KLT, one will obtain the decorrelated $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ with the ordered main diagonal elements $\mathbf{K}^{11}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \geq \mathbf{K}^{22}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \geq \ldots \geq \mathbf{K}^{NN}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ [16]. The difference with the RP consists in perfect decorrelation versus the asymptotic one in (8a) and the power–law decaying character of main diagonal elements versus approximately uniform ones in (8b) [16].

B. The Statistics of Digital Fingerprint Extracted from Query

The query $\mathbf{y}^N$ under the true hypothesis represents the distorted version of $\mathbf{x}^N$ that might undergo various distortions ranging from the simple addition of signal independent noise to signal dependent distortions such as lossy compression or even geometrical transforms. The statistical modelling of these distortions is quite a challenging task by itself. However to withstand geometrical transforms, many fingerprinting algorithms use robust features. In this case, the matching of geometrically distorted images is based on the matching of robust features based on Euclidean distance [28] where the effect of geometrical distortions is converted into the independent additive noise [29]. In the general case, one can assume that the resulting noise comes from the broad family of i.i.d. Generalized Gaussian distributions (GGD) with the shape parameter less than or equal to 2. The distribution parameters will impact the statistics of query fingerprint and its mismatch with the fingerprint of the original content. As will be shown below, the Gaussian distribution, that is a particular case of the GGD with the shape parameter equals to 2, produces the largest mismatch in terms of Hamming distance due to the highest cross-over probability among all GGDs with the bounded variance. Therefore, our analysis will be concentrated on the consideration of the worst case crossover probability provided by the Gaussian noise. Consider the query $\mathbf{y}^N$ to be a noisy version of a piece of content or its extracted feature that can be modeled as the AR(1) Gaussian process and is observed through an Additive White Gaussian Noise (AWGN) channel, $\mathbf{Y}^N = \mathbf{X}^N + \mathbf{Z}^N$, where $\mathbf{Z}^N \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$ and σ_Z^2 is the variance of the noise. At the output of the first step of digital fingerprinting, we have $\tilde{\mathbf{Y}}^L = \tilde{\mathbf{X}}^L + \tilde{\mathbf{Z}}^L$. From Proposition 1, we can assume that $\tilde{\mathbf{X}}^L$, asymptotically follows the distribution $\mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_L)$. To justify the statistics of $\tilde{\mathbf{Z}}^L$, we have the following corollary of Proposition 1.

Corollary 1. Let the elements of the RP matrix, $\mathbf{w}$, be generated as in Proposition 1, and $\mathbf{Z}^N$ is drawn i.i.d. from a common stationary distribution with variance σ_Z^2. Then, the diagonal elements of the covariance matrix of the projected noise $\tilde{\mathbf{Z}}^L = \mathbf{w}\mathbf{Z}^N$ are equal to σ_Z^2, i.e., $\forall i, \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}^{ii} = \sigma_Z^2$, and all off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$ satisfy:

$$\Pr\left\{\max_{i \neq j} |\mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}^{ij}| > \delta \sigma_Z^2\right\} < \frac{1}{L}, \tag{9}$$

where $\delta = \sqrt{\frac{12}{N} \ln L}$.

Proof: Appendix B. ■

Remark 3. 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. Since $\mathbf{Z}^N$ is i.i.d. Gaussian and RP is a linear transform, $\tilde{\mathbf{Z}}^L$ is jointly Gaussian whose elements are asymptotically uncorrelated, i.e., $\tilde{\mathbf{Z}}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}), \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} \approx \sigma_Z^2 \mathbf{I}_L$, thus $\tilde{\mathbf{Z}}^L$ asymptotically

follows i.i.d. Gaussian distribution. Consequently, the transformed channel is asymptotically a discrete memoryless channel, i.e., $p(\mathbf{b}_{\mathbf{y}}^L|\mathbf{b}_{\mathbf{x}}^L) = \prod_{i=1}^{L} p(b_y[i]|b_x[i])$.

Remark 4. The obtained results are also applicable to the noise modeled as the AR(1) Gaussian process.

Remark 5. From Proposition 1 and Corollary 1, $\tilde{\mathbf{Y}}^L$ is a sum of two independent random vectors $\tilde{\mathbf{X}}^L$ and $\tilde{\mathbf{Z}}^L$, where $\tilde{\mathbf{X}}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}), \mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} \approx \sigma_X^2 \mathbf{I}_L$ and $\tilde{\mathbf{Z}}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}), \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}} \approx \sigma_Z^2 \mathbf{I}_L$. Therefore, $\tilde{\mathbf{Y}}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}})$, where $\mathbf{K}_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} \approx (\sigma_X^2 + \sigma_Z^2)\mathbf{I}_L$, is a jointly Gaussian distributed random vector with asymptotically uncorrelated elements, which implies their convergent independence. Moreover, one can conclude that $\mathbf{B}_{\mathbf{y}}^L$ consists of L bits that are asymptotically i.i.d. $\mathrm{Bernoulli}(\frac{1}{2})$ due to symmetry of the zero mean Gaussian distribution. Conditioned on $\mathcal{H}_m$, the relation between $\mathbf{b}_{\mathbf{x}}^L(m)$ and $\mathbf{b}_{\mathbf{y}}^L$ can be modeled by the Binary Symmetric Channel (BSC) with crossover probability [5]:

$$P_b = \frac{1}{\pi} \arccos\left(\sqrt{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}}\right). \tag{10}$$

To demonstrate that the i.i.d. Gaussian noise indeed represents the worst case in terms of resultant P_b, we consider the projected data $\tilde{\mathbf{X}}^L \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_L)$ corrupted by additive noise that follows i.i.d. GGD, resulting in observation $\tilde{\mathbf{Y}}^L = \tilde{\mathbf{X}}^L + \tilde{\mathbf{Z}}^L$.

The crossover probability in this case is defined as:

$$\begin{aligned} P_b &= \Pr[\mathrm{sign}(\tilde{X}) \neq \mathrm{sign}(\tilde{Y})] = \Pr[\tilde{Y} \geq 0 | \tilde{X} \leq 0]\Pr[\tilde{X} \leq 0] + \Pr[\tilde{Y} \leq 0 | \tilde{X} \geq 0]\Pr[\tilde{X} \geq 0] \\ &\overset{(a)}{=} \Pr[\tilde{Z} \geq \tilde{x} | \tilde{X} \geq 0] = \int_0^\infty \int_{\tilde{x}}^\infty \frac{\theta}{2\omega\Gamma(1/\theta)} \exp\left(-\frac{|t-\mu|^\theta}{\omega}\right) \frac{2}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{x^2}{2\sigma_X^2}\right) \mathrm{d}t\mathrm{d}\tilde{x} \end{aligned} \tag{11}$$

where $\mu = 0$, θ is the shape parameter, $\omega^2 = \sigma_Z^2 \frac{\Gamma(1/\theta)}{\Gamma(3/\theta)}$, $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t$ is the Gamma function, and (a) follows from the fact that $\tilde{X}$ and $\tilde{Z}$ have symmetric distributions. Besides the Gaussian mentioned above, and Laplacian, which is a particular case of GGD with $\theta = 1$, where the P_b can be computed analytically:

$$P_b = \exp\left(\frac{\sigma_X^2}{\sigma_Z^2}\right) \mathrm{Q}\left(\sqrt{2\frac{\sigma_X^2}{\sigma_Z^2}}\right), \tag{12}$$

it is difficult to find closed form expressions for P_b for all other values of $\theta \leq 2$. For the comparison reasons, we numerically compute P_b for several values of the shape parameters, shown in Fig. 2. As expected, the Gaussian PDF ($\theta = 2$) is characterized by the highest crossover probability and it will be used in all future considerations.

Under these conditions, the corresponding hypotheses (2) become:

$$\begin{cases} \mathcal{H}_0 \ : \mathbf{B}_{\mathbf{y}}^L \sim \frac{1}{2^L}, \\ \mathcal{H}_m : \mathbf{B}_{\mathbf{y}}^L \sim P_b^{d_m}(1-P_b)^{L-d_m}. \end{cases} \tag{13}$$

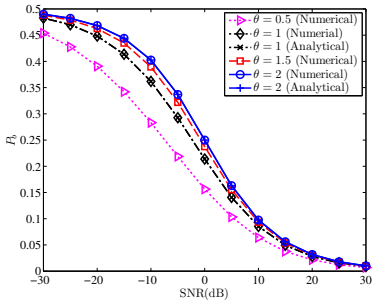


Fig. 2. The crossover probability computed for different GGD shape parameters.

where $d_m \triangleq d_H(\mathbf{b}_{\mathbf{y}}^L, \mathbf{b}_{\mathbf{x}}^L(m))$ is the Hamming distance between $\mathbf{b}_{\mathbf{y}}^L$ and $\mathbf{b}_{\mathbf{x}}^L(m)$ that was also considered in [30].

VI. Theoretical Performance Analysis of The Identification Setup

In this section, we analyze the performance of the identification setup based on P_d and P_f defined in Section IV.

A. Probability of Correct Detection

In this subsection, we evaluate the probability of correct detection P_d. From Remark 5, the log-normalized likelihood $\mathcal{L}_m$ conditioned on $\mathcal{H}_m$ is given by:

$$\mathcal{L}_m = \ln \frac{P_b^{d_m}(1-P_b)^{L-d_m}}{\left(\frac{1}{2}\right)^L}. \tag{14}$$

The log-normalized-likelihood is a decreasing function of d_m for $P_b \in [0, 0.5)$, where the Hamming distance d_m is a realization of the random variable D_m, which can be considered as a sufficient statistic in the analyzed setup to evaluate the probability of correct detection. According to Remark 1 all entries of the Database are considered to be i.i.d., moreover, we assume that all entries can be queried equally likely, i.e., $\Pr\{\mathcal{H}_m\} = \frac{1}{M}$. Therefore, the overall probability of correct detection does not depend on the particular index m. And, the analysis is accomplished only for the first index m_1:

$$\begin{aligned} P_d &= \Pr\{(m_1 \in \mathcal{N}_l) \cap (\mathcal{L}_1 \geq \gamma L)|\mathcal{H}_1\} \overset{(a)}{=} \Pr\{(\mathcal{L}_{(M-N_l:M-1)} < \mathcal{L}_1) \cap (\mathcal{L}_1 \geq \gamma L)|\mathcal{H}_1\} \\ &\overset{(b)}{=} \Pr\{(D_{(N_l:M-1)} > D_1) \cap (D_1 \leq \eta L)|\mathcal{H}_1\} = \sum_{d=0}^{\eta L} \Pr\{D_{(N_l:M-1)} > d|\mathcal{H}_1, D_1 = d\} p_{D_1}(d), \end{aligned} \tag{15}$$

where $p_{D_1}(d)$ denotes the PMF of D_1, $\eta = \frac{\gamma - \ln 2 - \ln(1-P_b)}{\ln(P_b/(1-P_b))}$ is obtained from the replacement of the condition $\mathcal{L}_m \geq \gamma L$ by $D_1 \leq \eta L$, (a) follows from the fact that the first event $m_1 \in \mathcal{N}_l$ occurs if and only if $\mathcal{L}_1$ is among the N_l largest of $\{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m\}$, which is equivalent to $(\mathcal{L}_{(M-N_l:M-1)} < \mathcal{L}_1)$, (b) holds because the log-normalized likelihood is a decreasing function of the Hamming distance.

By using Remarks 1 and 5, conditioned on $\mathcal{H}_1$, the sufficient statistics mentioned above have the following distributions for $1 \leq m \leq M$:

$$D_m \sim \begin{cases} \mathcal{B}(L, P_b), \text{for } m = 1, \\ \mathcal{B}(L, \frac{1}{2}), \text{ for } m \neq 1. \end{cases} \tag{16}$$

Consequently, by substituting (1) and (16) into (15), the correct detection probability over the BSC is given by:

$$P_d = \sum_{d=0}^{\eta L} \binom{L}{d} P_b^d (1-P_b)^{L-d} \left\{ \sum_{p=0}^{N_l-1} \binom{M-1}{p} \left[\left(\tfrac{1}{2}\right)^L \sum_{x=0}^{d} \binom{L}{x} \right]^p \left[\left(\tfrac{1}{2}\right)^L \sum_{x=d+1}^{L} \binom{L}{x} \right]^{(M-1)-p} \right\}. \tag{17}$$

B. Probability of False Acceptance

The main reason to consider the probability of false acceptance is to investigate the reliability of identification with respect to the acceptance of queries, which are unrelated to entries of the Database. To evaluate P_f, we define the following events:

$$E_{D_{(i:M)}} = \{D_{(i:M)} \leq \eta L | \mathcal{H}_0\}, \tag{18}$$

where $1 \leq i \leq N_l$ and $E_{D_{(i:M)}}$ happens if the i^{th} of M ascendingly ranked i.i.d. Hamming distances between the query and entries of the Database is smaller than the threshold. Moreover, from Remarks 1 and 5, conditioned on $\mathcal{H}_0$, $D_m \sim \mathcal{B}\left(L, \frac{1}{2}\right)$ for $1 \leq m \leq M$. The probability of false acceptance is:

$$P_f = \Pr\{\cup_{i=1}^{N_l} E_{D_{(i:M)}} | \mathcal{H}_0\} = 1 - \Pr\{\cap_{i=1}^{N_l} E^c_{D_{(i:M)}} | \mathcal{H}_0\} \overset{(a)}{=} 1 - \Pr\{E^c_{D_{(1:M)}} | \mathcal{H}_0\} = \Pr\{E_{D_{(1:M)}} | \mathcal{H}_0\} \tag{19}$$

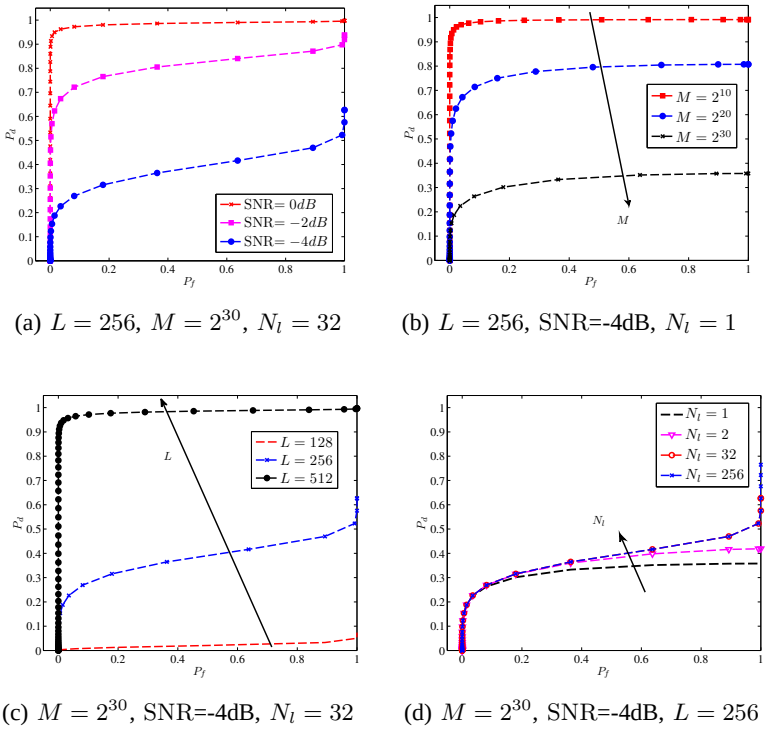
where $E^c_{D_{(i:M)}}$ is the complement of $E_{D_{(i:M)}}$, and (a) follows from the fact that if the event $E^c_{D_{(1:M)}}$ occurs the remaining events are certain. Then, the probability of false acceptance is given by:

$$P_f = \Pr\left\{ \min_{1 \leq m \leq M} D_m \leq \eta L | \mathcal{H}_0 \right\} = 1 - \left[1 - \left(\frac{1}{2}\right)^L \sum_{x=0}^{\eta L} \binom{L}{x} \right]^M. \tag{20}$$

Remark 6. The probability of false acceptance is independent of the primary list size and the channel crossover probability.

Remark 7. For the case $M = 1$ and $N_l = 1$, P_d and P_f coincide with the probabilities of detection and false alarm under the authentication setup using binary fingerprints defined in [14], [30].

In the following, we will present the results for large-scale identification applications to investigate the impact of database cardinality M, fingerprint length L and primary list size N_l on the performance of the CBI system. In Fig. 3, we demonstrate the receiver operational characteristic (ROC) computed using (17) and (20) for various values of the parameters L, M and N_l. Fig. 3a shows the ROC curves for the



(a) $L = 256$, $M = 2^{30}$, $N_l = 32$ (b) $L = 256$, SNR=-4dB, $N_l = 1$

(c) $M = 2^{30}$, SNR=-4dB, $N_l = 32$ (d) $M = 2^{30}$, SNR=-4dB, $L = 256$

Fig. 3. The CLB decoding performance analysis for binary fingerprints.

range of SNRs between -4dB and 0dB, where $\mathrm{SNR} = 10\log_{10}\frac{\sigma_X^2}{\sigma_Z^2}$, and primary list size $N_l = 32$ for $M = 2^{30}$ digital fingerprints of length 256 bits in the Database. We observe that as the SNR decreases, the probability of correct detection P_d reduces for a given probability of false acceptance P_f. Fig. 3b examines the influence of the number of fingerprints M in the Database on the decoder performance for a fixed fingerprint length $L = 256$ bits and the SNR=-4dB. As M increases, the probability of false acceptance increases and the probability of correct detection decreases. Consequently, for a given P_d, the P_f is higher, or equivalently, for a fixed P_f , the probability of correct detection is lower. However, this improvement occurs in the region with high P_f. Fig. 3c demonstrates the decoder performance enhancement using longer digital fingerprints for a given distortion level, allowing for the optimization of the CBI system design with respect to this parameter. Finally, Fig. 3d demonstrates the impact of list size on the ROC. Although we can increase the performance in term of P_d by increasing the list size N_l, this improvement is restricted by a certain range of primary list sizes. This confirms that CLB decoding can enhance the correct decoding at the cost of relaxing the constraint on the probability of false acceptance introduced via the increased list size.

C. Miss Error Exponent

In this section, we derive bounds on the probability of miss, which is complementary to P_d, for the DMC. Conditioned on $\mathcal{H}_1$, the probability of miss of the identification system based on the CLB decoding is given by:

$$P_m = \Pr\{m_1 \notin N_l \cup \mathcal{L}_1 < \gamma L | \mathcal{H}_1\} = \Pr\{m_1 \notin N_l \cap \mathcal{L}_1 \geq \gamma L | \mathcal{H}_1\} + \Pr\{\mathcal{L}_1 < \gamma L | \mathcal{H}_1\} = P_m^I + P_m^{II}, \tag{21}$$

The first term in (21) is referred to as the *probability of miss of the first kind*, P_m^{I}, and the second term is the *probability of miss of the second kind*, P_m^{II}.

Remark 8. Under variable list size decoding approach [11], where there is no restriction over the list of candidates, P_m^{I} equals to 0.

Proposition 2. Consider a DMC with a transition probability $p(y|x)$, and a Database of block length L with $M = e^{LR}$ entries independently and identically generated according to $p(x)$, and let $p(y) = \sum_{x\in\mathcal{X}} p(x)p(y|x)$. Suppose that the query $\mathbf{y}^N$ acts as an input to the content-based identification system and the CLB decoding is applied. The average CLB probability of miss is bounded by:

$$P_m \leq \exp[-N_l L(E(\gamma) + \gamma - R)] + \exp[-LE(\gamma)] \tag{22a}$$

$$\leq 2\exp[-L\min\{N_l(E(\gamma) + \gamma - R), E(\gamma)\}], \tag{22b}$$

where

$$E(\gamma) = \max_{0\leq s\leq 1} -\ln \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x)p(y)^s p(y|x)^{1-s} e^{s\gamma} \tag{23}$$

and γ is a fixed threshold.

Proof: Appendix C. ■

The miss error exponent is referred to as $E_m = \min\{N_l(E(\gamma) + \gamma - R), E(\gamma)\}$.

Remark 9. For, $N_l = 1$, (22a) coincides with the error probability given in [31] for communications.

Corollary 2. For any $\gamma < C$, where $C = \mathcal{D}(p(x,y)\|p(x)p(y))$ is the DMC capacity, $E(\gamma) > 0$ and $E(\gamma) \to 0$ as $\gamma \to C$.

Proof: Appendix D. ■

Corollary 3. For the binary symmetric channel with a crossover probability P_b, the list-decoding-based average probability of miss is bounded, for any η, $P_b < \eta < \frac{1}{2}$:

$$P_m \leq \{\exp[-L(\ln 2 - R - H_2(\eta))]\}^{N_l} + \exp[-L\mathcal{D}(\eta\|P_b)] \tag{24a}$$

$$\leq 2\exp[-L\min\{N_l(\ln 2 - R - H_2(\eta)), \mathcal{D}(\eta\|P_b)\}]. \tag{24b}$$

Proof: Appendix E. ■

The corresponding miss error exponent over the BSC is given by $E_m = \min\{N_l(\ln 2 - R - H_2(\eta)), \mathcal{D}(\eta\|P_b)\}$.

Remark 10. For the case $N_l = 1$, i.e., unique decoding mode, the obtained miss probability bound coincides with the result provided in [5], [14]. If $N_l > 1$, i.e., list decoding mode, P_m^{I} converges to 0 up to N_l times exponentially faster than for unique decoding.

Remark 11. For $P_b < \eta < \frac{1}{2}$ and $R < \ln 2 - H_2(\eta)$, there exist fingerprints with rate R and miss probability P_m such that $\lim_{L\to\infty} P_m = 0$.

D. False Acceptance Error Exponent

In this section, we derive the upper bound on P_f for the DMC.

Proposition 3. Consider a DMC with a transition probability $p(y|x)$, and a database of block length L with $M = e^{LR}$ entries independently and identically generated according to $p(x)$, and let $p(y) = \sum_{x\in\mathcal{X}} p(x)p(y|x)$. Suppose that the query $\mathbf{y}^N$ acts as an input to the content-based identification system and the CLB decoding is used. The average list-decoding-based probability of false acceptance is bounded by:

$$P_f \leq \exp[-L(E(\gamma) + \gamma - R)]. \tag{25}$$

where $E(\gamma)$ is defined in (23).

Proof: Appendix F. ■

The false acceptance error exponent is referred to as $E_f = E(\gamma) + \gamma - R$.

Remark 12. If we set $\gamma = R$ and $N_l = 1$, then $P_m \leq 2\exp[-LE(R)]$ and $P_f \leq \exp[-LE(R)]$. Since $E(R) > 0$ for $R < C$, $P_m \to 0$ and $P_f \to 0$ as $L \to \infty$. Moreover, this holds for R close to C, then one can conclude that the identification capacity is achievable.

Corollary 4. For the BSC with crossover probability P_b, the CLB average probability of false acceptance is bounded, for any η, $P_b < \eta < \frac{1}{2}$:

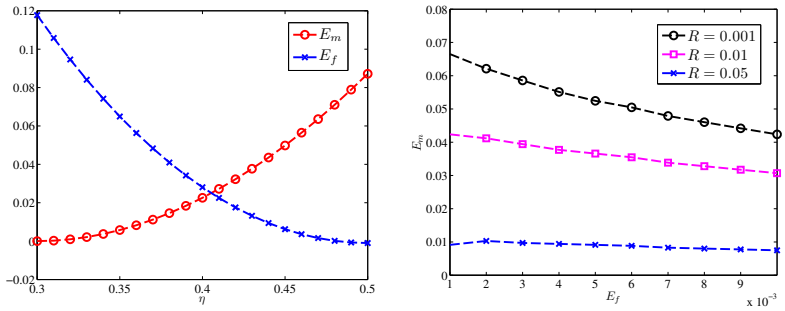
$$P_f \leq \exp[-L(\ln 2 - R - H_2(\eta))]. \tag{26}$$

Proof: Appendix G. ■

The corresponding false acceptance error exponent over the BSC is given by $E_f = \ln 2 - R - H_2(\eta)$.

Remark 13. 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 14. From Remarks 11 and 13, both P_m and P_f go to zero as $L \to \infty$. Moreover, this holds for η arbitrarily close to P_b. Therefore, the identification rate R approaches the identification capacity in the fingerprint domain $C = I(B_x; B_y) = \ln 2 - H_2(P_b)$ [32].



(a) $R = 0.001$, $N_l = 32$, SNR=0dB (b) $N_l = 32$, SNR=0dB

Fig. 4. Error exponents of the probabilities of miss and false acceptance as a function of (a) decision threshold η for a fixed R and (b) R.

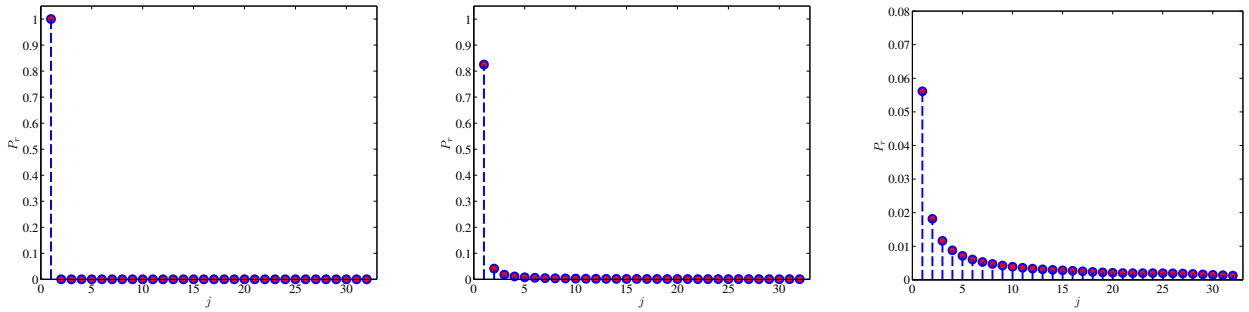
Fig. 4a and Fig. 4b exemplify the behaviour of error exponents (24) and (26) as functions of the decision threshold and the identification rate, respectively. These results were obtained using the following identification system operational parameters: $L = 64, R = 0.001, N_l = 32, P_b = 0.25$ obtained for $\mathrm{SNR} = 0$dB. Fig. 4a demonstrates that the probabilities of miss and false acceptance exponentially vanish to zero as functions of the fingerprint length with rates that exponentially decrease/grow with an increase in the decision threshold. In Fig. 4b, the obtained results demonstrate how the performance of the CBI degrades with the system identification rate.

E. Probabilistic Analysis of List Content

In this Section, we analyze the PMF of the correct entry position on the resulting list $\mathcal{N}_l'$. Because of the symmetric structure of the Database, we need only consider the case of the query being related to the first entry of the Database. Conditioned on $\mathcal{H}_1$, the probability that the first index falls in the j^{th} position of the primary list is given by:

$$\begin{aligned} P(j) &\overset{(a)}{=} \Pr\{(\mathcal{L}_{(M-j:M-1)} < \mathcal{L}_1) \cap (\mathcal{L}_{(M-j+1:M-1)} \geq \mathcal{L}_1) | \mathcal{H}_1\} \\ &\overset{(b)}{=} \sum_{d=0}^{L} \Pr\{(D_{(j:M-1)} > d) \cap (D_{(j-1:M-1)} \leq d) | \mathcal{H}_1, D_1 = d\} p_{D_1}(d) \\ &\overset{(c)}{=} \sum_{d=0}^{L} \binom{L}{d} P_b^d (1-P_b)^{L-d} \binom{M-1}{j-1} \left[\left(\tfrac{1}{2}\right)^L \sum_{x=0}^{d} \binom{L}{x} \right]^{j-1} \left[\left(\tfrac{1}{2}\right)^L \sum_{x=d+1}^{L} \binom{L}{x} \right]^{M-j}, \end{aligned} \tag{27}$$

where (a) follows from the fact that the first index falls in the j^{th} position, if $\mathcal{L}_1$ is smaller than $(j-1)$ largest of $\{\mathcal{L}_2, \ldots, \mathcal{L}_m\}$ and larger than the rest, (b) holds similar to (15), and (c) since the event $(D_{(j:M-1)} > d) \cap (D_{(j-1:M-1)} \leq d)$ occurs if exactly $j-1$ of $D_2, D_3, \ldots, D_M$ are less than or equal to d and exactly $M-j$ of $D_2, D_3, \ldots, D_M$ are bigger than d.



(a) SNR=0dB (b) SNR=-4dB (c) SNR=-12dB

Fig. 5. Primary list order statistics of the relevant query in CBI with $M = 2^{30}$ and $L = 512$ in channels with distortions.

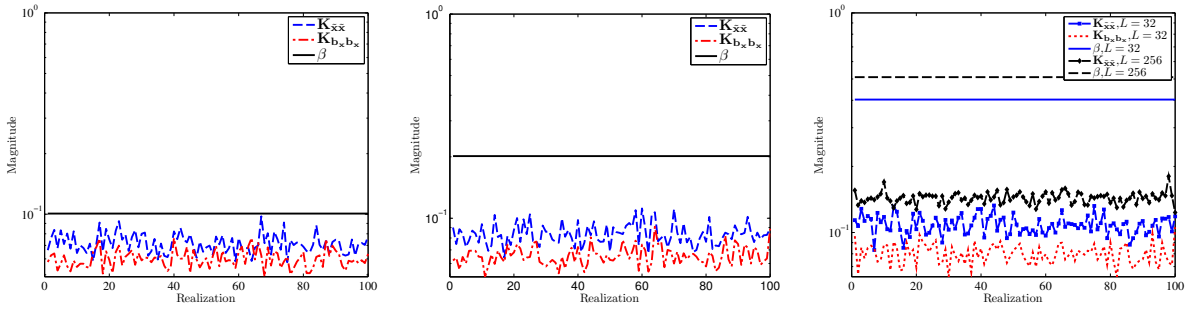
In Fig. 5, we show the impact of channel distortions on the position of the related entry in the primary list of candidates (27) for a codebook of $M = 2^{30}$ of length $L = 1024$ fingerprints. The primary list size N_l was fixed to 32, and the identification channel distortions were assumed to be originated from the BSC for the range of SNRs from 0dB to -12dB. One can observe the natural consequence that by increasing the channel degradation, the probability for the related entry falling into the first position decreases. This behaviour also justifies why the CLB decoding is able to return the correct index on the list for the proper SNR while the ML decoding, which always selects the candidate with the largest likelihood function completely fails. Moreover, in the strong distortion case, the likelihood of a correct candidate on the list is almost equal to the ones of other list members. This makes the query almost equally likely in relation to all elements on the list. Finally, the exponential character of the analyzed PMF allows for the determining of a reasonable list size that will guarantee the desired value of P_d for human-centric systems. In the case P_d is assumed to be equal 0.90, the average cardinalities of $\mathcal{N}_l$ are 1, 16, and 2^{30}, respectively, for 3 operational regimes specified in Fig. 5.

VII. Numerical Evaluation

We validate the theoretical results by experiments using a database of synthetically generated sequences and a real database of images.

A. Numerical results using Synthetic Database

In this Section, we perform the analysis of theoretical results presented in Sections V and VI based on synthetic database. The goal of this analysis is twofold. In the first part, we experimentally confirm our theoretical findings about the decorrelation property of RP. In the second part, we analyze the performance of the identification system in terms of P_d and P_f. For this purpose, we generated three databases with $M = 1024$ and length $N = 4096$ according to the AR(1) Gaussian process with $\rho = 0$, $\rho = 0.5$ and $\rho = 0.75$. The goal of the first experiment consists in the investigation of maximum residual correlation in the projected data $\tilde{\mathbf{x}}$ using 100 realizations of RP. According to the theoretical results (8a), it is

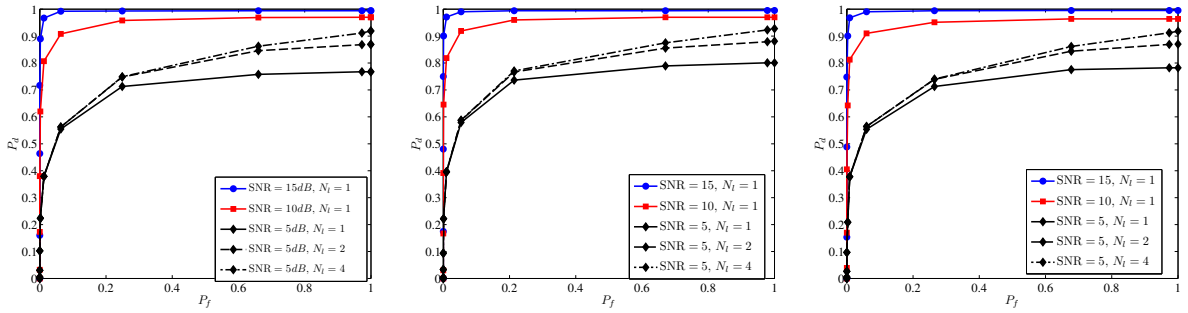


(a) $\rho = 0$ (b) $\rho = 0.5$ (c) $\rho = 0.75$

Fig. 6. The decorrelation and approximate i.i.d. preservation properties of the RP: (a), (b) and (c) the maximum off-diagonal elements of the covariance matrix of the projected data $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ and the binarized version of the projected data $\mathbf{K}_{\mathbf{b}_\mathbf{x}\mathbf{b}_\mathbf{x}}$, using 100 different realization of the RP, where the original data $\mathbf{x}$ is generated from the AR(1) Gaussian process with $\sigma_X^2 = 1$ and $\rho = 0$, $\rho = 0.5$ and $\rho = 0.75$, respectively; (c) also shows the impact of $L = 32$ and $L = 256$.

expected that the residual correlation in the projected data is negligible, which confirms the decorrelation property of RP. As in the enrollment stage, we first apply the RP defined in Section V to reduce the dimensionality of each item to $L = 32$. Each element of the projected items is then quantized to one bit according to whether it is greater and equal or less than zero. Fig. 6 shows the impact of the normalized correlation coefficient ρ and dimensionality reduction L on the maximum value of off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ and $\mathbf{K}_{\mathbf{b}_\mathbf{x}\mathbf{b}_\mathbf{x}}$ using 100 different realizations of $\mathbf{W}$. Based on the simulation results shown in Figs. 6, one can conclude that the elements of the randomly projected data, which are generated from the AR(1) Gaussian process, are approximately uncorrelated. Therefore, it is possible to assume that the resulting coefficients follow approximately i.i.d Gaussian distribution, due to the linearity of RP and the joint Gaussian distribution of the input. Moreover, it is important to point out the tightness of the bounds obtained in (8a) and (8b). The obtained results show that the bounds' tightness is a decreasing function of the correlation coefficient in the transform input. The main argument justifying such a behaviour is the use of Chernoff's bounding techniques to data obeying a residual correlation. The development of tighter bounds for ($\rho \geq 0.75$) is left for future research.

In the second part of our modelling, we experimentally validate the identification system using ROC curves under AWGN distortions. Fig. 7 shows the CLB decoding performance analysis for this type of degradation for the databases mentioned above. The target range of operational SNRs is limited to $[5, 15]$dB, assuming that the channel noise is zero-mean white Gaussian with variance σ_Z^2. Based on these experimental results, one can conclude that the CLB decoding performance under the proposed approach of digital fingerprint extraction is approximately independent of the correlation between elements of the data in the observation domain.



(a) $\rho = 0$ (b) $\rho = 0.5$ (c) $\rho = 0.75$

Fig. 7. The identification performance analysis using CLB decoding under AWGN distortions for a database of binary fingerprints of dimensionality $L = 32$ and cardinality $M = 1024$ which are extracted from the AR(1) Gaussian process with $\sigma_X^2 = 1$ $\rho = 0$, $\rho = 0.5$ and $\rho = 0.75$.

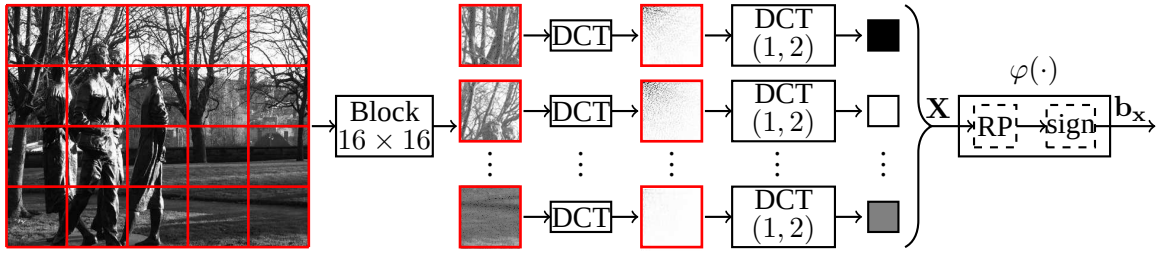


Fig. 8. Fingerprint extraction from images.

B. Numerical results using Image Database

In this subsection, we compare the performance predicted by the theoretical analysis with simulation results obtained using an image database. The used database, UCID [20] [3], consists of 1338 image of size 384 by 512. Fig. 8 illustrates a basic fingerprint extraction scheme under analysis based on 2D DCT. In this scheme, each image is converted to gray scale and divided in 16 by 16 blocks and the 2D DCT of each block is computed. The feature vector is constructed by concatenating the DCT coefficients at the coordinates $(1, 2)$ inside each block. Finally, the binary fingerprint of length $L = 32$ from each feature vector is extracted by applying RP and binarization. The main reason to use such a fingerprint extraction scheme is justified according to the well-known decorrelation properties of DCT. In the case of AR(1), the DCT closely approximates KLT [16] and provides almost perfect energy compaction and decorrelation of the transformed coefficients. However, the obtained coefficients will have different statistics according to the consideration of their covariance matrix, i.e., different variances in the case of considered AR(1)-model. Therefore, to satisfy the condition of stationarity, one can select the coefficients that possess the same statistics in each block of DCT. In the above scheme, we have just chosen the coefficient at the

[3]the same database was used in [4] for fingerprinting system validation

TABLE I

DATA STATISTICS IN DIFFERENT DOMAINS.

Feature domain	RP domain		Binary domain	
ρ	$\max_{i\neq j} \mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$	$\theta_{\tilde{X}}$	$\max_{i\neq j} \mathbf{K}^{ij}_{\mathbf{b}_x\mathbf{b}_x}$	$\hat{P}$
0.41	0.08	1.74	0.07	0.5

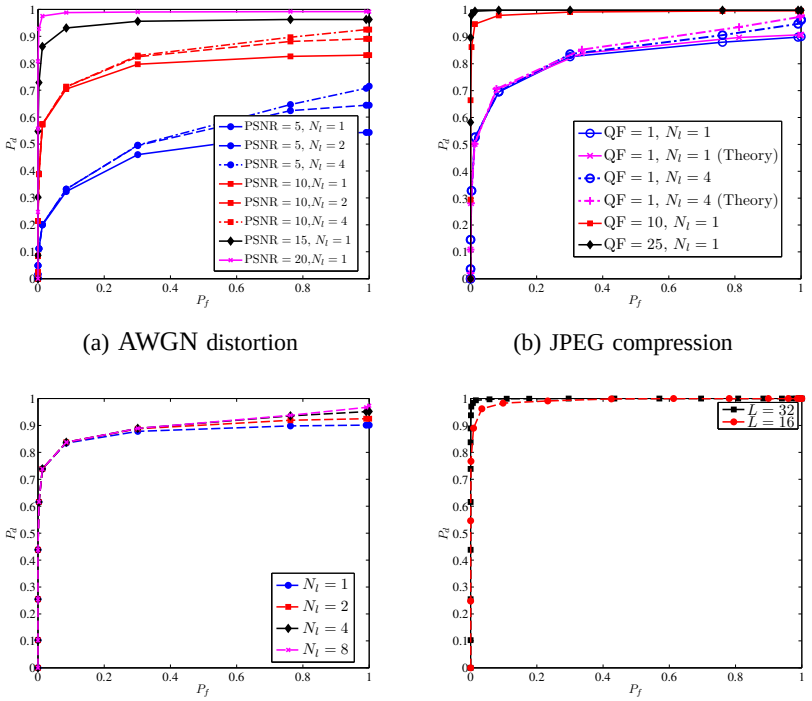
coordinate (1,2). Obviously, other choices are possible. However, to keep a good trade-off between the robustness and distinguishablity, the preferred coefficient should be in the range of middle frequencies.

Table I summarizes the statistics of feature vectors extracted from the UCID images. The AR(1) model parameter of feature vectors ρ is estimated by using the well-known least square method and Yule-Walker equation [16], and by averaging over all images in the database. As shown in this table, the coefficients obey a certain correlation with a normalized correlation coefficient $\rho = 0.41$. The correlation between components of the projected feature vectors are almost negligible and their marginal distribution approximately follows the Gaussian distribution. Finally, the extracted digital fingerprints follow $\mathrm{Bernoulli}(0.5)$. This confirms to the main assumptions used in the theoretical analysis of fingerprint performance and also demonstrates the ability of the fingerprinting system to generate an unique fingerprint for each image. We evaluate the ability of the identification system to correctly identify an image after it has undergone the potential malicious attacks listed in Table II, where $\mathrm{PSNR} = 10\log_{10}\frac{255^2}{\sigma_Z^2}$. We compare the predicted P_b, which is evaluated based on the AWGN model (10) with σ_X^2 and σ_Z^2 estimated in the RP domain, and the empirical $\hat{P}_b$, which is the average bit flipping in the binary domain. One can conclude a good match between the predicted and estimated crossover probabilities for all class of distortions except the histogram equalization. Furthermore, in the histogram equalization attack, the predicted P_b using the GGD model (11) with $\theta_{\tilde{Z}} = 1.19$ (Table II) equals to 0.11 that is quite close to the empirical $\hat{P}_b = 0.1$ (Table II). The results imply that the assumption of the additive and independent noise and the corresponding results in the RP domain is approximately valid.

TABLE II

LIST OF ATTACKS TESTED AND THE CORRESPONDING NOISE STATISTICS.

Attack	Parameters	Feature domain			RP domain			Binary domain			
		$\max_{i\neq j} \mathbf{K}^{ij}_{\mathbf{zz}}$	θ_Z	$\max \mathbf{K}_{\mathbf{zx}}$	$\max_{i\neq j} \mathbf{K}^{ij}_{\tilde{\mathbf{z}}\tilde{\mathbf{z}}}$	$\theta_{\tilde{Z}}$	$\max \mathbf{K}_{\tilde{\mathbf{z}}\tilde{\mathbf{x}}}$	$\max_{i\neq j} \mathbf{K}^{ij}_{\mathbf{b}_z\mathbf{b}_z}$	P_b	$\hat{P}_b$	$\max \mathbf{K}_{\mathbf{b}_z\mathbf{b}_x}$
AWGN	PSNR=5 dB	0.03	2	0.03	0.08	2	0.09	0.08	0.20	0.21	0.09
	PSNR = 10 dB	0.03	2	0.03	0.07	2	0.07	0.07	0.12	0.13	0.08
	PSNR = 15 dB	0.03	2	0.03	0.08	2	0.08	0.08	0.07	0.08	0.09
	PSNR = 20 dB	0.03	2	0.03	0.08	2	0.09	0.08	0.04	0.05	0.09
JPEG	QF = 1	0.04	1.2	0.09	0.1	1.93	0.1	0.1	0.10	0.11	0.09
	QF = 10	0.04	1.8	0.05	0.07	1.96	0.08	0.08	0.03	0.04	0.09
	QF = 25	0.03	1.95	0.06	0.06	1.99	0.09	0.07	0.01	0.01	0.09
Histeq		0.15	0.75	0.49	0.2	1.19	0.3	0.12	0.14	0.1	0.09



(a) AWGN distortion

(b) JPEG compression

(c) Histogram Equalization

(d) Histogram Equalization (authentication)

Fig. 9. (a), (b) and (c) the identification performance analysis using CLB decoding under the attacks listed in Table II for a database of binary fingerprints of dimensionality $L = 32$ and cardinality $M = 1338$ which are extracted from UCID images; (d) the detector performance for different length of fingerprints.

To evaluate the performance of the identification setup using the CLB decoding, the hypothesis testing described in Section IV is performed and P_m and P_f are evaluated. The obtained experimental and theoretical results demonstrate an excellent match. Figs. 9a, 9b and 9c show the ROC curves of the CLB decoding performance over the AWGN, JPEG compression and histogram equalization (Histeq) attacks, respectively, with different primary list sizes. Although increasing the primary list size can enhance the performance in terms of P_d, this improvement is restricted over the specific range of primary list sizes if low P_f is acceptable. Moreover, Fig. 9b demonstrates the simulation versus analytical results for JPEG compression with the quality factor equals to 1 using (17) and (20) assuming $P_b = 0.1$ (Table II), $M = 1338$, $N_l = 1$ and $N_l = 4$. Additionally, to validate the detection capabilities of the considered method, Fig. 9d shows the detector performance under authentication setup defined in Remark 7 using the advocated fingerprint extraction scheme under the Histeq attack for $L = 16$ and $L = 32$ in terms of P_m and P_f, where P_f is estimated using 1000 pairs of randomly chosen original fingerprints.

VIII. Conclusions

We analyzed the identification setup based on the CLB decoding framework. In light of this framework,

we have investigated the CLB decoding performance by deriving closed-from analytical expressions and bounds for the probabilities of correct detection and false acceptance. The theoretical results are also validated on synthetic data and a set of real images.

The simulation results show that, on the one hand, the CLB decoding can only improve the identifier performance in low SNR scenarios. On the other hand, this improvement is restricted by a certain range of list sizes. The fingerprinting algorithm proposed in this context demonstrated an excellent performance under the class of considered attacks that might even generate content dependent noise. In fact, we have shown that the assumption of independence is valid and one can evaluate the performance of the identification setup based on i.i.d. binary fingerprints.

Further improvement of the fingerprinting performance is envisioned by optimizing the feature extraction with respect to its robustness to malicious attacks, especially geometrical attacks that have not been directly addressed in this paper. In this respect, the use of robust features looks very promising; features such as SIFT and their combination with the block-based DCT considered in this manuscript. One can also predict an essential improvement in the complexity of an identification problem by using the BDD proposed in [5]. The complexity of this identification technique critically depends on P_b and the fingerprint length. The achievable values of P_b demonstrated in this paper, makes it feasible to perform a search through a large-scale database even on computers with moderate computational power. Moreover, an additional reduction in complexity is possible with the use of the bit reliability that was also considered in [33]. In this respect, the theoretical framework developed in the current paper should serve as a basis for the theoretical analysis of soft fingerprinting systems. Finally, we plan to test the identification setup on several databases of medical images and physical unclonable microstructure images to evaluate the uniqueness of fingerprints and their performance in terms of ROC curves. It is also very important that security and privacy analysis be included in the proposed framework under consideration, and that appears to be our next research challenge.

ACKNOWLEDGMENT

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

REFERENCES

[1] F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, "Performance analysis of identification system based on order statistics list decoder," in *IEEE International Symposium on Information Theory*, Austin, TX, June, 13-18 2010.

[2] F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, "Content identification based on digital fingerprint: what can be done if ml decoding fails?," in *IEEE International Workshop on Multimedia Signal Processing*, France, October 2010.

[3] Ol. Koval, S. Voloshynovskiy, F. Beekhof, and T. Pun, "Decision-theoretic consideration of robust perceptual hashing: link to practical algorithms," in *WaCha2007, /Third WAVILA Challenge/*, Saint Malo, France, June 15th 2007.

[4] A. L. Varna and M. Wu, "Modeling and analysis of correlated binary fingerprints for content identification," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1146 – 1159, September 2011.

[5] S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak, "Information-theoretical analysis of private content identification," in *Proc. of IEEE Inf. Theory Workshop*, Dublin, Ireland, August 2010.

[6] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun, "Conception and limits of robust perceptual hashing: toward side information assisted hash functions," in *Proc. of SPIE, Media Forensics and Security*, San Jose, USA, 2009.

[7] J. Fridrich, "Robust bit extraction from images," in *MCS*, july 1999, vol. 2, pp. 536 –540.

[8] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system.," in *ICM Information Retrieval,*, 2002.

[9] G. David F. Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *Information Theory, IEEE Transactions on*, vol. 14, no. 2, pp. 206 – 220, Mar. 1968.

[10] P. Elias, "List decodeing for noisy channels," Tech. Rep. 335, Research Labratoary of Electronics, M.I.T, 1955.

[11] P. Moulin, "Statistical modeling and analysis of content identification," in *ITA*, San Diego, CA, 2010.

[12] F. Willems, T. Kalker, J. Goseling, and J. Linnartz, "On the capacity of a biometrical identification system," in *Information Theory, 2003. Proceedings. IEEE International Symposium on*, 2003, p. 82.

[13] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[14] F. M.J. Willems, "Information theory and biometrics," in *Keynote Lecture at the Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing,*, Darmstadt, Germany, October 2010.

[15] A. Papoulis and S. U. Pillai, *Probabikity, Random Variables and Stochastic Processes*, McGrawHill, 2002.

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

[17] G. H. Golub and C. F. van Loan, *Matrix Computations*, North Oxford Academi, Oxford, UK, 1983.

[18] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mapping into a hilbert space," in *Conf. in Modern analysis and probability*. 1984, vol. 26 of *Contemporary Mathematics*, pp. 189–206, AMS.

[19] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 445–460, 2010.

[20] G. Schaefer and M. Stich, "Ucid - an uncompressed colour image database," in *Storage and Retrieval Methods and Applications for Multimedia*, 2004, Proc.of SPIE.

[21] H. A. David and H. N. Nagaraja, *Order Statistics*, Wiley-Interscience, 3ed. edition, 2003.

[22] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, pp. 671–687, June 2003.

[23] D.G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, vol. 2, pp. 1150 –1157.

[24] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *ECCV*, 2006, pp. 404–417.

[25] B.S. Reddy and B.N. Chatterji, "An fft-based technique for translation, rotation, and scale-invariant image registration," *Image Processing, IEEE Transactions on*, vol. 5, no. 8, pp. 1266 –1271, 1996.

[26] C. Kim, "Content-based image copy detection," *Signal Processing*, vol. 18, no. 3, pp. 169 – 184, 2003.

[27] L. Juan and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *Image Processing, IEEE Transactions on*, vol. 10, no. 11, pp. 1647–1658, nov 2001.

[28] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *international Journal of Computer Vision*, vol. 60, pp. 63–86, October 2004.

[29] E. McCarthy, F. Balado, G.C.M. Slvestre, and N.J. Hurley, "A framework for soft hashing and its application to robust image hashing," in *International Conference on Image Processing*, oct. 2004, vol. 1.

[30] A. L. Varna, A. Swaminathan, and M. Wu, "A decision theoretic framework for analyzing binary hash-based content identification systems," in *Proceedings of the ACM workshop on Digital rights management*, New York, NY, USA, 2008.

[31] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & sons, New York, NY, USA, 1968.

[32] P. Tuyls, B. Skoric, and T. Kevenaar, *Security with noisy data: On Private Biometrics, Secure Key Storage and Anti-Counterfeiting*, Springer, 2007.

[33] F. Beekhof, S. Voloshynovskiy, O. Koval, and T. Holotyak, "Fast identification algorithms for forensic applications," in *Proc. of IEEE Int. Workshop on Inf. Forensics and Security*, London, UK, 2009.

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

[35] 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: At first, we consider off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ that can be expanded as follows:

$$\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \sum_{r=1}^{N}\sum_{c=1}^{N} w_{ir}\mathbf{K}^{rc}_{\mathbf{xx}}w_{jc} = \sigma_X^2\Bigg[\underbrace{(w_{i1}w_{j1}+\cdots+w_{iN}w_{jN})}_{t_0^{ij}} + \rho\underbrace{(w_{i1}w_{j2}+w_{i2}w_{j1}+\cdots+w_{iN}w_{jN-1})}_{t_1^{ij}}$$
$$+\ldots+\rho^{N-1}\underbrace{(w_{i1}w_{jN}+w_{iN}w_{j1})}_{t_{N-1}^{ij}}\Bigg] = \sigma_X^2\sum_{k=0}^{N-1}\rho^k t_k^{ij}. \tag{28}$$

Due to symmetry of a covariance matrix, we just investigate upper off-diagonal elements, i.e., $1 \le i < j \le L$. In order to bound these elements, we evaluate an upper bound for the probability that the largest upper off-diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ is greater than $\sigma_X^2\left(\frac{1-\rho^N}{1-\rho}\right)\zeta$, where ζ is a positive real value:

$$\Pr\left\{\max_{i\neq j}|\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}| > \sigma_X^2\left(\frac{1-\rho^N}{1-\rho}\right)\zeta\right\} \overset{(a)}{\le} \frac{L(L-1)}{2}\Pr\left\{|\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}| > \sigma_X^2\left(\frac{1-\rho^N}{1-\rho}\right)\zeta\right\}$$
$$= \frac{L(L-1)}{2}\Pr\left\{\frac{1}{\sigma_X^2}\left(\frac{1-\rho}{1-\rho^N}\right)|\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}| > \zeta\right\} \overset{(b)}{\le} L(L-1)\exp(-s\zeta)\mathbb{E}\left[\exp\left(s\frac{1}{\sigma_X^2}\left(\frac{1-\rho}{1-\rho^N}\right)\mathbf{K}^{ij}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\right)\right]$$
$$\overset{(c)}{=} L(L-1)\exp(-s\zeta)\mathbb{E}\left[\exp\left(s\sum_{k=0}^{N-1}\kappa_k T_k^{ij}\right)\right] \overset{(d)}{\le} L(L-1)\exp(-s\zeta)\left[\sum_{k=0}^{N-1}\kappa_k\mathbb{E}\left[\exp\left(sT_k^{ij}\right)\right]\right]$$
$$\overset{(e)}{=} L(L-1)\sum_{k=0}^{N-1}\kappa_k\left[\exp(-s\zeta)\prod_{l=1}^{N_k}\mathbb{E}\left[\exp\left(sV_l^{ij}\right)\right]\right] \overset{(f)}{\le} L(L-1)\sum_{k=0}^{N-1}\kappa_k\left[\exp(-s\zeta)\prod_{l=1}^{N_k}\exp\left(\frac{s^2}{2N^2}\right)\right]$$
$$\overset{(g)}{=} L(L-1)\sum_{k=0}^{N-1}\kappa_k\left[\exp\left(-\frac{N^2\zeta^2}{2N_k}\right)\right] = L(L-1)\left[\kappa_0\exp\left(-\frac{N\zeta^2}{2}\right)+\sum_{k=1}^{N-1}\kappa_k\exp\left(-\frac{N^2\zeta^2}{(N-k)4}\right)\right]$$
$$\le L(L-1)\left[\kappa_0\exp\left(-\frac{N\zeta^2}{2}\right)+\sum_{k=1}^{N-1}\kappa_k\exp\left(-\frac{N^2\zeta^2}{(N-1)4}\right)\right] \le L(L-1)\exp\left(-\frac{N\zeta^2}{4}\right). \tag{29}$$

where $\kappa_k = \rho^k \left(\frac{1-\rho}{1-\rho^N}\right)$ and $\sum_{k=0}^{N-1} \kappa_k = 1$, (a) since there are only $\frac{L(L-1)}{2}$ such identically distributed random variables [34], (b) follows form the generalized Chernoff bound [31] for $s \geq 0$, (c) holds from (28), (d) holds due to the convexity of $\exp(\cdot)$, (e) follows that $\mathbb{E}\left[\exp(sT_k^{ij})\right]$ is the moment generating function of T_k^{ij} that is the sum of $N_k \in \{N, 2(N-1), \ldots, 2\}$ i.i.d. Bernoulli$(\frac{1}{2})$ random variables $V^{ij} = W_{ir}W_{jc} \in \{\frac{+1}{N}, \frac{-1}{N}\}$, (f) holds since V^{ij} is a bounded random variable [35], and (g) holds by choosing $s = \frac{N^2\zeta}{N_k}$. By substituting $\zeta = \sqrt{\frac{12}{N}\ln L}$ and setting $\beta = \left(\frac{1-\rho^N}{1-\rho}\right)\sqrt{\frac{12}{N}\ln L}$, (8a) is obtained.

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

$$\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \sigma_X^2 + \sum_{\substack{r=1\\ r\neq c}}^{N}\sum_{c=1}^{N} w_{ir}\mathbf{K}^{rc}_{\mathbf{xx}}w_{ic} = \sigma_X^2 + 2\sigma_X^2\rho\underbrace{(w_{i1}w_{i2} + \cdots + w_{iN-1}w_{iN})}_{d_1^{ii}}$$

$$+ 2\sigma_X^2\rho^2\underbrace{(w_{i1}w_{i3} + \cdots + w_{iN-2}w_{iN})}_{d_2^{ii}} + \ldots + 2\sigma_X^2\rho^{N-1}\underbrace{(w_{i1}w_{iN})}_{d_{N-1}^{ii}} = \sigma_X^2 + 2\sigma_X^2\sum_{k=1}^{N-1}\rho^k d_k^{ii},$$

Similar to (29), we evaluate an upper bound for the probability that the maximum deviation of diagonal elements of $\mathbf{K}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}$ from σ_X^2 exceeds $2\sigma_X^2\left(\frac{\rho-\rho^N}{1-\rho}\right)\epsilon$, where $\epsilon > 0$. This probability is given by:

$$\Pr\left\{\max_i |\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \sigma_X^2| > 2\sigma_X^2\left(\frac{\rho-\rho^N}{1-\rho}\right)\epsilon\right\} \overset{(a)}{\leq} L\Pr\left\{|\mathbf{K}^{ii}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \sigma_X^2| > 2\sigma_X^2\left(\frac{\rho-\rho^N}{1-\rho}\right)\epsilon\right\}$$

$$\overset{(b)}{=} 2L\exp(-s\epsilon)\mathbb{E}\left[\exp\left(s\sum_{k=1}^{N-1}\lambda_k D_k^{ii}\right)\right] \overset{(c)}{\leq} 2L\exp(-s\epsilon)\left[\sum_{k=1}^{N-1}\lambda_k\mathbb{E}\left[\exp\left(sD_k^{ii}\right)\right]\right]$$

$$\overset{(d)}{\leq} \sum_{k=1}^{N-1}\lambda_k\exp\left(-\frac{N^2\epsilon^2}{2(N-k)}\right) \leq \exp\left(-\frac{N\epsilon^2}{2}\right) \tag{30}$$

where $\lambda_k = \rho^k\left(\frac{1-\rho}{\rho-\rho^N}\right)$ and $\sum_{k=1}^{N-1}\lambda_k = 1$, (a) follows from the fact that there are only L such random variables which are identically distributed [34], (b) follows form the generalized Chebyshev inequality [31] for $s \geq 0$, (d) results from the convexity of $\exp(\cdot)$, and (e) holds following the same results of parts (f) and (g) in (29) and by choosing $s = \frac{N^2\zeta}{2(N-k)}$. By substituting $\epsilon = \sqrt{\frac{2}{N\rho}\ln L}$ and setting $\alpha = \left(\frac{1-\rho^{N-1}}{1-\rho}\right)\sqrt{\frac{8}{N}\rho\ln L}$, (8b) is obtained. ■

APPENDIX B
PROOF OF COROLLARY 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 (9) 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 $E_{\mathcal{I}_j}$ as the event that there exists a subset of indices $\mathcal{I}_j \subset \mathcal{M}' = \{2, \ldots, M\}$ with $|\mathcal{I}_j| = N_l$ whose all log-normalized-likelihoods, i.e., $\mathcal{L}_{j(i)}, 1 \leq i \leq N_l$, satisfies $\mathcal{L}_{j(i)} > \mathcal{L}_1$ and $\mathcal{L}_{j(i)} \geq \gamma L$. P_m^{I} can be bounded for any $0 \leq s \leq 1$ as follows:

$$P_m^{\mathrm{I}} = \Pr\{(m_1 \notin N_l) \cap (\mathcal{L}_1 \geq \gamma L)|\mathcal{H}_1\} = \Pr\left\{\bigcup_{j=1}^{C} E_{\mathcal{I}_j}\middle|\mathcal{H}_1\right\} \overset{(a)}{\leq} \sum_{j=1}^{C} \Pr\left\{E_{\mathcal{I}_j}\middle|\mathcal{H}_1\right\}$$

$$= \sum_{j=1}^{C} \Pr\{(\mathcal{L}_{j(1)} \geq \mathcal{L}_1 \cap \mathcal{L}_1 \geq \gamma L) \cap \cdots \cap (\mathcal{L}_{j(N_l)} \geq \mathcal{L}_1 \cap \mathcal{L}_1 \geq \gamma L)|\mathcal{H}_1\}$$

$$\overset{(b)}{=} C\Pr\{\mathcal{L}_{m\neq 1} \geq \mathcal{L}_1 \cap \mathcal{L}_1 \geq \gamma L|\mathcal{H}_1\}^{N_l} \overset{(c)}{\leq} (M-1)^{N_l}\Pr\{\mathcal{L}_{m\neq 1} \geq \gamma L|\mathcal{H}_1\}^{N_l}$$

$$\overset{(d)}{\leq} \left\{e^{LR}e^{-u\gamma L}\mathbb{E}[\exp[u\mathcal{L}_m]]\right\}^{N_l} = \left\{e^{LR}e^{-u\gamma L}\prod_{n=0}^{L-1}\mathbb{E}\left[\exp\left(u\ln\frac{p(y[n]|x(m)[n])}{p(y[n])}\right)\right]\right\}^{N_l}$$

$$= \left\{e^{LR}e^{-u\gamma L}\left\{\mathbb{E}\left[\exp\left(u\ln\frac{p(y|x)}{p(y)}\right)\right]\right\}^{L}\right\}^{N_l} = \left\{e^{LR}e^{-u\gamma L}\left\{\sum_{x\in\mathcal{X}}p(x)\sum_{y\in\mathcal{Y}}p(y)^{1-u}p(y|x)^{u}\right\}^{L}\right\}^{N_l}$$

$$= \left\{e^{LR}e^{-\gamma L}\left\{\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}p(x)p(y)^{s}p(y|x)^{1-s}e^{s\gamma}\right\}^{L}\right\}^{N_l} \leq \{\exp[-L(E_o(s,\gamma)+\gamma-R)]\}^{N_l} \tag{31}$$

where $C = \binom{M-1}{N_l}$, (a) follows from the union bound [13], (b) follows from the fact that the events are independent and $\mathcal{L}_m$s, $m \neq 1$, are i.i.d. random variables, (c) follows from the fact that $\Pr\{\mathcal{L}_{m\neq 1} \geq \mathcal{L}_1 \cap \mathcal{L}_1 \geq \gamma L|\mathcal{H}_1\} \leq \Pr\{\mathcal{L}_{m\neq 1} \geq \mathcal{L}_1|\mathcal{L}_1 \geq \gamma L, \mathcal{H}_1\} \leq \Pr\{\mathcal{L}_{m\neq 1} \geq \gamma L|\mathcal{H}_1\}$ and inequlaity $\binom{M-1}{N_l} \leq (M-1)^{N_l}$, (d) follows from applying the Chernov bound [31] for any $0 \leq u \leq 1$ and $0 \leq s \leq 1$. And, $E_o(s,\gamma)$ is defined as:

$$E_o(s,\gamma) = -\ln\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}p(x)p(y)^{s}p(y|x)^{1-s}e^{s\gamma}. \tag{32}$$

Finally, since s is arbitrary in (32), we get the tightest bound by choosing s to maximize $E_o(s,\gamma)$. Therefore, we have

$$P_m^{\mathrm{I}} \leq \{\exp[-L(E(\gamma)+\gamma-R)]\}^{N_l}, \tag{33}$$

where $E(\gamma)$ is defined in (23). Using the Chernov bound, P_m^{II} can be bounded for any $-1 \leq \nu \leq 0$ as:

$$P_m^{\mathrm{II}} = \Pr\{\mathcal{L}_1 < \gamma L|\mathcal{H}_1\} \leq e^{-\nu\gamma L}\mathbb{E}[\exp(\nu\mathcal{L}_1)] = e^{-\nu\gamma L}\prod_{n=0}^{L-1}\mathbb{E}\left[\exp\left(\nu\ln\frac{p(y[n]|x(1)[n])}{p[y[n]]}\right)\right]$$

$$= e^{-\nu\gamma L}\prod_{n=0}^{L-1}\sum_{x\in\mathcal{X}}p(x)\sum_{y\in\mathcal{Y}}p(y)^{-\nu}p(y|x)^{1+\nu} \overset{(a)}{=} \left\{\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}p(x)p(y)^{s}p(y|x)^{1-s}e^{s\gamma}\right\}^{L}$$

$$= \exp[-LE_o(s,\gamma)] \leq \exp[-LE(\gamma)] \tag{34}$$

where (a) follows from letting $s = 1 - \nu$, $0 \leq s \leq 1$. Combining (33) and (34), we obtain (22a). ■

APPENDIX D

PROOF OF COROLLARY 2

Proof: Using (32) one can see that $E_o(0,\gamma) = 0$, and it can be easily verified that $\frac{\partial E_o(s,\gamma)}{\partial s}\Big|_{s=0} = -\gamma + C$. Therefore, since for $\gamma < C$, $E_o(s,\gamma)$ is a non-decreasing function of s, then for any $0 \leq s \leq 1$, $E_o(s,\gamma) \geq 0$ implying $E(\gamma) \geq 0$. Moreover, as $\gamma \to C$, $\frac{\partial E_o(s,\gamma)}{\partial s}\big|_{s=0} \to 0$ leading to $E(\gamma) \to 0$. ■

APPENDIX E

PROOF OF COROLLARY 3

Proof: For inputs $p(0) = p(1) = \frac{1}{2}$, the output probability will be given by $p(0) = p(1) = \frac{1}{2}$. Then,

$$E_o(s,\gamma) = s\ln 2 - s\gamma - \ln\left[(1-P_b)^{1-s} + P_b^{1-s}\right]. \tag{35}$$

By setting the derivative of $E_o(s,\gamma)$ respective to s, equal to zero. We have

$$\gamma = \ln 2 + \eta \ln P_b + (1-\eta)\ln(1-P_b), \tag{36}$$

where $\eta = \frac{P_b^{1-s}}{P_b^{1-s}+(1-P_b)^{1-s}}$, was defined in (15). By substituting γ in (35), we get

$$E = \mathcal{D}(\eta \| P_b). \tag{37}$$

Finally, substituting (36) and (37) in (23), we obtain (24a). The equations are only valid for $P_b < \eta < \frac{1}{2}$.

APPENDIX F ■

PROOF OF PROPOSITION 3

Proof: Conditioned on $\mathcal{H}_0$, we define E_m, $1 \leq m \leq M$, as the event that there exists $\mathbf{x}^N(m)$ whose $\mathcal{L}_m \geq \gamma L$. The probability of false acceptance for any $0 \leq s \leq 1$ can be bounded as follows:

$$\begin{aligned}
P_f =& \Pr\left\{\bigcup_{j=m}^{M} E_m \middle| \mathcal{H}_0\right\} \overset{(a)}{\leq} \sum_{m=1}^{M} \Pr\left\{E_m \middle| \mathcal{H}_0\right\} = \sum_{m=1}^{M} \Pr\{\mathcal{L}_m \geq \gamma L | \mathcal{H}_0\} \overset{(b)}{\leq} M e^{-u\gamma L}\mathbb{E}[\exp[u\mathcal{L}_m]] \\
=& M e^{-u\gamma L} \prod_{n=0}^{L-1} \mathbb{E}\left[\exp\left(u \ln \frac{p(y[n]|x(m)[n])}{p(y[n])}\right)\right] = M e^{-u\gamma L} \prod_{n=0}^{L-1} \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y)^{1-u} p(y|x)^u \\
=& M e^{\gamma L} \left\{\sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x) p(y)^s p(y|x)^{1-s} e^{s\gamma}\right\}^L,
\end{aligned} \tag{38}$$

where (a) follows from the union bound, and (b) holds since the events are i.i.d. and using the Chernov bound for any $0 \leq s \leq 1$ and $0 \leq v \leq 1$. Finally substituting $M = e^{LR}$ in (38), we obtain (25). ■

APPENDIX G

PROOF OF COROLLARY 4

Proof: Similarly to the proof of Proposition 3, by substituting $\gamma = \ln 2 + \eta \ln P_b + (1-\eta)\ln(1-P_b)$ and $E = \mathcal{D}(\eta \| P_b)$ in (25), (26) is obtained. ■

Farzad Farhadzadeh received the M.Sc. degree in Signal Processing and Communication Systems from Electrical Engineering department at Amirkabir University of Technology, Tehran, Iran and the B.Sc. degree in electrical engineering from Isfahan University of Technology, Isfahan, Iran. Since 2009 he has joined Stochastic Image Processing group and has started the Ph.D. program in computer sciences department of University of Geneva. His research interests include information forensics and multimedia security.

Sviatoslav Voloshynovskiy (IEEE Senior Member'11) received the Radio Engineer degree from Lviv Polytechnic Institute, Lviv, Ukraine, in 1993 and the Ph.D. degree in electrical engineering from the State University *Lvivska Polytechnika*, Lviv, Ukraine, in 1996. From 1998 to 1999, he was a visiting scholar with University of Illinois at Urbana-Champaign. Since 1999, he has been with the University of Geneva, Switzerland, where he is currently an Associate Professor with the Department of Computer Science and head of the Stochastic Image Processing group. His current research interests are in information-theoretic aspects of digital data hiding, content fingerprinting, physical object security, stochastic image modeling and machine learning. He has coauthored over 200 journal and conference papers in these areas and holds ten patents. He has served as Associate Editor for IEEE Transactions on Information Forensics and Security. S. Voloshynovskiy is an elected member of the IEEE Information Forensics and Security Technical Committee (2011-2013) where he is area chair in information-theoretic security. He was a general chair of ACM Multimedia Security Conference, 2006. He has served as a consultant to private industry in the above areas.

Oleksiy Koval received his Master degree in Electrical Engineering from the National University *Lvivska Politechnika*, Lviv, Ukraine in 1996. In 1996-2001 he was with the Department of Synthesis, Processing and Identification of Images, Institute of Physics and Mechanics (Lviv Ukraine) as a researcher and Ph.D student. He received Ph.D degree in electrical engineering from the National University *Lvivska Politechnika* in 2002. Since 2002 he has been with Stochastic Image Processing Group, Computer Vision and Multimedia Laboratory, University of Geneva from which he received Ph.D degree in stochastic image modelling in 2004, where he is currently a Senior Lecturer and Researcher. His current research interests are in authentication and identification of objects, digital forensics, security of multimedia, machine learning, information-theoretic aspects of robust perceptual multimedia hashing, security, robustness and universality of robust perceptual multimedia hashing, information-theoretic aspects of data hiding. He has co-authored over 100 journal and conference papers in these areas as well as three patents. He has served as a consultant to private industry in the above areas.