Conception and limits of robust perceptual hashing: towards side information assisted hash functions

Sviatoslav Voloshynovskiy*, Oleksiy Koval, Fokko Beekhof and Thierry Pun
University of Geneva, Department of Computer Science,
7 route de Drize, CH 1227, Geneva, Switzerland

ABSTRACT

In this paper, we consider some basic concepts behind the design of existing robust perceptual hashing techniques for content identification. We show the limits of robust hashing from the communication perspectives as well as propose an approach capable to overcome these shortcomings in certain setups. The consideration is based on both achievable rate and probability of error. We use a fact that most of robust hashing algorithms are based on dimensionality reduction using random projections and quantization. Therefore, we demonstrate the corresponding achievable rate and probability of error based on the random projections and compare with the results for the direct domain. The effect of dimensionality reduction is studied and the corresponding approximations are provided based on Johnson-Lindenstrauss lemma. A side information assisted robust perceptual hashing is proposed as a solution to the above shortcomings.

Notations: We use capital letters to denote scalar random variables X and $\mathbf{X}$ to denote vector random variables, corresponding small letters x and $\mathbf{x}$ to denote the realizations of scalar and vector random variables, respectively. All vectors without sign tilde are assumed to be of the length N and with the sign tilde of length L with the corresponding subindexes. The binary representation of vectors will be denoted as $b_{\mathbf{x}}$ with the corresponding subindexing. We use $\mathbf{X} \sim p_{\mathbf{X}}(\mathbf{x})$ or simply $\mathbf{X} \sim p(\mathbf{x})$ to indicate that a random variable $\mathbf{X}$ is distributed according to $p_{\mathbf{X}}(\mathbf{x})$. $\mathcal{N}(\mu, \sigma_X^2)$ stands for Gaussian distribution with mean μ and variance σ_X^2. $||.||$ denotes Euclidean vector norm and $Q(.)$ stands for Q-function.

1. INTRODUCTION: BASIC DESIGNS

The robust perceptual hashing was originally considered as an alternative to the classical crypto based hashing algorithms known to be sensitive to any content modification. The main distinguishable feature of the robust perceptual hashing is the ability to withstand certain modifications while producing the same or at least very close (in a defined distance space) hash value. The applications of robust perceptual hashing are numerous and include content management (content identification, indexing and retrieval), content security (tracking of illegal copies, verification of authenticity, anticounterfeiting) as well as used as an assisting functionality for synchronization and alignment.

A common principle in the design of most of robust perceptual hashing is a mapping of the original data $\mathbf{x}$ to some *secure* but at the same time *robust* domain. This step is unavoidably accompanied by a dimensionality reduction also known as feature extraction:

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$, $\tilde{\mathbf{x}} \in \mathbb{R}^L$, $\mathbf{W} \in \mathbb{R}^{L \times N}$ and $L \leq N$ and $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_L)^T$ consists of a set of projection basis vectors $\mathbf{w}_i \in \mathbb{R}^N$ with $1 \leq i \leq L$. The second step uses also possibly a key-dependent labeling or the Grey codes to ensure the closeness of labels for the close vectors. Such kind of labeling is known as a soft hashing when only the most significant bit of the Grey code is used (achieved by a simple comparison with the threshold), it is known as binary or hard hashing.

The most simple quantization or binarization of extracted features is known as *sign random projections*:

$$b_{\mathbf{x}_i} = sign(\mathbf{w}_i^T \mathbf{x}), \tag{2}$$

*The contact author is S. Voloshynovskiy (email: svolos@cui.unige.ch). http://sip.unige.ch

where $b_{\mathbf{x}_i} \in \{0,1\}$, with $1 \leq i \leq L$ and $sign(a) = 1$, if $a \geq 0$ and 0, otherwise. The vector $\mathbf{b}_{\mathbf{x}} \in \{0,1\}^L$ computed for all projections represents a binary hash computed from the vector $\mathbf{x}$.

The example of mapping $\mathbf{W}$ used in the robust perceptual hashing are numerous and we will only mention some of them: Fridrich[7] uses block-based random projections generated from the uniform distribution and compares the resulted scalars with the threshold that can be considered as the second level bitplane in the Gray labeling (to enhance the robustness the projection vectors/fields are low-pass filtered that corresponds to the extraction of low-pass coefficients from the data $\mathbf{x}$); to enhance the security by randomized sampling Mihcak *et. al.*[18] use overlapping rectangles with the key dependent weights to compute the local statistics for further quantization (the overlapping rectangles can be also considered as the projection vectors/fields with zeros besides the support, where the random weights are generated, that makes them conceptually very close to Fridrich design); F. Lefebvre and B. Macq suggest to use Radon projections[12]; Kalker *at. al.*[8] consider overlapping blocks for audio hashing.

The performance analysis of robust perceptual hashing was mostly performed using computer simulation. Therefore, there is a real need in the thorough investigation of theoretical limits of robust hashing. The first efforts in this direction have been reported in.[3, 6, 16, 17, 24] However, the simultaneous impact of dimensionality reduction and binarization still remains uncovered in terms of both identification rate and average probability of error.

Another important aspect of robust perceptual hashing is security. The main belief behind the construction of good hashing algorithms was a randomization property, i.e., the good hash should have largest possible entropy. Swaminathan *et. al.* considered the entropy of different featured used in the state-of-the-art robust hashing algorithms[22] and later this analysis was extended to the crypto-based measures such as equivocation and unicity distance.[11, 14] The main analysis is performed along the line of investigating transformations that are difficult to invert. This links the robust perceptual hashing based on non-invertible transformations with similar transforms applied in biometric database protection against impersonation attack. At the same time, the growing number of publications in the recently emerged domain of *compressive sensing* demonstrates a possibility to accurately reconstruct some classes of sparse signals from low-dimensionality projections.[4, 20] This also raises certain security concerns about some transformations and their corresponding level of security.

Not less important problem of robust perceptual hashing is related to the fact that the entropy of source $H(X)$ exceeds the maximum number of reliably identifiable sequences under the certain distortions. The situation is inherently different to those in the digital communications where the construction of codebook is quite flexible.

Therefore, in this paper, we will make an attempt to consider the basic construction of state-of-the-art robust hashing techniques, evaluate the loss in performance in terms of achievable identification rate and probability of error and suggest alternative design capable to overcome the shortcomings of existing robust hashing methods.

2. SOURCE AND CHANNEL MODELS: DIRECT DOMAIN IDENTIFICATION CAPACITY AND PROBABILITY OF ERROR

To illustrate the limitations of robust perceptual hashing, we will refer to the identification setup in the scope of communication framework where some sequence should be identified based on its noisy observation on the output of channel. We will assume that the source is memoryless and produces the sequences that follow $p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{N} p_X(x_i)$. Assuming N is sufficiently large, we will use the concept of typicality,[5] according to which the maximum number of uniquely distinguishable sequences is limited by $M \leq 2^{NH(X)}$.

The robust hashing in the communication framework can be represented as in Figure 1. All sequences generated by the above source are indexed by the index m, $1 \leq m \leq M$. The sequences are communicated via a channel. We will assume the channel to be discrete memoryless channel (DMC), which is characterized by $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} p_{Y|X}(y_i|x_i)$. This channel is characterized by some capacity C_{id}. The decoder has to establish the index of input sequence $\mathbf{x}$ based on the channel output $\mathbf{y}$. The corresponding identification scheme based on practical implementation of robust hashing described in Section 1 is shown in Figure 2.

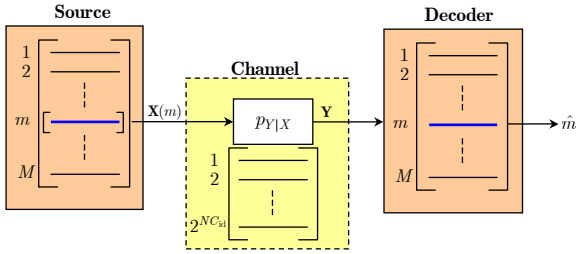


Figure 1. Robust hashing in the communication framework.

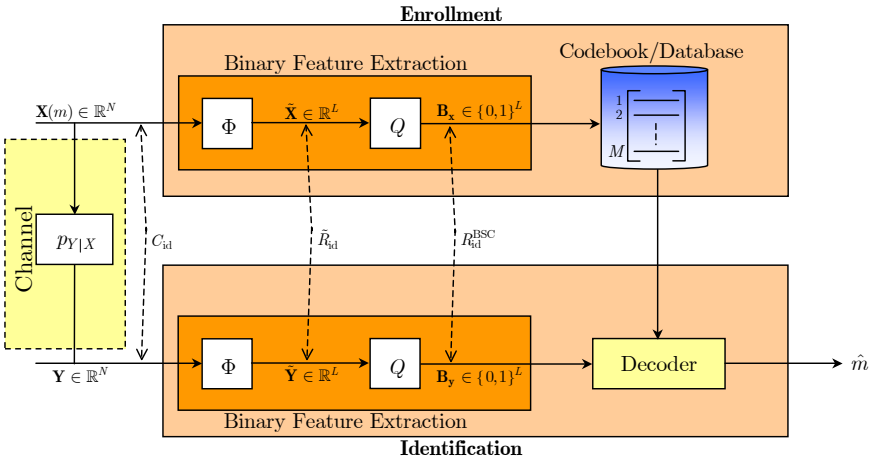


Figure 2. Robust hashing based identification with the corresponding identification capacities at each stage of processing.

2.1. Direct domain identification capacity

To estimate the maximum achievable errorless identification rate for the above channel $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, we will use the notion of *identification capacity* C_{id} defines as:

$$C_{\text{id}} = \frac{1}{N} I(\mathbf{X}; \mathbf{Y}), \tag{3}$$

where $I(\mathbf{X}; \mathbf{Y})$ is the mutual information between the input and output of channel[25] †. Accordingly, the maximum number of reliably distinguishable sequences on the output of such a channel is limited by $2^{NC_{\text{id}}}$. Since, the identification capacity (3) is upper bounded by:

$$C_{\text{id}} = \frac{1}{N}(H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y})) \leq \frac{1}{N} H(\mathbf{X}), \tag{4}$$

and for the case of i.i.d. sequences it means that about $2^{NH(X|Y)}$ sequences will not be distinguished due to the channel distortions. Therefore, the channel loss $H(X|Y)$ plays an essential role for the identification that will be addressed in our further analysis.

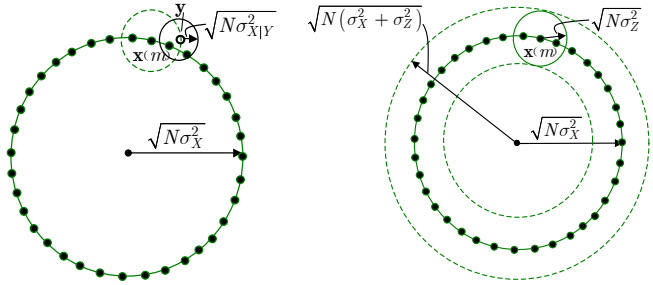
We demonstrate it on the example of i.i.d. Gaussian statistics assuming $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$. The differential entropy of this source is $h(X) = \frac{1}{2}\log_2(2\pi e \sigma_X^2)$. For the memoryless additive white Gaussian noise (AWGN) channel $\mathbf{y} = \mathbf{x}(m) + \mathbf{z}$ with $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbf{I}_N)$, the identification capacity is[5]:

$$C_{\text{id}} = \frac{1}{2}\log_2 \frac{1}{1 - \rho_{XY}^2} = \frac{1}{2}\log_2\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right), \tag{5}$$

†Note the difference with the channel capacity $C = \max_{p_X(x)} I(X;Y)$, where the maximization is performed with respect to the channel input distribution that is more flexible contrarily to the considered fixed input distribution identification setup.

where $\rho^2_{XY} = \frac{\sigma^2_X}{\sigma^2_X+\sigma^2_Z}$ is a squared correlation coefficient between X and Y. Here, $h(X|Y) = \frac{1}{2}\log_2(2\pi e\sigma^2_{X|Y})$ represents the conditional entropy with $\sigma^2_{X|Y} = \frac{\sigma^2_X\sigma^2_Z}{\sigma^2_X+\sigma^2_Z}$. This variance also corresponds to the variance of minimum mean square error (MMSE) estimator of $\mathbf{x}$ based on $\mathbf{y}$.

This situation can be presented graphically, if one assumes that N is sufficiently large for $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2_X\mathbf{I}_N)$. All the realizations $\mathbf{x}$ will be almost uniformly distributed on the surface of sphere of radius $\sqrt{N\sigma^2_X}$ with the probability close to one.[5] For the above AWGN channel, the noise will create a sphere of ambiguity around the communicated sequence $\mathbf{x}(m)$ with the radius $\sqrt{N\sigma^2_Z}$ that is schematically shown in Figure 3(a). The noisy realizations $\mathbf{y}$ with be located on the surface of this sphere. Under the proper codebook construction, the MMSE estimate should ensure the presence of a unique $\mathbf{x}(m)$ in the sphere of radius $\sqrt{N\sigma^2_{X|Y}}$ around $\mathbf{y}$. In the classical digital communications, this requirement is easily met by constraining the rate of source using optimal source coding, which minimizes source reconstruction distortion for a given rate, and by selecting a proper codebook of channel code for the specified statistics of energy constrained channel [‡]. In the robust hashing, the situation is inherently different due to inability to control the rate of source that is defined by nature. Moreover, the source output serves directly as an input to the channel, whose distribution is not necessarily optimal for specified channel statistics. Similar situation can be also considered for the sphere packing counterpart of the above considered coding framework that is presented in Figure 3(b) and corresponds to another interpretation of (3) in the form of $I(X;Y) = h(Y) - h(Y|X)$. In this case, the restriction is coming from the number of spheres of radius $\sqrt{N\sigma^2_Z}$ concentrated around all possible codewords that can be packed into the sphere of radius $\sqrt{N(\sigma^2_X+\sigma^2_Z)}$. Obviously, under the above conditions many codewords will be located within the sphere of ambiguity and can not be distinguished.



(a) (b)

Figure 3. The origin of identification ambiguity in robust hashing for the Gaussian setup: (a) coding with MMSE estimate-base decoder and (b) sphere packing.

2.2. Average probability of error

We will continue the analysis of theoretical limits of robust hashing algorithms by a performance analysis of robust hashing in terms of average probability of error. First, we will provide the analysis in the direct domain that goes along the line of classical communications and then extend it to the random projections domains.

The average probability of error is defined as:

$$P_e = \frac{1}{M}\sum_{m=1}^{M}\Pr[\hat{m} \neq m|M = m] = \frac{1}{M}\sum_{m=1}^{M}P_{e|\mathbf{x}(m)}, \tag{6}$$

i.e., that the decoded index $\hat{m}$ is not equal to the true index m, with $P_{e|\mathbf{x}(m)}$ to be the probability of error for the codeword $\mathbf{x}(m)$, which is computed according to:

[‡]It will be shown below that it is also equivalent to satisfying the proper minimum distance between codewords in the codebook.

$$P_{e|\mathbf{x}(m)} = \int_{\mathcal{R}_m^c} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}(m))d\mathbf{y}, \tag{7}$$

where $\mathcal{R}_m^c$ is the complementary decision region for the codeword $\mathbf{x}(m)$. This decision region is defined as:

$$\mathcal{R}_m^c = \bigcup_{n=1,n\neq m}^{M} D_{m,n}, \tag{8}$$

where the sets $D_{m,n} = \{\mathbf{y} : p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}(m)) \leq p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}(n))\}$, $m,n = \{1,\cdots,M\}$ $m \neq n$ correspond to the *maximum likelihood* (ML) decision region for the case of two codewords. We will also define the corresponding pairwise error probability of falsely accepting $\mathbf{x}(n)$ instead of $\mathbf{x}(m)$ as $P_{e|[\mathbf{x}(m)\to\mathbf{x}(n)]}$.

The exact average probability of error can be computed as[19]:

$$P_e = 1 - \int_{-\infty}^{+\infty} \left(1 - Q\left(\frac{t+\frac{1}{2}\epsilon_{\mathbf{x}}}{\sqrt{\sigma_Z^2\epsilon_{\mathbf{x}}}}\right)\right)^{M-1} \frac{1}{\sqrt{2\pi\sigma_Z^2\epsilon_{\mathbf{x}}}} \exp\left[-\frac{1}{2\sigma_Z^2\epsilon_{\mathbf{x}}}\left(t-\frac{1}{2}\epsilon_{\mathbf{x}}\right)^2\right] dt, \tag{9}$$

where $\epsilon_{\mathbf{x}} = \|\mathbf{x}\|^2$.

It is also useful to introduce the bounds on the above probability of error and investigate the identification performance in terms of minimum distance among the codewords for a given codebook. In particular, using the maximum pairwise probability and union bound, one can show that the probability of error $P_{e|\mathbf{x}(m)}$ in (6) can be bounded as:

$$\max_{n,n\neq m} P_{e|[\mathbf{x}(m)\to\mathbf{x}(n)]} \leq P_{e|\mathbf{x}(m)} \leq \sum_{n=1,n\neq m}^{M} P_{e|[\mathbf{x}(m)\to\mathbf{x}(n)]}. \tag{10}$$

For the AWGN channel, the pairwise error probability can be computed as[19]:

$$P_{e|[\mathbf{x}(m)\to\mathbf{x}(n)]} = Q\left(\frac{d_{m,n}}{2\sigma_Z}\right), \tag{11}$$

where $d_{m,n} = \|\mathbf{x}(m) - \mathbf{x}(n)\|$ is the distance between two codewords $\mathbf{x}(m)$ and $\mathbf{x}(n)$.

Assuming the symmetric construction of the codebook and the worst case distance among all codewords to be $d_{min} = \min_{m\neq n}\|\mathbf{x}(m) - \mathbf{x}(n)\|$, one can combine (11) with (10) and substitute them to (6) to introduce the bounds on the average probability of error:

$$Q\left(\frac{d_{min}}{2\sigma_Z}\right) \leq P_e \leq \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1,n\neq m}^{M} Q\left(\frac{d_{m,n}}{2\sigma_Z}\right). \tag{12}$$

Assuming $Q\left(\frac{d_{m,n}}{2\sigma_Z}\right) \leq Q\left(\frac{d_{min}}{2\sigma_Z}\right)$, the bounds on the average probability of error finally can be reduced to:

$$Q\left(\frac{d_{min}}{2\sigma_Z}\right) \leq P_e \leq (M-1)Q\left(\frac{d_{min}}{2\sigma_Z}\right). \tag{13}$$

Therefore, the average probability of error is determined by the worst case distance. Similarly to the analysis presented in the previous section for the achievable identification rate, there is no possibility to control and actually maximize $\|\mathbf{x}\|^2$ in (9) or d_{min} in (13) for the fixed M and N by the best codebook construction as in the digital communication since all codewords are generated by the source (natural randomness) with the statistical distribution defined by nature and their number is only limited by the source entropy.

3. IDENTIFICATION RATE AND PROBABILITY OF ERROR AFTER BINARY FEATURE EXTRACTION

In this section, we will consider the setup presented in Figure 2 using key-dependent mapping as defined in (1):

$$\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}. \tag{14}$$

We will try to introduce a formal approach to the random projections considered by Fridrich,[7] Mihcak *et. al.*,[18] F. Lefebvre and B. Macq.[12] Instead of following a particular consideration of mapping $\mathbf{W}$, we will assume that $\mathbf{W}$ is a random matrix. The matrix $\mathbf{W}$ has the elements $w_{i,j}$ that are generated from some specified distribution known as a random projections. $L \times N$ random matrices $\mathbf{W}$ whose entries $w_{i,j}$ are independent realizations of Gaussian random variables $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ represent a particular interest for our study. In this case, such a matrix can be considered as an *orthoprojector*, for which $\mathbf{W}\mathbf{W}^T \approx \mathbf{I}_L$ (almost orthogonal) [§].

The performed above analysis of identification capacity and probability of error concerns the so-called *direct domain* where the decision about the codeword index is deduced directly using the N-length sequences. Contrarily, practically all robust hashing algorithms use the above dimensionality reduction (1) thus converting all pairs of sequences $(\mathbf{x}, \mathbf{y})$ into the projections $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. In this section, we will demonstrate that the mutual information between the projected sequences is now redefined as $\tilde{R}_{\text{id}} = \frac{L}{N} I(\tilde{X}; \tilde{Y})$ and $\tilde{R}_{\text{id}} \leq C_{\text{id}}$, with the equality if and only if the transformation is invertible. Since it is not obviously a case for most of dimensionality reduction or feature extraction transforms where $L << N$, one is additionally facing the loss in performance in terms of maximum amount of uniquely distinguishable sequences. It should be pointed out that an extra loss is coming from the binarization stage that results in $R_{\text{id}}^{\text{BSC}}$ for the binary sequence representation.

3.1. Identification rate and probability of error in random projection domain

According to the definition of identification capacity (3), we will consider the transformation of random vectors $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$ and $\tilde{\mathbf{Y}} = \mathbf{W}\mathbf{Y}$ that results into the identification rate:

$$\tilde{R}_{\text{id}} = \frac{1}{N} I(\tilde{\mathbf{X}}; \tilde{\mathbf{Y}}). \tag{15}$$

For the Gaussian assumptions considered above $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_L)$ and $\tilde{\mathbf{Y}} \sim \mathcal{N}(\mathbf{0}, (\sigma_X^2 + \sigma_Z^2)\mathbf{I}_L)$, the identification rate in the random projections domain is:

$$\tilde{C}_{\text{id}} = \frac{1}{N} \frac{L}{2} \log_2 \left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) = \frac{L}{N} C_{\text{id}}. \tag{16}$$

Therefore, the dimensionality reduction transform $\mathbf{W}$ based on orthoprojector introduces a loss proportional to the ratio of signal dimensions after and before projection, i.e., $\frac{L}{N}$.

The average probability of error (9) will be the same with the only replacement of norm $\epsilon_{\mathbf{x}} = \|\mathbf{x}\|^2$ by $\tilde{\epsilon}_{\mathbf{x}} = \|\tilde{\mathbf{x}}\|^2 = \|\mathbf{W}\mathbf{x}\|^2$. The bounds on the average probability of error in the direct domain (13) can be readily rewritten for the random projections domain as:

$$Q\left(\frac{\tilde{d}_{min}}{2\sigma_Z}\right) \leq \tilde{P}_e \leq (M-1) Q\left(\frac{\tilde{d}_{min}}{2\sigma_Z}\right), \tag{17}$$

where $\tilde{d}_{min}^2 = \min_{m \neq n} (\mathbf{x}(m) - \mathbf{x}(n))^T \mathbf{W}^T \left(\mathbf{W}\mathbf{W}^T\right)^{-1} \mathbf{W} (\mathbf{x}(m) - \mathbf{x}(n))$ is the worst case squared distance among all pairs of codewords. Moreover, in the case of orthoprojection ($\mathbf{W}\mathbf{W}^T = \mathbf{I}_L$), it reduces to $\tilde{d}_{min}^2 = \min_{m \neq n} (\mathbf{x}(m) - \mathbf{x}(n))^T \mathbf{W}^T \mathbf{W} (\mathbf{x}(m) - \mathbf{x}(n)) = \min_{m \neq n} \|\mathbf{W}(\mathbf{x}(m) - \mathbf{x}(n))\|^2$.

Therefore, it is important to point out that the distance between the codewords in the random projections domain has reduced from d_{min} to $\tilde{d}_{min}$, and $\tilde{d}_{min} \leq d_{min}$.

To introduce the bounds on the norm $\|\tilde{\mathbf{x}}\|^2$ and distance $\tilde{d}^2$ we will use the results of Johnson-Lindenstrauss lemma,[10] which states that with high probability the geometry of a point cloud is not disturbed by certain Lipschitz mappings onto a space of dimension logarithmic in the number of points. In particular, some existing proofs of the lemma show that the mapping $\mathbf{W}$ can be taken as a linear mapping represented by an $L \times N$ matrix whose entries are randomly drawn from certain probability distributions. More particularly, M vectors in the Euclidean space can be projected down to $L = O(\zeta^{-2} \log_2 M)$ dimensions while incurring a distortion of at most $1 + \zeta$ in their pairwise distances, where $0 < \zeta < 1$. In principle, this can be achieved by a dense $L \times N$ matrix and such a mapping takes $O(N \log_2 M)$ (for fixed ζ). We refer interested readers to[1] for more details.

[§]Otherwise, one can apply special orthogonalization techniques to ensure perfect orthogonality.

According to Johnson-Lindenstrauss result,[10] one can use use the approximation for the random orthoprojector $\mathbf{W}$ as:

$$(1-\zeta)\sqrt{\frac{L}{N}}||\mathbf{x}|| \leq ||\mathbf{W}\mathbf{x}|| \leq (1+\zeta)\sqrt{\frac{L}{N}}||\mathbf{x}||. \tag{18}$$

Thus, with high probability one can approximate (17) as follows:

$$Q\left(\sqrt{\frac{L}{N}}\frac{d_{min}}{2\sigma_Z}\right) \leq \tilde{P}_e \leq (M-1)Q\left(\sqrt{\frac{L}{N}}\frac{d_{min}}{2\sigma_Z}\right). \tag{19}$$

Therefore, the random projections introduce the loss in the norm of projected codewords and distance between them proportional to $\sqrt{\frac{L}{N}}$ that also reflects the corresponding loss in terms of achievable identification rate.

3.2. Identification rate and probability of error after binarization

The next step in binary feature extraction according to Figure 2 corresponds to the binarization. In this section, we will consider binarization based on sign random projections introduced by (2). The link between the binary representation $\mathbf{b_x}$ of vector $\mathbf{x}$ and its noisy counterpart $\mathbf{b_y}$ of vector $\mathbf{y}$ is defined according to *binary symmetric channel* (BSC) model. It is assumed that noise in the direct domain might cause a bit flipping in the binary domain with a certain average probability $\bar{P}_b$. The corresponding identification rate can be readily found as[5]:

$$R_{\text{id}}^{\text{BSC}} = \frac{1}{N}I(\mathbf{B_x};\mathbf{B_y}) = \frac{L}{N}(1-H_2(\bar{P}_b)), \tag{20}$$

where $H_2(\bar{P}_b) = -\bar{P}_b\log_2\bar{P}_b - (1-\bar{P}_b)\log_2(1-\bar{P}_b)$ is the binary entropy.

The bit error probability indicates the mismatch of signs between $\tilde{x}_i$ and $\tilde{y}_i$, i.e., $\Pr[sign(\tilde{x}_i) \neq sign(\tilde{y}_i)]$. For a given vector $\mathbf{x}$ and defined projection vector $\mathbf{w}_i$, one can find the probability of bit error as:

$$P_{b|\tilde{x}_i} = \frac{1}{2}\Pr[\tilde{Y}_i \geq 0|\tilde{X}_i < 0] + \frac{1}{2}\Pr[\tilde{Y}_i < 0|\tilde{X}_i \geq 0], \tag{21}$$

or by symmetry as:

$$P_{b|\tilde{x}_i} = \Pr[\tilde{Y}_i < 0|\tilde{X}_i \geq 0]. \tag{22}$$

For a given $\tilde{x}_i$ and Gaussian assumption about the noise, the distribution of the projected vector is $\tilde{Y}_i \sim \mathcal{N}(\tilde{x}_i, \sigma_Z^2\mathbf{w}_i^T\mathbf{w}_i)$ that reduces to $\tilde{Y}_i \sim \mathcal{N}(\tilde{x}_i, \sigma_Z^2)$ for the orthoprojection case ($\mathbf{w}_i^T\mathbf{w}_i = 1$) and:

$$P_{b|\tilde{x}_i} = \int_{-\infty}^{0}\frac{1}{\sqrt{2\pi\sigma_Z^2}}e^{\frac{-(\tilde{y}_i-\tilde{x}_i)^2}{2\sigma_Z^2}}d\tilde{y}_i = Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right). \tag{23}$$

Thus, the average bit error probability is:

$$\bar{P}_b = 2\int_0^\infty P_{b|\tilde{x}_i}p(\tilde{x}_i)d\tilde{x}_i = 2\int_0^\infty Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right)\frac{1}{\sqrt{2\pi\sigma_X^2}}e^{\frac{-\tilde{x}_i^2}{2\sigma_X^2}}d\tilde{x}_i = \frac{1}{\pi}\arccos(\rho_{XY}), \tag{24}$$

with the statistics of projection $\tilde{X}_i \sim \mathcal{N}(0, \sigma_X^2)$. Remarkably, the average probability of error depends on the correlation coefficient between the direct domain data and is completely determined by the channel and source statistics. Similar result is also confirmed by McCarthy *et. al.*[17] and Doets and Lagendijk[6] for audio hashing based on binary fingerprinting method proposed in.[8]

Summarizing the above consideration, one can conclude that the identification rate deceases with each stage of processing $C_{\text{id}} \geq \tilde{R}_{\text{id}} \geq R_{\text{id}}^{\text{BSC}}$ that is obviously a source of serious restriction. To avoid such kind of ambiguity, it is well known according to the Shannon channel coding theorem that the number of codewords M should be restricted to $M \leq 2^{NC_{\text{id}}}$.[5] Therefore, there does not exist any robust hashing algorithm in the considered decoding sense capable to reliably distinguish more codewords than it is allowed by the identification capacity. The same conclusions are valid for the average probability of error.

Therefore, there is a high demand in solutions capable to resolve the above problem. We envision two possible solutions:

- Solution 1: **list decoding** when more than one index is allowed to be produced by a system; in this case, the task is to ensure that the correct index is always on a list, i.e., to minimize a probability of miss while controlling a list size;
- Solution 2: **side information assisted hashing** when the decoder produces only one index but it has an access to some side information that resolve the ambiguity created by the channel; in this paper, we will concentrate on this solution.

4. SIDE INFORMATION ASSISTED ROBUST PERCEPTUAL HASHING

As it was shown above for both identification capacity and average probability of error, the only way to decrease the probability of error for a single candidate decoder is to reduce the number of codewords in the codebook (or actually to expurgate the codewords with the smallest distance up to the limit defined by the identification capacity). This approach, typically used in the digital communications, is not acceptable for the content identification/authentication due to the main requirement to avoid collisions for $M > 2^{NC_{\text{id}}}$.

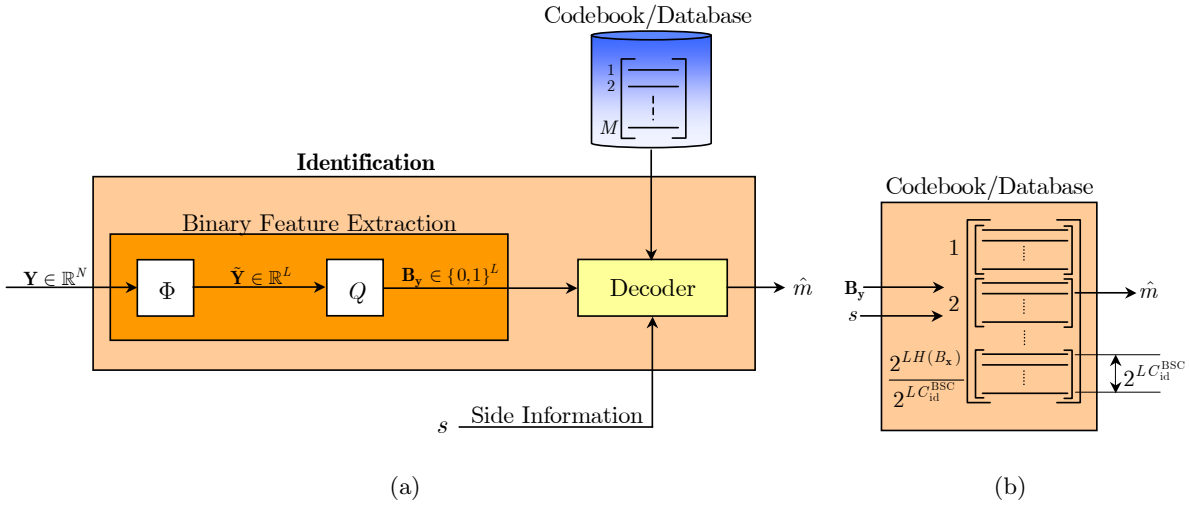


Figure 4. Side information based robust hashing: (a) identification system with the side information assisted robust hashing and (b) codebook construction and decoding based on binning.

That is why, we advocate an alternative approach based on side information for collision-free robust hashing (Solution 2 above). This approach is essentially inspired by Slepian-Wolf distributed source coding[21] and schematically presented for the binary data in Figure 4. The basic idea behind this approach consists in the partitioning (binning) of the entire codebook on the sets (bins) indexed by s. The bin index s, considered as the side information, is communicated to the decoder, which makes the decision about a particular sequence index m matched with the binary data $\mathbf{B_y}$ within the bin s. For the binary data $1 \le s \le 2^{L(H(B_{\mathbf{x}})-R_{\text{id}}^{\text{BSC}})}$ and $1 \le m \le 2^{LR_{\text{id}}^{\text{BSC}}}$ and the corresponding codebook design is shown in Figure 4(b). The efficient practical implementation of this theoretical decoding can be achieved using low-density parity check (LDPC) codes.

Similar interpretation can be introduced for the Gaussian data (considered in Section 2) that is shown in Figure 5. The codewords shown in the upper part of Figure can not be uniquely distinguished in the presence of noise. However, one can easily distinguish all of them (given that the number of codewords does not exceed $2^{NC_{\text{id}}}$) by knowing that the codewords belong to the certain set indexed by s. To achieve this goal, one can label the codewords in several groups (in this particular example 3) with the index s that are represented by triangles, circles and squares. Given the index s, one performs the identification by the ML decoding in the direct or random projections domains.

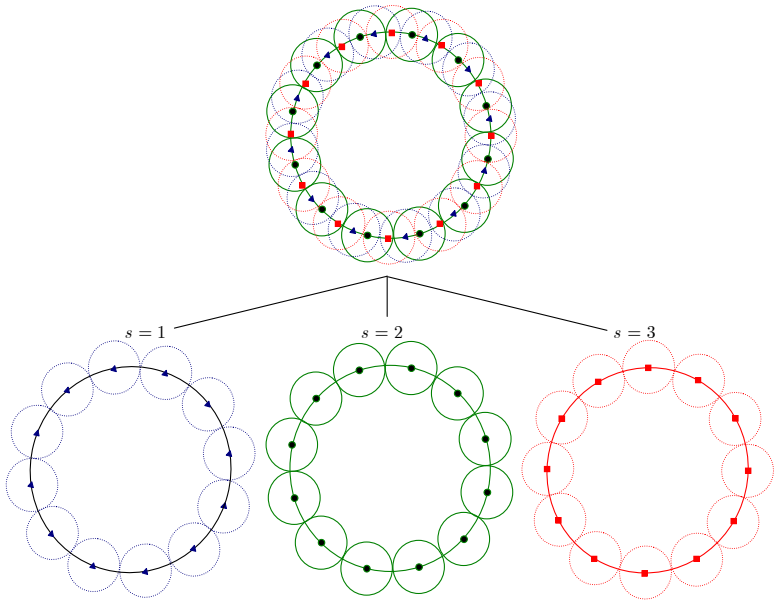


Figure 5. Side information based robust hashing: the codebook is partitioned on 3 sets labeled by triangles, circles and rectangles.

Summarizing the above consideration, one needs to compensate $H(X|Y)$ loss in $I(X;Y) = H(X) - H(X|Y)$ by providing this side information via some auxiliary channel to the decoder. The main question is how to provide this side information. In our belief, it is highly determined by a particular application and here we will provide only some examples:

1. Content identification:

   - *Content based image retrieval* (CBIR) systems: the codebook construction can be considered based on binning with the index s provided in the form of text annotations about the groups or categories such as for example, people, trees, animals, vehicles, nature, etc.; it is important to note that the clustering, typically required by the existing CBIR systems (that naturally disappears with the growth of the codebook cardinality), is not needed in the advocated approach; the samples from different groups can be mixed and overlapped in the direct space; additionally, the system can provide the list of indexes on the output contrarily to strict identification; other modalities can be used as the side information;
   - *DNA/RNA sequence and mass spectrometry protein identification*: the entire database is partitioned into groups according to the domain specific classification;
   - *Brand protection (item identification, tracking and tracing)*: the item identification features (e.g., microstructure images) are partitioned into the bins according to some classification information such as for example serial number, date of production, country of destination, etc.;
   - *Multimedia identification*: image, audio, video content identification is based on the robust hashing but the index of set s is communicated to the decoder as the hidden watermark; it should be pointed out the main difference with the known schemes such as for example proposed by Fridrich[7] where the hash is stored as a watermark as opposed to what is considered in the paper;

2. Content authentication:

   - Authentication architectures are designed in the similar way as identification with the only difference that consists in the providing additional authentication information to the decision device about the sequence index m within the set specified by the index s. In the authentication applications it is also known as *common randomness extraction*[2] and practically used in various architectures as reported in.[9, 13, 15, 23]

5. RESULTS OF COMPUTER MODELING

In this section, we will first demonstrate the loss in performance due to the dimensionality reduction and binarization for both identification rate and average probability of error. In the second part of modeling, we will highlight the impact of side information on the average probability of error for the Gaussian and binarized data.

We will start with the demonstration of loss in performance introduced by the different signal processing transformations in the procedure of robust hash computation. For this, we present identification capacity C_{id} (3) and identification rates $\tilde{R}_{\text{id}}$ (15) and $R_{\text{id}}^{\text{BSC}}$ (20) for the dimensionality reduction factors $\frac{L}{N} = \frac{1}{8}$ and $\frac{1}{16}$ as the functions of signal-to-noise ratio (SNR) defined as $\text{SNR} = 10\log_{10}\frac{\sigma_X^2}{\sigma_Z^2}$. The resulting curves are shown in Figure 6(a). As expected, the achievable rate decreases with each stage of processing. The first rate loss from the direct domain to the random projections domain is caused by the dimensionality reduction and is proportional to $\frac{L}{N}$. The second rate loss is additionally caused by the mapping of real data into binary counterparts. It is evident that for the errorless identification one should satisfy the condition $M \leq 2^{NC_{\text{id}}}$ and since $C_{\text{id}} \geq \tilde{R}_{\text{id}} \geq R_{\text{id}}^{\text{BSC}}$ the number of uniquely distinguishable sequences is reducing with each stage of hash computation.

To demonstrate the same impact on the average probability of error, we investigated the average probability of error in the direct domain, after sequential dimensionality reduction and binarization. The results of this study are presented in Figure 6(b). At the first stage, the average probability of error for the direct domain was computed according to the exact formula (9), upper bound (12) and experimentally simulated by averaging over 10 randomly generated Gaussian codebooks of size $M = 2^3$ and length $N = 1024$ with $\sigma_X^2 = 1$ under 50000 random noise realizations. The average minimum distance was estimated to be $d_{min} = 43$. At second stage, the same simulations were performed for the randomly projected data with $L = 128$ that are shown in the same Figure. The average minimum distance was $\tilde{d}_{min} = 15$. Both theoretical exact formula and upper bound on the average probability of error as well as the experimental results confirm the conclusion about the distance decrease proportional to $\sqrt{L/N}$ that causes the corresponding deterioration of performance. At the final stage, the randomly projected data were binarized thus preserving its dimensionality $L = 128$ and the experimentally computed average probability of error is presented to characterize the identification performance based on binary (hard) hashed values. These results are in the good match with the corresponding performance in terms of achievable identification rate.

The last part of simulation addresses the impact of side information on the identification system performance in terms of average probability of error. The results of simulation are shown in Figures 6(c) and (d). Figure 6(c) represents the results obtained for the Gaussian codebooks of size $M = 2^{10}$ and length $N = 1024$ with $\sigma_X^2 = 1$ under 50000 random noise realizations. The curve with label "0 bit" corresponds to the case when no side information is used for the identification in both direct and binary domains, i.e., the rate of side information is $R_{\text{si}} = 0$, and the rest of curves indicate the graduate increase in the side information up to 6 bits. The codebook was partitioned onto $2^{R_{\text{si}}}$ bins during simulation with the random allocation of codewords. The same simulation was also performed for the binary codebooks obtained by the random projections and binarization with $L = 128$ of the the previous Gaussian codebooks. In both cases, the presence of side information makes possible to decrease the average probability of error with the SNR gain about 2.5 dB.

6. CONCLUSIONS

In this paper, we investigated the fundamental restrictions of basic robust perceptual hashing algorithms based on dimensionality reduction and binarization in the identifications setup. Based on the well-known communication results we have linked the identification setup with the identification capacity and demonstrated the maximum number of uniquely identifiable sequences without any processing. Then we showed the gradual decrease of this number with each stage of hash computation as well as presented the accuracy of identification in terms of average probability of error. Due to the inherent necessity to identify larger number of sequences than those enabled by the identification capacity and taking into account the above loss due to dimensionality reduction and binarization in the hash computation, we suggested two possible solutions based on the list decoding and side information assisted hashing that was the subject of our study. The results of computer simulation performed for both Gaussian random codebooks and binary hashes demonstrate the positive impact of side information on the performance enhancement. In part of future research, we will concentrate on the alternative approach based on list decoding and compare the results with those obtained in this paper.

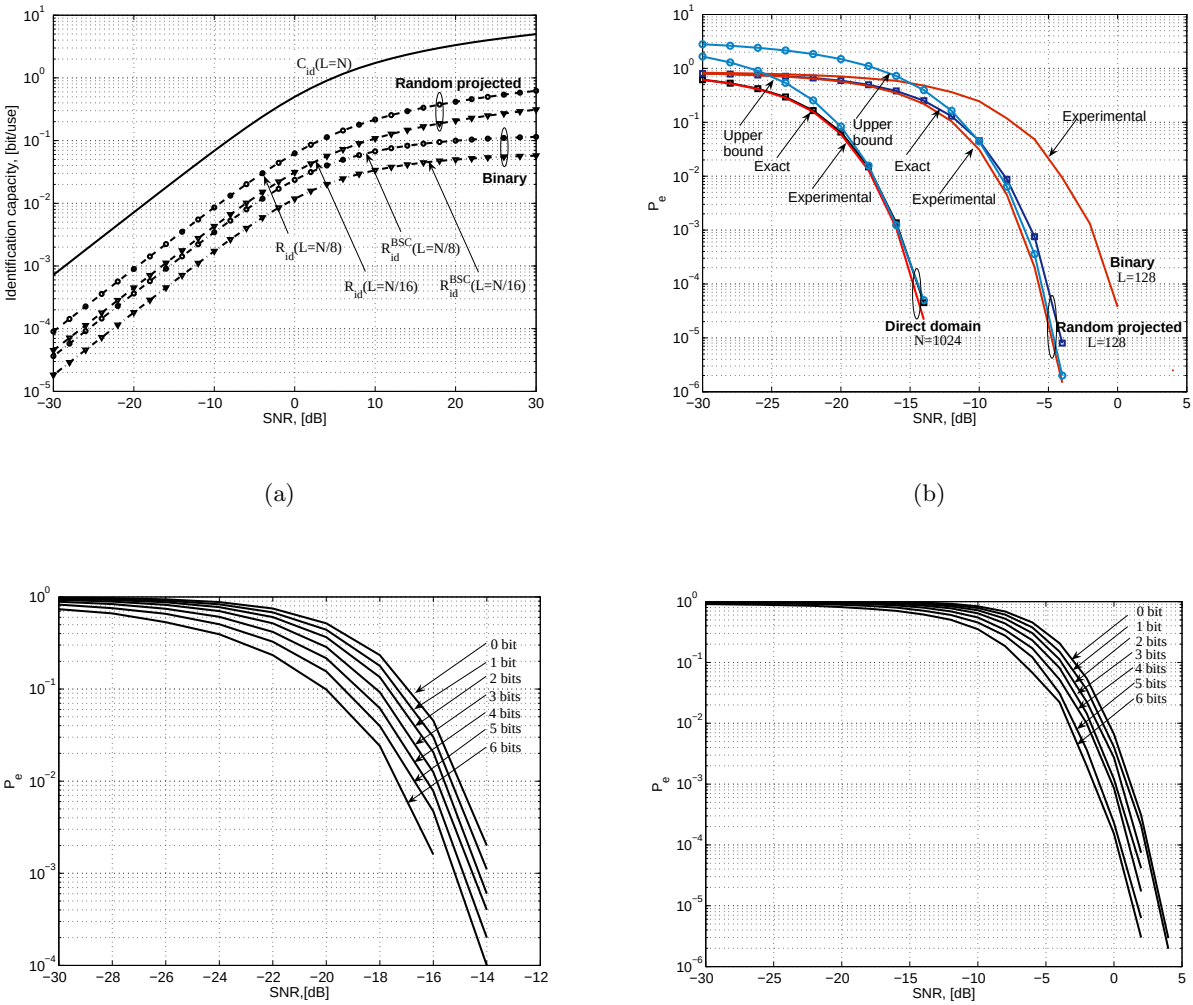


Figure 6. Side information based robust hashing: (a) achievable rates at different stages of hash computation, (b) the corresponding average probabilities of error for $M = 2^3$, (c) impact of side information on average probability of error in the direct domain ($N = 1024$) and (d) binary domain ($L = 128$) for $M = 2^{10}$.

7. ACKNOWLEDGMENT

This paper was partially supported by SNF Professorship grant 114613 and SNF projects 200021-111643, 200020-121635 and 200021-1119770.

REFERENCES

1. D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *JOURNAL OF COMPUTER AND SYSTEM SCIENCES*, 66(4):671–687, 2003.
2. R. Ahlswede and I. Csiszar. Common randomness in information theory and cryptography - Part I: secret sharing. *IEEE Trans. Inform. Theory*, 39(4):1121–1132, 1993.

3. Y. Altug, M. K. Mihcak, O. Ozyesil, and V. Monga. Reliable communications with asymmetric codebooks: An information theoretic analysis of robust signal hashing. *submitted to IEEE Transactions on Information Theory*, September 2008.
4. E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52(2):489–509, February 2006.
5. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, 1991.
6. P.J.O. Doets and R.L. Lagendijk. Distortion estimation in compressed music using only audio fingerprints. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):302–317, February 2008.
7. J. Fridrich. Robust bit extraction from images. In *Proceedings ICMCS'99*, volume 2, pages 536–540, Florence, Italy, June 1999.
8. J. Haitsma, T. Kalker, and J. Oostveen. Robust audio hashing for content identification. In *International Workshop on Content-Based Multimedia Indexing*, pages 117–125, Brescia, Italy, September 2001.
9. T. Ignatenko and F.M.J. Willems. On the security of xor-method in biometric authentication systems. In *The twenty-seventh symposium on Information Theory in the Benelux*, pages 197–204, Noordwijk, The Netherlands, June 8-9 2006.
10. W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into Hilbert space. *Contemporary Mathematics*, (26):189–206, 1984.
11. O. Koval, S. Voloshynovskiy, F. Beekhof, and T. Pun. Security analysis of robust perceptual hashing. In *Proceedings of SPIE-IS&T Electronic Imaging 2008, Security, Steganography, and Watermarking of Multimedia Contents X*, San Jose, USA, 26 Jan. – 1 Feb. 2008.
12. F. Lefebvre and B. Macq. Rash : RAdon Soft Hash algorithm. In *Proceedings of EUSIPCO - European Signal Processing Conference*, Toulouse, France, 2002.
13. Y.-C. Lin, D. Varodayan, and B. Girod. Image authentication based on distributed source coding. In *IEEE International Conference on Image Processing (ICIP2007)*, San Antonio, USA, September 2007.
14. Y. Mao and M. Wu. Unicity distance of robust image hashing. *IEEE Trans. on Information Forensics and Security*, 2(3):462–467, September 2007.
15. E. Martinian, S. Yekhanin, and J.S. Yedidia. Secure biometrics via syndromes. In *43rd Annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, USA, October 2005.
16. E. McCarthy, F. Balado, G. Silverstreand, and N. Hurley. A framework for soft hashing and its application to robust image hashing. In *In Procs. of the IEEE International Conference on Image Processing*, Singapore, Oct 2004.
17. E. McCarthy, F. Balado, G. Silvestre, and N. Hurley. A model for improving the performance of feature extraction based robust hashing. In *In Procs. of SPIE*, San Jose, CA, USA, Jan 2005.
18. M. K. Mihcak and R. Venkatesan. A perceptual audio hashing algorithm: A tool for robust audio identification and information hiding. In *Proceedings of 4th International Information Hiding Workshop*, Pittsburgh, PA, USA, April 2001.
19. J. G. Proakis. *Digital Communications*. McGraw-Hill, 1995.
20. S. Sarvotham, D. Baron, and R. Baraniuk. Sudocodes - fast measurement and reconstruction of sparse signals. In *EEE Int. Symposium on Information Theory (ISIT)*, Seattle, Washington, July 2006.
21. D. Slepian and J.K. Wolf. Noiseless encoding of correlated information sources. *IEEE Trans. Information Theory*, 19:471–480, July 1973.
22. A. Swaminathan, Y. Mao, and M. Wu. Security of feature extraction in image hashing. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'05)*, pages 1041–1044, Philadelphia, PA, March 2005.
23. P. Tuyls, B. Skoric, and T. Kevenaar (Eds.). *Security with Noisy Data: On Private Biometrics, Secure Key Storage and Anti-Counterfeiting*. Springer, 2007.
24. S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun. Robust perceptual hashing as classification problem: decision-theoretic and practical considerations. In *Proceedings of the IEEE 2007 International Workshop on Multimedia Signal Processing*, Chania, Crete, Greece, October 1–3 2007.
25. F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz. On the capacity of a biometrical identification system. In *Proc. 2003 IEEE Int. Symp. Inform. Theory*, page 82, Yokohama, Japan, June 29 - July 4 2003.