BINARY ROBUST HASHING BASED ON PROBABILISTIC BIT RELIABLITY

Sviatoslav Voloshynovskiy, Oleksiy Koval, Fokko Beekhof and Taras Holotyak

University of Geneva
Department of Computer Science
7 route de Drize, CH 1227, Geneva, Switzerland

ABSTRACT

In this paper, we consider robust hashing based on a bit reliability function that allows to enhance the performance in terms of both average probability of error and identification complexity. The obtained results demonstrate the high efficiency of the prosed approach.

1. INTRODUCTION

Robust hashing, a.k.a. as digital fingerprinting in some applications, 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 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 (identification, indexing and retrieval), security (tracking of illegal copies, verification of authenticity, anticounterfeiting), as well as assisting functionality for data synchronization.

The typical design of robust hashing consists of the dimensionality reduction and quantization or binarization that might be also followed by cryptographic encryption for the security enhancement. The performance analysis of robust hashing was mostly performed using computer simulation. Therefore, there is a real need for a thorough investigation of theoretical limits of robust hashing. The first efforts in this direction have been reported in [1] that mostly focused on the investigation of the average probability of error. The simultaneous impact of dimensionality reduction and binarization in terms of both identification rate and average probability of error was considered in [2].

A good robust hash should be of a sufficient length to ensure a relatively large minimum Hamming distance between the codewords stored in the database. However, searching for similar codewords or matching might be quite computationally expensive for long codewords. Therefore, the practical design of robust hashing is facing three main open issues: (a) relatively low performance in terms of average probability of bit error that requires the use of long codewords; (b) the complexity of searching in large scale systems; (c) prior ambiguity about the applied distortions. For example, a typical robust hash length L is about 3000-10000 bits and the database size M is about 0,5-2 Billions items.

In this paper, we will demonstrate that the common cause of the above problems is the data-independent or blind character of dimensionality reduction accomplished for the security reasons. The matrix of the dimensionality reduction transform is often generated from a secret key thus disregarding the statistical properties of the input data. That is why the dimensionality reduction as a feature extraction is performed blindly.

Therefore, in this paper, we select an alternative approach for the design of robust hashing with the overall goal of minimizing the average bit error probability and thus enhancing the identification accuracy while reducing the search complexity.

2. PROBLEM FORMULATION

In this paper, we will follow the information-theoretic approach allowing to estimate the maximum achievable number of uniquely distinguishable items. Considering robust hashing in identification applications, one can use the notion of *identification capacity* [3] for the evaluation of theoretical performance in terms of achievable rate. At the same time, we will evaluate the performance of a practical system according to an average probability of identification error P_e for a robust hash of a fixed length L and the search complexity, i.e., the maximum number of operations needed to establish the the data identity with the given P_e.

2.1. System performance: direct domain

Assuming that the data are independent or weakly dependent and can be treated as almost identically distributed, one can define the identification capacity as:

$$\bar{I}(X;Y) = \lim_{N\to\infty} E_{p(\mathbf{x},\mathbf{y})}[I_N], \tag{1}$$

The contact author is S. Voloshynovskiy (email: svolos@unige.ch). http://sip.unige.ch. This paper was partially supported by SNF projects 200021-111643 and 200021-1119770 and Swiss IM2 project.

where $I_N = \frac{1}{N}\log_2 \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}$ is the information density and $p(\mathbf{x},\mathbf{y})$, $p(\mathbf{x})$ and $p(\mathbf{y})$ are joint probability density of data stored in the database $\mathbf{X}$ and observed data $\mathbf{Y}$ and their marginals, respectively, with the length of data vectors N; provided that the limit is well defined.

In this case, the maximum number of classes that can be recognized with vanishing probability of error under the above conditions is bounded as:

$$M \leq 2^{N\bar{I}(X;Y)}. \tag{2}$$

The optimal decoder is a *maximum likelihood* (ML) decoder, which assumes a perfect knowledge of the observation model $p(\mathbf{y}|\mathbf{x}(m))$:

$$\hat{m} = \arg\max_{1\leq m\leq M} p(\tilde{\mathbf{y}}|\mathbf{x}(m)), \tag{3}$$

that is not always the case in practice.

For the case of i.i.d. Gaussian data $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2\mathbf{I}_N)$ and the memoryless additive white Gaussian observation model $\mathbf{y} = \mathbf{x} + \mathbf{z}$ with $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2\mathbf{I}_N)$, the recognition capacity is readily found as:

$$\bar{I}(X;Y) = \frac{1}{2}\log_2 \frac{1}{1-\rho_{XY}^2} = \frac{1}{2}\log_2\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right), \tag{4}$$

where $\rho_{XY}^2 = \frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}$ is the squared correlation coefficient between X and Y.

The ML decoder (3) is reduced to the minimum Euclidean distance decoder:

$$\hat{m} = \arg\max_{1\leq m\leq M} \left\|\tilde{\mathbf{y}} - \mathbf{x}(m)\right\|^2, \tag{5}$$

that requires $O(MN)$ operations for N-length vectors stored in the database of size M. Equivalently, the minimum Euclidean distance decoder can be reduced to the maximum cross-correlation decoder:

$$\hat{m} = \arg\max_{1\leq m\leq M} \mathbf{y}^T\mathbf{x}(m) - \frac{1}{2}\mathbf{x}(m)^T\mathbf{x}(m). \tag{6}$$

2.2. System performance: robust hash

To evaluate the achievable rate of an identification system based on robust hashing, we will consider a typical design of robust hashing algorithms based on a mapping of the original data $\mathbf{x}$ to some *secure* but at the same time *robust* domain. This step is accompanied by a dimensionality reduction:

$$\tilde{\mathbf{x}} = \mathbf{W}\mathbf{x}, \tag{7}$$

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 $1 \leq i \leq L$. 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. $L \times N$ random matrix $\mathbf{W}$ whose entries $w_{i,j}$ are independent realizations of Gaussian random variables $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ presents a particular interest for our study. In this case, such a matrix can be considered as an almost *orthoprojector*, for which $\mathbf{W}\mathbf{W}^T \approx \mathbf{I}_L$. [1]

The second step also uses a possibly key-dependent labeling or Grey codes to ensure closeness of labels for close vectors. The most simple quantization or binarization of extracted features is known as *sign random projections*:

$$b_{\mathbf{x}_i} = sign(\mathbf{w}_i^T\mathbf{x}), \tag{8}$$

where $b_{\mathbf{x}_i} \in \{-1, 1\}$, with $1 \leq i \leq L$ and $sign(a) = 1$, if $a \geq 0$ and -1, otherwise. The vector $\mathbf{b}_\mathbf{x} \in \{-1,1\}^L$ computed for all projections represents a binary hash from the vector $\mathbf{x}$. Since all projections are independent, it can be assumed that all bits in $\mathbf{b}_\mathbf{x}$ will be independent and equiprobable for the independent inputs.

Obviously, the hash computed from some distorted version $\mathbf{y}$ of $\mathbf{x}$ denoted as $\mathbf{b}_\mathbf{y}$ might contain some bits different from those in $\mathbf{b}_\mathbf{x}$. Therefore, the link between the binary representation $\mathbf{b}_\mathbf{x}$ of vector $\mathbf{x}$ and its noisy counterpart $\mathbf{b}_\mathbf{y}$ of vector $\mathbf{y}$ is defined according to a *binary symmetric channel* (BSC) model with a certain average probability $\bar{P}_b$.

The corresponding maximum number of recognizable classes (2) can be now estimated as:

$$M_b \leq 2^{L\bar{I}(B_\mathbf{x};B_\mathbf{y})}, \tag{9}$$

with $\bar{I}(B_\mathbf{x};B_\mathbf{y}) = H(B_\mathbf{x}) - H(B_\mathbf{x}|B_\mathbf{y})$.

It can be noticed that to maximize the M_b, one needs to maximize $I(B_\mathbf{x};B_\mathbf{y})$ for a given L. That can be achieved by: (a) maximization of $H(B_\mathbf{x})$ and (b) minimization of $H(B_\mathbf{x}|B_\mathbf{y})$. In the considered binary case, the maximum value of $H(B_\mathbf{x})$ is 1 that can be achieved for the equiprobable independent data, i.e., $P_{B_X}(-1) = P_{B_X}(1) = 0.5$.

The second term $H(B_\mathbf{x}|B_\mathbf{y})$ is defined by the average error probability of binary classification $\bar{P}_b$ and $H(B_\mathbf{x}|B_\mathbf{y}) = H_2(\bar{P}_b) = -\bar{P}_b\log_2\bar{P}_b - (1-\bar{P}_b)\log_2(1-\bar{P}_b)$ that is the binary entropy. In the considered setup with a blind fixed matrix $\mathbf{W}$ it is not possible to control $\bar{P}_b$. Therefore, we will consider an alternative design where $\bar{P}_b$ can be considerably reduced due to basis adaptation based on bit reliability.

The ML decoder for this setup is a minimum Hamming distance decoder:

$$\hat{m} = \arg\min_{1\leq m\leq M} \sum_{i=1}^{L} d^H(b_{\mathbf{x}_i}(m), b_{\mathbf{y}_i}), \tag{10}$$

where $d^H(.,.)$ stands for the Hamming distance. The complexity of this decoder is reduced to $O(ML)$. Obviously, selecting a large L one can approach the performance of the ML decoder in the direct domain but that would require higher complexity. Therefore, practical systems should ideally benefit from small L without a decrease of performance.

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

3. PROPOSED APPROACH

In this paper, we propose an alternative approach that would still require $O(ML)$ operations but is characterized by a smaller bit error probability. To introduce this approach we will first present the concept of bit reliability.

3.1. Bit error reliability

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 $\mathbf{x}$ and $\mathbf{w}_i$, the probability of bit error is:

$$P_{b|\tilde{x}_i} = \frac{1}{2}(\Pr[\tilde{Y}_i \geq 0|\tilde{X}_i < 0] + \Pr[\tilde{Y}_i < 0|\tilde{X}_i \geq 0]), \quad (11)$$

or by symmetry as:

$$P_{b|\tilde{x}_i} = \Pr[\tilde{Y}_i < 0|\tilde{X}_i \geq 0]. \quad (12)$$

For a given $\tilde{x}_i$ and Gaussian noise[2], 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 orthoprojector ($\mathbf{w}_i^T \mathbf{w}_i = 1$) and:

$$P_{b|\tilde{x}_i} = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{\frac{-(\tilde{y}_i - \tilde{x}_i)^2}{2\sigma_Z^2}} d\tilde{y}_i = Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right). \quad (13)$$

The origin of $P_{b|\tilde{x}_i}$ can be explained considering the mutual configuration of $\mathbf{x}$ and $\mathbf{w}_i$. The vector $\mathbf{x}$ forms an angle θ_{XW_i} with the basis vector $\mathbf{w}_i$ and the projection results into a scalar value $\tilde{x}_i$. The closer the angle θ_{XW_i} is to $\pi/2$, the smaller the value od $\tilde{x}_i$. This leads to the larger probability that the sign of $\tilde{y}_i$ will be different from the sign of $\tilde{x}_i$. One can immediately note that since the projections are generated at random there is generally no guaranty that two vectors can be collinear. However, at the same time some of the projections might form angles with $\mathbf{x}$ that deviate from $\pi/2$ thus leading to a smaller bit error probability. This observation makes it possible to assume that some projections can be more preferable than others and the equation (13) can be a good measure of bit *reliability*.

The above analysis only refers to a single realization of $\mathbf{x}$. Since $\mathbf{X}$ is a random vector following some distribution $p(\mathbf{x})$, one should find the average probability of error for all possible realizations. Assuming $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_N)$, the statistics of the data in the projected domain are $\tilde{X}_i \sim \mathcal{N}(0, \sigma_X^2)$ and:

$$\bar{P}_b = 2\int_{0}^{\infty} P_{b|\tilde{x}_i} p(\tilde{x}_i) d\tilde{x}_i \quad (14)$$

$$= 2\int_{0}^{\infty} Q\left(\frac{\tilde{x}_i}{\sigma_Z}\right) \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-\tilde{x}_i^2}{2\sigma_X^2}} d\tilde{x}_i = \frac{1}{\pi}\arccos(\rho_{XY}). \quad (15)$$

It should be noticed that all possible values $\tilde{x}_i$ in (14) originating from both "unreliable", i.e., values close to zero, and "reliable", i.e., values far away from zero, projections are taken into account with the same weight to form the resulting binary vector $\mathbf{b_x}$. Obviously, for a given set of enrollment data one can always find a set of vectors $\mathbf{w}_i$, $1 \leq i \leq L$ minimizing the overall bit error probability. However, keeping in mind the facts that the number of classes might be in the order of millions and constantly updated such an optimization problem looks highly unfeasible. Therefore, in the scope of this paper we will consider another approach when one generates an *overcomplete set of projections* J and select among them only those L projections that are the largest in the absolute magnitude, if the fixed number of bits L is requested by the complexity concerns.

3.2. Reliable bits in cross-correlation measure

The data coefficients and noise statistics in the projected domain will follow Gaussian distributions due to the selection of Gaussian basis vectors and the central limit theorem. Under this condition, the considered maximum cross-correlation decoder (6) will be optimal. Therefore, following the proposed strategy of overcomplete representation with J projections, it can be rewritten as:

$$\hat{m} = \arg\max_{1\leq m\leq M} \sum_{j=1}^{J} (\tilde{y}'_j \tilde{x}'_j(m) - \tilde{x}'^2_j(m)), \quad (16)$$

where $'$ denotes the ascending ordering according to the magnitude of $|\tilde{y}_j|$.

According to the above introduced concept of bit reliability, only reliable projections will have the largest magnitude and lowest probability of sign flipping. One can readily demonstrate that these components largely contribute in the sum (16). Therefore, we will approximate this sum by its L largest components:

$$\hat{m} = \arg\max_{1\leq m\leq M} \sum_{j=J-L}^{J} (\tilde{y}'_j \tilde{x}'_j(m) - \tilde{x}'^2_j(m)). \quad (17)$$

We will also assume that all projected vectors have the same norm $\|x(m)\|^2$ and thus we will skip the second term for the simplicity of further consideration. Furthermore, using a representation $\tilde{x} = sign(\tilde{x})\,|\tilde{x}| = b_{\mathbf{x}}\,|\tilde{x}|$, (17) yields for the L largest components:

$$\hat{m} = \arg\max_{1\leq m\leq M} \sum_{i=1}^{L} b_{\mathbf{y}_i}\,|\tilde{y}_i|\, b_{\mathbf{x}_i}(m)\,|\tilde{x}_i(m)|\,. \quad (18)$$

4. PRACTICAL DECODER

The practical implementation of identification based on robust hashes assumes that only binary templates $\mathbf{b_x}$ are stored

[2] In the case of assumed Gaussian random basis vectors $\mathbf{w}_i$ any distribution will be mapped into Gaussian one for both entry and noisy data.

in the database. Therefore, there are several possible practical implementations of the proposed framework: (a) based on soft information when $|\tilde{y}_j|$ is taken into account:

$$\hat{m} = \arg \max_{1 \leq m \leq M} \sum_{i=1}^{L} b_{\mathbf{y}_i} b_{\mathbf{x}_i}(m) \, |\tilde{y}_j| \, , \tag{19}$$

or (b) based on a hard decoder:

$$\hat{m} = \arg \max_{1 \leq m \leq M} \sum_{i=1}^{L} b_{\mathbf{y}_i} b_{\mathbf{x}_i}(m), \tag{20}$$

that is equivalent to a minimum Hamming distance decoder based on the reliable components only.

5. RESULTS OF COMPUTER SIMULATION AND CONCLUSIONS

Due restrictions on the paper length, in this Section, we will present only some obtained results for real images confirming the advantages of the proposed approach. We investigated the overall recognition accuracy according to the average probability of error P_e for $M = 2048$ images of size 60×60, i.e., $N = 3600$, under additive white Gaussian noise (AWGN) (Fig. 1) as a function of signal-to-noise ratio (SNR) defined as SNR $= 10 \log_{10} \frac{\sigma_X^2}{\sigma_Z^2}$ and lossy JPEG compression distortions (Fig. 2). All results are obtained for 100 noise and random projection matrix realizations. The random projection domain is of dimensionality $L = 128$ and the overcomplete domain of $J = 2500$.

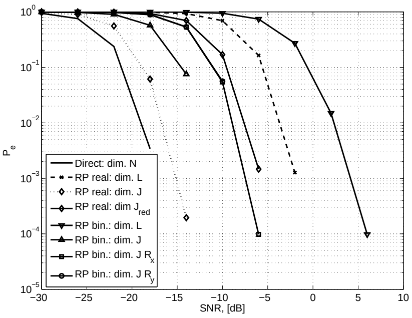

Fig. 1. The average probability of error for the AWGN.

The probability of identification error of the optimal ML decoder denoted as "Direct: dim. N" with the data dimensionality N represents the best achievable limit under the condition of completely known distributions for the AWGN that is obviously not the case for lossy JPEG compression. The dimensionality reduction based on random projections was performed for 3 lengths L, J and $J_{red} = J/8$ for comparison reasons correspondingly denoted as "RP real: dim. L", "RP real: dim. J" and "RP real: dim. J/8". Under both models of distortions, the dimensionality reduction, as expected, lead to an increase in probability of error. We tested also two classes of robust hashes, i.e., blind and reliability based ones. The blind robust hashes were tested for the lengths L and J denoted as "RP bin.: dim. L" and "RP bin.: dim. J", respectively. The L most reliable projections were selected out of J random projections for the reliability based hashing with the reliability estimation based on $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ denoted as "RP bin.: dim. JR_X" and "RP bin.: dim. JR_Y", respectively. The probability of error with blind binary hashing with the length L represents the upper bounds in the performed modeling while considerably enhances when the length is increased to J and the reliability information is provided and outperforms both blind cases. It is interesting to emphasize that the performance of the reliability based classifiers with the reliability estimation based on $\mathbf{R_x}$ and $\mathbf{R_y}$ practically coincide.

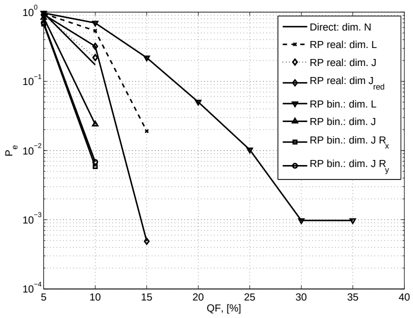

Fig. 2. The average probability of error for the lossy JPEG compression.

Therefore, the proposed approach closely approaches the performance of the optimal ML decoder operating on the full dimensionality data with the exact knowledge of the channel noise while the prosed identification based on the bit reliability uses the low dimensionality data representation and is characterized by lower complexity.

6. REFERENCES

[1] 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*, vol. 16, no. 2, pp. 302–317, February 2008.

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

[3] J. A. OSullivan and N. A. Schmid, "Performance analysis of physical signature authentication," *IEEE Trans. on Info. Theory*, vol. 47, pp. 3034–3039, 2001.