Decision-theoretic consideration of robust hashing: link to practical algorithms

Oleksiy Koval, Sviatoslav Voloshynovskiy, Fokko Beekhof, and Thierry Pun

CUI-University of Geneva, Stochastic Image Processing Group,
24, rue du Général-Dufour, 1211 Genève 4, Switzerland
{Oleksiy.Koval, svolos, Fokko.Beekhof, Thierry.Pun}@cui.unige.ch
http://sip.unige.ch

Abstract. In this paper we propose to consider the problem of robust perceptual hashing of multimedia data as composite hypothesis testing. Such a problem formulation is justified by prior ambiguity about source statistics and channel parameters that is usually the case in multiple practical scenarios. An asymptotically universal test approaching the performance of the classical maximum likelihood test performed under the exact knowledge of the mentioned statistics is proposed under the specific constraints on the assumed source and geometric channel models. Finally, we consider the problem of a practical hash construction under the constraints on complexity, robustness to geometrical transformations, universality and security. The proposed solution is based on a binary hypothesis testing for randomly or semantically selected blocks or regions in sequences or images.

1 Introduction

Recent advances of modern digital imaging and audio open new directions in modern imaging science, content management and secure communications. Evidently, this development is still underway and further success in the mentioned and possibly new directions can be easily foreseeing. Simultaneously, this avalanche progress is inalienably followed by a risk of various malicious illegal actions including violation of copyright, unauthorized prohibited usage, multiplication and distribution of digital media, high fidelity efficient counterfeiting of digital and analog content as well as goods and products justifying an urgent need for reliable document, product and person identification. The requirements one needs to satisfy developing such techniques include robustness and security in order to withstand various attacks, and at the same time preserve privacy as well as universality to provide asymptotic independence to a complete or partial lack of prior information defining the protocol design particularities.

Historically, a suggested solution to the above mentioned set of problems was based on a classical cryptographic hashes. Possessing excellent security level, authentication mechanism based on cryptographic hash functions appeared to be still not free from some shortcomings mostly concerning its robustness to various representations of multimedia files. For example, an image can be represented in different formats and would be perceptually or semantically the same although the two digital files would be entirely different leading to the entirely randomized hashes.

In order to overcome the mentioned drawbacks and weaknesses of a classical approach, robust perceptual hashing has been recently proposed and have constituted the core of a challenging and dynamically developing research area.

The robustness/invariance of multimedia geometric hashing as a problem of robust pattern matching have received a lot of attention in computer vision ([2] and the reference therein), the issue of security still remains to be an open and little-studied problem. Moreover, to our knowledge no results exist justifying robustness of the hash in the setup where the malicious or unintentional modifications of the hash as well as the statistics of the multimedia input of a hashing function are defined retaining a certain level of ambiguity. Thus, new information-theoretic and detection-theoretic approaches to secure hashing, as well as carefully designed attacks, should be proposed and investigated. This aspect will potentially have a great impact on security applications, such as content, object, person authentication and identification, tamper evidence, synchronization, forensic analysis and brand protection as well as might be of some interest for non-secure applications such as multimedia information retrieval.

The design of efficient robust hashing techniques is a challenging problem that simultaneously addresses a set of various conflicting requirements including:

- **robustness to distortions**, i.e., the ability of hash function to produce asymptotically the same output based on inputs that differ by legitimate distortion level that can be a consequence of signal processing and/or desynchronization transformations applied to a multimedia data;
- **security**, i.e., the ability of the attacker to learn the hash (index m) without the knowledge of key k based on the observed data y^N and knowledge of hash codebook construction (equivocation $H(M|Y^N) = H(M) - I(M;Y^N)$) or about the key k based on the observed data y^N (equivocation $H(K|Y^N) = H(K) - I(K;Y^N)$);
- **universality**, i.e., optimal or asymptotically optimal hash performance in the case of lack of prior knowledge about the statistics of input source distribution and channel that is related to the machine learning framework and universal hypothesis testing;

Thus, *a robust perceptual hash* can be defined as a one-way function, which takes multimedia objects as inputs, and generates sufficiently-short binary strings approximately invariant under perceptual-quality-preserving modifications.

The domain of robust image hashing is an active and rapidly developing research direction that attracts significant attention in data-hiding community. The main focus of the conducted research falls on the experimental justification of resistance of practical robust hashing schemes to diverse attacking strategies that lead to non-significant perceptual modifications of media files. This resilience is achieved due to the use of error correcting codes [5], quantized pseudorandom robust semi-global statistics [3], or randomly quantized perceptually invariant image feature points [4].

Despite the evident success that was achieved in development of practical robust perceptual hashing algorithms, there are still some common open problems of state-of-the-art in robust hashing that include:

- lack of systematic information-theoretic or decision-theoretic performance limits;
- lack of solid security understanding;
- optimal practical concerns the selection of the most representative and robust features and construction of the corresponding classifiers (joint classifier and feature optimization (JCFO)) that can provide the best attainable exponent;
- lack of theoretical link between random coding exponent and hypothesis testing problem for robust hashing as a joint design of optimal source-channel code.

Leaving the joint classification/feature extraction optimization outside of the scope, the main goal of this paper can be formulated as follows. First, a decision-theoretic framework for the analysis and design of perceptually robust hashing will be introduced and theoretical limits on performance and security of these systems will be established. Secondly, main open problems and challenges that will guide the development of future robust hashing methods will be formulated.

This paper has the following structure. Theoretical formulation of robust hashing as composite hypothesis testing is presented in Section 2. Performance of robust hashing under the protocol ambiguity is considered in Section 3. Some aspects of practical hash construction are analyzed in Section 4. Finally, Section 5 concludes this paper.

Notations We use capital letters to denote scalar random variables X, X^N to denote vector random variables, corresponding small letters x and x^N to denote the realizations of scalar and vector random variables, respectively. The superscript N is used to designate length-N vectors $x^N = [x[1], x[2], ..., x[N]]$ with k^{th} element $x[k]$. We use $X \sim p_X(x)$ or simply $X \sim p(x)$ to indicate that a random variable X is distributed according to $p_X(x)$. $p(x^N; H_m)$ denotes pdf/pmf of x^N under hypothesis H_m. The mathematical expectation of a random variable $X \sim p_X(x)$ is denoted by $E[X]$. Calligraphic fonts $\mathcal{X}$ denote sets $X \in \mathcal{X}$ and $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$.

2 Hashing as composite hypothesis testing

We propose to consider robust perceptual hashing problem as $|\mathcal{M}|$-ary hypothesis testing. Such a formulation can be justified by a certain level of ambiguity concerning the distribution of discrete memoryless source (DMS) $p_{X^N}(x^N)$ that generates the hashing inputs and the parameters of the channel that models the introduced distortions. In the scope of this paper it is supposed that the sequences x^N are generated from $p_{X^N}(x^N; s_X^J)$, where s_X^J are the parameters of distribution defined on a discrete set $\mathcal{S}_X^J$ with a finite cardinality.

In order to simultaneously cope with signal processing and geometrical desynchronization distortions, the channel is modeled as a cascade of a fixed memoryless part given by the transition probability $p(v|x)$ and an invertible global mapping T_θ (Figure 1). In this consideration we suppose that the family $\{T_\theta, \theta \in \Theta_N\}$ satisfies the conditions of: (a) invertibility $T_\theta : \mathcal{Y}^N \to \mathcal{V}^N$ for all N and for all $\theta \in \Theta_N$; (b) restricted cardinality that at most is growing subexponentially with N: $\limsup_{N\to\infty} \frac{1}{N} \ln |\Theta_N| = 0$. This also oncerns the source, i.e., $\limsup_{N\to\infty} \frac{1}{N} \ln |\mathcal{S}_X|^J = 0$.

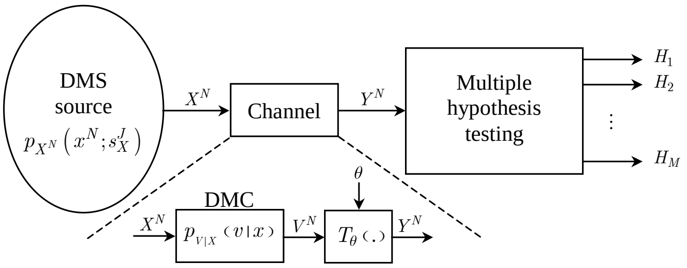

Fig. 1. Hashing as a multiple hypothesis testing.

According to the presented setup, one should decide based on the channel output which out of $|\mathcal{M}|$ hypotheses is in force:

$$H_m : Y^N \sim p(y^N; s_X^J, \theta, H_m), \tag{1}$$

where $1 \leq m \leq |\mathcal{M}|, s_X^J \in \mathcal{S}_X^J, \theta \in \Theta_N$.

The performance of $|\mathcal{M}|$-ary hypothesis testing is measured in terms of average probability of error for a given set of source s_X^J and channel θ parameters and a chosen decision rule ψ:

$$P_e(s_X^J, \theta, \psi) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} \Pr[\psi(Y^N) \neq m |\ m \text{ in force},\ s_X^J, \theta]. \tag{2}$$

If the statistics of source s_X^J and channel parameter θ are known, the probability of error (2) can be simplified to:

$$P_e = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} \Pr[\psi(Y^N) \neq m |\ m \text{ in force}]. \tag{3}$$

The test that minimizes the above error probability is the maximum likelihood (ML) decision rule:

$$\hat{H}_m = \psi_{ML}(y^N) = \arg \max_{1 \leq m \leq |\mathcal{M}|} p(y^N; H_m). \tag{4}$$

Hash generation procedure can be considered in the rate-distortion formulation when the DMS with the pmf $p_{X^N}(x^N)$ generates $2^{NH(X)}$ typical sequences that are mapped to $|\mathcal{M}| = 2^{NR}$ sequences as:

$$\psi_{ML} : \mathcal{X}^N \to \{1, 2, \cdots, 2^{NR}\}, \tag{5}$$

i.e., assigning an index m to all sequences x^N that are within some distance measure $d^N(x^N, \hat{x}^N(m))$ bounded by D to the sequence $\hat{x}^N(m)$ (a region of admissible distortions).

However, since some robustness is required to be possessed by a generated hash, one should address a question of the maximum number of uniquely recognizable sequences $\hat{x}^N(m)$ for a given D and the unknown θ under the condition that the source pmf $p_{X^N}(x^N)$ is selected such to be matched with the DMC. This number is defined by $2^{NR_{max}}$ where:

$$R_{max} = \min_{\theta \in \Theta_N} \max_{p_{X^N}} I(X;Y). \tag{6}$$

It should be noticed that there exists generally no decision rule that achieves P_e, if the DMS and channel parameters are not known.

3 Universal hypothesis testing

A capability of an algorithm to act at performance level independent of available prior information could be attractive in many practical applications including decision making. Unfortunately, to prove universality of decision rules that in the case of our formulation should be independent of unknown parameters s_X^J and θ is not a trivial task. Instead of proving the universality of the selected decision rule, we will rather demonstrate its *asymptotic universality* in the sense that it achieves exponential decay of error probability for all values of s_X^J and θ:

$$\lim \sup_{N \to \infty} \max_{s_X^J \in \mathcal{S}_X^J} \max_{\theta \in \Theta_N} \frac{1}{N} \ln \frac{P_e(s_X^J, \theta, \psi)}{P_e} = 0. \tag{7}$$

There are two ways of removing the performance dependence of the mentioned parameters of DMS s_X^J and channel state θ. In case, it is supposed that they are coming from some probability densities, one can apply Bayes approach using integration of $p(y^N; s_X^J, \theta, H_m)$ over the corresponding pmfs. However, this approach is not free from some drawbacks: (a) the lack of knowledge of prior distributions; (b) once the realizations of parameters are drawn, they remain fixed through the entire experiment and (c) the integrals are difficult to compute in practice. Therefore, usually the second approach based on the generalized ML (GML) is prefered in practice:

$$\psi_{GML}(y^N) = \arg \max_{1 \leq m \leq |\mathcal{M}|} \max_{s_X^J \in \mathcal{S}_X^J} \max_{\theta \in \Theta_N} p(y^N; s_X^J, \theta, H_m) \tag{8}$$

or

$$\psi_{GML}(y^N) = \arg \max_{1 \leq m \leq |\mathcal{M}|} p(y^N; \hat{s}_X^J, \hat{\theta}, H_m) \tag{9}$$

where $\hat{s}_X^J = \arg\max_{s_X^J \in \mathcal{S}_X^J} p(y^N; s_X^J, \theta, H_m)$ and $\hat{\theta} = \arg\max_{\theta \in \Theta_N} p(y^N; s_X^J, \theta, H_m)$ are the ML-estimates of s_X^J and θ, respectively.

One can find the conditions of GML universality under the assumptions about the parameter set $\mathcal{S}_X^J$ and index Θ_N considered in Section 2 according to:

$$\max_{s_X^J \in \mathcal{S}_X^J} \max_{\theta \in \Theta_N} \frac{P_e(s_X^J, \theta, \psi)}{P_e} \leq |\Theta_N| |\mathcal{S}_X|^J (N+1)^{|\mathcal{X}|(1+2|\mathcal{Y}|)}, \tag{10}$$

and thus the GML hypothesis testing rule is asymtotically universal [6, 7].

4 Practical hash construction

A practical implementation of the introduced $|\mathcal{M}|$-ary hypothesis testing is a very complex problem that covers:

- **computational complexity** that should be asymmetrically low for the authorizes users versus unauthorized ones;
- **robustness to geometrical transformations**. It can be attained in several different ways:
  - **exhaustive search** over Θ_N is possible without loss in performance under the specific constraints on the set Θ_N in price of computational complexity (10);
  - **selection of robust or invariant features** obtained in some transform domain or based on robust feature extraction. Such a strategy might lead to the performance loss in the case the dimensionality reduction is provided in an non-ivertible way due to data processing inequality.
- **priors about the source statistics** are very important in order to reduce variability of media data statistics. Fortunately, in most cases, some parametric families such as Generalized Gaussian are used in the transform domains such as DCT or DWT;
- **security** can be achieved by randomization of feature selection or key-dependent randomized codebook construction. A possible loss in performance accuracy should be carefully analyzed depending to the randomization scheme.

Evidently, to find an optimal trade-off among the mentioned conflicting requirements is not an easy task left outside of the scope of this paper. In order to relax the actual problem formulation, we will propose a suboptimal low-complexity hashing, which consists in replacement of $|\mathcal{M}|$-ary composite hypothesis testing by a set of binary counterparts. At first stage of the proposed approach the entire sequence y^N is splited into L possibly overlapping blocks as shown in Figure 2.

Then, the binary test ψ_B is applied to each block ℓ with $1 \leq \ell \leq L$ of a fixed length P (or to its transformed version/extracted features) to form the resulting

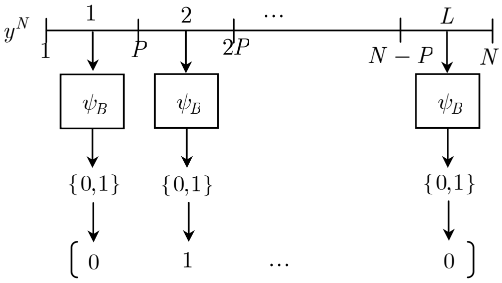

Fig. 2. Binary hypothesis based hash construction.

$\{0,1\}^L$-hash as a concatenation of the L-binary decisions. The test ψ_B is defined as:

$$\psi_B(y_\ell^P) = \arg \max_{1 \le m \le 2} \max_{s_X^J \in \mathcal{S}_X^J} \max_{\theta \in \Theta_N} p(y_\ell^P; s_X^J, \theta, H_m), \tag{11}$$

where $1 \le \ell \le L$. Further simplification might come from the fact that under the assumptions adopted in this paper the DMS statistics and the parameters of a global geometrical transformation remain the same in all L blocks and can be estimated only in one block.

The minimum average error probability for each block for the case of known source and channel parameters is bounded as:

$$P_e^B \le P(H_1)^{1-s} P(H_2)^s e^{-D_s(p(y^P;H_1),p(y^P;H_2))}, \forall\, 0 < s < 1, \tag{12}$$

where $P(H_1)$ and $P(H_2)$ are prior probabilities of hypothesis H_1 and H_2 and $D_s(p(y^P;H_1), p(y^P;H_2))$ is the Chernoff distance defined as:

$$D_s(p(y^P;H_1), p(y^P;H_2)) =$$

$$= -\ln \int_{\mathcal{Y}} p(y^P;H_1) \left(\frac{p(y^P;H_2))}{p(y^P;H_1)}\right)^s dy^P. \tag{13}$$

The total probability of error is the union of probabilities for each block. However, since average probabilities of errors are exponentially small, it is possible to demonstrate that the overall probability of error will be defined by the probability of error of a block that is characterized by the smallest Chernoff distance [1] among the all pairs.

A practical implementation of considered binary version of $\mathcal{M}$-ary hypothesis test includes the following steps:

1. **Transform** is required to provide robustness to geometrical distortions and asymptotically guarantee independent and identically distributed output. The randomized data sampling is considered as potential part of this stage.
2. The estimates of true moments based on the available feature pdf, characteristic functions or moment generation functions are obtained on the **Feature**

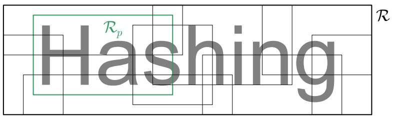

Fig. 3. Practical hashing with randomized region partition.

statistics computation step:

$$\Phi_Y(t) = \int_{-\infty}^{+\infty} p_Y(y)e^{jty}dy, \tag{14}$$

as:

$$\hat{m}_{n,\ell} = \frac{1}{P}\sum_{i=1}^{P} y_\ell[i]^n, \;\; M_{n,\ell} = \int_{-\infty}^{+\infty} \Phi_{Y,\ell}(t)t^n dt, \tag{15}$$

$$\hat{m}^A_{n,\ell} = \frac{1}{P}\sum_{i=1}^{P} |y_\ell[i]|^n, \;\; M^A_{n,\ell} = \int_{-\infty}^{+\infty} \Phi_{Y,\ell}|t|^n dt. \tag{16}$$

for $n \geq 1$. Depending on a particular application, in order to achieve the optimal performance in terms of P_e^B (12), the selection of low-dimensional features should be performed trying to maximize the Chernoff distance $D_s(.,.)$.

3. **Decision making** stage consists in deciding $\{0, 1\}$ between the alternative hypotheses H_1 and H_2.

Example. Application of the introduced binary hypothesis testing framework for robust hashing of text documents is presented in Figure 3. First, the text image area $\mathcal{R}$ is partitioned in a randomized way onto L blocks. Then the first empirical moment $\hat{m}_{1,1}$ is computed to be used for deciding $\{0, 1\}$. A possible modification of the applied transformation nay include semantic segmentation in order to simulate the use of object character recognition (OCR). Another extension consists in application of the presented strategy to more complex media data like grayscale images and audio signals that also includes the estimation of parameter θ in the scope of GML strategy to recover from potential geometrical distortions.

5 Conclusion

In this paper, we considered the problem of decision-theoretic analysis of robust perceptual hashing. The main obtained results can be summarized as follows. First, we propose a composite hypothesis testing formulation of the problem of hashing under source and channel ambiguity. Second, in order to link the considered theory with practice in a way that a number of conflicting requirements to complexity, robustness, lack of priors and security will be satisfied, we analyzed a practical hash construction that replaces an $\mathcal{M}$-ary formulation by a set of binary hypothesis tests and presented the details of its implemetation for robust hashing of text documents.

Acknowledgment

This paper was partially supported by SNF Professeur Boursier grants PP002-68653, 114613 and 200021-111643, by the European Commission through the IST Programme under IST-2002-507932-ECRYPT and Swiss IM2 projects.

References

1. D.H. Johnson C.C. Leang. On the asymptotic of m-hypothesis bayesian detection. *IEEE Trans. on Information Theory*, 43(1):280–282, October 1997.
2. M. Lifshits, I. Blayvas, R. Goldenberg, E. Rivlin, and M. Rudzsky. Rehashing for bayesian geometric hashing. In *Proceedings of ICPR 2004*, volume 3, pages 99–102, 23-26 August 2004.
3. M. K. Mihak, R. Venkatesan, and T. Liu. Watermarking via optimization algorithms for quantizing randomized semi-global image statistics. 2(11):185–200, Dec. 2005.
4. V. Monga and B. L. Evans. Robust perceptual image hashing using feature points. In *ICIP 2004*, pages 677–680, 2004.
5. R. Venkatesanan, S. Koon, M. Jacubowski, and P. Moulin. Robust image hashing. In *ICIP 2000*, Vancouver, BC, Canada, September 2000.
6. S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun. Geometrically robust perceptual image hashing. Technical report, University of Geneva, Feb. 2007.
7. S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun. Robust perceptual hashing as classification problem: decision-theoretic and practical considerations. In *IEEE Second Workshop on Multimedia Signal Processing (MMSP-07)*, Panorama Hotel, Chania, Crete, Greece, October 2007.