Robust perceptual hashing as classification problem: decision-theoretic and practical considerations

Sviatoslav Voloshynovskiy, Oleksiy Koval, Fokko Beekhof and Thierry Pun
CUI-University of Geneva,
24 rue Général-Dufour,
1211, Geneva, Switzerland
Email: {svolos, koval, beekhof, pun}@cui.unige.ch

***Abstract*— In this paper we consider the problem of robust perceptual hashing as composite hypothesis testing. First, we formulate this problem as multiple hypothesis testing under prior ambiguity about source statistics and channel parameters representing a family of restricted geometric attacks. We introduce an efficient universal test that achieves the performance of informed decision rules for the specified class of source and geometric channel models. Finally, we consider the practical hash construction, which compromises computational complexity, robustness to geometrical transformations, lack of priors about source statistics and security requirements. The proposed hash is based on a binary hypothesis testing for randomly or semantically selected blocks or regions in sequences or images. We present the results of experimental validation of the developed concept that justifies the practical efficiency of the elaborated framework.**

I. Introduction

New possibilities of digital imaging and audio open wide prospects in modern imaging science, content management and secure communications. However, despite the obvious advantages of modern digital technologies and their ongoing progress, these developments carry inherent risks, such as copyright violation, unauthorized prohibited usage and distribution of digital media, high fidelity efficient counterfeiting of digital and analog content as well as brand products. An urgent need for reliable document, product and person identification also calls for emerging necessity in robust and secure techniques, capable of withstanding various attacks, and at the same time preserving privacy. On the other hand, the issue of security is not necessarily emphasized in several other relevant applications, such as content indexing and retrieval, navigation, interaction with the physical world objects and scenes, but such tasks also require reliable and computationally efficient techniques for semantic content management. Thus, the need for such kind of techniques can be considered in secure and non-secure applications.

The solution to the above problems can be considered based on hash functions. Traditional cryptographic hash function based mechanisms have been found lacking for this purpose due to the peculiar nature of multimedia data. Namely, with multimedia data, the same content can have many different digital representations. 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.

Robust perceptual hashing (a.k.a. as fingerprinting in some contexts) has been recently proposed as primitives to overcome the above problems and have constituted the core of a challenging and dynamically developing research area.

Although the robustness/invariance aspects of multimedia hashing have received a lot of attention especially in computer vision, the issue of security still remains to be an open and little-studied problem. 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.

The design of efficient robust hashing techniques is very challenging problem that should address the compromise between various conflicting requirements that cover:

- *robustness to distortions*, i.e., the ability of hash function to produce either the same or close in some sense results under the legitimate distortions applied to the same data that include both signal processing and desynchornization transformations (that can be achieved using either special labeling or error correction decoding);
- *security*, i.e., the ability of the attacker to deduce the knowledge about 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 $H(K|Y^N) = H(K) - I(K;Y^N)$); this also includes the one-way hashing or non-invertibility property, i.e., computationally expensiveness in finding original data given a hash index m and a hash function (codebook construction), and collision-free property, which refers to the fact that given an input and a hash function, it is computationally hard to find a second image such that produces the same hash outside the regime of allowable distortions.
- *universality*, i.e., the practical aspects of optimal hash construction under the lack of statistics about input source distribution and channel distortions that is related to the machine learning framework and universal hypothesis testing.

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

The domain of robust image hashing is an active and rapidly developing research direction that attracts significant attention in data-hiding community. Most of elaborated robust image hashes are mainly targeting providing tolerance to a wide range of perceptually insignificant distortions. Such robustness might be granted due to the use of error correcting codes [3], hash computation as quantized robust pseudorandom robust semi-global statistics of an image [1], or using randomly quantized perceptually invariant image feature points [2]. In all above mentioned cases, a set of experiments is performed to justify the efficiency of the developed methods facing certain attacks.

The common open problems of state-of-the-art in robust hashing are:

- lack of systematic information-theoretic or decision-theoretic analysis in both construction and performance;
- lack of solid security understanding;
- optimal practical design 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.

That is why the goal of this paper is to introduce a decision-theoretic framework for the analysis and construction of perceptually robust hashing that is free from the above drawbacks. According to this framework we will formulate the main open problems and challenges that will guide the development of future robust hashing methods. Finally, we believe that this framework will help establish the theoretical limits on performance and security of these systems. Our main initial statements can be summarized as:

- robust hashing is a joint design of optimal semantic source-channel coding;
- a useful tool for its design and analysis is a multiple hypothesis testing framework.

This paper differs from the previous ones in part of:

- decision-theoretic analysis based on the multiple hypothesis testing framework;
- practical aspects of robust hashing construction including robustness to distortions, security, universality with respect to prior knowledge about source statistics and geometrical channel state and complexity;
- error bounds on robust hashing performance;
- practical joint design of classifier and feature optimization formulation of robust hashing.

This paper is organized as follows. The theoretical formulation of robust hashing as composite hypothesis testing is given in Section II. The universal hypothesis testing is considered in Section III. Practical hash construction is explained in Section IV. Finally, Section V 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. Calligraphic fonts $\mathcal{X}$ denote sets $X \in \mathcal{X}$ and $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$.

II. Hashing as composite hypothesis testing

The hashing problem can be considered as $|\mathcal{M}|$-ary hypothesis testing problem. The composite character of hashing problem comes from a fact that both the distribution of discrete memoryless source (DMS) $p_{X^N}(x^N)$ and the parameters of channel are unknown. We will assume that the source generates the sequences x^N from the pmf $p_{X^N}(x^N; s_X^J)$, where s_X^J are the parameters of distribution given on a set $\mathcal{S}_X^J$ that is assumed to be discrete with the fixed cardinality.

The channel is modeled as a cascade of a fixed memoryless channel given by the transition probability $p(v|x)$ and an invertible global mapping T_θ, which models a geometric transformation. We will assume that the family $\{T_\theta, \theta \in \Theta_N\}$ satisfies the conditions of: (a) mapping invertibility $T_\theta : \mathcal{Y}^N \to \mathcal{V}^N$ for all N and for all $\theta \in \Theta_N$; (b) restricted cardinality that is either fixed or grows subexponentially with N, i.e., $\limsup_{N\to\infty} \frac{1}{N} \ln |\Theta_N| = 0$.

The considered set-up is presented in Figure 1.

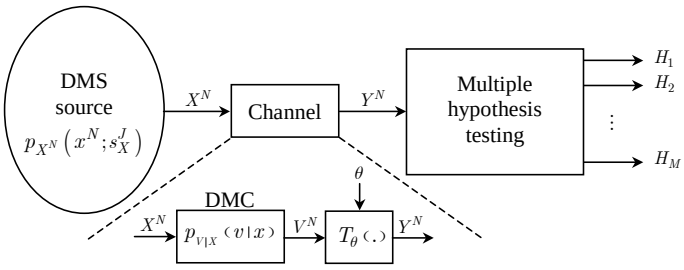

Fig. 1. Hashing as multiple hypothesis testing.

The corresponding $|\mathcal{M}|$-ary hypothesis testing is given in the form:

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

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

We will estimate the performance of hypothesis testing according to the average probability of error for a given set of source s_X^J and channel θ parameters in $|\mathcal{M}|$-ary composite

hypothesis testing and a chosen decision rule ψ:

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

In more general case, the analysis of performance might also include the expressions for probabilities of miss P_F and alarm P_M and the corresponding decision rules are chosen depending on the problem requirements.

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

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

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

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

This corresponds to 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:

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

thus 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 $\hat{x}^N(m)$, where $H(X)$ denotes the entropy and R stands for the rate.

Moreover, 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 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.

III. Universal hypothesis testing

The universal decision rules are independent of unknown parameters s_X^J and θ. However, in general the performance will depend on them. Therefore, a universal test is said to be *efficient*, if it achieves exponential decay of error probability for all values of s_X^J and θ:

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

In fact, considering the parameters of DMS s_X^J and channel state θ as random with some pmfs, one can apply Bayes approach using integration of $p(y^N; s_X^J, \theta, H_m)$ over the corresponding pmfs. However, this approach has some drawbacks related to: (a) the lack of knowledge of prior distributions; (b) once the realizations of parameters are drawn, they remain fixed through the experiment and (c) the integrals are difficult to compute in practice. Therefore, not always universal hypothesis testing based on generalized ML (GML) is used:

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

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-estimate 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 II 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 universal [4].

IV. Practical hash construction

The considered $|\mathcal{M}|$-ary hypothesis testing is a very complex problem that covers various aspects considered below:

- *computational complexity* for the authorized users should be low;
- *robustness to geometrical transformations* is very crucial for the robust media hashing. Practically, it can be solved 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 (the above considered GML strategy);
  - *selection of robust or invariant features* that assumes the transformation to some specific domain or application of robust feature extractors that might lead to the dimensionality reduction. If the transformation is not invertible, there might be the loss in performance due to data processing inequality. Such a transformation reduces the distance between the distributions of y^N under different hypothesis before and after transformation that increases inversely proportionally the probability of error;
- *priors about the source statistics* are very important due to high variability of image statistics. In the 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 the randomized feature selection or key-dependent randomized codebook construction. The loss in performance accuracy should be carefully analyzed depending to the randomization scheme.

To compromise the above practical requirements, we propose a suboptimal low-complexity hashing, which consists in replacement of $|\mathcal{M}|$-ary composite hypothesis testing by a binary counterpart. According to the proposed approach the entire sequence y^N is splitted into L possibly overlapping blocks as shown in Figure 2.

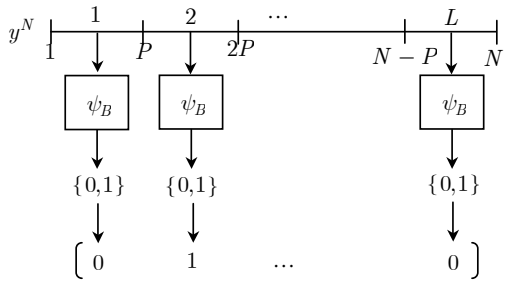

Fig. 2. Binary hypothesis based hash construction.

The binary test ψ_B is applied to each block ℓ with $1 \leq \ell \leq L$ to form the resulting $\{0,1\}^L$-hash as a concatenation of the L-binary decisions. The binary test ψ_B is:

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

where $1 \leq \ell \leq L$ and it is assumed that all blocks have the length P. Further simplification might come from the fact that the DMS statistics and θ are the same in all L blocks that can be estimated only in one block or over entire sequence y^N.

The minimum error probability for each block in assumption of known source and channel parameters can be bounded as:

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

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 distances defined as:

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

The total probability of error is the union of probabilities for each block.

The practical implementation of considered hash consists of following steps:

1) *Transform* step consists in the transformation of data y^N to some possible geometrically invariant or robust domain where samples are decorrelated and assumed to be independent and identically distributed. This also assumes the randomized data sampling. Figure 3 shows the randomized sampling of image region $\mathcal{R}$ with L randomly overlapping areas (for simplify assumed to be rectangles).

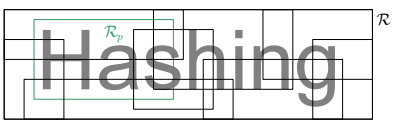

Fig. 3. Practical hashing with randomized region partition.

2) *Feature statistics computation* step consists in the computation of empirical moments (as unbiased estimates of true moments) from the selected feature pdfs and characteristic functions or their absolute value counterparts:

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

as:

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

$$\hat{m}_{n,\ell}^A = \frac{1}{P} \sum_{i=1}^{P} |y_\ell[i]|^n, \; M_{n,\ell}^A = \int_{-\infty}^{+\infty} \Phi_{Y,\ell} |t|^n dt. \quad (16)$$

for $n \geq 1$. Depending on a particular application, one should select informative low-dimensional features with the overall objective to minimize P_e^B in (12) that is achieved by maximizing the Chernoff distance $D_s(.,.)$.

3) *Decision making* step consists in deciding $\{0,1\}$ for each randomized area according to the hypothesis H_1 and H_2, respectively.

Example with the above text hashing includes the direct randomized partition of text image area $\mathcal{R}$ onto L blocks with the computation of the first moment $\hat{m}_{1,1}$ that are used for deciding $\{0,1\}$. More involved transform might also include a sort of semantic sampling where the statistics are computed only from the non-white regions that corresponds to the segmentation and mimics optical character recognition. This also includes the orientation estimation as a part of geometrical channel parameter θ in the scope of GML strategy. Similar functions can be extended to more complex grayscale images and audio signals.

V. Conclusion

In this paper, we dealt with the problem of decision-theoretic analysis of robust perceptual hashing. Firstly, we addressed the problem of hashing as composite hypothesis testing and considered the universal formulation of this problem and source and channel ambiguity. Secondly, we studied a practical hash construction that complies with a number of conflicting requirements to complexity, robustness, lack of priors and security. We have considered an example of text hashing and performed extended experimental validation that confirms the high efficiency of the proposed framework.

Acknowledgment

This paper was partially supported by SNF Professeur Boursier grants PP002-68653, 114613 and 200021-111643 and EU project ECRYPT and Swiss IM2 projects.

References

[1] 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.

[2] V. Monga and B. L. Evans. Robust perceptual image hashing using feature points. In *ICIP 2004*, pages 677–680, 2004.

[3] R. Venkatesanan, S. Koon, M. Jacubowski, and P. Moulin. Robust image hashing. In *ICIP 2000*, Vancouver, BC, Canada, September 2000.

[4] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun. Geormetrically robust perseptual image hashing. Technical report, University of Geneva, Feb. 2007.