Information-Theoretic Analysis of Electronic and Printed Document Authentication

Voloshynovskyy, Svyatoslav; Koval, Oleksiy; Villan Sebastian, Renato Fisher; Topak, Emre; Vila Forcen, Jose Emilio; Deguillaume, Frédéric; Rytsar, Yuriy; Pun, Thierry

Information-theoretic analysis of electronic and printed document authentication

Sviatoslav Voloshynovskiy*, Oleksiy Koval, Renato Villan, Emre Topak,
José Emilio Vila Forcén, Frederic Deguillaume, Yuriy Rytsar and Thierry Pun
University of Geneva, Department of Computer Science,
24 rue Général-Dufour, CH 1211, Geneva, Switzerland

ABSTRACT

In this paper we consider the problem of document authentication in electronic and printed forms. We formulate this problem from the information-theoretic perspectives and present the joint source-channel coding theorems showing the performance limits in such protocols. We analyze the security of document authentication methods and present the optimal attacking strategies with corresponding complexity estimates that, contrarily to the existing studies, crucially rely on the information leaked by the authentication protocol. Finally, we present the results of experimental validation of the developed concept that justifies the practical efficiency of the elaborated framework.

Keywords: document authentication, data-hiding, robust hashing, security leakage, equivocation, hypothesis testing, separation principle.

1. INTRODUCTION

Text documents are still the most common and almost unavoidable form of information communication among humans. Text documents are omnipresent everyday in the form of newspapers, books, web pages, contracts, advertisements, checks, identification documents, etc. At the same time, they can be widely distributed in electronic form via Internet communications.

The high significance of text documents justifies the importance of their copyright protection, authentication and tracking that still remain an open and challenging problem. One possible explanation of the current situation is that text media have a relatively small number of features that can be exploited in order to hide (or embed) information in comparison to images, audio or video. For example, while it is often possible to perform imperceptible modifications to an image, casual readers can easily notice extra letters or punctuation symbols in a text. Indeed, a text document can be seen as a form of a highly structured image, which is precisely the kind of images the human visual system is more sensitive to. For the same reason, the data embedding rate in text media is comparatively much smaller than in images, audio or video. Moreover, the hidden data can be always removed by the advanced methods of optical character recognition (OCR) due to the above mentioned high structuring of text information. Therefore, the copyright protection of text documents solely based on robust data-hiding or watermarking seems to be questionable. Fortunately, in the most important practical situations, one is more interested in the authentication of text information rather than in the protection of copyright contrarily to the images, video and audio (the exception is the copyright protection of printed papers and books).

That is why we formulate the main goal of this paper as information-theoretic study of theoretic limits of text document authentication in both electronic and printed forms.

A document authentication system should decide whether a given document is authentic or not. The decision about the document authenticity is performed at the global level by using a secret key K. Contrarily, if the decision is made at the local level, we will refer to it as to a tamper proofing system. Document authentication and document tamper proofing aim at guaranteeing the authenticity of a document and at indicating the corresponding modifications, if the document is suspected to be non-authentic. The most common solution consists in the generation of the document's hash computed from the document based on a secret key. At the authentication stage, the hash value is computed again from the document under investigation and compared with the

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

securely stored one. In the case of an authentic document, two hash values should be identical. Otherwise, the decision about non-authenticity of the document is declared. Obviously, the hash value should be designed to withstand various unintentional modifications that might occur during the document's life cycle. At the same time, the hash should be sensitive enough to various intentional modifications. The development of such hash functions is an active field of research known as *robust visual hashing*.[1] Contrarily to document authentication, where the hash is taken from the entire document, document tamper proofing is based on the concept of local hashing. Thus, if some modifications occurred, the tamper proofing technology should be capable of identifying the modifications locally. This might be useful to provide the interested party with some hints and evidence about the introduced modifications. Finally, the hash value can be computed from the entire document, which can include text, images, logos, drawings, etc.

One can construct three basic protocols for hash-based document authentication depending on the storage of the hash: hash storage in a database, direct storage of the hash onto the document, and hash storage in the document itself (self-embedding). The document authentication protocol based on the hash-storage in a database is shown in Figure 1.

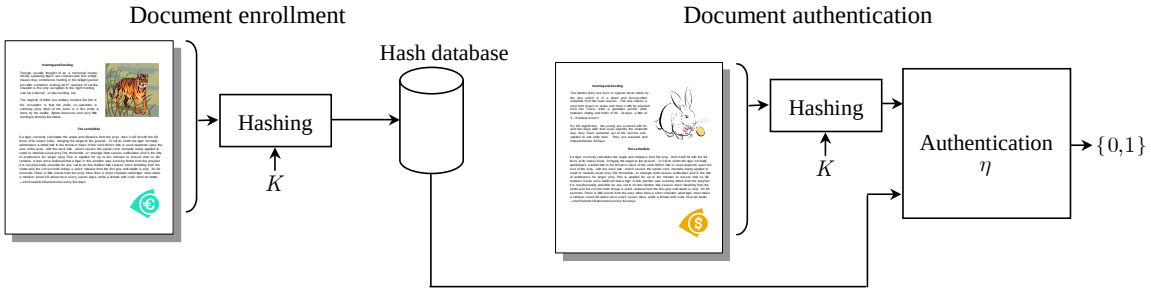

Figure 1. Document authentication protocol with the hash storage in a database.

The hash computed at the document enrollment stage is stored in the hash database and then is compared with a hash computed from the document under investigation at the authentication stage. The drawback of this solution consists in the necessity to have a direct access to the hash database that might be a practical limitations for numerous applications.

To overcome the above problem, the document hash can be stored directly onto the document using some special auxiliary data storage means. For instance, one can use some specific electronic memory chips, magnetic stripes, inks or even crystals to store the hash information. However, the most simple solution is to use barcode that can be either directly printed in one step together with the document reproduction or printed over the existing document at the second stage of document enrollment.

In such a protocol, the hash value is encrypted using a private key in order to prevent the possibility of generation of a new hash value from a tampered version of the document. For the reliable storage into the barcode, the obtained encrypted data is encoded using an appropriate channel code. The resulting barcode can be integrated directly into the digital document or can be printed onto the physical document.

One can distinguish two types of barcodes, namely, dense and sparse barcodes. The dense barcodes are printed in a localized position, for example at the margins of the document. Moreover, printing can be performed using either visible, ultraviolet, or infrared inks, depending on the application's concern about security. Barcodes can be read using low-resolution readers equipped with cheap charged coupled devices (CCDs) like those in flatbed scanners, handy scanners, digital photo cameras, web cameras, or even cell phone cameras. The drawback of this solution is related to specific aesthetic and security issues. In fact, the barcode can be simply removed from the document or can be easily copied if it was printed using ordinary inks. The latter does not provide copy evidence verification.[2] Alternatively, one can design a sparse barcode that is distributed over the document surface potentially in an invisible way using special inks or crystals or even specially chosen normal inks. Both approaches are shown in Figure 2. Although this protocol resolves the open issues of database hash storage

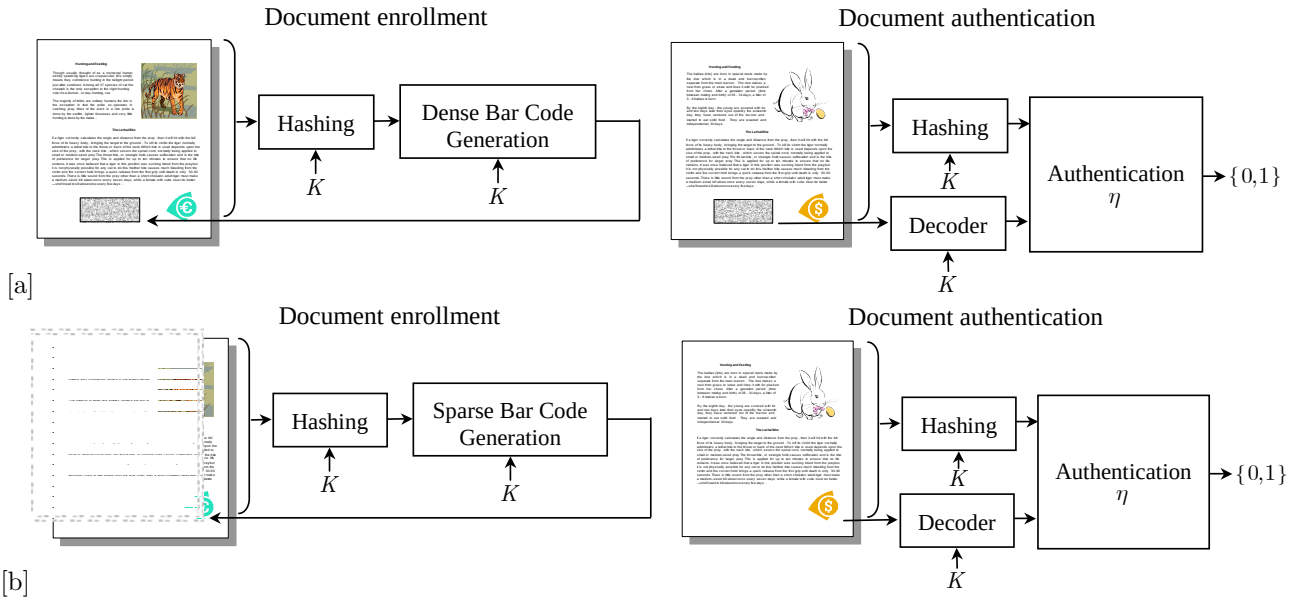

Figure 2. Document authentication protocol with the hash storage in barcodes: (a) dense barcodes and (b) sparse barcodes.

protocol, it possesses certain constraints in terms of aesthetics if dense barcodes are used, and flexibility handling electronic formats if sparse barcodes are used. Additionally, the storage capacity of sparse barcodes subject to visibility constraints is significantly lower those of dense barcodes that is an important factor for the security analysis as well as for the authentication power of the hash code.

Finally, the self-embedding approach that falls into the category of digital data-hiding seems to be very attractive for numerous reasons. The document authentication is performed directly based on the document without accessing a hash database similarly to the barcode protocol. However, contrarily to the barcode protocol the document is enrolled in a single step. Moreover, this approach can be easily integrated into any text/image editing tool and the resulting document can be stored in any suitable electronic format or even re-converted from one format to another. Additionally, the self-contained information cannot be easily separated from the document like in the case of dense barcodes. Finally, such an approach provides a unique copy evidence that was not possible with dense barcodes. It should be also mentioned that the storage capacity of data-hiding techniques might be superior than that of sparse barcodes but lower than that of dense barcodes. Taking into account all these advantages we concentrate our analysis only on this protocol. The document authentication protocol based on self-embedding is depicted in Figure 3 and interested readers can find more details about it in.[3,4]

The only drawback and main concern of the self-embedding approach is the limited data storage capacity resulting from the constraints on the document visible degradation as well as from the physical printing/scanning factors that might arise numerous security issues. Therefore, the optimal joint design of the corresponding system compromising between the limited storage capacity and security is of great practical importance as well as it represents a challenging theoretical problem.

The early methods of document authentication based on self-embedding hide the document ID as secret data. However, the embedding of document ID information is unsecure by itself. First, the hidden data is vulnerable to the copy attack, a protocol attack, which estimates the modulated features from the protected document (without knowing the key) and remodulates another document identically, thus creating an ambiguity.[2,5] Therefore, a document should be authenticated at the same time using content-dependent data, such as a hash code or a digital signature of the textual content and of the ID information. Secondly, tamper proofing (with localization capability) can be achieved by authenticating the text by parts or blocks. This means that a local modification

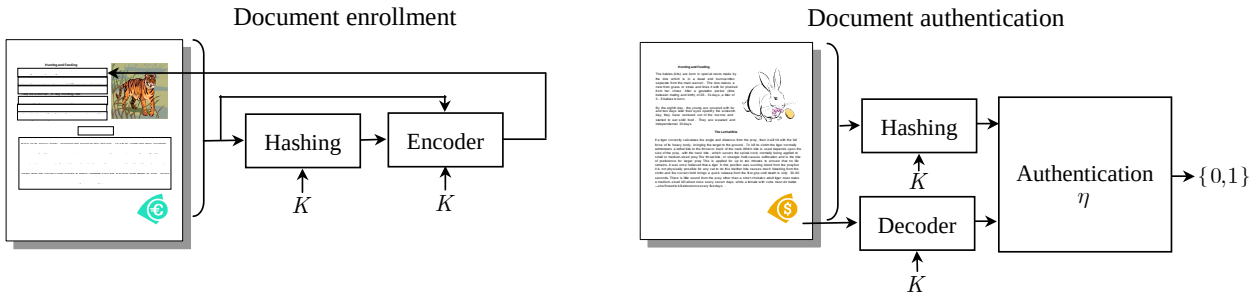

Figure 3. Document authentication protocol with the hash storage using self-embedding.

of the textual content can be detected with the resolution of one block, i.e., one word, one or several line(s), one paragraph, etc., depending on the embedding rate of the data-hiding algorithm.

Trying to be as general as possible from the point of view of the used apparatus of information theory, we distinguish three main practical scenarios:

- *Electronic document authentication*: the documents are stored and converted only in electronic format that is typical for Internet communications and database management;
- *Hybrid electronic-analog document authentication*: the documents are stored/communicated in electronic format as well as can be printed in certain circumstances that is typical for research, education, small and medium businesses;
- *Analog document authentication*: the document final destination is only envisioned in analog (printed) form that is typical for large-scale industrial applications.

Our goal is to analyze the performance and to consider the optimal design rules for these three scenarios. There are two possible designs of authentication systems based on data-hiding. The first approach assumes that the payload (message) is content-independent where it is either fixed for all documents or randomly generated from a user key.[6] The content authentication is based on the decision whether the embedded information is present or not in the document under analysis. However, this approach poses a lot of security concerns for both spread-spectrum and quantization based data-hiding techniques when the secret information (payload or key or both) can be estimated and then the tampered document can be remodulated similarly to the above mentioned copy attack.[2]

In the second approach, the payload (message) is assumed to be content dependent.[7,8] This prevents the usage of copy and remodulation attacks. Therefore, we will concentrate on the second approach. Although the approach we follow is also based on hashing-data-hiding, there are several important difference with.[7,8] First, we propose a generalized information-theoretic consideration of authentication problem, whereas the above approaches mostly advocate practical frameworks. Secondly, the major difference with the work[7] consists in the definition of authentication security solely expressed as the security of the hash based on the definition of its entropy. This has two important consequences: (a) the security leakage of data-hiding part was not taken into account in the analysis of the security of the complete system and (b) the security of the hash was defined solely based on the entropy of secret features whereas the actual security leak based on the additional observation of Y^N was disregarded. Finally, Fei *et. al.*[8] consider only key-dependent hashing while the data-hiding part security was not directly addressed, i.e., the key was not defined for the data-hiding part in Figure 1 of the cited paper. Moreover, depending on the application the attacker can learn more observing multiple documents, pages or paragraphs $Y_1^N, Y_2^N, \ldots$ protected with the same key or potentially by observing the same document protected with different keys. However, to be compliant with the above mentioned publications, we will constrain our analysis to the case of a single document Y^N.

The paper has the following structure. The basic formulation of authentication problem is considered in Section 2. Here we briefly review the fundamentals of authentication system performance analysis, analyze

possible system designs as well as introduce necessary data-hiding and hashing concepts in the scope of generalized Gel'fand-Pinsker[9] problem and robust hashing. In Section 3, we consider the main practical scenarios and corresponding channel models that define the data-hiding capacity. Section 4 contains the security analysis of document authentication protocols. Section 5 presents some preliminary results of system implementation and the concept validation. Finally, Section 6 concludes the paper and presents some future research perspectives.

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 $x^N = [x[1], x[2], ..., x[N]]^T$ 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 $p_X(x)$. 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. PROBLEM FORMULATION

2.1. Authentication problem based on self-embedding: global scale

In this Section, we consider the authentication problem based on self-embedding from the information-theoretic point of view. The generalized block-diagram of this set-up is shown in Figure 4.

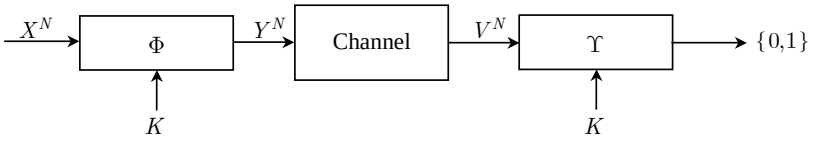

Figure 4. Generalized block diagram of authentication system based on a self-embedding.

According to the presented set-up, the data-hider has access to the uniquely assigned secret key $K = k$ that is uniformly distributed over the set $\mathcal{K} = \{1, 2, ..., |\mathcal{K}|\}$ and to the non-causal host $x^N \in \mathcal{X}^N$. The key k and the non-causal host realization x^N are used at the encoder Φ to generate the watermarked data $y^N \in \mathcal{Y}^N$. The watermarked data y^N is communicated through the attacking channel with some legitimate distortions described by the transition probability $p(v^N|y^N)$ resulting in $v^N \in \mathcal{V}^N$. The decoder Υ makes the decision about authenticity of v^N using k. Thus, the authentication system consists of the set $\{\mathcal{X}^N, \mathcal{K}, \mathcal{Y}^N, p(v^N|y^N), \mathcal{V}^N\}$, the data-hider allowable distortions D^E between X^N and Y^N, the attacking channel distortion D^A between Y^N and V^N and an encoder-decoder pair:

$$\Phi^N : \mathcal{X}^N \times \mathcal{K} \to \mathcal{Y}^N, \tag{1}$$

$$\Upsilon^N : \mathcal{V}^N \times \mathcal{K} \to \{0, 1\}. \tag{2}$$

The authentication problem, i.e., the problem of deciding whether the received data v^N are authentic or not, can be considered as a hypothesis testing problem.[10] One can assume that H_0 corresponds to the hypothesis that the received data is authentic, and H_1 corresponds to the hypothesis that v^N was generated by a fraudulent party. Therefore, the task of authentication as a hypothesis testing is to decide which of the two hypotheses is true given v^N. This generalized idea was first suggested by Maurer[10] and extended to steganographic applications by Cachin[11] and more recently elaborated by Wang and Moulin.[12] In the authentication application this approach was considered in.[7,8] We will assume that the test (2) is performed as:

$$\begin{cases} H_0, V^N \sim p_{V_0^N|K}(v^N|k), \\ H_1, V^N \sim p_{V_1^N|K}(v^N|k), \end{cases} \tag{3}$$

where H_0 corresponds to the decision 0 and H_1 to 1, respectively. Various tests can be performed,i.e.,Bayessian, minimax or Neyman-Pearson, however we will use the optimal Neyman-Pearson test in our formulation due to the particularities of authentication problem that will be discussed below. Disregarding the chosen testing

strategy, two types of errors are possible: type I error or a false alarm occurs denoted as P_f, if an authentic document is decided to be a counterfeited one, and type II error or a miss occurs denoted as P_m, if the mistaken decision is taken about a counterfeited document considering it as the authentic one.

According to the Neyman-Pearson test, the goal of the data-hider is to keep the probability P_F fixed and to minimize the probability P_M of missing a document counterfeiting. Contrarily, the objective of the counterfeiter is to modify the document keeping the distortions in the specified ranges of legitimate modifications in such a way that the data-hider cannot notice these modifications. Thus, the objective of the counterfeiter is to maximize the probability of miss P_M. These conflicting requirements can be formulated as a game between the data hider and attacker:

$$min_{\Phi,\Upsilon} max_{p_{V^N|Y^N}(.|.)} P_M(\Phi, \Upsilon, p_{V^N|Y^N}(.|.)) \geq min_{\Phi} max_{p_{V^N|Y^N}(.|.)} min_{\Upsilon} P_M(\Phi, \Upsilon, p_{V^N|Y^N}(.|.)), \tag{4}$$

which depends on the particular encoder/decoder pair Φ, Υ and the attacking channel $p_{V^N|Y^N}(.|.)$.

The Neyman-Pearson test states that for a given maximal tolerable probability P_F, P_M can be minimized by assuming hypothesis H_0 if, and only if, the log-likelihood ratio defined as:

$$\ell(v^N|k) \triangleq \log_2 \frac{p_{V_1^N|K}(v^N|k)}{p_{V_0^N|K}(v^N|k)}, \tag{5}$$

satisfies:

$$\ell(v^N|k) \geq T, \tag{6}$$

for some threshold T.

We will accordingly define the corresponding probabilities of miss and false detection for a given key k as:

$$P_M(k) \triangleq \Pr[\ell(v^N|k) < T|H_1], \tag{7}$$

$$P_F(k) \triangleq \Pr[\ell(v^N|k) > T|H_0]. \tag{8}$$

The relative entropy or discrimination $D(p_{V_1^N|K}||p_{V_0^N|K})$, defined as the expected value of the log-likelihood function in (5) with respect to $p_{V_0^N|K}(v^N|k)$, measures the level of distinguishibility between the two involved distributions:

$$D(p_{V_1^N|K}||p_{V_0^N|K}) = \sum_{k\in\mathcal{K}} p_K(k) \sum_{v^N\in\mathcal{V}^N} p_{V_1^N|K}(v^N|k) \log_2 \frac{p_{V_1^N|K}(v^N|k)}{p_{V_0^N|K}(v^N|k)}, \tag{9}$$

where $p_K(k)$ is the distribution of K on $\mathcal{K}$ that can be assumed to be uniform, i.e., $p_K(k) = 1/|\mathcal{K}|$.

In this case, the average error probabilities $P_F = \sum_{k\in\mathcal{K}} p_K(k)P_F(k)$ and $P_M = \sum_{k\in\mathcal{K}} p_K(k)P_M(k)$ satisfy[10]:

$$P_M \log_2 \frac{P_M}{1-P_F} + (1-P_M) \log_2 \frac{1-P_M}{P_F} \leq \sum_{k\in\mathcal{K}} p_K(k) D(p_{V_1^N|K}||p_{V_0^N|K}). \tag{10}$$

Fixing $P_M = 0$, one can obtain a lower bound on the miss probability:

$$P_F \geq 2^{-\sum_{k\in\mathcal{K}} p_K(k) D\left(p_{V_1^N|K}||p_{V_0^N|K}\right)}. \tag{11}$$

The complete system performance limits can be defined according to the Stein lemma.[13]

The above authentication system can be considered in the scope of a *hashing-communication problem* or a *hashing-data-hiding problem*. There are several possible system designs that can be based on *separation* or *joint* principles as an analogy to the Shannon source-channel communication problem.[13] Here, one is facing the same problems related to the optimality of performance, complexity as well as the additional issue of security. Since the optimal system structure still remains an open and little studied theoretical problem, we will try to analyze the the possible advantages of both approaches.

Separation approach: By analogy to the Shannon separation theorem, one can assume that the hashing-data-hiding problem can be nicely separated on hashing and data-hiding parts how it is done in the most practical authentication systems considered in the Introduction. This approach is schematically presented in Figure 5. The fact that the system can be separated in such a way without sacrificing optimality is conceptually and practically plausible. However, the analogy would not be complete without mentioning some differences. The Shannon source-channel separation theorem assumes that the source coding part does not require any knowledge about the channels statistics to produce the input to the channel coding part while for the channel coding part, the structure of the source code is irrelevant too. Contrarily, the hashing part should take into account legitimate channel distortion to produce a robust hash. This very important fact makes a significant difference between traditional hashing and hashing in multimedia authentication problems that especially concerns digital images, video and audio. This fact has also important implications for the system security.

Joint approach: While the separation approach provides a very nice solution to the classical source-channel communication problem, it does not claim to be unique. In fact, this approach is very expensive in terms of delay and complexity. That is why less expensive solutions can be found, which abandon the separation property and utilize *joint* source-channel coding. An extreme case of this approach is well-known as the *uncoded transmission*.[14] Similar consideration might be valid for the document authentication problem where some joint optimal approach could be suggested similarly to the setup shown in Figure 4 where no particular form of separation was assumed. We will consider this approach as a subject of future research concentrating on the separation approach in this paper.

Rate matching: By analogy with the source-channel communications suggested by the separation principle, one can consider the authentication problem based on separated hashing-data-hiding as a rate-matching problem. According to this interpretation, one should match the rate of the hash, selected to satisfy all requirements with respect to authentication, security and robustness for the defined legitimate distortions from one side and the rate of reliable hidden data communications from another. Contrarily to the source-channel communications based on the separation principle where the performance criterion is the distortion of the source for a given channel, in the hashing-data-hiding problem one has to deal with P_M for a fixed P_F and a given legitimate channel and a possible set of various counterfeiting attacks. In practice, it means that certain parameters should be optimally matched with others that include the document pmf $p_{X^N}(x^N)$ (respectively pdf $f_{X^N}(x^N)$), the distortion measure $d(.,.)$ and corresponding embedding distortion D^E, the channel conditional pmf $p_{V^N|Y^N}(v^N|y^N)$ (respectively conditional pdf $f_{V^N|Y^N}(v^N|y^N)$) and corresponding legitimate distortion D^A, the performance measures P_M and P_F as well as encoding and authentication functions. Moreover, one should also take into account possible security leakages of this protocol that can be efficiently used by the counterfeiter to produce a faked document using quite involved yet light complexity attacks.

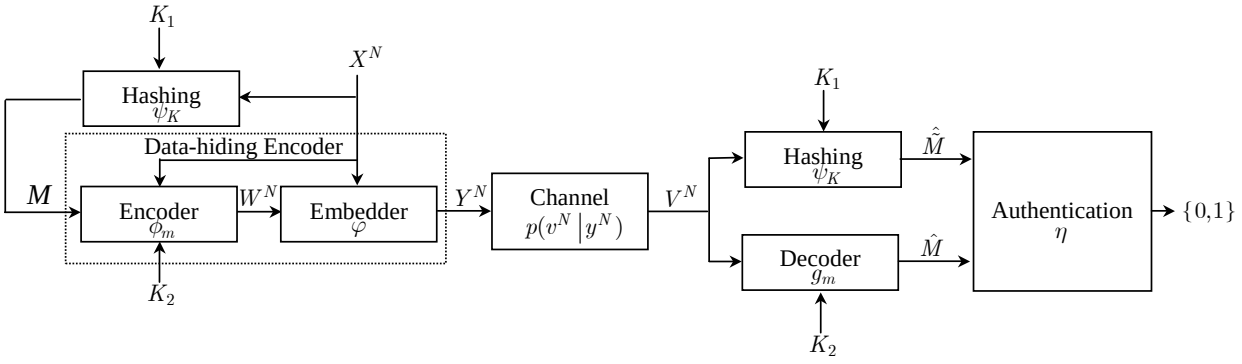

Figure 5. Document authentication based on the separation principle of hashing and data-hiding.

2.2. Authentication problem based on self-embedding: local scale

In this part, we will consider in details all elements of an authentication system based on the separation principle shown in Figure 5 and indicate the conditions of parameter matching. According to the above set-up, the data-hider has access to the uniquely assigned secret keys $K_1 = k_1$ and $K_2 = k_2$ that are uniformly distributed over the sets $\mathcal{K}_1 = \{1, 2, ..., |\mathcal{K}_1|\}$ and $\mathcal{K}_2 = \{1, 2, ..., |\mathcal{K}_2|\}$ and to the non-causal interference $x^N \in \mathcal{X}^N$. We also assume that X^N is distributed according to $p_{X^N}(x^N)$. The key k_1 and the non-causal host realization x^N are used to generate a hash message m that is encoded into the watermark w^N based on the non-causal host x^N and key k_2. The watermark w^N is embedded into the host data x^N, resulting in the watermarked data y^N. The watermarked data y^N is communicated through the attacking channel with some legitimate distortions described by $p(v^N|y^N)$. The decoder estimates the message $\hat{m}$ based on the attacked data v^N and the available key k_2 while the hash $\hat{\tilde{m}}$ is computed from v^N based on k_1. Finally, the decision about authenticity of v^N is made based on the comparison of $\hat{m}$ and $\hat{\tilde{m}}$. We will assume that the message $m \in \mathcal{M}$ and hash $\tilde{m} \in \tilde{\mathcal{M}}$ are uniformly distributed over $\mathcal{M} = \{1, 2, ..., |\mathcal{M}|\}$ and $\tilde{\mathcal{M}} = \{1, 2, ..., \left|\tilde{\mathcal{M}}\right|\}$, respectively, with $|\mathcal{M}| = 2^{NR}$ and $|\tilde{\mathcal{M}}| = 2^{NR'}$, where R and R' are the data-hiding and hash rates, respectively, and N is the length of all involved vectors x^N, w^N, y^N and v^N. It is important to point out that under the conditions of this paper where optimal joint data-hiding-hashing is performed according to the separation principle, later formulated in the form of Conjecture 1, that $\mathcal{M} = \tilde{\mathcal{M}}$ (Figure 5). However, for the sake of generality the current notations are exploited in the remaining part of this paper.

It is assumed that the stego and attacked data are defined on $y^N \in \mathcal{Y}^N$ and $v^N \in \mathcal{V}^N$, respectively. The distortion function is defined as:

$$d^N(x^N, y^N) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i), \tag{12}$$

where $d(x_i, y_i) : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+$ denotes the element-wise distortion between x_i and y_i.

Definition 1: A *discrete memoryless legitimate data-hiding channel* consists of four alphabets $\mathcal{X}$, $\mathcal{W}$, $\mathcal{Y}$, $\mathcal{V}$ and a probability transition matrix $p_{V^N|W^N,X^N}(v^N|w^N, x^N)$ that corresponds to the covert channel communication of the watermark W^N through the host image X^N (channel $p_{Y^N|W^N,X^N}(y^N|w^N, x^N)$) and the attacking channel $p_{V^N|Y^N}(v^N|y^N)$ such that $p_{V^N|W^N,X^N}(v^N|w^N, x^N) = \Sigma_{y^N} p_{Y^N|W^N,X^N}(y^N|w^N, x^N) p_{V^N|Y^N}(v^N|y^N)$. The attacking channel is subject to the distortion constraint D^A:

$$\sum_{y^N \in \mathcal{Y}^N} \sum_{v^N \in \mathcal{V}^N} d^N(y^N, v^N) p_{V^N|Y^N}(v^N|y^N) p_{Y^N}(y^N) \leq D^A, \tag{13}$$

where $p_{V^N|Y^N}(v^N|y^N) = \prod_{i=1}^{N} p_{V|Y} p(v_i|y_i)$. We will understand under legitimate distortions signal processing operations such as lossy compression, addition of noise, change of contrast; printing/scanning operations with corresponding halftoning and inverse halftoning; geometrical distortions such as translation, rotation and scaling.

Definition 2: A $(2^{NR}, N)$ code for the data-hiding channel consists of a *message set* $\mathcal{M} = \{1, 2, ..., 2^{NR}\}$, an *encoding function*:

$$\phi_m^N : \mathcal{M} \times \mathcal{X}^N \times \mathcal{K}_2 \to \mathcal{W}^N, \tag{14}$$

embedding function:

$$\varphi^N : \mathcal{W}^N \times \mathcal{X}^N \to \mathcal{Y}^N, \tag{15}$$

subject to the embedding distortion constraint D^E:

$$\frac{1}{|\mathcal{K}_2||\mathcal{M}|} \sum_{k_2 \in \mathcal{K}_2} \sum_{m \in \mathcal{M}} \sum_{x^N \in \mathcal{X}^N} d^N(x^N, \varphi^N(\phi_m^N(m, x^N, k_2), x^N)) p_{X^N}(x^N) \leq D^E \tag{16}$$

and a *decoding function*:

$$g_m^N : \mathcal{V}^N \times \mathcal{K}_2 \to \mathcal{M}. \tag{17}$$

We define the *average probability of error* for a $(2^{NR}, N)$ code as:

$$P_e^{(N)} = \frac{1}{|\mathcal{K}_2||\mathcal{M}|} \sum_{k_2 \in \mathcal{K}_2} \sum_{m \in \mathcal{M}} \Pr[g^N(V^N, K_2) \neq m | M = m]. \tag{18}$$

Definition 3: A rate $R = \frac{1}{N} \log_2 |\mathcal{M}|$ is achievable for distortions (D^E, D^A), if there exists a sequence of $(2^{NR}, N)$ codes with $P_e^{(N)} \to 0$ as $N \to \infty$.

Definition 4: The capacity of the data-hiding channel is the supremum of all achievable rates for distortions (D^E, D^A).

Theorem 1 (data-hiding capacity for a fixed channel)[9]: A rate R is achievable for the distortion D^E and the fixed attacking channel $p(v|y)$ with bounded distortion D^A, iff $R < C$, where:

$$C = \frac{1}{N} \max_{p(u^N, w^N | x^N)} \left[I(U^N; V^N) - I(U^N; X^N) \right], \tag{19}$$

and U^N to be a random variable $u^N \in \mathcal{U}^N(K_2 = k_2)$, with $|\mathcal{U}(K_2 = k_2)| \leq |\mathcal{X}||\mathcal{W}| + 1$. We also assume that $p(k_2, x^N, u^N, w^N, y^N, v^N) = p(k_2)p(x^N)p(u^N|x^N)\underline{\mathbf{1}}\{u^N \in \mathcal{U}(K_2 = k_2)\}p(w^N|u^N, x^N)p(y^N|x^N, w^N)p(v^N|y^N)$ to reflect the technicality behind the codebook and watermark generations as well as channel degradations, where $\underline{\mathbf{1}}\{.\}$ denotes the indicator function.

The proof of this theorem in the more general form of an active attacker is provided by Moulin and O'Sullivan[15] and the details can be found in the referred paper. However, it is important to emphasize that the main difference with our set-up is the codebook construction and the corresponding interpretation of the user key. In the scope of this paper, the key K_2 is considered uniquely as the index that defines the codebook of a particular user. Contrarily, Moulin and O'Sullivan have a broader understanding of the key as a sort of side information shared between the encoder and the decoder, where K_2 can be in some relationship with X^N. Therefore, we assume that K_2 is solely a cryptographic key that is independent of X^N. The details of the codebook construction, encoding, decoding and performance analysis of this part of the code are given in our previous publication.[16]

Definition 5: A hash consists of a *hash set* $\tilde{\mathcal{M}} = \{1, 2, ..., 2^{NR'}\}$ and a *hash function*:

$$\psi_K : \mathcal{X}^N \times \mathcal{K}_1 \to \tilde{\mathcal{M}}. \tag{20}$$

The construction of a hash should satisfy several conflicting requirements. To analyze these constraints we will assume that $X^N = [X_1, X_2, ..., X_N]$ is a discrete memoryless source (DMS) that emits a sequence x^N. The hash function produces the secure hash index $\tilde{m} \in \tilde{\mathcal{M}}$, i.e., a hash value, given x^N and k_1. Contrarily to the classical hashing, where two vectors that differ in only a single bit will have independent hash values, we will require that two vectors x_1^N and x_2^N that are perceived (respectively, understood) by the observer (respectively, by the reader) to be similar is some sense will have the same hash value, even if x_1^N and x_2^N have small bit-level differences. In practice, it also means that if a vector x_2^N is obtained via a mapping $p(x_2^N|x_1^N)$ of x_1^N, where $E[d^N(X_1^N, X_2^N)] \leq D^A$, i.e., the difference between the two vectors is defined by some value of legitimate variation D^A under certain transforms, one should expect $\psi_K(x_1^N) = \psi_K(x_2^N)$. Additionally, the hash should be secure in the sense that having the image x^N, the attacker cannot generate a hash without the knowledge of a secure key k_1. At this moment, we do not address the issues of computational complexity of hash calculation as well as collusion resistance.

Rather than follow some particular design of a hash function that is generally the case for most of the publications on this subject,[7,8,17] we will consider a generalized approach. We will assume that hashing is accomplished in some transform domain that is achieved by applying a transform $\mathbf{T}_K(.)$ that is key-dependent in the general case. We will denote a transformed vector as $\tilde{x}^L = \mathbf{T}_K(x^N)$ of length L and $\tilde{x}^L \in \tilde{\mathcal{X}}^L$ as shown in Figure 6.

The application of transform $\mathbf{T}_K(.)$, besides the security concerns, is additionally supposed to provide the hash robustness against various legitimate degradations within D^A. The robustness or invariance to legitimate distortions of the vector coefficients in the transform domain by applying transform $\mathbf{T}_K(.)$ might be achieved in

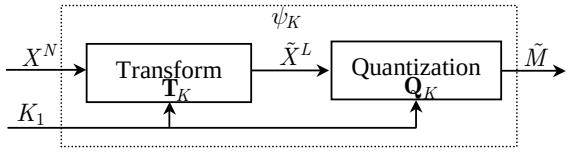

Figure 6. Robust hashing: generalized block diagram.

several different ways: (a) using an invariant domain[17–19] or (b) using robust features[20,21] that correspond to the $\mathbf{Q}_K$-block in Figure 6.

We will present a random code construction for hashing and analyze its performance assuming that the legitimate distortion D^A is the same in the transform domain, although it is not true in the general case but it can be assumed for all orthogonal transforms. Suppose we choose a mapping $p_{\hat{\tilde{X}}|\tilde{X}}(.|.)$ and compute $p_{\hat{\tilde{X}}}(.)$ as the marginal distribution of $p_{\hat{\tilde{X}}\tilde{X}}(.)$.

Hash code construction: Generate $2^{NR'}$ codewords $\hat{\tilde{x}}^L(\tilde{m}, k_1)$, $\tilde{m} = \{1, 2, \cdots, 2^{NR'}\}$ for each $k_1 \in \mathcal{K}_1$, with $\mathcal{K}_1 = \{1, 2, \cdots, |\mathcal{K}_1|\}$ by choosing each of the $L2^{NR'}|\mathcal{K}_1|$ symbols $\hat{\tilde{x}}[i](\tilde{m}, k_1)$ in the codebooks independently at random according to $p_{\hat{\tilde{X}}}(.)$. It can be shown that if the number $2^{NR'}|\mathcal{K}_1|$ is smaller than $2^{H(\hat{\tilde{X}}^L)}$, one can hope to have a unique set of sequences in each user codebook. Finally, assign indices $\tilde{m}$ to the codewords $\hat{\tilde{x}}^L(\tilde{m}, k_1)$ for each k_1 in such a way that the distance between the binary representation of indices $\tilde{m} \equiv \tilde{\mathbf{b}}$, $\tilde{b}[j] \in \{0, 1\}$ for the original and legitimately distorted vectors is minimal. One can consider the equivalent problem of robust labeling in the design of channel codes based on *multilevel coding*[22] where Gray labeling is chosen instead of a randomized assignment. Therefore, we can construct a set of codebooks for $|\mathcal{K}_1|$ users that can be also considered as a binning technique.

Hash encoding: Given x^N and k_1 or equivalently $\tilde{x}^L$ after transform $\mathbf{T}_K(.)$, try to find a codeword $\hat{\tilde{x}}^L(\tilde{m}, k_1)$ such that $(\tilde{x}^L, \hat{\tilde{x}}^L(\tilde{m}, k_1)) \in A_\epsilon^{*(L)}(\tilde{X}, \hat{\tilde{X}})$, i.e., two codewords are strongly jointly typical.[13] If one finds such a jointly typical pair, send or declare the corresponding index $\tilde{m}$. Otherwise, an error is declared.

In fact, one can show that the achievable rate R' of this hash will satisfy:

$$R'(D^A) = \min_{p_{\hat{\tilde{X}}|\tilde{X}}(.|.): E[d(\tilde{X}, \hat{\tilde{X}})] \leq D^A} I(\tilde{X}; \hat{\tilde{X}}), \tag{21}$$

if we allow the "collusion" of all vectors in the range of D^A using a proof similar to classical Shannon source coding theorem.[13] The sketch of the achievability part of the proof is based on the analysis of the bound on average distortion for three different cases: (a) $\tilde{x}^L \notin A_\epsilon^{*(L)}(\tilde{X})$, (b) $\tilde{x}^L \in A_\epsilon^{*(L)}(\tilde{X})$ but non of the $\hat{\tilde{x}}^L(\tilde{m}, k_1)$ satisfies $(\tilde{x}^L, \hat{\tilde{x}}^L(\tilde{m}, k_1)) \in A_\epsilon^{*(L)}(\tilde{X}, \hat{\tilde{X}})$ and (c) $\tilde{x}^L \in A_\epsilon^{*(L)}(\tilde{X})$ and we find a $\hat{\tilde{x}}^L(\tilde{m}, k_1)$ with $(\tilde{x}^L, \hat{\tilde{x}}^L(\tilde{m}, k_1)) \in A_\epsilon^{*(L)}(\tilde{X}, \hat{\tilde{X}})$. One can show that in all these cases the distortion will not exceed D^A with hight probability as long as $L \to \infty$.

In practice, one can consider the implementation of such a hash code using the keyed vector quantizer $\mathbf{Q}_K(.)$ shown in Figure 6, where the reconstruction points are selected properly for each key $k_1 \in \mathcal{K}_1$ to satisfy the distortion constraint. To simplify the implementation and analysis, one can design a dithered quantizer $\mathbf{Q}(\tilde{x}^L - d_{k_1}^L) + d_{k_1}^L$, which uses a fixed vector quantizer $\mathbf{Q}(.)$ in all cases and a dither vector $d_{k_1}^L$ generated from the key k_1. Further simplification can be achieved using uniform scalar quantization instead of vector quantization.

Definition 6: An authentication is defined as the binary decision $\{0, 1\}$ based on mapping:

$$\eta : \tilde{\mathcal{M}} \times \tilde{\mathcal{M}} \to \{0, 1\}. \tag{22}$$

The performance of the overall authentication system will be defined according to the results presented in Section 2. The authentication is performed based on the validation of two hypothesis H_0 and H_1 based on two

binary representations of the hash computed from the observed data v^N, namely $\hat{\tilde{m}} \equiv \hat{\tilde{\mathbf{b}}}$ and the decoded message $\hat{m} \equiv \hat{\mathbf{b}}$. The binary decision $\{0, 1\}$ is taken based on the comparison of the number of different bits with respect to a threshold defined according to the specified P_F.

The main issue in the information-theoretic analysis of this set-up consists in the derivation of direct and converse theorems for reliable document authentication under the assumed types of channels with the legitimate distortions.

Conjecture 1 (Authentication based on hashing-data-hiding principle): if X^N is a finite alphabet stochastic process that satisfies the asymptotic equipartition property (AEP)[13] then there is a hashing-data-hiding code with specified P_F and $P_M \to 0$, if the rate of the hash code R' satisfies $R' < C$. Conversely, for any stationary process, if $R' > C$, the P_M is bounded away from zero, and it is not possible to authenticate X^N with arbitrarily low probability of error.

The proofs try to establish the information-theoretic bounds on the rate of the hash R', providing reliable document authentication for the above channels and assumed intentional class of tampering attacks, as well as the rate of reliable message communications $R \leq C$, where C is the Gel'fand-Pinsker capacity for the text data-hiding. In fact, the solution to this problem indicates that $R \geq R'$ that assumes the design of hashing-data-hiding that is similar in spirit to Shannon joint source-channel coding based on the separation principle.[13]

Here, we sketch the proof of the achievability part of the above conjecture. Assuming that the encoder and decoder share the same pair of secret keys k_1, k_2, we will use three quantities defining the performance of the system, i.e., data-hiding, code according to P_e^N and the authentication code according to P_F and P_M. Since, we assume that X^N satisfies the AEP, we assume that the hash code with parameters specified in Definition 5 can be constructed with $2^{NR'}$ codewords for each key k_1, i.e., with the rate R'. According to the considered robust labeling, the index of the hash for a given key k_1 is fed to the input of the channel encoder assigning $m = \tilde{m}$. The encoder maps this index with the key k_2 into the sequence Y^N sent to the decoder according to Definition 2 with rate R that satisfies the conditions of Theorem 1.

One can transmit the hash index m to the decoder with probability of error less than ϵ, if:

$$R' + \epsilon = R < C. \tag{23}$$

The decoder estimates the sent message $\hat{m}$ with high probability, if $R < C$ according to Theorem 1. Finally, if the hash code is able to reliably extract the index $\hat{\tilde{m}}$ from the channel output V^N under specified allowable legitimate distortion D^A and assuming the knowledge of the key k_2, one can guarantee the authenticity of V^N with the specified P_F and P_M for a given rate R' with high probability.

3. MAIN PRACTICAL SCENARIOS AND CHANNEL MODELS

In this Section, we will consider three main practical scenarios for the document authentication systems. For each of them, we will introduce the corresponding channel models and briefly discuss the overall technical specification of the system.

3.1. Electronic document authentication

Here we assume that the source (cover) document is available in structured electronic form generated by a character-oriented encoding using popular edition/publication tools like Microsoft Word (DOC), Adobe Acrobat (PDF), PostScript (PS), LaTeX(TeX), Rich Text Format (RTF), Hypertext Markup Language (HTML), eXtensible Markup Language (XML), etc. The documents are only stored and circulated in electronic form with possible conversion among the above formats. Moreover, during a format conversion some information can be distorted or lost. The goal is to develop such an authentication technique that should be flexible and compatible with all the above existing or even future formats. The theoretical analysis of the data-hiding part of such a technique was performed in our previous publication.[4] The hash part of the code can be simplified to the OCR operation as the unkeyed feature extraction stage and corresponding key-based hashing. Any other reasonable hashing satisfying the conditions of Definition 5 could be applied here.

The main problem comes from the fact that even under deterministic (predefined) format conversions (for example, Microsoft Word to Adobe PDF), the most advanced OCR tools are not able to produce the unique recognition results. Thus, one has two opportunities that consist either in the use of special robust hashing procedures or in the definition of a special protocol, which can benefit from the prior knowledge of all possible distortions causing the ambiguity in the OCR results. The results of our modelling have indicated that the OCR tool of ABBYY[23] can produce the necessary OCR under the broad class of electronic document conversions including pdf, ps, djvu and processing including compression of text documents stored using image graphical formats. The fact of a prior knowledge of channel transformations was reflected by introducing the corresponding prediction module at the encoder according to Figure 7. Common unkeyed hash primitives, which can be used are Rivest's Message Digest version 5 (MD5)[24] or NIST's Secure Hash Algorithm version 1 (SHA-1).[25] The output hash-codes could be encrypted by a symmetric block cipher such as the Data Encryption Standard (DES) or its triple version (Triple-DES),[26] Lai and Massey IDEA,[27] or the recently accepted NIST's Advanced Encryption Standard (AES / Rijndael).[28] A keyed hash functions could be Hash Message Authentication Code (HMAC) based either on MD5 or on SHA-1 [†]. The results of the experimental validation of this setup are presented in Section 5.

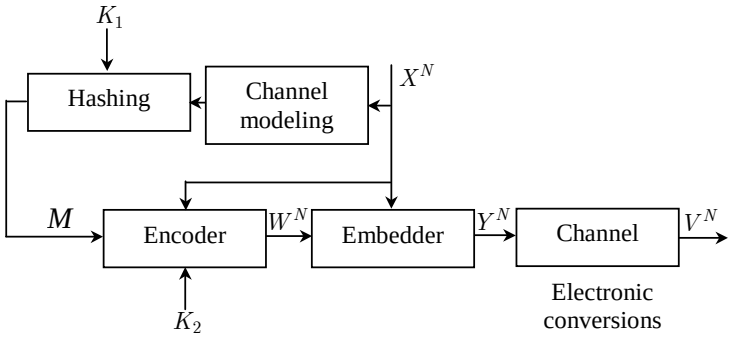

Figure 7. Electronic document authentication.

3.2. Hybrid electronic-analog document authentication

The analysis of the hybrid electronic-analog document authentication is the same as in the previous case besides that the channel additionally includes the halftoning process (Figure 8). The only difference comes from the mapping of modulated signals from the intensity space to the halftone space for the printed reproduction using a halftone encoder:

$$\phi_{HT} : \mathcal{Y}^N \to \mathcal{Y}'^{L \times N}, \tag{24}$$

where each character from $\mathcal{Y}$ is reproduced by a halftone pattern from $\mathcal{Y}'^L$ of size $n_1 \times n_2$ with $L = n_1 \cdot n_2$. The halftone space is defined to consist of black and white dots $\mathcal{Y}'^L \in \{0, 255\}^L$ that are reproduced by a printing device with resolution r_p. We assume that a scanner performs a mapping of this pattern into a new one with resolution r_s with a given size $n_1' \times n_2'$ and $L' = n_1' \cdot n_2'$.

Two possibilities exist depending on the relationship between the printing and scanning resolutions r_p and r_s.

If $r_s = r_p$, the so-called indirect decoding is applied where the halftone pattern $V'^{L' \times N}$ is first mapped to the intensity $\hat{Y}^N$ using an estimator of intensity (halftone decoder known as inverse halftoning):

$$g_{HT} : \mathcal{V}'^{L' \times N} \to \mathcal{Y}^N \tag{25}$$

[†]For security Triple-DES is preferable to DES which uses too short keys and has been officially abandoned. Moreover SHA (including all variants) is currently the only FIPS-approved method for secure hashing, and should be preferred to MD5 since it adds security measures.

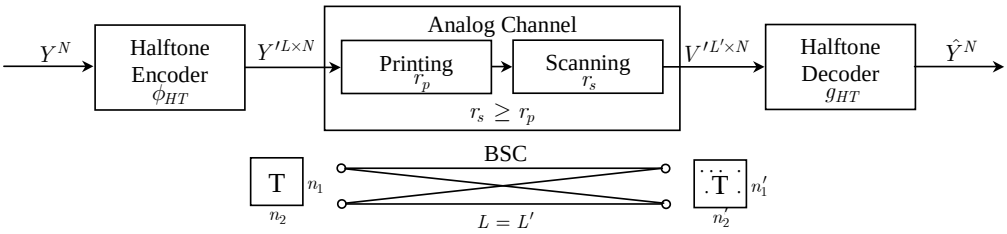

Figure 8. Hybrid electronic-analog document authentication.

and then decoded. In this case $L = L'$ and the channel is modeled as a parallel binary symmetric channel (BSC) with a given transition probability. The drawback of this decoding is related to the corresponding issues of the data-processing inequality. If $r_s > r_p$, one can perform direct decoding using a halftone pattern codebook thus avoiding the mapping to the intensity space.

Finally, it should be pointed out that the overall printing-scanning channel between Y^N and $\hat{Y}^N$ can be accurately modeled using a non-stationary Generalized Gaussian distribution (GGD) approximation. To validate this we performed numerous modelings for both inkjet and laser printers and various scanners under different resolutions. Here, we present the most typical results for two printers: laser printer HP Color LaserJet 4600 (BW mode), which will be denoted as p_1, and inkjet printer HP Color DeskJet 990Cxi (color mode) denoted as p_2. In both cases we have used the same Epson Perfection 3170 Photo CCD scanner. We simulated the range of inputs $y \in \mathcal{Y} = \{0, 1, \cdots, 255\}$ and $\hat{y} \in \mathbb{R}^+$. Figure 9 presents results for printing and scanning resolutions equal to 600 ppi.

The resulting channel model is:

$$\hat{Y} = \rho(Y) + Z, \tag{26}$$

with $\rho : \mathcal{Y} \to \mathbb{R}^+$ to be a non-linear function and Z represents zero-mean additive GGD noise whose parameters are determined by the input Y, i.e., $(Z|Y = y) \sim \mathcal{GGD}(0, \sigma_Z(y), \gamma_Z(y))$, where $\sigma_Z(y)$ and $\gamma_Z(y)$ are the noise standard deviation and shape parameter given $Y = y$. In our modeling, we have assumed that $\rho(y) = \hat{\mu}_{\hat{Y}|Y}(y)$, $\sigma_Z(y) = \hat{\sigma}_{\hat{Y}|Y}(y)$ and $\gamma_Z(y) = \hat{\gamma}_{\hat{Y}|Y}(y)$. The interested reader can find more details about this model in our previous publications.[4, 29]

For both models considered above one can compute the capacity of the data-hiding channel and perform the matching of corresponding hash rates.

3.3. Analog document authentication

In the case of analog document authentication we consider the industrial version of the system where the embedding is performed directly using halftone pattern codebooks thus omitting the intensity representation used in the previous versions. The advantage of this solution consists in the absence of intermediate processing operations that lead to a decrease of mutual information which impacts the achievable rate of message communications. The practical implementation of this system is under way.

4. SECURITY ANALYSIS OF DOCUMENT AUTHENTICATION

In this Section, we will consider the above authentication system from the security perspectives. As it was pointed out in Section 2.1, the overall objective of the counterfeiter is to modify the document in such way keeping the fact of modification undetectable, i.e., to increase the probability P_M operating in the range of legitimate distortions D^A. Contrarily to various attacking scenarios that in the general case assume the availability of several copies of protected documents $Y_1^N, Y_2^N, \cdots, Y_J^N$ marked with the same key, which have been intensively studied in the publications on authentication summarized by Maurer[10] providing corresponding bounds for impersonalization and authentication attacks, we will focus on a case of a single available document Y^N leaving the extension of our framework to multiple documents for future study.

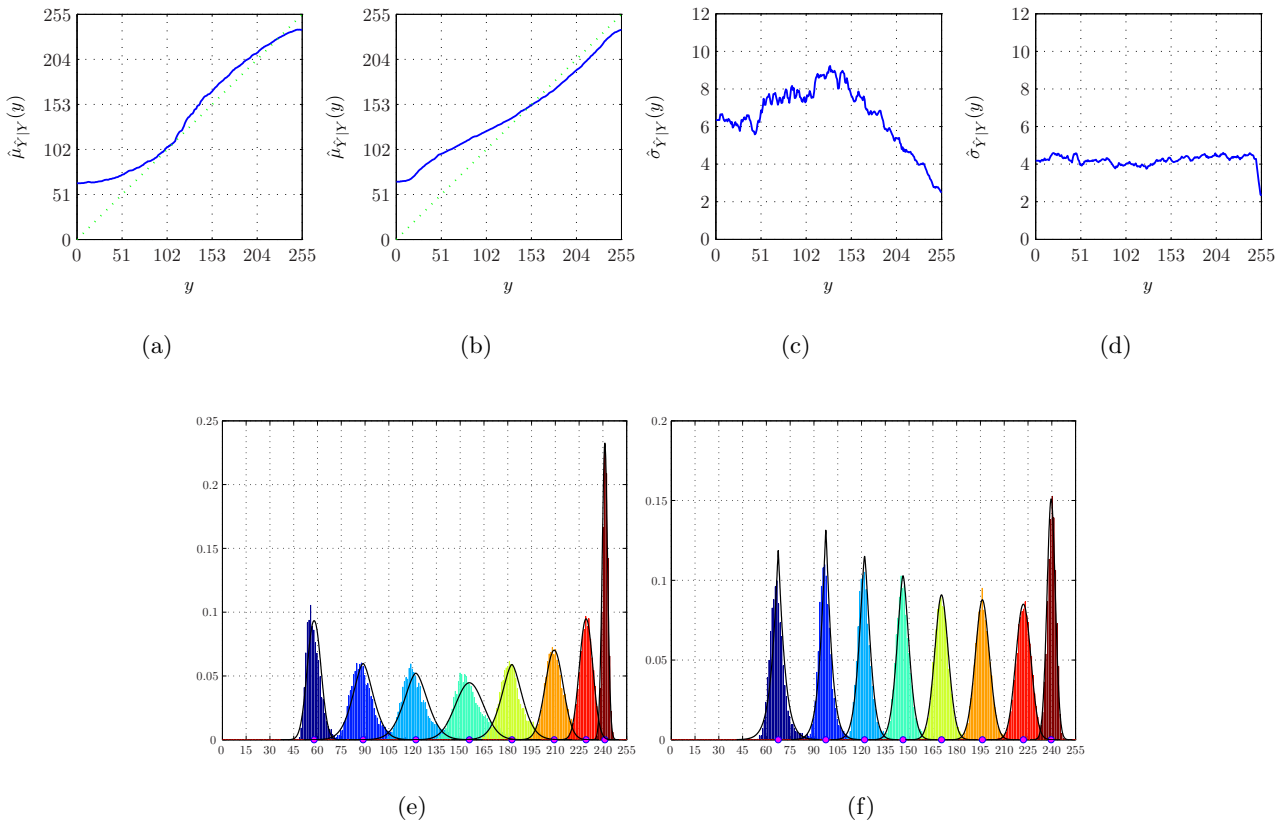

Figure 9. Conditional sample means (a) and (b) for printers p_1 and p_2, (c) and (d) corresponding conditional sample standard deviations and (e) and (f) normalized histograms of $\hat{Y}$ with the GGD approximation for the constellations $y \in \{0, 37, 73, 110, 146, 183, 219, 255\}$.

To fully benefit from the available copy of Y^N, we restate the Shannon equivocation principle considered in[30–32] and originally formulated with respect to the parameters of the scheme such as the secret key, the message, the host, the auxiliary random variable U^N and the hash $\tilde{M}$.

We consider the security analysis of the authentication system as a further development of the reversibility principles introduced in.[16] Here, we will adopt the security framework to the authentication application defining *security as the amount of trial efforts to reveal the secret information about the used scheme to design the worst counterfeiting strategy in the sense that the desirable document modification is unnoticeable by the authentication system, i.e., maximization of P_M combining all public information with the revealed secret one while keeping the distortions in the given ranges*. The trial efforts are considered in a broad sense for the generic applications. In the scope of this paper, we will only consider a particular aspect of complexity expressed by the number of checks to be performed to reveal some secret or to attack the scheme. This definition of security is also coherent with the definitions given in previous publications.[30–32]

In fact, we will show that the knowledge of $U^N = u^N(m, j, k_2)$ and $\hat{\tilde{X}}^L = \hat{\tilde{x}}^L(\tilde{m}, k_1)$ is sufficient to achieve the attacker goal under certain circumstances. Thus, the amount of efforts in terms of number of trials is understood as the real security of the robust data-hiding system. The larger this amount, the higher the security is. It should be also pointed out that a key difference of this approach with the previously considered ones consists in the fact that we analyze the security leak coming from the auxiliary random variable U and $\hat{\tilde{X}}^L$ rather than considering particular messages, keys or host. From the information about U^N and $\hat{\tilde{X}}^L$, one can deduce the

knowledge of their components, i.e., the key, message and even the host, and apply this knowledge selectively depending on a particular application scenario. The converse is not generally true.

Lemma 1 (Equivocation for the data-hiding code): The observation of Y^N reduces the ambiguity about U^N from $h(U^N)$ to $h(U^N|Y^N)$:

$$h(U^N|Y^N) = h(U^N) - I(U^N;Y^N), \tag{27}$$

where U^N is defined over $\mathcal{U}^N$.

Therefore, the corresponding complexity of the attacker in revealing the information about u^N used in the data-hiding code is:

$$\mathcal{O}_A^{DH} = 2^{h(U^N|Y^N)}. \tag{28}$$

In fact, it shows that the attacker can reduce the dimensionality of his exhaustive search space from $2^{h(U^N)}$ to $2^{h(U^N|Y^N)}$ trials by performing the jointly-typical decoding only with those codewords that are located within some "distance" from the estimate $E[U^N|Y^N]$ computed based on minimum mean square estimate (MMSE). It can be then argued that this distance is determined by the estimation variance $\sigma^2_{U^N|Y^N}$ that in fact defines the volume of ambiguity $h(U^N|Y^N)$. Thus, the attacker will perform the search not among all $u^N \in \mathcal{U}^N$ but only among those u^Ns that are within the above volume that considerably reduces his complexity thanks to the security leakage analysis. Finally, it should be pointed out that the knowledge of u^N provides also information about m, j and k_2 according to the assumed codebook construction.

Similar considerations can be deduced for the security of the hash part of the code.

Lemma 2 (Equivocation for the hash code): The observation of Y^N reduces the ambiguity about $\hat{\tilde{X}}^L$ from $h(\hat{\tilde{X}}^L)$ to $h(\hat{\tilde{X}}^L|Y^N)$:

$$h(\hat{\tilde{X}}^L|Y^N) = h(\hat{\tilde{X}}^L) - I(\hat{\tilde{X}}^L;Y^N), \tag{29}$$

where $\hat{\tilde{x}}^L \in \tilde{\mathcal{X}}^L$. Since $\hat{\tilde{X}}^L(\tilde{m},k_1)$ is a function of $\tilde{m}$ and k_1, the fact of revealing information about $\hat{\tilde{X}}^L$ implies the availability of information about $\tilde{m}$ and k_1.

Therefore, the corresponding complexity of the attacker in revealing the information about $\hat{\tilde{x}}^L$ used in the hash code is:

$$\mathcal{O}_A^{HS} = 2^{h(\hat{\tilde{X}}^L|Y^N)}. \tag{30}$$

The above considerations concerning the ambiguity volume are also valid here. It should be also pointed out the major difference between the presented approach and the approach suggested in[7] for the security evaluation based only on the entropy of $\hat{\tilde{X}}^L$, thus leading to the ambiguity volume overestimation. Such a definition of security disregards a possible security leakage from Y^N that is taken into account in our formulation. Finally, the security of data-hiding and hash codes is formulated using the same apparatus that makes this approach uniform for both parts.

Having defined the equivocations for the data-hiding and hash codes, we will consider several possible attacking scenarios against the authentication system for two different key management protocols. The consideration is performed according to the Kerckhoffs principle[33] assuming that the forger has access to all the particularities of the authentication protocol besides the secret keys and communicated messages.

4.1. The case of same keys

Within this key management protocol, we assume that $k_1 = k_2 = k$ that was a quite typical assumption for the majority of early authentication systems and investigate the possible security of this scheme.

The corresponding attacking strategy can be summarized as follows:

1. Given y^N, estimate the center of the ambiguity sphere about the sent codeword $E[U^N \,|Y^N]$ applying the MMSE strategy.

2. Find $\hat{u}^N$ based on the available y^N performing $2^{h(U^N|Y^N)}$ possible checks.
3. Find k since the knowledge of $\hat{u}^N$ reveals the complete information about m and k.
4. Apply reversibility, i.e., estimate $\hat{X}^N = E[X^N|Y^N, \hat{U}^N]$. Under special conditions considered in,[16] one can even achieve perfect reversibility $\hat{X}^N = X^N$.
5. Produce a faked copy X'^N of X^N according to the desire of the counterfeiter.
6. Compute a new hash value M' from the available X'^N and K.
7. Perform embedding of M' into X'^N using K.

The corresponding complexity of the attacker to perform the above operations is:

$$\mathcal{O}_1 = 2^{h(U^N|Y^N)}|_{U^N \in \mathcal{U}^N}, \tag{31}$$

which coincides with (28) indicating the important fact that the security of the hash does not play any role in this key management protocol.

The opposite is also valid, in case the attacker tries to reveal the information about k from the hash part of the code:

$$\mathcal{O}_2 = 2^{h(\hat{\tilde{X}}^L|Y^N)}|_{\hat{\tilde{X}}^L \in \tilde{\mathcal{X}}^L}. \tag{32}$$

Since, the attacker can always choose the strategy with lower complexity, the final estimate for the security of the considered authentication systems is:

$$\mathcal{O}_A = \min\{\mathcal{O}_1, \mathcal{O}_2\}. \tag{33}$$

4.2. The case of the different keys

Similar attacking strategy can be applied in the case of two different keys, i.e., $k_1 \neq k_2$. The attacker can first learn the embedded message based on Y^N with complexity $2^{h(U^N|Y^N)}|_{U^N \in \mathcal{U}^N}$ and then use the prior knowledge about the message to learn the key k_1 for the hash regeneration from the tampered data. This attack is similar in spirit to the so-called known message attack[30,32] and its complexity can be estimated as $2^{h(\hat{\tilde{X}}^L|Y^N)}|_{\hat{\tilde{X}}^L \in \tilde{\mathcal{X}}^L(\tilde{M}=\tilde{m},K_1),K_1 \in \mathcal{K}_1}$. Finally, the total efforts of the attacker are:

$$\mathcal{O}_1 = 2^{h(U^N|Y^N)}|_{U^N \in \mathcal{U}^N} + 2^{h(\hat{\tilde{X}}^L|Y^N)}|_{\hat{\tilde{X}}^L \in \tilde{\mathcal{X}}^L(\tilde{M}=\tilde{m},K_1),K_1 \in \mathcal{K}_1}. \tag{34}$$

Contrarily, the attacker can first estimate the message that was embedded based on the hash analysis and then apply the above mentioned known message attack to reveal the information about the secret key k_2 used for the data-hiding part. Knowing two keys, the attacker can perform the reversibility, regenerate the hash, and sequentially embed it into the tampered data. The total efforts of the attacker in this strategy can be estimated as:

$$\mathcal{O}_2 = 2^{h(U^N|Y^N)}|_{U^N \in \mathcal{U}^N(M=m,K_2),K_2 \in \mathcal{K}_2} + 2^{h(\hat{\tilde{X}}^L|Y^N)}|_{\hat{\tilde{X}}^L \in \tilde{\mathcal{X}}^L}. \tag{35}$$

The resulting efforts are equal to the minimum of (34) and (35):

$$\mathcal{O}_A = \min\{\mathcal{O}_1, \mathcal{O}_2\}. \tag{36}$$

5. SYSTEM IMPLEMENTATION AND CONCEPT VALIDATION

The presented system was implemented for the protection of text documents in both electronic and analog forms. To be robust against various protocol and cryptographic attacks we have followed the guidelines developed for the authentication of images in our previous work.[5] The particularities of the hash construction as well as block hashing can be found there. The text data-hiding code was implemented according to the technique presented in.[4] In this paper, we report the implementation of an authentication system according to the presented rate matching approach. The document enrollment is accomplished for two popular text editing tools Microsoft Word and LaTeX. The resulting document can be either stored in doc format or converted to pdf/ps formats or printed either from Microsoft Word or from Adobe PDF Reader/Ghostview.

The protected document is printed and scanned at 600 dpi and 600 ppi, respectively. The result of authentication of a randomly selected document fragment when no tampering was introduced is shown in Figure 10. The modifications have been introduced in this fragment by replacing word "Geneva" by "Paris V" in the electronic form and then the document was printed and scanned with the same parameters. The authentication has immediately revealed the fact of document tampering and the suspicious parts have been highlighted as shown in Figure 11.

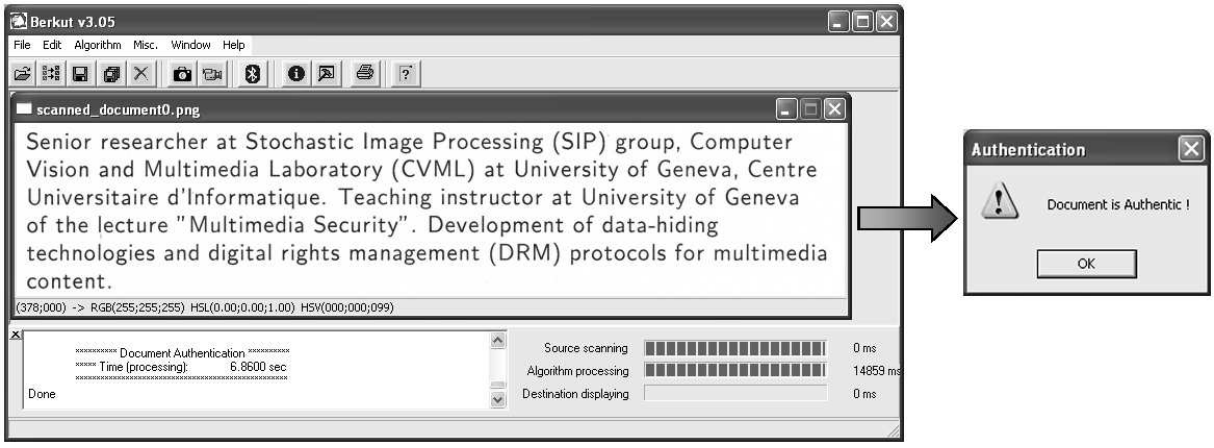

Figure 10. Document authentication without tampering.

6. CONCLUSIONS

In this paper, we considered the problem of electronic and printed documents authentication. In particular, we presented the information-theoretic framework for joint hashing-data-hiding and formulated the joint source-channel coding theorem with a sketched proof justifying the performance limits in such protocols. We performed the security analysis of document authentication methods and proposed the optimal attacking strategies with corresponding complexity estimates that are based on the information leakage analysis. Finally, the presented experimental results for a document authentication application justify the practical efficiency of the elaborated framework.

As possible extensions of the current status of our research that are already initiated we see the complete performance investigation of the reported system as well as a study of all related security issues. Despite a very promising character of the first obtained results (the system was capable to reliably detect all types of introduced tampering (more than 20) without producing a single false detection decision), all the tests where performed under a restricted number of printing/scanning setups. Therefore, further research will be focused on the experimental study of P_M and P_F for various printing/scanning conditions (different types of printing technologies including laser, ink jet, off-set as well as laser engraving on plastic, scanners and image acquisition

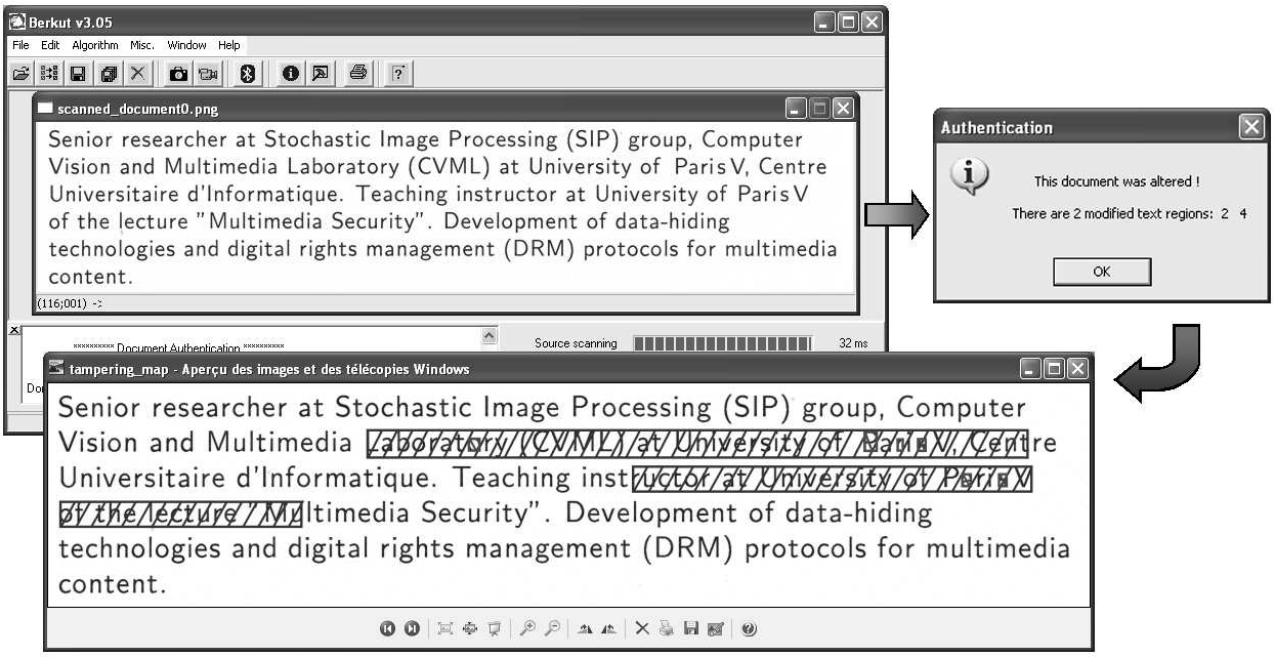

Figure 11. Document authentication from the tampered document.

devices including flat bed scanners, photo cameras and cameras of portable devices like mobile phones and PDAs). Finally, the attacking scenarios described in Section 4 will be also validated based on the developed prototype.

Acknowledgements

This paper was partially supported by SNF Professeur Boursier grant PP002–68653, by the European Commission through the IST Programme under contract IST-2002-507932-ECRYPT, FP6-507609-SIMILAR and Swiss IM2 projects.

The information in this document reflects only the authors views, is provided as is and no guarantee or warranty is given that the it is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

REFERENCES

1. J. Fridrich, "Visual Hash For Oblivious Watermarking," in *Proceedings of SPIE Photonic West Electronic Imaging 2000, Security and Watermarking of Multimedia Contents*, **3971**, pp. 286–294, (San Jose, California), Jan. 24-26 2000.
2. M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, **3971**, (San Jose, California USA), 23–28 jan 2000.
3. F. Deguillaume, Y. Rytsar, S. Voloshynovskiy, and T. Pun, "Data-hiding based text document security and automatic processing," in *IEEE International Conference on Multimedia Expo (ICME)*, (Amsterdam, The Netherlands), July 6-8 2005.
4. R. Villán, S. Voloshynovskiy, F. Deguillaume, Y. Rytsar, O. Koval, E. Topak, E. Rivera, and T. Pun, "A Theoretical Framework for Data-Hiding in Digital and Printed Text Documents," in *Proceedings of 9th IFIP TC-6 TC-11 International Conference on Communications and Multimedia Security*, **LNCS 3677**, pp. 280–281, (Salzburg, Austria), September 19–21 2005.

5. F. Deguillaume, S. Voloshynovskiy, and T. Pun, "Secure hybrid robust watermarking resistant against tampering and copy attack," *Signal Processing, Special Issue on Security of Data Hiding Technologies* **83**, pp. 2133–2170, October 2003.
6. J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *ICASSP 2001*, (Salt Lake City, Utah, USA), May 7–11 2001.
7. 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)*, **2**, pp. 1041–1044, (Philadelphia, PA, USA), March 2005.
8. C. Fei, D. Kundur, and R. Kwong, "Analysis and design of authentication watermarking," in *Proceedings of SPIE-IS&T Electronic Imaging 2004, Security, Steganography, and Watermarking of Multimedia Contents VI*, **53061**, (San Jose, USA), January 2004.
9. S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Probl. Control and Inf. Theory* **9**(1), pp. 19–31, 1980.
10. U. Maurer, "A unified and generalized treatment of authentication theory," in *Proc. 13th Symp. on Theoretical Aspects of Computer Science (STACS'96)*, *Lecture Notes in Computer Science* **1046**, pp. 387–398, Springer-Verlag, Feb. 1996.
11. C. Cachin, "An information-theoretic model for steganography," in *Information Hiding: Second International Workshop IHW'98*, (Portland, Oregon, USA), April 1998.
12. Y. Wang and P. Moulin, "Steganalysis of block-structured stegotext," in *Proceedings of the SPIE International Conference on Security and Watermarking of Multimedia Contents III*, **5306**, (San Jose, CA, USA), January 2004.
13. T. Cover and J. Thomas, *Elements of Information Theory.*, Wiley and Sons, New York, 1991.
14. T. Berger, *Rate-distortion theory: A mathematical basis for data compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.
15. P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. on Information Theory* **49**, pp. 563–593, March 2003.
16. S. Voloshynovskiy, O. Koval, E. Topak, J. Vila-Forcén, P. Comesana, and T. Pun, "On reversibility of random binning techniques: multimedia perspectives," in *9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security (CMS 2005)*, (Salzburg, Austria), September 2005.
17. A. Swaminathan, Y. Mao, and M. Wu, "Image hashing resilient to geometric and filtering operations," in *Proc. of IEEE Workshop on Multimedia Signal Processing (MMSP'04)*, pp. 355–358, (Siena, Italy), September 2004.
18. G. Bonmassar and E. Schwartz, "Space-variant fourier analysis: The exponential chirp transform," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, October 1997.
19. M. Wu and B. Liu, "Watermarking for image authentication," in *IEEE International Conference on Image Processing '98 (ICIP'98) Proceedings*, Focus Interactive Technology Inc., (Chicago, IL, USA), October 1998. TA10.11.
20. P. Bas, J. M. Chassery, and B. Macq, "Robust watermarking based on the warping of predefined regular triangular patterns," in *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, **3971**, pp. 99–109, (San Jose, CA, USA), 23-28 January 2000.
21. S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *IEEE Int. Conference on Image Processing 98 Proceedings*, Focus Interactive Technology Inc., (Chicago, Illinois, USA), October 1998.
22. J. Huber, U. Wachsmann, and R. Fischer, "Coded modulation by multilevelcodes: Overview and state of the art," in *ITG–Fachbericht: Codierung fur Quelle, Kanal und Ubertragung*, pp. 255–266, (Aachen, Germany), March 1998.
23. ABBY, "http://www.abbyy.com/," 2005.
24. R. Rivest, "The MD5 message-digest algorithm," April 1992. Request for Comments (RFC) 1321, MIT LCS and RSA Data Security, Inc.
25. NIST, "Secure Hash Standard," May 1993. Federal Information Processing Standards Publications (FIPS) PUB 180-1.

26. N. I. of Standards and T. (NIST), "Data Encryption Standard (DES)," October 1999. Federal Information Processing Standards Publications (FIPS) PUB 46-3.
27. X. Lai and J. L. Massey, "A proposal for a new block encryption standard," in *Lecture Notes in Computer Science (EUROCRYPT´90), Advances in Cryptology, I. B. Damgard editor, Springer-Verlag*, **473**, pp. 389–404, 1991.
28. N. I. of Standards and T. (NIST), "Advanced Encryption Standard (AES)," November 2001. Federal Information Processing Standards Publications (FIPS) PUB 197.
29. R. Villán, S. Voloshynovskiy, O. Koval, and T. Pun, "Multilevel 2D Bar Codes: Towards High Capacity Storage Modules for Multimedia Security and Management," in *Proceedings of SPIE-IS&T Electronic Imaging 2005, Security, Steganography, and Watermarking of Multimedia Contents VII*, **5681**, pp. 453–464, (San Jose, USA), January 16–20 2005.
30. L. Perez-Freire, P. Comesana, and F. Perez-Gonzalez, "Information-theoretic analysis of security in side-informed data hiding," in *7th Information Hiding Workshop (IHW2005)*, (Barcelona, Spain), June 2005.
31. P. Comesana, L. Perez-Freire, and F. Perez-Gonzlez, "Fundamentals of data-hiding security and their application to spread-spectrum analysis," in *7th Information Hiding Workshop (IHW2005)*, (Barcelona, Spain), June 2005.
32. F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Transactions on Signal Processing* , 2005. special issue "Supplement on Secure Media II".
33. A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires* , pp. 161–191, February 1883.