Active content fingerprinting: a marriage of digital watermarking and content fingerprinting

Sviatoslav Voloshynovskiy #1, Farzad Farhadzadeh #2, Oleksiy Koval #3, Taras Holotyak #4

Computer Science Department, University of Geneva
7, Route de Drize, CH-1227 Carouge (GE), Switzerland
1 svolos@unige.ch, 2 Farzad.Farhadzadeh@unige.ch, 3 Oleksiy.Koval@unige.ch, 4 Taras.Holotyak@unige.ch

***Abstract*—Content fingerprinting and digital watermarking are techniques that are used for the content protection and distribution monitoring. Nowadays, both techniques are well studied and their shortcomings are understood. In this paper, we introduce a new framework named as *active content fingerprinting* that takes the best from two worlds of content fingerprinting and digital watermarking to overcome some of fundamental restrictions of these techniques in terms of performance and complexity. The proposed framework extends the encoding of conventional content fingerprinting in the way similar to digital watermarking thus allowing to extract the fingerprints from the modified cover data. We consider several encoding strategies, examine the performance of the proposed schemes in terms of bit error rate and compare it with those of conventional fingerprinting and digital watermarking.**

I. Introduction

The *content fingerprinting*, a.k.a. robust perceptual hashing in some applications, has emerged as a tool for the content identification and integrity verification, filtering of user-generated content websites, content tracking, broadcast monitoring, upload control, etc. [1]. The content fingerprinting consists in the extraction of short, robust and distinctive fingerprint, which is in most cases a binary vector, allowing the operation with the data of lower dimensionality. The extracted fingerprints are stored in the databases, which can also contain some additional information about the fingerprint relations, clustering or binning, enabling efficient search of similar fingerprints.

In the conventional content fingerprinting, the fingerprint is computed directly from the original content and does not require any content modifications that preserves the original content quality. In this sense it can be considered as a *passive content fingerprinting* (pCFP).

The diagram of pCFP is shown in Fig.1a. The content owner provides the content $\mathbf{x} \in \mathcal{X}^N$ and assigns the ID number, $\mathrm{ID} \in \mathcal{M}$, $\mathcal{M} = \{1, 2, \cdots, |\mathcal{M}|\}$ to this content. The content owner also possesses a secret key $k \in \mathcal{K}$. The fingerprint extractor (FP) generates the fingerprint $\mathbf{b_x} \in \mathcal{B}^L$, where $\mathcal{B} = \{0, 1\}$, based on the mapping $\psi : \mathcal{X}^N \times \mathcal{K} \to \mathcal{B}^L$. The generated fingerprint $\mathbf{b_x}$ together with the assigned ID are stored in the database. For a given query $\mathbf{y}$, which might result either from the enrolled content $\mathbf{x}$ or unrelated one $\mathbf{x}' \in \mathcal{X}^N$, the FP estimates the fingerprint $\mathbf{b_y}$ by mapping $g : \mathcal{Y}^N \times \mathcal{K} \to \mathcal{B}^L$ and the decoder produces $\hat{\mathrm{ID}}$ or rejects the query.

Another approach to the content protection is based on *digital watermarking*. Nowadays, the digital watermarking is a well-studied domain were a lot of work was done on the investigation of its performance [2] and more recently security [3]. In this context, we only consider the main groups of watermarking methods that are based on known host, a.k.a. quantization index modulation or random binning-based methods, and known statistics methods, a.k.a. spread spectrum based methods [4]. To introduce the uniform consideration of pCFP and watermarking problems, we briefly consider the generalized diagram of digital watermarking as shown in Fig.1b. The essential difference between these two approaches is that in fingerprinting a content owner only assigns some ID number to the content $\mathbf{x}$, while in the digital watermarking one can mark every individual copy of content $\mathbf{x}$ by embedding a message $\mathbf{m}$ that by the analogy to the pCFP is assumed to encode the L character message from the alphabet $\mathcal{B}$, i.e., $\mathbf{m} \in \mathcal{B}^L$. The assigned ID number and the message $\mathbf{m}$ are stored in the database. Prior to the embedding, the message $\mathbf{m}$ is encoded into the codeword $\mathbf{b_x}$ based on the mapping $\omega_{\mathrm{ECC}} : \mathcal{B}^L \to \mathcal{B}^J$, where $J > L$ based on some error correction codes (ECC). The codeword $\mathbf{b_x}$ is embedded into the content $\mathbf{x}$, a.k.a. a host, thus resulting into the marked copy $\mathbf{v}$ according to some specified measure of distortions and embedding function $\psi : \mathcal{X}^N \times \mathcal{B}^J \times \mathcal{K} \to \mathcal{V}^N$. Obviously, two operations of message encoding and embedding can be combined together. However, to highlight the similarity with the pCFP and to reflect the way how most of practical digital watermarking methods are designed we separate these stages. Under this consideration, the role of considered embedder consists in the content modulation. Similarly to the pCFP, the extractor produces the estimate $\mathbf{b_y}$ based on mapping $g : \mathcal{Y}^N \times \mathcal{K} \to \mathcal{B}^J$. The estimation accuracy is evaluated based on probability of bit error $P_b = \Pr\{B_{\mathbf{x}} \neq B_{\mathbf{y}}\}$ that is similar to the pCFP. The goal of the next stage is to correct the errors in the estimate $\mathbf{b_y}$ using the error correction mapping $\omega_{\mathrm{ECC}}^{-1} : \mathcal{B}^J \to \mathcal{B}^L$, which should produce an estimate of the original message $\hat{\mathbf{m}}$. The performance of error correction

WIFS'2012, December, 2-5, 2012, Tenerife, Spain. 978-1-4244-9080-6/10/$26.00 *©2012 IEEE.*

decoder is evaluated based on $P_m = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \Pr\{\mathbf{M} \neq \hat{\mathbf{M}}|\mathbf{M} = \mathbf{m}\}$. The scheme is designed in such a way that the digital watermarking rate $R_{\text{DW}} = \frac{1}{N}\log_2|\mathcal{M}|$ approaches the watermarking capacity $C_{\text{DW}} = I(U;X) - I(U;Y)$ for host sequences of length N, where U denotes the auxiliary random variable [5]. The decoder should ensure the rejection option "$\oslash$" that was not considered in the original Gel'fand–Pinsker paper but that is automatically satisfied for strongly typical sequences in the theoretical analysis. In practice, the estimated message $\hat{\mathbf{m}}$ is matched with the database to deduce the estimate of ID number based on $\hat{\mathbf{m}}$. If the rate of the ECC is chosen properly according to P_b, the overall $P_m \to 0$ and there is no need in high complexity search procedures used in the pCFP to deduce the estimate of ID. Therefore, the digital watermarking possesses two advantages over the pCFP: (a) each copy of content $\mathbf{x}$ can be marked independently and (b) there is no need in complex search procedures due to the usage of structured error correction codes contrary to the random fingerprint codes.

At the same time, the practical digital watermarking techniques face the host interference. To achieve the host interference cancellation special binning or quantization techniques are used that are demonstrated to be very insecure in comparison to the spread spectrum based methods which suffer from the host interference [4]. This recalls for the trade off between the probability of error P_b, minimization of which requires an efficient host interference cancellation based on structured codes, and security, which, on contrary, is characterized by the leaks provided by any nonrandom code structures.

Unfortunately, it is little known about the security of practical pCFP. The secret key estimation in the pCFP is not a well studied problem besides some exceptions like [6]. At the same time, it is intuitive that since the content is not modified in the pCFP framework and if the database is handled properly, the attacker obtains much less information for the secret key estimation in contrast to the digital watermarking. Additionally, the pCFP does not require any embedding of messages into the host data and thus there is no need in the efficient host cancellation.

In this paper, we introduce a new hybrid technique that combines pCFP and digital watermarking to achieve better trade off between performance, complexity and security. We refer to this technique as *active content fingerprinting* (aCFP). The aCFP model is shown in Fig.1c. The aCFP essentially obeys the structure of pCFP except the only difference at the enrollment part. The modulator of aCFP observes the input vector $\mathbf{x}$ and generates the fingerprint $\mathbf{b_x}$ and modified vector $\mathbf{v}$ for a given key k and specified distortion between $\mathbf{x}$ and $\mathbf{v}$ based on this mapping $\psi : \mathcal{X}^N \times \mathcal{K} \to \mathcal{B}^L \times \mathcal{V}^N$. The identification are performed in the same way as for the pCFP.

It is important to point out that the resulting data $\mathbf{v}$ does not carry out any embedded message. The sole purpose of this modulation at the encoder is to decrease the probability of error P_b, which will have a crucial impact on both the overall system performance and complexity.

Therefore, the goal of this paper is twofold. The first goal is to evaluate the probability of bit error P_b for the aCFP under different modulation strategies to achieve better robustness and potentially faster search with respect to the conventional pCFP. The second goal consists in comparison of aCFP with the digital watermarking under the same distortion constraints.

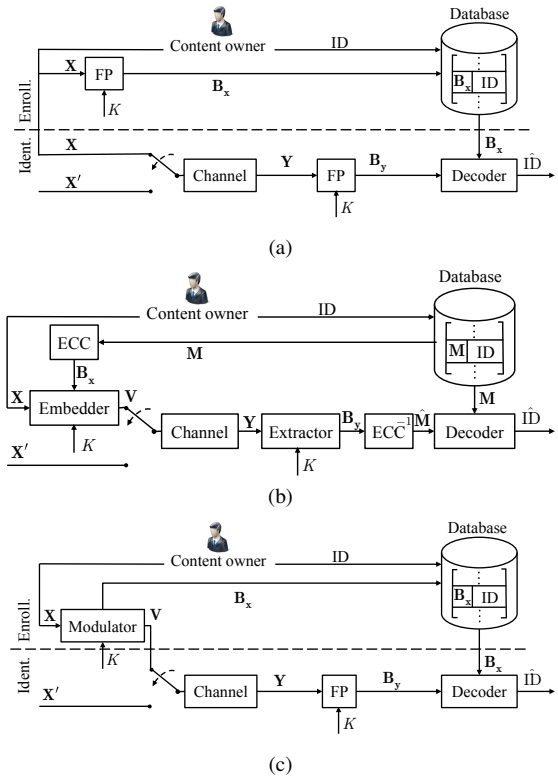

Fig. 1. Generalized models of (a) passive content fingerprinting, (b) digital watermarking and (c) active content fingerprinting.

II. A Common Basis for Secret Transform Domain Systems

Since most fingerprinting and digital watermarking systems operate in some transform domain, we define the direct and inverse transforms as:

$$\begin{cases} \tilde{\mathbf{x}} = \mathbb{W}\mathbf{x}, \\ \mathbf{x} = \mathbb{W}^{-1}\tilde{\mathbf{x}}, \end{cases} \tag{1}$$

for the orthonormal matrix $\mathbb{W}^{-1} = \mathbb{W}^T$.

We will assume that the matrix $\mathbb{W} \in \mathcal{R}^{N\times N}$, $\mathbb{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N)^T$ consists of a set of basis vectors $\mathbf{w}_i \in \mathcal{R}^N$ with $1 \le i \le N$. In the part of theoretical analysis, we will assume that this transform is based on any randomized orthogonal matrix $\mathbb{W}$ (random projection transform) whose elements $w_{i,j}$ are equal likely $\pm\frac{1}{\sqrt{N}}$ based on the secret key k. Such a matrix can be considered as an almost *orthoprojector*, for which $\mathbb{W}\mathbb{W}^T \approx \mathbb{I}_N$, and the basis vectors are asymptotically of a unit norm [1].[1]

In this paper, we will use an alternative representation of direct and inverse transforms (1):

$$\begin{cases} \tilde{x}_i = \mathbf{w}_i^T\mathbf{x}, \ 1 \le i \le N, \\ \mathbf{x} = \sum_{i=1}^N \tilde{x}_i\mathbf{w}_i = \sum_{i\in\mathcal{K}} \tilde{x}_i\mathbf{w}_i + \sum_{i\notin\mathcal{K}} \tilde{x}_i\mathbf{w}_i, \end{cases} \tag{2}$$

[1]The results also hold for other orthogonal matrices $\mathbb{W}$.

where a set $\mathcal{K} = \{i_1, i_2, \cdots, i_L\}$ represents a set of indices defined by the secret key k. This representation corresponds to the direct generation of a set of secret basis vectors or carriers $\mathbf{w}_i$, $i \in \mathcal{K}$, with unit norm $\|\mathbf{w}_i\|^2 = 1$, where the fingerprint is computed. This is also closely related to the digital watermarking techniques based on *spread spectrum* (SS), *spread transform* (ST) watermarking [7] and *subspace projections* (SSP) [8].

The selection of secret carriers is very important for both the robustness of the scheme and its security as well as statistics of projected coefficients [1].

III. Conventional passive content fingerprinting

A. Model of Passive Content Fingerprinting

In case of the conventional pCFP, the original content $\mathbf{x}$ is not modified and the fingerprint is computed directly from $\mathbf{x}$ using the projections onto a set of secret vectors $\mathbf{w}_i$, $i \in \mathcal{K}$ and quantization resulting into[2]

$$b_{x_i} = \text{sign}(\mathbf{w}_i^T \mathbf{x}) = \text{sign}(\tilde{x}_i), \tag{3}$$

where $\tilde{x}_i = \mathbf{w}_i^T \mathbf{x}$ and $\text{sign}(x) = +1$ for $x \geq 0$ and -1, otherwise. This process is distortionless and $\mathbf{x}$ remains intact.

Assuming an additive noise observation channel [3]:

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}, \tag{4}$$

where $\mathbf{Z}$ denotes the channel distortion, one is interested in estimating the level of degradations to the fingerprint extracted from the degraded content $\mathbf{Y}$.

For the comparison reasons we will consider the performance of all methods under the additive white Gaussian noise.

To quantify the level of signal distortions we will use the *document-to-noise ratio* (DNR) defined as:

$$\text{DNR} = 10\log_{10}\left(\frac{\frac{1}{N}E[\|\mathbf{X}\|_2^2]}{\frac{1}{N}E[\|\mathbf{Z}\|_2^2]}\right) = 10\log_{10}\left(\frac{\sigma_X^2}{\sigma_Z^2}\right), \tag{5}$$

where $\|\cdot\|_2$ stands for Euclidean norm, and we assumed that all signals are zero-mean Gaussian vectors, i.e., $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbb{I}_N)$ with the variance σ_X^2 and noise is also zero-mean Gaussian, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_Z^2 \mathbb{I}_N)$ with the variance σ_Z^2, where $\mathbb{I}_N$ is a unit matrix of size $N \times N$.

Given a secret key k and the corresponding set of secret carriers, the query fingerprint extraction is performed as[4]:

$$b_{y_i} = \text{sign}(\mathbf{w}_i^T \mathbf{y}) = \text{sign}(\tilde{x}_i + \tilde{z}_i),\ i \in \mathcal{K} \tag{6}$$

where $\tilde{z}_i = \mathbf{z}^T \mathbf{w}_i$.

The projected original content and noise coefficients are distributed as $\tilde{X} \sim \mathcal{N}(0, \sigma_X^2)$ and $\tilde{Z} \sim \mathcal{N}(0, \sigma_Z^2)$.

[2]In [1], it is shown that pCFP based on random projections and quantization asymptotically can approach theoretical performance limits.

[3]The additive model of distortions can be assumed for some robust feature extraction domains including SIFT descriptors, where the Euclidean metric or the Mahalanobis distance, which are the ML counterparts for the i.i.d. additive and correlated Gaussian noise, respectively, are used for the descriptor matching [9].

[4]In this work we do not consider *soft fingerprinting* where the extracted fingerprint also consists of additional information about bit reliabilities.

The performance of pCFP in terms of probability of bit error rate (BER) is measured as [10]:

$$P_{b-\text{pCFP}} = \Pr\{B_{\mathbf{x}} \neq B_{\mathbf{y}}\} \tag{7}$$

$$= E_{f(|\tilde{x}|)}\left[\text{Q}\left(\frac{|\tilde{X}|}{\sigma_Z}\right)\right] = \frac{1}{\pi}\arctan\left(\frac{\sigma_Z}{\sigma_X}\right), \tag{8}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{u^2}{2}}du$ indicates Q-function.

IV. Active content fingerprinting

In the aCFP, we will consider the modification of the original content in the space of secret carriers defined by the key k with the overall goal to minimize the BER defined according to (7). For this purpose, we define a general form of aCFP modulation:

$$\mathbf{v} = \sum_{i=1}^{N} \varphi_i(\tilde{x}_i)\mathbf{w}_i = \sum_{i \in \mathcal{K}} \varphi_i(\tilde{x}_i)\mathbf{w}_i + \sum_{i \notin \mathcal{K}} \tilde{x}_i \mathbf{w}_i, \tag{9}$$

where $\varphi_i(.)$ denotes a modulation function applied to the ith coefficient belonging to the set of secret carriers. In the following, we will assume that the same function $\varphi(.)$ is applied to all coefficients in the set of secret carriers.

Since the BER in (8) is defined by the magnitudes of projected coefficients $|\tilde{x}|$, one possible modulation is to increase these magnitudes limited by the defined distortion constraint.

Definition: A *distortion measure per dimension* between sequences $\mathbf{x}$ and $\mathbf{v}$ is defined by:

$$D = \frac{1}{N}E\left[\|\mathbf{V} - \mathbf{X}\|_2^2\right] = \frac{1}{N}E\left[\|\mathbf{S}\|_2^2\right], \tag{10}$$

where $\mathbf{S} = \sum_{i \in \mathcal{K}}\left(\varphi(\tilde{X}_i) - \tilde{X}_i\right)\mathbf{w}_i$ denotes the modulation signal that can be considered as a sort of a watermark by analogy to the digital watermarking. In the case of the pCFP, $D = 0$ while the aCFP will be characterized by the distortion determined by the modulation function $\varphi(.)$.

For coherence with digital watermarking, we will also define the *document-to-watermark ratio* (DWR):

$$\text{DWR} = 10\log_{10}\left(\frac{\frac{1}{N}E[\|\mathbf{X}\|_2^2]}{\frac{1}{N}E[\|\mathbf{S}\|_2^2]}\right) = 10\log_{10}\left(\frac{\sigma_X^2}{D}\right), \tag{11}$$

which should reflect the fact of content modification by the embedded "watermark" or actually the modulation signal $\mathbf{S}$. Obviously, there is an important difference between the watermark, which carries out the information about the content owner, and the modulation signal, which is solely used for the BER reduction. Therefore, it would be more correct to use the term *document-to-modulation signal ratio*. However, to make the comparison with the digital watermarking consistent, we will use the DWR assuming that the reader can clearly identify the difference between both techniques.

In following, we consider several modulation strategies in terms of their performance and distortions.

A. Additive active content fingerprinting

Definition: An *additive active content fingerprinting* (AddaCFP) is defined by the modulation function of the form:

$$\varphi_A(\tilde{x}_i) = \tilde{x}_i + \alpha\,\mathrm{sign}(\tilde{x}_i), \tag{12}$$

for $i \in \mathcal{K}$, where $\alpha > 0$ stands for the strength of aCFP. Substituting (12) into (9) yields:

$$\begin{aligned}\mathbf{v} &= \sum_{i\in\mathcal{K}} \left(\tilde{x}_i + \alpha\,\mathrm{sign}(\tilde{x}_i)\right)\mathbf{w}_i + \sum_{i\notin\mathcal{K}} \tilde{x}_i\mathbf{w}_i \\ &= \mathbf{x} + \alpha\sum_{i\in\mathcal{K}} \mathrm{sign}(\tilde{x}_i)\mathbf{w}_i = \mathbf{x} + \mathbf{s}_A,\end{aligned} \tag{13}$$

where $\mathbf{s}_A = \alpha\sum_{i\in\mathcal{K}}\mathrm{sign}(\tilde{x}_i)\mathbf{w}_i$.

The distortion of AddaCFP per content sample is:

$$\begin{aligned}D_A &= \frac{1}{N}E\left[\|\mathbf{S}_A\|_2^2\right] \\ &= \frac{1}{N}E\left[\alpha^2\left\|\sum_{i\in\mathcal{K}}\mathrm{sign}\left(\tilde{X}_i\right)\mathbf{w}_i\right\|_2^2\right] \overset{(a)}{=} \frac{L}{N}\alpha^2,\end{aligned} \tag{14}$$

where (a) follows from the fact that $\tilde{X}_i$ are i.i.d.. Consequentially, the DWR under the AddaCFP is $\mathrm{DWR}_{\mathrm{Add}} = 10\log_{10}\left(\frac{N}{L}\frac{\sigma_X^2}{\alpha^2}\right)$.

The fingerprint extraction for the enrollment can be performed based on $\mathbf{v}$ as:

$$b_{v_i} = \mathrm{sign}(\mathbf{w}_i^T\mathbf{v}) = \mathrm{sign}(\tilde{x}_i + \alpha\,\mathrm{sign}(\tilde{x}_i)),\ i \in \mathcal{K}.$$

The fingerprint computed at the verification stage is:

$$b_{y_i} = \mathrm{sign}(\mathbf{w}_i^T\mathbf{y}) = \mathrm{sign}(\tilde{x}_i + \alpha\,\mathrm{sign}(\tilde{x}_i) + \tilde{z}_i),\ i \in \mathcal{K}.$$

where $\mathbf{y} = \mathbf{v} + \mathbf{z}$ (Fig. 1c).

The performance of AddaCFP is determined by the BER and is given by:

$$\begin{aligned}P_{b-\mathrm{AddaCFP}} &= E_{f(\varphi_A(\tilde{X}))}\left[Q\left(\frac{|\varphi_A(\tilde{X})|}{\sigma_Z}\right)\right] \\ &\overset{(a)}{=} 2\int_0^\infty Q\left(\frac{\tilde{x}+\alpha}{\sigma_Z}\right)\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left(\frac{-\tilde{x}^2}{2\sigma_X^2}\right)\mathrm{d}\tilde{x}\end{aligned} \tag{15}$$

$$\overset{(b)}{\leq} \exp\left(-\frac{\alpha^2}{2\sigma_Z^2}\right)P_{b-\mathrm{pCFP}}, \tag{16}$$

where (a) results from $\varphi_A(\tilde{X})$ follows the following pdf:

$$f(\varphi_A(\tilde{x})) = \begin{cases}\mathcal{N}(\alpha,\sigma_X^2)\ \ ,\ \tilde{x} \geq \alpha, \\ \mathcal{N}(-\alpha,\sigma_X^2),\ \tilde{x} \leq -\alpha,\end{cases} \tag{17}$$

and (b) follows from the inequality $Q(x+t) \leq \exp(-\frac{t^2}{2})Q(x)$ for $x, t \geq 0$. The only difference between the BER of pCFP (8) and AddaCFP is in the positive bias α introduced by the additive modulation that reduces BER by at least the factor of $\exp(-\frac{\alpha^2}{2\sigma_Z^2})$.

Remark: (*Link to improved spread spectrum (ISS) watermarking*): The ISS is a watermarking method aiming at the host interference cancellation in the projected domain with the embedding rate $R_{\mathrm{DW}} = \frac{L}{N}$ bits [11][5]:

$$\mathbf{v} = \mathbf{x} + \nu\sum_{i\in\mathcal{K}}\mathbf{b}_{\mathbf{x}_i}\mathbf{w}_i - \lambda\sum_{i\in\mathcal{K}}\tilde{x}_i\mathbf{w}_i, \tag{18}$$

where ν and λ control the strength of the watermark and host cancellation, respectively, $b_{x_i} = (-1)^{m_i}$ and $m_i \in \{0,1\}$. The notation b_{x_i} is introduced by purpose to reflect the link with the extracted bits in the pCFP. Under the assumption of unit norm basis vectors, the embedding distortion is:

$$D = \frac{L}{N}\left(\nu^2 + \lambda^2\sigma_X^2\right). \tag{19}$$

It is not difficult to trace the link between the AddaCFP and ISS despite the different objectives behind both techniques. Since the AddaCFP does not target any data hiding, one can disregard the watermark embedding component by setting $\nu = 0$. Thus, the AddaCFP counterpart of ISS can be obtained as:

$$\mathbf{v} = \mathbf{x} + \lambda\sum_{i\in\mathcal{K}}\tilde{x}_i\mathbf{w}_i, \tag{20}$$

where the sign of interference cancellation is replaced by opposite to ensure host amplification.

The ISS-based modulation is only efficient for large values of $\tilde{x}_i$, while the coefficients close to zero do not obtain significant amplification. At the same time, these small-value coefficients represent the main source of bit errors due to the sign flipping. That is why contrary to the host interference cancellation in the digital watermarking, the AddaCFP increases these coefficients.

Finally, the BER of ISS corresponds to the mismatch of embedded and extracted bits and is defined as [11]:

$$P_{b-\mathrm{ISS}} = Q\left(\sqrt{\frac{(\alpha^2 - \lambda^2\sigma_X^2)}{\sigma_Z^2 + (1-\lambda^2)\sigma_X^2}}\right), \tag{21}$$

with the optimal $\lambda_{\mathrm{opt}} = \frac{1}{2}\left(1 + \frac{\sigma_Z^2}{\sigma_X^2} + \frac{\alpha^2}{\sigma_X^2}\right) - \frac{1}{2}\sqrt{\left(1 + \frac{\sigma_Z^2}{\sigma_X^2} + \frac{\alpha^2}{\sigma_X^2}\right)^2 - 4\frac{\alpha^2}{\sigma_X^2}}$ which minimizes the above BER for the embedding distortion $D_{\mathrm{ISS}} = D_{\mathrm{A}}$.

For the comparison purposes, we also consider SS-based watermarking which suffers from host interference and can be obtained from the ISS by assigning $\lambda = 0$ that results in:

$$P_{b-\mathrm{SS}} = Q\left(\sqrt{\frac{\alpha^2}{\sigma_Z^2 + \sigma_X^2}}\right). \tag{22}$$

B. Quantization-based active content fingerprinting

Definition: A *quantization-based active content fingerprinting* (QbaCFP) is defined by the modulation function of form:

$$\begin{aligned}\varphi_Q(\tilde{x}_i) &= c\,\mathrm{sign}(\tilde{x}_i) \\ &= \tilde{x}_i + c\,\mathrm{sign}(\tilde{x}_i) - \tilde{x}_i = \tilde{x}_i + (c - |\tilde{x}_i|)\mathrm{sign}(\tilde{x}_i).\end{aligned} \tag{23}$$

Substituting (23) into (9) yields:

$$\begin{aligned}\mathbf{v} &= \sum_{i\in\mathcal{K}}(\tilde{x}_i + (c - |\tilde{x}_i|)\mathrm{sign}(\tilde{x}_i))\mathbf{w}_i + \sum_{i\notin\mathcal{K}}\tilde{x}_i\mathbf{w}_i \\ &= \mathbf{x} + \sum_{i\in\mathcal{K}}(c - |\tilde{x}_i|)\mathrm{sign}(\tilde{x}_i)\mathbf{w}_i.\end{aligned} \tag{24}$$

[5]The presented multi-bit formulation of ISS has two differences with the originally proposed one-bit ISS: the third term is not normalized by $\|\mathbf{w}\|^2$ and L-bit embedding is considered.

The distortion of QbaCFP is:

$$D_Q = \frac{1}{N} E\left[\left\| -\sum_{i\in\mathcal{K}} \left(c - |\tilde{X}_i|\right) \operatorname{sign}(\tilde{X}_i)\mathbf{w}_i \right\|_2^2\right] = \frac{L}{N} E\left[\left(|\tilde{X}| - c\right)^2\right] \overset{(a)}{=} \frac{L}{N}\left(\sigma_X^2 - 2c\sigma_X\sqrt{\frac{2}{\pi}} + c^2\right), \tag{25}$$

where (a) follows from $E\left[|\tilde{X}|\right] = \sigma_X\sqrt{\frac{2}{\pi}}$ for the half-normal distribution.

Finally, the BER of QbaCFP is given by:

$$P_{b-\text{QbaCFP}} = E_{f(\varphi_Q(\tilde{X}))}\left[Q\left(\frac{|\varphi_Q(\tilde{X})|}{\sigma_Z}\right)\right] \overset{(a)}{=} Q\left(\frac{c}{\sigma_Z}\right), \tag{26}$$

where (a) follows from:

$$f(\varphi_Q(\tilde{x})) = \begin{cases} \frac{1}{2}\delta(\tilde{x} - c), & \tilde{x} \geq 0, \\ \frac{1}{2}\delta(\tilde{x} + c), & \tilde{x} < 0. \end{cases} \tag{27}$$

For the simple comparison of the QbaCFP with the AddaCFP, ISS and SS, we will make their average distortions equal by forcing $D_Q = D_A = \frac{L}{N}\alpha^2$ that results in:

$$P_{b-\text{QbaCFP}} = Q\left(\frac{\eta}{\sigma_Z}\right), \tag{28}$$

where $\eta = \sigma_X\sqrt{\frac{2}{\pi}}\left(1 + \sqrt{1 + \frac{\pi}{2}\left(\frac{\alpha^2}{\sigma_X^2} - 1\right)}\right)$.

It is also well known in the rate-distortion theory that the reconstruction level c that minimizes the distortion of one-bit scalar quantizer, which corresponds to the modulation function of QbaCFP, is equal to the mean value of the region, i.e., $c = E\left[|\tilde{X}|\right] = \sigma_X\sqrt{\frac{2}{\pi}}$ that yields:

$$D_Q = \sigma_X^2 \frac{L}{N}\left(1 - \frac{2}{\pi}\right), \tag{29}$$

$$P_{b-\text{QbaCFP}} = Q\left(\sqrt{\frac{2}{\pi}}\frac{\sigma_X}{\sigma_Z}\right). \tag{30}$$

Finally, the maximum achievable DWR under the QbaCFP, which corresponds to the minimum distortion (29), is $\text{DWR}_Q \leq 10\log_{10}\frac{N}{L\left(1-\frac{2}{\pi}\right)}$.

Remark: (*Link to spread-transform dither modulation (ST-DM) watermarking*): The ST-DM is a hybrid watermarking method aiming at the host interference cancellation in the projected domain, which combines a quantization (binning) strategy with spread transform with the embedding rate $R_{\text{DW}} = \frac{L}{N}$ bits [7]:

$$\mathbf{v}{=}\mathbf{x} + \mu\sum_{i\in\mathcal{K}}\left(\mathcal{Q}_{m_i}(\tilde{x}_i) - \tilde{x}_i\right)\mathbf{w}_i, \tag{31}$$

where $0 \leqslant \mu \leqslant 1$ is a distortion compensation parameter, and $\mathcal{Q}_{m_i}(.)$ is a scalar quantizer, which is defined by the bit m_i with the centroids defined by[6]:

$$c_i{=}\Delta\mathcal{Z} + (-1)^{m_i}\frac{\Delta}{4}, \text{for } m_i = 0, 1. \tag{32}$$

[6]We consider only binary embedding here.

TABLE I
COMPARISON OF DIGITAL WATERMARKING, PCFP AND ACFP METHODS.

	Technique	Distortion D	Probability P_b
Watermarking	SS	$\frac{L}{N}\alpha^2$	$Q\left(\frac{\alpha}{\sqrt{\sigma_X^2+\sigma_Z^2}}\right)$
	ISS	$\frac{L}{N}\alpha^2$	$Q\left(\sqrt{\frac{(\alpha^2-\lambda^2\sigma_X^2)}{\sigma_Z^2+(1-\lambda^2)\sigma_X^2}}\right)$
	Lower bound (LB)	$\frac{L}{N}\alpha^2$	$Q\left(\frac{\alpha}{\sigma_Z}\right)$
Fingerprinting	pCFP	0	$\frac{1}{\pi}\arctan\left(\frac{\sigma_Z}{\sigma_X}\right)$
	AddaCFP	$\frac{L}{N}\alpha^2$	$\text{E}\left[Q\left(\frac{\lvert\tilde{X}\rvert+\alpha}{\sigma_Z}\right)\right]$
	QbaCFP	$\frac{L}{N}\alpha^2$	$Q\left(\frac{\eta}{\sigma_Z}\right)$

Rewriting (24) in the form of (31) by introducing the distortion compensation parameter α, one obtains:

$$\mathbf{v}{=}\mathbf{x} + \alpha\sum_{i\in\mathcal{K}}(c\text{sign}(\tilde{x}_i) - \tilde{x}_i)\mathbf{w}_i, \tag{33}$$

with the remarkable correspondence in part of quantizations $\mathcal{Q}_{m_i}(.)$ and $c$sign(.). The fundamental difference between the ST-DM and QbaCFP consists in the absence of periodical structure of ST-DM quantizer depending on the message bit m_i, which should compensate the interference with the host signal. In the case of the QbaCFP similarly to the AddaCFP, one is only interested to increase the magnitudes of small components in the set $\mathcal{K}$ that is simply achieved by the quantization. Obviously, the cost for this simplicity are the distortions of all components whose values are larger than the centroid c and their decrease that might contribute the increase of the probability of bit error. Therefore, one might imagine more advanced modulation strategies that quantize the low-magnitude coefficients to some prettified levels while preserving the large-magnitude coefficients. This should provide an additional gain in the introduced distortion for the same robustness to the degradations. One can also assume multi-level quantization $Q(\tilde{x}_i)$ instead of one-bit scalar quantizer $c$sign$(\tilde{x}_i)$. It should be noted that $Q(\tilde{x}_i)$ does not depend on m_i as in (31). We leave this line of research out of scope of this paper targeting here only the introduction of basic principles and advantages of basic aCFP methods in light of existing pCFP and digital watermarking.

The BER of the ST-DM can be found in [4]. There are also other modifications of this basic scheme [8]. However, instead of considering all of them, that is obviously out of scope of this paper, we will provide a lower bound on the performance of both spread transform and quantization methods in the assumption of no host interference and binary modulation. This bound should serve us as a basis for the comparison with the aCFP methods.

Remark: (*BER lower bound on all watermarking methods*): The BER lower bound on all watermarking techniques in assumption of host interference absence and binary embedding is defined as:

$$P_{b-\text{LB}} = Q\left(\frac{\alpha}{\sigma_Z}\right). \tag{34}$$

To exemplify the effect of aCFP-based modulation, we have

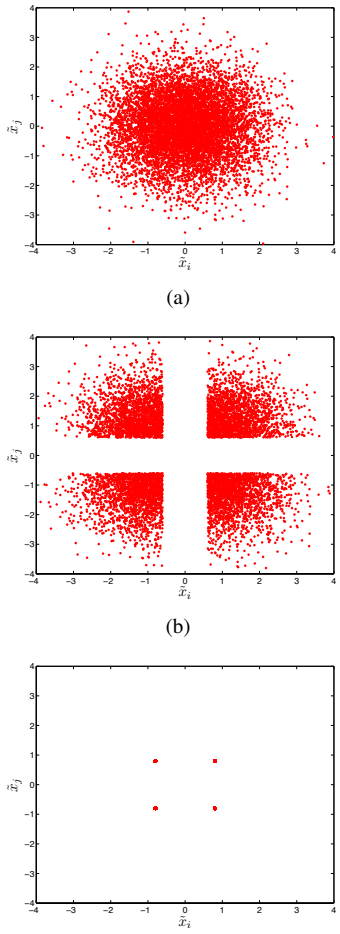

Fig. 2. PDFs of projected coefficients: (a) original, (b) AddaCFP and (c) QbaCFP for DWR=17dB.

performed the experiment with 1,000,000 Gaussian sequences of length $N = 2048$ that have been projected to the secret subspace with $L = 32$. The distribution of two such projections $\tilde{x}_i$ and $\tilde{x}_j$ is shown in Fig. 2a. The resulting projected coefficients have been modulated according to AddaCFP and QbaCFP, which are shown in Figs. 2b and 2c, respectively. The AddaCFP (13) produces a bias to all coefficients proportional to α while the QbaCFP (24) quantizes each coefficient based on one-bit scalar quantizer to the level c depending on the sign.

Finally, the considered techniques are summarized in Table I and Fig. 3 in terms of their BERs for the same embedding distortion besides the pCFP which has $D = 0$. These results are confirmed by the simulation on synthetic Gaussian sequences up to level of precision ensured by 100,000 sequences observed in 100 noise realizations. Both AdaCFP and QbaCFP considerably outperform pCFP and digital watermarking in terms of P_b. The reduction of P_b will have significant impact on both complexity and security of aCFP methods [10].

V. Conclusions

In this paper we intended to introduce the concept of aCFP and to consider its two practical implementations based on

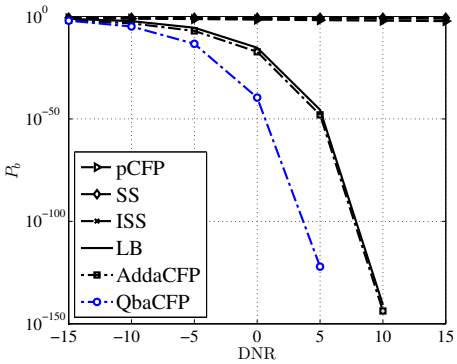

Fig. 3. Comparison of BERs in watermarking and fingerprinting approaches, for $\sigma_X^2 = 1$, $N = 2048$, $L = 32$ and DWR = 20dB.

additive and quantization methods.

It should be pointed out that besides its remarkable performance the aCFP does not intent to replace the digital watermarking in the copyright protection applications. However, it could be considered as a reasonable alternative in those applications that require content related management, tracking, tracing and monitoring.

In future, we will extend our analysis to consider security of aCFP and enhanced search strategies as proposed in [10].

Acknowledgment

This paper was partially supported by SNF projects 200020-134595.

References

[1] F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, "Performance analysis of content-based identification using constrained list-based decoding," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 5, pp. 1652 –1667, oct. 2012.

[2] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. Morgan Kaufmann Publishers (Academic Press), 2002.

[3] L. Pérez-Freire and F. Pérez-González, "Spread spectrum watermarking security," *IEEE Trans. on Information Forensics and Security*, vol. 4, no. 2–24, pp. 969–978, March 2009.

[4] F. Pérez-González, F. Balado, and J. R. Hernández, "Performance analysis of existing and new methods for data hiding with known-host information in additive channels," *IEEE Trans. on Signal Processing*, vol. 51, no. 4, April 2003.

[5] S. Gel'fand and M. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[6] O. Koval, S. Voloshynovskiy, P. Bas, and F. Cayre, "On security threats for robust perceptual hashing," in *Proceedings of SPIE Photonics West, Electronic Imaging / Media Forensics and Security XI*, San Jose, USA, 2009.

[7] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory.*, vol. 47, pp. 1423–1443, May 2001.

[8] R. F. H. Fischer and R. Bäuml, "Lattice costa schemes using subspace projection for digital watermarking," *European Trans. Telecommunications*, vol. 15, pp. 351–362, 2004.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, November 2004.

[10] S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak, "Information-theoretical analysis of private content identification," in *IEEE Information Theory Workshop, ITW2010*, Dublin, Ireland, Aug.30-Sep.3 2010.

[11] H. S. Malvar and D. A. F. Florncio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. on Signal Processing*, vol. 51, no. 4, pp. 898–905, APRIL 2003.