Improved robust watermarking through attack characterization

Deepa Kundur and Dimitrios Hatzinakos

Department of Electrical & Computer Engineering, University of Toronto
10 King's College Road, Toronto Ontario Canada M5S 3G4

deepa@comm.toronto.edu, dimitris@comm.toronto.edu

Abstract: In this paper, we propose an approach to improve the performance of a broad class of watermarking schemes through attack characterization. Robust and reference watermarks are both embedded into a signal. The reference watermark is used to characterize any modifications of the resulting marked signal, so that the robust watermark can be more reliably extracted. Analysis and simulations are provided to demonstrate the effectiveness of the approach.

OCIS codes: (100.2000) Digital image processing; (999.9999) Watermarking

References

1. N. Nikolaidis and I. Pitas, "Robust Image Watermarking in the Spatial Domain," Signal Process. **66**, 385-403 (1998).
2. J. F. Delaigle, C. De Vleeschouwer and B. Macq, "Watermarking Algorithm Based on a Human Visual Model," Signal Process. **66**, 319-335 (1998).
3. C. I. Podilchuk and W. Zeng, "Image-Adaptive Watermarking using Visual Models," IEEE J. Sel. Area in Commun. **16(4)**, 525-539 (1998).
4. X.-G. Xia, C. G. Boncelet and G. R. Arce, "A Multiresolution Watermark for Digital Images," Proc. IEEE Int. Conf. on Image Processing **1**, 548-551 (1997).
5. G. W. Braudaway, "Protecting Publicly-Available Images with an Invisible Image Watermark," Proc. IEEE Int. Conf. on Image Processing **1**, 524–527 (1997).
6. D. Kundur and D. Hatzinakos, "Digital Watermarking using Multiresolution Wavelet Decomposition," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing **5**, 2969-2972 (1998).
7. M. D. Swanson, M. Kobayashi and A. H. Tewfik, "Multimedia Data-Embedding and Watermarking Technologies," Proceedings of the IEEE **86(6)**, 1064-1087 (1998).
8. J. R. Hernández, F. Pérez-González and J. M. Rodriguez, "The Impact of Channel Coding on the Performance of Spatial Watermarking for Copyright Protection," Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing **5**, 2973–2976 (1998).
9. D. Kundur and D. Hatzinakos, "Towards a Telltale Watermarking Technique for Tamper-Proofing," Proc. IEEE Int. Conf. on Image Processing, **2**, 409–413 (1998).
10. D. Kundur and D. Hatzinakos, "Semi-Blind Image Restoration Based on Telltale Watermarking," to appear in Proc. 32nd Asilomar Conference on Signals, Systems, and Computers, (1998).

1. Introduction

Digital watermarking is the process by which a discrete data stream called a *watermark* is hidden within a multimedia signal by imposing imperceptible changes on the signal. In many proposed techniques this procedure entails the use of a secret key which must be used to successfully embed and extract the watermark. Watermarking has gained interest in applications involving the security of multimedia signals. One major driving force for research in this area is the need for effective copyright protection scenarios for digital imagery, sound and video. In such an application a serial number is watermarked into the signal to protect to mark ownership. It is expected that an *attacker* will attempt to remove the watermark by intentionally modifying the watermarked signal. Thus, we must strive to embed the mark such that it is difficult to remove (without the use of the key) unless the marked signal is significantly distorted. Our goal, therefore, parallels

#7118 - $15.00 US Received November 06, 1998; Revised December 04, 1998

that of cryptography in that we attempt to make the cost of watermark removal by an attacker much greater than the value of the multimedia signal itself. This problem is commonly called *robust* watermarking.

Various techniques for robust watermarking have been proposed in the literature [1,2,3,4,5,6 (Please note that this is not an exhaustive list)]. However, studies have shown that there still exists a need to improve reliability [7]. Previously proposed research has concentrated on sophisticated embedding strategies. In addition, the potential of error-correction codes has been assessed [8]. In this paper we consider the importance of a watermark extraction stage which makes use of information concerning the attacker's actions. To the best of the authors' knowledge this is the first paper which formally addresses the importance of characterizing the attacker's transformations on the marked signal to optimally estimate the watermark. By optimal we mean that the probability of bit error for watermark extraction is minimized. Our approach can be easily applied to a broad class watermarking methods to improve reliability. We assume that the original *host* signal is not available for watermark extraction and that the watermark is a binary data stream which is repeatedly embedded throughout the signal.

The contributions of this paper include 1) the incorporation of *reference* watermarking for attack identification prior to robust watermark extraction, 2) the use of a localized binary symmetric channel model to characterize the watermark attacks, and 3) the design of a weighted linear receiver structure which minimizes the probability of bit error for watermark extraction. We also demonstrate the improved performance of our technique through simulation results.

In the next section we briefly discuss the main elements to our novel approach for improved robust watermarking. Section 3. provides analytic results to show how the proposed scenario theoretically improves watermark extraction reliability. Simulation results are presented in Section 4., followed by concluding remarks in Section 5.

2. Robust Watermarking through Attack Characterization

2.1 Overview

In digital watermarking a host signal is transformed to a *watermark domain* in which modifications are imposed on the domain coefficients to embed the watermark. The modified coefficients are then inverse transformed to produce the marked signal[1]. Our proposed approach to improved robust watermarking is applicable to the general class of watermarking methods with the following basic properties:

- The watermark data stream consists of binary elements.
- The host signal (which refers to the original multimedia signal before watermarking) is not available or exploited for watermark extraction.
- The entire watermark is repeatedly embedded throughout the signal and each repetition of the watermark is positioned in a distinct localized region of the watermark domain. We will discuss this later in greater detail.

Many of the proposed techniques such as [4,5,6] fit the above criteria. Hence, the approach we present in this paper can be incorporated into already existing/implemented algorithms to enhance performance. We highlight the novel concepts of our approach in the next few sections.

[1]Popular transforms found in the watermarking literature are the wavelet transform and the DCT. The definition does not preclude techniques which embed the watermark in the time/spatial domain as the transformation will reduce to the identity operator.

#7118 - $15.00 US Received November 06, 1998; Revised December 04, 1998

2.2 *Reference Watermarking for Channel Identification*

Robust watermarking techniques which do not make use the host signal for watermark extraction are more practical than their counterparts which exploit the host. The trade-off, however, is that the former class of methods are, on average, less robust as little information is available on how the marked signal has been modified. We propose an approach to characterize the attacks on the watermark without use of the host signal. We define an attack as any signal modification, intentional or otherwise, which is applied to the marked signal and which effects the reliability of the extracted watermark. It has been shown in [9,10] that elementary characteristics of the signal distortions are easily estimated using a reference watermark. A reference watermark is one which is embedded into a signal for the purpose of detecting signal distortions. In this paper we show the importance of such tamper modeling to robust watermark extraction.

We propose the scenario shown in Figure 1. The host signal is embedded with both robust and reference watermarks. The two kinds of watermarks are placed orthogonally so that they do not interfere with one another. The trade-off is that fewer repetitions of the robust watermark can be placed in the signal as a portion of the watermark "bandwidth" is consumed by the presence of the reference watermark. Each embedded repetition of the robust watermark sequence, which we denote w_i, $i = 1, 2, \ldots M$ (where M is the total number of repetitions), has an associated binary reference watermark sequence v_i, with the same statistical properties as w_i [2]. Figure 1(b) demonstrates the embedding procedure where each w_i is placed in a localized region denoted D_i of the watermark domain. The bits of w_i are alternated with those of v_i such that an attack on the marked signal will reflect in the same way statistically on both w_i and v_i. Thus, if we let $\hat{w}_i$ and $\hat{v}_i$ be the extracted versions of w_i and v_i after an attack, it is expected that the probability of bit error for $\hat{w}_i$ is equal to that for $\hat{v}_i$.

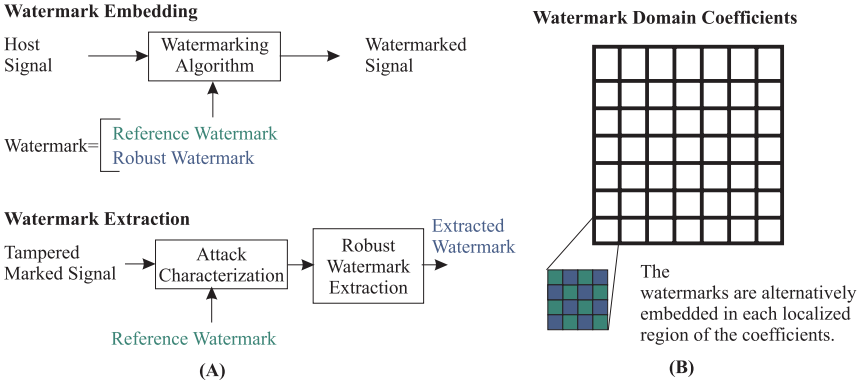

Figure 1. Combined Reference and Robust Watermarking for Channel Characterization and Reliable Watermark Extraction. **(A)** The watermark embedding and extracting scenarios, **(B)** We consider a 2-D host image. The watermark domain coefficients are divided into localized regions D_i (outlined with bold lines). Reference and robust watermark bits are alternatively embedded in each region.

The approach is similar to the concept of a "training sequence" or a "reference signal" used in digital communications in which a known data sequence is transmitted from the source to the destination to characterize the communications channel. We may consider watermarking to be analogous to digital communications in which each

[2]The reader is reminded that since w_i is a repetition of the robust watermark, $w_i = w_j$ for all i and j. The reference watermarks $\{v_i\}$ do not necessarily have to be identical as long as their individual bit elements have the same statistical properties as that of the robust watermark. Both $\{w_i\}$ and $\{v_i\}$ are generated using a pseudo-random number emulating the same probability distribution.

#7118 - $15.00 US
Received November 06, 1998; Revised December 04, 1998

repetition of the robust watermark in the marked signal has an associated channel, which we term a *watermark channel*. Proper identification of this channel will allow more accurate "transmission" (i.e., extraction) of the robust watermark as optimal processing may be incorporated at the receiver. The channel estimation is performed with the use of the reference watermark. In the next section we discuss the particular model of the channel we assume.

2.3 The Binary Symmetric Channel Model

Each watermark repetition w_i and its associated reference watermark v_i are embedded in the same localized region D_i of the watermark domain as shown in Figure 1(b). Assuming that the function used to transform the signal to the watermark domain is continuous, most degradations which maintain the perceptual quality of the signal will have a similar effect on both w_i and v_i. That is, we can assume that the degree of distortion experienced by both w_i and v_i due to an attack is the same; hence they have they same watermark channel.

We model the watermark channel for w_i and v_i as a binary symmetric channel (BSC) with probability of bit error p_{Ei}. Each bit of the embedded robust watermark $w_i(k)$, $k = 1, 2, \ldots, N$ (where N is the length of the watermark) is modeled as passing through a BSC to produce the corresponding extracted watermark bit $\hat{w}_i(k)$. We assume in our model that $0 \leq p_{Ei} \leq 0.5$. If $p_{Ei} > 0.5$ we merely complement the output and effectively use $0 \leq 1 - p_{Ei} < 0.5$ as the BSC parameter.

The reference watermark v_i is used to estimate the parameter p_{Ei} for each i. If we let N be the length of the binary stream v_i, and $\hat{v}_i$ be the corresponding extracted binary stream after an attack, we can approximate the probability of bit error for the watermark channel associated with D_i with

$$\hat{p}_{Ei} = \tfrac{1}{N}\textstyle\sum_{k=1}^{N} v_i(k) \oplus \hat{v}_i(k) \tag{1}$$

where $\oplus$ is the exclusive-OR operator and $v_i(k)$ and $\hat{v}_i(k)$ are the kth watermark bits of v_i and $\hat{v}_i$, respectively. It can be shown using the law of large numbers that the expected value of $\hat{p}_{Ei}$ is p_{Ei} and that the variance of estimate decreases for increasing N.

There are important advantages to using this model of the watermark channel. The model is simple and the parameter p_{Ei} is easy to accurately estimate using the associated reference watermark. In addition, a different parameter p_{Ei} for each w_i is incorporated which provides a localized assessment of the attack in the watermark domain. In most watermarking schemes, the extracted watermark repetitions $\hat{w}_i$ are averaged to produce the overall extracted watermark. Our attack characterization allows us to combine these repetitions based on a measure of their reliability to minimize the probability of watermark bit error. It should be emphasized that degradations such as filtering additive noise and lossy compression are reliably modeled using the BSC [9]. This characterization, however, is not appropriate for geometric transformations on the signal such as rotation and scaling.

2.4 A Weighted Receiver Structure for Watermark Extraction

In this section, we discuss how information about the BSC parameters can be used to obtain a more accurate estimate of the watermark information compared to simple averaging of the extracted repetitions. Our goal is to keep computational complexity low so we limit ourselves to linear estimation. The overall extracted watermark $\hat{w}$ is computed as the weighted sum of the individual extracted repetitions. That is,

$$\hat{w}(k) = \text{round}\left[\textstyle\sum_{i=1}^{M} \alpha_i \hat{w}_i(k)\right] \tag{2}$$

#7118 - $15.00 US Received November 06, 1998; Revised December 04, 1998

where $\hat{w}(k)$ and $\hat{w}_i(k)$ are the kth watermark bits of $\hat{w}$ and $\hat{w}_i$, respectively, and α_i is the associated scalar nonnegative weight dependent on p_{Ei} such that $\sum_{i=1}^{M} \alpha_i = 1$. The rounding operation makes sure that $\hat{w}$ is a binary data string comprised of zeros and ones. If the argument of round is 0.5, an arbitrary value of 0 or 1 is assigned. In any type of watermark attack, some regions in D_i are likely to undergo greater distortion than others. It is a direct advantage to be able to determine the regions which are less distorted and, hence, contain a more reliable watermark estimate. It is intuitively clear that a larger weighting for repetitions with a lower probability of bit error will improve the reliability of $\hat{w}$. In the next section we show how the following assignment for α_i minimizes the bit error of $\hat{w}$ to produce an optimal linear watermark extraction.

$$\alpha_i = \log\left(\frac{1-p_{Ei}}{p_{Ei}}\right) \Big/ \left(\sum_{j=1}^{M} \log\left(\frac{1-p_{Ej}}{p_{Ej}}\right)\right) \tag{3}$$

3. Weights for Optimal Linear Watermark Extraction

Let w be the embedded watermark of length N and let $\hat{w}_i$ represent the extracted bit stream corresponding to w_i. We denote the kth watermark bits of w and $\hat{w}_i$ as $w(k)$ and $\hat{w}(k)$, respectively. We define $b_i(k) \triangleq w(k) \oplus \hat{w}_i(k)$ which is equal to 1 if there is a bit error in the kth bit of the ith extracted watermark repetition and is 0 otherwise. Assuming the BSC model for the watermark channel, the $b_i(k)$ variables follow the Bernoulli random variable distribution with a probability of "success" (i.e., probability that $b_i(k) = 1$) of p_{Ei}. We also define the following variables used in our analysis: $\hat{\mathbf{w}}(\mathbf{k}) \triangleq [\hat{w}_1(k)\ \hat{w}_2(k)\ \cdots\ \hat{w}_M(k)]$ and $\mathbf{b}(\mathbf{k}) = [b_1(k)\ b_2(k)\ \cdots\ b_M(k)]$ for $k = 1, 2, \ldots, N$. Both $\hat{\mathbf{w}}(\mathbf{k})$ and $\mathbf{b}(\mathbf{k})$ are related by $b_i(k) \triangleq w(k) \oplus \hat{w}_i(k)$.

It can be shown that Equation 2 implies that there is no bit error in $\hat{\mathbf{w}}(\mathbf{k})$ if $\sum_{i=1}^{M} \alpha_i b_i(k) < 0.5$ (For simplicity, we let $\xi(\mathbf{b}(\mathbf{k})) \triangleq \sum_{i=1}^{M} \alpha_i b_i(k)$ for the remainder of the analysis.). In addition, Equation 2 with the constraints that $\sum_{i=1}^{M} \alpha_i = 1$ and $\alpha_i \geq 0$ imposes the restriction that $\xi(\mathbf{b}(\mathbf{k})) = 1 - \xi(\overline{\mathbf{b}(\mathbf{k})})$ where $\overline{(\cdot)}$ is the binary complement operator. This suggests that if the elements of $\hat{\mathbf{w}}(\mathbf{k})$ accurately estimate $\hat{w}(k)$ using Equation 2, then the elements of $\overline{\hat{\mathbf{w}}(\mathbf{k})}$ produce a bit error. To clarify this point we consider the sets A and $\tilde{A}$ in which $A \subset \{\hat{\mathbf{w}}(\mathbf{k}) | \xi(\mathbf{b}(\mathbf{k})) < 0.5\}$ and $\tilde{A} \subset \{\hat{\mathbf{w}}(\mathbf{k}) | \xi(\mathbf{b}(\mathbf{k})) > 0.5\}$. Any remaining complement pairs for which $\xi(\mathbf{b}(\mathbf{k})) = \xi(\overline{\mathbf{b}(\mathbf{k})}) = 0.5$ are arbitrarily distributed among A and $\tilde{A}$ such that their cardinalities are both 2^{M-1} and $\hat{\mathbf{w}}(\mathbf{k}) \in A$ if and only if $\overline{\hat{\mathbf{w}}(\mathbf{k})} \tilde{A}$.

The probability of bit error for $\hat{\mathbf{w}}(\mathbf{k})$ is given by $P_E(k) = \sum_{\hat{\mathbf{w}}(\mathbf{k}) \in \tilde{A}} P\{\hat{\mathbf{w}}(\mathbf{k})\}$ where $P\{\hat{\mathbf{w}}(\mathbf{k})\}$ is the probability of extracting the sequence $\hat{\mathbf{w}}(\mathbf{k})$ which is given by

$$P\{\hat{\mathbf{w}}(\mathbf{k})\} = \prod_{i=1}^{M} (1 - p_{Ei})^{\overline{b}_i(k)} {p_{Ei}}^{b_i(k)} \tag{4}$$

since $\hat{\mathbf{w}}(\mathbf{k})$ is an ordered set of Bernoulli random variables. It can be shown that the minimization of $P_E(k)$ is equivalent to selecting a set of weights $\{\alpha_i\}$ such that $\hat{\mathbf{w}}(\mathbf{k}) \in A$ implies that $P\{\hat{\mathbf{w}}(\mathbf{k})\} \geq P\{\overline{\hat{\mathbf{w}}(\mathbf{k})}\}$. Equivalently, for each complement pair $\{\hat{\mathbf{w}}(\mathbf{k}), \overline{\hat{\mathbf{w}}(\mathbf{k})}\}$, we want to place the element with the lower probability of occurrence in $\tilde{A}$ to minimize the overall bit error rate. It can be easily shown that a selection of α_i given in Equation 3 implies that $P\{\hat{\mathbf{w}}(\mathbf{k})\} \geq P\{\overline{\hat{\mathbf{w}}(\mathbf{k})}\}$ if and only if $\xi(\mathbf{b}(\mathbf{k})) \leq \xi(\overline{\mathbf{b}(\mathbf{k})})$ and thus $\overline{\hat{\mathbf{w}}(\mathbf{k})} \in \tilde{A}$ which suggests that $P_E(k)$ is minimized. The analysis is omitted for compactness.

4. Simulation Results

We apply our proposed approach to the wavelet-based watermarking technique proposed by the authors in [6]. Figure 2 displays the 256×256 pixel host and watermarked colour images used in our simulations. A 32 byte watermark, randomly generated with

#7118 - $15.00 US Received November 06, 1998; Revised December 04, 1998

a uniform probability distribution, is embedded in the luminance image component. The performance of optimal weighting during watermark extraction is compared to that of simple averaging. In each case, the correlation coefficient[3] of the extracted and the embedded watermarks were used to assess the robustness of the technique. The results for linear filtering with a radially symmetric blur of the form $h(m,n) = a^{\sqrt{m^2+n^2}}/K$ where $K = \sum_{\forall(m,n)} h(m,n)$ are displayed in Figure 3 for different values of a. Similar results are presented for JPEG compression. The pink line represents the correlation for the weighted extraction and the blue corresponds to averaging. In both cases a significant increase in the correlation coefficient is observed which indicates a higher accuracy in the extracted watermark. The weights $\{\alpha_i\}$ are calculated from Equation 3 using $\hat{p}_{Ei}$ from the reference watermark calculated by Equation 1.

Figure 2. The (a) host image and (b) watermarked image used for simulations.

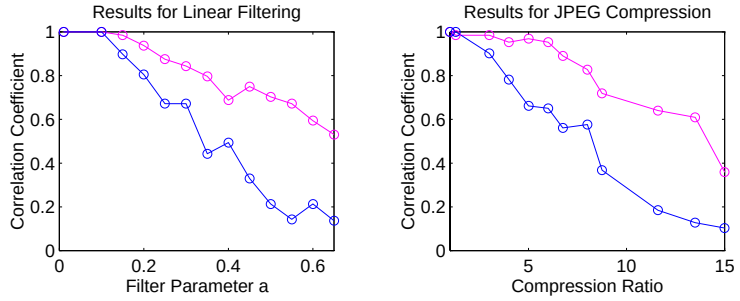

Figure 3. The improved performance of the proposed approach.

An improvement in performance was also noticed for other distortions such as median filtering. The only type of tampering for which little improvement was observed was that of additive white Gaussian noise. We believe that the whiteness of the noise made it difficult to predict its effects using the reference watermark.

5. Conclusion

In this paper we demonstrate how improved performance for robust watermarking can be achieved through assessment of attacker tampering. Watermark repetition throughout the signal provides diversity to combat a broad class of degradations. Characterization of the attacks can be used to optimally combine the extracted watermark repetitions to minimize the probability of error in watermark extraction. Future work involves extending the method to detect and identify geometric transformations on the marked signal to decrease the computational load required for watermark synchronization.

[3]The correlation coefficient of u and v is defined as $\sum_{\forall k} u(k)v(k)/(\sqrt{\sum_{\forall k} u^2(k)}\sqrt{\sum_{\forall k} v^2(k)})$.

#7118 - $15.00 US
Received November 06, 1998; Revised December 04, 1998