ATTACK CHARACTERIZATION FOR EFFECTIVE WATERMARKING

Deepa Kundur and Dimitrios Hatzinakos

10 King's College Road
Department of Electrical and Computer Engineering
University of Toronto,
Toronto, Ontario, Canada M5S 3G4
E-mail: {deepa,dimitris}@comm.toronto.edu

ABSTRACT

We present and analyze an approach to improve the performance of a broad class of watermarking schemes through attack characterization. Traditional robust watermarking methods use little information concerning the way in which the image is tampered to estimate the embedded watermark. In our novel scheme we propose adding two types of watermarks: reference and robust. The reference watermark is used to assess the way in which the marked image has been modified so that the robust watermark can be optimally extracted to maximize security. Analysis and simulation results are provided to demonstrate the significant performance improvement when the proposed approach is employed in an existing watermarking scheme.

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 which we strive to embed the mark robustly such that it is difficult to remove (without the use of the key) unless the marked signal is significantly distorted.

Previously proposed research has concentrated on sophisticated embedding strategies. In this paper we demonstrate the importance of a watermark extraction stage which characterizes tampering imposed on the image and which optimally extracts the watermark. We present an approach to improve the performance of a broad class of watermarking schemes through attack characterization. We go beyond our previous work [1] and formally analyze the proposed approach to relate the improvement in reliability to fundamental entropy concepts to obtain performance bounds. Subsequently, we derive general rules of thumb for effective watermark embedding. The significant performance improvement of our approach is also demonstrated through simulation results to verify our theoretical observations.

This work was supported by Communications and Information Technology Ontario (CITO).

2. ROBUST WATERMARKING THROUGH ATTACK CHARACTERIZATION

2.1. General 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. Our proposed approach to improved robust watermarking is applicable to the general class of watermarking methods with the following basic properties: 1) the watermark data stream consists of binary elements, 2) the host signal (which refers to the original multimedia signal before watermarking) is not available or exploited for watermark extraction, 3) the watermark is repeatedly embedded throughout the signal and each repetition of the watermark is positioned in a distinct localized region of the watermark domain. In many proposed watermarking techniques [2, 3, 4, 5] each repetition/segment of the watermark is separately extracted; the overall watermark estimate is calculated by averaging all of the repetitions or by using the one which produces the highest correlation coefficient with the embedded watermark, if known. Depending on the type of degradation, some repetitions of the watermark are more

severely distorted than others. We propose a technique which optimally weights the different repetitions before averaging to minimize the probability of watermark bit error. Such a weighting requires assessment of the tampering on the image, so that less distorted repetitions are given a higher weight.

There exists a parallelism between robust watermarking and communication through an imperfect channel. The watermarking process is likened to communicating a watermark signal through an associated *watermark channel*. Modifications of the watermarked signal (intentional or otherwise) impose and characterize the non-idealities of this associated watermark channel.

We implement our approach for improved robustness using attack characterization on an existing wavelet-based watermark method discussed in [5]. Watermark bits are embedded into a signal by quantizing the corresponding wavelet coefficients. We do not discuss the details of the technique here, but refer the reader to [5].

2.2. Robust and Reference Watermarks

The main distinction of our approach is that two types of watermarks, reference and robust, are embedded into the signal. Figure 1 provides an overview of the scenario. We define a reference watermark, which is assumed to be known at the extraction end, as one which is embedded into a signal for the purpose of detecting signal distortions. The robust watermark, which is not known at the extraction end, refers to the hidden information to be reliably extracted from the watermarked signal. Robust and reference watermarks are placed orthogonally so that they do not interfere with one another. Specifically, the watermarks are placed in separate host signal coefficients. Each embedded repetition of the robust watermark sequence, which we denote w_k, $k = 1, 2, \ldots M$ (where M is the total number of repetitions), has an associated N_w-bit binary reference watermark sequence r_k. Since w_k is the kth repetition of the robust watermark, $w_k = w_j$ for all k and j. The reference watermarks $\{r_k\}$ do not necessarily have to be identical, but for the implementation and simulations in this paper, we let $r_k = r_j$ for all k and j. Figure 1(b) demonstrates the embedding procedure where each w_k and r_k are placed in a localized region of the wavelet domain denoted R_k, where k is the index of the particular set of coefficients. This localized region is a rectangular spatial neighbourhood of coefficients at a given resolution level in the wavelet domain. For each resolution level, the corresponding spatial locations are segmented into non-overlapping blocks to represent these coefficient sets. The number of elements in R_k is equal to $2N_w$, so that one entire repetition of each of the robust and reference watermarks can be embedded in

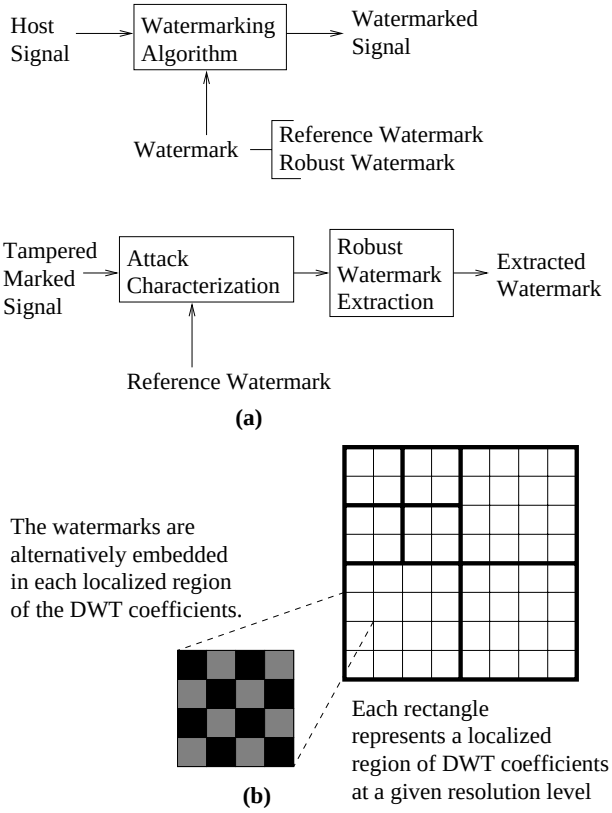

Figure 1: Combined Reference and Robust Watermarking for Channel Characterization and Reliable Watermark Extraction. (a) The watermark embedding and extracting scenarios. (b) For a two-dimensional host image, the watermark domain coefficients are divided into localized regions R_k (outlined with the thin lines) for each resolution level (shown in bold). Reference and robust watermark bits are alternatively embedded in each region.

the region. The bits of w_k are alternated with those of r_k in a checker board pattern such that an attack on the marked signal will reflect statistically in the same way on both w_k and r_k. Thus, if we let $\hat{w}_k$ and $\hat{r}_k$ be the extracted versions of w_k and r_k after an attack, we expect that the probability of bit error for $\hat{w}_k$ is equal to that for $\hat{r}_k$.

2.3. BSC Model of the Watermark Channel

We model the watermark channel for both w_k and r_k as a binary symmetric channel (BSC) with probability of bit error p_{Ek}. Each bit of the embedded robust watermark $w_k(i)$, $i = 1, 2, \ldots, N_w$, is modeled as passing through a BSC to produce the corresponding extracted watermark bit $\hat{w}_k(i)$. We assume in our model that $0 \leq p_{Ek} \leq 0.5$. If $p_{Ek} > 0.5$ we merely complement

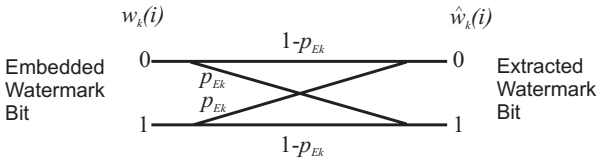

Figure 2: Binary Symmetric Channel Model. Each watermark repetition is considered to travel through a binary symmetric channel with probability of bit error p_{Ek}. The value of this parameter is estimated using the reference watermark.

the output and effectively use $0 \leq 1 - p_{Ek} < 0.5$ as the BSC parameter. Figure 2 provides an overview of the model.

The reference watermark r_k is used to estimate the parameter p_{Ek} for each k. If we let r_k be the corresponding extracted reference watermark after an attack, we can approximate the probability of bit error for the associated watermark channel by $\hat{p}_{Ek} = \frac{1}{N_w}\sum_{i=1}^{N_w} r_k(i) \oplus \hat{r}_k(i)$, where $\oplus$ is the exclusive-OR operator, and $r_k(i)$ and $\hat{r}_k(i)$ are the ith watermark bits of r_k and $\hat{r}_k$, respectively.

There are important advantages to using this model of the watermark channel. The model is simple and the parameter p_{Ek} is easy to accurately estimate using the associated reference watermark. In addition, a different parameter p_{Ek} for each w_k is incorporated which provides a localized assessment of the attack in the wavelet domain. In most watermarking schemes, the extracted watermark repetitions $\hat{w}_k$ 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 [6]. This characterization, however, is not appropriate for geometric transformations on the signal such as rotation and scaling.

2.4. An Effective Watermark Receiver

To keep computational complexity low, we limit ourselves to linear watermark estimation. The overall extracted watermark $\hat{w}$ is computed as the weighted sum of the individual extracted repetitions. That is,

$$\hat{w}(i) = \text{round}\left[\sum_{k=1}^{M} \alpha_k \hat{w}_k(i)\right] \tag{1}$$

where round$[\cdot]$ is the integer round operator, $\hat{w}(i)$ and $\hat{w}_k(i)$ are the ith watermark bits of $\hat{w}$ and $\hat{w}_k$, respectively, and α_k is the associated scalar nonnegative weight dependent on $\hat{p}_{Ek}$ such that $\sum_{k=1}^{M} \alpha_k = 1$. In any type of watermark attack, some regions R_k 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. We show in [1] that the following assignment for α_k minimizes the bit error of $\hat{w}$ to produce an optimal linear watermark extraction:

$$\alpha_k = \frac{\log\left(\frac{1-\hat{p}_{Ek}}{\hat{p}_{Ek}}\right)}{\left(\sum_{j=1}^{M} \log\left(\frac{1-\hat{p}_{Ej}}{\hat{p}_{Ej}}\right)\right)}. \tag{2}$$

3. ANALYSIS

We define the following:

$$e_k(i) \triangleq w(i) \oplus \hat{w}_k(i) \tag{3}$$
$$e(i) \triangleq w(i) \oplus \hat{w}(i). \tag{4}$$

If $e_k(i) = 1$, there is a bit error in the ith bit of the kth extracted watermark repetition. Equation (1) becomes

$$e(i) = \text{round}\left[\sum_{k=1}^{M} \alpha_k e_k(i)\right]. \tag{5}$$

We introduce the following *error statistic function*

$$\mathcal{E}(\mathbf{e}(i)) \triangleq \sum_{k=1}^{M} \alpha_k e_k(i). \tag{6}$$

A bit error in $\hat{w}(i)$ occurs if $\mathcal{E}(\mathbf{e}(i)) > 0.5$. We define the following bit error vector

$$\mathbf{e}(i) = [e_1(i)\ e_2(i)\ \cdots\ e_M(i)]^T. \tag{7}$$

From the independence of $e_k(i)$ for different k.

$$E\{\mathcal{E}(\mathbf{e}(i))\} = \frac{\sum_{k=1}^{M} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right) E\{e_k(i)\}}{\sum_{k=1}^{M} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)} \tag{8}$$
$$= \frac{\sum_{k=1}^{M} p_{Ek} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)}{\sum_{k=1}^{M} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)} \tag{9}$$

Using the fact that $\log\left(\frac{1-p_{Ek}}{p_{Ek}}\right) \geq (1-p_{Ek})\log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)$ for $p_{Ek} \leq 0.5$, we find that

$$E\{\mathcal{E}(\mathbf{e}(i))\} \leq \frac{\sum_{k=1}^{M} p_{Ek} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)}{\sum_{k=1}^{M} (1-p_{Ek}) \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)} \tag{10}$$

with equality if and only if and only if $p_{Ek} = 0$ for all k. Using the log-sum inequality [7] to the denominator, we can show

$$E\{\mathcal{E}(\mathbf{e}(i))\} \leq \frac{\sum_{k=1}^{M} p_{Ek} \log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)}{M(1-\overline{p}_E)\log\left(\frac{1-p_{Ek}}{p_{Ek}}\right)} \qquad (11)$$

where $\overline{p}_E = \frac{1}{M}\sum_{k=1}^{M} p_{Ek}$ with equality if and only if $p_{Ek} = 0$ for all k. The right hand side of (11) can be expanded, rearranged, factored and reduced to give

$$E\{\mathcal{E}(\mathbf{e}(i))\} \leq \frac{\overline{p}_E}{1-\overline{p}_E}\left[1 - \frac{D(q_a\|q_b)}{\log\left(\frac{1-\overline{p}_E}{\overline{p}_E}\right)}\right] \qquad (12)$$

where $D(q_a\|q_b)$ is the relative entropy measure [7], and $q_a(k) = p_{Ek}/(M\overline{p}_E)$ and $q_b(k) = (1-p_{Ek})/(M(1-\overline{p}_E))$.

4. EFFECTIVE EMBEDDING STRATEGIES TO COMBAT COMPRESSION

Analysis of (12) reveals that the following possible tactics may be incorporated into a watermarking scheme to lower the value of $E\{\mathcal{E}(\mathbf{e}(i))\}$ and, hence, improve the robustness of the watermarking system in some way:

1. Reduce the value of $\overline{p}_E$. Reducing the value of $\overline{p}_E$ decreases the term $\frac{\overline{p}_E}{1-\overline{p}_E}$ and increases the denominator term $\log\left(\frac{1-\overline{p}_E}{\overline{p}_E}\right)$ which both serve to lower the overall bound. Many proposed watermarking methods attempt to gain performance by diminishing this average bit error rate. Signal processing strategies to imperceptibly embed a higher energy and, hence, more robust watermark are commonly employed.

2. Embed the watermark such that the distributions q_a and q_b are dissimilar for a large class of distortions. Given a fixed value of $\overline{p}_E$, we may reduce the performance bound by increasing the value of $D(q_a\|q_b)$. The relative entropy $D(\cdot\|\cdot)$ is a measure of the distance between its two argument distributions [7]. Roughly, we can see that $D(q_a\|q_b)$ is large when $q_a(k) = p_{Ek}/(M\overline{p}_E)$ and $q_b(k) = (1-p_{Ek})/(M(1-\overline{p}_E))$ are dissimilar. Assuming a fixed $\overline{p}_E$, this requires that p_{Ek} vary in amplitude for different values of k, implying that we should embed the watermark in a domain for which the degree of distortion varies in each localized watermark domain coefficient region. This can be achieved by embedding the watermark in a domain which distributes the distortion more to certain coefficients, leaving others less effected.

3. Localize the distortions on the watermarked signal. The existence of $p_{Ek} = 0$ for at least one $k \in \{1, 2, \ldots, M\}$ implies that $\mathcal{E}(\mathbf{e}(i)) = 0$. Thus, if there exists a set of coefficients containing one watermark repetition which are unmodified by the distortion, then perfect watermark recovery is possible, even if $\overline{p}_E \neq 0$ as long as all $\{p_{Ek}\}$ are known. This translates to embedding the watermark in a domain which completely localizes the distortion to a few coefficients.

We discuss in [8] the implications of our analysis to watermarking in the presence of perceptual coding. It follows that complementary perceptual models should be used for watermarking and coding to increase robustness which is in direct contrast to other research in the area [9].

5. SIMULATIONS

We demonstrate the practical effectiveness of our approach through simulations. We compare watermark estimation through simple averaging with that of application of Equation (1). We employ our modified extraction stage in the watermarking technique of [5] for $\Delta = 3$, $L = 4$ and the Daubechies 10-point wavelet transform. The correlation coefficient [1] is used to compute the similarity between the embedded and extracted watermarks. Figure 3 demonstrates the improvement in correlation coefficient over simple averaging for robust reference watermarking using the 256×256 8-bit grey-scale tiger image used for the simulations in [1]. Randomly generated 256 bit binary robust and reference watermarks are embedded in the signal. Figure 3 presents the results for JPEG compression for varying compression ratio, additive white Gaussian noise for different signal-to-noise ratios, $M \times M$ median filtering for different M, and low pass filtering with filter $h(m,n) = a^{\sqrt{m^2+n^2}}/(\sum_{\forall m2,n2} a^{\sqrt{m2^2+n2^2}})$ for distinct values of a. The solid and dashed lines represents the correlation when optimal weights are employed and when simple averaging is applied, respectively. As we can see, application of reference watermarking to optimally weight the watermark repetitions significantly improves performance. Tests were also conducted to demonstrate that perfect watermark recovery is possible for cropping and localized spatial tampering of any form.

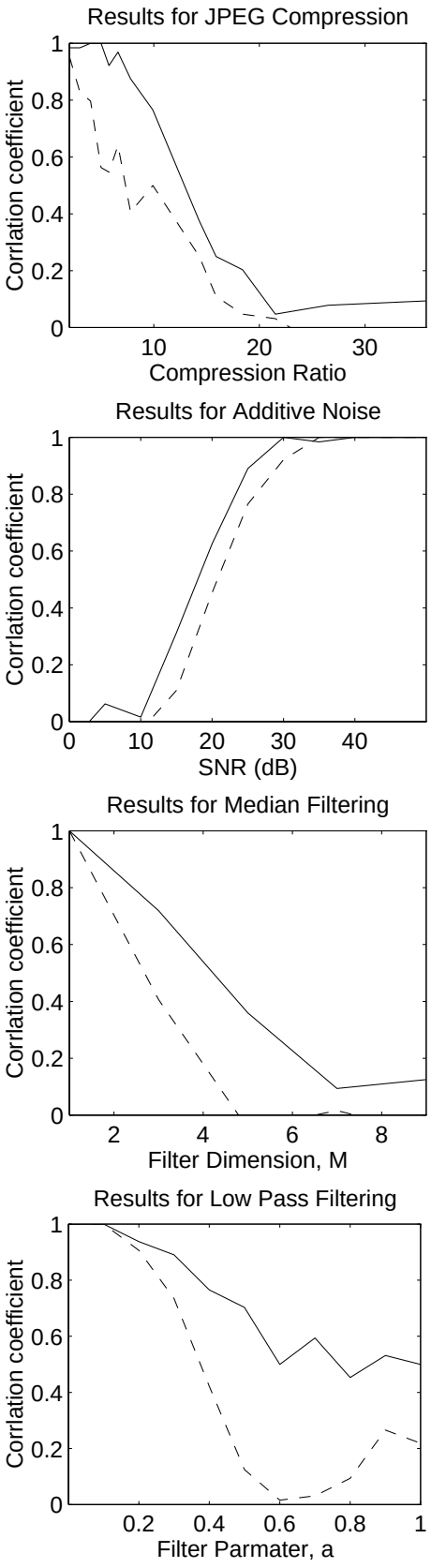

Figure 3: Results of the proposed approach.

6. CONCLUSIONS

This paper discusses a novel technique to improve the performance of a broad class of watermarking methods through attack characterization. Such an approach has the advantage that it improves robustness to a broad range of distortions by adaptively estimating the watermark. Analysis shows that embedding information in a domain which localizes any distortion on the watermarked signal results in a more robust watermark extraction.

7. REFERENCES

[1] D. Kundur and D. Hatzinakos. Improved robust watermarking through attack characterization. *Optics Express*, 3(12):485–490, December 7 1998.

[2] X.-G. Xia, C. G. Boncelet, and G. R. Arce. A multiresolution watermark for digital images. In *Proc. IEEE Int. Conference in Image Processing*, volume 1, pages 548–551, 1997.

[3] G. W. Braudaway. Protecting publicly-available images with an invisible image watermark. In *Proc. IEEE Int. Conference on Image Processing*, volume 1, pages 524–527, 1997.

[4] J. J. K. ÓRuanaidh, W. J. Dowling, and F. M. Boland. Phase watermarking of digital images. In *Proc. IEEE Int. Conference on Image Processing*, volume 3, pages 239–242, 1996.

[5] D. Kundur and D. Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2969–2972, 1998.

[6] D. Kundur and D. Hatzinakos. Towards a telltale watermarking technique for tamper-proofing. In *Proc. IEEE Int. Conf. on Image Processing*, volume 2, pages 409–413, 1998.

[7] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Toronto, 1991.

[8] D. Kundur and D. Hatzinakos. Mismatching perceptual models for effective watermarking in the presence of compression. In *Proc. SPIE, Multimedia Systems and Applications II*, September 1999.

[9] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp. The effect of matching watermark and compression transforms in compressed color images. In *Proc. IEEE Int. Conference in Image Processing*, volume 1, 1998.