Improved Digital Watermarking Through Diversity and Attack Characterization

Deepa Kundur

Department of Electrical & Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario
Canada M5S 3G4
Phone: (416) 946-5181
Fax: (416) 971-3020

deepa@comm.toronto.edu

ABSTRACT

In this work, we propose and evaluate the use of novel communication theory tool-sets to improve the performance of digital watermarking algorithms. The emphasis is on the application of basic channel estimation and communication diversity principles to the problem of robust data hiding in multimedia signals. An analytic framework is presented from which we derive general insights into strategies for effective watermarking.

KEYWORDS

Digital watermarking, content-based security, attack characterization, multimedia security, copy protection.

1 Introduction

Much of the previous work in digital watermarking has addressed the problem in a practical manner by presenting novel algorithms for a variety of applications. Improved performance was often gained through the use of tool-sets employed in communication theory.

In this work, we propose an analytic framework applicable to most general multimedia signals to understand effective embedding and extraction strategies for robust watermarking. This work is a preliminary investigation and is not intended, in its present form, to represent a complete methodology.

A great deal of the work on robust digital watermarking is based on spread spectrum (SS) principles [1-5]. In SS watermarking the embedded signal is generally a low energy pseudo-randomly generated white noise sequence. It is *detected* by correlating the known watermark sequence with either the extracted watermark or a transformed version of the watermarked signal itself (if the original *host* signal is not available for extraction). If the correlation factor is above a given threshold then the watermark is detected. The anti-jamming properties of SS signaling makes it attractive for application in watermarking since a low energy and hence imperceptible watermark, robust to narrow band interference, can be embedded [4].

However, SS approaches have the following limitations:

- SS allows detection of a known watermark, but the fundamentally large bandwidth requirement does not facilitate the *extraction* of a long bit sequence or logo from an audio signal or an image.
- SS approaches are specifically vulnerable to the "near-far" problem [6]. For watermarking this implies that if the energy of the watermark is reduced due to fading-like distortions on the watermark, any residual correlation between the host signal and watermark can result in unreliable detection.
- Most SS approaches are not adaptive. That is, they neither take into account spatial non-stationarity of the host signal and attack interference nor readily incorporate adaptive techniques to estimate the statistical variations.
- The correlator receiver structures used for watermark detection are not effective in the presence of fading. Although SS systems in general try to exploit spreading to average the fading, the techniques are not designed to maximize performance.

We also consider a communication paradigm to watermarking; communicating the watermark is analogous to transmission of the signal through an associated *watermark channel*. However, we hypothesize that common multimedia signal distortions including cropping, filtering, and perceptual coding are not accurately modeled as narrow band interference which is a common assumption in SS approaches. Instead, we believe that such signal modifications have the effect of fading on the embedded watermark. As a result, the watermark can be made more robust by employing effective diversity techniques and channel estimation. Previous work has demonstrated through practical implementation and simulations the improved performance of taking such an approach [7].

This paper provides analysis to demonstrate the advantages of incorporating these new principles for digital watermarking, and outlines approaches to improve algorithm performance to specific watermark attacks. It should be emphasized that the ideas presented in this work are meant to be employed within existing watermarking techniques, and are not intended to replace well-established watermarking strategies such as SS watermarking and modulation.

2 Context and Scope

We incorporate diversity and channel estimation into our analysis framework through the use of watermark repetition and attack characterization. In particular, we assume that the watermark is embedded many times throughout the host signal. Each repetition is assumed to be separately extracted. Attack characterization is the process of measuring the reliability of each extracted watermark repetition. We do not specify how the characterization is performed as this is an implementation issue, but assume a reliability factor is available. To perform analysis we limit the scope of our framework to the broad class of watermarking systems with the following basic characteristics:

1. The watermark w is binary and of length N_w bits.
2. The watermark information is repeatedly embedded $M \geq 1$ times within the host signal.
3. The embedding process occurs in the *watermark domain*. Specifically, an invertible transformation $\boldsymbol{T}_w$ is applied to the host signal to produce coefficients in which the watermark bits are repeatedly inserted.
4. Each repetition of the watermark is embedded in a localized region of coefficients in the watermark domain, so that most traditional distortions on multimedia signals such as perceptual coding and filtering will have a similar degree of distortion on all embedded watermark bits of a given repetition.
5. Each embedded watermark repetition is extracted separately.
6. The watermarked signal may undergo distortions that affect the integrity of the embedded watermark information. We assume that there exists a method of attack characterization such that each extracted watermark repetition has a known associated reliability. In our analysis, we make use of the probability of bit error measure.

Many proposed watermarking algorithms [2,8-13] (this is by no means an exhaustive list) are encompassed by this class of techniques or can be easily modified to fit this category.

The specific details of the data embedding and extraction processes are not relevant to our work. Although we restrict the watermark to be a bit sequence and the reliability measure to be the bit error rate, we believe the spirit of the results discussed in the paper holds for non-binary watermarks with a different reliability measure such as the signal-to-noise ratio.

3 Modeling and Estimation

3.1 Parallel BSC Model

Given the characterization presented in the previous section, we can extract the individual watermark repetitions to produce M estimates of the watermark, $\hat{w}_1$, $\hat{w}_2$, ..., $\hat{w}_M$. As discussed in the previous section each estimate $\hat{w}_k$ has an associated probability of bit error p_{Ek} . We assume that some sort of characterization is performed so that a good estimate of p_{Ek} is available. Such a technique is presented, implemented and tested in [7].

Our framework is analogous to transmitting the watermark simultaneously along M independent binary symmetric channels (BSC) as shown in Figure 1. The error probabilities $0 \leq p_{Ek} \leq 0.5$ are assumed to be known and independent of one another. If $p_{Ek} > 0.5$, then the output is complemented and $1 - p_{Ek}$ is used as the probability of error parameter value.

This type of localized characterization of the distortion in the watermark domain allows better modeling of non-stationary fading-like distortions. The new perspective provides insights on effective strategies for watermarking. Most of the theoretical work in the area so far has considered the attacks on the watermark to be stationary [2]. This basic assumption precludes the benefits that diversity can provide, and limits understanding into the advantages of using one watermarking domain over another.

3.2 Linear Watermark Estimation

To estimate the embedded watermark *w*, we choose to linearly weight and add the extracted repetitions so that the overall estimate of the *i*th watermark bit is given by

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

for $i = 1, 2, ..., N_w$, where round[·] is the integer round operator. It is shown in [14] that

$$\alpha_k = \frac{\log\left(\dfrac{1-p_{Ek}}{p_{Ek}}\right)}{\sum_{j=1}^{M}\log\left(\dfrac{1-p_{Ej}}{p_{Ej}}\right)},$$

minimizes the bit error rate of the overall extracted watermark estimate $\hat{w}(i)$.

This linear estimation procedure is by no means the only alternative for combining the various extracted repetitions, but it is computationally simple, and it has been successfully implemented and tested in [7]. The following section summarizes theoretical observations concerning the analysis of such watermark recovery.

4 Error Statistic Bound

We provide a sketch of the analytic work initially presented in [7,14-16] and attempt to present a more intuitive perspective of the theory in the subsequent section.

Consider the bit $e_k(i)$ defined as

$$e_k(i) \triangleq w(i) \oplus \hat{w}_k(i) = \begin{cases} 1 & \text{if bit error in } \hat{w}_k(i) \\ 0 & \text{otherwise} \end{cases},$$

where ⊕ is the exclusive-OR operator. Similarly, we let

$$e(i) \triangleq w(i) \oplus \hat{w}(i) = \begin{cases} 1 & \text{if bit error in } \hat{w}(i) \\ 0 & \text{otherwise} \end{cases}.$$

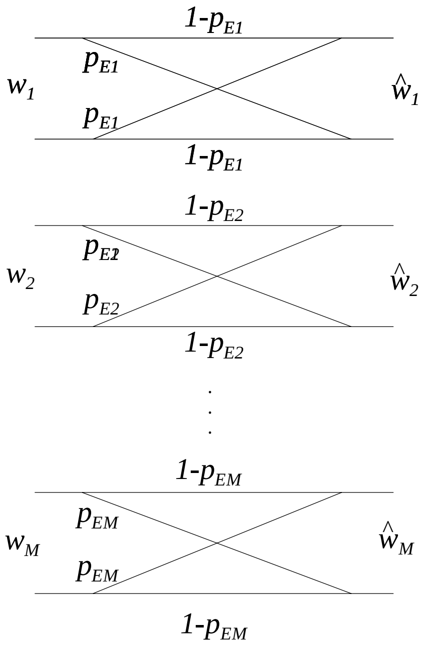

Figure 1: Parallel BSC model of the watermarking channel. Each repetition of the watermark is considered to undergo transmission through an independent BSC.

It follows that [14]

$$e(i) = \mathrm{round}\left[\sum_{k=1}^{M} \alpha_k e_k(i)\right],$$

which relates the bit errors of the individual extracted repetitions to the bit error of the overall watermark estimate. For perfect watermark extraction, we would like $e(i) = 0$ for $i = 1, 2, ..., N_w$. The bit error *e(i)* is a statistical quantity dependent on the probabilities p_{E1} , p_{E2} , ..., p_{EM} . This follows from the BSC model of the watermark channel in which $\hat{w}_k(i)$ (or equivalently $e_k(i) = w(i) \oplus \hat{w}_k(i) = 1$) occurs with probability p_{Ek} . The greater the degree of distortion on the region of the watermark domain in which the *k*th watermark repetition is embedded, the larger the value of p_{Ek} where $0 \le p_{Ek} \le 0.5$.

Analysis of the characteristics of $e(i)$ is not straightforward due to the presence of the integer round operator. Alternatively, we can consider the argument of this operator which is given by

$$E' \triangleq \sum_{k=1}^{M} \alpha_k e_k(i).$$

A bit error occurs in $\hat{w}(i)$ if $E' > 0.5$. We can analyze the mean value of E' , denoted $\mathscr{E}\{E'\}$, to assess the reliability of the watermark channel. Although this is not a precise measure of the error rate of the system since a smaller $\mathscr{E}\{E'\}$ does not necessarily guarantee a lower overall bit error rate, it does provide some useful insight into the watermarking problem.

The following *error statistic bound* is established in [16],

$$\mathscr{E}\left\{ E' \right\} \leq \frac{\bar{p}_E}{1-\bar{p}_E}\left[1 - \frac{D(q_a\|q_b)}{\log\left(\dfrac{1-\bar{p}_E}{\bar{p}_E}\right)}\right]$$

where

$$\bar{p}_E = \frac{1}{M}\sum_{k=1}^{M} p_{Ek} ,$$

is the average bit error rate, and the quantity $D(q_a\|q_b)$ is the relative entropy given by [17]

$$D(q_a\|q_b) = \sum_{k=1}^{M} q_a(k)\log\left(\frac{q_a(k)}{q_b(k)}\right) ,$$

where the arguments are $q_a(k) = p_{Ek}/(M\bar{p}_E)$, and $q_b(k) = (1-p_{Ek})/(M(1-\bar{p}_E))$.

We can see that q_a and q_b are probability-like distributions since their elements are nonnegative and sum to one. It is discussed in [14,16] the error bound is tight for small $\bar{p}_E$ and p_{Ek} close to a constant. The equality of the error bound holds if and only if p_{Ek}=0 for all k.

A smaller value for the bound on $\mathscr{E}\{E'\}$ implies that, for the most part, we can guarantee better accuracy of the extracted watermark, and, hence, greater robustness. In the next section, we intuitively discuss ways of diminishing the bound on the error statistic, which provides practical strategies for more effective watermarking.

6 Implications and Design Insights

From our analysis we find that the following possible tactics may be incorporated into a watermarking scheme to lower the value of the error statistic bound on $\mathscr{E}\{E'\}$ and, hence, improve the robustness of the watermarking system in some way:

1. Reduce the value of the average bit error rate. Reducing the value of $\bar{p}_E$ decreases the term $(\bar{p}_E/(1-\bar{p}_E))$ and increases the denominator term $\log((1-\bar{p}_E)/\bar{p}_E)$ 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, on average more robust watermark are commonly employed. The deficiency of most watermarking methods is that they solely rely on embedding a stronger watermark using sophisticated human perceptual mathematical models for improved performance. Our next two theoretical observations shed light on a different strategy to increase robustness.

2. Embed the watermark such that the distributions q_a and q_b are dissimilar for a large class of distortions. For a fixed value of $\bar{p}_E$, we may reduce the performance bound by increasing the value of $D(q_a\|q_b)$. The relative entropy is a measure of the distance between its two argument distributions [17]. Roughly, we can see that $D(q_a\|q_b)$ is relatively large when $q_a(k) = p_{Ek}/(M\bar{p}_E)$ and its corresponding $q_b(k) = (1-p_{Ek})/(M(1-\bar{p}_E))$ are *dissimilar*.

Assuming a fixed average probability of bit error, 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 region containing a repetition of the watermark. As a result, the amplitude of p_{Ek} will be different for distinct values of k. This can be achieved by inserting the watermark in a domain which distributes the distortion more to certain coefficients, leaving the others less affected.

3. Localize the distortions on the watermarked signal. It is shown in [15,16] that the existence of $p_{Ek} = 0$ for at least one $k \in \{1, 2, ..., M\}$ implies that $E' = 0$. Thus, if there exists a set of localized coefficients containing one complete repetition of the watermark which are unmodified by the distortion, then perfect watermark recovery is possible, as long as all the values of p_{Ek} are known. This translates to embedding the watermark in a domain which completely localizes the distortion to a finite and relatively small percent of the coefficients.

Both **2.** and **3.** relate the accuracy of the extracted watermark to the watermark domain in which the hidden data is embedded. By using diversity and attack characterization, it is possible to improve the effectiveness of the watermark to a specific class of distortions by inserting the mark in signal coefficients which localize these distortions. For example, to design a watermark robust against cropping, it would be wise to embed the mark in the spatial domain, which completely localizes the manipulation. Although a portion of the watermark is clipped out, the repetitions in the remaining signal are still accessible. Similarly, for robustness against filtering, the watermark should be embedded in the discrete Fourier domain which localizes the associated degradations. Mild linear filtering will affect some Fourier coefficients more than others. To make the watermark robust to both, a compromise would be to use the discrete wavelet domain for hiding the data.

More specific work on incorporation of particular wavelets to be robust against perceptual coding is presented in [16]. It is demonstrated that use of different domains for watermarking and perceptual coding improves the robustness of the embedded watermark. This work is in direct conflict with well-established principles which suggest the same domain is superior [18].

7 Final Remarks

In this paper, we take a different perspective on the problem of digital watermarking. We hypothesize that common watermark attacks are non-stationary and model the associated watermark channel as a set of parallel BSCs. We incorporate notions of diversity and channel estimation into our framework. Preliminary analysis provides new insights into the problem of robust data hiding.

Specifically we demonstrate how it is not only necessary to gain performance improvement by maximizing watermark signal energy, but it is better to embed the mark in a domain which localizes the distortions to a relatively small fraction of the coefficients. By assuming a localized non-uniform degradation model for the watermark we gain insight into appropriate domains in which to robustly hide data.

Future work involves using more sophisticated methods of estimating the overall embedded watermark from the various extracted repetitions by implying nonlinear order statistics [19]. We predict that analysis of such a system will lead to further performance improving approaches.

8 Acknowledgments

Part of this work was conducted by the author as part of her doctoral thesis under the supervision of Professor Dimitrios Hatzinakos.

9 References

1. R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, “A digital watermark,” *Proc. IEEE Int. Conference on Image Processing*, vol. 2, pp. 86-90, 1994.
2. I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for multimedia,” Tech. Rep. 95-10, NEC Research Institute, 1995.
3. R. B. Wolfgang and E. J. Delp, “A watermark for digital images,” *Proc. IEEE Int. Conference on Image Processing*, vol. 3, pp. 219-222, 1996.
4. J. R. Smith and B. O. Comiskey, “Modulation and information hiding in images,” *Proc. First Int. Workshop on Information Hiding* (R. Anderson, ed.), no. 1174 in Lecture Notes in Computer Science, pp. 207-226, May/June 1996.
5. X.-G. Xia, C. G. Boncelet, and G. R. Arce, “A multiresolution watermark for digital images,” *Proc. IEEE Int. Conference on Image Processing*, vol. 1, pp. 548-551, 1997.
6. P. G. Flikkema, “Spread-spectrum techniques for wireless communications,” *IEEE Signal Processing Magazine*, vol. 14, pp. 26-36, May 1997.
7. D. Kundur and D. Hatzinakos, “Improved robust watermarking through attack characterization,” *Optics Express focus issue on Digital Watermarking*, vol. 3, no. 12, pp. 485-490, Dec. 7, 1998.
8. E. Koch and J. Zhao, “Towards robust and hidden image copyright labeling,” *Proc. Workshop on Nonlinear Signal and Image Processing* (I. Pitas, ed.), pp. 452-455, June 1995.
9. J. Ohnishi and K. Matsui, “Embedding a seal into a picture under orthogonal wavelet transform,” *Proc. Int. Conference on Multimedia Computing and Systems*, pp. 512-521, June 1996.
10. C. I. Podilchuk and W. Zeng, “Image-adaptive watermarking using visual models,” *IEEE Journal in Selected Areas in Communications*, vol. 16, pp. 525-539, May 1998.

11. G. W. Braudaway, “Protecting publicly-available images with an invisible image watermark,” *Proc. IEEE Int. Conference on Image Processing*, vol. 1, pp. 524-527, 1997.
12. J. J. K. O'Ruanaidh and T. Pun, “Rotation, scale and translation invariant digital image watermarking,” *Proc. IEEE Int. Conference on Image Processing*, vol. 1, pp. 536-539, 1997.
13. S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, “Capacity issues in digital image watermarking,” *Proc. IEEE Int. Conference on Image Processing, 1998.*
14. D. Kundur and D. Hatzinakos, “Attack characterization for effective watermarking," to appear in *Proc. Int. Conf. on Image Processing*, Kobe, Japan, October 1999.
15. D. Kundur and D. Hatzinakos, “Mismatching perceptual models for effective watermarking in the presence of compression,” to appear in *Proc. SPIE -- Multimedia Systems and Applications II*, vol. 3845, Boston, Massachusetts, September 1999.
16. D. Kundur, “Mulitresolution digital watermarking: Algorithms and implications for multimedia signals,” Ph.D. Thesis, University of Toronto, 1999.
17. T. Cover and J. Thomas, *Elements of Information Theory*. Toronto: John Wiley & Sons, Inc., 1991.
18. R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, “The effect of matching watermark and compression transforms in compressed color images,” *Proc. IEEE Int. Conference on Image Processing*, vol. 1, 1998.
19. G. R. Arce, Personal Communication, August 9, 1999.