A STOCHASTIC QIM ALGORITHM FOR ROBUST, UNDETECTABLE IMAGE WATERMARKING

Pierre Moulin and Alexia Briassouli

U. of Illinois at Urbana-Champaign
Beckman Inst., Coord. Sci. Lab. & ECE Dept
405 N. Mathews Ave., Urbana, IL 61801, USA
Email: {moulin,briassou}@ifp.uiuc.edu

ABSTRACT

We propose a blind image watermarking algorithm that has three main features: (1) it satisfies an undetectability constraint with respect to a statistical image model, (2) it rejects host-signal interference, and (3) it is robust against additive and multiplicative noise, filtering, cropping, and random bending.

1. INTRODUCTION

In some applications, it is desired to secretly embed information into cover data, in such a way that the presence of embedded information is undetectable. Such is the case in steganography, where a sender wishes to secretly transmit data to a receiver [1]. It is often assumed that the stego data (output of the steganographic code) are received without transmission errors (the so-called passive warden scenario).

Another type of application, which is explored in this paper, is digital signature verification with the additional requirement that embedding of the digital signature into the host data be undetectable by the general public (by eavesdroppers). This property of the verification system could be useful in applications such as

- security applications. If signature embedding is statistically visible, then appropriate attacks can be often designed to remove the signature. [1]
- monitoring systems in which the document owner wishes to discretely verify the trustworthiness of users.
- copyright protection systems in which the document owner does not wish to publicly announce that he is protecting some documents, because of sensitive commercial or political issues.
- other forensic watermarking systems.

The difference between the above applications and passive-warden steganography is that (1) perhaps as little as one bit of information is to be hidden (because the receiver knows the digital signature to be verified), and (2) the system should be designed so that reliable verification is possible even when the marked data are degraded (attacked).

While there has been considerable literature on robust, blind image watermarking in the last ten years, the best methods (those that reject host-signal interference at the detector) do not satisfy an undetectability condition. This is true of most QIM (quantization index modulation) watermarking systems. Two recently developed stochastic QIM systems [4] satisfy an undetectability condition, and this paper explores the possibility of making such systems robust against signal-dependent noise and geometric attacks.

The method developed in this paper is in a sense also analogous to STDM (spread transform dither modulation [5]) or sparse QIM [6] methods: secret image components are coarsely quantized. The receiver knows the location of the secret components. Three differences with respect to STDM and sparse QIM methods are a) 1-bit quantization is used; b) the encoder is a stochastic encoder [4], introducing a source of randomness that is unknown to the detector; and c) the statistics of image components plays a central role in the system design and in the performance analysis.

2. HOST IMAGE MODEL

Our approach is based on a statistical model for the squared magnitude of the 2-D DFT of the image. Let $\Omega = \{0, 1, \cdots, N_1\} \times \{0, 1, \cdots, N_2\}$ be the domain over which the image $s(n), n \in \Omega$ is defined. The image is viewed as a realization from a

WORK SUPPORTED BY NSF GRANTS CCR 00-81268, CCR 02-08809, AND CDA 96-24396.

[1] Consider for example embedding of signatures at secret locations in the Fourier domain [2, 3]. Unless special precautions are taken, these locations can be guessed by an attacker following a simple spectral analysis.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

random process, to be defined. Its 2-D DFT is

$$S(k) = \sum_{n \in \Omega} s(n_1, n_2) e^{-j2\pi(k_1 n_1/N_1 + k_2 n_2/N_2)}, \quad k \in \Omega.$$

Its power spectrum is

$$\sigma_S^2(k) = \mathbb{E}|S(k)|^2, \quad k \in \Omega.$$

Let $\hat{\sigma}_S^2(k)$ be a *good* estimator of $\sigma_S^2(k)$. Define the normalized periodogram,

$$u(k) = \frac{|S(k)|^2}{\hat{\sigma}_S^2(k)}, \quad k \in \Omega. \tag{1}$$

Also define a DFT phase sequence $\phi_s(k)$ such that

$$S(k) = |S(k)| e^{\phi_s(k)}, \quad k \in \Omega.$$

Statistical Model: $u(k)$ are i.i.d. random variables with unit exponential distribution: $p_U(u) = e^{-u}$ for $u \geq 0$. The power spectrum $S(k)$ is assumed to be slowly varying.

Under the above assumption, the power spectrum can be reliably estimated from data by lowpass filtering the periodogram, so we have $\hat{\sigma}_S^2(k) \approx \sigma_S^2(k)$. The model above is motivated by spectral theory for wide-sense stationary random processes [7].

To illustrate the validity of the above model for $u(k)$, we computed the histogram of $u(k)$ for *Lena*, as shown in Fig. 1. The fit to the exponential pdf is nearly perfect. (Note that due to the 256×256 size of the image, probabilities of the order of 10^{-4} cannot be reliably estimated.)

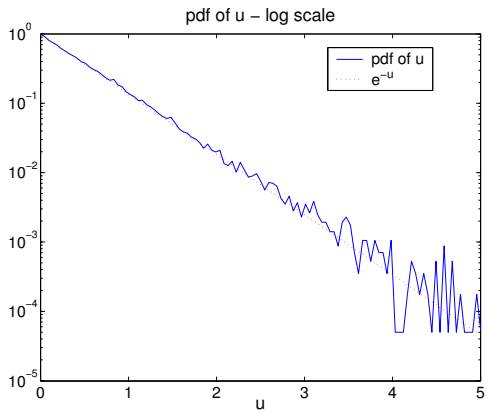

Fig. 1. Log histogram for normalized periodogram of *Lena*.

Attacks. Some simplified models for attacks on X are considered in this paper.

Linear Filtering: $Y(k) = X(k)H(k), \quad k \in \Omega$, where $H(k)$ is a slowly varying function.

Multiplicative Noise. Same as above, except that $H(k)$ is a noise sequence whose pdf is concentrated in the range $[0.9, 1.1]$.

Cropping. In the frequency domain, the effect of cropping is to convolve $X(k)$ with the 2-D DFT of the window function used for cropping.

Warping (or random bending). This causes a frequency-dependent spread of DFT components, which is benign at low frequencies but severe at high frequencies.

3. EMBEDDING MODEL #1

First consider a marking model that is resistant against noise and linear filtering attacks but not against strong cropping and warping.

Given the host signal s, compute the normalized periodogram as described in (1). Select a set $\mathcal{K}^*$ made of J secret frequencies $k_1, k_2, \cdots, k_J$ in the range $\mathcal{K} := [k_{\min}, k_{\max}]^2$. Let $\epsilon = \frac{|\mathcal{K}^*|}{|\mathcal{K}|}$ and $u^* = -\ln \epsilon$. Observe that $Pr[U \geq u^*] = \epsilon$. Define the three pdf's

$$p_0(u) = p_U(u) \tag{2}$$

$$p_{1a}(u) = \frac{1}{\epsilon} p_U(u) \, 1_{\{u \geq u^*\}} \tag{3}$$

$$p_{1b}(u) = \frac{1}{1-\epsilon} p_U(u) \, 1_{\{0 \leq u < u^*\}} \tag{4}$$

Therefore $p_U = \epsilon p_{1a} + (1-\epsilon) p_{1b}$. These three pdf's are shown in Fig. 2.

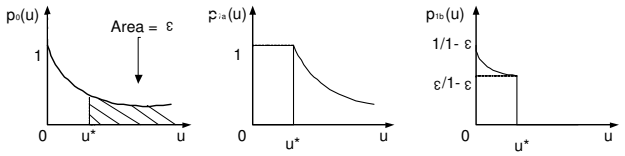

Fig. 2. Pdf's p_0, p_{1a} and p_{1b}.

The marking process is defined as follows. For $k \in \mathcal{K}^*$ let $\tilde{u}(k)$ be equal to $u(k)$ if $u(k) \geq u^*$; otherwise generate i.i.d $\tilde{u}(k)$ from the pdf p_{1a}. For $k \notin \mathcal{K}^*$ let $\tilde{u}(k)$ be equal to $u(k)$ if $0 \leq u(k) \leq u^*$; otherwise generate i.i.d. $\tilde{u}(k)$ from p_{1b}. Therefore $\tilde{u}(k)$ are independent, with pdf p_{1a} if $k \in \mathcal{K}^*$ and p_{1b} if $k \in \mathcal{K} \setminus \mathcal{K}^*$. Next let $|X(k)| = \sigma_S(k)\sqrt{\tilde{u}(k)}$. The marked signal is defined as

$$X(k) = |X(k)| e^{\phi_s(k)}, \quad k \in \Omega.$$

The set $\mathcal{K}^*$ is the secret shared between the embedder and detector. The embedding distortion is an increasing function of ϵ.

Observe that the marking scheme satisfies a steganographic constraint in the sense that according to our model [2], the pdfs of S and X (averaged with respect to $\mathcal{K}^*$) are

[2] More work is needed to see how well this model holds in practice.

identical. (Of course the pdf of X conditioned on $\mathcal{K}^*$ is markedly different from p_S.) From the attacker's perspective, reliably estimating the secret $\mathcal{K}^*$ is straightforward when $\epsilon = 1$ (the attacker would easily defeat our scheme in this case) but becomes impossible when $\epsilon \to 0$.

4. DETECTION

We have two hypotheses: H_0 (unmarked signal Y) and H_1 (marked signal Y). The detector shares the secret $\mathcal{K}^*$ with the embedder.

Given the data y, the detector first computes an estimate $\hat{\sigma}_Y^2(k)$ of the power spectrum of Y and next computes the normalized periodogram

$$u_Y(k) = \frac{|Y(k)|^2}{\hat{\sigma}_Y^2(k)}, \quad k \in \mathcal{K}. \tag{5}$$

Due to our smoothness assumptions and model for filtering attacks, $\hat{\sigma}_Y^2(k)$ is a reliable estimator of the power spectrum $\sigma_Y^2(k) = \sigma_S^2(k)|H(k)|^2$ of Y.

4.1. No Attacks

In the absence of attacks ($Y = X$), the rival pdfs of u_Y can be derived exactly:

$$\left\{ \begin{array}{lll} H_0 : & U_Y(k) \sim \text{i.i.d. } p_0 & , k \in \mathcal{K} \\ H_1 : & U_Y(k) \sim \text{i.i.d. } p_{1a} & , k \in \mathcal{K}^* \\ & U_Y(k) \sim \text{i.i.d. } p_{1b} & , k \in \mathcal{K} \setminus \mathcal{K}^* \end{array} \right. \tag{6}$$

A randomized likelihood ratio test (LRT) is optimal under a Neyman-Pearson setup for the hypothesis testing problem [8]. The probabilities of false alarm (false positives) and of correct detection are denoted by P_{FA} and P_D, respectively. The LRT is a comparison of the likelihood ratio $L(u_Y)$ with a threshold τ. We have

$$L(u_Y) = \frac{p_1(u_Y)}{p_0(u_Y)} = \left\{ \begin{array}{ll} e^{|\mathcal{K}|h(\epsilon)} & : \mathcal{K}_{\text{match}} = \mathcal{K} \\ 0 & : \text{else} \end{array} \right. \tag{7}$$

where $h(\epsilon) := -\epsilon \ln \epsilon - (1-\epsilon)\ln(1-\epsilon)$ is the binary entropy function (in nats), and $\mathcal{K}_{\text{match}} = \{k \in \mathcal{K}^* : u_Y(k) \geq u^*\} \cup \{k \in \mathcal{K} \setminus \mathcal{K}^* : u_Y(k) < u^*\} \subset \mathcal{K}$ is the set of "positive matches".

The Receiver Operating Characteristic (ROC) [8] for the detector is given by

$$P_D = \min(1, e^{|\mathcal{K}|h(\epsilon)} P_{FA}), \quad 0 \leq P_{FA} \leq 1. \tag{8}$$

The points in the range $0 < P_{FA} < e^{-|\mathcal{K}|h(\epsilon)}$ are achieved using a randomized LRT.

Remark. The ROC is independent of the energy of the host image. It depends only on the product of $|\mathcal{K}|$ and $h(\epsilon)$.

4.2. Attacks

Motivated by the analysis in the previous subsection, we use $N_{\text{match}} = |\mathcal{K}_{\text{match}}|$ as a test statistic. Our test is of the form

$$N_{\text{match}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

Observe that N_{match} is a sufficient statistic [8] for detection in the absence of attacks. Moreover, N_{match} is almost invariant to filtering attacks.

4.3. Numerical Results

We applied the detector (9) to the 256×256 *Lena* image. The region $\mathcal{K} := [k_{\min}, k_{\max}]^2$ was made of middle- and high-frequency DFT coefficients: $k_{\min} = 40$ and $k_{\max} = 220$. We chose $\epsilon = \frac{1}{25}$. For each experiment we performed 1000 Monte-Carlo simulations in which the noise realization (if applicable) and the set $\mathcal{K}^*$ were randomly varied. We tested our detector using the following degraded images:

- a heavily blurred version of *Lena*. The blurring filter was a 21×21 box filter.
- a noisy version of *Lena*. The noise was multiplicative and Gaussian with a mean of 1 and variance 0.02.

Not a single false positive or false negative was obtained.

5. EMBEDDING MODEL #2

The marking model needs to be modified in order to cope with strong cropping or warping attacks. Cropping results in blurring in the frequency domain. Warping causes frequency-dependent spread in the frequency domain. Let $\mathcal{K}_j$ denote a neighborhood of L frequencies around frequency k_j. Define

$$v(j) = \sum_{i \in \mathcal{K}_j} u(i), \tag{10}$$

which is half a χ^2 random variable with $2L$ degrees of freedom.

If L is large enough that most of the frequency spread is contained within $\mathcal{K}_j$, then the features $v(j)$ are relatively insensitive to frequency spreading. (In addition, they inherit the relative insensitivity of the original features $u(i)$ to linear filtering).

The embedding algorithm of Sec. 3 is modified as follows. Choose two "sufficiently large" integers J_1, J_2, and let $\mathcal{J} = \{1, 2, \cdots, J_1 J_2\}$ and $\epsilon = \frac{1}{J_2}$. Partition the frequency range $\mathcal{K} = [k_{\min}, k_{\max}]^2$ into J_1 regions $\mathcal{R}_1, \cdots, \mathcal{R}_{J_1}$. Each region $\mathcal{R}_{j_1}$ is further subdivided into J_2 size-L_{j_1} neighborhoods $\mathcal{K}_j$, where $(j_1 - 1)J_2 < j \leq j_1 J_2$. In our design,

the regions $\{\mathcal{R}_{j_1}\}$ and neighborhoods $\{\mathcal{K}_j\}$ have square shapes. The neighborhood sizes $L_1, \cdots, L_{J_1}$ are odd integers, with mean $\overline{L} := \frac{1}{J_1}\sum_{j_1=1}^{J_1} L_{j_1}$, and so $|\mathcal{J}|\,\overline{L} = |\mathcal{K}|$. Next select randomly a neighborhood $\mathcal{K}_j$ from each region $\mathcal{R}_{j_1}$, The set $\mathcal{J}^* \subset \mathcal{J}$ indexing the selected neighborhoods has size J_1 and is a secret shared with the detector. A 1-D version of the selection procedure is illustrated in Fig. 3.

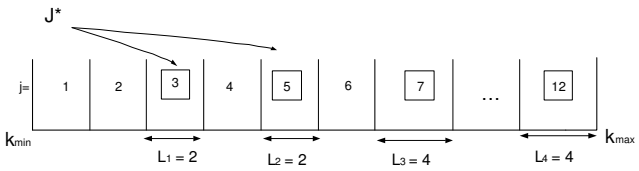

Fig. 3. (1-D) Selection of intervals $\mathcal{K}_j$, $j \in \mathcal{J} = \{1, 2, \cdots, 12\}$. Here $J_1 = 4$, $J_2 = 3$, $\{L_{j_1}\} = \{2, 2, 4, 4\}$, and $\mathcal{J}^* = \{3, 5, 7, 12\}$.

For any L, define v_L^* as half the ϵ-quantile of the χ^2_{2L} distribution, i.e., $Pr[\frac{1}{2}\chi^2_{2L} > v_L^*] = \epsilon$. Define the pdf's

$$p_{0,L} = \frac{1}{2}\chi^2_{2L} \tag{11}$$

$$p_{1a,L}(v) = \frac{1}{\epsilon} p_{0,L}(v)\, 1_{\{v \geq v_L^*\}} \tag{12}$$

$$p_{1b,L}(v) = \frac{1}{1-\epsilon} p_{0,L}(v)\, 1_{\{0 \leq v < v_L^*\}} \tag{13}$$

Therefore $p_{0,L} = \epsilon p_{1a,L} + (1-\epsilon) p_{1b,L}$ for all L.

A marking process similar to that described in Sec. 3 is applied to $\{v(j)\}$, resulting in marked coefficients $\{\tilde{v}(j)\}$ that have the same pdf (averaged over $\mathcal{J}^*$) as the original pdf. Next let $\tilde{u}(k) = \frac{\tilde{v}(j)}{v(j)} u(k)$ for all $k \in \mathcal{K}_j$, $j \in \mathcal{J}$, and $\tilde{u}(k) = u(k)$ for all other values of k. Obtain the marked signal X from $\{\tilde{u}(k)\}$ as described in Sec. 3.

The set $\mathcal{J}^*$ is the secret shared between the embedder and detector. The marking scheme satisfies a steganographic constraint in the sense that the pdfs of S and X (averaged with respect to $\mathcal{J}^*$) are identical.

For detection, we use the same assumptions about the detector's knowledge as before. The detector first computes the normalized periodogram (5) and then the statistics

$$v_Y(j) = \sum_{i \in \mathcal{K}_j} u_Y(i), \quad j \in \mathcal{J}. \tag{14}$$

The test is (9), with test statistic now given by $N_{\text{match}} = |\mathcal{J}_{\text{match}}|$, where $\mathcal{J}_{\text{match}} = \{j \in \mathcal{J}^* \,:\, v_Y(j) \geq v_L^*\} \cup \{j \in \mathcal{J} \setminus \mathcal{J}^* \,:\, v_Y(j) < v_L^*\}$.

We tested this detector using the following degraded images:

- a cropped version of *Lena*, in which only the 156×156 center part of her face is retained.
- a heavily warped, difformed version of *Lena* (see Fig. 4). The warping field was a 2-D AR(1) Gaussian process. The maximal pixel deviation for typical realizations of this process was 15.

Fig. 4. Warped *Lena*.

Applying the test (9) to cropped and warped versions of *Lena* and using $k_{\min} = 40$, $k_{\max} = 220$, $\epsilon = \frac{1}{25}$, $\{k_j, j \in \mathcal{J}\} = \{40, 60, \cdots, 220\}^2$, $\{k_j, j \in \mathcal{J}^*\} = \{80, 180\}^2$ and $\overline{L} = 5$, we have obtained no errors in the cropping case and error probabilities lower than 10% in the case of severe warping.

6. REFERENCES

[1] N. F. Johnson, Z. Duric and S. Jajodia, *Information Hiding. Steganography and Watermarking – Attacks and Countermeasures*, Kluwer, Boston, 2001.

[2] M. Kutter, "Watermarking Resisting to Translation, Rotation and Scaling," *Proc. SPIE*, Boston, Vol. 3528, pp. 423—431, 1998.

[3] S. Pereira and T. Pun, "Robust Template Matching for Affine Resistant Image Watermarks," *IEEE Trans. on Image Processing*, Vol. 9, No. 6, pp. 1123–1129, June 2000.

[4] Y. Wang and P. Moulin, "Steganalysis of Block-Structured Stegotext," *Proc. SPIE* Vol. 5306, San Jose, CA, Jan. 2004.

[5] B. Chen and G. W. Wornell, "Quantization Index Modulation Methods: A Class of Provably Good Methods for Digital Watermarking and Information Embedding," *IEEE Trans. Info. Thy*, Vol. 47, No. 4, pp. 1423—1443, May 2001.

[6] A. K. Goteti, P. Moulin and R. Koetter, "Optimal Sparse QIM Codes," *Proc. ICASSP*, Montreal, Canada, May 2004.

[7] M. B. Priestley, *Spectral Analysis of Time Series*, Vol. I, Academic Press, London, 1981.

[8] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd Ed., Springer-Verlag, 1994.