DETECTING SPREAD SPECTRUM WATERMARKS USING NATURAL SCENE STATISTICS

Kalpana Seshadrinathan, Hamid Rahim Sheikh, Alan C. Bovik

Laboratory for Image and Video Engineering; Dept. of Elec. and Comp. Engg.
The University of Texas at Austin, Austin, TX 78712-1084, USA.

ABSTRACT

This paper presents novel techniques for detecting watermarks in images in a known-cover attack framework using natural scene models. Specifically, we consider a class of watermarking algorithms, popularly known as spread spectrum-based techniques. We attempt to classify images as either watermarked or distorted by common signal processing operations like compression, additive noise etc. The basic idea is that the statistical distortion introduced by spread spectrum watermarking is very different from that introduced by other common distortions. Our results are very promising and indicate that this statistical framework is effective in the steganalysis of spread spectrum watermarks.

1. INTRODUCTION

The science of information hiding has received a lot of attention from both industry and academia over the past decade. The primary motivation behind information hiding has been copyright protection. Steganography hides secret messages in media in such a way that the very presence of the message cannot be detected. Digital watermarking systems have the additional requirement of being robust to common signal processing operations, although the presence of the watermark need not be hidden.

This problem of communicating secretly without anybody detecting the presence of a message was formalized as the *Prisoners' Problem* [1]. Two prisoners, Alice and Bob, are locked up in jail and wish to communicate with each other to hatch a plan to escape. All communication between them however passes through a warden, Willie. Willie will inspect all the messages that are exchanged and passes on a message only if he is sure it does not contain a secret message. The Prisoners' problem is hence for Alice and Bob to find some means of communicating secretly so they can coordinate their escape plans. Steganalysis assists Willie in detecting any secret messages automatically.

This paper presents statistical methods to detect the presence of spread spectrum watermarks in images, in a known cover attack framework, i.e., we assume that the reference image is available. Common distortions like JPEG compression, Gaussian blurring or additive white noise produce certain statistical distortions in an image that are distinctly different from those produced by spread spectrum watermarking. We propose a statistical framework using natural scene models to distinguish between images distorted by common operations like these and watermarked images, when the reference image is known.

A general purpose tool for steganalysis using image quality metrics was proposed in [2]. However, this general tool was designed to work across a range of algorithms and the accuracy of watermark detection is not very high. A technique to detect and remove only binary spread spectrum watermarks was proposed in [3]. Most steganalytic algorithms are tested using only cover and stego-images. Since images distorted by operations like JPEG compression are very commonly encountered in practice, we believe that using such images in testing is essential in making any claims about the performance of an algorithm. The algorithm presented in this paper attempts to distinguish between spread-spectrum based watermarks and other distortions like compression and blurring.

The noise that is added in watermarks based on spread spectrum ideas is additive and does not have a blur component. Also, this noise scales linearly with the coefficients of the image in the DCT or wavelet domain. This is in stark contrast to distortions like compression that have a significant blur component. Natural scene statistical models can be used effectively to model the image-dependent noise in watermarks. These statistical features are used in a hypothesis testing framework in the development of our classification algorithm.

2. NATURAL SCENE MODEL

In this section, we outline the statistical model for the wavelet coefficients of natural images that we use. These models are shown to be effective in modeling the image dependent noise in spread spectrum systems in this paper. The wavelet coefficients of an image can be modeled as a realization of a doubly stochastic process [4]. A related model which accounts for local dependencies was proposed in [5]. The wavelet coefficients are assumed to be conditionally independent zero-mean Gaussian random variables, given their variances. Let $\vec{X} = \{X_i, i \in \mathscr{I}\}$ denote the random field representing the wavelet coefficients of a particular sub-band of natural images. Here, $\mathscr{I}$ denotes a set of spatial indices for the random field. Then, $\vec{X}$ is modeled using $\vec{X} = \vec{Z} * \vec{U}$ where $*$ denotes element-wise multiplication of the vectors. $\vec{Z} = \{Z_i, i \in \mathscr{I}\}$ is the spatially varying, highly correlated random field, representing the local standard deviations and $\vec{U} = \{U_i, i \in \mathscr{I}\}$ is a zero-mean, white Gaussian random field. Each U_i can be modeled as a normal random variable of unit variance without any loss of generality as the variance of U_i can be absorbed into Z_i. It is assumed that the random field $\vec{X}$ is ergodic to make the coefficients obtained from the reference image representative of the underlying statistics. The Maximum Likelihood (ML) estimate of Z_i, denoted by $\hat{Z}_i$, is simply the standard deviation in a local neighborhood, under the assumption that the correlation between the standard deviations of neighboring coefficients is very high [4]. In our experiments, we use a square 5×5 window to compute $\hat{Z}_i$.

3. SPREAD SPECTRUM MODEL

We use the wavelet domain watermarking algorithm proposed in [6] in our experiments. The watermark $\vec{w} = \{w_1, w_2, \ldots, w_n\}$ is a realization of an iid Gaussian random vector $\vec{W}$. The image is decomposed using an R-level wavelet transform to generate $R+1$ resolutions. The watermark is added to all the coefficients at all resolutions except the DC coefficients. The coefficients in all these sub-bands are collected into a vector $\vec{x} = \{x_1, x_2, \ldots, x_n\}$. The watermark $\vec{w}$ is inserted into the coefficients $\vec{x}$ to obtain the watermarked coefficients $\vec{x}'$ using

$$x'_i = x_i \left(1 + \sigma_w w_i\right) \tag{1}$$

where σ_w is the strength of the watermark. Although we choose to analyze the algorithm presented in [6] due to its simplicity, the same principles can be extended to other spread spectrum based systems with minor modifications.

4. PROPOSED ALGORITHM

The key feature that we use is the fact that the noise that is added in spread spectrum algorithms scales with the coefficients of the image itself, which is evident from (1). The kind of noise that results due to common distortions like compression is, however, statistically different.

Let $\vec{Y_d} = \{Y_{d_i}, i \in \mathscr{I}\}$ denote a random field representing the coefficients of distorted images in one sub-band. Let $\vec{X} = \{X_i, i \in \mathscr{I}\}$ denote the random field representing the corresponding coefficients of the reference natural images. The model that we use for common distortions can be expressed as

$$Y_{d_i} = G_i X_i + N_i, \quad i \in \mathscr{I} \tag{2}$$

where $\vec{G} = \{G_i, i \in \mathscr{I}\}$ represents the signal attenuation and is a random gain field. $\vec{N} = \{N_i, i \in \mathscr{I}\}$ represents a stationary, white additive Gaussian noise field. Many distortion types that are present in the real world can be described locally by the combination of these two factors in the wavelet domain [7]. $\vec{G}$ is the blur factor that accounts for the loss of signal energy in sub-bands, which is common in distortions like compression and Gaussian blurring.

Denoting the random field representing the coefficients of the watermarked image by $\vec{Y_w}$, the watermarked image can be expressed as

$$Y_{w_i} = X_i + X_i W_i \tag{3}$$

where $\vec{W} = \{W_i, i \in \mathscr{I}\}$ is the random field representing the watermark and the variance of each W_i is σ_w^2. No assumptions are made about the distribution of W_i. We however assume that $\vec{W}$ is zero-mean and white. Also, $\vec{W}$ is independent of the random field $\vec{X}$. The proposed algorithm operates in three stages and we outline these in the following sections.

4.1. Regression Analysis of Blurred Images

The key feature that is used in this stage of the algorithm is the fact that the watermarking distortion is additive and does not have a blur component. We use this to first distinguish between images that are blurred (like compressed images) and ones that are not.

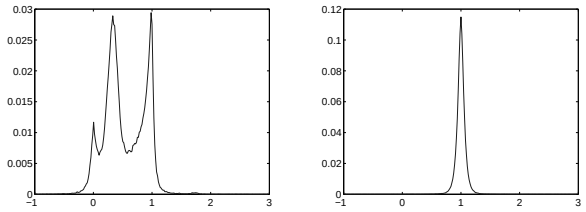


Fig. 1. Prior distribution of $\hat{G}$ for blurring(left) and watermarking(right) distortions

Linear regression is used to estimate G_i in (2) and the watermarked image is treated as a special case of (2) in this stage [8].

Let $\vec{Y} = \{Y_i, i \in \mathscr{I}\}$ denote the random field representing the coefficients of the test image that is suspected to be watermarked. We denote the least squares estimate of G_i by $\hat{G}_i$ and in the case of the watermarked signal given by (3), we will have

$$\hat{G}_i \simeq \frac{E(X_i Y_i)}{E(X_i^2)} = 1 \tag{4}$$

where (4) follows since X and W are assumed to be zero-mean and independent and from the linearity of expectation. We hence expect $\hat{G}_i$ to be 1 when the image is watermarked or if the distortion is purely additive. However, in the case of a compressed image, for example, the high frequencies in the image that the human visual system is less sensitive to are attenuated and we have $\hat{G}_i < 1$. This is also true of several other distortion types like blurring etc.

We use this difference in the distribution of G_i to eliminate "blurred" images in the first stage of our algorithm using a Bayes' classifier. Prior models for the distribution of G_i were derived using training data, for the cases when the distortion had a blur component and when it was purely additive. The histogram of $\hat{G}_i$ in one sub-band, obtained from training data for the two cases is shown in Fig. 1. The distribution clearly shows the properties that we expect.

Since the values of $\hat{G}_i$ are computed in overlapping blocks, we sub-sample the resulting values by a factor of 5. These can then be assumed to be independent samples from either one of the prior distributions. The blurring is more pronounced in the finer scale sub-bands as these contain the high frequencies that are attenuated in any low-pass filtering or compression operations. Independent decisions as to which class the image belongs to was made in six sub-bands at the finest scale and the majority value was chosen.

4.2. Hypothesis Testing using Natural Scene models

Once the blurred images have been identified, we are left with images contaminated with additive distortions, that are to be classified as watermarked or not. The key feature that we use here is the fact that the noise that is added in the watermarked images scales with the image coefficients.

The random field representing the reference image coefficients, $\vec{X}$, can be modeled using

$$\vec{X} = \vec{Z_x} \vec{U_x} \tag{5}$$

where $\vec{Z_x}$ and $\vec{U_x}$ are the random fields described in Section 2. $\hat{Z}_{x_i}$ is used to represent the estimated value of Z_{x_i}.

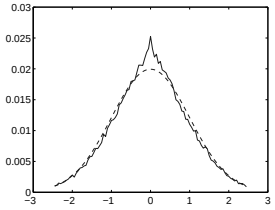


Fig. 2. Histogram of the normalized difference coefficients (solid line) and a discretized standard normal density (dashed line)

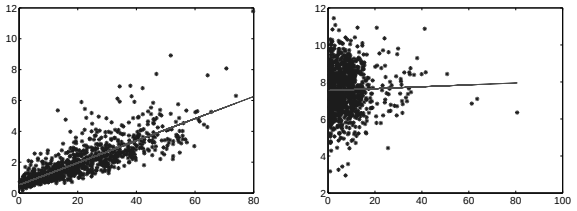


Fig. 3. Plot of $\hat{Z}_x$ vs. $\hat{Z}_d$ and the least-squares optimal linear fit for watermarked(left) and AWGN corrupted(right) images

Let $\vec{D} = \{D_i, i \in \mathscr{I}\}$ denote a random field representing the difference between reference and test image coefficients. We have

$$D_i = Y_i - X_i, \quad i \in \mathscr{I}$$

When the test image is watermarked, D_i is simply the watermark added to the image and we have $D_i = X_i W_i$. Since D_i is derived from the the reference image itself, the statistics of D_i exhibit similar properties that are characteristic of natural images. We discovered that $\vec{D}$ can be modeled reasonably well by the same model used for reference images and we have

$$\vec{D} = \vec{Z}_d \vec{U}_d \tag{6}$$

$\hat{Z}_{d_i}$ is used to denote the estimated value of Z_{d_i}. Fig. 2 shows the histogram of the difference coefficients normalized by the corresponding $\hat{Z}_{d_i}$ in one sub-band of a watermarked image. Also, shown is the discretized standard normal density. The fit is reasonably good and shows the effectiveness of (6) in modeling the noise that is added in spread spectrum watermarks. Notice that this also models the statistics of noisy images modeled using (2) that are for example, corrupted by gaussian noise. $\vec{Z}_d$ in this case will not be spatially varying, but almost a constant everywhere. $\vec{D}$ can be modeled using (6) for most commonly occurring additive distortions.

When the image is watermarked and $D_i = X_i W_i$, we expect

$$\hat{Z}_{d_i} \simeq \sqrt{E(X_i^2)E(W^2)} = \hat{Z}_{x_i}\sigma_w \tag{7}$$

at each coefficient i. Hence, a linear relationship exists between the local standard deviations of the reference image coefficients and the difference coefficients when the image is watermarked. A plot of $\hat{Z}_x$ versus $\hat{Z}_d$ is shown in Fig. 3, which clearly illustrates this linear behavior. Also shown is the same plot for an image contaminated by AWGN. $\hat{Z}_x$ and $\hat{Z}_d$ in this case are seen to be uncorrelated.

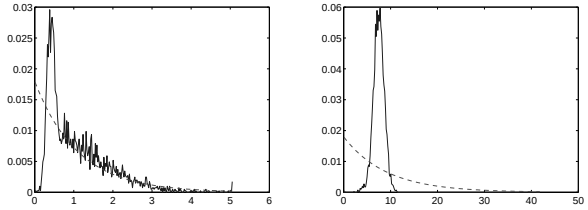


Fig. 4. Histogram of $\hat{Z}_d$ and the best fitting discretized exponential density for watermarked(left) and AWGN(right) images

We fit a simple linear regression model between $\hat{Z}_d$ and $\hat{Z}_x$ given by:

$$\tilde{Z}_{d_i} = \alpha \hat{Z}_{x_i} + \beta, \quad i \in \mathscr{I} \tag{8}$$

We then use an analysis of variance approach to test the significance of regression. Specifically, we test the hypothesis that $\alpha = 0$. If the hypothesis is true, the two quantities are uncorrelated and we can conclude that no watermark is present. However, if the hypothesis is rejected, a linear relationship does exist between $\hat{Z}_d$ and $\hat{Z}_x$ and we can conclude that the noise added does indeed scale with the image coefficients and the image is watermarked. Note from (7) that the least squares estimate of α that we obtain by solving (8) is approximately equal to the strength of the watermark σ_w, when the image is indeed watermarked.

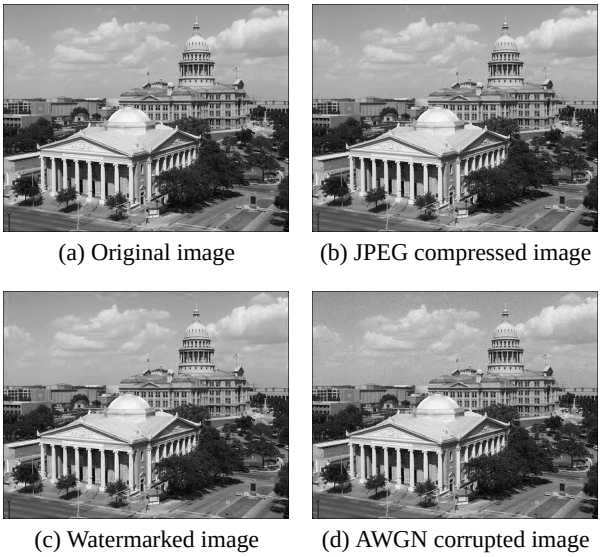
The test statistic is the ratio of the regression sum of squares and the error sum of squares and follows the F-distribution with 1 degree of freedom in the numerator and $n-1$ degrees of freedom in the denominator [8]. The test was carried out at a significance level of 0.01. In our experiments, we sub-sample the values of $\hat{Z}_d$ and $\hat{Z}_x$ that we obtain by a factor of 5, since these values are redundant as they are computed in overlapping blocks. Again, the F-test was carried out in each sub-band independently and the majority value was chosen.

4.3. Reducing False Positives

In our experiments, a good number of false positives were observed at this stage, i.e., although watermarked images were identified correctly, the null hypothesis was rejected for noisy images sometimes. This is because the number of samples that we use in the hypothesis testing is quite large and even a small difference between the empirical and hypothesized values leads to rejection of the null hypothesis.

To reduce the number of false positives, in the third stage, we use prior models to describe the distribution of the standard deviation field $\vec{Z}_d$. It has been shown that the exponential density is a reasonably good fit for the standard deviation field $\vec{Z}_d$ in the case of images [4]. Again, since the difference coefficients are derived from the image coefficients when the test image is watermarked, we observed that the exponential prior is a reasonably good fit in this case too. This however would not be true when the distortion in the image is not a watermark.

Fitting was done by minimizing the Kullback-Leibler (KL) divergence between the empirical histogram and a discretized version of the exponential density. Examples of the best fitting discretized exponential for the samples of $\vec{Z}_d$ when the image is watermarked and corrupted by AWGN is shown in Fig. 4.



(a) Original image (b) JPEG compressed image

(c) Watermarked image (d) AWGN corrupted image

Fig. 5. Original and distorted "Church and Capitol" images

A threshold was empirically obtained for the KL divergence between the best-fitting exponential density and the empirical histogram in the case of watermarked and distorted images using training data. Any image that rejected the null hypothesis in the previous stage of the algorithm is tested again in this stage. If the KL divergence of the fit is less than the obtained threshold, we declare the image to be watermarked.

5. RESULTS AND CONCLUSION

We tested the proposed algorithm using the watermarking algorithm proposed in [6]. The watermarked images in the test set were generated using different watermarking strengths and different wavelet bases. The distorted images included images that were JPEG compressed, JPEG 2000 compressed, printed and scanned, Gaussian blurred and AWGN and salt and pepper noise corrupted. All distortions were adjusted in strength such that the images look perceptually similar in quality. All reference images used in the simulations are available from the database in [9]. Approximately 50% of the images were used in the training phase to obtain prior distributions for $\hat{G}_i$ and to derive the threshold for the KL divergence. The results of running our algorithm on the remaining 50% of the images, comprising the test set, are described in Table 1.

In the implementation of our algorithm, we used a 3-level orthonormal wavelet decomposition using the Daubechies length-8 filters. The orthonormal representation was chosen because white noise processes in the space domain remain white in the wavelet coefficients, which is required in the statistical framework we have described.

It is seen that the algorithm performs well across a range of distortion types, although no assumptions are made about the nature of distortion. It should be noted that malicious attacks are not an issue in steganalytic techniques. Also, the proposed framework is independent of the distribution of the watermark and is applicable to other algorithms based on spread spectrum ideas with suitable modifications. No watermarked images are missed, although there are some false alarms which is desirable as the cost of a miss is usually much higher than that of a false alarm. The images used in the experiment were of very high quality perceptually as is seen from example images from the test set shown in Fig. 5. The classifier performed well despite the fact that the strength of the distortion was very low.

Watermark Type	No. images	Hits	%
D-4, 4 levels, σ_w = 0.1	14	14	100
D-8, 4 levels, σ_w = 0.2	14	14	100
D-16, 5 levels, σ_w = 0.1	14	14	100
Distortion Type	**No. images**	**True Neg.**	**%**
JPEG Compressed	14	14	100
JPEG 2000 compressed	14	14	100
Gaussian Blurred	14	14	100
AWGN corrupted	28	22	78.6
Salt and Pepper noise	14	11	78.6
Printed and Scanned	15	15	100

Table 1. Results: D-n refers to the Daubechies filter of length n

In conclusion, a new framework for the steganalysis of spread spectrum watermarks using natural scene statistics is proposed. The results are promising and efforts to extend these concepts to a stego-only framework are under-way. We also hope to analyze images that have been watermarked and distorted in the future.

6. REFERENCES

[1] G. J. Simmons, "The prisoners' problem and the subliminal channel," in *Proc. IEEE Workshop Communications Security CRYPTO'83*, Santa Barbara, CA, 1983, pp. 51–67.

[2] I. Avcibas, N. Memon, and B. Sankur, "Steganalysis using image quality metrics," *IEEE Trans. Image Processing*, vol. 12, no. 2, pp. 221–229, Feb. 2003.

[3] G. C. Langelaar, R. L. Lagendijk, and J. Biemond, "Removing spatial spread spectrum watermarks by nonlinear filtering," in *Proc. EUSIPCO*, 1998, pp. 2281–2284.

[4] M. K. Mihcak, I. Kozintsev, K. Ramachandran, and P. Moulin, "Low-complexity image denoising based on statistical modelling of wavelet coefficients," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 300–303, December 1999.

[5] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE Conf. 3813 on Wavelet Applications in Signal and Image Processing*, Denver, CO, July 1999.

[6] W. Zhu, Z. Xiong, and Y.-Q. Zhang, "Multiresolution watermarking for images and video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 545–550, June 1999.

[7] H. R. Sheikh, "Image quality assessment using natural scene statistics," Ph.D. dissertation, Univ. of Texas at Austin, May 2004.

[8] D. C. Montgomery and G. C. Runger, *Applied Statistics and probability for engineers*, 3rd ed. New York: John Wiley and Sons, Inc., 2003.

[9] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. (2003) Live image quality assessment database. [Online]. Available: http://live.ece.utexas.edu/research/quality