EFFECT OF JPEG COMPRESSION ON IMAGE WATERMARK DETECTION

Minghui Xia, Bede Liu

Information Sciences and Systems, Department of Electrical Engineering
Princeton University, Princeton, NJ 08544
Contact: {mxia,liu}@ee.princeton.edu

ABSTRACT

Compression on a watermarked image can significantly affect the detection of the embedded watermark. To show that a particular watermarking scheme is robust against compression, simulation is often relied upon. In this paper, we investigate this problem analytically. We first characterize the noise introduced by JPEG compression. We then propose a maximum a posteriori (MAP) detector, analyze its performance, and demonstrate that it has superior performance over the correlation detector.

1. INTRODUCTION

The detection of the presence or absence of a watermark embedded in an image is often affected if the watermarked image has undergone compression. Compression can also be considered as an attack on watermarked images [1]. Partly because it is difficult to analyze the effect of compression on the detection of watermarked image, past investigations of this problem have relied heavily on simulation. In this paper, we analyze the effect of JPEG compression on a well known watermark scheme. Noise is introduced by JPEG compression when the block-wise DCT coefficients are quantized. The statistics of this quantization noise generally depend on the statistics of the input source, and cannot be modeled merely as additive white Gaussian noise. Modeling of the quantization noise is presented in Section 3 and used in a maximum a posteriori (MAP) detector proposed in Section 4. The performance of the MAP detector is analyzed and, as expected, is shown to have better performance than the commonly used correlation detector. Experimental results are presented in Section 6.

We assume the image is of size NxN pixels. (The case of non-square arrays are handled similarly.) We need to deal with two kinds of DCT's. The first is that of the entire image; it is also a 2D array of size NxN. The second is a block-wise DCT or BDCT, which is obtained by first dividing the image into blocks of MxM pixels (M=8 in JPEG), and then taking the DCT of each block. The result is also a 2D array of size NxN. To study the statistical properties of images and its transforms, it is convenient sometimes to convert the 2D array into a 1D vector. This can be done by taking the rows in successive order. Another way is to divide the 2D array into blocks, arrange all blocks in a 1D array, and convert each block to a vector. When the vector $\boldsymbol{\xi}$ is the realization of a random variable ξ, we sometimes use $\boldsymbol{\xi}$ to represent either case for simplicity. When not all entries in a vector $\boldsymbol{\xi}$ are involved in a computation, it is convenient to make use of a select operation $\mathbf{S}$ on the vector to pick out only those element of interest: $\mathbf{S}\,\boldsymbol{\xi}$. For convenience, we may at times use the same symbol to denote the different representations of the same object, if the meaning is clear from the context. Thus, $\mathbf{y}$ may denote the NxN array of pixels, or a 1D vector of all the pixels. Similarly, $\boldsymbol{\xi}$ may denote a vector or a smaller vector $\mathbf{S}\,\boldsymbol{\xi}$.

This research is supported by an NJ State R&D excellence grant.

2. SPREAD SPECTRUM EMBEDDING IN FREQUENCY DOMAIN

In this section, we investigate a specific watermark scheme proposed in [2]. Similar analysis can be carried out for other schemes. In this scheme, the DCT of the entire host image is taken and the largest K coefficients, denoted by $\mathbf{x}$, are modified to embed a watermark $\mathbf{w}$. An often used watermark is a spreading sequence of normalized Gaussian distributed random numbers, generated from a key. It is embedded according to:

$$\mathbf{y} = (\mathbf{1} + \boldsymbol{\alpha} \cdot \mathbf{w})\mathbf{x}\,. \tag{1}$$

That is, the original K largest coefficients $\mathbf{x}$ is replaced by the K modified coefficients $\mathbf{y}$. The inverse DCT of the entire NxN array is the marked image. The masking factor $\boldsymbol{\alpha}$ is adjustable, to make the watermark invisible. This method inserts a single watermark in the entire image, similar to using spread spectrum for transmitting one bit information. The detector decides whether or not a test image contains the specific watermark. This is done by subtracting the original image from the test image, taking the DCT of the difference, and selecting the prescribed set of K coefficients to form the test watermark, denoted by $\hat{\mathbf{w}}$. It is

compared with the original watermark w using correlation:

$$sim(\mathbf{w}\,,\hat{\mathbf{w}}\,) = \frac{\langle \hat{\mathbf{w}}\ \cdot \mathbf{w}\,\rangle}{\sqrt{\langle \hat{\mathbf{w}}\ \cdot \mathbf{w}\,\rangle}}, \tag{2}$$

where $\langle\cdot\rangle$ denotes inner product.

3. NOISE DUE TO JPEG COMPRESSION

The block diagram of Figure 1 depicts watermark embedding, followed by JPEG compression and watermark detection. A watermark $\mathbf{w}$ is embedded in the DCT of the image to obtain $\mathbf{y}$, which is then transformed into image domain by inverse DCT. JPEG compression takes the 8x8 block DCT of the marked image $\mathbf{z}$, quantize the coefficients $\mathbf{f}$, and produce the distorted coefficients $\tilde{\mathbf{f}}$. An 8x8 block inverse DCT then gives the compressed and watermarked image, $\tilde{\mathbf{z}}$. To detect the watermark, the DCT of $\tilde{\mathbf{z}}$ is taken and subtracted from it the DCT of the original image $\mathbf{x}$.

As the blocks labeled with IDCT and 8x8 BDCT are linear operations, $\mathbf{f}$ can be related to $\mathbf{y}$ via

$$\mathbf{f}\ = \mathbf{H}\,\mathbf{y}\,, \tag{3}$$

where $\mathbf{H}$ is the linear transformation matrix. Similarly,

$$\tilde{\mathbf{y}}\ = \mathbf{H}^{-1}\tilde{\mathbf{f}} \tag{4}$$

The major source of error is due to quantization of the 8x8 block DCT coefficients in JPEG compression. The error introduced can be approximated as an additive noise $\mathbf{n_f}$, whose statistics depend in general on that of the input source and the quantization step size Δ. The value of Δ is position dependent in a 8x8 block. DCT coefficients of images without watermark can be modeled by Generalized Gaussian or Laplacian distribution [3]. When an image independent watermark is embedded in the image, the marked image can be viewed as dithered signal added before the quantizer. Dithered quantizer was analyzed in [4]. However, the watermark inserted according to Eq. 1, is dependent of the original image, making the analysis of [4] invalid.

3.1. Watermark absent

When there is no watermark, the first and second order statistics of the quantization noise $\mathbf{n_f}$ are given by

$$\begin{aligned} E\{n_{f_0}\} &= -j\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{(-1)^n}{u}\Phi_f(u)\ |_{u=\frac{2n\pi}{\Delta}}, \qquad (5)\\ E\{n_{f_0}^2\} &= \frac{\Delta^2}{12}+\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{2(-1)^n}{u^2}\Phi_f(u)\ |_{u=\frac{2n\pi}{\Delta}}, \end{aligned}$$

where $\Phi_f(u)$ is the characteristic function of the block DCT coefficients $\mathbf{f}$. Since Laplacian or generalized Gaussian distribution will be used to calculate the characteristic function, the quantization error in this case has zero mean.

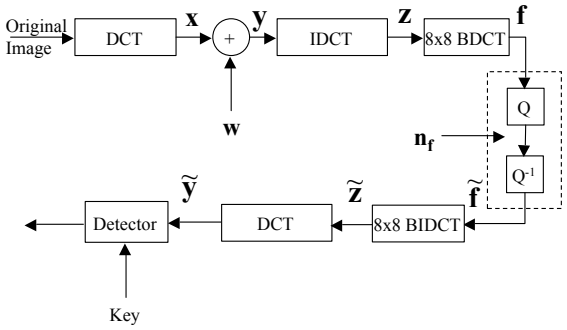

Fig. 1. Watermarked image going through JPEG compression-decompression

3.2. Watermark present

When watermark is present, we keep the notation $\mathbf{f}$ to represent the block DCT coefficients of the original image. The quantity to be quantized is now

$$\begin{aligned} \mathbf{f_w} &= \mathbf{H}\,\mathbf{y}\ = \mathbf{H}\,(\mathbf{x}\,(\mathbf{1}+\boldsymbol{\alpha}\,\mathbf{w}\,)) \qquad (6)\\ &= \mathbf{H}\,\mathbf{x}\ + \mathbf{H}\,(\boldsymbol{\alpha}\,\mathbf{w}\,\mathbf{x}\,) = \mathbf{H}\,\mathbf{x}\ + \mathbf{w_H}\,, \end{aligned}$$

where $\mathbf{w_H}\ = \mathbf{H}\,(\boldsymbol{\alpha}\,\mathbf{w}\,\mathbf{x}\,)$.

Note that $\mathbf{w}$ is Gaussian with zero mean. So for a given image, $\mathbf{f_w}$ is Gaussian with mean $\mathbf{H}\,\mathbf{x}$ and certain variance that can be computed from $\mathbf{w_H}$. We do so by modeling the blocks individually. As in the no-watermark case, the first and second moments of the quantization error can be calculated from

$$\begin{aligned} E\{n_{f_1}\} &= -j\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{(-1)^n}{u}\Phi_{f_w}(u)\ |_{u=\frac{2n\pi}{\Delta}}, \qquad (7)\\ E\{n_{f_1}^2\} &= \frac{\Delta^2}{12}+\sum_{\substack{n=-\infty\\ n\neq 0}}^{\infty}\frac{2(-1)^n}{u^2}\Phi_{f_w}(u)\ |_{u=\frac{2n\pi}{\Delta}},\\ Var\{n_{f_1}\} &= E\{n_{f_1}^2\}-E\{n_{f_1}\}^2, \end{aligned}$$

where $\Phi_{f_w}(u)$ is the characteristic function of $\mathbf{f_w}$, which can be obtained by computing the mean $\mathbf{H}\,\mathbf{x}$ and the variance of $\mathbf{w_H}$.

Note that $\mathbf{n_{f_1}}$ and $\mathbf{w_H}$ are dependent. It will be seen later in Section 5 that this dependency will affect the performance of the MAP detector. The correlation between $\mathbf{n_{f_1}}$ and $\mathbf{w_H}$ can be obtained by

$$E(n_{f_1}\cdot w_H) = \frac{d^2}{dudv}\Phi_{\tilde{f}_w f_w}\ |_{u=0,v=0} - E\{{f_w}^2\} - fE\{n_{f_1}\}, \tag{8}$$

where f is the block DCT coefficient of the original image,

$$\Phi_{\tilde{f}_w f_w} = \sum_{n=-\infty}^{\infty} \Phi_{f_w f_w}(u + \frac{2n\pi}{\Delta} v) \frac{\sin\frac{\Delta}{2}(u + \frac{2n\pi}{\Delta})}{\frac{\Delta}{2}(u + \frac{2n\pi}{\Delta})}. \quad (9)$$

$\Phi_{f_w f_w}(u, v)$ can be calculated from the auto-correlation function of $\mathbf{f_w}$.

The quantization noise $\mathbf{n_f}$ propagates to the input of the detector via linear transformation $\mathbf{H}$. Therefore, if the noise at the input of the detector $\mathbf{n_y}$ is $\tilde{\mathbf{y}} - \mathbf{y}$, its covariance matrix $\mathbf{C_{n_y}}$ can be expressed as

$$\mathbf{C_{n_y}} = \mathbf{H}^{-1}\mathbf{C_{n_f}}\mathbf{H}, \quad (10)$$

where $\mathbf{C_{n_f}}$ is the covariance matrix for $\mathbf{n_f}$.

4. MAXIMUM-A-POSTERIOR (MAP) DETECTOR

A MAP detector which uses statistical knowledge of the signal and noise would give superior performance over a correlation detector. The detector tests hypothesis H_0 (watermark not present) against hypothesis H_1 (watermark present):

$$\begin{cases} H_0 : \ no\ watermark\ (\hat{\mathbf{w}} = \mathbf{n_0}), \\ H_1 : \ watermark\ present\ (\hat{\mathbf{w}} = \mathbf{w} + \mathbf{n_1}). \end{cases} \quad (11)$$

Here $\mathbf{n_0}$ is the noise at the detector when there is no watermark and $\mathbf{n_1}$ is the noise when watermark is present. The detector uses the following log likelihood function

$$l(\hat{\mathbf{w}}) = \ln \frac{p_1(\hat{\mathbf{w}}\,|H_1)}{p_0(\hat{\mathbf{w}}\,|H_0)}, \quad (12)$$

where p_1 and p_0 are respectively the probability density functions of the test watermark ($\hat{\mathbf{w}}$) under the hypotheses H_0 and H_1. If the noises $\mathbf{n_0}$ and $\mathbf{n_1}$ are assumed to follow Gaussian distributions with covariance matrix $\mathbf{C_0}$ and $\mathbf{C_1}$ respectively, Eq. 12 reduces to

$$l(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^T\mathbf{C_0}^{-1}\hat{\mathbf{w}} - (\hat{\mathbf{w}} - \mathbf{w})^T\mathbf{C_1}^{-1}(\hat{\mathbf{w}} - \mathbf{w}). \quad (13)$$

The matrices $\mathbf{C_0}$ and $\mathbf{C_1}$ are quite large. For images of size 256×256, $\mathbf{C_0}$ and $\mathbf{C_1}$ are $(256)^2 \times (256)^2$, which would require 32GB storage for double-precision computation. To simplify subsequent calculations, we shall use the approximation that $\mathbf{C_0} = \mathbf{C_1} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2)$, where M is the number of rows/columns in $\mathbf{C_0}$ and $\mathbf{C_1}$.

That $\mathbf{C_0}$ and $\mathbf{C_1}$ are diagonal can be justified by observing that the noises $\mathbf{n_0}$ and $\mathbf{n_1}$ were generated by quantization and passed through the DCT step to reach the input of the detector and that DCT has the effect of de-correlation. That $\mathbf{C_0} = \mathbf{C_1}$ can be justified because invisible watermark is much smaller than the host signal so it is not expected to change the variance of the quantization error significantly. With this approximation, Eq 13 reduces to

$$l(\hat{\mathbf{w}}) = \sum_{i=1}^{K} \frac{\hat{w}_i w_i}{\sigma_i^2} / \sum_{i=1}^{K} \frac{w_i^2}{\sigma_i^2}, \quad (14)$$

where K is the length of the inserted watermark, and the term w_i^2/σ_i^2 is due to normalization.

Given a JPEG image with known compression quality factor (which determines the quantization matrix), the overall detector function can be computed as follows. First extract the watermark $\hat{\mathbf{w}}$ in DCT domain by subtracting the original image. Then we calculate the covariance matrix for $\mathbf{n_f}$ as in Figure 1, using Eqs. 5 and 7. The covariance matrices of the noise at the detector input can then be obtained by Eq. 10. σ_i^2 in Eq. 14 are just the diagonal elements of the resulting matrix. In fact, the intermediate computation can be simplified since we only need the diagonal elements of the final covariance matrices.

5. MAP DETECTOR PERFORMANCE ANALYSIS

When watermark is present, we have

$$\tilde{w}_i = w_i + n_i. \quad (15)$$

Eq. 14 becomes

$$l(\hat{\mathbf{w}}) = 1 + \sum_{i=1}^{N} \frac{n_i w_i}{\sigma_i^2} / \sum_{i=1}^{N} \frac{w_i^2}{\sigma_i^2}. \quad (16)$$

This is a sum of a large number of nearly independent variables, it may be approximated by a Gaussian variable, $l(\hat{\mathbf{w}}) \sim N(\mu_{l_1}, \sigma_{l_1}^2)$. Normalizing the variance of inserted watermark to 1, μ_{l_1} and $\sigma_{l_1}^2$ can be calculated by

$$\begin{aligned} \mu_{l_1} &= 1 + \sum_{i=1}^{N} \frac{E\{n_i w_i\}}{\sigma_i^2} / \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad (17) \\ \sigma_{l_1}^2 &= 1 / \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \end{aligned}$$

where $E\{n_i w_i\}$ can be obtained by first computing the noise-watermark correlation at the quantizer output from Eq. 8, then calculating the detector input noise-watermark covariance matrix by using Eq. 10.

Similarly, Under hypothesis H_0,

$$\begin{aligned} \mu_{l_0} &= 0, \quad (18) \\ \sigma_{l_0}^2 &= 1 / \sum_{i=1}^{N} \frac{1}{\sigma_i^2}. \end{aligned}$$

As a result, P_D and P_F can be expressed by

$$P_D = Q(\frac{\eta - \mu_{l_1}}{\sigma_{l_1}}), \quad (19)$$
$$P_F = Q(\frac{\eta}{\sigma_{l_0}}).$$

where $Q(x) \triangleq \frac{1}{2\pi}\int_x^\infty e^{-t^2/2}\,dt$.

6. RESULTS

A gray level Lena image of size 256×256 is used to test our theoretical analysis. The JPEG compression quality factor is set at 50%. Figure 2 shows the theoretical estimation of P_D and P_F of the MAP detector, as well as the simulation result. Note that the mean of the detector likelihood function $l(\hat{\mathbf{w}})$ is the threshold value at which $P_D = 0.5$. The ideal mean μ_{l_1} should be 1 because of normalization. However, it can be seen from the graph that the mean is 'shifted' from 1. This is due to the correlation between the quantization noise and the inserted watermark.

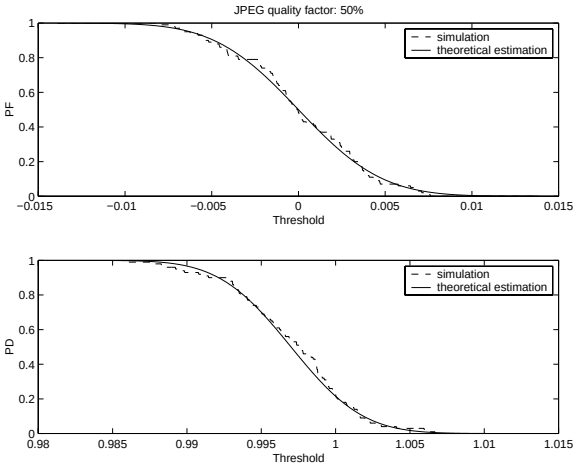

Fig. 2. MAP detector performance after JPEG: theoretical estimation vs. simulation

Figure 3 compares the simulation result of the false alarm P_F and detection probabilities P_D as functions of detector threshold for the MAP detector and the commonly used correlation detector. It is clear that the MAP detector has higher P_D and has lower P_F over the range where P_F is small.

7. CONCLUSIONS

We presented statistical modeling for JPEG compression noise in image watermarking. MAP detector was proposed to improve detector performance. Theoretical estimation of the detector performance was also given. Similar analysis approach can be applied to other watermarking schemes, such as multi-bit embedding.

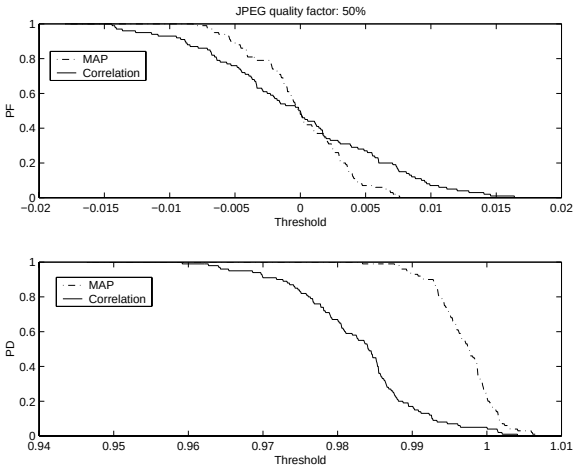

Fig. 3. Detector performance comparison with JPEG compression: MAP vs. Correlation

8. REFERENCES

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *2nd workshop on information hiding*, Portland, Oregon, Apr. 1998, pp. 218–238.

[2] I. J. Cox, J. K. F. T. Leighton, and T. Shamoon, "Secure spread specturm watermarking for multimedia," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1673–1686, 1997.

[3] Julia Minguillon and Jaume Pujol, "Uniform quantization error for laplacian sources with applications to jpeg standard," in *SPIE Conference on Mathematics of Data/Image Coding, Compression, and Encryption*, San Diego, California, Jul. 1998, pp. 77–88.

[4] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. on Communication Technology*, , no. 12, pp. 162–165, 1964.