Statistical Modelling and Steganalysis of DFT-Based Image Steganography

Ying Wang and Pierre Moulin[a]

[a]Beckman Institute, Coordinate Science Lab
and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA

Note: This is a revised version of the original SPIE 2006 paper in which a mistake was made in normalizing features before feeding them to the classifier: features from cover images and features from stego images were normalized differently. Due to this mistake, extra information about image classes was introduced to classification and the result was exceptionally good—the detection rate is close to 100%. We correct the mistake in this revision and most changes are made at the end of Sec. 4 and in Sec. 5 to present the correct results. Corresponding changes are also made in the conclusion (Sec. 6) while other sections remain intact.

ABSTRACT

An accurate statistical model of cover images is essential to the success of both steganography and steganalysis. We study the statistics of the full-frame two-dimensional discrete Fourier transform (DFT) coefficients of natural images and show that the independently and identically distributed model with unit exponential distribution is not a sufficiently accurate description of the statistics of normalized image periodograms. Consequently, the stochastic quantization index modulation (QIM) algorithm that aims at preserving this model is detectable in principle. To discriminate the resulted stegoimages from cover images, we train a learning system on them. Building upon a state-of-the-art steganalysis method using the statistical moments of wavelet characteristic functions, we propose new features that are more sensitive to data embedding. The addition of these features significantly improves the steganalyzer's receiver operating characteristic (ROC) curve.

Keywords: Steganalysis, steganography, full-frame DFT, statistical modelling, stochastic QIM, machine learning

1. INTRODUCTION

Hiding data in a transform domain is a popular practice in image watermarking. Commonly used transforms are the discrete wavelet transform (DWT), the discrete cosine transform (DCT), and the discrete Fourier transform (DFT). The above transforms may be applied to image blocks or to the whole image frame. Any hiding technique used in watermarking is a potential candidate for covert communication, i.e., steganography.

In recent work, the authors quantified the detectability of block-structured stegotext[1] and developed a steganalysis method based on the blockiness introduced by block-DCT embedding to cover images.[2] Fridrich[3] also proposed a steganalysis method based on increased blockiness to detect block-DCT embedding in JPEG images. Since now there are good methods for detection of block-based steganography, this motivates us to explore *full-frame transforms* as an alternative for steganographic embedding.

Moulin and Briassouli[4] proposed to use the stochastic quantization index modulation (QIM) algorithm[1] and embed data into the normalized two-dimensional (2-D) periodogram, i.e., the normalized absolute square of the 2-D DFT coefficients. The method was designed for watermark detection but can be easily modified and applied to steganographic data hiding scenarios. Based on the assumption that the normalized periodogram is independently and identically distributed (i.i.d.) with unit exponential distribution, this embedding is undetectable, hence a perfect steganography method. Results show that this embedding technique is robust against attacks such as blurring, additive and multiplicative noise, cropping, and warping.

Following up on Moulin and Briassouli's work, this paper focuses on the statistical modelling of the full frame DFT coefficients of natural images and the steganalysis of full frame DFT embedding. From the spectral theory of wide-sense stationary random processes,[5] it is known that the aforementioned i.i.d exponential model is accurate if the underlying image is a 2-D stationary Gaussian random field. However, most natural images deviate from this model: they are realizations of non-Gaussian and/or nonstationary processes. Nevertheless, preliminary result of Moulin and Briassouli[4] suggested that the above DFT-domain model approximately holds. This paper shows that this approximation is not good enough for steganography. To show this, we first evaluate the first- and higher-order statistics of the normalized image periodogram in Sec. 2. We briefly introduce the stochastic QIM embedding method in image periodograms in Sec. 3. In Sec. 4, we study the change in statistics caused by DFT embedding and propose enhancements to a feature-learning based nonparametric steganalysis method originally used by Xuan *et al.*.[6] We present steganalysis results by our proposed image features on 2-D DFT stochastic QIM embedding in Sec. 5. The improvement on steganalysis performance shows that our feature selection is indeed more sensitive to the noise imposed by data embedding to natural cover images. Finally, the paper is concluded in Sec. 6.

2. STATISTICAL MODEL OF 2-D DFT COEFFICIENTS OF NATURAL IMAGES

Let $\Omega = \{0, 1, \ldots, N_1 - 1\} \times \{0, 1, \ldots, N_2 - 1\}$ be the domain over which a gray-scale natural cover image $\mathbf{s} = \{s(n_1, n_2)\}$, $(n_1, n_2) \in \Omega$, is defined.

For any image $\mathbf{s}$, its 2-D DFT $\mathbf{S} = \{S(k_1, k_2)\}$ is defined as

$$S(k_1, k_2) = \sum_{(n_1,n_2)\in\Omega} s(n_1, n_2) e^{-j2\pi(k_1 n_1/N_1 + k_2 n_2/N_2)} \triangleq |S(k_1, k_2)| e^{\phi_S(k_1,k_2)},\ (k_1, k_2) \in \Omega, \tag{1}$$

where $\phi_S(k_1, k_2)$ is the DFT phase. The periodogram $I(k_1, k_2)$ is the squared magnitude of 2-D DFT coefficients

$$I(k_1, k_2) = |S(k_1, k_2)|^2. \tag{2}$$

2.1. Image Model

First assume that the image $\mathbf{s}$ is a 2-D stationary Gaussian random field

$$s(n_1, n_2) = \sum_{(m_1,m_2)} g(m_1, m_2)\epsilon(n_1 - m_1, n_2 - m_2), \tag{3}$$

where $\{\epsilon(n_1, n_2)\}$, $(n_1, n_2) \in \Omega$, is an i.i.d Gaussian random process with zero mean, finite variance σ_ϵ^2, and finite fourth moment; and $\{g(m_1, m_2)\}$ are linear prediction coefficients that are absolutely summable.

Define the normalized periodogram as

$$U(k_1, k_2) = \frac{I(k_1, k_2)}{4\pi H(k_1, k_2)},\ (k_1, k_2) \in \Omega, \tag{4}$$

where

$$H(k_1, k_2) = (\sigma_\epsilon^2/2\pi) \cdot |G(k_1, k_2)|^2 \tag{5}$$

is the spectral density function of the random process $\mathbf{s}(n_1, n_2)$ and

$$G(k_1, k_2) = \sum_{(m_1,m_2)} g(m_1, m_2) e^{-j(\omega_1 m_1 + \omega_2 m_2)} \Big|_{\substack{\omega_1 = 2\pi k_1/N_1 \\ \omega_2 = 2\pi k_2/N_2}} \tag{6}$$

is the transfer function of the linear prediction system.

Asymptotically, the set of random variables $\{U(k_1, k_2)\}$ are *independently* distributed.[5] When $k_1 \neq 0, N_1/2$ and $k_2 \neq 0, N_2/2$ (N_1 and N_2 even), we obtain the asymptotic distribution of $U(k_1, k_2)$

$$U(k_1, k_2) \overset{i.i.d}{\sim} \frac{1}{2}\chi_2^2, \tag{7}$$

where $\frac{1}{2}\chi_2^2$ is the unit exponential distribution: $p_U(u) = e^{-u}$ for $u \geq 0$. From (7), we obtain the asymptotic approximation of the expectation and variance of periodogram $I(k_1, k_2)$ as

$$E[I(k_1, k_2)] \sim 4\pi H(k_1, k_2), \quad \text{var}[I(k_1, k_2)] \sim 16\pi^2 H^2(k_1, k_2). \tag{8}$$

The derivation for the properties of the normalized periodogram of 1-D Gaussian random linear processes can be found in the Chapter 6 of.[5] The derivation can be generalized to obtain above properties for 2-D Gaussian random fields as well.

When the spectral density function $H(k_1, k_2)$ is smooth, we can approximate $E[I(k_1, k_2)]$ at one frequency point by averaging the periodogram over adjacent points:

$$\hat{E}[I(k_1, k_2)] = \frac{1}{(2J_1 + 1)(2J_2 + 1)} \sum_{k_1 - J_1 \leq l_1 \leq k_1 + J_1, k_2 - J_2 \leq l_2 \leq k_2 + J_2} I(l_1, l_2). \tag{9}$$

For the experiments in this paper, we take $J_1 = J_2 = 3$.

Let the estimated normalized periodogram be

$$\hat{U}(k_1, k_2) = I(k_1, k_2) / \hat{E}[I(k_1, k_2)], \tag{10}$$

we have approximately

$$\hat{U}(k_1, k_2) \approx U(k_1, k_2) \overset{i.i.d}{\sim} \frac{1}{2}\chi_2^2. \tag{11}$$

When the random field $\mathbf{s}(n_1, n_2)$ is stationary but not Gaussian, it can still be proved that $U(k_1, k_2)$ is asymptotically i.i.d exponential if $\mathbf{s}(n_1, n_2)$ has a rapidly decreasing autocorrelation function.[5] When the random field is neither stationary nor Gaussian, which is the case for most images, all conditions that are required to derive the i.i.d unit exponential model of (7) are violated, and we expect that $\hat{U}(k_1, k_2)$ is further away from the above model.

In the remainder of this section, we test the fitness of the i.i.d unit exponential model to natural images. Our test image set consists of 1370 256×256 8-bit graylevel images, which include standard test images such as *Lena* and images from the Uncompressed Colour Image Database (UCID) collected by Schaefer and Stich.[7] Images in our test image set contain a wide range of outdoor/indoor and daylight/night scenes, including nature (eg. landscapes, trees, flowers, animals), portraits, man-made objects (eg. ornaments, kitchen tools, architectures, cars, signs, neon lights), and so on.

2.2. Histogram

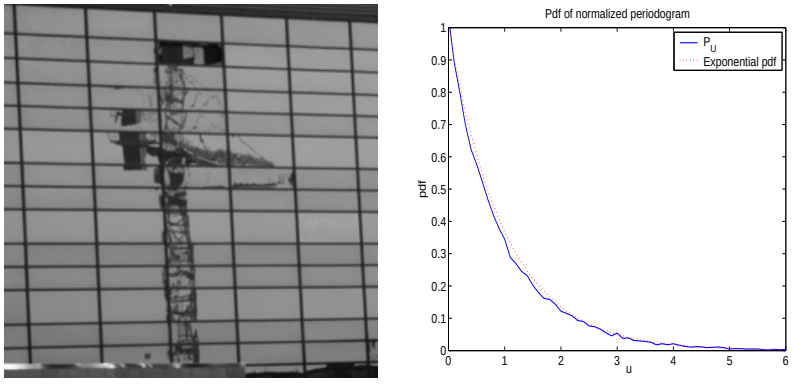


Figure 1. An image of reflective glass on a building and the histogram of its estimated normalized periodogram.

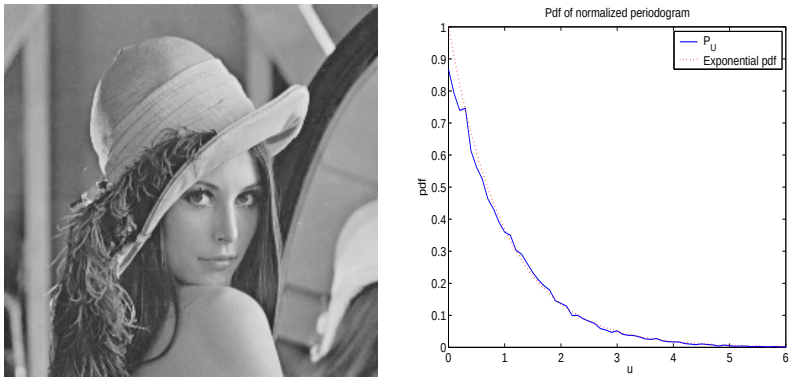


Figure 2. Image *Lena* and the histogram of its estimated normalized periodogram.

We first look at the first order statistics — $p_{\hat{U}}(u)$, the histogram of the estimated normalized periodogram $\hat{U}(k_1, k_2)$. Fig. 1 shows an image of reflective glass on a building and the histogram of its $\hat{U}(k_1, k_2)$. We see that the fit to the exponential probability density function (pdf) is very good. Except for the lattice structure and the reflection of a tower crane, most of the pixels in the image are gray background with stationary variations. Hence, the good fit of $p_{\hat{U}}(u)$ to the exponential pdf is not surprising. Fig. 2 shows the *Lena* image and the histogram of its $\hat{U}(k_1, k_2)$. The fit to the exponential pdf is slightly worse. *Lena* has large areas of background too. But, it is clear that it has many non-stationary elements such as the hat, the feather, and so on. We found that for most images, the histogram of $\hat{U}(k_1, k_2)$ is well fitted by a unit exponential pdf when $\hat{U}(k_1, k_2) \geq 2$, while there is a relatively large discrepancy for small values of $\hat{U}(k_1, k_2)$. The discrepancy is largest around $u = 0$.

We use the one-sample Kolmogorov-Smirnov (K-S) test[8] to quantify the goodness-of-fit of the i.i.d unit exponential distribution to $p_{\hat{U}}(u)$. Fig. 3 shows the percentage of test images versus their K-S test statistics. Among the 1370 test images, only 4.45% of them are accepted to have a unit exponential pdf with a significance level of 0.05, and only 8.1% even with a significance level of 0.01.* Thus for most images, the i.i.d unit exponential distribution model is not accurate in the sense of the K-S goodness-of-fit test.

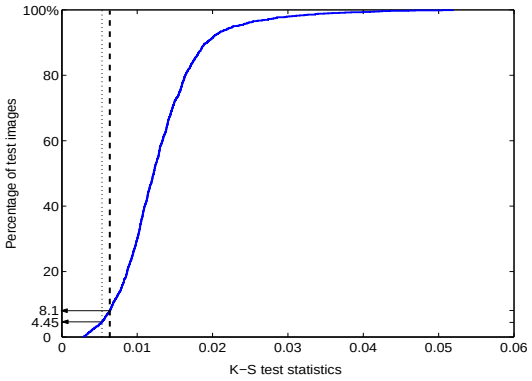


Figure 3. Percentage of test images vs. their K-S test statistics. The dotted line corresponds to the K-S test statistics threshold (0.00531) at the significance level of 0.05 and the dashed line corresponds to the K-S test statistics threshold (0.00637) at the significance level of 0.01.

*The significance level is the probability of rejecting the i.i.d unit exponential distribution model when it is true. The higher the significance level is, the test is more aggressive at rejecting the i.i.d unit exponential distribution model and more conservative on accepting it.

2.3. Higher-Order Statistics

To quantify the dependencies between adjacent normalized periodogram samples — a second or higher order statistics, we consider both the correlation coefficient and the mutual information. Given (U, V) and their joint pdf $p(u, v)$, the linear correlation coefficient is defined as

$$r(U;V) = \frac{E[UV] - E[U] \cdot E[V]}{\sqrt{var(U) \cdot var(V)}}. \tag{12}$$

The mutual information is defined as

$$I(U;V) = \int_{U \in \mathcal{U}} \int_{V \in \mathcal{V}} p(u,v) \log \frac{p(u,v)}{p(u)p(v)} du dv. \tag{13}$$

Although a linear correlation coefficient is sufficient for the dependency in Gaussian distributions and directly related to mutual information, it is typically insensitive to dependencies that do not manifest themselves through covariances. On the other hand, mutual information is zero if and only if the two random variables are truly independent. Liu and Moulin studied the mutual information between intra- and inter-scale wavelet coefficients to characterize the dependency among them and answered the question of which dependency, intra- or inter-scale, is larger.[9] We take a similar approach here.

We consider two groups of bivariate measurements from the image DFT coefficients: $\hat{U}(k_1, k_2)$ and its right neighbor

$$T_1 = \hat{U}(k_1, k_2 + 1), \tag{14}$$

and $\hat{U}(k_1, k_2)$ and the average of its eight neighbors

$$T_2 = \frac{1}{8} \sum_{-1 \le i \le 1, -1 \le j \le 1, (i,j) \ne (0,0)} \hat{U}(k_1 + i, k_2 + j). \tag{15}$$

Let $\mathcal{N}\hat{U}$ denote the eight neighbors of $\hat{U}$,

$$\hat{U} \to \mathcal{N}\hat{U} \to T = f(\mathcal{N}\hat{U}) \tag{16}$$

forms a Markov chain. The function $f(\cdot)$ is a selection function in (14) and an average function in (15), both of which reduce the eight dimensions of $\mathcal{N}\hat{U}$ to the one dimension of T. The data-processing theorem[10] implies that

$$I(\hat{U};T) \le I(\hat{U};\mathcal{N}\hat{U}). \tag{17}$$

Hence, $I(\hat{U};T)$ provides a lower bound to $I(\hat{U};\mathcal{N}\hat{U})$, for which the estimation accuracy suffers the curse of high dimensionality.[11] The bound is tight only when the statistics T is sufficient.[10]

To obtain the estimated correlation coefficient, we replace the expectations in (12) with the means of measurements, and the variance with the standard deviation of measurements. The estimation of mutual information and entropy from measurements with small error is a hard problem.[12] We adopt the mutual information estimator developed by Kraskov *et al.*,[13] which is based on the k-nearest neighbor statistics[†]. Fig. 4 shows the percentage of test images versus their estimated correlation coefficients and mutual information. Both the estimated correlation coefficients —$r(\hat{U};T_1)$ and $r(\hat{U};T_2)$— have similar distribution among test images. However, $I(\hat{U};T_2)$ is significantly larger than $I(\hat{U};T_1)$: an evidence that the mutual information can pick up additional dependency between random variables that is not reflected by the correlation coefficient. When compared to the reference values calculated from two sets of exponential-distributed independent data, test images clearly have nonlinear dependencies among estimated normalized periodogram samples.

[†]The code for their mutual information estimator can be downloaded from http://www.fz-juelich.de/nic/cs/software/.

3. STOCHASTIC QIM EMBEDDING

Assuming an i.i.d unit exponential distribution model on the normalized periodogram, Moulin and Briassouli[4] embeds one bit into an image through the following steps:

1. Select a range $\mathcal{K}$ from all normalized periodogram samples and a set $\mathcal{K}^*$ out of $\mathcal{K}$. Let $\epsilon = \frac{|\mathcal{K}^*|}{|\mathcal{K}|}$ and $u^* = -\ln \epsilon$. $Pr[\hat{U} \geq u^*] = \epsilon$ if $\hat{U}$ is indeed i.i.d exponential with pdf $p_U(u)$. Define $p_{1a}(u) = \frac{1}{\epsilon} p_U(u) 1_{\{u \geq u^*\}}$ and $p_{1b}(u) = \frac{1}{1-\epsilon} p_U(u) 1_{\{0 \leq u \leq u^*\}}$, then $p_U = \epsilon p_{1a} + (1-\epsilon) p_{1b}$.
2. For $(k_1, k_2) \in \mathcal{K}^*$, let $\tilde{u}(k_1, k_2) = \hat{u}(k_1, k_2)$ if $\hat{u}(k_1, k_2) \geq u^*$; otherwise generate $\tilde{u}(k_1, k_2)$ from the pdf p_{1a}. For $(k_1, k_2) \in \mathcal{K} \setminus \mathcal{K}^*$, let $\tilde{u}(k_1, k_2) = \hat{u}(k_1, k_2)$ if $0 \leq \hat{u}(k_1, k_2) \leq u^*$; otherwise generate $\tilde{u}(k_1, k_2)$ from the pdf p_{1b}.
3. The new DFT coefficient is $\tilde{s}(k_1, k_2) = \sqrt{\hat{E}[I(k_1, k_2)] \tilde{u}(k_1, k_2)} e^{\phi_S(k_1, k_2)}$.

We can check that the pdf of $\tilde{U}(k_1, k_2)$ is also the exponential pdf $p_U(u)$. Therefore, the above embedding is a perfect steganographic method if the samples $\hat{U}(k_1, k_2)$ are indeed i.i.d exponential. However, from Sec. 2, we learned from test images that the i.i.d exponential model for $\hat{U}(k_1, k_2)$ has limited accuracy. Although the above stochastic QIM embedding intends to preserve the assumed image statistical model, the embedding disrupts the true statistics of the original normalized periodogram. This change in statistics caused by embedding might be detectable. In practice, however, we need to construct a suitable detector.

4. FEATURE-BASED STEGANALYSIS METHOD

Since the normalized periodogram is a function of the DFT coefficients, the discrimination between the original and modified DFT coefficients is no less than the discrimination between the original and modified periodograms according to the data processing inequality.[10] The discrimination between two pdfs p_1 and p_2 is defined as

$$D(p_1 || p_2) = \int p_1(x) \cdot \log \frac{p_1(x)}{p_2(x)} \, dx.$$

We could develop a statistically optimal test on the DFT coefficients if we knew their before-and-after statistics. However, in reality this is not the case. Our approach is to develop and train a learning system on both cover and stegoimages in order to discriminate between them. The features used to represent the unknown statistics should have manageable size and be sensitive to changes in statistics.

Before we select appropriate features, we first characterize the changes in image statistics due to data embedding.

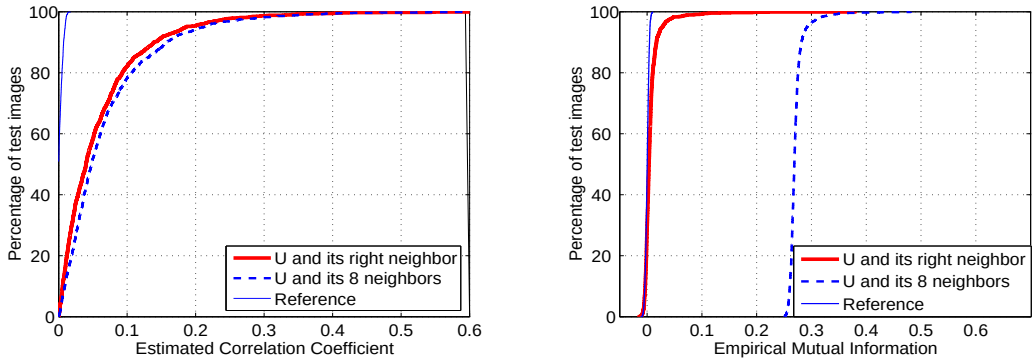


Figure 4. Percentage of test images versus estimated correlation coefficient (left) and empirical mutual information (right). The thick solid lines are for the estimated normalized periodogram sample $\hat{U}(k_1, k_2)$ and its right neighbor; the dashed lines are for $\hat{U}(k_1, k_2)$ and the average of its eight neighbors; and the thin solid lines are calculated from exponential-distributed independent bivariate measurements and for the purpose of reference.

4.1. Statistics of Image Wavelet Coefficients

Since the DFT is invertible, we do not lose any information when we transform back to the image pixel domain. We observe the difference between the cover image $\mathbf{s}$ and the stegoimage $\mathbf{x}$ resulted from the stochastic QIM embedding introduced in Sec. 3, and find that the information-bearing embedding noise

$$\mathbf{n} = \mathbf{x} - \mathbf{s}$$

has a 1-D histogram that can be fitted nearly perfectly by a discretized Gaussian probability mass function (pmf). However, the embedding noise is not white Gaussian. Its empirical correlation structure is similar to that of the cover image, suggesting that the DFT embedding method of[4] does adapt partially to the cover image statistics. The steganalysis of DFT embedding is reduced to the steganalysis of data hiding with colored Gaussian embedding noise.

Steganalysis of full-frame data embedding, in either spatial or transform domain, has been studied by Farid,[14] Harmsen and Pearlman,[15] Sullivan *et al.*,[16] and Xuan *et al.*[6] Although they use different features in their learning system, the essence is to exploit the dependency between image pixels. Wavelet transform is often used to decorrelate image pixels. The resulted wavelet coefficients in high-pass subbands are well modelled by a generalized Gaussian distribution (GGD)[17]

$$p_S(s) \approx p_{\alpha,\beta}(s) \triangleq \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \exp\left\{-\left(\frac{|s|}{\alpha}\right)^{\beta}\right\}, \tag{18}$$

where $\Gamma(\cdot)$ is the Gamma function, $\alpha > 0$ is the scale parameter, and $\beta > 0$ is the shape parameter. Most natural images have distributions with $\beta \leq 1$, i.e. Laplacian-like: the large amount of small magnitudes in smooth areas form the sharp peak at zero, while a few large magnitudes in "busy" areas give the heavy tails. For example, we estimate that a GGD with $\alpha = 1.9$ and $\beta = 0.68$ is a good fit for the histogram of the first-level diagonal subband coefficients of the *Lena* image when the Haar wavelet transform is used (Fig. 5(a)). The fitting criteria is the maximum difference between two cumulative distribution functions.

We observe that the DFT-embedding noise has a nearly-Gaussian histogram $p_N(n)$ (Fig. 5(b)). The histogram of the stegoimage wavelet coefficients in high-pass subbands $p_X(x)$ is close to the convolution of the original image histogram $p_S(s)$ and the embedding noise histogram $p_N(n)$:

$$p_X(x) \approx (p_S * p_N)(x).$$

This approximation is accurate only when the embedding noise is independent of the cover image. Nevertheless, the stego-histogram (Fig. 5(c)) is smoothed as a result of the embedding: it still has heavy tails, but at the vicinity of $x = 0$ the curvature is much smaller compared to the original image histogram as the embedding noise strength increases. Through experiments, we found that $p_X(x)$ is well approximated by a Cauchy distribution with scale parameter b:

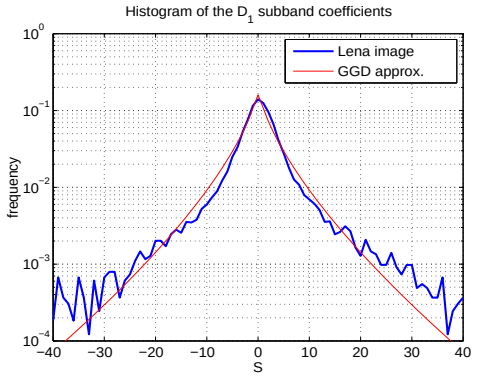
$$p_X(x) \approx p_b(x) \triangleq \frac{b}{\pi[x^2 + b^2]}.$$

$b = 2.79$ is a good estimate for the histogram of the first-level diagonal subband coefficients of a stego *Lena* image with PSNR [‡] of 38.8dB when the Haar wavelet transform is used (Fig. 5(c)).

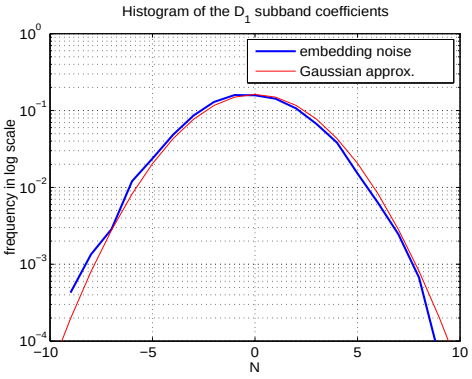
4.2. Feature Selection

Knowing the existence of differences between the histograms of original images and stegoimages, we could use histograms as features. Histograms of wavelet decomposition subbands do capture global image statistics fully, but at the cost of large memory requirements and computational complexity. Indeed, there are usually hundreds of bins in each histogram. So, we need to construct *low-dimensional* features that are informative for classification.

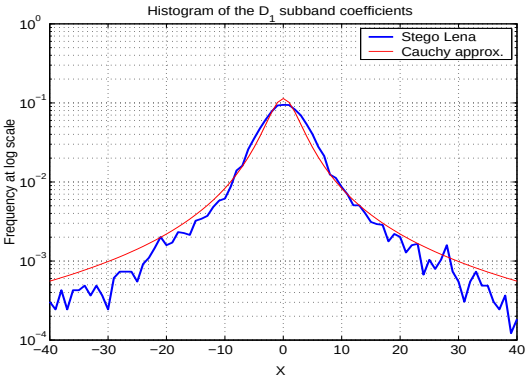
[‡]The peak-signal-to-noise ratio (PSNR) is defined as $PSNR_{(dB)} = 10 * \log_{10}[\frac{255^2}{\text{noise variance}}]$.



(a) Original *Lena*



(b) DFT-embedding noise



(c) Stego *Lena*

Figure 5. Histograms of the first-level diagonal subband coefficients of (a) the original *Lena* image, (b) the DFT-embedding noise with variance 8.6, and (c) a stego *Lena* image with PSNR 38.8dB. In each subfigure, the thick line is the histogram and the thin line is an analytical approximation.

For a histogram $\{h(m)\}$, where $m \in \{0, 1, \ldots, M\}$ denotes the m^{th} bin, its characteristic function is obtained through a K-point DFT

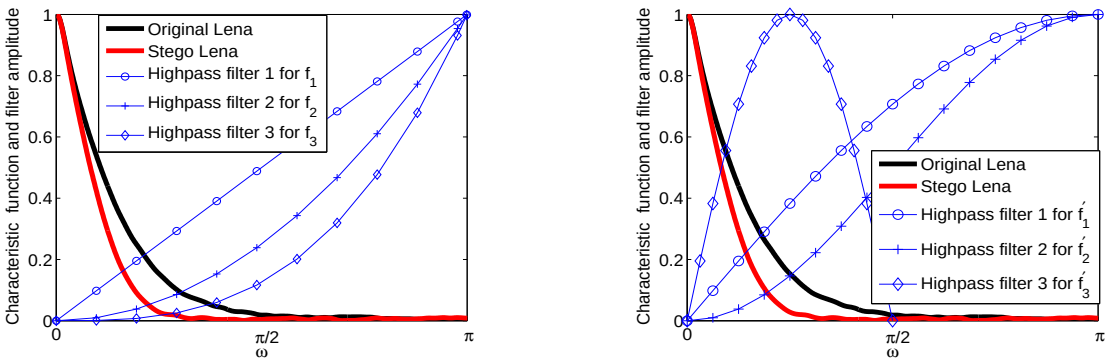
$$H(k) = \sum_{m=0}^{M-1} h(m) \exp\{-i2\pi mk/K\},\ 0 \leq k \leq K. \tag{19}$$

Fig. 6 shows the corresponding characteristic function amplitude plots for the histograms of original *Lena* image and stego *Lena* image in Fig. 5(a) and (c). Natural images have some significant mid- to high-frequency components due to the sharp peak in Laplacian-like distributions; stego images have fast decreasing mid- to high-frequency components while packing most of their energy in low-frequency components because of the smoother curvature at bin 0 in their Cauchy-like distributions.

Xuan *et al.*[6] proposed a selection of 39 features, consisting of the first three moments of the characteristic functions of the three-level Haar wavelet decomposition subbands (smooth LL_i, horizontal HL_i, vertical LH_i, diagonal HH_i, $i = 1, 2, 3$) as well as the test image (LL_0). They use the following three features for each histogram:

$$f_1 = \sum_{k=0}^{K/2} |H(k)| \cdot \frac{k}{K}, \tag{20}$$

$$f_2 = \sum_{k=0}^{K/2} |H(k)| \cdot \left(\frac{k}{K}\right)^2, \tag{21}$$



(a) Characteristic functions and highpass filters used by Xuan *et al.*

(b) Characteristic functions and our proposed new filters

Figure 6. Amplitude plots of characteristic functions and filters for calculating features.

and

$$f_3 = \sum_{k=0}^{K/4} |H(k)| \cdot \left(\frac{k}{K}\right)^3. \tag{22}$$

These features are obtained by essentially high-pass filtering the histogram $h(m)$, as shown in Fig. 6(a). f_2 is related to an upper bound on the second derivative (curvature) of the histogram at bin 0.[6] However, we observe that the largest difference between the characteristic functions of most original images and stegoimages lies at mid-frequencies. Especially, for the *Lena* images, both histograms do not have strong components at high frequencies. Hence, we propose three different features from each histogram: two that emphasize information about the high-frequency components and one that convey information about the low- to mid-frequency components. The three new features are defined as

$$f_1' = \sum_{k=0}^{K/2} |H(k)| \sin\left(\frac{\pi k}{K}\right), \tag{23}$$

$$f_2' = \sum_{k=0}^{K/2} |H(k)| \sin^2\left(\frac{\pi k}{K}\right), \tag{24}$$

and

$$f_3' = \sum_{k=0}^{K/4} |H(k)| \sin\left(\frac{4\pi k}{K}\right). \tag{25}$$

f_1' and f_2' are variants of f_1 and f_2 proposed by Xuan *et al.*, related to upper bounds on the first and second discrete derivatives (curvature) of the histogram at bin 0. f_3' is crucial to improve classification performance and—to the best of our knowledge—a new feature. From Fig. 6(b), we expect that for cover images, they have larger f_1', f_2' and f_3' than their corresponding stegoimages.

We also decompose an image into 3 levels using the Haar wavelet transform to obtain 12 subbands (lowpass LL_i, horizontal HL_i, vertical LH_i, diagonal HH_i, $i = 1, 2, 3$) as done by Xuan *et al.*[6] Therefore, we have a total of 39 features: 3 features for each of the 12 subbands plus the original image.

5. STEGANALYSIS RESULTS

To illustrate the effect of these new features, we experiment on our test image dataset: the 1370 8-bit graylevel images from the UCID and other standard test images. We use the stochastic QIM DFT embedding method[4] to generate 4 stegoimages from each cover image. Therefore, we have 5480 stegoimages with effective PSNR ranging from 21 to 70 dB. Fig. 7 shows the PSNR distribution of the generated stegoimages.

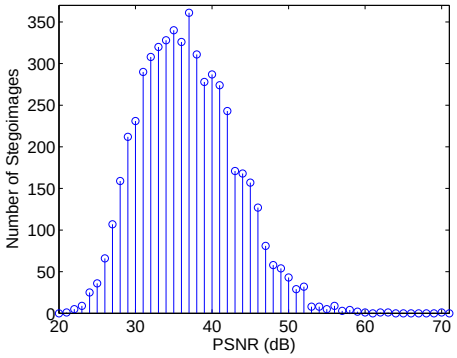


Figure 7. The number of stegoimages versus PSNR.

We randomly choose 800 cover images for training and the remaining 570 images for testing. This random splitting is repeated for 1000 times in order to avoid flukes for specific training/testing splits. We take out 1355 stegoimages with PSNR of 34-38 dB for training and test on the remaining stegoimages.

The Gaussian Bayes classifier[18] is adopted for training and testing. We scale the features so that they have comparable dynamic ranges. The (correct) scaling is done as follows. For any feature f, we find its maximum value $f_{\max}$ and its minimum value $f_{\min}$ from all the training images (including both cover images and stegoimages). For any image, the feature f is extracted and scaled according to[§]

$$\tilde{f} = \frac{f - f_{\max}}{f_{\max} - f_{\min}}.$$

Then, $\tilde{f}$ is fed to the classifier. Clearly, for all the training images, $\tilde{f} \in [0, 1]$; for most of test images, it is expected that $\tilde{f}$ will also be between 0 and 1. This scaling step prevents some features that have large numeric ranges from dominating those that have small numeric ranges, avoids numerical ill-conditioning, and dramatically improves classification accuracy.[19]

Table 1 compares the false alarm probabilities (P_{FA}) and the detection probabilities (P_D) of training and testing when the decision threshold is 0, which corresponds to equal priors for stegoimages and cover images. The false alarm probability is the ratio of the cover images that are misclassified as stegoimages; and the detection probability is the ratio of the stegoimages that are correctly classified. Also, the false alarm and detection probabilities in Table 1 are averaged over the 1000 training/testing splits on cover images. Compared to Xuan *et al.*'s feature selection, our feature selection improves steganalysis performance slightly: except for the test false alarm probability, the training and test detection probabilities and the training false alarm probability are all better than Xuan et al.'s results when the decision threshold is fixed 0. By adjusting the decision threshold, we generate the receiver operating characteristic (ROC) curves shown in Fig. 8. Our ROC curve for this data set is again better than the result by Xuan *et al.*, especially when the test false alarm probability is smaller than 0.1.

6. CONCLUSIONS

In summary, we showed that the i.i.d unit exponential distribution model is not a sufficiently accurate description of the statistics of the normalized periodogram of the full-frame 2-D image DFT coefficients. Consequently, the stochastic QIM embedding[4] aimed at preserving the above model is detectable. Based on Xuan *et al.*'s steganalysis method[6] using the statistical moments of wavelet characteristic functions, we proposed new features that are more sensitive to the change of statistics resulted from the data embedding to cover images. With our new features, the classification ROC curve is improved.

[§]The mistake we made in the original SPIE 2006 paper was to find $f_{\max,1}$ and $f_{\min,1}$ from training cover images and $f_{\max,2}$ and $f_{\min,2}$ from training stegoimages. Then we used $f_{\max,1}$ and $f_{\min,1}$ to scale test cover images and $f_{\max,2}$ and $f_{\min,2}$ to scale test stegoimages. By doing so, we introduced into the classification prior knowledge about the class that the test images belong to.

Probability vs. Features	Xuan et al.'s 39-D features	Our 36-D features
$P_{FA,Training}$	0.0794	0.0688
$P_{FA,Test}$	0.0839	0.0965
$P_{D,Training}$ (PSNR=34-38 dB)	0.9976	0.9985
$P_{D,Test}$ (PSNR=20 $\sim$ 70dB)	0.7559	0.7668

Table 1. Performance comparison of the 39-D feature selection proposed by Xuan *et al.*[6] and our proposed 36-D feature selection. Note that the Gaussian Bayes classifier decision threshold is set at 0.

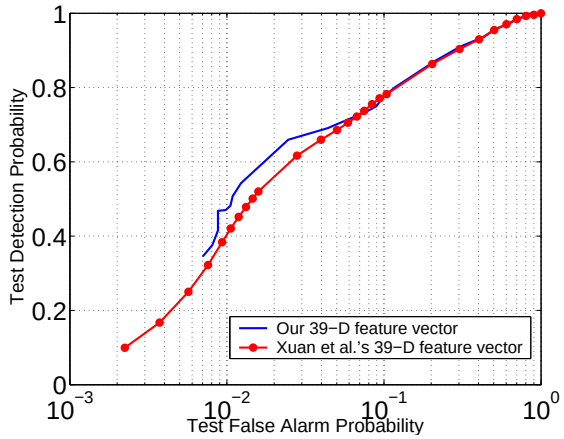


Figure 8. ROC curves for Xuan *et al.*'s 39-D steganalysis feature selection and our proposed 36-D feature selection. The training set consists of 800 randomly chosen cover images and 1355 stegoimages with PSNR of 34-38 dB. The test set consists of the remaining 570 cover images and 4125 stegoimages with PSNR distributed as shown by the histogram in Fig. 7.

REFERENCES

1. Y. Wang and P. Moulin, "Steganalysis of Block-Structured Stegotext," *Proc. SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, vol. 5306, pp. 477-488, San Jose, CA, Jan. 2004.
2. Y. Wang and P. Moulin, "Steganalysis of Block-DCT Steganography," *Proc. IEEE Workshop on Statistical Signal Processing*, pp. 339-342, St. Louis, MO, Sep. 2003.
3. J. Fridrich, "Feature-Based Steganalysis for JPEG Images and Its Implications for Future Design of Steganographic Schemes," *Proc. 6th Information Hiding Workshop*, pp. 67-81, Toronto, Canada, May 2004.
4. P. Moulin and A. Briassouli, "A Stochatic QIM Algorithm for Robust, Undetectable Image Watermarking," *Proc. ICIP*, vol. 2, pp. 1173-1176, Singapore, Oct. 2004.
5. M. B. Priestley, *Spectral Analysis and Time Series*. Orlando: Academic Press, Inc., 1981.
6. G. Xuan, Y. Q. Shi, J. Gao, D. Zou, C. Yang, Z. Zhang, P. Chai, C. Chen and W. Chen, "Steganalysis Based on Multiple Features Formed by Statistical Moments of Wavelet Characteristic Functions," *Proc. of Information Hiding Workshop (IHW05)*, Barcelona, Spain, June 2005.
7. G. Schaefer and M. Stich, "UCID - An Uncompressed Colour Image Database," *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia 2004*, pp. 472-480, San Jose, USA, Jan. 2004.
8. J. D. Gibbons, *Nonparametric Methods for Quantitative Analysis*. New York: Holt, Rinehart and Winston, 1976.
9. J. Liu and P. Moulin, "Information-Theoretic Analysis of Interscale and Intrascale Dependencies Between Image Wavelet Coefficients," *IEEE Trans. on Image Processing*, Vol. 10, No. 10, pp. 1647-1658, Nov. 2001.
10. R. E. Blahut, *Principles and Practice of Information Theory*. Preprint, 1999.
11. B. W. Silverman, *Density Estimation of Statistics and Data Analysis*. London, UK: Chapman and Hall, 1986.

12. J. Beirlant, E. J. Dudewicz, L. Gyorfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of the Mathematical Statistics Sciences*, no. 6, pp. 17-39, 2001.
13. A. Kraskov, H. Stgbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 066138-1-16, 2004.
14. H. Farid, "Detecting Hidden Messages Using Higher-order Statistical Models," *Proc. of IEEE International Conference on Image Processing*, Vol. 2, pp. 905-908, Rochester, New York, 2002.
15. J. J. Harmsen and W. A. Pearlman, "Steganalysis of Additive Noise Modelable Information Hiding," *Proc. of SPIE, Security and Watermarking of Multimedia Contents V*, vol. 5020, pp. 131-142, Santa Clara, CA, Jan. 2003.
16. K. Sullivan, U. Madhow, S. Chandrasekaran and B. S. Manjunath, "Steganalysis of Spread Spectrum Data Hiding Exploiting Cover Memory," *Proc. of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VII*, vol. 5681, pp. 38-46, San Jose, CA, Jan. 2005.
17. S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, July 1989.
18. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
19. C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification." Available at http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.