Steganalysis of additive noise modelable information hiding

Jeremiah J. Harmsen[a] and William A. Pearlman[a]

[a]Center for Image Processing Research,
Electrical Computer and Systems Engineering Department,
Rensselaer Polytechnic Institute, Troy, NY

ABSTRACT

The process of information hiding is modeled in the context of additive noise. Under an independence assumption, the histogram of the stegomessage is a convolution of the noise probability mass function (PMF) and the original histogram. In the frequency domain this convolution is viewed as a multiplication of the histogram characteristic function (HCF) and the noise characteristic function. Least significant bit, spread spectrum, and DCT hiding methods for images are analyzed in this framework. It is shown that these embedding methods are equivalent to a lowpass filtering of histograms that is quantified by a decrease in the HCF center of mass (COM). These decreases are exploited in a known scheme detection to classify unaltered and spread spectrum images using a bivariate classifier. Finally, a blind detection scheme is built that uses only statistics from unaltered images. By calculating the Mahalanobis distance from a test COM to the training distribution, a threshold is used to identify steganographic images. At an embedding rate of 1 b.p.p. greater than 95% of the stegoimages are detected with false alarm rate of 5%.

Keywords: Steganalysis, steganography, additive noise

1. DATA HIDING AS ADDITIVE NOISE

1.1. Motivation

The motivation to model the steganographic process as the addition of noise arises from a number of factors. In the process of sampling and transmitting signals there are numerous sources of noise such as quantization[1], sensor[2], and channel[3]. A number of steganographic hiding schemes have used this as a foundation for noise based data hiding. The goal is to disguise the message as a naturally present noise and add it to the coverimage.

While the additive noise framework is especially well suited to schemes which rely on noise based embedding, it can be easily generalized to any method which embeds data without consideration toward the covermessage. Sampled signals have a large amount of correlation present- both from the natural statistics of the signal and the sampling device. If data is hidden without regard to this correlation, it can be considered as an external force which corrupts the image. This formulation allows us to model many hiding methodologies which do not directly rely on additive noise.

1.2. Modeling

In additive noise modelable information hiding we model the embedding of a message as the addition of noise to the covermessage. The pseudo-noise containing the message is referred to as stegonoise. A block diagram of this framework is shown in Figure 1. We begin with the covermessage, which has a histogram $h_c[n]$. To that we add the stegonoise, which has a probability mass function (PMF) of $f_\Delta[n]$. This results in the stegomessage which has a histogram $h_s[n]$.

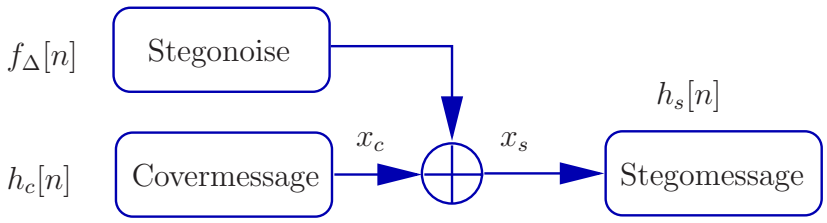

Figure 1. Additive Noise Steganography Model

1.3. Stegonoise Probability Mass Function

The stegonoise probability mass function is the distribution of the additive noise defined as,

$$f_\Delta[n] \triangleq p(x_s - x_c = n). \tag{1}$$

Where x_s is the pixel value after embedding, and x_c is the pixel value prior to embedding. Generally speaking, $f_\Delta[n]$ is the probability that a pixel will be altered by n. In this model it is assumed that the noise acts independently on each pixel. So $f_\Delta[0]$ is the probability that, after embedding, a pixel is unchanged. Whereas $f_\Delta[-1]$ is the probability that the pixel is decreased by one. Many times it is more convenient to work with a continuous probability density function, $f_\Delta(x)$, rather than the discrete probability mass function. Of course, when digital media is stored, the values must be quantized to a finite number of bits. When this is the case, we can consider transforming the *pdf* into a PMF using,

$$f_\Delta[n] = \int_{n-0.5}^{n+0.5} f_\Delta(x)\, dx. \tag{2}$$

1.4. Effects Of Additive Noise

We are interested in the effect that additive noise has on the statistics of a signal. More specifically we are interested in modeling these changes and exploiting them to detect steganographic content. The histogram, $h[n]$, of an image is the frequency count of the pixel intensities present in an image. The histogram can be viewed as the PMF multiplied by the number of pixels in the image. This allows us to state the primary theorem in additive noise modelable information hiding.

THEOREM 1.1 (HISTOGRAM CONVOLUTION). *In a hiding system where the additive noise is independent of the coverimage, the histogram of the stegoimage is equal to the convolution of the stegonoise PMF and the coverimage histogram,*

$$h_s[n] = h_c[n] * f_\Delta[n]. \tag{3}$$

Proof. Consider the histogram as a probability mass function multiplied by a constant. From stochastic theory [4, Chap. 3], we know the addition of two independent random variables results in a convolution of their probability mass functions. □

From Theorem 1.1 we see that the effect of the additive noise on the image histogram is equivalent to a convolution of the stegonoise PMF and the histogram. Thus, given knowledge of any hiding scheme in the form of $f_\Delta[n]$ as well as knowledge of $h_c[n]$, the histogram of the stegomessage is known.

In the analysis of embedding it will be more convenient to work in the frequency domain. We use the discrete Fourier transform (DFT) defined as,

$$X[k] = DFT(x[n]) = \sum_{n=0}^{N-1} x[n]\, e^{-\frac{2\pi jnk}{N}}. \tag{4}$$

Figure 2. Pout.tif

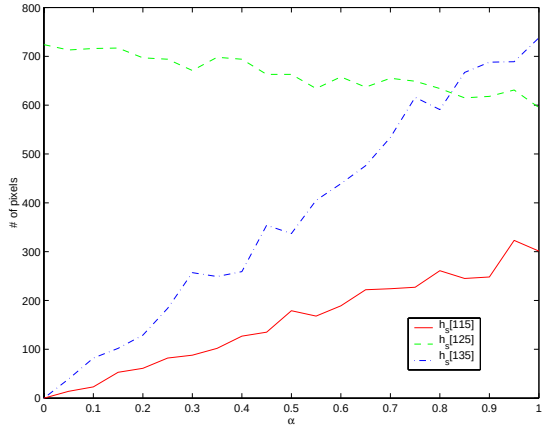

Figure 3. Various values of $h_\alpha[n]$ as embedding rate α changes.

Where N equals the largest intensity possible in the image. For example, in an 8 bit grayscale image N would be 2^8 or 256.

By taking the DFT of the PMFs involved, we have the characteristic functions defined as,

$$F_\Delta[k] \triangleq DFT\left(f_\Delta[n]\right), \tag{5a}$$

$$H_c[k] \triangleq DFT\left(h_c[n]\right), \tag{5b}$$

$$H_s[k] \triangleq DFT\left(h_s[n]\right). \tag{5c}$$

In particular the DFT of a histogram will be referred to as the histogram characteristic function, or $\mathcal{HCF}$. Using these definitions, we rewrite (3) in the frequency domain as,

$$H_s[k] = F_\Delta[k]H_c[k]. \tag{6}$$

Equation (6) gives us an insight into how embedding a message alters the $\mathcal{HCF}$ of an image. This will be particularly useful in the steganalysis explored in Section 3.

Thus far it has been assumed that the additive noise has operated on each pixel in the image. In practice the embedding rate may be reduced for a number of reasons, the most common is to increase the stealth of a hiding method. The following assumes that when only a fraction of the pixels are used for embedding, they are randomly chosen from the entire image. This prevents spatial/temporal-statistical attacks such as those discussed in [5].

THEOREM 1.2 (α-BITRATE EMBEDDING). *In a system where α is the fraction of pixels chosen at random for embedding and the stegonoise is independent of the coverimage. The stegoimage histogram is given by,*

$$h_\alpha[n] = \alpha\left(h_c[n] * f_\Delta[n]\right) + (1-\alpha)\, h_c[n]. \tag{7}$$

An illustration of this linearity is shown in Figure 3. In this figure we observe the contents of three histogram bins, (115, 125, and 135), as the embedded pixel rate, α, is varied from 0.0 to 1.0. The embedding method used is spread spectrum image steganography, described in Section 3.2. Here we see that the alterations of the histogram are roughly linear.

Equation (7) is easily extended to the frequency domain as,

$$H_\alpha[k] = \alpha H_c[k]F_\Delta[k] + (1-\alpha)\, H_c[k]. \tag{8}$$

To represent the addition of stegonoise at a bitrate of α as a single convolution we use the following theorem.

THEOREM 1.3 (UNIFIED α-BITRATE EMBEDDING). *In a system where α is the fraction of pixels chosen at random for embedding and the stegonoise is independent of the coverimage, the stegoimage histogram is given by,*

$$h_\alpha[n] = f_\Delta^\alpha[n] * h_c[n], \tag{9}$$

where,

$$f_\Delta^\alpha[n] \triangleq \alpha f_\Delta[n] + (1-\alpha)\delta[n].$$

2. THE HISTOGRAM CHARACTERISTIC FUNCTION

This section deals with the histogram characteristic function ($\mathcal{HCF}$). The $\mathcal{HCF}$ is a representation of the image histogram in the frequency domain. Much of the natural correlation as well as that introduced by the capturing device is apparent in the frequency domain. The histogram characteristic function center of mass (COM) is introduced as a measure of the energy distribution in an $\mathcal{HCF}$.

2.1. HCF Center of Mass

The $\mathcal{HCF}$ COM a simple metric which will be used in the steganalysis of images. We would like to use a metric which will show evidence of processing by $f_\Delta[n]$ or equivalently $F_\Delta[k]$. From this we choose to look at the center of mass of the $\mathcal{HCF}$,

$$\mathcal{C}\left(H[k]\right) \triangleq \frac{\sum\limits_{k\in\mathcal{K}} k\left|H[k]\right|}{\sum\limits_{i\in\mathcal{K}} \left|H[i]\right|}. \tag{10}$$

Where $\mathcal{K} = \{0, \ldots, \frac{N}{2}-1\}$ and N is the DFT length. The COM gives a general information about the energy distribution in the histogram characteristic function. The following provides a useful result for a class of additive noise modelable steganographic schemes.

THEOREM 2.1. *For an embedding scheme with a nonincreasing $|F_\Delta[k]|$ for $k = (0, \ldots, \frac{N}{2}-1)$, the $\mathcal{HCF}$ COM decreases or remains the same after embedding,*

$$\mathcal{C}\left(H_s[k]\right) \leq \mathcal{C}\left(H_c[k]\right), \tag{11}$$

with equality if and only if $|F_\Delta[k]| = 1,\ \forall\ k = 0, \ldots, \frac{N}{2}-1$.

Proof. By the discrete Čebyšev inequality [6, Chap. 4], for a nondecreasing sequence, $a = (a_0, \ldots, a_n)$, a nonincreasing sequence, $b = (b_0, \ldots, b_n)$, and a non-negative sequence, $p = (p_0, \ldots, p_n)$,

$$\sum_{k=0}^{n} p_k \sum_{k=0}^{n} p_k a_k b_k \leq \sum_{k=0}^{n} p_k a_k \sum_{k=0}^{n} p_k b_k. \tag{12}$$

Letting $a_k = k$, $b_k = |F_\Delta[k]|$, $p_k = |H_c[k]|$ and $\mathcal{K} = \{0, \ldots, \frac{N}{2}-1\}$ we have,

$$\sum_{k\in\mathcal{K}} \left|H_c[k]\right| \sum_{k\in\mathcal{K}} k\left|F_\Delta[k]\right|\left|H_c[k]\right| \leq \sum_{k\in\mathcal{K}} k\left|H_c[k]\right| \sum_{k\in\mathcal{K}} \left|F_\Delta[k]\right|\left|H_c[k]\right|, \tag{13}$$

or,

$$\frac{\sum\limits_{k\in\mathcal{K}} k\left|F_\Delta[k]\right|\left|H_c[k]\right|}{\sum\limits_{k\in\mathcal{K}} \left|F_\Delta[k]\right|\left|H_c[k]\right|} \leq \frac{\sum\limits_{k\in\mathcal{K}} k\left|H_c[k]\right|}{\sum\limits_{k\in\mathcal{K}} \left|H_c[k]\right|}. \tag{14}$$

Note that (13) holds with equality if and only if $|F_\Delta[k]| = 1,\ \forall\ k \in \mathcal{K}$. In the spatial domain, the equality condition is satisfied if $f_\Delta[n] = \delta[n]$. □

There exists a number of distributions having monotonically decreasing characteristic function magnitudes, these include the Gaussian and Laplacian.

2.2. HCF of Color Images

The above arguments can easily be extended for use with RGB color images as follows. We consider a pixel, $\mathbf{x}(n_1, n_2)$, as a vector of RGB intensities,

$$\mathbf{x}(n_1, n_2) = [x_r(n_1, n_2) \;\; x_g(n_1, n_2) \;\; x_b(n_1, n_2)].$$

We define an RGB histogram, $h[\mathbf{n}]$, where $\mathbf{n}$ is a vector of the RGB intensities, and the value of the histogram evaluated at $\mathbf{n}$ is the number of pixels with that RGB triplet. Taking the 3 dimensional discrete Fourier transform of $h[\mathbf{n}]$ we define the histogram characteristic function, $\mathcal{HCF}$ for an RGB image as

$$H[\mathbf{k}] \triangleq DFT_3 h[\mathbf{n}] \tag{15}$$

Since the length N DFT is of real data its magnitude is symmetric about $\frac{N}{2}$ such that we only need to observe $[0, \frac{N}{2} - 1]^3$ of the $[0, N-1]^3$ DFT coefficients.

We now consider the centers of mass for $H[\mathbf{k}]$ along each of it's three axes,

$$\mathcal{C}_{k_1}(H[\mathbf{k}]) \triangleq \sum_{\mathbf{k} \in \mathcal{K}} k_1 |H[\mathbf{k}]|, \tag{16a}$$

$$\mathcal{C}_{k_2}(H[\mathbf{k}]) \triangleq \sum_{\mathbf{k} \in \mathcal{K}} k_2 |H[\mathbf{k}]|, \tag{16b}$$

$$\mathcal{C}_{k_3}(H[\mathbf{k}]) \triangleq \sum_{\mathbf{k} \in \mathcal{K}} k_3 |H[\mathbf{k}]|. \tag{16c}$$

Where $\mathcal{K}$ is the set of first octant indices, i.e. $\mathbf{k} \in [0, \frac{N}{2} - 1]^3$. Combining the values of each of (16) we can define a point in 3 dimensional space to be a "center of mass" for the RGB $\mathcal{HCF}$.

3. MODELING SYSTEMS

In this section a number of information hiding methodologies are analyzed. The goal in each analysis is to derive the probability mass function of the stegonoise. Once we have this expression we use Theorem 1.1 to estimate the stegoimage histogram.

3.1. LSB

Least significant bit (LSB) steganography is the most simplistic form of steganography. It hides information by replacing the least significant bit of a pixels intensity with a message bit[7]. This system can be approximated as an additive noise scheme. First we consider the message bits (mb) to be i.i.d. with $p(mb = 0) = p(mb = 1) = \frac{1}{2}$. Likewise we assume that the LSBs of the coverimage (x_c^{LSB}) are i.i.d. with $p\left(x_c^{LSB} = 0\right) = p\left(x_c^{LSB} = 1\right) = \frac{1}{2}$. It is then easily shown,

$$f_\Delta[-1] = p(mb = 0)\, p\left(x_c^{LSB} = 1\right) = 0.25, \tag{17a}$$

$$\begin{aligned} f_\Delta[0] &= p(mb = 0)\, p\left(x_c^{LSB} = 0\right) \\ &\quad + p(mb = 1)\, p\left(x_c^{LSB} = 1\right) = 0.5, \end{aligned} \tag{17b}$$

$$f_\Delta[1] = p(mb = 1)\, p\left(x_c^{LSB} = 0\right) = 0.25. \tag{17c}$$

The LSB $|F_\Delta[k]|$ and $f_\Delta[n]$ for a DFT length $N = 256$ are shown in Figure 4.

Notice that this scheme acts as a lowpass filter on the histogram of the image. This filtering causes the histogram bins to "bleed" together, resulting in more unique intensities, as well as more close intensity pairs. These results are exploited in [8] to detect the presence of LSB steganography. In addition to being lowpass, $|F_\Delta[k]|$ is monotonically decreasing, which allows us to use Theorem (2.1).

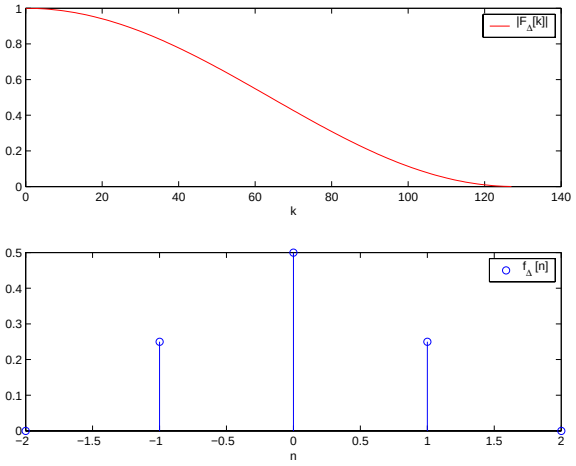

Figure 4. $F_\Delta[k]$ and $f_\Delta[n]$ for a LSB scheme

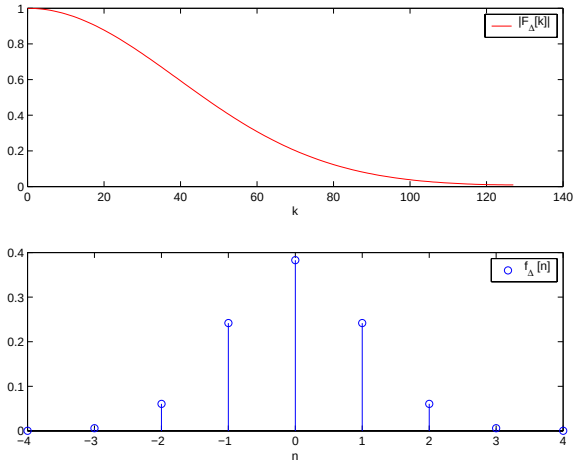

Figure 5. $|F_\Delta[k]|$ and $f_\Delta[n]$ for WGN

In this analysis, $f_\Delta[n]$ approximates the alterations caused by LSB embedding as an additive noise. The actual embedding is not independent of the coverimage, for example,

$$f_\Delta[n=-1] \neq f_\Delta[n=-1|x_c^{LSB}=0] = p(x_s - x_c = -1|x_c^{LSB}=0) = 0,$$

because when $x_c^{LSB} = 0$, only the addition of 0 or 1 can result.

3.2. Spread Spectrum Image Steganography

In this discussion we analyze spread spectrum image steganography (SSIS)[9]. The SSIS scheme hides data in a Gaussian stegonoise that is added to the coverimage. This additive noise signal is equivalent to a direct-sequence spread spectrum system [10] wherein the PN-code is distributed as $\mathcal{N}\left(\mu, \sigma^2\right)$ with a chip period of every pixel. The use of Gaussian noise in this scheme is motivated by the assumption that AWGN is a common distortion in images.

The distribution function of the pseudo-noise is defined as,

$$f_\Delta\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\cdot(x-\mu)^2} \tag{18}$$

For this discussion we will assume $\mu = 0$ and $\sigma^2 = 1$. To determine the effect this additive noise will have on the histogram of the coverimage we use (2) to find $f_\Delta[n]$. This yields the coefficients plotted in Figure 5 along with their corresponding frequency response for a DFT length $N = 256$.

Notice that the effect of the independent additive noise is a monotonically decreasing lowpass filter on the histogram. This is illustrated in the histogram in Figure 6 as well as the $\mathcal{HCF}$ magnitude in Figure 7.

To reduce error rate the stegonoise may be multiplied by a scale-factor, β, to adjust the power. From stochastic theory the variance of a scaled random variable behaves as,

$$\begin{aligned} \sigma^2_{scale} &= E[\beta\left(X-\mu\right)\beta\left(X-\mu\right)] \\ &= \beta^2 E[\left(X-\mu\right)^2] \\ &= \beta^2\sigma^2 \end{aligned} \tag{19}$$

As the variance of the additive noise increases by β^2, the stegonoise PMF will spread out. This spreading of $f_\Delta[n]$ yields a lower cutoff point in $|F_\Delta[k]|$. This effect is plotted in Figure 8 for $\beta = \{1, 2, 3, 4, 5\}$ and $\sigma^2 = 1$. The alteration of $h_c[n]$ becomes increasingly pronounced as β increases.

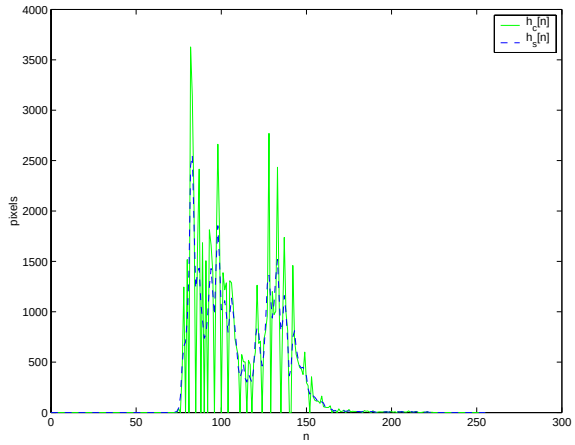

Figure 6. $h_c[n]$ and $h_s[n]$ for pout.tif

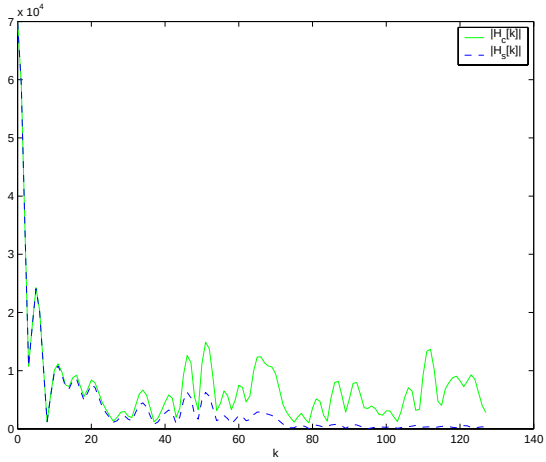

Figure 7. $|H_c[k]|$ and $|H_s[k]|$ for pout.tif

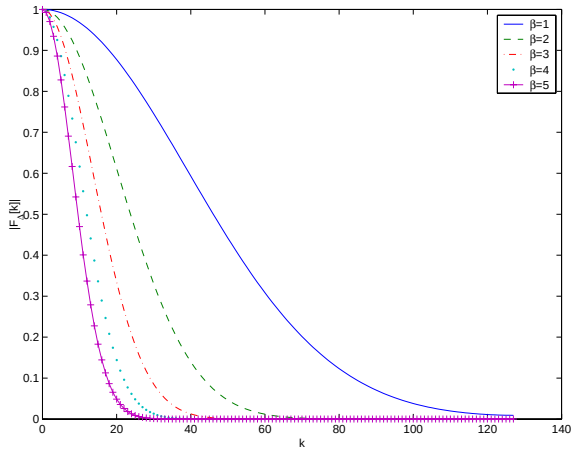

Figure 8. Effect of scaling factor β on $|F_\Delta[k]|$

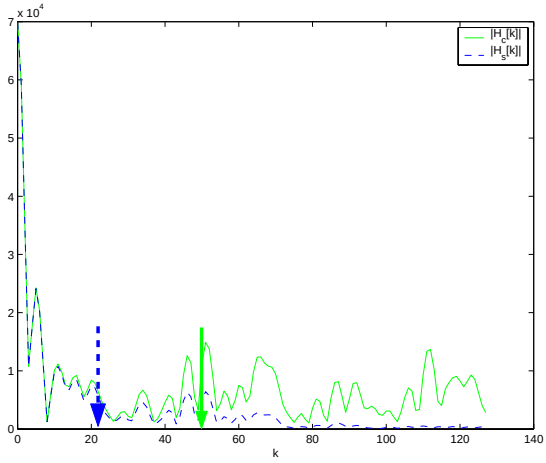

Figure 9. Center of mass for $\mathcal{HCF}$ magnitude

3.3. Discrete Cosine Transform Steganography

To improve robustness and stealth, many steganographic schemes utilize projections to embed data in an alternate space. In this section we consider the effects of hiding information as an additive noise in discrete cosine transform (DCT) coefficients. We choose the DCT as it is a common transform in image processing. The process we discuss is generally similar to the DCT hiding of [11], with the exception our model will hide data as an additive noise rather than a quantization.

The actual embedding process begins by decorrelating the image by reordering the pixels based on a keying variable. Next, the mean of the pixels is subtracted and an $L \times L$ block DCT [12] is taken over the image. The decorrelation of the pixels serves to whiten the image and increase the energy in the high frequency DCT coefficients, making them more useful in hiding data. Once in the frequency domain, an i.i.d. stegonoise is added to each coefficient *. An $L \times L$ block IDCT is performed the previously

*In [11] the DCT coefficients are quantized to hide information. The error introduced in this process is a deterministic function of the coefficients. As this error would be considered the stegonoise in our framework, the heavy dependence between the cover-coefficients and stegonoise does not allow for a direct additive noise analysis.

subtracted mean is added to each pixel. Finally, the pixels are rounded to integers and returned to their original order using the keying variable.

Considering the signals involved we have,

$$\mathcal{X}_c = DCT\{x_c\},$$

$$\mathcal{X}_s = \mathcal{X}_c + stegonoise,$$

$$x_s = IDCT\{\mathcal{X}_c + stegonoise\} = x_c + IDCT\{stegonoise\}.$$

The additive noise embedding in the frequency domain is modeled as the addition of spatial stegonoise, $IDCT\{stegonoise\}$.

We now present an informal argument that the spatial stegonoise is i.i.d Gaussian using properties of the DFT. In [13] it is shown that for a stationary sequence with finite second-order moments and mixing, the DFT elements are asymptotically independent. In [14, Chap. 2] it is shown that for sequences obeying the Lindeberg condition, the DFT elements asymptotically approach normal distributions. With this we can consider the spatial stegonoise to be roughly equivalent to i.i.d. Gaussian. This allows us to consider the addition of an i.i.d. stegonoise in the frequency domain, as approximately i.i.d. Gaussian stegonoise in the spatial domain. With these assumptions the effect of additive noise in the frequency domain is modeled as in Section 3.2, in particular the monotonically decreasing $|F_\Delta[k]|$.

4. DETECTION SCHEMES

This section uses the framework previously developed to build two classifiers that are able to differentiate altered images from original. The method presented in Section 4.1 builds a classifier which is trained on both the coverimages as well as stegoimages. A second classifier is presented in Section 4.2 which uses no explicit information about the hiding method.

4.1. Known Scheme Detection

In known scheme detection the method of hiding is assumed to be available in classifier construction. This provides a significant advantage in detection as a concrete notion of the effects of embedding can be developed.

Using results from Section 3.2 we create a simple classifier scheme to detect the presence of SSIS in a color image. The classification of a test image will be into one of two categories: containing SSIS data or unaltered.

Recalling that the addition of noise affects the $\mathcal{HCF}$ magnitude as lowpass filter shown in Figure 7 as well as the bound given in Theorem 2.1, we expect $\mathcal{C}(H_s[k])$ to be lower in the stegoimage. Indeed this phenomenon is shown in Figure 9. In the case of the 3 dimensional histogram of an RGB image, we would expect that the center of mass would move toward the origin.

To verify this, 24 images from the Kodak PhotoCD PCD0992 [15] are used. These images are 24-bit, 768x512 pixel, lossless truecolor images stored in the PNG format. For each image the $\mathcal{HCF}$ COM is computed for the original image as well as the SSIS stegoimage with $\mathcal{N}(0,1)$ and $\alpha = 1$ (full embedding). A 3-D scatter plot of these points is shown in Figure 10. As expected the centers of mass for the stegoimages are considerably lower than those of the originals.

To create the classifier, we first assume the distribution of COMs is Gaussian to make use of the Bayesian multivariate classifier [16]. The Bayesian multivariate classifier requires that the mean, μ, and covariance, $\mathbf{\Sigma}$, matrices of the source distributions be known or estimated. For our application we estimate these values using the estimators,

$$\mu_i = \frac{1}{S}\sum_{k=0}^{S-1} \mathbf{x}_i^{(k)} \tag{20}$$

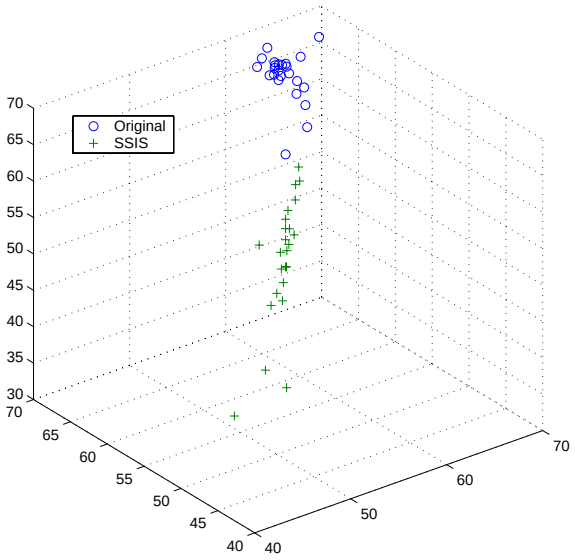

Figure 10. Center of Mass for Test Images

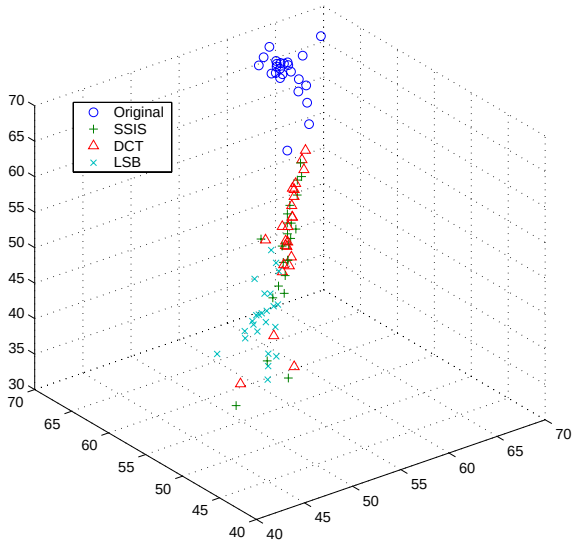

Figure 11. Centers of mass

$$\mathbf{\Sigma}_i = \frac{1}{S}\left(\mathbf{x}_i - \mu_i\right)^T \left(\mathbf{x}_i - \mu_i\right) \tag{21}$$

where $\mathbf{x}_i$ is the training set for the ith multivariate and S is the number of samples.

The general multivariate discriminant functions are then,

$$g_i\left(\mathbf{k}\right) = \mathbf{k}^t \mathbf{W}_i \mathbf{k} + \mathbf{w}_i^t \mathbf{k} + w_i, \tag{22}$$

with

$$\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}, \tag{23a}$$

$$\mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\mu_i, \tag{23b}$$

$$w_i = -\frac{1}{2}\mu_i^t \mathbf{\Sigma}_i^{-1}\mu_i - \frac{1}{2}\ln\left|\mathbf{\Sigma}_i\right|. \tag{23c}$$

To classify an unknown sample vector $\mathbf{x}$, each discriminant function is evaluated at $\mathbf{x}$. If $g_1(\mathbf{x}) > g_2(\mathbf{x})$ the pattern is assigned to ω_1, else it is assigned as ω_2.

For each trial the 24 test images were randomly placed into one of four groups.

1. 10 Unaltered image COMs used to find μ_1 and $\mathbf{\Sigma}_1$ for ω_1.
2. 10 SSIS image COMs embedding used to find μ_2 and $\mathbf{\Sigma}_2$ for ω_2.
3. 2 Unaltered image COMs classified.
4. 2 SSIS image COMs classified.

Where μ_1 and $\mathbf{\Sigma}_1$ are the estimated mean and covariance matrices of the original $\mathcal{HCF}$ COM class, ω_1. Likewise, μ_2 and $\mathbf{\Sigma}_2$ are the estimated mean and covariance matrices of the SSIS stegoimage $\mathcal{HCF}$ COM class, ω_2. Using these distributions, the remaining 4 images are classified by evaluating the discriminant functions of each class at the test COMs.

As shown in Table 1, with 10000 trials (40000 tests) the classifier was 94.68% correct. This equates to 2129 errors in classification. Of these, 1956 were Type I (false alarms), while only 173 of the 2129 errors were Type II (missed signals).

Table 1. Known Scheme Classification Performance (10000 Trials)

Tests	40000	
Errors	2129	
Correct	94.68%	
	Original	Stegoimage
Tests	20000	20000
Errors	1956	173
Correct	90.22%	99.13%

4.2. Unknown Scheme Detection

In practice it is desirable to detect the presence of a message regardless of the embedding method. The foremost reason for this is that the algorithm used in embedding may not be known. With this in mind we now describe an unknown scheme detection.

In contrast to the previous section where we made use of statistics from both original and modified images, we now only consider the availability of original images. It is worth emphasizing that *we assume no explicit knowledge of the hiding method* in the classifier construction. We only have what we consider to be "normal" images available to train on, and knowledge of Theorem (2.1).

Again we focus on the $\mathcal{HCF}$ COM as our feature in the detection scheme. As we would like to measure how similar (or dissimilar) a COM in question is to our trained statistic, we consider the Mahalanobis distance defined as,

$$d^2 = (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu). \tag{24}$$

Where μ and $\mathbf{\Sigma}$ are the mean and covariance estimates defined in (20) and (21), using measurements gathered from a training set.

The Mahalanobis distance essentially gives a statistical measure of how far a given point is from the estimated mean, with consideration toward the variance of each variable. Generally speaking, the greater the Mahalanobis distance, the less likely the test point is of the same distribution as the training set. The surface defined by $d^2 = 1$ is a surface where each point is one standard deviation away from the mean.

To test this classification scheme, the set of 24 images is randomly divided into 5 groups:

1. 20 original image $\mathcal{HCF}$ COMs used to estimate μ and $\mathbf{\Sigma}$
2. 1 Unaltered image COM classified
3. 1 SSIS image COM classified
4. 1 DCT image COM classified
5. 1 LSB image COM classified

The 20 unaltered COMs are used to form an estimate of the mean vector and covariance matrix. The multivariate described by these is considered to be a natural $\mathcal{HCF}$ COM distribution, and any images which differ significantly will be classified as containing steganographic data.

The first test image is the unaltered image in its original form without any modifications. The SSIS image has a message embedded using the method described in Section 3.2, that is equivalent to adding i.i.d. $\mathcal{N}(0, 1)$. The DCT image is created using the method in Section 3.3. A DCT block size of 2×2

Table 2. Unknown Scheme Classification Performance (20000 Trials)

Tests	80000			
Errors	3285			
Correct	95.89%			
	Original	SSIS	DCT	LSB
Tests	20000	20000	20000	20000
Errors	1024	626	1635	0
Correct	94.88%	96.87%	91.83%	100%

is used to project the image into the frequency domain, where an i.i.d. uniform noise over $[-2, 2]$ is added to each coefficient. The LSB image is formed as described in Section 3.1, by replacing the least significant bit of each pixel with the message bit. For each method of embedding, 1 bit was hidden in each pixel (or coefficient), i.e. $\alpha = 1$. Figure 11 shows a plot of the $\mathcal{HCF}$ COMs for all 24 images with each embedding.

A Mahalanobis cutoff of approximately 40 was chosen to yield a Type I, (false alarm), rate of approximately 5%. As can be seen the classifier performs very well, with a correct classification rate of approximately 95%.

5. CONCLUSION

A framework for modeling additive noise information hiding has been developed. This framework allows for an analysis of the effects of data hiding on the histogram of a signal. The histogram characteristic function center of mass is introduced as a simple metric that is predictably affected by a class of additive noise.

Three data hiding methodologies are analyzed in anticipation of constructing a detection scheme. Two detection schemes are built and tested, the first allows the classification of known embedding methods, while the second assumes no explicit knowledge of the additive noise. Both detection schemes show that the addition of a zero-mean, unit variance Gaussian noise can be readily detected in the test images. In addition, the unknown scheme detection performs very well on least significant bit, and additive noise discrete cosine transform hiding using a unified approach.

While the framework introduced in this paper has been used to explore steganography in images, the additive noise model is applicable to many media types. The continued development and application of additive noise modelable information hiding stands to offer many insights into the field of information hiding.

REFERENCES

1. R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. on Information Theory* **44**, pp. 2325 – 2383, Oct. 1998.
2. G. E. Healey and R. Kondepudy, "Radiometric ccd camera calibration and noise estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **16**, pp. 267–276, Mar. 1994.
3. C. E. Shannon, "Communication in the presence of noise," *Proceedings of the I.R.E.* **37**, pp. 10–21, Jan. 1949.
4. J. Woods and H. Stark, *Probability and Random Processes With Applications to Signal Processing*, Prentice-Hall, Upper Saddle River, NJ, 3 ed., 2001.
5. A. Westfeld and A. Phitzmann, "Attacks on steganographic systems," in *Proceedings* 3^{rd} *Information Hiding Workshop*, pp. 61–75, (Dresden, Germany), Sept. 28-Oct. 1 1999.

6. D. S. Mitrinović, J. E. Pečarić, and A. M. Fink, *Classical and New Inequalities in Analysis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
7. C. Kurak and J. McHugh, "A cautionary note on image downgrading," in *Computer Security Applications Conference*, (San Antonio, TX), Dec. 1992.
8. J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color, and gray-scale images," *IEEE Trans. Multimedia* **8**, pp. 22–28, Oct. 2001.
9. L. M. Marvel, C. G. Boncelet, Jr, and C. T. Retter, "Spread spectrum image steganography," *IEEE Trans. Image Processing* **8**, pp. 1075–1083, Aug. 1999.
10. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, "Theory of spread spectrum communications — a tutorial," *IEEE Trans. Comm.* **COM-30**, pp. 855–884, May 1982.
11. F. Alturki and R. Mersereau, "A novel approach for increasing security and data embedding capacity in images for data hiding applications," in *Information Technology: Coding and Computing*, pp. 228–233, (Las Vegas, NV), Apr. 2–4, 1997.
12. J. S. Lim, *Two-Dimensional Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1990.
13. D. R. Brillinger, "Fourier analysis of stationary processes," *Proceedings of the IEEE* **62**, pp. 1628–1643, Dec. 1974.
14. W. A. Pearlman, "Quantization error bounds for computer-generated holograms," Tech. Rep. 6503-1, Information Systems Laboratory, Stanford University, Stanford,CA, Aug. 1974.
15. R. Franzen, "Kodak lossless true color image suite: PhotoCD PCD0992," Mar. 27, 2002. Available: http://sqez.home.att.net/thumbs/Thumbnails.html.
16. R. O. Duda, P. E. Hart, and H. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, NY, 2 ed., 2000.