A Mathematical Approach to Steganalysis

R. Chandramouli

Multimedia Systems, Networking and Communications (MSyNC) Lab
Department of Electrical and Computer Engineering
Stevens Institute of Technology

ABSTRACT

A mathematical approach to steganalysis is presented in this paper with linear steganography being the main focus. A mathematically formal definition of steganalysis is given followed by definitions for *passive* and *active* steganalysis. The steganalysis problem is formulated as blind system identification and conditions for identifiability (successful steganalysis) are derived. A procedure to systematically exploit any available spatial and temporal diversity information for efficient steganalysis is also discussed.

Experimental results are given for steganalysis of Gaussian distributed, spread spectrum image steganography and watermarking. The proposed technique is observed to produce impressive results for a variety of performance measures. Based on the results we conclude that a common belief, namely, spread spectrum steganography/watermarking is secure because of the low strength, noise-like message carrier is not valid anymore within the current context. Therefore, new questions regarding steganography security that differ from the standard information theoretic notion are raised and some answers are provided.

Keywords: Steganalysis, steganography, spread spectrum, cumulants.

1. INTRODUCTION

Steganalysis is a relatively new branch of research. While steganography (which is somewhat different from watermarking) deals with techniques for hiding information, the goal of steganalysis is to detect and/or estimate potentially hidden information from observed data with little or no knowledge about the steganography algorithm and/or its parameters. It is fair to say that steganalysis is both an art and a science. The art of steganalysis plays a major role in the selection of features or characteristics a typical stego message might exhibit while the science helps in reliably testing the selected features for the presence of hidden information. While it is possible to design a reasonably good steganalysis technique for a specific steganography algorithm, the long term goal must be to develop a steganalysis framework that can work effectively at least for a class of steganography methods if not for all. Clearly, this poses a number of mathematical challenges and questions some of which are summarized below.

- Can the current and future steganography algorithms be categorized into distinct classes of mathematical techniques?
- What is a good mathematical definition of steganalysis?
- What a priori knowledge can we assume the steganalyst possesses?
- What mathematical properties a class of steganography algorithms must satisfy for which good steganalysis techniques can be developed? This will give rise to a new notion of security in steganography that could be quite different from the popular information theoretic definition.[1]
- What are the candidate *cost* or *risk* functions that a steganalyst must optimize during hidden data detection or extraction procedure?

To appear in Proc. SPIE Security and Watermarking of Multimedia Contents IV, California, Jan. 2002. For author information:
E-mail: rchandr1@stevens-tech.edu, URL: http://www.ece.stevens-tech.edu/∼msync

- What are the performance trade-offs involved if a steganalysis algorithm is designed only to detect, only to extract, or detect and extract the hidden message?

We attempt to address some of these questions in this paper and develop a formal theory of steganalysis. We note that in our present analysis we assume the steganalyst has reasonable computational resources and time.

In a traditional steganography set-up formulated as a prisoner's problem,[2] Alice wishes to send a secret message to Bob by hiding information in a cover message. The stego message (cover+message) passes through Wendy (a warden) who inspects it to determine if there is anything suspicious about it. Wendy could perform one or several tests to decide if the message from Alice to Bob contains any secret information. If her decision is negative then Wendy forwards the message to Bob—Wendy acts as a *passive warden*. On the other hand, Wendy can take a conservative approach and modify all the messages from Alice to Bob irrespective of whether any information is hidden by Alice or not. In this case, Wendy is called an *active warden*. Of course, Wendy will have constraints such as the maximum allowable distortion when modifying the message etc. For example, if the cover messages are digital images, then Wendy cannot modify the stego message to an extent that perceptually significant distortions are induced.

While current steganalysis techniques focus on detecting the presence/absence of a secret message in observed message, to our knowledge there seems to have been no attempt in extracting the secret message. In general, extraction of the secret message could be a harder problem than mere detection. Therefore, based on the ultimate outcome of the effort we classify steganalysis into two categories:

- **Passive steganalysis**: Detect the presence or absence of a secret message in an observed message.
- **Active steganalysis**: Extract a (possibly approximate) version of the secret message from a stego message.

Note that active steganalysis could be different from an active warden case. An active warden manipulates the stego message in the hopes of destroying the secret message (if any) but an active steganalyst attempts to estimate and extract the secret message without destroying it. In this paper, we discuss a mathematical framework for active steganalysis when a certain class of linear steganography algorithms are employed. We also discuss the strengths and limitations of the proposed framework and provide numerical examples. Without loss of generality we consider digital images as cover messages for our experiments. Our primary goal is to estimate the cover message, secret message, and even perhaps the steganography key using only the observed stego messages. During this process we exploit *spatial diversity* and *temporal diversity* information that will be explained in later sections.

The paper is organized as follows. A mathematical formulation of steganalysis is presented in Section 2 and the corresponding steganalysis algorithm is given in Section 3. Experimental results are discussed in Section 4 and concluding remarks can be found in Section 5.

2. STEGANALYSIS: PROBLEM SET-UP

We consider the scenario where the steganography key is the same for at least two stego messages. This is not a restrictive assumption because Alice and Bob could exchange a steganography key initially and later use this key to embed and extract multiple secret information. For example, the steganography key could be the pixels of an image where the secret message is hidden such as in LSB image steganography. This also simplifies the key management problem for Alice and Bob. In fact, a fixed key is prevalent in many watermarking algorithms and has found to be a cause for concern.[3] Our next assumption is that the secret message is statistically independent of the cover message and is embedded into the cover message in an additive fashion. For example, a message of length L bits could modulate the sign of a zero-mean, finite variance white Gaussian random vector of length L that is statistically independent of the cover image. If the sign of an element in the message carrying Gaussian random vector is positive then it stands for message bit 1 if not it represents the bit 0. The Gaussian random vector is then scaled by a positive value and then added to certain discrete cosine transform (DCT) coefficients of the cover image chosen based on a steganography key. Note that this method is similar to

the popular spread spectrum watermarking algorithm.[4] Of course it can be argued that watermarks carry one bit of information when their presence/absence is detected. The third assumption is that the steganalyst has access to at least to two stego messages with same cover message, same secret information and same key but differing in some other parameters. There are both practical and purely academic examples supporting these assumptions such as the following:

- Spread spectrum image steganography has been previously proposed by Smith et. al.[5] and Marvel et. al..[6] In such schemes use the message bits modulate a carrier function/vector (Gaussian random vector is a popular choice) and then the result is added to the cover message. Extraction of the message bits is the inverse of the embedding process after applying filtering and other types of image processing operations to the stego image. In these methods the steganography key and the message carrier are independent of the cover message. It is conceivable that the same key and message carrier signal are used for different images for practical simplicity, *e.g.*, fingerprinting.

- Some commercial products such as Digimarc's[7] image watermarking software allow a user to choose multiple watermark strengths for a fixed key and cover image.

- Some video watermarking techniques[8] spread the message sequence spatially before modulating a carrier and then adding it to a video frame. The carrier strength can be adjusted based on the characteristics of the local spatial location. Clearly, in these types of data embedding, information about the hidden message is spread across time (video frames). Note that in general a slow motion video leaks more information to a steganalyst from successive video frames compared to a high motion video sequence. This is because in a slow motion video successive frames have more or less similar statistical and perceptual characteristics.

- Alice uses an additive image steganography algorithm to send a message to Bob. Wendy, being an active warden compresses the image to a certain rate using JPEG[9] before forwarding the stego image to Bob. Since Alice is unaware of the compression rate, she initially chooses a random strength for the message carrier and hopes that it will survive Wendy's compression. If Bob is successfully able to retrieve the hidden message, he indicates this to Alice the next day at the prison's dining hall by drinking coffee instead of his usual tea! If not, Alice assumes that the message has been lost due to Wendy's compression attack and therefore re-sends it the next day after increasing the message strength hoping it survives Wendy's attack (this is a specific case of adaptive steganography[10]). Now, Wendy has access to two copies of the same stego image differing only in the strength factor of the message carrying signal.

Based on this discussion of information collection for steganalysis we classify steganalysis methods into two general categories:

- **Spatial diversity information based steganalysis**: Steganography methods could spread information in the spatial domain and this information repeats itself in various forms in different spatial locations (e.g., different blocks within an image or, in different images). We call this exploitation of repeated information for steganalysis spatial diversity based steganalysis.

- **Temporal diversity information based steganalysis**: Steganography information that appears repeatedly over time can also aid steganalysis. Such techniques are called *temporal diversity information based steganalysis*.

An excellent survey of steganalysis techniques that fall into the spatial diversity steganalysis framework and their implications is provided by Fridrich et. al.[11] One of the main effects of a good steganalysis technique is a reduction in the maximum number of bits that can be embedded (steganography capacity) without being detected. A mathematical formulation for computing the stego capacity in the presence of steganalysis for LSB image steganography was provided by Chandramouli et. al.[12] It is shown that a good steganalysis technique can significantly reduce the embedding capacity. Now that we have stated the assumptions and discussed the practical and theoretical validity of our assumptions, next we describe a mathematical formulation of the steganalysis problem.

2.1. Mathematical Formulation

In this section we first describe a generic linear additive steganography algorithm and then mathematically set-up the corresponding steganalysis problem. Note that an additive steganography model seems to fit a wide range of current steganography techniques such as the following. Suppose the data embedding method is based on employing two different quantizers to represent the message bits 0 and 1 then the quantization error can be modelled as additive noise interfering with the cover message. LSB embedding for image steganography changes the pixel values by ± 1. Finally, many steganography methods first use the message bits to modulate a carrier signal which is then added to the cover message.

2.1.1. Message Embedding and Extraction

Let $\{s(k)\} \in \Re$ denote a cover message, $\{w(k)\} \in \Re$ be the message carrier independent of the cover message and let the stego message be obtained as,

$$y(k) = s(k) + \alpha w(k), k = 1, 2, \ldots, N \tag{1}$$

$\{s(k)\}$ is continuous valued and $\alpha > 0$ denotes the message strength that could be adjusted based on perceptual characteristics, robustness properties etc. Some of the $w(k)$'s (also continuous valued) will be equal to zero based on the steganography key if that particular $s(k)$ does not carry a message bit. We assume $\{s(k)\}$ and $\{w(k)\}$ are samples from a stationary random vector. The steganography key and α are known to the decoder. Suppose the decoder has access to the cover message $\{s(k)\}$ then it is quite straightforward to extract the secret message by subtracting $s(k)$ from $y(k)$. On the other hand, if the decoder does have not access to $s(k)$ then filtering techniques can be employed to obtain an estimate of $s(k)$ and hence an approximate version $\hat{w}(k)$ which can then lead to bit errors.[6] A number of possibilities exists to minimize the bit error rate such as error control coding, better estimation techniques etc. We do not discuss these methods in detail as they are beyond the scope of this paper. In watermarking like techniques where it is not necessary to extract the individual message bits rather only the detection of the presence/absence of a message is of interest a correlation type detection technique can be applied to $y(k)$. We note that both in steganography and watermarking applications the genuine decoder may possess only a noisy copy of $\{y(k)\}$ say $\{\hat{z}(k)\}$ due to attacks but we assume the steganalyst has access to $\{y(k)\}$.

Some observations and beliefs about a popular choice for $\{w(k)\}$ in Eq. (1) are helpful at this juncture. It is quite common to choose $\{w(k)\}$ as a zero mean, white Gaussian process with finite variance. This gives rise to the so called spread spectrum steganography[5,6] and spread spectrum watermarking.[4] It is widely believed that this random noise-like message carrier with spread spectrum is secure against steganalysis attacks that aim at estimating it. This choice is also observed to be robust for watermarking applications.[4] Another reason for this popular choice for a message carrier comes from information theory. It is known from information theory that a Gaussian signal is the best choice for a Gaussian channel. Since most image steganography methods conveniently assume the image pixel distribution and common transform coefficient distribution to be Gaussian, the choice of $\{w(k)\}$ as Gaussian is justified. However, in reality many image related features are non-Gaussian. It is well known that the discrete cosine transform (DCT) coefficients of an image that are used widely as a message carrier have a generalized Gaussian distribution.[13] We'll later show that this fact can immensely aid in steganalysis. To summarize, we show through analysis and experiments that the *sense of security using spread spectrum type steganography is a myth if*:

- steganalysis systematically exploits the spatial and/or temporal information.
- non-Gaussian nature of a cover message can expose the presence of a Gaussian distributed secret message.

2.1.2. Steganalysis

Assume that the only knowledge available to the steganalysis is that the steganography model is of the form given in Eq. (1). Let the two copies of a stego message available to the steganalyst be $\{y_1(k)\}$ and $\{y_2(k)\}$. We

Figure 1: Steganalysis as a blind system identification problem.

can then write,

$$\mathbf{z}(k) = \begin{pmatrix} y_1(k) \\ y_2(k) \end{pmatrix} = \mathbf{A}\mathbf{r}(k) \tag{2}$$

$$= \begin{pmatrix} 1 & \alpha_1 \\ 1 & \alpha_2 \end{pmatrix} \begin{pmatrix} s(k) \\ w(k) \end{pmatrix} \tag{3}$$

$\mathbf{z}(k)$ is the random stego message vector observed by the steganalyst, $\mathbf{A}$ is the *strength matrix* and $\mathbf{r}(k)$ is the vector with the cover message and the secret message as its components. The steganalyst is now faced with the problem of inferring $\mathbf{A}$ from $\mathbf{z}(k)$. This can be viewed as a blind system identification problem as shown in Figure 1. If $\mathbf{A}^{-1}$ can be identified then we can obtain $\mathbf{r}(k)$ from $\mathbf{A}^{-1}\mathbf{z}(k)$, *i.e.*, the steganalysis problem is to find a linear transform such that the components of $\mathbf{r}(k)$ can be retrieved. We also notice the similarity between this version of steganalysis and a blind source separation (BSS) problem.[14] Therefore techniques from BSS can carry over here. While there are many ways of computing the linear transform to retrieve the cover message and the secret message from $\mathbf{z}(k)$ we choose the independent component analysis (ICA) method[14] as this seems to be applicable in a natural way. We use the fact that the message carrier is generated independently (statistically) of the cover message and attempt to estimate a linear transform that will maximize a measure of this independence at the output. Thus, the steganalysis problem for linear steganography under the stated assumptions can be described more formally as follows.

DEFINITION 2.1. *Steganalysis of a random vector $\mathbf{z}(k)$ in Eq. (2) is the computation of a linear transform $\mathbf{B}$ such that components $s(k)$ and $w(k)$ obtained from $\mathbf{r}(k) = \mathbf{B}\mathbf{z}(k)$ are as independent as possible under a suitable measure $F(.)$.*

This definition leads us to the question: when is steganalysis possible? Fortunately, we can adopt a known result from ICA[15] and show that steganalysis is possible if in addition to the statistical independence assumption of $\{s(k)\}$ and $\{w(k)\}$ we have the following:

Identifiability Condition:

- At least $\{s(k)\}$ or $\{w(k)\}$ must be non-Gaussian.
- The matrix $\mathbf{A}$ must be of full column-rank.

From the first condition we observe that using spread spectrum data embedding in the DCT domain with a Gaussian distributed message carrier can be identified because the DCT coefficients are non-Gaussian. This gives rise to a new constraint on secure steganography—*choose a Gaussian distributed cover message or pre-process the cover message so that it has a Gaussian distribution if Gaussian distributed spread spectrum image steganography is employed*. The second constraint for security is to make $\mathbf{A}$ rank deficient.

Now that we know the conditions for blind steganalysis the next question is whether the identified matrix $\mathbf{A}$ and the identified components of $\mathbf{r}(k)$ are unique? The answer is no because the columns of $\mathbf{A}$ and the independent components can be identified only up to a multiplicative constant. This is because multiplying a component of $\mathbf{r}(k)$ by a constant and dividing the corresponding column of $\mathbf{A}$ by the same constant will leave the problem unchanged, *i.e.*,

$$\mathbf{z}(k) = \mathbf{A}\mathbf{r}(k) = \sum_{p=1}^{2} \frac{\mathbf{a}_p}{\beta_p} \beta_p r_p(k) \tag{4}$$

where $\mathbf{a}_p$ is the pth column of $\mathbf{A}$, β_p is an arbitrary constant and $r_p(k)$ denotes the pth component of $\mathbf{r}(k)$. Without loss of generality if the components of $\mathbf{r}(k)$ are assumed to have unit variance then the identified components are unique up to a multiplicative sign.[15] We observe that this limitation is not serious for steganalysis. We explain this with an example. Let the cover message be the DCT coefficients of an image and signs of a Gaussian message carrier contains the embedded message. Suppose the DCT coefficients and the Gaussian message carrier are identified using the proposed technique. Then the constant β_1 in Eq. (4) can be estimated accurately by taking the inverse of the estimated DCT coefficients and comparing the resultant image with the stego image in terms of (say) peak signal to noise ratio (PSNR). If the PSNR is not a desired value then the estimated DCT coefficients can be scaled by a new value of β_1 which is chosen such that a high PSNR is obtained; meaning the estimated DCT coefficients scaled by the new β_1 is a better approximation of the cover message. This can be iterated a few number of times until a desired β_1 is obtained that produces a reliable estimate of the cover message. A similar method can also be used to compute β_2; however, note that if only the sign of the Gaussian carrier contains the hidden message computing β_2 is irrelevant. But, if the Gaussian carrier has unit variance (which is usually the case in steganography) then the signs of the computed coefficients of the Gaussian carrier can be the true signs (corresponding to the secret message bits) or they may be just the opposite signs. So, the message bits can be extracted either from the signs of the estimated carrier coefficients or by simply negating all the coefficients. One of these two is the original hidden message. Finally, we note that the proposed steganalysis method imposes no ordering on the identified independent components because

$$\mathbf{R}_{\mathbf{r}}(0) = \mathbf{I} \Rightarrow \mathbf{R}_{\mathbf{z}}(0) = E(\mathbf{z}(k)\mathbf{z}^t(k)) = \mathbf{A}\mathbf{A}^t \tag{5}$$

where $\mathbf{R}(.)$ stands for the correlation matrix, $\mathbf{I}$ is the identity matrix and t denotes the transpose of a matrix. Again, we note that this permutation indeterminacy is not serious for steganalysis. In view of these indeterminacies a more general definition of the proposed steganalysis problem can be obtained using the following definition of *essentially equal matrices*.[14]

DEFINITION 2.2. *Two matrices* $\mathbf{M}$ *and* $\mathbf{N}$ *are said to be* essentially equal *if there exists a matrix* $\mathbf{P}$ *such that* $\mathbf{M} = \mathbf{NP}$ *where* $\mathbf{P}$ *has exactly one non-zero entry in each row and column with unit modulus.*

With this definition the present steganalysis problem is the determination of a matrix essentially equal to $\mathbf{A}$. Next, we discuss an algorithm based on the steganlysis framework discussed here.

3. STEGANALYSIS ALGORITHM

We first note that if $\mathbf{A} = \begin{pmatrix} 1 & \alpha_1 \\ 1 & \alpha_2 \end{pmatrix}$ then it is full column-rank as long as $\alpha_1 \neq \alpha_2$ and therefore the proposed steganalysis technique can be applied. We then begin by whitening $\mathbf{z}(k)$, *i.e.*, apply a whitening transform $\mathbf{W}$ to $\mathbf{z}(k)$ such that

$$\begin{aligned} E(\mathbf{W}\mathbf{z}(k)\mathbf{z}^t(k)\mathbf{W}^t) &= \mathbf{W}\mathbf{R}_{\mathbf{z}}(0)\mathbf{W}^t & (6)\\ &= \mathbf{W}\mathbf{A}\mathbf{A}^t\mathbf{W}^t = \mathbf{I} \text{ using Eq.(5)} & (7) \end{aligned}$$

This means $\mathbf{WA}$ is a unitary matrix when $\mathbf{W}$ is a whitening matrix. It is also known that[14] for any whitening matrix $\mathbf{W}$ there exists a unitary matrix $\mathbf{U}$ such that $\mathbf{WA} = \mathbf{U}$. Therefore $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{W}^{\#}\mathbf{U} \tag{8}$$

where # denotes pseudo-inverse. The whitened process is still linear,

$$\begin{aligned} \mathbf{x}(k) &= \mathbf{W}\mathbf{z}(k) & (9)\\ &= \mathbf{W}\mathbf{A}\mathbf{r}(k) = \mathbf{U}\mathbf{r}(k) & (10) \end{aligned}$$

One of the main reasons for first whitening the stego message is that some computations performed during the steganalysis procedure is simplified. Therefore, the first step in our steganalysis procedure is the application of a

whitening transform $\mathbf{W}$ i to $\mathbf{z}(k)$ and making the resulting data spatially white. Now, by Definition 2.1 we want to compute a matrix $\mathbf{B}$ such that $\mathbf{Bz}(k)$ is spatially white meaning its covariance matrix is the identity matrix. Since we have spatially white data after the first step, in the second step we want to compute $\mathbf{U}$ since $\mathbf{W}$ can be computed from the observed stego message. Therefore from an implementation perspective the proposed steganalysis procedure can be described as follows:

Two step-steganalysis:

- Compute a whitening matrix $\mathbf{W} = \Gamma^{-1/2}\Xi^t$ where $\Gamma = diag(\gamma(1), \gamma(2), \ldots, \gamma(M))$ is a diagonal matrix with the eigenvalues of the covariance matrix $E(\mathbf{z}\mathbf{z}^t)$ and Ξ is a matrix with the corresponding eigenvectors as its columns. Apply $\mathbf{W}$ to $\mathbf{z}(k)$.
- Compute $\mathbf{U}$ using $\mathbf{Wz}(k)$ and hence $\mathbf{B}$.

Since $\mathbf{W}$ can be computed using the stego message (assuming a consistent estimate of its covariance matrix can be computed) we are now left with the problem of computing $\mathbf{U}$. While there are many ways of doing this we adopt a method based on higher order cumulants[14] and exploit certain algebraic structures which we described next.

Recall that if x_1, x_2, x_3 and x_4 are random variables with expected values equal to μ_1, μ_2, μ_3 and μ_4 and if $\tilde{x}_i = x_i - \mu_i$, $i = 1, 2, 3, 4$ then the fourth order cumulants are given by,

$$Cum(x_1, x_2, x_3, x_4) = E(\tilde{x}_1\tilde{x}_2\tilde{x}_3, \tilde{x}_4) - E(\tilde{x}_1\tilde{x}_2)E(\tilde{x}_3\tilde{x}_4) - E(\tilde{x}_1\tilde{x}_3)E(\tilde{x}_3\tilde{x}_4) - E(\tilde{x}_1\tilde{x}_4)E(\tilde{x}_2\tilde{x}_3) \quad (11)$$

For the $n \times 1$ random vector $\mathbf{x}$ and any $n \times n$ matrix $\mathbf{M}$ we define the associated cumulant matrix $\mathbf{Q}_\mathbf{x}(\mathbf{M})$ as the $n \times n$ matrix with components given by

$$[\mathbf{Q}_\mathbf{x}(\mathbf{M})]_{ij} = \sum_{k,l=1}^{n} Cum(x_i, x_j, x_k, x_l)\mathbf{M}_{kl} \quad (12)$$

Then it can be observed that[14]

$$\mathbf{Q}_\mathbf{x}(\mathbf{M}) = \mathbf{U}\mathbf{\Lambda}_\mathbf{M}\mathbf{U}^t \quad (13)$$

where $\mathbf{\Lambda}_\mathbf{M} = diag(K(r_1)\mathbf{u}_1^t\mathbf{M}\mathbf{u}_1, K(r_2)\mathbf{u}_2^t\mathbf{M}\mathbf{u}_2)$ is a diagonal matrix. Here, $\mathbf{u}_i$ denotes the ith column of $\mathbf{U}$ and $K(.)$ denotes kurtosis. So we see that any cumulant matrix is diagonalized by $\mathbf{U}$. Therefore the eigen vectors of the cumulant matrix left multiplied by $\mathbf{W}^\#$ gives the columns of $\mathbf{A}$. In practice the typical steps of the proposed steganalysis method are implemented as follows:

Steps involved in practical steganalysis:

- **Step 1**: Estimate an orthogonal matrix $\hat{\mathbf{W}}$ from the stego data.
- **Step 2**: Compute some empirical cumulants of $\hat{\mathbf{W}}\mathbf{z}$.
- **Step 3**: Compute an orthonormal estimate $\hat{\mathbf{U}}$ of $\mathbf{U}$ using the empirical cumulants.
- **Step 4**: Compute an estimate $\hat{\mathbf{A}}$ of $\mathbf{A}$ from $\hat{\mathbf{A}} = \hat{\mathbf{W}^\#}\hat{\mathbf{U}}$.

In the next section we discuss some experimental results based on the theoretical framework.

Figure 2: *Lenna*: Original host image.

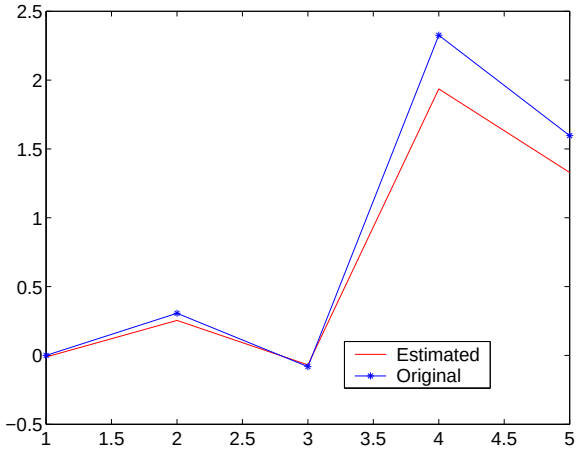

Figure 3. Original and estimated message carrier when carrier length is equal to 5. Signs of the carrier signal samples indicate the embedded message bits.

4. EXPERIMENTAL RESULTS

We apply the theory developed so far for spread spectrum steganalysis in the DCT domain. One of the main reasons for choosing spread spectrum steganography for our experiments is to test the commonly held belief that spread spectrum steganography is highly secure due to the noise-like message carrier.[4–6, 11] By spreading the spectrum of the message carrier throughout a wideband, spread spectrum steganography makes the carrier strength less than the noise strength in the band of interest thus making its detection by an intruder difficult. Usually a zero-mean Gaussian distributed message carrier is employed by these methods. This information (weakness) can be exploited by the proposed steganalysis method because the DCT co-efficients are non-Gaussian. Also, we can gather critical spatial diversity information from multiple images that carry this message to produce effective steganalysis. We again note that the steganalysis technique discussed in this paper can be applied to other signal types also.

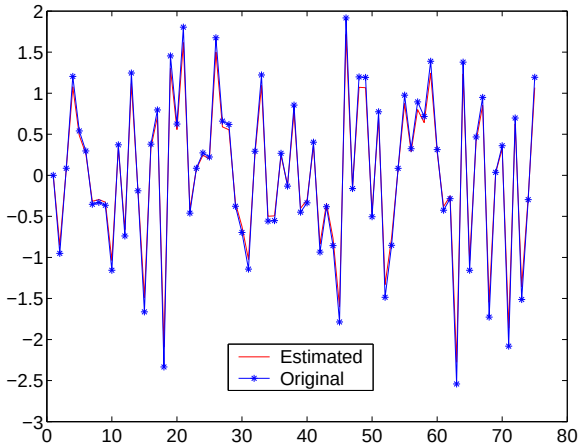

Figure 4. Original and estimated message carrier when carrier length is equal to 75. Signs of the carrier signal samples indicate the embedded message bits.

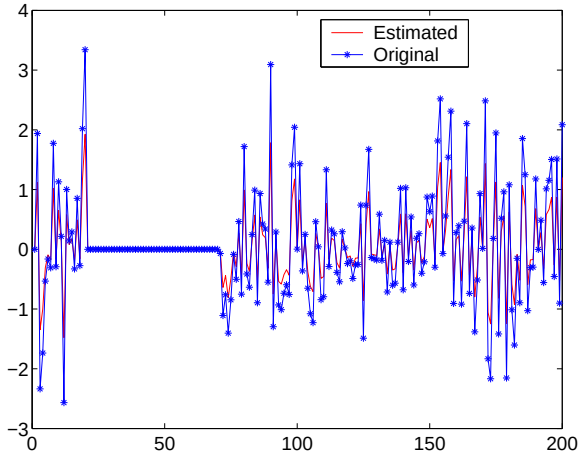

Figure 5. Original and estimated message carrier when carrier length is equal to 200. Signs of the carrier signal samples indicate the embedded message bits.

4.1. Spread Spectrum Steganography Encoder

We assume the steganography (and/or watermarking) encoder to take the form

$$y(k) = s(k) + \alpha w(k), k = 1, 2, \ldots, L \tag{14}$$

where $s(k)$ denotes the kth DCT coefficient of the host image, $w(k) \sim N(0, \sigma^2)$ is the kth sample of a Gaussian distributed message carrier, α is the carrier strength and L is the message length. For our experiments we take the 2-D DCT of the image shown in Figure 2 and ignoring the DC coefficient choose the L highest magnitude coefficients for embedding. We assume that the signs of $\{w(k)\}$ carry the message bits (positive implies bit 1 and negative stands for bit 0) and therefore the decoder is not interested in the magnitude of $w(k)$. Whereas

for spread spectrum watermarking applications the magnitude of the received message carrier is also of concern to the decoder. The role of α is to make the stego technique robust against noise attacks.

4.2. Spread Spectrum Steganalysis

We assume that two copies of the stego image are available with parameters $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$. Then, according to Eq. (2) we have,

$$\mathbf{z}(k) = \begin{pmatrix} y_1(k) \\ y_2(k) \end{pmatrix} = \mathbf{A}\mathbf{r}(k) \tag{15}$$

$$= \begin{pmatrix} 1 & 0.1 \\ 1 & 0.2 \end{pmatrix} \begin{pmatrix} s(k) \\ w(k) \end{pmatrix}, k = 1, 2, \ldots, N \tag{16}$$

where N is the total number of DCT coefficients (or size of the host image). Note that $N \geq L$. Since the steganalysis algorithm does not possess knowledge about the steganography key which in this case is the set of DCT coefficients that carry the embedded message, all the available $\mathbf{y}$ (spatial diversity information) are processed using steps discussed in Section 3 to gain knowledge about the message carrier .

4.2.1. Steganography Key and Message Length Estimation

The first goal of the steganalysis procedure is to estimate the steganography key. Note that this also gives an estimate of the embedded message length since the length of the key in this case is equal to the message length. Figures 3, 4 and 5 show the original and the estimated message carrier for message length equal to 5, 75 and 200, respectively. These message lengths were chosen to represent small, medium and reasonably large message sizes. We see from the figures that the proposed steganalysis algorithm produces good estimates of the message carrier from which the key can also be extracted. Due to the numerical nature of the experimental outputs, we assume that if the magnitude of an extracted carrier sample is less than 10^{-1} then that sample is not considered part of the carrier. For higher message lengths (≥ 1000) a threshold of 10^{-4} was used. These choices were seen to give good results through extensive experimentation. From Figure 5 we observe that even if the stego key does not contain continuously indexed DCT co-efficients the proposed method still produces good estimates.

Table 1 gives a comparative performance of the original message length versus the percentage error in the message length estimation. Note that these numbers vary from one run of the experiment to another and also depend on the random message carrier generated for that run. That's why it is not surprising to see that for $L = 5$ the estimation error is lower than some other numbers in the table. But, from the table we note that as the message length increases the steganalysis algorithm consistently produces length estimates with error percentages less than 0.13%.

Table 1: Original message length versus % error in estimated length.

L	5	25	50	75	100	1000	2000	4000	5000	6000	7000	9000
% Est. error	0	4	8	8	9	0	0.05	0.025	0.04	0.13	0.02	0.05

4.2.2. Message Estimation

By reading off the signs of the estimated carrier samples in Figures 3, 4 or 5 the embedded message bits can be estimated for the corresponding message lengths. But, we know that the carrier samples can be estimated only up to a multiplicative sign using our steganalysis technique. But, this is not a major problem because either the extracted signs give the message bits or the opposite signs. One can test both these options for the presence of an useful message.

In watermarking applications we are interested in estimating accurately both the magnitude and the sign of the embedded watermark. For this we devised a simple procedure that seems to work well in practice. Assume the steganalyst has access to the watermark detector (considered a black box). Then, the estimated watermark

can be given as input to the detector and based on the output the strength can be changed iteratively until the detector accepts the watermark. The same method can be used to fix the sign of the estimated watermark also. By using this technique we observed that the difference between the expected *Sim(.)* measure[4] and the computed Sim(.) measure using the estimated watermark was negligible for all the cases considered in the experiments. Thus watermark estimation is quite reliable and therefore can be successfully used to forge.

4.2.3. Bit Error Rate

We define bit error rate as p_e = Prob(sign(original message carrier sample)$\neq$ sign(estimated message carrier sample)). To compute p_e numerically we used the frequency based statistical estimator. p_e was computed using only min(estimated message length,L) number of samples. We found that p_e was equal to zero for all the experimental runs!

4.2.4. Estimating the Host Image

We observe that the steganalysis algorithm also produces an estimate of the DCT coefficients of the original host image. An estimate of the original host image can therefore be obtained by simply taking the inverse DCT. Before this process the magnitude and sign of the estimated DCT coefficients have to be fixed. This was achieved in a simple manner. Since the DC coefficient does not carry the message, we computed the ratio of the estimated DC coefficient to the DC coefficient of the stego image. This number was then used to scale all the estimated DCT coefficients. In order to fix the sign of the DCT coefficients we looked at the sign of the estimated DC coefficient. If this was positive the signs were not changed; however, if this was negative then all the estimated DCT coefficients were negated to obtain the final estimate. Using this procedure we obtained an estimate of the original Lenna image shown in Figure 6 for L=1000. The peak signal to noise ratio (PSNR) between the original host and this estimate is 47dB meaning that the estimate is fairly good.

Figure 6: Estimated host image when message carrier length is equal to 1000.

5. CONCLUSIONS

It is shown that by looking for the right type of spatial/time diversity information good steganalysis methods can be designed. Simple common knowledge such as the non-Gaussian distribution of the DCT coefficients can provide valuable information for steganalysis. Apart from the traditional information theory based notion of steganography security, the analysis provided here raises some interesting questions and produces some answers about other factors that also determine the true security of a steganography method.

Experimental results are provided to prove the practical utility of the proposed steganalysis method. Spread spectrum steganography is the object of the experimental study. Results show that it is fairly easy to break spread spectrum steganography and watermarking within the context of this paper.

REFERENCES

1. C. Cachin, "An information-theoretic model for steganography," *Proc. 2nd International Workshop on Information Hiding,* **LNCS 1525**, pp. 306–318, 1998.
2. G. J. Simmons, "Prisoners' problem and the subliminal channel," *Proc. CRYPTO83 - Advances in Cryptology* , pp. 51–67, August 1984.
3. M. Holliman, N. Memon, and M. Yeung, "On the need for image dependent keys for watermarking," *Proc. of IEEE Symposium on Content Security and Data Hiding in Digital Media* , May 1999.
4. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Processing* **6**(12), pp. 1673–1687, 1997.
5. J. Smith and B. Comiskey, "Modulation and information hiding in images," *Proc. First Information Hiding Workshop* **LNCS 1174**, pp. 207–226, May 1996.
6. L. Marvel, J. C.G. Boncelet, and C. Retter, "Spread spectrum image steganography," *IEEE Trans. on Image Processing* **8**, pp. 1075–1083, Aug. 1999.
7. U. http://www.digimarc.com
8. F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Processing* **66**, pp. 283–301, May 1998.
9. http://www.jpeg.org
10. R. Chandramouli and N. Memon, "Adaptive steganography," *To appear in Proc. SPIE Conf. on Security and Watermarking of Multimedia Contents* , 2002.
11. J. Fridrich and M. Goljan, "Practical steganalysis: State-of-the-art," *To appear in Proc. SPIE Conf. on Security and Watermarking of Multimedia Contents* , 2002.
12. R. Chandramouli and N. Memon, "Analysis of lsb based image steganography techniques," *Proc. IEEE International Conf. on Image Processing* **3**, pp. 1019 –1022, 2001.
13. R. Clarke, *Transform coding of images*, Academic Press, 1985.
14. J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE* **90**, pp. 2009–2026, Oct. 1998.
15. P. Comon, "Independent component analysis–a new concept?," *Signal Processing* **36**, pp. 287–314, 1994.