

A mathematical framework for active steganalysis

R. Chandramouli

Stevens Institute of Technology, B207, Department of E.C.E., Hoboken, NJ 07030, USA

Abstract. A mathematical framework for steganalysis is presented in this Paper, with linear steganography being the main focus. A mathematically formal definition of steganalysis is given. Then active steganalysis, defined as the extraction of a hidden message with little or no a priori information, is formulated as a blind system identification problem within this framework. Conditions for identifiability (i.e., 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namely, spread spectrum steganography/watermarking, is secure because the low strength, noise-like message carrier is no longer valid within the current context.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.

Key words: Steganography – Steganalysis – Security – Independent component analysis

1 Introduction

Steganalysis is a relatively new branch of research. While steganography deals with techniques for hiding information (such as fingerprinting), 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 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 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 must a class of steganography algorithms 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 [2].
- What are the candidate *cost* or *risk* functions that a steganalyst must optimize during hidden data detection or extraction procedure?
- What are the performance trade-offs 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. Note that in our analysis we assume that the steganalyst has reasonable computational resources and time.

In a traditional steganography setup formulated as a prisoner's problem [24], Alice wishes to send a secret message to Bob by hiding information in a cover message. The stego message (cover+secret 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; Wendy acts as a *passive warden*. If her decision is negative, then Wendy forwards the message to Bob. 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 mes-

sages 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 an observed message, to our knowledge there seems to have been little attempt in developing steganalysis methods that can extract the secret message. In general, extraction of the secret message is a harder problem than mere detection simply because the former outputs multiple bits of information while the latter results in 2-bit (secret message present or absent) information. Therefore, based on the ultimate outcome of the effort we classify steganalysis into two categories:

- **Passive steganalysis:** Detect presence/absence of hidden message in a stego signal, identify the stego embedding algorithm.
- **Active steganalysis:** Estimate the embedded message length, estimate locations of the hidden message, estimate the secret key used in embedding, estimate some parameters of the stego embedding algorithm, extract the hidden message (ultimate goal!).

Note that, according to our definition, *active steganalysis is 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 class of linear steganography algorithms are employed. We also discuss the strengths and limitations of the proposed framework and provide numerical examples to illustrate the performance. 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.

Passive steganalysis has been attempted previously by many researchers [1, 5, 6, 11, 13, 12, 19, 21, 23, 27]. The general theme behind these techniques is the exploitation of first-order and higher-order statistics depending on the steganography technique. A priori spatial and frequency domain information about the stego messages are used to arrive at a steganalysis strategy. When such a priori information is unavailable, two broad approaches are followed: (a) assume a priori model based on image characteristics, etc. or (b) learn the information using a large database of training set. Of course, each of these approaches has its pros and cons. It is a generally accepted fact that the more focused a steganalysis algorithm is on a specific steganography technique, the less its generalization capability. On the other hand, a very general method may not produce acceptable performance for detecting a specific steganography algorithm. Therefore, choosing the right steganalysis algorithm is in itself an open research problem.

The work closest in spirit to that presented in this paper can be found in the watermarking literature [20, 26, 28]. Here linear filtering techniques are employed to obtain an estimate of an embedded watermark. Conditions for resistance against this attack can also be found in these references. As we will see in later sections, the proposed work is significantly different from

these. The approach, analysis, assumptions, and conditions under which the proposed method fails are also different from the previous research.

The paper is organized as follows. The scope of this work is presented in Sect. 2, and a mathematical formulation of active steganalysis is introduced in Sect. 3. Based on the mathematical formulation, a steganalysis algorithm is given in Sect. 4. Experimental results are discussed in Sect. 5, and concluding remarks can be found in Sect. 6.

2 Scope of this work

We consider the scenario where the steganography key is the same for at least two stego messages. While in general this is a stronger assumption, we will see later that a nice mathematical theory can be developed for active steganalysis in this case. Clearly, our future goal is to relax this assumption. In some cases, however, it may be possible to satisfy this assumption. Alice and Bob could exchange a steganography key initially and later use this key to embed and extract multiple bits of secret information. For example, the steganography key could be the pixels of an image where the secret message is hidden such as in LSB (least significant bit) image steganography. This also simplifies the key management problem for Alice and Bob. In fact, fixed key embedding that is prevalent in many watermarking algorithms has found to be a cause for concern (e.g., Memon et. al. [16]).

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 the Gaussian random vector is positive, then it stands for message bit 1; otherwise, it represents the bit 0. The Gaussian random vector is then scaled by a positive value and added to certain discrete cosine transform (DCT) coefficients of the cover image based on a steganography key. Note that this is similar to the popular spread spectrum watermarking algorithm [10].

Finally, we assume that the steganalyst has access to at least two stego messages with the same (or highly correlated) cover message, the same secret information, and the same key but differing in some other parameters. There are both practical and academic examples supporting these assumptions, such as the following:

- Commercial products such as Digimarc's [17] image watermarking software assign each user a unique ID (or key). A user uses this fixed key to embed fingerprint information in her images. The user can choose between a wide range of data-embedding signal strengths. Clearly, the steganalyst posing as a legitimate user can buy Digimarc's software and create two stego images with a fixed key that differ only in the embedding strength.
- Spread spectrum image steganography has been previously proposed by Smith et al. [25] and Marvel et al. [22]. In such schemes, the message bits modulate a carrier function/vector (Gaussian random vector is a popular choice), and the result is then 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 video watermarking techniques [15] spread the message sequence spatially before modulating a carrier and adding it to a video frame. The carrier strength is adjusted based on the characteristics of the local spatial location. Clearly in these types of data embedding, information about the hidden message is found 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 [18] 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 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 resends it the next day after increasing the message strength and hoping it survives Wendy's attack (this is a specific case of adaptive steganography [7]). 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:** Steganalysis methods can look for information in the spatial domain that repeats itself in various forms in different spatial locations (e.g., different blocks within an image or in different images). We call this 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 within the spatial diversity steganalysis framework and their implications is provided by Fridrich et. al. [14]. One of the 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. [6]. 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, we

next describe a mathematical formulation of the steganalysis problem.

3 Mathematical formulation of active steganalysis

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 popular 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 modeled 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 that is then added to the cover message.

3.0.1 Message embedding and extraction

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

$$z(k) = s(k) + \alpha w(k), k = 1, 2, \dots, N \quad (1)$$

$\{s(k)\}$ 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 $\{w(k)\}$ to be samples from stationary random vectors. The steganography key and α are known to the legitimate 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 not have access to $s(k)$, then filtering techniques can be employed to obtain an estimate of $s(k)$ and hence an approximate version $\hat{w}(k)$ that can incur bit errors [22]. A number of possibilities such as error control coding, better estimation techniques, etc. can reduce the bit error rate. 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 but rather only the detection of the presence/absence of a message, a correlation-type detection technique that can be applied to $z(k)$ is of interest. We note that both in steganography and watermarking applications the genuine decoder may possess only a noisy copy of $\{z(k)\}$, say, $\{\hat{z}(k)\}$, due to a variety of reasons, but in this paper we assume the steganalyst has access to $\{z(k)\}$.

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 vector with finite variance. This gives rise to the so-called spread spectrum steganography [22, 25] and spread spectrum watermarking [10]. 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

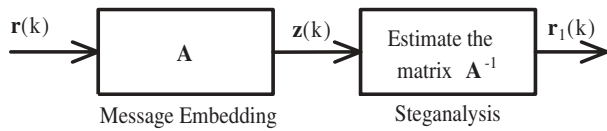


Fig. 1. Steganalysis as a blind system identification problem

observed to be robust for watermarking applications [10]. Another reason for choosing this 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 [8]. We'll later show that this fact can immensely aid in steganalysis. That is, the non-Gaussian nature of a cover message can expose the presence of a Gaussian distributed secret message.

3.0.2 Steganalysis set up

Assume that the only knowledge available to the steganalyst 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 $\{z_1(k)\}$ and $\{z_2(k)\}$. We can then write

$$\mathbf{z}(k) = \begin{pmatrix} z_1(k) \\ z_2(k) \end{pmatrix} = \mathbf{A}\mathbf{r}(k) \quad (2)$$

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

$\mathbf{z}(k)$ is the random stego message vector observed by the steganalyst, \mathbf{A} is the *strength matrix*, and $\mathbf{r}^t(k) = (r_1(k), r_2(k)) = (s(k), w(k))$ (superscript t denotes matrix transpose) 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}^{-1} from $\mathbf{z}(k)$. This can be viewed as a blind system identification problem as shown in Fig. 1. If \mathbf{A}^{-1} can be identified, then we can obtain an estimate of $\mathbf{r}(k)$, say, $\mathbf{r}_1(k)$, 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 [4]. Therefore, techniques from BSS 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 [4] 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 1 *Steganalysis of a random vector $\mathbf{z}(k)$ in Eq. 2 is the computation of a linear transform \mathbf{B} such that the*

estimates of the components $s(k)$ and $w(k)$ obtained from $\mathbf{r}_1(k) = \mathbf{B}\mathbf{z}(k)$ are as independent as possible under a suitable measure $F(\cdot)$.

Here the aim is to fit a model for the probability distribution of $\mathbf{z}(k)$ that captures the model of the independent components (whiteness constraint) of $\mathbf{r}(k)$. Suppose we assume that the probability density function of s is $f_s(\cdot)$ and that of w is $f_w(\cdot)$. Then the joint probability density function of the random vector \mathbf{r} is $f_{\mathbf{r}} = f_s f_w$. Then for a given \mathbf{A} the probability density of \mathbf{z} is given by:

$$f_{\mathbf{z}} = |\det \mathbf{A}|^{-1} f_{\mathbf{r}}(\mathbf{A}^{-1}\mathbf{z}) \quad (4)$$

From this we see that for L independent samples $\{\mathbf{z}(k)\}_{k=1}^L$ the normalized log likelihood is given by:

$$\mathcal{L}(L) = \frac{1}{L} \sum_{l=1}^L \ln f_{\mathbf{r}}(\mathbf{A}^{-1}\mathbf{z}(l)) - \ln |\det \mathbf{A}| \quad (5)$$

Under certain regularity conditions, by law of large numbers we then get $\mathcal{L}(L) \rightarrow E(\ln f_{\mathbf{r}}(\mathbf{A}^{-1}\mathbf{z})) - \ln |\det \mathbf{A}|$ as $L \rightarrow \infty$. Here $E(\cdot)$ stands for the expected value. This limit can be seen to be equal to $-H(\mathcal{P}_{\mathbf{z}}) - \mathcal{K}(\mathcal{P}_{\mathbf{r}_1}|\mathcal{P}_{\mathbf{r}})$ [3], where \mathcal{P} , $H(\cdot)$ and \mathcal{K} stand for probability distribution, differential entropy, and the Kullback-Leibler divergence, respectively. Since $H(\mathcal{P}_{\mathbf{z}})$ does not play any role in the steganalysis optimization procedure, we arrive at the contrast function $\phi(\mathbf{r}_1) = \mathcal{K}(\mathcal{P}_{\mathbf{r}_1}|\mathcal{P}_{\mathbf{r}})$ that has to be minimized by the steganalysis algorithm under the whiteness constraint. Note that the contrast function is a measure of the difference between the distribution of \mathbf{r}_1 and the distribution of the independent components in \mathbf{r} .

Now Definition 1 leads us to the question: when is steganalysis possible using this approach? Fortunately, we can adopt a known result from ICA [9] 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 by the proposed steganalysis framework because the DCT coefficients are non-Gaussian. This gives rise to a new constraint on secure steganography—*choose a Gaussian distributed cover message or preprocess 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 successful 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.,

$$\mathbf{z}(k) = \mathbf{A}\mathbf{r}(k) = \sum_{p=1}^2 \frac{\mathbf{a}_p}{\beta_p} \beta_p r_p(k) \quad (6)$$

where \mathbf{a}_p is the p -th column of \mathbf{A} , β_p is an arbitrary constant, and $r_p(k)$ denotes the p -th 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 [9]. 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 ICA. Then the constant β_1 in Eq. 6 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 higher 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 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 spread spectrum steganography), then the signs of the computed coefficients of the Gaussian message carrier can be the true signs (corresponding to the secret message bits) or they may be just the opposite signs. Thus 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.

We also note that the proposed steganalysis method imposes no ordering on the identified independent components because

$$\mathbf{R}_r(0) = \mathbf{I} \Rightarrow \mathbf{R}_z(0) = E(\mathbf{z}(k)\mathbf{z}^t(k)) = \mathbf{A}\mathbf{A}^t \quad (7)$$

where $\mathbf{R}(\cdot)$ stands for the correlation matrix and \mathbf{I} is the identity 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* [4].

Definition 2 *Two matrices \mathbf{M} and \mathbf{N} are said to be essentially equal if there exists a matrix \mathbf{P} such that $\mathbf{M} = \mathbf{N}\mathbf{P}$ where \mathbf{P} has exactly one nonzero entry in each row and column with unit modulus.*

Definition 3 *Steganalysis is the problem of determining a matrix essentially equal to \mathbf{A} .*

Next we discuss a steganalysis algorithm based on the framework proposed here.

4 Steganalysis algorithm

We first note that if $\mathbf{A} = \begin{pmatrix} 1 & \alpha_1 \\ & \alpha_2 \end{pmatrix}$, then it is full-column rank as long as $\alpha_1 \neq \alpha_2$, and therefore the proposed steganalysis

technique can be applied. We begin by whitening $\mathbf{z}(k)$, i.e., applying a whitening transform \mathbf{W} to $\mathbf{z}(k)$ such that

$$\begin{aligned} E(\mathbf{W}\mathbf{z}(k)\mathbf{z}^t(k)\mathbf{W}^t) &= \mathbf{W}\mathbf{R}_z(0)\mathbf{W}^t \\ &= \mathbf{W}\mathbf{A}\mathbf{A}^t\mathbf{W}^t = \mathbf{I} \text{ from Eq. 7.} \end{aligned} \quad (8)$$

This means $\mathbf{W}\mathbf{A}$ is a unitary matrix when \mathbf{W} is a whitening matrix. It is also known that [4] for any whitening matrix \mathbf{W} there exists a unitary matrix \mathbf{U} such that $\mathbf{W}\mathbf{A} = \mathbf{U}$. Therefore, \mathbf{A} can be factored as

$$\mathbf{A} = \mathbf{W}\#\mathbf{U} \quad (9)$$

where $\#$ denotes pseudoinverse. Note that here we use pseudoinverse so that we can also handle the general case when the number of rows of the matrix \mathbf{A} is greater than the number of columns. If the matrix \mathbf{A} is square (as in the current work), then the pseudoinverse is the regular inverse if \mathbf{A} is full rank. Now since the whitened process is still linear, we have

$$\mathbf{x}(k) = \mathbf{W}\mathbf{z}(k) \quad (10)$$

$$= \mathbf{W}\mathbf{A}\mathbf{r}(k) = \mathbf{U}\mathbf{r}(k) \quad (11)$$

The reasons for first whitening the stego message are the following. By taking an orthogonal ICA approach to steganalysis we are looking to compute a matrix \mathbf{B} such that $\mathbf{B}\mathbf{z}(k)$ is spatially white (by Definition 1), i.e., its covariance matrix is the identity matrix. We know that, in practice, components that are as independent as possible according to some metric of independence are not necessarily uncorrelated. Therefore, we enforce the decorrelation condition by using a whitening process as a first step. Once this whitening is performed, it is enough to compute an orthonormal transform to be applied to the whitened data to obtain the required estimates. This is because our goal is to compute a white vector that contains estimates of the original and the stego message, and only an orthonormal transform can preserve whiteness. Finally, we will see later in this paper that the whitening process results in reducing the computational complexity of the steganalysis process. Therefore, we first apply a whitening transform \mathbf{W} to $\mathbf{z}(k)$ and make the resulting data spatially white. Since we have spatially white data after this step, in the second step we want to compute \mathbf{U} from \mathbf{W} . 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 $\Gamma = \text{diag}(\gamma(1), \gamma(2), \dots, \gamma(M))$ 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 $\mathbf{W}\mathbf{z}(k)$ 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} . To this end, we start with the contrast function $\phi(\mathbf{r}_1)$. To compute this function, we observe that knowledge of \mathcal{P}_r is required. Since this may not be available to the steganalysis algorithm, it must be approximated. A useful tool in obtaining such an approximation is the usage of higher-order cumulants [4]. Recall

that if y_1, y_2, y_3 , and y_4 are random variables with expected values equal to μ_1, μ_2, μ_3 , and μ_4 , and if $\tilde{y}_i = y_i - \mu_i$, $i = 1, 2, 3, 4$, then the second- and fourth-order cumulants are respectively given by

$$Cum(y_1, y_2) = E(\tilde{y}_1 \tilde{y}_2) \quad (12)$$

and

$$Cum(y_1, y_2, y_3, y_4) = E(\tilde{y}_1 \tilde{y}_2 \tilde{y}_3, \tilde{y}_4) - E(\tilde{y}_1 \tilde{y}_2)E(\tilde{y}_3 \tilde{y}_4) - E(\tilde{y}_1 \tilde{y}_3)E(\tilde{y}_2 \tilde{y}_4) - E(\tilde{y}_1 \tilde{y}_4)E(\tilde{y}_2 \tilde{y}_3)$$

From these definitions we clearly see that the variance and kurtosis of a real random variable y are given by $\sigma_y^2 = Cum(y, y)$ and $K(y) = Cum(y, y, y, y)$. Then if the probability density function $f_y(\cdot)$ of a random variable y is close to the standard normal density, then by truncating the Edgeworth expansion we get the following approximation [9]:

$$f_y(t) \approx \frac{1}{\sqrt{2\pi}} \left(1 + \frac{\sigma_y^2 - 1}{2} h_2(t) + \frac{K(y)}{4!} h_4(t) \right) \quad (13)$$

where $h_2(t) = t^2 - 1$ and $h_4(t) = t^4 - 6t^2 + 3$ are the second- and fourth-order Hermite polynomials. Extending this result to the multivariate case we also get the following approximation:

$$\phi(\mathbf{r}_1) \approx \mathcal{K}(\mathcal{P}_{\mathbf{r}_1} | \mathcal{P}_{\mathbf{r}}) \quad (14)$$

$$\begin{aligned} &\approx \frac{1}{4} \sum_{ij} (R_{ij}^{\mathbf{r}_1} - R_{ij}^{\mathbf{r}})^2 \\ &+ \frac{1}{48} \sum_{ijkl} \left(Q_{ijkl}^{\mathbf{r}_1} - Q_{ijkl}^{\mathbf{r}} \right)^2, \end{aligned} \quad (15)$$

where $R_{ij}^{\mathbf{r}} = Cum(r_i, r_j)$ (correspondingly for \mathbf{r}_1) and $Q_{ijkl}^{\mathbf{r}} = Cum(r_i, r_j, r_k, r_l)$ (correspondingly for \mathbf{r}_1) and $i, j = 1, 2$. From this equation we see that the Kullback-Leibler distance can be approximated by mismatch in the cumulants between \mathbf{r} and \mathbf{r}_1 . Since the components of \mathbf{r} are independent, its cross cumulants are zero, thereby reducing Eq. 14 to the following:

$$\begin{aligned} \phi(\mathbf{r}_1) &\approx \frac{1}{4} \sum_{ij} (R_{ij}^{\mathbf{r}_1} - \sigma_{r_i}^2 \delta_{ij})^2 + \\ &\frac{1}{48} \sum_{ijkl} \left(Q_{ijkl}^{\mathbf{r}_1} - K(r_i) \delta_{ijkl} \right)^2 \end{aligned} \quad (16)$$

where $\delta_{iiii} = 1$ and 0 otherwise. From this we note that the steganalyst minimizes $\phi(\mathbf{r}_1)$ when $R_{ij}^{\mathbf{r}_1} = \sigma_{r_i}^2 \delta_{ij}$ and $Q_{ijkl}^{\mathbf{r}_1} = K(r_i) \delta_{ijkl}$. This means $\phi(\mathbf{r}_1)$ now becomes

$$\phi(\mathbf{r}_1) = \underbrace{\frac{1}{4} \sum_{ij \neq ii} (R_{ij}^{\mathbf{r}_1})^2}_{\text{Term I}} + \frac{1}{48} \sum_{ijkl \neq iiii} \left(Q_{ijkl}^{\mathbf{r}_1} \right)^2 \quad (17)$$

Since the whiteness constraint is imposed on \mathbf{r}_1 , $R_{ij}^{\mathbf{r}_1} = 0$, $\forall ij \neq ii$, and therefore term I in Eq. 17 is zero. Then by dropping the constant $\frac{1}{48}$ we arrive at the contrast function in terms of the cross cumulants

$$\phi(\mathbf{r}_1) = \sum_{ijkl \neq iiii} \left(Q_{ijkl}^{\mathbf{r}_1} \right)^2 \quad (18)$$

which is a measure of independence between the entries of \mathbf{r}_1 that we use in Definition 1.

For the 2×1 random vector \mathbf{z} and any 2×2 matrix \mathbf{M} we define the associated cumulant matrix $\mathbf{Q}_{\mathbf{z}}(\mathbf{M})$ as the 2×2 matrix with components given by

$$[\mathbf{Q}_{\mathbf{z}}(\mathbf{M})]_{ij} = \sum_{k,l=1}^2 Cum(z_i, z_j, z_k, z_l) \mathbf{M}_{kl} \quad (19)$$

Since $\mathbf{z} = \mathbf{A}\mathbf{r}$ and the components of \mathbf{r} are independent, we have

$$Cum(z_i, z_j, z_k, z_l) = \sum_{q=1}^2 K(r_q) a_{iq} a_{jq} a_{kq} a_{lq} \quad (20)$$

where a_{iq} denotes the iq -th entry of \mathbf{A} , etc. Therefore, from Eqs. 19 and 20 we see that

$$\mathbf{Q}_{\mathbf{z}}(\mathbf{M}) = \mathbf{A} \Delta(\mathbf{M}) \mathbf{A}^t \quad (21)$$

where $\Delta(\mathbf{M}) = \text{diag}(K(r_1) \mathbf{a}_1^t \mathbf{M} \mathbf{a}_1, K(r_2) \mathbf{a}_2^t \mathbf{M} \mathbf{a}_2)$ is a diagonal matrix and \mathbf{a}_i denotes the i th column of \mathbf{A} . Applying the same analysis with $\mathbf{U}(= \mathbf{W}\mathbf{A})$ in place of \mathbf{A} we get [4]

$$\mathbf{Q}_{\mathbf{x}}(\mathbf{M}) = \mathbf{U} \Lambda_{\mathbf{M}} \mathbf{U}^t \quad (22)$$

where $\Lambda_{\mathbf{M}} = \text{diag}(K(r_1) \mathbf{u}_1^t \mathbf{M} \mathbf{u}_1, K(r_2) \mathbf{u}_2^t \mathbf{M} \mathbf{u}_2)$, and \mathbf{u}_i are the i -th column of \mathbf{U} . So we see that any cumulant matrix is diagonalized by \mathbf{U} . This diagonalization procedure can be made to behave computationally [4] by choosing $\mathbf{M} = \mathbf{e}_k \mathbf{e}_k^t$ where \mathbf{e}_k is a column vector with 1 in the k -th position and 0 elsewhere. In this case, it is easily seen that $[\mathbf{Q}_{\mathbf{x}}(\mathbf{M})]_{ij} = Cum(x_i, x_j, x_k, x_l)$ and is therefore seen to be equivalent to using the contrast function ϕ for steganalysis. The eigenvectors of the cumulant matrix left multiplied by $\mathbf{W}^\#$ give the columns of \mathbf{A} . In practice, the typical steps of the proposed steganalysis method are implemented as follows:

- **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.

5 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 [10, 14, 22, 25]. By spreading the spectrum of the message carrier throughout a wide band, 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

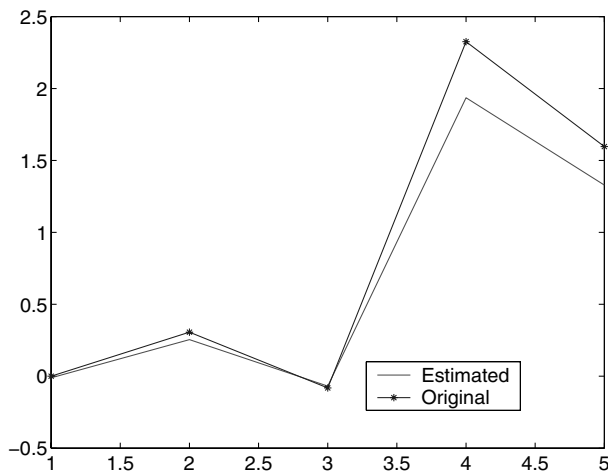


Fig. 2. Original and estimated message carrier when carrier length is equal to 5. Signs of the carrier signal samples indicate the embedded message bits

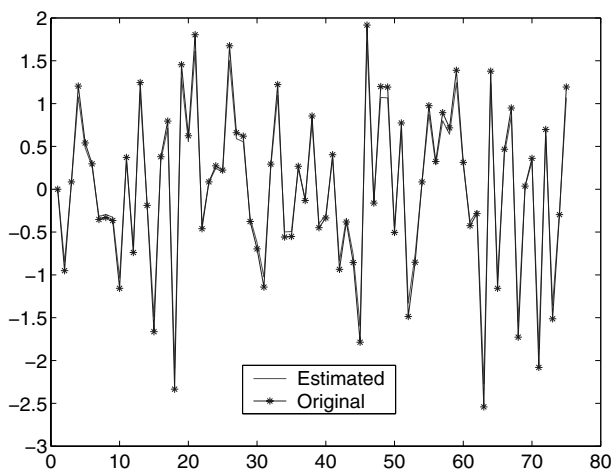


Fig. 3. Original and estimated message carrier when carrier length is equal to 75. Signs of the carrier signal samples indicate the embedded message bits

message carrier is employed by these methods. This information (weakness) can be exploited by the proposed steganalysis method because the DCT coefficients are non-Gaussian. We performed experiments on several images but present results only for one image for the sake of brevity. We however note that the results observed were more or less similar for several test images.

5.1 Spread spectrum steganography encoder

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

$$z(k) = s(k) + \alpha w(k), k = 1, 2, \dots, L \quad (23)$$

where $s(k)$ denotes the k -th DCT coefficient of the host image, $w(k) \sim N(0, \sigma^2)$ is the k -th 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

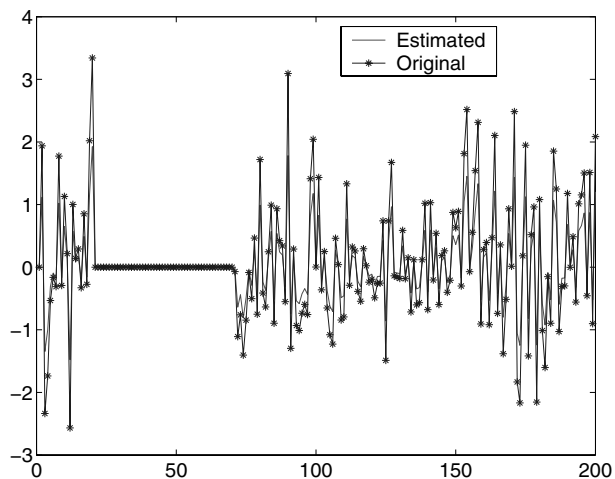


Fig. 4. Original and estimated message carrier when carrier length is equal to 200. Signs of the carrier signal samples indicate the embedded message bits

the 256×256 Lenna image 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.

5.2 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

$$\begin{aligned} \mathbf{z}(k) &= \begin{pmatrix} z_1(k) \\ z_2(k) \end{pmatrix} = \mathbf{A}\mathbf{r}(k) \\ &= \begin{pmatrix} 1 & 0.1 \\ 1 & 0.2 \end{pmatrix} \begin{pmatrix} s(k) \\ w(k) \end{pmatrix}, k = 1, 2, \dots, N \end{aligned} \quad (24)$$

where $N \geq L$ is the total number of DCT coefficients (or size of the host image).

5.2.1 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 2, 3, and 4 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

Table 1. Original message length vs. % error in estimated length

L	5	25	50	75	100	1000	2000	4000	5000	6000	7000	9000
% Average Est. error	0.01	0.03	0.027	0.09	0.05	0.04	0.05	0.025	0.04	0.13	0.02	0.05

less than 10^{-1} , then that sample is not considered part of the carrier (insignificant amplitude). For higher message lengths (≥ 1000) a threshold of 10^{-4} was used. These choices were seen to give good results through extensive experimentation. From Fig. 4 we observe that even if the stego key does not contain continuously indexed DCT coefficients, the proposed method still produces good estimates.

Table 1 gives a comparative performance of the original message length vs. the average percentage error in the message length estimates. These numbers were obtained by averaging over ten simulation runs. Note that these numbers depend on the random message carrier generated for each run. From the table we note that the steganalysis algorithm consistently produces length estimates with a negligibly small average error percentage. Of course, how small small can be depends on the type of steganography application. For example, for malicious use of steganography we would ideally like these percentages to be exactly equal to 0. But for most practical applications the numbers observed in our experiments seem to indicate that the proposed technique could be a very effective active steganalysis attack.

5.2.2 Message estimation

By reading off the signs of the estimated carrier samples in Figures 2, 3, or 4 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. However, this is not a major problem because either the extracted signs or the opposite signs give the message bits. One can test both these options for the presence of a 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 watermark estimate produced by the proposed algorithm can be given as input to the detector and, based on the detector output, the watermark strength can be changed iteratively until the detector accepts the watermark with a high probability. The same method can also be used to fix the sign of the estimated watermark. By using this technique we observed that the difference between the expected $Sim(\cdot)$ measure [10] and the computed $Sim(\cdot)$ 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.

5.2.3 Bit error rate

We define bit error rate (p_e) to be equal to the probability that the sign of the original message carrier sam-

**Fig. 5.** Estimated host image when message carrier length is equal to 1000

ple is not equal to the sign of the estimated message carrier sample. To compute p_e numerically we used a simple frequency-based statistical estimator. p_e was computed using only $\min(\text{estimated message length}, L)$ number of samples. We found that p_e was equal to zero for all the experimental runs! This means that we were able to extract all the hidden bits correctly in these experiments. Of course, this does not imply that the proposed algorithm can recover the message bits perfectly all the time.

5.2.4 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 is executed, 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 Fig. 5 for $L = 1000$. The peak signal-to-noise ratio (PSNR) between the original host and this estimate is 47 dB, meaning that the estimate is fairly good.

6 Conclusion

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. We are currently working on active steganalysis when only one copy of the stego message is available.

Acknowledgements. This material is based on research sponsored by the Air Force Research Laboratory (AFRL) under agreement number F306020-02-2-0193 and NSF DAS 0242417. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation therein. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of Air Force Research Laboratory or the U.S. government.

References

1. Avcibas I, Sankur B, Memon ND (2001) Steganalysis based on image quality metrics. Proceedings of the IEEE workshop on multimedia signal processing, Cannes, October 2001, pp 517–522
2. Cachin C (1998) An information-theoretic model for steganography. Proceedings of the 2nd international workshop on information hiding, Berlin, April 1998. Lecture notes in computer science, vol 1525. Springer, Berlin Heidelberg New York, pp 306–318
3. Cardoso JF (1997) Infomax and maximum likelihood for source separation. IEEE Signal Process Lett (4):112–114
4. Cardoso JF (1998) Blind signal separation: statistical principles. Proc IEEE 90(8):2009–2026
5. Chandramouli R, Memon N (2000) A distributed detection framework for watermark analysis. Proceedings of the ACM multimedia workshop on multimedia and security, Los Angeles, October 2000, pp 123–126
6. Chandramouli R, Memon N (2001) Analysis of lsb based image steganography techniques. Proceedings of the IEEE international conference on image processing, Thessaloniki, October 2001, pp 1019–1022
7. Chandramouli R, Memon N (2002) Adaptive steganography. Proceedings of the SPIE conference on security and watermarking of multimedia contents, San Jose, January 2002 (in press)
8. Clarke RJ (1985) Transform coding of images. Academic, New York
9. Comon P (1994) Independent component analysis—a new concept? Signal Process 36:287–314
10. Cox I, Kilian J, Leighton T, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
11. Fridrich J, Du R, Meng L (2000) Steganalysis of lsb encoding in color images. Proceedings of the IEEE international conference on multimedia and expo, New York, July 2000, vol 3, pp 1279–1282
12. Fridrich J, Goljan M, Du R (2001) Steganalysis based on jpeg compatability. Proceedings of SPIE multimedia systems and applications IV, Denver, August 2001, pp 275–280
13. Fridrich J, Goljan M, Du R (2001) Reliable detection of lsb steganography in grayscale and color images. Proceedings of the ACM workshop on multimedia security, Ottawa, October 2001, pp 27–30
14. Fridrich J, Goljan M (2002) Practical steganalysis: state-of-the-art. Proceedings of the SPIE conference on security and watermarking of multimedia contents, San Jose, January 2002
15. Hartung F, Girod B (1998) Watermarking of uncompressed and compressed video. Signal Process 66(3):283–301
16. Holliman M, Memon N, Yeung NM (1999) On the need for image dependent keys for watermarking. Proceedings of the IEEE symposium on content security and data hiding in digital media, Newark, May 1999 (preprint)
17. URL <http://www.digimarc.com>
18. URL <http://www.jpeg.org>
19. Johnson NF, Jajodia S (1998) Steganalysis of images created using current steganography software. Lecture notes on computer science, vol 1525. Springer, Berlin Heidelberg New York, pp 273–289
20. Kutter M, Voloshynovskiy S, Herrigel A (2000) The watermark copy attack. Proceedings of SPIE security and watermarking of multimedia content II, San Jose, January 2000, vol 3971, pp 371–381
21. Lyu S, Farid H (2002) Detecting hidden messages using higher-order statistics and support vector machines. Proceedings of the 5th international workshop on information hiding, Noordwijkerhout, The Netherlands (available at <http://www.cs.dartmouth.edu/farid/>)
22. Marvel Jr LM, Boncelet CG, Retter CT (1999) Spread spectrum image steganography. IEEE Trans Image Process 8(8):1075–1083
23. Provos N, Honeyman P (2002) Detecting steganographic content on the Internet. Proceedings of ISOC NDSS'02, San Diego, February 2002 (available at <http://www.citi.umich.edu/u/provos/>)
24. Simmons GJ (1984) Prisoners' problem and the subliminal channel. Proceedings of CRYPTO83 – Advances in cryptology, Santa Barbara, August 1984, pp 51–67
25. Smith JR, Comiskey BO (1996) Modulation and information hiding in images. Proceedings of the 1st information hiding workshop, Lecture notes in computer science, vol 1174, Springer, Berlin Heidelberg New York, pp 207–226
26. Su JK, Eggers JJ, Girod B (2001) Analysis of digital watermarks subjected to optimum linear filtering and additive noise. Signal Process 81(6):1141–1175
27. Voloshynovskiy S, Herrigel A, Ritsar Y, Pun T (2002) Stegowall: Blind statistical detection of hidden data. Proceedings of SPIE photonics west, security and watermarking of multimedia contents IV, Santa Clara, CA, pp 57–68
28. Voloshynovskiy S, Pereira S, Herrigel A, Baumgartner N, Pun T (2000) Generalized watermark attack based on watermark estimation and perceptual remodulation. Proceedings of SPIE security and watermarking of multimedia content II, San Jose, pp 358–370