IS IMAGE STEGANOGRAPHY NATURAL?

By

Alvaro Martín

Guillermo Sapiro

and

Gadiel Seroussi

IMA Preprint Series # 1965

(February 2004)

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
514 Vincent Hall
206 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Is Image Steganography Natural?[*]

[*]*Alvaro Martín*[†], *Guillermo Sapiro*[‡], and *Gadiel Seroussi*[§]

Abstract

Steganography is the art of secret communication. Its purpose is to hide the presence of information, using for example images as covers. We experimentally investigate if stego-images, bearing a secret message, are statistically "natural." For this purpose, we use recent results on the statistics of natural images and investigate the effect of some popular steganography techniques. We found that these fundamental statistics of natural images are, in fact, generally altered by the hidden "non-natural" information. Frequently, the change is consistently biased in a given direction. However, for the class of natural images considered, the change generally falls within the intrinsic variability of the statistics, and thus does not allow for reliable detection, unless knowledge of the data hiding process is taken into account. These results have consequences both in the art of steganography and in the mathematical modeling of natural images.

Index Terms— Steganography, Information Hiding, Image Models, Natural Images.
EDICS— 5-AUTH Authentication and Watermarking.

1 Introduction

In steganography, we study techniques to achieve secret communication between two parties that are interested in hiding not only the content of a secret message but also the act of communicating it. To this aim, steganography algorithms ("*stego algorithms*") embed the secret information into different types of "natural" cover data like sound, images, or video. The resulting altered data is referred to as *stego-data* and it must be perceptually indistinguishable from its natural cover. On the other hand, *stego-analysis* seeks to analyze (possibly altered) cover data to decide whether a message has been embedded in it or not. Thus, the problem can be seen as one of classification into two classes, namely, natural and stego-data.

In this paper we focus on the use of natural images, i.e., images that appear naturally in "real world" photographic scenes, as covers, and study how several recently proposed statistical models can be used for stego-analysis. We also experiment with basic statistics based on wavelet coefficients and block discrete cosine transform coefficients, exploiting (partial) knowledge of the data hiding

[*] This work is partially supported by the Office of Naval Research grants N000140310399 and N000140310176, by the Presidential Early Career Award for Scientists and Engineers (PECASE), and a National Science Foundation CAREER Award.

[†] Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Work done while the author was with Information Theory Research, Hewlett-Packard Laboratories, Palo Alto, CA 94304, and Electrical and Computer Engineering Department, University of Minnesota, Minneapolis, MN 55455.

[‡] Electrical and Computer Engineering and Digital Technology Center, University of Minnesota, Minneapolis, MN 55455.

[§] Information Theory Research, Hewlett-Packard Laboratories, Palo Alto, CA 94304.

technique. The goal then is to investigate if the act of embedding (hiding) a “non-natural” message into a “natural” image, changes some of the basic statistics of the image, thereby allowing for the detection (but not necessarily interpretation) of the presence of a hidden message. For instance, we will show that a model for the distribution of the differences between adjacent pixels, which fits natural images very accurately, is not a good model for images altered by one of the stego algorithms in S-Tools [1], a popular package we included in our experiments. Other algorithms, like Jsteg [2], however, do not significantly violate this property.

While previous works, [3, 4], had focused on rather simple image statistics, in [5], the authors proposed a stego-analysis technique based on image quality metrics while, in [6, 7], Farid proposed a technique based on high order statistics of wavelet coefficients. Recently, in [8], a stego algorithm invulnerable to Farid’s technique was introduced. This algorithm is a modified version of the Histogram-Preserving Data Mapping (HPDM) [9], and we will refer to it as MHPDM.

One of the main conclusions of this work is that embedding a stego message generally alters the studied statistics of its cover image. Moreover, in some cases, the hidden data causes a consistent bias in some of the statistical parameters. On the other hand, the effect is often not sufficient to “move” a significantly large set of images beyond what may be considered natural according to the studied statistical models, when the analysis is independent of the stego algorithm used. As we demonstrate below, better results, including statistically significant discrimination between natural and stego-images, can be obtained when (partial) knowledge of the stego algorithm is used in the analysis.

The remainder of this paper is organized as follows. Section 2 briefly describes the steganography algorithms that are considered in our experiments, and Section 3 introduces the models of natural images that are tested for sensitivity to steganography. Section 4 describes the general setting for the experiments, the specifics of each experiment, and the results obtained. Finally, the conclusions on the results, and directions for future research, are summarized in Section 5.

2 Steganography Algorithms

We consider three different stego algorithms in our experiments: Jsteg [2], the above mentioned MHPDM [8], and one of the algorithms in S-Tools [1]. Jsteg embeds a message in the least significant bit of JPEG DCT coefficients. The algorithm selected in S-Tools admits 8-bits palletized images (256 colors) as inputs, and maintains this range throughout processing. The algorithm operates in two stages. First it reduces the number of entries in the color palette of the cover image, and then it embeds a message in the least significant bits of the three RGB components, without expanding the number of colors beyond 256. Note of course that, as each RGB component of each pixel is altered independently, this technique is not directly suitable for gray images since it can be detected by simply observing that some colors in the color palette are not exactly gray. We experimented with this algorithm as an example of a scheme operating in the space domain. To study the effects of S-Tools purely on image statistics (our main focus in the paper), the mentioned color-shift issue was bypassed by transforming RGB stego-images back to gray scale, taking the rounded luminance of each pixel.

The MHPDM algorithm [9], as well as its predecessor HPDM [8], works by altering the least significant bit of a subset of the JPEG DCT coefficients of an image. If the 64 coefficients of each DCT block are indexed from zero following the usual zig-zag order [10], only coefficients 1 through 20 are candidates for modification. The rest are left untouched, since values of coefficient with index 0 (DC) are far from being independent, and coefficients 21 through 63 are highly quantized during the JPEG process.

Both MHPDM and HPDM preserve the zero-order histograms of each DCT frequency independently. Denoting by $x_{i,j}$ the value of DCT coefficient j at block i for a given image I, and $x'_{i,j}$ the corresponding value for a stego image I' with cover I, the histograms of $\{x_{i,j}\}$ and $\{x'_{i,j}\}$ are preserved for all fixed j in the range 0..63. In order to do that, it is necessary that the message bit stream to be embedded in the j-coefficients has the same memory-less empirical distribution as $\{lsb(x_{i,j})\}$, where $lsb(x)$ denotes the least significant bit of x. This is done by assuming that the input message b has approximately as many zeros as ones, and processing it with an entropy decoder designed for $P(b_i = 1) = \hat{P}(lsb(x_{i,j} = 1)$, the latter denoting the mentioned empirical distribution of the least significant bit of the j-th DCT coefficient. The value of this probability is included with the coded data, to allow for lossless decoding of the hidden data. In [8], the authors showed certain weakness of the HPDM algorithm with respect to Farid's stego-analysis (which is based on statistics of wavelets coefficients), and observed that it could be avoided by not modifying coefficients with values 0,1 and -1. This modification constitutes basically the MHPDM algorithm that we use in our experiments.

3 Models of Natural Images

Our experiments are based on statistics based on wavelet coefficients, block discrete cosine transform coefficients and three recently proposed statistical models of natural images. These models, which are briefly described below, reflect in general properties that are more global than those used in earlier stego-analysis works.

3.1 Areas of Connected Components Model

In [11, 12], it is observed that the distribution of the areas of connected components of bilevel (thresholded) images follow a power law which depends on just two parameters, an exponent α and a scaling factor C. More precisely, consider an image I whose gray levels are between 0 and N. For an integer k, define the bilevel (thresholded) images

$$I_l(i,j) := \begin{cases} 1 & \text{if } (l-1)\frac{N}{k} \leq I(i,j) \leq l\frac{N}{k}, \\ 0 & \text{otherwise.} \end{cases}$$

In [11, 12], the authors found that the total number $f(a)$ of connected components of the bilevel images I_l with area a is

$$f(a) \approx C a^{\alpha}$$

Furthermore, it was experimentally found and theoretically justified [12] that the exponent α is close to -2 for natural images. We refer to this model as the *Areas Model*. We should note that this is a strongly non-local statistical model, since it looks at areas and at all bilevel images simultaneously. This is in sharp contrast with models based on individual pixels statistics, which were common in earlier works.

3.2 Adjacent Pixel Values Model

In [13, 14, 15], a statistical model for the horizontal derivative $I_x = \frac{\partial I}{\partial x}$ of an image I is introduced. Based on the *transported generator model* [16], the authors model an image as a random number of profiles of the same object and each pixel is obtained as a linear combination of these profiles, weighted randomly. Mathematically,

$$I(z) = \sum_i a_i g(z - z_i)$$

where z and z_i are coordinates in $\mathbb{R}^2$ denoting a pixel location and an object profile location respectively, g is the profile of an object, and the coefficients a_i are random weights. Locations z_i are modeled as samples from a 2D Poisson process with uniform intensity, and weights a_i are modeled as independent and identically distributed (IID), also independent of the z_i-s.

Under this model and certain assumptions on $u(z) = \sum_i g_x^2(z - z_i)$, the authors show that the probability density function of I_x is

$$f(t) = \frac{1}{\sqrt{\pi}\Gamma(p)}(\frac{c}{2})^{-\frac{p}{2}-\frac{1}{4}}(2)^{-p+\frac{1}{2}}t^{p-\frac{1}{2}}K_{p-\frac{1}{2}}(\sqrt{\frac{2}{c}}t), \text{ for } p > 0,$$

where K is the modified Bessel function, Γ is the Gamma function, and p and c are two parameters referred to as *shape parameter* and *scale parameter* respectively. Furthermore, they show that p and c satisfy

$$p = \frac{3k_1^2}{k_2}, \quad c = \frac{k_2}{3k_1}$$

where $k_1 = E[I_x^2]$ and $k_2 = E[I_x^4]$.

Notice that given an image I, one can approximate I_x as the difference between adjacent pixel values and estimate k_1 and k_2, obtaining thereby an estimate of $f(t)$. We will refer to this model as the *PC Model.*

3.3 Laplacian Distribution Model

In [17], the author reports on an empirically observed property of natural images referred to as *Differentially Laplacian.* It is observed that for a reasonably small constant k, and any fixed set of k^2 coefficients adding up to 0, the linear combination of k^2 pixel intensities in a $k \times k$ square, using these k^2 coefficients as weights, tends to exhibit a Laplacian-like distribution for natural images (this is related to the well known Laplacian distribution of prediction errors in image coding [18]).

4 Experimental Results

4.1 Experimental Setting

For all experiments we used gray scale 1536×1024 natural images from Van Hateren's data base.[1] The 12-bits pixel values of all images are proportional to the light intensities in the scenes; however, the multiplying constant need not be the same for different images. In experiments where this disparity might affect the statistics of interest, we follow [19], and use *log-contrast* images. In the log-contrast image of I, the pixel at location (i, j) is calculated as $log^+(I(i, j)) - E(log^+(I))$, where $\log^+(x) = \log(x + 1)$, and $E(f(I))$ denotes the arithmetic mean of $f(I(i, j))$ when (i, j) ranges over all pixel coordinates in the image.[2] Cases where log-contrast was used will be explicitly identified in the sequel.

We experimented with a subset, which will be denoted $\mathcal{I}$, of 1400 images from the Van Hateren's data base. From this set of images we generated Jsteg and MHPDM stego images by first reducing the number of gray levels to a maximum of 256 (scaling by $255/max(I)$ and rounding) and then

[1] http://hlab.phys.rug.nl/imlib/index.html

[2] $\log^+$ is used to avoid problems with the logarithm of zero. The slight effect of this bias on eliminating the constant multiplier of the light intensity is secondary for the cases of interest.

compressing with JPEG and embedding a random message in JPEG DCT coefficients during the process.[3] For MHPDM in particular, the message satisfied $P(b_i = 1) = \hat{P}(lsb(x_{i,j} = 1))$ for every coefficient index $j = 1..20$. The amount of information embedded was always the maximum allowed by the image, i.e., a message as long, in bits, as the number of coefficient values suitable for modification according to the stego algorithm. When we used S-Tools to generate stego data, we also started from a 256 gray level version of the original image and adjusted the length of the embedded message to avoid visually perceptible artifacts. The amount of information embedded in an image in this case was significantly smaller than for the Jsteg or MHPDM counterparts. The S-tools images were always converted back to gray level images before computing statistics, by working on the image formed by the rounded luminance.

Since JPEG lossy compression may affect image statistics, when analyzing results for Jsteg and MHPDM we always compare stego images to clean *JPEG images*, i.e., images with no message embedded but that have been lossily compressed with JPEG (again reducing the number of gray levels to a maximum of 256 and using the same software and settings as for Jsteg and MHPDM). Similarly, we use the term *bitmap image* to refer to an image with no information embedded but whose number of gray levels has been reduced to a maximum of 256.

Some experiments rely on estimations of mean (μ), standard deviation (σ), skewness (γ_1) and kurtosis (β_2) of a random variable X based on an observed sample $x_1..x_n$. The skewness and kurtosis of X are defined (see e.g. [20]) as

$$\gamma_1 = \frac{E(x-\mu)^3}{\sigma^3}); \quad \beta_2 = \frac{E(x-\mu)^4}{\sigma^4}$$

We use estimators respectively calculated as

$$\bar{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}; \; \bar{\sigma} = (\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \bar{\mu})^2)^{1/2}; \; \bar{\gamma_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\mu})^3}{\bar{\sigma}^3}; \; \bar{\beta_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\mu})^4}{\bar{\sigma}^4},$$

where x_i ranges over all data values of interest.

4.2 Experiments

We now describe several experiments involving the different stego algorithms and natural image statistical models described above. We also present some additional experiments targeting MHPDM stego-analysis in particular. In this case, we include also an analysis of wavelet and DCT coefficients.

4.2.1 Areas Model Parameters

We explore the effect of stego algorithms on the values (α, C) of the Areas Model parameters. We observe that the power law holds in bitmap, JPEG, and stego images and, although the parameter values are often modified for individual images, they generally remain in the (relatively large) range of values observed for natural images. Thus, the variation does not allow us to clearly distinguish between natural and stego images. Moreover, there is not a clear bias effect, meaning that in contrast with other models (see below), this characterization of natural images is mostly "randomly" modified by the stego process. Figure 1 shows the distribution of connected components areas of a particular image from $\mathcal{I}$ as bitmap, JPEG, and covering a message embedded with Jsteg and S-Tools. We observe that the plots are very close and the values of the exponent α for the best linear fitting in each case are $-2.09, -2.06$, -2.05, and -2.06, respectively. Figure 2 shows,

[3] Jsteg and our implementation of MHPDM are both based on source code from the Independent JPEG Group's JPEG software, http://www.ijg.org/. We set the parameter *quality setting* to 75%.

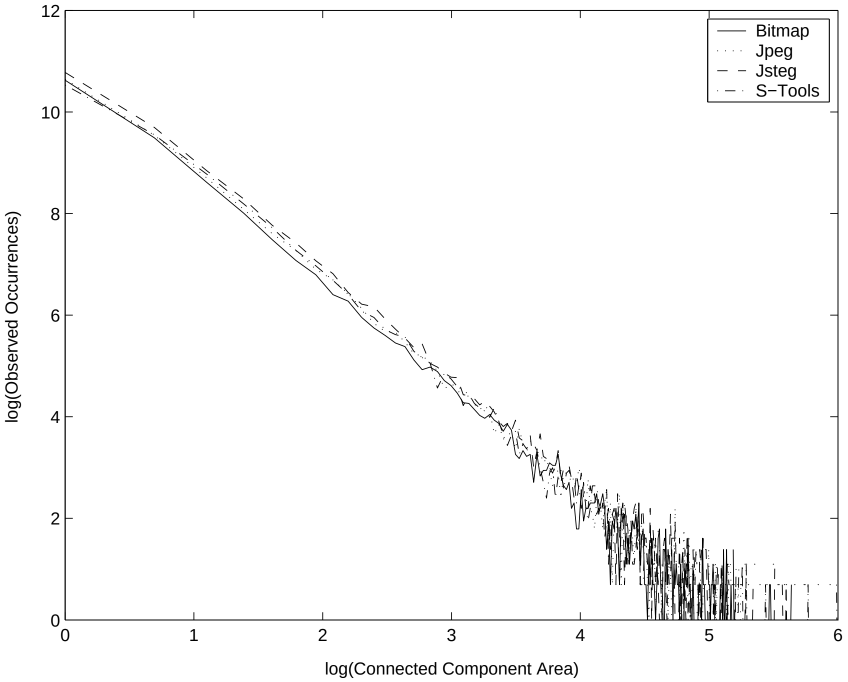

Figure 1: *Distribution of connected components areas for four versions of the same image, with and without hidden message. We observe that the exponential distribution is observed both by the original and the stego images, thereby limiting the use of this model for stego-analysis.*

enclosed in a rectangular frame, parameters values for a particular image from $\mathcal{I}$ as bitmap, JPEG, and covering a message embedded with MHPDM, Jsteg and S-Tools, together with a cloud of points obtained plotting parameters values for a subset $\mathcal{I}_{1000}$ of 1000 JPEG images from $\mathcal{I}$. The variation resulting from embedding a message is rather small as compared to the universe of observed values.

4.2.2 PC Model Parameters

This model has been found to fit accurately the distribution of differences between adjacent pixels. The model would be appropriate for classification if the observed fit deteriorated for stego images. Figure 3 shows the model fit for a given JPEG image, and the same image including a message embedded with MHPDM, Jsteg and S-Tools, respectively. As observed in the figure, Jsteg and MHPDM do not produce a noticeable departure from the model. However, the algorithm from S-tools does, and an image bearing a message embedded using this algorithm can easily be detected by observing the histogram of differences between adjacent pixels and its discrepancy with the model.

Figure 4 shows parameters values for the same four variations of the same image as Figure 3, and also the bitmap representation, immersed in a cloud of points obtained for parameters values of the subset $\mathcal{I}_{1000}$ of 1000 JPEG images from $\mathcal{I}$. Except for the values obtained for S-Tools, the rest, enclosed in a rectangular frame, show small differences as compared to the range of different values observed on JPEG images.

A closer examination of the effect on the whole data set $\mathcal{I}_{1000}$ reveals that the parameter p is altered in a consistent direction by the MHPDM algorithm, i.e., in more than 95% cases of 1000

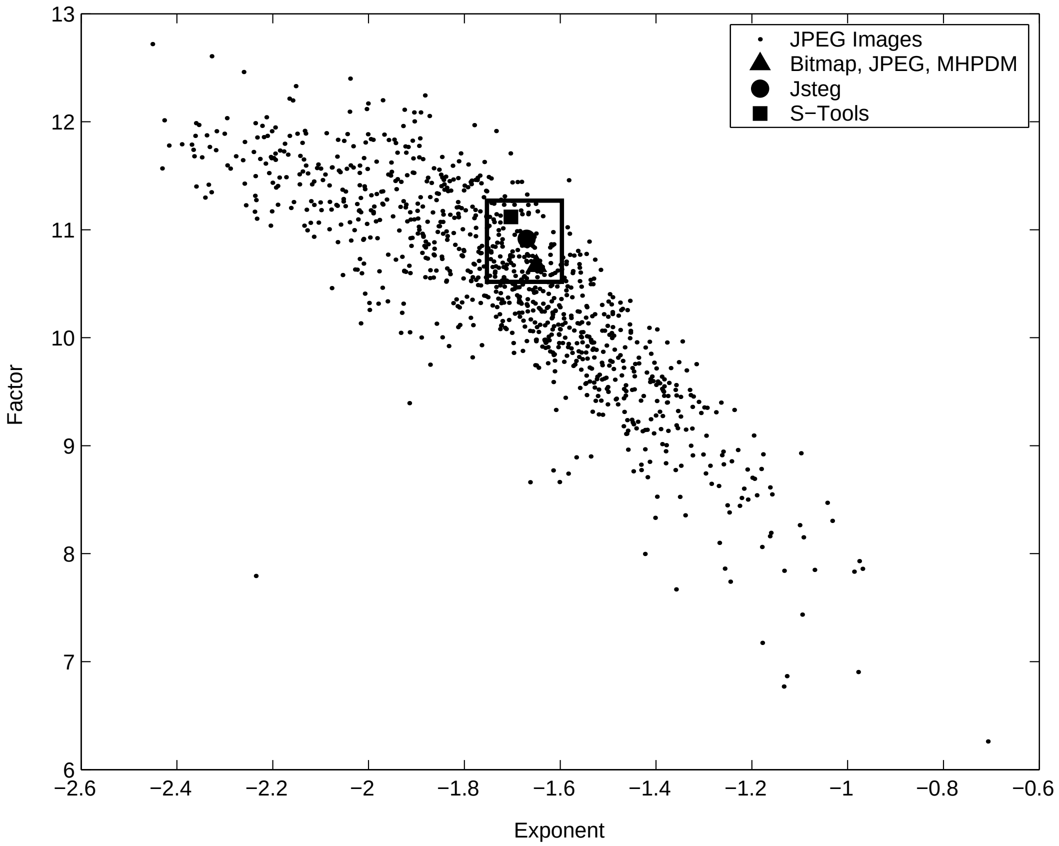

Figure 2: *Cloud of Areas Model parameters values for JPEG images and the effect of hiding information on one particular image. Note that the variation due to the hiding process is rather small compared to the intrinsic variability of the parameters for this class of natural images.*

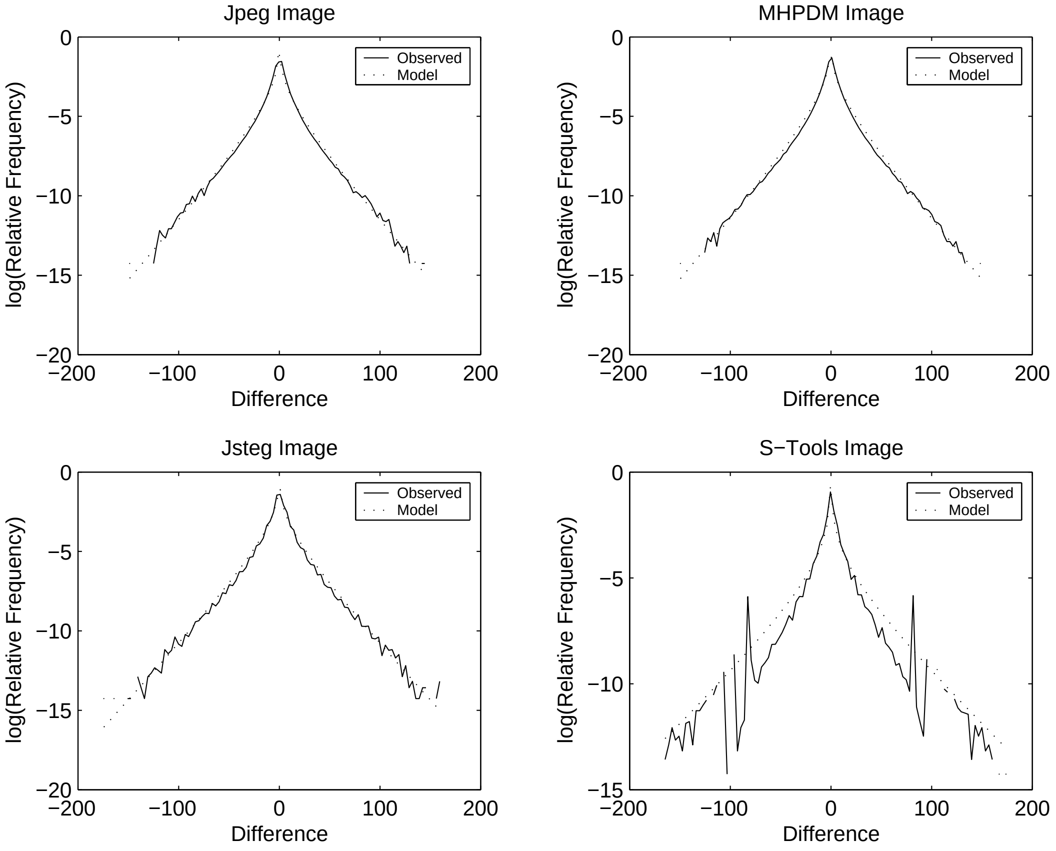

Figure 3: *PC Model fit to four versions of the same image. When the message is embedded using the S-Tools algorithm, this can be easily detected due to its discrepancy with the model. For MHPDM and Jsteg stego algorithms, there is a good to the natural images model.*

pairs of MHPDM / JPEG images from $\mathcal{I}_{1000}$, the stego algorithm causes an increase in the value of p. A histogram of relative differences of parameter p (the difference divided by the value of p for the JPEG image) is shown in Figure 5 where we observe that practically all values are positive. This consistent bias indicates a potential weakness of MHPDM with respect to stego-analysis based on this model. However, the shift is not large enough to achieve significant classification performance for this class of images, as can be appreciated in Figure 5, showing relative differences smaller than 5% in most cases, and Figure 6, showing very similar histograms of both parameters for 1000 JPEG images and 1000 MHPDM stego images from $\mathcal{I}_{1000}$.

4.2.3 Differentially Laplacian Model

For the Differentially Laplacian Model experiments we select k^2-1 coefficients pseudo-randomly with a uniform distribution in the interval (-1,1) and choose one more coefficient so that the overall coefficient sum is zero. As previously observed for the PC Model, the fit of the Differentially Laplacian Model does not deteriorate significantly when hidden data is embedded. This was the case observed for several values of parameter k and different images.

Also, for a fixed linear combination T, if we denote by $T(I) = \{T(block_{i,j}(I))\}$ where $block_{i,j}(I)$

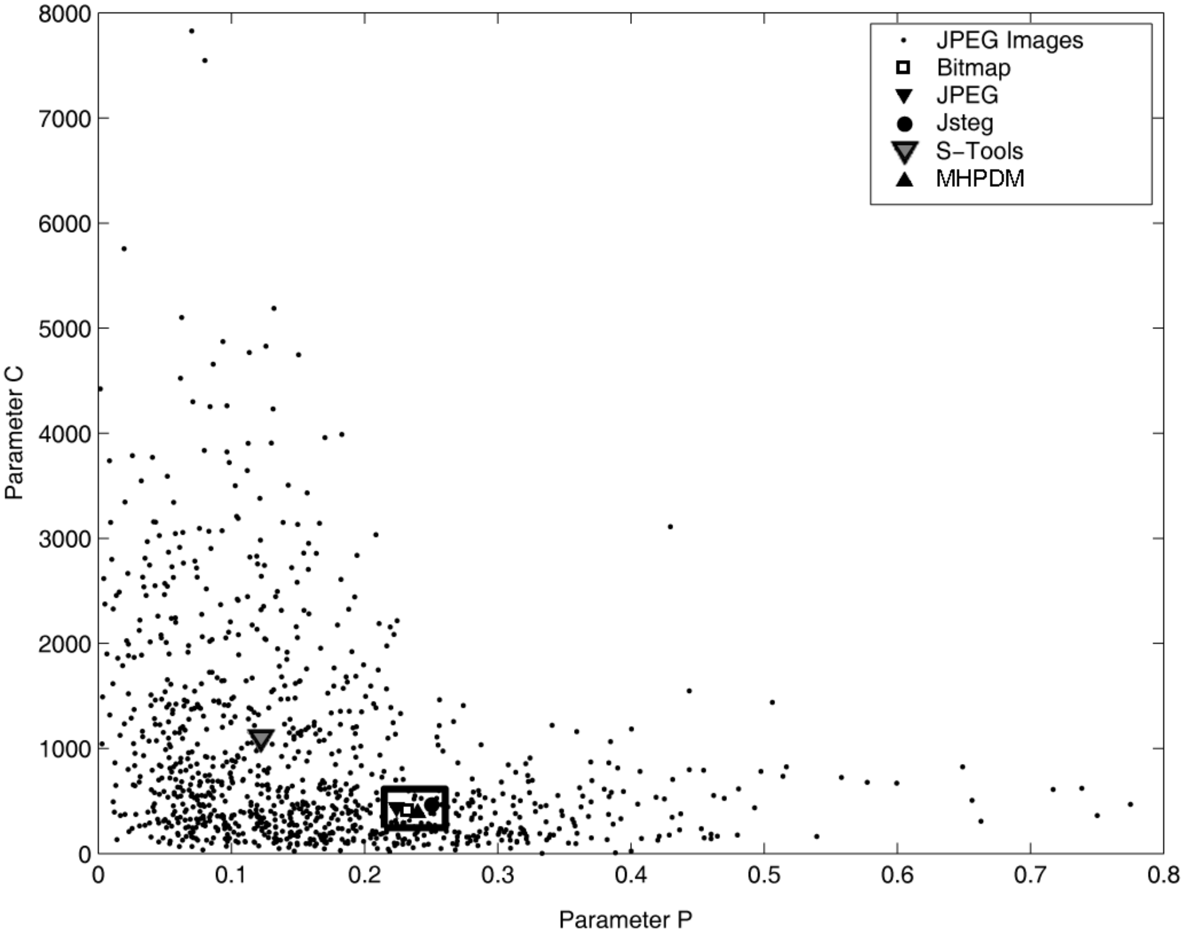

Figure 4: *Cloud of PC Model parameters values for JPEG images and the effect of hiding information on one particular image. Enclosed in a frame are values for bitmap, JPEG, Jsteg and MHPDM versions of the same image. The value for the same image processed with S-Tools, outside the frame, clearly shows that S-Tools produces easy to detect non-natural images (following this model), while MHPDM and Jsteg do produce what are considered legitimate natural images.*

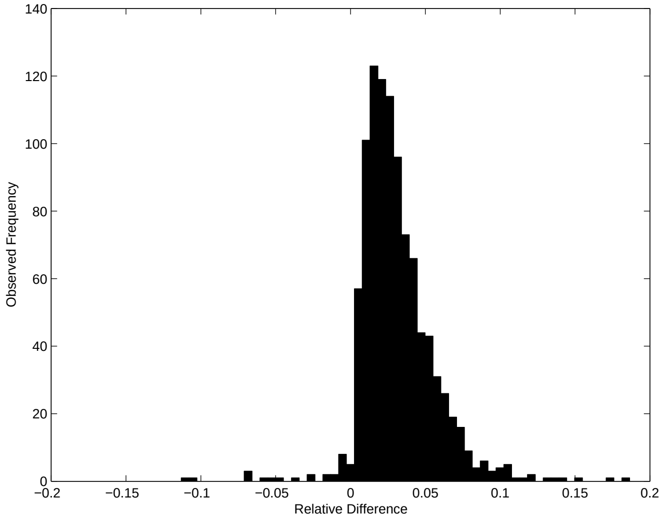

Figure 5: *Histogram of relative differences between parameter p of PC Model for a MHPDM image and its corresponding JPEG image. We observe that values of relative differences (difference divided by the value of p for JPEG image) are mostly smaller than 5% and practically all values are positive.*

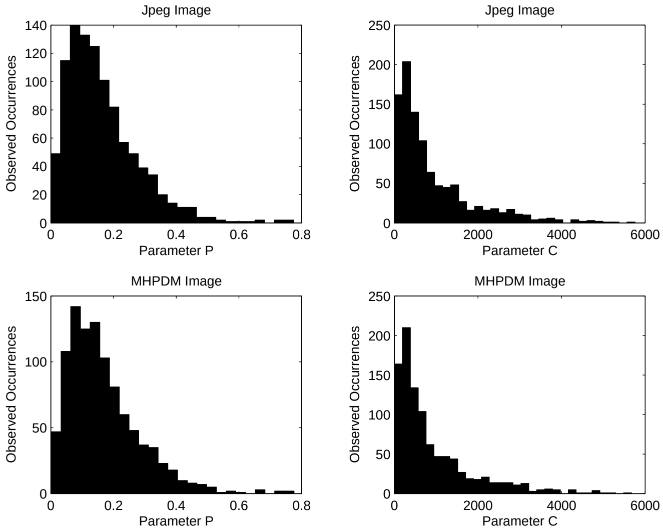

Figure 6: *Histograms of PC Model parameters for JPEG and MHPDM images. We observe that the parameters distributions are similar for natural and stego images.*

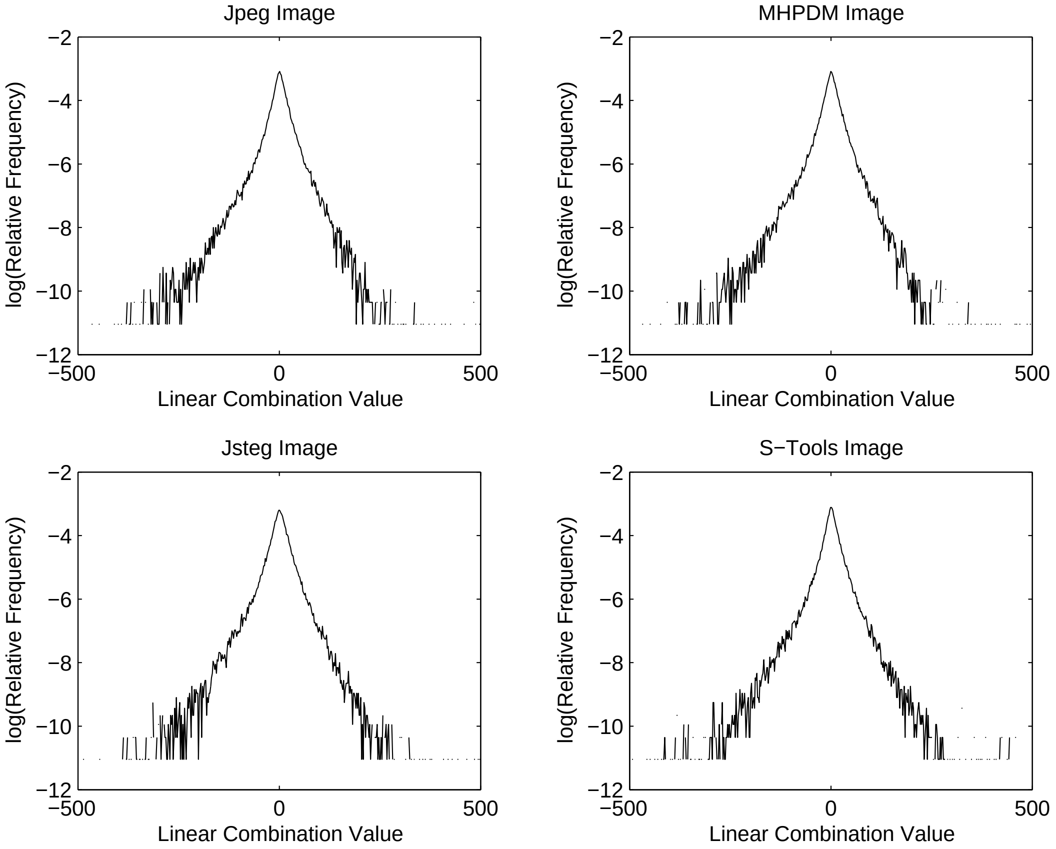

Figure 7: *Parameter distribution for the Differentially Laplacian Model for the four classes of images. We observe very similar distributions, indicating that the model is not powerful enough to detect stego images.*

varies along all $k \times k$ blocks of a partition of I, estimations of mean, standard deviation and kurtosis of $T(I)$ do not aid in the classification process.

Figure 7 shows a log normalized histogram of the values calculated for a fixed 5×5 linear combination of pixels values, on a JPEG image and the same image including a message embedded with MHPDM, Jsteg, and S-Tools. The four plots are very similar.

4.2.4 Statistics of Wavelet Coefficients

In this subsection, we consider the analysis of statistics on wavelet-transform coefficients of images. In particular, we considered, as features for classification, estimations of mean, standard deviation, skewness, and kurtosis of several statistics calculated from Haar wavelet coefficients on log-contrast images. The investigation focused on the MHPDM algorithm. We experimented with differences and sums of pairs of coefficients taken from horizontal, vertical and diagonal wavelet bands. In particular, denoting by $h_{i,j}$ a coefficient in the horizontal band of the first level decomposition of a $N \times M$ image, we found that the estimated kurtosis of $h_{i,j+1} - h_{i,j}$ with $0 \leq i < N/2$, $0 < j < M/2$, is consistently altered for stego images. Stego images showed a higher kurtosis than their corresponding JPEG images in more than 95% cases of the set $\mathcal{I}_{1000}$ of 1000 pairs of images

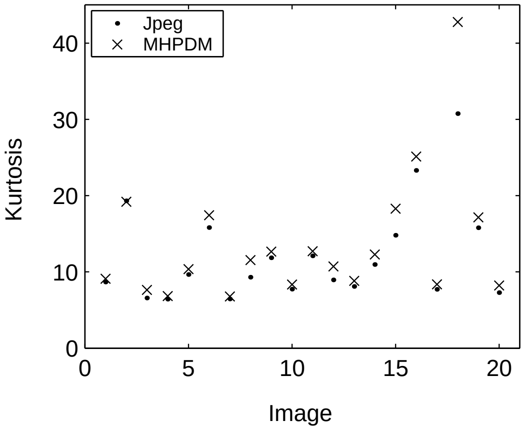

Figure 8: *Estimated kurtosis of differences of adjacent horizontal Haar wavelet coefficients. We observe a clear bias in the stego image.*

from $\mathcal{I}$. However, the kurtosis variability in this class of natural images is once again quite large, and it seems difficult to fix a threshold that could reliably discriminate between the two groups.

Figure 8 shows the estimated kurtosis of $h_{i,j+1} - h_{i,j}$ for 20 JPEG and MHPDM stego images from $\mathcal{I}$. Crosses representing kurtosis of stego images appear always above dots corresponding to JPEG images, but it seems difficult to choose a threshold that would separate the two series of values precisely. Nevertheless, the consistent bias shows that MHPDM images deviate from the natural images class, and can be regarded as a weakness of the method.

4.2.5 Comparing DCT Coefficients

An additional area explored for MHPDM was the information from higher order joint statistics of DCT coefficients. The fact that the histogram of each coefficient is preserved separately by the MHPDM algorithm opens the possibility that some joint distribution might be altered, thus aiding in stego-analysis. We consider the collection of 64-dimensional vectors obtained by applying the DCT on 8×8 blocks of a log-contrast image, and taking the absolute value of the resulting transform coefficients. We look at absolute values of each 8×8 DCT block as a vector in $\mathbb{R}^{64}$. Each image of size $N \times M$ brings $\frac{N}{8}\frac{M}{8}$ sample vectors. Given a JPEG image and the same image with a message embedded with MHPDM, let $J = \{j_i\}, S = \{s_i\}$, $0 \leq i < \frac{N}{8}\frac{M}{8}$, be the sample vectors in $\mathbb{R}^{64}$ obtained from each image respectively. We compute a vector $w \in \mathbb{R}^{64}$ that maximizes the empirical correlation $w = argmax\{\hat{\rho}(w.v', I_v)\}$, where v is a sample taken from J or S and I_v is valued 1 or -1 when v is a taken from J or S respectively. Averaging uniformly vectors w computed for several pairs of training images, we seek assigning a high weight to DCT coefficients that aid classification for many images whereas others would receive low weights. Once the average projection vector $W = mean(w)$ is determined, classification of an image I consists in calculating the arithmetic average $mean\{W.v_i'\}$, where $v_i \in \mathbb{R}^{64}$ ranges over vectors of absolute values of DCT coefficients of I, and finally using a threshold for the decision that must be fixed according to a trade off between false alarms and hit probabilities (i.e. respectively the probability of incorrectly classifying a natural image as stego and the probability of correctly classifying a stego image as such). The averaged projections were consistently higher for stego images than their corresponding JPEG images in more than 99% cases of a subset $\mathcal{I}_{test}$ of 1000 pairs of test images from $\mathcal{I}$ with

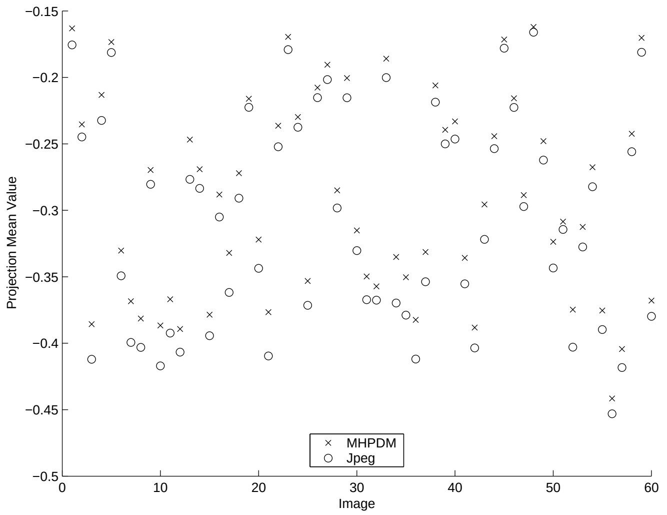

Figure 9: *Projection of DCT coefficients. Although the values are similar for the natural and stego images, there is a clear bias, being the value for the stego image always grater than the one for the corresponding natural one.*

a training subset $\mathcal{I}_{train}$ of 400 pairs of images from $\mathcal{I}$. Results for a subset of the testing set are shown in Figure 9 where crosses representing values for stego images appear always above circles representing JPEG images. This bias once again indicates a clear modification by MHPDM of the statistics of natural images. However, once again the variability for this large class of natural images is significant and it is impossible to fix a threshold that would work well for most pairs at the same time.

4.2.6 Coefficients Correlations based on Model Parameters

As expected, the MHPDM algorithm reduces the correlation between pairs of JPEG coefficients. However, these correlations vary considerable among natural images, thus they can not be used for classification unless they are tied to some other variables not affected by the steganography algorithm. One alternative is to exploit a possible correlation between Areas Model or PC Model parameters and the correlation between a pair of coefficients. To explore this idea, using canonical correlation analysis [21] on the set $\mathcal{I}_{1000}$ of 1000 JPEG images, we computed two projections that maximize empirical correlation between projected values: One from the four dimensional space of models parameters[4] and the other from the space of all empirical correlations between absolute values of pairs of coefficients altered by MHPDM. The empirical correlation obtained is high (> 0.9) yet the empirical correlation between the projected values and a variable indicating whether an image is natural or stego is practically zero.

[4] We recall that the parameters are exponent α and scaling factor C in Areas Model and shape parameter p and scale parameter c in PC Model

4.2.7 Coefficients Correlations Estimation: Exploiting algorithm knowledge in stego-analysis

Empirical correlations of DCT coefficients vary considerably among natural images. However, it is also possible to look at empirical correlations between empirical correlations for different pairs of coefficients. That is, images that have high correlation between coefficients, say a and b, might also have high correlation between a different pair of carefully chosen coefficients a' and b', with high probability. This fact can be exploited particularly for the MHPDM algorithm if we consider that only coefficients with indices 1 through 20 are modified. Based on a set of log-contrast training images, for each pair of absolute values of DCT coefficients a and b in $A = \{1..20\}$, we get an estimation $\hat{\hat{\rho}}(|a|,|b|)$ of $\hat{\rho}(|a|,|b|)$ (the empirical correlation between $|a|$ and $|b|$) based on the empirical correlations between pairs of absolute values of DCT coefficients taken from the set $B = \{0, 21..63\}$ and use $\hat{\hat{\rho}}(|a|,|b|) - \hat{\rho}(|a|,|b|)$ as a feature for classification. To calculate $\hat{\hat{\rho}}(|a|,|b|)$, we determine the projection from the vector v of values $v_i = \hat{\rho}(|a'|,|b'|)$ to a one-dimensional space that maximizes the empirical correlation with $\hat{\rho}(|a|,|b|)$. The set of pairs of coefficients (a', b') is the set of all possible pairs of coefficients from a subset $B' \subset B$, where highly quantized coefficients are discarded. Once this projection w is determined, we use a linear fitting from $w.v'$ to $\hat{\rho}(|a|,|b|)$ over the set of training images and use this polynomial to calculate $\hat{\hat{\rho}}(|a|,|b|)$. Having determined the estimator $\hat{\hat{\rho}}$ of $\hat{\rho}$ for all pairs $(|a|,|b|)$ we can calculate features $\hat{\hat{\rho}}(|a|,|b|)$ - $\hat{\rho}(|a|,|b|)$ for the set of training images and determine a projection from the space of features to a one dimensional space that maximizes the empirical correlation with a variable valued 1 for natural images and -1 for stego images. Classifying an image consist of comparing the projection of its features with a given threshold, which is chosen to determine an operating point in the "hit/false alarm" plane, as described below. This technique achieved the best classifications results. Figure 10 shows false alarm probability vs. hit probability for an experiment on a subset $\mathcal{I}_{train'}$ of 800 training pairs of JPEG/stego images from $\mathcal{I}$ and a subset $\mathcal{I}_{test'}$ of 600 test pairs from $\mathcal{I}$. The plot is obtained varying the classification threshold. Fixing its value is a trade off between these two probabilities, i.e., given the probability of correctly classifying an image as being stego (hit probability) there is an implicit probability of mistaking a natural image as stego (false alarm probability). Of course when designing a test the goal is for the plot to be as far apart as possible to the dotted line that represents simply selecting randomly with equal probabilities between the two classes (assuming they are equally probable). Thus, Figure 10 shows that the test described achieves significantly reliable detection of stego-images.

5 Conclusions

We have studied the effect of applying popular steganography algorithms on different statistical models of natural images. On one hand, we observed that some popular stego algorithms consistently bias these statistics for some of the most fundamental models. On the other hand, the intrinsic variability of these statistics is so high, for the class of images studied, that this bias induced by hiding "unnatural" information is not sufficient in general to move the results outside of the "natural" range, unless knowledge of the embedding algorithm is available and exploited. The best classification results were obtained in the latter case.

These experimental results lead us to conclusions in two directions. First, regarding faithful models of natural images, it seems that the reported efforts so far are not sufficient to clearly exclude some "non-natural" images, for example those obtained by artificially embedding hidden messages. Thus, there seems to be a need for further refinement of these models. In the stego arena,

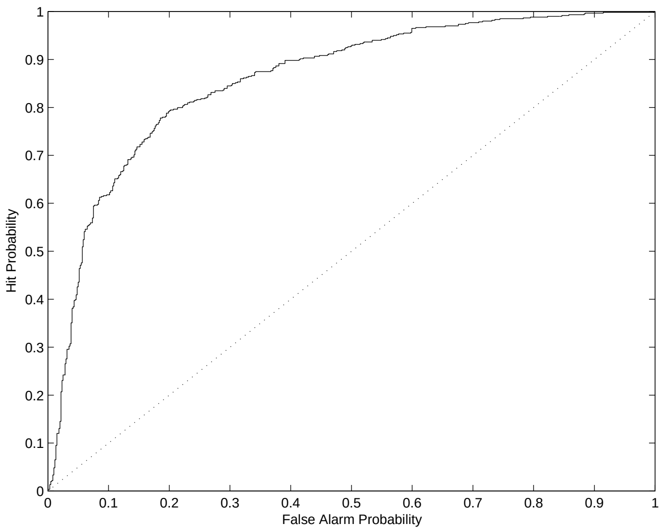

Figure 10: *Results using Coefficients Correlations Estimation. The graph shows that this technique, that uses information about the stego algorithm, can detect stego images with significant accuracy.*

it is first obvious that stego-analysis is a “cat and mouse” game: Knowing the stego algorithm, a technique can be devised to attack it; and knowing the attack, the stego algorithm can be further modified to mislead the detection procedure. An example is given by Farid’s stego-analysis approach [6, 7], which was overcome by MHPDM, which in turn, seems to be broken by the results in Section 4.2.7. It would therefore be desirable to have a more fundamental approach to the stego capacity in natural images, preferably based on universal properties and independent of the particular algorithm of choice. Some analysis has been done in this direction in [22, 23, 24, 25, 26]. An approach based on universal modeling and simulation [27, 28, 29, 30, 31] is currently being pursued. Results on this approach are reported elsewhere [32].

References

[1] A. Brown, “S-Tools for Windows..” Shareware, ftp://ftp.ntua.gr/pub/crypt/mirrors/idea.sec.dsi.unimi.it/code/s-tools4.zip, 1994.

[2] D. Upham, “JPEG-JSTEG - Modifications of the independent JPEG groups JPEG software for 1-bit steganography in JFIF output files.” ftp://ftp.funet.fi/pub/crypt/steganography/.

[3] N. F. Johnson, Z. Duric, and S. G. Jajodia, *Information Hiding : Steganography and Watermarking - Attacks and Countermeasures.* Kluwer Academic, Feb. 2001.

[4] P. Wayner, *Disappearing Cryptography, Second Edition - Information Hiding: Steganography and Watermarking.* Morgan Kaufmann, Apr. 2002.

[5] I. Avcibas, N. Memon, and B. Sankur, “Steganalysis using image quality metrics,” *IEEE Transactions on Image Processing*, vol. 12, pp. 221–229, Feb. 2003.

[6] H. Farid, “Detecting hidden messages using higher-order statistical models,” in *Proc. ICIP2002*, vol. 2, (Rochester, NY), pp. II–905–II–908, 2002.

[7] H. Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *Proc. of the 5th International Workshop of Information Hiding*, (Noordwijkerhout, The Netherlands), Oct. 2002.

[8] R. Tzschoppe, R. Bäuml, J. B. Huber, and A. Kaup, "Steganographic system based on higher-order statistics," in *Proc. SPIE Vol. 5020, Security and Watermarking of Multimedia Contents V*, (Santa Clara, CA), 2003.

[9] J. J. Eggers, R. Bäuml, and B. Girod, "A communications approach to image steganography," in *Proc. SPIE Vol. 4675, Security and Watermarking of Multimedia Contents IV*, (San Jose, CA), Jan. 2002.

[10] J. L. Mitchell and W. B. Pennebaker, *JPEG Still Image Data Compression Standard.* Van Nostrand Reinhold, 1993.

[11] Y. Gousseau and J. M. Morel, "Are natural images of bounded variation," *SIAM Journal on Mathematical Analysis*, vol. 33, no. 3, pp. 634–648, 2001.

[12] L. Alvarez, Y. Gousseau, and J. M. Morel, "The size of objects in natural images." CMLA preprint, Ecole Normale Sup. -Cachan, 1999.

[13] U. Grenander and A. Srivastava, "Probability models for clutter in natural images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 424–429, Apr. 2001.

[14] A. Srivastava, X. Liu, and U. Grenander, "Universal analytical forms for modeling image probabilities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1200–1214, Sept. 2002.

[15] U. Grenander, "Towards a theory of natural scenes," tech. rep., Brown University, Providence, RI, 2003. http://www.dam.brown.edu/ptg/publications.shtml.

[16] U. Grenander, M. I. Miller, and P. Tyagi, "Transported generator clutter models," monograph of center for imaging sciences, Johns Hopkins University, 1999.

[17] M. L. Green, "Statistics of images, the TV algorithm of Rudin-Osher-Fatemi for image denoising and an improved denoising algorithm," tech. rep., University of California, Los Angeles, CA, Oct. 2002. http://www.math.ucla.edu/ imagers/htmls/reports.html.

[18] A. Netravali and J. O. Limb, "Picture coding: A review," *Proc. IEEE*, vol. 68, pp. 366–406, 1980.

[19] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. Computer Vision and Pattern Recognition-Volume 1*, (Fort Collins, Colorado), June 1999.

[20] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th printing*, vol. 55 of *National Bureau of Standards Applied Mathematics*. Washington D.C.: U.S. Government Printing Office, Dec. 1972.

[21] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[22] C. Cachin, "An information-theoretic model for steganography," *Lecture Notes in Computer Science*, vol. 1525, pp. 306–318, 1998.

[23] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Transactions on Information Theory*, vol. 49, pp. 563–593, Mar. 2003.

[24] P. Moulin and M. K. Mihcak, "The parallel-gaussian watermarking game." To appear in *IEEE Transactions on Information Theory*, Feb. 2004, jun 2001.

[25] P. Moulin, M. K. Mihcak, and G.-I. Lin, "An information–theoretic model for image watermarking and data hiding," in *Proc. ICIP2000*, (Vancouver, B.C.), Sept. 2000.

[26] P. Moulin and M. K. Mihcak, "A framework for evaluating the data-hiding capacity of image sources," *IEEE Transactions on Image Processing*, vol. 11, pp. 1029–1042, Sept. 2002.

[27] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Transactions on Information Theory*, vol. 39, pp. 752–772, May 1993.

[28] T. S. Han and M. Hoshi, "Interval algorithm for random number generation," *IEEE Transactions on Information Theory*, vol. 43, pp. 599–611, Mar. 1997.

[29] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Transactions on Information Theory*, vol. 42, pp. 63–86, Jan. 1996.

[30] N. Merhav and M. J. Weinberger, "On universal simulation of information sources using training data," Tech. Rep. HPL-2002-263, Hewlett-Packard Laboratories, Sept. 2002.

[31] G. Seroussi, "Universal types and simulation of individual sequences." To appear in *Proc. LATIN'2004*, Buenos Aires, Argentina, April 2004.

[32] A. Martin, G. Seroussi, and G. Sapiro, "A universal approach to steganography." in preparation.