

Detecting Steganographic Messages in Digital Images

Hany Farid

Department of Computer Science
Dartmouth College
Hanover NH 03755

Techniques and applications for information hiding have become increasingly more sophisticated and widespread. With high-resolution digital images as carriers, detecting the presence of hidden messages has also become considerably more difficult. It is sometimes possible, nevertheless, to detect (but not necessarily decipher) the presence of embedded messages. The basic approach taken here works by finding predictable higher-order statistics of “natural” images within a multi-scale decomposition, and then showing that embedded messages alter these statistics.

1 Introduction

Information hiding techniques (e.g., steganography and watermarking) have recently received quite a bit of attention (see [9, 1, 7, 12] for general reviews). At least one reason for this is the desire to protect copyrights of digital audio, image and video. Other applications include unobtrusive military and intelligence communication, covert criminal communication, and the protection of civilian speech against repressive governments. Along with new and improved techniques for hiding information will come techniques for detecting (and possibly removing) such information.

Although messages embedded into an image are often imperceptible to the human eye, they often disturb the statistical nature of the image. Previous approaches to detecting such deviations [8, 23, 13, 15] examine first-order statistical distributions of intensity or transform coefficients (e.g., discrete cosine transform, DCT). In contrast, the approach to detection taken here relies on building higher-order statistical models for natural images [10, 5, 3, 17, 24, 11, 19] and looking for deviations from these models. Specifically, I show that, across a broad range of natural images, strong higher-order statistical regularities within a wavelet-like decomposition exist, and that when a message is embedded within an image, these statistics are significantly altered. The benefit of building a model based on higher-order statistics is that simple counter-measures that match first-order statistics are unlikely to entirely foil detection.

What follows is first a description of the image decomposition and statistical model, and then the classification scheme used to detect hidden messages. The efficacy of this approach is tested against messages hidden with Jsteg¹, EZStego², and OutGuess [13, 14].

¹Jsteg V4, by Derek Upham, is available at:
<ftp://ftp.funet.fi/pub/crypt/steganography>.

²EZStego, by Romana Machado, is available at
<http://www.stego.com/>.

2 Image Statistics

The decomposition of images using basis functions that are localized in spatial position, orientation, and scale (e.g., wavelets) has proven extremely useful in a range of applications (e.g., image compression, image coding, noise removal, and texture synthesis). One reason for this is that such decompositions exhibit statistical regularities that can be exploited (e.g., [18, 16, 2]). Described below is one such decomposition, and a set of statistics collected from this decomposition.

The decomposition is based on separable quadrature mirror filters (QMFs) [21, 22, 20] is employed. As illustrated in Figure 1, this decomposition splits the frequency space into multiple scales and orientations. This is accomplished by applying separable lowpass and highpass filters along the image axes generating a vertical, horizontal, diagonal and low-pass subband. For example, the horizontal subband is generated by convolving with the highpass filter in the horizontal direction and lowpass in the vertical direction, the diagonal band is generated by convolving with the highpass filter in both directions, etc. Subsequent scales are created by recursively filtering the lowpass subband. The vertical, horizontal, and diagonal subbands at scale $i = 1, \dots, n$ are denoted as $V_i(x, y)$, $H_i(x, y)$, and $D_i(x, y)$, respectively. Shown in Figure 2 is a three-level decomposition of a “disc” image.

Given this image decomposition, the statistical model is composed of, the mean, variance, skewness and kurtosis of the subband coefficients at each orientation and at scales $i = 1, \dots, n - 1$. These statistics characterize the basic coefficient distributions.

The second set of statistics collected are based on the errors in an optimal linear predictor of coefficient magnitude. As described in [2], the subband coefficients are correlated to their spatial, orientation and scale neighbors. For purposes of illustration, consider first a vertical band, $V_i(x, y)$, at scale i . A linear predictor for the magnitude of these coefficients in a

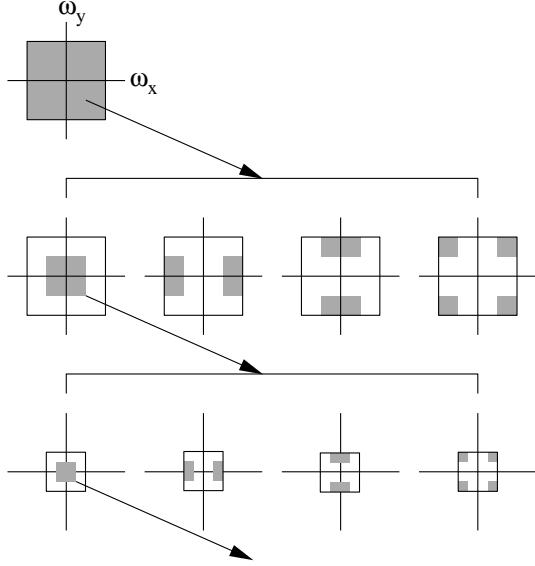


Figure 1: An idealized multi-scale and orientation decomposition of frequency space. Shown, from top to bottom, are levels 0, 1, and 2, and from left to right, are the lowpass, vertical, horizontal, and diagonal subbands.

subset of all possible neighbors³ is given by:

$$\begin{aligned} V_i(x, y) = & w_1 V_i(x - 1, y) + w_2 V_i(x + 1, y) \\ & + w_3 V_i(x, y - 1) + w_4 V_i(x, y + 1) \\ & + w_5 V_{i+1}(x/2, y/2) + w_6 D_i(x, y) \\ & + w_7 D_{i+1}(x/2, y/2), \end{aligned} \quad (1)$$

where w_k denotes scalar weighting values. This linear relationship may be expressed more compactly in matrix form as:

$$\vec{V} = Q\vec{w}, \quad (2)$$

where the column vector $\vec{w} = (w_1 \dots w_7)^T$, the vector \vec{V} contains the coefficient magnitudes of $V_i(x, y)$ strung out into a column vector, and the columns of the matrix Q contain the neighboring coefficient magnitudes as specified in Equation (1) also strung out into column vectors. The coefficients that minimize the squared error of the estimator is:

$$\vec{w} = (Q^T Q)^{-1} Q^T \vec{V}. \quad (3)$$

³The particular choice of spatial, orientation and scale neighbors was motivated by the observations of [2] and modified to include non-causal neighbors.

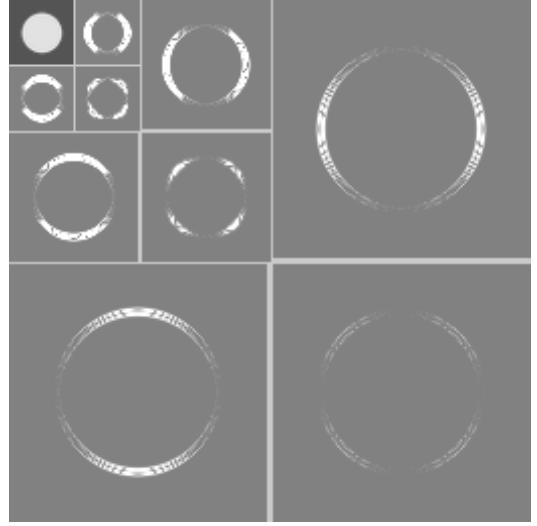


Figure 2: Shown are the absolute values of the subband coefficients at three scales and three orientations for a “disc” image. The residual lowpass subband is shown in the upper-left corner.

The log error in the linear predictor is then given by:

$$\vec{E}_v = \log_2(\vec{V}) - \log_2(|Q\vec{w}|). \quad (4)$$

It is from this error that additional statistics are collected, namely the mean, variance, skewness, and kurtosis. This process is repeated for each vertical subband at scales $i = 1, \dots, n-1$, where at each scale a new linear predictor is estimated. A similar process is repeated for the horizontal and diagonal subbands. The linear predictor for the horizontal subbands is of the form:

$$\begin{aligned} H_i(x, y) = & w_1 H_i(x - 1, y) + w_2 H_i(x + 1, y) \\ & + w_3 H_i(x, y - 1) + w_4 H_i(x, y + 1) \\ & + w_5 H_{i+1}(x/2, y/2) + w_6 D_i(x, y) \\ & + w_7 D_{i+1}(x/2, y/2), \end{aligned} \quad (5)$$

and for the diagonal subbands:

$$\begin{aligned} D_i(x, y) = & w_1 D_i(x - 1, y) + w_2 D_i(x + 1, y) \\ & + w_3 D_i(x, y - 1) + w_4 D_i(x, y + 1) \\ & + w_5 D_{i+1}(x/2, y/2) + w_6 H_i(x, y) \\ & + w_7 V_i(x, y). \end{aligned} \quad (6)$$

The same error metric, Equation (4), and error statistics computed for the vertical subbands, are computed for the horizontal and diagonal bands, for a total of $12(n - 1)$ error statistics. Combining these statistics with the $12(n - 1)$ coefficient statistics yields a total of $24(n - 1)$ statistics that form a “feature” vector which is used to discriminate between images that contain hidden messages and those that do not.

3 Classification

From the measured statistics of a training set of images with and without hidden messages, the goal is to determine whether a novel (test) image contains a message. To this end, Fisher linear discriminant analysis (FLD) [6, 4], a class specific method for pattern recognition, is employed. For simplicity a two-class FLD is described.

Denote column vectors \vec{x}_i , $i = 1, \dots, N_x$ and \vec{y}_j , $j = 1, \dots, N_y$ as exemplars from each of two classes from the training set. The within-class means are defined as:

$$\vec{\mu}_x = \frac{1}{N_x} \sum_{i=1}^{N_x} \vec{x}_i, \quad \text{and} \quad \vec{\mu}_y = \frac{1}{N_y} \sum_{j=1}^{N_y} \vec{y}_j. \quad (7)$$

The between-class mean is defined as:

$$\vec{\mu} = \frac{1}{N_x + N_y} \left(\sum_{i=1}^{N_x} \vec{x}_i + \sum_{j=1}^{N_y} \vec{y}_j \right) \quad (8)$$

The within-class scatter matrix is defined as:

$$S_w = M_x M_x^T + M_y M_y^T, \quad (9)$$

where, the i^{th} column of matrix M_x contains the zero-meaned i^{th} exemplar given by $\vec{x}_i - \vec{\mu}_x$. Similarly, the j^{th} column of matrix M_y contains $\vec{y}_j - \vec{\mu}_y$. The between-class scatter matrix is defined as:

$$S_b = N_x(\vec{\mu}_x - \vec{\mu})(\vec{\mu}_x - \vec{\mu})^T + N_y(\vec{\mu}_y - \vec{\mu})(\vec{\mu}_y - \vec{\mu})^T. \quad (10)$$

Finally, let \vec{e} be the maximal generalized eigenvalue-eigenvector of S_b and S_w (i.e., $S_b \vec{e} = \lambda S_w \vec{e}$).

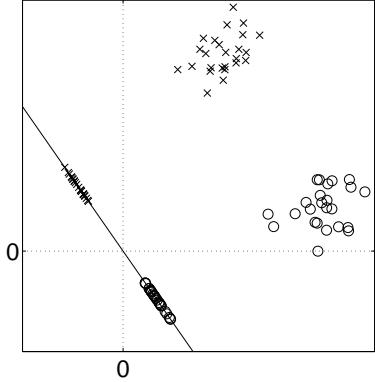


Figure 3: Shown are training exemplars from one of two classes ('x's and 'o's) in an initially two-dimensional space. These exemplars are projected onto the FLD projection axis (solid line) so as to minimize within-class scatter and maximize between-class scatter. Novel exemplars projected onto the same axis can be, in this example, categorized depending on which side of the origin they project.

When the training exemplars \vec{x}_i and \vec{y}_j are projected onto the one-dimensional linear subspace defined by \vec{e} (i.e., $\vec{x}_i^T \vec{e}$ and $\vec{y}_j^T \vec{e}$), the within-class scatter is minimized and the between-class scatter maximized, Figure 3. For the purposes of pattern recognition, such a projection is clearly desirable as it simultaneously reduces the dimensionality of the data and preserves discriminability.

Once the FLD projection axis is determined from the training set, a novel exemplar, \vec{z} , from the testing set is classified by first projecting onto the same subspace, $\vec{z}^T \vec{e}$. In the simplest case, the class to which this exemplar belongs is determined via a simple threshold, Figure 3. In the case of a two-class FLD, we are guaranteed to be able to project onto a one-dimensional subspace (i.e., there will be at most one non-zero eigenvalue). In the case of a N -class FLD, the projection may be onto as high as a $N - 1$ -dimensional subspace.

A two-class FLD is employed here to classify images as either containing or not containing a hidden message. Each image is characterized by its feature vector as described in

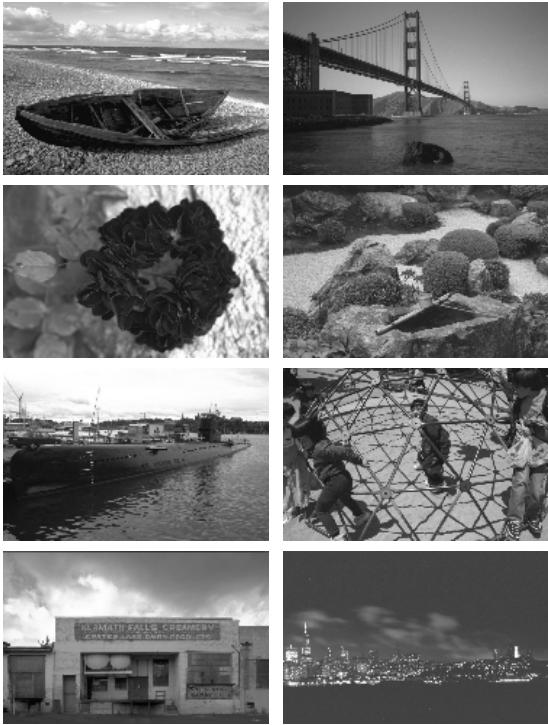


Figure 4: Sample images.

the previous section.

4 Results

Shown in Figure 4 are several examples taken from a database of natural images⁴. Each 8-bit per channel RGB image is 1000×1400 pixels in size. Statistics from 500 such images are collected as follows. Each image is compressed using Jsteg (with no steg message) to an average size of 250 Kb (quality 75). Each image is then converted from RGB to 8-bit gray-scale ($\text{gray} = 0.299R + 0.587G + 0.114B$). A four-level, three-orientation QMF pyramid is constructed for each image, from which a 72-length feature vector of coefficient and error statistics is collected. To reduce sensitivity to noise, only coefficient magnitudes greater than 1.0 are considered.

For the same set of 500 images, a message is

⁴Images were downloaded from: <http://philip.greenspun.com> and reproduced here with permission from Philip Greenspun.

embedded using Jsteg, a transform-based system that embeds messages by modulating the DCT coefficients [8]. Each message consists of a central 256×256 portion of a random image chosen from the same image database. After the message image is embedded into the full resolution cover image, the same transformation, decomposition, and collection of statistics as described above are performed on the “steg” image.

Shown in Figure 5 is an example cover and message image, and the result of embedding the message into the cover image. In this example, the mean of the absolute value of the difference between the cover and steg image is 3.1 intensity values with a standard deviation of 3.2. For display purposes the difference image is renormalized into the range $[0, 255]$.

The two-class FLD described in Section 3 is trained on a random subset of 400 images and then tested on the remaining 100 images. Shown in Figure 6 are results for the training and testing set. In this figure the ‘x’ corresponds to “steg” images and the ‘o’ corresponds to the “no-steg” images. The vertical axis corresponds to the value of an image feature vector after projecting onto the FLD projection axis. Results from the training set are shown to the left of the vertical line, and results from the testing set are shown to the right. In this example, 99% of the training set is correctly classified. In the testing set 98% of the steg images are correctly classified with 2% false positives (i.e., a no-steg image incorrectly classified as a steg image). The threshold for classification (horizontal line) is selected using the ROC curves shown in the lower panel of Figure 6. In this panel, the solid line corresponds to the percent of correctly classified steg images, and the dashed line corresponds to the percent of correctly classified no-steg images. The classification threshold is selected to be the point at which these curves cross (-0.73 in this example). A 0% false positive rate can be achieved by selecting a more conservative threshold, which in this example, would yield a detection rate of 97%.

Shown in Table 1 are the results averaged

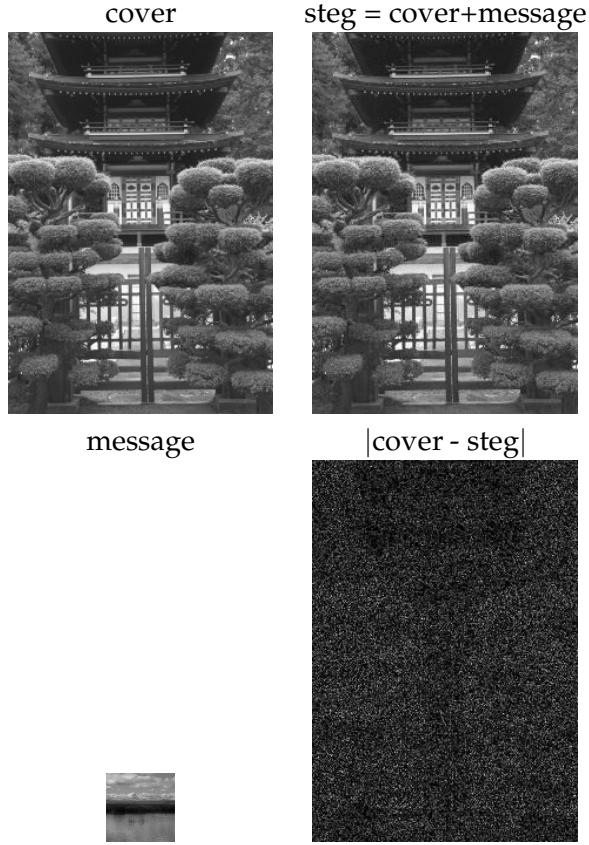


Figure 5: Shown is a cover image and a steg image containing an embedded message. Also shown is the the 256×256 message (at scale), and the absolute value of the difference between the cover and steg image (renormalized into the range [0,255] for display purposes).

across two hundred independent trials, where on each trial a random subset of 400 images are used for training, and the remaining 100 images for testing. The reported values, from the testing set, correspond to the accuracy of correctly classifying a steg image, and of incorrectly classifying no-steg images. Also shown in this table is the detection accuracy for a fixed false positive rate of 1% and 2%, and for message sizes ranging from 128×128 to 16×16 . As the message size decreases, detection rates fall correspondingly.

A second program, EzStego, was also tested. EzStego embeds information into GIF format images by modulating the least significant bits

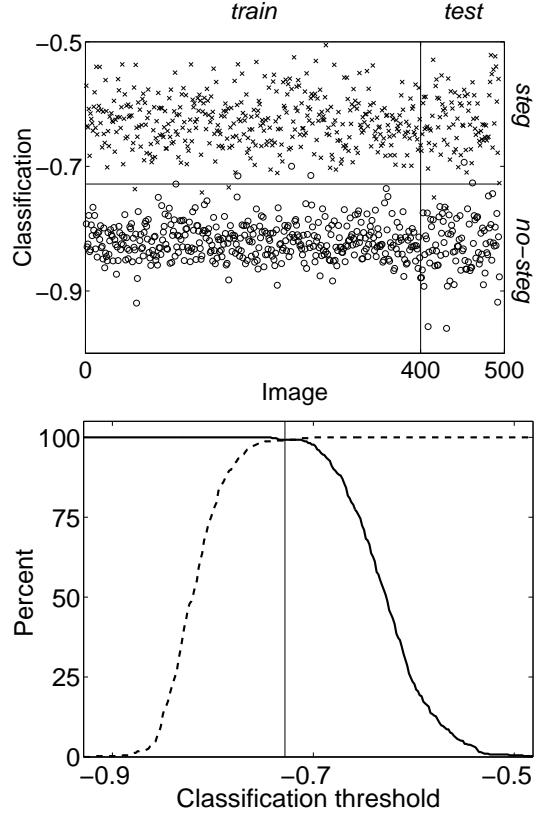


Figure 6: A perfect classifier would have placed all 'x's (steg images) above, and all 'o's (no-steg images) below, the horizontal line. The position of the horizontal line is determined from the ROC curves shown in the lower panel, where the solid and dashed lines correspond to the percent of correctly classified steg and no-steg images, respectively.

of the sorted color palette index. Statistics from the same set of 500 images are collected as follows. Each image is converted from its original RGB format to an 8-bit GIF format. As described above, each image is then converted to gray-scale, upon which the same multi-scale decomposition and feature vector computation is performed. For this set of 500 images, a message is embedded, where each message consists of a central 256×256 portion of a random image. Because EzStego embeds messages in scan-line order, the image is cropped to contain only the first 256 rows (i.e., 256×1400 in size). A simple counter-measure that

Embedding	Detection (%)	False positive (%)
Jsteg	97.8	1.8
Jsteg	95.0	1.0
Jsteg (128×128)	95.4	1.0
Jsteg (64×64)	88.5	1.0
Jsteg (32×32)	38.3	2.0
Jsteg (32×32)	26.6	1.0
Jsteg (16×16)	4.8	2.0
Jsteg (16×16)	3.0	1.0
EzStego	86.6	13.2
EzStego	62.7	2.0
EzStego	51.5	1.0
OutGuess ⁻	80.4	19.4
OutGuess ⁻	29.9	2.0
OutGuess ⁻	22.5	1.0
OutGuess ⁺	77.7	23.8
OutGuess ⁺	22.9	2.0
OutGuess ⁺	17.5	1.0

Table 1: Classification accuracy from three different steg programs. Unless noted, messages are 256×256 in size.

places the message at the middle or bottom of the image could easily be countered by analyzing all non-overlapping regions of the image. On average, 86.6% of the testing set is correctly classified, with 13.2% false positives, and 51.5% detection can be achieved with 1% false positives, Table 1.

A third program, OutGuess [13], was also tested. Similar to Jsteg, OutGuess embeds messages into JPEG format images by modulating the DCT coefficients (although it is not fundamentally limited to one image format). Unique to OutGuess, however, is a technique for first embedding only into one-half of the redundant bits and then using the remaining redundant bits to preserve the distribution of the DCT coefficients. This embedding is therefore not vulnerable to detection schemes based on an analysis of coefficient distributions only. As before, statistics for the same set of 500 images are first collected by converting each image from RGB to gray-scale, applying the multi-scale decomposition and extracting the feature

vector. A random 256×256 message is then embedded, followed by the same transformation, decomposition, and collection of statistics. OutGuess was tested⁵ with (+) and without (-) statistical correction, Table 1. Without statistical correction, detection rates are 80.4% with 19.4% false positives, or 22.5% detection with 2% false positives. With statistical correction, detection rates are slightly worse at 77.7% with 23.8% false positives, or 17.5% detection with 1.0% false positives.

While not all steg programs have been tested, it is likely that they will fall within the range of easily detectable (Jsteg) to less than twenty percent detectable (OutGuess). How effective these detection rates are depends, of course, on the specific applications.

5 Discussion

Messages can be embedded into digital images in ways that are imperceptible to the human eye, and yet, these manipulations can fundamentally alter the underlying statistics of an image. To detect the presence of hidden messages a model based on statistics taken from a multi-scale decomposition has been employed. This model includes basic coefficient statistics as well as error statistics from an optimal linear predictor of coefficient magnitude. These higher-order statistics appear to capture certain properties of “natural” images, and more importantly, these statistics are significantly altered when a message is embedded within an image. As such, it is possible to detect, with a reasonable degree of accuracy, the presence of steganographic messages in digital images. To avoid detection, however, one need only embed a small enough message. In the examples shown here, the message was typically 5% the size of the cover image. As the message size decreases, detection will become in-

⁵OutGuess was run with unlimited iterations to find the best embedding. When run with statistical testing, OutGuess imposes limits on the message size, as such only 219 of the 500 images could be used as cover images. Only these 219 images were used in the results of Table 1.

creasingly more difficult and less reliable.

Although not tested here, it is likely that the presence of digital watermarks could also be detected. Since one of the goals of watermarking is robustness to attack and not necessarily concealment, watermarks typically alter the image in a more substantial way. As such, it is likely that the underlying statistics will be more significantly disrupted. Although only tested on images, there is no inherent reason why the approaches described here would not work for one-dimensional audio signals or video sequences.

There are several directions that should be explored in order to improve detection accuracy. The particular choice of statistics is somewhat ad hoc, as such it would be beneficial to optimize across a set of statistics that maximizes detection rates. The two-class FLD should be replaced with a multi-class FLD that simultaneously distinguishes between no-steg images and steg images generated from multiple programs. However convenient, FLD analysis is linear, and detection rates would almost certainly benefit from a more flexible non-linear classification scheme. Lastly, the indiscriminant comparison of image statistics across all images could be replaced with a class-based analysis, where, for example, indoor and outdoor scenes, or images with similar frequency content, are compared separately.

One benefit of the higher-order models employed here is that they are not as vulnerable to counter-attacks that match first-order statistical distributions of pixel intensity or transform coefficients. There is little doubt, however, that counter-measures will be developed that can foil the detection scheme outlined here. The development of such techniques will in turn lead to better detection schemes, and so on.

Acknowledgments

We are grateful for the support from a National Science Foundation CAREER Award (IIS-99-83806), a Department of Justice Grant (2000-

DT-CS-K001), and a departmental National Science Foundation Infrastructure Grant (EIA-98-02068).

References

- [1] R.J. Anderson and F.A.P. Petitcolas. On the limits of steganography. *IEEE Journal on Selected Areas in Communications*, 16(4):474–481, 1998.
- [2] R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.
- [3] J.G. Daugman. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transaction on Biomedical Engineering*, 36(1):107–114, 1989.
- [4] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.
- [5] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.
- [6] R. Fisher. The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
- [7] N. Johnson and S. Jajodia. Exploring steganography: seeing the unseen. *IEEE Computer*, pages 26–34, 1998.
- [8] N. Johnson and S. Jajodia. Steganalysis of images created using current steganography software. *Lecture notes in Computer Science*, pages 273–289, 1998.
- [9] D. Kahn. The history of steganography. In *Proceedings of Information Hiding, First International Workshop*, Cambridge, UK, 1996.

- [10] D. Kersten. Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12):2395–2400, 1987.
- [11] G. Krieger, C. Zetsche, and E. Barth. Higher-order statistics of natural images and their exploitation by operators selective to intrinsic dimensionality. In *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 147–151, Banff, Alta., Canada, 1997.
- [12] E.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn. Information hiding - a survey. *Proceedings of the IEEE*, 87(7):1062–1078, 1999.
- [13] N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, Washington, DC, 2001.
- [14] N. Provos. Probabilistic methods for improving information hiding. Technical Report CITI 01-1, University of Michigan, 2001.
- [15] N. Provos and P. Honeyman. Detecting steganographic content on the internet. Technical Report CITI 01-1a, University of Michigan, 2001.
- [16] R. Rinaldo and G. Calvagno. Image coding by block prediction of multiresolution submimages. *IEEE Transactions on Image Processing*, 4(7):909–920, 1995.
- [17] D.L. Ruderman and W. Bialek. Statistics of natural image: Scaling in the woods. *Phys. Rev. Letters*, 73(6):814–817, 1994.
- [18] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.
- [19] E.P. Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *Proceedings of the 44th Annual Meeting*, volume 3813, Denver, CO, USA, 1999.
- [20] E.P. Simoncelli and E.H. Adelson. *Subband image coding*, chapter Subband transforms, pages 143–192. Kluwer Academic Publishers, Norwell, MA, 1990.
- [21] P.P. Vaidyanathan. Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques. *IEEE ASSP Magazine*, pages 4–20, 1987.
- [22] M. Vetterli. A theory of multirate filter banks. *IEEE Transactions on ASSP*, 35(3):356–372, 1987.
- [23] A. Westfeld and A. Pfitzmann. Attacks on steganographic systems. In *Proceedings of Information Hiding, Third International Workshop*, Dresden, Germany, 1999.
- [24] S.C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (frame) - towards the unified theory for texture modeling. In *IEEE Conference Computer Vision and Pattern Recognition*, pages 686–693, 1996.