

DETECTING HIDDEN MESSAGES USING HIGHER-ORDER STATISTICAL MODELS

Hany Farid

Department of Computer Science
Dartmouth College
Hanover NH 03755

ABSTRACT

Techniques for information hiding have become increasingly more sophisticated and widespread. With high-resolution digital images as carriers, detecting hidden messages has become considerably more difficult. This paper describes a new approach to detecting hidden messages in images. The approach uses a wavelet-like decomposition to build higher-order statistical models of natural images. A Fisher linear discriminant analysis is then used to discriminate between untouched and adulterated images.

1. INTRODUCTION

Information hiding techniques (e.g., steganography and watermarking) have recently received quite a bit of attention (see [8, 1, 6, 11] for general reviews). With digital images as carriers, detecting the presence of hidden messages poses significant challenges. Although the presence of embedded messages is often imperceptible to the human eye, it may nevertheless disturb the statistics of an image. Previous approaches to detecting such deviations [5, 7, 21, 13] typically examine first-order statistical distributions of intensity or transform coefficients (e.g., discrete cosine transform, DCT). The drawback of this analysis is that simple counter-measures that match first-order statistics are likely to foil detection. In contrast, the approach taken here relies on building higher-order statistical models for natural images [9, 15, 22, 10, 17] and looking for deviations from these models. I show that, across a large number of natural images, there exists strong higher-order statistical regularities within a wavelet-like decomposition. The embedding of a message significantly alters these statistics and thus becomes detectable.

The author can be reached at farid@cs.dartmouth.edu. I am 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).

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 is that such decompositions exhibit statistical regularities that can be exploited (e.g., [16, 14, 2]). Described below is one such decomposition, and a set of statistics collected from this decomposition.

The decomposition employed here is based on separable quadrature mirror filters (QMFs) [19, 20, 18]. 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 lowpass subband. 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.

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 is 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 subset of all possible neighbors¹ is given by:

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

¹The particular choice of spatial, orientation and scale neighbors was motivated by the observations of [2] and modified to include non-casual neighbors.

where w_k denotes scalar weighting values. This linear relationship is 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 are determined by minimizing the quadratic error function $E(\vec{w}) = [\vec{V} - Q\vec{w}]^2$. This error function is minimized analytically by differentiating with respect to \vec{w} : $dE(\vec{w})/d\vec{w} = 2Q^T[\vec{V} - Q\vec{w}]$, setting the result equal to zero, and solving for \vec{w} to yield:

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

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

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

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), a class specific method for pattern recognition, is employed [4, 3]. 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:

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

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

4. RESULTS

Shown in Fig. 1 are several examples taken from a database of natural images². Each 8-bit per channel RGB image is

²Images were downloaded from: philip.greenspun.com and reproduced here with permission from Philip Greenspun.



Fig. 1. Sample images.

cropped to a central 640×480 pixel area. Statistics from 1,800 such images are collected as follows. Each image is first converted from RGB to 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, Section 2. To reduce sensitivity to noise in the linear predictor, only coefficient magnitudes greater than 1.0 are considered. The training set of “no-steg” statistics comes from either 1,800 JPEG images (quality ≈ 75), 1,800 GIF images (LZW compression), or 1,800 TIFF images (no compression). The GIF and TIFF images are converted from their original JPEG format.

Messages are embedded into JPEG images using either Jsteg³ or OutGuess⁴ (run with (+) and without (–) statistical correction). Jsteg and OutGuess are transform-based systems that embed messages by modulating the DCT coefficients. Unique to OutGuess is a technique for embedding into only one-half of the redundant bits and then using the remaining redundant bits to preserve the first-order distribution of DCT coefficients [12]. Messages are embedded into GIF images using EzStego⁵ which modulates the least significant bits of the sorted color palette index. Messages are embedded into the TIFF images using a generic LSB embedding that modulates the least-significant bit of a random subset of the pixel intensities. In each case, a message consists of a $n \times n$ pixel ($n \in [32, 256]$) central portion of a random image chosen from the same image database. After the message is embedded into the cover image, the same transformation, decomposition, and collection of statistics

³Jsteg V4, by Derek Upham, is available at <ftp.funet.fi>

⁴OutGuess, by Niels Provos, is available at www.outguess.org

⁵EzStego, by Romana Machado, is available at www.stego.com

| Embedding | Message | JPEG | GIF | TIFF |
|-----------------------|------------------|------|------|------|
| Jsteg | 256×256 | 94.0 | - | - |
| Jsteg | 128×128 | 95.7 | - | - |
| Jsteg | 64×64 | 95.3 | - | - |
| Jsteg | 32×32 | 51.7 | - | - |
| OutGuess ⁻ | 256×256 | 92.8 | - | - |
| OutGuess ⁻ | 128×128 | 63.4 | - | - |
| OutGuess ⁻ | 64×64 | 27.7 | - | - |
| OutGuess ⁻ | 32×32 | 5.9 | - | - |
| OutGuess ⁺ | 256×256 | 74.4 | - | - |
| OutGuess ⁺ | 128×128 | 41.4 | - | - |
| OutGuess ⁺ | 64×64 | 14.0 | - | - |
| OutGuess ⁺ | 32×32 | 4.1 | - | - |
| EzStego | 194×194 | - | 45.2 | - |
| EzStego | 128×128 | - | 13.8 | - |
| EzStego | 64×64 | - | 2.9 | - |
| EzStego | 32×32 | - | 1.6 | - |
| LSB | 194×194 | - | - | 42.3 |
| LSB | 128×128 | - | - | 16.8 |
| LSB | 64×64 | - | - | 2.8 |
| LSB | 32×32 | - | - | 1.3 |

Table 1. Classification accuracy (percent) with less than 1% false positives for varying message sizes (the maximum message size for EzStego and LSB is 194×194).

as described above is performed.

The two-class FLD, Section 3, is trained separately to classify the JPEG, GIF and TIFF embeddings. In each case, the training set consists of the 1,800 “no-steg” images, and a random subset of 1,800 “steg” images embedded either with OutGuess⁺, EzStego or LSB, and with varying message sizes.⁶ For each classifier, the FLD projection axis and a threshold, yielding a 1% false-positive rate, is fixed and then used to classify all of the remaining previously unseen steg images of the same format, Table 1. In this table, the third through fifth columns correspond to the JPEG, GIF and TIFF classifiers, respectively. Note that the JPEG classifier generalizes to the different embedding programs not previously seen by the classifier. In general each image format, and possibly each class of embedding algorithms, will require separate training to learn the relevant statistical deviations.

5. DISCUSSION

Messages can be embedded into digital images in ways that are imperceptible to the human eye, and yet, these manipulations can significantly alter the underlying statistics of an

⁶OutGuess is run with unlimited iterations to find the best embedding. OutGuess imposes limits on the message size, so not all images were able to be used for cover. This is significant only for message sizes of 256×256 , where less than 300 steg images were generated.

image. To detect the presence of hidden messages a model based on higher-order 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. This makes it possible to detect, with a reasonable degree of accuracy, the presence of hidden messages in digital images. To avoid detection, of course, one need only embed a small enough message that does not significantly disturb the image statistics.

There are several directions that can be explored in order to improve detection accuracy. The particular choice of statistics is somewhat ad hoc, as such it would be beneficial to choose a set of statistics that optimize detection rates. However convenient, FLD analysis is linear, and detection rates would almost certainly benefit from a more flexible non-linear classification scheme. The indiscriminant comparison of image statistics across all images could be replaced with a class-based analysis, where, for example, indoor and outdoor scenes are compared separately. And lastly, although only tested on images, there is no inherent reason why the approaches described here would not work for audio signals or video sequences, arbitrary image file formats, or other hiding algorithms.

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. It is possible, 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.

6. 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] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.
- [4] R. Fisher. The use of multiple measures in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
- [5] J. Fridrich and M. Goljan. Practical steganalysis: State of the art. In *SPIE Photonics West, Electronic Imaging*, San Jose, CA, 2002.
- [6] N. Johnson and S. Jajodia. Exploring steganography: seeing the unseen. *IEEE Computer*, pages 26–34, 1998.
- [7] N. Johnson and S. Jajodia. Steganalysis of images created using current steganography software. *Lecture notes in Computer Science*, pages 273–289, 1998.
- [8] D. Kahn. The history of steganography. In *Proceedings of Information Hiding, First International Workshop*, Cambridge, UK, 1996.
- [9] D. Kersten. Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12):2395–2400, 1987.
- [10] G. Krieger, C. Zetzsche, 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.
- [11] E.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn. Information hiding - a survey. *Proceedings of the IEEE*, 87(7):1062–1078, 1999.
- [12] N. Provos. Defending against statistical steganalysis. In *10th USENIX Security Symposium*, Washington, DC, 2001.
- [13] N. Provos and P. Honeyman. Detecting steganographic content on the internet. Technical Report CITI 01-1a, University of Michigan, 2001.
- [14] R. Rinaldo and G. Calvagno. Image coding by block prediction of multiresolution subimages. *IEEE Transactions on Image Processing*, 4(7):909–920, 1995.
- [15] D.L. Ruderman and W. Bialek. Statistics of natural image: Scaling in the woods. *Phys. Rev. Letters*, 73(6):814–817, 1994.
- [16] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.
- [17] 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.
- [18] E.P. Simoncelli and E.H. Adelson. *Subband image coding*, chapter Subband transforms, pages 143–192. Kluwer Academic Publishers, Norwell, MA, 1990.
- [19] P.P. Vaidyanathan. Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques. *IEEE ASSP Magazine*, pages 4–20, 1987.
- [20] M. Vetterli. A theory of multirate filter banks. *IEEE Transactions on ASSP*, 35(3):356–372, 1987.
- [21] A. Westfeld and A. Pfitzmann. Attacks on steganographic systems. In *Proceedings of Information Hiding, Third International Workshop*, Dresden, Germany, 1999.
- [22] 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.