Feature-Based Steganalysis for JPEG Images and Its Implications for Future Design of Steganographic Schemes

Jessica Fridrich

Dept. of Electrical Engineering, SUNY Binghamton, Binghamton, NY 13902-6000, USA
fridrich@binghamton.edu
http://www.ws.binghamton.edu/fridrich

Abstract. In this paper, we introduce a new feature-based steganalytic method for JPEG images and use it as a benchmark for comparing JPEG steganographic algorithms and evaluating their embedding mechanisms. The detection method is a linear classifier trained on feature vectors corresponding to cover and stego images. In contrast to previous blind approaches, the features are calculated as an L_1 norm of the difference between a specific macroscopic functional calculated from the stego image and the same functional obtained from a decompressed, cropped, and recompressed stego image. The functionals are built from marginal and joint statistics of DCT coefficients. Because the features are calculated directly from DCT coefficients, conclusions can be drawn about the impact of embedding modifications on detectability. Three different steganographic paradigms are tested and compared. Experimental results reveal new facts about current steganographic methods for JPEGs and new design principles for more secure JPEG steganography.

1 Introduction

Steganography is the art of invisible communication. Its purpose is to hide the very presence of communication by embedding messages into innocuous-looking cover objects. Each steganographic communication system consists of an embedding algorithm and an extraction algorithm. To accommodate a secret message in a digital image, the original cover image is slightly modified by the embedding algorithm. As a result, the stego image is obtained.

Steganalysis is the art of discovering hidden data in cover objects. As in cryptanalysis, it is assumed that the steganographic method is publicly known with the exception of a secret key. Steganography is considered secure if the stego-images do not contain any detectable artifacts due to message embedding. In other words, the set of stego-images should have the same statistical properties as the set of cover-images. If there exists an algorithm that can guess whether or not a given image contains a secret message with a success rate better than random guessing, the steganographic

J. Fridrich (Ed.): IH 2004, LNCS 3200, pp. 67-81, 2004.
© Springer-Verlag Berlin Heidelberg 2004

system is considered broken. For a more exact treatment of the concept of steganographic security, the reader is referred to [1,2].

1.1 Steganalytic Methods

Several trends have recently appeared in steganalysis. One of the first general steganalytic methods was the "chi-square attack" by Westfeld [3]. The original version of this attack could detect sequentially embedded messages and was later generalized to randomly scattered messages [4,5]. Because this approach is based solely on the first order statistics and is applicable only to idempotent embedding operations, such as LSB (Least Significant Bit) flipping, its applicability to modern steganographic schemes, that are aware of the Cachin criterion [2], is rather limited.

Another major stream in steganalysis is based on the concept of a distinguishing statistic [6]. In this approach, the steganalyst first carefully inspects the embedding algorithm and then identifies a quantity (the distinguishing statistics) that changes predictably with the length of the embedded message, yet one that can be calibrated for cover images. For JPEG images, this calibration is done by decompressing the stego image, cropping by a few pixels in each direction, and recompressing using the same quantization table. The distinguishing statistic calculated from this image is used as an estimate for the same quantity from the cover image. Using this calibration, highly accurate and reliable estimation of the embedded message length can be constructed for many schemes [6]. The detection philosophy is not limited to any specific type of the embedding operation and works for randomly scattered messages as well. One disadvantage of this approach is that the detection needs to be customized to each embedding paradigm and the design of proper distinguishing statistics cannot be easily automatized.

The third direction in steganalysis is formed by blind classifiers. Pioneered by Memon and Farid [7,15], a blind detector learns what a typical, unmodified image looks like in a multi-dimensional feature space. A classifier is then trained to learn the differences between cover and stego image features. The 72 features proposed by Farid are calculated in the wavelet decomposition of the stego image as the first four moments of coefficients and the log error between the coefficients and their globally optimal linear prediction from neighboring wavelet modes. This methodology combined with a powerful Support Vector Machine classifier gives very impressive results for most current steganographic schemes. Farid demonstrated a very reliable detection for J-Steg, both versions of OutGuess, and for F5 (color images only). The biggest advantage of blind detectors is their potential ability to detect any embedding scheme and even to classify embedding techniques by their position in the feature space. Among the disadvantages is that the methodology will always likely be less accurate than targeted approaches and it may not be possible to accurately estimate the secret message length, which is an important piece of information for the steganalyst.

Introducing blind detectors prompted further research in steganography. Based on the previous work of Eggers [8], Tzschoppe [9] constructed a JPEG steganographic scheme (HPDM) that is undetectable using Farid's scheme. However, the same scheme is easily detectable [10] using a single scalar feature – the calibrated spatial

blockiness [6]. This suggests that it should be possible to construct a very powerful feature-based detector (blind on the class of JPEG images) if we used *calibrated* features computed directly in the *DCT domain* rather than from a somewhat arbitrary wavelet decomposition. This is the approach taken in this paper.

1.2 Proposed Research

We combine the concept of calibration with the feature-based classification to devise a blind detector specific to JPEG images. By calculating the features directly in the JPEG domain rather than in the wavelet domain, it appears that the detection can be made more sensitive to a wider type of embedding algorithms because the calibration process (for details, see Sec. 2) increases the features' sensitivity to the embedding modifications while suppressing image-to-image variations. Another advantage of calculating the features in the DCT domain is that it enables more straightforward interpretation of the influence of individual features on detection as well as easier formulation of design principles leading to more secure steganography.

The proposed detection can also be viewed as a new approach to the definition of steganographic security. According to Cachin, a steganographic scheme is considered secure if the Kullback-Leibler distance between the distribution of stego and cover images is zero (or small for ε-security). Farid's blind detection is essentially a reflection of this principle. Farid first determines the statistical model for natural images in the feature space and then calculates the distance between a specific image and the statistical model. This "distance" is then used to determine whether the image is a stego image. In our approach, we change the security model and use the stego image as a *side-information* to recover some statistics of the cover image. Instead of measuring the distance between the image and a statistical model, we measure the distance between certain parameters of the stego image and the same parameters related to the original image that we succeeded to capture by calibration.

The paper is organized as follows. In the next section, we explain how the features are calculated and why. In Section 3, we give the details of the detection scheme and discuss the experimental results for OutGuess [11], F5 [13], and Model Based Steganography [12,14]. Implications for future design of steganographic schemes are discussed in Section 4. The paper is summarized in Section 5.

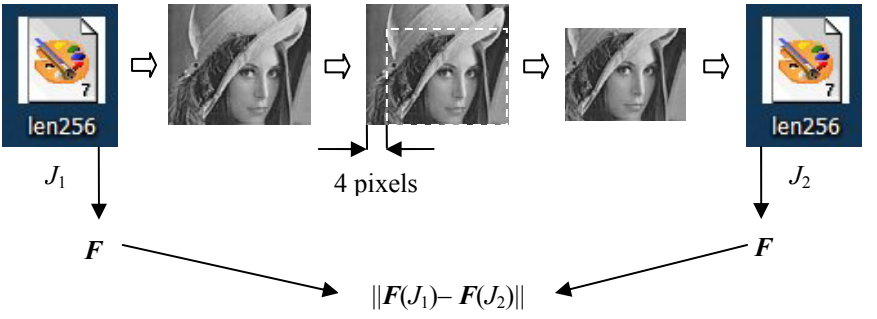
2 Calibrated Features

Two types of features will be used in our analysis – first order features and second order features. Also, some features will be constructed in the DCT domain, while others in the spatial domain. In the whole paper, scalar quantities will be represented with a non-bold italic font, while vectors and matrices will always be in bold italics. The L_1 norm is defined for a vector (or matrix) as a sum of absolute values of all vector (or matrix) elements.

All features are constructed in the following manner. A vector functional $\boldsymbol{F}$ is applied to the stego JPEG image J_1. This functional could be the global DCT coefficient histogram, a co-occurrence matrix, spatial blockiness, etc. The stego image J_1 is de-

compressed to the spatial domain, cropped by 4 pixels in each direction, and recompressed with the same quantization table as J_1 to obtain J_2. The same vector functional $\boldsymbol{F}$ is then applied to J_2. The final feature f is obtained as an L_1 norm of the difference

$$f = \left\| \boldsymbol{F}(J_1) - \boldsymbol{F}(J_2) \right\|_{L_1}. \tag{1}$$



The logic behind this choice for features is the following. The cropping and recompression should produce a "calibrated" image with most macroscopic features similar to the original cover image. This is because the cropped stego image is perceptually similar to the cover image and thus its DCT coefficients should have approximately the same statistical properties as the cover image. The cropping by 4 pixels is important because the 8×8 grid of recompression "does not see" the previous JPEG compression and thus the obtained DCT coefficients are not influenced by previous quantization (and embedding) in the DCT domain. One can think of the cropped /recompressed image as an approximation to the cover image or as a side-information. The use of the calibrated image as a side-information has proven very useful for design of very accurate targeted steganalytic methods in the past [6].

2.1 First Order Features

The simplest first order statistic of DCT coefficients is their histogram. Suppose the stego JPEG file is represented with a DCT coefficient array $d_k(i, j)$ and the quantization matrix $Q(i, j)$, $i, j = 1,\ldots,8$, $k = 1, \ldots, B$. The symbol $d_k(i, j)$ denotes the (i, j)-th quantized DCT coefficient in the k-th block (there are total of B blocks). The global histogram of all $64k$ DCT coefficients will be denoted as H_r, where $r = L, \ldots, R$, $L = \min_{k,i,j} d_k(i, j)$ and $R = \max_{k,i,j} d_k(i, j)$.

There are steganographic programs that preserve $\boldsymbol{H}$ [8,10,11]. However, the schemes in [8,9,11] only preserve the global histogram and not necessarily histograms of individual DCT modes. Thus, we add individual histograms for low frequency DCT modes to our set of functionals. For a fixed DCT mode (i, j), let h_r^{ij}, $r = L, \ldots, R$, denote the individual histogram of values $d_k(i, j)$, $k = 1, \ldots, B$. We only use histograms of low frequency DCT coefficients because histograms of coefficients from medium and higher frequencies are usually statistically unimportant due to the small number of non-zero coefficients.

To provide additional first order macroscopic statistics to our set of functionals, we have decided to include "dual histograms". For a fixed coefficient value d, the dual histogram is an 8×8 matrix g_{ij}^{d}

$$g_{ij}^{d} = \sum_{k=1}^{B} \delta(d, d_k(i,j)), \tag{2}$$

where $\delta(u,v)$=1 if u=v and 0 otherwise. In words, g_{ij}^{d} is the number of how many times the value d occurs as the (i, j)-th DCT coefficient over all B blocks in the JPEG image. The dual histogram captures how a given coefficient value d is distributed among different DCT modes. Obviously, if a steganographic method preserves all individual histograms, it also preserves all dual histograms and vice versa.

2.2 Second Order Features

If the corresponding DCT coefficients from different blocks were independent, then any embedding scheme that preserves the first order statistics – the histogram – would be undetectable by Cachin's definition of steganographic security [2]. However, because natural images can exhibit higher-order correlations over distances larger than 8 pixels, individual DCT modes from neighboring blocks are not independent. Thus, it makes sense to use features that capture inter-block dependencies because they will likely be violated by most steganographic algorithms.

Let I_r and I_c denote the vectors of block indices while scanning the image "by rows" and "by columns", respectively. The first functional capturing inter-block dependency is the "variation" V defined as

$$V = \frac{\sum_{i,j=1}^{8} \sum_{k=1}^{|I_r|-1} | d_{I_r(k)}(i,j) - d_{I_r(k+1)}(i,j) | + \sum_{i,j=1}^{8} \sum_{k=1}^{|I_c|-1} | d_{I_c(k)}(i,j) - d_{I_c(k+1)}(i,j) |}{| I_r | + | I_c |}. \tag{3}$$

Most steganographic techniques in some sense add entropy to the array of quantized DCT coefficients and thus are more likely to increase the variation V than decrease.

Embedding changes are also likely to increase the discontinuities along the 8×8 block boundaries. In fact, this property has proved very useful in steganalysis in the past [6,10,12]. Thus, we include two blockiness measures B_α, α = 1, 2, to our set of functionals. The blockiness is calculated from the decompressed JPEG image and thus represents an "integral measure" of inter-block dependency over all DCT modes over the whole image:

$$B_\alpha = \frac{\sum_{i=1}^{\lfloor (M-1)/8 \rfloor} \sum_{j=1}^{N} \left| x_{8i,j} - x_{8i+1,j} \right|^\alpha + \sum_{j=1}^{\lfloor (N-1)/8 \rfloor} \sum_{i=1}^{M} \left| x_{i,8j} - x_{i,8j+1} \right|^\alpha}{N \lfloor (M-1)/8 \rfloor + M \lfloor (N-1)/8 \rfloor}. \tag{4}$$

In the expression above, M and N are image dimensions and x_{ij} are grayscale values of the decompressed JPEG image.

The final three functionals are calculated from the co-occurrence matrix of neighboring DCT coefficients. Recalling the notation, $L \le d_k(i, j) \le R$, the co-occurrence matrix $\boldsymbol{C}$ is a square $D \times D$ matrix, $D = R - L + 1$, defined as follows

$$C_{st} = \frac{\sum_{k=1}^{|I_r|-1}\sum_{i,j=1}^{8}\delta\left(s, d_{I_r(k)}(i,j)\right)\delta\left(t, d_{I_r(k+1)}(i,j)\right) + \sum_{k=1}^{|I_c|-1}\sum_{i,j=1}^{8}\delta\left(s, d_{I_c(k)}(i,j)\right)\delta\left(t, d_{I_c(k+1)}(i,j)\right)}{|I_r| + |I_c|} \tag{5}$$

The co-occurrence matrix describes the probability distribution of pairs of neighboring DCT coefficients. It usually has a sharp peak at (0,0) and then quickly falls off. Let $\boldsymbol{C}(J_1)$ and $\boldsymbol{C}(J_2)$ be the co-occurrence matrices for the JPEG image J_1 and its calibrated version J_2, respectively. Due to the approximate symmetry of C_{st} around $(s, t) = (0, 0)$, the differences $C_{st}(J_1) - C_{st}(J_2)$ for $(s, t) \in \{(0,1), (1,0), (-1,0), (0,-1)\}$ are strongly positively correlated. The same is true for the group $(s, t) \in \{(1,1), (-1,1), (1,-1), (-1,-1)\}$. For practically all steganographic schemes, the embedding changes to DCT coefficients are essentially perturbations by some small value. Thus, the co-occurrence matrix for the embedded image can be obtained as a convolution $\boldsymbol{C} * P(q)$, where P is the probability distribution of the embedding distortion, which depends on the relative message length q. This means that the values of the co-occurrence matrix $\boldsymbol{C} * P(q)$ will be more "spread out". To quantify this spreading, we took the following three quantities as our *features*:

$$\begin{aligned} N_{00} &= C_{0,0}(J_1) - C_{0,0}(J_2) \\ N_{01} &= C_{0,1}(J_1) - C_{0,1}(J_2) + C_{1,0}(J_1) - C_{1,0}(J_2) + C_{-1,0}(J_1) - C_{-1,0}(J_2) + C_{0,-1}(J_1) - C_{0,-1}(J_2) \\ N_{11} &= C_{1,1}(J_1) - C_{1,1}(J_2) + C_{1,-1}(J_1) - C_{1,-1}(J_2) + C_{-1,1}(J_1) - C_{-1,1}(J_2) + C_{-1,-1}(J_1) - C_{-1,-1}(J_2)\,. \end{aligned} \tag{6}$$

The final set of 23 functionals (the last three are directly features) used in this paper is summarized in Table 1.

3 Steganalytic Classifier

We used the Greenspun image database (www.greenspun.com) consisting of 1814 images of size approximately 780×540. All images were converted to grayscale, the black border frame was cropped away, and the images were compressed using an 80% quality JPEG. We selected the F5 algorithm [13], OutGuess 0.2 [11], and the recently developed Model based Steganography without (MB1) and with (MB2) deblocking [12,14] as three examples of different steganographic paradigms for JPEG images.

Each steganographic technique was analyzed separately. For a fixed relative message length expressed in terms of bits per non-zero DCT coefficient of the cover image, we created a training database of embedded images. The Fisher Linear Discriminant classifier was trained on 1314 cover and 1314 stego images. The generalized eigenvector obtained from this training was then used to calculate the ROC curve for the remaining 500 cover and 500 stego images. The detection performance was evaluated using detection reliability ρ defined below.

Table 1. All 23 distinguishing functionals

Functional/feature name	Functional $\boldsymbol{F}$
Global histogram	$\boldsymbol{H} / \| \boldsymbol{H} \|_{L_1}$
Individual histograms for 5 DCT modes	$\frac{\boldsymbol{h}^{21}}{\| \boldsymbol{h}^{21} \|_{L_1}}, \frac{\boldsymbol{h}^{31}}{\| \boldsymbol{h}^{31} \|_{L_1}}, \frac{\boldsymbol{h}^{12}}{\| \boldsymbol{h}^{12} \|_{L_1}}, \frac{\boldsymbol{h}^{22}}{\| \boldsymbol{h}^{22} \|_{L_1}}, \frac{\boldsymbol{h}^{13}}{\| \boldsymbol{h}^{13} \|_{L_1}}$
Dual histograms for 11 DCT values (–5, …, 5)	$\frac{\boldsymbol{g}^{-5}}{\| \boldsymbol{g}^{-5} \|_{L_1}}, \frac{\boldsymbol{g}^{-4}}{\| \boldsymbol{g}^{-4} \|_{L_1}}, \ldots, \frac{\boldsymbol{g}^{4}}{\| \boldsymbol{g}^{4} \|_{L_1}}, \frac{\boldsymbol{g}^{5}}{\| \boldsymbol{g}^{5} \|_{L_1}}$
Variation	V
L_1 and L_2 blockiness	B_1, B_2
Co-occurrences	N_{00}, N_{01}, N_{11} (features, not functionals)

The reason why we used in our tests message lengths proportional to the number of non-zero DCT coefficients in each image was to create stego image databases for which the detection is approximately of the same level of difficulty. In our experience, it is easier to detect a 10000-bit message in a smaller JPEG file than in a larger JPEG file. The testing was done for the following relative embedding rates expressed in bpc (Bits Per non-zero DCT Coefficient), bpc = 0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8. If, for a given image, the bpc rate was larger than the maximal bpc rate $\text{bpc}_{\max}$ determined by the image capacity, we took $\text{bpc}_{\max}$ as the embedding rate. The only exception to this rule was the MB2 method, where we took $0.95 \times \text{bpc}_{\max}$ as the maximal rate because, for the maximal embedding rate, the deblocking algorithm in MB2 frequently failed to embed the whole message. Fig. 1 shows the capacity for all three methods expressed in bits per non-zero DCT coefficient.

The detection results were evaluated using 'detection reliability' ρ defined as

$$\rho = 2A - 1, \tag{7}$$

where A is the area under the Receiver Operating Characteristic (ROC) curve, also called an accuracy. We scaled the accuracy in order to obtain $\rho = 1$ for a perfect detection and $\rho = 0$ when the ROC coincides with the diagonal line (reliability of detection is 0). The detection reliability for all three methods is shown in Table 2.

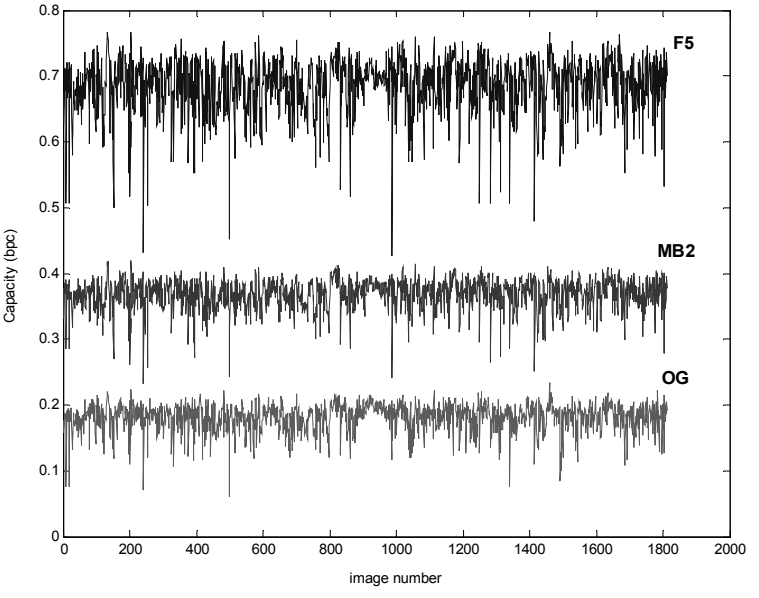


Fig. 1. Capacity for the tested techniques expressed in bits per non-zero DCT coefficient. The capacity for MB1 is double that of MB2. The F5 and MB1 algorithms provide the highest capacity

Table 2. Detection reliability ρ for F5 with matrix embedding $(1, k, 2^k - 1)$, F5 with turned off matrix embedding (1,1,1), OutGuess 0.2 (OG), Model based Steganography without and with deblocking (MB1 and MB2, respectively) for different embedding rates (U = unachievable rate)

bpc	F5	F5_111	OG	MB1	MB2
0.05	0.2410	0.6451	0.8789	0.2197	0.1631
0.1	0.5386	0.9224	0.9929	0.4146	0.3097
0.2	0.9557	0.9958	0.9991	0.7035	0.5703
0.4	0.9998	0.9999	U	0.9375	0.8243
0.6	1.0000	1.0000	U	0.9834	U
0.8	1.0000	1.0000	U	0.9916	U

One can clearly see that the OutGuess algorithm is the most detectable. Also, it provides the smallest capacity. The detection reliability is relatively high even for embedding rates as small as 0.05 bpc and the method becomes highly detectable for messages above 0.1 bpc. To guarantee a fair comparison, we have tested F5 both with and without matrix embedding because some programs could be easily adapted to incorporate it (e.g., OutGuess). Turning off the matrix embedding, the F5 algorithm still performs better than OutGuess. The matrix embedding significantly decreases the detectability for short messages. This is understandable because it improves the embedding efficiency (number of bits embedded per change). Because OutGuess needs to reserve a relatively large portion of coefficients for the correction step, its embed-

ding efficiency is lower compared to F5. This seems to have a bigger impact on the detectability than the fact that OutGuess preserves the global histogram of DCT coefficients.

Table 3. Detection reliability for individual features for all three embedding algorithms for *fully* embedded images (for fully embedded images, F5 with matrix embedding and without matrix embedding coincide)

Functional/feature	Method			
	F5	OutGuess 0.2	MB1	MB2
Global histogram	0.9936	0.8110	0.1224	0.0359
Indiv. histogram for (2,1)	0.9343	0.6625	0.6166	0.3775
Indiv. histogram for (3,1)	0.9940	0.7521	0.1018	0.0606
Indiv. histogram for (1,2)	0.8719	0.6353	0.4686	0.3828
Indiv. histogram for (2,2)	0.9827	0.7879	0.5782	0.3499
Indiv. histogram for (1,3)	0.9879	0.7718	0.0080	0.0095
Dual histogram for –5	0.1294	0.0853	0.1350	0.1582
Dual histogram for –4	0.1800	0.2727	0.0338	0.0448
Dual histogram for –3	0.2188	0.4239	0.6675	0.3239
Dual histogram for –2	0.2939	0.9921	0.2724	0.0733
Dual histogram for –1	0.4824	0.9653	0.7977	0.4952
Dual histogram for 0	0.9935	0.6160	0.2697	0.0859
Dual histogram for 1	0.5101	0.4068	0.6782	0.3336
Dual histogram for 2	0.2740	0.8437	–0.0058	0.0311
Dual histogram for 3	0.1990	0.7060	0.0904	0.1208
Dual histogram for 4	0.1421	0.1933	0.0169	0.0100
Dual histogram for 5	0.1315	0.1055	0.4097	0.2540
Variation	0.7891	0.5576	0.7239	0.2337
L_1 blockiness	0.9908	0.1677	0.5749	0.2737
L_2 blockiness	0.9411	0.1064	0.2485	0.2253
Co-occurrence N_{00}	0.9997	0.4180	0.8818	0.6088
Co-occurrence N_{01}	0.9487	0.9780	0.8433	0.5569
Co-occurrence N_{11}	0.9954	0.9282	0.7873	0.4957

Both MB1 and MB2 methods clearly have the best performance of all three tested algorithms. MB1 preserves not only the global histogram, but all marginal statistics (histograms) for each individual DCT mode. It is quite remarkable that this can be achieved with an embedding efficiency slightly over 2 bits per change (compared to 1.5 bits per change for F5 and roughly 1 for OutGuess 0.2). This is likely because MB1 does not avoid any other coefficients than 0 and its embedding mechanism is guaranteed to embed the maximal number of bits given the fact that marginal statistics of all coefficients must be preserved. The MB2 algorithm has the same embedding mechanism as MB1 but reserves one half of the capacity for modifications that bring the blockiness of the stego image to its original value. As a result, MB2 is less detectable than MB1 at the expense of a two times smaller embedding capacity. Both methods perform better than F5 with matrix embedding and are significantly better than F5 without matrix embedding. Even for messages close to 100% capacity, the detection of MB2 is not very reliable. An ROC with $\rho = 0.82$ does not allow reliable

detection with a small false positive rate (c.f., Fig. 2). Never the less, in the strict formulation of steganographic security, whenever the embedded images can be distinguished from cover images with a better algorithm than random guessing, the steganography is detectable. Thus, we conclude that the Model based Steganography is detectable using our feature-based approach on our test database.

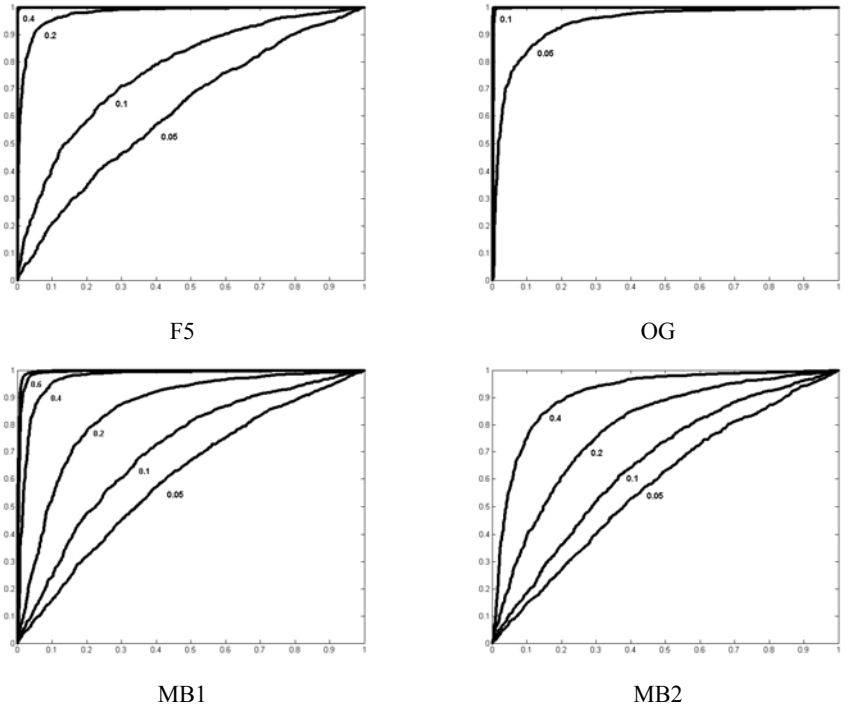


Fig. 2. ROC curves for embedding capacities and methods from Table 2.

For each steganographic method, we also measured the influence of each individual feature f as its detection reliability $\rho(f)$ obtained from the ROC curve calculated from the single feature f and no other features. We acknowledge that the collection of individual reliabilities $\rho(f)$ does not have to necessarily capture the performance of the whole detection algorithm in the 23 dimensional space. This is because it is possible that none of the individual features themselves has any distinguishing power, yet the collection of all features achieves a perfect detection. Never the less, we use $\rho(f)$ as an indication of how much each feature contributes to the detection.

In Table 2, we show the influence of each feature for each steganographic method for the maximal bpc rate. In the next section, we interpret the results and draw conclusions concerning the existing and future design principles of steganographic schemes for JPEG images.

We note that in our tests, we did not include double compressed images. It is likely that such images would worsen our detection results. In agreement with the conclusion reached in [6], the double compression needs to be first estimated and

then corrected for during the feature calibration. Although we have not tested this, we believe that the feature-based blind steganalysis would work in this case as well.

4 Implications for Steganography

The F5 algorithm uses a non-idempotent embedding operation (subtracting 1) to prevent the attacks based on the chi-square attack and its generalizations [3–5]. It also makes sure that the global stego image histogram is free of any obvious artifacts and looks "natural". In fact, it has been argued by its authors [13] that the stego image looks as if the cover image was originally compressed with a lower JPEG quality factor. However, the F5 predictably modifies the first order statistics and this is why the first six functionals are so influential (see Table 2). It is also not surprising that the dual histogram for 0 has a big influence because of the shrinkage. Note that the second-order statistics significantly contribute to the detection as well. Most features with the exception of dual histograms have high influence on detection.

OutGuess 0.2 was specifically designed to preserve the *global* coefficient histogram. However, OutGuess does not have to necessarily preserve the *individual* histograms or the dual histograms, which is reflected by a relatively large influence for these functionals in Table 2. The most influential functional is the dual histogram for the values –1 and –2. This is again, understandable, considering the embedding mechanism of OutGuess. The values –1 and –2 determine the maximum correctable capacity of the method and thus form the most changed pair of values during the embedding (and the correction step). Although the coefficient counts are preserved, their positions in the JPEG file are highly disturbed, which is why we see a very high influence of features based on dual histograms for values –1 and –2. Another reason why OutGuess is more detectable than F5 is its low embedding efficiency of 1 bit per change compared to 1.5 for F5.

Considering the large influence of the dual histogram, it seems feasible that one could design a targeted steganalytic scheme of the type described in [6] by using the dual histograms for values –1 and –2 as the *distinguishing statistic*. This is an example how the blind analysis may, in turn, give us direct ideas how to estimate the length of the embedded message.

What is somewhat surprising is that the global histogram also has quite a large influence on detection, despite the fact that it is preserved by OutGuess. We will revisit this peculiar finding when we discuss the results for Model Based Steganography below. Another seemingly surprising fact is that although L_1 blockiness proved very useful in designing successful attacks against OutGuess [6], its influence in the proposed detection scheme is relatively small (0.16). This fact is perhaps less surprising if we realize that the distinguishing statistic in [6] was the *increase* of blockiness after full re-embedding rather than the blockiness itself, which appears to be rather volatile.

Looking at the results in Table 1 and 2, there is no doubt that the Model Based Steganography [12,14] is by far the most secure method out of the three tested paradigms. MB1 and MB2 preserve not only the global histogram but also *all* histograms of individual DCT coefficients. Thus, all dual histograms are also preserved. More-

over, MB2 also preserves one second-order functional – the L_1 blockiness. Thus, we conclude that the more statistical measures an embedding method preserves, the more difficult it is to detect it. Consequently, our analysis indicates that it is possible to increase the security of JPEG steganographic schemes by identifying a set of key macroscopic statistical features that should be preserved by the embedding. It is most likely not necessary to preserve all 23 features to substantially decrease the detectability because many of the features are not independent.

One of the most surprising facts revealed by the experiments is that even features based on functionals that are preserved by the embedding may have substantial influence. One might intuitively expect that such features would have very small influence. However, as shown in the next paragraph, preserving a specific *functional* does not automatically mean that the *calibrated feature* will be preserved. Let us take a closer look at the L_1 blockiness as an example.

Preserving the blockiness along the original 8×8 grid (solid lines) does not mean that the blockiness along the shifted grid will also be preserved (see Fig. 2). This is because the embedding and deblocking changes are likely to introduce distortion into the middle of the blocks and thus disturb the blockiness *feature*, which is the difference between the blockiness along the solid and dashed lines. Consequently, it is not surprising that features constructed from functionals that are preserved still have some residual (and not necessarily small) influence in our feature-based detection. This is seen in Table 2 for both OutGuess 0.2 and the Model Based Steganography. Therefore, the designers of future steganographic schemes for JPEG images should consider adding *calibrated* statistics into the set of quantities that should be preserved during embedding.

We further point out that the features derived from the co-occurrence matrix are very influential for all three schemes. For the Model based Steganography, these features are, in fact, the most influential. The MB2 method is currently the only JPEG steganographic method that takes into account inter-block dependencies between DCT coefficients by preserving the blockiness, which is an "integral" measure of these dependencies. Not surprisingly, the scalar blockiness feature does not capture all higher-order statistics of DCT coefficients. Thus, it seems that the next generation of steganographic methods for JPEG images should preserve both the marginal statistics of DCT coefficients and the probability distribution of coefficient pairs from neighboring blocks (the co-occurrence matrix). Eventually, if the stego algorithm preserved all possible statistics of the cover image, the embedding would be presumably undetectable. Although this goal will likely never be achieved, as the embedding algorithm preserves more "orthogonal or independent" statistics, its detectability will quickly decrease. We firmly believe that incorporating a model for the co-occurrence matrices and preserving it would probably lead to significantly less detectable schemes. The Model based Steganography [14] seems to be an appropriate guiding principle to achieve this goal. However, the embedding operation should not be idempotent, otherwise targeted attacks based on re-embedding (c.f., the attack on OutGuess [6]) could likely be mounted.

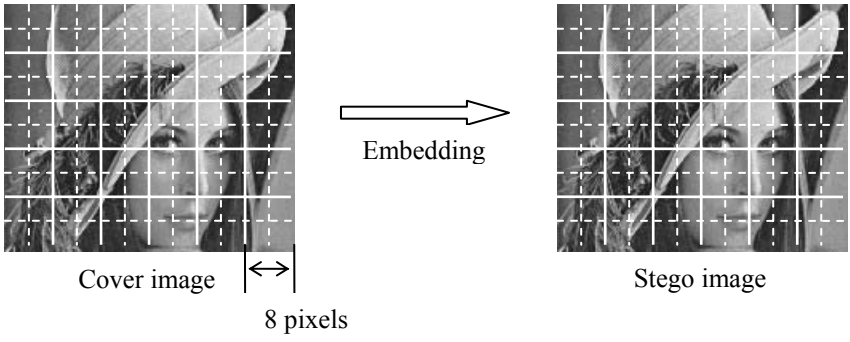


Fig. 2. Blockiness is preserved along the solid lines but not necessarily along the dashed lines

5 Summary and Future Research

In this paper, we developed a new blind feature-based steganalytic method for JPEG images. Each feature is calculated as the L_1 norm of the difference between a specific functional of the stego image and its cropped/recompressed version. This "calibration" can be interpreted as using the stego image as side information to approximately recover some parameters of the cover image. As a result, the calibration decreases image-to-image variations and thus enables more accurate detection.

The features were calculated directly in the DCT domain as first and higher order statistics of DCT coefficients. This enables easier explanation of the impact of embedding modifications on detection as well as direct interpretation of the detection results and easy formulation of design principles for future steganographic methods.

We have applied the detection to several current steganographic schemes some of which are aware of the Cachin criterion [2]. The experimental results were carefully evaluated and interpreted. Conclusions concerning current and future steganographic schemes for JPEGs were also drawn. In particular, we concluded that

1. Secure steganographic schemes must preserve as many statistics of DCT coefficients as possible. It is not enough to preserve the marginal statistics, e.g., the histograms. DCT coefficients exhibit block-to-block dependencies that must be preserved as well.
2. A scheme that preserves more statistics is likely to be more secure than a scheme that preserves fewer statistics. Surprisingly, preserving more statistics may not necessarily lead to small capacity, as shown by Model Based Steganography. This is also because many statistical features one can identify in an image are likely to be dependent.
3. Even though a scheme may preserve a specific statistic $\zeta(X)$ of the cover JPEG image X, the calibrated statistic $\zeta(\text{Compress}(\text{Crop}(X)))$ calculated from the cropped/recompressed image may not necessarily be preserved, thus opening the door for attacks. Future steganographic schemes should add calibrated statistics to their set of preserved statistics.

4. For all tested schemes, one of the most influential features of the proposed detection was the co-occurrence matrix of DCT coefficients (5), which is the probability distribution of coefficient pairs from neighboring blocks. We hypothesize that a scheme that preserves marginal statistics of DCT coefficients and the co-occurrence matrix (which captures block-to-block dependencies) is likely to exhibit improved resistance to attacks. For this purpose, we propose the Model Based Steganography paradigm [12,14] expanded by the model for joint probability distribution of neighboring DCT coefficients.

Although the calibration process is very intuitive, we currently do not have a quantitative understanding of how much information about the cover image can be obtained from the stego image by calibration. For example, for images that contain periodic spatial structures with a period that is an integer multiple of 8, the calibration process may give misleading results (c.f., the spatial resonance phenomenon [6]). In this case, it may be more beneficial to replace the cropping by other operations that will also break the block structure of JPEG images, such as slight rotation, scaling, or random warping. Further investigation of this issue will be part of our future research.

In the future, we also plan to replace the Fisher Linear Discriminant with more sophisticated classifiers, such as Support Vector Machines, to further improve the detection reliability of the proposed steganalytic algorithm. We also plan to develop a multiple-class classifier capable of recognizing stego images produced by different embedding algorithms (steganographic program identification).

Acknowledgements

The work on this paper was supported by the Air Force Research Laboratory, Air Force Material Command, USAF, under research grant number F30602-02-2-0093. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Air Force Research Laboratory, or the U. S. Government. Special thanks belong to Phil Sallee for many useful discussions during preparation of this paper and for providing the code for Model Based Steganography.

References

1. Anderson, R. J. and Petitcolas, F.A.P.: On the Limits of Steganography. IEEE Journal of Selected Areas in Communications. Special Issue on Copyright and Privacy Protection, vol. 16(4) (1998) 474–481
2. Cachin, C.: An Information-Theoretic Model for Steganography. In: Aucsmith, D. (ed.): Information Hiding. 2nd International Workshop. Lecture Notes in Computer Science, Vol. 1525, Springer-Verlag, Berlin Heidelberg New York (1998) pp. 306–318

3. Westfeld, A. and Pfitzmann, A.: Attacks on Steganographic Systems. In: Pfitzmann A. (eds.): 3rd International Workshop. Lecture Notes in Computer Science, Vol.1768. Springer-Verlag, Berlin Heidelberg New York (2000) 61–75
4. Westfeld, A.: Detecting Low Embedding Rates. In: Petitcolas, F.A.P. (ed.): Information Hiding. 5th International Workshop. Lecture Notes in Computer Science, Vol. 2578. Springer-Verlag, Berlin Heidelberg New York (2002) 324–339
5. Provos, N. and Honeyman, P.: Detecting Steganographic Content on the Internet. CITI Technical Report 01-11 (2001)
6. Fridrich, J., Goljan, M., Hogea, D., and Soukal, D.: Quantitative Steganalysis: Estimating Secret Message Length. ACM Multimedia Systems Journal. Special issue on Multimedia Security, Vol. **9**(3) (2003) 288–302
7. Farid H. and Siwei, L.: Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines. In: Petitcolas, F.A.P. (ed.): Information Hiding. 5th International Workshop. Lecture Notes in Computer Science, Vol. 2578. Springer-Verlag, Berlin Heidelberg New York (2002) 340–354
8. Eggers, J., Bäuml, R., and Girod, B.: A Communications Approach to Steganography. In Proc. EI SPIE Electronic Imaging SPIE Vol. 4675 (2002) 26–49
9. Tzschoppe, R., Bäuml, R., Huber, J.B., and Kaup, A.: Steganographic System based on Higher-Order Statistics. Proc. EI SPIE Electronic Imaging. Santa Clara (2003) 156–166
10. Tzschoppe, R.: Personal communication. February (2003)
11. Provos, N.: Defending Against Statistical Steganalysis. 10th USENIX Security Symposium. Washington, DC (2001)
12. Sallee, P.: Model Based Steganography. International Workshop on Digital Watermarking. Seoul, October (2003) 174–188
13. Westfeld, A.: High Capacity Despite Better Steganalysis (F5–A Steganographic Algorithm). In: Moskowitz, I.S. (eds.): Information Hiding. 4th International Workshop. Lecture Notes in Computer Science, Vol.2137. Springer-Verlag, Berlin Heidelberg New York (2001) 289–302
14. Sallee, P.: Model-based methods for steganography and steganalysis. Submitted to International Journal of Image and Graphics. Special issue on Image Data Hiding (2004)
15. I. Avcibas, N. Memon, and B. Sankur, "Steganalysis using Image Quality Metrics", SPIE Security and Watermarking of Multimedia Contents II, Electronic Imaging, San Jose, CA, Jan. 2001.