

# Universal Image Steganalysis Using Rate-Distortion Curves

Mehmet U. Celik<sup>a</sup>, Gaurav Sharma<sup>a</sup>, A. Murat Tekalp<sup>a,b</sup>

<sup>a</sup>University of Rochester, Rochester, NY, USA

<sup>b</sup>Koc University, Istanbul, Turkey

## ABSTRACT

The goal of image steganography is to embed information in a cover image using modifications that are undetectable. In actual practice, however, most techniques produce stego images that are perceptually identical to the cover images but exhibit statistical irregularities that distinguish them from cover images. Statistical steganalysis exploits these irregularities in order to provide the best discrimination between cover and stego images. In general, the process utilizes a heuristically chosen feature set along with a classifier trained on suitable data sets. In this paper, we propose an alternative feature set for steganalysis based on rate-distortion characteristics of images. Our features are based on two key observations: i) data hiding methods typically increase the image entropy in order to encode hidden messages; ii) data hiding methods are limited to the set of small, imperceptible distortions. The proposed feature set is used as the basis of a steganalysis algorithm and its performance is investigated using different data hiding methods.

**Keywords:** steganography, steganalysis, data hiding, image quality

## 1. INTRODUCTION

In many applications, concealing the message content using cryptographic techniques guarantees the security and secrecy of the communications. However, it is sometimes necessary to conceal the very existence of the communications—often to guard against traffic analysis. *Steganography* refers to the techniques that establish a covert (subliminal) communications channel within regular, innocuous message traffic.<sup>1-3</sup>

The basic principles of steganography are best demonstrated with the *Prisoners' Problem* by Simmons.<sup>4</sup> In this scenario, Alice and Bob would like to plan a break out from the prison. However, all their communications are monitored by Willy—the warden. Therefore, they can neither discuss their plans openly, nor can they establish an encrypted channel. In the latter case, Willy will suspect the worst and block their communication. So, the only avenue left to Alice and Bob is to establish a subliminal channel where they conceal their escape correspondence among innocuous looking messages. Note that the requirements for steganography are fundamentally different than cryptography,<sup>5,6</sup> where it is important to prevent the warden from deciphering messages. However, Alice and Bob may want to encrypt their messages before sending them through the cover channel as a secondary line of defense—in case their steganographic technique fails and the warden gets suspicious.

Stated version of the Prisoners' Problem, where Willy can only monitor the messages, is known as the *passive warden scenario*. In some cases, the warden is allowed to alter the messages as he sees fit, but often under some maximum distortion constraints. The latter scenario is known as the *active warden scenario* and often requires a different set of solutions. In this paper, we will focus on the passive warden scenario. As Alice and Bob try to establish a covert communications channel, the task for Willy is to prevent such communications. He has to inspect contents of each message and decide if it contains hidden secondary messages. The task of the passive warden is referred as *steganalysis*.

As digital communications and digital audio-visual signals came into widespread use, the research on steganography and steganalysis is focused on methods that use audio-visual signals as the cover medium. High data rates required by these signals make them especially attractive as a cover medium. For instance, significant amounts

---

Further author information: (Send correspondence to M.U. Celik)

M.U. Celik: E-mail: celik@ece.rochester.edu, Telephone: 1 585 275 8122

G. Sharma: gsharma@ece.rochester.edu, A.M. Tekalp: tekalp@ece.rochester.edu

of data can be hidden in digital images without causing any visible artifacts. A good steganography method, however, should also be undetectable by other means, such as steganalysis algorithms.

In this paper, we look at the image steganalysis problem and develop new algorithms based on rate-distortion concepts. In particular, we observe the effect of different steganographic methods on image rate-distortion characteristics and construct detectors to separate innocuous cover images from message bearing stego images. A brief overview of existing steganography and steganalysis methods is presented in Sec. 2. The proposed detection methods are described in Sec. 3, while performance evaluation of said methods is presented in Sec. 4. Finally, we conclude with a discussion on our future work.

## 2. BACKGROUND

### 2.1. Steganography using compressed vs. uncompressed images

Lossy compression is an effective means of reducing the channel capacity requirements for image transmission while simultaneously preserving the perceptual quality of the images. A compact representation is achieved by discarding the perceptually redundant portions of the image data. As steganographic methods are often limited to imperceptible changes of the image data, compressed images provide an unfavorable medium for image steganography with significantly lower stego-channel capacities. As the compression schemes approach to the ideal case where all perceptually redundant data is eliminated, the stego-channel capacity approaches to zero\*. As a result, uncompressed images pose a more challenging steganalysis problem and we focus on related methods in this paper.

### 2.2. Image steganography methods

A basic and well-known steganography technique for hiding messages into uncompressed digital images is the least significant bit (LSB) embedding method. In LSB embedding, image pixels are traversed in a pre-determined—often pseudo-random—order and LSB value of each pixel is modified to reflect the corresponding bit in the message payload. LSB embedding is preferred in many cases for its simplicity and high embedding rate (up to 1 bpp when all pixels are used). The process amounts to introducing a small amplitude noise signal and the resulting stego image is visually identical to the cover image (PSNR is above 48 dB for 8 bpp images). However, highly structured nature of the noise signal can be effectively used for steganalysis at high payloads.

Stochastic embedding<sup>7</sup> is proposed by Fridrich et al. to remedy the problems of LSB embedding. The embedding artifacts mimic statistics of noise processes naturally encountered in image acquisition devices. In this method, two pseudo-noise sequences with some specified statistics are generated. At each pixel position, one or the other noise sample is added to the pixel value based on the payload bit (and the position of the resulting value in a look-up table). If noise samples are equal, the process does not encode any payload bits, but the noise sample is added regardless. Stochastic embedding achieves high embedding rates (typically above 0.5 bpp) with relatively small distortions.

### 2.3. Image steganalysis methods

Although stego images are perceptually identical to cover images, they often exhibit statistical irregularities that distinguish them from cover images. Statistical steganalysis methods exploit these irregularities in order to provide the best discrimination power between stego and cover images.

Steganalysis algorithms can be classified as *universal* or *model-based* algorithms depending on the steganography methods they target. Universal steganalysis methods try to detect a number of steganographic methods—including yet unknown ones—in a common framework. While they may be preferred for their versatility, their performance is often inferior to model-based algorithms, which specifically target a single steganographic method.

Westfield and Pfitzmann's histogram analysis<sup>8</sup> and Fridrich's RS-steganalysis<sup>9</sup> algorithms are well-known model-based methods that target LSB embedding. Westfield and Pfitzmann's technique is based on analyzing

---

\*Nonetheless, as Anderson and Petitcolas have observed,<sup>1</sup> it is possible to achieve 100% stego-channel capacity in the existence of ideal compression. In that case, any stego-message will decompress into a regular stego image making it impossible to distinguish it from other cover images.

the pair of values (*PoVs*) in the image histogram. The presence of an embedded message is detected with a Chi-Square test<sup>8</sup> that evaluates the (dis)similarity between consecutive histogram bins. The method is most effective for images with high payloads, i.e. when most or all pixels are used for LSB embedding. In RS-steganalysis,<sup>9</sup> Fridrich et al. classify each pixel into *Regular* and *Singular* groups and perform detection based on the relative number of such groups. A pixel is classified into a *Regular* (*Singular*) group if its clique potential is more (less) than its LSB flipped version. Computation of the potential over different cliques takes the spatial distribution of pixels into account and imposes a smoothness constraint. As a result, the algorithm is especially accurate when images conform with the smoothness assumptions.

On the other hand, Farid<sup>10</sup> and Avcibas et al.<sup>11</sup> proposed universal steganalysis methods that are capable of detecting various steganography algorithms. These methods utilize a heuristically chosen image feature set along with a classifier trained on suitable data sets and a given steganographic method. Farid utilizes the first and higher order statistics of a multi-level wavelet transform. In particular, various moments of wavelet coefficients are computed. Resulting feature vector is used to train a linear classifier based on Fisher discriminant. Avcibas et al. have proposed a similar approach that utilizes image quality measures. In particular, various image quality metrics are used to compute the distance between the image under test and its low-pass filtered versions. A classifier is built using linear regression and can successfully detect LSB embedding and some digital watermarking methods.

### 3. PROPOSED METHOD

We propose novel steganalysis algorithms based on the effect of data hiding process on image rate-distortion characteristics. In particular, we make the following assumptions/observations about the data embedding process:

1. *Data embedding typically increases the image entropy:* In order to encode the hidden messages, steganography methods modify parts of the image data. These modifications typically do not conform with the existing image statistics and therefore result in a net increase in the image entropy.
2. *Data embedding methods are limited to the set of small, imperceptible distortions:* Typical steganography methods make only small modifications to ensure perceptual transparency. Perceptually significant parts of the image remain intact.

In the next sections, we built on these observations to develop steganalysis tools targeted at stochastic and LSB embedding methods.

#### 3.1. Detection of Stochastic Embedding

##### 3.1.1. Overview and Rate-Distortion Features

Stochastic embedding does not result in artifacts with known structures, therefore we develop a generalized method to detect the changes in the rate-distortion characteristics. Since real rate-distortion points for signals with unknown probability distributions—such as natural images—cannot be reliably calculated, we use the data rates achieved by lossy compression schemes.

The flow chart of the detection process is seen in Fig. 1. Detector structure is similar to Farid's and Avcibas's detectors. An image feature extraction phase is followed by a classifier that is trained on relevant data sets. As image features, we use the distortion values at different rate points. Mean square error (MSE), mean absolute error (MAE) and weighted mean square error (wMSE) are used as distortion metrics. In order to reduce the within class variability of cover image rate-distortion characteristics, the rate points are defined relative to the lossless rate. That is, instead of compressing images at fixed rates, we compress them at a rate that is a fixed percentage of the lossless rate. The advantage of this approach is empirically verified. We use a Bayesian classifier preceded by a KL transform, which reduces the dimensionality of the feature vector. A detailed explanation of the classifier and the KL transform is provided in the next section.

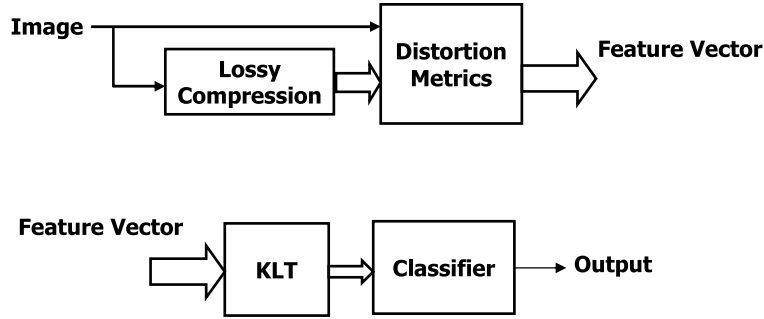


Figure 1. Flow chart describing detection of the stochastic embedding.

### 3.1.2. Classifier

Let us begin by denoting different classes by  $w_i$ , where each  $w_i$  corresponds to a different stego method. We also assume  $1 \leq i \leq M$  that  $M$  such classes exist. We denote the  $L$  dimensional feature vector by  $\mathbf{x}$ .

$$p(\mathbf{x}|w_i) = \frac{1}{(2\pi)^{L/2} |\Sigma_i|^{L/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right) \quad (1)$$

where  $\mu_i = E[\mathbf{x}]$  is the mean value of the  $w_i$  class and  $\Sigma_i$  is the covariance matrix defined as

$$\Sigma_i = E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T] \quad (2)$$

and  $|\Sigma_i|$  denotes the determinant of  $\Sigma_i$  and  $E[\cdot]$  denotes the expected value.

We also define the discriminant function in the logarithmic form as

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|w_i)P(w_i)) \quad (3)$$

$$= -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \ln(P(w_i)) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|) \quad (4)$$

Assuming equiprobable classes and eliminating constant terms, Eqn. 3 can be reduced to

$$g_i(\mathbf{x}) = (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \ln(|\Sigma_i|) \quad (5)$$

$\mu_i$  and  $\Sigma_i$  are estimated from the training samples for each class during the training phase.

When the classifier has to operate on a limited number of training samples with relatively small number of classes, the high dimensionality of the problem adversely affects the classifier performance. In particular, the covariance matrix becomes nearly singular and classification results become sensitive to acquisition noise. A method of reducing the dimensionality of the classification problem while keeping the discriminatory power of the feature vector is to project the feature vector onto a proper subspace.

Let us define the within class and between class scatter matrices,  $S_w$  and  $S_b$  as,

$$S_w = \sum_{i=1}^M P_i E[(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T] \quad (6)$$

$$S_b = \sum_{i=1}^M P_i (\mu_i - \mu_0)(\mu_i - \mu_0)^T \quad (7)$$

where  $\mu_0$  is the global mean vector

$$\mu_0 = \sum_{i=1}^M P_i \mu_i \quad (8)$$

We further define the scattering matrix criterion  $J_3$  as

$$J_3 = \text{trace}\{S_w^{-1}S_b\} \quad (9)$$

We can now define a linear projection from the  $L$  dimensional feature space to  $N$  dimensional sub-space.

$$\hat{\mathbf{y}} = C^T \mathbf{x} \quad (10)$$

The optimal projection matrix with respect to the scattering matrix criterion  $J_3$  is the eigenvectors corresponding to the largest eigenvalues of the system  $S_w^{-1}S_b$ . As the individual scatter matrices  $S_i$ , the within class scatter matrix may also be ill conditioned. Therefore, in practice we use the pseudo-inverse of  $S_w$  in our calculations.

### 3.2. Detection of LSB Embedding

As stated earlier, LSB embedding modifications are highly structured, and therefore result in statistical irregularities that may be exploited for steganalysis. In order to benefit from this structure, we employ an alternative, heuristic detection mechanism that is based on level-embedded (bit-plane scalable) compression.

Let us define the rate required for representing the 7 most significant bits of an image  $I$  as  $R_{7MSB}(I)$ . Similarly,  $R_{7MSB}(I + 1)$  denotes the rate required to represent 7 MSBs of the image after adding one to all pixels. An increase in the DC bias should have an insignificant effect on the rate-distortion characteristics of a natural image. As the statistics of a natural image is unrelated to its binary representation, we expect

$$R_{7MSB}(I) \approx R_{7MSB}(I + 1). \quad (11)$$

However, the LSB embedding process introduces an artificial relation between the image statistics and its binary representation. In particular, Eqn. 11 does not necessarily hold for the stego image  $I_s$ . In general,

$$R_{7MSB}(I_s) \leq R_{7MSB}(I_s + 1). \quad (12)$$

We define the difference between these two rates as

$$\Delta(I) = |R_{7MSB}(I + 1) - R_{7MSB}(I)|. \quad (13)$$

$\Delta(I)$  turns out to be a good discriminant for LSB steganalysis. However, this feature is susceptible to false positives if a cover image has under/over exposed regions. We partially remedy this problem by an additional normalization step. In particular, for a given image we randomize all the LSBs and obtain  $\tilde{I}$ . Note that  $\tilde{I}$  has similar statistics to a stego image with a full payload. As a result,

$$\Psi(I) = \frac{\Delta(I)}{\Delta(\tilde{I})} \quad (14)$$

provides a rough estimate of the payload ratio. In our experiments,  $\Psi(I)$  provided the necessary discrimination power. We provide details of our empirical findings in the next section.

## 4. EXPERIMENTAL RESULTS

### 4.1. Image Database

Successful evaluation of steganalysis algorithms largely depends on the selection of a proper image database. The chosen database should be representative of the images that will be observed in the real world. Nonetheless, the number of different image acquisition devices, common pre/post-processing operations, and the diversity of image subjects make the formation of such a database a significant challenge. While it is immensely important, undertaking such a challenge is beyond the scope of this paper. Therefore, we utilize an existing collection from Kodak's *ftp* site.<sup>12</sup> This collection consists of 108 images in Kodak PhotoCD format and spans a large variety of image subjects (see Fig. 2 for sample images). In fact, the collection even includes some digitally manipulated images (see Fig. 2 top-right). We assume that these images have not been modified by a steganography software and hence represent the set of cover images.

All images are converted from their original PhotoCD format to RGB TIFF images at the base resolution of  $768 \times 512$  pixels. As the proposed methods operate on mono-chrome images, only the Green channel is used. Furthermore, we crop the image to remove the black boundary regions (30 and 20 pixels).

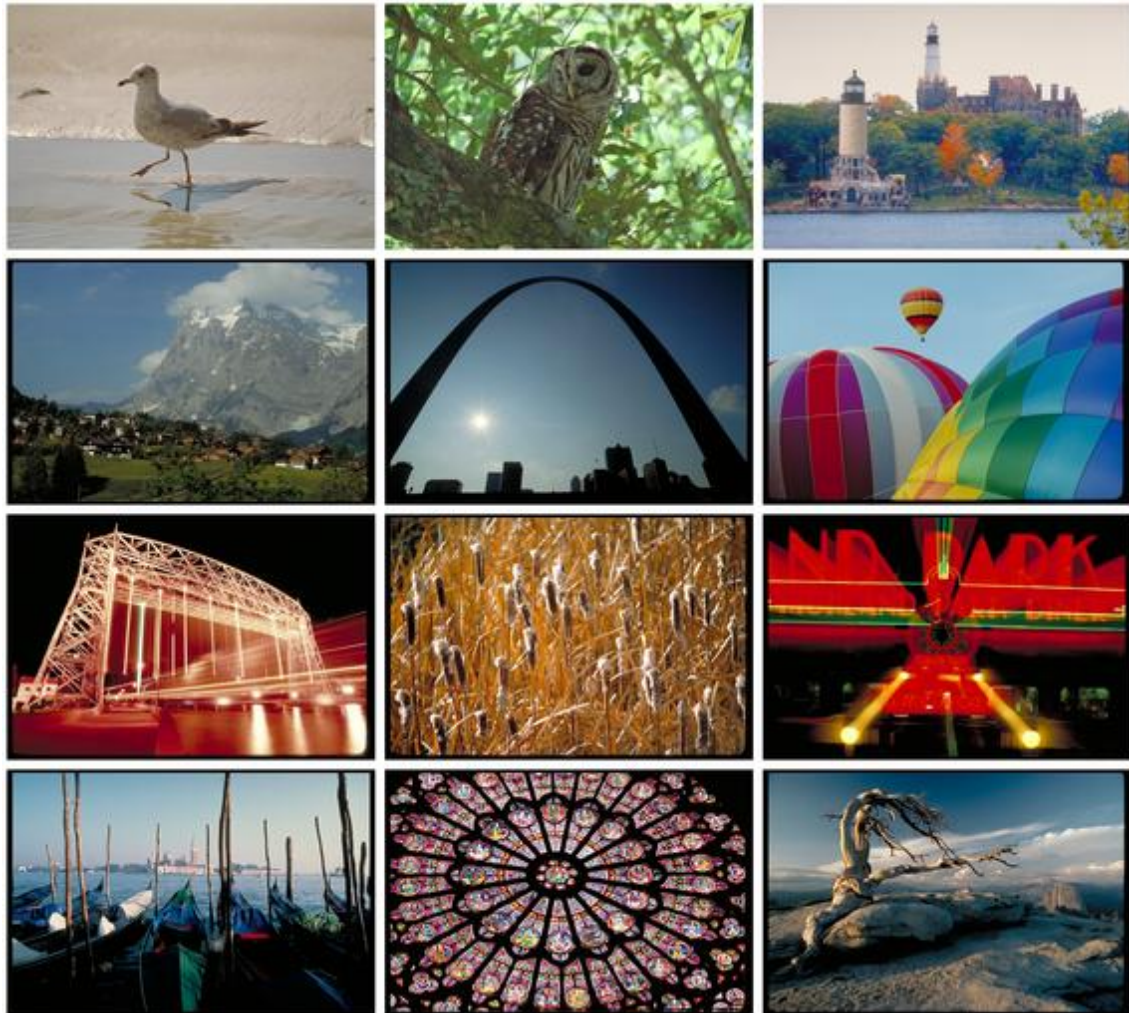
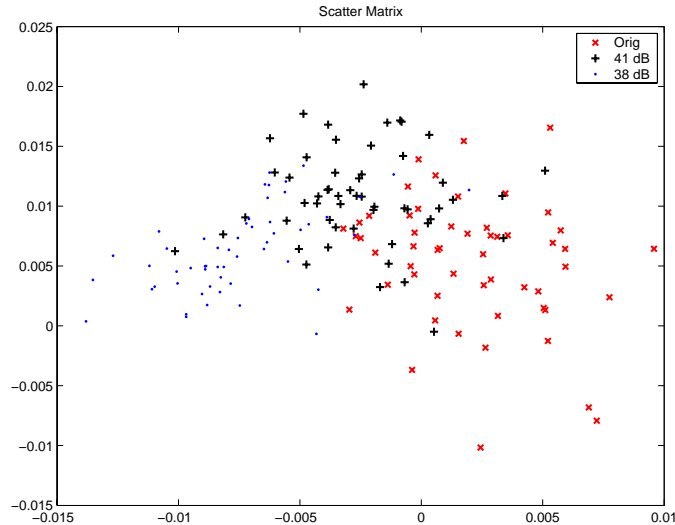


Figure 2. A subset of images used for performance evaluation.



**Figure 3.** Scatter matrix showing three classes. Axes correspond to feature values after projection into the two dimensional eigen-space.

## 4.2. Detection of Stochastic Embedding

Stochastic embedding process can be modeled by noise addition, without loss of generality. Although, the method allows for alternative noise statistics, we use a Gaussian noise in our experiments. We use two different embedding strengths at  $\sigma^2 = 3$  and  $\sigma^2 = 9$ , corresponding to PSNR of  $41dB$  and  $38dB$ , and embedding rates of  $0.84bpp$  and  $0.91bpp$ , respectively<sup>†</sup>.

Training and test sets are formed randomly and contain equal number of images. In addition to its original form, each image is modified by the low and high payload stochastic embedding processes. For each image, mean square error (MSE), mean absolute error (MAE) and weighted mean square error (wmSE) between the image and its compressed versions are computed. Compression is performed with JPEG2000<sup>13</sup> at 95, 90, 85, 80, 70, 60, 50% of the lossless rates.

During training, feature vectors are processed to obtain an optimal projection onto a two dimensional feature space. Then a Bayesian classifier is trained on the reduced features using three classes (namely no embedding, low embedding, high embedding). The scatter plot for these features are seen in Fig. 3. In the test phase, the projection matrix obtained in the training phase is used to reduce the feature vector dimensions. Afterwards, classification is performed using the previously learned parameters. The resulting confusion matrix is seen in Fig. 4.

In summary, 9 cover images out of 54 are mis-labeled as a stego-image, while 13 stego-images are mis-labeled as a cover image. Corresponding false alarm and miss rates are 16.7% and 12%, respectively.

## 4.3. Detection of LSB Embedding

We consider six different embedding strengths for LSB steganography. We set the embedding rate to 0.1, 0.2, 0.4, 0.6, 0.8, and 1.0 bits per pixel (bpp). An embedding rate below one bits per pixel implies that only a subset of pixels are used for embedding. For instance, at 0.2 bpp, we set 20% of total pixel LSBs to corresponding payload bits, flipping an estimated 10% of the image LSBs. All images are embedded at each of these rates.

We use the  $\Psi(\cdot)$  discriminant as defined earlier (rates are obtained with JPEG-LS<sup>14</sup>), and set the cover/stego image decision threshold to 0.1. That is, the algorithm is set to detect images with more than 10% payload.

<sup>†</sup>The spread-spectrum embedding method used in active warden scenarios may also be modeled as Gaussian noise addition. While the payload for spread-spectrum embedding is significantly less than the stochastic embedding, two schemes are equivalent for steganalysis purposes.

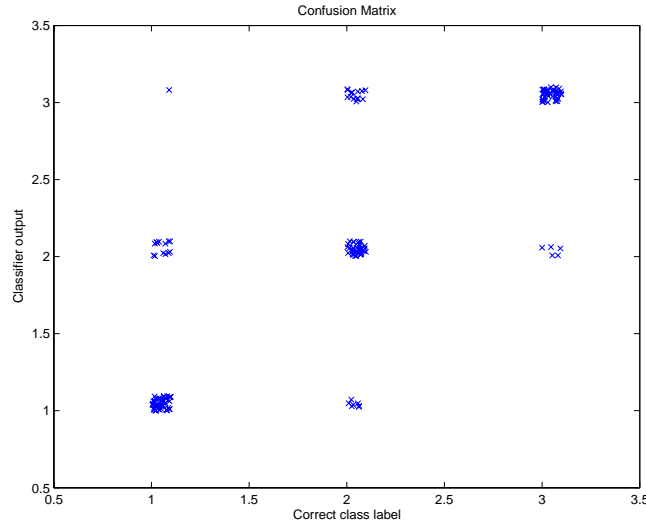


Figure 4. Confusion matrix showing three classes.

In Table. 1, we see that 27% of the cover images are labeled as stego-images, while the miss rate decreases with increasing embedding rate. False positive rate can be traded off for the miss rate by changing the decision threshold. The receiver operating characteristics (ROC) for the overall false alarm and miss rates—as the decision threshold is varied—is seen in Fig. 5.

Table 1. LSB detection results.

Embedding Rate (bpp)	0	0.1	0.2	0.4	0.6	0.8	1.0
Miss/ False Alarm (%)	27 (FA)	35	15	3	0	0	0

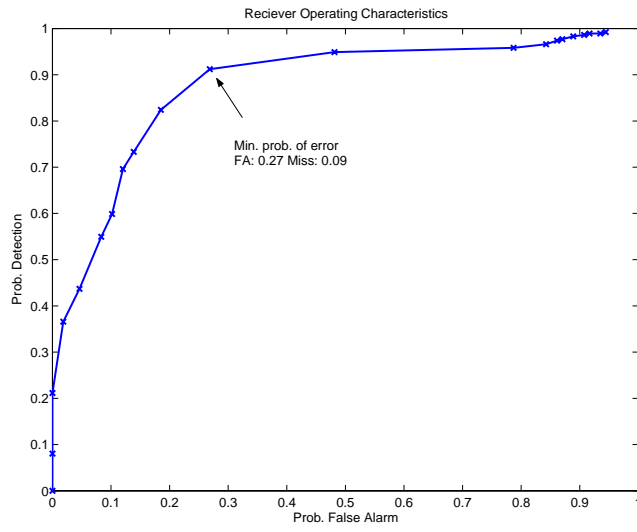
#### 4.4. Effect of pre-processing on detector performance

In Sec. 4.1, we stated the importance of using a representative database for performance evaluation, as steganalysis may be easier for a certain set of images. In general, steganalysis algorithms are more accurate for less noisy images. Equivalently, noisy images tend to generate more false positives. In order to demonstrate this dependence, we applied a number of pre-processing operations to our images. In particular, after converting images to RGB TIFF files, we compressed them with JPEG ( $QF = 75$ ), converted to gray-scale (Matlab *rgb2gray* command), and resized them to  $600 \times 400$  pixels (Matlab *imresize* with *bicubic* interpolation and anti-alias filtering). We tested the LSB steganalysis algorithm with the same threshold. As seen from Table. 2 and Fig. 6, the false alarm and miss rates have been significantly reduced.

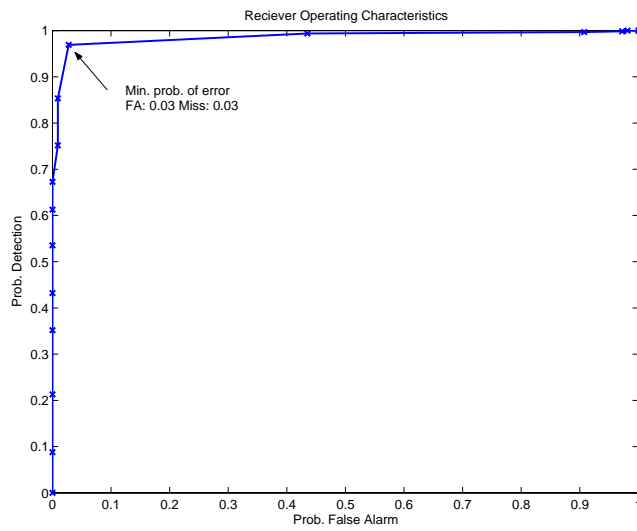
Table 2. LSB detection results with (last row) and without pre-processing.

Embedding Rate (bpp)	0	0.1	0.2	0.4	0.6	0.8	1.0
Miss/ False Alarm (%)	27 (FA)	35	15	3	0	0	0
Miss/ False Alarm (%)	3 (FA)	13	6	0	0	0	0





**Figure 5.** Receiver operating characteristics curve.



**Figure 6.** Receiver operating characteristics curve when cover images are pre-processed by JPEG compression, color to gray conversion, and re-sampling with anti-alias filtering.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed new steganalysis techniques based on rate-distortion arguments. Our techniques are based on the observation that the steganographic algorithms invariably disturb the underlying signal statistics and therefore change the rate-distortion characteristics of the signals. We demonstrated the effectiveness of the proposed approach against least significant bit (LSB) embedding and stochastic embedding algorithms with varying degrees of success. Finally, we highlighted the importance of image database selection on the evaluation of steganalysis algorithms. In our future work, we will continue to analyze the effect of different database selections. In addition, we will investigate the use of non-linear classifiers—such as support vector machines—to improve the performance of our detection algorithms.

## REFERENCES

1. R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE Journal of Selected Areas in Communications* **16**, pp. 474–481, May 1998. Special issue on copyright & privacy protection.
2. I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2002.
3. S. Katzenbeisser and F. A. P. Petitcolas, eds., *Information Hiding: techniques for steganography and digital watermarking*, Artech House, Boston, MA, 2000.
4. G. Simmons, "Prisoners' problem and the subliminal channel," in *CRYPTO83-Advances in Cryptology*, pp. 51–67, 1984.
5. D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, Florida, USA, 1995.
6. A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Florida, USA, 1997.
7. J. Fridrich and M. Goljan, "Digital image steganography using stochastic modulation," in *Proc. SPIE: Security and Watermarking of Multimedia Contents V*, E. J. Delp and P. W. Wong, eds., **EI23**, pp. 191–202, Jan. 2003.
8. A. Westfield and A. Pfitzmann, "Attacks on steganographic systems," in *3rd International Workshop on Information Hiding*, pp. 61–76, 1999.
9. J. Fridrich, M. Goljan, D. Hoge, and D. Soukal, "Quantitative steganalysis of digital images: Estimating the secret message length," *ACM Multimedia Systems Journal* **9**, Sept. 2003.
10. S. Lyu and H. Farid, "Detecting hidden messages using higher-order statistics and support vector machines," in *5th International Workshop on Information Hiding*, 2002.
11. I. Avcibas, N. Memon, and B. Sankur, "Steganalysis using image quality metrics," **12**, pp. 221–229, Feb. 2003.
12. "Kodak PhotoCD images." <ftp://ftp.kodak.com/www/images/pcd>.
13. "Jasper Project Home Page." <http://www.ece.uvic.ca/mdadams/jasper/>.
14. "HP Labs LOCO-I/JPEG-LS Home Page." <http://www.hpl.hp.com/loco/>.