Effective Steganalysis Based on Statistical Moments of Wavelet Characteristic Function

Yun Q. Shi[1], Guorong Xuan[2], Chengyun Yang[2], Jianjiong Gao[2], Zhenping Zhang[2], Peiqi Chai[2], Dekun Zou[1], Chunhua Chen[1], Wen Chen[1]
[1] New Jersey Institute of Technology, Newark, NJ, USA (shi@njit.edu)
[2] Tongji University, Shanghai, P.R. of China (grxuan@public1.sta.net.cn)

Abstract

In this paper, an effective steganalysis based on statistical moments of wavelet characteristic function is proposed. It decomposes the test image using two-level Haar wavelet transform into nine subbands (here the image itself is considered as the LL_0 subband). For each subband, the characteristic function is calculated. The first and second statistical moments of the characteristic functions from all the subbands are selected to form an 18-dimensional feature vector for steganalysis. The Bayes classifier is utilized in classification. All of the 1096 images from the CorelDraw image database are used in our extensive experimental work. With randomly selected 100 images for training and the remaining 996 images for testing, the proposed steganalysis system can steadily achieve a correct classification rate of 79% for the non-blind Spread Spectrum watermarking algorithm proposed by Cox et al., 88% for the blind Spread Spectrum watermarking algorithm proposed by Piva et al., and 91% for a generic LSB embedding method, thus indicating significant advancement in steganalysis.

1. Introduction

Recently, digital watermarking and data hiding have become a vibrant research area. Many watermarking software algorithms can be downloaded freely from the Internet. Terrorists might have seen this as an opportunity to communicate secretly with each other. Thus, various steganalysis methods have emerged as means to deter covert communication by terrorists. Steganalysis is the scientific technology to decide if a medium carries some hidden messages or not and, if possible, to determine what the hidden messages are. In addition to preventing secret communication among terrorists, steganalysis serves a way to judge the security performance of steganography techniques. In other words, a good steganography method should be imperceptible not only to human vision systems, but also to computer analysis.

Owing to the wide diversity of natural images, a large number of data embedding algorithms, and the many, possibly infinite number of, messages, steganalysis turns out to be a tough mission. The basic rationale of steganalysis is that there should have differences between an original cover medium and its stego version (with hidden messages inside the original cover medium). Normally, natural images tend to be continuous and smooth. The correlation between adjacent pixels is strong. Often, the hidden data will be independent to the cover media. The watermarking process may change the continuity, incur random variation or reduce the correlation among adjacent pixels, bit-planes and image blocks. Discovering the difference of some statistical characters between the cover and stego media becomes the key issue in steganalysis.

In [1], a steganalysis method based on the mass center (the first order moment) of histogram characteristic function is proposed. The second, third and fourth order moments are also considered for steganalysis. It is found in our investigation that the performance of method in [1] is not good enough since it adopts only a very small number of features. In [2], Farid proposed to use mean, variance, skewness and kurtosis of coefficients of wavelet subbands as features for steganalysis. Using the statistical moments of characteristic functions of wavelet subbands, our proposed work has achieved substantially superior performance in steganalysis over that achieved by [1] and [2]. Section 2 discusses the wavelet based feature selection. In Section 3, Bayes classifier is introduced to classify the feature vectors. Experimental results are presented in Section 4. Section 5 concludes our work and proposes some future research works.

2. Feature vectors

To decide if an image contains secret messages is equivalent to classifying a given image into two different categories: stego-image or non-stego-image. In this sense, steganalysis is actually a matter of pattern classification in which a key issue is how to select effective features. The features should be sensitive to

the hidden message while not sensitive to other operations such as compression. Besides, the features should be applicable to all kinds of images. Intuitively, it is hardly possible to use one single feature to achieve high correct classification rate. Multi-dimensional (M-D) feature vector should be used. It is better to have each dimension function independently to others. Each image is a sample point in the M-D feature space. Steganalysis has thus become a pattern classification in the M-D feature space.

2.1. De-correlation of wavelet transform

The histogram of a wavelet subband only reflects the statistical distribution of coefficients in the subband. It does not reflect the correlation of the coefficients within this subband. The wavelet transform is well known for its capability of multi-resolution decomposition and coefficients de-correlation. It is known that for discrete wavelet transform different high frequency subbands within one level will be uncorrelated to each other. The features extracted from one high frequency subband are thus uncorrelated to that extracted from another high frequency subband at the same level. Therefore, features from different subbands can form an M-D feature vector with different dimensions most likely uncorrelated to each other. From this point of view, this M-D feature vector will be suitable to represent the image for steganalysis purpose.

2.2 Statistical moments of characteristic function of wavelet subbads

Here, let us consider all of three major types of image data hiding techniques, i.e., spread spectrum, least significant bitplane and quantization index modulation. There is one thing in common for these three types of embedding techniques, i.e., the hidden data can be modeled as an additive signal, which is independent to the cover image. It is well-known that the addition of two independent random signals results in the convolution of two probability density functions (pdf's).

a) Because of the convolution mentioned above, the pdf, hence the histogram, of the stego-image is expected to be more flat than that of the original image.

b) It is well-known that the Fourier transform of a Gaussian distributed pdf is also Gaussian, but the kurtosis of these two Gaussian distributed pdf's is inversely proportional. That is, if the standard deviation of the Gaussian before the Fourier transformation becomes larger, then the standard deviation of the Gaussian distribution after the transformation becomes smaller. Therefore, the statistical moments of the characteristic function are good candidates of features for steganalysis.

c) It is well-known that the wavelet coefficients in the high frequency subbands obey Laplacian-like distribution. It can also be shown that a Lapalacian distribution can be viewed as the sum of two Gaussian distribution with different variances. Hence, the reasoning used in b) of this section can also be applied to wavelet high frequency subbands.

2.3 Proposed M-D feature vector

In our proposed system, a two-level Haar wavelet transform is performed to the image under analysis. Including the image itself (which can be viewed as the subband LL_0), we have nine subbands: LL_0 LL_1 HL_1 LH_1 HH_1 LL_2 HL_2 LH_2 and HH_2. For each subband, the characteristic function is obtained. That is, we calculate the DFT of the histogram of this subband. Then, the first order moment and the second order moment are extracted according to the following equations:

1st order moment: $$M_1 = \frac{\sum_{j=1}^{n} f_j A_j}{\sum_{j=1}^{n} A_j}$$

2nd order moment: $$M_2 = \frac{\sum_{j=1}^{n} f_j^2 A_j}{\sum_{j=1}^{n} A_j}$$

where A_j is the amplitude of j^{th} frequency component, f_j.

In this way, for each subband, we have two features. Totally, we have 18 features which form an 18-D feature vector for steganalysis. Next we will discuss how to classify the images with the proposed features.

3. Bayes Classifier

In addition to feature selection, the design of classifier is another key element in pattern recognition. It affects the classification performance in terms of success classification rate as well as computational complexity, hence, implementation speed. Therefore, the classifier plays an important role in steganalysis. In this paper, we use Bayes classifier because the embedded data basically follow Gaussian distribution or can be approximated by Gaussian distribution. Denote by X_i the 18-D feature vector, where i is the image index. The notations of w_1, w_2 represent the original image set and the stego image set, respectively. The mean vectors and covariance matrixes of w_1 and w_2 are represented by m_1, m_2 and Σ_1, Σ_2, respectively. The Bayes classification can be stated as:

if $P(w_1 / X_i) \geq P(w_2 / X_i)$,

$X_i \in w_1$

else $X_i \in w_2$

where

$$P(w_k / X_i) = \frac{P(w_k) p(X_i / w_k)}{\sum_{m=1}^{2} P(w_m) p(X_i / w_m)}, k = 1, 2$$

and

$$P(X_i / w_1) = N(X_i, m_1, \Sigma_1)$$

$$P(X_i / w_2) = N(X_i, m_2, \Sigma_2)$$

where *N* stands for normal (Gaussian) distribution.

4. Experimental results

Considering a large number of images are necessary for steganalysis if we want the steganalysis to make sense and to be practical, we use the CorelDraw image database [8] as the experimental image set. This database contains 1096 images in total, including images of leisure, place, animal, food, scenery, architecture and so on. Some sample images are contained in Appendix I. In the experiments, we randomly choose 100 images from the 1096 images for training purpose. Then, the remaining 996 images are used for testing purpose. To be reliable, the correct classification rate in steganalysis is obtained by averaging the results obtained in 10 times of such experiments.

4.1. Correct classification rate

The data are embedded into images with three types of data embedding methods, i.e., the non-blind spread spectrum (SS) method by Cox et al. [3], the blind spread spectrum method by Piva et al. [4], and the least significant bit-plane replacement (LSB) [5]. The non-blind SS method by Cox et al. is noted for its strong robustness. The hidden data are a random number sequence obeying Gaussian distribution with zero mean and unit variance. The data are embedded into the 1000 coefficients of global discrete cosine transform (DCT) coefficients of the largest magnitudes. The original cover image is needed for hidden data extraction. The SS method by Piva et al. is blind. That is, it does not need the original cover image for hidden data extraction. It embeds data into 16,000 middle frequency DCT coefficients. It is well-known that the LSB is one type of methods widely used by many commercial data hiding algorithms. A generic LSB data hiding method with embedding rate as 0.3 bpp (bit per pixel) is used in this experiment.

For each data hiding method, 1096 stego-images are generated. Now we have 1096 pairs of images, one is the original image, another is the stego-image. Then an 18-D feature vector as defined above is extracted from each of these images. The 18-D feature vectors corresponding to randomly selected 100 pairs of images are used for training the classifier. The feature vectors corresponding to the remaining 996 pairs of images are used as test images. The correct classification rate is reported by averaging over 10 times of this type of experiments. Two tables in Appendix II contain all of detailed test results. The arithmetically averaged test results are shown below in Table 1.

Table 1. Correct detection rates. (100 pairs of images serve as training set, and the remaining 996 pairs of images as testing set.)

Steganalysis / Watermarking	Proposed steganalysis method	Farid's method: reported in [2]	Farid's method: implemented by authors of this paper
Cox et al.'s SS alpha=0.1	79%	No mention	57%
Piva et al.'s SS	88%	No mention	66%
LSB(0.3 bpp)	91%	43%-90%	63%

From Table 1, it is observed that the correct classification rate achieved by our proposed system is 79% for Cox et al.'s non-blind SS method, 88% for Piva et al.'s blind SS method, and 91% for a generic LSB method,.

Among reported steganalysis investigations, only a very few have conducted experimental works on a reasonably large image set. For instance, only 24 images are used (20 for training and four for testing) in [1]. Therefore, in Table 1, we only list test results achieved by the algorithm reported in [2] for comparison. That is, according to [2], we list some test results achieved by the system proposed by the author of [2] in Table 1. We also used the feature set proposed in [2] and the Bayes classifier described above to conduct the steganalysis experiments on the 1096 CorelDRAW images by ourselves. The test results are also listed in Table 1. It is observed that the system proposed in [2] almost fails in steganalysis of Cox et al.'s SS method. This does not come out as a surprise since the spread spectrum technique is known to be very difficult to be detected since the signal behaves like random noise. In addition, the Cox et al.'s SS method is non-blind, indicating it is more difficult to be steganalyzed than the blind SS method. This has been

verified by the test results for Piva et al.'s method, as listed in Table 1.

It is observed from Table 1 that our proposed system has outperformed that in [2], and has performed reasonably well, though further improvements are necessary.

4.2 Analysis of contributions made by different features in steganalysis of LSB data hiding

To evaluate the contribution of each dimension of the 18-D feature vector to the classification performance, we separately apply each dimension as a one-dimensional (1-D) feature for steganalysis. Table 2 lists the error rate in classification. Here, the generic LSB algorithm with embedding rate of 0.3 bpp is used as the data hiding method. Similarly, only random selected 100 image pairs were used for training and the remaining 996 pairs of images were used for testing.

Table 2. Error rate of 1-D feature compared with the 18-D feature vector.

Order / Subband	1st order moment (%)	2nd order moment (%)
LL_0	49.02	48.81
LL_1	48.09	46.14
LH_1	44.77	44.23
HL_1	44.62	43.82
HH_1	34.62	33.33
LL_2	48.74	44.56
LH_2	47.67	46.73
HL_2	46.64	46.30
HH_2	42.76	41.09
9-D feature vector	11.97	10.94
18-D feature vector	8.75	

It can be observed that the error rates in steganalyzing the LSB method are different for different 1-D features. However, the error rates using only one feature are much larger than that using the 18-D feature vector. In other words, the 18 features collectively perform much better. This is expected.

This type of experiments has also been conducted for Cox et al.'s SS method. It is observed that the error rates obtained by applying each of these 18 features alone are different from that with the LSB method, reported in Table 2. This indicates that these 18 features are complementary in steganalysis of these two different embedding methods.

4.3 Demonstration of feature effectiveness

By using the Bhattacharyya distance feature selection method [7], all of the 18-D feature vectors used in the experiments described in the previous subsection have been reduced to 2-D feature vectors and shown in Figures 1, 2, and 3, respectively. The effectiveness of stegeanalysis is confirmed by the distributions of the 2-D feature vectors associated with the cover images and the stego-images in these figures.

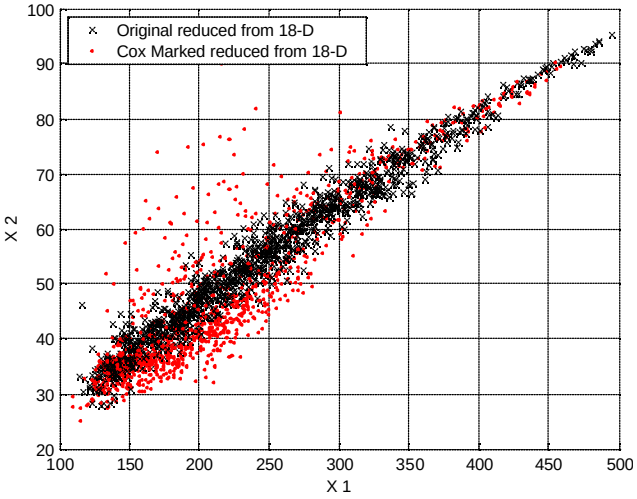

Figure 1. 2-D feature point of cover images and Cox et al.'s SS embedded images (alpha=0.1).

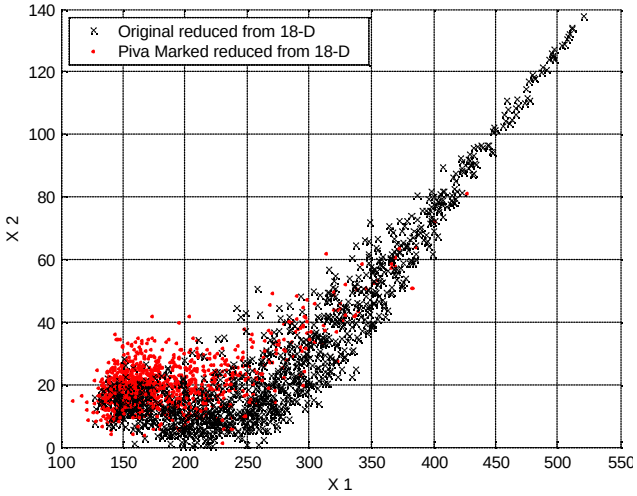

Figure 2. 2-D feature point of cover images and Piva et al's SS embedded images.

It is observed that the stego-images generated by using Cox et al.'s non-blind SS method are the hardest ones to be separated from the cover images, compared with Piva et al.'s blind SS and the generic LSB methods. This verifies what we pointed out previously. On the other hand, the stego-images generated by using the generic LSB method is easiest among these three

methods to be detected from the corresponding cover images.

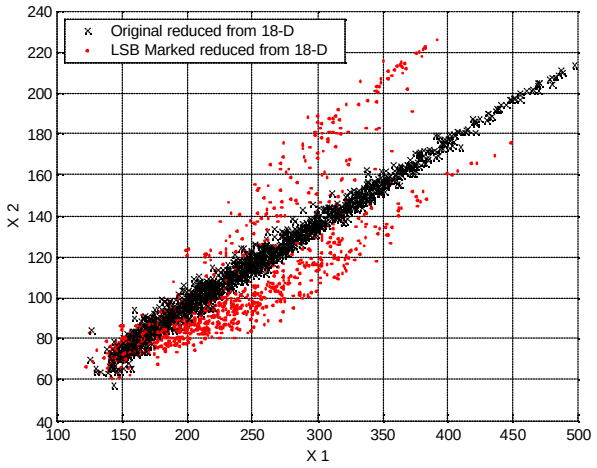

Figure 3. 2-D feature point of cover images and LSB embedded images.

5. Conclusions and future work

This paper proposed a wavelet based steganalysis system. Similar to the area of pattern recognition, the feature selection is the key issue. We choose statistical moments of characteristic function of wavelet subbands as the features, thus constructing an M-dimensional (currently M=18) feature vector. The design of classier is another key issue. We choose to use Bayes classifier in this system. Extensive experimental works have demonstrated that our steganalysis system based on the multi-dimensional feature vector is rather effective. For the non-blind Cox et al.'s SS, the correct detection rate reaches 79%. For the blind Piva et al.'s SS, the correct detection rate reaches 88%. For the generic LSB, we can achieve a correct classification rate of 91%. These results are superior over the existing steganalysis systems reported in the literature, that have been tested in a set of images with a sufficiently large size.

For future work, the feature set will be further improved in order to achieve a higher correct detection rate. For instance, the contribution made by each feature will be examined for various data hiding methods in order to gain more insight. The issues if we should further increase dimensionality, and what dimension will be the optimal are also on the list of our future investigation. Besides, further investigation on more powerful classifiers will be conducted in order to enhance the performance. In particular, artificial neural network will be examined. The consideration is when the embedded signal does not follow Gaussian distribution, the Bayes classifier will not work well. Furthermore, many more image data hiding algorithms will be investigated and tested. The goal is a steganalysis system that can blindly detect stego-images from the original images with a high and reliable success rate and that can handle various images and various image data hiding algorithms.

6. References

[1] J. J. Harmse, "Steganalysis of Additive Noise Modelable Information Hiding," Master Thesis of Rensselaer Polytechnic Institute, Troy, New York. Thesis advisor: W. A. Pearlman. May 2003.

[2] H. Farid, "Detecting hidden messages using higher-order statistical models," *Proc. of the IEEE Int'l. Conf. on Image Processing 02*, vol. 2. New York: IEEE, 2002. 905~908.

[3] I. J. Cox, J. Kilian, T. Leighton and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Processing*, 6, 12, 1673-1687, (1997).

[4] A. Piva, M. Barni, E Bartolini, V. Cappellini, "DCT-based watermark recovering without resorting to the uncorrupted original image," *Proceedings of the 1997 International Conference on Image Processing (ICIP '97)* Volume 1, pp.520-523.

[5] J. Fridrich, M. Goljan, R. Du, "Detecting LSB steganography in color and gray-scale images," *Magazine of IEEE Multimedia, Special Issue on Security*, 2001.8, 22~28.

[6] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Second Edition, John Wiley & Sons, 2001.

[7] G. Xuan, P. Chai and M. Wu, "Bhattacharyya distance feature selection," *Proceedings of the 13th International Conference on Pattern Recognition*, volume 2, pp. 195-199. IEEE, 25-29 Aug. 1996.

[8] CorelDraw Software, www.corel.com.

Acknowledgments

This research is partly supported by New Jersey Commission of Science and Technology via New Jersey Center of Wireless Networking and Internet Security (NJWINS), and by China NSFC, the Research of Theory and Key Technology of Lossless Data Hiding (90304017).

Appendix I. Sample Images from CorelDraw Database.


Place


Animal


Architecture


Food


Leisure


Nature

Appendix II. Detailed test results on three data hiding methods using the proposed method and the method reported in [2].

Table 3. Error rates in steganalysis of three data embedding methods using the 18-D feature vector.

Watermarking / Experiments	Cox et al.'s SS (alpfa=0.1) (%)	Piva et al.'s SS (%)	LSB (0.3 bpp) (%)
1st	(274+162)/(996+996)=21.89	(207+50)/(996+996)=12.9	(106+57)/(996+996)=8.18
2nd	(184+203)/(996+996)=19.43	(180+100)/(996+996)=14.06	(89+78)/(996+996)=8.38
3rd	(224+210)/(996+996)=21.79	(167+67)/(996+996)=11.85	(107+87)/(996+996)=9.74
4th	(191+209)/(996+996)=20.08	(146+95)/(996+996)=12.10	(123+51)/(996+996)=8.73
5th	(240+167)/(996+996)=20.43	(199+36)/(996+996)=11.80	(92+89)/(996+996)=9.09
6th	(249+177)/(996+996)=21.39	(158+64)/(996+996)=11.14	(103+75)/(996+996)=8.94
7th	(212+210)/(996+996)=21.18	(161+89)/(996+996)=12.55	(74+99)/(996+996)=8.68
8th	(189+189)/(996+996)=18.98	(175+75)/(996+996)=12.55	(97+74)/(996+996)=8.58
9th	(262+200)/(996+996)=23.19	(184+69)/(996+996)=12.70	(122+45)/(996+996)=8.38
10th	(194+195)/(996+996)=19.53	(167+73)/(996+996)=12.05	(96+79)/(996+996)=8.79
Average rates	Avg. error rate: 20.79% Avg. correct rate: 79.21%	Avg. error rate: 12.37% Avg. correct rate: 87.63%	Avg. error rate: 8.75% Avg. correct rate: 91.25%

Table 4. Error rate in steganalysis of three data embedding methods using Farid's 72-D feature vector.

Watermarking / Experiments	Cox et al.'s SS (alpha =0.1) (%)	Piva et al.'s SS (%)	LSB (0.3 bpp) (%)
1st	(412+473)/(996+996)=44.43	(353+349)/(996+996)=35.24	(245+544)/(996+996)=39.61
2nd	(481+402)/(996+996)=44.33	(330+362)/(996+996)=34.74	(447+262)/(996+996)=35.59
3rd	(515+349)/(996+996)=43.37	(285+350)/(996+996)=31.88	(429+325)/(996+996)=37.85
4th	(497+375)/(996+996)=43.78	(392+265)/(996+996)=32.98	(392+323)/(996+996)=35.89
5th	(360+525)/(996+996)=44.43	(297+370)/(996+996)=33.48	(183+541)/(996+996)=36.35
6th	(428+415)/(996+996)=42.32	(328+314)/(996+996)=32.23	(299+483)/(996+996)=39.26
7th	(436+457)/(996+996)=44.83	(508+265)/(996+996)=38.81	(246+479)/(996+996)=36.4
8th	(446+443)/(996+996)=44.63	(322+338)/(996+996)=33.13	(285+431)/(996+996)=35.94
9th	(387+494)/(996+996)=44.23	(388+265)/(996+996)=32.78	(232+517)/(996+996)=37.60
10th	(446+452)/(996+996)=45.08	(483+234)/(996+996)=35.99	(287+433)/(996+996)=36.14
Average rates	Avg. error rate: 44.14% Avg. correct rate: 56.86%	Avg. error rate: 34.13% Avg. correct rate: 65.87%	Avg. error rate: 37.06% Avg. correct rate: 62.94%