Steganalysis Based on Moments of Characteristic Functions Using Wavelet Decomposition, Prediction-Error Image, and Neural Network

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

Abstract

In this paper, a general blind steganalysis system is proposed, in which the statistical moments of characteristic functions of the prediction-error image, the test image, and their wavelet subbands are selected as features. Artificial neural network is utilized as the classifier. The performance of the proposed steganalysis system is significantly superior to the prior arts.

1. Introduction

In recent years, digital watermarking has emerged as an increasingly active research area. Information can be hidden into images, videos, and audios imperceptibly to human beings. It provides vast opportunities for covert communications. Consequently, methods to detect covert communications are called for. This task is especially urgent for law enforcement to deter the distribution of children pornography images/videos hidden inside normal images/videos, and for intelligence agencies to intercept communications of enemies. Steganalysis is the art and science to detect whether a given medium has hidden message in it. On the other hand, steganalysis can serve as an effective way to judge the security performance of steganographic techniques. In other words, a good steganographic method should be imperceptible not only to human vision systems, but also to computer analysis.

The huge diversity of natural images and the wide variation of data embedding algorithms make steganalysis a tough mission. However, an original cover medium and its stego-version (with hidden message inside) always differ from each other in some aspects since the cover medium is modified during the data embedding. Some data hiding method introduces a certain pattern in the stego-images. For example, in [1], Fridrich et al. have discovered that the number of zeros in the block DCT domain of a stego-image will increase if the F5 embedding method is applied to generate the stego-image. This feature can be used to determine whether there exist hidden messages embedded with the F5 method. There are some other findings regarding the steganalysis of a particular data hiding method [2, 3]. However, this type of steganalysis cannot cope with the real world since the data embedding method is often unknown in advance. A method designed to blindly detect stego-images is referred to as a general steganalysis method. From this point of view, the general steganalysis methods have more real value for deterring covert communications.

In [4], Farid proposed a general steganalysis method based on image high order statistics. These statistics are based on decomposition of images with separable quadrature mirror filters. The subbands' high order statistics are obtained as features for steganalysis. It can differentiate stego-images from cover images with a certain success rate. In [5], a steganalysis method based on the mass center (the first order moment) of histogram characteristic function is proposed. The second and third order moments are also considered for steganalysis. Compared with [4], its performance has been improved. However, the performance achieved by [5] is still not high enough since it adopts very limited number of features extracted from the test image. This paper proposes to select statistical moments of characteristic functions of the prediction-error image, the test image, and their wavelet subbands as features. Artificial neural network is used as the classifier. The proposed steganalysis system outperforms the existing techniques, say, [4,5] significantly.

The rest of this paper is organized as follows. Section 2 discusses the proposed features. In Section 3, the used neural network classifier is presented. Experimental results are presented in Section 4. Conclusion is drawn in Section 5.

2. Features for steganalysis

Because the dimensionality of image data is normally huge, it is unrealistic to use the image data directly for steganalysis. A feasible approach is to extract a certain amount of data from the image and use them to represent the image itself for steganalysis. In other words, they are features characterizing the image. Different tasks decide the different relation of features with respect to image. In the area of facial recognition, the features should reflect the shape of target faces in an image, i.e. the main content of the image. Minor distortions should not affect the final decision. However, in steganalysis, the main content of an image is not an issue to be considered since human eyes cannot tell the difference between an original image and its stego-version. On the contrary, those minor distortions introduced during data hiding stand up as the first priority. Therefore, the features for steganalysis should reflect those minor distortions associated with data hiding.

2.1. Moments of characteristic function

It is well-known that an image's histogram is essentially the probability mass function (pmf) of the image (only differing by a scalar). Multiplying each component of the pmf by a correspondingly shifted unit impulse results in the probability density function (pdf). Obviously, in the context of discrete Fourier transform (DFT), the unit impulses can be ignored, implying that we can treat pmf and pdf exchangeable. Thus, the pdf can be thought as the normalized version of a histogram. According to [6, pp. 145-148], one interpretation of characteristic function (CF) is that the CF is simply the Fourier transform of the pdf (with a reversal in the sign of the exponent).

Owing to the decorrelation capability of discrete wavelet transform (DWT), the coefficients of different subbands at the same level are kind of independent to each other. Therefore, the features generated from different wavelet subbands at the same level are kind of independent to each other. This property is desirable for steganalysis.

We propose to use the statistical moments of the CFs of both a test image and its wavelet subbands as features for steganalysis, which are defined as follows.

$$M_n = \sum_{j=1}^{(N/2)} f_j^n \left|H(f_j)\right| \Big/ \sum_{j=1}^{(N/2)} \left|H(f_j)\right| \qquad (1)$$

where $H(f_i)$ is the CF component at frequency f_i, N is the total number of points in the horizontal axis of the histogram. Note that we have purposely excluded the zero frequency component of the CF, i.e., $H(f_0)$, from calculating the moments because it represents only the summation of all components in the discrete histogram. For an image, it is the total number of pixels. For a wavelet subband, it is the total number of the coefficients in the subband. In either case, it does not change during the data hiding process. As shown below, its exclusion can enhance moments' sensitivity to data hiding.

2.2. Why moments of characteristic functions?

Denote histogram by $h(x)$, which is the inverse Fourier transform (in the above-mentioned sense) of the CF, $H(f)$. The following formula can be derived straightforwardly.

$$\left|\left(\frac{d^n}{dx^n} h(x)\Big|_{x=0}\right)\right| = \left|(-j2p)^n \int_{-\infty}^{\infty} f^n H(f)df\right| \qquad (2)$$
$$\le 2(2p)^n \int_0^{\infty} f^n \left|H(f)\right| df$$

This is to say that the magnitude of the n-th derivative of the histogram at $x=0$ is upper bounded by the n-th moments of the CF multiplied by a scalar quantity (simply stated below as "upper bounded by the n-th moments of the CF"). Using Fourier translation property, it can be shown that this upper bound is also valid for $x \neq 0$.

Assume the noise introduced by data hiding is additive, Gaussian distributed, and is independent to the cover image, which is valid in general for most data hiding methods, including spread spectrum (SS), least significant bit-plane (LSB), and quantization index modulation (QIM). This assumption leads to that the magnitude of the DFT sequence of the noise caused by data hiding is non-increasing. Obviously, sequence of the magnitude of CF is non-negative. Using the discrete Chebyshev inequality [5,7], we can show that the moments defined in Equation (1) are non-increasing after data hiding.

Combining the above two results, one can derive that the upper bound of the magnitude of the n-th derivative of the histogram will not increase after data hiding. This observation will be graphically illustrated in Section 2.4.

2.3. Prediction-error image

In steganalysis, we only care about the distortion caused by data hiding. It is known that this type of distortion may be rather weak and hence covered by other types of noises, including those due to the peculiar feature of the image itself. In order to enhance the noise introduced by data hiding, we propose to predict each pixel grayscale value in the original cover image by using its neighboring pixels' grayscale values, and obtain a prediction-error image by subtracting the predicted image from the test image. It is expected that this prediction-error image removes various information other than that caused by data hiding, thus making the steganalysis more efficient because the hidden data are usually unrelated to the cover media. In other words, the prediction-error image is used to erase the image content. The prediction algorithm is expressed below [8].

$$\hat{x} = \begin{cases} \max(a,b) & c \le \min(a,b) \\ \min(a,b) & c \ge \max(a,b) \\ a+b-c & otherwise \end{cases} \qquad (3)$$

where a, b, c are is the context of the pixel x under consideration, $\hat{x}$ is the prediction value of x. The location of a, b, c can be illustrated in Fig. 1.

x	***b***
a	***c***

Fig. 1 Prediction context.

2.4. Graphical illustration

In this section, we use some graphs to illustrate the effectiveness of the selected features: moments of CFs. In Fig. 2, an original color image from the CorelDraw image database [9] with serial no. 173037 is shown in the left. Its grayscale image obtained by using irreversible color transform is shown in the middle. The prediction-error image is shown in the right. The histograms of the four subbands at the 1st level Haar wavelet transform are shown in Fig. 3. The zoom in of Fig. 3 is shown in Fig. 4. The CFs of these four subbands are shown in Fig. 5. Note that due to the space limit, these figures are displayed in small size. However, readers are strongly recommended to view the figures, from Fig. 3 to Fig. 8, clearly by using zoom to 500%. In these several figures, the "Orig." means the graph is for original image, while the "cox" stands for stego-image produced by using Cox et al.'s SS method [10]. Two numbers are the 1st order moments of the corresponding CF's from the original and stego-image, respectively. It is observed that the histograms become flatter after data hiding, and this is reflected by the reduced 1st order moments, respectively, thus illustrating the effectiveness of the proposed features.

Similarly, Fig.'s 6, 7 and 8 provide illustration for prediction-error images. Similar observation can be obtained. It is noted that the LL_1 subbands in Fig. 3 and Fig. 6 are rather different, demonstrating the effectiveness of using prediction-error image as analyzed in Section 2.3. This will be further verified by experimental works presented in Section 4.

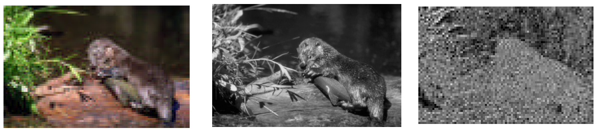
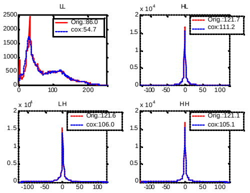


Fig. 2 CorelDraw image no.173037: Original color (left), Original grayscale (middle), prediction-error image (right).



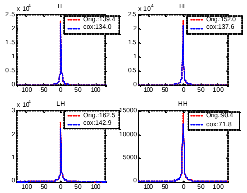
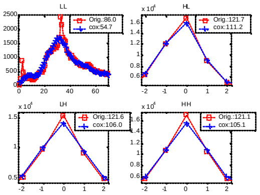


Fig. 3 Original grayscale **Fig.6 Prediction-error**



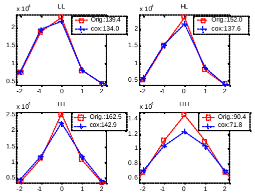
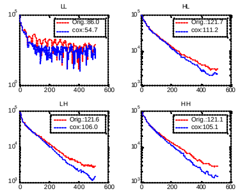


Fig.4 Zoom in of Fig. 3. **Fig.7 Zoom in of Fig. 6.**



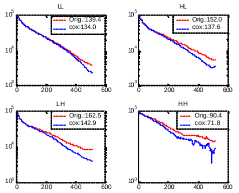


Fig. 5 CF of original grayscale **Fig.8 CF of prediction-error**

2.5. 78-D feature vector

In our work, a test image will be decomposed using a three-level Haar transform. For each level, there are four subbands, resulting in 12 subbands in total. If the original image is considered as level-0 *LL* subband, we have a total of 13 subbands. For each subband, the first three moments of characteristic functions are derived according to Equation (1), resulting in a set of 39 features. Similarly, for the prediction-error image, another set of 39 features can be generated. Thus, a 78-D feature vector is produced for the test image. Our extensive experimental study has shown that using more than three-level wavelet decomposition and including more than the first three order moments do not further improve the steganalysis performance, while leading to higher computational complexity. Hence the 78-D feature vectors are used in our proposed steganalysis system.

3. Neural network classifier

The design of classifier is another key element in steganalysis. In our work, an artificial neural network (NN) [11], specifically, the feed forward NN with back-propagation training algorithm is used as the classifier. It is expected that the powerful learning capability possessed by the NN will outperform the linear classifiers. The number of hidden layers is four. All hidden neurons use the tan-sigmoid function. For the one-neuron output layer, all three activation functions (linear, log-sigmoid, tan-sigmoid) have been tested in the simulation. In the training stage, the outputs of log-sigmoid and tan-sigmoid neuron have larger mean square error (MSE) than the linear neuron output. In the testing stage, the linear neuron output provides higher classification rate than the non-linear outputs. A heuristic explanation for this observation is given below. Because log-sigmoid function squeezes the output into the range from 0 to 1 and tan-sigmoid function squeezes the output into the range -1 to 1, more training exemplars or testing patterns may lie on the wrong side at the output. Therefore, a reasonable structure is composed of four tan-sigmoid neuron hidden layers and one linear neuron output layer. In the back-propagation training, the computation programming is based on the neural network toolbox of Matlab 6.5.

4. Experimental results

To evaluate the performance of the proposed steganalysis system, we use all the 1096 sample images included in the CorelDRAW Version 10.0 software CD#3 for experiments [9]. It contains pictures of Nature, Ocean, Food, Animals, Architecture, Places, Leisure and Misc. The following five typical data hiding methods are used in experiments: Cox et al.'s non-blind SS [10] ($a = 0.1$), Piva et. al's blind SS [12], Huang and Shi's 8 by 8 block SS [13], a generic QIM [14] (0.1 bpp (bit per pixel)), and a generic LSB (0.3 bpp, both the pixel position used for embedding data and the to-be-embedded bits are randomly selected). For each image in the CorelDRAW image database, five stego-images are generated with these five data hiding methods, respectively. For all the data hiding methods, different random signals are embedded into different images. The evaluation of the proposed steganalysis system is hence more general.

At first, we evaluate the system with each one of the five data hiding methods at a time. Randomly selected 896 original images and the corresponding 896 stego-images are used for training. The remaining 200 pairs of the cover images and stego-images are put through the trained neural network to evaluate the performance. The detection rate is defined as the ratio of the number of the correctly classified images with respect to the number of the overall test images. The 10-time average detection rates are listed in Table 1.

Next, we combine the five data hiding methods to evaluate the blind steganalysis ability of the proposed system. Similarly to the above, we start with 1096 6-tuple images. Each 6-tuple images consists of an original image, and the five stego-images generated by the five data hiding methods. We then randomly selected 896 6-tuple images for training, and use the remaining 200 6-tuples for testing. Again, the 10-time average correct detection rates are listed in Table 1. We also evaluate the test results of Farid's method [4] and Harmsen's method [5] under the same circumstances. Note that the NN converges to MSE<0.05 with our proposed features in less than 1,500 iterations during the training, while the NN with either Farid's or Harmson's features does not converge even after 100,000 iterations. Therefore, the Bayes classifier [5] is used to obtain the detection rates for these two methods listed in Table 1. It can be observed that the proposed system outperforms both Farid's and Harmsen's methods at a significant advantage.

Table 1 Testing results.

Detection rate	Farid [4]	Harmsen [5]	Proposed
Cox et al.'s SS	64.9%	78.3%	98.1%
Piva et. al's SS	87.8%	79.4%	98.7%
Huang and Shi' block SS	76.1%	81.5%	98.8%
Generic QIM (0.1 bpp)	99.7%	75.7%	99.0%
Generic LSB (0.3 bpp)	71.9%	56.5%	98.9%
5 methods combined	68.9%	72.8%	98.7%

Thirdly, to further evaluate our system, a data hiding method, which has not been used in the training process, is tested. We apply Hide4PGP [15] to 200 randomly selected CorelDraw images. The detection rate is 99.5%.

Fourthly, to evaluate the effectiveness of using the prediction-error image, we conduct the same evaluation as stated above to the first 39 features (generated from the test images) and the second 39 features (obtained from the prediction-error images), separately. Table 2 contains the comparison results, which has demonstrated the effectiveness of using the prediction-error images. That is, the performance of using features obtained from the prediction-error images is more effective than that obtained from the test images. This is expected as analyzed above.

Table 2 Effectiveness comparison of features from original images and features from prediction-error images.

Detection rate	39D (test mage)	39D(prediction -error image)
Cox et al.'s SS	96.2%	96.6%
Piva et. al's SS	95.2%	98.8%
Huang and Shi's block SS	95.4%	97.9%
Generic QIM (0.1 bpp)	97.9%	98.7%
Generic LSB (0.3 bpp)	94.5%	98.7%
5 methods combined	94.9%	98.4%

Finally, the effectiveness of using the neural network is evaluated. We conduct experiments with our proposed 78-D feature vectors but using the Bayes classifier and the neural network, respectively, for the five data hiding methods individually and jointly. Table 3 contains detection rate for Cox et al.'s SS data hiding method and for the combined testing. Comparing with the results obtained with the Bayes classifier, a 3% to 4% increase in terms of detection rate has been achieved by using the proposed neural network.

Table 3 Comparison of neural network with Bayes classifier.

Detection rate	Bayes classifier	Neural network
Cox et al.'s SS	95.2%	98.1%
5 methods combined	94.6%	98.7%

5. Conclusion

In this paper, a novel general steganalysis system is proposed. Our contributions are summarized below.

a) Statistical moments of wavelet characteristic functions (CF's) are proposed to be used for steganalysis for the first time. Our theoretical analysis and experimental work have pointed out that the moments of wavelet CF's can reflect the differentiation property of the associated histograms, hence, reflecting sensitively the changes caused by data hiding. b) Excluding zero frequency component of CF's from the calculation of moments has improved the effectiveness of moments in steganalysis. Our experimental works have shown more than three-percept increase in detection rate. c) Prediction-error images are able to enhance the changes caused by data hiding by reducing the effect caused by the diversity of natural images. d) Artificial neural network performs better in steganalysis than Bayes classifier due to its powerful learning capability. e) Our combined steganalysis approach has pointed out a promising way towards blind and practically powerful steganalysis. f) Our experiments are conducted over a large number of images, which is necessary for steganalysis. g) Our proposed steganalysis system has demonstrated a significant performance improvement over the prior-arts.

6. References

[1] J. Fridrich, M. Goljan and D. Hogea, "Steganalysis of JPEG Images: Breaking the F5 algorithm", *5th Information Hiding Workshop*, 2002, pp. 310-323.

[2] J. Fridrich, M. Goljan and R. Du, "Detecting LSB steganography in color and gray-scale images", *Magazine of IEEE Multimedia Special Issue on Security*, Oct.-Nov. 2001, pp. 22-28.

[3] R. Chandramouli and N. Memon, "Analysis of LSB based image steganography techniques", *Proc. of ICIP* 2001, Thessaloniki, Greece, Oct. 7–10, 2001.

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

[5] J. Harmsen, *Steganalysis of Additive Noise Modelable Information Hiding*, MS thesis, Rensselaer Polytechnic Institute, NY, thesis advisor William Pearlman, April 2003.

[6] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd Edition, Reading, MA: Addison-Wesley Publishing Company, 1994.

[7] D. S. Mitrinovic, J. E. Pecaric and A. M. Fink, *Classical and New Inequalities in Analysis*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1993.

[8] M. Weinberger, G. Seroussi, and G. Sapiro, "LOCOI: A low complexity context-based lossless image compression algorithm," *Proc. of IEEE Data Compression Conf.* 1996, pp.140-149.

[9] www.corel.com

[10] 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).

[11] C. M. Bishop, *Neural Network for Pattern Recognition*, Oxford, New York, 1995

[12] A. Piva, M. Barni, E Bartolini, V. Cappellini, "DCT-based watermark recovering without resorting to the uncorrupted original image", *Proc. ICIP 97*, vol. 1, pp.520

[13] J. Huang and Y. Q. Shi, "An adaptive image watermarking scheme based on visual masking," *IEE Electronic Letters*, vol. 34, no. 8, pp. 748-750, April 1998.

[14] B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," *Proceedings of IEEE MMSP 1998*, pp273 – 278.

[15] Available at http://www.jjtc.com/Security/stegtools.htm