Steganalysis of Images based on Randomness Metrics

Tao Zhang Xijian Ping

rbfn@sina.com

Department of Information Science
University of Information Engineering
Zhengzhou, P.R.China, 450002

Abstract

This paper introduces a steganalytic technique that can detect the existence of secret messages embedded by randomly scattered LSB replacement in raw lossless compressed images. This technique is based on the analysis of the difference between randomness metrics of two binary sequences formed by concatenating the least significant bits of carrier image and stego-image. A logistic regression model is constructed to determine the existence of secret messages. Experimental results show that for gray-level images even if secret message capacity is as less as 0.4 bit per pixel, it is possible to achieve a high degree of detection reliability.

1 Introduction

Information hiding techniques are gaining worldwide attention due to the increasing popularity of information technology [1]. There are two main branches: steganography and digital watermarking. As a new way of covert communication, the main purpose of steganography is to convey messages secretly by concealing the very existence of messages [2]. In steganography the carrier image and the stego-image is visually indiscernible. The basic requirement for steganography is undetectability; in addition, embedding capacity should also be considered. However, security often conflicts with embedding capacity. Studies on steganalysis of images can evaluate the security of a given steganographic tool and promote the presentations of more secure steganographic algorithms.

Detection of secret messages in images is usually broken down into two areas: signature detection and blind detection. N. Johnson [2] made a careful analysis of signatures introduced by current steganographic software. J. Fridrich [3] introduced a steganalytic technique that can be successfully used for raw high-color-depth images with randomly scattered messages. N. D. Memon [4] constructed a steganalyzer that can classify the embedded and non-embedded using multivariate regression on the selected image quality metrics.

This paper focuses on stego-only attack of steganography algorithm of LSB randomly scattered insertion, i.e., determining the existence of secret messages only by computer analysis of stego-images.

2 Randomness Metrics

The randomness of binary sequences is an abstract concept and usually described using probability model. Kolmogorov [5] defined the amount of randomness (Kolmogorov-complexity) of a binary sequence as the length of the shortest program for a universal Turing-machine that generates the sequence. A sequence can be considered “random” if one of the shortest descriptions is the sequence itself.

In the application of random bit generators as the secret-key source in cryptography, it is often necessary to decide whether the output of the given generator is "random" enough. When no theoretical proof based on the device's physical structure can be given, such a decision must be based on an observed sample output sequence of a given length *N*. Therefore, several kinds of statistical tests are introduced [6]. A statistical test is typically implemented by specifying an efficiently computable test function that maps the binary sequences to the real number set. Usually, the test function is chosen such that the test statistics is distributed according to a well-known probability distribution, most often the normal distribution or the chi-square distribution with *d* degrees of freedom. Given a significance level of test, we can decide the range of the test statistics when the sequence is a random sequence. On the other hand, the more the statistics deviate from this range, the worse the randomness of the sequence is. Therefore, we will construct randomness metrics based on those statistical test statistics. By experimental comparison we select Maurer's universal statistical test and runs test to construct randomness metrics. See [6-7] for the definitions of those statistical tests.

We select Maurer's approximation information entropy x_e and runs test statistic x_r as the randomness metrics of binary sequences. The more "random" the binary sequence, the bigger x_e and the smaller x_r; and vice versa. Note that there are an upper limit for x_e and a lower limit for x_r respectively.

By concatenating the least significant bits of each row of an image, we can get a binary sequence l^N, called LSB sequence. Experimental results show that for raw lossless compressed images more than 95 percent of them can not pass the statistical tests with significance level of 0.05. This fact indicates that their randomness are much weaker than that of a true random binary sequence. Therefore, we can utilize the randomness metrics to describe those differences on the randomness of LSB sequences between carrier image and stego-image quantitatively.

3 Detection of Secret Messages

Though it is the simplest way of steganography, LSB embedding is still one of the most practical algorithms due to its large embedding capacity, easy implementation and hard detection. We embed messages in images using the same methods as used in [3]: select a portion of points randomly in the LSB plane of the image, and replace all the bits on those points with the secret message to be embedded.

Define the embedding ratio β ($0 \le \beta \le 1$) as the proportion of the size of secret messages embedded to the maximum embedding capacity. It should be noted that secret messages are compressed and encrypted prior to embedding.

3.1 The Basis of Our Steganalytic Technique

The main idea of LSB embedding algorithm is that the LSB sequence l^N can be considered as random binary sequence and replacing l^N with encrypted secret messages will not bring any visual difference between carrier image and stego-image. However, statistical tests on LSB binary sequences of a large amount of images show

that most of them can not pass those tests with a significance level of 0.05, that is, they can not be viewed as true random binary sequences.

Experimental results also show that owing to the randomness of test messages and embedding positions the LSB sequence of stego-image l_1^N exhibits much stronger randomness than that of carrier image l^N. Describing quantitatively using randomness metrics introduced in section 2, generally, we have:

$$x_e(l^N) < x_e(l_1^N),\ x_r(l^N) > x_r(l_1^N) \quad (1)$$

Further analysis reveals that changes on the randomness metrics brought by embedding secret messages are closely related with the embedding ratio. Besides, those changes are also related to $x_e(l^N)$ and $x_r(l^N)$.

Without carrier image (stego-only attack) we can not find out those changes on randomness metrics of LSB sequences. However, we have noticed that if an image already contains a certain amount of secret message, embedding another test message in it will not modify the randomness metrics of the LSB sequence significantly; On the other hand, if the image does not contain a secret message, the randomness metrics of the LSB sequence will change significantly after embedding a test message. Thus, we can compute the Maurer entropy x_e and runs test statistic x_r of the LSB sequence of the image to be tested at first, and after embedding a curtain proportion of test messages using the same embedding method, compute Maurer entropy x_e^1 and runs test statistic x_r^1 again. A logistic regression model is used to describe the close relationship between the embedding ratio and x_e, x_e^1, x_r, x_r^1. Because of the predictability of regression model, we can predict the embedding ratio of secret messages in any image and consequently determine the existence of secret messages.

3.2 Logistic Regression Model Based on Randomness Metrics

The steganalytic technique proposed in this paper is based on changes on randomness metrics of LSB sequences. In this section, decision-making rules based on randomness metrics will be constructed and used to determine the existence of secret messages in the image. Logistic regression model [8] is selected in this paper to construct decision-making rules.

Let $\boldsymbol{Y}$ be the embedding ratio, x_e and x_r denote the Maurer's approximation information entropy and runs test statistic of LSB sequence of the image to be tested, respectively, x_e^1 and x_r^1 denote those of LSB sequence of the image after embedding test messages, respectively. The logistic regression model we used can be expressed as follows:

$$\boldsymbol{Y} = \frac{\boldsymbol{e}^{\vec{\boldsymbol{X}}\vec{\boldsymbol{b}}^{\boldsymbol{T}}}}{1+\boldsymbol{e}^{\vec{\boldsymbol{X}}\vec{\boldsymbol{b}}^{\boldsymbol{T}}}} \quad (2)$$

in which

$$\vec{X} = (1,\ x_e,\ x_e^1, \lg(x_r), \lg(x_r^1),\ x_e / x_e^1,\ \lg(x_r)/\lg(x_r^1)),\ \vec{b} = (b_0, b_1, ..., b_6).$$

4 Experimental Results

We test our steganalytic technique on an image database that contains 350 raw lossless compressed gray-scale images of 512*512 pixels. Those images in the database are collected from USC-SIPI image database, RPI image database and the website of KODAK company , in which 250 images are used as training data for fitting the logistic regression model, and the rest 100 images are used to test the predictive ability of regression model. Secret messages and test messages are all cut randomly from a piece of cipher-text. Experimental procedures are listed below:

4.1 Compute The Parameters of The Model

First, we embed variant size of secret messages in 250 training images. Ten embedding ratios are $\beta = 10\%, 20\%, ..., 100\%$, respectively. Then, compute the Maurer entropy and runs test statistic of LSB sequences of those images (L=8, Q=2560, K=30208 for computing Maurer entropy). Second, embed test messages in carrier images and stego-images, and compute Maurer entropy and runs test statistic again. Consider the embedding ratio as the probability of the existence of secret messages in the image, and fit the logistic regression model defined in equation (2). We use SPSS statistical software package to compute the parameters of the model.

The logistic regression model constructed can be used to predict the embedding ratio of secret messages in any images and consequently be used to determine the existence of secret messages. The embedding ratio also indicates the possibility of existence of secret messages.

4.2 Parameter Optimization

Change the test ratio(the ratio of test message size to the maximum LSB embedding capacity) to the following value: 0.01, 0.02, 0.03, 0.04, 0.08, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Using the area under the ROC curve and the false positive /negative rate (short for FPR/FNR, respectively) as criterion for evaluating the performance of the regression model, experimental results show that when the testing ratio is equal to 0.5 the performance of classifier is the best. Especially when the embedding ratio is between 0.4 and 0.8, the performance of the classifier is improved significantly; when the embedding ratio is above 0.9 or under 0.3, the size of test messages has no significant effect on the performance of the classifier.

When the testing ratio is equal to 0.5 the parameters of the model are: b_0=2917.68, b_1=472.38, b_2=-444.59, b_3=7.25, b_4=-7.29, b_5=-3101.30, b_6=-15.12. The model can predict the embedding ratio of secret messages in any images and consequently be used to classify the embedded and the non-embedded by selecting a simple threshold of the predicted embedding ratio. Figure 1 shows the ROC curves of the regression model, in which from left upper to right lower it in turn depicts the ROC curve when the embedding ratio is 100%, 90%, …, 10%. The FPR/FNR listed in Table I is the value of false positive /negative rate while the false positive rate is equal to the false negative rate.

4.3 The Predictive Ability of Regression Model

The following two experiments are designed to test the predictive ability of the logistic regression model:

1. Embed different embedding ratios of secret messages from those of section 4.1 in 250 training images first, and then utilize the model constructed in section 4.1 to predict the embedding ratio of secret messages and consequently classify the embedded from unembeded images. The embedding ratios are: 5%,15%,…,95%;

2. Embed the same embedding ratios of secret messages as those of section 4.1 in 100 test images, and then utilize the model constructed in section 4.1. to predict the embedding ratio of secret messages and consequently classify the embedded from unembeded images.

(a): results for variant embedding ratios: see Figure 2 and Table II; (b): results for test image database (100 images): see Figure 3 and Table III.

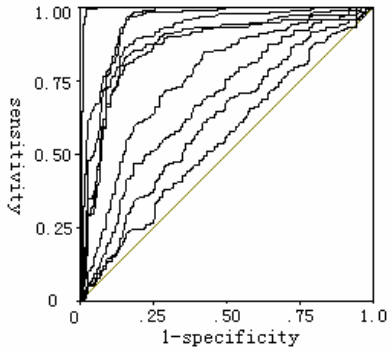

Figure 1.

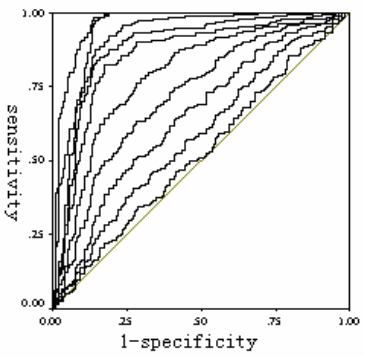

Figure 2.

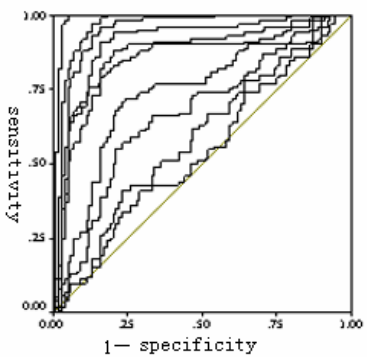

Figure 3.

Table I

Embedding Ratio	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Area Under ROC	0.990	0.948	0.935	0.898	0.874	0.866	0.757	0.677	0.610	0.550
FPR/FNR(%)	2.4	13.2	12.8	14.8	16.4	18.4	28.8	37.6	41.6	46.8

Table II

Embedding Ratio	0.95	0.85	0.75	0.65	0.55	0.45	0.35	0.25	0.15	0.05
Area Under ROC	0.963	0.941	0.911	0.883	0.851	0.789	0.716	0.645	0.579	0.523
FPR/FNR(%)	10	12.8	14.4	15.6	18	27.2	34	39.2	44.8	49.2

Table III

Embedding Ratio	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Area Under ROC	0.983	0.960	0.941	0.900	0.857	0.862	0.738	0.665	0.584	0.537
FPR/FNR(%)	3	9	12	14	19	17	28	34	46	49

4.4 Analysis of Experimental Results

From Figure 1, 2, 3 and Table I, II, III we can see that even if the embedding ratio β is as less as 0.4, it is possible to achieve a high degree of detection reliability. Moreover, The FPR/FNR listed in Table I,II,III is the values of FPR/FNR while the FPR is equal to the FNR. In many cases, people would pay more attention to FNR. From the ROC curves depicted in Figure 1, 2, 3 we can see most of them are close to the upper part of the rectangle; therefore, we can get a lower FNR while maintaining a certain FPR.

From the comparison of ROC curve and FPR/FNR between section 4.1 and section 4.2 (a), (b), we can conclude that the logistic regression model has an unusual ability to predict for unknown images and variant embedding ratios.

5 Conclusion

Starting from randomness metrics of binary sequences, we proposed a steganalytic technique that can detect the existence of secret messages embedded by randomly scattered LSB replacement in raw lossless compressed images. Experimental results show that for gray-level images even if secret message capacity is as less as 0.4 bit per pixel, it is possible to achieve a high degree of detection reliability.

References

[1] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn: “Information Hiding – A Survey”, Proceeding of IEEE, vol. 87, no. 7, pp. 1062-1078, June 1999.

[2] N. F. Johnson, S. Jajodia: “Steganalysis of Images Created Using Current Steganography Software”, LNCS Vol.1525, pp. 273-289, Springer-Verlag, 1998.

[3] Jiri Fridrich, M. Long: “Steganalysis of LSB Encoding in Color Images”, pp. 1279-1282, ICME 2000.

[4] N. D. Memon, et al.: “Steganalysis Based on Image Quality Metrics”, SPIE Vol. 4314, Jan. 2001.

[5] A. N. Kolmogorov: ”Three Approaches to the Quantitative Definition of Information”, Problemy Peredachi Informatsii, Vol. 1, No. 1, pp. 3-11, 1965.

[6] A. Menezes, P. van Oorschot, and S. Vanstone: Handbook of Applied Cryptography, CRC Press, 1996.

[7] U. M. Maurer: “A Universal Statistical Test for Random Bit Generations”, Journal of Cryptology, Vol. 5, No. 2, 1992, pp. 89-105.

[8] David W. Hosner, S.Lemeshow: Applied Logistic Regression, John Wiley & Sons, New York, 1989.