Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium, Volume 21, No. 1, ISSN 1726-9679
ISBN 978-3-901509-73-5, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2010
Make Harmony Between Technology and Nature, and Your Mind will Fly Free as a Bird
Annals of DAAAM International

RESEARCH ON ESTIMATION LENGTH OF HIDDEN MESSAGE

**RACUCIU, C[iprian] I[ulian]; MIHAILESCU, M[arius] I[ulian];
GARBAN, V[alentin] & PRAOVEANU, I[osif]**

Abstract: *In this paper we will discuss a strategy for estimation the length of hidden message from stego-images. A brief presentation of attacks (steganalysis) will be explained, which are based on statistical parameters of the image that is being analyzed and signatures which are embedded information in the images. This poses a threat to forensic analysts; the next step will lead us in methods for determining the existence of and potential locations of hidden information.*

Key words: *stego-image, steganography, stegnalysis, hidden message, embedded information*

1. INTRODUCTION

Steganography represent the art of *covered*, or *hidden, writing.*

Steganography hides the covert message but not the fact that two parties are communicating with each other. Using a transport in which the stego process involves placing a hidden message, named *carrier*. The workpaper shows that the secret message is embedded with the carrier to form the *stego environment.*

Using a stego key it can be used for encryption of the hidden message and/or for randomization in the stego system. Thus, we can resume:

stego_environment = hidden_message + carrier + stego_key

Steganalysis, the detection of this hidden information, is an inherently difficult problem and requires a thorough investigation.

Steganalysis is a relatively new branch of research. While steganography (which is somewhat different from watermarking) deals with techniques for hiding information, the goal of steganalysis is to detect and/or estimate potentially hidden information from observed data with little or no knowledge about the steganography algorithm and/or its parameters. It is fair to say that steganalysis is both an art and a science. The art of steganalysis plays a major role in the selection of features or characteristics a typical stego message might exhibit while the science helps in reliably testing the selected features for the presence of hidden information.

While it is possible to design a reasonably good steganalysis technique for a specific steganography algorithm, the long term goal must be to develop a steganalysis framework that can work effectively at least for a class of steganography methods if not for all. Clearly, this poses a number of mathematical challenges and questions.

2. DATA HIDING

As long as people have been able to communicate with one another, there has been a desire to do so secretly. Two general approaches to covert exchanges of information have been: communicate in a way understandable by the intended parties, but unintelligible to eavesdroppers; or communicate innocuously, so no extra party bothers to eavesdrop. Naturally both of these methods can be used concurrently to enhance privacy. The formal studies of these methods, cryptography and steganography, have evolved and become increasingly more sophisticated over the centuries to the modern digital age.

Methods for hiding data into cover or host media, such as audio, images, and video, were developed about a decade ago.

Steganography generally is subjected to less vicious attacks, however as much data as possible is to be inserted. Additionally, whereas in some cases it may actually serve a watermarker to advertise the existence of hidden data, it is of paramount importance for a steganographer's data to remain hidden.

3. LSB INFORMATION

An early method used to detect LSB hiding is the x^2 (chi-squared) technique, later successfully used by Provos' stegdetect for detection of LSB hiding in JPEG coefficients.

We first note that generally the binary message data is assumed to be independent and indentically distributed with the probability of 0 equality to the probability of 1. If the hider's intended message does not have these properties, a wise steganographer would use an entropy coder to reduce the size of the message; the compressed version of the message should fulfill the assumptions. Because 0 and 1 are equally likely, after overwriting the LSB, it is expected that the number of pixels in a pair of values which share all but the LSB are equalized.

Although, we would expect these numbers to be close before hiding, we do not expect them to be equal in typical cover data.

Due to this effect, if a histogram of the stego data is taken over all pixel values (e.g. 0 to 255 for 8-bit data), a clear "steplike" trend can be seen. We know then exactly what the histogram is expected to look like after LSB hiding in every pixel (or DCT coefficient). The x^2 test is a goodness-of-fit measure which analyzes how close the histogram of the image under scrutiny is to the expected histogram of that image with embedded data. If it is "close", we decide it has hidden data, otherwise not. In other words, x^2 is a measure of the likelihood that the unknown image is stego.

4. IMPROVED DIFFERENCE IMAGE HISTOGRAM STEGANALYSIS

For detection and estimation for length of hidden message, difference image histogram algorithm was primarily based on the statistical hypothesis that for natural images

$$\alpha_i \approx \gamma_i \tag{1}$$

and for a stego-images with the LSB plane fully embedded

$$\alpha_i \approx 1 \tag{2}$$

Obviously, the hypotheses given in Equations (1) and (2) will affect the precision of the Difference image histogram method. Once in these hypotheses there exists some initial bias, the estimate value via the Equation (1) will not be reliable. When the embedding ratio is low, the bias of these hypotheses will lead the incorrect decision, and if there are no embedding messages in images, the false alarm rate is high. Table 1 will show the mean and variance of the γ_i to α_i value.

With the increase in *i* the variance increases and the mean begins to deviate from 1. In some cases the detection lead to an incorrect decision of estimating more than 1% embedding, for the normal images.

5. IMPROVED DIFFERENCE IMAGE HISTOGRAM ALGORITHM

The Improved Difference Image Histogram algorithm consists in several steps, as follows:
Input is represented by a set of BMP images for detecting;
Output the embedded ratio estimate $p_{modified}$ *for each image;*
Step 1, from the set of the images, we select one single image;
Step 2, we obtain difference image histogram of the image before (h_i) and after flipping the LSB planes to „zero" (g_i);
Step 3, for steps 4 to 8 do for each value of i = 0,1,2;
Step 4, we calculate the statistical values for the image i.e. $\alpha_i = \frac{a_{2i+2,2i+1}}{a_{2i,2i+1}}, \beta_i = \frac{a_{2i+2,2i+3}}{a_{2i,2i}-1}$ and $\gamma_i = \frac{g_{2i}}{g_{2i+2}}$, where the co-efficient which represent the transition can be estimated using the following equation:

$$a_{0,1} = a_{0,-1} = \frac{g_0 - h_0}{2_{g0}}, \tag{3}$$

$$a_{2i,2i} = \frac{h_{2i}}{g_{2i}}, \tag{4}$$

$$a_{2i,2i-1} = \frac{h_{2i-1} - a_{2i-2,2i-1} g_{2i-2}}{g_{2i}}, \tag{5}$$

$$a_{2i,2i+1} = 1 - a_{2i,2i} - a_{2i,2i} - 1. \tag{6}$$

Step 5, we obtain de value „p" from the root of the below equation qhose absolute value is smaller.

$$2d_1p^2 + (d_3 - 4d_1 - d_2)P + 2d_2 = 0 \tag{7}$$

in which $d_1 = 1 - \gamma_i, d_2 = \alpha_i - \gamma_i, \ d_3 = 1 - \beta_i$;
Step 6, calculate the value $\alpha_i(0)$ which represents the estimation of α_i for zero embedded message length using the following equation

$$\alpha_i = \frac{2e_1^2 - 2e_1e_2 - (e_2 + e_3)(1 - e_1)}{2e_1^2} \tag{8}$$

in which $e_1 = 1 - p, e_3 = 1 - \alpha_i \ and \ e_3 = 1 - \beta_i$;
Step 7, we calculate the initial bias „ε" with the formula:

$$\varepsilon = \gamma_i - \alpha_i(0) \tag{9}$$

Step 8, now we will subtract the error „ε" from the *p* to obtaın the modified estimation ration $p_{modified}(i)$.

$$p_{modified}(i) = p - \varepsilon \tag{10}$$

Step 9, the average of $p_{modified}(i)$ for i = 0,1,2 will give final embedded ratio $p_{modified}$.

6. EXPERIMENTAL RESULTS

For experimental tests, we have selected more standard 512*x*512 test images (such as Lena, Peppers and so on). We have apply sequential and random LSB replacement to embed the images with the ratio of p= 0, 10%, 20%,. . . ,100% respectively with 10% increments we created two databases.

Then we have use the RS method, DIHmethod and GEFR method to estimate the embedding ratio of secret information respectively. The mask used in the RS method is [1,0; 0,1]. The results from testing the test images are obtained by DIH method and the proposed method (IDIH) which is shown in Table 1.

The leftmost column in Table 1 is the real embedding ratio, and column "IDIH", "DIH" represent the estimate embedding ratio got by Improve Difference Image Histogram method (proposed method) and Difference Image Histogram Method (DIH) respectively. The estimate precision of IDIH is higher than DIH obviously; this aspect is shown in the Table 1.

Embedding ratio (%)	Random		Sequential	
	IDIH	DIH	IDIH	DIH
0%	0.3052	1.6855	0.3052	1.6855
10%	14.7804	15.3881	15.6703	16.0368
20%	20.38	20.80	27.98	28.11
30%	20.3764	20.8017	27.9818	28.1124
40%	40.1524	42.9062	44.3258	44.922
50%	48.6793	52.2864	49.7154	48.5228
60%	62.245	63.8	60.5394	56.5979
70%	72.7311	66.67118	69.7919	68.726
80%	84.6388	73.4632	80.8796	72.2582
90%	90.9915	85.8664	84.8516	81.955
100%	98.6088	95.5193	98.6088	92.5193

Tab. 1. Comparison between IDIH and DIH

In the case of sequential embedding, the accuracy is much higher than the case of random embedding for the embedded ratios of greater than 40%. It is having a higher performance to all the other steganalytic techniques for entire range of possible embedding lengths.

7. CONCLUSIONS

This paper proposes a new detection algorithm, which represent an improved algorithm of the difference image histogram algorithm and performed tests on a group of raw lossless images. Results based on experimental tests, show that the improved difference image histogram steganalysis method is more accurate and guarantee the assurance than the conventional difference image histogram method. The algorithm described in this paper, reduces the mean error with 50% for embedding ratios greater than 40% when compared to the DIH algorithm.
Finally we have generally focused on grayscale still images. However the methods we presented here can be applied to the study of data hiding in color images, video, and audio.

8. REFERENCES

Anantharam V. A large deviations approach to error exponents in source coding and hypothesis testing. IEEE Trans. on Information Theory, 36(4):938–943, 2008

Neil F. Johnson, Sushil Jajodia: Steganalysis of Images Created Using Current Steganography Software, in David Aucsmith (Ed.): Information Hiding, LNCS 1525, Springer-Verlag Berlin Heidelberg 2007. pp. 32–47

Robert Tinsley, Steganography and JPEG Compression, Final Year Project Report, University of Warwick, 1999

Rowland, C.H. "Covert Channels in the TCP/IP Protocol Suite." *First Monday*, 2006.URL: http://www.firstmonday.dk/issues/issue25/rowland/.Last accessed: 2010-12-21. URL: http://www.guides.sk/psionic/covert/covert.tcp.txt. Last accessed: 2004-01-10

K. Bennett,Linguistic steganography: Survey, analysis and robustness cocerns for hiding information in text, Tech. Rep. TR 2004-13, Purdue CERIAS, May 2009