Mathematical and Software Engineering, Vol. 1, No. 2 (2015), 31-36
Varepsilon Ltd, http://varepsilon.com

Modified Algorithm for Steganalysis

Stanimir Zhelezov

Department of Computer Systems and Technologies
Faculty of Mathematics and Informatics
Konstantin Preslavski University of Shumen, 9712 Shumen, Bulgaria

Abstract

This paper proposes a modified algorithm for steganalysis based on data compression. The experimental results verify that the proposed steganalysis can detect the altered images with high accuracy.

Subject Codes: 68U10, 94A60.

Keywords: Steganalysis, Steganography, Information security.

1 Introduction

During the last years the use of Stenography methods for hiding and transmitting of secret messages got spread. These methods can be also used for unauthorized leak of confidential information to people outside the organization. To detect such activity it is necessary to use the other component of steganology – steganalysis.

The purpose of the steganalysis is to gather enough data for the existence of a hidden message, thus destroying the purpose of the steganography – the secret communication.

By efficiency of the steganalysis system we understand the degree of its adaptability in terms of finding the fact that there is a hidden message in a given container [5]:

$$E_{ss} = f\,(P(S/D),\, P_{br}, T_{sa},\, Q_m), \qquad (1)$$

where T_{sa} is the time of execution of the steganalysis, $P(S/D)$ is the probability of finding a hidden message in the container in a given time interval under certain conditions, P_{br} is the probability of reliable system operation, Q_m – expenses for input and operation of the hardware and software [5].

When checking the effectiveness of the steganalysis the first two parameters

T_{sa} and *P(S/D)* can be considered main criteria.

During the analysis of the steganalysis methods using classifiers that are typical for the method of steganalysis based on autocorrelation coefficients [7] and the blind steganalysis based on empirical matrices [3], one major drawback is noticed – the need for continuous training in advance. This leads, in turn, to significant difficulties in implementation of the algorithms that realize this type of methods in actual working steganalysis systems.

The methods, using calculation of the statistical dependencies between different elements of the analyzed files have a significant time advantage over the above. With them the time of algorithm execution depends mostly on the calculations, which are used as basis for defining these dependencies. For example, the methods using transformations (DWT and CWT) will take a lot more time for analysis than the methods using direct calculations of statistical dependencies. For example, in the method for universal blind steganalysis based on color wavelet decomposition [1], the time for calculating the dependencies is considerably more than it is in the method for histogram analysis (chi-square) [6], as the calculations with the second method are with considerably less mathematical complexity.

2 Modified Algorithm for Steganalysis

The main indicator for the effectiveness of a stego system is the accuracy of determining the existence of hidden information. For most of the steganalysis methods, based on calculating the statistical dependencies between different elements of the analyzed files, accumulation of preliminary statistics for analyzing each tangible object leads to a false indication of hidden information existence. This in turn leads to inaccuracies when determining the existence of a hidden message if it does not exceed 5-6% of the stego capacity of the file.

To overcome this disadvantage, from the analyzed methods in this work, the steganalysis method based on data compression [8] was selected for further study.

The proposed in this work algorithm is a modification of the method of steganalysis of image files by using this method [8]. It is based on the fact that the output container and the information added in it are statistically independent, so when adding hidden data to the container its size during compression is greater than the size of an empty output container during compression.

The developed algorithm for steganalysis of image data checks the statistical independence of the data. Widely spread archiving programs are used for compression.

Let *Arh(Z)* be an algorithm for compression, realized with the archiving program applied to the input sequence *Z* of a container and let *Steg(Z)* be the stego algorithm realized with the software, that hides the sequence of the message *M* in the container. Let $K_e(Z)$ be the compression coefficient of the input sequence, defined by the formula:

$$K_e(Z) = \frac{|Z|}{|Arh(Z)|} \quad (2)$$

Let $K_f(S)$ be the compression coefficient of S, defined by the formula:

$$K_f(S) = \frac{|S|}{|Arh(S)|} \quad (3)$$

Let ε be the subtraction of these two coefficients (2) and (3), set with:

$$\varepsilon = K_e(Z) - K_f(S) \quad (4)$$

To determine the fact of inserting information, a limit for δ is chosen and evaluation is made whether the resulting value exceeds the limit. In Fig.1 is shown a block diagram of the modified algorithm for steganalysis of image files using data compression.

The algorithm execution starts with entering input data. The files that will be analyzed are verified for correspondence with the input requirements. Difference is a reason to stop processing. In the absence of such difference it is proceeded with compression of the analyzed input file Z and defining the compression coefficient $K_e(Z)$. In the input sequence Z a control message M is embedded with the help of the stego algorithm $Arh(Z)$, it is subjected to compression and compression coefficient $K_f(S)$ is defined of the obtained stego file. In the next step, ε is defined – the difference between the two compression coefficients ($E<\delta$) and depending on the result a decision is made for the existence of a stego message or its absence. A decision to continue or not the analysis of image files is made based on the result of the performed check in last steps.

For conducting an analysis and evaluation of the effectiveness of the developed algorithm for finding steganography information, an experiment was conducted, including data processing and analysis of over 2500 files with different file formats (BMP, JPEG, DOC, TXT). For comparative analysis, the algorithm, based on the method of histogram analysis (chi-square) was used.

To determine the effectiveness of the proposed algorithm, a base of 1000 stego files was created with the help of different stego programs. This base was subjected to steganalysis using the method, based on data compression, and the results are the following:

1. During embedding of a control message in containers of the base there is a significant difference in compression coefficients (Fig. 1) of the "empty" containers $K_e(Z)$ and those with embedded control stego message $K_f(S)$.

2. This is not typical for the compression coefficients (Fig. 2) of the stego files $K_f(S)$ and the same with an extra embedded message $K_{ff}(S)$.

3. A clear differentiation can be made between ε of a container and ε_f of a

stego file, which is at the core of the proposed algorithm (Fig. 3).

4. There is a wild enough range of values (Fig. 3), that can be selected as limits values for . Therefore, the proposed algorithm allows clear determination of the existence of hidden information in arbitrary files.

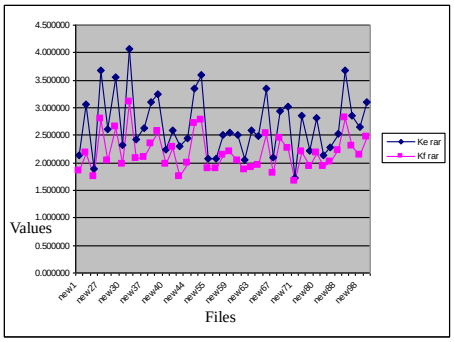

Fig. 1 Difference in compression coefficients in empty file and stego file

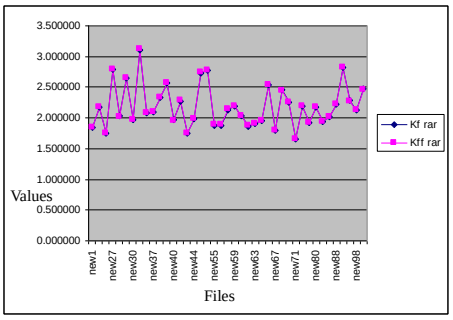

Fig. 2 Difference between stego file and the same with re-inserted message in it.

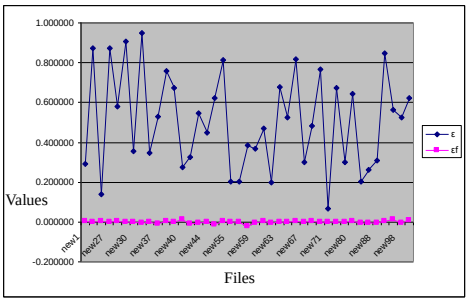

Fig. 3 Differentiation between container and stego file.

To determine the effectiveness of the algorithm of steganalysis of image files using the method χ^2 verification tests of 800 files are carried out. For the purposes of the experiment, the base with containers was used to create a base with stego files and for that purpose messages with different format and size were embedded. The resulting base with stego files was analyzed. And the result was (Fig.4):

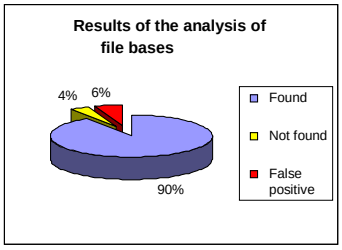

Fig. 4 Analysis of the stego files

For files with small amount of embedded information (embedding coefficient less than 5%), the proposed algorithm does not always find the stego embedding. This applies to all analogical algorithms that use initial statistical information gathering.

3 Conclusion

From the comparative analysis of the received results is clear that using the proposed algorithm for finding stenographic information significantly reduces the error rate (less than 1%). Based on the experimental results, the following main advantages and disadvantages of the analyzed algorithms were realized:

- from the point of view of the criterion of effectiveness in finding hidden information, the algorithm with data compression gives significantly better results than the one based on chi-square. For files with small amount of embedded information (embedding coefficient less than 5%), the chi-square based algorithm does not always find the stego embedding. This applies to all analogical algorithms that use initial statistical information gathering. Using this algorithm significantly reduces the error rate (less than 1%).

- from the point of view of the speed criterion, the algorithm based on chi-square significantly surpasses the second algorithm. In conducting the experiments described above, the execution time of the steganalysis of the stego files the base was also noted. For the first algorithm this time is in the range of 5 minutes (275 seconds). With the proposed algorithm this delay is slightly greater (345 seconds).

Therefore, the use of these algorithms for steganalysis in real time is not very appropriate, because it would lead to a significant delay in the data transfer. For this purpose it is necessary to use high-performance systems and parallel calculations for improving the reviewed algorithms and reducing the execution time of the steganalysis. Implementation of a program realization of the proposed algorithm, as a sub-module of the network security systems will increase their efficiency and will raise their security level [2], [4].

References

[1] Agaian, S., Cai, H., Color wavelet based universal blind steganalysis, San Antonio, Texas, USA, 2005.

[2] Boyanov, P., Using HTTP filter to analyze and monitor the vulnerability and security states in determined computer network, Journal Science Education Innovation, vol. 2 (2014), 45-51.

[3] Chen, X., Blind Image Steganalysis Based on Statistical Analysis of Empirical Matrix, Pattern Recognition, 2006, ICPR 2006, vol.3, 1107-1110.

[4] Kordov, K., Modified Chebyshev map based pseudo-random bit generator, AIP CP 1629, 432-436.

[5] Stanev, S., Galyaev, V., Smyslovoe sopostavlenie nauchnyh terminov na russkom i anglijskom jazykah v oblasti komp'juternoj steganografii, Sbornik materialov mezhdunarodnoj nauchno-prakticheskoj konferencii GAOU VPO Dagestanskij gosudarstvennyj institut narodnogo hozjajstva, Mahachkala, DGINH, 2013, 51-56.

[6] Stanley, C.A., Pairs of Values and the Chi-squared Attack, Department of Mathematics, Iowa State University, May 1, 2005.

[7] Yadollahpour, A., Naimi, H.M., Attack on LSB Steganography in Color and Grayscale Images Using Autocorrelation Coefficients, European Journal of Scientific Research, vol. 31 (2009), 172-183.

[8] Zhilkin, M.Ju., Information-theoretical methods of steganoanalysis of graphic data, Diss. Ph.D., Novosibirsk: Siberian State University of Telecommunications and Information Sciences, 2009, 153 p. (In Russian).