A Performance of Embedding Process for Text Steganography Method

BAHARUDIN OSMAN[1], ROSHIDI DIN[1], TUAN ZALIZAM TUAN MUDA[2],
MOHD. NIZAM OMAR[1],
School of Computing[1],
School of Multimedia Technology and Communication[2]
College of Arts and Sciences
Universiti Utara Malaysia
Sintok, 06000 Kedah
MALAYSIA
bahaosman@uum.edu.my, roshidi@uum.edu.my, zalizam@uum.edu.my, niezam@uum.edu.my
http://www.soc.uum.edu.my,

Abstract: - One of the main aspects on embedding process of any text steganography methods is the capacity text. It is because a better embedding ratio and saving space offers; a more text can be hidden. This paper tries to evaluate several format based techniques of text steganography based on their embedding ratio and saving space capacity. Thus, main objective of this paper is to analyze the performance of text steganography methods which are CALP, Vertical based and Quadruple methods based on these two capacity factors. It has been identified that vertical based method give a good effort performance compared to CALP and Quadruple based method. In future, a robustness of text steganography methods should be considered as a next effort in order to find a strength capability on text steganography.

Key-Words: - Text Steganography, Steganography Method, Format Based Method

1 Introduction

Information hiding is a general term encompassing various sub disciplines in information security field. One of the most popular sub disciplines in information hiding is steganography domain [1, 2].

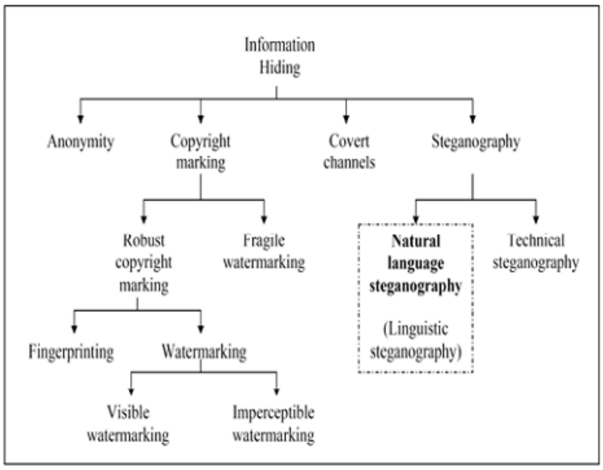


Fig.1. Disciplines in the area of Information Hiding (as depicted from [3])

Steganography is the science of hiding information with the goal is to hide the data from a third party in such way that no one suspects the existence of the message. The word steganography is derived from Greek (*steganos* + *graphy*) and it means covered or hidden writing. Actually, steganography is introduced to hide the existence of the communication by concealing a hidden message in an appropriate carrier which is divided into two domains such as Technical Steganography (image, audio, video, and network) and Natural Language Steganography as shown in Fig.1.

Meanwhile, text steganography is amongst the most challenges method in steganography domains because it is use only a small memory and simple communication compare to other media due to hide messages inside text [4]. A technique used in text steganography can be classified into three types which are format based, random and statistical generation and linguistic method [5, 6]. Format-based method used and change the physical formatting of the cover message to hide the data and maintain the existing word or sentence. A few techniques used in these methods are spacing between line, spacing between word, modifying feature of certain letter are discussed in [7] and also using an extra space to be added into the hidden text. Some of these techniques such as deliberate misspelling and space insertion might fool human reader but can often be easily detected by computer [8]. Random and statistical generation is used to generate cover message automatically according to the statistical properties of language. Such generation tries to simulate some property of normal text, usually by approximating some arbitrary statistical distribution found in real text. Some techniques used are character sequences, words sequences, statistical generation of sequences-text mimicking and etc.

A linguistic method considers the linguistic properties of the text to modify it. A linguistic structure is used to hide the data where the syntax or semantic of the language is used. In syntactic method, such as punctuation, comma and full stop are placed in a proper place in the document, whereas semantic method will replace the synonym word. In order to be a good text steganography, methods use should consider at least two capacity factors of the hidden text against cover text which are embedding ratio and saving space ratio factors. Thus, this paper tries to evaluate several methods of text steganography on format based techniques. These methods will be examined both from their embedding ratio and saving space capacity. Therefore, the main objective of this paper is to analyze the performance of text steganography methods based on these two capacity factors.

The rest of the paper is organized as follows. Section 2 describes a problem formulation of text steganography. In Section 3, this paper presents the methods uses on text steganography, capacity factors and dataset uses in this experimental work. Section 4 discusses a result of the experimental work. Finally, Section 6 is concluding this paper.

2 Problem Formulation of Text Steganography

A general idea of steganography process can be shown in Fig.2.

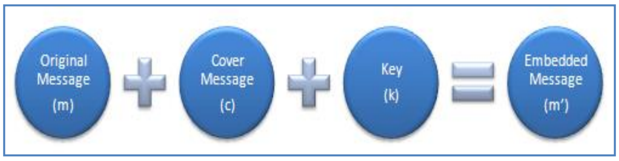


Fig.2 A general formula of steganography process

Firstly, the original message will be concealed in cover message by applying an embedding algorithm (using key) to produce an embedded message. Sender will send an embedded message via a communication channel to receiver. The receiver needs to use a recovering algorithm to extract embedded message. A key is used to control the hiding process so as to restrict detection and/or recover of the embedded data to parties who know it. The relationship of the process can be written as

$$m' = \{m, c, k\}$$

where,

m' *message that hold the hidden data*
m *covert message that one wishes to send*
c *text used to hide the embedded data*
k *function used to hide and unhide the hidden data*

A keyword hidden text, cover text, stego key and stego text will be used to represent an original message, cover message, key and embedded message respectively in further discussion. Based on Fig.3, a stego text is obtained by embedding a hidden text (original text) within a cover text using a stego key function. In this example, the cover text (c) was embedded with a hidden message using a key function. A key function is injected in the embedding process to hide the hidden message to produce a stego text.

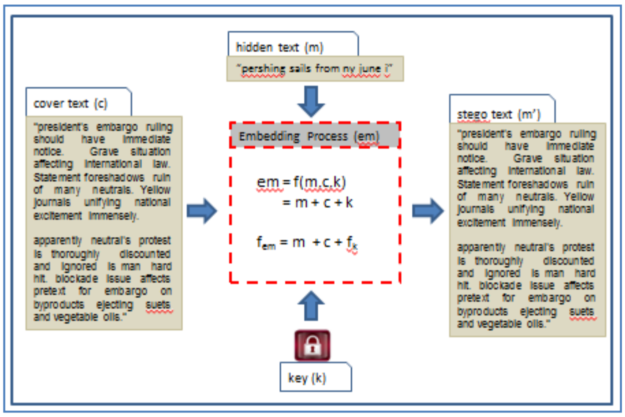

hidden text (m)
"pershing sails from ny june i"
cover text (c)
"president's embargo ruling should have immediate notice. Grave situation affecting international law. Statement foreshadows ruin of many neutrals. Yellow journals unifying national excitement immensely.
apparently neutral's protest is thoroughly discounted and ignored is man hard hit. blockade issue affects pretext for embargo on byproducts ejecting suets and vegetable oils."
Embedding Process (em)
em = f(m,c,k)
= m + c + k
f_{em} = m + c + f_k
stego text (m')
"president's embargo ruling should have immediate notice. Grave situation affecting international law. Statement foreshadows ruin of many neutrals. Yellow journals unifying national excitement immensely.
apparently neutral's protest is thoroughly discounted and ignored is man hard hit. blockade issue affects pretext for embargo on byproducts ejecting suets and vegetable oils."
key (k)

Fig.3 An embedding process of stego text

One of the main aspects should be considered when discussing the embedding process of any text steganography methods is the capacity of hidden text and cover text. At least, there are two capacity factors of the hidden text against cover text which are embedding ratio and saving space ratio factors. It is because a better embedding ratio and saving space offers; a more text can be hidden. Then, these factors will determine the capability of the stego text based on the embedding process. Since the capacity of the hidden text and cover text is one of the important factors, several studies have been done in order to measure fitness's performance of text steganography. However, a result shows that only a limited amount of hidden text can be embedded into cover text [6, 8]. Thus, this paper tries to evaluate several methods of text steganography on format based technique such as CALP, Vertical Based and Quadruple Based [9, 10].

3 Experimental Work

Firstly, this section discusses the selected methods uses on text steganography during embedding process. Then, the performance of the selected methods will be examined based on their embedding ratio and saving space factors. After that, their performance will be compared in order to find the best fit performance among each other.

3.1 Method Used

There are three methods are considered uses in this work namely, changing in alphabet letter patterns method known as CALP method, vertical straight line method and quadruple categorization. These methods uses text file containing hidden text and this hidden text is converted to binary bits before applying in embedding process. Changing in Alphabet Letter Patterns method known as CALP method is tries to manipulate English letters by mapping the binary sequence of the hidden text through pattern changes of several letter of the cover text during embedding process. These pattern changes have been incorporated using some unused symbols of the ASCII number system.

Meanwhile, vertical straight line method uses English letters into two group based on straight vertical line in a characters as the basis to group each letters. The letters contain one vertical straight line is identified as A group such as "*B, D, E, F, I, J, K, L, P, R, and T*" will be hide 1 bit hidden data. Whereas, a letter contain more than one single line or do not contain a vertical straight line is identified as B group such as "*A, C, G, H, M, N, O, Q, S, U, V, W, X, Y, and Z*" will be hide 0 bit hidden data.

Finally, quadruple categorization method utilizes an English letters into four group based on the letters pattern whether the letter has a curve, middle horizontal straight line, single vertical line or multiple straight vertical line. Each of these group will hide data bit either 00 bit, 01 bit, 10 bits or 11 bits depend on which group of the letter used belongs.

3.2 Capacity Factor

In this analysis, two factors of the capacity measurement have been used which are *Embedding Ratio (ER)* and *Saving Space Ratio (SSR)*.

a) *Embedding Ratio (ER)*

Embedding ratio is used to determine the total fitness of hidden text can be embedded in cover. This analysis is very important for steganographer to understand the fitness capability of cover text.

$$ER = \left[\frac{\textit{Total Bits of Stego Text} - \textit{Total Bits of Cover Text}}{\textit{Total Bits of Cover Text} + \textit{Total Bits of Hidden Text}}\right] X\ 100\%$$

$$ER = \left[\frac{\textit{Total Number of Embedded Bits}}{\textit{Total Bits of Expected Stego Text}}\right] X\ 100\%$$

$$\textit{Percent Embedded Bits (\%)} = \frac{\sum_{i=1}^{1<m<100} a_i}{\sum_{i=1}^{1<n<100} b_i} \times 100\%$$

where

a = *Total Number of Embedded Bits*
b = *Total Bits of Expected Stego Text* (1)

where

a = *Total number of embedded bits*
b = *Total bits of cover text*

b) *Saving Space Ratio (SSR)*

Saving space ratio is used to determine the total space of hidden text that can be saved during embedding process in cover text. This analysis is very important for steganographer to understand the capability of maximum space that can be utilized in cover text in order to embed the hidden text.

$$SSR = \left[\frac{\textit{Total Bits of Expected Stego Text} - \textit{Total Bits of Stego Text}}{\textit{Total Bits of Expected Stego Text}}\right] X\ 100\%$$

$$SSR = \left[\frac{\textit{Total Number of Saving Space Bits}}{\textit{Total Bits of Expected Stego Text}}\right] X\ 100\%$$

$$\textit{Percent Saving Space Bits (\%)} = \frac{\sum_{i=1}^{1<m<100} a_i}{\sum_{i=1}^{1<n<100} b_i} \times 100\%$$

where

a = *Total Number of Saving Space Bits*
b = *Total Bits of Expected Stego Text* (2)

where

a = *Total number of saving space bits*
b = *Total bits of hidden text*

3.3 Dataset Selection

Text chosen dataset is one of the important components in benchmarking steganography methods. Our study used a dataset of hidden text and cover text which includes a variety of textures and sizes. In order to evaluate the text steganography methods, various file sizes have been categorized in three phases during evaluation process from phase I to phase III, respectively. Phase I started with 765.455 bytes, followed by phase II with 832.361 bytes, phase III with 982.656 bytes of cover text files size. Meanwhile, there are several hidden text have been used during embedding proses such as 86 bytes until 663 bytes.

4 Result

The following figure has shown the performance of embedding process and saving space bit on text steganography based on CALP, Vertical based and Quadruple methods. It has been identified that Vertical based and Quadruple methods give a good effort performance compared to CALP method. Both are having a quite similar performance. The following result show the analysis using the hidden text with the size of 710.144 bytes and 16 varies of coverText.

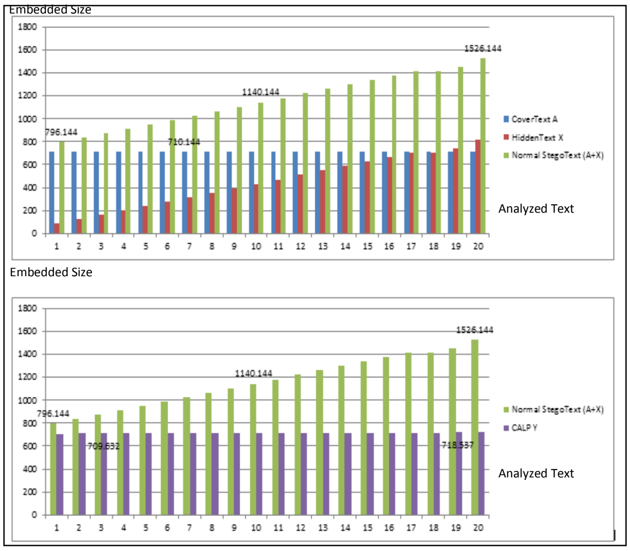


Figure 4.1: StegoText using CALP method

Figure 4.1a shown the size of normal stegoText after embeding hidden text and cover text which seen that the minimum and maximum size of normal stegoText is 796.144 and 1526.144 bytes respectively. The size of normal StegoText dramatically increase when the size of hidden text increased. However, Figure 4.1b show that CALP method constantly maintain the size of stegoText with the minimum and maximum of the StegoText size is 709.632 and 718.537 bytes respectively.

Figure 4.2a shown the size of normal stegoText after embeding hidden text and cover text which seen that the minimum and maximum size of normal stegoText is 796.144 and 1373.144 bytes respectively. The size of normal StegoText dramatically increase when the size of hidden text increased.

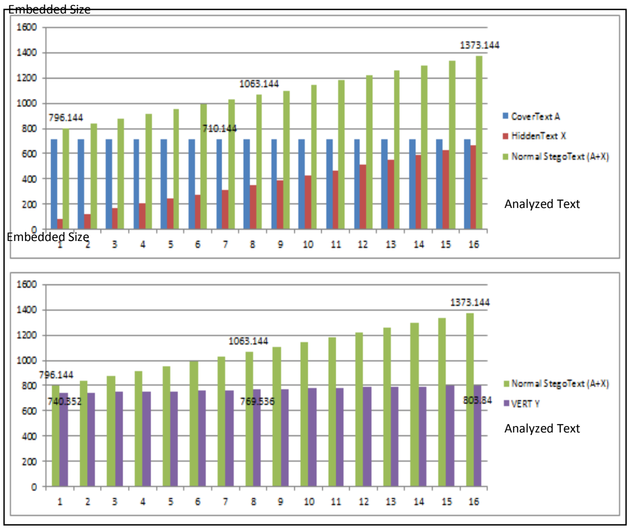


Figure 4.2: StegoText using VERT method

However, Figure 4.1b show that VERT method constantly maintain the size of stegoText with the minimum and maximum of the StegoText size is 740.552 and 803.84 bytes respectively.

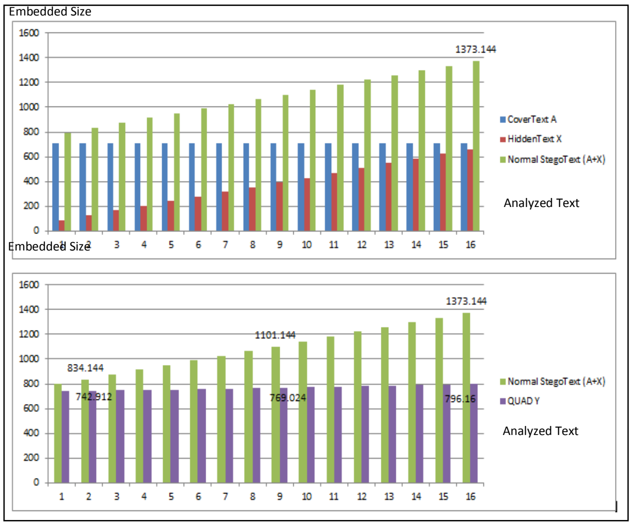
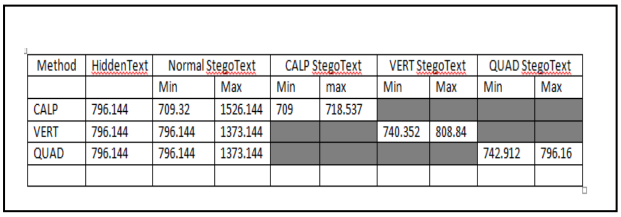


Figure 4.3: StegoText using QUAD method

Figure 4.3a shown the size of normal stegoText after embeding hidden text and cover text which seen that the minimum and maximum size of normal stegoText is 796.144 and 1373.144 bytes respectivelly. The size of normal StegoText dramatically increase when the size of hidden text increased. However, Figure 4.1b show that QUAD method constantly maintain the size of stegoText with the minimum and maximum of the StegoText size is 740.912 and 796.16 bytes respectively.

5 Conclusion

In this paper, three types of text steganography methods have been evaluated. All of these methods give a similar result for embedding ratio and saving space performance. In future, this paper proposes to evaluate a robustness of each method in order to find a strength capability on text steganography from steganalysis activities.



Method	HiddenText	Normal StegoText		CALP StegoText		VERT StegoText		QUAD StegoText	
		Min	Max	Min	max	Min	Max	Min	Max
CALP	796.144	709.32	1526.144	709	718.537	▨	▨	▨	▨
VERT	796.144	796.144	1373.144	▨	▨	740.352	808.84	▨	▨
QUAD	796.144	796.144	1373.144	▨	▨	▨	▨	742.912	796.16

References:

[1] Fabien A. P. Petitcolas, Ross J. Anderson and Markus G. Kuhn, Information Hiding – A Survey, *Proceedings of the IEEE, Special Issue on Protection of Multimedia Content*, July 1999, pp. 1062 – 1078.

[2] Mohammed Al-Mualla, Hussain Al-Ahmad, Information Hiding: Steganography and Watermarking. [Online] Available: *http://www.emirates.org/ieee/information_hiding.pdf* [Accessed]: December 12, 2012.

[3] Roshidi D., Azman S, Puriwat L., A Framework Components for Natural Language Steganalysis, *Journal of Computer Theory And Engineering*, Vol. 4, 2012, pp. 641 - 645.

[4] Shirali-Shahreza M. H., Sharali-Shahreza M., A New Approach to Persian/Arabic Text Steganography, *Proceeding of the 5th IEEE/ACIS International Conference on Computer and Information Science*, Honolulu, USA, July 10 - 12, 2006, pp. 310 - 315.

[5] Souvik B., Indradip B. and Gautam S., A Survey of Steganography and Steganalysis Technique in Image, Text, Audio and Video as Cover Carrier, *Journal of Global Research in Computer Sciences*, Vol. 2, April 2011.

[6] M. Grace V., Rao, M. Swapna, J. Sasi Kiran, Hiding the Text Information Using Stegnography, *International Journal of Engineering Research and Applications (IJERA)*, Vol. 2, Issue 1, Jan - Feb 2012, pp. 126 - 131. ISSN: 2248-9622.

[7] Sellars, Duncan, An Introduction to Steganography. *http://www.cs.uct.ac.za/courses/CS400W/NIS/papers99/dsellars/stego.html (March, 2003).*

[8] K. Bennett, Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text, Purdue University, *CERIAS Tech. Report*, 2004.

[9] Souvik B., Pabak I., Sanjana D., Indradip B. and Gautam S., Hiding Data in Text Changing in Alphabet Letter Patterns, *Journal of Global Research in Computer Sciences*, Vol. 2, No. 3, March 2011.

[10] Shraddha D., Devesh J., and Aroop D., Experimenting with the Novel Approaches in Text Steganography, *International Journal of Network Security and Its Applications (IJNSA)*, Vol.3, No. 6, November 2011.