A Novel Text Steganography System Using Font Color of the Invisible Characters in Microsoft Word Documents

Md. Khairullah
Department of Computer Science and Engineering
Shahjalal University of Science and Technology, Sylhet, Bangladesh
E-mail: khairul_cse@yahoo.com

Abstract- **Steganography can be defined as a method of hiding data within a cover media so that other individuals fail to realize their existence. Image, audio and video are some popular media for steganography. But text is ideal for steganography due to its ubiquity and smaller size compared to these media. However, text communication channels do not necessarily provide sufficient redundancy for covert communication. In this paper, a new approach for steganography in Microsoft Word documents is proposed. The main idea is that setting any foreground color for invisible characters such as the space or the carriage return is not reflected or viewed in the document.**

Keywords- Steganography, Microsoft Word, Font Color, Invisible Characters, Cover media

I. INTRODUCTION

Steganography is the art and science of writing hidden messages in such a way that no-one, apart from the sender and intended recipient, suspects the existence of the message, a form of security through obscurity [1]. Cryptography — the science of writing in secret codes — addresses all of the elements necessary for secure communication over an insecure channel, namely privacy, confidentiality, key exchange, authentication, and non-repudiation. But cryptography does not always provide safe communication. Consider an environment where the very use of encrypted messages causes suspicion. If a nefarious government or Internet service provider (ISP) is looking for encrypted messages, they can easily find them. Consider the following text file; what else is it likely to be if not encrypted? Steganography is the science of hiding information. Whereas the goal of cryptography is to make data unreadable by a third party, the goal of steganography is to hide the data from a third party [2].

Microsoft Word is popular word processing software which comes with Microsoft Office package. One of the reasons behind its popularity is huge number of text formatting features. Setting background color and font color is the simplest of text formatting features. We can set font color by choosing Format from menu bar, Font menu and then Font color. We can do this work also from the font color tool from the tool bar. We can choose some predefined colors or can set RGB values for the desired color by choosing the more color options and then choosing the custom option. These are depicted in figure-1 to 4. The R, G and B values are of 8 bits, which means the allowed range is 0 to 255. One of the interesting things is that we can also set font color of invisible characters such as the space, the tab or the carriage return characters. Most of the users or viewers never feel interest about the color values of these invisible characters. So we can very easily hide three bytes of information in each occurrence of the space or the tab or the carriage return characters without any risk of exposing the hidden information. More over this approach takes no extra information to hide the desired bits.

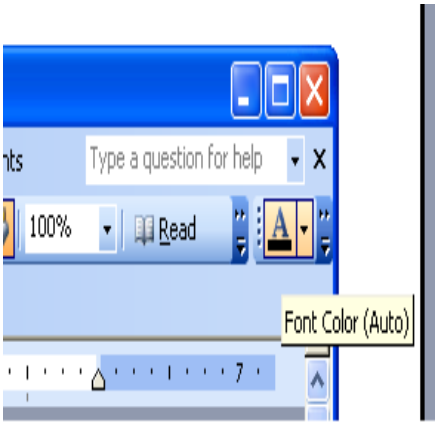

Figure-1 Choosing the Font Color option

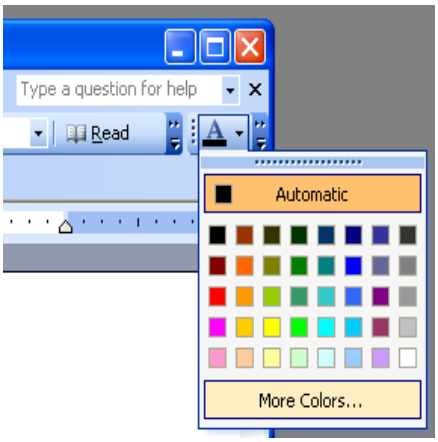

Figure-2: Choosing more color option to set RGB values

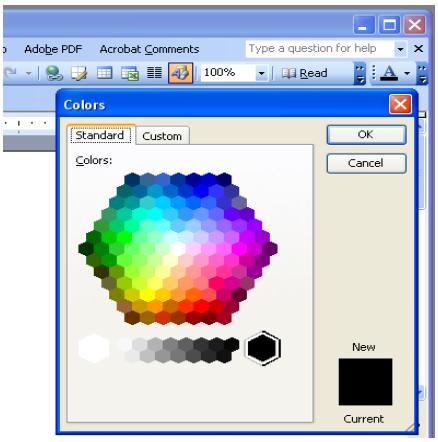

Figure-3: Choosing the custom option to set RGB values

978-0-7695-3925-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICCEE.2009.127

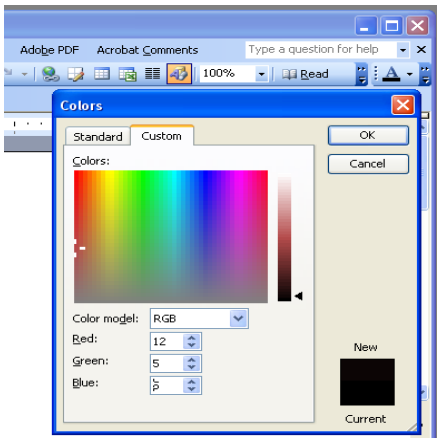

Figure-4: setting RGB values

II. Previous Works

Following is the list of works has been done on hiding information or text steganography carried out.

A. Text Steganography in Markup Languages

In this method, one of the features of markup languages is used to hide information [3]. For instance feature of HTML document is their tags case insensitivity. For example, the tag
 can be also used as
 and
. As a result one can do text steganography in HTML documents by changing the small or large case of letters in document tags. However these methods are not for those markup languages which has case sensitive tags. WML is an example of this category. In some cases the positions of tags are also used for text steganography. For example <B><U> </B></U> or like this <U><B> </U></B>. We extract information by comparing the tags positions. In WML the second text steganography method can be employed.

B. Text Steganography In Specific Characters In Words

In this method, some specific characters from certain words are selected [4]. For example the first words of each paragraph are selected in a manner that by placing the first characters of the selected words side by side, as a result it forms secret or hidden information is extracted. This method is a little bit analytical rather than automated. Hence this method is time-consuming.

C. Line Shifting Method

In this method, the lines of the text are vertically shifted to some degrees [5,6]. For example, some lines are shifted 1/300 inch up or down in the text and information are hidden by creating a hidden unique shape of the text. This method is feasible for printed texts. However, if the text is retyped or if character recognition programs (OCR) are used, the hidden information would get destroyed.

D. Word Shifting

In this method, by shifting words horizontally and by changing distance between words, information are hidden in the text [5, 7]. This method is acceptable for texts where the distance between words is varying. Although this method is very time consuming, there is a high probability of finding information hidden in the text. The same as in the method described under 2-3, retyping of the text or using OCR programs destroys the hidden information.

F. Syntactic Methods

By placing some punctuation signs such as full stop (.) and comma (,) in proper places, one can hide information in a text file [4]. This method requires identifying proper places for putting punctuation signs. The amount of information to hide in this method is trivial.

G. Semantic Methods

In this method, we use the synonym of words for certain words thereby hiding information in the text [6, 8]. A major advantage of this method is the protection of information in case of retyping or using OCR programs (contrary to methods listed under 2-3 and 2-4). However, this method may alter the meaning of the text.

H. Feature Coding

In this method, some of the features of the text are altered [9]. For example, the end part of some characters such as h, d, b or so on, are elongated or shortened a little thereby hiding information in the text. Retyping the text or using OCR program destroys the hidden information.

I. Abbreviation

Another method for hiding information is the use of abbreviations. In this method, very little information can be hidden in the text [4]. For example, only a few bits can be hidden in a file of several kilobytes.

J. Open Spaces

In this method, hiding information is done through adding extra white-spaces in the text [4, 10]. These white spaces can be placed at the end of each line, at the end of each paragraph or between the words. However, some text editor programs automatically delete extra white-spaces and thus destroy the hidden information.

K. Vertical Displacement of the Points in Arabic, Persian and Urdu Letters

In this method, text steganography is applied on Arabic, Persian and Urdu text [11, 12, 16]. One of the characteristics of these languages is abundance of points in its letter. One point letters are used to hide the information by shifting position of point a little bit vertically high with respect to the standard point position in the text.

L Vertical Displacement of the Diacritics in Arabic Language

In this method, text steganography is applied on Arabic text [13, 14]. One of the characteristics of Arabic languages is that the vowel symbols are located either on the upper or lower side of the character to be modified by the vowel. This characteristic is used to hide the information by shifting position of the vowel sign (diacritics) a little bit vertically high or low with respect to the standard diacritics position in the text.

M. Use of Extension Letters in Arabic Language

In this method, text steganography is applied on Arabic text [15]. Arabic language has a special extension character which can be arbitrarily inserted between characters for formatting purposes.

III. The Proposed Technique

The idea behind our technique is very simple. We can set foreground color for invisible characters as like as any other visible characters. But the color attribute is not noticed by the users where the font color is very much likely to be noticed by the user. So we can hide secret information in invisible characters in form of RGB values without any risk of exposure. For example consider the following stream
10101011010111010110001101010011001110100100111001010011100101010110001010
Here we divide the secret bit stream of groups of 8 bits
Groups are represented by alteration of italic and bold fonts.
10101101*01110101***10001101***01001100***11101001***00111001***01001110***01010101***10001010**
The RGB values of the first three invisible characters of a document should be
{173, 117, 141}, {76, 233, 57}, {78, 85, 138}

Algorithm-1: Hiding secret bits
Input: Cover Microsoft Word document, secret bit stream
Output: Stego Microsoft Word document

- Step-1: Arrange the secret bit stream in group of 8 bits from the right side of the stream. If there are less number of bits in the last group pad 0s to complete 8 bits. Have number of groups multiple of 3. If the number of group is not multiple of 3 add additional 0s to left of the stream so that number of groups are multiple of 3.
- Step-2: Group each three group to construct a super group.
- Step-3: Convert the binary value of each group to the equivalent decimal value.
- Step-4: For each invisible character in the cover document repeat steps 5 to 8.
- Step-5: Set R value of the color to the first decimal value of the super group.
- Step-6: Set G value of the color to the second decimal value of the super group.
- Step-7: Set B value of the color to the third decimal value of the super group.
- Step-8: Choose the next super group of secret bits for the next invisible character.

Algorithm-2: Extracting bits from the received Microsoft Word document
Input: Stego Microsoft Word document
Output: extracted secret bit stream

- Step-1: For each invisible character in the Stego document repeat step 2.
- Step-2: Get the RGB value of the font color and add them to the extracted hidden data in sequence of R, G and B
- Step-3: Convert every decimal value to the equivalent binary value to form the secret bit stream.

IV. Conclusion

For text steganography various methods have been proposed. We represent a novel technique of hiding information in Microsoft Word documents. Microsoft Word documents are very much common in every day life of today's digital world. The capacity of this method is very high. We hide 3 bytes of data in every invisible character. So, our proposed method can be a good choice for text steganography.

References

[1] Steganography, http://en.wikipedia.org

[2] Steganography: Hiding Data Within Data, http://sover.net

[3] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report 2004-13.

[4] T. Moerland, "Steganography and Steganalysis", May 15, 2003, www.liacs.nl/home/tmoerlan/privtech.pdf, last visited: 1 May 2006.

[5] S.H. Low, N.F. Maxemchuk, J.T. Brassil, and L. O'Gorman, "Document marking and identification using both line and word shifting", Proceedings of the FourteenthAnnual Joint Conference of the IEEE Computer andCommunications Societies (INFOCOM '95), 2-6 April 1995,vol.2, pp. 853 - 860.

[6] A.M. Alattar and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing ", Proceedings of SPIE -- Volume5306, Security, Steganography, and Watermarking ofMultimedia Contents VI, June 2004, pp. 685-695.

[7] Y. Kim, K. Moon, and I. Oh, "A Text Watermarking Algorithm based on Word Classification and Interword Space Statistics", Proceedings of the Seventh International Conference on Document Analysis and Recognition(ICDAR'03), 2003, pp. 775–779

[8] M. Niimi, S. Minewaki, H. Noda, and E.Kawaguchi, "A Framework of Text-based Steganography Using SDForm Semantics Model", Pacific Rim Workshop on Digital teganography 2003, Kyushu Institute of Technology, Kitakyushu, Japan, July 3-4, 2003.

[9] K. Rabah, "Steganography-The Art of Hiding Data", Information Technology Journal, vol. 3, Issue 3, pp.245-269, 2004.

[10] D. Huang, and H. Yan, "Interword Distance Changes Represented by Sine Waves for Watermarking Text Images", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 12, December 2001, pp. 1237-1245

[11] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", Proceedings of 5th IEEE/ACIS international Conference on Computer and Information Science and 1st IEEE/ACIS, June 2006.

[12] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A Robust Page Segmentation Method for Persian/Arabic Document", WSEAS Transactions on Computers, vol. 4, Issue 11, Nov. 2005, pp. 1692-1698.

[13] Mohammed Aabed, Sameh Awaideh, Abdul-Rahman Elshafei, and Adnan Gutub, Arabic Diacritics Based Steganography", IEEE International Conference on Signal Processing and Communications (ICSPC 2007), Pages: 756-759, Dubai, UAE, 24-27 November 2007

[14] Adnan Gutub, Yousef Elarian, Sameh Awaideh, and Aleem Alvi, "Arabic Text Steganography Using Multiple Diacritics", WoSPA 2008 – 5th IEEE International Workshop on Signal Processing and its Applications, University of Sharjah, Sharjah, U.A.E. 18 – 20 MARCH 2008.

[15] Adnan Abdul-Aziz Gutub, and Manal Mohammad Fattani, "A Novel Arabic Text Steganography Method Using Letter Points and Extensions", PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 21 MAY 2007 ISSN 1307-6884 pp 28-31

[16] Jibran Ahmed Memon, Kamran Khowaja, Hameedullah Kazi, "EVALUATION OF STEGANOGRAPHY FOR URDU /ARABIC TEXT", Journal of Theoretical and Applied Information Technology, pp 232-237