2014 1st International Congress on Computer, Electronics, Electrical, and Communication Engineering (ICCEECE2014)
IPCSIT vol. 59 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2014.V59.2

A Novel Approach to Text Steganography

Mr. Hitesh Singh [1], Mr. Anirudra Diwakar [1+], and Ms. Shailja Upadhyaya [1].

[1] HMRITM, Delhi.

Abstract. Steganography is an exciting field for research providing directions for safe and secure transmission of information. The useful information needs to be concealed from the outside world. For this purpose, techniques like cryptography and steganography are instigated. Steganographic techniques are preferred now a days to keep secretive information hidden from any spy's suspicion. Images [1, 2, 3], audio [4, 5], video and text are the mediums for steganography. Among these, text steganography proves to be the most secure and trustworthy one. It is challenging to detect a text which carries a hidden message. We focus on comparing new methods being introduced in feature encoding for hiding a message by exploiting "Text Formatting". Through this paper we propose innovative techniques to use text color to hide a message which is received by the intended receiver. A comparison between three feature encoding techniques is also carried out.

Keywords: Text steganography, RTF, color, feature encoding, Devanagari, Latin.

1. Introduction

Steganography [6, 7] is an ancient technique of hiding messages in such a manner that only the sender and the intended recipient know about the message hidden in it. It was first used in 480BC [7] and then the need for steganography extended to world wars also. The use of steganography is gradually increasing due to the fact that cryptography is no longer used. The main difference between cryptography and steganography is that cipher codes are generated as a result of the cryptographic algorithms and these cipher codes are easily interpreted by a third party. The cipher codes attract attention of the people reading it. On the other hand steganography is a technique which does not generates a cipher code. It hides information in ways a message is hidden in an innocent-looking cover media. The cover text does not have the capability to attract attention of the person reading it. It appears as a normal text. Hence the term 'cover writing' is coined for steganography.

There are various terms used in this paper: message, stego medium, stego object and Steganalysis [6]. A *Message* is to be hidden in a cover file. Stego medium is the transmission channel through which the stego object is transmitted. In this paper, algorithms are introduced through which a file is used to hide the message and a stego object is therefore created. Steganalysis is the method for detecting hidden message in the file.

Devanagari is written from left to right. It comprises of Nepali, Marathi, standard Hindi and other languages. A horizontal line is added to the top of letters in the languages of Devanagari script. The horizontal line acts as a recognizable pattern for the letters of this script. It is one of the impressive languages. In this paper both standard Hindi and English are used to mask a message.

This paper unravels new technologies in the field of text steganography. The colors of the text are changed and encoding of the message is done in such a manner that is undetectable to human eye. Multiple encodings are used to generate a cover text and increase the efficiency along with capacity.

[+] Corresponding author. Tel.: + (91)9958353874
E-mail address: anirudradiwakar@gmail.com

2. Types of Steganography

Steganography means forming covert channels [8]. This field has two possible directions: protection against detection [8] and protection against removal [8]. Protection against detection means that the visibility of the object is altered in a way undetectable by human but measurable by a computer. Protection against removal hinders the quality of the object.

Within protection against removal there are fingerprinting [8] and watermarking [8] techniques. Fingerprinting embeds a unique identifier into user's copy of the file which can be extracted to help identify the source of an unauthorized copy. Every object is marked specifically.

Watermarking is small alteration to a file in such a manner that it cannot be noticed. All objects are marked in the same way.

Hiding a text within a text is text steganography. It is the most difficult steganography. Text steganography is broadly classified as linguistic steganography and format based steganography. Linguistic steganography is divided into semantic and syntactic method. Format based steganography is divided into line shift encoding, word shift encoding, feature encoding and white space method. Linguistic encoding deals with the change in the grammatical sense of the text. Syntactic method is the change in the syntax of the text, for e.g. the American and British spellings can be used to distinguish a word. In the semantic method the word is itself changed. A primary and secondary meaning is provided to a word which is used to hide a message. Line shift encoding is the method under format based steganography which shifts the lines of the text vertically to hide a text message. In word shift encoding words are shifted horizontally within the text to hide a message. Feature encoding incorporates the modification in the features of the text. The height of the characters, their colors or fonts can be changed to hide a message. White space method manipulates the white spaces to yield the desired result in the cover text.

3. Previous Research

This section presents work previously undertaken to implement text steganography.

3.1. Semantic Method

This method is implemented by introducing a change in the meaning of the text. Semantic method [9] takes into account the synonyms of a word. The synonyms convey the same meaning so they can be used in a better way to hide a message. For example: primary and secondary meanings of a word can be used in a text to hide a message. This will prevent attacker from knowing that he is reading a cover text. Semantic method is the one which does not destroys the hidden information even if an Optical Character Recognition technique is used. Even OCR is not able to detect the hidden message.

Semantic and Syntactical methods are used in parallel to hide a message and these methods are more secure for implementing steganography.

3.2. Syntactical method

Syntactical method [9] as the name suggests focuses on the syntax of the text. The syntax of the text can be varied by inserting punctuations marks or by using different spellings of a word. As an example, comma (,) or full stop(.) can be used to hide secret information. Another example can be that American and British English change the spellings of the words. It is a better technique to use these languages for creating a stego object. The chance of detection of hidden message by the attacker is minimal. The attraction of the attacker will not be attracted if syntactical methods are used in an appropriate manner.

3.3. Feature encoding

The feature encoding [9, 10] deals with the change in the features of the text in such a manner that a meaningful message is hidden to produce a cover text. Features like height of the text, color of the text, font of the text are some of the ways which are used. A large volume of information is hidden using feature encoding. When the features of the text are altered then only the sender and the intended recipient detects the hidden message. The third party does not catch the attention of something concealed inside a text. OCR techniques and retyping are responsible for changing the text features and damaging the message.

3.4. Text Steganography using Hindi letters and its Diacritics

The letters and the letter diacritics of Hindi language are used to hide a secret message. A message is taken and it is converted into binary scheme. Bit ‘0’ is used for letters and bit ‘1’ is used for letter diacritics. A Hindi alphabet is represented as a four bit code and this code is replaced with a different word which starts with the respective letter that is assigned in the scheme. This method takes a large amount of time and it requires strong mental computations to hide a message in a text [9].

3.5. Text steganography in Random Character and Word Sequences

It focuses on hiding a message at random places in the text [9]. Various word sequences and characters are inserted at these random places but the method is not a suitable one. This is because it is easy to attract attention of the third party and so it can be thought of as an encryption instead of a steganography.

3.6. Text steganography in specific characters in words

As the name suggests specific characters from certain words are selected to hide a meaningful message [9]. This method requires strong mental computations for hiding messages in such a manner that recipient will not sense the hidden message. There are limited languages supporting this technique as characters of words of only certain languages can be used to hide messages. Hindi language is one of them.

For example: the characters of every first or second word of the paragraphs in text can be selected to create a hidden message.

3.7. Text steganography in Mark-Up Language

A markup language is a modern system for annotating a document in a way that is syntactically distinguishable from the text. Markup languages which are case insensitive can be easily used to create hidden meanings. HTML is used for this purpose. The html tags do not support case sensitivity. Tags like <BR> and <br> gives the same output without being affected by the case in which they are written. Other tags like <b>,<u> etc can be used so that there are more options available for steganography in HTML. XML is case sensitive language. This means the tags written in XML are affected by the capitalization of the alphabets. For steganography purposes HTML is used mostly [9]. It increases the range with which sender exploits the functionality of creating a secret message.

3.8. Random and Statistical generation

In this method some property of normal text is simulated by approximating some arbitrary random and statistical distribution found in real text [2]. Steganographers generate their own cover texts and this leads to excluding the comparison of this text with the plain text. These cover texts are a result of the random techniques applied so that a hiding of message is accomplished properly. Although this method is used but there is a disadvantage of implementing this technique the properties of the cover text which changes may give a sense of something hidden in it. An observer may find out the changes done to create a cover text.

3.9. Word Sequences

The problem arising is detection of the cover text in a normal text sometimes is not that difficult. Various non-lexical sequences can be detected and the power of steganography is decreased. To solve this problem actual detection items can be used to encode one or more bits of information per word. Mapping between the lexical sequences and the bit sequences may require a code-book [9]. This is because the bits in the texts are used to encode a lexical message. This method also has several problems that both human and computer can detect a string of words with no semantic structure. The anomalous behavior may attract attackers and destroy the essence of steganography.

3.10. Character Sequences

There are a number of characters in a text. The beauty of text steganography lies in the fact that characters of languages can be used to create a cover text. In the character sequence method [9] the character generation is to consider properties of word length and letter frequency in order to create words. This gives an outlook the same statistical property as actual words in a given language.

3.11. Open Spaces

Another work undertaken in text steganography is the use of white spaces [9]. A lot of white spaces can be included in the text to hide a message. These spaces can be inserted in the beginning of a line or at the end of a line. The beginning of the paragraph can also include the spaces. The paragraph ending may have spaces to create a cover text. In between the lines of the text message can be hidden. White spaces only add an extra advantage of hiding information. The integrity of the text is lost if text editors are used to delete white spaces.

3.12. Abbreviations

Abbreviation reduces the text and hides information by representing a word for phrase of words. The abbreviations can be used for hiding a message as number of words can be replaced by a single word and this will create a different meaning which helps to generate a cover text [9]. The intended recipient should know about the abbreviations used by the sender and the attacker may not detect the cover text. This method is useful only for short length texts. It is not applicable to texts with large amount of information.

3.13. Word Shifting

The method introduces the idea of shifting the words [10, 11] of the text in a horizontal manner to hide a meaningful message. The distance in between the words is altered so that information is hidden. The spaces within the words should be different in order to generate a cover text. The decoder should also know about how to shift the words to extract a message hidden using steganography. If there are variable places involving the word shifting method then the decoder should be given the original text so that detection of correct information at the decoders end is possible. This method has similar limitations as line shifting. If retyping of the cover text is done or if the OCR techniques are used then the message will be destroyed and this method will fail.

3.14. Line Shifting

This technique is used to modify the text by shifting the lines vertically [10, 11]. The shifting of lines led to a specific pattern in the text which is used to generate a cover text. Shifting is incorporated by moving the lines of the plain text by some degree in the vertical direction. Bits like 0, 1 and -1 can be used to denote the unmoved, shifted up and shifted down lines. At the decoders end either baseline shifting is found or the centroid shifting is found. Baseline shifting includes the hidden message in the baseline of the adjacent lines in text. If centroid shifting is done then the decoder searches for the hidden message in the centroids of the adjacent lines of the cover text. This technique has problems that if in the transmission of the cover text the lines are shifted or if OCR techniques are applied the hidden information may get destroyed.

4. Suggested Algorithm

The proposed algorithm exploits features of text formatting. When two identical alphabets are colored with successive or similar RGB values, it is visually impossible to differentiate between the two, as shown in Fig.1.

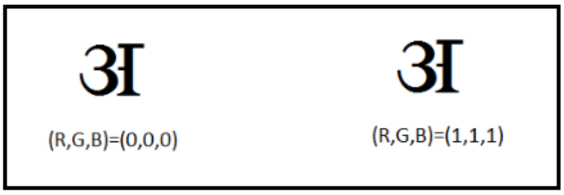

Fig. 1: Indiscernible changes

The above figure shows “Devanagari Letter A” (UNICODE) colored with RGB values 000 and 111, which are visually very similar to one another. This property is used to hide data in a text file. Text with RGB value (0, 0, 0) (hereby named as C0) may be assigned as binary 0, and text with RGB value (1, 1, 1) (hereby named as C1) may be assigned as binary 1. Thus a message 0110 can be hidden as four alphabets having colors (C0, C1, C1, C0). This maintains visual integrity of text file as well as allows detection of message using specially programmed software.

The capacity of this method is however limited due to the use of only two colors during the message encoding process. The storage capacity is directly proportional to number of alphabets present in input file and the number of colors used for hiding text. Thus by increasing number of available colors to 4, namely RGB (0, 0, 0), RGB (1, 1, 1), RGB (2, 2, 2) and RGB (3, 3, 3), the capacity can be doubled as two bits of message are parsed directly. In another variation, 16 colors, RGB (0, 0, 0) to RGB (15, 15, 15) were used to quadruple the storage capacity without disturbing visual integrity of the document.

The same procedure can also be applied to any script as long as text formatting options are available.

5. Result

Implemented algorithm successfully exploits text formatting on Devanagari, Latin scripts and hides data using RGB color values.

Table 1: 1-bit encoding results

Type of Input	Input file size	Output	Capacity (bits)	Message (bits)	Script used	Execution Time (ms)
TXT	1 KB	RTF	220	24	Latin	20
TXT	5.5 MB	RTF	210000	32768	Latin	700

Table 2: 2-bit encoding results

Type of Input	Input file size	Output	Capacity (bits)	Message (bits)	Script used	Execution Time (ms)
TXT	1 KB	RTF	110	24	Latin	22
TXT	5.5 MB	RTF	421615	32768	Latin	680

Table 3: 4-bit encoding results

Type of Input	Input file size	Output	Capacity (bits)	Message (bits)	Script used	Time(ms)
TXT	1 KB	RTF	110	24	Latin	20
TXT	5.5 MB	RTF	843230	32768	Latin	750
RTF	1 KB	RTF	304	32	Devanagari	30
RTF	5.5 MB	RTF	833889	40960	Devanagari	760

Best results were observed when more number of colors were used during the encoding process. Tables 1, 2, 3, describe results of 1-bit, 2-bit and 4-bit color encoding on RTF (Rich Text Format) and TXT (Text) files using Devanagari and Latin scripts.

1-bit encoding, an input file with capacity 220 bits and 210000 bits took 20 ms and 700 ms to execute. 2-bit encoding, an input file of 421615 bits took 680 ms to hide the message. Whereas in four bit encoding, 833889 bit file took only 760 ms to hide the message. Speed and capacity improvements using 4-bit encoding are clearly indicated through this data.

6. Conclusion

This paper presents novel techniques which provide implementation of steganography to text with an extension using features of text formatting. It focuses on innovative ideas which explore a new dimension for further enhancements. The storage capacity of the input file is increased using encoding techniques.

These techniques can be extended to other languages and scripts for broadening the scope of text steganography. This paper unravels new technologies to implement steganography which will prove to be beneficial from the security point of view.

7. References

[1] N Provos and P.Honeyman "hide and seek: an introduction to steganography”, *IEEE security and privacy*, p.p, 32-44 May/June 2003.

[2] T. Morkel, J.H.P. Eloff, M.S. Olivier “AN OVERVIEW OF IMAGE STEGANOGRAPHY”.

[3] W. Bender, D. Gruhl, N. Morimoto, A. Lu “Techniques for data Hiding” *IBM Systems Journal*, Volume 35,Issue 3 and 4,1996,p.p 313-336.

[4] Masoud Nosrati, Ronak Karimi, Mehdi Hariri, “Audio Steganography: A Survey on Recent Approaches”, *World Applied Programming*, Vol (2), No (3), March 2012. 202-205

[5] V. J. Rehna, M. K. Jeya Kumar, “A Strong Encryption Method of Sound Steganography by Encoding an Image to Audio”, *International Journal of Information and Electronics Engineering*, Vol. 2, No. 3, May 2012.

[6] Tayana Morkel, "Steganography and Steganalysis”, *2005.*

[7] J.C Judge, “Steganography: past, present, future”, *Sans white paper*, November 30, 2001

[8] Richard Popa, “An Analysis of Steganographic Techniques”, *The Politehnica University of Timisoara, Faculty of Automatics and Computers*, Department of computer science and Software Engineering,1998.

[9] Mrs. Kalavathi Alla, Dr. R. Siva Ram Prasad,”A Novel Hindi Text Steganography Using Letter Diatrics and its Compound Words”, *IJCSNS International Journal of Computer Science and National Security*, Vol.8 No.12, December 2008.

[10] K. Rabah, “Steganography-The Art of Hiding Data”, *Information Technology Journal*, vol. 3, Issue 3, pp. 245-269, 2004.

[11] S. H. Low, N. F. Maxemchuk, J. T. Brassil, L. O'Gorman, “Document Marking and Identification using Both Line and Word Shifting” , *Proceedings of the fourteenth annual joint conference of the IEEE Computer and Communication Societies*.

Hitesh Singh is an Assistant Professor at Department of Computer Science Engineering at HMR Institute of Technology and Management (HMRITM). He was born in Delhi and completed his Bachelor’s in Technology in 2007, Master’s in Technology in 2009 from C-DAC Noida, and Executive MBA in 2013 from Indian Institute of Technology, Mumbai. His current research interests include Steganography and Telecommunication.

Anirudra Diwakar is pursuing his Bachelor’s in Technology in the field of Computer Science Engineering from HMR Institute of Technology and Management (HMRITM). He was born in Delhi and trained in platforms such as Java, Salesforce and was awarded Academic Excellence Award in 2013 by HMRITM. His curent research interests include Network Security and Steganography.

Shailja Upadhyaya is pursuing her Bachelor’s in Technology in the field of Computer Science Engineering from HMR Institute of Technology and Management (HMRITM). She was born and brought up in Delhi. She was awarded Certificate of Merit in Mathematics by Central Board of Secondary Education in 2008 & graduated from school in 2010. Her training and internships include Salesforce, Java and current research interests are Steganography and Systems Design.