ISSN 1991- 8690 الترقيم الدولي 8690 - 1991

A New Text Steganography Method by Using Non-Printing Unicode Characters and Unicode System Characteristics in English/Arabic documents

Aliea Salman Saber AL - Mozani Wid Akeel Jawad Awadh

Computer Department - Science College - Basrah University

Abstract :

The massive explosion in computer technology and communication has encouraged the development of steganography and water marking, where breathed the spirit of steganography and made it the top security technique to protect and preserve the rights and privacy. So steganography is the one of important types of hide information techniques that has evolved a lot in the past years.

In this paper, we introduce a new approach for text steganography in text documents (word & Excel). In this method, two techniques are marged. First method, use the special Unicode standard characters, zero width non joiner (ZWNJ) and zero width joiner (ZWJ) characters (which have the non-printing properties) to embedding the bits of secret message into English letters, Arabic related letter, Arabic letters separated if they connect, and English/Arabic numbers. Second method; use Unicode system characteristics to embedding the bits of secret message into separate Arabic letters.

This method has high capacity, it can hide one bit in each letter or number in the cover_file, and it satisfies perceptual transparency,by dose not make any apparent changes in the original text. In our methoud we increase the level of security by encode the secret message before embedding it into the cover_text by using Advanced Encryption Standard Algorithm (AES-128).

Keywords: - Word or Excel documents, Cryptography, Unicode standard, Non_printing characters, Steg_word or Steg_Excel, Perceptual transparency.

أحرف اليونيكود غير المرئية وموالصفات نظام اليونيكود

علياء سلمان صابر الموزاني ود عقيل جواد عوض

قسم علوم الحاسبات - كلية العلوم - جامعة البصرة

الخلاصة:

أن ألأنفجار الهائل في تقنية الحاسوب وألأتصالات شجع على أحياء وتطوير الكتابة المخفية والعلامة المائية, حيث نفخت هذه التقنيات بروح الكتابة المخفية وجعلتها تتصدر تقنيات ألأمنية من حماية وضغط وحقوق وخصوصية. لذلك تعتبر الكتابة المخفية من أهم أنواع التضمين الذي تطور كثيرا في السنوات الماضية. أن هذا البحث يتناول طريقة جديده لأخفاء المعلومات في المستندات العربية وألأنكليزية من خلال دمج تقنيتين, التقنية ألأولى: أستخدام حروف من نظام اليونيكود تتصف بأنها غير مرئية عند الطباعة لترميز حروف اللغة ألأنكليزية و ألأرقام العربية وألأنكليزية وألأحرف العربية المتصلة وألأحرف العربية المنفصلة في حالة أتصالها, و التقنية الثانية: أستغلال مواصفات اليونيكود لترميز ألأحرف العربية المنفصلة. هذه الطريقة تمتلك سعة أخفاء عالية حيث تستطيع أخفاء كل بت من بتات الرسائل السرية في كل حرف أو رقم من ملف الغطاء, كذلك فأنها لا تسبب أي تغيير في شكل النص ألأصلي أي أنها تحقق شفافية عالية. في طريقتنا حققنا مستوى عالي من ألأمنية من خلال تشفير الرسالة قبل تضمينها بأستخدام طريقة التشفير القياسي المتقدم (AES-128).

1. Introduction :

The application of computer in real life is increasing every day. So, the need to secure data is becoming more and more essential part of message or data transfer. Information security became a part of our daily life. Among the different techniques, hidden exchange of information is one of the concerns in the area of information security. Various methods like cryptography, steganography, coding and so on have been used for this purpose. However, during recent years, steganography has attracted more attention [1].

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 [2]. The word "*steganography*" is of Greek origin and means "concealed writing" from the Greek words "*steganos*" meaning "covered or protected", and "*graphein*" meaning "to write".

In Steganography, the information is hidden in a cover media so nobody notices the existence of the secret information. Steganography works have been carried out on different medium such as images [3], video clips [4], music and sounds [5].

Steganography scheme is [6]:

Cover_medium + hidden_data + steg_key = steg_medium

There are three important parameters in designing steganography methods: perceptual transparency, robustness and hiding capacity. These requirements are known as "the magic triangle" [7].

Steganography may have different application. For example, it can be used by medical doctors to combine explanatory information with X-ray images. It can be useful in communication for codes self error correction. It can embed corrective audio or image data in case corruption occurs due to poor connection or transmission, copyright protection, preventing _document forging, and other application [8].

There is the major distinction between steganography and other methods of exchange hidden information. For example, in cryptography method, people become aware of the existence of information by observing coded information although they will be unable to comprehend the information. However, in steganography, nobody

will understand the existence of information in the resources [9].

Text steganography is the most difficult kind of steganograhy; this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [10]. The structure of text document is normally very similar to what is seen, while in all other cover media types (audio, picture, video), the structure is different than what we observe, making the hiding of information in other than text easy with out a notable alteration. The advantage of prefer text steganography over other media is its smaller memory occupation and simpler communication, send more information and need less cost for printing as well as some other advantages.

Today, the computer system has facilitated hiding information in texts. The range of using hiding information in text has also developed. From among the most important of these technologies, one can name of hiding information in electronic texts, web pages and documents.

2. Unicode Standard [11] :

The Unicode standard is the international character-encoding standard used for presenting the texts for computer processing. This standard is compatible to the second version of ISO/IEC 10646-1:2000 and have the same characters and codes as ISO/IEC 10646.

Unicode enables us to encode all the characters used in writing the languages of the world. This standard uses the 16-bit encoding, which provides enough space for 65536 characters; that is to say, it is possible to specify and define 65536 characters in different moulds such as numbers, letters, symbols, and a great number of current characters in all different languages of the world. This standerd covers a mathematical and technical symbols, punctuation marks, arrows, and miscellaneous marks. Moreover, because of the wideness of the space dedicated to the characters, this standard also includes most of the symbols necessary for high-quality typesetting. The languages whose writing systems can be supported by this standard are Latin (covering most of the European languages), Cyrillic (Russian and Serbian), Greek, Arabic (including Arabic, Persian, Urdu, Kurdish), Hebrew, Indian, Armenian, Assyrian, Chinese, Katakana, Hiragana (Japanese), and Hangeul (Korean).

An Arabic Unicode table (takes the range 0600-06FF, form_A) represents standard forms of all characters used in Arabic language, and another Unicode table (takes the range FE70-FEFF) represents Arabic presentation forms _B that has all Arabic characters with isolated form .

Unicode table has been developed to cover the characters of the languages which use Arabic writing system. Among these languages we can mention Persian, Urdu, Pashto, Sindhi, and Kurdish. This standard has detailed and careful explanations about the implementation methods including letters-connection method, the exhibition of the right-to-left and bi-direction texts.

3. Advanced Encryption Standsrd (AES) [12]:

This standard specifies the Rijndael algorithm, a symmetric block cipher that can process data blocks of 128 bits, using cipher keys with lengths of 128, 192, and 256 bits. Rijndael was designed to handle additional block sizes and key lengths; however they are not adopted in this standard.

Throughout the remainder of this standard, the algorithm specified herein will be referred to as “the AES algorithm.” The algorithm may be used with the three different key lengths indicated above, and therefore these different “flavors” may be referred to as “AES-128”, “AES-192”, and “AES-256”.

For the AES algorithm, the length of the input block, the output block and the State is 128 bits. This is represented by Nb = 4, which reflects the number of 32-bit words (number of columns) in the State. For the AES algorithm, the length of the Cipher Key, K, is 128, 192, or 256 bits. The key length is represented by Nk = 4, 6, or 8, which reflects the number of 32-bit words (number of columns) in the Cipher Key. For the AES

algorithm, the number of rounds to be performed during the execution of the algorithm is dependent on the key size. The number of rounds is represented by Nr, where Nr = 10 when Nk = 4, Nr = 12 when Nk = 6, and Nr = 14 when Nk = 8. The only Key-Block-Round combinations that conform to this standard are given in table (1).

Table (1): Key-Block-Round combinations

	Key Length (Nk words)	Block Size (Nb words)	Number of Rounds (Nr)
AES-128	4	4	10
AES-192	6	4	12
AES-256	8	4	14

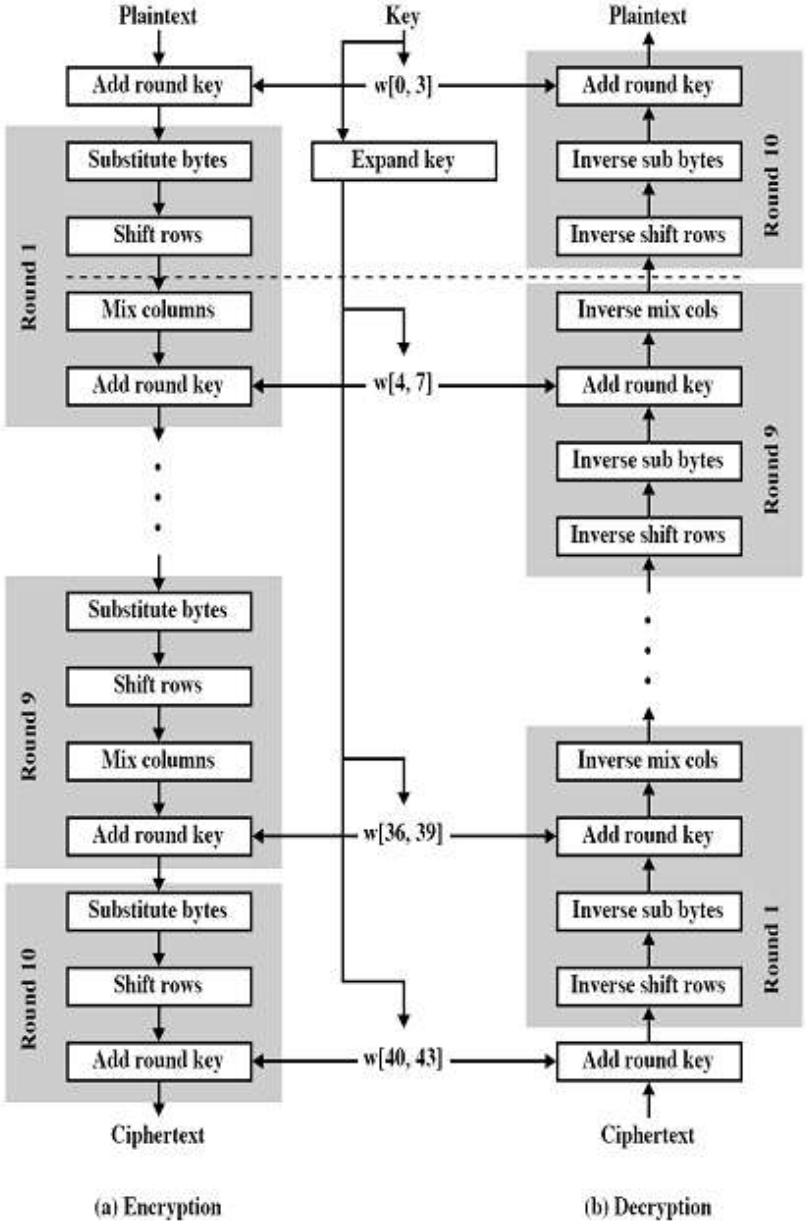

Figure (1): AES Encryption and Deccryption

4. Related Work:

In this section, the text Steganography methods that are especially designed for Persian and Arabic texts are surveyed. To the best knowledge of the authors, there are only two persian and Arabic text Steganography methods that are reported in the literatures.

4.1 Non-Printing Unicode Characters method:

This method is based on special unicode characters, zero width non joiner (ZWNJ) used to disconnect two characters (Unicode = U+200C) and zero width joiner (ZWJ) used to connect two characters (Unicode = U+200D), Which are also known as Pseudo-Space and Pseudo-Connection characters. This method is used by Hassan, Mohammad Shirali-Shahreza [13]. They are introduces a new method for Steganography in Persian and Arabic Unicode texts only.

In this method we hide one bit in each letter. For hiding data in this method, first we look whether the letter of a word is connected to the next letter or not. If it is connected to the next letter, we insert ZWJ letter between two letters for hiding bit 1 and do not add anything for hiding bit 0. Because the letters are connected together, adding ZWJ for connecting the letters together does not have any effects on the apparent of the text. But if the letter is not connected to the next letter, we insert ZWNJ letter between two letters for hiding bit 1 and do not add anything for hiding bit 0. Also in this case the apparent of the word is not changed; because the letters are not connected together and adding ZWNJ for separating the letters from each other does not have any effects on the apparent of the word. For hiding data in the last letter of a word, we always insert ZWNJ letter after it for hiding bit 1 and do not add anything forhiding bit 0.

To extract the information from the text having hidden information (stego text), we respectively investigate the letters of the text words. If after the letter there is ZWNJ or ZWJ character, it means that the bit 1 is hidden in that word. But if after the letter there is no ZWNJ and ZWJ, it means that the bit 0 is hidden in this letter. By putting all the bits of 0 and 1 next to each other we can extract the hidden information from the text.

This technique is not dependent on any special format and they can save the stego text in numerous formats such as HTML pages, Microsoft Word documents or even plain text format. Because the stego Unicode texts will not change during copy and paste between computer programs, the data hidden in texts remains intact during these operations. This technique satisfies both perceptual transparency , because it did not make any apparent changes in the original text by hiding data. And this method has high hiding capacity, because they can hide a bit of information in each Persian and Arabic letter. Disadvantage of this method the size of steg-text is greater than original text.

4.2 Unicode System Characteristics (Similar Letters With Different Codes):

This method is based on Unicode system standard, This technique is used by Hassan, Mohammad Shirali-Shahreza [14], and Auday Fawzi [15].

Hassan, Mohammad Shirali-Shahreza they are proposed a new method for steganography in Persian and Arabic texts by using the Unicode System Standard. In this method, the two characters of « ى» and « ک» have the same shape but different codes if they are used at the beginning or in the middle of words.

The method can be described as follows: If we come across any « ک» or «ى» (used at the beginning or in the middle) in the text, we choose one of the Persian or Arabic characters of « ک» or « ى» considering the information in question for hiding in the text. That is to say, if we want to hide the bit 0, we use the Persian characters of « ک» or « ى» and if we are going to hide the bit 1, we choose the Arabic characters of « ك» or « ي». To extract the information from the text containing information (stego text), we respectively investigate the letters of « ک» and « ى» in the text

(only if these two letters have been used at the beginning or in the middle of a word). If the character is Persian « ک» or « ى», it means that the bit 0 is hidden in the text. If the character is Arabic « ك» or « ي», it means that the bit 1 is hidden in the text. By putting all the bits of 0 and 1 next to each other, we can extract the hidden information from the text.

Auday Faawzi is proposed a new method to hide information in Arabic text only, this technique take each word in the paragraph, and check if there is an isolated character (ا, د, ذ, ر, ز, و), range 0600-06FF(form-A), then replacing it with the same glyph character but with Arabic code, Range FE70 – FEFF(form-B).

The main goal from using this technique is perceptual transparency. This method has an excellent perceptual transparency because the stego_text which the user sees is exactly similar to the original text. However, it is robust to digital copy-past operation, which means that copying and pasting the text between computer programs preserve hidden information. This technique can be used with HTML, Microsoft Word documents or even plain text format. Therefore, the hiding capacity of this method is not very high, because it can hidden bits of secret message in some characters (« ى »,« ک» and isolated characters) only . In addition, this technique is vulnerable to some attacks such as retyping.

5. Proposed Method

In our paper, we introduce a new approach for hide text message in text documents (word & Excel). Our method based on merging previous methods to take advantage of the good features of both techniques and minimize the disadvantage. Through, embedding plain text (English/Arabic (letters or digits), or any symbol in the keyboard) in text documents (Word & Excel) using Non-Printing unicode Characters and Unicode system characteristics, and the level of security of our method increasd by using Advanced Encryption Standard (AES-128) Algorithm for encryption the plain text. Our method work in two stages, embedding stage and extraction stag.

❖ **Proposed algorithm method for embedding :-**

Input: - Cover Microsoft Word Document or Microsoft Excel Sheet , plain-text.

Output: - Stego Microsoft Word Document or Stego Microsoft Excel Sheet.

Step1: Enter plain_text, and then, we encrypt the plain_text into secret message by using (AES-128) Algorithm.

Step2: Open cover_text (Microsoft Word document, or Microsoft Excel sheet).

Step3: Convert the secret message into binary bits (0,1).

Step4: Check if the bits of the secret message less than the bits of the cover_text, if condition is true continue to step 5, otherwise, goto step 7.

Step5: The size of the secret data is hidden at the beginning of text by the process describe in step 6, in order to prevent the extraction of the additional data at the time of the extraction from the stego_text .

Step6: To embed the size and secret data, take each word in the text, and check each letter in a word:

- If the letter is one of the types that we describe in Table (2), we use Non-printing Unicode characters for hiding bit "1" by adding (ZWJ or ZWNJ), and do not add any thing for hiding bit "0".

Table (2): describe laters and type of technique that use to embed it.

Letter	Example	Technique
Arabic isolated letters if they connect	ا،أ،ا،د،ذ،ر،ز،و،ؤ	ZWNJ
English letters	A, b, c ... z A,B,C ... Z	ZWJ
Arabic related letter	ب،ت،ث،ج،ح،خ،س،ش،ص،ض،ط،ظ،ع،غ، ف،ق ، ك ، ل، م ،ن ، ه ، و ، ي	ZWJ
English/Arabic numbers	1,2,3,4,5,6,,7,8,9,0 ٠، ١، ٢، ٣، ٤، ٥، ٦، ٧، ٨، ٩	ZWJ

- if the letter is Arabic isolated letter (ا, أ, د, ذ, ر, ز, و), we use Unicode system characteristics for hiding bit "1" by replacing it with the same glyph character but with Arabic code**, Range FE70 – FEFF,** in Unicode standard, and do not add any thing for hiding bit "0".

Step7: end embedding process

❖ **Proposed algorithm method For Extraction:-**

Input: - Stego-Microsoft Word document or Stego-Microsoft Excel sheet.

Output: - Extracted plain_text.

Step1: Open stego Microsoft Word document or Stego Microsoft Excel sheet.

Step2: Extract the size of hidden data that embedded at the beginning of the text and extract the hidden data by the process describe in step 3.

Step3: Take each word in the text, and check each letter in the word:

- If the letter is (ZWJ or ZWNJ or any character in **Range FE70 – FEFF** that mean the bit "1" has been hidden in the text.
- Else that mean the bit "0" has been hidden in the text.

Step4: Collacate the sequence of extracted bits.

Step4: Decoded the secret bits into plain text by (AES-128) decryption algorithm.

Step5: End extraction process.

6. Experiment Result :

In our method, the information can be hidden in text documents using Non-Printing Characters and Unicode system characteristics**.** This approach will embed one bit in each English/Arabic letter or digit in cover-text after encryption the secret message by (AES) algorithm.

We tested our method on some word & Excel files. The type of files and the capacity of each text for hiding data are shown in table (3). When comparing our method with previous methods we note that, our approach use all English/Arabic characters and numbers in the process of embedding, so the capacity is very high. Our approach like the pervious method that does not change the apparent of the text and does not required specific font, so the robustness is very high. Also, our method has high security, by increase the level of security by encryption the message and then embedding the secret message into the cover-text.

In case of printing the stego_text containing the hidden data, the hidden data will be lost, because, as mentioned earlier, due to the hiding of the data in the text, the text appearance remains unchanged and only the internal structure of the saved file is changed. Consequently, the hidden data will be lost because of losing the internal data of the file and we cannot extract the hidden data from the printed copy of the text. As a result, this method is limited to hiding data in electronic documents (e-document).

Table (3): Example of text steganography method proposed:

Secret Message	Cover File		
Number of bits	File Type	Number of characters	Capacity (%)
128 BIT	Word	593	21.5 %
1152 BIT	Word	2283	50.4 %
2688 BIT	Word	2821	95.2 %
4224 BIT	Word	5929	71.2 %
5632 BIT	Word	5929	94.9 %
128 BIT	Excel	477	26.8 %
512 BIT	Excel	797	64.2 %
1152 BIT	Excel	1180	97.6 %
1792 BIT	Excel	2996	59.8 %
2816 BIT	Excel	3159	89.1 %

7. Conclusion :

In this paper, anew approach for hide text message in (word & Excel) documents is presented by merging two techniques, First technique**,** Non-Printing unicode characters and second technique, Unicode system characteristics**.**

In this method, we hide data in Microsoft Word documents or Microsoft Excel sheets by embedding each bit in each (English/Arabic) letter or numbers by using non-printing characters or Unicode system characteristics based on the type of character or digit. Microsoft Word document is very much common in every day life of today digital world.

The main goal in designing this method is perceptual transparency. Our method has an excellent perceptual transparency because the stego_text which the user sees is exactly similar to the original text. Therefore, the hiding capacity of our method is very high. In addition, our method is vulnerable to some attacks such as retyping. However, it is robust to digital copy-past operation, which means that copying and pasting the text between computer programs preserve hidden information.

References:

[1]E. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Information Hiding -A Survey”, in Proceedings of the IEEE, 87(7)(1999), pp. 1062–1078.

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

[3] M. Shirali Shahreza, “An Improved Method for Steganography on Mobile Phone”, WSEAS Transactions on Systems, vol. 4, Issue 7, July 2005, pp. 955-957.

[4]G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", Signal Processing:

Image Communication, vol. 18, Issue 4, 2003, pp. 263-282.

[5]G. Doërr and J.L. Dugelay, "Security Pitfalls of Frameby-Frame Approaches to Video Watermarking", IEEE Transactions on Signal Processing, Supplement on Secure Media, vol. 52, Issue 10, 2004, pp. 2955-2964.

[6]Kessler, c. Gary, "An Overview of Steganography", the Computer Forensics Examiner issue of Forensic Science Communications, July 2004.

[7]N. Cvejic, Algorithms for Audio Watermarking and Steganography. Finland: Oulu University Press, 2004.

[8]N. F. Maxemchuk and S. Low, "Marking Text Documents", in Proceedings of the IEEE International Conference on Image Processing, Santa Barbara, CA, USA, 1997, pp. 13–16.

[9]J.C. Judge, "Steganography: Past, Present, Future", SANS white paper, November 30,2001, http://www.sans.org/rr/papers/index.php?id=552.

[10]Memon,jibran a. and khowaja, k. and K. Hameedullah, "EVALUATION OF STEGANOGRAPHY FOR URDU /ARABIC", Journal of Theoretical and Applied Information Technology, 2008.

[11]The Unicode Standard, URL: http://www.unicode.org, last visited: 31 September 2011.

[12]Specification for the ADVANCED ENRYPTION STANDARD (AES), Federal Information Processing Standard Publication 197, November 26, 2001.

[13]M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza, "STEGANOGRAPHY IN PERSIAN AND ARABIC UNICODE TEXTS USING PSEUDO-SPACE AND PSEUDO CONNECTION CHARACTERS", The Arabian journal for science and engineering, Iran, 2008, Pages: 682-687.

[14]M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza, "ARABIC/PERSIAN TEXT STEGANOGRAPHY UTILIZING SIMILAR LETTERS WITH DIFFERENT CODES", The Arabian journal for science and engineering, volume 35 , number 1B, Pages: 214-222, April 2010.

[15]Auday Jamal Fawzi, "Data Hiding in Arabic Text", University of Technology / Baghdad, 2007.