A Novel Arabic Text Steganography Method Using Letter Points and Extensions

Adnan Abdul-Aziz Gutub, and Manal Mohammad Fattani

***Abstract*—**This paper presents a new steganography approach suitable for Arabic texts. It can be classified under steganography feature coding methods. The approach hides secret information bits within the letters benefiting from their inherited points. To note the specific letters holding secret bits, the scheme considers the two features, the existence of the points in the letters and the redundant Arabic extension character. We use the pointed letters with extension to hold the secret bit 'one' and the un-pointed letters with extension to hold 'zero'. This steganography technique is found attractive to other languages having similar texts to Arabic such as Persian and Urdu.

***Keywords*—**Arabic text, Cryptography, Feature coding, Information security, Text steganography, Text watermarking.

I. Introduction

STEGANOGRAPHY**,** in today's electronic era, is the ability of hiding information in redundant bits of any unremarkable cover media. Its objective is to keep the secret message undetectable without destroying the cover media integrity. Steganography replaces unneeded bits in image, sound, and text files with secret data. Instead of protecting data the way encryption does, steganography hides the very existence of the data [12].

Capacity, security, and robustness [1], are the three main aspects affecting steganography and its usefulness. Capacity refers to the amount of data bits that can be hidden in the cover medium. Security relates to the ability of an eavesdropper to figure the hidden information easily. Robustness is concerned about the resist possibility of modifying or destroying the unseen data.

Steganography is different than cryptography and watermarking although they all have overlapping usages in the information hiding processes [12]. Steganography security hides the knowledge that there is information in the cover medium, where cryptography revels this knowledge but encodes the data as cipher-text and disputes decoding it without permission; i.e., cryptography concentrate the challenge on the decoding process while steganography adds the search of detecting if there is hidden information or not. Watermarking is different from steganography in its main goal. Watermarking aim is to protect the cover medium from any modification with no real emphasis on secrecy. It can be observed as steganography that is concentrating on high robustness and very low or almost no security.

Steganography may have different applications. For example, it can be used by medical doctors to combine explanatory information within X-ray images. It can be useful in communications for codes self-error correcting. It can embed corrective audio or image data in case corruption occurs due to poor connection or transmission. Steganography may be practical to form a secure channel for private communication, however, it does not cover the fact that the communication happened or the data is hidden. This makes steganography as a special technique of encryption or cryptography [3].

Steganography can also be utilized for posting secret communications on the Web to avoid transmission or to hide data on the network in case of a violation. It can be useful for copyright protection, which is, in reality, digital watermarking [12]. Copyright protection is to protect the cover medium from claiming its credit be others, with no real emphasis on secrecy.

Most of steganography research uses cover media as pictures [4], video clips [5] and sounds [6]. However, text steganography is not normally preferred due to the difficulty in finding redundant bits in text files [2,7]. The structure of text documents is normally very similar to what is seen, while in all other cover media types, the structure is different than what we observe, making the hiding of information in other than texts easy without a notable alteration. The advantage to prefer text steganography over other media is its smaller memory occupation and simpler communication [2].

Languages and their structures play differences in the preferred steganographic system. Normally no single technique is to be used for all languages [12]. Section 2 presents several schemes used for hiding data within electronic English text files. A specialized published [2] method for hiding information in Arabic texts is detailed in Section 3. Section 4 discusses our new Arabic text steganography technique using character extensions. The conclusion is presented in Section 5.

Manuscript received March 4, 2007. This work is discussed by the Crypto-Group at Computer Engineering Department of King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia.

Adnan Abdul-Aziz Gutub is with the Computer Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia (phone: +966-3-860-1723; fax: +966-3-860-3059; e-mail: gutub@kfupm.edu.sa).

Manal Mohammad Fattani is a third level student with the Department of Islamic & Arabic Studeis, Girls Collage in Dammam, Saudi Arabia.

II. Text Steganography Techniques

Some researches on hiding information in texts have been performed. Different methods, as examples, are presented in this section.

A. Particular Characters in Words

Hiding information can be performed by selecting characters in certain words. This method can range from simple to very complicated depending on the specifications. "In the simplest form, for example, the first words of each paragraph are selected in a manner that by placing the first characters of these words side by side, the hidden information is extracted" [2]. A more advanced example can be by selecting the first letter from the first word, second letter from the second word, third from the third, and so on, to hide the information in.

B. HTML Documents

Secret information can be hidden within HTML Tags [8] since they feature case insensitivity. For example, the tags <p align="center">, <p align="cenTER">, <p align="Center"> and <p aLigN="center">, are all similarly valid. Steganography in HTML documents can be done by varying the small or large case letters in document tags. Extraction of information can be by comparing these tags words with words in normal case. This HTML steganography security can be increased by choosing a certain letter sequence function. For example, the third capital letter within the tags, where most tags should have several randomly altered letters so eavesdropper is confused.

C. Line and Word Shifting

Shifting text lines vertically and shifting words horizontally [9] may help in hiding some information. Security of this method depends on the availability of varying the distances between words and lines to puzzle intruders. This method of steganography shifts the lines up or down slightly with a fixed space (say 0.003 inch) and modifies the distances between words, according to the intended hidden information. This text shifting steganography depends on constructing visual shapes for information to be hidden in spaces. The technique is appropriate for texts printed, since it faces problems against robustness. Whenever, the text is electronically rewritten or modified, there is great possibility for the hidden information to be destroyed. Furthermore, when using character recognition programs, such as OCR, the visual shapes hiding information are lost or cannot be traced accurately.

D. Abbreviations and Spaces

Abbreviations and spaces steganography can hide very little information in the text [7]. For instance, "only a few bits can be hidden in a file of several kilobytes" [2].

Space steganography, in particular, hides information by adding extra white-spaces between words, or at end of lines or paragraph of the text [7]. This technique may be used with any text and does not reveal secrecy to the normal reader, its security is good. However, its capacity and robustness is low. The method cannot hide too much information and some electronic text editors automatically remove extra white-spaces.

E. Semantic and Character Feature Methods

To protect hidden information among electronic retyping or OCR usage problems of the previous shifting approach, semantic [8] and character feature [10] steganography methods are suggested. Semantic method proposes using synonyms of words for certain words as for hiding information in the text. However, this method may alter the meaning of the text which will change the intended hidden information [2].

Character feature steganography changes some of the features of the text characters. For example, the most significant bits of some characters are extended to hold bits of the hidden information [10]. Character steganography can hold a large quantity of secret information without making normal readers aware of the existence of such information in the text.

III. Published Security Method Specified for Arabic Text

Shirali-Shahreza [2] proposed a special character feature security method for Arabic and Persian letters. Their scheme depends on the points inherited in the Arabic and Persian letters [11], which are some who very similar. The concentration in this study will be on the steganography related to Arabic language.

Although, both Arabic and English languages have points in their letters, the amount of pointed letters differ too much. English language has points in only two letters, small "i" and small "j", while Arabic has in 15 letters out of its 28 alphabet letters as shown in Fig. 1. This large number of points in Arabic letters made the points in any given Arabic text remarkable and can be utilized for steganography and information security as presented by Shirali-Shahreza in their "new approach to Persian/Arabic text steganography" [2].

un-pointed letters	pointed letters
ا ح د ر	ب ت ث
س ص	ج خ ذ ز
ط ع ك	ش ض
ل م هـ و	ظ غ ف
	ق ن ي

Fig. 1 Arabic letters

Shirali-Shahreza [2] point steganography hides information in the points of the letters. To be specific; they hide the information in the points' location within the pointed letters. First, the hidden information is looked at as binary with the first several bits (for example, 20 bits) to indicate the length of the hidden bits to be stored. Then, the cover medium text is scanned. Whenever a pointed letter is detected its' point location may be affected by the hidden info bit. If hidden value bit is one the point is slightly shifted up; otherwise, the concerned cover-text character point location remains unchanged.

This point shifting process is shown in Fig. 2 for the Arabic letter 'Fa'. "In order to divert the attention of readers, after hiding all information, the points of the remaining characters are also changed randomly" [2]. Note that, as mentioned earlier, the size of hidden bits is known and also hidden in the first 20 bits.

This method of point shifting may have its advantages in security and capacity; it features good secret storing of large number of hidden bits within any Arabic text. However, it has main drawback in robustness making it unpractical. For example, the hidden information is lost in any retyping or scanning. The output text has a fixed frame due to the use of only one font. In fact, this information security method is appropriate to be classified as watermarking instead of steganography.

Fig. 2 Point shift-up of Arabic letter 'Fa'

IV. Proposed Arabic Steganography

Benefiting from Shirali-Shahreza [2] point steganography and trying to overcome the negative robustness aspect, we propose a new method to hide info in any letters instead of pointed ones only. We use the pointed letters with extension to hold secret bit 'one' and the un-pointed letters with extension to hold secret bit 'zero'. Note that letter extension doesn't have any affect to the writing content. It has a standard character hexadecimal code: 0640 in the Unicode system. In fact, this Arabic extension character in electronic typing is considered as a redundant character only for arrangement and format purposes.

The only bargain in using the extension is that not all letters can be extended with this extension character due to their position in words and Arabic writing nature. The extension can only be added in locations between connected letters of Arabic text; i.e. extensions cannot be placed after letters at end of words or before letter at beginning. Our proposed steganography hypothesis is that whenever a letter cannot have an extension or found intentionally without extension it is considered not holding any secret bits.

This proposed steganography method can have the option of adding extensions before or after the letters. To be consistent, however, the location of the extensions should be the same through out the complete steganography document.

Assume we add the extensions after the letters. Fig. 3 shows an example to detail this steganography process. We first select the secret bits to be hidden (say 110010) looking from the least significant bits to be started with. The first secret bit found is '0' to be hidden in an un-pointed letter. The cover-text is scanned from right to left due to Arabic regular direction. The first un-pointed letter in the cover-text is found to be the first, known as 'meem'. This 'meem' should hold the first secret bit '0' noted by adding extension character after it.

The second secret bit is '1' and the second letter of the cover-text is pointed known as 'noon'. However, this letter position cannot allow extension, forcing us to ignore it. The next possible pointed letter to be extended is 'ta'. Note that a pointed letter 'noon' before 'ta' is not utilized due to its unfeasibility to add extension character after it.

Secret bits	110010
Cover-text	من حسن اسلام المرء تركه مالا يعنيه
Steganographic text	من حسن اسلام المرء تركه مالا يعنيه 1 1 0 0 1 0

Fig. 3 Steganography example adding extensions after letters

The same steganography example of securing: 110010 in the Arabic text, illustrated earlier, is readjusted assuming the extensions added are before the letters, as shown in Fig. 4.

To add more security and misleading to trespassers, both options of adding extensions before and after the letters can be used within the same document but in different paragraphs or lines. For example, the even lines or paragraphs use steganography of extensions after the letters and the odd use extensions before or visa versa.

Secret bits	110010
Cover-text	من حسن اسلام المرء تركه مالا يعنيه
Steganographic text	من حسن اسلام المرء تركه مالا يعنيه 1 1 0 0 1 0

Fig. 4 Steganography example adding extensions before letters

V. Conclusion

This paper presents a novel steganography scheme useful for Arabic language electronic writing. It benefits from the feature of having points within more than half the text letters. We use pointed letters to hold secret information bit 'one' and the un-pointed letters to hold secret bit 'zero'. Not all letters are holding secret bits since the secret information needs to fit in accordance to the cover-text letters. Redundant Arabic extension characters are used beside the letters to note the specific letters holding the hidden secret bits. The nice thing

about letter extension is that it doesn't have any affect to the writing content.

This method featured security, capacity, and robustness, the three needed aspects of steganography that makes it useful in hidden exchange of information through text documents and establishing secret communication. This steganography technique is also useful to other languages having similar texts to Arabic such as Persian and Urdu scripts, the official languages of Iran and Pakistan, respectively. These characteristics and features promises that this novel Arabic text steganography method using letter extensions attractive for information security.

ACKNOWLEDGMENTS

Thanks to King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, for its support of all research work. Thanks to Dr. Alaaeldin Amin, Dr. Talal Al-Kharobi, and the COE 509 students of the Applied Cryptography course for all their inputs, feedback, and comments.

REFERENCES

[1] B. Chen and G.W. Wornell, "Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding," *IEEE Trans. Information Theory*, Vol. 47, No. 4, pp. 1423-1443, 2001.

[2] M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography," *5th IEEE/ACIS International Conference on Computer and Information Science (ICIS-COMSAR 06)*, pp. 310- 315, July 2006.

[3] J.C. Judge, "Steganography: Past, Present, Future", *SANS white paper*, http://www.sans.org/rr/papers/, November 30, 2001.

[4] R. Chandramouli, and N. Memon, "Analysis of LSB based image steganography techniques", *Proceedings of the International Conference on Image Processing*, Vol. 3, pp. 1019 – 1022, Oct. 2001.

[5] G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", *Signal Processing: Image Communication*, Vol. 18, No 4, pp. 263-282, 2003.

[6] K. Gopalan, "Audio steganography using bit modification", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '03), Vol. 2, pp. 421-424, April 2003.

[7] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM Systems Journal*, Vol. 35, No 4, pp. 313-336, 1996.

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

[9] 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 Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '95)*, April 1995.

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

[11] 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.

[12] N. Provos and P. Honeyman, "Hide and Seek: An Introduction to Steganography", *IEEE Security & Privacy*, pp. 32-44, May/June 2003.

Adnan Abdul-Aziz Gutub obtained his PhD degree in Electrical and Computer Engineering from Oregon State University in 2002. He has been appointed as an Assistant Professor in Computer Engineering Department at KFUPM. His research interests have been in modeling, designing, and implementing crypto arithmetic operations in hardware where in 2005 he has been awarded the summer research grant through the British Council to work with a group at Brunel University in the UK.

Adnan Gutub is an official reviewer for conferences and journals of IEEE, IET (IEE), and CHES. He is a member of the following technical groups:

- Security Research Group, Information and Computer Science Department, KFUPM
- Bio-Inspired Intelligent Systems Team, Brunel University, UK
- Cryptography Research Group, Computer Engineering Department, KFUPM
- IET (IEE) Computers & Digital Techniques
- Cryptographic Hardware and Embedded Systems (CHES) Research Group
- Information Security Laboratory, at Oregon State University, Corvallis, Oregon, USA

He had been involved in organizing several scientific events such 18th IEEE International Conference on Microelectronics (ICM), First e-Service Symposium, 6th Saudi Engineering Conference, and 10th Annual IEEE KFUPM Technical Exchange Meeting.