Arabic Language text Steganography Besed on Microsoft Word documents

Assist.prof. Dr. Hanaa M. Ahmed
Computer Science Department
University of Technology
Baghdad, Iraq
salmanhanna2007@yahoo.com

PhD Students: Maisa'a A. A. Khohder
Computer Science Department
University of Technology
Baghdad, Iraq
maisaa.ali2007@yahoo.com

Abstract—

The fast evolution of internet innocent over communication in the net environment become an important research trend. Steganography is meaning that secure information is embedded into cover data unnoticeably for transmitted. Linguistic steganography is covering all the techniques that deal with using written natural language to hide secret message. This paper, it presents a linguistic steganography in arabic texts, using single quotation mark and double quotation mark that transform on the bases of using new technique entitled RSVD Image as allocation to hide secret message. The proposed that approach is an attempt to present a transform linguistic steganography using levels for hiding to improve implementation of quotation mark, and improve the security of the secret message using RSVD Image. The proposed algorithm is achieved perfect steganography properties such as capacity, security, transparency, and robustness.

Keywords— Arabic text, Linguistic Steganography, random Singular Value Decomposition(RSVD), quotation mark, Transform Based

اخفاء النصوص العربية بأسخدام وثائق المايكروا سوفت ورد

أ.م.د. هناء محسن احمد / م. ميساء عبد علي خضر / قسم علوم الحاسوب / الجامعة التكنولوجية

الملخص

ان تطورات السريعة في الانترنت والاتصالات في بيئة الشبكات اصبحت مهمة في اتجاه البحوث . اخفاء المعلومات بمعنى سرية المعلومات وتضمينها داخل غطاء البيانات بدون ملاحظة اثناء النقل .

اخفاء المعلومات اللغوية يغطي كل التقنيات التي تتعامل معها واستخدام كتابتها باللغات الطبيعية لإخفاء الرسالة السرية .

في هذا البحث ، استخدم اخفاء المعلومات اللغوية داخل النصوص العربية ونستخدم علامة quotation الفردية والمزدوجة في التحويل وتستخدم هذه الطريقة تقنية جديدة بعنوان فك القيم الفردية العشوائية للصورة كما في تحديد مواقع اخفاء الرسالة السرية .

الطريقة المقترحة هي محاولة تحويل اخفاء المعلومات اللغوية الموجودة باستخدام مستويين للإخفاء وأهميتها بارتقاء في تنفيذ علامة quotation والارتقاء في امنية الرسالة السرية بواسطة استخدام فك القيم الفردية العشوائية للصورة .

وان هذه الخوارزمية المقترحة هي اساسية وممتازة في خصائص اخفاء المعلومات مثل السعة والأمنية ، الشفافية ، والقوة .

مفاتيح الكلمات

النصوص العربية ، اخفاء المعلومات اللغوية ، فك القيم الفردية العشوائية ، علامة quotation ، اساس التحويل.

Introduction

Linguistic steganography is focused on apply changes to a cover text so as to embed secret message, in a way that the changes do not caused any unnatural or ungrammatical text. According to cover, text steganography can be categorized at three sets [1], as depicted in Figure (1), [2]:

1. Format-based methods: is use and change the formatting of the cover-text to hide data [2]. “It involves modifying physically the form of text to hide data. This manner have some flaws. When the stego file is opened in microsoft word handling, error spell and additional white spaces will bring discovery. Varied fonts sizes can burn suspiciousness to a human reader [3]. Format- based methods usually alter existing text for hiding the steganography text, by insertion of space or non-displayed, as depicted in Figure (2) [5].

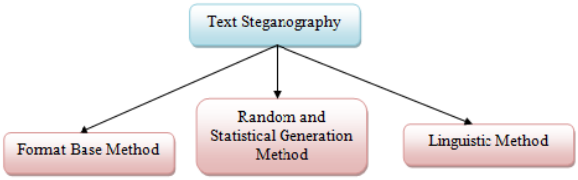


Figure (1): The types of Text Steganography [1].

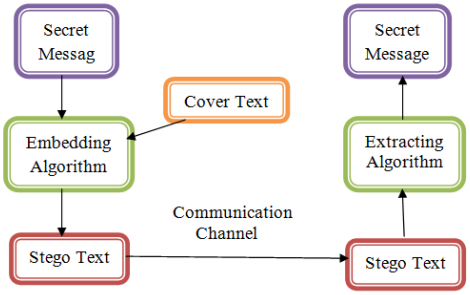


Figure (2): Mechanism of Text Steganography [5].

- There are many approaches of format-based:

1- Line Shift approach:

This approach secure letter is hidden via perpendicularly shifting the text lines to some degree (1/300 inch up or down) [5]. A line sign have two unsign monitoring lines one on each part of it discovery the trend of move of the sign line. When it hide zero bit, a line is shifted up and when it hide one bit , the line is shifted down. Design of case the line have shifted up or down is done via measure the range of the center of sign line and its monitoring lines. when the text is rewriting or applied a character recognition program (OCR). [4, 6].

2- Word Shift Approach:

This approach is modifying a scrip via horizontally shifting the positions of words inside text lines to encode the script uniquely.

"This encryption enable applied to any file or to the bitmap of a page image". Decoding might performed at the file or bitmap [5]. Secure letter is hidden via shifting the words horizontally, i.e. "left or right to represent zero bit or one bit separately" [4]. "Also rewriting of the text or using OCR programs broken the hidden information" [5, 6].

3- Feature Coding Approach:

This is a coding manner that is applied whereas a file or to picture of a script. The picture can be testing for selection text face, and those faces are modified, or not be modified, depending on codeword [5]. And saying, "dot in messages i and j not ability in place, extent of strike in messages f and t can be changed, or by extending or lessening rise of messages b, d, h". The failings of this manner is that if an OCR program is applied or when re-writing is well-done, the hidden content becomes broken [4, 6].

4- White Spaces Approach:

The white spaces is method uses for hiding a secret letter. There are three parts for hiding data exploitation white spaces. "whereas enter statement spacing, when site single

space into hide zero bit and two spaces to hide one bit at the end of each ending letter" [4].

Space steganography, hides data via white-spaces among statement adding extra, or end of lines or paragraph of the text [6]. This method can implemented on arbitrary text and not observe attention of the reader [6].

In this paper, the propose layers steganography technique for arabic language text using quotation mark. The proposed approach use Singular Value Decomposition (SVD) as generate random location, to embed the secret message bits using single-double quotation layer steganography followed by add quotation mark randomly . The proposed algorithm perfect steganography properties such as capacity, transparency, robustness, and security of the secret message for arabic text based secure communication.

The other sections of the paper are structured as: section II presents the literature review for quotation mark based linguistics steganography and explain fundamental used of proposed system. Section III explain algorithm for proposed system and results and discussions are done in section V, and IV deals with the conclusions.

I. Literature review and fundamental used in proposed system

A. Literature review

Kashida is an Arabic redundant character which is used to justify the text, without affect the meaning of words. The researchers suggested using one kashida as bit zero, and two kashida as bit one, or vice versa.

In 2007, Adnan Abdul-Aziz Gutub, and Manal Mohammad Fattani, introduced an enhanced Arabic script steganography technique for Arabic text using hides secure data bits into messages making use two sets points. Hiding a secret message in cover text. Where use the pointed character for expansion convene if secret bit ‘one’ and the un-pointed character for expansion convene ‘zero’. That manner

enhanced security, capacity, transparency and robustness for Arabic texts based secure communication [7].

In 2009, M. Hassan Shirali-Shahreza and Mohammad Shirali-Shahreza, introduced a novel manner to conceal data in Persian (Farsi) and Arabic languages. In Unicode scale, there are to conceal two characters for 'Ya' (ی) and 'Kaf' (ک). The two characters of (ی) and (ک) has itself shape however various codes which they are utilized at the start or in the center of words. The major aim in this manner is perception translucence. It have a stellar perception translucence cause the stego-text who the employee sight is aright like for the main text [8].

In 2009 Ammar Odeh and Khaled Elleithy, introduce on algorithm that have been used some of general characteristics Arabic texts, that explain as few vowel letters. Arabic Diacritics is a volitional ownership to every text and commonly is not at all used. This algorithm use ownership to conceal datum and minimize the prospect of suspicions conceal. And utilize a rendering metrical content, the file volume before and after extra Diacritics and possibility to conceal data together existence. conceal data in various covering media appear one of defy safety causes. The file volume and style alteration ought to growing the eventuality to existence detected utilize tools of steganalysis and that ought to command for detect the conceal data [9] .

In 2012 Vasil Kolev, Katya Tsvetkova, and Milcho Tsvetkov, introduced to virtual the m×n picture A from inspected astronomical photo plateful via utilize onward littel into than in the main model. By utilized degree of a model, k it delete the excessive data or fuss and used such as Wiener filter, when degree k < m or k< n. This virtual unless than 98% pressure ratio of picture of sidereal plateful wanting that picture specific, is acquired. The SVD in picture for astronomical photo plateful (SPP) that has considered and it possible picture pressure [10].

B. Carrier Arabic File Text in Microsoft Documets(CAFTMD)

The Arabic scripts ownership into employ about of its feature to conceal great offset of information into Unicode file (Arabic scripts). In this research using Arabic text without Harakat and using each character eight
bits. The secret message carrier also after convert to binary bit each

character is eight bits also. And using search algorithm to hide bit of secret message to found position location.

C. Singular Value Decomposition (SVD)

Singular Value Decomposition technique splits given matrix for output of orthonormal matrices and a diametrical matrix. The mathematical formula to Singular Value Decomposition is [11].

$$A = USV^T \quad \ldots\ldots (1)$$

$$A = [U1 \;\; U2 \;\; \ldots\ldots Um] \begin{bmatrix} s1 \ldots\ldots & 0 \ldots\ldots\ldots & 0 \\ 0 \ldots\ldots & s2 \ldots\ldots\ldots & 0 \\ 0 \ldots\ldots & \ldots\ldots\ldots & sn \end{bmatrix} \begin{bmatrix} v1T \\ v2T \\ . \\ . \\ VnT \end{bmatrix} \quad , \quad \ldots\ldots (2)$$

Let A be an *m*× *n* matrix. Performing SVD to A factorizes it for output of perpendicular matrix, a diametrical matrix and others perpendicular matrix as:

$$A = USV^T \quad \ldots\ldots (3)$$

Where,

A: original image matrix
U: *m*× *m* product of orthogonal matrix
S: *m*× *n* diagonal matrix
V: *n* × *n* orthogonal matrix

D. Random Singular Value Decomposition (RSVD)

Is a new technique to generate a set of random position (x_i, y_j) to apply the embedding algorithm, from decomposing original image (A) using SVD, the result is B. Detect the non-zeros elements, and converges into nearest integer the results is RSVD. Figure (3) is an example for the original image.

Figure (3): Original image example.

The original image A is decomposed into three matrices: U of size mi × mj matrix, V of size mi × nj matrix, and D of size ni × nj matrix. The new array B = S*V*D as depicted in Figure (4), indicates random location from original image A.

Figure (4):Location generated using RSVD.

In this algorithm, RSVD process to find the position for hide secret message bit depended on the partial image for SVD randomly. Explain algorithm below:

Algorithm RSVD:
Input: original image A Output: random number Process: Step1: input A Step2: Apply SVD algorithm to A Step1: $B = U*S*V^T$ Step2: For i= 1 to Length of B Step3: No= B[i]; Step4: No= absolute (No); Step5: While ((integer (No==0) Step6: No =No*10; Step7: RSVD[i]=(integer (No)); Step7: Next End of algorithm

III. proposed system

A. Idea for proposed system

The proposed approach main idea as depicted in Figure (5) the embedding, and Figure (6) the extraction, is to use RSVD as generated random offset location, to add random single or double quotation words to the rest Arabic word texts where that inject the secret message bits in cover Arabic text word, and then apply single quotation words when a secret message is 0, but apply double quotation words when a secret message is 1.

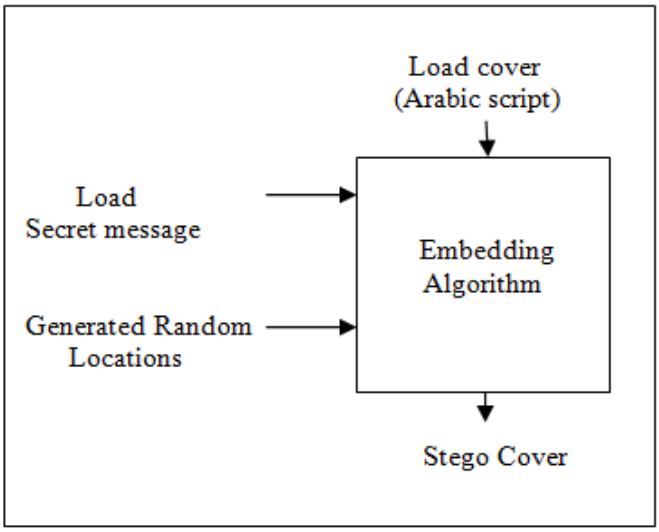


Figure (5): The proposed hiding process.

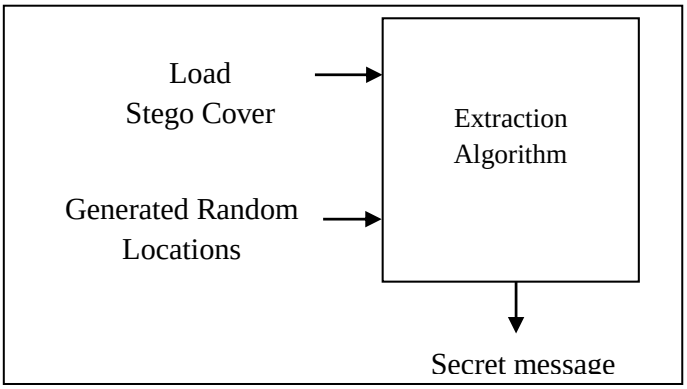


Figure (6): The proposed extraction process.

In this paper, using two algorithms: the first algorithm is embedding process. The Figure (7) represent the binarization process secret message. Figure (8) is a simple example of applying binarization process to secret message.

B. Embedding process

Embedding Algorithm:

Input: secret message, N, image A, set of Arabic texts.

Output: stego-cover.

Process:

Step1. Secret message is hiding in form (0)s, and (1)s, which represent each character is 8-bits, and represent N, is the total number of secret message.

Step2. Generate Random positions: The process of generated random positions, using RVSD, starts by using applying SVD algorithm to the input image (A) to generate a sequence of random values **C**

that represents offset of Arabic text words to start the embedding process. The total number of Generate Random positions is **(N),** where **N,** is the total number of secret message bits.

Step3. Cover selection: select Arabic text (cover) that can hold input secret message bits.

Step4. Do while not end of Arabic text words

Step5. Embedding secret message: For each secret message bit and Generate Random positions do

step6. Using hide bit in the words, the length words greater than 5.

Step7. Apply quotation if secret message bit is one put double quotation or if secret message bit is zero put single quotation.

Step8. End For

Step9. End

The embedding a secret message binarization depicted in Figure (7) and Figure (8).

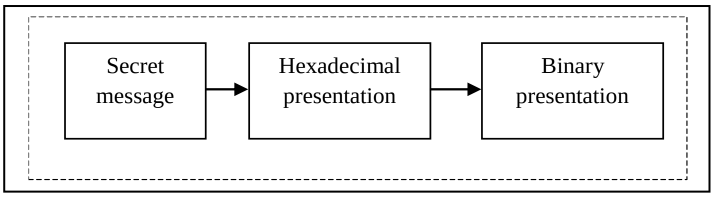


Figure (7): Secret message Binarization.

Secret Message	---------------- Arabic Text ---------------- الجمع بين الماء والنار في يذيب أصعب
Hexadecimal Representation	--------------- Hexa Decimal --------------- 20C7FBADEFDF2020A9FDF22020C7FBEFC7C12020E8C7FBF2C 7D12020BAFD2020FDCFFDA92020C3BEDFA9
Binary Representation	001000001100011111111101110101101111011111110111110010 000000100000101010011111111011111001000100000000100000 110001111111101111110111111000111110000001001000000010 0000111010001100011111111011111100101100011111010001 001000000001000001011101011111101001000000001000001111 1101110011111111110110101001001000000001000001100011 101111101101111110101001

Figure (8): Secret message binarization Example.

The second algorithm is extraction process using to extract secret message from cover arabic language text, depended on the position from RSVD.

C. *Extraction process*

Extraction Algorithm:

Input: secret message, image A, stego cover.

Output: secret message.

Process:

Step1. Generate Random positions: The process of generated random positions, using RSVD, start by using applying SVD algorithm to the input image (A) to generate a sequence of random values c

that represents offset of Arabic text words to start the extraction
process.
Step2. Loading: Load stego-cover, and Generate Random positions.
Step3. For each Generate Random Positions do
Step4. Extract secret message from stego-cover when found single or
Double quotation in Arabic text word.
Step5. End of For.
Step6.Converts each eight bits into one letter the result is the secret message.
Step 7. End

IV. RESULTS AND DISCUSSION

In this section discuss two cases to ensure the proposed technique security:

Case one: An example result for applying the proposed technique using
embedding secret message using single-double quotation, as depicted in
Figure (9) and figure (10). This case given high level information hiding
security, after apply Jaro-Winkler method, as depicted in table (1) and table (2).

- The Jaro-Winkler method is distance measures the similarity between two strings.

The Jaro distance is: $dj = \frac{1}{3}(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m})$

t = max{[|S1|,|S2|]/2}-1. Explain in table (1) below:

If the word is القدماء without stego, dj=1/3(7/7+7/7+7-2/7)=0.9047
where t = 2
else the word is 'القدماء' stego, dj= 1/3(9/9+9/9+9-2/9) = 0.9259

cover without stego

Stego -cover

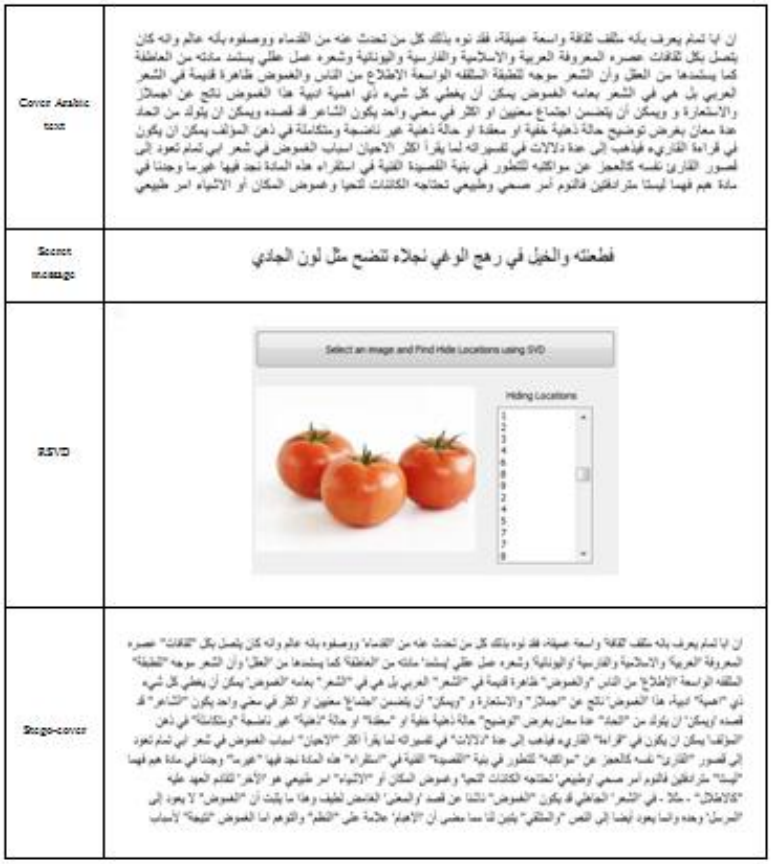
	ا	ل	ق	د	م	ا	ه
'	0	0	0	0	0	0	0
ا	1	0	0	0	0	0	0
ل	0	1	0	0	0	0	0
ق	0	0	1	0	0	0	0
د	0	0	0	1	0	0	0
م	0	0	0	0	1	0	0
ا	0	0	0	0	0	1	0
ه	0	0	0	0	0	0	1
'	0	0	0	0	0	0	0

Table (1): similarity between cover and stego cover.

No of cover	Secret messge size (Byte)	Secret messge size (KB)	Carrier file size (Byte)	Carrier file size (KB)	Average of hide capacity ratio %
1	20480	20	32768	32	0.625 B or KB
2	20480	20	34816	34	0.588 B or KB

Table (2): Explain hide capacity ratio in proposal algorithm.

Figure(9): Proposed technique example one of embeddin.



Figure(10): Proposed technique example two of embedding.

We can conclude from this example that it is not visually easy to find the locations of secret message that is embed in stego-cover.

Case two: An example result for applying the proposed technique using embedding single-double quotation the steganography have no change, this state indicates to robustness. As depicted in figure (11).

Figure (11): Proposed technique example robustness in quotation.

Case three: In this proposed technique, when delete all quotation retain hide information in the words, Because hide bit inside words. The attack cannot known on detected a secret message. This technique is given high security.

Case four: the proposed technique is very high transparency, because it is not seen in human vision and not clear for attack. Especialy when the text without quotation only programmer computers. As depicted in figure (12).

Figure (12): Proposed technique example Transparency quotation.

Case five: The capacity of the proposed technique is change during hiding a secret message, because the first state is convert secret message to binary and two state is addition and injection the quotation inside words. The amount of hiding data is increase in cover, because addition and injection in file carrier imply relative increase in stego cover. The equation below is given hidden ratio:
Hidden Ratio = amount of hidden data / carrier file size
For example:
Hide Capacity Ratio = hidden data1 / carrier file1 size
Hide Capacity Ratio1 = 20 KB/32 KB= 0.625 KB
Hide Capacity Ratio = Hidden data 2 / carrier file2 size
Hide Capacity Ratio2 = 20 KB/34 KB = 0.588 KB

Case six: The compare between previous method in Arabic language text the text steganography using letters points and un-points to hide secret message 0 or 1, the text steganography in other previous method using add kashida in letters point message is 0 bit and add kashida in letters un-point, and other previous method using harakaat/Araabs i.e. Fatha, Kasra, and Damma in Arabic and Urdu phrases, and replace in hide secret message Fatha in /, Kasre in \, Damma in ^. These method is easy detected whereas proposed method using add single-double quotation randomly is difficult to detected secret message.

V. Conclusion

In this paper a new layers Arabic language steganography is implemented using the single-double quotation as an embed process, and RSVD as random location generator to embed the Arabic text message in the Arabic text. The present some conclusions bellow:

1. Applying steganography methods to document (text) files as a cover which is written by Arabic language is difficult, due to the visually sensitivity of Arabic letters to any miner change as in single-double quotation.
2. The RSVD is fast search algorithm, which is improved to use as means to allocate randomly positions in the cover media (Arabic texts) to perform the embedding operation.

3. As embedding methods, domain method is harder against attack than other method, so using single-double quotation as embedding method, which improve its security agents attack.
4. Algorithm robustness: The proposed algorithm prohibits any change to carrier (Arabic text) during the transmission process since the hidden secret message does not change the cover (Arabic text) file properties such as, file size, content, and format during the transmission. Also in scanner not change.
5. Algorithm transparency: the proposed algorithm improvers the transparency property by hiding secret message inside the Arabic text using single-double quotation.
6. Algorithm security: the proposed algorithm improvers the security property by hiding secret message inside the Arabic text using single-double quotation and apply rest Arabic text.
7. Algorithm capacity: the proposed algorithm not change carrier file text Arabic before hide a secret message, and change carrier file text Arabic after hide a secret message .

References

[1] H. M. Salman, " A Natural Language Steganography Technique for Text Hiding Using LSB's", Eng.&Tech. Vol.26,No3,2008.

[2] X. Hu, Gang Luo, Yongjing Lu, and Lingyun Xiang, "A Steganography on Synonym Frequency Distribution", Advances in information Sciences and Service Sciences(AISS), Vol.5, no.10, May 2013.

[3] C. Yun Chang, and Stephen Clark, "Adjective Deletion for Linguistic Steganography and secret sharing", Proceedings of Coling 2012: Technical Papers, pages 493–510, Mumbai, December 2012.

[4] M. Agarwal,"Text Steganographic Approches: A Comparison", International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.1, January 2013.

[5] K. Alla, Christu Jayanthi Jubilcc, R. Siva Ram Prasad, and Acharya Nagarjunat University," A Novale Hindi Text Steganography Using Letter Diacritics and Its Compound Words", IJCSNS. International Journal of Computer Science and Network Security, Vol.8, No.12, December 2008.

[6] M. Shirali, M.Hassan Shirali,"Text Steganography in SMS", International Conference on Convergence Information Technology,2007 IEEE, DOI 10. 1109/ICCIT.2007.100.

[7] A. Abdul-Aziz Gutub, and Manal Mohammad Fattani," A Novel Arabic Text Steganography Method Using Letter Points and Extensions", International Journal of Computer, Information, Systems and Control Engineering Vol :1, No:3, 2007.

[8] H. Shirali-Shahreza and Mohammad Shirali-hahreza,"Arabic/Persian Text Steganography Utilizing Similar Letters With Different Codes",The Arabian Journal for Science and Engineering, Volume 35, Number 1B, December 9, 2009.
[9] A. Odeh and Khaled Elleithy," Steganography in Arabic Text Using Full Diacritics Text", 2009.
[10] V. Kolev, Katya Tsvetkova, and Milcho Tsvetkov," Singular Value Decomposition of Images From Scanned Photographic Plates", Proceedings of the VII Bulgarian-Serbian Astronomical Conference (VII BSAC) Chepelare, Bulgaria, June 1-4, 2010, and Publ. Astron. Soc. “Rudjer Bošković” No 11, 2012.
[11] K. Mounika, D. Sri Navya Lakshmi, K. Alekya," SVD Based Image Compression", International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015, ISSN 2091-2730.