Conversion of Bitmap Text Images for Data Hiding

Ali Mir Arif Mir Asif[1], Shaikh Abdul Hannan[1], R. R. Manza[2], R. J. Ramteke[3]
[1]*Lecturer, I.M.S.I.T., Aurangabad, India* (mirarifali@rediffmail.com)
[1]*Lecturer, Vivekanand College, Aurangabad, India* (hannan_7us@yahoo.com)
[2]*Lecturer, Vivekanand College, Aurangabad, India* (hannan_7us@yahoo.com)
[3]*Department of Computer Science and IT, NMU, Jalgaon* (rakeshramteke@gmail.com)

***Abstract*- In this paper, we propose the comparison of original text in the form of bitmap image and extracted image by using various fonts of text as a bitmap image. The purpose of data hiding is to embed relating textural description into an image. The textural description and original host image can be extracted and reconstructed from the stego-image in the data extraction process. However, the reconstructed host image will more or less be damaged or distorted by using traditional data hiding methods. The advantage of applying LSB data hiding is its simplicity in implementation and its robustness. However, it is not suitable for the data hiding of binary images because the original host image cannot be completely reconstructed due to the problem of information loss. Here we designed a program in Matlab using image processing toolbox which can encode the text images in a single color image. We can recover the image very easily.**

***Keywords*- Bitmap Image, Open Space, Data Hiding, Least Significant Bit.**

I. INTRODUCTION

Digital representation of media facilitates access and potentially improves the portability, efficiency, and accuracy of the information presented. Undesirable effects of facile data access include an increased opportunity for violation of copyright and tampering with or modification of content. The motivation for this work includes the provision of protection of intellectual property rights, an indication of content manipulation, and a means of annotation.

Data hiding represents a class of processes used to embed data, such as copyright information, into various forms of media such as image, audio, or text. Soft-copy text is in many ways the most difficult place to hide data. But the hard-copy text can be treated as a highly structured image and is readily amenable to a variety of techniques such as slight variations in letter forms, kerning, baseline, etc. This is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound bite. While it is often possible to make imperceptible modifications to a picture, even an extra letter or period in text may be noticed by a casual reader. Data hiding in text is an exercise in the discovery of modifications that are not noticed by readers. The three major methods of encoding data are:

1) Open space methods i.e. encode through manipulation of white space (unused space on the printed page)
2) Syntactic methods i.e. utilize punctuation and
3) Semantic methods i.e. encode using manipulation of the words themselves [1], [4], [7].

A. Open Space

There are two reasons why the manipulation of white space in particular yields useful results. First, changing the number of trailing spaces has little chance of changing the meaning of a phrase or sentence. Second, a casual reader is unlikely to take notice of slight modifications to white space. The three methods of using white space to encode data that exploit inter-sentence spacing, end-of-line spaces and inter-word spacing in justified text.

The first method encodes a binary message into a text by placing either one or two spaces after each terminating character, e.g., a period for English prose, a semicolon for C-code, etc. A single space encodes a "0", while two spaces encode a "1". An inter-sentence spacing method has a number of inherent problems. It is inefficient, requiring a great deal of text to encode a very few bits. (One bit per sentence equates to a data rate of approximately one bit per 160 bytes assuming sentences are on average two 80-character lines of text). Its ability to encode depends on the structure of the text. But some text, such as free-verse poetry, lacks consistent or well-defined termination characters. Many word processors automatically set the number of spaces after periods to one or two characters. Finally, inconsistent use of white space is not transparent [2], [8], [11].

A second method of exploiting white space to encode data is to insert spaces at the end of lines. The data are encoded allowing for a predetermined number of spaces at the end of each line. Two spaces encode one bit per line, four encode two, eight encode three, etc., dramatically increasing the amount of information we can encode over the previous method. In Figure 1, the text has been selectively justified, and has then had spaces added to the end of lines to encode more data. Rules have been added to reveal the white space at the end of lines. Additional advantages of this method are that it can be done with any text, and it will go unnoticed by readers, since this additional white space is peripheral to the text. As with the previous method, some programs, e.g., "sendmail," may inadvertently remove the extra space characters. A problem unique to this method is that the hidden data cannot be retrieved from hard copy [3], [5], [6], [9], [13].

A third method of using white space to encode data involves right-justification of text. Data are encoded by controlling where the extra spaces are placed. One space

between words is interpreted as a "0." Two spaces are interpreted as a "1." This method results in several bits encoded on each line. Because of constraints upon justification, not every inter-word space can be used as data. In order to determine which of the inter-word spaces represent hidden data bits and which are parts of the original text, are employed a Manchester-like encoding method. Manchester encoding groups bits in sets of two, interpreting "01" as a "1" and "10" as a "0." The bit strings "00" and "11" are null. For example, the encoded message "1000101101" is reduced to "001," while "110011" is a null string [10], [12].

Open space methods are useful as long as the text remains in an ASCII (American Standard Code for International Interchange) format. As mentioned above, some data may be lost when the text is printed. Printed documents present opportunities for data hiding far beyond the capability of an ASCII text file. Data hiding in hard copy is accomplished by making slight variations in word and letter spacing, changes to the baseline position of letters or punctuation, changes to the letter forms themselves, etc. Also, image data-hiding techniques such as those used by Patchwork can be modified to work with printed text.

B. Syntactic

That white space is considered arbitrary is both its strength and its weakness where data hiding is concerned. While the reader may not notice its manipulation, a word processor may inadvertently change the number of spaces, destroying the hidden data. Robustness, in light of document reformatting, is one reason to look for other methods of data hiding in text. In addition, the use of syntactic and semantic methods generally does not interfere with the open space methods. These methods can be applied in parallel [14], [16], [19], [21].

There are many circumstances where punctuation is ambiguous or when mispunctuation has low impact on the meaning of the text. For example, the phrases "bread, butter, and milk" and "bread, butter and milk" are both considered correct usage of commas in a list. We can exploit the fact that the choice of form is arbitrary. Alternation between forms can represent binary data, e.g., anytime the first phrase structure (characterized by a comma appearing before the "and") occurs, a "1" is inferred, and anytime the second phrase structure is found, a "0" is inferred. Other examples include the controlled use of contractions and abbreviations. While written English affords numerous cases for the application of syntactic data hiding, these situations occur infrequently in typical prose. The expected data rate of these methods is on the order of only several bits per kilobyte of text.

Although many of the rules of punctuation are ambiguous or redundant, inconsistent use of punctuation is noticeable to even casual readers. Finally, there are cases where changing the punctuation will impact the clarity, or even meaning, of the text considerably. This method should be used with caution.

Syntactic methods include changing the diction and structure of text without significantly altering meaning or tone. For example, the sentence "Before the night is over, I will have finished" could be stated "I will have finished before the night is over." These methods are more transparent than the punctuation methods, but the opportunity to exploit them is limited [24].

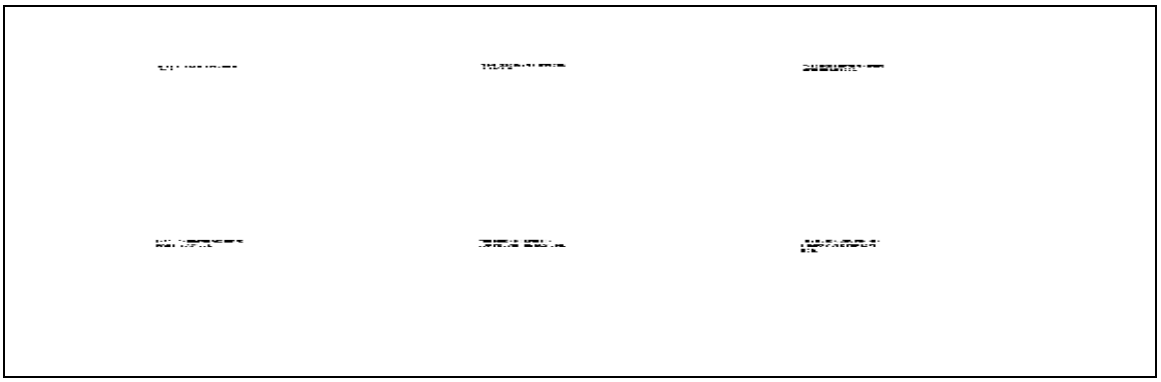
C. Semantic

A final category of data hiding in text involves changing the words themselves. Semantic methods are similar to the syntactic method. Rather than encoding binary data by exploiting ambiguity of form, these methods assign two synonyms primary or secondary value. For example, the word "big" could be considered primary and "large" secondary. Whether a word has primary or secondary value bears no relevance to how often it will be used, but, when decoding, primary words will be read as ones, secondary words as zeros.

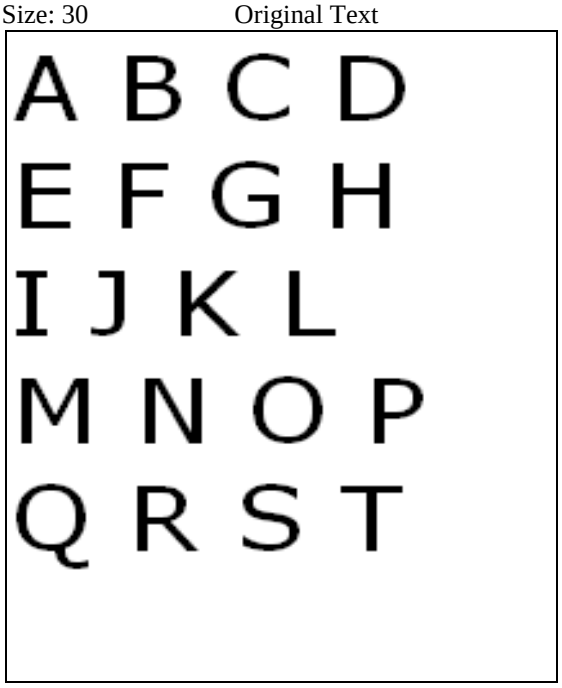
Word webs such as WordNet can be used to automatically generate synonym tables. Where there are many synonyms, more than one bit can be encoded per substitution. (The choice between "propensity," "predilection," "penchant," and "proclivity" represents two bits of data.) Problems occur when the nuances of meaning interfere with the desire to encode data. For example, there is a problem with choice of the synonym pair "cool" and "chilly." Calling someone "cool" has very different connotations than calling them "chilly." The sentence "The students in line for registration are spaced-out" is also ambiguous [15], [17], [23].

II. Experiment Analysis

Text to be encoded



Font : Verdana
Font Size: 30 Original Text



Extracted Text

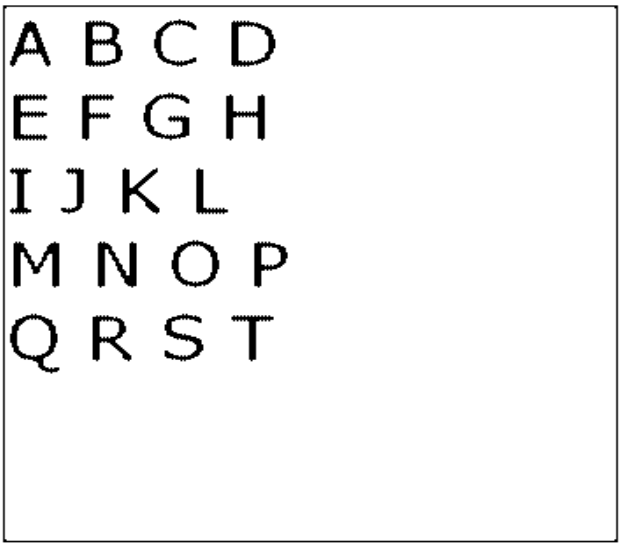


TABLE I
COMPARATIVE RESULTS FOR SINGLE FONT AND DIFFERENT FONT SIZES

Text Font	Font Size	Visible no. of Characters	
		Input	Output
Verdana	20	20	10
	22	20	14
	24	20	15
	26	20	18
	28	20	17
	30	20	18

III. CONCLUSION

In image the basic objective of data hiding is to store as much as data in the host image without degrading the quality of the host image and which will be reconstructed again without compromising the loss of source image data and the actual hided information. There are several well known hiding techniques used by the programmers. The data hiding in still images, in audio, echo and text are used. Out of which the most emerging area is hiding the data into different media files such as image, audio, video, etc. In these media files the image is considered as the most suitable file format for the data processing.

The security has become a very important task in the recent years with the increasing popularity of internet where the people are using internet as a medium to exchange the information at various levels of use. In this scenario there is a need of data hiding with respect to the nature of interest and the complexity of applications. The present paper work is based on the above considerations. The study has done the preparation of the text data set of size 20 characters in single font type, variable font sizes and the color of text as black. The bitmap image can be hided into any color image source which will acts as a medium of carrier of the text data. This color image is further decoded to get the actual data of 20 characters without any loss if possible.

The experimental steps used for text data hiding in images is done with above data set and image processing functions of MATLAB. The data set of a single color source image and the data set of 2 to 3 sentence each of 24 characters with above specification. The experiments resulted into 50% to 75% of reconstruction rate of actual hided text data for the font size of 20 to 24 with VERDANA font. The result is found to be most satisfactory and prominent in the font VERDANA in the font size of 26 to 30 resulted into 85% to 90% of reconstruction rate of actual hided text data, the comparative study is as shown in Table I. The current system can be used with appropriate coding techniques. Another area where more investigation is needed is in modeling of the print-scan channel with respect to the character edges.

REFERENCES

[1] A. V. Drake, Fundamentals of Applied Probability, McGraw-Hill, Inc., New York (1967).

[2] A. V. Oppenheim and R. W. Shaffer, Discrete-Time Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, NJ (1989).

[3] Abdulaziz, N.; Pang, K.K., "Coding techniques for data hiding in images", Symposium on, Signal Processing and its Applications, Sixth International, Volume 1, 13-16 Aug. 2001 Page(s):170 – 173.

[4] Adsumilli, C.B.; Farias, M.C.Q.; Mitra, S.K.; Carli, M., "A robust error concealment technique using data hiding for image and video transmission over lossy channels**"**, IEEE Transactions on, Circuits and Systems for Video Technology, Volume 15, Issue 11, Page(s):1394 - 1406 2005.

[5] B. P_tzmann, \Information hiding terminology." In Anderson [13], pp. 347{350, ISBN 3-540-61996-8, results of an informal plenary meeting and additional proposals.

[6] Beau Grantham, "Bitmap Steganography: An Introduction" COT 4810: Topics in Computer Science Dr. Dutton.

[7] Charkari, N.M.; Chahooki, M.A.Z., "A Robust High Capacity Watermarking Based on DCT and Spread Spectrum", International Symposium on, Signal Processing and Information Technology, 15-18 Dec. 2007 Page(s):194 – 197.

[8] D. L. Hecht, "Embedded Data Glyph, Technology for Hardcopy Digital, Documents," SPIE 2171 (1995).

[9] Digimarc Corporation, Identification/ Authentication Coding Method and Apparatus, U.S. Patent (1995).

[10] E. Adelson, Digital Signal Encoding and Decoding Apparatus, U.S. Patent No. 4,939,515 (1990).

[11] Fakhr, M.W., **"**A Novel Data Hiding Technique for Speech Signals with High Robustness", International Symposium on, Signal Processing & IT, 15-18 Dec. 2007 Page(s):379 – 384.

[12] Hesse, D.; Dittmann, J.; Lang, A., "Network based intrusion detection to detect steganographic communication channels - on the example of images", Euromicro Conference, 2004 Page(s):453 – 456.

[13] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," NECI Technical Report 95-10, NEC Research Institute, Princeton, NJ (1995).

[14] Jun Cheng; Kot, A.C.; Rahardja, S., **"**Steganalysis of Binary Cartoon Image using Distortion Measure", IEEE International Conference on, Acoustics, Speech and Signal Processing, Volume 2, 15-20 April 2007 Page(s):II-261 - II-264.

[15] L. R. Rabiner and R. W. Schaffer, Digital Processing of Speech Signal, Prentice-Hall, Inc., Englewood Cliffs, NJ (1975).

[16] Moon, S.K.; Kawitkar, R.S., "Data Security Using Data Hiding", International Conference on, Conference on Computational Intelligence and Multimedia Applications, Volume 4, 13-15 Dec. 2007 Page(s):247 – 251.

[17] Pingault, M.; Pellerin, D., "Motion transparency constraint equation based on a wavelet function decomposition", IEEE International Conference on, Multimedia and Expo, Volume 1, 26-29 Aug. 2002 Page(s):717 – 720.

[18] Placeway, P.; Schwartz, R.; Fung, P.; Nguyen, L., **"**The estimation of powerful language models from small and large corporation", International Conference on, Acoustics, Speech, and Signal Processing, Volume 2, 27April 1993 Page(s):33– 36.

[19] R. C. Dixon, Spread Spectrum Systems, John Wiley & Sons, Inc., New York (1976).

[20] R. J. Anderson, ed., Information hiding: First international workshop, vol. 1174 of Lecture Notes in Computer Science, Isaac Newton Institute, Cambridge, England, May 1996, Springer-Verlag, Berlin, Germany, ISBN 3-540-61996-8.

[21] S. K. Marvin, Spread Spectrum Handbook, McGraw-Hill, Inc., New York (1985).

[22] Sakaguchi, S.; Arai, T.; Murahara, Y., "The effect of polarity inversion of speech on human perception and data hiding as an application", International Conference on, Acoustics, Speech, and Signal Processing, Volume 2, 5-9 June 2000 Page(s):II917 - II920.

[23] Sang-Bong Lee; Tae-Jung Kim; Jae-Won Suh; Hyeon-Deok Bae, **"**Error Concealment for 3G-324M Mobile Videophones Over a WCDMA networks", International Conference on, Consumer Electronics, 10-14 Jan. 2007 Page(s):1 – 2.

[24] Solanki, K.; Dabeer, O.; Manjunath, B.S.; Madhow, U.; Chandrasekaran, S., "Joint source-channel coding scheme for image-in-image data hiding", International Conference on, Image Processing, Volume 2, 14-17 Sept. 2003 Page(s):II - 743-6.

[25] Srivastava, C, 2000, Fundamentals of Information Technology, New Delhi, Kalyani Publishers.

[26] Suhail, M.A.; Obaidat, M.S., "A watermarking technique for geometric manipulation attacks", 12th IEEE International Conference on, Electronics, Circuits and Systems, 11-14 Dec. 2005 Page(s):1 – 5.

[27] R.Z. Wang, C.F. Lin, J.C. Lin, Image hiding by optimal LSB substitution and genetic algorithm, Pattern Recognition 34 (3) (2001) 671–683.

[28] J. Brassil, S. Low, N. Maxemchuk, L. O'Gorman, Electronic marking and identification techniques to discourage document copying, IEEE Journal on Selected Areas in Communications 13 (1995) 1495–1504.

[29] S.H. Low, N.F. Maxemchuk, J.T. Brassil, L. O'Gorman, Document marking an identification using both line and word shifting, in: Proceedings of Infocom, Boston, MA, 1995, pp. 853–860.

[30] H. Lu, A.C. Kot, J. Cheng, Secure Data Hiding in Binary Document Images for Authentication, IEEE, New York, 2003.