# A REVIEW ON TEXT STEGANOGRAPHY

Amanpreet Kaur[1], Sukhvir Kaur[2], Gunjan Sethi[3]
[1]*Research Scholar, Department of Computer Engineering, CTIEMT, Jalandhar*
*(E-mail: amanpreet.kamboj231@gmail.com)*
[2]*Assistant Professor, Department of Computer Engineering, CTIEMT, Jalandhar*
*(E-mail: sukhsain.17@gmail.com)*
[3]*Assistant Professor, Department of Computer Engineering, CTIEMT, Jalandhar*
*(E-mail: gunjan.ctit@gmail.com)*

*Abstract*— Steganography relies on hiding message in unsuspected multimedia data and is generally used in secret communication between different parties. Text steganography uses text as a cover media for hiding message. Message can be hidden by shifting word and line, properties of a sentence such as number of words, number of characters, number of vowels, position of a vowel in a word are also used to hide secret message. The advantage of preferring text steganography over other steganographic technique is its smaller memory requirement and simpler communication medium. Steganography is the art and science of secret communication. It is the practice of encoding/embedding secret information in a manner such that the existence of the information is invisible. In this paper we have discussed the types, model of text steganography in detail.

***Keywords***— *Steganography, cryptography, plain text, encryption, decryption, cipher.*

## I. INTRODUCTION

In the current world we cannot imagine our lives without computers. However with the use of computers a question of secure data transfer appears rather soon. Information coding and cryptography is essential, but efficient privacy has been given by encryption and information hiding methods that can be misused for covering criminal activities. Therefore is important to develop tools and methods for forensic analysis. Steganography [1] and cryptography [2] are normally connected together. Cryptography is effective in the usage of the key and the message is somehow coded. If it is sent insecurely, an attacker will notice it immediately and will try to decode it. However there is a steganography, which helps with the secure transfer of encoded messages. It codes a message inside of a picture or another multimedia file. If you see a stenographic picture, you will not recognize the secret message inside of picture. And this is the point. Crackers will go through and will not pay attention to the message. Therefore it is necessary to have a method for its detection. To decode a message itself is another challenge, this thesis is aimed to reveal a secret message inside the picture. Steganography, as art of hiding information, has been known for over 2500 years. Back then steganography was mainly used for diplomatic, military and a very few people used it for personal purposes along with cryptography. Steganography as well as cryptography have a goal to secure transmitted information between the sender and the recipient, but both systems are used in a different way. Cryptography is aimed on transformation of input data into unreadable output. Level of information security depends on the quality of cryptographic algorithm and correct cipher key selection. Steganography has a different approach, stego messages also referred as steganograms are made in such a way that they do not attract attention to themselves. Even transfer remains undetected if steganography is used correctly. No matter how strong cipher can be used, there is always an attempt to wire tape the crypted message and try to break cipher or recover cipher key. However if it is not possible to determine message itself there is nothing to do. The very best solution for securing messages and transport medium is to use cryptography for transforming message into unintelligible gibberish, referred as cipher text, and steganography to cover a whole message along with transport medium. First documented steganography application was around 440 BC where Demaratus sent a warning about a forthcoming attack to Greece on a wax tablet. The message in that case was written on a wooden backing and then covered by beeswax. It appeared as unused. Second one was from that time too. But this time a different transfer medium was chosen. The steganogram was made as a message tattooed on slave's clean shaven head. Then they waited for hair to grow back and then send the slave to deliver the message. Steganography is the art of hiding information imperceptibly in a cover medium. The word "Steganography" is of Greek origin and means "covered or hidden writing". The main aim in steganography is to hide the very existence of the message in the cover medium. Steganography includes a vast array of methods of secret communication that conceal the very existence of hidden information. A message in cipher text might arouse suspicion on the part of the recipient while an "invisible" message created with stenographic methods will not. Anyone engaging in secret communication can always apply a cryptographic algorithm to the data before embedding it to achieve additional security. In any case, once the presence of hidden information is revealed or even suspected, the purpose of steganography is defeated, even if the message content is not extracted or deciphered. According to [1], "Steganography's niche in security is to supplement cryptography, not replace it. If a hidden message is encrypted, it must also be decrypted if discovered, which provides another layer of protection." Steganography became very popular during Second World War where there was a limited amount of usable communication

routes for resistance in Europe that made a perfect environment for developing methods of secret communication. Crucial was simplicity information exchange and high level of security. Messages were delivered through radio broadcast coded into birthday wishes, name day wishes or in advertisement that was the reason why Nazist baned Low frequency radio receivers under the death sentence. Among other steganography techniques used in Second World War were various kinds of invisible ink or microdots. Microdots are a text or an image substantially reduced in size onto a 1mm disc to prevent detection by unintended recipients. Microdots placed into regular text message would microdot provide a very good transportation medium with perfect protection of transferred secret message. All mentioned methods belong into so called mechanical steganography that can still be used nowadays but as technical development turned 21st century within computer revolution new methods of information transfers have become common for everyone. Computer data offers undepletable options for digital steganography. As mentioned above, steganography has many forms and can be divided into groups by used cover medium and embedding system for secret message.

In steganography the object to be transmitted is the embedded message, and the cover signal serves as an innocuous disguise chosen fairly arbitrarily by the user based on its technical suitability. In addition, the existence of the watermark is often declared openly, and any attempt to remove or invalidate the embedded content renders the host useless. The crucial requirement for steganography is perpetual and algorithmic undetectability. Robustness against malicious attack and signal processing is not the primary concern, as it is for watermarking. As mentioned, steganography deals with hiding of information in some cover source. On the other hand, Steganalysis is the art and science of detecting messages hidden using steganography; this is analogous to cryptanalysis applied to cryptography. The goal of steganalysis is to identify suspected packages, determine whether or not they have a payload encoded into them, and, if possible, recover that payload. Hence, the major challenges of effective steganography are:-

1. *Security of Hidden Communication*: In order to avoid raising the suspicions of eavesdroppers, while evading the meticulous screening of algorithmic detection, the hidden contents must be invisible both perceptually and statistically.

2. *Size of Payload*: Unlike watermarking, which needs to embed only a small amount of copyright information, steganography aims at hidden communication and therefore usually requires sufficient embedding capacity. Requirements for higher payload and secure communication are often contradictory. Depending on the specific application scenarios, a tradeoff has to be sought.

## II. TYPES OF STEGANOGRAPHY

Steganography has many forms and can be divided into groups by used cover medium and embedding system for secret message. Steganography examples divided by cover medium are as follows:

a) PHYSICAL STEGANOGRAPHY: In this steganography the secret message is physically encoded into the carrying medium.

• *Hiding one thing inside of another*

Basic techniques in physical steganography are e.g. safe place in walking stick, double bottom of carry-on bag or suitcase.

• *Microdots*

As mentioned above, microdots are a method used for reducing information into 1mm disk similar to period produced by typewriter.

• *Yellow dots*

Yellow dots are produced by color laser printers. Every printed paper is marked by almost invisible code of yellow dots representing printer name, date and time stamp.

• *Code recognition and automatic code extraction from Fujitsu*

Barcode is embedded into printed image and remain readable by portable devices [1], steganogram is combination of human understandable information represented by pictogram and embedded computer readable data (QR code). Information extraction is possible trough cell phone with camera or similar handled device.

• *The letter size, spacing, typeface*

The main idea of typography modification is to cover embedding methodology look like a typing error or unusual document layout. In fact, it is a very good method of message transportation trough public media such as newspapers, magazines, no matter if the text is printed or electronical [2].

b) DIGITAL STEGANOGRAPHY: Digital Steganography is the science that involves communicating secret data in an appropriate multimedia carrier, e.g., image, audio, and video files.

• *Text*

The method is used to conceal messages in ASCII text by appending whitespace to the end of lines. Spaces and tabs are generally not visible in text editors by default. The message is effectively hidden from casual observers. Text steganography has more techniques of hiding information, for example: Open space methods, Word shifting coding, Line shifting coding, Syntactic methods, Semantic methods, Feature coding [3]. To utilize text steganography has a several reasons. Firstly, a secret message coded into an internet article or email message will not attract any attention when transferred. Another reason may be afford to stylize cover text into specific form e.g. SPAM message. Propagation steganography can be used to generate an artificial message with content similar to SPAM message Existence of such message would not attract any attention due to common occurrence. Symantec in their MessageLabs Intelligence: 2010 Annual Security Report [4] announced that more than 89 percent of all email traffic worldwide is SPAM. SPAM message can be artificially made by mimic algorithm or mimicry which is an example of propagation steganography and can be used to make artificial

texts with look of average internet article or advertisement. Mimic texts are not linguistically correct, but statistically they are good enough to fool spam filters. Mimic is not capable of fooling a human, at least in most cases. If a mimic text is investigated it will strongly support the idea of SPAM message.

- *Image*

Graphic files are the most common data files on the Internet after text information files. Computer graphics is the ideal cover medium for covert communication because of the limitation of the human eye as well as the limited representation of digital technology. Usually the still image represented by a BMP image format has a large size, which is inefficient for transmitting over the Internet. This was the reason for a new type of stego tools for compressed graphic files such as GIF, PNG and JPEG. JPEG steganography is more complicated than usual still image steganography because raw image data is not directly accessible as for example on 24 bit RGB BMP model. JPEG steganography is generally based on Least Significant Bit (LSB) applied during the discrete cosine transformation (DCT), information are hidden in the frequency domain. The DCT algorithm is one of the most important components of the JPEG compression.

- *Audio*

Encoding replaces the least significant bit of information in each sampling point with a coded binary string. While this method can be efficiently employed to encode fairly large amounts of hidden data in a given audio signal, it does so at the expense of introducing significant noise at theoretical upper limits [8]. Phase coding works by substituting the phase of an initial audio segment with a reference phase that represents the data. The phase of subsequent segments is adjusted in order to preserve the relative phase between segments [9]. While phase coding is discrete in comparison to low-bit encoding, it is also a more complicated method [8].

- *Video*

Most of the previously mentioned methods are suitable for various kinds of cover mediums; video files and streams are not exception. Video files are nothing else than large number of images and sound data. Video steganography has an advantage over the static image because of the continuous stream of information and large space for hidden information.

c) TEXT STEGANOGRAPHY: Steganography can be classified into image, text, audio and video steganography depending on the cover media used to embed secret data. Text steganography can involve anything from changing the formatting of an existing text, to changing words within a text, to generating random character sequences or using context-free grammars to generate readable texts [7]. Text steganography is believed to be the trickiest due to deficiency of redundant information which is present in image, audio or a video file. Hiding information in text file is the most common method of steganography. The method was to hide a secret message into a text message. After coming of Internet and different type of digital file formats it has decreased in importance. Text stenography using digital files is not used very often because the text files have a very small amount of excess data.

The structure of text documents is identical with what we observe, while in other types of documents such as in picture, the structure of document is different from what we observe. Therefore, in such documents, we can hide information by introducing changes in the structure of the document without making a notable change in the concerned output [8]. Unperceivable changes can be made to an image or an audio file, but, in text files, even an additional letter or punctuation can be marked by a casual reader [9]. Storing text file require less memory and its faster as well as easier communication makes it preferable to other types of stenographic methods [10]. Text steganography can be broadly classified into three types: Format based, Random and Statistical generation, Linguistic methods.

- *Format Based Methods*: Format based methods involve altering physically the format of text to conceal the information. This method has certain flaws. If the stego file is opened with a word processor, misspellings and extra white spaces will get detected. Changed fonts sizes can arouse suspicion to a human reader. Additionally, if the original plaintext is available, comparing this plaintext with the suspected stenographic text would make manipulated parts of the text quite visible [7].

- *Random and Statistical Generation*: In order to avoid comparison with a known plaintext, stenographers often resort to generating their own cover texts [7]. One method is concealing information in random looking sequence of characters. In another method, the statistical properties of word length and letter frequencies are used in order to create words which will appear to have same statistical properties as actual words in the given language [2, 3].

- *Linguistic Steganography*: Linguistic steganography specifically considers the linguistic properties of generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden [7]. CFG create tree structure which can be used for concealing the bits where left branch represents '0' and right branch corresponds to '1'. A grammar in GNF can also be used where the first choice in a production represents bit 0 and the second choice represents bit 1. This method has some drawbacks. First, a small grammar will lead to lot of text repetition. Secondly, although the text is syntactically flawless, but there is a lack of semantic structure. The result is a string of sentences which have no relation to one another [7].

## III. EXISTING APPROACHES OF TEXT STEGANOGRPAHY

In this sub-section, we present some of the popular approaches of text steganography.

- *Line Shift*

In this method, secret message is hidden by vertically shifting the text lines to some degree [10, 11]. A line marked has two unmarked control lines one on either side of it for detecting the direction of movement of the marked line [12]. To hide bit 0, a line is shifted up and to hide bit 1, the line is shifted down [13]. Determination of whether the line has been shifted up or down

is done by measuring the distance of the centroid of marked line and its control lines [12]. If the text is retyped or if a character recognition program (OCR) is used, the hidden information would get destroyed. Also, the distances can be observed by using special instruments of distance assessment [10].

- *Word Shift*

In this method, secret message is hidden by shifting the words horizontally, i.e. left or right to represent bit 0 or 1 respectively [13]. Words shift are detected using correlation method that treats a profile as a waveform and decides whether it originated from a waveform whose middle block has been shifted left or right [12]. This method can be identified less, because change of distance between words to fill a line is quite common [10, 11]. But if someone knows the algorithm of distances, he can compare the stego text with the algorithm and obtain the hidden content by using the difference. Also, retyping or using OCR programs destroys the hidden information [10, 11].

- *Syntactic Method*

This technique uses punctuation marks such as full stop (.), comma (,), etc. to hide bits 0 and 1. But problem with this method is that it requires identification of correct places to insert punctuation marks [10, 11]. Therefore, care should be taken in using this method as readers can notice improper use of the punctuations [9].

- *White Space* This technique uses white spaces for hiding a secret message. There are three methods of hiding data using white spaces. In Inter Sentence Spacing, we place single space to hide bit 0 and two spaces to hide bit 1 at the end of each terminating character [9]. In End of Line Spaces, fixed number of spaces is inserted at the end of each line. For example, two spaces to encode one bit per line, four spaces to encode two bits and so on. In Inter Word Spacing technique, one space after a word represents bit 0 and two spaces after a word represents bit 1. But, inconsistent use of white space is not transparent [9].

- *SMS-Texting*

SMS-Texting language is a combination of abbreviated words used in SMS [8]. We can hide binary data by using full form of word or its abbreviated form. A codebook is made which contains words and their corresponding abbreviated forms. To hide bit 0, full form of the word is used and to hide bit 1, abbreviated form of word is used [8].

- *Feature Coding*

In feature coding, secret message is hidden by altering one or more features of the text. A parser examines a document and picks out all the features that it can use to hide the information [13]. For example, point in letters i and j can be displaced, length of strike in letters f and t can be changed, or by extending or shortening height of letters b, d, h, etc. [6, 14]. A flaw of this method is that if an OCR program is used or if re-typing is done, the hidden content would get destroyed.

- *SSCE (Secret Stenographic Code for Embedding)*

This technique first encrypts a message using SSCE table and then embeds the cipher text in a cover file by inserting articles or with the nonspecific nouns in English language using a certain mapping technique [15]. The embedding positions are encrypted using the same SSCE table and saved in another file which is transmitted to the receiver securely along with the stego file.

- *Word Mapping*

This technique encrypts a secret message using genetic operator crossover and then embeds the resulting cipher text, taking two bits at a time, in a cover file by inserting blank spaces between words of even or odd length using a certain mapping technique [6]. The embedding positions are saved in another file and transmitted to the receiver along with the stego object.

- *MS Word Document*

In this technique, text segments in a document are degenerated, mimicking to be the work of an author with inferior writing skills, with secret message being embedded in the choice of degenerations which are then revised with changes being tracked [7]. Data embedding is disguised such that the stego document appears to be the product of collaborative writing [7].

- *Cricket Match Scorecard*

In this method, data is hidden in a cricket match scorecard by pre-appending a meaningless zero before a number to represent bit 1 and leaving the number as it is to represent bit 0 [8].

- *CSS (Cascading Style Sheet)* This technique encrypts a message using RSA public key cryptosystem and cipher text is then embedded in a Cascading Style Sheet (CSS) by using End of Line on each CSS style properties, exactly after a semicolon. A space after a semicolon embeds bit 0 and a tab after a semicolon embeds bit 1 [9].

## IV. APPLICATIONS OF TEXT STEGANOGRAPHY

- *Secret Communications:* [13] the use of text steganography does not advertise secret communication and therefore avoids scrutiny of the sender, message, and recipient. A trade secret, blueprint, or other sensitive information can be transmitted without alerting potential attackers.

- *Feature Tagging:* Elements can be embedded inside the cover medium, such as the names of individuals or email id's. Copying the stego-cover also copies all of the embedded features and only parties who possess the decoding stego-key will be able to extract and view the features.

- *Copyright Protection*: Copy protection mechanisms that prevent data, usually digital data, from being copied. The insertion and analysis of watermarks to protect copyrighted material is responsible for the recent rise of interest in digital steganography and data embedding. [6, 7]

## V. CONCLUSION

In this paper we reviewed many papers on Steganography and various techniques based on it .By reviewing various papers it is concluded that steganography is means passing of secret

information within an innocent cover and send it to proper recipient who is aware of decoding process. Steganography is the art and science of covered writing. Steganography is the art of hiding of a message within another so that presence of hidden message is indistinguishable. The key concept behind steganography is that message to be transmitted is not detectable to the casual eye. This is also the advantage of steganography over cryptography. Modern digital steganography uses text, images, audio, video etc. as a cover medium. This paper presents a survey on a data hiding technique called 'Steganography', the terminology, the model, its types. This is followed by a discussion on various text-based data-hiding techniques where the primary focus remained on recently proposed/developed stenographic techniques. In physical steganography microdots and yellow dots are used for hiding the secret message which is very much complicated and time consuming. However image or video steganography also consumes lot of time when sending the large files over network. A lot of time is consumed while processing the large audio and video files. Finally we narrow down our search to the text based steganography which is efficient for storing large secret data and can be easily transferred over a network.

## VI. REFERENCES

[1] H.Kabetta, B.Y. Dwiandiyanta, Suyoto, "Information hiding in CSS: A secure scheme text-steganography using public key Cryptosystem", IJCIS, pp. 13-22, Vol.1, No.1, December 2011.

[2] Mohit Garg," A Novel Text Steganography Technique Based on Html Documents", International Journal of Advanced Science and Technology, pp.132-138, Vol. 35, October, 2011.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", pp. 313-336, IBM Systems Journal, vol. 35, Issues 3&4, 1996.

[4] P.Singh, R.chaudhary and A.Agarwal," A Novel Approach of Text Steganography based on null spaces", IOSRJCE, PP 11-17, Volume 3, Issue 4 (July-Aug. 2012).

[5] M.S. Shahreza," A New Method for Steganography in HTML Files", Computer, Information, and Systems Sciences, and Engineering, Proceedings IETA 2005, TeNe 2005, EIAE 2005, 247-251, Springer.

[6] R. Kumar, A. Malik, "A Space based reversible high capacity text steganography scheme using Font type and style", International Conference on Computing, Communication and Automation, IEEE, 2016.

[7] R. Saniei, K. Faez, "The Capacity of Arithmetic Compression Based Text Steganography Method", Iranian Conference on Machine Vision and Image Processing, 2013

[8] Proceedings of the 3rd National Conference; INDIACom-2009Computing For Nation Development, February 26 – 27, 2009 Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi

[9] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", IBM Systems Journal, vol. 35, Issues 3&4, 1996, pp. 313-336

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

[11] 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), vol.2, 2-6 April 1995, pp. 853 - 860.

[12] A.M. Alattar, and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing", Proceedings of SPIE -- Volume5306, Security, Steganography, and Watermarking of Multimedia Contents VI, June 2004, pp. 685-695.

[13] D. Huang, and H. Yan, "Inter word Distance Changes Represented by Sine Waves for Watermarking Text Images", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 12, December 2001, pp. 12371245.

[14] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", Proceedings of 5th IEEE/ACIS international Conference on Computer and Information Science and 1st IEEE/ACIS, June 2006.

[15] 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] J.A. Memon, K. Khowaja, and H. Kazi, "Evaluation of steganography for Urdu /Arabic text", Journal of Theoretical and Applied Information Technology, pp 232-237