Steganography using web documents as a carrier: A Survey

[1]Chintan Dhanani, [2]Krunal Panchal

[1]Mtech Scholar, [2]Lecturer

Department of computer engineering, Gujarat Technological University, Gujarat, India
[1]chintan00014@gmail.com, [2]krunaljpanchal@gmail.com

***Abstract* - In the world of information technology security of data is most important part. Everywhere there is a problem of security threats which are always looking to steal the information. So any how data protection is required. Cryptography, steganography and watermarking are some of the well known data protection techniques. Steganography and cryptography are data hiding techniques while watermarking is used to give unique identity to the objects like image, audio, video etc which prevents it from forgery. The benefit of steganography over cryptography is that no one except sender and intended user can see the message. This survey paper concentrates on the steganography techniques and mainly on the techniques that have used web document as a carrier to hide the data. The HTML steganography has its own benefit that data doesn't look suspicious because HTML web pages are fundamental elements of the modern internet technology and are very rapidly used in websites.**

***Keywords* – Steganography; Stego key; Stego data; Carrier; Embedding; Decoding**

I. INTRODUCTION TO STEGANOGRAPHY

'Steganography' is a Greek word which means concealed writing or hidden writing. Steganography is the art and science of encoding hidden messages in such a way that no one except the sender and intended recipient, suspects the existence of the message. Steganography and cryptography both are used to protect data from unauthorized person but there is only one difference between these two data protection techniques. In cryptography the unauthorized user can able to see the encrypted message but may not be able to decrypt it and view the original message while in steganography unauthorized person can't able to see the message because message is hidden in carrier and travel through the carrier. The carrier of the message may be plain text, image, audio, video, flash drive, Huffman tree, Binary file, Histogram, HTML code etc. The main carrier that are been used in current technology are text, audio, video, image. HTML steganography is a one part of text steganography.

Basic model of steganography

Fig-1 gives the idea about steganography scheme in which first step is to embed original message in the carrier using any embedding technique. Then embedded message travel through the transmission media. At the receiver side receiver decodes the message which is the reverse process of embedding and gets the original message.

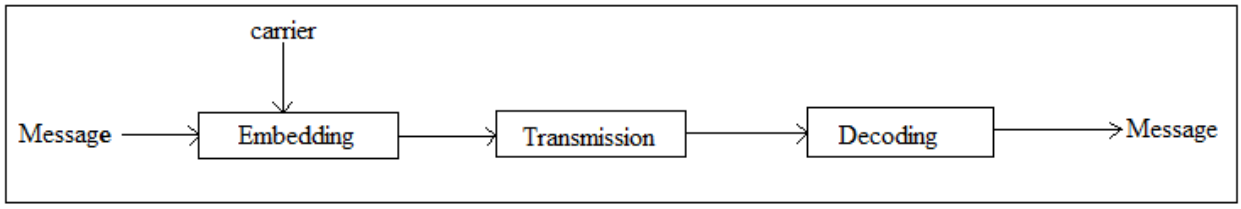

Fig 1 Basic model of steganography

Types of carrier used in steganography

Carrier is one type of transmission medium which hides information in it. Fig-2 gives idea about carrier used in steganography. The types of information carriers are as follows.

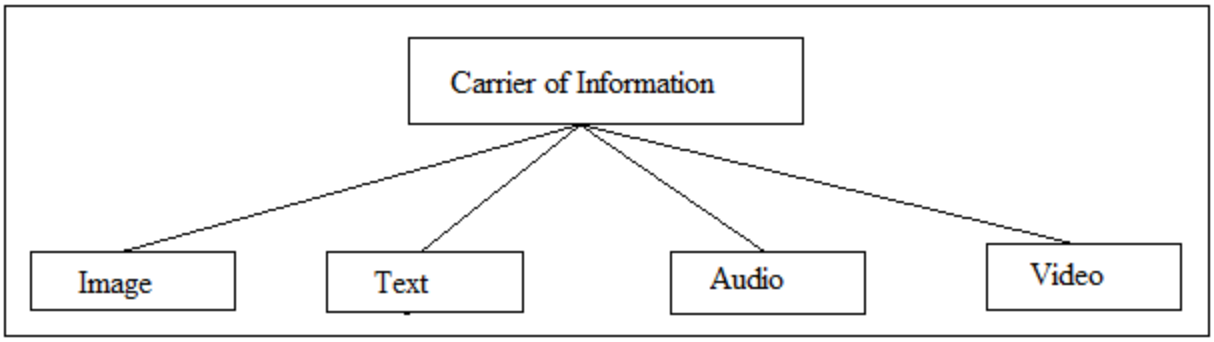

Fig 2 Types of information carrier

1. Image - It is one of the most carriers used in steganography. Human eye has some limitations. Because of these limitations data can be hidden in the image. The original image and the embedded image with data in it both look almost same.
2. Video - Video is nothing more than group of images. In which images flows at some specific speed measured in frames/second so that we can see a constant video. So to hide data in video is same as hide data in the images. It also takes benefit of some limitations of Human Visual System (HVS). In video steganography each frame contains some hidden information
3. Audio - Audio steganography is one of the difficult techniques of steganography because human can easily detect the minute change in the audio. It hides the data in analog or digital signals so audio steganography requires the more knowledge of signal processing too.
4. Text - Text steganography hides text in the text. Text steganography is difficult in the sense that there are very limited places in the text carrier where there is a possibility to hide the text message. HTML and binary file are also the type of text carrier.

II. CLASSIFICATION OF DATA HIDING TECHNIQUES

Fig-3 is about the classification of data hiding techniques. Main three types of data hiding techniques are steganography, covert channel hiding & copy right marking.

A. Steganography

Steganography can be classified based on the carrier used to hide data. Four types of carrier generally used in steganography are audio, video, text and images. We have already learned about these carriers in section 1.2. Text steganography can also further been classified as the carrier used. The carriers may be plain text, webpage text or binary file.

Webpage text steganography

Web page text contains HTML, CSS, XML, JavaScript etc as content. So anyone of these can be used as a carrier when webpage is used to hide data. Webpage text steganography uses tags, attributes of the tags of web document to hide data. Using attribute in different order, use tags in varying combination, hide data in id attribute of tags, whitespaces in the tags are some of the techniques used to hide data in web document.

Plain text steganography

Plain text steganography further classified in Format based (Technical) Steganography and Linguistic Steganography.

1. **Technical or Format based Steganography [5]**
   Format based (Technical) steganography methods generally modify existing text in order to hide the steganography text. White space insertion, line shift, word shift, feature encoding are some of the many format-based methods used in text steganography. Deliberate misspellings distributed throughout the text, and resizing of fonts are types of feature encoding.
2. **Linguistic Steganography [5]**
   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. A semagram is a secret message that is not in a written form. For example, a system can use long blades of grass in a picture as dashes in Morse code, with short blades for dots. Open codes are illusions or code words. In World War 1, for example, German spies used fake orders for cigars to represent various types of British warships – cruisers and destroyers. Thus 500 cigars needed in Portsmouth meant that five cruisers were in Portsmouth. In syntactic method by placing dot (.) and comma (,) in proper places, one can hide information in the text file. Identification of proper hiding places is required to use syntactic methods. Semantic methods hide information by replacing word by its synonyms. The synonym substitution may represent a single or multiple bit combination for secret information.

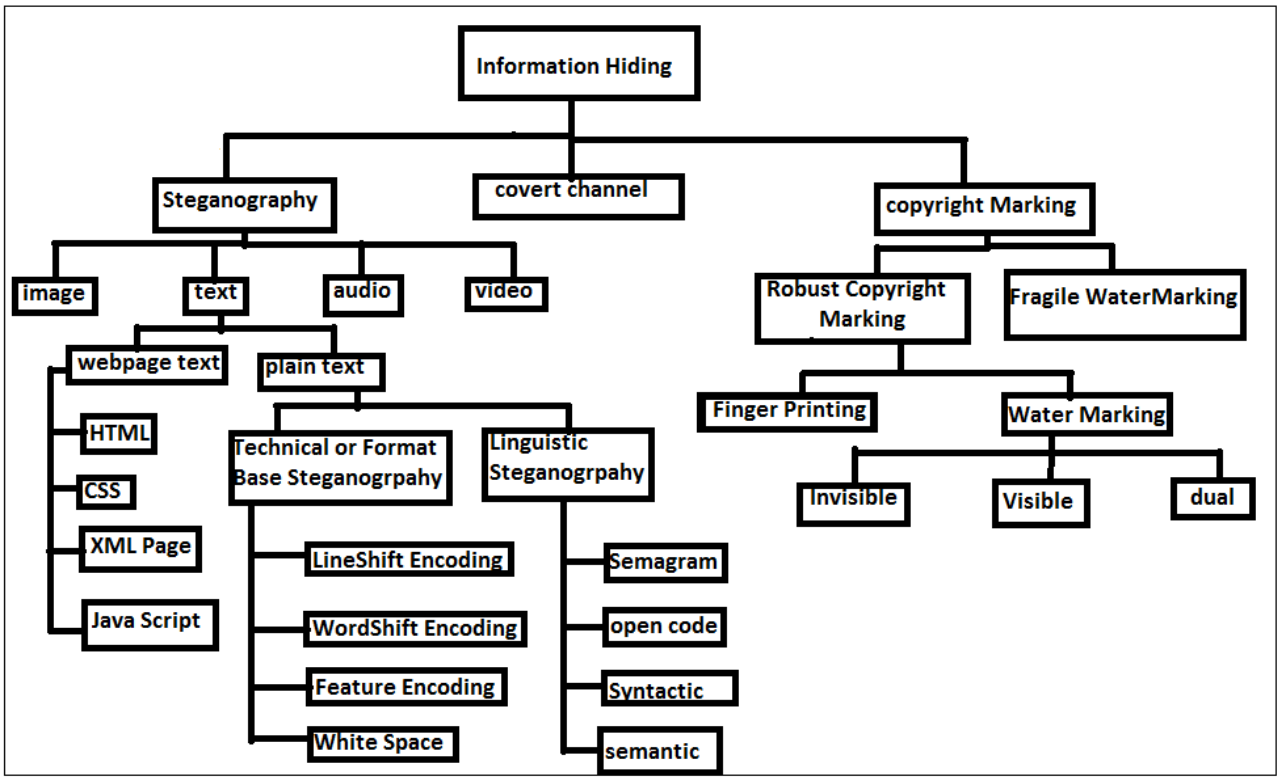

Fig 3 Classification of data hiding techniques [8] [9]

B. Covert channel steganography [9]

A covert channel is a communication channel that transfers some kind of information using a method originally not intended to transfer this kind of information. Observer can't suspect that a covert message is being transferred. Only the sender and receiver of the message notice it.

C. Copyright Marking

Copyright marking is used to give unique identity to object. Two types of copyright marking are robust copyright marking & fragile (semi) water marking.

1. **Fragile or semi water marking**
   A fragile (semi) watermark is a mark which is sensitive to a modification of the carrier. A fragile watermarking software scheme should be able to detect any change in the signal and identify where it has taken place and possibly what the signal was before modification. It serves at proving the authenticity of a document.
2. **Robust copyright marking**
   A robust photo watermark should be stuck to the document it has been embedded in, in such a way that any signal transform of reasonable strength cannot remove the watermark. Hence a pirate willing to remove the watermark will not succeed unless they debase the document too much to be of commercial interest. The latter form is the very challenging and attracts most research. Two types of robust copyright marking are finger printing and watermarking. Fingerprints [9] are characteristics of the object that distinguish it from other similar objects. They enable the owner to trace illegal distribution of that object. Watermarking is used to verify the identity and authenticity of the owner of a digital image. It is a process in which the information which verifies the owner is embedded into the digital image or signal. For example, famous artists watermark their pictures and images. If somebody tries to copy the image, the watermark is copied along with the image. Invisible, visible and dual watermarking are three types of watermarking.

III. STEGANOGRAPHY TECHNIQUES USING HTML AS A CARRIER

HTML steganography is one part of text steganography. In which it uses HTML file, CSS, java script etc. as a carrier of information. Some of the HTML steganography techniques that have already been used are as follows.

1. Two consecutive attribute of same HTML tag are used to hide data [1]. Because sequence of the attributes of the same tag doesn't make difference to output of HTML document.
   Example : <body **background**="image1.jpeg" **bgcolor**="#FFFFFF">……0
   <body **bgcolor**="#FFFFFF" **background**="image1.jpeg" >.....1
   As per above example if the sequence of attributes is (background,bgcolor) then it hides 0 and if sequence of attributes is (bgcolor,background) then it hides 1.

2. Relation between two consecutive attributes is considered for hiding secret data. Two hide '1' two consecutive attributes of same tag are taken and to hide '0' two consecutive attributes of different tags are taken [2].
   Example:
   Step 1: secret message is – 11100
   Step 2: take attributes (no of digits + 1) = 5 + 1 = 6
   bgcolor background alink vlink size src
   So the pair (bgcolor,background) hides 1
   (background,alink) hides 1
   (alink,vlink) hides 1
   (vlink,size) hides 0
   (size,src) hides 0
   Now the pair which hides 1 has both its attributes in same tag & the pair which hides 0 has both its attributes in different tags.
   Step 3 : take a key as 1650(11001110010) because it contains same number of 1's as the number of attributes considered in step 2
   Step 4: Put attributes of step-2 in place of 1 & leave 0 as it is
   bgcolor background 0 0 alink vlink size 0 0 src 0
   Step 5 : put any attribute in place of zero
   bgcolor background **text link** alink vlink size **face color** src **alt**
   Step 6 : Make HTML document

```
<body bgcolor =”#0000” background= “filename” text =“color” link =“filename” alink=“filename” vlink=“filename”>
    <font size=“value” face=“name” color=“color”>
        I am showing Example
    </font>
    <Img src=“url” alt=“text”>
</body>
```

   As per the rule written in step-2 this HTML document hides the message 11100.

3. Hide Coded data in the ID attribute of the HTML document tags [3].

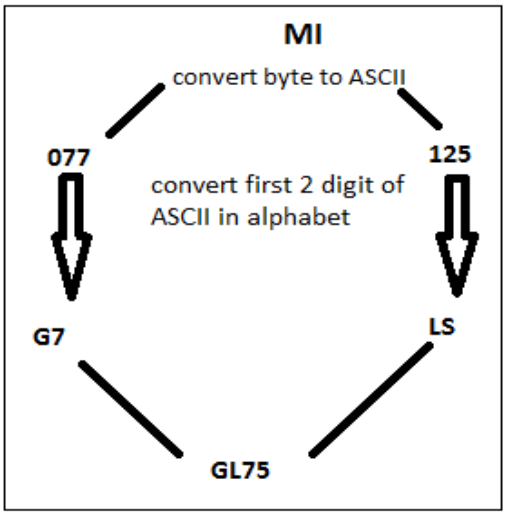

Fig 4 conversion of message in to code

   Above fig-4 gives the idea of conversion of word in to the code. These code looks like the id attributes of the HTML tags. So it can be placed as id to hide in HTML document.

4. HTML tags can be used in varying combinations to represent pattern of secret information bits[4].
   Example :
   Stego key : <img></img>.....0
   <img/>……1
   Stego data :

```
<img src=image1.jpg></img>
<img src=image2.jpg/>
<img src=image3.jpg/>
```

   Hidden bits: 100

5. White spaces in the tags [5].
   Example :
   Stego key:
   <tag>,</tag>, or <tag/>……0
   <tag >,</tag >, or <tag />…..1
   Stego data :
   <customer ><name>James </name ><id >2345</id></customer>
   Embeded data :
   101100

6. Appearing order of the Elements [5].
   stego key:
   <customer><name>NAME</name><id>ID</id></customer>……0
   <customer><id>ID</id><name>NAME</name></customer>…….1
   stego data:
   <customer><name>Minto</name><id>2354</id></customer>
   <customer><id>8976</id><name>Nikki</name></customer>
   Hidden Bits: 01

7. Appearing order of attributes [5].
   Example:
   stego key:
   <calendar month=”MONTH” date=”DATE”>EVENT</calendar>…….0
   <calendar date=”DATE” month=”MONTH”>EVENT</calendar>…….1
   stego data:
   <calendar month="JUL" date="7">charles’ birthday</calendar>
   <calendar date="24" month="JAN">merry’s birthday</calendar>
   Hidden bits:
   01

8. Elements Containing Other Element[5].
   Example:
   stego key:
   <favorite><game>NAME</game></favorite>…0
   <game><favorite>NAME</favorite></game> …1
   stego data:
   <game><favorite>CRICKET</favorite></game>
   <favorite><game>CRICKET</game></favorite>
   embedded data: 10

9. This Technique uses simple white space technique of steganography. But before transmitting data to receiver sender encrypts data using public key cryptography that provides more security to information [7].

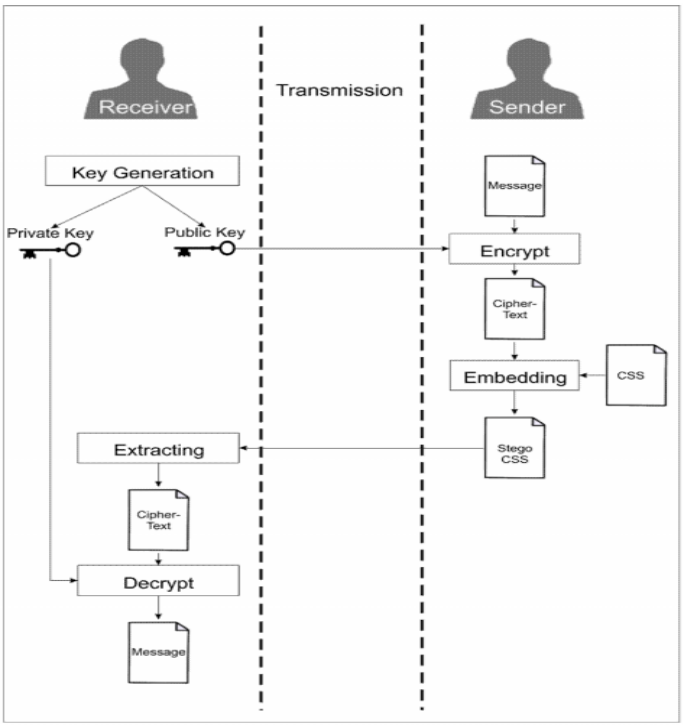

Fig-5: Steganography combined with cryptography

10. Random Character Technique [6].
    In this technique random characters are generated and inserted in XML tags.
    Original message:

```
<CD>
    <ITEM_ONE>
        <TITLE>Empire Burlesque</TITLE>
        <ARTIST>Bob Dylan</ARTIST>
        <COUNTRY>USA</COUNTRY>
        <COMPANY>Columbia</COMPANY
        >
        <PRICE>10.90</PRICE>
        <YEAR>1985</YEAR>
    </ITEM_ONE>
</CD>
```

    Embedded message:

```
<CKDBS>
    <IGTCLEFTWMOKLI_OGNPRESTN>
      <TFIXLTSBJLVHFJE>ESmwxpacgirtvjr
      stgmyefgpkhwBOuwvrdftlujifextbyusgfteb
      hqferovgtuocdenjirehgbiwntlx</>
      <AKRHYTHBJIEBOPSGFRDFTFGBEIT
      >BQovsbiex DPywglhdtawertnondtp</>
      <CTOXKUASRNFBHATQMPZFRSCTH
      INYOPWCTHD>URSFGANDT</>
      <CLOWHMSBUPXCFLAVKQISNDTU
      QLXYFVWTOLS>CTowglkrtudvhwmzn
      kspboastvjihstveunaqjdtvpsa</>
      <PARDFIGXOCSTHUELQIBS>14067.9
      3089</>
      <YDETVAWKORDFTE>109458379525 9
      0</>
    </IGTCLEFTWMOKLI_OGNPRESTN>
</CKDBS>
```

11. By changing the case of the characters of the tags data can be hidden. Fig-6 gives idea about this technique

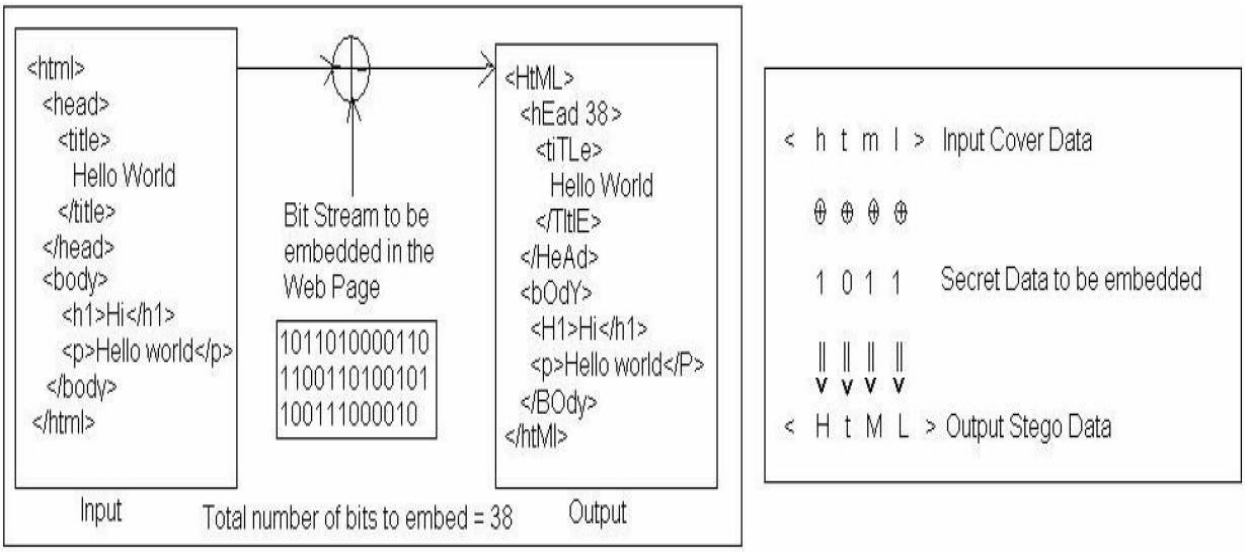

Fig-6: Hide data by changing the case of the characters

IV. Limitations of Web Page Text Steganography

1. HTML page contains less data hiding places & in most of the technique one hiding place hides only one bit either 1 or 0 so the message of large size is difficult to hide.
2. If white space is used to hide the data then the difficulty is that some of the HTML editors remove extra white spaces from the document & destroy the hidden data
3. Inserting additional spaces to represent information results in increase in stego-cover object size so it becomes suspicious.

V. Proposed Solution to Remove the Limitations & Provide More Security

Try to hide hexadecimal data in html document to overcome the problems of limited amount of hiding places & increase in size of document so one hiding place can hide equal to 4 bits.

1. To avoid the problem of limited amount of hiding places on time html page generation can be useful. It generates the hiding places as per stego-message requirement.
2. To provide more protection to stego-message steganogrpahy combined with cryptography can be used .

VI. Conclusion

Now a day, because of increasing amount of security threats protection of data is required. Steganography provides security of information by hiding it in carrier. This survey paper includes the classification of steganography techniques and techniques that already been implemented to hide information in web documents. Data hidden in the web document is less suspicious in compare of other carriers because HTML WebPages are now a routine part of everyone's life and html document contains the considerable number of tags, attributes & other elements in which data can be hidden. We can hide information in HTML, CSS, XML, JavaScript etc. So we have more options to secure the information.

REFERENCES

[1] Mohit Garg, "A Novel Text steganogrpahy Technique Based on HTML Document", International Journal of advanced Science and Technology Vol. 35, October 2011.
[2] Susmita Mahato, Dilip Kumar Yadav , Danish Ali Khan ,"A Modified Approach to Text Steganography using Hyper Text Markup language", Third international conference on Advanced Computing & Communication Technologies , 2013 IEEE
[3] Mohammad Shirali Shahreza, "A New Method for Steganogrpahy in HTML Files", Advance in computer, Information and System Science & Engineering, 247-251 , 2006 Springer
[4] Prem singh , Rajat Chaudhary and Ambika Agarwal, "A Novel Approach of Text Steganogrpahy based on null spaces", IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 4(July- Aug , 2012), Page 11-17
[5] Shingo Inoue, Ichiro Murase, Osamu Takizawa, Tsutomu Matsumoto, Hiroshi Nakagawa, "A Proposal on information Hiding Mehods Using XML"
[6] Aasma Ghani Menon , Sumbul Khwaja , Asadullah Shah, "Steganogrpahy: A new Horizon for safe Communication Through XML",Journal of Theoretical and Applied Information Technology(JATIT) 2008.
[7] Herman kabetta , B. Yudi Dwiandiyantaaa, Suyoto, "Information Hiding in CSS : A secure scheme Text Steganogrpahy Using Public Key Cryptosystem",International Journal on Cryptography and information Security (IJCIS),Vol.1,December 2011
[8] Hitesh Singh, Pradeep Kumar Singh, Kriti Saroha," A Survey on Text Based Steganography", School of Information Technology, Center for development of Advance Computing, Noida, India,2009.

[9] Sabu M Thampi," Information Hiding Techniques: A Tutorial Review", ISTE-STTP on Network Security & Cryptography, LBSCE 2004.