Secret Communication through Web Pages Using Special Space Codes in HTML Files

I-Shi Lee[a, c] and Wen-Hsiang Tsai [a, b,*]

[a]Department of Computer Science National Chiao Tung University, Hsinchu, Taiwan 30010.
[b]Department of Information Communication, Asia University, Taichung, Taiwan 41354.
[c]Department of Management Information Technology and Science, Institute of Northern Taiwan, Taipei, Taiwan.

Abstract: A secret communication method via web pages by embedding special space codes in HTML text to encode message bits in triplets is proposed. The codes, replacing the ASCII space code 20, appear as white spaces as well, creating a steganographic effect. Such codes are found by a systematic investigation of all the coding systems applicable in the HTML. Message hiding and recovery processes are also proposed. The character string of each message, before being embedded, is randomized with a secret key to enhance the security against illegal intercept and extraction. The embedded message is non-destructible unless the web page server is intruded. Experimental results show the feasibility of the proposed method.

Keywords: secret communication; data hiding; ASCII codes; space codes

1. Introduction

Data hiding with steganographic effects is a good way for secret communication [1]. Due to high accessibility on the Internet, it is convenient to use the web page as a communication channel by hiding secret messages in the HTML file of a cover web page. A merit here is that the secret message cannot be destructed illegally unless the website publishing the web page is intruded and the HTML file modified.

About hiding data in the HTML, Shirali-Shahreza [2] protects a Java applet in an HTML file from being copied by hiding a special 8-character string with a secret key within the Java applet. Wu and Lai [3] hide binary data in HTML files using various properties of tags like attributes for bit encoding. Wu, Chang, and Yang [4] use hash functions to compute digests of web page contents as fragile watermarks for tamper-proof. Chang and Tsai [5] insert extra white spaces in HTML text to encode bits for watermarking, as done by some commercial software [6].

In this paper, a new secret communication method by embedding special space codes in the HTML files of web pages is proposed. These codes appear as white spaces in the web page, and so may be used to encode secret message bits with steganographic effects. The codes are the result of a thorough investigation of all possible coding systems which can be applied in the HTML file. There are many of such codes, and each of them may be used to encode at least three message bits, increasing the data hiding capability and elimi-

[*] Corresponding author; e-mail: whtsai@cis.nctu.edu.tw

Accepted for Publication: December 25, 2008

nating the weakness of certain methods [5] of using more than two space codes to encode one bit and creating undesirable double spacing at originally single-spaced between-word locations.

The proposed method carries out the communication work between two sites, a sender and a receiver, through the Internet via web page publishing and downloading in the following way.

1. At the sender site:
   - 1.1 Create a web page containing mainly a piece of text.
   - 1.2 Hide the secret message to be transmitted in the HTML file of the page by the proposed method.
   - 1.3 Publish the web page on the Internet to make it accessible.
2. At the receiver site:
   - 2.1 Browse the web page on the Internet.
   - 2.2 Download its HTML codes by a code editor like UltraEdit or by a special program (not directly by the web browser using the “save as new file” command).
   - 2.3 Extract the secret message hidden in the codes by the proposed method.
   - 2.4 In the sequel, we describe how secret messages are encoded in Section 2, how they are hidden and recovered in Section 3. An experimental result is presented in Section 4, followed by a conclusion in Section 5.

2. Secret message coding using space characters in HTML

The HTML, Hypertext Markup Language, was created for describing the structure of a web page, including its appearance and semantics. Many coding systems are applicable in the HTML to specify characters used in the web pages. It is found in this study that there exist many codes in the HTML, all of which appear to be a *white space* in the window of the web page browser of the Internet Explore (IE). These codes come from two distinct types of space characters, named (*normal*) *space and non-breaking space*, and are specified in the following ways.

1. *Direct character entry of the (normal) space* ---

   A white space will appear in a line of HTML if the space bar on the keyboard is pushed during character typing, and the hexadecimal ASCII code 20 will be inserted in the program codes of the HTML file.
2. *Numeric character reference of the (normal) space* ---

   We can also represent a (normal) space character in the HTML using a so-called *numeric character reference*, by the form **&#xhhhh;**, where hhhh = 0020 is the hexadecimal value representing the character's Unicode scalar value; or by the form **&#dddd;**, where dddd = 0032 is the decimal value equivalent to the hexadecimal value. That is, we may represent the white space as **&#x20;** or **&#32;**. It is found in this study that the code **&#32** with the semicolon “;” missing is displayed as a space as well in the IE browser, while the code **&#x20** without the semicolon will *not* but as the code &#x20 *itself*, a peculiar phenomenon! A constraint to use **&#32** is that the character following it should not be a digit number; otherwise, it will become another code. We assume this constraint is satisfied in the HTML text in which this code is embedded.
3. *Numeric character reference of the non-breaking space* ---

   The non-breaking space with the hexadecimal ASCII code A0 is displayed in a web page browser like IE as a white space, too. Therefore, we may similarly represent it in the HTML using a numeric character reference, by one of the three forms **&#xA0;**, **&#160**;, and **&#160** (without a semicolon).
4. *Character entity reference of the non-breaking space* ---

The HTML accepts a third way of character specification, called *character entity reference*, which is a short-length text name used to identify a character. For the non-breaking space, it is **&nbsp;**. It is found that without the semicolon, the code **&nbsp** still appears to be a white space, so two codes are available for representing the white space.

Totally, nine distinct codes may be used to specify a character which appears to be a white space in the web page browser of the IE, as summarized in Table 1. They are called *space codes* subsequently. An illustration of the appearances of all the space codes is shown in Figure 1. The first eight space codes of the nine ones are used to encode three message bits in this study as shown in the last table column, although all nine of them may be used to encode a digit of a novenary number as well.

3. Message hiding and security enhance ment

During message hiding, we regard a given message as a sequence of characters, including letters, punctuations, white spaces, symbols, etc. Each character is represented as an 8-bit ASCII code, resulting in a string of bits which we encode three by three into the first eight space codes shown in Table 1. Each space code is then embedded at a between-word location in the *cover text* in the HTML file, replacing the original code 20h there, resulting in a *stego-text*. The embedded codes, after being extracted during message recovery, can be decoded uniquely by table lookup using Table 1.

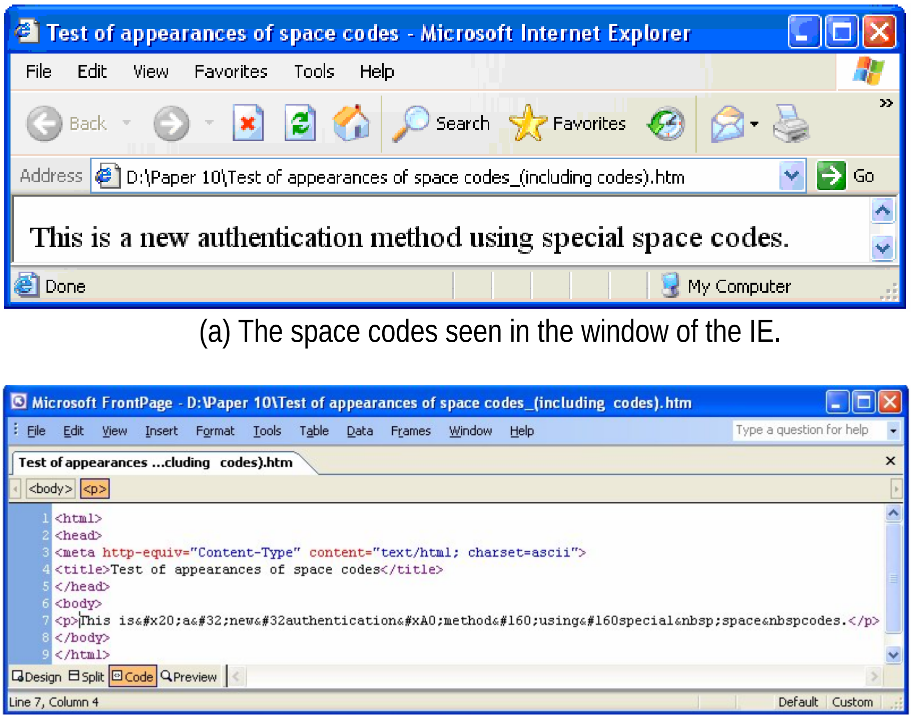
To increase the security of the embedded message, we use a random number generator to randomize the order of the characters in the message string before they are encoded sequentially. A secret key is provided as the seed for the generator. The key is used again in message recovery to re-arrange the order of the extracted characters. Without the key, if the hidden characters cannot be properly re-ordered to get the correct message.

Algorithm 1. Embedding of a secret message.

```
Input: a secret message S in the form of a character string, a cover HTML text T,
       a secret key K, and a random number generator f.
Output: a stego-HTML text $T'$ with $S$ embedded.
Steps:
1. Create a randomized version $S' = C_1'C_2'...C_n'$ of $S = C_1C_2...C_n$ in the
   following way, where $C_i$ and $C_i'$ represent characters of $S$ and $S'$,
   respectively, and $n$ is the number of characters in $S$.
   1.1 Generate n distinct random numbers k1, k2, ..., kn, within the range of 1
       through n using the generator f with the secret key K as the seed.
   1.2 For i = 1, 2, ..., n, take Ci' in S' to be Cki in S.
2. Convert the length n of S in the unit of character into a binary number and add
   leading 0's to it to form a 3m-bit binary string B, where m is a pre-selected
   integer such that 3m is no smaller than the length of any possible message to be hidden.
3. Transform each character in S' into its 8-bit binary ASCII code and concatenate
   them to form a binary string S1.
4. Concatenate B and S1 to form a binary string S".
5. Embed S" in T in the following way.
   5.1 Append zero, one, or two 0's to S" to form another binary string S2 with its
       length being a multiple of 3.
   5.2 Encode every three bits of $S_2$ into a space code $D$ according to the last
       column of Table 1.
   5.3 Embed $D$ in $T$ by replacing the (normal) space code 20h at a between-word
       location, starting from the top leftmost one in $T$ in a raster scanning order.
6. Take the resulting HTML text T' as the output.
```

Table 1. Character representations in HTML

No.	name	Reference type	Code type	Code inserted in HTML	Bits en-coded
1	(normal) space	direct character entry	ASCII	typed space (with **20h** inserted)	000
2	(normal) space	numeric character reference	Unicode	**&#x20;**	001
3	(normal) space	numeric character reference	Unicode	**&#32;**	010
4	(normal) space	numeric character reference	Unicode	**&#32**	011
5	non-breaking space	numeric character reference	Unicode	**&#xA0;**	100
6	non-breaking space	numeric character reference	Unicode	**&#160**;	101
7	non-breaking space	numeric character reference	Unicode	**&#160**	110
8	non-breaking space	character entity reference	HTML name	**&nbsp;**	111
9	non-breaking space	character entity reference	HTML name	**&nbsp**	unused



D:\Paper 10\Test of appearances of space codes_(including codes).htm

This is a new authentication method using special space codes.

(a) The space codes seen in the window of the IE.

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ascii">
<title>Test of appearances of space codes</title>
</head>
<body>
<p>This is&#x20;a&#32;new&#32authentication&#xA0;method&#160;using&#160special&nbsp;space&nbspcodes.</p>
</body>
</html>
```

(b) The codes inserted at between-word locations seen in the window of the FrontPage.

Figure 1. Appearances of nine space codes as white spaces in the window of the IE.

In the above algorithm we assume that the text T is long enough to embed the message S. Also, the length of the message is also embedded in the leading between-word locations in T. This is necessary for the later work of message recovery to extract a correct numbers of characters from the stego-text. The detailed algorithm for extracting the embedded message is as follows.

Algorithm 2. Extraction of a secret message.

Input: a stego-HTML text T' with a message S embedded, and a secret key K and a random number generator f as those used in Algorithm 1.

Output: the embedded message S.

Steps:

1. Extract the length n of the embedded message S in T' in the following way.
   - 1.1 For each of the m leading between-word locations in T' where m is a pre-selected integer mentioned in Algorithm 1, acquire the space code embedded there and decode it into three bits according to the last column of Table 1, resulting in a $3m$-bit binary string B.
   - 1.2 While ignoring the leading 0's in B, convert it into an integer n which presumably is the length of the embedded message S.
2. Compute the value $n_1 = \lceil n\times 8/3 \rceil$ which is the number of between-word locations in T' where S is embedded.
3. For each of the n_1 between-word locations after the m leading ones in T', acquire the space code there and decode it into three bits according to the last column of Table 1, resulting in $3n_1$-bit binary string S_2.
4. Take the leading $n\times 8$ bits of S_2 to form a string S' and transform every 8 bits of S' into an ASCII character.
5. Create a randomized version $S = C_1C_2...C_n$ of $S' = C_1'C_2'...C_n'$ in the following way, where C_i and C_i' represent characters of S and S', respectively, and n is the number of characters in S'.
   - 5.1 Generate n distinct random numbers k_1, k_2, ..., k_n, within the range of 1 through n using the generator f with the same secret key K as the seed.
   - 5.2 For i = 1, 2, .., n, take C_i in S to be C_{k_i}' in S', resulting in a string of characters $S = C_1C_2...C_n$ as the desired output.

For security consideration, the length of secret data should be long enough, e. g., more than 256 characters, to reduce the probability for a hacker to guess the message correctly. Otherwise, another way of security protection may be adopted, that is, to conduct the reordering operation in Step 1 of Algorithm 1 and Step 5 of Algorithm 2 in unit of bits instead of in unit of characters. Since there are normally very many bits, it is almost impossible to get a correct guess. If these measures of security enhancement are taken, it can be figured out from the above algorithm that without a correct key, the embedded message, even when the stego-text is intercepted, is almost impossible to be recovered by a hacker.

4. Experimental results

In order to have a clear illustration of the proposed method and to see clearly the embedded codes in web page and HTML editor windows, we report first a simple example of the experiments we conducted without embedding the length of the message and without using a secret key. Let the message to be embedded be “sky” whose three characters “s,” “k,” and “y” have 8-bit ASCII codes 01110011, 01101011, and 01111001, respectively. So the message in binary string form is 011 100 110 110 101 101 111 001 which includes eight 3-bit segments, and can be encoded into eight space codes &#32 &#xA0; &#160 &#160 &#160; &#160; &nbsp; &#x20;. We embedded these codes at eight consecutive between-word locations in the following HTML text:

This is a secret communication method through HTML files.

Then the result is:

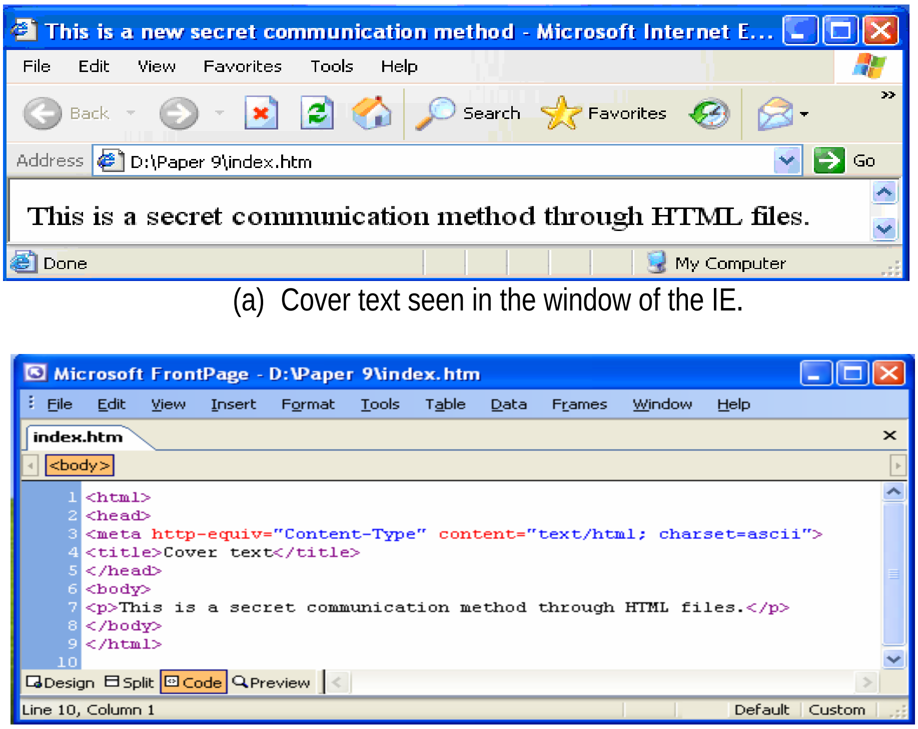
This&#32is&#xA0;a&#160new&#160communication&#160;method&#160;through&nbsp;HTML&#x20;files.

This stego-text, when observed in the web page browser of the IE, appears to be identical to that of the cover text, as shown in Figure 2.

Another example of our experimental results is shown in Figure 3, in which we show a cover text in the IE and the Frontpage windows in Figures 3(a) and 3(b), respectively; and a secret message in the Notepad window in Figure 3(c). The length of the message is 96 characters which are embedded first into the cover text as a 15-bit number. The stego-text appearing in the IE and the Frontpage windows is shown in Figures 3(d) and 3(e), repsectively. From the identicalness of Figures 3(a) and 3(d), the steganographic effect of the space codes is confirmed.

5. Conclusion

A new secret communication method via web pages using special space codes in HTML files has been proposed. These codes appear as white spaces in the web page, and so may be used to encode secret message bits with steganographic effects. The codes are the result of a thorough investigation of all possible coding systems which can be applied in the HTML file. The character string of each message, before being embedded, is randomized with a secret key to enhance the security against illegal intercept and extraction. The original message embedded in the HTML text is non-destructible unless the web page server is intruded. Our experimental results show that the proposed method is feasible. Future researches may be directed to utilizing the space codes in other data hiding applications.


This is a new secret communication method - Microsoft Internet E...
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Done
My Computer

D:\Paper 9\index.htm

This is a secret communication method through HTML files.

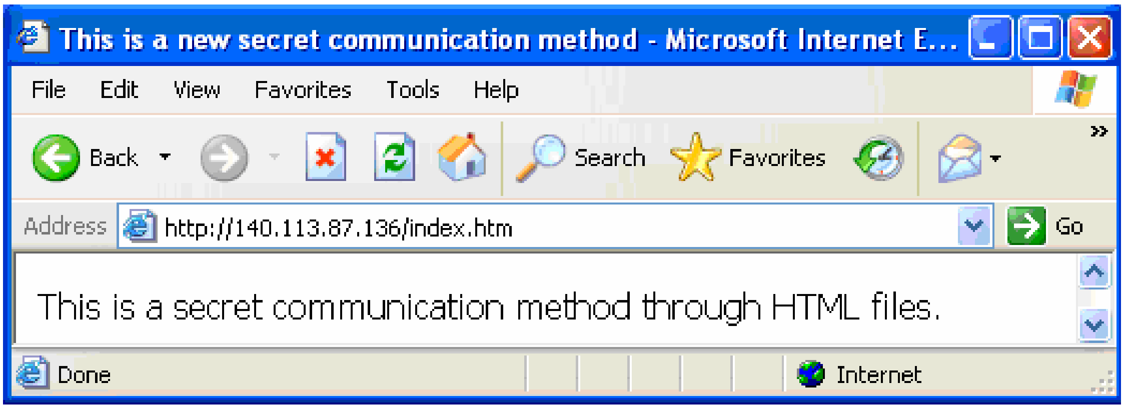
(a) Cover text seen in the window of the IE.

Microsoft FrontPage - D:\Paper 9\index.htm
File Edit View Insert Format Tools Table Data Frames Window Help
index.htm
<body>
Design Split Code Preview
Line 10, Column 1
Default Custom

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ascii">
<title>Cover text</title>
</head>
<body>
<p>This is a secret communication method through HTML files.</p>
</body>
</html>
```

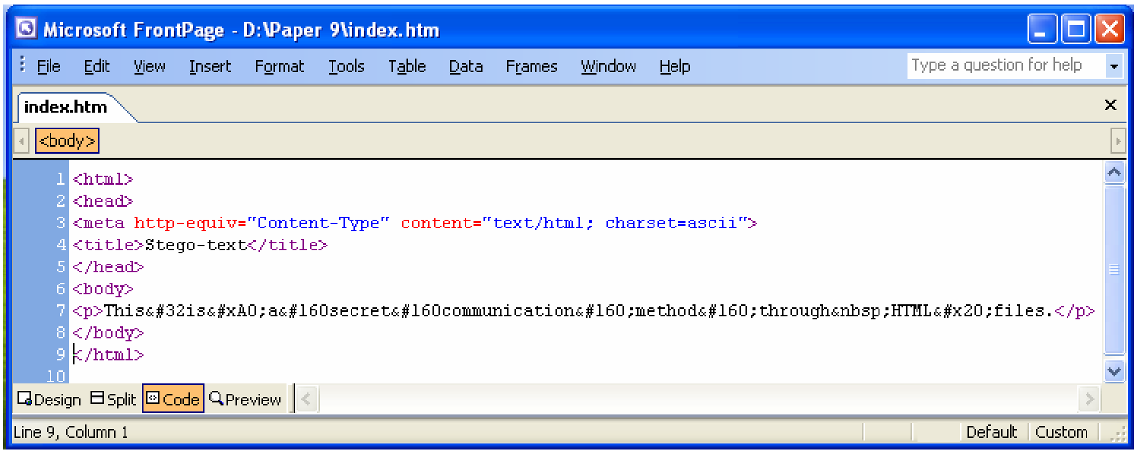
(b) Cover text seen in the window of the FrontPage editor.

Figure 2. Invisibility of space codes for the message “sky” in an HTML text.


This is a new secret communication method - Microsoft Internet E...
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://140.113.87.136/index.htm
Go
Done
Internet

This is a secret communication method through HTML files.

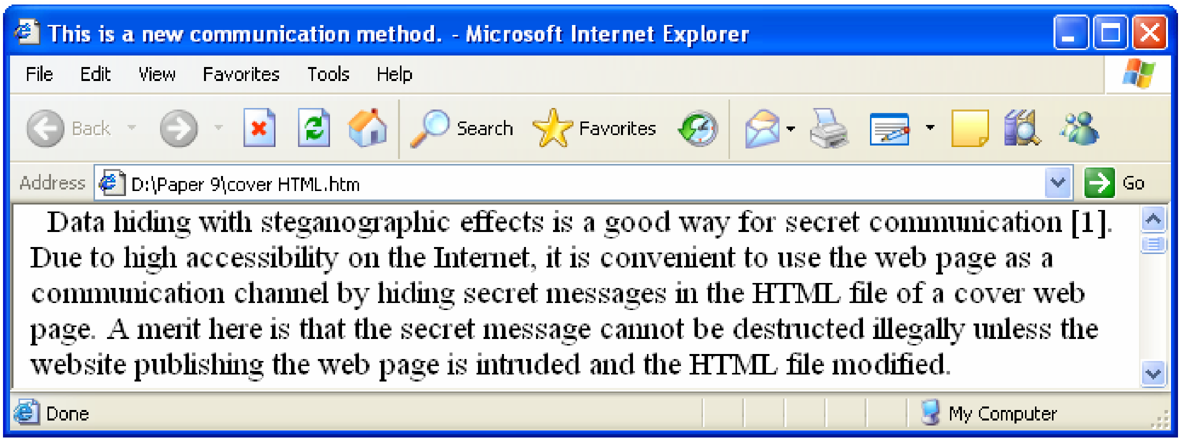
(c) Stego-text seen in the window of the IE.


Microsoft FrontPage - D:\Paper 9\index.htm
File Edit View Insert Format Tools Table Data Frames Window Help
Type a question for help
index.htm
<body>
Design Split Code Preview
Line 9, Column 1
Default Custom

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=ascii">
<title>Stego-text</title>
</head>
<body>
<p>This&#32is&#xA0;a&#160secret&#160communication&#160;method&#160;through&nbsp;HTML&#x20;files.</p>
</body>
</html>
```

(d) Stego-text seen in the window of the FrontPage editor.

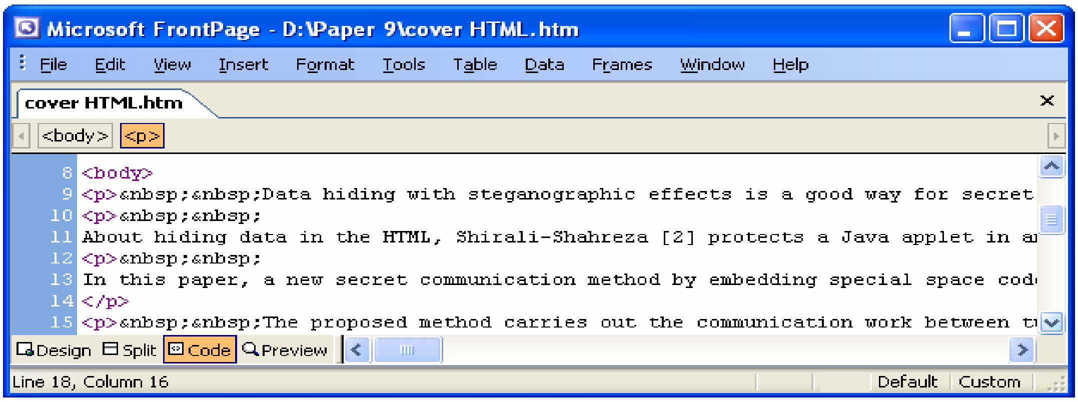
Figure 2. Invisibility of space codes for the message “sky” in an HTML text (cont’d).


This is a new communication method. - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
D:\Paper 9\cover HTML.htm
Go
Done
My Computer

Data hiding with steganographic effects is a good way for secret communication [1]. Due to high accessibility on the Internet, it is convenient to use the web page as a communication channel by hiding secret messages in the HTML file of a cover web page. A merit here is that the secret message cannot be destructed illegally unless the website publishing the web page is intruded and the HTML file modified.

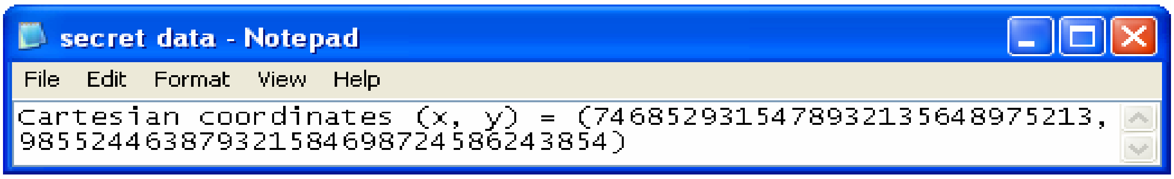
(a) Cover text seen in the window of the IE.

Figure 3. The embedded secret data.


Microsoft FrontPage - D:\Paper 9\cover HTML.htm
File Edit View Insert Format Tools Table Data Frames Window Help
cover HTML.htm
<body> <p>
Design Split Code Preview
Line 18, Column 16
Default Custom

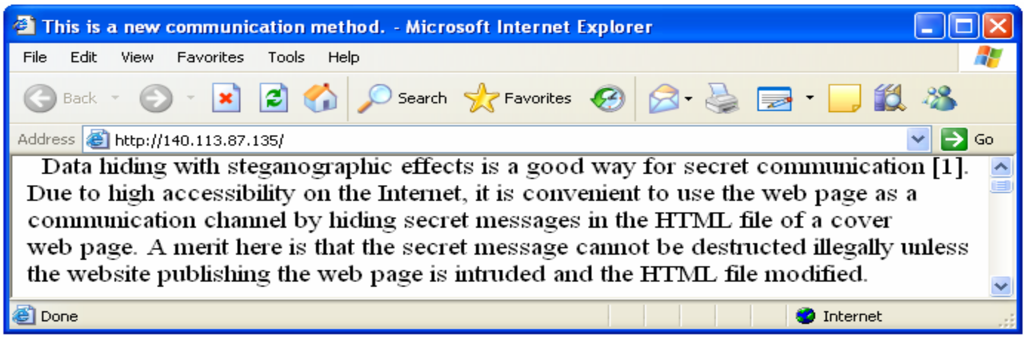
```
<body>
<p>&nbsp;&nbsp;Data hiding with steganographic effects is a good way for secret
<p>&nbsp;&nbsp;
About hiding data in the HTML, Shirali-Shahreza [2] protects a Java applet in a
<p>&nbsp;&nbsp;
In this paper, a new secret communication method by embedding special space cod
</p>
<p>&nbsp;&nbsp;The proposed method carries out the communication work between t
```

(b) Cover text seen in the window of the FrontPage editor.



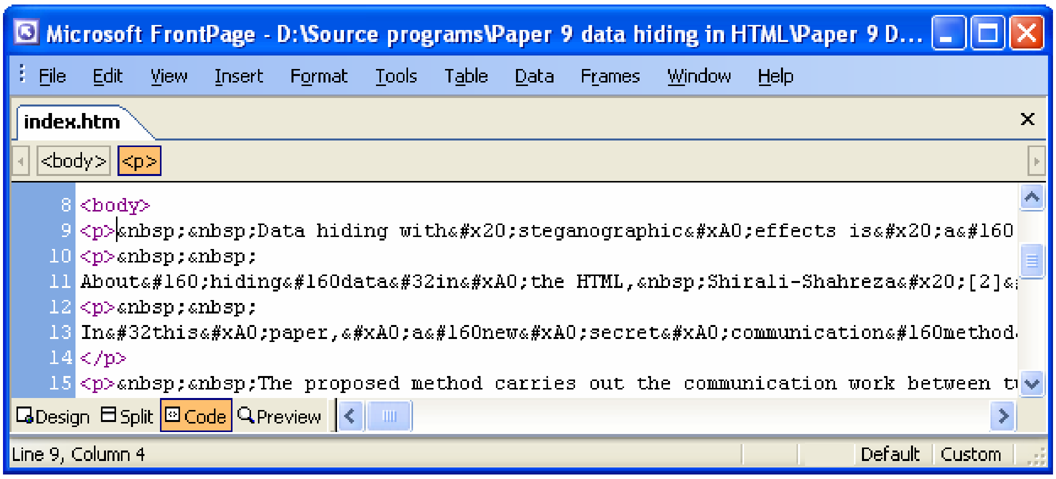
```
Cartesian coordinates (x, y) = (7468529315478932135648975213,
985524463879321584698724586243854)
```

(c) A secret message seen in the Notepad window.


This is a new communication method. - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://140.113.87.135/
Go
Done
Internet

Data hiding with steganographic effects is a good way for secret communication [1]. Due to high accessibility on the Internet, it is convenient to use the web page as a communication channel by hiding secret messages in the HTML file of a cover web page. A merit here is that the secret message cannot be destructed illegally unless the website publishing the web page is intruded and the HTML file modified.

(d) Stego-text seen in the window of the IE.


Microsoft FrontPage - D:\Source programs\Paper 9 data hiding in HTML\Paper 9 D...
File Edit View Insert Format Tools Table Data Frames Window Help
index.htm
<body> <p>
Design Split Code Preview
Line 9, Column 4
Default Custom

```
<body>
<p>&nbsp;&nbsp;Data hiding with&#x20;steganographic&#xA0;effects is&#x20;a&#160
<p>&nbsp;&nbsp;
About&#160;hiding&#160data&#32in&#xA0;the HTML,&nbsp;Shirali-Shahreza&#x20;[2]&
<p>&nbsp;&nbsp;
In&#32this&#xA0;paper,&#xA0;a&#160new&#xA0;secret&#xA0;communication&#160method
</p>
<p>&nbsp;&nbsp;The proposed method carries out the communication work between t
```

(e) Stego-text seen in the window of the FrontPage editor.

Figure 3. The embedded secret data (continued).

Acknowledgement

This work was supported jointly by the NSC projects No. 96-2752-E-009-006-PAE and No. 96-2422-H-009-001.

References

[1] Katzenbeisser, S., and Petitolas, F. A. P. 2000. *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House, Boston, Massachusetts, U.S.A.

[2] Shirali-Shahreza, M. Dec. 2006. "*Java applets copy protection by steganography*", *Proc. of 2006 Int'l Conf. on Inteligent Information Hiding & Multimedia Signal Processing*, Pasadena, California, U.S.A., 388-391.

[3] Wu, D. C., and Lai, P. H. June, 2005. "*Novel Techniques of Data Hiding in HTML Documents*", *Proc. 2005 Conf. on Digital Contents Managements & Applications*, 21-30, Kaohsiung, Taiwan.

[4] Wu,C. C., Chang, C. C., and Yang, S. R. 2007. "*An efficient fragile watermarking for web pages tamper-proof*", in *Advances in Web and Network Technologies, and Information Management*, *Lecture Notes in Computer Science*, Vol. 4537, Springer, Berlin, Germany, 654-663.

[5] Chang, Y. H., and Tsai, W. H. Dec. 2003. "*A steganographic method for copyright protection of HTML documents*", *Proc. of 2003 Nat'l Computer Symposium*, Taichung, Taiwan.

[6] Invisible Secrets. Retrieved May , 2008, from http://www.invisiblesecrets.com