A Watermarking for HTML Files Based on Multi-channel System

Yung-Chen Chou
Department of Computer Science and Information Engineering,
Asia University, Taichung 41354, Taiwan
yungchen@gmail.com

Iuon-Chang Lin
Department of Information Management,
National Chung Hsing University, Taichung, Taiwan
iclin@nchu.edu.tw

Ping-Kun Hsu
Department of Information Management,
National Chung Hsing University, Taichung, Taiwan
kelp11211@gmail.com

Abstract

In this paper a novel HTML file watermarking method is presented. An HTML file is to present the personality of a user or to advertise a company. In order to increase the robustness of the proposed watermarking, watermark data is concealed into the HTML file many times using a multi-channel system. The watermark data can be extracted from the watermarked HTML when arguing the copyright issue. Because the visual quality of watermarked HTML is one of the most important requirements in designing watermarking technique, the watermarked HTML will not have any perceptible distortion when watermark has been embedded into the HTML by the proposed method. Also, the voting strategy is adopted in the proposed method for achieving the robustness. Experimental results show that the proposed method is feasible for achieving copyright protection of HTML files.
Keywords : *HTML, multi-channel system, voting, robust, watermarking*

1 Introduction

As the digital age is dawning, any user can easily download/upload digital files from the Internet. Thus, many problems arise from the activities on the Internet, such as the copyright protection, the secret data delivery, and secure communications. Watermarking is one of the most popular techniques for protecting the copyright of the digital files. In recent years, the researchers focus on the digital image, video, and audio copyright protection. Generally, webpage is another medium for sharing knowledge or advertising products. How to embed watermark data into an HTML file to achieve the goal of copyright protection becomes a serious research topic.

Generally, the watermarking technique can be briefly classified into visible watermarking and invisible watermarking. Visible watermarking is to add a logo into the digital file for announcing the copyright. Invisible watermarking is focused on the quality of the watermarked medium so the watermark can't be distinguished by human's sense (e.g., eye and ear). Nowadays, the digital

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Watermarking for HTML file</title></head>
5      <body bgcolor="pink">
6          <p><font size="+2" color="#443389" face="Times New Roman">
7              This is an illustration about the attributes of tags in HTML.
8              </font></p>
9          Go to <a href="http://www.asia.edu.tw/" target="_blank">Asia</a>!
10     </body>
11 </html>
```

Figure 1. An illustration of HTML source code

watermarking techniques are concerned with concealing watermark data into host media with high quality and robustness [5, 6, 7, 8, 9, 10, 11, 12].

In addition to digital image, video, and audio, the text file should also be protected its copyright. A user builds up a colorful webpage to present his creation on the Internet. No one can copy someone's webpage without the permission from the author. So, the webpage's authors need to embed watermark information into the HTML file for declaration of the copyright. Because HTML file has no too much redundant information, it is hard to embed the watermark data into the HTML file with robustness. Further, HTML is composed by the tags and many multimedia resources (i.e., referring to Fig. 1) to show up the colorful contents. The copyright protection is trying to adjust the tags' attributes, values, quotation marks, etc. for implying watermark.

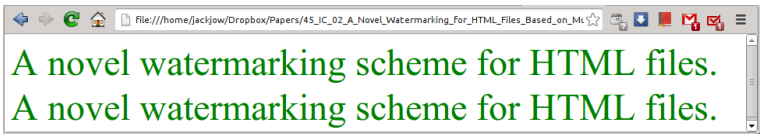
Recent years, many researchers presented a lot of data hiding techniques to achieve the secret data delivery. Sui and Luo presented a steganography method based on hypertext [3]. Sui and Luo indicated that the letter case of hyper text's tags can be used to imply a secret message. Yang and Yang presented a steganography method for webpage by adjusting the quotation marks of tag's attribute values [4]. Yang and Yang's method uses different quotation marks for tags' attributes to conceal secret data. Huang et al. pointed out that the arrangement of tags' attributes can be used to conceal secret data [1]. HTML tags may contain some attributes for showing personality of webpage. Lee and Tsai proposed a data hiding scheme for an HTML file by changing the space code of between-word locations [11]. The proposed method is inspired from the methods mentioned above. We design a multi-channel system to conceal watermark into an HTML file many times. In extracting phase, the watermark is extracted from the channels separately and the voting strategy is adopted to get the real watermark data.

The remaining sections are organized as follows. Section 2 briefly describes the steps of related works which were useful for understanding the proposed method. Further, the proposed watermarking method is detailed in Section 3. The performance evaluation is summarized in Section 4. Finally, some conclusions are summarized in Section 5.

2 Related Works

2.1 Case of Tag Letter

Sui and Luo presented a steganography method based on hypertext [3]. Sui and Luo indicated that the letter case of hyper text's tags can be used to imply secret message. The case of letters in the tag will not affect the representation when user browsing the web page. Thus, Sui and Luo's method is to adjust the letter case of a tag to imply secret data. For example, "<html>", "</html>", "<body>", and "</body>" are the fundamental tags for a simple HTML file. "<html>" tag will be modified as "</HTmL>" for implying secret "1101", according to Sui and Luo's method. Fig. 2 illustrates an example for Sui and Luo's method.



(a) the browsing result for testing difference tag letter case

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Case of Tag Letter</title></head>
5      <body>
6          <font size="+4" color="green">
7              A novel watermarking scheme for HTML files.
8          </font><br/>
9          <Font size="+4" color="green">
10             A novel watermarking scheme for HTML files.
11         </Font>
12     </body>
13 </html>
```

(b) The Source code of (a)

Figure 2. The illustration of Sui and Luo's scheme

2.2 Quotation Marks

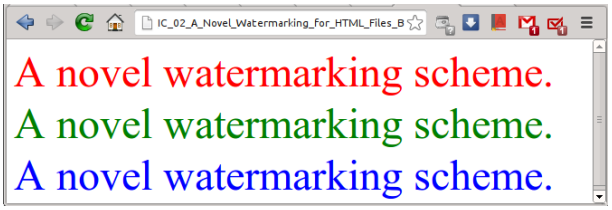
Yang and Yang presented a steganography method for webpage by adjusting the quotation marks of tag's attribute values [4]. Yang and Yang's method uses different quotation marks for tags' attributes to conceal secret data. Because the flexibility of tags' attributes value coding in HTML file, the tags' attribute value with different quotations can be used to imply secret data. Here, the quotation marks for tags' attribute values include a) value without quotation; b) value with a pair of single quotation; c) value with double quotations. For example, the tag "<font>" has "size" and "color" attributes, there has three ways for setting the font color as blue a) "<font color=blue>", "<font color = 'blue'>", and "<font color="blue">". Thus, sender and receiver can pre-share the rule of secret data embedding, for instance single quotation implying secret bit '1', otherwise implying secret bit '0'. Fig. 3 illustrates an example for Yang and Yang's method.

2.3 Attributes Sequence Permutation with Case of Name String

Huang et al. pointed out that the arrangement of tags' attributes can be used to conceal secret data [1]. HTML tags may contain some attributes for showing personality of webpage. For instance, "<font>" tag contains "face", "color", and "size" attributes. Also, different attributes arranging sequence will not affect the visual quality when browsing the webpage. Let $T(a_1, a_2, \ldots, a_n)$ is a tag T with attributes a_1, a_2, ..., a_n. Because T contains n attributes there has $n!$ sequences for arranging the attributes. After that, secret data can be embedded into an HTML file by permuting the attributes sequence in a tag.

3 The Proposed Scheme

The proposed method is inspired by the data hiding techniques mentioned in Section 2. For increasing the robustness of watermarking of an HTML file, the watermark is embedded into HTML four times by using different concealing strategies. Fig. 4 shows the proposed watermark embedding flowchart, where H and H' denote the original host HTML file and watermarked HTML file, respec-



(a) The browsing result for testing difference quotation marks

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Quotation Test</title></head>
5      <body>
6          <font size="+4" color=red>A novel watermarking scheme.</font><br/>
7          <font size="+4" color='green'>A novel watermarking scheme.</font><br/>
8          <font size="+4" color="blue">A novel watermarking scheme.</font><br/>
9      </body>
10 </html>
```

(b) The Source code of (a)

Figure 3. An example for Yang and Yang's method

tively. For simplicity, Table 1 summarizes the notations' definition for the proposed watermarking method.

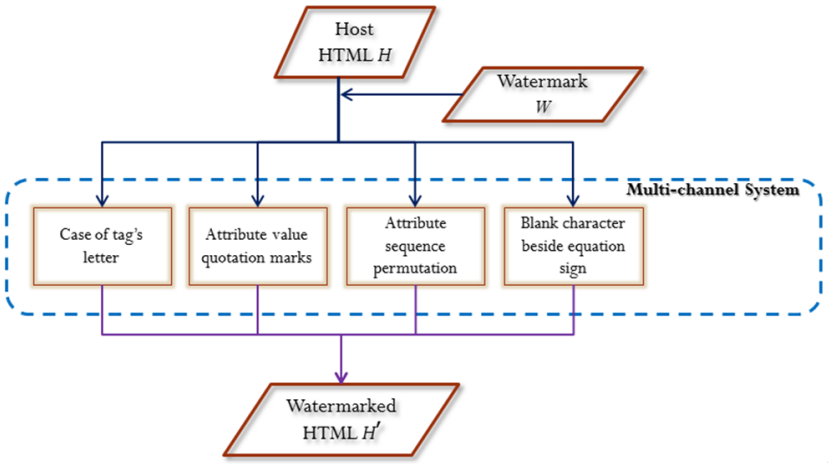


Figure 4. The embedding flowchart of the proposed method

3.1 Watermark Embedding Phase

The main idea of the proposed watermarking method is to conceal watermark data several times to increase the robustness. Thus, the case of tag name method, the attribute value quotation mark method, the tag's attributes sequence permutation method, and the blank characters beside equal sign method were adopted in the proposed watermarking scheme. The key steps of the proposed watermarking scheme for HTML file are summarized as follows:

Step 1: Figure out all of tags between tags "<body>" and "</body>". The upper case and lower case for tag's name are used to imply the watermark bit '1' and '0', respectively.

Table 1. The definition of notations

Notation	Definition
H	The original HTML file
H'	The watermarked HTML file
W	Watermark
W'	Extracted watermark
T_i	The i-th tag in H
$\|\|H\|\|$	The number of tags between tag "<body>" and "</body>"
S	The watermark string transferred from W
S_j	The j-th bit of S
S'	The extracted watermark string from H'
$\|\|S\|\|$	The length of S
$\|\|T_i\|\|$	The number of attributes of T_i

Step 2: For all tags staying between tags "<body>" and "</body>", the attribute value with different quotation marks will be used to imply watermark data. The single quotation and double quotation were used to imply watermark bit '0' and '1', respectively.

Step 3: In order to adopt the tag's attributes sequence permutation for embedding watermark data, the tag's attributed will be sorted by lexicographic order of attributes' name in ascending order. Further, for increasing embedding payload, the case of attributes' letter can also be used to conceal watermark data. Table 2 summarizes the embedding rules for concealing watermark data using attributes sequence permutation and case of attributes' name. Where $Ord(\cdot)$ denotes the value of input attribute name by lexicographic order. Fig. 5 shows an example embedding watermark data using attribute sequence permutation and the case of attributes' name.

Step 4: For tag's attribute value setting, the equal sign is staying between the attribute name and value. From programmer's point of view, adding blank characters beside the equal sign can increase the readability. Also, the blank characters used in the attribute setting will not affect the visual quality when browsing the watermarked HTML file. Thus, it has four cases can be used for implying watermark data.

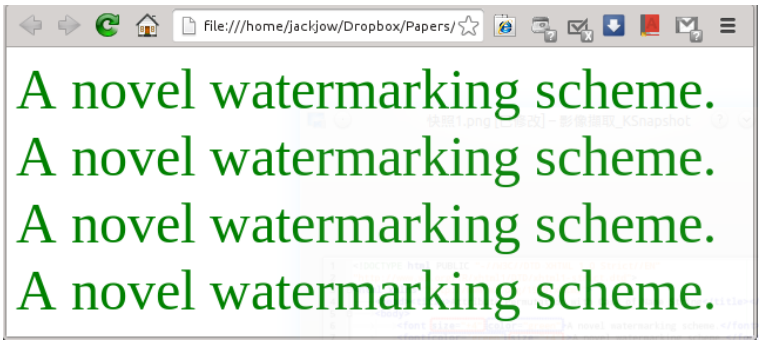
1. If "attribute = "value"", then implying watermark data "00";
2. If "attribute ="value"", then implying watermark data "01";
3. If "attribute= "value"", then implying watermark data "10";
4. If "attribute="value"", then implying watermark data "11".

Fig. 6 shows an example for embedding watermark data into an HTML file.

Table 2. The embedding rules for attributes sequence permutation

Attributes sequence	Letter case of a_1	Letter case of a_2	Watermark
$Ord(a_1) \leq Ord(a_2)$	Lower	Lower	000
		Upper	001
	Upper	Lower	010
		Upper	011
$Ord(a_1) > Ord(a_2)$	Lower	Lower	011
		Upper	100
	Upper	Lower	101
		Upper	111

According to the proposed multi-channel embedding method, the watermark is embedded four times in an HTML file. Table 3 summarizes the maximum payload for every channel.

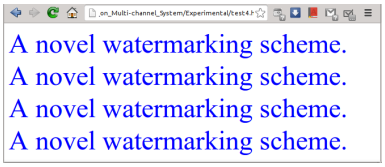


(a) The browsing result for testing difference attribute sequence permutation

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Attribute Permutation with Case of Name String</title></head>
5      <body>
6          <font size="+4" color="green">A novel watermarking scheme.</font><br/>
7          <font color="green" size="+4">A novel watermarking scheme.</font><br/>
8          <font SIZE="+4" COLOR="green">A novel watermarking scheme.</font><br/>
9          <font COLOR="green" SIZE="+4">A novel watermarking scheme.</font><br/>
10     </body>
11 </html>
```

(b) The source code for (a)

Figure 5. An example for embedding watermark data using attribute sequence permutation and the case of attributes' name



(a) The browsing result for testing difference blank beside equal sign

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Blank beside equal sign</title></head>
5      <body>
6          <font size="+4" color = "blue">A novel watermarking scheme.</font><br/>
7          <font size="+4" color ="blue">A novel watermarking scheme.</font><br/>
8          <font size="+4" color= "blue">A novel watermarking scheme.</font><br/>
9          <font size="+4" color="blue">A novel watermarking scheme.</font><br/>
10     </body>
11 </html>
```

(b) The source code for (a)

Figure 6. An example for concealing watermark using blank beside equal sign

3.1.1 Watermark Embedding Example

For protecting the copyright of HTML file, the watermark "000110" will be embedded into the HTML file. Figs. 7(a)-7(b) shows the source code of the original HTML file and the watermarked HTML file, respectively.

Table 3. The embeddable data of different situation

Embedding strategy	number	embeddable data
Case of attribute name	n_c	n_c
Attribute sequence permutation	n_s	n_s
quotation mark	n_q	n_q
Blank beside equal sign	n_e	$2n_e$

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Example</title></head>
5      <body bgcolor="gray">
6          <font size="+4" color="red">Hello</font><br/>
7          <font size="+4" color="green">World</font><br/>
8          <hr width="80%" align="center">
9      </body>
10 </html>
```

(a) The original HTML source code

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3  <html xmlns="http://www.w3.org/1999/xhtml">
4      <head><title>Example</title></head>
5      <body bgcolor="gray">
6          <font color="red" size="+4">Hello</font><br/>
7          <FONT size= "+4" color='green'>World</FONT><br/>
8          <HR ALIGN ="center" WIDTH ="80%">
9      </body>
10 </html>
```

(b) The watermarked HTML source code

Figure 7. The example of the proposed scheme

3.2 Watermark Extracting Phase

Watermark extracting is to extract the watermark data from the watermarked HTML file using four types extracting procedures. After that, a voting strategy is utilized for generating the final extracted watermarked. Thus, the proposed method can extract watermark data more correctly; even a part of the source code of watermark HTML has been altered by any reason. Fig. 8 illustrates the flowchart of the proposed watermark data extracting. The notation '$\bigotimes$' is the voting operation.

4 Experiment Results and Analysis

For evaluating the performance of the proposed method, an HTML file and a watermarked sized 6×6 are used for testing the proposed method. The HTML file and the watermark are shown in Figs. 9 and 10 respectively. Halftone watermark is easier to recognize the extracted watermark when arguing the copyright issue. Halftone image is more suitable for playing the watermark because the

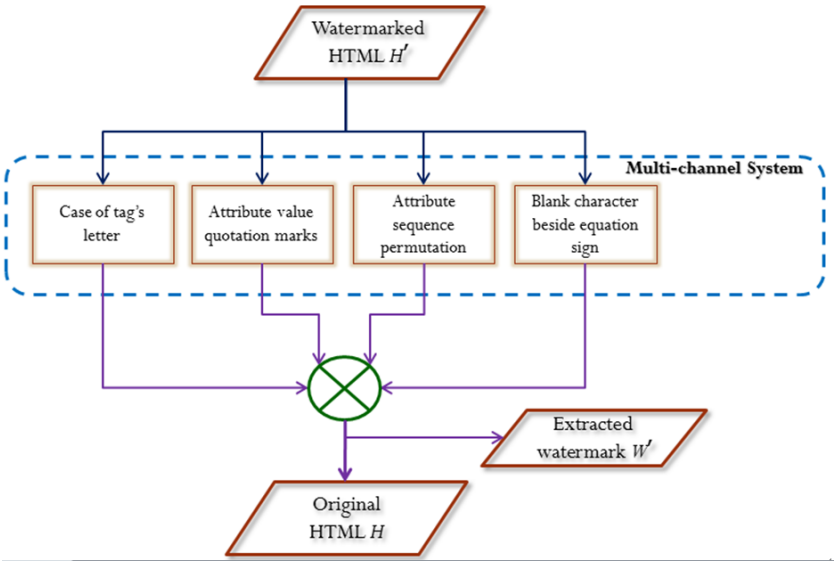


Figure 8. The proposed watermark extracting flowchart

size of halftone image is quite small comparing with grayscale image and color image. The source

original.html

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head><title>Watermarking Test</title></head>
<body bgcolor="#CDEFCE">
    <center><font size="7" color="green"><I>Welcome to my world</I></font></center>
    <hr width="70%" color="red">
    <table>
        <tr><td><font size="6" color="black"><U>Profile</U></font></td>
            <td></td></tr>
        <tr><td><table bgcolor="#ECFCEE">
                <tr><th align="right"><font size="4" color="gray">Name: </font></th>
                    <td><font size="4" color="#3937FF">Ping-Kun Hsu</font></td></tr>
                <tr><th align="right"><font size="4" color="gray">Age: </font></th>
                    <td><font size="4" color="#3937FF">Tweenty-four</font></td></tr>
                <tr><th align="right"><font size="4" color="gray">Gender: </font></th>
                    <td><font size="4" color="#3937FF">Male</font></td></tr>
                <tr><th align="right"><font size="4" color="gray">School: </font></th>
                    <td><font size="4" color="#3937FF">NCHU</font></td></tr>
                <tr><th align="right"><font size="4" color="gray">Major: </font></th>
                    <td><font size="4" color="#3937FF">M.I.S.</font></td></tr>
                <tr><th align="right"><font size="4" color="gray">E-mail: </font></th>
                    <td><a href="mailto:kelp11211@gmail.com">
                        <font size="4" color="gray">kelp11211@gmail.com</font></a>
```

Figure 9. The source code of a part of original HTML file

code of the watermarked HTML file is demonstrated in Fig. 11. Compared with Fig. 9 there are many different representations for the same tag in order to imply the watermark data. It is hard to distinguish the distortion when a user browses the watermarked HTML file. Table 4 gives the extracted results with four cases attack about removing watermark corresponding to one channel of multi-channel system in order.

Case 1: removing the secret data embedded in letter case of tag name

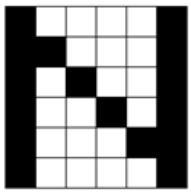
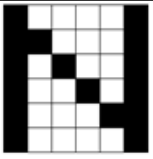
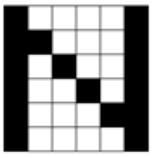
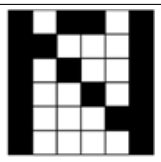
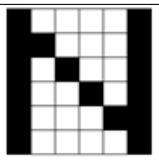
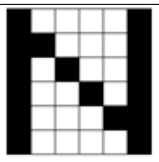
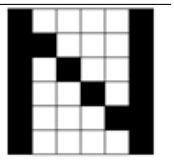
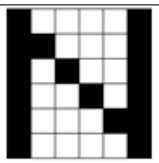
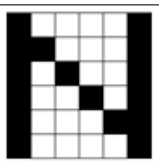
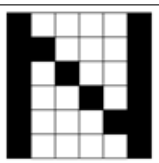
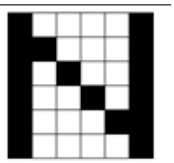


Figure 10. The watermark

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head><title>Watermarking Test</title></head>
<body bgcolor="#CDEFCE">
    <center><f0nT size= "7" color= 'green'><I>Welcome to my world</I></f0nT></center>
    <HR Width = '70%' Color = 'red'>
    <table>
        <tr><td><FoNt Color ='black' Size ="6"><U>Profile</U></FoNt></td>
            <td></td></tr>
        <tr><td><table bgcolor="#ECFCEE" border="0">
            <tr><TH align = "right"><FoNt Color ='gray' Size ="4">Name: </FoNt></TH>
                <td><f0nT size= "4" color= '#3937FF'>Ping-Kun Hsu</f0nT></td></tr>
            <tr><tH align= "right"><FoNt Color ='gray' Size ="4">Age: </FoNt></tH>
                <td><f0nT size= "4" color= '#3937FF'>Tweenty-four</f0nT></td></tr>
            <tr><Th align ="right"><FoNt Color ='gray' Size ="4">Gender: </FoNt></Th>
                <td><f0nT size= "4" color= '#3937FF'>Male</f0nT></td></tr>
            <tr><TH align = "right"><font size="4" color="gray">School: </font></TH>
                <td><f0nT size= "4" color= '#3937FF'>NCHU</f0nT></td></tr>
            <tr><TH align = "right"><FoNt Color ='gray' Size ="4">Major: </FoNt></TH>
                <td><f0nT size= "4" color= '#3937FF'>M.I.S.</f0nT></td></tr>
            <tr><TH align = "right"><FoNt Color ='gray' Size ="4">E-mail: </FoNt></TH>
                <td><a href="mailto:kelp11211@gmail.com">
                    <font size="4" color="gray">kelp11211@gmail.com</font></a></td></tr>
            </table></td>
```

Figure 11. A part of source code of the watermarked HTML file

Table 4. The four cases of extracting watermark

	Case 1	Case 2	Case 3	Case 4
First				
Second				
First ⋂ Second				

Case 2: removing the secret data embedded in quotation marks

Case 3: removing the secret data embedded in attributes permutation with upper-lower case

Case 4: removing the secret data embedded in equal sign side space

From **case 2** to **case 4**, the first and second extracted watermarks are the same to each other without any noise or data losing because the removing data is lesser than the remaining ones and don't affect the result of voting. Nevertheless, case 1 loses too much embedded watermark data, the first and second extracted watermarks are not the same even voting strategy had been adopted. For this situation, we regard the intersection of the two watermarks as the real watermark. It could filter out some noise came into being during voting phase.

5 Conclusions

In this paper, a novel watermarking scheme based on multi-channel system for HTML files is proposed. The proposed watermarking method employs several embedding strategies to conceal watermark four times. In case of an unexpected user copies the HTML file with little changes, the watermark can be approximately extracted by the proposed watermark extraction method. The voting strategy provides a good way to increase the survived watermark data when the HTML file encountered modification. The proposed method utilizes multi-channel embedding and voting strategy to achieve the goal of copyright protection of HTML files successfully.

References

[1] Huang, H.J., Zhong, S.H., and Sun, X.M., *An Algorithm of Webpage Information Hiding Based on Attributes Permutation*, Proceedings of the Fourth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Harbin, China, Aug. 2008, pp. 257-260.

[2] Lee, I.S. and Tsai, W.H., *Secret Communication through Web Pages Using Special Space Codes in HTML Files*, International Journal of Applied Science and Engineering, Vol. 6, No. 2, 2008, pp. 141-149.

[3] Sui, X.G. and Luo, H., *A New Steganography Method Based on Hypertext*, Proceedings of Asia-Pacific Radio Science Conference, Qingdao, China, Aug. 2004, pp. 181-184.

[4] Yang, Y. Y.J. and Yang, Y.M., *An Efficient Webpage Information Hiding Method Based on Tag Attributes*, Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, Aug. 2010, pp. 1181-1184.

[5] Automatic Information Processing Lab, *Digital Watermarking*, http://debut.cis.nctu.edu.tw/Demo/Watermarking/watermarking_new.html, available on Jun. 20, 2011.

[6] Kim, Y.W., Moon, K.A., and Oh, I.S., *A Text Watermarking Algorithm Based on Word Classification and Inter-word Space Statistics*, Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, Scotland , August 3-6,pp. 775-779, 2003.

[7] Wang, Z.H., Chang, C.C., Lin, C.C., and Li, M.C., *A Reversible Information Hiding Scheme Using Left-Right and Up-Down Chinese Character Representation*, Systems and Software, vol. 82, no. 8, pp. 1362-1369, 2009.

[8] Liu, T.Y. and Tsai, W.H., *A New Steganographic Method for Data Hiding in Microsoft Word Documents by a Change Tracking Technique*, IEEE Transactions on Information Forensics and Security, vol. 2, no. 1, pp. 24-30, 2007.

[9] Dey, S., Al-Qaheri, H., and Sanyal, Sugata, *Embedding Secret Data in HTML Web Page*, Image Processing & Communications Challenges, pp. 474-481, 2009.

[10] Qadir, M.A. and Ahmad, I., *Digital Text Watermarking: Secure Content Delivery and Data Hiding in Digital Documents*, IEEE Aerospace and Electronic Systems Magazine, vol. 21, no. 11, pp. 18-21, 2006.

[11] Lee, I.S. and Tsai, W.H., *Data Hiding in Emails and Applications Using Unused ASCII Control Codes*, Journal of Information Technology an Applications, vol. 3, no. 1, pp. 13-24, 2008.

[12] Chang, C.C., Wu, C.C., and Lin, I.C., *A Data Hiding Method for Text Documents Using Multiple-Base Encoding*, Communications in Computer and Information Science, Germany: SpringerLink,vol. 66, pp. 101-109, 2010.