Protection of Web Document using Watermarking: A Cryptographic Approach

Namrata R. Shaha
Technocrats Institute of Technology, RGPV Bhopal

Aishwarya Vishwakarma
Assistant Professor
Technocrats Institute of Technology, RGPV Bhopal

Bhupendra Verma, Ph. D
Professor Technocrats Institute of Technology, RGPV Bhopal

ABSTRACT

Watermarking is theory of protecting the documents from illegal copying, redistribution of the documents and to preventing from copyright violation. The document Watermark created in manner that it's inaccessible to the document viewer in normal circumstances. There are two main issues regarding web document watermarking is imperceptibility and robustness of watermark. In this paper, we proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes safer and sound for the reason that it the authentication for generation and extraction process. This system is elastic since it will take user defined string meant for watermark. An experimental result shows that system's work progress and it will accomplish fidelity, robustness, authenticity, and flexibility.

Keywords

Authentication, imperceptibility, robust, watermark, web document, Unicode.

1. INTRODUCTION

Over the past few years incredible enhancement of the internet and electronic networks, assist the circulation and exchange of information all over the world is rapid, effortless, and economical. Some crucial issues come into sight due to current data security schemes are inadequate. To control such issues there are several schemes comes in picture viz encryption, steganography, watermarking and so on. [1]

Digital watermarking is theory of hiding digital signal into digital document in a manner that it's imperceptible and inaccessible to the document viewer in normal circumstances. Digital watermarking is classified into visible and invisible watermarking. [2] Visible watermarking is viewable to normal human insight such as logo, bills. Invisible watermarking is not viewable to human insight such watermark requires some mathematical calculation or extra processing for retrieving by some authorized person.

Web page is the main component of web publications. Web document watermarking is to achieve the reliability of web pages which is a very famous and prosperous resource of information. Web document is primarily assembled by using HTML or XML and supplemented by other dynamic web techniques. So text document watermarking is applied to web document in a manner that it preserves the main features of web page that is web page display features, content, web page layout. We use the structural features of web document in contest with text watermarking for embedding watermark into html file. Extraction process needs some extra processing for retrieving watermark by authorized user. We found that there were no any general web document watermarking technique will provide requirement as fidelity, flexibility and robustness.

In this paper, we proposed the web document watermarking based on watermarking technique based on cryptographic Unicode (WTCU) for improving imperceptibility, flexibility and robustness of watermark. WTCU consist of watermark generation, embedding, detection, and extraction processes.

2. LITERATURE SURVEY

Brassil et al. [4] proposed the technique in which adjustment in text document layout information but it will affect the display effect of document and also had reduced capacity of watermark for attack. Mercan Topokara et al. [5] proposed the technique in which it carries out change the voice of text document. Because of it will alter the original content of the document. Katzenbeisser et al. [6] proposed Space-Tab Coding (STC) in which watermark was embedded as space and tab into web document. This technique doesn't affect the display of web document. But in this technique watermark detection process not succeed because of loss in any added space or tab in web document. It also has low watermark capacity because it uses only two tags to code the watermark. Zhao and Lu [7] proposed the technique in which it alters the case of HTML tags in HTML 4.01 or its previous version. This technique included the embedded watermark and also prevents the document size. But there are two negative aspects one is that latest W3C standards spotted out next version of HTML tags which are case sensitive and must be in lowercase. Other is that if any one tag is loss will cause whole watermark unrecoverable. John [8] produced the technique which is based on order of attribute pair. Embedding and detection process is based on order of attribute pair. It avoids the increase in document size and it also does not affect on web document display effect. If any tag loss then entire watermark become unrecoverable. Again it also required sufficient pair of tags in document for some watermark capacity. I-Shi Lee and Tsai [9] proposed technique included special space codes (SSC) to be encoded as watermark in html text. The codes are ASCII 0x20 or &#32 and &#160 emerge as white space in document. This technique doesn't impact on display effects and also condenses the coding length. But this technique has poor robustness because single space loss becomes reason for watermark detection collapse. Threshold-based watermarking is used to enhance the robustness of embedded watermark. Still these techniques mainly spotlight on image watermarking it was not available for web documents. Daojing Li and Bo Zhang [11] proposed dual watermarking scheme based on threshold cryptography (DWTC) for web document which give invisibleness and robustness of watermark. This technique involved three stages generation, embedding, and detection. In embedding process involves the vertical layers characteristics of web document which doesn't affect on display of web document. If any of watermark slices missing then it refers the backup watermark save at head of previous layer. Due to above different changes in previous schemes it will improve the robustness & invisibleness of document. The embedding process in both vertical and horizontal direction for two watermark segment in it can be viewed as dual digital watermarking scheme.

3. PROPOSED TECHNIQUE

3.1 Watermark Generation and Insertion in WTCU

Plain text, password, and size for watermark segment are the input to generation algorithm. Plain text is the data to be embedded into web document. Password is the applied for the providing authentication to system. Size specifies the length of watermark segments it's given manually or system may take default size.

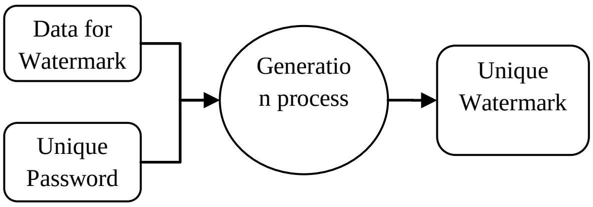


Fig 1: Generation Process of WTCU

Html tag coding table shown above is used to convert hexadecimal text to tag structure. Each hex digit is associated with certain tag of html coding structure.

Table 1: Html Tag Coding

Hex Digit	Html tag	Hex digit	Html code
0	<a></a>	8	<em></em>
1	<i></i>	9	<sup></sup>
2	<tr></tr>	A/a	<big></big>
3	<td></td>	B/b	<tt></tt>
4	<u></u>	C/c	<s></s>
5	<b></b>	D/d	<sub></sub>
6	<h1></h1>	E/e	<var>,/var>
7	<th></th>	F/f	<bdo></bdo>

Generation process describes in algorithm. For converting previously processed data into recommended form for web document, generation algorithm1.1 uses the html tag coding table 1.

1.1 Algorithm for Generation

Input: Password P, Size S Plain Text Pt
Output: Watermark W

```
i ← 0, j ← 0
Sb ← encrypt(P, Pt)
while (Sb <> null) {
            Sbb ← sb.substring(s)
            i←i+s
            }
while (Sbb <> null) {
        ss ← Sbb.substring(j,3)
        j←j+3
        st ← getunicode(ss)
            sd←gethexdigit(st)
            W←createtag() }
```

3.2 Watermark insertion in WTCU

Insertion can be done manually or system can also generate blank html file so that you can code that file according to your use.

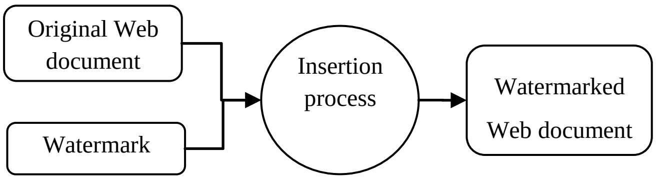


Fig 2: Insertion Process of WTCU

3.3 Watermark detection in WTCU

The detection process input is web document and the output is watermark but in encrypted format. Web document may possibly enclose watermark or not it will detect in this process. Following algorithm 1.2 shows the detection process.

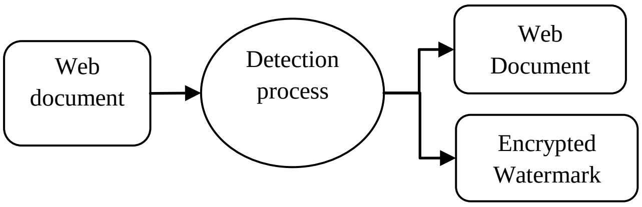


Fig 3: Detection Process of WTCU

1.2 Algorithm for Detection

Input: Web Document WD

Output: Watermark W

```
While (WD <> null) {
A ← readline()
If (a == "></") {
b←getindex("<")
c ← b+7
while(b<llength){
r←substring(b,c)
d←gettagtounicode(r) }
If (d=null) {

g←b+9
while(b<llength){
r←substring(b,c)
h←gettagtounicode(r) } }
If (h=null) {
g←b+11
while(b<llength){
r←substring(b,c)
k←gettagtounicode(r) } }
else :web document does not contained
watermark }
W ← merge(d,h,k)
}
```

3.4 Watermark Extraction in WTCU

Extraction process is done by some authorized person only or the web document owner. Password and encrypted watermark which was computed in detection process are the input to the system and it will provide us the original plain text that was applied to web document as ownership information or details. Recovery process fails if we applied wrong password so recovery process is totally depend on document owner. In such case this will protect the ownership details. Extraction process is totally depends on the document owner so that the ownership of document is highly secured.

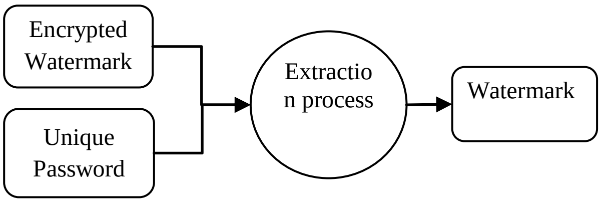
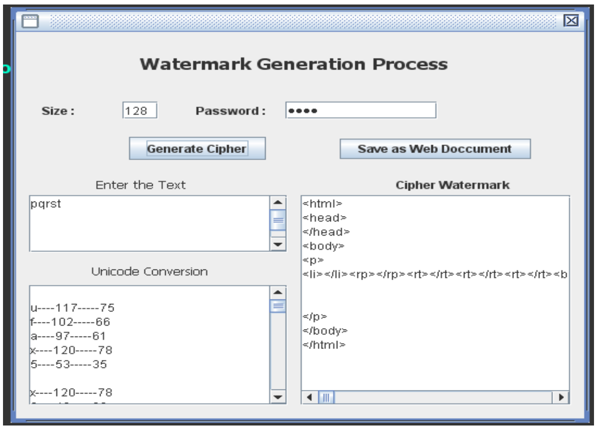


Fig 4: Extraction Process of WTCU

4. IMPLEMENTATIONS AND RESULT

Clear illustration of WTCU system is given. Figure 5 shows the generation process in which inputs are Text to be watermark and password and size. Generation process is done and generates the empty html file enclosed watermark with the html file. Following figure shows size box which is mentioned for creating segments of the ownership secure data, while password is for protecting document and that is known to document owner.



Size : 128

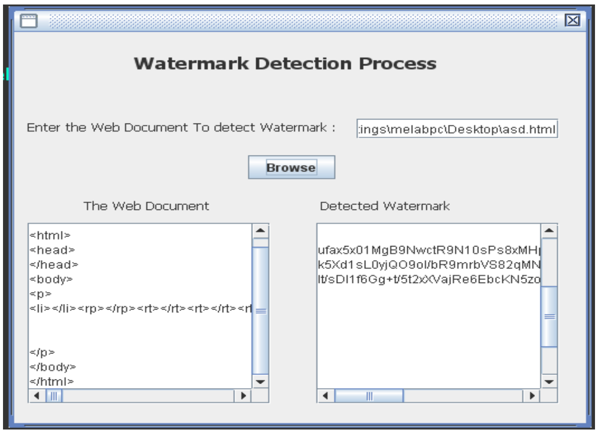
Password : ●●●●

Enter the Text: pqrst

```
u----117-----75
f----102-----66
a----97-----61
x----120-----78
5----53-----35

x----120-----78
```

```
<html>
<head>
</head>
<body>
<p>
<li></li><rp></rp><rt></rt><rt></rt><rt></rt><b


</p>
</body>
</html>
```

Fig 5: Watermark Generation Process of WTCU


Watermark Detection Process
Enter the Web Document To detect Watermark :
Browse
The Web Document
Detected Watermark

Enter the Web Document To detect Watermark : :ings\melabpc\Desktop\asd.html

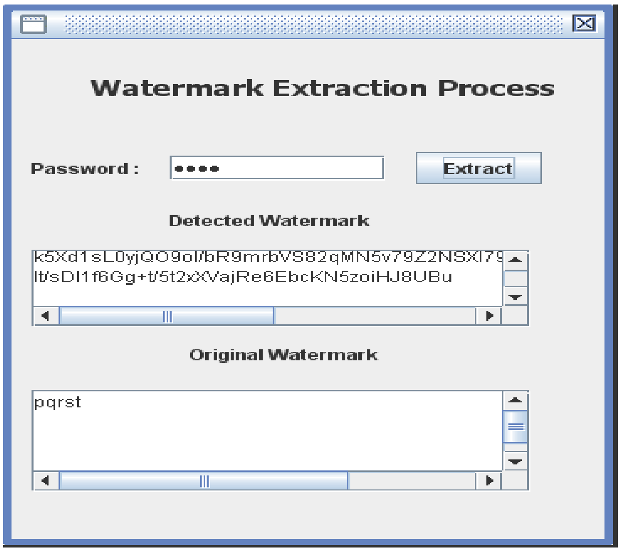
```
<html>
<head>
</head>
<body>
<p>
<li></li><rp></rp><rt></rt><rt></rt><rt


</p>
</body>
</html>
```

```
ufax5x01MgB9NwctR9N10sPs8xMH
k5Xd1sL0yjQO9ol/bR9mrbVS82qMN
lt/sDI1f6Gg+t/5t2xXVajRe6EbcKN5zo
```

Fig 6: Watermark Detection Process of WTCU

Figure 6 shows the detection process of WTCU is execute by any user using this system. WTCU detection process's goal is to determine that whether any document encloses watermark or not. If it has watermark the detection process is executed and result the watermark but in encrypted form.

Figure 7 show the extraction process of WTCU. This process performs the authentication and recovery of watermark. This process has input as encrypted watermark which was computed in detection process and the password. This process is executed only by document owner. This will provide the ownership details which is embedded in web document.



Password : ●●●●

```
k5Xd1sL0yjQO9ol/bR9mrbVS82qMN5v79Z2NSXI79
lt/sDI1f6Gg+t/5t2xXVajRe6EbcKN5zoiHJ8UBu
```

Original Watermark: pqrst

Fig 7: Watermark Extraction Process of WTCU

4.1 RESULT ANALYSIS

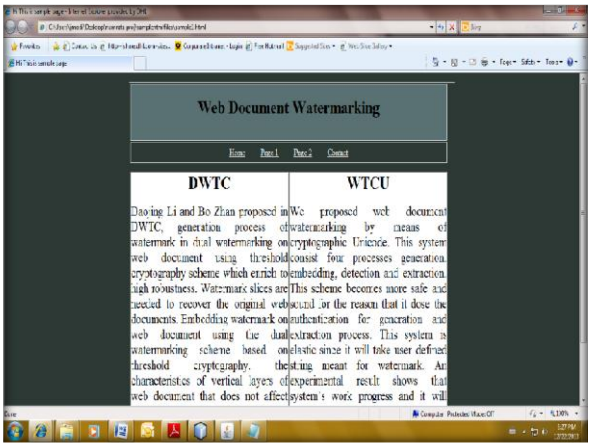
The capacity of water marked text in WTCU is compared with other watermarking schemes like Space-Tab coding (STC). In STC watermark capacity is limited to only two tags to code the watermarks while in case of WTCU have different tags in order to code the watermark. After comparison of STC and WTCU in accordance with capacity WTCU is preferable.

SSC and WTCU both techniques are compared on basis of robustness, SSC emerges white space in document using Special Space codes. Single space loss in document due some attack it cannot recover the watermark.

WTCU is robust because any alteration in document does not affect the extraction process. In comparison with flexibility issue WTCU and DWTC is embedding the tags after span tag only. In WTCU we can embed the tag anywhere in the document so in accordance with flexibility WTCU is preferable.

In case of security we compare the WTCU with DWTC; WTCU involves the encryption of text to be watermark and password protected so WTCU is preferable. Now we want to focus on robustness issue in which SSC, STC are not robust because when we modify the web document then these two are fails in detection process.

In DWTC has the backup tag in head of layers to recover whole watermark. In WTCU password is sufficient for the recovery of the whole watermark it also preserves the authenticity of document owner.



Web Document Watermarking

Home Page 1 Page 2 Contact

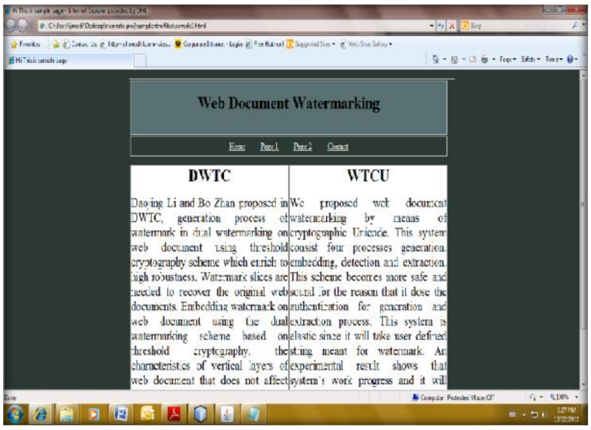
DWTC

Daojing Li and Bo Zhan proposed in DWTC, generation process of watermark in dual watermarking on web document using threshold cryptography scheme which enrich to high robustness. Watermark slices are needed to recover the original web documents. Embedding watermark on web document using the dual watermarking scheme based on threshold cryptography. the characteristics of vertical layers of web document that does not affect

WTCU

We proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes more safe and sound for the reason that it dose the authentication for generation and extraction process. This system is elastic since it will take user defined string meant for watermark. An experimental result shows that system's work progress and it will

Figure 8: Normal Web Document



Web Document Watermarking

Home Page 1 Page 2 Contact

DWTC

Daojing Li and Bo Zhan proposed in DWTC, generation process of watermark in dual watermarking on web document using threshold cryptography scheme which enrich to high robustness. Watermark slices are needed to recover the original web documents. Embedding watermark on web document using the dual watermarking scheme based on threshold cryptography. the characteristics of vertical layers of web document that does not affect

WTCU

We proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes more safe and sound for the reason that it dose the authentication for generation and extraction process. This system is elastic since it will take user defined string meant for watermark. An experimental result shows that system's work progress and it will

Figure 9: Result web page obtained by WTCU

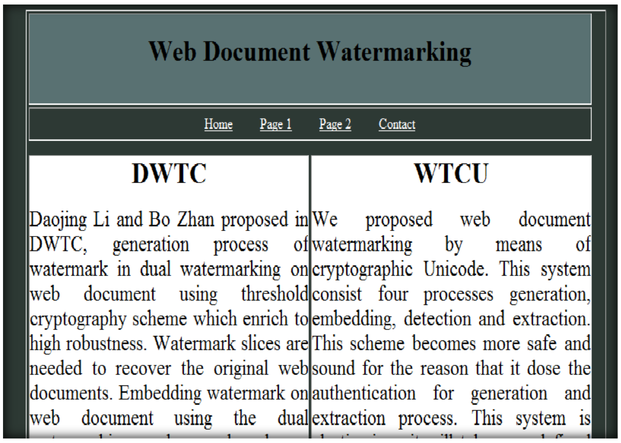
4.2 SECURITY ANALYSIS

There are two kinds of attack on watermark they are additive attack, distortive attack. In additive attack attacker insert his own watermark in watermarked document using his own technique so document contains two watermarks.

In WTCU can easily recover the watermark after applying this attack on web document because it is password protected and size is also associated with it. In distortive attack attacker made alteration or copying to the web documents.

WTCU is HTML tag based technique that's why it is hidden when the HTML is generated and when it is copied or altered the tags are embedded with the text. So that WTCU technique is more robust. WTCU provides the efficient way of embedding and recovery of the watermark without any long process thus it is highly adoptable.

Now additive attack is applied on to web document by adding one more hyper link tag with some additional space. The hyperlink tag is added for **About** and the space is added by using **&nbsp;** that both are shown in figure 11 web document attack. Then this web document is given to the detection process of WTCU and watermarking is detected from that web document. Watermark detection process is executed properly in attacked web document is shown in figure 12.



Web Document Watermarking

Home Page 1 Page 2 Contact

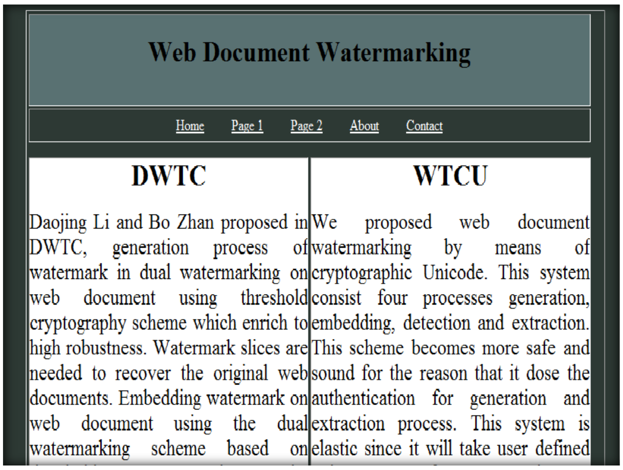
DWTC

Daojing Li and Bo Zhan proposed in DWTC, generation process of watermark in dual watermarking on web document using threshold cryptography scheme which enrich to high robustness. Watermark slices are needed to recover the original web documents. Embedding watermark on web document using the dual

WTCU

We proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes more safe and sound for the reason that it dose the authentication for generation and extraction process. This system is

Figure 10: Sample Web documet with watermark



Web Document Watermarking

Home Page 1 Page 2 About Contact

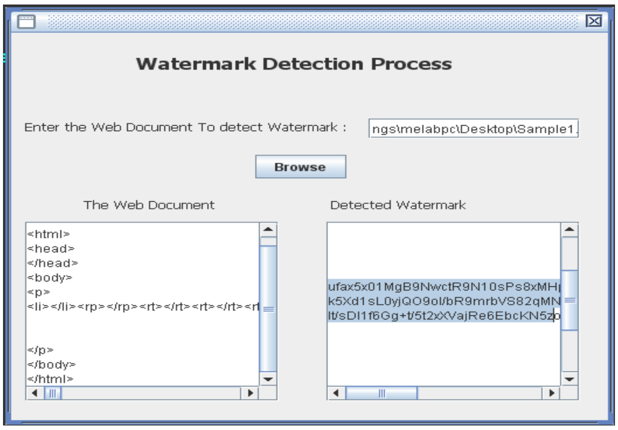
DWTC

Daojing Li and Bo Zhan proposed in DWTC, generation process of watermark in dual watermarking on web document using threshold cryptography scheme which enrich to high robustness. Watermark slices are needed to recover the original web documents. Embedding watermark on web document using the dual watermarking scheme based on

WTCU

We proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes more safe and sound for the reason that it dose the authentication for generation and extraction process. This system is elastic since it will take user defined

Fig 11: Web Document with additive attack as one hyperlink "ABOUT" is inserted to the document



Watermark Detection Process

Enter the Web Document To detect Watermark : ngs\melabpc\Desktop\Sample1

The Web Document

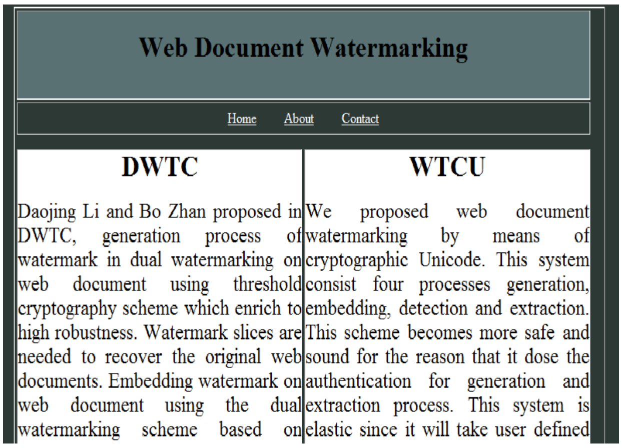
```
<html>
<head>
</head>
<body>
<p>
<li></li><rp></rp><rt></rt><rt></rt><r


</p>
</body>
</html>
```

Detected Watermark

```
ufax5x01MgB9NwctR9N10sPs8xMH
k5Xd1sL0yjQO9ol/bR9mrbVS82qMN
lt/sDl1f6Gg+t/5t2xXVajRe6EbcKN5z
```

Fig 12: Watermark detection on Additive attacked Sample web document

Distortive attack is applied on the same sample web document source code shown in figure 13 in which delete the two hyperlink tags for page 1 and page 2 with its space is deleted. Resulted web document which does not contained the hyperlinks for page 1 and page 2. Watermark detection process is executed properly in attacked web document is shown in figure 14.



Web Document Watermarking

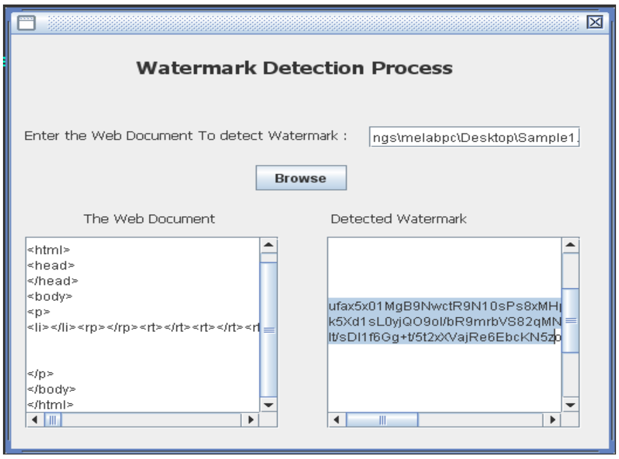
DWTC

Daojing Li and Bo Zhan proposed in DWTC, generation process of watermark in dual watermarking on web document using threshold cryptography scheme which enrich to high robustness. Watermark slices are needed to recover the original web documents. Embedding watermark on web document using the dual watermarking scheme based on

WTCU

We proposed web document watermarking by means of cryptographic Unicode. This system consist four processes generation, embedding, detection and extraction. This scheme becomes more safe and sound for the reason that it dose the authentication for generation and extraction process. This system is elastic since it will take user defined

Fig 13: Distrotive Attacked Sample web document



Watermark Detection Process

Enter the Web Document To detect Watermark : ngs\melabpc\Desktop\Sample1.

The Web Document

```
<html>
<head>
</head>
<body>
<p>
<li></li><rp></rp><rt></rt><rt></rt><rt


</p>
</body>
</html>
```

Detected Watermark

```
ufax5x01MgB9NwctR9N10sPs8xMH
k5Xd1sL0yjQO9ol/bR9mrbVS82qMN
lt/sDl1f6Gg+t/5t2xXVajRe6EbcKN5zo
```

Fig 14: Watermark detection on Distortive attacked Sample web document

Table 5: Comparison Between Different web documents watermark embedding and extraction and its correlation

Cover Web Document	Embedded Watermark Text	Resultant Web Document	Attack Applied	Extracted Watermark Text	Correlation between watermark and extracted watermark
S1	xyz	S 11	No attack	xyz	1
S2	abcdef	S12	Insertion	abcdef	1
S3	pqrst	S21	Deletion	pqrst	1

Following formula is used for calculating correlation between embedded watermark text and extracted watermark text. Correlation is exactly 1 that means watermarked embedded and extracted are exactly same.

$$r = \frac{n\,(\sum W_{em}\,W_{et}) - (\sum W_{em})\,(\sum W_{et})}{\sqrt{[(n\,\sum W_{em}{}^{2} - (\sum W_{em})^{2}\,)\,(n\,\sum W_{et}{}^{2} - \,(\sum W_{et})^{2}]}}$$

Where,
n= Number of web documents
W_{em}= Persentage value of Embedded Watermark Text
W_{et} = Persentage value of Extracted Wateramrk Text

5. CONCLUSION

Tremendous development and popularity of internet, data over the internet should be protected. We proposed the watermarking technique based on cryptographic Unicode (WTCU) for the web documents having better imperceptibility, robustness, flexibility and security. Experimental results are shown for successful implementations of system. WTCU embeds the secret watermark as copyright information by document owner into web document with authentication. WTCU technique applied on different web document in order to analyses the sizes of web documents changes and results are shown above. Watermark capacity, flexibility and authenticity are improved due to cryptographic Unicode. Generally this system is appropriate and effectual for all types' web documents protection. Correlation between embedded watermark text and extracted watermark text is determined by formula andits exactly one.

.

6. REFERENCES

[1] Zunera Jalil and Anwar M. Mirza, "A Review of Digital Watermarking Techniques for Text Documents", International Conference on Information and Multimedia Technology 2009.

[2] Yanqun Zhang, "Digital Watermarking Technology: A Review", IEEE International Conference on Future Computer and Communication, 2009.

[3] N. P. Sheppard et al, "On Multiple Watermarking", ACM Workshop on Multimedia and Security 2001, 1-3.

[4] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," IEEE L. Select. Areas Commun, vol. 13, pp. 1495-1504, Oct. 1995.

[5] Mercan Topkara, Umut Topkara, and Mikhail Atallah, "Words are not enough : Sentence level natural language watermarking," In Proceedings of the 41h ACM international workshop on Contents protection and security, pp. 37-45, Santa Barbara, California, USA, October 2006.

[6] S. Katzenbeisser, A. P Petitcolas, "Information Hiding Techniques for steganography and Digital watermarking," Boston, Artech House, 2000.

[7] Zhao, Q. and Lu, H. 2005, "A PCA-based watermarking scheme for tamper- proof of web Pages," Pattern Recognition 38 (2005), pp.1321-1323.

[8] C John, "Hiding binary data in HTML documents", [Online], Available: http://www.codeproject.com/csharplsteganodotnet 13 .asp, May, 2007.

[9] I-Shi Lee, Wen-Hsiang Tsai, "Secret Communication through Web Pages Using Special Space Codes in HTML Files," International Journal of Applied Science and Engineering. 2008.6, 2:141-149.

[10] C. C. Chang, C.Y. Lin and CS. Tseng, "Secret image hiding and sharing based on the (t, n)-threshold", Fundam. Inf. 76 (4) (2007), pp. 399-411.

[11] Daojing Li, Bo Zhang, "DWTC: A Dual Watermarking Scheme Based on Threshold Cryptography for Web Document," International Conference on Computer Application and System Modeling (ICCASM 2010).

[12] Nighat Mir, Sayed Afaq Hussain , "Web Page Watermarking: XML files using Synonyms and Acronyms" , World Academy of Science, Engineering and Technology 73 2011

[13] http://unicode.org/