Abstract—Nowadays, the security of web pages becomes very important. Zhao and Lu had proposed a PCA-based watermarking scheme for web pages; their scheme is an effective tool for tamper-proof of web pages. However, the PCA algorithm is time-consuming, especially when the size of web page is too large. In this paper, we proposed two novel fragile web page watermarking schemes, in which simple principal component analysis and random projections are used to generate watermarks of the web page. Experimental results show out that the proposed scheme is much faster than the PCA algorithm, and, at the same time, our schemes make some improvements on Zhao and Lu’s scheme.

Keywords—web page; watermark; PCA; simple principal component analysis; random projections; 

I. INTRODUCTION

With the rapid development of network technology, web pages published on the Internet often suffer from attacks. It is desired to have an automatic authentication scheme to check the integrity of concerned web pages. Fragile digital watermarking provides a convenient tool for authentication of digital information.

Recently, many fragile watermarking schemes for tamper-proof of images are presented [1, 2]. But there are very few studies on web page authentication using watermark technique. Luo and Sui proposed a steganography method, which hides secret information in hypertext by modifying the written states of the markup letters [3]. Katzenbeisser et al. proposed a watermark-based method by adding spaces and tags into the source code of web pages [4], but this method has its drawback of expanding file size. To solve the problem, Zhao and Lu proposed a PCA-based web page watermarking scheme [5, 6], where watermark is generated by principle component analysis and embedded by upper and lower cases of letters in HTML tags. PCA-based method can be an effective tool for tamper-proof of web pages, but principle component analysis is time-consuming, and the proposed PCA-based scheme may cause the malfunction of some hyperlink of web pages, which will prevent their scheme from practical use.

In this paper, we propose two new fragile watermark methods for web pages. We use simple principal component analysis (SPCA) [7] and random projections (RP) [8] respectively to generate our watermark and embed them into the tags of the source code of web pages. While embedding, we omit the double-quote parts of HTML tags, thus the proposed scheme can solve the problem of malfunctions of HTML tags. Experimental results show out that the proposed scheme is much faster than the PCA algorithm, especially for large web pages. Thus, the proposed scheme can be a practical tool for tamper-proof of web pages.

The rest of this paper is organized as follows: section 2 presents the drawbacks of PCA-based scheme and gives a simple description on SPCA and RP. The proposed watermarking schemes and its security analysis are given in the following section 3. In the end, you will see the experimental results and conclusions of this paper.

II. THE RELATED WORK

In this section, we will give a simple description on the drawbacks of PCA-based scheme, SPCA and RP.

A. The drawbacks of PCA-based scheme

The PCA-based watermarking scheme is proposed by Zhao and Lu. They use PCA to generate watermark of web page. While embedding, watermark is embedded into the tags of HTML, because of the case-insensitive of HTML tags. If a certain bit of the watermark is ‘0’, then the corresponding character is converted into lower case, otherwise, it is converted into upper case. Obviously, the scheme will not increase the file size of web pages.

The drawbacks of the PCA-based scheme are obvious:

1) PCA algorithm is time-consuming, which is shown in our experiment.
(2) Malfunction of some hyperlink. For some web pages, this problem will not occur, but for other web pages, will cause the hyperlink doesn’t work: 
the tag, “<a href="http://www.opensource.org/licenses/W3C.php" >” is transformed into “<a href="http://www.opensource.org/licenses/W3C.php" >”.

B. SPCA
The simple PCA (SPCA) method produces approximate solutions to PCA without the need for calculating a variance-covariance matrix and then diagonalizing it (as PCA does) and does not depend on learning parameters (as neural networks does) [7, 9]. Most importantly, SPCA is faster than other existing techniques and easy to implement.

C. RP
The random projections (RP) method provides a computationally feasible method for reducing the dimensionality of the data so that the mutual similarities between the data vectors are approximately preserved [8, 10, 11].

In the random projections method, the original data vector, denoted by $\mathbf{n} \in \mathbb{R}^N$, is multiplied by a random matrix $\mathbf{A}$. The projection
$$\mathbf{x} = \mathbf{A}\mathbf{n}$$
results in a reduced-dimensional vector $\mathbf{x} \in \mathbb{R}^d$. The matrix $\mathbf{A}$ consists of random values and the Euclidean length of each column has been normalized to unity.

III. THE PROPOSED SCHEMES
In this section, we will give a detail description on the SPCA-based watermarking scheme and the RP-based scheme, and meanwhile points out our improvements to the PCA-based scheme.

A. The watermark generation processes

Usually, the contents of the HTML codes are easier to be suffered from tampering than the tags of the HTML codes, and the most important message of the HTML codes are in its content parts and the double-quotes parts of its tags. So, when we extract message data from the web page to be protected, we can simply omit the non-double-quotes parts in HTML tags, which will save a lot of time for the watermark generation process. Then the characters at the range of protection in each line are changed into their corresponding decimal value according to its index in the code chart adopted by the document, ASCII or UNICODE. And we get a message data matrix $\mathbf{D}$ of the web page. A key $\mathbf{K}$ is also needed in the watermark generating. It can be either an image or a random sequence. Assume it is a matrix; we use the operation of convolution to diffuse the data of web page and combine the key by the following equation.

$$\mathbf{V} = \mathbf{D} \otimes \mathbf{K}$$

For SPCA-based scheme, we apply SPCA to matrix $\mathbf{V}$ to get the principle components of $\mathbf{V}$. Then the first $k$ principle components are changed into their binary form denoted by $\mathbf{W}_s$, which is the watermark of the web page.

For RP-based scheme, RP is applied to matrix $\mathbf{V}$, thus the dimension of $\mathbf{V}$ is reduced to $d$. The newly mapped matrix is then changed into its binary form denoted by $\mathbf{W}_r$, which is the watermark of the web page.

B. The watermark embedding and validating processes
We only embed the watermark into the tags of the HTML codes, making use of the case-insensitiveness of the HTML tags. And while embedding, we omit the double-quotes parts of the tags, so the proposed scheme can also solve the problem of malfunction of hyperlinks in tags. Then, if a certain bit of watermark is ‘0’, the corresponding character in tags is then converted into lower case; otherwise it is converted into upper case.

In Zhao and Lu’s scheme, they use cyclic embedding to embed watermark: if the length of watermark is less than the length of tags, they embed the watermark redundancy. We think that it is not necessary to do like that, because it is time-consuming and is of no use to security. So, while embedding, if length of tags is larger than the length of watermark, we don’t convert the remaining characters in tags any more.

The validating process is easy to understand. We firstly get a message data from the web page to be validated, and, at the same time, extract a watermark $\mathbf{W}'$.

For SPCA-based scheme, we use the message data and secret key to generate a new watermark $\mathbf{W}_s$. If $\mathbf{W}_r$ is equal to $\mathbf{W}'$, then we think that the web page is not tampered, otherwise, the web page is tampered.

For RP-based scheme, we use original data to generate a new watermark $\mathbf{W}_r$, if $\mathbf{W}_s$ is equal to $\mathbf{W}'$, then we think that the web page is not tampered, otherwise, the web page is tampered.
IV. SECURITY ISSUES OF THE PROPOSED SCHEMES

In general, there are three kinds of tampering for the watermarked web pages [5]:

(1) Web contents are tampered, but no new watermark is generated. (2) Web contents are tampered and the watermark is destroyed. (3) Web contents are tampered and a new watermark is generated.

For the first problem, because the watermark we embed is fragile, any change of the web content we protect will cause a different watermark in validating process, thus, the tamper can be detected. For the second problem, it is easier to determine whether the web page is tampered or not.

For the last one, we can focus on one question: whether the attacker can work out the key. Given a watermarked web pages, we can easily get its codes. So if the watermark is totally embedded, the attacker can get the message data and the watermark. But the attacker cannot get the secret key, which is guaranteed by the convolution operation we used.

V. EXPERIMENTAL RESULTS AND ANALYSIS

We randomly collected about 120 web pages in the Internet as our dataset. We use MATLAB to implement the proposed scheme and conduct a series of experiments on the effectiveness and time-efficiency of the proposed schemes. All of the experiments are based on the same machine configuration.

A. Effectiveness

As mentioned above, we use the SPCA and RP separately to generate a fragile watermark for a web page, and then embed into the original web page. Whenever the web page is tampered, the watermark in it will be destroyed or inconsistent with the content, thus the tamper can be found. This can be illustrated below.

The HTML codes of the original web page can be seen in figure 1, the SPCA-based watermarked HTML codes of is shown in figure 2. The RP-based watermarked HTML codes of is shown in figure 3. We can see that the watermarks are embedded into the tags of the web page and the file size does not increase. Figure 4 and figure 5 are the web pages that users can see in web browser before and after the web page is watermarked. We can see that there is no visual difference after watermark embedding. The tamper in figure 6 can be found by the validating process.

B. Efficiency

We can see from figure 7 and figure 8 that the two proposed schemes have much less time spent, it is all because of the less time spent of less message data got only from web page’s content parts and the double-quotes parts of its tags and the SPCA and RP we used, which is faster than PCA [7,11]. We should point out that the embedding time here and after includes watermark generating and embedding time. We have implemented the PCA-based scheme. The time spent of the PCA-based scheme. The comparisons between PCA, SPCA and RP can be seen in figure 7 and figure 8. We can see that the PCA-based scheme spends so much time both in embedding and validating process. So, the proposed scheme can be a practical tool for tamper-proof of web pages.
We also make comparisons between the embedding times of the two schemes and the convolution time in them. We can see from figure 9 that the convolution time consists most of the embedding time. So, the time spent of the two proposed schemes can be improved if we substitute more time-efficient methods such as confusion for the convolution operation, on condition that they do not lower the security of the schemes. We can also confuse the watermarks before embedding, which will make it more difficult to get the watermark from the watermarked web pages, thus strengthen the security of the schemes.

VI. CONCLUSIONS

In this paper, we propose two efficient web page watermarking schemes. The proposed schemes choose the double-quotes parts of HTML tags and the contents parts of HTML for protection, and use SPCA and RP to generate the watermark of web page separately. While embedding the watermark, we only choose the non-double-quotes parts of HTML tags and make use of the case-insensitivity of HTML tags. Experimental results show out that the proposed scheme not only can protect the integrity of web pages, but also reduces the time spent to a large extent, in comparison with the PCA-based scheme. So, the proposed schemes can be an effective and efficient tool for tamper-proof of web pages.

ACKNOWLEDGEMENT

This work was supported by NSFC under project No.60573033 and Shanghai Science and Technology Development Funds (No.08JG05002).

REFERENCES