To Design an Algorithm for Text Watermarking

Pradeep Kaur* & Pankaj Bambri**

**Assistant Professor, Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, New Delhi.

Abstract—Nowadays the wide use of various communication technologies and internet, it has become extremely easy to reproduce, communicate, and distribute digital contents. So there is need to authenticate the data and copyright protection issues resolve which arisen. Textual way of communication is the most widely used medium for travelling the data over the Internet. In this thesis, I have proposed a zero-watermarking approach towards text watermarking; propose a zero text watermarking algorithm based on occurrence frequency of vowel ASCII characters and articles for copyright protection of plain text. Uses of watermark for the watermark embedding process are smaller in length. The embedding algorithm makes use of frequency vowel characters and articles to generate a specialized author key. The extraction algorithm uses this key to extract watermark, hence identify the original copyright owner. Experimental results illustrate the effectiveness of the proposed algorithm on text documents encountering various tempering attacks like insertion and deletion and the results are also compared with the recent work on text watermarking.

Keywords—Authentication; Copyright Protection; Embedding Algorithm; Extraction Algorithm; Text Watermarking; Watermarking.

Abbreviations—Average Frequency of Articles (AFA); Copyright Authority (CA); Maximum Occurring Vowels (MOV).

I. INTRODUCTION

To provide the security to the digital systems has gain remarkable importance in contemporary era. World Wide Web has helped us in our day today life daily life for the movement of different forms of data like papers, emails, images, articles, videos, websites, and opinion blogs. Information over the electronic media mostly textual based which required the security for the text document, because text is the main issue for the creator. Text is the most important and core part of legal papers, journals, and reports which needed its security that has been critically ignored. The intimidation of internet are similar to re-distribution and prohibited copying of copyrighted objects, the copyright violation and different sources of copying require a security basically for the text part of documents.

1.1. Watermarking

Watermarking is used for hiding the data such as a audio, video, digital images, or a text and it is a branch of information hiding. It is a technique which is used for the embedding a given data in the form of secure data which can be any data like image, text or anything. The embedded watermark information is protective and not apparent by any human vision. In a watermark there is an identification code that can be visible or invisible which is eternally embedded in the information, for the transmission of hidden message. Watermark remains there within the in the data even that the decryption process. Watermarking usually embeds the watermark data which is unique for the creator and it provides copyright protection to the watermark information which is secured. That watermark is use later on for the identification of original copyright holder by certifying authorities.

1.1.1. Certifying Authority (CA)

Certifying Authority is documented organization or an official administration which acts as the impartial facilitator between all stake holders. Certifying authorities are the same as a registration authority in which the data is to be register with the creator’s name. For the protection of data or information in this electronic world every writer should embeds there watermark within their original information. After the process of watermark embedding the key is generated, this key is registered with the watermark by the original creator to the certifying Authority. So the registered watermark and the watermark key is imprinted by the certifying Authority on the bases of writer. After the embedding and registered process, it ensure logical that the property exactly is the possessions of the creator. The key, extracted watermark and instance can confirm claims towards the owner of the text.
1.2. Types of Digital Watermarking

The digital watermarking is classified into two parts the first is “visible watermarking” and the second is “invisible watermarking” [Jalil & Mirza, 2]. In perceptible or visible watermarking, watermarks are embedding process is done in such a way that the embedded watermark is visible to the user whenever the content is observed. The imperceptible or invisible watermarks cannot be perceived but recovering of watermark is achievable only by the suitable decoding algorithm or by the extraction process. The Imperceptible watermarks are more strong and robust than the perceptible watermarking. Further the watermarking can be robust/strong or fragile/delicate. In a robust watermarking, this technique does not affect the watermark even after the content is modified by the attacker in any way. But in the case of the fragile watermarking technique is subtle technique in which watermark gets damaged when the contents of watermark is altered or tampered.

1.3. Text Watermarking

Over the internet the most widely used medium for the communication is text. There are various components of text files such as books, websites, papers, articles, documents and many more, which required more security and safety from the various violators of copyright. From the past many years there is numerous algorithm for watermarking are designed for videos, audios, and digital images, but for the textual part the watermarking algorithms are not enough and mostly are unsuccessful.

Text watermarking is the method of embedding an identical watermark within the content for the reason to protect the content from unauthorized copying and from other copyright violation. The whole procedure in which to embedding the watermark within the content and extract the watermark from the content is to verify the original copyright author of that content or data is known as the digital text watermarking. The principles used by an image, video and audio watermarking is same as the principles of text watermarking.

For various tempering attacks the watermark should stay durable and they are untraceable to any other third party except the original creator of the text, at the same time the watermark can be simply and completely reproducible automatically through the watermark extraction algorithm. The major unease behind the text watermarking in the plain text is that it consists of less redundant data as comparative to other watermarking like in digital images, audio, or video which can be using in covert communication, like occur in steganography.

1.4. Contribution towards Thesis

A number of developments in the field of text watermarking have been made till knows. This thesis contributes towards text watermarking with the utility of text constituents like vowel characters and articles. Using these watermarks remains resistant to tampering attacks. The main contributions of this thesis are:

- A novel zero-watermarking approach has been adopted towards text watermarking.
- Watermarking and encryption are used for robust text watermarking results.
- The proposed technique provides optimal results using vowel characters and articles
- The algorithms are tested under the insertion and deletion tampering attacks
- Results are compared with the previous algorithm based on prepositions.

II. Related Work

In this image-based technique of text watermarking includes embedding the binary watermark in text image. According to this method embed the watermark using text document image. Usually the text documents are complicated to watermark for the reason of their sensitiveness, simplicity and small ability for watermark embedding. The very first step in text watermarking is treating the text as image. Watermark is embedded with the arrangement and emergence of the text image. Brassil et al., [5, 6] anticipated a small number of technique to watermark the text data with the use of text image, which is the line-shift coding algorithm, in which it modify the document image; the modification is done by shifting lines upward or downward in left or right according to the binary indicator watermark for the insertion.

The next method of image based is the word-shift coding algorithms which shift the words within the text in horizontal way therefore it increasing spaces to embed the watermark. This algorithm functions in both blind and non blind modes. After word shift coding algorithm the next method under the image based technique is the feature coding algorithm in which a little modification in the features like in the pixel characters, the span of the end line in the character to encode watermark bit into the text.

Syntactic methods is a methods in which includes the syntactic text structure and it used for embedding the watermark. Generally text is design through the characters, words, and then sentences. Every sentence has a diverse syntactic structure. To Applying syntactic transformation on the text structures for embedding the watermark has been another approach in the direction of text watermarking in the history. The natural language watermarking algorithm is proposed by the Meral et al., [4] and they perform the morpho-syntactic alterations to the textual data, which is shown in a figure 1.
In semantic technique in the watermark embedding process is done by utility of text semantics. There are number of algorithms are proposed which are based upon these methods. This watermarking method concentrates on the semantic structure of text for embedding a watermark [Topkara et al., 1; Khan Asifullah, 3]. The text contents, nouns, verbs, words and those spellings, grammar rules, sentence structure etc. have been oppressed for insertion of watermark within the text but not any of these confirm the flexibility and corrupt the worth of the text at the huge level.

This technique of text watermarking is recently used by the watermarking methods, in structural method the text structures are used to embed the watermarks. By using this technique, text is not altered while watermark is embedded in to the text.

Structural scheme of a text watermarking is another research field of watermarking. With using this technique on double letters have a various limitations. Generally the watermarking algorithms those are generated in the past for the simple text is to insert the watermark within the original text data which gives a text quality, significance, and value degradation as a result. So, it’s a new approach which proposed as a zero-watermarking method that includes no alteration in the original text document for embedding the watermark, relatively the component of text are used for the generation of unique watermark key to give protection to the text. Here I encompass the fundamental components of text are used like vowels and articles in for the proposed algorithm.

III. PRESENT WORK

3.1. Background of the Proposed Algorithm

The proposed algorithm uses vowel characters to watermark the text document. The original owner of the text generate key using an algorithm which is watermark embedding algorithm. This algorithm is known as zero watermarking algorithms in which text documents remain same when watermarking is done as it generates the author’s key by using properties of the text without changing it. The text document is first analyzed and then articles from the text are identified. Average frequency articles (AFP) are obtained and on that bases create the partition of text. Then count highest occurring vowel characters and makes a list of MOV that is maximum occurring vowel characters list. This list is used to generate the author key of a particular watermark given by the original owner.

The proposed algorithm is a merge of watermarking, and encryption. The original author embeds the copyright information in a text and it generates the watermark key using embedding algorithm, the existence of watermark remains hidden.

The watermarking process involves two stages, watermark embedding and watermark extraction. Watermark embedding is done by the original author and extraction done later by the copyright owner (CA) to prove ownership. The original copyright owner of text inputs a watermark. And unique key is generated using input text. This key is used later for extraction of watermark, whenever a copyright conflict arises in future. In the proposed algorithm, at the time of watermark extraction there is no need of original watermark is needed and there is no alteration in the text watermark. The original owner records there copyrights to the trusted certification authority, that authority take decisions whenever there is any copyright conflicts arises.

3.2. Parameters Settings

These are the parameters used for to check the performance of proposed algorithm. These parameters give the quality of an algorithm which shows the accuracy of watermark used for experiment. Parameters used for experiment are:

3.2.1. Watermark

Watermark should be carefully selected for the robustness against attacks. Experimental results show that the watermark consisting of minimum length is more robust against insertion and deletion attacks. In the proposed algorithm, watermark is restricted to only alphabetic characters and there is no numbers and special characters. Watermark length is smaller in as compared with previous based algorithm’s watermarks.

3.2.2. Accuracy

Accuracy represents how accurately we retrieve watermark from the attacked text. It depends on text length, watermark length, and the quantity of attack, which shows as:

\[ \text{Accuracy} = f (\text{TL}, \text{V}, \text{WL}) \]

Where, TL represents text length measured using sentence, V represents quantity of attack and WL is watermark length.

3.3. Embedding Algorithm

The algorithm in which the watermark is embedding into text is called embedding algorithm. The embedding algorithm logically embeds the watermark in text without making any changes in text document and it generates the author key. Flowchart for embedding algorithm is shown in figure 2.
3.4. Extraction Algorithm

The extraction algorithm is used to extract watermark from the text. It takes a key as input to extract the watermark from the document. Articles and vowels of entire document make the algorithm more resistant against attacks and watermark is still robust after various attacks. The watermark extraction process is shown in the figure 3.

IV. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed algorithm, there is a text samples to perform attacks on it by a different individuals. The characteristics of the original files can be altering by the attacks but the whole theme of the text is remaining same. Whenever the attacker try to ruin the copyrights then they will perform attacks to alter the text and various attack files were differ which is based on attack volume. To examine the tampering attacks on the text file by evaluating the accuracy of retrieved watermarks as well as experiments were performed to check the insertion and deletion attacks on the text files. To insert and delete the data from the text is the most common attacks on text documents.

Further, the proposed algorithm is compared with the previous algorithm which is based upon the prepositions.

To test the effect of tempering on text experiment were conducted to examine the accuracy of retrieved watermarks, to calculate the insertion and deletion attack, and to further noticed the impact of tempering on watermark. And also the proposed algorithm is compared with the previous algorithm. The performance of proposed algorithm is shown in table 1 and table 2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Attack file 1</th>
<th>Attack file 2</th>
<th>Attack file 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watermark Accuracy (%)</td>
<td>77</td>
<td>80.37</td>
<td>81.11</td>
<td>79.49</td>
</tr>
<tr>
<td>Inserted Words</td>
<td>72</td>
<td>68</td>
<td>67</td>
<td>69</td>
</tr>
<tr>
<td>Deleted Words</td>
<td>28</td>
<td>32</td>
<td>33</td>
<td>31</td>
</tr>
</tbody>
</table>

In the below figure 4, it shows the accuracy of watermark w1 which used to in this proposed algorithm. It shows the percentage accuracy of attack files which used watermark 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Attack file 1</th>
<th>Attack file 2</th>
<th>Attack file 3</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watermark Accuracy (%)</td>
<td>97</td>
<td>98</td>
<td>97.51</td>
<td>97.50</td>
</tr>
<tr>
<td>Inserted Words</td>
<td>61</td>
<td>58</td>
<td>62</td>
<td>60.33</td>
</tr>
<tr>
<td>Deleted Words</td>
<td>39</td>
<td>42</td>
<td>38</td>
<td>39.66</td>
</tr>
</tbody>
</table>

In the below figure 5, it shows the accuracy of watermark w2 which used to in this proposed algorithm.
From the below figure 6, it shows the resultant window of MATLAB. This figure shows the watermark which is used in this research and the key that is generated by algorithm.

![Figure 6: Resultant Window (MATLAB) showing Watermark and there Generating Key](image)

From the below figure 7, it shows the resultant window of MATLAB. This figure shows the watermark extraction process in which watermark is extracted from the text file.

![Figure 7: Watermark Extraction Resultant Window](image)

Figure 8 shows the watermark accuracy using watermark 1 with attack file 1.

![Figure 8: Watermark Accuracy using Watermark 1](image)

V. COMPARATIVE RESULTS

The performance of the proposed algorithm is compared with the previous algorithm which is based on prepositions. To facilitate the comparison, I have used three different watermarks and three different text files with varied length. Previous algorithm used a watermark with a longer in length for the watermarking, but I have used a watermark with a smaller in length. Table 3 shows the comparison result of two algorithms.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Previous Algorithm</th>
<th>Proposed algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W1</td>
<td>W2</td>
</tr>
<tr>
<td>Watermark Accuracy</td>
<td>75.92</td>
<td>67.22</td>
</tr>
<tr>
<td>Watermark Length</td>
<td>112</td>
<td>40</td>
</tr>
</tbody>
</table>

VI. CONCLUSION AND FUTURE SCOPE

Digital watermarking is a very effective solution for authentication and copyright protection of digital contents and text documents. Text documents gained very high importance for their security purpose. So, here I have proposed a zero-watermarking algorithm for copyright protection of text documents. The algorithm integrates the occurrence frequency of articles and vowels characters in the text to protect it. The algorithm with a zero-watermarking approach provides a robust solution for text watermarking problem. To check the frequency of occurrence of each vowels character in each text and generate a key using the intrinsic properties of the text. The key which is generated is registered with CA and that key is used when there is any conflict arises in the copyright claims, and then this watermark can be extracted from the digital content to identify the original owner.

I have tested the performance of an algorithm for tampering attacks like insertion and deletion attacks of different texts. I also compared the performance of the algorithm with the previous algorithms. The results show that my algorithm is also more robust even when the watermark length is shorter, as well as they are secure, and efficient with minimal computational requirements. The watermark remains resilient after attacks which make the watermark more efficient and robust.

In future, same algorithm is used for another language which consist of components like articles, prepositions, double letters and any other component.

REFERENCES


Punkaj Bhambri he’s assistant professor in Guru Nanak Dev Engineering College, Ludhiana in IT department.