Markov Matrix and Entropy based Tamper Detection Technique for Text Images

DOI:10.36909/jer.10947

Balkar Singh[1], M. K. Sharma[2]

[1]Computer Science and Engineering Department,

[2]School of Mathematics

Thapar Institute of Engineering and Technology

Patiala, Punjab INDIA

Email-balkar.singh@thapar.edu[1]; mksharma@thapar.edu[2]

ABSTRACT

In this paper, a novel watermarking technique for the tamper detection of text images is proposed. Entropy of every sentence is computed and Markov matrix using the occurrences of the characters is used to generate a character pattern. Entropy and character patterns are converted to Unicode Zero Width Characters (ZWCs) by using a lookup table. The *ZWCs* of entropy of each sentence is embedded at the end of every sentence after terminator. *ZWCs* of the character patterns are embedded in the end of the text of the image. On receiver side, ZWCs are extracted and converted to numerical form using the same lookup table. Entropy of every sentence and character patterns are recalculated and compared with extracted values for tamper detection. Comparison of technique with existing state-of-art techniques shows the effectiveness of the proposed technique.

Keywords:

Text watermarking, Tamper Detection, Unicode Zero Width Characters, Entropy, Integrity Rate.

INTRODUCTION

Application of digital media is increasing exponentially day by day. Digital media like text, audio, video and document are transferred through unsecured communication channel which needs a security malicious attack. Document and text media contains important information like personal information, bank account information and these types of information always

needs more security as compared to other media. Different methods like encryption and decryption are used to provide the security to digital media. Watermarking is one of the approaches which is used to provide protection and authentication to a digital media.

Entropy of the text data is one of the attributes that can be used to embed the watermark. Most frequently occurred with smallest font size in document are identified and suitable regions are selected based on entropy variations (Kurup *et al.,* 2007). Entropy of the sentences in Chineses text is utilized by Yingjie *et al.* (2010) to generate the watermark. The entropy is calculated by using the word frequency and further used to select the crucial sentences to embed the watermark. The watermark is constructed by using order of the crucial sentences. Given the complexity of Chinese text semantics, a sentence with a high word frequency possesses larger entropy.

Entropy is used to identify least distortive area in which embedding of the data can be done. Blocks with the small font size are selected to embed the secret bits (Khan *et al.,* 2011). Unicode space characters is utilized by Por *et al*. (2012) to increase the embedding efficiency and it also supports reversible method. Lossless watermarking and digital signatures are used to secure digital images (Umamageswari *et al*. 2014). The homogeneity of pixels is examined in order to discover possible cover image blocks to which watermark image blocks can be applied, preserving the visual quality of the watermarked images (Varghese *et al*. 2015). The features of Chinese sentences are utilized by Liu. *et al.* (2015). The text of document is divided into sentences and semantic code of each word is used to calculate its entropy. The sentence entropy, relevance, weighing function and length are used to calculate the weight of each sentence. The generated key is encrypted and stored with Certifying Authority. Al-Maweri *et al.* (2016) proposed a text watermarking on the basis of Unicode extended characters. The Unicode extended characters are used to embed the watermark bits in the text. Alotaibi and Elrefaei (2017) proposes a text watermarking on the basis of open word space for Arabic text. In the first method, Dotting feature in Arabic text is used to embed pseudo-space after and before the normal space. In the second method, embeds the pseudo-space and zero width space to increase the capacity. Naqvi *et al.* (2018) proposed a multilayer partially homomorphic encryption text steganography based on zero steganography approach. To increase capacity characters of cover message are replaced with the characters of secret message. Robustness is increased by using the multilayer encoding concept. An instance-based learning algorithm is proposed by Ahvanooey *et al.* (2020) proposed for Latin text watermarking. Zero width Unicode char- acters are used to convert the watermark into invisible form. This watermark

can be extracted to provide proof of ownership. Zero based text watermarking using Effective Characters List (ECL) is proposed by Saba *et al.* (2020). *ECL* is used to enhance the fragility of watermark. The technique is evaluated using attacks like deletion, reordering and insertion.

The existing authentication techniques applicable to text images are language dependent. Mostly techniques are applicable to English text and some are applicable to Arabic text. To remove this limitation, a novel text watermarking technique based on entropy of each sentence and character patterns of the cover text is proposed. *ZWCs* of entropy of every sentence and of character patterns is generated and embedded in the cover text image. These *ZWCs* values are extracted for the tampered detection process, when needed. This technique is language independent and is applicable to text images of any language. Only need to identify the terminator of sentence, as it changes from language to language.

This paper is organized as follows: Unicode standard and entropy used in proposed technique are discussed in preliminary section. Proposed technique is described followed by experimental results. Finally, conclusion and future scope are discussed.

Preliminary

This section presents a brief description of the Unicode standard as well as encoding pattern and entropy.

Unicode standard

The Unicode is a standard defined that is used to represent, encode and handle a digital text. Unicode have specific *ZWC* which are used to handle specific entities like zero width joiner combines two supportable characters together in particular language [1]. There are four *ZWCs* used to embed the encoding value in cover text, as shown in Table 1.

Table 1: Unicode zero width characters

ZWC	Hexadecimal Code	Symbol Used
Left to Right Mark	U + 200E	No symbol/width
Zero Width Non-Joiner	U + 200C	No symbol/width
POP Directional	U + 202C	No symbol/width
Left to Right Override	U + 202D	No symbol/width

Table 2 contains the encoding pattern of *ZWCs*. This pattern helps to generate a *ZWCs* of a value in the invisible form and image visual quality is not degrading after an embedding process.

Table 2: Encoding pattern of *ZWCs*

Value	Two ZW Hex Codes	Value	Two ZW Hex Codes
0	200E + 200E	5	200C + 202C
1	200E + 202C	6	200C + 200D
2	200E + 202D	7	202C + 202C
3	200E + 200C	8	202C + 200C
4	200C + 200E	9	202C + 202D
		'.'	200C + 200C

Entropy

Entropy is the measurement of randomness and it provides the amount of average information. It is defined as

$$H(X) = -\sum_{k} p_k \ log_2(p_k) \qquad (1)$$

where k is the number of gray levels and p_k is the probability associated with gray level k. Entropy of a sentence depends upon the occurrence of characters in the word and independent of the language used.

Character Pattern Generation using Markov Matrix

Characters are the smallest elements in a text's structure. In proposed technique, occurrence of the characters is used to create a Markov matrix. This matrix is used to create a character pattern. For matrix generation, all of the characters in the text are converted to small cases and a list of all of the document's characters is created. This list is used to generate a Markov Matrix, M, as shown below:

$$M = \left\{ \begin{vmatrix} * & c_1 & c_2 & .. & c_n \\ c_1 & t_{11} & t_{12} & .. & t_{1n} \\ c_2 & t_{21} & t_{22} & .. & t_{2n} \\ .. & .. & .. & .. & .. \\ c_n & t_{n1} & t_{n2} & .. & t_{nn} \end{vmatrix} \begin{bmatrix} P \\ P_1 \\ P_2 \\ . \\ P_n \end{bmatrix} \right\} \qquad (2)$$

Where, c_i refers to an element of list, t_{ij} is the number of immediate appearances of c_j after c_i in the document, * denotes the transition point, and P_i shows the transition pattern of each element of the list. The sum of each row in the given matrix is 1 and none of the list elements are negative.

Proposed Technique

In this section, embedding, and extraction algorithms of the proposed technique are discussed. Main goal of the text watermarking scheme is to tamper detection in the text image by embedding a *ZWcs* of entropy after every sentence. The basic idea of the proposed technique is that if during transmission, some words or even letters of the cover text are changed, then the frequency of the occurrence of the letter changes. This change is reflected in the value of the entropy. The value of the entropy embedded in the end of each sentence is compared with newly calculated value. This comparison is used to find the tamper of the text. Flowchart of proposed technique is shown in Figure 1.

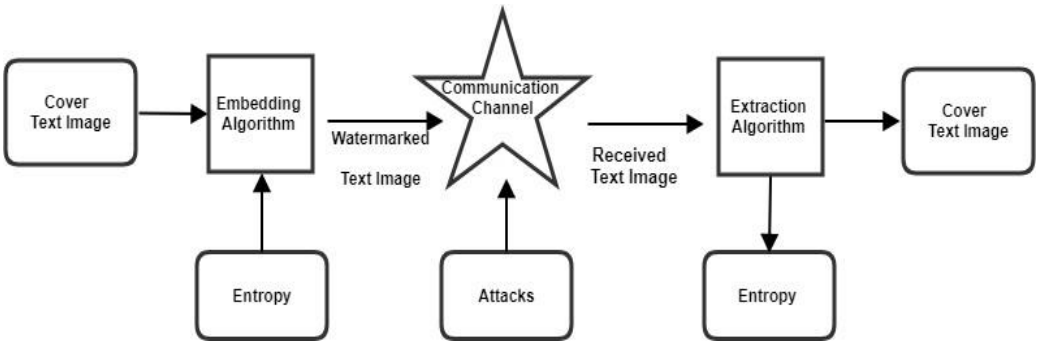

Figure 1: Text Watermarking Scheme

Embedding Algorithm

Step i. Calculate the entropy E_i of every sentence of the cover text using (1), where i =1, 2, ...

Step ii. A terminator is selected at end of every sentence that changes language to language like '.' is used in English language while '|' in Hindi as well as in Punjabi language.

Step iii. Convert E_i to *ZWCs* by using the mapping given in Table 2.

Step iv. Convert each character of the cover text into lower case and count the total occurrence of each character. Generate a Markov Matrix M, as shown in (2). Embed the pattern of the cover text in the end of the cover text by converting into *ZWCs* given in Table 2.

Step v. Embed *ZWCs* values of the entropies at end of sentence and *ZWCs* of the character pattern in the end of the cover text.

Step vi. Generate the Watermarked Text Image CT_w.

Extraction Algorithm

Step i. Read the cover text image CT_w.

Step ii. Calculate the entropy E_i of every sentence of the cover text using (1), where i =1,2, ...

Step iii. Generate the character pattern of each character *k by* using Markov Matrix M, as shown in (2) and store into P'_k.

Step iv. Extract *ZWCs* of entropy of every sentence of the cover text and *ZWCs* of each character pattern.

Step v. Convert *ZWCs* of the entropies of the i^{th} sentence to a number E'_i and character pattern of each character P_k by using Table 2.

Step vi. Compare E_i and E'_i for each i where i =1, 2, ...

if $E_i == E'_i$

then i^{th} sentence of the received cover text is not tampered, otherwise, it is tampered and use the character patterns P_k and P'_k to find the tampered words.

Experimental Results and Analysis

Proposed technique is implemented in R2019a version of MATLAB tool. Text images, having the contents used by the existing state-of-art techniques, are used as the cover images. Different watermark is also considered in this experimental analysis. This technique is language independent and can be used with the text of any language.

Invisibility

The proposed technique does not degrade the quality of cover text due to embedding of *ZWCs*. *ZWCs* of entropy is embedded at the end of every sentence according to the entropy of that sentence as shown in Figure 2. Human Vision Systems is not able to detect the effect of embedding process.

Entropy value 0.9401 0.9812 0.8987 1.0490 1.0746
ZWCs 200E+200E+200C+200C+202C+202D+200C+200E+200E+200E+200C+200C+202C+202D+202C+200C+ 200E+200E+200C+200C+202C+202D+200E+200E+200E+202C+200C+200C+200E+200E+200C+202C+ 200E+202C+200C+200C+200E+200E+202C+202C
'मैं भारत से प्यार करता हूँ। गुरुदेव रवींद्र नाथ ठाकुर भारत के बँगला साहित्य के शिरोमणि कवि थे। उनकी कविता में प्रकृति के सौंदर्य और कोमलतम मानवीय भावनाओं का उत्कृष्ट चित्रण है। "जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त है। गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है।'
Embed ZWCs of entropy after '|' of every sentence.

Figure 2: Embedding process Example

Distortion Robustness

Distortion Robustness (DR) is computed with the help of Loosing Probability (LP) [1]. *LP* is calculated by using the sentences and total length of string. *LP* is equal to the number of sentences divided by total length of string i.e $LP = \frac{NS}{LT}$ where NS is number of sentences through the *CT* and *LT* is length of *CT*. If *CT* = [Once you stop learning, you start dying.] then value of *LP* is determined by $LP = \frac{1}{40} = 0.025$. The probability of the *DR* is determined by P(DR) = [1-0.025] = 0.975 = 97.5 %.

Table 3: *DR* comparison with the existing techniques

Water-mark (W)	CT	Length of CT	Proposed Technique	[Ahva-nooey *et al.*, (2020)]	[Alot-aibi *et al.*, (2017)]	[Naqvi *et al.*, (2018)]	[Pati-burn *et al.*, (2017)]	[Por *et al.*, (2012)]	[Ri-zzo *et al.*, (2017)]
Albert	Once you stop learn-ing, you start dying.	40	97.5	97.5	83.3	90	85	83.3	72.5

Einstein	In the middle of difficulty lies opportunity.	45	97.7	97.7	83.3	84.4	85.6	83.3	57.7
Albert Einstein	Unthinking respect for authority is the greatest enemy of truth.	64	98.4	98.4	88.8	82.8	85.9	88.8	78.1
Steven Paul Jobs	Great things in business are never done by one person. They are done by a team of people.	89	97.7	97.7	94.1	86.5	80.6	94.1	74.1
Ali ibn Abi Talib	Beautiful people are not always good, but good people are always beautiful.	75	98.6	98.6	90.9	81.3	85.3	90.9	72
William Henry Gates III	Don't compare yourself with anyone in this world. If you do so, you are insulting yourself.	91	97.8	97.8	93	80.2	83.5	93.3	67

Table 3 describes the *DR* comparison with the existing technique. *DR* is calculated by using total number of sentences and total length of string. The proposed technique and existing technique [1] follow the same rule to determine the *DR*. This is reason *DR* of the proposed technique as well as existing technique [1] is same.

Table 4: Embedding Capacity comparison with the existing techniques

Water-mark (W)	CT	Length of CT	Proposed Technique	[Ahva-nooey *et al.*, (2020)]	[Alot-aibi *et al.*, (2017)]	[Naqvi *et al.*, (2018)]	[Pati-burn *et al.*, (2017)]	[Por *et al.*, (2012)]	[Ri-zzo *et al.*, (2017)]
Albert	Once you stop learning, you start dying.	40	6	6	3	4	6	1.7	2.8
Einstein	In the middle of difficulty lies opportunity.	45	6	8	3	7	6	1.7	3.8
Albert Ein-stein	Unthinkin g respect for authority is the greatest enemy of truth.	64	6	15	4.5	11	9	2.25	4
Steven Paul Jobs	Great things in business are never done by one person. They are done by a team of people.	89	12	32	8.5	12	11	4.25	7.2
Ali ibn Abi Talib	Beautiful people are not always good, but good people are always beautiful.	75	6	17	5.5	14	15	2.75	5.3
	Don’t compare								

William Henry Gates III	ourself with anyone in this world. If you do so, you are insulting yourself.	91	12	46	7.5	18	15	3.75	7.5

Embedding Capacity

The entropy value of each sentence is converted into *ZWCs* before the embedding process. The comparison with existing technique is shown in Table 4.

The comparison between the proposed technique and existing technique is shown in the Table 4. As shown in Table 4, different external watermark is used by existing techniques for every string. But the proposed technique embeds *ZWCs* of entropy value after every sentence. Embedding capacity of the proposed technique depends upon number of sentences in a text.

Prevention against malicious attacks

Deletion attack

If a malicious user deletes the content of the watermarked text, then it can be detected with the help of entropy value. It is cleared from Table 5 that value of entropy in both cases have changed which shows that received text has been tampered.

Table 5: Deletion Attack Example

Cover Text Once you stop learning, you start dying. In the middle of difficulty lies opportunity. Unthinking respect for authority is the greatest enemy of truth. Great things in business are never done by one person. They are done by a team of people.
Entropy value (Entropy of first sentence =4.0276, Entropy of second sentence=4.1313, Entropy of third sentence=4.0470, Entropy of fourth sentence= 3.8913, Entropy of fifth sentence= 3.6144)
Tampered Text Once stop learning, you start dying. In middle of difficulty lies opportunity. Unthinking respect for authority is greatest enemy of truth. Great things in business never done by one person. They done bya team of people.

Entropy Value (Entropy of first sentence =4.0487, Entropy of second sentence=4.1008, Entropy of third sentence=4.0932, Entropy of fourth sentence= 3.9086, Entropy of fifth sentence= 3.5988)

As shown in Table 5, red color words of cover text has been deleted by a malicious user attack and these words are missing in the tampered text. The value of entropy and character patterns are changed due to deletion of these words.

Figure 3 shows the deletion attack on Hindi cover text. Entropy of every sentence of cover text is calculated as shown in the Figure 3. Entropy of every sentence is embedded behind every sentence individually during the embedding process.

Cover Text
मैं भारत से प्यार करता हूँ |
गुरुदेव रवींद्र नाथ ठाकुर भारत के बँगला साहित्य के शिरोमणि कवि थे |
उनकी कविता में प्रकृति के सौंदर्य और कोमलतम मानवीय भावनाओं का उत्कृष्ट चित्रण है|
"जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त है|
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है|
Entropy value (Entropy of first sentence=0.9401, Entropy of second sentence=0.9812, Entropy of third sentence=0.8987, Entropy of fourth sentence=1.0490, Entropy of fifth sentence=1.0746)
Tampered Text
मैं भारत से प्यार हूँ |
गुरुदेव रवींद्र नाथ भारत के बँगला साहित्य के शिरोमणि कवि थे |
उनकी कविता में प्रकृति के सौंदर्य और कोमलतम भावनाओं का उत्कृष्ट चित्रण है|
"जन गण मन" उनकी रचित एक कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त है|
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय ओर से, आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है|
Entropy value (Entropy of first sentence=0.9738, Entropy of second sentence=1.0063, Entropy of third sentence=0.9217, Entropy of fourth sentence=1.0820, Entropy of fifth sentence=1.0751)

Figure 3: Deletion Attack Example

Figure 3 describe the tamper detection that is occurred due to deletion attack. The words shown with underline are deleted from the cover text during deletion attack. The different values of entropy of every sentence show that text has been tampered.

In the Figure 4, the cover text and its corresponding tampered text of Punjabi language are shown. The words with underline in cover text are deleted from the tampered text.

Cover Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏਕੁ ਹੈ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਏਕੋ ਸਬਦੁ ਹੈ ਨਾਨਕ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਅੰਮ੍ਰਿਤ ਹਰਿ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਮੀਠਾ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾਦੁ ਆਇਆ |
ਅੰਮ੍ਰਿਤੁ ਨੀਰੁ ਗਿਆਨਿ ਮਨ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.8060, Entropy of second sentence=1.1149, Entropy of third sentence=1.1745, Entropy of fourth sentence=1.1279, Entropy of fifth sentence=1.0412)
Tampered Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏਕੁ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਸਬਦੁ ਹੈ ਨਾਨਕ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਅੰਮ੍ਰਿਤ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾਦੁ ਆਇਆ |
ਅੰਮ੍ਰਿਤੁ ਨੀਰੁ ਗਿਆਨਿ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.7852, Entropy of second sentence=1.1692, Entropy of third sentence=1.1991, Entropy of fourth sentence=1.1642, Entropy of fifth sentence=1.0354)

Figure 4: Deletion Attack Example

Entropy values and Character Patterns each character of cover and tampered text are different. These differences show that both the text is different. On this basis, one can say that the proposed technique is able to detect the deletion attack.

Insertion Attack

A malicious user can insert extra word in the cover text during transmission time is known as an insertion attack. Due to change in the entropy value, the proposed technique is able to detect insertion type of tampering.

Table 6: Insertion Attack Example

Cover Text
Once you stop learning, you start dying. In the middle of difficulty lies opportunity. Unthinking respect for authority is the greatest enemy of truth. Great things in business are never done by one person. They are done by a team of people.
Entropy value
(Entropy of first sentence =4.0276, Entropy of second sentence=4.1313, Entropy of third sentence=4.0470, Entropy of fourth sentence= 3.8913, Entropy of fifth sentence= 3.6144)

Tampered Text
Once you stop the learning, you start dying. In the middle of difficulty lies a opportunity. Unthinking respect for authority is the greatest enemy of a truth. Great things inbusiness are never done by one from person. They are not done by a team of people.
Entropy Value
(Entropy of first sentence =4.0522, Entropy of second sentence=4.1650, Entropy of third sentence=4.0393, Entropy of fourth sentence= 3.9825, Entropy of fifth sentence= 3.6240)

Table 6 shows the result of insertion attack by a malicious user. Red words are inserted in the cover text during the attack. Values of entropy is changed due to insertion of words as shown in Table 6. Only one alphabet insertion effects entropy as shown in Table 6.

Figure 5 describe the example of insertion attack on Hindi language. Some words are shown with underline inserted within the original cover text. These words are inserted due to insertion attack by a malicious user.

Cover Text
मैं भारत से प्यार करता हूँ |
गुरुदेव रवींद्र नाथ ठाकुर भारत के बँगला साहित्य के शिरोमणि कवि थे |
उनकी कविता में प्रकृति के सौंदर्य और कोमलतम मानवीय भावनाओं का उत्कृष्ट चित्रण है|
"जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त है|
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है|
Entropy value
(Entropy of first sentence=0.9401, Entropy of second sentence=0.9812, Entropy of third sentence= 0.8987, Entropy of fourth sentence=1.0490, Entropy of fifth sentence=1.0746)
Tampered Text
मैं भारत से प्यार मे करता हूँ |
गुरुदेव रवींद्र नाथ ठाकुर भारत उनकी के बँगला साहित्य के शिरोमणि कवि थे |
उनकी कविता में प्रकृति के सौंदर्य प्यार और कोमलतम मानवीय भावनाओं का उत्कृष्ट चित्रण है|
"जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त चित्रण है|
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, कविता आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है|
Entropy value
(Entropy of first sentence=0.9308, Entropy of second sentence=0.9680, Entropy of third sentence=0.8854, Entropy of fourth sentence=1.0257, Entropy of fifth sentence=1.0553)

Figure 5: Insertion Attack Example

As shown in Figure 5, entropy of cover text is different as compared to tampered text. Change in the entropy is occurred due to insertion some words in the original text. A change in entropy means that cover text has been tampered.

Figure 6 shows the Punjabi language cover text and its tampered text. The entropy of both texts is different as shown in Figure. The words with underline are inserted during the insertion attack.

Cover Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏਕੁ ਹੈ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਏਕੋ ਸਬਦੁ ਹੈ ਨਾਨਕ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਅੰਮ੍ਰਿਤ ਹਰਿ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਮੀਠਾ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾਦੁ ਆਇਆ |
ਅੰਮ੍ਰਿਤੁ ਨੀਰੁ ਗਿਆਨਿ ਮਨ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.8060, Entropy of second sentence=1.1149, Entropy of third sentence=1.1745, Entropy of fourth sentence=1.1279, Entropy of fifth sentence=1.0412)
Tampered Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏਕੁ ਹੈ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਹੈ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਏਕੋ ਸਬਦੁ ਹੈ ਨਾਨਕ ਏਕੁ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਏਕੋ ਅੰਮ੍ਰਿਤ ਹਰਿ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਮੀਠਾ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾਦੁ ਆਇਆ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤੁ ਨੀਰੁ ਨਾਮੁ ਗਿਆਨਿ ਮਨ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.8226, Entropy of second sentence=1.0991, Entropy of third sentence=1.1885, Entropy of fourth sentence=1.1207, Entropy of fifth sentence=1.0203)

Figure 6: Insertion Attack Example

As shown in Figure 6, entropy value is changed due to insertion attack. Different entropy means that the cover text has been tampered.

Tampered Attack

In this type of attack, a word is altered from a malicious user. Due to alteration of a word, entropy value always effects as shown in Table 7.

Table 7: Tamper Attack Example

Cover Text
Once you stop learning, you start dying. In the middle of difficulty lies opportunity. Unthinking respect for authority is the greatest enemy of truth. Great things in business are never done by one person. They are done by a team of people.

Entropy value (Entropy of first sentence =4.0276, Entropy of second sentence=4.1313, Entropy of third sentence=4.0470, Entropy of fourth sentence= 3.8913, Entropy of fifth sentence= 3.6144)
Tampered Text Once you stops learning, you start dying. In the middles of difficulty lies opportunity. Unthinking respects for authority is the greatest enemy of truth. Great things in business are never do by one person. They are done by an team of people.
Entropy Value (Entropy of first sentence =4.0036, Entropy of second sentence=4.1525, Entropy of third sentence=4.0500, Entropy of fourth sentence= 3.9251, Entropy of fifth sentence= 3.6416)

The words with red color are altered due to tampered attack by a malicious user as shown in Table 7. The entropy of every sentence is changed due to alter a word. So, the proposed technique is able to detect these types of tampering with help of entropy value.

Cover and tampered text are shown in the Figure 7 of Hindi language. Entropy value of both the text is calculated as shown in the Figure.

Cover Text
मैं भारत से प्यार करता हूँ।
गुरुदेव रवींद्र नाथ ठाकुर भारत के बँगला साहित्य के शिरोमणि कवि थे।
उनकी कविता में प्रकृति के सौंदर्य और कोमलतम मानवीय भावनाओं का उत्कृष्ट चित्रण है।
"जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हमारे राष्ट्रीय गीत होने का गौरव प्राप्त है।
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, आप सबको यह कविता अपने मूल बंगला रूप में प्रस्तुत है।
Entropy value
(Entropy of first sentence=0.9401, Entropy of second sentence=0.9812, Entropy of third sentence=0.8987, Entropy of fourth sentence=1.0490, Entropy of fifth sentence=1.0746)
Tampered Text
मैं भारत से प्यार कर हूँ।
गुरुदेव रवींद्र नाथ ठाकुर भारत के बँगला साहित्य के शिरो कवि थे।
उनकी कविता में प्रकृति के सौंदर्य और कोमल मानवीय भावनाओं का उत्कृष्ट चित्रण है।
"जन गण मन" उनकी रचित एक विशिष्ट कविता है जिसके प्रथम छंद को हम राष्ट्रीय गीत होने का गौरव प्राप्त है।
गणतंत्र दिवस के शुभ अवसर पर, काव्यालय की ओर से, आप सब यह कविता अपने मूल बंगला रूप में प्रस्तुत है।
Entropy value
(Entropy of first sentence=0.9790, Entropy of second sentence=1.0093, Entropy of third sentence=0.9132, Entropy of fourth sentence=1.0681, Entropy of fifth sentence=1.0879)

Figure 7: Tamper Attack Example

As shown in the Figure 7, entropy value of both the text is different. The entropy value is different of both the text due to some words are tampered due to tampered attack. The words with underline as shown in Figure 7 are tampered of the cover text. At the base of entropy value, the proposed technique can detect the tampered text.

Figure 8 describe the cover and tampered text of Punjabi language. The words with underline are altered word of cover text. These words are altered due to tamper attack.

Cover Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏਕੁ ਹੈ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਏਕੋ ਸਬਦੁ ਹੈ ਨਾਨਕ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਅੰਮ੍ਰਿਤ ਹਰਿ ਬਾਣੀ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਮੀਠਾ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾਦੁ ਆਇਆ |
ਅੰਮ੍ਰਿਤੁ ਨੀਰੁ ਗਿਆਨਿ ਮਨ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.8060, Entropy of second sentence=1.1149, Entropy of third sentence=1.1745, Entropy of fourth sentence=1.1279, Entropy of fifth sentence=1.0412)
Tampered Text
ਨਾਨਕ ਅੰਮ੍ਰਿਤੁ ਏ ਹੈ ਦੂਜਾ ਅੰਮ੍ਰਿਤੁ ਨਾਹਿ |
ਅੰਮ੍ਰਿਤੁ ਏਕੋ ਸਬ ਹੈ ਨਾਨਕ ਗੁਰਮੁਖਿ ਪਾਇਆ |
ਅੰਮ੍ਰਿਤ ਸਬਦੁ ਅੰਮ੍ਰਿਤ ਹਰਿ ਬਾ |
ਅੰਮ੍ਰਿਤ ਨਾਮੁ ਸਦ ਮੀਠਾ ਲਾਗਾ ਗੁਰ ਸਬਦੀ ਸਾ ਆਇਆ |
ਅੰਮ੍ਰਿਤੁ ਨੀ ਗਿਆਨਿ ਮਨ ਮਜਨੁ ਅਠਸਠਿ ਤੀਰਥ ਸੰਗਿ ਗਹੇ |
Entropy value
(Entropy of first sentence=0.8306, Entropy of second sentence=1.1505, Entropy of third sentence=1.2235, Entropy of fourth sentence=1.1586, Entropy of fifth sentence=1.0693)

Figure 8: Tamper Attack Example

It is clear from the Figure 8, entropy value is changed due to tamper attack. Different entropy value means that cover text has been tampered as shown in Figure 8.

Integrity Rate

A malicious user can alter or add any extra word in the text during the transmission process. Integrity rate *IR* of a technique is ability to detect a malicious user's activity. The proposed technique can detect different types of activities like alter or add any word in the text. From the Tables 5, 6 and 7, it is cleared that proposed technique works for different types of attack and its *IR* is 100%.

IR of the proposed technique is compared with Ahvanooey et al. (2020). The results of this comparison are shown in Table 8.

Table 8: Comparison of IR with the existing technique

Original CT	Compromised CT	Ahvanooey et al. (2020)	Proposed Technique
Once you stop learning, you start dying.	Once you discontinue learning, you start dying.	86 %	100 %
In the middle of difficulty lies opportunity.	In the medium of hardship lies opportunity.	71 %	100 %

Unthinking respect for authority is the greatest enemy of truth.	No thinking about authority is the greatest enemy of truth.	82 %	100 %
Great things in business are never done by one person. They are done by a team of people.	Great earnings in business are never achieved by one person. They are done by a team of people.	84 %	100 %
Beautiful people are not always good, but good people are always beautiful.	Beautiful persons are not always good, but good persons are always look beautiful.	62 %	100 %
Don't compare yourself with anyone in this world. If you do so, you are insulting yourself.	Do not compare yourself with anyone in this world. If you do so, you are offending yourself.	74 %	100 %

Table 8 describes the comparison of *IR* between the proposed technique with the existing technique. The idea behind the *IR* to verify the integrity and originality of the watermarked text. If an unauthorized user manipulates CT_w and the technique can detect this manipulation process is known as *IR* of the technique. The proposed technique can detect tampered occurred due to different types of attacks. This is the reason that *IR* of the proposed technique is 100% in every case.

Visual attack

In the embedding process, Unicode *ZWCs* of entropy value are embedded into every sentence of *CT*. The embedding trace of *ZWCs* is usually invisible to the *HVS* and not easy to a malicious user to trace these invisible entropy values.

Conclusions

In this paper, a text watermarking technique for tampered detection of text images based on Uni- code *ZWC* and entropy has been published. *ZWCs* of entropy of each sentence are embedded at the end of the sentence. The quality of text will not degrade due to embedding of *ZWCs*. The proposed technique can detect the tamper occurred due to attack like insertion, deletion and tampering attack. The *IR* of the technique is 100% and provides good embedding capacity as well as good tamper detection capability against different attacks.

References

Ahvanooey, M.T., Li, Q., Zhu, X., Alazab, M., & Zhang, J. 2020. A novel intelligent text watermarking technique for forensic identification of spurious information on social media. Computers & Security 90: 1-14.

Al-Maweri, N.A.S., Adnan, W.A.W., Ramli, A.R., Samsudin, K., & Rahman, S.M.S.A.A. 2016. Robust digital text watermarking algorithm based on Unicode extended characters. Indian Journal of Science and Technology 9: 1-14.

Alotaibi, R.A., & Elrefaei, L.A. 2017. Improved capacity Arabic text watermarking methods based on open word space. Journal of King Saud University-Computer and Information Sciences 30: 236-248.

Khan, A., Khanam, M., Bashir, S., Khiyal, M.S.H., Iqbal, A., & Khan, F.H. 2011. Entropy based data hiding in binary document images. International Journal of Computer and Electrical Engineering 3: 503-506.

Kurup, S., Sridhar, G., & Sridhar, V. 2007. Entropy based data hiding for document images. World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering 1: 3582-3585.

Liu, Y., Zhu, Y., & Xin, G. 2015. A zero-watermarking algorithm based on merging features of sentences for Chinese text. Journal of Chinese Institute of Engineers 38(3): 391-398.

Naqvi, N., Abbasi, A.T., Hussain, R., Khan, M.A., & Ahmad, B. 2018. Multilayer partially Homomorphic encryption text steganography: A zero steganography approach. Wireless Personal Communications 103(2): 1563-1585.

Patiburn, S.A., Iranmanesh, V., & Teh, P.L. 2017. Text Steganography using Daily Emotions Monitoring. International Journal of Education and Management Engineering 7(3): 1-14. Por, L.Y., Wong, K., & Chee, K.O. 2012.

Por, L.Y., Wong, K., & Chee, K.O. 2012. UniSpaCh: A text-based data hiding method using Unicode space characters. The Journal of Systems and Software 85(5): 1075-1082.

Rizzo, S.G., Bertini, F., Montesi, D., & Stomeo, C. 2017. Text watermarking in social media. In the proceedings of International Conference on Advances in Social Networks Analysis and Mining, IEEE/ACM 208-211.

Saba, T., Bashardoost, M., Kolivand, H., Rahim, M.S.M., Rehman, A., & Khan, M.A. 2020. Enhancing fragility of zero-based text watermarking utilizing effective characters list. Multimedia Tools and Applications 79(1): 341-354.

Umamageswari, A., & Suresh, G.R. 2014. Secure medical image communication using ROI based lossless watermarking and novel digital signature. Journal of Engineering Research 2(3): 87-108.

Varghese, J., Subash, S., Hussain, O.B., Saddy, M.R., Babu, B., & Riazuddin, M. 2015. Image adaptive DCT-SVD based digital watermarking scheme by human visual characteristics. Journal of Engineering Research 3(1): 95-112.

Yingjie, M., Tao, G., Zhihua, G., & Liming, G. 2010. Chinese text zero-watermark based on sentence's Entropy. In the proceedings of international conference on multimedia technology, IEEE 1-4.