Implementation of a Novel Watermarking Technique for Devanagari Text

Nitin N. Patil and J. B. Patil

***Abstract*—In recent decades, Internet technology has gained tremendous importance across the globe to substitute the traditional paperwork of information dissemination with electronic form. This expansion of Internet has triggered security constraints of intellectual property rights. Various security issues like illegal distribution, copy, reproduction and authentication can pose threat to this global information available on Internet. Till now considerable attempts are made for developing watermarking techniques for numerous natural languages. Each attempt has its own distinguished approach with respect to the properties of concerned natural language. This paper introduces a novel text watermarking technique to secure the intellectual property rights of the original author of Devanagari text. We mainly focus on Devanagari Text for which no watermarking technique has been proposed so far. Our aim is to develop an embedding algorithm which cleverly uses the unique construct of Devanagari language 'sarvanam' (pronoun) for generation of the effective watermarks in combination with additional security phrases. The proposed extraction algorithm also supports the defined security checks to preserve the authentication and copyrights of the genuine author of the Devanagari text document. Our experimental results demonstrate that proposed novel technique is significant for Devanagari text of varying length.**

***Index Terms*—Intellectual property, Devanagari text, sarvanam, copyrights, natural languages.**

I. Introduction

The enormous use of Internet technology has given rise to electronic form of most of the user contents instead of conventional paperwork. All these user contents comprise different forms like image, audio, video and text which are universally available on Internet. Because of easy availability of this data, there are many security threats such as illegal redistribution, copying, authentication fingerprinting and tampering. Digital watermarking is an appropriate and dominant solution against these security issues for all types of the digital contents. A typical watermarking algorithm needs to possess the properties such as robustness, security, imperceptibility, detectability and capacity [1], [2]. Generally text information covers a large amount of digital content consisting literature, news articles, Government Resolutions, educational (teaching-learning) material, commercial data and a lot many. The process of embedding a watermark in text that uniquely identifies the original copyright owner of the text is called as Digital Text Watermarking [3], [4]. Till now various attempts are made by researchers to develop text watermarking techniques to meet the security requirements of digital text in some natural languages like English, Chinese, Arabic, German and Turkish [5]-[9]. Each of these text watermarking technique has unique approach to secure the contents of respective language. In this paper, we focus on watermarking for Marathi language text which is official language of Maharashtra and Goa state in India and one of the twenty-two official languages of the India. Basically the Devanagari script is used to write in many Indian and foreign languages. Marathi is the southern-most Indo-Aryan language. It is spoken by over 90 million people in India. This places Marathi among the top 15 languages of the world with respect to total number of speakers. Outside India Marathi is spoken in Israel and Mauritius [10]. A wide range of Marathi text of different categories is available across Internet. In this paper we propose a novel watermarking technique for Marathi language text. This proposed technique makes effective use of the sarvanam (pronoun) which is an important language construct to embed the watermark in the text. This technique belongs to the open-space method which follows structural approach and does not affect the value and meaning of the document. We organize this paper as follows: Section II provides previous work in the area of text watermarking and particularly in the linguistic text watermarking. Section III demonstrates our proposed embedding and extraction watermarking algorithms. It is followed by experimental results and discussion which focuses on frequency comparison of natural language watermarks in Section IV. The last section includes conclusion of this paper and indicates future directions.

Manuscript received November 3, 2014; revised January 6, 2015.

Nitin N. Patil is with the Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur (M.S.) India (e-mail: er_nitinpatil@rediffmail.com).

J. B. Patil is with the R. C. Patel Institute of Technology, Shirpur (M.S.) India (e-mail: jbpatil@hotmail.com).

II. Previous Work

Internet holds wide range of different type of media like image, audio, video and text at a large extent. The digital watermarking can be characterized as image watermarking, audio watermarking, video watermarking and text watermarking. Although the format of media is diverse, then also the main goal is to secure the contents against different attacks & illegitimate activities like redistribution, copying, authentication, fingerprinting and tampering. The text watermarking has significant scope in information hiding domain due to widespread accessibility of text documents over the Internet. Maxemchuk *et al.* suggested possibilities to restrict black and white image redistribution [11]. In this approach, to insert the watermark, the sentences were reformed by changing the position of individual characters or word. This approach was further evaluated by researchers [3], [12]. Several other methods, such as line-shifting,

DOI: 10.7763/IJIEE.2015.V5.557

word-shifting and character modification are used to embed data within text image. Though these methods are robust; they failed to secure the embedded data directly [12], [13]. MercanTopkara *et al.* discussed watermarking of natural language text by using structure of the sentence in insertion of the watermark. The text elements like words, characters or lines were modified to embed the information [5]. S. Ranganathan *et al*. have proposed a watermarking approach by combining image based and syntactic watermarking [14]. C. Culnane *et al*. developed a method with greater watermarking capacity for formatted text which are robust for printing and scanning purpose [15]. Few authors have suggested text watermarking techniques for English, Chinese, Turkish, German and Arabic language text based on feature characteristic of these natural languages as discussed below. Yingli Zhang *et al*. have suggested a robust watermarking for protecting copyright and controlling propagation of MS-Word document in English and Chinese languages. It focused on properties of word document objects. This method is applicable to English and Chinese language [6]. Hasan M. Meral *et al.* have explored syntax-based natural language watermarking scheme for Turkish language using morphosyntactic tools [9]. Adnan Abdul-Aziz Gutub *et al*. have developed a digital watermarking scheme for Arabic e-text files by focusing on extension Arabic character 'Kashida' [7]. Zunera JaliI *et al.* proposed a structural approach for copyright protection of plain text document for English language. A zero text watermarking algorithm based on occurrence frequency of non-vowel ASCII characters and words for copyright protection of plain text documents is implemented. This structural algorithm is not appropriate for all type of text documents [16]. Nighat Mir has focused on grammatical rules like verbs, articles and prepositions of English language. To provide authorship protection of web page with more robust imperceptible digital watermarking. Frequency order and common occurrence of English language text is studied. According to which verbs, articles and prepositions are found to be most common and first hundred words in English [17].

III. Proposed Algorithm

Our proposed algorithm concentrates on occurrence of special language construct 'sarvanam' in the Marathi text. Sarvanam is very useful language construct in Marathi language. Even though Marathi is a highly inflected language, the form of sarvanam remains unchanged during suffix change or modified use in the sentences [10], [16]. Our algorithm uses concatenation of three security phrases to generate a secured key comprising sarvanam count, Author ID and Timestamp. The key is embedded securely without any alteration in the text. The meaning and value of the original text is preserved at the time of watermark embedding. In the proposed technique, a unique user id for the document creator is created by specific alphanumeric combination. Additionally the document creator chooses a security phrase of maximum 16 characters alphabet string followed by the timestamp. After these required inputs, some preprocessing is done prior to generating the embedding watermark. From the input text, the occurrence of sarvanam is counted. These occurrences are concatenated with randomly generated Author ID and desired timestamp. Advanced Encryption Algorithm (AES) is applied on the above discussed string combination such that an encrypted form of the key is obtained to be used to embed as the watermark. In case if any conflict regarding the copyright of the document is raised, this unique key is used in watermark extraction algorithm to verify the original author to resolve the copyright issue. The embedding algorithm is used by the genuine author of text document whereas the extraction algorithm confirms authentication of the author by Certifying Authority or a third party involved in resolving the copyright issue.

A. The Embedding Algorithm

In this step the embedding algorithm is designed to generate the required watermark to be embedded along with the desired security measures.

1) Input Text and author name.
2) Generate Author_ID from A-Z, a-z, 0-9 of length equal to length of author name.
3) Provide file name, security phrase and Timestamp.
4) Count occurrence of Marathi sarvanam. Define and assign counts to respective variables.
5) Repeat step 4 for all occurrences of Marathi sarvanam.
6) Concatenate Author_ID and watermark timestamp to Marathi sarvanam.
7) Apply AES to raw string to generate encrypted key with security phrase.
8) Output protected file with Author key (watermark).

B. The Extraction Algorithm

The extraction algorithm separates original text from the watermark which is the encrypted form of concatenated string comprising occurrences of sarvanam, Author ID and timestamp. The encrypted string is first decrypted and further processed to separate from each other to extract the watermark completely from the input Devanagari text. The authorship verification third party or Certified Author is able to extract the watermark if and only if all security parameters are satisfied effectively.

1) Input protected file (containing watermark).
2) Provide Author_ID, security phrase and Timestamp sequentially.
3) Separate encrypted watermark and original text from protected file.
4) Get raw string from encrypted watermark stored in protected file
   Applying AES and security phrase to raw string.
5) Isolate Marathi sarvanam and create temp string from original test.
6) Repeat step 4 for all occurrences of Marathi sarvanam.
7) Concatenate Author_ID and watermark timestamp to Marathi sarvanam.
8) Compare temp string created from original text and raw string generated from watermark.
9) If match found, output "You are Genuine Author of this Content!" else "You are not Author of this Content!"

IV. Experimental Results

Text watermarking area generally lacks for standards which determine robustness of the proposed technique. To authenticate our proposed Devanagari text watermarking algorithms, we do not have any specific corpus available for the experiments. We validated our experiments by considering text of different lengths. We defined three different categories of Marathi language text. All these categories are based on types of Marathi texts such as kavita (poems), charoli, (short poems), laghukatha (short stories), utara (short passages), laghunibhand (short essays), dirghakatha (long stories), dirghalekh (detail articles/passages), educational teaching-learning materials etc. For Text Category 1, we consider text length by number of lines whereas for Text Category 2 and 3 we count number of sentences. Five sample texts of varying size for each category are taken from different Marathi text sources. Table 1 summarizes number of extracted watermarks for different Text Categories. One can observe that the more the number of watermarks available, the better watermarking can be experimented for that text category. According to Marathi language grammar rules, there are five types of sarvanam such as personal (PP), Demonstrative (DP), relative (RP), interrogative / indefinite / quantifiers (IP) and reflexive (RxP). All the extracted watermarks are grouped in one of these five categories respectively [16].

TABLE I: Summary of Watermarks Used for All Three Categories of Devanagari Text

Text Category	Sarvanam as Watermark				
	PP (Personal Pronoun)	DP (Demonstrative Pronoun)	RP (Relative Pronoun)	IP (Interrogative/ Indefinite pronoun/ Quantifiers)	RxP (Reflexive Pronoun)
Category 1	50%	30%	30%	10%	10%
Category 2	80%	60%	40%	10%	20%
Category 3	60%	70%	20%	10%	30%

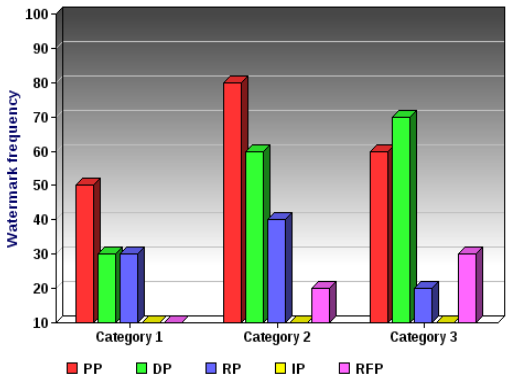

Fig. 1. Graph comparison of extracted watermarks.

The Fig. 1 illustrates the graph showing frequency comparison of watermarks for different text categories as defined above. We can observe category wise variation in occurrences of sarvanam. These observations can be more useful while studying the accuracy of retrieved watermarks in case of different possible attacks on the Devanagari text. The size of text in each category is also required to be considered as important factor during the experiments. The performance of the embedding algorithm can be influenced by the text size in each text category. Consequently it can affect the accuracy of retrieved watermarks.

The time stamping reduces possibility of unauthorized insertion. Moreover the security phrase required as input by original author at the time of document creation enhances security level of watermarking technique.

V. Conclusion and Future Work

Marathi language ranks in top 15 among all the languages across the world. But till now there does not exist any specific watermarking technique for Marathi text. We have developed a novel technique which can be applied to protect the intellectual property rights of the genuine authors. The proposed technique is specifically limited for Marathi language text. We are in progress of investigating our proposed technique against typical text watermarking attacks like insertion, deletion and reordering attack for varying text volumes. Continuing our work in future, we will be able to make this technique to excel most of the characteristics of digital text watermarking algorithm for Devanagari text. Also the proposed technique is needed to be tested for standard corpus to confirm its robustness against different types of possible attacks.

References

[1] F. A. P. Petitcolas, R. Anderson, and M. G. Kuhn, "Information hiding - A survey," in *Proc. the IEEE Special Issue on Protection of Multimedia Content*, 1999, pp. 1062-1078.

[2] F. Hartung and M. Kutter, "Multimedia watermarking techniques," in *Proc. the IEEE*, vol. 87, no. 7, pp. 1079-1107, 1999.

[3] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O. Gorman, "Electronic marking and identification techniques to discourage document copying," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1495-1504, 1995.

[4] Z. Jalil, "Copyright protection of text documents using digital watermarking," Thesis, Dept of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan, 2010.

[5] M. Topkara, C. M. Taskiran, and E. J. Delp, "Natural language watermarking," in *Proc. SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents*, 2005, pp. 441-452.

[6] Y. Zhang and H. Qin, "A novel robust text watermarking for word document," in *Proc. 3rd International Congress on Image and Signal Processing*, 2010, pp. 38-42.

[7] A. Gutub, F. Al-Haidari, K. Al-Kahsah, and J. Hamodi, "E-text watermarking: Utilizing 'Kashida' extensions in Arabic language

electronic writing," *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 1, pp. 48-55, 2010.
[8] O. Halvani, M. Steinebach, P. Wolf, and R. Zimmermann, "Natural language watermarking for German texts," in *Proc. the First ACM Workshop on Information Hiding and Multimedia Security*, 2013, pp. 193-202.
[9] H. M. Meral, E. Sevinc, E. Unkar, B. Sankur, A. Sumru Ozsoy, and T. Gungor, "Syntactic tool for text watermarking," presented at SPIE 6505, Security, Steganography, and Watermarking of Multimedia Contents IX, 2007.
[10] M. Berntsen and J. Nimbkar, *A Marathi Reference Grammar*: *South Asia Regional Studies*, University of Pennsylvania, 1975.
[11] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," in *Proc. the IEEE*, vol. 87, no. 7, pp. 1181-1196, 1999.
[12] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document identification for copyright protection using centroid detection," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 372-383, 1998.
[13] K. Bhattacharjya and H. Ancin, "Data embedding in text for a copier system," in *Proc. the IEEE International Conference on Image Processing*, 1999, pp. 245-249.
[14] S. Ranganathan, A. J. Ali, K. Kathirvel, and M. M. Kumar, "Combined text watermarking," *International Journal of Computer Science and Information Technologies*, vol. 1, no. 5, pp. 414-416, 2010.
[15] C. Culnane, H. Treharne, and A. T. S. Ho, "Improving mutli-set formatted binary text watermarking using continuous line embedding," in *Proc. 2nd International Conference on Innovative Computing, Information and Control*, 2007, pp. 287-293.
[16] M. R. Walambe, *Sugam Marathi Vyakaran-Lekhan*, Pune: Nitin Prakashan, 2011, pp. 68-75.
[17] N. Mir, "Robust techniques of web watermarking," *International Journal of Computer Science and Information Security*, vol. 9, no. 2, pp. 248-252, 2011.

Nitin N. Patil was born at Chopda, Maharashtra, India, on October 1, 1978. He has completed his master of technology in computer science & engineering, Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India in 2009 and pursuing his Ph.D. in computer engineering from North Maharashtra University, Jalgaon, Maharashtra, India.

He is working as the head and an associate professor in the Computer Engineering Department at R.C. Patel Institute of Technology, Shirpur (Maharashtra), India. He has 12 years of teaching experience. His area of research is information hiding, security particularly text and image watermarking.

Mr. Nitin N. Patil is a life member of Indian Society for Technical Education and also a life member of International Association of Computer Science and Information Technology. He has published many papers in international/national conferences and journals. He has also actively participated in numerous seminars and workshops.

Jayantrao B. Patil was born at Dahigaon Maharashtra, India, on June 11, 1963. He has completed master of technology in computer science & data processing from IIT, Kharagpur and Ph.D. in computer engineering from North Maharashtra University, Jalgaon, Maharashtra, India.

He is working as a principal at R. C. Patel Institute of Technology, Shirpur (Maharashtra), India. He has total 22 years of teaching and 2 years of industrial experience. His area of research is web catching and web prefetching, web data mining, text watermarking, web usage mining, web personalization, semantic web mining and web security.

Dr. J. B. Patil is the dean of Faculty of Engineering & Technology, the member of Academic Council, Member of Senate, and the chairman of Board of Studies in Computer Engineering and Information Technology of North Maharashtra University, Jalgaon from March 2011 onwards. He has worked as the chairman of Board of Studies in Computer Engineering and Information Technology and Member of Academic Council of North Maharashtra University, Jalgaon from January 2008 till date. He is the recognized PhD. guide in the subjects of computer engineering and electronics and telecommunication engineering of North Maharashtra University, Jalgaon and NMIMS University, Mumbai. He has published many research papers in international/national conferences and journals. He has also actively participated in numerous seminars and workshops. He is a life member of Indian Society for Technical Education (ISTE), Computer Society of India (CSI), the member of Institute of Engineers (IE), India and the senior member of International Association of Computer Science and Information Technology (IACSIT), Singapore.