A NOVEL APPROACH TO DIGITAL WATERMARKING FOR VIETNAMESE DOCUMENTS

Nguyen Van Doan, Dang Tran Khanh, Nguyen Thanh Son
University of Technology, VNU-HCM

ABSTRACT: *In this article, we propose a novel approach to digital watermarking for Vietnamese documents. The target documents need to be converted to PostScript format files before embedding information. Security information is embedded into the documents by shifting up or down a small gap of Vietnamese signs, which are special to Vietnamese documents. This embedded information can be used for copyright protection or tampering detection. The watermarked document will be converted to PDF format, containing only images instead of texts, before being published. During the detection phase, our approach does not need original documents in order to recover the watermarks. Theoretical analysis and empirical experiments show the efficiency and effectiveness of our proposed solution.*

Keywords: *Text watermarking, data hiding, copyright protection, tampering detection, Vietnamese documents.*

1.INTRODUCTION

Nowadays, digital watermarking is applied in a variety of applications such as copyright protection, tampering detection, fingerprinting, broadcast monitoring, data tracking, copy control, etc. [2,3,14]. This technique can be applied to various data formats like image, video, audio, text, XML, rational database, and many others [1,2,3,4,6,7].

Currently, text documents are widely used for information exchange. They are omnipresent in the form of newspapers, books, web pages, contracts, advertisements, cheques , identification documents, etc. At the same time, they can be widely distributed in electronic different forms via Internet communications [15]. Copyright protection and tampering detection for these text documents are indispensable problems.

Watermarking techniques for text documents can be classified according to the following methods: watermarking directly in plain text documents, watermarking in formatted text documents, and watermarking in text images. For plain text, there are three major approaches: open space based methods hide information in the document through manipulation of white space, syntactic based methods utilize punctuation, and semantic based methods employ the words themselves to do the same [5,9]. With formatted text, there are two typical approaches: first, changing the space between words, lines or letters to embed information; second, changing the format of document like font styles, names and size, etc. [2,8,10]. For text image, there is a possibility to apply methods of formatted text on a scanned document or feature coding methods [1,11,12,13]. As with Vietnamese documents, however, we use of special characteristics of the languages in order to strengthen the security degree as well as the effectiveness of watermarking.

In this article, we propose a novel approach to digital watermarking for Vietnamese documents. Security information is embedded into the documents by shifting up or down a small gap of Vietnamese signs in the documents. The security information can be used for copyright protection and tampering detection of Vietnamese documents. Our approach can also be applied to other languages that have Vietnamese like signs as Russian, Spanish, Portuguese, etc.

The rest of this article is organized as follows. Section 2 presents details of the proposed embedding phase. Section 3 presents details of the detection phase. Section 4 presents system implementation and experimental results. Our conclusions and future work are presented in the last section.

2.EMBEDDING PHASE

Vietnamese is the national and official language in Vietnam. Much vocabulary of Vietnamese has been borrowed from Chinese, and it was originally written based on the Chinese writing system. The Vietnamese writing system in use today is an adapted version of the Latin alphabet, with additional diacritics for tones and certain letters. The Vietnamese alphabet, called script of the national language is the current writing system for the Vietnamese language [16]. The Vietnamese alphabet has 29 letters as shown in Table 1. Letters 'F', 'J', 'W' and 'Z' are not part of the Vietnamese alphabet, but are used in foreign loan words.

Table 1. Vietnamese alphabet.

A	Ă	Â	B	C	D	Đ	E	Ê	G	H	I	K	L	M
a	ă	â	b	c	d	đ	e	ê	g	h	i	k	l	m
N	O	Ô	Ơ	P	Q	R	S	T	U	Ư	V	X	Y	
n	o	ô	ơ	p	q	r	s	t	u	ư	v	x	y	

Vietnamese vowels are pronounced with a tone. There are six tones in Vietnamese: level tone, hanging tone, sharp tone, asking tone, tumbling tone, heavy tone. Most of the tones are put upper the vowels; however, the heavy tone is put lower the vowels. Table 2 shows the description of Vietnamese six tones. Some Vietnamese vowels have special signs as circumflex accent on Latin alphabet as 'ă', 'â', 'ê', etc. We call the above six tones and the special signs Vietnamese signs.

Table 2. Description of Vietnamese six tones [16].

Name	Description	Diacritic	Vietnamese	English meaning
level tone	high level	(no mark)	ma	ghost
hanging tone	low falling	` (grave accent)	mà	but
sharp tone	high rising	´ (acute accent)	má	cheek, mother
asking tone	dipping-rising	̉ (hook)	mả	tomb, grave
tumbling tone	breaking-rising	~ (tilde)	mã	horse, code
heavy tone	constricted	. (dot below)	mạ	rice seedling

If Vietnamese signs are shifted up or down a small gap, the Vietnamese document will slightly be changed but it is very hard for reader by eyes to recognize the document has been changed. As a result, the secret information can be embedded into Vietnamese documents by shifting up or down a small gap of Vietnamese signs. Note that this still keeps the document's contents unchanged. This methodology could be applied to other language as Russian, Spanish, Portuguese, etc.

In Vietnamese documents, characteristic of the heavy tone is the same as punctuation mark's, and so it's hard to differentiate these signs. By this reason, they are out used in our proposed approach. For dot upper letters 'i' and 'j', they could be recognized by the body part

of these letters. Therefore, letters 'i' and 'j' could be used for embedding information in our approach.

Fig. 1 shows letters in Vietnamese documents can be employed to embed information. For letters having one sign upper like 'i', 'j', 'ă', 'â', 'ô', 'ê', 'Ă', 'Â', 'Ô', etc. the upper sign can be shifted. For letters having two signs upper like 'ằ', 'ắ', 'ẳ', 'ẵ', 'Ằ', 'Ắ', 'Ẳ', 'Ẵ', 'ầ', 'ấ', 'ẩ', 'ẫ', etc. the upmost sign is the one can be chosen to shift.

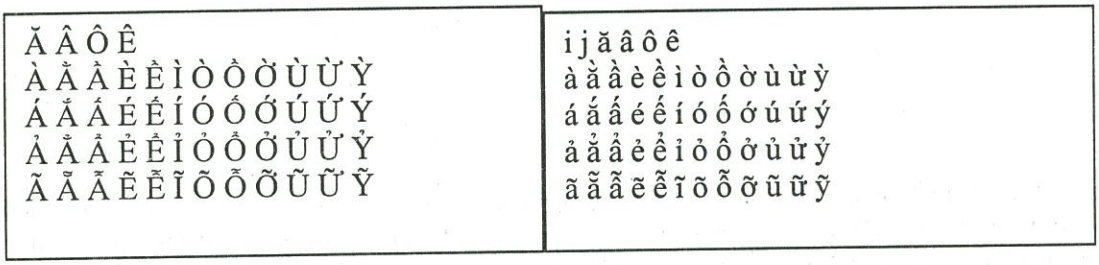


Fig. 1. Letters can be selected to embed information.

For Vietnamese documents, number of bit can be embedded depending on each word. Some words do not have any Vietnamese sign as 'anh', 'em', 'nam', etc., and thus they can not be embedded any bit. Some words as 'tháng', 'ngày', 'buồn', etc. have one letter containing a Vietnamese sign, so they can be embedded one bit. Words like 'tiếng', 'nói', 'việt', etc. can be embedded two bits because they contain two letters with special signs. Especially, words such as 'giải', 'giãi', 'giới', etc. can be embedded three bits. A statistical result for the embedding capacity of 10 casual Vietnamese documents is shown in Table 3. The second, third, and forth columns display the number of words in each document, the number of bit can be embedded into each document, and the embedding rate, respectively. We can observe that this is an encouraging result because the embedding rate (the number of embedded bits vs. the document length) is high – that is over 98% averagely. Fig. 2 graphically depicts this result more clearly.

Table 3. Capacity of embedding into Vietnamese documents.

No	Number of words	Number of embedded bits	Embedding rate (%)
1	123	128	104.07
2	74	75	101.35
3	108	96	88.89
4	206	206	100.00
5	127	128	100.79
6	170	166	97.65
7	162	155	95.68
8	184	185	100.54
9	189	179	94.74
10	77	75	97.40
Total	1420	1393	98.10

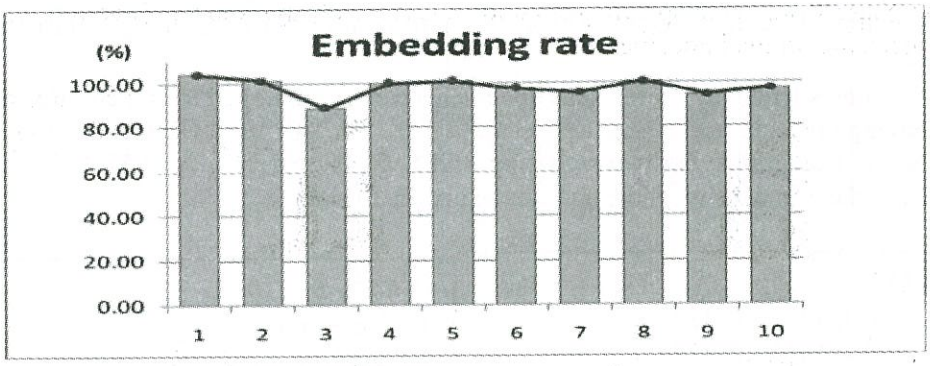


Fig. 2. Embedding rate.

Security information must be converted to binary data before embedded into Vietnamese documents. Binary data are embedded serially into a Vietnamese document from top to bottom on each line of documents. Vietnamese signs are slightly shifted up or down based on values of their corresponding watermark bits. If the watermark bit has value as 1, the sign is shifted up; otherwise, the sign is shifted down. The shifted gap for embedding job must be small enough to make small changes to the original document but spacious enough to retrieve the embedded information. In practice, a shifting gap value of 1/150 inch for Vietnamese signs can balance the two.

In our proposed approach, the PostScript format is used to embed security information into Vietnamese document. The reason why we determine to choose the PostScript format is that the underlying of this format is a page description programming language. In addition, PDF format is the popular file format in the Internet communities, so watermarked Vietnamese documents will be converted to this format, containing only image instead of texts, before being published.

To conclude this section, we summarize steps of our approach to watermark a Vietnamese document as follows: firstly, converting security information to binary data; secondly, converting the document to PostScript format; thirdly, embedding the binary data into the document as described above; finally, converting watermarked documents to PDF format.

3.DETECTION PHASE

Documents to be checked will be scanned and stored as images if we have their hard copies. Otherwise, their electronic versions can be used directly during this detection phase. The resulting images should be filtered with noises that may occur while scanning. After that, embedded information will be retrieved from the images without the original documents. The scanners can scan documents with various resolutions. If the documents are scanned with high resolution, Vietnamese signs can be recognized easily with high accuracy but the processing to recover the information need more time. If the documents are scanned with low resolution, the processing to recover the information is difficult. In practice, the scanned documents with 300 dpi resolutions are cost-efficient for a high precision detection.

Embedded information is recovered relying on absolute positions between Vietnamese signs and Latin characters. Let h, h' denote the gap between Vietnamese sign and Latin letter before and after embedding information. If h'>h, embedded data is 1, otherwise it is 0. Fig. 3 shows an example of detecting embedded information from letters 'à'. Part (a) illustrates for retrieving rule with bit 1 (h'>h) and part (b) illustrates for retrieving rule with bit 0 (h'<h).

Detection phase works with specific resolution and h are always known. Therefore, this phase does not need the original documents.

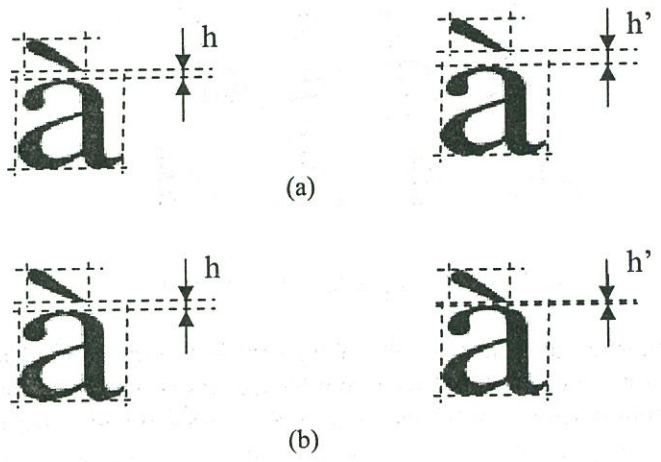


Fig. 3. Example of detecting embedded information.

For Vietnamese documents, gaps between signs and Latin letters are different depending on each signs and each letters. Therefore, to retrieve embedded information properly, this feature should be noted. Fig. 4 shows an example with letters in this case. These letters are not embedded with any security information but h1>h2>h3>h4.

In order to cope with this problem, to retrieve the security information from images, the decoder has to recognize Vietnamese signs based on a pre-built set of masks. If a sign is recognized, the decoder will calculate the gap from sign to lower vowel to determine the embedded data if have. Masks are different for different signs or the same signs but styles are different. Fig. 5 shows example masks of sharp sign with different styles.

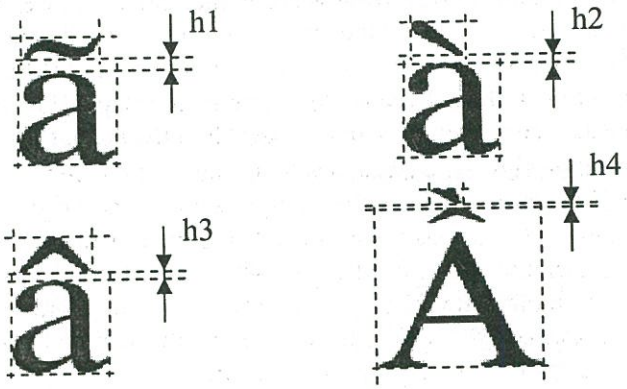


Fig. 4. Different gaps between signs and letters.

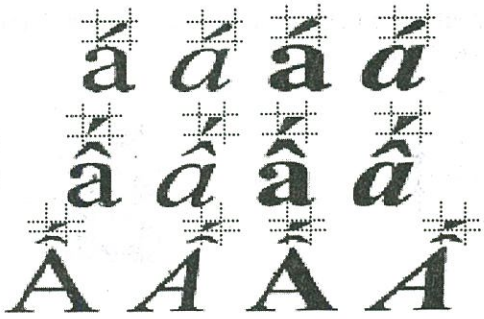


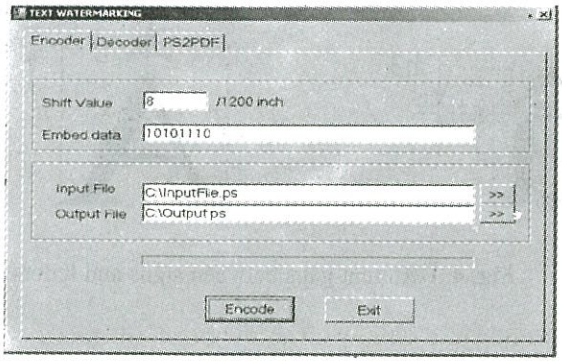
Fig. 5. Masks for sharp sign recognizing.

Overall, we summarize steps of the detection phase as follows: firstly, scanning the document and saving it as an image; secondly, filtering noises of this image; finally, detecting security information from the scanned image (to get the embedded watermark, if have).

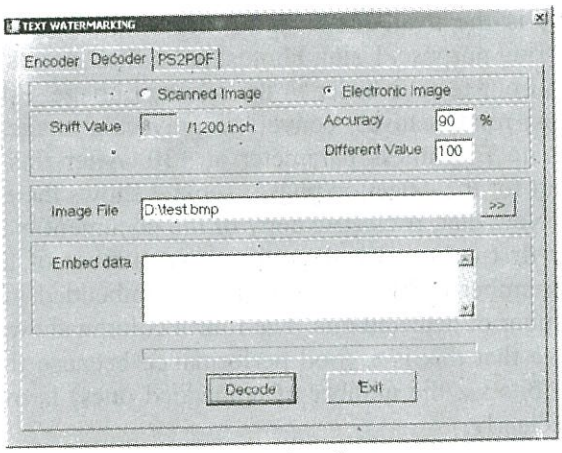
4.SYSTEM IMPLEMENTATION

To establish the practical value of our approach, we build a prototype system. The prototype works properly for Vietnamese documents in Microsoft Word format with Times New Roman font, size 13, 1.5 line spacing, and shifted value is set to 1/150 inch. If changing the size of the font or changing the font itself of the document, some program parameters need to be modified accordingly. The module for embedding information works fine. However, the decoder module only runs on normal style documents and works rather slowly. Fig. 6 shows example screenshots of the prototype system: Part (a) depicts a screenshot of the embedding phase and part (b) depicts a screenshot of the detection phase.

Microsoft Word is one of the most popular text editors so this software is chosen to create original Vietnamese documents and then they will be converted to PostScript format. To convert a Microsoft Word document to a PostScript file, there are many approaches as printing to a file by using Microsoft Word options, using Adobe Acrobat [17], etc. The PostScript codes generated from a Vietnamese Microsoft Word document using the above mechanisms would be very hard to examine and understand. We need only basic PostScript codes in order to carry out the watermarking process. DOC to Image Converter tool [18] is such a tool that is suitable for analyzing and recognizing Vietnamese signs in generated PostScript file. We used this tool in our tests to transform a Vietnamese document to a PostScript file.



a)Embedding phase



(b) Detection phase

Fig. 6 Example screenshots of the prototype system.

The PDF format is the most popular file format in the internet communities, so embedded Vietnamese documents should be converted to this format. There are some softwares to support the conversion from PostScript to PDF format document like Adobe Acrobat, Ghostscript [19], etc. Ghostscript, a popular free and open source product, is chosen in all our tests. Fig. 7 shows an example of Vietnamese text that was tested with the prototype system (cf. Appendix). Part (a) illustrates the original text, and part (b) illustrates the embedded text.

Trường Đại Học Bách Khoa Tp. Hồ Chí Minh là một trung tâm đào tạo cán bộ kỹ thuật công nghệ và các nhà quản lý có trình độ ngang tầm với các nước tiên tiến trong khu vực Đông nam Á, đáp ứng nguồn nhân lực có chất lượng cao cho sự nghiệp công nghiệp hóa và hiện đại hóa đất nước cũng như khu vực phía Nam.

Original text

Trường Đại Học Bách Khoa Tp. Hồ Chí Minh là một trung tâm đào tạo cán bộ kỹ thuật công nghệ và các nhà quản lý có trình độ ngang tầm với các nước tiên tiến trong khu vực Đông nam Á, đáp ứng nguồn nhân lực có chất lượng cao cho sự nghiệp công nghiệp hóa và hiện đại hóa đất nước cũng như khu vực phía Nam.

Embedded text

Fig. 7. An example of Vietnamese text.

Documents to be examined are scanned and stored as 8-bits grayscale images with 300 dpi resolution. The resulting images should be filtered with noises that may occur while scanning. Building a noisy filter application requires a good algorithm which is usually rather

complicated. Photoshop [20] is the available image processing software with many good functions suitable for the filtering purpose. Using Photoshop to filter noises is quite convenient and it is also employed for all our tests in order to the scanned image. Filtering noises using Photoshop requires four steps: rotate canvas, convert to grayscale image, calibrate Contrast parameter at +30, and calibrate Brightness parameter at +30. After this step, the resulting image can be used as the input for the detection phase. Table 4 shows the experimental results of eight scanned documents. Note that, if electric copy of the document is used instead, 100% of the bits are detected. The first column displays the order number of the documents. The second column displays the number of bits which had been embedded into documents. The third column displays the number of detected bits. The fourth column shows the correct rate of the detection. We can observe that this is a good performance because the correct detection rate (the number of detected bits vs. the number of embedded ones) is over 90% averagely. Fig. 8 shows this result more clearly.

Table 4. Experimental results with Vietnamese documents.

No	Number of embedded bits	Number of detected bits	Correct rate (%)
1	83	79	95.18
2	56	56	100.00
3	62	61	98.39
4	72	68	94.44
5	57	57	100.00
6	65	64	98.46
7	42	39	92.86
8	44	40	90.91
9	58	56	96.55
10	64	61	95.31
Total	603	581	96.35

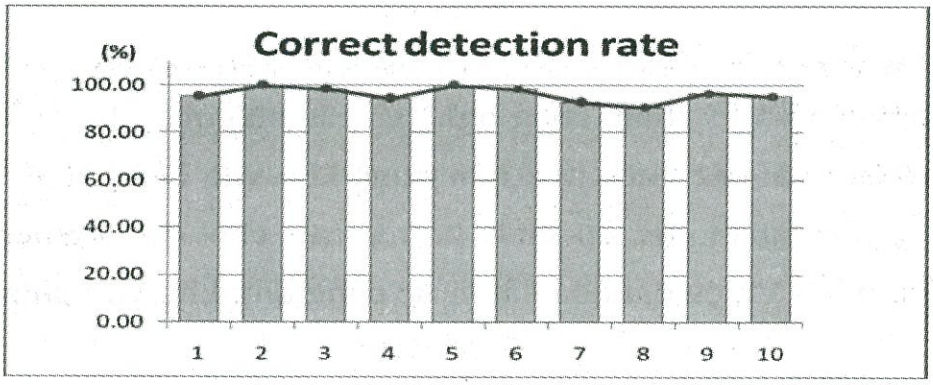


Fig. 8 Correct detection rate.

5.CONCLUSION AND FUTURE WORK

This article proposes a novel approach using Vietnamese signs to digital watermarking for Vietnamese documents. Security information can be extracted for both copyright protection and tempering detection purposes. The embedded document is nearly the same as the original one, and so the reader could not find any differences by eyes. To the best of our knowledge, the proposed approach is the vanguard solutions to digital watermarking for Vietnamese document that employing specific aspects of the language. Our theoretical analyses and empirical experiments have established the practical value of the approach.

In the future, embedding more information to Vietnamese documents by using letters with two signs, heavy tone, and implementing a mechanism for extracting security information effectively from Vietnamese documents (including scanned documents) such as using neural network are also of our great interest. Moreover, building security information with fault tolerance is a challenging issue and it is also among our planned future work. Last but not least, applying this research result to real-world application domains such as Vietnamese digital libraries is also one of our most interesting concerns.

Appendixes

Fig. 5's Vietnamese text is translated to English version:

HCMUT is a centre of technology industry and management training. Graduate students from HCMUT have strong professional skills, which are the same level as those of other developed countries in the Southeast Asia. Our training activities have made remarkable contributions to satisfy the requirements of man power for the industrialization and modernization in Vietnam generally and Southern Vietnam areas particularly.

PHƯƠNG PHÁP TIẾP CẬN MỚI TRONG VIỆC ỨNG DỤNG KỸ THUẬT THỦY VÂN SỐ (DIGITAL WATERMARKING) TRÊN CÁC TÀI LIỆU TIẾNG VIỆT

Nguyễn Văn Đoàn, Đặng Trần Khánh, Nguyễn Thanh Sơn
Trường Đại học Bách Khoa, ĐHQG-HCM

TÓM TẮT: *Trong bài báo này chúng tôi đề xuất một phương pháp tiếp cận mới trong ứng dụng kỹ thuật thủy vân số (digital watermarking) trên các tài liệu tiếng Việt. Cách tiếp cận này khai thác đặc điểm đặc biệt của chữ viết tiếng Việt để nhúng thông tin bảo mật vào tài liệu. Thông tin bảo mật được sử dụng để chứng minh bản quyền sở hữu hoặc chống giả mạo. Các tài liệu tiếng Việt phải chuyển sang định dạng PostScript trước khi tiến hành nhúng thông tin bảo mật. Thông tin bản mật được nhúng vào tài liệu thông qua việc dịch chuyển lên hoặc xuống các dấu tiếng Việt một khoảng cách rất nhỏ. Dữ liệu sau khi nhúng thông tin bảo mật sẽ được chuyển sang định dạng PDF, ở dạng hình ảnh, trước khi gửi đi hoặc công bố rộng rãi. Quá trình xử lý để lấy lại thông tin bảo mật từ tài liệu không cần phải sử dụng tài liệu tiếng Việt chưa nhúng thông tin ban đầu. Ngoài ra, chúng tôi cũng trình bày một số kết quả thực nghiệm để chứng minh cho sự hiệu quả của cách tiếp cận này.*

REFERENCES

[1]. F. Hartung, and M. Kutter, *Multimedia Watermarking Techniques,* Proceedings of the IEEE, Vol. 87, pp. 1079-1107, (1999).

[2]. M. Arnold, M. Schmucker, and S. D. Wolthusen, *Techniques and Applicaitons of Digital Watermaking and Content Protection,* ISBN 1-50853-111-3, Artech House, (2003).

[3]. M. J. Cox, M. L. Miller, J. A. Bloom, *Digital Watermarking*, ISBN 1-55860-714-5, Morgan Kaufmann Publishers, (2002).

[4]. C. S. Lu, *Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property*, ISBN 1-59140-193-3, Idea Group Publishing, (2004).

[5]. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, *Techniques for data hiding,* ISSN 0018-8670, IBM Systems Journal, (1996).

[6]. S. Inoue, K. Makino, I. Murase, O. Takizawa, T. Matsumoto, and H. Nakagawa, "A Proposal on Information Hiding Methods using XML", Proceedings of The First NLP and XML Workshop, pp. 55-62, 2001.

[7]. J. Su, F. Hartung, and B. Girod, *Digital Watermarking of Text, Image, and Video Documents,* Computers & Graphics, (1999).

[8]. J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, *Electronic Marking and Identification Techniques to Discourage Document Copying*, Proceedings of IEEE INFOCOM'94, Vol. 3, pp. 1278-1287, (1994).

[9]. D. Salomon, *Data Privacy and Security,* ISBN 0-387-00311-8, Springer Verlag, (2003).

[10]. J. Brassil, S. Low, N. F. Maxemchuk, L. O'Gorman, *Hiding Information in Document Images,* Proceedings of the 29th Annual Conference on Information Sciences and Systems, pp. 482-489, (1995).

[11]. R. Villán, S. Voloshynovskiy, O. Koval, J.E. Vila-Forcén, E. Topak, F. Deguillaume, Y. Rytsar, and T. Pun, *Text Data-Hiding for Digital and Printed Documents: Theoretical and Practical Considerations,* Proceedings of the SPIE, Vol. 6072, pp. 406-416, (2006).

[12]. D. Huang, and H. Yan, *Interword distance changes represented by sine waves for Watermarking Text Images,* School of Electrical and Information Engineering University of Sydney, (2006).

[13]. A.M. Alattar, and O. M. Alattar, *Watermarking Electronic Text Documents containing Justified Paragraphs and Irregular Line Spacing,* Proceedings of SPIE, Vol. 5306, pp. 685-695, (2004).

[14]. S. Katzenbeisser, and F. A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*, ISBN 1-58053-035-4, Artech House, (2000).

[15]. S. Voloshynovskiy, O. Koval, R. Villan, E. Topak, J. V. Forcén, F. Deguillaume, Y. Rytsar, and T. Pun, *Information-theoretic analysis of electronic and printed document authentication,* Proceedings of the SPIE, Vol. 6072, pp. 516-535, (2006).

[16]. *Wikipedia, the free encyclopedia*, http://www.wikipedia.org.

[17]. *Adobe - Adobe Acrobat Family - Create PDF file, Edit PDF file, Convert PDF to word, Convert PDF to doc*, http://www.adobe.com/products/acrobat/.

[18]. *Convert Word Doc RTF to Jpeg/Jpg/Tiff/Bmp/Eps/Ps, Doc to Image Converter*, http://www.pdf-convert.com/doc2img/.

[19]. *Ghostscript: Ghostscript Website*, http://www.ghostscript.com/awki.

[20]. *Professional photo editing software - the digital imaging software standard by Adobe*, http://www.adobe.com/products/photoshop/.