PAPER Special Section on Enriched Multimedia—Making Multimedia More Convenient and Safer—

Robust and Secure Data Hiding for PDF Text Document

Minoru KURIBAYASHI[†a)], ***Senior Member***, **Takuya FUKUSHIMA**[†], ***Student Member***, ***and*** **Nobuo FUNABIKI**[†], ***Member***

SUMMARY The spaces between words and paragraphs are popular places for embedding data in data hiding techniques for text documents. Due to the low redundancy in text documents, the payload is limited to be small. As each bit of data is independently inserted into specific spaces in conventional methods, a malicious party may be able to modify the data without causing serious visible distortions. In this paper, we regard a collection of space lengths as a one-dimensional feature vector and embed watermark into its frequency components. To keep the secrecy of the embedded information, a random permutation and dither modulation are introduced in the operation. Furthermore, robustness against additive noise is enhanced by controlling the payload. In the proposed method, through experiments, we evaluated the trade-off among payload, distortion, and robustness.

key words: *portable document format, quantization index modulation, dither modulation, permutation*

1. Introduction

The portable document format (PDF) [1] has been developed by Adobe Systems Society as a page description language that can preserve the formatting of a file. Because of its popularity, some data hiding methods for PDF files have been studied. Zhong et al. [2] change the preset distance between each word in PDF files. In [3], the spaces between words and paragraphs are used for embedding, using an extension of a basic embedding algorithm [4]. In a PDF file, a TJ operator displays text strings with the position and space lengths between characters. Therefore, it is easy to apply the method to a PDF file. The main drawback of this method is there is one-to-one correspondence between the space length and embedding bit. A malicious party will be able to edit the embedding watermark by modifying specific space lengths. As the hexadecimal text encoding of "20" and "A0" are displayed as blank characters in PDF files, such specific characters are used for embedding a watermark in [5]. Lin et al. [6] proposed an interesting method based on PDF files of ISO-8859 encoding. It combines a data hiding method for PDF files and an encryption method based on quadratic residue. However, the method is adversely affected by the removal of the inserted watermark. The spread transform dither modulation (STDM) is used for embedding information into PDF file in [7]. Each bit of the watermark is embedded into some x-coordinate (horizontal position) values. Because of its spreading effects, the method retains transparency and robustness. However, the payload is small. In [8], the object in each line of the PDF files is divided into some groups according to the watermark. The advantage of the method is that no visual distortion appears on the display of a PDF viewer. However, the payload is not large, and a malicious party can be easily find the irregularities in the watermarked PDF file if the descriptions of the document are observed.

In this paper, we propose a new data hiding method for a PDF file with large payload and low distortion, by embedding a watermark into the spaces among characters in each line. Unlike the conventional methods, the proposed method regards a collection of the space lengths as a host vector and embed a watermark into its frequency components. The preliminary version of this paper was presented at IIHMSP2017 [9]. To control the changes of space lengths induced by the embedding, the watermark is embedded into the frequency components converted from the vector by discrete cosine transform (DCT). In order not to change the sum of the space lengths in each line, the watermark is embedded only into its alternate current (AC) components, not its direct current (DC) component. To ensure the secrecy of the watermark, we use the dither modulation quantization index modulation (DM-QIM) [10], which introduces randomness to the embedding process based on a secret key. In addition, before performing the DCT, the host vector is permuted using a secret key to enhance the secrecy of the watermark. From the spectral efficiency point of view, the proposed method can be regarded as orthogonal frequency division multiplexing (OFDM). Each watermark bit is transmitted (embedded) by means of each frequency component as a carrier, and each carrier is a DCT basic vector. As DCT basic vectors are mutually orthogonal, watermark information is transmitted over multiple carrier frequencies, namely it is DCT-OFDM. The secrecy of the carried information is maintained well by the combination of the random permutation and the DM-QIM method. This combination is the most important advantage of the proposed method.

The rest of this paper is organized as follows. In Sect. 2, we briefly review the data hiding method and the structure of the PDF document file. In Sect. 3, the basic method is presented and the enhancement of robustness is considered. The experimental results are shown in Sect. 4. Finally, we

Manuscript received March 12, 2018.
Manuscript revised July 26, 2018.
Manuscript publicized October 19, 2018.

[†]The authors are with the Graduate School of Natural Science and Technology, Okayama University, Okayama-shi, 700–8530 Japan.
a) E-mail: kminoru@okayama-u.ac.jp

DOI: 10.1587/transinf.2018MUP0003

conclude this paper in Sect. 5.

2. Preliminaries

2.1 Data Hiding

To embed a watermark into digital content, the extraction of features is very important in terms of transparency as well as the secrecy of data hiding method. In the case of a digital image, the common feature is frequency components. For security reasons, some components are selected as a host vector for embedding the watermark according to a secret key. A data hiding algorithm takes a host vector, watermark, and a secret key as input, and outputs a watermarked vector. Typical algorithms are the spread spectrum method [11] and quantization index modulation (QIM) method [10].

Let $w \in \{0, 1\}$ be a watermark bit, and δ be a quantization step size to control the distortion level. Assume that d is an element of the selected host signal and it is a real number. In the QIM method [10], d is rounded to the nearest odd/even quantized value according to the watermark bit w using step size δ. Typically, the dither modulation (DM) is combined with the QIM method to prevent artifacts induced by quantization. Let k be an embedding key which value is a real number chosen randomly from a uniform distribution over $[-\delta/2, \delta/2]$. The embedding operation is performed as follows.

$$d' = \begin{cases} \delta \cdot \lfloor \frac{d+k}{\delta} \rfloor & \text{if } \lfloor \frac{d+k}{\delta} \rfloor \bmod 2 = w \\ \delta \cdot (\lfloor \frac{d+k}{\delta} \rfloor + 1) & \text{otherwise} \end{cases}. \quad (1)$$

The above operation quantizes $d+k$ into the nearest even/odd value according to the watermark bit w. If $w = 0$, $\lfloor d'/\delta \rfloor$ becomes an even value; otherwise, it becomes an odd value. Finally, k is subtracted from d' to obtain the output $d^\star$.

$$d^\star = d' - k. \quad (2)$$

Without the dither modulation, the quantized value becomes a multiple of δ. For example, if $\delta = 10$, the quantized value must be $\{0, \pm 10, \pm 20, \pm 30, \cdots\}$. If a malicious party knows the embedding algorithm, the DCT coefficients selected for embedding can be found by simply observing the values.

Note that $d^\star$ is not a multiple of δ. As the watermark bit w cannot be extracted without k, it is regarded as a secret key in this method. For convenience, we denote the above embedding algorithm as

$$d^\star = \text{DM-QIM}(d, w, k), \quad (3)$$

where $k \in [-\delta/2, \delta/2]$.

At the detection, the watermark w is extracted as follows.

$$w = \left\lfloor \frac{d^\star + k + \frac{\delta}{2}}{\delta} \right\rfloor \bmod 2. \quad (4)$$

```
10obj
<</Length 10>
stream

BT
/F1 24 Tf
10 500 Td
[(Hello)]TJ
0 -100 Td
[(He)-50(llo)]TJ
ET

endstream
endobj
```

Fig. 1 Example of text syntax in object.

2.2 PDF File Structure

A PDF file has a structured binary file format with four components: header, body, cross-reference table, and trailer.

The header is the first line of the PDF file and indicates the version of the PDF specification. In the body, there are objects including text streams, images, other multimedia elements, etc. The body is used to hold all the data of the documents to be shown on a PDF viewer. The cross-reference table contains the references to all the objects in the document. It allows random access to objects in the file. Each object is represented by one entry in the table. The trailer is used to find the cross-reference Table, which is like a dictionary indicating the link to each object.

An example of PDF syntax in an object is shown in Fig. 1. Each object has its unique number, and it appears between "obj" and "endobj". The script and data for displaying text, figures, and images appear between "stream" and "endstream". "BT" and "ET" represent begin text and end text, respectively. There are some operators to represent the text document. The "Tf" operator specifies the text style and font size, and the "Td" operator specifies the offset of the beginning of the current line. When a current coordinate is (x, y), the Td operator shifts the coordinate into $(x + \Delta x, y + \Delta y)$. The "TJ" operator shows the text characters and spaces between characters in which a number is expressed in thousandths of a unit of text space. The unit of text space is determined by a proportional font used in the file and its size specified by the Tf operator. In this example, the syntax "[(He)−50(llo)]TJ" represents the characters "He" and "llo" and the space lengths "−50", which means the space between them is shortened by 50/1000 unit of text space. Generally, the spaces between characters appear in PDF files when a proportional font is used.

3. Proposed Data Hiding Method for PDF File

The proposed method regards a collection of space lengths in each line as one dimensional signal, and embed a watermark into its frequency components.

3.1 Basic Method

Assume that the number of spaces among characters in the t-th line, denoted by ℓ_t, is more than 1. The collection of

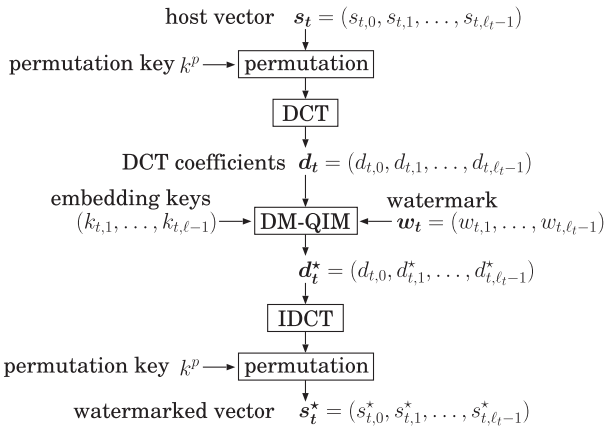


Fig. 2 Procedure of the proposed embedding operation.

space lengths in the t-th line is denoted by a vector $\boldsymbol{s_t} = (s_{t,0}, s_{t,1}, s_{t,2} \ldots, s_{t,\ell_t-1})$, we call it host vector. The proposed method enables us to embed an at most (ℓ_t-1)-bit watermark into $\boldsymbol{s_t}$.

Let $\boldsymbol{w_t} = (w_{t,1}, w_{t,2}, \ldots, w_{t,\ell_t-1})$ be a $(\ell_t - 1)$-bit watermark for the t-th line, where $w_{t,j} \in \{0, 1\}$, $(1 \le j \le \ell_t - 1)$. The following operation is performed to embed the watermark.

1. Extract a host vector $\boldsymbol{s_t}$ from a t-th line.
2. Randomly permute the elements of $\boldsymbol{s_t}$ according to permutation key k^p, where the superscript p stands for the permutation. The permuted vector is denoted by $\boldsymbol{s^p_t}$.
3. Perform DCT on the vector $\boldsymbol{s^p_t}$. The DCT coefficients are denoted by $\boldsymbol{d_t} = (d_{t,0}, d_{t,1}, d_{t,2}, \ldots, d_{t,\ell_t-1})$.
4. Embed the j-th watermark bit $w_{t,j}$ into the corresponding DCT coefficient $d_{t,j}$ for $1 \le j \le \ell_t - 1$ using the DM-QIM method with an embedding key $k_{t,j}$. The watermarked DCT coefficients are denoted by $\boldsymbol{d^\star_t} = (d_{t,0}, d^\star_{t,1}, d^\star_{t,2}, \ldots, d^\star_{t,\ell_t-1})$, where
$$d^\star_{t,j} = \text{DM-QIM}(d_{t,j}, w_{t,j}, k_{t,j}). \tag{5}$$
5. Perform IDCT on $\boldsymbol{d^\star_t}$, and permute the result in the inverse order to obtain $\boldsymbol{s^\star_t}$ by using k^p.
6. Replace the original space lengths in the t-th line with $\boldsymbol{s^\star_t}$.

Figure 2 illustrates the above operation. Note that the DC component $d_{t,0}$ is not used for embedding in the above operation. This ensures that the total length of the spaces is unchanged after the embedding, namely, $\sum s_{t,j} = \sum s^\star_{t,j}$. If the total length is changed, irregularities will clearly appear at the end of the lines, which will become noticeable distortions. By excluding the DC component, we avoid such distortions. The payload P of the above embedding method is
$$P = \sum_{t=1}^{n} (\ell_t - 1), \tag{6}$$
where n is the number of lines in the PDF file.

3.2 Security

Because of the random permutation and the DM-QIM method, it is difficult for a malicious party to modify the embedded watermark without the secret keys k^p and $k_{t,j}(1 \le t \le n, 1 \le j \le \ell_t - 1)$. Even if the number of spaces in a certain line is small, the DM-QIM method prevents a malicious party from getting useful information about the permutation and watermark.

For security reasons, the proposed method uses secret keys k^p and $k_{t,j}(1 \le t \le n, 1 \le j \le \ell_t - 1)$. It seems that the number of secret keys increases the management costs in the proposed method. However, it is possible to use a single secret key as a security parameter. We assume that a secure pseudo-random number generator (PRNG) is used to generate the secret keys k^p, and $k_{t,j}$. Without the secret key, the direct observation of (watermarked) host vector is difficult in the proposed method. Hence, attacker gets no useful information even when the algorithm of the method is revealed.

Due to the specification of PDF file, each element of host vector is integer, $s_{t,j} \in \mathbb{Z}$. Similar to the case of digital image, the elements obtained after inverse frequency transformation should be rounded into integers. Hence, $s^\star_{t,j}$ is rounded to its nearest integer. Even if no attack is performed to the watermarked PDF file, noise must be induced by the rounding operation. Such noise can be immunized by properly designing the step δ in the proposed method.

3.3 Enhancement of Robustness

A PDF file can be protected against modification by setting password. In such a case, it is difficult for attackers to delete/modify the embedded watermark. However, in some cases, no protection is done by mistake at a storage. Here, we assume that an attacker tries to modify the spaces slightly so as to fool a watermark detector without serious perceptual degradation to the PDF file.

It is possible for the proposed method to be robust against additive noise by reducing the bit-length of the watermark. Without the secret key used for the permutation, the noise energy will be spread over all DCT coefficients because direct modification of DCT coefficients is difficult. Specifically, a malicious party will be able to observe the watermarked space lengths $\boldsymbol{s^\star_t}$, not the watermarked DCT coefficients $\boldsymbol{d^\star_t}$ without k^p. If noise $\boldsymbol{\epsilon} = (\epsilon_0, \epsilon_1, \ldots, \epsilon_{\ell-1})$ is added to $\boldsymbol{s^\star_t}$, the energy $\sum_{j=0}^{\ell-1} \epsilon_j^2$ will be spread over $\boldsymbol{d^\star_t}$ after the transformation. If the random permutation is ideal, the observed noise at the DCT coefficients is approximated to be additive white Gaussian noise (AWGN). Due to the short length of spaces ℓ_t in each line, it is difficult to realize such an ideal permutation and have such a spreading effect. Nevertheless, the noise energy cannot directly affect the watermark signal in the proposed method.

Let $R(0 < R \le 1)$ be the parameter controlling the

embedding rate. Then, the payload becomes RP. In the case of $R = 1/2$, the bit-length of the watermark is $(\ell_t - 1)/2$ at the t-th line, and then, only half of the noise energy is expected to affect the embedded signal, which reduces the effects of the additive noise. By properly selecting the quantization step size δ and the rate R, it is possible to control the trade-off between perceptual distortions and robustness.

The proposed method multiplexes a $R(\ell_t - 1)$-bit watermark to embed it into vector $\boldsymbol{s_t}$ with ℓ_t elements, while each watermark bit is separately spread by the STDM. Since the watermark is embedded into DCT coefficients derived from $\boldsymbol{s_t}$ using secret keys, the carriers of the watermark bits are orthogonal with each other. Therefore, the proposed method is a kind of OFDM method that has the advantage of allowing high spectral efficiency.

4. Experimental Results

4.1 Condition

The text is selected from the first 5 chapters of Genesis in the Old Testament of the Bible, and the PDF file is created by a Latex file by using an extended version of dvipdfm-0.13.2c. The created PDF file has 3178 words with 15694 characters and 150 lines. Since the document of PDF file is compressed, we use the "PDFtk" toolkit[†] to decompress the file before embedding the watermark. The size of the original PDF file is 24767 bytes, while the size becomes 52445 bytes after decompression. The embedding payload in the PDF file is $P = 3810$ bits. We use 1000 randomly selected binary sequence patterns as a watermark, and embed them into the PDF file using quantization step size $\delta = \{3, 6, 12, 24, 48\}$. As the number of spaces ℓ_t is not always sufficiently large, we select the line for embedding that has $\ell_t \geq 10$ spaces.

4.2 Transparency and Payload

First, we consider the transparency under embedding rate $R = 1$. Figure 4 shows the screen-captured image of the watermarked PDF files. In the case of $\delta = 3$, the changes in the spatial domain are negligible, and it is hard to perceive them. Due to the permutation and DM-QIM method using a secret key, a malicious party will not be able to find any information from observing the frequency components without the secret key. Even in the case of $\delta = 48$, no visual degradation of the figure is observed. Although the changes becomes large with increasing δ, no irregularity is observed at the descriptions of the watermarked PDF file. In case of $\delta = 96$, visible distortion appears at some words. For instance, in Fig. 5 it is observed from the 3rd line that the word "light" coming up three times looks different. The spaces of 1st and 3rd ones are larger than the 2nd one. It is because of the spaces between the characters "h" and "t" are different. As the watermark energy is spread over spaces in one line, the control of the changes is not easy. In this experimental

[†]https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/

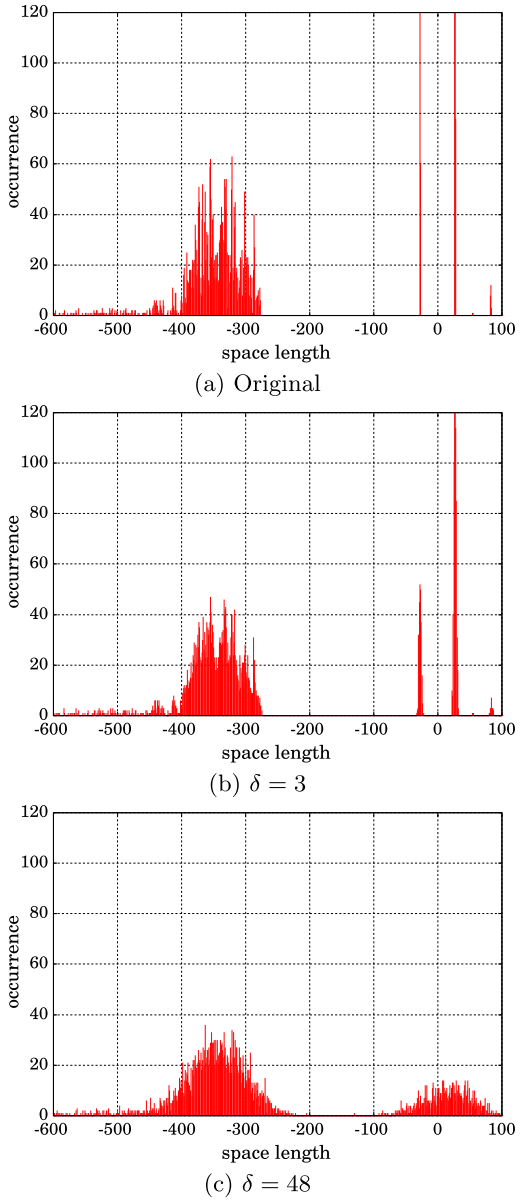


Fig. 3 Changes in the histogram of space lengths after embedding.

condition, the transparency of the watermarked PDF files is not assured if δ exceed 48. However, it is strongly dependent on one's subjective view. The investigation of objective measurement is left for our future work.

We generate PDF files by using 1000 different watermarks, and measure the average file sizes in the experiment, which results are shown in Table 1. As expected, the size of the compressed file increases with increasing step size δ because the distortion induced by embedding increases accordingly. From the information theoretical point of view, the embedding operation should increase the amount of entropy in the PDF files. On the other hand, the size of the uncompressed file is slightly decreased with the increase of δ. It is because of the statistical distribution of the space lengths in the PDF file used in this experiment. Figure 3

1 In the beginning God created the heaven and the earth. 2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. 3 And God said, Let there be light: and there was light. 4 And God saw the light, that it was good: and God divided the light from the darkness. 5 And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day. 6 And God said, Let there be a firmament in the midst of the waters, and

(a) Original

1 In the beginning God created the heaven and the earth. 2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. 3 And God said, Let there be light: and there was light. 4 And God saw the light, that it was good: and God divided the light from the darkness. 5 And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day. 6 And God said, Let there be a firmament in the midst of the waters, and

(b) $\delta = 3$

1 In the beginning God created the heaven and the earth. 2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. 3 And God said, Let there be light: and there was light. 4 And God saw the light, that it was good: and God divided the light from the darkness. 5 And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day. 6 And God said, Let there be a firmament in the midst of the waters, and

(c) $\delta = 48$

1 In the beginning God created the heaven and the earth. 2 And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. 3 And God said, Let there be light: and there was light. 4 And God saw the light, that it was good: and God divided the light from the darkness. 5 And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day. 6 And God said, Let there be a firmament in the midst of the waters, and

(d) $\delta = 96$

Fig. 4 Comparison of visual difference. PDF files are opened by Adobe Acrobat Reader DC version 2015.023.20070, and first 5 lines of document are screen-captured.

Table 1 Comparison of file size [bytes] when 3810-bit watermark is embedded.

	uncompressed	compressed
original	52445	24767
$\delta = 3$	52448.00	26128.84
$\delta = 6$	52448.00	26638.56
$\delta = 12$	52445.38	27091.19
$\delta = 24$	52372.03	27440.37
$\delta = 48$	52369.07	27802.03

light (a) 1st light (b) 2nd light (c) 3rd

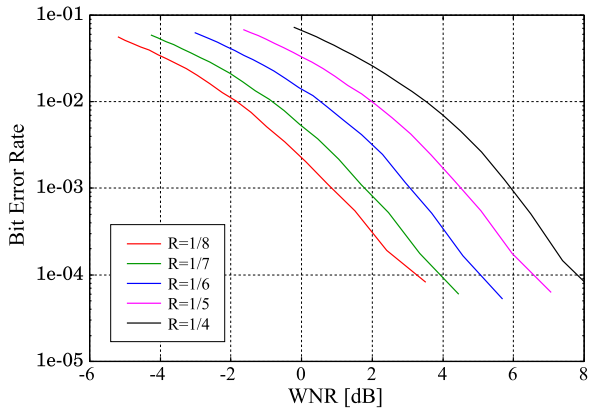
Fig. 5 Example of visual distortions appeared in a PDF file, where $\delta = 96$ and the term "light" appeared at 3rd line in the document.

shows the histogram of the space lengths. It is observed from Fig. 3 (a) that the space lengths are distributed over less than -280 and three spikes (around ± 27 and 83) are appeared. After embedding watermarks, the space lengths are increased or decreased while the average value is unchanged in the proposed method. In case of $\delta = 3$, the changes of space lengths at these spikes are small, and hence, the number of digits for representing the space lengths are not changed. On the other hand, in case of $\delta = 48$, some space lengths at two spikes (around ± 27) are changed within the range $[-9, 9]$. As the number of digits (e.g. 2-digit for 27 and 3-digit for -27) becomes small in such a case, the file size of uncompressed PDF file becomes smaller from the original. With the increase of δ, the change of digits are increased. In order to make the statistical analysis performed by attackers difficult, it can be said that the size of δ should be small from the above consideration. Note that the statistical distribution of space lengths is strongly dependent on the process how a PDF file is created.

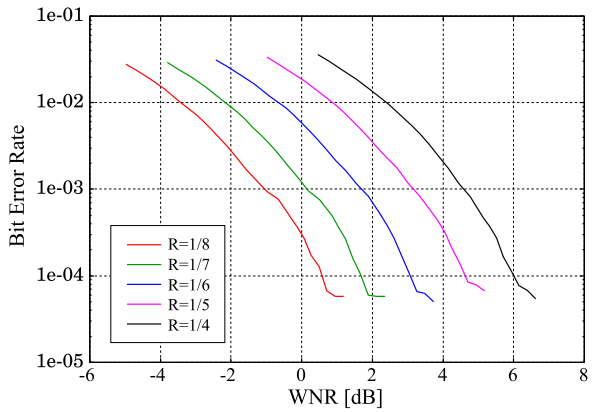
4.3 Robustness

Next, we evaluate the robustness against additive noise using the watermark-to-noise ratio (WNR). The WNR is calculated by using the ratio between the total amount of energy embedded into a PDF and the energy of AWGN:

$$WNR = \frac{total\ amount\ of\ watermark\ energy}{total\ amount\ of\ noise\ energy}$$



(a) $\delta = 12$



(b) $\delta = 48$

Fig. 6 Comparison of robustness against additive noise for different WNR.

$$= \frac{\sum_{j=0}^{\ell-1}(s_j^{\star} - s_j)^2}{\sum_{j=0}^{\ell-1}\epsilon_j^2} \tag{7}$$

The embedding rate R is changed from 1/4 to 1/8 in the experiment (Fig. 6). It is observed that the bit error rate (BER) is improved with decreasing embedding rate R. The dependency on the quantization step δ is also evaluated, which results are shown in Fig. 7.

With increasing δ, the robustness against additive noise is improved because the robustness in the QIM method is controlled by the quantization step size. As a consequence, we can control the robustness by sacrificing the payload and the transparency, namely, by changing the parameters R and δ. The selection of suitable values are strongly dependent on its application and original content because of the difficulty in the objective measurement for the transparency.

Due to its simplicity, the DM-QIM method is used in this paper, but the use of sophisticated methods, such as the distortion-compensated QIM method is possible. The use of such methods is our future work.

4.4 Comparison

Although conventional data hiding methods for text document [12], [13] considers the spaces between lines and words, the available host signal in a PDF file is different. Even if these methods are used in a PDF file, the selected host signal retains one-to-one correspondence between the signal and embedding bit, hence the secrecy is not kept in these methods.

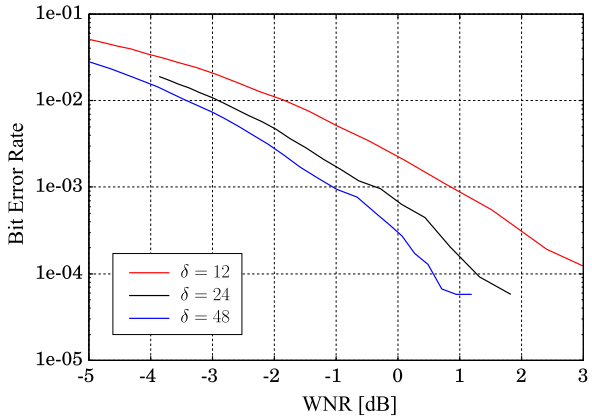


Fig. 7 Comparison of robustness against additive noise for different quantization step δ.

The positions indicated by the TD operator are changed by using the STDM for the QIM method in [7]. Although one TD operator usually appears at the first character of each line, the experiments are demonstrated under the assumption that each character has its respective TD operator, which is not normal for a PDF file. If the same text document as our experiment is used, the payload is much smaller. In addition, a malicious party will be able to find the irregularity at x-coordinates of characters modified by embedding. For instance, the left side at each line will slightly fluctuate in the left/right direction. As the number of lines is 150 in this experiment, the payload of the method [7] must be less than 150 bits in this case.

It is possible to use the space lengths specified by the TJ operator in the method [7]. However, the sum of the space lengths at each line will change after the embedding, so a malicious party can easily find the irregularity. On the other hand, the proposed method does not change the sum even if the strength of the watermark signal is large.

In [8], each line represented by the TJ operator is divided into two objects to embed a 1-bit watermark. When a 150-bit watermark is embedded into 150 lines, the file size becomes 53183 and 25107 bytes for uncompressed and compressed files, respectively, which is much worse than when using the proposed method.

5. Conclusion

In this paper, we proposed a novel data hiding method for PDF files that regards a collection of space lengths in each line as a one-dimensional signal. According to a secret key, randomized frequency components are used for embedding watermark with less distortion and high secrecy. By avoiding the DC component, the total length of each line is kept

unchanged to control the distortions induced by the embedding. We also consider the secrecy of the watermark by introducing the permutation and dither modulation according to a secret key. The robustness against attacks can be controlled by changing the embedding rate.

Although we use English documents, it is possible to embed watermark into documents written in other languages with a proportional font in which different characters have different pitches. The difference of space lengths between characters is available for the proposed method.

Acknowledgments

This research was partially supported by JSPS KAKENHI Grant Number JP16K00185.

References

[1] A.S. Incorporated, "Document management — portable document format — part 1: PDF 1.7." ISO 32000-1:2008, July 2008.
[2] S. Zhong, X. Cheng, and T. Chen, "Data hiding in a kind of PDF texts for secret communication," Int. J. Network Security, vol.4, pp.17–26, 2007.
[3] L.Y. Por and B. Delina, "Information hiding: a new approach in text steganography," Proc. ACACOS'08, pp.689–695, 2008.
[4] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," IBM Syst. J., vol.35, no.3-4, pp.313–336, 1996.
[5] I.-S. Lee and W.-H. Tsai, "A new approach to covert communication via PDF files," Signal Processing, vol.90, no.2, pp.557–565, 2010.
[6] H.F. Lin, L.W. Lu, C.Y. Gun, and C.Y. Chen, "A copyright protection scheme based on PDF," Int. J. Innovative Computing, Information and Control, vol.9, no.1, pp.1–6, 2013.
[7] A.W. Bitar, R. Darazi, J.-F. Couchot, and R. Couturier, "Blind digital watermarking in PDF documents using spread transform dither modulation," Multimedia Tools and Applications, vol.76, no.1, pp.143–161, 2017.
[8] T. Iwamoto and M. Kawamura, "Proposal for invisible digital watermarking method for PDF files by text segmentation (in japanese)," IEICE Tech. Rep., EMM2016-59, vol.116, no.303, pp.31–35, 2016.
[9] M. Kuribayashi, T. Fukushima, and N. Funabiki, "Data hiding for text document in PDF file," Proc. IIHMSP'17, vol.81, pp.390–398, 2017.
[10] B. Chen and G.Q. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," IEEE Trans. Inform. Theory, vol.47, no.4, pp.1423–1443, 2001.
[11] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamson, "Secure spread spectrum watermarking for multimedia," IEEE Trans. Image Processing, vol.6, no.12, pp.1673–1687, 1997.
[12] J.T. Brassil, S. Low, N.F. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," IEEE J. Selected Area in Comm., vol.13, no.8, pp.1495–1504, 1995.
[13] S.H. Low and N.F. Maxemchuk, "Performance comparison of two text marking methods," IEEE J. Selected Area in Comm., vol.16, no.4, pp.561–572, 1998.

Minoru Kuribayashi received B.E., M.E., and D.E degrees from Kobe University, Japan, in 1999, 2001, and 2004. From 2002 to 2007, he was a Research Associate in the Department of Electrical and Electronic Engineering, Kobe University. In 2007, he was appointed as an Assistant Professor at the Division of Electrical and Electronic Engineering, Kobe University. Since 2015, he has been an Associate Professor in the Graduate School of Natural Science and Technology, Okayama University. His research interests include digital watermarking, information security, cryptography, and coding theory. He received the Young Professionals Award from IEEE Kansai Section in 2014. He is a senior member of IEEE.

Takuya Fukushima received B.E. degree from Okayama University, Japan, in 2017. He is currently a student of Graduate School of Natural Science and Technology in Okayama University. His research interest includes digital watermarking and information security.

Nobuo Funabiki received the B.S. and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo, Japan, in 1984 and 1993, respectively. He received the M.S. degree in electrical engineering from Case Western Reserve University, USA, in 1991. From 1984 to 1994, he was with the System Engineering Division, Sumitomo Metal Industries, Ltd., Japan. In 1994, he joined the Department of Information and Computer Sciences at Osaka University, Japan, as as assistant professor, and became an associate professor in 1995. He stayed at University of Illinois, Urbana-Champaign, in 1998, and at University of California, Santa Barbara, in 2000–2001, as a visiting researcher. In 2001, he moved to the Department of Communication Network Engineering (currently, Electrical and Communication Engineering) at Okayama University as a professor. His research interests include computer network, optimization algorithm, image processing, educational technology, Web technology, and network security. Dr. Funabiki is a member of the IEEE and the IPS of Japan.