Robust text watermarking based on line shifting

Alexander Kozachok, Sergey Kopylov

***Abstract*— This article presents an approach to protection of printed text data by watermark embedding in the printing process. Data protection is based on robust watermark embedding that is invariant to text data format converting into image. The choice of a robust watermark within the confines of the presented classification of digital watermark is justified. The requirements to developed robust watermark have been formed. According to the formed requirements and existing restrictions, an approach to robust watermark embedding into text data based on a steganographic algorithm of line spacing shifting has been developed. The block diagram and the description of the developed algorithm of data embedding into text data are given. An experimental estimation of the embedding capacity and perceptual invisibility of the developed data embedding approach was carried out. An approach to extract embedded information from images containing a robust watermark has been developed. The limits of the retrieval, extraction accuracy and robustness evaluation of embedded data to various transformations have been experimentally established.**

***Tóm tắt*— Bài báo trình bày cách tiếp cận để bảo vệ dữ liệu văn bản in bằng cách nhúng vào văn bản một đoạn thủy vân trong quá trình in. Bảo vệ dữ liệu dựa trên việc sử dụng thủy vân bền vững có khả năng chống lại sự chuyển đổi định dạng dữ liệu văn bản sang dữ liệu hình ảnh. Sau quá trình phân tích các hệ thống thủy vân số hiện có, nhận thấy việc lựa chọn một mô hình thủy vân bền vững là hợp lý. Do yêu cầu thực tế và các hạn chế của phương pháp nhúng thủy vân vào dữ liệu văn bản hiện có, bài báo đưa ra phương pháp nhúng mới được phát triển dựa trên một thuật toán ẩn mã sử dụng cách thay đổi khoảng cách giữa các dòng trong văn bản. Bài báo đưa ra một sơ đồ khối và mô tả thuật toán nhúng thông tin vào dữ liệu văn bản. Các thực nghiệm về khả năng nhúng và khả năng che giấu thông tin với tri giác thông thường của dữ liệu nhúng cũng được trình bày. Bài báo cũng nêu cách tiếp cận để trích xuất thông tin được nhúng từ các hình ảnh có chứa thủy vân bền vững. Bên cạnh đó, chúng tôi cũng đưa ra các giới hạn về khả năng ứng dụng của phương pháp dựa trên các thực nghiệm, các đánh giá về độ chính xác của việc trích xuất được dữ liệu và độ mạnh của phương pháp nhúng mới này đối với các phép biến đổi ảnh khác nhau.**

***Keywords*— information security; digital watermarking; data hiding; text steganography.**

***Từ khóa*— an toàn thông tin, thủy vân số, giấu tin, ẩn mã văn bản.**

I. INTRODUCTION

The rapid development of information technology generates large amounts of personal information stored and processed in cloud repositories. The main direction of protecting data banks containing confidential information is protection them from an external violator. In turn, the analysis of information security incidents shows that more than 60 percent of security violations are committed by internal violators [1]. The most common destructive actions performed by employees include illegal copy, provision and distribution of confidential information [2]. One of the compromised data types is text data which contain sensitive information. However, improving the security

This manuscript is received on December 14, 2018. It is commented on December 17, 2018 and is accepted on December 24, 2018 by the first reviewer. It is commented on December 16, 2018 and is accepted on December 25, 2018 by the second reviewer.

of printed text data is not often considered with due attention.

Digital Rights Management (DRM) and Data Loss Prevention (DLP) technologies are widely used to ensure the copyright of text information owners and to protect from leakage. DRM is a technology for controlling access to digital data content [3]. These systems are used to protect against unauthorized use (copying, reproduction) and unauthorized changes of data. Data protection is provided by means of cryptographic encryption and use of a digital license. A digital license is a set of rights (actions) granted to the user when operating data. Using DRM technology in electronic text document protection systems, restrictions can be imposed on making changes to the source text and the structure of a document, the number of operations performed with a text document (copying, printing and transmitting). In addition, restrictions may be imposed on the number of devices from which reading is carried out and the use of programs or devices different from the recommended ones [4].

DLP systems are designed to control dissemination of confidential information by content analysis of data circulating in the network [5]. These systems identify confidential data sent to other networks, located in distributed repositories and data banks, as well as stored by the end user. Data identification is implemented by means of signature analysis. To form a signature, digital fingerprint and digital watermark technologies can be used. The transmission process is blocked in the event of detecting confidential information in the transmitted traffic by the corresponding signature.

The peculiarity of digital fingerprint signature formation is the allocation of unique properties characterizing the source document [6]. The use of this technology makes it possible to achieve high detection accuracy. Unlike fingerprint technology, digital watermark technology allows to embed the generated signature into the source document, which, in turn, enables to increase resistance to various transformations applied to the data. In this case, the process of extracting the embedded digital watermarks is based on the optical character recognition tools, characterized by recognition errors, which impede the correct data identification.

The considered information security technologies can be applied to electronic text data, but are ineffective for printed texts. In order to protect the information owners' copyright from leakage of printed texts, it is necessary to develop new methods of data identification. One of these methods can be presented by a steganographic embedding of the identifying digital watermark information in text data in the process of printing

II. FEATURE OF DIGITAL WATERMARK FORMATION

Digital watermark is a visible or invisible sign (information sequence) embedded in the source data. In the course of work, the classification of digital watermarks is considered by two aspects (Fig. 1) [7].

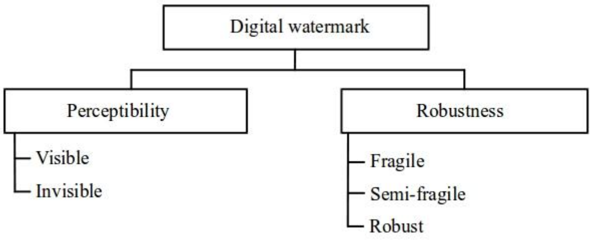


Fig 1. Classification of digital watermark

Embedding of an electronic signature is considered to be an example of embedding a visible digital watermark in text data. The disadvantage of using visible digital watermarks lies in the fact that embedded data can be deleted or modified, which, in turn, leads to incorrect owner identification. Invisible digital watermarks do not have this drawback, but to determine the presence of a watermark and to recover the data it containing additional needed steps in the process of extracting them. According to the reliability criterion, digital watermarks can be represented by the following groups:

- Robust watermarks provide resistance to distortion and implementation of various transformations;
- Fragile watermarks are destroyed by introducing distortions and implementing various transformations;
- Semi-fragile watermarks are destroyed by a separate type of distortion (transformation),

but provide resistance to other types of distortions (transformations) [8].

In case of introducing transformations and distortions into fragile and semi-fragile watermarks, it is impossible to extract embedded information due to the destruction of a watermark. This feature prevents from the use of digital watermark data in the process of copyright protection of the owner's text information from leakage. In turn, robust watermark is able to ensure the property of embedded data invariance in the event of transformations and distortions being introduced. Invisible robust watermarks are characterized by the following parameters [9-11]:

- Embedding capacity (payload) – the amount of information that can be embedded in a container;
- Invisibility (perceptual transparency) – a qualitative characteristic, which reflects the degree of container distortion by the embedded data. This characteristic is based on human perception;
- Non-detectability (detection complexity) – a qualitative characteristic, which reflects the degree of distortion of the container's statistical characteristics, that are not connected with human perception;
- Robustness – the ability of embedded data to retain the property of invariance after various transformations, substitution or deletion of data embedded in the container being implemented;
- Extractability – the ability to retrieve the embedded data from the container correctly.

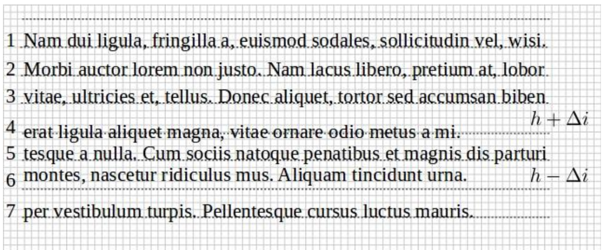
Based on the features of the presentation format of a printed text document, the invisible digital watermark requirements are determined as follows [12]:

- Invariance to the conversion of an electronic text document into a printed form;
- Extractability of the embedded digital watermark from the image containing the source text;
- Invisibility of embedded data to visual analysis;
- Independence of embedded data from the typeface and the semantics of the text.

In addition to the invariance of embedded data to transforming an electronic text document into a printed form, the developed robust watermark should possess the robustness of embedded data to distortion and various transformations being introduced to the container. To determine the limits of robustness of the robust watermark formed, it is necessary to consider possible impacts on the digital watermark systems.

III. THE APPROACH TO ROBUST WATERMARK EMBEDDING

Embedding of a robust watermark in text data, prepared for printing, is carried out through the approach to steganographic embedding of information. In the course of the analysis of steganographic methods for embedding information in text data [13-19], it was established that the embedding approach based on changing the value of line spacing (text line position) (Fig. 2) permits to meet the requirements for robustness of embedded data to format conversion and for independence from a headset and text semantics [20].



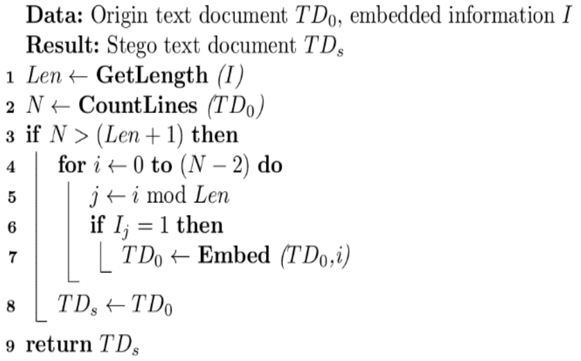
Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi.
Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobor
vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan biben
erat ligula aliquet magna, vitae ornare odio metus a mi.
tesque a nulla. Cum sociis natoque penatibus et magnis dis parturi
montes, nascetur ridiculus mus. Aliquam tincidunt urna.
per vestibulum turpis. Pellentesque cursus luctus mauris.

Fig 2. Line space shifting

The approach to steganographic embedding of information, based on changing the value of line spacing, makes it possible to encode 1 bit of information by changing line spacing by magnitude Δ. The algorithm of embedding information is implemented as follows:

- An increase in the value of line spacing by a set Δ value between adjacent lines of the text is interpreted as "1".
- The absence of changes in the value of line spacing between adjacent lines of the text is interpreted as "0".

The flowchart of the steganographic algorithm of embedding information in text data is presented in Fig 3.



```
Data: Origin text document $TD_0$, embedded information $I$
Result: Stego text document $TD_s$
1 $Len \leftarrow$ GetLength $(I)$
2 $N \leftarrow$ CountLines $(TD_0)$
3 if $N > (Len + 1)$ then
4     for $i \leftarrow 0$ to $(N - 2)$ do
5         $j \leftarrow i \bmod Len$
6         if $I_j = 1$ then
7             $TD_0 \leftarrow$ Embed $(TD_0, i)$
8     $TD_s \leftarrow TD_0$
9 return $TD_s$
```

Fig 3. Algorithm of embedding information in text document

The source data for the algorithm operation is a text document (container document) comprising text data. The use of the following text document formats is allowed:

- RTF (Rich Text Format) – specification [21];
- DOC (Word .doc binary format) – specification [22];
- DOCX (Word Extension to the Office Open XML .docx file format) – specification [23];
- PDF (Portable Document Format) – standard [24, 25].

The first step in the steganographic information embedding algorithm operation (steps 1-2) is to determine the size of embedded information and the achievable embedding capacity using functions *GetLength* and *CounLines*.

At the second stage (step 3), the possibility of embedding information in the source text data is determined. At the stage of coding embedded information (steps 4-7), the code sequence is embedded in the source text document as follows:

- Information is embedded from the first i to $(N-2)$ line, where N determines the achievable capacity of embedding;
- The symbol of the code combination I is read in the line i;
- If the read character corresponds to “1”, then the line spacing between i and $i+1$ lines is increased by Δ value.

To increase the robustness of embedded information, a cyclic embedding scheme has been implemented, which provides re-embedding of identification information a permissible number of times. At the end of the steganographic information embedding algorithm, a signed text document TD_s containing embedded information I is formed.

At the stage of format conversion, a signed text document TD_s printed on paper is scanned, which may include an additional processing procedure consisting of the following transformations: removal of characters, text lines, identification information; printer/scanner noise embedding, etc. As a result, a digital robust watermark-containing image Im is formed.

The process of extracting embedded data imposes additional requirements on the accuracy of detecting the embedded robust watermark and determining the values of line spacing. In addition, the extraction of the embedded robust watermark is implemented from the text-containing image. It requires introducing an image pre-processing stage into the extraction process.

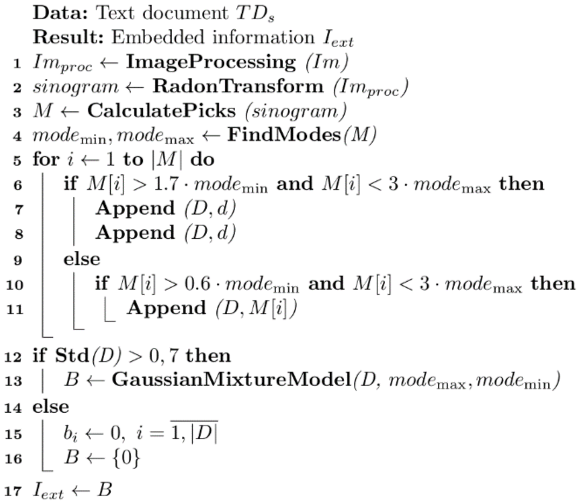
IV. THE APPROACH TO EXTRACTING INFORMATION FROM ROBUST WATERMARK-CONTAINING IMAGES

To implement these requirements and to ensure the veracity of embedded data extraction, an approach to extract information from images with an embedded robust watermark was developed. It consists of the following steps [26]:

- image pre-processing;
- detection of text lines;
- defining an array of line spacing values;
- error detection and correction;
- decoding of embedded information.

A distinctive feature of the developed approach is the use of the normal Radon transform to detect text lines.

The algorithm for extracting information embedded in text data from robust watermark containing images is presented in Fig 4.



```
    Data: Text document $TD_s$
    Result: Embedded information $I_{ext}$
 1  $Im_{proc} \leftarrow$ ImageProcessing $(Im)$
 2  $sinogram \leftarrow$ RadonTransform $(Im_{proc})$
 3  $M \leftarrow$ CalculatePicks $(sinogram)$
 4  $mode_{\min}, mode_{\max} \leftarrow$ FindModes$(M)$
 5  for $i \leftarrow 1$ to $|M|$ do
 6      if $M[i] > 1.7 \cdot mode_{\min}$ and $M[i] < 3 \cdot mode_{\max}$ then
 7          Append $(D, d)$
 8          Append $(D, d)$
 9      else
10          if $M[i] > 0.6 \cdot mode_{\min}$ and $M[i] < 3 \cdot mode_{\max}$ then
11              Append $(D, M[i])$
12  if Std$(D) > 0,7$ then
13      $B \leftarrow$ GaussianMixtureModel$(D,\ mode_{\max}, mode_{\min})$
14  else
15      $b_i \leftarrow 0,\ i = \overline{1, |D|}$
16      $B \leftarrow \{0\}$
17  $I_{ext} \leftarrow B$
```

Fig 4. Robust watermark extraction algorithm from images containing text data

At the first stage, the initial image Im_t is pre-processed (step 1). The function converts the color space of the image to a halftone image (in grayscale), the brightness component of which is subjected to arithmetic average filtering. As a result of processing the source image, a normalized image Im_{proc} is formed in grayscale.

In step 2, the normalized image Im_{proc} is converted into a *sinogram* by the normal Radon transform. A sinogram is a two-dimensional distribution of one-dimensional projections of an object layer as a function of the projection angle, where the projection angle is located on the ordinate axis, and the linear projection is located on the abscissa axis [27]. This conversion is performed by a function *RadonTransform*.

The sinogram contains a series with the most aligned alternation along the line of white stripes and black dots. This series determine the position of text lines (indices of sign change), and the image rotation angle. The array of line spacing values is calculated using the generated sinogram.

The stage of calculating the line spacing values of the image text includes detecting text lines and calculating the interline values of the function *CalculatePicks* (step 3). At the first stage, text lines are detected. A text line is a virtual line along which the text is aligned or merged [28]. To extract text lines, a search for character change indices that correspond to lines of the source text is performed. The distance between the character change indices determines the value of line spacing between two lines of the source text. An array M of line spacing values is calculated as the difference between two character change indices.

At the stage of error detection and correction (steps 4-11), the obtained values M are checked for errors of the first and second kind. Errors of the first kind are the extraction of line intervals that are not presented in the source text. Errors of the second kind are characterized by skipping line intervals in the source text. To detect errors, the function *FindModes* calculates the upper and lower modes $mode_{\max}, mode_{\min}$ of the array M. If the array element is larger than the value $1.7 \cdot mode_{\max}$, the second kind of error is assumed. To correct the detected error, the specified array element is replaced by two elements, each of which has a value $mode_{\max}$. Errors of the first kind corresponding to the array M line intervals values smaller than $0.6 \cdot mode_{\min}$. Correction of the first kind errors is carried out by removing this element from the array M.

Steps 12-17 convert the adjusted D array into a binary form. The resulting array D can be interpreted as a bimodal distribution (zero and one values). Thus, the problem of converting the obtained data sequence into a binary form can be solved by applying methods of mathematical statistics aimed at separating the mixture of distributions. The model of separating a mixture of normal distributions is used as a method of mixture separation [29].

The transformation of array D into a binary array B depends on the value of standard deviation of the array D elements values. The *Std* function calculates the standard deviation of array D elements values. If the obtained value of the standard deviation exceeds 0.7, the binarization is carried out using the function GaussianMixtureModel (step13 in Fig 4), which implements the model of separating the normal distributions mixture. Interpretation of the array D as a discrete bimodal distribution allows us to divide it into two classes. A class characterized by a smaller value of the element mode corresponds to “0”, and a class characterized by a larger value of the mode corresponds to “1” in the binary array B. If, $Std(D) < 0.7$ the array D is considered unimodal and the decision is made about the absence of embedded information. As a result, a binary array B consisting of “0” elements is formed. A binary array B is embedded data.

For the developed approach to the protection of text data to be printed, it is necessary to carry out experimental evaluation aimed at determining the maximum achievable embedding capacity and perceptual invisibility of embedded data, as well as to assess the accuracy of the embedded data extraction from images and robustness to the application of various transformations and distortion.

V. EMBEDDING CAPACITY AND INVISIBILITY OF EMBEDDING APPROACH

A set of experiments was carried out during the experimental evaluation of the developed approach to embedding robust watermark into text data. Within the first set of experiments, the value of the maximum achievable embedding capacity was determined and the dependence of the achievable embedding capacity on the font size and typeface, as well as the size of the line spacing was established. The second group of experiments allowed to estimate perceptual invisibility of embedded data and to establish threshold values of changing the line spacing size.

The source data was the text created in LaTeX, not less than 10 pages with the following margin sizes:

- upper, lower – 20 mm;
- left – 30 mm;
- right – 10 mm.

The maximum achievable embedding capacity of the algorithm is characterized by the number of line spacings of text on one page. To establish the relationship between the embedding capacity, font parameters and the size of the line spacing, the following variable parameters were selected:

- line spacing: 1, 1.25, 1.5.
- font size: 10 pt, 12 pt and 14 pt.
- typeface of the font used:

1) serif font – Computer Modern Roman (similar to Times New Roman).

2) sans serif font – Computer Modern Sans Serif (similar to Arial).

3) monospaced font (monospace) – Computer Modern Typewriter (similar to Courier New).

The values of the maximum achievable embedding capacity are presented in table 1.

TABLE 1. EMBEDDING CAPACITY DEPENDENCE

Font size (pt)	**Line spacing (multiple)**	**Maximum achievable embedding capacity (bit)**
10	1	60
10	1.25	48
10	1.5	40
12	1	49
12	1.25	39
12	1.5	33
14	1	42
14	1.25	33
14	1.5	28

The analysis of the results obtained allows us to make a conclusion about the dependence of the embedding capacity on the size of the line spacing and font size, as well as the absence of dependence on the typeface of the font used. The maximum achievable embedding capacity is 60 bits (when using a font of a 10 pt font size and a line spacing of 1), the minimum is 28 bits (when using a font of a 14 pt font size and a line spacing of 1.5).

During the second group of experiments, the perceptual invisibility of the embedded data was evaluated. The information encoding is realized in the following way:

- The value of the line spacing after each odd line of text does not change and is interpreted as a value of “0”;
- The value of the line spacing after each even line of text is increased by the set value and interpreted as a value of “1”.

During this experiment one paragraph of text with a Computer Modern Roman headset and 14 pt was used. The following variable parameters to determine the threshold of perceptual invisibility were used:

1) Computer Modern Roman 14 pt (4.94 mm) font with 1 (4.94 mm) line spacing:

- increasing the line spacing to 1.25 (6.17 mm) in 0.05(0.245 mm) step increments;
- decreasing the line spacing to 0.75 (3.71 mm) in 0.05 (0.245 mm) step increments;

2) Computer Modern Roman 14 pt (4.94 mm) font with line spacing of 1.25 (6.17~mm):

- increasing the line spacing to 1.4 (6.91 mm) in 0.05 (0.245 mm) step increments;
- decreasing the line spacing to 1.1 (5.43 mm) in 0.05 (0.245 mm) step increments.

3) Computer Modern Roman 14 pt (4.94 mm) font with line interval of 1.5 (7.41~mm):

- increasing the line spacing to 1.25 (8.64 mm) in 0.05 (0.245 mm) step increments;
- decreasing the line spacing to 0.75 (6.17 mm) in 0.05 (0.245 mm) step increments.

As a result of visual analysis of the obtained data, the following threshold of the embedded data perceptual invisibility was established by the expert analysis:

- 0.85-1.15 (4.20-5.68 mm) for a single line spacing;
- 1.10-1.40 (5.43-6.92 mm) for the line spacing of 1.25;
- 1.35-1.65 (6.67-8.15 mm) for a one-and-a-half line spacing.

To determine the applicability and to assess the accuracy of data extraction by means of the developed algorithm to protect text information from leakage due to format conversion, it is necessary to carry out an experimental evaluation of the embedded robust watermark extractability and robustness of the embedded data to possible transformations and distortions of the image.

VI. EXPERIMENTAL EVALUATION OF ROBUST WATERMARK EXTRACTION FROM IMAGES

In the course of determining the extractability of embedded data from images in the LaTeX desktop publishing system, PDF text documents (used typeface – Computer Modern Roman) with the following variable parameters were created:

- font size: 10 pt, 12 pt and 14 pt;
- line spacing: 1; 1.25 and 1.5.

The algorithm for extracting embedded information from the robust watermark containing image is implemented in the Python programming language. For the experimental evaluation of the embedded data extraction accuracy dependence on the value of changing the line spacing used in the process of robust watermark embedding, the integration of robust watermark into prepared text documents, the conversion of text documents from PDF to PNG image format and the extraction of embedded information are carried out. Embedding robust watermark into text documents is implemented by changing the value of the line spacing multiplier from 0.01 to 0.10 from the initial value in 0.01 step increments. When converting a text document from PDF to PNG image format, a DPI value of 200 is used.

During the experimental evaluation, more than 10,000 bits (more than 250 pages of text information) were extracted; it suggests that the confidence interval is 0.95 with an accuracy of 0.01. The results of data extraction accuracy for

TABLE 2. DATA EXTRACTION ACCURACY

Font size (pt)	**Line spacing shifting value (multiple)**	**Extraction accuracy (%)**	**False positive probability (%)**	**False negative probability (%)**
14	1.01	67.8	28.6	3.6
14	1.02	60.5	16.1	23.4
14	1.03	72.9	8.6	18.5
14	1.04	95.1	1.2	3.7
14	1.05	96.3	2.5	1.2
14	1.06	96.3	2.5	1.2
14	1.07	98.8	0	1.2
14	1.08	98.8	1.2	0
14	1.09	98.8	1.2	0
14	1.10	98.8	0	1.2

a text document with a size of 14 pt are presented in Table 2.

The obtained values of accuracy of extracting data from robust watermark containing images make it possible to draw a conclusion on the dependence of the result of extracting embedded information on the value of changing the line spacing used in the process of embedding. Thus, changing the value of the line spacing in the range of values from 0.01 (0.05 mm) to 0.03 (0.15 mm) interval does not allow to correctly extract the built-in information due to a high percentage of errors. At the same time, the change in the value of the line interval in the range from 0.04 (0.20 mm) to 0.10 (0.50 mm) interval is characterized by single extraction errors, which allows us to conclude that it is possible to correctly extract the embedded data from robust watermark containing images by means of the developed algorithm.

The obtained values of the embedded data extraction accuracy allow us to proceed to the experimental evaluation of the embedded data extraction from robust watermark containing images obtained by applying the operation “print-scan” to the text documents prepared for printing. In the course of the experimental evaluation of the embedded data extraction accuracy dependence on the image quality (DPI of the scanned image), embedding of robust watermark into the prepared text documents, printing text documents, scanning printed text documents and the extraction of embedded information was carried out.

The obtained values make it possible to draw a conclusion about high accuracy of extraction (over 95%) of embedded data from the images containing an embedded robust watermark for the DPI indicator of 150 points per pixel and above.

In addition to determining the applicability and assessment of data extraction accuracy, evaluation of resistance of robust watermark containing images to various transformations and distortions was made.

VII. ROBUSTNESS OF DEVELOPED WATERMARK

The evaluation of robustness of the developed algorithm for extracting information from robust watermark containing images consists in the possibility of reliably extracting information from images after applying the following transformations to them:

1) Image rotation;

2) Scaling;

3) Format conversion (JPEG, PNG, BMP, PDF);

4) Median filtering;

5) Gaussian blurring;

6) Averaged filtering.

In the process of determining the robustness of the developed algorithm for extracting data from robust watermark containing images, the 360-degree rotation of the signed image to its spindle was performed (with spacing – 5 degree). An example of a generated sinogam, the result of the Radon transform and an image after Gaussian blurring filter with a blur rate 8 are shown in Fig 5.

During experimental assessment of the developed algorithm robustness to image rotation for all cases, the angle of image rotation was correctly determined, a sinogram was constructed, and the line spacing values were correctly extracted. The obtained results led us to the conclusion that the developed algorithm for extracting information to any degree image rotation was robust.

To assess the robustness of the developed algorithm to image scaling, an experiment was conducted, during which the source image was scaled with a multiplier from 0.5 to 1.5 (with 0.05 spacing).

The analysis of the obtained results made it possible to conclude that the developed algorithm for extracting embedded data is robust to image scaling within a scaling factor of a multiplier up to 1.5, taking into account the reliable extraction of embedded information from transformed images.

The robustness of the developed algorithm for extracting information to image format conversion is due to implementation of image format conversion at the stage of preprocessing of the source image. During processing, an image, regardless of the original format (JPEG, PNG, BMP, GIF, PDF, etc.), is converted to PNG format.

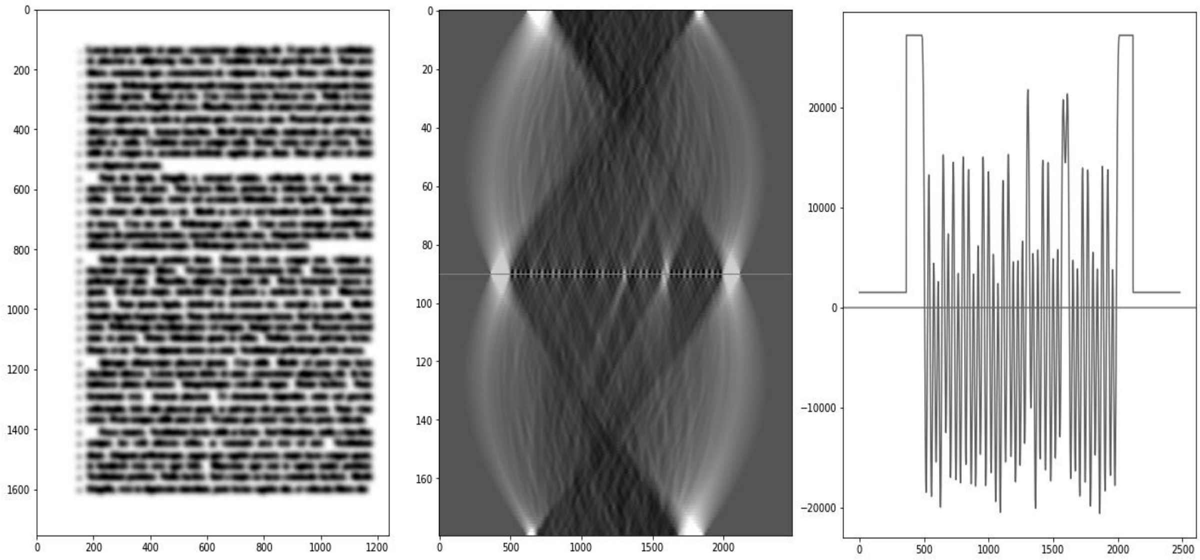


Fig 5. Data extraction from image filrering Gaussian filter with blur rate 8

Robustness to median filtering of a text data-containing image was performed by applying a median filter with a convolution kernel size increment from 1 to 9 pixels. The developed extraction algorithm is resistant to median filtering with a maximum value of a convolution kernel size equal to 9 pixels.

Robustness to Gaussian blurring of a text data-containing image was evaluated by applying a Gaussian blur filter. The limit of information extraction by the developed algorithm is restricted by 8 pixel blur radius.

Averaged filtering was implemented by a mode filter. The developed extraction algorithm provides robustness to mode filtering with a convolution kernel size of 5 pixels.

It should be noted that the boundary values of the estimated filters distort the source image so much that the text can no longer be read, but the developed algorithm permits to comprehend the degree of the source text document secrecy. The obtained results allow us to conclude that the developed method is robust to the specified types of filtering.

VIII. CONCLUSION

The developed approach to robust watermarking of the printed text data based on the line shifting makes it possible to increase text data security by embedding and subsequent extraction of identification information invariant to transformation of an electronic text document into a printed form. The obtained results proving the developed watermark durability make it possible to attribute it to the robust class due to its resistance to the considered transformations applied to images. Improving extraction accuracy and reducing the number of embedding errors are the directions of further research.

REFERENCES

[1]. Analytical center InfoWatch. “Global Data Leak Report 2017”, 2018. https://infowatch.com /report 2017 (accessed: 13/06/2018) (in Russian).

[2]. “Analytical center InfoWatch. Information Security Incidents Caused by Resigning Employees”. A Study by InfoWatch, 2018. https://infowatch.com/report%5CUEBA2017; (accessed: 13/06/2018) (in Russian).

[3]. A. Mostafa [et al.]. “Mostafa A. Consumer Privacy Protection in Digital Right Management: A Survey”, International Journal of Computer Information Systems and Industrial Management Applications, Vol. 9, pp. 218–231, 2017.

[4]. M. M. Azad, A. H. S. Ahmed, A. Alam. “Azad M. M. Digital Rights Management”0. International Journal of Computer Science and Network Security. Vol. 10, no. 11, pp. 24–33, 2010.

[5]. Kanagasingham P. “Data Loss Prevention P. Kanagasingham”, Sans institute, pp. 1–31, 2008.

[6]. Milano D. “Content control: Digital watermarking and fingerprinting”, D. Milano, White Paper, Rhozet, a business unit of Harmonic Inc.,Vol. 30, pp. 1–11, 2012.

[7]. Gribunin V. “Digital steganography V. Gribunin”, I. Okov, I. Turincev, Moscow: SOLON-Press, pp. 262 (in Russian), 2017.

[8]. A. V. Kozachok [et al.]. “Review of the current methods for robust image hashing”, Computer optics, Vol. 4, no. 5, pp. 743–755 (in Russian), 2017.

[9]. Salomon D. “Data privacy and security: encryption and information hiding”, Springer Science & Business Media, pp.469, 2003.

[10]. Woo C.-S. “Digital image watermarking methods for copyright protection and authentication”, Queensland University of Technology, pp. 197, 2007.

[11]. Phadikar “A. Robust Watermarking Techniques for Color Images”, April, 2009.

[12]. Kozachok A. V. “Robust watermark as technique to text data leakage prevention” A. V. Kozachok, S. A. Kopylov, M. V. Bochkov Information Security. INSIDE.Vol. 82, no. 4, pp. 1–8 (in Russian), 2018.

[13]. Rathore A, S. Rawat. “A Secure Image and Text Steganography Technique”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 3, no. 5, pp. 506–509, 2015.

[14]. Agarwal M. “Text steganographic approaches: a comparison”, International Journal of Network Security & Its Applications (IJNSA). Vol. 5, no. 1, pp. 91–106, 2013.

[15]. Pamulaparty L, N. Rao. “Text Steganography: Review”, International Journal of Computer Science and Information Technology & Security (IJCSITS),Vol. 6, no. 4, pp. 80–83, 2016.

[16]. Saraswathi V, ”, M. S. Kingslin. “Different Approaches to Text Steganography: A Comparison”, International Journal of Emerging Research in Management & Technology. Vol. 9359, no. 11, pp. 124–127, 2014.

[17]. Kumar K, S. Pabboju, N. M. S. Desai. “A. Advance text steganography algorithms: an overview”, International Journal of Research and Applications. Vol. 1, no. 1, pp. 31–35, 2014.

[18]. Rani N, J. Chaudhary. “Text Steganography Techniques: A Review”, International Journal of Engineering Trends and Technology (IJETT). Vol. 4, no. 7, pp. 3013–3015, 2013.

[19]. Singh H, A. Diwakar, S. Upadhyaya. “A Novel Approach to Text Steganography”, International Congress on Computer, Electronics, Electrical, and Communication Engineering (ICCEECE2014). Vol. 59, no. 1, pp. 8–12, 2014.

[20]. Kozachok A. V, S. A. Kopylov. “The embedding approach to robust watermarking in text data”, 2018.

URL:http://www.ruscrypto.ru/resource/archive/rc2018/files/11%5CKozachok%5CKopylov.pdf. (accessed: 13.06.2018) (in Russian).

[21]. Word 2007: Rich Text Format (RTF) Specification, version 1.9.1. – 2008. – URL: https://www.microsoft.com/en/us/download/details.aspx?id= 10725 (accessed: 13.06.2018).

[22]. [MS-DOC] Word (.doc) Binary File Format. – 2017. – URL: http://interoperability.blob.core.windows.net/files/MS-DOC/%5C%5BMS-DOC%5C%5D.pdf (accessed: 13.06.2018).

[23]. [MS-DOCX] Word Extensions to the Office Open XML (.docx) File Format. – 2017. – URL: http:// interoperability. blob . core . windows . net / files / MS - DOCX/%5C%5BMS-DOCX%5C%5D.pdf (accessed: 13.06.2018.

[24]. ISO TS. 171/SC 2: ISO 32000–1: 2008 Document Management-Portable Document Format-Part 1: PDF 1.7. – 2008.

[25]. ISO TS. 171/SC 2: ISO 32000–2: 2017 Document Management-Portable Document Format-Part 2: PDF 2.0. – 2017.

[26]. A. V. Kozachok [et al.]. “An approach to a robust watermark extraction from images containing text”, SPIIRAS Proceedings (in Russian),Vol. 5(60), pp. 128–155, 2018.

[27]. IEC/TR. 61948-2:2001 “Nuclear medicine instrumentation. Routine tests”. Part 2. Scintillation cameras and single photon emission computed tomography imaging. Moscow: Standartinform, (in Russian), pp.11, 2009.

[28]. Bahaghighat M. K, J. Mohammadi. “Novel approach for baseline detection and Text line segmentation”, International Journal of Computer Applications. Vol. 51, no. 2, pp. 9–16, 2012.

[29]. Reynolds D. “Gaussian mixture models”, Encyclopedia of biometrics, pp. 827–832, 2015.

ABOUT THE AUTHORS

PhD. Alexander Kozachok

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: alex.totrin@gmail.com

The education process: has received PhD. degree in Engineering Sciences in Academy of Federal Guard Service of the Russian Federation in Dec. 2012.

Research today: Information security; Unauthorized access protection; Mathematical cryptography; theoretical problems of computer science.

PhD. Sergey Kopylov

Workplace: The Academy of Federal Guard Service of the Russian Federation.

Email: gremlin.kop@mail.ru

The education process: graduated from Academy of Federal Guard Service of the Russian Federation in 2010.

Research today: Information security, Data leakage protection, Pattern recognition, Image processing.