ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.352.242, 004.056 **doi: 10.32620/reks.2022.1.12**

Konstantin DERGACHOV, Leonid КRASNOV, Vladislav BILOZERSKYI, Anatoly ZYMOVIN

National Aerospace University "Kharkiv Aviation Institute", Ukraine

METHODS AND ALGORITHMS FOR PROTECTING INFORMATION IN OPTICAL TEXT RECOGNITION SYSTEMS

The subject of the study. *A concept of OCR systems performance improvement is proposed, which is achieved through the integrated use of special algorithms for preliminary processing of documents picture, an extended set of service functions, and advanced techniques for information protection.* ***Study objectives****: development of algorithms that compensate for the influence of the unfavorable points like imperfect lighting conditions overshooting, images geometric deformation, noises, etc., which corrupt the pictured text, on the efficiency of that text recognition. It is needed to provide for a series of service procedures that would ensure adequate data handling while viewing, converting, and storing in standard formats the results, and ensuring the possibility to exchange data in open communication networks. Additionally, it is necessary to ensure the information protection against unauthorized use at the stage of data processing and provide secretiveness of their transmission through the communication channels.* ***Investigation methods and research results****: developed and tested algorithms for preliminary picture data processing, namely, for the captured image geometry transformation, picture noise correction with different filters, image binarization when using the adaptive thresholds reduced the negative influence of irregular image portions illumination; in the software, the special services ensure the data processing ease and information protection are affected. In particular, the interactive procedure for text segmentation is proposed, which implies the possibility of anonymizing its fragments and contributes to collecting confidentiality for documents treated. The package for processing document shots contains the face detection algorithm bringing the identification of such information features; it can be used further in the task of face recognition. After the textual doc is recognized, the received data encryption is provided by generating a QR-code and the steganography methods can deliver the privacy of this information transmission. The algorithms' structures are described in detail and the stability of their work under various conditions is investigated. Focused on the case study, docs' text recognition software was developed with the usage of Tesseract version 4.0 optical character recognition program. The program named "HQ Scanner" is written in Python using the present resources of the OpenCV library. An original technique for evaluating the efficiency of algorithms using the criterion of the maximum probability of correct text recognition is implemented in the software.* ***Conclusions****. The study results can constitute the basis for developing advanced specialized software for easy-to-use commercial OCR systems.*

Keywords: *Optical character recognition; probability of correct text recognition; text segmentation fragment anonymization; QR code; steganography algorithms.*

1. Introduction

Currently, most problems on characters recognition and classification are solved using resources of artificial intelligence and employing deep learning neural networks. One of the most popular and demanded areas in patterns recognition is optical character recognition (OCR) technique. Since it is customary nowadays to avoid the paper format while storing, reading and editing information, there exist the systems for converting textual and digital data from paper to electronic format. Both seeking and handling information over digital documents happen much easier and faster than with viewing handwritten or printed paper data. Therefore text data extraction from documents picture finds numerous useful applications.

The work is dedicated to the promising OCR system selection, the enhancement of the system operation quality in condition of interfering influences, as well as to the methods of protecting confidential information obtained as a result of text recognition.

1.1. Motivation for research

Today, there are a sufficient number of ready-made software solutions for optical text recognition. There are numerous free-to-use and customized text recognition applications available on the IT services market. The most popular are the *FineReader* by ABBYY Company and the OCR system *CuneiForm* by Cognitive Technologies; The Tesseract engine is a free OCR program developed by Hewlett-Packard and donated in 2006 by Google. Some other well-known software products are shown in Fig. 1.

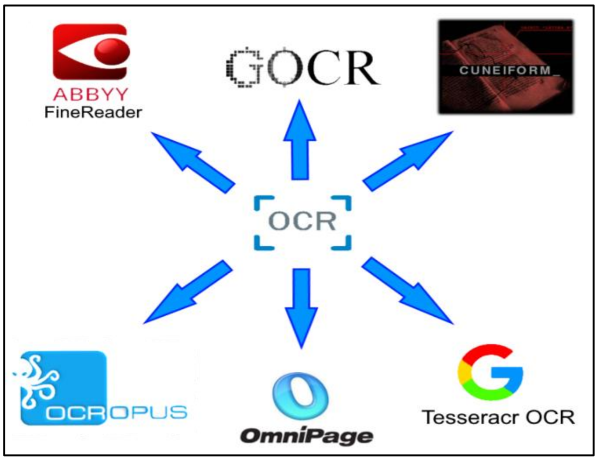


Fig. 1. Optical text recognition facilities

With the analysis of these means, the following conclusions can be drawn:

– several programs do not get an intuitive interface;

– the programs cannot run in the background;

– not all programs have image pre-processing resources to improve the quality of text recognition;

– none of the programs gives 100% result and perfect saving of formatting.

Taking into account the above things, the choice was made in favor of the Tesseract program.

The Tesseract program became a very popular tool for optical texts recognition recently [1, 2]. This is an open source OCR engine, which grounds neural networks to search and identify textual contents in images media. Since the version 4.0, Tesseract employs the Long Short-Term Memory architecture (LSTM) for recurrent neural networks.

In laboratory "glasshouse" conditions, i.e. under comfortable lighting while shooting the text document and absence of picture's geometric distortions while concluding recognition, the Tesseract software manifests perfective efficiency. The accuracy of text identification approaches 99%. However, external interfering impacts reduce sharply the system operation quality, up to complete loss of its performance.

To the current, the leading OCR experts have reviewed and analyzed in detail the works on assessment of certain factors negative influence on the efficiency of text recognition [3]. It is stressed that severe hard issues in realizing an OCR program is not the recognition procedure itself but the bound with that stages of the image preprocessing, noise reduction and cleaning. The experts have brought in a number of constructive solutions for such issues [4, 5]. These focus on algorithms, which correct image imperfections that appear due to wretched position of the text document with respect to the camera coordinate system while shooting. However, a real challenge of different light intensity for fragments when photographing a document is not properly discussed. Though these issues result in worse for text recognition and sometimes end with losing whole fragments if they appear to be shadowed while shooting. Unfortunately, at present there are no universal recommendations on this matter, let alone standards [6, 7].

1.2. Work related analysis

The examination of the complex problem on improving the quality of text recognition systems indicates a few main areas of work that can significantly improve the performance of existing systems, namely:

– keyword discovery, which is based on deep convolutional neural networks, to protect personal information in document images. The method involves the key characters detection and the vocabulary analysis; it is developed on the basis of RetinaNet and transfer learning. To search for the key characters, the RetinaNet neural network that consists of convolutional layers is employed. The latter constitute a pyramidal network and two subnets used to search key characters in a region of interest on the document picture. The proposed technique is significantly superior to the classic Tesseract OCR software (known facility used for detecting key characters in document images). The paper [18] presents an example of such an approach;

– compensation for the most significant negative external factors (images geometric distortion, shading of image fragments, poor contrast, noise, etc.). This demands pre-processing and picture enhancement to ensure correct identification and classification of text images. To estimate the level of projective distortion at the point of the reconstructed image, a theoretically substantiated method was developed [19]. On this basis, a new technique for binary estimation of the quality of reconstructed images is proposed. Notably useful results are contained in the studies, which use, before recognizing the text image, the picture binarization with adaptive thresholds applied [10, 20, 21];

– comprehensive measures aimed at protection of confidential information being received in the result of textual document recognition rely currently on QR coding [11, 12] and the use of various LSB steganography options to mask the fact of message transmission over open communication channels [11 – 13, 15].

In addition to the above-mentioned identification tasks related to the recognition of texts in photographs taken under non-standard conditions or low-resolution cameras, there are special tasks associated with providing additional service functions, such as ease of viewing, the ability to record data in a regular format, detection and recognition of faces in images found in texts, etc. All this determines the content of the work offered to your attention.

2. Formulation of the problem

When designing optical character recognition systems based on the Tesseract OCR engine, it is necessary not only to compensate for negative external influences, but also to provide additional service functions that allow you to view, sort and save text information obtained as a result of recognition in a standard form. The most important indicator of the quality of the system is also the availability of information protection functions, both at the stage of document recognition and during their transmission over open communication channels.

The system **design concept** implies significant improvement of the text recognition systems quality when relying on photographs of the paper documents, and provides for realizing the following mandatory provisions:

– assessment of the external factors negatively affecting the outcome of the text recognition system;

– formulation of criteria and indicators to quantitatively assess the quality of text recognition;

– creation of a set of algorithms to fulfil images preliminary treatment intended to effectively compensate for the influence of disturbing factors;

– providing a set of service functions for the convenience of processing source data, viewing them, converting and saving them in standard formats, and transmitting results via communication channels;

– ensuring the confidentiality of the received data at the stage of processing and ensuring secrecy during transmission over communication channels;

– designing and testing the software for practical implementation of the research results;

– to supply a guidance for further improvement of optical texts recognition techniques.

3. Tools and methods for the study

Hardware. The project is focused on the use of affordable and relatively cheap means of photography (phones, tablets, cameras, webcams and other nonvolatile devices). The operation of the optical text recognition system is implemented in a laptop; the system can be hosted on a Raspberry Pi single board computer if a mobile system option is needed.

Software resources. The Python programming language and the resources of the OpenCV library were used [8, 9]. This choice was done due to open access to software products and their compatibility with Windows, Linux and Android operating systems. For texts recognition, the Tesseract software version 4.0 is used as the dedicated program module.

Factors of negative effects:

− bad quality of original documents (poor print quality, insufficient text contrast, or poor paper quality);

− unfavorable lighting conditions for the scene when photographing;

− photographic equipment bad-quality and incorrect operator actions resulting in, for example, pictures defocusing and so on;

− shootings geometric distortion (photograph's disposition mismatch to the recognition facility coordinates system).

Text recognition fidelity assessment. To assess the effectiveness of text recognition in various conditions, a quantitative indicator of efficiency is used [10]. This index determines the accuracy of text recognition, which allows for a comparative analysis of various case studies; the evaluation takes into account such parameters as N_src (number of characters in the source text), N_dst (number of correctly recognized characters), N_extra (number of erroneous new characters after recognition), N_missed (number of unrecognized characters).

"HQ Scanner" system’s modules. During researches, a complex program was created, the block diagram of which is shown in Fig. 2. The program under name "HQ Scanner" is designed on a modular basis where the modules run around using common interface.

In building the program interface, the PyQt5 library and the associated Qt Designer environment were used.

PyQt is a set of extensions in the Qt graphics framework employing the Python programming language. PyQt also includes Qt Designer (Qt Creator) as a graphical user interface designer. The `pyuic` program generates code from files created in Qt Designer.

The content of individual modules and the results of a study of their effectiveness are described in detail by the authors in a previous publication [10].

Therefore, further we will consider only the algorithms used to protect information both in the process of creating an electronic version of documents, and in the process of transferring this data through communication channels.

Note it is not needed usually to run the system for text documentation identification in real-time format. Besides often, the interactive procedures like current viewing of transformation results, text partial segmentation, etc. are necessary to perform. These principally slow down the program, but essentially increase the text recognition authenticity.

Interface and main functions of the program "HQ Scanner". Fig. 3 displays the master window running theprogram basic facilities.

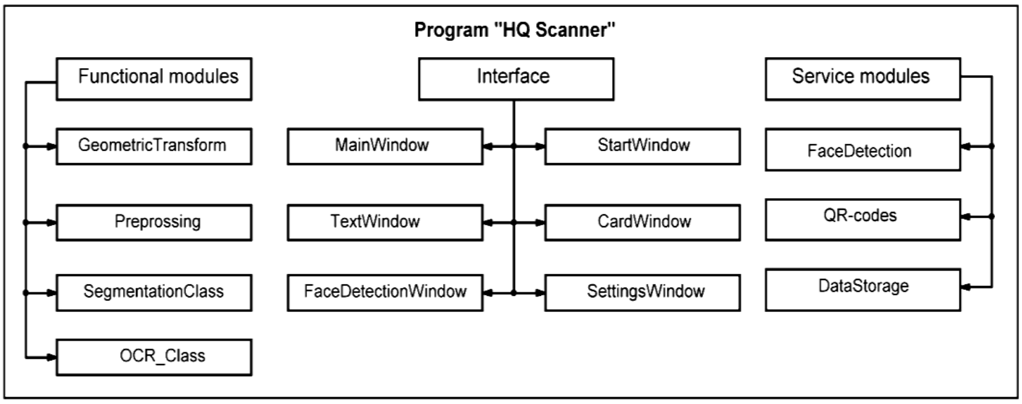
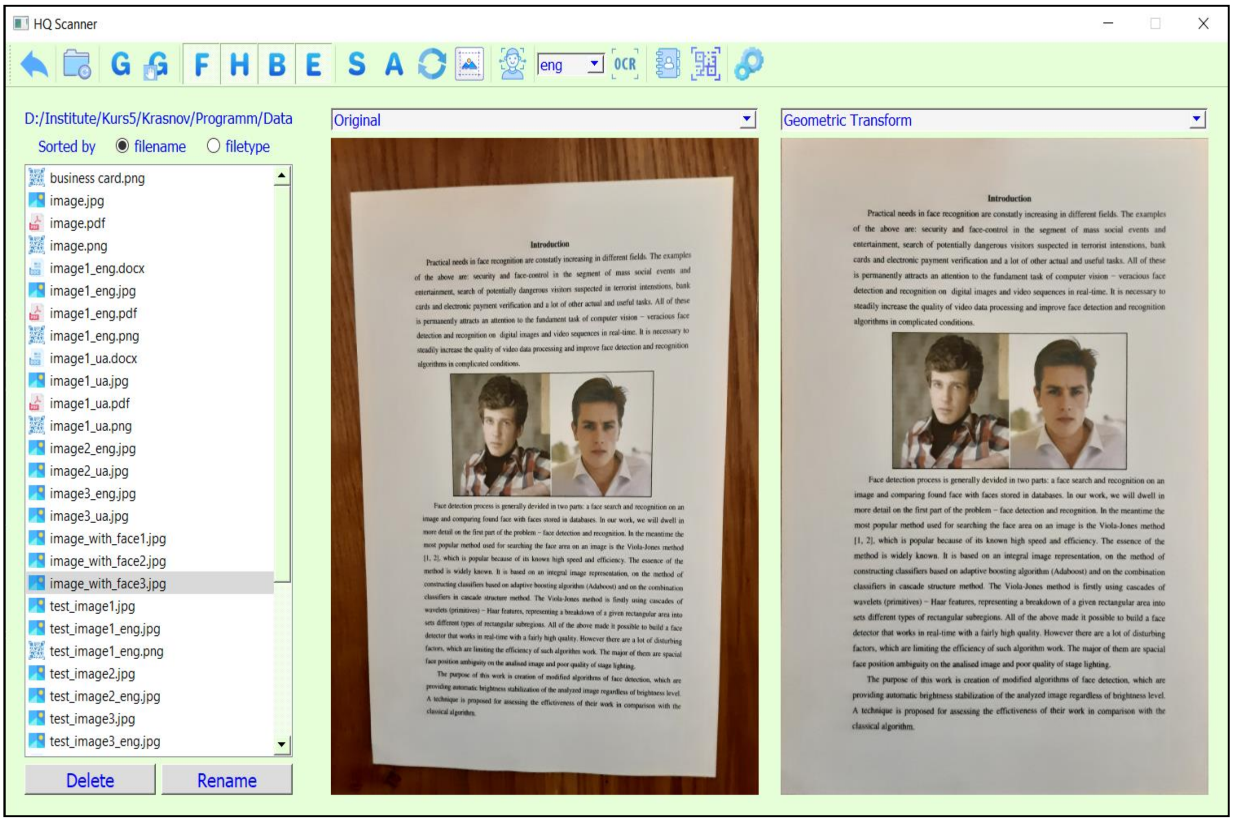


Fig. 2. “HQ Scanner” structure



Introduction

Practical needs in face recognition are constatly increasing in different fields. The examples of the above are: security and face-control in the segment of mass social events and entertainment, search of potentially dangerous visitors suspected in terrorist intenstions, bank cards and electronic payment verification and a lot of other actual and useful tasks. All of these is permanently attracts an attention to the fundament task of computer vision – veracious face detection and recognition on digital images and video sequences in real-time. It is necessary to steadily increase the quality of video data processing and improve face detection and recognition algorithms in complicated conditions.

Face detection process is generally devided in two parts: a face search and recognition on an image and comparing found face with faces stored in databases. In our work, we will dwell in more detail on the first part of the problem – face detection and recognition. In the meantime the most popular method used for searching the face area on an image is the Viola-Jones method [1, 2], which is popular because of its known high speed and efficiency. The essence of the method is widely known. It is based on an integral image representation, on the method of constructing classifiers based on adaptive boosting algorithm (Adaboost) and on the combination classifiers in cascade structure method. The Viola-Jones method is firstly using cascades of wavelets (primitives) – Haar features, representing a breakdown of a given rectangular area into sets different types of rectangular subregions. All of the above made it possible to build a face detector that works in real-time with a fairly high quality. However there are a lot of disturbing factors, which are limiting the efficiency of such algorithm work. The major of them are spacial face position ambiguity on the analised image and poor quality of stage lighting.

The purpose of this work is creation of modified algorithms of face detection, which are providing automatic brightness stabilization of the analyzed image regardless of brightness level. A technique is proposed for assessing the effictiveness of their work in comparison with the classical algorithm.

Fig. 3. "HQ Scanner" program master window

The program toolbar is on the top, and the window common workspace is shared conditionally into three parts. There is a box on the left that is used to open the required file for viewing; the panel has the option to sort files by either "file_name" or "file_type" feature (pointed for that are the related buttons "Filename" and "Filetype"). Over these, a path to locate the input doc is shown.

The toolbar displays a set of tab labels used to select the required operating mode. The name's first letter of the operation being called settled the icon label (Fig. 4).

Segmentation and anonymization. To avoid unnecessary effort, the program provides the ability to segment the page of the processed text document (after segmentation, only the key fragments are treated and identified).

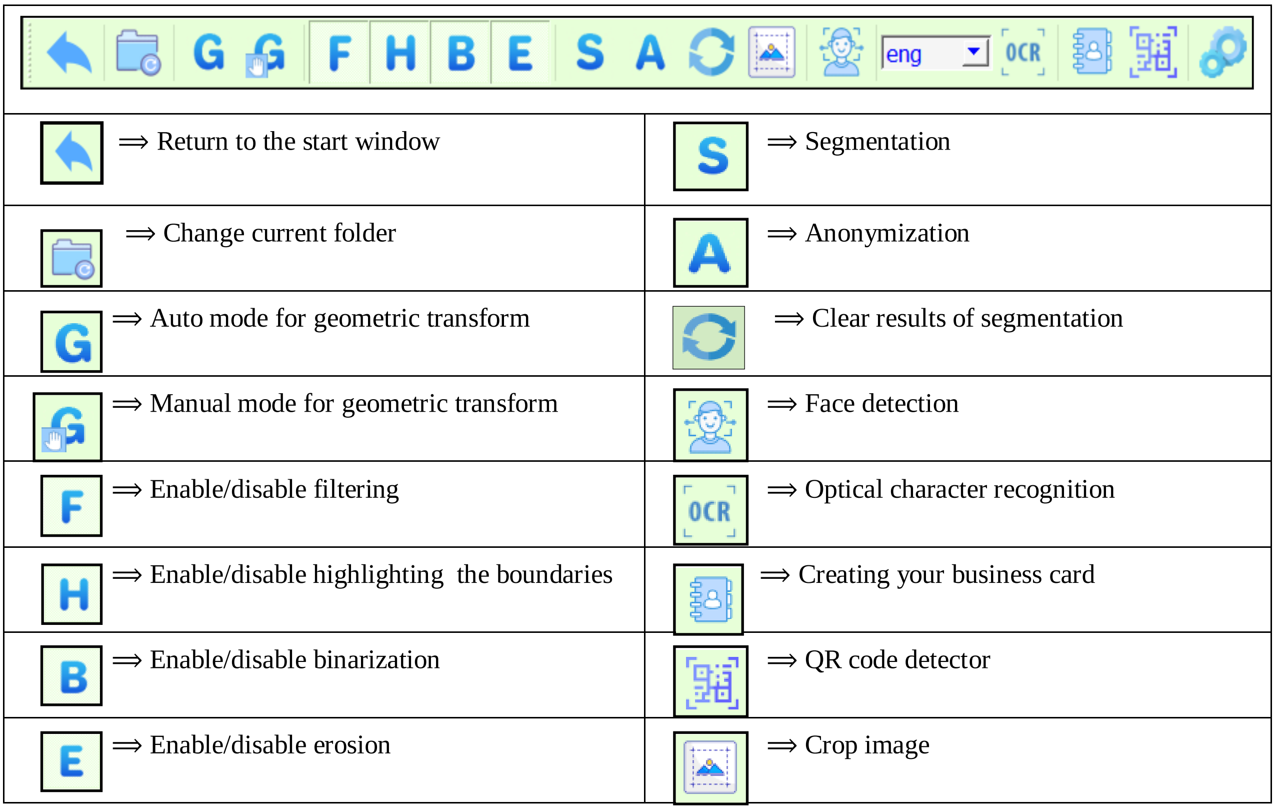
The segmentation routine is performed in the interactive mode with the help of the mouse on the “Segmentation” tab shown in Fig. 5, a.

Similarly, to maintain the confidentiality of the processed information, necessary image sections can be hidden handling the dedicated window after activation of the tab "Anonymization" (Fig. 5, b).

Display and save recognition results. You can call a dedicated window (Fig. 6) to display and next to save the recognition outcomes.

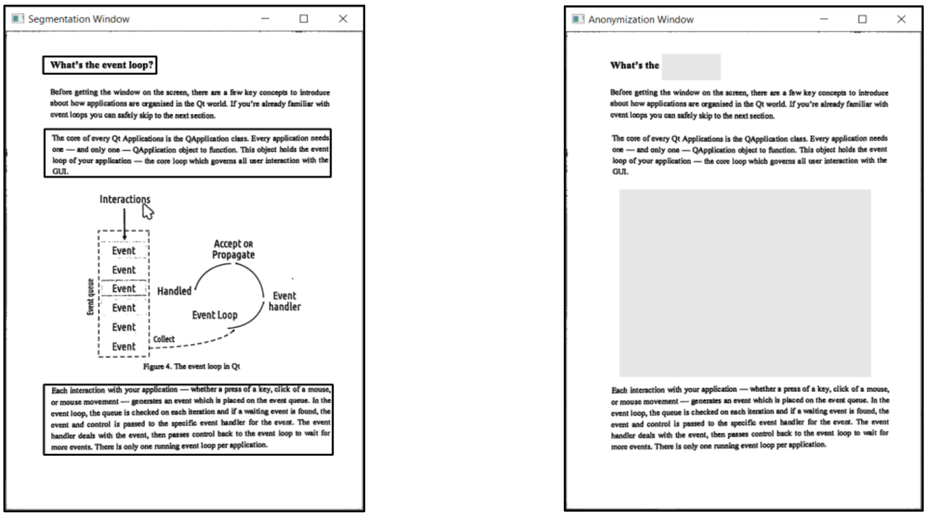
Also in this box, there is own little toolbar that allows one to select the text font size and to save the results of OCR in `*.docx` or `*.pdf` format with the use

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)



↩	⟹ Return to the start window	S	⟹ Segmentation
📁	⟹ Change current folder	A	⟹ Anonymization
G	⟹ Auto mode for geometric transform	🔄	⟹ Clear results of segmentation
G	⟹ Manual mode for geometric transform	👤	⟹ Face detection
F	⟹ Enable/disable filtering	OCR	⟹ Optical character recognition
H	⟹ Enable/disable highlighting the boundaries	📇	⟹ Creating your business card
B	⟹ Enable/disable binarization	▦	⟹ QR code detector
E	⟹ Enable/disable erosion	🖼	⟹ Crop image

Fig. 4. Toolbar of the program and the purpose of its elements



What's the event loop?

Before getting the window on the screen, there are a few key concepts to introduce about how applications are organised in the Qt world. If you're already familiar with event loops you can safely skip to the next section.

The core of every Qt Applications is the QApplication class. Every application needs one — and only one — QApplication object to function. This object holds the event loop of your application — the core loop which governs all user interaction with the GUI.

Figure 4. The event loop in Qt

Each interaction with your application — whether a press of a key, click of a mouse, or mouse movement — generates an event which is placed on the event queue. In the event loop, the queue is checked on each iteration and if a waiting event is found, the event and control is passed to the specific event handler for the event. The event handler deals with the event, then passes control back to the event loop to wait for more events. There is only one running event loop per application.

a

What's the

Before getting the window on the screen, there are a few key concepts to introduce about how applications are organised in the Qt world. If you're already familiar with event loops you can safely skip to the next section.

The core of every Qt Applications is the QApplication class. Every application needs one — and only one — QApplication object to function. This object holds the event loop of your application — the core loop which governs all user interaction with the GUI.

Each interaction with your application — whether a press of a key, click of a mouse, or mouse movement — generates an event which is placed on the event queue. In the event loop, the queue is checked on each iteration and if a waiting event is found, the event and control is passed to the specific event handler for the event. The event handler deals with the event, then passes control back to the event loop to wait for more events. There is only one running event loop per application.

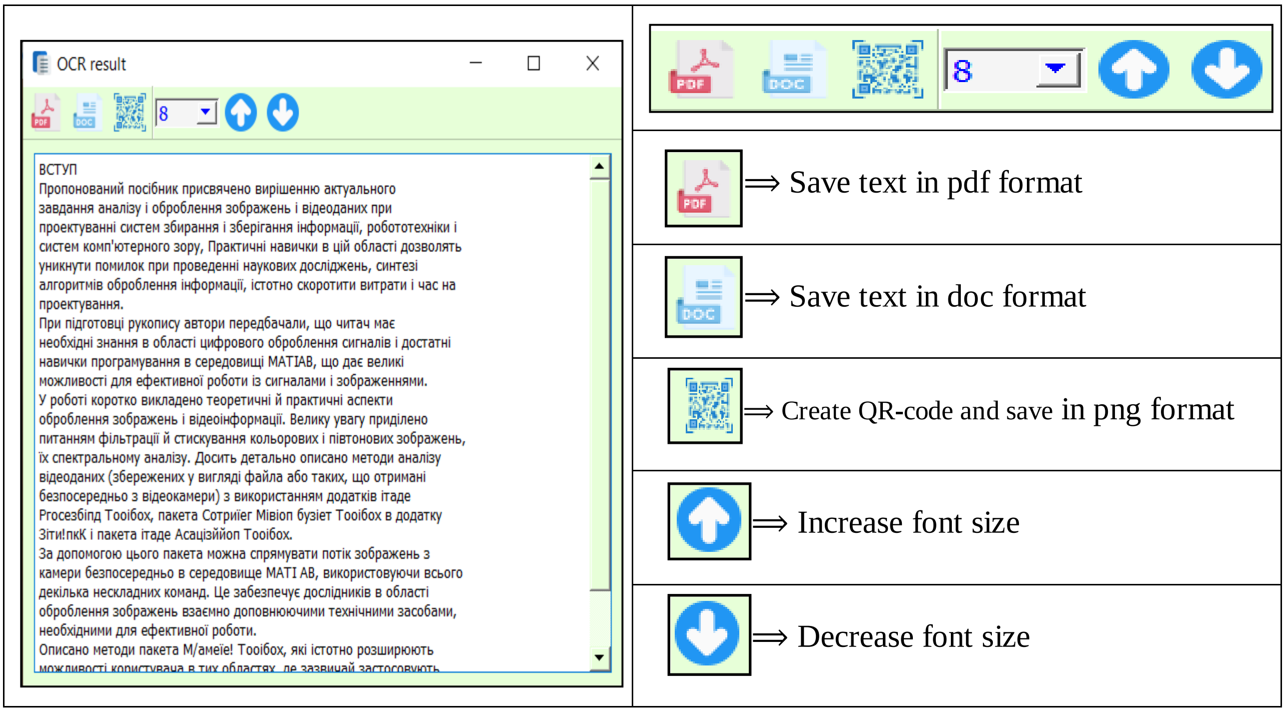
b

Fig. 5. Pop-up windows for segmentation and anonymization of text fragments:
a – segmentation window; b – anonymization window

of the dedicated buttons. Pushing either of the options yields a file being reflected on the navigation tab designed for viewing and sorting files (see Fig. 3) and having the same name as the original image, though getting the just chosen extension.

The most prospective function of the toolbar in the viewing and saving recognition results window is generation of the QR-code for the given piece of the text. QR-codes is a highly progressive tool for hiding text information in the form of a compact image [11, 12]. It is common to utilize the png format to store results in this technique.

The program's function for generating QR-codes is activated when it is necessary to preserve the confidentiality of information obtained during recognition. The images with QR-codes are used as embedded images in a cover image due to what the information transmitted via communication channel obtains protection, for instance, through the LSB-steganography technique [13, 14]. This method of encrypting messages and, simulta-



OCR result

8

ВСТУП
Пропонований посібник присвячено вирішенню актуального завдання аналізу і оброблення зображень і відеоданих при проектуванні систем збирання і зберігання інформації, робототехніки і систем комп'ютерного зору, Практичні навички в цій області дозволять уникнути помилок при проведенні наукових досліджень, синтезі алгоритмів оброблення інформації, істотно скоротити витрати і час на проектування.
При підготовці рукопису автори передбачали, що читач має необхідні знання в області цифрового оброблення сигналів і достатні навички програмування в середовищі MATIAB, що дає великі можливості для ефективної роботи із сигналами і зображеннями.
У роботі коротко викладено теоретичні й практичні аспекти оброблення зображень і відеоінформації. Велику увагу приділено питанням фільтрації й стискування кольорових і півтонових зображень, їх спектральному аналізу. Досить детально описано методи аналізу відеоданих (збережених у вигляді файла або таких, що отримані безпосередньо з відеокамери) з використанням додатків ітаде Ргосезбіпд Тооібох, пакета Сотриїег Мівіоп бузіет Тооібох в додатку Зіти!пкК і пакета ітаде Асацізййоп Тооібох.
За допомогою цього пакета можна спрямувати потік зображень з камери безпосередньо в середовище MATI AB, використовуючи всього декілька нескладних команд. Це забезпечує дослідників в області оброблення зображень взаємно доповнюючими технічними засобами, необхідними для ефективної роботи.
Описано методи пакета М/амеїе! Тооібох, які істотно розширюють можливості користувача в тих областях, де зазвичай застосовують

8

⟹ Save text in pdf format

⟹ Save text in doc format

⟹ Create QR-code and save in png format

⟹ Increase font size

⟹ Decrease font size

Fig. 6. Pop-up window to display and save OCR results

neously, ensuring the data transfer secrecy seems to be very promising and in demand in various applications.

4. Data protection using steganography techniques

The use of steganography to hide secret data implies their transfer using conventional multimedia (pictures, audio or video files). The main purpose of the approach is to hide the very existence of certain information. The general structure of the network for transmitting secret messages is shown in Fig. 7.

The most popular and adequate method to solve the problem is using an image as a message container and the option of steganography known as LSB steganography (LSB stands for 'least significant bit'). For clarity of further presentation, we will be using the term 'image-cover'. The idea of coding is to replace the least significant bits in the container picture with the bits of the message being a secret. The latter can be both textual data and a new embedded image.

Quality criteria for steganographic transformations. While high-quality steganography using, the an empty and filled container image difference should not be perceptible to the human visuals. However, the perception can differ significantly at various observers [16, 17]. Therefore, related quantitative indicators are required to evolve objective assessment of this difference when comparing image containers before and after steganography. In the work, these indicators will be:

− MSE (mean square error) − standard deviation of a mathematical expectation:

$$\mathrm{MSE} = \frac{1}{XY} \cdot \sum_{x,y} (C_{x,y} - S_{x,y})^2 \,,$$

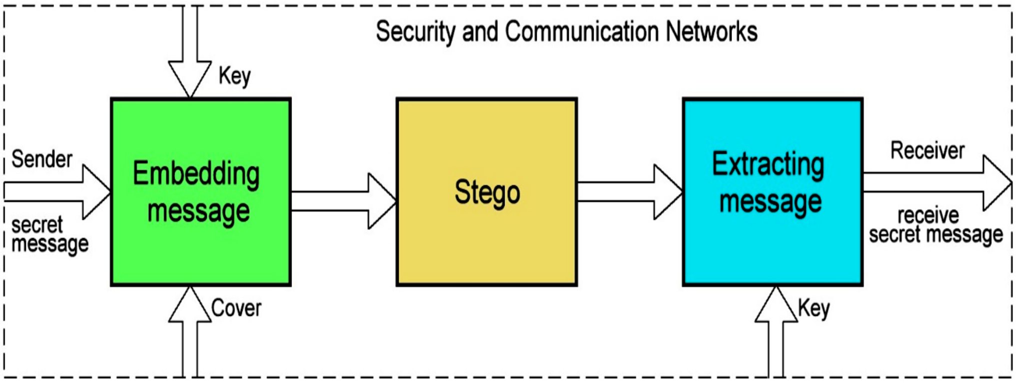


Fig. 7. Generic steganography scheme

where X, Y are the sizes of the cover image; C and S – cover image names before and after transformations;

− PSNR (peak signal-to-noise ratio) − the ratio of a signal possible maximum value and a power of the noise distorting the signal value; the indicator is calculated as:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{\left(\mathrm{C}_{x,y}^{\mathrm{MAX}}\right)^2}{\mathrm{MSE}},$$

where $\mathrm{C}_{x,y}^{\mathrm{MAX}}$ is the maximum value accepted by a pixel of the cover-image.

QR-codes as embedded image. One of the most promising techniques for encrypting text data, the authors consider using of QR-codes, and this has been implemented in the "HQ Scanner" release; in the preview window here, the option to generate a QR-code for a piece of the text after its recognition is available.

Revise that the newer the QR-code version, the more characters in the text can be embedded (version 40 can involve up to 3,000 characters, while a page of the dense A4 text contains about 2,500 characters). QR-code images are stored in *.png format. It is these QR-code image formats that are used as embedded-images in a cover-image if the transmitted over communication channel information needs to be protected by the LSB steganography method.

Secret data encoding algorithms. Consider an example. An occasional 600 x 400 pixel RGB picture was chosen as the cover-image (Fig. 8), and the embedded image would be a QR-code in three different versions for the encrypted text like “Most individuals don't think about numbers, or numerical representations of quantity, but they do play a major part in everyday life".

Let us accept the following constraints: the embedding image is treated as a grayscale picture (one channel), and the covering image will have the RGB format (three color channels) to enhance the total pixels, in which the attachment can be embedded.

In the example, we assume that embedding image has a fixed size (300 x 300 pixels), and each of its pixels is represented in a byte format. The encoding procedure for embedding (regardless of techniques) one image into another can be represented as the following algorithm:

1. Conversion of the two-dimensional array of pixels (300 x 300) for the embedding image into a one-dimensional array (1 x 90,000) (Fig. 9).

2. Conversion of the cover-image three-channel pixels array (600 x 400 x 3) into a one-dimensional array (1 x 720,000) (Fig. 10).

3. The numbers that represent each pixel's byte of the embedded-image are divided into groups with a definite number of bits. These bit groups are next used to replace the corresponding pixel bits of the cover-image. Over implementation, various replacement options are possible. The authors investigated 4 possible coding options, each has its own advantage and disadvantage. Consider each of the proposed options in more detail and choose the most advantageous one.

Algorithm 1. Herein, the 4 most significant bits from each byte of the embedding content are extracted to be located at the place of the 4 least significant bits of the cover image in the appropriate byte (Fig.11).

The scheme implementation results are shown in Fig. 12, and the corresponding calculated quality indicators are given in table 1.

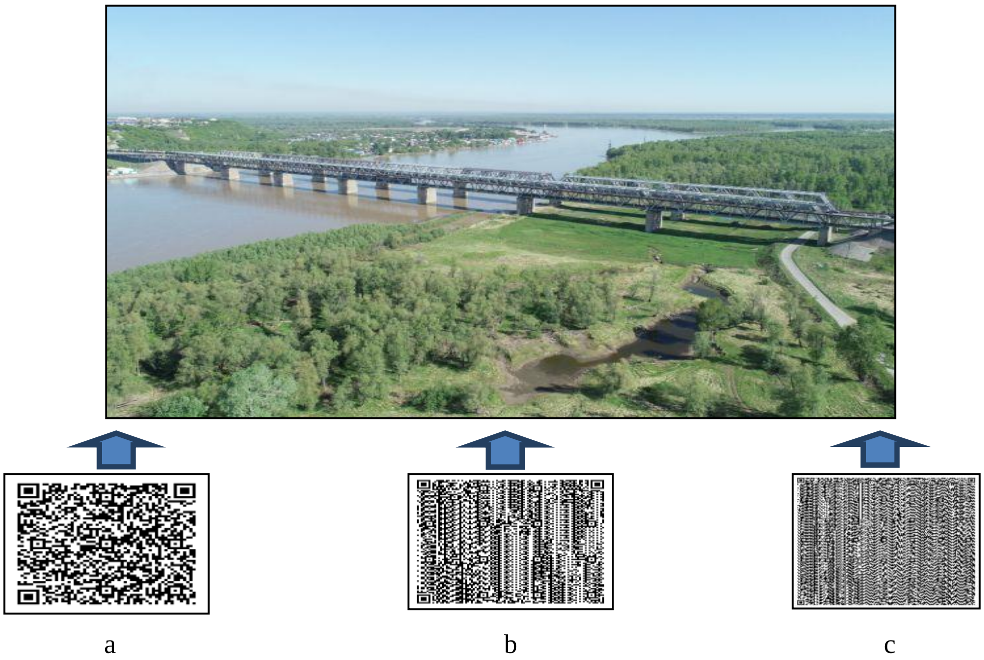


Fig. 8. General steganographic encoding scheme:
a − 10 version of the QR-code; b − 20 version of the QR-code; c − 40 version of the QR-code

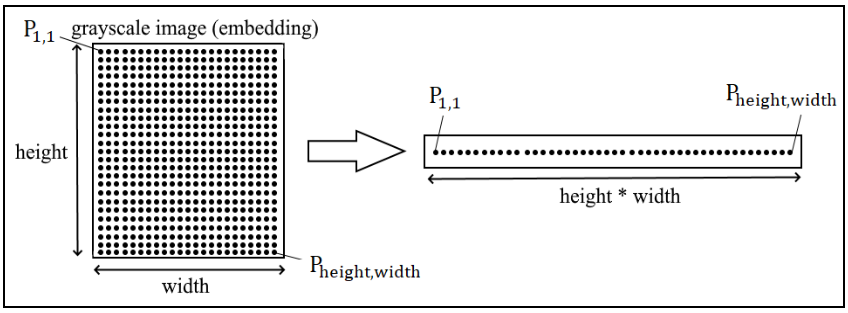


Fig. 9. Transformation of the embedding image

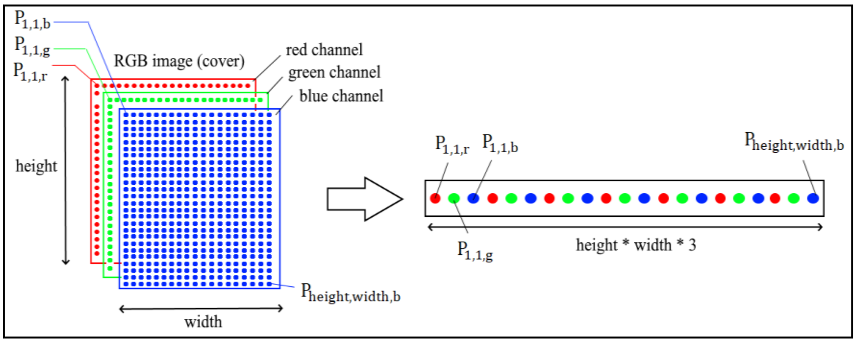


Fig.10. Cover image transformation

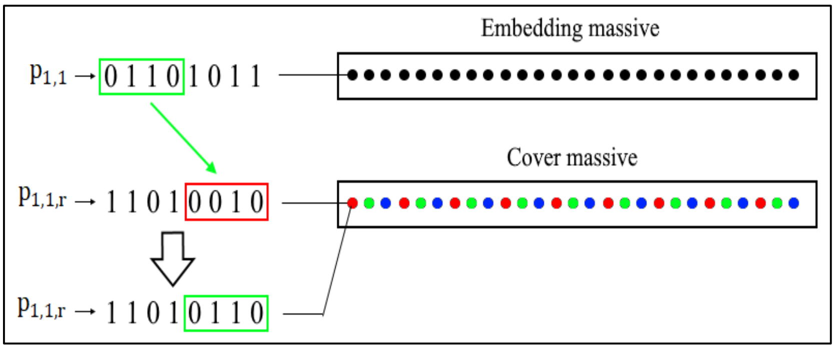
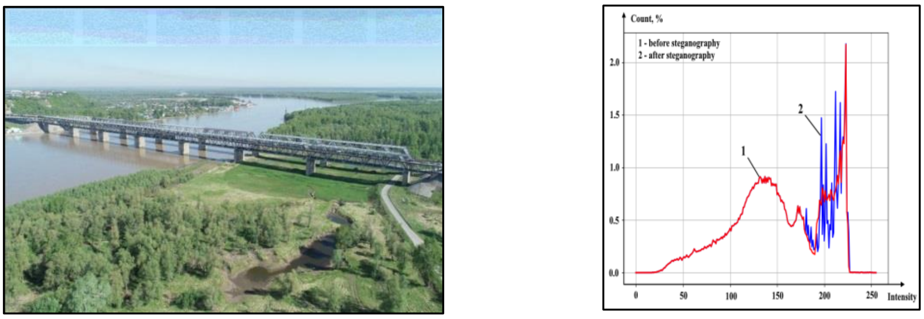


Fig. 11. Steganography algorithm 1: formatting cover image's 1st byte

Table 1

Statistical characteristics of steganographic transformations

QR-code type	MSE $_{\text{red channel}}$	MSE $_{\text{green channel}}$	MSE $_{\text{blue channel}}$	MSE $_{\text{total}}$	PSNR
QR-code (version 10)	8.9547	9.1357	11.7422	9.9442	38.1551
QR-code (version 20)	8.8306	9.0881	11.2795	9.7327	38.2485
QR-code (version40)	9.0308	9.3903	11.0065	9.8092	38.2144



a

b

Fig. 12. Cover image:
a – after steganographic encoding, b – corresponding histograms before and after transformation

Note that the encoded byte after decoding does not completely correspond to the original byte of the embedding-image after using this encoding option, since the least significant 4 bits are discarded during the encoding process. This feature leads to the fact that the embedded image

will be different, after decoding, from the original. So, this steganography option implies that the total number of bytes of the embedding image R_{emb} should not exceed the total number of bytes of the cover image R_{cov}:

$$R_{emb} \leq R_{cov} .$$

Algorithm 2. Each byte of the embedding image is hidden in the two bytes of the cover image, namely: the byte's 4 most significant bits have to replace the 4 lower bits of the cover image's first byte, and the least significant bits replace the 4 least significant bits of the cover-image's second byte (Fig. 13).

When using this steganography scheme, the bytes of the embedding image completely correspond after decoding to the original ones, since the 4 least significant bits are not discarded completely, as over previous scheme, but are stored in one more byte of the cover image pixel.

In this steganography technique, a set of bytes of the embedding image R_{emb} increased twice should not exceed the total number of bytes of the container image R_{cov}:

$$2R_{emb} \leq R_{cov} .$$

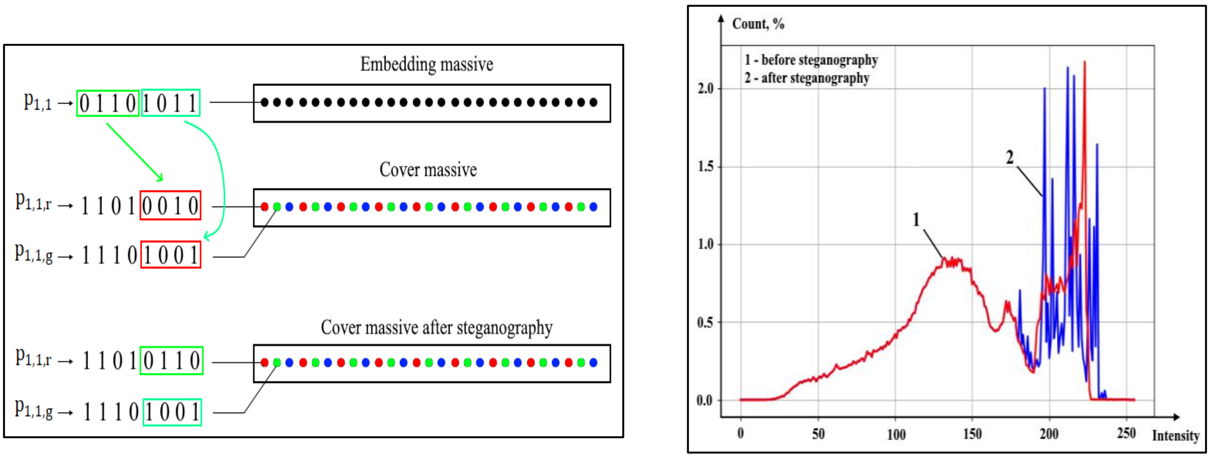
The scheme implementation results are shown in Fig. 13, and the corresponding quality indicators calculated are given in table 2.

Obviously, the coding results quality in the both options is not great, since after the attachment of the secret massive into the cover image consequences of the attachments are visually noticeable on it.

A characteristic visual marker of the encoding kind quality are the depicted jointly normalized brightness distribution histograms related to the cover image before and after encoding (with the corresponding images of embedded QR-codes; see Fig. 12 and Fig. 13). It is clearly seen that the brightness distribution function resulted due to the embedding takes on a more random behavior, where the brightness fluctuation are also clearly noticeable being almost the same in level in the both considered variants.

No less indicative are the outcomes in table 2, 3. The statistical data for the cover images difference before and after encoding are expedient to use them for selecting an optimal encoding scheme.

Algorithm 3. The embedding image byte is shared in the last two bits of the four bytes of the cover image (two bits in each) as shown in Fig. 13.



a

b

Fig. 13. Encoding algorithm 2:
a – forming first byte of the cover image, b – cover image histogram before and after transform

Table 2

Statistical characteristics of steganographic transformations

QR-code type	MSE $_{\text{red channel}}$	MSE $_{\text{green channel}}$	MSE $_{\text{blue channel}}$	MSE $_{\text{total}}$	PSNR
QR-code (version 10)	17.4515	18.4967	22.5156	19.4879	35.2331
QR-code (version 20)	18.1893	18.8327	22.3913	19.8044	35.1632
QR-code (version 40)	18.1153	18.4461	21.4748	19.3453	35.2651

Table 3

Statistical characteristics of steganographic transformations

QR-code type	**MSE** **red channel**	**MSE** **green channel**	**MSE** **blue channel**	**MSE** **total**	**PSNR**
QR-code (version 10)	1.6766	1.6723	1.6736	1.6742	45.8927
QR-code (version 20)	1.7251	1.7013	1.7046	1.7103	45.8000
QR-code (version 40)	1.6902	1.6530	1.6917	1.6783	45.8821

Similar to the previous technique, all information bits of the embedding image are completely retained while encoding, which makes it possible to obtain an identical image after decoding.

In this steganography technique, the total number of bytes of the cover image R_{emb} must cover the four times number of bytes of the embedding image R_{cov}:

$$4R_{emb} \le R_{cov} .$$

The results of the implementation scheme are shown in Fig. 14, quality indicators are given in table 3.

Note that the 3rd coding option gives much better performance than the previous two. First of all, footprints of the embedding image are not observed in the cover image, and the histograms brightness index dispersion after the secret image is inserted into the cover image is much smaller. Besides, the statistical indicators for the difference of the cover images before and after the encoding procedure provision are also significantly lower.

Algorithm 4. Each byte bit of the embedding image is hidden with the last bit in a set of eight bytes of the cover image.

This encoding-decoding routine is similar to the previous algorithm of option 3 with the difference that now one information byte is encoded into 8 bytes of the cover image instead of 4.

The technique demands that the total number of bytes of the cover image R_{cov} would cover the eight times number of bytes of the embedding image R_{emb}:

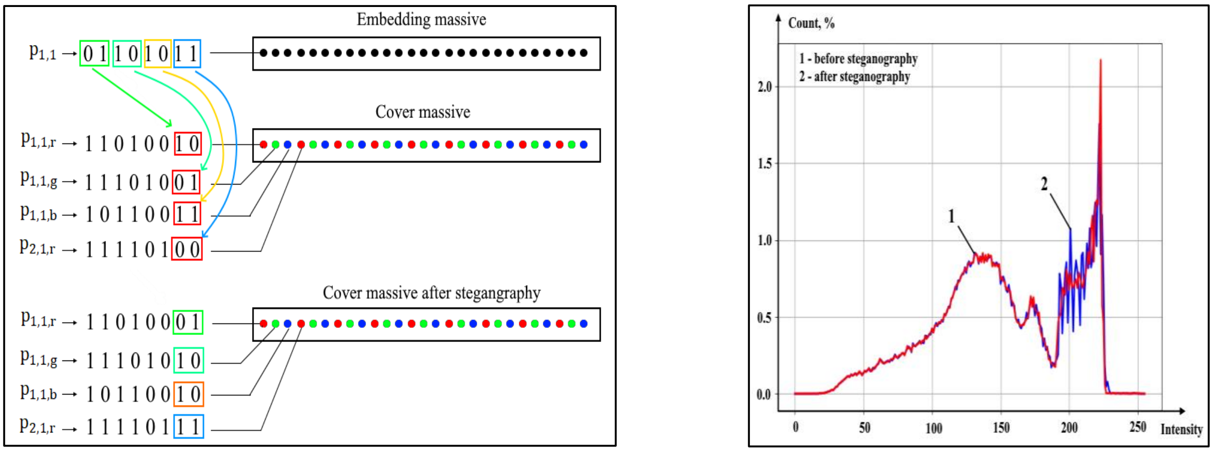
$$8R_{emb} \le R_{cov} .$$

The corresponding quality indicators calculated are given in table 4.

Fig. 15 shows the dependence graphs for the MSE and PSNR indicators on the steganography scheme related to the different options of the QR-code used in the embedded image. It can be seen that the graphs for different QR options coincide, which means that using a different version of the QR-code affects a little to the quality.

It is evident that the steganography techniques of the 1st and 2nd schemes with their maximum MSE and minimum PSNR are the least suitable for a practical application. This is due to the fact that the bits of the embedding image replace within the option 2 the 4 least significant bits in the 2 bytes of the cover image. This leads to seeable distortions of the cover image and, hence, to enhanced values of the MSE.

The option 1 (the bits of the embedding-image are also encoded in the least significant 4 bits of the cover image) results in a lower MSE value compared to the option 2, although it also leads to distortion of the embedding image after decoding, as the least significant 4 bits of each byte of the embedding image are again discarded.



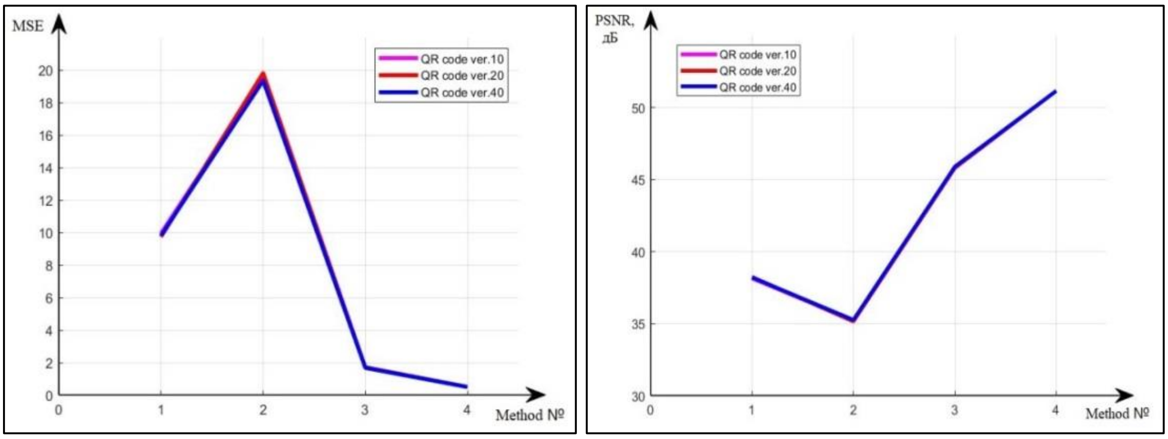
a b

Fig. 14. Encoding algorithm 3:
a – forming first byte of the cover image, b – cover image histogram before and after transform

Table 4

Statistical characteristics of steganographic transformations

QR-code type	MSE red channel	MSE green channel	MSE blue channel	MSE total	PSNR
QR-code (version 10)	0.4977	0.5011	0.4982	0.4989	51.1499
QR-code (version 20)	0.5027	0.4975	0.4980	0.4994	51.1463
QR-code (version 40)	0.5014	0.4981	0.4984	0.4993	51.1471



a b

Fig. 15. Embedding quality indicators dependence on the scheme and the volume QR-code: a – MSE, b – PSNR

The smallest MSE is achieved in the 4th coding option, since in this case only the very last significant bits of the cover image bytes are being distorted. This option has the best performance with respect to both MSE and PSNR, but it is worth to note, when using the embedding in the last bits, various external noises can produce a noticeable effect on the cover image. In turn, this can result in a loss of information in the embedding image. Also, this option requires significantly bigger number of bytes of the cover image for adequate encoding.

Therefore, the most acceptable is the steganography scheme 3 that yields a low MSE value, requires fewer bytes of the cover image compared to the scheme 4, and it is less sensitive to noises in the cover image transfer.

Cover image selection. In addition a special investigation was held that related the analysis of features of the cover images appropriate for steganography. Nine different cover images were chosen (Fig. 16) and the corresponding MSE and PSNR indicators at every of mentioned steganography techniques were calculated. In each scheme study, a 40-size QR-code was used to encode the embedding image. The results are shown in Fig. 17.

The graphs consideration shows that the use of the steganography options 1, 2 yields the dependence of the mean-square error and signal-to-noise ratio values on the container image choice (seen curvature of the corresponding plots). When using the options 3, 4 there are no such effects; this testifies getting of declared steganography outcomes independently on the cover image chosen to bring the procedure.

In the conclusion, we can constitute that using of the steganography scheme specified as Option 3 hereabove can be used to approach the result, which depends neither upon the QR code version encrypting the embedding image nor on the pattern picture chosen for the container.

5. Conclusion

The concept of significant improvement in the quality of the operation of OCR systems has been proposed and successfully implemented due to the complex use of algorithms for preliminary processing of document images, an expansion of the set of service functions and the use of methods and algorithms for information protection.

Pre-processing of the initial data made it possible to achieve almost 100% accuracy of text recognition. Particularly useful in this project are the service functions of interactive segmentation and anonymization of text documents, which allow ensuring data confidentiality at the stage of preliminary preparation.

It is possible to encode text information in the form of QR codes. This makes it possible not only to encrypt the documentation, but also to provide an original method for the secrecy of data transmission over open communication channels using steganography methods.

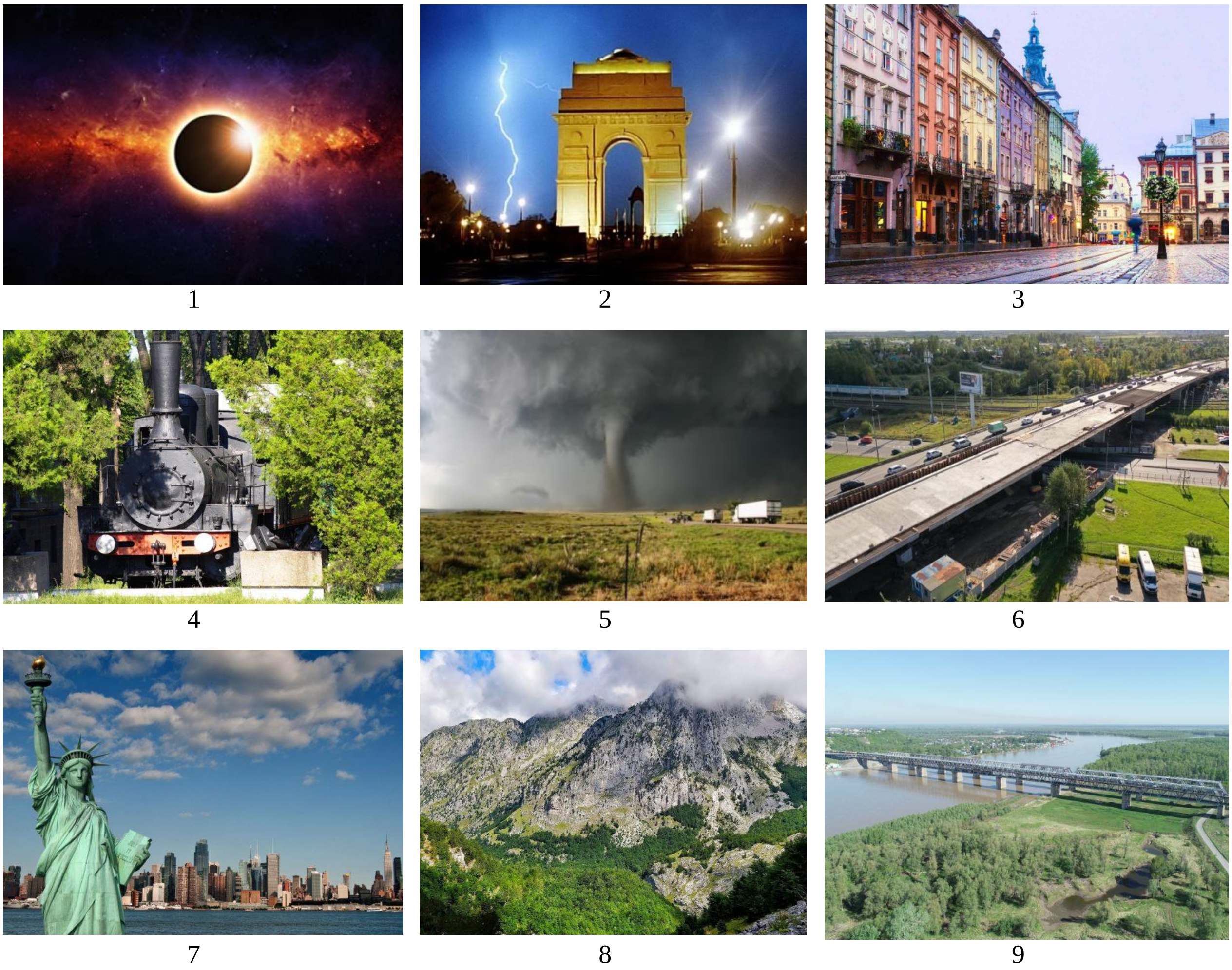


Fig. 16. A set of cover images

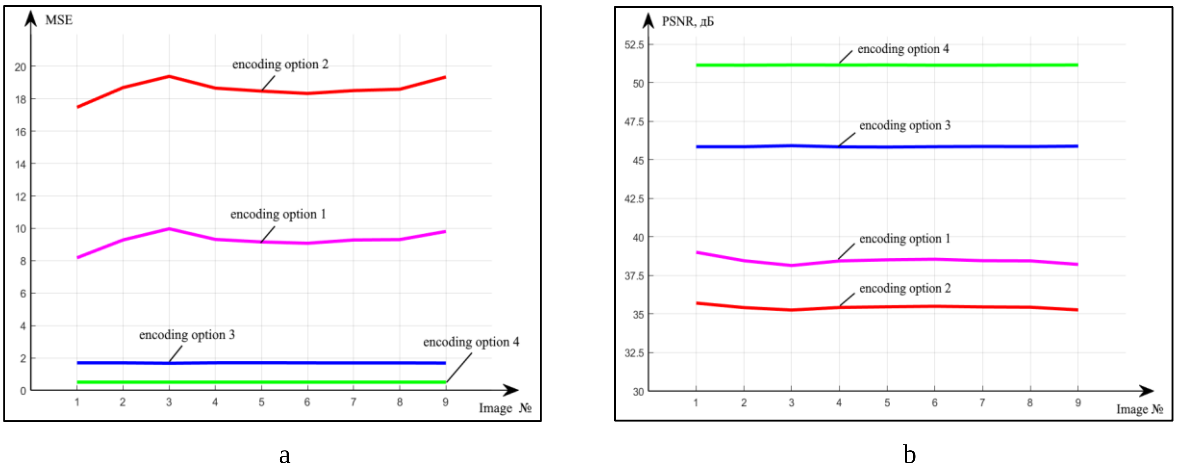


Fig. 17. Steganography quality variation gotten at cover image change: a – MSE, b – PSNR

In scientific terms, the authors of the project consider it expedient to further develop the LBS steganography technology using the presentation of text information in the QR code format for more effective information protection in optical text recognition systems.

Contribution of authors: analysis of present information resources related to improving quality of OCR systems. Formulating the problem of study the methods and algorithms required to protect information in optical text recognition systems; specifying requirements for their practical implementation – **K. Dergachov**; suggested to encrypt obtained in the result of optical text recognition confidential information with using QR-codes, and to use LSB-steganography algorithms while transmitting this information through open communication channels (like e-mails). Several proposed security algorithms have been implemented into practice – **L. Krasnov**; selecting, use and optimization of modern programming methods and relevant information resources for implementation of the information security algorithms proposed; testing of the created software; conducting relevant experimental studies – **V. Bilozerskyi**; formulation of the main conclusions based on the results of the research and recommendations for their practical use, carrying the paper general editing, translating the source materials from Ukrainian to English – **A. Zymovin**. All authors have read and agreed to the published version of the manuscript.

References (GOST 7.1:2006)

1. Tesseract − ocr/Tesseract [Electronic resource]. – Available at: https://github.com/tesseract-ocr/tesseract. − 18.05.2021.

2. Python-tesseract − Optical character recognition (OCR) tool for Python [Electronic resource]. – Available at: https://pypi.org/project/pytesseract/. − 18.05.2021.

3. A Study on Optical Character Recognition-Techniques [Text] / N. Sahu, M. Sonkusare // The International Journal of Computational Science, Information Technology and Control Engineering (IJCSITCE). – 2017. – Vol. 4, No. 1. – 14 p. DOI: 10.5121/ ijcsitce.2017.4101.

4. Offline optical character recognition (OCR) method: An effective method for scanned documents [Text] / Mujibur Rahman Majumder et al. // 22nd International Conference on Computer and Information Technology (ICCIT) – 2019. – P. 1-5. DOI: 10.1109/ ICCIT48885.2019.9038593.

5. Improved OCR quality for smart scanned document management system [Text] / Anh Phan Viet et al. / Journal of Science and Technique − Le Quy Don Technical University. – 2020. − No. 210. – P. 51-67.

6. Image to Text Conversion Using Tesseract [Text] / N. Pawar, Z. Shaikh, P. Shinde, Y. Warke // International Research Journal of Engineering and Technology (IRJET). – 2019, – Vol. 6, iss 02. – P. 516-519.

7. Scan.it − on Advances in Computing, Communication and Control (Text Recognition, Translation and Conversion) [Text] / M. Acharya, P. Chouhan, A. Deshmukh // International Conference (ICAC3). – 2019. – P. 1-5. DOI: 10.1109/ICAC347590.2019. 9036849.

8. OpenCV Tutorials − Image Processing (imgproc module) [Electronic resource]. – Available at: https://opencv.org/ − 18.05.2021.

9. OpenCV / dnn modules [Electronic resource]. –. Available at: https://github.com/opencv/opencv/tree/master/modules/dnn. − 18.05.2021).

10. Data pre-processing to increase the quality of optical text recognition systems [Text] / K. Dergachov, et al. // Радіоелектронні і комп'ютерні системи. − 2021. – № 4(100). – P. 183-198. DOI: 10.32620/reks.2021.4.15.

11. A Secure QR Code System for Sharing Personal Confidential Information [Text] / M. S. Ahamed, H. Mustafa Asiful // International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). – 2019. – P. 1-4, DOI: 10.1109/ IC4ME247184.2019.9036521.

12. Some Methods of QR code Transmission using Steganography [Text] / D. F. Pastukhov et al. // World of transport and transportation. − 2019. – Vol. 17, Iss. 3. – P. 16–39.

13. Efficiency Assessment of the Steganographic Coding Method with Indirect Integration of Critical Information [Text] / O. Yudin et al. // IEEE International Conference on Advanced Trends in Information Theory (ATIT). – 2019. – P. 36-40. DOI: 10.1109/ATIT49449.2019.9030473.

14. PSNR and MSE based investigation of LSB [Text] / K. Joshi, R. Yadav and S. Allwadhi // International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT). – 2016. – P. 280-285. DOI: 10.1109/ICCTICT. 2016. 7514593.

15. QR code image steganography (LSB BIT) with secret image (MSB BIT) using AES cryptography and JPEG compression [Text] / R. Rituraj et al. // Recent Scientific Research. − 2019. – Vol. 9, Issue, 7. – P. 27820-27826.

16. Li, F. Two-step providing of desired quality in lossy image compression by SPIHT [Text] / F. Li, S. Krivenko, V. Lukin // Радіоелектронні і комп'ютерні системи. − 2020. – № 2(94). – С. 22-32. DOI: 10.32620/reks.2020.2.02.

17. Objective Quality Metrics in Correlation with Subjective Quality Metrics for Steganography [Text] / R. Wazirali et al. // Asia-Pacific Conference on Computer Aided System Engineering. – 2015. – P. 238-245, DOI: 10.1109/APCASE.2015.49.

18. Keyword Detection Based on RetinaNet and Transfer Learning for Personal Information Protection in Document Image [Text] / G.-S. Lin et al. // Appl. Sci. – 2021. – Vol. 11. – Article No. 9528. DOI: 10.3390/app11209528.

19. A Method of Image Quality Assessment for Text Recognition on Camera-Captured and Projectively Distorted Documents [Text] / J. Shemiakina, et al. // Mathematics. – 2021. – Vol. 9. – Article No. 2155. DOI: 10.3390/ math9172155.

20. Business Process Automation: A Workflow Incorporating Optical Character Recognition and Approximate String and Pattern Matching for Solving Practical Industry Problems [Text] / C. de Jager et al. // Appl. Syst. Innov. – 2019. – Vol. 2, No. 4. – Article No. 33. DOI: 10.3390/asi2040033.

21. Manual character recognition with OCR [Text] / Sasmitha Kumari Sahu et al. // Project. – 2021. DOI: 10.13140/RG.2.2.32608.81927.

References (BSI)

1. *Tesseract − ocr / Tesseract.* Available at: https://github.com/tesseract-ocr/tesseract. (accessed 18.05.2021).

2. *Python-tesseract − Optical character recognition (OCR) tool for Python*. Available at: https: //pypi.org/project/ pytesseract/. (accessed 18.05.2021).

3. Sahu, N., Sonkusare, M. A Study on Optical Character Recognition-Techniques. *The International Journal of Computational Science, Information Technology and Control Engineering* (IJCSITCE), 2017, vol. 4, no. 1. 14 p. DOI: 10.5121/ijcsitce.2017.4101.

4. Mujibur Rahman Majumder et al. Offline optical character recognition (OCR) method: An effective method for scanned documents. *22nd International Conference on Computer and Information Technology (ICCIT)* – 2019, pp. 1-5. DOI: 10.1109/ICCIT48885. 2019. 9038593.

5. Viet, Anh Phan. et al. Improved OCR quality for smart scanned document management system. *Journal of Science and Technique − Le Quy Don Technical University*, 2020, no. 210, pp. 51-67.

6. Pawar, N., Shaikh, Z., Shinde, P., Warke, Y., Image to Text Conversion Using Tesseract. *International Research Journal of Engineering and Technology (IRJET),* 2019, vol. 6, iss 2, pp. 516-519.

7. Acharya, M., Chouhan, P., Deshmukh, A. Scan.it - on Advances in Computing, Communication and Control (Text Recognition, Translation and Conversion). *International Conference (ICAC3),* 2019, pp. 1-5. DOI: 10.1109/ICAC347590.2019. 9036849.

8. *OpenCV Tutorials − Image Processing (imgproc module).* Available at: https://opencv.org/ (accessed 18.05.2021).

9. *OpenCV / dnn modules.* Available at: https://github.com/opencv/opencv/tree/master/modules/ dnn (accessed 18.05.2021).

10. Dergachov, K. et al. Data pre-processing to increase the quality of optical text recognition systems. *Radioelektronni i komp'uterni sistemi – Radioelectronic and computer systems,* 2021, no. 4(100), pp. 183-198. DOI: 10.32620/reks.2021.4.15.

11. Ahamed, M. S., Asiful, Mustafa H. A Secure QR Code System for Sharing Personal Confidential Information. *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2),* 2019, pp. 1-4, DOI: 10.1109/IC4ME247184.2019.9036521.

12. Pastukhov, D. F. et al. Some Methods of QR code Transmission using Steganography. *World of transport and transportation,* 2019, vol. 17, Iss. 3, pp. 16–39.

13. Yudin, O. et al. Efficiency Assessment of the Steganographic Coding Method with Indirect Integration of Critical Information. *IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, 2019, pp. 36-40, DOI: 10.1109/ATIT49449.2019. 9030473.

14. Joshi, K. et al. PSNR and MSE based investigation of LSB. *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT),* 2016, pp. 280-285, DOI: 10.1109/ICCTICT.2016.7514593.

15. Rituraj, R. et al. QR code image steganography (LSB BIT) with secret image (MSB BIT) using AES cryptography and JPEG compression. *International Journal of Recent Scientific Research,* 2019, vol. 9, Issue, 7, pp. 27820-27826.

16. Li, F., Krivenko, S., Lukin, V. Two-step provsding of desired quality in lossy image compression by spiht. *Radioelektronni i komp'uterni sistemi – Radioelectronic and computer systems*, 2020, no. 2(94), pp. 22-32. DOI: 10.32620/reks.2020.2.02.

17. Wazirali, R. et al. Objective Quality Metrics in Correlation with Subjective Quality Metrics for Steganography. *Asia-Pacific Conference on Computer Aided System Engineering,* 2015, pp. 238-245, DOI: 10.1109/APCASE.2015.49.

18. Lin, G.-S. et al. Keyword Detection Based on RetinaNet and Transfer Learning for Personal Information Protection in Document Image. *Appl. Sci.*, 2021, vol. 11, article no. 9528. DOI: 10.3390/app11209528.

19. Shemiakina, J. et al. A Method of Image Quality Assessment for Text Recognition on Camera-Captured and Projectively Distorted Documents. *Mathematics*, 2021, vol. 9, article no. 2155. DOI: 10.3390/math9172155.

20. De Jager, C. et al. Business Process Automation: A Workflow Incorporating Optical Character Recognition and Approximate String and Pattern Matching for Solving Practical Industry Problems. *Appl. Syst. Innov.*, 2019, vol. 2, no. 4, article no. 33. DOI: 10.3390/asi2040033.

21. Sasmitha Kumari Sahu et al. Manual character recognition with OCR. Project, 2021 DOI: 10.13140/RG.2.2.32608.81927.

Надійшла до редакції 24.11.2021, розглянута на редколегії 16.02.2022

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

МЕТОДИ ТА АЛГОРИТМИ ЗАХИСТУ ІНФОРМАЦІЇ У СИСТЕМАХ ОПТИЧНОГО РОЗПІЗНАВАННЯ ТЕКСТІВ

К. Ю. Дергачов, Л. О. Краснов, В. О. Білозерський., А. Я. Зимовин

Предмет дослідження. Пропонується концепція підвищення якості роботи систем оптичного розпізнавання текстів за рахунок комплексного використання алгоритмів попередньої обробки зображень документів, розширення набору сервісних функцій та використання методів та алгоритмів захисту інформації. **Мета дослідження**. Запропонувати набір алгоритмів, що компенсують вплив зовнішніх негативних впливів (несприятливий геометричний фактор, погані умови освітлення під час фотографування, вплив шумів та ін.) на ефективність роботи систем розпізнавання текстів. Необхідно передбачити низку сервісних процедур, які забезпечують зручність обробки даних (їх перегляд, перетворення та збереження у стандартних форматах, забезпечення можливості обміну даними у відкритих комунікаційних мережах). Крім цього необхідно забезпечити захист інформації від несанкціонованого використання на етапі обробки даних та забезпечити скритність їх передачі каналами зв'язку. **Використані методи та результати досліджень:** Розроблено та протестовано алгоритми попередньої обробки даних (геометричні перетворення вихідних зображень, їх обробка набором різних фільтрів, бінаризація з адаптивними порогами перетворення для усунення впливу нерівномірного засвічення фото). Передбачено низку сервісних процедур, що забезпечують зручність обробки даних та їх інформаційний захист. Зокрема, запропоновано інтерактивну процедуру сегментації тексту з можливістю анонімізації окремих його фрагментів, що сприяє збереженню конфіденційності документів, що обробляються. У пакет алгоритмів обробки фотознімків документації підвищення можливостей з ідентифікації даних вбудований алгоритм детектування осіб (face detection), призначений подальшого їх розпізнавання (face recognition). Після розпізнавання текстової документації захист отриманих даних здійснюється шляхом генерації відповідних QR-кодів, а скритність передачі забезпечується методами стеганографії. Докладно описано структуру алгоритмів та досліджено стійкість їх роботи в різних умовах. За результатами проведених досліджень, розроблено програмне забезпечення для розпізнавання текстів, що базується на програмі оптичного розпізнавання символів (OCR) Tesseract версії 4.0. Програма отримала назву HQ Scanner, вона написана мовою Python з використанням сучасних ресурсів бібліотеки OpenCV. Програмно реалізовано оригінальну методику оцінки ефективності роботи алгоритмів за критерієм максимуму ймовірності правильного розпізнавання текстів. **Висновки.** Результати проведених досліджень є основою розробки програмного забезпечення та створення простих у використанні систем оптичного розпізнавання тексту для комерційного використання.

Ключові слова: Оптичне розпізнавання символів; можливість правильного розпізнавання тексту; сегментація тексту та анонімізація його фрагментів; QR-коди; алгоритми стеганографії.

МЕТОДЫ И АЛГОРИТМЫ ЗАЩИТЫ ИНФОРМАЦИИ В СИСТЕМАХ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ ТЕКСТОВ

К. Ю. Дергачёв, Л. А. Краснов, В. А. Билозерский, А. Я. Зимовин

Предмет исследования. Предлагается концепция повышения качества работы систем оптического распознавания текстов за счет комплексного использования алгоритмов предварительной обработки изображений документов, расширения набора сервисных функций и использования методов и алгоритмов защиты информации. **Цель исследования**. Предложить набор алгоритмов, компенсирующих влияние внешних негативных воздействий (неблагоприятный геометрический фактор, плохие условия освещения при фотографировании, влияние шумов и пр.) на эффективность работы систем распознавания текстов. Необходимо предусмотреть ряд сервисных процедур, обеспечивающих удобство обработки данных (их просмотр, преобразование и сохранение в стандартных форматах, обеспечение возможности обмена данными в открытых коммуникационных сетях). Кроме этого нужно обеспечить защиту информации от несанкционированного использования на этапе обработки данных и обеспечить скрытность их передачи по каналам связи. **Используемые методы и результаты исследований:** Разработаны и протестированы алгоритмы предварительной обработки данных (геометрические преобразования исходных изображений, их обработка набором различных фильтров, бинаризация с адаптивными порогами преобразования для устранения влияния неравномерной засветки фото). Предусмотрен ряд сервисных процедур, обеспечивающих удобство обработки данных и их информационную защиту. В частности, предложена интерактивная процедура сегментации текста с возможностью анонимизации отдельных его фрагментов, что способствует сохранению конфиденциальности обрабатываемых документов. В пакет алгоритмов обработки фотоснимков документации для повышения возможностей по идентификации данных встроен алгоритм детектирования лиц (face detection), предназначенный для дальнейшего их распознавания (face recognition). После распознавания текстовой документации защита полученных данных осуществляется путём генерации соответствующих QR-кодов, а скрытность передачи информации обеспечивается методами стеганографии. Подробно описана структура алгоритмов и исследована устойчивость их работы в различных условиях. По результатам проведенных исследований разработано программное обеспечение для распознавания текстов, базирующееся на программе оптического распознавания символов (OCR) Tesseract версии 4.0. Программа получила название «HQ Scanner», она написана на языке Python с использованием современных ресурсов библиотеки OpenCV. Программно реализована оригинальная методика оценки эффективности работы алгоритмов по критерию максимума вероятности правильного распознавания текстов. **Выводы.** Результаты проведенных исследований являются основой для разработки программного обеспечения и создания простых в использовании систем оптического распознавания текста для коммерческого использования.

Ключевые слова: Оптическое распознавание символов; вероятность правильного распознавания текста; сегментация текста и анонимизация его фрагментов; QR-коды; алгоритмы стеганографии.

Дергачов Костянтин Юрійович − канд. техн. наук, старш. наук. співроб., зав. каф. систем управління літальних апаратів, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Краснов Леонід Олександрович – канд. техн. наук, старш. наук. співроб., доцент каф. управління літальних апаратів, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Білозерський Владислав Олександрович – магістр каф. систем управління літальних апаратів, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Зимовин Анатолій Якович − канд. техн. наук, проф., проф. каф. управління літальних апаратів, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Konstantin Dergachov − Candidate of Technical Science, Senior Researcher, Head of the Department «Aircraft Control Systems», National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: k.dergachov@khai.edu, ORCID: 0000-0002-6939-3100.

Leonid Krasnov – Candidate of Technical Science, Senior Researcher, Assistant Professor of Department «Aircraft Control Systems», National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: leonid.krasnov.1947@gmail.com, ORCID: 0000-0003-2607-8423.

Vladislav Bilozerskyi – master student of Department «Aircraft Control Systems», National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e mail: b1llays123@gmail.com, ORCID: 0000-0002-5503-3163.

Anatoly Zymovin − Candidate of Technical Science, Professor of Department «Aircraft Control Systems», National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: zim301g@gmail.com, ORCID: 0000-0001-8580-2317.