

*Розглянуто особливості розробки нової інформаційної технології розпізнавання символів тексту, яка ґрунтується на новому типі клітинних автоматів, конкуруючих клітинних автоматах. Перевагами цієї технології є простота правил взаємодії, можливість легкого розпаралелювання процесу розпізнавання, можливість розпізнавання спотворених та частково накладених символів. Проведене порівняння якості роботи та швидкодії з комерційною системою ABBYY FineReader продемонструвало порівнянні експлуатаційні показники розробленої технології розпізнавання*

*Ключові слова: конкуруючий клітинний автомат, рухомий клітинний автомат, автомат Мура, граф переходів*

*Рассмотрены особенности разработки информационной технологии распознавания символов текста, базирующаяся на типе клеточных автоматов, конкурирующих клеточных автоматах. Преимущества такой технологии заключаются в простоте правил взаимодействия, возможности легкого распараллеливания процесса распознавания, возможность уверенной работы с искаженными и частично наложенными символами. Выполненные сравнительные эксперименты с коммерческой системой ABBYY FineReader продемонстрировали сравнимые эксплуатационные показатели разработанной технологии распознавания*

*Ключевые слова: конкурирующий клеточный автомат, подвижный клеточный автомат, автомат Мура, граф переходов*

UDC 004.93'12; 004.932; 004.932.75

DOI: 10.15587/1729-4061.2017.103550

# REALIZATION OF INFORMATION TECHNOLOGY OF CHARACTER RECOGNITION BASED ON COMPETING CELLULAR AUTOMATA

**I. Myroniv**

Postgraduate student\*

E-mail: ivan.myroniv@gmail.com

**V. Zhikharevich**

PhD, Associate Professor\*

**S. Ostapov**

Doctor of Physics and Mathematics Sciences,  
Professor, Head of Department\*

E-mail: sergey.ostapov@gmail.com

\*Department of Computer Systems Software  
Yuriy Fedkovych Chernivtsi National University

Kotsyubynskoho str., 2,  
Chernivtsi, Ukraine, 58012

## 1. Introduction

The systems of character recognition, which are part of the tasks on pattern recognition, are used today in many areas of human activity: from document management systems and all the way up to the modern-day robotics [1–5].

Similar to any intelligent system, character recognition requires a modular approach that allows the use of different principles and technologies of recognition, processing of incoming information, and interaction with hardware.

At present, there many different commercially available recognition systems, for example, CuneiForm, Finereader, Readiris and others. Each of these systems proposes its own version of solving the problems on pattern processing and recognition. In most cases, these are commercial program, which is why techniques underlying their operation, as well as their implementation, are known only to the developers, which greatly complicates the analysis and comparison of software from different manufacturers. As far as the recognition programs with open source are concerned, their functionality and requirements for computational capacities often exceed the capabilities of modern systems. Therefore, development of new methods and information technologies for the recognition of characters, which do not require significant computing resources and employ modern embedded software, is a relevant scientific and technical task.

## 2. Literature review and problem statement

Very attractive from the point of view of development of new information technologies of recognition is the use of cellular automata. In paper [6], author demonstrated indisputable advantages of applying cellular automata (CA) for the problems where there is a need for parallelization of computations that enables simple implementation of complex algorithms for pattern processing, which does not require significant computing resources. Despite these advantages, cellular automata have not been often employed in the tasks on recognition. We could not find a better-detailed research in this area than article [7], the main part of which addresses the study into characteristics of CA in the processes of text recognition. The author uses sequences of different CA to highlight characteristic attributes of text characters: loops, junctions, end points. Paper [8] examines cellular automata on the possibility to recognize handwritten characters. The main disadvantages of such approaches are that they are cumbersome and needed to be trained.

Other studies related to engaging CA to such problems are either outdated or address the use of CA in different tasks accompanied by recognition problems. Thus, in article [9], author proposed a new algorithm for the recognition from a JPEG image of watermarks based on cellular automata.

ata, while [10] constructed specialized cellular automata structures to analyze a contour image [9, 10].

However, there is a real way to apply a new, proposed by authors of [11–14], type of CA directly to the process of character recognition. Such approach is based on movable CA, which should implement all their states on the appropriate text symbol. Ambiguity in the interpretations of characters, which occurs in this case, is compensated for by the developed mechanism of competition when the “winner” is the CA with the maximal number of implemented states. This CA is actually the most correct reflection of the character that it recognizes. Thus, article [11] proposed the type of CA that allows recognition of correct symbols in the text of hexadecimal number system. Paper [12] improved CA, which made it possible to recognize the characters in the English alphabet, while article [13] added to the transition graphs of our CA the tags of states, which allowed the authors to identify character patterns with improved quality and better performance efficiency.

At the same time, all the examined methods and systems of recognition ineffectively work with handwritten and partially distorted characters, as well as the characters that partially overlap.

### 3. The aim and objectives of the study

The goal of present study is to develop information technologies for the recognition of text characters based on the advanced ideas about competing CA, which would allow, in addition to the printed characters, effective recognition of deformed characters, or those that partially overlap.

To accomplish the set goal, the following tasks have been solved:

- obtaining a graphic image of the text using scanning techniques or a camera;
- input image should be cleared from noise and brought to the form, which makes it possible to effectively highlight the characters and recognize them;
- the image has to be segmented, that is divided into rows and characters based on the peculiarities of its alignment;
- the image of the character is processed as a whole; for this purpose, a set of types of CA is applied to it, which, as a result of the interaction, would yield the most probable variant of the recognized character.

### 4. Stages of the proposed information technology

Fig. 1 shows a general scheme of the developed information technology of character recognition based on competing CA in the IDEF0 notation.

We shall consider in more detail the structure of the information technology. The input information for the developed information technology is the hard copies of documents (texts). It is also possible to apply the proposed method of character recognition for the characters of a handwritten text (the so-called “printed” letters).

Depending on the variant (desktop or mobile) of software, we scan or photograph the input text (block 1). Scanning is performed by the user by means of software modules that employ algorithms that are intended to do it. As a result, we receive a scanned image of the input text that is sent to

the input of block 2, which conducts pre-processing of the image: it makes it possible to change brightness/contrast, uses noise reduction algorithms for the optimization of recognition. Here the language of the document is selected. The optimized image, depending on the needs, is stored to a disk, or sent to the input of block 4. This block executes image segmentation algorithms; the image is split into lines and letters, which are basically the objects of recognition.

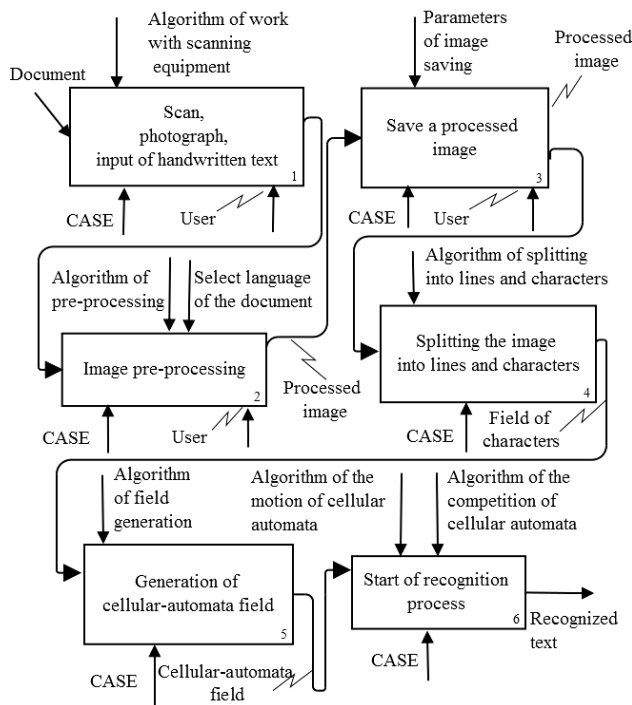


Fig. 1. General scheme of the developed information technology for character recognition based on the competing cellular automata

The field of letters (their images) is sent to the input of block 5, which is responsible for generating a cellular-automata field. All types of CA that are in charge of recognition of characters of the text in the language chosen are randomly arranged on a cellular-automata field. The result is sent to block 6.

Block 6 starts the algorithms of motion and competition of CA, which lead to the recognition of text characters. Data of the new quality, as a result of implementation of the given information technology, is the recognized text that is displayed in a certain assigned way.

Stages of actual recognition require additional details that are shown in Fig. 2. All steps are performed by the software and thus do not require user intervention.

Fig. 2, which demonstrates the process of character recognition, shows that a cellular-automata field is sent to the input of these stages, on which the generation of CA is initiated, which are responsible for every character of the language of the original. CA are generated randomly, uniformly all over the entire field with the images of letters. Next, the mechanism of the motion of CA over the images of characters is executed (block 2). If CA does not hit the image of the character, it remains stationary and is removed from the cellular-automata field.

The motion of CA is understood here as its transitions over the totality of states that it owns. The number and order of achieving the states of CA depends on which letter the

automaton describes. For example, Fig. 3 shows the graphs of states of CA, which describe numbers 0, 2 and letter C.

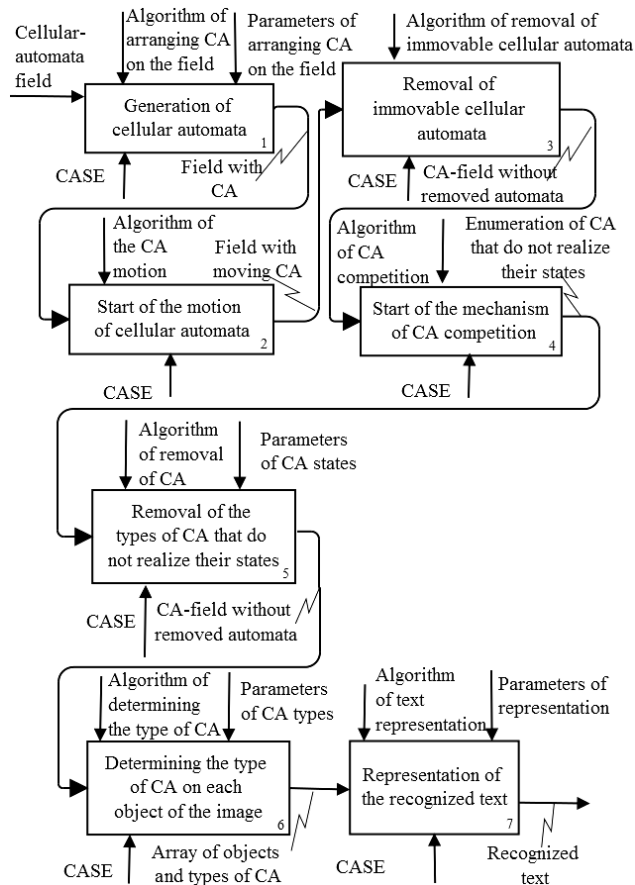


Fig. 2. Diagram of the stages of the process of character recognition based on cellular automata

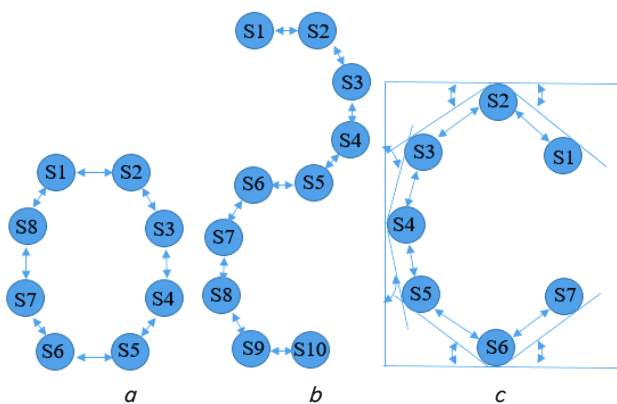


Fig. 3. State diagrams of a Moore machine that describe: a – number 0; b – number 2; c – character C

A set of CA matches a recognition language; the set of states is unique for each character. Thus, if, during the motion of CA over the image of a character, it achieves all its states, then there is a high probability that it describes it correctly. More details on the algorithm of CA motion can be found in [12].

However, the situations are likely when CA implements all its states, but does not correspond to the given character. A striking example is CA that describes number “1”,

which hit numbers “0”, “3”, “4”, “7”, “8” or “9”. This CA achieves all its states but do not describe the character completely.

In order to overcome such situations, they proposed the algorithm of CA competition (block 4), which selects, among CA that move over the given symbol, the automaton with the maximum number of reachable states. It is believed that such CA describes the given character in the most correct way. The rest of CA that move over this symbol are deleted (block 5), and the one remaining is accepted as the recognized current character.

Next, it remains to determine the unique parameter (for example, color) of those CA that remained on the images of letters (block 6) and then to represent the recognized text (block 7).

The benefits of such information technology is that the proposed algorithms are capable of recognition of distorted characters, of those that partially overlap, as well as a handwritten text with “printed” letters. In addition, as subsequently demonstrated, it shows high probability of recognition in all of these cases [13].

### 5. The architecture of the developed software

The proposed information technology was implemented as a software product in two variants: desktop, and a mobile version running on the Android OS. Libraries of cellular automata have been developed for the English language. We developed it using the Java programming language.

Block diagram of the algorithm for character recognition is shown in Fig. 4.

The input of the block of recognition receives a character brought to the states of transition, which we term a cellular-automata field. The next step is a cyclical motion over the set of states of the character (set listStan) and checking if the current status of the character matches its status in the transition graph (container baseGraph) of cellular automaton. If the state of the transition graph is satisfied, the cellular automaton passes over to the next state and checking is executed again. If the state of the transition graph is not satisfied, then the cellular automaton of this type will be removed from the cellular-automata field. The next step is to check dimensionality of the set of cellular automata (set listKA) against the dimensionality of the states area of the character domain. If the number of CA is less, then a new CA of the given type is born; otherwise, we initiate a block of competition of cellular automata. Next, there is a check of dimensionality of CA that remained in the domain of the given character (set listTKA): if the dimensionality of CA, selected from the set, happens to be larger than the dimensionality of the current CA, then the latter is removed from the character domain; otherwise, selected CA is removed from the set listTKA. In the case of equality in dimensionality, the user will be given the option to select from the set of possible variants.

A simplified class diagram of the developed software is shown in Fig. 5. It gives the interaction between classes only that describe cellular automata and their interaction, that is, what is really important for the process of recognition.

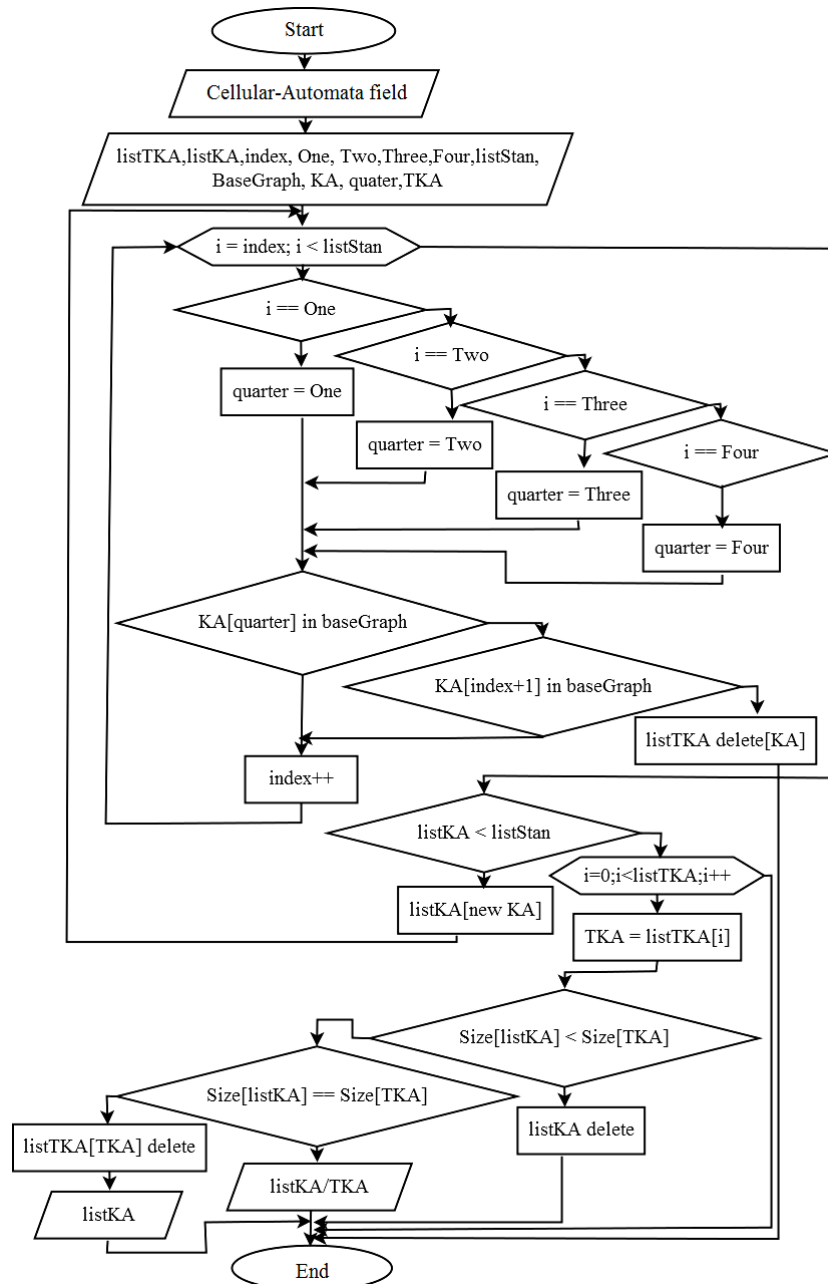


Fig. 4. Block diagram of the recognition algorithm

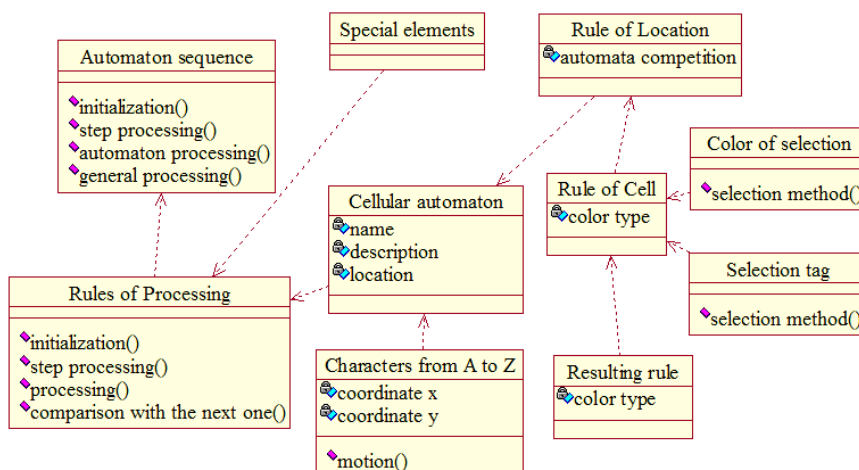


Fig. 5. Part of a class diagram of the developed software that implements operation with cellular automata

Table 1 gives a description of classes of the block of cellular automata.

Table 1

Description of classes of the block of cellular automata

Class	Description
Cellular automaton	Basic class that describes operation of cellular automaton
Rule of location	Class that describes the rule of cellular automaton in the transition graph
Rule of cell	Class that describes the condition for imposing on a cell within the rule of cellular automaton
Color of selection	Abstract class that checks color of a cell
Selection tag	Abstract class that checks tags of a cell in the transition graph
Resulting rule	Class that describes events on determining the color and tags of a cell within the rule of cellular automaton
Automaton sequence	Basic class that describes sequence operation of cellular automaton and verification of its competition
Rules of processing	Abstract class that describes operation of a sequence element
Special elements	Abstract class that describes operation of special element of the sequence
Characters from A to Z	Class that describes characters of the Latin alphabet using CA

The system consists of two basic classes: CellularAutomata and AutomataSequence, which describe operation of CA and their sequences to launch the mechanism of competition. The cellular-automata interaction includes such elements as splitting the images into separate characters, checking conditions in the transition graph, etc. States of each automaton, which matches a particular letter of alphabet and the rules of transitions in a graph, are implemented in the classes RuleItem, RuleCell and SymbolA\_Z. To simplify the diagram, descriptions of all the letters are shown in the same class, although they are actually implemented separately. These classes are responsible for the implementation of the motion of CA and describe the transition graph. The index of the type of automaton, which allows its unambiguous identification, is a color tag. Using the read-out of this tag, we can determine the recognized letter. This process is managed by the classes RuleCellResult, LabelsChecker, ColorChecker. The result of their operation is the text that is recognized using competing cellular automata.

The interaction with the hardware and representation of the recognized text is performed using the standard Windows and Android API functions. Processing of the images, received from scanning equipment, was carried out using libraries with open code OpenCV (for the OS Android).

**6. Description of the interface of the developed software**

The created software has a very simple interface as it was designed to verify the developed algorithms and recognition methods rather than to be distributed commercially. The main window of the desktop variant of the software under the mode of recognition of distorted handwritten characters is shown in Fig. 6.

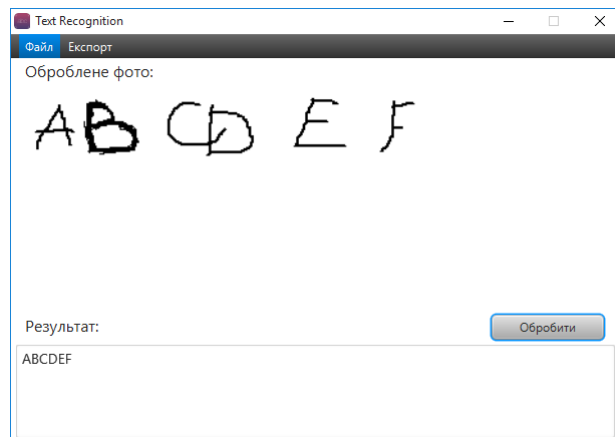
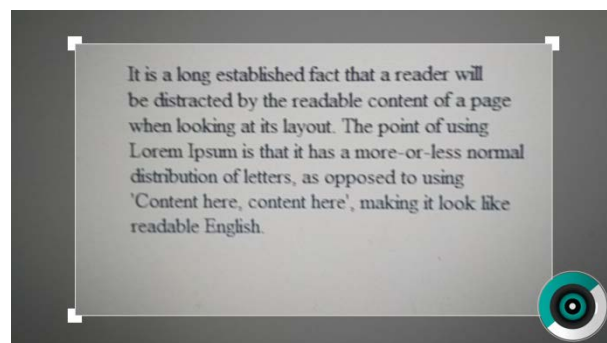
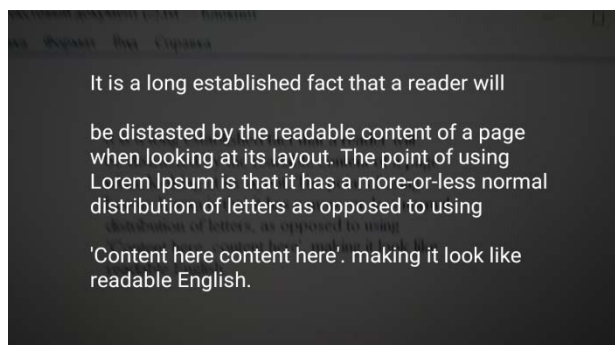


Fig. 6. Interface of the software under the mode of recognition of a handwritten text

Fig. 6, which displays the interface of the software, shows that the system successfully recognizes the characters that partially overlap. When a mobile version launches, the user is offered to select a text area for recognition using the smartphone's camera (Fig. 7, a) and to click the button to take a picture. Directly after this, the text recognition is enabled with the results displayed in the form (Fig. 7, b).



a



b

Fig. 7. Interface of the mobile version of the developed software: a – mode of capturing the text for recognition; b – result of the text recognition

Using the given software, we examined the quality of text recognition.

**7. Discussion of results of recognition**

Using the developed software, we conducted testing the quality of text character recognition, typed in different



fonts, handwritten symbols and characters that partially overlap. We also compared recognition quality of the developed software product with commercially available software FineReader.

To generate the input images, we used text generators. Research results are given in Tables 2, 3. Table 2 gives the results of recognition of different fragments of text by the developed software, Table 3 – results of comparison to the software FineReader.

While recognizing 1000 characters, Times New Roman and Arial fonts, the software showed 98 % of the detected characters. Characters that were not recognized are the characters r and n, which follow each other in the text. The result that we obtained was the symbol m. This error is typical for FineReader, too.

While recognizing 100 handwritten characters, the software showed 84 %, since the handwritten characters e and l are recognized depending on their proper writing.

If the characters have common lines, they are also poorly recognized, but in this case humans poorly recognize them as well. If the characters partially overlap, the result of recognition is 56 %.

The software in 68 % of the cases successfully recognized partially deformed characters. It should be noted that the success of recognition depends on the degree of deformation, and if this degree exceeds 30 %, then there is a high probability of false recognition of the character.

Table 2

Quality characteristics of text character recognition

Font	Image resolution	Number of characters	Recognition result
Times New Roman	760×276	1000	98 %
Arial	760×276	1000	98 %
Handwritten text	760×276	100	84 %
Overlap of handwritten characters	456×58	100	56 %
Deformed characters	456×58	100	68 %

Comparison of the developed software with the commercially available system FineReader (Table 3) reveals that the printed text in these systems is recognized approximately equally, however the performance efficiency of FineReader is higher.

While recognizing the deformed characters and the characters that overlap, the developed software demonstrates best results: 3 % of errors against 5 % and 40 % of errors against 50 %, respectively.

Based on the conducted experiments, we can draw several conclusions:

- method of competing cellular automata has demonstrated high quality of character recognition, especially those distorted and overlapped;
- speed characteristics of desktop variants of the commercially available software remain somewhat better.

High accuracy and speed of text recognition of existing fonts by the commercially available software FineReader are determined by the use of a complex of tools of recognition, including active use of dictionaries. One can also add that FineReader is a fully-fledged product that is being developed over many years by the company ABBYY. However, the software based on the algorithm, built by using competing cellular automata with tags of states, has shown some advan-

tages in the recognition of characters with a certain degree of deformation and of characters that partially overlap. Optimization of the algorithm will greatly increase performance efficiency and improve to some extent the percentage of recognition quality.

Table 3

Comparative characteristics of text recognition using the software FineReader

Test values	Developed software		FineReader	
	Operation time, s	Error percentage	Operation time, s	Error percentage
Representation of words "Hello World!" Resolution 236×30 Time New Roman Font	1	0 %	1	0 %
Representation of text characters (78 characters), resolution 561×56, Arial Font	3	1 %	3	1 %
Representation of a text page (784 characters)	9	3 %	7	3 %
Representation of handwritten characters (10 characters). Resolution 230×50	1	2 %	1	2 %
Representation of characters with a certain degree of deformation (10 characters). Resolution 230×50.	2	3 %	1	5 %
Representation of characters that overlap (10 characters). Resolution 230×50	2	40 %	1	50 %

8. Conclusions

Results of the studies conducted have shown the effectiveness of the proposed method of text character recognition based on competing cellular automata. The developed information technology allowed us to achieve a high quality of recognition and provided us with tools for further research.

The studies we conducted reveal that the proposed method of text character recognition based on competing cellular automata and the developed information technology have allowed us to achieve a high quality of recognition and provided us with tools for further research, in particular:

1. The algorithm of text character recognition, proposed in the present article, based on competing CA, demonstrated a high probability of recognition (up to 98 % for printed text and up to 84 % for handwritten text).
2. Results of the recognition of handwritten deformed characters or those that partially overlap using the given method reaches 68 % and 56 %, respectively.
3. Performance efficiency characteristics of the desktop variant of the developed software are at the level of commercially available products or somewhat worse. Optimization and subsequent parallelization of the recognition process will make it possible to improve the indicators of performance efficiency.
4. The information technology developed, and the software that was implemented based on it, demonstrated high indicators of character recognition even compared with the leader of commercial systems ABBYY FineReader.

## References

1. Zhang, P. A novel cascade ensemble classifier system with a high recognition performance on handwritten digits [Text] / P. Zhang, T. D. Bui, C. Y. Suen // Pattern Recognition. – 2007. – Vol. 40, Issue 12. – P. 3415–3429. doi: 10.1016/j.patcog.2007.03.022
2. Lauer, F. A trainable feature extractor for handwritten digit recognition [Text] / F. Lauer, C. Y. Suen, G. Bloch // Pattern Recognition. – 2007. – Vol. 40, Issue 6. – P. 1816–1824. doi: 10.1016/j.patcog.2006.10.011
3. Savas, B. Handwritten digit classification using higher order singular value decomposition [Text] / B. Savas, L. Elden // Pattern Recognition. – 2007. – Vol. 40, Issue 3. – P. 993–1003. doi: 10.1016/j.patcog.2006.08.004
4. Zheltov, S. Yu. Obrobka ta analiz zobrazen' v zadachakh mashynnoho zoru [Text] / S. Yu. Zheltov. – Moscow: Fizmatknyha, 2010. – 672 p.
5. Potapov, A. A. Noveyshie metody obrabotki izobrazheniy [Text] / A. A. Potapov, A. A. Pahomov, S. A. Nikitin, Yu. V. Gulyaev. – Moscow: Fizmatlit, 2008. – 496 p.
6. Wolfram, S. A New Kind of Science [Text] / S. Wolfram. – Wolfram Media. Inc., 2002. – 1197 p.
7. Oliveira, C. C. An Approach to Searching for Two-Dimensional Cellular Automata for Recognition of Handwritten Digits [Text] / C. C. Oliveira, P. P. B. de Oliveira // Lecture Notes in Computer Science. – 2008. – P. 462–471. doi: 10.1007/978-3-540-88636-5\_44
8. Suyasov, D. I. Vydelenie strukturnykh priznakov izobrazheniy simvolov na osnove kletochnykh avtomatov s metkami [Text] / D. I. Suyasov // Programmnye i apparatnye. – 2010. – Issue 4. – P. 39–45.
9. Wu, H. A new JPEG image watermarking algorithm based on cellular automata [Text] / H. Wu, J. Zhou, X. Gong, Y. Wen, B. Li // Journal of Information & Computational Science. – 2011. – Vol. 8, Issue 12. – P. 2431–2439.
10. Belan, S. N. Specialized cellular structures for image contour analysis [Text] / S. N. Belan // Cybernetics and Systems Analysis. – 2011. – Vol. 47, Issue 5. – P. 695–704. doi: 10.1007/s10559-011-9349-8
11. Myroniv, I. V. Rozrobka ta doslidzhennya alhorytmu rozpoznavannya symvoliv tekstu na osnovi konkuruyuchykh klitynykh avtomativ [Text] / V. V. Zhykharevych, S. E. Ostapov, I. V. Myroniv // Naukovy visnyk Chernivets'koho natsional'noho universytetu imeni Yuriya Fed'kovycha. Seriya: Komp'yuterni systemy ta komponenty. – 2010. – Vol. 1, Issue 2. – P. 47–52.
12. Zhykharevych, V. V. Analiz metodiv rozpoznavannya symvoliv tekstu [Text] / V. V. Zhykharevych, S. E. Ostapov, I. V. Myroniv // Radioelektronni i komp'yuterni systemy. – 2016. – Issue 5. – P. 137–142.
13. Zhykharevych, V. V. Alhorytm rozpoznavannya symvoliv tekstu na osnovi konkuruyuchykh klitynykh avtomativ [Text] / V. V. Zhykharevych, I. V. Myroniv, S. E. Ostapov // Radioelektronika, Informatyka, Upravlinnya. – 2015. – Issue 4 (35). – P. 31–45.
14. Myroniv, I. Development of the character recognition software on the base cellular automata [Text]: VI-th Intern. Conf. / I. Myroniv, V. Zhikharevich, S. Ostapov // Engineer of XXI Century. – 2016. – P. 229–240.