OCR/ICR Text Recognition Using ABBYY FineReader as an Example Text

A. R. Bagirzade, A. Sh. Najafova, S. M. Yessirkepova, E. S. Albert

***Abstract*—**This article describes a text recognition method based on Optical Character Recognition (OCR). The features of the OCR method were examined using the ABBYY FineReader program. It describes automatic text recognition in images. OCR is necessary because optical input devices can only transmit raster graphics as a result. Text recognition describes the task of recognizing letters shown as such, to identify and assign them an assigned numerical value in accordance with the usual text encoding (ASCII, Unicode). The peculiarity of this study conducted by the authors using the example of the ABBYY FineReader, was confirmed and shown in practice, the improvement of digital text recognition platforms developed by Electronic Publication.

***Keywords*—**ABBYY FineReader system, algorithm symbol recognition, OCR/ICR techniques, recognition technologies.

I. Introduction

FROM a technical point of view, OCR refers only to the subarea of comparing patterns of individual portions of an image as candidates for recognition of individual characters. This OCR process is preceded by global structure recognition, in which text blocks are first separated from graphic elements, linear structures are recognized, and finally individual characters are separated. When deciding which symbol is present, it is possible to take into account the linguistic context using additional algorithms.

Originally, specially designed fonts were developed for automatic text recognition, which were used, for example, to print blank checks. These fonts have been designed so that individual characters can be quickly and easily distinguished using text recognition. The OCR-A font (DIN 66008, ISO 1073-1) is distinguished by different characters, especially in the part of numbers. OCR-B (ISO 1073-2) is more like a non-proportional Sans serif font, while OCR-H (DIN 66225) is based on handwritten numbers and capital letters.

II. Historical Aspects

1968 was a revolutionary year not only politically, but also in the history of computers. Engelbart invented the computer mouse, the first predecessor of our personal computers appeared on the market, and gradually electronic data processing became in demand [1].

Initially, the use of OCR was limited. Because of the processing power, which seems ridiculous from today's point of view, to achieve useful results, it was necessary to use standardized, easy-to-read fonts with clearly distinguishable characters. One can find the most famous example of this font in Fig. 1for this article. The columns of numbers in the check number field are in the very first machine font called OCR-A. An important part of OCR fonts are additional control characters commonly referred to as "hook", "fork" and "chair". They provide important assistance to the scanner, for example, in recognizing the end of an information block [2].

After 5 years and a technical leap, OCR not only can read most of all fonts, but thanks to built-in speech recognition - usually with dictionaries - it checks whether the text read has a meaningful relationship or not. In case of doubt, the program does not solve "2weifels§all", but corrects itself. However, a prerequisite for a good result is still a good resolution of the digital template created by the scanner. Therefore, scanners must read documents with a resolution of at least 200 dpi, since lower resolutions have too few pixels in the original and lead to too many errors in the OCR process.

Even today, OCR still struggles with heavily darkened scans or faded receipts on thermal paper. The human eye is able to interpret illegible parts and establish a general context. The software (still) reaches its limits here.

Until a few years ago, recognized texts were displayed only without any information about the layout. The texts were recognized correctly, but displayed without information about their location (layout) on the page. This is no longer a problem nowadays, thanks to the extended hOCR standard, page structure and layout information can be stored based on XML tags - even without hooks and chairs. WebPDF has supported this output format since version 5.0. You can select it manually or set it as the default by changing the preset parameter [3].

Machine fonts such as B. OCR-A or OCR-B are still not obsolete. They are increasingly seen as a style element in modern design. These fonts appear over and over again, especially during the "retro design" era. OCR is also about text recognition with minimal errors.

III. How the Program Works in Practice

OCR is an area in which there are still relatively few "proprietary" Linux developments. The term "OCR" is used

Bagirzade A.R. master student of financial faculty of Plekhanov Russian University of Economics, 117997, Moscow, Russian Federation, (phone: 89256523996, e-mail: fedorbagirov47@gmail.com).

Najafova A.Sh. bachelor's degree student of the "International School of Business and World Economy" faculty of Plekhanov Russian University of Economics, 117997, Moscow, Russian Federation, (phone: 89776414103, e-mail: aisha.najafova@bk.ru).

Yessirkepova S.M. master student at programme "information systems" of Toraighyrov University, 140000, Pavlodar, Republic of Kazakhstan, (phone: 87076074477, e-mail: Saltuwa1993@mail.ru).

Albert E.S. master student at programme "Applied Informatics" faculty of the Plekhanov Russian University of Economics, 117997, Moscow, Russian Federation, (phone: 87057086305, e-mail: es_kuznetsova10@mail.ru).

synonymously with "Scanner", which strictly speaking includes only one part, namely character recognition. Recognition will be better if words are matched in addition to characters, which, however, requires a lot of programming effort (e.g. tesseract-ocr). In addition to the "classic" open source projects, most of which were created using source code by the same main contributor (gocr, Ocrad, etc.), there are further developments from those previously commercialized. And, later projects such as Cuneiform-Linux and tesseract-ocr or OCRopus, which have been migrated to an open source license that have been transferred to others based on tesseract-ocr. There is also Archivista, which is commercially developed, but also offers a complete paperless office solution under a free license; the company also offers eruption elimination equipment as a complete solution.

While pretty good results can be achieved for pure character recognition, layout analysis and font/size recognition in Linux is not very advanced yet. The overall development effort appears to be quite high; in the case of purely commercial applications, this is reflected in the prices of the programs. Projects such as tesseract-ocr or OCRopus are funded and partially funded by Google out of their own interests (for example, the creation of "e-books").

The increased performance of modern computers and improved algorithms also allow the recognition of "normal" printer fonts up to handwriting (for example, when distributing letters); however, if human readability is not a priority, barcodes are used that are easier to work with in terms of printing and identification [4].

Modern text recognition now covers not only OCR, in addition, Intelligent Character Recognition (ICR) context analysis techniques are used to correct the actual OCR results. A character that was actually recognized as "8" can be corrected to "B" if it is inside a word. Instead of "8aum", "tree" is recognized, but the transformation is "8th", i.e. an alphanumeric combination should not be produced. In the field of industrial text recognition systems, OCR/ICR systems are used. However, the boundaries of the term OCR are fuzzy as OCR and ICR also serve as marketing terms to better promote technical developments in the market. Intelligent Word Recognition (IWR) also falls into this category. This approach attempts to solve the problem of fluid handwriting recognition, in which individual characters cannot be clearly separated and therefore cannot be recognized using conventional OCR techniques [4].

A fundamentally different approach to text recognition is used to recognize handwriting input on touch screens or input fields (PDA, etc.). Vector templates are processed here, either "offline" in general, or "online" with additional analysis of the input stream (for example, an Apple inkwell) [5].

Inkwell (also abbreviated as Ink) is a handwriting recognition program that is fully integrated into Apple's Mac OS X text system. It is introduced in Mac OS X 10.2. Inkwell supports English, French and German. Inkwell was developed by Brandin Webb, but originally on Rosetta, a technology that was present in the Apple Newton PDA [5].

Inkwell only works with a graphics tablet as an input device. If this graphics tablet is connected to a computer, all settings for working with Inkwell can be specified in the control panel. When the ink tank is activated, a translucent, dynamically growing yellow note window appears in which the user can write. When the input is complete, the recognized text is automatically inserted into the currently active program. If, for example, handwriting recognition is not activated, a hand-drawn sketch can be inserted into the active program. In early 2008, computer maker Axiotron introduced the ModBook tablet laptop. ModBook is the first touchscreen laptop to use Inkwell [6].

A special form of text recognition leads, for example, to the automatic processing of incoming mail from large companies. One of the tasks is to sort the documents. To do this, it is not always necessary to analyze the content, but sometimes it is enough to recognize gross features such as the characteristic arrangement of forms, company logos, etc. As with OCR, the classification of certain types of text is done using pattern recognition, which, however, refers to the whole sheet or to specific places, and not to individual letters [7].

One of the most common image recognition tasks is handwriting recognition. Despite the fact that a large number of different programs for handwriting recognition have been created, the relevance of the development of new software tools does not decrease. This is mainly due to the fact that free software products are primarily focused on recognizing printed text, not handwritten text. The authors conducted research on some software tools. For example, the free Open OCR Cuneiform package (File version 12.0.0.58851) showed the following results: from handwritten Russian text (the text is written in block letters), consisting of 192 characters, 123 characters (64%) were correctly recognized, from printed Russian text, consisting of 191 characters, 186 characters (97%) were correctly recognized. Specially developed complexes for streaming data and document input, for example ABBYY FlexiCapture, demonstrate the accuracy of handwriting recognition (98%), but at the same time have a high cost (more than 200 thousand rubles) [8]. Based on this, it was decided to develop software using our own character recognition algorithms based on existing ones.

FineReader is a proprietary desktop OCR program from the Russian company ABBYY for Windows and Mac OS X. A command line version is available for Linux. With PDF Transformer integration from the same company, FineReader Standard and Corporate for Windows have also been offering PDF tools since 2017 [9].

The software offers a graphical user interface from which images can also be read directly by scanners and processed in batches with various automation functions. The user can take corrective actions at various stages of the process and, for example, correct recognized blocks in the page layout and check for characters that were found to be unsafe. Results can be output in various file formats (including PDF and Microsoft Office formats) or directly to a text editor. FineReader offers language models for a wide variety of natural and artificial languages.

A special version has been added to recognize the German

Fraktur font: FineReader XIX. Its development and sale is currently discontinued, but it can be read in text fragments through the recognition server and online conversion of FineReader Online. FineReader versions are often included with scanners. The first version was released in July 1993 [7].

Imagine you have a paper document - like a newspaper article, brochure, or contract - that your partner has sent you as a PDF attachment. A scanner is not enough to extract relevant information from these documents and, for example, reproduce it in an editable Microsoft Word format. All the scanner can do is take a snapshot of the document. And it is just a collection of black, white, or color pixels that are arranged in a table and are known in technical terms as bitmap graphics. To read and use information from scanned documents, digital images, or image-only PDFs, you need OCR software that recognizes letters in images, combines them into words, and uses them to construct entire sentences. The software allows you to access the actual content of the documents, which you can then edit [8].

First, let us take a look at how FineReader PDF OCR recognizes text. First, the program analyzes the structure of the document image. It divides the page into elements such as blocks of text, tables, images, etc. Then it divides lines into words and finally words into letters. Once individual letters have been identified, the program compares them to a series of sample images and generates numerous hypotheses about which letters are being used. Based on these hypotheses, the program explores different ways of splitting strings into words and words into letters. After processing a very large number of such probabilistic hypotheses, the program finally makes a decision and presents the recognized text [9]. In addition, ABBYY FineReader PDF offers dictionary support for some languages. This allows for secondary analysis of textual elements at the word level. Thanks to dictionary support, the program guarantees even more accurate analysis and recognition of documents, and also simplifies the verification of recognition results.

The most advanced text recognition systems, such as ABBYY FineReader OCR, aim to simulate object recognition as found in nature or in animals. Essentially, these systems are based on three pillars of unity, usability, and adaptability (Integrity, Purpose, and Adaptability, in short (IPA)) [10]. Based on these principles, the program uses an extremely flexible and intelligent recognition method that is very close to the human way of recognizing objects [11].

After years of research, ABBYY has successfully integrated the IPA principles described above into its OCR technology. ABBYY FineReader OCR is easy to use - the process usually goes through three stages: opening (scanning) the document, recognizing and saving it in the desired format (DOC, RTF, XLS, PDF, HTML, TXT etc.) or exporting the data directly in an office application, e.g. Microsoft Word, Excel or Adobe Acrobat [12].

The corporate version of ABBYY FineReader PDF also supports automated document processing with the Hot Folder Tool, which is particularly indispensable when processing regularly recurring tasks. With this feature, text recognition can run automatically without manual activation of the individual steps [13].

Thanks to FineReader OCR, the recognized documents have the same layout as the originals. Sophisticated and powerful OCR software saves a lot of time and effort when creating, processing and reusing many different documents. With ABBYY FineReader OCR, you can scan paper documents for editing and sending to colleagues and partners. You can take quotes from books and magazines and include them in your research and working papers without typing them. With a digital camera and FineReader OCR, you can capture text from banners, posters, and timetables on the go and use the information thus obtained. In the same way, you can get information from paper documents and books - for example, when the scanner is not at hand or it cannot be used. Alternatively, you can use OCR software to create searchable PDF archives [14].

The entire process of converting data from a paper document, image or PDF file takes less than a minute, and the final OCR document looks exactly like the original.

To test the algorithm described above, we developed specialized software for recognizing test forms containing handwritten Arabic numeral characters. During testing, more than 2,500 Arabic numeral characters were recognized. Table I presents experimental data on the probability of correct recognition of Arabic numerals from "0" to "9", depending on the number of loaded characters in the template, obtained using this algorithm [15]. Fig. 1 shows how the program works in practice.

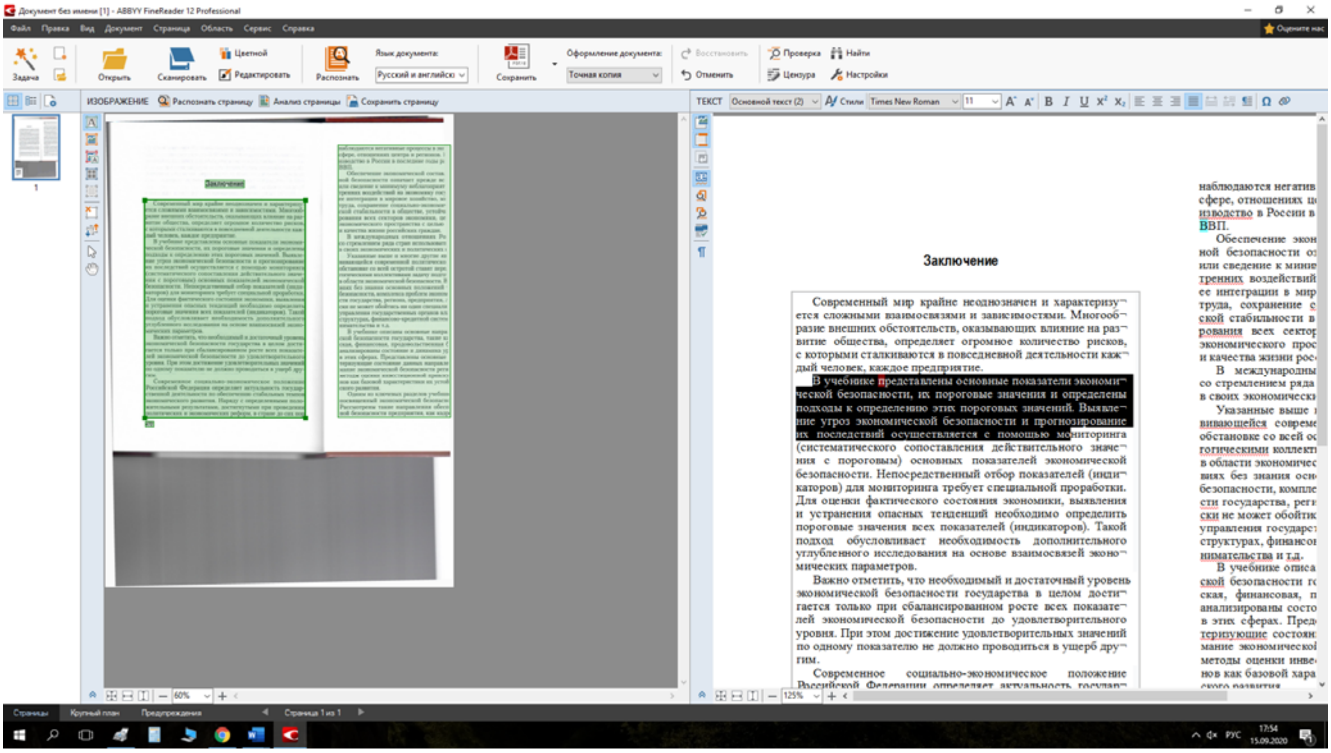
TABLE I
PROBABILITY OF CORRECT RECOGNITION OF DIFFERENT SYMBOLS DEPENDING ON THE NUMBER OF SYMBOLS UPLOADED IN THE TEMPLATE

Number of symbols loaded into the template	10	30	50	100	150	200
“0”	0.9	1	1	1	1	1
“1”	0.7	0.7	0.6	0.6	0.8	0.9
“2”	0.7	0.7	0.7	0.8	0.9	1
“3”	0.6	0.7	0.7	0.9	1	1
“4”	0.8	0.9	0.7	0.9	1	0.8
“5”	0.8	0.6	0.8	1	1	1
“6”	0.6	0.6	0.8	1	1	0.9
“7”	0.5	0.5	0.7	0.7	0.7	0.9
“8”	0.4	0.5	0.4	0.5	0.7	0.7
“9”	0.5	0.6	0.8	0.9	0.8	0.8

Based on the obtained data presented in Table I, an increase in the number of symbols loaded into a template does not always have a positive effect. For example, when recognizing characters "4" and "6", an increase in the number of loaded characters into the template leads to worse results. Therefore, when developing a template, you must take this point into account [16].

TABLE II
PROBABILITY OF CORRECT RECOGNITION OF AN ARBITRARY DIGIT SYMBOL DEPENDING ON THE NUMBER OF SYMBOLS UPLOADED IN THE TEMPLATE

Number of symbols loaded into the template	10	30	50	100	150	200
Recognition probability	0.65	0.68	0.72	0.83	0.89	0.91



Заключение

Современный мир крайне неоднозначен и характеризуется сложными взаимосвязями и зависимостями. Многообразие внешних обстоятельств, оказывающих влияние на развитие общества, определяет огромное количество рисков, с которыми сталкиваются в повседневной деятельности каждый человек, каждое предприятие.

В учебнике представлены основные показатели экономической безопасности, их пороговые значения и определены подходы к определению этих пороговых значений. Выявление угроз экономической безопасности и прогнозирование их последствий осуществляется с помощью мониторинга (систематического сопоставления действительного значения с пороговым) основных показателей экономической безопасности. Непосредственный отбор показателей (индикаторов) для мониторинга требует специальной проработки. Для оценки фактического состояния экономики, выявления и устранения опасных тенденций необходимо определить пороговые значения всех показателей (индикаторов). Такой подход обусловливает необходимость дополнительного углубленного исследования на основе взаимосвязей экономических параметров.

Важно отметить, что необходимый и достаточный уровень экономической безопасности государства в целом достигается только при сбалансированном росте всех показателей экономической безопасности до удовлетворительного уровня. При этом достижение удовлетворительных значений по одному показателю не должно проводиться в ущерб другим.

Современное социально-экономическое положение Российской Федерации определяет актуальность государ¬

Fig. 1 Practical use ABBYY FineReader

Table II shows experimental data on the probability of correct recognition of an arbitrary digit character, depending on the number of characters loaded into the template obtained using this algorithm [17]. As you can see from this table, the probability of correct character recognition is on average 0.91, which is comparable to the recognition results by the Abby FormReader 4.5 program. According to the research carried out, the correctness of the recognition of Arabic numerals when using the Abby FormReader 4.5 program is approximately 0.98 (when checking test forms). To improve this result, it is planned to use additional recognition algorithms, for example, neural networks, which will increase the likelihood of correct identification of symbols.

Based on this research topic, it can be said that technologies are rapidly developing and bearing fruit, and also contribute to simplifying people's lives [18].

VI. Results and Discussion

The actual detection is usually done with command line tools, which can usually only handle a very small number of image formats. There are often so-called "front-ends" with graphical user interfaces that can usually also convert data into the format required by a real OCR program. Scanning programs are also integrated so that complete processing from template creation to editing can take place right through to text recognition and output.

V. Conclusion

Using a hybrid neural network increases the efficiency of character recognition. This process is also influenced by the number of training examples (training sets/data sets).

The developed neural network model and software application allow using them in the real process of recognizing handwritten Kazakh text. It can be integrated into the electronic document management system and software solution designed for digitizing archival and other operational handwritten information in the Kazakh language.

Acknowledgment

This article was prepared as part of the government contract as requested by the Ministry of Science and Higher Education of the Russian Federation on the subject formulated as «Structural changes in economy and society as a result of achieving the target indicators of National projects, which provide opportunities to organize new areas of social and economic activity, including commercial, both in Russia and abroad» (project No. FSSW-2020-0010).

References

[1] Lowood, Henry Douglas Engelbart Interview 1, Stanford and the Silicon Valley: Oral History Interviews Stanford University (December 19, 1986) https://library.stanford.edu

[2] Technology visionary Doug Engelbart, inventor of computer mouse, dies at age of 88. The Washington Post (July 3, 2013) https://gigaom.com/2013/07/03/doug-engelbart-american-inventor-computing-legend-passes-away/

[3] Douglas C. Engelbart. Augmenting Human Intellect: A Conceptual Framework dougengelbart.org.

[4] Arlazarov V.L., Slavin O.A. "Recognition algorithms and technologies for entering texts into computers." - In-formation technologies and computing systems № 1, 1996

[5] Slavin O.A., Fedorov G.O. "Questions of recognition of text digitized with video cameras." ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic3/.

[6] N. E. Buzikashvili "Selection and presentation of pictures on non-monochrome images."ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic1/

[7] Vladimir Vezhnevets "Assessment of the quality of classifiers work" http://cgm.graphicon.ru/content/view / 106/60 /

[8] Kazem Taghva, Julie Borsack, Steven Lumos, Allen Condit "A comparison of automatic and manual zoning."

[9] International Journal on Document Analysis and Recognition, Volume 6, Number 4 / April, 2003

[10] Lobachev A. A., Kulikova O. V. Choice of language for teaching programming / / Information technologies in education. XVIII international Conf. - exhibition: sat. Tr. participants Conf. CH. VI. - M.: MEPhI, 2008. - P. 45-47.

[11] Intelligent Document Recognition. (Electronic resource) URL: http://idr.in.ua/article/12_2013/12.html

[12] Kazem Taghva, Julie Borsack, Steven Lumos, Allen Condit "A comparison of automatic and manual zoning."

[13] International Journal on Document Analysis and Recognition, Volume 6, Number 4 / April, 2003

[14] He K., Zhang X., Ren S., et al. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (Las Vegas, NV, USA, 27–30 June 2016), 2016. pp. 770–778 DOI: 10.1109/CVPR.2016.90.

[15] Szegedy C., Liu W, Jia Y. et al. Going Deeper with Convolutions. IEEE Conference on Computer Vision and Pattern Recognition (Boston, MA, USA, June 7–12, 2015), 2015. pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[16] Psichogios, D.C. and Ungar, L.H. (1992), A hybrid neural network-first principles approach to process modeling. AIChE J., 38: pp. 1499-1511. doi:10.1002/aic.690381003

[17] ABBYY FineReader: an outward glance / / 3DNews. (Electronic resource) URL: https://3dnews.ru/632560

[18] Belchusov A. A., Stepanov A.V. Improving the effectiveness of programming training in schools and universities // Materials V All-Russian. scientific-practical Conf. "Problems of Informatization of education: regional aspect", Cheboksary, April 25-27, 2007-Cheboksary, 2007. - Pp. 27-33.