UDC 004.051

RESEARCH OF AUTOMATIC INFORMATION RECOGNITION SYSTEMS CONSTRUCTION APPROACHES

B. Havrysh[1], Z. Selmenska[2], B. Kovalskyi[2], I. Izonin[1]

[1] Lviv Polytechnic National University,
12, S. Bandera St., Lviv, 79013, Ukraine

[2] Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine

One of the earliest attempts to create a system capable of reading texts was created in 1870. It was a retina scanner, the work of which was based on photocells. Later Fourier d'Albe's Optophone appeared in 1912 and Thomas tactile relief device in 1926. Optical text reading systems appeared in the middle of the twentieth century as a result of the digital computers development. David Shepard, the founder of Intelligent Machine Research, is considered the father of commercial OCR systems.

In the early 60's the automatic reading systems have become widespread, despite the ability to read a very limited number of fonts and the limitations imposed on the orientation of the characters. With the development of microelectronics, these systems have been constantly improved.

Keywords: *optical recognition, system, reading, symbols, technology.*

Relevance of research. The task of recognizing textual information during the transformation of printed and handwritten text into machine codes is one of the most important components of projects aimed at document automation. However, this task is one of the most complex and knowledge-intensive in the field of automatic image analysis. Even a person reading a handwriting, in isolation from the context, makes an average of 4% of errors. As for the systems of reading printed documents, the difficulty here is that in the relevant applications, such as, for example, automation of passport and visa information entry, it is necessary to ensure high reliability of recognition (more than 98-99%) even with poor print quality and digitization source text [1, 2].

Thanks to the use of modern advances in computer technology, in recent decades new methods of image processing and pattern recognition ([*l*]-[*ll*]) have been developed, making it possible to create such systems of printed text recognition that would meet the basic requirements of automation systems document management. However, the creation of each new application in this area still requires an additional research in connection with the specific requirements for resolution, speed, reliability of recognition and memory, which characterizes each specific task of developing a problem-oriented system of automatic input to computer paper documentation.

Presenting the main material. The various technologies, united by the general term *"character recognition"*, are divided into the real-time recognition and batch recognition, each of which is characterized by its own hardware and its own recognition algorithms [2-4].

In a typical *optical text recognition system*, entered characters are read, and digitized by an optical scanner. After that, each symbol can be localized and selected, and the resulting matrix is a matter to process, i.e. smoothing, filtering and normalization. As a result of preliminary processing, the corresponding signs are allocated and then, the classification is carried out.

There are numbers of significant problems associated with handwriting and typed character recognition. The most important are the following:

– variety of drawing symbols forms;
– image distortion;
– variations in character size and scale.

Each individual character can be written in different standard fonts, for example (Gothic, Elite, Courier, Orator), special fonts used in OCR systems, as well as many non-standard fonts. In addition, different characters may have similar shapes [5]. For example, "U", "V", "S" and "5", "Z" and "2", "G" and "6".

Distortion of digital images of symbols can be of the following types (Fig. 4):

– Form distortion:
  • line breaks;
  • unprinted characters;
  • isolation of individual points;
  • non-planar nature of the information carrier (for example, the effect of distortion);
  • offset of characters or their parts relative to the location in the line;
  • rotation with a change in the slope of the characters;
  • distortion of the symbol shape due to the digitization of the image with a "rough" discrete.
– Radiometric distortions:
  • lighting defects;
  • shadows;
  • glare;
  • uneven background;
  • errors while scanning or shooting with your video camera.

The influence of the original scale of printing is also significant. In accepted terminology, a scale of 10, 12 or 17 means that 10, 12 or 17 characters are placed in an inch of string. In this case, for example, the symbols of scale 10 are usually larger and wider than the symbols of scale 12.

In addition to these problems, the optical text recognition system must select text areas in the image, highlight individual characters in them, recognize these characters and be insensitive to the method of printing (layout) and line spacing [6, 7].

Structure of optical text recognition systems. Typically, OCR systems consist of several blocks that involve hardware or software implementation:

– optical scanner;

– block of localization and selection of text elements;
– image pre-processing unit;
– feature selection unit;
– recognition unit;
– post-processing unit of recognition results.

As a result of the optical scanner work, the source text is entered into the computer in the form of a halftone or binary image.

In order to save memory and reduce the time spent on information processing, OCR systems typically use the conversion of a halftone image to black and white. This ope ration is called binarization. However, it should be taken into account that the binarization operation may lead to a deterioration in the recognition efficiency.

Software in OCR systems is responsible for presenting data digitally and splitting the coherent text into individual characters.

After partitioning, the symbols presented in the form of binary matrices are subject to smoothing, filtering to eliminate noise, normalization of size, and other transformations in order to highlight the features that are subsequently used for recognition.

Character recognition occurs in the process of comparing the selected characteristics with the reference characteristics, which are selected during the statistical analysis of the results obtained in the process of learning the system.

Thus, semantic or contextual information can be used both to resolve uncertainties that arise during the recognition of characters having identical dimensions, and to correct words and phrases in general [8].

Reprocessing of text symbols images and post-processing of recognition results. Reprocessing is an important step in the process of pattern recognition and allows for smoothing, normalization, segmentation and approximation of line segments (Fig. 1).

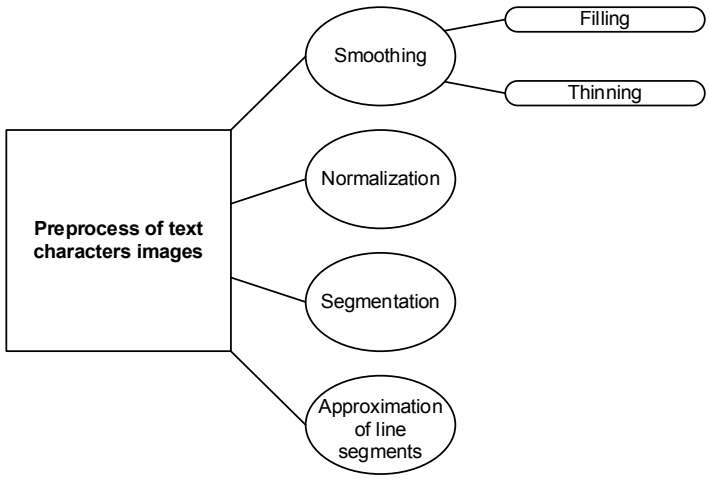

Fig. 1. Image pre-processing

Smoothing consists of operations of *filling and thinning*.
Filling eliminates small severance and gaps.

Thinning is the process of reducing the thickness of a line in which several pixels are matched to only one pixel. Sequential, parallel and hybrid thinning algorithms are known. The most common methods of thinning are based on iterative blurring of the contours, in which the window (3 × 3) moves along the image, and inside the window, the corresponding operations are performed. At the end of each step, all selected points are deleted.

Normalization consists of algorithms that eliminate the skew of individual characters and words, and also includes procedures that normalize the characters in height and width after their appropriate processing.

Segmentation divides the image into separate areas. As a rule, first of all it is necessary to clear the text of graphics and handwritten marks, as the listed methods allow to process only the noisy text. The text cleared of various marks can already be segmented.

Most optical recognition algorithms divide text into characters and recognize them separately.

This simple solution is really effective, as long as the characters of the text do not overlap. Character merging can be caused by the typeface in which the text was typed, by the poor resolution of the printer, or by the high brightness level selected to recover broken characters.

The division of the text into words is possible if the word is a component feature, according to which the segmentation is performed. Such an approach is difficult to implement due to the large number of elements to be recognized, but it can be useful if the set of words from the code dictionary is limited by the condition of the problem.

Approximation of line segments is a process of drawing a graph of the description of the symbol in the form of a set of vertices and straight edges, which directly approximate the pixel chains of the original image. This approximation is performed to reduce the amount of data and can be used in the recognition based on the selection of features that describe the geometry and topology of the image.

In high-precision OCR systems, such as passport and visa documents, the quality of recognition of individual characters read by machines is not considered sufficient. Contextual information must also be used in such systems. The use of contextual information allows not only to find errors, but also to correct them.

There are many OCR applications that use global and local position charts, trigrams, *n*-grams, dictionaries, and various combinations of these methods. Consider two approaches to solving this problem: a dictionary and a set of binary matrices that approximate the structure of the dictionary.

It is proved that dictionary methods are one of the most effective in determining and correcting errors in the classification of individual characters. In this case, after recognizing all the characters of a word, the dictionary is searched in search of this word, taking into account the fact that it may contain an error. If the word is in the dictionary, it does not indicate the absence of errors. An error can turn one word in the dictionary into another that is also in the dictionary. Such an error cannot be detected without the use of semantic contextual information, but it only can confirm the correct spelling. If the word is missing in the dictionary, it is considered that an error has been made in the

word. To correct the error, they resort to replacing such a word with a similar word from the dictionary. The correction is not performed if several suitable replacement options are found in the dictionary. In this case, the interface of some systems offers the user different solutions, such as correcting the error, ignoring it and continuing to work, or make the word in the dictionary.

The main disadvantage of using a dictionary is that the search and comparison operations used to correct errors require significant computational costs, which increase with the volume of the dictionary.

Some developers, in order to overcome the difficulties associated with the use of the dictionary, try to extract information about the structure of the word from the word itself. Such information indicates the degree of plausibility of *n*-grams (for example, pairs and triplets of letters) in the text. *n*-grams can also be globally positioned, locally positioned or not positioned at all.

Conclusions. This paper considers the existing approaches to the problem of text recognition. Recognition data can be obtained in two different ways – offline and online. The main difference between them is the set of recognizable features. If the first method is simple and can work with data obtained without the use of any specialized equipment, the second is more accurate, but requires a much more complex data set. The combination of different methods of recognition of signs should be considered as the main direction of development in this area.

An important part of any character recognition system is the segmentation subsystem. Distinguishing between written words in an image and distinguishing letters from words is a complex task that requires no less attention than the actual recognition process. Even more important is the system of selection of features, which should find significant properties of the selected letters and discard the unimportant.

A modern text recognition machine cannot exist without a dictionary and a context recognition subsystem. They allow the machine to use external data to resolve conflict situations, such as distinguishing between lowercase and uppercase letters or understanding a vaguely written sign.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Козел В. О. Методи та етапи автоматичного розпізнавання тексту. *Вісник Черкаського університету. Серія прикладна математика. Інформатика.* Вип. 172. С. 75–86.
2. Застосування методів захисту інформації в інформаційних системах / Гавриш Б. М., Дурняк Б. В., Полусин О., Семенова О. Є. *Моделювання та інформаційні технології*. 2019. Вип. 89. С. 224–229.
3. Гавриш Б. М., Тимченко О. В., Дурняк Б. В. Побудова ієрархічних сценаріїв опрацювання даних. *Моделювання та інформаційні технології*. 2019. Вип. 86. С. 86–90.
4. Гавриш Б. М., Тимченко О. В., Ковальський Б. М. Дослідження методу коригування роздільної здатності зображення для відтворення цифровими вивідними поліграфічними пристроями. *Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту* : матеріали Міжнародної наукової конференції (с. Залізний Порт, 21–25 травня 2019 р.). 2019. С. 174–175.

5. Recognition of handwritten digits using template and model matching / Gader P. D., Forester B., Ganzberger M., Billies A., Mitchell B., Whalen M., Youcum T. *Pattern Recognition*, 1991. 5 (24). 421–431.
6. Arica N., Yarman-Vural F. T. An Overview of Character Recognition Focused on Off-Line Handwriting. Machine Intelligence, 22 (1). 4–37. *IEEE Trans. on Systems, Man, and Cybernetics*. Part C: Applications and Reviews, Vol. 31. No. 2, May 2001.
7. Васин Д., Ершов М. Распознавание символов на базе низкоуровневых моделей описания графических зображений. *Графикон-2014* : 24-я Междунар. конф. по компьютерной графике и зрению. 2014. С. 62–64.
8. Steinbuch K., Piske U. A. W. Learning matrics and their application. *IEEE Transaction of Electronic Computers.* Dec. 1963. Vol. EC–12. № 6.

REFERENCES

1. Kozel, V. O. Metody ta etapy avtomatychnoho rozpiznavannia tekstu: Visnyk Cherkaskoho universytetu. Seriia prykladna matematyka. Informatyka, 172, 75–86 (in Ukrainian).
2. Havrysh, B. M., Durniak, B. V., Polusyn, O., & Semenova, O. Ie. (2019). Zastosuvannia metodiv zakhystu informatsii v informatsiinykh systemakh: Modeliuvannia ta informatsiini tekhnolohii, 89, 224–229 (in Ukrainian).
3. Havrysh, B. M., Tymchenko, O. V., & Durniak, B. V. (2019). Pobudova iierarkhichnykh stsenariiv opratsiuvannia danykh: Modeliuvannia ta informatsiini tekhnolohii, 86, 86–90 (in Ukrainian).
4. Havrysh, B. M., Tymchenko, O. V., & Kovalskyi, B. M. (2019). Doslidzhennia metodu koryhuvannia rozdilnoi zdatnosti zobrazhennia dlia vidtvorennia tsyfrovymy vyvidnymy polihrafichnymy prystroiamy. Intelektualni systemy pryiniattia rishen ta problemy obchysliuvalnoho intelektu : materialy Mizhnarodnoi naukovoi konferentsii (s. Zaliznyi Port, 21–25 travnia 2019 r.), 174–175 (in Ukrainian).
5. Gader, P. D., Forester, B., Ganzberger, M., Billies, A., Mitchell, B., Whalen, M., & Youcum, T. (1991). Recognition of handwritten digits using template and model matching: Pattern Recognition, 5 (24), 421–431 (in English).
6. Arica, N., & Yarman-Vural, F. T. (May 2001). An Overview of Character Recognition Focused on Off-Line Handwriting. Machine Intelligence, 22 (1), 4–37. IEEE Trans. on Systems, Man, and Cybernetics. Part C: Applications and Reviews, 31, 2 (in English).
7. Vasin, D., & Ershov, M. (2014). Raspoznavanie simvolov na baze nizkourovnevyh modelej opisanija graficheskih zobrazhenij. Grafikon-2014 : 24-ja Mezhdunar. konf. po komp'juternoj grafike i zreniju, 62–64 (in Russian).
8. Steinbuch, K., & Piske, U. A. W. (Dec. 1963). Learning matrics and their application: IEEE Transaction of Electronic Computers, EC–12, 6 (in English).

doi: 10.32403/0554-4866-2021-1-81-21-27

ДОСЛІДЖЕННЯ ПІДХОДІВ ДО ПОБУДОВИ СИСТЕМ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ СИМВОЛЬНОЇ ІНФОРМАЦІЇ

Б. М. Гавриш[1], З. М. Сельменська[2], Б. М. Ковальський[2], І. В. Ізонін[1]

[1]Національний університет «Львівська політехніка», вул. С. Бандери, 12, Львів, 79013, Україна

[2]Українська академія друкарства, вул. Під Голоском, 19, Львів, 79020, Україна
dana.havrysh@gmail.com, zorselm@gmail.com, bkovalskyy@ukr.net, ivanizonin@gmail.com

Одна з найбільш ранніх спроб створити систему, здатну зчитувати тексти, була створена в 1870 році. Вона представляла собою сканер-сітківку, робота якого була заснована на фотоелементах. Надалі з'явилися Fourier d'Albe's Optophone в 1912 р. і Thomas tactile relief device в 1926 р. Системи оптичного зчитування текстів з'явилися в середині XX ст. в результаті розвитку цифрових комп'ютерів. Девід Шепард, засновник компанії Intelligent Machine Research, вважається родоначальником створення комерційних систем OCR.

На початку 60-х р.р. системи автоматичного зчитування набули широкого поширення, незважаючи на можливість зчитування дуже обмеженої кількості шрифтів і на обмеження, що накладаються на орієнтацію символів. З розвитком мікроелектроніки ці системи постійно вдосконалювалися.

Системи оптичного розпізнавання тексту вимагають калібрування для роботи з конкретним шрифтом; в ранніх версіях для програмування було необхідне зображення кожного символу, програма одночасно могла працювати тільки з одним шрифтом. В даний час найпоширеніші так звані «інтелектуальні» системи, які з високим ступенем точності розпізнають більшість шрифтів. Деякі системи оптичного розпізнавання тексту здатні відновлювати початкове форматування тексту, враховуючи зображення, колонки та інші нетекстові компоненти.

Оптичне розпізнавання тексту є досліджуваною проблемою в областях розпізнавання образів, штучного інтелекту та комп'ютерного зору. Оптичне розпізнавання символів дозволяє редагувати текст, здійснювати пошук слів чи фраз, зберігати його в більш компактній формі, демонструвати або роздруковувати матеріал, не втрачаючи якості, аналізувати інформацію, а також застосовувати до тексту електронне переведення, форматування або перетворення в усне мовлення.

Ключові слова: *оптичне розпізнавання, система, зчитування, символи, технологія.*

Стаття надійшла до редакції 28.01.2021.
Received 28.01.2021.