RECOGNITION OF TEXT INFORMATION IN THE BRONCHOPULMONARY DISEASES DIAGNOSIS SYSTEM

G.R. Shakhmametova[1], A.A. Evgrafov[1], R.Kh. Zulkarneev[2]
Computer Science and Robotics Department, Ufa State Aviation Technical University Ufa, Russia[1]
Faculty of General Medicine, Bashkir State Medical University Ufa, Russia[2]
e-mail: shakhgouzel@mail.ru, evgrafov.alexander92@yandex.ru, zrustem@ufanet.ru

***Abstract**: In article the problem of recognition of text information is considered. The obtained information is processed in the form of the scanned pages with a large number of medical terms for the purpose of further processing in the system of diagnosis of bronchopulmonary diseases.*

1. Introduction

Active penetration of information technologies into medical branch stimulated creation of program complexes which main goal is essential improvement of quality of rendering medical services due to introduction of modern technologies in process of interaction of medical personnel with patients of clinics. One of such developments – the system of diagnosis of bronchopulmonary diseases. This system assumes existence of the following functions: the analysis of data of the patient in the form of the block of text information, statement of the estimated diagnosis on the basis of the carried-out analysis and the revealed regularities, formation and issue of medical recommendations for increase in efficiency of treatment of the patient [1].

2. Problem definition

One of functional blocks of the developed system is the block of recognition of text information. Existence of this block is caused by specifics of collecting and storage of information on the patients asking for the help in medical institutions. The individual data of patients including both the general information, and the data obtained during delivery of health care (primary survey, holding necessary procedures and their results), are often stored in a type of the unstructured electronic, or printing text and also occupy essential volumes. At the same time it is necessary to understand that qualitative and timely analysis of this sort of information is capable to lead to essential increase in efficiency of diagnosing and treatment of patients. A prime task is transformation of the saved-up information on patients to a format of electronic text documents that will allow to start the further analysis.

3. State of Art

Process of recognition of the text represents a special case of a problem of recognition of images. Recognition of images - the scientific direction connected with development of the principles and creation of the systems intended for definition of belonging of a concrete object to one of in advance allocated classes of objects [2]. The possibility of recognition relies on similarity of the same objects. In spite of the fact that all objects and situations are unique in strict sense, between some of them it is always possible to find similarities on this or that sign. From there is a concept of classification — splitting all set of objects into not crossed subsets - classes which elements have some similar properties distinguishing them from elements of other classes [3].

There are several main approaches described in references to process of recognition of images:

- Heuristic (approach experience and an intuition of the person is the cornerstone, means transfer of members of a class and the principle of community of properties);
- Mathematical (approach rules of classification which are formulated and removed within a certain mathematical formalism by means of the principles of community of properties and a clustering are the cornerstone);
- Linguistic (The image can be described by means of hierarchical structure of subimages, similar syntactic structure of language) [4].

In relation to a solvable task preference is given to mathematical methods as they are the most universal and effective.

4. Proposed solution

During the work with text information the problem of optical recognition of symbols which consists in mechanical or wire transfer of images of the hand-written, typewritten or printing text in the text data which are used for representation of symbols in the computer is solved (for example, in a text editor). Recognition is widely applied to transformation of books and documents to an electron look, to automation of accounting systems in business or to the publication of the text on the web page. Optical recognition of symbols allows to edit the text, to carry out search of words or phrases, to store it in more compact form, to show or unpack material, without losing quality, to analyze information and also to apply to the text wire transfer, formatting or transformation to the speech [5].

In the developed system of diagnostics of bronchopulmonary diseases as entrance data for recognition of the text use of the files of the JPEG format containing the scanned documents of the A4 format with existence of the printing text is supposed. These documents contain necessary information on the patient, symptoms of an estimated disease and also results of different types of the carried-out analyses.

During primary analysis of the documents which are subject to processing the following aspects complicating recognition of text information have been marked out:

- symbols of Cyrillics can settle down, both in lower, and in top registers;
- tracing of symbols can significantly differ, owing to use at the press of various fonts or their sizes;
- as a result of scanning documents with a non-uniform brightness of symbols and also distortions of the text and perspective distortions can be received;
- the recognized text may contain various schedules, charts, tables, designations using Latin symbols, and the additional signs coded, for example, in Unicode.

The above-stated aspects considerably complicate process of recognition of text information and require special attention at the solution of tasks of this sort.

We will consider in more detail process of recognition of texts, having divided him into separate blocks from which it is accepted to distinguish the following:

- segmentation block (localizations and allocations) text elements;
- block of preprocessing of the image;

- block of allocation of signs;
- block of recognition of symbols;
- block of post-processing of results of recognition [6].

It is easy to notice that process of recognition of symbols is only one of components of the described task. First of all the initial image is exposed to the analysis which is designed to localize all significant fragments containing text information with the highest probability. At the same time it should be noted that the block realizing preprocessing of images can be carried out not only after the segmentation block, but, including, to it and, in some cases, together with a problem of localization of symbols.

Process of segmentation of elements of the text comes down to splitting the image of the document for certain areas, at the same time there is an allocation of structural text units, such as lines, words and symbols. Allocation of fragments of high levels (lines and words) can be carried out on the basis of the analysis of intervals between dark areas [7]. At first there is a division of the text into lines by means of the procedure of search between them white strips. Further the approximate value of average distance between symbols is set that allows to allocate separate letters, and distances, big in comparison with averages, will confirm the gaps concluded between words. At the same time it is supposed that offers of the text are located horizontally and have no crossings with each other.

There are several widespread methods of segmentation:

1. Dummy algorithm - the dummy segmentation algorithm takes the whole page as one segment. The purpose of this algorithm is to see how well we can perform without doing anything. Then the performance of other algorithms can be seen as gains over that achieved by the dummy algorithm;
2. The X-Y cut segmentation algorithm [8], also referred to as recursive X-Y cuts (RXYC) algorithm, is a tree-based top-down algorithm. The root of the tree represents the entire document page. All the leaf nodes together represent the final segmentation. The RXYC algorithm recursively splits the document into two or more smaller rectangular blocks which represent the nodes of the tree. At each step of the recursion, the horizontal and vertical projection profiles of each node are computed.
3. Run-length smearing algorithm RLSA algorithm (RLSA) [9] works on binary images where white pixels are represented by 0's and black pixels by 1's.
4. The whitespace analysis algorithm described by Baird [10] analyzes the structure of the white background in document images. The first step is to find a set of maximal white rectangles (called covers) whose union completely covers the background. Breuel's algorithm for finding the maximal empty whitespace is used in our implementation for this step.
5. Constrained text-line detection - the layout analysis approach by Breuel [11] finds text-lines as a two step process: Find tall whitespace rectangles and evaluate them as candidates for gutters, column separators, etc. The whitespace rectangles representing the columns are used as obstacles in a robust least square, globally optimal text-line detection algorithm [12]. Then, the bounding box of all the characters making the text-line is computed.
6. The algorithm Docstrum – is carried out a clustering of symbols, distances between them are calculated, at the same time three minimum distances on increase are allocated: 1 – distance between letters in a word, 2 – interlower case distance, 3 – a space;
7. The Voronoi-diagram based segmentation algorithm by Kise et al. [13] is also a bottom-up algorithm [14].

Finally the image is divided into certain sectors (squares) in which some symbol which is subject to further recognition with guarantee contains.

Further there is a problem of preprocessing of the image which consists in carrying out operations of smoothing over him, filtrations and normalization. Smoothing is meant as process of restoration of the symbols containing the gaps or gaps received during scanning. Also the return process – reduction of the lines significantly differing from others on thickness that allows to simplify further recognition of a symbol is carried out. Methods of binary filtration which give the chance to get rid of roughnesses of borders of a symbol are applied. Normalization of the image, in turn, is intended for elimination of distortions of separate lines and symbols and also reduction of symbols to the uniform height and width after their processing.

The following two operations realized by blocks of allocation of signs and recognitions of symbols have various purpose, but at the same time are closely connected among themselves as directly depends on quality of the marked-out signs whether the system will be able to distinguish this or that symbol. At this stage allocate four main techniques of recognition:

1. The search method – consists in comparing of a defined image with numerous number of the reference samples which are contained in certainly made database;

The deep analysis of characteristics of an image – a method is the cornerstone determination of different geometrical characteristics of characters, their key features;

Use of artificial neural networks - this method requires essential quantity of examples of the task of recognition when training and also special structure of the neural network considering specifics of an objective. It is necessary to mark that use of artificial neural networks is the most perspective and effective method;

The expert method – a method is based on the continuous training of expert system in use [15].

As a rule, above-mentioned techniques are applied not only separately, but also often integrate in a complex of the programs providing effective implementation of an objective on recognition of texts.

The unit of post-processing of results which purpose is improvement of quality of character recognition at the expense of assessment of the context information finishes recognition process. Use of methods of post-processing also allows to find the mistakes made in the text and to correct them. As a rule, two basic approaches are applied to the decision of tasks of this kind: use of dictionaries and sets of binary matrixes. The entity of the first approach consists search of the recognized words in certainly created dictionary. In the absence of coincidence to the dictionary the conclusion that the word contains an error is drawn, and the word is subject to changeover by the most similar to dictionary. The second approach is based on the N-grams representing the character strings consisting of 2-3 letters of a separate word. This method allows to obtain additional information on structure of the checked word, passing the analysis of a context and calculating the reliability level characterizing probability of absence of errors in the checked word. It is necessary to mark that data application of methods requires essential computing resources.

5. The choice of the used way of recognition of the text

The above-mentioned units executing text recognition process can be realized by different modules, programs and components. We will consider some of them.

Among software products today the leader in the field of recognition of the Cyrillic text it is possible to call the development of the Russian company "ABBYY" named "FineReader". The program allows to transfer images of documents (photos, results of scanning, the PDF files) to the electronic edited formats, such as Microsoft Word, Microsoft Excel, Microsoft Powerpoint, Rich Text Format, HTML, PDF/A, PDF, CSV and text files [16].

Also there is a set of other different software products in the field of text recognition: Office Lens (recognizes pictures of the camera and saves them in the format docx, pptx, PDF) [17], Adobe Scan (saves files only in the PDF format) [18], Free OCR to Word (Recognizes: JPG, TIF, BMP, GIF, PNG, EMF, WMF, JPE, ICO, JFIF, PCX, PSD, PCD, TGA and other formats, saves files in the DOC, DOCX, TXT formats) [19], Online OCR (recognizes: JPG, BMP, TIFF, GIF, PDF, saves: DOCX, XLSX, TXT) [20], Microsoft OneNote (recognizes: JPG, BMP, converts into files of OneNote) [21].

Unfortunately, use of above-mentioned products in the developed system is impossible because of absence of a possibility of their adaptation and implementation in system. So, the most perspective decision is use as means of recognition of text information the special modules adapted to different programming languages, in particular to Python.

The Python programming language has been chosen for development as the basic for a number of reasons:

- has the created modules for work in the field of recognition and information processing;
- has rather high speed of performance;
- it is optimized and is widely applied in the solution of problems of the required range.

As the module for recognition of text information use of the pytesseract module is offered. The kernel of the Tesseract program was developed in the Bristol laboratory Hewlett Packard in 1985 — 1994. In 1996 the considerable changes were carried out and the port for Windows is prepared [22]. The main advantage of this software package is the possibility of his adaptation under recognition of the Russian-language text.

Nevertheless, use of this program will also demand additional adaptation in view of specifics of the applied area. It is necessary to provide additional training of a product by means of medical dictionaries for increase in overall performance.

6. Conclusions

In conclusion it should be noted that the choice and setup of the block of recognition of text information is an important and responsible task. As this block is the first in the system of diagnosis of bronchopulmonary diseases, his incorrect work can lead to mistakes in all subsequent blocks of the program. As a result, at setup of the block of recognition it is necessary to pay special attention to his parameters to achieve maximum efficiency and to transfer extremely useful data to the information processing block.

References

1. Shakhmametova, G.R., Zulkarneev K. Kh., Evgrafov A.A., Analytical processing of medical data in the system of diagnosis of bronchopulmonary diseases. ITIDS'2018, International Scientific Issue, Volume 1, p.256-261 (2018)
2. Chichvarin, N.V., Recognition of images, Material from National library of N.E. Bauman, URL: https://ru.bmstu.wiki/Распознавание_образов (date of the address: 7/24/2018)
3. Zhuravlev, Yu., I., Recognition. Classification. Forecast. Mathematical methods and their application. Issue 2. M.: Science, 1989
4. Nazarov D.A., Methods of recognition of images. Basic principles, M.: Science, 2010
5. Optical character recognition, URL: https://en.wikipedia.org/wiki/Optical_character_recognition (date of the address: 7/24/2018)
6. Optical character recognition, URL: http://wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_(OCR) (date of the address: 7/29/2018)
7. Gritsai A, Methods of recognition of texts, URL: https://habr.com/post/112442 (date of the address: 8/1/2018)
8. Nagy, G., Seth, S., Viswanathan, M.: A prototype document image analysis system for technical journals. Computer 7 (1992)
9. Antonacopoulos, A., Gatos, B., Bridson, D.: ICDAR 2005 page segmentation competition. In: Proc. ICDAR, Seoul, Korea (2005)
10. Baird, H.S.: Background structure in document images. In: Document Image Analysis, World Scientific, (1994)
11. Breuel, T.M.: Two geometric algorithms for layout analysis. In: Document Analysis Systems, Princeton, NJ. (2002)
12. O Gorman, L.: The document spectrum for page layout analysis. IEEE TPAMI 15 (1993)
13. Kise, K., Sato, A., Iwata, M.: Segmentation of page images using the area Voronoi diagram. CVIU 70 (1998)
14. Shafait F., Keysers D., Breuel T., Performance Comparison of Six Algorithms for Page Segmentation (2012)
15. Kolesnikov S., Recognition of images. General information, URL: http://old.ci.ru/inform03_06/p_24.htm (date of the address: 7/26/2018)
16. Official site of the ABBYY software product, URL: https://www.abbyy.com/ (date of the address: 8/20/2018)
17. Official site of the Office Lens software product, URL: https://www.microsoft.com/ru-ru/p/office-lens/9wzdncrfj3t8?activetab=pivot%3aoverviewtab (date of the address: 8/20/2018)
18. Official site of the Adobe Scan software product, URL:https://acrobat.adobe.com/us/en/mobile/scanner-app.html (date of the address: 8/20/2018)
19. Official site of the Free OCR to Word software product, URL: www.ocrtoword.com/ (date of the address: 8/20/2018)
20. Official site of the Online OCR software product, URL: https://www.onlineocr.net (date of the address: 8/20/2018)
21. Official site of the Microsoft OneNote software product, URL: https://products.office.com/en-us/onenote/digital-note-taking-app (date of the address: 8/20/2018)
22. Official site of the pytesseract software product, URL: https://pypi.org/project/pytesseract/ (date of the address: 9/03/2018)