R. Gavrilyuk, Student
O. Pasichnyk, PhD in Educ., As. Prof., language advisor
Khmelnitsky National University

METHODS OF TEXT RECOGNITION

In the course of everyday activities, government structures, business, and academic institutions educational institutions use a large number of paper documents, most of which are handwritten. A large amount of data and knowledge is contained in printed or handwritten documents that are archived. The need is growing digitization of paper documents in order to further process their content computerized computer systems.

Text recognition can be divided into several areas that are sufficient significantly differ in their methods of solving. The text can be printed either manuscript. Any of them can be extra structured. For example, Formulas can contain different levels of records, such as superscripts, sublines, special marks for mathematical actions, etc.

To date, there are a number of methods that solve the problem print text recognition, but there are still no systems capable of to recognize any handwritten text [1]. Existing systems can suffice not qualitatively recognize specific handwriting. Therefore, the task of development is relevant handwriting recognition method that will allow you to process it handwritten documents.

Before the text is recognized, there is always a previous one processing the input image. The first step is to improve the quality image. At this stage, increase the contrast and sharpness of the image, as well as filtration from noise. The next step is segmentation [2], by which the structure of the text is determined. Segmentation somewhat different for printed and handwritten text. In both cases lines, words and letters are highlighted, but for the printed text segmentation letters are much simpler and occurs similarly to segmentation of words - for using the method of horizontal and vertical profiles [3]. For handwriting segmentation at the level of the letters is more complex: the letters can merge into one segment or vice versa, one letter disintegrate into several segments. It’s greatly complicates the recognition task.

After preliminary preparation, methods of recognition of handwritten and the printed text is more significant. For printed text the comparison of segmented letters with different standards is used fonts. Finding a match with one of them, for example, the first letter, The remainder of the text is cyclically recognized by comparing all the selected ones Segments with letters of a specified font.

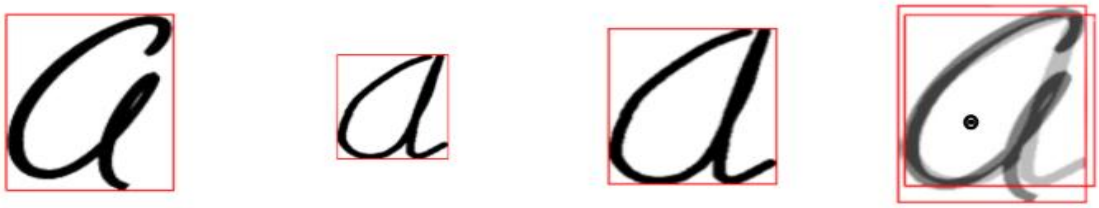
There are two approaches to recognizing handwriting recognition online and offline [4]. The first involves recognition directly at a time writing text and using algorithms to write characters that take into account the trajectory of the movement of the pen - the subject of writing. This approach is called online recognition. To date, the task This recognition for most languages can be considered solved. Contemporary Electronic notebooks use it extensively. The second approach is aimed at handwriting recognition that was written in advance. Offline recognition is much more important because the number is already written the text is enormous. The roblem is this recognition in the general case still not resolved.

There are two main types of methods for solving the problem of text recognition offline - structural and standards, as well as their combinations [4].

Structural methods are based on the selection and analysis of various structural elements of the symbol, their features and properties. Each letter is divided into knots and curves that connect them. On the basis of a set of such data is done conclusion which letter is written. However, there is a problem with that most of the letters are written not calligraphically and, accordingly, are not clear connections.

First, you need to conduct a preliminary study for each individual sample handwriting by creating the necessary reference base for this purpose. Letters fit into the rectangle in such a way that its sides are tangent to them. The letter shown in Fig.1.a - standard, in fig. 1.b - a letter separated from text using segmentation. An additional condition for comparison is the same scales of letters. Since the size of the selected segment for the letter, which need to recognize, and the letters from the database may differ, then in the second step you need to scale (Fig. 1.c). Often the linear dimensions are highlighted segment and template in horizontal and vertical planes simultaneously may not coincide, as even letters written by one person in part are different. However, there is enough coincidence horizontally or vertically (Fig. 1.a, 1.c).

The comparison of the selected segment with the standards takes place in such a way that their centers of mass coincide (Fig. 1.d). The litera has the most coincidence with a letter from the base being a copy of it, and may also be added to the base for further training.



a) b) c) d)

Fig.1. An example of horizontal-aligned letters.

The proposed method provides the ability to recognize handwritten text given sample. Its advantage is the ability to compare letters that do not have identical linear sizes. Due to the fact that it is enough to compare only centers of mass, the speed of accurate overlay of letters increases, due to reduce the number of overlays and mutually shifted images for the correct one positioning.

REFERENCES

1. Кучуганов А. В., Лапинская Г. В. Распознавание рукописных текстов / А. В. Кучуганов, Г.В. Лапинская // Материалы международной научной конференции, Ижевск, 13–17 июля 2006. – С. 98 – 103.

2. Shafait F. Performance Comparison of Six Algorithms for Page Segmentation / F. Shafait, D.Keysers, T. Breuel // Image Understanding and Pattern Recognition (IUPR) research group. –2006. – pp. 12.

3. Запрягаев С. А., Сорокин А. И. Сегментация рукописных и машинописных текстов методом диаграмм Воронного / С. А. Запрягаев, А. И.

Сорокин // Вестник ВГУ, серия: системный анализ и информационные технологии, No 1, 2010. – С. 160 – 165.

4. Васильев С. Распознавание непрерывного рукописного текста в режиме off-line[Електронний ресурс]. – Режим доступу: https://geektimes.ru/post/136165/ – Електрон.тестові дані (дата доступу 02.03.2016).