MATNLARNI TANIB OLISH MUAMMOLARI VA YONDASHUVLARI

Umarova Benazir Botirjon qizi
NamMTI Avtomatika va energetika fakulteti
Informatsion texnologiyalar kafedrasi assistenti

ANNOTATSIYA

Kletkali avtomatlar, birinchi navbatda, soddaligi tufayli alohida qiziqish uyg‘otadi: oddiy qoidalarga asoslanib, kletkali avtomatlar murakkab hatti-harakatlarni yaratishi mumkin. Bundan tashqari, kletkali avtomatlar parallel hisoblash uchun mos variantdir, ular ko‘p prosyessorli tizimlarda samarali ishlatilishi yoki apparatda amalga oshirilishi mumkin, chunki ularning qoidalarining asosiy xususiyati mahalliylik va bir xillikdir. Kletkali avtomatlarning afzalliklari matnni aniqlash tizimida foydali bo‘lishi mumkin. Qoidalarning soddaligi va bir xilligi bir nechta mantiqiy yoki matematik elementlarga asoslangan murakkab tizimlarni yaratish va kamroq hisoblash resurslari va xotira bilan natijalarga erishish imkonini beradi.

Kalit so‘zlar. Matnni aniqlash, XADARA loyihasi, segmentatsiya, belgilarni ajratib olish, tanib olish, belgilar vektori, pochta kodlari, literal ma’lumotlar.

PROBLEMS AND APPROACHES OF TEXT RECOGNITION

Benazir Umarova
NamIET Faculty of Automation and Energy
Assistant of the Department of Information Technology

ABSTRACT

Cellular automatons are of particular interest primarily because of their simplicity: based on simple rules, cellular automaton can generate complex behaviors. In addition, cellular automatons are a suitable option for parallel computing, they can be efficiently used in multiprocessor systems or implemented in hardware, because the main feature of their rules is locality and uniformity. The advantages of cellular automaton can be useful in a text recognition system. The simplicity and uniformity of the rules allows to create complex systems based on few logical or mathematical elements and achieve results with less computing resources and memory.

Keywords: Text recognition, XADARA project, Segmentation, Character extraction, Recognition, Character vector, ZIP codes, literal data.

Text recognition is one of the most active and leading branches of image processing and pattern recognition. The achievements in this field can be usefully used in tasks such as searching for information from large texts, making changes to them, making them easy to read and edit, and editing documents containing handwritten information. The problem of text recognition has been solved to some extent for many languages of the world. In our work, we focused on the issue of creating a system that recognizes texts in the Uzbek language.

There is an active community of scholars engaged in text recognition. One of the largest conferences in this field is the International Conference on Frontiers in Handwriting Recognition (ICFHR). It occurs in even years. Another conference, International Conference on Document Analysis and Recognition, is held in odd-numbered years.

One of the notable works on text recognition was published in 2000 by R. Plamodon done by Srihari [1]. Srihari and his colleagues were the first in the world to create a system for recognizing handwritten addresses [2]. This system would read the numbers and letters separately, by first breaking the text into segments. It then compares the numbers to the ZIP codes in the database, and the literals to the region names, and returns the results with the highest probability.

Also noteworthy is the XADARA project [3] developed at the Braunschweig Institute in Germany in 2014. XADARA is a semi-automatic software system for working with ancient Arabic manuscripts. XADARA project implementers are a team of scientists in signal processing, computer science, history and linguistics. The essence of the X KHADARA system is a system of “Equipment” that facilitates work on ancient sources through the organization of computerization and information retrieval. This system is extremely useful not only for librarians, but also for historians. Because they expand the scope of analyzing the history and features of Arabic writing.

Now let’s briefly review the above steps.

Image input is usually done by optical scanning. A digital image of the original document is obtained through the scanning process. Recognition usually uses optical scanners, which consist of a transport mechanism and a sensitive device that converts light intensity into grayscale surfaces.

Initial processing. The image obtained as a result of the scanning process may contain a certain amount of “Noise”. Depending on the quality of the scanner and technology, characters may be left or distorted. Some of these defects may later cause poor quality results. These problems can be overcome by using binary, padding, and thinning operations.

In the process of binary, the input image is converted into a binary image. That is, pixels in the background get a value of 0, and pixels assumed to contain text get a value

of 1. Then it “Walks” through the binary image of the symbol, first determining the scales. In the process of filling, small gaps that appear in the image are closed. In thinning, the widths of the lines are reduced.

Segmentation. Segmentation is a process in which the components of an image are determined. When applied to text, segmentation is the act of isolating characters and words. Most recognition algorithms divide words into discrete characters and recognize them individually. Usually segmentation is done by isolating each connected component, which means a single unbroken black area.

Character extraction. The purpose of character extraction is to record the salient features of characters, and it is generally considered to be one of the most difficult problems in character recognition. In this case, a set of symbols related to the symbol is created. Then the range of characters is determined. More important characters are given a higher value in the compatibility check.

Recognition. Neural networks are often used in the text recognition stage. Compared to other mesh models, the advantage of neural networks is that they consist of many layers of elements with internal connections. The feature vector is the input layer in this mesh. Each element of the layer calculates the overall importance level of this input parameter and converts this into an output result via non-linear functions.

During the “Training” process, the importance of each link is changed until the desired output is obtained. The disadvantage of neural networks in text recognition is their limited prediction and generalization, and their advantage is their adaptive nature.

Pattern recognition. OCR (Optical Character Recognition) software usually works with a large raster image of a scanned page. However, most systems have templates created for different styles. After a few recognized words, the program detects the font in use and searches for matching pairs only for that font. In some cases, the software uses numerical values of the parts (proportions) of the characters to determine the new font. This can improve recognition performance.

TypeReader recognition software uses machine-specific algorithms based on a template approach.

This approach requires creating a template for each font. For example, TypeReader uses 2100 different character styles.

Structural approach. The world’s best-selling system OCR - Caere OmniPage Professional uses an algorithm based on finding common character features.

Conclusion. In this brochure, we talked about text recognition technologies. We have seen several real examples. As we have witnessed, there are many things that need to be done in this field in the Uzbek language, and there are many issues that are waiting to be solved. In our future research, we will try to raise the work done on Uzbek textual

data recognition to the level of the world community, and add it to the list of systems recognized at conferences such as ICDAR, ICFHR.

REFERENCES

1. Plamondon R., Srihari S. N. On-line and off-line handwriting recognition: a comprehensive survey //In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1), pp. 63-84.
2. Wang C.-H., Palumbo P.W., Srihari S.N. Performance evaluation of a system to recognize address block on mail pieces //AAAI-88: Seventh National AI Conference, Minneapolis, 1988, pp. 837-841.
3. Pantke W and etc. HADARA - A Software System for Semi-Automatic processing of Historical Handwritten Arabic Documents //In Proc. Archiving Conf. 2013, pp.161-166.