PERIODICA
Journal of Modern Philosophy, Social Sciences and Humanities
ISSN NO:2720-4030 Volume 8, July, 2022

Methods of the Pattern Recognition in the Uzbek Language on the Basis of Cellular Automata

Turkmenova Rukhiye Tokhirovna
Assistant of the department "Software engineering" of the Samarkand branch of the Tashkent University of Information Technologies
Samarkand, Uzbekistan
(e-mail: ruha.5202155@yandex.ru)

ABSTRACT
The article describes the usage of the method of cellular automata in the OCR in the Uzbek language on the basis of the intersecting allocation algorithms, ends and symbol loops, allocation of endpoints and intersections

ARTICLE INFO
Received:1st May 2022
Revised: 1st June 2022
Accepted: 6th July 2022

KEYWORDS: cellular automata, pattern recognition, image, signs of symbols, algorithms, selection of endpoints, the selection of intersection.

In the age of information technology, the so-called "information explosion" or the growth of the disproportion between the amount of information produced by mankind and the amount of information that people are able to consume and assimilate, not an unimportant problem is both the accuracy of information, its presentation and design, and grammatical "correctness" submission of information or the reliability of the transmission of the message. The operation of information systems, the correct operation of enterprises, etc. depend on the reliability of information. The reliability of information often decreases with errors made by a person, a machine, communication failures, and errors in recognition. When processing a large amount of information, for example in electronic document management systems, spelling errors are common. [1]

Error detection in the context is directly related to the use of recognition methods, when the controlled sections of the text must be selected and verified with reference samples stored in the databases of the relevant information systems.

One of the most difficult sections of the field of image recognition is the field of text recognition. There are many techniques for text recognition. One of them is the method of text recognition by the method of cellular automata.

Cellular automata are discrete dynamic systems whose behavior is completely defined in terms of local dependencies. The space is represented by a uniform grid, each cell or cell of which contains several bits of data; any cell at each step calculates its new state from the states of its close neighbors. The calculation methods in such a system are local and the same everywhere. [2]

Before the process of extracting character features in a text recognition program, it is necessary to solve several problems: it is necessary to process the text image from noise, bring it to a state that allows you to fulfill the conditions of recognition algorithms, and extract individual images and character features from it.

Each character of the text, in particular in the Uzbek language, has its own unique features. These features uniquely distinguish characters from each other.

Text symbols have a large number of signs: the position and slope of lines, arcs, the presence of loops, vertical - horizontal lines, protrusions and their slope, intersections. The main features can be considered protrusions, loops and intersections, as well as their relative position.

After processing and filtering, signs of symbols are highlighted. After that, a classification process is supposed, which, based on the obtained features, will determine the name of the symbol. Classification is most often based on methods that, based on the accumulated information about the features, determine the symbol.

The recognition system assumes the presence of a training block. The training of the system is directly related to the classification, it allows you to change and correct the coefficients based on the association of the classification result with the name of the symbol.

Cellular automata are also involved in the feature classification process. The idea of classification is the creation of a characteristic cellular automaton for each feature and its correction, taking into account certain features in the learning process.

Formally, a cellular automaton can be defined as a set

{G, Z, N, f},

Where G is the metric of the field on which the cellular automaton acts;

Z is the set of states of each cell;

N is the neighborhood of the cell, which affects the state of this cell;

f is cellular automaton rules, which can be written mathematically

$Z \times Z^{w} \wedge Z$

The properties of a cellular automaton are: the locality of the rules, the homogeneity of the system, the finiteness of the set of cell states, the simultaneity of changes for all cells.

Since the main elements of symbols are loops, intersections, the position of the ends, these elements are used for segmentation and image identification, correlation with a specific symbol. There are many strategies for extracting the described features based on cellular automata. Two such strategies that use labeled cellular automata are described below. [3]

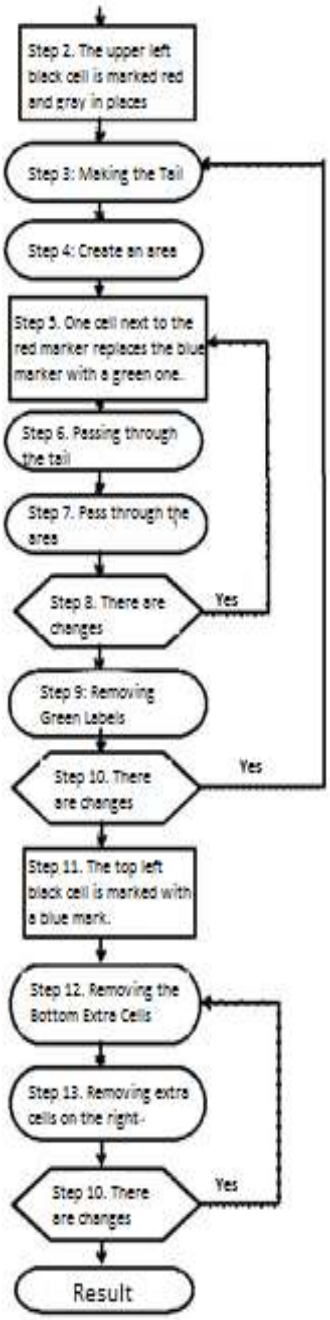

Figure 1. Scheme of operation of a sequence of cellular automata for the first algorithm for extracting features of symbols

The strategy for selecting endpoints and intersections. The strategy is to echo from the top of the symbol along the dots that make up the symbol. This echo is divided into components, following the outline of the character image. At some point, echo components meet or fade at the end of a character. [4]

When creating text recognition algorithms in the Uzbek language, the segmentation echo components are defined as follows:

1. "Echo area" - character dots that move from one end of the character image to the other.
2. "Echo tail" - these are the image points where the echo area was located at the previous moment in time.
3. "Points of the traveled path" - these are the points of the image where the "echo area" was present, and then the "echo tail", at these points the process does not resume.

Initially, all points are not labeled. After the first echo point is marked, the algorithm starts its work.

The algorithm is based on the following idea. As the echo passes, at some point the "echo area" will die out while the tail is still present. This event can only happen at the end of a character or at the meeting point of two echo components. The position of the echo tail at this moment is remembered.

Also, the event of the meeting of two echo components is registered based on the fact that the tails of the two echo components at the time of the meeting are not connected. Thus, the position of the symbol loop is remembered (figure 1).

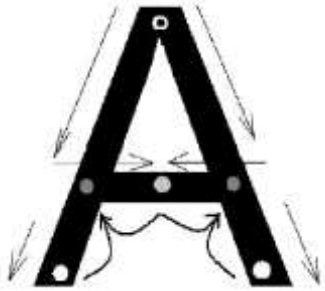

Figure 2. The direction of propagation of the "echo" along the character "A" with marking the positions of the ends of the characters and loops, as well as the propagation of the "echo" with marking the position of the intersections

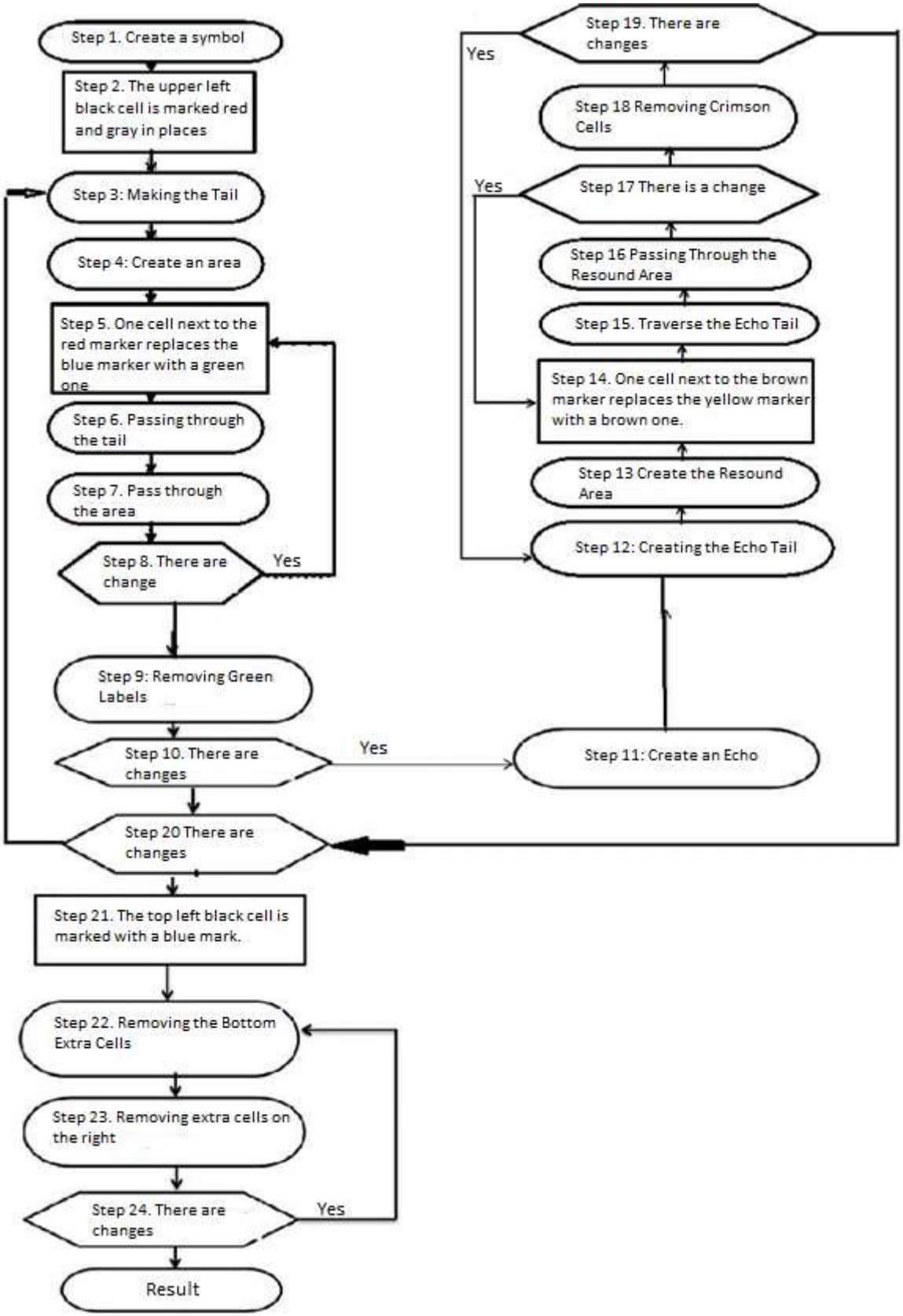

Figure 3. Scheme of operation of a sequence of cellular automata for the algorithm for extracting features of symbols

Strategy for highlighting intersections, ends and loops of the symbol . This strategy is an improvement and development of the endpoint and intersection detection strategy. In the process of propagation of the

"echo" at the moment when the cell of the end of the symbol or the cell of the meeting of two components of the "echo" is determined, a response "echo" is generated - "echo" along the already passed cells. The "echo" travels the path in the opposite direction and marks the cells in which the initial echo was divided into components. This allows you to find the positions of the intersections of the segments that make up the characters, as well as the loops that are in the letters of the Uzbek alphabet.

The strategy for highlighting intersections, ends and loops of a symbol determines the positions of the intersections of lines in the image of the symbol (Figures 2 and 3).

The algorithm may seem cumbersome, but it contains two components similar to the first algorithm. The rules of the automata of this algorithm are similar to the rules of the automata of the first algorithm for extracting the features of symbols. The advantage of this algorithm is the allocation of more character features than with the previous algorithm, so it takes more steps and lasts longer.

Bibliography.

1. М. М. Камилов, А. Р. Ахатов Система контроля достоверности текстовой информации на основе n-граммных парсинговых моделей.
2. Тоффоли Т., Марголус Н. Машины клеточных автоматов. М.: Мир, 1991. 280 с.
3. Д.И.Суясов Разработка алгоритмов распознавания текстов на основе клеточных автоматов, Санкт-Петербург. 2007. 34 с.
4. Д.И.Суясов Разработка алгоритмов распознавания текстов на основе клеточных автоматов, Санкт-Петербург. 2007. 35 с.