Text characters recognition on the basis of rival cellular automata functioning in parallel

Myroniv I.V., Zhikharevich V.V. , Ostapov S.E.

Yuriy Fedkovych Chernivtsi National University, Ukraine

Ivan.Myroniv@gmail.com, Vzhikhar@mail.ru, sergey.ostapov@gmail.com

Abstract. *The research of possibility of cellular automata application in the tasks of text characters recognition is conducted in this paper. For this purpose the notion of rival cellular automata functioning in parallel is introduced and algorithms of their functioning and interaction are developed. A model program was created for the realization of proposed algorithms. It allowed to evaluate the effectiveness of cellular automata algorithms, experiment with text characters recognition and demonstrate some advantages in comparison with other methods.*

Keywords

Text characters recognition, rival cellular automata, probabilistic Moore machine, characters overlapping introduction.

1 Introduction

Text images recognition is a very difficult task from theoretical and practical points of view. Human, for instance, involves for this full range of knowledge and experience. He full defines text from the unity of tactile organs signals distinguish each symbol, and typical symbol features and on the basis of his experience comes to the conclusion of the symbol and the whole text meaning.

Computer is mistaken in the process of recognition more often than a human. Nowadays perfectly correct method of text and symbol definition upon their image doesn't exist. Many developed commercial projects use their patented methods and can not boast of a perfect task solution.

In many cases good solution gives a comprehensive approach to the set problem. The task of text characters recognition divides into subtasks: image noise filtering, symbols images of the text image selection, symbols features selection and these features comparison with the saved samples. Each task has multiple solutions of which only some are more or less optimal.

The main task of this work is an attempt to use the theory of cellular automata in solving the problem of text characters recognition. Nowadays theory of cellular automata is a sufficiently developed branch of science [1, 2]. Special interest to cellular automata is shown first of all because of their simplicity: on the basis of the simple rules cellular automata can beget complex behavior. Besides, cellular automata are perfect variant for parallel computing and can be effectively used in multiprocessor systems or can be hardware realized as the main features of their rules are locality and homogeneity.

Advantages of cellular automata can be useful in the text recognition system. Simplicity and homogeneity of the rules can allow to create complex systems on the basis of several logical or mathematical elements and to achieve result with lesser losses of both calculating resources and the memory.

The major tendency of researches that are conducted in the work is the definition of cellular automata applicability in tasks and subtasks bounds, which appear in the text recognition process. For this the analysis of cellular automata and their features should be fulfilled, the main cellular automata characteristics should be selected which are necessary for the text recognition tasks solution and appropriate algorithms should be developed.

In the process of investigation a model and a program on the basis of it for the developed ideas and algorithms realization should be created.

2 Problem and its solution

Text recognition system presumes availability of the image with the text on its entry (in the graphic file data format). The text, selected from this image should be formed in the output of the system.

In general the task of text recognition includes such subtasks and subprocesses.

1. Image that gets to the system input has to be noise refined and reduced to a form which allows effective symbols selection and recognition.
2. System has to divide images into text blocks basing on peculiarities of its alignment and assignment into several columns.
3. Image with texts should be divided into line images and further into symbol images in order to process each symbol separately.

After this step different recognition systems work according to their peculiar algorithms.

4. Symbol image can be processed entirely, for this purpose it is compared with available patterns. Another variant is the displayed symbol characteristics singling out: peculiar features selection and their classification by available criteria in system.

The possible variant of the letter appears on the output of the fourth step. Although systems usually don't stop on this step and continue working on the basis of other methods ascertaining the received result.

5. Recognition result may not be satisfactory. For better results achievement a learning block can be built in a system. With the help of this block different letters images examples of the given font can be assigned in the system. Text recognition better quality is expected after the learning process. Text recognition system should not always follow on all described steps but the recognition process main actions are general for evens algorithm.

There are several text recognition systems. All of them are commercial products and to many internal algorithms of their work general access is forbidden. The principle such systems work is based on several strategies [4] but often recognition algorithm in general consists of consecutive submission and verification of hypothesis. Herewith the order of their submission is managed by knowledge about the investigated object and results of pervious hypotheses verification.

Attempts of recognition task solution were also made on the basis of cellular automata [5, 6]. Unfortunately, given works contain only expressions about cellular automata construction possibility for text recognition.

Let's consider propounded cellular automata recognition algorithm on the example of hexadecimal notation symbols, which are represented as a combination of horizontal and vertical lines (Fig.1).

Fig.1. Set of hexadecimal notation symbols 0123456789ABCDEF

Idea of cellular automata recognition algorithm in our case is based on the characters text nature. First of all symbols differ one from another not by size and lines thickness but by peculiar position of lines relative to each other. Secondly, symbol feature doesn't change on two symbols superposition, lines slope or whatever distortion (Fig.2), which don't change lines mutual position. For example in Fig.2a number 3469, is clearly visible, in Fig.2b - number 28, in Fig. 2c numbers 6 and 9 are clearly visible, in Fig. 2 d - number 444.

Here should be mentioned the simplicity of indicated examples recognition by a human being unlike artificial intellect systems that exist nowadays for which given symbols with a high degree of probability will be mistakenly recognized.

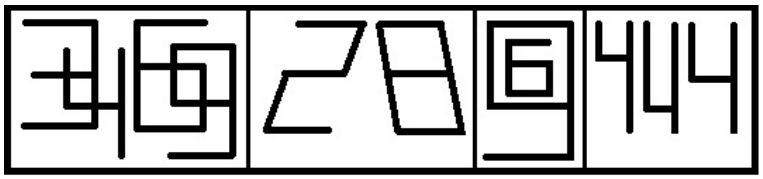

Fig.2. Examples of symbols distortion (a - superposition, b - slope, c - size changing, d - proportion changing)

Besides, typical for symbols is the mutual location of symbol lines not only relative to each other but also to the dimensional coordinate system. For example sufficiently complicated symbol recognition is shown in Fig.3 (Fig.3a – 6 or 9; Fig. 3b – 3 or E).

Fig.3. Examples of ambiguous symbol recognition

In this work we use a system of certain cellular automata types which describes proper symbols, in other words, the motion trajectory of a specific automaton type coincides with the appropriate symbol. Besides, such rules of functioning and cellular automata interaction are set that from any initial state of cellular automaton field transition in steady-state condition is occurred which in the vicinity of the specific symbol is a set of certain type automata. So the task of symbol recognition turns to the task of cellular automata set types. It is rather convenient and visible to perform this with the help of cell type certain color confrontation. Then separate symbols during the recognition algorithm will gain typical proper color.

Text symbols are depicted as an appropriate cell set in cellular automaton field. Cellular automata can move only along trajectory which coincides with symbol fragments. It is quite evident that a system of different cellular automata types each of which would describe appropriate symbol can be build.

Let's present algorithms of cellular automata functioning as transition graphs of probabilistic Moore automaton (Fig. 4).

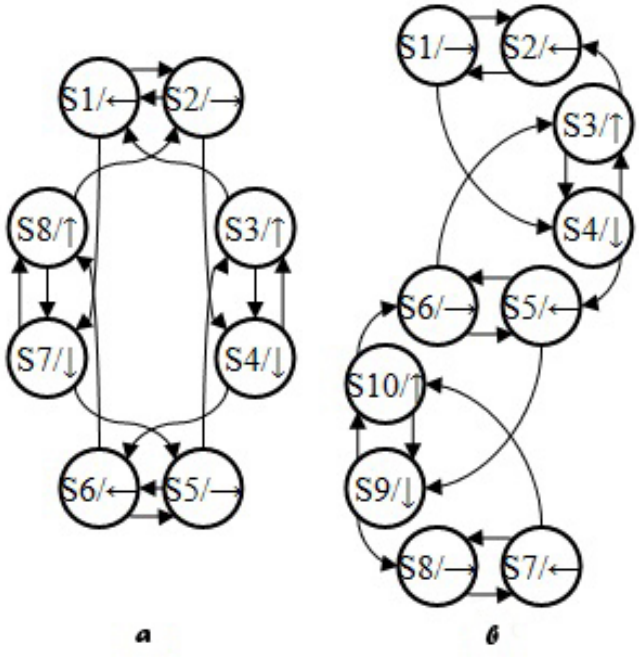

Fig.4. Examples of probabilistic Moore automata transition graph (a for 0, b for 2)

Here the input signal (another state of automaton transfer condition) is either the end of the symbol line cell or reaching cellular automaton in the point of ramification which takes place in symbols: 3, 4, 6, 8, 9, A, B, D, E and F. At the same time automaton transfers into one of the equiprobable state sets (according to the transition graph).

Outgoing reaction of cellular automaton is a signal about cells direction in given moment of time (shown by arrows near states of the transition graph: → movement to the right, ←movement to the left, ↓ down motion, ↑ up motion). Herewith cellular automata can move only within boundaries of cells that respond cells of symbols.

It is quite evident that cellular automata specified by graph shown in Fig.4a, will describe number "0" and in Fig. 4B - number "2". Similarly cellular automata transfer graphs can be build which will describe other symbols, shown in Fig.1.

On the other hand, recognition task doesn't foresee a priory information concerning relation of symbols to the specific class. That is why as it has been mentioned such an algorithm of cellular automata functioning and interaction should be provided that in process of algorithm work certain automata types accumulate in cells of those symbols which are in a line with these types.

Lets assume that in the cells field there formed symbol images as it is shown in Fig. 5. Let’s fill in cells field with cellular automata of different types and different initial states as it is shown in Fig. 6.

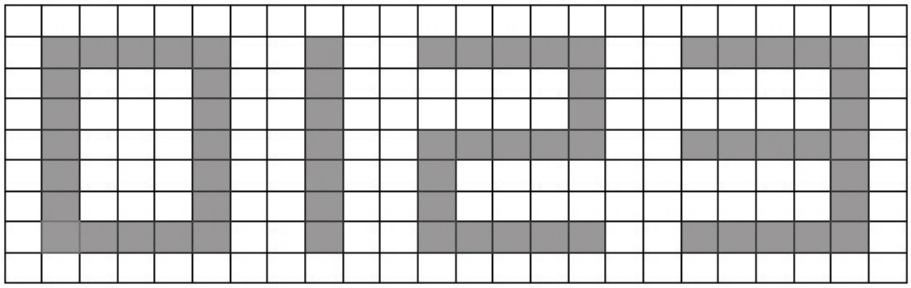

Fig.5. Symbol images in a cellular automata field

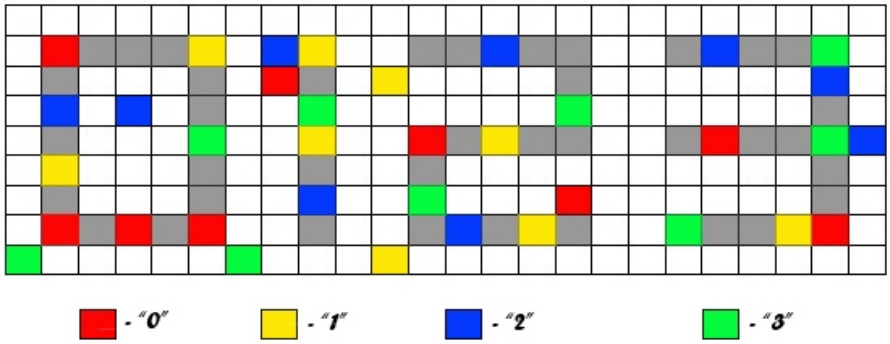

Fig.6. Random filling of the field by different types of cellular automata

Algorithm of cellular automata functioning provides statistical analysis of states in which cellular automaton remains. In the case of unattainable states, appropriate automaton is deleted from a field. This situation takes place in cases when cellular automaton doesn't remains in a cell that refers to the symbol (automaton remains in the initial state) or when cellular automaton describes a symbol fragment which doesn't refer to an appropriate class. For example, if automaton of a "0" type remains in the upper part of symbol "2" it never gets to a lower part of and also never gets S7 and S8 states (see Fig. 4a). Besides, if an automaton remains in the initial state which conforms to the moving directions perpendicular to the symbol fragment, it won't also move and all states excepting the initial one will be unattainable so such automaton will be also deleted from the field.

After certain quantity of cellular automata interactions automata with unattainable states will be removed and cell field will look in a way as it is shown in Fig.7.

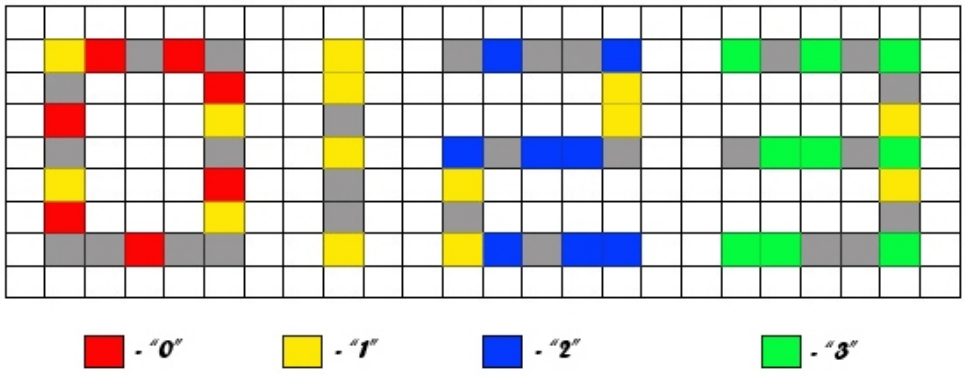

Fig.7. Figure of cellular automata field after the unattainable states automata removal

As it is seen from the picture one automata of "0", "1", "2" and "3" types remain in cells of appropriate symbols. It is also seen that automata of "1" type describe vertical fragments of all symbols. Indeed, cellular automates graph of "1" type provides for the availability of only two states with output signals: ↓ down motion and ↑ up motion. So moving within boundaries of any symbols vertical fragments, automata of "1" type will always run all their states. So they will never remove. At the same time it is quite obvious that symbols mentioned in Fig.7 should be perceived as 0, 1, 2 and 3 in other words cellular automata of appropriate types should "eject" the rest types, even those for which a condition of all states attainable is fulfilled.

In connection with a described phenomenon it is necessary to complete algorithm of cellular automata interaction with the aim of their "competition" providing. Indeed, average quantity of cellular automata of "0", "2" and "3" types

prevails average quantity of "1" type automata in appropriate symbols. So the removing of "1" type automata with the fragments of these symbols is quite natural.

The simplest way of this problem solving is the providing of random equiprobable peculiarities transfer from one cell to another during their change. In other words, while meeting of two cells there has to exist a probability that either one cell will copy its features (type, state etc.) to another or vice versa, so "infection" of one cell by another will take place.

Besides, cells "duplication" of that types which successfully describe, appropriate symbols can be provided. In this case competition algorithm can consist in a random equiprobable absorption of one cellular automaton by another.

In both cases because of prevailing quantity of one cells comparing to another and existence of symbol fragments where there is no competition, automata of "1" type will be ejected from symbols 0, 2 and 3. So cellular automata field becomes the way it is shown in Fig.8.

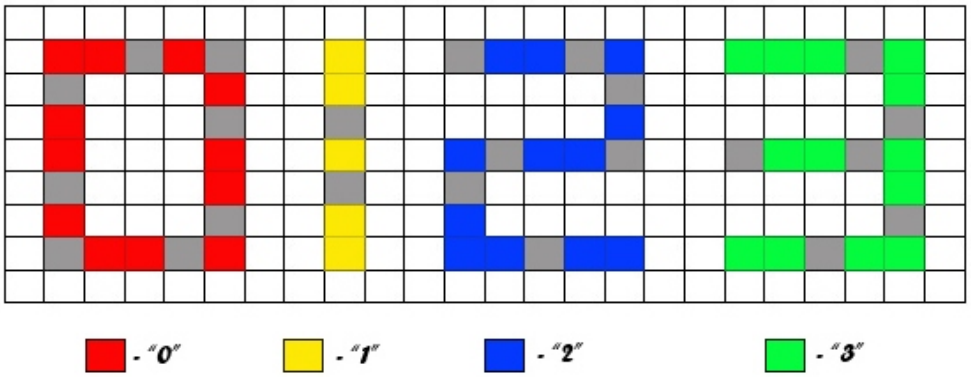

Fig.8. Figure of cellular automata field in case of automata "competition"

As it is seen in Fig.8, now automata of "1" type remain only in cells of symbol 1 and from fragments of the rest of symbols they are ejected. Thus into the cellular automata sets of types analysis in the appropriate areas of a field characters recognition can be performed.

Similar approach can be called as a method of "rival cellular automata". This method concerning text symbols recognition can be modified with the aim of recognition not only symbols that consist of vertical and horizontal lines but of any others in particular of hand written text. It is obvious that in such case automata graphs that describe cells behavior will be more complicated than those investigated in this work (see Fig. 4). On the other hand, in individual cases it is possible to transform symbols into rectangular view.

Besides, with growth of automata graphs complication their selection can be performed that is to say organization of genetic search procedure of the most optimal, effective and accurate automata graphs which describe one or another symbol (in a similar with work result [7]).

3 Results and analysis

Cellular automata functioning and interaction algorithms described above were realized as a modeling program. It allowed estimating efficiency of cellular automata algorithms and implementing the experiments of text symbols recognition. Successful recognition of hexadecimal symbols set shown in Fig.1 has been carried out. In particular an experiment of symbols recognition which partially superpose upon each other (Fig.9) has been successfully performed.

Symbols recognition was rather distinct in cases when symbols didn't have mutual lines but were only intersected. Indeed, on rectilinear mutual areas of symbols a high competition degree of automata takes place which complicates stable appropriate types cellular automata set formation process. In case of mutual intersection of symbols by lines-fragments, probability of rival automata interactions considerably lessens. Here it should be noted that cases when symbols have mutual lines are difficult to recognize even for human being.

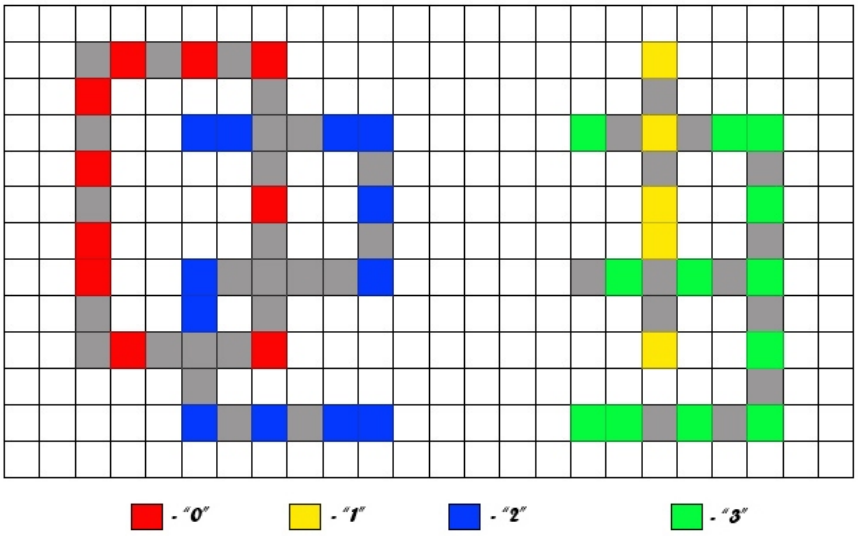

Fig.9. Symbols, with partly superpose recognition (0 2 1 3) example.

Symbols superposition sets of similar cellular automata had to be analyzed for the purpose of the set center location in the space of cellular field for the determination of the order of symbols disposition.

Besides order of dislocation quantitative characteristics of cell sets were analyzed. So for instance in the case depicted on the right in Fig.9 the availability of stable noticeable quantity of "9" type automaton was observed. It is connected with peculiar combination of numbers 1 and 3 that at superposition depicted on fig. 9, form a considerable cell space for number 9. Moreover, if symbol 1 is moved to the far left side of number 3, cellular automata of "3" type will disappear (to any person such situation will remind symbols 19 and not 13).

So a conclusion about duality and efficiency of offered cellular automata method of text symbols recognition can be made. Besides, it should be mentioned advantages of method in comparison with methods that use artificial neural systems for which symbol position recognition is rather difficult process. For method developed in the given work such symbol parameters as size, thickness of lines and fragment proportions are absolutely insignificant as well as for human being.

In conclusion it should be mentioned that from the point of work speed view the offered method as well as any other cellular automata algorithm can be realized the most optimally in computing systems with parallel architecture. If there isn't such a possibility additional measures concerning optimization of cellular automata program model should be taken.

In particular, taking into consideration high degree of tenuity of cellular automata field (elements of symbols and accordingly cellular automata occupy low percentage of field cells) it is necessary to foreknow auxiliary index massive in which all information about cellular automata should be kept: automaton coordinates on field, automaton type, state, maximum number of states, output signal, state counters (for an analysis of statistics of remaining in one or another state) etc.

Cellular automata field should be organized as a two-layer structure where the first layer corresponds to cellular automaton type (for the quick displaying) and the second one corresponds to automaton number in index array (for the quick search of interacting cellular automata).

At algorithm working it is necessary to choose one of cellular automata set on every next stage that is to say one element from auxiliary index array and avoid "empty" cycles in such a way.

Also such situations are possible when cell types that have to describe corresponding symbols are accidentally deleted from cells that correspond to symbols. In order to avoid such phenomena it is necessary to enter new automata types in different points of cellular automata field randomly and with low probability degree (in order not to break stability of formed cellular automata sets).

References

[1] Wolfram S. A., New Kind of Science. Wolfram Media. Inc. , 2002.
[2] Smith R.A., Real-Time Language Recognition by One-Dimensional Cellular Automata. J. of Computer and System Sciences, v. 6 (1972)
[3] Buchholz T., Klein A., Kutrib M., Real-Time Language Recognition by Alternating Cellular Automata. IFIG Research Report 9904. 1999. March.
Guttmann W., Simulated Evolution of Artificial, Path-Following Ants using a Genetic Algorithm. 2003