BEST: International Journal of Management, Information
Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X,
Vol. 4, Issue 6, Jun 2016, 61-68

RECOGNITION TEXT INFORMATION

MUHAMEDIYEVA D.T[1], MAMATOV A[2] & ISKANDAROVA S[3]

[1]Center for development hardware and software products under TUIT, Uzbekistan

[2,3]Guliston state university, Uzbekistan

ABSTRACT

The paper describes the basic ideas and methods to solve sub problems such as Sellisele algorithm, Neural network structures, Invariant numbers and their modifications used in modern recognition systems. The main objective of the research is the development of software complexes of recognition of text data on the basis of genetic algorithms.

KEYWORDS: Neural Network, Sellisele Algorithm, Recognition Block, OCR

INTRODUCTION

The task of recognition of textual information in the translation of printed and handwritten text into machine code is an essential part of projects aimed at automation of the document flow. However, this task is one of the most complex and high-tech in the field of automatic image analysis. Even the person reading handwritten text, divorced from context, makes an average of 4% error. With regard to reading printed documents, the difficulty lies in the fact that in critical applications, such as, for example, automation of the input of the passport and visa information, you need to ensure the highest reliability of recognition (more than 98-99%) even in poor quality printing and digitizing of the source text.

In recent decades, through the use of modern achievements of computer technologies, we developed new methods of image processing and pattern recognition, making it possible to create such systems for recognition of printed text that satisfies the basic requirements of workflow systems. However, the creation of each new application in this area remains a creative task and requires further research in connection with specific requirements for resolution, speed, reliability, recognition, and memory that characterize each particular task the development of problem-oriented systems of automatic input of paper documents.

The various technologies grouped under the General term "character recognition", divided into recognition in real-time and recognition in batch mode, each of which has its own hardware and its own recognition algorithms.

In a typical optical character recognition system (OCR), the characters are read and digitized by an optical scanner. Thereafter, each symbol is subjected to localization and separation, and the resulting matrix is subjected to the preprocessing, i.e. smoothing, filtering, and normalization. The result of preprocessing singles out characteristic features, followed by classification.

STATEMENT OF THE TASK

Identification task of textual information when translating printed and handwritten text into machine code is one of the most important components of projects, aimed at the automation of document circulation. However, this task is one of the most complex and high-tech in the field of automatic image analysis. Even person who read your handwriting, in isolation from the context makes an average of 4% of errors. As far as the reading of printed documents, the difficulty lies in the fact that in critical applications, such as automation of input of the passport and visa information, it is necessary to ensure high reliability of recognition (over 98-99%) even in poor print quality and digitization of the source text.

In recent decades, thanks to the use of modern achievements of computer technologies were developed new methods of image processing and pattern recognition, what made possible the creation of such systems recognize printed text, which would satisfy the essential requirements of systems of automation of document circulation. However, creation of each new application in this area remains a creative task and requires additional research in connection with the specific requirements of the resolution, performance, and reliability of recognition and the amount of memory that characterize each specific task to develop problem-oriented systems of automatic input of paper documentation. Different technologies, under the general term "recognition", are divided into recognition in real time and recognition in batch mode, each characterized by its own hardware and self-recognition algorithms.

In typical optical character recognition (OCR) software input characters are read and digitized by an optical scanner. After that, each character undergoes localization and allocation, and the resulting matrix is pretreatment, i.e., smoothing, filtering and normalization. Result of preprocessing of the characteristic signs, followed by the classification. The work describes the basic ideas and methods to solve these subtasks, as well as their modifications, used in modern systems of recognition.

MAIN PART

Theory of machine vision there is not the first day, on this in the reference one can find plenty of approaches and solutions. For a start, here are some of them:

Sellisele Algorithm

It is a method of recognition single binary images, based on the build skeletons of these images and the selection of the skeletons of edges and nodes. Next on the ratio of ribs, their number and the number of nodes is constructed a table of figures. For instance, the skeleton of a circle will be one node, the skeleton U - three ribs and two nodes and edges are like 2:2:1. In programming, this method has several possible implementations; more information on the method of skeletal can be found below in the references section.

Neural Network Structures

This area was very popular in the ' 60s and ' 70s, in consequence of interest to them slightly diminished, since a considerable number of neurons requires considerable computing power that are not normally present at a simple mobile platforms. However, we must bear in mind that neural networks are sometimes give very interesting results, due to its nonlinear patterns, moreover some of the neural network is able to recognize the images are invariant under the rotation of a preprocessing. For example, network-based neocognitron able to allocate some characteristic features of the images, and to recognize them as if the images have not been tapped.

Invariant Numbers

The geometry of the image it is possible to allocate some numbers that are invariant to the size and rotation of images, then you can make the table of conformity of these numbers to a specific way (almost as in algorithm of skeletal). Examples of invariant numbers - Euler number, eccentricity, orientation (in the sense of the location of the principal axis of inertia about something too invariant).

Point Wise Percentage Comparison with the Standard

There must be some pretreatment to obtain invariance relative size and position, then a comparison is made with the prepared base of standards images - if a match more than a mark, then believe the image is detected.

There are a number of significant problems associated with the recognition of handwritten and printed characters. The most important of them are the following:

- the diversity of the glyph;
- distortion of images;
- Variations of size and scale of the characters.

Each individual character can be written in different default fonts, for example: (Gothic, Elite, Courier, Orator), special fonts used in OCR systems, as well as many non-standard fonts. In addition, different symbols can have similar shapes. For example, 'U and 'V, 'S' and '5', 'Z' and '2', 'G' and ‘6’

Distortion of digital images of characters can be of the following types:

Distortion: the disruption of rows, not impregnated characters, isolation of separate points, non-planar nature of the information media (for example, the effect of distortion), offset characters or parts thereof relative to the location in the string; rotation with the change in the slope of characters; gross increment the digitization of images;

Besides, it is necessary to allocate radiometric distortion: defects of lighting, shadows, reflections, uneven background error when scanning or when shooting a video camera.

Significant is the effect of the original scale printing. In the accepted terminology scale of 10, 12 or 17 means that in inches placed 10, 12 or 17 characters. Thus, for example, character 10 scales usually larger and wider than the character of scale 12.

In addition to these problems, an optical character recognition (OCR), you must select the image text areas, to select individual characters, to recognize these characters and must be insensitive to the way of printing (layout), and the distance between rows.

As a rule, OCR system consists of several units, involving hardware or software implementation:

- optical scanner;
- localization block and highlight the elements of the text;
- images preprocessing block;
- signs allocation block;
- recognition block;
- Recognition results post-processing block.

As a result of work of an optical scanner source text is entered into the computer as grayscale or binary image.

In order to save memory and reduce expenses of time for information processing, OCR systems typically applied converting a grayscale image to black and white. This operation is called a binarization. However, it should be borne in mind that the operation of the binarization can lead to a loss of efficiency of recognition.

Software in OCR systems is responsible for presenting data in digital form and split coherent texts on individual characters.

After splitting the characters presented in the form of binary matrices, are smoothing, filtering to eliminate noise, the normalization of size, as well as other reforms aimed at highlighting the characteristics used subsequently for recognition.

Character recognition is the process of comparing the selected characteristic features with reference signs selected in the statistical analysis of the results obtained in the process of training system.

Thus, the meaning or context information can be used for the resolution of uncertainties that arise in recognition of symbols, having identical sizes, and for the adjustment of words and phrases in general.

For the classification of characters, you must first create library of standard eigenvectors. To do this, at the stage of training, the operator or the developer enters into the system OCR large number of samples of the glyph. For each sample, the system allocates signs and stores them in a corresponding vector of characteristics. A set of vectors of characteristics that describe the character class is called a cluster.

During operation of the system OCR may need to expand the knowledge base. To realize this purpose, some systems have the ability to study in real time. However, the learning process requires human intervention and time-consuming, though research is aimed at automation of the process of learning, which in future will allow minimizing the participation of the human operator. The task of classification is the definition of the class that owns the attribute vector obtained for a given character.

Classification algorithms based on the determination of the degree of closeness of a set of signs character to be examined each of the classes. The credibility of the result depends on the metric space of characteristics. One of the most important metrics is the Euclidean distance:

$$D_j^E = \sqrt{\sum_{i=1}^{N} (F_{ij}^L - F_i^L)^2} ,$$

Where F_{ij}^L - i-th sign of j-th standard (etalon) vector; F_i^L - i-th sign of testing image symbol.

One technique that can improve similarity metric based on the use of the genetic algorithm. To solve the set of problems and a search for new optimization algorithms. Proposed quite recently - in 1975 – by John Holland genetic algorithms (GA) is based on the principles of natural selection including Darwin. GA belongs to stochastic methods. These algorithms have been successfully used in various fields of activity (Economics, physics, engineering sciences, and so on). Created various modifications HA and developed a number of test functions. Considered how work GA, and what issues remain unresolved - the purpose of this work.

Genetic algorithms belong to the field of soft computing. The term "soft computing" was introduced by Lotfi Zadeh in 1994 [1]. This concept unites areas such as fuzzy logic, neural networks, probabilistic reasoning, trust networks and evolutionary algorithms, which complement each other and are used in different combinations or separately for the development of hybrid intelligent systems. The first scheme of genetic algorithm was proposed in 1975 at the University of Michigan's John Holland (John Holland) [2], and the prerequisites for this was works of Ch.Darwin [3] (theory of evolution) and research L. J.Fogel, A.J. Owens, M. j. Volsh [4] on the evolution of the simple machines, foretelling the characters in the numeric sequence (1966). A new algorithm called "reproductive plan of Holland" and later actively used as a basic algorithm in evolutionary computation. The idea of Holland has developed his disciples Kenneth De Jong (Kenneth De Jong) from George Mason University (Virginia), [5] and David Goldberg (David E. Goldberg) from the laboratory GA of Illinois [6]. Thanks to them, was created classic HA described all operators and studied the behavior of a group of test functions (which algorithm Goldberg and got the name of "genetic algorithm"). Genetic algorithms are adaptive search methods, which are used to solve optimization problems. They use as an analogue of the mechanism of genetic inheritance and similar natural selection. This preserves the biological terminology in a simplified form and the basic concepts of linear algebra.

In the classification process to more reliable signs given higher priority

$$D_j^E = \sqrt{\sum_{i=1}^{N} w_i (F_{ij}^L - F_i^L)^2} \to \min$$

Where w_i, - the weight of the i-th sign.

Reformulate optimization task as the task of finding the minimum of a function D_j^E that calls fitness function. It must take nonnegative values on a limited scope (so that we could for each individual to assume its fitness, which cannot be negative), however, did not require the continuity and differentiability.

Each parameter is the function of adaptation is encoded by a string of bits.

Individual will be called a string that is the concatenation of strings ordered set of parameters:

1010 10110 101 … 10101

| F1 | F2 | F3 | … | FN |

Step of the Algorithm Consists of Three Phases

- Generation of intermediate populations (intermediate generation) through selection of the current generation.
- Recombination of individuals of intermediate populations by *crossover* (crossover), which leads to the formation of a new generation.

Pre-defined value P_K – the probability of crossing-over and put the box FG two-state "run", "not to do". Original state FG "not to do". When performing crossover consistently reviewed the loci selected pairs of chromosomes. With probability P_k "the box" FG state is run. If FG passed into the state "run", that is the exchange of genes between a pair of chromosomes in the current locus, hereinafter the "flag" has a status of "not run", and then it moves to the next locus.

- Mutation of the new generation.

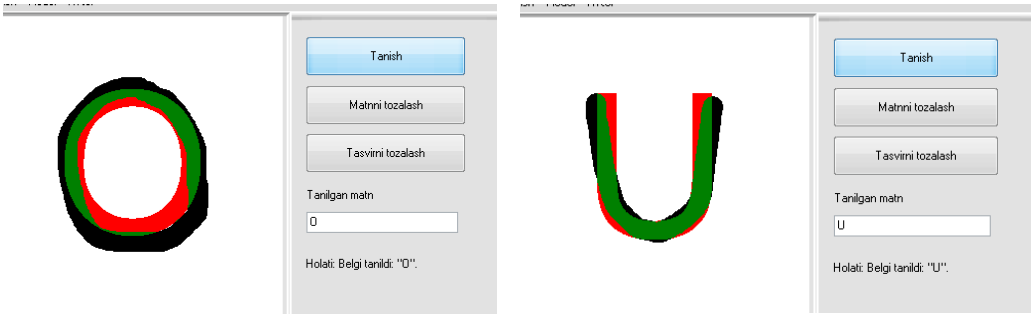
Mutation algorithm is implemented as follows.

Consistently selected chromosomes from the current population. Within the selected chromosomes consistently viewed genes. After navigating to another gene, FG with probability P_M state is run. If FG passed into the state "run", then randomly gene g_n takes one of the values in a given range, with the exception of the value gene has before mutation. Next FG enters the state of "not run" and selects the next gene, chromosome, or the next chromosome.

Such process of evolution, generally speaking, can continue indefinitely. Criterion stop can serve as a set number of generations or *convergence* of population.

Convergence is called the state of the population, when all the strings of the population are in the area of some extreme and almost the same. That is a crossover practically does not change the population, and mutating individuals tend to die out, as less adapted. Thus, the convergence of the population means that reached the decision is close to optimal.

Was written program which partially implements the basic units of automatic recognition of letters



Tanilgan matn: O

Holati: Belgi tanildi: "O".

Tanilgan matn: U

Holati: Belgi tanildi: "U".

Figure 1: Recognition of Handwritten Letters

CONCLUSIONS

Thus, at present, the development and application of genetic algorithms is intensively developing direction. Thanks to the versatility of the computational scheme, parallel implementation of stability to noise, genetic algorithms are successful practical application while solving complex nonlinear multidimensional optimization tasks. Unlike traditional methods of multi-criteria optimization, many of which are often characterized by a sharp increase computational costs as the number of variable parameters, genetic algorithms have proven themselves in the large-scale tasks. Currently, they are widely used for a wide range of practical tasks, and their use is constantly expanding.

REFERENCES

1. Zadeh L. A. a new approach to the analysis complex systems and decision processes//Math today. –M.: Knowledge. 1974. –pp.5-49.
2. Holland J. H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence [Text]/J. H. Holland. — The MIT Press, Cambridge, 1992. — ISBN 0262581116.
3. Darwin CH., On the origin of species by means of natural selection, or the preservation of favored races in the struggle for life [Text]/ C. Darwin. — M.: an SSSR, 1939. — Vol.3.
4. Barseghyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I. I. Technology for data analysis: Data Mining,

Visual Mining, Text Mining, OLAP. Saint Petersburg, Publishing office BHV-Petersburg, 2007.

5. Association on Genetic Algorithms, George Mason University. — http://www.cs.gmu.edu/research/gag

6. Goldberg D. Genetic Algorithms in Search, Optimization, and Machine Learning [Текст]/ D. Goldberg. — Massachusetts: Addison-Wesley, 1989. — ISBN 0201157675.