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, 51-60

METHODS OF IMAGE PREPROCESSING OF TEXT CHARACTERS

MUHAMEDIYEVA D.T[1], ABDURAIMOV D[2] & NIYOZMATOVA N.A[3]

[1, 3]Center for Development hardware and software products under TUIT, Uzbekistan

[2]Guliston state universities, Uzbekistan

ABSTRACT

Discusses methods of image preprocessing of text characters, as well as creating recognition of words of a natural language on the basis of calculation of the correlation coefficient using neural networks

KEYWORDS: Test Sample, Character Recognition, Hidden Layer, Binary Image

STATEMENT OF THE TASK

The objective of this work is the study of methods, as well as creating recognition of words of a natural language on the basis of calculation of the correlation coefficient using neural networks. To accomplish it was chosen a neural network of error back propagation, which will be implemented in the programming language in the form of a class. The class will contain a constructor, which will be submitted, the following settings:

- training sample
- test sample;
- the number of neurons of the hidden layer;
- coefficient training;
- The number of training epochs**.**

For translation training and test words in a binary form and also save the words to the exception in the hash table will be provided for separate classes.

We will also create a class that produces splitting text into individual tokens and their recognition by means of neural networks and hash tables.

Accordingly, to solve this problem it is necessary to create four classes:

- class of neural network and means of processing;
- class of text translation in a binary form;
- class hash table and methods of work

The generated classes are decorated as a library component.

To test the generated class will create a Windows application using visual components. The tester program should carry out input of text information from a file and from the keyboard, and the input parameters of the neural network: the

number of training epochs, number of neurons in the hidden layer, the time factor and learning.

TYPICAL PROBLEMS ASSOCIATED WITH THE RECOGNITION TEXT DATA

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

- variety of forms of character shape;
- distortion of images;
- Variations in the size and scale of the characters.

Each individual symbol can be written in a variety of standard fonts, such as Gothic, Elite, Courier, Orator, special fonts used in the OCR systems, and also many non-standard fonts. In addition, different characters can have similar contours. For example, 'U and 'V, 'S' and '5', 'Z' and '2', 'G' and ‘6’

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

Distortion: the fragmentation of lines, nepropitannoy characters, the isolation of individual points, the non-planar character of the information carrier (for example, the effect of warping), the offsets of the symbols or their parts relative to the location in the string; rotating with the change of inclination of the characters; rough discrete digitization of the images;

In addition, it is necessary to allocate radiometric distortion: the defects of lighting, shadows, and glare, and uneven background, errors in the scanning or shooting a video camera.

Important also is the impact of the initial scale printing. In the accepted terminology of the scale 10, 12 or 17 means that the inch lines are placed 10, 12 or 17 characters. Thus, for example, the symbols of the scale 10 are usually larger and wider character of the scale 12.

In addition to these problems, optical character recognition (OCR), you must allocate the image text field, to select individual symbols, to recognize these symbols and to be insensitive to the method of printing (DTP) and distance between rows.

The Structure of Systems of Optical Recognition of Texts

Typically, OCR systems consist of several blocks, involving hardware or software implementation:

- optical scanner;
- block localization and selection of text elements;
- block preprocessing of the images;
- block feature extraction;
- unit recognition;
- Block post processing the recognition results.

The result of the optical scanner of the source text is entered into the computer in the form of gray scale or binary image.

In order to save memory and reduce time spent on information processing, OCR systems, usually used for

converting grayscale image to black and white. Such an operation is called binarization. However, you must keep in mind that the operation of binarization can lead to the deterioration of the efficiency of recognition.

Software systems OCR are responsible for representing data in digital form and partitioning of coherent text into individual characters.

After splitting the characters represented as binary matrices, are subjected to a smoothing filter to eliminate noise, normalize size, as well as other reforms with the aim of feature extraction, to be used later for recognition.

Character recognition occurs in the comparison process highlighted the characteristic features with the reference features selected through statistical analysis of the results obtained in the learning process of the system.

Thus, semantic or context information can be used to resolve uncertainties arising from the recognition of the symbols with identical dimensions, and corrections of words and phrases in General.

METHODS OF IMAGE PREPROCESSING THE TEXT CHARACTERS

Preprocessing is an important step in the process of pattern recognition and allows for the smoothing, normalization, segmentation and approximation of line segments.

Smoothing consists of lling and thinning. Filling eliminates small breaks and gaps.

Thinning is the process of reducing the thickness of the line in which multiple pixels are associated with only one pixel. Known serial, parallel and hybrid algorithms of thinning. The most common methods of thinning based on iterative erosion of contours, in which the window (3x3) moves the image inside the window and performs the corresponding operation. After the completion of each stage, all selected points are deleted.

Skeletization is a reconstruction of the trajectory of the writing tool as input information which is used or the skeleton image, or contour image. Despite the greater efficiency of the use of "carcass" restoration trajectory on the skeleton image has a significant drawback: critical information for reconstruction of the trajectory is the shape of the boundary kennel in the vicinity of the areas of intersection of the strokes. From the skeleton image this information is not available (see Fig. 1). The disadvantages of the "carcass" can also include greater sensitivity to noise at the boundary of the image and, as a consequence, the occurrence of random "spikes" in the skeleton image. Therefore, methods of recovery trajectory, based on the receipt of the skeleton, leading to errors when processing images with a fairly complex trajectory.

An alternative to the "carcass" is the method of preprocessing, which consists of the following: a character image is divided into stripes of black dots corresponding to disjoint intervals of the strokes (regular region) and the region of intersection of the lines (nodal region).

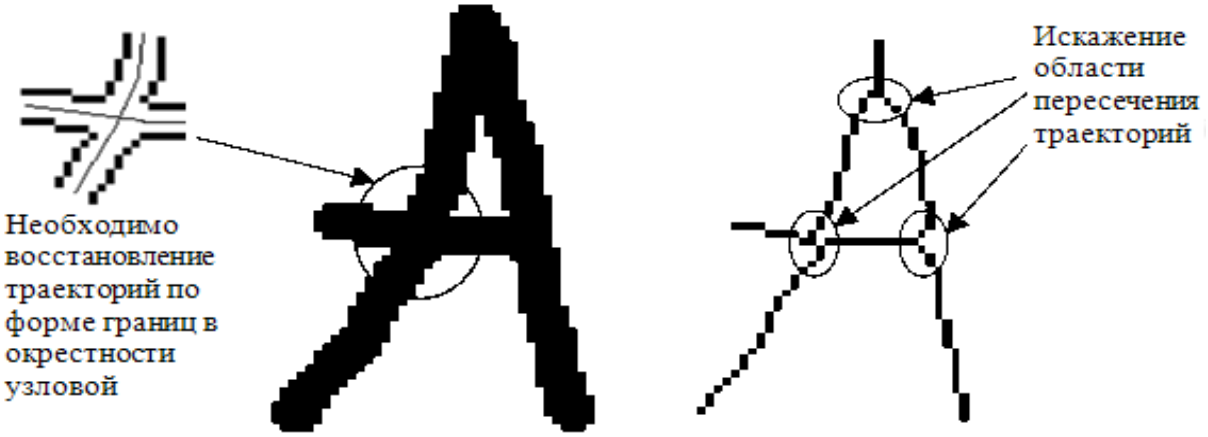

Figure 1: Image of the Character and His Skeleton

Existing algorithms for detection of regular fields based on scatter imaging. In this paper, we propose an algorithm based only on using the polyline approximating an outline of an image.

The main features of the proposed method are:

- high speed image processing, which is achieved due to the absence of spot treatment;
- Screening of images for which the method of allocation of the trajectories is unacceptable (images with spots and swims) at the stage of preprocessing.

Normalization consists of algorithms, eliminating the distortions of the individual characters and words, and also includes procedures that the normalization of characters in height and width after appropriate processing.

Segmentation provides a partition image in a separate region. As a rule, you first need to clear the text from graphics and handwritten annotations, as these methods allow handling not only noisy text. Purified from various marks, the text could be segmented.

Most of optical character recognition algorithms divide the text into characters and recognize them individually.

This simple solution is really effective only if text characters do not overlap each other. Merge symbols can be caused by font type, which was the text, poor resolution printer or high level of brightness selected to restore broken characters.

Splitting text into words is possible, if the word is a wealthy symptom, in which segmentation is performed. This approach is difficult to implement due to the large number of elements to be recognized, but it can be useful if the set of words in the code dictionary is limited by the condition of the problem.

Under the approximation of line segments understand the compilation graph description of the character in the form of a set of vertices and straight edges that directly approximiert chain of pixels of the original image. This approximation is performed in order to reduce the amount of data and can be used for recognition based on selection of features describing the geometry and topology of the image.

IMPLEMENTATION OF THE RECOGNITION ALGORITHM OF THE TEXT DATA

In the study, a program was written to partially implements the main blocks of the automatic recognition of letters.

Figure 2 shows the block diagram of the program executed.

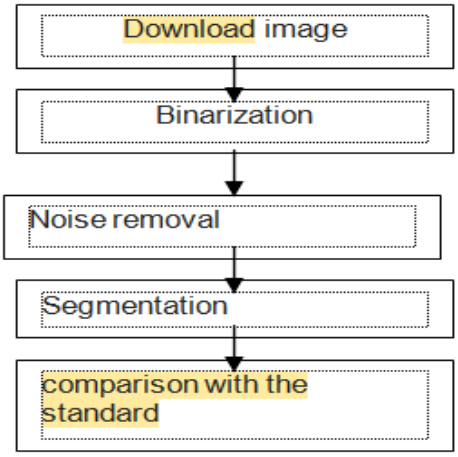

Figure 2: Block Diagram of the Program

Recognition Process

Download of the image. To start we need an image, which then can work. Downloadable its standard methods of Windows.

This program recognizes only binary images, so the second stage after receiving the pictures, she binarized. When working with a color image conversion from color to black-and-white image is by the standard formula

*Y: =0.3*R+0.59*G+0.11*B.*

Further, the Algorithm is Quite Simple

There is some plank, if the color of the shade of grey above it is considered white, if lower - it is considered black. As can be seen, the binarization is very simple, but serious quality improvement recognition, and reduction of time of work of the subsequent modules, at this point, it is better to incorporate some sort of filter, even the most simple. This program does not use any filter, but a place where he can be included is indicated.

Often the resulting image is replete with interference, have no relation to symbols, and only hinder the recognition process. Using the simplest method, which considers the density of outlets in a given area, it is possible to dispose a larger amount of interference.

Split image into pieces, each of which contains its own unique object is called segmentation.

This program is implemented pixel-by-pixel comparison with the reference characters for the specified font. If the percentage of coincidence of the reference and the selected symbol is not below a set limit, the character is considered recognized.

Testing Program

Open the image (Figure 3). Fits any image in BMP format.

Figure 3: Open the Image

For recognition, you should select the menu item "Recognize". After this binarization, removal of noise (Figure 4) and begins the process of character recognition (Figure 5).

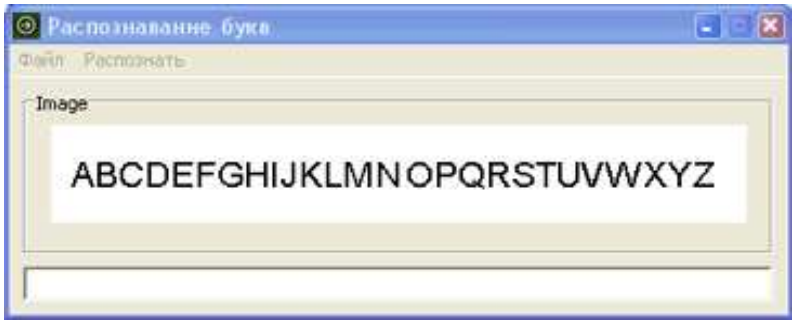

Figure 4: Binarization and Noise Removal

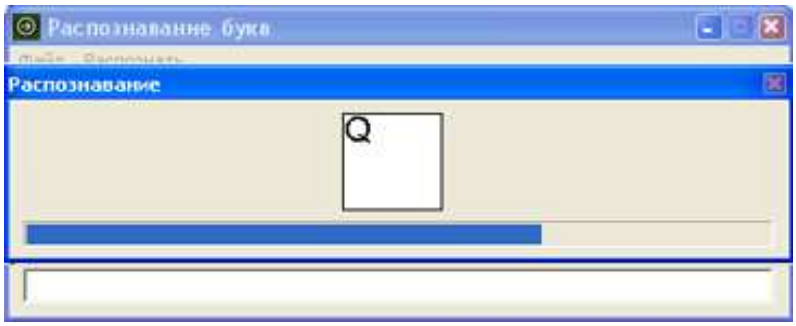

Figure 5: Character Recognition

Recognized characters after the end of program operation are displayed in the line under the figure (Figure 6).

Figure 6: Recognized Characters

OBJECT RECOGNITION ON THE BASIS OF CALCULATION OF THE CORRELATION COEFFICIENT

Consider the method of recognition of objects in the image based on the use of the calculation of the correlation coefficient. In this case, for the solution of such problems is necessary in addition to the original image to have an image of

the object to be detected on the original image.

Choose as the source image containing a collection of letters.

Figure 7: The Original Image 100x300

We also need to have reference images of objects (characters) that you want to recognize.

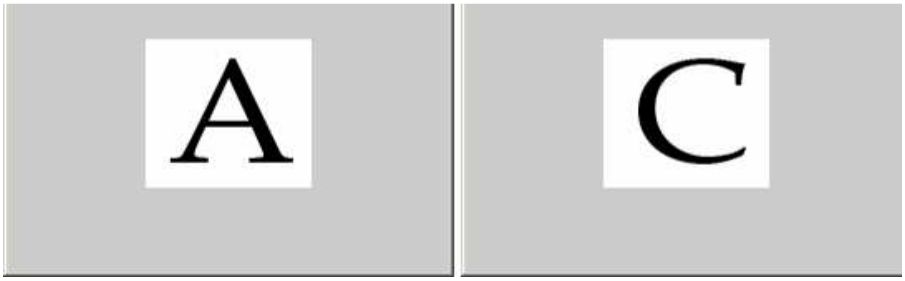

Figure 8: The Reference Image

The next step is to calculate the correlation coefficient between the matrices of the original image and the corresponding reference.

The correlation coefficient is calculated according to the formula:

$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\sum_m \sum_n (A_{mn} - \overline{A})^2 \sum_m \sum_n (B_{mn} - \overline{B})^2}},$$

Where $\overline{A} = mean2(A), \overline{B} = mean2(B)$

It then computes the correlation coefficient between each part of the input image and each reference.

Figure 9: The Selection of Part of the Image

The values of the correlation coefficients for the first pattern (the letter a) and three parts of the original image is presented on the chart

The location of the maximum correlation coefficient suggests that this part of the source image most similar to the benchmark.

Define the location of the maximum value of the correlation coefficient. As the figure shows, the maximum correlation value obtained between the matrix of the first standard (the letter a) and the first part of the input image, where he posted the letter A. similarly calculated correlation coefficients between the other references(second and third) and the original image.

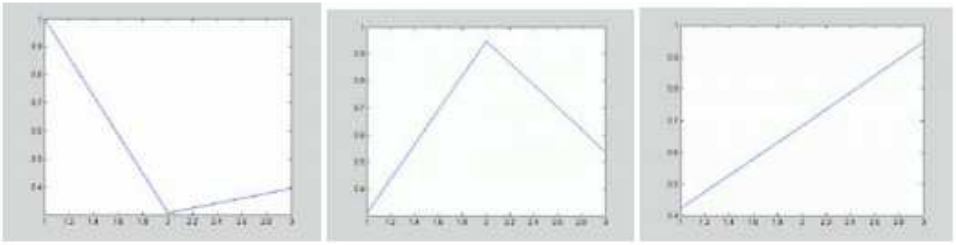

Figure 10: Plot of Values of Correlation Coefficients for the First -, Second – b and Third – With Standard and Three Parts of the Original Image

From the above two figures it is seen that the maximum values of the correlation coefficients reach in the second and third parts of the image. This corresponds to the highest similarity, i.e. the detection of the second Etalon and a second portion of the source image and the third reference and the third part of the original image.

Thus, we have considered an approach to object recognition based on computing the correlation coefficient between the matrices of their images. This approach has become rather widespread. However, recognition of real objects correlation method is characterized by high computational complexity. This is due to scaling and rotations of the detected image.

After successful completion of the segmentation, each segment falls into the recognition module. For what would the images to be recognized is invariant under position and rotation need to be attached to their structure.

Each binary image we can calculate a few characteristics that are not dependent on its rotation or size. Describe the application of neural networks (NN) for image recognition.

APPLICATION OF NEURAL NETWORKS (NN) FOR IMAGE RECOGNITION

NN consists of elements, called formal neurons, which themselves are very simple and are linked to other neurons. Each neuron converts the set of signals supplied thereto at the input to the output signal. It is the ties between neurons, encode weights, play a key role. One of the advantages of the national Assembly (and the drawback to their implementation on a serial architecture) is that all elements can operate in parallel, thereby significantly improving the efficiency of solving the tasks, especially in image processing. Except that NA can effectively solve many tasks, they provide a powerful flexible and universal mechanisms of learning, which is their main advantage over other methods [1,2] (probabilistic methods, linear separators, decisive trees, etc.). Training eliminates the need to select key features, their

importance and relationships between characteristics. But, nevertheless, the original representation of the input data (a vector in n-dimensional space of frequency features, wavelets, etc.), significantly affects the quality of decisions and is a separate topic. NA have good generalizing ability (better than the decisive tree [2]), i.e. can successfully disseminate the experience gained on the target training set, the whole set of images.

CONCLUSIONS

Thus, the analysis of recognition algorithms showed that the invariance to rotation of the neural networks takes a long time to show the deviation is necessary to create a table of invariant numbers for images. The proposed work recognizes only binary images, so the second stage after receiving the pictures, she binarized.

REFERENCES

1. Petrou M. Learning in Pattern Recognition. Lecture Notes in Artificial Intelligence – Machine Learning and Data Mining in Pattern Recognition, 1999, pp. 1-12.
2. Jacobsen X., Zscherpel U. and Perner P. A Comparison between Neural Networks and Decision Trees. Lecture Notes in Artificial Intelligence – Machine Learning and Data Mining in Pattern Recognition, 1999, pp. 144-158.