BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X,
Vol. 4, Issue 7, Jul 2016, 27-36

DEVELOPMENT OF SOFTWARE COMPLEXES OF TEXT DATA RECOGNITION

MUHAMEDIYEVA D. T[1], ABDURAIMOV D[2] & PRIMOVA KH.A[3]

[1,3]Center for Development Hardware and Software Products under TUIT, Uzbekistan

[2]Gulistan State University, Guliston, Uzbekistan

ABSTRACT

The technique of development of software complexes of recognition based on neural networks. It is shown that the training eliminates the need to select key features, their importance and relationships between characteristics. But, nevertheless, the original representation of the input data greatly affects the quality of solutions. The neural network has good generalizing ability, i.e. can successfully disseminate the experience gained on the target training set, the whole set of images

KEYWORDS: Text Recognition, Printed Symbols, A Neural Network Algorithm Back-Propagation of Error, Training

INTRODUCTION

In the case when it comes to the recognition of printed characters should be mentioned that an almost infinite variety of printed products manufactured using a limited set of original symbols which are grouped by style (set of art-making) that distinguishes this group from others. One group, which includes all alphabetical characters, numbers and a standard set of special characters, called headset. However, widely spread other term font, this term will be used in the future. Any printed text is the primary property - the fonts, which he printed. From this point of view there are two classes of algorithms for the recognition of printed characters and without font (omnifont). Font or settotaltime algorithms use a priori information about the font printed letters. This means that OCR must be presented to the full sample of text printed in that font.

ANALYSIS OF ALGORITHMS FOR OPTICAL CHARACTER RECOGNITION

Program measures and analyzes various characteristics of the font and stores them in its database reference characteristics. At the end of this process, font program optical character recognition (OCR) is ready to recognize this particular font. This process can be called a training program. Further training occurs for a particular set of fonts, which depends on the area of application of the program.

- **Disadvantages of this Approach Should Include the Following Factors:** The algorithm needs to know in advance the font that you are introduced to him for recognition, i.e. it needs to store in a database the various features of this font. Recognition quality of text printed in an arbitrary font is directly proportional to the correlation of the characteristics of this font with the fonts available in the database of the program. Given the existing wealth of printed materials in the learning process it is impossible to cover all fonts and their modifications. For example, in modern computer systems, layout of documents uses more than 100 fonts. In other words, this factor limits the versatility of such algorithms.

- For program recognition, you need the power settings on a specific font. Obviously, this unit will pay its share of errors in the integral assessment of the quality of recognition or the font installer feature will have to assign to the user.
- The program is based on the font algorithm for character recognition, require the user any special knowledge about fonts in General, their groups and their differences from each other, the script, which printed the document of the user. Note that if a paper document is not created by the user, and came to it from the outside, there is no regular way of knowing what fonts the document was printed. Factor need special knowledge narrows the range of potential users and shifts it in the direction of the organizations employing the relevant specialists.

On the other hand, the type approach has the advantage, thanks to which it is actively used and likely to use in the future. Namely, having detailed a priori information about the symbols, you can build very accurate and reliable recognition algorithms. In General, when building a font recognition algorithm the reliability of the recognition of the symbol is intuitively clear and mathematically expressible value. This value is defined as the distance in any metric space of the reference symbol that is presented to the program in the process of learning the character, which the program tries to recognize.

Second class of algorithms – without font or font, i.e. algorithms that do not have a priori knowledge about the symbols available to them at the entrance. These algorithms measure and analyze various characteristics (features) inherent in the letters as such without regard to absolute font size (point size), which they printed. In the limiting case font for the algorithm, the learning process may be missing. In this case, the characteristics of the characters measures, encodes and puts into the database program itself. In practice, however, cases when such a path is exhaustively solves the problem, are rare. A more generic way of creating a database of characteristics is the training program on a sample of real characters.

The Disadvantages of this Approach Include the Following Factors:

- **Achievable quality of recognition is lower than the font of the algorithms:** This is due to the fact that the level of aggregation when measuring the characteristics of the characters much higher than in the case of font dependent algorithms. Practically, this means that various tolerances and coarsening measurements of the characteristics of the characters to work without font algorithms can be 2-20 times larger compared to the font.
- Should be considered lucky if without font algorithm has adequately and physically proved, i.e., naturally arising from the basic procedure of the algorithm, a coefficient of reliability of recognition. Often have to put up with the fact that accuracy assessment is either missing or is artificial. Under artificial evaluation refers to the fact that it is significantly not the same as probability of correct recognition, which provides this algorithm.

The Advantages of this Approach are closely linked to his Weaknesses. Main Advantages Are as Follows:

- **Versatility:** This means on the one hand the applicability of this approach in cases where the potential variety of characters that can be sent to the input of the system is high. On the other hand, at the expense of the inherent ability to generalize these algorithms can extrapolate the knowledge beyond the training sample, i.e., stable to recognize the characters by sight distant from those present in the training set.
- **Adaptability:** The process of font learning algorithms is usually easier and more integrated in the sense that the training sample is not fragmented into different classes (fonts, pins, etc.). There is no need to maintain our

database of characteristics of various conditions of existence of these classes (nekorrelirovannogo, not Miscibility, unique naming, etc.). A manifestation of adaptability is also the fact that it is often possible to create almost fully automated training procedure.

- **Ease in Use of the Program:** If the program is built on font-algorithms, the user is not required to know anything about the page that he wants to enter in the computer memory and notify about this knowledge program. Also simplified the user interface due to the lack of options and dialogs that serve the training and handling characteristics. In this case, the recognition process can be presented to the user as a “black box” (in this case, the user is completely unable to manage or in any way modify the course of the recognition process). In the end it leads to the expansion of the range of potential users due to the inclusion of people with minimal computer skills.

Software Implementation of Neural Network of Error Back Propagation

Algorithm was implemented as a class Neuro Network. To describe the neuron structure has been established neuron type containing the following fields:

List of weights for the connections between this neuron and all neurons of the previous layer (or input, if the neuron is in the input layer). Each weighting factor is a real number (in the 1 weight coefficient of the neuron of the previous layer);

–Threshold level;

– Error value is used only on stage of learning;

– Change of error, and is used only during training.

Output signal of the neuron is stored in a field (the so-called activity of the neuron). The neuron should respond to input of suspended relations, calculating the output signal. To transform the input signals to weekend necessary function.

Activation function $x_i^{[n+1]} = f\left(a_i^{[n]}\right)$ was implemented in a public method sigmoid. In this method, use the following formula $f(x) = \frac{1}{1+e^{-\alpha x}}$.

Biological neurons do not work (do not issue an output signal as long as the input signal level reaches a threshold value, i.e. the input neuron receives the weighted sum signal minus some value. The resulting value passes through an activation function.

The activity of each neuron of the previous layer multiplied by the appropriate weighting factor, the results of the multiplication are summed $a_i^{[n]} = \sum_j w_{ij}^{[n]} x_j^{[n]}$, subtracted a threshold value, calculate the value of a sigmoid function $x_i^{[n+1]} = f\left(\sum_j w_{ij}^{[n]} x_j^{[n]}\right)$.

Calculation of the sum of weighted signals for the hidden and output layers (methods run_hidden_layer and run_output_layer) is similar.

Each layer of neurons is based on the output of the previous layer (except the input layer based directly on the

requirements of the network input (in the code - pattern test_pat). This means that the values of the input layer needs to be completely calculated before calculating the values of the hidden layer, which in turn must be calculated to calculate the values of the output layer.

Outputs of the neural network values of activity (a field) of neurons in the output layer. Program that simulates the work of a neural network, the learning process will compare them with the values that should be at the output of the network.

The Complete Learning Algorithm, the NA using the Procedure of Back Propagation is as Follows:

- To initialize the thresholds and weights are small random values (not more than 0.4). Initialization of the weighting factors of random real values using the Random class is produced as a function random_weights.
- To apply to the inputs of a network one of possible images and in the normal functioning of the national Assembly, when signals propagate from inputs to outputs to calculate values of the latter. Method run_the_network.
- Calculate error for output layer (calculate_output_layer_errors). In this case use the formula $E_i=(t_i-a_i).a_i.(1-a_i)$ (here E_i - error for the i-th node of the output layer, ai is the activity of this node, t_i - required for the activity. the desired output value for each i-th values of the output layer. Below are the corresponding function:

The calculation of the error for the hidden and input layers is performed by methods calculate_input_layer_errors and calculate_hidden_layer_errors by the formula $E_i = a_i. (1 - a_i). \sum_j E_j.w_{ij}$.

Using the formulas get feature, learning the weights and threshold levels.

Next, combine the above functions in a single method back_propogate().

For clearing the previous values, use the function blank_changes.

To simplify the time complexity of the network obtained weights will be recorded in separate files (method AddWeightsToFile()), the names of which are given automatically by the program To read the saved settings will be applied the method ExtractWeights().

Class of Translate Text to Binary form designed for the binarization of the original data, i.e. the translation of words from natural language into a set of ones and zeros. This procedure is necessary because the neural network is error back-propagation works only with binary data.

The constructor parameter is an array of strings for translation in a binary form.

Algorithm of the methods GetBinarizeWord and GetBinarizeText this class in General consists of the following steps:

– Encoding of words: summation of the ASCII codes of the letters (i+4 where i is the number of letters in the word;

– Convert the resulting decimal numbers in binary form using the method DecToBin;

– Processing of the received data.

Recognition class is the main. His task is recognition. It uses objects of all the above classes.

The constructor parameters are:

– text for analysis – sText;

– parameters of the neural network N_HID, beta, m, Epoch;

– indicator of the need of training the neural network – flag.

Of all of the above methods, the most important are: the constructor, Scaning and GetNeuroResult. The analyzed text is first supplied to the constructor. There it is broken down into individual tokens whose type is determined either immediately (if it is punctuation mark or name) or by the method of Scaning. Scaning method logically can be divided into 3 blocks: – word search in the hash table;

– Identification of a word at the end and determine its type; – identification of a word at the end and an in-depth analysis using a neural network.

- **Computational Experiment**

  During solving the problem of recognizing printed characters can be divided into two stages:

  - Localization of the symbol in the image.
  - Character recognition on the sign.

  To simplify further calculations and software implementation converted the original color images to grayscale.

Figure 1: Converting Image to Grayscale

Now, as in most other tasks, it is necessary to pre-process raw images. This stage is especially important when solving problems of recognition. The quality of the decisions of the pre-processing stage will depend largely on the effectiveness of the recognition problem in General.

The most common defects that may be present on the source images are noise and low contrast level.

For elimination of pulse emissions using median filtering.

The treatment result is presented in the picture below.

Figure 2: The Image with Impulse Noise and The Image after Eliminating Impulse Noise

Filter that increases the sharpness of the image has a mask defined by the following expression:

$$h = \frac{1}{a+1}\begin{bmatrix} -a & a-1 & -a \\ a-1 & a+5 & a-1 \\ -a & a-1 & -a \end{bmatrix},$$

Where the parameter a is chosen in the range from 0 to 1.

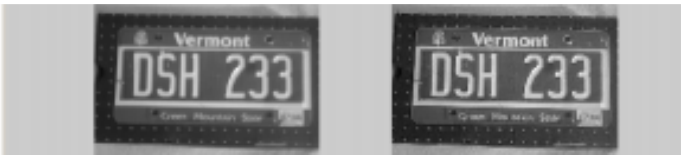

Figure 3: Eliminating Image Blur

In addition, to improve the quality of image you can use the histogram.

The histogram improves the image contrast by mapping pixel values of the original image so that the histogram of brightness of pixels of the resulting image approximately corresponded to some predefined histogram. This function is designed to convert grayscale or palette images.

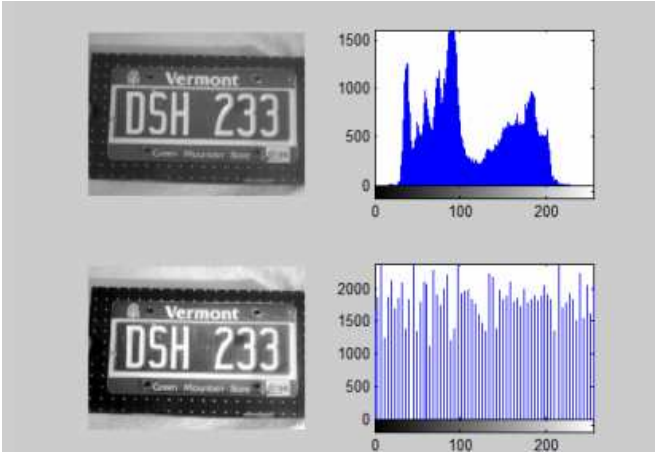

Figure 4: Histogram Equalization

Determining the location of the printed characters in the image After pre-processing of the image to determine the location of printed characters. On the image are searched for the so-called associated region of the pixels and generates a matrix L, where each element is equal to the number of the object that owns the corresponding pixel in the image. The size of the matrix units and L equals the image size. Objects are numbered in order starting from 1. Items that have a value of 1, belong to the first object, having a value of 2 are assigned to the second object, etc. If the element in matrix L is equal to 0, it means that the corresponding pixel of the original image belongs to the background. The parameter n specifies the criterion of connectedness used to locate connected regions objects. The parameter n can take the values 4 or 8 (the default value).

Figure 5: The Result of the Program

Next step is calculation of the characteristics of found objects. The elements of the matrix L, having a value of 1 belong to the first object, having a value of 2 are assigned to the second object, etc. If the element in matrix L is equal to 0, then it belongs to the background. Let N be the number of pixels belonging to the object. The set of all pixels p(x, y) belonging to the object, denoted by Q. Then the coordinates of the center of mass of the object is calculated as

$$x_c = \frac{1}{N} \sum_{p(x,y)\in\Omega} x \quad y_c = \frac{1}{N} \sum_{p(x,y)\in\Omega} y$$

Then we can calculate some auxiliary quantities:

$$U_x = \frac{1}{12} + \frac{1}{N} \sum_{p(x,y)\in\Omega} (x - x_c)^2$$

$$U_y = \frac{1}{12} + \frac{1}{N} \sum_{p(x,y)\in\Omega} (\mathrm{y} - y_c)^2$$

$$c = \sqrt{(U_x - U_y)^2 + 4 \cdot U_{xy}^2}$$

Then the length of the maximum A_{max} and the minimum A_{min} of the axes of inertia are calculated as:

$$A_{\min} = 2\sqrt{2} \cdot \sqrt{U_x + U_y - C}$$

The lengths of the principal axes of inertia are used to calculate the eccentricity and orientation of the object. The eccentricity is determined by the ratio

$$E = \frac{2 \cdot \sqrt{(0.5 \cdot A_{\max})^2 - (0.5 \cdot A_{\min})^2}}{A_{\max}}$$

Orientation is defined as the angle in degrees between the maximum axis of inertia and the axis of X. If, $U_y > U_x$ the orientation O is calculated using the formula

$$O = \frac{180}{\pi} \cdot arctg\left(\frac{U_y - U_x + C}{2 \cdot U_{xy}}\right)$$

otherwise, computed as

$$O = \frac{180}{\pi} \cdot arctg\left(\frac{2 \cdot U_{xy}}{U_y - U_x + C}\right)$$

Figure 6: The Numbers Mark the Detected Objects

We present the results of calculations of characteristics for all objects in the image. As the parameter was chosen as the fill factor, which is the ratio of the object area to the area of the bounding rectangle.

Character Recognition

After localization of printed characters in the image, executes the second stage is character recognition. The first step is the selection of the image. The function execution result is shown in Figure 7.

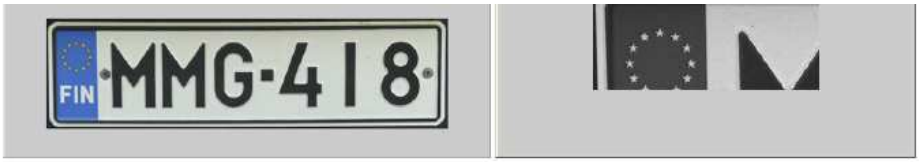

Figure 7: Original Image and License Plate Image with a Selected Area

Get the coordinates of the location of the first character on the image.

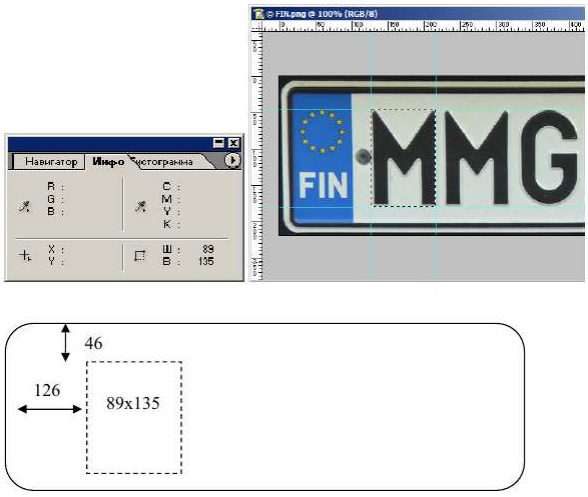

Figure 8: The Determination of the Coordinates of the First Character

Using the original image and the Photoshop program, create the reference image.

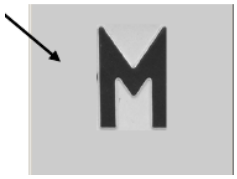

Figure 9: Source and Reference Picture

Proposed work recognizes only binary images, so the second stage after receiving the pictures, she binarized. When working with a color camera, the transformation from color to black and white goes 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. At the stage of binarization converts an object image into a binary data matrix, and then the work begins not with the object images and binary matrix.

If you do not provide a partitioning of the image into parts, none of the above algorithms will not work correctly. Split image into pieces, each of which contains its own unique object is called segmentation.

It should be noted that segmentation clearly separated black and white images, binary and gray. Working here is completely different as the speed and complexity of algorithms, however, it is intuitively clear that any image with shades of gray can be binary soul according to some rules. 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 (NS) for image recognition. NS 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

Study of methods and hardware and software systems of optical character recognition allows to formulate the following conclusions:

- The current state of technology for the automatic recognition of printed text (OCR) allows solving the problem of automation of information input with the required level of reliability.
- Building an OCR system, comprising the optical device capturing, the block localization and selection of text elements, the block preprocessing of the image; a block feature extraction unit character recognition unit and the post processing of the recognition results, it is necessary to use methods and algorithms with high robustness to arkosta geometric distortions and complex textured backgrounds.
- Such methods and algorithms can be used: procedure detect the lines of familiarity on the basis of modifications

of the Hough transform; methods based on the study of stable statistical distributions of points; methods using integral transforms and structural analysis of characters.

- In the development of modern systems of OCR to enhance recognition of characters and words it is necessary to consider contextual information. The use of contextual information allows not only detect errors but also correct them.

Turning to the program developed during the studies it should be noted that although it does not apply artificial intelligence systems (perceptrons and neural networks), and used a fairly simple method of comparison with the reference symbols, the algorithm gives a reasonable result for a pre-known set of standards.

The application of this method would be appropriate in cases when it is necessary to recognize large volumes of text printed in one font in one size. Under these conditions, the recognition results can compete with the methods based on the use of neural networks and not to give them the speed of recognition.

With all of this, the method of comparison with the standard much easier for other algorithms uses a simple mathematical apparatus. However, a small deviation of the input data from the reference values lead to a sharp drop in the recognition quality.

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.