Character recognition algorithms

A A Valke, D G Lobov
Omsk State Technical University, 11, Mira Ave., Omsk, 644050, Russia

Abstract. The article discusses various algorithms for character recognition, describes their advantages and disadvantages. The modified algorithm of character recognition based on the template method is proposed, its block diagram is given, the results of testing are presented, as well as the possibility of its application in the system of automated test forms verification. The article presents experimental data on the probability of correct recognition of a symbol of an arbitrary number, depending on the number of loaded characters in the pattern obtained using this algorithm. The algorithm developed by the authors is planned to be used in the development of a testing forms verification system, both for entrance tests and for various disciplines studied at the University. This system will allow to automate the test results verification, create a test results database and on its basis to create various reports types.

1. Introduction
Recognition of images, including symbols, is one of the urgent tasks arising in various fields of human activity today. For example, it can be used in industry for automatic part recognition, in education for automated testing forms verification and in other fields [1-3].

Despite the fact that the first publications on the topic of image recognition appeared in the 60s of the 20th century [4,5], the practical application of image recognition algorithms began only in the 1990s, which is associated with a significant increase in computer performance.

2. Problem statement
One of the most common problems of image recognition is the problem of handwriting recognition. Despite the large number of different programs designed for handwriting recognition, the relevance of the development of new software is not reduced. This is primarily due to the fact that free software products are mainly focused on the recognition of printed text, not handwritten. The authors conducted research of some software products. Thus, the free packet Open OCR Cuneiform (file version 12.0.0.58851) showed the following results: from the handwritten Russian text (the text is written in block letters) consisting of 192 characters 123 characters (64%) were correctly recognized, from the printed Russian text consisting of 191 characters 186 characters (97%) were correctly recognized.

Specially designed systems for streaming data and documents input, such as ABBYY FlexiCapture, show the 98% accuracy of handwriting recognition, but have a high cost (more than 200 thousand rubles).

Based on this, it was decided to develop software that uses its own character recognition algorithms, built on existing ones.

3. Theory
There are several classes of character recognition algorithms, but they all require a high-quality input image. An unsharp image with a lot of noise will be detected incorrectly with a higher probability than a high-contrast image with no noise. Also, the recognition quality is affected by the slope and size of characters. Therefore, before using the recognition algorithm, it is necessary to pre-process the image. Currently, the most common methods of character recognition are [6].
3.1. The template method
One of the classical methods of character recognition is the template method. The template method basis is a comparison of the input image and a number of recognized characters template images. The dimensionality of the input image and templates should be the same. For this purpose the normalization of the input image should be performed.

To identify the symbol, special correlation coefficients are calculated, which show the difference degree between the input image and the template. In this case, if one of the correlation coefficients is much more important than the others, then this input image is clearly identified. The correlation coefficient can be calculated by the formula [6]:

$$r = \frac{\sum \sum (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum \sum (A_{mn} - \bar{A})^2)(\sum \sum (B_{mn} - \bar{B})^2)}}$$

(1)

where $A_{mn}$ is the input image, $B_{mn}$ is one of the templates.

The advantages of this method are the simplicity of the algorithm implementation, relatively high speed of recognition, as well as good recognition of symbols with various defects, for example, shredded or glued.

The disadvantage of this method is that you need to write characters using a pre-defined template for correct recognition.

3.2. The indicative method
In the indicative method, the N-dimensional feature vector is constructed, for example, the number of closed areas, the percentage of occupancy, etc. The constructed feature vector is compared with a set of reference vectors of the same dimension by calculating the Euclidean distance between the reference vector and the feature vector of the recognized text.

The advantage of this method is the immunity to the symbol's shape, its inclination, proportions, etc.

The disadvantage of this method is the instability to various image defects.

3.3. The structural recognition method
The structural recognition method converts the input image into its topological representation, which displays information about the relative position of the symbol structural elements. This information is most often presented in graphs.

This method has advantages and disadvantages similar to the indicative method.

3.4. The artificial neural network
Recently one of the most common symbol recognition methods is the recognition using a neural network. The main difference between the use of neural network and the classical recognition methods is that neural networks are not programmed, but are trained. The training possibility is one of the main neural networks advantages over traditional algorithms. In addition, this method has high efficiency and performance [7,8]. The neural networks disadvantages include the fact that this recognition method requires a large data amount of used for training and the special structure development of the neural network, taking into account this task specifics.

4. Experimental result
This paper presents a modified template method for recognizing Arabic numerals in test forms. A feature of the proposed template method is the ability to edit the template in the recognition process.
This allows to dynamically adjust the template for different handwriting. The recognition algorithm is shown in figure 1.

**Figure 1.** Character recognition algorithm.

The program loads a database of symbols and forms templates from a hard drive. The detection of the character boundaries located in the recognition block and its scaling takes place after loading. The most upper black pixel and the lowest black pixel coordinates belonging to the symbol are determined for scaling. After that, the height of the symbol is found. Then the symbol width is determined in the same way, and the scale factors are calculated horizontally and vertically. Character
scaling is necessary to adapt the dimension of the recognized character to the dimension of the template.

Then the correlation coefficient for all patterns is calculated. The symbol template is a probability distribution of a black pixel appearing at a point with specified coordinates. To calculate the correlation coefficient, a pixel-by-pixel comparison of the symbol image with the current template is performed.

The symbol pattern with the maximum correlation number is considered correct. In this case, the user has the ability to edit the template in the recognition process. So, if the symbol is written correctly, the information about its writing can be entered into the template, that is, to change the probability distribution and write it to the template database. In this, the number of templates does not increase. This allows to increase the probability of correct character recognition at the same recognition rate.

5. Discussion results
To test the above algorithm, specialized software was developed for test forms recognition that include handwritten symbols of Arabic numerals. During the test, more than 2500 symbols of Arabic numerals were recognized. Table 1 presents experimental data on the correct recognition probability of Arabic numerals from "0" to "9" depending on the loaded characters number in the template obtained using this algorithm.

Table 1. The correct recognition probability of different symbols depending on the number of symbols loaded in the template.

| Number of loaded characters into the template | «0» | «1» | «2» | «3» | «4» | «5» | «6» | «7» | «8» | «9» |
|---------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 10                                          | 0,9 | 0,7 | 0,7 | 0,6 | 0,8 | 0,8 | 0,6 | 0,5 | 0,4 | 0,5 |
| 30                                          | 1   | 0,7 | 0,7 | 0,7 | 0,9 | 0,6 | 0,6 | 0,5 | 0,5 | 0,6 |
| 50                                          | 1   | 0,6 | 0,7 | 0,7 | 0,7 | 0,8 | 0,8 | 0,7 | 0,4 | 0,8 |
| 100                                         | 1   | 0,6 | 0,8 | 0,9 | 0,9 | 1   | 1   | 0,7 | 0,5 | 0,9 |
| 150                                         | 1   | 0,8 | 0,9 | 1   | 1   | 1   | 1   | 0,7 | 0,7 | 0,8 |
| 200                                         | 1   | 0,9 | 1   | 1   | 0,9 | 1   | 0,9 | 0,9 | 0,7 | 0,8 |

As can be seen from table 1, the increase in the number of symbols loaded into the template does not always have a positive effect. For example, if you recognize characters "4" and "6", increasing the number of loaded characters in the template gives a worse result. Therefore, when developing a template, you should consider this point.

Table 2 presents the correct recognition experimental data probability of correct recognition of an random digit symbol depending on the loaded symbols number in the template obtained using this algorithm.

Table 2. The correct recognition probability of a random digit symbol depending on the number of symbols loaded in the template.

| Number of characters loaded into the template | The probability of detecting |
|---------------------------------------------|------------------------------|
| 10                                          | 0,65                         |
| 30                                          | 0,68                         |
| 50                                          | 0,72                         |
| 100                                         | 0,83                         |
As can be seen from table 2, the correct symbol recognition probability averages 0.91, which is comparable with the program Abby FormReader 4.5 recognition results. According to the experiments carried out, the accuracy of the Arabic numerals symbols recognition when using the Abby FormReader 4.5 program is approximately 0.98 (when checking the test forms). To improve this result, it is planned to use additional recognition algorithms, such as neural networks, which will increase the correct symbol definition probability.

6. Conclusion
The algorithm developed by the authors is planned to be used in the development of a testing forms verification system, both for entrance tests and for various disciplines studied at the University. This system will allow to automate the test results verification, create a test results database and on its basis to create various reports types.

References

[1] Zhong Zhang, Hong Wang, Shuang Liu and Baihua Xiao 2018 Consecutive Convolutional Activations for Scene Character Recognition IEEE Access pp 35734-35742
[2] Jiang, Yuxiang & Dong, Haiwei & El Saddik, Abdulmotaleb 2018 Baidu Meizu Deep Learning Competition: Arithmetic Operation Recognition Using End-to-End Learning OCR Technologies IEEE Access pp 1-1. 10.1109/ACCESS.2018.2876035
[3] Jack Greenhalgh, Majid Mirmehdi 2015 Recognizing Text-Based Traffic Signs 2015 16 3 IEEE Transactions on Intelligent Transportation Systems pp 1360-1369
[4] Steinbuch K Piske U 1963 Transaction of Electronic Computers EC-12 6 DOI: 10.1109/PGEC.1963.263588
[5] Abramson N Braverman D 1962 Learning to recognize patters in a random environment Trans ITE IT-8 5
[6] Yafang Xue 2014 Optical character recognition
[7] Ahmad Muqeem Sheri Hyunsang Hwang Moongu Jeon Byung-geun Lee Neuromorphic Character Recognition System With Two PCMO Memristors as a Synapse 2014 61 IEEE Transactions on Industrial Electronics pp 2933-2941
[8] Mehdi L A, Solimani A Dargazany 2009 41st Symposium Southeasten on system theory (Tullahoma, TN, USA. Tullahoma, TN, USA, 2009) pp 44-48