UDC 004.352

O. Polshakova, E. Malchenko

TEXT ON AN IMAGE RECOGNITION METHOD, BASED ON THE SYMBOLS STRUCTURAL MODELS SIMILARITY CRITERIA

Abstract: The main methods of text recognition in the image are considered and analyzed in the work. An expedient approach to solving the problem of character recognition in the image using the method of comparing structural models of symbols, which is based on the criterion of similarity through the search for the maximum pairing of the minimum weight.

Keywords: character recognition, structure model, similarity criteria, дводольний граф.

Introduction

All existing universal approaches to solving the problem of character recognition are focused on the use of the reference base of reference images, which for high recognition accuracy, must be large enough. At the same time, there are a number of tasks in which the number of pre-known graphical representations of characters is very limited. Examples of such tasks are: recognition of forms filled in atypical handwriting, selection of textual information on available in a single copy of historical documents, identification of signatures in bank documents, identification of the user by handwritten signature, etc.

The purpose of this article is to describe a method that will allow you to solve the problem of character recognition with high accuracy, using a reference sample of the minimum volume.

Formulation of the problem

For effective text recognition in the image in the absence of a large training sample, it is necessary to present the development of a method that will allow high accuracy to classify characters regardless of their language, handwriting, etc.

Description of existing solutions

Existing approaches can be divided into the following categories:

- methods using feature classifiers;
- statistical methods;
- methods using structural components.

The first group of methods is one of the most common in the scientific community. In developing such methods, one can abstract from the details of the image comparison process by placing this problem on one of the mathematical tools for classification. Artificial neural networks have become widespread in solving problems of character and image recognition. At present, on the World Wide Web, there are a huge number of libraries that implement certain feature classifiers. Given this, the implementation process becomes relatively time-consuming and the solution of the recognition problem is reduced to the correct choice of feature space, as

 ISSN 1560-8956

well as a representative training and test sampling of images. However, a significant disadvantage of this group of methods is the inability to provide a person with a reasoned and understandable explanation of why the image or symbol is assigned to a particular class. In addition, methods based on characteristic classifiers usually require large-scale training samples, which contradicts the task, taking into account the view proposed in this paper.

The second group of methods is also simple in terms of software implementation. The initial graphical representation of the image is analyzed in order to identify various statistical values. Then, on the basis of the received data, recognition by means of both the simple sign classifier, and the heuristic approach with use of the revealed regularities is carried out. Histograms of the location of the pixels of the symbol are much more obvious to human perception and can be useful in the process of debugging or improving the operation of the algorithm. However, the basis of such methods is the establishment of statistical patterns, which in turn requires a large initial training sample.

The third group of methods differs in that the search for patterns and dependencies of the graphical appearance of symbols is entrusted to the developer of the method. The graphical representation of the symbol highlights the key elements - the structural components. They can be represented, for example, by graphic primitives or a sequence of actions of the operator when depicting a symbol. Extracting such components from a graphical representation is a rather time-consuming task, but it is even more difficult to use them to compare two symbols. Both the development and software implementation of such methods usually takes a significant amount of time. However, such algorithms have a higher recognition accuracy in problems where there is one or more reference forms of graphical representation of the symbol. The key difficulty of such an approach is to identify a reliable and not too time-consuming algorithm for comparing structural models of graphical representation of symbols.

Problem solution

One approach to solving the problem of character recognition is to alternately compare the graphic image with each of the standards, the classification of which is known, and the subsequent selection of the class is made from the most similar reference image. In the general case, to assess the degree of similarity of graphical representations of two symbols based on structural models, you need to build a structural model for each of these symbols separately, and then compare the two obtained models. However, for all graphical representations of the base of the reference images, you can create structural models and not spend computing power to build them each time you start the process of character recognition. Next, we turn to a detailed consideration of the proposed algorithm.

Denote the terms that will be used to describe the extracted structural components of the representation of the symbol (Fig. 1), as well as to describe the algorithm for obtaining them.

Key pixels are black pixels of a skeletalized binary image of a symbol that correspond to key points or bends of its structural model.

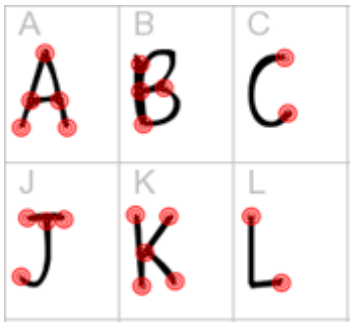

Figure 1. Examples of letters of the Latin alphabet with marked key elements

A connecting edge is a structural component of a symbol representation that describes a sequence of black pixels on a skeletalized binary image of a symbol connecting two key pixels.

The key point is the structural component of the symbol representation, which is the junction of one, three or more connecting edges. Also, the key point can be called the junction of two connecting edges, the guide vectors of which, when selected as the starting position of this key point, are located at an angle of less than 120 degrees.

Bending (bending point) - a structural component of the symbol representation, which can not be attributed to key points, but the line connecting the two key pixels and passing through this pixel, significantly changes the direction of the graphical representation of the symbol corresponding to this pixel. In fact, the bend is the junction of two connecting edges, the guides of the vector which, when selected as the initial position of the bend, are located at an angle of at least 120 degrees.

Composite edge - a sequence of connecting edges that begins and ends at key points, which does not contain any key point, except the start and end.

Thus, black pixels, which are not key pixels or bends, are combined into connecting edges, which, in turn, are combined into composite edges.

Assessing the degree of similarity of two composite edges can be a difficult task, given that they may differ: the number of components of their graphic primitives, the types of these primitives, their total length, as well as the location of connecting bends and key points.

However, there is a way to determine the area of the figure enclosed between two composite edges, despite significant differences in their characteristics. This area with the opposite sign can be used as a measure of similarity between the edges: the larger the area, the more different these edges are.

In addition to the area, the degree of similarity should be influenced by the distance between the corresponding points of the composite edges - the farther these points are, the less similar the corresponding composite edges are.

Consider a composite edge as a route along which an object passes in one second. The speed of the object is such that in this second it will travel the full path from one end of this composite edge to the other. Let the objects begin their movement simultaneously along each of the edges. At each of the possible moments of time t we draw a segment connecting the positions of the objects. The set of such segments also forms the required area concluded between two composite edges (Fig. 2).

 ISSN 1560-8956

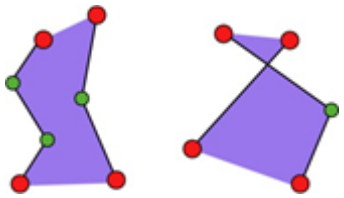

Figure 2. The area between two composite ribs as a resemblance to these edges

Even if the composite edges have common points, this method of calculating the area between them can be used to assess the degree of similarity.

Consider the case when the described method is to find the area enclosed between two segments. Let the first segment be given by its two ends (x_a, y_a) and (x_b, y_b), second – by two edges (x_c, y_c) and (x_d, y_d). Then at some point in time t the object on the first segment will be at a point (x_1, y_1), on the second – in the point (x_2, y_2). The coordinates of the point (x1, y1) can be found using formula (1).

$$\begin{aligned} x_1 &= x_a + (x_b - x_a) \cdot t \\ y_1 &= y_a + (y_b - y_a) \cdot t \end{aligned} \tag{1}$$

Similarly, you can determine the coordinates of a point (x_2, y_2):

$$\begin{aligned} x_2 &= x_c + (x_d - x_c) \cdot t \\ y_2 &= y_c + (y_d - y_c) \cdot t \end{aligned} \tag{2}$$

The square of the distance between them is described by formula (3):

$$d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 \tag{3}$$

The value of the first term of formula (3) can be represented using formula (1):

$$\begin{aligned} (x_1 - x_2)^2 &= ((x_a - x_c) + (x_b + x_c - x_a - x_d) \cdot t)^2 = \\ &= (A_x + B_x \cdot t)^2 = B_x^2 t^2 + 2 \cdot A_x \cdot B_x \cdot t + A_x^2 \end{aligned} \tag{4}$$

Similarly, the value of the second term:

$$\begin{aligned} (y_1 - y_2)^2 &= ((y_a - y_c) + (y_b + y_c - y_a - y_d) \cdot t)^2 = \\ &= (A_y + B_y \cdot t)^2 = B_y^2 t^2 + 2 \cdot A_y \cdot B_y \cdot t + A_y^2 \end{aligned} \tag{5}$$

The final formula for calculating the distance between objects at any time will look like:

$$\begin{aligned} d &= \sqrt{\left(B_x^2 + B_y^2\right) \cdot t^2 + 2 \cdot \left(A_x B_x + A_y B_y\right) \cdot t + \left(A_x^2 + A_y^2\right)} = \\ &= \sqrt{a \cdot t^2 + b \cdot t + c} \end{aligned} \tag{6}$$

Using formula (6), you can find the area between the two considered segments:

$$\begin{aligned} S &= \int_{t_1}^{t_2} \sqrt{a \cdot t^2 + b \cdot t + c}\, dt = \\ &= \frac{1}{8 \cdot a^{3/2}} \cdot \Big[\Big(2 \cdot \sqrt{a} \cdot (2 \cdot a \cdot t_2 + b) \cdot \sqrt{t_2 \cdot (a \cdot t_2 + b) + c}\Big) - \\ &-(b^2 - 4 \cdot a \cdot c) \cdot log\left(2 \cdot \sqrt{a} \cdot \sqrt{t_2 \cdot (a \cdot t_2 + b) + c} + 2 \cdot a \cdot t_2 + b\right) - \\ &- \Big(2 \cdot \sqrt{a} \cdot (2 \cdot a \cdot t_1 + b) \cdot \sqrt{t_1 \cdot (a \cdot t_1 + b) + c}\Big) + \\ &+(b^2 - 4 \cdot a \cdot c) \cdot log\left(2 \cdot \sqrt{a} \cdot \sqrt{t_1 \cdot (a \cdot t_1 + b) + c} + 2 \cdot a \cdot t_1 + b\right)\Big] \end{aligned} \tag{7}$$

Formula (7) can be used for areas of a pair of composite edges from t_1 to t_2, where there are no joints of the two connecting edges, provided that both connecting edges on this gap are segments. To avoid overly laborious calculations of integrals for arcs and elliptical arcs, they can be approximated by a sequence of segments. In this case, the composite edge will be represented by the same sequence of connecting edges, each of which will be a segment, but their number can be much larger (Fig. 3).

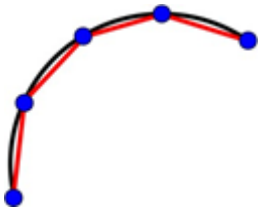

Figure 3. Approximation of the arc by segments

This approach will not use the exact values of the area, but approximate. However, such an error is not very critical for the task of estimating the degree of similarity, but the development and implementation of such an approach is much simpler than a similar implementation without approximation of arcs by sets of segments.

After the area between them is calculated for each pair of composite edges, this area can be used as a measure of the difference between these edges: the larger this area, the more different these two edges are. To obtain a measure of similarity, it is enough to multiply the resulting area by minus one.

For each pair of composite edges from different structural models, their degree of similarity is known. You can construct a bipartite graph G, in which the vertices will be composite edges of the original structural models. The first fraction of the graph G will include the vertices corresponding to the composite edges of the first structural model, the second fraction will include the vertices corresponding to the second model.

In order to assess the degree of similarity of the two structural models, it is necessary to compare the vertices of the first lobe with the vertices of the second lobe so that the largest possible number of pairs was formed. From all such distributions on pairs it is necessary to choose partition at which the sum of degrees of similarity will be the maximum. Since the degree of similarity is essentially the magnitude of the area with the opposite sign, such a task will be equivalent to the problem of finding the maximum pair of minimum weights in a bipartite graph, provided that the edge of such a graph has a weight equal to the area between the corresponding composite edges.

It should be noted that the fractions of the constructed graph can contain a different number of vertices. In this case, some vertices will not be included in the maximum pairing. For each of these vertices to the total weight of the found pair, it is necessary to add twice the minimum weight of the incident incident to it, as a "penalty" for the mismatch of the number of composite edges in the two models.

 ISSN 1560-8956

The process of finding the maximum pair of the minimum weight in a bipartite graph is reduced to the task of finding the maximum flow of the lowest value in the network. To do this, you need to build a new graph G', which will be a network with one source and one drain. To obtain such a graph from the original bipartite graph it is enough to add two vertices. The first of these vertices is a source, it must be connected by oriented edges with all vertices from the first part. All such edges must be directed from the source to the vertices of the first lobe. The second vertex that is added - the drain, it must be connected with oriented edges with all the vertices of the second part. The direction of such edges will always be from the vertices of the second particle to the drain. All edges between the particles are replaced by similar oriented edges from the first part to the second. The structure of the obtained graph G' is presented in Fig. 4:

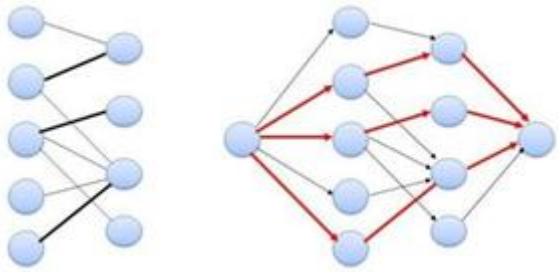

Figure 4. The scheme of the grid obtained from a bipartite graph

So the graph G 'is a network for which it is necessary to determine the value of the maximum flow, for each of its edges is determined by the amount of bandwidth. In this case, it is enough to assign a single bandwidth to each of the edges. This is because if any edge is already involved in the current pair, it should no longer be involved in the formation of new pairs.

In the received network it is necessary to find the maximum stream of the lowest cost. Cost - a value that characterizes each of the edges of this network. The cost of an edge reflects the cost of including this edge in the found stream. For the constructed network G 'the value of all edges from the source and all edges in the drain is zero. The values of the edges held between the vertices of the fate correspond to the weights of the edges in the original bipartite graph G.

One of the most well-known methods of finding the maximum flow of the lowest cost is the algorithm of increasing paths. To find the maximum flow of the lowest value in the network G 'you need to perform the following steps:

1. Pre-prepare the graph G ', adding for each edge a return edge with zero bandwidth and a value equal to the value of the output edge with the opposite sign.

2. Using as the weights of the edges of their value, find the shortest path from source to drain.

3. If there is no path from source to drain, the algorithm is considered complete. The flow found before this iteration is the maximum flow of the minimum cost.

4. Start the flow on the found route. To do this, consider all its ribs. Determine the edge whose bandwidth is minimal, denote it by f. Reduce the bandwidth of all edges of the found path by f, while increasing the bandwidth of the opposite edges by the same amount.

5. Perform step 2 again for the upgraded network.

The computational complexity of the above algorithm can be represented by the formula: O (V3E), where V is the number of vertices in the network, E is the number of edges in it, including the inverse edges added in the first step of the algorithm. Since the number of vertices V in the resulting network G 'is equal to E1 + E2 + 2, with a large number of composite edges in the two structural models, finding the maximum flow at the lowest cost can significantly slow down the process of character recognition. In practice, the number of composite edges in each of the models rarely exceeds 10, so the algorithm has little effect on the overall performance of the recognition process.

Conclusion

The article offers a detailed description and mathematical substantiation of the algorithm for estimating the degree of similarity of structural models of handwritten symbols based on the method of maximum pairing of minimum weight to solve the problem of text recognition in the image in a small reference (training) sample. Based on the review and analysis of methods of text recognition in the image, it is possible to conclude about the usefulness of the proposed algorithm for solving the problem of text recognition in the image in a small reference (training) sample. Based on it, you can get a more appropriate and efficient algorithm for recognizing characters with high accuracy and a relatively small sample of reference structural models.

REFERENCES

1. Kavallieratou E. Handwritten Character Recognition based on Structural Characteristics

2. Chan K.-F. Recognizing on-line handwritten alphanumeric characters through flexible structural matching

3. Lucas S. A comparison of syntactic and statistical techniques for off- line OCR

4. Lucas S. A comparison of syntactic and statistical techniques for off- line OCR

5. Майника Э. Optimization algorithms on networks and graphs. M .: Mir, 1981

 ISSN 1560-8956