Research on Automatic Splicing Technology of Chinese File Fragment based on MATLAB Rules

Liang He

Xi'an International University, Xi'an, Shaanxi, 71007, China

Keywords: Rule fragmentation; 0-1 planning; cluster analysis; text feature extraction; L1-norm

Abstract: In order to realize the splicing of Chinese text fragments in rules, this paper studies the characteristics of Chinese character text in the rule fragment file, proposes the extraction method of text line information in the file fragment, and defines the concept of fragment boundary degree based on L1-norm, which is based on 0. -1 Planned file fragmentation model and using cluster analysis to reduce algorithm complexity. Compared with the existing similar algorithms, the algorithm of this paper can complete the correct splicing without manual intervention.

1. Introduction

The splicing and restoration of file fragments has important applications in information acquisition and file repair [1]. In the absence of computer assistance, the splicing of a small amount of debris can be done by hand alone, but it takes a relatively long time. When the size of the debris increases, the difficulty of manual splicing and the labor time are greatly increased. Therefore, designing an appropriate algorithm to achieve computer splicing can solve this problem. In reality, file splicing is roughly divided into two categories: irregularly shaped splicing [2-6] (depending on the boundary contour and information on the shard) and regular splicing of shards [7-10], the latter The fragments do not have obvious boundary shape information and must therefore be spliced based on the information on the fragments.

In this paper, for the second type of splicing problem, the automatic splicing technology of the Chinese file fragment of the rule is studied. Firstly, the fragmentation is digitized and characterized, and the clustering algorithm is used to group the data according to the text line information, which reduces the complexity of the splicing algorithm. Then, the concept of fragment boundary degree based on L1-norm is defined, and the 0-1 planning model with the minimum total difference is established accordingly. Finally, the file information is restored using the line information again.

The fragment samples used in this paper are from a printed Chinese document. They are cut into 209 pieces of the same size by the shredder. The Chinese characters printed in the fragments are the same size, the word spacing and line spacing are constant, and the text is horizontal. arrangement. In order to use the computer for fragmentation, the physical fragments of the file are scanned to obtain the image file, and then each image file is converted into a matrix of size 180×72 (hereinafter referred to as a fragment matrix) and numbered by software. The element value in the fragment matrix corresponds to the gray value of the pixel in the image file, and the value range of the matrix element is [0, 255].

2. Demarcation boundary definition

Regularly cut file fragments do not have identifiable outline information and must be stitched by textual information in the image. In this paper, we use the difference size of the two picture boundary pixels to judge whether the splicing should be performed, which is called the boundary difference degree in the following. Let the last column vector of the i-th fragment matrix (ie, the gray value of a column of pixels on the right edge of the i-th fragment) be recorded as r_i, the first column vector of the j-th fragment matrix (ie, the left edge of the j-th fragment) A column of gray values is recorded as l_j, and the two pieces are spliced left and right (shard i is on the left and fragment j is on the right), and the boundary difference of horizontal splicing is defined as

  DOI: 10.25236/dpaic.2018.018

$$h_{ij} = \sum_{k=1}^{180} \left(\left| r_{ik} - l_{jk} \right| \right), i, j = 1,2, \ldots ,209$$

Where r_ik and l_jk are the kth elements of the vectors r_i and l_j. Similarly, it is also possible to define the boundary difference between two upper and lower longitudinal stitches as

$$v_{ij} = \sum_{k=1}^{180} \left(\left| u_{ik} - d_{jk} \right| \right), i, j = 1,2, \ldots ,209$$

Where u_ik and d_ik are the kth element of the vector u_i (the vector corresponding to the upper edge of the i-th fragment) and d_j (the vector corresponding to the lower edge of the j-th fragment). It can be seen from the definition that the smaller the value of the boundary difference is, the higher the probability that the two boundaries are correctly spliced.

Mathematical model based on 0-1 programming

In theory, there should be no difference between the two adjacent boundaries obtained by the cropping. Therefore, the sum of the differences in the boundary of the fragments when correctly splicing should be zero. However, considering the change of image resolution and pixel gray value, there is a possibility that there is a boundary difference in the actual cutting. Therefore, from the perspective of probability, the optimal splicing method is a method that minimizes the sum of the differences of all splicing boundaries.

Literature [10] proposed a mathematical model of 0-1 programming to describe this problem, and its objective function is

$$\min \sum_{i=1}^{209} \sum_{j=1}^{209} \left(h_{ij} x_{ij} + v_{ij} y_{ij} \right)$$

$$x_{ij} = \begin{cases} 0, & \text{Indicates that the i} - \text{th fragment and the j} - \text{th fragment are not stitched left and right} \\ 1, & \text{Indicates that the ith fragment and the jth fragment are stitched left and right} \end{cases}$$

$$y_{ij} = \begin{cases} 0, & \text{Indicates that the i} - \text{th fragment and the j} - \text{th fragment are not stitched up and down} \\ 1, & \text{Indicates that the i} - \text{th fragment and the j} - \text{th fragment are spliced up and down} \end{cases}$$

It can be seen from the objective function of the model that this model considers both horizontal splicing and vertical splicing of all fragments. It is a two-dimensional splicing problem, but because the model needs to set more variables, the constraints are complex, and the model needs to be solved. The memory space is large, which can cause a memory leak in the normal configuration of the computer when solving the model. In order to solve the memory problem solved by the model, a natural idea is to group large-scale data and degrade the two-dimensional fragment mosaic problem into a one-dimensional fragment mosaic problem that is, only horizontal line stitching or vertical column stitching of the fragments. For example, for a set of horizontally stitched pieces, the mathematical model of the 0-1 plan is simplified to:

$$\min \sum_{i=1}^{N} h_{ij} x_{ij}$$
$$s.t. \sum_{i=1}^{N} x_{ij} \le 1$$
$$\sum_{j=1}^{N} x_{ij} \le 1$$
$$\sum_{j=1}^{N} x_{ij} \le N-1$$
$$x_{ij} = 0 \text{ or } 1$$

Where x_ij=1 indicates that the ith fragment and the jth fragment are spliced left and right, and x_ij=0 indicates that the two are not spliced, and N is the number of fragments to be spliced.

All fragments must be grouped before using the model (4) for fragment splicing. Since the file fragment is a regular cut of the text file, it can be ensured that the line of the text is horizontally oriented horizontally and perpendicular to the longitudinal cutting direction, so that the line information of the text in each fragment can be accurately grouped.

3. Line information feature extraction and row clustering

In order to perform an accurate classification, it is necessary to determine valid line information features. The following characteristics can be found by observing a given file fragment:

The height of each piece can contain three lines of text, and some pieces are only printed one line or two lines of text;

The upper and lower ends of the debris may contain only half of the lines that have been cut;

The height of the complete text in each line of most fragments is equal. After testing, the text height is 41; the line height is 68;

The lines in individual fragments contain only words such as "one", and their height is less than 41;

Among them, the (1), (2) and (4) bring the randomness of the row information to the row classification. The method of the literature [10] will lead to the partial classification of some special fragments. According to the analysis, since the line height of the text is constant, as long as the line contains a complete line of text, the information of other lines on the fragment can be generated. In this paper, the position of the center line of the first line (the ordinate in the coordinate system) is used to represent the line information characteristics of the fragment. Then, the center line of the first line of the two pieces that should be spliced to the left and right should coincide. In this paper, the sliding window summation method is used to obtain the position of the first axis of the debris. The calculation method is as follows.

The row mark vector of the fragment file is V=(v_1,v_2,...,v_180)^T, and v_i=0 means that the i-th row element of the fragment matrix is all 255 (that is, a blank line), and v_i=1 indicates that there is an ith row. An element smaller than 255 (that is, a black pixel with text in the line). 2(a) is a fragment matrix, FIG. 2(b) is a row marker vector V, and FIG. 2(c) is a sliding window sum vector W=(w_1, w_2, . . . , w_180)^T, which is calculated as

$$w_k = \sum_{l=k-20}^{k+20} v_l \quad k = 1,2,\ldots,180$$

The window width of the sliding window is 41 (the height of the text). Then the position of the central axis of the first line of each fragment is

$$Mid = mod\left(\underset{k}{\text{argmax}}(w_k), 68\right)$$

Where argmax ┬ k (w_k) represents k which takes w_k to the maximum value. The first horizontal line in Fig.1 is referred to as the center line of the first line, and its vertical axis coordinate is 170.

Through the above method, the coordinates of the first row of the central axis of all the fragments can be obtained. The result is shown in Fig. 3. The abscissa of each point in the figure is the fragment number, and the vertical axis coordinate is the position of the central axis of the first row of the fragment. . As can be seen from the figure, the first row of the central axis has a very obvious clustering feature, and 10 thresholds are set (as indicated by the horizontal lines in the figure), so that all the fragments can be grouped. From the grouping results, 209 pieces were divided into 11 groups with 19 pieces in each group. Compared with the method adopted in [10], the clustering method of this paper can complete correct grouping at one time without manual intervention.

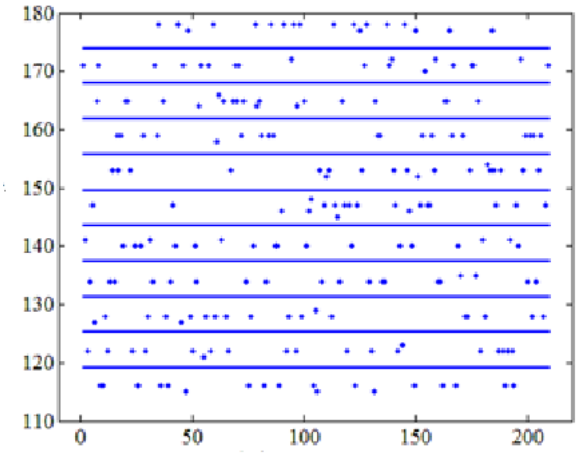

Figure 1 The first line of text fragmentation and the clustering threshold

After correct grouping, horizontal splicing of each group of debris utilization models (4) can obtain the correct splicing results of all the groups.

4. File stitching recovery

After getting the correct grouping and the stitching result (line fragmentation) of each group, 11 line fragments need to be stitched to restore the original text file. Since the horizontal cutting line has a higher probability of being located between the two lines of text when cutting the document, the boundary pixel values of the line fragments cannot be matched at this time, and thus, without using other auxiliary means, Application 0-1 planning can not achieve the correct splicing of line fragments. This paper proposes a search strategy that uses the existing first-line central axis information to complete the stitching recovery of the final file.

Let the position of the first row of the i-th and j-th row fragments be f_i and f_j, respectively. If the i-th row fragment and the j-th row fragment should be spliced up and down, then the first line of the two lines is fragmented. The distance between the central axes should be an integer multiple of the row height, ie the expression should be satisfied

$$f_i + (180 - f_j) = 68 * N$$

Among them, N=3, the line height is 68. Simulation results show that this method can automatically complete the correct line fragmentation.

5. Conclusion

In this paper, the feature fragmentation of the spliced file fragments is obtained. The boundary vector of the fragment is obtained. The line feature information of Chinese characters on the fragment is extracted. All the fragments are grouped by the line feature, and then the boundary

vector difference is used for each group. Fragment splicing based on 0-1 planning is performed. Finally, the spliced packet fragments are spliced again using row information, and finally restored to the original file. The splicing algorithm proposed in this paper is fast and effective. As long as it can guarantee at least one line of complete text in each fragment, that is, the height of the fragment is not less than the sum of the height of the line and the height of the text, the algorithm can automatically complete the splicing without human intervention.

References

[1] Yang Luobin. Research and application of shape matching technology in cultural relics restoration [D]. Xi'an: Northwest University, 2002.

[2] Jia Haiyan, Zhu Liangjia, Zhou Zongtan, et al. Shape matching method in automatic splice splicing [J]. Computer Simulation, 2006, 23(11): 180-183.

[3] Luo Zhizhong. Research on debris boundary detection algorithm based on line segment scanning [J]. Journal of Scientific Instrument, 2011, 23(2): 289-294.

[4] Xie Ping, Ma Xiaoyong, Zhang Xianmin, et al. A Fast Complex Polygon Matching Algorithm[J]. Computer Engineering, 2003, 29(16): 177-181.

[5] Zhu Yanjuan, Zhou Laishui. Matching Algorithm for 2D Irregular Fragments[J]. Computer Engineering, 2007, 33(24): 7-9.

[6] Zhang Xinbu, Yan Long, Zhu Liangjia, et al. Research on shredded paper contour extraction technology in physical evidence restoration system[J]. Computer Simulation, 2006, 23(11): 184-187.

[7] Efthymia T, Ioannis P. Automatic color based reassembly of fragmented images and paintings [J]. IEEE Transactions on Image Processing, 2009, 19(3): 680-690.

[8] Nasir M, Anandabrata P. Automated reassembly of file fragmented images using greedy algorithms [J]. IEEE Transactions on Images Processing, 2006,15(2):385-393.

[9] Luo Zhizhong. Semi-automatic stitching of document shredded paper based on text features[J]. Computer Engineering and Applications, 2012, 48(5): 207-210

[10] Shen Hongping, Zhang Yipeng, Wang Yikang, Research on Chinese fragmentation mosaic restoration based on 0 - 1 planning model [J]. Electronic Technology, 2014(27)13-16