Establishment and research of shredded paper mosaic restoration model

Jungang Chen

Ningbo University, Ningbo, China

dove_cn@sina.com

Keywords: Automatic stitching, feature extraction, assignment problems, optimized matching

Abstract: The splicing of broken documents has important applications in the fields of judicial material verification, historical document restoration and military intelligence acquisition. With the development of computer technology, the computer can develop the automatic splicing technology of shredded paper to improve the splicing recovery efficiency. In order to solve this problem, we first carry out the data processing on the fragments given in the attachment, use Matlab to program, obtain the binarized original data of the picture, then calculate the distance matrix between the two pictures, and finally output the restored picture.

1. Introduction

The splicing of broken documents has important applications in the fields of judicial material verification, historical document restoration and military intelligence acquisition. Traditionally, the stitching recovery work needs to be done manually, with high accuracy but low efficiency. Especially when the number of pieces is huge, it is difficult for manual stitching to complete the task in a short time. With the development of computer technology, people have tried to develop automatic splicing technology for shredded paper to improve the efficiency of splicing recovery. Try to establish a mathematical model analysis on the relevant information of the given fragments to study the following problems:

Question 1:

(1) According to the shredder shredder in the attachment (several slitting), a shredded paper stitching restoration model and algorithm are established.

(2) Splicing and restoring the fragment data of Annex 1 and Annex 2. If the recovery process requires manual intervention, write out the intervention method and the time node of the intervention, and express the results of the restoration in the form of pictures and tables.

Question 2:

(1) For the case where the shredder is slit and cross-cut, design a scrap paper stitching recovery model and algorithm.

(2) Splicing and restoring the fragment data given in Annex 3 and Annex 4. If the recovery process requires manual intervention, write the intervention method and the time node of the intervention. The expression of the recovery result is the same as above.

2. The problem analysis

Shredded paper splicing can be converted into an assignment problem [1], in order to assign each piece of paper to a fixed position, forming a correct order, treating the piece of paper as a person, ie each person completes, its workflow [2]] Generally:

(1) Pre-treating the shredded paper that is, digitizing the object fragments.

(2) Optimize matching through various matching methods

(3) Flattening scraps of paper

Question 1: The problem requires splicing and restoration of the fragmentation data of each page in Chinese and English given in Annex I and Annex II, and establishing a mosaic restoration model and algorithm.

  DOI: 10.25236/isrme.2019.115

Assume that there are $j\ (j=1,2,\cdots,19)$ spaces available for $i\ (i=1,2,\cdots,19)$ fragment selection. In order to assign 19 images to a fixed and reasonable position, sort them by optimization matching. First, use the Matlab to binarize the fragments and determine by manual intervention. The first fragment is solved by Matlab programming.

Question 2: The problem requires splicing and restoration of the fragmentation data of each page of Chinese and English documents given in Annex III and Annex IV, and establishing a mosaic restoration model and algorithm.

Suppose there are $j\ (j=1,2,\cdots,209)$ spaces for $i\ (i=1,2,\cdots,209)$ pieces to choose. In order to assign 11*19 pictures to a fixed and reasonable position, sort them by optimizing matching. Firstly, observe the 209 pieces of debris and find out that they are located in the first place. 11 pictures of a column, then read and binarize the picture, record 11*19 picture words and the top and bottom positions of the line and list the matrix, and then filter the remaining pieces through the already obtained borders. Then use Matlab programming to solve.

3. Model hypothesis

1) Assume that the text on the scraps of paper is oriented upwards;
2) Assume that the shape of the shredded paper is regular;
3) Assume that the shredded papers have the same height and the same spacing.

4. The establishment and solution of the model

4.1 Establishing shredded paper stitching restoration model and algorithm

- Problem 1: Mathematical model for the restoration of silted scraps

(1) Model preparation

Grayscale refers to the color depth of dots in black and white images. The range is generally from 0 to 255, white is 255, black is 0, so black and white images are also called grayscale images. The binarization of images is the pixel on the image. The gray value is set to 0 or 255, which means that the entire image is presented with a distinct black and white visual effect [2].

According to the above principle, each piece of fragment is read by the imread function in Matlab software. Matlab stores the grayscale image as a data matrix. The elements in the data matrix represent the pixels in the image, and the values are the gray values of the color. . Then use the im2bw function in Matlab software to binarize the gray value and convert it into image data represented by only 0-1. These data are used as the basic data for modeling below.

In order to solve the problem of slitting paper recovery, we first digitize the fragments given in the attachment, use Matlab to program, read the image and get the binarized raw data of the image, because the first piece of scrap paper is left. The value of the side must be all 1, so you can find the picture of the first column by observing the binarized data, then subtract the absolute value from the two pictures, and then sum the absolute value as the distance between the two pictures, the minimum distance Indicates that the two pictures match the most.

The positions of the pictures are numbered from 1 to 19, from left to right, and are recorded as matrix $[1,2,\cdots,19]$.

(2) Establishment of model one

Asked to splicing 19 scraps of paper into the original image.

When the problem is converted to an assignment problem, the mathematical programming expression for the assignment problem is given below:

Objective function: Find a task assignment with the smallest distance

$$\text{Min: } \sum_{j=1}^{18}\sum_{k=1}^{19}\sum_{i=1}^{19} A_{ij} d_{ik} A_{kj+1}$$

S.t. Each fragment must have one and only one space:

$$\sum_{j=1}^{19} A_{ij} = 1(i=1,2,\cdots,19)$$

Each space must and only accept one image:

$$\sum_{i=1}^{19} A_{ij} = 1(j=1,2,\cdots,19)$$

Each decision variable is a 0-1 variable:

$$A_{ij} = 0 \text{ or } 1,(i, j=1, 2, \ldots 19)$$

(3) Solution of model one

(a) Through the imread function to obtain the binary data of the fragment

Find the debris on the far left of a graph by manual intervention

It is concluded that the left boundary fragment number of Annex I Chinese page is 9 and the left boundary fragment number of Annex 2 English page is 4.

(b) Using Matlab software programming, find the distance matrix between the two fragments.

Solve the distance between the left and right boundaries of the fragments, and then sum the absolute values of the distance to obtain a matrix of 19×19. The smaller the value, the more matching the two.

(c) Through Matlab software to solve, get sorted, and display the restored picture

4.2 Problem 2: Mathematical model for splicing restoration of silted and cross-cut scraps

- Modeling preparation

(1) Acquisition of distance matrix 11

If $k > i : d_{i,k}$ indicates the distance between the kth picture and the i-th picture adjacent to the left and right.

If $k < i : d_{i,k}$ indicates the distance between the kth picture and the i-th picture adjacent to the top and bottom.

If $k = i \ : d_{i,k} = 0$.

(2) The position of the first column of pictures is numbered from 1 to 11 from top to bottom, and the position of the second column of pictures is numbered from top to bottom, 12-22, 198, 209, which is recorded as a matrix. $\begin{pmatrix} 1 & 12 & \cdots & 198 \\ 2 & 13 & \cdots & 199 \\ \vdots & \vdots & \cdots & \vdots \\ 11 & 22 & \cdots & 209 \end{pmatrix}$.

(3) According to the characteristics of the problem, we have relatively more blank edges according to the picture. We can calculate all the scraps of the four borders of the picture as follows:

$$leftA = \{i_1, i_2, i_{3,} \ldots, i_{11}\}$$

$$rightA = \{i_1, i_2, i_{3,} \ldots, i_{11}\}$$

$$upA = \{i_1, i_2, i_{3,} \ldots, i_{19}\}$$

$$downA = \{i_1, i_2, i_{3,} \ldots, i_{19}\}$$

- Establishment of Model 2

The second question is the problem of splicing 1119 pieces of scrap paper into the original picture.

Splicing to find the minimum distance as the goal. In the splicing process, the splicing scheme is planned in stages according to the following steps:

1) Calculate the collection of left, right, upper, and lower boundaries according to the principle of more debris blanks at the edges.

2) According to each blank line height of each fragment and the height of each line of text, 209 pieces are divided into 11 groups of 19 groups each.

3) Sort the 11 groups according to the upper and lower boundaries, the standard blank line height of each group and the upper and lower boundaries.

4) Splicing the fragments in each group. The splicing method is performed according to the following objective function.

Min: $\sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1}^{198} A_{ij}d_{ik}A_{k,j+1} + \sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1...209}^{j\neq 11\text{的倍数}} A_{ij}d_{ik}A_{k,j+1}$

S.t. Put all empty sums for each fragment i:

$$\sum_{j=1}^{209} A_{ij} = 1, i=1,2,\cdots,209$$

For each space, all images can only be merged into one: $\sum_{i=1}^{209} A_{ij} = 1, j=1,2,\cdots,209$

Each decision variable is a 0-1 variable:

$$A_{ij} = 0 \text{ or } 1, (i, j=1, 2, \ldots 209)$$

$$A_{i1} \in leftA \quad i=1,2,...,19;$$

$$A_{i19} \in rightA \quad i=1,2,...,19;$$

- Solution of model two

(1) Part 1

(a) Using Matlab programming to find 11 pictures in the first column, and then sorting them, the first line of pictures is manually intervened 3 times, then read and binarize the picture, and put the workspace in Matlab running. The binarized data is extracted into excel, the text is displayed by zooming out the enlarged table, and then the word height and line spacing of the Chinese character are calculated by manual search, and the 11 picture words and the upper and lower positions of the line are recorded and the matrix is listed, which is recorded as :

$$A_1 = \begin{bmatrix} 37 \\ 79 \\ 106 \\ 147 \\ 174 \end{bmatrix}, A_2 = \begin{bmatrix} 35 \\ 63 \\ 103 \\ 132 \\ 171 \end{bmatrix}, A_3 = \begin{bmatrix} 19 \\ 60 \\ 88 \\ 129 \\ 155 \end{bmatrix}, A_4 = \begin{bmatrix} 17 \\ 44 \\ 85 \\ 112 \\ 153 \end{bmatrix}, A_5 = \begin{bmatrix} 69 \\ 110 \\ 137 \\ 177 \end{bmatrix}, A_6 = \begin{bmatrix} 93 \\ 134 \\ 162 \end{bmatrix},$$

$$A_7 = \begin{bmatrix} 23 \\ 50 \\ 91 \\ 118 \\ 159 \end{bmatrix}, A_8 = \begin{bmatrix} 9 \\ 46 \\ 75 \\ 116 \end{bmatrix}, A_9 = \begin{bmatrix} 32 \\ 72 \\ 101 \\ 140 \\ 171 \end{bmatrix}, A_{10} = \begin{bmatrix} 28 \\ 57 \\ 97 \\ 124 \\ 165 \end{bmatrix}, A_{11} = \begin{bmatrix} 84 \\ 121 \end{bmatrix}$$

Calculate the height and line spacing of each Chinese character in 11 pictures, which are approximately 40, 30 respectively. Through these matrices, you can use MATLAB programming to filter the border of the complete picture, that is, the top and bottom of the picture, thus determining the four pictures.

Table 1 11 lines remaining fragments

All left columns	7	14	29	38	49	61	71	89	94	125	168
All right columns	18	36	43	59	60	74	123	141	145	176	196

(b) The remaining fragments are classified by these data and assigned to each line, which is divided into 11 lines.

(c) The pictures in the line are sorted, and the correct rate after the arrangement is about 70%.

(d) Manual intervention of local errors: Intervention after the results appear in the running process, the text needs to be re-matched after each intervention.

The method is to visually test the results, enter the correct two adjacent numbers, and then the program will rematch.

Rationality evaluation: There are four cut planes in the horizontal and vertical cuts. When matching, it needs to match at the same time. Therefore, the difficulty of matching is increased. The judgment factor of the program is also increased. It is necessary to satisfy more conditions at the same time. Therefore, the error becomes large and requires labor. The intervention adjusts the results and finally determines the integrity and accuracy of the stitching based on the resulting restoration map.

(2) Part 2

(a) Using Matlab programming to find 11 pictures in the first column, then sorting them, reading and binarizing the pictures, and extracting the binarized data in the workspace after running in Matlab to excel. Display the text by reducing the enlarged table, then calculate the word height and line spacing of the Chinese character by manual search, record the 11 picture words and the top and bottom positions of the line and list the matrix, which is recorded as:

$$B_1=\begin{bmatrix}38\\75\end{bmatrix}\ B_2=\begin{bmatrix}37\\62\\100\\125\\164\end{bmatrix},B_3=\begin{bmatrix}61\\99\\126\\175\end{bmatrix},B_4=\begin{bmatrix}13\\46\\73\\109\\135\\171\end{bmatrix},B_5=\begin{bmatrix}88\\120\\146\end{bmatrix},B_6=\begin{bmatrix}34\\66\\92\\130\end{bmatrix},$$

$$B_7=\begin{bmatrix}53\\92\\116\\142\end{bmatrix},B_8=\begin{bmatrix}25\\116\\154\end{bmatrix},B_9=\begin{bmatrix}4\\37\\64\\100\\125\\163\end{bmatrix},B_{10}=\begin{bmatrix}14\\46\\84\\110\\135\\173\end{bmatrix},B_{11}=\begin{bmatrix}18\\56\\94\\132\end{bmatrix}$$

Calculate the height and line spacing of each of the 11 pictures, which are approximately 48 and 37 respectively. Through these matrices, you can use MATLAB programming to filter the border of the complete picture, that is, the top and bottom of the picture, thus determining the four sides of the picture.

Table 2 13 lines remaining fragments

All left columns	19	20	70	81	86	132	146	149	159	171	191	201	208
All right columns	28	31	44	82	109	112	115	127	143	146	147	178	184

(b) Through the above data, the writing baseline of each line (the third line of the English four-line book) can be determined. As a standard, the remaining fragments can be classified and assigned to 11 lines according to the conditions.

(c) The pictures in the line are sorted, and the correct rate after the arrangement is about 70%.

(d) Manual intervention of local errors: Intervention after the results appear in the running process, the text needs to be re-matched after each intervention.

The method is to visually test the results, enter the correct two adjacent numbers, and then the program will rematch.

4.3 Question 3: Mathematical model of shredded paper stitching restoration of double-sided printed documents

- Modeling preparation

Acquisition of distance matrix $d_{i,k}$

If k>i: $d_{i,k}$, the distance between the kth picture on the front side and the left side of the i-th picture is adjacent.

$d_{i,k}$ denotes the distance when the kth picture on the reverse side is adjacent to the left and right of the i-th picture.

If k<i: $d_{i,k}$, the distance between the kth picture on the front side and the i-th picture is adjacent to the top and bottom.

$d_{i,k}$ denotes the distance when the kth picture on the reverse side is adjacent to the i-th picture.

If k = i: $d_{i,k}$ = 0.

The position of the first column of pictures is numbered from 1 to 11 from top to bottom, and the position of the second column of pictures is numbered from top to bottom 12 - 22, ⋯, 198 - 209, recorded as matrix $\begin{pmatrix} 1 & 12 & \cdots & 198 \\ 2 & 13 & \cdots & 199 \\ \vdots & \vdots & \cdots & \vdots \\ 11 & 22 & \cdots & 209 \end{pmatrix}$。

- Establishment of Model 3

The third question is to splicing 1119 pieces of scrap paper into the original picture and considering the problem of the front and back. There are a total of 418 pieces.

Splicing to find the minimum distance as the goal. In the splicing process, the splicing scheme is planned in stages according to the following steps:

1) Calculate the collection of left, right, upper, and lower boundaries according to the principle of more debris blanks at the edges.

2) According to each blank line height of each fragment and the height of each line of text, 418 pieces are divided into 22 groups of 19 groups each.

3) According to the upper and lower boundaries, the standard blank row height of each group and the upper and lower boundaries, the 22 groups are sorted into two groups, the positive and negative groups, and the eleven groups for each major component. They are:

$$L_{正} = \{l_1^0, l_2^0, l_3^0, \cdots, l_{11}^0\}$$

$$L_{反} = \{l_1^1, l_2^1, l_3^1, \cdots, l_{11}^1\}$$

4) Each time the two corresponding pieces in the two rows of L are reversed at the same time, the front side starts to splicing from the left side to the right side, and the reverse side starts to splicing from the right side to the left side, and the figure numbers of the front and back sides are the same, only a , b number is not the same (for example, the first one from the left to the right 23a, the reverse from the right to the first must be placed 23b), the objective function is as follows:

$$\text{Min:} \sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1}^{198} A_{ij} d_{ik} A_{kj+1} + \sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1\cdots209}^{j\neq11\text{的倍数}} A_{ij} d_{ik} A_{kj+1} + \sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1}^{198} A_{ij} d_{ik}^{'} A_{kj+1} + \sum_{i=1}^{209}\sum_{k=1}^{209}\sum_{j=1\cdots209}^{j\neq11\text{的倍数}} A_{ij} d_{ik}^{'} A_{kj+1}$$

- Solution of model three

(1) Manually intervening to observe 209 pieces of debris to find 11 pictures in the first column, then

Read and binarize the picture, record 11 picture words and the top and bottom positions of the line and list the matrix, calculate the height and line spacing of each of the 11 pictures, and use the matrix to select the complete picture by MATLAB programming. The border, that is, the top and bottom of the picture.

Table 3 related data

All left columns	3b	5b	13b	23b	78b	86b	88b	89a	91b	99a	100a	114a	146a	165b	172b	186b	199b					
All right columns	3a	5a	9b	13a	23a	35a	54b	78a	83a	88a	89b	90a	99b	105a	114b	136b	143b	146b	165a	172a	186a	199a

(2) Through the above data, the writing baseline of each line (the third line of the English four-line book) can be determined. As a standard, the remaining fragments can be classified and assigned to 11 lines according to the conditions.

(3) Sorting each image in the line

(4) Manual intervention of local errors: Intervention after the results appear in the running process, the text needs to be re-matched after each intervention.

Grouping 19 lines per line, can be repeated, some pictures are not suitable in any line, they are used as a common collection alone. When they are arranged in a row, if you can't find a suitable one, you still have to find it in this collection. Line 22 is a public collection.

5. Model evaluation and promotion

In this paper, from the aspect of gray value and assignment, several models are built for the splicing problem of fragments. There are advantages and disadvantages.

The advantages of the model in this article:

For the splicing problem of fragments, starting from the pixels of the fragments and transforming the fragments into data matrices for splicing and restoration, we solve the distance matrix between the two fragments by comparing the gray values between the left and right borders of the shredded paper. In this way, it is judged whether the two are adjacent, the problem is digitized, and the expression is clearer.

The shortcomings of the model in this article:

Because the read picture is binarized, the pixel is strictly defined, 1 is white, 0 is black, there is a difference in the calculation, and it is necessary to manually sort the filtered pictures. When there are a lot of shredded paper, the workload is very large, and for the second problem, the picture is not only slit but also cross-cut. After determining the first column, the matrix is selected by the matrix of the word height and the line spacing to filter the fragments of each line. In the screening process, due to blank lines, errors will occur and people need to intervene.

References

[1] Xue Yi., Mathematical Modeling. Beijing, Science Press, 2011. Start page number 243 - end page number 252.

[2] Xianyi, Research on Image Fragment Restoration Method, January 2010.

[3] Zhang Qiang, Wang Zhenglin, Proficient in MATLAB Image Processing (Second Edition), Beijing: Publishing House of Electronics Industry, 2012. Start page number 98 - end page number 102.