Research on Splicing and Restoration of Paper Fragments Based on MATLAB

Yikun Bai, Dopwang Pu*

The School of Science, Tibet University, Lhasa, 850000, China
**Corresponding author: pdw@utibet.edu.cn*

Abstract: *The splicing technology of broken files is a very important and highly applicable technology, which has important applications in the fields of judicial evidence restoration, historical document restoration, and military intelligence acquisition. Traditionally, splicing and restoration work needs to be completed manually, with high accuracy but low efficiency. Especially when the number of fragments is huge, manual splicing is difficult to complete the task in a short period of time. In today's rapidly developing computer technology, it is increasingly necessary to focus on developing automatic splicing technology for shredded paper to improve the efficiency of splicing and restoration. This article delves into this issue in depth, making full use of the powerful mathematical software $Matlab$. Ultimately, a total of 5 pieces of shredded paper, including horizontal and vertical cutting, Chinese and English, and single-sided and double-sided splicing and restoration work, were achieved for three major categories. For shredding paper sheets (only vertically cut) of printed text files on the same page, considering the small degree of paper fragmentation, the method of using software to obtain image data is not considered for now, but the macro stitching method is prioritized. The so-called macro stitching method is based on the cutting situation of each vertical cutting line on each line of text, and uses simple and effective methods to solve problems from a macro perspective. Subsequently, the probability of occurrence of concatenated duplicate codes in this method was provided to demonstrate its applicability.*

Keywords: *Macro stitching method, Least squares method, MATLAB, Splicing of shredded paper pieces*

1. Introduction

The splicing of broken files has important applications in the fields of judicial evidence restoration, historical document restoration, and military intelligence acquisition [1]. Traditionally, splicing and restoration work needs to be completed manually, with high accuracy but low efficiency. Especially when the number of fragments is huge, manual splicing is difficult to complete the task in a short period of time. With the development of computer technology, people are trying to develop automatic splicing technology for shredded paper to improve the efficiency of splicing and restoration[2]. This article mainly establishes a model and algorithm for the splicing and restoration of shredded paper sheets (only vertically cut) from the same printed text file in a shredder[3]. The fragmented data of one page of Chinese and one page of English files are spliced and restored. If the restoration process requires manual intervention, study the intervention methods and time nodes. The restoration results are presented in the form of images and tables. For the situation where the shredder cuts both vertically and horizontally, design a model and algorithm for splicing and restoring shredded paper, and perform splicing and restoration on the fragmented data of one page of Chinese and one page of English files.

The fragmented data provided in this article are all single-sided printed files. From a practical perspective, there may also be issues with the splicing and restoration of shredded paper in double-sided printed files that need to be addressed. Fragmented data for a double-sided printed document with one page of English printed text. Design corresponding models and algorithms for splicing and restoring shredded paper, and provide splicing and restoration results for fragmented data. (https://www.mcm.edu.cn/problem/2013/2013.html)

2. Analysis of paper fragments from shredders of printed text files on the same page

For a given shredder that breaks paper sheets (only vertically) from the same printed text file, considering the small degree of paper fragmentation, we do not consider using software to obtain image data for now. Instead, we prioritize using the cutting situation of each vertical cutting line for each line of text as a basis, and solve the problem from a macro perspective using simple and effective methods. And provide the probability of occurrence of concatenated duplicate codes in this method to demonstrate its applicability. Due to the fact that this method obtains data from a macro perspective, it is temporarily referred to as the macro stitching method.

In daily life and study, the following rules can be found to play a very important role in fragment splicing and reorganization, and are also the main theoretical basis of the so-called macro splicing method: if there are only a few missing first lines (especially two Chinese characters), and there are no punctuation marks on the left side of the blank part, it is likely to be caused by the indentation of the first line; If there are only a few words in the first line, it is likely due to special formatting issues (such as the title used when writing a letter), and the other possibility is mainly the extension and remainder of the text on the previous page; Excluding the space of the leftmost four characters (two Chinese characters), the more lines there are on each piece of paper, the further left it was originally placed on the paper.

2.1 Model Establishment and Solution

As shown in Figure 1, for some Chinese fragmented data, two sets of vectors can be obtained, one left and one right, based on whether more than one-third of the missing parts are due to cutting (in English, whether the cut letters can be recognized and whether they affect recognition are used as standards). If the degree of loss is not severe, it is marked as 1 if it has not been cut, otherwise it is marked as 0, and spaces are marked as not cut. From this, we get two vectors, namely a_i and b_i.

For English fragmented data, the same approach can be applied. By analyzing the cutting patterns and the recognizability of letters, two sets of vectors are generated to represent the left and right edges of each fragment. These vectors provide a foundation for matching fragments based on their compatibility. To further enhance the accuracy of splicing, additional factors such as font size, line spacing, and alignment are considered. The algorithm prioritizes fragments with higher similarity in their vector representations, ensuring that the most compatible pieces are joined first. This step-by-step process continues until all fragments are successfully reassembled or until manual intervention is required to resolve ambiguities. Through this method, the efficiency and accuracy of document restoration can be significantly improved, offering a practical solution for handling large-scale fragmented data in various fields.

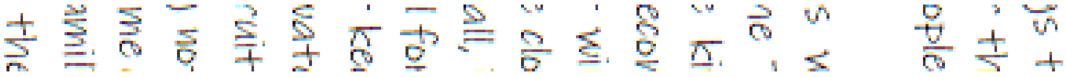


Figure 1 Fragment diagram example

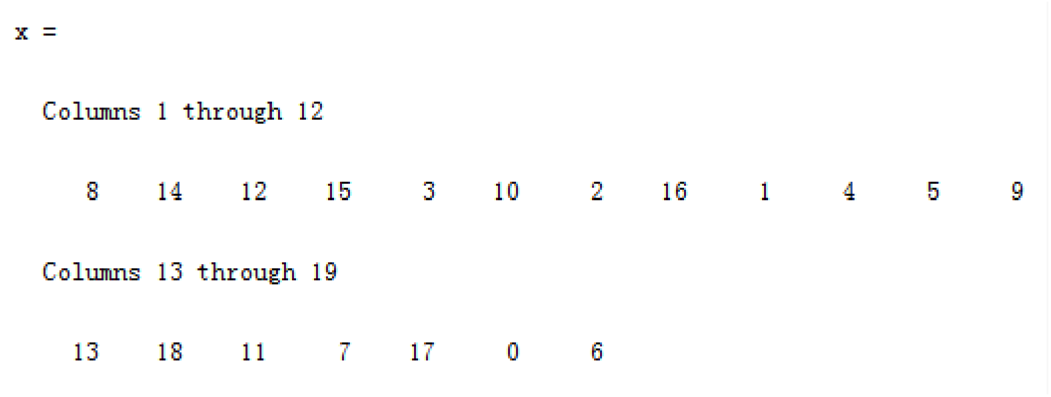
In this way, this article can obtain a_i total of 19 pairs of vectors, a and $b_i (i = 1,2,3,\cdots 19)$, to describe 19 pieces of fragmented data. If the fragmented data labeled i, j can be concatenated, then ab must hold true. Therefore, the algorithm process is as follows:

(1) Filter out a_i whose elements are all 1 and mark the value of i, using this bar fragment as the starting point at the left end;

(2) Using b_i as the "paired gene chain", select the vector a in a that is equal to b_i, mark the value of j, and then the bar fragment with sequence number j can be pieced after the fragment data with sequence number i;

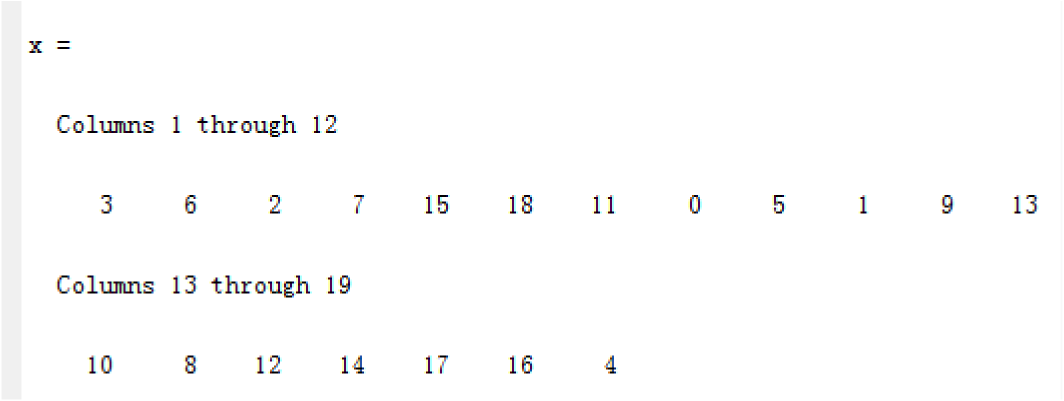
(3) Repeat the above two steps, and so on, until all elements of b_i are 1, at which point the rightmost end of the original paper has been reached.

By programming with $Matlab$ software, the following results can be obtained, as shown in Figure 2 and Figure 3:



```
x =

  Columns 1 through 12

     8    14    12    15     3    10     2    16     1     4     5     9

  Columns 13 through 19

    13    18    11     7    17     0     6
```

Figure 2: Running Results of Chinese Longitudinal Fragment Data Recovery Software



```
x =

  Columns 1 through 12

     3     6     2     7    15    18    11     0     5     1     9    13

  Columns 13 through 19

    10     8    12    14    17    16     4
```

Figure 3: Running Results of the English Longitudinal Fragment Data Recovery Software

2.2 Summary of solution results for longitudinal fragments

The final order of the restored Chinese and English longitudinal fragment numbers is summarized in Table 1:

Table 1 Restoration order of longitudinal fragment numbers in Chinese and English

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Chinese	8	14	12	15	3	10	2	16	1	4	5	9	13	18	11	7	17	0	6
English	3	6	2	7	15	18	11	0	5	1	9	13	10	8	12	14	17	16	4

From the effect of restoring the image, it can be seen that the results obtained by macro stitching method are completely accurate and do not require manual intervention. When there is a concatenation of duplicate codes, it means that the right vector of one fragment has at least two left vectors of other fragments that are equal to it. For 28 lines of Chinese text, the probability of occurrence of splicing duplicate codes is $P_1 = \frac{19}{2^{28}}$, while for 30 lines of English text, the probability of occurrence of splicing duplicate codes is $P_2 = \frac{19}{2^{30}}$. Therefore, it can be seen that macro splicing is an ideal and effective method for handling small sample splicing with light fragmentation.

3. Research on the degree of damage to fragments

Due to the high degree of damage to the fragments, the large number of fragments, and the small number of rows contained in each fragment, the macro stitching method is no longer applicable. At the same time, the method of converting images into data can only be used to read their pixels through software, and *Matlab* is used to process the data. Considering that the characteristics of the obtained data are complex and large in quantity, but with a more detailed characterization of fragmented features, even for the spliceable parts, the two vectors can only guarantee approximate equality and cannot guarantee absolute equality.

If we consider the cutting line, the right vector of a certain fragment, and the left vector of the undetermined piece that can be spliced, we will find that the necessary and sufficient condition for determining whether the undetermined piece can be spliced is that the sum of the distances from the left vector of the fragment and the right vector of the predetermined piece to the cutting line is minimized. This links the stitching problem with the least squares method. Therefore, in solving problem two, the core method selected in this article is the least squares model[4].

In addition, observing the characteristics of the image, it can be seen that the horizontal cutting line runs exactly between two lines of text without damaging the text, that is, two horizontal cutting lines (with other horizontal cutting lines between them) cut the text into a complete strip. Therefore, in the process of stitching and restoring, it is not possible to stitch all the fragments in one step, but rather to first piece them into several strip-shaped texts. Then, through manual intervention, the internal logical relationships of the text are analyzed to obtain a complete restored image[5].

3.1 Model Establishment and Solution

(1) Determination of initial position

To start the stitching work, the first step is to select the basic points, which means selecting a relatively special fragment to fix it, and then starting to search for adjacent images. Due to the absence of text damage on the left and upper ends of the fragments in the upper left corner, the fragment data in the upper left corner is generally selected when selecting the basic point. However, this article takes into account the effectiveness of the adopted method and intentionally avoids weakening the influence of lateral cutting lines. Therefore, when selecting the basic points, the next best option is to choose the fragment data that may appear at the leftmost end, which has the feature of $a_i = 255\alpha$ (a full 1 array of the same dimension). Through $Matlab$ implementation, the fragment number that may be located at the far left of the search is:

A.Chinese text:7,14, 29,38,49,61,62,67,71,80,89,94,125,135,143,168;

B.English text:2,19,20,70, 81,86,132,146,149,159,171,191,201,202,208.

Next, we will conduct the first manual intervention in this article, as shown in Table 2:

Table 2 Explanation of the first manual intervention situation

Intervention time node	Intervention methods	Intervention effect	
Determination of initial position	Manual splicing	Chinese left 5 fragments: 49 is above 61; 168 is above 38; 71 is above 14; 94 on top, 125 in the center, 29 on the bottom; 7 is above 89	Fragment not on the left end:62,67, 80,135,143

(2) Establishment of Splicing Model

According to the principle of least squares, a_j concatenation model is established as follows: assuming the basic point number is i, the right-hand vector b_i can be obtained. If for the left-hand vector a of the remaining fragment, there is:

$$X = b_i - a_j \tag{1}$$

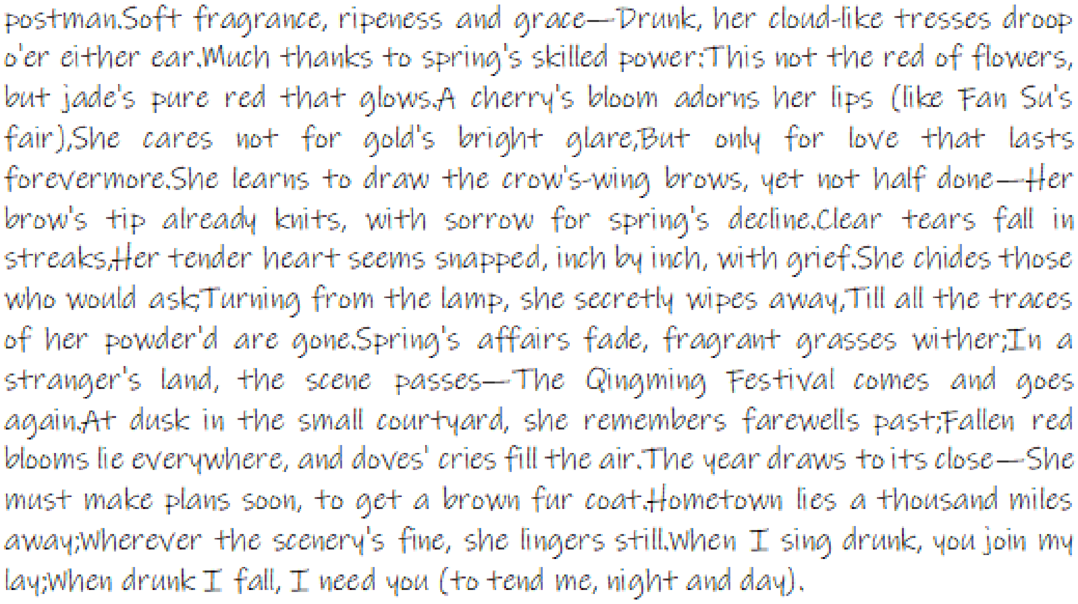
$$s.t. for \forall k \neq j, \text{Always have}: \left|X\right| \leq \left|b_j - a_k\right| \tag{2}$$

The number of the spliceable fragments to the right of index i is j and can be used as new basic points, and so on.

(3) The generation and sorting of strip text

When using *Matlab* implementation, what is obtained is not the final complete puzzle result, but several strip-shaped texts. The following is an example of a Chinese file(For ease of reading, the content has been translated into English):

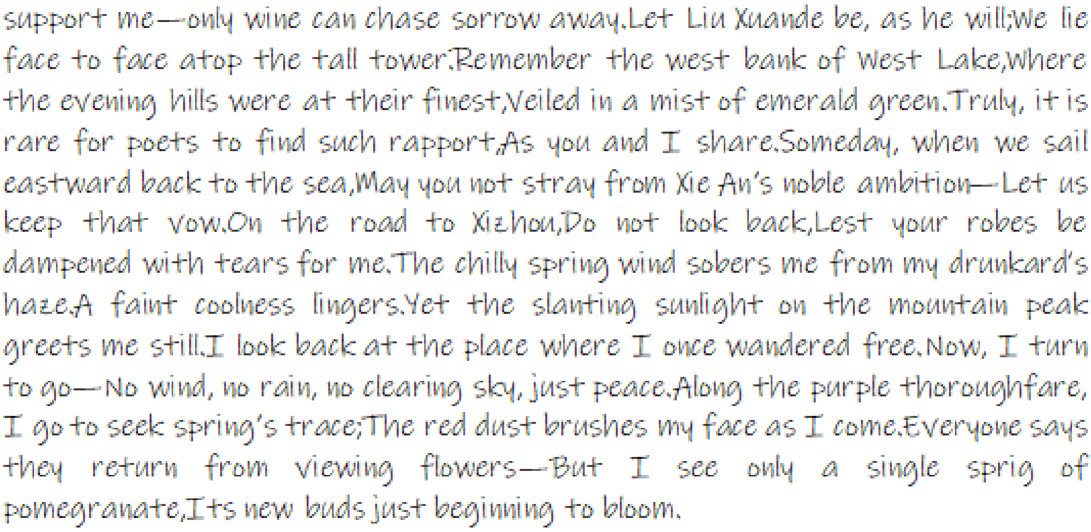
a) The first strip-shaped text generated by MM, with 49 and 61 on the left, as shown in Figure 4:



postman.Soft fragrance, ripeness and grace—Drunk, her cloud-like tresses droop o'er either ear.Much thanks to spring's skilled power:This not the red of flowers, but jade's pure red that glows.A cherry's bloom adorns her lips (like Fan Su's fair),She cares not for gold's bright glare,But only for love that lasts forevermore.She learns to draw the crow's-wing brows, yet not half done—Her brow's tip already knits, with sorrow for spring's decline.Clear tears fall in streaks,Her tender heart seems snapped, inch by inch, with grief.She chides those who would ask;Turning from the lamp, she secretly wipes away,Till all the traces of her powder'd are gone.Spring's affairs fade, fragrant grasses wither;In a stranger's land, the scene passes—The Qingming Festival comes and goes again.At dusk in the small courtyard, she remembers farewells past;Fallen red blooms lie everywhere, and doves' cries fill the air.The year draws to its close—She must make plans soon, to get a brown fur coat.Hometown lies a thousand miles away;Wherever the scenery's fine, she lingers still.When I sing drunk, you join my lay;When drunk I fall, I need you (to tend me, night and day).

Figure 4: The first strip of text generated by Matlab

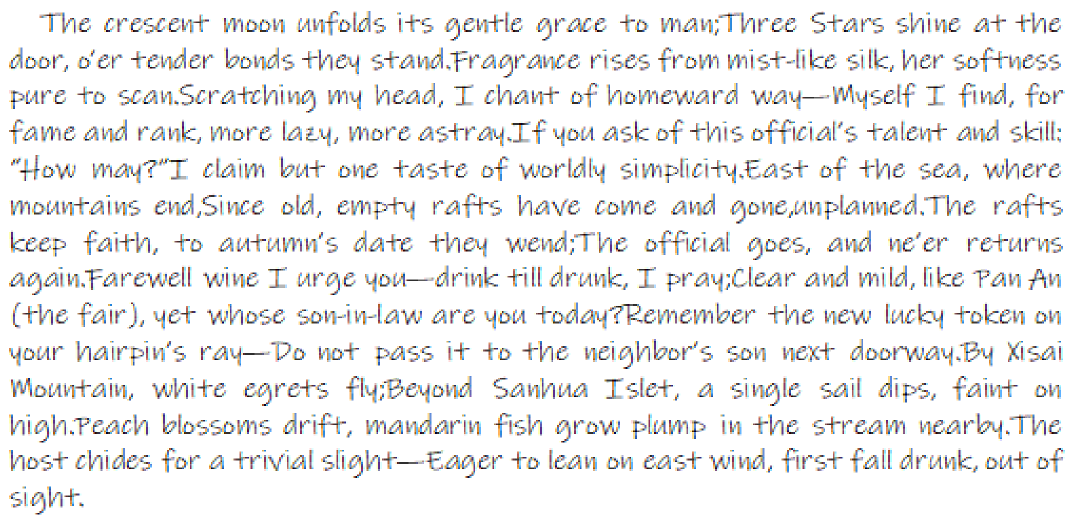
b) The second strip-shaped text generated by MM, with 168 and 38 on the left, as shown in Figure 5:



support me—only wine can chase sorrow away.Let Liu Xuande be, as he will;We lie face to face atop the tall tower.Remember the west bank of West Lake,Where the evening hills were at their finest,Veiled in a mist of emerald green.Truly, it is rare for poets to find such rapport,As you and I share.Someday, when we sail eastward back to the sea,May you not stray from Xie An's noble ambition—Let us keep that vow.On the road to Xizhou,Do not look back,Lest your robes be dampened with tears for me.The chilly spring wind sobers me from my drunkard's haze.A faint coolness lingers.Yet the slanting sunlight on the mountain peak greets me still.I look back at the place where I once wandered free.Now, I turn to go—No wind, no rain, no clearing sky, just peace.Along the purple thoroughfare, I go to seek spring's trace;The red dust brushes my face as I come.Everyone says they return from viewing flowers—But I see only a single sprig of pomegranate,Its new buds just beginning to bloom.

Figure 5: The second strip-shaped text generated by Matlab

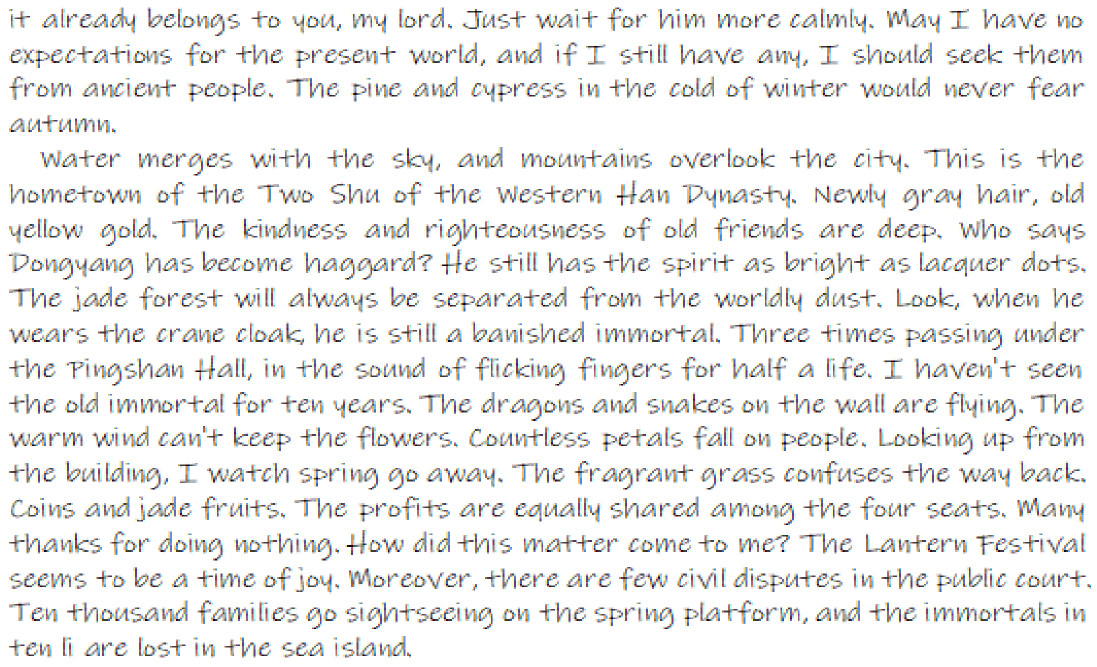
c) The third strip of text generated by MM, with 71 and 14 on the left, as shown in Figure 6:



The crescent moon unfolds its gentle grace to man;Three Stars shine at the door, o'er tender bonds they stand.Fragrance rises from mist-like silk, her softness pure to scan.Scratching my head, I chant of homeward way—Myself I find, for fame and rank, more lazy, more astray.If you ask of this official's talent and skill: "How may?"I claim but one taste of worldly simplicity.East of the sea, where mountains end,Since old, empty rafts have come and gone,unplanned.The rafts keep faith, to autumn's date they wend;The official goes, and ne'er returns again.Farewell wine I urge you—drink till drunk, I pray;Clear and mild, like Pan An (the fair), yet whose son-in-law are you today?Remember the new lucky token on your hairpin's ray—Do not pass it to the neighbor's son next doorway.By Xisai Mountain, white egrets fly;Beyond Sanhua Islet, a single sail dips, faint on high.Peach blossoms drift, mandarin fish grow plump in the stream nearby.The host chides for a trivial slight—Eager to lean on east wind, first fall drunk, out of sight.

Figure 6: The third strip-shaped text generated by $Matlab$

d) The fourth strip of text generated by $Matlab$, with 94125 and 29 on the left, as shown in Figure 7:

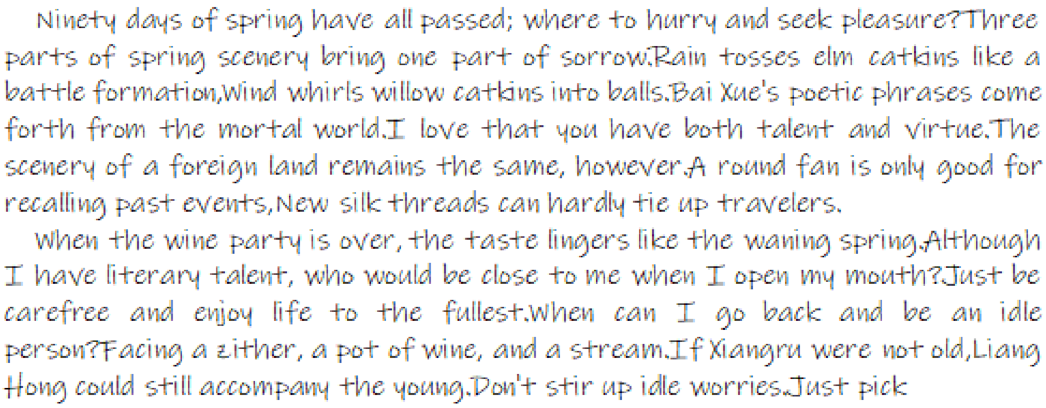


it already belongs to you, my lord. Just wait for him more calmly. May I have no expectations for the present world, and if I still have any, I should seek them from ancient people. The pine and cypress in the cold of winter would never fear autumn.

Water merges with the sky, and mountains overlook the city. This is the hometown of the Two Shu of the Western Han Dynasty. Newly gray hair, old yellow gold. The kindness and righteousness of old friends are deep. Who says Dongyang has become haggard? He still has the spirit as bright as lacquer dots. The jade forest will always be separated from the worldly dust. Look, when he wears the crane cloak, he is still a banished immortal. Three times passing under the Pingshan Hall, in the sound of flicking fingers for half a life. I haven't seen the old immortal for ten years. The dragons and snakes on the wall are flying. The warm wind can't keep the flowers. Countless petals fall on people. Looking up from the building, I watch spring go away. The fragrant grass confuses the way back. Coins and jade fruits. The profits are equally shared among the four seats. Many thanks for doing nothing. How did this matter come to me? The Lantern Festival seems to be a time of joy. Moreover, there are few civil disputes in the public court. Ten thousand families go sightseeing on the spring platform, and the immortals in ten li are lost in the sea island.

Figure 7: The fourth strip-shaped text generated by $Matlab$

e) The third strip of text generated by $Matlab$, with 7 and 89 on the left, as shown in Figure 8:



Ninety days of spring have all passed; where to hurry and seek pleasure?Three parts of spring scenery bring one part of sorrow.Rain tosses elm catkins like a battle formation,Wind whirls willow catkins into balls.Bai Xue's poetic phrases come forth from the mortal world.I love that you have both talent and virtue.The scenery of a foreign land remains the same, however.A round fan is only good for recalling past events,New silk threads can hardly tie up travelers.

When the wine party is over, the taste lingers like the waning spring.Although I have literary talent, who would be close to me when I open my mouth?Just be carefree and enjoy life to the fullest.When can I go back and be an idle person?Facing a zither, a pot of wine, and a stream.If Xiangru were not old,Liang Hong could still accompany the young.Don't stir up idle worries.Just pick

Figure 8: The fifth strip-shaped text generated by *Matlab*

By analyzing the relationship between the five banded texts, it can be concluded that "When drunk I fall, I need you (to tend me, night and day)-support me" is a whole sentence that cannot be separated, so regions a and b are adjacent. "Qie Zhe" is seen in "Qie Zhe Shuang Rui Jin Yu Pei", and "Bian You（postman）" is seen in "Shy to Ask Bian You". Therefore, e is not before "a", and "a" can only be used as the first banded text, while "e" can only be used as the last banded text. The strip-shaped text after b should include the beginning of the paragraph, so the five strip-shaped order can be obtained as: a, b, c, d, e. Furthermore, the serial numbers of the fragments restored in Attachment 3 can be obtained, as shown in Table 3:

Table 3 Restoration numbers of single-sided horizontal and vertical cut Chinese fragments

049	054	065	143	186	002	057	192	178	118	190	095	011	022	129	028	091	188	141
061	019	078	067	069	099	162	096	131	079	063	116	163	072	006	177	020	052	036
168	100	076	062	142	030	041	023	147	191	050	179	120	086	195	026	001	087	018
038	148	046	161	024	035	081	189	122	103	130	193	088	167	025	008	009	105	074
071	156	083	132	200	017	080	033	202	198	015	133	170	205	085	152	165	027	060
014	128	003	159	082	199	135	012	073	160	203	169	134	039	031	051	107	115	176
094	034	084	183	090	047	121	042	124	144	077	112	149	097	136	164	127	058	043
125	013	182	109	197	016	184	110	187	066	106	150	021	173	157	181	204	139	145
029	064	111	201	005	092	180	048	37	75	55	44	206	010	104	098	172	171	059
007	208	138	158	126	068	175	045	174	000	137	053	056	093	153	070	166	032	196
089	146	102	154	114	040	151	207	155	140	185	108	117	004	101	113	194	119	123

Similarly, a list of restored sequence numbers for English fragmented documents can be obtained, but due to the occurrence of overlapping codes, manual intervention is required. The time node for this intervention is determined by the sequence number, and the intervention method is manual splicing. Finally, the sequence number list of the restored English fragmented document is shown in Table 4:

Table 4 Restoration numbers of single-sided horizontal and vertical cut Chinese fragments

191	075	011	154	190	184	002	104	180	064	106	004	149	032	204	065	039	067	147
201	148	170	196	198	094	113	164	078	103	091	080	101	026	100	006	017	028	146
086	051	107	029	040	158	186	098	024	117	150	005	059	058	092	030	037	046	127
019	194	093	141	088	121	126	105	155	114	176	182	151	022	057	202	071	165	082
159	139	001	129	063	138	153	053	038	123	120	175	085	050	160	187	097	203	031
020	041	108	116	136	073	036	207	135	015	076	043	199	045	173	079	161	179	143
208	021	007	049	061	119	033	142	168	062	169	054	192	133	118	189	162	197	112
070	084	060	014	068	174	137	195	008	049	172	156	096	023	099	122	090	185	109
132	181	095	069	167	163	166	188	111	144	206	003	130	034	013	110	025	027	178
171	042	066	205	010	157	074	145	083	134	055	018	056	035	016	009	183	152	044
081	077	128	200	131	052	125	140	193	087	089	048	072	012	177	124	000	102	115

4. Conclusions

The shredded paper reconstruction problem discussed in this paper is primarily limited to horizontal and vertical cuts. However, in real-world scenarios, achieving perfectly vertical or horizontal cuts is often challenging. Therefore, based on the methodology presented here, the problem of reconstructing diagonally cut shredded paper fragments could be considered. By performing operations such as straightening the diagonal fragments, the problem can be transformed into the horizontal and vertical cut shredded paper reconstruction scenario addressed in this study. Moreover, regarding the issue of reassembling fragments with double-sided printed content, this model exhibits an oversimplification that leads to relatively significant errors and suboptimal applicability.

The shredded paper fragments in this study exhibit very clear residual text edges. However, in practical applications, the edges of mechanically shredded paper fragments typically contain "burrs," which significantly reduce the accuracy of identifying residual text edges. Thus, building upon the methods presented in this paper, image denoising techniques could be implemented to handle more general cutting conditions.

References

[1] Jia Haiyan. Research on Key Technologies for Automatic Splicing of Shredded Paper [D]. National University of Defense Science and Technology, 2005

[2] Tang Qiaoling, Chen Jia. Research on Paper Fragment Splicing and Restoration Technology Based on MATLAB [J]. Science and Technology Innovation, 2021, (18):106-107.

[3] Yu Dan. Research and Implementation of Simulated Paper Fragment Splicing Algorithm Based on Horizontal and Vertical Cutting [D]. Foshan University of Science and Technology, 2021.

[4] Cui Jing. Research on Jaccard based Plant Leaf Image Recognition Method [D]. Lanzhou University, 2021 DOI: 10.27204/d.cnki.glzhu.2021.000738.

[5] Chen Zhigang, Su Zhou, Fang Jia. Paper shredding and splicing algorithm based on triplet network and hybrid particle swarm optimization [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2024, 52 (02): 22-28. DOI: 10.13245/j.gust.240203