# Text Document Fragments Reconstruction Algorithm Based on Human-Computer Interaction

NIU Wen-Ya, WENG Ning-Xin, LI Zhi-Wei*
School of Math and Computer Science
Quanzhou Normal University
Quanzhou, China
* Corresponding author: wei2785801@qztc.edu.cn

**Keywords:** ragment reconstrucion; minesweeper strategy; adjacent distance; human-computer interaction

**Abstract.** This paper proposes a novel human-computer interactive algorithm to reconstruct fragments from a text document paper. Three types of adjacent information, namely, distances based on (1) border pixels matching, (2) baseline matching and (3) letter templates matching, are proposed. Average of these distances is used as the adjacent information for fragment reconstruction. The human-computer interactive algorithm is similar to the strategy of "minesweeper" game, the most valuable adjacent information is selected by computer, and human's responds is used to modify the adjacent information. Experiments on 2013 a MCM contest problem shows that 98% of the adjacent information given by computer is correct, and our algorithm can solve fragment reconstruction fairly well.

## Introduction

Fragment reconstruction is not only interesting as a game, but also important in fields such as finance archaeology and forensics. The work of document fragment reconstruction by hand is not only inefficient, but also difficult especially when a large amount of fragments are required to be completed in a short time. Therefore, a computer algorithm is essential to reduce human labors and improve reconstruction efficiency.

Many computer-aid methods have been proposed by some scholars to deal with the reconstruction of text document fragments in the past few years. Just like the methods from [1-3], some authors try to solve this problem in two steps, text fragments from a same row are first selected and reconstructed according to the row adjacent information, (horizontal adjacent information is more accurate than that from vertical), then the row fragments is assembled into a single column in the second step to finish the reconstruction. Such methods don't fully utilize the adjacent information, since only one type of adjacent information is used in each step. Consequently, the result of such methods is not completely correct. And the correctness of such result usually require much human load.

Some artificial intelligent methods, such as Genetic algorithm [4-5] and ant colony algorithm [6-7], are applied for fragment construction. When the number of fragments is large, the computation time is usually unacceptable, and there might be some mismatching fragments in their results because locally optimal solution, instead of global optimal solution, is often achieved by these methods.

Some fragment construction methods are based on human-computer interaction algorithms [8-9]. However, these methods didn't fully utilize the adjacent information, they are still rather inefficient to solve the problem of constructing text document fragments.

Inspired by the methods mentioned above, a human-computer interactive algorithm is proposed in this paper. Our work is briefly summarized in Fig.1. And the rest of this paper is organized as follows. Section 2 describes the definition of three types of distance matrix on adjacent information. In Section 3, we introduce the strategy of "minesweeper", and present a human-computer interactive scheme to reconstruct the shredded document. Experimental results of this algorithm are provided in Section 4. Finally, in Section 5, we conclude this paper.

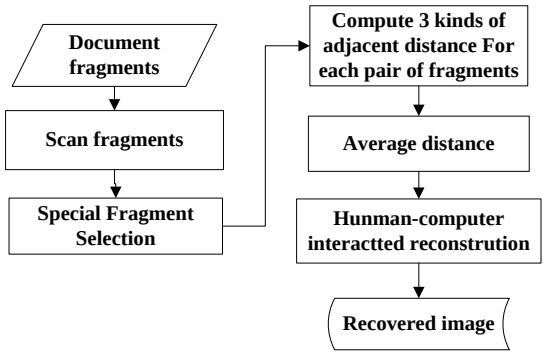



Figure 1. The proposed text document fragment reconstruction flow chart. Three kinds of adjacent distance and their average, which indicates two pieces of fragments is adjacent or not, are evaluated. These distances are then used in performing the computer-aid reconstruction.

## Pariwise Relation between Fragmetns

To calculate the pairwise matching degrees to which a pair of fragments adjoins each other, is the most important step in fragment reconstruction. Such pairwise relation can be expressed as a distance metric; whose elements evaluate the degree of adjacency between any two fragments. For such distance-based matching, a lower score on the matching metric generally means a higher possibility of being an adjacent pair of fragments. It should be noted that there exist two kinds of adjacency, namely, horizontal and vertical adjacency. In the following paper, we call it type-1 and type-2 adjacency. Accordingly, two matrixes are needed to store these 2 types of matching degrees. And it should also be noted that these matrix are generally not symmetric. Since there are two permutations, two kinds of distances should be computed between any pair. Namely, the distance of from $F_i$ to $F_j$, and the distance of from $F_j$ to $F_i$.

Existing distance computation usually consider geometry-based pairwise matching, which relies on matching the curve contours of the boundary, or color-based pairwise matching, which analysis the color information on the boundary. However, these methods cannot be applied in our problem because all of fragments have rectangular boundaries; there is few color information on the boundary since most of documents is usually printed black and white.

For such reasons, we propose three approaches for the adjacency information computation in this work. One is to use the matching degree of left-most (down-most) and right-most (up-most) border of the fragments. Another is to calculate the matching degree of baseline of each fragment. And the other is to judge whether a letter appears on the borders of a pair of fragments.

### *A. Special Fragment Selection*

Two types of special fragment are different from the majority of other fragments in a shredded document, namely, the leftmost and rightmost fragments [10-11].The leftmost fragment is characterized by a considerable blank area in its left border, and a little texture on its right border. Similarly, the rightmost fragment is characterized by the large area of blank on its right border, and little texture on its left. Such fragments can be easily selected if one examines the gray levels of the pixels near the border. To improve the computation efficiency, these border fragments are singled out. For a leftmost fragment from a document paper, no fragment should be adjacent to it from the left. Similarly, no one should be adjacent to a rightmost fragment from the right. Fig.2 shows some typical examples of these fragments from the border of a document paper.

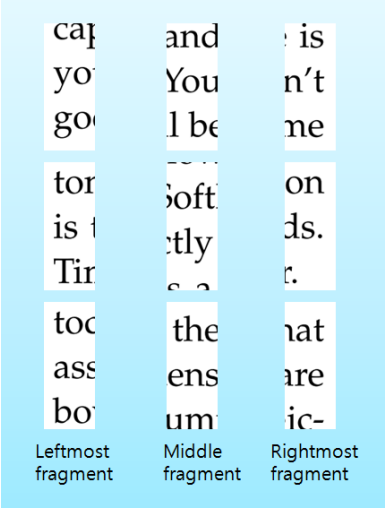



Figure 2. illustration of fragments on the border. From the left to the right: the leftmost fragment, the ordirnary fragment, the rightmost fragment.

### *B. Pairwise Distance Base on Border Pixels Matching*

The matching degree of gray levels from the borders of a pair of fragments is one of the most evident information to judge whether this pair of fragments is adjacent. Use the Euclidean distance of gray-level vectors, the type-1 (horizontal) distance of $F_i, F_j$ is defined as

$$Dist^{(1)}(1,i,j) = \left\| B_{right}(i) - B_{left}(j) \right\| \qquad (1)$$

where $B_{right}(i), B_{left}(j)$ are grey-level vectors of the right border of $F_i$, and left border of $F_j$, respectively; $\|\ \|$ means the Euclidean distance[12].

Similarly, the type-2 distance of (vertical) distance of $F_i, F_j$ is defined as

$$Dist^{(1)}(2,i,j) = \left\| B_{down}(i) - B_{up}(j) \right\| \qquad (2)$$

where $B_{down}(i), B_{up}(j)$ are grey-level vectors of the down and up border of $F_i$ and $F_j$, respectively.

### *C. Pairwise Distance Base on Baseline Matching*

The horizontal projection method, which is frequently used in many references [13], finds the baseline according to the local maxima of histogram obtained from a horizontal projection of fragments. But when the width of a rectangle fragment is short, information in such histogram may be incomplete. Thus we suggest using a letter matching method to find the baseline of a fragment. Such measurement procedure is illustrated in Fig.3. Repeatedly change the position of any given letter template, once the template overlaps a same letter, the baseline is determined by the y-axis of the template.

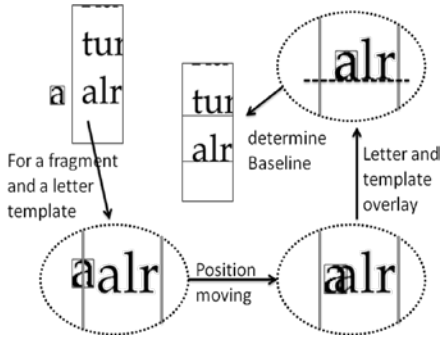



Figure 3. illustration of the measure of baseline. Repeatly move the letter template, if this template is overlay on a same letter, stop iteration and output the position of the template as the position of baseline.

Type-1 and type-2 baseline distances of $F_i, F_j$ is defined as

$$Dist^{(2)}(1,i,j) = \min_{k \in \mathbb{Z}} \left( \left| x(i) - x(j) + k \cdot d_{skip} \right| \right) \qquad (3)$$

and

$$Dist^{(2)}(2,i,j)=\min_{k\in\mathbb{Z}}\left(\left|x(i)+height-x(j)+k\cdot d_{skip}\right|\right) \quad (4)$$

Where $d_{skip}$ the baseline skip of the text document, height is is refers to the height of a fragment (unit: pixels).

### *D. Pairwise Distance on Letter Template Matching*

Though $Dist^{(2)}(k,i,j)$ is always near zero when $F_i$ and $F_j$ are adjacent, but the reverse is not true. For example, when $F_i$ and $F_j$ are from a same row, their baseline distance $Dist^{(2)}(k,i,j)$ is still near zero. Letter template matching is an effective way to avoid such instances. As illustrated in Fig.4, $F_i$ and $F_j$ are probably horizontal adjacent, since half of a letter "A" appears on right part of $F_i$, and the other half appears on the left part of $F_j$. We call it, $F_i$ horizontally matches $F_j$ via template "A". Thus Type-1 and type-2 letter template distances of $F_i, F_j$ is defined as

$$Dist^{(3)}(1,i,j)=\begin{cases}0, F_i \text{ horizontally matches } F_j \text{ via a template}\\ 0, \text{otherwise}\end{cases} \quad (5)$$

And

$$Dist^{(3)}(2,i,j)=\begin{cases}0, F_i \text{ vertically matches } F_j \text{ via a template}\\ 0, \text{otherwise}\end{cases} \quad (6)$$

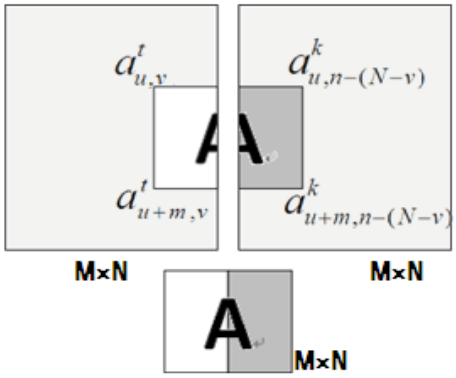



Fig.4 illustration of baseline matching is flawed. $F_i$ and $F_j$ are from a same row, their baseline distance $Dist^{(2)}(k,i,j)$ is still near zero, but $F_i$ and $F_j$ are not horizontal adjacent.

### *E. Average of Pairwise Distacne*

The common average function

$$Dist(k,i,j)=\sum_{t=1}^{3}\frac{Dist^{(t)}(k,i,j)}{3} \quad (7)$$

is used to utilize the advantages of these types of distances functions in our experiment, where $Dist^{(t)}(k,i,j)$ is calculated by (1) to (6).

## Text Document Fragment Reconstruction

The fragment reconstruction based on human-computer interaction in this paper is inspired by the game strategy of minesweeper.

### *F. Game Strategy of Minesweeper Game*

Minesweeper game is one of the most classic computer games of all time. It has been installed on all windows PCs since the days of windows 95 and is one of the simplest, most fun games one can play.

The minesweeper game rule is simple, to clear a rectangular board as fast as possible without detonating any hidden "mines" among the board. Click the left mouse button if you are sure the lattice contains a mine; otherwise click the right mouse button, and this lattice will open, the number of mines neighboring this lattice will be shown on this lattice. These numbers serve as clues to help one select and click the next lattice. Repeatedly use these clues to judge whether a lattice contains a

mine, one will win eventually with great probability. A simple example is shown in Fig. 5. Such strategy can be summarizing as follow:

- Step1: Use the numbers and mines to judge whether a neighboring lattice is a mine;
- Step2. Left-click the lattice to get more number if one is sure it is not a mine; Right-click the lattice to denote a mine, if one is sure it is a mine.
- Step3: go to Step1; until all the mines have been swept.

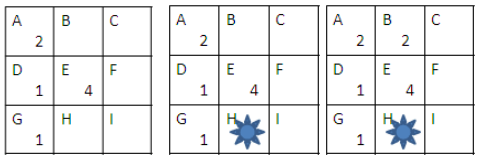



Figure 5. minswepper stragege. from the numbers in Lattice D and G, we can infer that lattice H is mine (Fig.b). Use this new information and the number in D, us can confirm B is a number. Click B, we got a "2"(Fig.c), which enables us to know C, F and I are all mines if the number in E is properly considered.

### *G. Human-Computer Interaction Algorithm*

This section proposes a novel algorithm similar to the strategy of "minesweeper" to implement the reconstruction of fragments, by the help of a distance matrix we created in Section 2.

The row and column distance matrix $Dist(k,i,j)$ are similar to the "numbers" in "minesweeper", they provides information about the likelihood of horizontal of vertical adjacency for each pair of fragments $(i,j)$, and the pair which is most likely to be adjacent is then shown on the screen, waiting for human to determine whether it is really type-*k* adjacent. Such algorithm can be summarized by the following steps:

- step 1: Initiates $Confirm(k,i,j)=0$, $Block(i)=\{F_i\}$;
- step 2: for every $(k,i,j)$ such that
$$Confirm(k,i,j)=1$$
(In such case, $F_i, F_j$ is already type-*k* adjacent), Find $(K,I,J)$,
Where $(K,I,J)=\arg\min_{(k,i,j)} Dist(k,i,j)$
- step 3: On a screen, show $F_I, F_J$ (type-*K* adjacent), to make human determines the correctness of this adjacency, $Block(I)$ and $Block(J)$ are shown simultaneously.
- step 4: If human determines $F_I, F_J$ are not type-*K* adjacent, Let $Dist(K,I,J)=\inf$. Else if human determines $F_I, F_J$ are really type-*K* adjacent, Let
$$Dist(K,I,J)=0,\quad Confirm(K,I,J)=1,$$
$$Dist(K,I,j)=\inf,\quad \text{for all } j\neq J,$$
$$Dist(K,i,J)=\inf,\quad \text{for all } i\neq I,$$
- step 5: Go to step 2, until all $Confirm(k,i,j)=1$.

## Experiments

In order to validate the proposed algorithm, we use the database of shredded documents in the problem of 2013 China Mathematical Contest. 209 fragments from a text document have a common size of $180\times72$ pixel. Fig.6 shows how the fragments are concatenated during several steps in the process of construction. After 243 times of judgments, the reconstruction is finished (as shown in Fig. 7), 95.5% of the adjacent pair given by computer is correct.

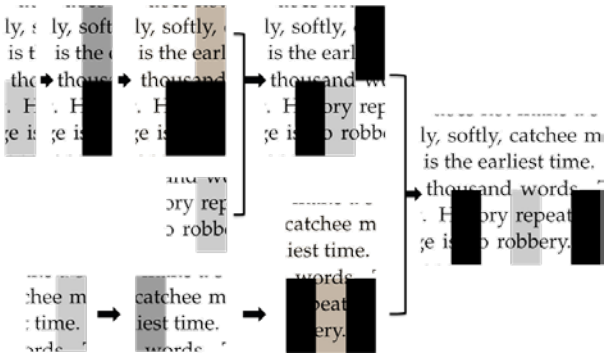
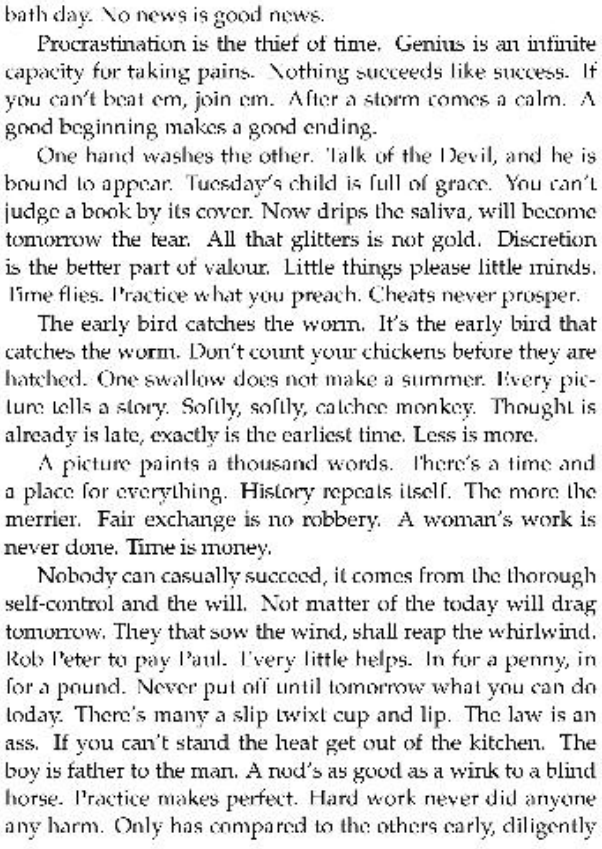


Figure 6. Some steps in the concatenation process. The fragments highlighted are the current matching pair suggested by computer. All of the pairs suggested by computer in these steps are correct.

## Discussions and Conclusions

Adjacent Information is the most important step in fragment construction. So long as all adjacent distances are correctly given, this problem can be solved by many methods efficiently. Unfortunately, it is usually hard to find the adjacency distances, especially when the fragments are small. In order to find as much information as possible, three types of adjacent distances are defined in this paper, and human response are added to modify the adjacent information in our algorithm. The most possible adjacent pair is provided by Computer, while Human's response is integrated in the algorithm to modify adjacent information so that computer can find a better adjacent pair. And experiment result shows that such scheme can solve this problem fairly well. However, our scheme has it limit when the shape of fragments is irregular. And adjacent distance based on letter template matching is often computationally intensive.



bath day. No news is good news.

Procrastination is the thief of time. Genius is an infinite capacity for taking pains. Nothing succeeds like success. If you can't beat em, join em. After a storm comes a calm. A good beginning makes a good ending.

One hand washes the other. Talk of the Devil, and he is bound to appear. Tuesday's child is full of grace. You can't judge a book by its cover. Now drips the saliva, will become tomorrow the tear. All that glitters is not gold. Discretion is the better part of valour. Little things please little minds. Time flies. Practice what you preach. Cheats never prosper.

The early bird catches the worm. It's the early bird that catches the worm. Don't count your chickens before they are hatched. One swallow does not make a summer. Every picture tells a story. Softly, softly, catchee monkey. Thought is already is late, exactly is the earliest time. Less is more.

A picture paints a thousand words. There's a time and a place for everything. History repeats itself. The more the merrier. Fair exchange is no robbery. A woman's work is never done. Time is money.

Nobody can casually succeed, it comes from the thorough self-control and the will. Not matter of the today will drag tomorrow. They that sow the wind, shall reap the whirlwind. Rob Peter to pay Paul. Every little helps. In for a penny, in for a pound. Never put off until tomorrow what you can do today. There's many a slip twixt cup and lip. The law is an ass. If you can't stand the heat get out of the kitchen. The boy is father to the man. A nod's as good as a wink to a blind horse. Practice makes perfect. Hard work never did anyone any harm. Only has compared to the others early, diligently

Figure 7. The text document after 243 times of judgments, among which 95.5% of adjacent pairs given by computer is correct.

**Acknowledgment**

This work is sponsored by the National Undergraduate Training Programs for Innovation and Entrepreneurship Project(121301045), and scientific research project of quanzhou normal university for undergraduates.

Corresponding author: Li Zhi-wei.

**References**

[1] LIU Men-Juan. Reconstruction of Ripped-up Documents Based on Clustering Analysis and Grey Value Matching [J]. Value Engineering, 2013(32): 209-211

[2] WANG Chen &ZEN Jan. The research on the reversion of paper fragments which are in paper shredder [J]. Popular Science. 2014(16): 15-17.

[3] YU Xiang &XIAO Xiang& LI Lu & XU Bo-Sheng &GU Xi. Splicing and Recovery of Scrapped Paper with Longitudinal Cutting Based on Image Gray Value[J]. Journal of Shanghai University of Engineering Science , 2014,28(3):266-269.

[4] PAN Bin & GUO Xiao-Ming &CHEN Ming-Ming & YU Jing-Xian &ZHAO Xiao-Ying &CHEN Wei. Reconstruction of Regular Ripped-Up Documents [J]. Joumal of Liaoning Shihua University, 2014(5):70-78

[5] SU Xiao-Peng& YANG Xi-Yang. A GA Based Automatic Stitching Technology of Strip-shaped Piece [J]. Journal of Sanming University, 2014, 31(2): 39-42

[6] LI Yi-Ying. Double sided stitching recover of text fragments based on ant colony algorithm [J]. Science & Technology Vision, 2013 (28):11

[7] YANG Ling & WANG Lin-Lin &LIU Chong-Chong& SU Si-Mei. Chopped paper Splicing Restoration Model Based on SACO Algorithm [J]. Journal of Taiyuan Normal University Natural Science Edition. Vol.12 No. 4Dec.2013 : 65-68.

[8] LI Xuan-Jie. Scraps of paper splicing recovery based on human-computer interaction interface [J]. Wireless technology. 2013(12): 87-89

[9] LU Jia-Qi. Algorithm for torn paper restoration based on character information [J]. Modern Electronics Technique. 2014(4) :28-31

[10] ZHANG Guo-Lin. Research on Scraps Paper Stitching Recovery Model Based on Chinese Character Recognition [J]. Science Mosaic 2014 (1) : 62-64.

[11] LIU Ci-De& MIAO Nan-Qian& CHANG Qing &WANG Zhi-Peng. The improved standard fragments spliced recovery algorithm [J]. Journal of Nanyang Normal University, Vol.13 No. 3Mar. 2014 : 22-24

[12] BI Kai-Ming. Construction of Mathematical Model of Splicing Scrap Recovery [J]. Value Engineering 2014, 33(25) : 238-239

[13] Huei-Yung Lin& Wen-Cheng Fan-Chiang. Reconstruction of shredded document based on image feature matching [J]. Expert Systems with Applications 2012(39): 3324–3332