A Pipeline for Reconstructing Cross-Shredded English Document

Guanghao Chen, Jue Wu
School of Computer Science and Engineering
Northeastern University
Shenyang, China
e-mail: guanghaochen@stumail.neu.edu.cn

Cundi Jia, Yunzhou Zhang*
College of Information Science and Engineering
Northeastern University
Shenyang, China
e-mail: zhangyunzhou@ise.neu.edu.cn

***Abstract*—Document shreds reconstruction is of great significance in the fields of file confidentiality, anti-disclosure, and investigative science. In this paper, a complete and practical pipeline is designed to reconstruct cross-shredded English documents. The pipeline firstly classifies the shreds into several clusters based on an improved K-means algorithm to reduce clustering imbalance. Especially, a preprocessing is needed before extracting feature vector for shredded English document because of the unaligned characters. Owing to its successful performance in reconstructing the strip-shredded documents, Hungarian algorithm is applied into the permutation for the cross-shredded shreds in the same row. Eventually the location of the connective horizontal paper slips are arranged by considering the complementary relationship of edge vectors between two neighboring shreds. Reconstruction experiment results indicate that the designed pipeline can acquire high precision and efficiency.**

Keywords-cross-shredded document; reconstruction; improved K-means algorithm; hungarian algorithm; pipeline

I. INTRODUCTION

Ever since the paper shredder was invented, people have worked on sticking the pieces back together again. Recently, techniques permitting the purely electronic storage and transmittal of sensitive documents have been developed, however, because of convenience or for legal reasons, many sensitive documents are still printed and eventually shredded. Traditionally, the reconstructing works have to be completed by hand, which could reach a high accuracy but low efficiency. Especially, with a large amount of cutting shreds, manual work seems to be an impossible mission. With the rapid development of computer technology, ones try to develop more automatic fragment assembly technologies to complete the reconstruction task effectively and accurately. The field of document reconstruction can be classified into various subdomains including the restoration of hand torn papers [1], [2], the reconstruction of strip-shredded documents [3]-[5], and the cross-shredded ones [6]-[12]. Nowadays, most documents are mechanically shredded rather than hand torn, and thus the strip-cut and cross-cut variants are major concerns. Some researches have been done on solving the problem of reconstructing strip-cut documents. Prandtstetter *et al*. [8]-[13] proposed a method that used a specific variable neighborhood search approach to reconstruct documents, which involved the user in the construction process to enhance the results. Ukovich *et al*. [4] proposed an algorithm for the reconstruction of strip-cut shredded documents, paying particular attention to the possibility of using mpeg-7 descriptors. Marques *et al*. [5] used boundary features and utilized the nearest neighbor algorithm to calculate the Euclidean distance between the feature vectors corresponding to the concerned strips. Azzam Sleit [12] proposed a clustering approach to reconstruct cross-cut shredded documents based on a cost function.

In the present work, we focus on designing a complete and practical pipeline to reconstruct the cross-shredded English documents. Our work has the following specific contributions: (1) a complete pipeline is proposed for reconstructing cross-shredded English documents. (2) An improved K-means algorithm is proposed to reduce the clustering imbalance caused by uneven sample distribution among classes. (3) Hungarian algorithm is applied into the permutation for cross-shredded paper slips in the same row based on its significant performance in reconstructing the strip-shredded documents

II. PROPOSED METHODOLOGY

A. Hungarian Algorithm for Strip-Shredded Document

Although the reconstruction of strip-shredded document has been studied extensively, it is still worthwhile to investigate because the successful application of Hungarian algorithm could provide the basis for the cross-cut shredded document reconstruction.

As we know, the repeating emergence of gray value in space forms the image texture, indicating that the variation of gray values between two neighboring pixels should be continuous. In other words, the gray values in the edges between two neighboring shreds would reveal the same feature.

The permuting method is analyzed explicitly. The leftmost shred can be determined firstly by searching for the one has the largest blank space at the left edge. After that, the neighboring shred can be found by searching the one that has the biggest value of edge gray similarity with the leftmost. Then, it is necessary to arrange the locations of the rest shreds. Since the neighboring shreds have the strongest similarities between the contiguous edges based on the similarity of gray value, the needed edge vectors are extracted and the digital images are deemed to the digital matrix in computer view, as shown in Fig. 1. During preprocessing, both the leftmost and rightmost edge gray

978-1-5090-6238-6/17/$31.00 ©2017 IEEE

vectors are extracted for every single paper slip and stored into two columns, i.e., $[L_1 \cdots L_k \cdots L_{19}]$ and $[R_1 \cdots R_k \cdots R_{19}]$. Where L_k is the leftmost edge gray vector of kth shred, and R_k is the rightmost edge gray vector of kth shred.

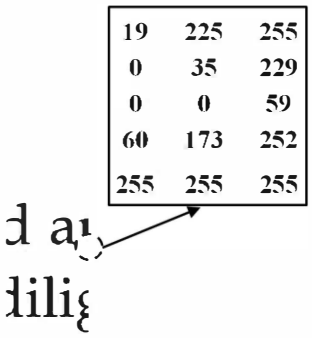

Figure 1. Digital matrix in computer view for digital image.

Then, a method is designed to permute the exact order for the strip-cut documents. Here, Hungarian algorithm for the assignment problem is applied to the reconstruction of strip-shredded documents. For two neighboring shreds, by defining the similarity between the right edge gray vector and left-edge one as the total cost, the goal of our task is to find the permutation result with the lowest cost. The established mathematical model is described as follows.

Equation (1) gives the solution space (x_{ij}) determined by zero or one matrix.

$$x_{ij} = \begin{cases} 1: \text{ assign segment j to be the right side of segment i} \\ 0: \text{ assign segment j not to be the right side of segment i} \end{cases} \quad (1)$$

Thus, target function and constraint conditions are depicted in (2).

$$\min z = \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} x_{ij}$$
$$S.T. \begin{cases} \sum_{i=1}^{n} x_{ij} = 1 \\ \sum_{j=1}^{n} x_{ij} = 1 \\ x_{ij} = 0 \text{ or } 1, i=1,2,..,n \end{cases} \quad (2)$$

where C_{ij} is the cost matrix containing similarities between the rightmost gray vector of shred *i* and the leftmost gray vector of shred *j*. Nevertheless, the values on the diagonal in matrix C_{ij} are meaningless because they represent the similarities of the rightmost and leftmost gray vectors between the same shreds. Therefore, the values on the diagonal are set to be a relatively large number to reduce the possibilities for selecting the same shred as its neighbor.

B. Cross-Shredded Document

1) Classification

It's necessary to propose a proper method to treat with the cross-cut shredded documents because the above permuting method for strip-cut shredded papers cannot be used directly. By observing the paper slip in the shreds pool, in can be noted that the variation of vertical gray value for fragments in the same row are similar, indicating that the shreds can be firstly separated based on the vertical gray value vector. Therefore, in order to classify the shreds into different classes (or rows), a robust feature vector is extracted to enlarge the inter-class variance while reduce the intra-class difference.

The lineation algorithm is proposed to decide the marked locations of character zone and line space, and the vectors composed of these marked locations will be used. The lineation algorithm executes in the following steps: (1) Set the number of vertical pixels as the loop time; (2) Scan each fragment with the pixels in one row. If the portion of white pixels in one row is over the threshold, it is regarded as a blank line or a character zone; (3) Jump to step (1) if the scanner doesn't reach the bottom.

Here, it's necessary to clarify that a threshold is set to judge whether a pixel row is blank spacing or not, and a portion threshold should be set to avoid alienating mistakenly

Unlike Chinese characters, the height of English letter is not uniform such as 'h' with a high crown while 'j' with a long tail, indicating that English letters can not be aligned horizontally. For this case, the similarities between the shreds in the same rows are undistinguishable for classification. Therefore, the English document shreds need to be preprocessed before extracting lineation and thus a morphological processing is conducted for the shreds to extract better feature vectors. The adopted morphological processing method is an opening operation which is used to remove some of the foreground pixels from the edge regions and smooth the sharp-edged part.

After lineation, the feature representation of each shred is extracted to be used as the standard to classify. The clustering algorithm adopted is K-means algorithm. As we know, the main idea behind the K-means algorithm is the minimization of an objective function usually used as a function of the deviations between all patterns from their respective cluster centers. The mathematical representation of K-means is depicted in (3), (4) and (5). K-means algorithm is simple and efficient, and has good flexibility for large data. However, K-means algorithm has its limitations such as clustering imbalance caused by uneven sample distribution among different clusters. Thus, the K-means algorithm was improved in the present pipeline to reduce the clustering imbalance. In our pipeline, the shreds in head row are set as the initial cluster center to get more accurate clustering results. Based on the extracted feature vector and the determined shreds in head row, all shreds are clustered into several clusters. However, the shreds in head row could be re-clustered into other clusters rather than the originally correct cluster due to the reassignment of cluster center during each iteration process, resulting in severe unbalance among different clusters. To overcome the weakness, we designed an algorithm to decide whether the shreds belonged

to the head row or not during each reassignment of cluster center, and then the determined shreds in head row would be kept their original cluster number. Fig. 2 presents the flowchart for the improved K-means algorithm.

$$C_i = \arg\min \left\| x^{(i)} - \lambda_j \right\|^2 \tag{3}$$

$$\lambda_j = (\sum_{i=1}^{m} 1\{C_{i=1} = j\} x^{(i)}) / (\sum_{i=1}^{m} 1\{C_{i=1} = j\}) \tag{4}$$

$$J(C, \lambda) = \sum_{i=1}^{m} \left\| x^{(i)} - \lambda(i) \right\|^2 \tag{5}$$

where x denotes sample points, μ denotes cluster center, C denotes clusters, J denotes objective function.

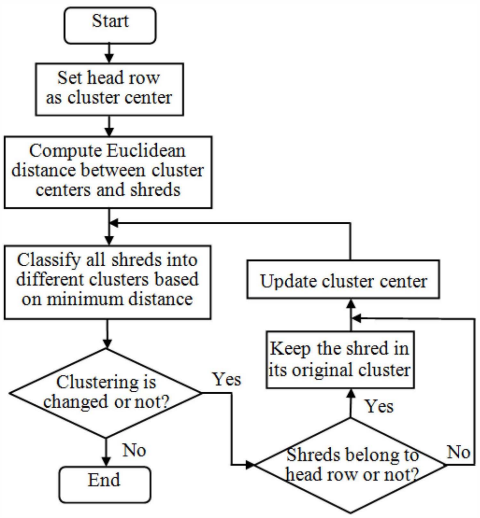

Figure 2. Flowchart for improved K-means clustering algorithm.

2) Permutation in row

Our task is to permute the shreds in the same row after separating the cross-cut shreds into several clusters. Since the successful performance of converting the permutation problem into an assignment one has been verified, Hungarian algorithm is used to complete the permutation work.

3) Permutation in column

Firstly, we attempt to rotate the jointed paper slips anticlockwise by 90 degree so that the above-discussed method for reconstructing strip-shredded document can be used. Nevertheless, the result is not reasonable because most top and bottom edges of horizontal paper slip are all-white so that two paper slips can't be distinguished just based on no-difference edges. Therefore, a model is established to reflect both blank space and character region. As mentioned above, two neighboring horizontal paper slips are complementary because the total heights of character and blank space are fixed in a document. For example, supposing that the character height is 30 pixels while the blank spacing is 40 pixels respectively, the bottom gray vector of slip A and the top gray vector of slip B should meet (6) and (7) if slip A is on the top of slip B.

$$(\text{Character height})_{slipA} + (\text{Character height})_{slipB} = 30\,px \tag{6}$$

$$(\text{blank})_{slipA} + (\text{blank})_{slipB} = 40\,px \tag{7}$$

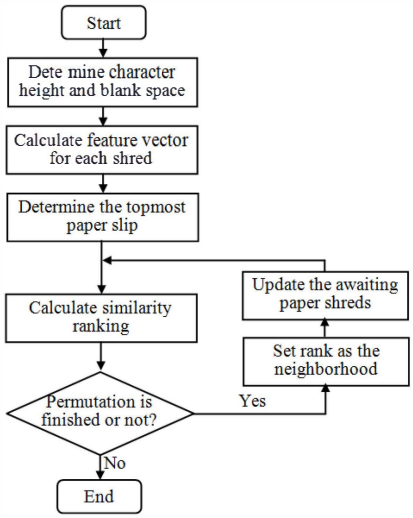

Figure 3. Flowchart for permutation in column.

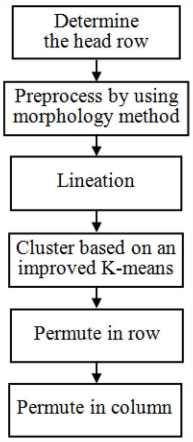

Figure 4. Pipeline to reconstruct cross-cut shredded document.

(i) Determine the character height and blank space which are expressed as h_c and h_b respectively.

(ii) Define a matrix. Of which, each row representing one head row of 11 paper slips stores Euclidean distances of [h_c-bottom character, h_b-bottom blank] and sub-vector comprised of the 1st and 2nd elements among the rest 10 slips.

(iii) Determine the topmost paper slip.

(iv) Update the awaiting paper slip.

(v) Search similarities rank and set 1st slip as the neighboring slip.

(vi) Repeat steps (iv) and (v) till finish the reconstruction of whole document.

Figure 3 represents the flowchart for permutation in column. The designed pipeline to reconstruct cross-shredded English documents is shown in Fig. 4.

III. RECONSTRUCTION EXPERIMENT

In order to validate the designed pipeline, the reconstruction experiments were conducted for both strip and cross-shredded English documents by using the dataset from reference [14].

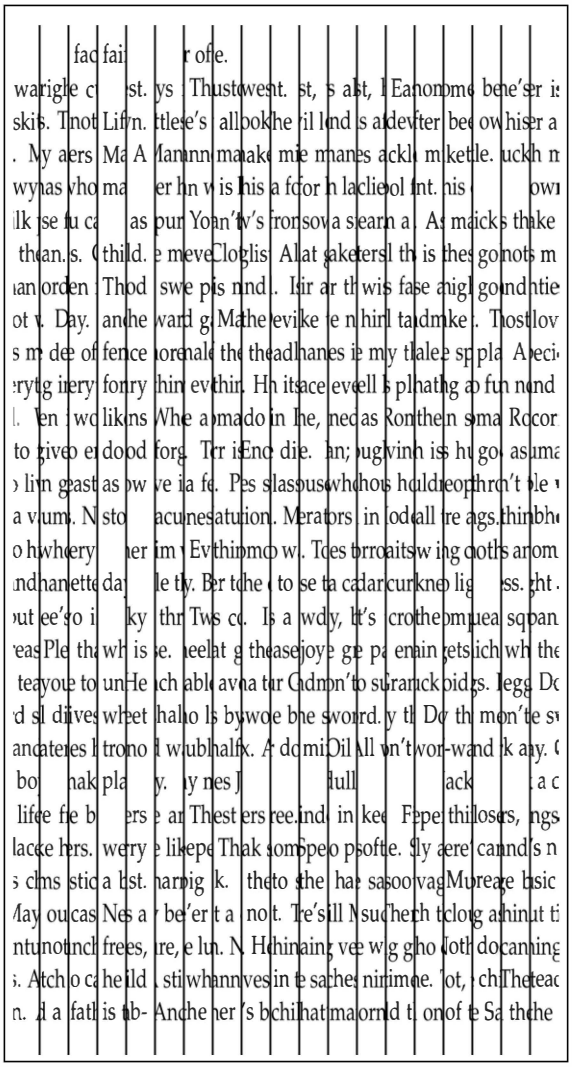

Figure 5. Overview of strip-shredded English document before reconstruction.

A. Reconstruction of Strip-Shredded English Document

The size of shreds is 1980×72 pixels. The overview of original strip-shredded English document before reconstruction is shown in Fig. 5. It is confirmed that the leftmost (head row) paper slip for the English document was number 003, and the corresponding permutation sequences are listed in Table I. By implementing the improved K-means algorithm, the reconstructed English document is shown in Fig. 6.

The customer is always right. East, west, home's best. Life's not all beer and skittles. The devil looks after his own. Manners maketh man. Many a mickle makes a muckle. A man who is his own lawyer has a fool for his client.

You can't make a silk purse from a sow's ear. As thick as thieves. Clothes make the man. All that glisters is not gold. The pen is mightier than sword. Is fair and wise and good and gay. Make love not war. Devil take the hindmost. The female of the species is more deadly than the male. A place for everything and everything in its place. Hell hath no fury like a woman scorned. When in Rome, do as the Romans do. To err is human; to forgive divine. Enough is as good as a feast. People who live in glass houses shouldn't throw stones. Nature abhors a vacuum. Moderation in all things.

Everything comes to him who waits. Tomorrow is another day. Better to light a candle than to curse the darkness.

Two is company, but three's a crowd. It's the squeaky wheel that gets the grease. Please enjoy the pain which is unable to avoid. Don't teach your Grandma to suck eggs. He who lives by the sword shall die by the sword. Don't meet troubles half-way. Oil and water don't mix. All work and no play makes Jack a dull boy.

The best things in life are free. Finders keepers, losers weepers. There's no place like home. Speak softly and carry a big stick. Music has charms to soothe the savage breast. Ne'er cast a clout till May be out. There's no such thing as a free lunch. Nothing venture, nothing gain. He who can does, he who cannot, teaches. A stitch in time saves nine. The child is the father of the man. And a child that's born on the Sab-

Figure 6. Overview of strip-shredded English document after reconstruction.

TABLE I. PERMUTATION OF STRIP-SHREDDED DOCUMENTS

Document Type	Sequence
English	003-006-002-007-015-018-011-000-000-005 -001-009-013-010-008-012-014-016-004

B. Reconstruction of Cross-Shredded Document

Figure 7 shows the overview of the cross-shredded documents before reconstruction. According to proposed method, we should firstly pick out 11 head rows from the paper slip pool composed of 209 size-identical shreds and set them as the cluster centers in the improved K-means algorithm in each iteration. Table II shows the eleven head rows determined for the English document.

In order to use the above-mentioned lineation algorithm, it is needed to determine the value of partition threshold. Since the improved K-means algorithm is used to cluster the shreds, our goal is to make the clustering results as balance as possible, namely, the most proper result is that each cluster should contain 19 shreds. Here, the variation of balance value calculated by module of each cluster is used to evaluate the balance of clustering result, defined in (8). The

value of threshold is determined to be 0.2 by iterating within a small range, indicating that a row can be regarded as a character line once the number of non-white pixels in one row is over 20% of the total pixels, as shown in Fig. 8. So far, the improved K-means algorithm is implemented to classify 209 shreds into 11 clusters and the clustering results are listed in Table III.

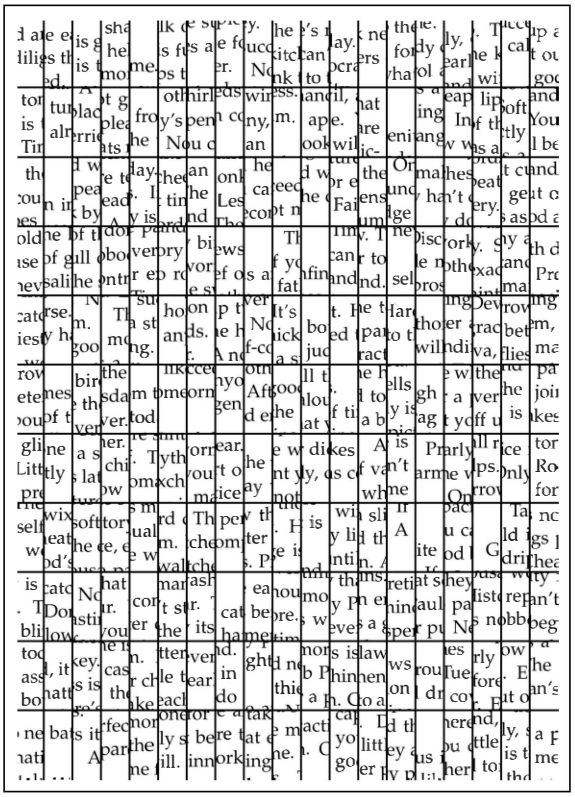

Figure 7. Overview of cross-shredded document before reconstruction.

TABLE II. DETERMINED HEAD ROW FOR BOTH ENGLISH DOCUMENT

Document Type	Number of Head Row					
English	019	020	070	081	086	132
	159	171	191	201	208	/

$$Balance=var(num_1, \ldots, num_{11}) \quad (8)$$

From Table III, it can be noted that the module members of all rows are 19. Based on the classification results in Table III, we apply the method for permutation in row and export the results to check the legibility.

For the English case, a preprocessing is needed and then the preprocessed shreds are clustered. Identically, there are some unexpected situations for the results. Based on the adjusted results, we permute and obtain the row-clustering results for English document. Fig. 9 shows the overview of two typical row-matching results for English document.

TABLE III. K-MEANS RESULTS FOR EACH CLUSTER IN ENGLISH DOCUMENT

Row	Members
1	002 004 011 032 039 064 065 067 075 104
	106 147 149 154 180 184 190 191 204
2	006 017 026 028 078 080 091 094 100 101
	103 113 146 148 164 170 196 198 201
3	005 024 029 030 037 040 046 051 059 085
	086 092 098 107 117 127 150 158 186
4	019 022 057 071 082 088 093 105 114 121
	126 141 151 155 165 176 182 194 202
5	001 031 038 050 053 063 085 097 120 123
	129 138 139 153 159 160 175 187 203
6	015 020 036 041 043 045 073 076 079 108
	116 135 136 143 161 173 179 199 207
7	007 021 033 049 054 061 062 112 118 119
	133 142 162 168 169 189 192 197 208
8	008 014 023 047 060 068 070 084 090 096
	099 109 122 137 156 172 174 185 195
9	003 013 025 027 034 069 095 110 111 130
	132 144 163 166 167 178 181 188 206
10	009 010 016 018 035 042 044 055 056 066
	074 083 134 145 152 157 171 183 205
11	000 012 048 052 072 077 081 087 089 102
	115 124 125 128 131 140 177 193 200

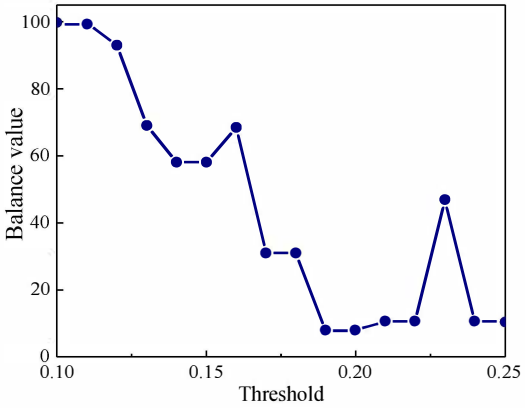

Figure 8. Variance of balance value with thresholds.

horse. Practice makes perfect. Hard work never did anyone any harm. Only has compared to the others early, diligently
(a)
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
(b)

Figure 9. Overview of three row-matching results for English document: (a) row 6 and (b) row 7.

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 10. Reconstructed cross-shredded English document.

Since English shreds in the corresponding rows have been organized into an intact row, the next problem is to these row-matching paper slips are permuted and reconstructed a complete document. Here, a permutation algorithm depicted in Fig. 3 is used. Fig. 10 shows the overview of reconstructed English document by applying the method into the permutation in column.

IV. Conclusions and Future Work

This paper investigated the reconstruction of cross-shredded English document and a complete pipeline is designed to reconstruct the cross-shredded English documents. The pipeline firstly classifies the shreds into several clusters based on an improved K-means algorithm for reducing the clustering imbalance. Meanwhile, a preprocessing is needed before extracting feature vectors for shredded English documents due to the unaligned characters. In addition, Hungarian algorithm is applied into the permutation for the cross-shredded slips in the same row. Eventually the location of the connective horizontal paper slips are arranged by considering the complementary relationship of edge vectors between two neighboring shreds. The present pipeline can be regarded as a black box which can export a completely reconstructed document by just inputting the information of shreds.

As one solution for reconstructing the cross-shredded documents, the pipeline proposed in this paper can be improved by designing a better method to extract a more distinguishable feature vector to cluster the cross-cut shreds. In the future, in order to enhance the automation extent, we will also focus on the determining method of segmentation scale by the pipeline itself instead of knowledge from human being.

References

[1] P. De Smet, "Reconstruction of Ripped-up Documents Using Fragment Stack Analysis Procedures", Forensic Science International, vol. 176, Apr. 2008, pp. 124-136, doi:10.1016/j.forsciint.2007.07.013.

[2] E. Justino, L. S. Oliveira and C. Freitas, "Reconstructing Shredded Documents through Feature Matching" Forensic Science International, vol. 160, Jul. 2006, pp.140-147, doi: 10.1016/j.forsciint.2005.09.001.

[3] A. Ukovich and G. Ramponi, "Features for the Reconstruction of Shredded Notebook Paper" Proc. IEEE International Conference on Image Processing (ICIP 2005), IEEE Press, Sep. 2005, pp. 93-96, doi:10.1109/ICIP.2005.1530336.

[4] A. Ukovich, G. Ramponi, H. Doulaverakis, Y. Kompatsiaris, and M. Strintzis, "Shredded Document Reconstruction Using MPEG-7 Standard Descriptors", Proc. the Fourth IEEE International Symposium on Signal Processing and Information Technology, IEEE Press, Dec. 2004, pp. 334-337, doi: 10.1109/ISSPIT.2004.1433788.

[5] Marlos A. O. Marques and Cinthia O. A. Freitas, "Reconstructing Strip-shredded Documents Using Color as Feature Matching", Proc. the 2009 ACM symposium on Applied Computing (SAC'09), ACM Press, Mar. 2009, pp. 893-894, doi:10.1145/1529282.1529475.

[6] C. Schauer, M. Prandtstetter, and G. R. Raid, "A Memetic Algorithm for Reconstructing Cross-Cut Shredded Text Documents", Proc. International Workshop on Hybrid Metaheuristics (HM 2010), Springer Press, Oct. 2010, pp. 103-117, doi: 10.1007/978-3-642-16054-7_8.

[7] M. Prandtstetter and G. R. Raidl, "Meta-heuristics for Reconstructing Cross Cut Shredded Text Documents", Proc. the 11th Annual Conference on Genetic and Evolutionary Computation (GECCO 2009), ACM Press, Jul. 2009, pp. 349-356, doi:10.1145/1569901.1569950.

[8] M. Prandtstetter, "Hybrid Optimization Methods for Warehouse Logistics and the Reconstruction of Destroyed Paper Documents", Ph.D. Thesis, Vienna University of Technology, 2009.

[9] C. Schauer, "Reconstructing Cross-cut Shredded Documents by Means of Evolutionary Algorithms", Master's Thesis, Vienna University of Technology, 2010.

[10] A. Deever and A. Gallagher, "Semi-automatic Assembly of Real Cross-cut Shredded Documents", Proc. 19th IEEE International Conference on Image Processing (ICIP 2012), IEEE Press, Sep. 2012, pp. 233-236, doi:10.1109/ICIP.2012.6466838.

[11] B. Biesinger, C. Schauer, B. Hu, and G. R. Raidl, "Enhancing a Genetic Algorithm with a Solution Archive to Reconstruct Cross Cut Shredded Text Documents", Proc. International Conference on Computer Aided Systems Theory Extended Abstracts of the 14th International Conference on Computer Aided Systems Theory, Springer Press, Feb. 2013. pp.380-387, doi: 10.1007/978-3-642-53856-8_48.

[12] A. Sleit, Y. Massad and M. Musaddaq, "An Alternative Clustering Approach for Reconstructing Cross Cut Shredded Text Documents", Telecommunication Systems, vol. 52, Mar. 2013, pp.1491-1501, doi: 10.1007/s11235-011-9626-x.

[13] M. Prandtstetter and G. R. Raidl, "Combining Forces to Reconstruct Strip Shredded Text Documents", Proc. HM'08-the 5th International Workshop on Hybrid Metaheuristics, Springer Press, Oct. 2008. pp.175-189, doi: 10.1007/978-3-540-88439-2_13.

[14] J. Zhu, X. Hu, C. Zheng, A Concise Tutorial on Mathematical Models, 1st ed., Science Press, 2015, pp.199.