AENSI Journals

Journal of Applied Science and Agriculture

ISSN 1816-9112

Journal home page: www.aensiweb.com/jasa/index.html

Combinational Method for Shredded Document Reconstruction

[1]Rouhollah Mostafaei, [2]Vasif V.Nabiyev, [3]Parisa Gouchkhani

[1]*Computer Engineering Department, Islamic Azad University, Khoy Branch, Khoy, Iran*
[2]*Computer Engineering Department, Karadeniz Teknink University, Trabzon, Turkey.*
[3]*Computer Engineering Department, Islamic Azad University, Khoy Branch, Khoy, Iran*

ARTICLE INFO

Article history:
Received 20 January, 2014
Received in Revised form 16 April, 2014
Accepted 25 April 2014
Available Online 5 May, 2014

Keywords:
Shredded Documents
Reconstruction
Feature Extraction
Matching

ABSTRACT

Background: Shredded document reconstruction can provided necessary information in forensic investigations but is currently time consuming and requires significant human labor. **Objective:**Over the past decade researchers have been improving automated reconstruction techniques but it is still far from a solved problem. **Results:**In this paper we propose a combinational method for reconstructing documents that are shredded by hand and by machine. Our proposed method is based on both character identification and feature matching techniques. **Conclusion:** Practical results of this hybrid approach are excellent. . The preliminary results reported in this paper, which take into account a limited amount of shredded pieces (10–15), demonstrate that proposed approach produces interesting results for the problem of document reconstruction.

To Cite This Article: Rouhollah Mostafaei, Vasif V.Nabiyev, Parisa Gouchkhani., Combinational Method for Shredded Document Reconstruction. ***J. Appl. Sci. & Agric.***, *9(4): 1452-1457, 2014*

INTRODUCTION

Reconstruction of shredded document is a tedious and laborious task that should be performed by forensic document examiners quite often. The amount of time necessary to reconstruct a document depends on the size and the number of fragments, and it can be measured in days or even weeks. A typical shredded document reconstruction system can be decomposed into two major parts, pairwise matching and a reconstruction strategy. Pair- wise matching measure show well any two shredded pieces fit together. The reconstruction strategy then attempts to use the pairwise matching results to reassemble the original document. In order to alleviate the manual effort of the forensic examiner, some methods for reducing the complexity of reconstruction and reassembly problems using digital images have been proposed in the literature. Most of these methods were developed for solving related problems, such as the jigsaw puzzles Yao and Shao (2003). In general, they are based on specific shape and color features as well as the relationships that may exist between several jigsaw puzzle pieces.

An efficient algorithm for puzzle solving was proposed by Wolfsonin, Wolfson(1990). In this work, the author presents two curve matching algorithms where the boundaries are represented by shape feature strings which are obtained by polygonal approximation. It fails when the number of puzzle pieces becomes larger, though. Another interesting strategy is proposed by Kong and Kimia (Kongand &.Kimia., 2001) They resample the boundaries by using a polygonal approximation in order to reduce the complexity of the curve matching. Dynamic programming is used to align the pieces.

These techniques have been applied to other fields such as archeology and art restoration. In these cases, the goal is to reconstruct two-dimensional objects that have been broken or torn into a large number of irregular fragments. Willis and Cooper (Willis and Cooper., 2008) address the problem of artifact reconstruction discussing 2D and 3D approaches. Leit˜ao and Stolfi (Leitao and Stolfi, 2002) propose an algorithm based on incremental dynamic programming to reconstruct ceramic tiles. Interesting results also have been reported by Papaodysseus *et al* (Papaodysseus., 2002) where the focus is the reconstruction of archaeological wall-paintings.

Regarding the reconstruction of documents for forensics purposes, few works can be found in the literature. In a more recent work, Smet(2008) discuss a formal analysis of the problem of reconstructing ripped-up documents when the remnants can be recovered as an ordered stack of fragments. Justino *et al* (2005) propose a local reconstruction of shredded documents based on polygonal approximation and feature matching.

However, very often questioned documents suffer damages at several levels, such as, torn edges, moisture, obliteration, charring, and shredding. In the latter case, shredding can be performed by a machine or by hand (Fig.1). In both cases, documents need to be reconstructed so that forensic examiners can analyze them. The

Corresponding Author: Rouhollah Mostafaei, Computer Engineering Department, Islamic Azad University, Khoy Branch, Khoy, Iran.

amount of time necessary to reconstruct a document depends on the size and the number of fragments, and it can be measured in days or even weeks. Sometimes some fragments of the document can be missing, and for this reason, the document can be only partially reconstructed.

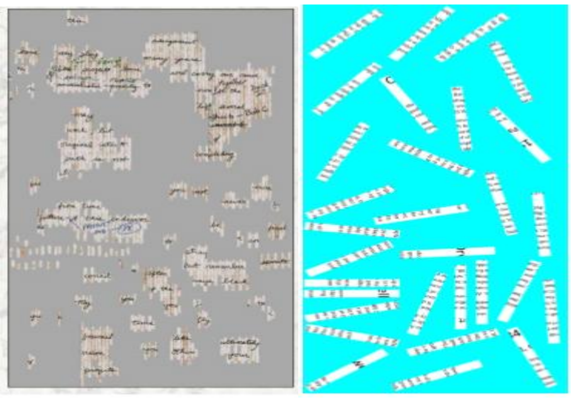

Fig.1: Differnt kinds of shredding.

In this work we introduce a methodology based on both character identification and feature matching techniques for shredded document reconstruction. First we focus on detecting shredded pieces type (shredded by hand or machine). This can be performed by calculating the angles of shredded pieces that we can consider them as polygon.

After shredding type detection we call one of two functions for shredded document reconstruction. Details will discuss at remain of this paper. If documents are shredded by hand we use Feature Extraction (FE) method based on angles and etc. Otherwise Character Identification (CI) based method for document reconstruction is use.

1. The proposed methodology:

Our methodology is composed of two distinct parts as depicted in Fig. 2. Initially, shredding kind of each piece of the document is detected so pieces of each group (shredded by hand or machine) pre-processed distinctly. Then, a set of features is extracted from pieces of each group in order to carry out the matching. Finally suitable reconstruction is performed. In the following sections, we describe in details each component of the methodology.

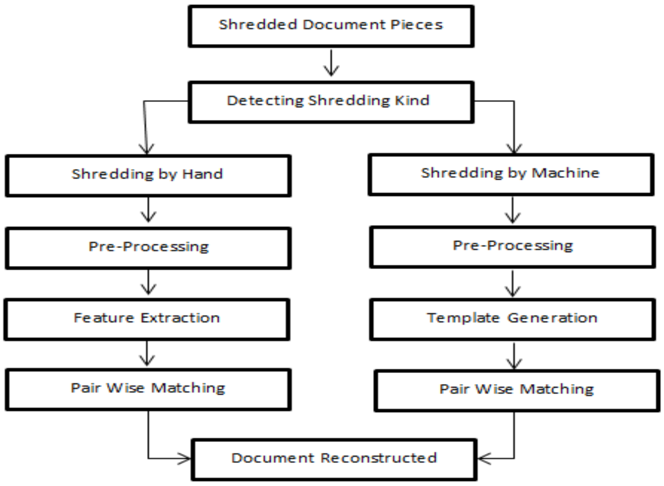

Fig. 2: The block diagram of the proposed methodology.

1.1. Detecting Shredding Kind:

In general, as fig1 shows two types of shredded pieces exist, rectangular or polygonal. Documents that shredded by machine will have rectangular form and documents that shredded by hand will have polygonal form. To distinguish between two types of pieces, we should compute at least four distinct angles of one of pieces (for reducing complexity). If all of angles have degree 90, document was shredded by machine otherwise was shredded by hand. For computing degree of angle we consider follow equation:

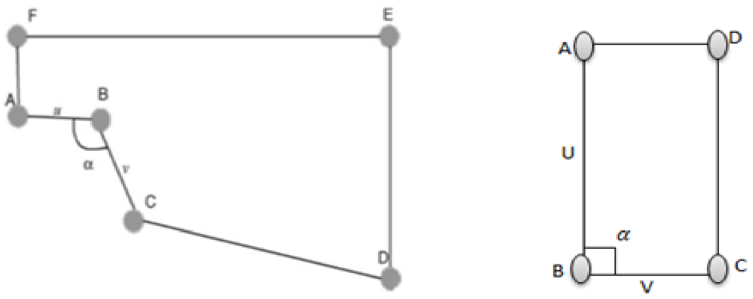

Fig. 3: Pieces of documents that shredded by hand and machine respectively.

Consider, i.e., the vertices A and B in the polygon and rectangle depicted in Fig.3. The angle α is given by:

$$\cos\alpha = {}^{\mathrm{UV}}\!\!/_{|\mathrm{U}||\mathrm{V}|} \tag{1}$$

1.2. Shredding by Hand:

1.2.1. Shredding by Hand Pre-processing:

Traditional puzzle solving algorithms usually take into account smooth edges and well defined corners. However, dealing with shredded documents is quite more complex. The act of shredding a piece of paper by hand often produces some irregularities in the boundaries, which makes it impossible to get a perfect curve matching. To overcome this kind of problem, the best results were the well-known Douglas-Peucker (DP) algorithm Douglas and Peucker(1973). This algorithm implements a polyline simplification and it is used extensively for both computer graphics and geographic information systems. Figure4 shows an example of this process using different levels of approximation Pimenta *et al* (2009).

1.2.2. Shredding by Hand Feature extraction:

The feature extraction can be seen also as a complexity reduction process, since it converts the polygon in a sequence of features. The first feature is the angle of each vertex with respects its two neighbors. Consider for example the vertices A and B in the polygon depicted in Figure3 (leftmost). The angle α is given by formula (1). We also verify whether such an angle is convex or concave. For example, in Figure 3(leftmost), vertex B has a convex angle while vertex C has a concave one. To complete our feature set, we compute the distances between the vertex and its neighbors (next and previous in a clock wise sense). Such distances are achieved by means of the well-known Euclidean distance.

1.2.3. Shredding by Hand Matching:

Here, the main goal is to compute a degree of similarity between the boundaries of each fragment of the image. To perform this task, the LCS (Longest Common Subsequence) algorithm, which is a dynamic programming algorithm, devoted to find the longest subsequence common to all sequences in a set of sequences. In our case, the sequences are the features extracted previously. The LCS starts with a matrix E of size (M+1)×(N+1),where M and N are the length of the two sequences (X and Y) being analyzed. The first row and column are filled with zeros. The remaining values respect the following definition:

$$E_{ij} = \begin{cases} E_{i-1j-1} + S \\ \max(E_{i-1j} + P, 0) \\ \max(E_{ij-1} + P, 0) \end{cases}$$

Where S =1 if Xi = Yj and 0 otherwise. P is the penalty value.

1.2.4. Shredding by Hand Reconstruction:

In the end of the matching process we have a list of all possible matches with its respective scores, which should be used to reconstruct the image of the document. To perform the reconstruction in a more structured way, a graph representation can use, which shows clearly the relationship among all fragments Pimenta *et al* (2009).The original graph can have cyclic links, which makes the reconstruction problem more complex. To mitigate such a complexity, the matching graph has been transformed into a minimum weight spanning tree, using Prim’s algorithm Prim(1957).

After building the Prim’s tree, the reconstruction of the document is straightforward. Starting from node 1 and visit all other nodes using the information about rotation and translation found previously.

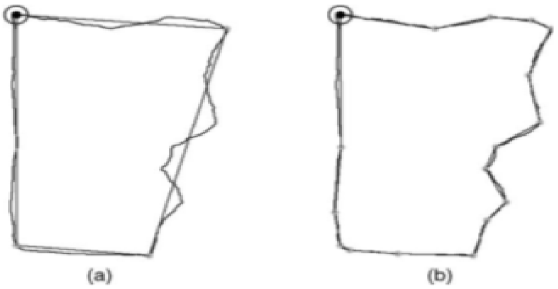

Fig. 4: Pre-processing using Douglas-Peucker algorithm.

1.3. Shredding by Machine:

1.3.1. Shredding by Machine Pre-Processing:

Preprocessing begins with an image of shredded pieces against a solid colored background. This input image is assumed to be perspective corrected as would result from a flatbed scanner or similar method of acquisition. The background color should not be present on any of the pieces and the pieces should not overlap.

The first step of preprocessing generates a mask that will be used to locate and extract the pieces. First, an optional box blur is applied to reduce high-frequency noise. Next, a flood fill is performed with a configurable delta value. The delta value specifies the maximum amount each channel may differ from the respective channel in an adjacent pixel for the two to be considered neighbors while filling. This delta value allows for minor local variations in background color and larger variations across the entire image. The flood fill is started from each of the four corners and it is assumed that the entire background region is reachable from at least one corner.

The next preprocessing step aligns the components such that their major axis is aligned vertically and extracts them from the image. This is done by computing an affine transformation that rotates the minimum area bounding rectangle the minimum amount needed, for alignment and translates it so that the top left corner is at the origin. Applying this transformation to the raw input image, results in an aligned image of each piece cropped by the minimum area bounding rectangle (see fig5). It is assumed that the flow of text is horizontal so that, characters in the aligned piece are either upright or inverted. At this point, the pieces are converted to gray scale because later stages only require luminance.

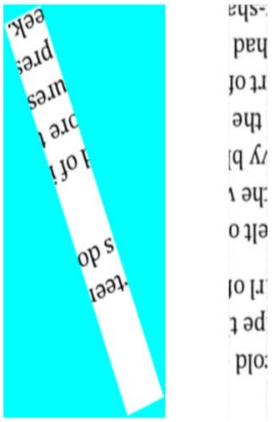

Fig. 5: Piece extraction and alignment (left to right respectively).

Finally, it is necessary correct the orientation of pieces containing inverted text by rotating them 180 degrees. At this point, the preprocessing and template generation stages become somewhat interleaved with the reorientation process relying on the templates and related information. As a result, this process will be described after template generation despite it belonging conceptually to the preprocessing stage.

1.3.2 Shredding by Machine Template Generation:

The goal of this stage is to generate a set of character templates that are likely to match the characters in the document being reconstructed. In order to generate templates, it is necessary to know the font sizes used in the document. The sizes can be estimated from horizontal projections of the lines. More specifically, the x-height can be estimated for each line on each piece and an x-height ratio can be computed for each font face allowing the template render size for the font to be estimated for each line. The estimated x-heights from all lines on all pieces are stored in a set to avoid generating duplicate templates.

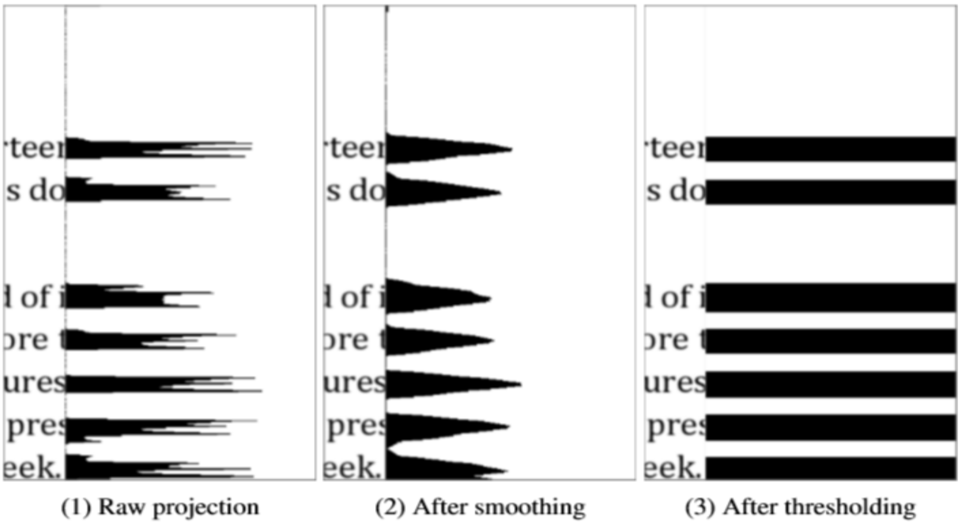

Fig. 6: Extraction of rough line bounds.

First, the shred luminosity is inverted to produce light text on a dark background and the horizontal projection of each shred is computed by summing the rows of the shred image. A box-blur with a user configurable window is then applied to the projection in order produce a filtered version with major peaks roughly identifying the location of individual lines. The filtered projection is then threshold to produce runs of line and non-line regions. For each line region, the raw projection is extracted and processed. (See fig6)

1.3.3. Shredding by Machine Reconstructing:

As in Perletal (2011), reconstruction is done using randomized exclusive matching. This is capable of reconstructing entire strip shredded documents but only produces chains when working with cross-cut shredded pieces. At the start of the process, each shred exists as a chain of length one and all edges are placed in the set of unmatched edges. Next, a random right edge is selected from the unmatched pool and the best matching unpaired left edge is found. The two chains are then joined and the left and right edges involved are removed from the unmatched pool. If there are no matching left edges remaining, the selected right edge is simply removed from the pool. This process of joining chains continues until there are no edges remaining in the unmatched pool. The mean match value for the set of resulting chains is then calculated from the matches that are actually used. The entire process is repeated a number of times and the set of chains with the highest mean match value is used as the best reconstruction.

Table1: Comparative results.

Strategy	Strip Shredding Reconstruction (%)	Hand Shredding Reconstruction (%)
Proposed	65	71
Pimenta *et al*	8	75
Perletal *et al*	70	12

Experiments:

To validate the proposed methodology we have used a database composed of 50 documents. Those documents were shredded into 5-20 fragments and their size range from 1cm × 2 cm to 5cm× 10 cm. The same database was used in Pimenta*et al* (2009).

Table 1 compares our results to those reported in Pimenta*et al* (2011), as we can notice; the methodology proposed in this work brings a considerable boost in the performance of the algorithm. In the case of strip shredding reconstruction performance of proposed method in average is in average 65%, but performance of Pimenta *et al* method is 8% and in the case of hand shredding reconstruction performance of proposed method in average is in average 71% against Perletal *et al* method, that is 12%. Generally, as we can see in table1 our proposed method has better performance in both shredding types. Notice that each of other methods has weak performance in one of the shredding types.

Conclusion:

In this paper, we have proposed combinational method for reconstructing documents that are shredded by hand and by machine. Initially we detected the shredding tape (by shredder machine or hand). Next base on the shredding type we used one of reconstruction methods. Specialty of our proposed method is reconstruction of any types of shreds. The results reported in the paper show an important boost in the reconstruction rate. However, there is a lot of room for improvement.

REFERENCES

.Justino, E., L.S. Oliveira and C. Freitas, 2005. “Reconstructing shredded documents through feature matching,” Forensic Science Intern, 160.

De, P., Smet, 2008.“Reconstruction of ripped-up documents using fragment stack analysis procedures,” Forensic Science Intern, 176: 124-136.

Douglas, D. and T. Peucker, 1973. “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature,” The Canadian Cartographer, 10: 112-122.

Johannes Perletal, 2011.“Strip shredded document reconstruction using optical character recognition.” In: Imaging for Crime Detection and Prevention, 2011. (ICDP 2011), 4th International Conference on, pp: 1-6.

Kongand, W., B. Kimia, 2001. “Onsolving2Dand3Dpuzzles under curve matching,” in CVPR, pp: 583-590.

Leitao H.C.G. and J. Stolfi, 2002. A multiscale method for the reassembly of two-dimensional fragmented objects,” IEEE Trans. Pattern Anal and Machine Intel, 24: 1239-1251.

Papaodysseus, C., T. Panagopoulos, M. Exarhos, C. Tri-antafillou, D. Fragoulis and C. Doumas, 2002. “Contour- shape based reconstruction of fragmented,” IEEE Trans. Signal Processing, 50: 1277-1288.

Pimenta, A., E. Justino, S. Luiz, Oliveira and R. Sabourin, 2009. “ Document reconstruction using dynamic programing Forensic Science Intern, 25.00 IEEE.

Prim, R.C., 1957. “Shortest connection networks and some generalizations,” Bell System Technical Journal, 36: 1389-1401.

Willis, A.R. and D.B. Cooper, 2008. “Computational reconstruction of ancient artifacts,” IEEE Signal Processing Magazine, 25: 65-83.

Wolfson, H., 1990. “On curve matching,” IEEE Trans. Pattern Anal and Machine Intell, 12: 483-489.

Yao F.H. and G.F. Shao, 2003. “A shape and image merging technique to solve jigsaw puzzles,” Pattern Recognition Letters, 24: 1819-1835.