Comparative Study on Reconstruction of Shredded Documents Based on Various Image Processing Techniques

Shefali Srivastava[1] Mukund Gohil[2] Viraj Dabhi[3] Kruti J.Dangarwala[4]
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Shri S'ad Vidya Mandal Institute of Technology, Bharuch - 392001, Gujarat, India

Abstract— The shredded document reconstruction is a standout amongst the most widely utilized techniques for semi computerized or mechanized recovery of detailed records from tore up documents. Reconstructing the shredded files manually is similar to resolve a jigsaw puzzle however because of uneven edges, shapes and measurement this assignment turns into very boring and intricate. It's a helpful job for forensic investigation science, historical artefact reconstructions. A generalized architecture for the shredded document reconstruction is proposed on this paper. On this paper we have now talked concerning the routines to reconstruct the ripped documents using various algorithms with their advantages and disadvantages.

Key words: Document Reconstruction, Algorithms, Image Processing

I. INTRODUCTION

Document can be categorized as foremost, private, personal, evidence or secret and contains touchy knowledge of the owner. These documents are of maximum value to forensic, criminal investigation, intelligence gathering and military operations. Documents store, organize and give an explanation for understanding for schooling, illumination and advancement of human progress. Documents are shredded for a variety of motives, but the foundational motive is to smash the information on that document and to keep the content of documents secret.

Shredding may also be guide like hand torn, torn edges, moisture, obliteration, charring or can be automatic using a shredder machine. Documents get worse due to insects, moisture, temperature, humidity, consistent handling and weathering. There are a variety of shredders and shredding methods [1]. Typically, a shredder that produces extra smaller pieces, or shreds, from the document is more secure. Whereas Least secure are the strip shredding (slicing the document into strips that span the length of the document) or hand-shredding into big pieces.

Right here comes the role of de-shredding i.e. reconstruction of those shredded documents. Reconstructing them manually is a time consuming job, and needs hard work. Automation of reconstruction made the undertaking less complicated and extra mighty. It yields image processing algorithms. This paper describes more than a few methods of reconstructing a document through assembling the shreds from a shredder or hand-torn documents. Reconstruction of shredded document is particularly predominant to import expertise which has broad software in forensic science, artwork conservation, spying, military and archaeology. The shredded documents will also be of various types in order with the sort of shredding is completed.

In 2011, DARPA launched a shredder challenge [2] to beef up capabilities in reconstructing shredded documents. The assignment concerned five puzzles of increasing difficulty, every of which incorporated color-scanned portraits of shredded pieces (chads) from one or more strip-shredded files. Proposed solutions to that challenge revealed that present strategies could not deal with the complexity of reassembling shredded documents. The puzzles have become the normal dataset for research in shredded document reconstruction.

Reconstruction of the shredded document is very similar to solving jigsaw puzzle [3]. The difference between shredded document and jigsaw puzzle is conveniently that the shredded document does now not have delicate contours. The difficulty additionally arises as a result of their shape and dimension of the chads. An extra obstacle concerns the wide variety of shredded parts that must be regarded during reconstruction.

Computerized document reconstruction strategies had been proposed to relief this issues, which will also be roughly divided into two categories established on whether or not the piece outlines are sampled uniformly [4]. One is string-matching based methods, in which fragment contours are represented by means of uniformly sampled points. Another one is feature-matching based approach, where the fragment contours are typically represented with the aid of utilizing critical points or polygonal approximations. To completely reconstruct the original document, a global technique is required to eliminate the paradox due to regional curve matching.

II. ARCHITECTURE

A common structure for reconstruction of ripped document is proposed on this paper which is based on the framework shown in [5]. The generalized process diagram is given in Figure: 1. The algorithm for the architecture is given as follows:

1) In the first phase we take shredded or torn document from scanner as our input.
2) Extract features i.e. extract the chads from the input image.
3) Arrange them in correct order by making use of various algorithms like Global Search, OCR(Optical Character Recognition) etc.
4) Recombine the shredded document image.
5) Apply distortion reduction methods to enhance the quality of document image.
6) At last the high resolution de-shredded document image is accomplished.

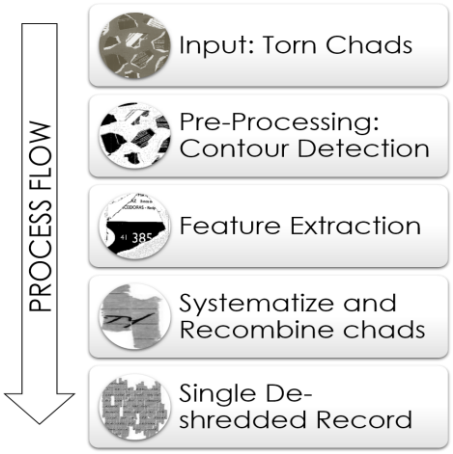

Fig. 1. Generalized Process Diagram for de-shredding of shredded documents

III. Related Work

Image Mosaicing means mixing together of a few arbitrarily formed images to create one enormous image so that the boundaries between the established image aren't noticeable [10]. This technique can be used for reconstruction of shredded document. Creation of shredded document from some torn piece views is an efficient means of getting a larger view of document than the torned view. In [5], presents 13 various Image Mosaicing Methods using Image Processing techniques to reconstruct the shredded document. They surveyed a lot of papers and enlisted the reviews on different applications of image mosaicing mainly in the area of document image mosaicing. A generalized framework is proposed by them.

In [4] reconstructing documents that have been shredded through hands or machine is discussed. First of all each chad of document is pre-processed via polygon approximation in an effort to slash the complexity of boundaries with Douglas-Peucker (DP) algorithm. DP algorithm implements a polyline simplification with inner and outer boundaries produced through shredding. In feature extraction phase, they extracted angles and coordinates of the vertex in the image. Now, matching is achieved by means of computing the similarity between polygons and eventually reconstructing the whole document is done with global search algorithm. It eliminates the ambiguities due to local reconstruction. Advantage of this process is that to make approximation to be able to cut back the complexity & get rid of the ambiguity and overcome designated problems faced in document reconstruction. Drawback of this strategy is that the act of shredding a piece of paper by using hand typically produces some irregularities in boundaries. Performance drops as the number of fragments gets greater due to scale used during polygon approximation.

In [6], a Semi-automated Toolset is used for reconstruction of ripped up documents. A Global reconstruction approach is used. This process consists of various steps, first of all chain code is used for feature and page border computation adopted through filling in the frames with non-border fragments by adapting String Matching method. Secondly, Contour Segments is used for matching page border fragments and eventually, Digital Glue is applied using gap and overlap computations. Ultimately outcome shows that [6] also presented a framework of image processing and computer vision techniques that can be used to automate the reconstruction of shredded documents and there additionally exists a limitation which indicates that merging fragments have some space between non uniform borders.

In [7], A. Pimenta and E. Justino, proposed dynamic programming for reconstruction of shredded documents. They used polygonal approximation to shrink the complexity of the boundaries. Thereafter, these aspects are used to feed the LCS dynamic programming algorithm. Thereafter the LCS matching graph is changed right into a minimal weight spanning tree utilizing the modified Prim's algorithm. After constructing the prim's tree, reconstruction of the document is easy.

In [1], an automated algorithm is used for segmenting and orienting individual shreds from a scanned image, as well as for computing features and rating potential matches for every shreds. This paper described reconstructing a document by means of assembling the shreds from a cross-cut shredder in a semi-automated trend where the human and the computer collaborate. They have followed three unique step approach. Firstly, the pre-processing which parsing the shreds to extract and orient each fragment of shred image with the connected accessories. Secondly, feature extraction and matching is performed to symbolize the looks of every shred and to check possible fits for every shred. Eventually, user enters the loop by means of semi-automatic assembly. Human interface is needed to evaluate the right fits. They used techniques of computer vision for suggesting fits that are then confirmed via human to build up completed document.

In [8], makes an attempt to strengthen a computerized reconstruction system by dividing the matching algorithm into three steps. First, Blank area searching algorithm used to pick up the pieces of ripped part for a given grey level matrix. This algorithm determine the portions column via column and the portions which have a column of clean area within the left of the gray stage matrix is authorized and reserved, rest is discarded. Second, Rightward Education algorithm search for the right hand aspect adjacent document pieces from the primary column pieces until each and every row is reconstructed thoroughly. Sooner or later Revised Education algorithm is used searching in special directions to reconstruct the original document in correct order. Also false searching disorders or sample recognition approach is used for scanning system which is potent but time consuming.

In [9], a semi-automatic procedure to reconstruct shredded document is present using curve matching technique. They carried out pairwise matching of chads. They divided chad contours into curves using corner detection and offered a method to access the match of two curves. The curve matching process is effective to translation and rotation and can take care of shape deformations due to shredding by means of permitting overlapping of chads for the period of matching. To support matching efficiency they used text lines alignment, crossing characters and color knowledge on the chads. And they designed a visual interface for user input in selecting correctly matching chad pairs and reconstructing the document. They solved first and second puzzles of DARPA shredder challenge for demonstration.

Pub./Year	Title	Method	Advantages	Disadvantages
IEEE/2005	Document Reconstruction Based on Feature Matching [4]	1) Inner and outer boundaries produced by shredding is recognized. 2) Feature matching is applied. 3) Douglas-Peucker (DP) algorithm is used for Polygon approximation. 4) Apply local matching. 5) Angle features extracted from the polygon. 6) Computes the similarities between polygons. 7) Distance feature is also extracted from the polygon. 8) Global search algorithm is used for best matching method.	1) Eliminates the ambiguity. 2) Reduces complexity of boundaries to improve performance. 3) More reliable identification of relevant matching. 4) Human will dispend considerably less efforts to finish reconstructing the document than starting from scratch. 5) Correct reconstruction ratio is 57%.	1) False Positive ratio is 24 %. 2) Error Ratio is 19 %. 3) Performance drops as the number of fragments gets bigger.
IEEE/2009	Semi-automatic Forensic Reconstruction of Ripped-up Documents[6]	1) Fragments are scanned and segmented using Recursive First In First Out Queue Flooding Algorithm. 2) Contours are used as feature for Feature matching. 3) Chain code method. 4) String matching method. 5) Digital glue.	1) Continuous handling of the evidence is avoided. 2) An interactive and iterative process are used to obtained partial reconstruction results.	1) Edges of the pieces remain blank. 2) Recursion in algorithm may form infinite loop.
IEEE/2009	Document Reconstruction Using Dynamic Programming[7]	1) In Preprocessing, they used Douglas-Peucker(DP) algorithm. 2) Feature Set is computed by Euclidean distance. 3) Longest Common Subsequence Algorithm for Feature Matching. 4) Modified Prim's Algorithm for minimum weight spanning tree to reconstruct the document.	1) Compared To Global Search Algorithm, 18% Improvement Is Seen In The Number Of Fragments Reconstructed. 2) Result Shows An Important Boost In The Reconstruction Rate. 3) Correct Reconstruction Ratio Is 75%.	1) False Positive ratio is 14%. 2) Error Ratio is 11%.
IEEE/2012	Semi-automatic Assembly of Real Cross-cut Shredded Documents[1]	1) Automatic algorithms for segmentation and orientation are proposed. 2) Shred orientation features. 3) Fast matching procedure.	1) Broader evaluation of partially reconstructed regions. 2) Rapid identification & confirmation for correct matches. 3) Easy and quick visual verification & confirmation. 4) Visual inspection can very	1) Difficult puzzle because the number of pieces can large. 2) Overlap was at least 35% of the size of smaller two pieces. 3) Speed can be confirmed on bases of

		4) Pre-processing: parsing the shreds. 5) Connected component algorithm. 6) Principle Component analysis. 7) Up-down orientation of shred. 8) Feature extraction and matching. 9) Human-computer assembly.	quickly confirm the accuracy of the match.	matches. 4) Document have not been completely reconstructed.
IEEE/2014	Shredded Document Reconstruction Based on Intelligent Algorithms[8]	1) Blank area searching algorithm. 2) Rightward education analysis 3) Pattern recognition method. 4) Colony algorithm. 5) Document Preprocessing. 6) Education algorithm	1) Faster Optimization. 2) Frame the original document.	1) Endless loop and false searching is seen in the method. 2) Scanning process is time consuming process. 3) Limited to the rectangle shaped chads.
IEEE/2014	A Semi-automatic Deshredding Method Based on Curve Matching[9]	1) Horizontal Projection to correct its orientation. 2) Smallest Univalue Segment Assimilating Nucleus (SUSAN) operator is used to determine chad corners. 3) Chad pair matching algorithm.	1) The percent of correct matches in top 15% of all possible matching from Puzzle 1 and Puzzle 2 were 54% and 35% respectively. 2) The technique used is robust to translation & rotation and can cope with shape deformations due to shredding by allowing overlapping of chads during matching. 3) Alignment of text lines and crossing characters and color information on the chads is also utilized to improve matching performance. 4) Visual interface for confusion chad pairs.	1) It is semi-automatic and need the help of human to completely reconstruct the document. 2) May increase chad pairs for re-construction

Table 1: The Comparative Study On Shredded Document Reconstruction

IV. Conclusion And Future Work

We've studied quite a lot of ways to reconstruct the shredded document and summarized the study in the Table: I. Reconstruction of shredded document, is very foremost within the fields of archeology, investigation, military, forensic science and day-to-day lives too. By means of our study on this systems we conclude that for much less quantity of chads a completely computerized system may also be desired however for higher quantity of chads or for smaller size of chads we will want human interaction with the computerized procedure. We hope this paper may encourage different researchers.

Furtherly, a neural network can also be developed for this process to scale back the human interaction. To achieve entire accuracy, present algorithm still ought to be improved. And within the output areas between the chads desires to be reduced and the line merging them must be eliminated to fortify total excellent of the reconstructed image.

References

[1] A. Deever and A. Gallagher, "Semi-automatic assembly of real crosscut shredded documents", in Image Processing (ICIP), 2012 19th IEEE International Conference on, Sept 2012, pp. 233–236.
[2] "DARPA Shredder Challenge", http://archive.darpa.mil/shredderchallenge/
[3] F.-H. Yao and G.-F. Shao, "A shape and image merging technique to solve jigsaw puzzles", Pattern Recogn. Lett., vol. 24, no. 12, pp. 1819 – 1835, Aug. 2003.
[4] C. Solana, "Document reconstruction based on feature matching", in Computer Graphics and Image Processing, 2005. SIBGRAPI 2005. 18 th Brazilian Symposium on, Oct 2005, pp. 163–170.
[5] D. Prajapati and K. Dangarwala, "Various document image mosaicing method in image processing: A survey", in Signal Processing And Communication Engineering Systems (SPACES), 2015 International Conference on, Jan 2015, pp. 281–285.

[6] P. De Smet, "Semi-automatic forensic reconstruction of ripped-up documents", in Document Analysis and Recognition, 2009. ICDAR '09. 10 th International Conference on, July 2009, pp. 703–707.
[7] A. Pimenta, E. Justino, L. Oliveira, and R. Sabourin, "Document reconstruction using dynamic programming", in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, April 2009, pp. 1393–1396.
[8] Y. Liu, H. Qiu, J. Lu, and Y. Fang, "Shredded document reconstruction based on intelligent algorithms", in Computational Science and Computational Intelligence (CSCI), 2014 International Conference on, vol. 1 , March 2014, pp. 108–113.
[9] S. Shang, H. Sencar, N. Memon, and X. Kong, "A semi-automatic deshredding method based on curve matching", in Image Processing (ICIP), 2014 IEEE International Conference on, Oct 2014, pp. 5537 – 5541.
[10] R. Inampudi, "Image mosaicing", in Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS '98. 1998 IEEE International, vol. 5, Jul 1998, pp. 2363–2365.

