Assessing Cross-Cut Shredded Document Assembly

Priscila Saboia* and Siome Goldenstein

University of Campinas (Unicamp), Campinas, SP, Brazil
{psaboia,siome}@ic.unicamp.br

Abstract. In this paper we address the problem of quantitative evaluation of cross-cut shredded document reconstruction. We propose quantitative metrics using graph theory and classic information retrieval concepts to compare the neighborhood connectivity graph of a reassembled document shredded by a cross-cut machine against the neighborhood graph of the ground-truth. These metrics focus entirely on the proper relative positioning of the shredded pieces. To do so, we have shredded 12 documents containing diverse content, such as handwriting, printed text, images and photographs. We then scanned, extracted the pieces, and reassembled them into the ground-truth. This dataset is available to the readers, with the original documents, the digital representation of the shreds, and the scripts that provide the quantitative evaluation of the users reconstructions.

Keywords: assessing document assembly, cross-cut, de-shredding dataset.

1 Introduction

Reconstruction of shredded documents is an important problem in forensic investigations, but usually involves many hours of fastidious human labor. The shredding of a document can be either performed by hand or with a machine. In the case of machines, there are two most common types of cuts, strip and cross cuts, which generate shreds of different shapes.

In this work, we decided to avoid any approaches that make naive simplifications of the problem of reconstructing shredded documents. Thus, we sought not to use any kind of simulation of the generations process of shreds. Additionally, we preferred not to face the problem of reconstructing documents torn by hands, and also chose not to deal with fragments generated by strip-cut machines. Both situations have plenty of solutions to their re-assemblage.

Instead, we opted to deal with the scenario yet little investigated of pages shredded by concrete cross-cut machines, as a manner to be closer to forensic problems of the real world. In this sense, we observed that, after 2011, some algorithms to reconstruct documents shredded by cross-cut machines [1–4] emerged

* Thanks the CNPq for financial support.

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 272–279, 2014.

from the *DARPA Shredder Challenge* [5] – a competition with five tasks of content retrieval in real handwritten cross-cut shredded documents. Nevertheless, the available database in that competition is not annotated, and this complicates systematic assessments between existing techniques that aim to solve the problem.

Therefore, we tackle these problems by introducing new quantitative metrics to assess the quality of a reconstruction, as long as the ground-truth is available. We address this issue providing a new unpublished dataset of 12 documents shredded on an off-the-shelf cross-cut shredding machine. The dataset has a scan of the original documents, the digital shreds (shape and texture), as well as the scripts to evaluate the user's reconstruction algorithm which is publicly registered on the adress: `http://dx.doi.org/10.6084/m9.figshare.1136126`. The results of this paper will allow researchers to scientifically advance the state-of-the art in the problem.

2 Related Work

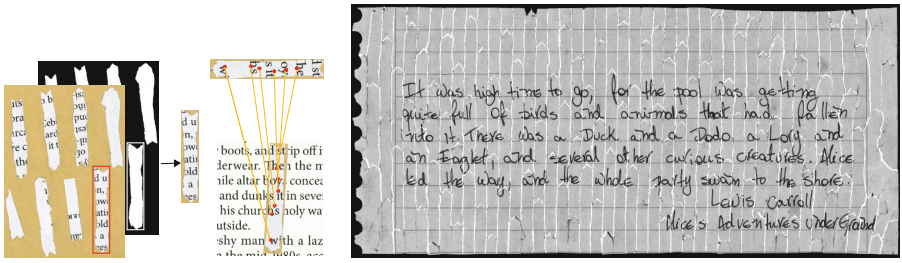
Techniques to automatically solve the problem of reconstructing fragmented documents generally focus on a single type of shredding. In the last decade, many studies focused on the reconstruction of documents shredded by hands [6–9], and by strip-cut machines [10–12]. In the case of cross-cut shredded document reconstruction, existing approaches have initially been tested only on simulated shredded text documents [13–17], in which the shreds have a rectangular shape and its corners are aligned perfectly. These authors have achieved good results for this type of data.

As of interest of this paper, focusing on the researches dealing with fragments generated by real machines, where the corners of the pieces are not aligned when the document is reconstructed, there is only the dataset available on the DARPA Shredder Challenge [5], which contains real cross-cut shredded handwritten documents. To complete each problem in this competition, participants were required to answer questions supported by the info embedded in the shredded pages. As a consequence, a complete reconstruction of the pages was not really necessary, and even human aiding and crowd-sourcing were allowed, since their objective was not to compare the different computational methods used to solve the problem.

Therefore, when a researcher needs to evaluate and/or to compare quantitatively different reconstruction algorithms, it is essential to have a systematic form of reconstruction assessment, less context-dependent with respect to the content of the reassembled documents.

3 Ground-Truth Dataset

Our dataset of real cross-cut shredded documents has, as of this moment, 12 diverse documents containing handwriting, printed text, and photographs. Each



(a) (b) (c)

Fig. 1. Groud-truth content. In (a), shredded pieces acquired by scanning, and a mask example useful to locate the exact positions of the shreds. In (b), control points and feature matching step. In (c), reconstruction example.

document has been shredded into 120–530 pieces, depending on its original physical size. We use a Staples 18-Sheet Cross-Cut Shredder, and each shredded piece has, on average, dimensions of 32 × 4mm. Shred images were acquired with a Plustek OpticPro A320 scanner, at 600 dpi, RGB color mode, all properly placed on a homogeneous colored background.

3.1 Preprocessing

On the preprocessing stage, we segmented the shreds from background using thresholds on the value of the Cb component, in the YCbCr color space. On the resulting binary image, the shreds were identified by connected components with size equal to or greater than the smallest possible shred. Around each identified shred was established a respective minimum bounding rectangle, whose limits oriented the cropping of two former sub-images: the binary mask of the shred, and the colored version of the shred (see Fig. 1 (a)).

3.2 Ground-Truth Establishing

On the establishing of the ground-truth for the reconstruction of the shredded documents, our goal was to find the relative positioning of the fragments in a correct reconstruction. In this case, we used image registration techniques to find the geometric alignment of the shred images with the intact document image (see Fig. 1 (b)), previously scanned before being shredded.

Four steps were applied to establish the ground-truth: feature detection, feature matching, transform model estimation and re-sampling. In the first one, SURF (*Speeded-Up Robust Features*) detector [18] was used to find and extract features from shred and intact document images. These features are represented by their control points in Fig. 1 (b). Such strategy worked well to around 80% of shreds; to the remaining ones, we have manually selected control points. On the next step, the set of shred features were matched with the set of intact document features using the nearest neighbor-based matching strategy. Then, we computed the geometric transformation matrix to the matched control point

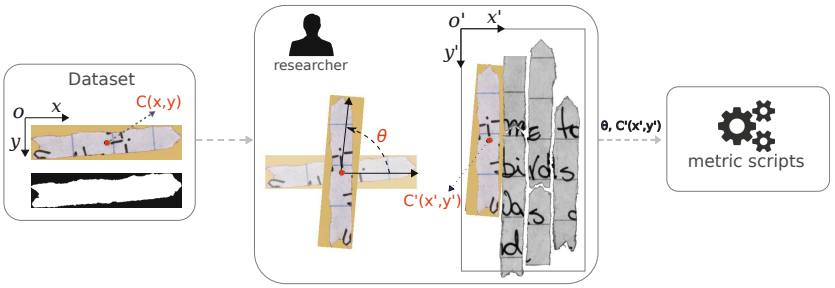


Fig. 2. Proposed assessment pipeline

pairs using the RANSAC algorithm [19] to eliminate outliers. Finally, we used the inverse of the geometric transform matrix to recover the positioning of the shred image in the original document image plane.

3.3 Assessment Pipeline

From the preprocessed 12 dataset documents, 12 correspondent image sets were generated (one set per shredded document), and are to be made available to the research community. Fig. 2 (a) illustrates a sample shred in our set, which has two image files associated to it: one for its colored version, and one for a binary mask discerning its format and position. In addition, there is also a text file whose lines individually contain the *id* of the related shred, followed by its *area* (in number of pixels), and by the x and y coordinates of its centroid C.

The reconstruction's results should be a simple text file, where each line should have the shred id, x' and y' coordinates of its centroid C' on the space of the proposed reconstruction, and the rotation angle θ necessary to reposition the fragment on such space (Fig. 2 (b)). To complete the assessment pipeline (Fig. 2), just use this paper's provided scripts, which implement and calculate the metrics of Section 4, comparing your solutions to the established ground-truth.

4 Metrics

This section presents two metrics for the cross-cut shredded document reconstruction problem. One of them uses the concepts of precision and recall which are common in information retrieval areas. The other estimates the minimum number of edit operations needed to rectify an arbitrary reconstruction.

We look at a reconstruction as a mapping of shreds $F = \{f_1, f_2, ..., f_n\}$ into $\mathbb{R}^3$. For each shred, the reconstruction assigns a coordinate centroid and an angle that informs how much the shred was rotated from its initial position. Two shreds f_i, f_j are neighbors if their mapping is placed side-by-side such that there are at least two points on their contours that lie close without other pieces between them. We define a reconstruction graph $G = (V, E)$ as:

- The vertex is $V = \{v_1, v_2, ..., v_n\}$, such that $v_i \in \{1, \ldots, n\}$, $i \neq j \leftrightarrow v_i \neq v_j$, and v_i represents f_i.
- The edge set $E = \{v_i, v_j\}$, such that $v_i, v_j \in V$, $i \neq j$, and f_i is a neighbor of f_j, we also use $e_{i,j} = (v_i, v_j)$.

We denote as $G_{gt} = (V, E_{gt})$ the reconstruction graph generated by using the ground-truth information. We denote as $G_s = (V, E_s)$ the reconstruction graph that represents a given reconstruction which we want to evaluate. They have the same vertex set V and different sets of edges E_{gt} and E_s.

4.1 Precision and Recall

The first metric is called *precision* and indicates how much of the neighborhood in G_s is correct according to the ground-truth G_{gt}. We start by counting the number of edges $e_{i,j}$ that are in E_{gt} and E_s. Those edges are what we call correct neighborhood. Then, we divide this number by the total amount of edges in E_s, which gives us a number in the range $[0, 1]$. If precision is close to zero, the reconstruction is poor – it is grouping together shreds that were not side by side in the original document.

$$precision = \frac{|\{e_{i,j} : e_{i,j} \in E_s \wedge e_{i,j} \in E_{gt}\}|}{|E_s|}. \tag{1}$$

The second metric is called *recall* and indicates how much of the neighborhood in G_{gt} was computed by the algorithm that reconstructed the document. We again use the correct neighborhood, but in this case we divide it by the total amount of edges in E_{gt}, which also gives us a number in the range $[0, 1]$. If recall is close to zero, the reconstruction is also poor, because it is failing on grouping together shreds that were side by side in the original document.

$$recall = \frac{|\{e_{i,j} : e_{i,j} \in E_s \wedge e_{i,j} \in E_{gt}\}|}{|E_{gt}|}. \tag{2}$$

The precision and recall rates were combined by simple arithmetic average, denoted by $PairingScore = (precision + recall)/2$. This measure is useful because it estimates the reconstruction quality in a single number in the range $[0, 1]$.

4.2 Number of Editions

In this section we present a measure of dissimilarity between an evaluated reconstruction and the ground-truth reconstruction. We compute the number of edit operations necessary, *EDOP*, to bring an arbitrary reconstruction to the ground-truth reconstruction. This measure is based on the notion of *graph edit distance*. Since graphs G_s and G_{gt} have the same set of vertices, the editing operations are restricted to insertion and deletion of edges in the graph G_s. The number of deletion operations DOP is equal to the number of edges in G_s that are not in G_{gt},

$$DOP = |\{e_{i,j} : e_{i,j} \in E_s \wedge e_{i,j} \notin E_{gt}\}|. \tag{3}$$

The number of insertion operations IOP is equal to the number of edges in G_{gt} that are not in G_s,

$$IOP = |\{e_{i,j} : e_{i,j} \in E_{gt} \wedge e_{i,j} \notin E_s\}|. \quad (4)$$

The total amount of edits, $EDOP$, to transform an arbitrary reconstruction into a ground-truth reconstruction is just the sum of the DOP and IOP values.

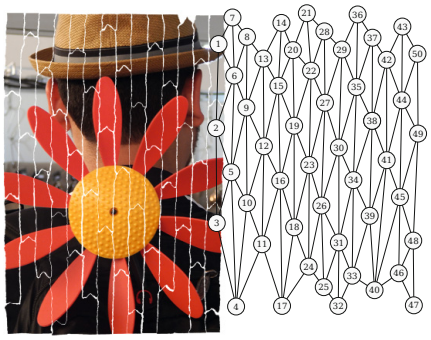
5 Use Cases

In this section, we illustrate our metrics with 50-piece toy-reconstruction examples. The ground-truth reconstruction established to the toy use case is shown in Fig. 3(a), and its corresponding reconstruction graph G_{gt} is presented in Fig. 3(b). There is a vertex in G_{gt} for each shred, and edges represent the relationship between neighbor shreds. The metric values of these reconstructions are in Table 1.

In the first case (Fig. 4), we intended to measure how the metrics behave in situations where some shreds were left out of the proposed reconstruction, but all the remaining others had no mismatch. Thus, we made a reconstruction (Fig. 4 (a)) that differed from the ground-truth baseline in 10 shreds, properly removed from the reconstructed image. As shown on Fig. 4(b), the graph for this case did not have edges connecting the vertices related to such shreds. In that scenario, the *recall* value was 0.64, emphasizing the absence of some shreds in the reconstruction. $precision = 1$ indicates that all the shreds used in the solution were correct. Additionally, the total number of editions ($EDOP = 43$) is the number of edges that should be added to the graph ($IDOP = 43$), since no incorrect pairing was made in the reconstruction.

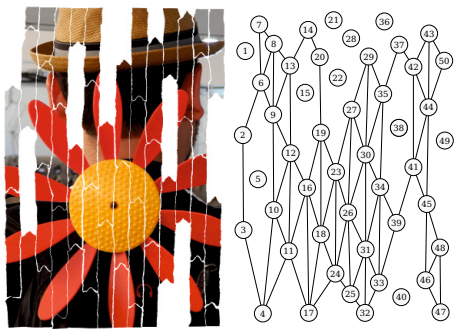
In the second case (Fig. 5), we evaluate how the metrics behave in situations where there was a large percentage of shreds left out of the proposed reconstruction, and others have been placed incorrectly on the solution. For that, we have made a reconstruction (Fig. 5 (a)) that differed from the ground-truth baseline on 25 shreds: 20 were missing in the reconstructed image, and 5 were intentionally placed in wrong positions. As one can see in Fig. 5(b), we highlighted in red the vertices related to the 5 shreds wrongly placed in the solution, as well as the 20 fragments not used resulted in 20 disconnected vertices. In this scenario, the *recall* value was 0.32, since there was a larger amount of missing shreds. The *precision* value was 0.7, indicating that some shreds had a wrong neighborhood. Finally, in the solution proposed in case 2, 81 edges should be added, and 16 others should be extracted to transform it in the correct reconstruction. In numbers: $EDOP = 97$, $IOP = 81$ and $DOP = 16$.

Comparing the reconstructions of the two scenarios, the metrics indicate that the first is better than the second in terms of correct $pairingScore$ and amount of editing, which concurs with a visual assessment of the reconstruction's quality.



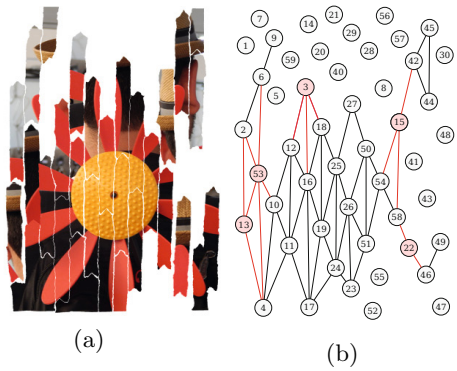
(a) (b)

Fig. 3. Illustrative scenario of a document that was fragmented in 50 cross-cut shreds. In (a), ground-truth reconstruction. In (b), related reconstruction graph G_{gt}.



(a) (b)

Fig. 4. Case 1: missing pieces. In (a), proposed reconstruction. In (b), reconstruction graph.



(a) (b)

Fig. 5. Case 2: missing and wrongly placed pieces. In (a), evaluated reconstruction. In (b), reconstruction graph.

Table 1. Metric values obtained in the scenarios

	Case 1	Case 2
recall	0.64	0.32
precision	1	0.7
pairingScore	0.82	0.51
IOP	43	81
DOP	0	16
EDOP	43	97

6 Conclusions

This paper introduces three graph-theory inspired quantitative metrics to evaluate the reconstruction of shredded documents. For a new reconstruction, we build a neighborhood graph based on the positioning of the fragments, and compare it to the ground-truth neighborhood graph. To make these concepts useful to the community, we provide a series of scripts to evaluate your reconstructions and a new unpublished dataset with 12 (at this moment) real cross-cut shredded documents containing handwriting, printed text, and images and photographs. The dataset has not only the shredded documents, but also their digital representation and reconstruction ground-truth.

Although powerful, these metrics might not be exhaustive for the shredded document problem, but by providing the code, and dataset to evaluate your

algorithms, we strongly believe we are empowering the community to do strong strides in the solution of real cross-cut document de-shredding.

References

1. Deever, A., Gallagher, A.: Semi-automatic assembly of real cross-cut shredded documents. In: 19th IEEE Intl. Conf. on Image Processing, pp. 233–236 (2012)
2. Zhang, H., Lai, J.K., Bcher, M.: Hallucination: A mixed-initiative approach for efficient document reconstruction. In: 4th Human Computation Workshop, pp. 54–60 (2012)
3. Butler, P., Chakraborty, P., Ramakrishan, N.: The Deshredder: A visual analytic approach to reconstructing shredded documents. In: 2012 IEEE Conference on Visual Analytics Science and Technology, pp. 113–122 (2012)
4. Geller, T.: DARPA shredder challenge solved. J. Commun. ACM. 55, 16–17 (2012)
5. DARPA Shredder Challenge, `http://archive.darpa.mil/shredderchallenge/`
6. Justino, E., Oliveira, L.S., Freitas, C.: Reconstructing shredded documents through feature matching. J. Forensic Science Intl. 160(2–3), 140–147 (2006)
7. Zhu, L., Zhou, Z., Hu, D.: Globally consistent reconstruction of ripped-up documents. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1–13 (2008)
8. Stieber, A., Schneider, J., Nickolay, B., Krüger, J.: A contour matching algorithm to reconstruct ruptured documents. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) Pattern Recognition. LNCS, vol. 6376, pp. 121–130. Springer, Heidelberg (2010)
9. Richter, F., Ries, C.X., Cebron, N., Lienhart, R.: Learning to reassemble shredded documents. J. IEEE Transactions on Multimedia. 15, 582–593 (2013)
10. Marques, M.A.O., Freitas, C.O.A.: Reconstructing strip-shredded documents using color as feature matching. In: 24th ACM Symp. on Applied Computing, pp. 893–894 (2009)
11. Lin, H.-Y., Fan-Chiang, W.-C.: Reconstruction of shredded document based on image feature matching. J. Expert Syst. with Applic. 39, 3324–3332 (2012)
12. Ranca, R., Murray, I.: A Composable strategy for shredded document reconstruction. In: Wilson, R., Hancock, E., Bors, A., Smith, W. (eds.) CAIP 2013, Part II. LNCS, vol. 8048, pp. 324–331. Springer, Heidelberg (2013)
13. Prandtstetter, M., Raidl, G.R.: Meta-heuristics for reconstructing cross cut shredded text documents. In: 11th Annual Conf. on Genetic and evolutionary computation, pp. 349–356 (2009)
14. Schauer, C., Prandtstetter, M., Raidl, G.R.: A memetic algorithm for reconstructing cross-cut shredded text documents. In: Blesa, M.J., Blum, C., Raidl, G., Roli, A., Sampels, M. (eds.) HM 2010. LNCS, vol. 6373, pp. 103–117. Springer, Heidelberg (2010)
15. Sleit, A., Massad, Y., Musaddaq, M.: An alternative clustering approach for reconstructing cross cut shredded text documents. J. Telecommunication Systems. 52, 1491–1501 (2013)
16. Biesinger, B., Schauer, C., Hu, B., Raidl, G.R.: Enhancing a genetic algorithm with a solution archive to reconstruct cross cut shredded text documents. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) EUROCAST. LNCS, vol. 8111, pp. 380–387. Springer, Heidelberg (2013)
17. Liu, Y., Qiu, H., Lu, J., Fang, Y.: Shredded document reconstruction based on intelligent algorithms. In: Intl. Conf. on Computational Science and Computational Intelligence, pp. 108–113 (2014)
18. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: SURF: Speeded up robust features. J. Computer Vision and Image Understanding. 110, 346–359 (2008)
19. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. J. Commun. ACM. 24, 381–395 (1981)