Student Research Abstract: Reconstructing shredded documents

Razvan Ranca
University of Edinburgh
10 Crichton Street
Edinburgh, UK
+44-797-5521236
ranca.razvan@gmail.com

ABSTRACT
This paper looks at the challenges involved in the automatic reconstruction of strip (vertically cut) and cross-cut (both vertically and horizontally cut) shredded documents. The unshredding problem is of interest in the fields of forensics, investigative sciences, and archeology. A novel probabilistic score function is proposed and shown to perform well in comparison with the standard cost functions used in literature. Additionally, several tractable search heuristics are analyzed. The paper attempts to identify which aspect of the reconstruction pipeline is currently the bottleneck and therefore which element can most benefit from future work. The fundamental security of several types of shredders is also touched upon.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]:Problem Solving, Control Methods, and Search – g*raph and tree search strategies, heuristic methods.*
I.4.9 [**Image processing and computer vision**]:Applications.
I.7.m [**Document and text processing**]: Miscellaneous.

General Terms
Algorithms, Security.

Keywords
Document Reconstruction, Shred, Cross-Cut

1. 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 but, because of convenience or for legal reasons, many sensitive documents are still printed and eventually shredded.

The recent DARPA Unshredding challenge has shown that shredders may not be as secure as people normally think. Further research in this area could help with such notable projects as the current effort to recover the shredded archives of the East German secret police. These consist of a total of 16,000 bags of shredded documents and, so far, it took three dozen people six years to reconstruct 300 of them. At that rate, it would take 11,520 man-years to finish this project.

2. LITERATURE REVIEW
In previous work [1, 2, 3, 4, 5], the solution to the reconstruction problem has been formulated with the help of two functions.
A cost function is used to represent how well two pieces match, the values varying from 0 for a perfect match to infinity for an impossible one. A search function is then used to find a permutation that minimizes the sum of the costs of all adjacent pieces. This permutation is considered the most likely solution.

Significant effort has been made towards finding a good search function. In [1] it is shown that the problem is NP hard by reduction to the Traveling Salesman Problem. In [2] an exact solution is attempted via Integer Linear Programming which yields very good results but is intractable for any number of strips above 150. Furthermore, several attempts [3, 4] have been made at applying evolutionary algorithms to the search problem.

In contrast, relatively little progress has been made in developing the cost function. Most papers settled on a simple formulation which does a weighted difference of each pair of adjacent pixels on either side of the proposed join and increases the cost of the join if the pixels are too dissimilar. See [1] for details.
One refinement for the cost function is proposed in [5], where the authors notice that the previous formulation places too much emphasis on matching adjacent white pixels which can obscure the real solution (see Figure 1). As such they base the cost on how well the black pixels match and also add a heuristic which looks at the positioning of the rows of text in the proposed match.

Figure 1. Left: correct match. Right: white-on-white match which would get a perfect cost and be incorrectly chosen

3. PROBABILISTIC SCORE
This paper proposes a departure from the previous cost function definitions, looking instead at using a probabilistic model to directly estimate the likelihood of two edges matching.
The central idea is to estimate the probability that a pixel is white or black given several of its neighbor pixels (the pixel's context). These conditional probabilities are estimated by analyzing the distribution of pixels in the shreds the algorithm is trying to piece together, which should be representative of the distribution in the reconstructed document.

Once these conditional probabilities are known, the probability of two edges matching can be calculated by sliding the context down the proposed edge and multiplying all the individual probabilities. Additionally, there is exactly one correct match for every edge, so the sum of the probabilities of all matches along one edge should sum up to one. In order to satisfy this condition the probabilities are normalized along every edge, which ends up mitigating the predisposition towards whitespace that some of the other cost functions suffer from. This is because there are usually many white edges available at any time and the probability mass will therefore be diluted between them. (see Table 1).

Table 1. After normalization, edge 2 is considered a better bet than edge 1, which could match equally well with pieces 1,2,3

	Raw probabilities			Normalized probabilities		
	Piece 1	Piece 2	Piece 3	Piece 1	Piece 2	Piece 3
Edge 1	1.00	1.00	1.00	0.33	0.33	0.33
Edge 2	0.50	0.10	0.10	0.71	0.14	0.14

Preliminary results suggest that this probabilistic score compares favorably to the cost functions used in [1] and [5] (see Figure 2)

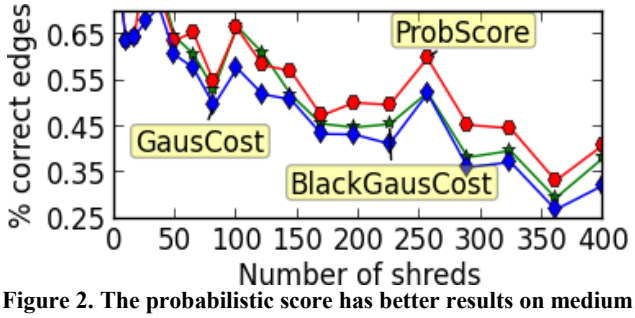

Figure 2. The probabilistic score has better results on medium and large instances than the cost functions

4. HEURISTIC SEARCH

Extending the graph-based heuristics introduced in [5], this paper looks at a Kruskal variant which tries to always pick the most probable edge and match the two pieces along that edge. This method creates multiple clusters of pieces which it must be able to later merge. If the best edge it has found would lead to a merge of two clusters, the potential match is checked for a superimposition of pieces, which would lead to the merge being rejected.

A problem with the Kruskal heuristic presented here as well as the heuristics from [5] is that they cannot correct previous errors and therefore are prone to a cascading effect (see Figure 3).

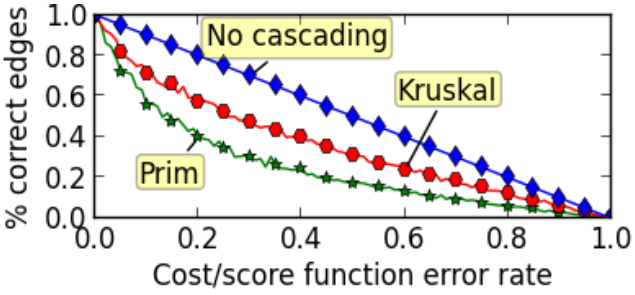

Figure 3. Error of cost vs. error of search for the heuristics

5. CONCLUSIONS & FUTURE WORK

Preliminary results, which assume the correct orientation of the shreds is known a priori, are shown (see Figure 4)

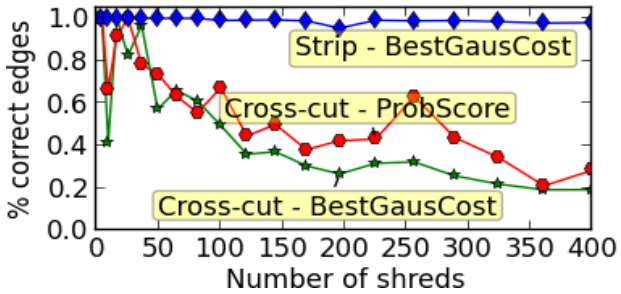

Figure 4. Performance on cross-cut shreds shown with the probabilistic score and the best performing cost function. Performance on strip shreds shown with the best cost function (the probabilistic score performs similarly, and is not shown)

The strip shred variant seems significantly easier to solve than the cross-cut version. This discrepancy can be explained by the fact that the shorter edges produced by cross-cutting are much more prone to noise because of the reduced number of edge pixels. Additionally, for text documents, horizontal cuts have a significant chance of falling between two lines of text, in which case the cost/score function has no information on how to order the lines. The problems with the cost/score functions are compounded by the cascading effect observed in the search heuristics which require an accuracy of at least 95% in the cost/score function to obtain an 80% accuracy in the solution, and the cascading only gets worse as the size of the problem increases.

Future work will look at ameliorating the cascading effect by introducing an error correcting heuristic. The possibility of incorporating higher level features, such as computer vision techniques, into the cost/score function should also be considered. Here the probabilistic score function has an advantage in that it can be easily coupled with other probabilistic models.

6. REFERENCES

[1] Prandtstetter, M., and Raidl, G. - Combining Forces to Reconstruct Strip Shredded Text Documents *In HM '08: proceedings of the 5th international workshop on hybrid metaheuristics (pp. 175–189). Berlin: Springer.*

[2] Prandtstetter, M. - Two Approaches for Computing Lower Bounds on the Reconstruction of Strip Shredded Text Documents *Technical Report TR–186–1–09–01, Technishe Universitat Wien, Institut für Computergraphik und Algorithmen, 2009*

[3] Prandtstetter, M., and Raidl, G. - Meta-Heuristics for Reconstructing Cross Cut Shredded Text Documents *In GECCO '09:proceedings of the 11th annual conference on genetic and evolutionary computation (pp. 349–356). New York: ACM Press.*

[4] Schauer, C., Prandtstetter, M. and Raidl, G. - A Memetic Algorithm for Reconstructing Cross-Cut Shredded Text Documents *In Hybrid Metaheuristics, pp. 103–117, 2010.*

[5] Sleit, A., Massad, Y. and Musaddaq M. - An alternative clustering approach for reconstructing cross cut shredded text documents *Telecommunication Systems, 2011 - Springer*