A Modular Framework for the Automatic Reconstruction of Shredded Documents

Razvan Ranca
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, Scotland, United Kingdom
ranca.razvan@gmail.com

Abstract

The unshredding problem is of interest in various domains such as forensics, investigative sciences and archeology, and has therefore been approached in many different ways. This paper tries to bridge the gap between previous, disparate, efforts by proposing a modular, probabilistic, solution. Novel approaches to two of the independent subproblems are presented and shown to both have good theoretical properties and to empirically outperform previously published methods.

Introduction

The US Federal Trade Commission recommends the shredding of sensitive documents as a good protection method against becoming one of the millions of victims of identity theft[1]. However, the successfully solved DARPA Shredder challenge[2] and the appearance of commercial document reconstruction services[3], puts the security of the shredder into question. Further research in this area will benefit projects such as the ongoing effort to recover the shredded archives of the East German secret police, which consist of 16,000 bags of shredded documents. Since it took 3 dozen people 6 years to reconstruct 300 of these bags (Heingartner 2003), at this rate, the project would require 11,520 person-years.

The unshredding challenge can be approached by splitting the problem into three independent subproblems. A *pre-processing* step transforms the noisy images of scanned shreds into uniform "ideal shreds". A *scoring function* then evaluates the potential match between every pair of shreds and, finally, a *search method* finds the permutation of shreds that obtains a good global score. In this paper I present a novel, composable, probabilistic scoring function and an original, graph-inspired, search heuristic which balances speed and accuracy while achieving modularity.

Related Work

Significant effort has been made towards finding a good search function. (Prandtstetter and Raidl 2008) formally define the search subproblem and show that it is NP hard. (Prandtstetter 2009) attempts an Integer Linear Programming solution which yields good results but is intractable for more than 150 shreds. (Prandtstetter and Raidl 2009; Schauer, Prandtstetter, and Raidl 2010) try instead to formulate the search as an evolutionary algorithm.

In contrast, relatively little progress has been made in developing the score function. All previously mentioned papers settled on a formulation which selects a score based on a weighted difference of the adjacent pixels on either side of the proposed join. (Biesinger 2012) provides a formal definition. (Sleit, Massad, and Musaddaq 2011) refine this method by placing an emphasis on black pixels, thus discounting the information content of white pixels. (Perl et al. 2011) try a different approach, by employing optical character recognition techniques. Their method is however left as a proof-of-concept, since it is not integrated with a search function or evaluated against any other scoring methods.

Finally, pre-processing can be split into several independent functions, which have been previously explored. For instance, (Skeoch 2006) extracts the shreds from scanned input images via rectangle and polynomial fitting, while (Butler, Chakraborty, and Ramakrishan 2012) fix the skew of the extracted shreds by framing the issue as an optimization problem. Up/down orientation of documents is explored in (Caprari 2000; Aradhye 2005) with good results, though the methods are only evaluated on full documents, not shreds.

Probabilistic Score

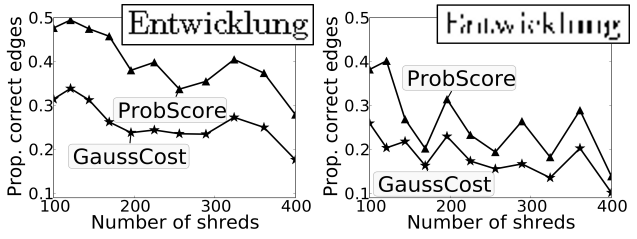
I propose a novel score function formulation, which uses a probabilistic model to directly estimate the likelihood of two edges matching. Employing a probabilistic model offers several advantages, such as an increase in robustness given by the ability to train the model on the document shreds and easy composability of different models, simply achieved by multiplying their probabilities and re-normalizing.

I test this idea by implementing a basic probabilistic model, based on the conditional probability that a pixel is white or black given a few of its neighboring pixels. Formally, given edge *Et*, *ProbScore* returns the best match for Et and the probability of that match. ProbScore is defined:

[1] http://consumer.ftc.gov/features/feature-0014-identity-theft
[2] http://archive.darpa.mil/shredderchallenge/
[3] eg: http://www.unshredder.com/

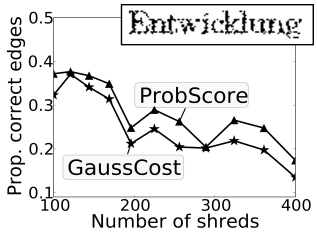
procedure PROBSCORE(Et)
  Initialize ps ▷ *probabilities of matches, initially all 1*
  for all $Ex \in Edges$ **do**
    for all $pixel \in Ex$ **do**
      $ps_{Ex} \leftarrow ps_{Ex} * \Pr(pixel|Neighbors_{Et}^{pixel})$
  Normalize ps ▷ *probabilities must sum up to 1*
  return $\arg\max ps, \max ps$

Empirical results show that this probabilistic score compares favorably to the most common scoring function used in literature (Biesinger 2012), both on noise-less documents and on several noisy documents (see Figures 1a, 1b, 1c). Additionally, in order to showcase the benefits offered by the modular nature of ProbScore, I compose it with another probabilistic model called *RowScore* which applies a Gaussian model to the distance between rows in neighboring shreds. Even such a simple model gives ProbScore a small but consistent boost (see Figure 1d). ProbScore could also be composed with more complex models, such as that proposed by (Perl et al. 2011).

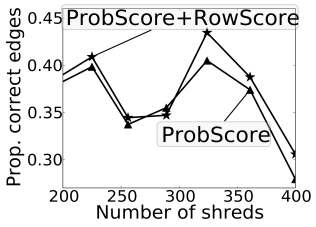


(a) Original document.

(b) Downsampled document.



(c) Noisy document.



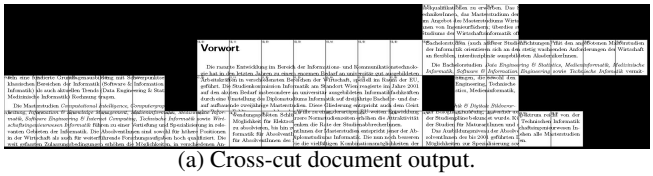
(d) Score function composition.

Figure 1: Figures **a**,**b** and **c** show comparisons between our method and the most common function used in literature. A sample word from each document is shown in the upper right corners. Figure **d** shows the improvement obtained by composing our function with another probabilistic model.

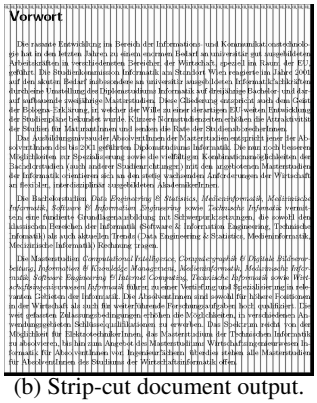
"Kruskal-Inspired" Heuristic Search

Extending the heuristic introduced in (Sleit, Massad, and Musaddaq 2011), I implement a search method inspired by the minimum spanning tree "Kruskal's algorithm" (Kruskal 1956). The method greedily unites the two best matching shreds, as indicated by the scoring function. This process creates multiple clusters of shreds which will eventually be merged into a single solution. Before performing a merge, I check if the move would result in two shreds being superimposed, in which case the merge is aborted.

This method outperforms previously proposed bottom-up heuristics but is still significantly more tractable than any of the top-down optimizing search functions. A novel aspect of the heuristic is that, if the next move is uncertain, execution can be stopped. This functionality achieves my overarching goal regarding modularity, since it means the method can reduce the search space of a problem, such that a more complex search function may become feasible (see Figure 2).



(a) Cross-cut document output.



(b) Strip-cut document output.

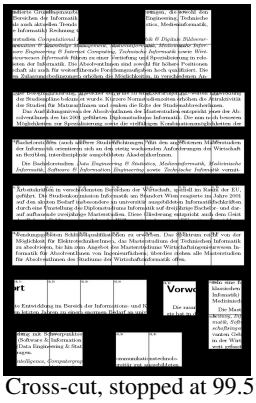

(c) Cross-cut, stopped at 99.5%.

Figure 2: Figures **a** and **b** show full reconstructions on the cross-cut variant (64% correct) and the strip-cut variant (100%correct). Figure **c** shows a partial reconstruction (stopped at 99.5% certainty) which successfully reduces the search space from 49 to 10 shreds while introducing 0 errors.

It's worth noting that cross-cut documents are significantly harder to solve than strip-cut ones. This is due to the short edges produced by cross-cutting, which are harder to model accurately. Horizontal cuts also have a significant chance of falling between two lines of text, in which case the score function has no information on how to order the lines.

Conclusions and Future Work

This paper presents a modular and composable framework for the shredded document reconstruction problem and provides sample solutions for 2 of its 3 components. Specifically, I propose a probabilistic scoring function which outperforms currently used alternatives and a tractable search heuristic which can solve simpler reconstruction problems and reduce the search space for more complex ones.

Future work will look at implementing more advanced score and search functions. Solving the cross-cut domain will likely require scoring functions which employ computer vision techniques and search function which perform a partial exploration of the search tree. The performance of the independent, pre-processing, components will also be tested as the size of their input documents decreases.

References

Aradhye, H. 2005. A generic method for determining up/down orientation of text in roman and non-roman scripts. *Pattern Recognition* 38(11):2114–2131.

Biesinger, B. 2012. Enhancing an evolutionary algorithm with a solution archive to reconstruct cross cut shredded text documents. Bachelor's thesis, Vienna University of Technology, Austria.

Butler, P.; Chakraborty, P.; and Ramakrishan, N. 2012. The deshredder: A visual analytic approach to reconstructing shredded documents. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, 113–122. IEEE.

Caprari, R. 2000. Algorithm for text page up/down orientation determination. *Pattern Recognition Letters* 21(4):311–317.

Heingartner, D. 2003. Back together again. *New York Times*.

Kruskal, J. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* 7(1):48–50.

Perl, J.; Diem, M.; Kleber, F.; and Sablatnig, R. 2011. Strip shredded document reconstruction using optical character recognition. In *Imaging for Crime Detection and Prevention 2011 (ICDP 2011), 4th International Conference on*, 1–6. IET.

Prandtstetter, M., and Raidl, G. 2008. Combining forces to reconstruct strip shredded text documents. In *Hybrid Metaheuristics*. Springer. 175–189.

Prandtstetter, M., and Raidl, G. 2009. Meta-heuristics for reconstructing cross cut shredded text documents. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, 349–356. ACM.

Prandtstetter, M. 2009. Two approaches for computing lower bounds on the reconstruction of strip shredded text documents. Technical Report TR18610901, Technishe Universitat Wien, Institut fur Computergraphik und Algorithmen.

Schauer, C.; Prandtstetter, M.; and Raidl, G. 2010. A memetic algorithm for reconstructing cross-cut shredded text documents. In *Hybrid Metaheuristics*. Springer. 103–117.

Skeoch, A. 2006. An investigation into automated shredded document reconstruction using heuristic search algorithms. Ph.D. thesis, University of Bath, UK.

Sleit, A.; Massad, Y.; and Musaddaq, M. 2011. An alternative clustering approach for reconstructing cross cut shredded text documents. *Telecommunication Systems* 1–11.