Unveiling the Secrets: Reconstruction of Shredded Documents using Deep Learning

Thiago M. Paixão[†*] , Maria C. S. Boeres[‡], and Thiago Oliveira-Santos[‡]
[†]Instituto Federal do Espírito Santo (IFES)
[†]Email: thiago.paixao@ifes.edu.br
[‡]Universidade Federal do Espírito Santo (UFES)

***Abstract*—This work addresses the intricate task of reconstructing mechanically-shredded documents with potential application in forensic investigation. Our primary contributions consist of two novel deep learning approaches for fully automatic reconstruction tested on real-world shredded data that achieved state-of-the-art accuracy in more realistic scenarios. As a second major contribution, we introduce a novel framework for semi-automatic reconstruction inspired by the principles of active learning. The core of our proposal is a recommendation module that smartly flags potential errors in the reconstruction output (permutation of shreds) for human review, enabling even more enhanced reconstructions. The mentioned contributions and additional outcomes (datasets and experimental protocols) resulted in five relevant publications: three journal articles and two international conferences, including the premier IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR).**

I. Introduction

Paper shredders have become a popular instrument to discard large amounts of documents regarding privacy issues. However, in some scenarios, the sensitive content of such documents must be revealed, demanding some sort o reconstruction process. It was through manual reconstruction of shredded documents that investigative journalists exposed the influence-buying scandal involving a Korean ambassador and members of the American Congress [1] (Figure 1). Typical reconstruction cases may involve thousands of shreds, therefore, computational reconstruction [2], [3] is of great value for forensic investigators, historians, or even individuals who, amidst a collection of numerous irrelevant documents, have inadvertently damaged crucial ones.

A standard reconstruction framework comprises two primary tasks (Figure 2): pairwise compatibility evaluation of the shreds based on image analysis and an optimization procedure to deliver the final solution (a permutation of shreds in the case of strip shredding). Most of the advances in reconstructions are related to optimization algorithms [4], [5], and a more in-depth survey of the literature has revealed a pronounced demand for image processing techniques that can effectively verify compatibility for real-shredded data. Several works apply *low-level* (dis)similarity measures (*e.g.* Euclidean metric) on boundary pixels [6], [7]. Nonetheless, the intrinsic pixel data is susceptible to substantial corruption due to the mechanical shredding process. As an alternative, *shape-based* fitting has been proposed to leverage higher-level information [8], [9]. Notably, our character-matching technique yielded the most promising outcomes [9][1], however, the reconstruction accuracy depends significantly on the density of textual content. *Supervised learning* has been utilized for symbol recognition in order to determine the fitting degree of the shreds. However, the applicability of this approach is substantially constrained because of two main reasons: (i) inherent instability in recognition accuracy due to the document corruption [8], and (ii) the reliance on specific languages [10].

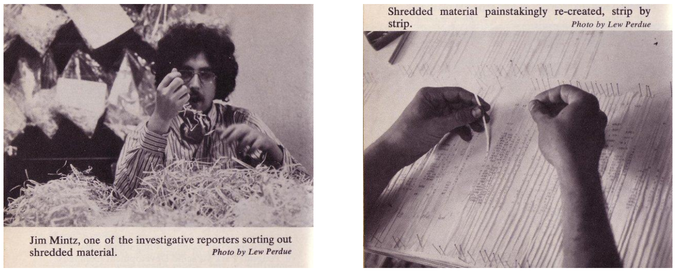

Jim Mintz, one of the investigative reporters sorting out shredded material.
Photo by Lew Perdue
Shredded material painstakingly re-created, strip by strip.
Photo by Lew Perdue

Fig. 1. Manual reconstruction of documents implying members of the American Congress in influence-buying scandal. Credits to Lewis Perdue [11].

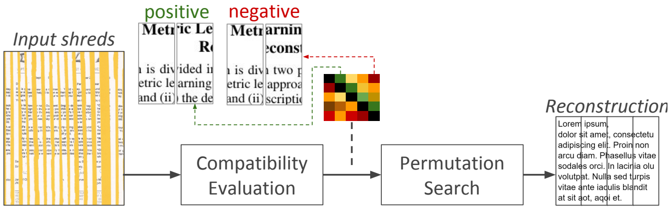


Fig. 2. Overview of a typical reconstruction framework. The assessment of compatibility between shreds occurs in a pairwise manner. The computed compatibility values drive the optimization search procedure which aims to determine the permutation of shreds that most accurately embodies the original document.

Our primary contributions extend to the domain of deep learning for compatibility evaluation, where we pioneered the exploration of deep models trained in a self-supervised manner for the reconstruction of shredded documents [12]–[14]. Unlike the competitors, the methods developed in our thesis were rigorously validated on real-world data, therefore, issues arising from physical shredding can not be simply

[*] This work presents the main contributions of the Ph.D. thesis defended by Thiago M. Paixão.

[1] This work carries the first relevant findings of our thesis, even though it does not fit the deep learning paradigm.

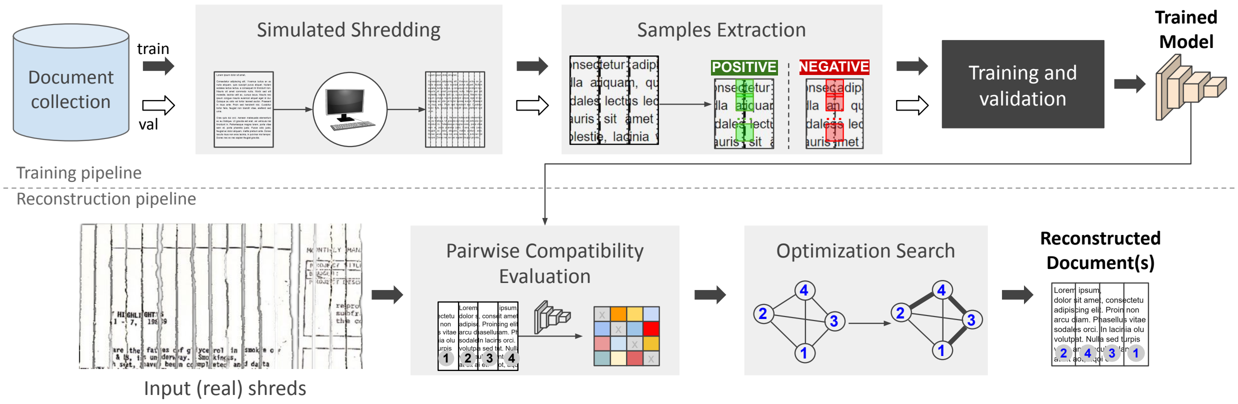


Fig. 3. Overview of the classification-based reconstruction approach. Top flow: self-supervised training of a deep model capable of estimating compatibility. Bottom flow: the reconstruction of shreds potentially from mixed documents.

overlooked. To enable the investigation in more real scenarios, we produced (and shared) a dataset comprising 120 shredded documents (2,797 shreds). To the best of our knowledge, this is the largest current collection of strip-shredded documents within this specific domain.

Human participation might be useful to obtain more accurate reconstructions [3], [15]. In this regard, we introduced an interactive framework where a human judge reviews pairs of adjacent shreds within a solution and decides whether they should stay together or be separated. The novelty of our proposal is the smart selection of cases where mistakes are more likely to occur, thereby maximizing the effectiveness of human effort.

To summarize, these are the **contributions** of our thesis discussed in the rest of this paper:

- **A deep classification-based approach** [12], [13] (Section II). The problem of evaluating compatibility is formulated as a classification problem using a deep model: the obtained results demonstrate the effectiveness of this approach in reconstructing 100 documents with an accuracy superior to 90%;
- **A deep metric-learning approach** [14] (Section III). This formulation decouples network inference from pairwise calculations, reducing drastically the time cost to compute compatibilities: approx. 22 times for 505 shreds, and more pronounced gains for larger numbers of shreds;
- **A human-in-the-loop reconstruction framework** [16] (Section IV). This is a scalable framework designed for interactive reconstruction of strip-shredded documents: our findings suggest that a manual review of just 25% of the solution (shreds permutation) may yield a reduction of over 40% in the mistakes resulting from using the metric learning approach.

Other contributions include: (i) a collection comprising a total of 120 documents (2,797 shreds); a novel methodology to assess multi-page (mixed-shreds) reconstruction; and (iii) a novel methodology to evaluate the impact of human labor on the reconstruction accuracy.

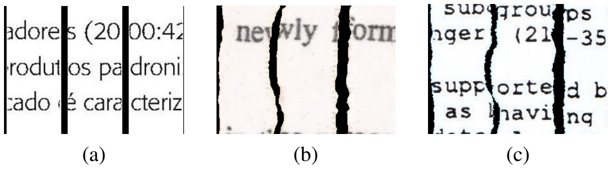

(a) (b) (c)

Fig. 4. Samples from the datasets used in the experiments: (a) S-MARQUES [18]; (b) S-ISRI-OCR [9]; (c) S-CDIP [13]

II. A DEEP CLASSIFICATION-BASED APPROACH

The underlying idea of the classification-based approach is training a model capable of estimating the (softmax) probability of two shreds being adjacent in the original document. The training and reconstruction pipelines are illustrated in Figure 3. In the training pipeline (top flow), a collection of digital documents is "artificially" shredded (simulated shredding), so that the adjacency relationship comes for free from the data. To put it plainly, one may say that the algorithm knows – without human assistance – which pairs are adjacent (positive class) and non-adjacent (negative class), enabling self-supervision. Small samples are extracted top-down from the image resulting from horizontally stacking two shreds. The model is trained to distinguish (classify) positive and negative samples.

In the reconstruction pipeline (bottom flow), the input comprises real digitized shreds whose compatibility is determined by the trained model (specific details are omitted due to page limit). The reconstruction instance is then modeled as a weighted asymmetric graph, where the nodes represent the shreds and an arc (i, j) with weight w indicates that the compatibility value between the shreds i and j, with i on the left of j, is w. This graph is converted into a Traveling Salesman Problem (TSP) instance so that an optimization solver [17] is applied to estimate the permutation of shreds representing the reconstructed document.

A. Results on multi-page reconstruction

A significant contribution of our thesis was the investigation on multi-page reconstruction, particularly, the experi-

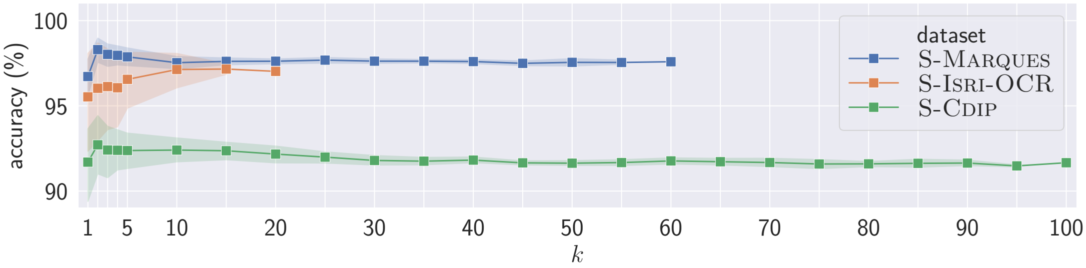


Fig. 5. Multi-page reconstruction accuracy: accuracy w.r.t. number of mixed pages (k).

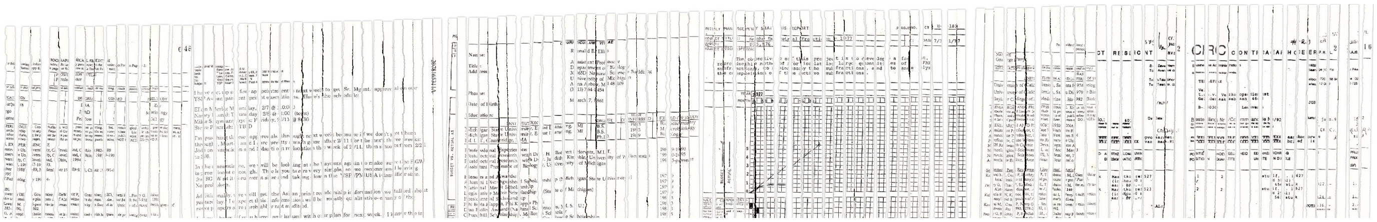


Fig. 6. Reconstruction of five pages from S-Cdip (accuracy of 82.93%). The full reconstruction involving all the 100 pages of this dataset is available at https://htmlpreview.github.io/?https://github.com/thiagopx/docs/blob/master/results_s-cdip.html.

mental protocol itself which involves cross-database training-test assessment, and progressive reconstruction of $k = 1, 5, 10, 20, \ldots, n_p$ pages, where n_p depends on the collection oh shredded documents. Experiments were conducted on three collections (Figure 4), from which two (S-Isri-OCR and S-Cdip) are contributions of our work. As shown in Figure 5, our method performs with accuracy higher than 90% for the three datasets with becomes stable as k increases. This observation suggests that the introduction of new documents has a negligible impact on the quality of the reconstruction process. S-Cdip comprises the most challenging instances due to its diverse content and intricate layout. Figure 6 exhibits the reconstruction of five pages from S-Cdip, which reached accuracy of 82.93%.

B. Comparative analysis

Our method (DeepRec-CL) was compared to three key techniques in the field, denoted by their lead authors: *Paix ao*, which represents our initial approach grounded in shape matching; *Liang*, an OCR-based technique [19]; and *Marques*, which relies on edge pixel dissimilarity [18]. Importantly, due to computational constraints, the Paix ao evaluation was confined to a dataset of only 5 documents, and the Liang experiments were restricted to the S-Marques and S-Isri-OCR datasets, encompassing a subset of 3 documents. To highlight the role of the compatibility evaluation, we tested our reconstruction framework with the Marques' nearest neighbor optimizer. This adapted version was named DeepRec-CL-NN. As depicted in Figure 7, the proposed method consistently outperformed the competitors in terms of average accuracy. Notably, our approach demonstrated heightened robustness, as evidenced by the stable trajectory of the accuracy curve. Remarkably, DeepRec-CL-NN significantly outperformed Marques, despite sharing the same optimizer, and also outperformed Paix ao, which employed a more powerful optimizer.

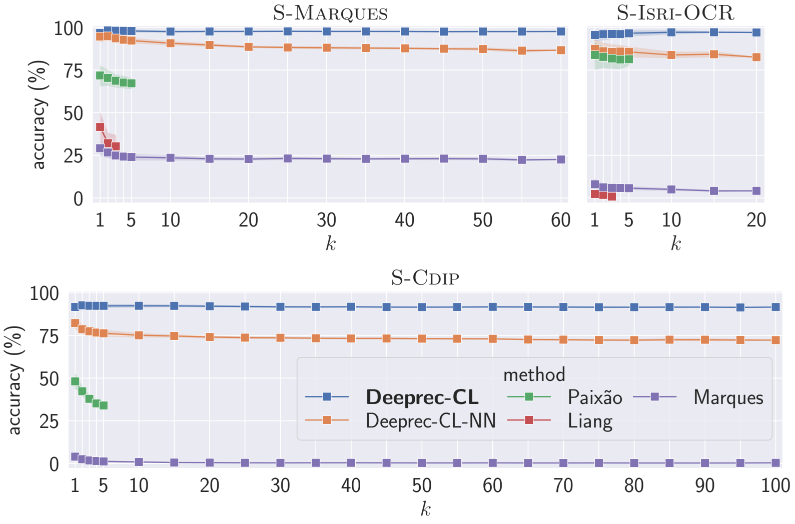
C. Time performance

DeepRec-CL also exhibits enhanced scalability in terms of execution time compared to Paix ao and Liang, as seen in Figure 8. This aspect is of great significance in practical scenarios since the input is expected to have far more than just five shredded pages. While Marques demonstrates desirable time efficiency, its compromised accuracy (Figure 7) is unsuitable for real-world data applications.

III. A Deep Metric-Learning Approach

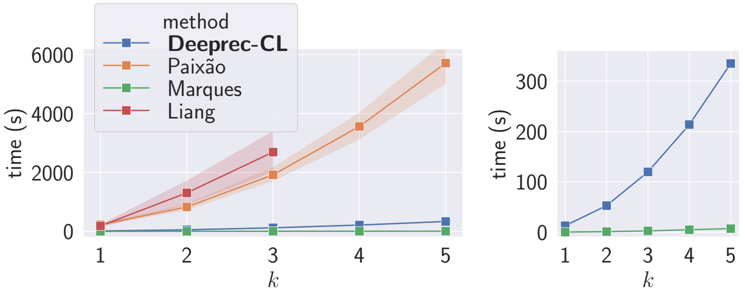
The time performance of DeepRec-CL depends heavily on the number of inferences performed during the compatibility evaluation. For a test instance of size n (*i.e.* the number of shreds), the number of inferences is $n(n-1) = O(n^2)$. In contrast, the novel metric-learning approach (DeepRec-ML) decouples the inference step from the pairwise compatibility evaluation. This strategic separation renders the inference cost linear with respect to the instance size. A neural network is trained in a self-supervised manner to produce embedded representations for the left boundary of the shreds, whereas a second network specializes in the right boundary. As a consequence, two inferences are performed for each shred, totaling $2n = O(n)$ inferences.

Figure 9 illustrates the underlying principle of the metric-learning approach. Similarly to DeepRec-CL, local samples ($\boldsymbol{x}$) are collected from boundary zones. Instead of applying raw pixel-to-pixel comparison, these samples are transformed into an intermediary representation ($\boldsymbol{e}$) through projection onto a shared embedding space $\mathbb{R}^d$. This projection is executed via two dedicated CNNs: f_{left} and f_{right}, with $f_\bullet : \boldsymbol{x} \mapsto \boldsymbol{e}$, where



(c)

Fig. 7. Comparative accuracy performance. Due to memory and processing consumption, the curves for Liang and Paixão were not fully computed.



(a)

(b)

Fig. 8. Comparative time performance.

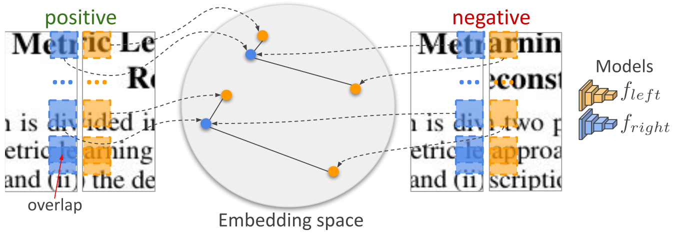


Fig. 9. Metric learning approach for shreds' compatibility evaluation.

the subscript indicates on which boundary (left or right) the network is specialized.

Considering a pair of shreds, each vertical position gives rise to a pair of samples, as depicted by the blue and orange squares in Figure 9. An ideal training would lead samples from positive pairings to be projected nearly in the embedded space, while embeddings produced from negative pairings would be set apart. Consequently, the overall compatibility of a pair is quantified by considering the distances between their corresponding embeddings.

1) Comparison with DEEPREC-CL.*:* The performance of DEEPREC-CL and DEEPREC-ML (the metric learning approach) was evaluated with respect to accuracy and time efficiency. For this analysis, the 1,370 shreds from S-MARQUES were mixed to compose the first instance. Similarly, the second instance consisted of the 505 shreds extracted from S-ISRI-OCR. DEEPREC-ML achieved accuracy of 94.81% and 97.22% for S-MARQUES and S-ISRI-OCR, respectively. In comparison, DEEPREC-CL achieved an accuracy of 97.08% and 95.24% for the corresponding datasets.

In general, both approaches produced high-quality reconstructions, and the variation in accuracy between them was minor (around ± 2 p.p.). This marginal difference suggests that the methods have similar performance in terms of accuracy. Concerning time efficiency, the methods behave notably differently, as evidenced in Figure 10. On the left, it is shown the elapsed time to reconstruct the 505 shreds from S-ISRI-OCR for each stage: projection (pro) – applicable only for DEEPREC-ML–, pairwise compatibility evaluation (pw), and the optimization process (opt). Clearly, the optimization expense was negligible in comparison to the execution time

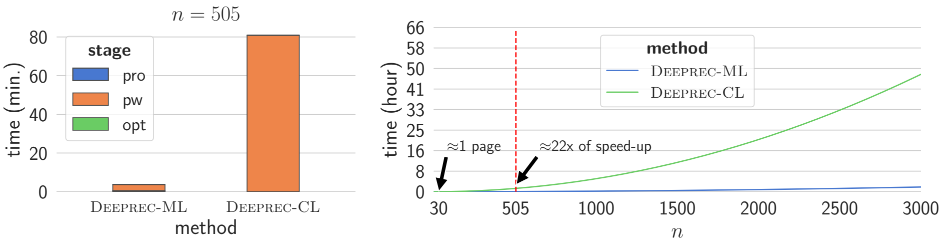


Fig. 10. Time performance for multi-page reconstruction.

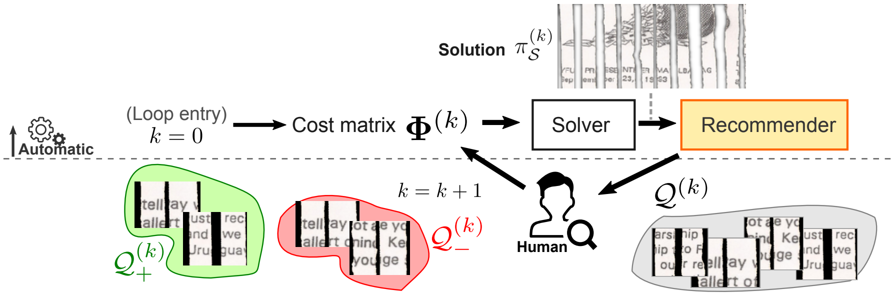


Fig. 11. Overview of the proposed HIL reconstruction framework. In each iteration, a recommender module indicates potential wrong pairings for human review.

required for pairwise evaluation. The overall time falls drastically from over 80 minutes to finalize the evaluation stage (DEEPREC-CL) to less than 4 minutes with the metric learning approach, resulting in a remarkable speed-up of approximately 22 times. Furthermore, as observed on the right chart, the estimated growth trajectory for DEEPREC-ML (represented by the blue curve) is noticeably slower than that of the alternative method.

IV. A HUMAN-IN-THE-LOOP RECONSTRUCTION FRAMEWORK (HIL)

The reconstruction can be enhanced by adding the human in the process. In our proposal, a human is required to review shreds (in pairs) which were attached after the optimization step. The reconstruction framework, inspired in the field of active learning [20], is structured as a repetitive procedure where, in each iteration, a *recommender* module flags potential mistakes for human review: assign the pair as positive or confirm that the shreds should remain unattached.

A. On the impact of the human effort (workload)

A permutation of the n shreds of a reconstruction instance results in $n - 1$ pairs of adjacent shreds. The human effort (workload) was defined as the percentage of those $n - 1$ pairs which has to be reviewed by the human referee. Figures 12 and 13 show the accuracy in function of the workload (α_{load}) for a single iteration of the framework configured with DEEPREC-ML and DEEPREC-CL, respectively. The reader can notice that accuracy growth is roughly linear as α_{load} increases. The proposed strategies to select shreds for review (OPT-R, OPT-RL, UNC-R, and UNC-RL) outperform significantly the random-choice baseline employed in [4]. OPT-R was able to increase the original solution accuracy of the S-CDIP dataset on approx. 3.80 p.p. when $\alpha_{load} = 0.25$, which means that 87/220 mistakes were eliminated (error reduction of approx. 39.50%).

B. Distributing the workload across iterations

We also investigated whether distributing the workload across a few iterations may yield a faster convergence: more corrections achieved with the same overall workload. Tests were conducted for up to three iterations due to the processing burden caused by running the solver multiple times. Figure 14 presents the results for the strategy OPT-R and the reconstruction method DEEPREC-ML. On initial observation, transitioning from one to two iterations led to a more consistent increase in accuracy. However, two or more iterations ($n_{iter} \geq 2$) seem to be advantageous for higher values of workload. As observed in the chart, the accuracy for S-CDIP increased by approx. 4.43 p.p. with two iterations, resulting in a substantial error reduction of about 46.10%.

V. CONCLUSION

This work discussed the contributions of our thesis on the field of reconstruction of documents fragmented by paper shredders. The major contribution lies in the robust self-supervised compatibility evaluation of shreds by using neural networks. Two approaches were proposed in this matter. The first models the compatibility evaluation as a classification problem, whereas the second fits in the metric learning paradigm. Additionally, it was discussed our proposal for semi-automatic reconstruction in which the human is required to provide some feedback on the results to improve the reconstruction accuracy. In future work, we plan to investigate reconstruction with more damaged and missing shreds, extend the proposed methods for cross-cut documents, adapt

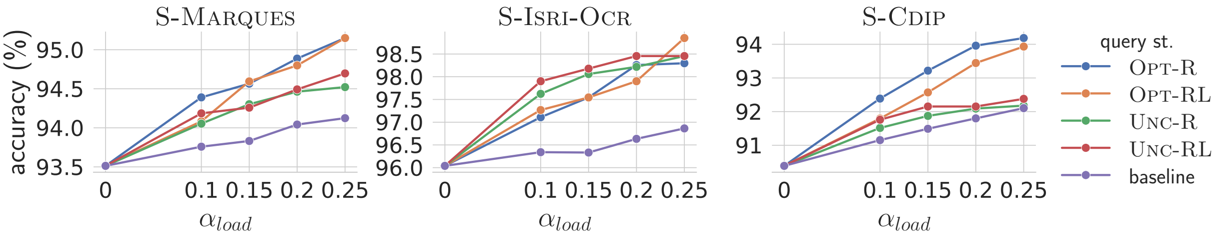


Fig. 12. Reconstruction accuracy w.r.t. workload (Deeprec-ML).

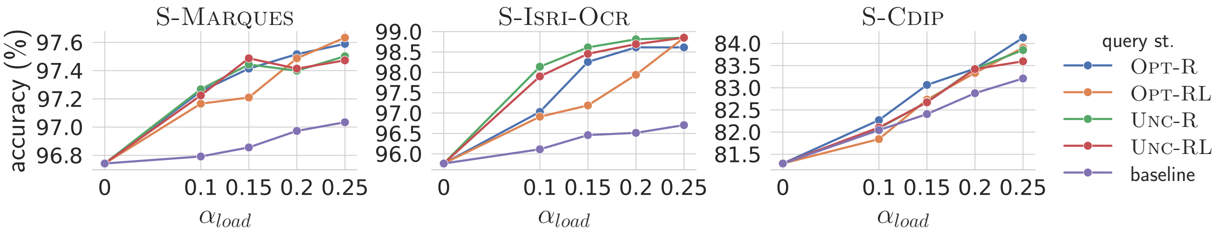


Fig. 13. Accuracy w.r.t. workload (Deeprec-CL).

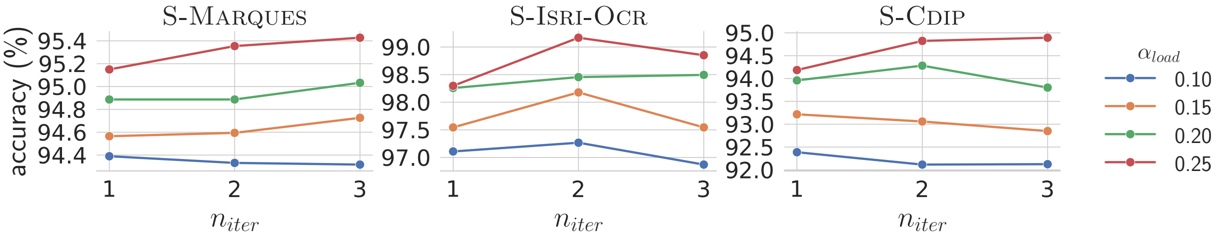


Fig. 14. Reconstruction accuracy w.r.t. the number of iterations (Deeprec-ML, Opt-R). Each curve represents a different workload (α_{load}).

and combine new active learning techniques for interactive reconstruction, and, finally, work on generalizing our methods for correlated problems (i.e., ancient papyrus reconstruction).

Acknowledgment

We thank all co-authors for their contributions. Special thanks to Alessandro L. Koerich and Rodrigo F. Berriel for their exceptional collaboration. We acknowledge the support of CNPq (process 310330/2020-3), as well as FAPES and CAPES (process 2021-2S6CD, FAPES no. 132/2021) through the PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Awards and Publications

This thesis received the third place award in the **XXXVI Concurso de Teses e Dissertações** (CTD-CSBC, 2023). It was also chosen by the Postgraduate Program in Computer Science (PPGI) at UFES to be their representative for the **Prêmio CAPES de Tese 2023**. The following listed publications directly resulted from this thesis:

- T. M. Paixão, M. C. S. Boeres, C. O. A. Freitas, and T. Oliveira-Santos, "Exploring Character Shapes for Unsupervised Reconstruction of Strip-shredded Text Documents," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 7, pp. 1744–1754, 2019
- T. M. Paixão, R. F. Berriel, M. C. S. Boeres, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "A deep learning-based compatibility score for reconstruction of strip-shredded text documents," in *Conf. on Graph., Patterns and Images*, 2018, pp. 87–94
- T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Self-supervised deep reconstruction of mixed strip-shredded text documents," *Pattern Recognit.*, vol. 107, p. 107535, 2020a
- T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. D. Souza, and T. Oliveira-Santos, "Fast(er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning," in *IEEE/CVF Conf. on Comp. Vision and Pattern Recognit.*, 2020b, pp. 14 343–14 351
- T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "A human-in-the-loop recommendation-based framework for reconstruction of mechanically shredded documents," *Pattern Recognit. Letters*, vol. 164, pp. 1–8, 2022

References

[1] C. R. Babcock, "Tongsun park's paper jigsaw puzzle solved," *The Washington Post*, 18 Sep 1977, available at: https://www.ft.com/content/6eba4d10-ba43-11e8-94b2-17176fbf93f5 (Accessed: April 24th, 2023).

[2] A. Ukovich, G. Ramponi, H. Doulaverakis, Y. Kompatsiaris, and M. Strintzis, "Shredded document reconstruction using MPEG-7 standard descriptors," in *Symp. on Signal Process. and Info. Technol.*, 2004, pp. 334–337.

[3] P. Butler, P. Chakraborty, and N. Ramakrishan, "The Deshredder: A visual analytic approach to reconstructing shredded documents," in *IEEE Conf. on Vis. Analytics Sci. and Technol.* IEEE, 2012, pp. 113–122.

[4] M. Prandtstetter and G. R. Raidl, "Combining forces to reconstruct strip shredded text documents," in *Int. Workshop on Hybrid Metaheuristics*. Springer, 2008, pp. 175–189.

[5] M. Prandtstetter, "Two approaches for computing lower bounds on the reconstruction of strip shredded text documents," TR1860901, Technishe Universitat Wien, Institut fur Computergraphik und Algorithmen, Tech. Rep., 2009.

[6] J. Chen, M. Tian, X. Qi, W. Wang, and Y. Liu, "A solution to reconstruct cross-cut shredded text documents based on constrained seed K-means algorithm and ant colony algorithm," *Expert Syst. with Appl.*, vol. 127, pp. 35–46, 2019.

[7] D. Pomeranz, M. Shemesh, and O. Ben-Shahar, "A fully automated greedy square jigsaw puzzle solver," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, 2011, pp. 9–16.

[8] J. Perl, M. Diem, F. Kleber, and R. Sablatnig, "Strip shredded document reconstruction using optical character recognition," in *Int. Conf. on Imag. for Crime Detection and Prevention*, 2011, pp. 1–6.

[9] T. M. Paixão, M. C. S. Boeres, C. O. A. Freitas, and T. Oliveira-Santos, "Exploring Character Shapes for Unsupervised Reconstruction of Strip-shredded Text Documents," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 7, pp. 1744–1754, 2019.

[10] N. Xing and J. Zhang, "Graphical-character-based shredded chinese document reconstruction," *Multimedia Tools and Appl.*, vol. 76, no. 10, pp. 12 871–12 891, 2017.

[11] L. Perdue, "What the argo movie got wrong about shredded documents," https://lewisperdue.com/archives/4052, April 2013, accessed: June 5, 2023.

[12] T. M. Paixão, R. F. Berriel, M. C. S. Boeres, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "A deep learning-based compatibility score for reconstruction of strip-shredded text documents," in *Conf. on Graph., Patterns and Images*, 2018, pp. 87–94.

[13] T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Self-supervised deep reconstruction of mixed strip-shredded text documents," *Pattern Recognit.*, vol. 107, p. 107535, 2020a.

[14] T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. D. Souza, and T. Oliveira-Santos, "Fast(er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning," in *IEEE/CVF Conf. on Comp. Vision and Pattern Recognit.*, 2020b, pp. 14 343–14 351.

[15] D. Pöhler, R. Zimmermann, B. Widdecke, H. Zoberbier, J. Schneider, B. Nickolay, and J. Krüger, "Content representation and pairwise feature matching method for virtual reconstruction of shredded documents," in *9th IEEE Int. Symp. Image and Signal Process. and Anal.*, 2015, pp. 143–148.

[16] T. M. Paixão, R. F. Berriel, M. C. S. Boeres, A. L. Koerich, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "A human-in-the-loop recommendation-based framework for reconstruction of mechanically shredded documents," *Pattern Recognit. Letters*, vol. 164, pp. 1–8, 2022.

[17] D. Applegate, R. Bixby, V. Chvatal, and W. Cook, "Concorde: A code for solving traveling salesman problems," http://www.math.uwaterloo.ca/tsp/concorde, 2001, accessed on: October 19, 2020.

[18] M. Marques and C. Freitas, "Document decipherment-restoration: Strip-shredded document reconstruction based on color," *IEEE Latin America Trans.*, vol. 11, no. 6, pp. 1359–1365, 2013.

[19] Y. Liang and X. Li, "Reassembling Shredded Document Stripes Using Word-path Metric and Greedy Composition Optimal Matching Solver," *IEEE Trans. on Multimedia*, vol. 22, no. 5, pp. 1168–1181, 2020.

[20] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan, "Active learning in recommender systems," in *Recommender systems handbook*. Springer, 2015, pp. 809–846.