Deep Learning-based Reconstruction of Shredded Documents

Thiago M. Paixão[1,2], **Maria C. S. Boeres** [2], **Thiago Oliveira-Santos**[2]

[1]Instituto Federal do Espírito Santo (IFES)
Av. dos Sabiás, 330, Morada de Laranjeiras, Serra - ES, Brazil

[2]Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, Brazil

thiago.paixao@ifes.edu.br, {boeres,todsantos}@inf.ufes.br

Abstract. *This thesis addresses the reconstruction of shredded paper documents, a relevant task in various domains such as forensic investigation and history reconstruction. Despite previous research, dealing with real-shredded data is a sensitive issue in the literature. To face this challenge, we proposed two deep learning approaches that have achieved state-of-the-art accuracy in more realistic scenarios. As a second major contribution, human interaction was explored to improve reconstruction. Our framework, inspired by the field of active learning, automatically selects potential mistakes in the solution for user analysis enabling better accuracy in a scalable way. The results yielded works in top-tier publications such as CVPR and the Pattern Recognition journal.*

1. Introduction

Historically, shredding is also associated with the destruction of espionage content, as in the Iran hostage crisis [Derian 1989] portrayed in the movie "Argo", or in the case of the documents left behind by the official state security service of former East Germany (*Stasi*) after the fall of the Berlin Wall. Additionally, shredding may be illicitly motivated when the objective is to destroy evidence of fraud and other sorts of crimes. In this context, revealing the original content of shredded papers is of great relevance for forensic investigation, which can be achieved by first joining coherently the shreds (pieces) as in a jigsaw puzzle.

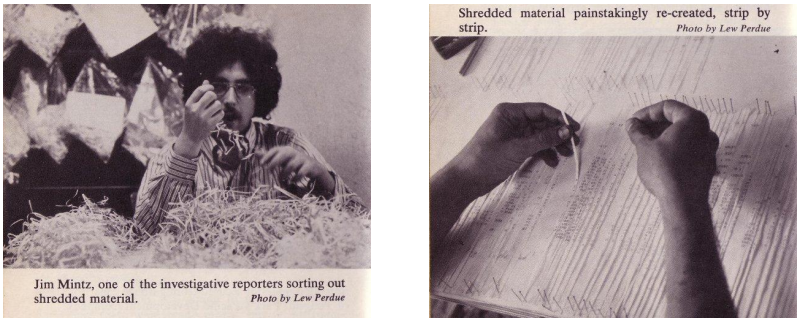

Figure 1. Manual reconstruction (1979 Iran hostage crisis). Credits to Lewis Perdue [Perdue 2013].

Regardless of its importance, manual reconstruction, as illustrated in Figure 1 is potentially damaging to the paper due to the direct contact with the shreds, besides being a slow and tedious task for humans. These factors motivated the development of the digital

and automatic reconstruction process [Ukovich et al. 2004, Butler et al. 2012]. Commercial software, such as "Unshredder"[1], is available to assist individuals and corporations in recovering destroyed documents or to aid in criminal investigations. On the other hand, such technology enables the use of disposed or robbed documents (*e.g.* industrial espionage) for invasion of privacy and illicit use of sensitive data. Therefore, reconstruction technologies might help assess the security level of shredding and disposal services provided by specialized companies.

This work advances the state-of-the-art in the reconstruction of shredded documents in two main directions. The first direction is the development of deep learning (DL) methods for high-accuracy reconstruction in more realistic scenarios. To the best of our knowledge, we were the first to use DL to solve this kind of puzzle. The other is the project of a human-in-the-loop framework that benefits from human interaction to improve solutions obtained with automatic methods. Although focus is on the reconstruction application, the proposed techniques can be extended to other related applications: *e.g.* solving jigsaw puzzles with eroded borders [Paumard et al. 2020, Li et al. 2021, Rika et al. 2022] and reconstruction of ancient papyrus [Abitbol et al. 2021, Pirrone et al. 2021].

Motivation. Traditionally, the reconstruction algorithms perform two main tasks (one at a time): measuring the compatibility of the shreds by using image features and grouping them to maximize the overall compatibility (combinatorial optimization). In digital reconstruction, the shreds are manipulated only during the acquisition step. After this, human participation is restricted to specific interventions (semi-automatic reconstruction [Butler et al. 2012, Pöhler et al. 2015]), or even not required at all (automatic reconstruction). Considering these facts, we identified four main limitations on the related literature that motivated our work: (i) the literature has mostly focused on improving the optimization process relegating to the background the compatibility analysis of the shreds; (ii) most works test only with simulated-shredded documents and (iii) with few test instances (usually ≤ 3 documents), which also yields biased conclusions; and (iv) it lacks studies on multi-page reconstruction.

Contributions. The bulk of this thesis is on evaluating the compatibility of shreds, the major research gap in the literature. Within this direction, these are the main contributions:

- A self-supervised classification-based approach (DEEPREC-CL) for multi-page reconstruction: results have shown that our method is capable of reconstructing 100 documents with accuracy superior to 90%;
- A self-supervised metric learning approach (DEEPREC-ML) that improves the time performance of DEEPREC-CL without losing accuracy: it might yield a speed-up of ≈ 22 times for 505 shreds, and higher for more shreds;

We also proposed a framework for semi-automatic (human-in-the-loop) reconstruction where human feedback is leveraged in a smart way. In this direction, the main contributions are:

[1] `https://www.unshredder.com`.

- A scalable human-in-the-loop (HIL) recommendation-based framework for reconstruction of strip-shredded documents;
- Four query strategies for recommending pairs of shreds to be annotated;
- A novel experimental methodology that assesses the impact of human labor on the quality of the reconstructions: results have shown that a user workload of 25% can lead to more than 4 p.p. of accuracy improvement ($> 40\%$ of error reduction) on the deep learning methods.

Additionally, it is worthy to mention the release of **a new public dataset**[2] with 100 real strip-shredded documents (totaling $2{,}292$ shreds). This addresses the lack of publicly available collections representing real scenarios. The contributions appear in relevant publications, such as the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), the top-tier in *enginering and computer science* and the Pattern Recognition journal (more detailed in the by-products document).

Scope. The scope of our work includes: strip-shredded documents (cuts in the vertical direction), correctly-oriented shreds (the shreds are set upwards, possibly slightly skewed), single-sided shreds (content is restricted to one paper face), black-and-white appearance, and shreds with nearly the same dimensions. These are reasonable assumptions commonly adopted in most works addressing text documents. Nonetheless, it is noteworthy that the cross-cut reconstruction is more complex from the optimization perspective. The works addressing this, however, are not robust for real-shredded data, which is more complex from the image analysis perspective. Thus, the first step for solving cross-cutting is a robust approach to strip-shredding.

2. Related works

This section covers different compatibility evaluation approaches and the topic of semi-automatic reconstruction. We present some representative works and the respective limitations that motivated our work.

2.1. Compatibility evaluation approaches.

Pixel-level fitting. Several works explore distance metrics (*e.g.* Euclidean) at pixel level for compatibility evaluation [Chen et al. 2019, Pomeranz et al. 2011]. They are more sensitive to the corruption on the edges of the shreds.

Shape-based unsupervised fitting. Compatibility is measured by using shapes, such as strokes [Perl et al. 2011] or characters [Paixão et al. 2019] (our previous work) without any learning process or supervised learning. Using character was proven one of the most effective reconstruction approaches, however, it requires segmenting textual information.

Supervised learning-based fitting. In this approach, the matching of the shreds is mainly determined by classification tasks (*e.g.* recognition of symbols). Typical issues are: instability in character recognition (*e.g.* OCR-based matching [Perl et al. 2011])

[2]Available at `https://github.com/thiagopx/deeprec-pr20`.

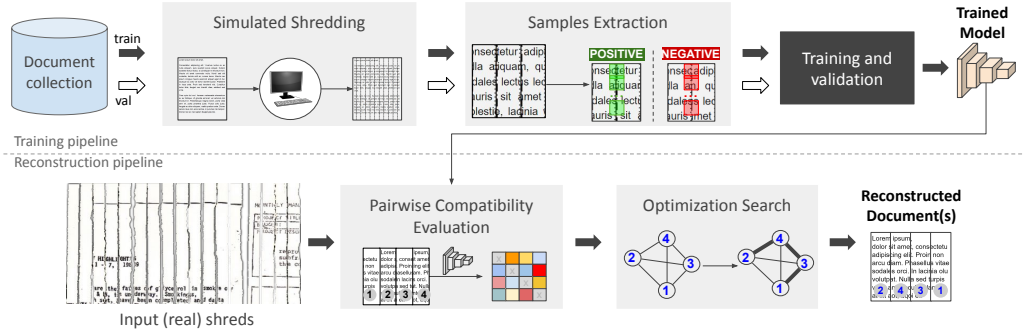

Figure 2. Overview of the classification-based method.

and dependency on specific languages and/or text segmentation [Xing and Zhang 2017]. When it comes to deep learning methods, **we were the first to explore deep models to reconstruct shredded documents** [Paixão et al. 2020a, Paixão et al. 2020b]. Our approach is able to cope with more heterogeneous content because the fitting of patterns is learned in a self-supervised fashion from large-scale data without segmenting symbols, as discussed in the following sections.

2.2. Semi-automatic reconstruction.

Few works address semi-automatic reconstruction. Most of them fit the active paradigm [Butler et al. 2012, Shang et al. 2014], where the user is an inherent part of the reconstruction process. The process is predominantly manual, being the user responsible for tedious activities such as moving shreds, correcting their orientation, and analyzing the adjacency of shreds assisted by GUI tools (*e.g.* zoom-in/out, drag-and-drop). Conversely, in the passive paradigm (ours), user intervention is optional since a preliminary solution can be automatically obtained. User inputs are leveraged to improve an initial/intermediate solution [Prandtstetter and Raidl 2008]. The novelty in our approach consists in recommending which parts of the solution should be analyzed so that simple (binary) feedback (confirm or correct) may yield more accurate solutions.

3. Classification-based reconstruction

The classification-based approach for document reconstruction (Figure 2) comprises two pipelines. The *training pipeline* (top flow) aims to produce a model to quantify the compatibility between shreds based on small samples extracted from around the cutting sections of digitally-cut documents. This local approach mimics the manual reconstruction process, where humans analyze the fitting of shreds based on the local matching of fragmented patterns, primarily at the text line level. Positive samples are obtained from adjacent shreds and negative samples from non-adjacent pairs. The learning process is *self-supervised* since adjacency is automatically inferred in simulated shredding. After sampling, a *convolutional neural network* (CNN) is trained to distinguish between positive and negative samples. In the *reconstruction pipeline* (bottom flow), the trained model is used to evaluate the pairwise compatibility of the scanned shreds (reconstruction instance). The resulting values are used as input for a graph-based optimizer [Applegate et al. 2001] that estimates the permutation of shreds representing the final reconstruction.

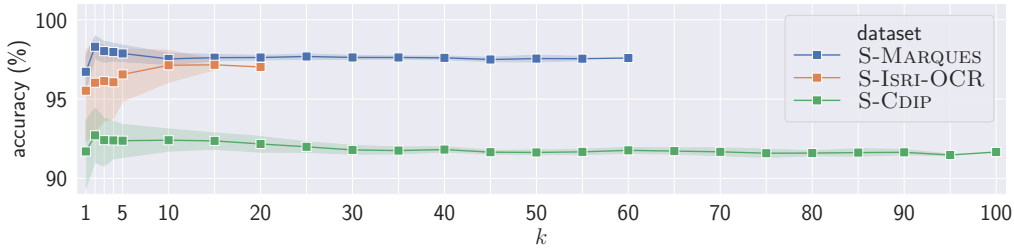

Figure 3. Multi-page reconstruction accuracy.

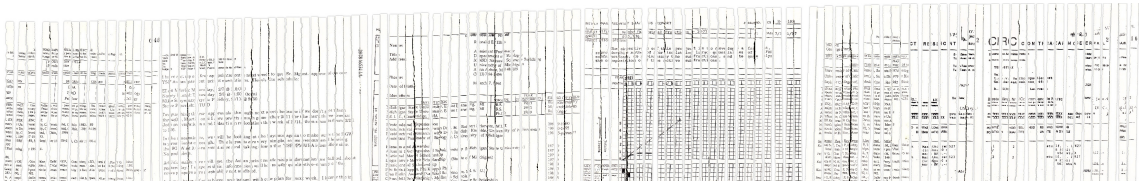

Figure 4. Reconstruction of a S-CDIP instance.

Results on multi-page reconstruction. Three datasets were used for evaluation: S-MARQUES (60 instances), S-ISRI-OCR (20 instances), and S-CDIP (100 instances), being the latter two contributions of our thesis. The evaluation is performed incrementally by mixing k instances (shredded pages), for different values of k. Figure 3 shows the accuracy (mean and 95% confidence interval) given the number of instances. **Overall, the proposed method performed above** 90% for the three datasets. **The accuracy tends to stabilize for large** k, indicating that the insertion of new documents does not degrade accuracy, although it increases the complexity of the problem considerably. S-CDIP, as expected, was verified as the most challenging dataset given the variability of content and layout complexity. Figure 4 shows an example of reconstruction comprising shreds from $k = 5$ documents of S-CDIP (accuracy of 82.93%)[3].

Comparison with state-of-the-art. Our method (DEEPREC-CL) was compared to three relevant methods of literature (referred to by the name of the first author): *Paixão*, our preliminary method based on shape matching; *Liang*, an OCR-based method [Liang and Li 2020]; and *Marques*, which relies on edge pixel dissimilarity [Marques and Freitas 2013]. Due to memory scalability issues, testing with Paixão was limited to only 5 documents. As for Liang, we were able to run experiments only on the S-MARQUES and S-ISRI-OCR datasets limited to 3 documents due to the high OCR overdue. To emphasize the role of compatibility evaluation in producing accurate reconstructions, we also modified our method by coupling the Marques' nearest neighbor optimizer: the modified version was named DEEPREC-CL-NN. As shown in Figure 5, **the average accuracy of the proposed method (DEEPREC-CL) was consistently superior to the compared methods**. Additionally, **it demonstrated greater robustness**, which is mainly evidenced by the stability of the accuracy curve. **DEEPREC-CL-NN greatly outperformed Marques, which has the same optimizer, and also Paixão, which leverages a more powerful optimizer**.

[3]The full reconstruction ($k = 100$) is available at `https://htmlpreview.github.io/?https://github.com/thiagopx/docs/blob/master/results_s-cdip.html`.

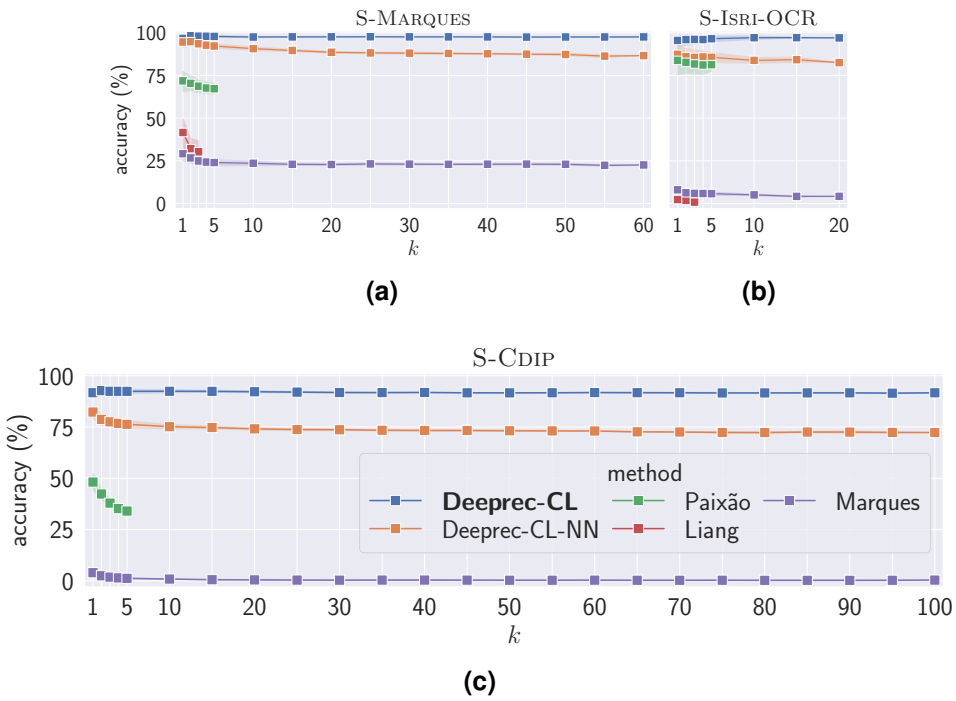

Figure 5. Comparative accuracy performance.

Time efficiency. DEEPREC-CL is also remarkably **more scalable in terms of execution time than Paixão and Liang**, as seen in Figure 6. This is critical in real scenarios since much more than 5 shredded pages are expected as input. Although Marques is very time efficient, as shown in Figure 5, its low accuracy prevents it from being used in real data.

4. Metric learning-based reconstruction

Considering a network inference as the time unit cost, we can say that DEEPREC-CL scales quadratically with the number of shreds. To deal with this, we proposed a metric-learning approach that scales linearly rather than quadratically. As illustrated in Figure 7, the rationale of the new approach is that two side-by-side shreds are globally compatible if they locally fit each other along the touching boundaries. The local approach relies on small samples ($\boldsymbol{x}$) cropped from the boundary regions. Instead of pixel comparison, the samples are first converted to an intermediary representation ($\boldsymbol{e}$) by projecting them onto a common embedding space $\mathbb{R}^d$. This is accomplished by two CNNs: f_{left} and f_{right}, $f_\bullet : \boldsymbol{x} \mapsto \boldsymbol{e}$, specialized on the left and right boundaries, respectively.

Assuming that these models are properly trained, boundary samples (orange and blue regions in Figure 7) are then projected, so that embeddings generated from compatible regions (mostly found on positive pairings) are expected to be closer in this metric space, whereas those from non-fitting regions should be farther apart. Therefore, the global compatibility of a pair of shreds is measured in function of the distances between corresponding embeddings. The interesting property of this approach is that the projection step (network inference) can be decoupled from the distance computation, thus, each shred is processed once by each model, and pairwise evaluation relies on the produced

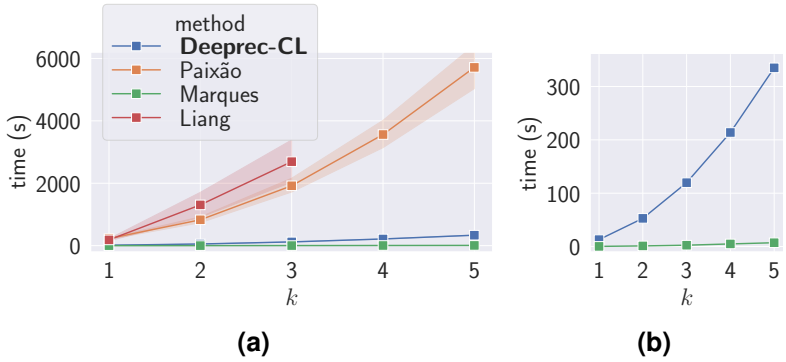

Figure 6. Comparative time performance.

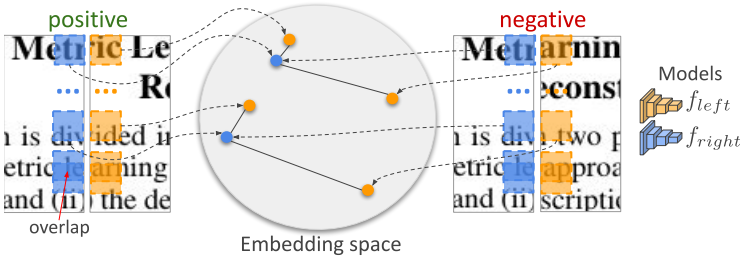

Figure 7. Metric learning approach for shreds' compatibility evaluation.

embeddings.

Comparison with DEEPREC-CL. DEEPREC-CL and DEEPREC-ML (the metric learning approach) were compared in terms of accuracy and time efficiency. We grouped the 1,370 shreds of S-MARQUES as the first instance, and the 505 shreds of S-ISRI-OCR as the second. DEEPREC-ML achieved 94.81 and 97.22% of accuracy for S-MARQUES and S-ISRI-OCR, respectively, whereas DEEPREC-CL achieved 97.08 and 95.24%. **Overall, both methods yielded high-quality reconstructions with a low difference in accuracy (approx. ± 2 p.p.), which is an indication that there is no significant difference in accuracy**. Concerning time efficiency, the methods behave notably differently, as evidenced in Figure 8. The left chart shows the average elapsed time of each stage to process the 505 shreds of S-ISRI-OCR: projection (pro) – applicable only for DEEPREC-ML–, pairwise compatibility evaluation (pw), and optimization process (opt). In this case, the optimization cost was negligible when compared to the execution time for pairwise evaluation. Remarkably, **DEEPREC-CL demanded more than 80 minutes to complete the evaluation stage, whereas our method took less than 4 minutes (speed-up of $\approx$ 22 times). Comparatively, the estimated growth for DEEPREC-ML (blue curve) is significantly slower than the competing method (right chart)**.

5. A human-in-the-loop reconstruction framework

Full automatic reconstruction usually leads to imperfect reconstructions. A particular way to improve solutions is to introduce active human supervision (semi-automatic reconstruction). Inspired by the active learning literature [Rubens et al. 2015], the reconstruction process can be modeled as a loop where, in each iteration, the human is queried to pro-

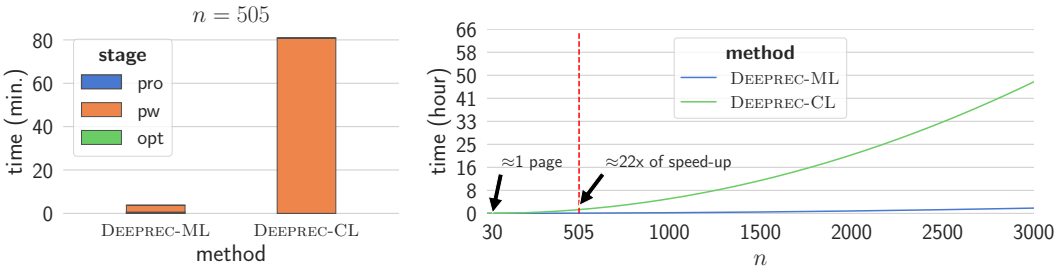

Figure 8. Time performance for multi-page reconstruction.

vide inputs, and a new solution is attained. Ideally, the human effort (workload) should be minimal. In view of this, our framework (Figure 9) includes a *recommender module* which detects potential mistakes for human analysis. The human is responsible for setting apart the negative (wrong) pairings and confirming the positives.

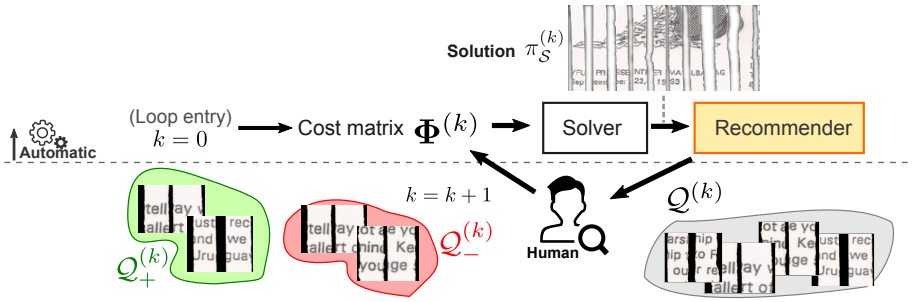

Figure 9. Overview of the proposed HIL reconstruction framework.

The impact of the workload on the accuracy. Quantitatively, workload (α_{load}) means the fraction of pairs of adjacent shreds in a solution to be analyzed. We have previously verified that the workload has a greater impact on the accuracy than the number of iterations, thus, the iterations was set to 1. Figures 10 and 11 show that the accuracy increases roughly linearly with the human workload. **The proposed query strategies (OPT-R, OPT-RL, UNC-R, and UNC-RL) outperform significantly the random-based selection (baseline)** [Prandtstetter and Raidl 2008]. Remarkably, **OPT-R was able to increase the original solution accuracy of the S-CDIP dataset on** ≈ 3.80 **p.p. for** $\alpha_{load} = 0.25$**:** 87 **pairs were corrected from a total of** 220 **mistakes** (≈ 39.50**% of error reduction**).

6. Concluding remarks and future work

This thesis presented a corpus of contributions for (semi-)automatic reconstruction of mechanically-shredded documents. Our effort was initially towards robust compatibility evaluation between shreds for fully automatic reconstruction, so that the optimization process might yield improved reconstructions. In a second moment, it was investigated the introduction of the human as part of the reconstruction process. Future work includes the extension of proposed methodologies to cross-cut documents. Concerning human-assisted reconstruction, a promising direction is the development/adaptation of query strategies to the reconstruction application, including the use of ensemble of strategies for more valuable user feedback. Finally, from a generalization perspective, there are correlated problems that should benefit from our findings as mentioned in the introduction.

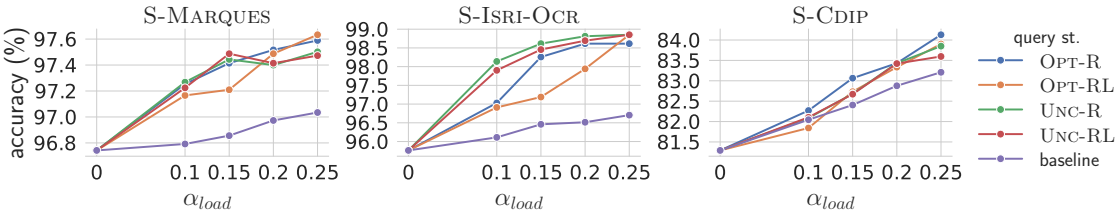

Figure 10. Reconstruction accuracy w.r.t. workload (DEEPREC-CL).

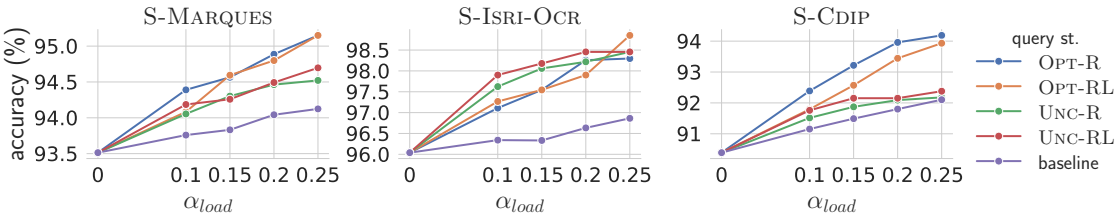

Figure 11. Reconstruction accuracy w.r.t. workload (DEEPREC-ML).

7. Acknowledgments

The authors would like to acknowledge the support of FAPES and CAPES (process 2021-2S6CD, FAPES no. 132/2021) through the PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

References

[Abitbol et al. 2021] Abitbol, R., Shimshoni, I., and Ben-Dov, J. (2021). Machine learning based assembly of fragments of ancient papyrus. *Journal on Comput.and Cultural Heritage*, 14(3):1–21.

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

[Butler et al. 2012] Butler, P., Chakraborty, P., and Ramakrishan, N. (2012). The Deshredder: A visual analytic approach to reconstructing shredded documents. In *IEEE Conf. on Vis. Analytics Sci. and Technol.*, pages 113–122. IEEE.

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

[Derian 1989] Derian, J. D. (1989). Arms, Hostages, and the Importance of Shredding in Earnest: Reading the National Security Culture (II). *Social Text*, (22):79–91.

[Li et al. 2021] Li, R., Liu, S., Wang, G., Liu, G., and Zeng, B. (2021). Jigsawgan: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *IEEE Trans. on Image Processing*, 31:513–524.

[Liang and Li 2020] Liang, Y. and Li, X. (2020). Reassembling Shredded Document Stripes Using Word-path Metric and Greedy Composition Optimal Matching Solver. *IEEE Trans. on Multimedia*, 22(5):1168–1181.

[Marques and Freitas 2013] Marques, M. and Freitas, C. (2013). Document decipherment-restoration: Strip-shredded document reconstruction based on color. *IEEE Latin America Trans.*, 11(6):1359–1365.

[Paixão et al. 2020b] Paixão, T. M., Berriel, R. F., Boeres, M. C. S., Koerich, A. L., Badue, C., Souza, A. F. D., and Oliveira-Santos, T. (2020b). Fast(er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning. In *IEEE/CVF Conf. on Comp. Vision and Pattern Recognit.*, pages 14343–14351.

[Paixão et al. 2019] Paixão, T. M., Boeres, M. C. S., Freitas, C. O. A., and Oliveira-Santos, T. (2019). Exploring Character Shapes for Unsupervised Reconstruction of Strip-shredded Text Documents. *IEEE Trans. Inf. Forensics Secur.*, 14(7):1744–1754.

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

[Paumard et al. 2020] Paumard, M.-M., Picard, D., and Tabia, H. (2020). Deepzzle: Solving visual jigsaw puzzles with deep learning and shortest path optimization. *IEEE Trans. on Image Processing*, 29:3569–3581.

[Perdue 2013] Perdue, L. (2013). What the argo movie got wrong about shredded documents. `https://lewisperdue.com/archives/4052`. Accessed: June 5, 2023.

[Perl et al. 2011] Perl, J., Diem, M., Kleber, F., and Sablatnig, R. (2011). Strip shredded document reconstruction using optical character recognition. In *Int. Conf. on Imag. for Crime Detection and Prevention*, pages 1–6.

[Pirrone et al. 2021] Pirrone, A., Beurton-Aimar, M., and Journet, N. (2021). Self-supervised deep metric learning for ancient papyrus fragments retrieval. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):219–234.

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

[Pomeranz et al. 2011] Pomeranz, D., Shemesh, M., and Ben-Shahar, O. (2011). A fully automated greedy square jigsaw puzzle solver. In *IEEE Conf. Comput. Vision and Pattern Recognit.*, pages 9–16.

[Prandtstetter and Raidl 2008] Prandtstetter, M. and Raidl, G. R. (2008). Combining forces to reconstruct strip shredded text documents. In *Int. Workshop on Hybrid Metaheuristics*, pages 175–189. Springer.

[Rika et al. 2022] Rika, D., Sholomon, D., David, E., and Netanyahu, N. S. (2022). Ten: Twin embedding networks for the jigsaw puzzle problem with eroded boundaries. *arXiv preprint arXiv:2203.06488*.

[Rubens et al. 2015] Rubens, N., Elahi, M., Sugiyama, M., and Kaplan, D. (2015). Active learning in recommender systems. In *Recommender systems handbook*, pages 809–846. Springer.

[Shang et al. 2014] Shang, S., Sencar, H. T., Memon, N., and Kong, X. (2014). A semi-automatic deshredding method based on curve matching. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5537–5541. IEEE.

[Ukovich et al. 2004] Ukovich, A., Ramponi, G., Doulaverakis, H., Kompatsiaris, Y., and Strintzis, M. (2004). Shredded document reconstruction using MPEG-7 standard descriptors. In *Symp. on Signal Process. and Info. Technol.*, pages 334–337.

[Xing and Zhang 2017] Xing, N. and Zhang, J. (2017). Graphical-character-based shredded chinese document reconstruction. *Multimedia Tools and Appl.*, 76(10):12871–12891.