Digitization of Handwritten Ledgers: An Integrated Approach to Improve Handwritten Text Recognition

Submitted on: **March 6, 2024**

Roderick Majoor
roderick.majoor@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Nanne van Noord
n.j.e.vannoord@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

1 INTRODUCTION

The Bank of Amsterdam was an early bank established in 1609, described by some as the first central bank [10]. The bank played an important role in the financial world of the 17th and 18th century. Transactions made at the bank from 1650 to 1800 were recorded in ledgers and are still preserved to this day. These ledgers contain information about transactions, both debit and credit, of customers of the bank. The ledgers can thus be very important in analyzing the money flow at the time. Digitizing these ledgers may help in preserving them in a much easier to analyze format which would allow for more research into the contents of the ledgers. However, these ledgers are all handwritten making it hard to retrieve the information out of them on large scale. Current efforts to retrieve the information are done by using Handwritten Text Recognition (HTR) tools to extract the text in the ledgers. The problem is that the currently used HTR techniques have too many inaccuracies to be used effectively. Because of this, the handwritten text cannot be recognized correctly and can thus not be digitized properly. To improve digitization efforts, it is crucial to find out why some handwritten text cannot be recognized, what kind of errors are made during the HTR process and how the HTR process can be optimized to improve these inaccuracies. The research question we wish to answer is:

How can the categorization of errors in handwritten ledger analysis be used to enhance the identification and resolution of text recognition inaccuracies?

Subquestions belonging to this research question are:

- What factors contribute to text recognition errors in handwritten ledgers?
- What are the different types of errors encountered in handwritten text recognition?
- How can specific preprocessing methods improve text detection and segmentation in handwritten ledgers?
- What strategies can be implemented to mitigate common text recognition errors in handwritten ledgers?

2 RELATED WORK

The research gap being addressed in this project revolves around the challenges associated with the digitization of handwritten ledgers. The proposed research will contribute in closing the research gap within the HTR domain, specifically looking at historical handwritten documents.

Earlier research into the domain of digitizing historical handwritten documents show a few steps that should typically be followed to get to the desired result [5]. These steps are typically outlined as:

(1) Pre-processing
(2) Document layout analysis (DLA)
(3) Text recognition
(4) Post-processing

Studies have shown that pre-processing images can lead to better results when using them as input for document layout analysis or OCR. Thus, we want to optimize our image material such that the document layout analysis can be performed more accurately. There are different methods for doing this such as adjusting the contrast to remove artifacts and noise or fixing the image alignment [1, 3, 5]. It is crucial to find out which pre-processing steps might be useful to our problem at hand.

To perform DLA on our dataset, we essentially perform table detection and recognition since our complete page includes a tabular structure.

Correia & Luck show that enhancing optical character recognition (OCR) engines with pre- and post-processing methods to digitize large-scale historical micro-data can be effective [3]. However, their methods are not suitable for handwritten documents.

Liang et al. show that existing OCR systems like Breuel's OCRopus method struggle to effectively analyze the layout of scanned historical handwritten documents with a tabular layout. While traditional image processing techniques can partially extract table layouts, machine learning classification yields the best results. Transfer learning and deep convolutional networks demonstrate high accuracy [7].

Prieto et al. address the challenge of document image understanding in documents with complex layouts like tables. They compare two machine learning based approaches and show promising results [9]. Their approach however, is tailored to printed tables with handwritten content, and requires the need for straight lines of the tables. Our approach should go a step further and be able to recognize the tabular structure merely from handwritten contents.

Oliveira et al. propose dhSegment, an open-source implementation of a CNN-based pixel-wise predictor for document segmentation. The approach aims to address various document processing tasks simultaneously, including page extraction, baseline extraction, layout analysis, and illustration and photograph extraction. They show that a single CNN architecture can be used across tasks with competitive results [2].

Multiple studies describe the use of Transformer based methods to perform layout analysis and optical character recognition (OCR) [6, 8, 13]. Furthermore, Smock et al. show that transformer-based object detection models can produce excellent results for the tasks of detection, structure recognition, and functional analysis of tables [11]. While these approaches show promising results, they are not tested for our specific task at hand.

Droby et al. propose a holistic method that applies Mask R-CNN for text line extraction in historical documents. Their work achieved state-of-the-art results on well known historical datasets [4]. While their work does not directly transfer to our task, it does show the potential for document segmentation and possibly table recognition abilities of Mask R-CNN.

3 METHODOLOGY

This part describes the dataset, methods used and the evaluation metrics used to assess the performance of the models.

3.1 Data

The dataset used in this study consists of pages from the collection of ledgers from the Bank of Amsterdam. The full collection can be found on the website of The Amsterdam City Archives. In the series of ledgers, the current accounts of the customers were maintained. Each customer had one or more pages, with smaller traders having a portion of a page. An alphabetical list of the traders and the page of their current account can be found in the index. The ledgers kept track of the amounts that were debited and credited from and to a customer. The ledger pages consist of several columns containing information such as the date, account holder name, account number and amount debited/credited. Figure 1 shows an example of what a page looks like. The corresponding annotated page with ground truth values is shown in Figure 2. The ground truth values better show the tabular structure of the pages.

3.2 Method

The first step in our methodology is to identify what current problems exist when performing HTR on the ledger pages. The currently used HTR tool is loghi, which is trained specifically on old handwriting including the ones used in our ledger documents. Using this tool and our ground truth data, we can see where mistakes are currently made in the documents.

Once we have discovered where the mistakes are made on the page, we can further analyse them to find out what the errors are and why they are made. Subsequently, a systematic error categorization framework can be made. This framework would consider common types of errors encountered in handwritten text recognition tasks, such as misinterpretation of characters, incorrect line segmentation, and misclassification of symbols.

Once we have identified the various inaccuracies within the HTR process, we can start to look at possible solutions for each of the different error types. Here, we will test different methods such as novel pre-processing steps and document layout parsing to see whether they yield improvements in the outcome of the HTR quality. Pre-processing steps could include techniques such as contrast adjustment, noise reduction, image normalization, image binarization, skew correction and image scaling. These techniques can be useful to improve character recognition. Since our ledger pages contain a tabular structure, we expect to see errors in recognizing the correct order of words in the HTR. We will employ models like the layout-parser tool to test whether it is possible to recognize the tabular structure of the ledger successfully such that we can get better results with respect to the word order in the HTR process.

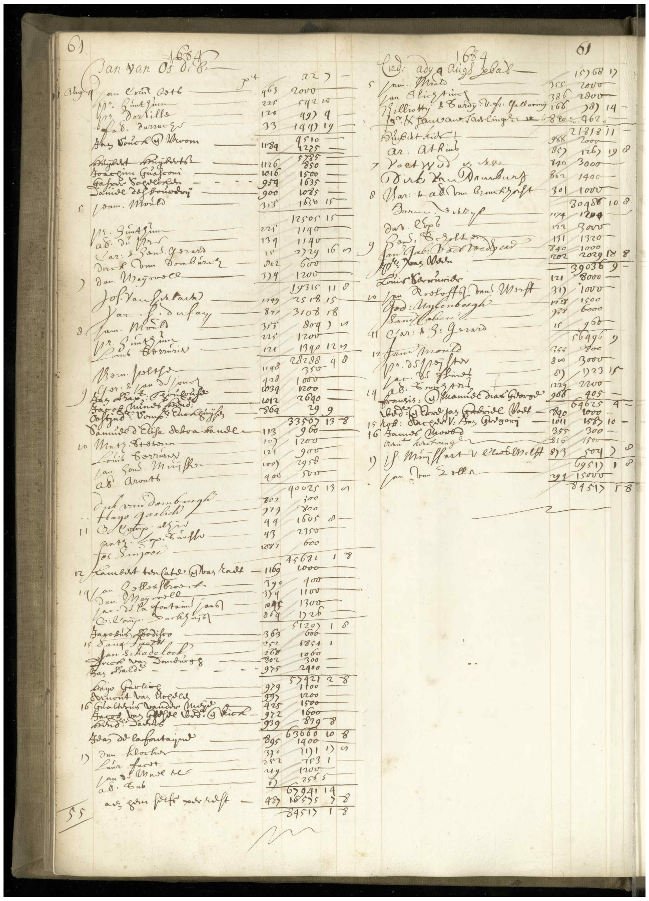


Figure 1: An example page from the ledger collection.

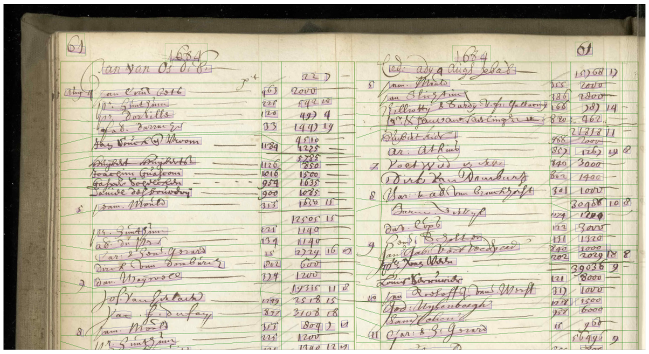


Figure 2: A part of the annotated ledger page.

3.3 Evaluation

To evaluate the layout parser tool, we would need to compare the model output bounding boxes to the ground truth bounding box values. For this we use intersection over union (IoU), precision, recall and f1-score. The IoU score tells us more about what predicted bounding box has most overlap with a ground truth bounding box. Once every predicted box is assigned to a ground truth box, we can use the other scores to measure the performance. The formulas for these scores are shown below.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

$$Precision = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{2}$$

$$Recall = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{3}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

We can measure the performance of the HTR process on our pages, using the word error rate (WER) and the character error rate (CER) [12].

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{5}$$

Where S_w, D_w, I_w are the number of word substitutions, deletions and insertions respectively and N_w is the total number of words in the ground truth.

$$CER = \frac{S_c + D_c + I_c}{N_c} \tag{6}$$

Where S_c, D_c, I_c are the number of character substitutions, deletions and insertions respectively and N_c is the total number of characters in the ground truth.

We can then look at the WER and CER of the currently used HTR process and compare it to the scores for our proposed method to see whether our method yields an improvement.

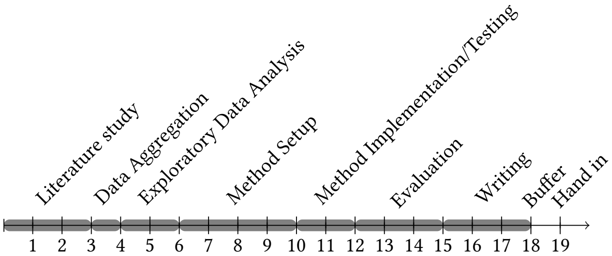
4 RISK ASSESSMENT

Potential risks may reveal themselves during this project. In this section, we will list some of the potential risks and the possible solutions to mitigate them.

A potential risk could be a lack of high quality labeled data on time. While the ledger pages are publicly available, there is no current dataset containing the annotated pages with ground truth values. If the data is not ready on time, a different dataset with similar features should be used to still test whether our models could be effective on the ledger documents. Datasets shown on PRimA can be useful in that case. Some of these contain historic documents with tabular structures. We could try our models using this data and see whether they can effectively capture the correct structure from these documents.

5 PROJECT PLAN

This thesis will span thirteen weeks from the 1st of April until 30th of June full-time. Additionaly, part-time research will be done in roughly the 6 weeks prior to the 1st of April, starting with this thesis design. A rough outline for this project is shown below.



REFERENCES

[1] Yasser Alginahi. 2010. *Preprocessing Techniques in Character Recognition.* https://doi.org/10.5772/9776

[2] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhSegment: A Generic Deep-Learning Approach for Document Segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE. https://doi.org/10.1109/icfhr-2018.2018.00011

[3] Sergio Correia and Stephan Luck. 2023. Digitizing historical balance sheet data: A practitioner's guide. *Explorations in Economic History* 87 (2023), 101475. https://doi.org/10.1016/j.eeh.2022.101475 Methodological Advances in the Extraction and Analysis of Historical Data.

[4] Ahmad Droby, Berat Kurar Barakat, Reem Alaasam, Boraq Madi, Irina Rabaev, and Jihad El-Sana. 2022. Text Line Extraction in Historical Documents Using Mask R-CNN. *Signals* 3, 3 (2022), 535–549. https://doi.org/10.3390/signals3030032

[5] Constantin Lehenmeier, Manuel Burghardt, and Bernadette Mischka. 2020. *Layout Detection and Table Recognition – Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data.* Springer International Publishing, 229–242. https://doi.org/10.1007/978-3-030-54956-5_17

[6] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv:2109.10282 [cs.CL]

[7] Xusheng Liang, Abbas Cheddad, and Johan Hall. 2021. Comparative Study of Layout Analysis of Tabulated Historical Documents. *Big Data Research* 24 (2021), 100195. https://doi.org/10.1016/j.bdr.2021.100195

[8] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2023. Kosmos-2.5: A Multimodal Literate Model. arXiv:2309.11419 [cs.CL]

[9] Jose Ramón Prieto, José Andrés, Emilio Granell, Joan Andreu Sánchez, and Enrique Vidal. 2023. Information extraction in handwritten historical logbooks. *Pattern Recognition Letters* 172 (2023), 128–136. https://doi.org/10.1016/j.patrec.2023.06.008

[10] Stephen Quinn and William Roberds. 2009. *An economic explanation of the early Bank of Amsterdam, debasement, bills of exchange and the emergence of the first central bank.* Cambridge University Press, 32–70.

[11] Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4634–4642.

[12] Enrique Vidal, Alejandro H. Toselli, Antonio Ríos-Vila, and Jorge Calvo-Zaragoza. 2023. End-to-End page-Level assessment of handwritten text recognition. *Pattern Recognition* 142 (2023), 109695. https://doi.org/10.1016/j.patcog.2023.109695

[13] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. arXiv:2104.08836 [cs.CL]