



# When Tesseract Meets PERO

## Open-Source Optical Character Recognition of Medieval Texts

Vít Novotný  and Aleš Horák 

Faculty of Informatics, Masaryk University  
Botanická 68a, 602 00 Brno, Czech Republic  
witiko@mail.muni.cz, haless@fi.muni.cz

**Abstract.** Conversion of scanned images to the text form, denoted as optical character recognition or OCR, for contemporary printed texts is widely considered a solved problem. However, the optical character recognition of early printed books and reprints of medieval texts remains an open challenge.

In our previous work, we developed an end-to-end image-to-text pipeline (via optical character recognition) for medieval texts, named AHISTO OCR, and we released it together with our test dataset under open licenses. However, the published system relied on the closed-source Google Vision AI service as one component, which made the experiments less reproducible. In this work, we replace Google Vision AI with an open-source OCR algorithm named PERO and we show that this not only makes the AHISTO OCR pipeline open, but also improves the performance of the system. We release the updated AHISTO OCR system and its test results again under open licenses.

**Keywords:** optical character recognition, OCR, medieval texts, AHISTO project

## 1 Introduction

In recent decades, public access to historical artifacts and documents has improved greatly via generally accessible photo banks and scanned sources such as the database of the Czech medieval sources online [7] provided by the Centre for Medieval Studies of the Czech Academy of Sciences. The utility of published documents further increases in case the data are not provided just in the image form but also with the recognized texts accessible to further content analysis via human search and automated natural language processing techniques.

The aim of the AHISTO project [1] is to make documents from the Hussite era (1419–1436) available to the general public through a web-hosted searchable portal and database. Although scanned images of modern letterpress reprints from the 19th and 20th century are available, accurate optical character recognition (OCR) algorithms are required to extract searchable text from the scanned images.

In our previous work [8,10], we have developed the AHISTO OCR pipeline for the image-to-text conversion of medieval texts in multilingual settings. We

have shown that the open-source Tesseract 4 OCR algorithm [14] was the second fastest and the most accurate among five different algorithms. [8] We have also shown that we can further improve the performance of the AHISTO OCR system by using Tesseract 4 for one-column pages and the closed-source Google Vision AI service [2] for two-column pages. [10] However, using a closed-source OCR service made our results difficult to investigate and reproduce.

In the current article, we present a new version of the AHISTO OCR pipeline where the Google Vision AI component is replaced with the open-source PERO OCR system [13]. We show that this not only makes the pipeline completely open, but the change also improves the performance of AHISTO OCR.

In Section 2, we introduce PERO OCR. In Section 3, we describe the experiments we have conducted and the dataset and measures that we used in our evaluation. In Section 4, we report the results of the evaluation. In Section 5, we summarize our contribution and outline the ideas for future work in the OCR of medieval texts.

## 2 Related Work

At ICDAR 2021, Michal Hradiš and his colleagues from the Brno University of Technology have presented different aspects of the new PERO OCR system [13,12]. Kodym and Hradiš [5] have introduced a page layout analysis algorithm for early modern and modern hand-written texts. Their algorithm achieved results comparable to state-of-the-art algorithms on a baseline detection task.

Kišš, Beneš, and Hradiš [4] have presented self-training and masked augmentation techniques for OCR algorithms. Their techniques led to significant improvements in performance on the optical character recognition of early modern and modern hand-written and printed texts.

Kohút and Hradiš [6] have showcased an adaptive instance normalization technique that can reconcile different transcription styles in OCR datasets produced by different annotators. Their technique makes it possible to use heterogeneous datasets to train OCR algorithms.

The PERO OCR system uses the above mentioned techniques for the optical character recognition of early modern and modern texts. The system is available as a web demo [13] and also as an open-source code at GitHub [12] with pre-trained models [3].

## 3 Methods

In our current work, we replace the Google Vision AI component of the AHISTO OCR system with two variants of the PERO OCR system: the web demo, with cloud-like service hosted at the Brno University of Technology, and the open-source code at GitHub with pre-trained models prepared for deployment at an independent server. In our previous work, we used the Google Vision AI model from October 2, 2020. To provide a fair comparison, we also report results with

a more recent model from August 11, 2022. As a baseline, we also report results for Google Vision AI and PERO OCR alone.

To evaluate the performance of the new AHISTO OCR version, we use the word error rate (WER) on the 120 human-annotated pages from the AHISTO dataset [11]. As in our previous work, we lower-cased and deaccented the texts in the dataset and in the predictions of our system to simulate a full-text search use case.

## 4 Results

In Table 1, we show that replacing Google Vision AI with PERO OCR improved the AHISTO OCR WER by 1.06%, to a very acceptable **2.08%**, even when the more recent Google Vision AI model from 2022 is used in the comparison.

The main complication in direct application of the two tested OCR systems was caused by the special format of two-column pages that appear in the scanned document collection with a non-negligible frequency. Whereas Google Vision AI achieved an extremely high WER of 78.35% on two-column pages with the earlier model from 2020, the more recent model is much better in analysing these page formats reaching an acceptable WER of 10.52%. To provide a fair comparison of PERO and Google Vision, we use AHISTO OCR with the 2022 Google Vision AI model in the evaluation.

The two variants of PERO OCR achieved different WER. This shows that the web demo of PERO OCR is not necessarily equivalent to the open-source code from GitHub with the published pre-trained models. To honor our intention to keep the AHISTO OCR system open-source, we employ the GitHub version of PERO OCR in the published AHISTO OCR pipeline.

Compared to the more recent Google Vision AI model alone, AHISTO OCR with PERO OCR improved WER by 1.54%. Compared to the web demo of PERO OCR alone, AHISTO OCR with PERO OCR improved WER by 3.97%.

Table 1: Word error rates (%) of Google Vision AI, PERO OCR, and AHISTO OCR evaluated on the AHISTO dataset [11]. For AHISTO OCR with Google Vision AI (the second column from the right), we report results with the more recent model from 2022-08-11. For AHISTO OCR with PERO OCR (the rightmost column), we report results with the open-source variant of PERO OCR available at GitHub and using the published pre-trained models. Best results in each row are **bold**.

	Google Vision AI		PERO OCR		AHISTO OCR	
	2020-10-02	2022-08-11	Demo	GitHub	with Google	with PERO
<b>One column</b> (103)	4.88%	3.79%	2.83%	<b>2.08%</b>	3.79%	<b>2.08%</b>
<b>Two columns</b> (17)	78.35%	10.52%	31.51%	49.38%	<b>7.43%</b>	9.93%
<b>All pages</b> (120)	16.23%	4.83%	7.26%	9.39%	4.35%	<b>3.29%</b>

## 5 Conclusion

In this work, we have shown that we can replace the closed-source Google Vision AI service with the open-source PERO OCR system to make the results of the image-to-text AHISTO OCR pipeline reproducible and also to improve the overall performance of the OCR system. We release the updated AHISTO OCR system [9] and its test results [11] under open licenses.

**Acknowledgements** This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

## References

1. Elbel, P., Novotný, R., Horák, A.: Accessible historical sources. Making medieval written documents available in the form of a contextual database, AHISTO, <https://nlp.fi.muni.cz/projects/ahisto>
2. Google: Vision ai, <https://cloud.google.com/vision>
3. Hradiš, M.: PERO EU-CZ print newspapers. Department of Computer Graphics and Multimedia, Faculty of Information Technology, Brno University of Technology, [https://www.fit.vut.cz/~ihradis/pero/pero\\_eu\\_cz\\_print\\_newspapers\\_2020-10-09.tar.gz](https://www.fit.vut.cz/~ihradis/pero/pero_eu_cz_print_newspapers_2020-10-09.tar.gz), [cit. 2022-08-09]
4. Kišš, M., Beneš, K., Hradiš, M.: AT-ST: Self-training adaptation strategy for OCR in domains with limited transcriptions. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR. pp. 463–477. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86337-1\\_31](https://doi.org/10.1007/978-3-030-86337-1_31)
5. Kodým, O., Hradiš, M.: Page layout analysis system for unconstrained historic documents. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR. pp. 492–506. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86331-9\\_32](https://doi.org/10.1007/978-3-030-86331-9_32)
6. Kohút, J., Hradiš, M.: TS-Net: OCR trained to switch between text transcription styles. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR. pp. 478–493. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86337-1\\_32](https://doi.org/10.1007/978-3-030-86337-1_32)
7. Novotný, R.: Czech medieval sources online. Centre for Medieval Studies, Institute of Philosophy, CAS CR, <https://sources.cms.flu.cas.cz>
8. Novotný, V.: When Tesseract does it alone: Optical character recognition of medieval texts. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) RASLAN 2020. pp. 3–12 (2020)
9. Novotný, V.: ocr: An OCR engine for the AHISTO project. Faculty of Informatics, Masaryk University (Sep 2022), <https://gitlab.fi.muni.cz/nlp/ahisto-modules/ocr>
10. Novotný, V., Seidlová, K., Vrabcová, T., Horák, A.: When Tesseract brings friends: Optical character recognition of medieval texts. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) RASLAN 2021. pp. 29–39 (2021)
11. Novotný, V., Seidlová, K., Vrabcová, T., Horák, A.: A human-annotated dataset of scanned images and OCR texts from medieval documents. Faculty of Informatics, Masaryk University (2022), <https://nlp.fi.muni.cz/projects/ahisto/ocr-dataset>, [cit. 2022-12-09]
12. PERO contributors: pero-ocr, <https://github.com/DCGM/pero-ocr>, commit 8908759
13. PERO contributors: PERO OCR demonstration application. Department of Computer Graphics and Multimedia, Faculty of Information Technology, Brno University of Technology (2022), <https://pero-ocr.fit.vutbr.cz/>, [cit. 2022-06-22]

14. Smith, R.: Tesseract blends old and new OCR technology. Tutorial at DAS (2016)