Evaluating the performance and usability of a Tesseract-based OCR workflow on French-Dutch bilingual historical sources

DH Benelux 2022 | short form paper

Alec Van den broeck[1,2,3]
0000-0002-3593-7851
@alecvdbroeck

Tess Dejaeghere[1]
0000-0002-8472-4457
@DejaeghereTess

Vincent Ducatteeuw[1,4]
0000-0003-4493-6268
@VDucatteeuw

Lise Foket[1]
0000-0003-3983-0469
@LFoket

Julie M. Birkholz[1,5]
0000-0003-1193-0847
@juliebirkholz

Sally Chambers[1,5]
0000-0002-2430-475X
@schambers3

Christophe[1] Verbruggen
0000-0003-0849-6365
@cvbrugg

Frederic Lamsens[1]

Julie Landuyt[6]
0000-0002-3728-8883

[1] Ghent Centre for Digital Humanities (GhentCDH), Ghent University
[2] Internet Technology and Data Science Lab (IDLab), Ghent University
[3] Vlaamse Kunstcollectie (VKC)
[4] Antwerp Cultural Heritage Sciences (ARCHES), University of Antwerp
[5] KBR - Royal Library of Belgium
[6] Master's student Art History, Ghent University

Introduction

The study of texts using a qualitative approach remains the dominant modus operandi in humanities research *(D. Nguyen et al., 2020)*. While most humanities researchers emphasize the critical examination of texts, digital research methodologies are gradually being adopted as complementary options *(Levenberg et al., 2018)*. These computational practices allow researchers to process, aggregate and analyze large quantities of texts. Analytical techniques can help humanities scholars uncover principles and patterns that were previously hidden or identify salient sources for further qualitative research *(Bod, 2013; Aiello & Simeone, 2019)*. However, to support these and more advanced use cases such as Natural Language Processing (NLP), sources must be digitized and transformed into a machine-readable format through Optical Character Recognition (OCR) *(Lopresti, 2009)*.

Despite the fact that OCR software is frequently used to convert analogue sources into digital texts, off-the-shelf OCR tools are usually less adapted to historical sources leading to errors in text transcription *(Martínek et al., 2020; Nguyen et al., 2021; Smith & Cordell, 2018)*. Another disadvantage to these models is that they are very susceptible to noise, resulting in relatively low text detection accuracy. Methods of digital text analysis have the potential to further expand the field of humanities *(Blevins & Robichaud, 2011; Kuhn, 2019; Nguyen et al., 2021)*. However, as OCR quality has a profound impact on these methods, it is important that OCR-generated text is as accurate as possible to avoid bias *(Traub et al., 2015; Strien et al., 2020)*. Adapting OCR systems to distinct historical sources is not only expensive and time-consuming, but the technical knowledge required to (re)train OCR models is often perceived as a hurdle by humanists *(Nguyen et al., 2021; Smith & Cordell, 2018)*. Consequently, research efforts are often geared towards improving the output of the off-the-shelf OCR tools through a process of error analysis and post-correction *(Nguyen et al., 2019)*. These efforts have resulted in streamlined, domain-specific OCR workflows including OCR4all, Escriptorium and OCR-D *(Reul et al., 2019; Kiessling et al., 2019; Neudecker et al., 2019)*. Despite these efforts, there are limited OCR workflows for non-English and multilingual texts *(Strien et al., 2020; Reynaert et al., 2020)*.

In this short paper we present our OCR workflow approach that proposes a user-friendly solution for bilingual historical texts. We test this on a corpus of art exhibition catalogs from INSERT EXACT PERIOD. These texts from the 19th and 20th century, a time period marked by a major expansion of the printed word, a context that makes OCR highly meaningful as manually processing these texts would be very laborious *(Taunton, 2014)*. This is a corpus of catalogs that record works present at specific exhibitions, the so-called *salontentoonstellingen*, which were held from 1792 to 1914 in Antwerp, Ghent and Brussels. The catalogs are bilingual - French and Dutch - printed texts.

Approach

This approach aims to develop a workflow that brings together a number of off the shelf tools and newly developed methods to enhance the quality of OCR on historical materials.

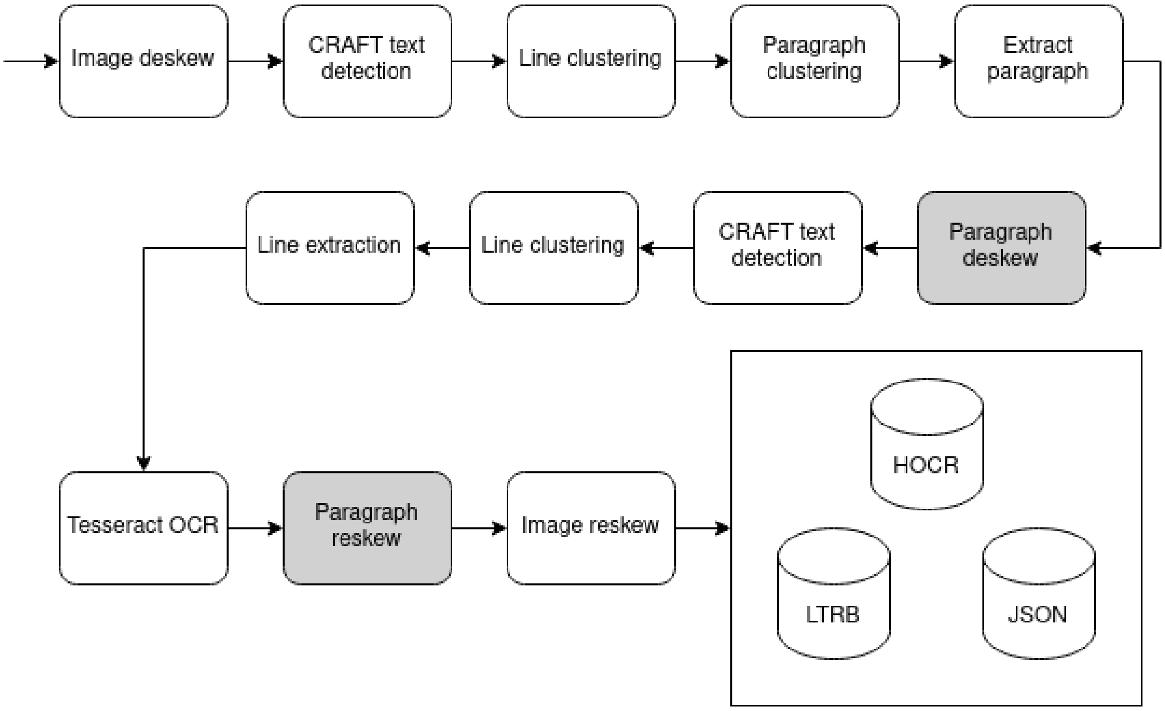


Figure 1: Overview of the OCR workflow

Firstly, we evaluated the performance of a Tesseract-based OCR workflow compared to manually created ground truth (GT) using the OCR-D Ground Truth Guidelines[1]. Secondly, the OCR evaluation tool CLEval was used to examine the OCR output.[2] The choice for CLEval was motivated by the fact that it shows both the text detection accuracy alongside the text recognition accuracy. This allows for a more clear interpretation of end-to-end accuracy of an OCR workflow.

[1] https://ocr-d.de/en/gt-guidelines/trans/

[2] https://github.com/clovaai/CLEval

To ensure that the user-generated OCR remains accessible and is easy to share with others, we convert the OCR output to IIIF-compliant format, which can be linked to the manifests (i.e. HOCR, ALTO). IIIF is a set of open standards for storing both images and metadata related to particular digital objects.

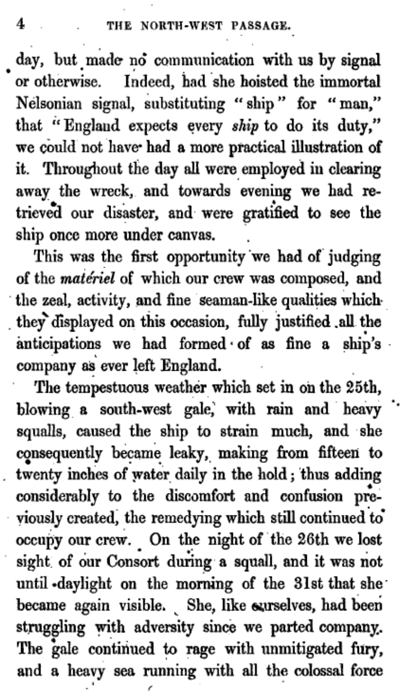
This open source approach emphasizes the need for preprocessing techniques to account for the specific characteristics of historical documents such as noise and text skew. The first results, as shown in Table 1, indicate that it outperforms the off-the-shelf version of Tesseract in terms of detection already and we expect that recognition can be even further improved with post-correction pipelines such as PICCL.

	Detection			**Recognition**			
	P	**R**	**F1**	**P**	**R**	**F1**	**CER**
Tesseract	92.5%	97.04 %	94.72%	90.44%	93.56%	91.97%	4.14%
OCR workflow	95.64%	96.09%	95.87%	91.81%	92.47%	92.14%	4.93%

Table 1: Comparison of our workflow with off-the-shelf Tesseract (P: Precision, R: Recall, F1: F1-score, CER: Character Error Rate)

Acknowledgements
We would like to thank a number of people and organizations for supporting this project. We thank our colleagues of the Ghent Centre for Digital Humanities (GhentCDH) for supporting us with their knowledge and feedback. We are grateful to the Ghent University Library (UB), the Department of Art History at Ghent University and specifically Prof. Marjan Sterckx and Malika M'rani Alaoui, Flemish Art Collection (VKC), and the Museum of Fine Arts (MSK) for providing us with the digitized source material.

Appendices



4 THE NORTH-WEST PASSAGE.

day, but made no communication with us by signal or otherwise. Indeed, had she hoisted the immortal Nelsonian signal, substituting "ship" for "man," that "England expects every *ship* to do its duty," we could not have had a more practical illustration of it. Throughout the day all were employed in clearing away the wreck, and towards evening we had retrieved our disaster, and were gratified to see the ship once more under canvas.

This was the first opportunity we had of judging of the *matériel* of which our crew was composed, and the zeal, activity, and fine seaman-like qualities which they displayed on this occasion, fully justified all the anticipations we had formed of as fine a ship's company as ever left England.

The tempestuous weather which set in on the 25th, blowing a south-west gale, with rain and heavy squalls, caused the ship to strain much, and she consequently became leaky, making from fifteen to twenty inches of water daily in the hold; thus adding considerably to the discomfort and confusion previously created, the remedying which still continued to occupy our crew. On the night of the 26th we lost sight of our Consort during a squall, and it was not until daylight on the morning of the 31st that she became again visible. She, like ourselves, had been struggling with adversity since we parted company. The gale continued to rage with unmitigated fury, and a heavy sea running with all the colossal force

Appendix 1: Art catalog, *salontentoonstelling.*

Bibliography

Aiello, K., & Simeone, M. (2019). Triangulation of History Using Textual Data. *Isis*, *110*(3), 522-537. https://doi.org/10.1086/705541

Blevins, C., & Robichaud, A. (2011). 2: A Brief History » Tooling Up for Digital Humanities. *Tooling Up for Digital Humanities*. http://toolingup.stanford.edu/?page_id=197

Bod, R., & Richards, L. (2015). *A new history of the humanities*. Oxford University Press

Kiessling, B. (2019). Kraken - a universal text recognizer for the humanities. DH 2019 Utrecht. https://dev.clariah.nl/files/dh2019/boa/0673.html

Kuhn, J. (2019). Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, *53*(4), 565–602. https://doi.org/10.1007/s10579-019-09459-3

Levenberg, I., Neilson, T., & Rheams, D. (2018). *Research Methods for the Digital Humanities*. Palgrave Macmillan.

Lopresti, D. (2008). Optical character recognition errors and their effects on natural language processing. *Proceedings Of The Second Workshop On Analytics For Noisy Unstructured Text Data - AND '08*. https://doi.org/10.1145/1390749.1390753

Martínek, J., Lenc, L., & Král, P. (2020). Building an efficient OCR system for historical documents with little training data. *Neural Computing and Applications*, *32*(23), 17209–17227. https://doi.org/10.1007/s00521-020-04910-x

Moretti, F. (2013). *Distant reading*. Verso.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., & Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 53–58. https://doi.org/10.1145/3322905.3322917

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence*, *3*, 62. https://doi.org/10.3389/frai.2020.00062

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, *54*(6), 1–37. https://doi.org/10.1145/3453476

Nguyen, T.-T.-H., Jatowt, A., Coustaty, M., Nguyen, N.-V., & Doucet, A. (2019). *Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing* (p. 38). https://doi.org/10.1109/JCDL.2019.00015

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., & Puppe, F. (2019). OCR4all—An Open-Source Tool Providing a

(Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences*, *9*(22), 4853. https://doi.org/10.3390/app9224853

Reynaert, Martin., Van Gompel, Maarten., van der Sloot, Ko. PICCL: Philosophical Integrator of Computational and Corpus Libraries. (2020). GitHub repository, https://github.com/LanguageMachines/PICCL.

Smith R., "An Overview of the Tesseract OCR Engine," *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007, pp. 629-633, https://doi.org/10.1109/ICDAR.2007.4376991

Smith, A. D., & Cordell, R. (2018). *A Research Agenda for Historical and Multilingual Optical Character Recognition—DRS*. https://repository.library.northeastern.edu/files/neu:f1881m035

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. *Proceedings Of The 12Th International Conference On Agents And Artificial Intelligence*. https://doi.org/10.5220/0009169004840496

Taunton, M. (2014). *Print culture*. The British Library; The British Library. https://www.bl.uk/romantics-and-victorians/articles/print-culture

Traub M.C., van Ossenbruggen J., Hardman L. (2015) Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In: Kapidakis S., Mazurek C., Werla M. (eds) Research and Advanced Technology for Digital Libraries. TPDL 2015. Lecture Notes in Computer Science, vol 9316. Springer, Cham. https://doi.org/10.1007/978-3-319-24592-8_19

Wick et al. (2018). Calamari − A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly*, 14(2).