A quantitative and qualitative evaluation of large scale handwriting recognition models for Norwegian

Solène Tarride, Yngvil Beyer, Per Erik Solberg, Mélodie Boillet, Christopher Kermorvant

The National Library of Norway's largest collection of private archive material from authors, artists, researchers, and other cultural figures. The majority of the collection stems from the 19th and 20th century. In the context of the ongoing Hugin Munin project [1], our goal is to develop datasets and methods that will enable good quality Handwritten Text Recognition (HTR) for the majority of the documents, in order to make them searchable and accessible with a separate digital text. The collection of hand written letters and manuscripts is catalogued and described either in an old card catalogue or in a digital catalogue [2]. The digital catalogue contains documents from more than 34,000 different writers. Parts of the documents in the digital catalogue are also accessible as digital objects (images and metadata) in the online library [3]. In the Digital Library there are currently 31,042 digitized letters and manuscripts, stemming from 4,468 of the 34,000 writers. Close to 13,000 of the digitized documents originate from the 19th century, and approximately the same number are from 1900-1950. During the project period, we have transcribed more than 5,000 of the digitized documents, and added and indexed ALTO-XML files in the digital library. Hence, they are now searchable in full text. Many of these are also used as training data in the NorHand dataset (see below).

While there are many open datasets available to train Handwritten Text Recognition models, they do not cover all languages equally. In their survey, Nikolaidou et al [4] conducted an extensive investigation of historical datasets dedicated to historical document recognition. While there are many datasets for Handwritten Text Recognition in English, French, Latin, Spanish, Italian and German, there is a noticeable lack of representation for Scandinavian languages. To fill this gap, we are releasing NorHand, the first dataset specifically tailored for Norwegian Handwritten Text Recognition. The NorHand dataset contains Norwegian letter and diary documents from the 19th and early 20th century. A small first version of our dataset, which contains about 800 pages, was released last year [5]. The latest version consists of more than 11,000 pages written by 393 different authors. Each page has been carefully annotated and transcribed, with detailed information on the localization of text lines and paragraphs, along with their transcriptions and reading order. In addition, each page is accompanied by meta-information such as the author's name and date of creation. To ensure reproducibility, we publish an official partitioning of our dataset into training, validation, and test sets. This data partitioning strategy was designed to maximize diversity in the test set: each test page was written by a different author. Additionally, the test set includes unseen writers and unseen document types, allowing us to evaluate out of sample generalisation. The latest NorHand dataset is available on Zenodo [6].

Our goal is to develop a robust document recognition system specifically tailored to the Norwegian language. To achieve this objective, we train and compare three recent handwritten text recognition (HTR) models on NorHand: Pylaia [7], trOCR [8] and DAN [9]. All the models trained as part of this project will be made publicly available on Hugging Face. Both Pylaia and trOCR work at the line level, which requires the training of a text line detector for comprehensive automatic processing. In our experiments, we train the Doc U-FCN model [10] for text line detection. In contrast, DAN operates at the page level, which eliminates the need for the initial segmentation step. Our experiments show that DAN is more reliable in an end-to-end evaluation setting, since any text line detection error is likely to trigger text recognition errors. For other

languages, these HTR models typically achieve a 5% character error rate and a 15-20% word error rate [4]. As a result, we expect similar performance on Norwegian documents. With a large dataset available for training, we observe that text recognition systems achieve very high quality. We perform a detailed analysis of the results for each page in the test set to identify the most difficult pages for all three HTR systems. Our analysis shows that the main difficulties lie in complex layouts, e.g. pages containing text with different orientations, combining handwritten and printed text, or containing illustrations.

We hope that the NorHand dataset will promote research on text recognition on historical Norwegian documents. The most effective model identified through our evaluation will be used to process handwritten documents from the National Library of Norway. This initiative aims to enhance their searchability in full text, ultimately improving accessibility for the general public.

Bibliography

[1] National Library of Norway's Letters and Manuscript collection https://beta.nb.no/dhlab/privatarkiv_navn/
[2] National Library of Norway's Digital Library https://www.nb.no/search?mediatype=brev-og-manuskripter
[3] Nikolaidou, K., Seuret, M., Mokayed, H. et al. A survey of historical document image datasets. IJDAR 25, 305–338 (2022). https://doi.org/10.1007/s10032-022-00405-8
[4] Maarand, M., Beyer, Y., Kåsen, A., Fosseide, K.T., Kermorvant, C. (2022). A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian. In: Uchida, S., Barney, E., Eglin, V. (eds) Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science, vol 13237. Springer, Cham. https://doi.org/10.1007/978-3-031-06555-2_27
[5] Beyer, Y., & Solberg, P. E. (2023). NorHand v3 / Dataset for Handwritten Text Recognition in Norwegian [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10255840
[6] J. Puigcerver, "Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 67-72, doi: 10.1109/ICDAR.2017.20.
[7] Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D.A., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. AAAI Conference on Artificial Intelligence.
[8] D. Coquenet, C. Chatelain and T. Paquet, "DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 07, pp. 8227-8243, 2023.
[9] Boillet, Mélodie & Kermorvant, Christopher & Paquet, Thierry. (2020). Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks.