Combining Automatic Text Recognition with Digital Paleography: New Trajectories for Digital Diplomatics

Tobias Hodel, University of Bern

Digitization led to many new possibilities for analyzing charters: Image reproduction allows for quick visual comparisons, text recognition makes documents searchable and instantly legible, and computer vision identifies visual entities such as heraldry. These technologies replicate "traditional" research questions and try to apply them to larger masses, higher accuracies, or "objective" accounts. The used technologies, mainly based on machine learning, would offer other research trajectories when the output of machine learning algorithms is taken seriously.

Up to 2014, within the DigiPal ERC project, an annotation suit for visual features was developed (see, e.g., Stokes 2012), formalizing the analysis of visual features and annotating manually. In the meantime, automatic text recognition received much attention, and new frameworks and engines emerged. To only name a few: Transkribus (Muehlberger et al. 2019), eScriptorium (Stokes et al. 2021), OCR4all (Reul et al. 2019). As a result, we can rely on new possibilities and a growing number of available documents; see HTR united (online: https://htr-united.github.io/) for some glimpses.

In the mentioned frameworks, text recognition of handwritten documents is based on neural networks/supervised deep learning. For supervised machine learning, training material (correctly transcribed texts) must be aligned to images for training specific models. The developed frameworks showed even to be capable of recognizing unknown hands and thus, dealing with writing styles (Hodel et al. 2021). For medieval charters, this opens up the possibility of processing large amounts of documents in similar writings and basing diplomatics and history, as well as paleography, on a broader basis.

The proposed paper takes on the challenge of taking the output of a machine learning algorithm seriously and not only as a potentially wrong assumption that needs to be corrected by a human. When applying text recognition, some recognition engines, like HTR+ by Planet AI (and usable within the Transkribus platform), not only decide on the most likely character at a particular image slide but also the likelihood of all characters appearing in training set at the image slide in a so-called confidence matrix. This information can be used to search for extremely high recalls (over 99,5%) and – the focal point of this proposal – to extract paleographical information about documents.

Combining the DigiPal approach (but relying on computer output) with text recognition promises new insights into paleographical structures of documents, especially charters. A stable recognition is possible with large text recognition models for specific scripts (Latin minuscule) or times (German, 19th century). At the same time, we can use such models to compare writing styles by scoring similarities on identified characters and adjacent incorrect alternatives stored in confidence matrices. In theory, the information could be harvested to identify scribes. However, for our purpose, it is paleography on a scale, allowing us to analyze characters' similarities from a computational point of view. On a methodological level, this will help us explain how a deep learning algorithm "sees" a text line or even parts of single characters.

The proposed paper is planned as a presentation and not as a poster.

Bibliography:

Hodel, Tobias, David Schoch, Christa Schneider, und Jake Purcell. „General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an

Example“. *Journal of Open Humanities Data*, Journal of Open Humanities Data, 7 (2021). https://doi.org/10.5334/johd.46.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, u. a. "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study ". *Journal of Documentation* 75, Nr. 5 (9. September 2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, und Frank Puppe. „OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings“. *Applied Sciences* 9, Nr. 22 (Januar 2019): 4853. https://doi.org/10.3390/app9224853.

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, und Gargem El Hassane. "The EScriptorium VRE for Manuscript Cultures – Classics@ Journal ". *Classics@* 18 (2021). https://classics-at.chs.harvard.edu/the-escriptorium-vre-for-manuscript-cultures/.

Tobias Hodel is an assistant professor on the tenure track for digital humanities at the University of Bern. Hodel is a Medieval Historian and got a Ph.D. from the University of Zurich in 2016. His research focuses among others, on applications of machine learning in the humanities.