MONK in Practice:
Indexing Heterogeneous Handwritten Collections

Anna Caceres[1], Andreas Weber[2], Lambert Schomaker[3]

[1] Leiden University, Postbus 9500, 2300 RA Leiden, The Netherlands
annamcaceres@googlemail.com
[2] University of Twente, BMS-STePS, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl
[3] University of Groningen, Bernoulli Institute, Nijenborgh 9, 9747 AG Groningen, The Netherlands
l.r.b.schomaker@rug.nl

Abstract: This short paper describes how MONK, a machine-learning driven handwriting recognition system, can be used to rapidly index a heterogeneous handwritten collection with the help of volunteers. We discuss the setup and results of an event which saw volunteers come together to enrich a subset of the digitized Prize paper collection, a collection of historical handwritten documents of the High Court of Admiralty (1652-1815).

Keyword: *handwriting recognition, user study, heterogeneous archives, archives, active learning, Prize papers, machine learning*

Over the last decades archives, museums, research institutions and publishers have undertaken major efforts to index their digitized handwritten collections. This paper describes the setup and results of an event which saw 14 expert volunteers come together to enrich a digitized collection of a visually heterogeneous archive – the Prize Papers - (see figure 1) using MONK, a machine-learning driven handwriting recognition system developed at the University of Groningen. MONK does not require prior training. It starts from scratch and actively and continuously learns from the input of users (Schomaker, 2016 and 2019). The event took place in the offices of Brill publishers in Leiden in October 2019 and took less than one working day, with time for instruction.

Figure 1: *Snippets from the Prize paper collection in the MONK system to show heterogeneity of script-styles.*

An indexing rather than line-by-line transcription focus, meant targeting labels on words known to be of indexable significance, or targeting areas of the document where such words were predicted to appear. Here, the format of the archive itself is of significance. The Prize Papers are the records of the High Court of the Admiralty, a British maritime legal body, and date from 1652 - 1815. (Van Lottum & Zanden, 2014). The archive consists of two parts: standardised interrogations of crew members on one hand, and miscellaneous seized documents from the ships on the other. For our use case we took a sample of 2111 pages from the interrogations, which were valuable because they presented a variety of script styles, paired with a highly standardised content, consistently asking a set of roughly 32 questions.

When identifying target zones for labelling then, we knew for example, that questions 7 and 8 enquire about the name, destination and origin of the ship so index-focused labelling should concentrate on these areas. Indexation is more valuable in the short and medium term, as it immediately increases the searchability, and thus useability, of historical documents (Zant et al., 2009; Colavizza, Ehrmann and Bortoluzzi, 2019). It further provides continuous learning systems with targeted training for word classes which typically appear less frequently, such as place names, people names and objects.

Volunteers were assigned different labelling activities targeting both breadth (number of word classes recognised) and depth (accuracy of recognition) of knowledge in MONK. MONK generates suggestions both for word zones (beginning and ends of words) and word classes (alphabetic content) which users confirm or reject in various formats. Figure 2 shows a single-word hit list for "Brigantine", one example of machine-generated and human-corrected labelling. Training in different functions allowed both specific words and specific pages to be targeted.

***Figure 2**: Example of a resulting hit list for word 'Brigantine' after labeling, using an LSTM recognizer in Monk (Ameryan & Schomaker, 2019). Green samples were used for training, Samples in light red correspond to the new harvest. A previously misrecognized sample 'Friancourt' (dark red) is now correctly recognized as Brigantine.*

The labelling efforts were primarily successful along the breadth axis with 113 new word classes labelled and a doubling of total transcribed lines - from 1143 to 2224. Along the depth axis there was an overall increase in word accuracy of 3.87% thanks to 761 newly labelled instances. This short paper details how MONK can facilitate the rapid indexation of heterogeneous archival material with a very limited involvement of volunteers. However, in order to make more general statements about the system's efficiency a much larger benchmarking study would be necessary.

References:

Ameryan, M. & Schomaker, L. (2019). A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Colavizza, G., Ehrmann, M. and Bortoluzzi,F., “Index-Driven Digitization and Indexation of Historical Archives,” Frontiers in Digital Humanities, 6:4 (2019). https://doi.org/10.3389/fdigh.2019.00004

Lottum J. van & Zanden, J.L., “Labour Productivity and Human Capital in the European Maritime Sector of the Eighteenth Century,” Explorations in Economic History, vol. 53, pp. 83-100, 2014.

van der Zant, T., Schomaker, L., Zinger, S., & van Schie, H. (2009). Where are the Search Engines for Handwritten Documents? Interdisciplinary Science Reviews, 34(2-3), 224-235. https://doi.org/10.1179/174327909X441126

Schomaker, L. (2019). Lifelong learning for text retrieval and recognition in historical handwritten document collections, arXiv:1912.05156 [chapter in book]

Schomaker, L. (2016). Design considerations for a large-scale image-based text search engine in historical manuscript collections. Information Technology, 58(2), 80-88. https://doi.org/10.1515/itit-2015-0049