Transcription Alignment for Highly Fragmentary Historical Manuscripts: The Dead Sea Scrolls

Daniel Stökl Ben Ezra
AOrOc (UMR 8546)
EPHE, PSL
Paris, France
daniel.stoekl@ephe.psl.eu

Bronson Brown-DeVost
Faculty of Theology
Georg-August-University
Göttingen, Germany
https://orcid.org/0000-0003-3655-7807

Nachum Dershowitz
School of Computer Science
Tel Aviv University
Ramat Aviv, Israel
nachum@tau.ac.il

Alexey Pechorin
School of Computer Science
Tel Aviv University
Ramat Aviv, Israel
alexey.pechorin@gmail.com

Benjamin Kiessling
AOrOc (UMR 8546)
EPHE, PSL
Paris, France
benjamin.kiessling@ephe.psl.eu

***Abstract*—Most of the Dead Sea Scrolls have now been digitally transcribed and imaged to very high standards. Our goal is to align the transcriptions with the text visible in the image, glyph by (often fragmentary) glyph. This involves several tasks, normally considered in isolation: (A) Baseline segmentation. (B) Line polygon extraction. (C) Automated transcription by handwritten character recognition, to aid in alignment. (D) Alignment of the Unicode characters in a line transcription with the characters in the image of that line. The task is frustrated by the degraded nature of the frequently very small and/or warped fragments with many broken letters, substantially different allographs, ligatures, and scribal idiosyncrasies. Furthermore, a great number of inconsistencies between current cataloguing systems for the data need to be resolved. For each task, we apply state-of-the-art machine-learning methods in addition to more traditional techniques, each presenting significant difficulties on account of the poor state of most fragments' preservation. We have built ground-truth datasets and have managed to achieve good results with well-preserved fragments by leveraging heavily augmented transfer learning from prior work with medieval manuscripts.**

***Index Terms*—historical manuscripts; transcription alignment; image segmentation**

Dedicated to the late Yaacov Choueka (1936–2020), pioneer in natural language processing and historical manuscript analysis. May his memory be blessed.

I. Background

The Dead Sea Scrolls (dated to the turn of the Common Era) are of enormous historical significance. They are the oldest witnesses of the biblical books and contain a treasure-trove of texts that have shed light on ancient Judaism shortly before and up into the time of Jesus and Paul. Their study continues to revolutionize our understanding of the evolution of Judaism and the emergence of Christianity. Unfortunately, the scrolls, or rather the fragments, were discovered in the 20th century CE in very poor condition, having deteriorated over the millennia. Few are large enough to contain even several columns.[1] The vast majority show only a low number of words or even just a few letters, many of which are only partially visible. Over the past decades, all texts have been painstakingly transcribed by scholars. The texts are so fragmentary that the editions have developed systems to distinguish between certain, probable, possible, and entirely restored letters. Taking only the certain and probable letters into consideration, the average fragment contains about 53 letters. However, the few scrolls with almost entirely preserved columns skew the mean as the median fragment contains only 13 letters. The extant fragments have recently been digitized by the Israel Antiquities Authority (IAA) using state-of-the-art multispectral imaging.[2] Older infrared images, photographed under the auspices of the Palestine Archaeological Museum (PAM) in the 1950s, are likewise available in retrodigitized form.

Virtually all the texts have been transcribed and most appeared in the *Discoveries in the Judaean Desert (DJD)* series,[3] in Qimron's edition,[4] and in the *Qumran-Wörterbuch (QWB)* database of the Akademie der Wissenschaften zu Göttingen.[5] As the only resource that was truly computationally accessible at the start of this project, we based our work on the latter. See Fig. 2.

II. Introduction

Our objective is to develop an automated system to align transcriptions of the texts of the scroll fragments with the visible glyphs on the scroll images on the individual glyph level. Achieving this end requires isolating the fragments in an image from the background so that its text lines can be identified. The alignment of transcribed letters with glyphs appearing in the

[1] For an example of a fragment from the book of Leviticus, see Fig. 1.
[2] https://www.deadseascrolls.org.il/explore-the-archive
[3] http://orion.mscc.huji.ac.il/resources/djd.shtml
[4] https://zenodo.org/record/3737950#.XoXRs6gzaiM
[5] http://www.qwb.adw-goettingen.gwdg.de

images of each line is then aided by recognizing at least some of the letters and spaces. The processes developed here are closely related to the ongoing work of several major research projects.

The DIP *Scripta Qumranica Electronica*[6] (SQE) aims to provide the scholarly community with an open source web-based portal for the purposes of material analysis of the scrolls. It combines the high-resolution image database of the IAA with the QWB lexical database. A feature-rich suite of digital tools and computational methods are brought together to create an infrastructure for the production of digital editions [1].

The University of Groningen project, *The Hands that Wrote the Bible*,[7] is dedicated to investigating the paleography of the Dead Sea Scrolls. It has so far produced a benchmark study in writer identification [2] and advances in dating [3] and in binarization techniques of these manuscripts [4].

Nearly all of the Dead Sea Scrolls are written in Hebrew letters on animal skin, i.e. parchment, but phenomenologically they are very close to papyrus, which is mostly from ancient Egypt and written in Greek. The University of Basel project, *Reuniting fragments, identifying scribes and characterizing scripts: the Digital palaeography of Greek and Coptic papyri*,[8] organized a binarization competition [5] and published a dataset on writer identification [6]. In addition, the Wuerzburg-Heidelberg-Paris project, PapyroLogos, works on text-image alignment of literary and documentary Greek papyri [7].

Other major projects on Hebrew manuscript material include the Friedberg Genizah Project, which digitized hundreds of thousands of fragments of medieval manuscripts, mostly in Hebrew, Judeo-Arabic, and Aramaic [8].[9] State-of-the-art computational tools were developed for segmentation [9], paleography [10], matching fragments by handwriting and codicological features [11], and word spotting [12].

The Sofer Mahir project strives to create open source transcriptions of ca. 6000 pages of 18 substantial manuscripts of the earliest Rabbinic literature (Mishnah, Tosefta and Midreshei Halakhah).[10] In the Tikkoun Sofrim project, crowdsourcing and machine learning is used to correct errors in the automatic transcriptions of manuscripts of medieval exegetical literature [13].[11]

III. Methods and Related Work

Different infrastructures allow automatic interaction with historical manuscripts (a brief overview is given in [14]). Most notable are Transkribus [15][12] and MONK [16],[13] which are however not open source and, at least in the case of Transkribus also commercial,[14] and therefore much more difficult or even impossible to include in a full treatment pipeline.

[6]http://qumranica.org
[7]https://cordis.europa.eu/project/id/640497
[8]https://altegeschichte.philhist.unibas.ch/de/digpaleo
[9]https://fgp.genizah.org
[10]https://sofermahir.hypotheses.org
[11]https://tikkunsofrim.hypotheses.org
[12]https://transkribus.eu/Transkribus
[13]http://www.ai.rug.nl/~lambert/Monk-collections-english.html
[14]https://readcoop.eu/transkribus-pricing

Fig. 1: Manuscript fragment (Leviticus 3) after imperfect foreground segmentation. All images of fragments are courtesy of the Leon Levy Dead Sea Scrolls Digital Library, Israel Antiquities Authority. Photos: Shai Halevi.

[...]על המזבח סביב והקריב מזבח השלמים[...]	1
[...]חלב המכסה את ה[...]העצה יסירנה ואת[...]	2
[...] הכליות ואת החלב אשר עליהן אשר ע[...]	3
[...] נה והקו[ט]יר הכהן המזבחה ל[...]	4
[...]ני יהוה וסמך את [...]	5
[...]ל[המז]בח סביב והז[...]	6
[...]ל[...]	7

Fig. 2: Scholarly transcription of the fragment (4Q24 fr. 8) in Fig. 1.

The only cutting edge and fully open-source infrastructure for historical document analysis we know of is eScriptorium [17].[15] Accordingly, we have made use of its tools and have performed the following procedures.

A. Line Segmentation

After a long predominance of methods relying on traditional computer vision approaches to perform text line extraction from handwritten documents, machine learning based systems have seen wider use recently [18]–[22]. The majority of these methods utilize combinations of CNNs and LSTMs. Still, traditional methods from computer vision can have advantages for certain tasks or types of manuscripts [23]–[25]. We tested several hand-crafted line segmentation algorithms without success, settling on a trainable method described in [7], [18], as implemented in kraken [26] and eScriptorium. Layout analysis is independent of binarization and works very well even on highly fragmentary and damaged material (such as the Dead Sea Scrolls and the Genizah); see Fig. 3.

B. Automated Transcription

In line with the state of the art in text image classification we utilize a hybrid CNN-RNN trained in a supervised manner to classify sequences of characters on whole text lines using the connectionist temporal classification loss [27]. The kraken OCR engine's recognizer with default parameters is used instead of a custom implementation.

Due to the challenging nature of the material such as high script variability and extensive degradation, even the best modern OCR engines perform quite poorly ($<$ 90% CER). While

[15]https://escripta.hypotheses.org

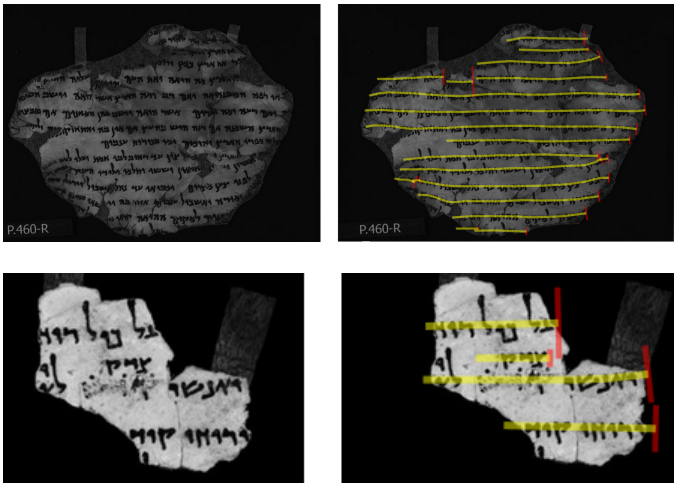

Fig. 3: Automatic segmentation result (left without, right with baselines marked in yellow and an additional right vertical bar marking the beginning) of a large (top) medium (bottom) size fragment.

unsuitable for close reading, even poor-quality OCR output can be serviceable for novel applications like intertextuality and search. In contrast to exact search as implemented in standard search engines, which yields very limited results, approximate search can be both applied to finding individual phrases [28], [29] and to matching against an existing corpus [30]. Our method for text identification based on approximate search is detailed in Section IV-C.

C. Transcription Alignment

Early work on aligning OCR text with ground truth is presented in [31]. More recent work includes [32]–[39]. We experimented with (a) optical SIFT-flow [35], (b) alignment with OCR results – by means of minimal edit distance, and (c) a combination – using anchors obtained from the OCR to constrain the optical flow.

For optical SIFT-flow we first render the known Unicode transcription as a line image in a manner and font that is similar to the manuscript. Next, a visual alignment is made between the synthetic transcription image and the original manuscript image by the SIFT flow image matching algorithm introduced in [40]. Since we have information regarding the letter boundaries in the rendered image, these boundaries translated into the manuscript image by the retrieved optical alignment result in an approximation of the letter boundaries in the manuscript image, thus resulting in a glyph alignment. To enhance the visual alignment, we may use previously discovered correspondences between the rendered image and the manuscript image, which we call anchors for the optical alignment, in order to align the images more precisely. These anchors might be gained, for instance, from character or inter-word bounding boxes found by the OCR algorithm.

In the OCR based method, we first train a recognition model on the known transcription-line pairs with kraken until the system overfits the data. We then apply the same model on the data on which it was trained. We can extrapolate the approximate x-coordinate of the character boundaries based on the highest activation time-stop returned by the system for a given character. The y coordinates can be estimated from the line polygon.

IV. Experimental Results

A. Corpus Sample

The base data for the following analyses are the images from the Leon Levy Dead Sea Scrolls Digital Library and the text transcription of the Qumran Wörterbuch Project. These projects made use of two different and only sometimes overlapping cataloguing systems, which complicated the correlation of image to a specific set of transcriptions. After aligning the two systems by applying various adaptive rules for entry matching and some manually specified correlations, images were selected for which reliable matches to the textual transcriptions were available.[16]

For many reasons, the catalogue remains perfectible. The definition of what is a fragment is not straightforward and the fragments continued to "live" and change after their publication. In the new photographs, a fragment is a physical unit that can be lifted in one piece from its archival plate. However, such a unit may constitute several different fragments in the editions that have been joined, for instance, with Japanese rice paper in a later conservation process. In other cases, the editor gave a single identifier to what constitute several distinct physical units depicted on distinct images. Other fragments, still in one piece in the edition, have since broken up or disintegrated into several pieces. Some images in the image database are identified as representing a specific fragment while in fact the current fragment only contains a fraction of the published text. Several identifications were incorrect, and a few imaged fragments were not identified at all. Therefore, we had to verify the identification of each image with its corresponding transcription from the QWB database.

B. Line Segmentation

As a first step, we transfer-learned a new baseline segmentation model on top of models trained initially on medieval manuscripts and Greek papyri. We bootstrapped the training material following a common procedure: Firstly, we manually annotated 100 images of Qumran fragments, used them as training material, and afterwards applied the results to 300 more images of Qumran fragments. We then manually corrected the automatic results in the eScriptorium web interface and used that larger corpus to train another model to apply to ca. 500 more images. The ergonomic interface of eScriptorium makes this usually cumbersome process very easy. While manually annotating the baselines of an image from scratch takes approximately 90 seconds, the average manual correction time for an automatically segmented image is less than 30 seconds for the first stage and less than 15 for the second stage. However, depending on the complexity of the layout, the time

[16]The results of the merging of these two catalogues can be accessed through the SQE web API https://api.qumranica.org/swagger.

Fig. 4: Imageline to textline alignment result as displayed in eScriptorium. Baselines are depicted in yellow, boundary polygons in alternating red and blue.

needed for an image can differ markedly. Many images require few or no corrections. See Fig. 4 for an example.

Some fragments have been imaged at a rotation angle other than upright. Consequently, we determine the correct reading order based on the median principal writing angle of the baselines, taking into consideration the writing direction. In the near future, we will add the new kraken and escriptorium feature for automatic segmentation of regions to the pipeline to improve the results for multi-column fragments, especially regarding the reading order [41].

C. Automated Transcription

In a second step, we extracted textual data from the QWB database and matched it linewise to the segmented lines on the images. We retained only fragments written in Judean square script leaving out any fragment written in paleo-Hebrew, Cryptic C, Greek, or Nabatean. Still the hands of the fragments vary widely in register, formality, and period and represent many different scribal habits. The dates the scrolls were written could vary by 300 years in a very "hot" period, that is, a period with massive changes in ductus according to local schools after the disintegration of the relatively unified Imperial writing system of the Persian Empire. We discarded all letters marked as restored or as merely possible readings, keeping only the probable and certain instances. Due to the aforementioned complications inherent to the fragments, editions, and the database, the "zipping" together of the image and the textual data is not a trivial process. Therefore, the rough OCR of the extant text in the next step, provided a welcome check, (1) whether the identification of the fragment image with the corresponding text was correct, (2) whether the fragment was still complete, and/or (3) whether it had been joined with other fragments.

The third step consisted of training a transcription model on the selected ground truth with a 90/10 training/testing split on the grayscale images (without binarization). Discarding misidentified items and fragments in other scripts, the final training material comprised 33075 characters on 2474 lines from 440 images. The testing material read 3403 characters on 247 lines from 44 images. On the average, we can count 5–6 lines and ca. 75 characters (including spaces) per fragment. These are thus relatively large fragments. New models were trained on top of the models previously trained on medieval Hebrew manuscripts.

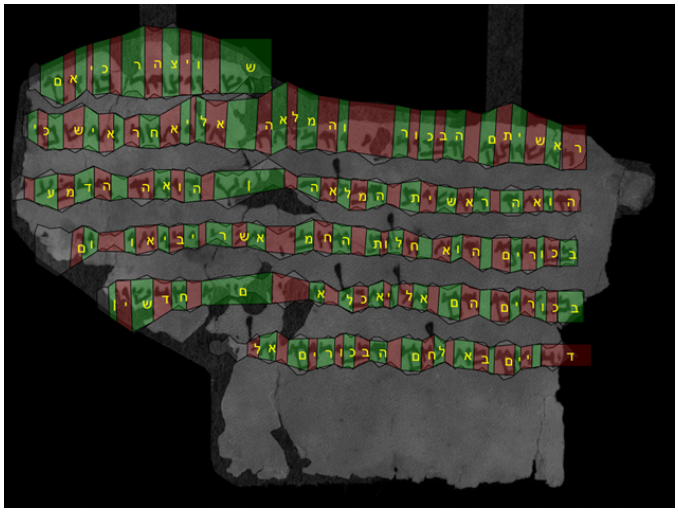

Fig. 5: Aligned glyphs of a whole fragment. Alternating red and green polygons indicate areas. Yellow overlay indicates identified letter.

The best model reached an accuracy of 67.9% on the test material after 21 epochs. While this may seem very low, we applied the trained models to fragments outside of the training and test corpora, and the automatic transcriptions were extremely convincing for most fragments. The results are in fact better than the numbers indicate because frequently the transcriptions include very partial letters of which sometimes only scant remains are visible, in particular in the top and the bottom row of fragments, but not only. Even experts would typically have to expend significant effort evaluating the best reading possibilities.

In particular, the OCR results are sufficient to identify the fragments. With a bag of words approach for identification and with rotations every 90° to choose the best angle for recognition, the system was able to identify 22 out of 24 available fragments comprising more than 100 characters. In other words, given the imperfect OCR of each fragment and searching for the words among all 5756 transcriptions in QWB, the best match was indeed the actual scholarly transcription of the fragment in question. The two exceptions were fragments for which the system preferred a fragment of the same composition but from a different manuscript.

D. Transcript Alignment

To evaluate the various transcription alignment algorithms' performance, we compared the automatic alignment with the bounding boxes from a set consisting of 1278 letters that had been expertly segmented by hand using the Scripta Qumranica Electronica website. We denote a glyph as correctly aligned if the correct glyph in the manuscript is the closest one to the computed glyph location. To measure the distance between letters, we use Euclidean distance between the centers of the bounding boxes of the glyphs.

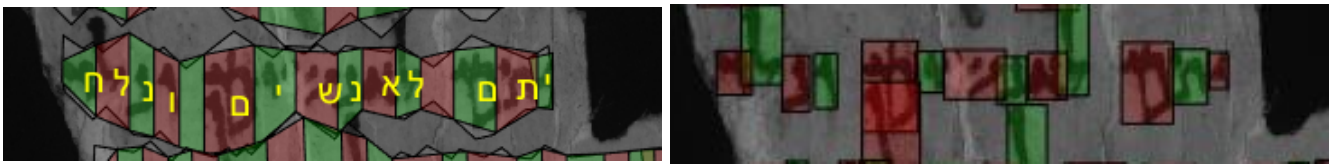

Fig. 6: Aligned glyphs of a single line. Left: Automatic alignment with alternating red and green polygons indicate areas. Yellow overlay indicates identified letter. Right: Corresponding human annotated ground truth (no interword spaces, no letter overlay).

Transcription alignment accuracy	
Method	Accuracy
Optical flow without anchors	48.1%
Optical flow with added anchors	74.0%
OCR derived alignment	**90.3**%

To further examine the performance of our leading transcription alignment method, we measured the proportion of the intersection area of the bounding polygon of the glyph of the OCR system and the human annotated ground truth:

OCR bounding boxes overlap with ground truth		
Statistic	Area	Area percentage
Average	31.0	81.0%
Median	23.8	87.1%
Standard deviation	30.8	20.5%

As can be seen, the accuracy of the OCR based transcription alignment method is the highest among the methods we've used, and the intersection of the recognized bounding polygon with the original glyph bounding box is high as well. An alignment example is displayed in Fig. 5. The interword space in line 2, for instance, has been well detected and shows that our alignment method provides excellent results on the word level. Fig. 6 shows aligned glyphs of a single line compared to the ground truth. Finally, the ground truth allows for overlapping glyph bounding boxes, a feature impossible for the current and all other known algorithms.

An analysis of the errors shows that the results can be further improved as some letters and some positions quite consistently reveal a higher error proportion. The letter lamed, which has a high ascender, is frequently cut below its top by the seamcarve algorithm, often because of the deterioration of the writing material. Similarly final mem has a long descender and can be cut too high. Finally, the seamcarve algorithm uses the neighboring lines to limit the height of rows. This is impossible for the upper boundary of the top and the lower boundary of the bottom rows. For all of these problems with y coordinates, obvious solutions tailored to the type of script and material are available. Otherwise, the method use should be able to be applied to other sequential scripts.

V. Discussion

We have put together an end-to-end automated pipeline for processing images and transcriptions of the very fragmentary Dead Sea Scroll manuscripts. Despite the many difficulties posed by the often seriously degraded material, the quality of segmentation and character recognition were sufficient to allow a glyph-by-glyph alignment of existing transcriptions to the new, high-quality images. The successful identification of fragments based on automatic transcriptions holds promise of helping to identify some of the remaining unidentified fragments of the image database with their counterparts in the text database. Each of the stages of the pipeline, viz. (A) baseline layout analysis of the fragment and (B) segmentation into line polygons, (C) rough automated transcription of the text in each of the fragment lines, and (D) alignment of the rough automatic transcription to the scholarly transcriptions to the image of the fragment, can be improved further.

The successful automated alignment of transcriptions to images will allow a textual layer to be added to the IAA images. This means that scholars and laypersons alike will be able to enter search terms and retrieve images containing them. It will also supply additional training data for improved character recognition and future paleographical analyses.

Acknowledgments

This research was supported in part by Grant BE 5916/1-1 KR 1473/8-1 from the Deutsch-Israelische Projektkooperation (DIP) and by Grant Agreement No. 871127 from the European Union's Horizon 2020 Research and Innovation Programme. It was made possible thanks to images taken by Shay Halevi and provided by the Leon Levy Dead Sea Scrolls Digital Library of the Israel Antiquities Authority, all rights reserved. We thank the SQE project members, especially Oren Ableman, Adiel Ben-Shalom, and Lior Wolf.

References

[1] B. Brown deVost, "Scripta Qumranica Electronica (2016–2021)," *Hebrew Bible and Ancient Israel*, vol. 5, pp. 307–315, 2016.

[2] M. A. Dhali, S. He, M. Popovic, E. Tigchelaar, and L. Schomaker, "A digital palaeographic approach towards writer identification in the Dead Sea scrolls," in *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2017, Porto, Portugal, February 24-26, 2017*, M. D. Marsico, G. S. di Baja, and A. L. N. Fred, Eds. SciTePress, 2017, pp. 693–702.

[3] M. A. Dhali, C. N. Jansen, J. W. de Wit, and L. Schomaker, "Feature-extraction methods for historical manuscript dating based on writing style development," *Pattern Recognition Letters*, vol. 131, pp. 413–420, 2020.

[4] M. A. Dhali, J. W. de Wit, and L. Schomaker, "Binet: Degraded-manuscript binarization in diverse document textures and layouts using deep encoder-decoder networks," *CoRR*, vol. abs/1911.07930, 2019. [Online]. Available: http://arxiv.org/abs/1911.07930

[5] I. Pratikakis, K. Zagoris, X. Karagiannis, L. T. Tsochatzidis, T. Mondal, and I. Marthot-Santaniello, "ICDAR 2019 competition on document image binarization (DIBCO 2019)," in *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*. IEEE, 2019, pp. 1547–1556.

[6] H. A. Mohammed, I. Marthot-Santaniello, and V. Märgner, "Grk-papyri: A dataset of Greek handwriting on papyri for the task of writer identification," in *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, 2019, pp. 726–731.

[7] B. Kiessling, D. Stökl Ben Ezra, R. Ast, and H. Essler, "Aligning extant transcriptions of documentary and literary papyri with their glyphs," 2019, 29th International Congress of Papyrology (Lecce). [Online]. Available: https://d-scribes.philhist.unibas.ch/en/events-179/neo-paleography-conference/poster-session-copy-1-237/

[8] Y. Choueka, "Computerizing the Cairo Genizah: Aims, methodologies and achievements," *Ginzei Qedem*, vol. 8, pp. 9*–30*, 2012.

[9] R. Shweka, Y. Choueka, L. Wolf, and N. Dershowitz, "Automatic extraction of catalog data from digital images of historical manuscripts," *Literary and Linguistic Computing*, vol. 28, no. 2, pp. 315–330, Feb. 2013.

[10] L. Wolf, N. Dershowitz, L. Potikha, T. German, R. Shweka, and Y. Choueka, "Automatic paleographic exploration of Genizah manuscripts," in *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*, ser. Schriften des Instituts für Dokumentologie und Editorik, F. Fischer, C. Fritze, and G. Vogeler, Eds. Germany: Norderstedt: Books on Demand, 2011, vol. 3, pp. 157–179.

[11] L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka, and Y. Choueka, "Identifying join candidates in the Cairo Genizah," *International Journal of Computer Vision*, vol. 94, no. 1, pp. 118–135, Aug. 2011.

[12] A. Ben-Shalom, Y. Choueka, N. Dershowitz, and L. Wolf, "Querying the Cairo Genizah images with word-spotting algorithm," in *The Twelfth Annual Jerusalem Conference on the Digitisation of Cultural Heritage*, Jerusalem, Israel, Nov. 2015, (Abstract).

[13] T. Kuflik, M. Lavee, D. Stökl Ben Ezra, A. Ohali, V. Raziel-Kretzmer, U. Schor, A. Wecker, E. Lolli, and P. Signoret, "Tikkoun Sofrim – combining HTR and crowdsourcing for automated transcription of Hebrew medieval manuscripts," in *Digital Humanities (DH2019)*, Utrecht, The Netherlands, 2019.

[14] B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra, "eScriptorium: An open source platform for historical document analysis," in *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22–25, 2019*, 2019, pp. 19–24.

[15] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, "Transkribus – a service platform for transcription, recognition and retrieval of historical documents," in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[16] L. Schomaker, "Lifelong learning for text retrieval and recognition in historical handwritten document collection," in *Handwritten Historical Document Analysis, Recognition, and Retrieval – State of the Art and Future Trends*, ser. Machine Perception and Artificial Intelligence, A. Fischer, M. Liwicki, and R. Ingold, Eds. World Scientific, 2020.

[17] B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra, "eScriptorium: An open source platform for historical document analysis," in *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, Sydney, Australia, Sep. 2019, p. 19.

[18] B. Kiessling, D. Stökl Ben Ezra, and M. T. Miller, "BADAM: A public dataset for baseline detection in Arabic-script manuscripts," in *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2019, Sydney, NSW, Australia, September 20-21, 2019*. ACM, 2019, pp. 13–18.

[19] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, "Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents," in *13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 351–356.

[20] M. Fink, T. Layer, G. Mackenbrock, and M.-a. Sprinzl, "Baseline detection in historical documents using convolutional u-nets," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 37–42.

[21] M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos, "cBAD: ICDAR2017 competition on baseline detection," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 1355–1360.

[22] B. Barakat, A. Droby, M. Kassis, and J. El-Sana, "Text line segmentation for challenging handwritten document images using fully convolutional network," in *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, pp. 374–379.

[23] G. Sadeh, L. Wolf, T. Hassner, N. Dershowitz, and D. Stökl Ben Ezra, "Viral transcript alignment," in *ICDAR*, 2015, pp. 711–715.

[24] M. Seuret, D. Stökl Ben Ezra, and M. Liwicki, "Robust heartbeat-based line segmentation methods for regular texts and paratextual elements," in *HIP@ICDAR*, 2017, pp. 71–76.

[25] D. Stökl Ben Ezra and H. Lapin, "Z-profile: Holistic preprocessing applied to Hebrew manuscripts for HTR with Ocropy and kraken," *Manuscript Cultures*, to appear.

[26] B. Kiessling, "Kraken – an universal text recognizer for the humanities," in *Digital Humanities (DH 2019)*, 2019.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[28] M. Christodoulakis, G. Brey, Uppal, and R. Ahmed, "Evaluation of approximate pattern matching algorithms for OCR texts," in *Proceedings of the 4th Annual Conference on Advances in Computing and Technology (AC&T)*, The School of Computing and Technology, University of East London, 2009, pp. 35–42.

[29] T. Badamdorj, A. Ben-Shalom, N. Dershowitz, and L. Wolf, "Fast search with poor OCR," *arXiv:1909.07899v2 [cs.IR]*, 2019, DH 2020.

[30] A. Zhicharevich, "Tools to aid OCR of Hebrew character manuscripts," Master's thesis, Tel Aviv University, 2012.

[31] J. D. Hobby, "Matching document images with ground truth," *International Journal on Document Analysis and Recognition*, vol. 1, pp. 52–61, 1998.

[32] S. Feng and R. Manmatha, "A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books," in *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006, Chapel Hill, NC, USA, June 11-15, 2006, Proceedings*, G. Marchionini, M. L. Nelson, and C. C. Marshall, Eds. ACM, 2006, pp. 109–118.

[33] I. Z. Yalniz and R. Manmatha, "A fast alignment scheme for automatic OCR evaluation of books," in *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*. IEEE, 2011, pp. 754–758.

[34] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2011, Beijing, China, September 16-17, 2011*. ACM, 2011, pp. 29–36.

[35] T. Hassner, L. Wolf, and N. Dershowitz, "OCR-free transcript alignment," in *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR 2013, Washington, DC)*, Aug. 2013, pp. 1310–1314.

[36] Y. Leydier, V. Eglin, S. Bres, and D. Stutzmann, "Learning-free text-image alignment for medieval manuscripts," in *14th International Conference on Frontiers in Handwriting Recognition (ICFHR 2014, Crete, Greece, September 1–4, 2014)*. IEEE Computer Society, 2014, pp. 363–368.

[37] T. Hassner, L. Wolf, N. Dershowitz, G. Sadeh, and D. Stökl Ben Ezra, "Dense correspondences and ancient texts," in *Dense Image Correspondences for Computer Vision*, T. Hassner and C. Liu, Eds. Switzerland: Springer-Verlag, 2016, pp. 279–295.

[38] Y. Leydier, V. Eglin, S. Bres, and D. Stutzmann, "Alignement texte-image sans apprentissage pour les manuscrits médiévaux," in *CORIA 2016 – Conférence en Recherche d'Informations et Applications – 13th French Information Retrieval Conference. CIFED 2016 Colloque International Francophone sur l'Ecrit et le Document, Toulouse, France, March 9–11, 2016*, S. Calabretto, B. Coüasnon, L. Goeuriot, and S. Barrat, Eds. ARIA-GRCE, 2016, pp. 481–496.

[39] M. Boillet, M. Bonhomme, D. Stutzmann, and C. Kermorvant, "HORAE: an annotated dataset of books of hours," in *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2019, Sydney, NSW, Australia, September 20–21, 2019*. ACM, 2019, pp. 7–12.

[40] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[41] B. Kiessling, "A modular region and text line layout analysis system," in *17th International Conference on Frontiers in Handwriting Recognition (ICFHR 2020, Dortmund, Germany, September 7–10, 2020)*. IEEE Computer Society, 2020.