EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2024)
M. Corsini, D. Ferdani, A. Kuijper, and H. Kutlu (Editors)

Computational Tools for Dead Sea Scrolls

B. Kurar-Barakat and N. Dershowitz

Tel Aviv University

Abstract

The Dead Sea Scrolls, which include the oldest known biblical manuscripts, have been digitized by the Israel Antiquities Authority using high-resolution multispectral imaging. As part of the ongoing Scripta Qumranica Electronica project, we aim to develop computational tools for editing and analyzing these images. A core component of our work involves aligning transcriptions to images, which will automate the generation of script charts, suggest readings for damaged characters, create fonts for visualizing reconstructed text, and provide additional training data for improved character recognition. This will also facilitate paleographical analyses, automated suggestions for fragment joins, and fragment matching. To achieve these goals, we enhance OCR-based transcription alignment and tackle the challenges of segmenting Dead Sea Scroll images through the use of energy minimization techniques and multispectral image analysis. Full-resolution ink and parchment segmentation results are available online at https://www.cs.tau.ac.il/~berat/dss_ink_parchment_segmentation.html.

CCS Concepts

• ***Applied computing*** → ***Digital libraries and archives;*** • ***Computing methodologies*** → ***Image segmentation;***

1. Introduction

The discovery of the Dead Sea Scrolls (DSS) over 60 years ago is one of the greatest archaeological breakthroughs in modern history. These ancient texts, written or copied mainly between the 2nd century BCE and the 2nd century CE, provide the oldest known written record of the Hebrew Bible. To ensure their preservation and accessibility, the Israel Antiquities Authority (IAA) has digitized the fragments using multispectral high-resolution imaging (Figure 1).

Our work is part of the Scripta Qumranica Electronica (SQE) project, which was financed through the German-Israeli Project Coordination (DIP) of the German Research Foundation (DFG). SQE connects the Leon Levy Dead Sea Scrolls Digital Library of the IAA and the Qumran-Wörterbuch-Projekt (Qumran Dictionary Project, QWB) of the Göttingen Academy of Sciences and Humanities—which contains transcriptions of all Dead Sea Scrolls, alongside computational tools developed by our computer science team at Tel Aviv University (headed by Nachum Dershowitz and Lior Wolf). SQE will create an online platform for accessing and manipulating the texts and images and for preparing digital scholarly editions, which will be made available as part of the IAA's digital library.

Our team is working on automatically aligning the transcribed text of a fragment, letter by letter, from the QWB to its precise location on the IAA images. This will enable users to automatically generate script charts, suggest readings for damaged characters, and even create a font for visualizing reconstructed text. The successful alignment of transcriptions to images will add a textual layer to the IAA images, allowing scholars and laypersons alike to enter search terms and retrieve corresponding images. It will also provide additional training data for improved character recognition, paleographical analyses, automated suggestions for fragment joins, and fragment matching.

2. Methods and Results

We have already developed an OCR-based transcription alignment method [SBEBDD*20]. This method trains a recognition model on the known transcription-line pairs with kraken [Kie19], which returns character identifications and detections on the fragment images. However, the detection boxes returned by kraken are not always accurate; they are approximate and sometimes include partial characters or parts of other characters or binarization artifacts.

We use the energy minimization framework [BVZ01] to improve kraken's rough character detections (Figure 2). We assume that each character bounding box from kraken corresponds to one and only one character, but each character might consist of several connected components. The minimum of the energy function corresponds to an optimal pixel segmentation, which urges the assignment of component pixels to the label of the closest character detection box while striving to assign closer component pixels to the same label.

The refined segmentation has problems due to binarization artifacts. There are several challenges in binarizing the DSS images. Similar to many other historical manuscripts, the DSS collection truly suffers from document degradation problems. Due to aging,

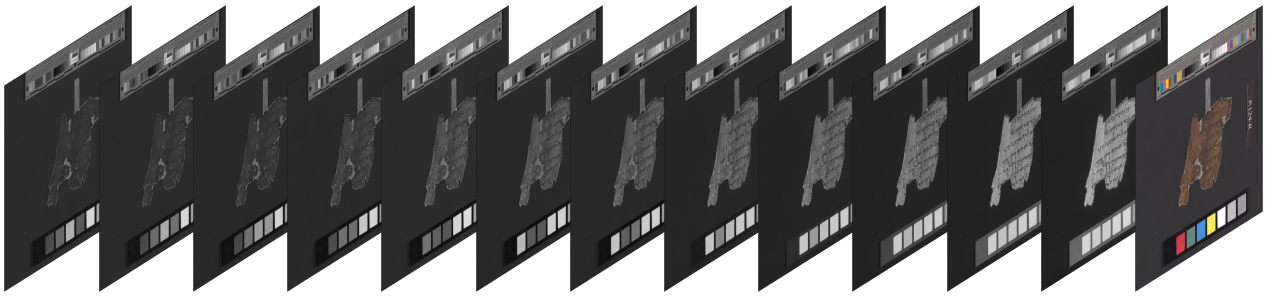

Figure 1: *Multispectral imaging of a DSS fragment at different wavelengths. The images represent bands 1 to 12 from left to right, followed by the color image on the far right. The first seven images are in the visible light spectrum, and the remaining five are in near-infrared.*

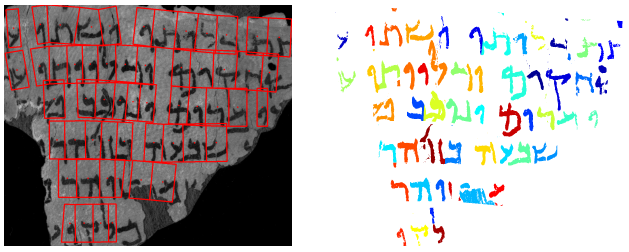

Figure 2: *Left: The red bounding boxes indicate kraken's rough character detections. Note that a bounding box may include pixels from adjacent characters or binarization artifacts, and may exclude some pixels of the detected character. Additionally parts of characters may be missing due to binarization issues. Right: Refined character segmentation using energy minimization guided by kraken's rough detections. The refined segmentation may still include pixels from binarization artifacts or miss parts of characters due to binarization issues.*

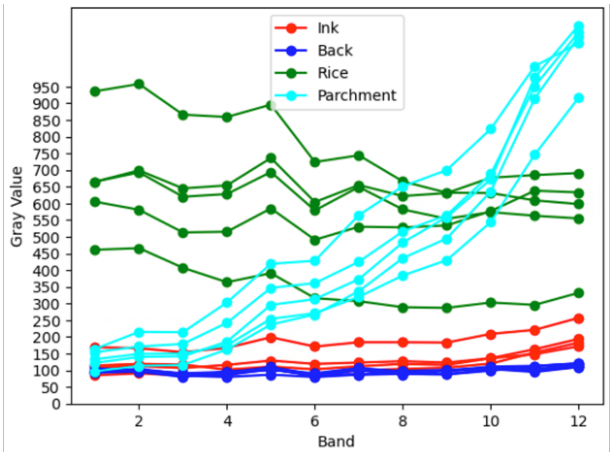

Figure 3: *Pixel intensity value trends across 12 bands of the multispectral images for different regions of a DSS fragment.*

the ink often show fading effects and parchment shows darkening effects [MSA20]. On top of all these, black was chosen as the background for the photographs because it does not reflect light, which could affect the image and reduce the system's stability. Hence the most severe issue is the low contrast between ink and background, shadows, holes, and halos [MSA17].

Otsu's method [Ots79] is one of the most commonly used. This unsupervised and non-parametric method automatically selects a global threshold based on the grayscale histogram of a given image with no prior information. The Otsu method performs well when the image has a uniform background; unfortunately, most historical manuscripts do not contain a uniform background. This can be handled using small local patches of the target image by local adaptive thresholding such as Sauvola [SSHP97]. It performs well compared to the global thresholding techniques but often shows poor performance in discriminating among multiple classes.

Maruf et al. [DdWS19] propose BiNet, an end-to-end binarization approach for Dead Sea Scrolls fragments based on the U-Net architecture [RFB15]. This method aims to segment out only the ink, defining the ink as the foreground class and all other materials as the background class. While their approach significantly outperforms unsupervised methods, a major challenge is that the labeling task requires approximately 4–8 hours of a paleographic expert's time per fragment.

Due to the high cost of labeling, we employed multispectral imaging provided by the IAA. Each fragment was photographed in 12 different wavelengths, seven in the visible light spectrum and five in near-infrared. We first analyze the pixel value trends of different regions across the 12 bands (Figure 3). Based on these trends, we use the 12th and 1st band values and their differences to threshold and identify approximate regions for parchment, background, rice paper, holes, halos, shadows, and ink contours. We then apply energy minimization to refine the ink contours by stretching noisy ink contours between parchment and all other regions: rice paper, holes, halos, shadows, and background. Subsequently, we use energy minimization to determine the ink regions by stretching the inverse parchment between clean ink contours and parchment. Finally, we combine the ink regions and parchment regions to achieve parchment segmentation (Figure 4).

Our method is complementary, exhaustively segmenting each of the 12 million pixels in an IAA image and assigning each pixel to one class only. It does not rely on morphological or edge detection

Otsu Sauvola MSEM

Figure 4: *Comparison of segmentation methods on a DSS fragment. The result of multispectral thresholding with energy minimization (MSEM) effectively segments not only the ink but also the parchment from all other elements, including background, holes, halos, shadows, and rice paper.*

operations and does not make assumptions about component size. Thus, the segmentation of a part of an image is consistent with its segmentation as part of the entire image.

3. Future Work

We will use the ink and parchment segmentation to convert kraken's [Kie19] rough character detections into precise character segmentations. We will then utilize the segmented and identified characters, along with the segmented parchments, to enable further analysis of DSS images. This includes automatically generating script charts, suggesting readings for damaged characters, and creating fonts for visualizing reconstructed text. The successful automated alignment of transcriptions to images will add a textual layer to the IAA images, allowing scholars to enter search terms and retrieve corresponding spatial locations on the images. This will also provide additional training data for improved character recognition, paleographical analyses, and automated suggestions for fragment joins and fragment matching.

References

[BVZ01] BOYKOV Y., VEKSLER O., ZABIH R.: Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 11 (2001), 1222–1239. 1

[DdWS19] DHALI M. A., DE WIT J. W., SCHOMAKER L.: Binet: Degraded-manuscript binarization in diverse document textures and layouts using deep encoder-decoder networks. *arXiv preprint arXiv:1911.07930* (2019). 2

[Kie19] KIESSLING B.: Kraken–an universal text recognizer for the humanities. In *ADHO, Éd., Actes de Digital Humanities Conference* (2019). 1, 3

[MSA17] MAOR Y., SHOR P., AIZENSHTAT Z.: White halos surrounding the Dead Sea scrolls. *Journal of Cultural Heritage 28* (2017), 90–98. 2

[MSA20] MAOR Y., SHOR P., AIZENSHTAT Z.: Parchment browning and the Dead Sea Scrolls – part i: Artificial aging. *Polymer Degradation and Stability 176* (2020), 109109. 2

[Ots79] OTSU N.: A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics 9*, 1 (1979), 62–66. doi:10.1109/TSMC.1979.4310076. 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, part III 18* (2015), Springer, pp. 234–241. 2

[SBEBDD*20] STÖKL BEN EZRA D., BROWN-DEVOST B., DERSHOWITZ N., PECHORIN A., KIESSLING B.: Transcription alignment for highly fragmentary historical manuscripts: The Dead Sea scrolls. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (2020), IEEE, pp. 361–366. 1

[SSHP97] SAUVOLA J., SEPPANEN T., HAAPAKOSKI S., PIETIKAINEN M.: Adaptive document binarization. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition* (1997), vol. 1, IEEE, pp. 147–152. 2