

Prediction of Paleographical Features in Ashkenazi Square Script for Identifying Subclusters Within the Style

Daria Vasyutinsky Shapira

Nachum Dershowitz

03.05.2024

Tel Aviv University

Abstract

This paper presents a model for the paleographical features of document images written in Ashkenazi square script, structured hierarchically with multi-labels and mutually exclusive sub-labels. We train a convolutional neural network (CNN) to predict these labels and use the predictions as paleographical feature vectors for each document image (Figure 1). These vectors serve as the basis for clustering the document images, thereby unveiling hidden subgroups within the Ashkenazi square script style.

The study employs a curated dataset from 55 manuscripts, each contributing four pages. These manuscripts lack specific annotations about date or region but include distinct groups believed to be from France and Germany, as well as unique manuscripts from England prior to 1290. This approach addresses the limitations of conventional algorithms, such as the bag of words method, which was originally developed for natural scene objects that possess a consistent number of vital features—a condition not applicable to the paleographical features of document images. These traditional methods have been relatively ineffective in identifying the critical features necessary for paleographical clustering. By employing a deterministic methodology, as guided by expert paleographers, and executing a brute-force search to optimize cluster formations, this study not only identifies subclusters but also enhances the overall understanding of sub-clustering, with potential applications to other Medieval Hebrew script types like Byzantine and Yemenite. This systematic method addresses the challenge paleographers face

