Towards binarization of Iron Age ostraca from multispectral weakly-annotated imaging

Ohr Dallal

Abstract:

Image binarization is one of the essential and preliminary steps towards many document processing tasks. We aim to binarize Iron Age Hebrew ostraca, which are of great importance to the historical study of ancient Israel and Judah. To this end, a new and unique dataset is introduced, consisting of multispectral ostraca images taken at different camera wavelengths. The data poses severe challenges such as faded text, degradations, contrast variation and characters rarity. The task is further complicated by sparse and partial labels, where only a fraction of image pixels is manually annotated by human experts. The underlined assumption of this research is that the multispectral signature of the ink differs from the one of the clay, due to material difference. We propose to apply end-to-end deep neural networks (DNN) for exploiting complex shape and spectral cues available in the data, which allow for better discrimination of ink and background and possibly even the reconstruction of ink invisible to the naked-eye. We develop and test DNN binarization models, based on segmentation feedforward convolutional neural networks, alongside methods for pre-processing and enriching the data representation through weakly-supervised learning, semi-supervised learning and visual augmentations. These techniques are employed to deal with the available data sparsity and improve generalization performance. We show the contribution of the multispectral nature of the data and demonstrate the effectiveness of our method on Arad ostraca unearthed in the Judahite desert dated ca. 600 BCE.