Character Recognition in Ancient Greek Papyrus

Gwenael Gendre

Master thesis in Computer Science

The conservation of historical documents is very important, digitization helps to maintain them. Due to their possible degradation, handwriting recognition and thus complete transcription become difficult. One way to tackle this problem is Keyword Spotting, where all instances of a query keyword are retrieved. Graphs are a possible representation of handwriting and its variations and can therefore be used for Keyword Spotting.

This work focuses on a collection of ancient greek papyri, without word or character segmentation: we introduce a new graph representation, Contour graphs, and we successfully adapt a graph-based Keyword Spotting framework for segmentation-free use. We first test it on a well-studied document, the George Washington letters, and then on the papyri.

The experiences highlight two difficulties tied to segmentation-free graph matching on papyri: the complexity and needed time for keyword spotting greatly increase, not allowing us to fully utilize this framework. We also show that papyri are a really challenging document type, however a well-crafted binarization can almost double the obtained results.

Prof. Rolf Ingold