Deep learning meets the old times

December 12, 2017

Outline

1. Extracting Fragment's Physical Attributes
   - Preprocessing
   - Physical measurements

2. An Automatic System for Discovering Joins

Outline

1. Extracting Fragment's Physical Attributes
   - Preprocessing
   - Physical measurements

2. An Automatic System for Discovering Joins

Computing image dpi

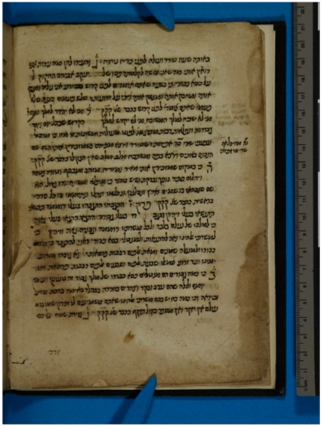
- A reference image of the ruler was photographed independently
- The identification is done by employing a randomized algorithm: RANSAC (Fischler and Bolles, 1981), combines with scale-invariant feature transform (SIFT) keypoint matching (Lowe, 2004)
- For the latter task, a classical method such as the binarization method of (Sauvola and Pietikäinen, 2000) that also contains a method for identifying textual regions could be used

Separating foreground from background

The goal of this step is to detect the image of the fragment itself, isolating it from the accompanying background

- an automatic classifier was first applied to identify foreground pixels (in contrast to background ones Figure 1) based on RGB color values (or HSV values)
- create a region-based segmentation of the fragment(s), the connected components of the detected foreground pixels were marked
- the convex hull of each component calculated (connected component = a contiguous region of foreground pixels; convex hull = the smallest possible encompassing polygon with angles opening inward)

(a)



(b)

Figure 1: (a) Fragment from the Strasbourg collection with label, clip and weight bags vs. (b) one from the British Library using a contrasting color

Detecting and removing irrelevant components

- In some collections, The system detected the binders (Figure 2) by the combination of their color and shape, and removed them from the image
- In other collections, images included a label (Figure 1a) with the fragment's shelfmark. These labels were also detected by the system and ignored

Figure 2: Fragment from the JTS collection with a black binder.

Separating multi-fragment images into components

- In many cases, more than one fragment was captured in a single image (Figure 3)
- Each fragment (a "component" of the image) was identified and given a unique identifier (serial number) and handled independently.
- However, there was a need to relate the components in the recto image of a fragment to the ones in its verso image (so as to have the same identifiers for both images)
- This was done automatically by mirroring one image and matching the components in both images by size and shape

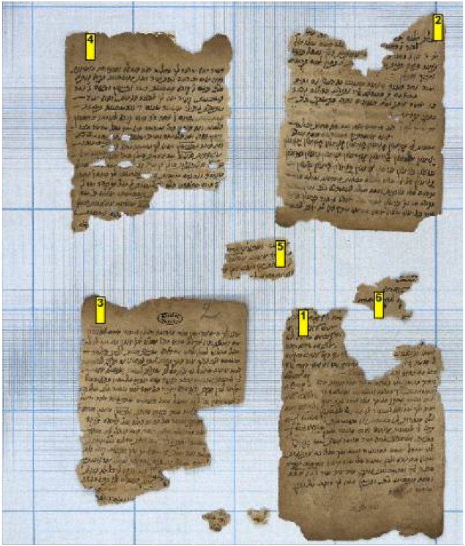


Figure 3: Multiple fragments in one image.

Binarization

- The regions detected are then binarized, that is, every ink pixel is assigned a value of 1(representing black), and all other pixels are assigned a value of 0 (for white)
- This is done using the auto-binarization tool of the ImageXpress 9.0 package by Accusoft Pegasus
- ??

Auto-alignment

- Most cases fragments were imaged placed upright, in many other cases the fragment was tilted.
- The need for alignment is two-fold:
  - first, to enable the correct measurement of the fragment's various attributes
  - second, to enable proper application of the handwriting-matching algorithm
- Alignment is achieved by rotating the image until the lines of text are horizontal, using a simple method akin to those in (Baird, 1992, Srihari and Govindaraju, 1989)