Efficient OCR Training Data Generation with Aletheia

Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos

Challenge: Methods for Optical Character Recognition (OCR) often need to be trained for new fonts or symbols. Training data can be either synthesised or extracted from document images with given text. Especially in the case of historical documents a synthesis is usually not feasible because font descriptions are not available.

Extracting training data of sufficient quality and quantity is cumbersome. It requires a precise representation of shape and class (character code) for a large amount of glyphs.

Aletheia is a tool for creating page layout and text content ground truth and for viewing document analysis results. We demonstrate the use of Aletheia to generate training data for the Gamera open source OCR toolkit. The same principle can be applied to train other OCR engines as well but may require conversion to the corresponding training data formats.

Ground truth is stored in the PAGE XML format wherein text objects are represented by (arbitrary) polygons and Unicode text content.

Regions

Text Lines

Words

Glyphs

Export to Gamera is achieved by transforming layout, text content, and the corresponding document image to a valid Gamera training data description.

Glyph shapes are translated from polygons to run-length encoding by scanning the pixel data of the area inside a polygon. Character classes are represented hierarchically in Gamera using a dot-separated name pattern, which usually corresponds to the respective textual Unicode descriptions (look-up table required).

Text input · Special characters/ligatures · Customisable virtual keyboard · Text overlay

Text propagation

Gamera · Successful application of the trained classifier

This work has been supported in part through the EU 7th Framework Programme grant Europeana Newspapers (Ref: 297380).