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 Especially in the case of historical documents a synthesis is usually not feasible because font descriptions are not available.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.

Layout ground truth

can be created efficiently with semi-automatic tools such as:

- Full pre-production using Tesseract (an open-source OCR engine)
- Tools that "shrink-wrap" around a selection
- Easy merging and splitting of layout objects (text lines and words can usually be separated with one mouse click)

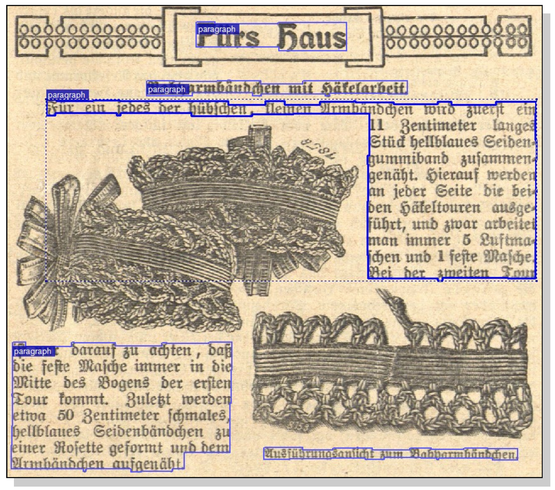
The degree of manual intervention required depends on the quality of the document image (noise, scanning artefacts, etc.).

Text content

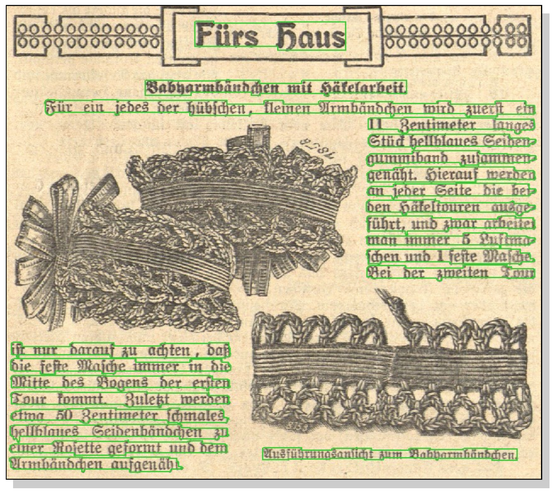
can be entered conveniently at region level and is then propagated automatically to glyph level by matching the text elements with the corresponding layout objects.

The matching only requires a consistent handling of white spaces, punctuations, and ligatures. If, for example, a punctuation character is not separated from the adjacent word by a space, the corresponding glyph object should also be part of the respective word object. Aletheia highlights segmentation inconsistencies to speed up their correction.

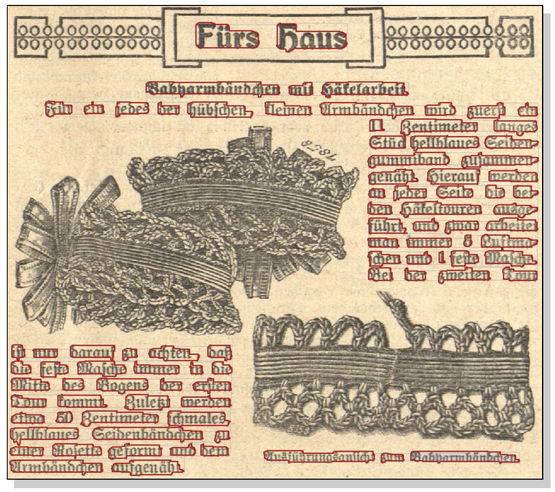
Regions



Text Lines



Words

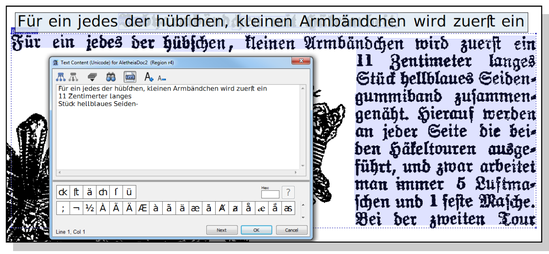


Glyphs

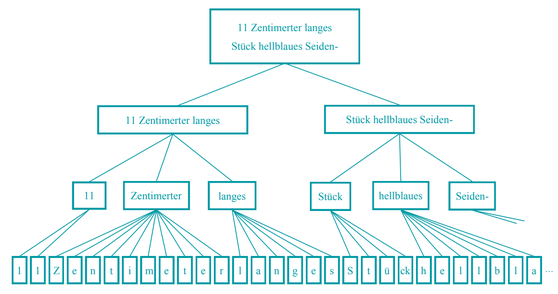
Export

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

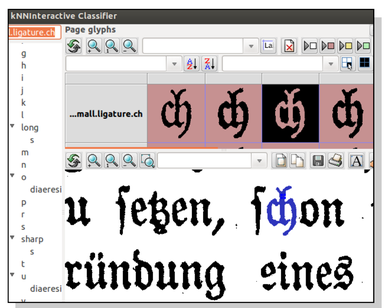
Glyph shapes are translated from polygons to run-length encoding by scanning the pixel data of the area inside a polygon. Character classes are represented hierarchically in Gamera using a dot-separated name pattern, which usually corresponds to the respective textual Unicode descriptions (look-up table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



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



Text propagation



Gamera · Successful application of the trained classifier

University of Salford MANCHESTER PRImA Research Lab

Pattern Recognition and Image Analysis Research Lab, School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom, www.primaresearch.org

www.primaresearch.org/tools/Aletheia

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

europeana newspapers