Efficient OCR Training Data Generation with Aletheia*

Christian Clausner, Stefan Pletschacher and Apostolos Antonacopoulos
PRImA Lab, School of Computing, Science and Engineering, University of Salford,
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

***Abstract*—We present how the ground-truthing tool Aletheia can be used to efficiently create training data for an open-source text recognition engine. The labelling process is sped up considerably through a top-down approach. Text content is thereby entered on region level. The characters are then propagated automatically to glyph objects. In addition, segmentation is simplified by several semi-automated tools.**

Keywords—Optical character reconition; training data generation; ground truthing

I. Introduction

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 Especially in the case of historical documents a synthesis is usually not feasible because font descriptions are not available.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

In this paper we demonstrate the use of the ground-truthing tool Aletheia [1] to generate training data for the Gamera open source OCR toolkit [2]. The same principle can be applied to train other OCR engines as well but may require conversion to the corresponding training data formats.

II. Training Data Generation

A. Ground Truth Creation

Aletheia is an advanced tool for creating page layout and text ground truth for document images. It supports top-down (from regions to glyphs) as well as bottom-up (from glyphs to regions) workflows.

Ground truth is stored in the PAGE XML format [3] wherein text objects are represented by (arbitrary) polygons and Unicode text content. Four levels of objects are available: Regions (blocks), text lines, words, and glyphs.

Semi-automated tools help to reduce manual labour and thereby increase the efficiency. Figure 1 shows the final stage of the top-down approach where word objects are split into glyph objects. In most cases this requires just one click in the gap between each pair of adjacent glyphs. The outlines are then automatically calculated based on the pixel information.

Wrongly generated object outlines can be corrected by directly manipulating the polygons. The degree of manual intervention required depends on the quality of the document image (noise, scanning artefacts, etc.).

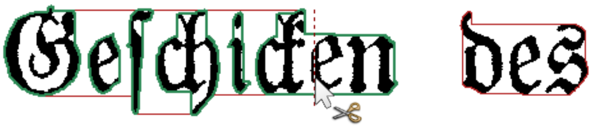


Fig. 1. Splitting word objects into glyph objects in Aletheia (top-down approach).

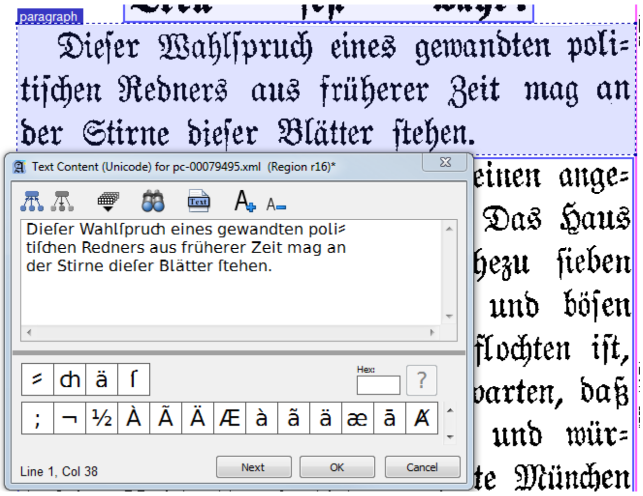
Under certain circumstances (for instance heavily degraded images), the bottom-up workflow is favourable. Glyphs are marked and subsequently grouped into words, text lines, and regions.

Text region, text line, and word objects are usually not necessary for OCR training data generation. They can, however, significantly speed up the task of assigning the character classes to glyph objects. Text can be entered conveniently at region level and can then be propagated down to glyph level. Aletheia automatically matches layout objects with their corresponding text content. Figure 2 shows an example for text entry.

The matching requires a consistent handling of white spaces 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. Similar considerations are required for ligatures. They can either be represented by one single pre-composed character or using a decomposed sequence of characters. In both cases, the glyph segmentation has to match the number of characters given in the transcribed text. Aletheia highlights segmentation inconsistencies to speed up their correction.

A virtual keyboard simplifies the input of special characters and symbols. The selection and layout of the keys is fully customisable.

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



paragraph
Text Content (Unicode) for pc-00079495.xml (Region r16)*
Hex:
?
Line 1, Col 38
Next
OK
Cancel
⸗ ch ä ſ
; ¬ ½ À Ã Ä Æ à ã ä æ ā Ⱥ

Dieſer Wahlſpruch eines gewandten poli=
tiſchen Redners aus früherer Zeit mag an
der Stirne dieſer Blätter ſtehen.

einen ange=
Das Haus
hezu ſieben
und böſen
flochten iſt,
parten, daß
und wür=
te München

Dieſer Wahlſpruch eines gewandten poli⸗
tiſchen Redners aus früherer Zeit mag an
der Stirne dieſer Blätter ſtehen.

Fig. 2. Entering text content in Aletheia.

B. *Conversion to Gamera Format*

Generated ground truth needs to be converted to an XML format specific to Gamera [4]. For this purpose, a tool has been developed that transforms a PAGE XML file (the output of Aletheia) and the corresponding document image to a valid Gamera training data description.

Glyph shapes need to be translated from polygons to run-length encoding. This is done by scanning the image pixel data of the area inside a polygon. A bi-level image is therefore mandatory. Aletheia provides a basic set of binarisation and noise removal methods to this end.

Character classes are represented hierarchically in Gamera using a dot-separated name pattern, which usually corresponds to the respective textual Unicode descriptions (see Table I for examples). A look-up table has to be defined, containing all characters that may occur in the documents that are to be processed.

TABLE I. EXAMPLES OF CHARACTER CLASSES IN PAGE XML AND GAMERA XML

PAGE (Unicode)	Gamera (generic)
004E	latin.capital.letter.n
002C	latin.punctuation.comma
0030	digit.zero

Once converted, the training data can be applied to a Gamera classifier. This process is not limited to one file. Data from multiple pages can be added incrementally.

III. EXPERIMENTAL VALIDATION

For proof of concept a document page with about 3000 glyphs has been processed. On average it took 2.1 seconds to mark and label one glyph. The result has then been applied to train a classifier using the Gamera interactive classifier tool. The application of this classifier to a document image that was not part of the training is shown in Figure 3.

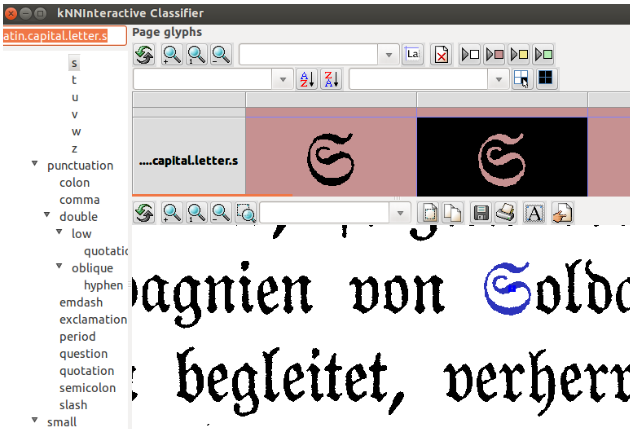


Fig. 3. Classification in Gamera after training with data that has been extracted using Aletheia.

IV. CONCLUSION AND FUTURE WORK

It has been presented how the Aletheia ground-truthing tool can be used for efficient generation of OCR training data. Semi-automated tools in combination with a mature user interface can speed up the extraction process in comparison to other tools (for instance the Gamera interactive classifier tool).

Future work will include the investigation and development of the conversion to formats required for training further OCR engines such as Tesseract.

REFERENCES

[1] C. Clausner, S. Pletschacher, A. Antonacopoulos, “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments”, Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, pp. 48-52, September 2011

[2] M. Droettboom, K. MacMillan, I. Fujinaga, “The Gamera framework for building custom recognition systems”, Symposium on Document Image Understanding Technologies, pp. 275-286, 2003.

[3] S. Pletschacher, A. Antonacopoulos, “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework”, Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), IEEE-CS Press, pp. 257-260, Istanbul, Turkey, August 23-26, 2010.

[4] Documention for Gamera XML format version 2.0, http://gamera.sourceforge.net/doc/html/xml_format.html (date accessed: 18/12/2013)