Aletheia - An Advanced Document Layout and Text Ground-Truthing System

PRImA Research Lab, University of Salford, United Kingdom
Overview

- Ground truth is the basis for any performance analysis workflow
- Aletheia:
  - Image Operations
  - Border / Print Space
  - Layout Regions
  - Unicode text content
  - Reading Order / Layers
  - Dewarping
  - Validation
  - Evaluation
  - XML

Diagram:
- Document Images
- Ground Truth Production
- Performance Evaluation
  - Rating / Decision
- Layout Analysis, OCR
  - A
  - B
- Evaluation
- XML
Ground Truth Production

• Accurate ground truth is crucial for evaluation
• Aletheia targets production environments (large scale digitisation)
• Goals: Efficiency, accuracy, ease of use, flexibility and robustness
Image Operations

• Binarisation
  – Threshold (variable)
  – Otsu
  – Sauvola

• Noise removal
  – Pepper noise (variable)
  – Selected components
Border and Print Space

- **Border**: Single polygon marking the edge of a scanned document.
- **Print Space**: Polygon marking the main text body of a document without page numbers, marginalia, etc.
Layout Regions

- Polygonal shape
- 11 types: Text, image, table, separator, ...
- Sub-types:
  - Text: paragraph, heading, page number, ...
- Other attributes (e.g. text content, language, orientation)
Creating Regions

- Manual tools (rectangle, arbitrary and isothetic polygons)
- Semi-automated tools (shrinkers)
  - Based on bounding boxes of connected components
  - Smearing based
Modifying Regions

- Correction of pre-produced data
- Add, move and delete polygon points
- Split and merge regions (attributes are preserved)
Region Attributes

• Additional properties for each type of region
• Can be modified for multiple regions at once (e.g. setting the language for all text regions)
Text Content

- Unicode
- Special font (support of characters that are not (yet) part of the Unicode standard)
- Virtual keyboard (customisable)
- Text search
Text Overlay

- Interactive text overlay on the document image for quality assurance
Reading Order

- Groups with ordered or unordered relations between regions
- Nested groups
- Drag & drop
Layers

- Layers are an additional level of abstraction to group regions by Z-order, allowing the definition of overlapping regions
- Drag & drop
Text Lines

• Marking text lines:
  – By splitting regions (one click split)
  – By shrinking around selected connected components
  – By merging line fragments
  – By combining words
  – By manually drawing
Words and Glyphs

- Similar to marking text lines:
  - Split tools
  - Merge tool
  - Shrinkers
  - Manual drawing

- Text can be propagated to all levels (e.g. from regions to glyphs)
Dewarping

- Grid-based method for geometric correction of document images
- Create and save dewarping ground truth as well as load existing dewarping data from XML files.
Validation

• Ground truthing rules and guidelines validator
• 3 message levels (error, warning, info)
• Easy identification of regions that cause problems (highlighted in document view)
Performance Evaluation

- Measure quality of segmentation methods
- In-depth
- Scenario-based
XML Format

• Mature XML schema which is part of the PAGE (Page Analysis and Ground truth Elements) format framework

• Backward compatibility (all tools can handle older versions of the XML schema)

 PAGE structure consisting of a root instance linking to task-specific output or ground truth files. Note that gts XML instances may link to further resources (e.g. a dewarped image) depending on the nature of the respective method.
Page Collections

• Can represent all pages of a printed document (e.g. a book) or a loose collection of unrelated pages

• Stored as METS XML file, linking to the individual page images and content PAGE XML files
Further Usage

• Aletheia:
  • Ground truth production system
  • Result viewer (layout analysis results)
  • Layout and text editor (e.g. showcase production and training data generation)
University of Salford PRImA Research Lab

www.primaresearch.org