Document Layout Evaluation

Command Line Tool

Created: 27/07/2010
Last Change: 10/05/2016

www.primaresearch.org

Table of Contents

Introduction

Layout evaluation is used to benchmark results of layout segmentation methods and uncover specific problems of the algorithms to help developers to improve them.

The command line tool takes the ground truth and segmentation result of one document as input and produces an evaluation result file (.evx). To distinguish between foreground and background, the black-and-white document image is needed as well. The result file contains the evaluation raw data (error values, ...), some statistics (number of regions, ...) and several numbers and figures indicating the quality of the segmentation. For more information on the output format see the document 'Layout Evaluation XML Schema'.
The evaluation process can be adjusted by using different profiles. A profile specifies which region types to highlight and which to ignore (weights) and some general settings. Several predefined profiles for common evaluation scenarios are provided together with the tool.

The evaluation tool also includes an option to output the most important results as comma separated values. That way, multiple evaluation results can be collected in a CSV file and can then be easily imported into a spreadsheet.

Command Line Arguments

```
layoutevalcmd <argument1> <value1> <arg2> <val2> ... <opt1> <opt2> ...

Arguments and options:

 -gt <file path>  - Ground truth input file in PAGE XML format or compatible
 -res <file path> - Page analysis / OCR result in PAGE XML format or compatible
 -img <file path> - Document image (.tif, .png, .jpg).
           Use black-and-white (bitonal) image if available. Colour
           or grey-scale images will be binarised internally (Otsu method).
           For border and baseline evaluation the image is optional.
 -profile <file path> - File containing the evaluation profile (.evx format).
                        Specifies weights and settings for the evaluation.
 -evalRes <file path> - Full path for the evaluation result output file (.evx).
                        (optional)
 -eval <L1[,L2,...]> - Specifies at which levels to evaluate (comma-separated).
                       (optional, if none is specified, all available levels
                       will be used)
                Levels:
                  regions   - Layout regions (blocks / zones)
                  textLines - Text line objects
                  words     - Word objects
                  glyphs    - Glyph objects
                  border    - Page border
                  readingOrder - Reading order of text regions
                  baselines    - Baselines of text line objects
 -overwrite - Option to overwrite existing files
```

```
-printWarnings - Option to print warnings and messages
-csvValues - Option to output the evaluation results as comma-separated values (CSV)
-csvHeaders - Option to output the column headers for the comma-separated values
-csvGlyphStats - Option to output alternative CSV data with glyph statistics (matches,
                 mismatches, misses). Only applicable in conjunction with '-eval glyphs'
-seqReadingOrder - Option to use a sequential reading order if no order defined
```

Recommended System Configuration

Operating System:	Windows 7 – 64bit (or higher)
CPU:	3.0 GHz dual Core
RAM:	8 GB
Hard disk space:	100 GB

Evaluation Profile

The evaluation profile is used to specify weights and other parameters to customize the layout evaluation. In some scenarios text regions may be important, in others it may be image regions. The weights can be adjusted from a general level (e.g. merge) down to the most detailed level (e.g. merge of text paragraph with text headline).
A weight can have values from 0.0 to 10.0, whereas 1.0 is the default. A value of 0.0 means that the region or error type is not regarded at all for the evaluation results. A value higher than 1.0 means that the region or error type is emphasized in comparison to other region or error types.
The profile is stored together with the evaluation results. So the weights that were used for the evaluation can always be examined.
The profile will usually be defined in the Graphical Layout Evaluation Tool and can then be referenced in the command line tool.

The evaluation profile contains following weights and parameters:

- error type weights stored hierarchical per error type (merge, split, ...), region type (text, image, ...) and for text regions also subtype (paragraph, headline, ...). The weights for merge and split have an extra value for 'allowable' (see chapter on 'Allowable Merges and Splits').
- region type weights stored per region type (text, image, ...)
- reading order weight (influence of the reading order to the overall success rate)
- general parameters
    - 'Use Foreground Area' – if TRUE, the number of black pixels is used for the error calculations instead of the polygon area
    - 'Reading Orientation Threshold' – Threshold of how much the reading orientation of two regions can differ to be allowable (see chapter on 'Allowable Merges')
    - 'Default Reading Direction' – Used for 'Allowable Merge' detection. See next parameter

- ‘Reading Direction Usage’ – Can be either ‘Ground-truth’ – always use reading direction as specified in the ground-truth; ‘Default if not set in Ground-Truth’ – uses the default value if the reading direction isn’t defined for the regarded region; ‘Default’ – always uses the default value, regardless which value is defined for the regarded region
- ‘Default Reading Orientation’ - Used for ‘Allowable Merge’ detection. See explanation above
- ‘Reading Orientation Usage’ - Used for ‘Allowable Merge’ detection. See explanation above
- ‘Default text type’ – Text region type that is to be used if not defined (use <undefined> to not use a default text type)
- ‘Ignore embedded text misclassification’ – if TRUE, misclassification is not penalised if a chart, image, graphic, line drawing, or table region was detected as text region and the ‘embedded text’ attribute is set to TRUE in the ground truth region

Several profile presets are delivered with the command line tool (\data\profiles):

General Document Recognition

All region and error types are regarded. Allowable merges and splits are considered less problematic.

All non-allowable merge weights = 1.5
All other non-allowable weights = 1.0
All allowable weights = 0.5
Miss and partial miss = 2.0

Exceptions:
Split of noise = 0.5
Merge noise – noise = 0.5

General Document Recognition (strict)

All region and error types are regarded. There is no difference between allowable and non-allowable merges and splits.

All weights = 1.0
Merge = 1.5
Miss and partial miss = 2.0

Exceptions:
Split of noise = 0.5
Merge noise – noise = 0.5

Images, Graphics and Charts

Only image, graphic, chart and line drawing regions are regarded.

All region type weights except image, graphics, chart and line drawing = 0.0
Merge = 1.5
Miss and partial miss = 2.0
Reading Order weight = 0

Full Text Recognition

For applications with full text recognition. Only text regions are evaluated. Allowable merges and splits are considered less problematic.

All region type weights except text = 0.0
All non-allowable merge weights = 1.5
All other non-allowable weights = 1.0
All allowable weights = 0.5

Miss and partial miss = 2.0

Keyword Search

For applications targeting keyword search. Only text regions are evaluated. Merges, splits and the reading order are not of interest.

All region type weights except text = 0.0
All merge weights = 0 (not critical, keywords will still be found)
All split weights = 0.5 (can be critical, if words are split)
Misclassification text to text = 0
Miss and partial miss = 2.0
Reading Order weight = 0

Document Structure

For document structure retrieval (table of contents, ...). Only structural text regions are of interest (heading, caption and page number).

All region type weights except heading, page number and caption = 0.0
Merge = 1.5
Miss and partial miss = 2.0

Plain

All weights = 1.0

Evaluation Statistics and Error Metrics

Based on the raw data, the actual error values and success rates are calculated. The metrics are calculated for a specified structure level (region, text line, word or glyph).
Following figures are produced:

- Statistics
  - Image Area
  - Number of black pixels within the image
  - Overall number of regions (ground-truth and segmentation result)
  - Number of regions per type (text, image, ...) (ground-truth and segmentation result)
  - Overall region area (ground-truth and segmentation result)

    - Overall number of black pixels in regions (ground-truth and segmentation result)
- Performance Indicators
    - Overall weighted area error per error type (merge, split, ...)
    - Overall weighted area error per region type (text, image, ...)
    - Overall weighted area error
    - Weighted area success rate per error type
    - Overall weighted count error per error type (merge, split, ...)
    - Overall weighted count error per region type (text, image, ...)
    - Overall weighted count error
    - Weighted count success rate per error type
    - Reading order error
    - Reading order success rate
    - Overall weighted area success rate (arithmetic mean, harmonic mean)
    - Overall weighted count success rate (arithmetic mean, harmonic mean)
    - Recall per type
    - Recall (strict / non-strict)
    - Precision per type
    - Precision (strict / non-strict)
    - F-Measure (strict / non-strict)
    - Region count deviation (absolute / relative)

The same figures are also calculated for each region type (text, image, ...) separately. Only merges are a small exception here. If we look for the merge errors of graphics, we also take into account merges of graphic regions with other region types (e.g. separator).

The following paragraphs describe the different values in detail.

Image Area

Image Width * Image Height

Number of Black Pixel within the Image

Number of black pixels within the black-and-white image.

Overall Number of Regions

Number of regions of the chosen type level (block, text line, word or glyph) within the document layout.

Number of Regions per Type

Number of regions of for each region type (text, image, ...) within the document layout. This value is only available in block level and not in text line, word or glyph level.

Overall Region Area

The combined area of all regions regarded for the 'Overall Number of Regions' value. Possible region overlaps are not left out. So if there are overlaps, some image parts are counted twice.

Overall Number of Black Pixels in Regions

The combined count of black pixels of all regions regarded for the 'Overall Number of Regions' value. Possible region overlaps are not left out. So if there are overlaps, some image parts are counted twice.

Weighted Errors

For the weighted area and count the weights defined in the evaluation profile are being used. There are two types of weighted errors: the 'Weighted Area' and the 'Weighted Count'. The weighted area is based on the assumption that bigger regions are more important than smaller ones. The error value is the region area (or the number of black pixels) multiplied with the weight. The weighted count only takes into account the error quantity. A misclassified region for instance is counted as one. A ground-truth region split into three regions is counted as 3. The count is then also multiplied with the weight.

Overall Success Rates

The overall success rates combine all error type success rates to one number. There are two different types of overall success rates: the arithmetic mean and the harmonic mean. And for each type there are again two success rates: One including the weighted area success rates and the reading order and the other one with the weighted count success rates and the reading order.

The general formula for the weighted arithmetic mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}.$$

The general formula for the weighted harmonic mean is:

$$\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Where n is the number of values, x_i are the values (in our case the success rates for the error types) and w_i are the weights. For the reading order the weight is directly the one defined in the evaluation profile. The other weights are defined by:

$$w_i = (5 * (1-x_i) + 1) / 6$$

This highlights low error type success rates and diminishes high success rates, without erasing them completely. The influence of a partial success rate to the overall rate is somewhere between 1/6 and 1.

Note: If one of the error type success rates is zero, the harmonic mean is also zero.
The harmonic mean is always smaller than or equal to the arithmetic mean.

Precision and Recall

Precision and Recall are generally defined as follows:

Precision is the number of relevant documents retrieved by a search divided by the total number of relevant documents.

Recall is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

In terms of document image evaluation this can be interpreted as follows (example for text regions):

Recall is the number of pixels within ground-truth text regions that are also within a text-region in the segmentation result divided by the overall number of pixels in ground-truth text regions.

Precision is the number of pixels within ground-truth text regions that are also within a text-region in the segmentation result divided by the overall number of pixels in segmentation result text regions.

For the overall recall and precision we differentiate between strict and non-strict. For the strict recall and precision the region type must be matched correctly. That means a ground-truth text region overlapped by a segmentation result image region does not count as recall. For non-strict recall and precision the region type doesn't matter. You could also say that strict means 'with classification' and non-strict means 'without classification'.

F-Measure

The F-Measure combines precision and recall to one quality measure. It is defined by:

F-Measure = 2 * precision * recall / (precision + recall)

Region Count Deviation

The region count deviation is simply the difference between the number of ground-truth regions and the number of segmentation result regions.

absolute region count deviation = |#ground-truth regions - #segmentation result regions|

relative region count deviation = absolute deviation / #ground-truth regions

Note: If there are no ground-truth regions, the relative value is the same as the absolute value.

Glyph Statistics for Character Mismatches

The command line options "GM" enable an alternative CSV output (a table) for OCR-related statistics. The table has following columns:

- Ground truth file: Filename of the ground truth PAGE XML file
- OCR result file: Filename of the OCR result PAGE XML file
- Ground truth character: A specific character that was found in the ground truth
- Ground truth Unicode: The ground truth character in Unicode hexadecimal notation (e.g. 0x00A3)
- OCR result character: Corresponding OCR result character (as it was recognised by the OCR – correctly or incorrectly)
- OCR result Unicode: The OCR result character in Unicode hexadecimal notation
- OCR error: One of 'None' (no error – correct OCR result), 'Miss' (the glyph was left out by the OCR), 'Mismatch' (the character was misrecognised by the OCR)
- Count: Number of occurrences of the same condition (for example "B, 0x0042, 8, 0x0025, Mismatch, **5**" means character B was misrecognised **five** times by the OCR as an 8)

Example:

Ground truth file	OCR result file	Ground truth character	Ground truth Unicode	OCR result character	OCR result Unicode	OCR error	Count
…	…	,	0x002c	,	0x002c	None	0
		,	0x002c			Miss	0
		,	0x002c	*	0x002a	Mismatch	3
		,	0x002c	»	0x00bb	Mismatch	2
		-	0x002d	-	0x002d	None	30
		-	0x002d			Miss	104
		-	0x002d	"	0x0022	Mismatch	1
		-	0x002d	*	0x002a	Mismatch	2

Statistics for the point of view from the OCR result (including false detections that are not contained in the ground truth) can be obtained by reversing the order in which the two PAGE XML files are passed to the command line tool.