ICDAR2009 Page Segmentation Competition (2009)

A Resume

November 29, 2018

Abstract

Describes the Page Segmentation competition held in the context of ICDAR2009 and presents the results of the evaluation of four submitted methods & three from ICDAR2007.

1 Introduction

Layout Analysis is the first major step in a Document Analysis workflow where, after Image Enhancement, a higher (than pixel-level) representation of the page structure is obtained. Homogeneous printed regions are identified (Page Segmentation) and labelled according to the type of their content (Region Classification). The correctness of the output of Page Segmentation is crucial as it forms the basis for all subsequent analysis and recognition processes.

2 ICDAR competition

For The page segmentation competition we have *Three objectives :*
- Comparative evaluation on realistic datasets,
- A detailed analysis of the performance of each method in different scenarios,
- A placement of the participating methods into context by comparing them to the state of the art.

Competition proceeds as follows : The authors of candidate methods registered their interest in the competition and downloaded the example dataset. One week before the competition closing date, registered authors of candidate methods were able to download the document images of the evaluation dataset. At the closing date, the organisers received both the executables and the results of the candidate methods on the evaluation dataset on a pre-defined format.

2.1 Dataset

A special dataset with rectangular regions for an easier comparaison, PRImA contemporary dataset is specific to the competition with a wide selection of contemporary document.

The ground truth is stored in an XML image representation framework. For each region on the page there is a description of its outline in the form of an isothetic polygon (i.e. a polygon having only horizontal and vertical edges). Such a representation enables a very accurate and efficient geometric description, especially for complex-shaped regions. A range of metadata is recorded for each different type of region. For example, text regions hold information about language, font, reading direction, text colour, background colour, logical label (e.g. heading, paragraph, caption, footer, etc.).

2.2 Evaluation

A new performance evaluation methodology is used instead of a pixel-based precision/recall, The new evaluation system comprises three stages:

1. *Region representation:* Ground truth and segmentation regions are transformed into an interval based representation.
2. *Region correspondence determination:* Using the interval-based representation, correspondence between parts of ground truth, segmentation and background regions is established.
3. *Error qualification and quantification:* Errors in correspondence between ground truth and segmentation regions are examined in the context of application scenarios and their significance is established.

The following conditions are identified:

- wrongly detected region = a segmentation region that has no overlap with any ground truth region.
- correctly detected region = a ground truth region that has been completely overlapped by a segmentation region.
- split region = a ground truth region that has been overlapped completely or partially by more than one segmentation region.
- merged regions = more than one ground truth region has been overlapped completely or partially by a single segmentation region
- partially or wholly missed region = A ground truth region that has not been completely (or not at all) overlapped by any number of segmentation regions.

The evaluation depends on the types of error made and the type of document, some errors can be allowable for specific cases / contexts: a merger of text regions within the reading order is allowable and A merger between regions of different type is non-allowable

3 Participating methods Brief

3.1 The ocument Image Content Extraction (DICE) system

a method that performs pixel classification rather than region segmentation in order to avoid the arbitrariness and restrictiveness of limited families of region shapes. The DICE system comprises two main steps. First individual pixels are classified primarily into machineprint text, handwriting text and photograph. Next, a post-classification methodology is used which enforces local uniformity without imposing a restricted class of region shapes.

To produce polygonal regions, a sequence of mathematical morphology operations are applied, First, masks are extracted for each content type. Second, isolated pixels are cleaned. Third, iterated open and close operations are used to remove small regions. Finally, interior pixels are removed and contours of polygons are extracted.

3.2 The Fraunhofer Newspaper Segmenter

applies the following modules to the image, in sequence:

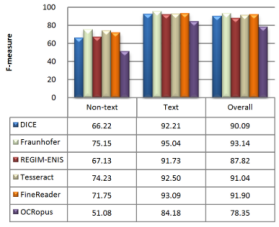
1. Pre-processing. Global optimal binarisation
2. Black separator detection
3. White separator detection
4. Page segmentation. A hybrid approach is applied comprising a bottom-up process, guided by top-down information given in the form of logical column layout of the page
5. Text line and region extraction

3.3 The REGIM-ENIS method

First, the page image is segmented based on a steerable pyramid transform. The features extracted from pyramid sub-bands serve to locate and classify regions into text and non-text regions. The second stage performs script identification to the printed and handwritten regions.

3.4 The Tesseract method

The page layout analysis method uses bottom-up methods, including binary morphology and connected component analysis, to estimate the type (text, image, separator, or unknown) of connected components. Two of the key methods employed include neighbourhood stroke-width measurement, and appropriateness of overlap between adjacent connected components



	Non-text	Text	Overall
DICE	66.22	92.21	90.09
Fraunhofer	75.15	95.04	93.14
REGIM-ENIS	67.13	91.73	87.82
Tesseract	74.23	92.50	91.04
FineReader	71.75	93.09	91.90
OCRopus	51.08	84.18	78.35

Figure 3. F-measure of the four submitted and two state-of-the-art methods.

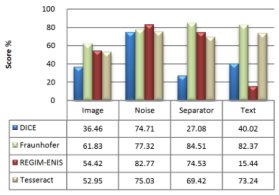
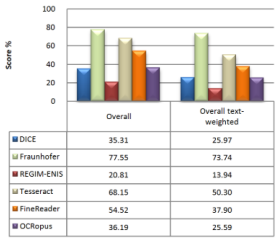


	Image	Noise	Separator	Text
DICE	36.46	74.71	27.08	40.02
Fraunhofer	61.83	77.32	84.51	82.37
REGIM-ENIS	54.42	82.77	74.53	15.44
Tesseract	52.95	75.03	69.42	73.24

Figure 4. PRImA measure for different region types for the four submitted methods.



	Overall	Overall text-weighted
DICE	35.31	25.97
Fraunhofer	77.55	73.74
REGIM-ENIS	20.81	13.94
Tesseract	68.15	50.30
FineReader	54.52	37.90
OCRopus	36.19	25.59

Figure 5. PRImA measure (standard and text-weighted) of the four submitted and two state-of-the-art methods.

Figure 1: Results

References

[1] A. Antonacopoulos, S. Pletschacher, D. Bridson and C. Papadopoulos Pattern, *ICDAR2009 Page Segmentation Competition*