Textline segmentation datasets, papers and results

October 25, 2018

1 cBAD dataset

[CBAD: ICDAR2017 Competition on Baseline Detection](#)

Explanation: This competition contains two datasets, Simple Documents and Complex Documents. Only task is baseline detection.

[Download dataset](#)
[Download evaluator](#)

Paper	Simple			Complex		
	P-Val	R-Val	F-Val	P-Val	R-Val	F-Val
dhSegment: A generic deep-learning approach for document segmentation	0.943	0.939	0.941	0.826	0.924	0.872
DMRZ	0.973	0.970	0.971	0.854	0.863	0.859
BYU	0.878	0.907	0.892	0.773	0.820	0.796
UPVLC	0.937	0.855	0.894	0.833	0.606	0.702
IRISA	0.883	0.877	0.880	0.692	0.772	0.730
LITIS	0.780	0.836	0.807	-	-	-

2 HisDoc dataset

[ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts](#)

Explanation: This competition contains 3 datasets (CB55, CSG18, CSG865) and two tasks. Task 2 is baseline detection, Task 3 is bounding polygon detection.

[Download dataset](#)
[Download evaluator for task 2](#)
[Download evaluator for task 3](#)

Paper	Task 2			Task 3		
	CB55 (F1-scr)	CSG18 (F1-scr)	CSG863 (F1-scr)	CB55 (PIU)	CSG18 (PIU)	CSG863 (PIU)
Baseline Detection in Historical Documents using Convolutional U-Nets	99.91	99.25	98.52	-	-	-
System-8	98.96	98.53	97.16	93.75	94.47	90.81
System-2	95.97	98.79	98.79	80.23	75.31	93.68
System-7	95.34	87.34	97.51	-	-	-
System-6	30.53	54.52	46.09	-	-	-

3   ICDAR 2013 Textline Segmentation Contest

ICDAR2013 Handwriting Segmentation Contest

Explanation: Metrics are detection rate (DR), recognition accuracy (RA) and F measure (FM)

Download dataset

Download evaluator

Paper	DR	RA	FM
A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents	96.75	96.21	96.48
Winner	98.68	98.64	98.66

4   ICFHR 2010 Handwriting Segmentation Contest

ICFHR 2010 Handwriting Segmentation Contest

Explanation: Metrics are detection rate (DR), recognition accuracy (RA) and F measure (FM)

Download dataset

Download evaluator

Paper	DR	RA	FM
A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents	98.10	97.86	97.98
Text Line Detection for Heterogeneous Documents	97.18	96.94	97.06
Winner	97.54	97.72	97.63

5   ICDAR2009 Handwriting Segmentation Contest

ICDAR2009 Handwriting Segmentation Contest

Explanation: Metrics are detection rate (DR), recognition accuracy (RA) and F measure (FM).

Download dataset

Download evaluator

Paper	DR	RA	FM
A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents	99.21	99.13	99.17
Text Line Detection for Heterogeneous Documents	98.59	98.59	98.59
Language-Independent Text-Line Extraction Algorithm for Handwritten Documents	99.60	99.63	99.62
Winner	99.55	99.50	99.53

6   Saint Gall dataset

Ground truth creation for handwriting recognition in historical documents

Explanation: Metrics are detection rate (DR), recognition accuracy (RA) and F measure (FM). Icdar 2009 contest evaluator can be used to measure these metrics.

Download dataset

Download evaluator

Paper	DR	RA	FM
Complete Text Line Extraction with Convolutional Neural Networks and Watershed Transform	96.39	96.52	96.46
Text Line Detection for Heterogeneous Documents	98.59	98.59	98.59
Language-Independent Text-Line Extraction Algorithm for Handwritten Documents	99.60	99.63	99.62

7 Parzival dataset

Ground truth creation for handwriting recognition in historical documents

Explanation: Metrics are detection rate (DR), recognition accuracy (RA) and F measure (FM). Icdar 2009 contest evaluator can be used to measure these metrics.

Download dataset

Download evaluator

Paper	DR	RA	FM
Complete Text Line Extraction with Convolutional Neural Networks and Watershed Transform	98.65	98.86	98.75