Text Line Extraction Based on Distance Map Features and Dynamic Programming

Vicente Bosch, Verónica Romero, Alejandro H. Toselli, Enrique Vidal – {vbosch/vromero/ahector/evidal}@prhlt.upv.es

Motivation

- Text line Segmentation (TLS) is a basic layout document task that is a pre-requisite for most KWS and HTR systems.
- TLS is usually tackled in two steps: detection and extraction
- The document layout community has currently shifted the focus to baseline detection only.
- This focus change creates the need for extraction methods that are able to capitalize on the results yielded by these new baseline detection systems.
- **We present a robust binarization-free approach inspired in path planning algorithms that uses the baseline information and a distance map in order to calculate equidistant separation frontiers**

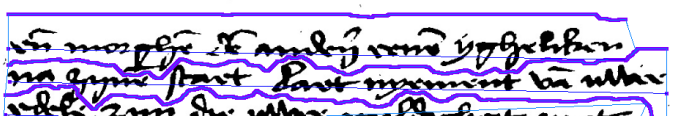
The Distance Model

- Approach inspired in DTOCS and WDTOCS ideas
- Distance map calculated on grey-scale image of page

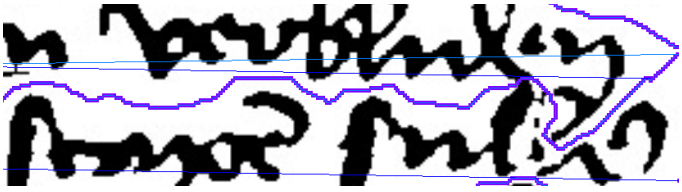
Baseline Usage and Frontier Calculation

Two levels of modelling:

- Use baselines to delimit search areas:



- Forward-Backward dynamic programming algorithm calculates best 8-connected path:



Issue Resolution

- No hard frontiers implies an optimal path will always be computed
- Collisions with black pixels can be detected and corrected

ICDAR'13 Competition Dataset Results

- ICDAR 2013 Competition corpus with standard measures used
- Two baseline scenarios: ground-truth vs automatically detected
- Automatically detected baselines were yielded by a system based on extremely randomized trees and the dbscan algorithm
- Two extraction polygon scenarios reviewed: simple projection vs dynamic programming

Method	D_R(%)	R_A(%)	F_M(%)
REGIM	40.38	35.70	37.90
AegeanUniv	77.59	77.21	77.40
PRHLT-17 + Simple Projection	89.84	83.56	86.59
ETS	86.66	86.68	86.67
Jadavpur Univ	87.78	86.90	87.34
GT. Base lines + Simple Projection	89.27	89.24	89.25
LRDE	96.70	88.20	92.25
PPSL	94.00	92.85	93.42
PRHLT-17 + Proposed Method	95.8	93.10	94.43
PortoUniv	94.47	94.61	94.54
CASIA-MSTSeg	95.86	95.51	95.68
URO-17	96.75	96.21	96.48
CVC-14	98.40	95.00	96.67
CMM	98.54	98.29	98.42
PAIS	98.49	98.56	98.52
INMC	98.68	98.64	98.66
ILSP-LWSeg-09	99.16	98.94	99.05
GT. Baselines + Proposed Method	99.62	99.58	99.60

Segmentation Results vs. HTR Results

- Experiments were carried out using the *C5 Hattem Manuscript*
- From the total of 572 leaves, a subset of 40 pages was used
- WER and CER results were calculated in a 8-block cross-validation experiment for each scenario
- Graphical error competition measure was calculated to review correlation

Extr. Method	GT	Simple Proj.		DP	
Baseline Type	NA	Straight	Line Seg.	Straight	Line Seg.
o2o	1592	1217	1306	1376	1405
F_M (%)	100	76.4	82.0	88.4	93.1
WER	34.8	36.3	35.4	37.83	35.18
CER	15.8	18.1	17.3	17.9	16.2

Conclusions

- We present a text line extraction approach that is applicable to printed as well as historical handwritten text
- The algorithm generates separation frontiers that are equidistant to the two adjacent text lines
- The method is able to capitalize on the detected baselines provided by other methods
- Our solution yields better results proportionally to the quality of the provided baselines
- We have experimentally proved that baseline detection performs the brunt of the work required for text line segmentation
- Our experimentation provides insight into the lack of correlation between the graphical error measure and the word error measure

Acknowledgements

This work has been supported by the EU project READ (Horizon-2020 programme, grant Ref. 674943).