Adaptive Methods for Robust Document Image Understanding
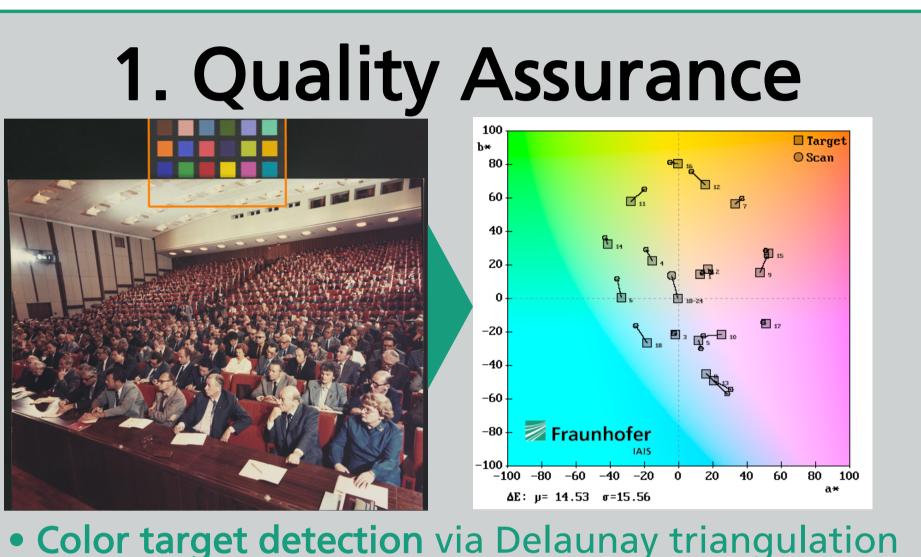
Iuliu Konya

Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Sankt Augustin, Germany

iuliu.konya@iais.fraunhofer.de; kiuliu@yahoo.com

First Examiner: Prof.Dr. Christian Bauckhage (Fraunhofer IAIS/Univ. Bonn)
Second Examiner: Prof.Dr. Michael Clausen (Univ. Bonn)
Advisor: Dr. Stefan Eickeler (Fraunhofer IAIS)

Working DIU System Data Flow Diagram

Page Splitting,

Page Flattening

Border Removal

1. Quality Assurance

• **Color target detection** via Delaunay triangulation matching
• **97.1% precision** on 239 image dataset: books, newspaper excerpts, photographs; 300-600dpi; varying target orientations

2. Paper Degradation and Artifact Removal

Classic

Proposed

• **FastICA** demixing to text, paper texture, degradation/ noise

• **Results: an order of magnitude OCR**

	Proposed	Otsu	Sauvola
Tesseract	2182	15193	13754
	5%	35%	32%
Finereader	17184	35207	38360
	40%	82%	85%

Levenshtein distance and error rate on a 22 document dataset containing 5833 words and 42825 characters

3. Color Reduction

a) Original image b) Ground truth
c) Result of using IPT d) Result of using RGB
e) Result of using CIELAB f) Result of using S-CIELAB

• **Comparison** of color spaces/ color appearance models: RGB, CIELAB, S-CIELAB, CIECAM (IPT) for **text extraction**

• Tested using 4 **different color reduction** algorithms: K-Means, Median Cut, SRM, Mean Shift against pixel-accurate ground truth (MSc thesis)
• Conclusion: CIECAM slight advantage but evaluation measure **(F-measure) not strong enough**

4. Binarization

Original   Binarized

• **Constant-time locally optimal** adaptive binarization framework (proof-of-concept tests: Otsu, Kittler)
• Fully **automatic parameter** estimation
• Using threshold interpolation **run-time <1s** per image (*2400x3500 pixels*)

5. Skew& Orientation Detection

• **Generic, parameter-free** skew detection via Euclidean MST

• **Orientation** detection using same algorithm(!), specific to Roman script

• **Orientation:** UW-I: 99.4%, OCRopus 100%, own dataset 99.9%
• **Skew:** 111,000 image dataset (born-digital + UW-I, 360x1° rot./image) – basic alg: 0.23° avg err
impr. alg. 0.16° avg err

6. Page Segmentation
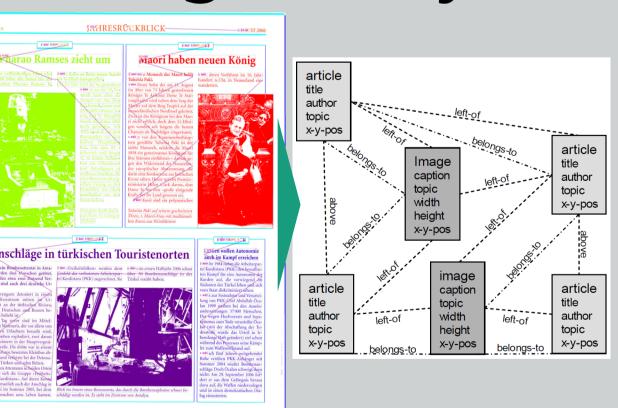

• **Production system** – Fraunhofer Newspaper Segmenter (>1,000,000 pg. projects)
• **Target:** complex **(multi-column)** layouts, **isothetic polygonal** regions

• Evaluation: **newspaper** set (22 page) – 97% prec, 92% recall; ICDAR'09 **magazine/journal** set (55 page) – PRImA measure 77.5% (winner); ICDAR'11 **historical doc.** set (100 pages) - ?

7. Logical Layout Analysis

• **Rule-based LLA** works well on simple to **medium complexity, unchanging** layouts (see evaluation)
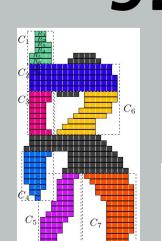
• **Trainable LLA** – proposed graph-structured CRF
• Evaluation (rule-based alg.): **100** chronicle images;
621 titles (incl. sub-, roof-, intermediary-)– 91% F-measure
255 captions – 97.6% F-measure
311 articles – correct: 85%; errors: 13% split, 2% merge

8. Front Page Detection

• **Model-based,** using both structural and statistical features (Delaunay triangulation, DCT coef.)
• **Semi-automatic** stable component selection
• Proposed **weighting function** for (sub-)graph matching (otherwise NP-hard)
• Evaluation: 17542 real-life images, 1141 newspaper issues – **100% precision, 99.9% recall**

9. Font Enhancement

• **Target: historical documents** printed using **hot metal typesetting** (beginning until mid-20th century)
• **DSCCs** (see fig.) as initial candidates, heuristic filtering
• **Levenshtein distance** as evaluation measure

	Affected dataset 33460 chars 5034 words 27 images	Unaffected dataset 30036 chars 4730 words 25 images
Tesseract	4793	743
Tess + Enhance	2288	756
Tess + Enhance relative diff.	52.3%	-1.7%
FineReader	1720	424
FR + Enhance	1074	441
FR + Enhance relative diff.	37.5%	-4%
Overall relative diff.	44.9%	-2.8%

OCR – Printed and Handwritten Character Recognition

Larger-than-Life Digital Repository