

2016 12th IAPR Workshop on Document Analysis Systems (DAS 2016)

**Santorini, Greece
11-14 April 2016**



**IEEE Catalog Number: CFP1670E-POD
ISBN: 978-1-5090-1793-5**

**Copyright © 2016 by the Institute of Electrical and Electronics Engineers, Inc
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

******This publication is a representation of what appears in the IEEE Digital Libraries. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP1670E-POD
ISBN (Print-On-Demand):	978-1-5090-1793-5
ISBN (Online):	978-1-5090-1792-8

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2016 12th IAPR Workshop on Document Analysis Systems

DAS 2016

Table of Contents

Message from the General Chairs and Program Chairs	xi
Conference Organization	xii
Program Committee	xiii
Additional Reviewers	xv
Sponsors	xvi

Camera-Based Document Image Analysis

Delaunay Triangulation-Based Features for Camera-Based Document Image Retrieval System	1
<i>Q. B. Dang, Marçal Rusiñol, M. Coustaty, M. M. Luqman, C. D. Tran, and Jean-Marc Ogier</i>	
High Performance OCR for Camera-Captured Blurred Documents with LSTM Networks	7
<i>Fallak Asad, Adnan Ul-Hasan, Faisal Shafait, and Andreas Dengel</i>	
Text Detection in Arabic News Video Based on SWT Operator and Convolutional Auto-Encoders	13
<i>Oussama Zayene, Mathias Seuret, Sameh M. Touj, Jean Hennebert, Rolf Ingold, and Najoua E. Ben Amara</i>	

Deep NN for Document Analysis

RNN Based Uyghur Text Line Recognition and Its Training Strategy	19
<i>Pengchao Li, Jiadong Zhu, Liangrui Peng, and Yunbiao Guo</i>	
CNN Based Transfer Learning for Historical Chinese Character Recognition	25
<i>Yejun Tang, Liangrui Peng, Qian Xu, Yanwei Wang, and Akio Furuhashi</i>	
Complete System for Text Line Extraction Using Convolutional Neural Networks and Watershed Transform	30
<i>Joan Pastor-Pellicer, Muhammad Zeshan Afzal, Marcus Liwicki, and María José Castro-Bleda</i>	

Document Analysis for Digital Humanities

Historical Document Dating Using Unsupervised Attribute Learning	36
<i>Sheng He, Petros Samara, Jan Burgers, and Lambert Schomaker</i>	
Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis	42
<i>Théodore Bluche, Dominique Stutzmann, and Christopher Kermorvant</i>	
Large Scale Continuous Dating of Medieval Scribes Using a Combined Image and Language Model	48
<i>Fredrik Wahlberg, Lasse Mårtensson, and Anders Brun</i>	
An Interactive Transcription System of Census Records Using Word-Spotting Based Information Transfer	54
<i>Joan Mas, Alicia Fornés, and Josep Lladós</i>	

Forensic Document Analysis

General Pattern Run-Length Transform for Writer Identification	60
<i>Sheng He and Lambert Schomaker</i>	
Banknote Counterfeit Detection through Background Texture Printing Analysis	66
<i>Albert Berenguel, Oriol Ramos Terrades, Josep Lladós, and Cristina Cañero</i>	
Performance of an Off-Line Signature Verification Method Based on Texture Features on a Large Indic-Script Signature Dataset	72
<i>Srikanta Pal, Alireza Alaei, Umapada Pal, and Michael Blumenstein</i>	

Document Image Analysis (DIA) Systems

Word Segmentation Using the Student's-t Distribution	78
<i>Georgios Louloudis, Giorgos Sfikas, Nikolaos Stamatopoulos, and Basilis Gatos</i>	
MSIO: MultiSpectral Document Image BinarizatIon	84
<i>Markus Diem, Fabian Hollaus, and Robert Sablatnig</i>	
SDK Reinvented: Document Image Analysis Methods as RESTful Web Services	90
<i>Marcel Würsch, Rolf Ingold, and Marcus Liwicki</i>	
A Compliant Document Image Classification System Based on One-Class Classifier	96
<i>Nicolas Sidère, Jean-Yves Ramel, Sabine Barrat, Vincent Poulain D'Andecy, and Saddok Kebairi</i>	

Graphics Recognition and Applications

Automatic Hyperlinking of Engineering Drawing Documents	102
<i>Purnendu Banerjee, Sumit Choudhary, Supriya Das, Himadri Majumdar, Rahul Roy, and Bidyut B. Chaudhuri</i>	
Understanding Line Plots Using Bayesian Network	108
<i>Rathin Radhakrishnan Nair, Nishant Sankaran, Ifeoma Nwogu, and Venu Govindaraju</i>	
What You See is What You Get? Automatic Image Verification for Online News Content	114
<i>Sarah Elkasrawi, Andreas Dengel, Ahmed Abdelsamad, and Syed Saqib Bukhari</i>	

Semi-automatic Text and Graphics Extraction of Manga Using Eye Tracking Information	120
<i>Christophe Rigaud, Thanh-Nam Le, J.-C. Burie, Jean-Marc Ogier, Shoya Ishimaru, Motoi Iwata, and Koichi Kise</i>	

Performance Evaluation and Ground Truthing

Creating Ground Truth for Historical Manuscripts with Document Graphs and Scribbling Interaction	126
<i>Angelika Garz, Mathias Seuret, Fotini Simistira, Andreas Fischer, and Rolf Ingold</i>	
Document Image Quality Assessment Based on Texture Similarity Index	132
<i>Alireza Alaei, Donatello Conte, Michael Blumenstein, and Romain Raveaux</i>	
Quality Prediction System for Large-Scale Digitisation Workflows	138
<i>Christian Clausner, Stefan Pletschacher, and Apostolos Antonacopoulos</i>	

Handwriting Recognition and Word Spotting

Keyword Retrieval Using Scale-Space Pyramid	144
<i>Irina Rabaev, Klara Kedem, and Jihad El-Sana</i>	
A Segmentation-Free Handwritten Word Spotting Approach by Relaxed Feature Matching	150
<i>Anders Hast and Alicia Fornés</i>	
Increasing Robustness of Handwriting Recognition Using Character N-Gram Decoding on Large Lexica	156
<i>Martin Schall, Marc-Peter Schambach, and Matthias O. Franz</i>	
Word Spotting in Historical Document Collections with Online-Handwritten Queries	162
<i>Christian Wieprecht, Leonard Rothacker, and Gernot A. Fink</i>	

OCR Systems

QATIP — An Optical Character Recognition System for Arabic Heritage Collections in Libraries	168
<i>Felix Stahlberg and Stephan Vogel</i>	
OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters	174
<i>Adnan Ul-Hasan, Syed Saqib Bukhari, and Andreas Dengel</i>	
Error Detection in Indic OCRs	180
<i>V. S. Vinitha and C. V. Jawahar</i>	
Multilingual OCR for Indic Scripts	186
<i>Minesh Mathew, Ajeet Kumar Singh, and C. V. Jawahar</i>	

Poster Papers 1

A Fine-Grained Approach to Scene Text Script Identification	192
<i>Lluís Gómez and Dimosthenis Karatzas</i>	

OCR Error Correction Using Character Correction and Feature-Based Word Classification	198
<i>Ido Kissos and Nachum Dershowitz</i>	
New Sharpness Features for Image Type Classification Based on Textual Information	204
<i>K. S. Raghunandan, Palaiahnakote Shivakumara, G. Hemantha Kumar, Umapada Pal, and Tong Lu</i>	
Removal of Gray Rubber Stamps	210
<i>Soumyadeep Dey, Jayanta Mukhopadhyay, and Shamik Sural</i>	
Evaluation of the Stability of Four Document Segmentation Algorithms	215
<i>Sébastien Eskenazi, Petra Gomez-Krämer, and Jean-Marc Ogier</i>	
Combination of Structural and Factual Descriptors for Document Stream Segmentation	221
<i>Romain Karpinski and Abdel Belaïd</i>	
Document Image Quality Assessment Using Discriminative Sparse Representation	227
<i>Xujun Peng, Huaigu Cao, and Prem Natarajan</i>	
Modified X-Y Cut for Re-Ordering Strokes of Online Handwritten Mathematical Expressions	233
<i>Anh Duc Le, Hai Dai Nguyen, and Masaki Nakagawa</i>	
Automatic Synthesis of Historical Arabic Text for Word-Spotting	239
<i>Majeed Kassis and Jihad El-Sana</i>	
Analysis of Stroke Intersection for Overlapping PGF Elements	245
<i>Yan Chen, Xiaoqing Lu, Jingwei Qu, and Zhi Tang</i>	
Recognition-Based Approach of Numeral Extraction in Handwritten Chemistry Documents Using Contextual Knowledge	251
<i>Nabil Ghanmi and Abdel Belaïd</i>	
Entity Local Structure Graph Matching for Mislabeled Correction	257
<i>Nihel Kooli, Abdel Belaïd, Aurélie Joseph, and Vincent Poulain D'Andecy</i>	
Unsupervised Word Clustering Using Deep Features	263
<i>Mandar Kulkarni, Shirish Subhash Karande, and Sachin Lodha</i>	
An Interactive Approach with Off-Line and On-Line Handwritten Text Recognition Combination for Transcribing Historical Documents	269
<i>Emilio Granell, Verónica Romero, and Carlos D. Martínez-Hinarejos</i>	
Handwriting Transcription and Keyword Spotting in Historical Daily Records Documents	275
<i>Verónica Romero, Alejandro H. Toselli, Joan Andreu Sánchez, and Enrique Vidal</i>	
Text Extraction in Document Images: Highlight on Using Corner Points	281
<i>Vikas Yadav and Nicolas Ragot</i>	
A Table Detection Method for PDF Documents Based on Convolutional Neural Networks	287
<i>Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang</i>	
Efficient Document Image Segmentation Representation by Approximating Minimum-Link Polygons	293
<i>George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos</i>	

Page Segmentation for Historical Document Images Based on Superpixel Classification with Unsupervised Feature Learning	299
<i>Kai Chen, Cheng-Lin Liu, Mathias Seuret, Marcus Liwicki, Jean Hennebert, and Rolf Ingold</i>	
Isolated Handwritten Digit Recognition Using oBIFs and Background Features	305
<i>Abdeljalil Gattal, Chawki Djeddi, Youcef Chibani, and Imran Siddiqi</i>	
Fuzzy Integral for Combining SVM-Based Handwritten Soft-Biometrics Prediction	311
<i>Nesrine Bouadjenek, Hassiba Nemmour, and Youcef Chibani</i>	
Poster Papers 2	
OCR Accuracy Prediction Method Based on Blur Estimation	317
<i>Van-Cuong Kieu, Florence Cloppet, and Nicole Vincent</i>	
Marginal Noise Reduction in Historical Handwritten Documents — A Survey	323
<i>Arpita Chakraborty and Michael Blumenstein</i>	
Preserving Text Content from Historical Handwritten Documents	329
<i>Arpita Chakraborty and Michael Blumenstein</i>	
Globally Optimal Text Line Extraction Based on K-Shortest Paths Algorithm	335
<i>Luan Wang, Seiichi Uchida, Wei Fan, and Jun Sun</i>	
Natural Scene Character Recognition Using Robust PCA and Sparse Representation	340
<i>Zheng Zhang, Yong Xu, and Cheng-Lin Liu</i>	
Recognition of Greek Polytonic on Historical Degraded Texts Using HMMs	346
<i>Vassilis Katsouros, Vassilis Papavassiliou, Fotini Simistira, and Basilis Gatos</i>	
Effective Candidate Component Extraction for Text Localization in Born-Digital Images by Combining Text Contours and Stroke Interior Regions	352
<i>Kai Chen, Fei Yin, and Cheng-Lin Liu</i>	
Interactive Definition and Tuning of One-Class Classifiers for Document Image Classification	358
<i>Nathalie Girard, Roger Trullo, Sabine Barrat, Nicolas Ragot, and Jean-Yves Ramel</i>	
Election Tally Sheets Processing System	364
<i>Juan Ignacio Toledo, Alicia Fornés, Jordi Cucurull, and Josep Lladós</i>	
Human-Document Interaction Systems — A New Frontier for Document Image Analysis	369
<i>Dimosthenis Karatzas, Vincent Poulain D’Andecy, Marçal Rusiñol, Antonio Chica, and Pere-Pau Vazquez</i>	
Named Entity Recognition from Unstructured Handwritten Document Images	375
<i>Chandranath Adak, Bidyut B. Chaudhuri, and Michael Blumenstein</i>	
Visual Analysis System for Features and Distances Qualitative Assessment: Application to Word Image Matching	381
<i>Frédéric Rayar, Tanmoy Mondal, Sabine Barrat, Fatma Bouali, and Gilles Venturini</i>	
An Adaptive Zoning Technique for Word Spotting Using Dynamic Time Warping	387
<i>A. Papandreou, Basilis Gatos, and K. Zagoris</i>	
Visual Script and Language Identification	393
<i>Angelos Nicolaou, Andrew D. Bagdanov, Lluís Gómez, and Dimosthenis Karatzas</i>	

Handwritten and Machine-Printed Text Discrimination Using a Template Matching Approach	399
<i>Mehryar Emambakhsh, Yulan He, and Ian Nabney</i>	
Making Europe's Historical Newspapers Searchable	405
<i>Clemens Neudecker and Apostolos Antonacopoulos</i>	
Keyword Spotting in Handwritten Documents Using Projections of Oriented Gradients	411
<i>George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos</i>	
Online Arabic Handwriting Recognition with Dropout Applied in Deep Recurrent Neural Networks	417
<i>Rania Maalej, Najiba Tagougui, and Monji Kherallah</i>	
Automatic Selection of Parameters for Document Image Enhancement Using Image Quality Assessment	422
<i>Ritu Garg and Santanu Chaudhury</i>	
A Simple and Effective Solution for Script Identification in the Wild	428
<i>Ajeet Kumar Singh, Anand Mishra, Pranav Dabral, and C. V. Jawahar</i>	
Camera-Based System for User Friendly Annotation of Documents	434
<i>Yusuke Oguma and Koichi Kise</i>	
Searching Corrupted Document Collections	440
<i>Jason Soo and Ophir Frieder</i>	
Author Index	446