

Generic Layout Analysis of Diverse Collection of Documents

Syed Saqib Bukhari

Advisor: Prof. Dr. Thomas M. Breuel

Image Understanding and Pattern Recognition (IUPR)

Department of Computer Science, Technical University of Kaiserslautern, Germany

bukhari@iupr.com, bukhari@informatik.uni-kl.de



Diverse Collection of Documents



Text-line is the dominant geometrical layout structure for collection of diverse documents.

Problem:

A large number of text-line finding methods are proposed, but most of them are **domain specific**.

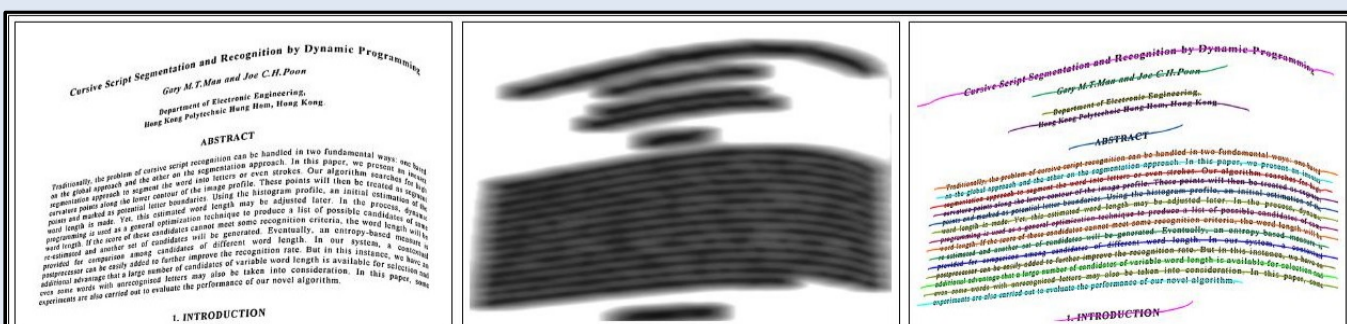
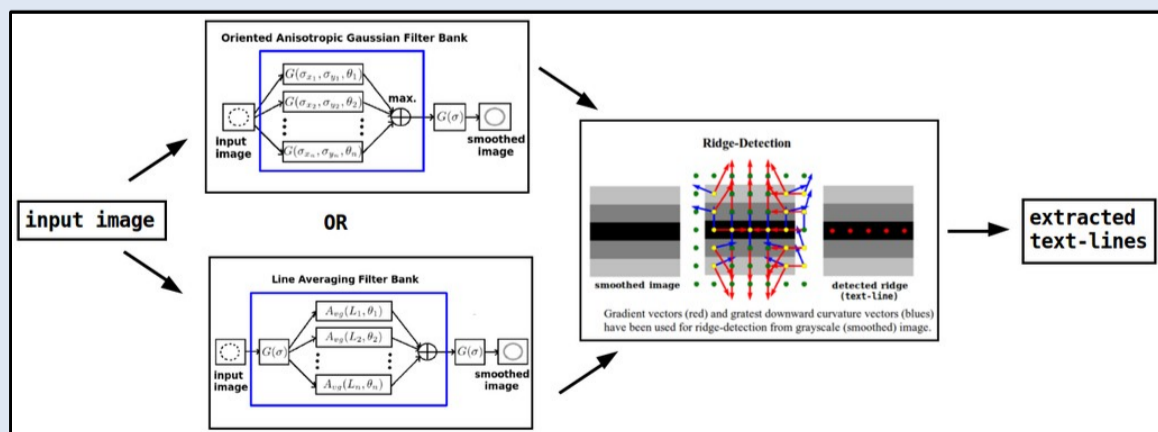
Solution:

A **generic text-line extraction method** that can be equally applied on a collection of diverse documents.

Generic Text-Line Extraction Method

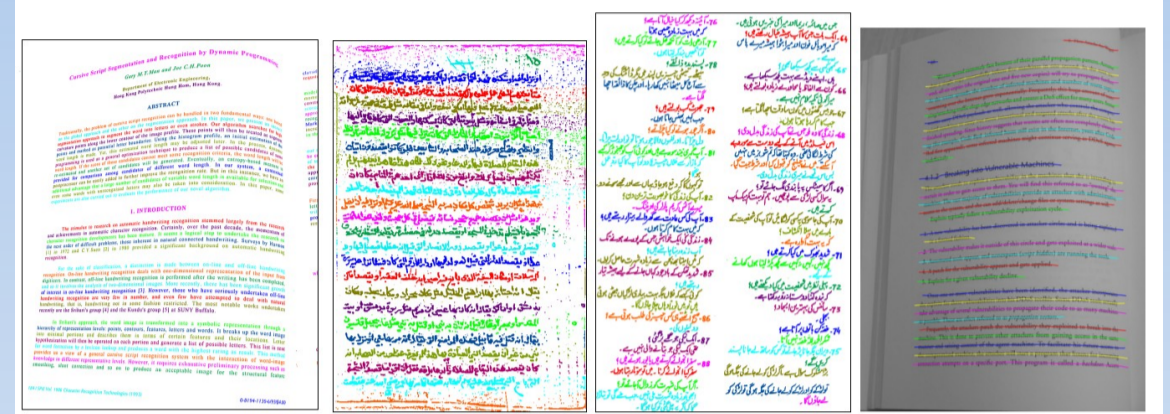
It is a combination of two standard image processing techniques [1, 4, 17]:

- Filter Bank Smoothing
- Ridge Detection



Applications

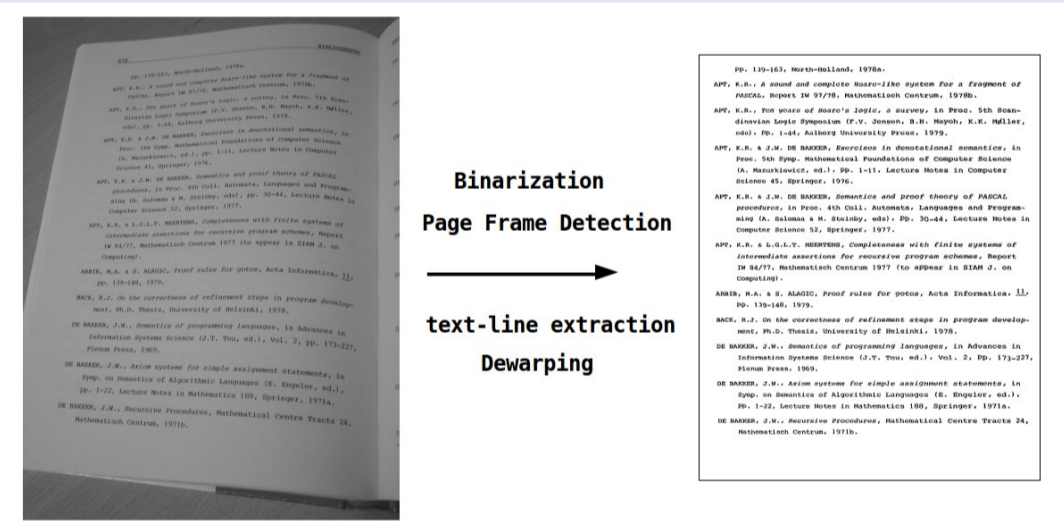
Text-Line Extraction of Diverse Documents Collection [1]



High Performance Layout Analysis of Arabic Documents [3, 5, 9]



Pepprocessing of Camera-Captured Documents [7, 12, 15, 16]



Publications

- [1] S. S. Bukhari, F. Shafait, T. M. Breuel, "An Algorithm for Text-Line Extraction from Diverse Document Collections", IEEE Transaction of Image Processing (under review).
- [2] S. S. Bukhari, F. Shafait, T. M. Breuel, "Coupled Snakelets for Curled Text-Line Segmentation from Warped Document Images", Int. Journal on Document Analysis and Recognition, IJDAR (minor revision).
- [3] S. S. Bukhari, F. Shafait, T. M. Breuel, "Layout Analysis of Arabic Script Documents", in Guide to OCR for Arabic Scripts, Springer-Verlag, 2011.
- [4] S. S. Bukhari, F. Shafait, T. M. Breuel, "Text-Line Extraction using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters", 11th Int. Conf. on Document Analysis and Recognition, ICDAR'11, China, 2011.
- [5] S. S. Bukhari, F. Shafait, T. M. Breuel, "High Performance Layout Analysis of Arabic and Urdu Document Images", 11th Int. Conf. on Document Analysis and Recognition, ICDAR'11, China, 2011.
- [6] S. S. Bukhari, F. Shafait, T. M. Breuel, "An Image Based Performance Evaluation Method for Page Dewarping Algorithms using SIFT Features", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
- [7] S. S. Bukhari, F. Shafait, T. M. Breuel, "Border Noise Removal of Camera-Captured Document Images using Page Frame Detection", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
- [8] S. S. Bukhari, F. Shafait, T. M. Breuel, "The IUPR Dataset of Camera-Captured Document Images", 4th Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'11, China, 2011.
- [9] S. S. Bukhari, F. Shafait, T. M. Breuel, "Improved Document Image Segmentation Algorithm using Multiresolution Morphology", Document Recognition and Retrieval XVIII, SPIE 2011, USA, 2011.
- [10] S. S. Bukhari, F. Shafait, T. M. Breuel, "Document Image Segmentation using Discriminative Learning over Connected Components", 9th IAPR Workshop on Document Analysis Systems, DAS'10, USA, 2010.
- [11] S. S. Bukhari, F. Shafait, T. M. Breuel, "Performance Evaluation of Curled Textlines Segmentation Algorithms on CBDAR 2007 Dewarping Contest Dataset", Int. Conf. on Image Processing, ICIP'10, Hong kong, 2010.
- [12] S. S. Bukhari, F. Shafait, T. M. Breuel, "Adaptive Binarization of Unconstrained Hand-Held Camera-Captured Document Images", Journal of Universal Computer Science, 2009.
- [13] S. S. Bukhari, F. Shafait, T. M. Breuel, "Coupled Snakelet Model for Curled Textline Segmentation of Camera-Captured Document Images", 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09, Spain, 2009.
- [14] S. S. Bukhari, F. Shafait, T. M. Breuel, "Script-Independent Handwritten Textlines Segmentation using Active Contours", 10th Int. Conf. on Document Analysis and Recognition, ICDAR'09, Spain, 2009.
- [15] S. S. Bukhari, F. Shafait, T. M. Breuel, "Dewarping of Camera-Captured Document Images", 3rd Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'09, Spain, 2009.
- [16] S. S. Bukhari, F. Shafait, T. M. Breuel, "Foreground-Background Regions Guided Binarization of Camera-Captured Document Images", 3rd Int. Workshop on Camera Based Document Analysis and Recognition, CBDAR'09, Spain, 2009.
- [17] S. S. Bukhari, F. Shafait, T. M. Breuel, "Ridges based Curled Textline Region Detection from Grayscale Camera-Captured Document Images", 13th Int. Conf. on Computer Analysis of Images and Patterns, CAIP'09, Germany, 2009.
- [18] S. S. Bukhari, F. Shafait, T. M. Breuel, "Curled Textline Information Extraction from Grayscale Camera-Captured Document Images", Int. Conf. on Image Processing, ICIP'09, Egypt, 2009.
- [19] S. F. Rashid, S. S. Bukhari, F. Shafait, T. M. Breuel, "A Discriminative Learning Approach for Orientation Detection of Urdu Document Images", 13th IEEE Int. Multitopic Conf., INMIC'09, Pakistan, 2009
- [20] S. S. Bukhari, F. Shafait, T. M. Breuel, "Segmentation of Curled Text Lines using Active Contours", 8th IAPR Workshop on Document Analysis Systems, DAS'08, Japan, 2008.