PageX: An Integrated Document Processing and Management Software for Digital Library

Hanchuan Peng, Zheru Chi, Wan-Chi Siu, and David D. Feng

Department of Electronic & Information Eng.,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong.
Email: phc@eie.polyu.edu.hk
Abstract

For digital libraries it is very important to design and to implement powerful engines to convert information on paper to electronic format. In this paper a PageX software is proposed as the integration of such engines, which include a set of intelligent document processing functions, a set of compact document management strategies, and a set of advanced accessories. With this software, a paper document will first be input as an optical image, which may be a mixture of graphics and text and may be skewed. The image will then be analyzed and decomposed into a series of component blocks, and encoded and stored in a structured and compact format. Well-developed accessory functions, including block editing and page annotation, page reconstruction and virtual editing, page matching and registration, document retrieval, etc., are provided to support advanced applications. With these carefully designed functions and strategies, PageX minimizes manual operations to a minimal degree.
Main Process of Working

1. Image Acquiring Devices
   - Binarization
   - Skew correction
   - Text blocks extraction
   - Script determination
   - Character segmentation
   - Thinning
   - Feature extraction

2. Character Recognition
   - PageX
   - Page Format Database

3. Page/Form Reformatting
   - Unrecognizable Block/Image Coding
   - Page Synthesis/Reconstruction

4. Form Registration
   - PageX
Main Structure of the System

Optical Images

Engine Set I: Page Analysis and Decomposition
(1) Foreground Extraction and Binarization
(2) Correction of Unknown Skew Angle
(3) Page Blocking
(4) Language Separation
(5) Character Separation
(6) Character Recognition
(7) ......

Engine Set II: Block Coding and Page Management
(1) Block Compression and Decoding
(2) Page Layout Control
(3) Dynamically Adjustable Data Structures
(4) ......

Engine Set III: Advanced Accessories
(1) Virtual Editing
(2) Page Matching
(3) Database Linking
(4) Retrieval
(5) ......

Compact e-Doc
Several adaptive and manual methods of binarization are included.
Both adaptive and manual skew correction methods are included.
Page Blocking

Auto Page Blocking produces the basic components of the page image for further processing.
Scripts of different languages can be distinguished and further encoded with different methods.

Registered Design, Copyright can help to protect your new shielding from imitators.

Language Separation Assistant

Binarization
- Script determination
- Character recognition
- Skew correction

- Redo from the binary image
- Just show current results
Virtual Editing

The character image and text are manipulated in a unified way.

Wrongly separated text image

Wrongly recognized text

Text image after virtual editing

Correctly recognized text
Pages are parsed as blocks and the block sequences are matched.