DOCLIB:
A Software Library for Document Processing

Stefan Jaeger, Guangyu Zhu, David Doermann
Institute for Advanced Computer Studies
Laboratory for Language and Media Processing
University of Maryland, College Park, MD 20742, USA

Kevin Chen
Booz | Allen | Hamilton
134 National Business Pkwy
Annapolis Junction, MD 20701, USA
DocLib - Goals

- **Efficient Technology Transfer**
  - software compatibility
  - balance of academia, government, and industry needs
  - common framework for document processing

- **Scalability**
  - rapid prototyping of new methods
  - simple algorithm comparison

- **Robustness and Stability**
  - high quality standards
  - platform-independence
  - accommodation of frequently changing requirements
DocLib

- **Project start:**
  October 2003

- **Manpower:**
  2-3 part-timers at LAMP and BAH

- **Software modules:**
  partly based on libraries from LAMP and government sources (e.g. TB, Eyepoke).

- **Milestones:**
  First release version: March 2004
DocLib Software

- contains standard C++ code without MFC etc.
- developed under Microsoft’s .net environment
- compiles and runs under Windows and Linux
- accessible via cvs server by core developers
- consistent coding conventions
- dynamic add-on concept
Software Development

- Source Control
- Bug Tracking
- Software Integrity
- Documentation Standards
- Distribution Portal
Software Development Process

1. Accept Bug/Feature
2. Implement
3. Unit Test
4. Bug?
   - Y: Debug
   - N: Run Regression Test
5. External Failure?
   - Y: Submit Bug
   - N: Bug?
6. Y: Debug
   - N: Submit Bug
7. Submit Bug
8. Resolve Bug
9. Check into CVS
Image Factory

Design Factors:
- Image Type objects are static/singleton objects created on startup
- DLImageFactory is a static/singleton object
- Image Type objects registers itself with the DLImageFactory during startup
- DLImageFactory keeps a list of supported Image objects as each image type calls the register function
- Additional image types can be plugged into DOCLIB without modifying existing DOCLIB code.
DocLib Architecture

DocLib’s two pillars:

**DLImage:**
- Image Processing

  - e.g. image rotation
  - image deskewing
  - image conversions
  - cc calculation
  - shape drawing

**DLDокумент:**
- Document Processing

  - e.g. page segmentation
  - text line extraction
  - logo detection
  - XML input/output
  - page layout analysis

DLImage

DLDoc.
DocLib Architecture

DLImage

Connected Component
Histogram
Image filtering
Color Conversion
Moments
Thresholding
Morphological
Edge Detection
Visualization
Transformation
Skeleton
Contour

DLDdoc.

Zone classification
Page Segmentation
Text Line
Table
Handwriting
Logo
OCR
Deskew
Script Identification
Signatures
Logical Structure

Stefan Jaeger
**DLImage**

1. Open TIFF, JPEG, GIF, PPM, PNG, BMP, PBM (P1 and P4)
   - 1a. Memory based read TIFF
   - 1b. Memory based read GIF
   - 1c. Memory based read PNG
   - 1d. Memory based read BMP
   - 1e. Memory based read PPM
   - 1f. Memory based read JPEG
   - 1g. Memory based read PBM
   - 1h. Supports multiple page TIFF file

2. Save TIFF, JPEG, GIF, PPM, and PNG images to disk
   - 2a. Memory based write TIFF
   - 2b. Memory based write BMP

3. Calculate connected components

4. Save individual components of an image to disk

5. Resize an image

6. Rotate an image (large image takes long time)

7. Copy an image

8. Extract sub-image

9. Flip image

10. Contour image

11. Reverse image

12. Paste Image

13. Set Pixel of an image

14. Convert images to and from BIT (1 bit, black/white), BYTE (8 bits, gray), COLOR (24 bits)

15. Draw shapes within images
   - a. Line (color)
   - b. box (color)

16. Dilate image

17. Erode image

18. Sharpen image

19. Blur image

20. Mask image

21. Convert to YCrCb

22. YCrCbBinarization

23. 256Color quantized

24. PercentThresholdBinarization percent (i.e. 90 or .9)

25. ThresholdBinarization rThresh gTresh bThresh

26. Color2Gray_global

27. Gray2Binary_global thresh

28. Binary2Color

29. Gray2Color

30. Binary2Gray

31. loadDocFile - Load an image as a document. Flips bits if black pixels are more than white pixels.

32. Histogram for 1 and 8 bit images

33. Projection

34. Skeletonize

35. deskew

36. Loading an unknown image from memory or from file (Unknown image type must be one of the supported images)

37. Skeletonize

38. DLDocument

39. Centroid Calculation
DLDocument - Document Hierarchy

Document:

Page 1:

Zone 1:

Zone 1_1:

Zone 1_2:

Zone 1_3:

Page 1:

Page 2:

... Zone M:

Page N:

DLDoc.
Document Hierarchy

**DLDocument class:**
- main document class
- stores document-specific information (number of pages, language, ...)
- contains a list of pages

**DLPage class:**
- general class for page information
- contains a list of zones

**DLZone class:**
- implements general container concept (bounding box)
- provides a tree hierarchy plus functions for adding or deleting zones and sub-zones
- performs consistency checks on zone boundaries
- can be sub-classed by the user (e.g. DLLine, DLCharacter, ...)
Add-Ons

DOCLIB Core

DLImage

DLDoc.

Add-on 1

Add-on 2

Add-on 3

Add-on 4

Add-on 5
Add-Ons

- Script Identification
- Layout Analysis
- Page Segmentation
- Line Processing
  - Text Line Detection
  - Line Removal
- Object Recognition
  - Logo Recognition
  - Stamp Detection
- Miscellaneous Applications
  - Classifier Combination
  - Degradation
  - Performance Evaluation
  - OCR Scansoft Interface (Windows)
GEDI

Ground-Truth Editor and Document Interface:
GEDI

- allows users to label and display rectangular zones in images
- supports user specified zone types
- handles type-specific attribute lists
- offers a graphical interface for editing and displaying zones
- enables users to create and distribute configuration files
- provides hotkeys for faster labeling
- can list multiple images in thumbnail views
- saves ground-truth and metadata as XML (compatible with DocLib)
GEDI – Java Interface

Browser Window

Type Window

Attribute Window

Image Window
GEDI - Windows

- **Browser Window**
  - lists all files currently loaded
  - allows users to easily browse through an image list

- **Type Window**
  - displays all types available for labeling
  - allows users to easily filter types or change their colors

- **Attribute Window**
  - shows the attributes for each type currently selected

- **Image Window**
  - contains the view of the current image
DocLib Status

- DocLib is available for core members and selected partners
- Number of DocLib users is on the increase
- Decisions on making DocLib publicly available are underway
- GEDI is available upon request