CAMINOVA
The Document Professionals

Caminova HC-PDF Creation using DjVu Segmentation technology abiding by PDF ISO standard

PDF Specification

Portable Document Format (PDF) is a file format created by Adobe Systems in 1993 for document exchange. PDF is used for representing two-dimensional documents in a manner independent of the application software, hardware, and operating system. Each PDF file encapsulates a complete description of a fixed-layout 2D document that includes the text, fonts, images, and 2D vector graphics which compose the documents.

Formerly a proprietary format, PDF was officially released as an open standard on July 1, 2008, and published by the International Organization for Standardization as ISO/IEC 32000-1:2008.

Raster images in PDF

Raster images in PDF (called Image XObjects) are represented by dictionaries with an associated stream. The dictionary describes properties of the image, and the stream contains the image data. (Less commonly, a raster image may be embedded directly in a page description as an inline image) Images are typically filtered for compression purposes. Image filters supported in PDF include the general purpose filters:

- **ASCII85Decode** a deprecated filter used to put the stream into 7-bit ASCII
- **ASCIIHexDecode** similar to ASCII85Decode but less compact
- **FlateDecode** a commonly used filter based on the DEFLATE or Zip algorithm
- **LZWDecode** a deprecated filter based on LZW Compression
- **RunLengthDecode** a simple compression method for streams with repetitive data using the Run-length encoding algorithm

and the image-specific filters:

- **DCTDecode** a lossy filter based on the JPEG standard
- **CCITTFaxDecode** a lossless filter based on the CCITT fax compression standard
- **JBIG2Decode** a lossy or lossless filter based on the JBIG2 standard, introduced in PDF 1.4

- **JPXDecode** a lossy or lossless filter based on the JPEG 2000 standard, introduced in PDF 1.5

Normally all image content in a PDF is embedded in the file. But PDF allows image data to be stored in external files by the use of External Streams or Alternate Images. Standardized subsets of PDF, including PDF/A and PDF/X, prohibit these techniques.

Compression and PDF

To PDF files, compression refers to image compressing. PDF formats are usually designed to compress information as much as possible (since these can tend to become very large files). Compression can be either lossy (some information is permanently lost) or lossless (all information can be restored).

PDF is a page description language, like PostScript but simplified with restricted functionality to be more lightweight, which due to not only a better data structure but also very efficient compression algorithms to reduce the file size to about half the size of an equivalent PostScript file. As mentioned earlier, PDFs use the following compression algorithms:

- LZW (Lempel-Ziv-Welch)
- FLATE (ZIP, in PDF 1.2)
- JPEG and JPEG2000 (PDF version 1.5)
- CCITT (the facsimile standard, Group 3 or 4)
- JBIG2 compression (PDF version 1.4)
- RLE (Run Length Encoding)

All of these compression filters produce binary data, which can be further converted to ASCII base-85 encoding if a 7-bit ASCII representation is required.

Compression algorithm introduction (Only the algorithms of interest)

The compression algorithms can be described in detail below:

CCITT

(International Coordinating Committee for Telephony and Telegraphy) is appropriate for black-and-white images made by paint programs and any images scanned with an image depth of 1 bit. CCITT is a lossless method. Acrobat provides the CCITT Group 3 and Group 4 compression options. CCITT Group 4 is a general-purpose method that produces good

compression for most types of monochrome images. CCITT Group 3, used by most fax machines, compresses monochrome images one row at a time.

JPEG/JPEG2000

JPEG stands for Joint Photographic Experts Group, which is a standardization committee. It also stands for the compression algorithm that was invented by this committee.

There are two JPEG compression algorithms: the oldest one is simply referred to as "JPEG" within this page. The newer is JPEG 2000 algorithm

JPEG is a lossy compression algorithm that has been conceived to reduce the file size of natural, photographic-like true-color images as much as possible without affecting the quality of the image as experienced by the human sensory engine. We perceive small changes in brightness more readily than we do small changes in color. It is this aspect of our perception that JPEG compression exploits in an effort to reduce the file size

JPEG is suitable for grayscale or color images, such as continuous-tone photographs that contain more detail than can be reproduced on-screen or in print. JPEG is lossy, which means that it removes image data and may reduce image quality, but it attempts to reduce file size with the minimum loss of information. Because JPEG eliminates data, it can achieve much smaller file sizes than other compression.

JBIG2

JBIG2 compression is superior to the CCITT or Zip algorithms when compressing scanned monochromatic copy. JBIG2 (Joint Bi-level Image Experts Group) encodes compresses monochrome (1 bit per pixel) image data from 20:1 to 50:1 for pages full of text. Like other dictionary-based algorithms (LZW, ZIP) JBIG2 creates a table of unique symbols and when a subsequent symbol matches one in the table, it substitutes a token pointing to the table index. JBIG2 also compresses the entire table.

CAMINOVA
The Document Professionals

Caminova Implementation of HC (High Compression) PDF

Caminova provides MRC (Mixed Raster Content) functions specifically for PDF compression through DjVu segmentation. The resulting MRC compressed PDF can be loaded into any PDF viewer that supports standard PDF files and it meets all PDF specification. This process creates a PDF with better compression and quality than any other raster PDF file. HC-PDF was specifically created to improve the size and quality of scanned PDF.
Caminova DEE 7.5 offers several ways to automatically segment or break down images that consist of text and images. Image segmentation is an important feature for improving OCR recognition rates and efficient compression within complex formats such as standard MRC T.44, and PDF.

Key Features of HC-PDF Segmentation and Compression:

- Automatic segmentation of image area with user defined optimization options.
- Manual profile selection of segmentation schemes.
- Save multiple pages.
- Compression can be specified for each area type:

Caminova Segmented PDF

The feature to generate a composite HC (High Compression) PDF

PDF1.4	Three Layer: JPEG BG /FG + G4 Mask Photo : JPEG Bitonal: JBIG2 / G4
PDF1.6	Three Layer: JPEG 2000 + JBIG2 Photo: JPEG 2000

Note: We are also planning to support PDF-A in the future release.

CAMINOVA
The Document Professionals

T.44 MRC Segmentation Standard:

T.44 MRC separates an image into three different layers: foreground, background, and mask. Each layer is compressed separately using the best type of compression for that data type and is later uncompressed and recombined to restore the original image.

Key Features of T.44 MRC:

- Automatic segmentation of image area with user defined optimization options.
- Manual segmentation of the image.
- Compression can be specified for each segment type.
- Choose from CCITT G3 1D, CCITT G4 2D, CCITT G4 or JBIG/JBIG2 for best

compression of 1 bit areas.

- Compress color areas with JPEG/JPEG2000.
- Save multiple pages.

DjVu segmentation scheme follows ITU MRC/T.44 standard (Mixed Raster Content MRC, 1997) with 3 layer segmentation: foreground, background and mask. It uses JBIG2 compression for 1 bit areas and JPEG2000 for color areas.

In summary, Caminova's HC-PDF segmentation and Compression technology meets 100% of PDF ISO specification and can be viewed from any PDF viewer that supports standard PDF files and it meets all PDF specification.