DjVu Technology Primer

NOVEMBER 2004

LIZARDTECH, INC.

OVERVIEW

LizardTech's Document Express products are powered by DjVu®, a technology developed in the late 1990s by a team of researchers at AT&T Labs. DjVu is specifically designed to enable the creation of digital libraries of high-visual-quality documents (either scanned from paper or produced in other digital formats).

DjVu (pronounced "day·zha·voo") is a compression technique, a file format, and a platform for delivering information in the form of color-rich documents. DjVu uses a mixed raster content (MRC) imaging model, an approach consistent with that endorsed by the International Telecommunications Union (ITU). It relies on advanced content analysis techniques to segment a scanned color page into layers and achieve high compression ratios, low memory utilization, extremely fast rendering on screen and indexing of the material.

LizardTech acquired the DjVu technology in the year 2000 in order to commercialize and further develop its utility in specific commercial markets. DjVu offers a means not only of preserving documents, but also making them truly accessible and usable across and throughout the digital world regardless of available bandwidth or the origin, complexity or size of the originals.

HOW DJVU DIFFERS FROM OTHER TECHNOLOGIES

File size and image quality are the perennially opposing forces in digital documents. There are several ways to digitally encode an existing document so that its visual quality is preserved in digital form (TIFF and PDF, for example). The resulting file is typically too large to transfer and use efficiently over networks. Conversely, there are several ways to compress a digital document such that it becomes small enough to send and download relatively quickly (classical JPEG, for example). However, quality usually decreases along with file size to the point that the usefulness of the compressed document is severely reduced.

Only DjVu technology addresses both issues. By design, DjVu preserves the visual fidelity of documents while reducing their file size significantly more than any other compression technology.

Limitations of Common Compression Technologies

Documents are composed of many different elements including text, images, printed textures and background colors. The limitation of competing compression technologies is that they apply a single compression method to every document – and to every part of a given document –

regardless of the document's make-up. This means that an office memo composed entirely of black and white text is compressed the same way as a page from a book of color advertisements.

A compression method optimal for bitonal documents will not be optimal for rich color documents. To preserve legibility, text elements require a higher dot-per-inch (dpi) resolution and lower compression ratio than graphics and other typical color elements. Applying a blanket compression method results in files that are still quite large or have degraded visual quality, or both.

The DjVu Approach

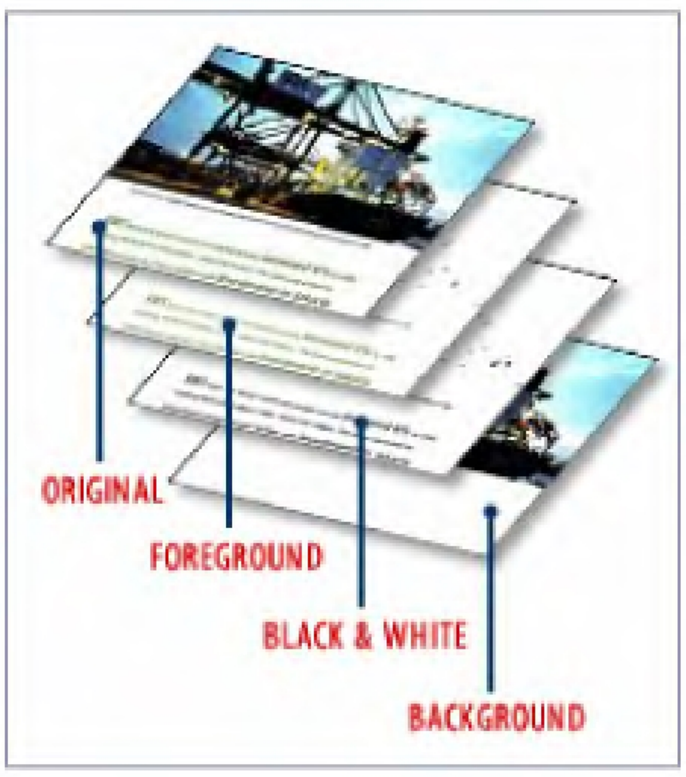
DjVu's architects asked, "Why use 200 dpi or more for image and graphical elements when a much lower resolution would be sufficient, and why use a full-color compression technique for text, which is typically black and white or low-color?"

Instead, they separated or "segmented" the document image into elements that could then be compressed independently using the best possible approach – and resolution – for each element. These include the compression methods known as JB2 and IW44.

Since the greatest obstacles to compression in a document are created by the sharp edges of the text component, segmenting the text into a separate layer enables DjVu to apply aggressive compression techniques to the relatively soft-edged graphics that remain, resulting in DjVu's high overall compression ratios while preserving a visually lossless appearance. In addition, the segmenter's ability to accurately separate text from graphics leads to dramatically improved optical character recognition (OCR) results when scanning complex color layouts.

How DjVu Works

DjVu starts by segmenting a page into layers:



- Foreground Layer – The foreground layer includes text, line art and other thin, low-color page elements. To take advantage of further opportunities for storage efficiency, this layer is itself separated into a black-and-white ("foreground mask") layer and a foreground color layer that captures the colors of the page elements in the mask layer.
- Background Layer – The background layer includes photos, graphics, tint, and paper texture. Areas of the background covered by foreground components are smoothly interpolated so as to minimize their coding cost.

Once a scanned color page is segmented into layers, DjVu uses a range of sophisticated compression techniques to represent the image with the smallest possible number of bits. Following is a typical, simplified example of how each layer is treated.

Foreground Mask Compression

For the black-and-white ("foreground mask") layer, an approach called "JB2" based on pattern matching is used: page shapes (alphabet characters, chiefly) that are found to be similar within defined parameters are considered identical and a single bitmap is used to represent all of them.

A JB2 compressed representation is derived from this, which consists of a "dictionary" of shapes plus the list of positions where each shape appears on a given page.

- Resulting compression ratios are often in excess of 100:1. A bitonal scanned page measuring 1MB as a TIFF is routinely represented in 10KB or less.
- Since multiple pages (by default, 20) are considered together during this clustering, compression ratios increase with longer documents. An improvement factor of 300 percent over industry standard CCITT-Group 4 is common, and a factor of up to 1000 percent improvement over CCITT-Group 4 is not unusual for long multiple-page documents such as books.

Foreground Color Compression

The “color” of colored text is typically represented as a separate, low-resolution layer called the “foreground color layer” and compressed using the wavelet-based compression technique IW44, which is the same method as the one used for the background layer.

IW44 is a progressive format which supports in-place updates; when DjVu files are viewed over a slow connection, background layers, text colors and images are decoded progressively and the view is gradually refined in the plug-in window as more data arrives. IW44 encoding is very similar to that of JPEG 2000, but it also supports a half-decoding that enables browser plug-ins to be very efficient while using very little memory, it enables the masking feature that guarantees that no IW44 bits will be wasted encoding pixels that will be covered by text/foreground pixels, and it is usually about three times faster.

In some situations, color is associated with each pattern position ("Color JB2") instead of the creation of a separate foreground color layer.

Background Compression

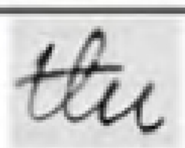
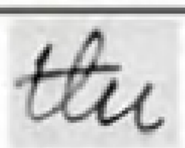
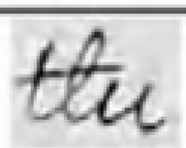
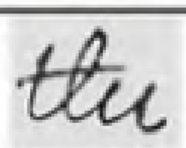
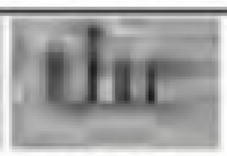
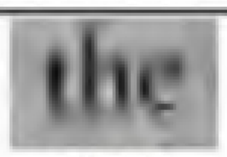
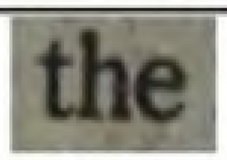
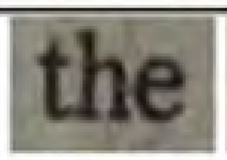
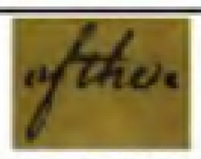
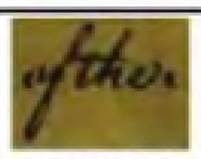
For standalone photos or for background layers of compound documents, DjVu uses IW44 compression. An IW44-compressed document image is routinely 3 to 10 times smaller than a classical JPEG (.jpg) file of equivalent quality.

Segmentation, JB2 compression and IW44 compression all work together seamlessly to produce a single compressed DjVu file with various data chunks. Overall compression rates are typically between 300:1 and 1000:1, far superior to those of any other solution for scanned color documents.

Effects on Visual Quality and Transfer Speed

The following chart illustrates the differences in quality and file size between raw, classical JPEG, IW44 (i.e., unsegmented DjVu) and regular (i.e., segmented) DjVu. The second image column shows JPEG at higher quality (300 dpi) and a correspondingly larger file size, while the third column shows JPEG when the file size is equal to that of a DjVu document. The fourth column shows IW44 compression without DjVu’s segmentation process.

The difference in visual quality is obvious in the table below. For the difference in speed, test-drive the technology yourself by downloading the free DjVu Browser Plug-in (www.lizardtech.com/download) and viewing our sample images.

Image Description	Raw image detail	JPEG, 300dpi, quality 20	JPEG, 100dpi, size=DjVu	IW44, 300dpi, size=DjVu	DjVu compressed
Magazine Add % image= 56 ads-freehand-300	 20640K	 292K 70:1	 50K 412:1	 61K 338:1	 52K 396:1
Brattain Notebook % image= 22 brattain-0001	 9534K	 116K 82:1	 17K 560:1	 20K 476:1	 19K 501:1
Scientific Article % image= 46 graham-001	 22013K	 383K 57:1	 41K 536:1	 43K 511:1	 38K 579:1
Newspaper Article % image= 50 lrr-wpost-1	 12990K	 250K 51:1	 38K 341:1	 42K 309:1	 40K 324:1
Cross-Section of Jupiter % image= 73 planets-jupiter	 24405K	 284K 85:1	 47K 519:1	 52K 469:1	 47K 519:1
XVIIIth Century book % image= 45 cuisine-p006	 12128K	 206K 58:1	 35K 346:1	 39K 310:1	 37K 327:1
US First Amend-ment % image= 30 usa-amend1	 31059K	 388K 80:1	 77K 403:1	 78K 398:1	 73K 425:1

OCR, Indexing and Full-Text Search

Creating an optical character recognition (OCR) layer for a scanned document enables keyword searches, indexing and retrieval. Because of DjVu technology's superior segmentation, its OCR engine can work on color text as well as black and white, which puts DjVu in a different league from solutions that can address only black and white text.

The Segmenter and "Hidden Text"

DjVu stores OCR information as a separate layer referred to as "hidden text," which can be programmatically exposed, exported and imported in XML for easy indexing and integration, and directly accessed by searching and indexing engines. The OCR layer is all that is needed to support keyword searches from within a document management application or from a viewer.

DjVu's OCR results on grayscale or color scans are often far superior to those produced using alternative solutions, which typically do not work well against color document images. The DjVu segmenter is able to deal with colored text, text on tint, text on images, reverse-video text – basically any text on a page. Accurately extracting this text as a bitonal layer (DjVu's foreground

mask) is key to producing high-quality OCR results. In contrast, other technologies are only able to deal with black-on-white text.

Exposing the OCR Layer for Searching and Indexing

A variety of tools can be used to expose the XML hidden text layer of any DjVu file. Many of these tools are included LizardTech's Document Express products. Such tools make it very easy to integrate full-text search for DjVu files into any document management system or searching and indexing engine. DjVu technology also allows for the integration of a different OCR engine; the Document Express SDK offers a simple, clean API for that purpose.

Large collections have been put on the Web in DjVu format with full-text search capabilities, including the 12 volumes and 10,000 pages of the Century Dictionary (www.global-language.com/century) several national library collections and content from commercial providers around the world. DjVu is currently used by thousands of users to publish and exchange scanned documents on the Web.

Efficient Universal Web Viewing

From the beginning, one of the goals in creating DjVu was to deliver a technology platform that would make it as easy to browse scanned documents as it is to browse HTML. Everything in the design of DjVu is optimized to reduce the delay between the user's decision to view a page and the display of that page on the screen, thus replicating the "page turning" experience of paper.

DjVu technology enables a number of capabilities that combine to provide an optimal viewing and browsing experience, largely by virtue of the fact that it is not necessary to fully decode a DjVu file into TIFF or an equivalent raster format before you can view or print it.

- Progressive Download and Display – A DjVu document is organized in such a way that the plug-in can immediately begin displaying the text of the first page, then progressively refine the view by adding foreground color and background color, until all the data chunks have arrived. This means that "time to first read" is very short, even on slow connections.
- Prefetching, Predecoding, Caching – While you are reading a particular page of a DjVu document, the next and previous pages are being downloaded ("prefetched") and decoded into an intermediate "semi-compressed" representation that is cached by your browser. When you navigate to one of these pages, viewing is instantaneous.
- Very Small Memory Requirements – Only the pixels of the current view are fully decoded; the rest are kept in the intermediate representation, which is decoded "on the fly" when zooming or panning.
- Individual Page Serving Capabilities – DjVu documents can be stored in an "indirect" format that makes it possible for a user to instantly jump to any page of a long document. With this approach, DjVu documents as long as several thousand pages can be efficiently accessed over the Internet or any other network.
- Embedded DjVu – Because Internet Explorer on Windows supports the <object> and <embed> HTML tags, DjVu documents can easily be embedded in HTML documents.
- Full Hyperlink Support – Viewers may not even realize it when their browser takes them to a DjVu document. Also, embedding hyperlinks in DjVu documents is easy.
- Customized Web Serving and Viewing – The DjVu technology allows you to precisely specify how a particular document should be presented to the user's browser. For example, the URL

  http://www.lizardtech.com/test.djvu?djvuopts&zoom=page&toolbar=no

specifies that the image called “test.djvu” should be displayed in the user's plug-in without any toolbar and using a zoom level that shows the entire page.

DIGITAL TO DJVU

Digital to DjVu refers to a component of DjVu technology designed for the encoding of digital documents (e.g., Word, Excel, HTML pages, PowerPoint presentations, PDF files, etc.) as opposed to scanned documents.

One way to encode an electronic document is to render it as a bitmap and then convert the bitmap into DjVu format. This is a valid approach but it requires segmenting the bitmap, which can generate artifacts. While these artifacts are generally unnoticeable on scanned documents, they can be disturbing on electronic documents where the user's expectations regarding quality are much higher.

Instead of rendering the document into a bitmap, Digital to DjVu considers page elements (words, pictures, graphics, lines, etc.) one at a time. For each such element, after occlusions are processed, the algorithm considers its shape and color content and decides whether to place it in the foreground layer or in the background layer. This in effect replaces the segmentation process used for scanned documents. Compression then proceeds normally.

This approach has several advantages including:

- Speed – Producing DjVu documents in this manner is several times faster than going through a full bitmap rendering.
- Memory Efficiency – There is never a need to render the page into a full-color raster image. Rasterization is slow and inefficient because it entails converting each page to a raster image, typically at 25 MB per page.
- No OCR Needed – Text information can usually be captured directly as foreground and background layers are being built, so no OCR is needed, and the hidden text layer is extremely accurate.
- Hyperlink Extractions – Depending on source document type, hyperlinks can usually be captured as well and inserted in the resulting DjVu document.
- Quality – The resulting DjVu files are of extremely high quality, consistent with expectations for electronic documents.

Documents created using Digital to DjVu offer maximum portability across networks, since they do not depend on any installed font packages.

Two LizardTech products currently include Digital to DjVu technology:

1. The DjVu Virtual Printer, included in Document Express Professional Edition, is a Windows application that installs as a printer and can be used to convert any kind of Windows-based document to DjVu format.
2. The PDF-to-DjVu Converter is included in Document Express Enterprise Edition and can be used to batch convert PDF and PostScript files into DjVu format.

FURTHER INFORMATION

For more detailed technical information about DjVu, see downloadable white papers at www.lizardtech.com/products/doc/techinfo.php.