Преглед НЦД 6 (2005), 45–52

Filip Brčić
(Matematički fakultet, Beograd)

DjVu – THE STANDARD FOR SCANNED DOCUMENTS

Abstract: DjVu (pronounced "déjà vu") is a document standard for efficient storing the scanned paper documents that could include color graphics, photographs in relatively small files suitable for web publishing.
Key words: DjVu, scanned, text, documents, National Center for Digitization

1. Introduction

From the ancient times, people used to write down their ideas, thoughts, drawings, sketches, plans and projects. First they wrote using the chalk on the walls of the caves. Later on, they invented papyrus and wrote on it. When they built their temples, they drew scenes from the lives of their Gods, Saints, their rulers on the walls. Through time a lot of that cultural heritage was destroyed in the wars, raids, great fires and so on. Only the ones that had a lot of copies survived. In the modern times the Internet was invented and a lot of stuff that nobody needs got copied millions of times and therefore condemned to eternity. But, the Internet brought us an opportunity to save our cultural heritage from being vanquished by time. And now a lot of books have their digital representation either in a simple textual form, or in form of a web page, a PDF file, a DOC file and so on. Those modern digital books are important for this time because they hold the knowledge of this time. But, the millions and zillions of books from the history are still in some sort of a paper form. They exist only in the real world, not in the brave new digital world. And getting these documents to the web by scanning is a bit problematic.

2. Current standards for digitally scanned documents

In this section currently available standards for storing digitally scanned documents, their advantages and their dissadvantages will be discussed.

There are many ways to store the digitally scanned documents. Since primary goal is to save the document from being vanquished, the scanned version of the document must preserve the quality of the original document. Therefore the document must be scanned at the high resolution. And high resolution scan makes big files. The uncompressed bitmap format (scanners usually use the TIFF format) is far too expensive for making digital versions of the documents. It would make documents space consuming when stored on the hard discs and bandwidth consuming when being transfered over the

Internet. That is why a lot of compressed image formats were developed, such as JPEG[1], PNG, TIFF, etc. Those algorithms are good for certain purposes, such as image compression, but for storing the scanned text with graphics, they are not efficient enough. And that is why DjVu was invented.

Here is an example that can be found on the Internet (see [1]). A typical A4 color page scanned at 300DPI would occupy about 24MB in an uncompressed format. Traditional image compression algorithms, such as JPEG, are inefficient for storing that textual page for the following reasons:

- typical JPEG file sizes vary between 400KB and 2MB, which is much better than the uncompressed 24MB, but still is too big for conventional dial-up Internet connections,
- JPEG is made for compressing general purpose pictures, not textual pictures (that is pictures with textual contents) and it uses a lossy DCT[2] algorithm on picture blocks. The DCT algorithm is an approximative algorithm and it makes the picture a little bit blurred. For most pictures this is not visible by eye because most pictures are naturaly smooth and don't have sharp edges. But, this is not the case for the scanned text which has quite a lot of sharp edges, because every character is made usualy with dark ink on the light surface to create as much contrast as possible. After regenerating such an image, we cannot get the edge to be as sharp as it used to and this blur effect causes the text to be less readable.
- JPEG requires the processor to regenerate the bitmap image to be able to display it. Therefore, the small JPEG file actually requires a large memory buffer for being displayed. That means that the gain of using JPEG is only in easier transport and storing of the image, but it still must be decompressed to be displayed. Maybe you don't have experience in managing JPEG images, but you have all watched the DivX/MPEG films and you should probably all know that if you try to play the movie in full-screen in high resolution (1024x768 or more) on a computer with a slow CPU (900MHz or less) and little RAM (128MB or less) it wouldn't be displayed well. That is because MPEG is based on the same algorithm as JPEG and it requires the frame and the music to be decompressed before being displayed.
- JPEG knows nothing about the text that is stored in the image. Therefore the text cannot be OCRed, indexed, searched or read by computer. And what is the benefit of having digitized books if they cannot be searched and easily read.
- And the least of all problems, JPEG is not made for multi-page documents and to store a multi-page document, one would be required to encapsulate those images into a container format, such as PDF, and therefore add another layer of inefficiency.

3. Benefits of DjVu

In this section the benefits of using the DjVu image compression format for storing the digitally scanned documents will be discussed.

---

[1]By JPEG image format we refer to the MIME image/jpeg format which is, in fact, JFIF or EXIF file format.

[2]DCT stands for Discrete Cosine Transform.

Now that we have convinced you that neither bitmap (TIFF[3], PNG, BMP, ...) nor JPEG file formats (JPEG embeded into JFIF, EXIF, TIFF, ...) are good for scanned text, we should probably tell you what is a better solution. And the answer is DjVu (a French word déjà vu which means already seen). That is a lossy (non-bit-preserving) compression method and file format for coding color, gray-scale and bilevel images, particularly suitable for compound images consisting of foreground text and background photographic or graphic images. These algorithms are designed for efficient storage, retrieval and display of the scanned document images on the Internet. The key benefit of DjVu is that it provides high compression rates by handling text and images differently. It provides the so called "hidden-text" layer which represents the OCRed text made from the foreground layer.

The final effect is that we can get:

- Smaller file sizes than other conventional image compression methods. Typical bitonal documents occupy from 5 to 30KB per page at 300DPI. Low-color images such as icons are typically 2 times smaller than with GIF, but can be even 10 times smaller if they contain lots of text. Photos are 2 times smaller than JPEG and also easier to be rendered. Scanned color and gray-scale documents are typically 30 to 100KB per page at 300DPI.
- Separated searchable text layer which provides easy access for the search engines to browse through the documents and index them.
- Very suitable for web presentation of a book because the books can really be read online! When one wants to read the book, he only downloads the current page in a very short time period (practically on-the-fly over the low-cost dial-up connection if the book is conventional black ink and white paper book and with just a little glitch if it is a color book).

4. Comparison between DjVu and other image formats

Here is one example that we have made in comparing several image formats with DjVu (see Table 1). It is not hard to spot that the standard bitmap formats (microsoft BitMaP - BMP, Tag Image File Format – TIFF – in this case, the uncompressed variation, Portable PixMap - PPM, Portable GrayMap - PGM, Portable BitMap - PBM, Portable aNyMap - PNM, Graphics Interchange Format - GIF and Portable Network Graphics - PNG) have the biggest files. That is because those formats have an exact bit-by-bit copy of the original image with lossless compression at most. The formats that introduce lossy compression (JPEG and DjVu) have much better performance. That information is not something that is unexpected and too important. DjVu and JPEG have very similar color and grayscale image compression algorithms and the DjVu files are only slightly smaller than the JPEG files. The advantage of DjVu image file format over JPEG is in it's ability to be rendered in parts therefore allowing faster panning and zooming of the large images.

The main advantage of DjVu image compression format is in bitonal images (that is in scanned text images). If you take a look to the the section called “The

[3]TIFF can use lossless compression, like all the other bitmap formats, but it can also be used with JPEG compressed image, although we wouldn't advise you to do so since some TIFF viewers cann handle that last feature well.

selection layer (Sjbz)" you will see that DjVu is specialized in spotting the symbols and then putting them to the paper as they appear. That is the reason why in bitonal image DjVu has drastically smaller file size than any other format.

Image	Colors	Image type	Image size
	Color	BMP	901KB
		TIFF	902KB
		PPM	901KB
		PGM	NA
		PBM	NA
		PNM	901KB
		GIF	341KB
		PNG	459KB
		JPEG	66KB
		DjVu	60KB
	Gray-scale	BMP	302KB
		TIFF	301KB
		PPM	901KB
		PGM	301KB
		PBM	NA
		PNM	301KB
		GIF	341KB
		PNG	155KB
		JPEG	54KB
		DjVu	44KB
$R = Y + \frac{3}{2}C_r$ $G = Y - \frac{1}{4}C_b - \frac{3}{4}C_r$ $B = Y + \frac{7}{4}C_b$	Bitonal	BMP	148KB
		TIFF	2.1KB
		PPM	111KB
		PGM	37KB
		PBM	4.8KB
		PNM	4.8KB
		GIF	21KB
		PNG	1.6KB
		JPEG	6.1KB
		DjVu	348B

Table 1. Image formats comparison

5. Comparison between DjVu and PDF

In the previous section the benefits of DjVu over the other compressed image formats have been discussed. Since the Portable Document Format - PDF is not a compressed image format, it didn't have the place in that discussion. But, it can be compared to DjVu as a format for holding multi-page documents. PDF is a compressed vector format with support for embedded images. When used as a format for natively digital documents (such as this document), DjVu has very little advantages. Those advantages come from the fact that PDF was not designed for online documents and it requires the document to be downloaded to the local computer to be viewed. The document can be viewed while being downloaded, but only subsequently. Take for an example reading an online encyclopedia. One cannot take a look only at the 1st, 43rd, 512th and 1324th page of that encyclopedia, but would have to download at least first 1324 pages to be able to take a look at those four pages. In the case of DjVu, one could download only those four pages that one requests.

The main advantage of DjVu over PDF is in fact in scanned documents. Consider the book scanned into JPEGs that are embedded into one PDF and on the other side that same book compressed with DjVu. First, PDF isn't designed to support hidden text layers over the image formats and therefore that kind of book wouldn't be searchable. On the other hand, DjVu supports the hidden text layers, and if the scanned images are OCRed, the contained text can be saved in DjVu which makes that DjVu searchable and indexable. Those abilities of DjVu are very important because having a lot of scanned and stored books can be just a lot of mess if one cannot find the information he seeks. Second great ability of DjVu is that it is easily readable online because of it's layered structure (for more information take a look at the the section called “Some technical details about DjVu”) and one can download only the text layer without images and background paper if one has a slow connection. And third, as seen in the previous section, DjVu has up to about 30% smaller files for color or grayscale images and almost 20 times better compression in the case of bitonal images. That would make the resulting DjVu document apparently smaller than the corresponding PDF document. As an example of that size difference, this document occupies 272KB in PDF format and 102KB in DjVu format which is almost three times smaller!

6. Some technical details about DjVu

In this section the technical details about DjVu compression format will be reviewed. The DjVu Specification has about 40 pages and some complex algorithms including several versions of new arithmetic coding algorithms (DjVu uses a version called Z'-coder), Electronic Arts' Interchange File Format (EA IFF 85), wavelet transforms, etc. We didn't want to get too much into details that are useful only to those who would like to write a (de)coder for DjVu. Whoever wants to take a closer look into DjVu technical details, should consult [2].

As was said in the earlier section, DjVu handles different types of data differently. Here is how it is accomplished.

7. DjVu image types

DjVu supports multilayered and single-layered image types. Three-layered images are called DJVU Images and there are three types of those images:

- Compound DJVU Image contains a selection layer (which is named Sjbz), a foreground layer (FG44) and a background layer (BG44). Foreground and background layers can consist of one or three color components. The selection layer has a purpose to select between those two layers. By default, only the background layer is visible, but the selection layer selects where the foreground layer should be visible and therefore defines the symbols in a document.
- Bilevel DJVU Image contains only the selection layer (Sjbz). The background layer is implicitly white and the foreground layer is implicitly black (which is useful for progressive loading of a page therefore allowing us to see the document in black&white before the color is loaded).
- Photo DJVU Image contains only a background layer (BG44).

DjVu also supports two single-layered image types, called IW44 Images:

- Color IW44 Image contains a single layer consisting of three color components (PM44).
- Gray-scale IW44 Image contains only a single layer consisting of one color component (BM44).

8. Foreground and background layers

Coding of the color and gray-scale image chunks BG44, FG44, PM44 and BM44 is based on the same algorithm. The colors are coded using the Y, C_b and C_r components, where the gray-scale images contain only the Y component. Each color layer is coded using the Dubuc–Deslauriers–Lemire (4,4) Interpolative Wavelet Transform. Within one layer coding is divided into series of slices. In the case of background layer or the IW44 Image, the slices may be coded in several chunks therefore allowing the progressive rendering. One slice contains refinement data for one color band for each color component. The image is divided in the blocks (32x32 pixels or less) starting from the lower left corner of the image.

$$\begin{aligned} R &= Y + \frac{3}{2}C_r \\ G &= Y - \frac{1}{4}C_b - \frac{3}{4}C_r \\ B &= Y + \frac{7}{4}C_b \end{aligned}$$

Equation 1. Conversion from YC_bC_r color space to RGB color space

The IW44 image compression algorithm is on par with JPEG2000 in terms of signal-to-noise ratio, but it's decoder/renderer is very memory efficient and optimized for speed. It uses a new binary adaptive arithmetic coder called the Z'-coder. Another good thing about the IW44 wavelet codec is that it allows on-the-fly decompression/rendering of

the area visible in the display window (and not more), therefore allowing faster zooming and panning of the document with less memory requirements.

9. The selection layer (Sjbz)

The selection layer is the core of DjVu image type. This layer builds a compressed library of repeating shapes in the document, such as characters, and codes the locations where they appear on each page. That is the main reason for the name déjà vu or already seen. The bitonal or bilevel DjVu documents consist of a white background layer, black foreground layer and a selection layer which defines the places that should be white or black. The data of the selection layer (Sjbz) is coded using arithmetic coding algorithm. The data in the Sjbz layer is actually an ordered list of records that define the symbols contained in the DjVu document. The following is the list of record types:

1. Start of image
2. New symbol, add to image and library
3. New symbol, add to library only
4. New symbol, add to image only
5. Matched symbol with refinement, add to image and library
6. Matched symbol with refinement, add to library only
7. Matched symbol with refinement, add to image only
8. Matched symbol, copy to image without refinement
9. Non-symbol data
10. Image refinement data
11. Comment
12. End of data

As you might have seen, not only that DjVu builds the symbol library, but it also builds new symbols by comparing them to similar ones. In a scanned document almost all the characters will be different in some point. By using DjVu, one would create only one base for the character and all the other occurrences of that characters would be interpreted as a difference therefore lowering the amount of space used by those characters.

10. Existing solutions for DjVu

Here is a short list of DjVu software sources online:

- Source code of the Open Source DjVu plug-in, independent viewer and encoders licensed under the GNU Public License can be obtained at http://djvu.sourceforge.net/.
- Plug-ins, compressors, SDKs, and commercial software can be obtained at DjVu.com.
- Servers that can convert documents in any format to DjVu are available at DjVu OpenLib (http://openlib.djvuzone.org/), Bib2Web (http://bib2web.djvuzone.org/), and Any2DjVu (http://any2djvu.djvuzone.org/).
- Papers, examples, benchmarks, pointers and other useful resources can be found at DjVuZone (http://www.djvuzone.org/).

Bibliography

1. Yann LeCun, Léon Bottou, and Patric Haffner, *DjVu: a Short Technical Introduction,* Source URL is http://djvu.sourceforge.net/abstract.html.
2. *Specification of DjVu Image Compression Format,* Version of 1999-04-29 15:46 EDT. Copyright © 1999 AT&T.
3. *What is DjVu,* Source URL is http://www.djvuzone.com/wid/index.html.

brcha@users.sourceforge.net