New File Formats in Archiving: JPEG2000, High Compressed PDF, JBIG2 with Real World Examples

Simon McPartlin
LuraTech GmbH
Berlin, Germany
Carsten Heiermann
LuraTech, Inc.
Redwood City, CA, USA

Abstract

We review the different JPEG2000 file formats for image and document compression, and describe the new functionalities of PDF in the context of scanned document compression. The suitability of JBIG2 as a component in a document compression format is highlighted and a number of document archiving scenarios are described.

JPEG2000/Part1

JPEG2000/Part1 [1] is the most recent still image compression standard developed by the Joint Picture Experts Group (JPEG). It has been designed to offer high performance lossless and lossy compression, and to offer a range of advanced features. A few of the most important features are described here.

JPEG2000/Part1 is a flexible standard, allowing a fully reversible, or lossless, compression to be performed, or the use of quantization to create smaller files using lossy compression. A single compression algorithm is used for both lossless and lossy coding, making JPEG2000/Part1 suitable for a wide range of application fields, from the lossless compression of sensitive image data to the high compression rates required for use in low bandwidth environments.

The compression rates achieved by many alternative compression techniques, including JPEG [2], depend on the image contents and desired rates cannot always be exactly achieved. JPEG2000/Part1, in contrast, allows the exact compression rate, file size or quality for an image to be specified.
The hierarchical construction of JPEG2000/Part1 files enables images to be progressively decoded in an optimal way. JPEG2000/Part1 allows the quality and resolution of the reconstructed image to be controlled without any manipulation of the compressed image data. For example, the reconstruction of a quality corresponding to a 1:100 compressed image can be quickly achieved from a 1:10 compressed image by only processing the first $1/10^{th}$ of the data. This allows images to be stored losslessly and accessed at a quality and speed adapted to individual users requirements. In addition, the wavelet transformations used in JPEG2000/Part1 enable lower resolution images to be efficiently reconstructed, with only the necessary wavelet coefficients decoded and re-transformed. Fast access to lower resolution preview or thumbnail images is therefore possible using only a fraction of the compressed image data, and without the use of additional compressed preview or thumbnail image data.

JPEG2000/Part1 combines compression efficiency with practical features and looks likely, in the next few years, to over take traditional JPEG as the most commonly used still image compression method.

JPEG2000/Part6

JPEG2000/Part6 [3] is designed to compress compound image documents based on the multi-layer mixed raster content (MRC) imaging model [4]. Compound image documents contain multiple images, both contiguous tone and bi-level, together with composition models describing how the images are combined. Such documents may, for example, contain text images with more or less bi-level content, graphic images using a limited number of colors, as well as photo realistic areas requiring a more extensive colors range.
General image compression schemes, including JPEG and JPEG2000/Part1, are not well suited to the high compression of documents containing such mixed content. In particular, the performance on text, simple drawings and

areas of high contrast is far from optimal at high compression rates [Fig. 1].

To address this problem, the new JPEG2000 standard was

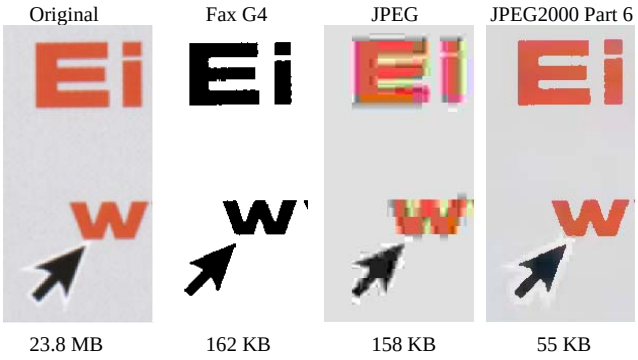

Fig. 1. Comparison of document compression results. Shown is a detail of a A4 size color scan compression.

supplemented with Part6 – a compound image file format for document imaging, archiving, pre-press, fax-like and medical imaging applications.

The multi-layer mixed content model supported by Part6 allows a suitable compression method appropriate to the content to be selected. Bi-level text, for example, can be compressed using JBIG2 or Fax Group 4, while natural color images can be efficiently compressed using JPEG2000.

The individual layers in a Part6 file are represented by layout objects, each consisting of an image and a transparency mask. The transparency masks, either bi-level or contiguous tone, determine how the images are to be combined to reconstruct the document image. The size, position and resolution of the images and masks can be freely chosen. This can be used, for example, to store important text content with a higher resolution than less important logo content.

A typical Part6 application may segment the content of a scanned document into two layout objects [Fig. 2]:

- a background object consisting of an opaque mask and an image containing natural image data; and
- a foreground object consisting of a mask defining the shape of text and graphic content, with an image defining the color for this content.

The layout images may be compressed using JPEG2000, or even the old JPEG [2], while the layout masks may be compressed using one of the Part6 supported mask compression methods: JBIG2, Fax Group 3, Fax Group 4, JBIG and JPEG2000.

The ability of an accurate segmentation algorithm to store the shape of text and high contrast content in a mask object allows high compression ratios to be achieved without affecting the text legibility. A good segmentation ensures not only that the text remains readable to the human eye but also often increases OCR accuracy [5].

A method for segmenting a document into individual layout objects is not defined in the Part6 standard. Designing a good and efficient segmentation algorithm is not trivial, and depends to some extent on the type of document and the application field.

Part6 is a step towards a new generation of flexible compression formats. The ability to apply optimal compression algorithms to the mixed-content found in documents makes possible significantly higher compression ratios than could be achieved using more traditional methods such as JPEG or JPEG2000/Part1.

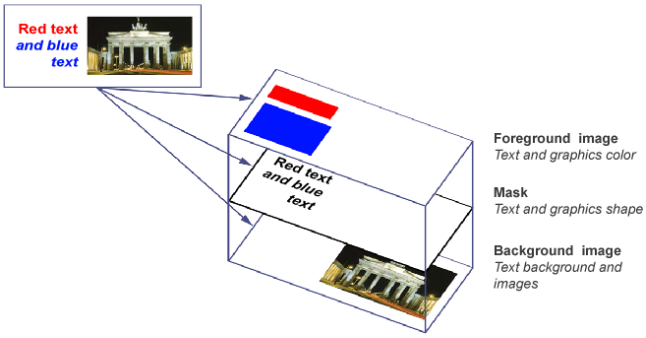

Fig. 2. Text and image segmentation scheme.

High compressed PDF

Adobe's Portable Document Format (PDF) [6,7] has established itself as one of the most widely supported digital document formats. In contrast to JPEG2000/Part6, PDF is currently predominately used as an efficient vector format, storing text with associated vector font (e.g. Postscript, true-type, etc.) information. PDF does, however, also provide a number of facilities for the storage of plain raster graphic input, making it useful for applications dealing with scanned documents.

The ability to store multiple images in a PDF file, together with the transparency masks introduced in PDF 1.4 and scaling features that can be used to combine them, gives PDF the functionality to support the multi-layer mixed raster content (MRC) imaging model [4].

The range of compression formats supported by PDF, including JPEG, Fax Group 4, JBIG2 (PDF 1.4 and higher) and JPEG2000/Part1 (PDF 1.5 and higher), together with the wide distribution of Adobe's Acrobat Reader, make PDF

an interesting – though significantly more complex and proprietary – alternative file format for the storage of compound documents.

JBIG2

The JBIG2 standard has been developed by the Joint Bi-level Image Experts Group (JBIG) for the efficient lossless and lossy compression of bi-level content [8, 9].

Thanks to higher compression performance it should supersede the widely used Fax Group 4 method for bi-level compression. In particular, the use of symbol dictionaries and symbol matching in JBIG2 enables very efficient adaptive compression of documents containing recurring symbols, making JBIG2 ideal for compressing text documents.

The compression rates achievable using JBIG2 are heavily dependent on the bi-level content, with lossless compression typically creating files 30-50% smaller than the Fax Group 4 equivalents.

JBIG2 is unique among bi-level compression methods in supporting both lossless and lossy compression. Lossy compression is extremely efficient when insignificant pixel differences are acceptable, generating files up to 40% smaller than lossless JBIG2 [10].

JBIG2 has been designed to be embedded in a number of existing file formats, including JPEG2000/Part6, PDF and TIFF. The flexibility of JBIG2, together with the efficiency with which it can compress text, makes it ideally suited as a compression method for mask images in both Part6 and high compressed PDF.

Applications and Scenarios in Archiving

A major decision when choosing a compressed format for archiving scanned documents is whether the image data should be losslessly stored, ensuring that each pixel in the compressed image is identical to the original scan.

JPEG2000/Part1 offers excellent lossless compression performance, creating files on average three times smaller than uncompressed TIFF format. Higher compression rates are possible when small differences are acceptable, with compression rates of 10-15:1 resulting in visually lossless images. This, together with the progressive and region decoding features of JPEG2000/Part1, make it a good choice for a flexible and efficient lossless and visually lossless archiving format.

Not all users of a digital document archive may require, or be permitted, full access to the high resolution, high quality archived documents. Faster access to a lower quality or lower resolution version of an archived document may, in certain circumstances, be preferable. An intelligent image viewer can achieve this by using the progressive features of JPEG2000/Part1 to decode only the required resolution or quality.

Unfortunately, when dealing with documents containing text, a high quality and high resolution is often necessary to generate an image containing legible text. The amount of data required to be transferred and decoded to generate such an image may be prohibitively large for many applications.
In addition there are circumstances where direct access to the digital document archive may not be possible (or desirable) and distribution of the large lossless JPEG2000/Part1 files, for example by email, is not practical.

One possible solution is to allow access to a highly compressed Part6 file based on the high quality JPEG2000/Part1 file. The Part6 file would combine text legibility with a small file size, allowing fast access to the archived documents. Direct access to the documents in the digital archive would no longer be necessary, increasing security and making distribution of the highly compressed Part6 files more practical.

Part6 files can be created by transcoding JPEG2000/Part1 files, using the progressive features to generate Part6 files of an appropriate quality and resolution. Single JPEG2000/Part1 files can also be grouped together in a multi-page Part6 file, enabling convenient access to related archived documents. The common JPEG2000 file format family used by both Parts also allows any meta-data in the JPEG2000/Part1 files to be simply reused in Part6 files without any format conversion.

The Part6 file format is fairly new and in some cases it may be important to use a more commonly supported format to access archived documents. The wide distribution of Adobe's Acrobat Reader, together with the mixed raster content support, make highly compressed PDF a suitable alternative as an efficient access file format for digital document archives.

Conclusion

We described Part1 and Part6 of the JPEG2000 ISO standard. The flexible lossless and lossy compression support offered by the JP2 file format of JPEG2000/Part1 make it ideally suited as an integrated storage and access format in digital document archives. A losslessly compressed file can act as a “digital original” in the storage. Tools and applications that make intelligent use of the progressive and regional decoding features of JPEG2000/Part1 can also use the digital original as a convenient access format, with fast zooming and panning of the document possible without having to download or decode the complete stored compressed data.
The JPM file format defined by JPEG2000/Part6 is ideally suited as access format for digital document archives. It

achieves very high compression ratios, even higher when JBIG2 is used to compress the bi-level mask layers, while still maintaining good visual quality and preserving fine detailed structures, including text. JPM files are very small and therefore very fast to download and to send by email. JPM to JP2 transcoder tools can use the JPEG2000/Part1 features such as progressive decoding to control the type of access to the stored digital original, limiting, for example, the quality or the resolution of the access view. In some archiving scenarios this is more secure than working directly with the digital original. As both formats JP2 and JPM are part of JPEG2000, the usage of metadata is identical and not an issue while transcoding. JPM can also be used as a storage format in archiving scenarios where no lossless compressed digital original need be stored. In this case, pure JPM based archives without a JP2 file of the same data are used.

An alternative to the native JPM file format of JPEG2000/Part6 are the JPM-style PDF solutions introduced by LuraTech, where high compressed PDF files are generated by applying the methods of MRC and JPEG2000/Part6 to the uncompressed data and then storing the results into a standard PDF file. The advantage of this solution over JPM is the wide availability of the Acrobat Reader. Version 6 or higher of the standard Acrobat Reader can be used as a viewer for JPM-style PDFs or, with some quality limitations, even using version 5. The reliance on Acrobat Reader does, however, have some drawbacks. JPM viewers are generally smaller, faster to start and more focused on the required specialized MRC compression support.

The choice between the use of the native JPM file format of JPEG2000/Part6 or the version framed in a PDF file depends on the application requirements and intended user group, with JPM providing powerful and flexible MRC support while PDF offers a convenient and widely supported format.

References

1. ISO/IEC 15444-1:2000, Information technology - JPEG 2000 image coding system - Part 1: Core coding system, www.iso.ch
2. D.Santa-Cruz, T.Ebrahimi, J.Askelof, M.Larsson, C. Christopoulos, "JPEG2000 still image coding versus other standards", SPIE's 45th annual meeting, San Diego, August 2000, Vol.. 4115, pp.446-454.
3. ISO/IEC 15444-6:2003, Information technology - JPEG 2000 image coding system - Part 6: Compound image file format, www.iso.ch
4. ISO/IEC 16485:2000, Information technology – Mixed Raster Content (MRC), www.iso.ch
5. K.Jung and T.Zellman, "JPEG2000/Part6 for Scanned Documents in Archiving Applications", IS&T Archiving Conference 2004, San Antonio, April 2004, pp.281-285.
6. PDF Reference, Third Edition, Adobe Portable Document Format Version 1.4, www.adobe.com.
7. PDF Reference, Fourth Edition, Adobe Portable Document Format Version 1.5, www.adobe.com.
8. ISO/IEC 14492:2001, Information technology – Lossy/Lossless Coding of Bi-level Images, www.iso.ch
9. P.Howard, F.Kossentini, B.Martins, S.Forchhammer, W.J.Rucklidge, F.Ono, "The Emerging JBIG2 Standard", IEEE Trans. on Circuit and Systems for Video Technology, September 1998, Vol. 8, No. 5, pp. 838-848.
10. Y.Ye and P.Cosman, "Dictionary design for text image compression with JBIG2", IEEE Trans. on Image Processing, June 2001, Vol. 10, No. 6, pp. 818-828.

Biography

Simon McPartlin was born in Scotland in June 1970. He studied Computer Science at the University of Edinburgh, receiving an honours degree in 1992. His interests include digital image processing and still image compression. He is a member of the JPEG group and co-editor of the JPEG2000/Part6 International Standard.

Carsten Heiermann participated in the founding of LuraTech's precursor company Algo Vision in Berlin in 1997 and was promoted to the position of Managing Director in 2000. He joined the Algo Vision plc. Group in 1995. Carsten Heiermann was responsible for the merger of LuraTech with Algo Vision's Berlin operations in 2001 and is one of the shareholders of LuraTech since the Management-Buy-Out in 2004. Prior to joining LuraTech Carsten Heiermann worked in the technical planning department of the cable TV company RKS-Rhein-Ruhr and as a database and software developer with responsibility for control techniques for different industrial customers. He has his technical background in software engineering and holds a degree in electrical engineering, specialized in information techniques.