New Technology for Raster Document Image Compression

Kai Uwe Barthel, Simon McPartlin, Michael Thierschmann
LuraTech GmbH, Rotherstraße 20, 10245 Berlin, Germany

ABSTRACT

This paper describes in detail the LuraDocument technique, a recently developed, high performance technique for compressing and archiving scanned documents, particularly those containing text and image. LuraDocument offers higher compression rates and quality in comparison to traditional document compression methods, preserving text legibility even at extremely high compression rates. This various stages of LuraDocument compression are described in detail, including image quantization and text detection procedures.

Keywords: Document compression, wavelets, mixed raster content, text detection, adaptive image quantization

1. INTRODUCTION

High quality, scanned documents are increasingly used in a wide range of applications, including archiving systems and document management systems. The storage requirements of uncompressed high quality color scanned documents are large; a standard DIN A4 color document, scanned with a resolution of 300 dpi, requires around 25 million bytes of storage space, a 600 dpi scan around 100 million bytes. The storage or transfer of documents of these sizes is, for many applications and environments, impractical. The use of compression in such applications and environments is essential.

Current compression techniques can be classified as either “lossless” or “lossy”. Lossless compression techniques, such as the lossless mode of the JPEG Standard (JPEG-LS) or the Lempel-Ziv Welch (LZW) algorithm, guarantee that the reconstructed data is identical to the original data but offer only low compression rates. The compression rates achievable using lossy compression techniques are higher, with the reconstructed data only approximating the original data.

Traditional image compression schemes assume that image signals are characterized by high spatial correlation. For many scanned documents, particularly those containing text, this assumption is incorrect. The use of such image compression schemes at high compression rates strongly alters the textual parts of documents, leaving the text illegible.

A common solution is the black and white scanning of documents and storage using the lossless CCITT fax compression standards such as “Fax Group 3” or “Fax Group 4”. This enables a high compression rate to be achieved and preserves text legibility but results in the loss of contrast and color information.

An alternative strategy for document compression is the described in the proposed Mixed Raster Content Standard (MRC) (ITU Recommendation T.44). In this planned standard it is possible, a split a document into different regions with different resolutions, and coded with different coders. One mode of the MRC-standard is a multi-layers coding method whereby documents are split into three layers, background, foreground and text mask layers. The proposed MRC-standard exclusively defines the decoding process, leaving the coding process and the generation of the three layers in the multi-layer mode undefined.

The LuraDocument technique describes a system of algorithms for compressing documents according to the proposed multi-layer mode of the MRC-standard.
2. LURADOCUMENT CONCEPT

The LuraDocument compression technique is based on the analysis and segmentation of scanned documents. A bitonal mask layer is generated, classifying each pixel in the original document as either belonging to a text or non-text region. In addition foreground and background images are created to store the intensity or color of the document text and non-text pixels. These three layers are individually coded using appropriate image coders, to generate a LuraDocument data stream.

![Figure 1: Block diagram of the LuraDocument compression stages](image)

2.1. Generation of the bitonal mask

The bitonal mask is generated by first applying an adaptive threshold procedure to the original document, creating a quantized document. The original and quantized documents are then used by a text detection procedure to generate the final bitonal mask. The following sections describe these procedures in detail.

2.1.1. Quantization of the document

Firstly, the original document is converted to a gray-scale image and reduced in size to a resolution of approximately 75 dpi using a low pass filter and sub-sampling. This reduction in size reduces the effect of noise, dither and raster artifacts on the generation of the threshold values, as well as reducing the computation time for document quantization.

A local dynamic activity analysis is carried out by applying a 3x3 minimum filter and a 3x3 maximum filter to the reduced grayscale image. The difference between the minimum and maximum images is used to generate a dynamic activity image. The quantization threshold for each pixel in the dynamic activity image with a value less than a minimum activity value is set to null. For all other pixels the quantization threshold value is calculated as the half the sum of the minimum and maximum image at this pixel position.

These calculated quantization threshold values have a strong local variance and the maximum values are always situated on the brighter part of an edge. An additional 3x3 averaging filter is thus performed, using all non-zero thresholds. These averaged threshold values are further propagated using a 5x5 averaging filter. In the last stage all the pixels with a quantization threshold of null are set using a 7x7 filter which averages all non-null pixels.

The quantization threshold image is increased to the original document dimensions using bilinear interpolation and the bitonal quantization image is finally calculated by comparing each pixel with the quantization threshold value. A pixel in the bitonal quantization image is set to 1 if the quantization threshold value is less than the corresponding grayscale image value, otherwise the pixel is set to 0. Figure 2 shows a typical scanned document and the generated quantized image.
2.1.2. Text detection

The text detection procedure uses the bitonal quantization image and the original document to generate a bitonal mask layer, identifying the text and non-text regions.

Our text detection scheme is based on classifying connected regions with of the same value in the bitonal quantization image. During the text detection procedure these connected regions are identified and analyzed.

The first stage identifies connected regions in the quantization image, using 4-neighbor-connectivity for documents with resolutions greater than 150 dpi, and 8-neighbor-connectivity for documents with a resolution less than or equal to 150 dpi.

Once the connected regions have been identified, the regions are filtered with respect to their size. Maximum and minimum regions sizes, dependent on the documents resolution, are used to eliminate identified regions. All pixels in eliminated regions are classified as non-text.

A non-linear edge detection process is next performed on all pixels on the boundary of connected regions. The edge activity for a boundary pixel is calculated by applying a horizontal Sobel filter, a vertical Sobel filter and a Laplace filter and taking the maximum absolute result. In order to better distinguish edges from raster and noise, the edge activity value is compared with a threshold value. The edge activity of any pixels with edge activities less than a fixed threshold value are set to 0. In a further step the average and maximum edge activity values for the boundary pixels of each of the remaining connected regions are calculated along with the internal variance within the boundary of each region.

Next each region is checked to determine whether the maximum edge activity value is above a fixed maximum threshold value and whether the internal variance is below a maximum variance value. If a region satisfies these two checks then the region is classified as text, unless the region touches a region already classified as text in which case the region is classified as hole. If a region fails either of the checks then the region is classified as non-text.

The bitonal mask layer can be generated once all the connected regions have been classified. All pixels in text regions are marked in the bitonal mask as text pixels, while all pixels in non-text and hole regions are marked as non-text pixels. Note that this text detection procedure is suitable for finding light text on dark backgrounds as well as dark text on light backgrounds. Figure 3 shows the mask layer for our example document.
2.2. Generation of the foreground and background images

The foreground and the background images are generated from the bitonal mask and the original data of the scanned document. Two aspects have to be fulfilled. The foreground and background images have to represent the intensities from the original document as closely as possible and misclassifications in the text detection stage must not lead to visual artifacts. On the other hand the fore- and background images have to be as simple as possible in order to enable an efficient encoding.

The bitonal mask and the original document are used to generate the foreground image. The first stage of this procedure identifies all the text regions. It can be observed that the intensity and color of pixels at the border of a text region are not always typical for pixels within the region. To reduce any problems associated with such border pixels, all text regions are thinned by one pixel, preserving at least the skeleton of each region. The thinned mask image is then used to scale the original document to a ninth of the original size. This is achieved by partitioning the thinned mask image and the original document into 3x3 boxes. The average intensity or color of the thinned text region pixels in each 3x3 box is used to generate the downscaled foreground image. Any downscaled pixels corresponding to 3x3 boxes containing no thinned text region pixels are marked with null. In the next stage a 5x5 averaging filter is used to propagate the average pixel values to any null marked pixels. A 5x5 averaging filter is finally used to damp the foreground color in non-text regions towards gray.

The background image is calculated in a similar way, using the original document and inverted bitonal mask. Figure 3 shows the foreground and background images for our example document.

![Figure 3: Mask, foreground and background image layers](image)

2.3. Coding the individual image layers

Once the mask, background and foreground layers have been generated, they are encoded using appropriate image coders. Suitable coders for the binary mask include the CCITT Fax Group 4 or similar binary coders. The background and foreground layers may be compressed using a standard image coder such as JPEG or a more efficient wavelet coder.

The standard LuraDocument technique uses a special version of the high performance LuraWave wavelet coder for the background and foreground layers, supports Fax Group 4 as the binary mask coder, and uses a TIFF compatible file format, ensuring that at least the mask layer can be decoded by all TIFF compatible applications.
2.4. Decoding a compressed document

The document can be reconstructed by decoding each of the three image layers and using the bitonal mask to select the color for each pixel from either the foreground or background images. Visually pleasing results can be obtained by first rescaling the background and foreground images to the original document dimensions using a bilinear interpolation.

![Figure 4: LuraDocument decompression](image)

3. SUMMARY

The LuraDocument technique uses document quantization, text detection and a variety of image processing procedures to provide extremely high quality document compression and high compression rates. The described LuraDocument technique successfully combines the high compression rates achievable with bitonal coders together with the high intensity and color quality offered by standard image coders. The following figures clearly illustrate the advantages of the LuraDocument technique over standard document coding techniques.