International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 1 103 - 108

Scanned Document Compression Technique

Deeksha kumari
M.Tech Scholar, *CS&E* Branch,
Govt. Women Engineering College
Ajmer, Rajasthan, India
meetasharma15@rediffmail.com

Meeta Sharma
Assistant Professor, Dept. of CS&E
Govt. Women Engineering College
Ajmer, Rajasthan, India
deeksha.choudhary695@gmail.com

Ankur Raj
Assistant Professor, Dept. of CS&E
JECRC College
Jaipur, Rajasthan, India
ankur.raj.10nov@gmail.com

***Abstract*—** These days' different media records are utilized to impart data. The media documents are content records, picture, sound, video and so forth. All these media documents required substantial measure of spaces when it is to be exchanged. Regular five page report records involve 75 KB of space, though a solitary picture can take up around 1.4 MB. In our paper, fundamental center is on two pressure procedures which are named as DjVU pressure strategy and the second is Block-based Hybrid Video Codec. In which we will chiefly concentrate on DjVU pressure strategy. DjVu is a picture pressure procedure particularly equipped towards the pressure of checked records in shading at high determination. Run of the mill magazine pages in shading filtered at 300dpi are compacted to somewhere around 40 and 80 KB, or 5 to 10 times littler than with JPEG for a comparative level of subjective quality. The frontal area layer, which contains the content and drawings and requires high spatial determination, is isolated from the foundation layer, which contains pictures and foundations and requires less determination. The closer view is packed with a bi-tonal picture pressure system that exploits character shape similitudes. The foundation is compacted with another dynamic, wavelet-based pressure strategy. A constant, memory proficient variant of the decoder is accessible as a module for famous web programs. We likewise exhibit that the proposed division calculation can enhance the nature of decoded reports while at the same time bringing down the bit rate.

Keywords- *Segmentation, Compression, Image Segmentation, MRC Compression, Multiscale Image analysis.*

*****

I. INTRODUCTION

Archive pictures have regularly ended up simpler and less expensive to control in electronic structure than in paper structure. Conventional libraries are turning out to be progressively advanced as the expenses of filtering and stockpiling are declining. With the summed up utilization of email and the Internet, the favored approach to convey records is electronic, and the favored presentation medium is quick turning into the PC screen.

A run of the mill page from a book, magazine, or old record examined in shading at 300dpi contains on the request of 8 million pixels, and involves 24MB uncompressed. Conventional pressure methods, for example, JPEG are famously wasteful on a few checks:

1. Typical file sizes for a page will be between 400kb and 2MB at the best, which is totally impractical for remote access.
2. Sharp edges (such as character outlines) are the cause of numerous wasted bits and or unpleasant ringing artifacts.
3. Such large images are very slow to render, require a very large memory buffer for the decomposed image in the client, and are not easily zoom able or pan able with current web browser technology.
4. The text is not normally separated from the image and therefore cannot be indexed or searched.
5. No provision is made for multipage documents, unless one encapsulates the images into container format such as PDF, thereby adding additional layers of inefficiencies.

The DjVu system alleviates these problems and can handle bitonal documents, low-color (palettized) images, photos and other continuous-tone images, scanned color or grayscale documents, as well as digitally produced documents (from PostScript or PDF) [1].

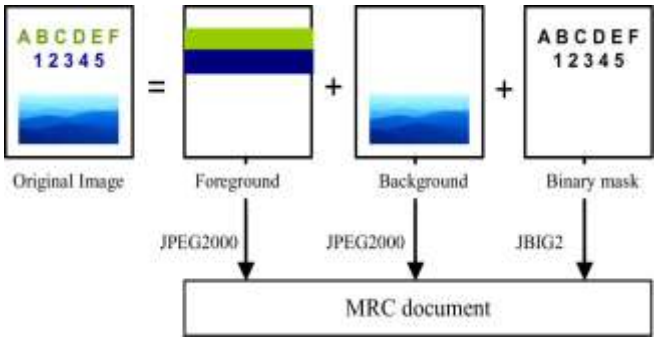

Figure 1: Layers in a MRC mode

Bitonal archives are encoded with a strategy named JB2, which assembles a library of rehashing shapes in the record, (for example, characters), and codes the areas where they show up on every page. Low-shading pictures are packed the same path, with the expansion of a shading palette, and a shading file for every shape. Ceaseless tone pictures are compacted with a dynamic wavelet-based strategy named IW44 that is keeping pace with JPEG2000 as far as sign to commotion proportion, yet whose decoder/renderer is

exceptionally memory proficient, and to a great degree quick (3 times speedier than the quickest JPEG-2000 mode) . Checked shading reports are decayed into a frontal area plane and a foundation plane.[2] The frontal area plane contains the content and the line drawings packed as a bitonal or low-shading picture at greatest determination (utilizing JB2), in this way saving the sharpness and meaningfulness of the content. The foundation plane contains the photos and paper surfaces compacted at diminished determination with IW44. Ranges of the foundation secured by closer view segments are easily added in order to minimize their coding cost.[3]The forefront/foundation sectioned first identifies pointedly differentiated territories, and after that changes them with a few criteria, for example, their shading consistency, their geometry, and an estimation of their coding cost.

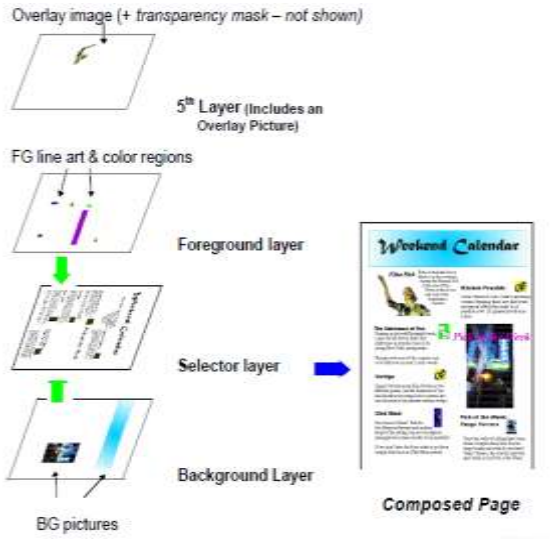

Figure2: DIR Multilayer representation example

In figure 1 we can see the different layers that are present in a document, if we bifurcate them there are basically 3 types of layers: 1.Foreground Layer 2. Selector Layer and 3. Background Layer.

II. LITERATURE REVIEW

We at first made a review out of existing progressions of altered visual assessment of works that are starting now achieved for the Compression of records or a picture document. We will now see the works that have been as of now done on this procedure. Picture based report exchange underpins filtered or electronic archives.[4] New and "legacy" sources, Guarantee appearance and format, fast rendering for survey and printing. Shading or grayscale examined archive pages can't be packed well utilizing standard systems E.g., JPEG compacted report pictures remain extensive, and the content zones are not very much saved Known pressure plans work preferable on some page components over on others. For instance, JPEG pressure procedure is sufficient for pictures, MMR pressure just takes a shot at double content, and Images typically take a gander at lower determination.

JPEG at 100dpi DjVu

Figure 3: The file sizes for the complete pages are 82 KB for JPEG and 67 KB for DjVu.

III. OBJECTIVE OF OUR WORK

Although there are various techniques have been already developed for the compression of image or file, our work is done on the following points.

1. Implementation of an improvised mathematical model optimized for scanned document compression.
2. Use of prebuilt signal processing operation such as DWT (Discrete Wavelength Transmitter) and DCT(Discrete Cosine Transmitter)
3. Creation of a new file format (non-image) for highly compact transmission of scanned document over internet.
4. Proposed file system will have an extension .SDC
5. As .SDC file will contain only numerical parameters which are frequency components of images. It will impossible to intersect without application of same sequences of inverse transfer and thus will provide an additional layer of security.
6. Use of extensive tool such as MATLAB eliminates the need for other decoding software such as DJVU.
7. Reduction of space and time complexity of exiting algorithm.
8. Our algorithm will be indifferently applicable on grayscale as well as color image also we indeed to provide GUI driven quality control option for user.

IV. PROPOSED WORK

Our test setup is essentially programming based framework in which a UI is produced in MATLAB which is appeared in Figure 3, through this video test is taken as information and experienced specific number of steps that we will see quite recently and after that we will acquire the article picture which is without shadow and through we can undoubtedly decide the real measurements of the item.

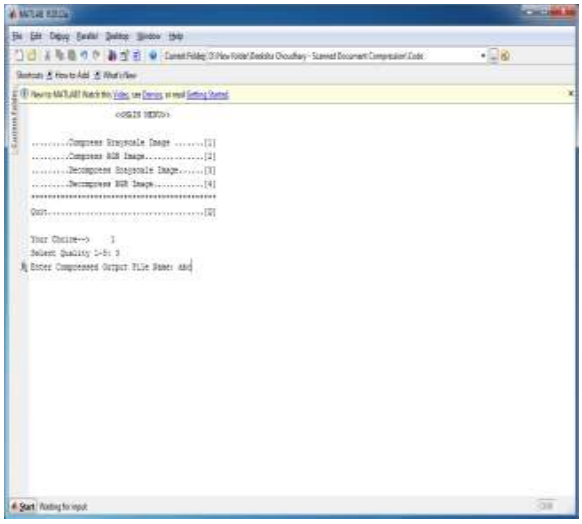

Figure 4: UI of the system

Figure 4 shows the UI of the system , which facilitates us for 4 major options:

a) *To Compress Grayscale Image*
b) *To compress RGB Image*
c) *To Decompress Grayscale Image*
d) *To decompress RGB Image*

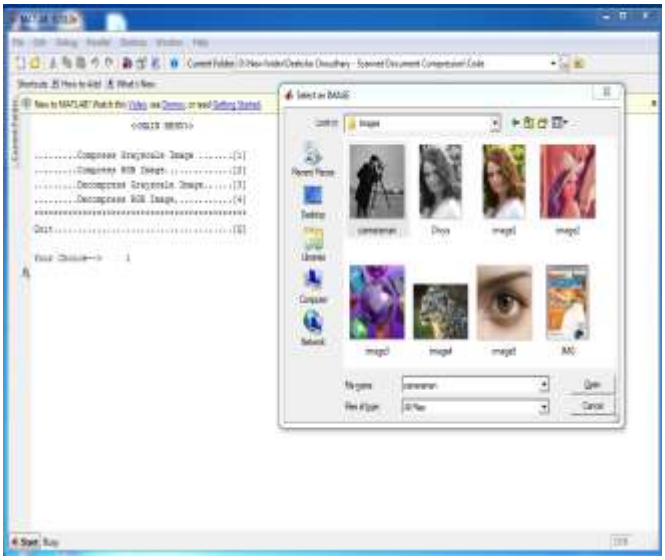

Figure:5: Selection of Image for conversion

The process of the software module is characterized in various numbers of steps which is mentioned below:

1.In the first step we have given four options as we have mentioned above either to compress or decompress a Grayscale or a coloured image.
2. For any of the method first we have to select an image from the source.
3. After selection of the proper image it will compress the image pixels and thus the size will get reduced.
4. At last we will obtain the same image with different image size.

Above mentioned steps are very important and just with the help of this image we can find out the shadow compressed image file.

V. METHODOLOGY AND PROCEDURE ADOPTED

Our proposed algorithm is mainly based on MATLAB and its GUI gives a user friendly environment through which any user can compress any colored or gray scale image easily and in 4-6 times smaller size. We will now explain our technique that we are going to develop for the Scanned Image compression. Mainly our research is based on the following points.

1. We will develop a new file extension for our own algorithm to store compressed images.
2. Use of frequency domain tools such as DWT (Discrete wavelength transfer) and a DCT (Discrete cousin transform) in compilation with code book method to achieve high compression ratio for internal distortion and streaming.
3. Decrease space up to four time complexity of existing algorithm to achieve higher performance of low end system.
4. Our GUI enable user to select the image type which can be either RGB (Colored) or Gray scale and the user can control the compression parameter either by manually baring the compression algorithm.
5. Variable compression ratio as peruse required or constraints such as bandwidth limitation, size limit (Internet size limit) or fit on a embedded media for educational distribution.

Above mentioned points describe the points on which we are going to work. Now we will see the work flow of our system.

- DECODING A COMPRESSED DOCUMENT

In our Methodology we used to decode a compressed document in following way. 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.

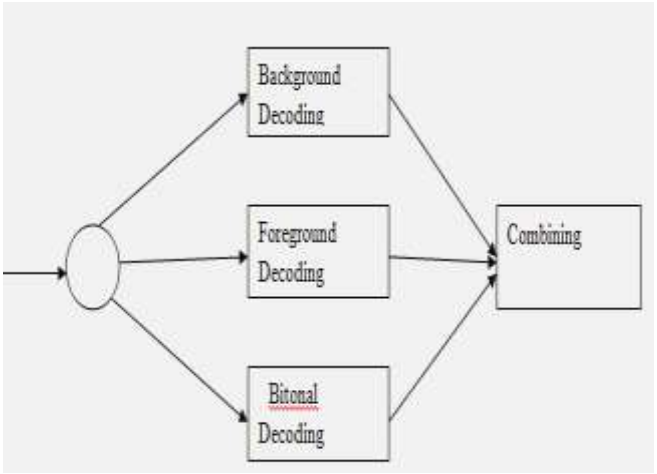

Figure 5: Document Compression Technique

➢ DIFFERENT COMPRESSION METHODS

1. The **DjVu document image compression** technique responds to all the problems as Magazine compression/document compression, Image compression and some more. [5]With DjVu, pages checked at 300dpi in full shading can be compacted down to 30 to 80 KB documents from 25 MB firsts with superb quality. This puts the span of fantastic examined pages in the same request of size as a normal HTML page (44 KB as indicated by the most recent measurements).[6] DjVu pages are shown inside of the program window through a module, which permits simple panning and zooming of huge pictures. The fundamental thought behind DjVu is to isolated the content from the foundations and pictures and to utilize distinctive strategies to pack each of those segments. Customary strategies are either intended to pack regular pictures with few edges (JPEG), or to pack highly contrasting archive pictures altogether made out of sharp edges (CCITT G3, G4, and JBIG1). The DjVu strategy enhances both and consolidates the best of both methodologies. A closer view foundation division calculation produces and encodes three pictures independently from which the first picture can be recreated: the foundation picture, the forefront picture and the veil picture. The initial two are low-determination shading pictures (for the most part 100dpi), and the recent is a high-determination bi-level picture (300dpi). A pixel in the decoded picture is built as takes after: if the comparing pixel in the veil picture is 0, the yield pixel takes the estimation of the relating pixel in the fittingly up inspected foundation picture. In the event that the veil pixel is 1, the pixel shading is picked as the shade of the joined segment (or taken from the closer view picture). The frontal area foundation representation is likewise a key component of the MRC/T.44 standard.

2. **Block Based Video Codec:** Pressure of examined archives can be dubious. The checked archive is either packed as a ceaseless tone picture, or it is binarized before pressure. The twofold record can then be compacted utilizing any accessible two level lossless pressure calculation, (for example, JBIG and JBIG2), or it might experience character acknowledgment.[7] Binarization might bring about solid debasement to question forms and surfaces, such that, at whatever point conceivable, nonstop tone pressure is favored. In single/multi-page archive pressure, every page may be independently encoded by a few persistent tone picture pressure calculations, for example, JPEG or JPEG2000. Multi-layer methodologies, for example, the blended raster content (MRC) imaging model are additionally tested by delicate edges in filtered reports, frequently requiring pre-and post-handling. Normal content along a record regularly introduces dull images such that lexicon based pressure routines turn out to be extremely productive. For constant tone symbolism, the repeat of comparative examples is Nevertheless; a proficient word reference construct encoder depending in light of consistent tone design coordinating is not that trifling. We propose an encoder that investigates such a repeat through the utilization of example coordinating indicators and effective change encoding of the remaining information.

➢ WORK FLOW OF IMAGE COMPRESSION TECHNIQUE

In the above sections we had explained what are the main issues on which we are going to focus now we will see what are the steps through which a common Image compression technique is used or what are the basic techniques that we will use.

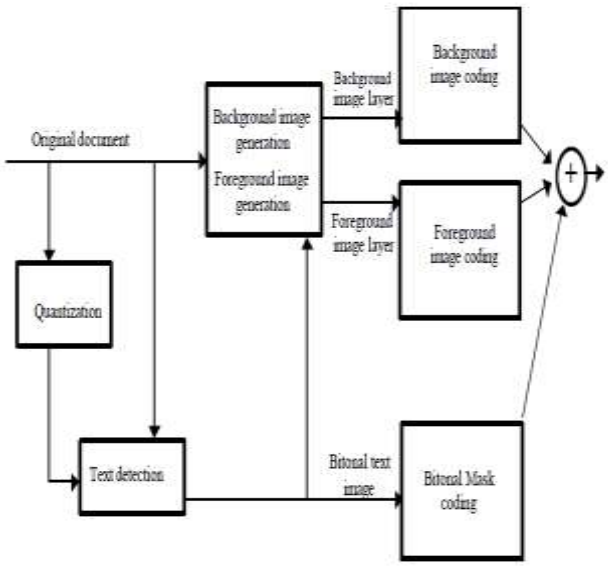

Figure 6: Work Flow of Image Compression Technique

This can be easily explained with the help of a work flow diagram which includes the initial step from input of original document to the final compressed document. At first the original document quantized and then it detects the text and image, later on foreground and background image generation will occur which is the most important step and in the final step coding is done of main three things i.e Background Image Coding, Foreground image coding and Bitonal Mask coding.

VI. RESULTS AND IMPLEMENTATION

In this section we will discuss the overall implemented result of our running system. Each output and the process of the system will be observed with the relevant snap-shots. Let's study each of them steps vise. UI which has been used in this Experiment is MATLAB based software. The software that we have developed here is named as "Scanned Document Compression Technique". And the complete process of our research is involved in this software step wise which begins from the input of the picture sample or a file sample.

As we know in our system it's completely custom automated, means here user can input the Image file and can compress or decompress the Image file as per as usability. One can also have the choice of the quality, that what quality of image he wants.

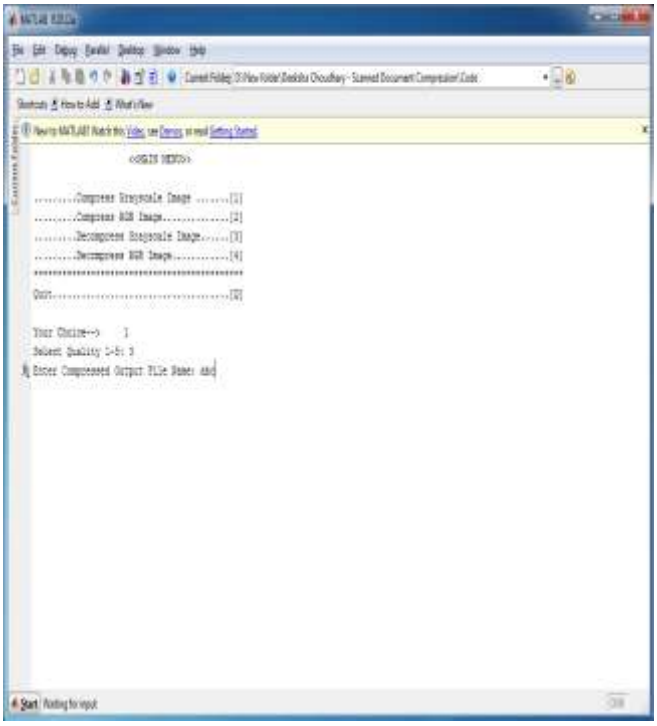

Figure 7: Selection of suitable image manually

Above figure facilitates the user to get the perfect image selection and the quality selection and also it will ask about the type of image, i.e. either grayscale or color.

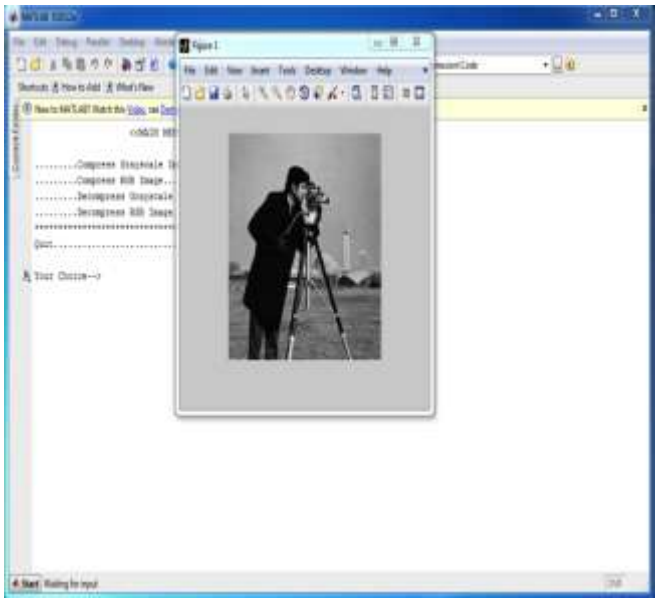

Figure 8: Compression of a Grayscale Image

After perfect compression the size of the image is reduced as per as required, which is to be done by compression of the pixels of the image which is arranged in the form of Matrix.

VII. CONCLUSION

We exhibited a novel division calculation called SMART for filtered, complex report pictures going for effective pressure. Division are arranged into binarizable and non-binarizable segments, where encoding plans suitable for their sorts are utilized. Keen can deal with picture segments of different shapes, numerous foundations of diverse dark levels, distinctive relative grayness of content to the foundation, tilted picture parts, and content of diverse dim levels. It includes preprocessing stage, where a shading space change may be performed, square arrangement into dynamic pieces and dormant pieces, macroblock development which gathers dynamic squares and macroblock grouping binarizable and non-binarizable macroblocks. Its adequacy in division and its advantages to pressure is illustrated.

DjVu, another pressure procedure for shading report pictures is depicted. It fills the crevice between the universe of paper and the universe of bits by permitting filtered report to be effortlessly distributed on the Internet. With the same level of intelligibility (300 specks for each inch), DjVu accomplishes pressure proportions 5 to 10 times higher than JPEG. DjVu can likewise be considered as an empowering innovation for some report examination procedures. To accomplish ideal pressure, it legitimizes the improvement of

complex content/picture partition calculations. The expansion of content design examination and optical character acknowledgment (OCR) will make it conceivable to file and alter content separated from DjVu-encoded reports.

VIII. FUTURE WORKS

Since we know that the experiments and research have no end points so we can have some future works.

1. Electronic document should be predominate

1. Electronic documents predominate

- Most authoring done with computers
  - MS office, e-mail, web, latex..
  - Text generation: bills, form letters…
- Most documents exchanged electronically
  - Web(news sites, scientific publication, government publication, public and business form….)
  - E-mail and email attachments
  - Groupware and document repositories
- Some laggards
  - Books(DRM concerns), legal documents and bill presentment (legal issues, reliability, user reluctance…)

Structure electronic documents

- Used by office suites, web browsers, presentation packages, form…)
- Contains
  - The text and its reading order
  - Annotation about the logical functions of chucks of text (heading, page number, title, author, etc….)
  - Annotations about appearance (italics, bold, font size, etc.)
- Semantics of content formally specifies
- Examples
  - HTML, XML, Latex, MS word

Images- based electrons documents

- Obtained by scanning, temporarily created during printing, screen display
- Can represent arbitrary images
  - Usually pixel- based(but could be vector-based ,e.g. link)
  - Little or no information about reading order, logical fuctation.
  - May contain text for searching, but the images is what the user sees

Semantics determined by user's interpretation

IX. REFERENCES

[1] R. N. Ascher and G. Nagy. A means for achieving a high degree of compaction on scan-digitized printed text. *IEEE Trans. Comput.*, C-23:1174–1179, November 1974.

[2] L. Bottou, P. Haffner, P. G. Howard, P. Simard, Y. Bengio, and Y. LeCun. High quality document image compression with djvu. *Journal of Electronic Imaging*, 7(3):410–428, 1998.

[3] L. Bottou, P. G. Howard, and Y. Bengio. The Z-coder adaptive binary coder. In *Proceedings of IEEE Data Compression Conference*, pages 13–22, Snowbird, UT, 1998.

[4] L. Bottou and S. Pigeon. Lossy compression of partially masked still images. In *Proceedings of IEEE Data Compression Conference*, Snowbird, UT, March-April 1998.

[5] P. G. Howard. Text image compression using soft pattern matching. *Computer Journal*, 40(2/3):146–156, 1997.

[6] W. N. J. Sheinvald, B. Dom and D. Steele. Unsupervised image segmentation using the minimum description length principle. In *Proceedings of ICPR 92*, 1992.

[7] MRC. Mixed rater content (MRC) mode. ITU Recommendation T.44, 1997.

[8] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[9] G. Story, L. O'Gorman, D. Fox, L. Shaper, and H. Jagadish. The RightPages image-based electronic library for alerting and browsing. *IEEE Computer*, 25(9):17–26, 1992.

[10] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes:Compressing and Indexing Documents and Images.*Van Nostrand Reinhold, New York, 1994.

[11] L. Bottou, P. Haffner, P. G. Howard, P. Simard, Y. Bengio, and Y. LeCun. High quality document image compression with djvu. *Journal of Electronic Imaging*, **7(3)**, pages 410–428, (1998).

[12] MRC. Mixed rater content (MRC) mode. ITU Recommendation T.44, (1997).

[13] L. Bottou and S. Pigeon. Lossy compression of partially masked still images. In *Proceedings of IEEE Data Compression Conference*, Snowbird, UT, March-April (1998).

[14] L. Bottou, P. G. Howard, andY. Bengio. The Z-coder adaptive binary coder. In *Proceedings of IEEE Data Compression Conference*, pages 13–22, Snowbird, UT, (1998).

[15] E. H. Adelson, E. Simoncelli, and R. Hingorani. Orthogonal pyramid transform for image coding. In *Proc. SPIE vol 845: Visual Communication and Image Processing II.*, pages 50–58, Cambridge, MA, October 1987.

[16] J. M. Shapiro. Embedded image coding using zerotrees of wavelets coefficients. *IEEE Transactions on Signal Processing*, **41**, pages 3445–3462, December (1993).

[17] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Journal of Applied Computing and Harmonic Analysis*, **3**, pages 186–200, (1996).