International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June-2018 www.irjet.net p-ISSN: 2395-0072

Compression of Old Manuscript Document Images Using Hybrid Encoding Technique

Pallavi S. Metkar[1], S. S. Thakare[2]

[1]*Electronics and Telecommunication Department, GCOE, Amravati (MH), India*
[2]*Assistant Professor, Electronics and Telecommunication Department, GCOE, Amravati (MH), India*

Abstract - *The word manuscript is derived from the Latin "manu scriptus", meaning "written by hand". It is an old document that was written by hand before the printing invented. As the time passes deteriorations in such documents increases and thus the preservation of manuscript is important. Hence such type of documents is scanned at the very high resolution to capture information correctly. This paper presents the compression technique for old manuscript document images. Here, illustrated manuscripts are taken. First the image is converted into YCbCr and then each layer is separated. The separated layers are segmented using Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) and statistical redundancies are removed using arithmetic coding. This hybrid combination of compression techniques gives better Compression Ratio (CR) and Peak Signal to Noise Ratio (PSNR) at the acceptable quality of the reconstructed manuscript images.*

Key Words: **Manuscript document image, DWT, DCT, arithmetic coding, CR, PSNR.**

1. INTRODUCTION

Manuscript documents are the witness of past occurred events. But, with time there is much degradation in such documents, hence preservation of such documents becomes important. Manuscripts are scanned at high resolution to capture information correctly and then stored in the digital format. Automatically, the file size increases along with memory requirement and the costing increases. By considering these requirements, the proposed method presented the efficient compression algorithm to compress manuscript document images. The paper focuses not only on good compression ratio and PSNR but also on the quality of the reconstructed document image.

There are two types of manuscript document images: illustrated manuscript document image and the non-illustrated manuscript document image. Non-illustrated manuscript document images have text on the other hand illustrated document images have text and the graphics part. Compression techniques are available for the non-illustrated manuscript. But for the illustrated manuscript document images a very little work is found. Hence the main focus of the presented paper is on illustrated document images. The proper compression technique is needed to compress illustrated manuscript document images.

Data compression is the system designed to remove the statistical redundant bit from an image [1]. Data compression estimates the statistical redundancies which ultimately increases the accuracy. In order to make historical document images precise, the comprehensive review of image enhancement methods is provided in the paper [2]. Old manuscript document images are preprocessed to enhance the quality of an image [3]. An algorithm is based on content retrieval of the historical manuscript indexing system, developed for the vast content of manuscripts. The documents which are having textual content along with pictures are called compound documents. Such types of documents have taken in [4], here mixed raster content multi-layer technique is used to compress compound documents. But, the results so far are not decisive when the complexity is not acceptable for the given application. An algorithm used in [5] segments a scanned document image into 8x8 blocks of different classes and compresses each class with an algorithm specifically designed for that class. The process explained in this paper is complex and very time-consuming. Paper [6], used hybrid pattern matching/ transform based compression method for scanned documents. This is the latest technique used for the compression of scanned document limited to English language document only. Paper [7], presents B-splines curve fitting method is used to compress the handwritten documents. Splines are the special function defined piecewise by polynomials. B-splines are very simple for their reconstruction and capacity to approximate complex shapes. B-splines also minimize errors.

The new compression strategy based on multilayer segmentation is presented. Each layer is compressed differently with different degradation criterions to get fine representations. The results demonstrated give much higher compression and the PSNR with the acceptable quality of the reconstructed image.

2. MANUSCRIPT DOCUMENT IMAGES OVERVIEW

Man-yoo-script - means text written by hands. The word manuscript is derived from the Latin manu scriptus, meaning "written by hand". It stands for a very old document that was written by hand before printing was invented or reproduced in some other way. If the scripts are written on palm leaves then those are named as palm-leaf manuscripts and when on paper those are called paper manuscripts. The term manuscript may also be used for information that is hand-recorded, other ways than writing for instance, inscriptions

that are chiseled upon durable material or scratched (the original meaning of graffiti) as with the help of knifepoint in plaster or with the help of a stylus on a waxed tablet, (the way Romans made notes), or are in cuneiform writing, impressed with a pointed stylus in a flat tablet of unbaked clay. The handwritten scripts are called pandulipi in Bengali; the material on which they were inscribed was either grey or pale yellow in colour. This is one of the reasons behind the name pandulipi since pandtu stands for yellow in Bangla and lipi for writing.

Figure 1: Old Manuscript document image

A manuscript has certain anatomy; while the chief portion comprised the main body text, the part where the scribe used to flaunt his own wordsmithery comprise of the following parts: Prasasti or namaskriya mangalacaranam - that means the ornamental beginning of die manuscript. Bhonita or preamble: Describes the name of the book and the writer, a variant of Pushpika. Mangala Gatha or Pushpika or colophon: Generally found on the first leaf, at the end of the chapter or act, and on the last folio, the Pushpika is a short, composite work made of a brief autobiography of the writer, die name of the manuscript the date of writing/copying and the name of the person who is the provider of the task. When it is on the first folio or at the end of a chapter or an act the Pushpika contains the names of the scribe and the book whereas in the last folio, the former content is placed along with the dates of writing or copying.

The manuscripts can be classified according to illustration into two categories such as: Illustrated manuscript and the Non-illustrated manuscript. When the manuscripts have pictures or diagrams, then these are called illustrated manuscripts. Such types of manuscripts are also called as compound manuscript. It has two parts, one is the text and the other is graphics. When there is no picture, no diagram but only the scripts are written, then these are called non-illustrated manuscripts.

Discrete Wavelet Transform

DWT is the mathematical tool used to transform the image from one domain to another domain. DWT is used to get different coefficients of image along with the details. Wavelet based technique is best suitable for transmission and error decoding. Wavelet segments an image into the approximated image and the detailed components. The HH sub-image is nothing but the background of the image. 2D-DWT of image with NxM pixels is performed by using following expression:

$$W_\phi(j_0,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}p(x,y)\,\phi_{j_0,m,n}(x,y)$$

where

$$\phi_{j_0,m,n}(x,y)=2^{j_0/2}\phi(2^{j_0}x-m,2^{j_0}y-n)$$

$$W_\psi^i(j,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}p(x,y)\,\psi_{j,m,n}^i(x,y)$$

where

$$\psi_{j,m,n}^i(x,y)=2^{j/2}\psi^i(2^jx-m,2^jy-n)$$

for $i=\{H,V,D\}$

p(x, y) is the pixel value of the image. j0 is the starting scale, the coefficients Wϕ (j0,m, n) define an approximation of $p(x,y)$. The 2D-IDWT is given by,

$$p(x,y)=\frac{1}{\sqrt{MN}}\sum_m\sum_n W_\phi(j_0,m,n)\phi_{j_0,m,n}(x,y)+\frac{1}{\sqrt{MN}}\sum_{i=H,V,D}\sum_{j=j_0}^{\infty}\sum_m\sum_n W_\psi^i(j,m,n)\,\psi_{j,m,n}^i(x,y)$$

In DWT the truncated coefficients usually belong to high frequency sub-images depending on the given image and the family of the wavelet used.

Discrete Cosine Transform

DCT is the real transform based on the set of cosine values. DCT divides an image into N by N block. The advantage of dividing an image into the block is that the pixels exist within the local blocks (i.e. in individual blocks) shows more homogeneity than the pixels exist within the global blocks. Local pixel values show more correlation than that of global pixel values. When the image is transformed the amount of information that will be required to represent an image is excessively high, so some kind of quantization is perform using quantizer in DCT. The quantizers are designed to exploit the psychovisual redundancy. Hence in DCT, quantization is done. The two-dimensional DCT of an M-by-N matrix is defined as follows:

$$Y[j,k]=C[j]C[k]\sum_{m=0}^{N-1}\sum_{n=1}^{N-1}x[m,n]\cos\frac{(2m+1)j\pi}{2N}\cos\frac{(2n+1)k\pi}{2N}$$

Where: j, k = 0, 1, 2,…., N −1 and. The inverse transform is defined as:

$$X[m,n]=\sum_{j=0}^{N-1}\sum_{k=1}^{N-1}C[j]C(k)y[j,j]\cos\frac{(2m+1)j\pi}{2N}\cos\frac{(2n+1)k\pi}{2N}$$

Where: m, n = 0, 1, 2, …, N −1. And c[n] is as it is as in 1-D transformation.

3. PROPOSED METHOD

The compression of the entire image has a major disadvantage, particularly for the document image. The quality of the image may be degraded due to the significant loss of information. We made the compression block by block after extraction and segmentation process. The paper presents the hybrid encoding technique for the compression of old manuscript document images. In the proposed technique color document image is converted into YCbCr and

then each YCbCr components are separated so that each component should be compressed properly without losing information. The statistical redundancies are removed using arithmetic encoding technique. The block diagram for the proposed system is as shown below:

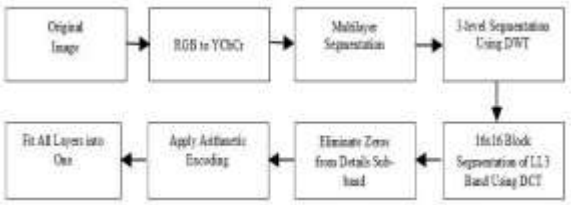

Figure 2: Block diagram of compression technique

RGB color model is suitable for color display not for color scene segmentation and image analysis because of high correlation among the RGB components. It means color components are changes with intensity. Hence YCbCr is used to separate the color information of an image from its intensity information. The rate-distortion performance of YCbCr is better as compared with the other color models. Thus the RGB image is converted into YCbCr. After conversion of RGB to YCbCr, each component separated into luminance, Chrominance-red, and Chrominance-blue. To get better Compression Rate and PSNR each component is compressed separately. Luminance is very similar to the grey version of the original image. Then the image is segmented into sub-images corresponding to their frequency components using DWT, nothing but the system generated segmentation. The approximated image is again segmented into 16x16 blocks using DCT and then compressed using DCT to get higher compression. Zeros eliminated from the details of the sub-band and each sub-image is compressed using arithmetic encoding to remove the statistical redundancies from the document image.

Design Metrics

The performance of the proposed system is based on the following design metrics:

PSNR (Peak Signal to Noise Ratio): PSNR is the ratio of maximum possible power of a signal and the power of corrupting noise. It is expressed in dB. Here, a signal refers the given original image and noise refers to the error occurred in reconstructed image. PSNR is given by,

$$PSNR = 10.\log_{10}\left(\frac{255^2}{MSE}\right)$$

MSE is the mean square error refers to the some sort of average or sum (or integral) of squares of the error between two images.

The MSE is given by,

$$MSE = \frac{1}{MN}\sum_{x=1}^{N}\sum_{y=1}^{M}[p_{xy} - \hat{p_{xy}}]^2$$

Compression Ratio: Compression ratio is defined as the ratio between the original image size and compressed image size.

$$\text{Compression Ratio} = \frac{\text{Uncompressed image size}}{\text{Compressed Image size}}$$

4. EXPERIMENTAL RESULTS

The proposed technique is tested on the several images. The document images test is carried out and run using matlab. Proposed compression technique gives high compression ratio and PSNR. And the reconstructed image is approximately equal to the original.

Figure 3: Original Image

The segmentation is applied up to 3rd level using DWT followed by the steps explained in the block diagram. Finally, all the components combined to one. In the proposed algorithm the compressed data is stored in a file. After decompression we will get the reconstructed image which is approximately same as that of the original image.

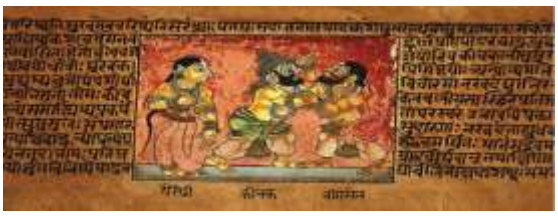

Figure 4: Reconstructed Image

In each test the resultant Compression Ratio and PSNR are calculated, after each test. The following table shows the results of proposed hybrid encoding technique. The experimental result shows high compression ratio. PSNR for the corresponding documents lies between 40-50 dB which is good for manuscript document images.

Sr. no.	Original image size (KB)	Compressed image size (KB)	Compression Ratio	PSNR (dB)
1.	2796.52	186	15:1	41.138
2.	4229.12	247	17:1	43.684
3.	3932.16	251	16:1	43.12

4.	7485.44	286	26:1	45.79
5.	7372.8	326	23:1	44.571
6.	9031.68	460	20:1	47
7.	3932	179	22:1	42
8.	2222.08	114	19:1	40.88
9.	9154.56	475	20:1	46.85
10.	9881.6	364	27:1	47.53
11.	11264	492	23:1	47.51
12.	13209.6	681	19:1	46.59
13.	8335.36	344	24:1	47.89
14.	11878.4	728	16:1	45.46
15.	12492.8	760	17:1	45.71
16.	8663.04	450	19:1	46.45
17.	10956.8	529	21:1	47.46
18.	13209.6	533	25:1	48.23
19.	13107.2	590	22:1	47.86
20.	17510.4	852	21:1	46.46
21.	8826.88	389	23:1	46.93
22.	7004.16	322	22:1	44.63
23.	9113.6	451	20:1	48.46
24.	10444.8	459	23:1	46.38
25.	8355.84	460	18:1	45.40
26.	14028.8	637	22:1	47.48
27.	15360	706	21:1	47.56
28.	13824	656	21:1	47.29
29.	9004.44	449	20:1	48.25
30.	8032.69	357	22:1	45.13

Table: Results of document image compression using Hybrid Encoding technique

5. CONCLUSIONS

The main objective of this paper is to compress old manuscript document images, with the great concern of preservation than achieving only high compression ratio. Generally, preprocessing of manuscript images involves quality enhancement through noise removal. Quality images need large memory and high bandwidth for storage and transmission that adversely affects transmission speed. Thus the compression scheme has become mandatory to be implemented on such type of document images. We used a hybrid algorithm to fulfill the aim of preserving all quality related components of the document. Compression becomes challenging when it deals with degraded manuscript images. Hence the efficient compression technique is proposed for the compression of manuscript document images. The proposed hybrid compression technique gives high compression ratio and also good PSNR values. The compression rate achieved by this technique is ranging between 90% to 95% and PSNR for the same ranges 40dB to 50dB.

REFERENCES

[1] Jorma Rissanen, "A Universal Data Compression System", IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT - 29, NO. 5, SEPTEMBER 1983.

[2] Sitti Rachmawati Yahya, S. N. H. Sheikh Abdullah, K. Omar, M. S. Zakaria, C. Y. Liong, "Review on Image Enhancement Methods of Old Manuscript with the Damaged Background", "International Conference on Electrical Engineering and Informatics", 5-7 August 2009.

[3] Wasfi G. Al-Khatib, S.A. Shahab, Sabri A. Mahmoud, "Digital Library Framework for Arabic Manuscripts", 2007 IEEE.

[4] D. Mukherjee, N. Menon, A. Said, "JPEG-matched MRC compression of compound documents", IEEE conference, 2001.

[5] Hui Cheng and Charles A. Bouman, "Multilayer Document Compression Algorithm", 1999 IEEE.

[6] Alexandre Zaghetto, Ricardo L. de Queiroz, "Scanned Document Compression Using Block-Based Hybrib Video Codec", IEEE conference VOL. 22, NO. 6, JUNE 2013.

[7] Kamal Gupta, Manish Bansal, SantanuChaudhury, "A Compression Scheme for Handwritten Patterns Based on Curve Fitting", 2011 IEEE.

[8] Umesh P. Akare, DR.N.G.Bawane, "Efficient Compression of Manuscript Images", 06 April, 2017