International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

A Review of Scanned Handwritten Document Compression in Devnagari script

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 - *A large number of documents of different languages are available on the server as the use of digital libraries increases day by day. The memory requirement for these documents is large. The speed of communication becomes slow if we want to send the document of large size from one server to another. Document image compression is the key factor used for speedy communication over the network. Also, the memory size requirement is less as compared to the original one. For easy accessing document image compression performs an important role. In the context of document image compression lot of work is done for the printed textual image of different languages. Less work is reported yet for the compression of handwritten document images. Hence this is the motivation of behind the work, to develop efficient compression technique for handwritten document image compression.*

Key Words: **Digital libraries, image compression, textual image, handwritten document, compression technique.**

1. INTRODUCTION

Digital libraries are very popular nowadays. A number of scanned documents are available in digital libraries and publish over the web. To store all these documents we require a large space for storage and minutes of time required for accessing these documents. If the size of a file is large then it takes the considerable amount of time, because of this the speed of transmission becomes slow. To avoid all these problems compression is the very important key factor. We can store a large number of the compressed scanned documents in available memory space. The number of scanned documents of different languages is available in printed text or in handwritten text on a server. Compression strategies are available for printed textual document and also for the handwritten textual document of English, Arabic, and Chinese. Very less work is reported for the compression of the handwritten textual document in Devanagari script and all this work is for gray level images only. Sometimes color documents itself contains important information. After conversion documents may loss its original identity. Devanagari script consists of more than 20 languages. Devanagari script may consist of important information. Many handwritten documents are in Indian languages used generally in official work so preservation of these documents becomes important. To store all these documents memory requirements is large and accessing becomes difficult because of file size. As the file increases transmission speed is also increases to transfer the file from one server to another. So to avoid all these difficulties image compression strategy plays an important role. Compression of handwritten Devanagari script is different than that of the printed textual document as the handwriting varies from person to person. Generally old documents in India are available in handwritten Devanagari language. So compression methodology can useful for the preservation of an ancient data in Devanagari script. The absence of any compression methodology for handwritten document color images in the context of Indian language is the motivation behind the present work.

2. RELATED WORK

In paper [1], SPM compression technique is used for the compression of gray level images. The main advantage of SPM is to avoid substitution error occurs in PM & S. proposed system is mainly design for facsimile technology involves scanning and transmitting bi-level images. Proposed technology can also be used for archiving in which scanning and storing of documents done. Described technique is used for both lossy and lossless compression. SPM technique removes all the errors which are occurs due to PM & S technique. In SPM, new symbol added into dictionary without compressing it which takes place in PM & S method.

Paper [2], addresses the problem of compressing text images with JBIG2. Given work proposes two symbol dictionary design techniques: 1.Class-based technique and 2. Tree-based technique. JBIG2 standard is for bi-level image compression. Bi-level images have only one bit plane in which each pixel takes one color out of two. Proposed system gives the comparison between the coding efficiency of PM & S based technique and SPM based technique, also gives the comparison between reconstructed image quality in lossy compression and the system complexity. And proposes the methods to change the class and tree-based dictionary sizes and trace out the bit rate as a function of dictionary size.

Paper [3], presents a new compression scheme for Indian language textual documents images. As OCR may not be able to compress documents due to unavailability of data set in most of the Indian language scripts and even if the data set is available, it is not efficient enough to perform the conversion job. In this paper new compression technique presented first

time for Indian languages. Proposed compression technique lossy in nature, compresses document images up to readable level. This method is based on the symbolic compression technique. Proposed technique has been accomplished with an efficient segmentation-based clustering approach.

In paper [4], proposes compression strategy for handwritten documents or receipts. In this foreground and background compressed separately. Based on the assumption that the foreground represents the most important information hence degraded less while more degradation is allowed for background, as background provides only the sense of reality of the document. To reduce data, foreground, background and color tones are down-sampled automatically and then RLE coding compression is applied to encode sub images.

Paper [5], designed compression technique for low quality color which are in various forms starting from gray-scale to color, printed and handwritten script. This technique is also applicable to those documents which are suffering from degradation like aging, uneven illumination. Here foreground and background are compressed separately to achieve better compression. Numerous works is done for the binarization of gray scale documents. To fill this gap this paper presents the binarization technique for color document. Binarization technique designed for the variety of documents like uneven illumination, noise, aging etc. Proposed technique is applicable to printed as well as handwritten documents.

Focused of paper [6] is on the compression strategy for handwritten document images in Devanagari script which are in gray level. In this paper compression is done by identifying different zones of Devanagari script and then compression technique is applied to different zones to compressed handwritten textual documents. In this paper each step of compression follows different technique, scanned documents must be in gray level image.

Paper [7], proposed the procedure of HCR for different Indian languages. This paper gives the limitations of OCR, as the OCR does not have sufficient data set for the compression of Indian language document images.

Paper [8], used hybrid pattern matching/ transform based compression method for scanned documents. This is the latest technique used for the compression of scanned document in English language. Coder is designed to compress scanned documents.

Paper [9], 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.

In paper [10], hybrid compression method is used for the compression of document. In first phase data is compressed with the help of dynamic bit reduction technique and in second phase Huffman coding is used to achieve better compression to produce final output. Proposed technique based on lossless data compression approach.

3. DEVNAGARI SCRIPT OVERVIEW

A lot of ancient data are available in Devnagari language. Devnagari script is a basic script for many of the language in India such as Hindi, Marathi and Sanskrit. In Devnagari all letters are equal. There is no concept of capital or small letters.

Fig. Non-compound characters in Devnagari script

A syllable is formed with vowel or any combination of the consonants and vowel. Figure 2 shows the sample set of non-compound set of characters.

4. CONCLUSIONS

Many of Devanagari languages consist of historic data so the storage and preservation is important. Also for fast transmission compression technique is very important. Most of the techniques based on JBIG2 standard. This review paper also focuses on the handwritten document image compression which is very important for archiving data. There are many applications where we need handwritten data compression techniques like offices, form documents, historic documents, file transmission, storage.

REFERENCES

[1] P.G.Howard, "Text image compression using soft pattern matching", The Computer Journal, vol. 40,pp.146-156,1997.

[2] Yan Ye and Pamela Cosman. "Dictionary Design for Text Image Compression with JBIG2,2001 IEEE transaction on image processing".

[3] U. Garain, S. Debnath, Compression of Scan-Digitized Indian Language Printed Text: A Soft Pattern Matching Technique

[4] X.Danhua, B.Xudong, "High efficient compression strategy for scanned receipts and handwritten

documents", IEEE International Conference on Information and Engineering,2009, pp.1270-1273

[5] Utpal Garain, Thierry Paquet, Laurent Heutte, "On foreground – background separation in low quality document images", International Journal of Document Analysis (2006) 8(1): 47–63

[6] Smita V. Khangar, Dr. Latesh G. Malik, "Compression Method for Handwritten Document Images in Devnagari Script", (IJCSIT)Vol. 3 (3) , 2012

[7] Swital J.Macwan, Archna N.Vyas, "Classification of offline Gujarati Handwritten Characters",2015 IEEE conference.

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

[9] Kamal Gupta, Manish Bansal, Santanu Chaudhury, "A Compression Scheme for Handwritten Patterns Based on Curve Fitting",1520-5363/11 $26.00 © 2011 IEEE

[10] Amandeep Singh Sidhu, Er. Meenakshi Garg, "Text Data Compression Algorithm using Hybrid Approach", IJCSMC, Vol. 3, Issue. 12, December 2014.