Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 213 – 221

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

Visualizing CCITT Group 3 and Group 4 TIFF Documents and Transforming to Run-Length Compressed Format Enabling Direct Processing in Compressed Domain

Mohammed Javed *[a], Krishnanand S.H.[a], P. Nagabhushan [a], B. B. Chaudhuri [b]

[a]*Department of Studies in Computer Science, University of Mysore, Mysore, India*
[b]*CVPR Unit, Indian Statistical Institute, Kolkata, India*

Abstract

Compression of data could be thought of as an avenue to overcome Big data problem to a large extent particularly to combat the storage and transmission issues. In this context, documents, images, audios and videos are preferred to be archived and communicated in the compressed form. However, any subsequent operation over the compressed data requires decompression which implies additional computing resources. Therefore developing novel techniques to operate and analyze directly the contents within the compressed data without involving the stage of decompression is a potential research issue. In this context, recently in the literature of Document Image Analysis (DIA) some works have been reported on direct processing of run-length compressed document data specifically targeted on CCITT Group 3 1-D documents. Since, run-length data is the backbone of other advanced compression schemes of CCITT such as CCITT Group 3 2-D (T.4) and CCITT Group 4 2-D (T.6) which are widely supported by TIFF and PDF formats, the proposal in this paper is to intelligently generate the run-length data from the compressed data of T.4 and T.6, and thus extend the idea of direct processing of documents in Run-Length Compressed Domain (RLCD). The generated run-length data from the proposed algorithm is experimentally validated and 100% correlation is reported with a data set of compressed documents. In the end, text segmentation and word spotting application in RLCD is also demonstrated.

Peer-review under responsibility of the Organizing Committee of CMS 2016

Keywords: Run-length compressed domain processing, Run-length data, Modified Huffman(MH), Modified Read(MR), Modified Modified Read(MMR)

* Corresponding author: Mohammed Javed, Tel.:+919741161929;
E-mail address: javedsolutions@gmail.com

Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.214

1. Introduction

In today's digital era, data compression is the technique generally employed to overcome the volume aspects of the Big data. In fact, on daily basis this results in large scale of compressed data being stored and transferred in the compressed formats. On the contrary, as generally witnessed, any operation or analytics over the compressed data is executed after decompression. If this reversing stage of decompression could be avoided and the analytics could be carried out directly in the compressed version, then it will be an additional breakthrough. Towards this, deeper understanding of the nature of the compression would provide some useful clues. Recently, this novel idea of operating directly over the compressed data has attracted many researchers and as a result latest books and research papers on compressed domain techniques[1,2,3,4] on texts, images and videos have been published. The Document Image Analysis (DIA) community is yet to gain thrust in the area.

In the literature of DIA, there have been a few initial attempts to explore the possibility of operating directly over the compressed formats such as CCITT Group 3[1,5,6,7] CCITT Group 4[3,8, 9] JPEG[4] and JBIG[10]. However, the proposed methods and operations are limited to a particular compressed format. In the recent literature, lot of interesting and deeper works like feature extraction[1,5,11], page segmentation[6,12], text segmentation[1,7], font size detection[11], etc have been reported on the run-length compressed data of CCITT Group 3 1-D compressed documents in Run-Length Compressed Domain(RLCD). Incidently, the other advanced compression schemes of CCITT are also based on Run-Length Encoding (RLE) technique. Based on the variations in the RLE encoding process, CCITT (International Telegraph Telephone Consultative Committee) has introduced a series of compression standards and transfer protocols for black and white images over telephone lines and data networks[13, 14]. They are popularly known as CCITT Group 3 1-D(MHModified Huffman), CCITT Group 3 2-D (MR-Modified Read) and Group 4 2-D (MMR-Modified Modified Read). These compression algorithms are widely supported by TIFF and PDF formats for handling printed and handwritten text documents. CCITT Group 3 contains synchronization codes and hence was developed for network communications, whereas CCITT Group 4 was designed for archival purpose, applicable in large databases because of its high compression ratio. Overall, it can be observed that the run-length compressed data is the backbone of CCITT compression schemes. Therefore, the proposal in the research paper is to extend the idea of directly operating on compressed documents in RLCD to advanced compression schemes like MR and MMR by intelligently generating run-length code. Towards this purpose, a novel algorithm is proposed in this paper.

In this backdrop, the proposed research paper aims at (i) getting deeper understanding of the compressed data of RLE flavored advanced compression schemes like MH, MR and MMR of CCITT, (ii) transforming the compressed data of MR and MMR to Run-length data, and (iii) demonstrating direct operations and analytics on the generated run-length compressed data.

Rest of the paper is organized as follows. Section 2 is dedicated for discussing background information related to this research work such as TIFF data format, MH, MR, MMR encoding schemes from the perspective of compressed domain processing. Section 3 demonstrates visualization of TIFF compressed data, section 4 introduces the novel algorithm of transforming MR and MMR compressed data to run-length data and subsequently discusses the Run-Length Compressed Domain processing. Section 5 reports experimental results and section 6 summarizes the research work.

2. Background

2.1. Structure

TIFF[15] is a graphical format which stands for Tagged Image File Format and a typical TIFF file organization is shown in Fig-1. In the figure IFH stands for Image File Header, Bitmap data actually contains the black and white pixels data either in raw or compressed form, IFD stands for Image File Directory, and EoB indicates End of Byte.

Fig. 1.File organization of TIFF

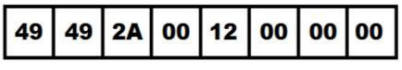

Fig. 2. Image File Header

A TIFF file always begins with an 8-byte IFH that points to an IFD which is shown in Fig-2. In the figure, the first two bytes indicate the byte order, where 4949H in hexadecimal notation represents litte-endian and 4D4DH