OPTIMIZING LOSSLESS COMPRESSION BY NORMALIZED DATA LENGTH IN HUFFMAN ALGORITHM

TONNY HIDAYAT

DOCTOR OF PHILOSOPHY

2022

Faculty of Information and Communication Technology

OPTIMIZING LOSSLESS COMPRESSION BY NORMALIZED DATA LENGTH IN HUFFMAN ALGORITHM

Tonny Hidayat

Doctor of Philosophy

2022

OPTIMIZING LOSSLESS COMPRESSION BY NORMALIZED DATA LENGTH IN HUFFMAN ALGORITHM

TONNY HIDAYAT

A thesis submitted
in fulfillment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2022

DECLARATION

I declare that this thesis entitled "Optimizing Lossless Compression by Normalized Data Lenght in Huffman Algorithm" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

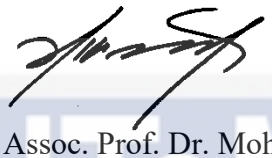
Signature : 

Name : Tonny Hidayat

Date : 14/04/2022

APPROVAL

I hereby declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality for the award of the Doctor of Philosophy.

Signature	:	
Supervisor Name	:	Assoc. Prof. Dr. Mohd Hafiz Zakaria
Date	:	14/04/2022

DEDICATION

I dedicated my thesis work to my family.

Special thanks to my loving parents, Aflahah and Wardah, who have always been the motivation and encouragement of my life.

For my little family, Elfin Aquarista Putri (Wife) was the one who sacrificed the most for me during this study, Levano Arsakha Valin (First Child) forgive Ayah because time and attention for you is reduced, Revalinka Aluna Saira (Second Child) you were born when Ayah is studying to be a great blessing, my love for you all can never be quantified.

Hopefully, this happiness and pride can be felt by all those who always never give up and work hard.

ABSTRACT

Due to the grown need of storage space, the demand for efficient compression scheme becomes increasingly important. One of the lossless data compression goals is to archive raw audio data to ensure the file is restored to the original form when it is to be reused. Generally, raw data is stored as 16-bit (65,536 difference possible values). Huffman Algorithms is currently still very effective at compressing 8-bit data, which can be grouped into Static, Dynamic, and Adaptive extensions, however its performance cannot be determined if it is performed on data that has several variables and probabilities. Based on the literature review, the measurement of the compression performance for files archives is to use the Compression Ratio (CR) and Compression Time (CT) indicators. These two indicators are used to calculate and analyse the file size reduction and the ability of the file to be reconstructed back to its original form without compromising its quality. This research produces a new scheme called Quaternary Arity (4-ary) Modification Quadtree (MQ) or 4-ary/MQ based on entropy coding which has its roots in other variants of Huffman schemes such as Binary / Static, Quadtree, Octatree, and Hexatree. The 4-ary/MQ method employs the characteristics of the Quadtree structure and extends the Dynamic Huffman coding mechanism (FGK rule) in node arrangement while adopting the Adaptive Huffman method that uses additional variable data. The novelty of this scheme is the work of adding additional variables to maintain the branch root to ensure it is always consistent with four branches. A descriptive analysis of the 4-ary/MQ was performed on several audio datasets (Music, Mono Music, Stereo Music, Ripping CD, Speech, Noise, Sound Effects, and Instruments) to compare with the Huffman Schematic Variant. A comparative analysis with several lossless compression applications has significantly shown that CR is more optimal than PKZIP, WinZip, 7-Zip, and Monkeys Audio. It was found that the 4-ary/MQ compression benefits the compressed data that is stored in local storage media as well as for hosting and optimizing bandwidth. The new algorithm also has a good performance in producing optimal CR with fast CT in most of the 16-bit WAV audio datasets. The proposed new algorithm has more optimal CR than the various variants of the Huffman-based lossless application. It is also expected that this new algorithm scheme may potentially work well on data above 16-bit for future research.

MENGOPTIMUMKAN PEMAMPATAN TANPA HILANG MELALUI PANJANG DATA TERNORMAL DALAM ALGORITMA HUFFMAN

ABSTRAK

Keperluan ruang simpanan yang semakin meningkat telah menjadikan keperluan untuk skema pemampatan yang efisien semakin penting. Salah satu tujuan pemampatan data berjenis tanpa hilang adalah untuk penyimpanan data audio berjenis raw bagi memastikan fail dapat dikembalikan ke bentuk asal apabila ingin digunakan semula. Secara amnya, data raw disimpan sebagai 16-bit (65,536 perbezaan kemungkinan nilai). Algoritma Huffman pada masa ini masih sangat berkesan dalam memampatkan data 8-bit, yang dapat dikategorikan sebagai berjenis statik, dinamis, dan adaptif, namun prestasinya tidak dapat ditentukan pada data yang memiliki sejumlah pemboleh ubah dan kebarangkalian. Berdasarkan tinjauan literasi, pengukuran prestasi pemampatan untuk fail adalah dengan menggunakan indikator Nisbah Pemampatan (NP) dan masa pemampatan (MP). Kedua-dua indikator ini digunakan untuk mengira dan menganalisis pengurangan ukuran fail dan kemampuan fail untuk disusun kembali ke bentuk asalnya tanpa menjejaskan kualitinya. Penyelidikan ini menghasilkan skema baharu yang disebut Quarternary Arity (4-ary) Modification Quadtree (MQ) atau 4-ary/MQ berdasarkan pengekodan entropi yang berasal dari pelbagai varian lain dari skema Huffman seperti Binary / Static, Quadtree, Octatree, dan Hexatree. Kaedah 4-ary/MQ menggunakan ciri struktur Quadtree dan memperluaskan mekanisme pengekodan Huffman dinamik (peraturan FGK) dalam susunan nod, dengan mengadaptasi kaedah Huffman adaptif yang menggunakan data pemboleh ubah tambahan. Keunikan skema ini adalah pada penambahan pemboleh ubah tambahan untuk mengekalkan akar cabang bagi memastikannya sentiasa konsisten dengan empat cabang. Analisis deskriptif ke atas 4-ary/MQ dilakukan pada beberapa set data audio (Muzik, Muzik Mono, Muzik Stereo, Ripping CD, Ucapan, Kebisingan, Kesan Bunyi, dan Instrumen) untuk dibandingkan dalam varian Skema Huffman. Analisis perbandingan dengan beberapa aplikasi pemampatan tanpa hilang lain telah menunjukkan bahawa NP lebih optimum berbanding PKZIP, WinZip, 7-Zip, dan Monkeys Audio. Kajian juga mendapati bahawa pemampatan 4-ary/MQ memberi keuntungan kepada data termampat yang disimpan di media penyimpanan tempatan dan juga untuk hosting dan mengoptimumkan lebar jalur. Algoritma baru ini juga mempunyai prestasi yang baik dalam menghasilkan NP yang optimum dengan MP pantas di kebanyakan set data audio WAV 16-bit. Algoritma baru yang dicadangkan mempunyai NP yang lebih optimum daripada pelbagai varian aplikasi tanpa hilang berasaskan Huffman. Diharapkan skema algoritma baru ini berpotensi untuk berfungsi dengan baik pada data di atas 16-bit untuk penyelidikan lain di masa akan datang.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious, the Most Merciful

First and foremost, all praise and glory to Almighty Allah (Subhanahu Wa Taalaa), my Creator, my Sustained, for everything I received since the beginning of my life. Shalawat and Salam to our Prophet Muhammad (Peace Be upon Him) who has brought us out from the darkness into the brightness world that full of knowledge.

I would like to extend my appreciation to the Universitas Amikom Yogyakarta and Universiti Teknikal Malaysia Melaka (UTeM) for providing the research platform. My utmost appreciation goes to my main supervisor Assoc. Prof. Dr. Mohd Hafiz Zakaria (UTeM) for all his support, advice and inspiration. His constant patience for guiding and providing priceless insights will forever be remembered. Also, to my co-supervisor, Assoc. Prof. Dr. Ahmad Naim Bin Che Pee (UTeM) who constantly supported my journey.

Last but not least, from the bottom of my heart a gratitude to my beloved wife, Elfin Aquarista Putri, for her encouragements and who have been the pillar of strength in all my endeavours. My eternal love also to all my children, Levano Arsakha Valin, and Revalinka Aluna Saira, for their patience and understanding. I would also like to thank my beloved parents for their endless support, love, and prayers. Finally, thank you to all the individual(s) who had provided me the assistance, support, and inspiration to embark on my study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES