Emre Sonrez
Timothy Blumberg
CompSci 201: Data Structures and Algorithms
Dr. Tabitha Peck
Huffman Assignment: Analysis.pdf
Due: 17/04/2014

Huffman Compression: An Analysis of the Compression Effectiveness

Introduction

The Huffman compression algorithm was first published in 1950 by David Huffman, who was then a graduate student at MIT. It relies upon analyzing which bit sequences of a file occur the most frequently, so that they can then be compressed the most in order to save the most disk space possible.

An Analysis of Compression Effectiveness

Using this compression algorithm, we were able to compress practically any file that could be stored on a computer. An interesting question can then be posed by this wide-range of possible file types; do certain types of files compress better than others? In order to address this problem most analytically, we chose to assess the differences in compression between .tiff and .txt files.

We found that .txt files actually compress better than .tiff files on average. Figure 1 shows the spectrum of compression degree as the number of characters (out of the 256 possible for the 8 bits per letter that we used) increases. Whenever the files use a similar number of the binary sequences out of the possible 256. As competent reader might notice, the low letter count TIFF files compress about as much as the average TXT files, but most TIFF files are high letter count files, so the average compression rate is much lower than the average TXT file compression rate. Table 1 outlines the basic statistics that were calculated during the analysis of the two different file formats.

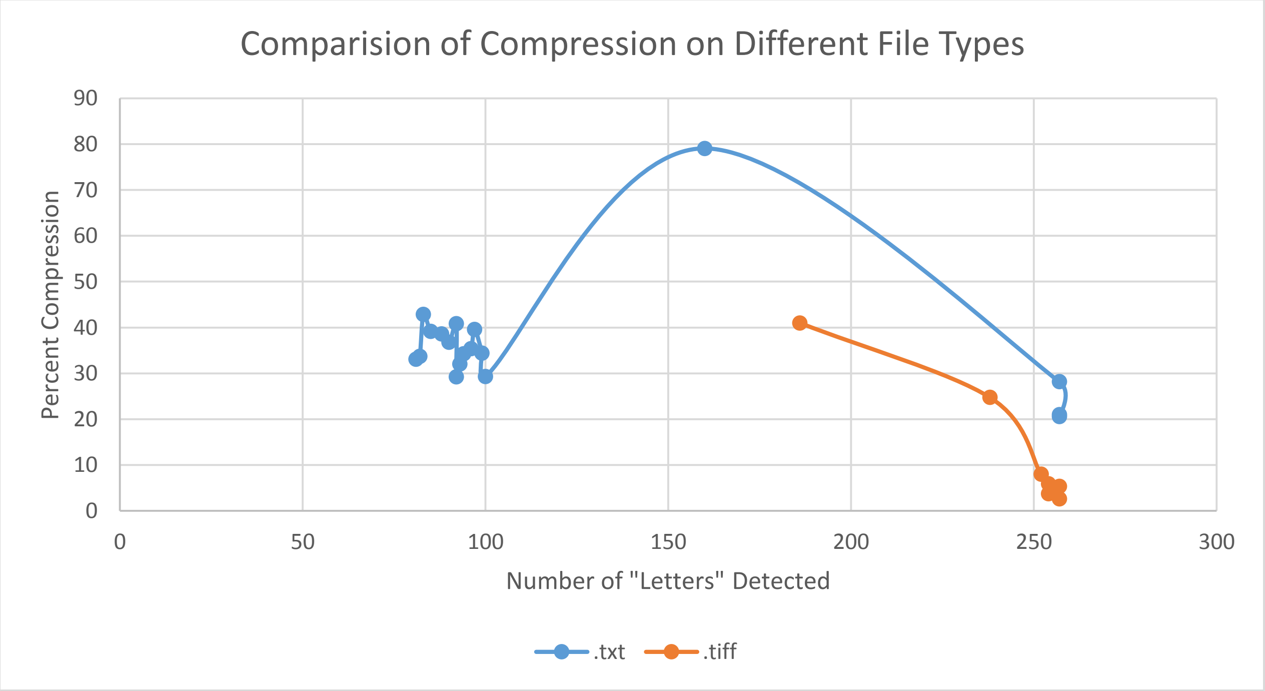

Figure 1: Comparison of Compression as a Function of Letter Count

Emre Sonrez
Timothy Blumberg
CompSci 201: Data Structures and Algorithms
Dr. Tabitha Peck
Huffman Assignment: Analysis.pdf
Due: 17/04/2014

Table 1: Comparison of Compression Statistics Between File Formats

	TXT Files	TIFF Files
Total % Compression:	43.241	18.137
Compress speed (bits / sec)	158037	384205

Therefore, the TXT files compress far better than the TIFF files because they contain fewer of the possible binary sequences, so the total file can be compressed more. An interesting finding that is shown in Table 1 is that the TIFF files actually compress faster by well over a factor of two than the TXT files that we tested. This faster compression is most likely due to the fewer number of compression actions that need to be taken to correctly compress the TIFF files as compared to the TXT files which are being more completely compressed.

Analyzing Multiple Compression

The second area of exploration that we undertook was how thoroughly could files be compressed before it was no longer useful? To explore this topic, we compressed files more than once and then compared how the compression changed as the number of compressions increased. We ran a modified version of the compression analysis utility `HuffMark` that would compress the files several times and measure how things changed as the number of compressions increased.

	Initially uncompressed	One Compression	Two Compressions
Total % Compression	43.24	1.96	1.96
Time Taken			
Compress spd. (bits/s)			