Data Compression

Jorge Solorzano

The most important advancement in the sharing of Data would have to be the internet. The amount of information that was available practically overnight was overwhelming for the slow internet connections that relied on dial-up. As with all technology that came before it, the internet had to somehow make it possible for users to retrieve information over this slow connection faster.

Downloading information, files, movies, songs etc. is common place in present day anywhere in the world. Many remember the hours upon hours it took to download a file even if it was something as small as a text file or a picture. Many minds decided to get together and find a solution to slow website download times. The solution was: Data Compression.

Data Compression is something many people are familiar with since it is common language on the World Wide Web. Zip files are the most common form of compressed files and an easy way to compress multiple folders all into one. Decreasing the size of a file from a gigabyte to a couple megabytes made downloading files a lot simpler and it was astounding that the compressed file could be restored to the original or change to a point where the changes made by the compressor program are not noticeable.

With the internet speeding up with DSL and cable connections, compressed files are now even faster to download and open. Even with the internet speeding up, it would feel like the download speed was back to dial-up if the files were not compressed or changed to be smaller size and that is something no one wants. No matter how amazing Data Compression is, it means nothing if no one knows how it works. How it is possible for a 5000 page text book online to be compressed into something less than its size without loss of data? How can pictures size be cut down drastically without the picture quality changing? How many types of Data Compressions are out there on the internet? All these are important questions that deserve an answer.

Categories and Subject Descriptors: Lossy [**Lossless**]:

General Terms: Text,Picture,Video,Sound

Additional Key Words and Phrases: Data Compression

1. INTRODUCTION

The textbook definition of Data Compression is the processes of encoding, converting a piece of information into another form of representation, information using fewer bits than an un-encoded representation would use, through use of specific encoding schemes. Present day society enjoys internet that has speeds that makes downloading files or loading possible in seconds and because of that, data compression does not get the respect it deserves by the new generation. In order to truly appreciate the greatness of data compression, a person must look back to the time when the internet was first made public also known as the times of dial up aka 56 kilobytes per second. Practically overnight there was a surge of information available to any user but with such a slow speed downloads of this information took hours which is not convenient to the busy lifestyle of modern society. How does society solve the problem of downloading files at slow speeds when people need them right away?

Data compression in its early days was known as Morse code and was introduced in 1838 as a way to quicken messages by breaking down words in to short segments of code. Modern advancements in data compression didnt start until the 1940s when Claude Shannon and Robert Fano devised a systematic way to assign code words based on probabilities of blocks. This method was later optimized by David Huffman in 1951 who used this method on hardware with specific choices of code words being made as compromises between compression and error collection. With online storage of text files becoming common in the 1970s, data compression programs were starting to come into fruition based on the early models by Hoffman.

Abraham Lempel and Jacob Ziv suggested point-based encoding in 1977 which was later followed by Terry Welch in the mid-1980s with his LZW algorithm which quickly became the method of choice for most general purpose compression systems. As digital images became more common in the late 1980s, data compression methods were extended to picture formats to lower the size of pictures in storage. The first common day data compression algorithm used today was introduced in the early 1990s as the lossy compression methods which brought in file types like JPG, GIF, BMP.

Data compression plays an important role in the society we live. The internet is the most common place to find compressed files but does not cover the whole spectrum that has some form of data compression. The modem that a person uses in their household uses data compression, HDTV uses video compression called MPEG-2, the application on a person phone use data compression as well. In reality, video processing would be virtually impossible without compression. A movie or workout program would not fit in a DVD if data compression didnt exist. These are things used every day that the common person doesnt even consider how it is possible and what to thank for the convenience of the appliance.

The huge increases in the speed of internet with first the introduction of DSL and later Cable have made one forget the beauty and importance that is data compression. With images and videos being loaded in just seconds one tends to forget how less than 15 years ago it took minutes if not hours to load video. To truly understand the amazing thing that is data compression it is important to go back to the 90s when dial-up internet was the wave of the future.

2. LOSSLESS COMPRESSION

With the introduction of the internet there was a world filled with information that was just on the fingertips of the user. As websites were being built left and right more books, articles, and newspapers were being made available online. The idea that a person could store the article that they found online onto their own computer led to the introduction of downloading. At this point in time people were downloading articles or portions of books if not whole books wherever they could find them, this, however, posed a problem. Dial-Up internet only worked with a theoretical speed of 56kpbs so it would not be possible for full sized files to be downloaded with such a slow speed since it would take far too much time and time was something that was becoming short in supply for corporate America. This lead to the question: How can we make this data accessible on such a low speed?

The answer was, of course, data compression. If the server could provide a compressed (smaller) version of the article then it would increase the download speed. When the internet was young, the only information available was text and it was being compressed in a style called Lossless compression. Lossless compression means that the original file will be compressed and when it is opened it will go back to the original size with no data loss. How does this work?

Lossless compression is truly useful for text files and in present days only really works for text files. The way it works is that the compressor program will take the original file and the program will find patterns and set a code value for each pattern found and this process will reduce the size of the file. So basically the file will have some words and a bunch of code values that represent bigger strings and

the file is now ready for download. A user then is able to download this compressed file along with the integrated key that will then allow a word processor program to use the key to extend the file back to the original size and most importantly, the original content.

Take into consideration the following example:

In John F. Kennedy's 1961 inaugural address, he delivered this famous line:

> "Ask not what your country can do for you – ask what you can do for your country."

The quote has 17 words, made up of 61 letters, 16 spaces, one dash and one period. If each letter, space or punctuation mark takes up one unit of memory, then this sentence is a total file size of 79 units. To get the file size down, the compression program would have to look for redundancies. The compression program will notice that ask, what, your, country, can, do, for, and you all appear twice and if the program ignores capitals and lower-case letters then roughly half of the phrase is redundant. To construct the second half of the phrase the program can point to the words in the first half and fill in the spaces and punctuation.

The program will then make a menu for the words being repeated:

(1) ask

(2) what

(3) your

(4) country

(5) can

(6) do

(7) for

(8) you

Each number represents a word that was repeated in the sentence and then the program will then replace each corresponding word with its number counter-part. So now the sentence reads:

"1 not 2 3 4 5 6 7 8 1 2 8 5 6 7 3 4"

This would be the compressed version of this sentence depending on the key being used but how much space was actually saved by compressing this sentence? Assuming every character takes one unit of memory then the new sentence takes up 37 units but the key has to be saved with the compressed file as well and the key takes up 37 units as well so the total size for this compressed sentence is 74 units versus the 79 units of the non-compressed sentence. That is not much saved space but this is only one sentence. Where ever the words in the key appear on the whole of the speech then the key will replace it with the corresponding number value providing which would lower the size of the file. This, however, may not be the most efficient key or the most efficient way to compress this sentence and there could be more effective ways since it all depends on the compression programs algorithm.

This process is very limited in what it can compress. It can take any data input but if it cannot find a pattern then it will spit out the original without compressing it at all. In other words this only works for things like text that have letter patterns or number patterns. For this reason other compression types were created for things like pictures, movies and music.

3. LOSSY

Lossy Compression is defined as a compression method that, unlike lossless, does throw out excess data to decrease the size of the file that is being compressed. Once the file is compress the excess data that the compression program took out as pointless is lost for good and cannot be retrieved. This is the type of compression mostly used for pictures.

Sharing pictures is now something that anyone can do with websites like facebook, myspace or tinypic. Any user that remembers the times of dial-up still remembers how slowly pictures loaded. These were already compressed pictures so it is easy to imagine how long a non-compressed picture would have taken to load. This is something that has been lost in the age of high-speed internet connections and download time on pictures has decreased to practically nothing. Despite this, most pictures today are still compressed in the same way.

Lossy is a type of compression which is used a lot for reducing the file size of bitmap pictures, which are usually fairly bulky. To understand how it works, consider how a computer might compress a scanned photograph.

A lossless compression program can't do much with this type of file. While large parts of the picture may look the same with the whole sky being blue, for example, most of the individual pixels are a little bit different. To make this picture smaller without compromising the resolution, the compressor program would have to change the color value for certain pixels. If the picture had a lot of blue sky, the compression program would pick one color of blue that could be used for every pixel. Then, the program rewrites the file so that the value for every sky pixel refers

back to this information. If the compression scheme works well, you won't notice the change, but the file size will be significantly reduced.

An important fact to take into consideration with lossy compression is that the original file cannot be brought back to its original size after it has been compressed. The user is stuck with the compression program's reinterpretation of the original. For this reason, the user can't use this sort of compression for anything that needs to be reproduced exactly, including software applications, databases and presidential inauguration speeches.

4. VIDEO COMPRESSION

A video is the same as a whole series of pictures put together in a sequence that gives the impression of movement through a movie. In other words, if a computer user was to break up videos into frames then the user would have a series of pictures. Imagine a pack of post-its. If someone was to draw frame by frame a movement sequence then flip through the post-its quickly they would see a moving picture. This is the basic idea of video with, of course, a much faster frame rate. The digital images can be compressed and sometimes left out, which allows for the size of the video to be smaller. This enables for less bandwidth needed to transmit videos, which is the technology used by youtube. Video compression works because a huge portion of the data present in digital video is not essential for the human brain's ability to process it. DVDs, for example, use a compression system called MPEG-2 that can shrink the data of a two-hour movie by 15 to 30 times without lowering the quality of the video for standard televisions.

Video compression can compress images spatially and temporally. Spatial com-

pression works because the human eye is less able to distinguish between small differences in color than changes in brightness. Spatial compression then removes small differences in color in areas of the picture that are next to each other. Temporal compression works by comparing the digital images from frame to frame. Areas that are very similar between frames are then ignored. Ignored areas are left out of the encoding and as a result aren't included in the data. Both of these methods are able to minimize the amount of data needed to encode a video clip.

5. SOUND COMPRESSION

The most common compression system for music would have to be mp3. With the explosion of mp3 players in the early years of the millennium to the download engines that provided peer to peer access to huge amounts of songs at the expense of copyright laws, mp3s have been filling up the internet for many years now. A CD stores a song as digital information. The data on a CD uses an uncompressed, high-resolution format. Here's what happens when a CD is created music is sampled 44,100 times per second. The samples are 2 bytes (16 bits) long. Separate samples are taken for the left and right speakers in a stereo system. So a CD stores a huge number of bits for each second of music:

44,100 samples/second * 16 bits/sample * 2 channels = 1,411,200 bits per second

Breaking that down: 1.4 million bits per second equals 176,000 bytes per second. If an average song is three minutes long, then the average song on a CD consumes about 32 million bytes (or 32 megabytes) of space. Even with a high-speed cable or DSL modem, it can take several minutes to download just one song. Over a 56K dial-up modem, it would take close to two hours.

The MP3 format is a compression system for music. The goal of using MP3 is to compress a CD-quality song by a factor of 10 to 14 without noticeably affecting the CD-quality sound. With MP3, a 32-megabyte song on a CD compresses down to about 3 MB. This lets you download a song much more quickly, and store hundreds of songs on your computer's hard disk.

To make a good compression algorithm for sound, a technique called perceptual noise shaping is used. It's perceptual partly because the MP3 format uses characteristics of the human ear to design the compression algorithm. Some human hearing characteristics taken into account such as certain sounds that the human ear cannot hear, certain sounds that the human ear hears much better than others, and if there are two sounds playing simultaneously, we hear the louder one but cannot hear the softer one.

Using facts like these, certain parts of a song can be eliminated without significantly hurting the quality of the song for the listener. Compressing the rest of the song with well-known compression techniques shrinks the song considerably – by a factor of 10 at least. When you're done creating an MP3 file, what you have is a near CD quality song. The MP3 version of the song does not sound exactly the same as the original CD song because some of it has been removed.

6. CONCLUSION

Compression is a very useful tool and deserves a lot of credit for the internet and technologies enjoyed by the public today. Something that a user must always remember when using a compression program is if the program is the most efficient way to compress the file the user wants to compress?

7. BIBLIOGRAPHY

Gonzalez, Rafeal and Woods, Richard. 'Digital Image Processing' Addison-Wesley 1992.

Harris, Tom. 'How Stuff Works: Data compression' http://computer.howstuffworks.com/mp34.htm

Motta, Giovanni and Salomon, David. 'The Handbook of Data Compression' Springer, New York .

Nelson, Mark and Gailly, Jean-Loup. 'The Data Compression Book' MnT Books. New York. 1996