RESEARCH PAPER | Engineering | Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

A Survey Paper on Different Data Compression Techniques

KEYWORDS: Shannon-Fano, Huffman coding, Lemple Ziv

Saumya Mishra | shraddha singh

ABSTRACT *This paper describes the different types of compression techniques such as lossless and lossy data compression. This survey paper has been written with the help of different types of algorithm like run length, Huffman coding, Shannon-Fano, Lempel ziv, vector quantization etc. The reference of these algorithms has been taken from various books and research papers on Data compression. Today data compression is very useful in our life. The main purpose or aim of data compression is to compress any type of data that is transfer over the communication channel, because of the limited channel bandwidth and data storage capacity. The use of lossless and lossy techniques for data compression means that the numbers of bits are reduced in the original information. By the use of lossless data compression there is no loss in the original information but while using lossy data compression technique some numbers of bits are loss.*

Introduction

Data compression is use in the encoding system. The bit rate reduction is use in the encoding system. By the use of bit rate reduction algorithm, the minimum bits are used to compare the original information.

The reduction algorithm is use to transfer minimum bits over the network, it save the storage capacity and the bandwidth, and also transmit the data easily and efficiently. There are two types of data compression: lossy data compression and second one is lossless data compression. Different-different types of compression techniques are involved in the lossless and lossy data compression, such as Shannon-Fano, run-length Encoding, Huffman coding, Arithmetic coding, Lemple Ziv, Vector quantization etc. reduces the number of bits. Data compression is the part of information theory.

The history of data compression involve:-

Data compression started form the year 1838:- In this year telegraphy was used. Telegraphy used the morse code for data compression which is based on small code words for letters for example "e" and "t" that are used as common letters in English.

And in 1940:- the development of information theory evolved. The data compression is the basic part of information theory. Data compression when used in information theory, it reduces the redundant information. Redundancy in the information lay additional bits for encoding, then remove the extra bit in the information and the size of information.

Claude Shannon and Robert Fano invent a algorithm based on probabilities in year 1949.

David Huffman develops the optimal technique for data compression in 1951 after the Shannon Fano algorithm. It depends on the probability of each symbol's appearance in information.

In 1970, the adaptive Huffman coding for data compression took its advanced form. It is almost same as the Huffman coding but not exactly the same.

Later in 1980 the digital image for data compression grew.

And in 1990 the lossy compression technique came which included digital image compression.

Categorization of data compression Technique:-

There are mainly two types of data compression technique.

1. Lossy data compression
2. Lossless data compression

This paper discusses the run length encoding, Huffman coding, adaptive Huffman coding, LZW coding, arithmetic coding as used in lossless data compression. And Vector quantization, in lossy data compression technique.

Lossy data compression: - when Lossy data compression is used for compressing the text or image or any other type of information, some number of bits are loss in the original information. In other words compressing data and decompressing data are not same. For example, lossy data compression is digital cameras; it decreases the quality of picture but increase the storage capacity.

Lossless data compression: - lossless data compression, transmit the message over the network in encoded from and when this message is received by the receiver there is no loss in the message.

3. Different compression techniques

Shannon-Fano coding:-It is the first method for compression, developed by Claude Shannon and RM Fano at MIT & Bell Lab. It is based on the probabilities of symbol in the message [1] .And construct the table on the basses of probabilities that contain some number of probabilities.

- Different codes have different numbers of bits.[1]
- More bits are provided to low probability symbol, and fewer bits are provided to high probability symbol.
- The codes that have different bit lengths in message are uniquely decoded.

Whenever the probability of symbol is larger than the

Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

codes they vary in length.

In this case of encoding the data compression is important for the use of binary tree to solve the problem of decoding for variable length codes.

Huffman coding:- Huffman coding holds the probabilities of Shannon-Fano coding. It creates variable-length codes that are an integral number of bits [1]. But it is different from Shannon-Fano coding, because in the Shannon-Fano the top down approach has been used to design the binary tree, but Huffman coding uses the down to up approach for designing the binary tree.

Adaptive Huffman coding
We use Huffman coding for data compression but the snag of Huffman coding is, to send the probability table with the compress information, because without the probability table decoding is not possible. To remove this disadvantage in Huffman coding, the adaptive Huffman coding developed. This table requires the addition of 0 an extra bytes to the output table, and consequently it usually doesn't make much difference in the compression ratio [1]

Arithmetic coding technique
After the adaptive Huffman coding, the Arithmetic coding developed .It removes the disadvantage of Huffman coding.

Arithmetic coding is used to change the method of replacing the each bit with a codeword [2].It replaces the input data (in the form of string) to a single floating point number. Huffman coding provides the codeword, particularly to each symbol but in the arithmetic coding codeword is provided to the whole string.

Lempel Ziv algorithms:- Jacob Z iv and Abraham Lempel[1] developed the different compression algorithm based on dictionary concept like LZ77 and LZ78 . This algorithm, popularly referred to as LZ78, was published in "Compression of Individual Sequences via Variable-Rate Coding" in *IEEE Transactions on Information Theory* (September 1978)[1]. LZ78 is based on the text window. Term of dictionary in LZ77 was based on fixed window, before seen text. Under LZ78, the dictionary is a potentially unlimited list of previously seen phrases [1].

Run length coding
Run-length coding works when symbols do not occur independently but are influenced by their predecessors. Given that a symbol has occurred, that symbol is more likely than others to occur next. If this is not the case, coding runs (rather than symbols) will not compress the information. The same effect can be achieved in a more general way by other coding techniques, but run-length coding uses very little overhead when the runs are long.

Speech Compression
Speech compression is the important part of data compression. This compression technique, compresses the human speech into encoded format. The encoded human speech is send over a network and receiver receive the encoded speech that will decode the message in approximate, the original from. By using this compression method, we can save the storage space, bandwidth, transmission power and energy. Speech is transfer from one place to another place with the use of limited storage space and bandwidth. The multimedia communication demands to use transmission bandwidth and storage space efficiently. To overcome in the signal we need to compress the speech signal [3].

Dictionary-based compression
This technique is also use for data compression. In this technique, the algorithm does not encode single symbols as variable-length bit strings; it encodes variable-length strings of symbols as single tokens [1,]. It uses the LZW algorithm. There are two types of dictionary based algorithm, static and dynamic .In static, the size of dictionary is fixed for encoding and decoding process And in dynamic there are size of dictionary for decoding and encoding process is not fixed.

Vector Quantization
Vector quantization is a type of lossy data compression because after decoding the message, some numbers of bites are loss. It is based on "principle of block coding [4]".This method is designed for multi dimensional data. Linde, Buzo and gray proposed the LBG algorithm for vector quantization. It is based on input training sequences.

4. Overview of the Research
Gaurav Gupta et al [5] In this paper, describe the image compression using lossless data compression. He also introduces the lossless and lossy data compression techniques. This paper discusses the "Principle of compression which "involves spatial redundancy, Spectral Redundancy, Temporal Redundancy, And also describe the performance of image on the bases of quality, amount and speed of image over the network. It also discusses the different types of lossless and lossy data compression such as Run length encoding, Huffman coding etc.

H. B. Kekre et al [6], this paper studies , the combination of vector quantization and hybrid wavelet transform for data compression and throws light on the implementation part of this compression technique .The reconstructed image of hybrid wavelet is transferred to vector quantization for compression, means that the output of hybrid wavelet transfer as a input to the vector quantization. According this paper the size of codebook is 16 and 32 used by vector quantization.

LBG, KPL and KEVR algorithm used for vector quantization. According to this paper, combination of vector quantization with hybrid wavelet transformation provides the better output comparison to hybrid wavelet transforms on the compression ratio 32.

These are the lossy image compression, and three vector quantization algorithm (LBG, KPE, KEVR) used for reconstructing the image again. And compare the output of these 3 algorithms. After the implementation of these algorithms the output is KPM which is better than the LBG algorithm.

Asmita A. Bardekar et al [7], discusses in this paper, the lossy and loosless data compression method but focuses only on the vector quantization technique for image compression. In vector quantization the training set of input as the codeword and set of codeword's are the codebook. In this paper, image as the input vector called codeword and block of image is the codebook. Three steps that follow vector quantization, first one is codebook design second one is image encoding and last step is decoding the image.

Manjeet Kaur et. al [8] describes lossy and lossless data

Volume : 6 | Issue : 5 | May 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

compression techniques in this paper. And then apply the modified Huffman coding. It works on text data for compression by using 3 steps. In the first step data is compressed with the help of dynamic bit reduction method and in second step unique words are found to compress the data and in final step, Huffman coding are used to compress the data and then produce the final output .

R. R. Khandelwal et. al[9] opines that vector quantizers (VQ) are a group of vectors mapped to a codeword. Codebook is a collection of codeword's. Vector Quantization is not having a regular shape while a lattice vector quantization is a regular arrangement of points. In lattice quantizer, the codeword's are lattice points. In lattice quantizer the codebook are obtained by selecting finite number of lattice points or codewords out of infinite lattice points. The performance of Lattice Vector Quantization depends on the length of the codebook. In a codebook, if the number of lattice points is less, time required to get quantized lattice point will be less but at the same time quality of reconstructed image decreases. But if the size of the codebook is larger than the time required to search nearest vector will be more. So to avoid these situations there is one solution, which we have used in this work, expansion of codebook as and when needed. This technique eliminates overload errors in the codebook. When compared with single codebook system it requires less time for encoding and decoding.

Vimal Kishore et al [10] relates Speech compression to multimedia communication. It is a very important part of compression because today, the multimedia communication is very popular. This compression method converts the human speech to compress encoded format. Use of this compression method, saves the storage space, bandwidth, energy over the communication channel. Satellite communications, internet communications, transmission of biomedical signals and other applications are involve in the speech compression. The paper also discusses the DWT algorithm for data compression.

5. Conclusion

This survey paper discusses the different type of lossy and lossless data compression. And different type of research paper on the based on data compression. This paper only discusses the general idea of data compression. Today, many compression techniques are developed and some techniques are in process .But this paper only discusses the general idea about the Shannon-Fano, Huffman coding, arithmetic coding, and vector quantization. This paper has been written to understand vector quantization in the better manner and relate it to the future work .

References

[1] Mark nelson ,Sean –Loup Gailly, "The data compression book", second edition, " Publisher: IDG books Worldwide, Inc."

[2] Rajinder Kaur, Mrs. Monica Goyal, "A Survey on the different text data compression techniques" ,Journal of Engineering Technology and Computer Research 2013(IJARCET)

[3] Vimal Kishore Yadav1, Alok Jain2, Lenka Bhargav3," Analysis and Comparison of Audio Compression Using Discrete Wavelet Transform", International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 4, Issue 1, January 2015

[4] http://www.datacompression.com/vq.html [17.10.2015]

[5] Gaurav Gupta, Parul Thakur," Image Compression Using Lossless Compression Techniques", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 2 Issue: 12

[6] H. B. Kekre ,T. Sarode, P. Natu," Image Compression using Fusion of Hybrid Wavelet Transform and Vector Quantization", African Journal of Computing & ICT ,Vol 7., No. 5 – December, 2014

[7] Asmita A. Bardekar et al "A Review on LBG Algorithm for Image Compression" / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2584-2589

[8] Manjeet Kaur , Er. Upasna Garg "Lossless Text Data Compression Algorithm Using Modified Huffman Algorithm ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015

[9] R. R. Khandelwal, P. K. Purohit and S. K. Shriwastava ,"LATTICE VECTOR QUANTIZATION FOR IMAGE CODING USING EXPANSION OF CODEBOOK" The International Journal of Multimedia & Its App.lications (IJMA), Vol.4, No.4, August 2012

[10] Vimal Kishore, Yadav Alok, Jain, Lenka Bhargav, "Analysis and Comparison of Audio Compression Using Discrete Wavelet Transform", International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 4, Issue 1January 2015