Applying Distortion Measure and Classification of Data Mining to Image Compression

S. Radhika
M.Sc., M.Phil.,
Assistant Professor,
Thiruthangal Nadar College,
Chennai - 600 051.

Edith Maria Delilah
M.C.A.,M.Phil.,
Assistant Professor,
Dr.MGR Janaki College of Arts & Science,
Chennai - 600 028.

Abstract:- **Vector quantization is one of the techniques to compress images. Image is divided into vectors. These vectors are called as image vectors. Code vectors are created. For each code vector, image vectors are classified based on distortion measure. In this way, Image vectors are grouped. Image contains groups of vectors. Mean vector is calculated for each group. Image vectors are replaced by its corresponding mean vector. Also code vectors are updated by its corresponding mean vectors. This process is repeated until decompressed image is as close as to original image. Each code vectors are assigned some unique address. In the compression process, image vectors are replaced by its code vector address. In the decompression process, address is replaced by its code vector.**

I. INTRODUCTION

Data Compression has been studied a topic in the field of information theory, encompassing the study of representation, storage, transmission and transformation of data. Data Compression technique can be divided into two categories such as lossless compression and loss compression.

Lossless compression technique involves no loss of information. If data have been compressed using lossless compression, it can be recovered exactly from compressed file.

Examples for lossless scheme include text compression, run-length coding and Huffman coding,

Spread sheet, word processor files, database files and program execution files usually contains repeated sequence of characters. When decompressed the compressed files, repeated characters are reinstated. Loss compression technique involves some loss of information. Data that have compressed using loss compression technique cannot be recovered original data exactly from decompressed file. Example for loss coding scheme includes vector quantization, transform coding, Loss compression techniques are used for gray scale or color images, compressing audio and video objects in which data accuracy are not essential.

In the lossless compression technique, compression ratio is much lower and where as in the loss compression technique, compression ratio is much higher. Data compression is a tradeoff between compression performance and distortion performance: between time and efficiency.

In this work, a method of vector formation, a technique of distortion measure, a method of calculating mean vector and optimization of code book are discussed.

II. VECTOR FORMATION

Each dot of binary image occupies 1-bit of information. If the pixel is black or white, it is represented by the value 0 or 1 in the computer storage respectively. Consecutive eight dots on the same row take the value from 0 to 255 in the computer storage of raster scan image. Sample is defined as a number ranging from 0 to 255 which represents eight consecutive dots. Now the image is divided into N*N samples. This N*N sample is an image vector. A 4*4 input vector consisting of 32 pixels in width and 4 pixels in height.

For color image of 256 colors, each dot is represented by eight bits. Here a vector is the square image consisting of 4 pixels in width and 4 pixels in height. An image vector in black and white image contains 128 pixels where as an image vector in color image contains 16 pixels. But vector size in both images is 16 bytes.

III. DISTORTION MEASURE (DATA MINING)

To fully define the encode mapping, we need to specify the distortion measure that will be employed in the code book search. Distance between two u and v is defined as number of unlike bits between u and v.

IV. MEAN VECTOR

We need to discuss about the method of finding mean vector of the finite set of vectors that is employed in the updating of code book. First we are going to find the mean of finite set of all samples and then that mean can be extended to finite set of vectors. Let V be the finite set of all samples. Method of finding mean of set V is as follows.

Count the numbers 1’s and 0‘s of corresponding bit position of all the integers in the set V. If 1‘s are more, then the bit position of mean of V is set 1. If 0‘s are more, set that position of mean of V is 0. Otherwise, that is set to1 or 0. If the number of zeros equal to the number of ones, then that position of mean of V is set to 1 and 0 alternatively.

Now the mean can be extended to finite set of vectors where each vector is an n- tuple sample. Let V be the finite set of all vectors. Each sample of mean vector of V is the mean of corresponding samples of all the vectors in a set.

V. OPTIMIZATION OF CODE BOOK

Initial code book is fixed and it is independent of of a given image. Reconstruction of image from the compressed file by using initial code book may or may not faithfully reproduce the original image. This problem is to optimize this code book such that the reconstruction of original image from the compressed file is as close as to original image. Code book is to be updated continuously during the process until quality of image reconstructed from the code book is constant for successive updating. Then at this stage, the code book is optimum.

VI. RESULT

Here 1056 * 104 bytes of rectangular image has been taken for compression. This image is divided into 4*4 vectors. 128 code vectors are predefined. Eight bits are used to index each code vector. Here each vector is compressed into one byte. So size of the compressed image is 6864 bytes and size of the code book is 2048 bytes. Finale file is compressed image together with the code book. So size of final file is 9912 bytes. Note that in this prototype image file is very small, compression ratio is around 12. If the image file is very large, size of the code book is comparatively very small. So for very large file, we can get the compression ratio around 15.

In this experiment, 5500 bytes are different between the image reconstructed from the compressed fileand the original image. Loss of information is around 5 percent. If we reduce the loss of information 2 or 3 percent more, the reconstructed image obtained from the compressed file seems to be original image. Classification of code vectors is necessary to reduce the loss of information.

A pictorial representation of this process is shown below

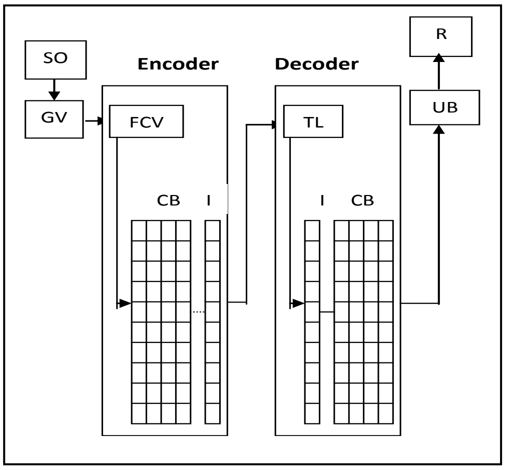

Fig. 1. The Vector quantization procedure

SO	- Source Output	GV	- Group into Vectors
FCV	- Find Closest Code Vector	TL	- Table Lookup
R	- Reconstruction	UB	- UnBlock
CB	- Code Book	I	- Index

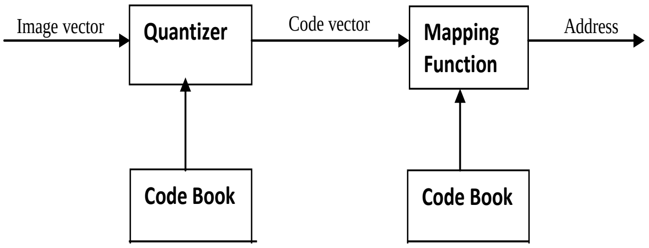

Fig. 2. Encoder

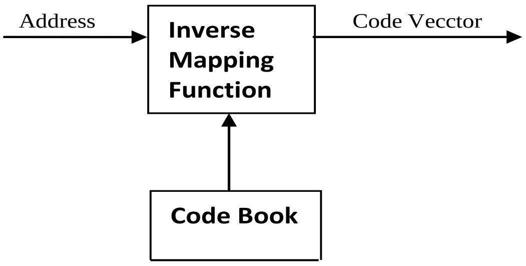

Fig. 3. Decoder

REFERENCES

[1] Khalid Sayood : Introduction to Data Compression Third edition, Morgan Kaufmann publishers.

[2] Cappellini , V,Ed. 1985. Data Compression and Error Control Techniques with Applications.

[3] Cortesi,D.1982 : An effective Text Compression Algorithm, Byte, 7, 1 (Jan) 397-403.

[4] Aravind R,GershoA: Image Compression based on vector quantization with finite memory: Optical engineering 1987,26(7), 570-590.

[5] Goldberg M, Sun H: Image sequence coding using vector quantization , IEEE Transaction on communications1986, 24(7)703-10.

[6] Hang H, Woods J W: Predictive vector quantization of images, IEEE Transaction on communications 1985, 33(11), 1208 – 19.

[7] Linde Y, Buzo A,Gray R M:An algorithm for vector quantization coding, IEEE Transaction on communications 1980,28(1) 85-95.

[8] Netravali A N : Interpolative picture code using a subjective criterion. IEEE Transactions on communications 1977, 25(5), 503-7

[9] Pickholtz R L. Loew M H : Combined transform coding scheme for image data compression, IEEE Transaction on Consumer electronics 1991, 37(1), 45-51.

[10] Ramamurthy B, Gersho A: Classified vector quantization of image, IEEE Transaction on communication 1986, 34(11), 1105-15

[11] Storer J A : Data Compression : Methods and theory, Computer science press, Brandeis University, USA , 1986.

[12] Margaret H. Dunham : Data Mining , Introductory and Advance Topics