Unit - III

Data Compression

Coding Requirement

→ Compression algorithm

Implement a suitable compression algorithm
Ex- Huffman coding.

→ Data Encoding

Write routines to encode and decode data using the chosen algorithm

→ Error Handling

Implement error detection and correction mechanism. Ex- Checksum

→ Speed And Efficiency

Ensure fast compression and decompression speed.

→ Memory Management

Manage memory usage during compression and decompression.

Source

Here are some common source

→ Text files

Documents, logs, and configuration files.

→ Images

Photos, graphics and icons
Ex- JPEG, JPG, PNG

→ Audio files

Music, voice recording and podcast Ex - (MP3)

→ Video files

Movies, TV shows, video clips Ex (MP4 ...)

→ Network traffic

web traffic, email and other network configuration

→ Back up data

Back up files, archives etc

→ Cloud storage

Data stored in cloud storage such as google cloud, Azure.

Entropy Coding

Entropy in data compression refers to the measure of uncertainty or randomness in data.

It represent the amount of information in the data that cannot be compressed further.

Entropy is typically measured in bits per byte or bits per symbol.

→ In data compression, Entropy is used to :-

- Quantify compressibility
- Choose compression algorithm
- Optimize compression

→ Entropy coding techniques

- Huffman Coding

  Assign shorter codes to more frequent symbols.

- Arithmetic Coding

  Encodes data by representing a probability distribution over a large number of possibilities.

- Rice Coding

A simple and efficient entropy coding techniques used in various compression algorithms.

→ Entropy Coding is used in many data compression algorithm and formats :-

- JPEG (Image compression)
- MP3 (Audio compression)
- MPEG (Video compression)

Hybrid Coding

Hybrid coding in data compression combines multiple coding techniques to achieve better compression ratio and efficiency.

→ Hybrid Coding technique

- Transform coding + Entropy coding

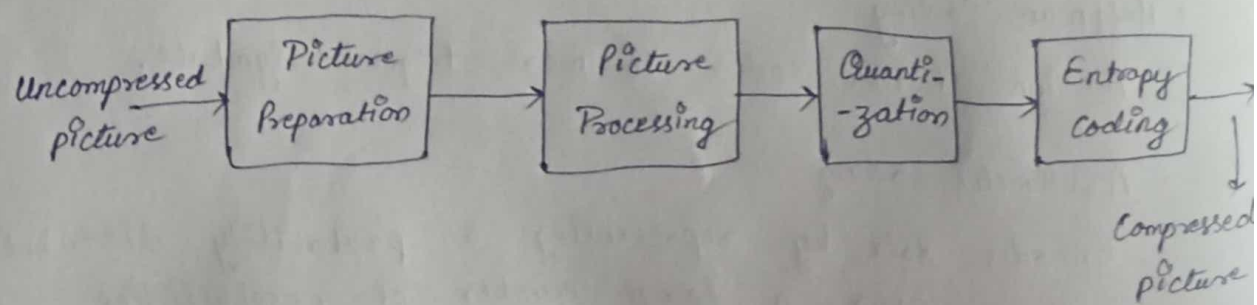
  Combines transform coding with entropy coding for image and video compression.

- Predictive coding + Entropy coding

  Combines predictive with entropy coding for audio compression.

→ Hybrid Coding is used in various data compression Ex- JPEG, MP3, MPEG, AVC etc.

Major Steps of Data Compression



JPEG

- Joint photographic Expert Group
- JPEG is a widely used image compression standard for photographic images.
- Some Requirements of JPEG

→ JPEG implementation should be independent of image size.

→ JPEG implementation should be applicable to any image and pixel aspect ratio.

→ Color representation itself should be independent.

Steps in JPEG Compression

(1) Image Division into blocks

JPEG compression process start by dividing the image into smaller block of pixels. Standard block size is 8×8 pixel. This means each block contain 64 pixels.

(2) Color Space Conversion

Image is often converted from RGB color space (Red, Green, Blue) to YCbCr color space. In YCbCr, the Y component represent the brightness, Cb and Cr represent the color information.

(3) Subsampling

Chromiance component (Cb and Cr) are subsampled.

Ex- Common Subsampling method is 4:2:0

④ Block Preparation

Each 8x8 block of the Y, Cb and Cr component is processed separately.

⑤ Discrete Cosine Trans-form (DCT)

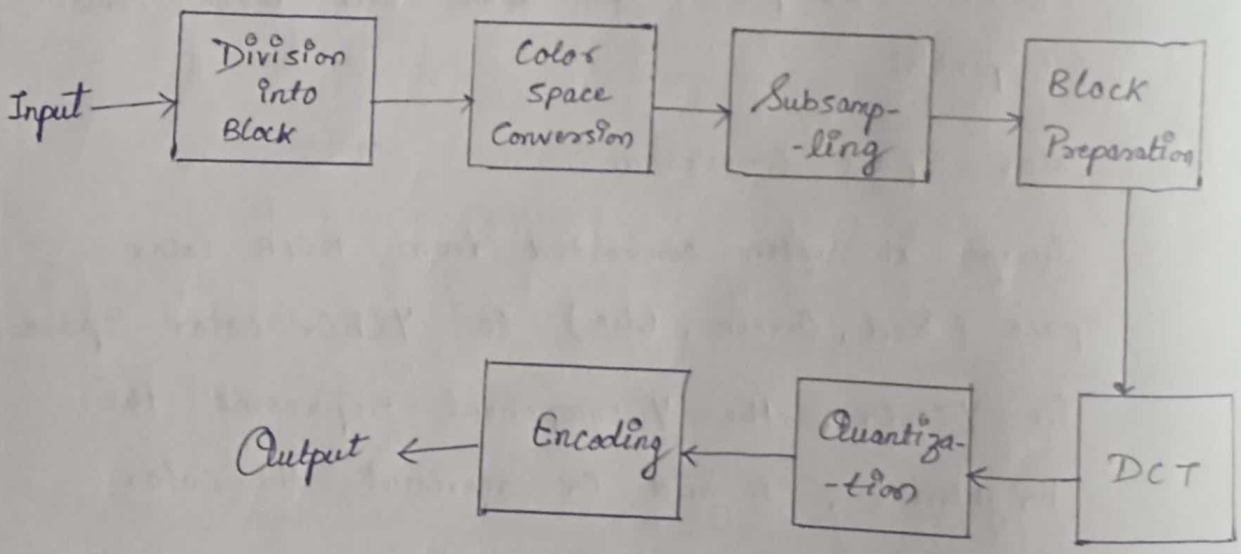
After preparing the block, the Discrete Cosin transform is applied to each block. DCT Convert the spatial domain data into frequency domain data.

⑥ Quantization

DCT coefficient are then quantized using a quantization matrix.

⑦ Encoding

The quantized value are encoded using technique like Huffman coding to reduce file size.



MPEG

→ Moving Picture Expert Group

→ MPEG is a working group that develops standard for audio and video compression and transmission.

→ The MPEG standard are widely used for compressing and encoding audio and video file.

Steps in MPEG Conversion

① Sampling

This is the first step where the analog audio or video signal is converted into digital format.

② Quantization

This process reduces the precision of the sampled data. In audio, this means reducing the number of bits. In video, it involves reducing the color depth.

③ Motion Compensation

This technique is used to exploit temporal redundancy in video sequence.

④ Transform Coding

This involves converting the original pixel value into a frequency domain using mathematical transformation, often Discrete Cosine transform. (DCT)

⑤ Entropy Coding

This step involves applying lossless compression technique to reduce the file size.

Ex :- Huffman coding, Run-length Encoding (RLE)

⑥ Bitstream formatting

After compression, the data is organized into a specific format that can be easily stored or transmitted.

Huffman Coding

Huffman Coding is a data compression algorithm that is used to reduce the size of data without losing any information.

Types

① Static Huffman Coding

The static huffman technique is a method used in data compression to assign variable-length codes to symbols based on their frequencies of occurrence.

Techniques

- Frequency Calculation

  Calculate the frequency of each symbol in dataset.

- Huffman Tree construction

  Build a binary tree where each leaf node represent a symbol and the path from the root to the leaf node represent code.

- Code Assignment

  Assign shorter code to more frequent symbol and longer codes to less frequent symbol.

- Encoding

  Replace each symbol with its corresponding Huffman code

Characterstics

- Static - Built once and remain unchanged
- Variable - length code - Codes have different length based on symbol frequencies.
- Prefix - free - Ensuring unique decoding.

Advantages

- Efficient compression
- fast encoding and decoding

Disadvantage

- Limited adaptability
- Sub optimal performance.

(2) Dynamic Huffman Coding

Dynamic Huffman coding is an advanced compression technique that allows for the efficient encoding of data by adapting to changes in symbol frequencies as the data is processed.

-

Technique

① Initialization

Start with an intial empty Huffman tree or a tree that include a predefined set of symbols

② Processing Symbol

As each symbol is read from the data stream, its frequency count is updated.

③ Tree Update

When a new symbol is added, the tree may need to be restructed to maintain the properties of huffman coding.

④ Encoding & Decoding

Each time a symbol is processed, it code is output immediately.

⑤ Output

The output include both the encoded data and the necessary information to reconstruct the tree for decoding.

Advantage

- Adaptability
- No preprocessing

Disadvantage

- Complexity
- Overhead.

Statistical Coding technique

Statistical coding technique method are used in data compression that rely on the statistical properties of the data to present it more efficiently

* Common Statistical Coding techniques :-

① Huffman Coding

A widely used method that creates a binary tree based on the frequencies of symbol. It can be static or dynamic.

② Arithmetic Coding

Represent the entire message as a single number in a specific range. Compressing the entire message into a single value.

③ Run length Encoding

This technique is effective for data with long sequence of repeated symbols.

Ex :- The sequence 'AAAAA' could be encoded as '5A'.

* These technique are widely used in various application :-

- Text compression
- Image compression
- Video compression
- Audio compression

Types of Statistical Coding technique :-

→ Lossless compression

Reconstruct the original data exactly.

Ex :- Huffman coding

→ Lossy Compression

Losses some data during compression, but achieves higher compression ratio.

Ex :- JPEG, MP3