Why is Encoding Important?

Video encoding is important because it allows us to more easily transmit video content over the internet. In video streaming, encoding is crucial because the compressing of the raw video reduces the bandwidth making it easier to transmit, while still maintaining a good quality of experience for end viewers. If all the video content was not compressed, available bandwidth on the Internet would be inadequate to transmit all of it and prevent us from deploying widespread, distributed video playback services. The fact that we can stream video on multiple devices in our homes, on-the-go using mobile, or even while video chatting with loved ones across the globe, even with low bandwidth, is owed to video encoding.

Scanned with CamScanner

An **inter frame** is a frame in a video compression stream which is expressed in terms of one or more neighboring frames. The "inter" part of the term refers to the use of *Inter frame prediction*. This kind of prediction tries to take advantage from temporal redundancy between neighboring frames enabling higher compression rates.

Scanned with CamScanner

20.3 Motion estimation and motion compensation

While explaining the block diagram of a generic video codec in Section 20.1, we identified motion estimation and motion compensation as two major blocks which are new additions as compared to the building blocks of an image codec.

The motion estimation block in a video codec computes the displacement between the current frame and a stored past frame that is used as the reference. Usually the immediate past frame is considered to be the reference. More recent video coding standards, such as the H.264 offer flexibility in selecting the references frames and their combinations can be chosen. Fig. 20.2 illustrates the basic philosophy of motion estimation. We consider a pixel belonging to the current frame, in association with its neighborhood as the candidates and then determine its best matching position in the references frame. The difference in position between the candidates and its match in the reference frame is defined as the *displacement vector* or more commonly, the *motion vector*. It is called a vector since it has both horizontal and vertical components of displacement. We shall offer a more formal treatment to motion estimation in the next sections.

After determining the motion vectors one can predict the current frame by applying the displacements corresponding to the motion vectors on the reference frame. This is the role of the motion compensation unit. The motion compensation unit therefore composes how the current frame should have looked if corresponding displacements were applied at different regions of the reference frame.

Scanned with CamScanner

Motion estimation is the process of determining motion vectors that describe the transformation from one 2D image to another; usually from adjacent frames in a video sequence. It is an ill-posed problem as the motion is in three dimensions but the images are a projection of the 3D scene onto a 2D plane. The motion vectors may relate to the whole image (global motion estimation) or specific parts, such as rectangular blocks, arbitrary shaped patches or even per pixel. The motion vectors may be represented by a translational model or many other models that can approximate the motion of a real video camera, such as rotation and translation in all three dimensions and zoom.

Scanned with CamScanner

2.1.1 Block Diagram of Encoder/Decoder

Compression system involves two pair, a compressor (encoder) and a de-compressor (decoder). The encoder converts the source data into a compressed form (occupying a reduced number of bits) prior to transmission or storage and the decoder reconverts the compressed form back to the original data. The encoder/decoder pair is often described as a CODEC and its block diagram has shown in figure 2.1.

Figure 2.1: Block Diagram of CODEC (Encoder/Decoder)

2.1.2 Compression Ratio

Compression ratio, also known as compression power, is a computer science term used to quantify the reduction in data-representation size produced by a data compression algorithm. The data compression ratio is analogous to the physical compression ratio used to measure physical compression of substances.

Compression ratio is defined as the ratio between the uncompressed size and compressed size:

$$Compression\ Ratio = \frac{Uncompressed\ Size}{Compressed\ Size} \tag{2.1}$$

Thus a representation that compresses a 10 MB file to 2 MB has a compression ratio of 10/2 = 5, it can also be represented as an explicit ratio, 5:1 (read as "five" to "one"), or as an implicit ratio, 5/1.

Sometimes the space savings is given instead compression size, which is defined as the reduction in size relative to the uncompressed size:

Scanned with CamScanner

$$Space\ Saving = 1 - \frac{Uncompressed\ Size}{Compressed\ Size} \quad (2.2)$$

Thus a representation that compresses a 10MB file to 2MB would yield a space savings of 1 - 2/10 = 0.8, often notated in percentage as 80%.

2.2 Types of Compression

There are two type of compression involved in information technology.

1. Lossless Compression
2. Lossy Compression

2.2.1 Lossless Compression

A lossless compression is a class of compression in which the original data is reconstructed from compressed data exactly. It involves no loss of information. Lossless compression is used when it is important that the original and the reconstructed data are identical. For examples, text, message, zip file format, PNG (Portable Network Graphics) or GIF (Graphics Interchange Format).

2.2.2 Lossy Compression

A lossy compression is one where compressing and decompressing data retrieve may be different from original but it is close enough to be useful in many ways. Lossy compression involves some loss of information and data cannot be reconstructed exactly. In lossy compression, a higher compression ratio is achieved as compared to lossless compression. The examples of lossy compression are sound, image or video.

Scanned with CamScanner

2.4.1Video CODEC

A video codec is an electronic circuit or hardware that compresses and decompresses digital video. It converts raw (uncompressed) digital video to a compressed format or vice-versa. In the context of video compression, codec is a concatenation of encoder and decoder. A device that only compresses is typically called an encoder, and one that only decompresses is a decoder.

The block diagram of video encoder isgiven in figure 2.2. It consists of three main functional units: a temporal model, a spatial model and an entropy encoder. The input to the temporal model is an uncompressed video sequence. The temporal model reduces temporal redundancy by exploiting the similarities between neighbouring video frames, usually by constructing a prediction of the current video frame(figure 2.2). In MPEG-4 Visual and H.264 video standards, the prediction is formed from one or more previous or future frames (figure2.3) and is improved by compensating for differences between the frames (motion compensated prediction). The output of the temporal model is a residual frame (figure 2.4) (created by subtracting the prediction from the actual current frame) and a set of model parameters, typically a set of motion vectors describing how the motion was compensated. The residual frame forms the input to the spatial model which makes use of similarities between neighboringsamples in the residual frame to reduce spatial redundancy. This is achieved by applying a transform coding to the residual samples and quantifying the results. The transform converts the samples into another domain in which they are represented by transform coefficients.

Scanned with CamScanner

The coefficients are quantized to remove insignificantvalues, leaving a small number of significant coefficients that provide a more compact representation of the residual frame. The output ofthespatial model is a set of quantized transformcoefficients. The parameters of the temporal model (typically motion vectors) and the spatial model (coefficients) are compressed by the entropy encoder. This removes statistical redundancy in the data (for example, representing commonly-occurring vectors and coefficients by short binary codes) and produces a compressed bit stream or file that may be transmitted and/or stored. A compressed sequence consists of coded motion vector parameters, coded residual coefficients and header information.

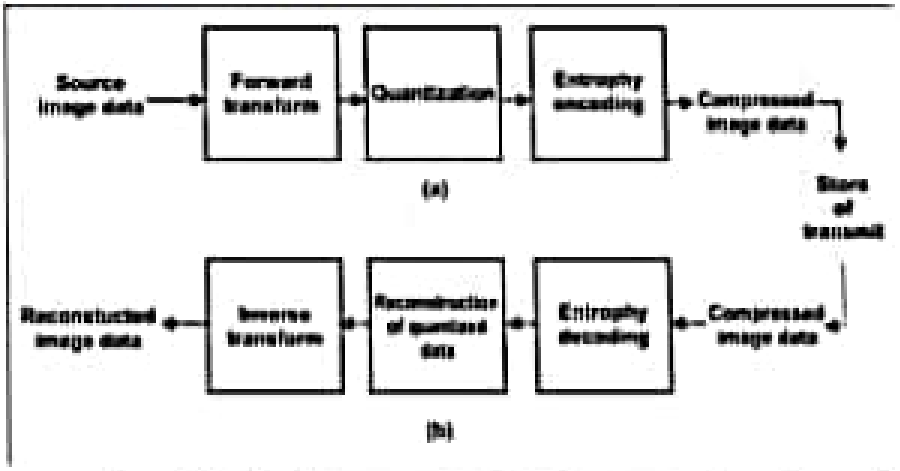

Figure 2.5: Block Diagram of VideoCodec: (a) Encoder, (b) Decoder

The video decoder reconstructs a video frame from the compressed bit stream. The coefficients and motion vectors are decoded by an entropy decoder after which the spatial model is decoded to reconstruct a version of the residual frame. The decoder uses the motion vector parameters, together with one or more previously decoded frames, to create a prediction of the current frame and the frame itself is reconstructed by adding the residual frame to this prediction.

Scanned with CamScanner

2.4.2 Types of Video Compression

Video compression refersto reducing the quantity of data used to represent, and is the combination of spatial image compression and temporal motion compensation. Most of video compressions are lossy. It operates on the premise that much of the data present before compression is not necessary for achieving good perceptual quality [2, 3]. For example, DVDs use a video coding standard called MPEG-2 that can compress around two hours of video data by 15 to 30 times, without compromising picture quality that is generally considered high-quality for standard-definition video. Video compression is a tradeoff between disk space, video quality, and the cost of hardware required to decompress the video in a reasonable time. However, if the video is over compressed in a lossy manner, visible (and sometimes distracting) artifacts can appear.

Video compression typically operates on square-shaped groups of neighboring pixels, often called macro blocks. These pixel groups or blocks of pixels are compared from one frame to the next and the video compression codec (encode/decode scheme) sends only the differences within those blocks. This works extremely well if the video has no motion. A still frame of text, for example, can be repeated with very little transmitted data. In areas of video with more motion, more pixels change from one frame to the next. When more pixels change, the video compression scheme must send more data to keep up with the larger number of pixels that are changing. If the video content includes an explosion, flames, a flock of thousands of birds, or any other image with a great deal of high-frequency profiles leads to the quality will decrease or to increase in the variable bitrates that must be increased to render this added information with the same level profiles[7, 8].

Scanned with CamScanner

2.4.2.1 Intra-frame Video Compression

Intra-frame compression is a compression applied to still images, such as photographs and diagrams, and exploits the redundancy within the image, known as spatial redundancy. Intra-frame compression techniques can be applied to individual frames of a video sequence.

2.4.2.2 Inter-frame Video Compression

Inter-frame compression is compression applied to a sequence of video frames, rather than a single image. Inter-frame compression exploits the similarities between successive frames, known as temporal redundancy that assists in reducing the volume of data required to describe the sequence.

There are several inter-frame compression techniques, of various degrees of complexity. Most of which attempt to more efficiently describe the sequence by reusing parts of frames the receiver already has, in order to construct new frames.

2.5 Video Compression Standards

Two International bodies are responsible for representing the standards in video compression/coding. First one is International Standards Organization (ISO) and another is International Telecommunication Union (ITU). They have developed a series of video compression standards that have shaped the development of the visual communication industry. The following two groups have significant contributions:

- Moving Picture Expert Group (MPEG) developed by ISO/IEC
- Video Coding Expert Group (VGEG) developed by ITU-T

Scanned with CamScanner

2.5.1 Moving Picture Expert Group (MPEG)

In 1992, MPEG-1 was created as the first standard for encoding moving pictures accompanied by sound. The aim was to achieve a picture quality close to that of VHS at CD data rates (< 1.5 Mbps). MPEG-1 was provided only for applications on storage media (CD, hard disk) and not for transmission (broadcasting) and its data structures correspond to this objective. The audio and video coding of MPEG-1 is quite close to that of MPEG-2 and all the fundamental algorithms and methods are already in place. There are I, P and B frames, i.e. forward and backward prediction, and naturally there is the DCT-based irrelevance reduction methods already found in JPEG. The picture resolution, however, is limited to about half the VGA resolution (352 × 288). Neither there is any necessity for field encoding (interlaced scanning method). In MPEG-1, there is only the so-called Program Stream (PS) which is composed of multiplexed packetized elementary stream (PES) packets of audio and video. The variable-length (64 Kbytes max) audio and video PES packets are simply assembled interleaved in accordance with the present data rate to form a data stream. This data stream is not processed any further since it is only intended to be stored on storage media and not used for transmission. A certain number of audio and video PES packets are combined to form a so-called pack which consists of a header and the payload just like the Packetized Elementary System (PES) packets themselves [13].

The MPEG-1 standard consists of three parts. Part-1 deals with system issues (including the multiplexing of coded video and audio). Part 2 deals with compressed video and Part 3 with compressed audio. Part 2 (video) was developed with the aim of supporting efficient coding of video

Scanned with CamScanner

for CD playback application and achieving video quality comparable to, or better than, VHS video tape at CD bit rate (around 1.2Mbps for video). There was need to minimize decoding complexity since most consumer applications were envisaged to involve decoding and playback only, not encoding. Hence MPEG-1 decoding is considerably simpler than encoding.

MPEG-2 is an extension of the MPEG-1 international standard for digital compression of audio and video signals. MPEG-1 was designed to code progressively scanned video at bit rates up to about 1.5 Mbps for applications such as CD-i (compact disc interactive). MPEG-2 is directed at broadcast formats at higher data rates. It provides extra algorithmic tools for efficiently coding interlaced video, supports a wide range of bit rates and provides for multichannel surround sound coding. This tutorial paper introduces the principles used for compressing video according to the MPEG-2 standard, outlines the general structure of a video coder and decoder, and describes the subsets (profiles) of the toolkit and the sets of constraints on parameter values (levels) defined to date [13].

MPEG-2 consists of three main sections: Video (described below), Audio (based on MPEG-1 audio coding) and System ((MPEG-1 System, multiplexing and transmission of the coded audio/visual stream). MPEG-2 is almost a superset of MPEG-1 video standards. Most MPEG-1 video sequences should be decodable by an MPEG-2.

MPEG-4 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), the committee that also developed standards known as MPEG-1 and MPEG-2. These standards made interactive video on CD-ROM, DVD and Digital Television possible. MPEG-4 is the result of another international effort involving hundreds of researchers and

Scanned with CamScanner

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos.[1] Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Scanned with CamScanner

2 MOTION SEGMENTATION

A simple model for segmentation may consist of grouping pixels that have similar velocity. In scenes where objects are undergoing simple translation, this model may provide a sufficient description. However, this model when applied on general image sequences will result in a highly fragmented description. For example, when an object is rotating, each point on the object exhibits a different velocity resulting a segmentation map that consists of many small regions. The problem lies in the image model used in the segmentation. This model, although it requires a small encoding overhead, is insufficient for describing typical image sequences. The ideal scene segmentation, however, requires 3-D object and shape estimation which remains difficult and computationally intensive. A model less complicated than the 3-D model must be employed.

A reasonable solution can be found by extending the simple translation motion model to allow for linear change in motion over spatial dimensions. This affine motion model consists of only six parameters and can describe motions commonly encountered in video sequences. These include: translation, rotation, zoom, shear, and any linear combination of these. Affine motion is defined by the equations:

$$V_x(x, y) = a_{x0} + a_{xx}x + a_{xy}y \quad (1)$$

$$V_y(x, y) = a_{y0} + a_{yx}x + a_{yy}y \quad (2)$$

where V_x and V_y are the x and y components of velocity, and the a_{ij} are the parameters of the transformation. We use the affine motion model in our decomposition of video data into the layered representation.

Image segmentation based on the affine motion model result in identifying piecewise linear motion regions. Affine motions have been shown to provide adequate description of object motion while being easily computable.[2,4,6] It can be shown that motion of 3-D planar surfaces under orthographic projection induce affine motions, thus, affine motion regions have a physical interpretation.

Scanned with CamScanner