

Multimedia Communications

Video compression



Video compression

- Of all the different sources of data, video produces the largest amount of data
- There are some differences in our perception with regard to image and video
- Motion involved in video may mask some of the artifacts and on the other hand artifact that are not visible in still images can be annoying in video.
- A random change in average intensity of the pixels is not important in images but can be quite annoying in video sequences.

Video compression

- Video compression algorithms have different applications
- The application determines the features to be used and the values of parameters.
- Two-way communication: coding delay should be minimal, compression and decompression should have about the same level of complexity
- Broadcast applications: complexity can be unbalanced, more tolerance for encoding delay
- Encoding is generally not done in real time, so the encoder can be quite complex.
- When video is transmitted over packet networks, effects of packet loss have to be taken into account.

Video compression

- In most video sequences there is little change in the content of the image from one frame to the next
- Most video compression schemes take advantage of this redundancy by using previous frame to generate a prediction for the current frame.
- In order to use a previous frame to predict the pixel values in the frame being encoded, we have to take the motion of objects in the image into account.
- The practical approach to do this is block-based motion compensation

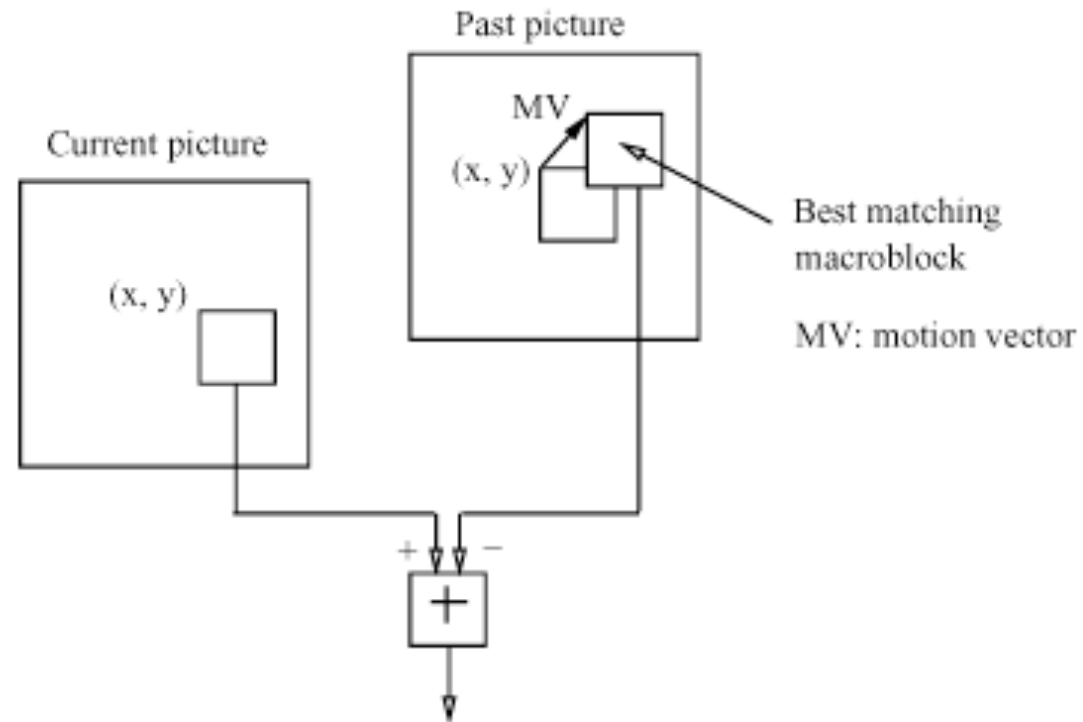
Motion compensation

- Frame being encoded is divided into blocks of size $M \times M$.
- For each block, we search the previous reconstructed frame for the $M \times M$ block that most closely matches the block being encoded.
- We can measure the closeness of a match using a distance measure (e.g., sum of absolute differences)
- If the distance from the block being encoded to the closest block in the previous reconstructed frame is greater than some prespecified threshold, the block is coded without the prediction. This decision is communicated to the decoder.

Motion compensation

- If the distance is below the threshold, motion vector is transmitted to the receiver.
- Motion vector: relative location of the block to be used for prediction
- Motion vector is obtained by subtraction the coordinate of the upper-left corner pixel of the block being coded from the coordinates of the upper-left corner pixel of the block used for prediction.
- There are algorithms in which motion vectors are measured in half pixels.
- In order to do this, pixels of the coded frame being searched are interpolated to obtain twice as many pixels as in the original frame.

Motion compensation



Motion compensation

- Motion compensation requires a large amount of computations.
- Consider finding a match for an 8x8 block. Each comparison requires taking 64 differences and then computing the sum of the absolute values of the differences.
- If we search the previous frame within 20 pixels in horizontal or vertical directions we should perform 1681 (41x41) comparison.
- How to cut the number of computations?

Motion compensation

- Increase the block size: more computations per comparison but less comparison.
 - Drawback: probability that a block contains objects moving in different directions increases
 - This means a less effective prediction
- Reduce the search space: increases the probability of missing a match
- There is tradeoff between computation and the amount of compression

Video signal

- Composite color signal consists of a luminance component and two chrominance components.

$$Y=0.299R+0.587G+0.114B$$

$$C_b=B-Y$$

$$C_r=R-Y$$

- CCIR has developed a standard for sampling of analog video signals in its recommendation 601-2
- Sampling rate is a multiple (up to 4) of a base sampling frequency of 3.725 MHz.
- Sampling rate is represented as a triple of integer numbers, the first one corresponds to the sampling of the luminance component and the remaining two corresponds to chrominance components.

Video signal

- 4:2:2 : luminance is sampled at 13.5 MHz, and the chrominance components at 6.75.
- Common interchange format (CIF): $Y=288 \times 352$, C_b and C_r are 144×176 .
- In QCIF (Quarter CIF), we have half the number of pixels in both rows and columns
- SIF: Y is 360×240 and C_r and C_b are 180×120 .

Video compression

| Standard Organization | Video-coding standard | Typical range | Typical application |
|-----------------------|-----------------------|--|--|
| ITU-T | H.261 | $p \times 64$ Kb/s $p=1,2,\dots,30$ | ISDN video phone |
| ISO | MPEG-1 | 1.2 Mb/s | CD-ROM |
| ISO | MPEG-2 | 4-80 Mb/s | HDTV |
| ITU-T | H.263 | 64 Kb/s or below | PSTN video phone |
| ISO | MPEG-4 | 24-1024 Kb/s | Interactive audio/ video |
| ITU-T | H.263 + | <64 Kb/s | PSTN video phone |
| ITU-T/ISO | H.264 | <64 Kb/s | Network-friendly packet-based video |

H.261

- Earliest DCT-based video coding standard is ITU-T H. 261.
- It assumes the input to be in CIF or QCIF format
- Cb and Cr have half the resolution of Y (in each of directions)
- An input frame is divided into blocks of 8x8 pixels.
- One block of Cb and Cr correspond to 4 blocks of Y
- Collection of these six blocks is called a macroblock (MB)
- In H.261 33 MBs are grouped together and called a Group of Blocks (GOB)
- For a given block, we subtract the prediction generated using the previous frame

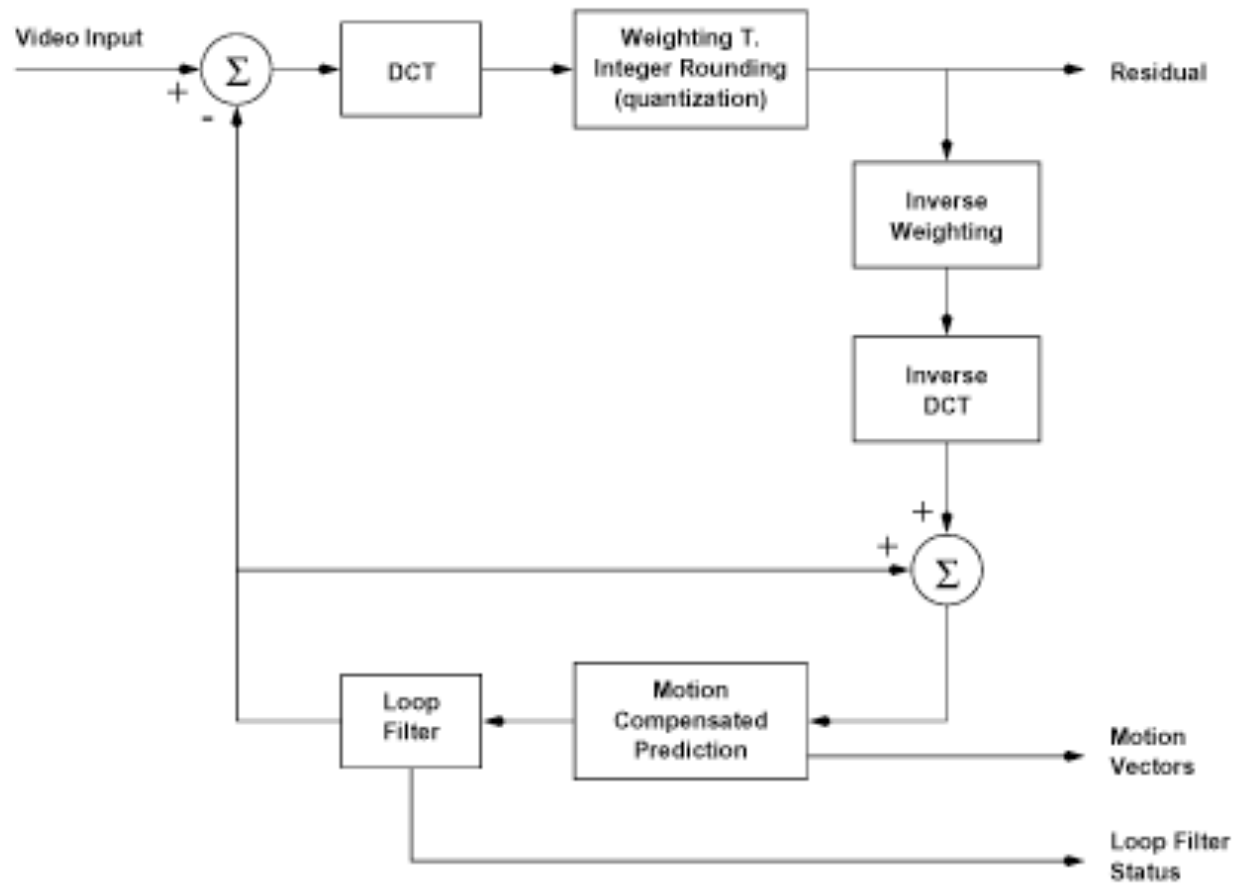
H.261

- The difference between the block being encoded and the prediction is transformed using DCT.
- Transform coefficients are quantized and the quantization label encoded using a variable length code.
- Motion compensation: four 8x8 luminance blocks of a macroblock are used
- For each macroblock an area of ± 15 in the previous reconstructed frame is search
- Motion vectors of the chrominance blocks are set to half of motion vectors for luminance.

H.261

- Since the prediction of the current frame is composed of blocks at various locations in the reference frame, the prediction may contain coding noise and blocking artifacts
- This in turn, causes high values for the high-frequency coefficients in the transform which can increase the rate
- To avoid this, prior to taking the difference, prediction block can be smoothed by using a 2-D spatial filter (loop filter)
- The filter is separable and can be implemented as a 1-D filter operated on rows and then on the columns.
- Coefficients of the filter are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$.

H.261



H.261

- DCT transform is performed on 8x8 blocks of pixels or pixel differences and then quantized
- Encoder is said to be in intra mode if it operates directly on the input image without the use of motion compensation. Otherwise it is said to be in inter mode.
- Depending on how good or poor the prediction is, we can get a wide variation in the characteristics of the coefficients that are to be quantized
- In the case of intra blocks, the DC coefficients will take on much larger values than the other coefficients.

H.261

- H.261 has 32 different quantizer.
- One quantizer is reserved for intra DC coefficient while the remaining 31 are used for other coefficients.
- Intra DC quantizer is a uniform mid-rise quantizer with a step size of 8.
- The other quantizers are midtread with a step size of an even value between 2 and 62.
- Smaller step size for quantizer => larger number of nonzero coefficients => higher rate
- The quantizer used for each macroblock is determined in its header

H.261

Encoding of quantization labels:

- Labels are scanned in zigzag fashion. Nonzero labels are coded along with the number (run) of coefficients quantized to zero.
- 20 most commonly occurring combinations of (run, label) are coded with VLC.
- All other combinations of (run,label) are coded with a 20-bit word, made up of 6 bit escape sequence, 6-bit denoting the run and 8-bit code for the label.
- In order to avoid transmitting blocks that have no nonzero quantized coefficient, the header preceding each macroblock can contain a variable length code called coded block pattern (CBP)

H.261

- CBP indicates which of the six block contain nonzero labels.
- CBP can take on one of 64 different pattern numbers which are encoded using VLC.
- The pattern number is given by:
$$\text{CBP} = 32P_1 + 16P_2 + 8P_3 + 4P_4 + 2P_5 + P_6$$
$$P_i \text{ is } 1 \text{ if its corresponding block has a nonzero quantized coefficient.}$$

H.261

- Binary codewords form the input to a transmission buffer.
- Transmission buffer: to keep the output rate of the encoder fixed
- If the buffer starts to fill up faster than transmission rate, it sends a message back to the coder to reduce the bit.
 - This can be done by using larger step sizes for quantizer or by dropping some of the frames.
- If the buffer is in danger of becoming empty because the coder is providing bits at a rate lower than the transmission rate, the transmission buffer can request a higher rate from the coder

H.261 & H.263

- H.261 was primarily designed for videophone and videoconferencing.
- It operates at very low bit rate: $p \times 64$ kbit/s (p is a number between 1 to 30)
- H.263: a video coding standard for very low bit rate applications (less than 64 Kb/s)
- In terms of SNR, H.263 can provide a 3 to 4 dB gain over H.261
- Since H.263 was built on top of H.261, the main structure of two standards are essentially the same

H.263

- Major differences between them include:
 - H.263 supports more picture formats (in addition to CIF and QCIF it supports sub-QCIF, 4CIF and 16 CIF)
 - H.263 uses a different GOB structure (a GOB is at least one full row of MBs)
 - H.263 uses half-pel motion compensation, but does not support loop filtering (Unlike H.261 motion vectors might be non-integers such as (4.5,-2.5). Bilinear interpolation is used to find the corresponding pel values for prediction)
 - In addition to basic coding algorithms, four options in H.263 that are negotiable between encoder and decoder. These are: unrestricted motion vector, syntax-based arithmetic coding, advanced prediction mode and PB-frame mode.

H263

- Unrestricted Motion Vector mode: motion vectors are allowed to point outside the picture.
 - Edge pels are used as prediction for the "not existing" pels.
 - A significant gain is achieved if there is movement along the edge of the pictures, especially for the smaller picture formats.
- Advanced Prediction mode: Four 8x8 vectors instead of one 16x16 vector are used for some of the macro blocks in the picture, and motion vectors are allowed to point outside the picture as in the UTMV mode above.
 - The encoder has to decide which type of vectors to use. Four vectors use more bits, but give better prediction.
- Syntax-based Arithmetic Coding mode: Arithmetic coding is used instead of VLC coding. The SNR and reconstructed frames will be the same, but generally fewer bits will be produced. The average gain for inter frames is 3-4%. This gain depends on the sequence, the bit rate and other options used. For intra blocks and frames, the gain is higher, on average about 10%.

H263

- **PB-frames** mode: A PB-frame consists of two pictures being coded as one unit.
 - A PB-frame consists of one P-picture which is predicted from the last decoded P-picture and one B-picture which is predicted from both the last decoded P-picture and the P-picture currently being decoded.
 - This last picture is called a B-picture, because parts of it may be bi-directionally predicted from the past and future P-pictures.
 - For relatively simple sequences, the frame rate can be doubled with this mode without increasing the bit rate much.
- These options are negotiable. This means the decoder signals the encoder which of the options it has the capability to decode. If the encoder has any of these options, it can then turn them on, and for each of the options used the quality of the decoded video-sequence will increase.

H.263 version 2 (H.263+)

- H.263+ contains approximately 12 new features that do not exist in H.263
- These include new coding modes that improve compression efficiency, support for scalable bit streams, several new features to support packet networks and error prone environments, added functionality and support for a variety of video frames.

MPEG

- In some applications it is cost effective to shift more of the computational burden to the encoder
 - A video sequence compressed and stored on a CD
 - Broadcast applications
- Standards developed for this type of applications MPEG (Moving Picture Expert Group)
- MPEG-1: 1.5 Mbit/s
- MPEG-2: 10 Mbit/s and above
- MPEG-4: object-oriented framework for encoding of multimedia

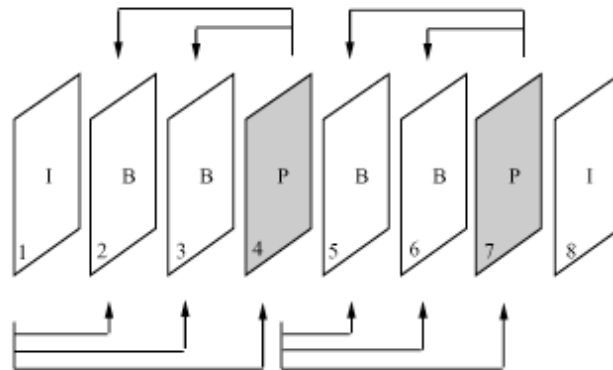
MPEG-1

- Basic structure of the compression algorithm is very similar to H.261.
- Differences between MPEG-1 and H.261 stems from the application each one is designed for: videoconferencing for H.261 and digital storage and retrieval for MPEG-1
- One of the differences is that in storage applications the user might want to access a frame in the middle of a video sequence (random access)
- In MPEG-1 this capability is provided by requiring that there be frames periodically that are coded without any reference to past frames (I frames)

MPEG-1

- Because I frames do not use temporal correlation, the compression rate is low
- Number of frames between two consecutive I frames is a tradeoff between compression efficiency and fast picture acquisition capability of I frames.
- MPEG-1 has two other kinds of frames: predictive coded (P frames) and bidirectionally predictive coded (B frames)
- P frames are coded using motion compensation prediction from the last I or P frame, whichever happens to be closest.
- Compression efficiency of P frames is substantially higher than I frames.
- I and P frames are called anchor frames.

MPEG-1



MPEG-1

- B frames achieve a high level of compression by using motion compensation prediction from the most recent anchor frame and the closest future anchor frame.
- By using both past and future for prediction, generally we can get better compression than if we only use prediction based on the past.
- A B-frame can only be generated after the future anchor frame has been generated.
- A B-frame is not used for predicting any other frame
- A group of pictures (GOP) is the smallest random access unit in the video sequence.

MPEG-1

- A GOP has to contain at least one I frame
- A GOP can begin with either a B frame or an I frame and must end with either an I or a P frame.
- Because of reliance of B frames on future anchor frames, there are two different sequence order in MPEG-1: display sequence and bitstream order.

| | | | | | | |
|---|---|---|---|---|---|---|
| I | B | B | P | B | B | P |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

| | | | | | | |
|---|---|---|---|---|---|---|
| I | P | B | B | P | B | B |
| 1 | 4 | 2 | 3 | 7 | 5 | 6 |

MPEG-1

- MPEG committee has provided some suggested values for various parameters (constrained parameters)
- Horizontal picture size: less than or equal to 768 pixels
- Vertical picture size: less than or equal to 576 pixels
- Pixel rate less than 396 macroblock/frame (@ 25 frames per second) and 330 macroblock/frame (@ 30 frames per second)
- MPEG-1 achieves bit rates between 1 and 1.5 Mbit/s with VHS quality for moderate to low motion video sequences, and worse than VHS for high-motion sequences.

MPEG-2

- The idea behind MPEG-2 was to provide a generic application independent standard.
- MPEG-2 has different options and for a particular application user can select from a set of profiles and levels.
- Profile: algorithm to be used
- Level: constraints on the parameters
- There are five profiles: simple, main, snr-scalable, spatially scalable, and high.
- Each higher profile is able to decode video encoded using all profiles up to and including that profile.

MPEG-2

- Simple profile does not use B frames.
- Main profile is very similar to MPEG-1
- snr-scalable, spatially scalable, and high profile use more than one bitstream to encode the video.
- The base bitstream is a lower-rate encoding of the video sequence that can provide a reconstruction of the sequence.
- The other bitstream is used to enhance the quality of reconstruction.
- This layered approach is useful when transmission video over a network where some connections only permit lower rates.

MPEG-2

- Layered approach allows us to increase the quality if bandwidth is available (quality is scalable).
- If the enhancement layer is used to reduce the error between original and reconstructed video, it is called snr-scalability.
- If the enhancement layer contained a coded bitstream corresponding to frames that would occur between frames of the base layer, the system is called temporally scalable.
- If the enhancement allowed an upsampling of the base layer, the system is spatially scalable.

MPEG-2

- Levels: low, main, high 1440, and high.
- Low: 352x240, main: 720x480, high 1440: 1440x1152, high: 1920x1080 (all at 30 frames/s)
- Compared to MPEG-1, MPEG-2 has several additional motion compensation prediction modes: field prediction and dual prime prediction
- MPEG-2 allows interlace video.
- In interlace video, each frame is divided into two fields.
- One field contains the odd lines of the frame and the other one the even ones.

MPEG-4

- MPEG-4 views a multimedia scene as a collection of objects
- Objects can be visual (e.g., background, talking head) or aural (e.g., speech, music)
- Each object can be coded independently using different techniques to generate separate bitstreams.
- Bitstreams are multiplexed along with a scene description
- A language called Binary Format for Scenes (BIFS) based on Virtual Reality Modeling Language (VRML) has been developed for scene description

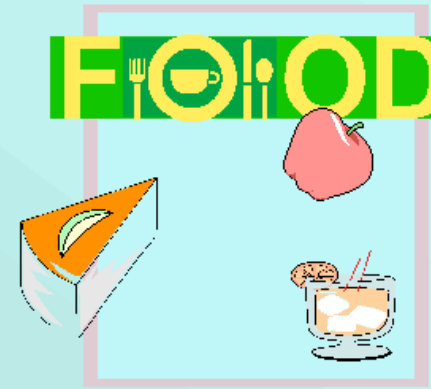
Object-based Coding

Object Manipulation

- Original Decoded



- Decoded and Manipulated



MPEG-4

- MPEG-4 achieves object-based representation by defining visual objects and coding them into separate bit streams.



- A binary alpha plane indicates the shape of the object.
- MPEG-4 video object coding:
 - Shape coding (for arbitrarily shaped video objects),
 - Motion compensated prediction,
 - DCT-based texture coding.

Pixel-based & Object-based Coding

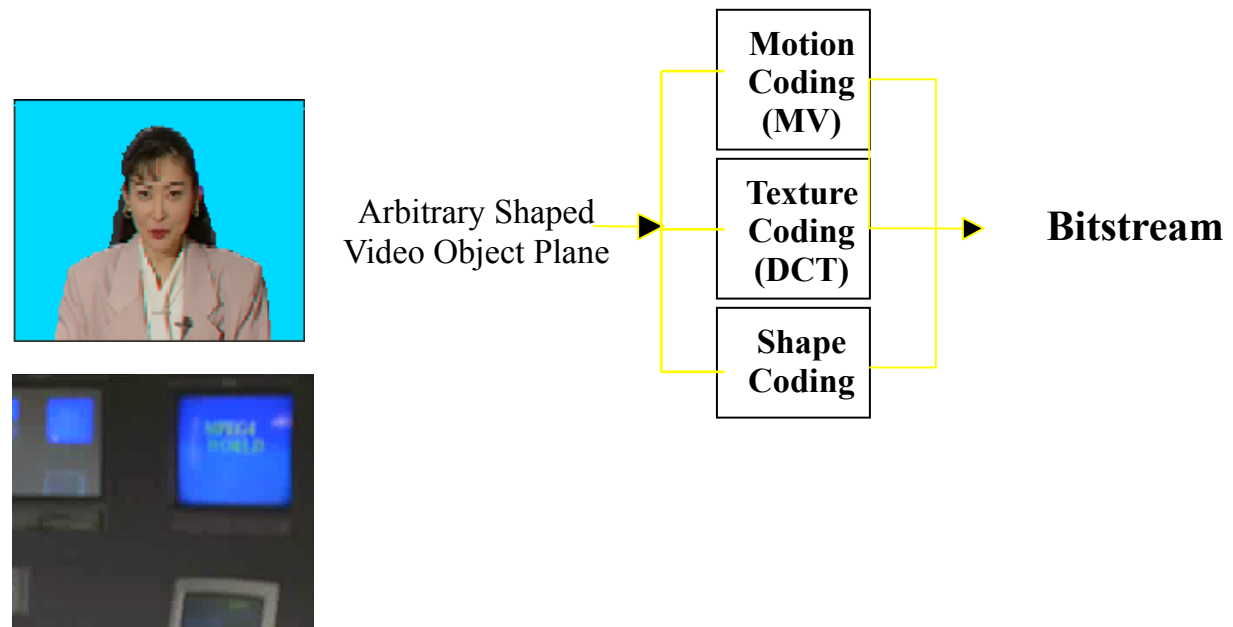
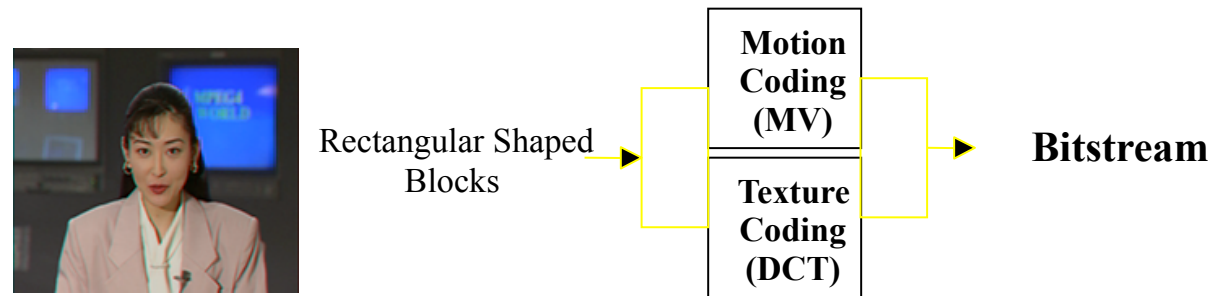
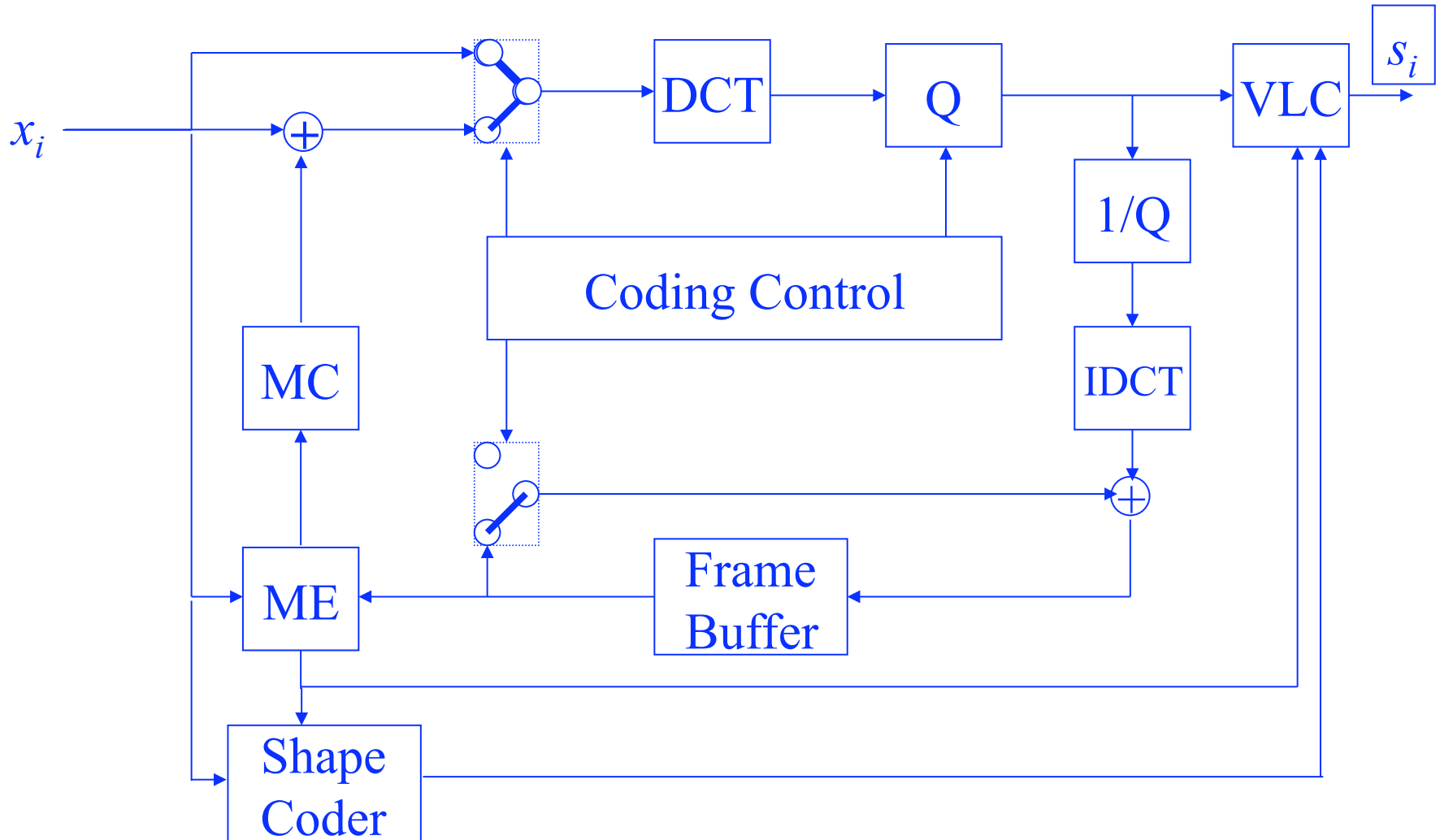


Image and Video Coding



MPEG-4

- Motion compensated predictor can use object-based motion compensation
- Video coding in MPEG-4 can also use a background “sprite” (a large still image that forms the background)
- Sprite is transmitted once and moving foreground objects are placed in front of the sprite based on the information provided by the encoder
- Model-based coding: a triangular mesh representing the moving object is transmitted followed by texture information for covering the mesh
- Information about movement of the mesh nodes can be transmitted to animate video object

MPEG-4 & MPEG-7

- Texture is coded using EZW
- Facial animation object: used to render an animated face
- Shape, texture, and expression of the face are controlled using facial definition parameters (FDPs) and facial action parameters (FAPs)

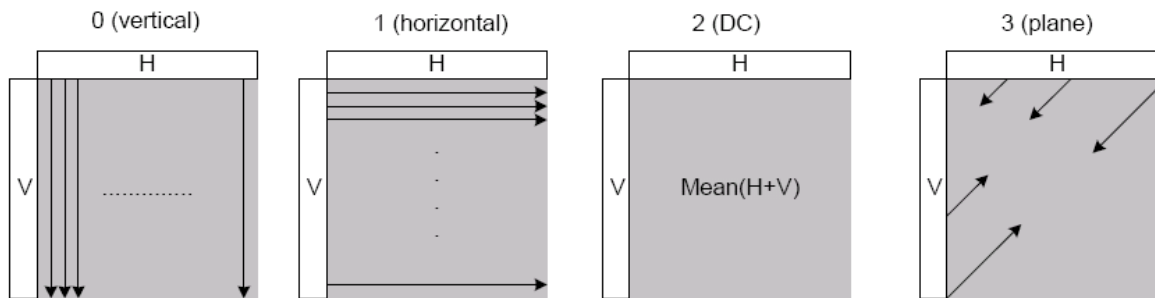
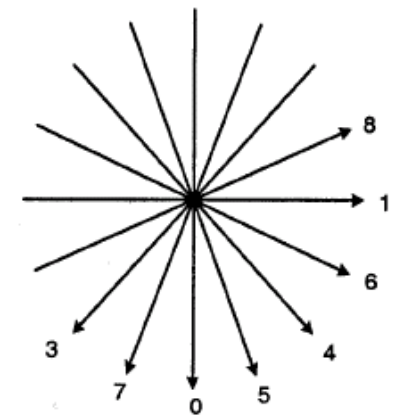
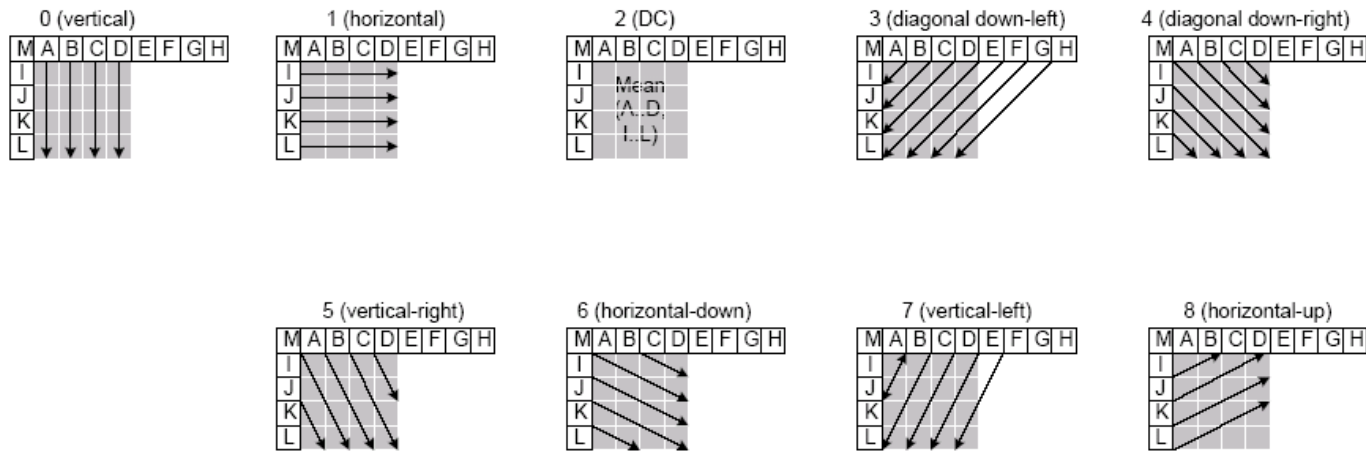
- MPEG-7: deals with development of a multimedia content description interface

H.264

- The basics are very similar to previous schemes
- H.264 allows 8x8 macroblocks to be further divided into sub-macroblocks of size 8x4, 4x8, and 4x4
- Smaller blocks allow tracking of finer details.
- Motion compensation is accomplished using quarter-pixel accuracy
- Standard allows for searching up to 32 pictures to find the best matching block.
- The transform is a 4x4 integer DCT.
- H.264 allows Intra prediction: a number of spatial prediction modes are available

H264

- Intra4x4



H.264

- H.264 uses a uniform scalar quantization for quantizing the coefficients.
- H.264 contains two options for binary coding
- The first uses exponential Golomb codes to encode the parameters and a context-adaptive variable length code (CAVLC) to encode the quantizer labels.
- The second binarizes all the values and then uses a context-adaptive binary arithmetic code (CABAC)

Packet video

- In dedicated communication, a channel was dedicated only to transferring information between two points
- Even if there is no information transfer going on during a particular period, the channel could not be used by anyone else.
- In asynchronous transfer mode (ATM) network, the users divide their information into packets, which are transmitted over channels that can be used by more than one user
- Availability of transmission capacity is affected by factors that are outside our control.
- Little traffic on the network: available capacity high
- Congestion on the network: available capacity low

Packet video

- To prevent congestion from impeding the flow of vital traffic, networks prioritize the traffic
- High priority traffic is permitted to move ahead of lower-priority traffic
- Guaranteed traffic is expensive
- Compressed-video transmission over ATM: compression scheme should have the ability to cope with network congestions
- Layered video coding is a solution: a low-rate high-priority layer is used to reconstruct the video, low-priority enhancement layers enhance the quality of reconstruction