Video Compression

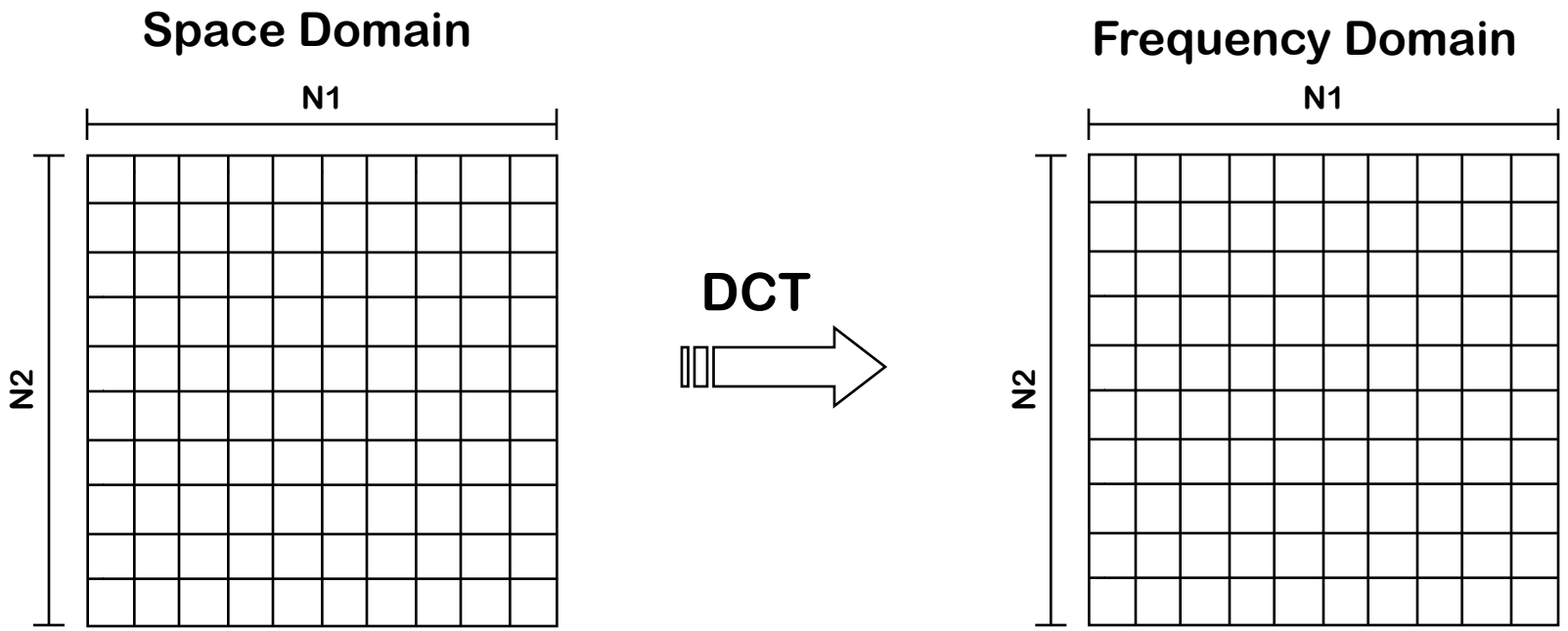
Video Compression Techniques

- Exploiting spatial correlation within a frame
  - Representing the picture in the Frequency domain using Transforms: e.g., Discrete Cosine Transform (DCT)
- Exploiting temporal redundancy in successive frames
  - Differential encoding
  - Motion Estimation and Compensation
- Quantization of Transform Coefficients
- Entropy Coding: run-length coding and Huffman coding of Transform Coefficients
- Intracoding and Intercoding of video frames

Discrete Cosine Transform (1)

- Maps values from time domain to frequency domain
- Given an image of size N1xN2
- $s_{y,x}$ - Value of pixel at position (x,y)
- $S_{v,u}$ - DCT coefficients
  - $S_{0,0}$ - DC Coefficient - Fundamental Color
  - $S_{N2-1,0}$ - Highest vertical frequency
  - $S_{0,N1-1}$ - Highest horizontal frequency
  - $S_{N2-1,N1-1}$ - Highest frequency that appears equally in both dimensions

Discrete Cosine Transform (2)



Each position represents an image pixel

Each position represents a frequency component

N1 x N2 samples in the space domain are converted into N1 x N2 samples in the frequency domain

Discrete Cosine Transform (3)

Forward DCT (compression)

$$S_{vu} = \frac{1}{\sqrt{N1N2}} c_u v_u \sum_{x=0}^{N1-1} \sum_{y=0}^{N2-1} s_{yx} \cos\frac{(2x+1)u\pi}{2N1} \cos\frac{(2y+1)v\pi}{2N2}$$

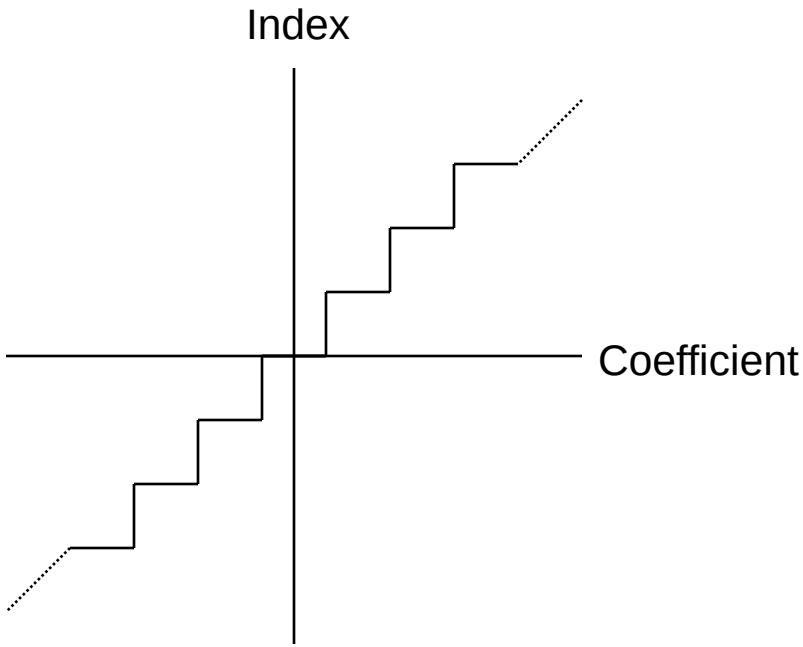
$$c_u, v_u = \frac{1}{\sqrt{2}}, u = 0, v = 0$$

$$c_u, v_u = 1, else$$

Inverse DCT (decompression)

$$s_{x,y} = \frac{1}{\sqrt{N1N2}} \sum_{u=0}^{N1-1} \sum_{v=0}^{N2-1} c_u c_v S_{vu} \cos\frac{(2x+1)u\pi}{2N1} \cos\frac{(2y+1)v\pi}{2N2}$$

Quantization



i [u,v] = 8 ∗ c[u,v] // (q ∗ m[u,v])

i [u,v] : quantized coefficient

c[u,v] : original coefficient

m[u,v] : quantization matrix

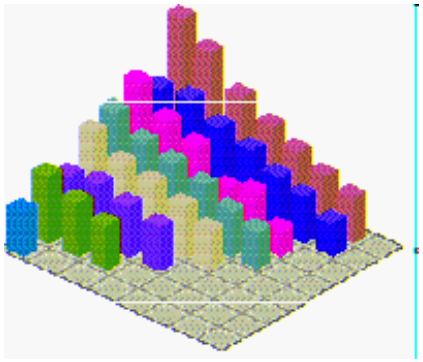
q: quantizer scale

// : division and rounding to the nearest integer

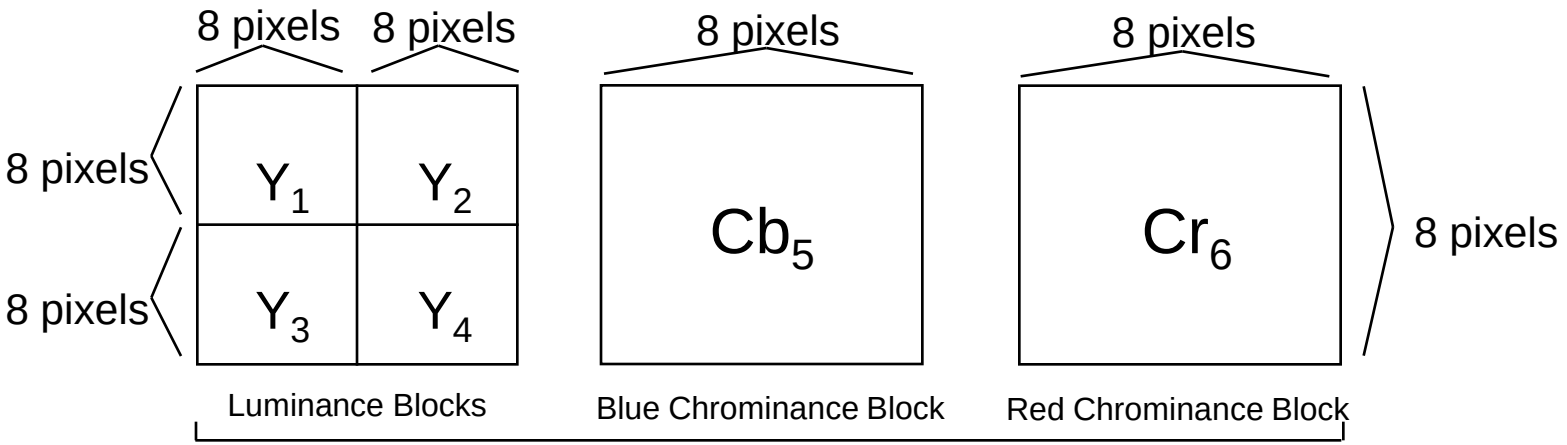
After the DCT, the resulting coefficients are quantized (i.e., truncated from floating point to an integer representation).

Why does it compress?

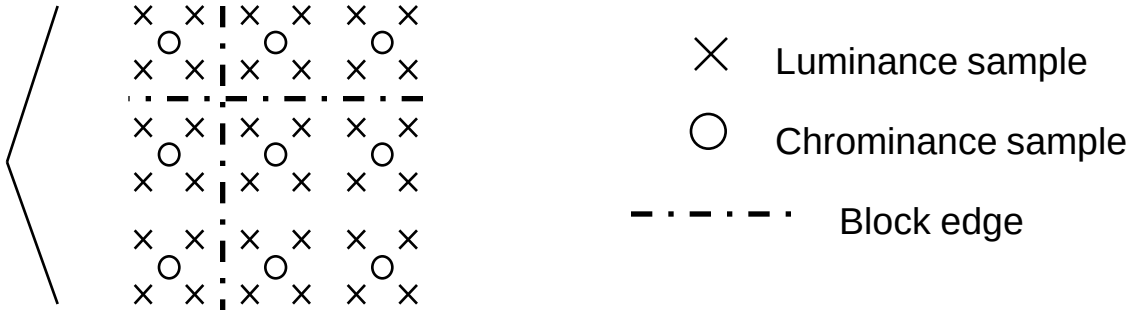
- Most images have the majority of their information in the low frequency components.
- Moreover, the high frequency components are less perceptible to the human eye - can be quantized coarser.
- The quantization is the lossy part of the process.



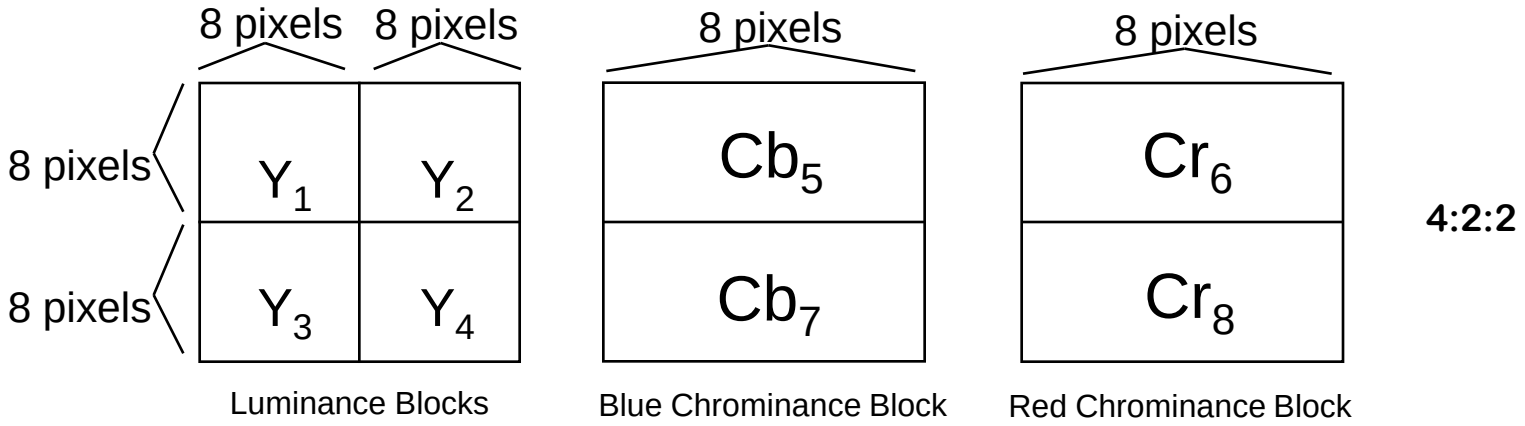
Blocks and Macroblocks
(4:2:0 format)

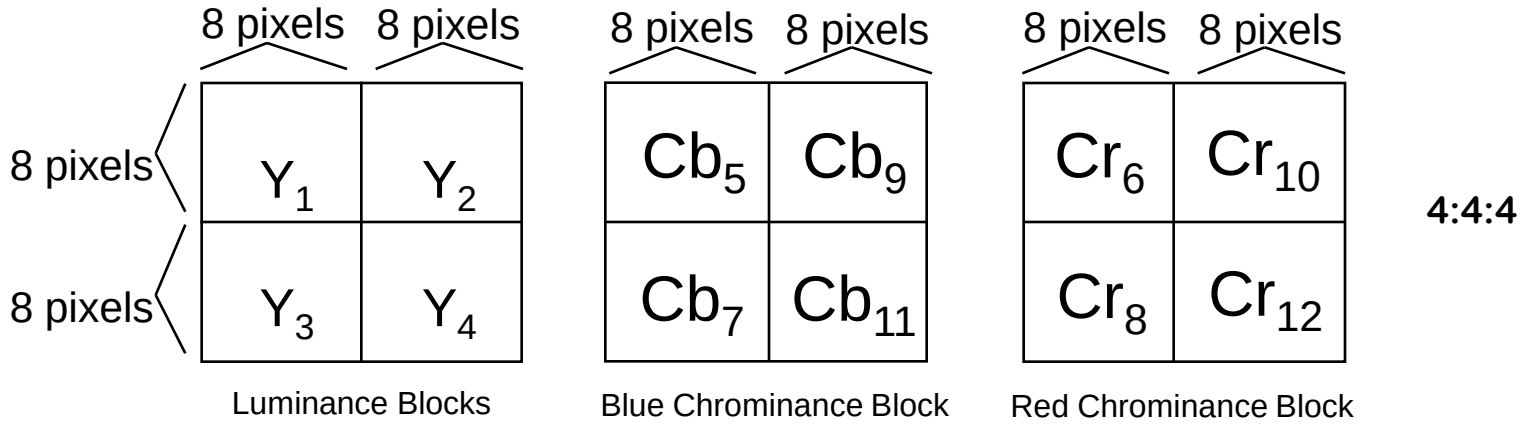


Positioning of Luminance and Chrominance Samples

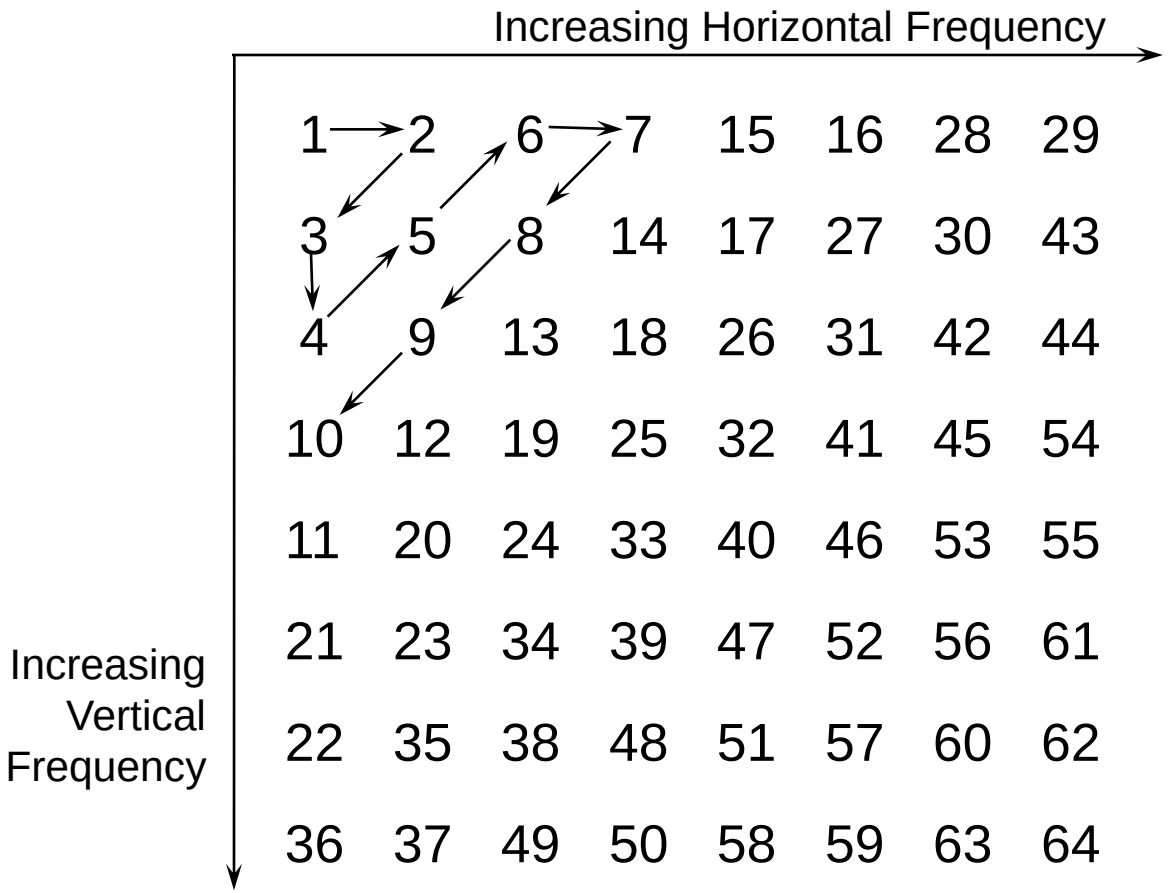


Blocks and Macroblocks
(4:2:2 and 4:4:4 formats)





Ordering of the DCT Coefficients in Coded Bit Stream



Entropy-Coding

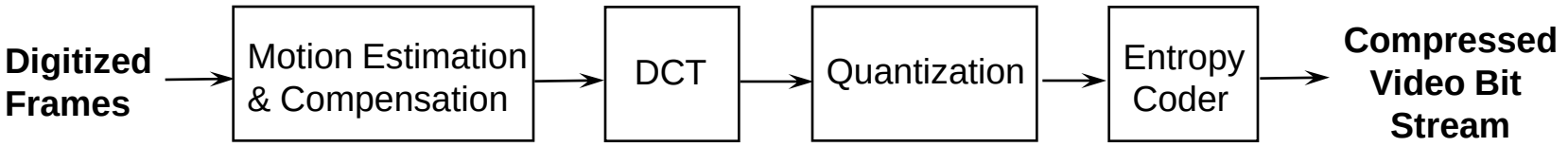
- The quantized coefficients are further compressed (in a lossless way):
  - Huffman (RLE) coding
  - Fixed tables (non-optimal but robust)
- Reading the coefficients out in a zig-zag pattern causes the lower-value coefficients to be grouped together, improving the performance of this step.

Exploiting Temporal Correlation

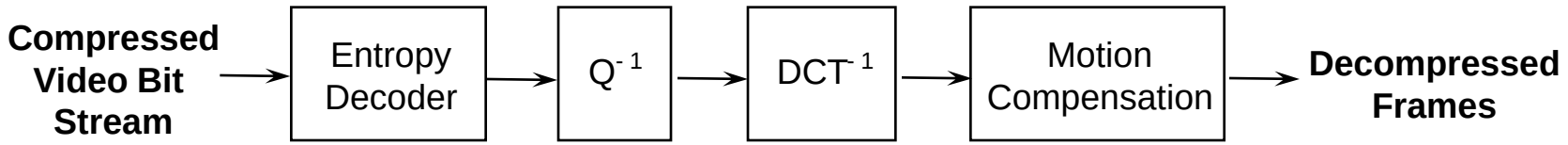
- Successive images are typically similar.
- Results can be improved by coding the ***difference*** between this image and the previous (and/or next) image.
- Motion compensation:
  - A block may move from one frame to another (example: camera pan)
  - Simple difference won't catch that.
  - Encoder needs to search the "best match" for a block prior to coding the difference.

Encoder and Decoder Block Diagrams

ENCODER....



DECODER....



Video Compression Schemes

- H.261
- MPEG-1
- MPEG-2
- H.263
- Motion JPEG

H.261

Objective:

- Video coding and decoding methods for audiovisual services at the rates of p x 64 kbit/s ($1 \leq p \leq 30$), to be carried over digital circuit-switched links
- Primary target applications are videophone and videoconference, hence low delay

H.261: Group of Blocks (GOB) and Frame Structure

Arrangement of macroblocks in a GOB

1	2	3	4	5	6	7	8	9	10	11
12	13	14	15	16	17	18	19	20	21	22
23	24	25	26	27	28	29	30	31	32	33

GOB within CIF

288 pixels

1	2
3	4
5	6
7	8
9	10
11	12

352 pixels

GOB within SIF (NTSC)

1	2
3	4
5	6
7	8
9	10

240 pixels

352 pixels

GOB within QCIF

1
3
5

144 pixels

176 pixels

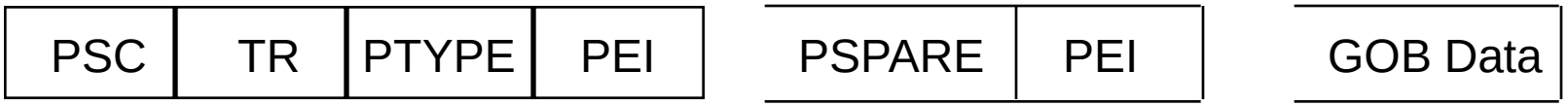
CIF=Common Intermediate Format
QCIF=Quarter CIF
SIF=Standard Interchange Format

H.261 Syntax

4 Layers:

- Picture layer
- GOB layer
- Macroblock layer
- Block layer

H.261 Picture Layer Syntax



PSC: Picture Start Code

TR: Temporal Reference: time elapsed between the current frame and the previously encoded one (in increments of 33 ms)

PTYPE: Type information
- Frame format (CIF or QCIF)
- Split screen indicator
- Document camera indicator
- Freeze Picture Release

PSPARE: Extra information placed by the encoder

H.261 GOB Layer Syntax

GBSC: GOB Start Code

GN: Group Number (position of GOB in the frame)

GQUANT: Quantizer scale (can be overridden at the macroblock layer)

GSPARE: Optional data field

H.261 Macroblock Layer Syntax

MBA: Macroblock Address
(Position of macroblock within GOB)

MTYPE: Macroblock Type
- Intracoded or Intercoded
- Motion Compensation used or not
- Loop filter used or not

MQUANT: Quantizer scale

MVD: Motion Vector Data

CBP: Coded Block Pattern
(Indicates which blocks are actually coded)

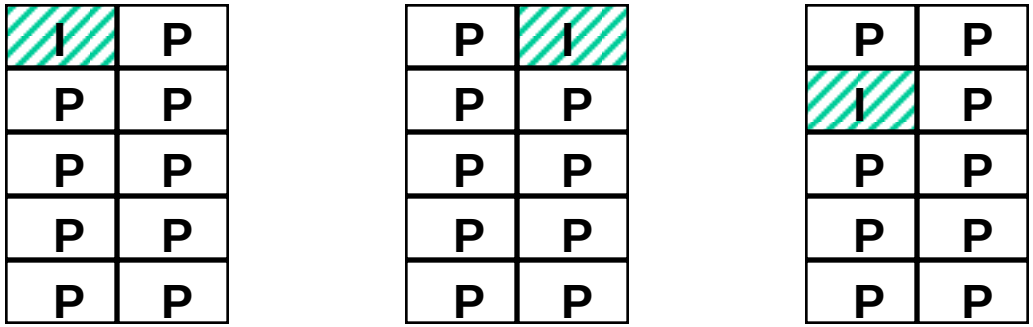
H.261 Block Layer Syntax

TCOEFF: Transform Coefficient Data

EOB: End of Block

H.261: Intra/Inter Coding of GOBs (1)

Approach 1: Rotational intracoding of GOBs in successive frames for reducing burstiness and hence latency

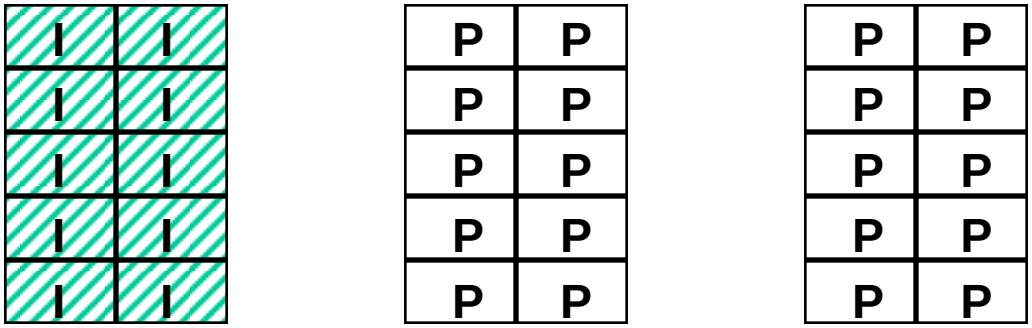


I: Intracoded

P: Predictive coded (Intercoded)

H.261: Intra/Inter Coding of GOBs (2)

Approach 2:



I: Intracoded

P: Predictive coded (Intercoded)

H.261 Restrictions

Frame Formats: QCIF, CIF, SIF

Frame rates: 30/n fps

- 30, 15, 10, 7.5, 6, 5, 4.28, etc.

Motion Vector:

- accuracy of 1 pixel
- range limited to ±16 pixels

Quantization Matrix:

- specified to be 16 for all coefficients

GOB Structure: fixed, 11 x 3 macroblocks

MPEG-1

- Intended for storage media
  - continuous transfer rate of about 1.5 Mbit/s (CD-ROM rate)
- VCR-level quality
- Coding algorithm generic enough to be used for different purposes:
  - at other bit rates
  - live encoding/transmission of video

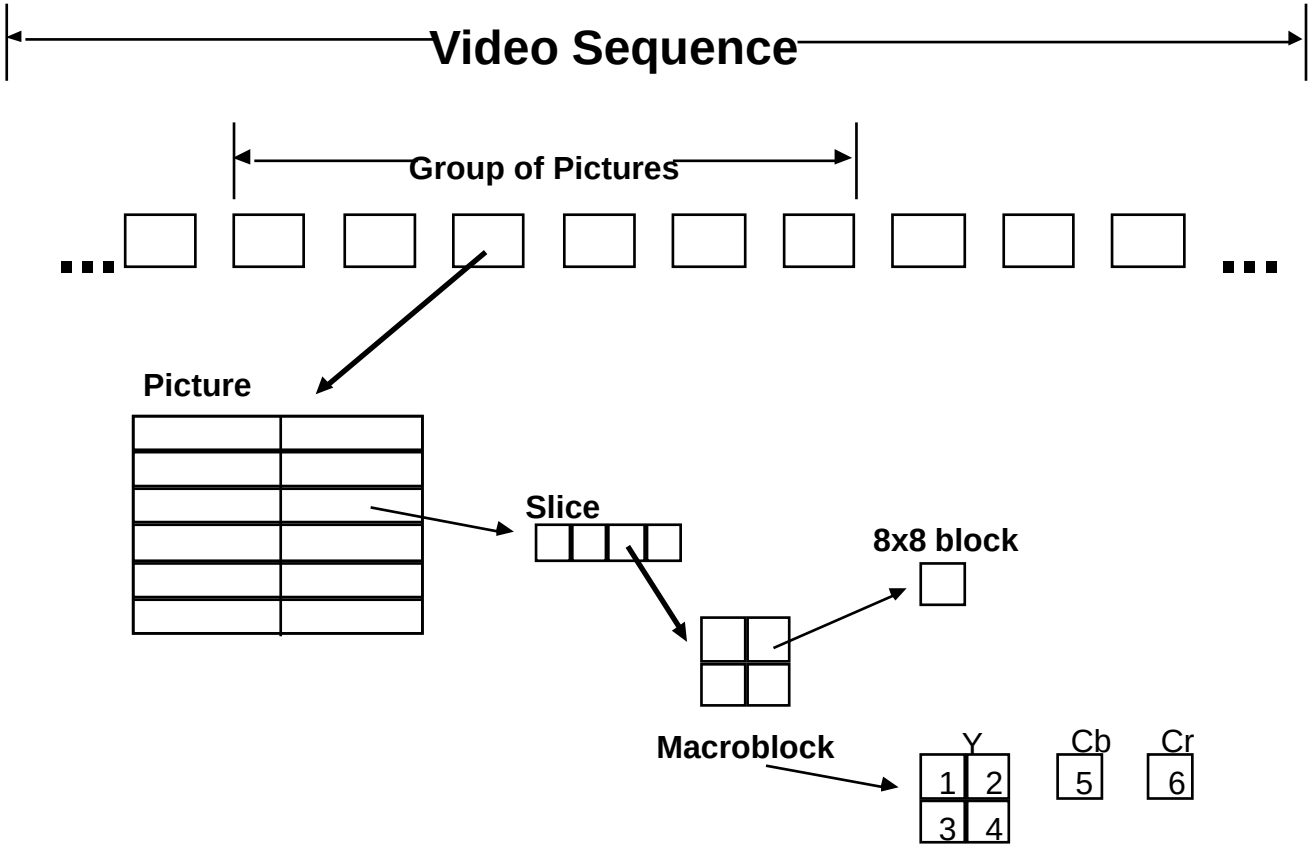
MPEG-1 Features (1)

- Frame Types:
  - I, P, and B
    (bidirectionally-predictive coded)
- Frame Format:
  - any number of pixels in each dimension
- Frame Rate:
  - any rate
- Motion Vector Data:
  - accuracy of 1/2 pixel
  - wider range, up to 128 pixels

MPEG-1 Features (2)

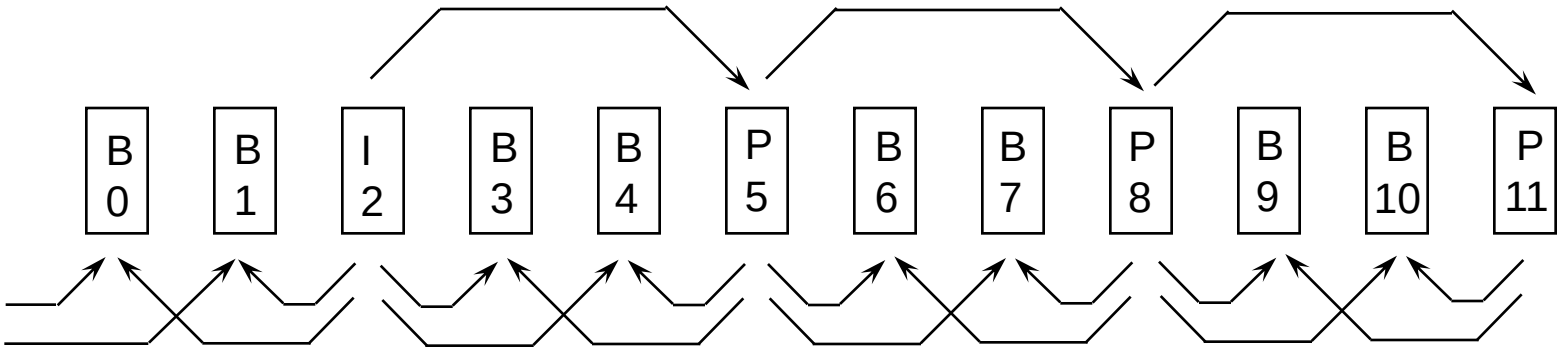
- Quantization Matrix:
  - two matrices, one for intracoded frames, one for intercoded frames
  - elements of a matrix may be different
  - specified for entire sequence
- Slice Structure:
  - flexible
  - may change dynamically during a sequence

MPEG Data Hierarchy



MPEG: Group of Pictures

Dependency relationship between I, B, and P-pictures

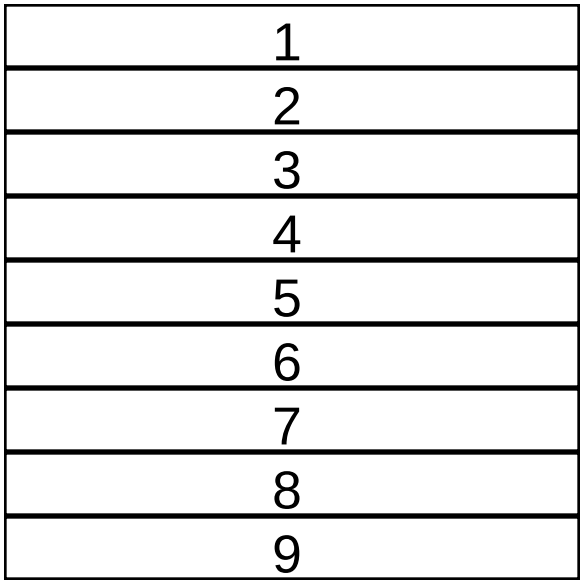


I-Frame: Intra-coded
P-Frame: Predictive
B-Frame: Bidirectional

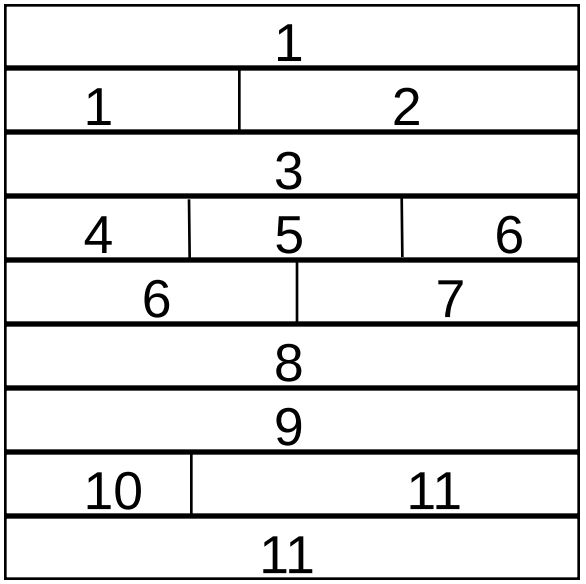
MPEG: Slices

Slice: any integral number of consecutive macroblocks in a frame

Example 1:



Example 2:



Coding of P-Frames

- P-Frames are coded based on the previous I- or P-frame.
- Processing:
  - Look for a block which is a "close" match
  - Take the DCT of the difference
  - If no good close match is found, intra-code
  - Use Huffman coding on the resulting data stream

Coding of B-Frames

- B-Frames are coded based both on the *previous* I or P frame and the *next* I or P frame.
- Try the following and pick what works best:
  - Use as reference a block from the previous frame
  - Use as reference a block from the next frame
  - Average the previous/next blocks and subtract
  - Intra-code
- Use the DCT step as before.

Typical MPEG-1 Statistics

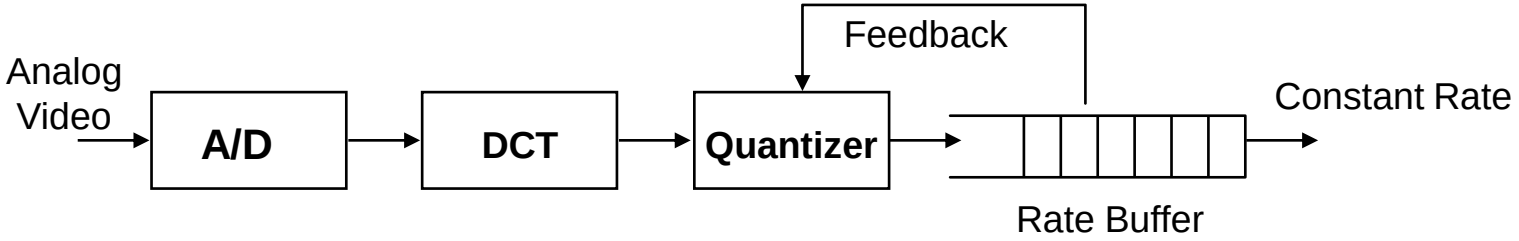
- Typical GOP is 15 frames:

  IBBPBBPBBPBBPBB

- Table shows the sizes for each type of frame at 1.15 Mb/s (video-CD rate)
- Table also shows glitch duration if the frame is lost

Frame Type	Average Size (bytes)	Stream fraction (% total data)	Glitch Duration (seconds)
I-Frames	13,660	19.00	0.500
P-Frames	6,079	33.63	0.250
B-Frames	3,420	47.37	0.033

MPEG Rate Control



- To achieve constant rate, the quantizer step is varied
- Quality is uneven: more complex scenes have lower quality
- It is also possible to code using variable bit rate to achieve “constant quality”

MPEG-1 Syntax

6 Layers:

- Sequence layer
- GOP layer
- Picture layer
- Slice layer
- Macroblock layer
- Block layer

MPEG Audio

- MPEG-1 defines three layers of audio coding.
- Basic idea: sub-band coding combined with psycho-acoustic models
  - Divide the signal into 32 bands
  - Assign bits dynamically based on energy
- Sampling rates: 32 kHz, 44.1 kHz, 48 kHz
- Data rates: from 32 kb/s to 384 kb/s; around 128 kb/s audio is transparent
- Modes of operation: Mono, Stereo, Joint-Stereo

MPEG-2 (also H.262)

Objective:

- Generic method of video coding, covering a wide range of applications. In particular, applications requiring better quality than what 1.5 Mb/s MPEG-1 can offer.

Typical Applications:

- Digital Television Broadcast
- Interpersonal Communications (videoconferencing, videophone, etc.)
- Networked Database Services
- Interactive Storage Media (optical disks, etc.)

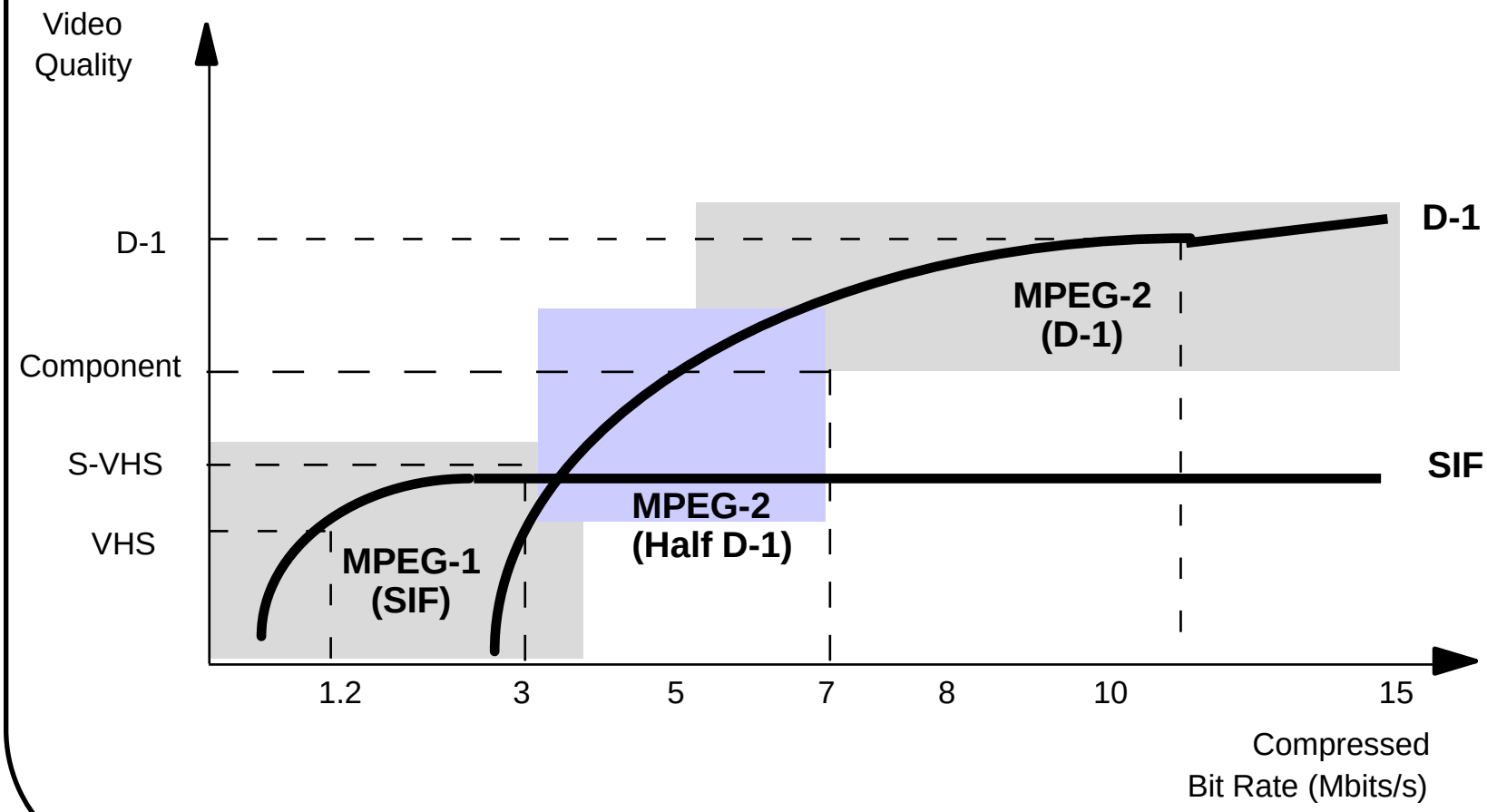
Why not MPEG-1 at higher bit rates?

- MPEG-1 does not support interlaced video
- Quality does not improve with increasing data rates

MPEG-2: Differences from MPEG-1

- Interlaced scan format supported
- Different chrominance sampling formats (4:2:0, 4:2:2, 4:4:4) can be represented
- Low delay mode available (intracoded slices)
- Scalability is supported
  - SNR Scalability
  - Multiple Resolution Scalability
  - Temporal Scalability
  - Combinations of the above

Comparison MPEG-1/MPEG-2



MPEG System Streams

- MPEG Audio and Video streams are multiplexed together in a "system" (combined) stream.
- Elementary Audio and Video streams are first converted into PES (Packetized Elementary Streams), which contain synchronization information.
- The Audio/Video PES are then combined into a single system stream.
- System streams may contain multiple audio and video streams.

MPEG System Stream Formats

- **MPEG-1 System Stream**
  - Defined in the original MPEG-1 standard.
  - Used in Video-CDs and MPEG-1 files.
- **MPEG-2 Program Stream**
  - Defined in the MPEG-2 standard.
  - Virtually identical to the MPEG-1 System Stream.
  - Used in DVDs.
- **MPEG-2 Transport Stream**
  - Defined in the MPEG-2 standard
  - Uses fixed-size (188-byte) “transport packets”
  - Used in satellite, digital cable, and HDTV

H.263

Objective:

- Video encoding at low bit rates (e.g., 10-20 kbit/s) for videophone/videoconferencing applications over analog telephone lines.

Features that enable efficient coding at low bit rates:

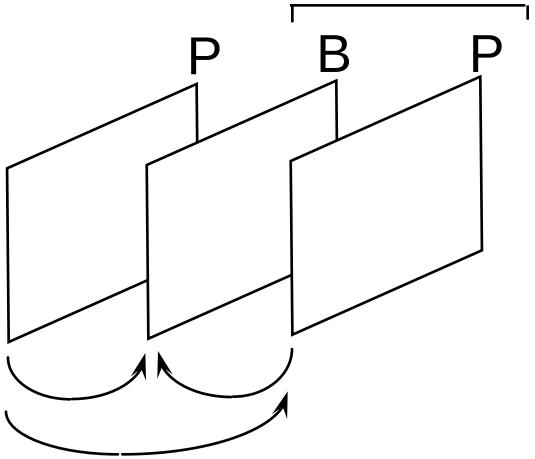
- PB frames mode
- Advanced prediction mode
  - Four motion vectors for each luminance block are used as opposed to a single motion vector
- Syntax-Based Arithmetic Coding Mode
  - More efficient compared to Huffman coding
- Motion vectors with half-pixel accuracy
  - as opposed to one-pixel for H.261
- Low header overhead by means of relatively restricted syntax

H.263: Other Differences with H.261

- In H.263, quantizer scale can only be incremented/decremented by 2 per macroblock
- GOB structure is different
- More frame formats can be used in H.263

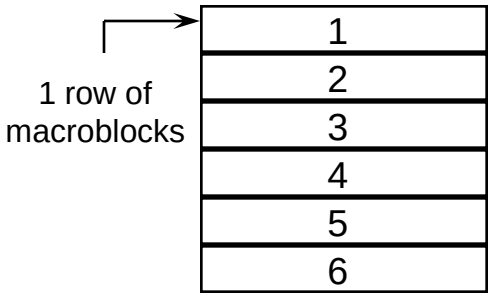
H.263 PB Frames

"PB Frame"

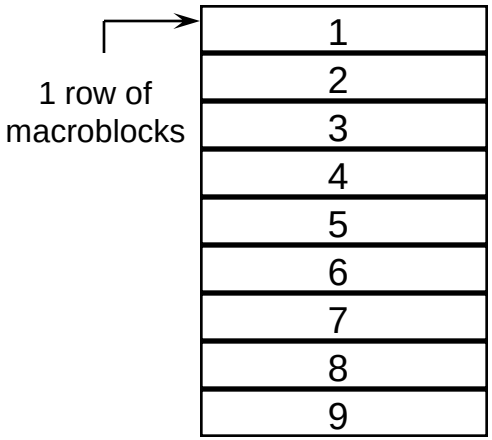


- Macroblocks for a P and a B frame are interleaved
- B frames:
  - Some macroblocks may be differentially encoded with respect to both previous and next frames; others are differentially encoded with respect to previous frame.
  - Motion vectors for a B macroblock are interpolated from the motion vectors for the corresponding P macroblock in the PB frame.

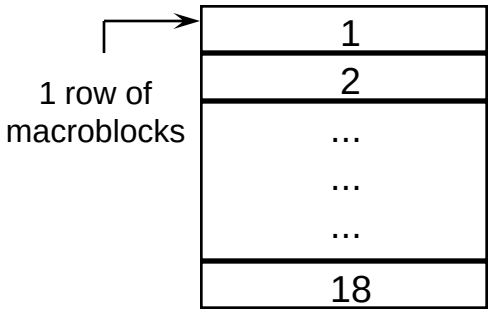
H.263
Frame Formats and GOB Structure



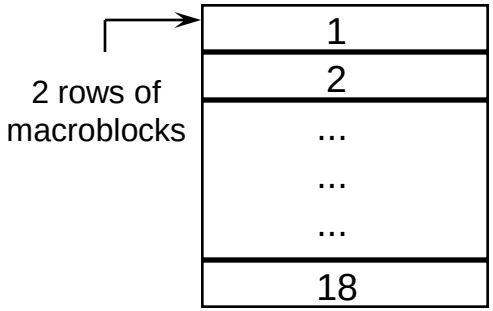
SUB-QCIF (128 x 96)



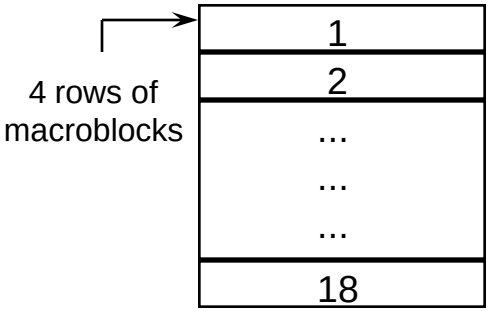
QCIF (176 x 144)



CIF (352 x 288)



4 CIF (704 x 576)



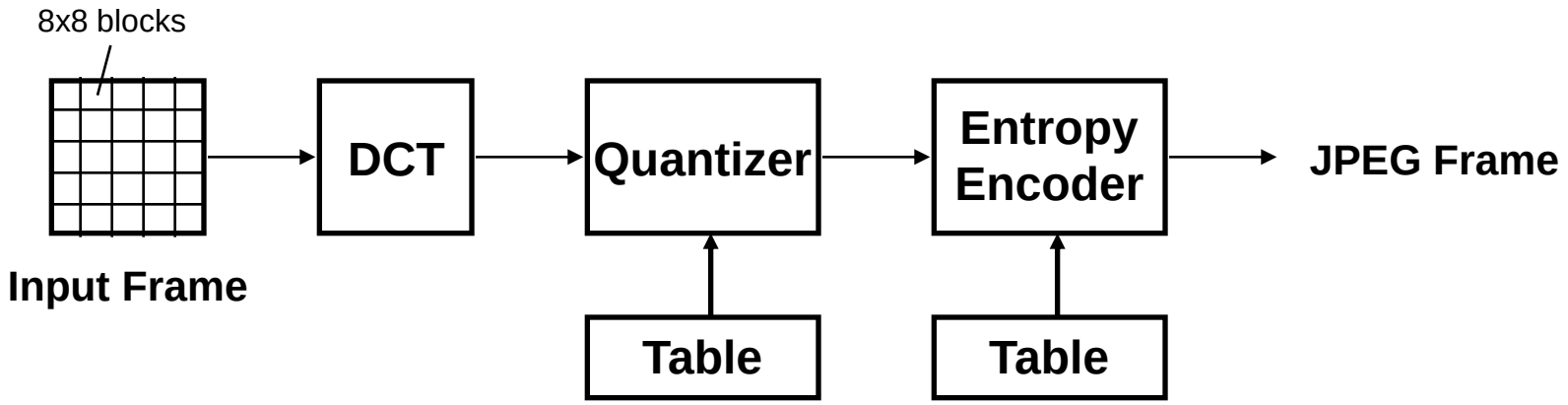
16 CIF (1408 x 1152)

H.263 Syntax

4 Layers (same as in H.261):

- Picture layer
- GOB layer
- Macroblock layer
- Block layer

JPEG



- JPEG is defined for still images
- Basically, JPEG uses DCT followed by entropy coding.

Motion - JPEG

- Motion-JPEG:
  - Each frame is JPEG encoded independently of each other (i.e., no interframe coding).
  - While JPEG includes a wide variety or encoding algorithms, only sequential DCT mode is used in current motion-JPEG encoders (i.e., very similar to MPEG-1 encoding with I-frames only).
  - Quantizer scale can only be specified on a frame-by-frame basis.
  - Video-related information (e.g., frame rate) must be communicated externally to the JPEG syntax.
  - No standard for audio.

Video Encoding Control Schemes

Video Compression

- Lossy algorithms ––> quality degradation
- Encoder parameters can be adjusted to trade-off quality and bit rate
- For given encoder parameter values, both quality and bit rate depend on scene content
- In order to achieve certain quality, bit rate, and/or delay objectives over time, encoder parameters must be dynamically adjusted.

==> *Encoder Control Schemes*

Video Quality Measures: Signal-to-Noise Ratio

$$SNR(n) = 10 \log_{10} \frac{\sum_{i=1}^{N_p} o_i^2(n)}{\sum_{i=1}^{N_p} [o_i(n) - d_i(n)]^2}$$

$o_i(n)$: Luminance value for the i'th pixel of the n'th original frame

$d_i(n)$: Luminance value for the i'th pixel of the n'th encoded/decoded frame

N_p: Number of pixels in a frame

ITS Video Quality measure
Ŝ

- Quantitative video quality measure, developed by the Institute for Telecommunication Sciences (ITS)
- Measures spatial and temporal information loss
- 94% correlation with subjective evaluations
- Validated using 10 second sequences, wide range of impairments including H.261 encoding
- Quality degradation measured on a scale of 1 to 5:
  - 5: Imperceptible
  - 4: Perceptible but not annoying
  - 3: Slightly annoying
  - 2: Annoying
  - 1: Very annoying

Video Quality Measure
$\hat{s}$

- $\hat{s} = 4.77 - 0.992\ m_1 - 0.272\ m_2 - 0.356\ m_3$
- m_1:

  spatial distortion measure, detects blurring and false edges
- m_2:

  temporal distortion measure, detects lost motion (e.g., dropped frames)
- m_3:

  temporal distortion measure, detects added (false) motion (e.g., jerkiness, block errors)

For more information, refer to: A. Webster et. al, "An objective video quality assessment system based on human perception," proceedings of the IS&T/SPIE 1993 Int. Sym. Elect. Imaging: Science & Tech., SPIE vol. 1913, pp. 15-26

Video Quality Measure
Ŝ

- Spatial Information (SI) for frame F_n:

$$SI(F_n) = STD_{space}\{Sobel[F_n]\}$$

- Temporal Information (TI) for frame F_n:

$$TI(F_n) = STD_{space}[F_n - F_{n-1}]$$

$$m_1 = RMS_{time}\left(5.81\left|\frac{SI[O_n] - SI[D_n]}{SI[O_n]}\right|\right)$$

$$m_2 = f_{time}\left[0.108\max\left\{\left(TI[O_n] - TI[D_n]\right),0\right\}\right]$$

$$f_{time}(x_t) = STD_{time}\left\{CONV\left(x_t,[-1,2,-1]\right)\right\}$$

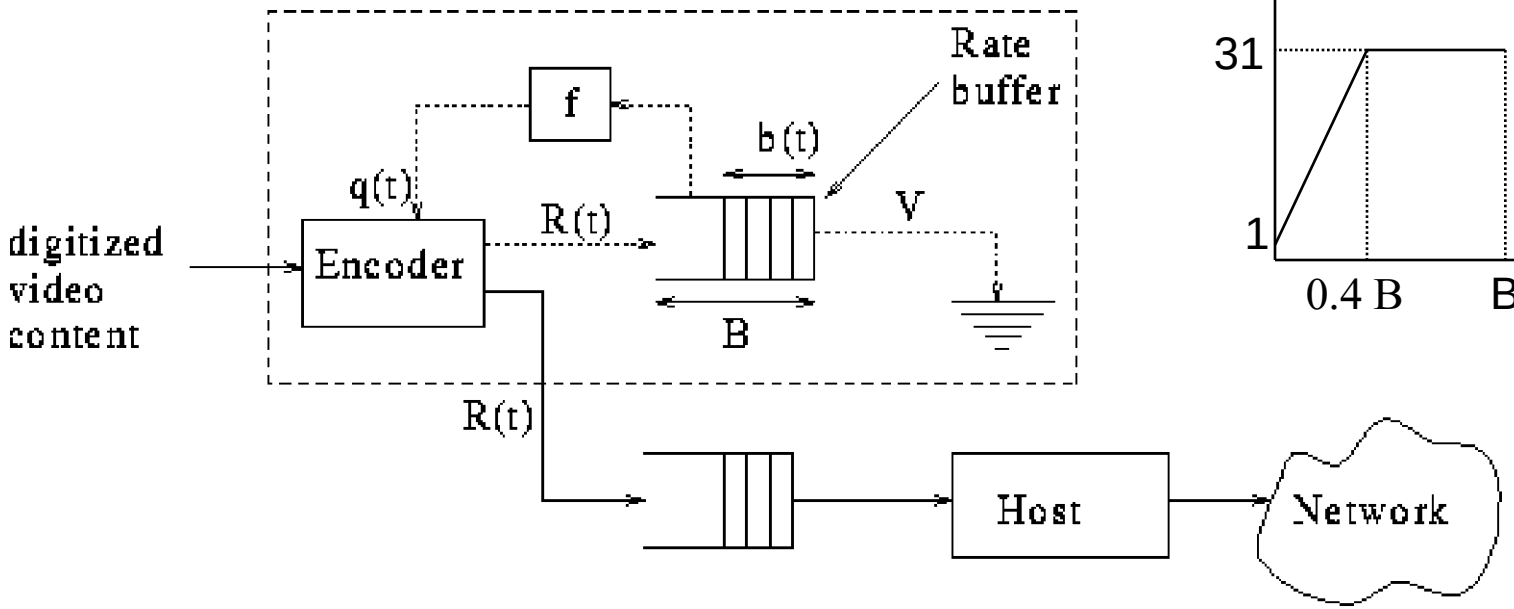
$$m_3 = \max_{time}\left\{4.23\log_{10}\left(\frac{TI[D_n]}{TI[O_n]}\right)\right\}$$

O_n: *n'th orig. frame*
D_n: *n'th degraded frame*

SNR versus ŝ

- SNR does not capture well the quality degradation due to video compression.
- Useful for comparing the same video content encoded in different ways.
- No relation between the absolute magnitude of the SNR and the perceived quality.
- Example: encoding a text against a flat background versus encoding a flower garden
  - Text may look awful (degradation around the letters) but will have high SNR
  - Garden may have low SNR while still looking acceptable

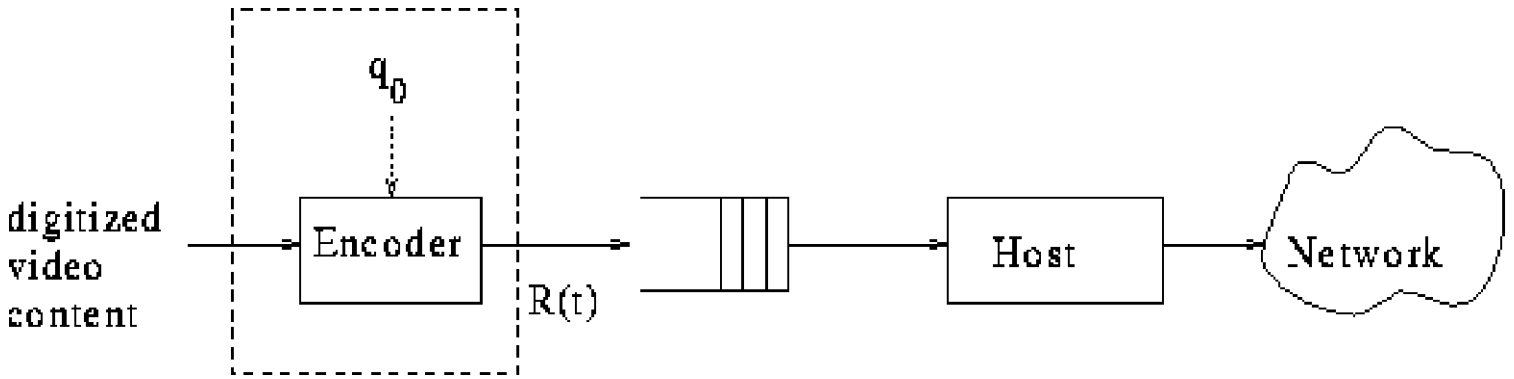
Constant Bit Rate Video Encoder Control



CBR feedback function

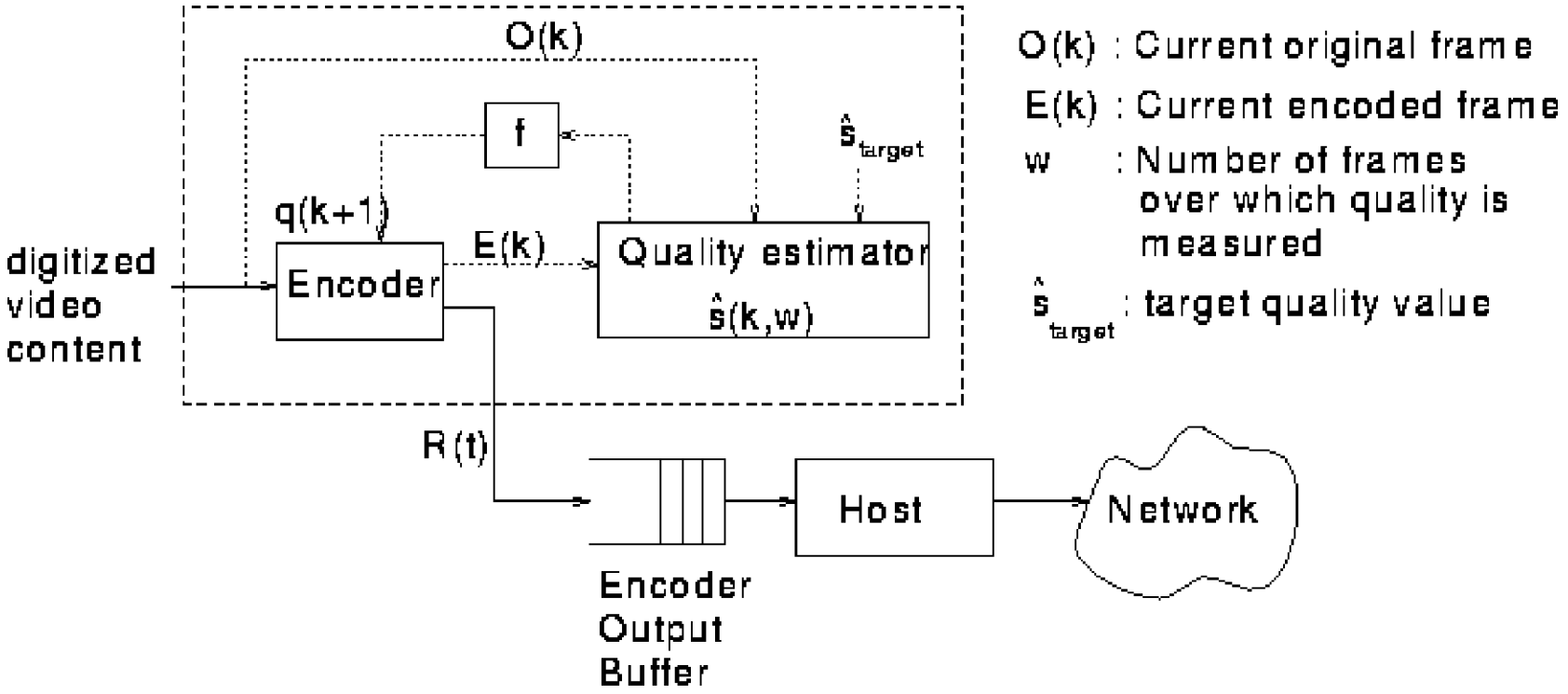
Design Parameters: V, B

Open-Loop Variable Bit Rate Video Encoder Control



Design Parameter: q_0

Constant Quality VBR Encoding



Design Problem: determine $f(\hat{s})$, w

Further Reading

- I. Dalgiç and F. Tobagi, "Characterization of Quality and Traffic for Various Video Encoding Schemes and Various Encoder Control Schemes," Technical Report CSL-TR-96-701, August 1996.
  `http://elib.stanford.edu/Dienst/UI/2.0/Describe/stanford.cs%2fCSL-TR-96-701`
- I. Dalgiç and F. Tobagi, "Performance Evaluation of 10Base-T and 100Base-T Ethernets Carrying Multimedia Traffic," IEEE JSAC, Vol. 14, No 7, Sept. 1996.
- I. Dalgiç and F. Tobagi, "Performance Evaluation of ATM Networks Carrying Constant and Variable Bit-Rate Video Traffic," IEEE JSAC, Vol. 15, No 6, August 1997.