Lecture 11
Basics of Digital Video – Part 1

Reference: Chapter 1: Video Formation, Perception, and Representation, in Video Processing and Communications, by Wang, Ostermann and Zhang

First "Motion Picture"

- In 1887, using a series of trip wires, Eadweard Muybridge created the **first high speed photo series** which can be run together to give the effect of motion pictures.
- High speed photography is the science of **taking pictures of very fast** phenomena. In 1948, the Society of Motion Picture and Television Engineers (SMPTE) defined high-speed photography as any set of photographs captured by a camera capable of **128 frames per second** or greater, and of at least three consecutive frames.
- Also see: http://www.youtube.com/watch?feature=player_embedded&v=F1i40rnpOsA

Basics of Video

Static scene capture ➔ Image

Bring in motion ➔ Video

Monochromatic image sequence: 3-D signal

- 2 spatial dimensions & 1 time dimension
- Continuous $I(x, y, t)$ ⇨ discrete $I(m, n, t_k)$

Color video → 4-D signal

Video Capture and Display

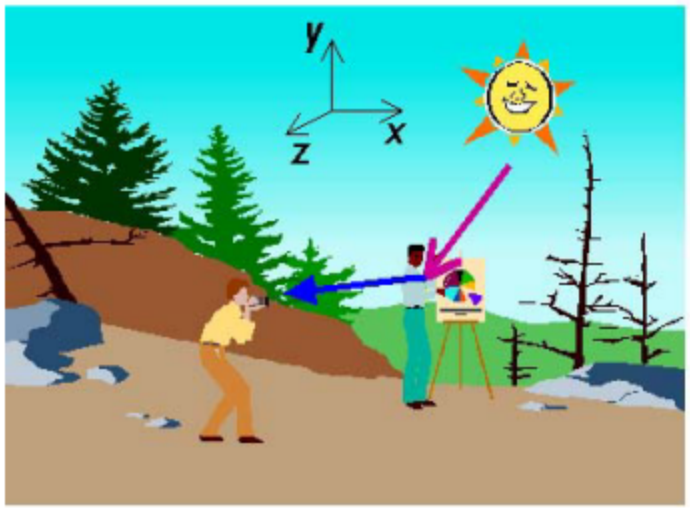
Involves the following components:

- Light reflection physics
- Imaging operator
- Color capture
- Color display
- Component vs. composite video

Video Capture

Scene and light source:

- For natural images we need a light source *(λ: wavelength of the source)*
  - $E(x, y, z, \lambda)$: incident light on a point *(x, y, z world coordinates of the point)*
- Each point in the scene has a reflectivity function.
  - $r(x, y, z, \lambda)$: reflectivity function
- Light reflects from a point and the reflected light is captured by an imaging device.
  - $c(x, y, z, \lambda) = E(x, y, z, \lambda) \times r(x, y, z, \lambda)$: reflected light. or emitted light intensity



$$E(x, y, z, \lambda)$$

$$c(x, y, z, \lambda) = E(x, y, z, \lambda) \cdot r(x, y, z, \lambda)$$

$$\text{Camera}(c(x, y, z, \lambda)) = X$$

Courtesy of Onur Guleryuz

Analog Video

- Video raster
- Progressive vs. interlaced raster
- Analog TV systems

Old and New Video Displays

- **CRT** (Cathode ray tube) display: vacuum tube containing one or multiple electron guns and a phosphorescent screen. Electron beam(s) are accelerated onto the screen to create images.

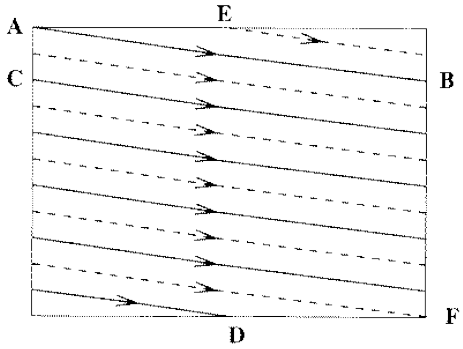
- **LCD** (Liquid-crystal display) is a flat display that uses the light modulating properties of liquid crystals.

- Need three light sources to input red, green, blue components.

- **QLED, OLED** (+foldable displays)

Raster Scan

- Real-world scene is a continuous monochrome 3-D signal (temporal, horizontal, vertical)
- Analog video is stored in the raster format
  - Sampling in time: consecutive sets of frames
    - To render motion properly, >=30 frame/s is needed
  - Sampling in vertical direction: a frame is represented by a set of scan lines
    - Number of lines depends on maximum vertical frequency and viewing distance, 525 lines in the NTSC system
  - Video-raster = 1-D signal consisting of scan lines from successive frames



NTSC: National Television System Committee

Raster scan produces either progressive or interlaced videos

Progressive

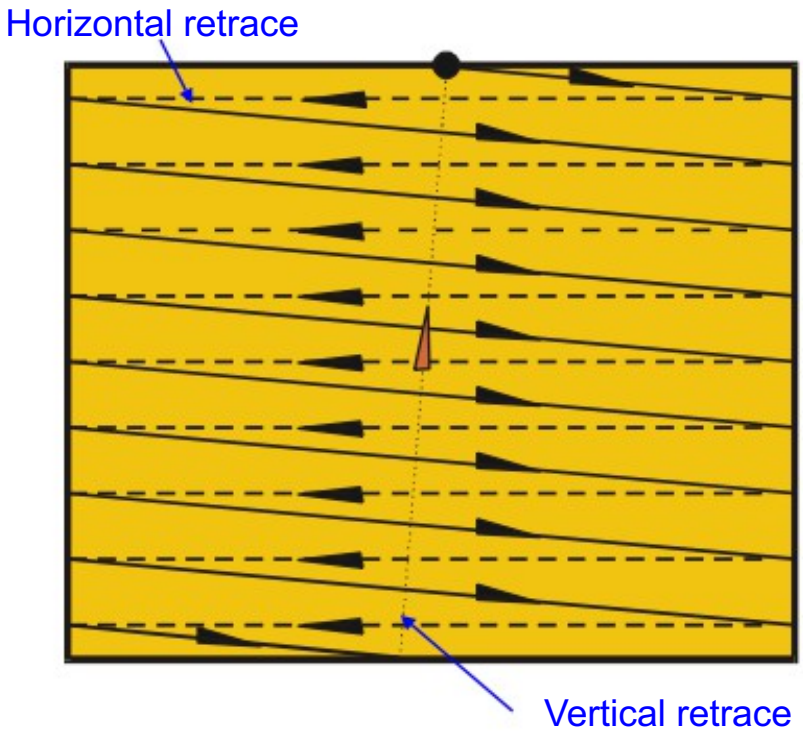
- Every pixel on the screen is either refreshed in a sequential order (monitors) or simultaneously (films)

Interlaced (old to obsolete)

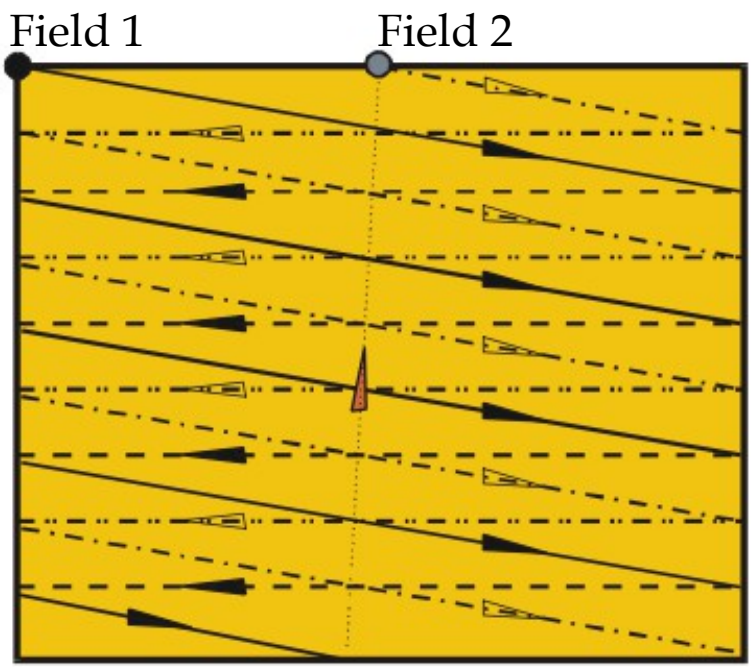
- Each frame is refreshed twice: the little gun at the back of the CRT shoots all the correct phosphors on the even numbered rows of pixels first and then odd numbered rows
- NTSC frame-rate of 29.97 means the screen is redrawn 59.94 times a second
- In other words, 59.94 half-frames per second or 59.94 fields per second

Progressive and Interlaced Scans

Progressive Frame

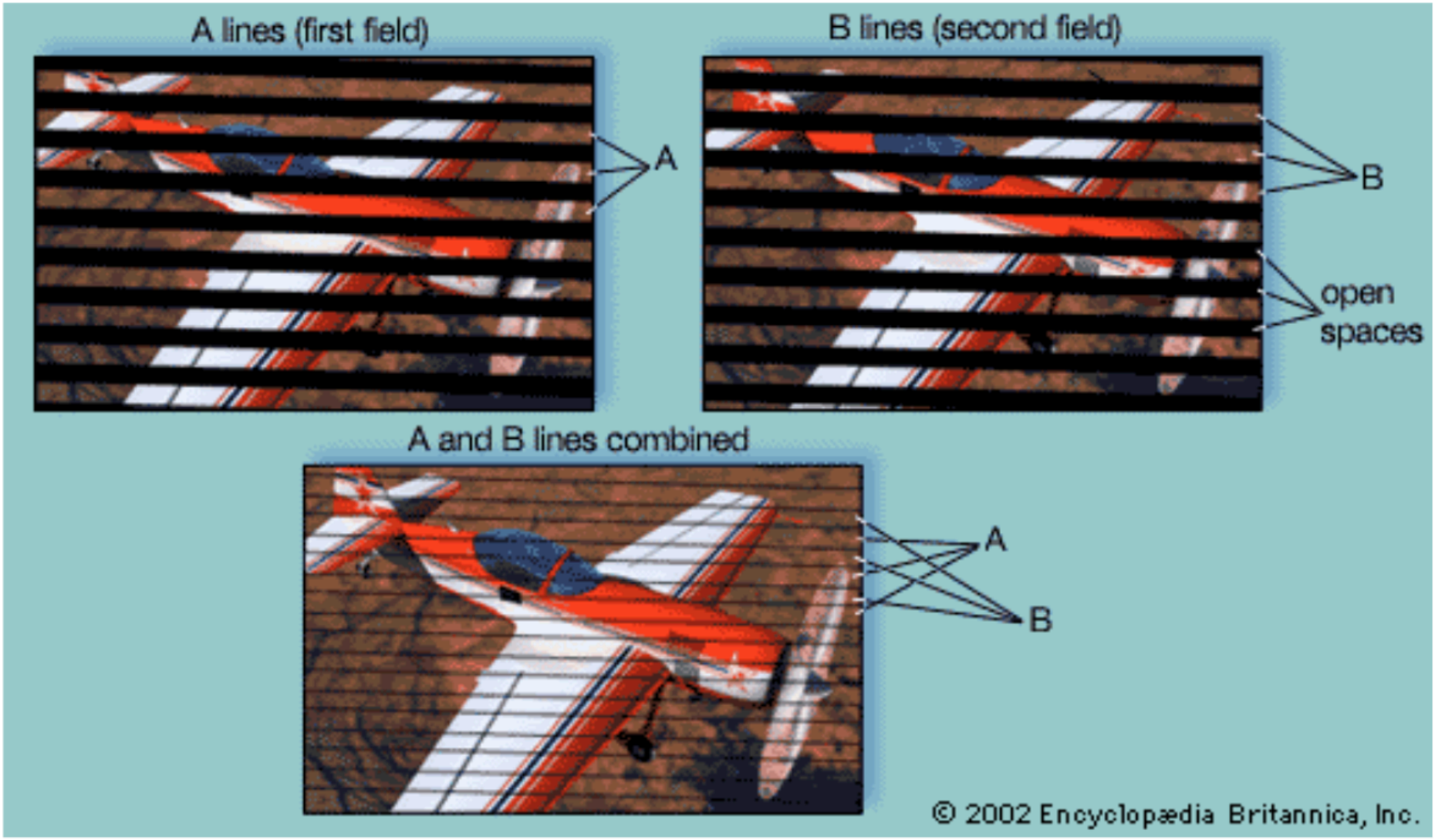


Interlaced Frame

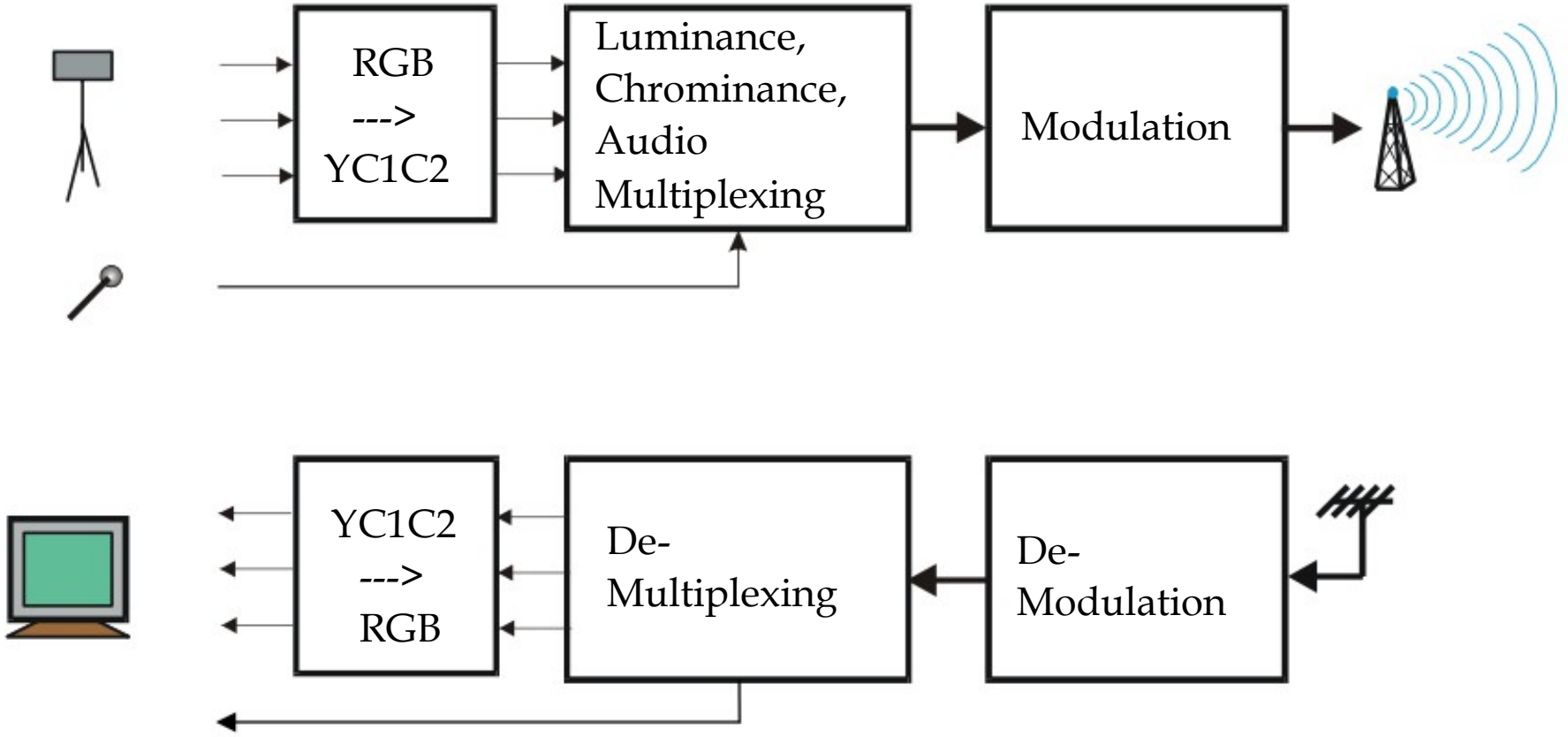


Interlaced scan is developed to provide a trade-off between temporal and vertical resolution, for a given, fixed data rate (number of line/sec).

Interlaced Videos



Color TV Broadcasting and Receiving



Why not using RGB directly?

- R,G,B components are correlated (How to verify this? Hint: where is information about brightness in RGB?)
  - Transmitting R,G,B components separately is redundant
  - More efficient use of bandwidth is desired
- RGB->YC1C2 transformation (recall YUV for analog)
  - Decorrelating: Y,C1,C2 are uncorrelated
  - C1 and C2 require lower bandwidth (WHY? See s.33)
  - Y (luminance) component can be received by B/W TV sets
- YIQ in NTSC
  - I: orange-to-cyan
  - Q: green-to-purple (human eye is less sensitive)
    - Q can be further band-limited than I
  - Hue = Arctan(Q/I) (Phase); Saturation =SQRT(I^2+Q^2) (Magnitude)
  - Hue is better retained compared to Saturation

Y: luminance
C1 (or Cb): blue difference
C2 (or Cr): red difference
YCbCr is a rotated and scaled version of RGB

Why not using RGB directly?

Example:

(a) (b)

Figure A.1: Sample colour image (a); and its grey level version (b).

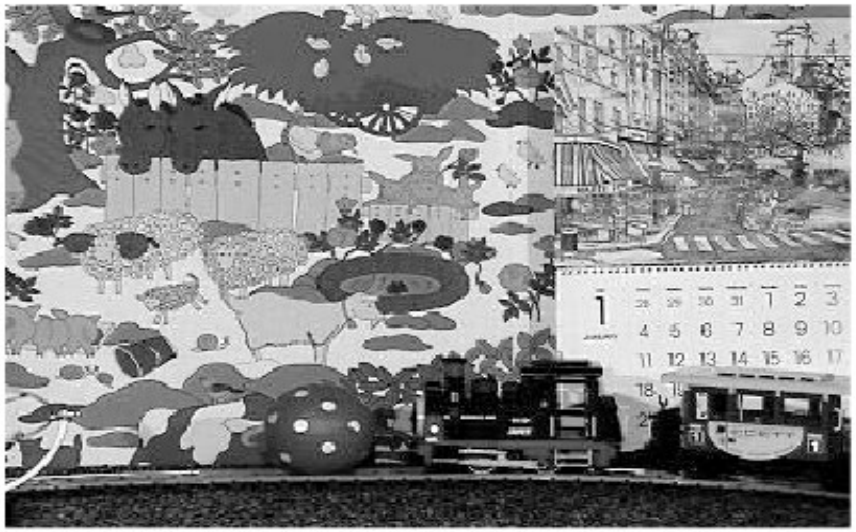
R G B

Figure A.2: RGB channels of image in Figure A.1(a) shown separately.

In RGB representation, the R, G and B channels are correlated, as all of them include a representation of brightness. This is illustrated in Figure A.1 and A.2, in which the brightness information can be recognized from R, G and B channels shown separately

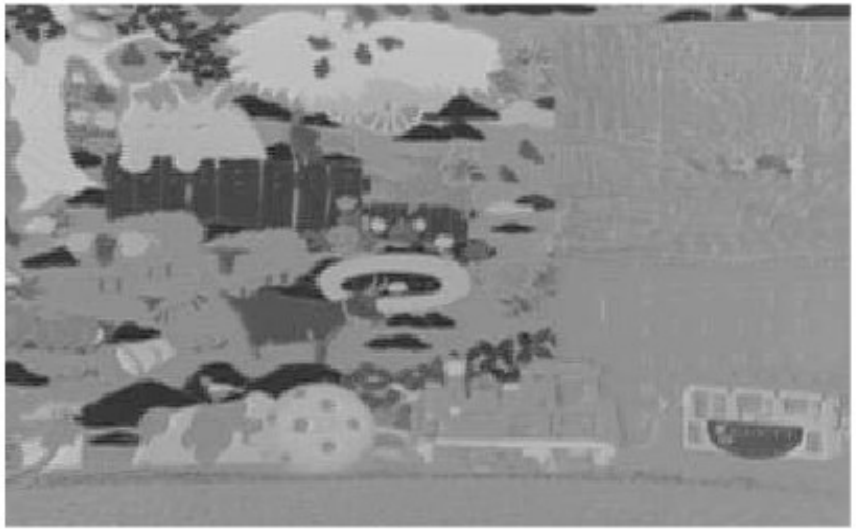
Another example: YIQ components. More correlation between I and Q than I and Y or Q and Y!

Color Image



Y image

I image (orange-cyan)



Q image (green-purple)

Conversion between RGB and YIQ, YCbCr

RGB -> YIQ

$$Y = 0.299\,R + 0.587\,G + 0.114\,B$$
$$I = 0.596\,R - 0.275\,G - 0.321\,B$$
$$Q = 0.212\,R - 0.523\,G + 0.311\,B$$

YIQ -> RGB

$$R = 1.0\,Y + 0.956\,I + 0.620\,Q$$
$$G = 1.0\,Y - 0.272\,I - 0.647\,Q$$
$$B = 1.0\,Y - 1.108\,I + 1.700\,Q$$

RGB -> YCbCr

$$\mathbf{Y} = 16 + 65.738*R/256 + 129.057*G/256 + 25.064*B/256$$
$$\mathbf{Cb} = 128 - 37.945*R/256 - 74.494*G/256 + 112.439*B/256$$
$$\mathbf{Cr} = 128 + 112.439*R - 94.154*G/256 - 18.285*B/256$$

YCbCr -> RGB

$$\mathbf{R} = 298.082*Y/256 + 408.583*Cr/256 - 222.921$$
$$\mathbf{G} = 298.082*Y/256 - 100.291*Cb/256 - 208.120*Cr/256 + 135.576$$
$$\mathbf{B} = 298.082*Y/256 + 516.412*Cb/256 - 276.836$$

Bandwidth of Chrominance Signals

- Theoretically, for the same line rate, the chrominance signal can have as high frequency as the luminance signal
- However, with real video signals, the chrominance component typically changes much slower than luminance
- Furthermore, the human eye is less sensitive to changes in chrominance than to changes in luminance
- The eye is more sensitive to the orange-cyan range (I) (the color of face!) than to green-purple range (Q)
- The above factors lead to
  - I: bandlimited to 1.5 MHz
  - Q: bandlimited to 0.5 MHz

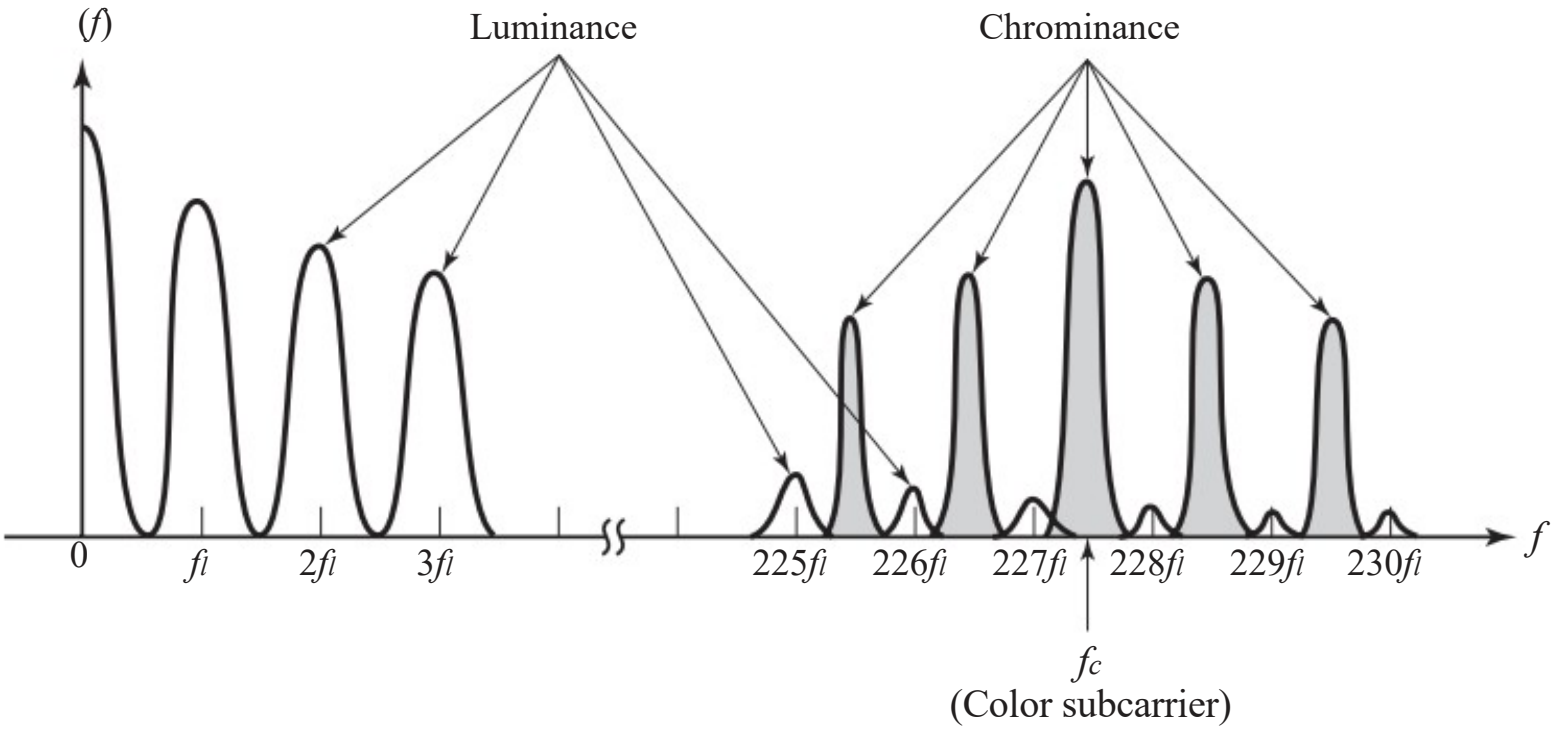
Multiplexing of Luminance and Chrominance

- Chrominance signal can be bandlimited
  - it usually has a narrower frequency span than the luminance and the human eye is less sensitive to high frequencies in chrominance
- The two chrominance components (I and Q) are multiplexed onto the same sub-carrier using QAM
  - The upper band of I is limited to 0.5 MHz to avoid interference with audio
- Position the bandlimited chrominance at the high end spectrum of the luminance, where the luminance is weak, but still sufficiently lower than the audio (at 4.5 MHz=286 f_l) (f_l is the line freq)
- The actual position should be such that the peaks of chrominance spectrum interlace with those of the luminance

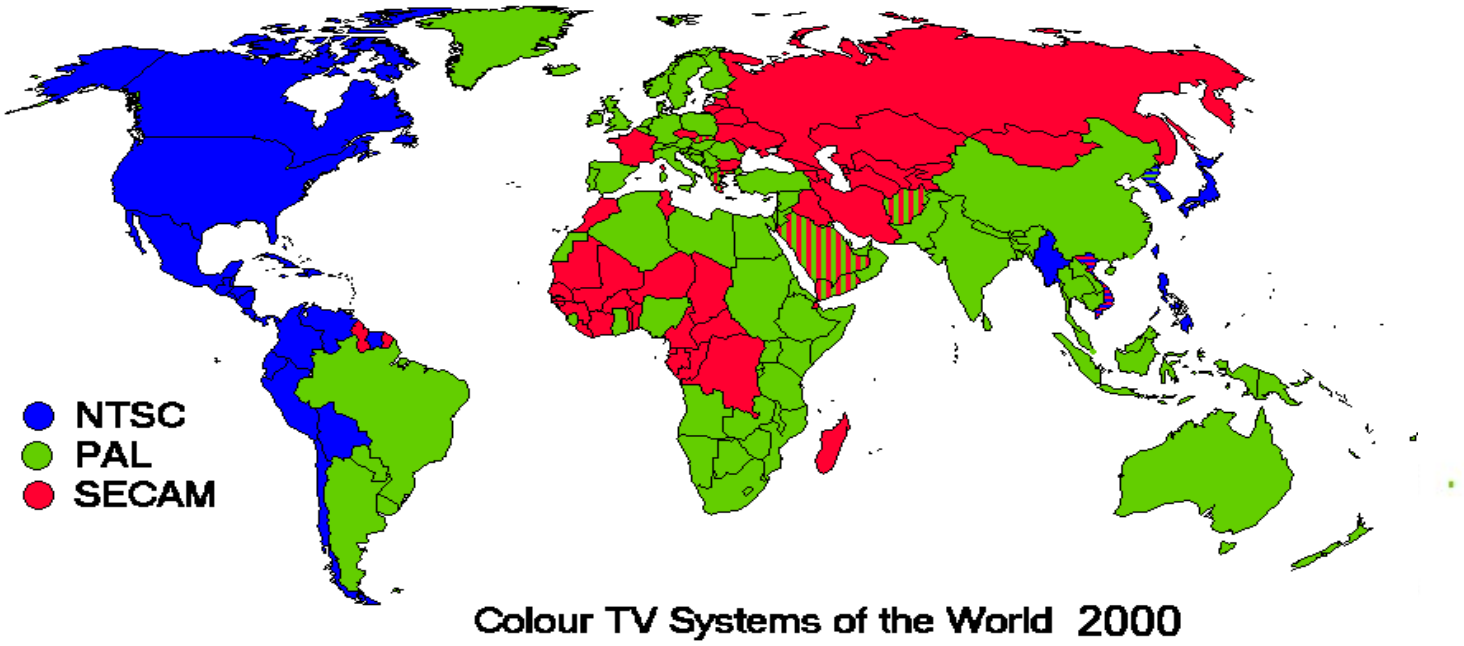
$f_c = 455 f_l/2$ (= 3.58 MHz for NTSC) (f_c is the color subcarrier freq)

QAM: Quadrature amplitude modulation modulates two signals, by changing the amplitudes of two carrier waves. The two carrier waves, usually sinusoids, are out of phase with each other by 90o

Spectrum Illustration



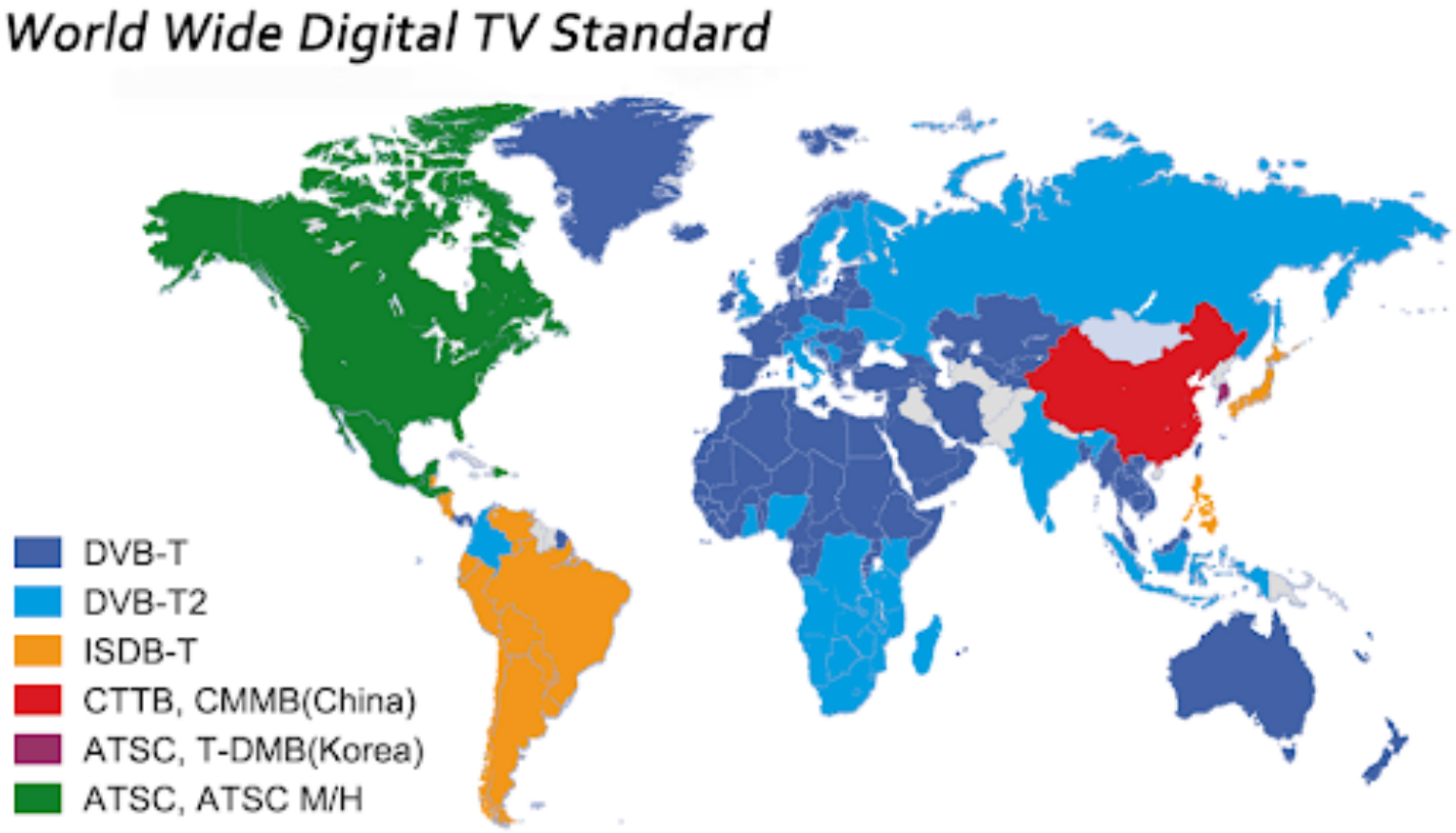
Who uses what?



NTSC: National Television System Committee
PAL: Phase Alternating Line
SECAM: Sequential Color with Memory

From http://www.stjarnhimlen.se/tv/tv.html#worldwide_0

Today's digital tv broadcasting



http://mww.dibvision.asia/info/world-wide-digital-tv-standard-introduction-25210898.html

DIGITAL VIDEO

Why Digital Video?

"Exactness"

- Exact reproduction without degradation
- Accurate duplication of processing result

Convenient & powerful computer-aided processing

- Can perform rather sophisticated processing through hardware or software

Easy storage and transmission after proper coding/compression

- Earlier, it was quite impressive how a DVD can store a three-hour movie, but now mobile phones come with 1TB storage)
- Transmission of high-quality video through network in reasonable time

Digital Video Raw Data Size

TABLE 1.3 DIGITAL VIDEO FORMATS FOR DIFFERENT APPLICATIONS

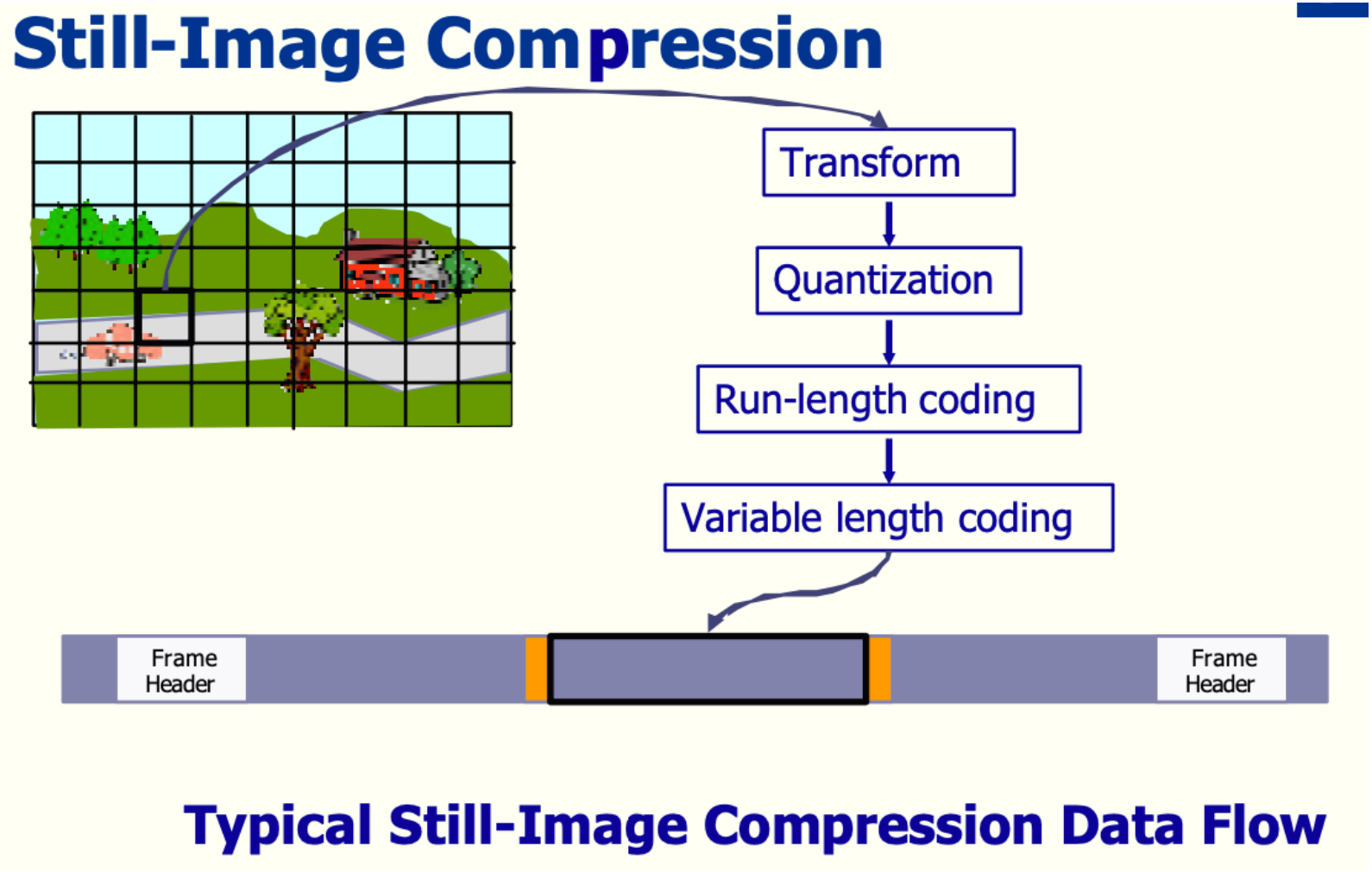
Video format	Y size	Color sampling	Frame rate	Raw data (mbps)
HDTV over air, cable, satellite, MPEG-2 video 20–45 mbps				
SMPTE 296M	1280 × 720	4:2:0	24P/30P/60P	265/332/664
SMPTE 295M	1920 × 1080	4:2:0	24P/30P/60I	597/746/746
Video production, MPEG-2, 15–50 mbps				
BT.601	720 × 480/576	4:4:4	60I/50I	249
BT.601	720 × 480/576	4:2:2	60I/50I	166
High-quality video distribution (DVD, SDTV), MPEG-2, 4–8 mbps				
BT.601	720 × 480/576	4:2:0	60I/50I	124

Digital Video Coding Principles

Still-Image Compression

- Still-image compression
  - Still-image techniques provide a basis for video compression
  - Video can be compressed using still-image compression individually on each frame
  - E.g., "Motion JPEG" or MJPEG
- But modern video codecs go well beyond this
  - Start with still-image compression techniques
  - Add motion estimation/compensation
    - Takes advantage of similarities between frames in a video sequence

Digital Video Coding Principles



Digital Video Coding Principles

The basic idea is to find and remove redundancy in video and encode it.

Three important types of redundancies:

A) Perceptual redundancy:

- The Human Visual System is less sensitive to color and high frequencies

B) Spatial redundancy:

- Neighboring pixels have close luminance levels
  - Low frequency

C) Temporal redundancy:

- Differences between adjacent frames can be small. Shouldn't we exploit this?

Hybrid Video Coding

"Hybrid" combination of Spatial, Perceptual, & Temporal redundancy removal

Issues to be handled

- Not all regions are easily inferable from previous frame
  - Occlusion is solved by backward prediction using future frames as reference
  - The decision of whether to use prediction or not is made adaptively
- Drifting and error propagation
  - Solved by encoding reference regions or frames at constant intervals of time
- Random access
  - Solved by encoding frame without prediction at constant intervals of time
- Bit allocation
  - according to statistics (more frequently used values are encoded with fewer bits!)
  - constant and variable bit-rate requirement

MPEG combines all of these features !!!

RGB-YCbCr Conversion

$$Yd = 0.257\,Rd + 0.504\,Gd + 0.098\,Bd + 16$$
$$Cb = -0.148\,Rd - 0.291\,Gd + 0.439\,Bd + 128$$
$$Cr = 0.439\,Rd - 0.368\,Gd - 0.071\,Bd + 128$$

And the inverse transform is:

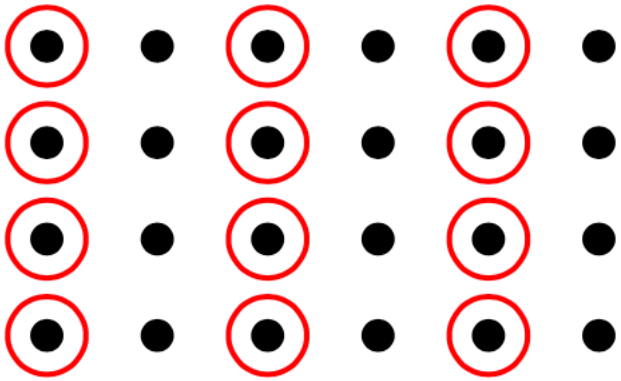
$$Rd = 1.164\,Yd' + 0.0\,Cb' + 1.596\,Cr'$$
$$Gd = 1.164\,Yd' - 0.392\,Cb' - 0.813\,Cr'$$
$$Bd = 1.164\,Yd' + 2.017\,Cb' + 0.0\,Cr'$$

Where

$$Yd' = Yd - 16,\ Cb' = Cb - 128,\ Cr' = Cr - 128$$

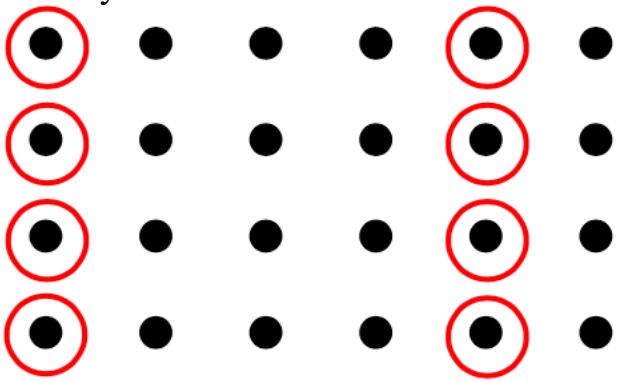
Chrominance Subsampling Formats

- Humans have a subjectively lower sensitivity to color,
- Color information is typically sampled at a lower rate than the intensity information



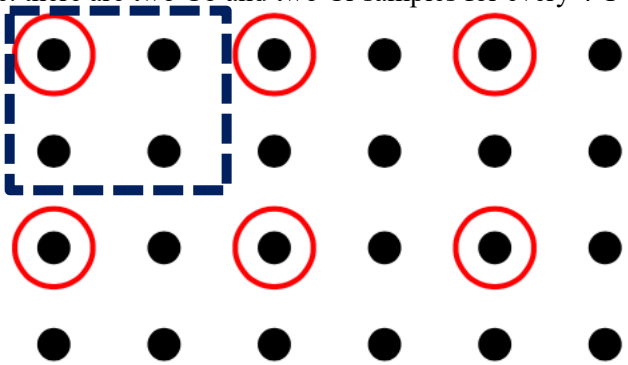
4:2:2

The color information is down-sampled by a factor of 2 horizontally from the full resolution intensity image
i.e. there are two Cb and two Cr samples for every 4 Y samples



4:1:1

4:1:1 sampling yields 1 Cb and 1 Cr sample for every 4 horizontal Y samples, notice the asymmetric resolution



4:2:0

Pixels with only Y value

Pixels with Y, Cb and Cr values

The color information is down-sampled by a factor of 2 horizontally and vertically from the full resolution intensity image, again 1 Cb and 1 Cr for every 4 Y samples, notice the more symmetric resolution

The **4:4:4** sampling structure represents video in which the color components of the signal are sampled at the same rate as the luminance signal

Note on the notation

The subsampling scheme is commonly expressed as a three-part ratio ***J*:*a*:*b***

- *J*: horizontal sampling reference (width of the conceptual region, usually 4)
- *a*: number of chrominance samples (Cr, Cb) in the first row of *J* pixels
- *b*: number of changes of chrominance samples (Cr, Cb) between first and second row of *J* pixels.

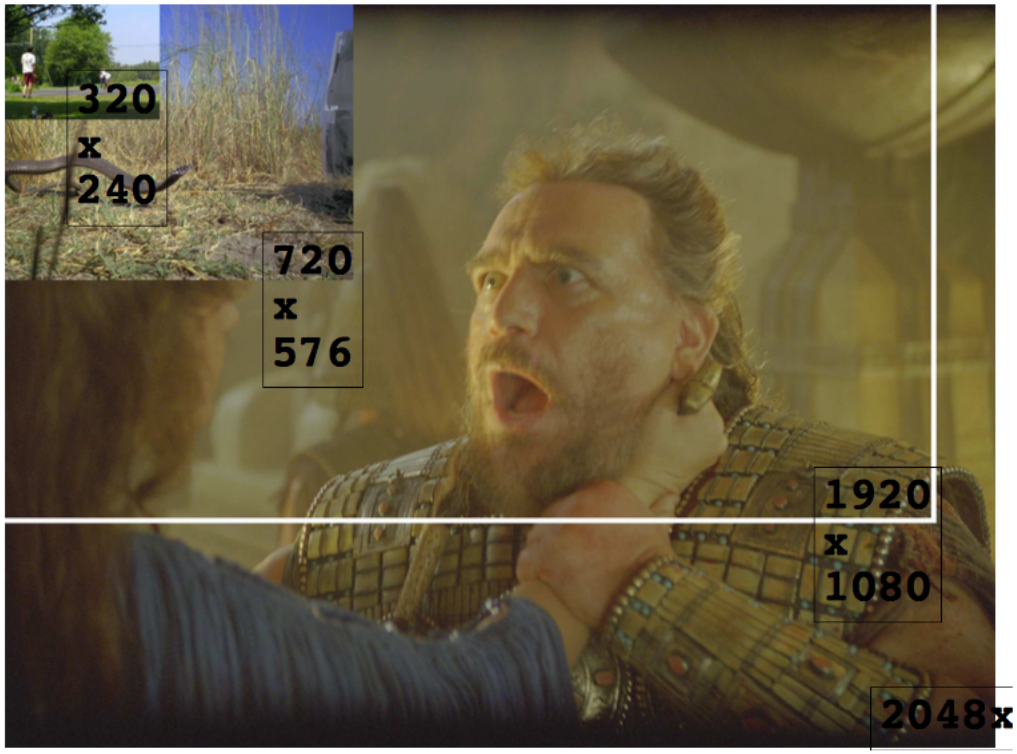
Digital Video Formats
(defined by the CCIR Recommendation 601)

Format	Total Resolution	Active Resolution	MB/sec
CCIR 601 30 frames/sec, 4:3 Aspect Ratio, 4:2:2, NTSC			
QCIF	214 × 131	176 × 120	1.27
CIF	429 × 262	352 × 240	5.07
Full	858 × 525	720 × 485	20.95
CCIR 601 25 frames/sec, 4:3 Aspect Ratio, 4:2:2, PAL			
QCIF	216 × 156	176 × 144	1.27
CIF	432 × 312	352 × 288	5.07
Full	864 × 625	720 × 576	20.74

CIF = Common Interchange Format, QCIT is Quarter CIF

High Definition Formats:
1920 x 1080 and 1280 x 720 (both at 16:9 aspect ratio)
1080i, 1080p and 720i, 720p (i is for interlaced and p for progressive, referring to the scan pattern used in recording and displaying the picture)

Digital Video Formats



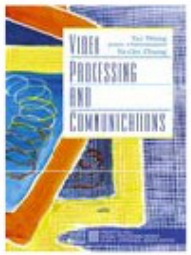
CIF

Full

HD

4K

Figure 5: From small to large : Approximately CIF, SD (PAL), HD (1080p) and Digital Film Picture Sizes.



Key Ideas in Video Compression

- Predict a new frame from a previous frame and only code the prediction error --- **Inter-frame prediction**
- Predict a current block from previously coded blocks in the same frame --- **Intra-frame prediction** (introduced in the latest standard H.264)
- Prediction error will be coded using a transform such as DCT
- Prediction errors have smaller energies than original pixel values and can be coded with fewer bits
- Those regions that cannot be predicted well will be coded directly using the transform (e.g. DCT) --- **Intra coding without intra-prediction**
- Work on each macroblock (MB) (16x16 pixels) independently for reduced complexity
  - Motion compensation done at the MB level
  - DCT coding of the difference image is done at the block level (usually 8x8 pixels)

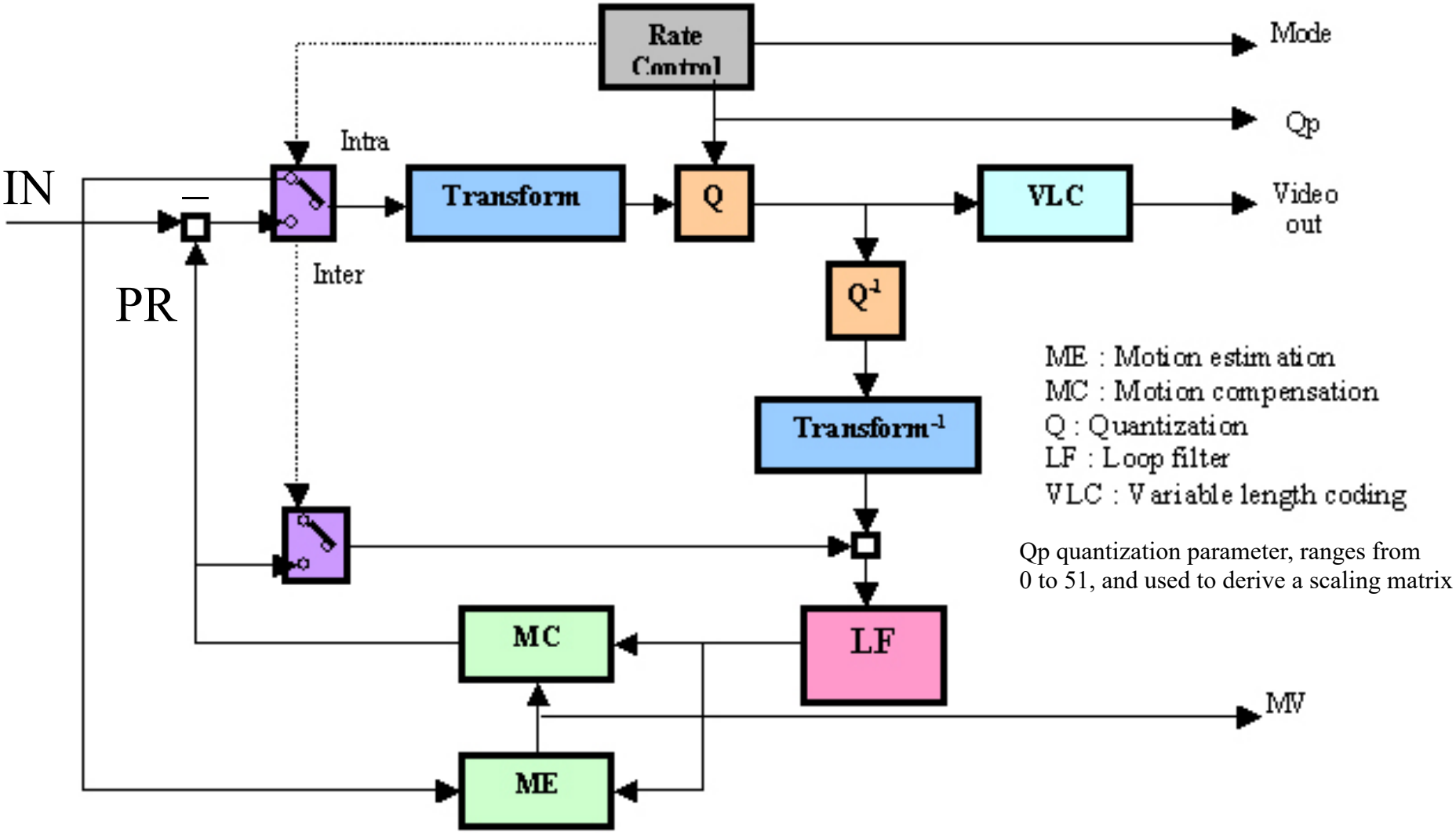
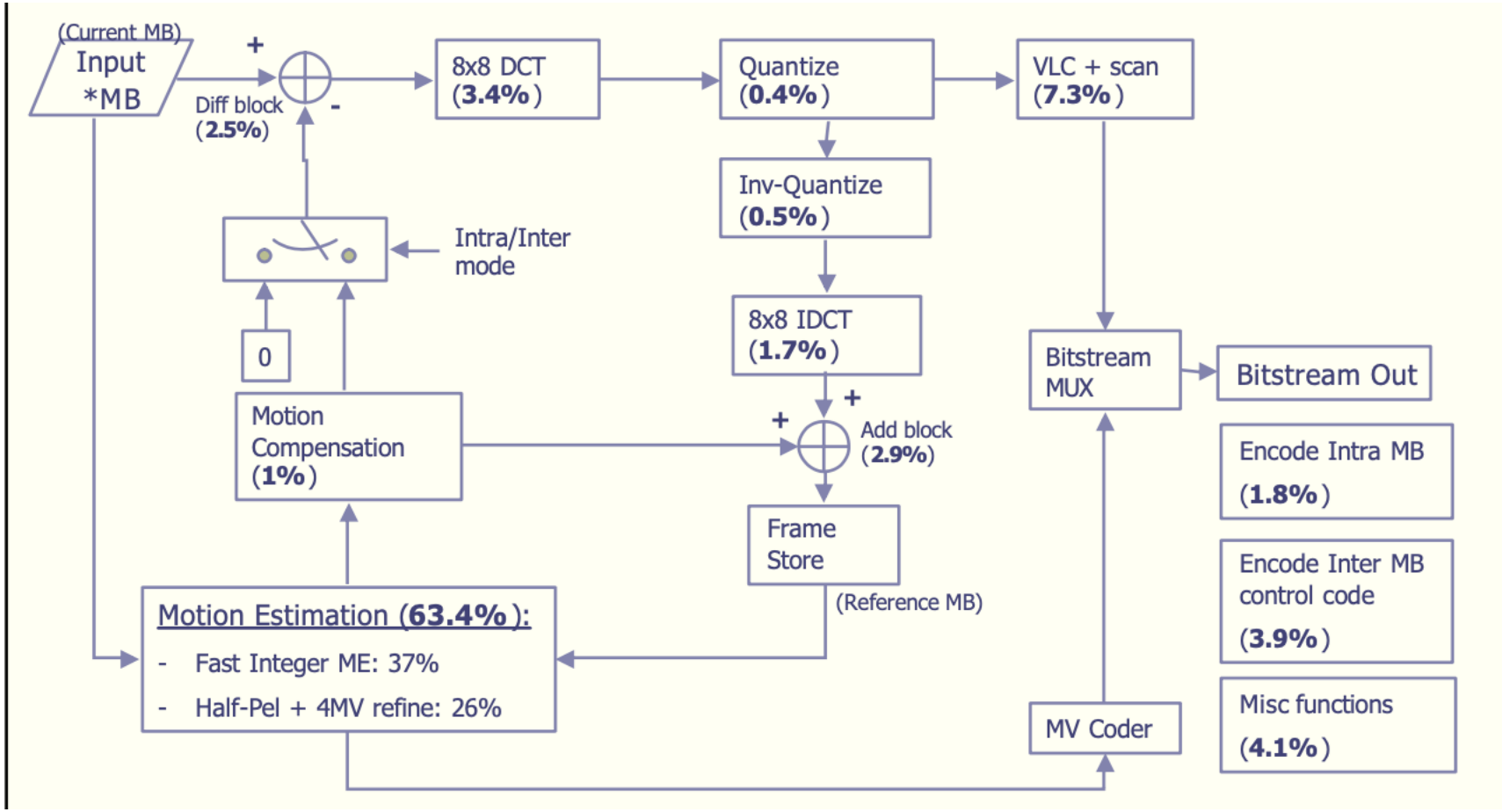
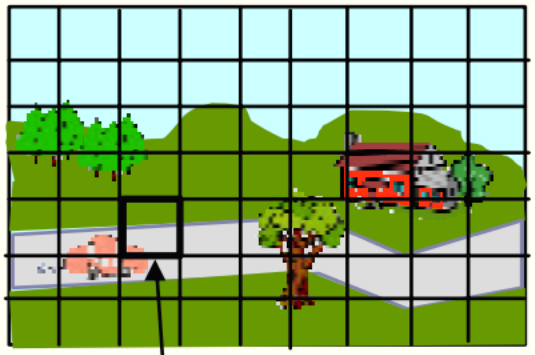


Figure 1: Generic Video encoder's block diagram

Video Encoder Block Diagram

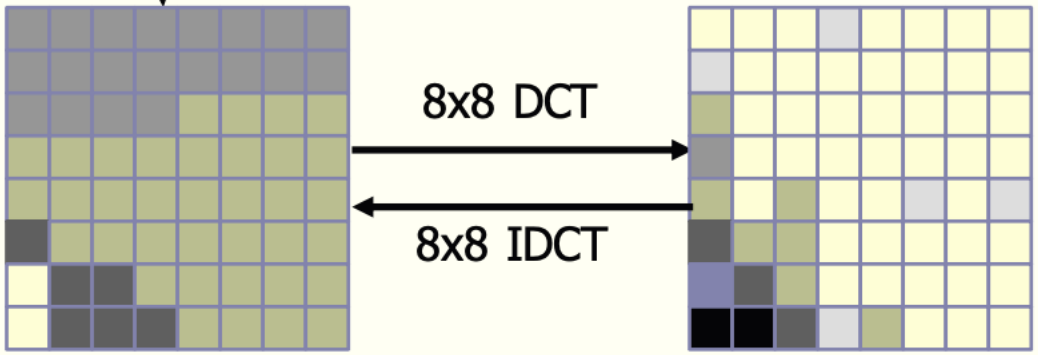


Block Transform: 8x8 DCT

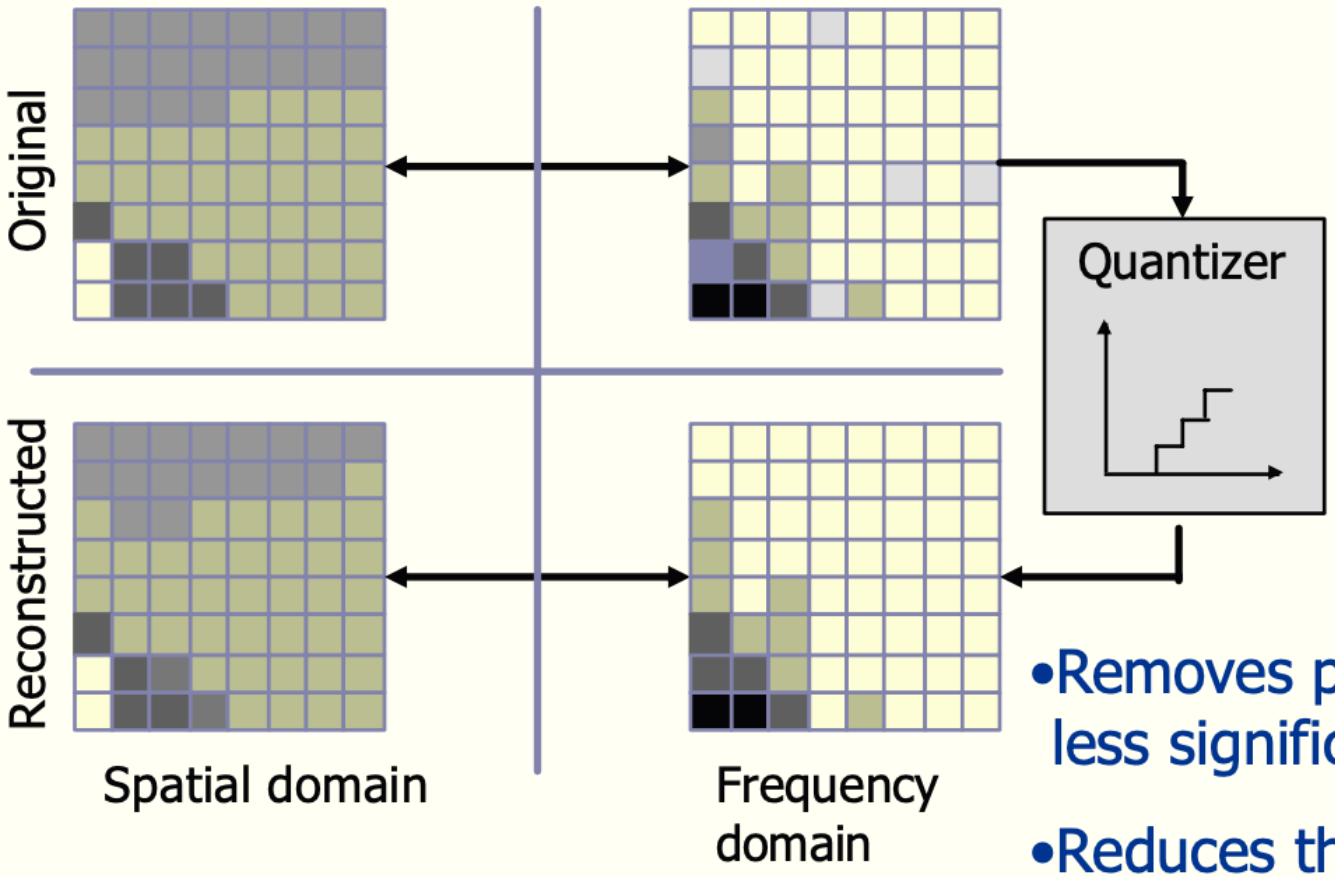


Y values

- 8x8 DCT blocks applied on Y, U, and V planes individually
- The energy is concentrated in the low frequencies
- Perceptual information also concentrated in low frequencies



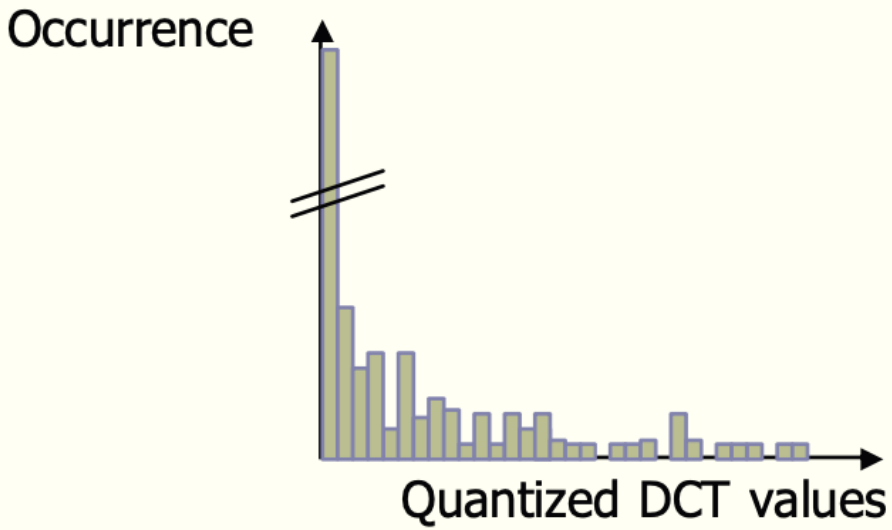
Quantization



- Removes perceptually less significant data
- Reduces the number of bits per DCT coefficient

Coding Quantized DCT Coefficients

Goal: Reduce the number of bits required to transmit the quantized coefficients

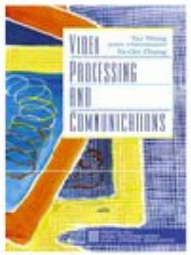


Observation: Unequal distribution of quantized DCT coefficient values

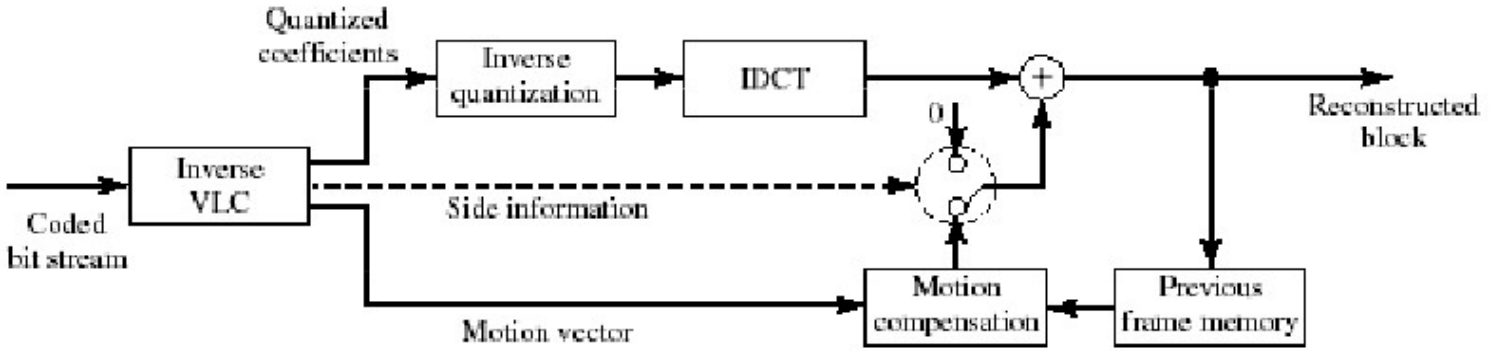
Variable Length Coding (VLC/VLD)

- Allocates fewer bits to the most frequent symbols (e.g., using Huffman)
- Integer number of bits per symbol
  - Not the most efficient coding method
  - Arithmetic coding more efficient, but expensive
  - Run-length coding improves efficiency of VLC/VLD for image and video coding

Symbol	Frequency	Code
A	22	1
B	16	011
C	9	0101
D	7	0100
E	4	0011
F	2	0010
...	...	...



Decoder Block Diagram



Decoder

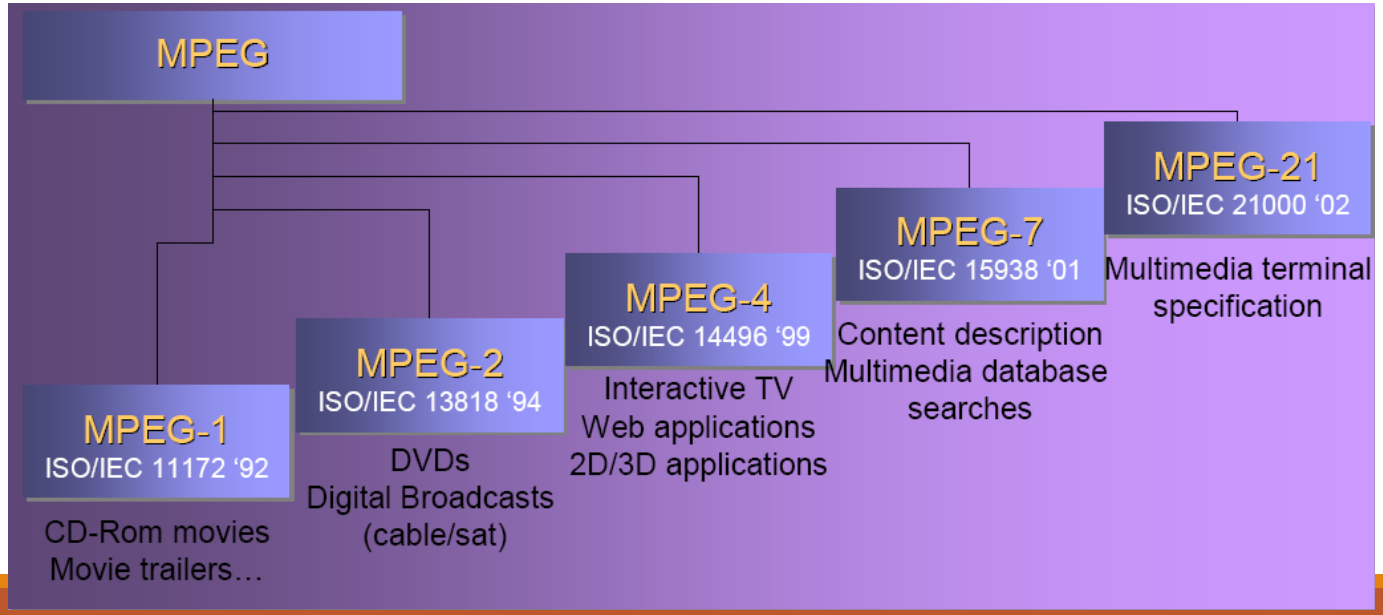
MPEG

MPEG – Moving Picture Experts Group

- Coding of moving pictures and associated audio

Picture part

- Can achieve compression ratio of about 50:1 through storing only the difference between successive frames
-



Bit Rate (ITU-R BT.601)

According to the standard, the spatial resolution was set to

$$f_s = 13.5 \text{ MHz}$$

4:2:2 there are 2 chrominance samples per 2 Y samples:

→ equivalent bitrate for each Y sample is $N_b = 16$ bits and the raw data rate is

$$f_s N_b = 16b \times 13.5MHz = 216 \text{ Mbps}$$

4:2:0, raw bit rate is 162 Mbps

4:4:4, raw bit rate is 324 Mbps

MPEG-1 Compression Aspects

Lossless and Lossy compression are both used for a high compression rate

Down-sampled chrominance
- Perceptual redundancy

Intra-frame compression
- Spatial redundancy
- Correlation/compression within a frame
- Based on “baseline” JPEG compression standard

Inter-frame compression
- Temporal redundancy
- Correlation/compression between frames

Audio compression
- MP3: reduces the accuracy of certain parts of a sound that are considered to be beyond the auditory resolution ability of most people.

Perceptual Redundancy

Here is an image represented with 8-bits per pixel

Perceptual Redundancy

The same image at 7-bits per pixel

Perceptual Redundancy

At 6-bits per pixel

Perceptual Redundancy

At 5-bits per pixel

Perceptual Redundancy

At 4-bits per pixel

Perceptual Redundancy

It is clear that we don’t need all these bits!

- Our previous example illustrated the eye’s sensitivity to luminance

We can build a perceptual model

- Give more importance to what is perceivable to the Human Visual System
  - Usually this is a function of the spatial frequency