CMPT 365 - Multimedia systems

January 8th - First day, on my phone

Our teacher is a phd student supervised by the teacher who usually teaches this course? (Dr. Jiangchaun Liu)

What is media?
Text, images, videos, etc. We have different data representations of information. Multimedia is the combination of these. Example: JPEG standard for pictures, H.264 video, ascii standard for text characters, HTTP protocol, etc.

Discrete media: time independent. Text, images, graphics, etc.
Continuous: time dependent. Animation, audio, video

Analog vs. Digital
Analog: the time-varying feature of the signal is a continuous representation of the input. I.e. analogous to the input audio, image, or video signal

We will focus a lot on compression, and signal processing.

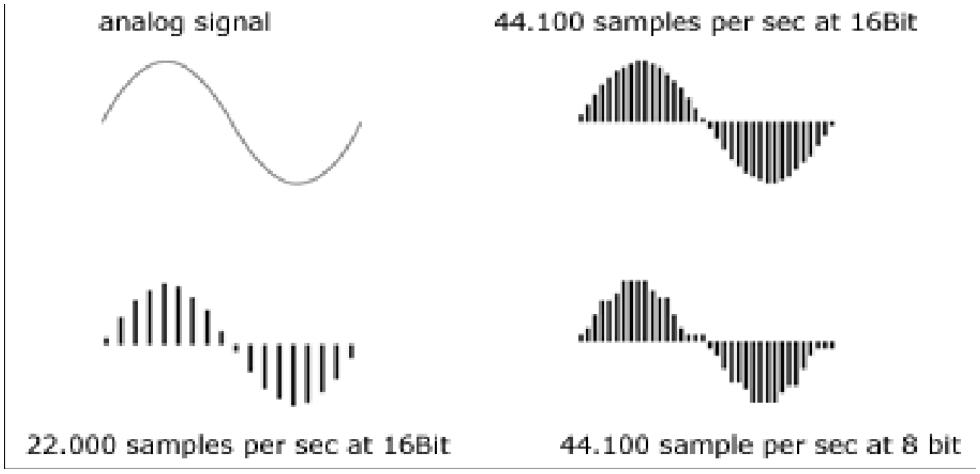
Representation (audio/video)
- Digitization
- Quanization: different data types that round and estimate (floats)

Compression
- Transform
- Entropy coding
- Coding standards

January 15th - Second class

Compression methods have a tradeoff between compressed file size (how much data is saved from being stored) vs. quality of compression (lossy compression)

Compressions remove redundant information:
Spatial redundancy: • Neighboring samples have similar values
Temporal redundancy: • Neighboring frames in a video sequence are similar



For this audio signal, each virtual bar is the amplitude and each sample is a measurement of the amplitude.

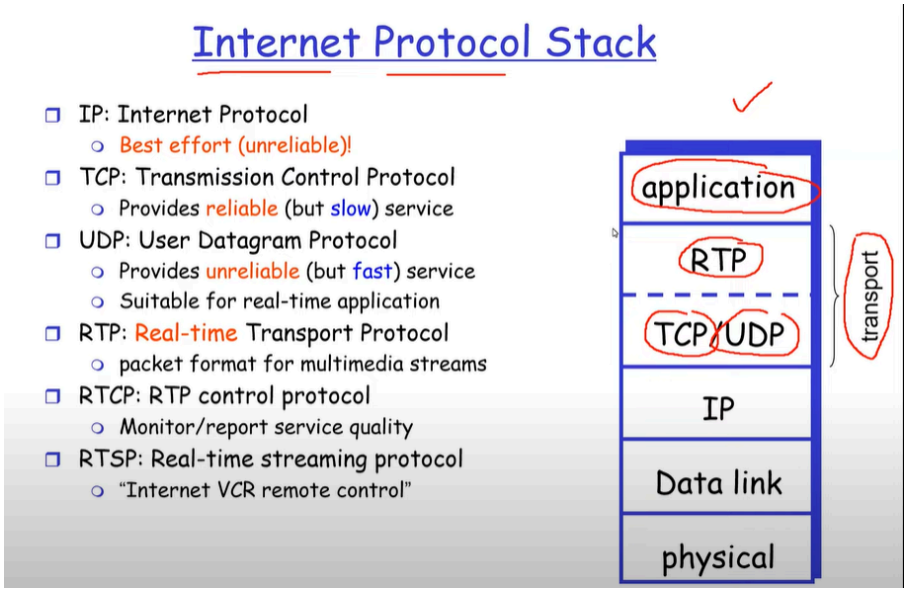
Why have we digitized? Analog has

- Bulky storage (space, cost, lifetime)
- Poor quality
- Poor/no compression
- Poor portability/mobility/editibility
- MP3 player, iPod, YouTube ? No way
  Film -> Polaroid -> Digital camera

January 22nd - Transit strike online class

Audio representation with multimedia needs to account for real life transmission delay, synchronizing with the media and be tolerant to transmission errors.

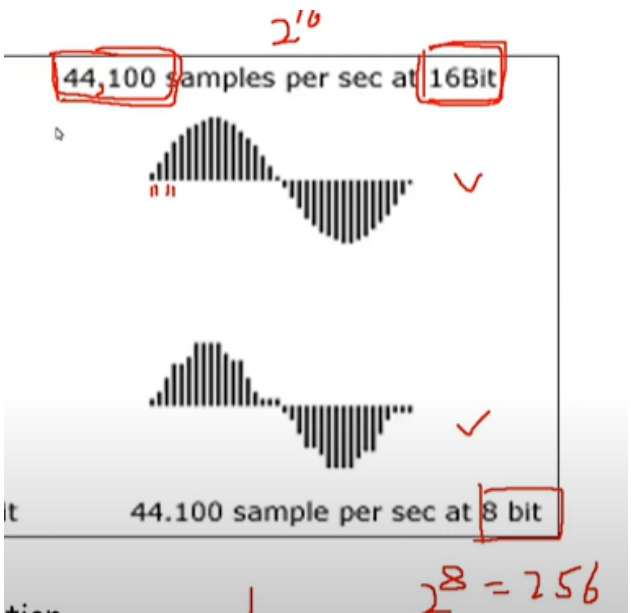
Internet is packet switched, meaning there's no dedicated connection from user to user, all resources are shared.
The internet will add redundancy at encoder with ecc and layered code.



The internet has quality of service protocols

- End-to-end delay
    - Time required for the end-to-end transmission of a single data element
- Jitter
    - Variation of delay
- Packet loss rate
    - Proportion of data elements that are dropped
- Bandwidth: bits/second

Now let's switch gears to audio.
We have to digitize an analog signal with amplitude and time variation. The digitization must be in both time and amplitude with sampling. In this course we will focus on constant frequency sampling.



At 16 bits, we can have 2^16 integer options for the amplitude, compared to 2^8 for the 8 bit option.
This is obviously a drop in sound quality, but 16 bits comes at a cost of file size increases.

Pulse code modulation (PCM) - audio signal represented by amplitude and time graph.
We have several questions, like what should be the sampling rate? For each sample, there will be a quantization range. Meaning if the amplitude is in a certain range, it will be assigned a digitized value. A uniform quantization has equal length intervals.
Obviously, this is lossy and will experience some rounding error. We also have quantization noise, which is the difference between the actual value of the analog signal and the nearest quantization interval value. At most this error can be as much as half the interval.
We can measure the quality of quantization by signal to quantization noise ratio (SQNR). Define signal to noise ratio as $10*\log_{10}(V^2_{signal}/V^2_{noise}) = 20*\log_{10}(V_{signal}/V_{noise})$. Where V is voltage of signal, and voltage represents amplitude of sound. Even though quantization will introduce a roundoff error, it is called quantization noise even if it really isn't "noise".

For a quantization accuracy of N bits per sample, the peak SQNR can be simply expressed:

$$SQNR = 20\log_{10}\frac{V_{signal}}{V_{quan\_noise}} = 20\log_{10}\frac{2^{N-1}}{\frac{1}{2}}$$

$$= 20 \times N \times \log 2 = 6.02N(\text{dB}) \qquad (6.3)$$

Meaning 6.02*N is worst case. This is where max signal is mapped to 2^(N-1) - 1 and most negative signal is to -2^(N-1).
And finally, we have dynamic range: the ratio of maximum to minimum absolute values of signal. V_{max}/V_{min}. The max abs. value V_{max} gets mapped to 2^(N−1) − 1; the min abs. value V_{min} gets mapped to 1. V_{min} is the smallest positive voltage that is not masked by noise. The most negative signal, $-V_{max}$, is mapped to −2^(N−1).

If sampling rate is equal to frequency, a false (constant) signal is detected. If the sampling rate is 1.5 times frequency, aliasing will occur and produce an incorrect smaller frequency.
For correct sampling we must use a sampling rate equal to at least twice the maximum frequency content in the signal. This rate is called the nyquist rate.

If the lower frequency limit is f_L and the upper limit is f_U, the sampling rate should be at least $2*(f_U - f_L)$.

We also have mono/stereo sound, where we have 1/2 data channels. In 2 channel audio, one channel is what you should hear from the right ear speaker and the other is for the left ear speaker.

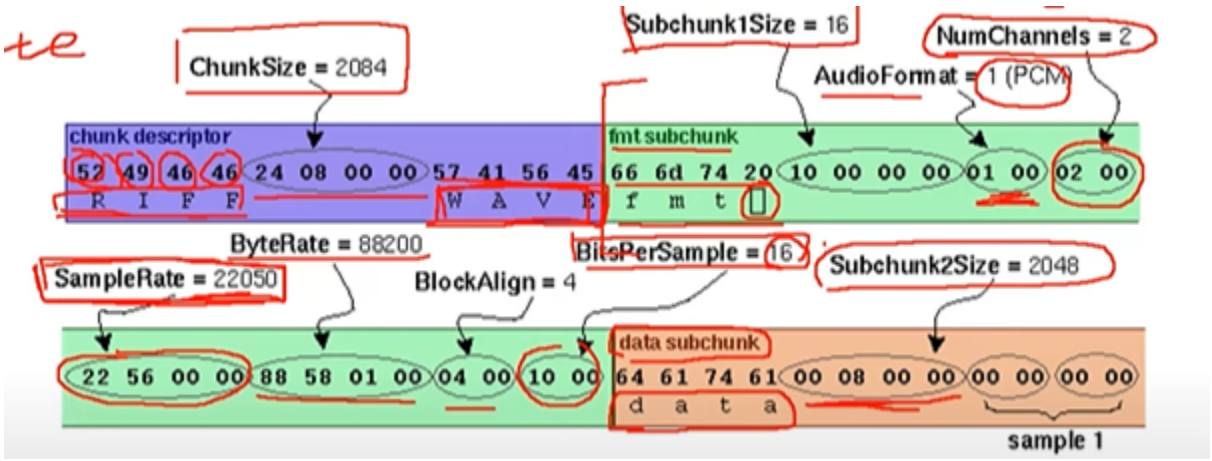
January 29th - Audio and image representation

WAV audio format

PCM audio format, first 12 bytes is a descriptor.

- First 4 bytes - chunk descriptor
- Next 4 bytes: Chunk size (file size basically)
- Last 4 bytes: Spells WAVE (format)

First subchunk: fmt[] reads format



After sample 1 we have 6 more 2 byte samples.

We have a wavread format in matlab.

For python plotting we have canvas library with create-line().

MIDI format

MIDI doesn't store the actual signal, it stores the instructions on how to recreate

the signal. It’s like a script.

Human audio perceptron

By understanding human interpretation we can do compressions on this.Our average range of human hearing is 20Hz to 20kHz, minimal sampling rate for music is 40 kHz by Nyquist frequency.

CD audio is a 44.1 kHz sampling rate standard where each sample is represented by a 16-bit signed integer. 2 channel stereo system.
44100*16*2 = 1.411,200 bits / second.

Speech signal: 300 Hz - 4kHz.

There are certain sounds that make us uncomfortable. Part 3 audio rep video.

Critical bands: Our brains perceive the sounds through 25 distinct critical bands, where frequencies in the same critical band cannot be perceived as different.

Masking: Our brain will also mask and filter other sounds. Band pass filter can filter out frequencies we can’t hear or don’t want. So we can have different quantizations for different critical bands.

Media representations for images

361 and 203 review.

PPM and PGM formats have fixed identifiers, comments, image size, max value (for rgb or grayscale), PPM has max value (R,G,B) of pixel (1,1) and (R,G,B) of (1,2).

We can also convert color space from rgb to yuv, as in yuv y represents brightness and u and v represent color.

U, V represent chrominance components (blue and red), the difference between a color and reference.

Chrominance refers to the difference between a color and a reference white at the same luminance. → use color differences U, V:

$$U = B' - Y', \qquad V = R' - Y' \tag{4.27}$$

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.299 & -0.587 & 0.886 \\ 0.701 & -0.587 & -0.114 \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} \tag{4.28}$$

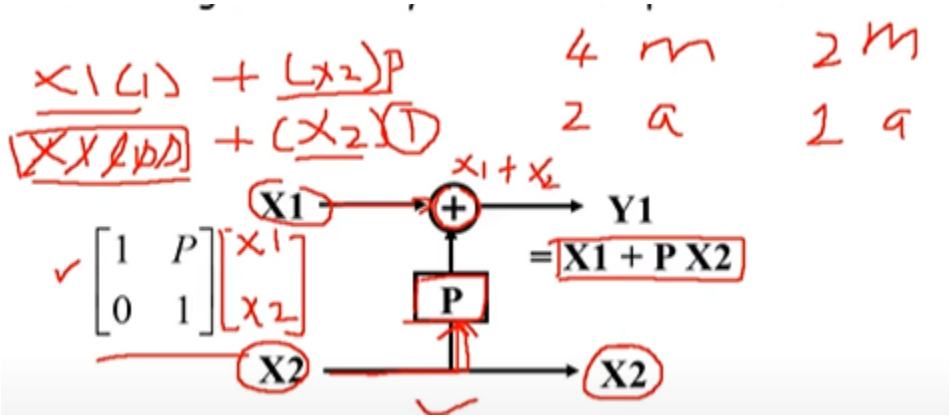
For a gray color, $R' = G' = B'$, the luminance Y' equals to that gray, since 0.299+0.587+0.114 = 1.0. And for a gray ("black and white") image, the chrominance (U, V) is zero.

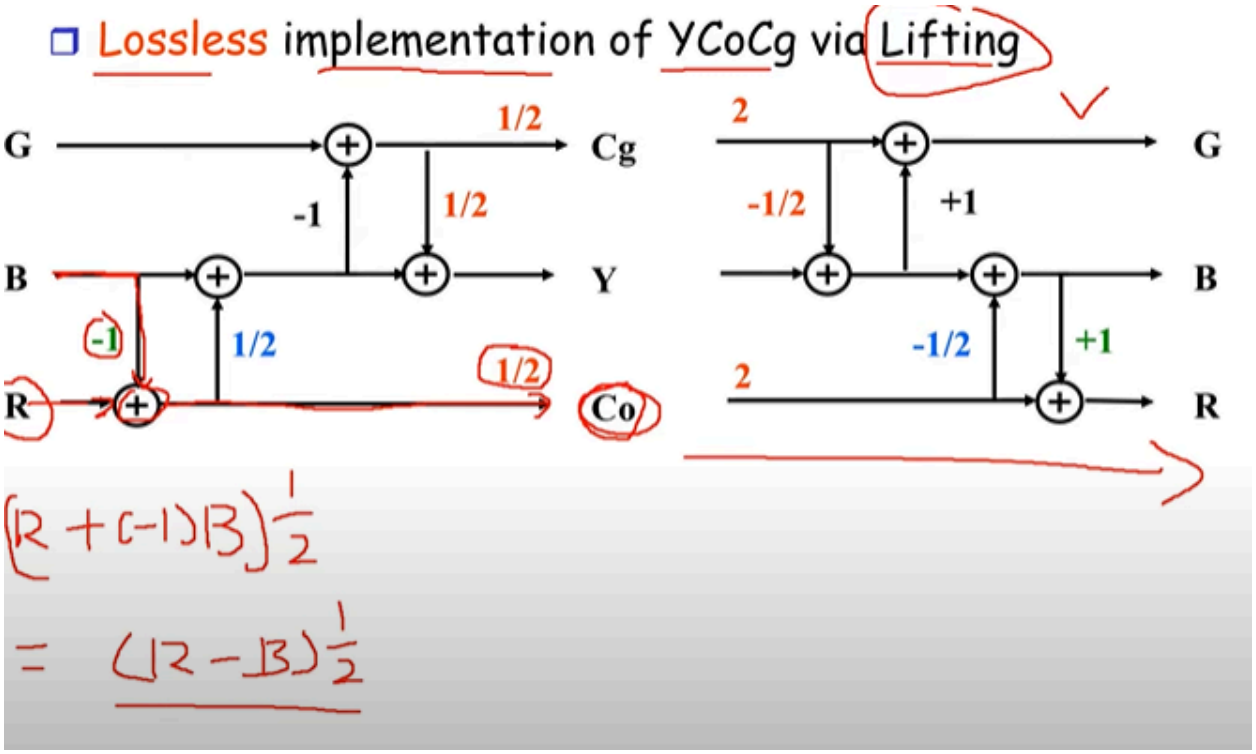
YUV has floating point implementation and there is rounding error during rgb → yuv → rgb.

YCoCg space

$$\begin{bmatrix} Cg \\ Y \\ Co \end{bmatrix} = \begin{bmatrix} 1/2 & -1/4 & -1/4 \\ 1/2 & 1/4 & 1/4 \\ 0 & -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} G \\ B \\ R \end{bmatrix} \qquad \begin{bmatrix} G \\ B \\ R \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} Cg \\ Y \\ Co \end{bmatrix}$$

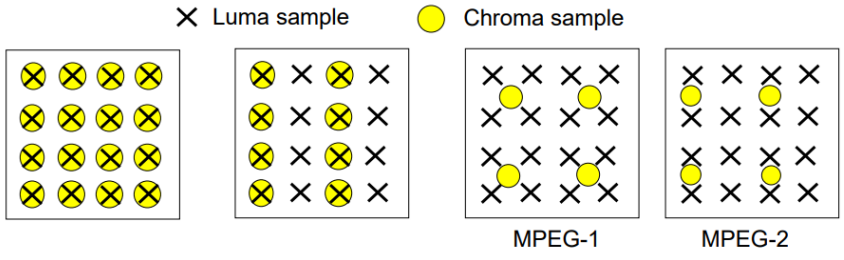
Matrix lifting: Instead of doing matrix multiplication, we simplify it by looking at matrix and getting result directly.





February 5th

Down sampling colour components to improve compression.



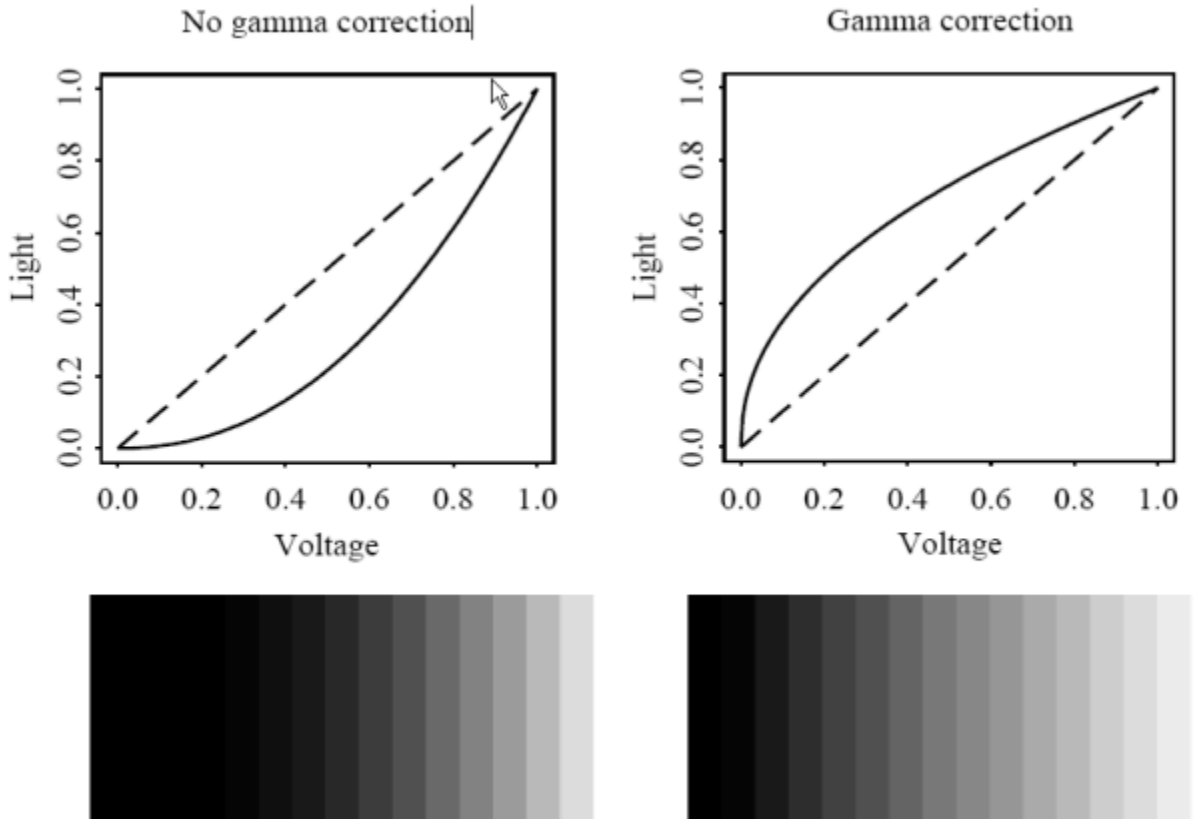
Gamma correction: Brightness alteration

Light emitted is roughly proportional to the voltage raised to a power (exponent γ) -- gamma,

- If the file value in the red channel is R, the screen emits light proportional to R^γ .
- Typical gamma is around 2.2 but depends on display device
- Gamma-correction: raising to the power (1/ γ) before transmission. Thus arrive at linear signals:

$$R \rightarrow R' = R^{1/\gamma} \Rightarrow (R')^{\gamma} \rightarrow R$$

Getting a lightblub with a light level 5, having double the lightbulbs would have a light level of 5.5-6.
Without gamma correction, voltage to light level is too low.



"Just raise the power of the value to some gamma" - Assignment

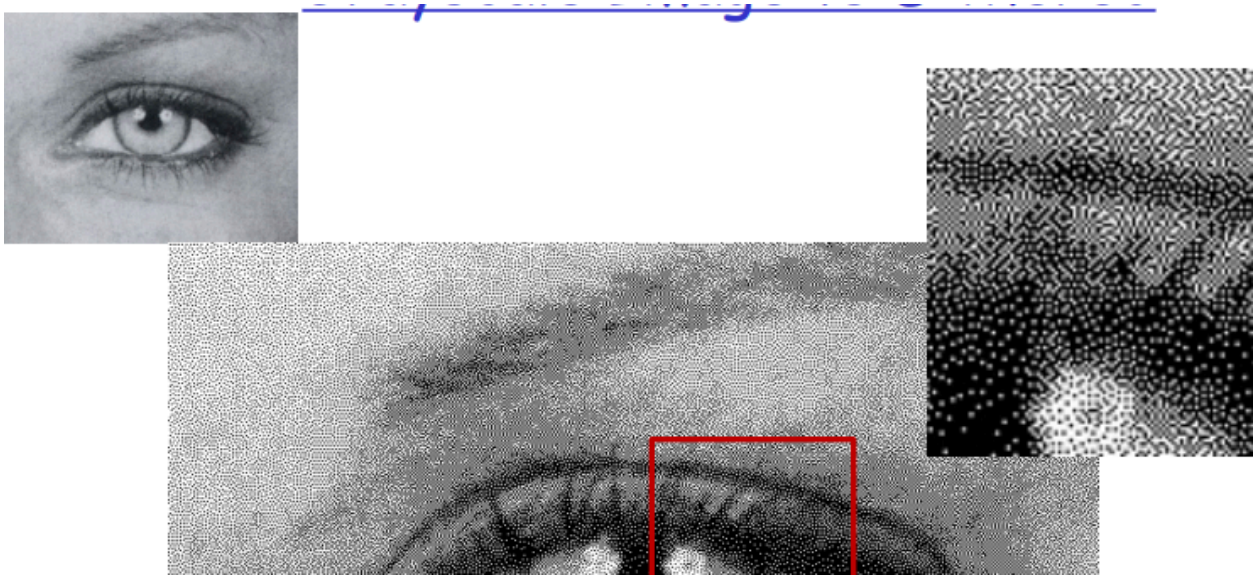
8-bit Gray-level images

Bitmap: The two-dimensional array of pixel values that represents the graphics/image data.

Frame buffer: Hardware used to store bitmap. Video card (actually a graphics card) is used for this purpose. The resolution of the video card does not have to match the desired resolution of the image, but if not enough video card memory is available then the data has to be shifted around in RAM for display.

8-bit image can be thought of as a set of 1-bit bit-planes, where each plane consists of a 1-bit representation of the image at higher and higher levels of "elevation": a bit is turned on if the image pixel has a nonzero value that is at or above that bit level.

Each pixel -> a byte (a value between 0 to 255) 640x480 grayscale image requires 300 kB of storage (640 x 480 = 307, 200).
Print a grayscale image on Black/White newspaper (or laser printer)? Dithering is used, which trades intensity resolution for spatial resolution to provide the ability to print multi-level images on 2-level (1-bit) printers. Also known as halftone printing
Rationale: for printing gray level on a 1-bit printer, calculate square patterns of dots such that gray level from 0 to 255 (i.e., darkness) correspond to patterns that are more and more filled at darker pixel values.



Strategy: Replace a pixel value by a larger pattern, say 2x 2 or 4 x 4, such that the number of printed dots approximates the varying-sized disks of ink used in analog, in halftone printing (e.g., for newspaper photos).

1) Half-tone printing is an analog process that uses smaller or larger filled circles of black ink to represent shading, for newspaper printing.
2) For example, if we use a 2 x 2 dither matrix

$$\begin{pmatrix} 0 & 2 \\ 3 & 1 \end{pmatrix}$$

The rule is:

If the intensity is > the dither matrix entry then print an on dot at that entry location: replace each pixel by an n x n matrix of dots.

Meaning replacing the pixel with the 2x2 dither matrix, it would 4x data size. A clever trick can get around this data size increase. Suppose we wish to use a larger, 4 x 4 dither matrix, such as

$$\begin{pmatrix} 0 & 8 & 2 & 10 \\ 12 & 4 & 14 & 6 \\ 3 & 11 & 1 & 9 \\ 15 & 7 & 13 & 5 \end{pmatrix}$$

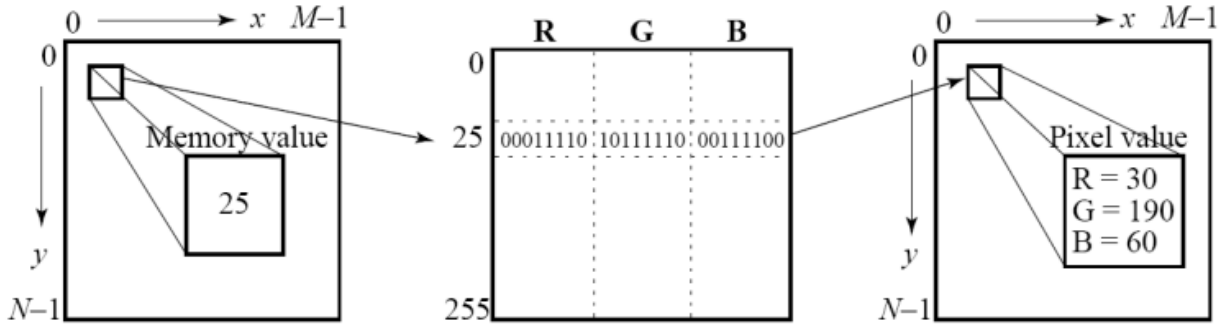
An ordered dither consists of turning on the printer out-put bit for a pixel if the intensity level is greater than the particular matrix element just at that pixel position.

Each entry in the ordered dither matrix will correspond to one pixel in the image, meaning if the image pixel is greater to the corresponding matrix entry, output printer bit. After the 4x4 matrix area is exhausted, we shift the matrix to the right.

```
BEGIN
    for x = 0 to x_max                  // columns
        for y = 0 to y_max              // rows
            i = x mod n
            j = y mod n
            // I(x, y) is the input, O(x, y) is the output,
            //D is the dither matrix.
            if I(x, y) > D(i, j)
                O(x, y) = 1;
            else
                O(x, y) = 0;
END
```

We can do coloured dithering by applying dithering to each colour.

For RGB images, even though we can have 256*256*256 possible colours, we don't need to save all these possible colours. We can instead store the colours stored in a table, and index by the possible colours for each pixel.
This is called a lookup table.
"Basically, the image stores not colour, but instead just a set of bytes, each of which is actually an index into a table with 3-byte values that specify the colour for a pixel with that lookup table index."



The idea used in 8-bit color images is to store only the index, or code value, for each pixel. Then, e.g., if a pixel stores the value 25, the meaning is to go to row 25 in a color look-up table (LUT).

February 12th

An analog signal f(t) samples a time-varying image. For analog video we have
Progressive scanning

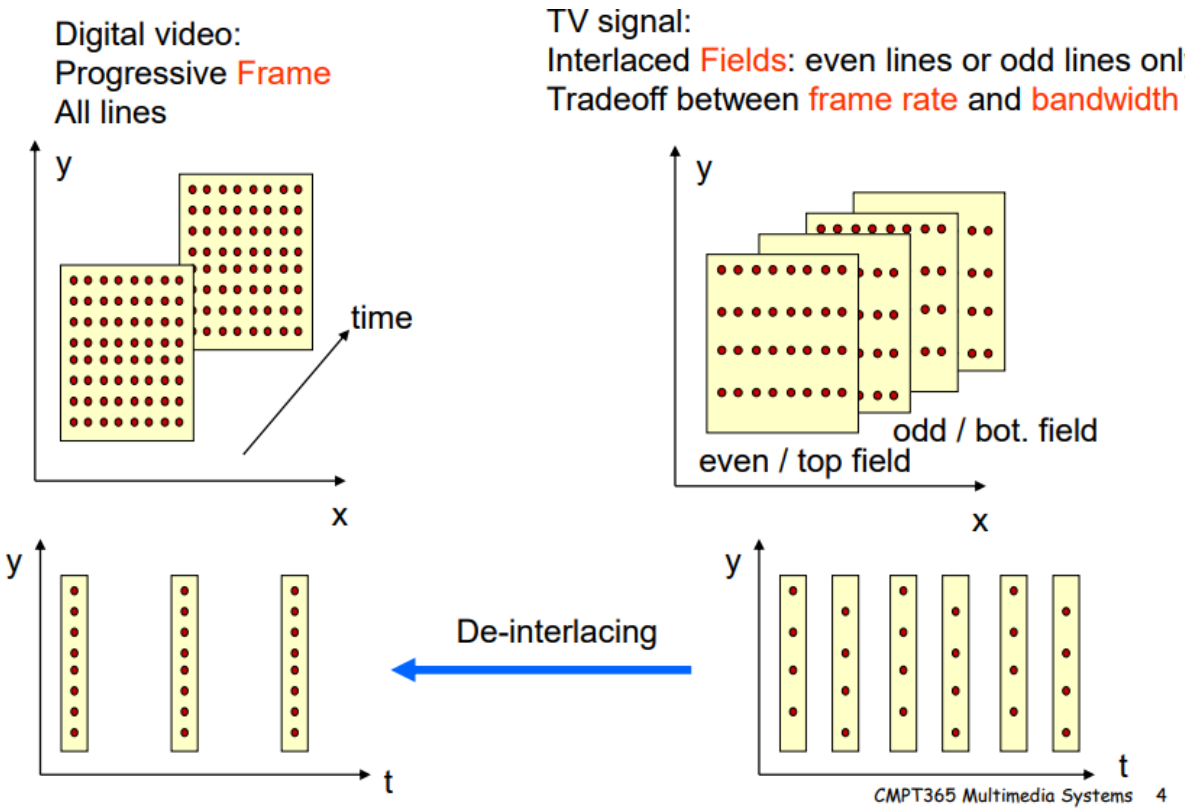
traces through a complete picture (a frame) row-wise for each time interval.

Interlaced scanning

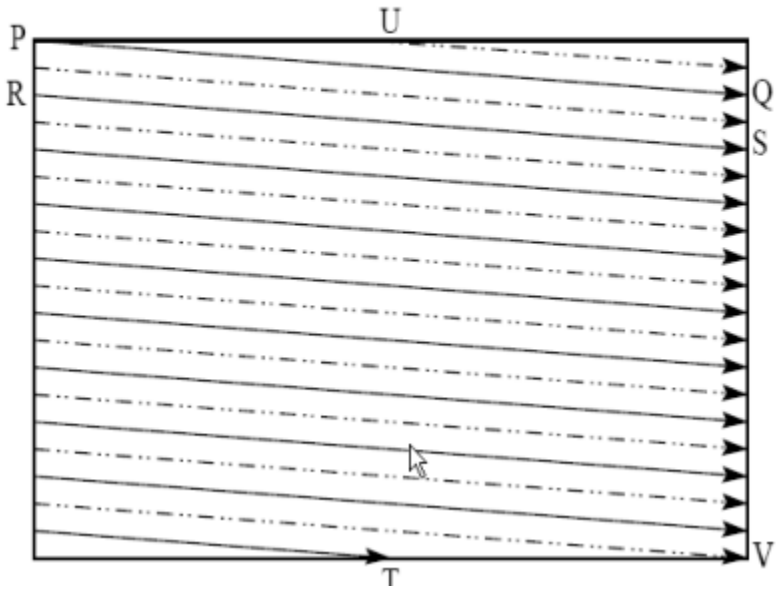
Odd-numbered lines traced first, and then the even numbered lines.

"odd" and "even" fields - two fields make up one frame

Widely used in traditional (non-digital) TV



Interlaced Scan



• First the solid (odd) lines are traced, P to Q, then R to S, etc., ending at T; then the even field starts at U and ends at V.

• The jump from Q to R, etc. is called the horizontal retrace, during which the electronic beam in the CRT is blanked.

• The jump from T to U or V to P is called the vertical retrace

NTSC (National Television System Committee) TV standard is mostly used in North America and Japan

- YIQ color model
- 4:3 aspect ratio (i.e., the ratio of picture width to its height)
- 525 scan lines per frame at 30 frames per second (fps).

- Interlaced scanning, and each frame is divided into two fields, with 262.5 lines/field
    - horizontal sweep frequency is 525x29.97 = 15,734 lines/sec
    - each line is swept out in 1/15,734 = 63.6 us the horizontal retrace takes 10.9 us
    - this leaves 52.7 us for the active line signal during which image data is displayed

Now for digital video

Advantages
- Stored on digital device or in memory
- Faithful duplication in digital domain
    - Good or bad ? "We do not care about the law, only the technology"
- Direct (random) access, • nonlinear video editing achievable as a simple, rather than a complex task
- Ease of manipulation (noise removal, cut and paste, etc.)
- Ease of encryption and better tolerance to channel noise
    - Multimedia communications
- Integration to various multimedia applications

ITU-R digital video specifications

	CCIR 601 525/60 NTSC	CCIR 601 625/50 PAL/SECAM	CIF	QCIF
Luminance resolution	720 x 480	720 x 576	352 x 288	176 x 144
Chrominance resolution	360 x 480	360 x 576	176 x 144	88 x 72
Colour Subsampling	4:2:2	4:2:2	4:2:0	4:2:0
Fields/sec	60	50	30	30
Interlaced	Yes	Yes	No	No

CCIR-601 old standard, ITU-R-601 new standard adopted by dvd format.

CIF: Common Intermediate Format

- Specified by CCITT (Comité Consultatif International Téléphonique et Télégraphique).

A format for lower bitrate

- CIF is about the same as VHS quality.
- Progressive (non-interlaced) scan.

QCIF: “Quarter-CIF"

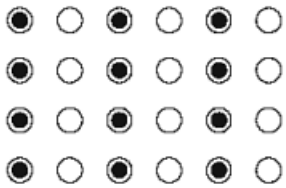
CIF/QCIF resolutions are evenly divisible by 8, and all except 88 are divisible by 16; this provides convenience for block-based video coding in H.261 and H.263, discussed later

Since humans see color with much less spatial resolution than they see black and white, it makes sense to subsample chrominance signal

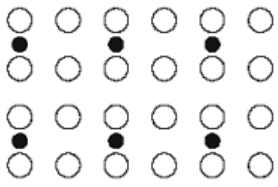
The chroma subsampling scheme 4:4:4 indicates that no chroma subsampling is used: each pixel's Y, Cb and Cr values are transmitted, 4 for each of Y, Cb, Cr.

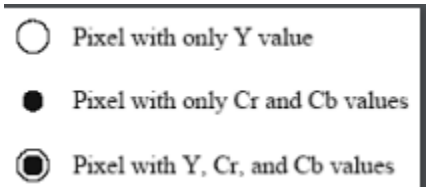
4:2:2: *horizontal subsampling* of the Cb, Cr signals by a factor of 2.

- of four pixels horizontally labelled as 0 to 3, all four Ys are sent, and every two Cb's and two Cr's are sent, as (Cb0, Y0)(Cr0, Y1)(Cb2, Y2)(Cr2, Y3)(Cb4, Y4), and so on (or averaging is used).



- 4:2:0: subsamples in *both the horizontal and vertical* dimensions by a factor of 2.
  - an average chroma pixel is positioned between the rows and columns.





Analog Video Display Interfaces

Component video: three separate video signals for the red, green, and blue image planes. Each color channel is sent as a separate video signal.

- For higher-end video systems
- Supported by most computer systems
- Best color reproduction
  - • no “crosstalk“ between the three channels.

But more bandwidth and good synchronization

Composite video: color (“chrominance") and brightness (“luminance") signals are mixed into a single wire

- Chrominance (I and Q, or U and V).
  - • Combined into a chroma signal, and then put at the high-frequency end of the signal shared with the luminance signal Y.
- Chrominance and luminance components are separated at the receiver end and then two color components are further recovered.
- Only one wire for video signal
  - • Audio signals added through separate wires

Interference is inevitable.

We also have S-Video (page 21)

DVI

Uncompressed digital video

Almost a ubiquitous computer display link replacing VGA (since 1999)

Uncompressed video only

- R, G, B (both digital and analog)
- plus clock, syn, power, control etc.

Single link: 1920x1080 60Hz

Dual link: 2560x1600 60Hz

HDMI

(2002) backward compatible with DVI

RGB or YCbCr + digital audio

+ bidirectional audio, ethernet
High bandwidth digital content protection (HDCP)
HDMI 1.3 2560x1600

Display port
(2006) Packetized transmission (like Internet)
4K video support
Royalty-free (HDMI is not!) ‘

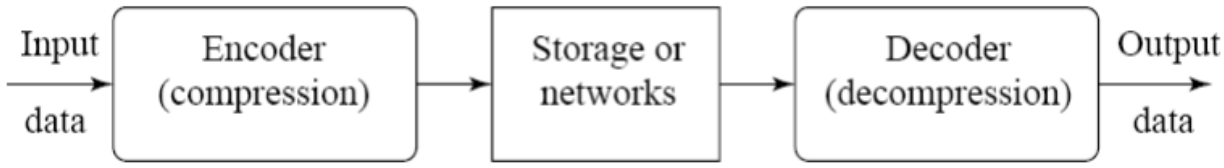
HDTV

Not necessary to increase “definition” in each unit area but increase visual field especially in it’s width.
MUSE (MUltiple sub-Nyquist Sampling Encoding)
an improved NHK HDTV with hybrid analog/digital technologies in the 1990s.
1,125 scan lines, interlaced (60 fields per second), and 16:9 aspect ratio.
Need for compressions
uncompressed HDTV will easily demand more than 20 MHz bandwidth, which will not work in the current 6 MHz or 8 MHz channels
high quality HDTV signals would be transmitted using more than one channel even after compression.

For video, MPEG-2 is chosen as the compression standard.
For audio, AC-3 is the standard
supports 5.1 channel Dolby surround sound -- 5 surround channels plus a subwoofer channel
Difference between conventional TV and HDTV:
Much wider aspect ratio of 16:9 instead of 4:3.
All progressive (non-interlaced) scan
• interlacing introduces serrated edges to moving objects and flickers along horizontal edges
• Upsampling ? 120 Hz, 200 Hz ?
In 1993, the FCC made the decision to go all digital. The FCC (Federal Communications Commission) has planned to replace all analog broadcast services

with digital TV broadcasting by the year 2006. later delayed to June 12,2009 in US Canada: August 31, 2011 (one year extension for some CBC transmitters)

Lossless compression



- **Compression ratio:**

$$compression\ ratio = \frac{B_0}{B_1}$$

B_0 – number of bits before compression

B_1 – number of bits after compression

We can express redundant information with run-length coding:
000 → 30 (three zeroes)
111 → 31 (three ones)

Information is related to probability
Information is a measure of uncertainty (or “surprise”)

Self-information: information of the event itself

$$i(A) = \log_b \frac{1}{P(A)} = -\log_b P(A)$$

if b = 2, unit of information is bits

Entropy: Average self information of data set
a data source generates output sequence from a set {A1, A2, …, AN}
P(Ai): Probability of Ai

$$H = \sum_i -P(A_i)\log_2 P(A_i)$$

The first-order entropy represents the minimal number of bits needed to losslessly represent one output of the source.

Bring 1 paper double sided cheat sheet, scientific calculator no graphing calculator.

Practice exercises:

Dithering: Dither matrix 2x2 matrix ((1 0), (3 2)). Image 2x2 matrix ((0 2), (3 2))

For 1,1 of image, 0 is nothing so ((0 0), (0 0))
For 1,2 of image, 2 want to print values lower so ((1 1), (0 0))
For 2,1 of image, 3 so ((1 1), (0 1))
For 2,2 of image, 2 ((1 1), (0 0))

Final image is 4x4 matrix

Ordered Dithering: Dither matrix 2x2 matrix ((1 0), (3 2)). Image 2x2 matrix ((0 2), (3 2))

For 1,1 of image, 0 compare element wise 0
For 1,2 of image, 2 so 1
For 2,1 of image, 3 so 0
For 2,2 of image, 2 so 0

For larger image, ordered dithering works better

Suppose a sound from source A is 80dB and another sound from source B is 40 dB.

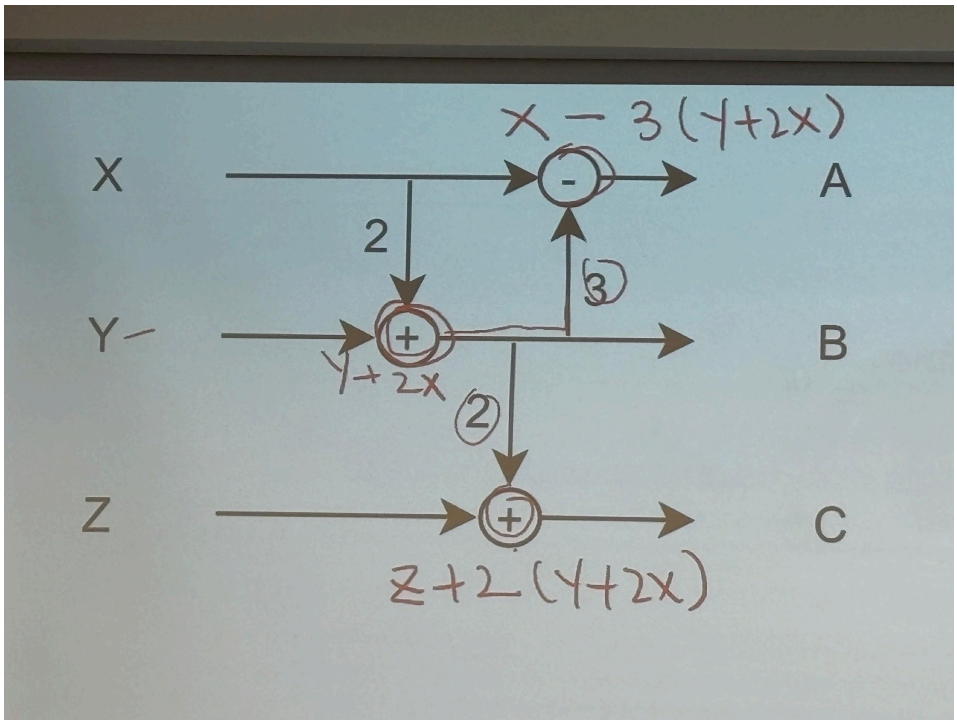
How much is the source A sound larger than the source B sound in terms of power? In terms of voltage?

Use snr equation

$10\log_{10}(r_A) = 80$ $\quad$ $10\log_{10}(r_B) = 40$
$\log_{10}(r_A) = 8$ $\quad$ $\log_{10}(r_B) = 4$
$r_A = 10\text{^}8$ $\quad$ $r_B = 10\text{^}4$

$r_A = P_A/P_{noise}$ $\quad$ $r_B = P_B/P_{noise}$
$r_A/r_B = (P_A/P_{noise})*(P_{noise}/P_B)$
$= (P_A/P_B)$
$= 10\text{^}8/10\text{^}4 = 10\text{^}4$

Since V^2 proportional to P. V = sqrt(P)
$\Rightarrow$ sqrt(10^4) = 10^2.



We can convert to matrix form with matrix times (x y z) vector

Consider a 6 bit gray level of value 23. If gamma value is 2, what is the output after gamma correction.

Need to normalize. 2^6 = 64, 0..63. Also need to denormalize

max value*((value/max value)^1/gamma) ⇒ 63*(23/63)^1/2

Consider a 6-bit gray level of value 23. If output after gamma correction is 58, what is the value of gamma?

63*(23/63)^1/gamma = 58

Assume one violin can generate a sound of 55dB

a) What is the sound level (in dB) generated by 3 such violins?
b) If mic output when recording 2 violins is 1mV, what is the mic output for 5 violins?

a) $10\log_{10}(r) = 55$
   $10\log_{10}(3r) = 10\log_{10}(3) + 10\log_{10}(r) = 10\log_{10}(3) + 55$
b) 5/2 = 2.5 is power ratio from 2 to 5
   $2.5*P_2 = P_5$ where P_2 is power of 2 violins
   $\text{sqrt}(2.5*P_2) = \text{sqrt}(P_5)$
   $\text{sqrt}(2.5)*V_2 = V_5$
   $\text{sqrt}(2.5)*1 = V_5$