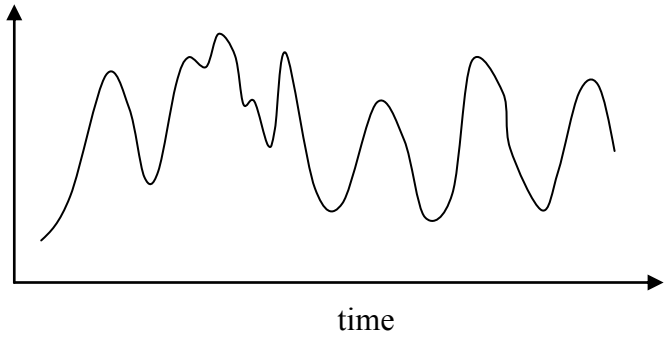
Chapter 3 - Multimedia Data Representation

In this chapter, we will cover basic multimedia data representation for audio, images, and video. These will form the basis of the compression, systems, and networking techniques we will describe in later chapters.

3.1 Sound

What is sound? Sound is pressure variations from vibrating matter. Audio speakers generate sound by moving a baffle (the speaker) in and out to create variations in pressure that our ears can distinguish as sound. Sound is an analog, continuous signal. An example is shown in the figure below.

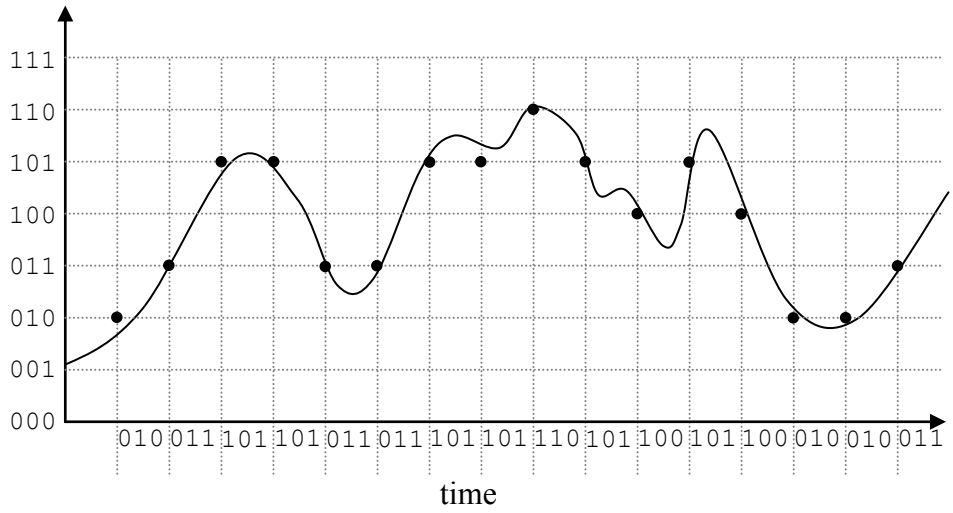


Obviously, for computers the sound needs to be represented somehow in binary format. This conversion is referred to as "A to D" conversion, or A/D conversion. This conversion is sometimes referred to as *pulse code modulation*, or PCM. The samples from this translation are sometimes referred to as *PCM samples.* There are two parameters that control the A/D conversion: the sampling rate and the number of bits used to represent each sample.

The amplitude of the signal is converted into an *n-bit* number. The higher the number of bits in the sample, the more accurately, the signal can be captured without loss. The CD audio standard uses 16-bit samples, allowing for 65536 unique levels to be represented.

The samples per second that are taken impact the frequencies that are represented in the captured signal. Nyquist's limit tells us that in order to capture a signal with maximum frequency, *f*, it is necessary to have 2*f samples* per second in order to capture it accurately. Because humans arecapable of hearing up to approximately 21 kHz, the sampling of the signal must be higher than 42 kHz. Incidentally, the CD audio sampling rate is 44,100 Hz, resulting in a signal that just captures the entire frequency range of the human ear.

As an example of the A/D conversion, we have graphed the sampling of a signal using 3-bits per sample below:



As shown in this example, digitizing audio needs to balance both the sampling frequency and the bit per sample used. If we double the sampling rate, then we need to halve the number of bits used per sample in order to keep the same bit rate.

Representing Sound

Sound is represented digitally by its sampling depth (bits used), sampling rate, and number of channels.

For sampling depth, the number of bits per sample determines the accuracy of the sample. The more bits used, the more accurate the samples. In the example above, we see that in the first sample, the actual signal is somewhere between 001 and 010. The A/D converter then has to choose the closest value to represent the sound. Increasing the representation from 3-4 bits will halve the distance between samples, decreasing the maximum error in half as well. The only other parameter for sampling depth is the "spacing" of the sampling. One can divide the amplitude into fixed intervals (as in the example above); this is also referred to as *linear sampling*. Alternatively, one can use a *perceptually uniform* spacing of the samples. In such an approach, the distance between adjacent values is much larger with the larger amplitudes. The distance between adjacent values is much smaller with the smaller amplitudes. This logarithmic spacing of samples corresponds more closely to the power in the signal. To understand this better, consider a standard stereo system. As the volume knob is changed upward, small changes initially make the sound much louder. As the sound gets louder, turning the knob further requires much more power. Thus, perceptually uniform spacing of samples corresponds more closely to the human's ability to distinguish between loudness of signal.

For the sampling frequency, samples are taken evenly spaced over time. As previously mentioned, the sampling frequency needs to be twice the largest frequency that needs to be represented. While it is beyond the scope of this text, frequencies outside of this range typically need to be run through a high (or low) pass filter in order to avoid aliasing of those frequencies not being represented.

The number of channels represented in the signal allows the system to represent the "placement" of sound. Stereo (or two channel) signals allow the system to represent sound

coming from the left and right. More recently surround sound systems allow multiple channels to be played in front of, behind, and to the left and right of the listener. Typically, the channels are interleaved. That is, for stereo audio at a given sampling depth and sampling frequency, a left sample appears followed by the corresponding right sample. This is then followed by the next left and right samples, and so on.

Basic Stereo Representations

To represent sound, a number of basic approaches have been used. These include u-law, CD-audio, and DVD-audio. We will discuss more advanced compression techniques for audio (such as MP3 technology in Chapter 4).

The u-law representation, commonly found in the telephony network, consists of 8-bits per sample, sampled at 8 kHz with only one channel (mono). The amplitude values (sample values) are spaced in a perceptually uniform way.

The CD-audio representation uses 16-bits per sample, sampled at 44,100 Hz with two channels (stereo).

The DVD-audio specification allows for a number of representations. The format allows for 6 channels with each channel being sampled at 44.1 kHz, 48 kHz, or 96 kHz. Each PCM sample can be 16, 20, or 24-bits in size. Furthermore, there is an option for 2-channel, 24-bit per sample audio using a sampling rate 192 kHz, well beyond anything humans can hear.

3.2 Image Representation

Digital images are represented as a matrix (2 dimensional array) of numerical values that represent quantized (digitized) intensity values. Each entry in the matrix represents a *pixel* of value(s). Each pixel can represent a grayscale value or can represent a tuple of values (e.g. red, green, blue). The key properties for digital images are similar to audio.

- *Resolution* is essentially the sampling frequency in two dimensions. Resolution is typically represented as pixels per inch or total pixels.
- *Quantization* is essentially the sampling depth for each pixel. The number of levels represents the intensity values of a particular pixel. 8 to 12-bit samples per channel are fairly common in digital imaging.
- *Channels* represent the actual values that are being used to hold the image data. For grayscale images, there is typically one-channel of 8 to 12-bits per sample. Color images are typically represented with three channels (red, green, blue) or some equivalent representation such as YIQ or YUV.

Representing Color in Image Data

Color spaces refer to the method in which colors are actually represented. The *red, green, blue* (RGB) color space is perhaps the most common of the color spaces being used. For each color a separate intensity value is stored. The RGB color space is an additive color space, in that, a mixture of all 0 values for *R, G,* and *B* yields black, while a mixture of the maximum intensity values for *R, G,* and *B* yields white. RGB color spaces are used for light-emitting devices such as CRT screens and projection TVs.

The *cyan, magenta, yellow* (CMY) color space is known as a *subtractive* color space. That is, a mixture of all 0 values for cyan, magenta, and yellow yields white, while maximum intensity cyan, magenta, and yellow yields black. In order to generate color, cyan, magenta, and yellow intensity filters are used to remove color (i.e., subtracting color through the filter). The CMY color space is typically used for printers. Theoretically:

- Cyan absorbs red, while reflecting (or passing through) green and blue
- Magenta absorbs green, while reflecting (or passing through) red and blue
- Yellow absorbs blue, while reflecting (or passing through) red and green

Thus, in order to generate red, magenta and yellow are used to remove the green and blue in the system leaving only red.

The *cyan, magenta, yellow, black* (CMYK) color space extends the CMY color space to have an explicit channel for black. CMYK is pretty much used only for printing. Theoretically CMY at full intensity should yield black, but in most printing systems this mixture yields a dark brown; the reason for this is that the pigment printed on the paper is not truly transparent nor is the paper a light source. In addition, using three colors at full intensity to represent black (which is perhaps the most printed color) yields significant use of the color ink / dye.

The YIQ[1] and YUV[2] color spaces are primarily used in television broadcasting systems. The YIQ was a color space that was originally used in NTSC broadcasts. It is now obsolete and most television broadcasting systems use YUV. The importance of these signals with regard to video will be discussed later. For now, the important concept is that the Y channel for a pixel represents the brightness (grayscale) value. The UV channels add the appropriate color to image; thus, a YUV image can be transmitted and displayed in color and grayscale with very little processing. The relation between the YUV and RGB color spaces for a given pixel can be described by the following transformation:

$$\begin{bmatrix} y \\ u \\ v \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

Representing Image Data

There are a number of ways to represent the data within digital images. For the purposes of our discussion, we will limit ourselves to red, green, blue (RGB) images. For the most part, substituting any of the above color spaces in this discussion is fine. To represent the actual pixel data a number of techniques can be used.

- Each pixel can have a red, green, and blue intensity value associated with it. Thus, the image is represented as an array of pixel values, each with a red, green, blue sample tuples stored with it.
- Each pixel can have a red, green, and blue pointer to a table of red colors, green, colors, and blue colors, respectively. Here, each pixel will have the index of the

[1] Y represents luminance / brightness, I is intermodulation, and Q is qudrature
[2] Y represents luminance / brightness, U/V represent chrominance / color

appropriate color value in the color table. The color table can be sorted by color intensity value or by some other metric such as most frequently used.

- Each pixel can have a single pointer that points to a unique red, green, blue tuple stored in a color table / palette. Rather than having three unique color palettes, this approach has a single color palette that represents the entire color for the pixel.

The main purpose of the latter two representations is that they may allow some additional compression to be had. For example, making the indices clustered around the colors that occur more frequently and assigning shorter codewords to them.

There are a number of image formats and compression technologies. We will describe the algorithms that they use in the next chapter.

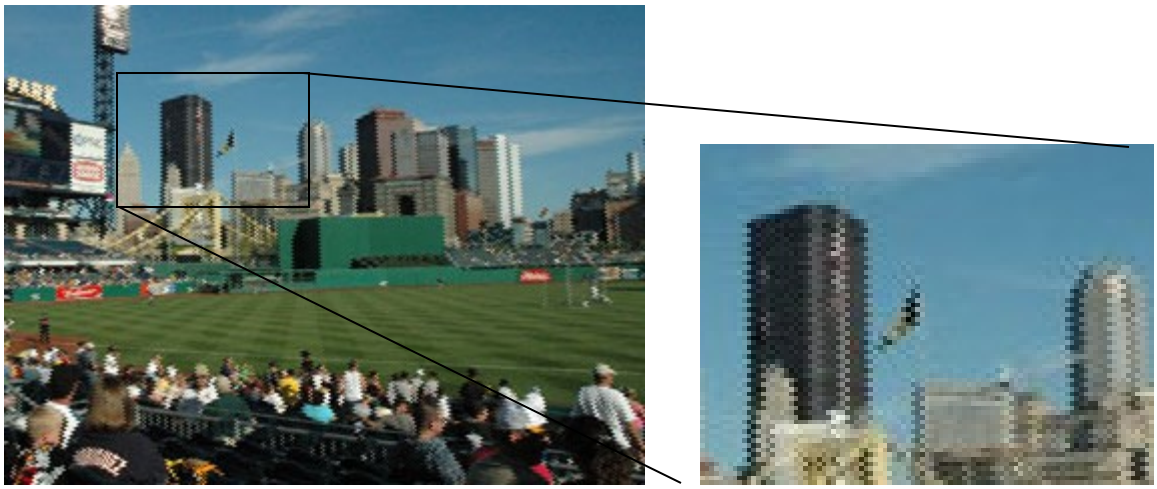
3.3 Video Representation

Video can essentially be thought of as a set of images over time. As a result, many of the ideas from digital images can be applied here. One of the key differences is that video as we know it has been steadily transitioning from an analog representation for broadcast television to one that is completely digital. This has ramifications for how the digital representation is managed. As we will see, there are some legacy issues to deal with the transition from analog video to the new digital video formats.

3.3.1 Analog Video

There are several analog broadcast television standards that exist and are continuing to evolve. In the U.S., the main television broadcast standard is the National Television System Committee (NTSC) format which was defined in 1953. The NTSC standard has 525 horizontal scan lines (analog) of which 486 are typically visible. The remaining scanlines are used for the sync pulse, which moves the electron gun from the bottom of the screen back up to the top. Because of CRT technology, the screen alternates between refreshing even and odd lines of the picture, displaying 59.94 half frames (or fields) per second. Thus, the NTSC standard delivers approximately 30 full frames per second to the user. The Phase Alternation by Line (PAL) standard is used is most of Europe except France. PAL has a higher resolution than NTSC with 625 lines of which 576 are visible. PAL uses 50 interlaced fields per second. The Systeme Electronique Couleur Avec Memoire (SECAM) standard is similar in resolution to PAL. The main difference is in the handling of the color channels. SECAM is primarily used in France.

While interlacing provided a mechanism for the broadcast of television on older CRT technologies, they make the bridge to digital technologies more difficult. In particular, motion can cause significant visual artifacts to appear when one pair of interlaced frames are combined together for systems (e.g., in modern progressive display screens). As an example, consider a video camera that is panning from left to right. The time between the even and odd lines being displayed is approximately 16 milliseconds. If the camera is moving, the position of the objects across the two fields will be in different places. As a result, this creates jagged edges when combined. An example of this is shown below.



3.3.2 Digital Video

Each *frame* of digital video is represented by a matrix of pixels, just like their digital image counterparts. In general, the key parameters for video include resolution, frame rate, quantization, and number of channels.

Resolution is typically reported in pixels (e.g. 640x480 pixel video). There is great variation in the amount of pixels used for video. Several common resolutions include: 320x240 pixel, 352x288 pixel, 640x480 pixel, 720x480 pixel, 1280x720 pixel, and 1920x1080 pixel video. Some of these are built around legacy analog systems. For example, in converting analog video into a digital counterpart, the natural video resolution is approximately 480 pixels or lines. This is because there are approximately 480 lines in viewable NTSC video signals. As a result of this legacy engineering, the DVD format and DV video camera format also use 480 pixels. In order to keep the 4:3 width to height ratio of NTSC video the analog scanlines are converted into approximately 640 or 720 pixels.

Related to resolution is the video's *aspect ratio*. The aspect ratio is the ratio of width to height for the video. For NTSC, the aspect ratio is 4:3. The newer HDTV standard and "widescreen" formats are 16:9.

Frame rate is the sampling rate in the temporal domain. Humans perceive continuous motion at approximately 15 frames per second, while smooth video motion is generally acceptable at 30 frames per second. Most motion pictures are shot at 24 progressive (full) frames per second.

Quantization is essentially the sampling depth for each pixel. The number of levels represents the intensity values of a particular pixel. Almost all color video systems use 24-bits per pixel, 8-bits each for red, green, and blue.

Channels represent the actual values that are being used to hold the image data. For video data, this is RGB or YUV.

3.3.3 Video Representation

The representation of video is very similar to that of digital images. Typically, digital video is represented in YUV format. There are a number of reasons for this. First, the luminance channel (Y) represents the overall brightness of a pixel. Using only the Y pixels of a particular frame yields a grayscale image. This was important as it allowed the television signal to be handled by "black and white" televisions as well as color televisions. Second, because the human eye is less sensitive to changes in the U and V channels compared to the luminance channel, subsampling can be used to save the storage space required to represent the video frame. We will discuss this further in the video compression chapter.

3.4 Chapter Summary

In this chapter, we have discussed various representation schemes for audio, image, and video data. Audio is characterized by its sampling depth (bits per sample) and sampling frequency (Hz). Images are characterized by their resolution (size or pixels per unit), quantization (bit sample depth), and the number of channels. Video is characterized by its resolution, frame rate (temporal frequency), quantization, and the number of channels.

3.5 Problems

1. A typical CD-audio disc holds approximately 74 minutes of PCM audio data. MP3's can be encoded at a number of bit-rates. For 128 kbps MP3 streams, how many minutes of audio can be stored on the same disc? What is the compression ratio for the MP3 stream?

2. Suppose we have the ideal stereo audio signal of 16-bit samples and a sampling rate of 44,100 Hz. Further, suppose we have a process that converts this data into the u-law format. What is the effective compression ratio of this process?

3. The hearing range of cats is approximately 20Hz to 60,000 Hz. Suppose we want to make a CD for our cat Garfield and want to have a high fidelity signal, how much audio can be stored on a single CD audio disc for our cat.

4. How does one generate green using CMY filters?

5. How does one generate blue using CMY filters?

6. Suppose we have an 8 x 11 inch document that we want to scan using 300 dpi. How large is the resulting image, assuming 24-bits per pixel? How long would the sample take to transfer over USB1 (approximately 10 megabits per second) and USB2 (approximately 450 megabits per second)?

7. Suppose we represent an image using a separate red, green, and blue palette for the colors. Further, suppose we only represent 32 unique hues of red, 64 unique hues of green, and 256 unique hues of blue. Assuming the palette table is well known (i.e.

specified in a standards document, what is the effective compression ratio of this process?

8. Estimate the bandwidth (bits per second) required to support uncompressed 720x480 pixel, full-color video over a network, assuming a frame rate of 30 fps?

9. USB1 bandwidth is approximately 11 megabits per second. Suppose we have a camera that we have attached to our computer that is capable of capturing 640x480 pixel video at 30 fps. Further, suppose we would like to capture the highest frame rate video (30 fps). What is the maximum sized 4:3 aspect ratio video that can be captured over the USB channel?

10. For problem 9, suppose we would like to capture the highest quality video (640x480). What is the maximum frame rate achievable over the USB channel? What is the answer for 320x240 video?

11. For problem 9, suppose we have a compression algorithm running between the web camera and the host computer. What compression ratio must be achieved in order to support the capture of 640x480 video at 30 fps over the USB channel?