ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

Multimedia Databases: Fundamentals, Retrieval Techniques, and Applications

A Short Course for Doctoral Students
University of Bologna

Multimedia Data and Content Representations

Ilaria Bartolini - DEIS

Bologna, September 7-10, 2010

Outline

- Multimedia (MM) data and applications
- MM data coding
- MM data content representation

Media (or medium)

- A way to distribute and represent information such as books, newspapers, music, radio news, TV news, etc.
  - E.g.: **text**, **graphics**, **images**, **voice**, **sound**, **music**, **animation**, **video**, etc.

text

sound

image

graphic

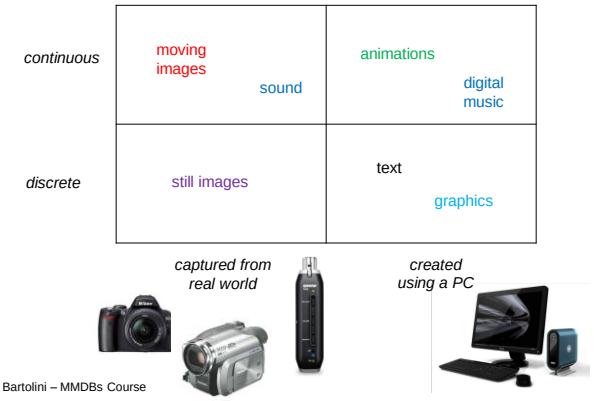
video

animation

Media description

- Perception
  - auditory media (voice, audio, music)
  - visual media (text, graphics, images, moving images)
- Representation
  - ASCII (text), JPEG (images), MP3 (audio), etc.
- Presentation
  - input: keyboard, mouse, digital camera, scanner
  - output: paper, monitor, printer, speaker
- Storage
  - disks (floppy, hard, optical), magnetic tapes, CD-ROM, DVD-ROM
- Transmission
  - coaxial cable, optical fiber, satellite
- Information exchange
  - CD, JAZ-Drives, optical fiber

Media types (1)



Media types (2)

- Represented in term of the *dimensions of the space* the data are in:
  - *0-dimensional data*: this type of data is the regular, alphanumeric data (e.g., text)
  - *1-dimensional data*: this type of data has one dimension (i.e., *time*) of the space imposed into them (e.g., audio)
  - *2-dimensional data*: this type of data has two dimensions (i.e., *x*, *y*) of the space imposed into them (e.g., images and graphics)
  - *3-dimensional data*: this type of data has tree dimensions (i.e., *x*, *y*, and *time*) of a space imposed into them (e.g., video and animation)

Multimedia data

- A combination of a number of media objects (i.e., text, graphics, sound, animation, video, etc.) that must be presented in a coherent, synchronized manner
  - It must contains at least a discrete and a continuous media
- Multimedia system or application
  - A system/application that uses both discrete and continuous media

Application domains (1)

- An effective and efficient management of MM data is required in a variety of application domains, including
- "General purpose" applications
  - E-commerce (where electronic catalogues have to be browsed and/or searched)
  - Digital libraries (text, images, audio interviews)
  - Edu-tainment (for example, to search in clipart repositories, or to search and organize personal photo albums in mobile phones or PDAs)
  - On line and print advertising
  - Personal and public photo/media collections
  - (semi-)automatic media object annotation techniques (which can be based on assigning to a unlabelled object the keywords associated to the objects most similar to a given one)
  - Media object classification (for example, to search for similar logo images for copyright infringement issues and for the detection of pornography images)

Application domains (2)

- "specific" applications
  - Medical DBs (ECG's, X-rays, Magnetic Resonance Images (MRI))
  - Biometric systems (fingerprints, faces, handwriting)
  - Molecular DBs (DNA sequences, proteins)
  - Scientific DBs (sensor data, e.g., traffic control, surveillance)
  - Financial DBs (stock prices)

Multimedia standards (1)

- Various standards are available to facilitate authoring of complex MM objects or documents
  - *SGML/XML*: *Standard Generalized Markup Language*, standard ISO (1986) for describing the structure of documents
    - Separation of document content and structure from the presentation of the document; document structure defined using *Document Type Definition* (DTDs) based on a formal grammar
    - One of the most notable applications of SGML standard is the *HyperText Markup Language* (HTML), current standard for publishing on the Internet (dates back to 1992)
    - *Extensible Markup Language* (XML) has been developed by the W3C as a follow-up of SGML
      - Especially suitable for creating interchangeable, structured Web documents
  - *HyTime*: *Hypermedia/Time-based Structuring Language*, an international multimedia standard ISO based on SGML
    - aims to describe not only the hierarchical and link structures of multimedia documents, but also temporal synchronization between objects to be represented to the user as part of the document

Multimedia standards (2)

- *SMIL*: *Synchronized Multimedia Integration Language*, synchronization standard developed by W3C based on XML
  - like HyTime, defines a language for interactive multimedia presentations
- *MPEG7* and *MPEG21*: unlike standard just mentioned, which aim to describe the content of authored documents, the main focus of MPEG7 (*Multimedia Content Description Interface*) is to describe the content of captured media objects, such as video
  - Follow-up of the previous MPEG standards MPEG1, MPEG2, and MPEG4
    - mainly concerned with audio/video compression
  - Includes content-based description mechanisms for images, graphics, 3D objects, audio, and video stream
  - Low-level visual descriptors for media include color, texture, shape, and motion
  - The standard also enables description of how to combine heterogeneous media content into one unified multimedia complex object
  - The follow-up standard MPEG21 aims to provide additional content management and usage services, such as caching, archiving, distribution, and intellectual property management for multimedia objects

Managing MM data

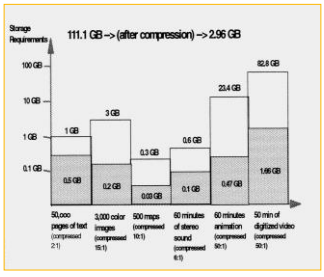
- There are several issues concerning the "management" of MM data (due to their complex and heterogeneous nature), such as:
  - **Representation**: formats, compression (e.g., JPEG, MPEG, WAV)
  - **Storage**: physical layout on disk (e.g., BLOB)
  - **Search and retrieval**
  - **Generation**, **acquisition**, **transmission**, **delivery**
- Although "*multimedia*" refers to the multiple modalities and/or multiple media types of data, conventionally each medium is studied separately, (from the *representation*, *searching*, and *indexing* points of view)
  - the features used for media-based retrieval are specific to each media type (e.g., image, audio, and video)
- *In this course we concentrate on aspects related to*
  - *representation of specific media types:*
    - *images*
    - *audios*
    - *videos*
  - *search/retrieval of generic MM objects*

MM data coding

- For a personal computer (PC) handling MM data requires a transformation process that *digitize* or *discretize* the original information to the digital representations known to the PC as *data*
  - *e.g., an image can be represented as a set of binary numbers for each byte in the original representation*
- MM data require a *vast amount of data for their representation*
- 3 main reasons for compression
  - Large storage requirement
  - Slow devices which do not allow playing back uncompressed MM data (especially video) in real time
  - Network bandwidth (not allow real-time video data transmission)
- Compression techniques are classified in two basic categories:
  - *Lossless* (e.g., *Huffman* coding)
    - capable to recover the original representation perfectly
  - *Lossy* (e.g., quantization, DCT)
    - recover the presentation to be similar to the original one
  - *Hybrid* (e.g., JPEG, MPEG)

Encyclopedia example

- Storage requirements for the multimedia application encyclopedia:
  - 500,000 pages of text (2 KB per page) - total 1 GB;
  - 3000 color picture (in average 640x480x24 bits = 1MB/picture) - total 3 GB;
  - 500 maps (in average 640x480x16 bits = 0.6 MB/map) - total 0.3 GB;
  - 60 minutes of stereo sound (176 KB/sec) - total 0.6 GB;
  - 30 animations, in average 2 minutes in duration (640x480x16 bits x 16 frames/sec = 6.5 MB/sec) - total 23.4 GB;
  - 50 digitized movies, in average 1 minute in duration (640x480x24 bits x 30 frames/sec = 27.6 MB/sec) – total 82.8 GB.



…for a total of 111.1 GB storage capacity!!

MM content representation (1)

- We can always represent the multimedia data in their original raw formats (e.g., images in their original formats such as JPEG, TIFF, or even the raw matrix representation)
  - considered as awkward representations, and thus are rarely used in a multimedia application for two basic reasons:
    - typically *take much more space than necessary*
      - more processing time and more storage space
    - *such formats are designed for best archiving the data*
      - e.g., for minimally losing the integrity of the data while at the same time for best saving the storage space
    - *…but not for fulfilling the MM research purpose*, i.e., to represent the MM data as useful information that would facilitate different processing and mining operations, having knowledge on the "what the data is", that is its semantic knowledge

MM content representation (2)

- Example:

Original format: JPEG

Actual content: binary numbers for each byte in the original representation

…but this does not tell anything about what this image is!!!

Ideally semantic representation

- 3 hierarchical levels of MM content representation:
  - High-level: semantic knowledge - bridge the semantic gap by integrating high level concepts (sites, objects, events) and low-level visual/audio features
  - Mid-level: text annotations/attributes (e.g., "JPEG", "bear", "grass", …)
  - Low-level: low level visual/audio features (color, texture, shape and structure, layout; motion; audio - pitch, energy, etc.)
- Instead of representing MM data in term of semantic knowledge (ideally representation), we first represent MM data as features

Categories of features

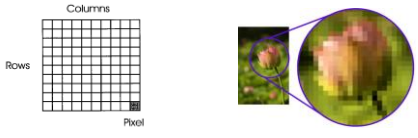
- 3 categories of features: *statistical*, *geometric*, *meta features*
- Except for some meta features, most of the feature representation methods are applied to a *unit of MM data* instead of the whole MM data
  - e.g., for an image collection a unit is an image, for an audio stream, a unit is an audio frame, and for a video is a video frame
- Statistical features: focus on statistical description of the original MM data in term of specific aspects, such as the *frequency counts* for each of the values of a specific quantity of data
  - e.g., histograms, transformation coefficients
- Geometric features: applied to segmented objects within a MM data unit
  - e.g., moments, Fourier descriptors
- Meta features: include the typical meta data to describe a MM data unit
  - e.g., scale of the unit, number of objects in the data unit

One image is worth 1,000 words…

- Undoubtedly, images are the most wide-spread MM data type, second only to text data
- Their representation is far more complex than the text one and needs more storage resources
- In the following we provide details on
  - physical image representations
  - image formats (e.g., BMP, GIF, JPEG, TIFF, …)
  - some basic features, such as color, texture, and shape and structure
    - considering *general purpose* images, i.e., no assumptions on the working domain
    - global features (related to the whole image)
    - local features (related to specific objects within the image)

Image representation (1)

- Physically speaking a digital image represents a 2-D array of samples, where each sample is called pixel



- The word pixel is derived from the two words “picture” and “element” and refers to the smallest element in an image
- Color depth is the number of bits used to represent the color of a single pixel in a bitmapped image or video frame buffer (also known *as bits per pixel – bpp*)
  - Higher color depth gives a broader range of distinct colors

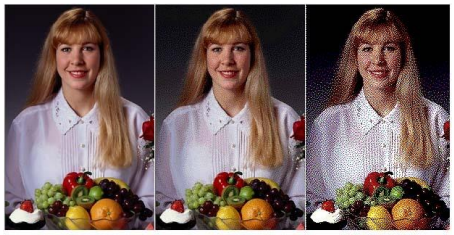
Image representation (2)

- According to the color depth, images can be classified into:
  - Binary images: 1 bpp (2 colors), e.g, black white photographic
  - Computer graphics: 4 bpp (16 colors), e.g., icon
  - Grayscale images: 8 bpp (256 colors)
  - Color images: 16 bpp, 24 bpp or more, e.g., color photography
- The table shows the color depths used in PCs today:

Color depth	# displayed colors	Bytes of storage per pixel	Common name
4-bit	16	0.5	Standard VGA
8-bit	256	1.0	256-Color Mode
16-bit	65.536	2.0	True Color
24-bit	16.777.216	3.0	High Color

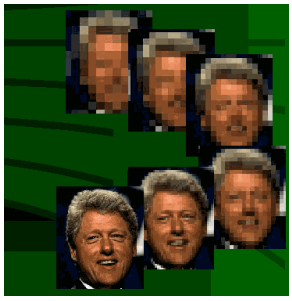
- Dimension is the number of pixels in an image; identified by the *width* and *height* of the image as well as the *total number of pixels* in the image (e.g., an image 2048 wide and 1536 high (2048 x 1536) contains 3,145,728 pixels - 3.1 Mp)
- Spatial resolution is the *number of pixels per inch – bpi*; the higher the bpi, the better the resolution (clarity) of the image. Resolution changes according to the size at which the image is being reproduced
- Size [Byte] = *(width * high) * color depth/8*

Color depth



Spatial resolution

Example: these images of Former President Clinton demonstrate the effects of different spatial resolutions. Each higher level of resolution allows you to distinguish more detail

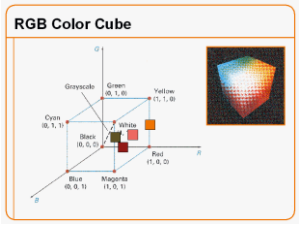


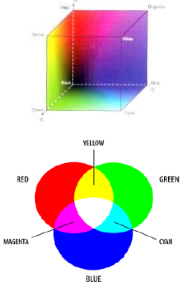
Color

- According to the tri-chromatic theory, the sensation of color is due to the stimulation of 3 different types of receptors (*cones*) in the eyes
- Consequently, *each color can be obtained as the combination of 3 component values* (one per receptor type)
- A *color space* defines 3 color channels and how values from such channels have to be combined in order to obtain a given color
- There is a large variety of color spaces (e.g, RGB, CMY, HSV, HSI, HLS, Lab), each designed for specific purposes, such as displaying (RGB), printing (CMY), compression (YIQ), recognition (HSV), etc.
- It is important to understand that *a certain "distance" value in a color space does not directly correspond to an equal difference in colors' perception*
  - E.g., distance in the RGB space badly matches human's perception

Color spaces: RGB

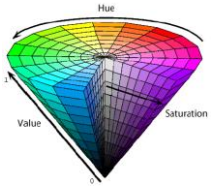
- The RGB space is a 3-D cube with coordinates Red,Green, and Blue
- The line of equation R=G=B corresponds to gray levels
- It can represent only a small range of potentially perceivable colors

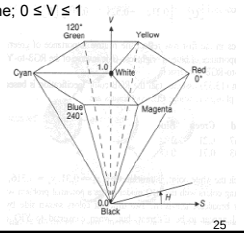




Color spaces: HSV

- The HSV space is a 3-D cone with coordinates Hue,Saturation, and Value:
- Hue is the "color", as described by a wavelength
  - Hue is the angle around the circle or the regular hexagon; $0 \le H \le 360$
- Saturation is the amount of color that is present (e.g., red vs. pink)
  - Saturation is the distance from the center; $0 \le S \le 1$
    - The axis S = 0 corresponds to gray levels
- Value is the amount of light (intensity, brightness)
  - Value is the position along the axis of the cone; $0 \le V \le 1$





Saturation of colors

Original image | Saturation decreased by 20% | Saturation increased by 40%

What the 3 channels represent

- The figure contrasts the information carried out by each channel of the RGB and HSI color spaces
  - HSI: similar to HSV, the color space is a "bi-cone"

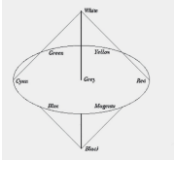


Figure 56. Color separations from a real-world color image of flowers:
a) original;
b) red component;
c) green component;
d) blue component;
e) hue component;
f) intensity component;
g) saturation component.

BMP format

- Bitmap format encodes images without compression:

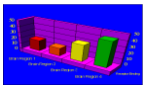
$$size = (number\ of\ pixels * bpp)$$

  - Example:
    a BMP image 640x480 (= 307200 pixels) with color depth 24 bpp
    has a size of = 307200 * 24 / 8 = 921600 bytes = ~0.9 MB

- The most important compressed formats are:
  1. GIF (*Graphics Interchange Format*)
  2. PNG (*Portable Network Graphics*)
  3. JPG (*Joint Photographer Expert Group*)
  4. TIFF (*Tagged Image File Format*)

GIF format

- GIF (*Graphics Interchange Format*)
- Introduced by *CompuServe* in 1987 is *one of the most used and supported format*
  - *8 bpp* image format, i.e., the color palette is limited to a maximum of 256 colors from the 24-bit RGB color space
- GIF images are compressed using the *Lempel-Ziv-Welch (LZW) lossless* data compression technique to reduce the file size without degrading the visual quality
- It also supports *animations* and allows a separate palette of 256 colors for each frame
- The color limitation makes the GIF format *unsuitable for reproducing color photographs* and other images with continuous color, but it is *well-suited for simpler images such as graphics or logos with solid areas of color*



PNG format

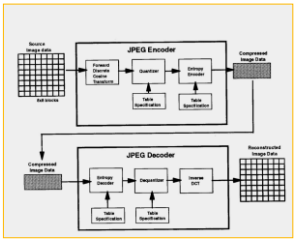
- PNG (*Portable Network Graphics*) was created to improve upon and replace GIF
  - It is pronounced “ping”, or “pee-en-jee”. The PNG acronym is optionally recursive, unofficially standing for PNG’s Not GIF!! ;-)
  - PNG supports palette-based (palettes of RGB 24-bit or RGB 32-bit colors), and grey-scale images
  - PNG was designed for transferring images on the Internet, not for print graphics
  - Better compression than GIF
  - PNG does not support animation like GIF does

JPEG format

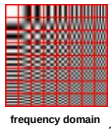
- JPEG (*Joint Photographic Experts Group*), standard issued in 1992 with the aim of improving and replacing previous image formats
- *JPEG images are full-color images* (24-bit, or “true color”), unlike GIFs that are limited to a maximum of 256 colors in an image
  - there is a lot of interest in JPEG images among photographers, artists, graphic designers, … and where color fidelity cannot be compromised
- JPEG can *achieve incredible compression ratios*, squeezing graphics down to as much as 100 times smaller than the original file. This is possible because the JPEG algorithm discards “unnecessary” data as it compresses the image
- There is also *an interlaced “Progressive JPEG” format*, in which data is compressed in multiple passes of progressively higher detail
  - This is *ideal for large images that will be displayed while downloading over a slow connection*, allowing a reasonable preview after receiving only a portion of the data

JPEG compression

- The standard specifies the codec, which defines how an image is compressed into a stream of bytes and decompressed back into an image
- The compression method is usually lossy, meaning that some original image information is lost and cannot be restored (possibly affecting image quality). There is an optional lossless mode defined in the JPEG standard; however, that mode is not widely supported in products
  - *Discrete Cosine Transform* (DCT) - *lossless*
  - *Quantization – lossy*
  - *Entropy coding – lossless*



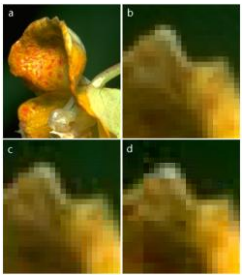
pixel domain



frequency domain

Levels of JPEG compression

- The figure shows an original photograph (a), and three detail views at different levels of JPEG compression:
  - "excellent" quality (b),
  - "good" quality (c), and
  - "poor" quality (d) (notice the *boxy* quality of this image)



Compression ratio

- The basic measure for the performance of a compression algorithm is the compression ratio (C_R):
  - C_R = *(Orig. size/Compressed size)*
- Higher compression ration will produce lower picture quality and vice versa

JPEG 2000 format

- JPEG 2000 is an image compression standard and coding system
- It was created by the *Joint Photographic Experts Group* committee in 2000 with the intention of superseding their original DCT-*based* JPEG standard (created in 1992) with a newly designed *wavelet-based method*
  - Higher compression rate (and implicit information loss) without the "boxy" effect induced by JPEG

TIFF format

- TIFF (*Tagged Image File Format*) is a file format for storing images, popular among *Apple Macintosh* owners, graphic artists, the publishing industry
- As of 2009, it is under the control of *Adobe Systems*
- TIFF is a flexible and adaptable file format:
  - Can handle multiple images and data in a single file through the inclusion of "tags" in the *file header*
    - Tags represent the basic geometry of the image (e.g., the size), or define how the image data is arranged and whether various image compression options are used
- TIFF format is widely supported by *image-manipulation applications, by publishing and page layout applications, by scanning, faxing, word processing, optical character recognition* and other applications

EXIF format

- EXIF (*Exchangeable Image file Format*) is a *specification for the image file format used by digital cameras*
- The specification uses the existing JPEG, TIFF, and WAV file formats, with the *addition* of specific metadata tags
- It is not supported in JPEG 2000, PNG, or GIF
- Used to store *photos parameters*:

Texture

- Unlike color, texture is not a property of the single pixel, rather it is a collective property of a pixel and its, suitably defined, "neighborhood"

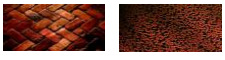
"mosaic" effect

"blinds" effect

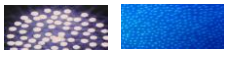
- Intuitively, texture provides information about the uniformity, granularity and regularity of the image surface
- *It is usually computed just considering the gray-scale values of pixels* (i.e., the V channel in HSV)

What texture measures

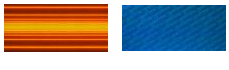
- A common model to define texture is based on the properties of coarseness, contrast e directionality:
  - Coarseness - coarse vs. fine: it provides information about the "granularity" of the pattern



  - Contrast - high vs. low contrast: it measures the amount of local changes in brightness



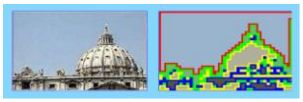
  - Directionality - directional vs. non-directional: it's a global property of the image



Shape

- Strictly speaking, an image has no relevant shape at all ☺
- When we talk about shape, we refer to that of the "object(s)" represented by the image
- Object recognition is a hard task, hardly solvable by any algorithm that operates in a general scenario (i.e., no knowledge about what to look for)
- In practice, *shape information is often obtained by "segmenting" the image into a set of "regions", and then recovering the contours of such regions*
  - …and *segmentation is typically performed by analyzing color and texture information…*

An example of segmentation



- A classical problem with segmentation is the trade-off between homogeneity of a region and number/significance of regions:

  How many regions?

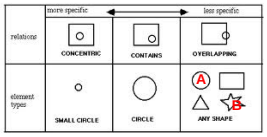
  How "homogeneous" pixels within a same region should be?

  No general answer!

- In the limit cases: a single region(!?), each pixel is a region(!?)

Spatial relations

- Given image objects, we can identify local properties:
  - position;
  - area;
  - perimeter;
  - …
- and/or global properties, such as
  - spatial relations (trough *spatial constraints* definition)
    - To the left, to the right
      - *Object A is to the left of B*
    - Above of, below of
      - *Object A is above object B*

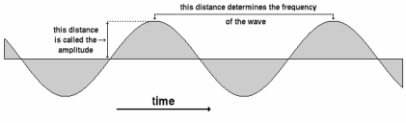


Audio

- Audio data are often viewed as *1-D continuous or discrete signals*
  - Many of the models that are applicable to 2-D images has their counterpart in audio data
- With respect to images, audio maintains temporal information
- In the following we detail on
  - physical audio representations
  - audio formats (e.g., WAV, MP3, MIDI, …)
  - some domain specific audio features, such as *pitch, loudness, beat, rhythm,* etc.

Audio technology (1)

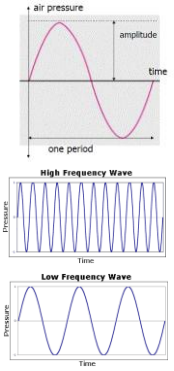
- *Sound is an oscillation of pressure transmitted through a solid, liquid, or gas, composed of frequencies within the range of hearing* and of a level sufficiently strong to be heard, or *the sensation stimulated in organs of hearing by such vibrations*
- Basic sound characteristics:
  - Frequency: *pitch*
  - Amplitude: *loudness*



Audio technology (2)

- Frequency of a wave is the number of cycles per second
  - Corresponds to the *pitch*
  - Measured in *Hertz* (*Hz*)
    - E.g., 1 Hz simply means one cycle per second

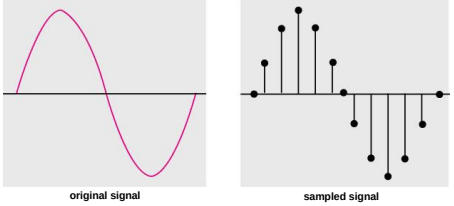
    Infrasonic: 0 ~ 20 Hz
    Audiosonic: 20 Hz ~ 20K Hz (what we hear; e.g., voice: 600 Hz ~ 6K Hz)
    Ultrasonic: 20K Hz ~ 1G Hz
    Hypersonic: 1G Hz ~ 10 THz
- Amplitude of a wave describes the maximum disturbance of a medium in a wave cycle
  - Corresponds to *loudness*
  - Measured in *Decibel* (*dB*)
    - E.g., *20 dB (quite home); 60 dB (conversation), 120 dB (loud rock band), 139 dB (loudest band on the planet: Manowar!!!)*



Audio representation

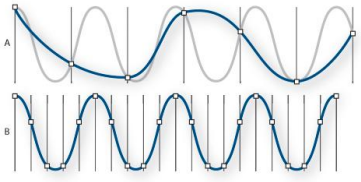
- Analog to digital-sampling theory…
  - *Low-pass filtering*: remove high frequencies information
  - *Sampling*: measure single value
  - *Quantization*: relate value to interval
  - *Encode*: assign binary code
- Important factors:
  - Sampling rate
    - *number of points used to capture the sound wave in 1 second*
    - unit: *Hz*
  - Quantization depth
    - *amount of information used to store the round-off amplitude of each sample*
    - unit: *bits* (usually 8/16 bits)

Sampling



original signal **sampled signal**

Sampling rate

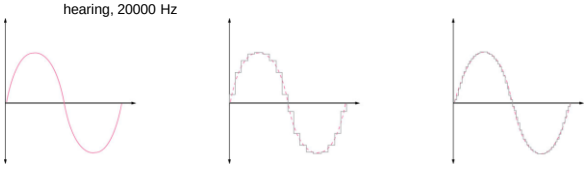


A: low sampling rate distorts the original sound wave

B: high sampling rate perfectly reproduces the original sound wave

Quantization depth

- Each so obtained sample is assigned the amplitude value closest to the original wave's amplitude.
  - *Higher bit depth provides more possible amplitude values*, producing greater dynamic range, a lower noise floor, and higher fidelity
  - To reproduce a given frequency, the *sample rate must be at least twice that frequency* (*Nyquist frequency* theorem)
    - E.g., CDs have a *sample rate of 44100* samples per second, so they can *reproduce frequencies up to 22050 Hz*, which is just beyond the limit of human hearing, 20000 Hz



Why audio compression?

- Till now, the hypothesis was to consider monophonic audio data
  - one audio channel
- Stereophonic sound (or stereo) is the reproduction of sound *using two or more independent audio channels*, through a symmetrical configuration of loudspeakers, in such a way as to create a pleasant and natural impression of sound heard from various directions, as in natural hearing
- Example:

  Let's go back to our CDs example; to store uncompressed CD quality (i.e., 1 sec., bandwidth of 22050 Hz, 16 bit, stereo), we need:
  - *44100 samples * 16 bit * 2 / 8 = 176400 B ~ 172KB (for 1 sec.)*
  - *for a song of 4 min (240 sec.): 40 MB*

- As for images, audio data need to be compressed!!

WAV format

- WAVE or WAV (*Waveform Audio File Format*) is a *Microsoft* and *IBM* audio file format standard for storing an audio bitstream on PCs
- *.WAV* files are the *default uncompressed audio format* on Windows; it is recognized by almost all computer systems
- Not useful as file sharing format over the Internet
  - still a commonly used and suitable for retaining "first generation" archived files of high quality (by professional users or audio experts), for use on a system where disk space is not a constraint, or in applications such as audio editing, where the time involved in compressing and uncompressing data is a concern
- It is limited to *files that are less than 4 GB* in size due to its use of a *32 bit* unsigned integer to record the *file size* header
  - a full pop song in WAVE format may take up to 40 MB of disk space
    - remember our previous example!!
  - …anyway, 4GB are still enough for storing 6.8 hours of CD-quality audio (44.1 kHz, 16-bit, stereo)!!

MP3 format

- The *Moving Picture Experts Group*, commonly referred to as simply MPEG, is a working group charged with the development of video and audio encoding standards
- MPEG has standardized the following compression formats for audio/video:
  - MPEG-1 (1993) - *Coding of moving pictures and associated audio*
  - MPEG-2 (1995) - *Coding of moving pictures and associated audio*
  - MPEG-4 (1998) - *Coding standard for audio and video*
  - MPEG-7 (2002) - *Multimedia content description interface*
  - MPEG-21 (2001) - *Multimedia framework*
- MP3 (*MPEG-1 or MPEG-2 Audio Layer 3*) is an audio encoding format that uses a lossy compression algorithm
- Common audio format for *consumer audio storage*
- *De facto* standard of digital audio compression for the transfer and playback of music on digital audio players

MP3 compression algorithm

- MP3 uses a *lossy compression* algorithm
  - it greatly reduces the amount of data required to represent the audio recording and still sound like a faithful reproduction of the original uncompressed audio for most listeners
- Bit rate specifies *how many kilobits the file may use per second of audio*
  - The higher the bit rate, the larger the compressed file will be, and, generally, the closer it will sound to the original file
- E.g., *an MP3 file that is created using the bit rate setting of 128 kbit/s will result in a file that is about 11 times smaller than the CD file created from the original audio source*
- The compression works by reducing accuracy of certain parts of sound that are deemed beyond the auditory resolution ability of most people
  - then records the remaining information in an efficient manner
- Similar principles used by JPEG

MIDI format

- MIDI (*Musical Instrument Digital Interface*) is a universally adopted language to *exchange musical information* between synthesizers and computers
- At minimum a MIDI representation of a sound includes values for the note's *pitch*, *length*, and *volume*
- It can also include additional characteristics, such as *attack* and *decay* time
- Since a MIDI file *only represents control information*, it is far more concise than formats that record the sound directly
  - Advantage: very small file size!

Audio features

- *Pitch* represents the *perceived fundamental (or lowest) frequency* of the audio data
  - While frequency can be analyzed and modeled using frequency analysis (e.g., DCT) of the data, perceived frequency needs psychophysical adjustments
    - For *frequencies lower than 1 KHz the human hear tones with linear scale*, whereas for *higher frequency values she hears in a logarithmic scale*
  - Two scales (i.e., *Mel* (or melody) and *Bark*) are commonly used for audio feature analysis rather than the original frequency scale
- *Loudness* measures the *sound level as a ratio of the power of the audio signal with respect to the power of the lowest sound that the human ear can recognize*
  - If we denote this lowest audible power with $P_{\perp}$ then the loudness of an audio signal with P power is measured (in dB) as $10log_{10}(P/P_{\perp})$
- *Beat* (or tempo) is the *perceived periodicity of the audio signal*
- *Rhythm* is the *repeated patterns in audio*

Video

- A video can be seen as a *sequence of still images representing scenes in motion*
- Thus, it maintains temporal information (as in audio) + objects and motion
  - Many of the representation techniques that we saw for images and audio data can apply
- In the following we detail on
  - physical video representations
  - video formats (e.g., M-JPEG, MPEG, AVI, DivX…)
  - some basic features

Video representation (1)

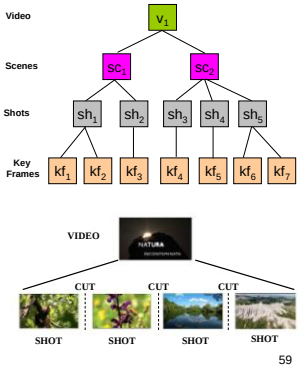
- A video can be represented as a 3-D array of color pixels
  - two dimensions serve as spatial (horizontal and vertical) directions of the moving pictures, and one dimension represents the time domain
- A data *frame* is a set of all pixels that correspond to a single time moment (i.e., a still *image)* of the complete moving picture
  - *The individual frames are separated by frame lines*
- When the moving picture is displayed, each frame is flashed on a screen for a short time (nowadays, usually 1/24th, 1/25th or 1/30th of a second) and then immediately replaced by the next one
- *Persistence of vision* (*POV*) is the *phenomenon of the eye by which an afterimage is thought to persist for approximately 1/25th of a second on the retina*
  - *POV* blends the frames together, producing the *illusion of a moving* image

Video representation (2)

- *Frame rate* is the *number of still images per unit of time of video*
- Ranges from *6 or 8* frames per second (frame/s) for old mechanical cameras to *120* or more frames per second for new professional cameras
  - The minimum frame rate to achieve the *illusion of a moving image* is about *15 frame/s*
  - In order to obtain *good quality* of motion the frame rate has to be *30 frame/s*
- *Aspect ratio* describes the dimensions of video screens and video picture elements
  - is measured as the *ratio between width and height* of video picture elements
    - e.g., 4/3, 16/9

Which problems with video streams?

- *Video streams are collection of objects*, synchronized through temporal and spatial constraints
- *Shot detection* (or *video segmentation*) gives a set of *frames* which are
  - *atomic* and
  - *share similar visual features*
- Each frame needs individual coding
- *Frame by frame representation* is too costly
  - *30 frame per second, at least!!*



Video compression

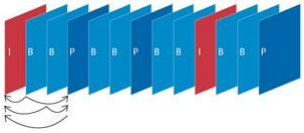
- Video data contains *spatial and temporal redundancy* making uncompressed video streams extremely inefficient
  - spatial redundancy is reduced by registering differences between parts of a single frame
    - this task is known as *intraframe* compression (closely related to image compression)
  - temporal redundancy can be reduced by registering differences between frames
    - this task is known as *interframe* compression (e.g., *motion compensation*)
- The obvious solution is to *encode each frame with image compression techniques*
  - MPEG group studied the *M-JPEG* (*Motion JPEG*) standard
    - *uses JPEG for each frame*
    - ☹ no exploitation of temporal redudancy!!

MPEG format

- The most common modern standard for video compression is MPEG
  - *MPEG-2*, used for DVD and satellite television, and
  - *MPEG-4*, used for home video
- Example:
  - DVDs use MPEG-2 video coding standard that *can compress around two hours of video data by 15 to 30 times*, while still producing a picture quality that is generally considered high-quality for standard-definition video

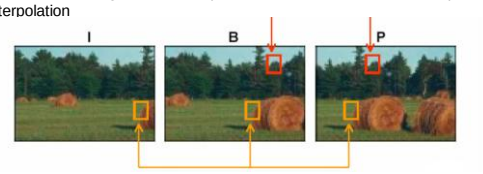
MPEG algorithm

- In the MPEG standard frames in a sequence are coded using 3 different algorithms:
  - **I** frame (intra images): coded using DCT-based technique similar to JPEG (*intraframe encoding*)
    - Are used as random access points in MPEG streams and they give the lowest compression ratios within MPEG
  - **P** frame (predicted images): coded using forward predictive coding, where the actual frame is coded with reference to a previous frame (**I** or **P**)
    - Compression ratio higher than of **I** frames
  - **B** frame (bidirectional or interpolated images): coded using two reference frames, a past and a future frame (**I** or **P**)
    - Highest compression ratio



MPEG interframe encoding

- The coding phase for **P** and **B** frames includes the *motion estimation*, which find the best matching block in the available reference frames
  - P frames are always using forward prediction
  - B frames are using bidirectional prediction, also called motion-compensation interpolation



- *Motion estimation is used to extract the motion information* of the video sequence
  - For every block (16x16) of P and B frames, 1 or 2 *motion vectors* are computed

AVI format

- AVI (*Audio Video Interleaved*) is a *multimedia container format* introduced by *Microsoft* in 1992 as part of its *Video for Windows technology*
- AVI files *can contain both audio* and video data in a file container that allows synchronous audio-with-video playback
- An AVI file may carry audio/visual data in any compression scheme, including M-JPEG, and MPEG-4