Modeling and Processing of Multimedia Data

International Second cycle degree programme (LM) in
Digital Humanities and Digital Knowledge (DHDK)
University of Bologna

Multimedia
Information Retrieval – Part I

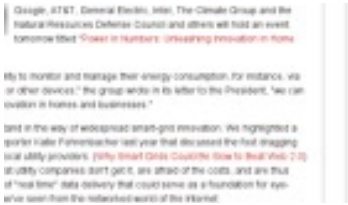
Home page: http://www-db.disi.unibo.it/courses/DMMMDB/
Electronic version: 2.01.MultimediaInformationRetrieval-I.pdf
Electronic version: 2.01.MultimediaInformationRetrieval-I-2p.pdf

Outline

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

Media (or medium)

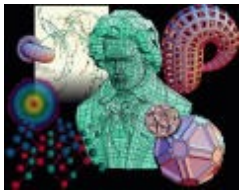
- A way to distribute and represent information such as books, newspapers, music, radio news, TV news, etc.
  - E.g.: **text**, **graphics**, **images**, **voice**, **sound**, **music**, **animation**, **video**, etc.



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)

continuous	moving images sound	animations digital music
discrete	still images	text graphics

captured from real world

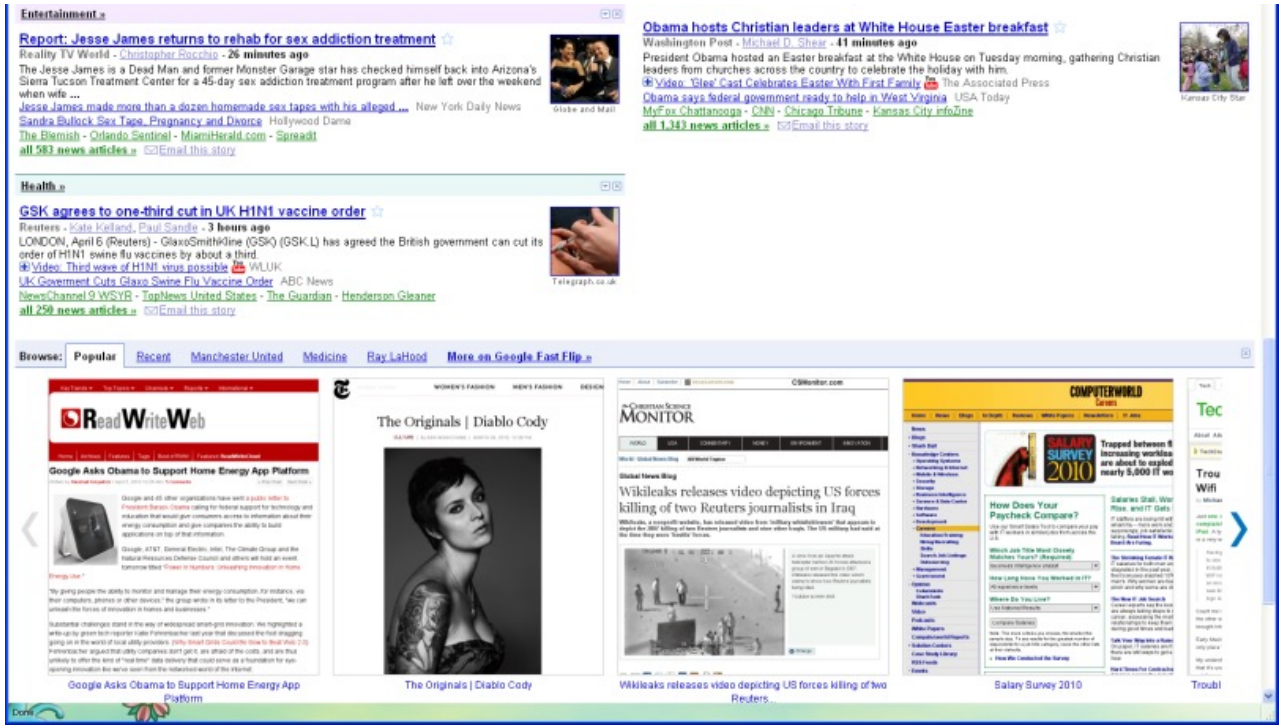
created using a PC

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

- 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/application:
  - a system/application that uses both discrete and continuous media

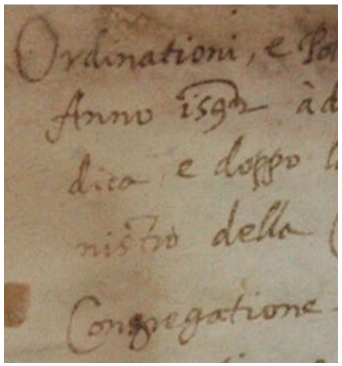


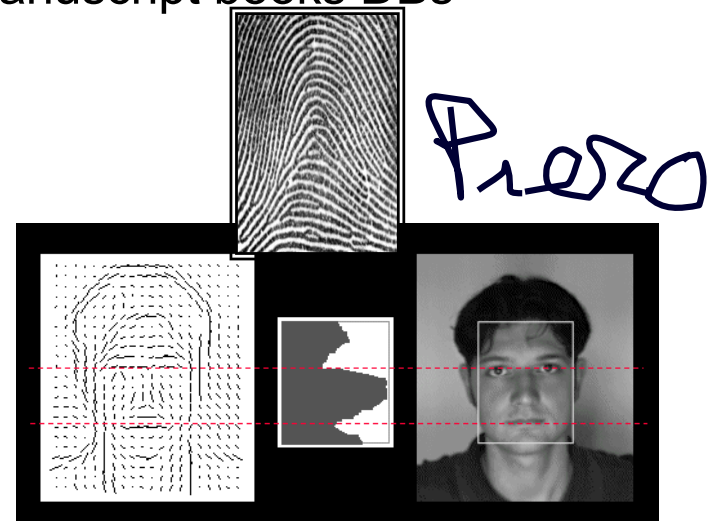
Application domains (1)

- 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)
  - Museum DBs (e.g., digital representation of objects)
  - Fashion DBs (e.g., potographs collections of clothing)
  - Ancient manuscript books DBs
  - …





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, and video)
- *Here we concentrate on aspects related to*
  - *representation of specific media types such as:*
    - *images*
    - *videos*
  - *search and 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)
    - recover the presentation to be similar to the original one
  - *Hybrid* (e.g., JPEG, MPEG)

Encyclopedia example (1)

- 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!!

Encyclopedia example (2)

- Let's assume to apply compression algorithms to the different media of the encyclopedia in order to obtain the following compression ratios:
  - Text 2:1;
  - Color picture 15:1;
  - Maps 10:1;
  - Stereo sound 6:1;
  - Animations 50:1;
  - Digitized movies 50:1.

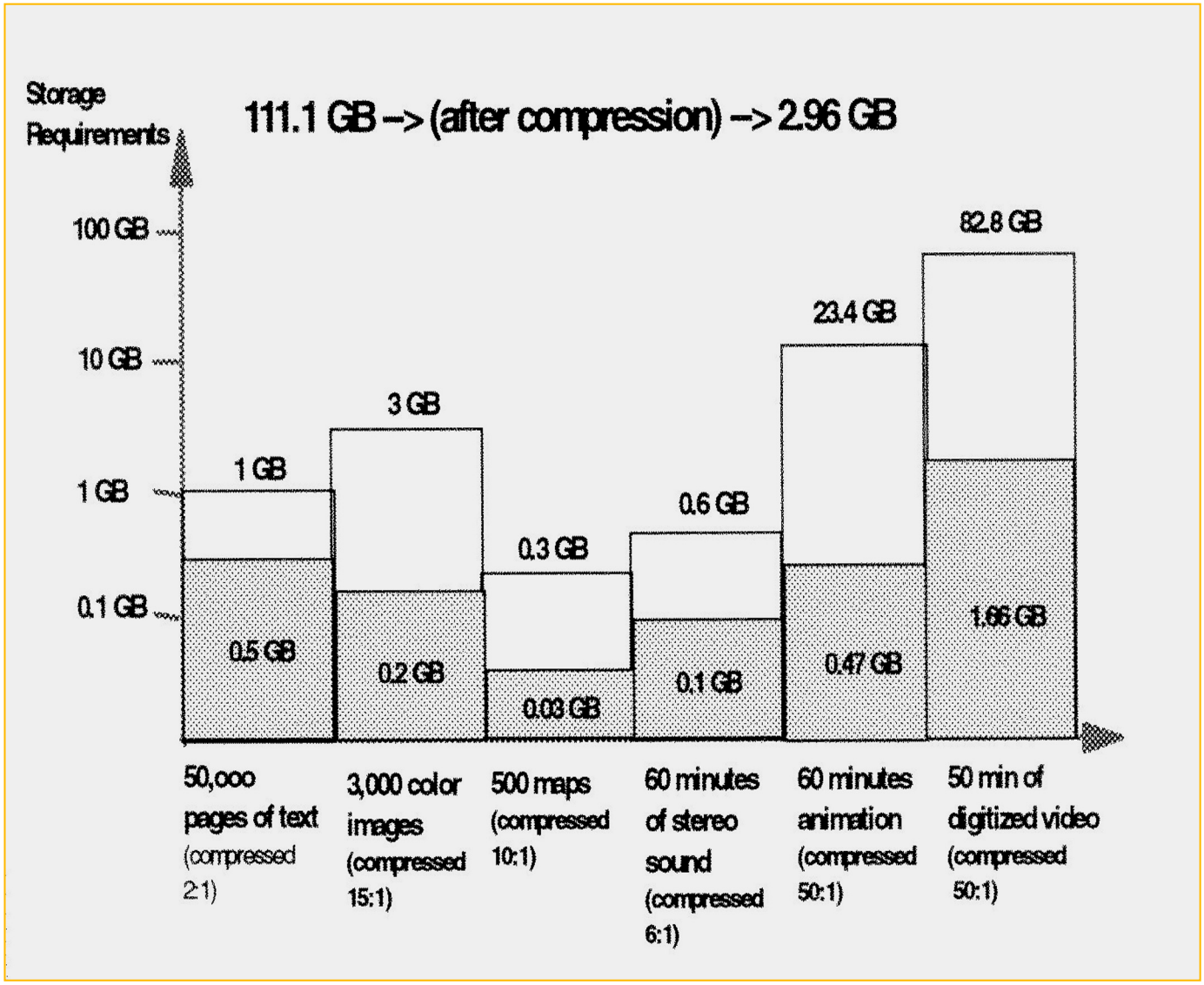
…the amount of saved memory is from 111.1 GB to 2.96 GB!!!!

Compression ratio:

$$C_R = \textit{uncompressed size / compressed size}$$

(C_R is inversely proportional to compression quality)

Encyclopedia example (3)



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, 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

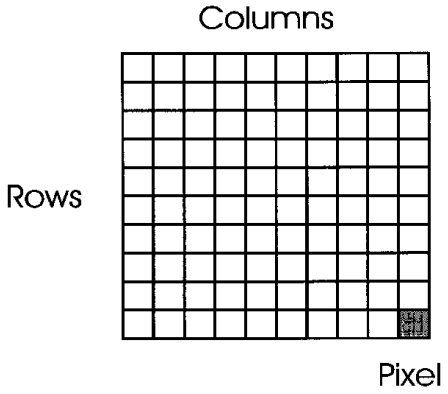
- 3 hierarchical levels of MM content representation:
  - High-level: semantic knowledge - bridge the semantic gap by integrating high level concepts (sites, objects, events) and low-level visual/audio features
  - Mid-level: text annotations/attributes (e.g., “JPEG”, “bear”, “grass”, …)
  - Low-level: 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

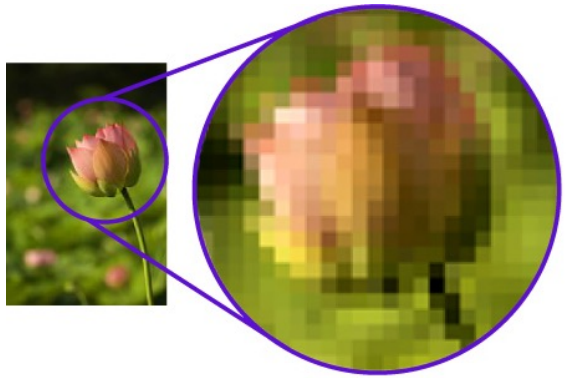
One image is worth 1,000 words…

- 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
  - 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

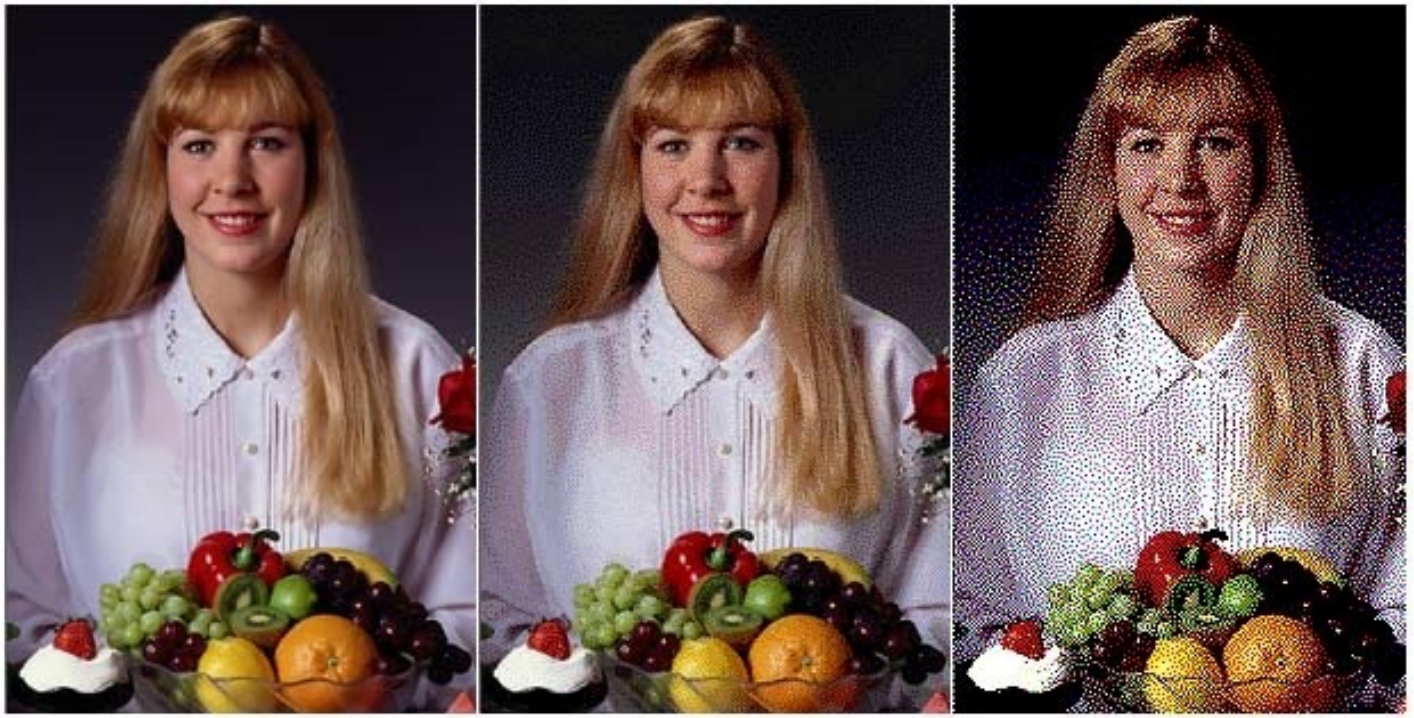
Image representation (2)

- 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

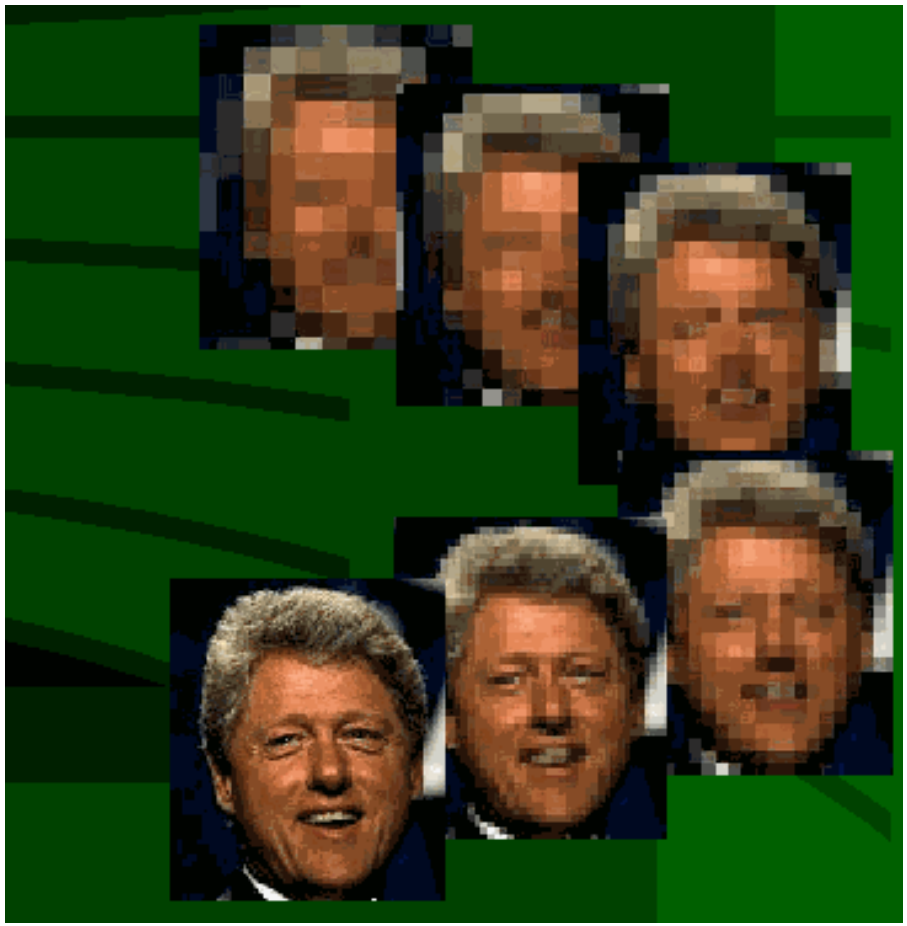
- Dimension is the number of pixels in an image; identified by the *width* and *height* of the image as well as the *total number of pixels* in the image (e.g., an image 2048 wide and 1536 high (2048 x 1536) contains 3,145,728 pixels - 3.1 Mp)
- Spatial resolution is the *number of pixels per inch – 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

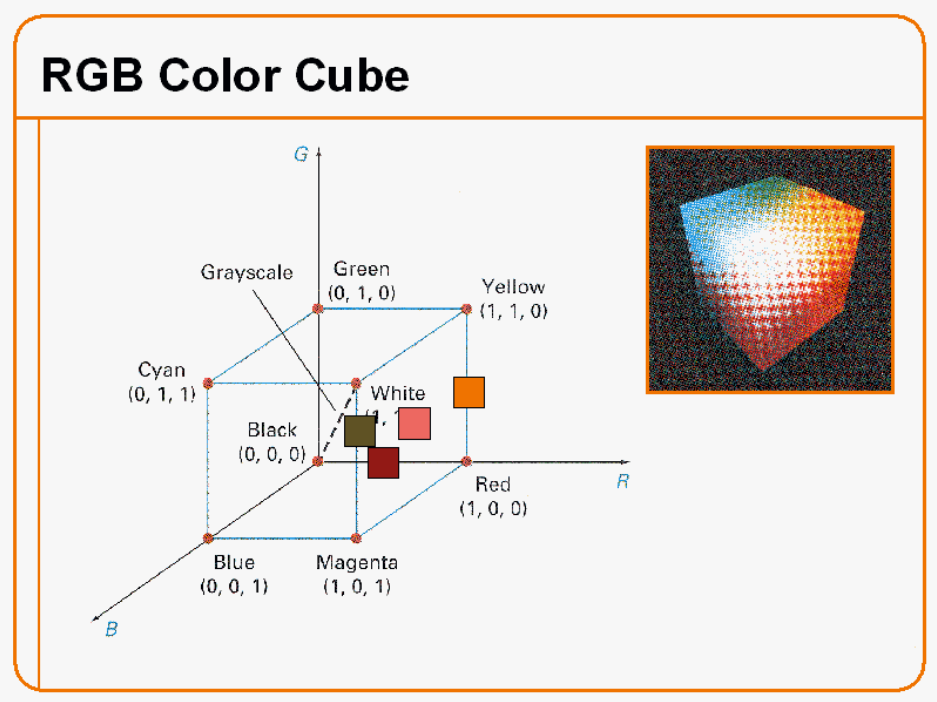


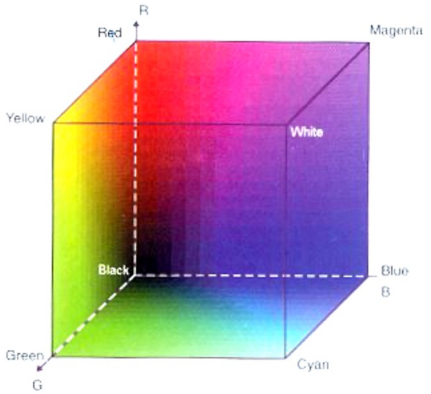
Color

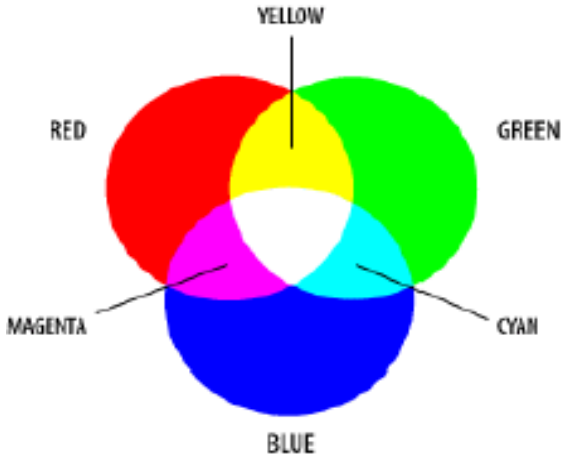
- According to the tri-chromatic theory, the sensation of color is due to the stimulation of 3 different types of receptors (*cones*) in the eyes
- Consequently, *each color can be obtained as the combination of 3 component values* (one per receptor type)
- A *color space* defines 3 color channels and 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

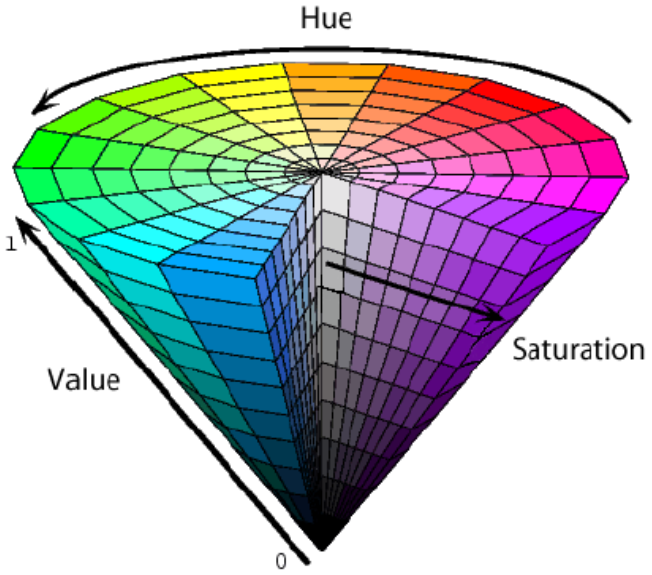


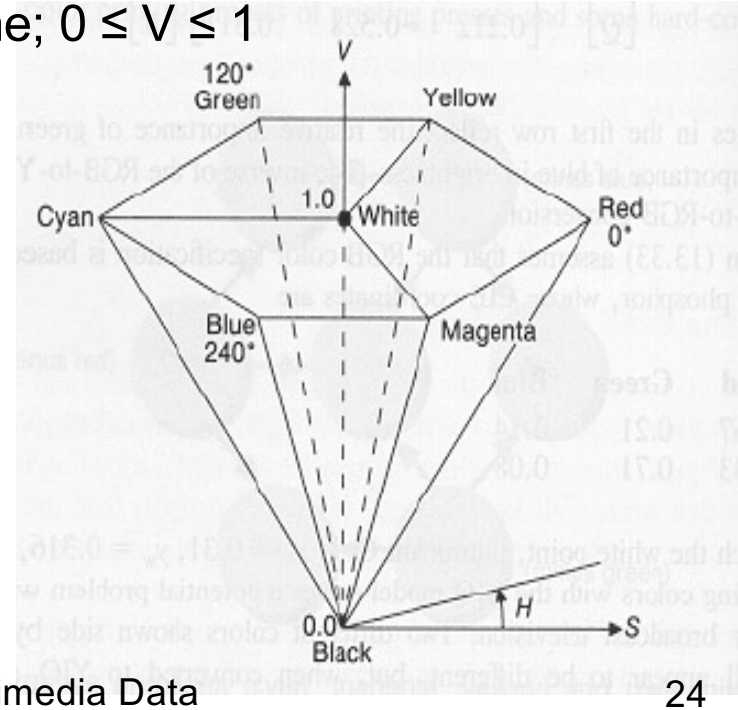




Color spaces: HSV

- 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 ≤ H ≤ 360
- Saturation is the amount of color that is present (e.g., red vs. pink)
  - Saturation is the distance from the center; 0 ≤ S ≤ 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 ≤ V ≤ 1





Saturation of colors

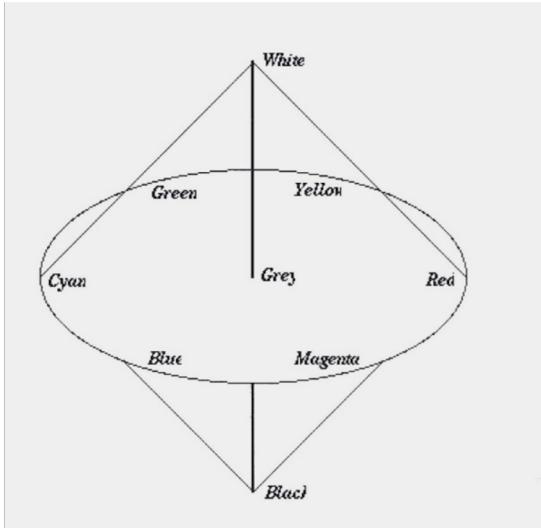
Original image

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”



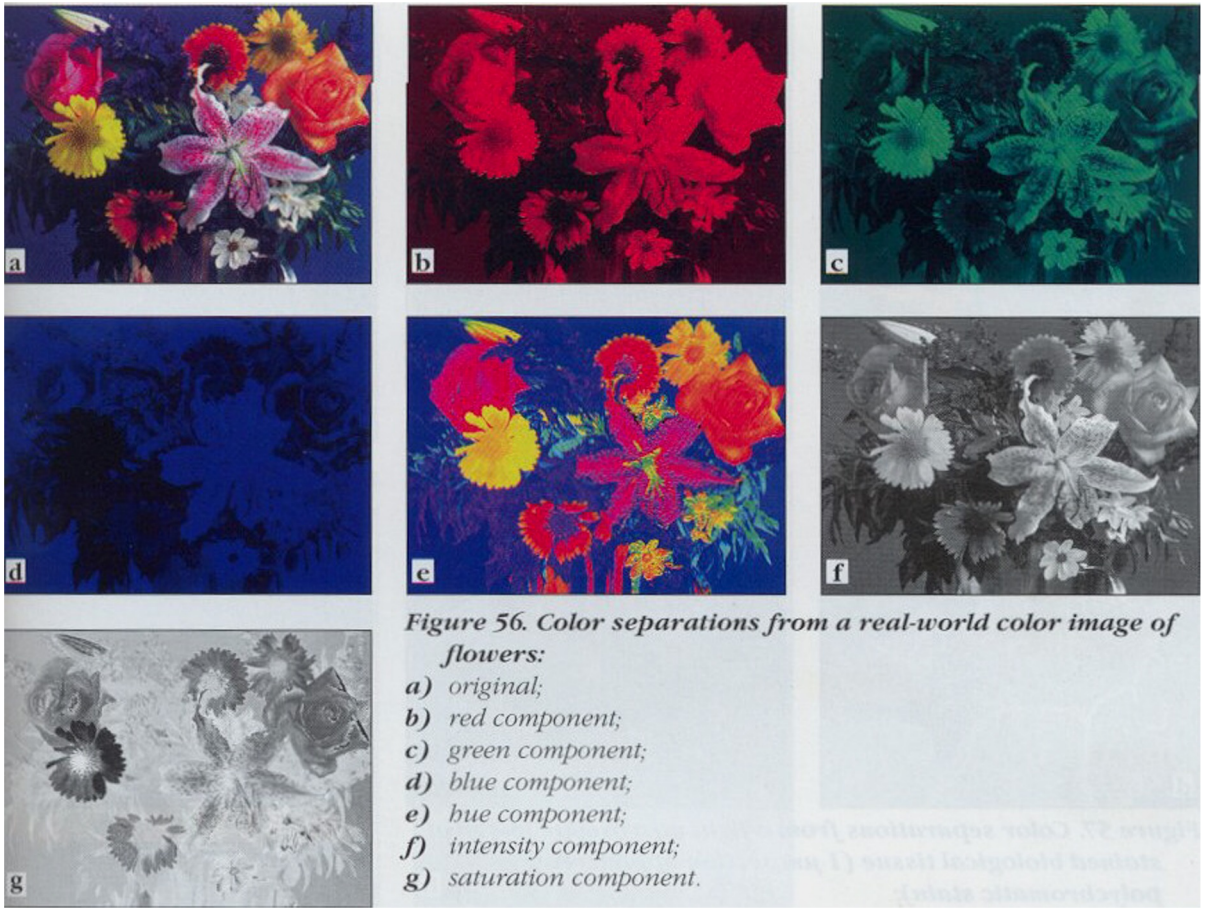


Figure 56. Color separations from a real-world color image of flowers:
a) original;
b) red component;
c) green component;
d) blue component;
e) hue component;
f) intensity component;
g) saturation component.

Color spaces: from RGB to HSV

- The conversion from RGB to HSV values is based on the following equations:

$$H = \cos^{-1} \frac{[(R-B)+(R-G)]/2}{[(R-G)^2+(R-B)(G-B)]^{1/2}}$$

$$S = 1 - 3 \times \min\{R, G, B\}/(R+G+B)$$

$$V = (R+G+B)/3$$

- HSV is much more suitable than RGB to support similarity search, since it better preserves perceptual distances

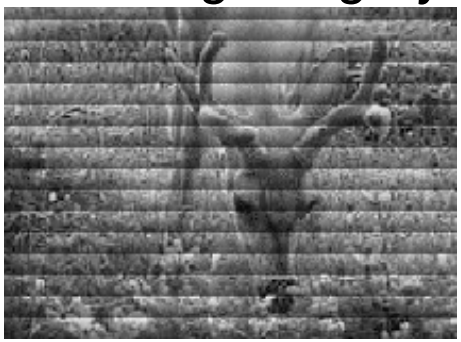
Texture

- 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

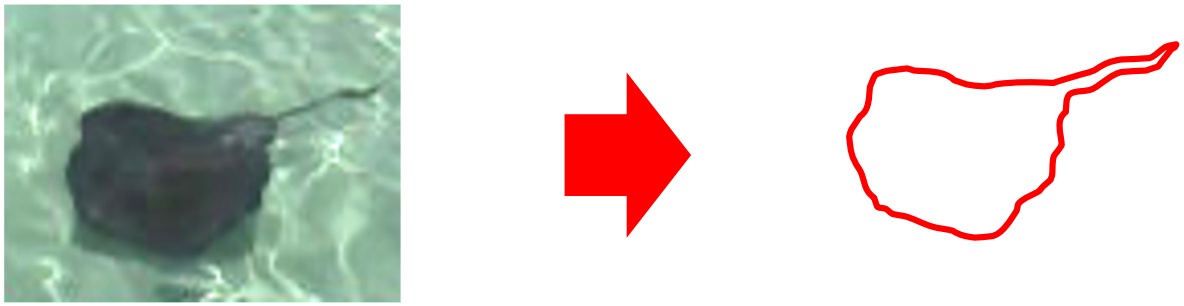
- Intuitively, texture provides information about the uniformity, granularity and regularity of the image surface
- *It is usually computed just considering the gray-scale values of pixels* (i.e., the V channel in HSV)



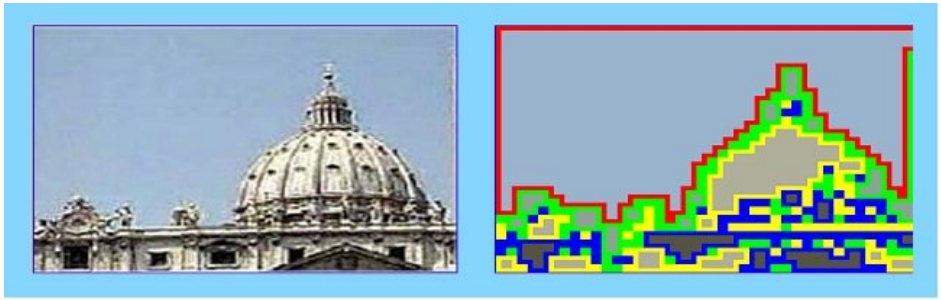
Shape

- 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…*



Example of image segmentation



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

  How many regions?

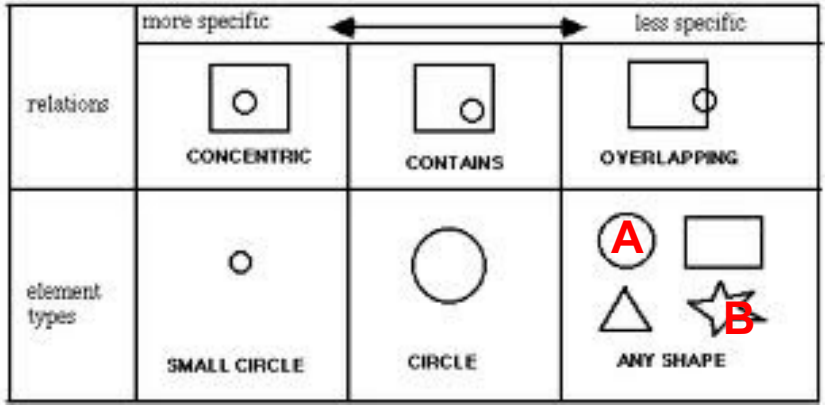
  How "homogeneous" pixels within a same region should be?

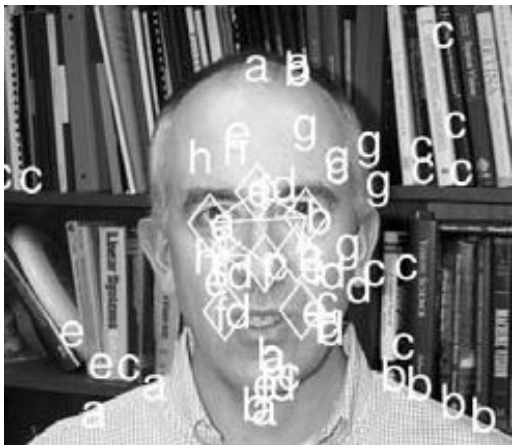
  No general answer!

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

Spatial relations of image objects

- 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*





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 can apply
- In the following we detail on
  - physical video representations
  - some basic features

Video representation (1)

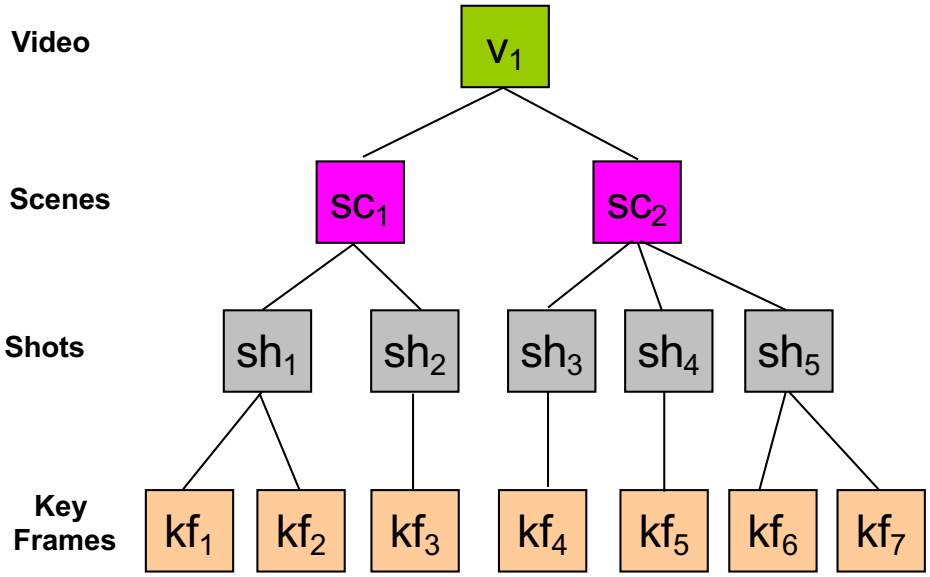
- A video can be represented as a 3-D array of color pixels
  - two dimensions serve as spatial (horizontal and vertical) directions of the moving pictures, and one dimension represents the time domain
- A data *frame* is a set of all pixels that correspond to a single time moment (i.e., a still *image)* 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 $1/24^{th}$, $1/25^{th}$ or $1/30^{th}$ 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* $1/25^{th}$ *of a second on the retina*
  - *POV* blends the frames together, producing the *illusion of a moving* image

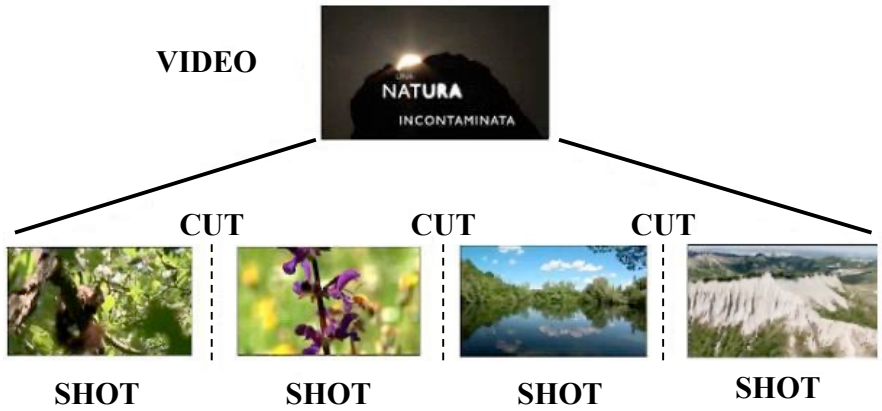
Video representation (2)

- *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" features*
    - *e.g., visual content*
- Each frame needs individual coding
- *Frame by frame representation* is too costly
  - *30 frame per second, at least!!*





Free exercise 1.B

- Starting from **descriptions in natural language** of relevant **structured**, **semi-structured**, and **unstructured data** selected for your examples of Exercise 1.A, **model the data** according to:
  - **relational model** (for **structured data**), and
  - **XML model** (for **semi-structured data**)
- Provide a **definition accurate as much as possible of the low-level features** you chose for describing the “content” of involved MM data (**unstructured data**)

Free exercise 1.B: students to do

- Prepare an electronic version of your proposals
  - .ppt file

  similarly, to the examples proposed during lectures

Archivio Storico Fiat

- **Trimotore Fiat G212**
- Data: **1947**
- Collezione: **Tema di cultura industriale**
- Tipologia: **Immagine**
- **Aereo**, **Motore**, **Ali**

Low-level features description per unstructured data
- Distribuzione colore dell'immagine
- Forma degli oggetti nell'immagine
- …