University of Trento

Part 1:
Images and Videos

Nicola Conci
nicola.conci@unitn.it

Definition

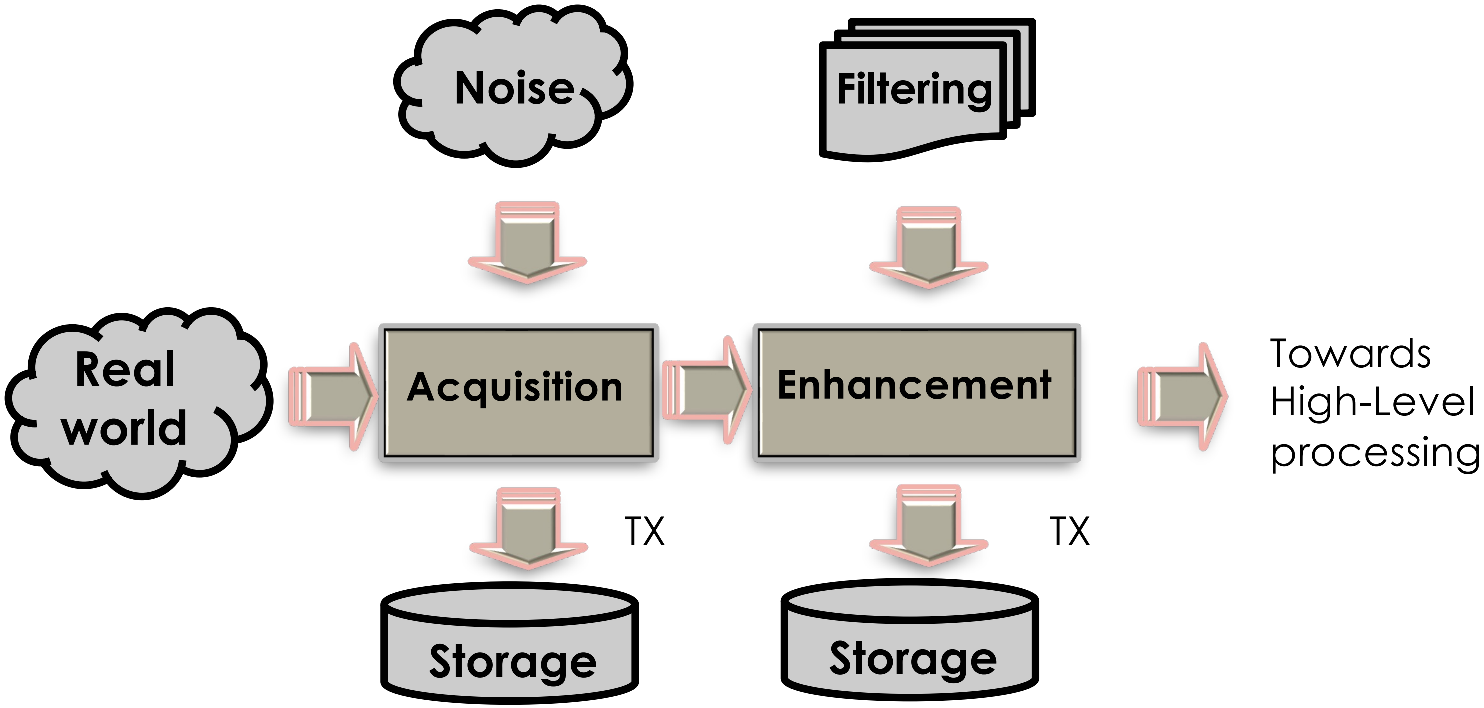
Computer vision (from Wikipedia)

Computer vision is the **science and technology of machines that see**, where *"see" in this case means that the machine is able to extract information from an image that is necessary to solve some task. […] The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner. […]*

Examples of applications of computer vision include systems for:

- **Controlling processes** (e.g., an industrial robot or an autonomous vehicle)
- **Detecting events** (e.g. for visual surveillance or people counting)
- **Organizing information** (e.g. for indexing databases of images and image sequences)
- **Modeling objects or environments** (e.g., industrial inspection, medical image analysis or topographical modeling)
- **Interaction** (e.g., as the input to a device for computer-human interaction).

The processing chain

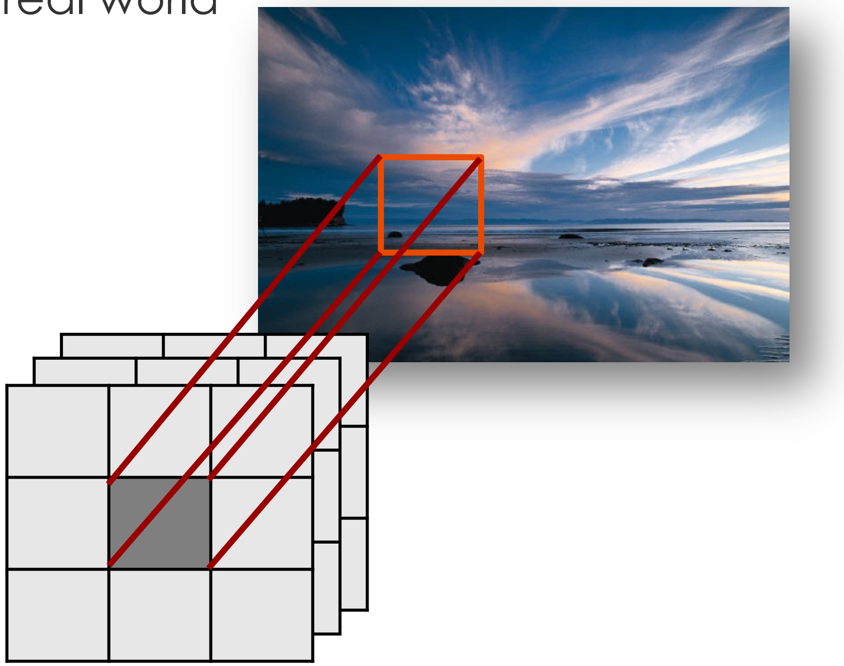


Acquisition

- Transformation of a physical signal into an electrical one (by means of a sensor)
  - **What** we measure: light intensity (B/W), wavelength (color), temperature (infrared), etc.
- It refers to the process of transferring a portion of the real 3D world onto a 2D surface
- It brings a continuous-parameter real world into a discrete-parameter one
- Representation in a standard format

Digital Images

- Collection of 2D coordinates
- Coordinates are known as pixels, picture elements
- A pixel represents a projection of a portion of the real world
- Pixels can be
  - Grayscale
    - One component, 8bit
  - Color
    - Typically 3 components, 24bits



Sampling

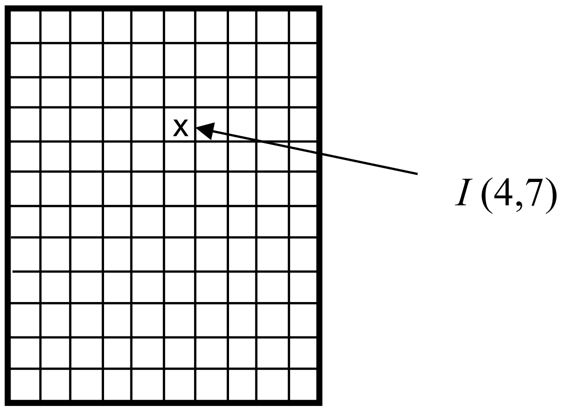
- The "real world" is a continuous function
- Analog video is a 1-D continuous function where one spatial dimension is mapped onto time by means of the scanning process
- Digital video is instead sampled in all three dimensions

Sampling in a nutshell

- Continuous signal $s_c(x_1,x_2)$
- Spatial rectangular sampling:
  - $x_1 = n_1 \Delta x_1$
  - $x_2 = n_2 \Delta x_2$
- $s(n_1,n_2) = s_c(n_1 \Delta x_1, n_2 \Delta x_2)$
- A 2D matrix of numbers, of pixels
- Once we have the digital format, we can manipulate data and:
  - Apply filters
  - Change colors
  - Store
  - Transmit

Handling images

- Pixels are numbered starting from the **top left** corner
- The value of a pixel in a certain position is defined as $I(r,c)$
  - r is the row index
  - c is the column index
- Numbering starts usually from 0 or 1, depending on habits…



Handling images

- Monochrome images
  - Values normalized in the range 0-1
  - 0 is black
  - 1 is white
  - The intensity is called *grey level*.
- Color pictures work the same way:
  - Three channels
  - Each of them in the range 0-1

Color

- What is color?
  - The attribute the human visual system associates to objects
  - A mathematical relationship that combines different wavelengths
- Why is it important?
  - To check whether something we see is what we expect
  - To recognize objects
  - To distinguish similar objects

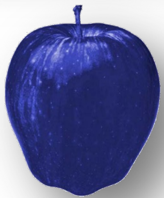
Examples

- The bus is red, no doubt.
- What would the computer say?
  - Sure there's some red, BUT
  - It's also
    - Black
    - Yellow
    - White
    - Grey
    - …

Examples

- (based on color) Can we say it's the same cat?


You would eat this:

And this?!

Color perception

- The human eye “is a camera” with a focal length of about 20mm, where the iris controls the amount of light by adjusting the size of the pupil
- The perception of color is possible through *cones* in the fovea
- It has around 100M receptors
- Cones have peak responses on three main wavelengths
  - Red (700nm)
  - Green (546.1nm)
  - Blue (435.8nm)

Image Compression

- Data can be:
  - Processed locally
  - Transmitted remotely
  - Archived on a storage unit
- Images and videos require a lot of bandwidth
- To compress a picture or a video, we need a codec
- A codec allows:
  - **CO**dec: encoding in the compressed domain
  - co**DEC**: decoding from the compressed domain
- Examples:
  - JPEG
  - MPEG
  - DIVX
- **Processing** is typically done in the **uncompressed** domain

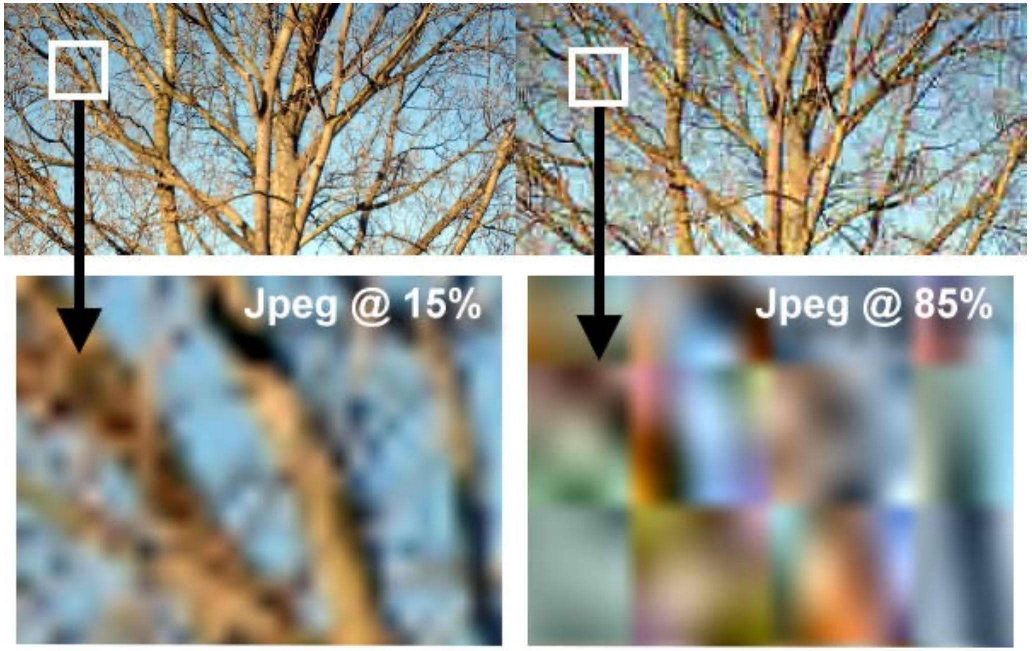
Raw vs compressed

- Raw images
  - vector of pixels
  - Usually stored in 1D vector, of size 1x(NxM)
  - If 8 bit per pixel (bpp), storage requires MxN bytes
  - Color image in HDTV format needs: 1920x1080x3 > 6MBytes
  - If it is a video at 25 pictures per second → 6x25 = 150MB
- Compressed images
  - Reduces the dimension of the file
  - With losses (most used)
  - Without losses (low compression)

Compression

- Standards
  - JPEG, compression around 10-20 or more
  - 1 Mbyte can be stored in 50 Kbytes without losing too much in quality
  - The choice of the standard typically depends on the data
- Compression reduces quality
- Compression introduces visual artifacts
- Be careful when you compress data!

Compression: example



By increasing the compression ratio, artifacts appear such as blocking, blurring, chromatic aberrations

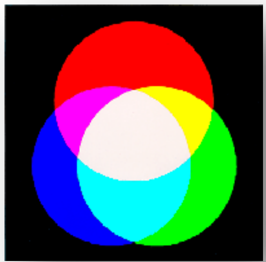
Compression or conspiracy?

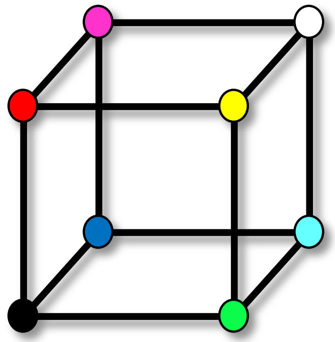
Translation:
I took a picture of the sky and zoomed, but what is the sky made of? It's similar to a huge circuit board or a hologram? You can see perfect geometries, an L upside down, a perfect rectangle, etc etc.

OR…
We can call it compression artifact + noise of the sensor

Additive Color Model

- Colored beams are projected onto a black surface
- Beams overlap
- Human eye receives the stimula without generating interference
- The eye mixes the components and perceives the resulting color
- Starting from the primary colors RGB we can obtain:
  - R+G = Yellow
  - R+B = Magenta
  - B+G = Cyan
  - R+G+B = White



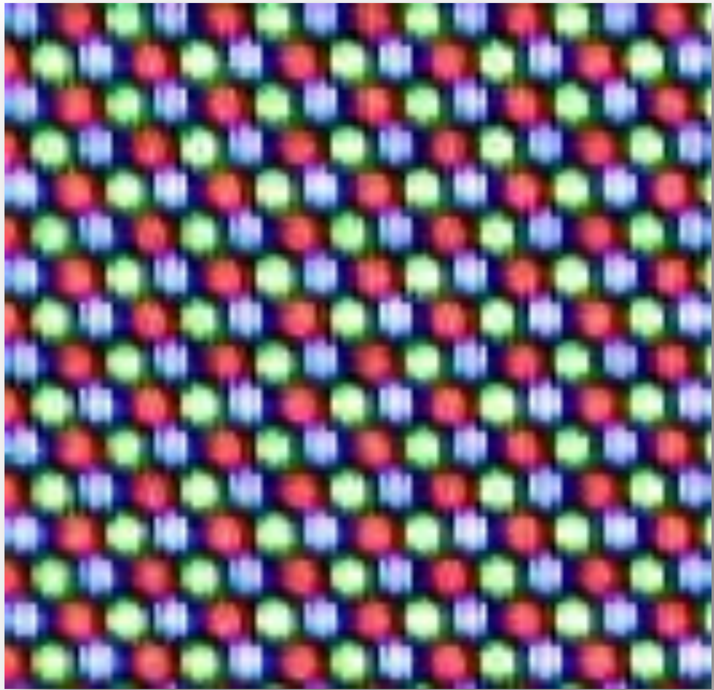


- ***Subtractive color* is the inverse process**

How are colors combined?

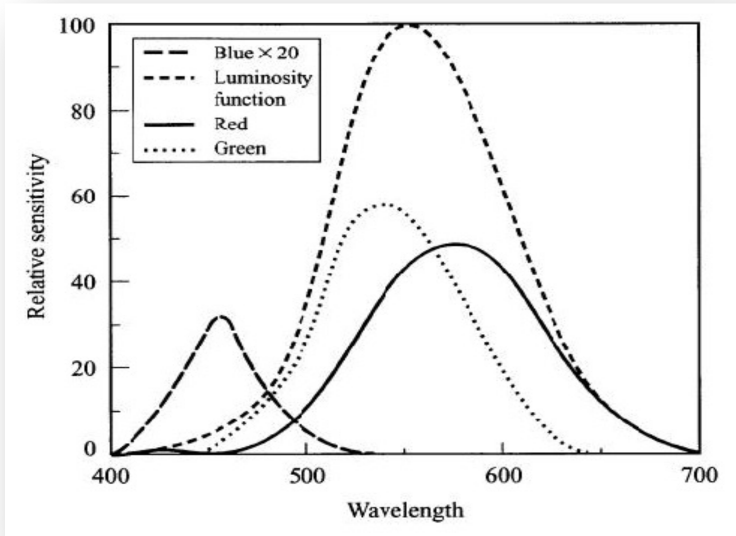
- Black: RGB (0,0,0)
- Green: RGB (0, 1, 0)
- Yellow: RGB (1, 1, 0)
- White: RGB (1, 1, 1)
- Grey: RGB (0.5, 0.5, 0.5)

Display



Color spaces: why RGB?

- Major response in the green component
- Red and Blue are *less* relevant
- Higher response to light than color
- Maybe a different representation would be more effective



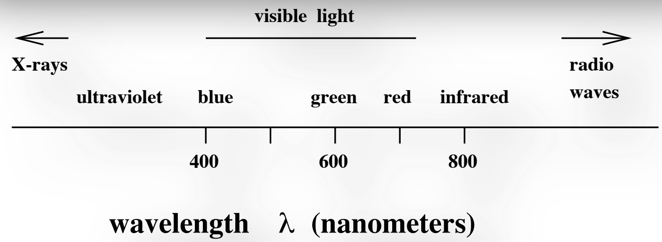


Image perception

- In general the human eye is more sensitive to luminance than color
- It is also more sensitive to **contrast variations** with respect to absolute values
- The internal square of the left image appears brighter than the one on the right
- Even though it's the same color…

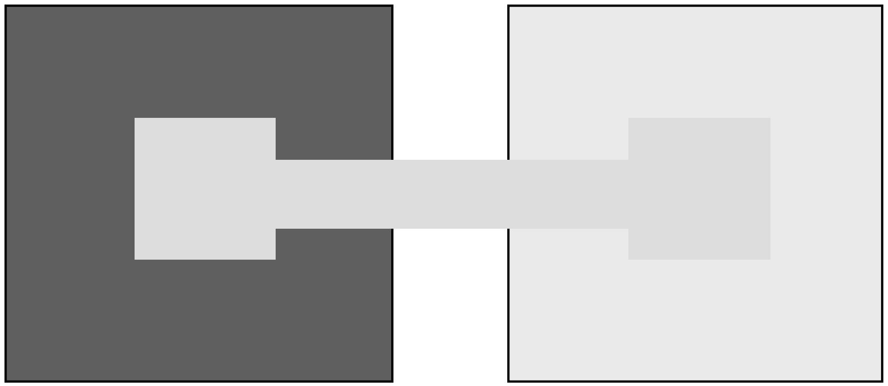
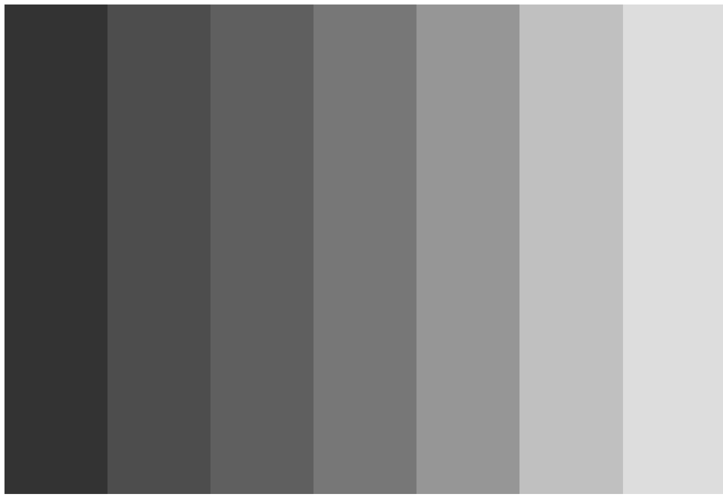


Image perception

- What is the luminance level of the stripes?
- Is it constant?
- It's actually constant, but perceived as over/undershoot of intensity because of the contrast



Color spaces: RGB

- If we separate the three components and generate single images we notice:
  - Components are correlated
  - This means that the three greyscale images carry *almost* the same amount of information

RGB **R** **G** **B**

Color spaces: YCbCr

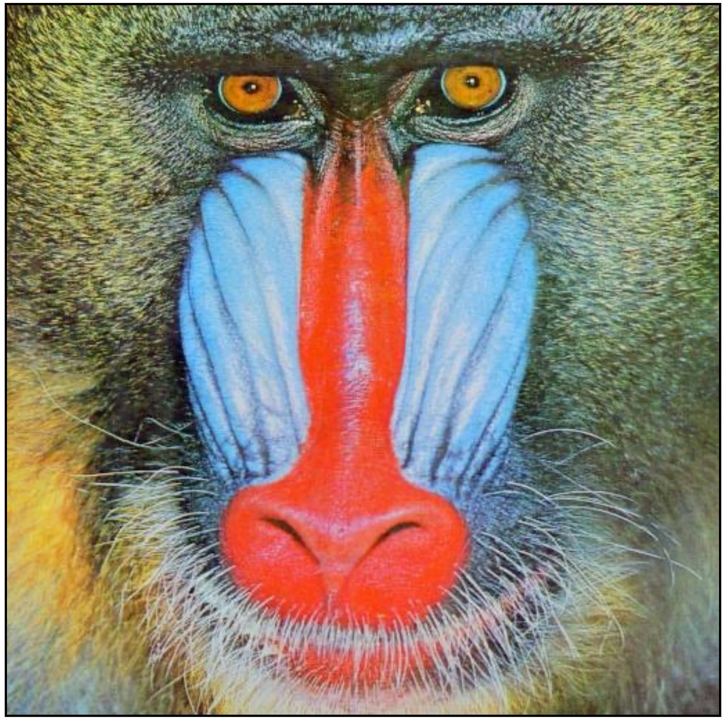
- More effective, since luminance (Y) is separated from the chrominance components (Cb Cr)
- Compatible with the Human Visual System:
  - Rods (120M) are used to discriminate light levels
  - Cones (6-7M) respond to color stimula in the RGB wavelengths.

Color spaces: YUV

- Different formats:
  - 4:4:4 → each component is fully represented
  - 4:2:0 → chrominances are downsampled by a factor 2
  - …
- YCbCr is a generalization of YUV, just a matter of conversion matrices
- Downsampling of color is done for convenience (storage)

Color representation

Original picture $\mathbf{YC_bC_r}$, no downsampling

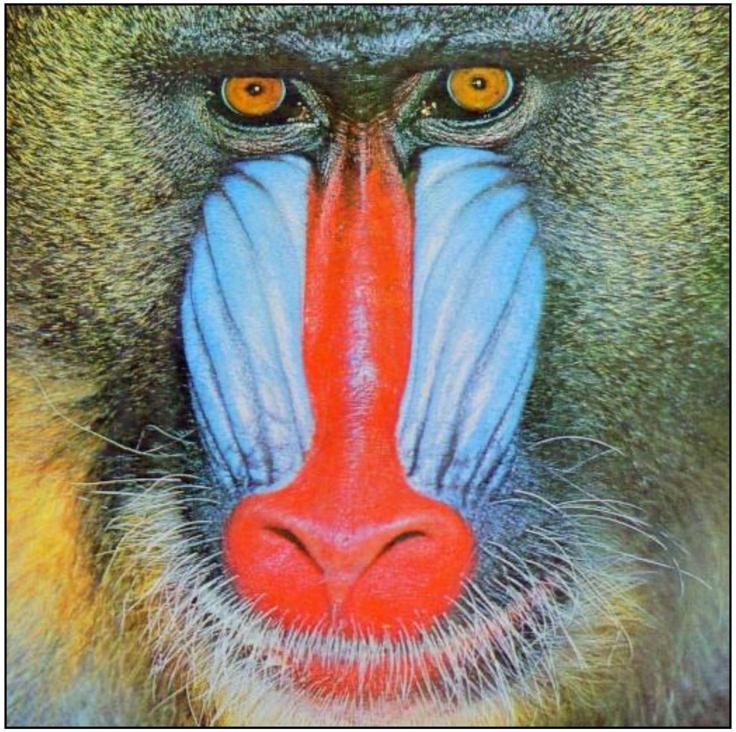


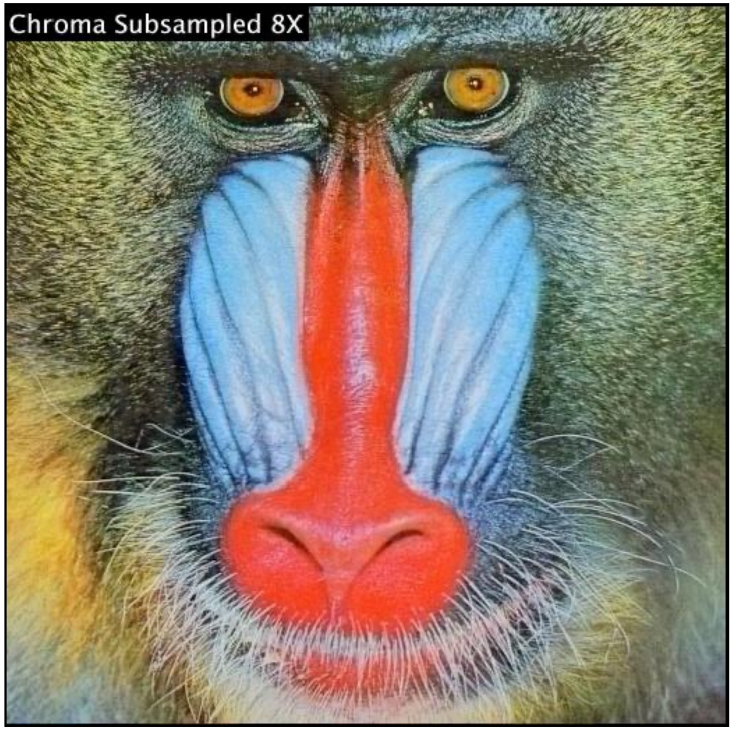
Color representation

Image $\mathbf{YC_bC_r}$ downsampling on Y

Color representation

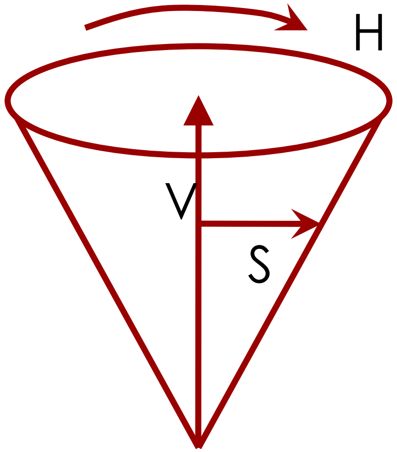
Image $\mathbf{YC_bC_r}$, downsampling on $\mathbf{C_bC_r}$

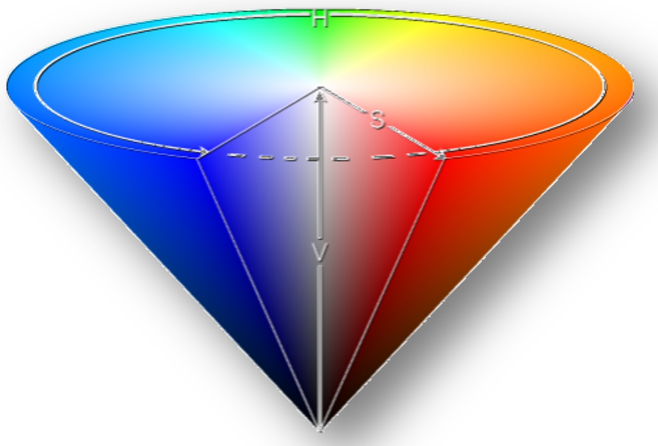




Color spaces: HSV

- Color is represented through:
  - Hue
  - Saturation
  - Value





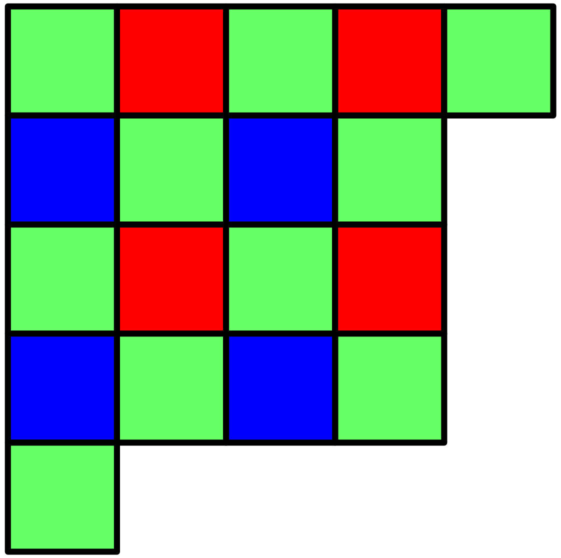
Different color spaces for different applications

- RGB is used in general for visualization
  - In displays each pixel is composed by three phosphors (CRT) or LEDs (LCD)
- YUV is suitable for compression
  - We are less sensitive to chrominance variations
  - U and V can be downsampled
- HSV is robust for computer graphics and image analysis
  - H is the "color"
  - V is "brightness"
  - S is the "intensity"

Bayer Pattern

- In the acquisition phase, light is captured by the CCD (Charge Coupled Device)
- The CCD is an array of cells
- The best solution would be to have devices with 3 different CCDs
- Most cameras are single-chip
- To correctly exploit the human eye response:
  - Three types of photosensors
  - 50% green
  - 25% red
  - 25% blue

Bayer Pattern



- Green sensors are defined as *luminance-sensitive elements*
- Red and Blue sensors are defined as *chrominance-sensitive elements*

Quantization

- Like in the mono-dimensional case, signals need to be quantized
- Quantization implies the definition of a number of levels to define our signal
- Typically, the range 0-1 is quantized using 8bpp
- Other representations with 10-12 bpp are available

Why 8bpp?

- 8bpp represent 256 levels, which is fine for the human eye
- 7bpp (128 levels) would still be ok
- What happens if we quantize with less than 6bpp (64 levels, minimum to ensure “smooth” pictures)
- False contours appear → contouring

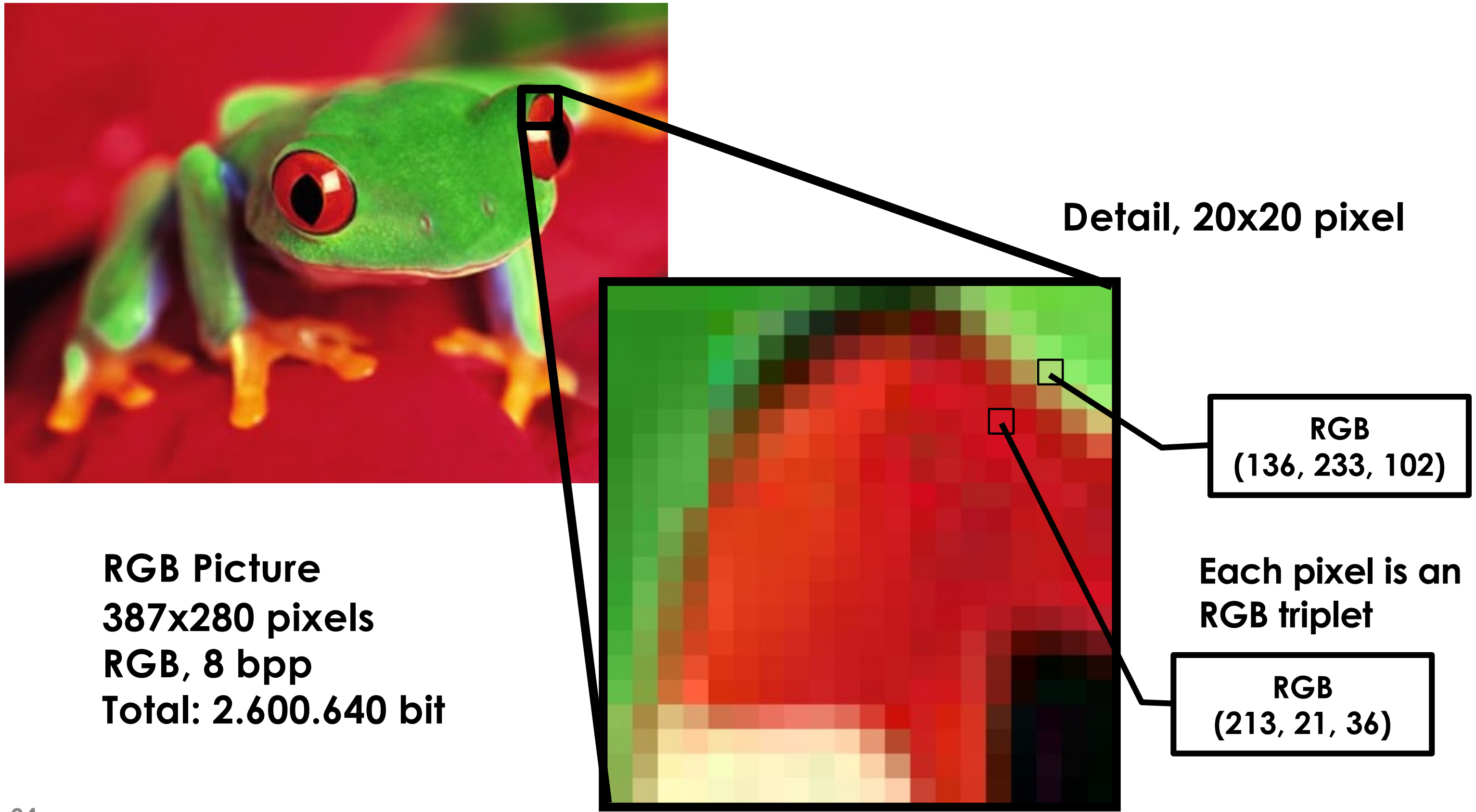
Contouring

- At 5bpp

Contouring

- At 4bpp

Images: representation



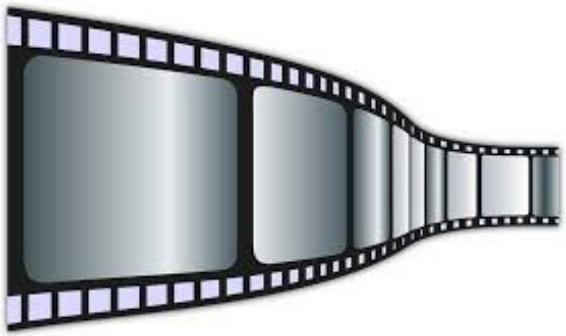
RGB Picture
387x280 pixels
RGB, 8 bpp
Total: 2.600.640 bit

Limits of the 2D

- Still images provide a reliable information about static scenes
- We lose motion information:
  - Temporal evolution of the scene
  - Rapid changes
  - Dynamics of motion (qualitative and quantitative)
  - How subjects/objects relate one to each other
- Analyzing a video provides a more consistent representation of the scene
- It’s closer to what humans do every day

What is a video?

- Sequence of 2D images that represents a projection of a moving 3D scene onto the video camera image plane.



It is expected that adjacent frames are strongly correlated

Resolution in images and video

- Images are up to 50MP
- Resolution in video is typically lower:
  - Full HD → 1920x1080 =~ 2MP
  - 4k → 3840 x 2160 =~ 8.3MP
  - 8k → 7680 x 4320 = ~33MP
- Reasons:
  - Videos can last hours
  - Videos have a higher frame rate, up to 60fps (also 120 or more)
  - Single frames last for a short time
  - Storage could be troublesome in terms of capacity and access to disk

Relevant features

- Once the image is acquired, what are the most relevant features we could be interested in?
  - Color and its distribution
  - Presence of edges and contours
  - ...
- If the analysis concerns a video, instead of an image, what is the added value?
  - Consistency of the features mentioned above over time
  - Evolution of the scene and objects displacement
  - Objects may enter/exit the scene

Examples

- Static background
- Low motion
- “Controlled” environment

Examples

- Noise
- Distortion
- Light artifacts

Examples

- Background with slow changes
- Consistent motion due to the presence of people
- Shadows
- Long temporal range of analysis

Comments

- Why are these scenes complex to analyze/process?

  - Edges are moving also if the scene “appears” static
  - Pixels are not constant even though they “look” constant
  - Shadows can be seen as moving objects depending on their intensity
  - Environmental conditions can significantly “disturb” the analysis modules
  - We have then highlights, reflections, occlusions, masking