Institut
Mines-Telecom

Introduction to compression

Marco Cagnazzo

MN907 - Multimedia Compression

Outline

The human visual system

The sound perception

Image and video representation

Compression principles

Outline

The human visual system

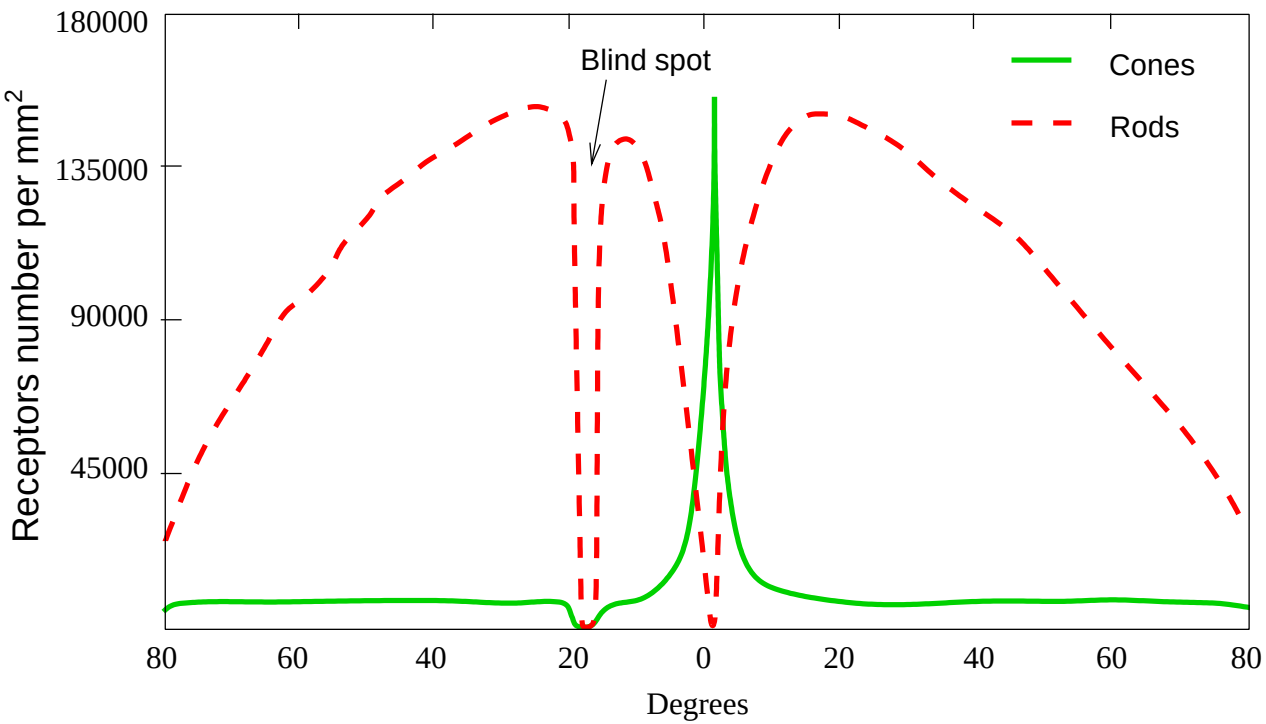
The sound perception

Image and video representation

Compression principles

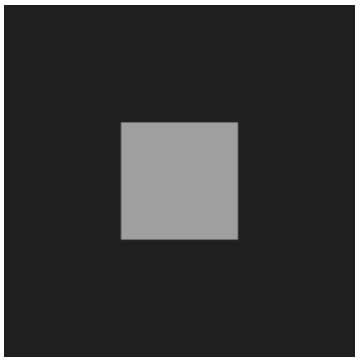
The eye

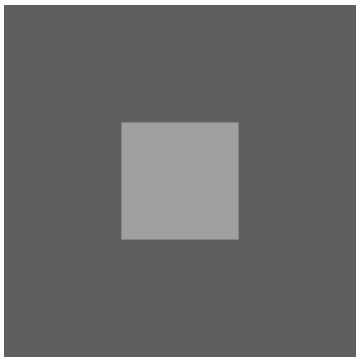
- Light is transformed into neural implusions by the retinal receptors
  - **Cones** (6÷7 millions, center of the retina) : very sensitive to cololrs, good resolutionm need high illumination
  - **Rods** (75÷150 millions) : sensitive to light intensity low resolution, very sensitive in low illumination

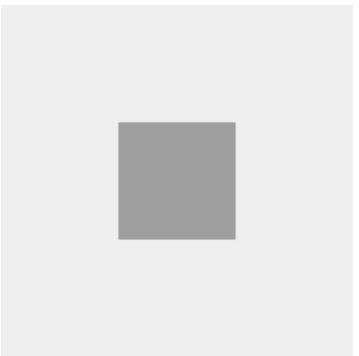


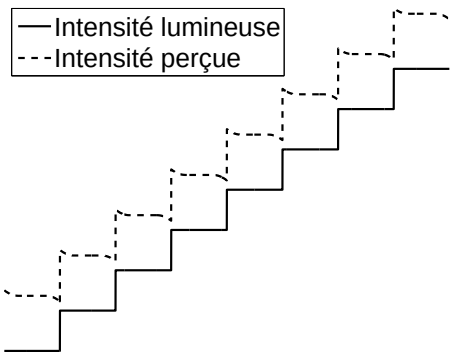
Light perception

- Perceived intensity : logarithmic function of intensity
- Intensity dynamic range: $\approx 10^{10}$ (100dB)
- The humaun visual system (HVS) cannot operate into such a large range at once
- Global lighting changesm reduced range dynamics
- Perceived light intensity:it is not just a function of light intensity

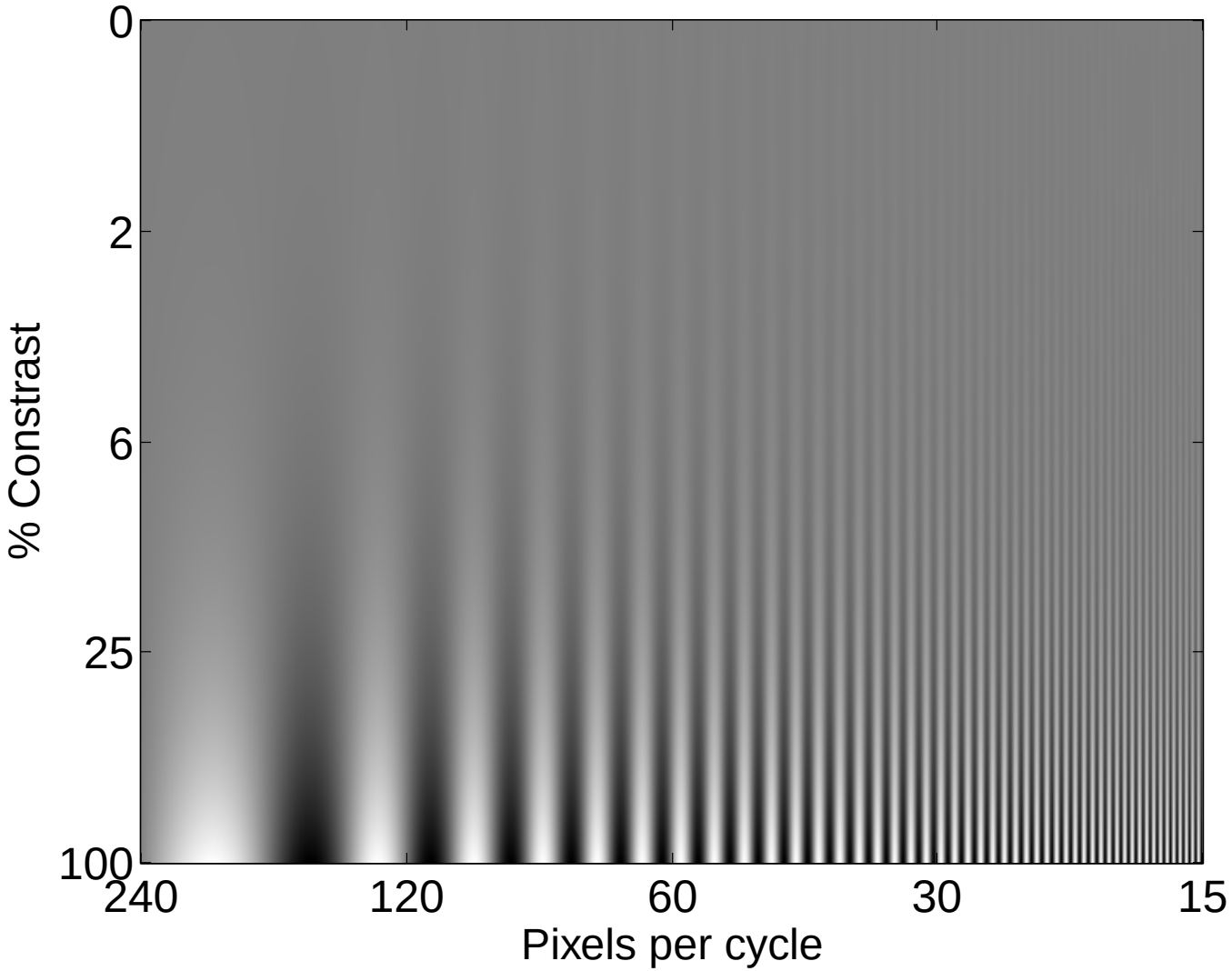








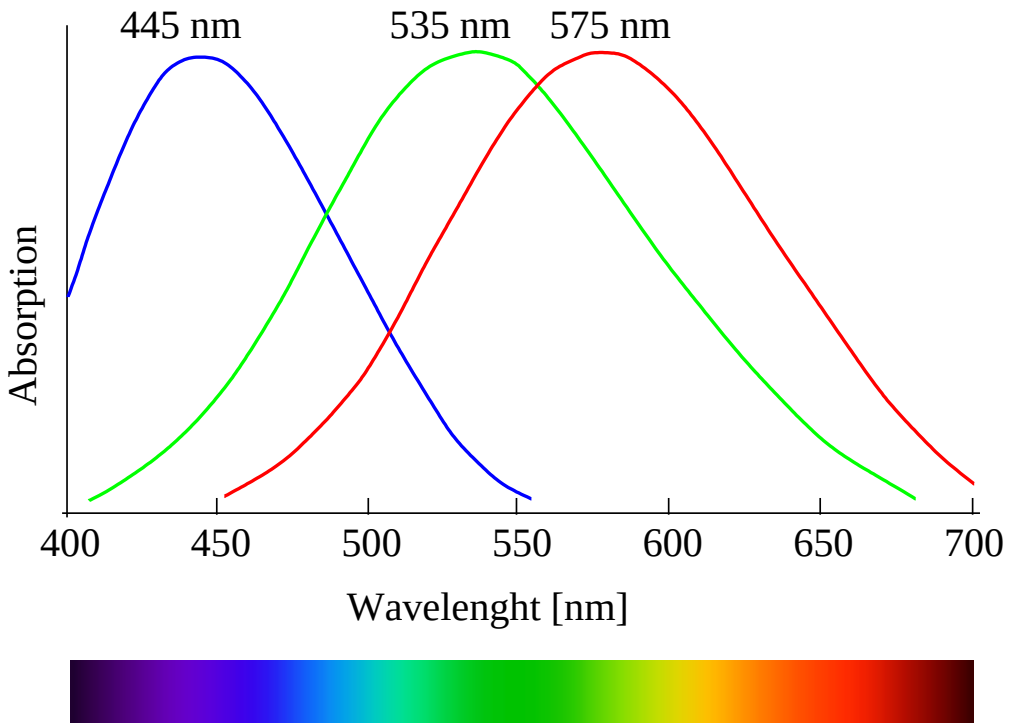
Contrast sensitivity function (CSF)



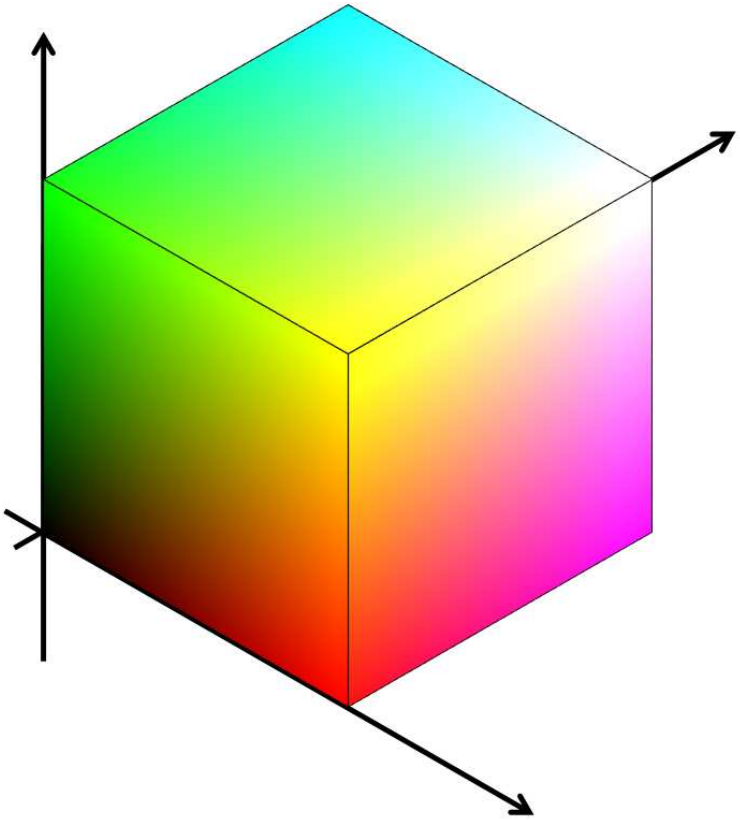
- Maximum at 2-5 cycles per degree

Color Perception

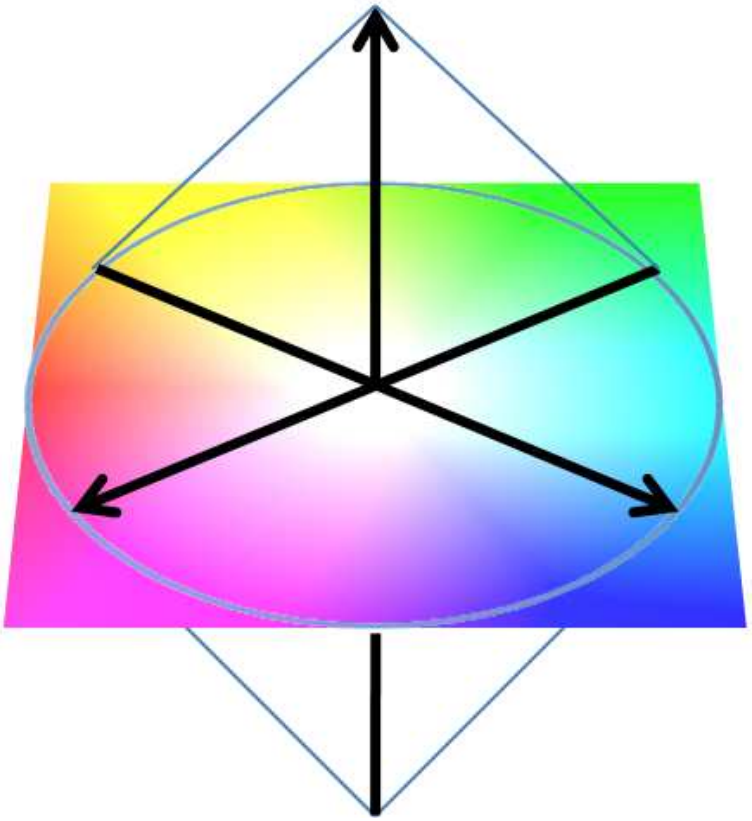
- Visible light spectrum: 400÷700 nm
- Cones sensitivity:
  - 65% sensitive to red
  - 33% sensitive to green
  - 2% sensitive to blue (but very sensitive)
- Color sensation: corresponds to the *tristimulus*
- Color is obtained as a combination of the three primary colors



Color spaces



RGB



HSV

Outline

The human visual system

The sound perception

Image and video representation

Compression principles

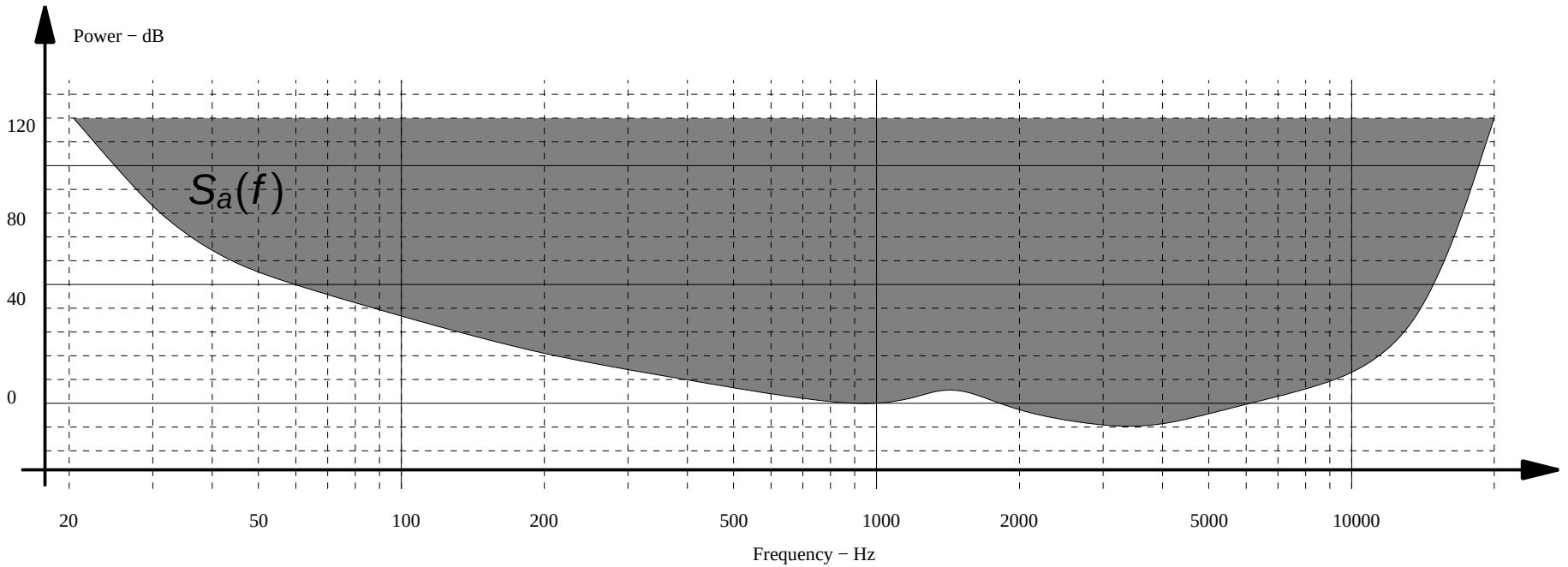
Pure tone perception

- Pure tone : $x(t) = a\sin(2\pi f_1 t)$ sinusoid with power $\sigma^2 = \frac{a^2}{2}$
- This sound excites several nerves (power spreading)
- Model : filterbank with M filters
  - The k-th filter corresponds to the k-th nervous fiber
  - Frequency response of the k-th filter: $H_k(f) = A_k(f)\exp^{j\phi_k(f)}$
  - Respons to the f_1 sinusoid:

$$y_k(t) = aA_k(f_1)\sin\left[2\pi f_1 t + \phi_k(f_1)\right]$$

- The power ratio is called spreading function: $S_E(k) = A_k^2(f_1)$

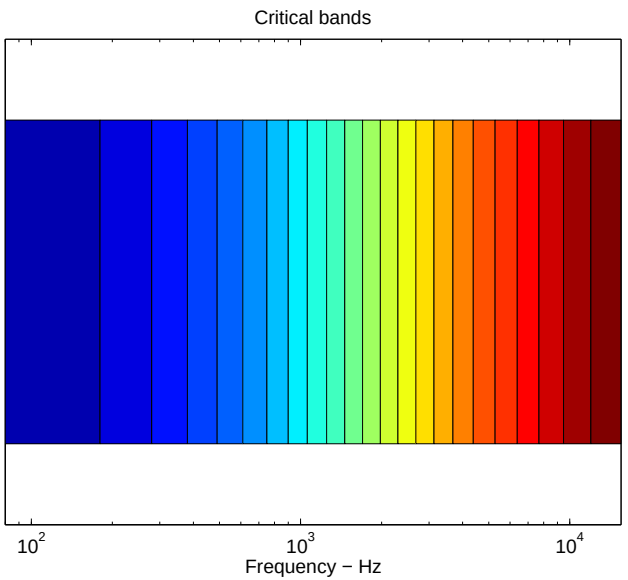
Hearing threshold



- The minimum power minimale for a tone at frequency f to be audible is $S_a(f)$
- $S_a(f)$ is a function of f and has a minimum between 1 and 4kHz (speech)

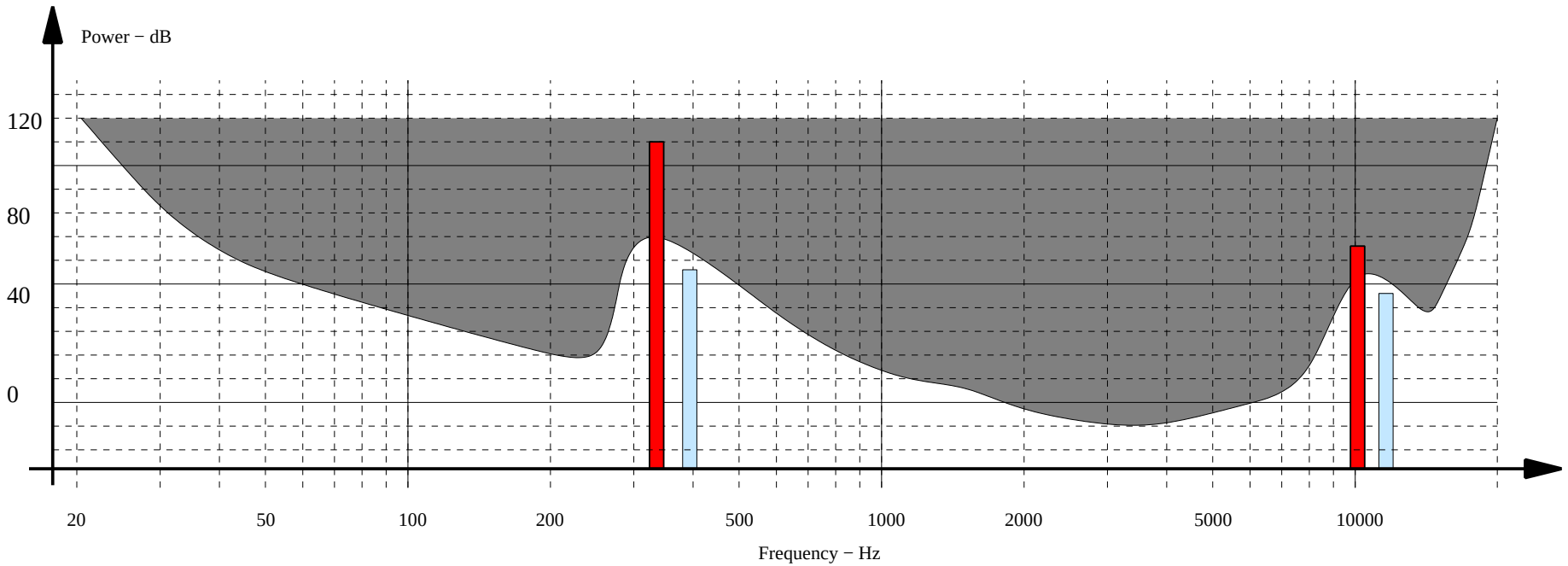
Critical band (CB)

- A pure tone at frequency f_1 must have a minimum power $\sigma_1^2 > S_a(f_1)$ to be heared
- For N sinusoids *near* f_1 we need only $\sum_i \sigma_i^2 > S_a(f_1)$
- Sinusoids are *near* if they are in the *critical band*
- CB amplitude is a function of f_1



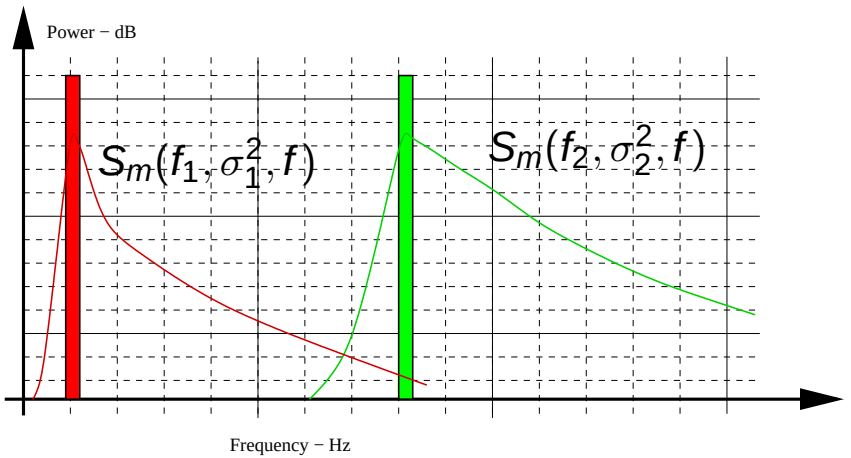
Masking curves

Frequency masking



- We define $S_m(f_0, \sigma^2, f)$ the minumum power of a pure tone at frequency f to be audible when a pure sound at frequency f_0 with power σ^2 is played, with $\sigma^2 > S_a(f_0)$
- Similar masking curve is observed for narrowband sounds

Frquency masking functions $S_m(f_0, \sigma^2, f)$

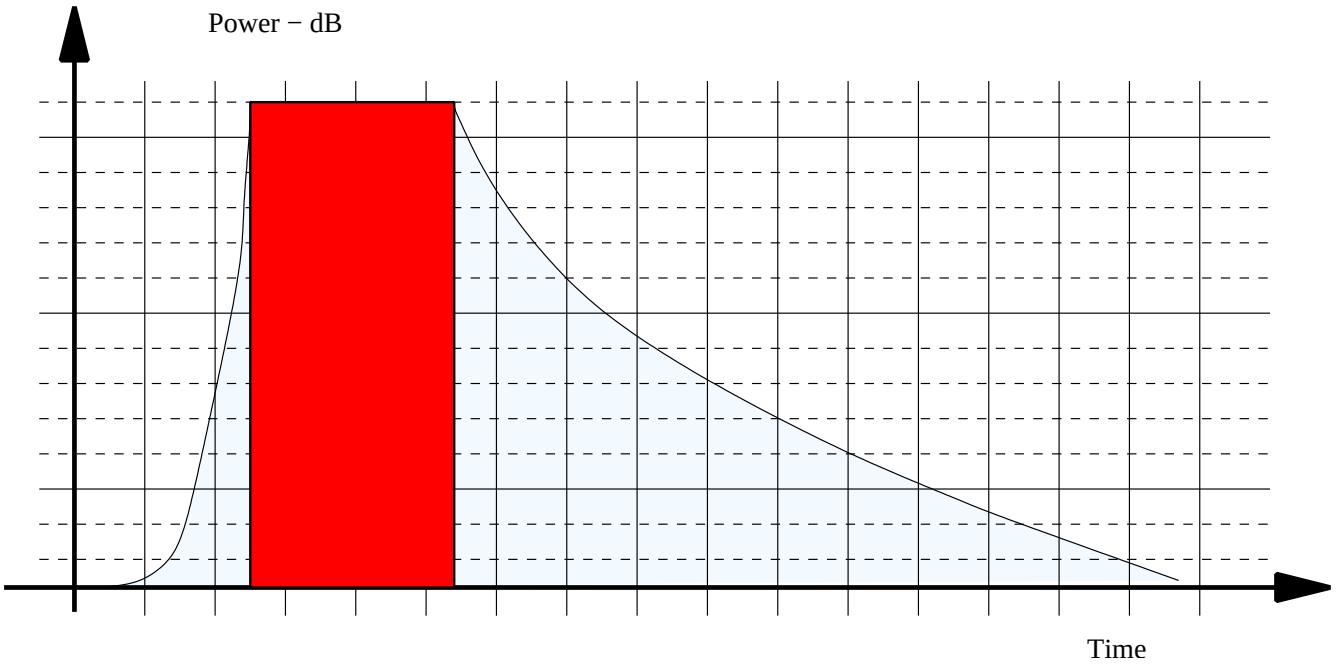


- For given f_0 adn σ^2 , $S_m(f)$ has a triangular shape
- The maximum is achieved at $f = f_0$
- Masking index: $S_m(f, \sigma^2, f) - \sigma^2$

- we observe that the second sound has not necessarily to by more powerful than the first to be heared: $S_m(f, \sigma^2, f) < \sigma^2$
- Decrease is slower for increasing f_1
- Decrease slope is proportional to CB
- Slope (toward increasing frequencies) is a decrasing function of σ^2

Masking curves

Time masking



- Pre-Masking : 2÷5 ms
- Post-Masking : 100÷200 ms

Model application

- The psychoacustical model allows to find out some non-audible parts of the signal
- We may then allow some quantization noise, as soon as it is masked by the rest of the signal
- Nevertheless, the model is not perfect:
  - Only pure tones or narrowband sounds are considered
  - We are only able to assess the influence of 3 sounds at a time
  - Real-life signals are much more complex: how do they interact?
- In practice, compression algorithms parameters are determined in an experimental way, after a large number of tests

Outline

The human visual system

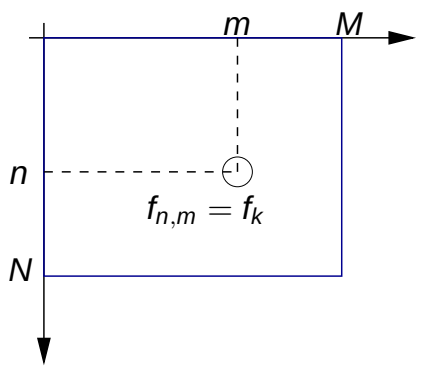
The sound perception

Image and video representation

Compression principles

Gray level images

- Discrete grid, $N \times M$ pixels
- A given position (m, n) is scanned in raster order k
- $k = (n-1)M + m$
- $f_{n,m} = f_k$



RGB representation

Color images have three components, each represented as a gray scale image.




YUV representation

Color images: one luminance component and two chrominance components

Color Sampling

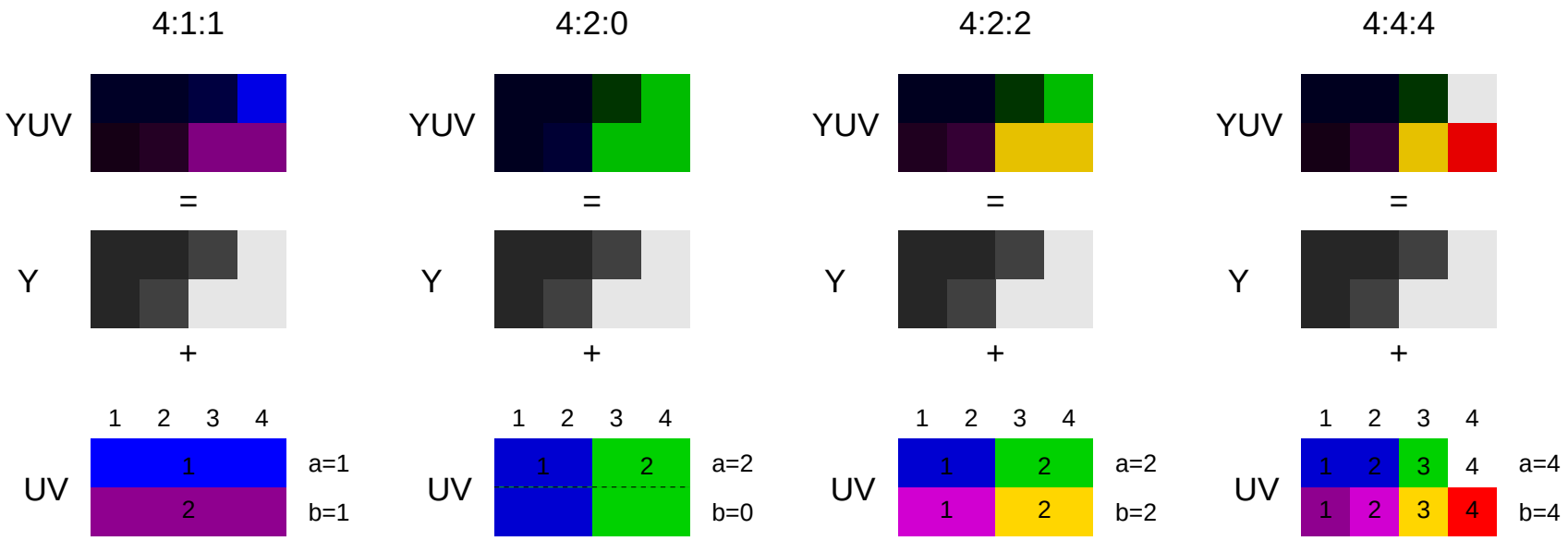
Sampling schemes

- The sampling scheme is represented with three integers

$$J : a : b$$

J Refernce horizontal size, usually equal to 4
a Number of chroma sample on the first line of the reference pattern
b Number of further sumples on the second line of the reference pattern

Color sampling



1/4 horizontal resolution
Full vertical resolution

1/2 horizontal resolution
1/2 vertical resolution

1/2 horizontal resolution
Full vertical resolution

Full horizontal resolution
Full vertical resolution

Quantization

- Samples represente on a discrete set
- Uniform quantization (rounding)
- L = number of levels
- $b = \log_2 L$ quantizer dynamics
- Typically $b = 8$ per component
  - 256 gray levels (8 bpp)
  - 16M colors (24 bpp)
- High dynamics range : 32 to 64 bits per channel

Resolution

SECAM	384×576	50 Hz
PAL	450×576	50 Hz
NTSC	323×486	60 Hz
QCIF	144×176	N/A
CIF	288×352	N/A
4CIF	576×704	N/A
SD/PAL	720×576	50 Hz
HD 720p/i	1280×720	50/100 Hz
HD 1080p/i	1920×1080	50/100 Hz
2K	2048×1556	24 Hz
4K	4096×2160	24 Hz
UHD	7680×4320	60 Hz

Representation of digital video

- Sequence of digital images
- Time dependency
- Three color components
- RGB or YUV representation
- Subsampling of color components

$$I : (n, m, T, c) \rightarrow x \in \left\{0, 1, \ldots, 2^b\right\}$$

Outline

The human visual system

The sound perception

Image and video representation

Compression principles

Compression: Motivations

- HD DVB System
  1 luminance component 1920×1080
  2 chrominance components 960×540
  8 bits quantization
  25 images per second
  $R \approx 622$ Mbps
- 2-hours movie ≈ 560 GB

Compression fundamentals

Why is it possible to compress?

- Statistical redundancy
  - images are spatially homogeneous
  - successive images are similar one to another
- Psychovisual redundancy
  - Spatial frequency sensitivity
  - Masking effects
  - Contours importance
  - Other limits of the HVS
- A compression algorithm should take into account both kinds of redundancy ot maximize its performance

Lossless and lossy algorithms

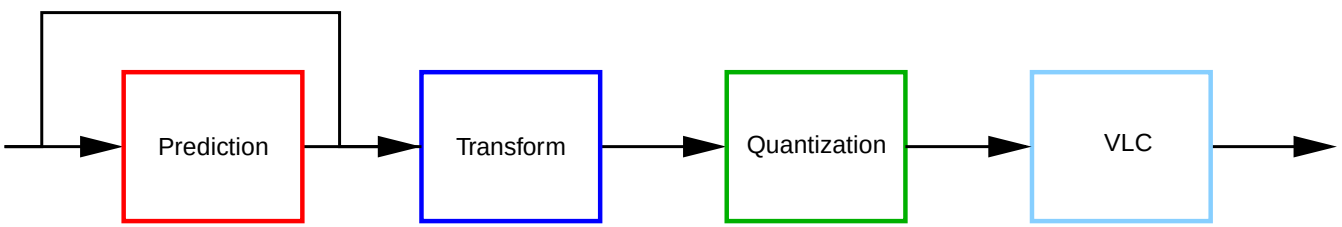
- Lossless algorithms
  - Perfect reconstruction
  - Based on statistics
  - Small compression ratio
- Lossy algorithms
  - Decoded $\neq$ original
  - Based on quantization
  - Psychovisual redundancy: “visually lossless”
  - High compression ratio

Symmetric and asymmetric algorithms (video)

- Symmetric algorithms
  - Same complexity for encoder and decoder
  - No motion estimation/compensation
  - Low compression ratio
  - Possibly real-time
- Asymmetric algorithms
  - Encoder (much) more complex than decoder
  - Motion Estimation/Compensation
  - High compression ratio
  - Typically “off line”, or hardware implementations

Basic tools for compression

- Transform
  - It concentrates information in a few coefficients
- Prediction
  - Alternative (and sometimes additional) method for information concentration
- Quantization
  - Rate reduction: rough representation of less important coefficients
- Lossless coding, or variable length coding (VLC)
  - Residual redundancy reduction



Rate

Compression ratio

- $T = \frac{B_{\text{in}}}{B_{\text{out}}} = \frac{R_{\text{in}}}{R_{\text{out}}}$

Coding rate

- Image : $R = \frac{B_{\text{out}}}{NM}$ [bpp]
- Video, audio : $R = \frac{B_{\text{out}}}{T}$ [bps]

Losslelss image coding: $T \leq 3$
Lossy image coding: $T \approx 5 \rightarrow ?$
Lossy vide ocoding: $T \approx 20 \rightarrow ?$

Quality and distortion

Criteria for image quality evaluation

- **Ojective criteria** are mathematical functions of:
  - $f_{n,m}$: original; and
  - $\widetilde{f}_{n,m}$: decoded image
- Non-perceptual objective criteria
  - Do not take into account the HVS characteristics
- Perceptual objective criteria
  - Based on perception models

Criteria for image quality evaluation

Non-perceptual objective criteria (NP-OC)

- Error image: $\mathcal{E}(f, \widetilde{f}) = f - \widetilde{f}$
- **Mean Square Error** (MSE) $\mathcal{D}$:

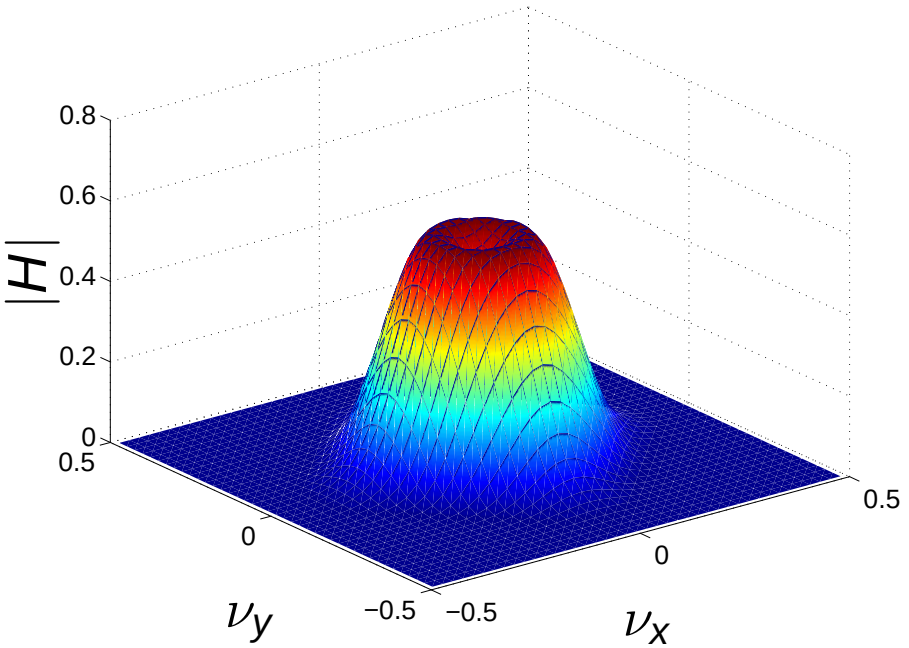
$$\mathcal{D}(f, \widetilde{f}) = \frac{1}{N\,M} \|\mathcal{E}\|^2 = \frac{1}{N\,M} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathcal{E}_{n,m}^2$$

- **Peak signal-to-noise ratio** :
  $\mathrm{PSNR}(f, \widetilde{f}) = 10 \log_{10} \left(\frac{255^2}{\mathcal{D}(f, \widetilde{f})} \right)$
- Simple, derivable, linked to the $\mathcal{L}^2$ norm

Criteria for image quality evaluation

Perceptual objective criteria

Weighted PSNR : Given a frequency weighting function (filter) h :



$$\mathrm{WPSNR}(f,\widetilde{f}) = 10\ \log_{10}\left(\frac{255^2}{\mathcal{D}_{\mathrm{W}}(f,\widetilde{f})}\right) \qquad \text{où}$$

$$\mathcal{D}_{\mathrm{W}}(f,\widetilde{f}) = \frac{1}{N\,M}\,\|h * \mathcal{E}\|^2$$

Criteria for image quality evaluation

Perceptual objective criteria

Structural Similarity Index (SSIM Index) between two blocks x et y :

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \qquad \text{Luminance}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \qquad \text{Contraste}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \qquad \text{Structure}$$

For thesake of simplicity, $\alpha = \beta = \gamma = 1$, $C_3 = C_2/2$

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

SSIM between images is the average of blocks' SSIM

Criteria for image quality evaluation

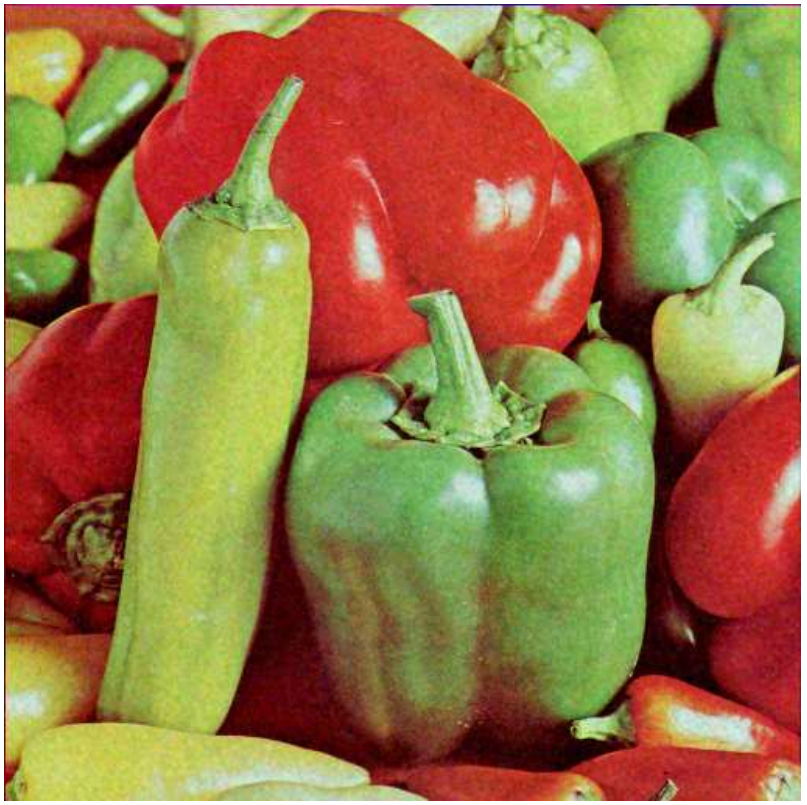
Subjective criteria (SC)

- Subjective criteria are based on the image assessment performed by human observers
  - Difficulty of HVS modelling for objective criteria
  - Statistical analysis of results
  - Long, difficult and costly evaluations
- Often NP-OC are used
  - Simplicity
  - Geometrical interpretation (norm)
  - analytical Optimisation
  - Correlation avec SC?

Error perception

Distributed error, white noise $\sigma = 4$

MSE: 16



SSIM: 0.906

Error perception

Error concentrated over 100×100 pixels

MSE: 16

SSIM: 0.972

Error perception

Error concentrated on the contours (estimation by Sobel's filter)

MSE: 16

SSIM: 0.987

Error perception

Noise on high spatial frequencies

MSE: 16

SSIM: 0.882

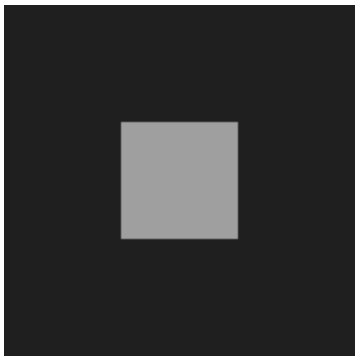
Error perception

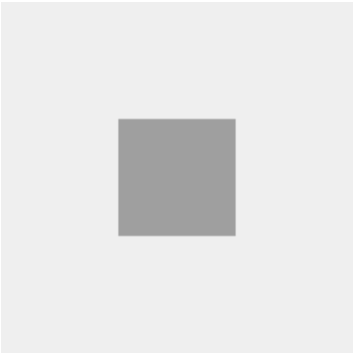
Chroma subsampling

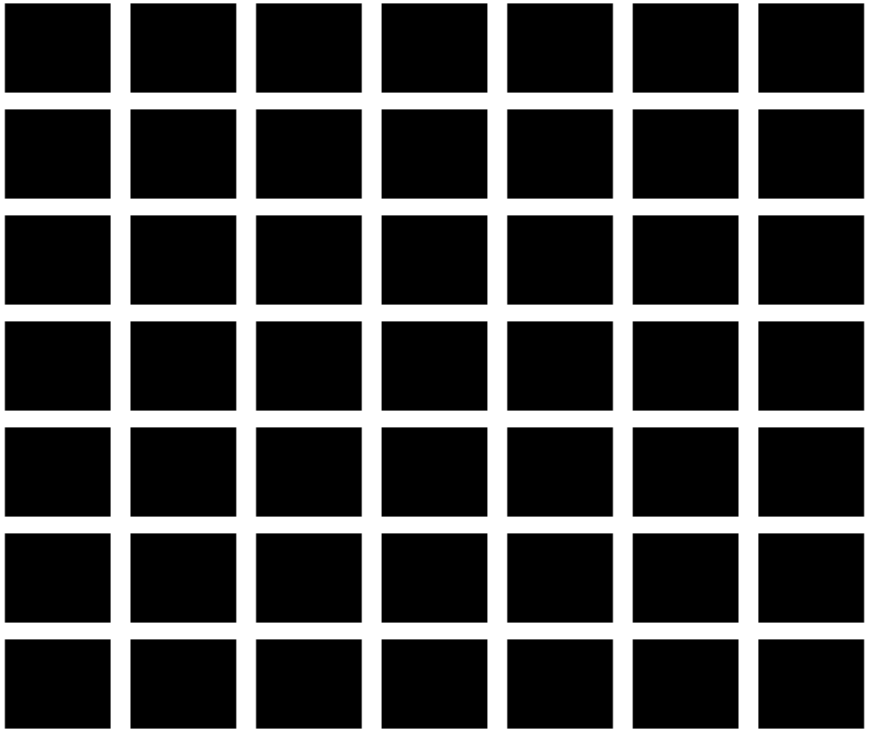
MSE: 21.27

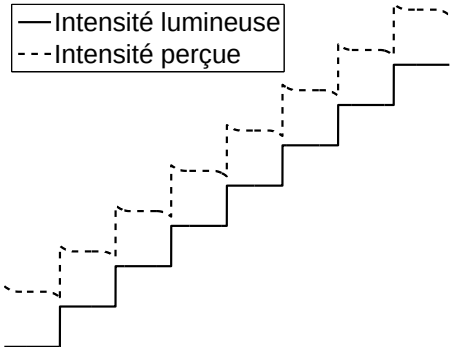
SSIM: —

Spatial effects



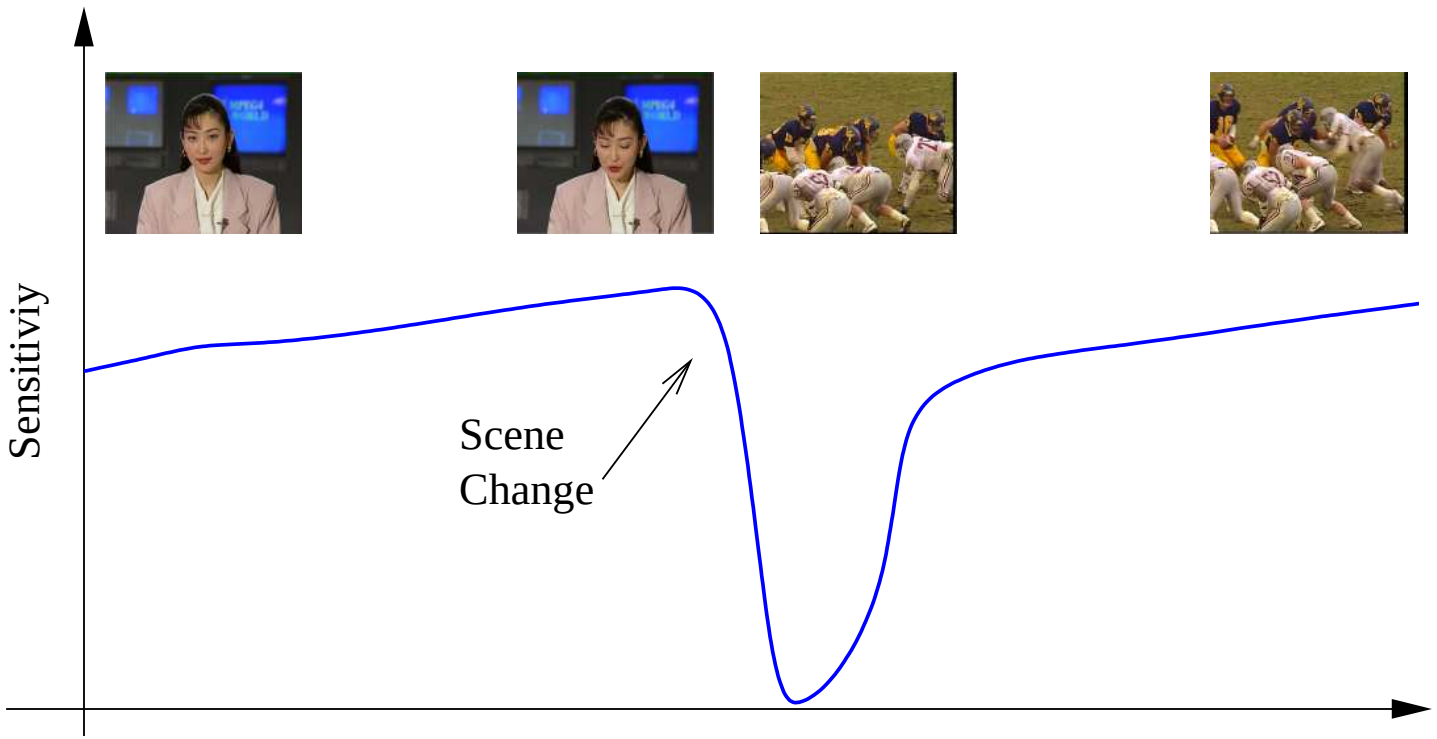






Video Perception

- Sensitivity spatio-temporal frequencies
- Spatial and time masking



Perception and quality: summary

- Perceptual models are needed in order to achieve good compression performance
- Auditive system is relatively well understood and exploited in audio coders (see later on)
- HVS less well understood
- No P-OC totally reliable
- Nevertheless, the best performance can be achieved only taking into account HVS

Complexity, delay and robustness

- The complexity of a compression algorithm may be limited by:
  - Real-time constraints
  - Hardware limitations
  - Economical cost
- Delay is usually measured at the encoder
  - It is related to complexity ...
  - ... but mainly affected by the *coding order*
- Robustness: sensitivity of the compressed stream to losses

Performance criteria: summary

Contrasting issues:

$$\begin{matrix} \Uparrow \text{Quality} \\ \Uparrow \text{Robustness} \end{matrix} \quad \begin{matrix} \Downarrow \text{Rate} \\ \Downarrow \text{Complexity} \\ \Downarrow \text{Delay} \end{matrix}$$