Information and Coding

Armando J. Pinho

Departamento de Electrónica, Telecomunicações e Informática

Universidade de Aveiro

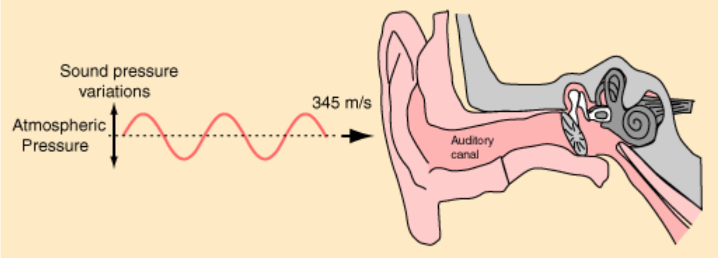
ap@ua.pt

Contents

1. Perceptual redundancy: auditory system
2. Some audio coding standards
3. Perceptual redundancy: visual system
4. Transform coding
5. Video coding standards

The human auditory system

- Humans perceive **sound** by the sense of hearing. By sound, we commonly mean the vibrations that travel through air and are audible to humans.



The human auditory system

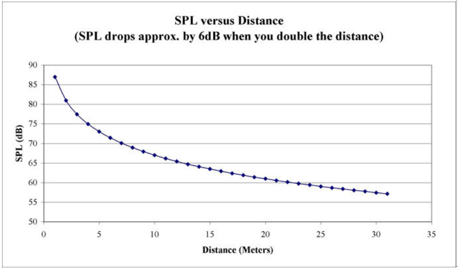
- **Audio** is the electrical representation of sound.
- Generally, humans can perceive variations in sound pressure from 16-20 Hz to 20-22 kHz.
- However, our capacity for perceiving sounds of very small amplitude varies according to frequency, being maximum between 2 and 4 kHz.
- The human voice produces frequencies approximately between 200 Hz and 8 kHz. Telephone communications limit this range from 300 Hz to 3.4 kHz (200 Hz to 3.2 kHz in the USA) .

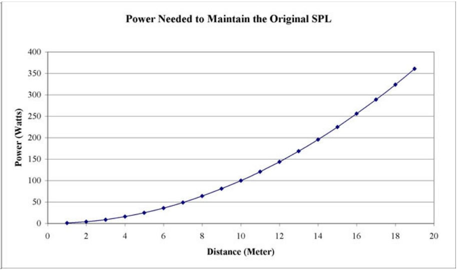
The human auditory system

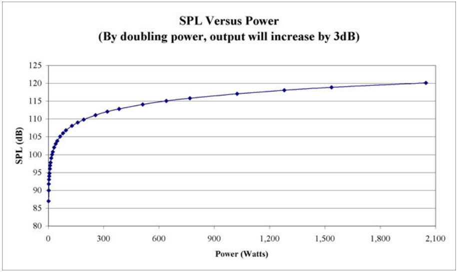
- Normally, the amplitude range that we can hear is about 100 dB:

Source of sound	Sound pressure level (dB)
Jet engine at 30 m	150
Jet engine at 100 m	140
Threshold of pain	125–130
Hearing damage (short-term exposure)	120
Maximum output of some MP3 players	110
Hearing damage (long-term exposure)	100
Major road at 10 m	80–90
TV (at home level) at 1 m	60
Normal talking at 1 m	40–60
Very calm room	20–30
Calm breathing	10
Auditory threshold at 2 kHz	0

The human auditory system

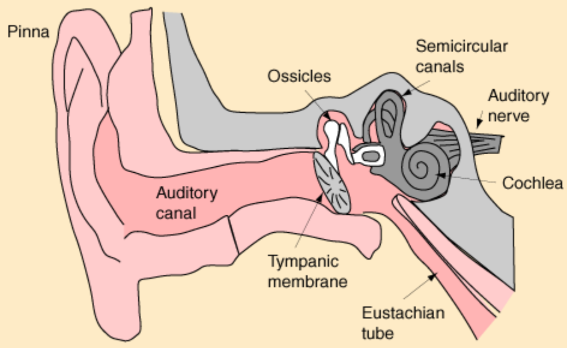






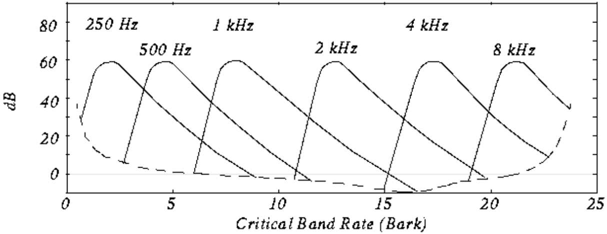
The human auditory system

- The auditory system can roughly be described as a **bandpass** filter-bank, consisting of strongly overlapping bandpass filters.



The human auditory system

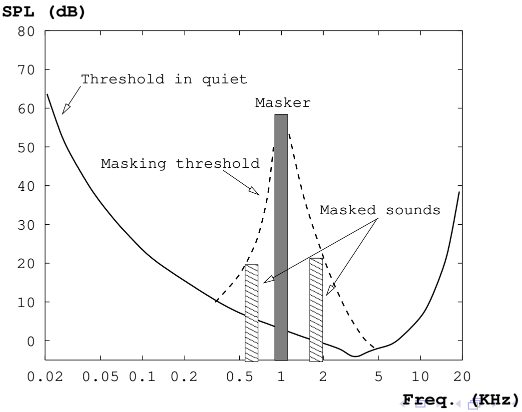
- These “filters” have bandwidths in the order of 50 to 100 Hz for signals below 500 Hz and up to 5000 Hz for signals at high frequencies.



- They are called **critical bands**.
- Twenty-five critical bands, covering frequencies of up to 20 kHz, are normally taken into account.

The human auditory system

- **Simultaneous masking** is a frequency domain phenomenon where a low-level signal (maskee) can be made inaudible (masked) by a simultaneously and close in frequency stronger signal (masker).

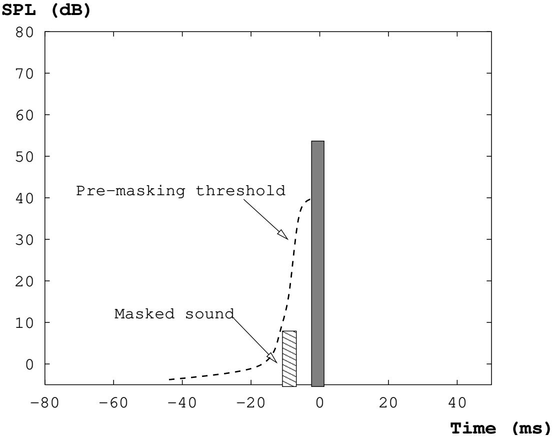


The human auditory system

- In addition to simultaneous masking, the time domain phenomenon of **temporal masking** plays an important role in human auditory perception.
- Temporal masking may occur when two sounds appear within a small interval of time.
- Depending on the individual sound pressure levels, the stronger sound may mask the weaker one, even if the maskee precedes the masker…

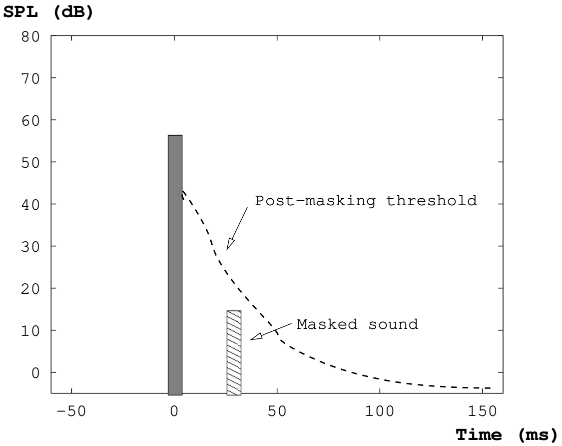
The human auditory system

- The **pre-masking** has a duration of about 5 to 20 ms:



The human auditory system

- The **post-masking** has a duration of about 50 to 200 ms:



Contents

1 Perceptual redundancy: auditory system

2 Some audio coding standards

3 Perceptual redundancy: visual system

4 Transform coding

5 Video coding standards

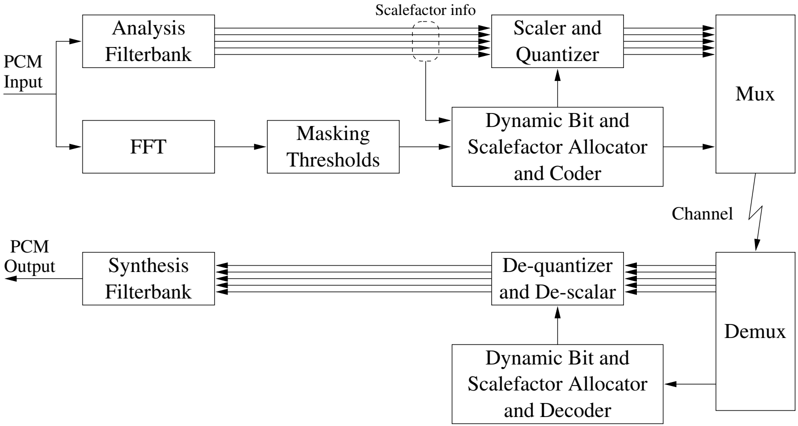
MPEG–1

- MPEG–1 audio coding is organized in three layers, I, II and III, with increasing performance, but also complexity and delay.
- It allows sampling frequencies of 32, 44.1 and 48 kHz, and bitrates between 32 kb/s (mono) and 448 kb/s (Layer I), 384 kb/s (Layer II) and 320 kb/s (Layer III).
- In terms of transparent CD (stereo) quality, the bitrates and compression rates are, approximately,

Layer	Bitrate	Compression rate
I	384 kb/s	4
II	192 kb/s	8
III	128 kb/s (VBR)	12

MPEG–1

MPEG–1 layer I and II



MPEG–1

- The analysis filterbank has 32 subbands, equally spaced in frequency.
- Each block is formed of 384 audio samples (8 ms for $f_s = 48$ kHz), meaning that each subband contains 12 samples.
- For $f_s = 48$ kHz, the width of each subband is 750 Hz.
- Usually, the bits are dynamically assigned to the coefficients of the subbands according to a psychoacoustic model (how to do this is not part of the standard).
- For each block of 12 coefficients (subband), an uniform quantizer (from 15 available) is selected, according to predefined levels of quality and compression.

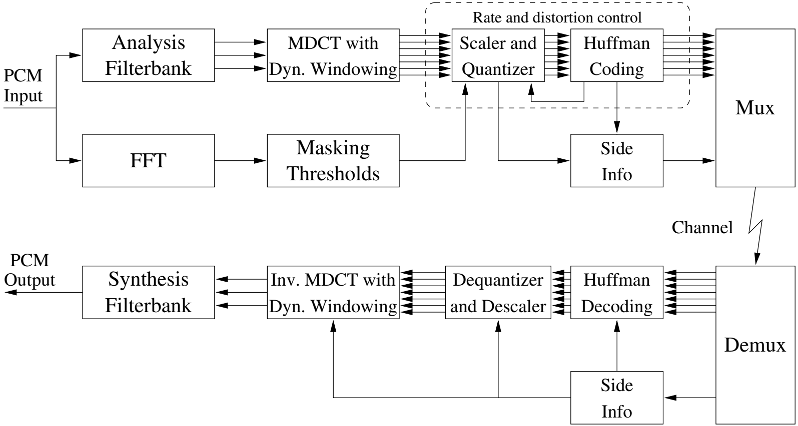
MPEG–1

MPEG–1 layer III

- MPEG–1 layer III has several differences in relation to the other two layers, being much more complex.
- It is based on hybrid coding: subband and transform.
- It allows **variable bitrate coding (VBR)**.
- It relies on a technique designated analysis by synthesis for dynamic bit assignment.
- It uses an advanced **pre-echo control**.
- It uses non-uniform quantization and statistical coding.

MPEG–1

MPEG–1 layer III



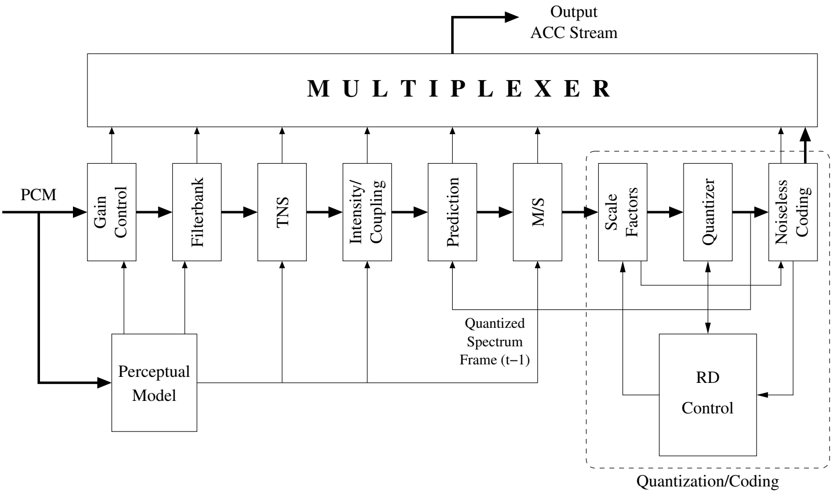
MPEG–2

- The MPEG–2 audio coding standard includes the MPEG–1 audio and introduces extensions for multi-channel configurations.
- MPEG–2 provides two audio coding standards: one forward and backward compatible with MPEG–1, the other incompatible.
- Forward compatibility means that a multi-channel decoder understands MPEG–1 mono and stereo streams.
- Backward compatibility means that a MPEG–1 decoder is able to extract stereo audio from a MPEG–2 multi-channel stream.
- The second standard is MPEG–2 AAC (Advanced Audio Coding).

MPEG–2 AAC

- Some parts are identical to those of MPEG–1/2 layer III:
  - Filterbank with dynamic windowing.
  - Non-uniform quantizers.
  - Huffman coding.
- MPEG–2 AAC defines 3 coding profiles:
  - The low complexity profile.
  - The main profile.
  - The scalable sampling rate profile.

MPEG–2 AAC


Output
ACC Stream
M U L T I P L E X E R
PCM
Gain Control
Filterbank
TNS
Intensity/ Coupling
Prediction
M/S
Scale Factors
Quantizer
Noiseless Coding
Perceptual Model
Quantized Spectrum Frame (t−1)
RD Control
Quantization/Coding

MPEG–4

- The MPEG–4 coding standard provides tools for coding audio objects, such as **natural audio** (for example, speech and music) and **synthetic audio**, aiming several applications:
  - Telephone over the Internet.
  - High quality music.
  - Text-to-speech conversion.
  - Synthesized music.
  - ...
- The synthesized audio can be obtained through text (TTS) or instrumental descriptions.

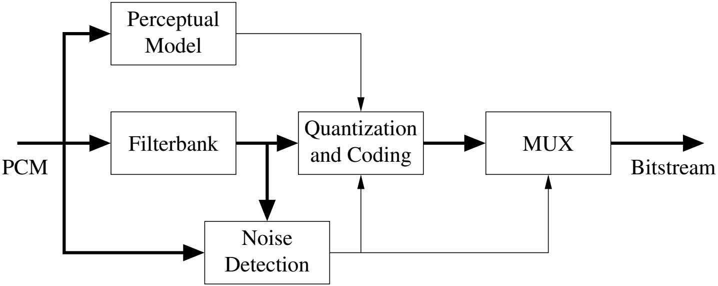
MPEG–4

- The encoding of the **natural audio** relies on several techniques:
  - Harmonic vector excitation coding (HVXC), for $f_s \leq 8$ kHz, and bitrates between 2 and 4 kbp/s (until 1.2 kbp/s, for VBR).
  - Code excited linear predictive (CELP), for $8 \leq f_s \leq 16$ kHz, and bitrates between 4 and 24 kbp/s.
  - Transform-domain weighted interleave vector quantization (TwinVQ) and AAC, for $f_s \geq 8$ kHz, and bitrates greater than 6 kbp/s.

MPEG–4

Perceptual noise substitution

- Produces a perceptual equivalent signal, instead of trying to reproduce the original waveform.
- It is used for audio components similar to noise.
- If this type of signal is detected in a certain band, then instead of coding the coefficients, it is encoded their total power.



MPEG–4

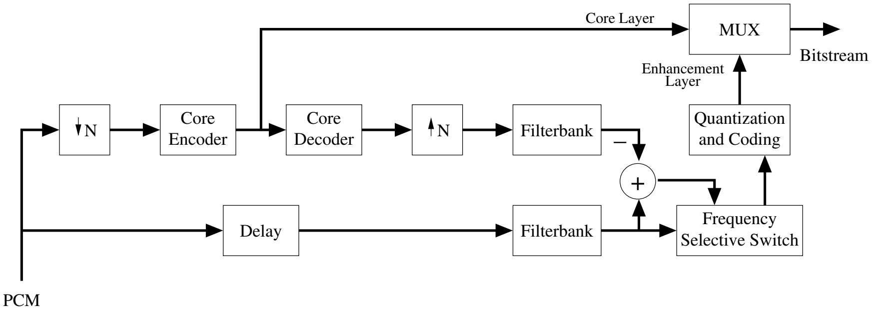
Low-delay coding

- Mode used in bi-directional, real-time, communications, where long delays are not acceptable (the "standard" encoder may introduce delays of several hundreds of ms).
- It uses windows with half the normal size.
- It does not use dynamic window adaptation (this implies long delays).
- The bit reservoir is minimized or even eliminated.

MPEG–4

Scalable coding

- The MPEG–4 allows a limited number of layers (typically 2 to 4). This is the **large-step scalability**.



Contents

1 Perceptual redundancy: auditory system

2 Some audio coding standards

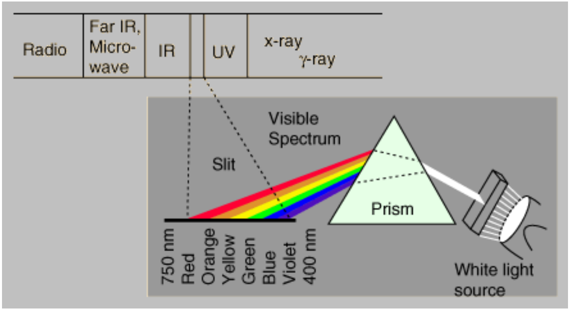
3 Perceptual redundancy: visual system

4 Transform coding

5 Video coding standards

The visible spectrum

- The typical human eye senses electromagnetic wavelengths between 400 and 700 nm, and has maximum sensitivity around the 555 nm (green zone).

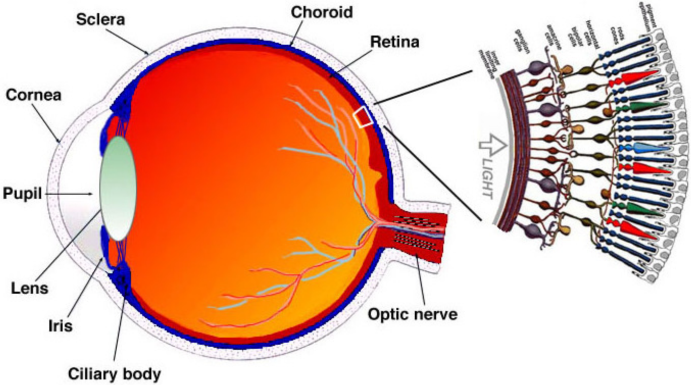


The human perception of color

- Normally, the characteristics that allow colors to be distinguished are:
  - The **brightness** (how bright is the color).
  - The **hue** (the dominant color).
  - The **saturation** (how pure is the color).
- Together, the hue and the saturation define the **chromaticity**.
- Therefore, a color can be characterized by the brightness and the chromaticity.

The human perception of color

- The human eye has **photoreceptors** that are sensitive to short wavelengths (S), medium wavelengths (M) and long wavelengths (L), also known as the blue, green and red photoreceptors.



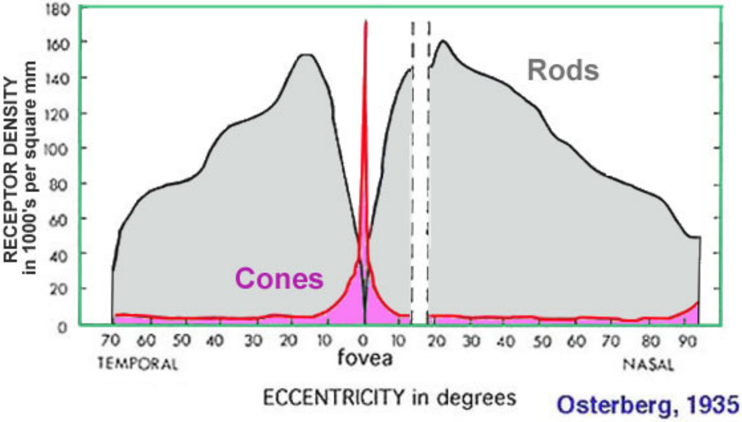
The photoreceptors: cones and rods

- The **cones**:
  - They provide the photopic vision.
  - They are between 6 and 7 million.
  - They are responsible for the perception of color.
  - There are three types:
    - Sensitive to the blue ($\approx$ 2%)
    - Sensitive to the green ($\approx$ 33%)
    - Sensitive to the red ($\approx$ 65%)
  - They are positioned mainly in the central part of the retina (fovea $\approx$ 0.3 mm diameter).

The photoreceptors: cones and rods

- The **rods**:
  - They provide the scotopic vision (under low light conditions).
  - They are between 75 and 150 million.
  - They are much more sensitive than the cones, but they are unable to distinguish colors.
  - They allow vision at low levels of light.
  - Because several rods are connected to the same nerve, they provide less spatial resolution.

Spatial distribution of the photoreceptors



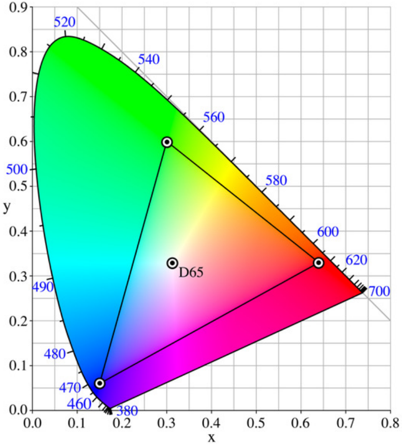
Additive primaries

- The red, green and blue are the three additive primary colors.

- Adding these three colors produces white.

The standard *RGB* (*sRGB*) color space

- Chromaticity diagram and corresponding color gamut of *sRGB* (proposed by HP and Microsoft):



The *sRGB* color space

R component

G component

B component

The *CMY* color space

- *CMY* is based on the subtractive properties of inks.
- The cyan, magenta and yellow are the **subtractive primaries**. They are the complements, respectively, of the red, green and blue. For example, the cyan subtracts the red from the white.

- Conversion from *RGB* to *CMY*: $C = 1 - R$, $M = 1 - G$, $Y = 1 - B$.

The *CMY* color space

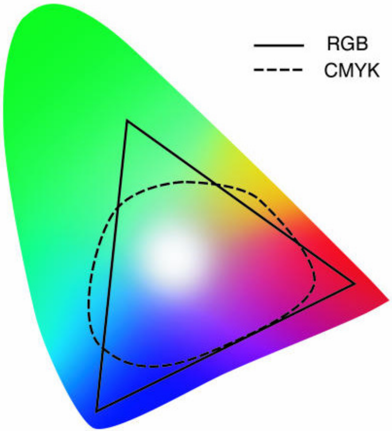
C component

M component

Y component

The *CMYK* color space

- Due to technological difficulties regarding the reproduction of black, it is generally used the *CMYK* color space for printing.



The *CMYK* color space

C component

M component

Y component

K component

The YUV and YC_bC_r color spaces

- The YUV color space is used by the PAL television standard.
- Y is the luminance component:

$$Y = \quad 0.299R + 0.587G + 0.114B$$

- Components U and V represent the chrominance:

$$\begin{aligned} U &= -0.147R - 0.289G + 0.436B = 0.492(B - Y) \\ V &= \quad 0.615R - 0.515G - 0.100B = 0.877(R - Y) \end{aligned}$$

- YC_bC_r is a family of color spaces, related to the YUV color space, used mainly in **digital image/video** systems.

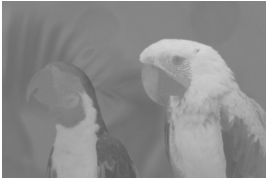
Some advantages of the *YUV* color space

- The *YUV* color space allowed to maintain the compatibility with the old "black and white" television receivers.
- The human eye is more sensitive in the green zone, which is represented mainly by the Y component (the U and V components are related to the blue and red.
- Because the human eye is less sensitive to the blue and red, it is possible to reduce the bandwidth used to represent the U and V components, without introducing significant perceptual degradation.

The YC_bC_r color space

Y component

C_b component



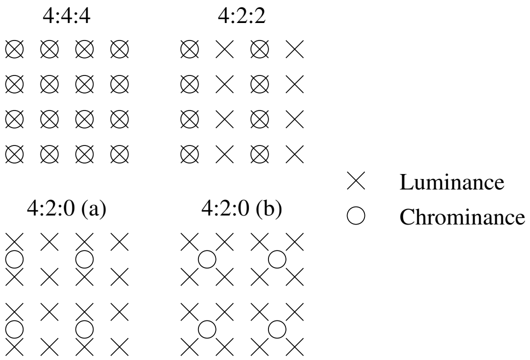
C_r component

Chrominance sub-sampling

- The YUV or YC_bC_r color spaces separate the chrominance component (UV / C_bC_r) from the luminance component (Y).
- The human eye is more sensitive to the greens, which are represented mainly by the Y component.
- For this reason, it is common to sub-sample the chrominance components UV / C_bC_r, producing a reduction in the data rate.
- This reduction is used by both the video coding standards (H.261, MPEG-1, MPEG-2, ...) and the image coding standards (JPEG).

Chrominance sub-sampling

- The most common types of chrominance sub-sampling:



- The 4:2:0 mode has two variants: (a) used by MPEG-2; (b) used by JPEG, MPEG-1, H.261,. . .

Example YUV 4:2:0


RGB


YC_bC_r 4:2:0

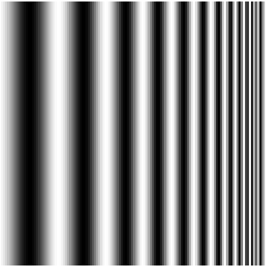

Y component

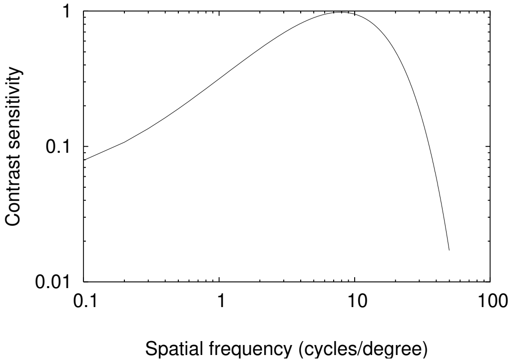

C_b component


C_r component

Spatial frequency

- The human visual system is characterized by a **bandpass** behavior in the spatial frequency domain:



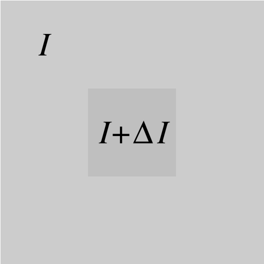


Weber's law

- Non-linear response to the light intensity (Weber's law):

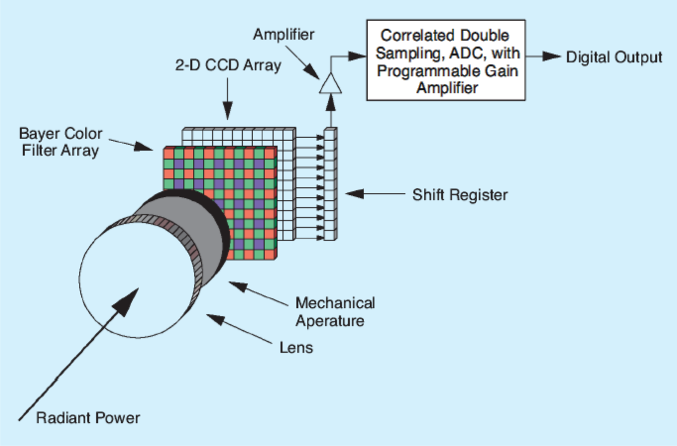
$$\frac{\Delta I}{I} \approx d(\log I) \approx \text{const.}$$

where ΔI represents the minimum intensity variation that can be perceived on a background of intensity I.

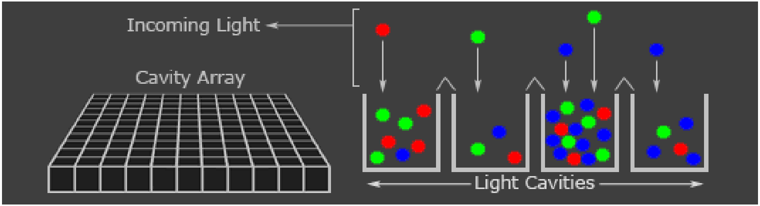


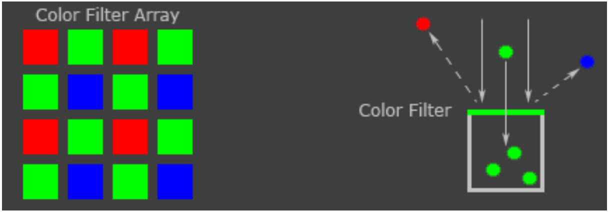
Digital camera

- Image acquisition using a digital camera:
  (IEEE SP Magazine, Jan 2005)

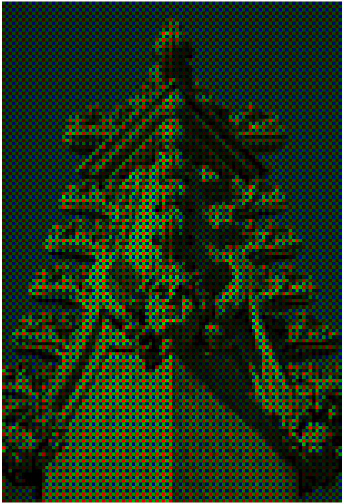


The Bayer matrix





The Bayer matrix



Contents

1 Perceptual redundancy: auditory system

2 Some audio coding standards

3 Perceptual redundancy: visual system

4 Transform coding

5 Video coding standards

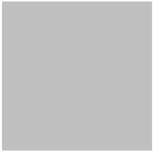
Motivation

- The main objective of using transforms in the context of data compression is to convert the original data into a new data set **more simple to quantize and encode**.
- Transforms are used **to reduce the statistical dependencies** among the original data (Ideally, the resulting coefficients should be statistically independent).
- Transforms are also used **to separate the relevant information from the irrelevant**, in order to permit course quantization or even removal of the irrelevant information.

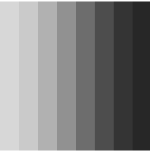
The DCT 2D

- The DCT 2D is obtained through a separable extension of the 1D version:

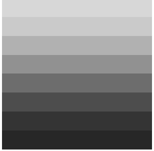
$$S_{q_1,q_2} = (\mathbf{s}_{q_1,q_2})_{p_1,p_2} = c_{q_1} c_{q_2} \cos\left(2\pi f_{q_1}(p_1 + 0.5)\right) \cos\left(2\pi f_{q_2}(p_2 + 0.5)\right)$$



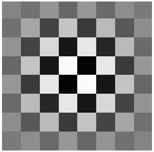
$S_{0,0}$



$S_{0,1}$



$S_{1,0}$



$S_{7,7}$

- For calculating a single coefficient of a non-separable transform we need m^2 operations.
- In the separable case, only $2m$ operations are required.

The DCT 2D

- Let us see how the DCT attains energy compaction, considering, for example, the following 8×8 block of pixels:

$$\mathbf{X} = \begin{bmatrix} 183 & 160 & 94 & 153 & 194 & 163 & 132 & 165 \\ 183 & 153 & 116 & 176 & 187 & 166 & 130 & 169 \\ 179 & 168 & 171 & 182 & 179 & 170 & 131 & 167 \\ 177 & 177 & 179 & 177 & 179 & 165 & 131 & 167 \\ 178 & 178 & 179 & 176 & 182 & 164 & 130 & 171 \\ 179 & 180 & 180 & 179 & 183 & 169 & 132 & 169 \\ 179 & 179 & 180 & 182 & 183 & 170 & 129 & 173 \\ 180 & 179 & 181 & 179 & 181 & 170 & 130 & 169 \end{bmatrix}$$

The DCT 2D

- The coefficients (rounded to the integers), resulting from applying the DCT to the block (after subtracting $2^7 = 128$ to each pixel), are:

$$\mathbf{Y} = \begin{bmatrix} 313 & 56 & -27 & 18 & 78 & -60 & 27 & -27 \\ -38 & -27 & 13 & 44 & 32 & -1 & -24 & -10 \\ -20 & -17 & 10 & 33 & 21 & -6 & -16 & -9 \\ -10 & -8 & 9 & 17 & 9 & -10 & -13 & 1 \\ -6 & 1 & 6 & 4 & -3 & -7 & -5 & 5 \\ 2 & 3 & 0 & -3 & -7 & -4 & 0 & 3 \\ 4 & 4 & -1 & -2 & -9 & 0 & 2 & 4 \\ 3 & 1 & 0 & -4 & -2 & -1 & 3 & 1 \end{bmatrix}$$

The JPEG standard

- The JPEG (Joint Photographic Experts Group) standard is a family of coding methods for images of continuous tones of grays or colors.
- The group was established in 1986, the standard was proposed in 1992 and approved in 1994 (ISO 10918-1).
- The JPEG standard comprises **four coding methods**: sequential, progressive, hierarchical and lossless.
- The JPEG standard is based on a number of compression techniques, such as the DCT, statistical coding and predictive coding.

The sequential mode of JPEG

- Every codec should include this mode in order to be considered JPEG-compatible (it is also known as the “baseline” mode).
- The sequential mode of JPEG comprises the following steps:
  - Calculation of the DCT.
  - Quantization of the DCT coefficients, in order to eliminate less relevant information, according to the characteristics of the human visual system.
  - Statistical coding (Huffman or arithmetic) of the quantized DCT coefficients.

The sequential mode of JPEG

- Calculation of the DCT:
  - The image is partitioned into 8×8 blocks of pixels. If the number of rows or columns is not multiple of 8, then they are internally adjusted (using padding).
  - Subtract 2^{b-1} to each pixel value, where b is the number of bits used to represent the pixels.
  - Calculate the DCT 2D of each block.

The sequential mode of JPEG

- Quantization of the DCT coefficients:
  - The DCT coefficients are quantized using a quantization matrix, Q, previously scaled by a compression quality factor.
  - Next, the coefficients are organized in a one-dimensional vector according to a zig-zag scan.
- Statistical coding:
  - The non-zero AC coefficients are encoded using Huffman or arithmetic coding, representing the value of the coefficient, as well as the number of zeros preceding it.
  - The DC coefficient of each block is predictively encoded in relation to the DC coefficient of the previous block.

Quantization of the coefficients

- Compression is obtained due to the **low-pass characteristic of the human visual system**.
- Because of this characteristic, generally more bits are assigned to the low frequencies (those appearing in the upper left corner of the transformed block).
- This is done using **threshold coding** (non-linear approximation).

Quantization of the coefficients

- Generally, the elements of Q are 8 bit integers that determine the quantization step according to the position of each coefficient.
- **Example:** quantization matrix of JPEG (luminance):

$$Q = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 129 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

Quantization of the coefficients

- By applying this quantization matrix to the block that we have previously used as example, we obtain the matrix $\widetilde{Y}$:

$$\widetilde{Y} = \begin{bmatrix} 20 & 5 & -3 & 1 & 3 & -2 & 1 & 0 \\ -3 & -2 & 1 & 2 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- In this example, 45 of the 64 coefficients are eliminated.

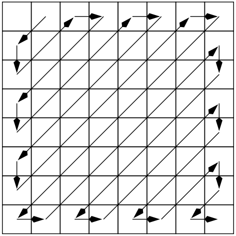
Quantization of the coefficients

- Because the sensitivity of the human eye to the colors is different from that of the luminance, JPEG provides a different quantization matrix for the chrominance components:

$$Q = \begin{bmatrix} 17 & 18 & 24 & 47 & 99 & 99 & 99 & 99 \\ 18 & 21 & 26 & 66 & 99 & 99 & 99 & 99 \\ 24 & 26 & 56 & 99 & 99 & 99 & 99 & 99 \\ 47 & 66 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \end{bmatrix}$$

Coefficient coding

- JPEG uses a zig-zag scanning of $\widetilde{Y}$ in order to encode the quantized coefficients, except for the $(0, 0)$ position, i.e., the DC coefficient.



- The objective of this scanning is to group together the zero coefficients, allowing a more efficient representation.
- This efficiency is obtained using a variant of run-length coding.

Coefficient coding

- Using again the same example, a JPEG encoder would generate the following codewords:

$$(0,5),(0,-3),(0,-1),(0,-2),(0,-3),(0,1),(0,1),(0,-1),(0,-1),$$
$$(2,1),(0,2),(0,3),(0,-2),(0,1),(0,1),(6,1),(0,1),(1,1),\mathrm{EOB}$$

$$\begin{bmatrix} 20 & 5 & -3 & 1 & 3 & -2 & 1 & 0 \\ -3 & -2 & 1 & 2 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The visual effect of the coding blocks

- The coding techniques that are based on a partition of the image into blocks are generally affected by a visual phenomenon known as the **blocking artifact**.

This artifact is more visible when the compression ratio is high and happens because the blocks are encoded independently (except for the DC coefficients).

Example: 8×8 DCT, 0.31 bpp.

Contents

1. Perceptual redundancy: auditory system
2. Some audio coding standards
3. Perceptual redundancy: visual system
4. Transform coding
5. Video coding standards

H.261

- H.261 (1990) is a ITU-T video coding standard (Video Codec for Audiovisual Services at p×64 kbit/s) that was developed with the aim of being used
  - In video-phone applications.
  - In video-conference applications.
  - Over ISDN links at $p \times 64$ kbps, $p = 1, \ldots, 30$.
- For example, $p = 1$ (64 kbps) would be appropriated for video-phone, where the video signal was transmitted at 48 kbps and the audio signal at 16 kbps.
- Generally, video-conference required better image quality, implying typically $p \geq 6$ (384 kbps).
- For $p = 30$ we have 1.92 Mbps, which was sufficient for a video quality similar to the old VHS tapes.

H.261

- The encoded stream has the following structure:
  - At the top, the **frame**.
  - Each frame is partitioned into several **groups of blocks**.
  - Each group of blocks is formed of several **macroblocks**.
  - The macroblock is the smallest region that can have a particular coding mode assigned to.
  - The macroblock is composed of four basic **blocks** (a basic block is 8×8) of luminance (Y) and by the corresponding 8×8 chrominance blocks (C_r and C_b).

H.261

- The H.261 uses two compression modes:
  - **Intraframe**: similar to the JPEG compression, i.e., relies on DCT applied to 8×8 blocks of pixels.
  - **Interframe**: temporal prediction (motion compensation), followed by DCT of the prediction residuals.
- Motion compensation (MC) is performed in macroblocks, within a search area of 15 pixels around the macroblock.
- It has 32 quantizers, one of them dedicated to the DC coefficient in intraframe mode (quantization step of 8). The others have quantization steps from 2 to 62.
- Statistical coding is performed with Huffman codes.

MPEG–1

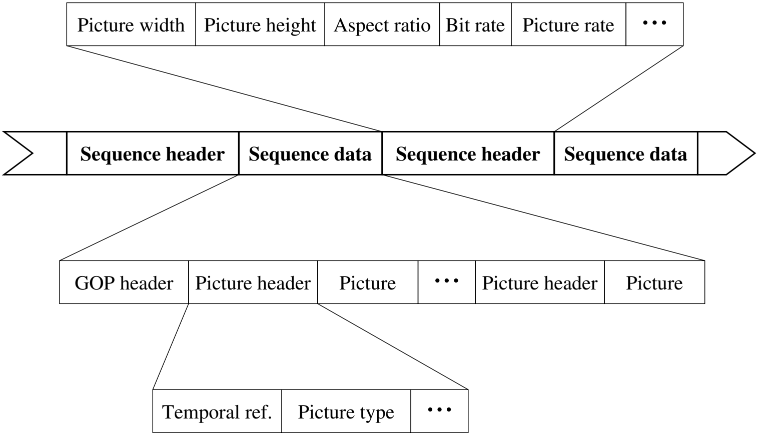
- MPEG–1 (1992) is a ISO/IEC (11172) coding standard that has been developed with the aim of storing video and audio in CD-ROMs.
- Target bitrates were around 1.5 Mbps, which was the bitrate associated to the early CD-ROM readers.
- The main objective of MPEG–1 was to provide means for encoding audio and video for interactive multimedia applications.
- For video segments having a moderate motion content, quality similar to VHS could be attained for MPEG–1 video at 1.2 Mbps.

MPEG–1

- The algorithms used in MPEG–1 are similar to those of H.261, although having some additional characteristics, such as
  - Random access (using type I frames)
  - Fast forward and reverse.
  - Backwards playing.
- Generally, the input is in the CCIR 601 format (576 $\times$ 720, for a 50 fps or 480 $\times$ 720 for 60 fps), and is converted to SIF (Source Input Format) before encoding (luma with 288(240) $\times$ 352 pixels and chroma with 144(120) $\times$ 176 pixels).

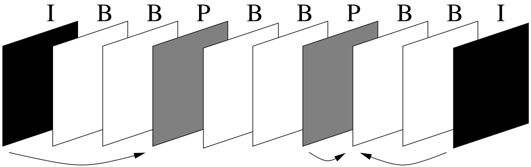
MPEG–1

Organization of the bitstream



MPEG–1

- MPEG–1 allows three types of **frames**:
  - **Type I**: encoding is similar to that of JPEG. These frames serve as entry points for random access.
  - **Type P**: frames encoded in predictive mode, using as reference previous frames of type I or P.
  - **Type B**: frames encoded in predictive mode, using both reference frames from the past and from the future (of type I or P).
- The number of I, P and B frames composing a **group of frames** depends on the application.



MPEG–2

- MPEG–2 (1994) has been developed aiming applications such as
  - Transmission of television signals in standard definition formats (PAL, SECAM, NTSC).
  - High definition television (HDTV).
  - Electronic cinema.
  - Games and high quality multimedia applications.
  - ...
- Some characteristics of MPEG–2 video:
  - Bitrates up to 100 Mbps.
  - More choices in terms of spatial and temporal resolution.
  - Support for interlaced video (notion of even and odd field).
  - More possibilities for the chrominance sub-sampling.
  - More coding and quantization options.
  - Support for bitstream **scalability**.

H.263

- Initially (1993), the MPEG–4 group started developing a video coding standard for bitrates < 64 kbps, i.e., for **very low bitrates**.
- However, some time after, this line was reformulated into a much more ambitious objective: that of creating a standard for coding audiovisual objects.
- Due to the urgent need for a low bitrate standard (for example, for enabling video over the analog public telephone network or over wireless channels), the work was divided in two phases:
  - One, for the immediate development of a video coding standard for very low bitrates: recommendation H.263 (1995).
  - The other, directed to a more vast set of tools, originated the MPEG–4 standard.

H.263

- Recommendation H.263 specifies an algorithm for video coding, similar to that of H.261, for bitrates of about 22 kbps of a total of 28.8 kbps.
- The main differences between H.261 and H.263 are:
  - New formats available: sub-QCIF, 4CIF and 16CIF, in addition to those already supported by H.261, CIF and QCIF.
  - Possibility of using a motion vector per block as well as one motion vector per macroblock.
  - Half-pixel precision motion estimation and prediction of motion vectors.
  - Arithmetic coding.
  - PB-frames (bi-directional prediction, similar to that used in MPEG).

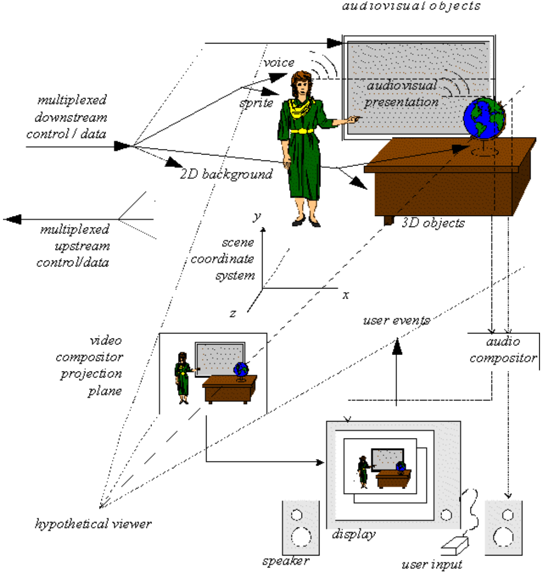
MPEG–4

- MPEG–4 (initial version in 1998) is a ISO/IEC standard providing tools for:
  - **Representing** audio, video or audiovisual data through **media objects** that can be natural (i.e., captured by a microphone or video camera) or synthetic (i.e., computer generated).
  - Describing the **composition** of these objects for creating composed objects and audiovisual scenes.
  - **Multiplexing and synchronizing** the data associated to the media objects, for transmission through the communication channels, providing an appropriate quality of service (QoS) to each object.
  - Enabling the **interaction** of the clients (receptor) with the audiovisual scene.

MPEG–4

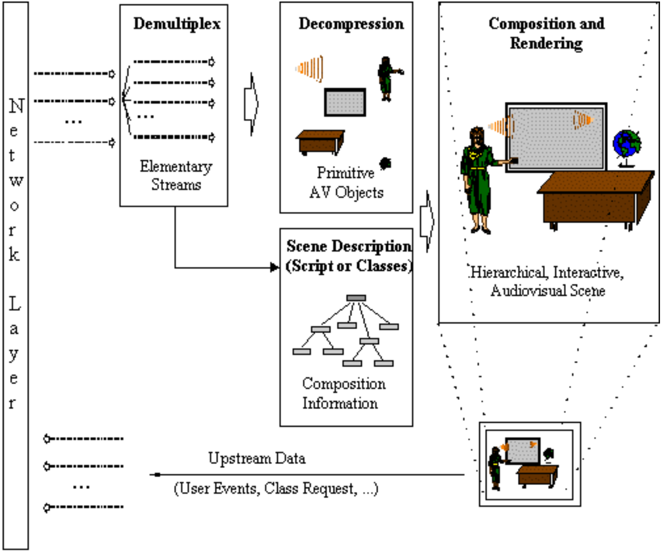
- MPEG–4 defines several primitive objects for representing **natural** and **synthetic** information, as well as **2D** and **3D** data.
- The **audiovisual scenes** are composed of these media objects, hierarchically organized:
    - Images (for example, a fixed background).
    - Video objects (for example, a person talking).
    - Audio objects (for example, the voice of the person, background music, . . .).
    - Text and graphics.
    - Synthetic talking heads and the corresponding text used by the speech synthesizer; animated synthetic bodies.
    - Synthetic sound.
    - . . .

MPEG–4



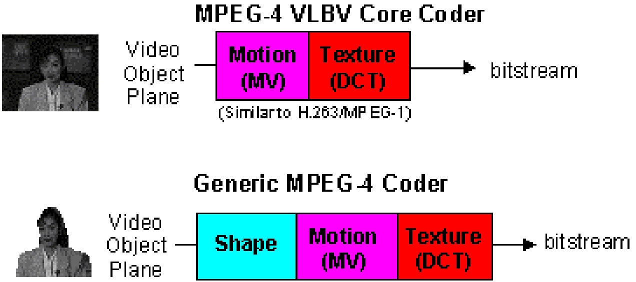
audiovisual objects
voice
sprite
audiovisual presentation
multiplexed downstream control / data
2D background
3D objects
multiplexed upstream control/data
y
scene coordinate system
x
z
user events
video compositor projection plane
audio compositor
hypothetical viewer
display
speaker
user input

MPEG–4


Network Layer
Demultiplex
Elementary Streams
Decompression
Primitive AV Objects
Scene Description (Script or Classes)
Composition Information
Composition and Rendering
Hierarchical, Interactive, Audiovisual Scene
Upstream Data
(User Events, Class Request, ...)

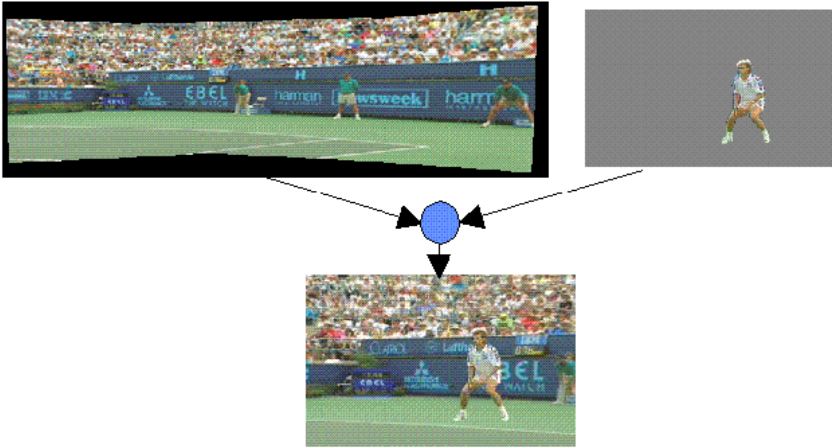
MPEG–4

- Conventional video coding is performed as in MPEG–1/2.
- In **content-based** coding, it is possible to encode regions with arbitrary shape, but, it this case, the shape of the object also needs to be efficiently represented.
- **Shape** is represented using a 8 bit transparency component or a binary mask.



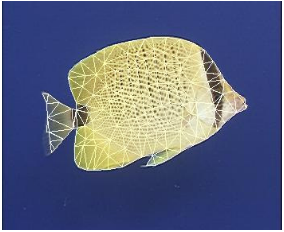
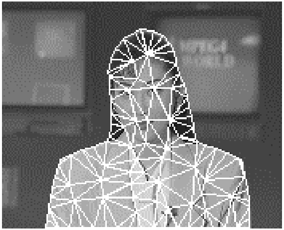
MPEG–4

- Sprites:



MPEG–4

- MPEG–4 supports **synthetic visual objects**:
  - Parametric description of **human heads and bodies** (also body animation in Version 2).
  - Parametric description of **static or dynamical meshes** with texture mapping.

  - **Scalable texture coding.**

H.264/AVC

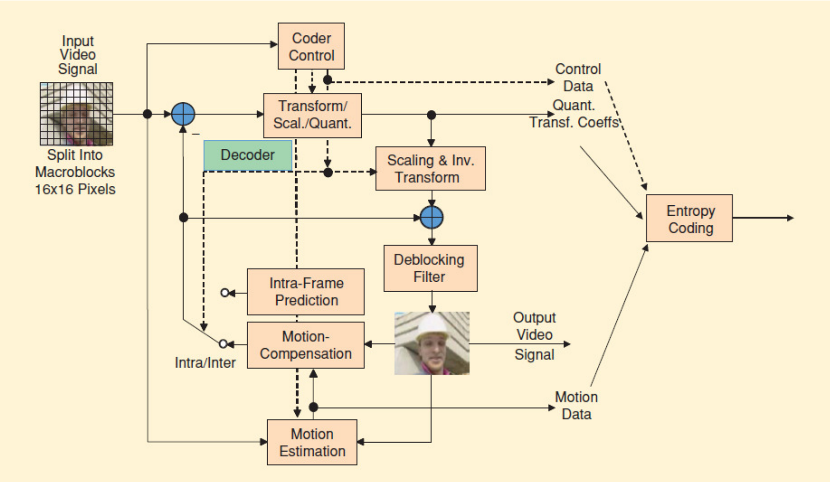
Overview

- H.264/AVC (Advanced Video Coding) was jointly developed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC MPEG.
- It was finalized in March 2003 and approved by the ITU-T in May 2003.
- H.264/AVC provides gains in compression efficiency of up to 50% over a wide range of bit rates and video resolutions compared to previous standards.
- The decoder complexity is about four times that of MPEG-2 and two times that of MPEG-4 Visual Simple Profile.

H.264/AVC

Overview

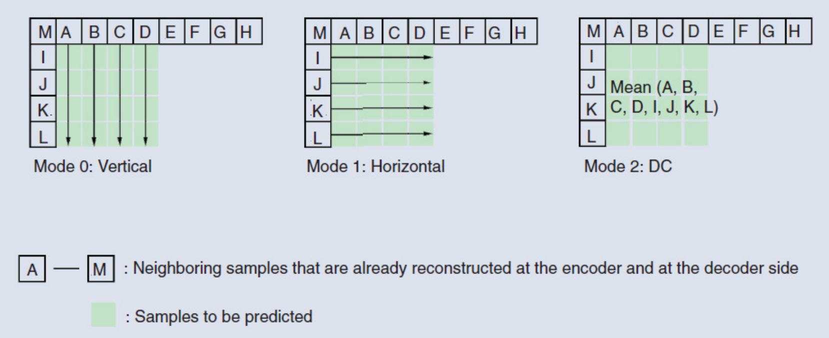
Block diagram of a typical encoding process of H.264/AVC



H.264/AVC

Intra prediction

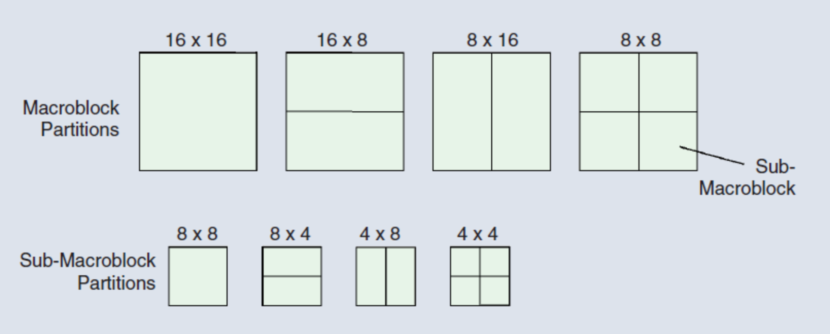
Three out of nine possible intra prediction modes for INTRA_4×4



H.264/AVC

Motion-compensated prediction

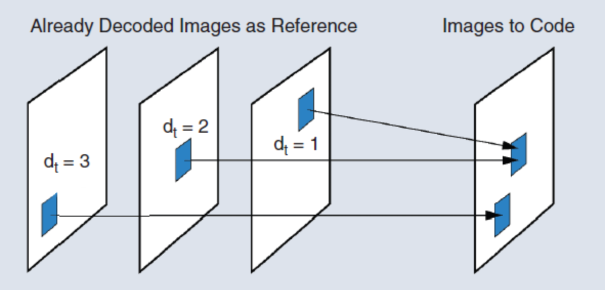
Partition of macroblock/sub-macroblock for motion-compensation



H.264/AVC

Motion-compensated prediction

Motion-compensated prediction with multiple reference images



H.264/AVC

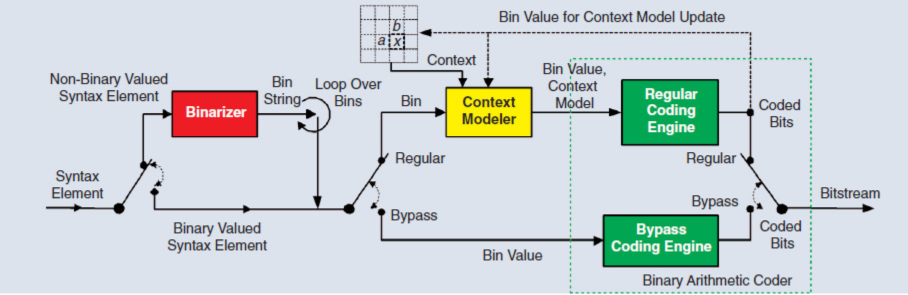
Entropy coding

- H.264/AVC provides two methods for entropy coding:
  - CAVLC, a low-complexity technique based on context-adaptive sets of variable length codes.
  - CABAC, a context-based adaptive binary arithmetic encoder.
- By incorporating context modeling, both methods offer a high degree of adaptation to the underlying source.
- CAVLC relies on 32 different VLCs. For typical coding conditions, it is 2–7% better than conventional codes.
- Typically, CABAC provides bit rate reductions of 5–15% compared to CAVLC.

H.264/AVC

CABAC

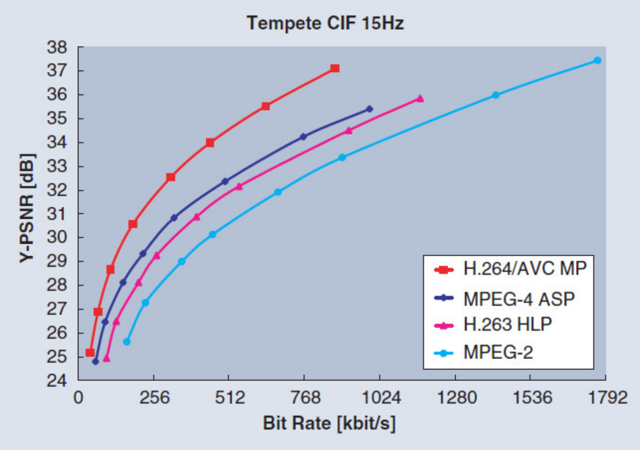
Context-based adaptive binary arithmetic coding



H.264/AVC

Performance

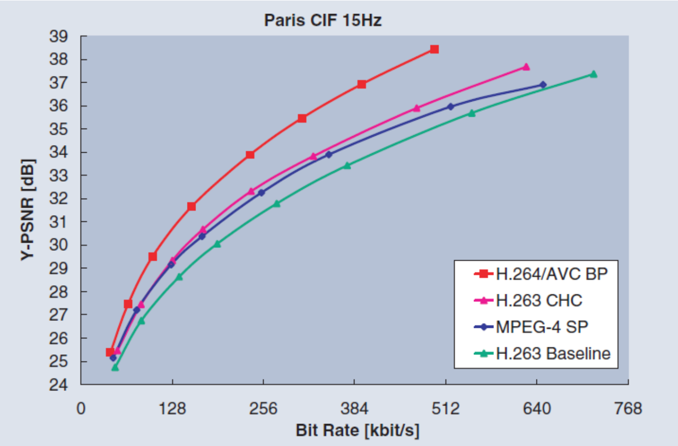
Video streaming application



H.264/AVC

Performance

Video conferencing application



H.265/HEVC

Overview

- H.265/HEVC (High Efficiency Video Coding) was (again) the result of a collaboration between the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC MPEG.
- It is also known as the MPEG-H Part 2 and the first version was finalized in 2013.
- H.265/HEVC can provide gains in compression efficiency of about 50%, when compared to H.264/AVC.
- This is mostly attained by further exploring existing techniques, but at a cost of increasing the complexity of the encoder.
- As with H.264/AVC, H.265/HEVC is dependent of a considerable number of patents, which is preventing its wide use. . .

H.265/HEVC

Block diagram